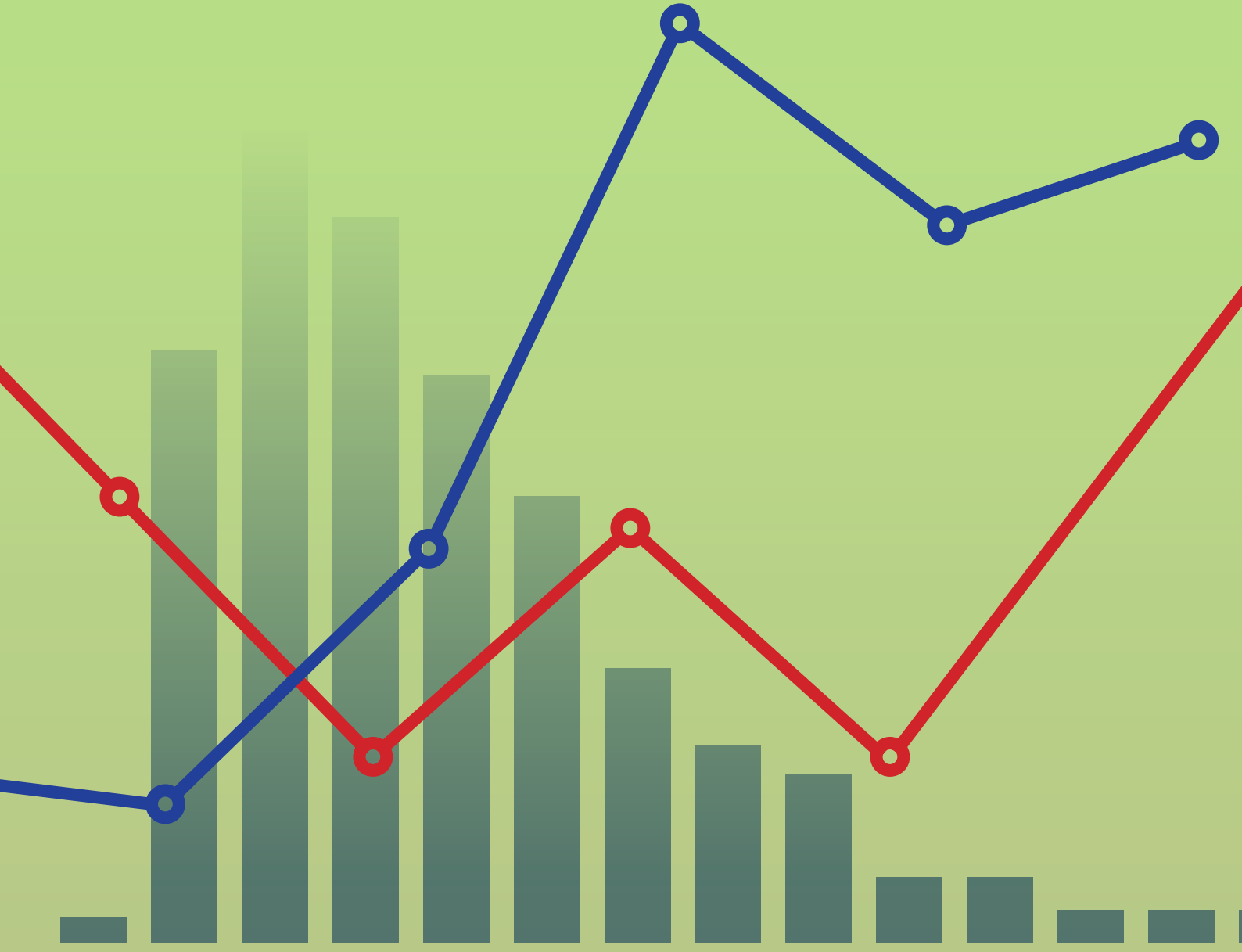
Μιχαήλ Τσαγρής
Επίκουρος Καθηγητής Πανεπιστημίου Κρήτης

Μίνως Κουκουριτάκης
Αναπληρωτής Καθηγητής Πανεπιστημίου Κρήτης

Στατιστική με τη χρήση των IBM SPSS 26 και Eviews 11



ΚΑΛΛΙΠΟΣ
σύγχρονες ακαδημαϊκές εκδόσεις

Εθνικό Πρόγραμμα Ανάπτυξης 2021-2025

ΜΙΧΑΗΛ ΤΣΑΓΡΗΣ
Επίκουρος Καθηγητής Πανεπιστημίου Κρήτης

ΜΙΝΩΣ ΚΟΥΚΟΥΡΙΤΑΚΗΣ
Αναπληρωτής Καθηγητής Πανεπιστημίου Κρήτης

Στατιστική με τη χρήση των IBM SPSS 26 και Eviews 11

ΚΑΛΛΙΠΟΣ
ανοικτές ακαδημαϊκές εκδόσεις

Στατιστική με τη χρήση των IBM SPSS 26 και Eviews 11

Συγγραφή

Μιχαήλ Τσαγρής

Μίνως Κουκουριτάκης

Συντελεστές έκδοσης

Γλωσσική Επιμέλεια: Γεωργία Τριανταφυλλίδου

Γραφιστική Επιμέλεια: Ελένη Τσακμάκη

Κεντρική ομάδα υποστήριξης

Γραφιστικός Έλεγχος: Αλεξάνδρα Θεοδωράκη

Βιβλιοθηκονομική Επεξεργασία: Έλενα Αδαμοπούλου

ΚΑΛΛΙΠΟΣ

Εθνικό Μετσόβιο Πολυτεχνείο

Ηρώων Πολυτεχνείου 9, 15780 Ζωγράφου

www.kallipos.gr

ISBN: 978-618-5667-05-4

Βιβλιογραφική Αναφορά: Τσαγρής, Μ., & Κουκουριτάκης, Μ. (2022). *Στατιστική με τη χρήση των IBM SPSS 26 και Eviews 11* [Προπτυχιακό εγχειρίδιο]. Κάλλιπος, Ανοικτές Ακαδημαϊκές Εκδόσεις. http://dx.doi.org/10.57713/kallipos-68

Στη Βάσω

Περιεχόμενα

Περίληψη

Το βιβλίο αυτό αποτελείται από 8 κεφάλαια. Στο πρώτο κεφάλαιο περιέχονται σύντομες αναφορές σε διάφορα θέματα της στατιστικής, όπως ένας σύντομος ορισμός της στατιστικής συνοδευόμενος από μία ιστορική αναδρομή του όρου. Τονίζεται, επίσης, η σημαντικότητα της στατιστικής και παρουσιάζονται οι δύο βασικές σχολές της στατιστικής, η κλασική και Μπεϋζιανή (Bayesian). Στη συνέχεια, παρουσιάζονται συνοπτικά τα διάφορα είδη μεταβλητών κάποιων βασικών δειγματοληπτικών τεχνικών, κάποια είδη μελετών και κάποια είδη δεδομένων. Οι πληροφορίες αυτές στοχεύουν στην ενημέρωση, αλλά και στην κατανόηση κάποιων βασικών αρχών της στατιστικής από τον/την αναγνώστη/-τρια.

Στο δεύτερο κεφάλαιο παρουσιάζονται κάποιες βασικές λειτουργίες του IBM SPSS 26 και του Eviews 11, προκειμένου να εξοικειωθεί ο/η αναγνώστης/-τρια με αυτές. Το τρίτο κεφάλαιο είναι αφιερωμένο στην περιγραφική στατιστική και παρουσιάζει κάποια περιγραφικά μέτρα για διάφορα είδη μεταβλητών, καθώς και τις γραφικές απεικονίσεις αυτών.

Η επαγωγική στατιστική αναλύεται στο τέταρτο κεφάλαιο, στο οποίο παρουσιάζονται κάποιοι βασικοί έλεγχοι υποθέσεων, όπως ο έλεγχος κανονικότητας, ο έλεγχος των μέσων, οι συντελεστές συσχέτισης και τα διαστήματα εμπιστοσύνης. Στο πέμπτο κεφάλαιο αναλύονται η τεχνική της απλής και πολλαπλής γραμμικής παλινδρόμησης, οι παραβιάσεις των υποθέσεων της παλινδρόμησης και οι μέθοδοι επιλογής μεταβλητών. Στο έκτο κεφάλαιο παρουσιάζονται η τεχνική της ανάλυσης διακύμανσης με έναν και δύο παράγοντες, καθώς και η περίπτωση των επαναλαμβανόμενων μετρήσεων.

Στο έβδομο κεφάλαιο περιλαμβάνονται προχωρημένα θέματα στατιστικής, όπως η μη γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση, η χαρακτηριστική λειτουργική καμπύλη και η διαχωριστική ανάλυση. Τέλος, το όγδοο κεφάλαιο είναι προσανατολισμένο σε θέματα κοινωνικής στατιστικής, όπως είναι η ανάλυση αξιοπιστίας ερωτηματολογίου και η παραγοντική ανάλυση.

Πρόλογος

Το βιβλίο αυτό γράφτηκε κυρίως για τους μη στατιστικούς, οι οποίοι ασχολούνται με τη στατιστική και ειδικότερα με την εφαρμοσμένη στατιστική. Στο επίκεντρο του βιβλίου βρίσκεται η χρήση των προγραμμάτων SPSS και Eviews για την υλοποίηση στατιστικών αναλύσεων, ενώ η θεωρία που εμπεριέχεται σε αυτό αποτελεί τη βάση της ανάλυσης. Πιστεύουμε ότι το συγκεκριμένο βιβλίο θα βοηθήσει τον/την αναγνώστη/-τρια να εκτελέσει μία βασική στατιστική ανάλυση, την οποία στη συνέχεια θα μπορέσει να επεκτείνει ερευνώντας λίγο περισσότερο. Δεν θα επεκταθούμε στο σημείο αυτό περισσότερο, παρά μόνο θα αναφέρουμε ότι αυτή η έκδοση είναι μία αναθεωρημένη έκδοση της αρχικής του Απριλίου 2008 (που αφορούσε μόνο το SPSS 15), όπου έχουν διορθωθεί πολλά λάθη. Θα ήμασταν υπόχρεοι για οποιαδήποτε υπόδειξη σφάλματος στο παρόν κείμενο, ώστε να βελτιωθεί ακόμα περισσότερο.

Η δεύτερη έκδοση του βιβλίου αυτού ήταν τον Ιανουάριο του 2010 και αφορούσε ξανά μόνο το SPSS 15, στην οποία όμως είχε προστεθεί και το αγγλοελληνικό λεξικό στατιστικών όρων, ενώ η τρίτη έκδοση (και πάλι μόνο για το SPSS 15) ήταν τον Νοέμβριο του 2010. Στην τέταρτη έκδοση του βιβλίου (Ιούλιος 2011), η οποία αφορούσε το IBM SPSS 19, αναφερθήκαμε και στη μεθοδολογία bootstrap που αποτελεί μία σχετικά νέα και πολύ χρήσιμη τεχνική στη στατιστική. Πιθανότατα, όμως, η συγκεκριμένη τεχνική θα αργήσει να διδαχτεί στις κοινωνικές επιστήμες, εκτός ίσως από κάποια πανεπιστημιακά τμήματα στα οποία διδάσκεται μόνο το SPSS.

Στο IBM SPSS 19, εκτός από την ενσωμάτωση της τεχνικής bootstrap, έχει αλλάξει και το συνολικό περιβάλλον για τις μη παραμετρικές εκτιμήσεις. Στο βιβλίο αυτό, θα αναλύσουμε τις κλασικές επιλογές για τη μη παραμετρική στατιστική, αλλά ο/η αναγνώστης/-τρια μπορεί, αν θέλει, να εστιάσει και στην άλλη επιλογή για τις ίδιες αναλύσεις. Η τελευταία οδηγεί στο ίδιο σημείο τερματισμού, αλλά ακολουθεί άλλη, ίσως πιο χρονοβόρα, πορεία. Στην πέμπτη έκδοση του βιβλίου (Ιανουάριος 2013), η οποία και πάλι αφορά το IBM SPSS 19, προστέθηκε ο έλεγχος του McNemar για την αλλαγή κατάστασης μιας δίτιμης μεταβλητής, καθώς και η καμπύλη λειτουργικού χαρακτηριστικού δέκτη (ROC curve).

Πλέον βρισκόμαστε στην έκδοση IBM SPSS 26, στην οποία έχουν αλλάξει κάποια γραφικά (λόγω και της εξέλιξης των Windows) και μερικά μενού, αλλά, όπως θα φανεί στην πορεία, οι αλλαγές αυτές είναι μικρές. Στην ανάλυση για το IBM SPSS 26 επεκτάθηκαν τα ήδη υπάρχοντα κεφάλαια, προστέθηκε ένα νέο κεφάλαιο, ενώ έγινε και ενδελεχής γραμματικός έλεγχος. Η πιο σημαντική συνεισφορά της παρούσας έκδοσης του βιβλίου είναι η προσθήκη της ανάλυσης για το Eviews 11 της IHS Markit, το οποίο απευθύνεται κυρίως σε οικονομολόγους και είναι περισσότερο εξειδικευμένο σε σχέση με το IBM SPSS. Ελπίζουμε ότι αυτή η επαυξημένη έκδοση του βιβλίου θα έχει αμελητέα λάθη και θα βοηθήσει σε σημαντικό βαθμό τον/την εκάστοτε αναγνώστη/-τρια. Για περισσότερες σημειώσεις στατιστικής (οι οποίες περιλαμβάνουν το SPSS) ο/η αναγνώστης/-τρια μπορεί να επισκεφτεί το website statlink.tripod.com

Τσαγρής Μιχαήλ και Κουκουριτάκης Μίνως

Νοέμβριος 2021

Κεφάλαιο 1 Εισαγωγή

Σύνοψη

Το κεφάλαιο αυτό αποτελείται από επτά σύντομες ενότητες, στις οποίες αναλύεται η σημαντικότητα της στατιστικής και παρουσιάζονται οι δύο βασικές προσεγγίσεις που έχουν διαμορφωθεί τις τελευταίες δεκαετίες. Στη συνέχεια, περιγράφονται τα είδη των μεταβλητών που απαντώνται στην πράξη, τα τέσσερα βασικά είδη δειγματοληπτικών τεχνικών, κάποια βασικά είδη μελετών και τέλος, κάποια είδη δεδομένων. Η σημασία του κεφαλαίου αυτού έγκειται στο να καταστεί ο/η αναγνώστης/-τρια ενήμερος/-η για κάποιες βασικές έννοιες της στατιστικής.

Προαπαιτούμενη γνώση

Δεν χρειάζονται εξειδικευμένες γνώσεις για την κατανόηση του κεφαλαίου.

1.1 Σύντομη εισαγωγή στη Στατιστική

Υπάρχουν δύο ενδεχόμενα σχετικά με την προέλευση του όρου «στατιστική». Το ένα ενδεχόμενο είναι ότι προέρχεται από την αρχαία ελληνική λέξη «στατίζω» που σημαίνει τοποθετώ, ταξινομώ, συμπεραίνω. Το άλλο ενδεχόμενο είναι ότι προέρχεται από τη λατινική λέξη “status” που σημαίνει πολιτεία, κράτος. Η λέξη αυτή αρχικά χρησιμοποιήθηκε για δεδομένα σχετικά με τον πληθυσμό μίας χώρας. Έρευνες έχουν δείξει ότι η αρχαιότερη συλλογή στοιχείων έγινε στην Κίνα το 2238 π.Χ. και αφορά μία απογραφή του πληθυσμού, η οποία διεξήχθη υπό την αυτοκρατορία του Yao. Υπάρχουν ενδείξεις ότι απογραφές είχαν διενεργηθεί και από άλλους λαούς κατά την αρχαιότητα, όπως από τους Αιγύπτιους και τους Πέρσες. Γνωρίζουμε, επίσης, ότι απογραφή διεξήχθη και από τον αυτοκράτορα της Ρώμης (Καίσαρα) Οκτάβιο ή Οκταβιανό Αύγουστο κατά την περίοδο γέννησης του Χριστού. Επίσης, ο όρος «στατιστική» αναφέρεται στο έργο του Ξενοφώντος «Απομνημονεύματα», καθώς και στο έργο του Πλάτωνα «Πολιτεία».

Η στατιστική με την πάροδο των χρόνων αναπτύχθηκε με αργούς ρυθμούς στην αρχή και ραγδαία από τα τέλη του 19^{ου} αιώνα και μετά, προκειμένου να φτάσει στο σήμερα. Η στατιστική είναι η επιστήμη που ασχολείται με τη συλλογή δεδομένων, την περιγραφή τους και την εξαγωγή τεκμηριωμένων αποτελεσμάτων με τη χρήση επιστημονικά αποδεκτών τεχνικών. Αν θέλαμε να δώσουμε έναν άλλο ορισμό στον όρο «στατιστική», θα επιλέγαμε αυτόν που έδωσε ο πατέρας της σύγχρονης στατιστικής, Ronald Fisher (1890-1962):

Στατιστική είναι ένα σύνολο αρχών και μεθοδολογιών για:

- τον σχεδιασμό της διαδικασίας συλλογής δεδομένων,
- τη συνοπτική και αποτελεσματική παρουσίασή τους,
- την ανάλυση και εξαγωγή αντίστοιχων συμπερασμάτων.

Οι βασικές μορφές της στατιστικής είναι αυτές της περιγραφικής στατιστικής και της επαγωγικής στατιστικής. Η πρώτη ασχολείται με την περιγραφή των δεδομένων του δείγματος και η δεύτερη με την εξαγωγή χρήσιμων συμπερασμάτων για τον πληθυσμό.

Στη σημερινή εποχή έχουν κάνει την εμφάνισή τους η μηχανική μάθηση και η επιστήμη δεδομένων, οι οποίες είναι άρρηκτα συνδεδεμένες -κατά την προσωπική μας άποψη- με τη στατιστική. Η μηχανική μάθηση αναπτύχθηκε στο πλαίσιο της επιστήμης των υπολογιστών και αποτελεί ένα σύνολο εργαλείων που έχουν ως απώτερο σκοπό την ανάλυση των δεδομένων, συνήθως όμως χωρίς το πνεύμα της στατιστικής συμπερασματολογίας. Η επιστήμη δεδομένων προέκυψε πιο πρόσφατα ως ένας πιο γενικός όρος, ο οποίος θα μπορούσαμε να πούμε ότι περικλείει τη στατιστική, τη μηχανική μάθηση, τα εφαρμοσμένα μαθηματικά, καθώς και άλλες επιστήμες. Μία ειδοποιός διαφορά μεταξύ της στατιστικής και των προαναφερθεισών

ορολογιών είναι ο σκοπός της καθεμίας επιστήμης. Η στατιστική εξετάζει το είδος της σχέσης μεταξύ κάποιων μεγεθών, καθώς και τον τρόπο επίδρασης κάποιων μεγεθών πάνω σε κάποια άλλα, ενώ η μηχανική μάθηση, για παράδειγμα, αξιοποιεί τη σχέση μεταξύ κάποιων μεγεθών, προκειμένου να κάνει κάποια πρόβλεψη, αδιαφορώντας για το είδος της σχέσης αυτών.

1.2 Η σημαντικότητα της Στατιστικής

Η στατιστική επιστήμη αποτελεί εργαλείο για πολλές άλλες επιστήμες, ενώ πολλές φορές η ορολογία προσαρμόζεται στο εκάστοτε επιστημονικό πεδίο. Σε πολλές περιπτώσεις μάλιστα θα μπορούσε να ισχυριστεί κάποιος πως αναπτύσσεται κάπως ανεξάρτητα από τον κλάδο της καθαυτής στατιστικής.

Για παράδειγμα, έχει ενσωματωθεί πλήρως στην επιστήμη των οικονομικών και ονομάζεται οικονομετρία. Ουσιαστικά, η μαθηματικο-οικονομική και η στατιστική ανάλυση χρησιμοποιούνται από κοινού με απώτερο σκοπό την εμπειρική εκτίμηση των «αφηρημένων» σχέσεων της οικονομικής θεωρίας. Η αφετηρία, όμως, είναι η ύπαρξη ή η κατασκευή ενός υποδείγματος οικονομικής θεωρίας, το οποίο εκτιμάται εμπειρικά με στατιστικά εργαλεία και ερμηνεύεται με οικονομικούς όρους. Στην κοινωνιολογία ο όρος που χρησιμοποιείται είναι η κοινωνική στατιστική και η λογική είναι παρόμοια με αυτή της οικονομικής επιστήμης. Στην ψυχολογία η στατιστική αποτελεί αναπόσπαστο κομμάτι της έρευνας, ενώ η υπολογιστική γλωσσολογία χρησιμοποιεί πολλές φορές στατιστικές μεθόδους. Επίσης, η στατιστική συναντάται και στις πολιτικές επιστήμες, κυρίως για την ανάλυση συμπεριφορών ψηφοφόρων.

Στην ιατρική και στη φαρμακευτική ο δόκιμος όρος είναι βιοστατιστική και ο ρόλος της είναι πολύ σημαντικός στις μελέτες για την αποτελεσματικότητα των φαρμάκων και των εμβολίων. Εφαρμογές της στατιστικής υπάρχουν και στη βιοπληροφορική και στην υπολογιστική βιολογία, προκειμένου να αναλυθούν τα βιολογικά δεδομένα. Προφανώς, η λίστα με τα επιστημονικά πεδία είναι μεγάλη και δεν περιορίζεται στα προαναφερθέντα πεδία. Απόδειξη αποτελεί το γεγονός ότι το μάθημα της στατιστικής εμφανίζεται σε πάρα πολλά προγράμματα πανεπιστημιακών ιδρυμάτων, καθώς και πιστοποιήσεων.

1.3 Σύντομη αναφορά στις δύο σχολές της Στατιστικής

Οι βασικές σχολές ή προσεγγίσεις της στατιστικής είναι δύο: η σχολή της κλασικής στατιστικής (frequentist approach) και η σχολή της Μπεϋζιανής στατιστικής (Bayesian statistics). Η κλασική στατιστική, της οποίας πατέρας θεωρείται ο Ronald Fisher, αναπτύχθηκε στις αρχές του $20^{ού}$ αιώνα και βασίζεται στη λογική της πιθανοφάνειας.

Ο θεμέλιος λίθος της κλασικής στατιστικής είναι ότι μία παράμετρος θ (η οποία δεν είναι γνωστή και θέλουμε να την εκτιμήσουμε), χρησιμοποιείται περισσότερο σαν να είναι μία σταθερά και όχι ως μια τυχαία μεταβλητή. Αυτό, όμως, οδηγεί σε πολλά προβλήματα ερμηνείας. Για παράδειγμα, όταν υπολογίζουμε ένα 95% διάστημα εμπιστοσύνης και το βρίσκουμε ίσο με [0.08,0.12], εννοούμε πως υπάρχει 95% πιθανότητα αυτή η παράμετρος θ να βρίσκεται μεταξύ του 0.08 και του 0.12. Όμως, σε αυτήν την περίπτωση υπάρχει πρόβλημα ερμηνείας, καθώς το θ δεν είναι τυχαίο: είτε θα ανήκει στο διάστημα αυτό είτε δεν θα ανήκει. Σύμφωνα με την κλασική στατιστική, το μόνο τυχαίο στοιχείο στο υπόδειγμα πιθανότητας είναι τα δεδομένα που έχουμε συλλέξει. Οπότε, η σωστή ερμηνεία του διαστήματος εμπιστοσύνης είναι πως, αν επαναλάβουμε τη διαδικασία πολλές φορές και κατασκευάσουμε πολλά 95% διαστήματα εμπιστοσύνης, αναμένουμε ότι το 95% των κατασκευασμένων διαστημάτων θα συμπεριλαμβάνουν την παράμετρο θ. Όλες οι συμπερασματολογίες που βασίζονται στην κλασική θεωρία έχουν αυτήν την ερμηνεία, παρόλο που στην πράξη έχουμε μόνο ένα διάστημα εμπιστοσύνης να ερμηνεύσουμε.

Η συμπερασματολογία κατά Bayes αναπτύχθηκε τις τελευταίες δεκαετίες, ενώ η ραγδαία εξέλιξη των υπολογιστών βοήθησε στην ανάπτυξή της, καθώς σε δύσκολα προβλήματα απαιτούνται ισχυροί υπολογιστικοί πόροι. Το πλαίσιο στο οποίο κινείται η συγκεκριμένη μεθοδολογία είναι παρόμοιο με αυτό της κλασικής στατιστικής, με μία θεμελιώδη όμως διαφορά: το θ χρησιμοποιείται ως μια τυχαία ποσότητα και όχι ως μία σταθερά. Αν και η διαφορά αυτή μπορεί να μην φαίνεται και τόσο ουσιαστική, οδηγεί σε μία τελείως διαφορετική προσέγγιση ως προς την ερμηνεία, σε σχέση με αυτήν την κλασικής στατιστικής. Στην ουσία, η συμπερασματολογία μας θα βασιστεί στην πιθανότητα της κατανομής της παραμέτρου δεδομένων

των δεδομένων και όχι των δεδομένων δεδομένης της παραμέτρου. Σε πολλές περιπτώσεις, αυτό οδηγεί σε περισσότερο φυσικά συμπεράσματα σε σχέση με την κλασική στατιστική. Για να μπορέσει όμως να επιτευχθεί αυτό, θα πρέπει να καθορίσουμε τη λεγόμενη *a-priori* κατανομή (*prior probability distribution*), η οποία αντιπροσωπεύει τις «πεποιθήσεις» μας για την κατανομή του θ, προτού αποκτήσουμε οποιαδήποτε πληροφορία για τα δεδομένα μας. Ο συνδυασμός της *a-priori* κατανομής και της πιθανοφάνειας μας οδηγεί στην *a-posteriori* κατανομή (*posterior probability distribution*), πάνω στην οποία γίνεται η συμπερασματολογία για την παράμετρο θ.

Η ιδέα της *a-priori* κατανομής της παραμέτρου θ αποτελεί και την «καρδιά» της θεωρίας κατά Bayes, και ανάλογα με το αν μιλάμε με έναν υπέρμαχο ή με έναν αντίπαλο της συγκεκριμένης μεθοδολογίας, η *a-priori* κατανομή μπορεί να αποτελέσει το μεγαλύτερο πλεονέκτημα ή το σοβαρότερο μειονέκτημα σε σχέση με την κλασική στατιστική. Η εξήγηση είναι απλή: η επιλογή της *a-priori* κατανομής απαιτεί ιδιαίτερη τεχνική, καθώς επηρεάζει την *a-posteriori* κατανομή και συνεπώς και τη συμπερασματολογία, η οποία δεν είναι καθόλου εύκολη.

1.4 Σύντομη αναφορά στα είδη των μεταβλητών

Τα μεγέθη ή τα χαρακτηριστικά τα οποία μετράμε ονομάζονται μεταβλητές. Οι μεταβλητές χωρίζονται σε δύο μεγάλες κατηγορίες: στις ποσοτικές και στις ποιοτικές. Οι ποσοτικές αναφέρονται σε μεγέθη τα οποία είναι ποσοτικά, δηλαδή αριθμητικά, ενώ οι ποιοτικές σε μεγέθη που δεν είναι αριθμητικά. Ποσοτικά μεγέθη είναι, για παράδειγμα, το βάρος, το ύψος, η θερμοκρασία, η ηλικία, το μηνιαίο εισόδημα, η ετήσια παραγωγή κάποιου αγροτικού προϊόντος κλπ. Οι ποσοτικές μεταβλητές με τη σειρά τους χωρίζονται σε συνεχείς και διακριτές. Οι συνεχείς μεταβλητές μπορούν να πάρουν οποιαδήποτε τιμή σε ένα περιορισμένο ή απεριόριστο εύρος τιμών, όπως για παράδειγμα, τα προαναφερθέντα μεγέθη. Από την άλλη πλευρά, οι διακριτές μεταβλητές παίρνουν, όπως λέει και το όνομά τους, διακριτές τιμές. Για παράδειγμα, όταν μετράμε την ηλικία σε έτη, έχουμε αυστηρά ακέραιους αριθμούς. Ένα άλλο παράδειγμα είναι ο αριθμός των εργαζομένων σε μία εταιρεία.

Ένας περαιτέρω διαχωρισμός των ποσοτικών μεταβλητών έχει να κάνει με την κλίμακα μέτρησης. Η κλίμακα διαστήματος αναφέρεται σε ποσοτικές μετρήσεις, για τις οποίες κάποιο νόημα έχει μόνο η απόσταση. Στη θερμοκρασία, για παράδειγμα, οι 20 βαθμοί Κελσίου είναι κατά 10 βαθμούς περισσότεροι από τους 10 βαθμούς Κελσίου, όμως δεν είναι σωστό να πούμε ότι θερμοκρασία των 20 βαθμών είναι διπλάσια από τη θερμοκρασία των 10 βαθμών. Ένα άλλο παράδειγμα είναι αν οι βαθμοί δύο φοιτητών σε ένα μάθημα είναι 6 και 8. Στην περίπτωση αυτή, μπορούμε να πούμε ότι η διαφορά μεταξύ των δύο βαθμών είναι 2 βαθμοί, αλλά όχι ότι ο δεύτερος βαθμός είναι κατά 33.3% μεγαλύτερος από τον πρώτο. Στην κλίμακα λόγου, όμως, μπορούμε να υπολογίσουμε τον λόγο μεταξύ δύο αριθμών. Για παράδειγμα, αν το μηνιαίο εισόδημα δύο εργαζομένων είναι 800 και 1000 ευρώ, μπορούμε να πούμε ότι ο δεύτερος αμείβεται με 200 ευρώ παραπάνω από τον πρώτο, αλλά και ότι λαμβάνει 25% μεγαλύτερη αμοιβή από τον πρώτο.

Από την άλλη πλευρά, ποιοτικά μεγέθη είναι, για παράδειγμα, το φύλο, η ομάδα αίματος, η κλίμακα της φορολογίας («$1^η$ κλίμακα εισοδήματος», «$2^η$ κλίμακα εισοδήματος», «$3^η$ κλίμακα εισοδήματος») και ο βαθμός ικανοποίησης του/της φοιτητή/-τριας σε ένα μάθημα (π.χ. «μη ικανοποιημένος/-η», «αδιάφορος/-η», «ικανοποιημένος/-η»). Προφανώς, σε κανένα παράδειγμα δεν μπορούμε να κάνουμε αριθμητικές πράξεις. Δεν μπορούμε, για παράδειγμα, να υπολογίσουμε τον μέσο όρο ανάμεσα σε 40 άνδρες και 50 γυναίκες. Οι δύο πρώτες μεταβλητές (φύλο και ομάδα αίματος) ονομάζονται και ονομαστικές, καθώς μετριούνται στην ονομαστική κλίμακα (δηλαδή, δεν υπάρχει κάποια διάταξη ανάμεσά τους). Για παράδειγμα, δεν μπορούμε να κατατάξουμε την ομάδα αίματος. Δεν μπορούμε να πούμε ότι η ομάδα Α βρίσκεται υψηλότερα από την ομάδα Β. Στην περίπτωση, όμως, της κλίμακας φορολογίας ή του βαθμού ικανοποίησης σε ένα μάθημα υπάρχει η έννοια της διάταξης, οπότε μιλάμε για διατακτικές μεταβλητές, καθώς μετριούνται στη διατακτική κλίμακα. Υπάρχει νόημα διάταξης, εφόσον η ικανοποίηση του/της φοιτητή/-τριας σε ένα μάθημα έχει διατακτικό χαρακτήρα. Θα πρέπει να δοθεί, ωστόσο, προσοχή στο σημείο αυτό, καθώς η έννοια της απόστασης μεταξύ των βαθμών ικανοποίησης δεν είναι η ίδια μεταξύ μη ικανοποιημένου/-ης φοιτητή/-τριας-αδιάφορου/-ης φοιτητή/-τριας και αδιάφορου/-ης φοιτητή/-τριας-

ικανοποιημένου/-ης φοιτητή/-τριας. Ο λόγος που η έννοια της απόστασης δεν υφίσταται είναι ότι οι συγκεκριμένες μεταβλητές, τα μεγέθη, οι μετρήσεις δεν είναι αριθμοί.

1.5 Σύντομη αναφορά στις δειγματοληπτικές τεχνικές

Στην ενότητα αυτή θα αναφερθούμε συνοπτικά στη δειγματοληψία, δηλαδή στη διαδικασία συλλογής δεδομένων από έναν πληθυσμό. Το σύνολο των δεδομένων που θα συλλεχθούν, το οποίο προφανώς θα αποτελεί ένα υποσύνολο του πληθυσμού, ονομάζεται δείγμα. Ένα δείγμα καλείται αντιπροσωπευτικό, όταν όλα τα στοιχεία του πληθυσμού έχουν την ίδια πιθανότητα επιλογής. Γενικά, υπάρχουν διάφορες τεχνικές με τις οποίες μπορούμε να αντλήσουμε δεδομένα από έναν πληθυσμό. Όμως, οι κύριες δειγματοληπτικές τεχνικές είναι τέσσερις:

- Η απλή τυχαία δειγματοληψία είναι η πιο απλή περίπτωση. Επιλέγουμε τυχαία στοιχεία (ή μονάδες) από το σύνολο του πληθυσμού.
- Η στρωματοποιημένη δειγματοληψία είναι η περίπτωση κατά την οποία χωρίζουμε τον πληθυσμό σε στρώματα και μετά επιλέγουμε τυχαία στοιχεία από κάθε στρώμα.
- Η δειγματοληψία κατά ομάδες είναι μία τεχνική δειγματοληψίας στην οποία χωρίζουμε τον πληθυσμό σε πολλές ομάδες (όχι στρώματα), όπου η κάθε ομάδα περιέχει ένα πλήθος στοιχείων. Επιλέγουμε τυχαία ομάδες από το σύνολο των ομάδων και συμπεριλαμβάνουμε στο δείγμα όλα τα στοιχεία των επιλεγμένων ομάδων.
- Η τέταρτη περίπτωση είναι η συστηματική δειγματοληψία. Ας υποθέσουμε ότι έχουμε στη διάθεσή μας έναν μακρύ κατάλογο με όλα τα στοιχεία του πληθυσμού αριθμημένα. Στην περίπτωση αυτή μπορούμε να εφαρμόσουμε το εξής: διαλέγουμε ένα στοιχείο στην αρχή του καταλόγου, έστω το στοιχείο που βρίσκεται στην $4^{η}$ γραμμή του καταλόγου. Τα επόμενα στοιχεία θα επιλεγούν με ένα συγκεκριμένο βήμα, για παράδειγμα το 10. Δηλαδή, θα επιλεγούν το $14^{ο}$ στοιχείο, το $24^{ο}$ στοιχείο, το $34^{ο}$ στοιχείο κοκ.

Όπως είναι φυσικό, οι διάφορες τεχνικές μπορούν να συνδυαστούν και να προκύψουν πιο σύνθετα δειγματοληπτικά σχήματα. Οι περιπτώσεις που αναφέρθηκαν παραπάνω είναι οι πιο απλές και συνάμα οι πιο δημοφιλείς τεχνικές δειγματοληψίας. Ωστόσο, στο σημείο αυτό πρέπει να κάνουμε οπωσδήποτε αναφορά στα είδη των πληθυσμών προς αποφυγή λανθασμένων συμπερασμάτων: στον αντικειμενικό πληθυσμό (target population), στον υπό μελέτη πληθυσμό (study population) και στο δειγματοληπτικό πλαίσιο (sampling frame), καθώς και στην έννοια του δειγματοληπτικού σφάλματος.

Ο αντικειμενικός πληθυσμός είναι το σύνολο των ατόμων ή των στοιχείων, των οποίων ένα ή περισσότερα χαρακτηριστικά θέλουμε να εξετάσουμε. Ο υπό μελέτη πληθυσμός είναι ένα υποσύνολο συνήθως του αντικειμενικού πληθυσμού (μπορεί όμως και να ταυτίζεται με τον τελευταίο σε διάφορες περιπτώσεις). Για παράδειγμα, ο αντικειμενικός πληθυσμός μιας μελέτης θα μπορούσε να είναι το σύνολο των Ελλήνων μαθητών του δημοτικού σχολείου, ενώ ο υπό μελέτη πληθυσμός να αφορά μόνο τους μαθητές της Αττικής. Το δειγματοληπτικό πλαίσιο είναι το σύνολο των ατόμων (ή στοιχείων) που έχουν πραγματικά δυνατότητα επιλογής στο δείγμα (η πηγή του δείγματος). Στο ίδιο παράδειγμα, οι μαθητές κάποιων απομακρυσμένων περιοχών της Αττικής δεν έχουν τη δυνατότητα να συμπεριληφθούν στο δείγμα λόγω κόστους. Ουσιαστικά, δηλαδή, το δειγματοληπτικό πλαίσιο αποτελεί μία υποδιαίρεση του υπό μελέτη πληθυσμού στο συγκεκριμένο παράδειγμα, ενώ μπορεί να ταυτίζεται με αυτόν σε άλλες περιπτώσεις.

Το δειγματοληπτικό σφάλμα είναι η διαφορά ανάμεσα στα αποτελέσματα μιας δειγματοληψίας και μιας απογραφής (100% δείγμα). Κατά την περίοδο των εκλογών, για παράδειγμα, είναι το περιθώριο σφάλματος που αναφέρουν οι εταιρείες δημοσκοπήσεων, όταν κάνουν πρόβλεψη για το ποσοστό ψήφων που θα συγκεντρώσει ένα πολιτικό κόμμα: συνήθως ±2,5 ποσοστιαίες μονάδες.

1.6 Σύντομη αναφορά στα είδη των ερευνών

Οι διάφορες στατιστικές μελέτες και έρευνες χωρίζονται σε δύο κατηγορίες. Η πρώτη κατηγορία είναι οι πειραματικές μελέτες, στις οποίες ο ερευνητής παρεμβαίνει ενεργητικά στον τρόπο σχεδιασμού του

πειράματος, του καθορισμού των ομάδων κλπ. Χαρακτηριστικό παράδειγμα τέτοιων μελετών είναι οι κλινικές δοκιμές στην ιατρική. Η δεύτερη κατηγορία είναι οι μελέτες παρατήρησης ή μη πειραματικές μελέτες, στις οποίες ο ερευνητής δεν παρεμβαίνει, αλλά απλά παρατηρεί και καταγράφει. Σε αυτήν την κατηγορία περιλαμβάνονται οι διατμηματικές μελέτες, οι προοπτικές μελέτες (όπου διαμήκεις-διαχρονικές, μελέτες κοόρτης, μελέτες παρακολούθησης ή follow-up studies αποτελούν εναλλακτικές ονομασίες αυτών των ερευνών) και οι αναδρομικές μελέτες ή μελέτες μαρτύρων-ασθενών (retrospective or case-control studies). Στις διατμηματικές μελέτες τα στοιχεία του δείγματος κατηγοριοποιούνται με βάση την έκθεσή τους σε μία νόσο, χωρίς να λαμβάνουμε υπόψη τον παράγοντα χρόνο. Από την άλλη πλευρά, οι προοπτικές μελέτες λαμβάνουν υπόψη τους τον παράγοντα χρόνο και έχουν ως βασικό χαρακτηριστικό ότι επιλέγεται μία ομάδα ανθρώπων, της οποίας η εξέλιξη παρατηρείται και καταγράφεται στο πέρασμα του χρόνου. Θα πρέπει να επισημάνουμε στο σημείο αυτό πως υπάρχουν έρευνες που διήρκησαν πολλά χρόνια. Χαρακτηριστικό παράδειγμα αποτελεί η έρευνα θνησιμότητας των Βρετανών ιατρών Doll & Hill (1964), στην οποία συμμετείχαν 40.637 ιατροί και είχε διάρκεια 10 ετών. Τέλος, οι αναδρομικές μελέτες ακολουθούν κατά κάποιο τρόπο μια αντίθετη μεθοδολογία σε σχέση με αυτή των προοπτικών μελετών. Οι συγκεκριμένες μελέτες εστιάζουν στο παρελθόν και κατηγοριοποιούν τα στοιχεία ανάλογα με κάποιο χαρακτηριστικό. Είναι προφανές ότι το είδος της μελέτης καθορίζει τόσο την τεχνική που θα υιοθετηθεί όσο και τον τρόπο ανάλυσης των δεδομένων.

1.7 Σύντομη αναφορά στα είδη των δεδομένων

Όπως έχουμε ήδη αναφέρει, το είδος των δεδομένων, η τεχνική δειγματοληψίας, καθώς και το είδος των ερευνών επηρεάζουν την τεχνική που θα χρησιμοποιηθεί. Για παράδειγμα, το εύρος των συνεχών μεταβλητών μπορεί να είναι από το $-\infty$ έως το $+\infty$, μπορεί να είναι από το 0 έως το $+\infty$, ή να είναι περιορισμένο εύρος και να κυμαίνεται από το 0 έως το 1.

Για παράδειγμα, στη γεωγραφία και στη σεισμολογία κάποιες μεταβλητές είναι σε ζεύγη, όπως το γεωγραφικό πλάτος και το γεωγραφικό μήκος. Η μέτρηση του ανέμου περιλαμβάνει την ταχύτητα του ανέμου, αλλά και τη γωνία της κατεύθυνσης. Σε αυτά τα δύο παραδείγματα ενδείκνυται η υιοθέτηση τεχνικών για δεδομένα κατεύθυνσης (directional data). Επίσης, στην οικονομική επιστήμη η μεταβλητή ενδιαφέροντος μπορεί να είναι η ποσοστιαία κατανομή των εξόδων των νοικοκυριών σε διάφορες κατηγορίες (π.χ. ενοίκιο, λογαριασμοί κοινής ωφέλειας, διατροφή, διασκέδαση, συντήρηση οχήματος κλπ.) Τέτοιου τύπου δεδομένα ονομάζονται δεδομένα σύστασης (compositional data) και χρήζουν ανάλογων τεχνικών.

Στην οικονομική επιστήμη είναι, επίσης, συνηθισμένη η μελέτη δεδομένων που περιλαμβάνει τόσο διαστρωματικά στοιχεία όσο και χρονολογικές σειρές (panel data), που αναφέρονται σε επαναλαμβανόμενες μετρήσεις στον χρόνο. Επίσης, μία άλλη περίπτωση είναι τα δεδομένα κατάταξης (π.χ. κατάταξη οικονομικών περιοδικών ανάλογα με κάποιο σκορ ή κατάταξη ομάδων και αθλητών). Είναι προφανές ότι αυτό που έχει σημασία δεν είναι το απόλυτο σκορ, αλλά η σειρά κατάταξης.

Είναι πλέον εμφανές ότι ο σκοπός, τα ερευνητικά ερωτήματα, το είδος των μεταβλητών, το είδος των δεδομένων, καθώς και η μέθοδος συλλογής τους επηρεάζουν το είδος της ανάλυσης που θα εφαρμοστεί. Οπότε, είναι εξαιρετικά σημαντικό να καθορίζονται οι παραπάνω παράμετροι πριν από την έναρξη του πειράματος ή της οποιασδήποτε διαδικασίας.

Βιβλιογραφία

Ξενόγλωσση

Doll, R., & Hill, A.B. (1964). Mortality in relation to smoking: ten years' observations of British doctor. *British Medical Journal*, *1*(5395), 1399-1410. https://doi.org/10.1136/bmj.1.5395.1399

Κεφάλαιο 2 Επισκόπηση του SPSS και του EVIEWS

Σύνοψη

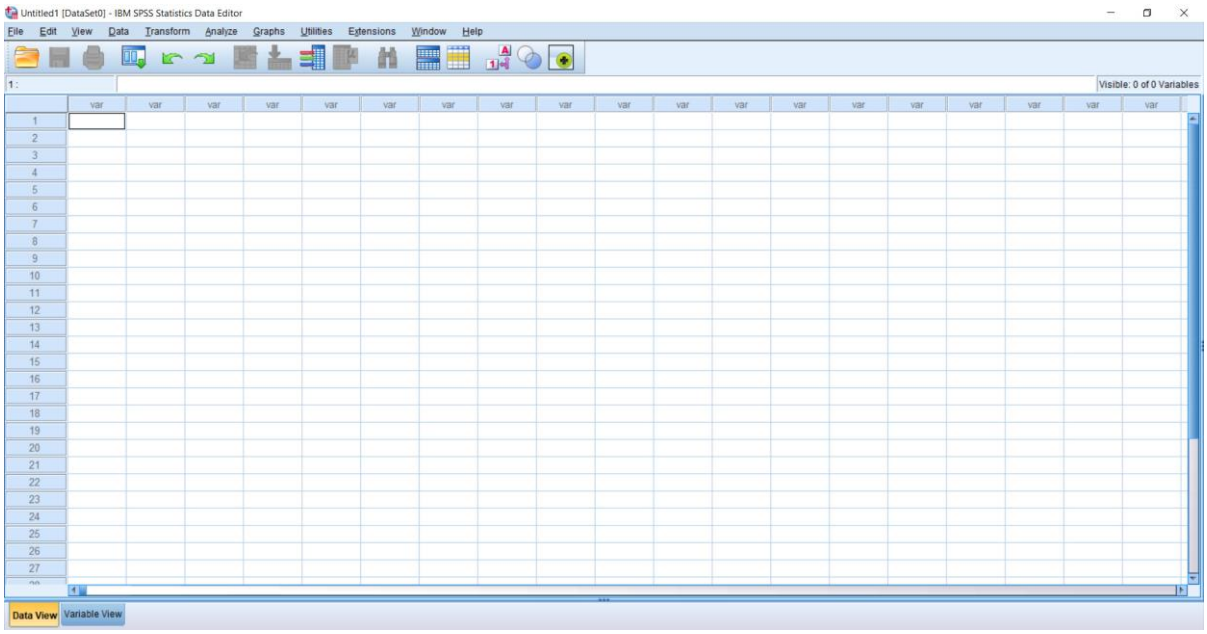
Το συγκεκριμένο κεφάλαιο αποτελείται από έντεκα ενότητες, οι οποίες στοχεύουν στην εξοικείωση του/της αναγνώστη/-τριας με τα στατιστικά προγράμματα IBM SPSS 26 και Eviews 11. Παρουσιάζεται το περιβάλλον των δύο προγραμμάτων, καθώς και οι διάφορες διαθέσιμες επιλογές για στατιστικές/οικονομετρικές αναλύσεις. Τέλος, δίνεται επιπλέον έμφαση σε κάποιες λειτουργίες των προγραμμάτων αυτών, όπως είναι η εισαγωγή των δεδομένων, η επιλογή τιμών και ο μετασχηματισμός των δεδομένων.

Προαπαιτούμενη γνώση

Δεν χρειάζονται εξειδικευμένες γνώσεις για την κατανόηση του κεφαλαίου.

2.1 Ανοίγοντας το IBM® SPSS® Statistics software ("SPSS") 26

Έχοντας εγκατεστημένο το στατιστικό πρόγραμμα SPSS, μπορούμε να προχωρήσουμε στη συνοπτική παρουσίαση των λειτουργιών που μας αφορούν, καθώς και στη χρήση του για τη διεξαγωγή στατιστικών αναλύσεων. Για να ανοίξουμε το SPSS, είτε κάνουμε διπλό κλικ πάνω στο αντίστοιχο εικονίδιο που βρίσκεται στην επιφάνεια εργασίας (αν υπάρχει τέτοιο εικονίδιο) ή πατάμε *Start button* και επιλέγουμε **IBM Statistics SPSS 26** (η ονομασία μπορεί να είναι διαφορετική ανάλογα με τον χρήστη). Μόλις ανοίξει το πρόγραμμα, θα εμφανιστεί μία οθόνη γεμάτη κελιά που αποτελεί ένα κενό φύλλο εργασίας (**IBM SPSS Data Editor**), όπως και στην περίπτωση του Microsoft Excel, καθώς και ένα πλαίσιο διαλόγου **(Εικόνα 2.1)**.



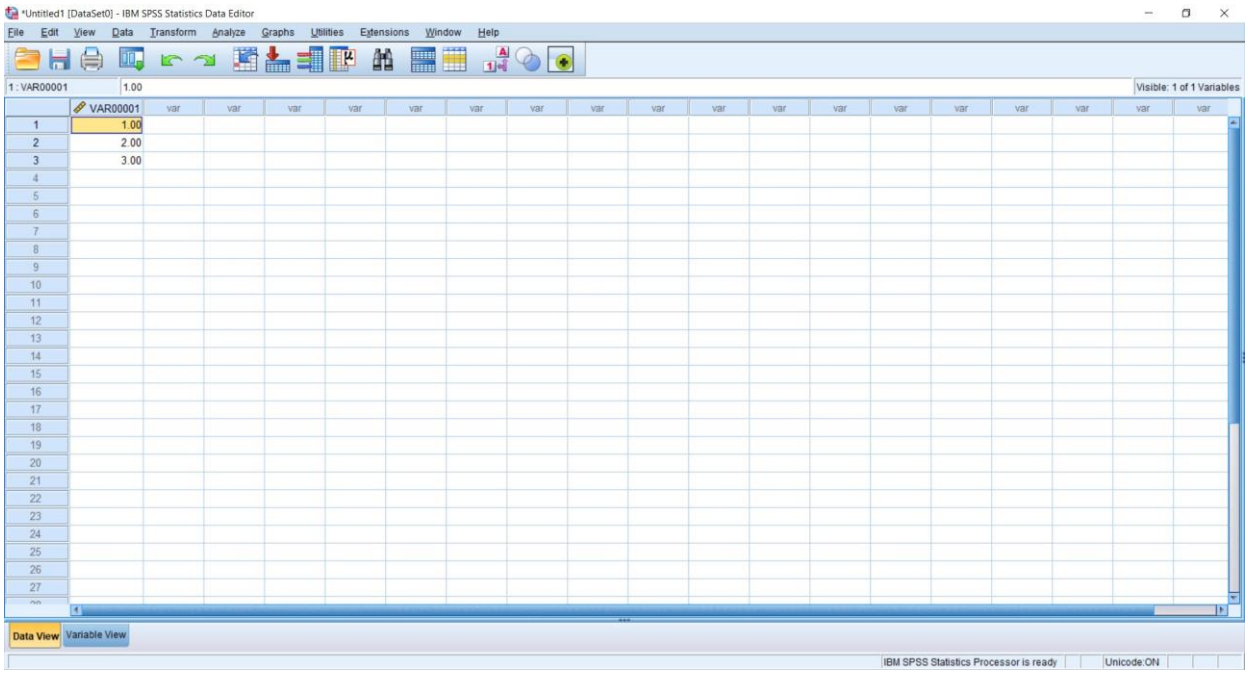
Εικόνα 2.1 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

2.2 Τα παράθυρα του SPSS και οι επιλογές τους

Το παράθυρο IBM SPSS Data Editor παρουσιάζεται στην **Εικόνα 2.2**. Η γραμμή τίτλου είναι η γκρι γραμμή που εμφανίζεται στο πάνω μέρος του παραθύρου. Το μενού επιλογών είναι παρόμοιο με αυτό που συναντάται στο Micorsoft Office. Εμφανίζεται κάτω από τη γραμμή τίτλου και περιλαμβάνει τις εξής επιλογές: **File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, Extensions**, **Window** και **Help**. Η γραμμή εργαλείων βρίσκεται κάτω από το μενού επιλογών και αποτελείται από εικονίδια που είναι πολύ χρήσιμα για λειτουργίες που χρησιμοποιούνται συχνά, όπως είναι η αποθήκευση, η εκτύπωση και το

άνοιγμα κάποιου υπάρχοντος αρχείου. Οι γραμμές κύλισης βρίσκονται στα δεξιά και στο κάτω μέρος του παραθύρου και μας βοηθάνε να μετακινηθούμε πάνω-κάτω και δεξιά-αριστερά. Στο κάτω μέρος του παραθύρου (δεξιά) εμφανίζεται ένα μήνυμα που αναφέρει ότι **IBM SPSS Statistics Processor is ready**. Η γραμμή στην οποία εμφανίζεται αυτό το μήνυμα είναι η γραμμή κατάστασης. Όταν το SPSS διεξάγει κάποιον υπολογισμό, πραγματοποιεί μία διεργασία σε εξέλιξη ή τερματίσει μία οποιαδήποτε διεργασία, εμφανίζεται το αντίστοιχο μήνυμα στη γραμμή κατάστασης. Επίσης, στο πάνω μέρος των κελιών εμφανίζεται το όνομα των κελιών. Από τη στιγμή που δεν έχουμε δώσει ονόματα στα κελιά, απλά εμφανίζεται η λέξη **var**.

Στο κάτω αριστερό μέρος του παραθύρου εμφανίζονται δύο επιλογές. Η πρώτη είναι η **Data View** και η δεύτερη είναι η **Variable View**. Στην **Εικόνα 2.2** φαίνεται ενεργοποιημένη η πρώτη επιλογή. Δηλαδή, το παράθυρο στο οποίο μπορούμε να εισάγουμε δεδομένα είναι αυτό που φαίνεται στην οθόνη. Τα δεδομένα εισάγονται με τον ίδιο τρόπο που εισάγονται στο Microsoft Excel, δηλαδή κάθετα αν πρόκειται για τιμές της ίδιας μεταβλητής. Πληκτρολογούμε τον αριθμό που επιθυμούμε και πατάμε **Enter**. Ενδεικτικά, έχουμε πληκτρολογήσει 3 αριθμούς.

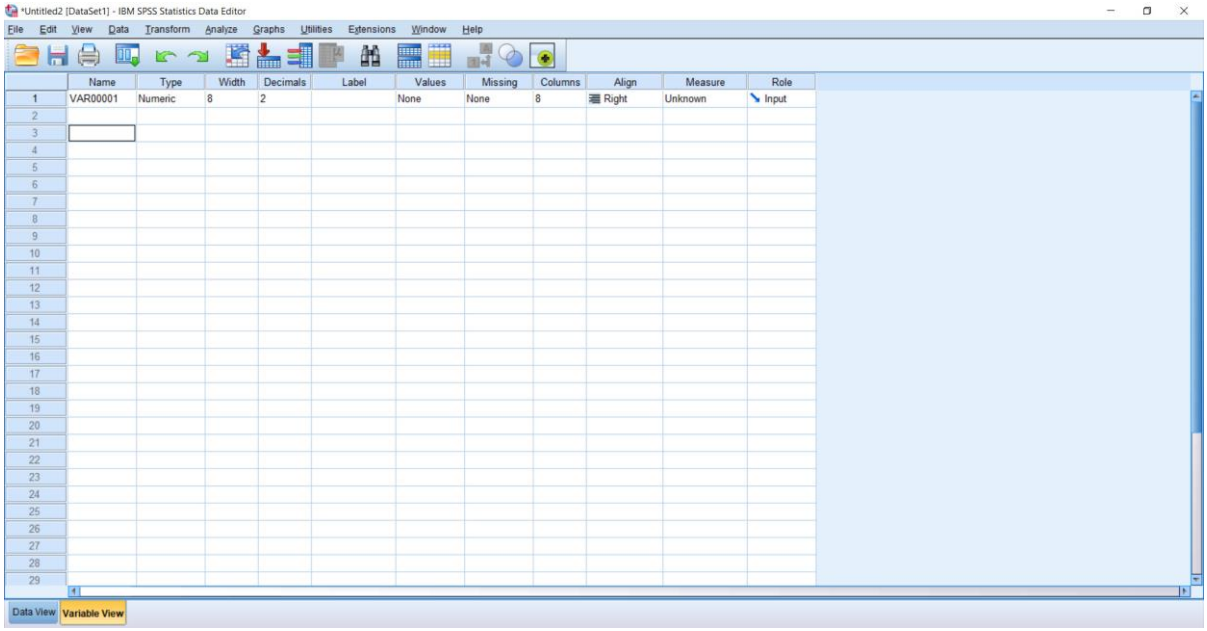


Εικόνα 2.2 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

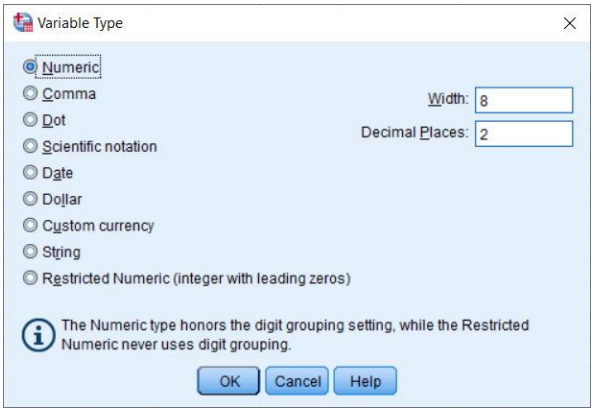
Αν ενεργοποιήσουμε τη δεύτερη επιλογή (**Variable View**), θα εμφανιστεί το παράθυρο **της Εικόνας 2.3.** Η πρώτη στήλη έχει τίτλο **Name**. Στα κελιά της πρώτης στήλης δίνουμε τα ονόματα των στηλών των δεδομένων που βρίσκονται στο Data Editor. Έτσι, στο πρώτο κελί αντιστοιχεί το όνομα της πρώτης στήλης των δεδομένων, στο δεύτερο κελί αντιστοιχεί το όνομα της δεύτερης στήλης κοκ. Αν σε ένα κελί της δεύτερη στήλης (**Type**) πατήσουμε στο δεξί μέρος του κελιού (το μικρό εικονίδιο με τις τρεις τελείες), θα εμφανιστεί το παράθυρο της **Εικόνας 2.4**. Το παράθυρο αυτό μας δίνει τη δυνατότητα να επιλέξουμε τον τύπο των δεδομένων για κάθε στήλη του Data Editor. Ξανά το πρώτο κελί αντιστοιχεί στα δεδομένα της πρώτης στήλης του Data Editor, το δεύτερο κελί στα δεδομένα της δεύτερης στήλης κοκ. Η επιλογή **Numeric** είναι προεπιλεγμένη από το πρόγραμμα, γιατί τα δεδομένα μας είναι συνήθως αριθμητικά. Αν επιλέξουμε **String**, τότε τα δεδομένα μας θα είναι σε μορφή χαρακτήρα ή απλά θα είναι γράμματα. Με το **Width** ορίζουμε το μέγιστο πλήθος των ψηφίων που θα έχουν τα αριθμητικά δεδομένα και με το **Decimal Places** ορίζουμε το πλήθος των δεκαδικών ψηφίων.

Το τρίτο και το τέταρτο κελί του παραθύρου της **Εικόνας 2.3** αναφέρονται στο πλήθος των ψηφίων αριστερά και δεξιά της υποδιαστολής, όπως ήδη αναφέραμε. Το κελί με τίτλο **Label** είναι η ετικέτα των δεδομένων. Αν δηλώσουμε ονόματα στα δεδομένα μας, στους πίνακες που θα προκύψουν με τα

αποτελέσματα από το SPSS θα εμφανίζονται οι στήλες των δεδομένων με τα ονόματά τους. Αν, όμως, δηλώσουμε και ετικέτες ή μόνο ετικέτες στις στήλες των δεδομένων μας, στα αποτελέσματα που θα προκύψουν κάθε στήλη δεδομένων θα έχει για όνομα αυτό που έχουμε ορίσει στις ετικέτες των στηλών. Το τι θα εμφανίζεται μπορεί να επιλεγεί από τον χρήστη μέσω της υπο-επιλογής **Options**, η οποία βρίσκεται μέσα στην επιλογή **Edit** που θα εξετάσουμε παρακάτω.



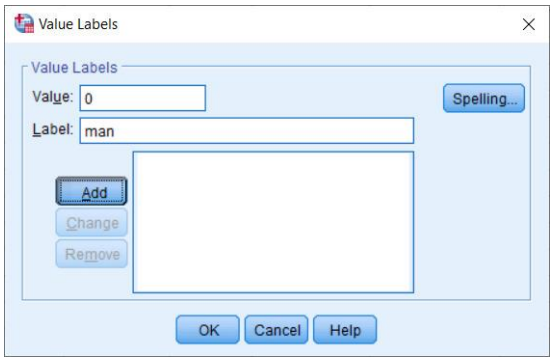
Εικόνα 2.3 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Εικόνα 2.4 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

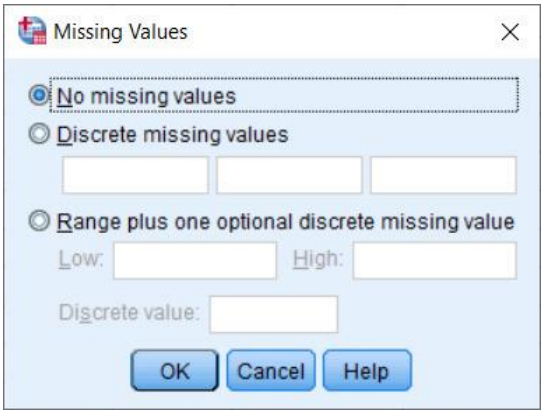
Κάνοντας κλικ σε ένα κελί της στήλης **Values** (και στη συνέχεια πάλι κλικ πάνω στο μικρό εικονίδιο με τις τρεις τελείες δεξιά ενός κελιού της στήλης αυτής), θα οδηγηθούμε στο παράθυρο της **Εικόνας 2.5**. Έστω, λοιπόν, ότι έχουμε ποιοτικά δεδομένα και θέλουμε να τα εισάγουμε στο SPSS. Για να γίνει αυτό, θα πρέπει να κωδικοποιήσουμε εξαρχής τα δεδομένα δίνοντάς τους τιμές για καθεμία κατηγορία. Για παράδειγμα, έστω ότι θέλουμε να καταχωρίσουμε σε μία στήλη του SPSS Data Editor το φύλο κάποιων ανθρώπων για τους οποίους έχουμε συλλέξει κάποια στατιστικά στοιχεία. Η πιο συνήθης κωδικοποίηση είναι της μορφής 0 για τους άνδρες και 1 για τις γυναίκες. Για να θυμόμαστε, λοιπόν, πού αντιστοιχεί ο κάθε αριθμός, θα πρέπει να εισάγουμε και το φύλο στο SPSS. Οπότε, γράφουμε 0 στο πρώτο λευκό τετραγωνάκι που αντιστοιχεί στο **Value** και στη συνέχεια, στο **Value Label** πληκτρολογούμε το φύλο, δηλαδή man. Στη συνέχεια, πατάμε **Add**, ώστε να καταχωριστεί το φύλο στον αριθμό. Το ίδιο κάνουμε και για τις γυναίκες. Προφανώς, η διαδικασία επαναλαμβάνεται όσες φορές χρειαστεί. Για παράδειγμα, σε ένα ερωτηματολόγιο

που υπάρχουν περισσότερες από δύο πιθανές απαντήσεις, η διαδικασία θα επαναληφθεί τόσες φορές όσες είναι και οι απαντήσεις.



Εικόνα 2.5 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Κάνοντας κλικ σε ένα κελί της στήλης **Missing** (και στη συνέχεια πάλι κλικ πάνω στο μικρό εικονίδιο με τις τρεις τελείες δεξιά ενός κελιού της στήλης αυτής) μπορούμε να ορίσουμε τις χαμένες παρατηρήσεις. Για παράδειγμα, κάποιοι εκ των ερωτηθέντων σε ένα ερωτηματολόγιο δεν έχουν απαντήσει σε όλες τις ερωτήσεις. Το παράθυρο που θα εμφανιστεί είναι αυτό που εμφανίζεται στην **Εικόνα 2.6**. Όμως, θα πρέπει να είμαστε ιδιαίτερα προσεκτικοί στο πώς θα ορίσουμε τις χαμένες τιμές. Αν σε ένα ερωτηματολόγιο οι απαντήσεις είναι σε κλίμακα Likert από 1 έως 5, τότε στις χαμένες τιμές θα δώσουμε έναν αριθμό που δεν βρίσκεται μεταξύ 1 και 5. Θα πρέπει, δηλαδή, να είναι ένας αριθμός που δεν βρίσκεται στα δεδομένα της κάθε στήλης ξεχωριστά.

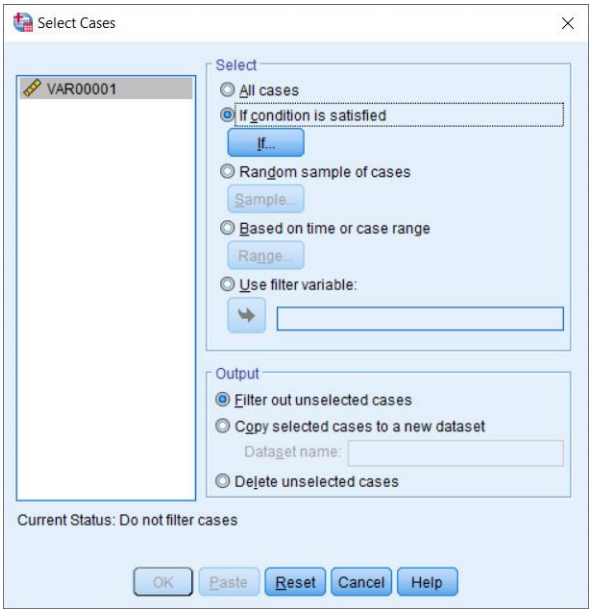


Εικόνα 2.6 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

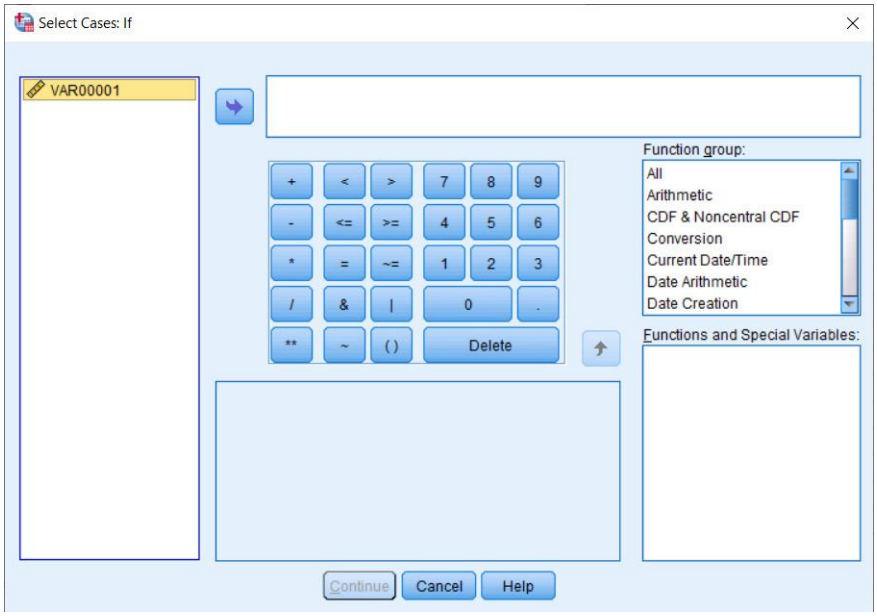
Το κελιά **Columns** και **Align** αναφέρονται στο μέγεθος της στήλης και στη στοίχιση των δεδομένων σε κάθε στήλη. Τέλος, η τελευταία στήλη (**Measure**) αναφέρεται στον τύπο των δεδομένων, δηλαδή αν τα δεδομένα αφορούν ποσοτικές μετρήσεις (**Scale**), διατεταγμένες (**Ordinal**) ή ονομαστικές (**Nominal**).

2.3 Η εντολή Select Cases

Επιλέγοντας **Data → Select Case** θα εμφανιστεί το παράθυρο της **Εικόνας 2.7**. Η επιλογή αυτή δίνει τη δυνατότητα στον χρήστη να επιλέξει ένα μόνο μέρος των δεδομένων, προκειμένου να χρησιμοποιήσει αυτό στις αναλύσεις. Η επιλογή **All cases** είναι προεπιλεγμένη από το πρόγραμμα. Όμως, μπορούμε να επιλέξουμε **If condition is satisfied** και στη συνέχεια **If…** Στην περίπτωση αυτή, θα οδηγηθούμε στο παράθυρο της **Εικόνας 2.8**.



Εικόνα 2.7 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Εικόνα 2.8 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

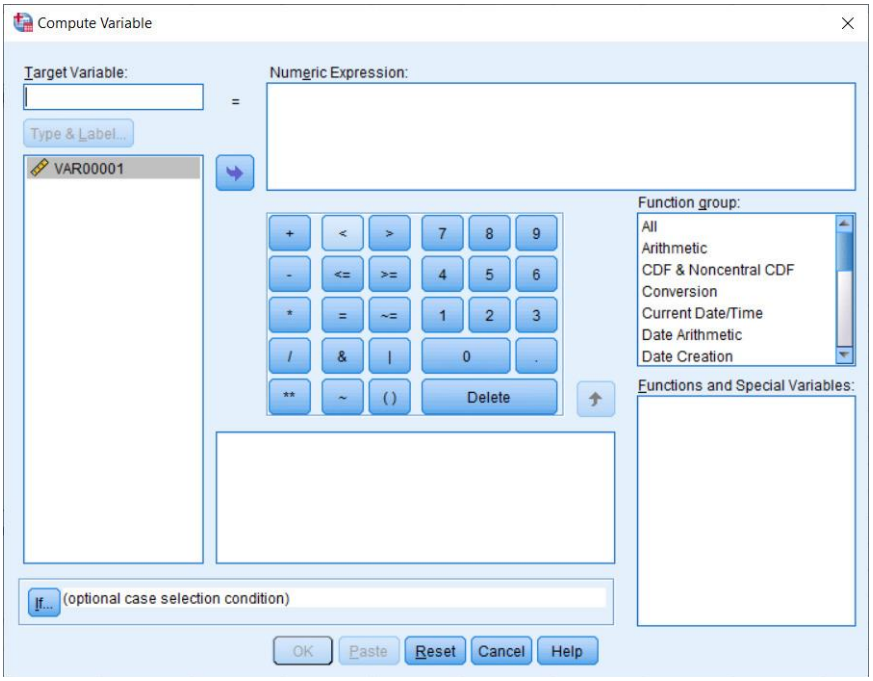
Η επιλογή αυτή μας επιτρέπει να επιλέξουμε από μία ή περισσότερες στήλες δεδομένων κάποια δεδομένα που ικανοποιούν μία συγκεκριμένη συνθήκη. Περνώντας την/τις μεταβλητή/-ές από το αριστερό κουτάκι στο δεξιό κουτάκι του παραθύρου της **Εικόνας 2.8**, δίνουμε τη δυνατότητα στο SPSS να «αντιληφθεί» για ποιες στήλες δεδομένων θέλουμε τα δεδομένα τους να ικανοποιούν κάποια συνθήκη. Υπάρχει μία λίστα από συναρτήσεις λογικές και μη, τις οποίες μπορούμε να χρησιμοποιήσουμε. Για παράδειγμα, οι τελεστές > , < , >= , <= δηλώνουν ανισοϊσότητες. Μπορούμε, δηλαδή, να επιλέξουμε τα δεδομένα μίας στήλης που να είναι μικρότερα ή μεγαλύτερα από μία τιμή. Χρησιμοποιώντας τις μαθηματικές συναρτήσεις, μπορούμε να ζητήσουμε από το SPSS να επιλέξει τα δεδομένα εκείνα για τα οποία η απόλυτη τιμή τους είναι μικρότερη από μία καθορισμένη τιμή.

Η επιλογή **random sample of cases** που εμφανίζεται στο παράθυρο της **Εικόνας 2.7** μας δίνει τη δυνατότητα της τυχαίας επιλογής είτε ενός ποσοστού των δεδομένων είτε ενός δείγματος από τα δεδομένα, καθορίζοντας φυσικά το μέγεθος του δείγματος. Η επιλογή **Based on time or case range** μας

επιτρέπει να επιλέξουμε δεδομένα, τα οποία βρίσκονται μέσα σε κάποια περιοχή ή κάποια όρια. Όμως, θα πρέπει να είμαστε ιδιαίτερα προσεκτικοί σχετικά με τα δεδομένα που δεν επιλέγονται με βάση κάποια από τις προηγούμενες επιλογές. Η επιλογή **Filter out unselected cases** είναι προεπιλεγμένη από το πρόγραμμα. Στην περίπτωση αυτή τα συγκεκριμένα δεδομένα απλά δεν θα περιλαμβάνονται στις επόμενες αναλύσεις. Θα εμφανιστεί μία νέα στήλη που θα περιέχει τις τιμές 0 και 1, ανάλογα με το αν τα δεδομένα ικανοποιούν ή όχι τη συνθήκη. Επίσης, στη στήλη που περιέχει την αρίθμηση των γραμμών θα εμφανιστεί μία διαγώνιος γραμμή που θα έχει «διαγράψει» κατά κάποιο τρόπο τις γραμμές των δεδομένων που δεν ικανοποιούν τη συνθήκη. Αν επιλέξουμε **Delete unselected cases**, τότε τα δεδομένα θα διαγραφούν από το αρχείο. Αν όμως επιλέξουμε **Copy selected cases to a new dataset**, τότε τα δεδομένα που ικανοποιούν τη συνθήκη θα αποθηκευτούν σε ένα νέο αρχείο δεδομένων του SPSS, στο οποίο θα πρέπει να δώσουμε ένα όνομα πληκτρολογώντας το στο λευκό κουτάκι που θα ενεργοποιηθεί (**Dataset name**).

2.4 Η επιλογή Transform

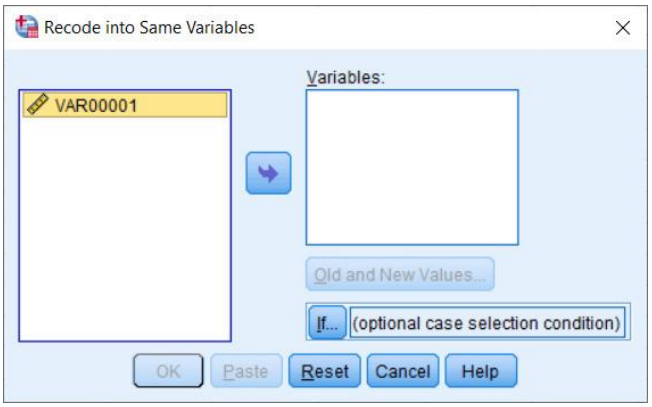
Μία πολύ χρήσιμη επιλογή από το μενού επιλογών είναι αυτή της μετατροπής των δεδομένων (**Transform**). Η πρώτη εντολή που εμπεριέχεται στην επιλογή **Transform** είναι η **Compute Variable**. Επιλέγοντας την εντολή αυτή θα εμφανιστεί το παράθυρο της **Εικόνας 2.9**. Το λευκό κουτάκι που λέγεται **Target Variable** θα πρέπει να συμπληρωθεί με ένα όνομα, καθώς εκεί θα αποθηκευτεί η μετασχηματισμένη στήλη δεδομένων. Η μετασχηματισμένη στήλη μπορεί να αποθηκευτεί είτε στην ίδια στήλη είτε σε διαφορετική. Περνώντας τις στήλες από το αριστερό κουτάκι στο κουτάκι που λέγεται **Numeric Expression**, μπορούμε να ορίσουμε τις στήλες που θα μετασχηματιστούν. Το κουτάκι **Function group** περιέχει διάφορα είδη συναρτήσεων, όπως μαθηματικές, στατιστικές, μετατροπής και άλλες. Για κάθε είδος συναρτήσεων που επιλέγουμε, στο κουτάκι που βρίσκεται ακριβώς κάτω από αυτές βλέπουμε τις διαθέσιμες συναρτήσεις. Οι συναρτήσεις αυτές μας βοηθάνε στον μετασχηματισμό των δεδομένων. Βεβαίως, μπορούμε να γράψουμε μία δική μας συνάρτηση μετατροπής, η οποία δεν βρίσκεται στη λίστα με τις ήδη υπάρχουσες συναρτήσεις.



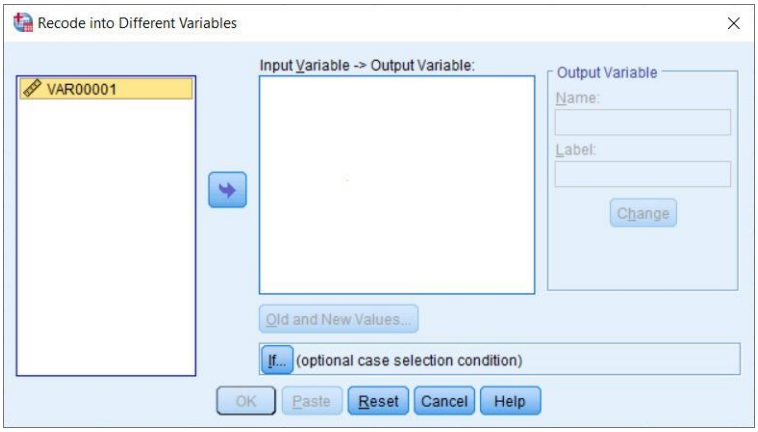
Εικόνα 2.9 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Η εντολή επανακωδικοποίησης είναι εντολή κωδικοποίησης των ήδη υπαρχουσών στηλών δεδομένων. Ένα παράδειγμα χρησιμοποίησης αυτής της εντολής είναι το εξής: έστω ότι έχουμε συλλέξει ηλικίες ατόμων από 20 έως 90+ έτη. Αντί να δουλεύουμε με τις ηλικίες αυτές καθαυτές, έστω ότι θέλουμε να τις κατηγοριοποιήσουμε σε ομάδες ηλικιών (για παράδειγμα, σε 7 ομάδες). Μπορούμε να επιλέξουμε να

σώσουμε τις ομάδες ηλικιών είτε στη στήλη των ήδη υπαρχουσών ηλικιών (οπότε θα χαθούν οι ηλικίες) είτε σε μία άλλη στήλη. Θα επιλέξουμε, δηλαδή, είτε **Recode into Same Variables** είτε **Recode into Different Variables** αντίστοιχα. Αν επιλέξουμε να σώσουμε τη νέα στήλη των ομάδων ηλικιών στη στήλη των ήδη υπαρχουσών ηλικιών, διαγράφοντας ουσιαστικά τις ηλικίες, θα εμφανιστεί το παράθυρο της **Εικόνας 2.10**. Αν επιλέξουμε να αποθηκεύσουμε τη στήλη των ηλικιακών ομάδων σε άλλη στήλη, θα εμφανιστεί το παράθυρο της **Εικόνας 2.11**. Και στις δύο περιπτώσεις θα πρέπει να περάσουμε τη στήλη των δεδομένων που θέλουμε να μετασχηματίσουμε από το αριστερό στο δεξιό λευκό κουτάκι. Μόλις το κάνουμε αυτό, θα ενεργοποιηθεί η επιλογή **Old and New Values**, η οποία βρίσκεται κάτω από το δεξιό κουτάκι. Επιλέγοντάς την, θα εμφανιστεί το παράθυρο της **Εικόνας 2.12**.

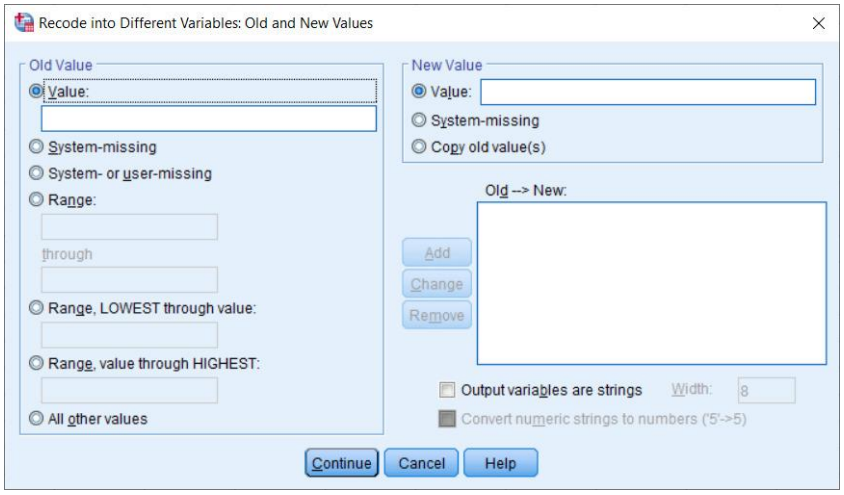


Εικόνα 2.10 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Εικόνα 2.11 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Στην περιοχή **Old Value**, η οποία βρίσκεται στο αριστερό μέρος του παραθύρου, θα επιλέξουμε να πληκτρολογήσουμε τις τιμές των δεδομένων που θα μετασχηματίσουμε. Στο παράδειγμα με τις ηλικιακές ομάδες θα επιλέξουμε το **Range** και θα πληκτρολογήσουμε το εύρος των τιμών στα δύο λευκά κουτάκια που θα ενεργοποιηθούν. Έστω ότι η πρώτη ηλικιακή ομάδα είναι οι ηλικίες από 20 έως 30 έτη. Οπότε, θα πληκτρολογήσουμε το 20 στο πρώτο κουτάκι και το 30 στο δεύτερο κουτάκι. Με τον τρόπο αυτό δηλώνουμε το εύρος των τιμών που θέλουμε να μετασχηματίσουμε. Υπάρχουν άλλες δύο επιλογές, με τις οποίες μπορούμε να ορίσουμε εύρη ή διαστήματα τιμών: είτε από τη χαμηλότερη τιμή έως κάποια τιμή είτε από κάποια τιμή έως την υψηλότερη. Στη συνέχεια, μεταφερόμαστε στο δεξιό μέρος του παραθύρου, στην περιοχή **New Value**. Στο λευκό κουτάκι που βρίσκεται δεξιά της προεπιλογής **Value** θα πληκτρολογήσουμε τη νέα τιμή για τη συγκεκριμένη ηλικιακή ομάδα. Καθώς είναι η πρώτη ηλικιακή ομάδα, θα βάλουμε τον αριθμό 1. Στη συνέχεια, θα πατήσουμε το κουτάκι **Add**, προκειμένου να καταχωριστεί η συγκεκριμένη αλλαγή στο SPSS.



Εικόνα 2.12 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

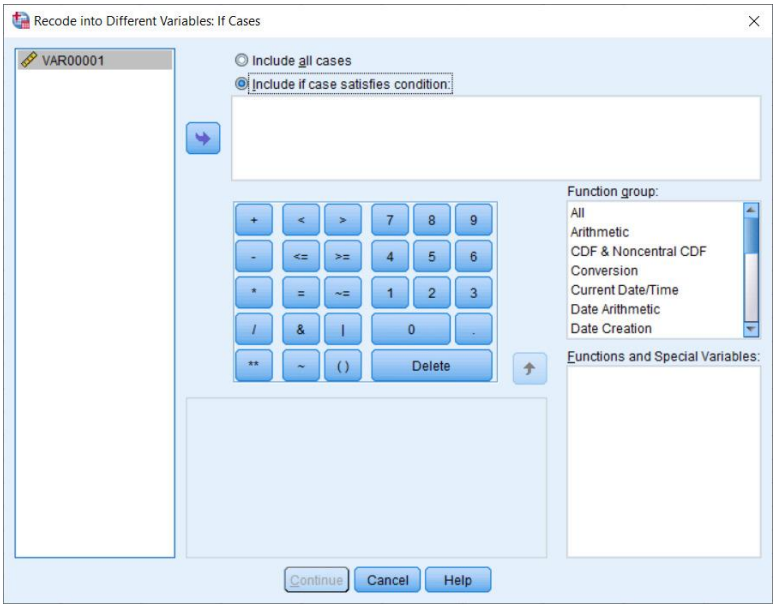
Μόλις προβούμε στη συγκεκριμένη ενέργεια, θα εμφανιστεί στο μεγάλο λευκό κουτάκι η καταχωρισμένη αλλαγή. Οπότε, συνεχίζουμε με τον ίδιο τρόπο και για τις υπόλοιπες ηλικιακές ομάδες. Αν έχουμε επιλέξει οι νέες κωδικοποιημένες τιμές να αποθηκευτούν στη στήλη με τις ήδη υπάρχουσες τιμές, δεν χρειάζεται να κάνουμε τίποτα άλλο. Αν όμως έχουμε επιλέξει να αποθηκεύσουμε τις νέες τιμές σε διαφορετική στήλη, τότε πατώντας **Continue** θα επιστρέψουμε στο παράθυρο της **Εικόνας 2.11**, το οποίο είναι διαφορετικό από αυτό της **Εικόνας 2.10**. Κάτω από το **Output Variable** θα πρέπει να δώσουμε στη νέα στήλη ένα όνομα και μία ετικέτα (οι συγκεκριμένες εντολές ενεργοποιούνται στην αρχή, μόλις περάσουμε τη στήλη των δεδομένων από το αριστερό στο δεξιό κουτάκι). Στη συνέχεια, πατώντας **Change** γίνεται η αλλαγή στο όνομα. Όταν πλέον τελειώσουμε, θα εμφανιστεί στο SPSS Data Editor μία νέα στήλη, η οποία θα περιέχει τις κωδικοποιημένες τιμές της αρχικής στήλης. Και στις δύο περιπτώσεις, το SPSS θα αντιστοιχίσει τις νέες τιμές στις ήδη υπάρχουσες τιμές, ανάλογα με το διάστημα στο οποίο βρίσκονται. Για τις ηλικίες δηλαδή από 20 έως και 29, θα αντιστοιχίσει την τιμή 1 (η ηλικία 30 θα συμπεριληφθεί στη δεύτερη ομάδα, όπως και κάθε άνω άκρο των κλάσεων ή ομάδων). Για τις ηλικίες από 30 έως 39 θα αντιστοιχίσει την τιμή 2. Οι κλάσεις που θα δημιουργηθούν θα είναι τύπου [,).

Καλό θα είναι να επιλέξουμε στο Data Editor τη δυνατότητα να εμφανιστεί το παράθυρο Variable View, προκειμένου να καθορίσουμε την κάθε κωδικοποιημένη τιμή. Στο παράδειγμα με τις ηλικίες, επιλέγοντας **Values** και ακολουθώντας τη διαδικασία που έχει ήδη περιγραφεί, μπορούμε να ορίσουμε σε ποια ηλικιακή ομάδα αντιστοιχεί κάθε τιμή. Οπότε, για την τιμή 1 μπορούμε να πληκτρολογήσουμε στο παράθυρο που θα εμφανιστεί (δείτε την **Εικόνα 2.5**) **ages between 20 and 30**. Αυτό είναι εξαιρετικά βοηθητικό, καθώς δίνει τη δυνατότητα να μην εμφανίζονται αριθμοί στις αναλύσεις (π.χ. 1, 2, 3,) και συνεπώς, να μην χρειάζεται να εξηγούμε κάθε φορά ότι η τιμή 1 αντιστοιχεί στην πρώτη ηλικιακή ομάδα κοκ. Με τον τρόπο αυτό, το μήνυμα που έχουμε πληκτρολογήσει στο παράθυρο (δείτε την **Εικόνα 2.5**) θα εμφανίζεται για κάθε τιμή.

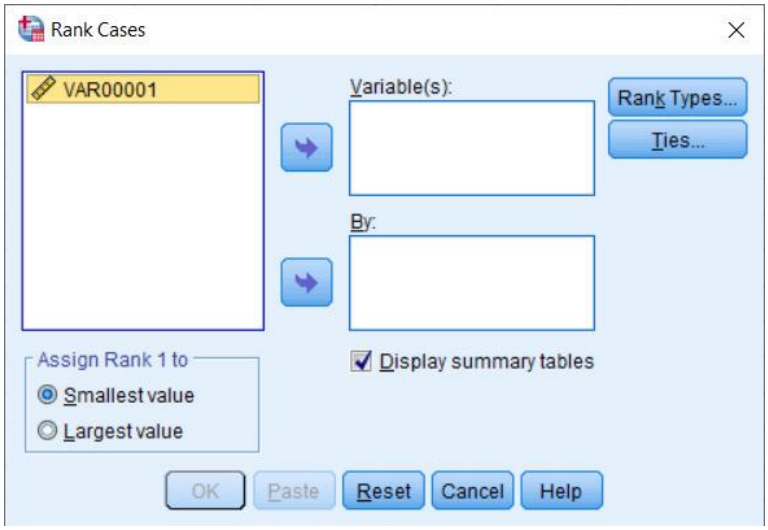
Ανεξάρτητα από την επιλογή αποθήκευσης της μετασχηματισμένης στήλης δεδομένων, αν πατήσουμε την επιλογή **If**, θα εμφανιστεί το παράθυρο της **Εικόνας 2.13**. Με την επιλογή αυτή κωδικοποιούμε μόνο τις τιμές που ικανοποιούν κάποια λογική συνθήκη. Όπως και στο παράθυρο της **Εικόνας 2.9**, πρώτα επιλέγουμε τη συνάρτηση από τη λίστα των συναρτήσεων, στη συνέχεια την ανεβάζουμε πάνω χρησιμοποιώντας το βελάκι και τέλος, περνάμε δεξιά τη στήλη με την οποία θα εργαστούμε.

Με την εντολή **Rank Cases** αναθέτουμε τάξεις μεγέθους στα δεδομένα των στηλών που επιλέγουμε. Στην περίπτωση αυτή εμφανίζεται το παράθυρο της **Εικόνας 2.14**. Στο πάνω δεξιό κουτάκι περνάμε τη στήλη στης οποίας τα δεδομένα θέλουμε να τοποθετήσουμε τάξεις μεγέθους. Κάτω αριστερά μας δίνεται η δυνατότητα να επιλέξουμε από πού θα αρχίζουν οι τάξεις μεγέθους. Η προεπιλογή του SPSS είναι αυτή που δίνει στη μικρότερη τιμή την τιμή 1, στην αμέσως επόμενη την τιμή 2 κοκ. Αν θέλουμε, μπορούμε να επιλέξουμε απλά τάξεις μεγέθους για τα δεδομένα ή ποσοστά από το **Rank Types**. Επιλέγοντας το **Ties**, το

SPSS μας ρωτάει τι τάξη μεγέθους να αναθέσει στην περίπτωση ισοβαθμουσών τάξεων μεγέθους. Η προεπιλογή είναι ο μέσος όρος των τάξεων. Έστω, για παράδειγμα, ότι η πέμπτη και η έκτη τιμή είναι ίσες. Οι τάξεις μεγέθους που θα αντιστοιχούσαν στις δύο αυτές τιμές, αν ήταν διαφορετικές, θα ήταν η 5 και η 6 αντίστοιχα. Στην περίπτωση αυτή που έχουμε δύο ίσες τιμές, το SPSS θα αναθέσει και στις δύο τιμές τον μέσο όρο των δύο τάξεων μεγέθους, δηλαδή το 5,5. Οι εντολές **Create Time Series** και **Replace Missing values** αφορούν πιο προχωρημένη ανάλυση, οπότε η ανάλυσή τους προς το παρόν παραλείπεται.



Εικόνα 2.13 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Εικόνα 2.14 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

2.5 Η τεχνική του Bootstrap

Η IBM έχει πλέον ενδυναμώσει σε πολύ σημαντικό βαθμό το στατιστικό πρόγραμμα SPSS, καθώς έχει πλέον ενσωματώσει την επιλογή bootstrap. Η διαδικασία του bootstrap μπορεί να χαρακτηριστεί «συγγενής» με αυτή του Monte Carlo. Οπότε, στο σημείο αυτό, θεωρούμε πως θα ήταν ιδανικό να προσπαθήσουμε να κατανοήσουμε τι είναι αυτή η τεχνική, ή καλύτερα, αυτός ο αλγόριθμος. Η ιδέα δημοσιεύτηκε για πρώτη

φορά από τον Efron (1979). Πρόκειται για τη θέα της στατιστικής από μία άλλη οπτική, πιο υπολογιστική. Θα προσπαθήσουμε να εξηγήσουμε τον συγκεκριμένο αλγόριθμο, χωρίς να γίνουμε πολύ τεχνικοί. Για την ιστορία, θα πρέπει να αναφέρουμε ότι το όνομα προήλθε από το παραμύθι «Οι περιπέτειες του Βαρόνου Μινχάουζεν». Σε μία από τις περιπέτειές του ο Βαρόνος βυθιζόταν στη θάλασσα. Προκειμένου να σωθεί, θα έπρεπε να τραβήξει κάποιο σκοινί, το οποίο όμως δεν είχε. Οπότε, άρχισε να τραβάει τα κορδόνια της μπότας του και έτσι «τράβηξε τον εαυτό του» επάνω στην επιφάνεια της θάλασσας.

Στη στατιστική, όταν υπολογίζουμε τον μέσο (όρο) ενός δείγματος, υποστηρίζουμε ότι εκτιμάμε την πραγματική τιμή του μέσου του πληθυσμού από τον οποίο προήλθε το δείγμα. Στη συνέχεια, θεωρούμε ότι ο μέσος ακολουθεί ασυμπτωτικά (δηλαδή, καθώς το μέγεθος του δείγματος μεγαλώνει και τείνει προς το άπειρο) την κανονική κατανομή. Οπότε, όταν κατασκευάζουμε ένα 95% διάστημα εμπιστοσύνης για την πραγματική τιμή του μέσου, η ερμηνεία βασίζεται σε ασυμπτωτικά αποτελέσματα.

Η ερμηνεία του 95% διαστήματος εμπιστοσύνης για τον μέσο είναι η εξής: αν είχαμε τη δυνατότητα να επαναλάβουμε τη δειγματοληψία *n* φορές και κάθε φορά να εκτιμήσουμε τον μέσο και να κατασκευάσουμε 95% διάστημα εμπιστοσύνης, θα αναμέναμε το 95% αυτών των διαστημάτων να περιλαμβάνει την πραγματική τιμή του μέσου. Ο αλγόριθμος bootstrap μας δίνει τη δυνατότητα αυτή, δηλαδή να προσομοιώσουμε αυτή τη δυνατότητα επαναδειγματοληψίας βασισμένοι στο αρχικό μας δείγμα. Όπως ο Βαρόνος έσωσε τον εαυτό του τραβώντας τον, έτσι και ο αλγόριθμος bootstrap θα μας δώσει μία εκτίμηση για την κατανομή του δειγματικού μέσου, η οποία θα βασίζεται στο ίδιο το δείγμα.

Θα περιγράψουμε τη συγκεκριμένη διαδικασία με τη βοήθεια μιας λοταρίας. Έστω ότι βάζουμε *n* μπαλάκια με αριθμούς σε ένα βάζο. Βάζουμε το χέρι μας μέσα και τραβάμε ένα μπαλάκι. Σημειώνουμε τον αριθμό του σε ένα χαρτί και το ρίχνουμε πάλι μέσα. Στη συνέχεια, ξαναβάζουμε το χέρι μας μέσα και τραβάμε και πάλι ένα μπαλάκι, σημειώνουμε τον αριθμό του και το ρίχνουμε μέσα στο βάζο (αυτή η διαδικασία λέγεται δειγματοληψία με επανατοποθέτηση). Η διαδικασία αυτή θα επαναληφθεί *n* φορές, καθώς *n* μπαλάκια έχουμε στη διάθεσή μας. Προφανώς, κάποια μπαλάκια μπορεί να έχουν επιλεγεί περισσότερες από μία φορές, αλλά αυτό δεν αποτελεί πρόβλημα. Αφού, λοιπόν, έχουμε επιλέξει *n* μπαλάκια, έχουμε τελειώσει την πρώτη δειγματοληψία bootstrap και έχουμε το πρώτο δείγμα bootstrap. Στη συνέχεια, παίρνουμε τα νούμερα από τα μπαλάκια και υπολογίζουμε, για παράδειγμα, τον μέσο όρο τους. Αυτός θα είναι ο πρώτος μέσος bootstrap. Επαναλαμβάνουμε τη διαδικασία που μόλις περιγράψαμε πολλές φορές, έστω 1.000. Τελικά, θα έχουμε υπολογίσει 1.000 μέσους (bootstrap). Με τον τρόπο αυτό θα έχουμε «κατασκευάσει» την κατανομή του δειγματικού μέσου. Αυτή η κατανομή θα «μιμείται» την πραγματική κατανομή του μέσου που είναι η κανονική. Για εμάς, προφανώς τα μπαλάκια είναι οι παρατηρήσεις που έχουμε, δηλαδή το δείγμα μεγέθους *n*.

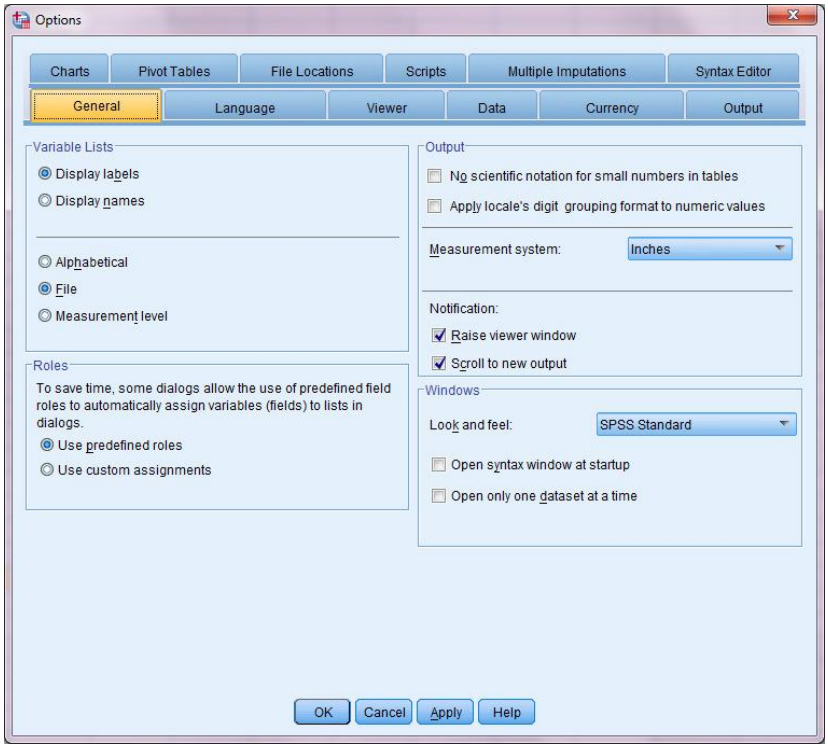
Έχοντας, λοιπόν, την (ψευδο-)κατανομή του μέσου, μπορούμε να εκτιμήσουμε το τυπικό σφάλμα του (υπολογίζοντας την τυπική απόκλιση των 1.000 αυτών τιμών bootstrap) και να κατασκευάσουμε το διάστημα εμπιστοσύνης για την πραγματική τιμή του μέσου. Αφού έχουμε 1.000 τιμές για τον μέσο, μπορούμε να τις κατατάξουμε από τη μικρότερη στη μεγαλύτερη και να πάρουμε το 95% των κεντρικών τιμών. Δηλαδή, θα αφήσουμε εκτός τις 25 χαμηλότερες και τις 25 υψηλότερες τιμές. Οπότε, η $26^{η}$ και η $975^{η}$ τιμή θα αποτελούν το κάτω και το άνω άκρο του 95% διαστήματος εμπιστοσύνης για τον μέσο, αντίστοιχα. Επιπλέον, μας δίνεται η δυνατότητα να εκτιμήσουμε τη μεροληψία του δειγματικού μέσου, δηλαδή την απόσταση του εκτιμημένου μέσου από τον πραγματικό μέσο. Ο μέσος του δείγματος θα παίξει τον ρόλο του πραγματικού μέσου, ενώ ο μέσος που θα προκύψει από το bootstrap θα παίξει τον ρόλο του εκτιμημένου μέσου. Συνεπώς, η διαφορά τους θα αποτελεί μια εκτίμηση της πραγματικής μεροληψίας.

2.6 Το μενού της επιλογής Analyze

Πριν αναλύσουμε την επιλογή **Analyze** από το μενού επιλογών του **Data Editor**, ας εξετάσουμε τις επιλογές που εμφανίζονται στο μενού:

- Η επιλογή **File** χρησιμοποιείται προκειμένου να δημιουργήσουμε ή να ανοίξουμε ένα νέο αρχείο δεδομένων ή να αποθηκεύσουμε/εκτυπώσουμε το υπάρχον αρχείο δεδομένων.
- Η επιλογή **Edit** χρησιμοποιείται για την επεξεργασία δεδομένων, όπως αντιγραφή, επικόλληση κ.ά.

- Η επιλογή **Options** οδηγεί στο παράθυρο της **Εικόνας 2.15**. Παρέχονται γενικές επιλογές του SPSS, όπως η εμφάνιση των αποτελεσμάτων και των πινάκων.
- Η επιλογή **View** παρέχει λίγες πληροφορίες, μία από τις οποίες είναι η παρουσίαση των δεδομένων χωρίς κελιά.

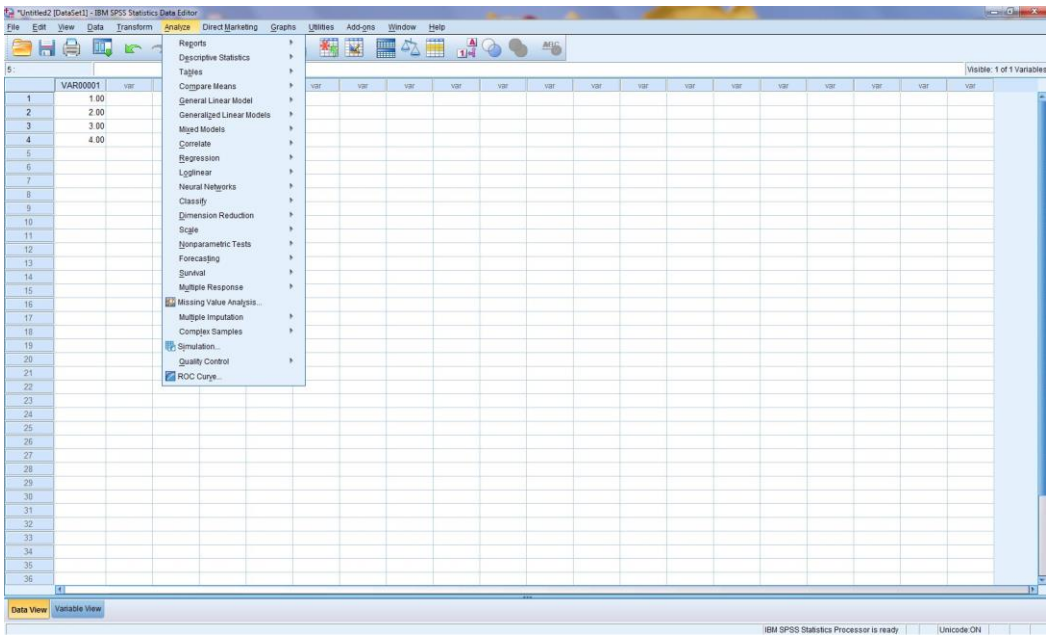


Εικόνα 2.15 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

- Η επιλογή **Data** μας παρέχει τη δυνατότητα να επιλέξουμε τις στήλες δεδομένων με τις οποίες θέλουμε να εργαστούμε (την είδαμε προηγουμένως), τη διάταξη των δεδομένων, τον ορισμό νέων μεταβλητών κ.ά.
- Η επιλογή **Transform** παρέχει δυνατότητες που αναλύσαμε παραπάνω.
- Η επιλογή **Analyze** είναι η «καρδιά» των επιλογών του SPSS, καθώς περιέχει όλες σχεδόν τις εντολές ανάλυσης δεδομένων, και θα την αναλύσουμε παρακάτω.
- Η επιλογή **Direct Marketing** περιέχει χρήσιμα εργαλεία σε αυτούς που αναλύουν τεχνικές *marketing*.
- Η επιλογή **Graphs** μας επιτρέπει τη δημιουργία διαγραμμάτων και θα αναλυθεί στη συνέχεια.
- Η επιλογή **Utilities** μας επιτρέπει να δημιουργήσουμε ομάδες στηλών, να δούμε πληροφορίες για κάθε στήλη ξεχωριστά κ.ά.
- Η επιλογή **Add-ons** έχει διάφορες επιλογές που παραπέμπουν στην ηλεκτρονική διεύθυνση του SPSS.
- Η επιλογή **Window** μας επιτρέπει να κάνουμε *split* του φύλλου δεδομένων ή να ελαχιστοποιήσουμε το παράθυρο εργασίας.
- Η τελευταία επιλογή του μενού επιλογών είναι αυτή της βοήθειας (**Help**) και είναι πάρα πολύ χρήσιμη.

Επιλέγοντας **Analyze (Εικόνα 2.16)** από το μενού, θα εμφανιστεί το υπομενού αυτής της επιλογής που περιέχει σχεδόν όλες τις στατιστικές τεχνικές του SPSS. Πιο συγκεκριμένα, περιέχονται οι ακόλουθες εντολές:

- **Reports**: περιέχει δυνατότητες παρουσίασης κάποιων στοιχείων των δεδομένων.
- **Descriptive Statistics**: περιέχει δυνατότητες εμφάνισης περιγραφικών μέτρων των δεδομένων, γραφημάτων, πινάκων δεδομένων κ.ά.
- **Bayesian Statistics**: περιλαμβάνει συγκεκριμένες αναλύσεις Μπεϋζιανής στατιστικής.
- **Tables**: περιλαμβάνει δυνατότητες δημιουργίας πολύπλοκων πινάκων.
- **Compare Means**: περιλαμβάνει εντολές ελέγχων υποθέσεων για τους μέσους, οι οποίες θα εξεταστούν αναλυτικότερα στη συνέχεια.
- **General Linear Model**: περιλαμβάνει υποδείγματα ανάλυσης διακύμανσης, τα οποία θα εξεταστούν αναλυτικότερα στη συνέχεια.
- **Generalized Linear Models**: περιέχει πληθώρα δυνατοτήτων σχετικά με την εκτίμηση γενικευμένων γραμμικών υποδειγμάτων.
- **Mixed Models**: αφορά μεικτά γραμμικά υποδείγματα.
- **Correlate**: περιέχει συντελεστές συσχέτισης, μερικής συσχέτισης και υπολογισμού αποστάσεων.
- **Regression**: περιέχει δυνατότητες εκτίμησης απλής και πολλαπλής γραμμικής και μη γραμμικής παλινδρόμησης, λογιστικής παλινδρόμησης κ.ά.
- **Loglinear**: παρέχει δυνατότητες χρησιμοποίησης λογαριθμικών υποδειγμάτων.
- **Neural Networks**: περιέχει εργαλεία για νευρωνικά δίκτυα.
- **Classify**: περιλαμβάνει πολλές πολυμεταβλητές, στατιστικές και μη, τεχνικές ομαδοποίησης δεδομένων ή μεταβλητών.
- **Dimension Reduction**: περιέχει πολυμεταβλητές τεχνικές μείωσης μεταβλητών, όπως παραγοντική ανάλυση και ανάλυση αντιστοιχιών.
- **Scale**: περιέχει τεχνικές πολυδιάστατης κλιμακοποίησης και ανάλυσης αξιοπιστίας, οι οποίες χρησιμοποιούνται κυρίως σε ψυχομετρικούς ελέγχους και ελέγχους προσωπικότητας, ικανοτήτων.
- **Nonparametric Tests**: περιλαμβάνει μη παραμετρικές στατιστικές τεχνικές, η χρησιμότητα των οποίων θα αναλυθεί στη συνέχεια.
- **Forecasting**: περιέχει διάφορες τεχνικές ανάλυσης χρονολογικών σειρών.

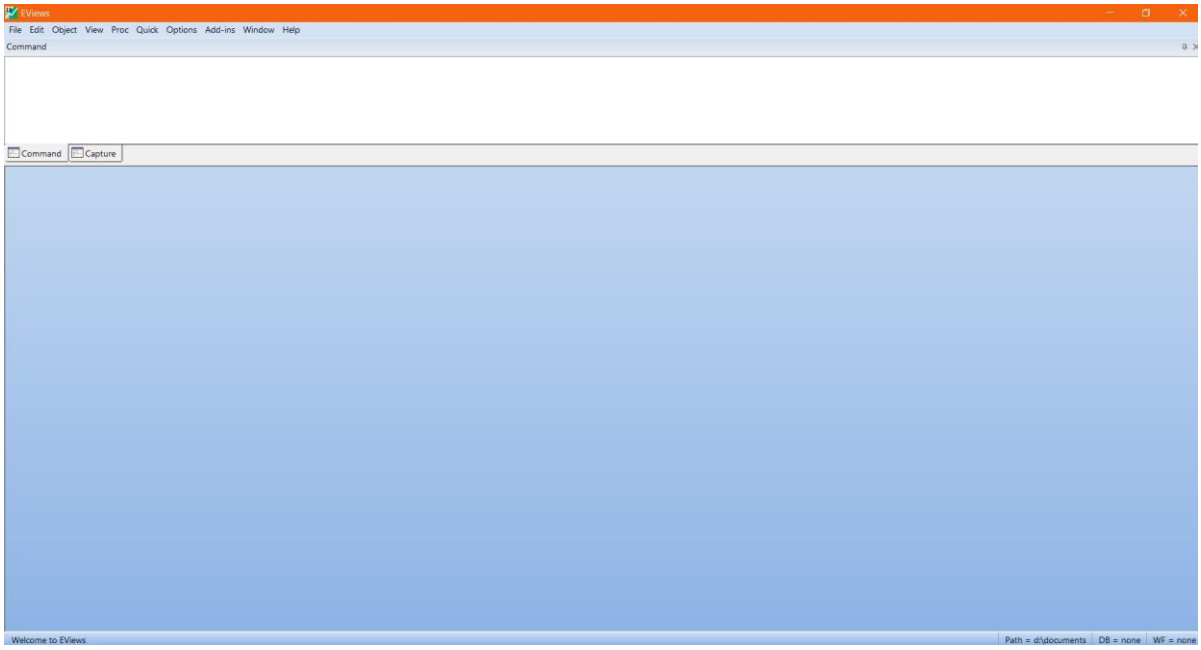


Εικόνα 2.16 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

- **Survival**: περιλαμβάνει τεχνικές ανάλυσης ετών ζωής από ιατρικές μελέτες.
- **Multiple Response**: παρέχεται η δυνατότητα δημιουργίας διχοτομικών μεταβλητών (ή ψευδομεταβλητών, όπως αλλιώς ονομάζονται), δηλαδή μεταβλητών που παίρνουν τις τιμές 0 και 1, από μεταβλητές με πολλές κατηγορίες.
- **Missing Value Analysis**: αφορά την ανάλυση εκλιπουσών τιμών.
- **Multiple Imputation**: αφορά την ανάλυση εκλιπουσών τιμών.
- **Complex Samples**: περιέχει ένα σύνολο διαδικασιών δειγματοληψίας.
- **Simulation**: περιλαμβάνει διαδικασίες προσομοίωσης.
- **Quality Control**: περιέχει διαδικασίες στατιστικού ελέγχου ποιότητας.
- **ROC Curve**: αφορά την ανάλυση καμπυλών χαρακτηριστικού λειτουργικού δείκτη.

2.7 Ανοίγοντας το IHS Markit® Eviews® software ("Eviews") 11

Το Eviews 11 (αλλά και όλες οι προηγούμενες εκδόσεις του) είναι περισσότερο εστιασμένο στην οικονομετρική ανάλυση. Έχοντας εγκατεστημένο το συγκεκριμένο πρόγραμμα, με διπλό κλικ πάνω στο αντίστοιχο εικονίδιο που βρίσκεται στην επιφάνεια εργασίας (αν υπάρχει τέτοιο εικονίδιο) ή πατώντας *Start button* και επιλέγοντας **Eviews 11**, εμφανίζεται μια οθόνη όπως αυτή της **Εικόνας 2.17**.



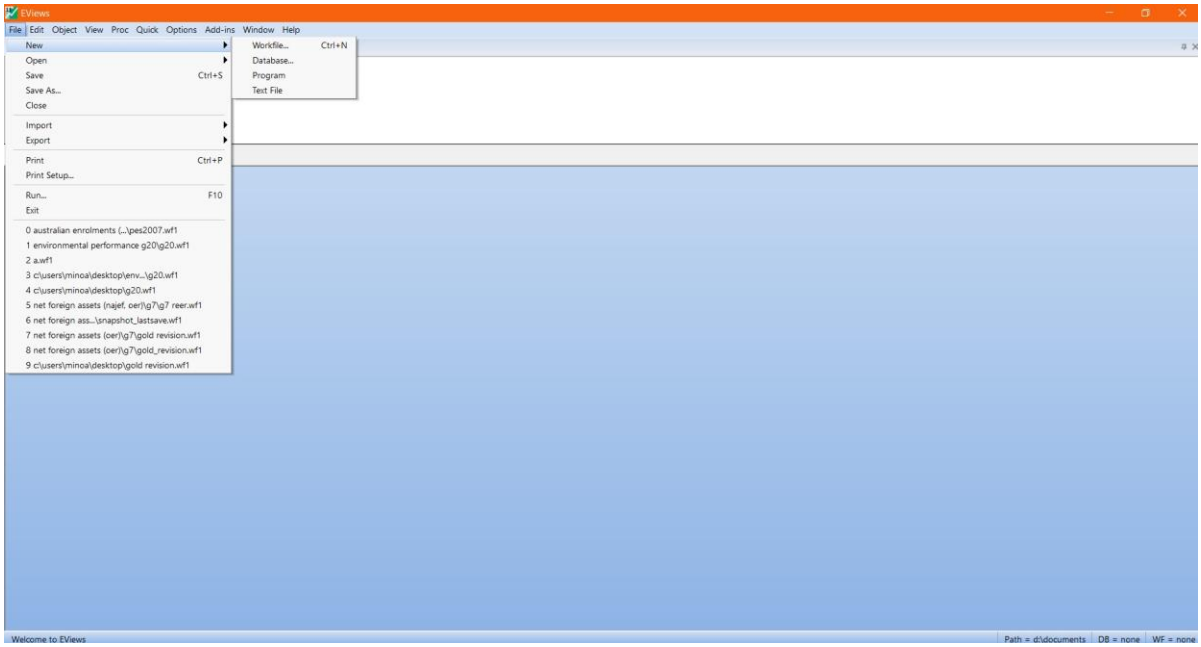
Εικόνα 2.17 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

2.8 Τα παράθυρα του Eviews και οι επιλογές τους

Όπως φαίνεται στην **Εικόνα 2.17**, το μενού επιλογών είναι στην ίδια λογική με αυτό που συναντάται στο SPSS, αλλά και στο Microsoft Office. Εμφανίζεται κάτω από τη γραμμή τίτλου και περιλαμβάνει τις εξής επιλογές: File, Edit, Object, View, Proc, Quick, Options, Add-ins, Window και Help. Κάτω από το μενού επιλογών εμφανίζεται ένα λευκό πλαίσιο, στο οποίο υπάρχουν δύο «κουμπιά»: Command και Capture. Επιλέγοντας Command μας δίνεται η δυνατότητα να γράψουμε εντολές στο λευκό πλαίσιο, το οποίο, όπως θα δούμε στη συνέχεια, είναι πολύ σημαντικό, ενώ επιλέγοντας Capture βλέπουμε το αρχείο ή τα αρχεία Eviews που έχουμε ανοίξει. Στην παρούσα φάση, αν πλοηγηθούμε στις διάφορες επιλογές του μενού επιλογών, θα διαπιστώσουμε ότι αρκετές επιλογές εμφανίζονται με αχνά γράμματα, κάτι που σημαίνει ότι τη δεδομένη στιγμή δεν μπορούν να χρησιμοποιηθούν. Επίσης, αν επιλέξουμε View ή Proc, θα μας εμφανίσει το μήνυμα "None available for this window", κάτι που επίσης σημαίνει ότι την παρούσα στιγμή οι συγκεκριμένες επιλογές δεν μπορούν να χρησιμοποιηθούν. Επίσης, κάτω δεξιά εμφανίζονται κάποιες

επιπλέον πληροφορίες, όπως Path=…, η οποία αφορά την τοποθεσία στον υπολογιστή όπου αποθηκεύεται το αρχείο που επεξεργαζόμαστε, DB=…, η οποία αφορά κάποια πιθανή βάση δεδομένων που χρησιμοποιούμε και WF=…, η οποία αντιστοιχεί στο όνομα του αρχείου Eviews που δουλεύουμε.

Επιλέγοντας **File** από το μενού επιλογών **(Εικόνα 2.18)** μας δίνεται η δυνατότητα να δημιουργήσουμε (**New**) ένα νέο αρχείο εργασίας του Eviews (**Workfile**). Όλα τα workfiles του Eviews έχουν κατάληξη **.wf1** και είναι αυτά που θα μας απασχολήσουν στην πορεία. Επιπλέον, μπορούμε να δημιουργήσουμε μια νέα βάση δεδομένων (**Database**), ένα νέο πρόγραμμα (**Program**) ή ένα νέο αρχείο κειμένου (**Text File**). Επίσης, από την επιλογή **File** μας δίνεται η δυνατότητα να ανοίξουμε ένα ήδη αποθηκευμένο workfile (**Open**) και να αποθηκεύσουμε το αρχείο Eviews που ήδη δουλεύουμε (**Save**). Από την επιλογή **File** μπορούμε, επίσης, να εισάγουμε δεδομένα (**Import**), όπως θα δούμε παρακάτω, να εξάγουμε δεδομένα (**Export**), να εκτυπώσουμε (**Print**), να διαμορφώσουμε τη δομή της εκτύπωσης (**Print Setup**) και να εκτελέσουμε κάποιο πρόγραμμα που έχουμε ήδη δημιουργήσει στο Eviews (**Run**). Τέλος, στο κάτω μέρος της λίστας εμφανίζονται τα πιο πρόσφατα αρχεία Eviews που έχουμε επεξεργαστεί.

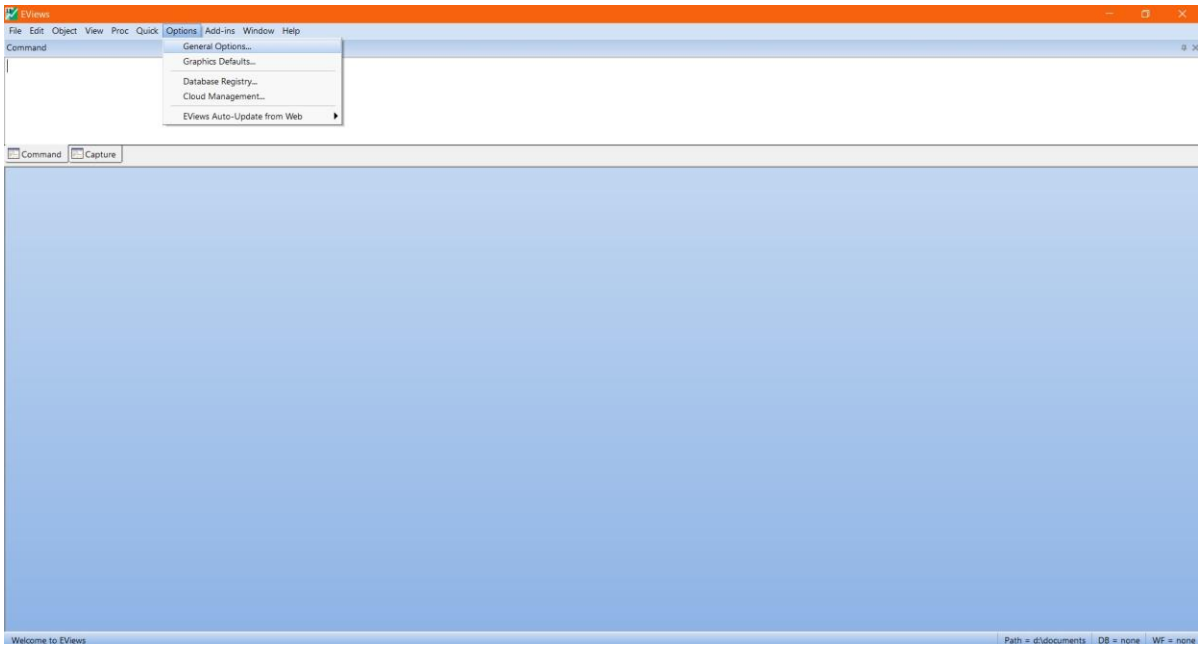


Εικόνα 2.1 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

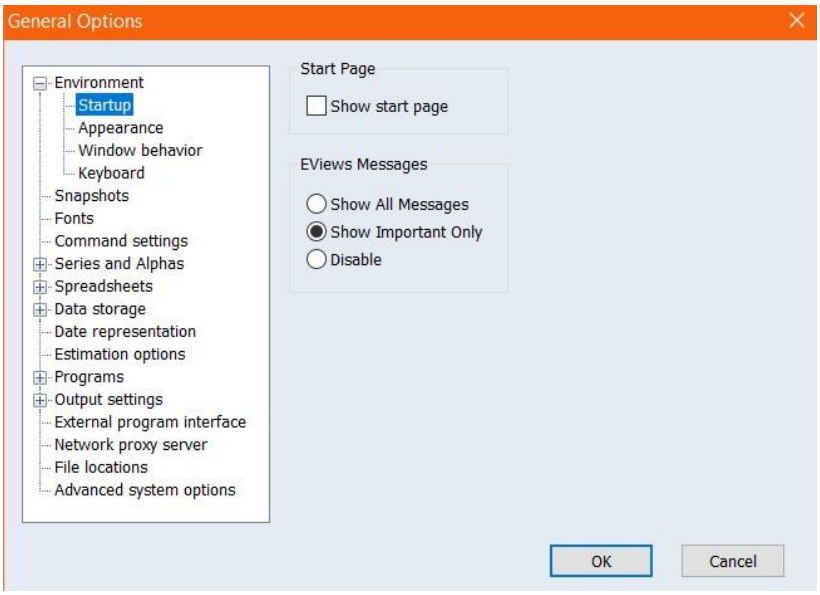
Η επιλογή **Edit** του μενού επιλογών είναι αντίστοιχη με αυτή που υπάρχει στο Microsoft Office. Χρησιμοποιώντας τη συγκεκριμένη επιλογή, μπορούμε να κάνουμε αναίρεση εντολής (**Undo**), αποκοπή (**Cut**), αντιγραφή (**Copy**), ειδική αντιγραφή (**Copy Special**), επικόλληση (**Paste**), ειδική επικόλληση (**Paste Special**) και διαγραφή (**Delete**) αντικειμένων που δουλεύουμε σε ένα workfile του Eviews. Όλα τα αντικείμενα που βρίσκονται σε ένα workfile του Eviews μπορούν να μεταφερθούν άμεσα σε ένα αρχείο του Microsoft Word, με τη χρήση των συγκεκριμένων εντολών. Επίσης, από την επιλογή **Edit** μπορούμε να βρούμε (**Find**) και να αντικαταστήσουμε (**Replace**) κάποιο μέρος κειμένου. Με τις επιλογές **Object**, **View**, **Proc** και **Quick** θα ασχοληθούμε στη συνέχεια, καθώς στην παρούσα φάση δεν μας επιτρέπουν να προβούμε σε κάποια ενέργεια.

Η επιλογή **Options** από το μενού επιλογών **(Εικόνα 2.19)** μας επιτρέπει να διαμορφώσουμε τη μορφή με την οποία θα εμφανίζεται το Eviews στον υπολογιστή μας, καθώς και τη μορφή των δεδομένων και των αντικειμένων σε κάθε workfile. Επίσης, η συγκεκριμένη επιλογή μας επιτρέπει να μορφοποιήσουμε τα διάφορα διαγράμματα που θα δημιουργήσουμε με το Eviews. Πιο αναλυτικά, επιλέγοντας **General Options** από το **Options**, εμφανίζεται ένα παράθυρο, όπως αυτό της **Εικόνας 2.20**. Καθώς οι επιλογές του συγκεκριμένου παράθυρου είναι πάρα πολλές, εμείς θα επικεντρωθούμε σε κάποιες βασικές που είναι απαραίτητο να αναλυθούν, καθώς οι υπόλοιπες απαιτούν μεγάλη εξοικείωση του χρήστη με το Eviews. Στο υπομενού **Startup** του μενού **Environment** μπορούμε να επιλέξουμε αν θέλουμε να μας εμφανίζει το Eviews κάποια αρχική σελίδα ή μηνύματα με την έναρξή του. Στο υπομενού **Appearance** του μενού

Environment (Εικόνα 2.21) μπορούμε να επιλέξουμε το χρώμα με το οποίο θα εμφανίζεται το Eviews στον υπολογιστή μας και αν θα εμφανίζονται κάποιες μπάρες με κουμπιά (**display button bars**). Προτείνεται το συγκεκριμένο κουτάκι να είναι επιλεγμένο, καθώς, όπως θα δούμε στη συνέχεια, οι button bars είναι πολύ βοηθητικές για τον χρήστη του Eviews.

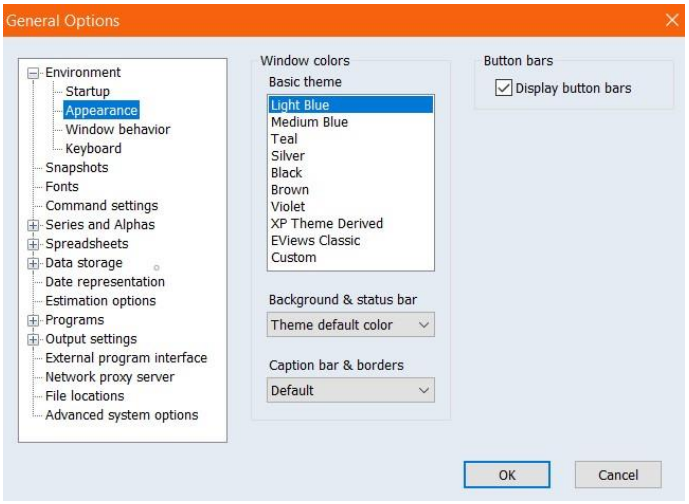


Εικόνα 2.19 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

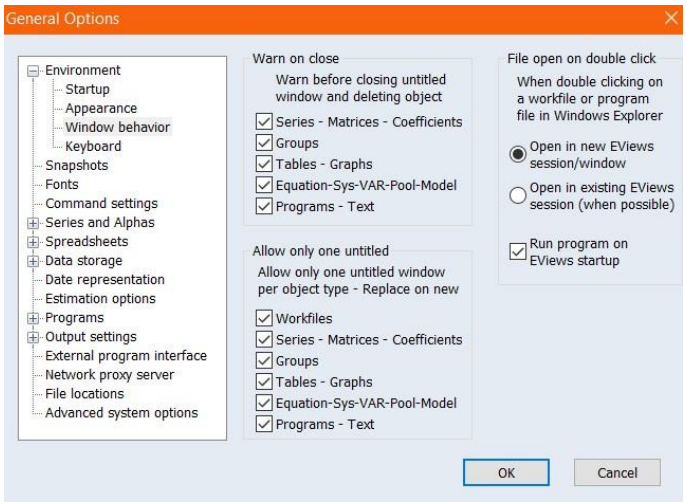


Εικόνα 2.20 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Το υπομενού **Windows behavior** του μενού **Environment (Εικόνα 2.22)** μας επιτρέπει να επιλέξουμε αν θα λαμβάνουμε προειδοποιήσεις, όταν πρόκειται να κλείσουμε ή να διαγράψουμε κάποια αντικείμενα του Eviews, αν θα μπορούμε να διατηρούμε χωρίς όνομα μόνο ένα ή και παραπάνω αντικείμενα στο workfile του Eviews, καθώς και αν θα έχουμε τη δυνατότητα να ανοίγουμε πολλαπλά workfiles του Eviews σε ένα ή σε διαφορετικά παράθυρα. Καλό θα είναι οι χρήστες του Eviews να κάνουν τις επιλογές που εμφανίζονται στην **Εικόνα 2.22**, τουλάχιστον σε αρχικό στάδιο, καθώς με τον τρόπο αυτό θα αποφύγουν την κατά λάθος διαγραφή αντικειμένων του Eviews που έχουν ήδη δημιουργήσει.



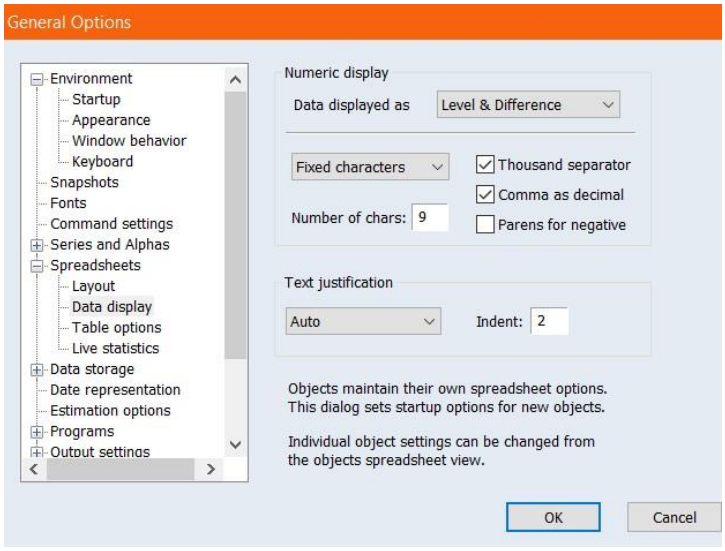
Εικόνα 2.21 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



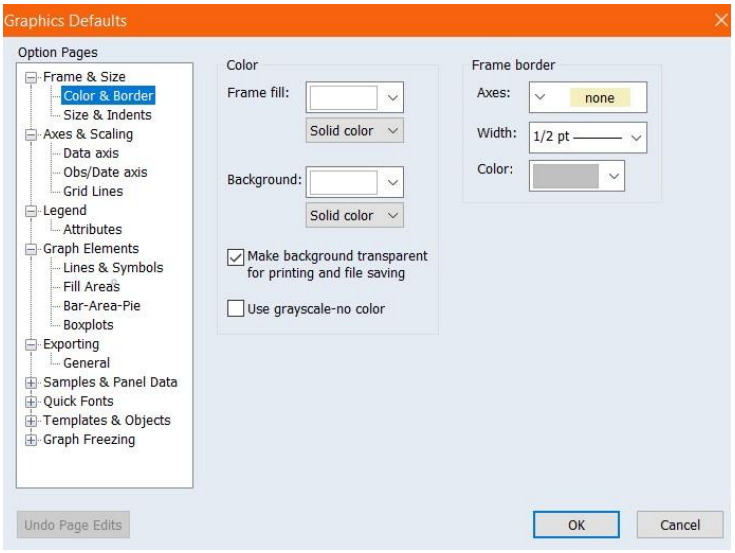
Εικόνα 2.22 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Το μενού **Fonts** μας επιτρέπει να επιλέξουμε το είδος και το μέγεθος της γραμματοσειράς στην οποία θα εμφανίζονται τα αποτελέσματα της ανάλυσης στο workfile του Eviews. Τέλος, είναι σημαντικό να αναφερθεί στο σημείο αυτό ότι για τα στατιστικά δεδομένα που εισάγονται στο Eviews είτε πληκτρολογώντας τα είτε εισάγοντάς τα από κάποιο άλλο πρόγραμμα, όπως το Microsoft Excel, υπάρχει η προεπιλογή (default) το δεκαδικό σύμβολο να έχει τη μορφή τελείας, όπως συμβαίνει στην αγγλική γλώσσα, και να μην υπάρχει διαχωριστικό χιλιάδων. Παρόλο που συνιστάται να παραμείνει αυτή η επιλογή, αν κάποιος χρήστης επιθυμεί να την αλλάξει, μπορεί να επιλέξει **Spreadsheets → Data display** και στη συνέχεια **Thousand separator**, αν θέλει να υπάρχει διαχωριστικό χιλιάδων στα δεδομένα του, ή/και **Comma as decimal** αν θέλει το δεκαδικό σύμβολο να έχει τη μορφή κόμματος, όπως συμβαίνει στην ελληνική γλώσσα **(Εικόνα 2.23)**. Τις υπόλοιπες επιλογές του **General Options** δεν θα τις αναλύσουμε στο σημείο αυτό, καθώς είναι αρκετά εξειδικευμένες και ξεφεύγουν από τους στόχους του παρόντος εγχειριδίου.

Επιλέγοντας **Graphic Defaults** από την επιλογή **Options (Εικόνα 2.19)**, εμφανίζεται το μενού της **Εικόνας 2.24**. Το συγκεκριμένο μενού μας επιτρέπει να διαμορφώσουμε τις επιθυμητές επιλογές μας σχετικά με τη μορφή που θα έχουν τα διαγράμματα που θα δημιουργήσουμε στο Eviews (χρώμα, κλίμακα αξόνων, πάχος γραμμών, υπομνήματα διαγραμμάτων, είδος και μέγεθος γραμματοσειρών κλπ.) και είναι στην ίδια λογική με το μενού **General Options** που αναλύσαμε παραπάνω. Δεν θα αναλύσουμε παραπάνω το μενού **Graphic Defaults**, καθώς είναι αρκετά εύχρηστο και μοιάζει πάρα πολύ με το αντίστοιχο του Microsoft Excel.

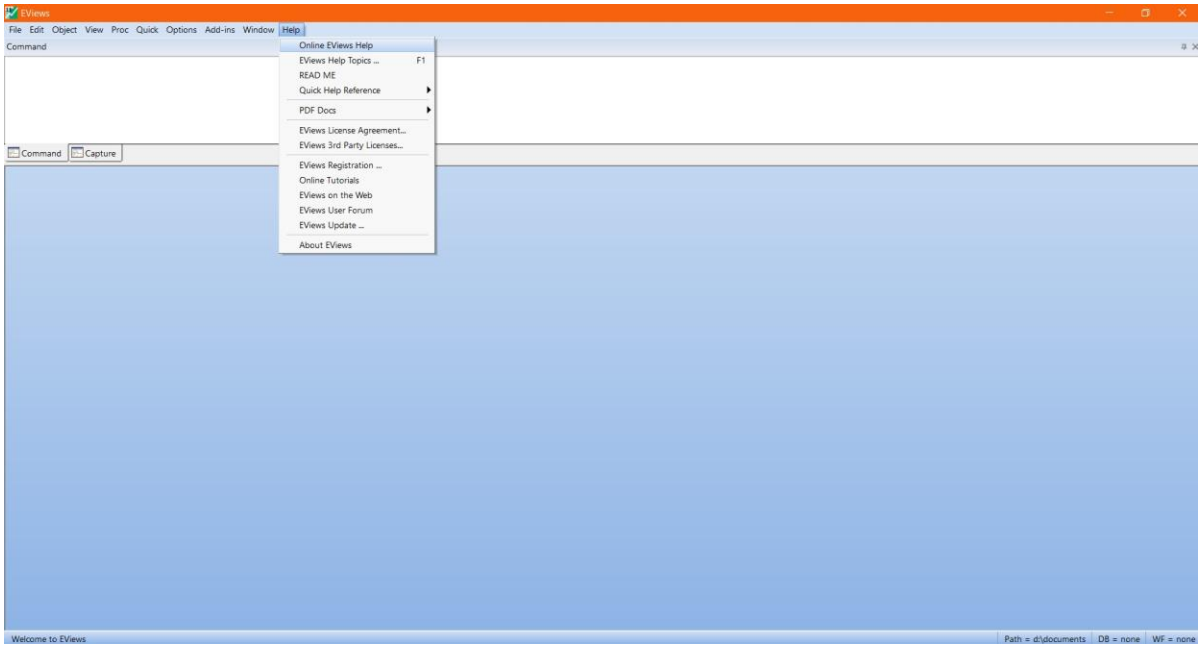


Εικόνα 2.23 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Εικόνα 2.24 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Τις επιλογές **Add-ins** και **Window** του μενού επιλογών δεν θα τις αναλύσουμε, καθώς δεν προσφέρουν κάτι παραπάνω στην ανάλυσή μας, ενώ η επιλογή **Help** του μενού επιλογών **(Εικόνα 2.25)** μας επιτρέπει, μεταξύ άλλων, να έχουμε πρόσβαση στις οδηγίες χρήσης του Eviews. Πιο συγκεκριμένα, αν επιλέξουμε **Online Eviews Help** από την επιλογή **Help**, θα μεταβούμε αυτόματα στη σελίδα http://www.eviews.com/help/helpintro.html στην οποία μπορούμε να ψάξουμε για οδηγίες σχετικά με τα αντικείμενα και τις λειτουργίες του Eviews που μας ενδιαφέρουν. Αν επιλέξουμε **Eviews Help Topics** από την επιλογή **Help**, θα μεταβούμε στις οδηγίες χρήσης που έχουν εγκατασταθεί από το Eviews στον υπολογιστή μας. Οι οδηγίες αυτές είναι ακριβώς οι ίδιες με αυτές που παρέχονται από το Online Eviews Help, ενώ και στις δύο περιπτώσεις υπάρχει η επιλογή **"Index"**, προκειμένου να χρησιμοποιήσουμε κάποια λέξη-κλειδί για το αντικείμενο ή τη λειτουργία του Eviews που μας ενδιαφέρει. Επίσης, αν επιλέξουμε **Quick Help Reference** από την επιλογή **Help**, θα μεταβούμε σε σύντομες οδηγίες χρήσης, χωρισμένες ανά ενότητα, οι οποίες είναι επίσης εγκατεστημένες στον υπολογιστή μας. Επιπλέον, επιλέγοντας **PDF Docs** από την επιλογή **Help**, αποκτάμε πρόσβαση στα εγχειρίδια χρήσης του Eviews, τα οποία είναι σε μορφή **pdf**, και μπορούμε να τα αποθηκεύσουμε στον υπολογιστή μας ή/και να τα εκτυπώσουμε. Τέλος, οι επιλογές **Online Tutorials** και **Eviews User Forum** από την επιλογή **Help** επιτρέπουν στον χρήστη να έχει πρόσβαση σε επιπλέον βοηθητικό υλικό, το οποίο είναι διαθέσιμο online (Microsoft PowerPoint slides, αντιμετώπιση προβλημάτων κλπ.).

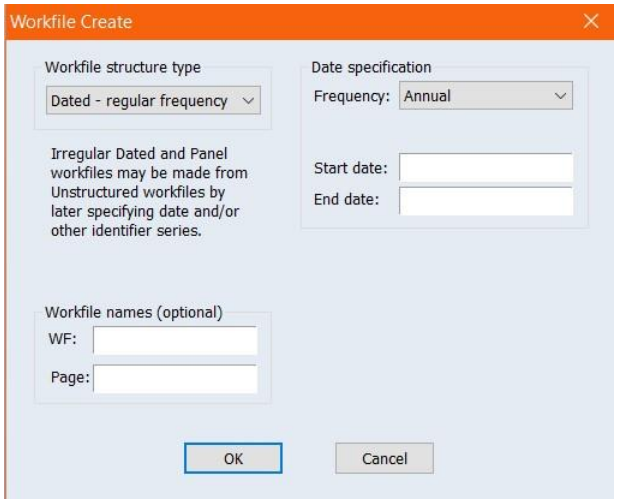


Εικόνα 2.25 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

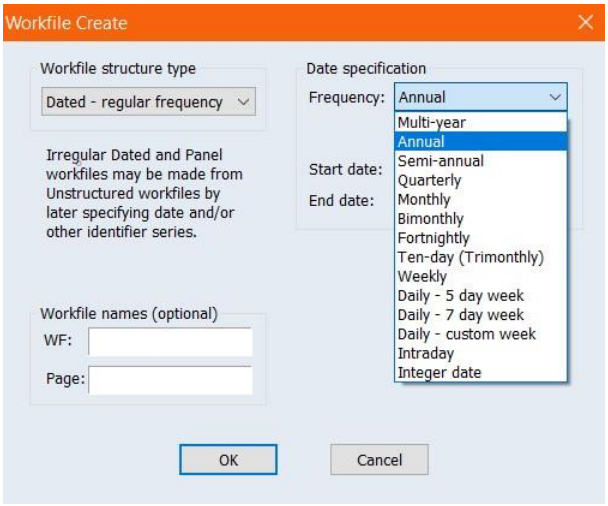
2.9 Δημιουργία Workfile του Eviews

Προκειμένου να δημιουργήσουμε ένα νέο workfile του Eviews, επιλέγουμε **File → New → Workfile**, με αποτέλεσμα να εμφανιστεί ένα παράθυρο όπως αυτό της **Εικόνας 2.26**. Στο παράθυρο αυτό μπορούμε να επιλέξουμε τη δομή που θα έχουν τα δεδομένα μας (**Workfile structure type**), τη συχνότητά τους (**Frequency**), καθώς και το εύρος του δείγματος (**Start date/End date**). Πιο συγκεκριμένα, η επιλογή Workfile structure type μας επιτρέπει να εισάγουμε 3 ειδών δεδομένα:

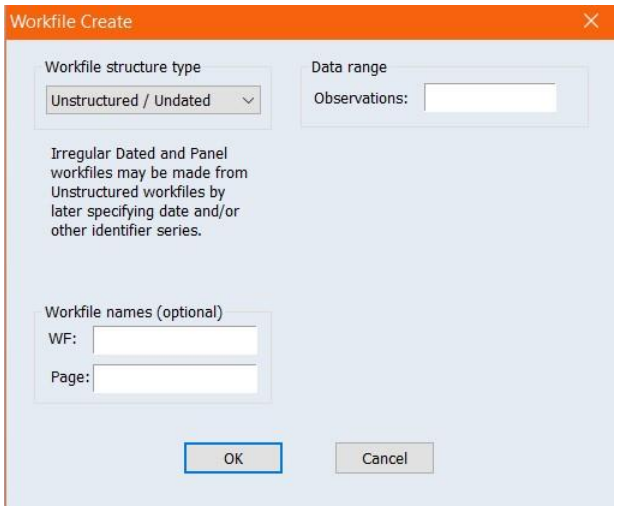
- Dated – regular frequency δεδομένα **(Εικόνα 2.26)**, δηλαδή δεδομένα που αντιστοιχούν σε κάποιο χρονικό εύρος. Αν τα δεδομένα μας είναι αυτής της μορφής, το επόμενο βήμα είναι να επιλέξουμε τη συχνότητά τους. Η επιλογή Frequency του Eviews μας δίνει πάρα πολλές επιλογές **(Εικόνα 2.27)**, όπως ετήσια, τριμηνιαία, μηνιαία, εβδομαδιαία, ημερήσια κλπ. Τέλος, θα πρέπει να επιλέξουμε τη χρονική περίοδο που αρχίζουν και τελειώνουν τα δεδομένα μας (Start date/End date). Αν, για παράδειγμα, τα δεδομένα μας είναι ετήσια και ξεκινούν από το 1980, ενώ τελειώνουν το 2020, επιλέγουμε Annual στο Frequency και συμπληρώνουμε 1980 στο Start date και 2020 στο End date. Αν τα δεδομένα μας είναι τριμηνιαία και ξεκινούν από το 1^{ο} τρίμηνο του 1980, ενώ τελειώνουν στο 3^{ο} τρίμηνο του 2020, επιλέγουμε Quarterly στο Frequency και συμπληρώνουμε 1980Q1 στο Start date και 2020Q3 στο End date. Τέλος, αν τα δεδομένα μας είναι μηνιαία και ξεκινούν από τον Ιανουάριο του 1980, ενώ τελειώνουν τον Ιούνιο του 2020, επιλέγουμε Monthly στο Frequency και συμπληρώνουμε 1980M1 στο Start date και 2020M3 στο End date.
- Unstructured/Undated δεδομένα **(Εικόνα 2.28)**, δηλαδή δεδομένα που δεν έχουν κάποια διάρθρωση ή δεν αντιστοιχούν σε κάποιες χρονικές περιόδους. Στην περίπτωση αυτή, όπως φαίνεται και από την **Εικόνα 2.28**, θα πρέπει απλά να εισάγουμε τον αριθμό των παρατηρήσεων που έχουμε (Observations). Αν έχουμε τέτοιου είδους δεδομένα, θα χρειαστεί να εξειδικεύσουμε τη μορφή τους στην πορεία της ανάλυσης.



Εικόνα 2.26 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

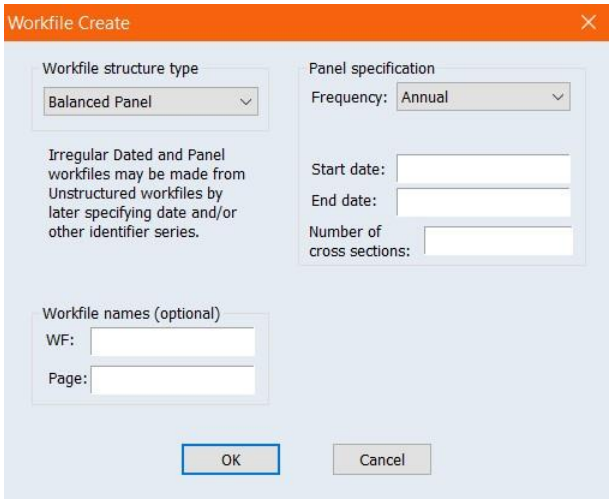


Εικόνα 2.27 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Εικόνα 2.28 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

- Balanced panel δεδομένα **(Εικόνα 2.29)**, δηλαδή δεδομένα που περιλαμβάνουν τόσο χρονολογικές σειρές (time series) όσο και διαστρωματικά στοιχεία (cross sections). Στην περίπτωση αυτή θα πρέπει να επιλέξουμε τη συχνότητα και το χρονικό εύρος των δεδομένων μας, καθώς και τον αριθμό των cross sections. Αν, για παράδειγμα, τα δεδομένα μας είναι ετήσια για τη χρονική περίοδο 1980-2020 και αφορούν μια σειρά μεταβλητών για 8 χώρες, θα επιλέξουμε Annual στο Frequency, θα συμπληρώσουμε 1980 στο Start date και 2020 στο End date, ενώ στο Number of cross sections θα συμπληρώσουμε 8.

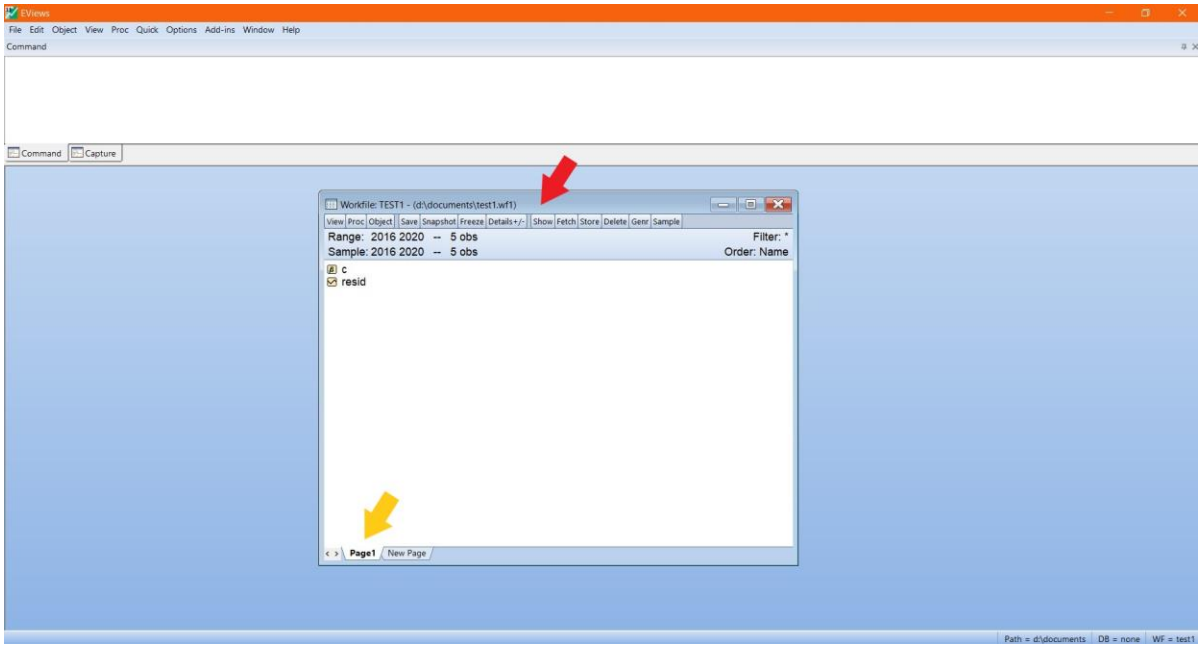


Εικόνα 2.29 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Τέλος, όποια επιλογή και αν κάνουμε, έχουμε τη δυνατότητα να δώσουμε εξαρχής όνομα στο workfile που δημιουργούμε, καθώς και στη σελίδα που θα περιλαμβάνει τα συγκεκριμένα δεδομένα (η σελίδα στο περιβάλλον του Eviews εμφανίζεται με αντίστοιχη μορφή όπως τα sheets στο περιβάλλον του Microsoft Excel). Οι επιλογές αυτές είναι προαιρετικές και εμφανίζονται στο κάτω αριστερό μέρος των **Εικόνων 2.26-2.29**, όπου το **WF** αντιστοιχεί στο Workfile και το **Page** στη σελίδα.

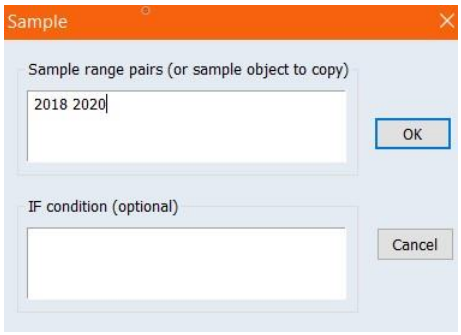
Έστω, λοιπόν, ότι έχουμε ετήσια στοιχεία για 5 έτη, από το 2016 έως το 2020. Για να δημιουργήσουμε το αντίστοιχο workfile, επιλέγουμε **Dated – regular frequency** στο **Workfile structure type**, **Annual** στο **Frequency**, **2016** στο **Start date** και **2020** στο **End date**. Προαιρετικά, ονομάζουμε **Test1** το workfile (στην επιλογή **WF**) και **Page1** τη σελίδα (στην επιλογή **Page**). Πατώντας **OK** μας εμφανίζεται το workfile που έχουμε φτιάξει **(Εικόνα 2.30)**. Επιλέγοντας **File → Save** από το μενού επιλογών μπορούμε να αποθηκεύσουμε το συγκεκριμένο workfile σε όποιο μέρος του υπολογιστή μας θέλουμε. Στο πάνω μέρος του workfile που έχει δημιουργηθεί εμφανίζεται ένα **toolbar** με μία σειρά από κουμπιά (που επισημαίνεται με το κόκκινο βέλος) και το οποίο, όπως θα δούμε στη συνέχεια, είναι εξαιρετικά χρήσιμο. Στο πάνω μέρος του workfile παρουσιάζεται, επίσης, το εύρος (**Range**) και το δείγμα (**Sample**) των δεδομένων μας (**2016 2020**), καθώς και ο αριθμός των παρατηρήσεων (**5 obs**). Στην **Εικόνα 2.30** το δείγμα και ο αριθμός των παρατηρήσεων είναι ίδια, αλλά αυτό δεν είναι απαραίτητο. Για να διαμορφώσουμε ένα μικρότερο δείγμα (για παράδειγμα, για την περίοδο 2018-2020), έχουμε 5 εναλλακτικούς τρόπους:

- Επιλέγουμε **Proc → Set Sample** από το μενού επιλογών.
- Επιλέγουμε **Quick → Sample** από το μενού επιλογών.
- Πατάμε το κουμπί **Sample** που είναι το τελευταίο στο toolbar του Eviews workfile.
- Κάνουμε διπλό κλικ στη λέξη **Sample** που βρίσκεται κάτω από το **Range**.
- Γράφουμε **smpl 2018 2020** στο **Command line** και πατάμε **Enter**.



Εικόνα 2.30 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Στις τέσσερις πρώτες περιπτώσεις θα εμφανιστεί ένα παράθυρο, όπως αυτό της **Εικόνας 2.31**, στο οποίο θα συμπληρώσουμε **2018 2020**, και θα πατήσουμε **OK**. Το νέο μας δείγμα έχει δημιουργηθεί. Πλέον, στο Eviews workfile εμφανίζεται **Range: 2016 2020 – 5 obs** και **Sample: 2018 2020 – 3 obs**. Επίσης, μπορούμε να δημιουργήσουμε ένα δείγμα που θα αποτελείται από 2 ή και περισσότερα μέρη των συνολικών παρατηρήσεων. Αν, για παράδειγμα, θέλουμε το νέο μας δείγμα να περιέχει τις παρατηρήσεις που αφορούν τα έτη 2016-2017 και 2019-2020, γράφουμε **2016 2017 2019 2020** στο παράθυρο της **Εικόνας 2.31** και πατάμε **OK** ή **smpl 2016 2017 2019 2020** στο **Command line** και πατάμε **Enter**. Το νέο μας δείγμα έχει δημιουργηθεί και στο Eviews workfile εμφανίζεται **Range: 2016 2020 – 5 obs** και **Sample: 2016 2017 2019 2020 – 4 obs**. Τέλος, γράφοντας **@all** στο παράθυρο της **Εικόνας 2.31** και πατώντας **OK** ή **smpl @all** στο **Command line** και πατώντας **Enter** επιστρέφουμε στην αρχική κατάσταση, όπου το δείγμα μας περιλαμβάνει το σύνολο των παρατηρήσεων. Θα πρέπει να επισημανθεί στο σημείο αυτό ότι, αν δημιουργήσουμε ένα νέο δείγμα, όλη η ανάλυση που θα πραγματοποιηθεί στη συνέχεια (στατιστικές εκτιμήσεις, παλινδρομήσεις, πίνακες, διαγράμματα κλπ.) θα αφορά το νέο αυτό δείγμα.



Εικόνα 2.31 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Στο κάτω μέρος του workfile που έχει δημιουργηθεί εμφανίζεται η σελίδα **Page1** που βρίσκονται τα δεδομένα μας (και η οποία επισημαίνεται με το κίτρινο βέλος). Μπορούμε, αν θέλουμε, να δημιουργήσουμε ξεχωριστές σελίδες επιλέγοντας **New Page**, όπου σε καθεμία από αυτές το δείγμα μας να είναι εντελώς διαφορετικό όσον αφορά τη διάρθρωσή του, το χρονικό εύρος και τη συχνότητα. Πιο αναλυτικά, κάνοντας κλικ στο **New Page** εμφανίζεται μια σειρά επιλογών, όπου επιλέγοντας **Specify by Frequency/Range** εμφανίζεται ξανά η **Εικόνα 2.26** και μπορούμε πλέον να ξεκινήσουμε τη διαδικασία από την αρχή.

Πλέον, στο κύριο παράθυρο του workfile **(Εικόνα 2.30)** εμφανίζονται 2 αντικείμενα: το **c** που αντιστοιχεί στον σταθερό όρο (constant term) και το **resid** που αντιστοιχεί στα κατάλοιπα (residuals). Κάνοντας κλικ στο **c** θα εμφανιστεί μια σειρά από μηδενικά, ενώ κάνοντας κλικ στο **resid** θα εμφανιστεί μια χρονολογική σειρά από το 2016 έως το 2020 (που είναι το εύρος του δείγματός μας), όπου σε όλα τα κελιά υπάρχει η ένδειξη **NA**. Η συγκεκριμένη ένδειξη σημαίνει ότι δεν υπάρχει διαθέσιμη πληροφορία (non-available). Στο **c** και στο **resid** θα εμφανιστούν αριθμοί, όταν θα πραγματοποιήσουμε την πρώτη στατιστική/οικονομετρική εκτίμηση. Θα πρέπει να επισημανθεί στο σημείο αυτό ότι στο Eviews το εικονίδιο  που υπάρχει αριστερά του **c** υποδηλώνει πάντα συντελεστή (coefficient), ενώ το εικονίδιο  που βρίσκεται αριστερά του **resid** υποδηλώνει πάντα στατιστική σειρά δεδομένων.

2.10 Εισαγωγή δεδομένων (data) στο Eviews και μετασχηματισμός τους

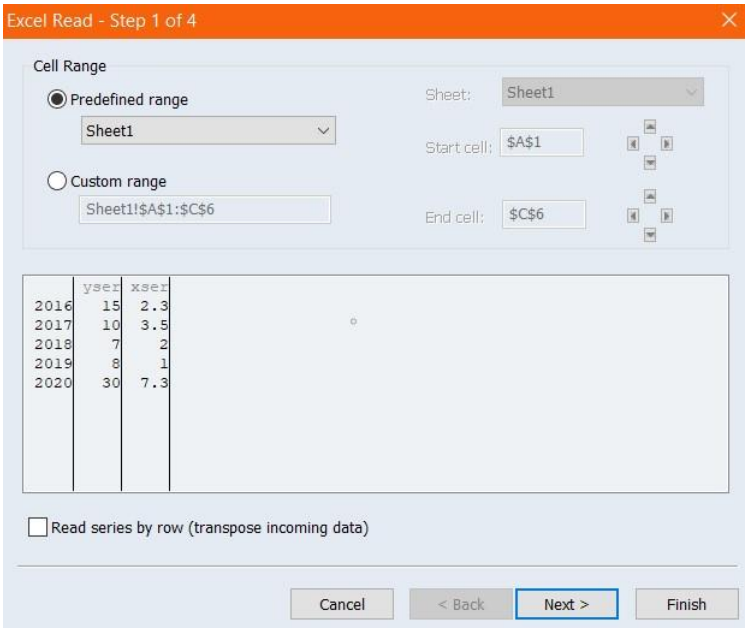
Προκειμένου να εισάγουμε στατιστικά δεδομένα (data) στο Eviews, έχουμε 3 εναλλακτικούς και ισοδύναμους τρόπους: (α) εισάγοντάς τα από αρχείο του Microsoft Excel, (β) χρησιμοποιώντας την επιλογή copy/paste από το Microsoft Excel στο Eviews, και (γ) πληκτρολογώντας τα δεδομένα απευθείας στο Eviews. Προφανώς, ο τρίτος τρόπος είναι εύχρηστος μόνο στην περίπτωση που τα δεδομένα που θέλουμε να εισάγουμε είναι πολύ λίγα, καθώς σε αντίθετη περίπτωση καθίσταται εξαιρετικά χρονοβόρος και εμπεριέχει μεγάλη πιθανότητα λάθους.

- Εισαγωγή μέσω αρχείου του Microsoft Excel: Έστω ότι έχουμε φτιάξει ένα αρχείο Microsoft Excel, στο οποίο έχουμε ετήσια δεδομένα δύο χρονολογικών σειρών για την περίοδο 2016-2020 (όπως και το workfile Test1 που έχουμε ήδη δημιουργήσει). Οι σειρές αυτές έχουν ονομαστεί yser και xser και είναι σε στήλες **(Εικόνα 2.32)**.

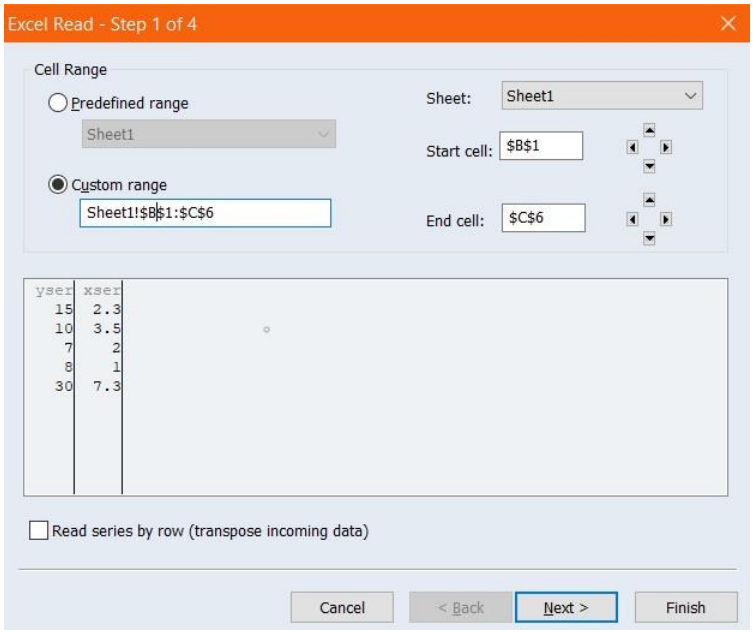
	A	B	C	D
1		yser	xser	
2	2016	15	2.3	
3	2017	10	3.5	
4	2018	7	2	
5	2019	8	1	
6	2020	30	7.3	
7				

Εικόνα 2.32 *Microsoft Excel Worksheet.*

Για να εισάγουμε τα δεδομένα μας στο Eviews, επιλέγουμε από το μενού επιλογών **File → Import → Import from file**, βρίσκουμε στον υπολογιστή μας το αρχείο του Microsoft Excel από το οποίο θέλουμε να εισάγουμε τα **data** και επιλέγουμε **Open**. Αμέσως μας εμφανίζεται ένα παράθυρο με την ονομασία **Excel Read - Step 1 of 4 (Εικόνα 2.33)**, το οποίο μας δείχνει τα **data** που πρόκειται να εισάγουμε στο Eviews. Στο παράθυρο αυτό, είναι προεπιλεγμένο το **Predefined range**, κάτι που σημαίνει ότι το Eviews θα «τραβήξει» όλα τα data που βρίσκονται στο Sheet1 του Microsoft Excel (άλλωστε δεν έχουμε data σε κάποιο άλλο sheet). Αντί για το **Predefined range**, μπορούμε να επιλέξουμε το **Custom range** όπου μας δίνεται πλέον η δυνατότητα να επιλέξουμε ορισμένα data από το Sheet1 ή data από άλλα sheets (εφόσον υπάρχουν). Στο παράδειγμά μας θα επιλέξουμε **Custom range** και θα γράψουμε στο αντίστοιχο κουτάκι **Sheet1!$B$1:$C$6 (Εικόνα 2.34)**. Κάνοντας αυτήν την επιλογή, δίνουμε εντολή στο Eviews να μην «τραβήξει» τα data της στήλης A (καθώς αυτή περιέχει τα έτη 2016-2020 που αποτελούν το δείγμα του Eviews workfile και δεν μας ενδιαφέρει να τα εισάγουμε ως χρονολογική σειρά), αλλά να «τραβήξει» τα data που ξεκινούν από το 1^{ο} κελί της στήλης B μέχρι το 6^{ο} κελί της στήλης C **(Εικόνες 2.32** και **2.34)**.

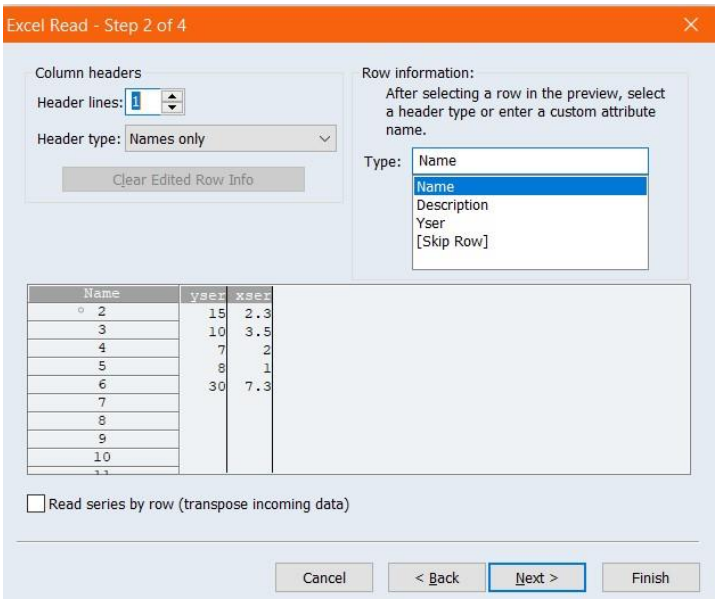


Εικόνα 2.33 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



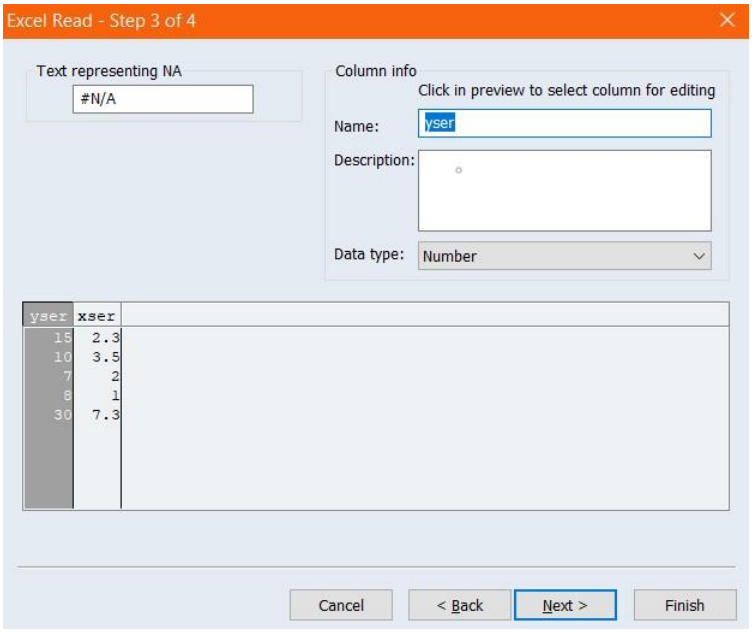
Εικόνα 2.34 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Θα πρέπει να επισημάνουμε στο σημείο αυτό ότι οι επιλογές Sheet:, Start cell: και End cell ενεργοποιούνται μόνο όταν επιλέξουμε **Custom range**. Τέλος, στο κάτω μέρος του παραθύρου των **Εικόνων 2.33-2.34** υπάρχει η επιλογή **Read series by row (transpose incoming data)**. Αν «τσεκάρουμε» το συγκεκριμένο κουτάκι, τα δεδομένα μας θα αναστραφούν και θα εμφανιστούν σε γραμμές. Οπότε, δεν το «τσεκάρουμε» και επιλέγουμε **Next**. Θα εμφανιστεί ένα παράθυρο με την ονομασία **Excel Read - Step 2 of 4 (Εικόνα 2.35)**, στο οποίο μπορούμε να μορφοποιήσουμε τα ονόματα των χρονολογικών σειρών (**Header type**) και να εισάγουμε πληροφορίες για κάθε έτος. Επίσης, παραμένει η επιλογή **Read series by row (transpose incoming data)**. Δεν κάνουμε τίποτα παραπάνω στο βήμα αυτό και επιλέγουμε **Next**.



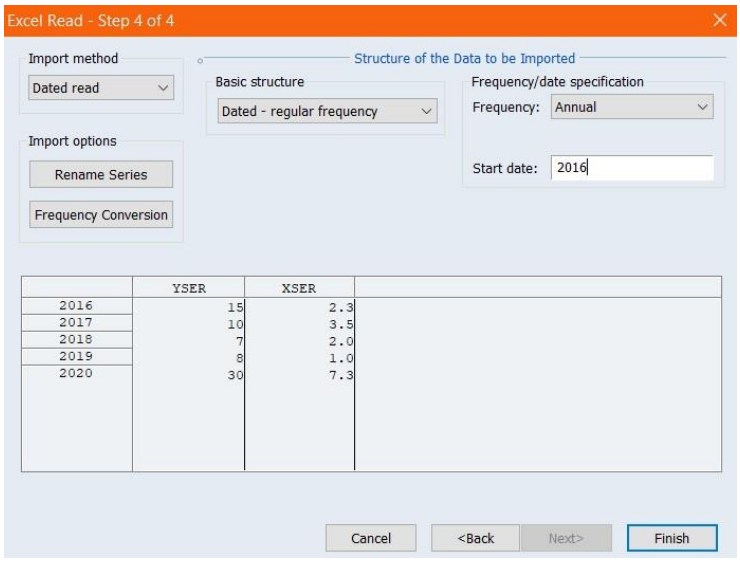
Εικόνα 2.35 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Στο παράθυρο **Excel Read - Step 3 of 4** που εμφανίζεται **(Εικόνα 2.36)** μπορούμε να επιλέξουμε το κείμενο που θα εμφανίζεται σε κάποιο κελί, όταν δεν υπάρχουν δεδομένα σε αυτό. Επίσης, κάνοντας κλικ σε κάθε σειρά μπορούμε να αλλάξουμε το όνομά της και να της δώσουμε κάποια περιγραφή. Δεν κάνουμε τίποτα στο βήμα αυτό και επιλέγουμε **Next**. Το τελευταίο παράθυρο που εμφανίζεται (**Excel Read - Step 4 of 4**) μας επιτρέπει να μετονομάσουμε τις στατιστικές σειρές που πρόκειται να εισάγουμε στο Eviews και να αλλάξουμε, αν θέλουμε, τη διάρθρωση και τη συχνότητα των δεδομένων μας. Πρέπει, επίσης, να εισάγουμε και την ημερομηνία από την οποία ξεκινάει το δείγμα μας. Γράφουμε 2016 **(Εικόνα 2.37)** και επιλέγουμε **Finish**. Στη συνέχεια επιλέγουμε **No** στο ερώτημα "**Link imported series and alpha object(s) to external source?**" που θα εμφανιστεί, προκειμένου να αποφύγουμε τη σύνδεση των data με άλλα προγράμματα, και η εισαγωγή των data έχει ολοκληρωθεί. Όπως φαίνεται στην **Εικόνα 2.38**, οι χρονολογικές σειρές **yser** και **xser** έχουν πλέον εμφανιστεί στο κεντρικό παράθυρο του Eviews workfile.

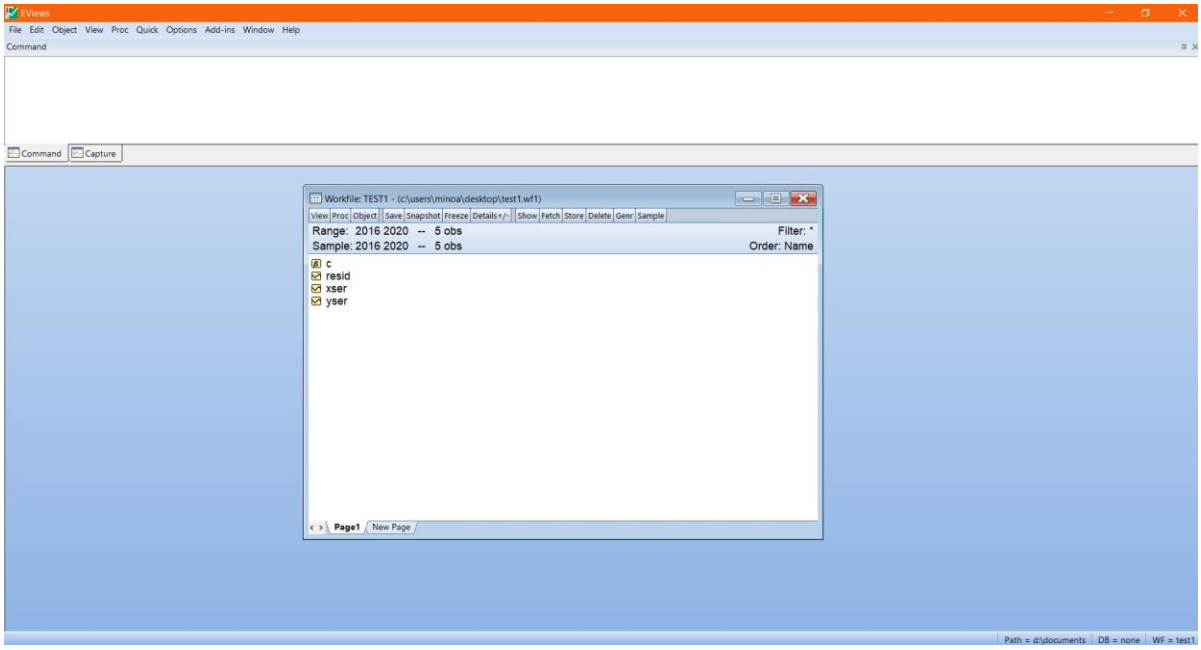


Εικόνα 2.36 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Καθώς οι επιλογές που έχει καθένα από τα παραπάνω 4 βήματα είναι πολλές και δεν είναι δυνατή η ανάλυση καθεμίας από αυτές στο παρόν εγχειρίδιο, συστήνεται ο χρήστης του Eviews να προβεί σε αρκετές δοκιμές σχετικά με την εισαγωγή δεδομένων από το Microsoft Excel στο Eviews.



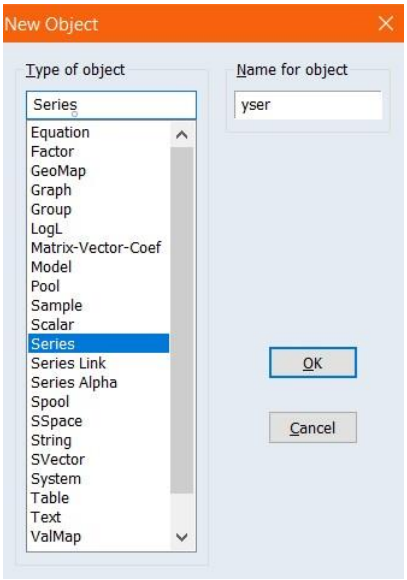
Εικόνα 2.37 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



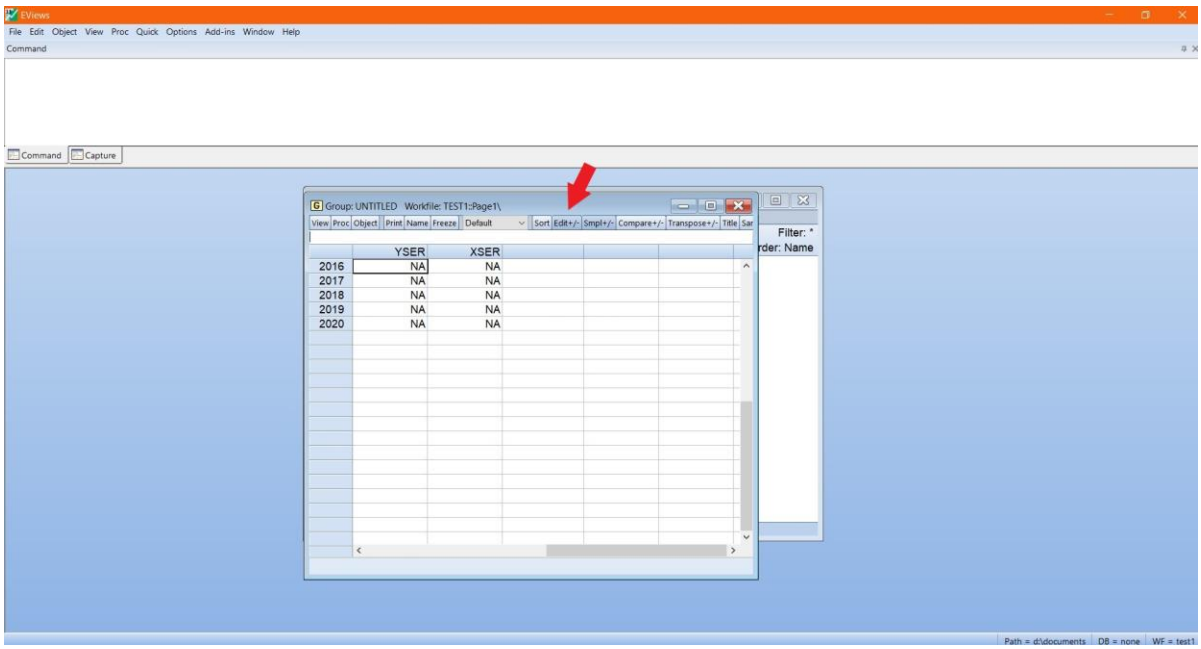
Εικόνα 2.38 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

- **Copy/paste από το Microsoft Excel στο Eviews:** Για να εισάγουμε data με αυτόν τον τρόπο στο Eviews, χρησιμοποιούμε την επιλογή **Object** από το μενού επιλογών, η οποία αποτελεί μία από τις πιο χρηστικές εντολές του Eviews. Επιλέγοντας **Object → New Object**, ανοίγει ένα παράθυρο που μας δίνει τη δυνατότητα να εισάγουμε διάφορα αντικείμενα (**objects**) στο Eviews workfile, όπως εξισώσεις (**Equation**), διαγράμματα (**Graph**), μήτρες και διανύσματα (**Matrix-Vector-Coef**), μοντέλα (**Model**), στατιστικές σειρές (**Series**), αρχεία κειμένου (**Text**) κλπ. Στην παρούσα φάση και προκειμένου να εισάγουμε τις στατιστικές μας σειρές, επιλέγουμε **Series**, δίνουμε το όνομα **yser** στο **Name for object** και επιλέγουμε **OK (Εικόνα 2.39)**. Επαναλαμβάνουμε τη διαδικασία και για τη σειρά **xser** και προφανώς για όσες επιπλέον

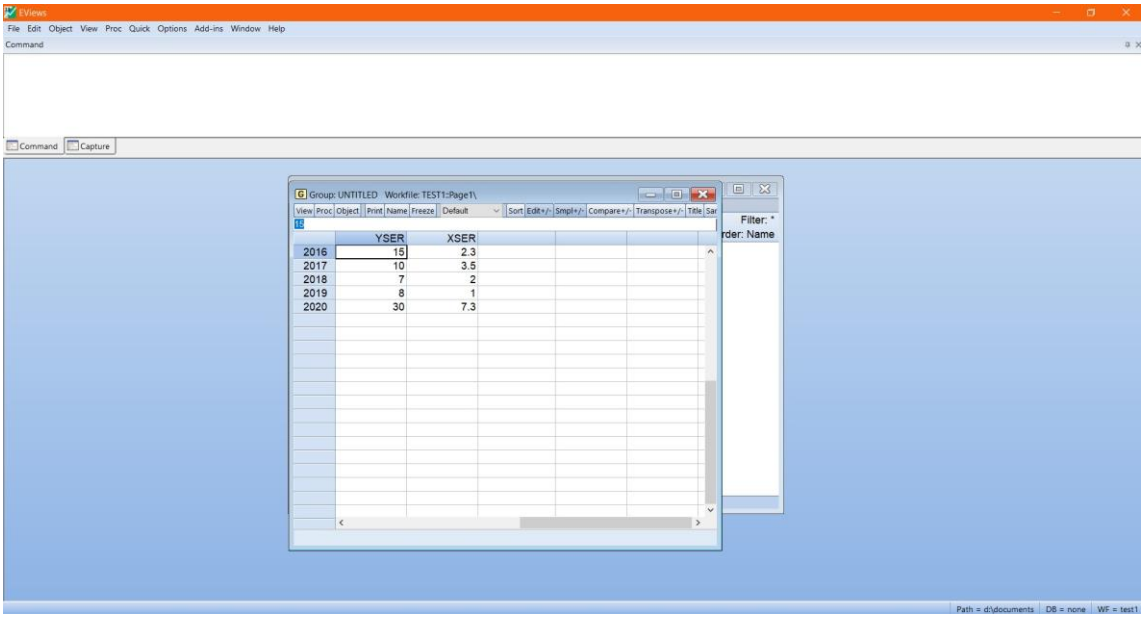
σειρές θέλουμε. Με αυτόν τον τρόπο, έχουν δημιουργηθεί οι 2 χρονολογικές σειρές **yser** και **xser**, οι οποίες όμως είναι προς το παρόν κενές, καθώς, αν τις μαρκάρουμε και τις ανοίξουμε (πατώντας **Enter**), θα δούμε ότι σε όλα τα κελιά υπάρχει η ένδειξη **NA** (non-available), όπως φαίνεται και στην **Εικόνα 2.40**. Για να είναι δυνατή η διαδικασία copy/paste από το Microsoft Excel, θα πρέπει οπωσδήποτε να είναι ενεργοποιημένο το **editing mode** του Eviews και αυτό γίνεται πατώντας το κουμπί **Edit+/-**, το οποίο βρίσκεται στο σημείο που υποδεικνύει το κόκκινο βέλος. Οπότε, επιλέγουμε τα κελιά που μας ενδιαφέρουν από το Microsoft Excel (B2 έως C6 στην **Εικόνα 2.32**), δίνουμε την εντολή copy, μεταβαίνουμε στο περιβάλλον της **Εικόνας 2.40** και δίνουμε την εντολή paste. Πλέον οι σειρές μας είναι έτοιμες **(Εικόνα 2.41)**. Αν θέλουμε, μπορούμε να απενεργοποιήσουμε το **editing mode** του Eviews, πατώντας και πάλι το κουμπί **Edit+/-**, προκειμένου να αποφύγουμε μια κατά λάθος τροποποίηση των δεδομένων μας.



Εικόνα 2.39 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Εικόνα 2.40 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

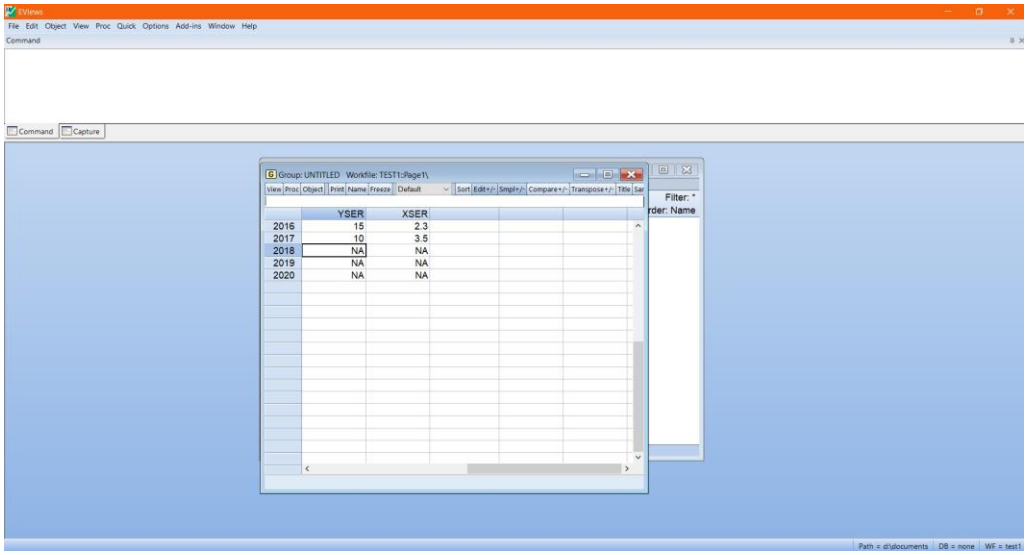


Εικόνα 2.41 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

- **Απευθείας πληκτρολόγηση των δεδομένων στο Eviews:** Για να εισάγουμε δεδομένα (data) με αυτόν τον τρόπο στο Eviews, ακολουθούμε μέχρι κάποιο σημείο την ίδια διαδικασία που εφαρμόσαμε προηγουμένως. Δημιουργούμε δύο νέες κενές σειρές (**yser** και **xser**) επιλέγοντας **Object → New Object**, τις ανοίγουμε και ενεργοποιούμε το **editing mode** του Eviews πατώντας το κουμπί **Edit+/**. Αντί, όμως, να δώσουμε την εντολή copy στο Microsoft Excel, αρχίζουμε και πληκτρολογούμε τα δεδομένα μας **(Εικόνα 2.42)**. Όταν ολοκληρωθεί η πληκτρολόγηση, απενεργοποιούμε το **editing mode** του Eviews πατώντας και πάλι το κουμπί **Edit+/,** και οι στατιστικές σειρές μας είναι πλέον έτοιμες. Όπως είπαμε και παραπάνω, αυτό ο τρόπος εισαγωγής δεδομένων είναι χρηστικός μόνο όταν τα δεδομένα μας είναι λίγα σε αριθμό, καθώς σε διαφορετική περίπτωση είναι αρκετά χρονοβόρος και εμπεριέχει μεγάλη πιθανότητα λάθους.

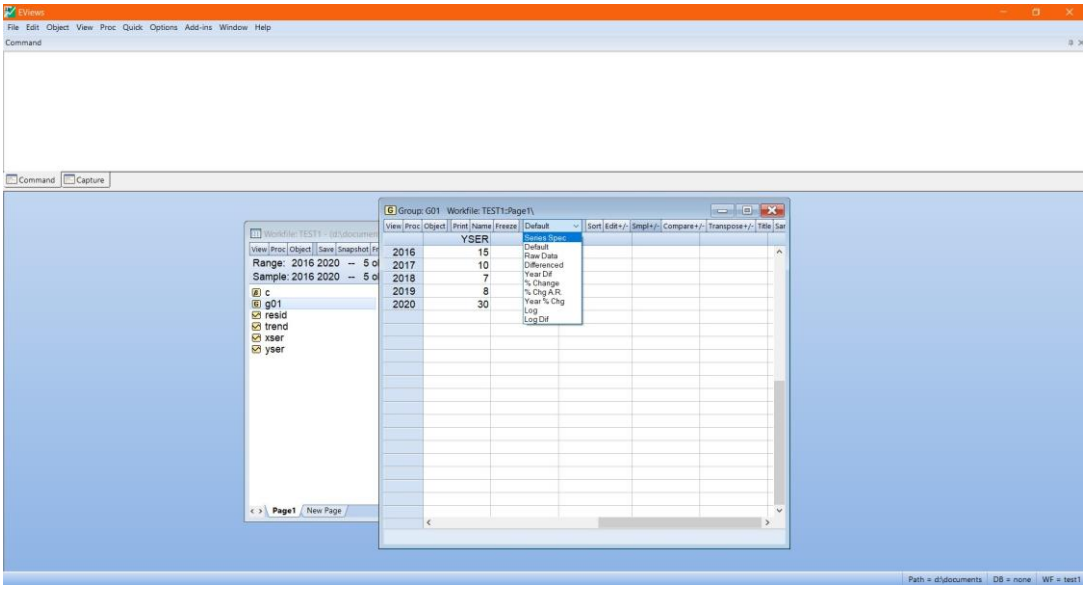
Θα πρέπει να επισημανθεί στο σημείο αυτό ότι, όταν βρισκόμαστε σε ένα περιβάλλον όπως αυτό της **Εικόνας 2.41**, και αποφασίσουμε να «κλείσουμε» τις σειρές, για να βρεθούμε στο αρχικό παράθυρο του Eviews workfile, θα εμφανιστεί το μήνυμα "**Delete Untitled GROUP?**". Ο λόγος είναι ότι το Eviews θεωρεί ότι τα δεδομένα μας είναι ανοικτά ως Group, στο οποίο δεν έχει δοθεί τίτλος (**untitled**) και το οποίο δεν έχει αποθηκευτεί. Αν θέλουμε να αποθηκεύσουμε το συγκεκριμένο **Group**, πατάμε **No** στο παραπάνω ερώτημα, και στη συνέχεια το κουμπί **Name** που βρίσκεται στο toolbar του workfile, προκειμένου να του δώσουμε κάποιο όνομα (έστω **G01**). Σε μια τέτοια περίπτωση, το συγκεκριμένο **Group** αποθηκεύεται αυτόματα και εμφανίζεται ως αντικείμενο στο Eviews workfile με ένα εικονίδιο  στα αριστερά του. Αν πάλι δεν μας ενδιαφέρει κάτι τέτοιο, επιλέγουμε **Yes** στο παραπάνω ερώτημα, με αποτέλεσμα το Group να εξαφανιστεί, χωρίς προφανώς να έχουν διαγραφεί οι σειρές **yser** και **xser** από το workfile.

Εναλλακτικά, ένα Group σειρών μπορεί να δημιουργηθεί ως εξής: στην αρχική οθόνη του workfile επιλέγουμε τις σειρές που θέλουμε να εντάξουμε σε αυτό, πατάμε δεξί κλικ, επιλέγουμε **Open → as Group** και πατάμε **Enter**. Εμφανίζονται οι σειρές που έχουμε επιλέξει μαζί με το περιεχόμενό τους. Στη συνέχεια, πατώντας **Name** στο toolbar του workfile το ονομάζουμε όπως επιθυμούμε (έστω **G01**, όπως παραπάνω). Το συγκεκριμένο Group θα εμφανιστεί αυτόματα ως αντικείμενο στο Eviews workfile, με το εικονίδιο  στα αριστερά του.



Εικόνα 2.42 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Με διπλό κλικ σε κάθε αντικείμενο του Eviews μπορούμε να δούμε το περιεχόμενό του, ενώ με δεξί κλικ ανοίγεται μια λίστα που έχει διάφορες επιλογές, όπως Copy, Paste σε κάποια άλλη page του Eviews workfile, Delete, Rename κλπ. Επίσης, όταν βρισκόμαστε σε ένα περιβάλλον όπως αυτό της **Εικόνας 2.41** και ανοίξουμε τη λίστα που βρίσκεται στην επιλογή **Default**, μπορούμε να δούμε διάφορους μετασχηματισμούς των στατιστικών μας σειρών, όπως τις πρώτες διαφορές τους, τις ποσοστιαίες μεταβολές τους, τους φυσικούς λογαρίθμους τους κλπ. **(Εικόνα 2.43)**.



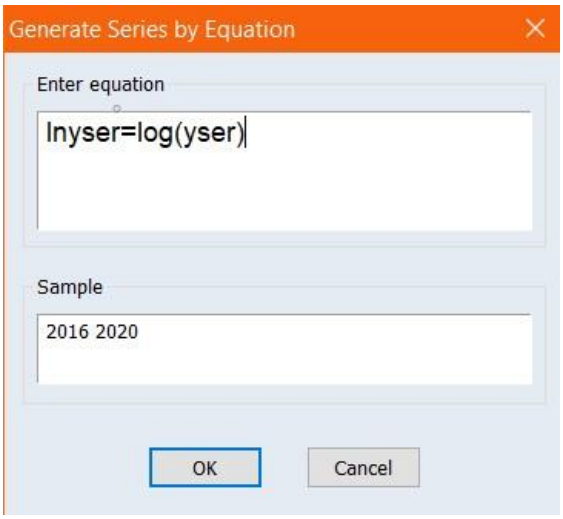
Εικόνα 2.43 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Μία πολύ σημαντική λειτουργία που έχει το Eviews είναι ότι μας παρέχει τη δυνατότητα να δημιουργήσουμε νέες στατιστικές σειρές από τον μετασχηματισμό των ήδη υπαρχουσών σειρών. Έστω, λοιπόν, ότι θέλουμε να δημιουργήσουμε τέσσερις νέες στατιστικές σειρές από τη σειρά **yser**, οι οποίες θα είναι **(α) ο φυσικός λογάριθμός της, (β) η πρώτη διαφορά της, (γ) η αντίστροφή της, και (δ) η ίδια σειρά με μια χρονική υστέρηση**. Οι τέσσερις νέες σειρές μπορούν να δημιουργηθούν με 3 τρόπους:

- Ο πρώτος τρόπος είναι να δημιουργήσουμε τις τέσσερις νέες σειρές, επιλέγοντας **Object → Generate Series** ή **Quick → Generate Series** για καθεμία από αυτές. Στο παράθυρο που θα εμφανιστεί θα γράψουμε την κατάλληλη συνάρτηση στο **Enter equation** και θα πατήσουμε

OK: **lnyser=log(yser)** για τον φυσικό λογάριθμο**, fdyser=d(yser)** για την πρώτη διαφορά**, invyser=1/yser** για την αντίστροφη σειρά και **olyser=yser(-1)** για τη σειρά με μια χρονική υστέρηση. Θα πρέπει να σημειωθεί στο σημείο πως τα ονόματα που δώσαμε στις νέες σειρές (**lnyser, fdyser, invyser** και **olyser**) είναι ενδεικτικά, καθώς ο χρήστης μπορεί να ονομάσει τις σειρές αυτές όπως επιθυμεί. Η περίπτωση του φυσικού λογαρίθμου της σειράς **yser** εμφανίζεται στην **Εικόνα 2.44**.

- Ο δεύτερος τρόπος είναι να δημιουργήσουμε τέσσερις νέες κενές σειρές, επιλέγοντας **Object → New Object → Series** για καθεμία από αυτές**,** και να τις ονομάσουμε όπως παραπάνω: **lnyser, fdyser, invyser, olyser**. Στη συνέχεια, για καθεμία από αυτές γράφουμε την κατάλληλη συνάρτηση στο **Command line** και πατάμε **Enter**: **lnyser=log(yser), fdyser=d(yser), invyser=1/yser** και **olyser=yser(-1)**.



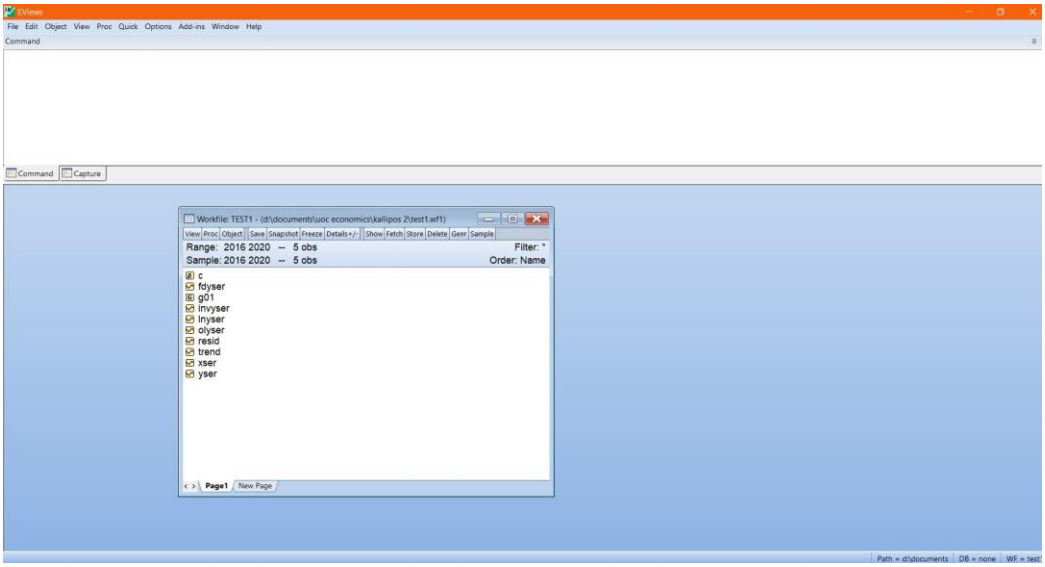
Εικόνα 2.44 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

- Ο τρίτος τρόπος είναι να γράψουμε κατευθείαν τις κατάλληλες συναρτήσεις στο **Command line** και να πατήσουμε **Enter** μετά από καθεμία: **genr lnyser=log(yser), genr fdyser=d(yser), genr invyser=1/yser** και **genr olyser=yser(-1)**. Στην περίπτωση αυτή, πριν από κάθε συνάρτηση θα πρέπει να γραφτεί η εντολή **genr** (από το generate που σημαίνει δημιουργώ), καθώς οι αντίστοιχες χρονολογικές σειρές δεν έχουν προηγουμένως δημιουργηθεί στο workfile μας. Η περίπτωση του φυσικού λογαρίθμου της μεταβλητής **yser** εμφανίζεται στην **Εικόνα 2.45**.

Εικόνα 2.45 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Οι τέσσερις νέες σειρές έχουν πλέον δημιουργηθεί. Εκτός από τις συναρτήσεις που αφορούν τον μετασχηματισμό των υπαρχουσών στατιστικών σειρών, το Eviews περιλαμβάνει και μια σειρά από αυτοματοποιημένες συναρτήσεις. Η πιο συνηθισμένη από αυτές είναι η γραμμική τάση

(**linear trend**), δηλαδή μια μεταβλητή που παίρνει την τιμή 0 στην πρώτη παρατήρηση και αυξάνεται κατά 1 μονάδα ανά παρατήρηση. Η συνάρτηση του Eviews είναι **@trend(YYYY)**, όπου YYYY είναι το έτος από το οποίο ξεκινάει η γραμμική τάση. Οπότε, για να εισάγουμε μια γραμμική τάση στο Eviews workfile που έχουμε δημιουργήσει και το οποίο περιλαμβάνει τις σειρές **yser** και **xser,** επιλέγουμε **Quick → Generate Series**. Στο παράθυρο που εμφανίζεται γράφουμε τη συνάρτηση **trend=@trend(2016)** στο **Enter equation** και πατάμε **OK**. Πλέον η γραμμική τάση έχει δημιουργηθεί και εμφανίζεται με το όνομα trend, το οποίο προφανώς είναι ενδεικτικό, καθώς ο χρήστης μπορεί να ονομάσει τη σειρά αυτή όπως επιθυμεί. Το Eviews workfile με τις δύο αρχικές σειρές, το Group που έχει φτιαχτεί, τις τέσσερις νέες σειρές και τη γραμμική τάση εμφανίζεται στην **Εικόνα 2.46**. Συνιστάται ο χρήστης του Eviews να διερευνήσει μέσα από την εντολή **Help** τόσο τις διάφορες συναρτήσεις που αφορούν τον μετασχηματισμό των στατιστικών σειρών όσο και τις αυτοματοποιημένες συναρτήσεις.

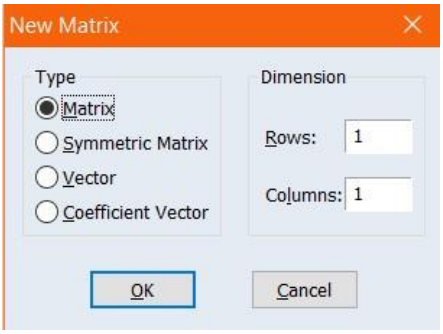


Εικόνα 2.46 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

2.11 Εισαγωγή μητρών και πράξεις μητρών στο Eviews

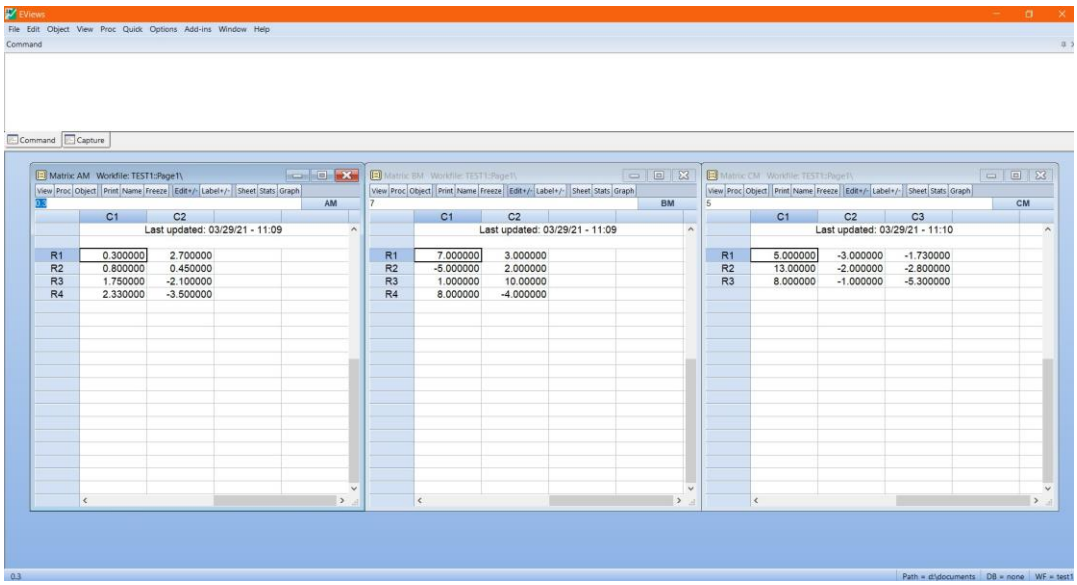
Το Eviews μας παρέχει, επίσης, τη δυνατότητα να εισάγουμε μήτρες σε αυτό και να πραγματοποιήσουμε πράξεις με αυτές. Για να εισάγουμε μια μήτρα, επιλέγουμε **Object → New Object** από το μενού επιλογών ή από το toolbar του Eviews workfile. Μας εμφανίζεται το παράθυρο που φαίνεται στην **Εικόνα 2.39**, στο οποίο επιλέγουμε **Matrix-Vector-Coef**. Στη συνέχεια, αφού δώσουμε στη νέα μήτρα το όνομα που επιθυμούμε και πατήσουμε **OK**, εμφανίζεται ένα νέο παράθυρο **(Εικόνα 2.47**) που μας ζητάει να επιλέξουμε τη μορφή της μήτρας που θέλουμε να εισάγουμε, καθώς και τις διαστάσεις της, δηλαδή τον αριθμό των γραμμών (**Rows**) και στηλών (**Columns**) που θα έχει.

Μπορούμε να εισάγουμε μια απλή μήτρα (**Matrix**), μια συμμετρική μήτρα (**Symmetric Matrix**), ένα διάνυσμα (**Vector**) ή ένα διάνυσμα συντελεστών (**Coefficient Vector**). Έστω, λοιπόν, ότι θέλουμε να εισάγουμε τρεις νέες απλές μήτρες, τις **am** και **bm**, διαστάσεων 4×2 και τη **cm** διαστάσεων 3×3. Για την πρώτη μήτρα, επιλέγουμε **Object → New Object → Matrix-Vector-Coef**, δίνουμε το όνομα **am** και πατάμε **OK**. Στο παράθυρο που εμφανίζεται **(Εικόνα 2.47)**, επιλέγουμε **Matrix,** ενώ στο κελί **Row** γράφουμε 4, στο κελί **Column** γράφουμε 2 και πατάμε OK. Αντίστοιχη διαδικασία ακολουθούμε, για να εισάγουμε και τις άλλες δύο μήτρες. Οι τρεις νέες μήτρες έχουν πλέον δημιουργηθεί ως αντικείμενα στο Eviews workfile, ενώ αριστερά τους εμφανίζεται το εικονίδιο , το οποίο υποδηλώνει Matrix, Symmetric Matrix ή Vector (αν στο παράθυρο της **Εικόνας 2.47** είχαμε επιλέξει Coefficient Vector, θα εμφανιζόταν το εικονίδιο).



Εικόνα 2.47 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Προκειμένου να εισάγουμε αριθμούς στα κελιά των μητρών (είτε μέσω copy/paste από κάποιο άλλο πρόγραμμα, όπως το Microsoft Excel, ή πληκτρολογώντας τους), «ανοίγουμε» τις μήτρες με διπλό κλικ και ενεργοποιούμε το **editing mode** πατώντας **Edit +/-** στο toolbar του Eviews workfile **(Εικόνα 2.48)**. Στο παράδειγμά μας έχουμε εισάγει κάποιους τυχαίους αριθμούς. Θα πρέπει, επίσης, να σημειωθεί ότι, αν επιλέξουμε να εισάγουμε μια Symmetric Matrix, αρκεί να εισάγουμε αριθμούς στα μισά κελιά, καθώς τα συμμετρικά κελιά θα συμπληρωθούν αυτόματα.



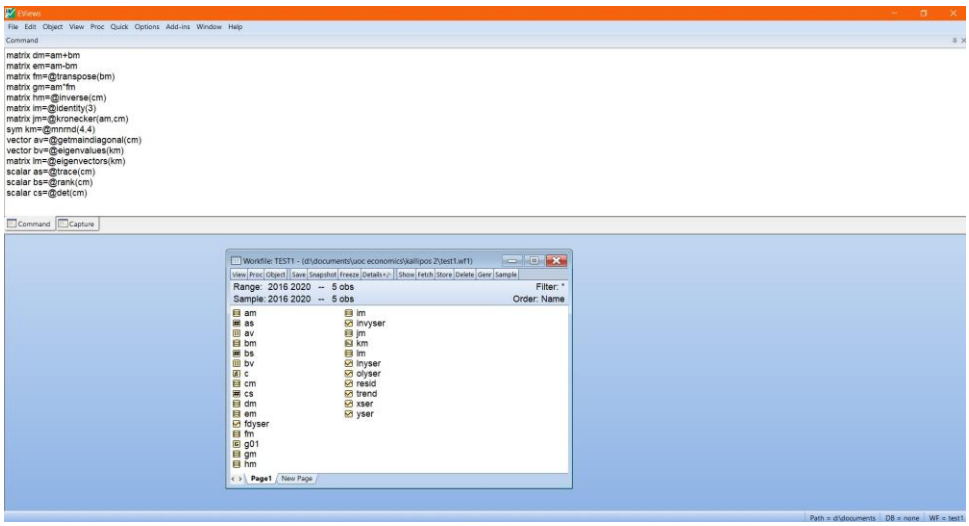
Εικόνα 2.48 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Έχοντας δημιουργήσει τις τρεις νέες μήτρες, μπορούμε πλέον να κάνουμε διάφορες πράξεις με αυτές. Η συγκεκριμένη διαδικασία είναι πολύ απλή και την πραγματοποιούμε εισάγοντας την αντίστοιχη «συνάρτηση» στο **Command line** και πατώντας **Enter**. Επίσης, το Eviews παρέχει μια πληθώρα αυτοματοποιημένων συναρτήσεων που αφορούν την άλγεβρα μητρών (προτείνεται ο χρήστης να διερευνήσει μέσα από το **Help** του Eviews τις συναρτήσεις αυτές). Ενδεικτικά, αναφέρουμε μερικές από αυτές τις συναρτήσεις:

- **matrix dm=am+bm** → δημιουργία μήτρας **dm** που είναι το άθροισμα των μητρών **am** και **bm**. Οι μήτρες **am** και **bm** θα πρέπει υποχρεωτικά να έχουν τις ίδιες διαστάσεις, καθώς, αν προσπαθήσουμε να αθροίσουμε μήτρες διαφορετικών διαστάσεων, οι οποίες δεν μπορούν να αθροιστούν, θα λάβουμε το μήνυμα "**Sizes do not match in matrix function**".
- **matrix em=am-bm** → δημιουργία μήτρας **em** που είναι η διαφορά των μητρών **am** και **bm**. Και στην περίπτωση αυτή, αν προσπαθήσουμε να πάρουμε τη διαφορά των μητρών διαφορετικών διαστάσεων, θα λάβουμε το μήνυμα "**Sizes do not match in matrix function**".
- **matrix fm=@transpose(bm)** → δημιουργία μήτρας **fm** που είναι η ανάστροφη (transpose) της μήτρας **bm**.

- **matrix gm=am*fm** → δημιουργία μήτρας **gm** που είναι το γινόμενο των μητρών **am** και **fm**. Και στην περίπτωση αυτή, αν οι διαστάσεις των μητρών δεν επιτρέπουν τον πολλαπλασιασμό τους, θα λάβουμε το μήνυμα "**Sizes do not match in matrix function**".
- **matrix hm=@inverse(cm)** → δημιουργία μήτρας **gm** που είναι η αντίστροφη (inverse) της τετραγωνικής μήτρας **em**. Αν προσπαθήσουμε να αντιστρέψουμε μια μη-τετραγωνική μήτρα, θα λάβουμε το μήνυμα "**Attempt to invert or decompose non square matrix**". Επίσης, αν προσπαθήσουμε να αντιστρέψουμε μια μήτρα με ορίζουσα (determinant) ίση με 0, δηλαδή μια μήτρα που δεν αντιστρέφεται, θα λάβουμε το μήνυμα "**Near singular matrix**".
- **matrix im=@identity(3)** → δημιουργία μοναδιαίας μήτρας (identity matrix) **im** διαστάσεων 3×3.
- **matrix jm=@kronecker(am,cm)** → δημιουργία μήτρας **jm** που περιέχει το προϊόν Kronecker (Kronecker product) των μητρών **am** και **cm**.
- **sym km=@mnrnd(4,4)** → δημιουργία συμμετρικής μήτρας **km** διαστάσεων 4×4, η οποία περιέχει τυχαίους αριθμούς κανονικής κατανομής. Σε ένα Eviews workfile τα αντικείμενα που είναι συμμετρικές μήτρες συμβολίζονται με το εικονίδιο  .
- **vector av=@getmaindiagonal(cm)** → δημιουργία διανύσματος (vector) **av** που περιέχει τα στοιχεία της κύριας διαγωνίου της μήτρας **cm**. Σε ένα Eviews workfile τα αντικείμενα που είναι vectors συμβολίζονται με το εικονίδιο  .
- **vector bv=@eigenvalues(km)** → δημιουργία διανύσματος (vector) **ab** που περιέχει τις ιδιοτιμές (eigenvalues) της συμμετρικής μήτρας **km**.
- **matrix lm=@eigenvectors(km)** → δημιουργία μήτρας **lm** που περιέχει τα ιδιοδιανύσματα (eigenvectors) της συμμετρικής μήτρας **km**.
- **scalar as=@trace(cm)** → δημιουργία βαθμωτού μεγέθους (scalar) **as** που έχει υπολογίσει το ίχνος (trace) της μήτρας **cm**. Σε ένα Eviews workfile τα αντικείμενα που είναι scalars συμβολίζονται με το εικονίδιο  .
- **scalar bs=@rank(cm)** → δημιουργία βαθμωτού μεγέθους (scalar) **bs** που έχει υπολογίσει τον βαθμό (rank) της μήτρας **cm**.
- **scalar cs=@det(cm)** → δημιουργία βαθμωτού μεγέθους (scalar) **cs** που έχει υπολογίσει την ορίζουσα (determinant) της μήτρας **cm**.

Στην **Εικόνα 2.49** παρουσιάζεται το Eviews workfile Test1.wf1, όπως έχει πλέον διαμορφωθεί, με τις παραπάνω εντολές να εμφανίζονται στο **Command line** και τα αντίστοιχα αντικείμενα να έχουν δημιουργηθεί. Όπως έχουμε ήδη δείξει παραπάνω, με διπλό κλικ σε καθένα από τα αντικείμενα αυτά μπορούμε να δούμε το περιεχόμενό του.



Εικόνα 2.49 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Βιβλιογραφία

Ξενόγλωσση

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics, 7*(1), 1-26. https://doi.org/10.1214/aos/1176344552

Κεφάλαιο 3 Περιγραφική στατιστική

Σύνοψη
Το τρίτο κεφάλαιο του βιβλίου καλύπτει την περιγραφική στατιστική, δηλαδή τα περιγραφικά μέτρα για τις συνεχείς και τις κατηγορικές μεταβλητές. Επίσης, δείχνει πώς δημιουργούνται στο SPSS και το Eviews κάποια συχνά χρησιμοποιούμενα διαγράμματα για τις μεταβλητές αυτές. Οι στόχοι του κεφαλαίου είναι ο/η αναγνώστης/-τρια να μπορεί να διακρίνει το είδος της κάθε μεταβλητής και να επιλέγει το κατάλληλο περιγραφικό μέτρο/διάγραμμα, προκειμένου να την αναλύσει.

Προαπαιτούμενη γνώση
Απαιτούνται βασικές γνώσεις στατιστικής.

3.1 Περιγραφικά μέτρα για συνεχείς μεταβλητές

Πριν μιλήσουμε για τον τρόπο εξαγωγής των περιγραφικών μέτρων στο SPSS και το Eviews, καλό θα ήταν να αλλάξουμε τους όρους «στήλη δεδομένων» για το SPSS και «στατιστική σειρά» για το Eviews σε «μεταβλητή», αφού κάθε στήλη ή σειρά αναπαριστά μία μεταβλητή που περιλαμβάνει διάφορες τιμές. Τα δεδομένα τα οποία θα χρησιμοποιήσουμε στην ανάλυσή μας αφορούν πιστωτικές κάρτες (**credit.sav**) και είναι διαθέσιμα από τη βιβλιοθήκη *AER* της *R*.[1] Περιέχουν 1319 παρατηρήσεις για τις εξής 12 μεταβλητές:

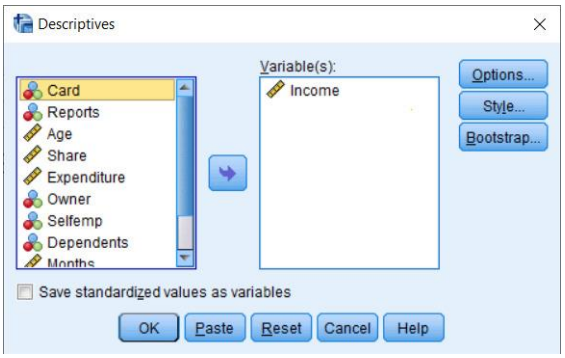
Card: *παίρνει την τιμή 1 αν η αίτηση για πιστωτική κάρτα έγινε δεκτή, και 0 αν απορρίφθηκε.*
Reports: *αριθμός κύριων υποτιμητικών αναφορών.*
Age: *έτη + 12.*
Income: *ετήσιο εισόδημα (σε δεκάδες χιλιάδες US$).*
Share: *μηνιαία έξοδα μέσω πιστωτικής κάρτας ως % του ετήσιου εισοδήματος.*
Expenditure: *μέση μηνιαία δαπάνη με τη χρήση πιστωτικής κάρτας (σε χιλιάδες US$).*
Owner: *παίρνει την τιμή 1 αν το άτομο είναι ιδιοκτήτης της τρέχουσας κατοικίας του, και 0 αν δεν είναι.*
Selfemp: *παίρνει την τιμή 1 αν το άτομο είναι αυτοαπασχολούμενο, και 0 αν δεν είναι.*
Dependents: *αριθμός εξαρτώμενων μελών +1.*
Months: *αριθμός μηνών διαβίωσης στην τρέχουσα κατοικία.*
Majorcards: *αριθμός πιστωτικών καρτών που έχει το άτομο στην κατοχή του.*
Active: *αριθμός ενεργών πιστωτικών λογαριασμών.*

Συνοπτικά, τα περιγραφικά μέτρα χωρίζονται σε μέτρα κεντρικής τάσης ή θέσης, μέτρα διασποράς και μέτρα ασυμμετρίας και κύρτωσης. Τα μέτρα θέσης δίνουν πληροφορίες για τις κεντρικές τιμές του δείγματος, όπως είναι ο μέσος, η διάμεσος, η επικρατούσα τιμή και τα εκατοστημόρια. Τα εκατοστημόρια είναι οι τιμές του δείγματος που «κόβουν» το δείγμα σε συγκεκριμένα συνήθως σημεία. Στην περίπτωση των τεταρτημορίων, για παράδειγμα, το πρώτο τεταρτημόριο είναι η τιμή του δείγματος όπου το 25% των παρατηρήσεων βρίσκεται κάτω από αυτήν. Το δεύτερο τεταρτημόριο είναι η τιμή όπου το 50% των παρατηρήσεων βρίσκεται κάτω από αυτή. Το τρίτο τεταρτημόριο είναι η τιμή όπου το 25% των παρατηρήσεων βρίσκεται πάνω από αυτή. Η διάμεσος είναι η τιμή που «κόβει» τις παρατηρήσεις του δείγματος στη μέση (και άρα ταυτίζεται με το δεύτερο τεταρτημόριο), ενώ η κορυφή είναι η παρατήρηση με τη μεγαλύτερη συχνότητα εμφάνισης.

[1] Τα δεδομένα είναι διαθέσιμα προς αποστολή στον/ην αναγνώστη/-τρια που δεν μπορεί να τα αποκτήσει μέσω της *R*.

Τα μέτρα διασποράς μας δίνουν πληροφορίες για το πώς διασπείρονται οι παρατηρήσεις γύρω από το «κέντρο» τους. Τα μέτρα αυτά είναι το εύρος, η τυπική απόκλιση, η διακύμανση, ο συντελεστής μεταβλητότητας και το ενδοτεταρτημοριακό εύρος. Ο συντελεστής μεταβλητότητας ορίζεται ως ο λόγος της τυπικής απόκλισης προς τον μέσο, πολλαπλασιασμένος επί 100. Αποτελεί ένα μέτρο ομοιογένειας του δείγματος και σχετικής διασποράς, αλλά όχι απόλυτης διασποράς. Χρησιμοποιείται και για τη σύγκριση μεταβλητών εκφρασμένων σε διαφορετικά μεγέθη. Γενικά, αποδεχόμαστε ότι ένα δείγμα είναι ομοιογενές, όταν η τιμή του συντελεστή μεταβλητότητας δεν ξεπερνάει το 10%. Τέλος, τα μέτρα ασυμμετρίας και κύρτωσης είναι ο συντελεστής ασυμμετρίας και ο συντελεστής κύρτωσης αντίστοιχα. Τα συγκεκριμένα μέτρα σχετίζονται με τη μορφή της κατανομής των δεδομένων και θα συζητηθούν παρακάτω.

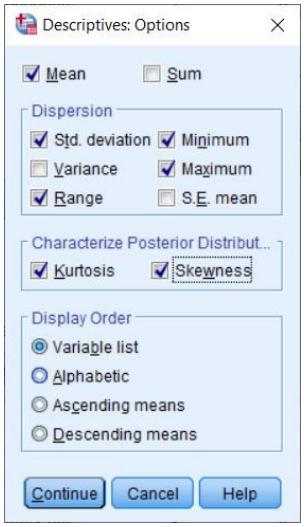
- Στο **SPSS**: Επιλέγοντας **Analyze → Descriptive Statistics → Descriptives,** θα εμφανιστεί το παράθυρο της **Εικόνας 3.1**, στο οποίο θα περάσουμε στο δεξιό μέρος τις μεταβλητές των οποίων τα περιγραφικά μέτρα θέλουμε να εμφανιστούν. Έστω, λοιπόν, ότι επιλέγουμε τη μεταβλητή “income”, η οποία αφορά το ετήσιο εισόδημα.



Εικόνα 3.1 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Πατώντας **Options** στο παράθυρο της **Εικόνας 3.1**, θα εμφανιστεί το παράθυρο της **Εικόνας 3.2**, στο οποίο μας δίνεται η δυνατότητα να επιλέξουμε ποια περιγραφικά μέτρα θέλουμε να εμφανιστούν. Το SPSS έχει ως προεπιλεγμένα μέτρα τον μέσο (Mean), την τυπική απόκλιση (Std. deviation), την ελάχιστη τιμή (Minimum) και τη μέγιστη τιμή (Maximum). Επίσης, το μενού **Display Order** μας επιτρέπει να επιλέξουμε με ποια σειρά θα εμφανιστούν τα αποτελέσματά μας.

Επιλέγουμε, λοιπόν, κάποια μέτρα και στη συνέχεια πατάμε **Continue**. Με τον τρόπο αυτό θα γυρίσουμε στο αρχικό παράθυρο της **Εικόνας 3.1**. Στο κάτω μέρος του παράθυρου αυτού υπάρχει η επιλογή **“Save standardized values as variables”**. Με την επιλογή αυτή το SPSS δημιουργεί μία νέα στήλη για κάθε μεταβλητή που έχουμε επιλέξει, η οποία στήλη θα περιέχει τις τυποποιημένες τιμές της συγκεκριμένης μεταβλητής. Οι τυποποιημένες τιμές μίας μεταβλητής είναι οι τιμές της μετασχηματισμένες με τέτοιον τρόπο που να έχουν μέση τιμή ίση με το μηδέν και διακύμανση ίση με τη μονάδα. Ο συγκεκριμένος τύπος μετασχηματισμού μιας μεταβλητής είναι ο εξής: $(x_i - \mu)/\sigma$, όπου το x_i είναι μία τιμή της μεταβλητής X, το μ είναι ο μέσος της και το σ είναι η τυπική της απόκλιση.

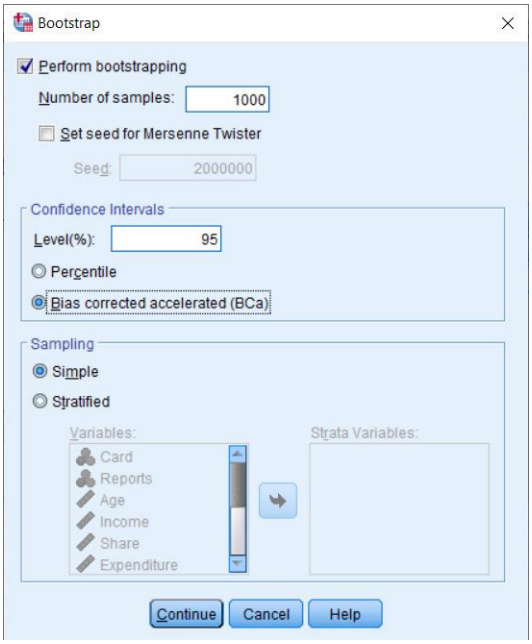


Εικόνα 3.2 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Στο παράθυρο της **Εικόνας 3.1** υπάρχει επιπλέον η επιλογή **Bootstrap**. Αν κάνουμε τη συγκεκριμένη επιλογή, θα οδηγηθούμε στο παράθυρο της **Εικόνας 3.3**. Στο παράθυρο αυτό θα κάνουμε τικ στην επιλογή **Perform bootstrapping**, ενώ λίγο παρακάτω στο **Confidence Intervals** θα αφήσουμε το 95% (εκτός και αν θέλουμε άλλο βαθμό εμπιστοσύνης, για παράδειγμα 90%). Ακριβώς από κάτω θα επιλέξουμε **Bias corrected accelerated (BCa)**, το οποίο είναι πιο ακριβές από το **Percentile**, και στη συνέχεια **Continue**, προκειμένου να επιστρέψουμε στο παράθυρο της **Εικόνας 3.1**. Εκεί θα πατήσουμε **OK** και θα προκύψουν τα αποτελέσματα, τα οποία εμφανίζονται στον **Πίνακα 3.1**.

Τα περιγραφικά μέτρα που παρουσιάζονται στον **Πίνακα 3.1** είναι με τη σειρά τα εξής: το πλήθος των στοιχείων (*N*), το εύρος (*Range*) που υπολογίζεται ως η διαφορά της μικρότερης τιμής (*Minimum*) από τη μεγαλύτερη τιμή (*Maximum*), ο μέσος (*Mean*) της μεταβλητής, καθώς και η τυπική της απόκλιση (*Std. Deviation*). Η διακύμανση είναι ο μέσος όρος των τετραγωνικών αποκλίσεων των τιμών από τη μέση τιμή, ενώ η τυπική απόκλιση προκύπτει από την τετραγωνική ρίζα της διακύμανσης.

Ο συντελεστής ασυμμετρίας (skewness) μας παρέχει πληροφορίες για την ασυμμετρία της κατανομής των δεδομένων. Τιμές κοντά στο μηδέν αποτελούν ένδειξη ότι η κατανομή των παρατηρήσεων είναι συμμετρική. Αρνητικές τιμές του συντελεστή ασυμμετρίας είναι ένδειξη ότι η κατανομή παρουσιάζει αρνητική ή αριστερή ασυμμετρία. Τέλος, η κατανομή είναι θετικά ή δεξιά ασύμμετρη, όταν ο συγκεκριμένος συντελεστής παίρνει θετικές τιμές. Όταν η κατανομή είναι θετικά ασύμμετρη, ο μέσος των παρατηρήσεων είναι μεγαλύτερος από τη διάμεσο, η οποία με τη σειρά της είναι μεγαλύτερη από την κορυφή. Το ακριβώς αντίθετο ισχύει στην περίπτωση της αρνητικής ασυμμετρίας. Δηλαδή, ο μέσος είναι μικρότερος από τη διάμεσο, η οποία είναι μικρότερη από την κορυφή. Στην περίπτωση της συμμετρικής κατανομής τα τρία αυτά μέτρα ταυτίζονται.



Εικόνα 3.3 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Ο συντελεστής κύρτωσης αναφέρεται στην κυρτότητα της κατανομής των δεδομένων. Αρνητικές τιμές του συντελεστή αυτού σημαίνουν ότι η κατανομή είναι πλατύκυρτη, ενώ οι θετικές τιμές αντιστοιχούν σε μια λεπτόκυρτη κατανομή. Τιμές κοντά στο μηδέν αποτελούν ένδειξη ότι η κατανομή είναι μεσόκυρτη. Όταν αναφερόμαστε στην κυρτότητα μιας κατανομής, αναφερόμαστε στα άκρα της κατανομής ή στις «ουρές» της, όπως αλλιώς λέγονται. Οι «παχιές» ουρές είναι ένδειξη πλατύκυρτης κατανομής, ενώ οι «λεπτές» ουρές αποτελούν ένδειξη πως η κατανομή είναι λεπτόκυρτη.

Πίνακας 3.1 *Περιγραφικά μέτρα για το εισόδημα.*

Descriptive Statistics

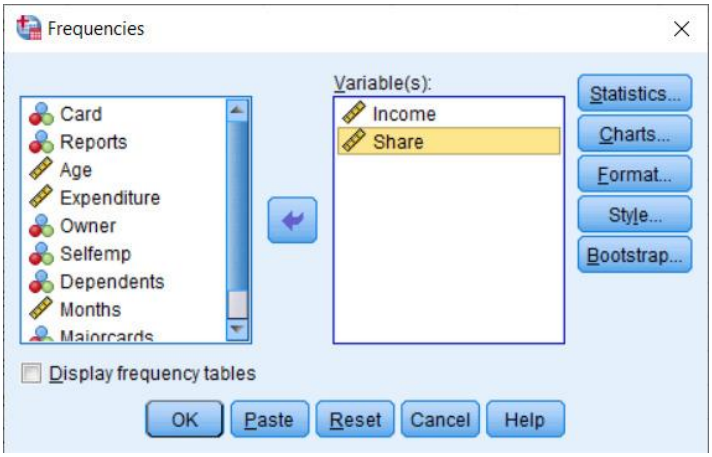
				Bootstrap[a]			
						BCa 95% Confidence Interval	
		Statistic	Std. Error	Bias	Std. Error	Lower	Upper
Income	N	1319		0	0	.	.
	Range	13.2900					
	Minimum	2100					
	Maximum	13.5000					
	Mean	3.365376		.000363	.045750	3.279866	3.454675
	Std. Deviation	1.6939017		-.0006529	.0582240	1.5733575	1.8090391
	Skewness	1.928	.067	-.011	.125	1.701	2.143
	Kurtosis	4.933	.135	-.071	.803	3.622	6.224
Valid N (listwise)	N	1319		0	0	.	.

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

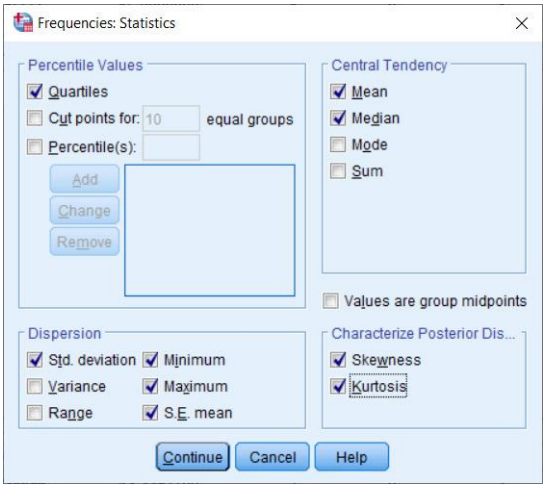
Όπως μπορούμε να δούμε στον **Πίνακα 3.1**, οι διαφορές μεταξύ των τιμών που υπολογίστηκαν από το δείγμα και των τιμών που υπολογίστηκαν από το bootstrap δεν διαφέρουν σε σημαντικό βαθμό (εκτός από τη διακύμανση). Η στήλη **Bias** περιέχει τις διαφορές αυτές, αλλά όχι και τις τιμές που υπολογίστηκαν από το bootstrap. Οι συγκεκριμένες διαφορές υπολογίζονται αφαιρώντας τις δειγματικές τιμές από τις τιμές bootstrap. Οπότε, για να υπολογίσουμε τις εκτιμήσεις από το bootstrap, απλά προσθέτουμε στις εκτιμήσεις που έχουμε, τις τιμές της στήλης **Bias**. Θα πρέπει να τονίσουμε στο σημείο αυτό ότι το bootstrap για τη

διακύμανση ή την τυπική απόκλιση καλό θα είναι να αποφεύγεται. Αυτό υποστηρίζουν οι Casella & Berger (2002) και δεν έχουμε κανένα λόγο να τους αμφισβητήσουμε. Ο λόγος που γενικά οι διαφορές είναι πολύ μικρές είναι γιατί το μέγεθος του δείγματος είναι μεγάλο (1319 παρατηρήσεις). Αν έχουμε λιγότερες παρατηρήσεις, για παράδειγμα 40 ή 50, αναμένουμε να παρατηρήσουμε μεγαλύτερες διαφορές.

Ας εξετάσουμε στη συνέχεια μία επίσης πολύ σημαντική επιλογή από το μενού επιλογών, η οποία μας δίνει τη δυνατότητα να υπολογίσουμε περισσότερα περιγραφικά μέτρα. Επιλέγουμε **Analyze → Descriptive Statistics → Frequencies**, με αποτέλεσμα να εμφανιστεί στην οθόνη μας το παράθυρο της **Εικόνας 3.4**. Στο κάτω δεξιό μέρος υπάρχει η επιλογή να εμφανιστούν οι πίνακες συχνοτήτων για τις επιλεγμένες μεταβλητές. Καθώς στην παρούσα φάση δεν τους χρειαζόμαστε, δεν κάνουμε τικ στη συγκεκριμένη επιλογή. Αν πατήσουμε **Charts**, θα εμφανιστεί ένα παράθυρο με διάφορες επιλογές γραφημάτων, τα οποία όμως θα αναλύσουμε παρακάτω. Οπότε, πατώντας **Statistics** θα εμφανιστεί το παράθυρο της **Εικόνας 3.5**.



Εικόνα 3.4 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Εικόνα 3.5 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Υπάρχουν διάφορες επιλογές εμφάνισης όλων των περιγραφικών μέτρων, τα οποία είναι επίσης διαχωρισμένα ανάλογα με το είδος τους (κεντρική τάση, διασπορά, κατανομή και ποσοστιαία σημεία). Στη συνέχεια, πατώντας **Continue** επιστρέφουμε στο παράθυρο της **Εικόνας 3.4**, όπου επιλέγοντας **Bootstrap** θα εμφανιστεί ξανά το παράθυρο της **Εικόνας 3.3**. Πατώντας **Continue** και στη συνέχεια **OK** θα εμφανιστεί ο **Πίνακας 3.2** στο Output του SPSS. Αν επιλέξουμε **Quartiles**, θα εμφανιστούν, επίσης, το πρώτο (25%), το δεύτερο (50% διάμεσος) και το τρίτο (75%) τεταρτημόριο. Θα πρέπει να υπενθυμίσουμε στο σημείο αυτό

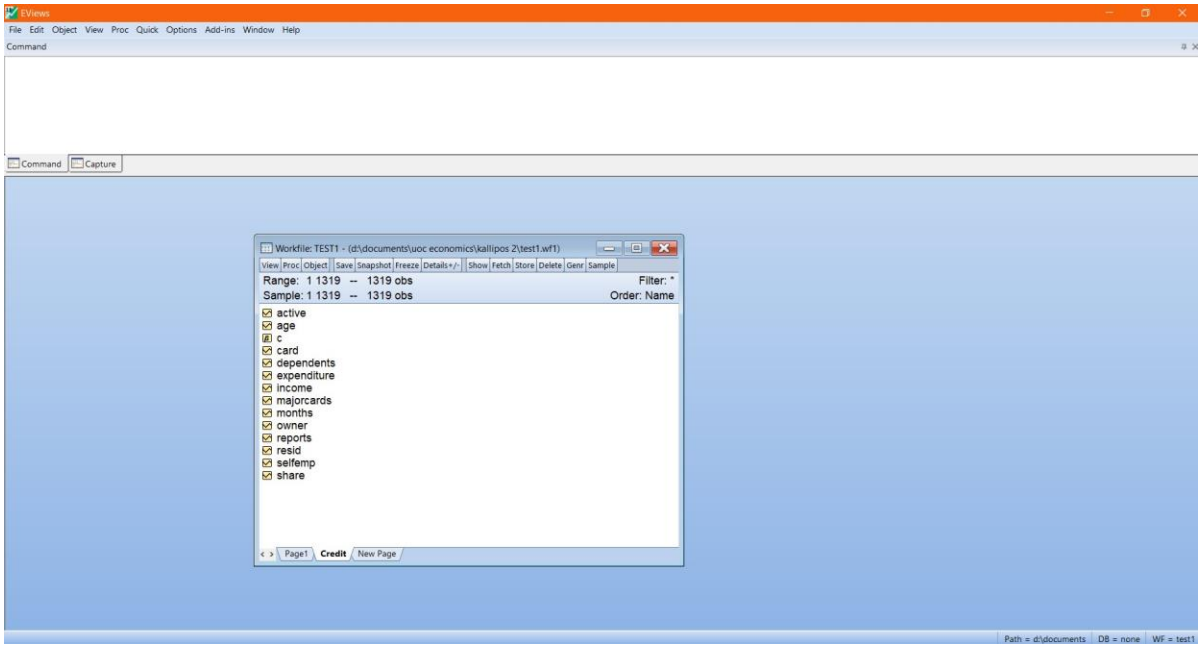
πως το τυπικό σφάλμα ή τυπική απόκλιση του μέσου ορίζεται ως ο λόγος της τυπικής απόκλισης του δείγματος με την τετραγωνική ρίζα του μεγέθους του δείγματος (*N*).

Πίνακας 3.2 *Περιγραφικά μέτρα για δύο συνεχείς μεταβλητές.*

Statistics			
		Income	Share
N	Valid	1319	1319
	Missing	0	0
Mean		3.365376	.0687321731
Std. Error of Mean		.0466408	.00260629611
Median		2.900000	.0388272200
Std. Deviation		1.6939017	.09465556526
Skewness		1.928	3.168
Std. Error of Skewness		.067	.067
Kurtosis		4.933	16.254
Std. Error of Kurtosis		.135	.135
Minimum		.2100	.00010909
Maximum		13.5000	.90632050
Percentiles	25	2.237500	.0022279990
	50	2.900000	.0388272200
	75	4.000000	.0936211100

Έχοντας πλέον αναλύσει τους δύο βασικούς τρόπους εξαγωγής κάποιων περιγραφικών μέτρων για τις συνεχείς μεταβλητές, καθώς και την τεχνική bootstrap, μπορούμε πλέον να πούμε μερικά λόγια σχετικά με τα αποτελέσματα του bootstrap. Καταρχάς, η τεχνική bootstrap δεν μας έδωσε εκτιμήσεις για όλα τα στατιστικά μέτρα. Το 95% διάστημα εμπιστοσύνης που υπολογίζεται δεν βασίζεται στον κλασικό τύπο που θα δούμε στη συνέχεια, δηλαδή στον εκτιμητή $+/-1{,}96$ φορές το τυπικό σφάλμα του. Αντίθετα, υπολογίζεται με βάση έναν άλλο τύπο που αναφέραμε, όταν εξηγούσαμε τον υπολογιστικό αλγόριθμο bootstrap. Έστω, λοιπόν, ότι η διαδικασία έχει επαναληφθεί B φορές. Οπότε, έχουμε B τιμές που τις διατάσσουμε από τη μικρότερη στη μεγαλύτερη. Στη συνέχεια, βρίσκουμε κάποιες συγκεκριμένες τιμές από αυτές τις διατεταγμένες τιμές. Αν, για παράδειγμα, θέλουμε βαθμό εμπιστοσύνης ίσο με a, θα πάρουμε την $(B+1)\times(a/2)$ μικρότερη και μεγαλύτερη τιμή. Δηλαδή, θα κόψουμε από την κατανομή των B τιμών το $a\%$ των τιμών ($a/2\,\%$ από κάτω και $a/2\,\%$ από πάνω). Έτσι, αν, για παράδειγμα, $B = 1.000$, θα πρέπει να υπολογίσουμε την 25η και 976η τιμή κατά αύξουσα σειρά, σύμφωνα με τη μέθοδο Percentile. Ωστόσο, η μέθοδος BCa διορθώνει τα προβλήματα μεροληψίας αυτής της μεθόδου και για τον λόγο αυτό είναι προτιμότερη.

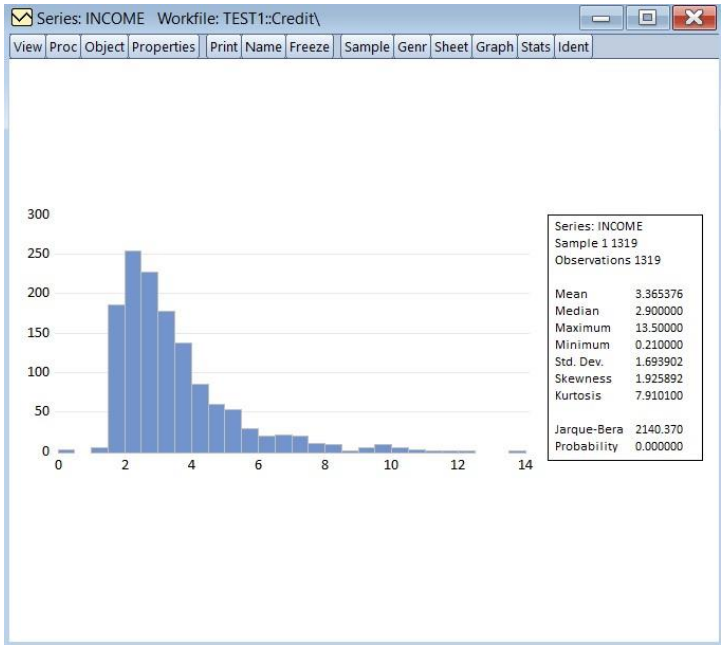
- Στο **Eviews**: Προκειμένου να εισάγουμε τις συγκεκριμένες μεταβλητές στο αρχείο **Test1** που έχουμε ήδη δημιουργήσει **(Εικόνα 2.49)**, επιλέγουμε **New Page**, το οποίο βρίσκεται κάτω αριστερά στο Eviews workfile, και στη συνέχεια **Specify by Frequency/Range**. Στο παράθυρο που εμφανίζεται επιλέγουμε **Unstructured/Undated (Εικόνα 2.28)**, καθώς οι μεταβλητές μας δεν είναι ούτε χρονολογικές σειρές ούτε έχουν τη μορφή panel, και στη συνέχεια εισάγουμε τον αριθμό των **observations** που είναι 1319. Προαιρετικά, μπορούμε να δώσουμε κάποιο όνομα στο νέο page (έστω **Credit** στο παράδειγμά μας). Πατώντας **OK**, θα εμφανιστεί το συγκεκριμένο page με το όνομα **Credit**, ενώ, για να εισάγουμε τις μεταβλητές μας, ακολουθούμε ακριβώς τη διαδικασία που περιγράψαμε στην ενότητα 2.9. Το page **Credit** εμφανίζεται πλέον ολοκληρωμένο στην **Εικόνα 3.6**.



Εικόνα 3.6 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

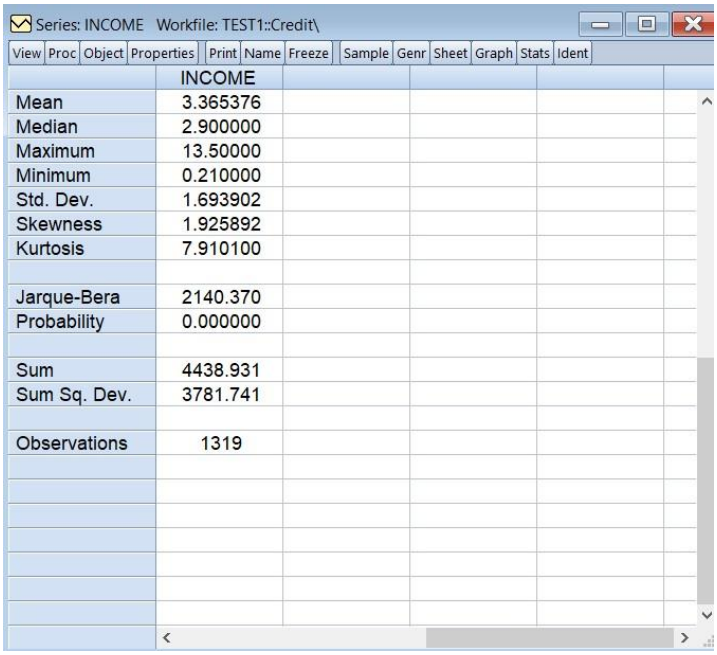
Για να υπολογίσουμε τα περιγραφικά στατιστικά μέτρα για τη μεταβλητή income (ετήσιο εισόδημα), κάνουμε διπλό κλικ στη συγκεκριμένη μεταβλητή έτσι ώστε να εμφανιστεί στην οθόνη μας. Στη συνέχεια, επιλέγουμε **View** (που είναι το πρώτο κουμπί στο toolbar) → **Descriptive Statistics & Tests**. Πλέον, έχουμε τρεις επιλογές:

- **Histogram and Stats:** Εμφανίζεται το ιστόγραμμα της συγκεκριμένης μεταβλητής (το οποίο θα αναλυθεί στην ενότητα 3.3), καθώς και τα βασικά περιγραφικά μέτρα της, όπως ο μέσος (*Mean*), η διάμεση τιμή (*Median*), η μεγαλύτερη τιμή (*Maximum*), η μικρότερη τιμή (*Minimum*), η τυπική απόκλιση (*Std. Dev.*), ο συντελεστής ασυμμετρίας (*Skewness*) και ο συντελεστής κύρτωσης (*Kurtosis*). Επίσης, εμφανίζεται η τιμή της στατιστικής Jarque-Bera με την αντίστοιχη πιθανότητά της. Η συγκεκριμένη στατιστική ελέγχει το κατά πόσο η συγκεκριμένη μεταβλητή ακολουθεί μια κανονική κατανομή και θα αναλυθεί στην ενότητα 4.1 **(Εικόνα 3.7)**.



Εικόνα 3.7 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

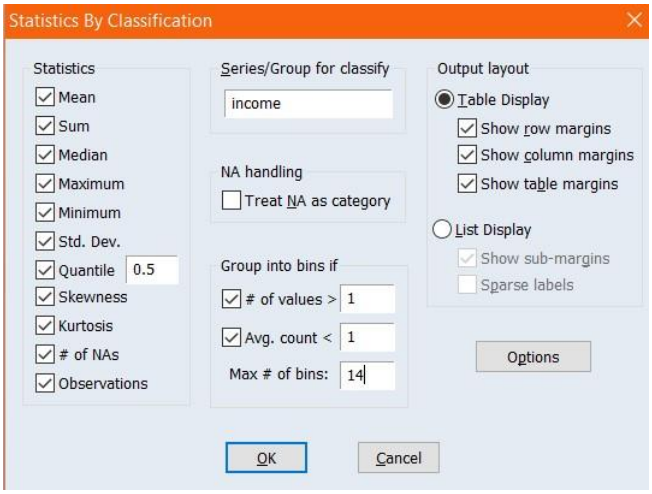
- **Stats Table:** Εμφανίζονται όλα τα παραπάνω βασικά περιγραφικά μέτρα, το άθροισμα των τιμών της μεταβλητής (Sum), καθώς και το άθροισμα των τετραγωνικών αποκλίσεων (Sum Sq. Dev.) των τιμών της από τον μέσο **(Εικόνα 3.8)**.



Εικόνα 3.8 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

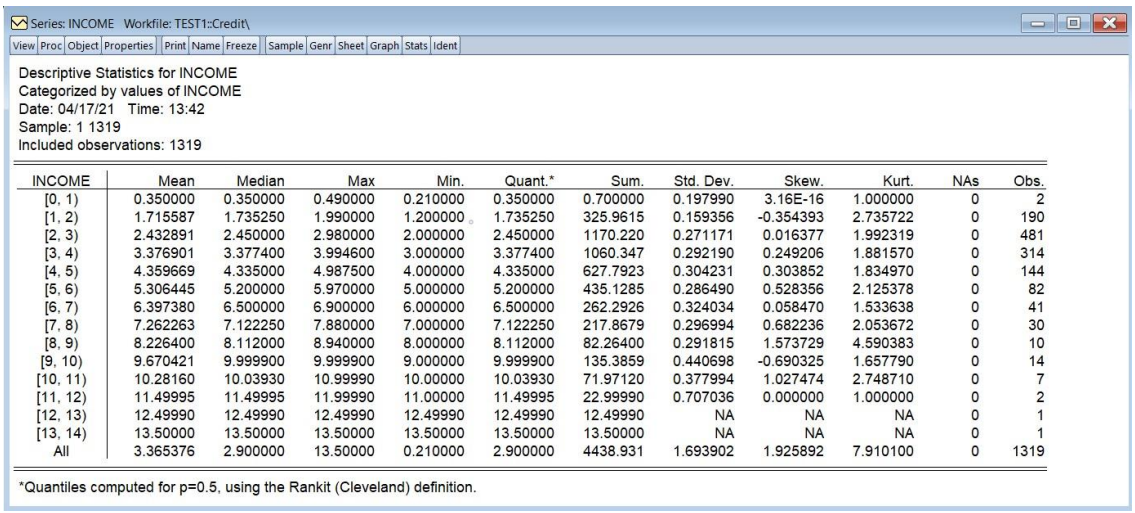
- **Stats by Classification:** Η συγκεκριμένη επιλογή μας επιτρέπει (α) να χωρίσουμε σε κατηγορίες τα δεδομένα μας, προκειμένου να μελετήσουμε καλύτερα την κατανομή τους, και (β) να επιλέξουμε ποια περιγραφικά μέτρα θέλουμε. Έστω, για παράδειγμα, ότι θέλουμε να υπολογίσουμε όλα τα περιγραφικά μέτρα για κάθε δέκα χιλιάδες δολάρια ετήσιο εισόδημα. Στο παράθυρο που θα εμφανιστεί **(Εικόνα 3.9)** επιλέγουμε όλα τα στατιστικά μέτρα στο αριστερό μέρος (τα NAs αντιστοιχούν σε μη διαθέσιμες παρατηρήσεις) και συμπληρώνουμε income στο κελί Series/Group for classify, καθώς αυτή είναι η μεταβλητή που μας ενδιαφέρει. Επίσης, «τσεκάρουμε» τα 2 κουτάκια “# of values >” και “Avg. Count <” και συμπληρώνουμε 1

και στα δύο αντίστοιχα κελιά, καθώς μας ενδιαφέρει να υπάρχει έστω μία παρατήρηση σε κάθε κατηγορία εισοδήματος. Τέλος, καθώς το ελάχιστο εισόδημα είναι 0.21 και το μέγιστο είναι 13.5 **(Εικόνα 3.8)** προκύπτουν 14 κατηγορίες εισοδήματος και άρα συμπληρώνουμε 14 στο κελί "Max # of bins:". Πατώντας **OK** εμφανίζονται όλα τα περιγραφικά μέτρα που επιλέξαμε **(Εικόνα 3.10)**.



Εικόνα 3.9 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Επιπλέον, έχοντας ανοικτή τη μεταβλητή "income" και επιλέγοντας από το toolbar **View → One-Way Tabulation,** μπορούμε να πάρουμε επιπλέον πληροφορίες για την κατανομή της συγκεκριμένης μεταβλητής, όπως είναι ο αριθμός των παρατηρήσεων σε κάθε κατηγορία (**Count**), το ποσοστό των παρατηρήσεων κάθε κατηγορίας στο σύνολο των παρατηρήσεων (**Percent**), ο σωρευτικός αριθμός των παρατηρήσεων (**Cumulative Count**) και το σωρευτικό ποσοστό των παρατηρήσεων (**Cumulative Percent**). Όπως και πριν, αν θέλουμε να υπολογίσουμε τα παραπάνω για κάθε δέκα χιλιάδες δολάρια ετήσιο εισόδημα, στο παράθυρο που θα εμφανιστεί **(Εικόνα 3.11)** επιλέγουμε όλα τα output στο αριστερό μέρος, «τσεκάρουμε» τα 2 κουτάκια "# of values >" και "Avg. Count <" και συμπληρώνουμε 1 και στα δύο αντίστοιχα κελιά. Επίσης, συμπληρώνουμε 14 στο κελί "Max # of bins:". Πατώντας **OK** εμφανίζονται οι συγκεκριμένες πληροφορίες **(Εικόνα 3.12)**. Αν αποεπιλέξουμε τα 2 κουτάκια "# of values >" και "Avg. Count <", θα πάρουμε μια πολύ πιο αναλυτική πληροφόρηση σχετικά με την κατανομή της συγκεκριμένης μεταβλητής.


Series: INCOME Workfile: TEST1::Credit\

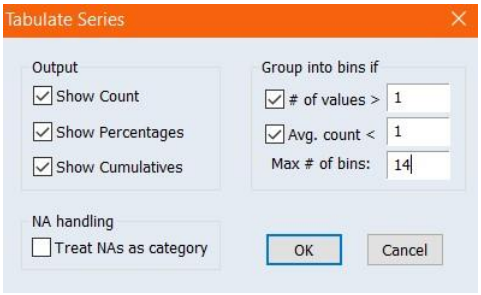
View | Proc | Object | Properties | Print | Name | Freeze | Sample | Genr | Sheet | Graph | Stats | Ident

Descriptive Statistics for INCOME
Categorized by values of INCOME
Date: 04/17/21 Time: 13:42
Sample: 1 1319
Included observations: 1319

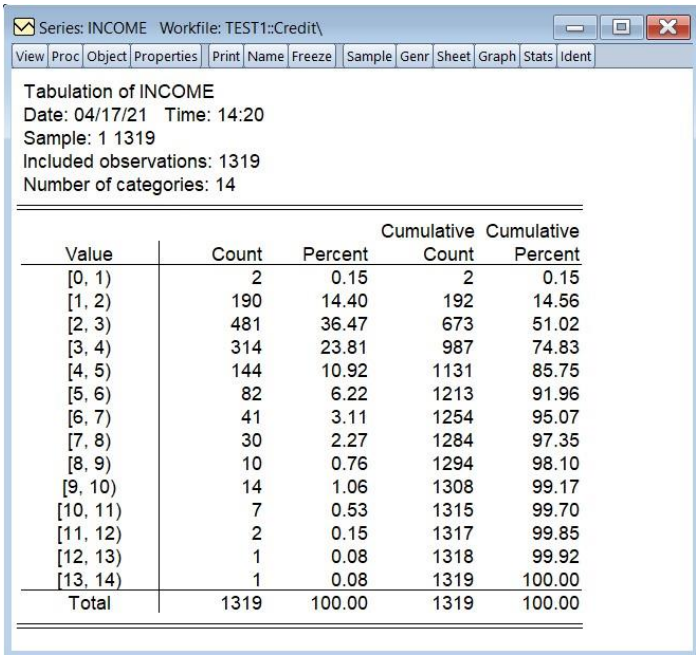
INCOME	Mean	Median	Max	Min.	Quant.*	Sum.	Std. Dev.	Skew.	Kurt.	NAs	Obs.
[0, 1)	0.350000	0.350000	0.490000	0.210000	0.350000	0.700000	0.197990	3.16E-16	1.000000	0	2
[1, 2)	1.715587	1.735250	1.990000	1.200000	1.735250	325.9615	0.159356	-0.354393	2.735722	0	190
[2, 3)	2.432891	2.450000	2.980000	2.000000	2.450000	1170.220	0.271171	0.016377	1.992319	0	481
[3, 4)	3.376901	3.377400	3.994600	3.000000	3.377400	1060.347	0.292190	0.249206	1.881570	0	314
[4, 5)	4.359669	4.335000	4.987500	4.000000	4.335000	627.7923	0.304231	0.303852	1.834970	0	144
[5, 6)	5.306445	5.200000	5.970000	5.000000	5.200000	435.1285	0.286490	0.528356	2.125378	0	82
[6, 7)	6.397380	6.500000	6.900000	6.000000	6.500000	262.2926	0.324034	0.058470	1.533638	0	41
[7, 8)	7.262263	7.122250	7.880000	7.000000	7.122250	217.8679	0.296994	0.682236	2.053672	0	30
[8, 9)	8.226400	8.112000	8.940000	8.000000	8.112000	82.26400	0.291815	1.573729	4.590383	0	10
[9, 10)	9.670421	9.999900	9.999900	9.000000	9.999900	135.3859	0.440698	-0.690325	1.657790	0	14
[10, 11)	10.28160	10.03930	10.99990	10.00000	10.03930	71.97120	0.377994	1.027474	2.748710	0	7
[11, 12)	11.49995	11.49995	11.99990	11.00000	11.49995	22.99990	0.707036	0.000000	1.000000	0	2
[12, 13)	12.49990	12.49990	12.49990	12.49990	12.49990	12.49990	NA	NA	NA	0	1
[13, 14)	13.50000	13.50000	13.50000	13.50000	13.50000	13.50000	NA	NA	NA	0	1
All	3.365376	2.900000	13.50000	0.210000	2.900000	4438.931	1.693902	1.925892	7.910100	0	1319

*Quantiles computed for p=0.5, using the Rankit (Cleveland) definition.

Εικόνα 3.10 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Εικόνα 3.11 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Value	Count	Percent	Cumulative Count	Cumulative Percent
[0, 1)	2	0.15	2	0.15
[1, 2)	190	14.40	192	14.56
[2, 3)	481	36.47	673	51.02
[3, 4)	314	23.81	987	74.83
[4, 5)	144	10.92	1131	85.75
[5, 6)	82	6.22	1213	91.96
[6, 7)	41	3.11	1254	95.07
[7, 8)	30	2.27	1284	97.35
[8, 9)	10	0.76	1294	98.10
[9, 10)	14	1.06	1308	99.17
[10, 11)	7	0.53	1315	99.70
[11, 12)	2	0.15	1317	99.85
[12, 13)	1	0.08	1318	99.92
[13, 14)	1	0.08	1319	100.00
Total	1319	100.00	1319	100.00

Εικόνα 3.12 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

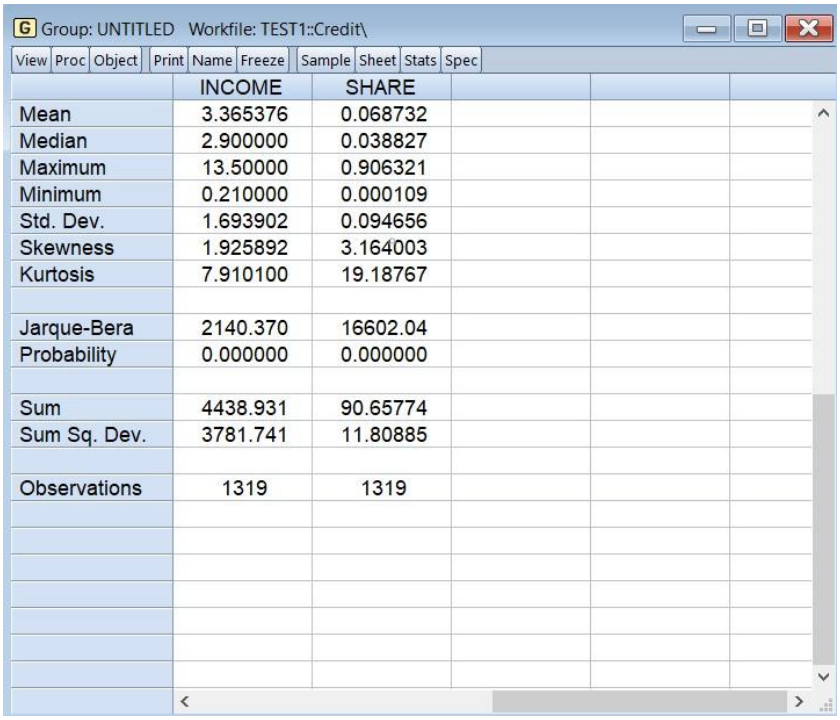
Τα αποτελέσματα που εμφανίζονται στις **Εικόνες 3.7, 3.8, 3.10** και **3.12** μπορούν να αποθηκευτούν στο Eviews workfile είτε ως γράφημα (στην περίπτωση της **Εικόνας 3.7**) είτε ως πίνακας στις άλλες περιπτώσεις. Προκειμένου να συμβεί αυτό, στο παράθυρο που εμφανίζονται τα αποτελέσματα επιλέγουμε **Freeze** από το toolbar. Στο νέο παράθυρο, που αναδύεται, επιλέγουμε **Name** από το toolbar, δίνουμε το όνομα που επιθυμούμε στο συγκεκριμένο αντικείμενο και αυτό πλέον εμφανίζεται στο Eviews workfile. Σε ένα Eviews workfile τα γραφήματα συμβολίζονται με το εικονίδιο και οι πίνακες με το εικονίδιο .

Όπως συμβαίνει στο SPSS, έτσι και στο Eviews μπορούμε να πάρουμε τις τυποποιημένες τιμές μιας μεταβλητής, δηλαδή τις ίδιες τιμές μετασχηματισμένες με τέτοιον τρόπο έτσι ώστε να έχουν μέση τιμή ίση με το μηδέν και διακύμανση ίση με τη μονάδα. Το Eviews παρέχει 2 εναλλακτικές: (α) η διακύμανση του δείγματος να είναι ίση με τη μονάδα, όπου η γενική μορφή της συνάρτησης είναι **@stdize(x, "sample")** και (β) η διακύμανση του πληθυσμού να είναι ίση με τη μονάδα, όπου η γενική μορφή της συνάρτησης είναι **@stdizep(x, "sample")**. Οπότε, αν στο page Credit θέλουμε να δημιουργήσουμε τις δύο αυτές νέες μεταβλητές, οι οποίες θα αφορούν τις τυποποιημένες τιμές της μεταβλητής "income" και θα ονομάζονται "incomest" και "incomestp" αντίστοιχα, θα γράψουμε στο **Command line**:

genr incomest=@stdize(income,"1 1319") και θα πατήσουμε **Enter**, και
genr incomestp=@stdizep(income,"1 1319") και θα πατήσουμε **Enter**.

Οι δύο νέες μεταβλητές έχουν πλέον δημιουργηθεί.

Με τον ίδιο τρόπο που υπολογίσαμε τα βασικά περιγραφικά μέτρα για τη μεταβλητή “income”, μπορούμε να τα υπολογίσουμε για δύο ή περισσότερες μεταβλητές ταυτόχρονα. Έστω, για παράδειγμα, ότι επιλέγουμε τις μεταβλητές “income” και “share” και με διπλό κλικ τις κάνουμε να εμφανιστούν στην οθόνη μας. Επιλέγουμε **View → Descriptive Stats → Common Sample** ή **Individual Samples**. Στο σημείο αυτό χρειάζεται προσοχή, καθώς, αν επιλέξουμε Common Sample, το Eviews θα υπολογίσει τα βασικά περιγραφικά μέτρα για όλες τις μεταβλητές χρησιμοποιώντας κοινό δείγμα. Με άλλα λόγια, αν έχουμε δύο μεταβλητές εκ των οποίων η μία έχει 500 παρατηρήσεις και η άλλη έχει 600 παρατηρήσεις, το Eviews θα υπολογίσει τα περιγραφικά μέτρα μόνο για το μέρος του δείγματος όπου υπάρχουν παρατηρήσεις και για τις δύο μεταβλητές, δηλαδή για δείγμα ίσο με 500. Αντιθέτως, αν επιλέξουμε Individual Samples, το Eviews θα υπολογίσει τα βασικά περιγραφικά μέτρα χρησιμοποιώντας ξεχωριστά το δείγμα κάθε μεταβλητής, δηλαδή 500 για την πρώτη μεταβλητή και 600 για τη δεύτερη μεταβλητή. Προφανώς, αν οι μεταβλητές έχουν τον ίδιο αριθμό παρατηρήσεων, όπως στο παράδειγμά μας όπου οι μεταβλητές “income” και “share” έχουν 1319 παρατηρήσεις η καθεμία, τότε είτε επιλέξουμε Common Sample είτε επιλέξουμε Individual Samples θα πάρουμε ακριβώς τα ίδια αποτελέσματα **(Εικόνα 3.13)**. Η επιλογή **View → N-Way Tabulation** δεν θα εξεταστεί στο σημείο αυτό, καθώς περιλαμβάνει μια σειρά στατιστικών ελέγχων που δεν έχουν ακόμα αναλυθεί.


Group: UNTITLED Workfile: TEST1::Credit\

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

	INCOME	SHARE
Mean	3.365376	0.068732
Median	2.900000	0.038827
Maximum	13.50000	0.906321
Minimum	0.210000	0.000109
Std. Dev.	1.693902	0.094656
Skewness	1.925892	3.164003
Kurtosis	7.910100	19.18767
Jarque-Bera	2140.370	16602.04
Probability	0.000000	0.000000
Sum	4438.931	90.65774
Sum Sq. Dev.	3781.741	11.80885
Observations	1319	1319

Εικόνα 3.13 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

3.2 Περιγραφικά μέτρα για κατηγορικές μεταβλητές

Στην προηγούμενη ενότητα αναλύσαμε τα περιγραφικά μέτρα για τις συνεχείς μεταβλητές. Στην παρούσα ενότητα θα αναλυθεί το τι γίνεται στην περίπτωση που οι μεταβλητές είναι κατηγορικές, όπως, για παράδειγμα, η ομάδα αίματος, το φύλο, ο βαθμός ικανοποίησης ή αρεσκείας, η χώρα που κατοικεί ο πληθυσμός, η ομάδα που υποστηρίζει σε κάποιο ομαδικό άθλημα ή η χώρα προέλευσης των αυτοκινήτων που κυκλοφορούν. Προφανώς, αυτές οι μεταβλητές δεν παίρνουν αριθμητικές τιμές 1, 2 κλπ. Θα πρέπει να τονίσουμε, επίσης, ότι, αν υπάρχει κάποια διάταξη στις μεταβλητές αυτές, τότε λέμε ότι είναι διατακτικές μεταβλητές ή κατηγορικές διατακτικής κλίμακας. Ένα παράδειγμα που φανερώνει διάταξη είναι ο βαθμός ικανοποίησης ενός ατόμου από τη δουλειά του: «*καθόλου*», «*μέτρια*», «*λίγο*» και «*πολύ*». Αν, όμως, δεν υπάρχει διάταξη, όπως στην περίπτωση του φύλου (άνδρας ή γυναίκα) ή της ομάδας αίματος, τότε οι μεταβλητές αυτές καλούνται ονομαστικές μεταβλητές ή κατηγορικές ονομαστικής κλίμακας.

- Στο **SPSS**: Στην περίπτωση των κατηγορικών μεταβλητών, ξεκινάμε και πάλι από παράθυρο της **Εικόνας 3.4**. Ενώ την επιλογή **Display frequency tables**, που βρίσκεται στο κάτω αριστερό μέρος, την είχαμε αποεπιλέξει στην περίπτωση των συνεχών μεταβλητών, θα την «τσεκάρουμε» στην περίπτωση των κατηγορικών μεταβλητών, γιατί αυτό κυρίως μας ενδιαφέρει. Όμως, στην επιλογή **Statistics** και **Bootstrap** (δεν χρειάζεται bootstrap στη συγκεκριμένη περίπτωση) δεν θα επιλέξουμε τίποτα. Πατώντας **OK** προκύπτουν τα αποτελέσματα που εμφανίζονται στον **Πίνακα 3.3**.

Πίνακας 3.3 *Περιγραφικά μέτρα για μία κατηγορική μεταβλητή.*

Card					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Not accepted	296	22.4	22.4	22.4
	Accepted	1023	77.6	77.6	100.0
	Total	1319	100.0	100.0	

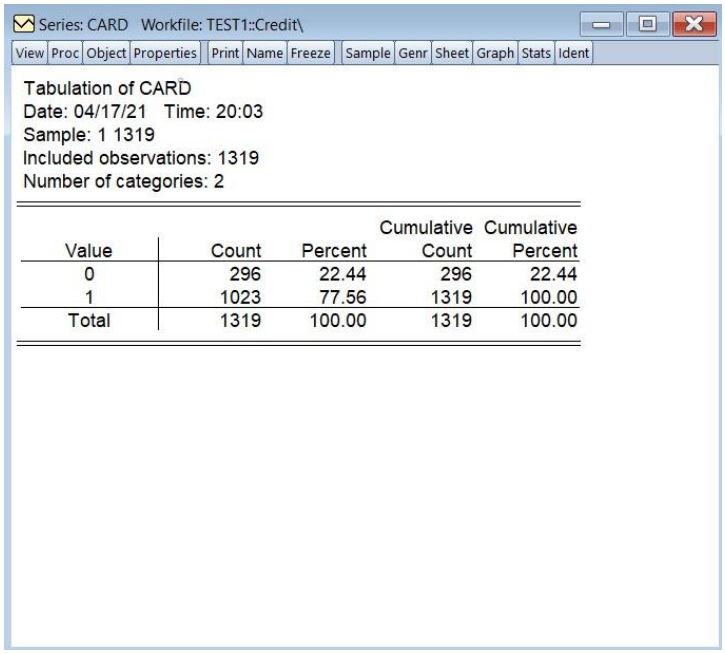
Ο **Πίνακας 3.3** περιέχει τη συχνότητα εμφάνισης κάθε απάντησης στην αίτηση για πιστωτική κάρτα, δηλαδή αν η αίτηση έγινε δεκτή ή απορρίφθηκε, τόσο σε απόλυτους αριθμούς όσο και σε σχετικούς. Παρατηρούμε ότι η πιο συχνή απάντηση είναι η αποδοχή της αίτησης. Η δεύτερη στήλη περιέχει τα ποσοστά των τιμών αυτών. Τα ποσοστά αυτά έχουν υπολογιστεί διαιρώντας κάθε αριθμό με το 1319 και στη συνέχεια πολλαπλασιάζοντας με το 100. Η τρίτη στήλη (**Valid Percent**) περιέχει ξανά τα αποτελέσματα της διαίρεσης με τον αριθμό 1319. Τα αποτελέσματα της δεύτερης και της τρίτης στήλης είναι ακριβώς τα ίδια, καθώς δεν υπάρχουν ελλείπουσες τιμές (*missing values*). Αν υπήρχαν ελλείπουσες τιμές, τότε τα νούμερα της τρίτης στήλης θα ήταν μεγαλύτερα από αυτά της δεύτερης, καθώς η διαίρεση θα γινόταν με τον αριθμό των διαθέσιμων πλήρων παρατηρήσεων, ο οποίος θα ήταν προφανώς μικρότερος από 1319. Τέλος, η τέταρτη στήλη παρουσιάζει τα αθροιστικά ποσοστά.

- Στο **Eviews**: Προκειμένου να πάρουμε τις αντίστοιχες πληροφορίες με αυτές του **Πίνακα 3.3** για τη μεταβλητή "Card", κάνουμε αρχικά διπλό κλικ στη συγκεκριμένη μεταβλητή, προκειμένου να εμφανιστεί στην οθόνη μας. Στη συνέχεια, επιλέγουμε **View → One-Way Tabulation** και στο παράθυρο που θα εμφανιστεί **(Εικόνα 3.11)** επιλέγουμε όλα τα output στο αριστερό μέρος, «τσεκάρουμε» τα κουτάκια "# of values >" και "Avg. Count <" και αφήνουμε τις προσυμπληρωμένες επιλογές, ενώ συμπληρώνουμε 2 στο κελί "Max # of bins:". Πατώντας **OK** εμφανίζονται οι συγκεκριμένες πληροφορίες **(Εικόνα 3.14)**, όπου στην πρώτη στήλη (**Value**) το 0 αντιστοιχεί στην απάντηση "Not accepted" και το 1 στην απάντηση "Accepted". Η πρώτη στήλη περιέχει τη συχνότητα εμφάνισης καθεμίας από τις δύο απαντήσεις στην αίτηση για πιστωτική κάρτα, η δεύτερη στήλη παρουσιάζει τα αντίστοιχα ποσοστά, η τρίτη στήλη εμφανίζει τις σωρευτικές συχνότητες, ενώ η τέταρτη στήλη δείχνει τα σωρευτικά ποσοστά.

3.3 Ιστογράμματα

Πριν αναλύσουμε τον τρόπο κατασκευής ιστογραμμάτων στο SPSS και το Eviews, είναι χρήσιμο να αναφερθούν κάποια ζητήματα σχετικά με τα ιστογράμματα. Έστω, λοιπόν, ότι έχουμε τιμές από μία ποσοτική μεταβλητή. Αν το πλήθος των τιμών αυτών είναι πολύ μεγάλο, τότε μπορούμε να τις απεικονίσουμε διαγραμματικά χρησιμοποιώντας ένα ιστόγραμμα συχνοτήτων. Στο συγκεκριμένο ιστόγραμμα, ο οριζόντιος άξονας θα περιλαμβάνει τις κλάσεις των τιμών (ή αλλιώς τις ομάδες των τιμών, τις οποίες έχουμε κατηγοριοποιήσει), ενώ ο κάθετος άξονας θα περιλαμβάνει τις συχνότητες εμφάνισης των ομαδοποιημένων τιμών. Με τον τρόπο αυτό σχηματίζονται ορθογώνια, το μήκος των οποίων είναι ίσο με το εύρος των τιμών που έχουν συμπεριληφθεί σε κάθε ιστόγραμμα. Τα ορθογώνια είναι «κολλημένα» το ένα στο άλλο. Αν ενώσουμε το μέσο της πάνω πλευράς όλων των ορθογωνίων με μία γραμμή, θα

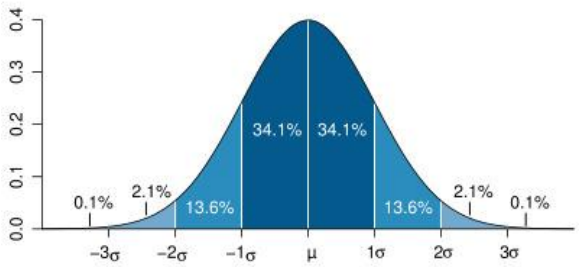
καταλήξουμε στο πολύγωνο συχνοτήτων. Όταν ο αριθμός των κλάσεων τείνει στο άπειρο, η πολυγωνική γραμμή γίνεται με τη σειρά της ομαλή και καταλήγει σε μια γραμμή που ονομάζεται καμπύλη συχνοτήτων.



Value	Count	Percent	Cumulative Count	Cumulative Percent
0	296	22.44	296	22.44
1	1023	77.56	1319	100.00
Total	1319	100.00	1319	100.00

Εικόνα 3.14 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Η πιο γνωστή κατανομή, αλλά και η πιο εύχρηστη είναι η κανονική κατανομή. Στην κατανομή αυτή, η καμπύλη συχνοτήτων των δεδομένων σχηματίζει μία «καμπάνα». Η κατανομή αυτή εξετάστηκε σε πολύ σημαντικό βαθμό από τον Γερμανό μαθηματικό Carl Friedrich Gauss, γι' αυτό και μερικές φορές συναντάται με το όνομα κατανομή Gaussian ή Γκαουσιανή κατανομή. Η συγκεκριμένη κατανομή έχει τη μορφή που παρουσιάζεται στο **Διάγραμμα 3.1**.



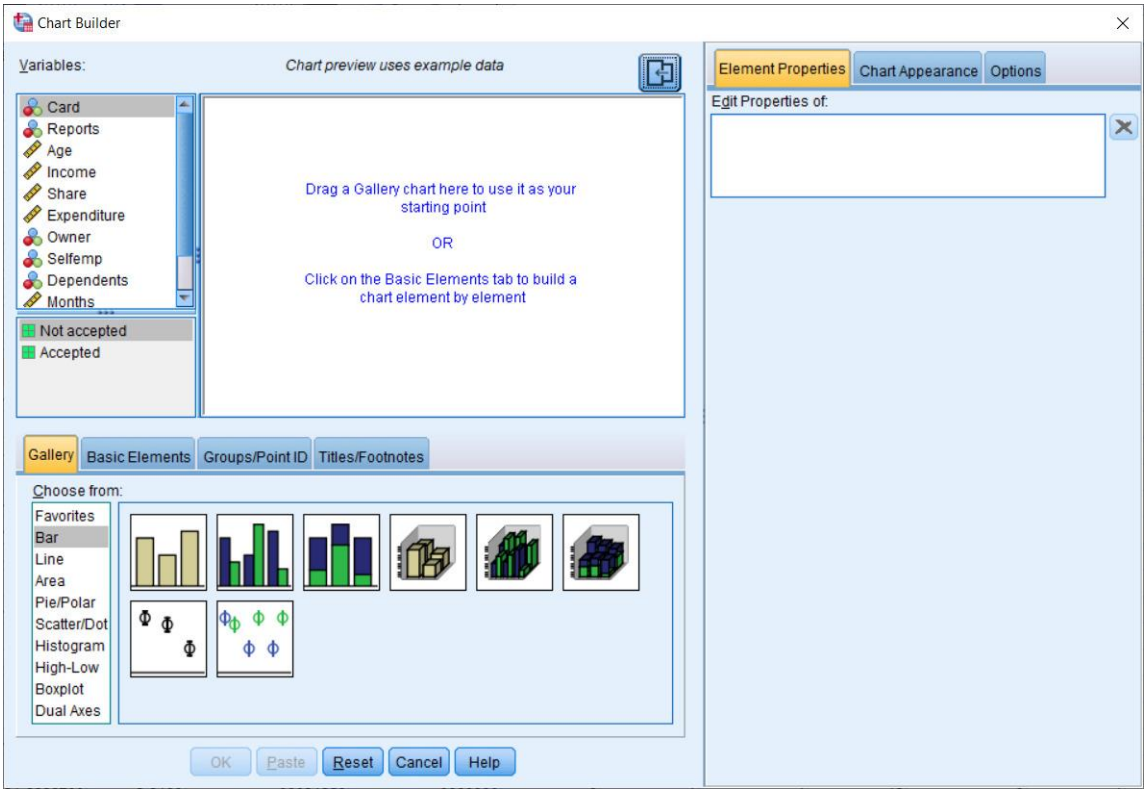
Διάγραμμα 3.1 *Η κανονική κατανομή.*

Η κανονική κατανομή είναι συμμετρική και μεσόκυρτη κατανομή και συνεπώς, η διάμεσος, η επικρατούσα τιμή και η μέση τιμή της ταυτίζονται. Επίσης, μία άλλη χρήσιμη ιδιότητα της κανονικής κατανομής, η οποία ισχύει και για άλλες μη κανονικές συμμετρικές κατανομές, είναι ότι περίπου το 68% των παρατηρήσεων βρίσκεται στο διάστημα $(\mu - \sigma, \mu + \sigma)$, περίπου το 95% των παρατηρήσεων βρίσκεται στο διάστημα $(\mu - 2\sigma, \mu + 2\sigma)$ και περίπου το 99.7% των παρατηρήσεων βρίσκεται στο διάστημα $(\mu - 3\sigma, \mu + 3\sigma)$. Με μ συμβολίζεται ο μέσος και με σ η τυπική απόκλιση της κατανομής.

Αν, όμως, έχουμε μία ποιοτική μεταβλητή ή μία ποσοτική μεταβλητή με μικρό εύρος διακριτών τιμών ή με λίγες κλάσεις ομαδοποιημένων τιμών, τότε μπορούμε να χρησιμοποιήσουμε το κυκλικό διάγραμμα (ή διάγραμμα πίτας). Το κυκλικό διάγραμμα χρησιμοποιείται όταν έχουμε ποιοτικές μεταβλητές, προκειμένου να απεικονίσουμε τις συχνότητες εμφάνισης των κατηγοριών ή το ποσοστό εμφάνισης που αντιστοιχεί σε κάθε κατηγορία μίας ποιοτικής μεταβλητής. Η κατασκευή του συγκεκριμένου διαγράμματος είναι απλή. Διαιρούμε τη συχνότητα εμφάνισης μίας κατηγορίας της ποιοτικής μεταβλητής με το άθροισμα των συχνοτήτων όλων των κατηγοριών της μεταβλητής αυτής και στη συνέχεια πολλαπλασιάζουμε με το 360^{o}.

Με τον τρόπο αυτό καθορίζουμε τις μοίρες της κάθε «φέτας» στο διάγραμμα, η οποία αντιστοιχεί σε κάθε κατηγορία. Αν, για παράδειγμα, μία κατηγορία μίας ποιοτικής μεταβλητής εμφανίζεται σε ποσοστό 50%, το κομμάτι της «πίτας» που «ανήκει» σε αυτήν την κατηγορία είναι ίσο με $50\% \times 360^{o} = 180^{o}$.

- Στο **SPSS**: Η επιλογή **Graphs** έχει δύο υποεπιλογές, οι οποίες μας επιτρέπουν να κατασκευάσουμε ένα ιστόγραμμα συχνοτήτων. Η πρώτη υποεπιλογή έχει ως εξής. Αρχικά, επιλέγουμε **Graphs → Chart Builder**, οπότε θα εμφανιστεί το παράθυρο της **Εικόνας 3.15**.



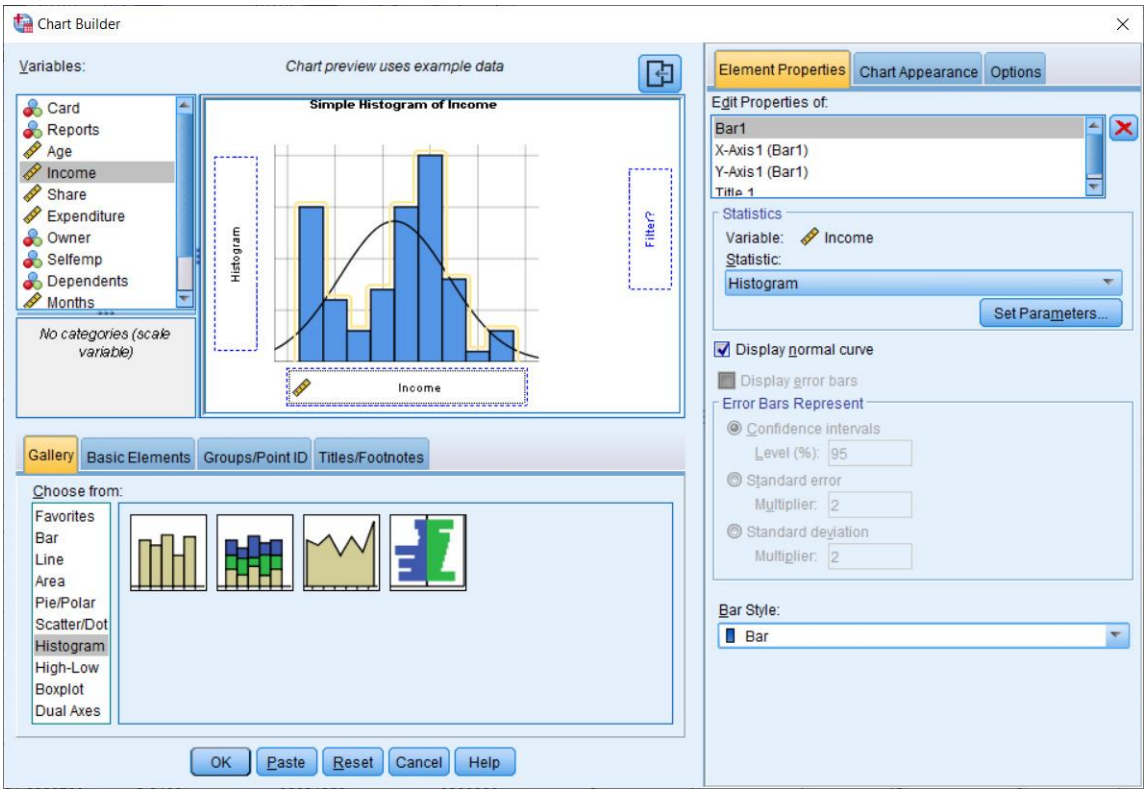
Εικόνα 3.15 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Οι μεταβλητές που φαίνονται στο αριστερό κουτάκι αφορούν τα δεδομένα μας. Αρχικά, πρέπει να επιλέξουμε τον τύπο γραφήματος που θέλουμε, μέσω της επιλογής **Gallery**. Θα επιλέξουμε **Histogram** και στη συνέχεια την πρώτη επιλογή δεξιά στο λευκό κουτάκι (όπως αναφέρεται και στις οδηγίες με τα μπλε γράμματα). Στη συνέχεια, επιλέγουμε τη μεταβλητή της οποίας το ιστόγραμμα θέλουμε να κατασκευάσουμε, και την σύρουμε με το ποντίκι στον **X-axis?** (οριζόντιος άξονας). Επιλέγουμε **Display normal curve** στη δεξιά πλευρά του παραθύρου αυτού **(Εικόνα 3.16)** και μετά **Apply**. Στην επιλογή **Titles/Footnotes** μπορούμε να δώσουμε τίτλο/υπότιτλο/υποσημείωση στο διάγραμμά μας. Πατώντας **OK** θα εμφανιστεί το ιστόγραμμα στο **Διάγραμμα 3.2**.

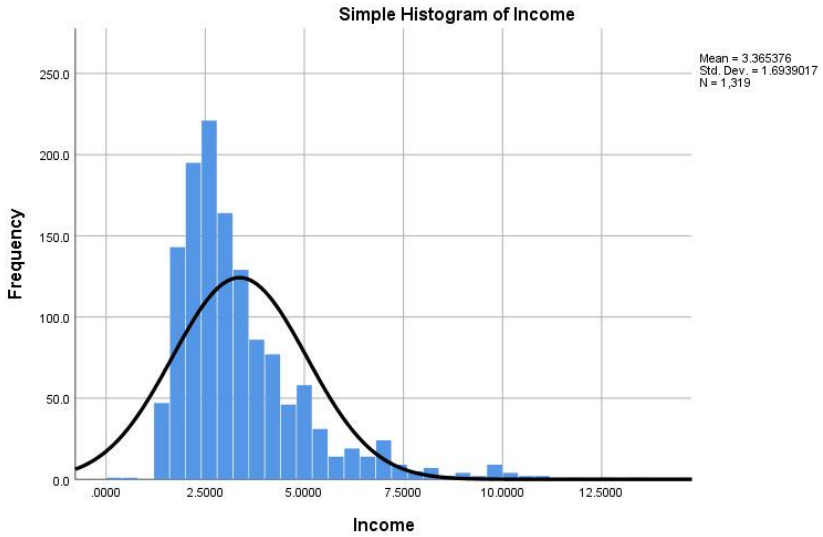
Η δεύτερη υποεπιλογή είναι η ακόλουθη. Επιλέγουμε **Legacy Dialogs** και στη συνέχεια **Histogram**, προκειμένου να εμφανιστεί το παράθυρο της **Εικόνας 3.17**. Στη συνέχεια, περνάμε τη μεταβλητή της οποίας το ιστόγραμμα θέλουμε να κατασκευάσουμε στο δεξιό ορθογώνιο κουτάκι. Ακριβώς πιο κάτω, η επιλογή **Display normal curve** μας επιτρέπει να επιλέξουμε, αν θέλουμε, να εμφανιστεί η γραμμή της κανονικής κατανομής. Όμως, και στην περίπτωση αυτή δεν μπορούμε να κατασκευάσουμε δύο ιστογράμματα με μία μόνο επιλογή. Πατώντας **OK** θα εμφανιστεί το ιστόγραμμα στο **Διάγραμμα 3.2**.

Όμως, ίσως ο πιο βολικός τρόπος κατασκευής ιστογραμμάτων συχνοτήτων είναι μέσω της επιλογής **Analyze**. Επιλέγοντας **Analyze → Descriptive Statistics → Frequencies**, θα εμφανιστεί το παράθυρο της **Εικόνας 3.9**. Αν σε αυτό το παράθυρο επιλέξουμε την **Charts**, θα εμφανιστεί το παράθυρο της **Εικόνας 3.18**. Στο παράθυρο της **Εικόνας 3.4** επιλέγαμε για ποιες μεταβλητές θέλουμε να εμφανιστούν τα περιγραφικά μέτρα. Οπότε, θα πρέπει να έχουμε περάσει τουλάχιστον μία μεταβλητή στο δεξιό κουτάκι, προκειμένου να

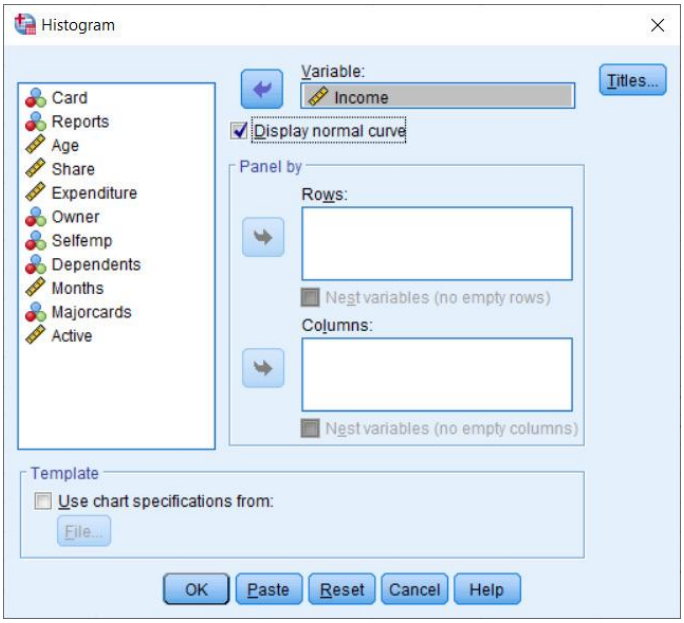
εμφανιστεί το ιστόγραμμα συχνοτήτων της. Επιλέγουμε, λοιπόν, **Histograms** και κάνουμε κλικ στην επιλογή **With normal curve**, αν επιθυμούμε την εμφάνιση της καμπύλης της κανονικής κατανομής. Μπορούμε, αν θέλουμε, να επιλέξουμε να μην εμφανιστεί κανένα περιγραφικό μέτρο. Στην περίπτωση αυτή θα εμφανιστεί μόνο το ιστόγραμμα συχνοτήτων. Ένα πλεονέκτημα της επιλογής αυτής σχετικά με την κατασκευή ιστογραμμάτων συχνοτήτων είναι ότι μπορούμε να «ζητήσουμε» την εμφάνιση ιστογραμμάτων συχνοτήτων για περισσότερες από μία μεταβλητές. Οπότε, επιλέγοντας **Continue** γυρίζουμε στο παράθυρο της **Εικόνας 3.4** και πατώντας **OK** (έχοντας αποεπιλέξει την επιλογή εμφάνισης του πίνακα συχνοτήτων, αφού αναλύουμε συνεχή μεταβλητή) εμφανίζεται ξανά το **Διάγραμμα 3.2**.



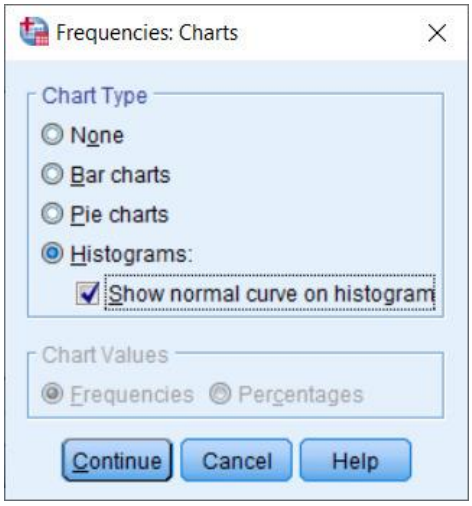
Εικόνα 3.16 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Διάγραμμα 3.2 *Ιστόγραμμα συχνοτήτων με την καμπύλη της κανονικής κατανομής.*



Εικόνα 3.17 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

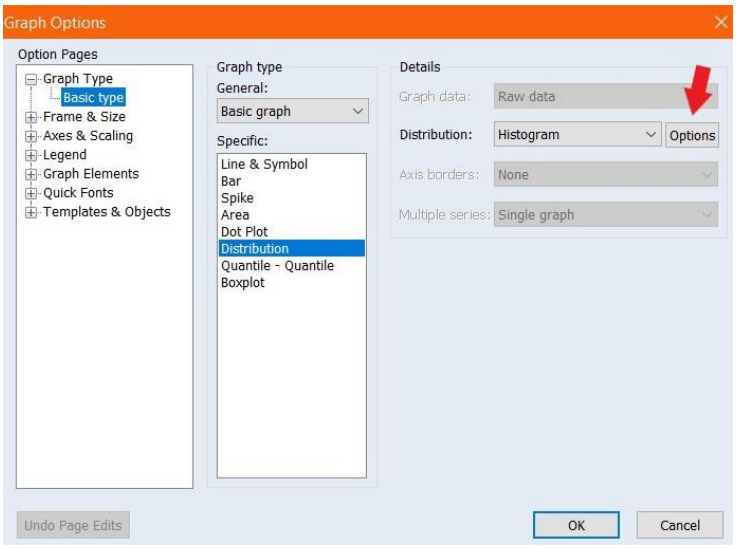


Εικόνα 3.18 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

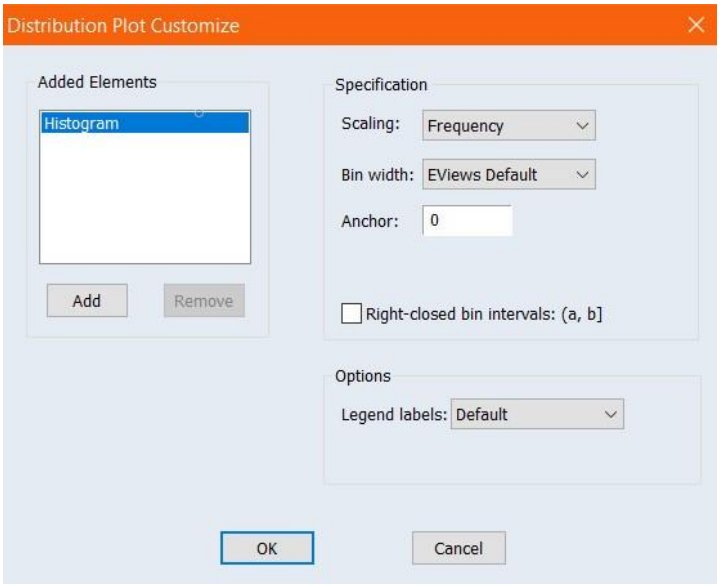
- Στο **Eviews:** Όπως δείξαμε στην προηγούμενη ενότητα, ένας πρώτος τρόπος κατασκευής ιστογράμματος για μια μεταβλητή (έστω για τη μεταβλητή "income"), είναι να «ανοίξουμε» τη συγκεκριμένη μεταβλητή και στη συνέχεια να επιλέξουμε **View → Descriptive Statistics & Tests → Histogram and Stats.** Θα εμφανιστούν το ιστόγραμμα της μεταβλητής "income" και τα βασικά περιγραφικά της μέτρα **(Εικόνα 3.7)**.

Προκειμένου, όμως, να αξιοποιήσουμε τις πολλές επιλογές που έχει το Eviews σχετικά με την κατασκευή ιστογραμμάτων, ο καλύτερος τρόπος είναι να «ανοίξουμε» τη μεταβλητή "income" και να επιλέξουμε **View → Graph**. Στο παράθυρο "Graph Options" που θα εμφανιστεί, επιλέγουμε **Basic type** στην κατηγορία "Option Pages", **Basic Graph** στην κατηγορία "General:", **Distribution** στην κατηγορία "Specific:" και **Histogram** στην κατηγορία "Distribution:" που βρίσκεται δεξιά **(Εικόνα 3.19)**. Πατώντας **OK** θα εμφανιστεί το ιστόγραμμα. Αν θέλουμε να εμφανιστεί και η γραμμή της κανονικής κατανομής, στο παράθυρο της **Εικόνας 3.19** πατάμε το κουμπί **Options** (που υποδεικνύεται από το κόκκινο βέλος) και εμφανίζεται το παράθυρο "Distribution Plot Customize" **(Εικόνα 3.20)**. Στην κατηγορία "Added Elements" επιλέγουμε **Add** και στις επιλογές που θα εμφανιστούν σχετικά με το "Element Type" επιλέγουμε

Theoretical Density και πατάμε **OK**. Πλέον, στην κατηγορία "Added Elements" της **Εικόνας 3.20** θα εμφανίζονται τα elements "Histogram" και "Theoretical Distribution". Πατώντας **OK** και ξανά **OK** (καθώς επιστρέφουμε στο παράθυρο της **Εικόνας 3.19**), εμφανίζεται το ιστόγραμμα της μεταβλητής "income" μαζί με τη γραμμή της κανονικής κατανομής **(Εικόνα 3.21)**.

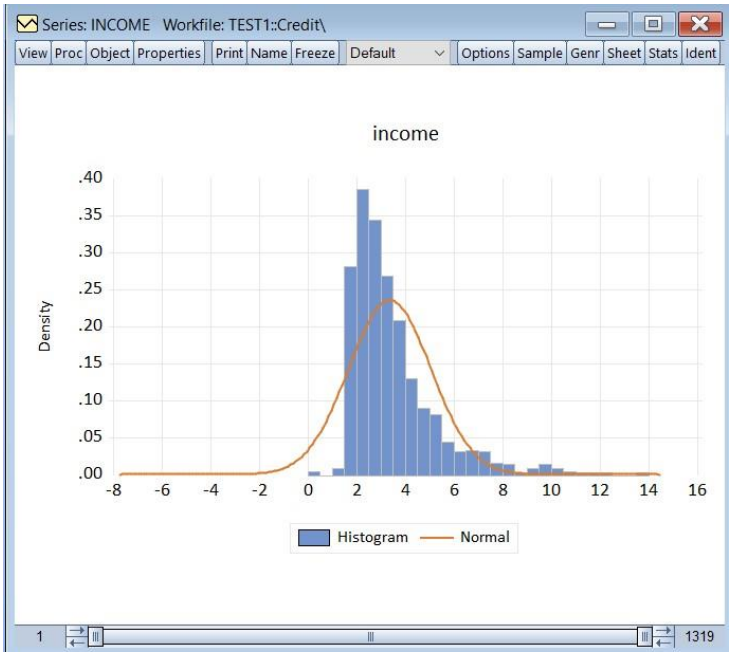


Εικόνα 3.19 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



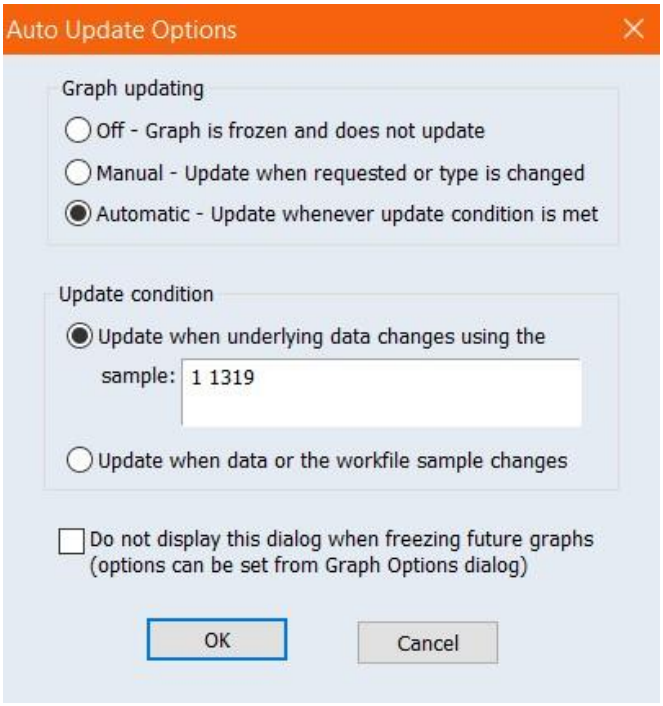
Εικόνα 3.20 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Την εμφάνιση του ιστογράμματος της **Εικόνας 3.21** μπορούμε να την προσαρμόσουμε στις δικές μας προτιμήσεις. Με **διπλό κλικ** πάνω στο ιστόγραμμα ή στη γραμμή της κανονικής κατανομής, εμφανίζεται το παράθυρο "Graph Options" της **Εικόνας 3.19**. Επιλέγοντας "Graph Elements" στην κατηγορία "Option Pages", μπορούμε να τροποποιήσουμε το χρώμα και το πάχος της γραμμής της κανονικής κατανομής, το χρώμα και το εσωτερικό σχέδιο που θα έχουν οι μπάρες του ιστογράμματος κλπ. Με **δεξί κλικ** στους άξονες του γραφήματος και επιλέγοντας **Options**, εμφανίζεται και πάλι το παράθυρο "Graph Options" της **Εικόνας 3.19**, εστιασμένο αυτή τη φορά στην επιλογή "Axes & Scaling". Εδώ μπορούμε να τροποποιήσουμε την κλίμακα των αξόνων, να καθορίσουμε το εύρος τους κλπ. Τέλος, με **διπλό κλικ** στον τίτλο του γραφήματος, στους τίτλους των αξόνων, καθώς και στο υπόμνημα (legend), το Eviews μας εμφανίζει συγκεκριμένα παράθυρα που μας επιτρέπουν να τροποποιήσουμε το αντίστοιχο κείμενο, το είδος και το μέγεθος της γραμματοσειράς κλπ.



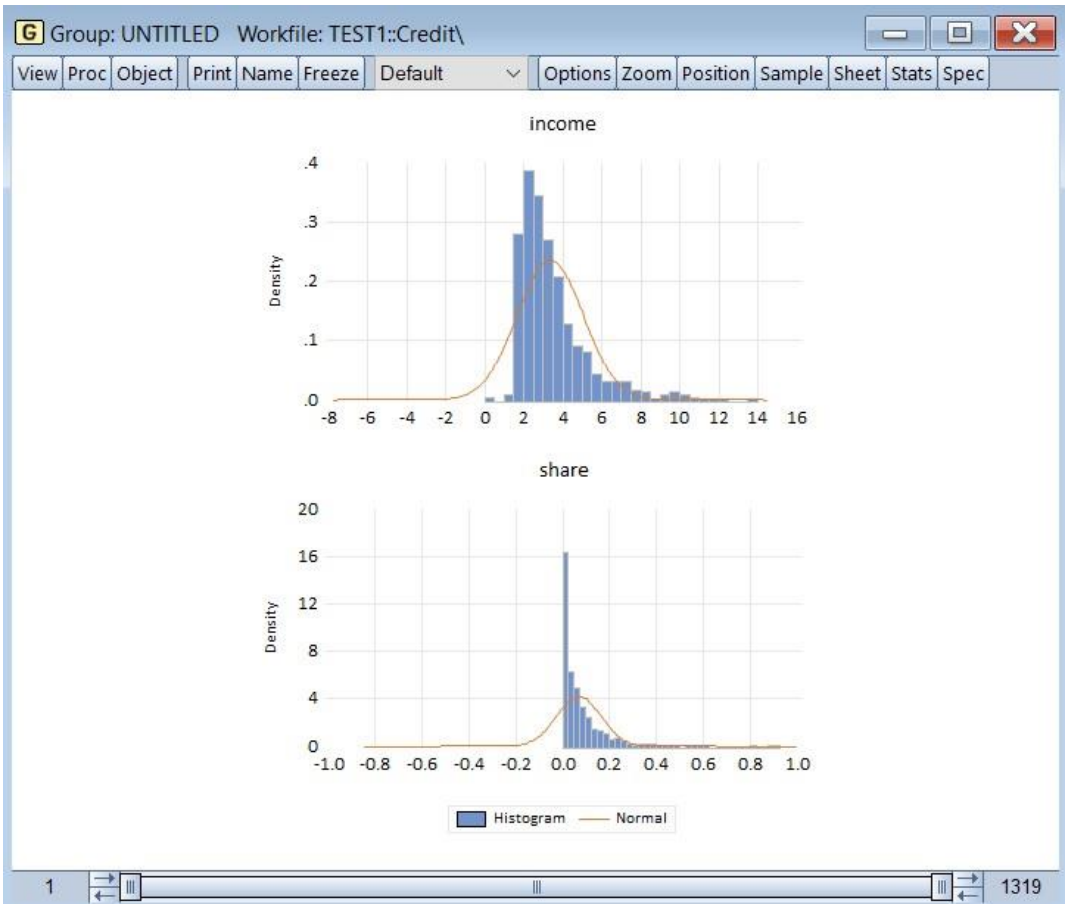
Εικόνα 3.21 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Το ιστόγραμμα που παρουσιάζεται στην **Εικόνα 3.21** μπορούμε να το αποθηκεύσουμε ως γράφημα στο Eviews workfile. Για να συμβεί αυτό, επιλέγουμε **Freeze** από το toolbar, με αποτέλεσμα να εμφανιστεί ένα παράθυρο που ονομάζεται "Auto Update Options" **(Εικόνα 3.22)**. Το συγκεκριμένο παράθυρο έχει τρεις επιλογές σχετικά με την επικαιροποίηση του ιστογράμματος: (α) να μην επικαιροποιείται, (β) να επικαιροποιείται, όταν το επιθυμούμε, ή (γ) να επικαιροποιείται αυτόματα. Στην περίπτωση που επιλέξουμε (β) ή (γ) πρέπει, επίσης, να καθορίσουμε τη συνθήκη επικαιροποίησης, δηλαδή σε ποιες μεταβολές του δείγματος θα γίνεται η συγκεκριμένη επικαιροποίηση. Κάνουμε την κατάλληλη επιλογή και πατάμε **OK**. Στο νέο παράθυρο που αναδύεται, επιλέγουμε **Name** από το toolbar, δίνουμε κάποιο όνομα στο συγκεκριμένο γράφημα και αυτό πλέον εμφανίζεται στο Eviews workfile με το εικονίδιο  .



Εικόνα 3.22 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Τέλος, το Eviews μας επιτρέπει την ταυτόχρονη δημιουργία ιστογραμμάτων για πολλές μεταβλητές. Επιλέγουμε τις μεταβλητές που επιθυμούμε (έστω τις μεταβλητές “income” και “share”), τις «ανοίγουμε» με διπλό κλικ και ακολουθούμε τη διαδικασία που περιγράψαμε παραπάνω. Η μόνη διαφορά είναι ότι στο παράθυρο “Graph Options” της **Εικόνας 3.19** είναι πλέον ενεργοποιημένη στο δεξί μέρος η επιλογή “Multiple series:”, στην οποία επιλέγουμε **Multiple graphs**. Αν θέλουμε να εμφανιστεί και η γραμμή της κανονικής κατανομής για κάθε μεταβλητή, ακολουθούμε τα ίδια βήματα που περιγράψαμε παραπάνω **(Εικόνα 3.23)**.



Εικόνα 3.23 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

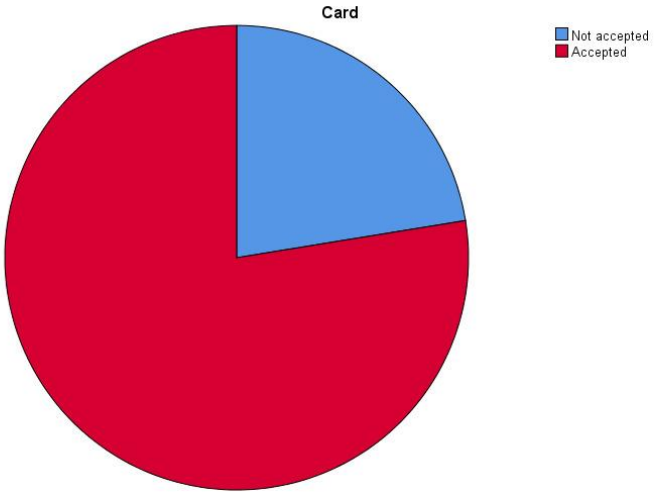
3.4 Κυκλικά διαγράμματα

Στο **SPSS:** Προκειμένου να κατασκευάσουμε ένα κυκλικό διάγραμμα (κυρίως για κατηγορικές μεταβλητές), το μενού του SPSS μας δίνει και πάλι πολλές επιλογές. Εμείς, όμως, θα επικεντρωθούμε στην ανάλυση της πιο χρήσιμης από αυτές. Πιο αναλυτικά, στο παράθυρο της **Εικόνας 3.18** κάνουμε κλικ στην επιλογή **Pie charts**. Στο κάτω μέρος του παραθύρου μπορούμε να επιλέξουμε αν θέλουμε να εμφανιστούν οι συχνότητες ή τα ποσοστά των αποτελεσμάτων των αιτήσεων για χορήγηση πιστωτικής κάρτας. Επιλέγουμε τα ποσοστά, καθώς είναι προτιμότερο να έχουμε την ποσοστιαία κατανομή παρά την απόλυτη (φανταστείτε, για παράδειγμα, τα αποτελέσματα των εκλογών να δίνονταν ως αριθμός ψήφων αντί για ποσοστά!).

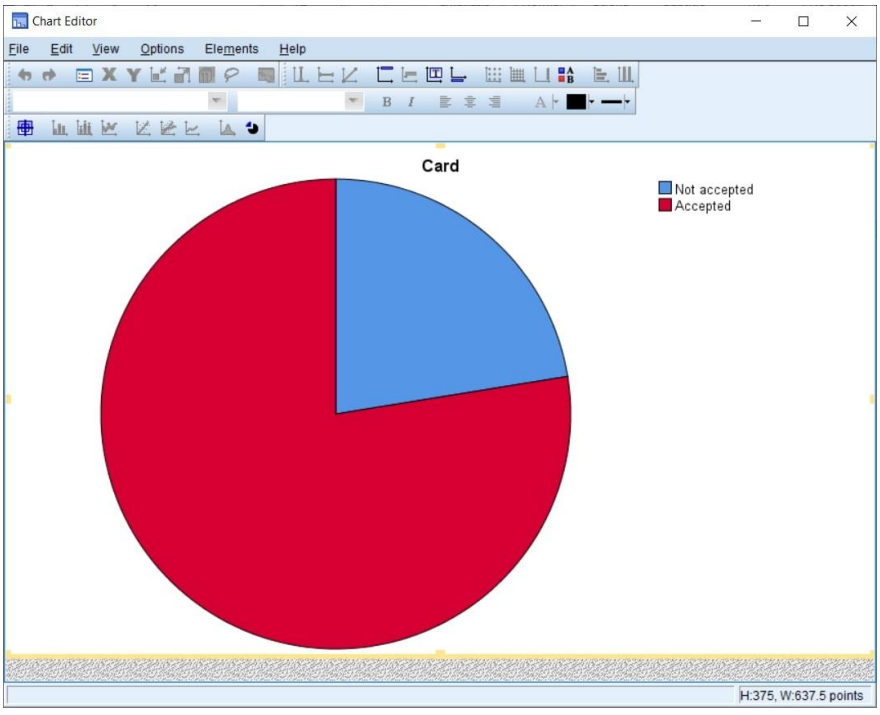
Αν στο παράθυρο της **Εικόνας 3.4** επιλέξουμε **Display frequency tables**, μαζί με το διάγραμμα θα εμφανιστεί και ο πίνακας συχνοτήτων και σχετικών συχνοτήτων για κάθε κατηγορία της ποιοτικής μεταβλητής που έχουμε επιλέξει. Επιλέγουμε, όπως και προηγουμένως, τη μεταβλητή “card”, η οποία αφορά τα αποτελέσματα των αιτήσεων για χορήγηση πιστωτικής κάρτας. Το κυκλικό διάγραμμα εμφανίζεται στο **Διάγραμμα 3.3**. Παρατηρούμε ότι, όπως και στην περίπτωση του ιστογράμματος που κατασκευάστηκε μέσω της συγκεκριμένης επιλογής, έτσι και στην περίπτωση του κυκλικού διαγράμματος

εμφανίζεται ένας πίνακας στο πάνω μέρος του διαγράμματος, ο οποίος παρουσιάζει το πλήθος των τιμών της μεταβλητής που συμμετέχουν στην κατασκευή του διαγράμματος. Το υπόμνημα που εμφανίζεται στο δεξιό μέρος του κυκλικού διαγράμματος δείχνει σε ποια απόφαση σχετικά με την αίτηση χορήγησης πιστωτικής κάρτας αντιστοιχεί το κάθε χρώμα.

Παρατηρούμε, όμως, ότι στο συγκεκριμένο διάγραμμα δεν έχουν εμφανιστεί τα ποσοστά σχετικά με τα αποτελέσματα των αιτήσεων. Για να εμφανιστούν τα συγκεκριμένα ποσοστά, υπάρχουν δύο επιλογές. Η πρώτη επιλογή είναι να κάνουμε διπλό κλικ πάνω στο κυκλικό διάγραμμα. Η δεύτερη επιλογή είναι να κάνουμε δεξί κλικ πάνω στο κυκλικό διάγραμμα και να επιλέξουμε **Edit Content → In Separate Window**. Το αποτέλεσμα θα είναι να εμφανιστούν τα παράθυρα των **Εικόνων 3.24α** και **3.24β**. Επιλέγοντας **Elements → Show Data Labels** (δείτε το παράθυρο της **Εικόνας 3.25**) στο παράθυρο της **Εικόνας 3.24α,** θα εμφανιστεί το **Διάγραμμα 3.4**, το οποίο είναι ακριβώς το ίδιο με το **Διάγραμμα 3.3**, αλλά σε αυτό εμφανίζονται τα ποσοστά.

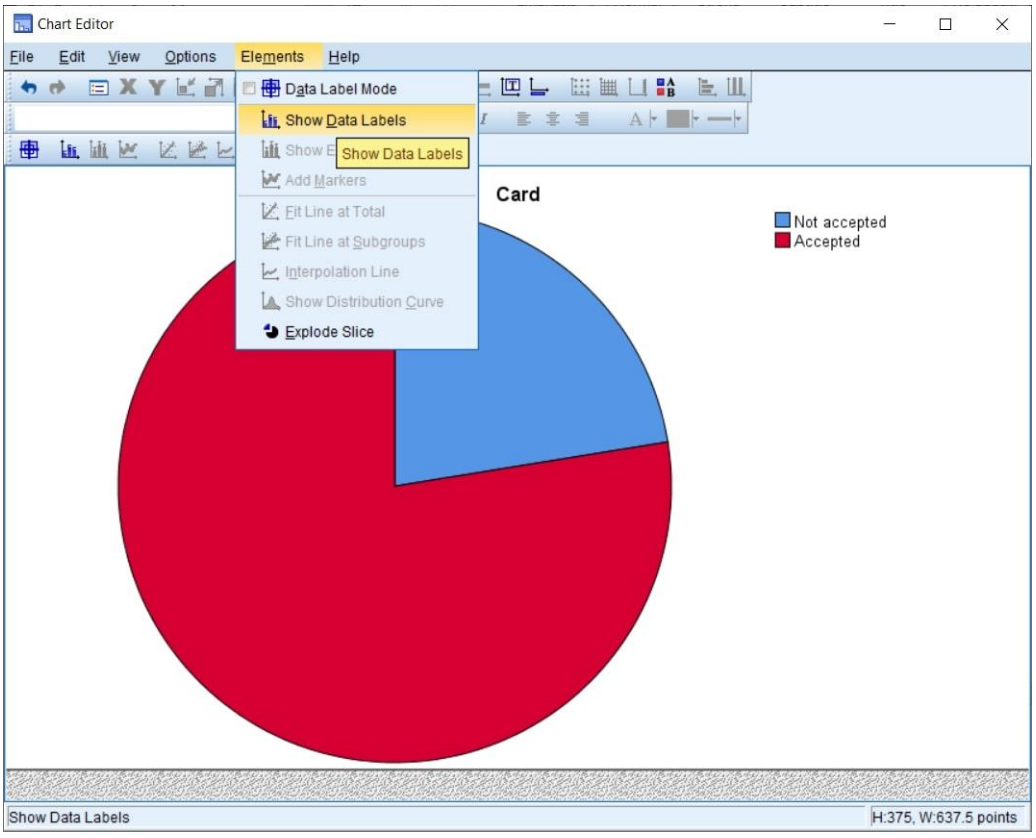


Διάγραμμα 3.3 *Κυκλικό διάγραμμα.*

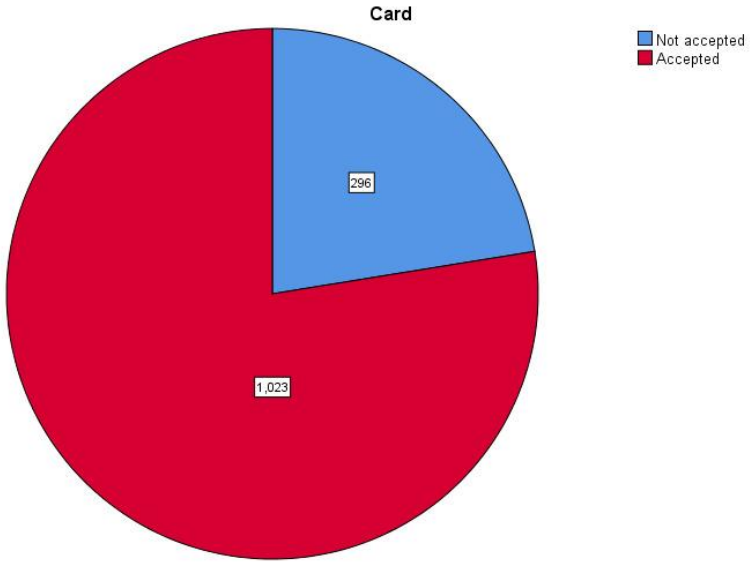


Εικόνα 3.24α *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Εικόνα 3.24β *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

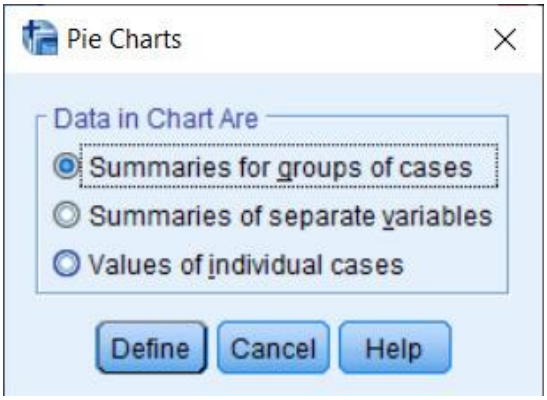


Εικόνα 3.25 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Διάγραμμα 3.4 *Κυκλικό διάγραμμα με ποσοστά.*

Ένας εναλλακτικός τρόπος κατασκευής ενός κυκλικού διαγράμματος είναι να επιλέξουμε **Graphs → Legacy Dialogs → Pie** και στο παράθυρο της **Εικόνας 3.26** που θα εμφανιστεί να επιλέξουμε **Summaries of group of cases** και μετά **Define**.

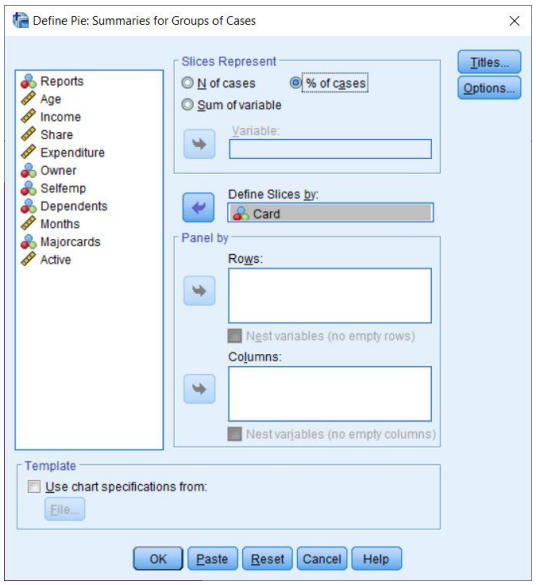


Εικόνα 3.26 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

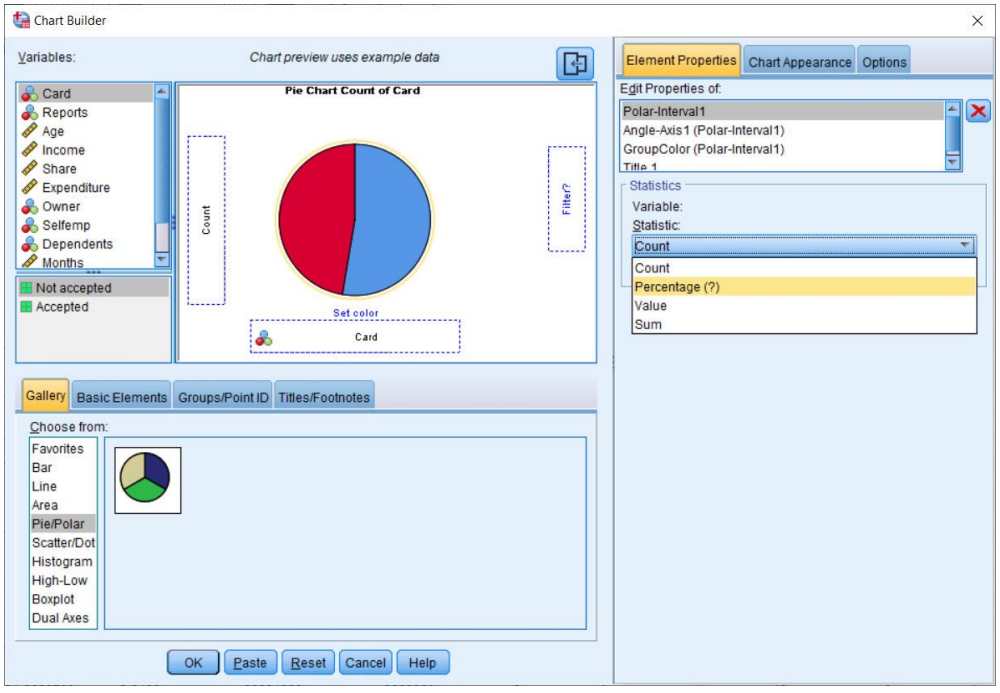
Στο παράθυρο που θα εμφανιστεί **(Εικόνα 3.27)**, περνάμε την ποιοτική μεταβλητή, της οποίας το κυκλικό διάγραμμα θέλουμε να κατασκευάσουμε, στο λευκό κουτί κάτω από την ένδειξη **Define Slices by**. Επίσης, στο πάνω μέρος του παραθύρου **Slices Represent** επιλέξαμε **% of cases**, προκειμένου να εμφανιστούν τα ποσοστά στο διάγραμμα. Στο κάτω δεξιό μέρος του παραθύρου μας δίνεται η δυνατότητα να δώσουμε κάποιον τίτλο στο διάγραμμα (επιλογή **Titles**). Επίσης, στο δεξιό μέρος του παραθύρου της **Εικόνας 3.27** υπάρχει η επιλογή **Options**, η οποία μας επιτρέπει να επιλέξουμε αν θέλουμε οι εκλιπούσες τιμές να συμπεριληφθούν στο διάγραμμα ή όχι. Το SPSS έχει ως προεπιλογή να μην συμπεριλαμβάνονται οι συγκεκριμένες τιμές στο κυκλικό διάγραμμα. Οπότε, πατώντας **OK** θα εμφανιστεί το **Διάγραμμα 3.3**. Και στην περίπτωση αυτή, αν θέλουμε να εμφανίζονται τα ποσοστά, θα πρέπει να ξανακάνουμε την ίδια διαδικασία που περιγράψαμε προηγουμένως.

Προηγουμένως αναλύσαμε τους τρόπους κατασκευής κυκλικών διαγραμμάτων ή διαγραμμάτων «πίτας», όπως αλλιώς ονομάζονται. Όμως, το μειονέκτημα των συγκεκριμένων τρόπων είναι ότι δεν εμφανίζουν τα ποσοστά των τιμών των μεταβλητών στο κυκλικό διάγραμμα, με αποτέλεσμα να πρέπει να τα τοποθετήσουμε στη συνέχεια. Εναλλακτικά, επιλέγοντας **Graphs → Chart Builder** θα εμφανιστεί το παράθυρο της **Εικόνας 3.15**. Στην επιλογή **Gallery** θα επιλέξουμε **Pie/Polar** και θα σύρουμε την επιλογή αυτή στο λευκό κουτί. Στη συνέχεια, θα επιλέξουμε την κατηγορική μεταβλητή, της οποίας οι τιμές θέλουμε

να εμφανιστούν στο κυκλικό διάγραμμα, και θα την σύρουμε στο λευκό κουτί με την ένδειξη **Slice by**. Επιλέγουμε **Statistic**, στη συνέχεια **Percentage(?)** (δείτε παράθυρο **Εικόνας 3.28**) και μετά **Apply**. Τέλος, πατώντας **OK** θα προκύψει το αποτέλεσμα, το οποίο είναι ακριβώς το ίδιο με αυτό του **Διαγράμματος 3.3**. Και στην περίπτωση αυτή, ο τρόπος για να εμφανιστούν τα ποσοστά είναι ο ίδιος με αυτόν που εφαρμόστηκε προηγουμένως.



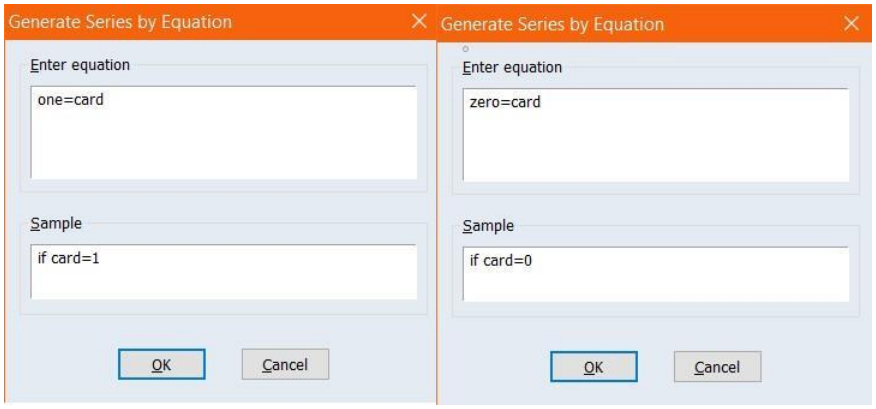
Εικόνα 3.27 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Εικόνα 3.28 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

- Στο **Eviews**: Η διαδικασία δημιουργίας κυκλικού διαγράμματος («πίτας») είναι σχετικά εύκολη, όταν πρόκειται για ένα group δύο ή περισσότερων μεταβλητών. Στην περίπτωση, όμως, που θέλουμε να κατασκευάσουμε κυκλικό διάγραμμα για μία μεταβλητή, η διαδικασία

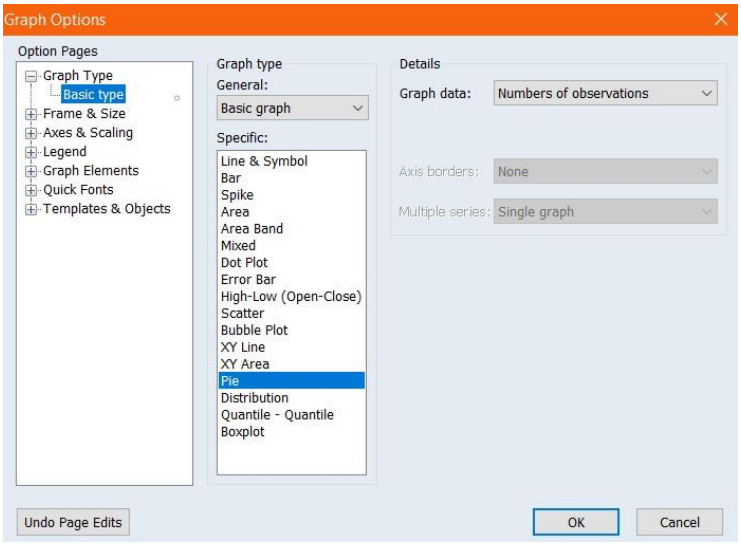
είναι έμμεση, καθώς το Eviews δεν παρέχει τη δυνατότητα κατασκευής τέτοιου διαγράμματος για μία μόνο μεταβλητή. Έστω, λοιπόν, ότι θέλουμε να δημιουργήσουμε ένα τέτοιο διάγραμμα για την κατηγορική μεταβλητή "card", η οποία δείχνει το αποτέλεσμα των αιτήσεων χορήγησης πιστωτικής κάρτας. Αρχικά, θα πρέπει να δημιουργήσουμε δύο νέες κατηγορικές μεταβλητές, με τον τρόπο που περιγράψαμε στην ενότητα 2.10 **(Εικόνα 2.39)**: μία μεταβλητή που θα περιλαμβάνει τις παρατηρήσεις με τιμή 1 (έστω "one") και μία μεταβλητή που θα περιλαμβάνει τις παρατηρήσεις με τιμή 0 (έστω "zero"). Στο παράθυρο "Generate Series by Equation" που θα εμφανιστεί, γράφουμε στο πάνω κουτί **one=card** και στο κάτω κουτί **if card=1** (αριστερό μέρος της **Εικόνας 3.29**) και πατάμε **OK**. Η μεταβλητή "one" έχει πλέον δημιουργηθεί, και όπως βλέπουμε «ανοίγοντάς» την, έχει την ένδειξη NA στις παρατηρήσεις εκείνες τις μεταβλητής "card" που παίρνουν την τιμή 0. Ακολουθούμε την ίδια διαδικασία και για τη μεταβλητή "zero", γράφοντας στην περίπτωση αυτή **zero=card** στο πάνω κουτί του παραθύρου "Generate Series by Equation" και **if card=0** στο κάτω κουτί (δεξιό μέρος της **Εικόνας 3.29**).



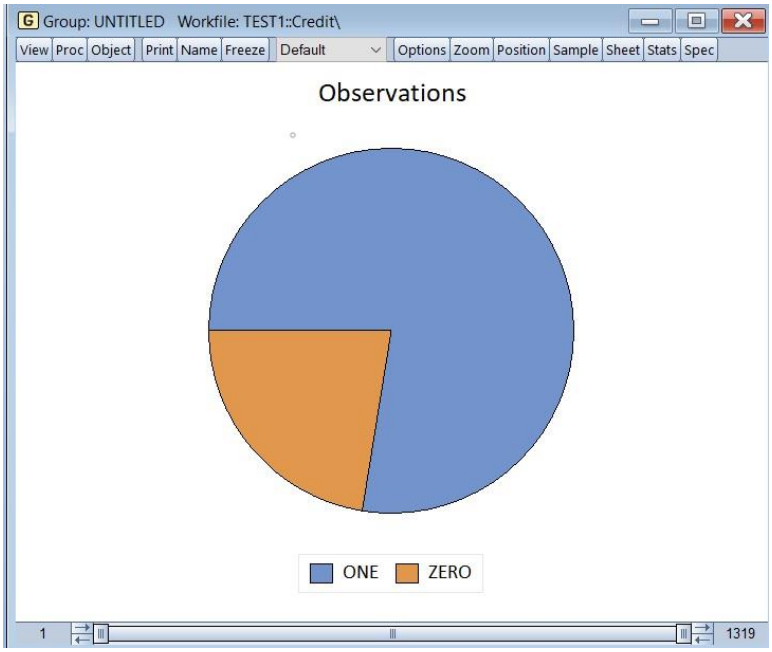
Εικόνα 3.29 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Στη συνέχεια, «ανοίγουμε» με διπλό κλικ τις μεταβλητές "one" και "zero" ως group, και επιλέγουμε από το toolbar **View → Graph**. Στο παράθυρο "Graph Options" που θα εμφανιστεί, επιλέγουμε **Basic graph** στο "General:", **Pie** στο "Specific:" και **Numbers of observations** στο "Graph data:" **(Εικόνα 3.30)**. Πατάμε **OK** και το κυκλικό διάγραμμα είναι έτοιμο **(Εικόνα 3.31)**. Με **διπλό κλικ** πάνω στην «πίτα» εμφανίζεται το παράθυρο "Graph Options" της **Εικόνας 3.29**, όπου επιλέγοντας "Graph Elements" στην κατηγορία "Option Pages" μπορούμε να διαμορφώσουμε την «πίτα» με βάση τις προτιμήσεις μας. Με **διπλό κλικ** πάνω στον τίτλο "Observations", εμφανίζεται ένα παράθυρο που μας επιτρέπει τη μετονομασία του σε "Card" (ή σε ό,τι άλλο θελήσουμε) και την εν γένει τροποποίησή του (είδος και μέγεθος γραμματοσειράς, χρώμα κλπ.). Τέλος, με **διπλό κλικ** στο υπόμνημα (legend), εμφανίζεται ξανά το παράθυρο "Graph Options" της **Εικόνας 3.29**, στο οποίο μπορούμε να μετονομάσουμε το "ONE" σε "Accepted", το "ZERO" σε "Not accepted" και εν γένει να κάνουμε όποιες τροποποιήσεις θέλουμε.

Και στην περίπτωση του Eviews, τα ποσοστά σχετικά με τα αποτελέσματα των αιτήσεων **(Εικόνα 3.14)** δεν εμφανίζονται στην «πίτα». Μπορούμε να εισάγουμε καθένα από τα δύο ποσοστά κάνοντας **δεξί κλικ** πάνω στην «πίτα» και επιλέγοντας **Add text**. Στο πλαίσιο κειμένου που θα εμφανιστεί, θα πρέπει να γράψουμε τα συγκεκριμένα ποσοστά και στη συνέχεια να σύρουμε τα δύο πλαίσια κειμένου σε όποιο σημείο της «πίτας» επιθυμούμε. Η τελική μορφή του κυκλικού διαγράμματος παρουσιάζεται στην **Εικόνα 3.32**. Το διάγραμμα αυτό μπορούμε να το αποθηκεύσουμε ως γράφημα στο Eviews workfile, ακολουθώντας τη διαδικασία που περιγράψαμε στην ενότητα 3.3 σχετικά με την αποθήκευση ιστογραμμάτων.



Εικόνα 3.30 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

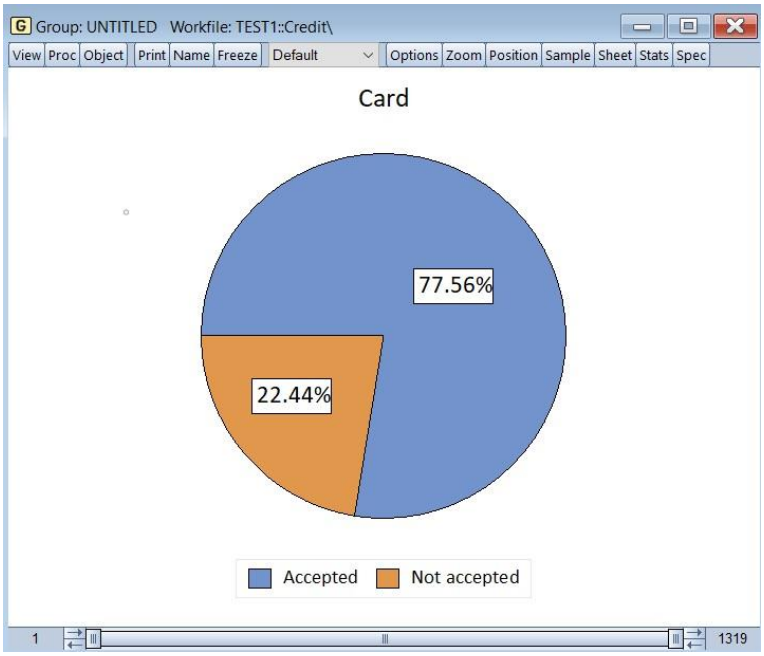


Εικόνα 3.31 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

3.5 Ραβδογράμματα

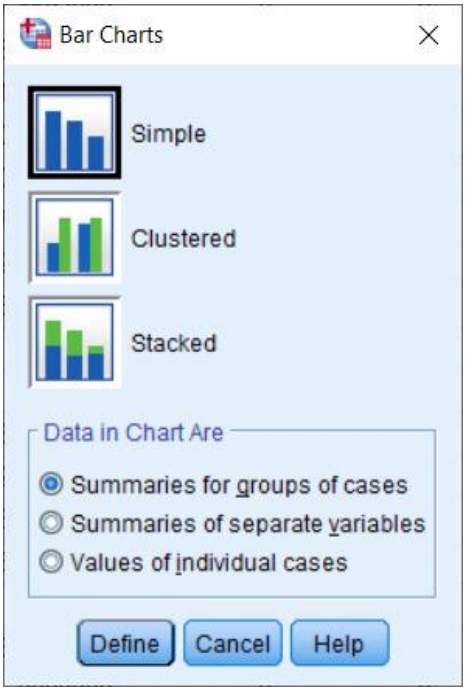
Στο **SPSS**: Ας υποθέσουμε και πάλι ότι έχουμε κατηγορικές μεταβλητές και θέλουμε να ερευνήσουμε τη μορφή που θα έχει το κυκλικό διάγραμμα, αν είναι σε στήλες. Οπότε, θα κατασκευάσουμε το λεγόμενο ραβδόγραμμα, το οποίο θυμίζει λίγο το ιστόγραμμα ως προς τα ορθογώνια (ράβδους), αλλά έχει μία σημαντική διαφορά από το τελευταίο, καθώς τα ορθογώνια δεν είναι «κολλημένα» μεταξύ τους. Για να κατασκευάσουμε ραβδογράμματα, υπάρχουν πολλές επιλογές. Εμείς θα εστιάσουμε στην πιο εύχρηστη από αυτές. Επιλέγοντας **Graphs → Legacy Dialogs → Bar**, εμφανίζεται το παράθυρο της **Εικόνας 3.33**. Στο παράθυρο αυτό θα επιλέξουμε το πρώτο εικονίδιο (**Simple**) και συνέχεια **Define**, προκειμένου να μεταβούμε σε ένα παράθυρο που είναι παρόμοιο με αυτό της **Εικόνας 3.27**. Στο παράθυρο αυτό θα περάσουμε την ποιοτική μεταβλητή, για την οποία θέλουμε να κατασκευάσουμε το ραβδόγραμμα, στο λευκό κουτί με την ένδειξη **Category Axis:**. Επίσης, μας δίνεται η δυνατότητα εμφάνισης των ποσοστών αντί

των συχνοτήτων. Στην περίπτωση αυτή, το διάγραμμα θα είναι το ίδιο, όμως στον κατακόρυφο άξονα θα βρίσκονται τα ποσοστά των τιμών (ή των επιπέδων) της κατηγορικής μεταβλητής αντί για τις συχνότητες.

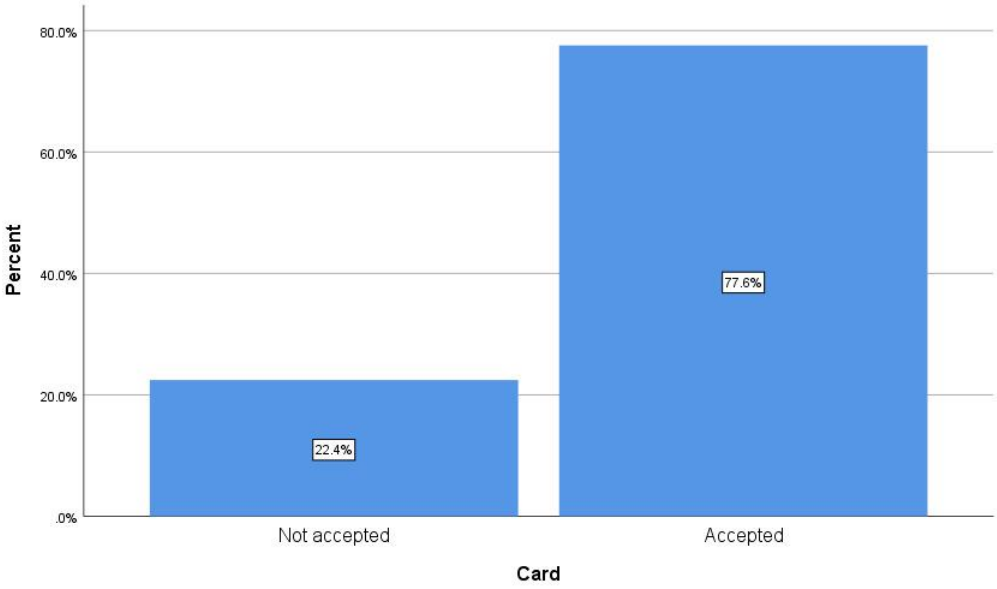


Εικόνα 3.32 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Επιπλέον, η επιλογή **Titles** μας δίνει τη δυνατότητα να δώσουμε κάποιον τίτλο της επιλογής μας στο διάγραμμα, ενώ η επιλογή **Options** μας επιτρέπει να εμφανίσουμε (αν επιθυμούμε) ένα ακόμα ραβδόγραμμα, το οποίο θα περιέχει το πλήθος των χαμένων τιμών. Στο παράδειγμά μας, έχουμε επιλέξει τη μεταβλητή που δείχνει το αποτέλεσμα των αιτήσεων χορήγησης πιστωτικής κάρτας. Πατώντας, λοιπόν, **OK** το αποτέλεσμα φαίνεται στο **Διάγραμμα 3.5.** Στο συγκεκριμένο διάγραμμα εμφανίζονται και τα ποσοστά των τιμών, έχοντας ακολουθήσει και πάλι τη διαδικασία που περιγράψαμε προηγουμένως.



Εικόνα 3.33 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



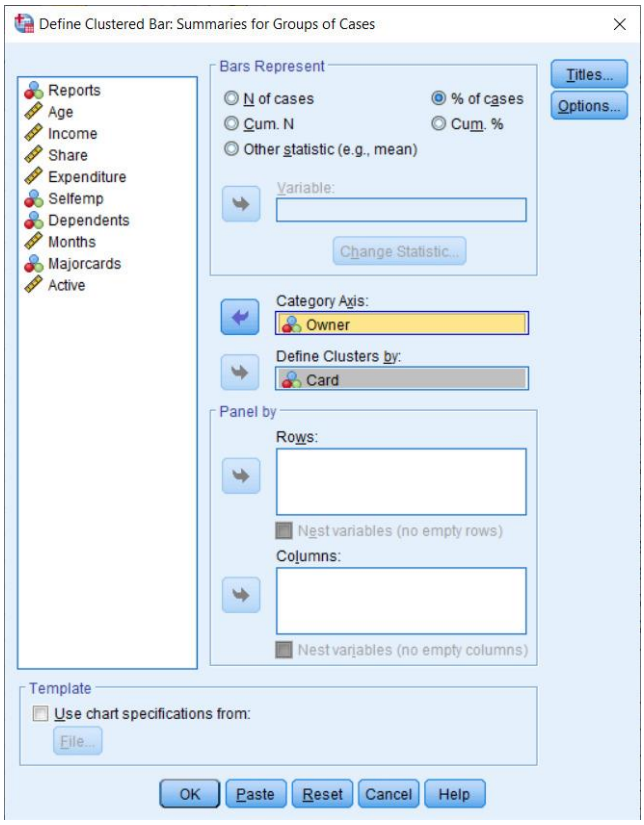
Διάγραμμα 3.5 *Ραβδόγραμμα συχνοτήτων.*

Στην περίπτωση που θέλουμε να κατασκευάσουμε ένα ραβδόγραμμα για δύο ποιοτικές μεταβλητές μαζί, στο παράθυρο της **Εικόνας 3.33** θα επιλέξουμε το δεύτερο εικονίδιο **Clustered** και στη συνέχεια **Continue**, προκειμένου να οδηγηθούμε στο παράθυρο της **Εικόνας 3.34**, το οποίο είναι ελαφρώς διαφορετικό από αυτό της **Εικόνας 3.27**. Οι ποιοτικές μεταβλητές θα περαστούν στα λευκά κουτιά με τις ενδείξεις **Category Axis:** και **Define Clusters by:**. Στο πρώτο κουτί επιλέγουμε τη μεταβλητή που δηλώνει αν το άτομο είναι ιδιοκτήτης της τρέχουσας κατοικίας του ή όχι (“owner”), ενώ στο δεύτερο κουτί επιλέγουμε τη μεταβλητή που δηλώνει το αποτέλεσμα των αιτήσεων χορήγησης πιστωτικής κάρτας (“card”). Το ραβδόγραμμα εμφανίζεται στο **Διάγραμμα 3.6** και παρουσιάζει το αποτέλεσμα της αίτησης ενός ατόμου για πιστωτική κάρτα, ανάλογα με το αν το συγκεκριμένο άτομο είναι ιδιοκτήτης της τρέχουσας κατοικίας του ή όχι. Από την άλλη πλευρά, αν θέλουμε να παρουσιάσουμε την ιδιοκτησία της κατοικίας ανάλογα με το αποτέλεσμα της αίτησης για πιστωτική κάρτα, θα αλλάξουμε τη σειρά με την οποία συμπληρώνουμε τις μεταβλητές στα λευκά κουτιά **(Εικόνα 3.34)**.

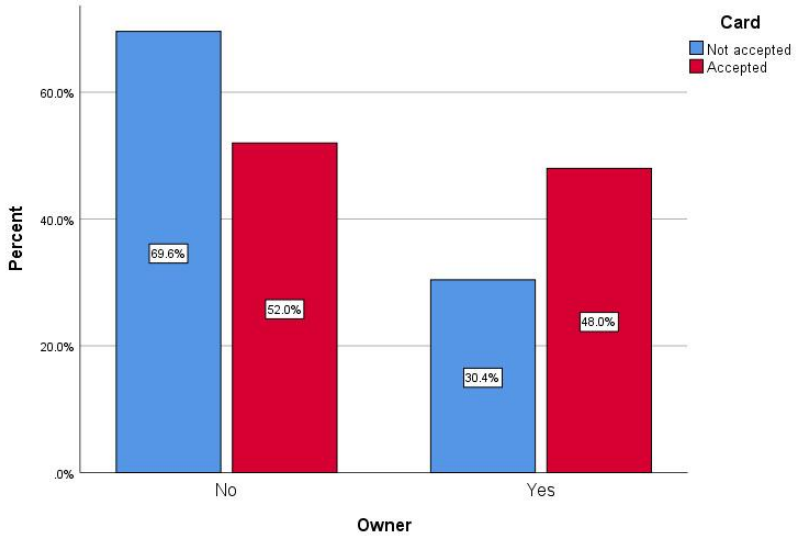
- Στο **Eviews**: Έστω ότι θέλουμε να δημιουργήσουμε το ραβδόγραμμα της κατηγορικής μεταβλητής “card”, η οποία δείχνει το αποτέλεσμα των αιτήσεων χορήγησης πιστωτικής κάρτας. «Ανοίγουμε» τη μεταβλητή αυτή και επιλέγουμε από το toolbar **View → Graph**. Στο παράθυρο “Graph Options” που θα εμφανιστεί, επιλέγουμε **Categorical graph** στο “General:”, **Bar** στο “Specific:” και **Numbers of observations** στο “Graph data:”. Επίσης, στην κατηγορία “Factors – series defining categories”, στην επιλογή “Within graph:” γράφουμε **card (Εικόνα 3.35)**. Αυτό συμβαίνει, για να μπορέσει το Eviews να προσδιορίσει τον παράγοντα (“factor”) με βάση τον οποίο θα «χωρίσει» τις παρατηρήσεις της μεταβλητής “card” στο ραβδόγραμμα (αν αφήσουμε κενό το συγκεκριμένο κελί και πατήσουμε OK, θα εμφανιστεί ένα ραβδόγραμμα με μία μόνο ράβδο που θα περιλαμβάνει τον συνολικό αριθμό των παρατηρήσεων). Πατάμε **OK** και το ραβδόγραμμα εμφανίζεται στην **Εικόνα 3.36**.

Όπως και στην περίπτωση της «πίτας», με **διπλό κλικ** πάνω στο ραβδόγραμμα εμφανίζεται το παράθυρο “Graph Options” της **Εικόνας 3.35**, όπου επιλέγοντας “Graph Elements” στην κατηγορία “Option Pages”, μπορούμε να το διαμορφώσουμε με βάση τις προτιμήσεις μας. Αν, για παράδειγμα, στην υποκατηγορία “Bar-Area-Pie” επιλέξουμε το “Label above bar”, θα εμφανιστεί πάνω από κάθε ράβδο ο αντίστοιχος αριθμός των παρατηρήσεων. Με **διπλό κλικ** στον τίτλο “Observations of CARD by CARD”, θα εμφανιστεί ένα παράθυρο που μας επιτρέπει τη μετονομασία του σε “Card” (ή σε οτιδήποτε άλλο επιθυμούμε) και την εν γένει τροποποίησή του (είδος και μέγεθος γραμματοσειράς, χρώμα κλπ.). Τέλος, με **διπλό κλικ** σε οποιοδήποτε σημείο του ραβδογράμματος, εμφανίζεται ξανά το παράθυρο “Graph Options”

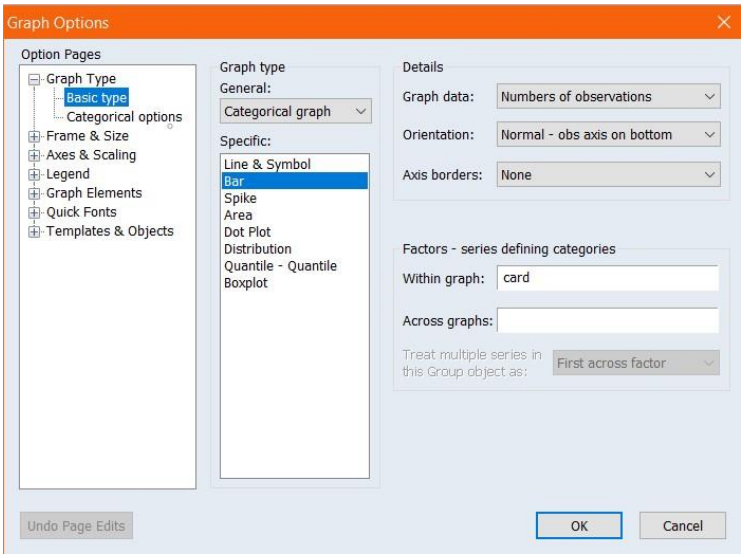
της **Εικόνας 3.35**, στο οποίο επιλέγοντας “Legend” στο “Option Pages” μπορούμε να δημιουργήσουμε ένα υπόμνημα. Πλέον, το ραβδόγραμμα θα έχει τη μορφή της **Εικόνας 3.37**.



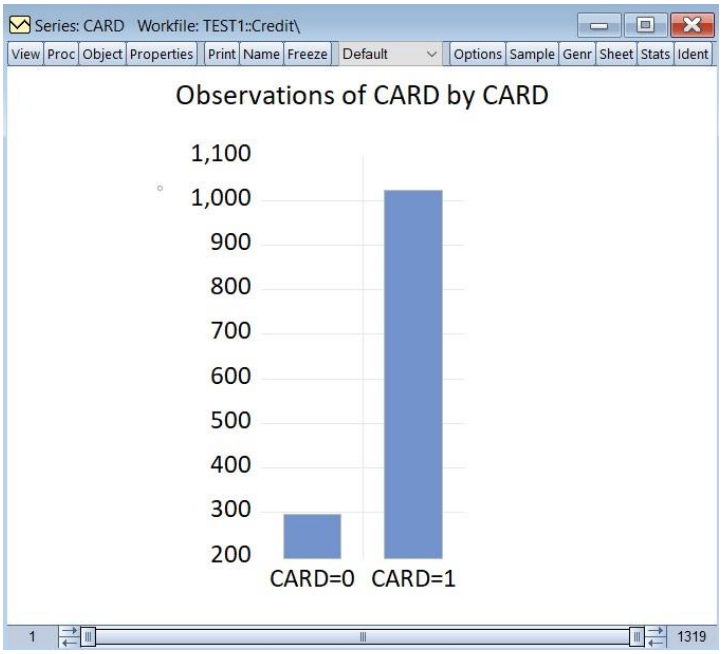
Εικόνα 3.34 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Διάγραμμα 3.6 *Ραβδόγραμμα δύο ποιοτικών μεταβλητών.*

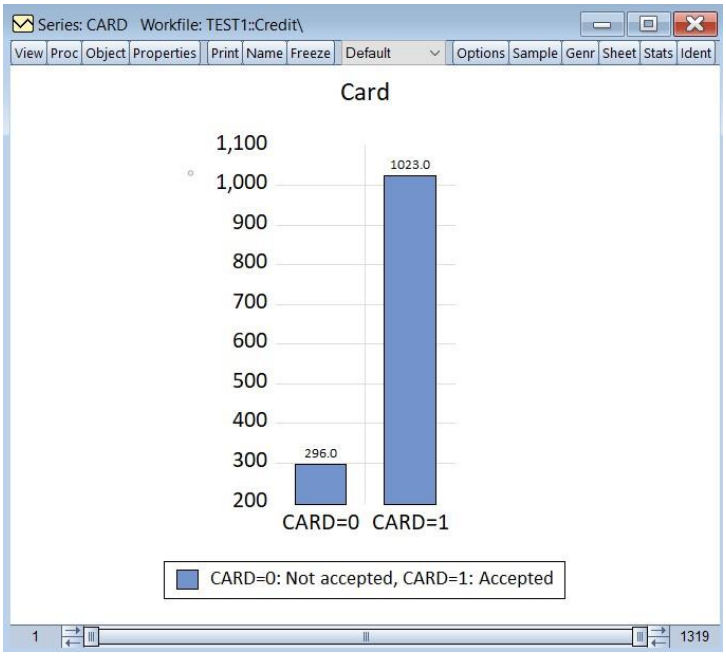


Εικόνα 3.35 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

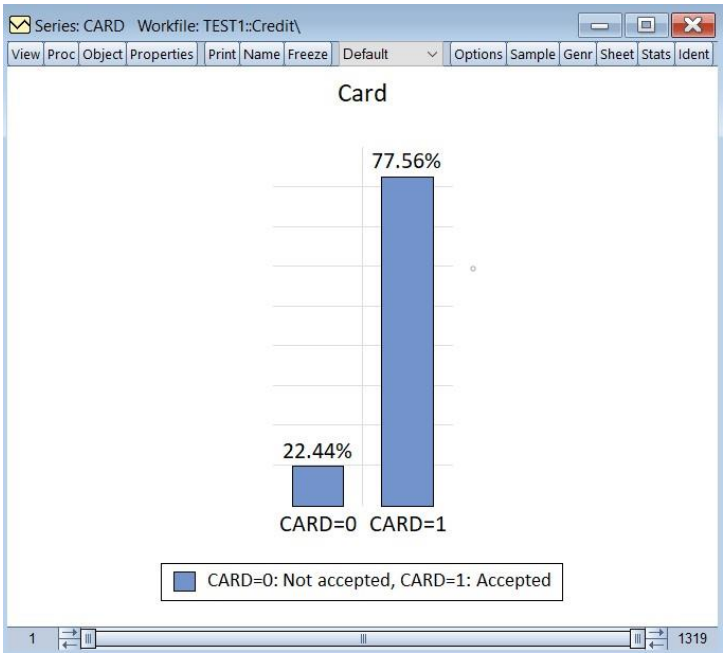


Εικόνα 3.36 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Το Eviews είναι εν γένει δύσχρηστο σχετικά με τις κατηγορικές μεταβλητές. Στην περίπτωση του παραπάνω ραβδογράμματος δεν παρέχεται η δυνατότητα εμφάνισης των ποσοστών της **Εικόνας 3.14** αντί των συχνοτήτων. Οπότε, αν θέλουμε να τα εμφανίσουμε, θα το κάνουμε με έμμεσο τρόπο. Αρχικά, κάνουμε **διπλό κλικ** στο ραβδόγραμμα και επιλέγουμε Option Pages → Graph Elements → Bar-Area-Pie → **No bar labels**, προκειμένου να εξαφανιστεί πάνω από κάθε ράβδο ο αντίστοιχος αριθμός των παρατηρήσεων. Στη συνέχεια, κάνουμε **διπλό κλικ** στον αριστερό άξονα, επιλέγουμε "Data axis labels" και «τσεκάρουμε» την επιλογή **Hide labels**. Με τον τρόπο αυτό, ο αριστερός άξονας που δεν είναι εκφρασμένος σε ποσοστά θα εξαφανιστεί από το γράφημα. Τέλος, εισάγουμε το καθένα από τα δύο ποσοστά στο πλαίσιο κειμένου που θα εμφανιστεί, αν κάνουμε **δεξί κλικ** πάνω στο ραβδόγραμμα και επιλέξουμε **Add text**. Όταν ολοκληρώσουμε τη διαδικασία, σύρουμε τα δύο πλαίσια κειμένου σε όποιο σημείο του γραφήματος επιθυμούμε. Η τελική μορφή του ραβδογράμματος παρουσιάζεται στην **Εικόνα 3.38**, ενώ μπορούμε να το αποθηκεύσουμε ως γράφημα στο Eviews workfile, ακολουθώντας τη διαδικασία που περιγράψαμε στην ενότητα 3.3.



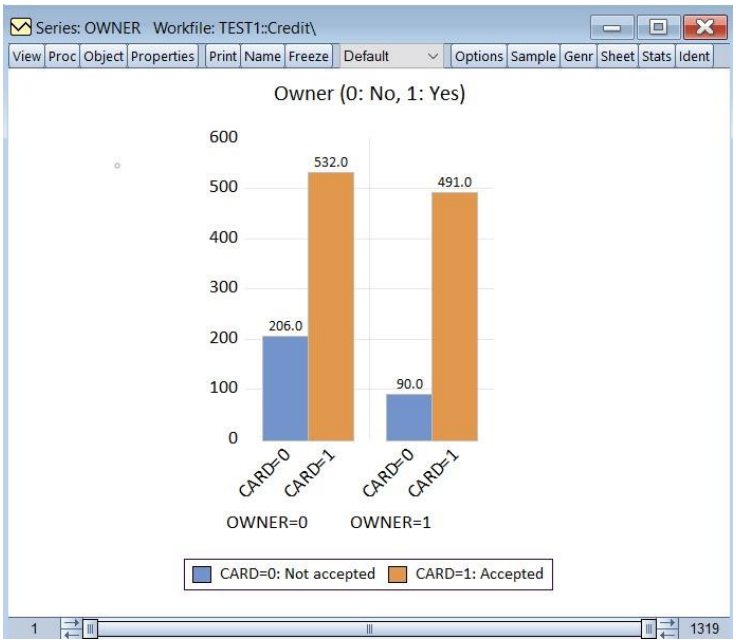
Εικόνα 3.37 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Εικόνα 3.38 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

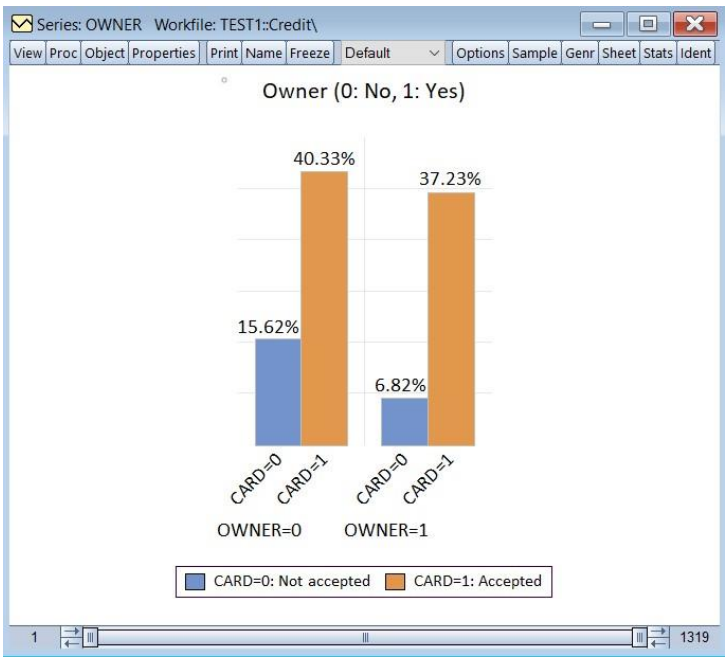
Μπορούμε, επίσης, να κατασκευάσουμε ραβδόγραμμα για δύο ποιοτικές μεταβλητές μαζί. Έστω, όπως και στο παράδειγμα του SPSS, ότι θέλουμε να παραστήσουμε το αποτέλεσμα της αίτησης ενός ατόμου για πιστωτική κάρτα (“card”) ανάλογα με το αν είναι ιδιοκτήτης ή όχι της τρέχουσας κατοικίας του (“owner”). Στην περίπτωση αυτή, «ανοίγουμε» τη μεταβλητή “owner” και επιλέγουμε από το toolbar **View → Graph**. Στο παράθυρο “Graph Options” που θα εμφανιστεί, κάνουμε ακριβώς τις ίδιες επιλογές με αυτές της **Εικόνας 3.35**, με μόνη διαφορά ότι στην κατηγορία “Factors – series defining categories”, στην επιλογή “Within graph:” γράφουμε **owner card**. Με τον τρόπο αυτό, το Eviews θα προσδιορίσει τους παράγοντες (“factors”) με βάση τους οποίους θα «χωρίσει» τις παρατηρήσεις της μεταβλητής “owner” στο ραβδόγραμμα. Και πάλι, με **διπλό κλικ** πάνω στο ραβδόγραμμα μπορούμε να το διαμορφώσουμε με βάση τις προτιμήσεις μας. Ακολουθώντας τη διαδικασία που περιγράψαμε προηγουμένως, μπορούμε να

επιλέξουμε να εμφανιστεί πάνω από κάθε ράβδο ο αντίστοιχος αριθμός των παρατηρήσεων, να τροποποιήσουμε τον τίτλο του γραφήματος και να εισάγουμε ένα υπόμνημα. Πλέον, το ραβδόγραμμα θα έχει τη μορφή της **Εικόνας 3.39**.



Εικόνα 3.39 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Και στο συγκεκριμένο ραβδόγραμμα, το Eviews δεν παρέχει τη δυνατότητα εμφάνισης των ποσοστών αντί των συχνοτήτων. Οπότε, αν θέλουμε να τα εμφανίσουμε, θα ακολουθήσουμε και πάλι τον έμμεσο τρόπο που περιγράψαμε προηγουμένως. Η τελική μορφή του ραβδογράμματος παρουσιάζεται στην **Εικόνα 3.40**, ενώ μπορούμε να το αποθηκεύσουμε ως γράφημα στο Eviews workfile, ακολουθώντας και πάλι τη διαδικασία που περιγράψαμε στην ενότητα 3.3.



Εικόνα 3.40 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Βιβλιογραφία

Ξενόγλωσση

Casella, G., & Berger, R.L. (2002). *Statistical Inference* (2nd ed.). Pacific Grove, California: Duxbury.

Κεφάλαιο 4 Έλεγχοι υποθέσεων

Σύνοψη
Η ανάλυση της επαγωγικής στατιστικής ξεκινάει από το συγκεκριμένο κεφάλαιο και πιο συγκεκριμένα αφορά έλεγχο κανονικότητας, διαστήματα εμπιστοσύνης, συντελεστές γραμμικής συσχέτισης, ελέγχους υποθέσεων για έναν και δύο μέσους, καθώς και ελέγχους ανεξαρτησίας μεταξύ δύο κατηγορικών μεταβλητών. Οι στόχοι του κεφαλαίου αυτού είναι ο/η αναγνώστης/-τρια να μπορεί να επιλέξει και να εφαρμόσει τους κατάλληλους ελέγχους υποθέσεων για τα δεδομένα του/της.

Προαπαιτούμενη γνώση
Απαιτούνται βασικές γνώσεις στατιστικής.

4.1 Έλεγχος κανονικότητας

Αναφέρθηκε στο προηγούμενο κεφάλαιο πως, όταν το ιστόγραμμα συχνοτήτων των ποσοτικών μεταβλητών έχει το σχήμα «καμπάνας», τότε θεωρούμε ότι τα αντίστοιχα δεδομένα ακολουθούν την κανονική κατανομή ή, με άλλα λόγια, κατανέμονται κανονικά. Το ιστόγραμμα, όμως, δεν είναι «ικανό» να απαντήσει στο ερώτημα αν είναι τα δεδομένα είναι κανονικά ή αν προέρχονται από μία κανονική κατανομή με έναν μέσο και μία διακύμανση. Το SPSS μας επιτρέπει να κατασκευάσουμε δύο διαφορετικά γραφήματα, το **P-P Plot** και το **Q-Q Plot** (επιλέγοντας **Analyze** → **Descriptive Statistics** → **P-P Plots** ή **Q-Q Plots**), προκειμένου να ελέγξουμε οπτικά την ύπαρξη κανονικότητας στα δεδομένα. Όσο πιο κοντά στην ευθεία είναι τα σημεία του σχήματος που θα προκύψει, τόσο περισσότερες είναι οι ενδείξεις ότι τα δεδομένα ακολουθούν την κανονική κατανομή. Όμως, ο οπτικός έλεγχος μπορεί να «πέσει έξω» και να μας κάνει να ξεγελαστούμε. Για τον λόγο αυτό, πραγματοποιούμε έλεγχο κανονικότητας στα δεδομένα μας, προκειμένου να διαπιστώσουμε αν ακολουθούν την κανονική κατανομή.

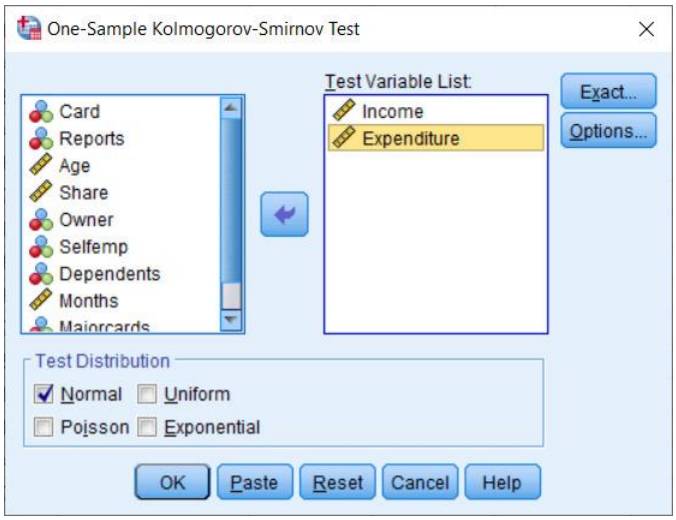
Ο έλεγχος κανονικότητας υπάγεται σε μία ευρύτερη οικογένεια ελέγχων, η οποία ονομάζεται «έλεγχοι υποθέσεων». Οι έλεγχοι υποθέσεων σχετίζονται με τη μηδενική υπόθεση (**Null Hypothesis** ή **Ho**), την εναλλακτική υπόθεση (**Alternative Hypothesis** ή $\mathbf{H_1}$), το επίπεδο στατιστικής σημαντικότητας (***α***) και το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας (***p*-value** ή **Significance**). Η μηδενική και η εναλλακτική υπόθεση είναι της ακόλουθης μορφής:

H_0: Η κατανομή των δεδομένων δεν διαφέρει από την κανονική κατανομή.
H_1: Η κατανομή των δεδομένων διαφέρει από την κανονική κατανομή.

Για τη διεξαγωγή των ελέγχων υποθέσεων χρησιμοποιούνται κάποιοι μαθηματικοί τύποι, οι οποίοι ονομάζονται ελεγχοσυναρτήσεις. Με βάση το αποτέλεσμα που προκύπτει από αυτές, οδηγούμαστε στο συμπέρασμα σχετικά με το αν η μηδενική υπόθεση απορρίπτεται ή όχι. Στη συγκεκριμένη περίπτωση, η μηδενική υπόθεση την οποία θέλουμε να ελέγξουμε είναι ότι τα δεδομένα μας ακολουθούν την κανονική κατανομή ή ότι προέρχονται από έναν πληθυσμό που ακολουθεί την κανονική κατανομή. Η εναλλακτική υπόθεση είναι ότι τα δεδομένα μας δεν ακολουθούν την κανονική κατανομή. Το επίπεδο στατιστικής σημαντικότητας ορίζεται συνήθως ίσο με 0,05 ή 5%. Το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας (*p*-value) ορίζεται ως η πιθανότητα η τιμή του ελέγχου (δηλαδή της ελεγχοσυνάρτησης) να πάρει μία τιμή τόσο ακραία ή περισσότερο ακραία από αυτή που πήρε στο συγκεκριμένο δείγμα υπό τη μηδενική υπόθεση. Αν η *p*-value είναι μικρότερη του 0,05, τότε λέμε ότι η μηδενική υπόθεση απορρίπτεται. Αντιθέτως, αν η *p*-value είναι μεγαλύτερη ή ίση του 0,05, τότε λέμε ότι η μηδενική υπόθεση δεν μπορεί να απορριφθεί.

Τα δεδομένα που θα χρησιμοποιήσουμε σε αυτό το κεφάλαιο είναι τα δεδομένα των πιστωτικών καρτών (**credit.sav**).

- Στο **SPSS**: Στο SPSS, οι τιμές των παρατηρηθέντων επιπέδων στατιστικής σημαντικότητας ονομάζονται (**Asymptotic**) **Significances**. Η κανονικότητα των δεδομένων είναι αναγκαία, για να έχουν ισχύ κάποιες στατιστικές τεχνικές που θα χρησιμοποιήσουμε, όπως οι έλεγχοι υποθέσεων για τους μέσους, η γραμμική παλινδρόμηση, η ανάλυση διακύμανσης κ.ά. Στο SPSS, οι έλεγχοι κανονικότητας διεξάγονται με τον ακόλουθο τρόπο. Επιλέγουμε **Analyze → Non Parametric Tests → Legacy Dialogs → 1-Sample K-S**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 4.1**.



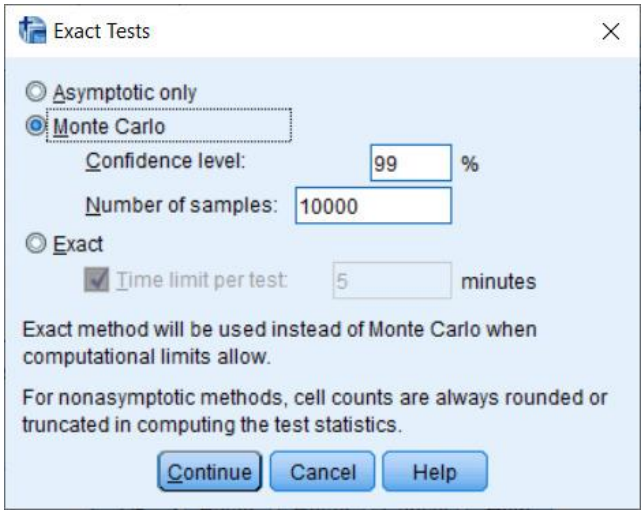
Εικόνα 4.1 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Όπως φαίνεται στο συγκεκριμένο παράθυρο, έχουμε περάσει δύο μεταβλητές ("income" και "expenditure") στο δεξιό κουτάκι, προκειμένου να ελέγξουμε αν οι τιμές τους ακολουθούν την κανονική κατανομή. Παρατηρούμε ότι η επιλογή για τον έλεγχο κανονικότητας (**Normal**) είναι ήδη προεπιλεγμένη από το SPSS. Επιλέγοντας **Options** εμφανίζεται ένα άλλο παράθυρο, στο οποίο μπορούμε να επιλέξουμε επιπλέον την εμφάνιση ενός πίνακα με κάποια περιγραφικά μέτρα που αφορούν τις συγκεκριμένες μεταβλητές. Οπότε, πατώντας **Exact** θα εμφανιστεί το παράθυρο της **Εικόνας 4.2**. Το SPSS έχει ως προεπιλογή το **Asymptotic only**, δηλαδή ότι θα διεξάγει τον έλεγχο κανονικότητας των **Kolmogorov-Smirnov** που έχουμε άλλωστε επιλέξει.

Αν αντί για **Asymptotic only** επιλέξουμε την παρακάτω επιλογή, δηλαδή το **Monte Carlo**, θα ενεργοποιηθούν και τα επόμενα δύο λευκά κουτάκια, το **Confidence level** και το **Number of samples**. Με την επιλογή **Monte Carlo** «ζητάμε» από το SPSS να χρησιμοποιήσει και την τεχνική της προσομοίωσης, για να διεξάγει τον έλεγχο κανονικότητας. Στην παρούσα φάση δεν θα επεκταθούμε περισσότερο στην τεχνική της προσομοίωσης, παρά μόνο θα αναφέρουμε ότι διεξάγει 10.000 (ως προεπιλογή) ελέγχους κανονικότητας και για καθέναν από αυτούς υπολογίζει την *p*-value. Όταν τελειώσει η διαδικασία, εμφανίζεται ο μέσος όρο αυτών των 10.000 *p*-values, καθώς και ένα 99% διάστημα εμπιστοσύνης για τον μέσο όρο αυτών των *p*-values, το οποίο είναι προφανώς βασισμένο στις 10.000 *p*-values. Ο αλγόριθμος bootstrap δεν χρειάζεται στη συγκεκριμένη περίπτωση, ενώ για τα διαστήματα εμπιστοσύνης θα αναφερθούμε εκτενώς στην επόμενη ενότητα. Στη συνέχεια, πατώντας **Continue** επιστρέφουμε στο αρχικό παράθυρο της **Εικόνας 4.1**, ενώ πατώντας **OK** εμφανίζονται τα αποτελέσματα στο Output του SPSS **(Πίνακας 4.1)**.

Η υπόθεση που θέλουμε να ελέγξουμε είναι ότι οι μεταβλητές ακολουθούν την κανονική κατανομή. Στα αποτελέσματά μας εμφανίζεται το μέγεθος του δείγματος (1319) και για τις δύο μεταβλητές, για τις οποίες δεν υπάρχουν εκλιπούσες τιμές.

Στα αποτελέσματα εμφανίζονται, επίσης, ο μέσος και η τυπική απόκλιση για κάθε μεταβλητή. Για τον έλεγχο της κανονικότητας μας ενδιαφέρουν δύο τιμές, η **Asymp. Sig. (2-tailed)** και η **Monte Carlo Sig**. Πρόκειται για τις *p*-values που υπολογίζονται ξεχωριστά από κάθε μέθοδο. Προφανώς, ο έλεγχος των **Kolmogorov-Smirnov** είναι ένας, απλά στη μία περίπτωση η *p*-value υπολογίζεται με βάση τη «συμβατική» μέθοδο, ενώ στη δεύτερη περίπτωση βασίζεται στην τεχνική **Monte Carlo**. Παρατηρούμε ότι, και για τις δύο μεταβλητές, οι *p*-values που υπολογίστηκαν και με τις δύο μεθόδους είναι ίσες με το μηδέν. Όπως αναφέραμε προηγουμένως, αν η *p*-value είναι μικρότερη από το 0,05, τότε απορρίπτουμε την υπόθεση της κανονικότητας των δεδομένων. Άρα, η υπόθεση ότι οι μετρήσεις για το εισόδημα και τις δαπάνες κατανέμονται κανονικά απορρίπτεται σε επίπεδο στατιστικής σημαντικότητας $a = 0{,}05$ ή $a = 5\%$. Οπότε, μπορούμε να συμπεράνουμε ότι υπάρχουν ενδείξεις ότι οι τιμές των μεταβλητών "income" και "expenditure" δεν ακολουθούν την κανονική κατανομή.



Εικόνα 4.2 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

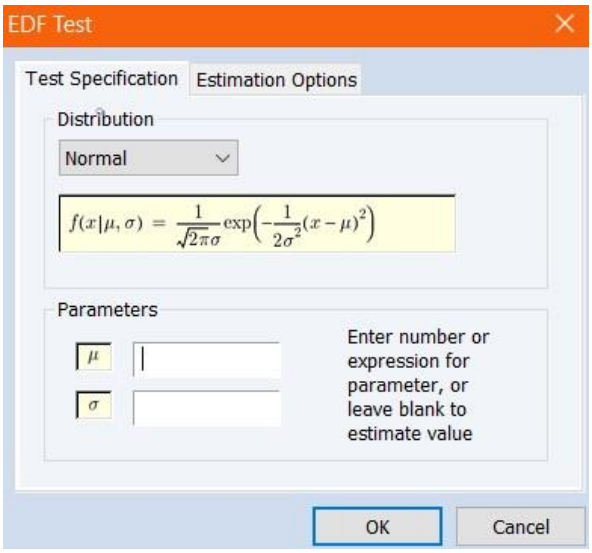
Πίνακας 4.1 *Έλεγχος κανονικότητας.*

One-Sample Kolmogorov-Smirnov Test				
			Income	Expenditure
N			1319	1319
Normal Parameters[a,b]	Mean		3.365376	185.057070778
	Std. Deviation		1.6939017	272.2189174955
Most Extreme Differences	Absolute		.144	.248
	Positive		.144	.175
	Negative		-.130	-.248
Test Statistic			.144	.248
Asymp. Sig. (2-tailed)			.000[c]	.000[c]
Monte Carlo Sig. (2-tailed)	Sig.		.000[d]	.000[d]
	99% Confidence Interval	Lower Bound	.000	.000
		Upper Bound	.000	.000
a. Test distribution is Normal. b. Calculated from data. c. Lilliefors Significance Correction. d. Based on 10000 sampled tables with starting seed 2000000.				

- Στο **Eviews**: Η διαδικασία ελέγχου κανονικότητας για τις τιμές μιας μεταβλητής ακολουθεί την ίδια λογική με αυτή του SPSS. Το Eviews παρέχει μια σειρά από στατιστικές ελέγχου, όπου σε όλες αυτές η μηδενική υπόθεση είναι ότι τα δεδομένα κατανέμονται κανονικά. Έστω, λοιπόν,

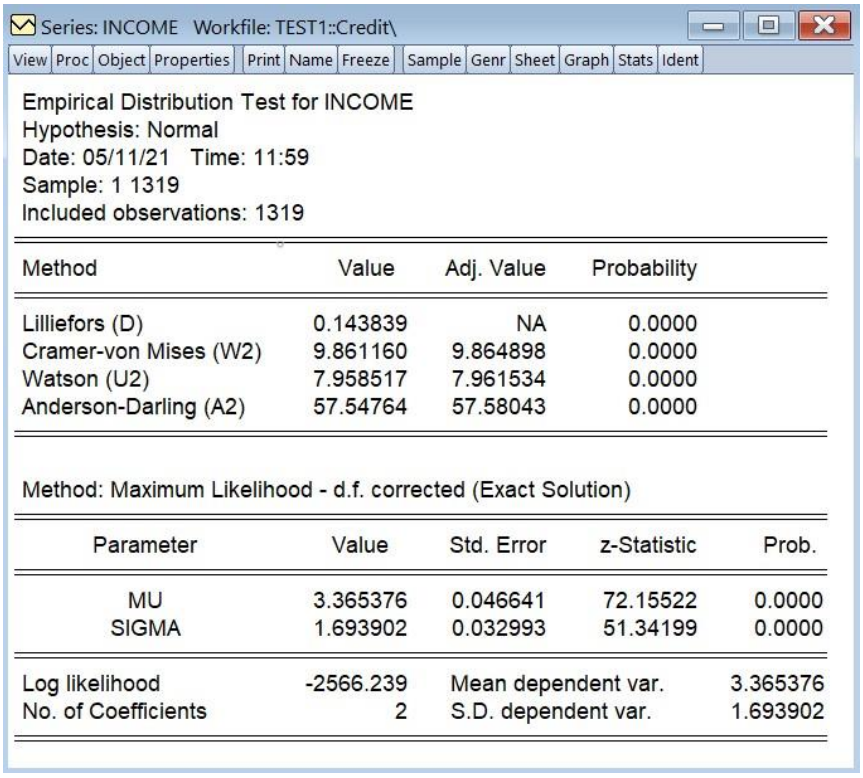
ότι θέλουμε να ελέγξουμε αν οι τιμές της μεταβλητής "income" ακολουθούν την κανονική κατανομή. Αρχικά, στον πίνακα **Stats Table** της **Εικόνας 3.8** εμφανίζονται η τιμή της στατιστικής ελέγχου κανονικότητας Jarque-Bera που είναι ίση με 2140.370 και ακριβώς από κάτω η αντίστοιχη *p*-value που είναι ίση με 0.000. Καθώς η συγκεκριμένη *p*-value είναι μικρότερη από το 0.05, απορρίπτουμε την υπόθεση ότι η μεταβλητή "income" κατανέμεται κανονικά σε επίπεδο στατιστικής σημαντικότητας $a = 0{,}05$ ή $a = 5\%$.

Βεβαίως, το Eviews μας επιτρέπει τον υπολογισμό επιπλέον στατιστικών ελέγχου κανονικότητας για τις τιμές μιας μεταβλητής. Επιλέγοντας **View → Descriptive Statistics & Tests → Empirical Distribution Tests**, εμφανίζεται το παράθυρο της **Εικόνας 4.3**. Επιλέγουμε **Normal** στο "Distribution", ενώ στα 'Parameters' μπορούμε να δώσουμε τιμές για τον μέσο (μ) και την τυπική απόκλιση (σ), αν τις γνωρίζουμε. Αν δεν γνωρίζουμε τις συγκεκριμένες τιμές, όπως συμβαίνει συνήθως, αφήνουμε κενά τα συγκεκριμένα κελιά, επιτρέποντας με τον τρόπο αυτό στο Eviews να τις υπολογίσει από τα δεδομένα μας. Το tab "Estimation Options" μας επιτρέπει να καθορίσουμε τον αριθμό των επαναλήψεων που θα πραγματοποιηθούν, το πόσο ακριβής θα είναι η σύγκλιση των τελικών αποτελεσμάτων, καθώς και τις αρχικές τιμές ("starting values") που θα χρησιμοποιήσει το Eviews. Συνιστάται η χρήση του συγκεκριμένου tab μόνο όταν το Eviews αποτυγχάνει να πραγματοποιήσει τους συγκεκριμένους υπολογισμούς. Οπότε, αφήνουμε το συγκεκριμένο tab ως έχει, πατάμε **OK** και εμφανίζονται τα αποτελέσματά μας **(Εικόνα 4.4)**.



Εικόνα 4.3 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Τα αποτελέσματα αυτά χωρίζονται σε δύο μέρη. Στο πάνω μέρος εμφανίζονται οι στατιστικές ελέγχου κανονικότητας Lilliefors, Cramer-von Mises, Watson και Anderson-Darling. Στη στήλη "Value" παρουσιάζονται οι εκτιμημένες τιμές τους, στη στήλη "Adj. Value" οι προσαρμοσμένες τιμές τους ως προς το μέγεθος του δείγματος και την παραμετρική αβεβαιότητα (δηλαδή, όταν οι παράμετροι του μέσου και της τυπικής απόκλισης έχουν εκτιμηθεί), ενώ στη στήλη "Probability" εμφανίζονται οι αντίστοιχες *p*-values. Όπως φαίνεται στην **Εικόνα 4.4**, όλες οι στατιστικές ελέγχου απορρίπτουν την υπόθεση ότι η μεταβλητή "income" κατανέμεται κανονικά σε επίπεδο στατιστικής σημαντικότητας $a = 5\%$, καθώς σε όλες τις στατιστικές η αντίστοιχη *p*-value είναι μικρότερη από το 0,05.



Εικόνα 4.4 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Στο κάτω μέρος της **Εικόνας 4.4** παρουσιάζονται οι παράμετροι για τον μέσο (MU) και την τυπική απόκλιση (SIGMA) της μεταβλητής "income", οι οποίες χρησιμοποιήθηκαν για τον υπολογισμό της συνάρτησης της θεωρητικής κατανομής. Οι εκτιμημένες τιμές τους, με τη μέθοδο της μεγίστης πιθανοφάνειας, εμφανίζονται στη στήλη "Value", ενώ τα τυπικά τους σφάλματα παρουσιάζονται στη στήλη "Std. Error". Δίνεται, επίσης, η τιμή της *z*-στατιστικής για τις δύο αυτές παραμέτρους, καθώς και οι αντίστοιχες *p-values* τους, οι οποίες βασίζονται στην ασυμπτωτική κανονική κατανομή.

Θα πρέπει να σημειωθεί στο σημείο αυτό πως η στατιστική Lilliefors υπολογίστηκε, γιατί επιτρέψαμε στο Eviews να εκτιμήσει τον μέσο (μ) και την τυπική απόκλιση (σ) της κανονικής κατανομής για τη μεταβλητή "income". Αν, αντιθέτως, δώσουμε εκ των προτέρων συγκεκριμένες τιμές για τις παραμέτρους αυτές, τότε στο παράθυρο της **Εικόνας 4.4** αντί της στατιστικής Lilliefors θα εμφανιστούν οι στατιστικές ελέγχου κανονικότητας Kolmogorov-Smirnov και Kuiper.

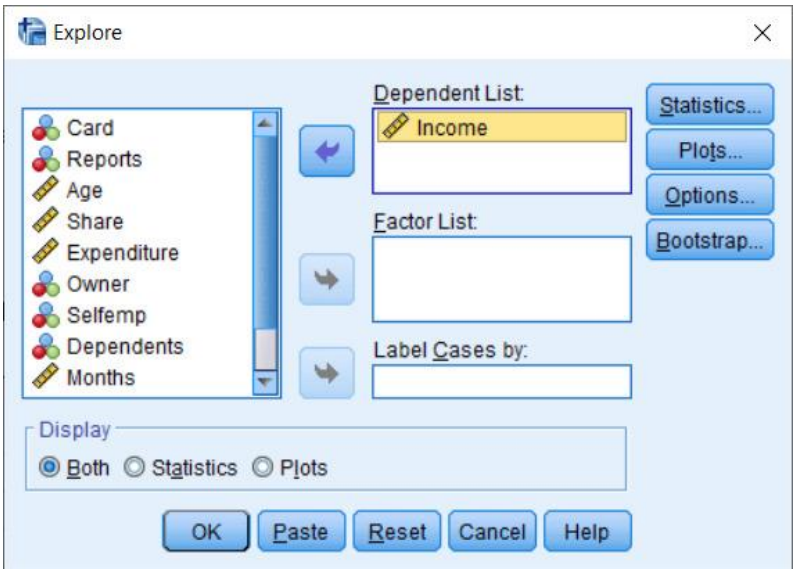
4.2 Διαστήματα εμπιστοσύνης

Πριν αναφερθούμε στη χρησιμότητα των διαστημάτων εμπιστοσύνης και στο πώς αυτά κατασκευάζονται στο SPSS και το Eviews, θα πρέπει να δώσουμε τον σωστό ορισμό τους. Αναλυτικότερα, στην προσπάθειά μας να εκτιμήσουμε την πραγματική τιμή του μέσου, χρησιμοποιούμε τον μέσο ενός δείγματος. Στη συνέχεια κατασκευάζουμε ένα 95% διάστημα εμπιστοσύνης με βάση έναν μαθηματικό τύπο, ο οποίος αφορά ουσιαστικά 2 τυπικά σφάλματα αριστερά και δεξιά της τιμής του μέσου που υπολογίσαμε από το δείγμα. Αν επαναλάβουμε τη δειγματοληψία *ν* φορές, θα εκτιμήσουμε *ν* διαφορετικούς μέσους και προφανώς *ν* διαφορετικά (αν και πολλά από αυτά θα είναι αλληλεπικαλυπτόμενα) διαστήματα εμπιστοσύνης. Ευελπιστούμε ότι στο 95% των *ν* περιπτώσεων, τα διαστήματα εμπιστοσύνης που έχουμε υπολογίσει θα έχουν συμπεριλάβει, περικλείσει, ή «πιάσει» την τιμή του πραγματικού μέσου. Οπότε, αν, για παράδειγμα, «παίρνουμε» κάθε φορά 1.000 δείγματα από έναν πληθυσμό και κατασκευάζουμε 1.000 διαστήματα εμπιστοσύνης για τη μέση τιμή μίας μεταβλητής, ενώ επαναλαμβάνουμε τη διαδικασία άπειρες φορές, στο 95% των περιπτώσεων θα έχουμε φτιάξει διαστήματα εμπιστοσύνης που θα έχουν «πιάσει» τον πραγματικό μέσο του πληθυσμού. Το 95% ονομάζεται βαθμός ή επίπεδο εμπιστοσύνης. Το υπόλοιπο 5% είναι αυτό που έχει ήδη οριστεί ως επίπεδο στατιστικής σημαντικότητας.

- Στο **SPSS**: Για να κατασκευάσουμε ένα 95% διάστημα εμπιστοσύνης για τη μέση τιμή μίας μεταβλητής εργαζόμαστε με τον ακόλουθο τρόπο. Επιλέγουμε **Analyze → Descriptive Statistics → Explore**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 4.5**. Από την επιλογή **Plots** μπορούμε να επιλέξουμε (αν επιθυμούμε) να εμφανιστεί ένα ιστόγραμμα των μεταβλητών που θα περάσουμε στο πάνω λευκό κουτάκι (**Dependent List:**). Στο κάτω αριστερό μέρος, στην επιλογή **Display** επιλέγουμε **Statistics**, καθώς δεν επιθυμούμε την εμφάνιση του ιστογράμματος. Πατώντας **OK** θα προκύψει ένας πίνακας που δίνει πληροφορίες για το δείγμα, καθώς και ο **Πίνακας 4.2.** Στο παράθυρο της **Εικόνας 4.5** υπάρχει στα δεξιά η επιλογή bootstrap. Δεν την επιλέξαμε αυτήν τη φορά για το συγκεκριμένο παράθυρο, το οποίο είναι ίδιο με αυτό της **Εικόνας 3.3**. Όμως, ο χρήστης του SPSS καλείται να το επιλέξει, προκειμένου να εξοικειωθεί με τη χρήση του συγκεκριμένου αλγόριθμου.

Ο **Πίνακας 4.2** παρουσιάζει τα περιγραφικά μέτρα, στα περισσότερα από τα οποία έχουμε ήδη αναφερθεί. Η πρώτη γραμμή του πίνακα περιέχει τη μέση τιμή της μεταβλητής "income". Οι επόμενες δύο τιμές είναι το κάτω και το άνω άκρο του 95% διαστήματος εμπιστοσύνης για τον πραγματικό μέσο του πληθυσμού. Η επόμενη γραμμή αναφέρεται στη μέση τιμή της μεταβλητής "income", από την οποία όμως έχουμε αφαιρέσει το 5% των μεγαλύτερων και των μικρότερων τιμών της. Το Interquartile range (ενδοτεταρτημοριακό εύρος) είναι η διαφορά μεταξύ του τρίτου και του πρώτου τεταρτημόριου. Μέσα στο εύρος αυτό βρίσκεται το 50% των κεντρικών παρατηρήσεων της μεταβλητής "income".

- Στο **Eviews**: Στο συγκεκριμένο πρόγραμμα δεν υπάρχει κάποια αυτοματοποιημένη διαδικασία, προκειμένου να κατασκευαστεί ένα διάστημα εμπιστοσύνης για τη μέση τιμή μίας μεταβλητής. Οπότε, το κάτω και το άνω άκρο του 95% διαστήματος εμπιστοσύνης για τον μέσο του πληθυσμού μιας μεταβλητής μπορούν να υπολογιστούν από τον γνωστό τύπο $\bar{\mu} \pm z\frac{s}{\sqrt{N}}$, όπου $\bar{\mu}$ είναι ο μέσος του δείγματος, z η θεωρητική τιμή της z-στατιστικής ή της t-στατιστικής για επίπεδο στατιστικής σημαντικότητας 95%, s η τυπική απόκλιση του δείγματος και N το μέγεθος του δείγματος. Για τη μεταβλητή "income", από τον πίνακα **Stats Table** της **Εικόνας 3.8** προκύπτει ότι $\bar{\mu} = 3.365376$, $\mathrm{s} = 1.693902$, $\mathrm{N} = 1319$ και $z = 1.96$. Συνεπώς, το κάτω και το άνω άκρο του 95% διαστήματος εμπιστοσύνης για τον μέσο του πληθυσμού της συγκεκριμένης μεταβλητής είναι 3.273960 και 3.456792, αντίστοιχα.



Εικόνα 4.5 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Πίνακας 4.2 *Περιγραφικά μέτρα με 95% διάστημα εμπιστοσύνης, για το εισόδημα.*

Descriptives				
			Statistic	Std. Error
Income	Mean		3.365376	.0466408
	95% Confidence Interval for Mean	Lower Bound	3.273878	
		Upper Bound	3.456874	
	5% Trimmed Mean		3.181872	
	Median		2.900000	
	Variance		2.869	
	Std. Deviation		1.6939017	
	Minimum		.2100	
	Maximum		13.5000	
	Range		13.2900	
	Interquartile Range		1.7625	
	Skewness		1.928	.067
	Kurtosis		4.933	.135

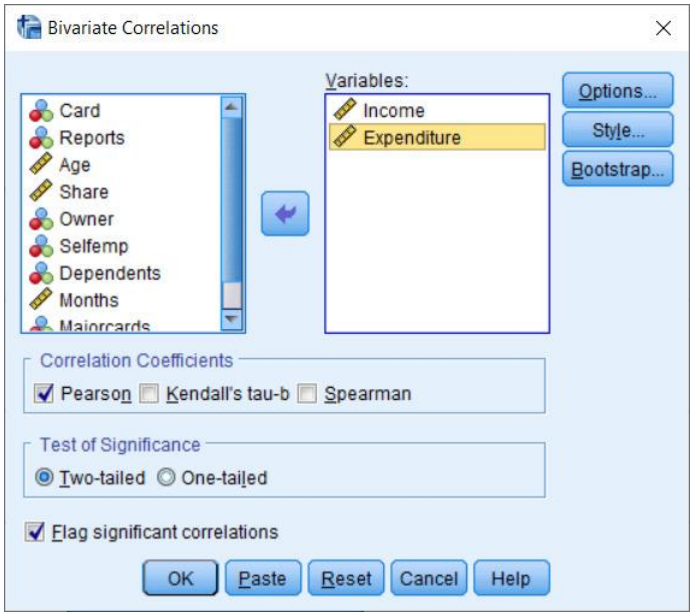
4.3 Συντελεστές γραμμικής συσχέτισης

Πριν προχωρήσουμε στη διεξαγωγή των ελέγχων υποθέσεων για τους μέσους των μεταβλητών, είναι απαραίτητο να αναφερθούμε στη γραμμική συσχέτιση μεταξύ δύο ποσοτικών μεταβλητών. Οι συντελεστές που θα παρουσιαστούν στη συνέχεια αναφέρονται στη *γραμμική* συσχέτιση που μπορεί να συνδέει τις δύο μεταβλητές. Οι τιμές που μπορεί να πάρει ένας συντελεστής συσχέτισης είναι από -1 έως +1. Αρνητικές τιμές του συντελεστή γραμμικής συσχέτισης μεταξύ δύο μεταβλητών υποδηλώνουν αρνητική γραμμική συσχέτιση. Δηλαδή, οι μεγαλύτερες τιμές της μίας μεταβλητής τείνουν να αντιστοιχούν στις μικρότερες τιμές της άλλης μεταβλητής. Θετικές τιμές του συντελεστή γραμμικής συσχέτισης αποτελούν ένδειξη θετικής γραμμικής συσχέτισης μεταξύ των δύο μεταβλητών. Δηλαδή, οι μεγαλύτερες τιμές της μίας μεταβλητής τείνουν να αντιστοιχούν στις μεγαλύτερες τιμές της άλλης μεταβλητής. Αν οι τιμές του συντελεστή γραμμικής συσχέτισης βρίσκονται κοντά στο μηδέν, αυτό αποτελεί ένδειξη ότι δεν υπάρχει στατιστικά σημαντική γραμμική συσχέτιση μεταξύ των δύο μεταβλητών. Όσο πιο μεγάλες είναι οι τιμές του συντελεστή συσχέτισης ή, με άλλα λόγια, όσο πιο κοντά βρίσκονται στη μονάδα (*σε απόλυτη τιμή πάντα*), τόσο πιο ισχυρή είναι η γραμμική συσχέτιση μεταξύ των αντίστοιχων μεταβλητών. Οι πιο γνωστοί συντελεστές γραμμικής συσχέτισης είναι οι συντελεστές του **Pearson**, του **Spearman** και του **Kendall**. Η μηδενική και η εναλλακτική υπόθεση για τον συγκεκριμένο έλεγχο είναι οι ακόλουθες:

H_0: $\rho = 0$ (δεν υπάρχει γραμμική συσχέτιση μεταξύ των δύο μεταβλητών).
H_1: $\rho \neq 0$ (υπάρχει γραμμική συσχέτιση μεταξύ των δύο μεταβλητών).

Ο συντελεστής συσχέτισης του Pearson υποθέτει κανονικότητα των δεδομένων, ενώ οι άλλοι δύο συντελεστές συσχέτισης (Spearman και Kendall) δεν υποθέτουν κανονικότητα των δεδομένων. Βέβαια, όταν πρόκειται για μεγάλα δείγματα, δηλαδή για δείγματα μεγέθους 30 παρατηρήσεων και πάνω, η στατιστική θεωρία δείχνει ότι, όσο μεγαλώνει το μέγεθος του δείγματος, οι τιμές των συγκεκριμένων συντελεστών «πλησιάζουν» η μία την άλλη. Οπότε, για τα δεδομένα μας που αφορούν τις πιστωτικές κάρτες, μπορούμε να χρησιμοποιήσουμε οποιονδήποτε συντελεστή συσχέτισης, ανεξαρτήτως της κατανομής των δεδομένων. Η βασική διαφορά μεταξύ των συγκεκριμένων συντελεστών είναι ότι ο συντελεστής του Pearson υπολογίζεται με βάση τα δεδομένα, ενώ οι άλλοι δύο υπολογίζονται με βάση τις τάξεις μεγέθους των δεδομένων. Ειδικότερα, ο συντελεστής του Spearman είναι, στην ουσία, ο συντελεστής του Pearson υπολογισμένος για τις τάξεις μεγέθους των δεδομένων. Το γεγονός, λοιπόν, ότι οι συντελεστές του Spearman και του Kendall υπολογίζονται με βάση τις τάξεις μεγέθους των δεδομένων είναι αυτό που επιτρέπει την ελευθερία σχετικά με τη μη ικανοποίηση της κανονικότητας των μεταβλητών. Παρόλα αυτά και ασχέτως από το ζήτημα της κανονικότητας, ο συντελεστής του Pearson είναι προτιμητέος.

- Στο **SPSS**: Για να υπολογίσουμε τους τρεις αυτούς συντελεστές συσχέτισης, επιλέγουμε **Analyze → Correlate → Bivariate**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 4.6**.



Εικόνα 4.6 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Στο δεξιό κουτάκι πρέπει να περάσουμε τουλάχιστον δύο μεταβλητές, καθώς οι συντελεστές συσχέτισης υπολογίζονται για ζεύγη μεταβλητών. Επίσης, αν περάσουμε περισσότερες από δύο μεταβλητές, τότε το SPSS θα υπολογίσει τους συντελεστές γραμμικής συσχέτισης για όλα τα πιθανά ζεύγη των μεταβλητών. Όπως φαίνεται στην **Εικόνα 4.6**, μόνο ο συντελεστής του Pearson είναι επιλεγμένος. Αν θέλουμε να εμφανιστούν και οι άλλοι δύο συντελεστές, «τσεκάρουμε» τις αντίστοιχες επιλογές. Παρατηρήστε, επίσης, ότι στο κάτω αριστερό μέρος του παραθύρου είναι «τσεκαρισμένη» η επιλογή **Flag significant correlations**. Τέλος, η επιλογή **Options** μας δίνει τη δυνατότητα της εμφάνισης των μέσων, των τυπικών αποκλίσεων, καθώς και του πλήθους των τιμών για καθεμία μεταβλητή. Πατώντας **OK**, προκύπτουν τα αποτελέσματα **(Πίνακας 4.3)**.

Θα πρέπει να επισημάνουμε ότι αυτήν τη φορά έχουμε επιλέξει την εφαρμογή του αλγόριθμου bootstrap, ο οποίος μας έδωσε μία εκτίμηση της μεροληψίας. Υποθέτοντας ότι η τιμή του δείγματος είναι η πραγματική τιμή, τότε η μέση τιμή των 999 bootstrap τιμών για τον συντελεστή συσχέτισης αποτελεί μία εκτίμηση για τη δειγματική τιμή του συντελεστή. Η διαφορά τους θα είναι μία εκτίμηση της μεροληψίας. Βεβαίως, όπως αναφέραμε και προηγουμένως, στην περίπτωση που έχουμε μεγάλα δείγματα, όπως στο παράδειγμά μας, ο αλγόριθμος bootstrap δεν θα μας δώσει διαφορετικά αποτελέσματα.

Πίνακας 4.3 *Συντελεστής συσχέτισης του Pearson.*

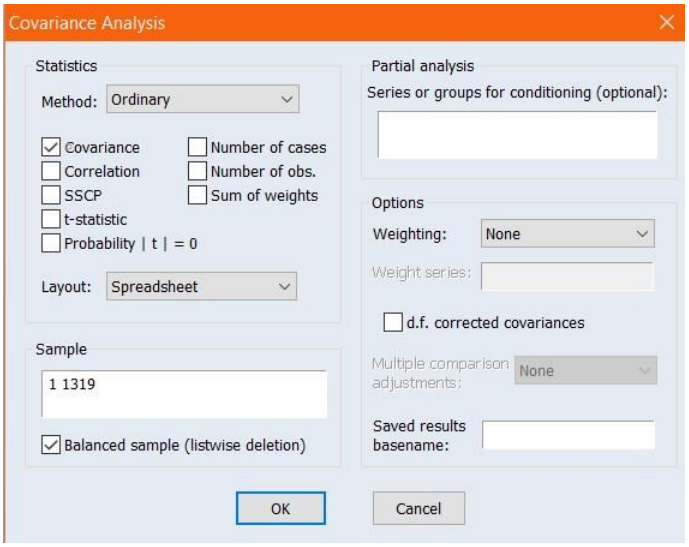
Correlations			
		Income	Expenditure
Income	Pearson Correlation	1	.281**
	Sig. (2-tailed)		.000
	N	1319	1319
Expenditure	Pearson Correlation	.281**	1
	Sig. (2-tailed)	.000	
	N	1319	1319
**. Correlation is significant at the 0.01 level (2-tailed).			

Όπως φαίνεται στον **Πίνακα 4.3,** υπάρχουν δύο αστεράκια. Αυτά προκύπτουν μέσω της επιλογής **Flag significant correlations**. Οι συντελεστές συσχέτισης που υπολογίστηκαν για το ζεύγος των μεταβλητών "income" και "expenditure" ανίχνευσαν στατιστικά σημαντική συσχέτιση μεταξύ τους. Κάτω από την τιμή του συντελεστή συσχέτισης εμφανίζεται μία υπολογισμένη *p*-value (**Sig. (2-tailed**). Η συγκεκριμένη *p*-value αναφέρεται στον έλεγχο της υπόθεσης ότι στο συγκεκριμένο ζεύγος μεταβλητών δεν υπάρχει γραμμική συσχέτιση (δηλαδή, ότι ο συντελεστής συσχέτισης για το συγκεκριμένο ζεύγος είναι ίσος με το μηδέν).

Αφού το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας είναι μικρότερο του 0,05, μπορούμε να συμπεράνουμε ότι η συγκεκριμένη υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας $a = 0{,}05$. Οπότε, υπάρχει στατιστικά σημαντική γραμμική συσχέτιση μεταξύ των μεταβλητών "income" και "expenditure". Στην περίπτωση που η *p*-value είναι μικρότερη του 0,01, τότε ο συντελεστής γραμμικής συσχέτισης εμφανίζεται με δύο αστεράκια αντί για ένα. Ενδεικτικό είναι το μήνυμα που υπάρχει στο κάτω μέρος του **Πίνακα 4.3** που εξηγεί τι σημαίνουν τα δύο αστεράκια, καθώς στο παράδειγμά μας η *p*-value είναι μικρότερη του 0,01.

Στο σημείο αυτό είναι χρήσιμο να αναφερθεί ότι ο συντελεστής του Kendall μπορεί να χρησιμοποιηθεί και στην περίπτωση που έχουμε κατηγορικές μεταβλητές. Στην περίπτωση όμως αυτή, θα πρέπει οι συγκεκριμένες μεταβλητές να είναι υποχρεωτικά σε κλίμακα διάταξης. Δηλαδή, να είναι διατακτικές κατηγορικές μεταβλητές. Επίσης, θα επαναλάβουμε ότι με τους συντελεστές γραμμικής συσχέτισης ελέγχουμε αν σε ένα ζεύγος μεταβλητών υπάρχει γραμμική συσχέτιση και μόνο. Δηλαδή, οι συγκεκριμένοι συντελεστές δεν μπορούν να ανιχνεύσουν συσχέτιση μη γραμμικής φύσεως για το συγκεκριμένο ζεύγος μεταβλητών. Οπότε, θα πρέπει να δίνεται μεγάλη προσοχή στην ερμηνεία τους. Τέλος, πρέπει να υπενθυμίσουμε ότι η λογική με την οποία απορρίπτουμε ή όχι μία μηδενική υπόθεση είναι πάντα η ίδια. Δηλαδή, αν το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας είναι μικρότερο του 0,05, τότε η συγκεκριμένη μηδενική υπόθεση απορρίπτεται πάντα.

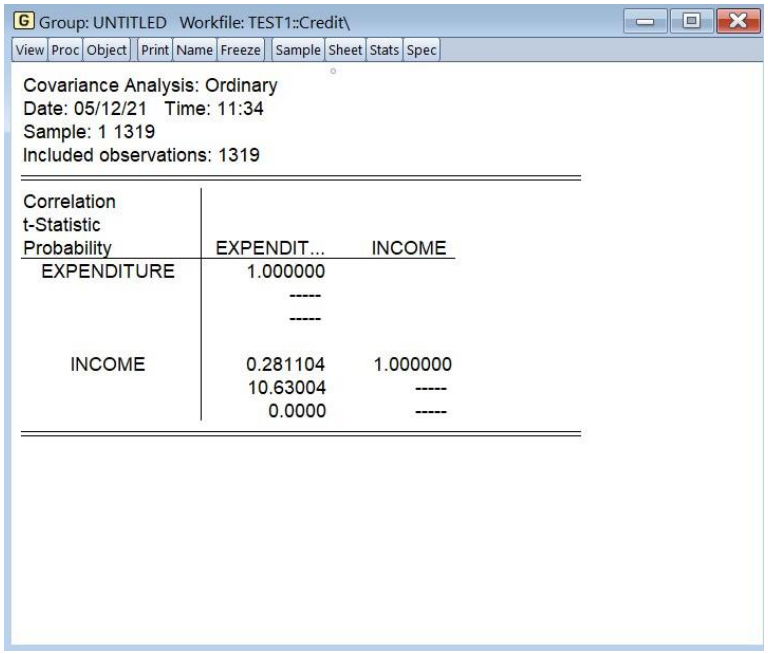
- Στο **Eviews**: Στο συγκεκριμένο πρόγραμμα, οι συντελεστές του **Pearson**, του **Spearman** και του **Kendall** υπολογίζονται με πολύ εύκολο τρόπο. Έστω, λοιπόν, ότι θέλουμε να υπολογίσουμε τους συγκεκριμένους συντελεστές για τις μεταβλητές "expenditure" και "income", προκειμένου να διαπιστώσουμε αν υπάρχει γραμμική συσχέτιση μεταξύ τους. Αρχικά, «ανοίγουμε» τις δύο αυτές μεταβλητές ως Group και στη συνέχεια επιλέγουμε **View → Covariance Analysis**, οπότε θα εμφανιστεί ένα παράθυρο, όπως αυτό της **Εικόνας 4.7**. Από το μενού **Method:** μας δίνεται η δυνατότητα να επιλέξουμε μεταξύ 4 διαφορετικών συντελεστών συσχέτισης: Ordinary (που είναι ο συντελεστής του Pearson), Ordinary (uncentered), Spearman rank-order και Kendall's tau. Στη συνέχεια, εμφανίζεται μια σειρά από κουτάκια που μας επιτρέπουν να επιλέξουμε ποιες πληροφορίες θα εμφανιστούν στα τελικά μας αποτελέσματα. Τα κουτάκια αυτά αφορούν συνδιακύμανση (covariance), συντελεστή συσχέτισης (correlation), *t*-στατιστική, *p*-value (probability | t | = 0), αριθμό παρατηρήσεων (number of obs.) κλπ. Όπως θα δούμε στη συνέχεια, τα κουτάκια αυτά διαφοροποιούνται πλήρως στην περίπτωση του συντελεστή του Kendall. Επίσης, το μενού **Layout:** μας επιτρέπει να επιλέξουμε πώς θα εμφανιστούν τα αποτελέσματά μας (φύλλο εργασίας, απλός πίνακας, πολλαπλοί πίνακες, λίστα), ενώ στη συνέχεια εμφανίζεται το δείγμα μας. Το κουτάκι **Balanced sample (listwise deletion)** πρέπει να είναι πάντα «τσεκαρισμένο», προκειμένου όλες οι στατιστικές να υπολογιστούν χρησιμοποιώντας τον ίδιο αριθμό παρατηρήσεων και να αποφευχθούν προβλήματα που σχετίζονται με ελλείπουσες τιμές (missing values). Οι επιλογές που εμφανίζονται στο δεξί μέρος του παραθύρου της **Εικόνας 4.7** δεν μας αφορούν προς το παρόν, οπότε δεν θα τις αναλύσουμε στη συγκεκριμένη ενότητα. Τέλος, το κουτάκι **Saved results basename:** μας επιτρέπει να αποθηκεύσουμε τα αποτελέσματά μας στο Eviews workfile. Εφόσον δώσουμε κάποιο όνομα, τα αποτελέσματά μας θα αποθηκευτούν με τη μορφή μητρών.



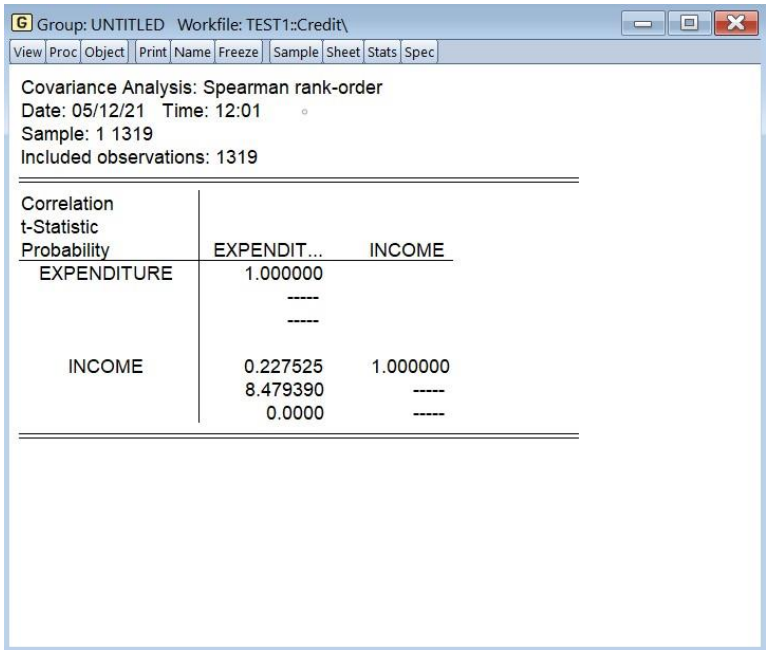
Εικόνα 4.7 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Προκειμένου, λοιπόν, να υπολογίσουμε τον συντελεστή συσχέτισης του Pearson, μαζί με την *t*-στατιστική του και την αντίστοιχη *p*-value, στο παράθυρο της **Εικόνας 4.7** επιλέγουμε **Ordinary** στο μενού **Method:** και «τσεκάρουμε» τα κουτάκια **Correlation**, **t-statistic** και **Probability | t | = 0**. Επίσης, αποεπιλέγουμε το κουτάκι Covariance. Προσέξτε στο σημείο αυτό ότι το μενού **Layout:** άλλαξε αυτόματα σε **Single table**. Πατώντας **OK** εμφανίζονται τα αποτελέσματα **(Εικόνα 4.8)**. Στο πάνω μέρος του πίνακα της **Εικόνας 4.8** παρουσιάζεται η σειρά που εμφανίζονται τα αποτελέσματα. Πιο συγκεκριμένα, ο συντελεστής γραμμικής συσχέτισης του Pearson είναι 0.281104, η *t*-στατιστική του είναι 10.63004 και η αντίστοιχη *p*-value είναι 0.0000. Με βάση τα αποτελέσματα αυτά, η μηδενική υπόθεση ότι δεν υπάρχει γραμμική συσχέτιση μεταξύ των μεταβλητών “expenditure” και “income” απορρίπτεται σε επίπεδο στατιστικής σημαντικότητας $a = 5\%$. Συνεπώς, υπάρχει στατιστικά σημαντική γραμμική συσχέτιση μεταξύ των δύο αυτών μεταβλητών.

Για να υπολογίσουμε τον συντελεστή γραμμική συσχέτισης του Spearman, ακολουθούμε ακριβώς την ίδια διαδικασία με τον συντελεστή του Pearson, με μόνη διαφορά ότι επιλέγουμε **Spearman rank-order** στο μενού **Method:**. Τα αποτελέσματα εμφανίζονται στη **Εικόνα 4.9**. Όπως φαίνεται στη συγκεκριμένη εικόνα, ο συντελεστής γραμμικής συσχέτισης του Spearman είναι 0.227525, η *t*-στατιστική του είναι 8.479390 και η αντίστοιχη *p*-value είναι 0.0000. Οπότε, και με τον συγκεκριμένο συντελεστή, η μηδενική υπόθεση ότι δεν υπάρχει γραμμική συσχέτιση μεταξύ των μεταβλητών “expenditure” και “income” απορρίπτεται σε επίπεδο στατιστικής σημαντικότητας $a = 5\%$.



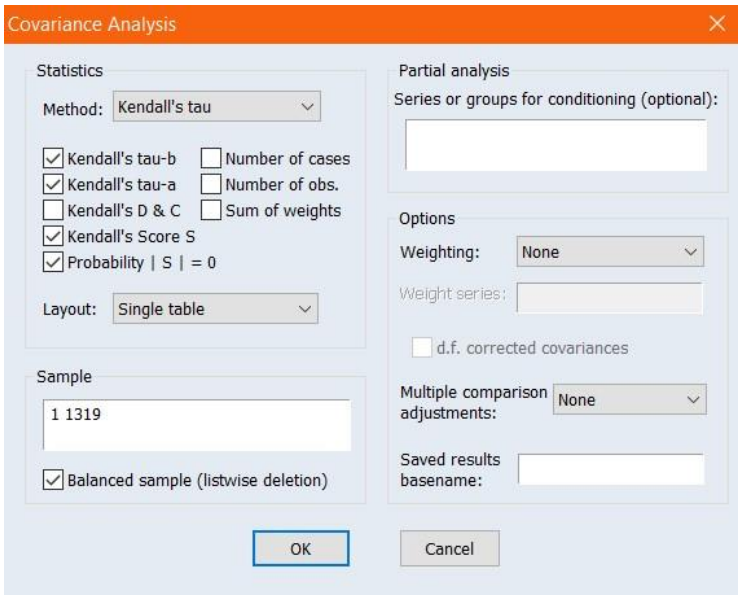
Εικόνα 4.8 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



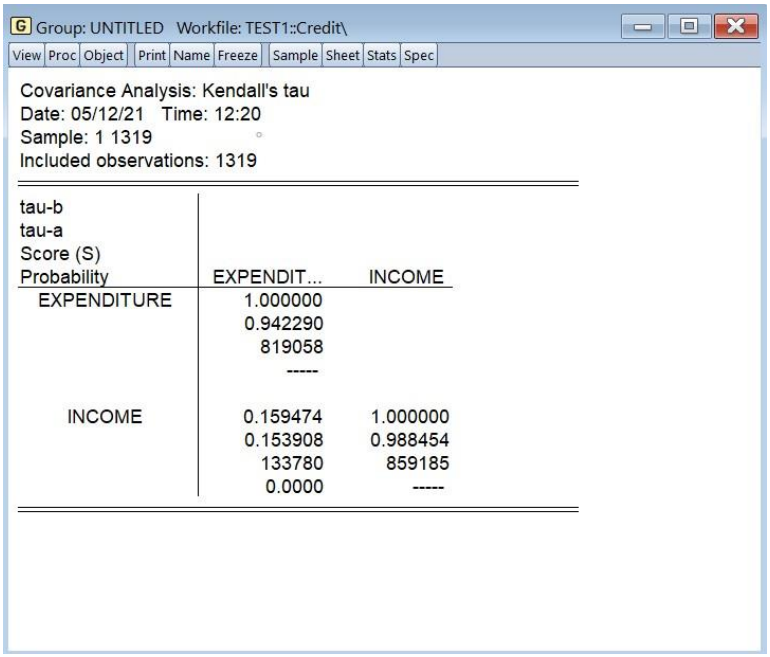
Εικόνα 4.9 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Στην περίπτωση του συντελεστή γραμμικής συσχέτισης του Kendall, τα πράγματα είναι λίγο διαφορετικά. Όπως αναφέραμε προηγουμένως, αν επιλέξουμε **Kendall's tau** στο μενού **Method:**, τα κουτάκια που εμφανίζονται είναι πλήρως διαφορετικά από τις περιπτώσεις των συντελεστών συσχέτισης του Pearson και του Spearman. Στην περίπτωση αυτή, «τσεκάρουμε» τα κουτάκια **Kendall's tau-b**, **Kendall's tau-a, Kendall's Score S** και **Probability | S | = 0 (Εικόνα 4.10)**. Και στην περίπτωση αυτή, το μενού **Layout:** άλλαξε αυτόματα σε **Single table**. Τα αποτελέσματα του συγκεκριμένου ελέγχου εμφανίζονται στην **Εικόνα 4.11**. Όπως φαίνεται στη συγκεκριμένη εικόνα, για τη μηδενική υπόθεση ότι η στατιστική Score (S) του Kendall είναι μηδέν, η αντίστοιχη *p*-value είναι 0.0000. Οπότε, όπως και προηγουμένως, η μηδενική υπόθεση ότι

δεν υπάρχει γραμμική συσχέτιση μεταξύ των μεταβλητών “expenditure” και “income” απορρίπτεται σε επίπεδο στατιστικής σημαντικότητας $a = 5\%$.



Εικόνα 4.10 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Εικόνα 4.11 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

4.4 Έλεγχοι υποθέσεων για τον μέσο ενός πληθυσμού

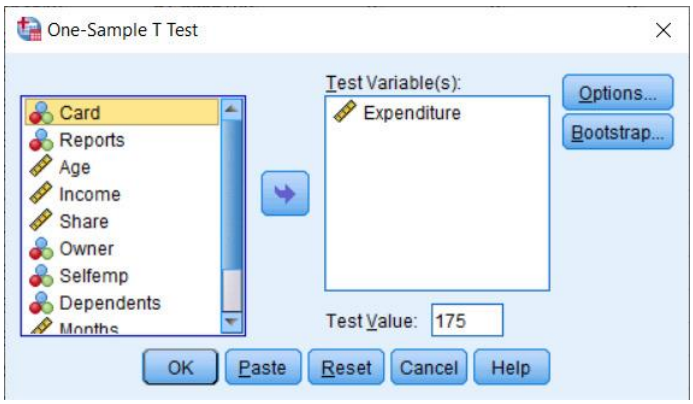
Ας υποθέσουμε στη συνέχεια, ότι έχουμε συλλέξει ένα δείγμα από παρατηρήσεις που αφορούν τις μετρήσεις μίας μεταβλητής. Όπως και προηγουμένως, θα χρησιμοποιήσουμε τα δεδομένα των πιστωτικών καρτών και συγκεκριμένα τη μεταβλητή που αφορά τη δαπάνη (“expenditure”). Όπως φαίνεται στον **Πίνακα 4.1**, ο μέσος της συγκεκριμένης μεταβλητής είναι 185.0571. Έστω, λοιπόν, ότι ενδιαφερόμαστε να ελέγξουμε την υπόθεση ότι η μέση δαπάνη του πληθυσμού είναι ίση με 175. Δηλαδή, οι υποθέσεις (μηδενική και εναλλακτική) θα έχουν την ακόλουθη μορφή:

$$\mathbf{H_0}: \boldsymbol{\mu} = \mathbf{175}$$
$$\mathbf{H_1}: \boldsymbol{\mu} \neq \mathbf{175}$$

Βασική προϋπόθεση για τη διεξαγωγή του ελέγχου *t*, είναι αυτή της κανονικότητας των δεδομένων. Δυστυχώς, όμως, ο έλεγχος κανονικότητας των Kolmogorov-Smirnov απέρριψε την υπόθεση της κανονικότητας για τις μετρήσεις της δαπάνης. Παρόλα αυτά, το Κεντρικό Οριακό Θεώρημα μας βοηθάει να ξεπεράσουμε το πρόβλημα της παραβίασης της κανονικότητας. Το συγκεκριμένο θεώρημα αναφέρει ότι, καθώς το μέγεθος του δείγματος τείνει στο άπειρο, η κατανομή του δειγματικού μέσου τείνει στην κανονική κατανομή. Με βάση τη θεωρία, ένα δείγμα της τάξης των 30 παρατηρήσεων είναι ικανοποιητικό, προκειμένου να εξασφαλιστεί η ισχύς του θεωρήματος. Προφανώς, η ισχύς του Κεντρικού Οριακού Θεωρήματος εξασφαλίζεται στο παράδειγμά μας, καθώς το δείγμα της μεταβλητής "expenditure" αποτελείται από 1319 παρατηρήσεις.

- Στο **SPSS**: Προκειμένου να διεξάγουμε τον έλεγχο *t*, επιλέγουμε **Analyze → Compare Means → One-Sample T-test**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 4.12**. Περνάμε τη μεταβλητή "expenditure" στο λευκό κουτάκι με την ένδειξη **Test Variable(s)** και στη συνέχεια, μεταφερόμαστε στο λευκό κουτάκι με την ένδειξη **Test Value**. Στο συγκεκριμένο κουτάκι θα διαγράψουμε το μηδέν και θα τοποθετήσουμε την τιμή που θέλουμε να ελέγξουμε (δηλαδή την τιμή της μηδενικής υπόθεσης). Στο παράδειγμά μας, η τιμή αυτή είναι ίση με 175.[2] Πατάμε **OK** και το αποτέλεσμα του ελέγχου εμφανίζεται στον **Πίνακα 4.4.** Θα πρέπει να επισημάνουμε στο σημείο αυτό πως ο πρώτος πίνακας που εμφανίζεται στο Output του SPSS έχει παραλειφθεί, καθώς περιέχει πληροφορίες που δεν είναι ιδιαίτερα σημαντικές για την παρούσα ενότητα (τιμή του μέσου, τιμή της τυπικής απόκλισης, τυπικό σφάλμα του μέσου και μέγεθος του δείγματος).

Η πρώτη στήλη του **Πίνακα 4.4** αναφέρεται στην υπό εξέταση μεταβλητή, ενώ η δεύτερη περιέχει μία τιμή. Όπως αναφέραμε σε προηγούμενη παράγραφο, οι στατιστικοί έλεγχοι χρησιμοποιούν συγκεκριμένους μαθηματικούς τύπους. Η συγκεκριμένη τιμή 1.342 είναι η τιμή του ελέγχου *t* για τη μεταβλητή "expenditure", η οποία χρησιμοποιήθηκε, προκειμένου να υπολογιστεί η *p*-value (Sig. (2-tailed)). Στην πρώτη γραμμή του πίνακα εμφανίζεται η τιμή του μέσου που ελέγξαμε μέσω του συγκεκριμένου στατιστικού ελέγχου. Η στήλη δίπλα από την *p*-value παρουσιάζει τη διαφορά ανάμεσα στην τιμή της μηδενικής υπόθεσης και στη μέση τιμή της μεταβλητής (10.057), ενώ οι δύο επόμενες στήλες περιέχουν ένα 95% διάστημα εμπιστοσύνης για τη διαφορά αυτή. Ουσιαστικά, αν η συγκεκριμένη διαφορά βρίσκεται εκτός των ορίων του διαστήματος εμπιστοσύνης, απορρίπτουμε τη μηδενική υπόθεση.



Εικόνα 4.12 *Reprint Courtesy of International Busines Machines Corporation, © International Business Machines Corporation.*

[2] Δεν έχουμε επιλέξει το bootstrap, διότι δεν είμαστε σίγουροι για το τι ακριβώς κάνει σε αυτήν την περίπτωση (ελέγχους υποθέσεων γενικά).

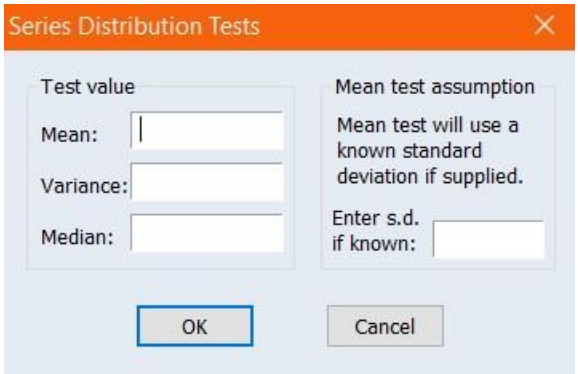
Το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας είναι ίσο με 0.180. Εφόσον είναι μεγαλύτερο του 0.05, μπορούμε να συμπεράνουμε ότι δεν υπάρχουν ενδείξεις απόρριψης της μηδενικής υπόθεσης ότι ο πραγματικός μέσος είναι ίσος με 175. Διαφορετικά, μπορούμε να πούμε ότι η διαφορά του μέσου από την τιμή της μηδενικής υπόθεσης δεν είναι στατιστικά σημαντική, δηλαδή ο μέσος του δείγματός μας δεν διαφέρει στατιστικά σημαντικά από την τιμή που ελέγξαμε και η οποία είναι ίση με 175.

Η τρίτη στήλη του **Πίνακα 4.4** ονομάζεται **df** και συμβολίζει τους βαθμούς ελευθερίας μίας κατανομής (**degrees of freedom**). Στον συγκεκριμένο έλεγχο, ο μαθηματικός τύπος (δηλαδή η ελεγχοσυνάρτηση) δεν βασίστηκε στην κανονική κατανομή, αλλά στην κατανομή *t*. Αυτή η κατανομή (όπως και κάποιες άλλες) έχουν τους βαθμούς ελευθερίας ως βασικό χαρακτηριστικό. Με τον όρο «βαθμοί ελευθερίας» ορίζουμε τον μέγιστο αριθμό μεταβλητών, οι οποίες μπορούν να προσδιοριστούν αυθαίρετα υπό κάποιες συνθήκες έτσι ώστε να ισχύουν οι συνθήκες αυτές. Με άλλα λόγια, για μία δεδομένη τιμή του μέσου σε ένα δείγμα μεγέθους *v* μπορούμε να επιλέξουμε αυθαίρετα *v-1* τιμές, όμως η *v*-οστή τιμή που θα επιλέξουμε θα είναι τέτοια, ώστε να καταλήξουμε στον ήδη γνωστό μέσο.

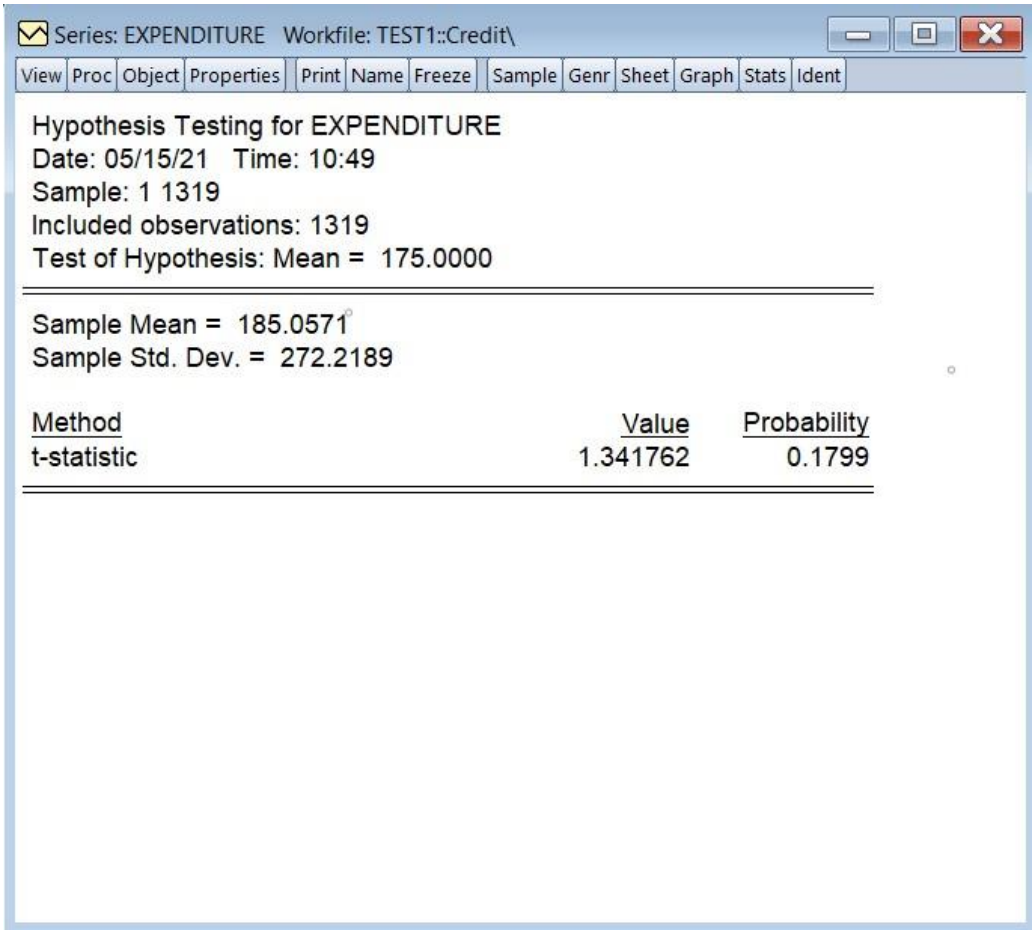
Πίνακας 4.4 *Αποτελέσματα ελέγχου t.*

One-Sample Test						
	Test Value = 175					
	t	Df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference Lower	Upper
Expenditure	1.342	1318	.180	10.0570707784	-4.647183367	24.761324924

- Στο **Eviews**: Στο συγκεκριμένο πρόγραμμα, η διαδικασία ελέγχου για τον μέσο του πληθυσμού είναι αρκετά απλή. Θα χρησιμοποιήσουμε και πάλι τη μεταβλητή "expenditure", η οποία, όπως έχουμε ήδη αναφέρει, αφορά τη μέση μηνιαία δαπάνη με τη χρήση πιστωτικής κάρτας**.** Προκειμένου να ελέγξουμε την υπόθεση ότι η μέση δαπάνη του πληθυσμού είναι ίση με 175, «ανοίγουμε» με διπλό κλικ τη συγκεκριμένη μεταβλητή και επιλέγουμε **View → Descriptive Statistics & Tests → Simple Hypothesis Tests**. Το αποτέλεσμα θα είναι να εμφανιστεί ένα παράθυρο όπως αυτό της **Εικόνας 4.13**, όπου στο πεδίο **Mean:** συμπληρώνουμε 175. Τα πεδία **Variance:** και **Median:** τα αφήνουμε κενά, καθώς αφορούν αντίστοιχους ελέγχους για τη διακύμανση και τη διάμεσο, αντίστοιχα. Επίσης, αφήνουμε κενό το πεδίο **Enter s.d. if known:**, το οποίο μας επιτρέπει να ορίσουμε την τιμή της τυπικής απόκλισης, αν αυτή είναι γνωστή. Πατώντας **OK** εμφανίζονται τα αποτελέσματα **(Εικόνα 4.14)**, όπου αρχικά παρουσιάζονται ο μέσος και η τυπική απόκλιση του δείγματος, στη συνέχεια η τιμή της *t*-στατιστικής που είναι ίση με 1.341762, καθώς και η αντίστοιχη *p*-value που είναι ίση με 0.1799. Οπότε, με βάση τα συγκεκριμένα αποτελέσματα, μπορούμε να συμπεράνουμε πως η μηδενική υπόθεση ότι η μέση δαπάνη του πληθυσμού είναι ίση με 175 δεν μπορεί να απορριφθεί σε επίπεδο στατιστικής σημαντικότητας $a = 5\%$.



Εικόνα 4.13 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Εικόνα 4.14 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

4.5 Έλεγχος υποθέσεων για τη διαφορά των μέσων δύο ανεξάρτητων δειγμάτων

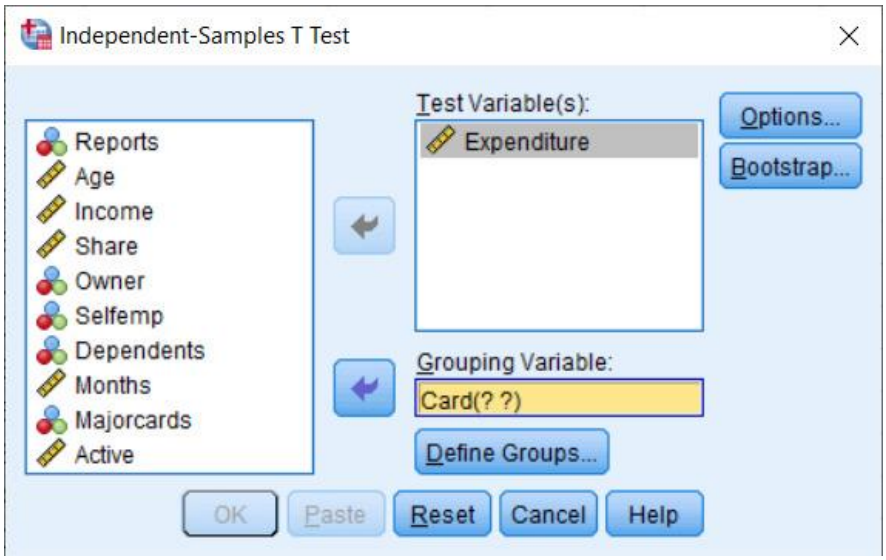
Ο έλεγχος *t* μπορεί, επίσης, να εφαρμοστεί και για τον έλεγχο ισότητας μεταξύ των μέσων δύο πληθυσμών, οι οποίοι είναι ανεξάρτητοι (ή υποθέτουμε ότι είναι ανεξάρτητοι). Έστω, λοιπόν, ότι θέλουμε να ελέγξουμε αν η μέση δαπάνη των ατόμων που έχουν πιστωτική κάρτα είναι ίση με τη μέση δαπάνη των ατόμων που δεν έχουν πιστωτική κάρτα. Στα δεδομένα μας, το 1 δηλώνει τις τιμές της μεταβλητής που προέρχονται από το πρώτο δείγμα (δηλαδή, τα άτομα που έχουν πιστωτική κάρτα) και το 0 τις τιμές που προέρχονται από το δεύτερο δείγμα (δηλαδή, τα άτομα που δεν έχουν πιστωτική κάρτα). Τα μεγέθη των δύο δειγμάτων δεν είναι απαραίτητο να είναι ίσα. Ουσιαστικά, αυτό που θέλουμε να ελέγξουμε είναι αν οι μέσοι των δύο πληθυσμών από τους οποίους προέρχονται τα αντίστοιχα δείγματα διαφέρουν. Οι υποθέσεις (μηδενική και εναλλακτική) θα έχουν την ακόλουθη μορφή:

$$\mathbf{H_0}: \boldsymbol{\mu_1} = \boldsymbol{\mu_2}$$
$$\mathbf{H_1}: \boldsymbol{\mu_1} \neq \boldsymbol{\mu_2}$$

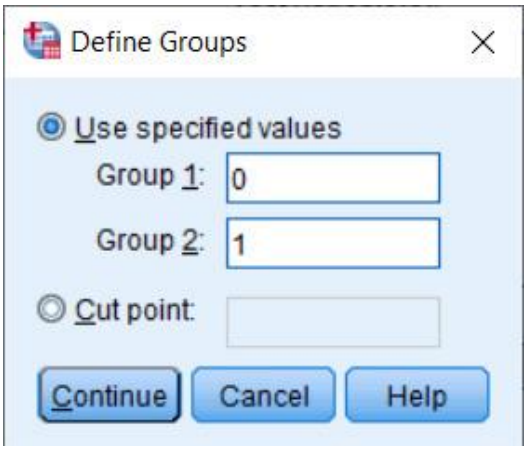
όπου $\boldsymbol{\mu_1}$ είναι ο μέσος του πληθυσμού του πρώτου δείγματος και $\boldsymbol{\mu_2}$ ο μέσος του πληθυσμού του δεύτερου δείγματος.

- Στο **SPSS**: Προκειμένου να πραγματοποιήσουμε τον συγκεκριμένο (παραμετρικό) έλεγχο *t*, επιλέγουμε **Analyze → Compare Means → Independent-Samples T Test**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 4.15**. Περνάμε τη μεταβλητή "expenditure", η οποία περιέχει τις μετρήσεις και για τα δύο δείγματα, στο λευκό κουτί κάτω από την ένδειξη **Test variable(s):**, και τη μεταβλητή "card", η οποία δηλώνει σε ποιο δείγμα ανήκουν οι παρατηρήσεις, στο λευκό κουτί κάτω από την ένδειξη **Grouping Variable:**. Ωστόσο, δεν

μπορούμε ακόμα να πατήσουμε το **OK**. Ο λόγος ότι πρέπει να «κάνουμε» το SPSS να «αντιληφθεί» ότι η μεταβλητή που περιέχει την απόφαση σχετικά με την αίτηση χορήγησης πιστωτικής κάρτας (“card”) θα χρησιμοποιηθεί για τον διαχωρισμό της μέσης δαπάνης (“expenditure”) ανάλογα με τον αν η συγκεκριμένη αίτηση έγινε αποδεκτή ή απορρίφθηκε Αυτό γίνεται επιλέγοντας **Define Groups**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 4.16**.



Εικόνα 4.15 *Reprint Courtesy of International Busines Machine Corporation, © International Business Machines Corporation.*



Εικόνα 4.16 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Καθώς θέλουμε να ελέγξουμε αν υπάρχουν διαφορές στις μέσες δαπάνες των ατόμων ανάλογα με το αν έχουν πιστωτική κάρτα (1) ή όχι (0), θα πρέπει να δηλώσουμε ότι στη μεταβλητή “card” η τιμή 1 αναφέρεται στο πρώτο δείγμα και η τιμή 0 αναφέρεται στο δεύτερο δείγμα. Πατάμε **Continue** προκειμένου να επιστρέψουμε στο αρχικό παράθυρο της **Εικόνας 4.15**, και στη συνέχεια πατάμε **OK**, για να εμφανιστούν τα αποτελέσματα στο Output του SPSS. Τα αποτελέσματα αυτά περιέχονται σε δύο πίνακες, εκ των οποίων παραθέτουμε μόνο τον δεύτερο **(Πίνακας 4.5)**, καθώς ο πρώτος περιέχει κάποια στατιστικά μέτρα για τα δύο δείγματα, τα οποία δεν έχουν ιδιαίτερη σημασία για την παρούσα ανάλυση.

Πίνακας 4.5 *Αποτελέσματα ελέγχου t.*

Independent Samples Test										
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	T	Df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Expenditure	Equal variances assumed	211.001	.000	-14.264	1317	.000	-238.602	16.727	-271.417	-205.787
	Equal variances not assumed			-26.525	1022	.000	-238.602	8.995	-256.253	-220.951

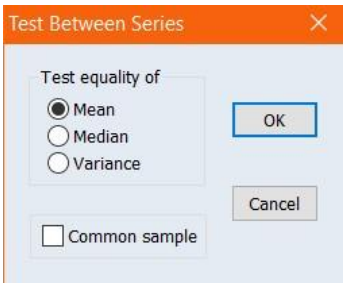
Ο παραπάνω έλεγχος *t* έχει δύο εναλλακτικές κατευθύνσεις: στην πρώτη υποθέτει ότι οι διακυμάνσεις των δύο δειγμάτων είναι περίπου ίσες, ενώ στη δεύτερη υποθέτει ότι οι διακυμάνσεις των δύο δειγμάτων δεν είναι ίσες. Ο **Πίνακας 4.5** παρουσιάζει δύο γραμμές αποτελεσμάτων. Η πρώτη αναφέρεται στην περίπτωση όπου γίνεται η υπόθεση της ισότητας των δύο διακυμάνσεων, ενώ η δεύτερη αναφέρεται στην περίπτωση όπου δεν γίνεται η συγκεκριμένη υπόθεση. Ο **Πίνακας 4.5** είναι χωρισμένος σε δύο κατηγορίες αποτελεσμάτων. Η μία αφορά τον έλεγχο του **Levene** για την ισότητα των διακυμάνσεων, ενώ η άλλη παρουσιάζει τα αποτελέσματα του ελέγχου *t* που έχουμε επιλέξει να κάνουμε.

Ο έλεγχος του Levene εξετάζει την υπόθεση της ισότητας των δύο διακυμάνσεων και υπολογίζει μία *p*-value. Αν η συγκεκριμένη *p*-value είναι μικρότερη του 0,05, η υπόθεση της ισότητας των διακυμάνσεων απορρίπτεται, ενώ, αν είναι μεγαλύτερη του 0,05, η παραπάνω υπόθεση δεν απορρίπτεται. Όπως φαίνεται στον **Πίνακα 4.5**, η *p*-value για τον έλεγχο της ισότητας των δύο μέσων είναι ίση με μηδέν (Sig. (2-tailed)). Οπότε, η μηδενική υπόθεση απορρίπτεται, δηλαδή οι μέσοι των δύο πληθυσμών από τα οποία προήλθαν τα δύο δείγματα διαφέρουν στατιστικά σημαντικά σε επίπεδο σημαντικότητας $a = 5\%$. Θα πρέπει να τονιστεί στο σημείο αυτό ότι, όπως προκύπτει από τις προσομοιώσεις των Tsagris et al. (2020), είναι προτιμότερο να επιλέγουμε τη δεύτερη γραμμή αποτελεσμάτων (όπου δεν γίνεται η υπόθεση της ισότητας των δύο διακυμάνσεων) ανεξαρτήτως του αποτελέσματος που προκύπτει από τον έλεγχο του Levene.

- Στο **Eviews**: Η διαδικασία ελέγχου για την ισότητα των μέσων δύο ανεξάρτητων δειγμάτων είναι σχετικά απλή. Έστω, λοιπόν, ότι θέλουμε να ελέγξουμε αν το ετήσιο εισόδημα "income" των ατόμων που έχουν πιστωτική κάρτα είναι ίσο με το ετήσιο εισόδημα των ατόμων που δεν έχουν πιστωτική κάρτα. Όπως και στην περίπτωση του SPSS, τα μεγέθη των δύο δειγμάτων δεν είναι απαραίτητο να είναι ίσα. Αρχικά, πρέπει να δημιουργήσουμε δύο νέες μεταβλητές με τον τρόπο που έχουμε ήδη περιγράψει παραπάνω. Η πρώτη θα περιλαμβάνει τις παρατηρήσεις της μεταβλητής "income" για αυτούς των οποίων η αίτηση για πιστωτική κάρτα έγινε δεκτή, και άρα η μεταβλητή "card" παίρνει την τιμή 1 (έστω "incone"), ενώ η δεύτερη τις παρατηρήσεις της μεταβλητής "income" για αυτούς των οποίων η αίτηση δεν έγινε δεκτή, και άρα η μεταβλητή "card" παίρνει την τιμή 0 (έστω "inczero"). Στο παράθυρο "Generate Series by Equation" **(Εικόνα 3.29),** για την πρώτη μεταβλητή γράφουμε στο πάνω κουτί **incone=income** και στο κάτω κουτί **if card=1** και πατάμε **OK**, ενώ για τη δεύτερη μεταβλητή γράφουμε στο πάνω κουτί **inczero=income** και στο κάτω κουτί **if card=0** και πατάμε **OK**. Οι μεταβλητές έχουν πλέον δημιουργηθεί.

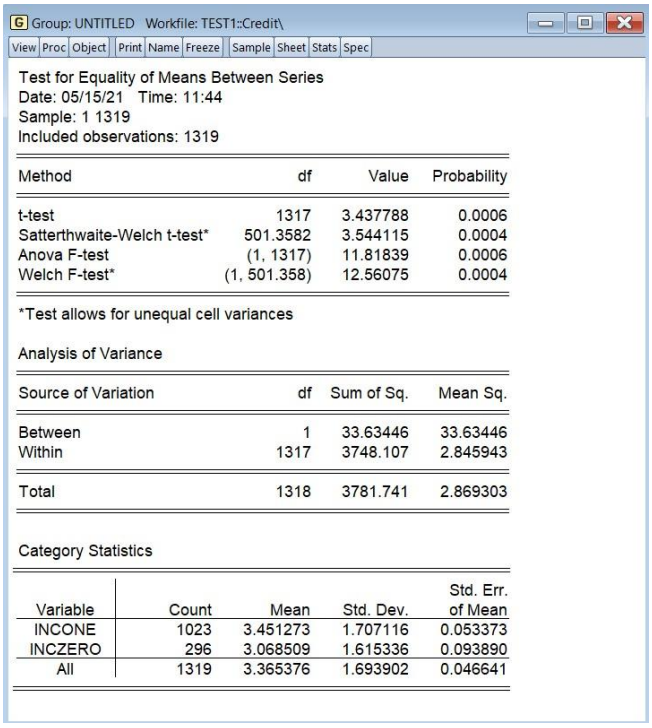
Προκειμένου, λοιπόν, να ελέγξουμε την υπόθεση ότι οι μέσοι των δύο αυτών μεταβλητών είναι ίσοι, τις «ανοίγουμε» ως Group και στη συνέχεια επιλέγουμε **View → Tests of Equality**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 4.17**, στο οποίο θα επιλέξουμε **Mean**. Προσέξτε ότι στο κάτω μέρος

του συγκεκριμένου παραθύρου υπάρχει η επιλογή **Common sample**. Η επιλογή αυτή έχει σημασία μόνο όταν τα δύο δείγματα έχουν διαφορετικό αριθμό παρατηρήσεων. Αν συμβαίνει αυτό και «τσεκάρουμε» τη συγκεκριμένη επιλογή, το Eviews θα υπολογίσει τις στατιστικές ελέγχου μόνο για τις παρατηρήσεις εκείνες που υπάρχουν στα δείγματα και των δύο μεταβλητών. Διαφορετικά, θα υπολογιστούν οι στατιστικές ελέγχου χρησιμοποιώντας το σύνολο των παρατηρήσεων για καθένα από τα δύο δείγματα. Στο παράδειγμά μας, όπου οι μεταβλητές "incone" και "inczero" δεν έχουν τον ίδιο αριθμό παρατηρήσεων, δεν «τσεκάρουμε» τη συγκεκριμένη επιλογή, καθώς θέλουμε να χρησιμοποιηθεί το σύνολο των παρατηρήσεων στις δύο αυτές μεταβλητές.



Εικόνα 4.17 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Πατώντας **OK** εμφανίζονται τα αποτελέσματα του συγκεκριμένου ελέγχου **(Εικόνα 4.18)**. Στο πάνω μέρος της **Εικόνας 4.18** εμφανίζονται τα αποτελέσματα των στατιστικών ελέγχου. Το Eviews παρέχει 4 εναλλακτικές στατιστικές: 2 *t*-tests (το απλό και το **Satterthwaite-Welch**) και 2 *F*-tests (το **Anova** και το **Welch**). Στη στήλη **df** εμφανίζονται οι βαθμοί ελευθερίας για καθεμία από τις τέσσερις στατιστικές ελέγχου, στην επόμενη στήλη (**Value**) παρουσιάζονται οι εκτιμημένες τιμές για τις συγκεκριμένες στατιστικές, ενώ στην τελευταία στήλη (**Probability**) εμφανίζονται οι αντίστοιχες *p*-values. Κάτω από τις στατιστικές ελέγχου παρουσιάζονται κάποιες επιπλέον πληροφορίες σχετικά με την ανάλυση της διακύμανσης και τις υπό εξέταση μεταβλητές, τις οποίες θα προσπεράσουμε καθώς δεν παρουσιάζουν ιδιαίτερο ενδιαφέρον στην παρούσα στιγμή. Όπως προκύπτει από τα αποτελέσματα της **Εικόνας 4.18**, η *p*-value είναι μικρότερη του 0,05 σε όλες τις στατιστικές ελέγχου. Αυτό σημαίνει πως η μηδενική υπόθεση ότι οι μέσοι των μεταβλητών "incone" και "inczero" είναι ίσοι απορρίπτεται σε επίπεδο στατιστικής σημαντικότητας $a = 5\%$.


Group: UNTITLED Workfile: TEST1::Credit\

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Test for Equality of Means Between Series
Date: 05/15/21 Time: 11:44
Sample: 1 1319
Included observations: 1319

Method	df	Value	Probability
t-test	1317	3.437788	0.0006
Satterthwaite-Welch t-test*	501.3582	3.544115	0.0004
Anova F-test	(1, 1317)	11.81839	0.0006
Welch F-test*	(1, 501.358)	12.56075	0.0004

*Test allows for unequal cell variances

Analysis of Variance

Source of Variation	df	Sum of Sq.	Mean Sq.
Between	1	33.63446	33.63446
Within	1317	3748.107	2.845943
Total	1318	3781.741	2.869303

Category Statistics

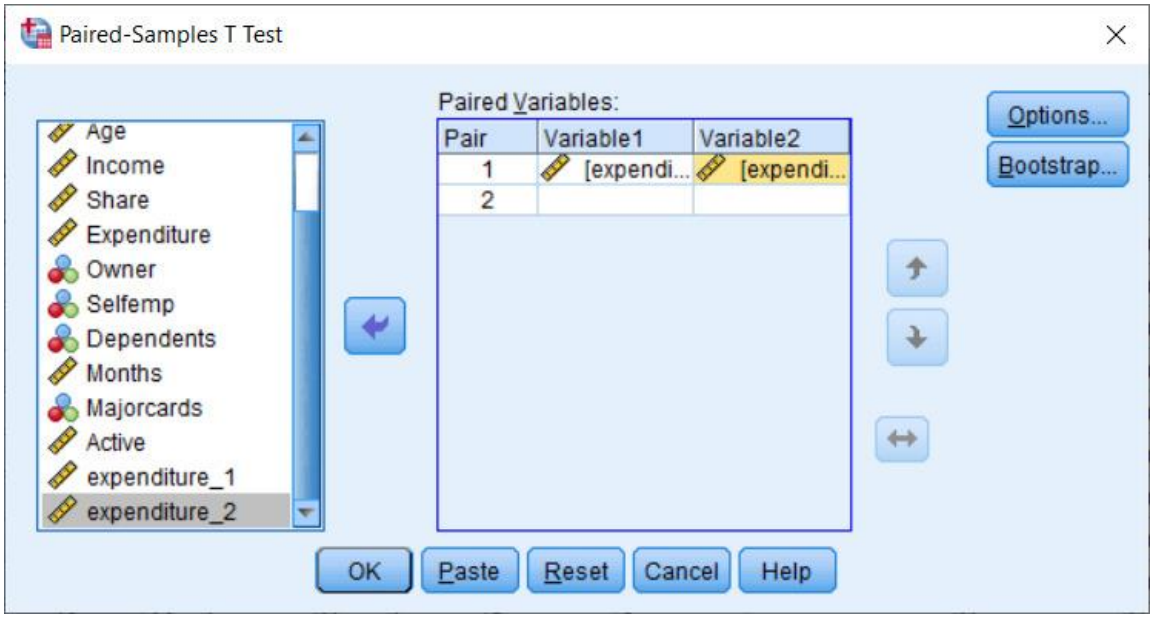
Variable	Count	Mean	Std. Dev.	Std. Err. of Mean
INCONE	1023	3.451273	1.707116	0.053373
INCZERO	296	3.068509	1.615336	0.093890
All	1319	3.365376	1.693902	0.046641

Εικόνα 4.18 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

4.6 Έλεγχος υπόθεσης για τη διαφορά των μέσων δύο εξαρτημένων πληθυσμών

Αναλύσαμε προηγουμένως την περίπτωση του ελέγχου υποθέσεων, όταν έχουμε δύο ανεξάρτητα δείγματα. Όμως, τι γίνεται στην περίπτωση που τα δύο δείγματα δεν είναι, ή δεν μπορούμε να υποθέσουμε ότι είναι, ανεξάρτητα; Η απάντηση στο ερώτημα αυτό δίνεται από τον έλεγχο *t* για ζεύγη παρατηρήσεων. Ένα κλασικό παράδειγμα εφαρμογής του ελέγχου αυτού αφορά την ιατρική, στην περίπτωση που έχουμε μετρήσεις για κάποια άτομα πριν και μετά από μία δίαιτα, και ενδιαφερόμαστε να εξετάσουμε κατά πόσο η συγκεκριμένη δίαιτα ήταν αποτελεσματική ή όχι. Στην περίπτωση αυτή, το βάρος των ατόμων μετριέται και τις δύο φορές, οπότε είναι σαφές ότι τα μεγέθη των δύο δειγμάτων πρέπει απαραίτητα να είναι ίσα. Μία άλλη περίπτωση εφαρμογής του συγκεκριμένου ελέγχου είναι όταν εξετάζουμε αδέρφια ως προς ένα ποσοτικό χαρακτηριστικό. Η υπόθεση της εξάρτησης των τιμών του χαρακτηριστικού ανάμεσα στα αδέρφια είναι εύλογη. Και στον συγκεκριμένο έλεγχο, η μηδενική και η εναλλακτική υπόθεση είναι ίδιες με αυτές της προηγούμενης ενότητας.

- Στο **SPSS**: Για τον συγκεκριμένο έλεγχο *t* εργαζόμαστε με τον ακόλουθο τρόπο. Αρχικά, πρέπει να φτιάξουμε τα δεδομένα. Οι τιμές των δύο δειγμάτων πρέπει να καταχωριστούν σε δύο στήλες, όπου η μία θα βρίσκεται δίπλα από την άλλη. Για λόγους παρουσίασης και μόνο, χρησιμοποιήσαμε 30 τιμές δαπανών και φτιάξαμε δύο νέες στήλες, στις οποίες τοποθετήσαμε 15 τιμές στην καθεμία. Επιλέγοντας **Analyze → Compare Means → Paired-Samples T Test**, θα εμφανιστεί το παράθυρο της **Εικόνας 4.19**. Στο συγκεκριμένο παράθυρο πρέπει πρώτα να επιλέξουμε το ζεύγος μεταβλητών εκείνων, τους μέσους των οποίων θα χρησιμοποιήσουμε, προκειμένου να διεξάγουμε τον συγκεκριμένο έλεγχο και να καταλήξουμε σε συμπεράσματα. Στη συνέχεια, περνάμε το ζεύγος μεταβλητών που έχουμε επιλέξει στο δεξιό κουτί κάτω από την ένδειξη **Paired Variables:** και πατάμε **OK**. Τα αποτελέσματα θα εμφανιστούν στο Output του SPSS σε τρεις πίνακες, από τους οποίους παρουσιάζουμε τους δύο τελευταίους. Ο πρώτος πίνακας περιέχει το μέγεθος των δειγμάτων, τους μέσους, τις τυπικές αποκλίσεις των μέσων και τις τυπικές αποκλίσεις κάθε δείγματος, ενώ ο δεύτερος και ο τρίτος (**Πίνακες 4.6** και **4.7**, αντίστοιχα) παρουσιάζουν ιδιαίτερο ενδιαφέρον.



Εικόνα 4.19 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Όπως φαίνεται στον **Πίνακα 4.6,** ο συντελεστής γραμμικής συσχέτισης που υπολογίστηκε μεταξύ των δύο μεταβλητών έχει μικρή και στατιστικά μη σημαντική τιμή. Το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας για τον έλεγχο της μηδενικής υπόθεσης ότι δεν υπάρχει γραμμική συσχέτιση μεταξύ των δύο δειγμάτων είναι ίσο με 0,798, γεγονός που ακριβώς σημαίνει ότι ο αντίστοιχος συντελεστής γραμμικής συσχέτισης είναι στατιστικά μη σημαντικός. Οπότε, με βάση τα αποτελέσματα αυτά, μπορούμε να συμπεράνουμε ότι δεν υπάρχει γραμμική συσχέτιση μεταξύ των τιμών των δύο δειγμάτων.

Πίνακας 4.6 *Συντελεστής συσχέτισης του Pearson μεταξύ των δύο μετρήσεων.*

Paired Samples Correlations				
		N	Correlation	Sig.
Pair 1	expenditure_1 & expenditure_2	15	-.072	.798

Σχετικά με τον **Πίνακα 4.7**, η τελευταία στήλη του περιέχει το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας, το οποίο είναι 0,461. Καθώς είναι μεγαλύτερο του 0,05, μπορούμε να συμπεράνουμε ότι οι δύο μέσοι δεν έχουν στατιστικά σημαντική διαφορά μεταξύ τους.

Πίνακας 4.7 *Αποτελέσματα ελέγχου t για το ζεύγος τιμών.*

Paired Samples Test									
		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	T	Df	Sig.(2-tailed)
Pair 1	expenditure_1 - expenditure_2	-92.2692	471.3749	121.708	-353.307	168.7694	-.758	14	.461

- Στο **Eviews**: Στο συγκεκριμένο πρόγραμμα δεν υπάρχει κάποια αυτοματοποιημένη διαδικασία, προκειμένου να ελεγχθεί στατιστικά η διαφορά των μέσων δύο εξαρτημένων πληθυσμών. Οπότε, η συγκεκριμένη διαδικασία θα πρέπει να γίνει βήμα-βήμα. Χρησιμοποιούμε και πάλι 30 παρατηρήσεις της μεταβλητής "expenditure" (έστω τις 1-30) και φτιάχνουμε 2 νέες σειρές με 15 παρατηρήσεις η καθεμία (ακολουθώντας τον τρόπο που περιγράψαμε στην ενότητα 2.10). Η πρώτη σειρά θα περιέχει τις παρατηρήσεις 1-15 και η δεύτερη σειρά τις παρατηρήσεις 16-30. Προκειμένου να υπολογίσουμε τον συντελεστή γραμμικής συσχέτισης του Pearson για τις 2 αυτές σειρές, ακολουθούμε τη διαδικασία που περιγράψαμε στην ενότητα 4.3. Από τις εκτιμήσεις που πραγματοποιήσαμε, προκύπτει ότι η τιμή του συντελεστή γραμμικής συσχέτισης του Pearson είναι -0,072199, η τιμή της t-στατιστικής είναι -0,260998 και η αντίστοιχη *p*-value είναι 0,7982. Καθώς η συγκεκριμένη *p*-value είναι μεγαλύτερη του 0,05, η μηδενική υπόθεση ότι ο συντελεστής γραμμικής συσχέτισης είναι στατιστικά ίσος με μηδέν δεν μπορεί να απορριφθεί σε επίπεδο στατιστικής σημαντικότητας $a = 5\%$. Τέλος, προκειμένου να ελέγξουμε αν οι μέσοι των δύο αυτών σειρών είναι ίσοι, ακολουθούμε τη διαδικασία που περιγράψαμε στην ενότητα 4.5. Από τις εκτιμήσεις μας προέκυψε ότι η τιμή της t-στατιστικής είναι -0,780963 και η αντίστοιχη *p*-value 0,4414 > 0,05. Οπότε, η μηδενική υπόθεση ότι οι μέσοι των 2 σειρών είναι ίσοι δεν μπορεί να απορριφθεί σε επίπεδο στατιστικής σημαντικότητας $a = 5\%$.

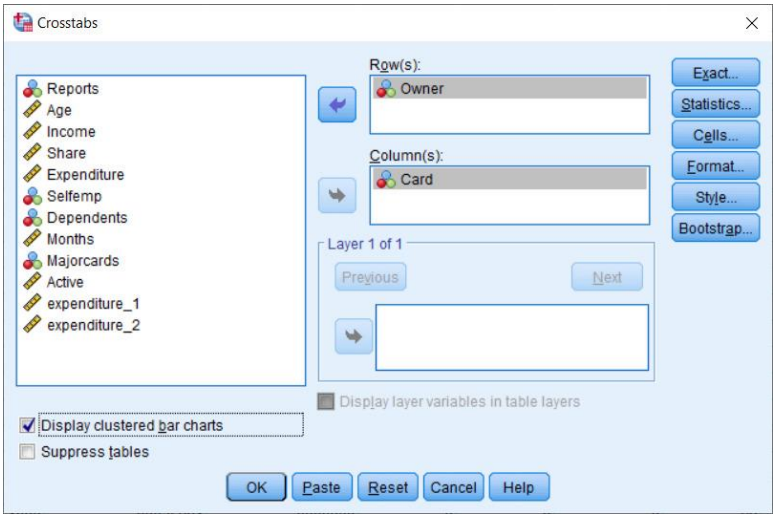
4.7 χ^2 και G^2 έλεγχοι ανεξαρτησίας για κατηγορικές μεταβλητές

Αναλύσαμε σε προηγούμενη ενότητα πώς υπολογίζεται ο συντελεστής γραμμικής συσχέτισης στην περίπτωση των ποσοτικών μεταβλητών. Στην περίπτωση, όμως, που έχουμε κατηγορικές μεταβλητές χρησιμοποιούμε διαφορετικούς ελέγχους και συγκεκριμένα τους ελέγχους ανεξαρτησίας χ^2 ή G^2 (likelihood ratio). Η απαιτούμενη κλίμακα μέτρησης των κατηγορικών μεταβλητών είναι η ονομαστική, παρόλο που και μεταβλητές με διατακτική κλίμακα είναι δυνατόν να χρησιμοποιηθούν. Οι δύο προαναφερθέντες έλεγχοι ανεξαρτησίας χρησιμοποιούνται για τον έλεγχο της μηδενικής υπόθεσης ότι δύο κατηγορικές μεταβλητές είναι ανεξάρτητες μεταξύ τους. Προφανώς, οι κατηγορικές μεταβλητές μπορούν να έχουν αρκετά επίπεδα (τιμές ή κατηγορίες), είναι όμως απαραίτητο η κάθε μεταβλητή να έχει τουλάχιστον δύο επίπεδα. Όπως θα δούμε στη συνέχεια, όταν διεξάγουμε τον συγκεκριμένο έλεγχο ανεξαρτησίας είτε με το SPSS είτε με το Eviews, εμφανίζεται ένας επιπλέον πίνακας που περιέχει τις συχνότητες εμφάνισης όλων των δυνατών συνδυασμών ζευγών των επιπέδων των κατηγορικών μεταβλητών. Στους συγκεκριμένους ελέγχους ανεξαρτησίας, οι υποθέσεις (μηδενική και εναλλακτική) έχουν την ακόλουθη μορφή:

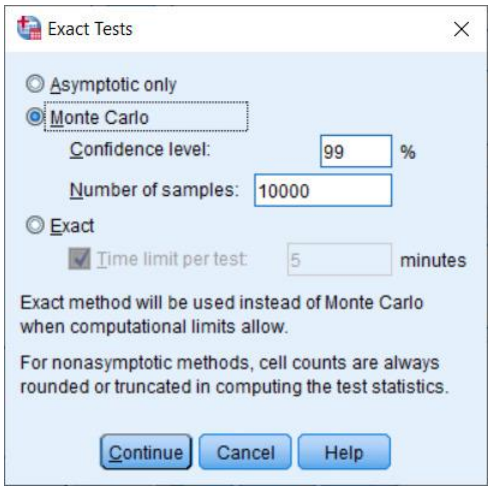
H_0: υπάρχει ανεξαρτησία μεταξύ δύο κατηγορικών μεταβλητών.
H_1: δεν υπάρχει ανεξαρτησία μεταξύ δύο κατηγορικών μεταβλητών.

Οπότε, χρησιμοποιώντας ξανά τα δεδομένα των πιστωτικών καρτών (**credit.sav**), μπορούμε να εξετάσουμε αν υπάρχει εξάρτηση μεταξύ των μεταβλητών ιδιοκτήτης σπιτιού ("owner") και κάτοχος πιστωτικής κάρτας ("card"). Η μηδενική υπόθεση είναι πάντα αυτή που *δεν* υποθέτει εξάρτηση (δηλαδή, υποθέτει ανεξαρτησία μεταξύ των μεταβλητών). Βασική προϋπόθεση του χ^2 ελέγχου ανεξαρτησίας είναι οι συχνότητες των κελιών να είναι τουλάχιστον ίσες με 5. Θα πρέπει να επισημάνουμε στο σημείο αυτό πως το SPSS χρησιμοποιεί ένα άλλο είδος υπόθεσης που προϋποθέτει ότι οι αναμενόμενες συχνότητες των κελιών να είναι τουλάχιστον ίσες με 5. Ένα αποδεκτό ποσοστό κελιών με συχνότητες μικρότερες του 5 είναι το 25%. Αυτό σημαίνει ότι το πολύ ένα στα τέσσερα κελιά μπορεί να έχει τιμή μικρότερη του 5, χωρίς αυτό να μειώνει σημαντικά την αποτελεσματικότητα του ελέγχου. Προφανώς, το παραπάνω ισχύει και για πίνακες που έχουν περισσότερα κελιά. Στην περίπτωση που η υπόθεση αυτή δεν ικανοποιείται, τότε εστιάζουμε στην *p*-value που υπολογίζεται με βάση τον ακριβή έλεγχο του Fisher (**Fisher's exact test**) ή το Monte Carlo.

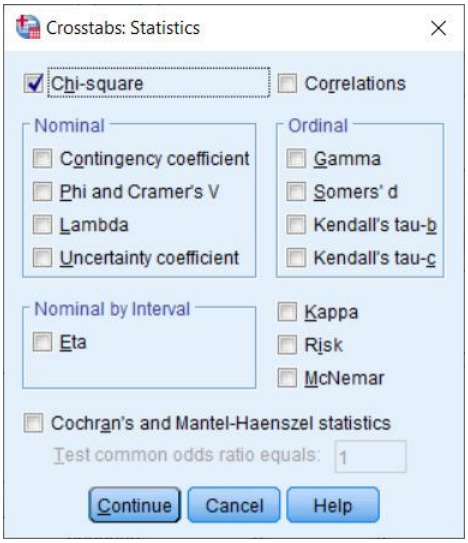
- Στο **SPSS**: Η διαδικασία ελέγχου ανεξαρτησίας είναι η ακόλουθη. Επιλέγοντας **Analyze → Descriptive Statistics → Crosstabs**, θα εμφανιστεί το παράθυρο της **Εικόνας 4.20**. Στο παράθυρο αυτό θα περάσουμε τις δύο μεταβλητές στα δεξιά κουτιά κάτω από τις ενδείξεις **Row(s)** και **Columns(s)**. Μία μεταβλητή θα τοποθετηθεί στο πάνω κουτί και μία στο κάτω. Παρατηρήστε, επίσης, ότι έχει «τσεκαριστεί» η επιλογή **Display clustered bar charts** στο κάτω αριστερό μέρος. Η επιλογή αυτή μας επιτρέπει την εμφάνιση ενός ραβδογράμματος. Επιλέγοντας **Exact**, θα εμφανιστεί το παράθυρο της **Εικόνας 4.21**, από όπου και θα επιλέξουμε τη διεξαγωγή του ελέγχου Monte Carlo. Ο αλγόριθμος bootstrap δε είναι απαραίτητος στη συγκεκριμένη περίπτωση. Ο ακριβής έλεγχος του Fisher θα πραγματοποιηθεί μόνο στην περίπτωση που έχουμε 2×2 πίνακες, όπως στο παράδειγμά μας. Στην περίπτωση αυτή, το Monte Carlo δεν θα υπολογιστεί. Πατώντας **Continue** θα επιστρέψουμε στο παράθυρο της **Εικόνας 4.20**. Στο συγκεκριμένο παράθυρο θα πατήσουμε **Statistics**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 4.22**, στο οποίο θα επιλέξουμε **Chi-square**. Πατώντας **Continue**, επιστρέφουμε στο αρχικό παράθυρο, της **Εικόνας 4.20**, όπου πατώντας **OK** θα εμφανιστούν στο Output του SPSS οι **Πίνακες 4.8** και **4.9**, καθώς και το ραβδόγραμμα του **Διαγράμματος 4.1**. Ο **Πίνακας 4.8** περιέχει τις συχνότητες της κάθε μεταβλητής, ενώ ο **Πίνακας 4.9** περιέχει τα αποτελέσματα των ελέγχων ανεξαρτησίας.



Εικόνα 4.20 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Εικόνα 4.21 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Εικόνα 4.22 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Πίνακας 4.8 *Πίνακας συνάφειας των δύο μεταβλητών.*

Owner * Card Crosstabulation				
Count				
		Card		
		Not accepted	Accepted	Total
Owner	No	206	532	738
	Yes	90	491	581
Total		296	1023	1319

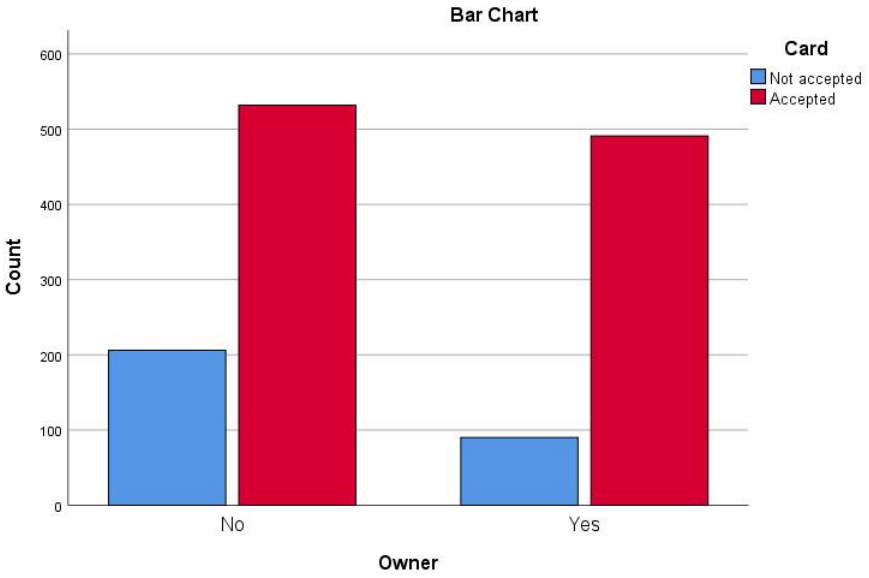
Παρατηρήστε ότι το SPSS εμφανίζει μια σειρά από μηνύματα στο κάτω μέρος του **Πίνακα 4.9**. Ιδιαίτερη σημασία έχει το τελευταίο μήνυμα, στο οποίο αναφέρεται ότι για την περίπτωση των 2×2 πινάκων υπολογίζεται ο έλεγχος του Fisher αντί του Monte Carlo. Επίσης, το δεύτερο μήνυμα μας πληροφορεί για το αν ικανοποιείται η προϋπόθεση ισχύος του χ^2 ελέγχου. Όπως αναφέρθηκε παραπάνω, το πολύ το 25% των κελιών μπορεί να έχει τιμές μικρότερες από 5. Αν δεν ισχύει αυτό, τότε δεν μπορούμε να εμπιστευτούμε τα αποτελέσματα του χ^2 ελέγχου, παρά μόνο αυτά του ελέγχου του Fisher για την περίπτωση πινάκων 2×2 και του Monte Carlo για την περίπτωση πινάκων με περισσότερες από δύο γραμμές ή/και στήλες. Στη βιβλιογραφία αναφέρεται και μία πιο αυστηρή προϋπόθεση σχετικά με τα κελιά που έχουν τιμές μικρότερες του 5, και η οποία απαιτεί όλα τα κελιά να έχουν τιμές μεγαλύτερες του 5.

Πίνακας 4.9 *Αποτελέσματα χ^2 και G^2 ελέγχων ανεξαρτησίας.*

Chi-Square Tests[c]						
	Value	Df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)	Point Probability
Pearson Chi-Square	**28.823[a]**	**1**	**.000**	**.000**	**.000**	
Continuity Correction[b]	28.114	1	.000			
Likelihood Ratio	29.613	1	.000	.000	.000	
Fisher's Exact Test				.000	.000	
Linear-by-Linear Association	28.802d	1	.000	.000	.000	.000
N of Valid Cases	1319					

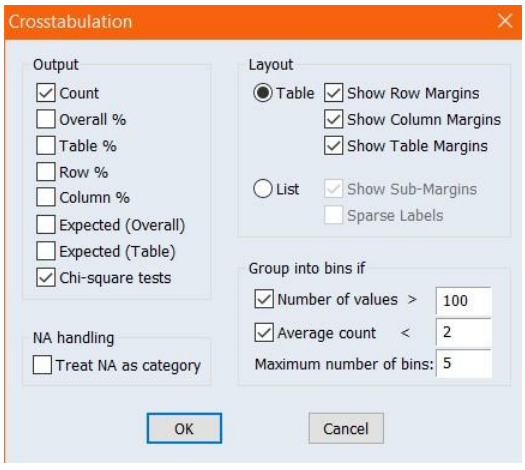
a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 130.38.
b. Computed only for a 2x2 table
c. For 2x2 crosstabulation, exact results are provided instead of Monte Carlo results.

Στον **Πίνακα 4.9** επικεντρωνόμαστε στις *p*-value των διάφορων ελέγχων. Δηλαδή, εστιάζουμε στις στήλες **Asymp. Sig. (2 sided)** και το **Exact Sig. (2-sided)** για τους ελέγχους **Pearson Chi-Square (χ^2 έλεγχος)** και **Fisher's Exact Test**. Προκειμένου να εξάγουμε τα συμπεράσματά μας, μπορούμε, επίσης, να δώσουμε βαρύτητα και στην *p*-value για το **Likelihood Ratio (G^2 έλεγχος)**.



Διάγραμμα 4.1 *Ραβδόγραμμα ποιοτικών μεταβλητών.*

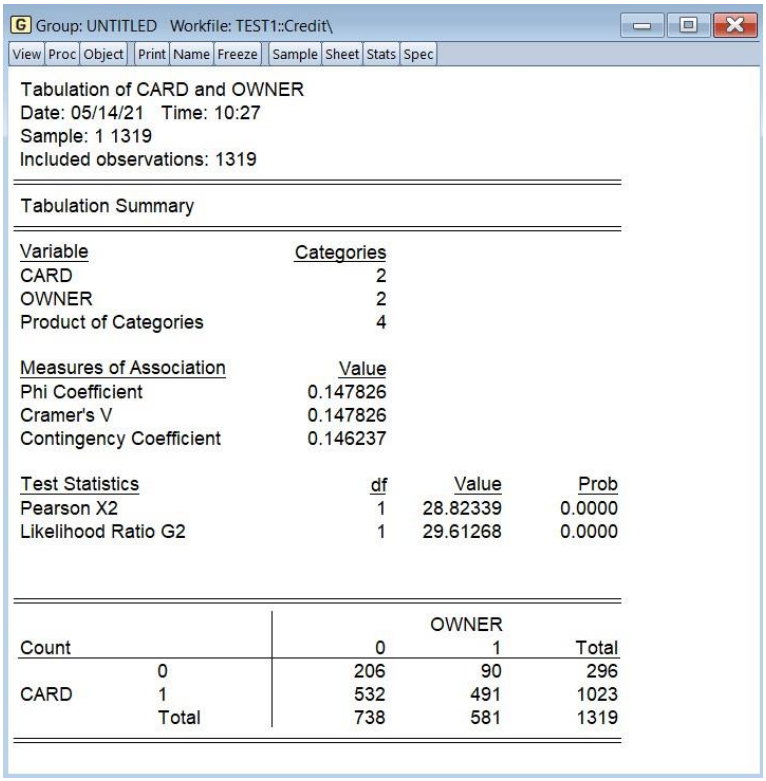
- Στο **Eviews**: Η διαδικασία ελέγχου ανεξαρτησίας μεταξύ δύο κατηγορικών μεταβλητών είναι αρκετά απλή. Και στο συγκεκριμένο πρόγραμμα, η μηδενική υπόθεση είναι πάντα αυτή που υποθέτει ανεξαρτησία μεταξύ των μεταβλητών. Επίσης, δεν υπάρχει καμία προϋπόθεση σχετικά με το ελάχιστο μέγεθος συχνοτήτων ή αναμενόμενων συχνοτήτων των κελιών. Έστω, λοιπόν, ότι θέλουμε να ελέγξουμε αν υπάρχει ανεξαρτησία μεταξύ των κατηγορικών μεταβλητών "card" (που παίρνει την τιμή 1 αν η αίτηση για πιστωτική κάρτα έγινε δεκτή, και την τιμή 0 αν απορρίφθηκε) και "owner" (που παίρνει την τιμή 1 αν το άτομο που έκανε την αίτηση είναι ιδιοκτήτης της τρέχουσας κατοικίας του, και την τιμή και 0 αν δεν είναι). «Ανοίγουμε» τις δύο αυτές μεταβλητές ως Group και επιλέγουμε **View → N-Way Tabulation**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 4.23**. Στη στήλη **Output** εμφανίζεται μια σειρά από «κουτάκια», τα οποία μας επιτρέπουν να επιλέξουμε ποιες πληροφορίες θα εμφανιστούν στο τελικό αποτέλεσμα. Το «κουτάκι» **Count** εμφανίζει τον αριθμό παρατηρήσεων που περιλαμβάνονται σε καθέναν από τους 4 δυνατούς συνδυασμούς 0,0 – 0,1 – 1,0 – 1,1 (όπως φαίνονται στην **Εικόνα 3.38**), το «κουτάκι» **Overall %** εμφανίζει τα αντίστοιχα ποσοστά (όπως φαίνονται στην **Εικόνα 3.39**), ενώ το «κουτάκι» **Chi-square tests** εμφανίζει τα αποτελέσματα των στατιστικών ελέγχου χ^2. Τα υπόλοιπα «κουτάκια» (**Table %**, **Row %**, **Column %**, **Expected (Overall)** και **Expected (Table)**) δεν έχουν ιδιαίτερη σημασία και για αυτό και δεν θα αναλυθούν περισσότερο τη δεδομένη στιγμή.



Εικόνα 4.23 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Στη στήλη **NA handing**, δεν «τσεκάρουμε» την επιλογή **Treat NA as category**, καθώς, αν το κάνουμε και υπάρχουν ελλείπουσες τιμές (missing values), το Eviews θα τις χειριστεί ως ξεχωριστή κατηγορία. Η στήλη **Layout** μας επιτρέπει να επιλέξουμε πώς θέλουμε να εμφανιστούν τα αποτελέσματά μας (πίνακας ή λίστα). Τέλος, η στήλη **Group into bins if** έχει σημασία μόνο στην περίπτωση που κάποιες από τις μεταβλητές που εξετάζουμε είναι ποσοτικές και παίρνουν διάφορες τιμές, με αποτέλεσμα ο αριθμός των κελιών να είναι εξαιρετικά μεγάλος. Οπότε, με τις επιλογές της συγκεκριμένης στήλης μπορούμε να χωρίσουμε τις παρατηρήσεις των συγκεκριμένων μεταβλητών σε υποομάδες. Καθώς στο παράδειγμά μας χρησιμοποιούμε δύο μόνο κατηγορικές μεταβλητές, μπορούμε είτε να αφήσουμε την προεπιλογή του Eviews είτε να αποεπιλέξουμε όλα τα κελιά της συγκεκριμένης στήλης, χωρίς να επηρεαστούν οι εκτιμήσεις μας. Οπότε, διατηρούμε μόνο τις επιλογές της **Εικόνας 4.23**, πατάμε **OK** και εμφανίζονται τα αποτελέσματα **(Εικόνα 4.24)**.

Στη στήλη **Variable** της **Εικόνας 4.24** εμφανίζεται ο αριθμός των κατηγοριών καθεμίας από τις 2 μεταβλητές που εξετάζουμε, καθώς και το σύνολο των πιθανών συνδυασμών (**Product of Categories**). Στη στήλη **Measure of Association** παρουσιάζονται τα ακόλουθα τρία μέτρα συσχέτισης: **Phi Coefficient, Cramer's V** και **Contingency Coefficient**. Τα μέτρα αυτά παίρνουν τιμές από 0 μέχρι και 1 και είναι ανάλογα του συντελεστή συσχέτισης. Όσο πιο κοντά στο 1 βρίσκονται, τόσο υψηλότερη συσχέτιση υπάρχει μεταξύ των υπό εξέταση μεταβλητών. Επίσης, τα συγκεκριμένα μέτρα είναι μη παραμετρικά και συνεπώς, είναι αξιόπιστα και σε περιπτώσεις μη γραμμικής συσχέτισης, κάτι που δεν συμβαίνει με τον συντελεστή συσχέτισης. Στη στήλη **Test Statistics** το Eviews έχει υπολογίσει 2 στατιστικές ελέγχου, τη χ^2 **του Pearson** και την **likelihood ratio** (G^2). Στη στήλη **df** εμφανίζονται οι βαθμοί ελευθερίας για καθεμία από τις στατιστικές αυτές, στην επόμενη στήλη (**Value**) παρουσιάζονται οι εκτιμημένες τιμές των στατιστικών, ενώ στην τελευταία στήλη (**Probability**) εμφανίζονται οι αντίστοιχες *p*-values. Όπως φαίνεται στα αποτελέσματα της **Εικόνας 4.24** η *p*-value είναι μικρότερη του 0,05 και στις 2 στατιστικές ελέγχου. Αυτό σημαίνει πως η μηδενική υπόθεση ότι υπάρχει ανεξαρτησία μεταξύ των μεταβλητών "card" και "owner" απορρίπτεται σε επίπεδο στατιστικής σημαντικότητας $a = 5\%$. Τέλος, στο κάτω μέρος της **Εικόνας 4.24** εμφανίζεται ένας πίνακας που παρουσιάζει τον αριθμό των παρατηρήσεων που περιλαμβάνονται σε καθέναν από τους τέσσερις δυνατούς συνδυασμούς που προκύπτουν από τις δύο υπό εξέταση μεταβλητές (όπως και στην **Εικόνα 3.38**).


Group: UNTITLED Workfile: TEST1::Credit\

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Tabulation of CARD and OWNER
Date: 05/14/21 Time: 10:27
Sample: 1 1319
Included observations: 1319

Tabulation Summary

Variable	Categories
CARD	2
OWNER	2
Product of Categories	4

Measures of Association	Value
Phi Coefficient	0.147826
Cramer's V	0.147826
Contingency Coefficient	0.146237

Test Statistics	df	Value	Prob
Pearson X2	1	28.82339	0.0000
Likelihood Ratio G2	1	29.61268	0.0000

		OWNER		
Count		0	1	Total
CARD	0	206	90	296
	1	532	491	1023
	Total	738	581	1319

Εικόνα 4.24 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Βιβλιογραφία

Ξενόγλωσση

Tsagris, M., Alenazi, A., Verrou, K. M., & Pandis, N. (2020). Hypothesis testing for two population means: parametric or non-parametric test?. *Journal of Statistical Computation and Simulation*, *90*(2), 252-270. https://doi.org/10.1080/00949655.2019.1677659

Κεφάλαιο 5 Γραμμική παλινδρόμηση

Σύνοψη

Το πέμπτο κεφάλαιο περιλαμβάνει επτά ενότητες και εστιάζει στην ανάλυση της (απλής και πολυμεταβλητής) γραμμικής παλινδρόμησης. Πιο συγκεκριμένα, επικεντρώνεται στην ερμηνεία των παραμέτρων και στις παραβιάσεις των υποθέσεων του γραμμικού υποδείγματος. Επίσης, παρουσιάζει τις μεθόδους επιλογής μεταβλητών που είναι διαθέσιμες στο SPSS και στο Eviews. Οι βασικοί στόχοι του κεφαλαίου αυτού είναι η κατανόηση της έννοιας της γραμμικής παλινδρόμησης, καθώς και η ερμηνεία των αποτελεσμάτων της.

Προαπαιτούμενη γνώση

Απαιτούνται βασικές γνώσεις στατιστικής.

5.1 Γραμμική παλινδρόμηση

Συνήθως, όταν αναφερόμαστε στην απλή γραμμική παλινδρόμηση, αναφερόμαστε και στη σχέση της με τον γραμμικό συντελεστή συσχέτισης, παρόλο που σε αυτές τις σημειώσεις αυτό δεν συμβαίνει. Η απλή γραμμική παλινδρόμηση χρησιμοποιείται για να εκτιμήσει τη σχέση που υπάρχει μεταξύ μίας ανεξάρτητης μεταβλητής *(X)* και μίας εξαρτημένης μεταβλητής *(Y)*. Με τον όρο εξαρτημένη μεταβλητή εννοούμε ότι οι τιμές της συγκεκριμένης μεταβλητής εξαρτώνται από τις τιμές της ανεξάρτητης μεταβλητής. Αυτό σημαίνει ότι η σχέση που υπάρχει μεταξύ των δύο μεταβλητών είναι στοχαστική, δηλαδή σε κάθε τιμή του *X* μπορεί να αντιστοιχούν περισσότερες από μία τιμές στην *Y*. Αν δεν ισχύει κάτι τέτοιο, τότε έχουμε την περίπτωση μαθηματικών ή συναρτησιακών σχέσεων μονοσήμαντα ορισμένων.

Για να μετρηθεί η ένταση της παραπάνω γραμμικής σχέσης, χρησιμοποιείται ο συντελεστής γραμμικής συσχέτισης, τον οποίο αναλύσαμε στην ενότητα 4.3. Γίνεται πλέον εμφανές ότι απαραίτητη προϋπόθεση εφαρμογής της απλής γραμμικής παλινδρόμησης ή της προσαρμογής ενός απλού γραμμικού υποδείγματος σε δύο μεταβλητές, είναι η ύπαρξη γραμμικής σχέσης μεταξύ των δύο αυτών μεταβλητών. Ο πιο συνηθισμένος διαγραμματικός τρόπος, για να ελέγξουμε τη γραμμικότητα της σχέσης μεταξύ δύο μεταβλητών, είναι το λεγόμενο διάγραμμα διασποράς (**scatter plot**). Πριν όμως μιλήσουμε για τον τρόπο με τον οποίο κατασκευάζουμε ένα διάγραμμα διασποράς στο SPSS και στο Eviews, είναι απαραίτητο να αναφερθούμε στις υπόλοιπες υποθέσεις της απλής και της πολλαπλής γραμμικής παλινδρόμησης, οι οποίες θα αναλυθούν στη συνέχεια, χρησιμοποιώντας και πάλι τα δεδομένα για τις πιστωτικές κάρτες.

Σε μία γραμμική παλινδρόμηση, οι εκτιμημένες (ή προβλεφθείσες) τιμές της εξαρτημένης μεταβλητής θα είναι προφανώς διαφορετικές από τις πραγματικές τιμές της. Οι αποκλίσεις των πραγματικών τιμών της εξαρτημένης μεταβλητής από τις αντίστοιχες εκτιμημένες τιμές της ονομάζονται κατάλοιπα (ή σφάλματα) και συμβολίζονται με $\boldsymbol{e_i}$, όπου $i = 1,2,3, \dots, n$ είναι δείκτης και αναφέρεται στην i-οστή τιμή. Οι υποθέσεις των γραμμικών υποδειγμάτων αφορούν κυρίως τα κατάλοιπα, και είναι οι ακόλουθες:

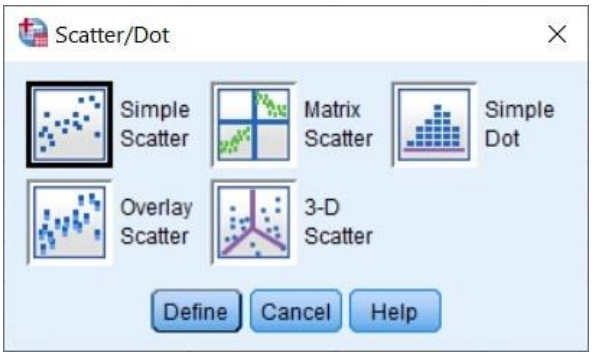
- Υπόθεση της **κανονικότητας των καταλοίπων**, δηλαδή $\boldsymbol{e_i \sim N(0, \sigma^2)}$, όπου $\boldsymbol{N}$ είναι ο συμβολισμός της κανονικής κατανομής (**normal distribution**), ενώ $\boldsymbol{0}$ **(μηδέν)** και $\boldsymbol{\sigma^2}$ είναι ο μέσος και η διακύμανση της κατανομής.
- Υπόθεση της **ανεξαρτησίας των καταλοίπων**, δηλαδή ότι $\boldsymbol{Cov(e_i, e_j) = 0}$ αν $\boldsymbol{i \neq j}$. Αυτό σημαίνει ότι η συνδιακύμανση (**covariance**) για κάθε ζεύγος τιμών των καταλοίπων θα πρέπει να είναι μηδέν. Με άλλα λόγια, δεν θα πρέπει να υπάρχει **σειριακή συσχέτιση (serial correlation)** μεταξύ των καταλοίπων.

- Υπόθεση της **ομοσκεδαστικότητας των καταλοίπων**, δηλαδή $\boldsymbol{Cov(e_i, e_j) = \sigma^2}$ (σταθερή), αν $\boldsymbol{i = j}$. Δηλαδή, η διακύμανση των καταλοίπων θα πρέπει να είναι σταθερή και ίση με $\boldsymbol{\sigma^2}$ για όλες τις τιμές των καταλοίπων.
- Στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης θα πρέπει να μην υπάρχει **συγγραμμικότητα**, δηλαδή οι ανεξάρτητες μεταβλητές να μην συσχετίζονται μεταξύ τους σε μεγάλο βαθμό.
- Τέλος, για να μπορέσει να εκτιμηθεί το γραμμικό υπόδειγμα, θα πρέπει ο αριθμός ($\boldsymbol{p}$) των υπό εκτίμηση συντελεστών να είναι μικρότερος από τον αριθμό των παρατηρήσεων του δείγματος ($\boldsymbol{n}$)**.**

Στη συνέχεια του κεφαλαίου αυτού θα αναλύσουμε το πώς ελέγχουμε τις παραπάνω υποθέσεις, καθώς και τι μπορούμε να κάνουμε στις περιπτώσεις που αυτές δεν ικανοποιούνται. Τα δεδομένα που θα χρησιμοποιήσουμε είναι τα δεδομένα των πιστωτικών καρτών (**credit.sav**).

5.2 Διάγραμμα διασποράς

- Ο πρώτος τρόπος είναι μέσω της επιλογής **Graphs → Legacy Dialogs → Scatter/Dot**, προκειμένου να εμφανιστεί το παράθυρο της **Εικόνας 5.1**. Στο παράθυρο αυτό επιλέγουμε **Simple Scatter** και στη συνέχεια **Define**, με αποτέλεσμα να οδηγηθούμε στο παράθυρο της **Εικόνας 5.2**, στο οποίο θα ορίσουμε τις μεταβλητές που θα απεικονιστούν στο διάγραμμα διασποράς (στο παράδειγμά μας, τις μεταβλητές "expenditure" και "income"). Στο παράθυρο αυτό θα «σύρουμε» τις μεταβλητές αυτές, προκειμένου να περαστούν στα δεξιά κουτιά κάτω από τις ενδείξεις **Y Axis:** και **X Axis:**.



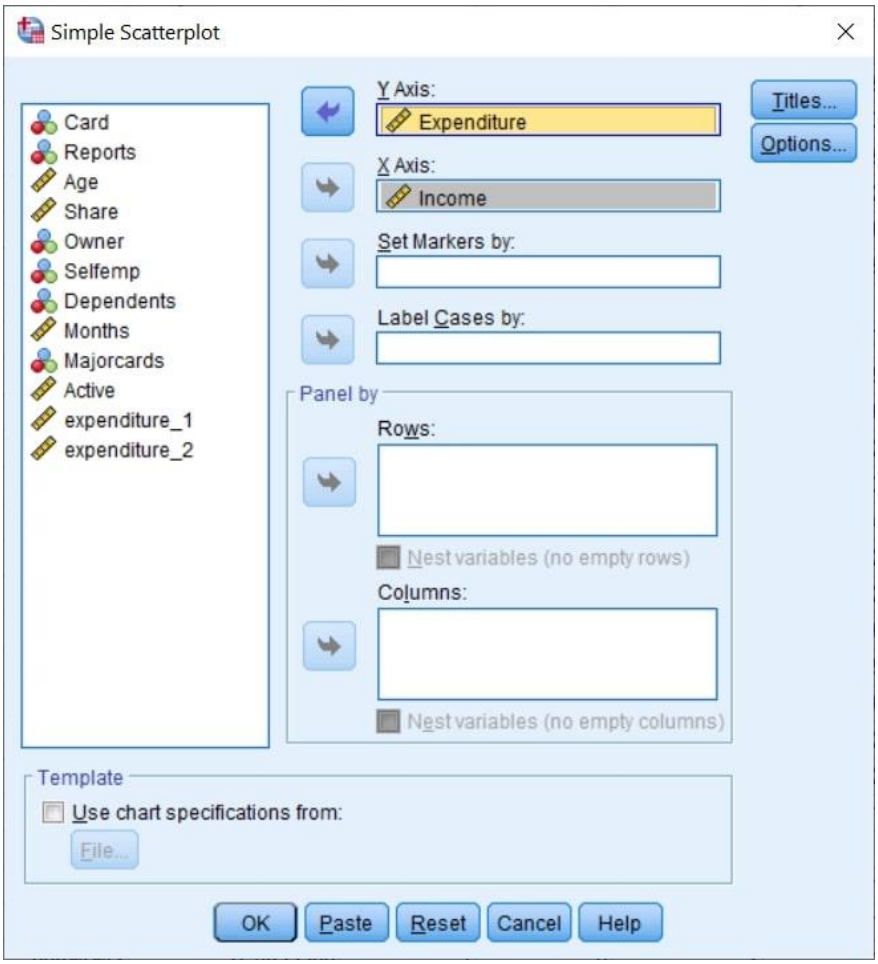
Εικόνα 5.1 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Η επιλογή **Titles** μας επιτρέπει να δώσουμε κάποιον τίτλο στο διάγραμμα, ενώ η επιλογή **Options** μας επιτρέπει να επιλέξουμε αν θέλουμε να εμφανίζονται οι εκλιπούσες τιμές. Θα πρέπει να αναφέρουμε στο σημείο αυτό πως το SPSS έχει ως προεπιλογή τη μη χρησιμοποίηση των συγκεκριμένων τιμών. Επίσης, αν περάσουμε μία κατηγορική μεταβλητή στο λευκό κουτί κάτω από την ένδειξη **Set Markers by:**, το SPSS θα χρωματίσει τους κύκλους του διαγράμματος ανάλογα με τις τιμές της κατηγορικής μεταβλητής.

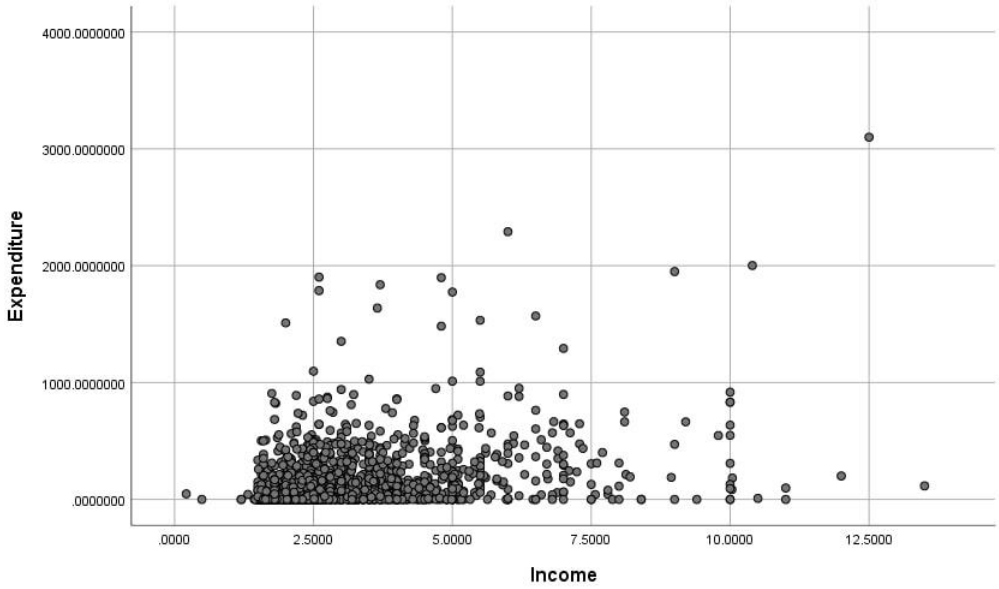
Πατώντας **OK**, εμφανίζεται το διάγραμμα διασποράς για τις μεταβλητές "expenditure" και "income" **(Διάγραμμα 5.1)**. Στο διάγραμμα αυτό μπορούμε να παρατηρήσουμε ότι υπάρχει μια μικρή σχέση μεταξύ των δύο μεταβλητών και συνεπώς, μπορούμε να υποθέσουμε γραμμική σχέση μεταξύ τους. Αν υπολογίσουμε τον συντελεστή γραμμικής συσχέτισης μεταξύ των δύο αυτών μεταβλητών, θα διαπιστώσουμε ότι, παρόλο που έχει χαμηλή τιμή (0,281), είναι στατιστικά σημαντικός (p-value < 0.01).

Ο δεύτερος τρόπος κατασκευής του διαγράμματος διασποράς είναι μέσω της επιλογής των γραφημάτων. Επιλέγοντας **Graphs → Chart Builder**, θα εμφανιστεί το παράθυρο της **Εικόνας 5.3**, το οποίο είναι παρόμοιο με αυτό του Microsoft Excel. Παρατηρήστε ότι έχουμε επιλέξει το **Scatter/Dot** από την Gallery κάτω δεξιά, και το οποίο πρέπει να σύρουμε στο μεγάλο κουτί πάνω δεξιά (**Drag a Gallery chart to use it as your starting point**). Μόλις περάσουμε το συγκεκριμένο διάγραμμα μέσα σε αυτό το κουτί, θα

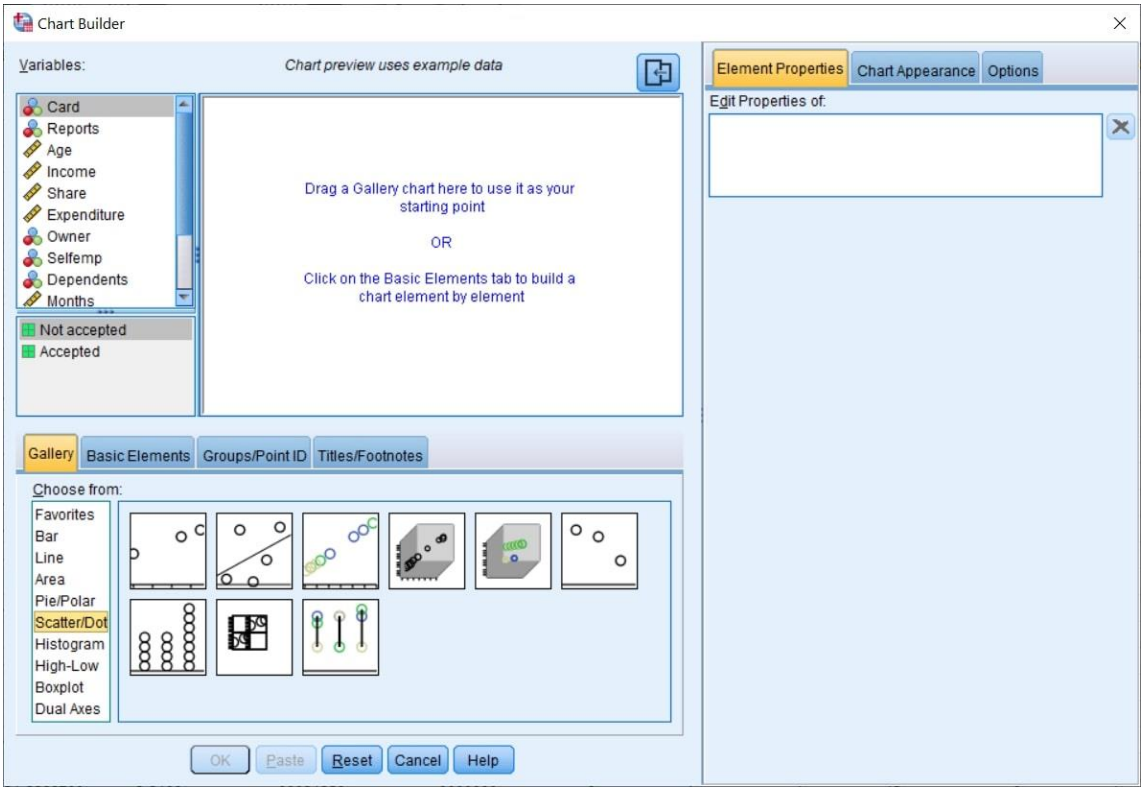
εμφανιστεί το παράθυρο της **Εικόνας 5.4**, το οποίο είναι σχεδόν ίδιο με αυτό της **Εικόνας 5.3**, με μόνη διαφορά ότι το δεξιό πλαίσιο έχει αλλάξει. Στη συνέχεια, θα περάσουμε στον κατακόρυφο και στον οριζόντιο άξονα τις μεταβλητές που επιθυμούμε ("expenditure" και "income", αντίστοιχα) και, πατώντας **OK**, θα εμφανιστεί και πάλι το γράφημα του **Διαγράμματος 5.1**.



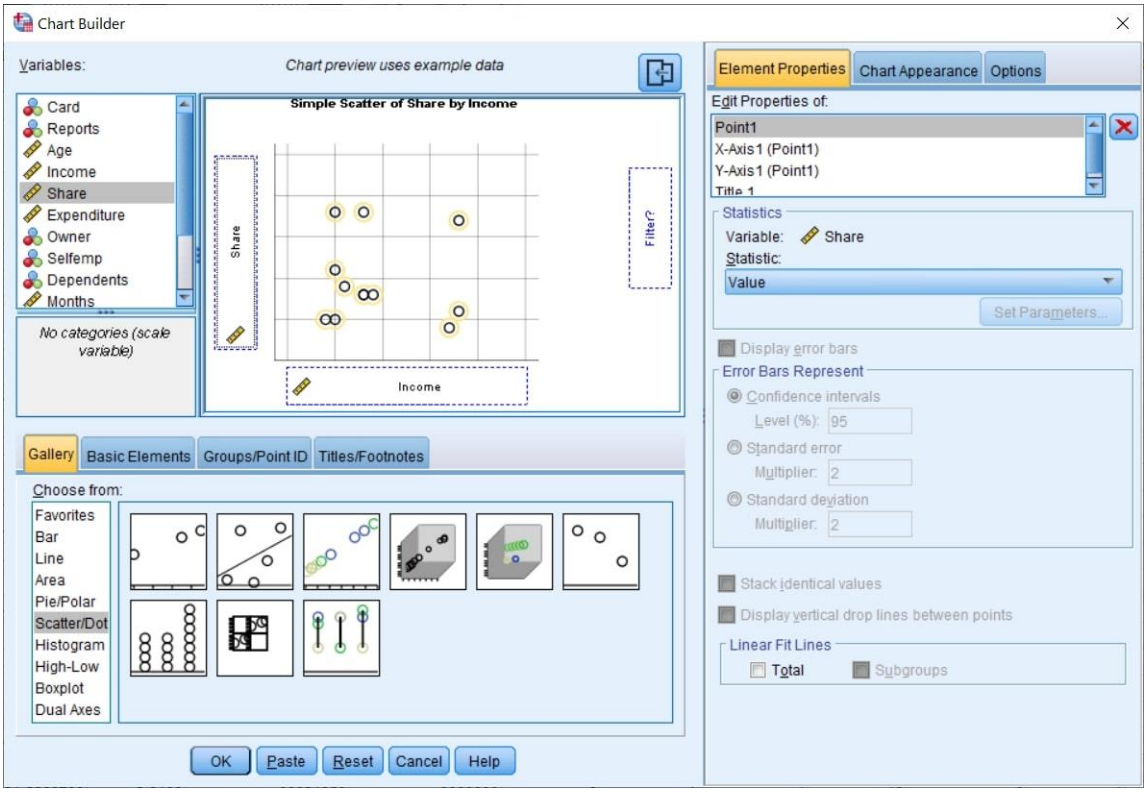
Εικόνα 5.2 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Διάγραμμα 5.1 *Διάγραμμα διασποράς.*

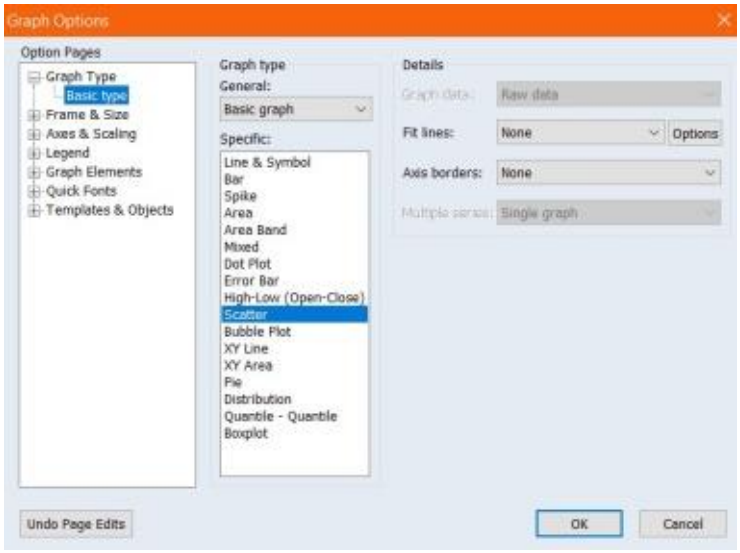


Εικόνα 5.3 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

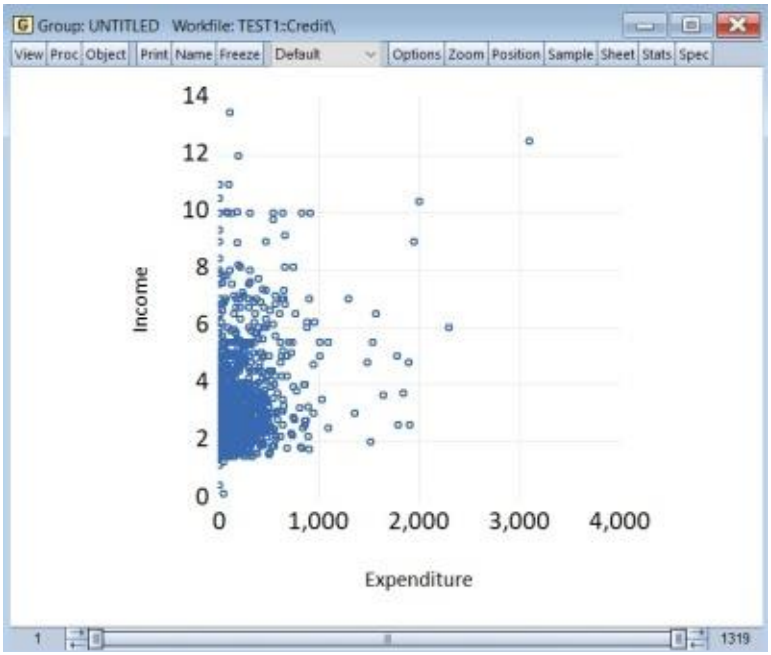


Εικόνα 5.4 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

- Στο **Eviews**: Για να κατασκευάσουμε ένα διάγραμμα διασποράς μεταξύ δύο μεταβλητών, η διαδικασία είναι πολύ απλή. Έστω, λοιπόν, ότι θέλουμε να δημιουργήσουμε ένα τέτοιο διάγραμμα για τις μεταβλητές "expenditure" και "income". Αρχικά, τις «ανοίγουμε» ως Group και στη συνέχεια επιλέγουμε **View → Graph**. Στο παράθυρο "Graph Options" που θα εμφανιστεί **(Εικόνα 5.5)**, επιλέγουμε **Basic type** στην κατηγορία "Option Pages", **Basic Graph** στην κατηγορία "General:" και **Scatter** στην κατηγορία "Specific". Στις επιλογές "Fit lines:" και "Axis borders:" αφήνουμε την προεπιλογή "None" του Eviews και στη συνέχεια πατάμε **OK**. Το διάγραμμα διασποράς είναι πλέον έτοιμο και εμφανίζεται στην **Εικόνα 5.6**. Όπως και στην περίπτωση των διαγραμμάτων που κατασκευάσαμε στο κεφάλαιο 3, με **διπλό κλικ** πάνω στο διάγραμμα διασποράς εμφανίζεται και πάλι το παράθυρο "Graph Options" της **Εικόνας 5.5,** το οποίο μας επιτρέπει να διαμορφώσουμε το συγκεκριμένο διάγραμμα με βάση τις προτιμήσεις μας. Επίσης, μπορούμε να το αποθηκεύσουμε ως γράφημα στο Eviews workfile ακολουθώντας τη διαδικασία που περιγράψαμε στην ενότητα 3.3 σχετικά με την αποθήκευση ιστογραμμάτων.



Εικόνα 5.5 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



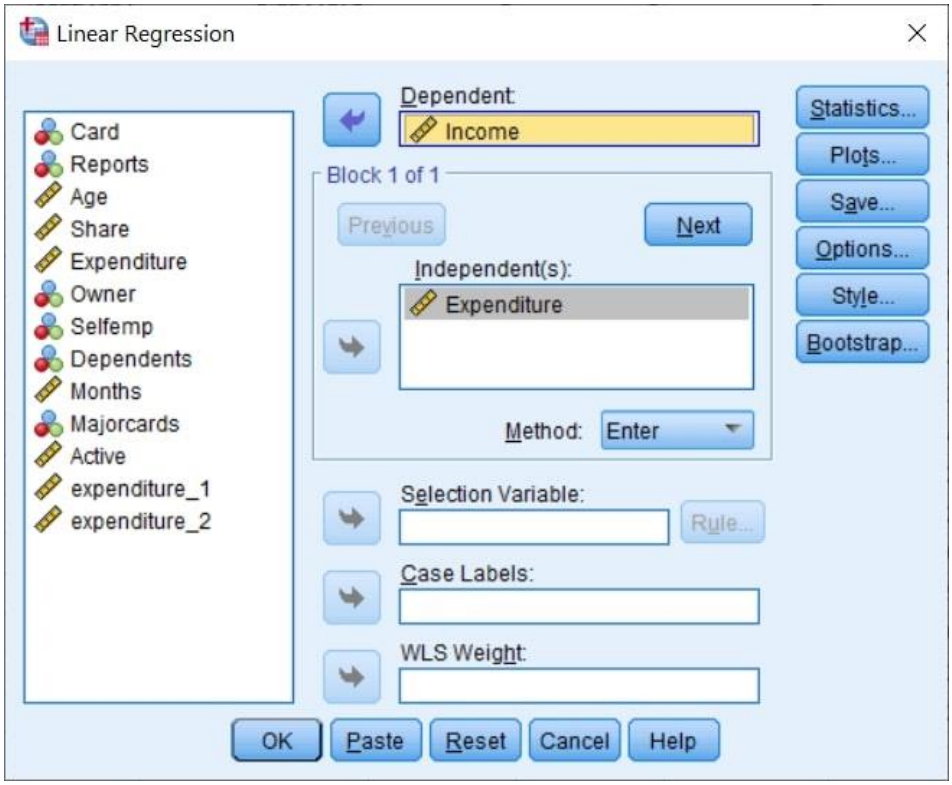
Εικόνα 5.6 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

5.3 Απλή γραμμική παλινδρόμηση

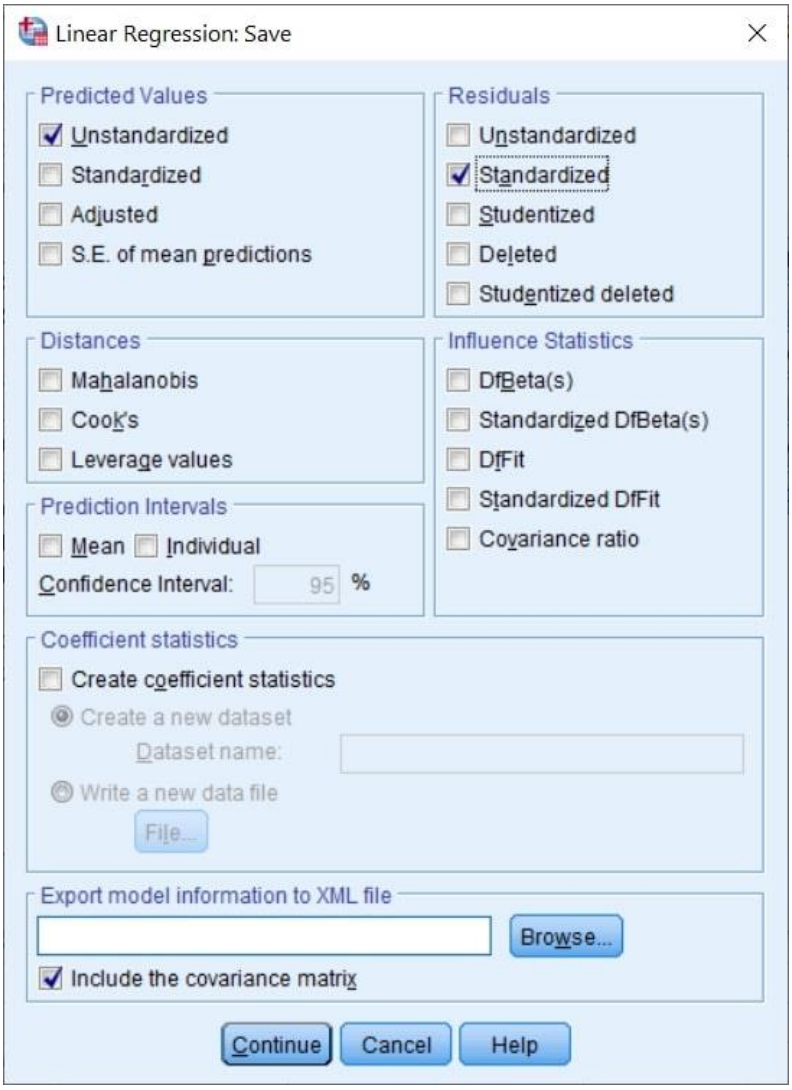
Με την απλή γραμμική παλινδρόμηση προσπαθούμε να εκτιμήσουμε τις τιμές μίας μεταβλητής που θεωρούμε ως εξαρτημένη, χρησιμοποιώντας τις τιμές μιας άλλης μεταβλητής που θεωρούμε ως ανεξάρτητη. Ως παράδειγμα στη συγκεκριμένη ενότητα, θα εκτιμήσουμε τη σχέση που συνδέει το ετήσιο εισόδημα (“income”), το οποίο μετριέται σε δεκάδες χιλιάδες $ και αποτελεί την ανεξάρτητη μεταβλητή *X*, με τη μέση μηνιαία δαπάνη μέσω της χρήσης πιστωτικής κάρτας (“expenditure”), η οποία μετριέται σε χιλιάδες $ και αποτελεί την εξαρτημένη μεταβλητή *Y*.

- Στο **SPSS**: Προκειμένου να εμφανιστεί η ευθεία της γραμμικής παλινδρόμησης, μαζί με κάποια διαγνωστικά μέτρα, επιλέγουμε από το μενού επιλογών **Analyze → Regression → Linear**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 5.7**. Στο λευκό κουτί κάτω από την ένδειξη **Dependent:** θα περάσουμε την εξαρτημένη μεταβλητή και στο λευκό κουτί κάτω από την ένδειξη **Independent(s):** θα περάσουμε την ανεξάρτητη μεταβλητή, της οποίας την

επίδραση πάνω στην εξαρτημένη μεταβλητή θέλουμε να εκτιμήσουμε. Αν γνωρίζουμε τη χρονική σειρά με την οποία έγιναν οι μετρήσεις, μπορούμε από την επιλογή **Statistics** να επιλέξουμε να εμφανιστεί ο έλεγχος των **Durbin-Watson**, ο οποίος χρησιμοποιείται προκειμένου να ελεγχθεί η πιθανή ύπαρξη σειριακής συσχέτισης των καταλοίπων. Από την ίδια επιλογή μπορούμε να επιλέξουμε να εμφανιστούν τα διαγνωστικά μέτρα σχετικά με την πιθανή ύπαρξη συγγραμμικότητας, στην οποία θα αναφερθούμε, όταν αναλύσουμε την πολλαπλή γραμμική παλινδρόμηση (δηλαδή, όταν έχουμε περισσότερες από μία ανεξάρτητες μεταβλητές). Επιλέγοντας **Save** θα εμφανιστεί το παράθυρο της **Εικόνας 5.8**, στο οποίο μπορούμε να επιλέξουμε να αποθηκεύσουμε τις μη τυποποιημένες (**Unstandardized**) εκτιμημένες τιμές της εξαρτημένης μεταβλητής, καθώς και τα τυποποιημένα (**Standardized**) κατάλοιπα. Στο παράθυρο της **Εικόνας 5.7** μπορούμε, επίσης, να ανοίξουμε το παράθυρο της επιλογής **Plots**, προκειμένου να κατασκευάσουμε τα απαραίτητα διαγράμματα σχετικά με τους ελέγχους των υποθέσεων.



Εικόνα 5.7 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

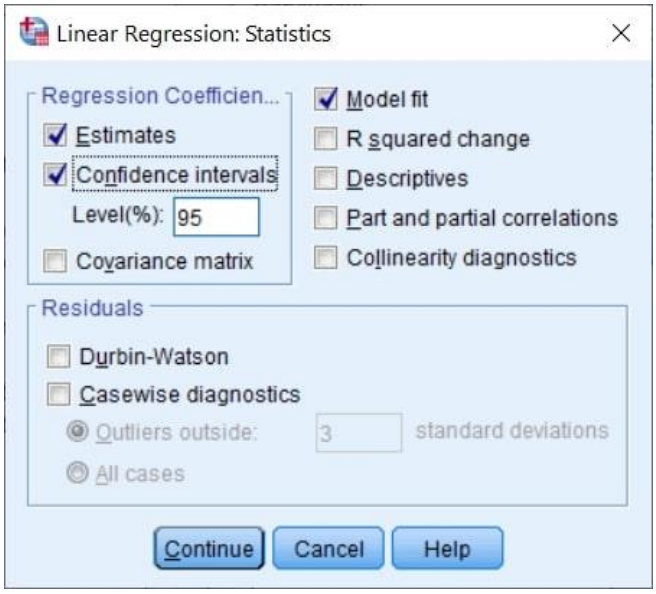


Εικόνα 5.8 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Προφανώς και στην τελευταία αυτή περίπτωση, θα χρειαστεί να αποθηκεύσουμε τα κατάλοιπα, για να διενεργήσουμε τον έλεγχο κανονικότητας των Kolmogorov-Smirnov. Γενικά, επιλέγουμε τα τυποποιημένα κατάλοιπα και όχι τα μη τυποποιημένα, καθώς με τον τρόπο αυτό μπορούμε να αντιμετωπίσουμε το πρόβλημα των ακραίων τιμών. Πιο αναλυτικά, αν κάποιες εκτιμημένες τιμές της εξαρτημένης μεταβλητής δίνουν κατάλοιπα με μεγάλες, κατά απόλυτη τιμή, τιμές (δηλαδή, πάνω από 2 ή 2,5) αυτό αποτελεί ένδειξη ότι οι συγκεκριμένες τιμές είναι ακραίες.

Επιλέγοντας **Statistics** στο παράθυρο της **Εικόνας 5.7** θα εμφανιστεί το παράθυρο **της Εικόνας 5.9**, στο οποίο θα «τσεκάρουμε» την επιλογή **Confidence intervals**, και στη συνέχεια επιλέγουμε **Continue**, προκειμένου να γυρίσουμε στο παράθυρο της **Εικόνας 5.7**. Πατώντας **OK** στο παράθυρο της **Εικόνας 5.7**, θα εμφανιστούν στο Output του SPSS τα αποτελέσματα που παρουσιάζονται στους **Πίνακες 5.1-5.4**.

Στον **Πίνακα 5.1,** η τιμή **R** αναφέρεται στην **απόλυτη τιμή** του συντελεστή γραμμικής συσχέτισης. Το **R Square** είναι το τετράγωνο του συντελεστή γραμμικής συσχέτισης και ονομάζεται συντελεστής προσδιορισμού. Ο συντελεστής προσδιορισμού φανερώνει το ποσοστό της μεταβλητότητας των δεδομένων που εξηγείται από την προσαρμογή τους στο γραμμικό υπόδειγμα. Όπως φαίνεται από τον **Πίνακα 5.1**, το συγκεκριμένο υπόδειγμα εξηγεί μόλις το 7,9% της μεταβλητότητας των δεδομένων. Ο προσαρμοσμένος ή διορθωμένος συντελεστής προσδιορισμού (**Adjusted R Square**) έχει παρόμοια ερμηνεία με τον συντελεστή προσδιορισμού, λαμβάνοντας όμως υπόψη και το μέγεθος του δείγματος.



Εικόνα 5.9 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Πίνακας 5.1 *Συντελεστής προσδιορισμού απλής παλινδρόμησης.*

Model Summary[b]				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.281[a]	.079	.078	261.3414792980
a. Predictors: (Constant), Income b. Dependent Variable: Expenditure				

Στον **Πίνακα 5.2** παρουσιάζεται ο έλεγχος *F* (που βασίζεται στην *F* κατανομή), ο οποίος ελέγχει τη μηδενική υπόθεση ότι όλες οι παράμετροι κλίσης του υποδείγματος (δηλαδή, ο συντελεστής παλινδρόμησης ή αλλιώς, το *β* της παλινδρόμησης) είναι στατιστικά μηδέν. Η εναλλακτική υπόθεση στον συγκεκριμένο έλεγχο είναι ότι έστω μία από τις παραμέτρους κλίσης είναι στατιστικά διαφορετική από το μηδέν. Στο παράδειγμά μας, η τιμή του συγκεκριμένου ελέγχου είναι ίση με **112,998**, ενώ η αντίστοιχη *p*-value είναι μικρότερη του 0,01. Οπότε, μπορούμε να συμπεράνουμε ότι ο συντελεστής του εισοδήματος είναι στατιστικά διαφορετικός από το μηδέν, σε επίπεδο στατιστικής σημαντικότητας ίσο με 5%.

Πίνακας 5.2 *Πίνακας ανάλυσης διακύμανσης απλής παλινδρόμησης.*

ANOVA[a]						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	**Regression**	**7717668.546**	1	7717668.546	**112.998**	.000**b**
	Residual	89950268.712	1317	68299.369		
	Total	**97667937.258**	1318			
a. Dependent Variable: Expenditure **b. Predictors: (Constant), Income**						

Στον πίνακα της ανάλυσης διακύμανσης **(Πίνακας 5.2)** έχουμε επισημάνει με έντονη γραφή δύο ακόμα αριθμούς. Ο πρώτος αριθμός (**7717668,546**) αντιστοιχεί στη διακύμανση που εξηγείται από το υπόδειγμα που προσαρμόσαμε, ενώ ο δεύτερος αριθμός (**97667937,258**) αντιστοιχεί στη συνολική διακύμανση των δεδομένων. Προφανώς, η διαφορά τους (89950268,712) αντιστοιχεί στη διακύμανση που δεν εξηγείται από το υπόδειγμα, ενώ το πηλίκο των δύο αριθμών είναι ουσιαστικά ο συντελεστής προσδιορισμού.

Το υπόδειγμα που προσαρμόσαμε στις δύο μεταβλητές (ή, με άλλα λόγια, η ευθεία των ελαχίστων τετραγώνων, όπως αλλιώς λέγεται) είναι της μορφής $\boldsymbol{y = \alpha + \beta x + e_i}$, όπου $\boldsymbol{y}$ είναι η εξαρτημένη μεταβλητή, $\boldsymbol{x}$ η ανεξάρτητη μεταβλητή, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ οι παράμετροι του υποδείγματος που εκτιμάμε και $\boldsymbol{e_i}$ το κατάλοιπο της i-οστής τιμής.

Πίνακας 5.3 *Εκτιμήσεις παραμέτρων απλής παλινδρόμησης.*

Coefficients[a]								
		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	**33.027**	16.010		2.063	.039	1.618	64.435
	Income	**45.175**	4.250	.281	10.630	.000	36.838	53.512
a. Dependent Variable: Expenditure								

Στο παράδειγμά μας **(Πίνακας 5.3)** η εξίσωση της γραμμικής παλινδρόμησης έχει την ακόλουθη μορφή:

$$\widehat{\boldsymbol{Expenditure}} = \mathbf{33.027} + \mathbf{45.175} * \boldsymbol{Income}.$$

Στον συγκεκριμένο πίνακα εμφανίζονται με έντονη γραφή οι εκτιμημένες τιμές των παραμέτρων. Η τιμή **33,027** (**Constant**) είναι η τιμή στην οποία η ευθεία ελαχίστων τετραγώνων που προσαρμόσαμε τέμνει τον κάθετο άξονα των $\boldsymbol{y'y}$. Η τιμή **45,175** είναι η κλίση της ευθείας, ενώ δείχνει και την επίδραση της ανεξάρτητης μεταβλητής στην εξαρτημένη. Γενικά, για κάθε αύξηση της ανεξάρτητης μεταβλητής κατά 1 μονάδα, η εκτιμημένη μέση τιμή της εξαρτημένης μεταβλητής μειώνεται ή αυξάνεται κατά $\boldsymbol{\beta}$ μονάδες. Οπότε, στο παράδειγμά μας, αν το εισόδημα αυξηθεί κατά 1 μονάδα (ή \$10.000), η εκτιμημένη μέση δαπάνη θα αυξηθεί κατά 45,175 ή \$4.517,5. Η στήλη *Sig*. περιέχει τα παρατηρούμενα επίπεδα στατιστικής σημαντικότητας, τα οποία είναι απαραίτητα, προκειμένου να εξάγουμε συμπεράσματα σχετικά με τη στατιστική σημαντικότητα των παραμέτρων $\boldsymbol{\alpha}$ και $\boldsymbol{\beta}$ του υποδείγματος. Οι υποθέσεις που ελέγχονται για τις συγκεκριμένες παραμέτρους είναι της ακόλουθης μορφής:

$$\mathbf{H_0}: \boldsymbol{\alpha} = \mathbf{0} \qquad \textbf{και} \qquad \mathbf{H_0}: \boldsymbol{\beta} = \mathbf{0}$$
$$\mathbf{H_1}: \boldsymbol{\alpha} \neq \mathbf{0} \qquad\qquad \mathbf{H_1}: \boldsymbol{\beta} \neq \mathbf{0}$$

Καθώς και οι δύο *p*-values είναι μικρότερες του 0,05, μπορούμε να συμπεράνουμε ότι και οι δύο μηδενικές υποθέσεις απορρίπτονται. Συνεπώς, και οι δύο εκτιμημένοι συντελεστές είναι στατιστικά σημαντικοί και άρα απαραίτητοι για το συγκεκριμένο υπόδειγμα.

Οι δύο τελευταίες στήλες του **Πίνακα 5.3** παρουσιάζουν τα 95% διαστήματα εμπιστοσύνης για τους δύο συντελεστές του υποδείγματος. Αν το διάστημα εμπιστοσύνης για κάποιον συντελεστή περιέχει το μηδέν, τότε μπορούμε να πούμε ότι η μηδενική υπόθεση ότι η τιμή του συγκεκριμένου συντελεστή είναι στατιστικά μηδέν δεν μπορεί να απορριφθεί. Για παράδειγμα, το 95% διάστημα εμπιστοσύνης για την πραγματική τιμή του συντελεστή $\boldsymbol{\beta}$ είναι (36,838, 53,512). Καθώς δεν περιέχει το μηδέν, η υπόθεση ότι το $\boldsymbol{\beta}$ μπορεί να πάρει την τιμή μηδέν απορρίπτεται σε επίπεδο σημαντικότητας 5%. Αν επιλέξουμε να εφαρμόσουμε τον αλγόριθμο bootstrap, τότε τα κατάλοιπα και οι εκτιμημένες τιμές δεν θα αποθηκευτούν, αλλά θα υπολογιστούν τα διαστήματα εμπιστοσύνης για τις τιμές των παραμέτρων.

Ο τελευταίος πίνακας **(Πίνακας 5.4)** παρουσιάζει κάποια περιγραφικά μέτρα σχετικά με τα κατάλοιπα. Όπως φαίνεται, ο μέσος τους είναι ίσος με το μηδέν. Θα πρέπει να υπενθυμίσουμε στο σημείο αυτό ότι η βασική υπόθεση σχετικά με τα κατάλοιπα είναι ότι ακολουθούν την κανονική κατανομή με μέσο ίσο με το μηδέν. Για να έχουν, όμως, τα κατάλοιπα μηδενικό μέσο, θα πρέπει η σταθερά να είναι μέσα στο υπόδειγμα. Δηλαδή, ακόμα και αν η *p*-value για τη σταθερά είναι μεγαλύτερη του 0,05 (και άρα δεν μπορεί να απορριφθεί η υπόθεση ότι η πραγματική της τιμή είναι ίση με το μηδέν), αυτή δεν πρέπει να αφαιρεθεί από το υπόδειγμα. Ουσιαστικά, δεν έχει καμία σημασία η στατιστική σημαντικότητα της σταθεράς, παρά μόνο η σημαντικότητα του συντελεστή της μεταβλητής.

Πίνακας 5.4 *Περιγραφικά μέτρα καταλοίπων απλής παλινδρόμησης.*

Residuals Statistics[a]					
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	42.513294220	642.887634277	185.057070778	76.5218315429	1319
Residual	-529.9503784180	2501.79687500	.0000000000	261.2423172837	1319
Std. Predicted Value	-1.863	5.983	.000	1.000	1319
Std. Residual	-2.028	9.573	.000	1.000	1319
a. Dependent Variable: Expenditure					

Στη συνέχεια, θα εξετάσουμε τους τρόπους με τους οποίους μπορούμε να ελέγξουμε αν ικανοποιούνται οι βασικές υποθέσεις του υποδείγματος. Στο παράθυρο της **Εικόνας 5.8** επιλέξαμε να αποθηκεύσουμε στο SPSS Data Editor τα κατάλοιπα και τις εκτιμημένες τιμές για την ευθεία των ελαχίστων τετραγώνων. Σχετικά με την κανονικότητα των καταλοίπων, η συγκεκριμένη υπόθεση ελέγχεται είτε γραφικά (**P-P Plot** ή **Q-Q Plot**) είτε με τον έλεγχο των Kolmogorov-Smirnov που αναλύσαμε στο προηγούμενο κεφάλαιο. Αν δεν μπορούμε να υποθέσουμε κανονικότητα των καταλοίπων, τότε πρέπει να προσπαθήσουμε να μετασχηματίσουμε τις εξαρτημένες μεταβλητές, προκειμένου να οδηγηθούμε στην κανονικότητα αυτή. Σε περίπτωση που το μέγεθος του δείγματος είναι αρκετά μεγάλο, μπορούμε να βασιστούμε στην ασυμπτωτική προσέγγιση της κανονικής κατανομής. Σε αντίθετη περίπτωση καταφεύγουμε σε άλλες τεχνικές, όπως είναι τα εύρωστα γραμμικά υποδείγματα, η παλινδρόμηση του Theil ή η παλινδρόμηση πάνω στις τάξεις μεγέθους των τιμών των μεταβλητών. Η τελευταία τεχνική είναι διαθέσιμη στο SPSS.

Προκειμένου να ελέγξουμε τις υποθέσεις της ανεξαρτησίας και της ομοσκεδαστικότητας των καταλοίπων, χρησιμοποιούμε συνήθως ένα διάγραμμα διασποράς, το οποίο θα περιέχει στον οριζόντιο άξονα τις εκτιμημένες τιμές της γραμμής παλινδρόμησης και στον κάθετο άξονα τα κατάλοιπα. Αν τα κατάλοιπα είναι ανεξάρτητα, τότε αναμένουμε να εμφανιστεί στο διάγραμμα ένα «σύννεφο» σημείων. Δηλαδή, δεν θα πρέπει να εμφανίζεται στο διάγραμμα κάποιο «σχήμα» (pattern). Ένας εναλλακτικός τρόπος, προκειμένου να διεξάγουμε τον παραπάνω έλεγχο, είναι να υπολογίσουμε τη συνδιακύμανση μεταξύ των καταλοίπων και των εκτιμημένων τιμών. Αν υπάρχει ανεξαρτησία των καταλοίπων, τότε η συνδιακύμανση θα είναι ίση με το μηδέν. Δυστυχώς, το αντίστροφο δεν είναι πάντα αληθές. Οπότε, αν η συνδιακύμανση είναι ίση με το μηδέν, αυτό δεν σημαίνει απαραίτητα ότι έχουμε και ανεξαρτησία των καταλοίπων. Αν, παρόλα αυτά, υποψιαζόμαστε ότι η υπόθεση της ανεξαρτησίας των καταλοίπων δεν ικανοποιείται, τότε η γραμμική παλινδρόμηση δεν μπορεί να εφαρμοστεί και είμαστε αναγκασμένοι να χρησιμοποιήσουμε άλλες τεχνικές. Αν τα κατάλοιπα έχουν σειριακή συσχέτιση, τότε μπορούμε να χρησιμοποιήσουμε τη μονότονη παλινδρόμηση η οποία είναι, επίσης, διαθέσιμη στο SPSS.

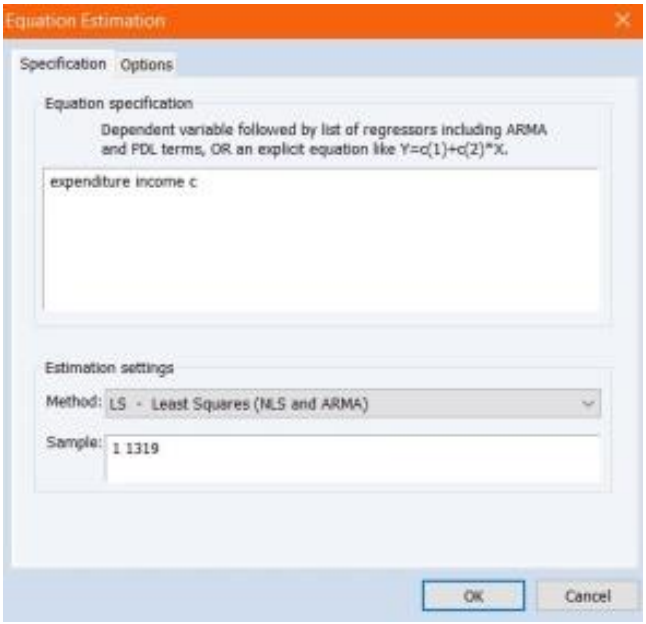
Αν η υπόθεση της ομοσκεδαστικότητας ικανοποιείται, τότε θα πρέπει στο ίδιο διάγραμμα, αν υποθέσουμε δύο παράλληλες γραμμές στον οριζόντιο άξονα, το «σύννεφο» των σημείων να βρίσκεται ανάμεσα στις δύο παράλληλες γραμμές. Με άλλα λόγια, το εύρος των σημείων στον κατακόρυφο άξονα θα πρέπει να παραμένει σταθερό, καθώς «κινούμαστε» στον οριζόντιο άξονα. Αν το εύρος αυτό μεγαλώνει ή μικραίνει, καθώς μετακινούμαστε δεξιά του οριζόντιου άξονα (σχηματίζοντας ένα «χωνί»), τότε δεν μπορούμε να υποθέσουμε ομοσκεδαστικότητα των καταλοίπων. Η αντιμετώπιση αυτού του προβλήματος γίνεται με μετασχηματισμό των ανεξάρτητων μεταβλητών ή ακόμα και των εξαρτημένων. Αν και πάλι δεν λυθεί το συγκεκριμένο πρόβλημα, τότε μπορούμε να δοκιμάσουμε μη παραμετρικές μεθόδους παλινδρόμησης (παλινδρόμηση του Theil ή παλινδρόμηση πάνω στις τάξεις μεγέθους των τιμών των μεταβλητών). Θα πρέπει να τονίσουμε στο σημείο αυτό πως και στις συγκεκριμένες περιπτώσεις η ετεροσκεδαστικότητα αποτελεί πρόβλημα (ίσως όχι τόσο σοβαρό, όπως στις παραμετρικές μεθόδους). Οπότε, ένας εναλλακτικός τρόπος αντιμετώπισης του συγκεκριμένου προβλήματος είναι η χρησιμοποίηση γενικευμένων γραμμικών υποδειγμάτων.

Στο παράδειγμα με τις πιστωτικές κάρτες, υπάρχουν κάποιες ενδείξεις στο διάγραμμα διασποράς των καταλοίπων με τις εκτιμημένες τιμές της γραμμής παλινδρόμησης, ότι η υπόθεση της ομοσκεδαστικότητας δεν ικανοποιείται. Οπότε, εφαρμόζοντας λογαριθμικό μετασχηματισμό στις τιμές της ανεξάρτητης

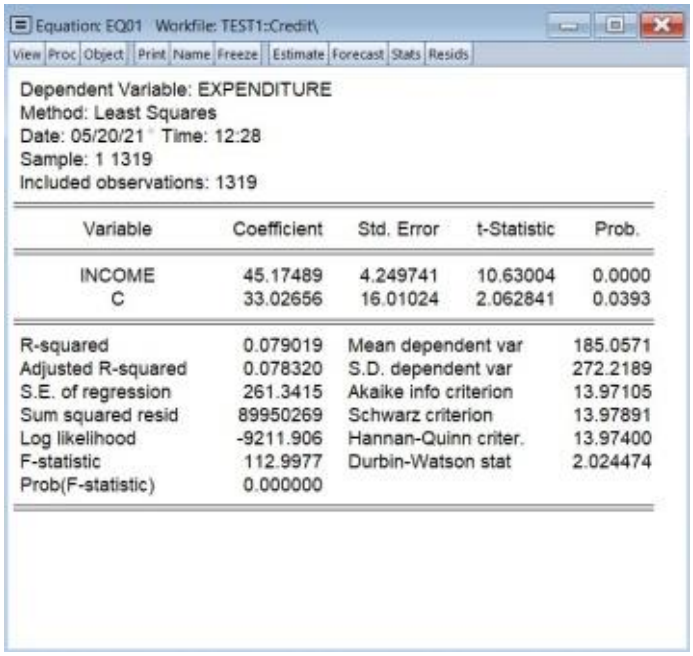
μεταβλητής και εκτιμώντας την εξίσωση της ευθείας παλινδρόμησης πάνω στη μετασχηματισμένη μεταβλητή, το διάγραμμα βελτιώθηκε σε σημαντικό βαθμό.

- Στο **Eviews**: Για να εκτιμήσουμε την ευθεία της γραμμικής παλινδρόμησης μεταξύ της εξαρτημένης (dependent) μεταβλητής "expenditure" και της ανεξάρτητης (independent) μεταβλητής "income", μαζί με κάποια διαγνωστικά μέτρα, μπορούμε:
- Να επιλέξουμε τις δύο αυτές μεταβλητές στο Eviews workfile και με δεξί κλικ να επιλέξουμε **Open → as Equation**.
- Να «ανοίξουμε» τις δύο μεταβλητές ως Group και στη συνέχεια να επιλέξουμε **Proc → Make Equation**.

Και στις δύο περιπτώσεις θα εμφανιστεί το παράθυρο "Equation Estimation" που φαίνεται στην **Εικόνα 5.10**. Στο tab "Specification", στην επιλογή "Equation specification" εμφανίζεται η εξίσωσή μας, **όπου η πρώτη στη σειρά μεταβλητή θα είναι πάντα η εξαρτημένη μεταβλητή** (στο παράδειγμά μας, η μεταβλητή "expenditure"). Στη συνέχεια, ακολουθούν η ανεξάρτητη μεταβλητή "income" και ο σταθερός όρος "c". Στη συγκεκριμένη επιλογή, η εξίσωσή μας μπορεί εναλλακτικά να γραφεί και στην κανονική της μορφή, δηλαδή **expenditure=c(1)+c(2)*income**. Θα πρέπει να σημειώσουμε στο σημείο αυτό πως οι ανεξάρτητες μεταβλητές και ο σταθερός όρος μπορούν να γραφούν με όποια σειρά επιθυμούμε. Στη συνέχεια, η επιλογή "Estimation settings" μας επιτρέπει να επιλέξουμε τη μέθοδο εκτίμησης της παλινδρόμησης (**Method:**) και το δείγμα για το οποίο θέλουμε να γίνει η εκτίμηση (**Sample:**). Στο παράδειγμά μας, επιλέγουμε **LS – Least Squares (NLS and ARMA)** που αντιστοιχεί στη μέθοδο των ελαχίστων τετραγώνων, και **1 1319** προκειμένου η εκτίμηση να αφορά το σύνολο του δείγματός μας. Με το tab "Options" δεν θα ασχοληθούμε προς το παρόν, καθώς οι επιλογές που παρέχει δεν αφορούν την απλή εκτίμηση της γραμμικής παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων. Πατώντας **OK**, εμφανίζεται η εκτίμηση της εξίσωσης, καθώς και κάποια βασικά διαγνωστικά μέτρα **(Εικόνα 5.11)**.



Εικόνα 5.10 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Εικόνα 5.11 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Τη συγκεκριμένη εκτίμηση **(Εικόνα 5.11)** την έχουμε ήδη αποθηκεύσει στο Eviews workfile με το όνομα eq01, προκειμένου να έχουμε τη δυνατότητα να την επεξεργαστούμε όποτε θελήσουμε. Για να το κάνουμε αυτό, επιλέξαμε **Name** στο toolbar που εμφανίστηκε στο πάνω μέρος των αποτελεσμάτων της εκτίμησης και στη συνέχεια δώσαμε το όνομα eq01. Με τον τρόπο αυτό, η εκτίμηση της εξίσωσης αποθηκεύτηκε στο Eviews workfile, ενώ στα αριστερά της εμφανίζεται το εικονίδιο  που είναι το σύμβολο του Eviews για τις εξισώσεις. Ένας γρηγορότερος τρόπος, για να εκτιμήσουμε τη συγκεκριμένη εξίσωση με τη μέθοδο των ελαχίστων τετραγώνων, είναι να γράψουμε απευθείας στο **Command line**:

equation eq01.ls expenditure income c

και να πατήσουμε **Enter**. Με τον τρόπο αυτό θα πάρουμε και πάλι τα αποτελέσματα της **Εικόνας 5.11,** ενώ η εξίσωσή μας θα αποθηκευτεί αυτόματα στο Eviews workfile με το όνομα eq01.

Γενικά, αν θέλουμε να δημιουργήσουμε στο Eviews μια νέα εξίσωση, επιλέγουμε **Object → New Object → Equation** και στη συνέχεια δίνουμε κάποιο όνομα της επιλογής μας. Πατώντας **OK** θα εμφανιστεί το παράθυρο της **Εικόνας 5.10**, όπου η επιλογή "Equation specification" θα είναι κενή. Οπότε, μπορούμε πλέον να γράψουμε όποια εξίσωση επιθυμούμε και στη συνέχεια να την εκτιμήσουμε.

Όπως και στην περίπτωση του SPSS, το υπόδειγμα που προσαρμόσαμε στις δύο μεταβλητές είναι της μορφής $\boldsymbol{y = \alpha + \beta x + e_i}$, όπου $\boldsymbol{y}$ είναι η εξαρτημένη μεταβλητή, $\boldsymbol{x}$ η ανεξάρτητη μεταβλητή, $\boldsymbol{\alpha, \beta}$ οι παράμετροι του υποδείγματος που εκτιμάμε, ενώ ο όρος $\boldsymbol{e_i}$ αναφέρεται στο κατάλοιπο της i-οστής τιμής. Στο πάνω μέρος των αποτελεσμάτων της **Εικόνας 5.11** αναφέρονται η εξαρτημένη μεταβλητή, η μέθοδος εκτίμησης, καθώς και το δείγμα που χρησιμοποιήθηκε για την εκτίμηση αυτή. Στη συνέχεια, παρουσιάζονται οι εκτιμημένοι συντελεστές για τη μεταβλητή "income" (β) και τον σταθερό όρο (α), τα τυπικά τους σφάλματα, οι *t*-στατιστικές τους και οι αντίστοιχες *p*-values. Οι *t*-στατιστικές αφορούν τις μηδενικές υποθέσεις $\boldsymbol{\beta = 0}$ και $\boldsymbol{\alpha = 0}$. Καθώς και οι δύο *p*-values είναι μικρότερες από το 0,05, μπορούμε να εξάγουμε το συμπέρασμα ότι τόσο ο εκτιμημένος συντελεστής της μεταβλητής "income" όσο και ο σταθερός όρος είναι στατιστικά σημαντικοί σε επίπεδο σημαντικότητας $a = 5\%$. Οπότε, η εκτιμημένη γραμμική παλινδρόμηση θα έχει την ακόλουθη μορφή:

$$\widehat{\boldsymbol{Expenditure}} = \mathbf{33.02656} + \mathbf{45.17489} * \boldsymbol{Income}.$$

Οι εκτιμημένοι συντελεστές έχουν, επίσης, αποθηκευτεί στο αντικείμενο **c** του Eviews workfile (που στα αριστερά του έχει το εικονίδιο ). Η τιμή 33,02656 είναι η τιμή στην οποία η εκτιμημένη ευθεία των

ελαχίστων τετραγώνων τέμνει τον κάθετο άξονα των $\boldsymbol{y}$, ενώ η τιμή 45,17489 είναι η κλίση της ευθείας και προσδιορίζει, επίσης, την επίδραση της ανεξάρτητης μεταβλητής στην εξαρτημένη. Για κάθε αύξηση της ανεξάρτητης μεταβλητής κατά μία μονάδα, η εκτιμώμενη μέση τιμή της εξαρτημένης μεταβλητής μεταβάλλεται κατά $\boldsymbol{\beta}$ μονάδες. Για παράδειγμα, αν αυξηθεί το ετήσιο εισόδημα κατά μία μονάδα (δηλαδή, \$10.000), τότε η μέση μηνιαία δαπάνη με τη χρήση πιστωτικής κάρτας θα αυξηθεί κατά 45,17489 ή \$4.517,489.

Στο κάτω μέρος των αποτελεσμάτων της **Εικόνας 5.11** εμφανίζονται τα ακόλουθα διαγνωστικά μέτρα της εκτίμησης:

- **R-squared**: Είναι ο συντελεστής προσδιορισμού R^2, ο οποίος δείχνει το ποσοστό της μεταβλητότητας των δεδομένων που εξηγείται από την εκτιμημένη παλινδρόμηση. Στην εκτίμησή μας το R^2 είναι 0.079019, γεγονός που σημαίνει ότι το υπόδειγμά μας ερμηνεύει μόνο το 7.9019% της μεταβλητότητας των δεδομένων.
- **Adjusted R-squared**: Είναι ο διορθωμένος συντελεστής προσδιορισμού $\bar{R}^2$, ο οποίος λαμβάνει υπόψη και το μέγεθος του δείγματος.
- **S.E of regression**: Είναι το τυπικό σφάλμα (standard error) της παλινδρόμησης.
- **Sum squared resid**: Είναι το άθροισμα τετραγώνων των καταλοίπων.
- **Log likelihood**: Είναι η τιμή της συνάρτησης log likelihood, η οποία έχει υπολογιστεί με βάση τους εκτιμημένους συντελεστές.
- **F-statistic**: Είναι η τιμή της F-στατιστικής για τη μηδενική υπόθεση ότι όλοι οι συντελεστές κλίσης (εκτός του σταθερού όρου) είναι μηδέν. Η εναλλακτική υπόθεση στην περίπτωση αυτή είναι ότι έστω ένας από τους συντελεστές κλίσης δεν είναι στατιστικά μηδέν. Η **Prob(F-statistic)** είναι η αντίστοιχη *p*-value για τον έλεγχο αυτό. Καθώς και η συγκεκριμένη *p*-value είναι μικρότερη από το 0,05, αυτό σημαίνει ότι η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας $a = 5\%$. Το αποτέλεσμα αυτό είναι αναμενόμενο, καθώς από την ατομική *t*-στατιστική του εκτιμημένου συντελεστή της μεταβλητής "income" προκύπτει ότι είναι στατιστικά σημαντικός για $a = 5\%$.
- **Mean dependent var**: Είναι ο μέσος της εξαρτημένης μεταβλητής.
- **S.D dependent var**: Είναι η τυπική απόκλιση (standard deviation) της εξαρτημένης μεταβλητής.
- **Akaike info criterion, Schwarz criterion και Hannan-Quinn criter**: Τα τρία αυτά κριτήρια αφορούν την επιλογή του κατάλληλου υποδείγματος. Δεν θα προβούμε σε περαιτέρω ανάλυση των κριτηρίων αυτών (καθώς αυτή ξεφεύγει από τους στόχους του παρόντος κεφαλαίου), απλά θα αναφέρουμε ότι επιλέγουμε το υπόδειγμα εκείνο στο οποίο τα κριτήρια αυτά παίρνουν τη μικρότερη τιμή.
- **Durbin-Watson stat**: Η συγκεκριμένη στατιστική ελέγχει την ύπαρξη σειριακής συσχέτισης πρώτου βαθμού στα κατάλοιπα. Ο έλεγχος γίνεται χρησιμοποιώντας την εκτιμημένη τιμή της στατιστικής αυτής και τις αντίστοιχες θεωρητικές τιμές που προκύπτουν από τους πίνακες που βρίσκονται στο τέλος κάθε βιβλίου Στατιστικής ή Οικονομετρίας. Εμπειρικά, όταν η τιμή της στατιστικής Durbin-Watson είναι κοντά στο 2, μπορούμε να συμπεράνουμε ότι δεν υπάρχει σειριακή συσχέτιση πρώτου βαθμού στα κατάλοιπα.

Τα αποτελέσματα της εκτιμημένης γραμμικής παλινδρόμησης **(Εικόνα 5.11)** μπορούν να αποθηκευτούν και ως πίνακας στο Eviews workfile, πατώντας το κουμπί **Freeze** που βρίσκεται στο toolbar της eq01. Στον πίνακα που θα εμφανιστεί πατάμε το κουμπί **Name** από το toolbar του, προκειμένου να του δώσουμε κάποιο όνομα (έστω **table01**), και στη συνέχεια επιλέγουμε **OK**. Ο συγκεκριμένος πίνακας έχει πλέον αποθηκευτεί Eviews workfile με το εικονίδιο στα αριστερά του, το οποίο υποδηλώνει πίνακα.

Στη συνέχεια, έχοντας ανοικτή την eq01 και επιλέγοντας **Proc → Specify/Estimate** (είτε από το μενού του Eviews είτε από το toolbar της eq01), εμφανίζεται το παράθυρο "Equation Estimation" **(Εικόνα 5.10)**, όπου μπορούμε να τροποποιήσουμε τις παραμέτρους της εξίσωσης που θέλουμε να εκτιμήσουμε. Το

παράθυρο “Equation Estimation” εμφανίζεται και όταν πατήσουμε το κουμπί **Estimate** από το toolbar της eq01.

Είναι σημαντικό να επισημάνουμε στο σημείο αυτό πως, αν τροποποιήσουμε τις παραμέτρους της eq01, προκειμένου να την επανεκτιμήσουμε, η σειρά **resid** στο Eviews workfile θα επικαιροποιηθεί αυτόματα με τη νέα εκτίμηση. Οπότε, αν θέλουμε να δημιουργήσουμε και να αποθηκεύσουμε τα κατάλοιπα από την αρχική εκτίμηση, επιλέγουμε αρχικά **Proc → Make Residual Series** (είτε από το μενού του Eviews είτε από το toolbar της eq01). Στο παράθυρο “Make Residuals” που εμφανίζεται **(Εικόνα 5.12)**, μπορούμε να δημιουργήσουμε τα κανονικά (ordinary), τα τυποποιημένα (standardized) και τα γενικευμένα (generalized) κατάλοιπα και να δώσουμε κάποιο όνομα στη συγκεκριμένη σειρά (έστω **resid01**). Στην περίπτωση που η εκτίμηση της γραμμικής παλινδρόμησης έχει γίνει με τη μέθοδο των ελαχίστων τετραγώνων (όπως στο παράδειγμά μας), οι επιλογές “standardized” και “generalized” εμφανίζονται αχνές, καθώς το Eviews επιτρέπει τη δημιουργία μόνο των κανονικών καταλοίπων. Πατώντας **OK**, η συγκεκριμένη σειρά δημιουργείται αυτόματα στο Eviews workfile. Εναλλακτικά, τα κανονικά κατάλοιπα μπορούν να δημιουργηθούν, αν γράψουμε στο **Command line**

eq01.makeresid resid01

και στη συνέχεια πατήσουμε **Enter**.

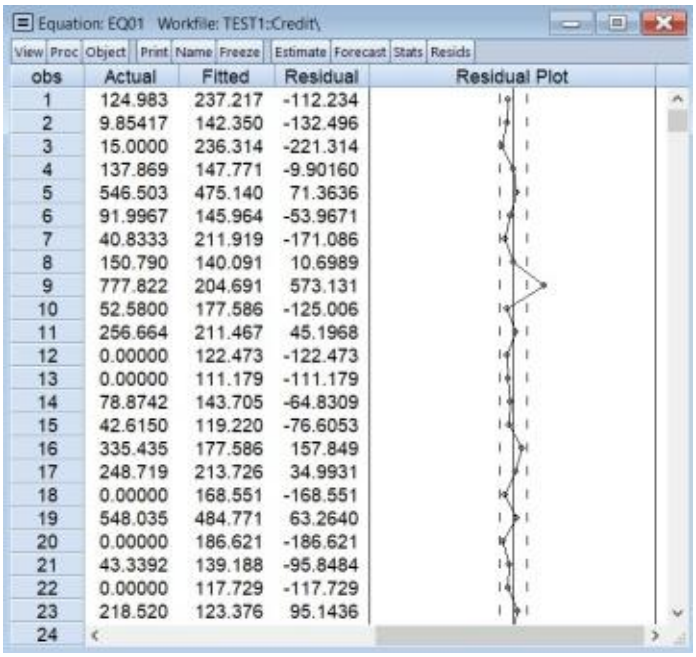
Εικόνα 5.12 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Τα τυποποιημένα κατάλοιπα, τα οποία προκύπτουν διαιρώντας τα κανονικά κατάλοιπα με την τυπική τους απόκλιση, μπορούμε εύκολα να τα υπολογίσουμε με την ακόλουθη διαδικασία. Προκειμένου να υπολογίσουμε την τυπική τους απόκλιση, «ανοίγουμε» με διπλό κλικ τη σειρά **resid01** και επιλέγουμε **View → Descriptive Statistics & Tests → Stats Table**. Όπως προκύπτει, η τυπική τους απόκλιση είναι 261,2423. Στη συνέχεια, δημιουργούμε μια νέα σειρά με τον τρόπο που έχουμε ήδη περιγράψει παραπάνω (έστω **resid02**), όπου στο παράθυρο “Generate Series by Equation” **(Εικόνα 3.29)** γράφουμε **resid02=resid01/261.2423** και πατάμε **OK**. Η σειρά των τυποποιημένων καταλοίπων έχει πλέον δημιουργηθεί.

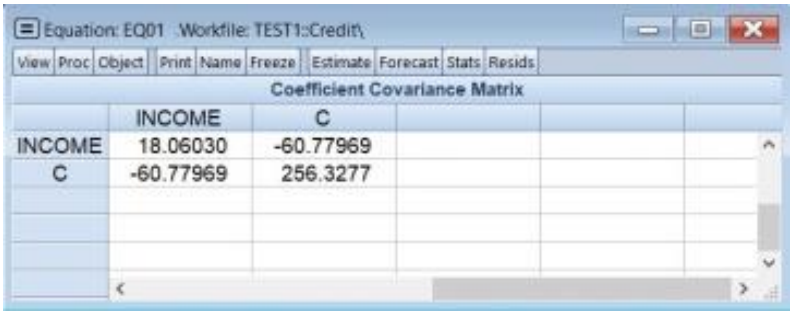
Στη συνέχεια, έχοντας ανοικτή την eq01 και επιλέγοντας **View** (είτε από το μενού του Eviews είτε από το toolbar της eq01), μπορούμε να πάρουμε μια σειρά από πληροφορίες, διαγνωστικά μέτρα και αποτελέσματα στατιστικών ελέγχων σχετικά με την εκτιμημένη γραμμική παλινδρόμηση. Πιο συγκεκριμένα, μεταξύ άλλων:

- **Actual, Fitted, Residual → Actual, Fitted, Residual Table**: Εμφανίζεται ένας πίνακας που παρουσιάζει τις πραγματικές τιμές της εξαρτημένης μεταβλητής “expenditure” (Actual), τις εκτιμημένες τιμές της (Fitted), καθώς και τα κανονικά κατάλοιπα (Residual) που αποτελούν τη διαφορά τους **(Εικόνα 5.13)**.
- **Actual, Fitted, Residual → Actual, Fitted, Residual Graph**: Εμφανίζονται σε διάγραμμα οι σειρές που παρουσιάζονται στον πίνακα της **Εικόνας 5.13**.
- **Actual, Fitted, Residual → Residual Graph**: Εμφανίζεται ένα διάγραμμα που παρουσιάζει τα κατάλοιπα της εκτιμημένης γραμμικής παλινδρόμησης.
- **Actual, Fitted, Residual → Standardized Residual Graph**: Εμφανίζεται ένα διάγραμμα που παρουσιάζει τα τυποποιημένα κατάλοιπα της εκτιμημένης γραμμικής παλινδρόμησης.

- **Covariance Matrix**: Εμφανίζεται η μήτρα διακυμάνσεων – συνδιακυμάνσεων των εκτιμημένων συντελεστών της παλινδρόμησης **(Εικόνα 5.14)**, η οποία είναι συμμετρική. Στην κύρια διαγώνιό της εμφανίζονται οι διακυμάνσεις, ενώ στα κελιά εκτός της διαγωνίου οι συνδιακυμάνσεις.
- **Coefficient Diagnostics → Scaled Coefficients**: Παρουσιάζονται οι εκτιμημένες τιμές των συντελεστών, όταν *τυποποιήσουμε* τις μεταβλητές της γραμμικής παλινδρόμησης *(standardized* ή *scaled coefficients)*. Τα αποτελέσματα παρουσιάζονται στον πίνακα της **Εικόνας 5.15**. Με βάση την τυποποιημένη κλίση και καθώς η τυπική απόκλιση της μεταβλητής "expenditure" είναι 272,2189 και της μεταβλητής "income" 1,693902, αν αυξηθεί η μεταβλητή "income" κατά 1,693902 μονάδες, τότε η εξαρτημένη μεταβλητή "expenditure" αναμένεται να αυξηθεί κατά $0{,}281104 \times 272{,}2189 = 76{,}5218$ μονάδες. Θα πρέπει να επισημανθεί ότι η τιμή του σταθερού όρου δεν εμφανίζεται (**NA**), καθώς έχει τυποποιηθεί και η εξαρτημένη μεταβλητή. Επίσης, στον ίδιο πίνακα παρουσιάζονται και οι ελαστικότητες των μεταβλητών στον μέσο.



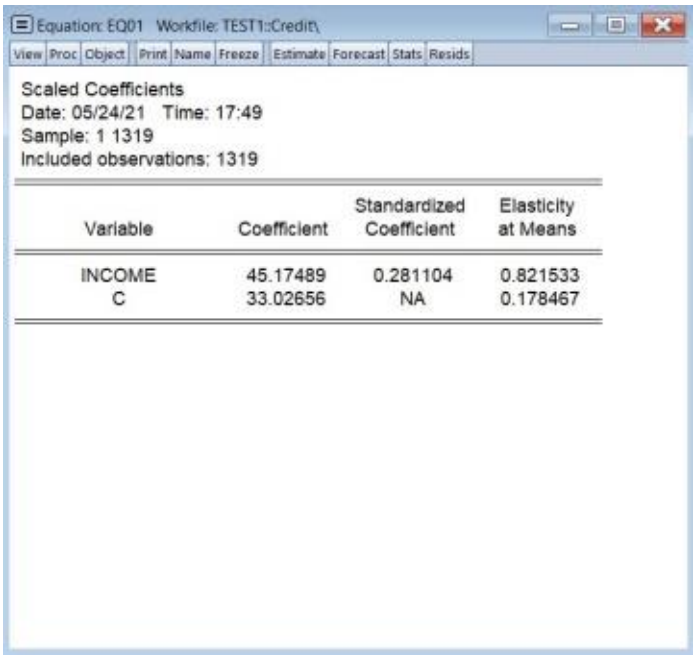
Εικόνα 5.13 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



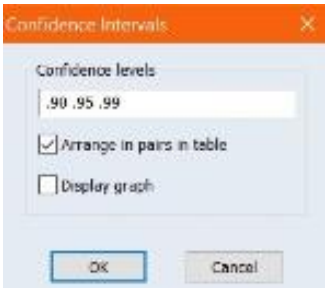
Εικόνα 5.14 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

- **Coefficient Diagnostics → Confidence Intervals**: Υπολογίζονται τα διαστήματα εμπιστοσύνης για τους εκτιμημένους συντελεστές της γραμμικής παλινδρόμησης. Κάνοντας τη συγκεκριμένη επιλογή, θα εμφανιστεί ένα παράθυρο **(Εικόνα 5.16)**, στο οποίο θα πρέπει να επιλέξουμε ποια διαστήματα εμπιστοσύνης θα υπολογιστούν. Το Eviews έχει προεπιλεγμένα τα 90%, 95% και 99% διαστήματα εμπιστοσύνης, αλλά εμείς μπορούμε να επιλέξουμε όποια θέλουμε.

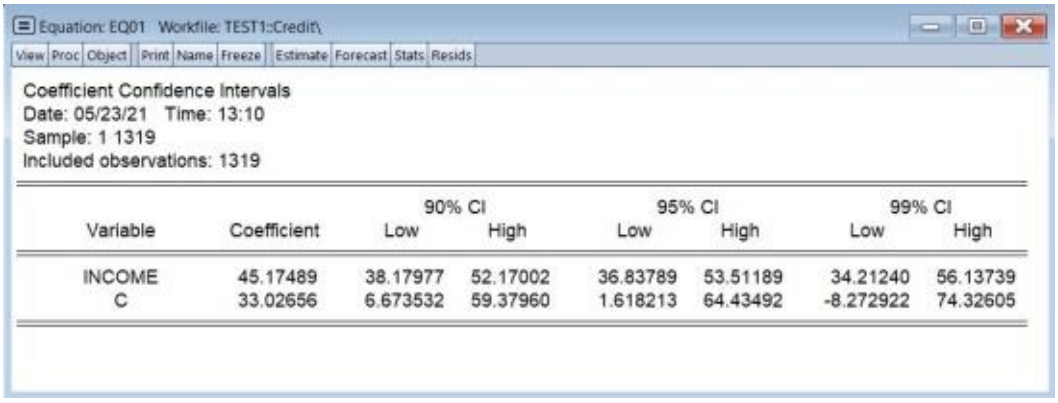
Πατώντας **OK**, εμφανίζονται τα διαστήματα εμπιστοσύνης για τους εκτιμημένους συντελεστές $\boldsymbol{\beta}$ και $\boldsymbol{\alpha}$ **(Εικόνα 5.17)**.



Εικόνα 5.15 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Εικόνα 5.16 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

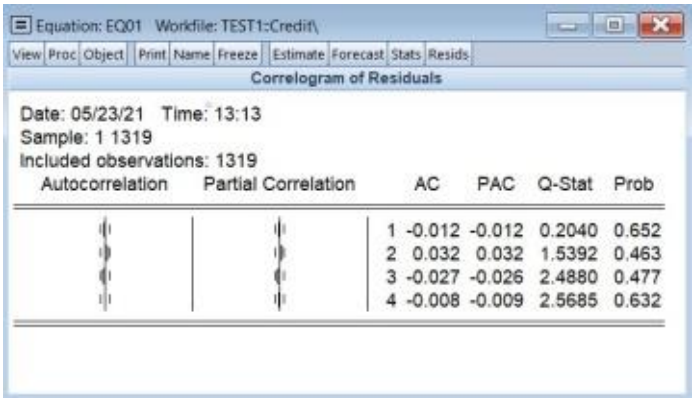


Εικόνα 5.17 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

- **Residuals Diagnostics → Correlogram – Q-statistics**: Η συγκεκριμένη επιλογή μας επιτρέπει να ελέγξουμε αν υπάρχει συνολική αυτοσυσχέτιση (**Autocorrelation - AC**) ή μερική αυτοσυσχέτιση (**Partial Correlation - PAC**) στα κατάλοιπα, ενώ υπολογίζει και τη στατιστική ελέγχου **Q-statistic**. *Γενικά, οι έλεγχοι για αυτοσυσχέτιση ή σειριακή συσχέτιση στα κατάλοιπα εφαρμόζονται, όταν το δείγμα μας αποτελείται από χρονολογικές σειρές.* Για αυτό και όταν κάνουμε τη συγκεκριμένη επιλογή, εμφανίζεται το παράθυρο της **Εικόνας 5.18**, στο οποίο

πρέπει να προσδιορίσουμε τον αριθμό των χρονικών υστερήσεων στα κατάλοιπα, προκειμένου να γίνουν οι παραπάνω έλεγχοι (έστω 4 στο παράδειγμά μας). Στην **Εικόνα 5.19** εμφανίζονται τα αποτελέσματα για καθεμία από τις 4 χρονικές υστερήσεις. Οι διακεκομμένες γραμμές στα διαγράμματα της AC και της PAC αποτελούν το εύρος 2 τυπικών σφαλμάτων. Αν οι τιμές των AC και PAC (που εμφανίζονται δίπλα) είναι μέσα σε αυτά τα εύρη (όπως συμβαίνει στο παράδειγμά μας), τότε μπορούμε να συμπεράνουμε ότι δεν υπάρχει αυτοσυσχέτιση σε επίπεδο σημαντικότητας $a = 5\%$. Οι τιμές της AC αφορούν την πιθανή ύπαρξη αυτοσυσχέτισης μέχρι και την αντίστοιχη χρονική υστέρηση, ενώ οι τιμές της PAC αφορούν την πιθανή ύπαρξη αυτοσυσχέτισης για τη δεδομένη χρονική υστέρηση, έχοντας αφαιρέσει την αυτοσυσχέτιση από τις ενδιάμεσες υστερήσεις. Τέλος, εμφανίζονται οι τιμές της Q-statistic και οι αντίστοιχες *p*-values. Η συγκεκριμένη στατιστική έχει μηδενική υπόθεση ότι δεν υπάρχει αυτοσυσχέτιση στα κατάλοιπα μέχρι και την αντίστοιχη χρονική υστέρηση, ενώ κατανέμεται ασυμπτωτικά ως χ^2. Καθώς όλες οι *p*-values είναι μεγαλύτερες του 0,05, η μηδενική υπόθεση δεν μπορεί να απορριφθεί σε επίπεδο σημαντικότητας $a = 5\%$ και άρα μπορούμε να συμπεράνουμε ότι δεν υπάρχει αυτοσυσχέτιση στα κατάλοιπα μέχρι και την τέταρτη χρονική υστέρηση. Η επιλογή **Residuals Diagnostics → Correlogram Squared Residuals** πραγματοποιεί ακριβώς τους ίδιους ελέγχους για τα τετράγωνα των καταλοίπων.

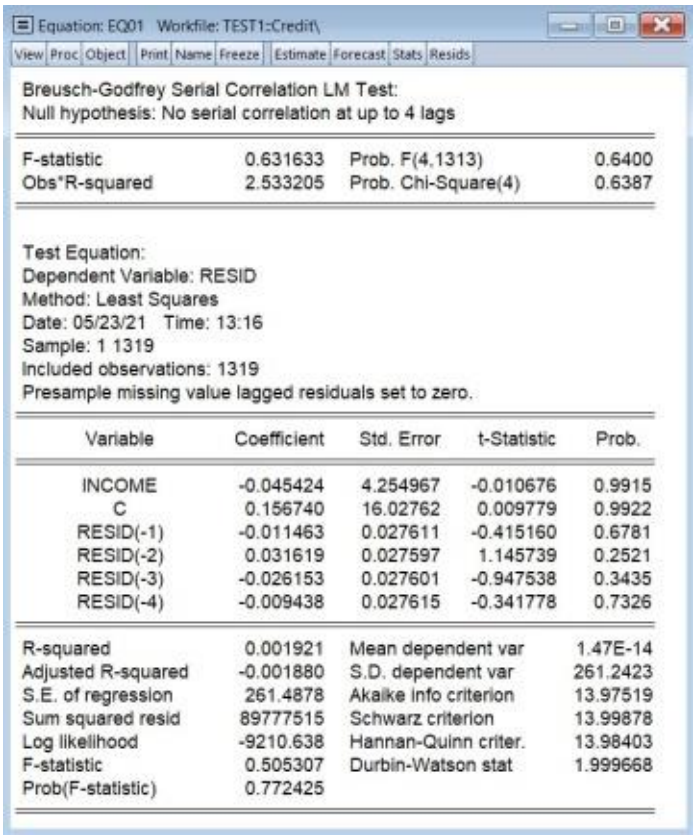
Εικόνα 5.18 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Εικόνα 5.19 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

- **Residuals Diagnostics → Histogram – Normality Test**: Εμφανίζονται το ιστόγραμμα και τα βασικά περιγραφικά μέτρα για τα κατάλοιπα, καθώς και η τιμή της στατιστικής ελέγχου κανονικότητας Jarque-Bera. Φυσικά, αν θέλουμε να υπολογίσουμε και τις υπόλοιπες στατιστικές ελέγχου για τα κατάλοιπα, ακολουθούμε τη διαδικασία που περιγράψαμε στην ενότητα 4.1.
- **Residuals Diagnostics → Serial Correlation LM Test**: Η συγκεκριμένη επιλογή μας επιτρέπει να ελέγξουμε αν υπάρχει σειριακή συσχέτιση στα κατάλοιπα, *στην περίπτωση που το δείγμα μας αποτελείται από χρονολογικές σειρές*. Χρησιμοποιείται η στατιστική ελέγχου **Breusch-Godfrey**, της οποίας η μηδενική υπόθεση είναι ότι δεν υπάρχει σειριακή συσχέτιση στα κατάλοιπα μέχρι και τη χρονική υστέρηση που επιλέγουμε, ενώ κατανέμεται ασυμπτωτικά ως

χ^2. Η συγκεκριμένη στατιστική υπολογίζεται με την ακόλουθη διαδικασία. Η διαδικασία ελέγχου είναι η ακόλουθη. Αρχικά εκτιμάται μια βοηθητική (auxiliary) παλινδρόμηση με τη μέθοδο των ελαχίστων τετραγώνων, όπου εξαρτημένη μεταβλητή είναι τα εκτιμημένα κατάλοιπα, ενώ ανεξάρτητες μεταβλητές είναι αυτές του αρχικού υποδείγματος, καθώς και τα εκτιμημένα κατάλοιπα με όσες χρονικές υστερήσεις έχουμε επιλέξει. Στη συνέχεια, η τιμή της στατιστικής ελέγχου προκύπτει από το γινόμενο $N \times R^2$, όπου N είναι ο αριθμός των παρατηρήσεων και R^2 ο συντελεστής προσδιορισμού της βοηθητικής παλινδρόμησης. Κάνοντας τη συγκεκριμένη επιλογή στο παράδειγμά μας, εμφανίζεται το παράθυρο της **Εικόνας 5.18**, στο οποίο πρέπει να προσδιορίσουμε τον αριθμό των χρονικών υστερήσεων στα κατάλοιπα (έστω 4). Πατώντας **OK**, προκύπτουν τα αποτελέσματα **(Εικόνα 5.20)**. Στο πάνω μέρος των αποτελεσμάτων, εμφανίζονται η μηδενική υπόθεση (**Null hypothesis**), η τιμή της στατιστικής Breusch-Godfrey (**Obs*R-squared**), καθώς και η αντίστοιχη *p*-value. Το Eviews εμφανίζει και την τιμή της *F*-στατιστικής μαζί με την αντίστοιχη *p*-value. Θα πρέπει να επισημανθεί στο σημείο αυτό πως η συγκεκριμένη *F*-στατιστική δεν έχει γνωστή κατανομή, αλλά χρησιμοποιείται πολλές φορές ως άτυπος έλεγχος της μηδενικής υπόθεσης. Καθώς η *p*-value της στατιστικής Breusch-Godfrey είναι 0,6387 > 0,05, η μηδενική υπόθεση δεν μπορεί να απορριφθεί σε επίπεδο σημαντικότητας $a = 5\%$. Συνεπώς, δεν υπάρχει αυτοσυσχέτιση στα κατάλοιπα μέχρι και την τέταρτη χρονική υστέρηση. Στο κάτω μέρος των αποτελεσμάτων της **Εικόνας 5.20** εμφανίζεται η εκτιμημένη βοηθητική παλινδρόμηση, μαζί με τα βασικά διαγνωστικά της μέτρα.


Equation: EQ01 Workfile: TEST1::Credit\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Breusch-Godfrey Serial Correlation LM Test:
Null hypothesis: No serial correlation at up to 4 lags

F-statistic	0.631633	Prob. F(4,1313)	0.6400
Obs*R-squared	2.533205	Prob. Chi-Square(4)	0.6387

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Date: 05/23/21 Time: 13:16
Sample: 1 1319
Included observations: 1319
Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
INCOME	-0.045424	4.254967	-0.010676	0.9915
C	0.156740	16.02762	0.009779	0.9922
RESID(-1)	-0.011463	0.027611	-0.415160	0.6781
RESID(-2)	0.031619	0.027597	1.145739	0.2521
RESID(-3)	-0.026153	0.027601	-0.947538	0.3435
RESID(-4)	-0.009438	0.027615	-0.341778	0.7326

R-squared	0.001921	Mean dependent var	1.47E-14
Adjusted R-squared	-0.001880	S.D. dependent var	261.2423
S.E. of regression	261.4878	Akaike info criterion	13.97519
Sum squared resid	89777515	Schwarz criterion	13.99878
Log likelihood	-9210.638	Hannan-Quinn criter.	13.98403
F-statistic	0.505307	Durbin-Watson stat	1.999668
Prob(F-statistic)	0.772425		

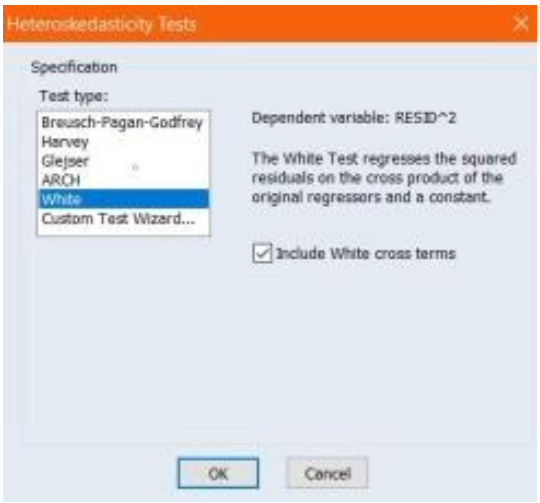
Εικόνα 5.20 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

- **Residuals Diagnostics → Heteroskedasticity Tests**: Η συγκεκριμένη επιλογή μας επιτρέπει να ελέγξουμε αν υπάρχει ετεροσκεδαστικότητα στα κατάλοιπα, *στην περίπτωση που το δείγμα μας αποτελείται από διαστρωματικά στοιχεία*. Το Eviews παρέχει 5 διαφορετικές στατιστικές ελέγχου: **Breusch-Pagan-Godfrey**, **Harvey**, **Glejser**, **ARCH** και **White (Εικόνα 5.21)**. Όλες αυτές οι στατιστικές ελέγχουν τη μηδενική υπόθεση της ομοσκεδαστικότητας στα κατάλοιπα, ενώ

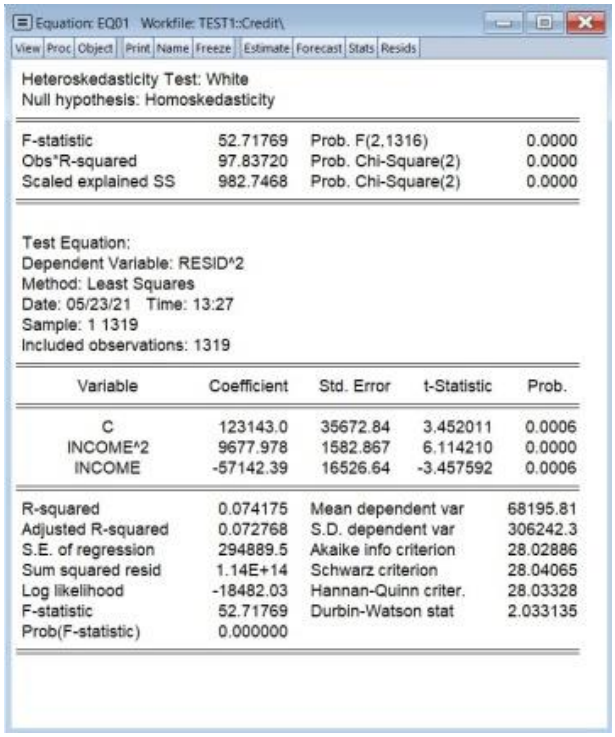
κατανέμονται ασυμπτωτικά ως χ^2. Οι τιμές τους προκύπτουν από το γινόμενο $N \times R^2$, όπου N είναι ο αριθμός των παρατηρήσεων και R^2 ο συντελεστής προσδιορισμού μιας βοηθητικής παλινδρόμησης. Η βοηθητική αυτή παλινδρόμηση είναι διαφορετική σε κάθε στατιστική ελέγχου:

- **Breusch-Pagan-Godfrey**: Εξαρτημένη μεταβλητή είναι το τετράγωνο των εκτιμημένων καταλοίπων, ενώ ανεξάρτητες μεταβλητές είναι αυτές του αρχικού υποδείγματος και ένας σταθερός όρος.
- **Harvey**: Εξαρτημένη μεταβλητή είναι ο φυσικός λογάριθμος του τετραγώνου των εκτιμημένων καταλοίπων, ενώ ανεξάρτητες μεταβλητές είναι αυτές του αρχικού υποδείγματος και ένας σταθερός όρος.
- **Glejser**: Εξαρτημένη μεταβλητή είναι η απόλυτη τιμή των εκτιμημένων καταλοίπων, ενώ ανεξάρτητες μεταβλητές είναι αυτές του αρχικού υποδείγματος και ένας σταθερός όρος.
- **ARCH:** Εξαρτημένη μεταβλητή είναι το τετράγωνο των εκτιμημένων καταλοίπων, ενώ ανεξάρτητες μεταβλητές είναι το τετράγωνο των εκτιμημένων καταλοίπων με όσες χρονικές υστερήσεις επιθυμούμε και ένας σταθερός όρος.
- **White**: Εξαρτημένη μεταβλητή είναι το τετράγωνο των εκτιμημένων καταλοίπων, ενώ ανεξάρτητες μεταβλητές είναι αυτές του αρχικού υποδείγματος στην πρώτη δύναμη και στο τετράγωνο, καθώς και ένας σταθερός όρος. Επίσης, όπως φαίνεται και στην **Εικόνα 5.21**, εμφανίζεται ένα κουτάκι στο οποίο αναφέρεται “Include White cross terms”. Η συγκεκριμένη επιλογή έχει νόημα μόνο στην πολλαπλή παλινδρόμηση, καθώς, αν κάνουμε τικ στο κουτάκι αυτό, το Eviews θα συμπεριλάβει ως ανεξάρτητες μεταβλητές στη βοηθητική παλινδρόμηση και τα γινόμενα μεταξύ των ανεξάρτητων μεταβλητών.

Σε καθέναν από τους παραπάνω ελέγχους, εκτός από την τιμή της αντίστοιχης στατιστικής, το Eviews εμφανίζει την τιμή της *F*-στατιστικής για άτυπο έλεγχο της μηδενικής υπόθεσης (η οποία όπως αναφέρθηκε παραπάνω είναι χωρίς γνωστή κατανομή), καθώς και τη στατιστική **Scale explained SS**, η οποία κατανέμεται ασυμπτωτικά ως χ^2. Πατώντας, λοιπόν, **OK**, στο παράθυρο της **Εικόνας 5.21**, προκύπτουν τα αποτελέσματα του ελέγχου White **(Εικόνα 5.22)**. Στο πάνω μέρος των αποτελεσμάτων, εμφανίζονται η μηδενική υπόθεση (**Null hypothesis**), η τιμή της στατιστικής White (**Obs*R-squared**), καθώς και η αντίστοιχη *p*-value που είναι 0,0000 < 0,05. Οπότε, μπορούμε να εξάγουμε το συμπέρασμα ότι η μηδενική υπόθεση της ομοσκεδαστικότητας απορρίπτεται σε επίπεδο σημαντικότητας $a = 5\%$. Στο κάτω μέρος των αποτελεσμάτων εμφανίζεται η εκτιμημένη βοηθητική παλινδρόμηση, μαζί με τα βασικά διαγνωστικά της μέτρα.



Εικόνα 5.21 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Εικόνα 5.22 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

5.4 Πολλαπλή γραμμική παλινδρόμηση

Όταν έχουμε περισσότερες από μία ανεξάρτητες μεταβλητές και θέλουμε να εξετάσουμε την επίδρασή τους σε μία εξαρτημένη μεταβλητή, χρησιμοποιούμε την πολλαπλή γραμμική παλινδρόμηση. Θα πρέπει να τονίσουμε στο σημείο αυτό ότι, όταν χρησιμοποιούμε το όρο «γραμμική», εννοούμε «γραμμική» ως προς τις παραμέτρους του μοντέλου $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Οπότε, η συνάρτηση της ευθείας των ελαχίστων τετραγώνων στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης θα είναι της μορφής:

$$\boldsymbol{Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + e_i},$$

όπου με $\boldsymbol{p}$ συμβολίζουμε το πλήθος των ανεξάρτητων μεταβλητών, ενώ ο όρος $\boldsymbol{e_i}$ αναφέρεται στο κατάλοιπο της i-οστής τιμής.

Οι υποθέσεις που πρέπει να ικανοποιούνται στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης είναι οι ίδιες με αυτές της απλής παλινδρόμησης. Επιπλέον, μία απαραίτητη προϋπόθεση για όλα γενικά τα υποδείγματα με περισσότερες από μία ανεξάρτητες μεταβλητές είναι η έλλειψη συγγραμμικότητας. Η τελευταία αποτελεί ένα σοβαρό πρόβλημα για την πολλαπλή γραμμική παλινδρόμηση και αφορά την περίπτωση όπου μία ανεξάρτητη μεταβλητή συσχετίζεται με μία άλλη ανεξάρτητη, με αποτέλεσμα μέσω της μίας να μπορούμε να εκτιμήσουμε τις τιμές της άλλης. Επομένως, η ύπαρξη και των δύο μεταβλητών στο υπόδειγμα δεν είναι δυνατή. Σκεφτείτε, για παράδειγμα, την περίπτωση στην οποία έχουμε δύο ανεξάρτητες μεταβλητές, το εισόδημα και την ηλικία, και ενδιαφερόμαστε να δούμε πώς επιδρούν πάνω σε μία εξαρτημένη μεταβλητή που είναι οι δαπάνες. Προφανώς και είναι λογικό να υπάρχει σχέση μεταξύ ηλικίας και εισοδήματος. Αν η μεταξύ τους συσχέτιση είναι πολύ υψηλή (κατά απόλυτη τιμή), τότε δεν χρειάζεται να γνωρίζουμε και τις δύο μεταβλητές, αφού η γνώση της μίας μας είναι αρκετή (αφού μέσω αυτής μπορούμε να εκτιμήσουμε τις τιμές της άλλης). Θα πρέπει να επισημάνουμε στο σημείο αυτό πως η τοποθέτηση «άχρηστων» μεταβλητών στο υπόδειγμα μπορεί φαινομενικά να είναι καλή, όμως ουσιαστικά οδηγεί στο λεγόμενο πρόβλημα της υπερ-προσαρμογής του υποδείγματος. Δηλαδή, στο παράδειγμά μας, αν κρατήσουμε και τις δύο μεταβλητές που αναφέραμε σε ένα υπόδειγμα, ενώ φαινομενικά το βελτιώνουμε, ουσιαστικά το χειροτερεύουμε. Οπότε, ή αφαιρούμε μία εκ των δύο μεταβλητών ή χρησιμοποιούμε την τεχνική της κεντροποίησης των τιμών των

μεταβλητών πριν την πολλαπλή γραμμική παλινδρόμηση. Επίσης, μπορούμε να χρησιμοποιήσουμε και άλλες τεχνικές αντί της γραμμικής παλινδρόμησης. Ένα βασικό μέτρο διάγνωσης της συγγραμμικότητας που παρέχεται τόσο από το SPSS όσο και από το Eviews είναι το **VIF**, το οποίο θα αναλυθεί στη συνέχεια. Εναλλακτικοί τρόποι είναι ο υπολογισμός του συντελεστή γραμμικής συσχέτισης, το Added Variable Plot, καθώς και η παλινδρόμηση ανάμεσα σε ζεύγη ανεξάρτητων μεταβλητών, για τις οποίες υποψιαζόμαστε συγγραμμικότητα.

- Στο **SPSS**: Ο τρόπος με τον οποίο διεξάγουμε την πολλαπλή γραμμική παλινδρόμηση είναι ο ίδιος με την περίπτωση της απλής γραμμικής παλινδρόμησης. Στο παράθυρο της **Εικόνας 5.7** θα περάσουμε δύο ανεξάρτητες μεταβλητές στο λευκό κουτί κάτω από την ένδειξη **Independent(s):** αντί για μία. Έστω ότι η επιπλέον μεταβλητή είναι αυτή που αναφέρεται στην ηλικία των συμμετεχόντων ("age"). Επιπλέον, επιλέγοντας **Statistics** θα εμφανιστεί το παράθυρο της **Εικόνας 5.9**, το οποίο μας δίνει τη δυνατότητα να υπολογίσουμε τα διαγνωστικά μέτρα της συγγραμμικότητας (**Collinearity diagnostics**). Επίσης, αν επιλέξουμε **Confidence Intervals**, θα εμφανιστούν τα 95% διαστήματα εμπιστοσύνης για τις εκτιμήσεις των παραμέτρων. Πατώντας **Continue** και στη συνέχεια **OK**, θα εμφανιστούν τα αποτελέσματα στο Output του SPSS σε τέσσερις πίνακες **(Πίνακες 5.5-5.8)**.

Ο **Πίνακας 5.5** παρουσιάζει τους συντελεστές προσδιορισμού, ενώ ο **Πίνακας 5.6** την ανάλυση της διακύμανσης στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης.

Πίνακας 5.5 *Συντελεστής προσδιορισμού πολλαπλής παλινδρόμησης.*

Model Summary[b]				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.292a	.086	.084	260.5151243415
a. Predictors: (Constant), Age, Income				
b. Dependent Variable: Expenditure				

Πίνακας 5.6 *Πίνακας ανάλυσης διακύμανσης πολλαπλής παλινδρόμησης.*

ANOVA[a]						
Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	8353478.164	2	4176739.082	61.542	.000[b]
	Residual	89314459.094	1316	67868.130		
	Total	97667937.258	1318			
a. Dependent Variable: Expenditure						
b. Predictors: (Constant), Age, Income						

Ο **Πίνακας 5.7** περιέχει τις εκτιμήσεις του υποδείγματος και είναι ίδιος με αυτόν στην περίπτωση της απλής γραμμικής παλινδρόμησης.

Πίνακας 5.7 *Εκτιμήσεις παραμέτρων πολλαπλής παλινδρόμησης.*

Coefficients[a]										
		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Collinearity Statistics	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	94.089	25.548		3.683	.000	43.969	144.208		
	Income	49.626	4.479	.309	11.080	.000	40.839	58.412	.895	1.118
	Age	-2.289	.748	-.085	-3.061	.002	-3.757	-.822	.895	1.118
a. Dependent Variable: Expenditure										

Οπότε, το υπόδειγμα που προσαρμόστηκε στα δεδομένα μας είναι το ακόλουθο:

$$\widehat{Expenditure} = 94.089 + 49.626 * Income - 2.289 * Age.$$

Όπως φαίνεται στον **Πίνακα 5.7**, όλοι οι εκτιμημένοι συντελεστές είναι στατιστικά σημαντικοί, ενώ η ερμηνεία τους είναι παρόμοια με την περίπτωση της απλής γραμμικής παλινδρόμησης, όπου η ανεξάρτητη μεταβλητή είναι μόνο μία. Η σταθερά (94,089) είναι η τιμή στην οποία η ευθεία των ελαχίστων τετραγώνων τέμνει τον κατακόρυφο άξονα συντεταγμένων. Ο συντελεστής του εισοδήματος (49,626) δείχνει το πόσο θα αυξηθεί η αναμενόμενη τιμή των δαπανών, αν αυξηθεί το εισόδημα κατά μία μονάδα, διατηρώντας σταθερή την επίδραση της ηλικίας (ceteris paribus). Αποτελεί, δηλαδή, την κύρια επίδραση του εισοδήματος πάνω στις δαπάνες. Ο συντελεστής της ηλικίας (-2,289) αναφέρεται στην κύρια επίδραση της ηλικίας πάνω στα δαπάνες. Για κάθε αύξηση της ηλικίας κατά μία μονάδα, η αναμενόμενη μέση δαπάνη μειώνεται κατά 2,289 μονάδες, διατηρώντας σταθερή την επίδραση του εισοδήματος.

Οι δύο τελευταίες στήλες του **Πίνακα 5.7**, καθώς και ο **Πίνακας 5.8**, αναφέρονται σε διαγνωστικά συγγραμμικότητας. Το **VIF** (Variation Inflation Factor) αποτελεί ένα βασικό μέτρο διάγνωσης της συγγραμμικότητας, στο οποίο τιμές μεγαλύτερες του 2 αποτελούν ένδειξη ότι υπάρχει πρόβλημα συγγραμμικότητας. Σχετικά με το μέτρο του **Tolerance**, η τιμή του για μία μεταβλητή εκφράζει το ποσοστό της διακύμανσης της μεταβλητής αυτής που εξηγείται από τις υπόλοιπες ανεξάρτητες μεταβλητές του υποδείγματος. Πιο συγκεκριμένα, το ποσοστό αυτό είναι ίσο με (1-Tolerance)%. Οπότε, τιμές της Tolerance μικρότερες του 0,5 αποτελούν ένδειξη ότι υπάρχει πρόβλημα συγγραμμικότητας. Όπως φαίνεται στον **Πίνακα 5.7**, οι τιμές των μέτρων αυτών υποδηλώνουν ότι οι δύο ανεξάρτητες μεταβλητές "Income" και 'age" δεν δημιουργούν πρόβλημα συγγραμμικότητας στο υπόδειγμα. Ο **Πίνακας 5.8** παρουσιάζει τα **Variance Proportions** για τις ανεξάρτητες μεταβλητές του υποδείγματος, τα οποία αποτελούν ένα επιπλέον διαγνωστικό μέτρο του προβλήματος της συγγραμμικότητας. Τιμές μεγαλύτερες του 15 φανερώνουν πιθανό πρόβλημα συγγραμμικότητας, ενώ τιμές άνω του 30 υποδηλώνουν σοβαρό πρόβλημα συγγραμμικότητας.

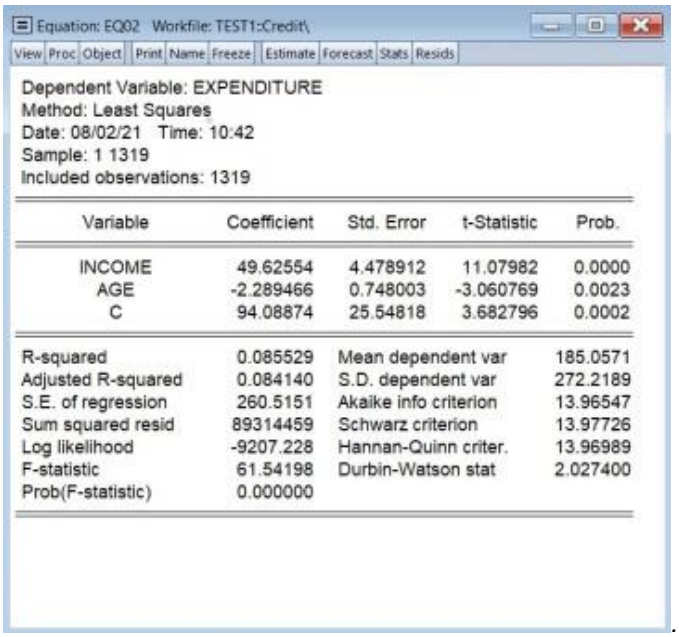
Πίνακας 5.8 *Διαγνωστικά συγγραμμικότητας για την πολλαπλή παλινδρόμηση.*

Collinearity Diagnostics[a]						
				Variance Proportions		
Model	**Dimension**	Eigenvalue	Condition Index	(Constant)	Income	Age
1	1	2.831	1.000	.01	.02	.01
	2	.125	4.759	.11	.98	.09
	3	.044	8.067	.88	.00	.90
a. Dependent Variable: Expenditure						

Οι υποθέσεις σχετικά με την ισχύ του υποδείγματος είναι προφανώς οι ίδιες σε σχέση με την περίπτωση της μίας ανεξάρτητης μεταβλητής και ελέγχονται με τις ίδιες μεθόδους.

- Στο **Eviews**: Για να εκτιμήσουμε μια πολλαπλή γραμμική παλινδρόμηση ακολουθούμε ακριβώς την ίδια διαδικασία με την απλή παλινδρόμηση. Έστω, λοιπόν, ότι, όπως και στην περίπτωση του SPSS, θέλουμε να εκτιμήσουμε την ευθεία γραμμικής παλινδρόμησης μεταξύ της εξαρτημένης (dependent) μεταβλητής "expenditure" και των ανεξάρτητων (independent) μεταβλητών "income" και "age". Όπως αναφέρθηκε παραπάνω, το υπόδειγμά μας θα είναι πλέον της μορφής $\boldsymbol{y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e_i}$. Οπότε, δημιουργούμε μια νέα εξίσωση στο Eviews workfile με τον τρόπο που δείξαμε στην προηγούμενη ενότητα (έστω eq02) και την εκτιμάμε με τη μέθοδο των ελαχίστων τετραγώνων. Από τα αποτελέσματα **(Εικόνα 5.23)** προκύπτει ότι τόσο οι εκτιμημένοι συντελεστές των μεταβλητών "income" και "age" όσο και ο σταθερός όρος είναι στατιστικά σημαντικοί σε επίπεδο σημαντικότητας $a = 5\%$, καθώς οι αντίστοιχες *p*-values είναι μικρότερες από το 0,05. Οπότε, η εκτιμημένη γραμμική παλινδρόμηση θα είναι της μορφής:

$$\widehat{Expenditure} = 94.08874 + 49.62554 * Income - 2.289466 * Ag$$



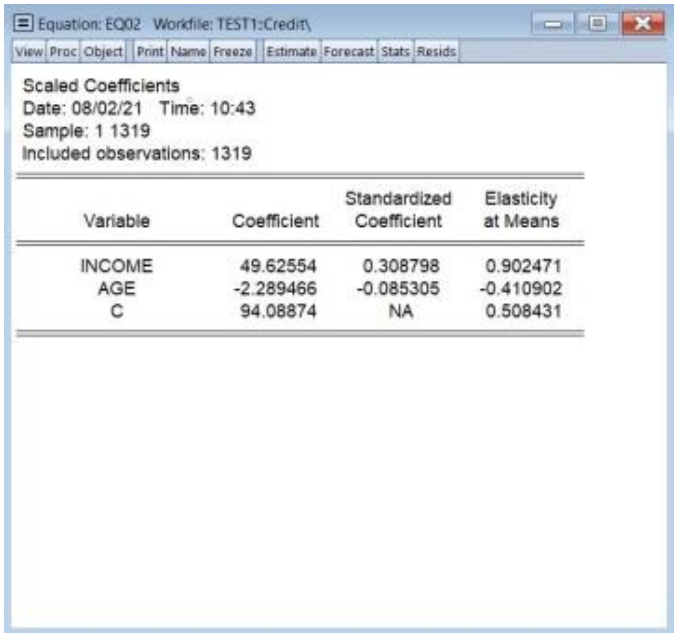
Εικόνα 5.23 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Στο κάτω μέρος των αποτελεσμάτων εμφανίζονται τα βασικά διαγνωστικά μέτρα της εκτιμημένης γραμμικής παλινδρόμησης, όπου το R^2 έχει αυξηθεί με την προσθήκη της μεταβλητής "age" ως ανεξάρτητης μεταβλητής και είναι πλέον 0,085529. Πλέον, το υπόδειγμά μας ερμηνεύει το 8,5529% της μεταβλητότητας των δεδομένων. Όπως και προηγουμένως, από το **View → Coefficient Diagnostics → Scaled Coefficients** προκύπτουν οι τυποποιημένοι συντελεστές *(standardized* ή *scaled coefficients)* της γραμμικής παλινδρόμησης **(Εικόνα 5.24)**. Όπως προκύπτει από τα βασικά περιγραφικά μέτρα, η τυπική απόκλιση της μεταβλητής "expenditure" είναι 272,2189, της μεταβλητής "income" 1,693902 και της μεταβλητής "age" 10,14278. Οπότε, αν αυξηθεί η μεταβλητή "income" κατά 1,693902 μονάδες, τότε η εξαρτημένη μεταβλητή "expenditure" αναμένεται να αυξηθεί κατά $0{,}308798 \times 272{,}2189 = 84{,}0607$ μονάδες. Ενώ, αν αυξηθεί η μεταβλητή "age" κατά 10,14278 μονάδες, τότε η εξαρτημένη μεταβλητή "expenditure" αναμένεται να μεταβληθεί κατά $-0{,}085305 \times 272{,}2189 = -23{,}2216$ μονάδες. Συνεπώς, η σχετική επίδραση της μεταβλητής "income" είναι, σε απόλυτους όρους, μεγαλύτερη από τη σχετική επίδραση της μεταβλητής "age".

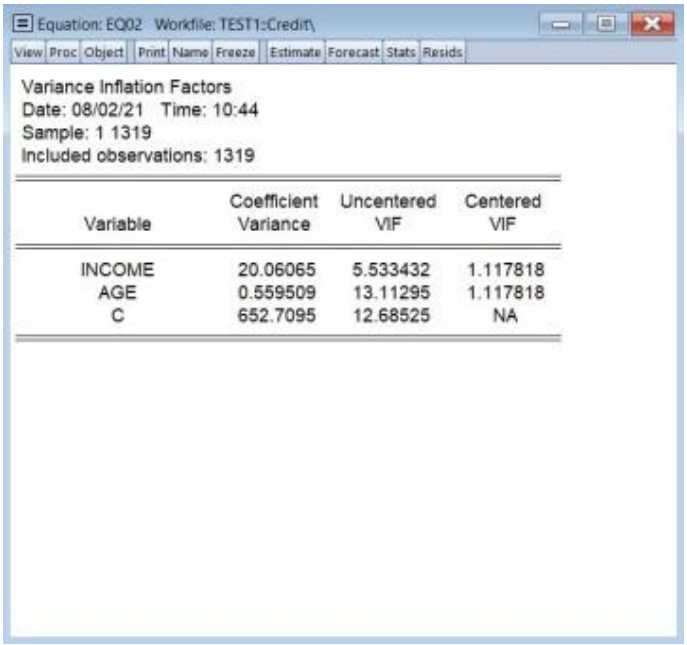
Τα υπόλοιπα διαγνωστικά μέτρα της εκτιμημένης παλινδρόμησης, καθώς και τα αποτελέσματα των στατιστικών ελέγχων σχετικά με τα κατάλοιπα (για κανονικότητα, σειριακή συσχέτιση και ομοσκεδαστικότητα), προκύπτουν με τον ίδιο ακριβώς τρόπο που περιγράψαμε στην προηγούμενη ενότητα. Όμως, όπως αναφέρθηκε παραπάνω, στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης θα πρέπει να γίνει επιπλέον έλεγχος για την πιθανή ύπαρξη συγγραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών. Το βασικό διαγνωστικό μέτρο είναι το **Variance Inflation Factor (VIF)**, το οποίο δείχνει πόσο έχει αυξηθεί η διακύμανση του εκτιμημένου συντελεστή μιας ανεξάρτητης μεταβλητής εξαιτίας της συσχέτισής της με τις άλλες ανεξάρτητες μεταβλητές. Το **VIF** υπολογίζεται, αν διαιρέσουμε τη διακύμανση του εκτιμημένου συντελεστή μιας ανεξάρτητης μεταβλητής με τη διακύμανση του ίδιου συντελεστή, όταν οι υπόλοιπες ανεξάρτητες μεταβλητές δεν περιλαμβάνονται στη γραμμική παλινδρόμηση. Εμπειρικά, τιμές του **VIF** μεγαλύτερες του 2 αποτελούν ένδειξη ότι έχουμε πρόβλημα συγγραμμικότητας.

Οπότε, για την εξίσωση eq02, επιλέγουμε **View → Coefficient Diagnostics → Variance Inflation Factors** και τα αποτελέσματά μας εμφανίζονται στον πίνακα της **Εικόνας 5.25**. Η στήλη **Coefficient Variance** δείχνει τη διακύμανση καθενός από τους εκτιμημένους συντελεστές. Στη συνέχεια, η στήλη **Uncentered VIF** παρουσιάζει τον λόγο της διακύμανσης του συγκεκριμένου εκτιμημένου συντελεστή από την αρχική εξίσωση προς τη διακύμανση του ίδιου συντελεστή, αν στην εξίσωση συμπεριληφθεί μόνο αυτός ο συντελεστής και όχι σταθερός όρος. Τέλος, η στήλη **Centered VIF** παρουσιάζει τον λόγο της διακύμανσης του συγκεκριμένου εκτιμημένου συντελεστή από την αρχική εξίσωση προς τη διακύμανση του ίδιου

συντελεστή, αν στην εξίσωση συμπεριληφθούν μόνο αυτός ο συντελεστής και ένας σταθερός όρος. Καθώς η τιμή του Centered VIF είναι 1,117818 < 2 τόσο για τη μεταβλητή "income" όσο και για τη μεταβλητή "age", μπορούμε να συμπεράνουμε ότι οι δύο αυτές μεταβλητές δεν δημιουργούν πρόβλημα συγγραμμικότητας στην εξίσωση eq02.



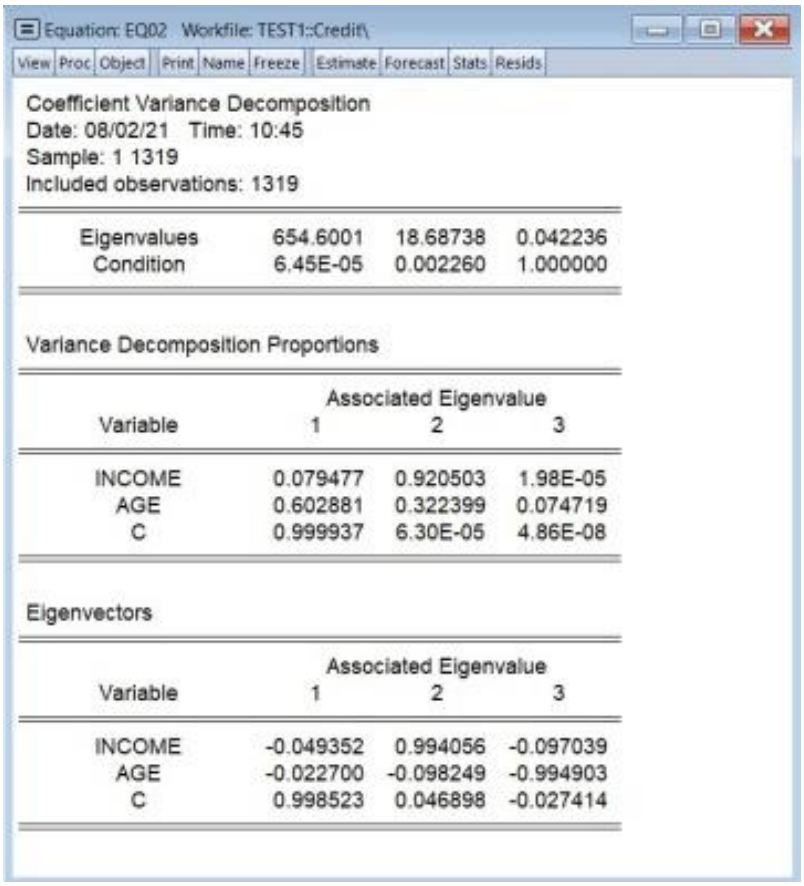
Εικόνα 5.24 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Εικόνα 5.25 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Ένα δεύτερο διαγνωστικό μέτρο για τη συγγραμμικότητα είναι το **Coefficient Variance Decomposition**. Για την εξίσωση eq02, επιλέγουμε **View → Coefficient Diagnostics → Coefficient Variance Decomposition** και τα αποτελέσματά μας εμφανίζονται στον πίνακα της **Εικόνας 5.26**. Στο πάνω μέρος του συγκεκριμένου πίνακα εμφανίζονται τα **Eigenvalues (Ιδιοτιμές)** της εκτιμημένης γραμμικής παλινδρόμησης, ταξινομημένα από τα μεγαλύτερα στη μικρότερα, καθώς και τα **Condition Numbers**. Το τελευταίο condition number είναι πάντα ίσο με το 1. Το γεγονός ότι μια eigenvalue έχει condition number μικρότερο του 0.001 αποτελεί κάποια ένδειξη ότι πιθανόν να υπάρχει πρόβλημα συγγραμμικότητας στην εξίσωση eq02. Στο μέσο του

πίνακα παρουσιάζονται τα αντίστοιχα **Variance Decomposition Proportions**, όπου στην πρώτη στήλη βρίσκονται τα proportions που αντιστοιχούν στο μικρότερο condition number. Καθώς το proportion για τη μεταβλητή "income" είναι 0,079477 < 0,5, ενώ για τη μεταβλητή "age" είναι 0,602881 και άρα λίγο μεγαλύτερο από το 0,5, ενώ απέχει πολύ από το 1, μπορούμε να συμπεράνουμε ότι δεν υπάρχει σημαντικό πρόβλημα συγγραμμικότητας στην εξίσωση eq02. Τέλος, στο κάτω μέρος του πίνακα εμφανίζονται τα αντίστοιχα **Eigenvectors (Ιδιοδιανύσματα)** αποκλειστικά για λόγους σύγκρισης, καθώς δεν προσφέρουν κάποια επιπλέον πληροφορία.


Equation: EQ02 Workfile: TEST1::Credit\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Coefficient Variance Decomposition
Date: 08/02/21 Time: 10:45
Sample: 1 1319
Included observations: 1319

Eigenvalues	654.6001	18.68738	0.042236
Condition	6.45E-05	0.002260	1.000000

Variance Decomposition Proportions

	Associated Eigenvalue		
Variable	1	2	3
INCOME	0.079477	0.920503	1.98E-05
AGE	0.602881	0.322399	0.074719
C	0.999937	6.30E-05	4.86E-08

Eigenvectors

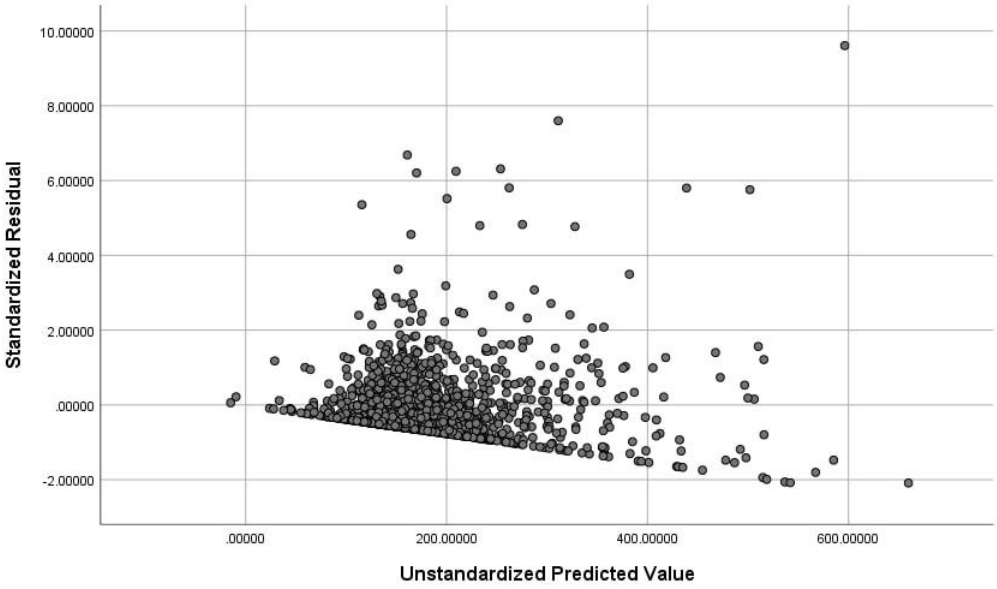
	Associated Eigenvalue		
Variable	1	2	3
INCOME	-0.049352	0.994056	-0.097039
AGE	-0.022700	-0.098249	-0.994903
C	0.998523	0.046898	-0.027414

Εικόνα 5.26 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

5.5 Παραβίαση των υποθέσεων της γραμμικής παλινδρόμησης

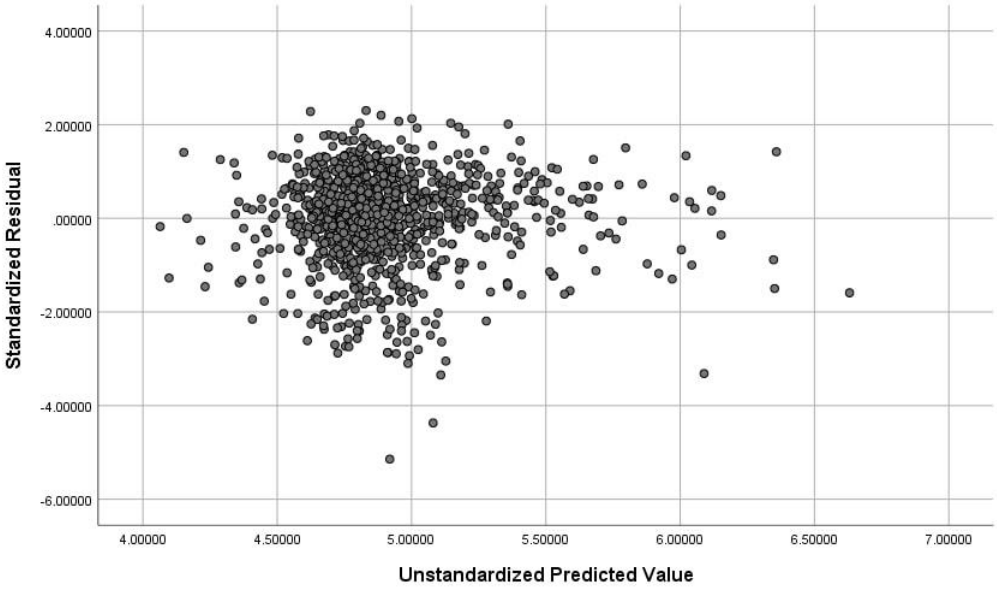
Αναφερθήκαμε προηγουμένως στις υποθέσεις της γραμμικής παλινδρόμησης, οι οποίες είναι η ομοσκεδαστικότητα, η ανεξαρτησία και η κανονικότητα των καταλοίπων. Προκειμένου να ελέγξουμε τις υποθέσεις αυτές, αρχικά εκτιμάμε την πολλαπλή γραμμική παλινδρόμηση με εξαρτημένη μεταβλητή τη μέση δαπάνη ("expenditure") και ανεξάρτητες μεταβλητές το εισόδημα ("income") και την ηλικία ("age"), και στη συνέχεια εργαζόμαστε με τον ακόλουθο τρόπο.

Στο **SPSS**: Αρχικά, κατασκευάζουμε το διάγραμμα διασποράς των τυποποιημένων καταλοίπων με τις εκτιμημένες τιμές της εξαρτημένης μεταβλητής και το αποτέλεσμα παρουσιάζεται στο **Διάγραμμα 5.2**. Παρατηρούμε ότι, καθώς κινούμαστε προς τα δεξιά του συγκεκριμένου διαγράμματος, το εύρος των καταλοίπων απλώνεται, δημιουργώντας με τον τρόπο αυτό ένα «χωνί». Συμπεραίνουμε, λοιπόν, ότι μάλλον υπάρχει πρόβλημα σχετικά με την υπόθεση της σταθερής διακύμανσης των καταλοίπων. Οπότε, διεξάγουμε τον μη παραμετρικό έλεγχο κανονικότητας των Kolmogorov-Smirnov για τα κατάλοιπα και παίρνουμε μία *p*-value που είναι μικρότερη του 0,01. Συνεπώς, υπάρχουν ενδείξεις ότι η κατανομή των καταλοίπων διαφέρει στατιστικά σημαντικά από την κανονική κατανομή και άρα δύο βασικές υποθέσεις της γραμμικής παλινδρόμησης παραβιάζονται.



Διάγραμμα 5.2 *Διάγραμμα διασποράς καταλοίπων με εκτιμημένες τιμές για την πολλαπλή παλινδρόμηση.*

Ένας τρόπος αντιμετώπισης του συγκεκριμένου προβλήματος είναι να μετασχηματίσουμε τις τιμές της εξαρτημένης μεταβλητής "expenditure" και να πάρουμε τον φυσικό τους λογάριθμο. Στο παράθυρο της **Εικόνας 2.9**, επιλέγουμε **Arithmetic** στο **Function group** και μετά πατάμε το βελάκι, προκειμένου η συγκεκριμένη επιλογή να μεταφερθεί προς τα πάνω. Στη συνέχεια, επιλέγουμε τη μεταβλητή που θέλουμε να μετασχηματίσουμε (την "expenditure" στο παράδειγμά μας) και την περνάμε δεξιά μέσα στη συνάρτηση του λογαρίθμου (**LN(Expenditure)**). Στην επιλογή **Target variable** δίνουμε κάποιο όνομα στη νέα μεταβλητή. Τέλος, εκτιμάμε την απλή γραμμική παλινδρόμηση των λογαριθμοποιημένων δαπανών πάνω στο εισόδημα και την ηλικία, και στη συνέχεια κατασκευάζουμε το διάγραμμα διασποράς των τυποποιημένων καταλοίπων με τις νέες εκτιμημένες τιμές **(Διάγραμμα 5.3)**.



Διάγραμμα 5.3 *Διάγραμμα διασποράς καταλοίπων με εκτιμημένες τιμές για την πολλαπλή παλινδρόμηση έχοντας χρησιμοποιήσει τον λογάριθμο των δαπανών.*

Πλέον, το σχήμα του **Διαγράμματος 5.3** δεν είναι πια ένα «χωνί», αλλά το εύρος των τυποποιημένων καταλοίπων έχει σταθεροποιηθεί, καθώς κινούμαστε από τα αριστερά προς τα δεξιά. Το κυκλάκι που φαίνεται στο κάτω μέρος και των δύο διαγραμμάτων αποτελεί ένα ακραίο σημείο. Παρόλα αυτά, ο έλεγχος

κανονικότητας των Kolmogorov-Smirnov δίνει και πάλι μία *p*-value μικρότερη του 0,01. Οπότε, παραμένουν οι ενδείξεις ότι η υπόθεση της κανονικότητας των καταλοίπων δεν ικανοποιείται. Ωστόσο, η παραβίαση της κανονικότητας των καταλοίπων δεν συνιστά σοβαρό πρόβλημα για την εκτίμηση του υποδείγματος, σε αντίθεση με την παραβίαση της ομοσκεδαστικότητας που αποτελεί πολύ σημαντικό πρόβλημα.

Βεβαίως, τα συμπεράσματα που πλέον θα προκύψουν σχετικά με τις παραμέτρους του υποδείγματος θα αφορούν τις λογαριθμοποιημένες τιμές της εξαρτημένης μεταβλητής. Προκειμένου να επιστρέψουμε από τους φυσικούς λογαρίθμους στις κανονικές τιμές, μπορούμε να χρησιμοποιήσουμε τη συνάρτηση **Exp** στην επιλογή **Compute**. Αν αντί των φυσικών λογαρίθμων χρησιμοποιήσουμε λογαρίθμους με βάση το 10, η διαδικασία επιστροφής στις αρχικές τιμές είναι διαφορετική.

Τέλος, όπως φαίνεται στο **Διάγραμμα 5.3**, μία τιμή έχει κατάλοιπο μεγαλύτερο από 5 (σε απόλυτη τιμή), και το αντίστοιχο κυκλάκι εμφανίζεται κάτω αριστερά. Η συγκεκριμένη τιμή δείχνει ότι η παρατήρηση στην οποία αντιστοιχεί είναι ακραίο σημείο. Το πώς αντιμετωπίζουμε τέτοιες περιπτώσεις θα το παραλείψουμε. Ο ενδιαφερόμενος χρήστης μπορεί να εστιάσει στις εύρωστες στατιστικές μεθόδους (robust statistics).

- Στο **Eviews**: Έστω, λοιπόν, ότι θέλουμε να κατασκευάσουμε το διάγραμμα διασποράς μεταξύ των τυποποιημένων καταλοίπων και των εκτιμημένων τιμών της μεταβλητής "expenditure" στην eq02 (πολλαπλή γραμμική παλινδρόμηση), προκειμένου να δούμε τι συμβαίνει με τα κατάλοιπα. Τον τρόπο κατασκευής ενός διαγράμματος διασποράς τον περιγράψαμε στην ενότητα 5.1, ενώ το πώς μπορούμε να υπολογίσουμε τα τυποποιημένα κατάλοιπα το δείξαμε στην ενότητα 5.2. Τη σειρά με τις εκτιμημένες τιμές της eq02 μπορούμε να την δημιουργήσουμε με δύο τρόπους:
- Να δημιουργήσουμε μια νέα σειρά στο Eviews workfile (έστω **expenditure_fit**) και να κάνουμε **copy/paste** σε αυτή τα **fitted values** που προκύπτουν, όταν έχουμε «ανοίξει» την eq02, και να επιλέξουμε **View → Actual, Fitted, Residual → Actual, Fitted, Residual Table**.
- Να γράψουμε στο **Command line** εναλλακτικά μία από τις ακόλουθες εντολές

eq02.fit expenditure_fit

και στη συνέχεια να πατήσουμε **Enter**. Η σειρά των εκτιμημένων τιμών της μεταβλητής "expenditure" στην eq02 θα δημιουργηθεί αυτόματα στο Eviews workfile.

Οπότε, μπορούμε πλέον να κατασκευάσουμε το συγκεκριμένο διάγραμμα διασποράς και να ελέγξουμε τι συμβαίνει με τα κατάλοιπα. Στην περίπτωση που παραβιάζονται οι βασικές υποθέσεις του γραμμικού υποδείγματος, το πρόβλημα συνήθως αντιμετωπίζεται με τον μετασχηματισμό των τιμών της εξαρτημένης μεταβλητής και την επανεκτίμηση του υποδείγματος με τις νέες τιμές. Για παράδειγμα, αν θέλουμε να χρησιμοποιήσουμε την πρώτη διαφορά της μεταβλητής ''expenditure'' ως εξαρτημένη μεταβλητή, μπορούμε εναλλακτικά:

- Να δημιουργήσουμε μια νέα σειρά στο Eviews workfile που θα περιέχει τη μεταβλητή "expenditure" σε πρώτη διαφορά (με τον τρόπο που περιγράψαμε στην ενότητα 2.10).
- Να γράψουμε κατευθείαν στην επιλογή 'Equation specification' της **Εικόνας 5.10 d(expenditure)** και στη συνέχεια **income** και **c**.

Στη συνέχεια, εκτιμάμε ξανά την eq02 χρησιμοποιώντας την πρώτη διαφορά της μεταβλητής "expenditure" ως εξαρτημένη μεταβλητή, και ελέγχουμε ξανά τι συμβαίνει με τα κατάλοιπα. Όπως αναφέρθηκε και προηγουμένως, χρειάζεται ιδιαίτερη προσοχή στην ερμηνεία των αποτελεσμάτων που προκύπτουν, καθώς αυτά αναφέρονται στην πρώτη διαφορά της μεταβλητής "expenditure".

5.6 Μέθοδοι πολλαπλής παλινδρόμησης

- Στο **SPSS**: Προκειμένου να εκτιμήσουμε την πολλαπλή παλινδρόμηση με τις δύο ανεξάρτητες μεταβλητές, χρησιμοποιήσαμε το παράθυρο που εμφανίζεται στην **Εικόνα 5.7**. Κάτω από το λευκό κουτί με την ένδειξη **Independent(s):**, στο οποίο περάσαμε τις ανεξάρτητες μεταβλητές,

υπάρχει η επιλογή σχετικά με τη μέθοδο της παλινδρόμησης που επιθυμούμε να χρησιμοποιήσουμε (**Method**). Το SPSS έχει προεπιλέξει τη μέθοδο **Enter**, με βάση την οποία εκτιμήσαμε την πολλαπλή γραμμική παλινδρόμηση με τις δύο ανεξάρτητες μεταβλητές. Όπως είδαμε προηγουμένως, όταν έχουμε περισσότερες από μία ανεξάρτητες μεταβλητές, ελέγχουμε αρχικά την ύπαρξη συγγραμμικότητας μεταξύ τους, χρησιμοποιώντας τον δείκτη **VIF**. Τι συμβαίνει, όμως, στην περίπτωση που δύο ανεξάρτητες μεταβλητές δεν είναι συγγραμμικές, αλλά παρόλα αυτά μία από τις δύο πρέπει να φύγει από το υπόδειγμα; Ή με άλλα λόγια, αν υπάρχει ήδη η μία μεταβλητή στο υπόδειγμα, μήπως η χρησιμοποίηση της δεύτερης δεν προσφέρει παραπάνω πληροφορία; Η συγκεκριμένη δυνατότητα του SPSS μας παρέχει τις ακόλουθες εναλλακτικές μεθόδους.

- Τη μέθοδο **Enter**, η οποία εκτιμά την πολλαπλή γραμμική παλινδρόμηση χρησιμοποιώντας όλες τις ανεξάρτητες μεταβλητές. Θα γίνει προφανές στη συνέχεια ότι η συγκεκριμένη μέθοδος θα πρέπει να χρησιμοποιείται όσο το δυνατόν λιγότερο.
- Τη μέθοδο **Forward**, η οποία ακολουθεί κάτι καλύτερο από τη μέθοδο **Enter**, καθώς ελέγχει με βάση ένα συγκεκριμένο κριτήριο (*p*-value ή *t*-στατιστικής) ποια είναι η «καλύτερη» μεταβλητή που πρέπει να εισαχθεί πρώτη στο υπόδειγμα. Αν αυτή η «καλύτερη» μεταβλητή δεν ικανοποιεί το συγκεκριμένο κριτήριο, για να εισαχθεί στο υπόδειγμα, τότε καμία μεταβλητή δεν θα εισαχθεί σε αυτό. Εφόσον επιλεγεί η «καλύτερη» μεταβλητή, για να εισαχθεί στο υπόδειγμα, η διαδικασία συνεχίζεται ψάχνοντας τη δεύτερη «καλύτερη» μεταβλητή, για να εισαχθεί στο υπόδειγμα, με δεδομένο ότι ήδη μία μεταβλητή έχει εισαχθεί σε αυτό. Αν η δεύτερη αυτή μεταβλητή ικανοποιεί το συγκεκριμένο κριτήριο εισάγεται στο υπόδειγμα κ.ο.κ.
- Τη μέθοδο **Backward**, η οποία ακολουθεί την αντίστροφη διαδικασία από τη μέθοδο **Forward**, καθώς ξεκινάει με ένα υπόδειγμα που περιέχει όλες τις ανεξάρτητες μεταβλητές και στη συνέχεια αρχίζει να αφαιρεί τις μεταβλητές που δεν ικανοποιούν κάποιο συγκεκριμένο κριτήριο (*p*-value ή *t*-στατιστικής), μέχρι να καταλήξει σε ένα υπόδειγμα που όλες οι ανεξάρτητες μεταβλητές ικανοποιούν το κριτήριο αυτό. Τόσο, όμως, η μέθοδος **Forward** όσο και η μέθοδος **Backward** έχουν το ίδιο μειονέκτημα: η εισαγωγή μίας μεταβλητής στο υπόδειγμα ή η αφαίρεσή της από αυτό είναι οριστική. Δηλαδή, μία μεταβλητή που εισάγεται δεύτερη στο υπόδειγμα μπορεί να κάνει μία μεταβλητή που εισάγεται στη συνέχεια να θεωρηθεί «άχρηστη», αλλά επειδή εισήχθη νωρίτερα στο υπόδειγμα, δεν μπορεί πλέον να αφαιρεθεί.
- Τη μέθοδο **Stepwise**, η οποία ακολουθεί μία πιο σύνθετη διαδικασία που αποτελείται από έναν συνδυασμό των δύο μεθόδων **Forward** και **Backward**, προκειμένου να αντιμετωπίσει το βασικό τους μειονέκτημα. Η μέθοδος αυτή βρίσκει ποια είναι η «καλύτερη» μεταβλητή για να εισαχθεί στο υπόδειγμα και, αφού αυτή εισαχθεί, γίνεται έλεγχος αν πρέπει να βγει από το υπόδειγμα. Η διαδικασία αυτή συνεχίζεται για όλες τις ανεξάρτητες μεταβλητές. Δηλαδή, αφού εισαχθούν κάποιες από αυτές στο υπόδειγμα, γίνεται έλεγχος μήπως κάποια ή κάποιες μεταβλητές πρέπει να αφαιρεθούν από το υπόδειγμα. Ουσιαστικά, ξεκινάει με τη **Forward** μέθοδο και συνεχίζει με την **Backward** μέθοδο. Είναι προφανές ότι αυτή είναι η καλύτερη μέθοδος που πρέπει να ακολουθηθεί, προκειμένου να εκτιμηθεί μια γραμμική παλινδρόμηση.

Όλες, όμως, οι παραπάνω μέθοδοι έχουν το προφανές μειονέκτημα ότι δεν ελέγχουν τις υποθέσεις του γραμμικού μοντέλου. Τους συγκεκριμένους ελέγχους θα πρέπει να τους πραγματοποιήσει ο χρήστης ξεχωριστά. Όταν με τη χρήση των παραπάνω μεθόδων αφαιρούνται μεταβλητές από το υπόδειγμα, το SPSS εμφανίζει στα αποτελέσματα της εκτίμησης της γραμμικής παλινδρόμησης έναν επιπλέον πίνακα που εμφανίζει αυτές τις μεταβλητές (**Excluded Variables**).

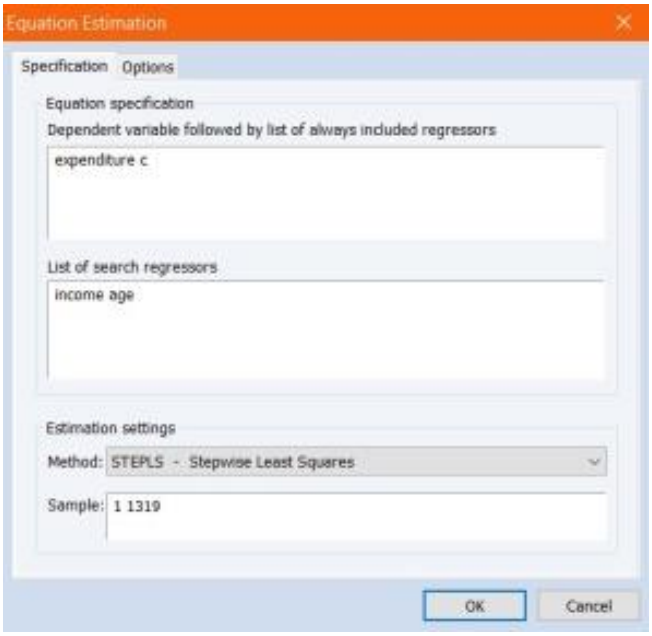
- Στο **Eviews**: Ουσιαστικά, η μέθοδος **Enter** του SPSS δεν είναι τίποτα παραπάνω από τη συνηθισμένη μέθοδο που ακολουθούμε στο Eviews, προκειμένου να εκτιμήσουμε μια γραμμική παλινδρόμηση με τη μέθοδο των ελαχίστων τετραγώνων. Δηλαδή, επιλέγουμε τις ανεξάρτητες μεταβλητές και εκτιμάμε την εξίσωση, χωρίς όμως να έχουμε κάνει κάποιον έλεγχο σχετικά με το αν αυτές οι μεταβλητές είναι κατάλληλες. Το πρόβλημα αυτό

αντιμετωπίζεται στο Eviews με τη χρήση της διαδικασίας **Stepwise Least Squares**, η οποία μας επιτρέπει να χρησιμοποιήσουμε εναλλακτικά είτε τη **Forward** μέθοδο (που προσθέτει σταδιακά στην εξίσωση τις κατάλληλες ανεξάρτητες μεταβλητές) είτε την **Backward** μέθοδο (που ξεκινάει με όλες τις ανεξάρτητες μεταβλητές και αφαιρεί σταδιακά από την εξίσωση αυτές που δεν είναι κατάλληλες).

Έστω, λοιπόν, ότι θέλουμε να χρησιμοποιήσουμε τη μεταβλητή "expenditure" ως εξαρτημένη μεταβλητή και να ελέγξουμε αν οι μεταβλητές "income" και "age" είναι κατάλληλες να χρησιμοποιηθούν ως ανεξάρτητες μεταβλητές. Δημιουργούμε μια νέα εξίσωση με τους τρόπους που έχουμε περιγράψει (έστω eq03) και στο κάτω μέρος του παραθύρου "Equation Estimation" (στο tab "Specification") επιλέγουμε ως **Method:** το **STEPLS - Stepwise Least Squares**. Κάνοντας την επιλογή αυτή, το παράθυρο "Equation Estimation" αλλάζει και εμφανίζονται οι ακόλουθες επιλογές **(Εικόνα 5.27)**:

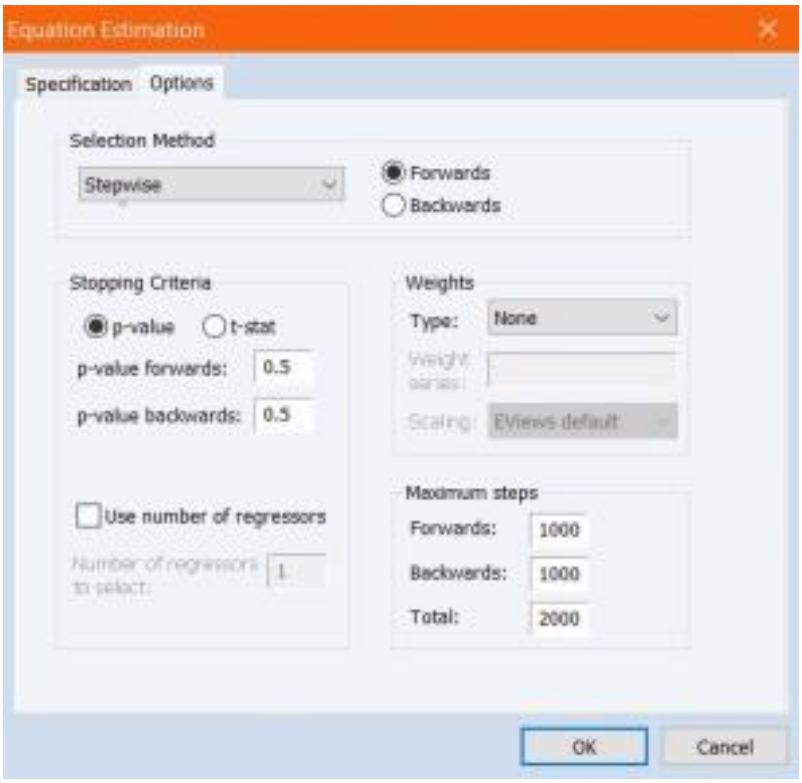
- Η επιλογή "Equation specification" (η οποία είναι η ίδια με προηγουμένως), στην οποία γράφουμε πρώτα την εξαρτημένη μεταβλητή "expenditure" και στη συνέχεια όσες μεταβλητές θέλουμε οπωσδήποτε να συμπεριληφθούν ως ανεξάρτητες μεταβλητές στην εξίσωσή μας (έστω ο σταθερός όρος "c").
- Η επιλογή "List of search regressors", στην οποία γράφουμε τις μεταβλητές που θέλουμε να ελέγξουμε αν είναι κατάλληλες, προκειμένου να χρησιμοποιηθούν ως ανεξάρτητες μεταβλητές στην εξίσωσή μας (έστω οι μεταβλητές "income" και "age").

Στη συνέχεια, από το παράθυρο της **Εικόνας 5.27** επιλέγουμε το tab "Options" **(Εικόνα 5.28),** όπου στο μενού "Selection Method" επιλέγουμε **Stepwise** και στη συνέχεια **Forwards** ή **Backwards**. Το πεδίο "Stopping Criteria" μας επιτρέπει να επιλέξουμε αν το Eviews θα χρησιμοποιήσει *p*-values ή *t*-στατιστικές, προκειμένου να προσθέσει ή να αφαιρέσει ανεξάρτητες μεταβλητές, καθώς και με βάση ποια επίπεδα σημαντικότητας θα κάνει τις συγκεκριμένες επιλογές. Επίσης, αν «τσεκάρουμε» το κουτάκι "Use number of regressors", μπορούμε να επιλέξουμε να σταματήσει η συγκεκριμένη διαδικασία, όταν ένας συγκεκριμένος αριθμός ανεξάρτητων μεταβλητών έχει προστεθεί στο υπόδειγμά μας ή έχει αφαιρεθεί από αυτό. Τέλος, το μενού "Weights" μας επιτρέπει να σταθμίσουμε τις μεταβλητές του υποδείγματος χρησιμοποιώντας την τυπική απόκλιση ή τη διακύμανση, ενώ το πεδίο "Maximum steps" καθορίζει τον μέγιστο αριθμό ανεξάρτητων μεταβλητών που θα προστεθούν στο υπόδειγμα ή θα αφαιρεθούν από αυτό.



Εικόνα 5.27 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

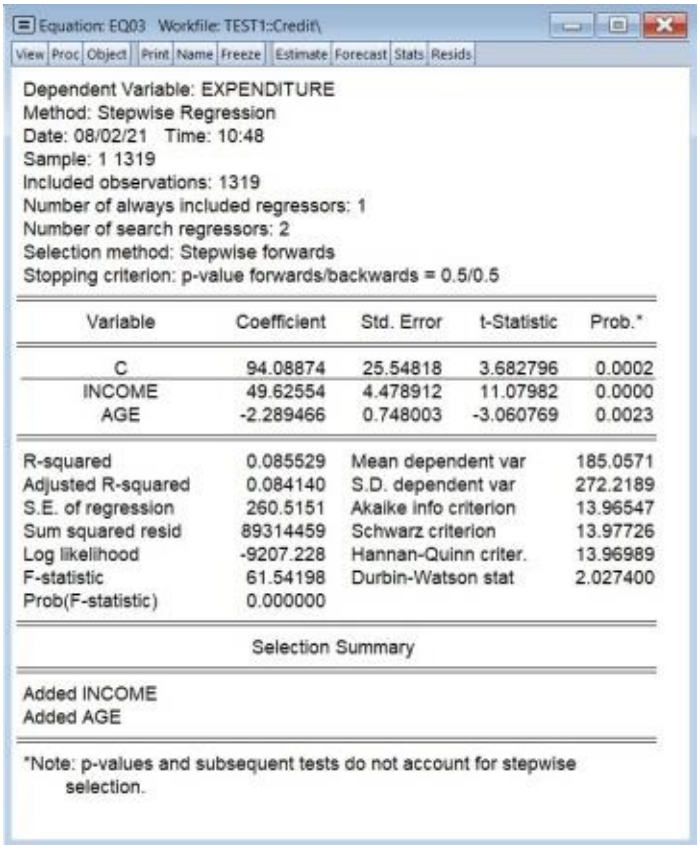
Κρατώντας τις επιλογές που φαίνονται στην **Εικόνα 5.28**, πατάμε **OK** και τα αποτελέσματα εμφανίζονται στον πίνακα της **Εικόνας 5.29**. Στο κάτω μέρος του συγκεκριμένου πίνακα, στο "**Selection Summary**", εμφανίζονται τα μηνύματα "**Added INCOME**" και "**Added AGE**". Αυτό σημαίνει ότι χρησιμοποιώντας τη μέθοδο **Stepwise forwards**, το Eviews επέλεξε τις μεταβλητές "income" και "age" ως κατάλληλες ανεξάρτητες μεταβλητές για την εξίσωση eq03. Στο πάνω μέρος του συγκεκριμένου πίνακα παρουσιάζεται η εκτιμημένη γραμμική παλινδρόμηση, η οποία έχει πλέον συμπεριλάβει τις μεταβλητές "income" και "age" ως ανεξάρτητες, καθώς και τα βασικά διαγνωστικά της μέτρα.



Εικόνα 5.28 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Αν στο παράθυρο της **Εικόνας 5.28** επιλέξουμε **Backwards** αντί για **Forwards** και πατήσουμε **OK**, θα εμφανιστεί ένας πίνακας σχεδόν ίδιος με αυτόν της **Εικόνας 5.29**, με μόνη διαφορά ότι στο κάτω μέρος του στο "**Selection Summary**", θα αναγράφεται το μήνυμα "**No regressors were chosen by the stepwise routine**". Καθώς η μέθοδος **Stepwise backwards** ξεκινάει με όλες τις ανεξάρτητες μεταβλητές στο υπόδειγμα και σταδιακά αφαιρεί αυτές που δεν είναι κατάλληλες, το συγκεκριμένο μήνυμα υποδηλώνει ότι καμία από τις ανεξάρτητες μεταβλητές "income" και "age" δεν επελέγη για να *αφαιρεθεί* από το συγκεκριμένο υπόδειγμα. Και πάλι, στο πάνω μέρος του νέου αυτού πίνακα εμφανίζεται η εκτιμημένη γραμμική παλινδρόμηση με τις μεταβλητές "income" και "age" ως ανεξάρτητες, καθώς και τα βασικά διαγνωστικά της μέτρα.

Βεβαίως, όπως και στην περίπτωση του SPSS, είτε επιλέξουμε τη μέθοδο **Stepwise forwards** είτε τη μέθοδο **Stepwise backwards**, προκειμένου να εκτιμήσουμε την πολλαπλή γραμμική παλινδρόμηση, θα πρέπει στη συνέχεια να ελέγξουμε τις υποθέσεις του γραμμικού υποδείγματος.



Equation: EQ03 Workfile: TEST1::Credit\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: EXPENDITURE
Method: Stepwise Regression
Date: 08/02/21 Time: 10:48
Sample: 1 1319
Included observations: 1319
Number of always included regressors: 1
Number of search regressors: 2
Selection method: Stepwise forwards
Stopping criterion: p-value forwards/backwards = 0.5/0.5

Variable	Coefficient	Std. Error	t-Statistic	Prob.*
C	94.08874	25.54818	3.682796	0.0002
INCOME	49.62554	4.478912	11.07982	0.0000
AGE	-2.289466	0.748003	-3.060769	0.0023

R-squared	0.085529	Mean dependent var	185.0571
Adjusted R-squared	0.084140	S.D. dependent var	272.2189
S.E. of regression	260.5151	Akaike info criterion	13.96547
Sum squared resid	89314459	Schwarz criterion	13.97726
Log likelihood	-9207.228	Hannan-Quinn criter.	13.96989
F-statistic	61.54198	Durbin-Watson stat	2.027400
Prob(F-statistic)	0.000000		

Selection Summary

Added INCOME
Added AGE

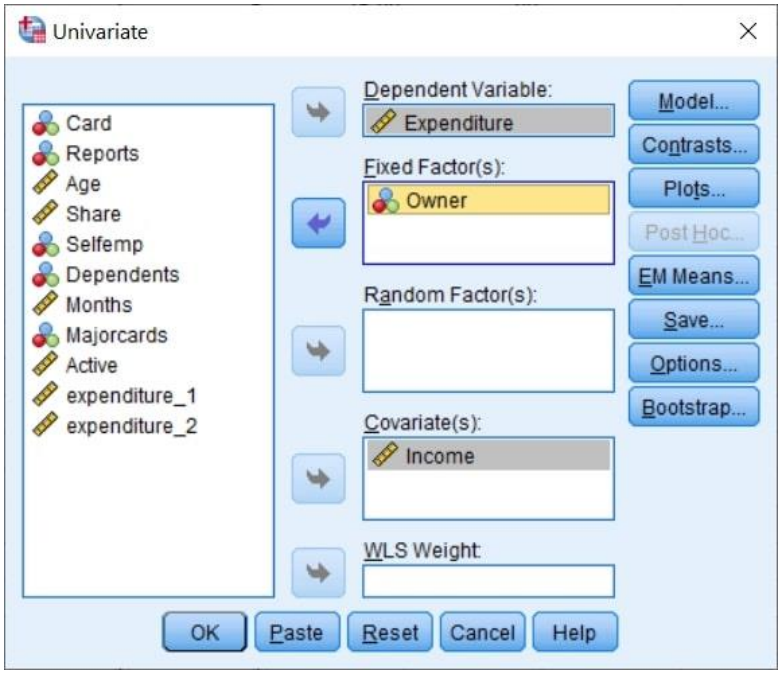
*Note: p-values and subsequent tests do not account for stepwise selection.

Εικόνα 5.29 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

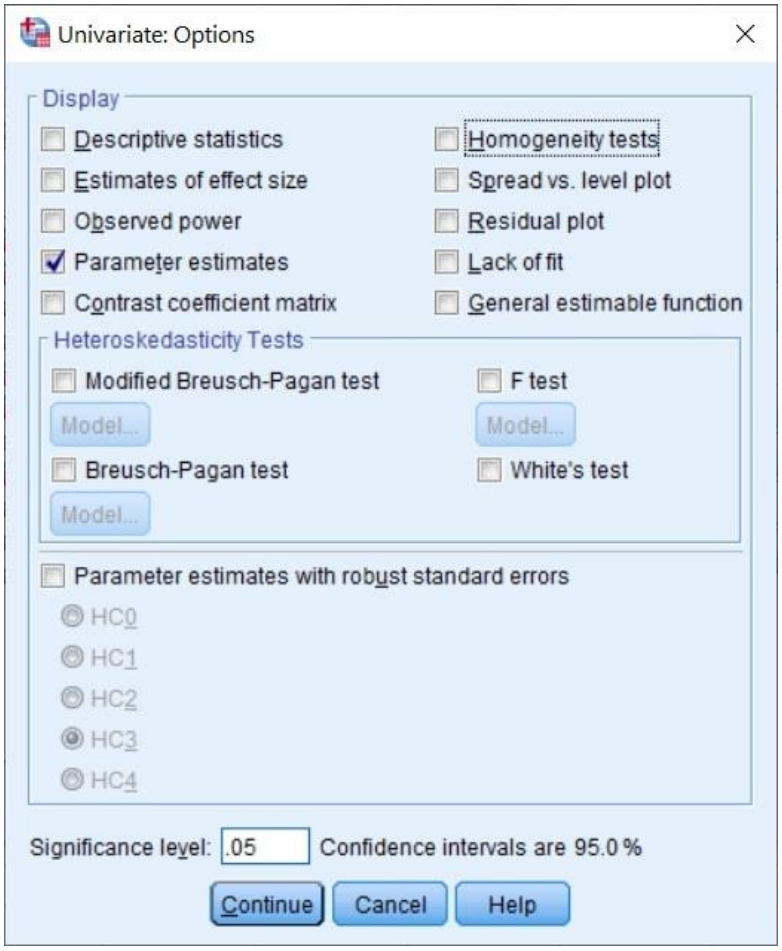
5.7 Πολλαπλή γραμμική παλινδρόμηση με κατηγορική/-ές μεταβλητή/-ές

Στην ενότητα αυτή θα εξετάσουμε τι γίνεται στην περίπτωση που έχουμε μία κατηγορική μεταβλητή στο υπόδειγμα. Θυμηθείτε ότι στην περίπτωση της πολλαπλής γραμμικής παλινδρόμησης που αναλύσαμε στην ενότητα 5.4 το υπόδειγμα περιλάμβανε δύο συνεχείς μεταβλητές.

- Στο **SPSS**: Στο παράδειγμα που θα χρησιμοποιήσουμε, θα αντικαταστήσουμε τη συνεχή μεταβλητή “age” με την κατηγορική μεταβλητή “owner”, η οποία μας δείχνει αν το άτομο που συμμετέχει στο δείγμα μας είναι ιδιοκτήτης της τρέχουσας κατοικίας του ή όχι. Στην περίπτωση που έχουμε έστω και μία κατηγορική μεταβλητή στο υπόδειγμά μας δεν θα ακολουθήσουμε τη διαδικασία εκτίμησης της παλινδρόμησης που χρησιμοποιήσαμε στις ενότητες 5.3 και 5.4, αλλά μία διαφορετική διαδικασία. Πιο συγκεκριμένα, επιλέγουμε **Analyze → General Linear Model → Univariate**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 5.30**. Στη συνέχεια, θα περάσουμε την εξαρτημένη μεταβλητή στο κουτάκι με την ένδειξη **Dependent Variable:**, την κατηγορική μεταβλητή (αλλά και όλες γενικά τις κατηγορικές μεταβλητές που θέλουμε να συμπεριλάβουμε στο υπόδειγμά μας) στο κουτί με την ένδειξη **Fixed Factor(s):** και τη συνεχή ανεξάρτητη μεταβλητή (αλλά και όλες γενικά τις συνεχείς μεταβλητές) στο κουτί με την ένδειξη **Covariate(s):**. Κάνοντας κλικ στην επιλογή **Save** θα εμφανιστεί ένα παράθυρο παρόμοιο με αυτό της **Εικόνας 5.8**, στο οποίο μπορούμε να αποθηκεύσουμε τα τυποποιημένα κατάλοιπα και τις εκτιμημένες τιμές της εξαρτημένης τιμής, όπως και προηγουμένως. Στη συνέχεια, επιλέγουμε **Options** και στο παράθυρο της **Εικόνας 5.31** που θα εμφανιστεί, «τσεκάρουμε» την επιλογή **Parameter estimates**. Πατάμε **Continue** για να επιστρέψουμε στο βασικό παράθυρο της **Εικόνας 5.30** και στη συνέχεια **OK**, προκειμένου να εμφανιστούν τα αποτελέσματα στο Output του SPSS **(Πίνακες 5.9 και 5.10)**.



Εικόνα 5.30 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Εικόνα 5.31 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Η τελευταία στήλη του **Πίνακα 5.9** παρουσιάζει τη στατιστική σημαντικότητα όλων των μεταβλητών (**Sig.**). Η *p*-value για το εισόδημα είναι μικρότερη του 0,01, γεγονός που υποδηλώνει ότι είναι στατιστικά σημαντικό. Όμως, για την ιδιοκτησία κατοικίας, η *p*-value είναι ίση με 0,940. Οπότε, μπορούμε να συμπεράνουμε ότι αυτή η (κατηγορική) μεταβλητή δεν είναι στατιστικά σημαντική. Τέλος, στο κάτω μέρος του **Πίνακα 5.9** εμφανίζεται η τιμή του συντελεστή προσδιορισμού, καθώς και η τιμή του διορθωμένου συντελεστή προσδιορισμού.

Πίνακας 5.9 *Πίνακας ανάλυσης διακύμανσης.*

Tests of Between-Subjects Effects					
Dependent Variable: Expenditure					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	7718056.173[a]	2	3859028.086	56.459	.000
Intercept	263542.451	1	263542.451	3.856	.050
Income	6870051.369	1	6870051.369	100.511	**.000**
Owner	387.626	1	387.626	.006	**.940**
Error	89949881.085	1316	68350.973		
Total	142838568.806	1319			
Corrected Total	97667937.258	1318			
a. R Squared = .079 (Adjusted R Squared = .078)					

Ο **Πίνακας 5.10** παρουσιάζει τις εκτιμημένες τιμές των συντελεστών του εισοδήματος. Όπως αναμενόταν, η επίδραση του εισοδήματος είναι θετική. Θα πρέπει να επισημάνουμε στο σημείο αυτό ότι, επειδή η μεταβλητή "owner" είναι κατηγορική (δηλαδή, μη αριθμητική), δεν μπορούμε να την χρησιμοποιήσουμε για μαθηματικές πράξεις. Για παράδειγμα, δεν μπορούμε να πούμε ιδιοκτησία – μη ιδιοκτησία = 0. Για τον λόγο αυτό δημιουργούμε τη λεγόμενη ψευδομεταβλητή (dummy variable). Αν η κατηγορική μας μεταβλητή έχει 2 τιμές, τότε θέλουμε 1 ψευδομεταβλητή. Γενικά, ο αριθμός των ψευδομεταβλητών θα πρέπει να είναι πάντα ίσος με τον αριθμό των τιμών της κατηγορικής μεταβλητής μειωμένος κατά 1. Στο παράδειγμά μας, η ψευδομεταβλητή που έχουμε δημιουργήσει θα έχει τη μορφή:

$$\text{Owner} = 1 \text{ αν το άτομο είναι ιδιοκτήτης της τρέχουσας κατοικίας του και} = 0 \text{ αν δεν είναι.}$$

Πώς, όμως, μπορούμε να ερμηνεύσουμε την τιμή -1.155 που εμφανίζεται στη στήλη **B** σχετικά με την επίδραση της ιδιοκτησίας της κατοικίας στις μηνιαίες δαπάνες; Το συγκεκριμένο αποτέλεσμα σημαίνει ότι τα άτομα που δεν είναι ιδιοκτήτες της τρέχουσας κατοικίας τους δαπανούν, κατά μέσο όρο, λιγότερα χρήματα από τα άτομα που είναι ιδιοκτήτες. Αν το δούμε από την αντίστροφη σκοπιά, τα άτομα που είναι ιδιοκτήτες κατοικίας δαπανούν κατά μέσο όρο περισσότερα χρήματα από τα άτομα που δεν είναι ιδιοκτήτες κατοικίας. Όμως, η διαφορά αυτή δεν είναι στατιστικά σημαντική (*p*-value = 0,940). Το SPSS θέτει ως επίπεδο αναφοράς την ιδιοκτησία της κατοικίας (owner = 1). Αν θέλουμε να αλλάξουμε την τιμή αναφοράς, θα πρέπει να αλλάξουμε την κωδικοποίηση και να ορίσουμε ως τιμή αναφοράς αυτήν που έχει τη μικρότερη τιμή.

Πίνακας 5.10 *Πίνακας με τις εκτιμήσεις των παραμέτρων της παλινδρόμησης.*

Parameter Estimates						
Dependent Variable: Expenditure						
					95% Confidence Interval	
Parameter	B	Std. Error	T	Sig.	Lower Bound	Upper Bound
Intercept	34.043	20.941	1.626	.104	-7.039	75.124
Income	**45.065**	4.495	10.026	.000	36.247	53.883
[Owner=0]	**-1.155**	15.331	-.075	.940	-31.231	28.922
[Owner=1]	0[a]	.	.	.	.	.
a. This parameter is set to zero because it is redundant.						

Πλέον, το υπόδειγμα της γραμμικής παλινδρόμησης έχει την ακόλουθη μορφή:

$$\widehat{Expenditure} = 34.043 + 45.065 * Income - 1.155 * Owner.$$

Στον **Πίνακα 5.11** παρουσιάζεται ο τρόπος με τον οποίο αλλάζει το υπόδειγμά μας ανάλογα με την ιδιοκτησία της κατοικίας. Στην ουσία, η μη ιδιοκτησία της κατοικίας οδηγεί σε ένα υπόδειγμα με διαφορετική σταθερά. Οπότε, θα έχουμε ένα υπόδειγμα που αποτελείται από 2 παράλληλες ευθείες γραμμές.

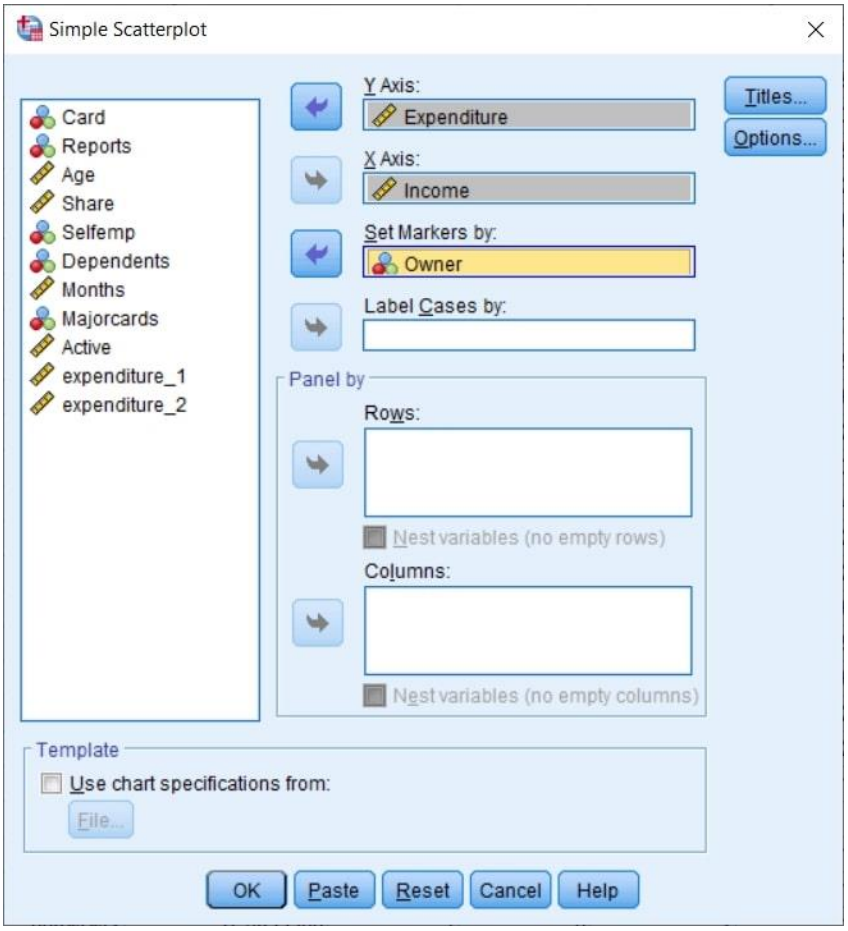
Πίνακας 5.11 *Μοντέλο παλινδρόμησης ανάλογα με την ιδιοκτησία της κατοικίας.*

Ιδιοκτήτης κατοικίας	Μοντέλο
Όχι	$34.043 - 1.155 + 45.065 * Income$
Ναι	$34.043 + 45.065 * Income$

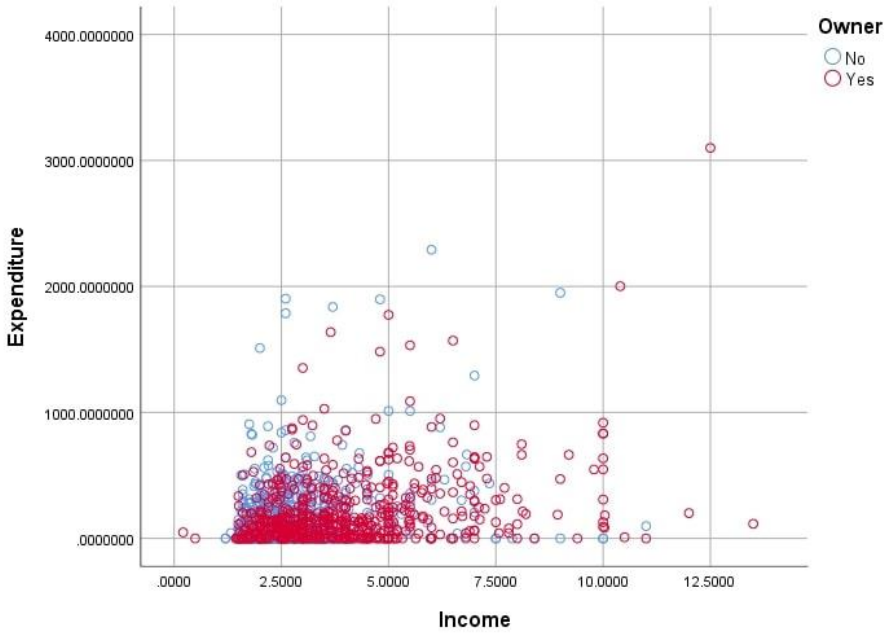
Καθώς έχουμε δύο μεταβλητές στο υπόδειγμά μας (μία κατηγορική και μία συνεχή), μπορούμε να δημιουργήσουμε ένα διάγραμμα, προκειμένου να εξηγήσουμε τη σχέση μεταξύ τους. Όπως έχουμε ήδη αναφέρει, το εισόδημα επιδρά θετικά στις δαπάνες. Αυξάνοντας το εισόδημα κατά μία μονάδα, αναμένουμε οι δαπάνες να αυξηθούν κατά 45 μονάδες περίπου, ανεξαρτήτως αν η κατοικία είναι ιδιόκτητη ή όχι (στο σημείο αυτό θα πρέπει να δοθεί προσοχή στις διαφορετικές μονάδες μέτρησης). Δηλαδή, αναμένουμε την ίδια αύξηση στις δαπάνες τόσο για τους ιδιοκτήτες των κατοικιών όσο και για τους μη ιδιοκτήτες. Μήπως, όμως, είναι πιο λογικό να υποθέσουμε ότι η επίδραση του εισοδήματος στις δαπάνες διαφέρει ανάμεσα σε αυτούς που είναι ιδιοκτήτες κατοικίας και σε αυτούς που δεν είναι; Μήπως, δηλαδή, οι δύο ευθείες γραμμές (που εμφανίζονται στον **Πίνακα 5.11**) δεν πρέπει να είναι παράλληλες, αλλά να έχουν διαφορετική κλίση η καθεμία;

Ας κατασκευάσουμε αρχικά ένα διάγραμμα διασποράς, προκειμένου να δούμε τι συμβαίνει. Στο παράθυρο της **Εικόνας 5.32** (το οποίο είναι ίδιο με αυτό της **Εικόνας 5.2**) θα χρησιμοποιήσουμε ως εξαρτημένη μεταβλητή (στην ένδειξη **Y Axis:**) τις δαπάνες, ως ανεξάρτητη μεταβλητή (στην ένδειξη **X Axis:**) το εισόδημα, ενώ στο κουτί με την ένδειξη **Set Markers by:** θα περάσουμε την κατηγορική μεταβλητή "owner". Πατώντας **OK** θα εμφανιστεί το διάγραμμα διασποράς **(Διάγραμμα 5.4).** Παρατηρώντας προσεκτικά το συγκεκριμένο διάγραμμα διασποράς, διαφαίνεται μία μικρή διαφοροποίηση στη συσχέτιση μεταξύ εισοδήματος και δαπανών, ανάλογα με το αν το άτομο είναι ιδιοκτήτης της κατοικίας ή όχι. Οπότε, μήπως μπορούμε να υπολογίσουμε τον συντελεστή συσχέτισης του Pearson μεταξύ των δύο συνεχών μεταβλητών, ξεχωριστά για τους ιδιοκτήτες κατοικίας και ξεχωριστά για τους ενοικιαστές, προκειμένου να ελέγξουμε κατά πόσο ισχύει αυτό που διακρίνεται γραφικά; Η απάντηση είναι θετική και η διαδικασία είναι η ακόλουθη. Επιλέγοντας **Data → Split File**, θα εμφανιστεί το παράθυρο της **Εικόνας 5.33**, στο οποίο θα επιλέξουμε το **Organize output by groups**, προκειμένου να εμφανιστεί ένα λευκό κουτί ακριβώς από κάτω. Στο κουτί αυτό θα περάσουμε την κατηγορική μεταβλητή "owner", καθώς με βάση τις τιμές της θέλουμε να χωρίσουμε τα δεδομένα μας. Πατώντας **OK** θα διαπιστώσουμε ότι η σειρά των δεδομένων μας έχει αλλάξει, καθώς έχουν ταξινομηθεί ανάλογα με το αν το άτομο είναι ιδιοκτήτης της κατοικίας ή όχι (δηλαδή, ανάλογα με την κατηγορική μεταβλητή).

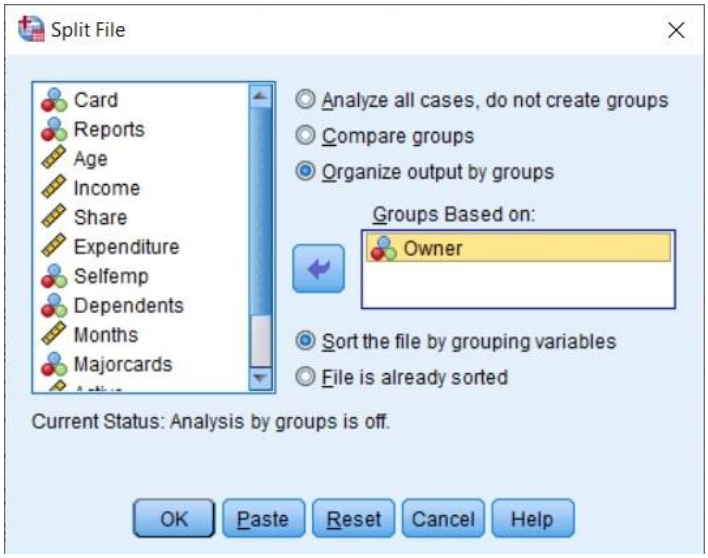
Μπορούμε πλέον να υπολογίσουμε τον συντελεστή συσχέτισης του Pearson (δείτε το παράθυρο της **Εικόνας 4.6**) και το αποτέλεσμα εμφανίζεται στους **Πίνακες 5.12α** και **5.12β**. Όπως προκύπτει από τους συγκεκριμένους πίνακες, για τα άτομα που δεν είναι ιδιοκτήτες κατοικίας ο συντελεστής συσχέτισης μεταξύ των δύο μεταβλητών είναι θετικός (0,197) και στατιστικά σημαντικός, καθώς η αντίστοιχη *p*-value είναι μικρότερη του 0,01 **(Πίνακας 5.12α)**. Κάτι αντίστοιχο ισχύει και για τα άτομα που είναι ιδιοκτήτες **(Πίνακας 5.12β)**. Θα πρέπει, επίσης, να υπενθυμίσουμε ότι, ανεξαρτήτως αν τα άτομα είναι ιδιοκτήτες ή όχι, ο συντελεστής συσχέτισης μεταξύ των δύο αυτών μεταβλητών είναι, επίσης, θετικός (0,281) και στατιστικά σημαντικός (*p*-value < 0,01).



Εικόνα 5.32 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Διάγραμμα 5.4 *Διάγραμμα διασποράς ανάλογα με την ιδιοκτησία της κατοικίας.*



Εικόνα 5.33 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Πίνακας 5.12α *Συντελεστής συσχέτισης μεταξύ εισοδήματος και δαπανών για τους μη ιδιοκτήτες.*

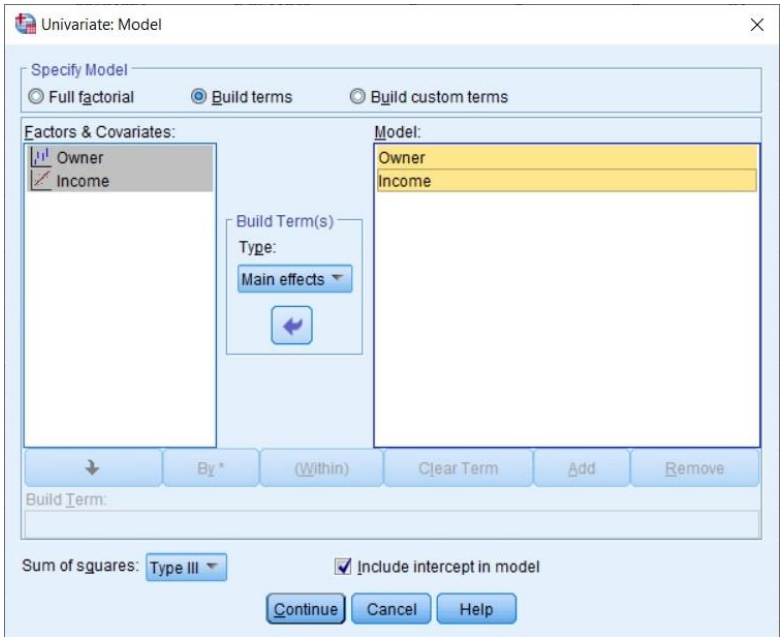
Correlations[a]			
		Income	Expenditure
Income	Pearson Correlation	1	**.197****
	Sig. (2-tailed)		.000
	N	738	738
Expenditure	Pearson Correlation	.197**	1
	Sig. (2-tailed)	.000	
	N	738	738
****. Correlation is significant at the 0.01 level (2-tailed).**			
a. Owner = No			

Πίνακας 5.12β *Συντελεστής συσχέτισης μεταξύ εισοδήματος και δαπανών για τους ιδιοκτήτες.*

Correlations[a]			
		Income	Expenditure
Income	Pearson Correlation	1	**.320****
	Sig. (2-tailed)		.000
	N	581	581
Expenditure	Pearson Correlation	.320**	1
	Sig. (2-tailed)	.000	
	N	581	581
****. Correlation is significant at the 0.01 level (2-tailed).**			
a. Owner = Yes			

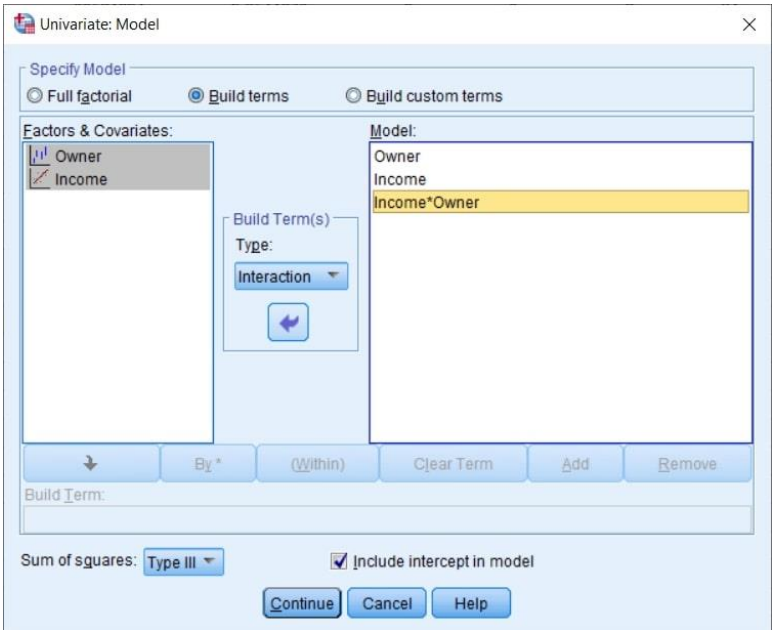
Οπότε, ίσως τελικά η επίδραση του εισοδήματος στις μηνιαίες δαπάνες να μην είναι η ίδια μεταξύ ιδιοκτητών και μη ιδιοκτητών. Προκειμένου αυτό να εξεταστεί σε περισσότερο βάθος, θα πρέπει η αντίστοιχη πληροφορία να ενσωματωθεί στο υπόδειγμά μας. Αρχικά, θα πρέπει να επανενώσουμε τα δύο σετ της μεταβλητής "owner" που δημιουργήσαμε. Για να γίνει αυτό στο SPSS, θα μεταφερθούμε στο παράθυρο της **Εικόνας 5.33**, θα επιλέξουμε το **Analyze all cases, do not create groups** και στη συνέχεια θα πατήσουμε **OK**. Θα επανέλθουμε στο παράθυρο της **Εικόνας 5.30** όπου θα επιλέξουμε **Model**, προκειμένου να εμφανιστεί το παράθυρο της **Εικόνας 5.34.** Στο παράθυρο αυτό θα επιλέξουμε το **Build terms** και θα

περάσουμε τις δύο μεταβλητές “income” και “owner” στο δεξιό λευκό κουτί (χρησιμοποιώντας και πάλι το βελάκι). Στη συνέχεια, θα επιλέξουμε ξανά τις δύο αυτές μεταβλητές και στο κουτί **Type:** θα επιλέξουμε **Interaction**. Οπότε, το παράθυρο της **Εικόνας 5.34** θα αλλάξει σε αυτό της **Εικόνας 5.35.** Τέλος, πατάμε **Continue**, προκειμένου να επιστρέψουμε στο αρχικό παράθυρο της **Εικόνας 5.30** και μετά **OK**. Τα αποτελέσματα που θα εμφανιστούν στο Output του SPSS παρουσιάζονται στους **Πίνακες 5.13** και **5.14**.



Εικόνα 5.34 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Όπως φαίνεται στον **Πίνακα 5.14**, οι συντελεστές της στήλης **Β** έχουν αλλάξει σε μικρό βαθμό σε σχέση με αυτούς του **Πίνακα 5.10**. Αυτό είναι λογικό, καθώς το ίδιο το υπόδειγμα έχει αλλάξει, αφού προσθέσαμε τον όρο της αλληλεπίδρασης. Επίσης, παρατηρήστε ότι η τιμή του διορθωμένου συντελεστή συσχέτισης δεν έχει αλλάξει.



Εικόνα 5.35 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Πινακας 5.13 *Πίνακας ανάλυσης διακύμανσης με την αλληλεπίδραση.*

Tests of Between-Subjects Effects					
Dependent Variable: Expenditure					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	7753995.613[a]	3	2584665.204	37.801	.000
Intercept	288755.743	1	288755.743	4.223	.040
Owner	25968.256	1	25968.256	.380	**.538**
Income	5799454.912	1	5799454.912	84.818	**.000**
Owner * Income	35939.440	1	35939.440	.526	**.469**
Error	89913941.645	1315	68375.621		
Total	142838568.806	1319			
Corrected Total	97667937.258	1318			
a. R Squared = .079 (Adjusted R Squared = .077)					

Πίνακας 5.14 *Εκτιμήσεις των παραμέτρων της παλινδρόμησης με αλληλεπίδραση.*

Parameter Estimates						
Dependent Variable: Expenditure						
					95% Confidence Interval	
Parameter	B	Std. Error	t	Sig.	Lower Bound	Upper Bound
Intercept	24.856	24.480	1.015	.310	-23.168	72.880
[Owner=0]	21.293	34.552	.616	**.538**	-46.489	89.076
[Owner=1]	0[a]	.	.	**.**	.	.
Income	47.370	5.507	8.602	**.000**	36.567	58.173
[Owner=0] * Income	-6.914	9.536	-.725	**.469**	-25.622	11.794
[Owner=1] * Income	0[a]	.	.	.	.	.
a. This parameter is set to zero because it is redundant.						

Οπότε, το υπόδειγμα με την αλληλεπίδραση έχει την ακόλουθη μορφή:

$$\widehat{\mathbf{Expenditure}} = \mathbf{24.856} + \mathbf{21.293} * \mathbf{Owner} + \mathbf{47.370} * \mathbf{Income} - \mathbf{6.914} * \mathbf{Owner} * \mathbf{Income.}$$

Συνεπώς, ο **Πίνακας 5.11** θα έχει πλέον τη μορφή του **Πίνακα 5.15**.

Πίνακας 5.15 *Μοντέλο παλινδρόμησης ανάλογα με το καθεστώς ιδιοκτησίας και την αλληλεπίδραση.*

Ιδιοκτήτης κατοικίας	Μοντέλο
Όχι	$\mathbf{24.856 + 21.293 + (47.370 - 6.914)} * \boldsymbol{Income}$
Ναι	$\mathbf{24.856 + 47.370} * \mathbf{I}\boldsymbol{ncome}$

Από τους **Πίνακες 5.14** και **5.15** προκύπτει ότι η γραμμή παλινδρόμησης είναι διαφορετική ανάλογα με το αν το άτομο είναι ιδιοκτήτης της κατοικίας του ή όχι. Οι μη ιδιοκτήτες (δηλαδή, οι ενοικιαστές) φαίνεται να δαπανούν λιγότερα σε σχέση με αυτούς που είναι ιδιοκτήτες. Ωστόσο, όπως φαίνεται στον **Πίνακα 5.14**, η διαφορά αυτή δεν είναι στατιστικά σημαντική (*p*-value = 0,469).

- Στο **Eviews**: Στην περίπτωση του Eviews, όταν έχουμε μια κατηγορική μεταβλητή (**dummy variable**) μεταξύ των ανεξάρτητων μεταβλητών, ακολουθείται η ίδια ακριβώς διαδικασία με αυτήν της πολλαπλής γραμμικής παλινδρόμησης. Στο παράδειγμα που αναλύσαμε στο SPSS, όπου χρησιμοποιήθηκε η κατηγορική μεταβλητή "owner" ως ανεξάρτητη μεταβλητή, τα αποτελέσματα έδειξαν ότι, αν και τα άτομα που είναι ιδιοκτήτες της κατοικίας τους δαπανούν, κατά μέσο όρο, περισσότερα χρήματα μηνιαίως από τα άτομα που δεν είναι ιδιοκτήτες, η

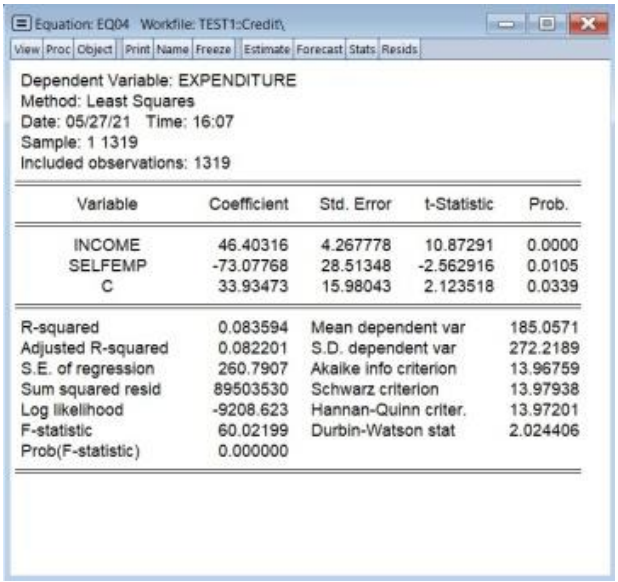
διαφορά αυτή δεν είναι στατιστικά σημαντική. Στην ανάλυση για το Eviews, αντί για την κατηγορική μεταβλητή "owner" θα χρησιμοποιήσουμε ως ανεξάρτητη μεταβλητή την κατηγορική μεταβλητή "selfemp", η οποία παίρνει την τιμή 1 αν το άτομο είναι αυτοαπασχολούμενο, και 0 αν δεν είναι. Έστω, λοιπόν, ότι θέλουμε να εκτιμήσουμε την ευθεία γραμμικής παλινδρόμησης μεταξύ της εξαρτημένης μεταβλητής "expenditure" και των ανεξάρτητων μεταβλητών "income" και "selfemp". Το υπόδειγμά μας θα είναι πλέον της μορφής $\boldsymbol{y = \alpha + \beta_1 x_1 + \beta_2 D_1 + e_i}$, όπου $\boldsymbol{D_1}$ είναι η κατηγορική μεταβλητή "selfemp". Οπότε, δημιουργούμε με τους τρόπους που έχουμε ήδη περιγράψει μια νέα εξίσωση στο Eviews workfile (έστω eq04), την οποία εκτιμάμε με τη μέθοδο των ελαχίστων τετραγώνων. Τα αποτελέσματα που προκύπτουν **(Εικόνα 5.36)** δείχνουν ότι τόσο οι εκτιμημένοι συντελεστές των μεταβλητών "income" και "selfemp" όσο και ο σταθερός όρος είναι στατιστικά σημαντικοί σε επίπεδο σημαντικότητας $a = 5\%$, καθώς οι αντίστοιχες *p*-values είναι μικρότερες από το 0,05.

Ειδικότερα για την κατηγορική μεταβλητή "selfemp", η συγκεκριμένη εκτίμηση υποδηλώνει ότι τα άτομα που είναι αυτοαπασχολούμενα δαπανούν, κατά μέσο όρο, λιγότερα χρήματα μηνιαίως από τα μη αυτοαπασχολούμενα, ενώ η διαφορά αυτή είναι στατιστικά σημαντική. Η εκτιμημένη γραμμική παλινδρόμηση θα είναι στην περίπτωση αυτή η ακόλουθη:

$$\widehat{\boldsymbol{Expenditure}} = \mathbf{33.93473} + \mathbf{46.40316} * \boldsymbol{Income} - \mathbf{73.07768} * \boldsymbol{Selfemp}.$$

Στο κάτω μέρος των αποτελεσμάτων της **Εικόνας 5.36** εμφανίζονται τα βασικά διαγνωστικά μέτρα της εκτιμημένης γραμμικής παλινδρόμησης, όπου το R^2 είναι 0,083594, γεγονός που σημαίνει ότι το υπόδειγμά μας ερμηνεύει μόνο το 8,3594% της μεταβλητότητας των δεδομένων. Τα υπόλοιπα διαγνωστικά μέτρα, καθώς και τα αποτελέσματα των στατιστικών ελέγχων σχετικά με τα κατάλοιπα προκύπτουν με τον ίδιο ακριβώς τρόπο που περιγράψαμε στις ενότητες 5.2 και 5.3. Καθώς η κατηγορική μεταβλητή "selfemp" παίρνει τις τιμές 0 και 1, η εκτιμημένη γραμμική παλινδρόμηση θα αποτελείται από δύο παράλληλες ευθείες γραμμές (μία για την περίπτωση που το άτομο είναι αυτοαπασχολούμενο και μία για την περίπτωση που δεν είναι):

$$\textbf{Αν } \boldsymbol{Selfemp} = \mathbf{1}, \textbf{ τότε } \widehat{\textbf{Expenditure}} = \mathbf{33.93473} + \mathbf{46.40316} * \textbf{Income} - \mathbf{73.07768} \Rightarrow$$
$$\widehat{\textbf{Expenditure}} = -\mathbf{39.14295} + \mathbf{46.40316} * \textbf{Income}.$$
$$\textbf{Αν Selfemp} = \mathbf{0}, \textbf{ τότε } \widehat{\textbf{Expenditure}} = \mathbf{33.93473} + \mathbf{46.40316} * \textbf{Income}.$$


Equation: EQ04 Workfile: TEST1::Credit\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

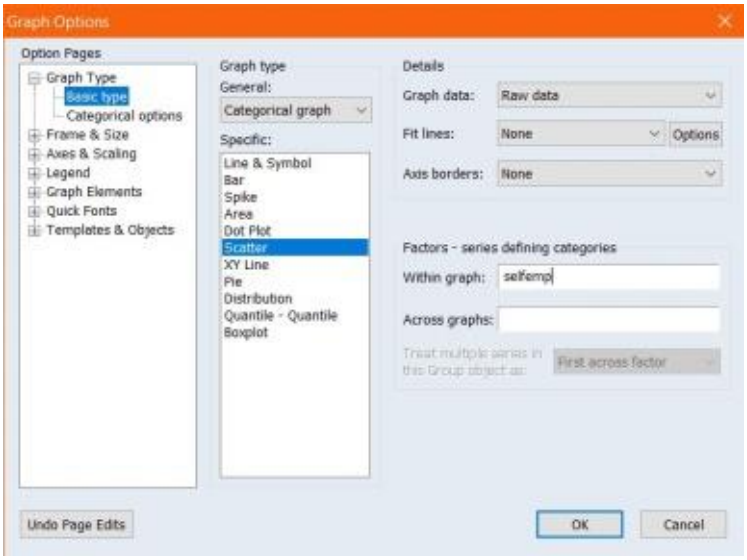
Dependent Variable: EXPENDITURE
Method: Least Squares
Date: 05/27/21 Time: 16:07
Sample: 1 1319
Included observations: 1319

Variable	Coefficient	Std. Error	t-Statistic	Prob.
INCOME	46.40316	4.267778	10.87291	0.0000
SELFEMP	-73.07768	28.51348	-2.562916	0.0105
C	33.93473	15.98043	2.123518	0.0339

R-squared	0.083594	Mean dependent var	185.0571
Adjusted R-squared	0.082201	S.D. dependent var	272.2189
S.E. of regression	260.7907	Akaike info criterion	13.96759
Sum squared resid	89503530	Schwarz criterion	13.97938
Log likelihood	-9208.623	Hannan-Quinn criter.	13.97201
F-statistic	60.02199	Durbin-Watson stat	2.024406
Prob(F-statistic)	0.000000		

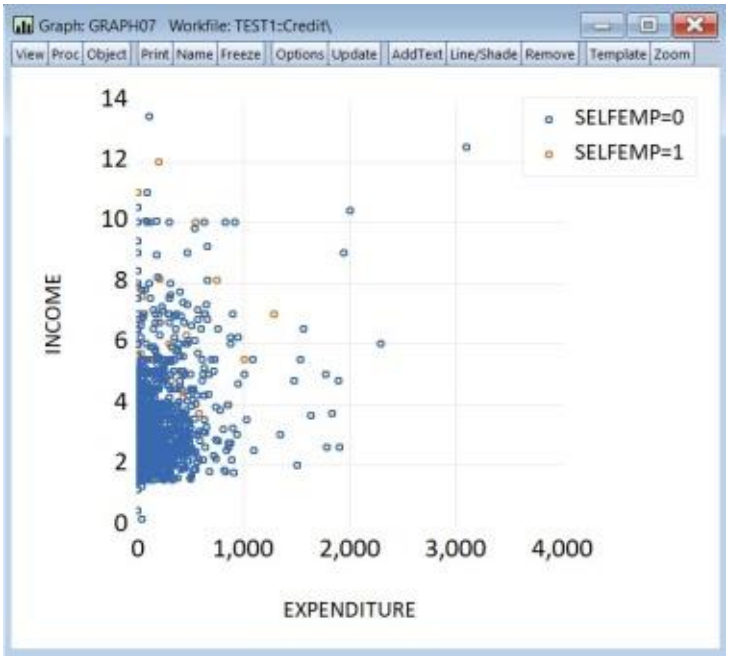
Εικόνα 5.36 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Με βάση τα παραπάνω, αν αυξηθεί το ετήσιο εισόδημα κατά μία μονάδα (ή $10.000), η εκτιμημένη μέση μηνιαία δαπάνη με τη χρήση πιστωτικής κάρτας θα αυξηθεί κατά 46,40316 ή $4.640,316. Δηλαδή, αναμένουμε την ίδια αύξηση στη μέση μηνιαία δαπάνη είτε το άτομο είναι αυτοαπασχολούμενο είτε όχι. Όμως, μπορεί τελικά να μην ισχύει κάτι τέτοιο, δηλαδή οι δύο ευθείες γραμμές να μην είναι παράλληλες, αλλά να έχουν διαφορετική κλίση η καθεμία. Για να το ελέγξουμε αυτό, κατασκευάζουμε αρχικά ένα διάγραμμα διασποράς. «Ανοίγουμε» τις μεταβλητές "expenditure" και "income" ως Group, επιλέγουμε στη συνέχεια **View → Graph** και στο παράθυρο "Graph Options" που θα εμφανιστεί **(Εικόνα 5.37)**, επιλέγουμε **Basic type** στην κατηγορία "Option Pages", **Categorical Graph** στην κατηγορία "General:", **Scatter** στην κατηγορία "Specific", ενώ συμπληρώνουμε **selfemp** στο πεδίο "Within graph:". Στις επιλογές "Fit lines:" και "Axis borders:" αφήνουμε το "None" που έχει προεπιλέξει το Eviews, και στη συνέχεια πατάμε **OK**. Το διάγραμμα διασποράς εμφανίζεται στην **Εικόνα 5.38**.



Εικόνα 5.37 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Από το συγκεκριμένο διάγραμμα διασποράς δεν μπορούμε να εξάγουμε κάποιο ξεκάθαρο συμπέρασμα σχετικά με το αν η επίδραση του ετήσιου εισοδήματος στη μέση μηνιαία δαπάνη διαφέρει ανάλογα με το αν το άτομο είναι αυτοαπασχολούμενο ή όχι. Οπότε, ως επόμενο βήμα, μπορούμε να υπολογίσουμε τον συντελεστή συσχέτισης του Pearson μεταξύ των μεταβλητών "expenditure" και "income", τόσο για την περίπτωση που η μεταβλητή "selfemp" παίρνει την τιμή 1 (δηλαδή, το άτομο είναι αυτοαπασχολούμενο) όσο και για την περίπτωση που παίρνει την τιμή 0 (δηλαδή, το άτομο δεν είναι αυτοαπασχολούμενο), προκειμένου να εξετάσουμε τι ισχύει τελικά. Στο Group των μεταβλητών "expenditure" και "income", επιλέγουμε **View → Covariance Analysis** και στο παράθυρο της **Εικόνας 4.7** «τσεκάρουμε» τα κουτιά **Correlation**, **t-statistic** και **Probability | t | = 0**. Για την περίπτωση που η μεταβλητή "selfemp" παίρνει την τιμή 1 γράφουμε στο πεδίο "Sample" **if selfemp=1**, ενώ για την περίπτωση που η μεταβλητή "selfemp" παίρνει την τιμή 0 γράφουμε στο πεδίο "Sample" **if selfemp=0**. Τα αποτελέσματα που προκύπτουν παρουσιάζονται στον **Πίνακα 5.16** και, όπως φαίνεται, ο συντελεστής συσχέτισης είναι και στις δύο περιπτώσεις θετικός και στατιστικά σημαντικός (καθώς οι αντίστοιχες *p*-values είναι μικρότερες του 0,05), ενώ οι τιμές του δεν παρουσιάζουν μεγάλη διαφορά. Συνεπώς, δεν φαίνεται η επίδραση του ετήσιου εισοδήματος στη μέση μηνιαία δαπάνη να διαφοροποιείται ανάλογα με το αν το άτομο είναι αυτοαπασχολούμενο ή όχι.

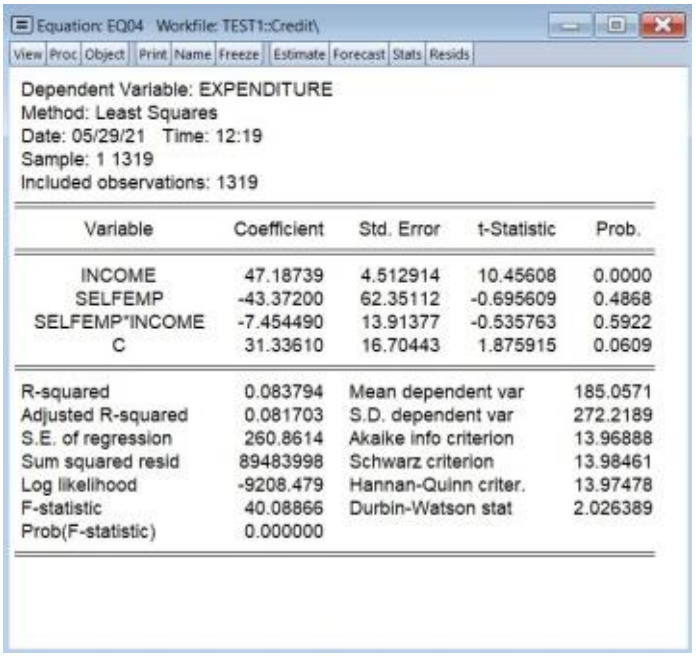


Εικόνα 5.38 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Πίνακας 5.16 *Συντελεστής συσχέτισης του Pearson.*

	Συντελεστής Pearson	*t*-στατιστική	*p*-value
selfemp=1	0,345360	3,471737	0,0008
selfemp=0	0,283840	10,36473	0,0000

Η εκτίμηση της eq04, στην οποία έχουμε συμπεριλάβει τη μεταβλητή "**selfemp*income**" ως ανεξάρτητη μεταβλητή, η οποία δείχνει την αλληλεπίδραση, επιβεβαιώνει τα παραπάνω συμπεράσματα. Αν ο εκτιμημένος συντελεστής της συγκεκριμένης μεταβλητής είναι στατιστικά σημαντικός, τότε η επίδραση του ετήσιου εισοδήματος στη μέση μηνιαία δαπάνη διαφοροποιείται ανάλογα με το αν το άτομο είναι αυτοαπασχολούμενο ή όχι. Από τα αποτελέσματα της **Εικόνας 5.39** προκύπτει ότι ο εκτιμημένος συντελεστής της μεταβλητής "selfemp*income" δεν είναι στατιστικά σημαντικός σε επίπεδο σημαντικότητας $a = 5\%$, καθώς η αντίστοιχη *p*-value είναι μεγαλύτερη από το 0,05. Επιπλέον, η προσθήκη της νέας αυτής μεταβλητής ουσιαστικά «χαλάει» το υπόδειγμά μας, καθώς έχει ως αποτέλεσμα ο εκτιμημένος συντελεστής της μεταβλητής "selfemp" να μην είναι πλέον στατιστικά σημαντικός (*p*-value = 0.4868 > 0,05). Τέλος, τόσο ο συντελεστής προσδιορισμού R^2 όσο και ο διορθωμένος συντελεστής προσδιορισμού $\bar{R}^2$ παραμένουν χαμηλοί (0,083794 και 0,081703, αντίστοιχα). Οπότε, μπορούμε να συμπεράνουμε ότι η επίδραση του ετήσιου εισοδήματος στη μέση μηνιαία δαπάνη δεν διαφοροποιείται ανάλογα με το αν το άτομο είναι αυτοαπασχολούμενο ή όχι. Συνεπώς, το υπόδειγμά μας περιγράφεται από τις δύο παράλληλες ευθείες γραμμές που αναλύσαμε παραπάνω.



Εικόνα 5.39 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Αν οι εκτιμημένοι συντελεστές των μεταβλητών *"selfemp"* και *"selfemp*income"* ήταν στατιστικά σημαντικοί, τότε η εκτιμημένη γραμμική παλινδρόμηση θα ήταν η ακόλουθη:

$$\widehat{\textbf{Expenditure}} = \textbf{31.33610} + \textbf{47.18739} * \textbf{Income} - \textbf{43.37200} * \textbf{Selfemp} - \textbf{7.454490} * \textbf{Selfemp} * \textbf{Income.}$$

Καθώς η κατηγορική μεταβλητή "selfemp" παίρνει τις τιμές 0 και 1, το υπόδειγμά μας θα αποτελούνταν από δύο διαφορετικές ευθείες (μία για την περίπτωση που το άτομο είναι αυτοαπασχολούμενο και μία για την περίπτωση που δεν είναι). Οπότε, θα είχε την ακόλουθη μορφή:

$$\textbf{Αν Selfemp} = \textbf{1, τότε } \widehat{\textbf{Expenditure}} = \textbf{31.33610} + \textbf{47.18739} * \textbf{Income} - \textbf{43.37200} - \textbf{7.454490} * \textbf{Income} \Rightarrow$$
$$\widehat{\textbf{Expenditure}} = -\textbf{12.0359} + \textbf{39.7329} * \textbf{Income.}$$
$$\textbf{Αν Selfemp} = \textbf{0, τότε } \widehat{\textbf{Expenditure}} = \textbf{31.33610} + \textbf{47.18739} * \textbf{Income.}$$

Βιβλιογραφία

Ξενόγλωσση

Draper, N.R., & Smith, H. (1981). *Applied Regression Analysis* (2nd ed.). New York: John Wiley.

Wooldridge, J.M. (2013). *Introductory Econometrics: A Modern Approach* (5th ed.). Mason, Ohio: South-Western, Cengage Learning.

Κεφάλαιο 6 Ανάλυσης διακύμανσης

Σύνοψη

Το έκτο κεφάλαιο του βιβλίου επικεντρώνεται στην ανάλυση της διακύμανσης. Στο κεφάλαιο αυτό παρουσιάζονται διάφορες τεχνικές, παραμετρικές και μη, είτε για έναν είτε για δύο παράγοντες. Επιπλέον, η ανάλυση που ακολουθεί καλύπτει και την περίπτωση των εξαρτημένων δειγμάτων (δηλαδή, των επαναλαμβανόμενων μετρήσεων). Οι στόχοι του κεφαλαίου αυτού είναι ο χρήστης του SPSS ή του Eviews να είναι σε θέση να επιλέγει τον κατάλληλο έλεγχο για το πείραμα ή τη μελέτη του και να μπορεί να τον εφαρμόζει σε καθένα από τα προγράμματα αυτά.

Προαπαιτούμενη γνώση

Απαιτούνται βασικές γνώσεις στατιστικής.

6.1 Ανάλυση διακύμανσης κατά έναν παράγοντα (One-way ANOVA) στο SPSS

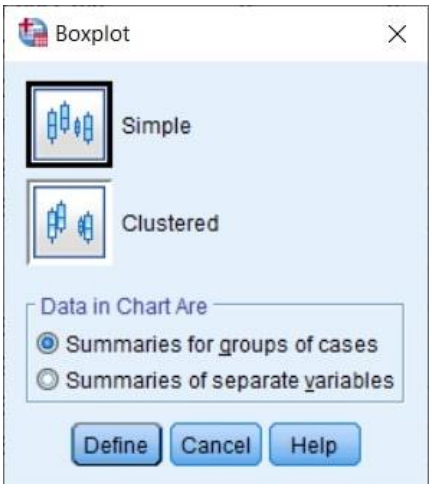
Στο τέταρτο κεφάλαιο αναλύσαμε τους ελέγχους υποθέσεων για τις περιπτώσεις ενός και δύο ανεξάρτητων (ή εξαρτημένων) δειγμάτων. Στην περίπτωση, όμως, που έχουμε δείγματα που προέρχονται από τρεις πληθυσμούς και για τα οποία θέλουμε να συγκρίνουμε τους μέσους τους, χρησιμοποιούμε την τεχνική της ανάλυσης διακύμανσης κατά έναν παράγοντα. Μία εναλλακτική μέθοδος είναι χρησιμοποιήσουμε τον έλεγχο *t* που αναλύσαμε προηγουμένως σε όλα τα πιθανά ζεύγη δειγμάτων. Με τον τρόπο αυτό, όμως, μειώνουμε αισθητά την πιθανότητα να ισχύουν και οι δύο έλεγχοι ταυτόχρονα. Αντιθέτως, η ανάλυση διακύμανσης διατηρεί την πιθανότητα αυτή σταθερή, καθώς πραγματοποιεί έναν μόνο έλεγχο. Το πρόβλημα αυτό μπορεί να εκφραστεί διαφορετικά ως ο έλεγχος ισότητας των μέσων για μία ποσοτική μεταβλητή με παράγοντα διαφοροποίησης μία κατηγορική μεταβλητή με τρία επίπεδα. Οπότε, ελέγχουμε αν η κατηγορική μεταβλητή (ή ο παράγοντας) επηρεάζει την ποσοτική μεταβλητή.

Ωστόσο, οι υποθέσεις εφαρμογής της ανάλυσης διακύμανσης κατά έναν παράγοντα είναι πιο αυστηρές σε σχέση με την περίπτωση των δύο δειγμάτων. Είναι οι ίδιες με την περίπτωση της απλής γραμμικής παλινδρόμησης, δηλαδή κανονικότητα, ανεξαρτησία και ομοσκεδαστικότητα των καταλοίπων. Στην παρούσα ανάλυση, η ομοσκεδαστικότητα των καταλοίπων σημαίνει ότι τα κατάλοιπα που δημιουργούνται πρέπει να έχουν ίσες διασπορές για κάθε επίπεδο του παράγοντα. Επίσης, η τυχαιότητα εννοείται στη στατιστική, καθώς διαφορετικά δεν υπάρχει νόημα διεξαγωγής των στατιστικών ελέγχων που έχουμε ήδη αναφέρει. Στην περίπτωση, όμως, που δεν ισχύουν οι υποθέσεις για τα κατάλοιπα, υπάρχουν διάφορες στατιστικές τεχνικές, οι οποίες περιλαμβάνονται στο SPSS και μας βοηθάνε στο να εξαγάγουμε συμπεράσματα. Εναλλακτικά, μπορούμε να χρησιμοποιήσουμε και κάποιο είδος μετασχηματισμού στην ποσοτική ή στην εξαρτημένη μεταβλητή. Ουσιαστικά, τα αποτελέσματα της ανάλυσης διακύμανσης κατά έναν παράγοντα δεν χάνουν την ισχύ τους, όταν έχουμε μικρές αποκλίσεις από την κανονικότητα. Γενικά, στις περιπτώσεις που δεν ισχύει η υπόθεση της ομοσκεδαστικότητας συνιστάται η χρησιμοποίηση του ελέγχου του **Welch** ή του ελέγχου των **Brown-Forsythe**, καθώς οι δύο αυτοί έλεγχοι αντιμετωπίζουν το ζήτημα της ετεροσκεδαστικότητας.

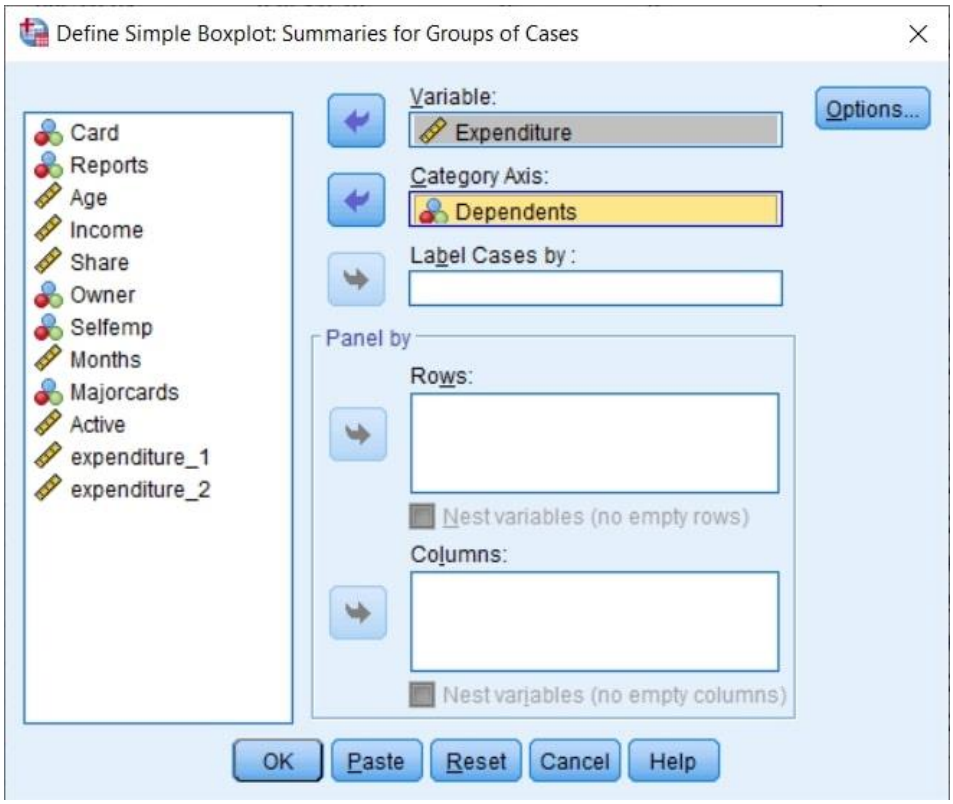
Θα χρησιμοποιήσουμε πάλι τα δεδομένα των πιστωτικών καρτών (**credit.sav**). Έστω, λοιπόν, ότι θέλουμε να ελέγξουμε αν οι μηνιαίες δαπάνες του ατόμου επηρεάζονται από τον αριθμό των εξαρτώμενων μελών του. Ο αριθμός των εξαρτώμενων μελών δεν είναι κατηγορική μεταβλητή, αλλά στο παράδειγμά μας θα παίξει τον ρόλο αυτό, προκειμένου να παρουσιάσουμε την τεχνική της ανάλυσης διακύμανσης. Οπότε, οι υποθέσεις (μηδενική και εναλλακτική) ορίζονται ως εξής:

$$\mathbf{H_0}: \boldsymbol{\mu_1 = \mu_2 = \cdots = \mu_7}.$$
$$\boldsymbol{H_1}: \textbf{ένα τουλάχιστον ζεύγος μέσων διαφέρει.}$$

Ωστόσο, πριν πραγματοποιήσουμε τον έλεγχο της ανάλυσης διακύμανσης, έχει ενδιαφέρον να παρατηρήσουμε γραφικά πώς κατανέμονται οι μηνιαίες δαπάνες ανάλογα με τον αριθμό των εξαρτώμενων μελών. Αυτό γίνεται με την κατασκευή του λεγόμενου **Box Plot** γραφήματος. Προκειμένου, λοιπόν, να το κατασκευάσουμε, επιλέγουμε **Graphs → Legacy Dialogs → Boxplot**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 6.1**. Στο παράθυρο αυτό επιλέγουμε το πρώτο εικονίδιο (**Simple**) και στη συνέχεια **Define**, προκειμένου να οδηγηθούμε στο παράθυρο της **Εικόνας 6.2**.

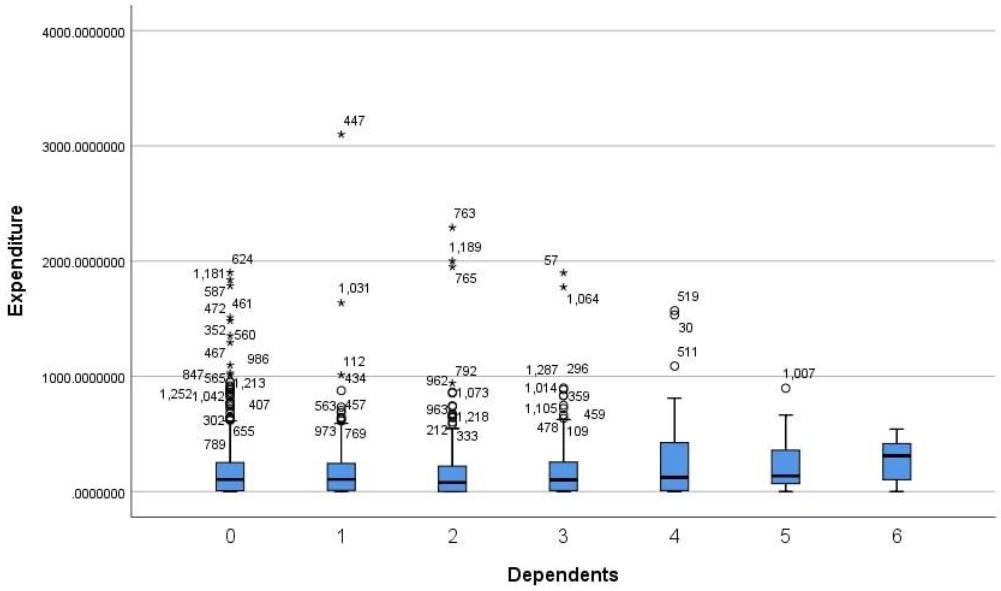


Εικόνα 6.1 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



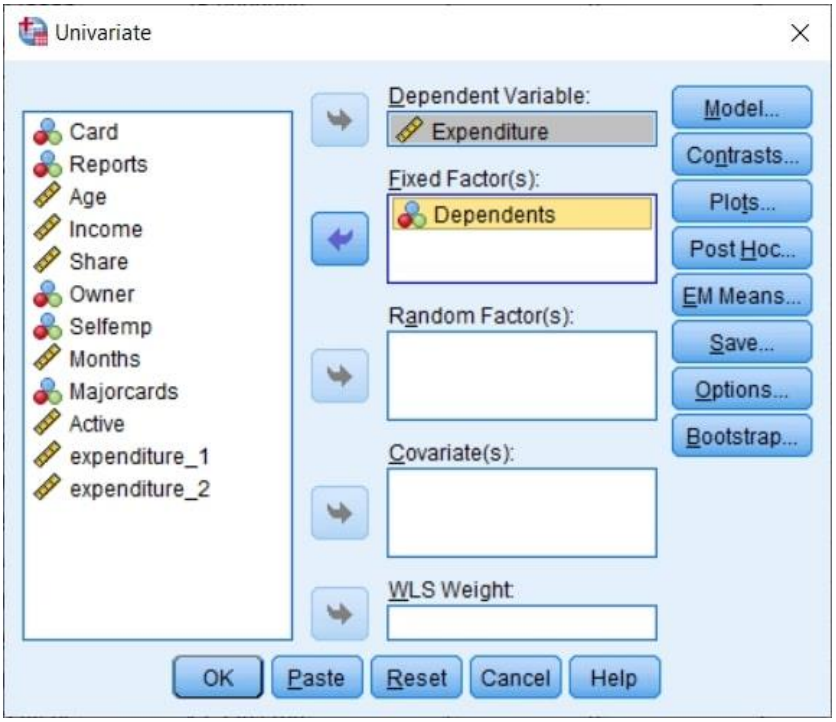
Εικόνα 6.2 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Στο παράθυρο αυτό, περνάμε την εξαρτημένη μεταβλητή στο πρώτο λευκό κουτί με την ένδειξη **Variable:** και τον παράγοντα στο δεύτερο λευκό κουτί με την ένδειξη **Category Axis:**. Στη συνέχεια, πατάμε **OK**, προκειμένου να εμφανιστεί το Box Plot γράφημα **(Διάγραμμα 6.1)**. Η οριζόντια γραμμή που εμφανίζεται μέσα σε κάθε ορθογώνιο είναι η διάμεσος και όχι ο μέσος. Το μήκος των ορθογωνίων αυτών, τα οποία έχουν κατασκευαστεί για καθέναν αριθμό εξαρτώμενων μελών ξεχωριστά, υπολογίζεται με βάση τα τεταρτημόρια. Το κάθε ορθογώνιο υποδηλώνει το 25% και το 75% των παρατηρήσεων, ενώ η γραμμή μέσα σε αυτό είναι το 50% των παρατηρήσεων (διάμεσος). Πάνω και κάτω από κάθε ορθογώνιο υπάρχουν κάθετες γραμμές, τα άκρα των οποίων ονομάζονται «φράχτες». Οι παρατηρήσεις που βρίσκονται έξω από το ορθογώνιο και απεικονίζονται με κυκλάκι ονομάζονται ήπια ακραία σημεία, ενώ οι παρατηρήσεις που απεικονίζονται με αστερίσκο ονομάζονται εξαιρετικά ακραία σημεία. Στη στατιστική, με τον όρο ακραίο σημείο εννοούμε μία παρατήρηση (ή τιμή) που απέχει περισσότερο από δύο ή τρεις τυπικές αποκλίσεις από τη μέση τιμή.



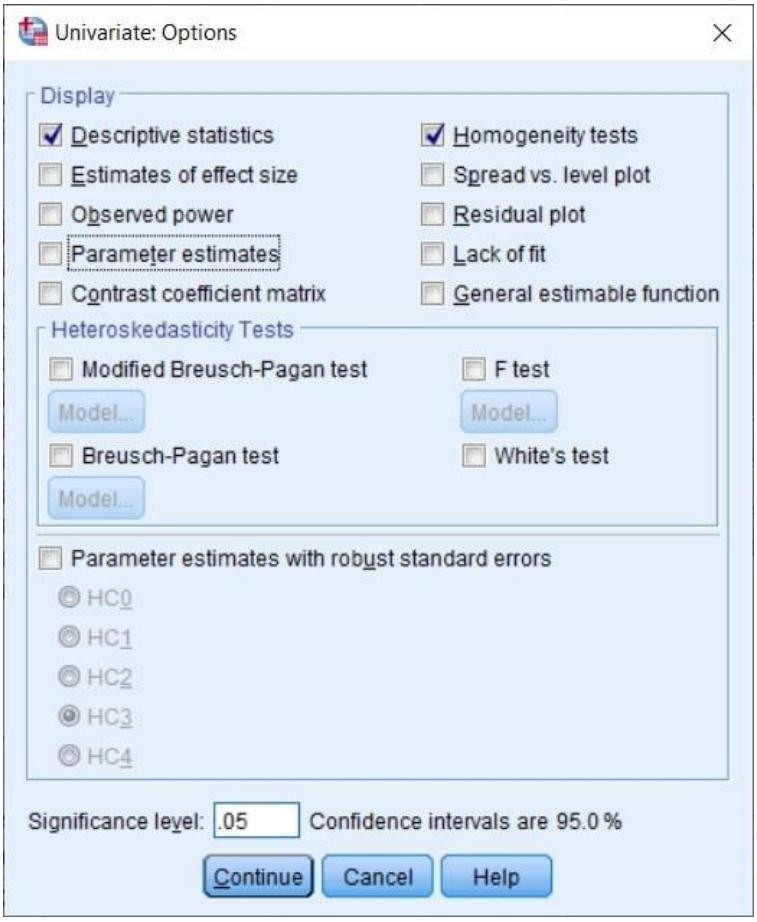
Διάγραμμα 6.1 *Box plot των μηνιαίων δαπανών ανάλογα με τον αριθμό των εξαρτώμενων μελών.*

Στη συνέχεια, προκειμένου να πραγματοποιήσουμε τον έλεγχο της ανάλυσης διακύμανσης στο SPSS, επιλέγουμε **Analyze → General Linear Model → Univariate**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 6.3**. Το παράθυρο αυτό είναι το ίδιο με αυτό της **Εικόνας 5.30**, ενώ και η διαδικασία που θα ακολουθηθεί είναι σχεδόν η ίδια.



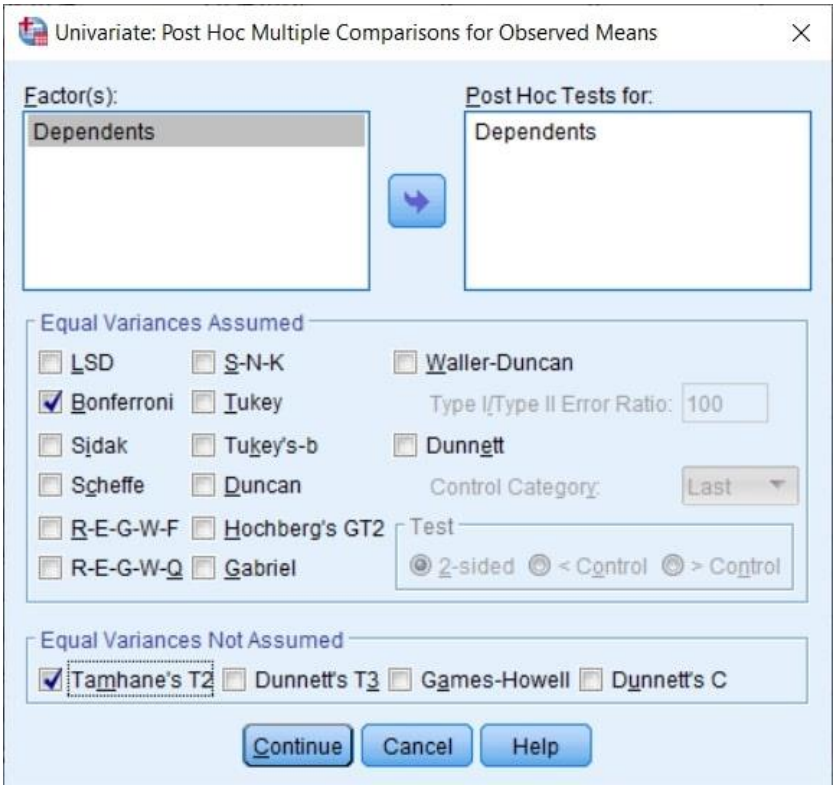
Εικόνα 6.3 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Στο λευκό κουτί με την ένδειξη **Dependent Variable:** περνάμε την εξαρτημένη μεταβλητή, δηλαδή τις μηνιαίες δαπάνες. Στο λευκό κουτί με την ένδειξη **Fixed Factor(s):** που βρίσκεται ακριβώς από κάτω, περνάμε την κατηγορική μεταβλητή της οποίας τα επίπεδα αντιστοιχούν στον αριθμό των εξαρτώμενων μελών. Πατώντας **Save** θα εμφανιστεί ένα παράθυρο, στο οποίο επιλέγουμε να αποθηκεύσουμε τα κατάλοιπα και τις εκτιμημένες τιμές της εξαρτημένης μεταβλητής, όπως ακριβώς κάναμε και στην περίπτωση της γραμμικής παλινδρόμησης, προκειμένου να ελέγξουμε την υπόθεση της κανονικότητας (και της ανεξαρτησίας) των καταλοίπων. Στη συνέχεια, επιλέγουμε **Options**, προκειμένου να εμφανιστεί το παράθυρο της **Εικόνας 6.4**.



Εικόνα 6.4 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Στο παράθυρο της **Εικόνας 6.4** εμφανίζονται διάφορες επιλογές, από τις οποίες θα επιλέξουμε μόνο τα περιγραφικά μέτρα (**Descriptive statistics**), καθώς και τον έλεγχο ισότητας των διασπορών του Levene (**Homogeneity tests**). Στη συνέχεια, πατάμε **Continue** και επιστρέφουμε στο παράθυρο της **Εικόνας 6.3**. Πατώντας **Post Hoc** στο παράθυρο της **Εικόνας 6.3** προκύπτει το παράθυρο της **Εικόνας 6.5**. Περνάμε τον παράγοντα στο δεξιό κουτί με την ένδειξη **Post Hoc Tests for:** και στη συνέχεια επιλέγουμε τον έλεγχο του **Bonferroni** για την περίπτωση που η υπόθεση της ισότητας των διασπορών ικανοποιείται, και τον έλεγχο **Tamhane's T2** για την περίπτωση που η υπόθεση της ομοσκεδαστικότητας δεν είναι εύλογη. Πατώντας **Continue** και μετά **OK** προκύπτουν τα αποτελέσματα στο Output του SPSS, μέρος των οποίων παρουσιάζουμε στη συνέχεια **(Πίνακες 6.1-6.3)**.



Εικόνα 6.5 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Πίνακας 6.1 *Περιγραφικά μέτρα για τις δαπάνες ανάλογα με τις τιμές της ανεξάρτητης μεταβλητής (παράγοντα).*

Descriptive Statistics			
Dependent Variable: Expenditure			
Dependents	Mean	Std. Deviation	N
0	180.096083875	247.6850077190	659
1	178.012246434	271.4225284819	267
2	171.662922367	295.8939905045	218
3	210.447835417	309.9055246599	115
4	271.652233318	376.3024568974	44
5	267.711485556	315.1957019757	9
6	270.223928571	219.2153018949	7
Total	185.057070778	272.2189174955	1319

Πίνακας 6.2 *Έλεγχος του Levene για την ισότητα των διακυμάνσεων.*

Levene's Test of Equality of Error Variancesa,b					
		Levene Statistic	df1	df2	Sig.
Expenditure	Based on Mean	2.397	6	1312	**.026**
	Based on Median	1.409	6	1312	**.208**
	Based on Median and with adjusted df	1.409	6	1220.687	**.208**
	Based on trimmed mean	1.956	6	1312	**.069**

Tests the null hypothesis that the error variance of the dependent variable is equal across groups.
a. Dependent variable: Expenditure
b. Design: Intercept + Dependents

Ο **Πίνακας 6.1** περιέχει κάποια περιγραφικά μέτρα, ενώ ιδιαίτερο ενδιαφέρον παρουσιάζει ο **Πίνακας 6.2** που παρουσιάζει τα αποτελέσματα του ελέγχου του Levene, ο οποίος χρησιμοποιείται για τον έλεγχο της υπόθεσης της ισότητας των διασπορών για όλα τα επίπεδα του παράγοντα. Η μηδενική και η εναλλακτική υπόθεση για τον συγκεκριμένο έλεγχο είναι οι ακόλουθες:

$$\mathbf{H_0}: \boldsymbol{\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_7^2}.$$
$$\boldsymbol{H_1}: \textbf{τουλάχιστον μία διακύμανση διαφέρει από τις άλλες.}$$

Το παρατηρηθέν επίπεδο στατιστικής σημαντικότητας για τον έλεγχο του Levene κυμαίνεται από 0,026 έως 0,208, ανάλογα με το στατιστικό μέτρο που έχει χρησιμοποιηθεί. Συνεπώς, μπορούμε να συμπεράνουμε ότι η παραπάνω μηδενική υπόθεση μπορεί και να απορρίπτεται, και άρα η υπόθεση της ισότητας των διακυμάνσεων να μην ικανοποιείται. Ο **Πίνακας 6.3** παρουσιάζει το αποτέλεσμα του ελέγχου *F* της ανάλυσης διακύμανσης για τον έλεγχο της ισότητας των μέσων. Το σκιασμένο παρατηρηθέν επίπεδο στατιστικής σημαντικότητας είναι ίσο με 0,246. Οπότε, μπορούμε να συμπεράνουμε πως η μηδενική υπόθεση ότι οι μηνιαίες δαπάνες δεν διαφέρουν ανάλογα με τον αριθμό των εξαρτώμενων μελών δεν μπορεί να απορριφθεί. Θα πρέπει, όμως, να έχουμε υπόψη μας ότι το αποτέλεσμα σχετικά με την υπόθεση της ισότητας των διασπορών δεν είναι ξεκάθαρο, οπότε καλό θα είναι να πραγματοποιήσουμε τον έλεγχο του Welch ή/και των Brown-Forsythe, οι οποίοι μπορούν να αντιμετωπίσουν τέτοιες περιπτώσεις. Εναλλακτικά, μπορούμε να μετασχηματίσουμε τις τιμές της εξαρτημένης μεταβλητής (δηλαδή, των μηνιαίων δαπανών), προκειμένου να σταθεροποιήσουμε τις διακυμάνσεις. Στη συνέχεια της ανάλυσης θα δείξουμε πώς γίνονται αυτοί οι ανθεκτικοί ή εύρωστοι (**robust**) έλεγχοι.

Πίνακας 6.3 *Αποτέλεσμα ελέγχου F ανάλυσης διακύμανσης.*

Tests of Between-Subjects Effects					
Dependent Variable: Expenditure					
Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	584922.655[a]	6	97487.109	1.317	.246
Intercept	8135387.946	1	8135387.946	109.943	.000
Dependents	584922.655	6	97487.109	1.317	**.246**
Error	97083014.602	1312	73996.200		
Total	142838568.806	1319			
Corrected Total	97667937.258	1318			
a. R Squared = .006 (Adjusted R Squared = .001)					

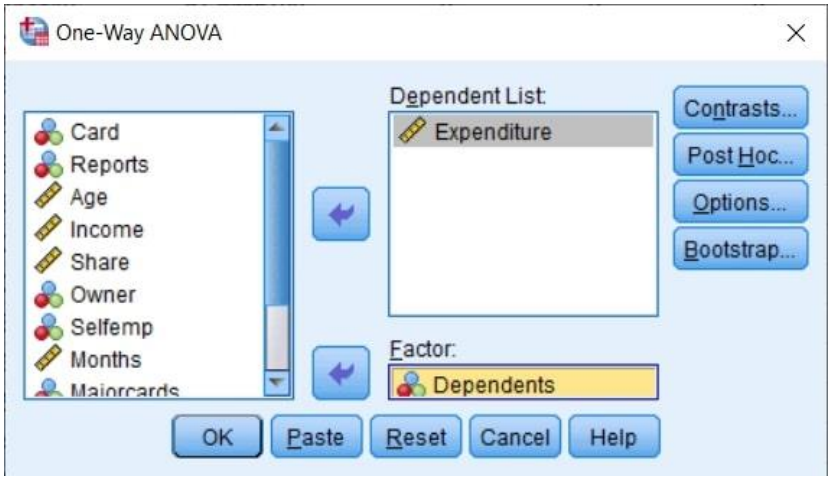
Συνοπτικά, τόσο οι έλεγχοι που πραγματοποιήθηκαν με την τεχνική του Bonferroni, η οποία βασίζεται στην κλασική ανάλυση διακύμανσης, όσο και οι έλεγχοι που έγιναν με την τεχνική του Tamhane, η οποία εφαρμόζεται στις περιπτώσεις που δεν ισχύει η υπόθεση της ομοσκεδαστικότητας, δεν ανιχνεύουν στατιστικά σημαντικές διαφορές μεταξύ των ζευγών των τιμών του παράγοντα. [3]

Γενικά, υπάρχουν δύο τρόποι για να αντιμετωπίσουμε το πρόβλημα της μη ισότητας των διακυμάνσεων των καταλοίπων κατά μήκος των επτά δειγμάτων. Ο ένας τρόπος είναι να μετασχηματίσουμε τις τιμές της εξαρτημένης μεταβλητής, χρησιμοποιώντας, για παράδειγμα, τον φυσικό λογάριθμο. Ο συγκεκριμένος τρόπος προσέγγισης χρησιμοποιήθηκε στην πολλαπλή γραμμική παλινδρόμηση. Ο δεύτερος τρόπος είναι να μεταβούμε σε πιο ανθεκτικές στις υποθέσεις του υποδείγματος μεθόδους, όπως είναι η ανάλυση διακύμανσης με τη μέθοδο του Welch και των Brown-Forsythe.

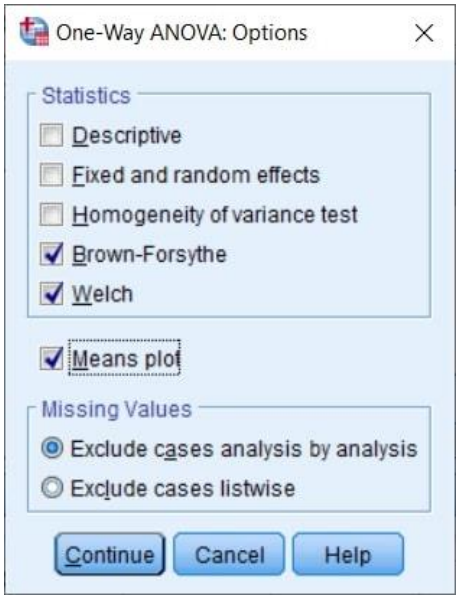
[3] Για λόγους παρουσίασης παραλείπεται ο αντίστοιχος πίνακας, καθώς είναι πολύ μεγάλος και καλύπτει σχεδόν 2 σελίδες σε εύρος.

6.2 Ανάλυση διακύμανσης με τις μεθόδους των Welch και Brown-Forsythe στο SPSS

Η ανάλυση διακύμανσης κατά έναν παράγοντα μπορεί να διεξαχθεί και με έναν εναλλακτικό τρόπο. Επιλέγοντας **Analyze → Compare Means → One-way ANOVA**, θα εμφανιστεί το παράθυρο της **Εικόνας 6.6**, στο οποίο περνάμε την εξαρτημένη μεταβλητή ("expenditure") και τον παράγοντα ("dependents") στα λευκά κουτιά που βρίσκονται στο δεξιό μέρος. Επιλέγοντας **Post Hoc**, θα εμφανιστεί το παράθυρο της **Εικόνας 6.5**, το οποίο μας επιτρέπει να διεξάγουμε πολλαπλούς ελέγχους μόνο, όμως, μέσω της επιλογής **Tamhane's T2**. Στη συνέχεια, πατώντας **Options** θα εμφανιστεί το παράθυρο της **Εικόνας 6.7**.



Εικόνα 6.6 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



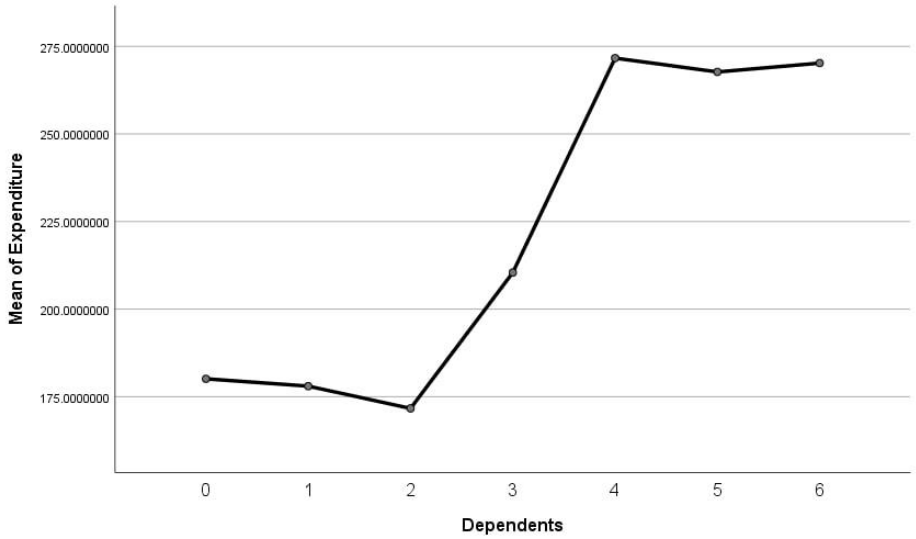
Εικόνα 6.7 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Στο παράθυρο της **Εικόνας 6.7** μπορούμε να επιλέξουμε όλες τις διαθέσιμες επιλογές εκτός από τις **Fixed and random effects** και **Homogeneity of variance test**, καθώς γνωρίζουμε ήδη ότι οι διακυμάνσεις διαφέρουν και συνεπώς δεν επιθυμούμε να εμφανιστούν και πάλι τα αποτελέσματα του ελέγχου του Levene. Επίσης, επιλέγουμε το **Means plot** (επιλογή που είναι, επίσης, διαθέσιμη και στην κλασική ανάλυση της διακύμανσης που περιγράψαμε στην προηγούμενη ενότητα). Πατώντας **Continue**

επιστρέφουμε στο παράθυρο της **Εικόνας 6.6** και πατώντας **ΟΚ** προκύπτουν τα αποτελέσματα που εμφανίζονται στον **Πίνακα 6.4** και στο **Διάγραμμα 6.2**. Επίσης, στο Output του SPSS εμφανίζεται και ο πίνακας ανάλυσης της διακύμανσης, τον οποίο, όμως, δεν παρουσιάζουμε, καθώς δεν έχει κάποιο ιδιαίτερο ενδιαφέρον.

Πίνακας 6.4 *Αποτελέσματα των εύρωστων ή ανθεκτικών ελέγχων.*

Robust Tests of Equality of Means				
Expenditure	Statistic[a]	df1	df2	Sig.
Welch	.875	6	49.355	**.520**
Brown-Forsythe	1.096	6	133.031	**.368**
a. Asymptotically F distributed.				



Διάγραμμα 6.2 *Διάγραμμα των μέσων.*

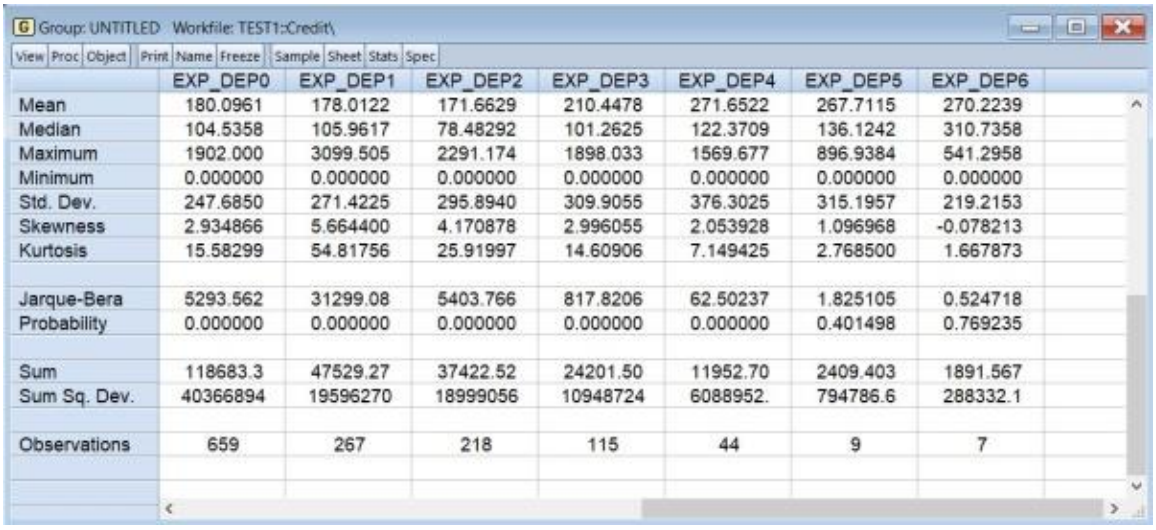
Ο **Πίνακας 6.4** περιλαμβάνει τα αποτελέσματα των ανθεκτικών (ή εύρωστων) ελέγχων. Όπως φαίνεται, τα παρατηρούμενα επίπεδα στατιστικής σημαντικότητας (*p*-values) είναι μεγαλύτερα του 0,05 και για τους δύο ελέγχους. Οπότε, τα συμπεράσματα είναι τα ίδια, όπως και προηγουμένως. Δηλαδή, ο αριθμός των εξαρτώμενων μελών του ατόμου δεν επηρεάζει στατιστικά σημαντικά τις μέσες μηνιαίες δαπάνες. Τέλος, στο **Διάγραμμα 6.2** παρουσιάζονται οι μέσοι ανά αριθμό εξαρτώμενων μελών.

6.3 Ανάλυση διακύμανσης κατά έναν παράγοντα (One-way ANOVA) στο Eviews

Η ανάλυση διακύμανσης κατά έναν παράγοντα, η οποία περιγράφηκε για το SPSS στις ενότητες 6.1 και 6.2, πραγματοποιείται με αρκετά εύκολο τρόπο στο Eviews. Όπως αναφέρθηκε και παραπάνω, όταν έχουμε δείγματα που προέρχονται από δύο ή και περισσότερους πληθυσμούς, οι υποθέσεις εφαρμογής της μεθόδου ανάλυσης διακύμανσης κατά έναν παράγοντα είναι οι ίδιες με την περίπτωση της απλής γραμμικής παλινδρόμησης (δηλαδή, κανονικότητα, ανεξαρτησία και ομοσκεδαστικότητα των καταλοίπων).

Έστω, λοιπόν, ότι θέλουμε να ελέγξουμε αν διαφοροποιούνται οι μέσοι, οι διάμεσες τιμές και οι διακυμάνσεις της μέσης μηνιαίας δαπάνης με τη χρήση πιστωτικής κάρτας (μεταβλητή “expenditure”) ανάλογα με το αν ο αριθμός εξαρτώμενων μελών του ατόμου είναι 0, 1, 2, 3, 4, 5 ή 6 (μεταβλητή “dependents”). Αρχικά, θα πρέπει να δημιουργήσουμε επτά νέες μεταβλητές με τον τρόπο που έχουμε ήδη περιγράψει (χωρίς τα μεγέθη των επτά δειγμάτων να είναι απαραίτητα ίσα). Η πρώτη θα περιλαμβάνει τις παρατηρήσεις της μεταβλητής “expenditure” για τα άτομα των οποίων ο αριθμός των εξαρτώμενων μελών

είναι 0 (έστω “exp_dep0”), η δεύτερη τις παρατηρήσεις της μεταβλητής “expenditure” για τα άτομα των οποίων ο αριθμός των εξαρτώμενων μελών είναι 1 (έστω “exp_dep1”) κ.ο.κ. Στο παράθυρο “Generate Series by Equation” **(Εικόνα 3.29)**, για την πρώτη μεταβλητή γράφουμε στο πάνω κουτί **exp_dep0=expenditure** και στο κάτω κουτί **if dependents=0** και πατάμε **OK**, για τη δεύτερη μεταβλητή γράφουμε στο πάνω κουτί **exp_dep1=expenditure** και στο κάτω κουτί **if dependents=1** και πατάμε **OK**, και η διαδικασία συνεχίζεται μέχρι να δημιουργήσουμε και τις επτά νέες μεταβλητές. Τα περιγραφικά στατιστικά μέτρα για τις επτά νέες μεταβλητές αυτές παρουσιάζονται στην **Εικόνα 6.8** και προκύπτουν, αν «ανοίξουμε» τις μεταβλητές αυτές ως Group και στη συνέχεια κάνουμε την επιλογή **View → Descriptive Stats → Individual Samples**.


Group: UNTITLED Workfile: TEST1::Credit\

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

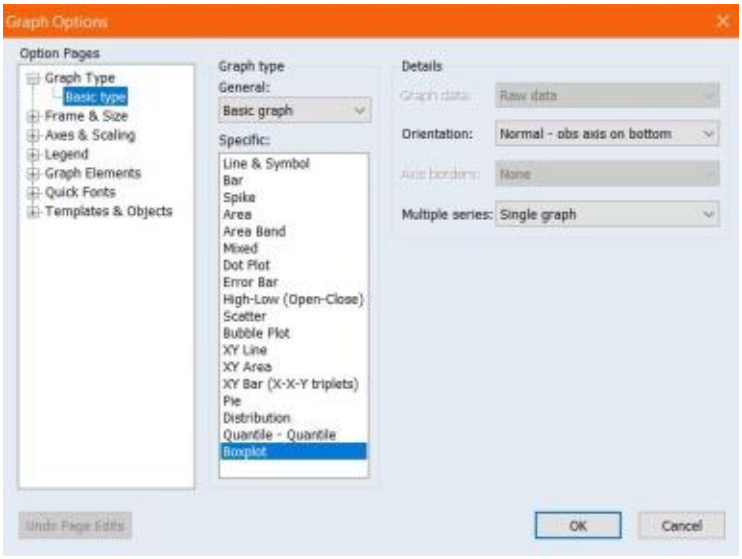
	EXP_DEP0	EXP_DEP1	EXP_DEP2	EXP_DEP3	EXP_DEP4	EXP_DEP5	EXP_DEP6
Mean	180.0961	178.0122	171.6629	210.4478	271.6522	267.7115	270.2239
Median	104.5358	105.9617	78.48292	101.2625	122.3709	136.1242	310.7358
Maximum	1902.000	3099.505	2291.174	1898.033	1569.677	896.9384	541.2958
Minimum	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Std. Dev.	247.6850	271.4225	295.8940	309.9055	376.3025	315.1957	219.2153
Skewness	2.934866	5.664400	4.170878	2.996055	2.053928	1.096968	-0.078213
Kurtosis	15.58299	54.81756	25.91997	14.60906	7.149425	2.768500	1.667873
Jarque-Bera	5293.562	31299.08	5403.766	817.8206	62.50237	1.825105	0.524718
Probability	0.000000	0.000000	0.000000	0.000000	0.000000	0.401498	0.769235
Sum	118683.3	47529.27	37422.52	24201.50	11952.70	2409.403	1891.567
Sum Sq. Dev.	40366894	19596270	18999056	10948724	6088952.	794786.6	288332.1
Observations	659	267	218	115	44	9	7

Εικόνα 6.8 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

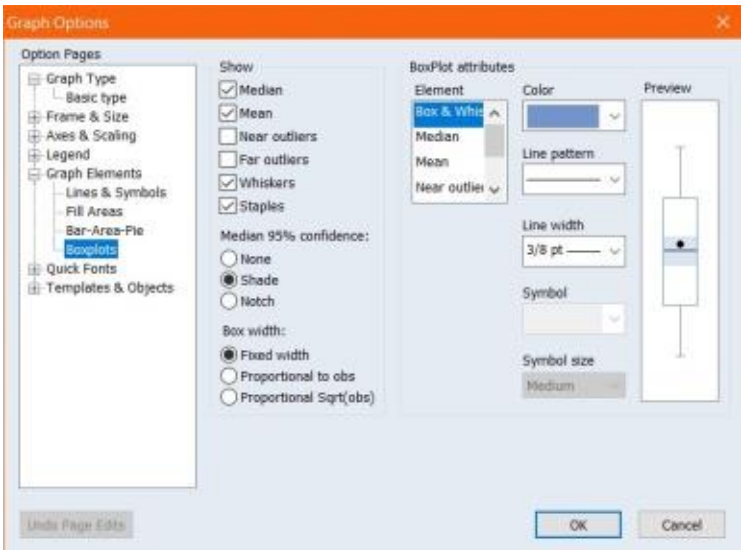
Πριν, όμως, διεξάγουμε τους παραπάνω ελέγχους, είναι χρήσιμο να δούμε πώς κατανέμεται γραφικά η μέση μηνιαία δαπάνη ανάλογα με τον αριθμό των εξαρτώμενων μελών, με τη χρήση ενός διαγράμματος **Box Plot**. Για να κατασκευάσουμε ένα τέτοιο διάγραμμα, «ανοίγουμε» τις επτά αυτές μεταβλητές ως Group και στη συνέχεια επιλέγουμε **View → Graph**. Στο παράθυρο “Graph Options” που θα εμφανιστεί **(Εικόνα 6.9)**, επιλέγουμε **Basic type** στην κατηγορία “Option Pages”, **Basic Graph** στην κατηγορία “General:” και **Boxplot** στην κατηγορία “Specific”. Επίσης, στο μενού “Option Pages” επιλέγουμε **Graph Elements → Boxplots** και αποεπιλέγουμε τα κουτιά “**Near outliers**” και “**Far outliers**”, αφήνοντας όλες τις υπόλοιπες επιλογές ως έχουν **(Εικόνα 6.10)**. Πατώντας **OK** εμφανίζεται το διάγραμμα **Box Plot (Εικόνα 6.11)**, στο οποίο η οριζόντια γραμμή μέσα σε κάθε ορθογώνιο είναι η διάμεσος, η σκιασμένη περιοχή γύρω από τη διάμεσο είναι το 95% διάστημα εμπιστοσύνης της, ενώ η τελεία είναι ο μέσος. Όπως αναλύθηκε προηγουμένως, τα άκρα κάθε ορθογωνίου είναι οι «εσωτερικοί φράχτες», ενώ τα άκρα των κάθετων γραμμών πάνω και κάτω από κάθε ορθογώνιο είναι οι «εξωτερικοί φράχτες». Σημεία εκτός του ορθογωνίου, αλλά εντός των «εσωτερικών φραχτών» είναι τα ήπια ακραία σημεία (“near outliers”), ενώ σημεία εκτός των «εξωτερικών φραχτών» είναι τα εξαιρετικά ακραία σημεία (“far outliers”), καθώς απέχουν περισσότερο από δύο ή τρεις τυπικές αποκλίσεις από τη μέση τιμή.

Όπως αναλύθηκε και στην ενότητα 6.1, αν έχουμε δείγματα που προέρχονται από πολλούς πληθυσμούς και θέλουμε να συγκρίνουμε τους μέσους τους, θα μπορούσαμε να χρησιμοποιήσουμε τον έλεγχο t για όλα τα πιθανά ζεύγη δειγμάτων. Έτσι, όμως, μειώνεται αισθητά η πιθανότητα να μην απορριφθεί η μηδενική υπόθεση σε όλους τους ελέγχους ταυτόχρονα. Για τον λόγο αυτό χρησιμοποιούμε την τεχνική της ανάλυσης διακύμανσης κατά έναν παράγοντα. Στο παράδειγμά μας, αν θέλουμε να ελέγξουμε την ισότητα των μέσων που προέρχονται από τους επτά πληθυσμούς, η μηδενική υπόθεση θα είναι $\boldsymbol{H_0}$: $\boldsymbol{\mu_1 = \mu_2 = \cdots = \mu_7}$, ενώ στην εναλλακτική υπόθεση $\boldsymbol{H_1}$ **δύο τουλάχιστον μέσοι δεν είναι ίσοι μεταξύ τους**. Το $\boldsymbol{\mu_1}$ είναι ο μέσος του πληθυσμού του πρώτου δείγματος (**dependents=0**), το $\boldsymbol{\mu_2}$ ο μέσος του πληθυσμού του δεύτερου δείγματος (**dependents=1**) κ.ο.κ. Προκειμένου, λοιπόν, να ελέγξουμε τη

συγκεκριμένη μηδενική υπόθεση, «ανοίγουμε» τις επτά μεταβλητές ως Group και στη συνέχεια επιλέγουμε **View → Tests of Equality**. Οπότε, θα εμφανιστεί το παράθυρο της **Εικόνας 4.17**, στο οποίο θα επιλέξουμε **Mean** και δεν θα «τσεκάρουμε» το κουτί **Common sample**, καθώς καθένα από τα επτά δείγματα έχει διαφορετικό αριθμό παρατηρήσεων. Στη συνέχεια, πατάμε **OK** και εμφανίζονται τα αποτελέσματα **(Εικόνα 6.12)**.



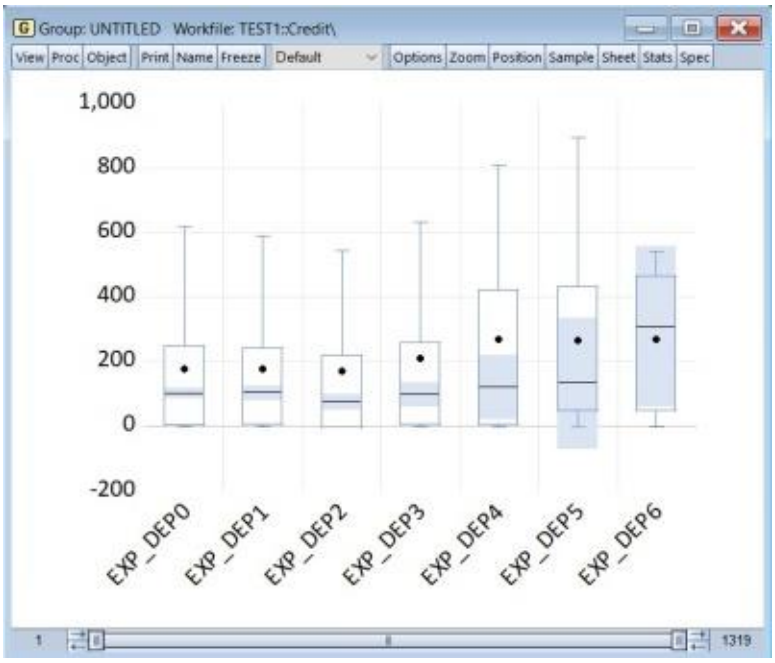
Εικόνα 6.9 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



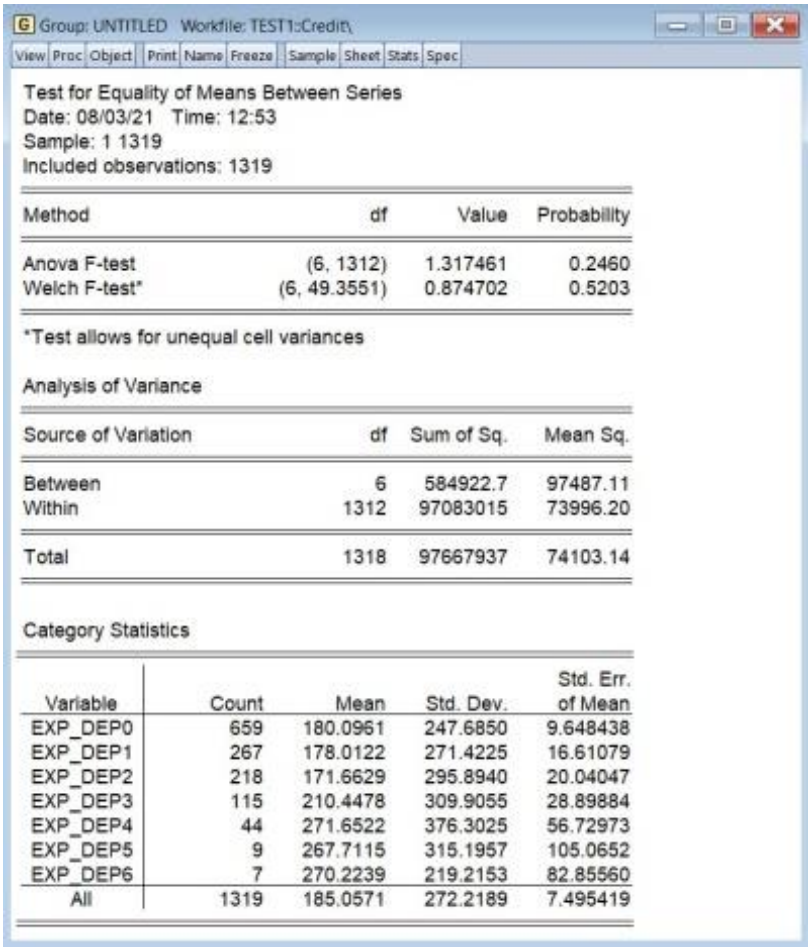
Εικόνα 6.10 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Στο πάνω μέρος της **Εικόνας 6.12** εμφανίζονται τα αποτελέσματα των στατιστικών ελέγχου. Το Eviews παρέχει δύο *F*-tests, το **Anova** και το **Welch**. Όπως αναφέρθηκε στην ενότητα 6.1, ο έλεγχος **Welch** προτιμάται στην περίπτωση που δεν ικανοποιείται η υπόθεση της ομοσκεδαστικότητας παρά μόνο αυτή της κανονικότητας, καθώς είναι ανθεκτικός σε περιπτώσεις ετεροσκεδαστικότητας. Στη στήλη **df** εμφανίζονται οι βαθμοί ελευθερίας για καθεμία από τις στατιστικές ελέγχου, στη στήλη **Value** οι εκτιμημένες τιμές των στατιστικών αυτών, ενώ στη στήλη **Probability** παρουσιάζονται οι αντίστοιχες *p*-values. Όπως προκύπτει από τα αποτελέσματα του πίνακα της **Εικόνας 6.12**, οι *p*-values είναι πολύ μεγαλύτερες του 0,05 και στις δύο στατιστικές ελέγχου, γεγονός που σημαίνει πως η μηδενική υπόθεση της ισότητας των επτά μέσων δεν μπορεί να απορριφθεί σε επίπεδο στατιστικής σημαντικότητας $a = 5\%$.

Επίσης, κάτω από τις στατιστικές ελέγχου παρουσιάζονται κάποιες επιπλέον πληροφορίες σχετικά με τις υπό εξέταση μεταβλητές.



Εικόνα 6.11 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Group: UNTITLED Workfile: TEST1::Credit\

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Test for Equality of Means Between Series
Date: 08/03/21 Time: 12:53
Sample: 1 1319
Included observations: 1319

Method	df	Value	Probability
Anova F-test	(6, 1312)	1.317461	0.2460
Welch F-test*	(6, 49.3551)	0.874702	0.5203

*Test allows for unequal cell variances

Analysis of Variance

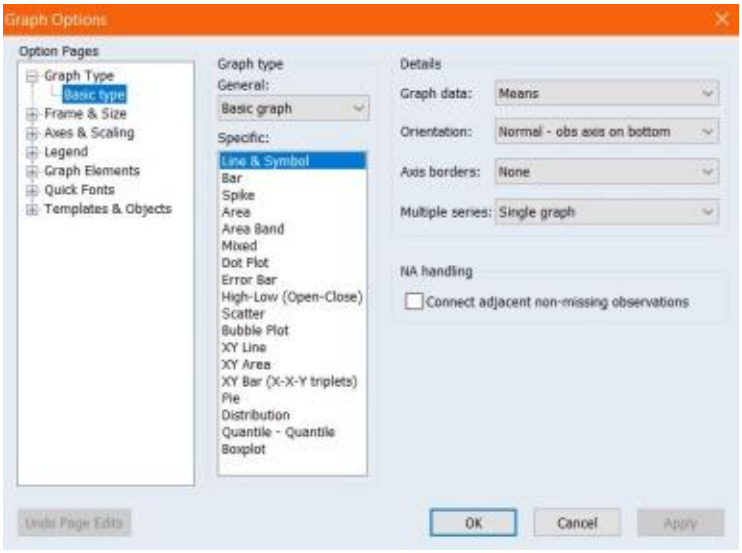
Source of Variation	df	Sum of Sq.	Mean Sq.
Between	6	584922.7	97487.11
Within	1312	97083015	73996.20
Total	1318	97667937	74103.14

Category Statistics

Variable	Count	Mean	Std. Dev.	Std. Err. of Mean
EXP_DEP0	659	180.0961	247.6850	9.648438
EXP_DEP1	267	178.0122	271.4225	16.61079
EXP_DEP2	218	171.6629	295.8940	20.04047
EXP_DEP3	115	210.4478	309.9055	28.89884
EXP_DEP4	44	271.6522	376.3025	56.72973
EXP_DEP5	9	267.7115	315.1957	105.0652
EXP_DEP6	7	270.2239	219.2153	82.85560
All	1319	185.0571	272.2189	7.495419

Εικόνα 6.12 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

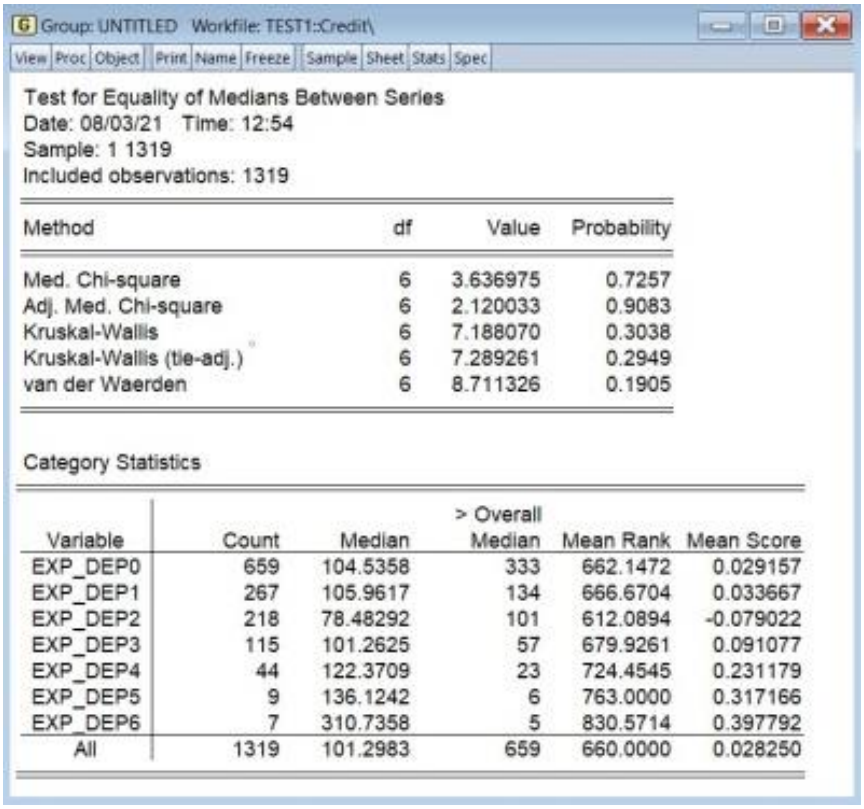
Στην ανάλυσή μας, μπορούμε, επίσης, να κατασκευάσουμε ένα διάγραμμα στο οποίο θα εμφανίζονται οι μέσες τιμές των επτά δειγμάτων. Για να το κάνουμε αυτό, «ανοίγουμε» τις επτά αυτές μεταβλητές ως Group και στη συνέχεια επιλέγουμε **View → Graph**. Στο παράθυρο "Graph Options" που θα εμφανιστεί, επιλέγουμε **Basic type** στην κατηγορία "Option Pages", **Basic Graph** στην κατηγορία "General:", **Line & Symbol** στην κατηγορία "Specific" και **Means** στο πεδίο "Graph data:" του μενού "Details", αφήνοντας όλες τις υπόλοιπες επιλογές ως έχουν **(Εικόνα 6.13)**. Πατώντας **OK**, εμφανίζεται το διάγραμμα με τις επτά μέσες τιμές **(Εικόνα 6.14)**, στο οποίο κάθε κυκλάκι αντιστοιχεί στη μέση τιμή των μηνιαίων δαπανών ανά αριθμό εξαρτώμενων μελών. Όπως φαίνεται στο συγκεκριμένο διάγραμμα, οι μέσες μηνιαίες δαπάνες αυξάνονται καθώς αυξάνεται ο αριθμός των εξαρτώμενων μελών.



Εικόνα 6.13 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Εικόνα 6.14 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Προκειμένου να ελέγξουμε την ισότητα των διάμεσων που προέρχονται από τους επτά πληθυσμούς, η μηδενική υπόθεση θα είναι $\boldsymbol{H_0}$: $\boldsymbol{median_1 = median_2 = \cdots = median_7}$, ενώ στην εναλλακτική υπόθεση $\boldsymbol{H_1}$ **δύο τουλάχιστον διάμεσοι δεν είναι ίσες μεταξύ τους**. Οπότε, «ανοίγουμε» και πάλι τις επτά μεταβλητές ως Group, επιλέγουμε **View → Tests of Equality** και στο παράθυρο της **Εικόνας 4.17** επιλέγουμε **Median** και δεν θα «τσεκάρουμε» το κουτί **Common sample**, όπως και στην περίπτωση του ελέγχου ισότητας των μέσων. Πατώντας **OK** εμφανίζονται τα αποτελέσματα **(Εικόνα 6.15)**.



Group: UNTITLED Workfile: TEST1::Credit\

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Test for Equality of Medians Between Series
Date: 08/03/21 Time: 12:54
Sample: 1 1319
Included observations: 1319

Method	df	Value	Probability
Med. Chi-square	6	3.636975	0.7257
Adj. Med. Chi-square	6	2.120033	0.9083
Kruskal-Wallis	6	7.188070	0.3038
Kruskal-Wallis (tie-adj.)	6	7.289261	0.2949
van der Waerden	6	8.711326	0.1905

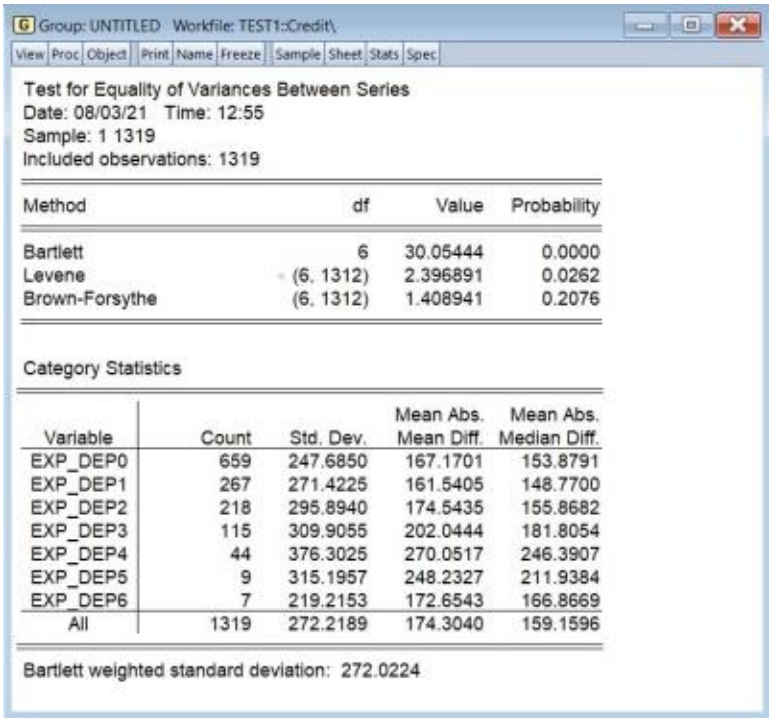
Category Statistics

Variable	Count	Median	> Overall Median	Mean Rank	Mean Score
EXP_DEP0	659	104.5358	333	662.1472	0.029157
EXP_DEP1	267	105.9617	134	666.6704	0.033667
EXP_DEP2	218	78.48292	101	612.0894	-0.079022
EXP_DEP3	115	101.2625	57	679.9261	0.091077
EXP_DEP4	44	122.3709	23	724.4545	0.231179
EXP_DEP5	9	136.1242	6	763.0000	0.317166
EXP_DEP6	7	310.7358	5	830.5714	0.397792
All	1319	101.2983	659	660.0000	0.028250

Εικόνα 6.15 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Στο πάνω μέρος του πίνακα της **Εικόνας 6.15** εμφανίζονται πέντε χ^2-tests: ο κλασικός έλεγχος για τη διάμεσο που βασίζεται στην ANOVA και η διορθωμένη μορφή του για συνέχεια, ο έλεγχος **Kruskal-Wallis** και η διορθωμένη μορφή του για συνέχεια, καθώς και ο έλεγχος **van der Waerden**. Όπως αναφέρθηκε στην ενότητα 6.1, ο έλεγχος **Kruskal-Wallis** είναι μη παραμετρικός και αναλύει τη διακύμανση κατά έναν παράγοντα βασισμένος στις τάξεις μεγέθους των τιμών της εξαρτημένης μεταβλητής. Υποθέτει, επίσης, ότι οι κατανομές των τιμών της εξαρτημένης μεταβλητής, οι οποίες δημιουργούνται για κάθε επίπεδο του συγκεκριμένου παράγοντα, έχουν το ίδιο σχήμα. Ο έλεγχος **van der Waerden** είναι ανάλογος του **Kruskal-Wallis**, με μόνη διαφορά ότι εξομαλύνει τα επίπεδα του συγκεκριμένου παράγοντα σε normal quantiles. Στη στήλη **df** εμφανίζονται οι βαθμοί ελευθερίας των στατιστικών ελέγχου, στη στήλη **Value** οι εκτιμημένες τιμές τους, ενώ στη στήλη **Probability** οι αντίστοιχες *p*-values και στις δύο στατιστικές ελέγχου. Καθώς οι *p*-values είναι πολύ μεγαλύτερες του 0,05 και για τις πέντε στατιστικές ελέγχου, η μηδενική υπόθεση της ισότητας των επτά διαμέσων δεν μπορεί να απορριφθεί σε επίπεδο στατιστικής σημαντικότητας $a = 5\%$. Επίσης, κάτω από τις στατιστικές ελέγχου εμφανίζονται κάποιες επιπλέον πληροφορίες σχετικά με τις υπό εξέταση μεταβλητές.

Τέλος, προκειμένου να ελέγξουμε την ισότητα των διακυμάνσεων που προέρχονται από τους επτά πληθυσμούς, η μηδενική υπόθεση θα είναι $\boldsymbol{H_0}: \boldsymbol{\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_7^2}$, ενώ στην εναλλακτική υπόθεση $\boldsymbol{H_1}$ **δύο τουλάχιστον διακυμάνσεις δεν είναι ίσες μεταξύ τους**. Και πάλι, «ανοίγουμε» τις επτά μεταβλητές ως Group, επιλέγουμε **View → Tests of Equality**, ενώ στο παράθυρο της **Εικόνας 4.17** θα επιλέξουμε **Variance** και δεν θα «τσεκάρουμε» το κουτάκι **Common sample**, όπως και προηγουμένως. Πατώντας **OK** εμφανίζονται τα αποτελέσματα **(Εικόνα 6.16)**.



Group: UNTITLED Workfile: TEST1::Credit\

View | Proc | Object | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Test for Equality of Variances Between Series
Date: 08/03/21 Time: 12:55
Sample: 1 1319
Included observations: 1319

Method	df	Value	Probability
Bartlett	6	30.05444	0.0000
Levene	(6, 1312)	2.396891	0.0262
Brown-Forsythe	(6, 1312)	1.408941	0.2076

Category Statistics

Variable	Count	Std. Dev.	Mean Abs. Mean Diff.	Mean Abs. Median Diff.
EXP_DEP0	659	247.6850	167.1701	153.8791
EXP_DEP1	267	271.4225	161.5405	148.7700
EXP_DEP2	218	295.8940	174.5435	155.8682
EXP_DEP3	115	309.9055	202.0444	181.8054
EXP_DEP4	44	376.3025	270.0517	246.3907
EXP_DEP5	9	315.1957	248.2327	211.9384
EXP_DEP6	7	219.2153	172.6543	166.8669
All	1319	272.2189	174.3040	159.1596

Bartlett weighted standard deviation: 272.0224

Εικόνα 6.16 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Στο πάνω μέρος του πίνακα της **Εικόνας 6.16** εμφανίζονται τρεις στατιστικές ελέγχου: ο χ^2 έλεγχος **Bartlett** και οι *F*-στατιστικές **Levene** και **Brown-Forsythe**. Ο έλεγχος **Levene** βασίζεται στην ANOVA της απόλυτης διαφοράς από τον μέσο, ενώ ο έλεγχος **Brown-Forsythe** τροποποιεί τον έλεγχο **Levene** αντικαθιστώντας την απόλυτη διαφορά από τον μέσο με την απόλυτη διαφορά από τη διάμεσο. Γενικά, αποδεικνύεται ότι ο έλεγχος **Brown-Forsythe** είναι προτιμότερος, καθώς είναι ανώτερος του ελέγχου **Levene** σε όρους ευρωστίας (**robustness**) και δύναμης (**power**). Στη στήλη **df** εμφανίζονται οι βαθμοί ελευθερίας των στατιστικών ελέγχου, στη στήλη **Value** οι εκτιμημένες τιμές τους και στη στήλη **Probability** οι αντίστοιχες *p*-values. Με βάση τα αποτελέσματα του συγκεκριμένου πίνακα, καθώς η *p*-value του ελέγχου **Brown-Forsythe** είναι 0,2076 > 0,05, η μηδενική υπόθεση τις ισότητας των επτά διακυμάνσεων δεν μπορεί να απορριφθεί σε επίπεδο στατιστικής σημαντικότητας $a = 5\%$. Επίσης, κάτω από τις στατιστικές ελέγχου εμφανίζονται κάποιες επιπλέον πληροφορίες σχετικά με τις υπό εξέταση μεταβλητές.

6.4 Ανάλυση διακύμανσης κατά δύο παράγοντες (Two-way ANOVA)

Όμως, τι γίνεται στην περίπτωση που έχουμε δύο κατηγορικές μεταβλητές, τη στατιστική σημαντικότητα των οποίων θέλουμε να εξετάσουμε; Είδαμε στο προηγούμενο παράδειγμα πως ο αριθμός των εξαρτώμενων μελών δεν επηρεάζει τις μηνιαίες δαπάνες. Όμως, πώς μπορούμε να ενσωματώσουμε και μία δεύτερη κατηγορική μεταβλητή; Δηλαδή, να εξετάσουμε, για παράδειγμα, αν οι μέσες μηνιαίες δαπάνες επηρεάζονται και από το αν κάποιος είναι αυτοαπασχολούμενος (δηλαδή, ελεύθερος επαγγελματίας) ή όχι. Με άλλα λόγια, να εξετάσουμε αν ο αριθμός των εξαρτώμενων μελών, αλλά και το είδος της απασχόλησης επιδρούν στατιστικά σημαντικά στις μηνιαίες δαπάνες. Ο συγκεκριμένος έλεγχος γίνεται μέσω της ανάλυσης διακύμανσης κατά δύο παράγοντες.

- Στο **SPSS**: Μεταβαίνουμε στο παράθυρο της **Εικόνας 6.3** και περνάμε και τη δεύτερη κατηγορική μεταβλητή (“selfemp”) στο κουτί **Fixed Factor(s):**, στο οποίο έχουμε ήδη περάσει την πρώτη κατηγορική μεταβλητή (“dependents”). Στη συνέχεια, ακολουθούμε την ίδια διαδικασία με προηγουμένως και προκύπτουν τα αποτελέσματα στο Output του SPSS. Η ανάλυσή μας θα εστιαστεί στον πίνακα της ανάλυσης διακύμανσης **(Πίνακας 6.5)**.

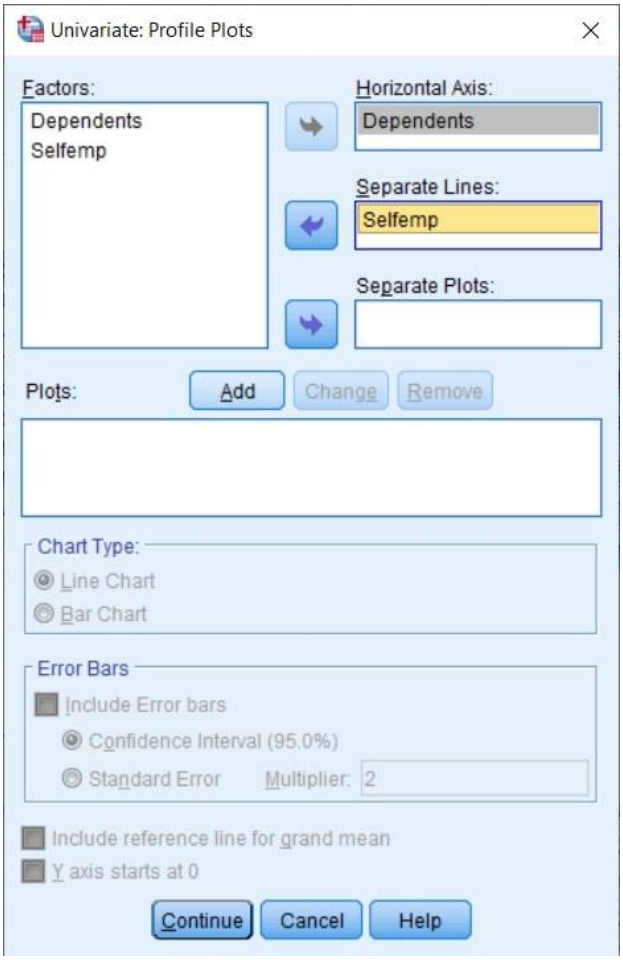
Πίνακας 6.5 *Ανάλυση διακύμανσης κατά δύο παράγοντες.*

Tests of Between-Subjects Effects					
Dependent Variable: Expenditure					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	965253.065[a]	12	80437.755	1.086	.368
Intercept	3024970.835	1	3024970.835	40.853	.000
Dependents	138356.281	6	23059.380	.311	**.931**
Selfemp	135327.562	1	135327.562	1.828	**.177**
Dependents * Selfemp	232363.831	5	46472.766	.628	**.679**
Error	96702684.193	1306	74044.934		
Total	142838568.806	1319			
Corrected Total	97667937.258	1318			
a. R Squared = .010 (Adjusted R Squared = .001)					

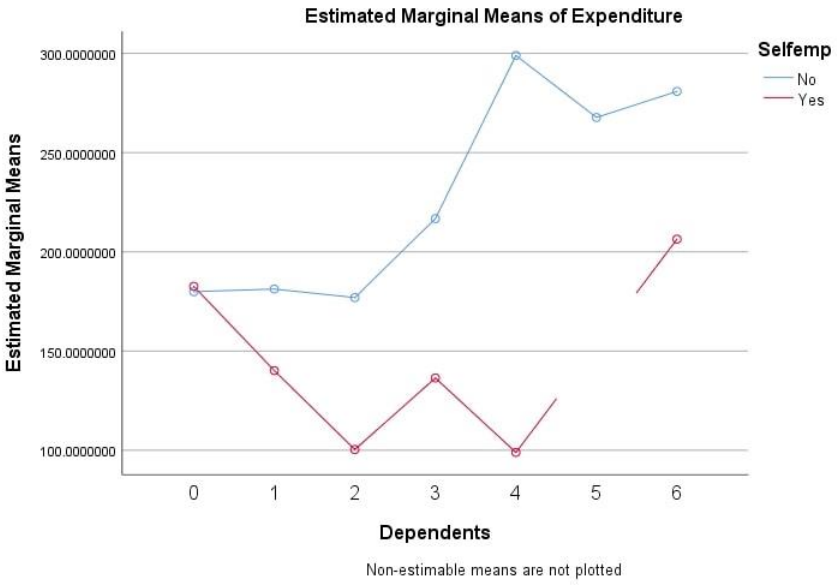
Όπως φαίνεται στον συγκεκριμένο πίνακα, η *p*-value και για τους δύο παράγοντες είναι μεγαλύτερη του 0,05. Αυτό σημαίνει ότι ούτε ο αριθμός των εξαρτώμενων μελών ούτε το είδος της εργασίας επηρεάζει τις μηνιαίες δαπάνες. Οι πολλαπλοί έλεγχοι είναι οι ίδιοι με προηγουμένως, με αποτέλεσμα να εξάγουμε σχεδόν τα ίδια συμπεράσματα. Παρατηρήστε, επίσης, ότι στον **Πίνακα 6.5** εμφανίζεται και μία επιπλέον γραμμή με αποτελέσματα (**Dependents * Selfemp**). Η συγκεκριμένη γραμμή αντιπροσωπεύει την αλληλεπίδραση μεταξύ του αριθμού των εξαρτώμενων μελών και του είδους της εργασίας, η οποία, όμως, δεν είναι στατιστικά σημαντική. Το τι αντιπροσωπεύει η συγκεκριμένη αλληλεπίδραση (που γενικά αναλύθηκε στην ενότητα 5.7) μπορούμε να το κατανοήσουμε καλύτερα γραφικά. Οπότε, στο παράθυρο της **Εικόνας 6.3** (και αφού έχουμε περάσει και τις δύο μεταβλητές στο δεξιό κουτί) επιλέγουμε **Options**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 6.17.**

Στο παράθυρο της **Εικόνας 6.17** θα περάσουμε τις δύο μεταβλητές στα δύο πάνω δεξιά κουτιά (**Horizontal Axis:** και **Separate Lines:**). Το ποια μεταβλητή θα περαστεί σε κάθε κουτί δεν έχει σημασία, καθώς η επεξήγηση θα είναι η ίδια, παρόλο που θα έχει διαφοροποιηθεί το γράφημα. Στη συνέχεια, θα πατήσουμε **Add** και μετά **Continue** (αν δεν πατήσουμε **Add**, θα εμφανιστεί ένα μήνυμα που θα μας προειδοποιεί πως ό,τι έχουμε κάνει μέχρι τώρα θα χαθεί και, επομένως, δεν θα κατασκευαστεί το διάγραμμα που επιθυμούμε). Πατώντας **OK** στο παράθυρο της **Εικόνας 6.17** θα εμφανιστεί το **Διάγραμμα 6.3**, στο οποίο υπάρχουν δύο γραμμές, μία για κάθε είδος απασχόλησης (δηλαδή, αν το άτομο είναι αυτοαπασχολούμενο ή όχι). Κάθε γραμμή έχει τόσα σημεία (κυκλάκια) όσα είναι και τα επίπεδα (ή οι τιμές) της κατηγορικής μεταβλητής, που στο παράδειγμά μας είναι ο αριθμός των εξαρτώμενων μελών. Κάθε κυκλάκι αντιπροσωπεύει τη μέση τιμή των μηνιαίων δαπανών, η οποία αντιστοιχεί στον κάθε αριθμό εξαρτώμενων μελών και στο είδος της απασχόλησης.

Στο **Διάγραμμα 6.3** παρατηρούμε ότι οι μέσες μηνιαίες δαπάνες αυξάνονται καθώς αυξάνεται ο αριθμός των εξαρτώμενων μελών, τόσο για τους ελεύθερους επαγγελματίες (κόκκινη γραμμή) όσο και για τους μισθωτούς (μπλε γραμμή). Οι δύο αυτές γραμμές ακολουθούν παράλληλες ή σχεδόν παράλληλες πορείες. Αυτό αποτελεί ένδειξη ότι η αλληλεπίδραση που εμφανίζεται στον **Πίνακα 6.5** δεν είναι στατιστικά σημαντική (όπως φαίνεται και από την αντίστοιχη *p*-value, η οποία είναι ίση με 0,679). Αυτό σημαίνει ότι ο τρόπος με τον οποίο αλλάζει η μέση τιμή της εξαρτημένης μεταβλητής για έναν παράγοντα (αριθμός εξαρτώμενων μελών) είναι παρόμοιος για όλα τα επίπεδα του άλλου παράγοντα (είδος απασχόλησης).

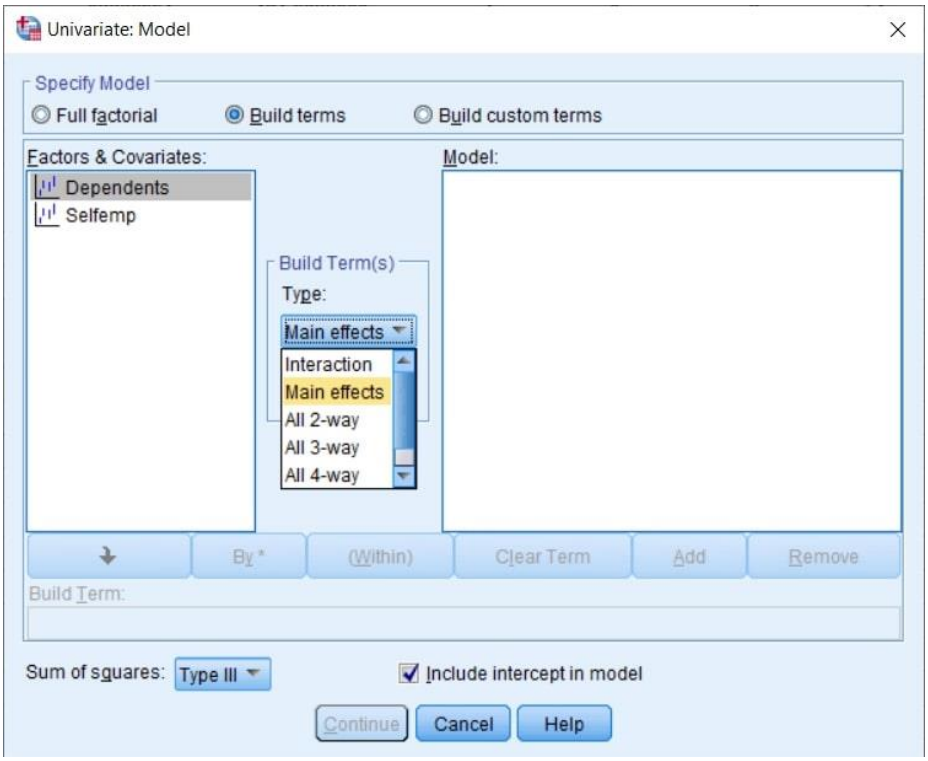


Εικόνα 6.17 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Διάγραμμα 6.3 *Γραφική αναπαράσταση αλληλεπίδρασης μεταξύ δύο κατηγορικών μεταβλητών.*

Καθώς η αλληλεπίδραση δεν είναι στατιστικά σημαντική, ας εξετάσουμε στη συνέχεια πώς μπορούμε να την αφαιρέσουμε από το υπόδειγμα. Καθώς έχουμε ήδη εκτιμήσει το υπόδειγμα (και άρα οι επιλογές μας δεν έχουν σβηστεί), επιστρέφουμε στο παράθυρο της **Εικόνας 6.3** και επιλέγουμε το **Model**, προκειμένου να εμφανιστεί το παράθυρο της **Εικόνας 6.18.** Στο συγκεκριμένο παράθυρο θα επιλέξουμε και πάλι **Build terms** και στην επιλογή **Type:** θα επιλέξουμε **Main effects**. Στη συνέχεια, θα περάσουμε τις δύο κατηγορικές μεταβλητές δεξιά και θα πατήσουμε **Continue**, προκειμένου να επιστρέψουμε στο παράθυρο της **Εικόνας 6.3**. Πατώντας **OK** θα εμφανιστούν στο Output του SPSS τα αποτελέσματα **(Πίνακας 6.6**), από τα οποία προκύπτει ότι και οι δύο παράγοντες εξακολουθούν να μην είναι στατιστικά σημαντικοί.



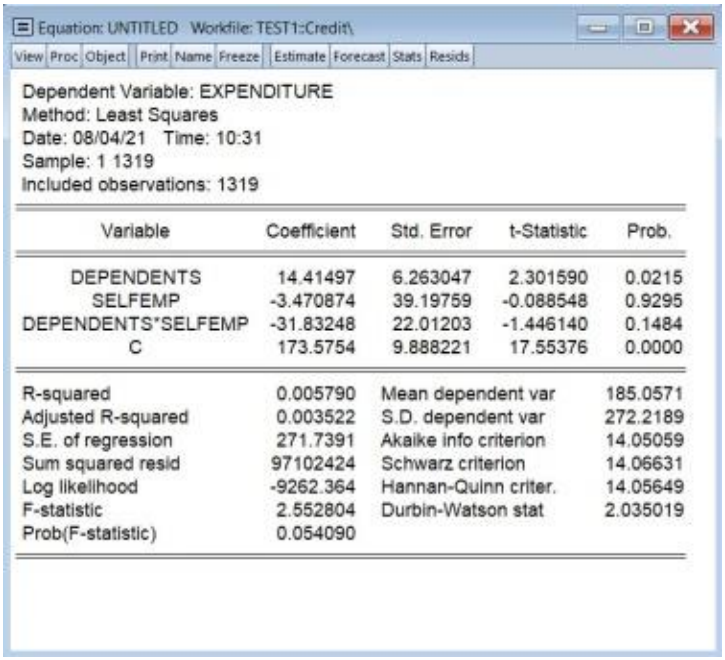
Εικόνα 6.18 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Πίνακας 6.6 *Ανάλυση διακύμανσης κατά δύο παράγοντες χωρίς αλληλεπίδραση.*

Tests of Between-Subjects Effects					
Dependent Variable: Expenditure					
Source	Type III Sum of Squares	Df	Mean Square	F	Sig.
Corrected Model	732889.233[a]	7	104698.462	1.416	.195
Intercept	5122501.948	1	5122501.948	69.279	.000
Dependents	608847.104	6	101474.517	1.372	**.222**
Selfemp	147966.578	1	147966.578	2.001	**.157**
Error	96935048.025	1311	73939.777		
Total	142838568.806	1319			
Corrected Total	97667937.258	1318			
a. R Squared = .008 (Adjusted R Squared = .002)					

- Στο **Eviews**: Έστω ότι, όπως και στην περίπτωση του SPSS, θέλουμε να εξετάσουμε αν οι δύο κατηγορικές μεταβλητές "dependents" (που υποδηλώνει τον αριθμό των εξαρτώμενων μελών) και "selfemp" (που υποδηλώνει το είδος της απασχόλησης) επιδρούν στατιστικά σημαντικά στις μηνιαίες δαπάνες ("expenditure"). Προκειμένου να πραγματοποιήσουμε τον

συγκεκριμένο έλεγχο, δημιουργούμε μία νέα εξίσωση (με τον τρόπο που έχουμε περιγράψει στην ενότητα 5.4) και επιλέγουμε ως εξαρτημένη μεταβλητή τη μεταβλητή "expenditure" και ως ανεξάρτητες μεταβλητές τις δύο αυτές κατηγορικές μεταβλητές, καθώς και το γινόμενό τους προκειμένου να εξετάσουμε αν η αλληλεπίδραση μεταξύ τους έχει στατιστικά σημαντικό αποτέλεσμα στις μέσες μηνιαίες δαπάνες. Όπως προκύπτει από τα αποτελέσματα που εμφανίζονται στον πίνακα της **Εικόνας 6.19**, ο αριθμός των εξαρτώμενων μελών ("dependents) επηρεάζει στατιστικά σημαντικά τις μέσες μηνιαίες δαπάνες, καθώς η αντίστοιχη *p*-value είναι μικρότερη του 0,05. Αντιθέτως, τόσο το είδος της απασχόλησης ("selfemp") όσο και η αλληλεπίδραση μεταξύ των δύο κατηγορικών μεταβλητών ("dependents*selfemp") δεν επηρεάζουν στατιστικά σημαντικά τις μέσες μηνιαίες δαπάνες, καθώς οι αντίστοιχες *p*-values είναι μεγαλύτερες του 0,05. Οπότε, η συγκεκριμένη εξίσωση θα πρέπει να επανεκτιμηθεί χωρίς το γινόμενο των δύο κατηγορικών μεταβλητών. Όπως προκύπτει από τα νέα αποτελέσματα **(Εικόνα 6.20)**, η επίδραση της μεταβλητής "dependents" παραμένει στατιστικά σημαντική, ενώ η επίδραση της μεταβλητής "selfemp" εξακολουθεί να μην είναι στατιστικά σημαντική.


Equation: UNTITLED Workfile: TEST1::Credit\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: EXPENDITURE
Method: Least Squares
Date: 08/04/21 Time: 10:31
Sample: 1 1319
Included observations: 1319

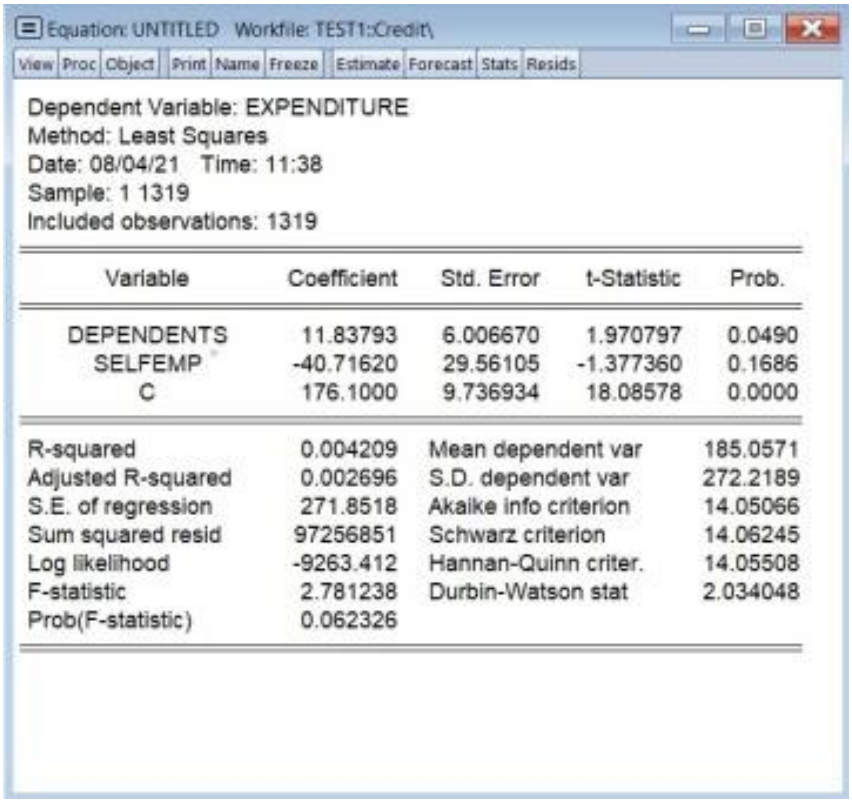
Variable	Coefficient	Std. Error	t-Statistic	Prob.
DEPENDENTS	14.41497	6.263047	2.301590	0.0215
SELFEMP	-3.470874	39.19759	-0.088548	0.9295
DEPENDENTS*SELFEMP	-31.83248	22.01203	-1.446140	0.1484
C	173.5754	9.888221	17.55376	0.0000

R-squared	0.005790	Mean dependent var	185.0571
Adjusted R-squared	0.003522	S.D. dependent var	272.2189
S.E. of regression	271.7391	Akaike info criterion	14.05059
Sum squared resid	97102424	Schwarz criterion	14.06631
Log likelihood	-9262.364	Hannan-Quinn criter.	14.05649
F-statistic	2.552804	Durbin-Watson stat	2.035019
Prob(F-statistic)	0.054090		

Εικόνα 6.19 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

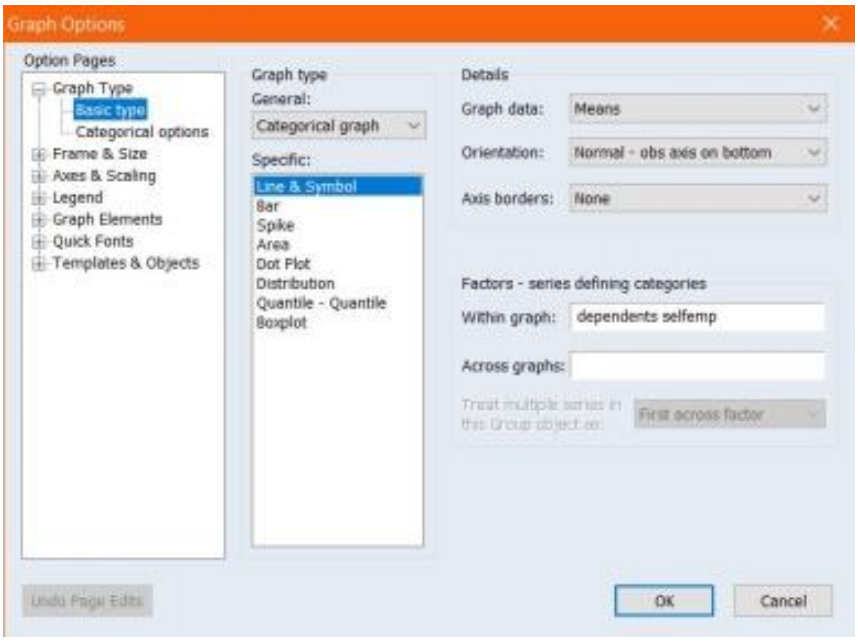
Προκειμένου να κατανοήσουμε καλύτερα το πώς διαφοροποιείται ο μέσος των μηνιαίων δαπανών ανάλογα με τον αριθμό των εξαρτώμενων μελών και το είδος της απασχόλησης, μπορούμε να δημιουργήσουμε το αντίστοιχο γράφημα. Για να το κάνουμε αυτό, «ανοίγουμε» τη μεταβλητή "expenditure" και στη συνέχεια επιλέγουμε **View → Graph**. Στο παράθυρο "Graph Options" που θα εμφανιστεί, επιλέγουμε **Basic type** στην κατηγορία "Option Pages", **Categorical Graph** στην κατηγορία "General:", **Line & Symbol** στην κατηγορία "Specific" και **Means** στο πεδίο "Graph data:" του μενού "Details". Επίσης, στο πεδίο "Within graph:" του μενού "Factors – series defining categories" γράφουμε με τη σειρά **dependents selfemp (Εικόνα 6.21)**. Η σειρά που θα γραφτούν οι μεταβλητές στο συγκεκριμένο πεδίο έχει σημασία, προκειμένου να δημιουργήσουμε το γράφημα που επιθυμούμε. Πατώντας **OK**, εμφανίζεται το διάγραμμα με τους επτά μέσους **(Εικόνα 6.22)**.[4] Κάθε κυκλάκι αντιπροσωπεύει τη μέση τιμή των μηνιαίων δαπανών, η οποία αντιστοιχεί στον κάθε αριθμό εξαρτώμενων μελών και στο είδος της απασχόλησης.

[4] Αν στο πεδίο "Within graph:" του μενού "Factors – series defining categories" γράψουμε μόνο **dependents,** τότε θα προκύψει το γράφημα της **Εικόνας 6.14.**

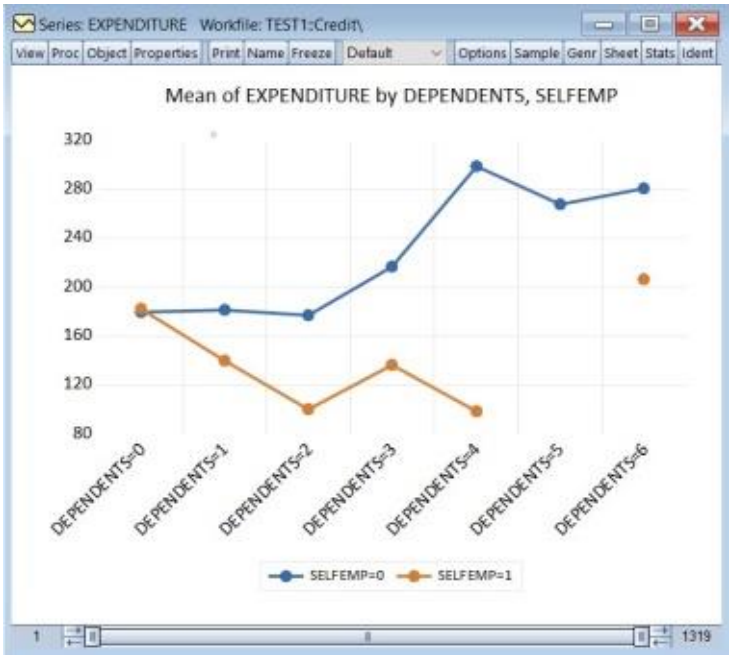


Εικόνα 6.20 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Όπως φαίνεται στο διάγραμμα της **Εικόνας 6.22**, οι μέσες μηνιαίες δαπάνες αυξάνονται καθώς αυξάνεται ο αριθμός των εξαρτώμενων μελών, τόσο για τους αυτοαπασχολούμενους (πορτοκαλί γραμμή) όσο και για τους μισθωτούς (μπλε γραμμή). Οι δύο αυτές γραμμές ακολουθούν περίπου παράλληλες πορείες, επιβεβαιώνοντας τα αποτελέσματα της **Εικόνας 6.19**, στα οποία φαίνεται ότι η αλληλεπίδραση μεταξύ των δύο κατηγορικών μεταβλητών δεν είναι στατιστικά σημαντική (*p*-value = 0,1484). Με άλλα λόγια, ο τρόπος με τον οποίο αλλάζει η μέση τιμή της εξαρτημένης μεταβλητής για τον έναν παράγοντα ("dependents") είναι παρόμοιος για κάθε επίπεδο του άλλου παράγοντα ("selfemp").



Εικόνα 6.21 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Εικόνα 6.22 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

6.5 Ανάλυση διακύμανσης για εξαρτημένα δείγματα

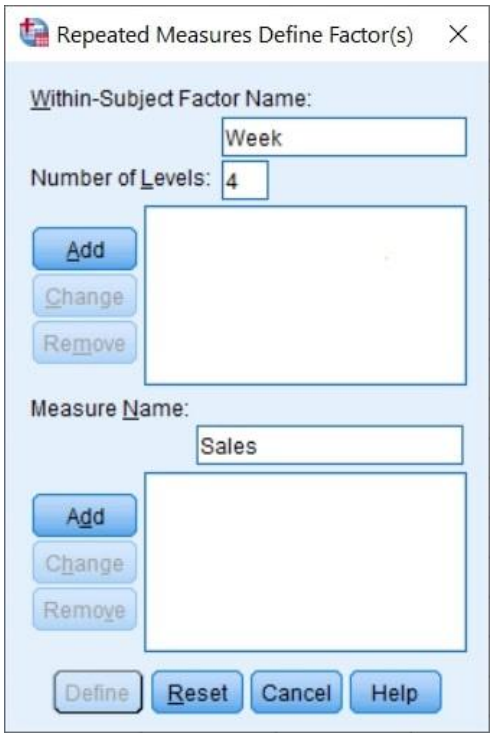
Μέχρι τώρα έχουμε εξετάσει τι συμβαίνει στην περίπτωση που τα δείγματα είναι ανεξάρτητα μεταξύ τους. Στην ενότητα αυτή, η ανάλυση θα επικεντρωθεί στην περίπτωση που τα δείγματα δεν είναι ανεξάρτητα. Στο κεφάλαιο 5 εξετάσαμε μία περίπτωση με δύο δείγματα, χρησιμοποιώντας τον έλεγχο *t* για δύο εξαρτημένα δείγματα. Στην ενότητα αυτή θα αναλύσουμε τι μπορούμε να κάνουμε, όταν έχουμε παραπάνω από δύο εξαρτημένα δείγματα. Η συγκεκριμένη τεχνική ονομάζεται στα αγγλικά *Repeated Measures ANOVA,* δηλαδή ανάλυση διακύμανσης για επαναλαμβανόμενες μετρήσεις.

Στον **Πίνακα 6.7** παρουσιάζεται η μορφή με την οποία θα πρέπει να καταχωριστούν τα δεδομένα στο SPSS. Τα δεδομένα βρίσκονται στο αρχείο **market.sav**. Θα πρέπει τα δείγματα να είναι το καθένα σε μία στήλη. Θυμηθείτε ότι στην περίπτωση των ανεξάρτητων δειγμάτων (ανάλυση διακύμανσης και έλεγχος των μέσων δύο ανεξάρτητων δειγμάτων) όλες οι μετρήσεις μας ήταν σε μία στήλη και υπήρχε και μία άλλη στήλη που παρουσίαζε το δείγμα στο οποίο ανήκε η κάθε μέτρηση. Στην περίπτωση, όμως, που έχουμε δύο εξαρτημένα δείγματα θα χρησιμοποιήσουμε οπωσδήποτε δύο στήλες.

Πίνακας 6.7 *Δεδομένα επαναλαμβανόμενων μετρήσεων.*

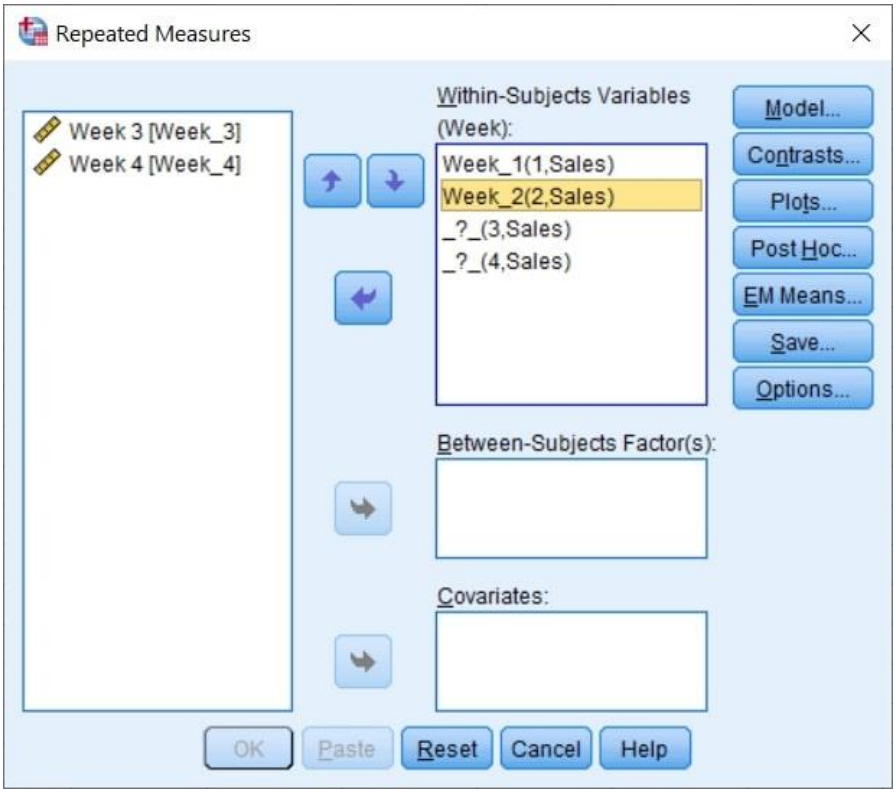
Προϊόν	Week1	Week2	Week3	Week4
1	18	14	12	6
2	19	12	8	4
3	14	10	6	2
4	16	12	10	4
5	12	8	6	2
6	18	10	5	1
7	16	10	8	4
8	18	8	4	1
9	16	12	6	2
10	19	16	10	8
11	16	14	10	9
12	16	12	8	8

Στον **Πίνακα 6.7** περιέχονται μερικές τιμές για την αγορά 12 προϊόντων στη διάρκεια 4 εβδομάδων. Ο σκοπός της ανάλυσής μας είναι να εξετάσουμε αν υπάρχει διαφορά μεταξύ των μέσων τιμών των 4 εβδομάδων. Προκειμένου να διεξάγουμε την ανάλυση αυτή στο SPSS, επιλέγουμε **Analyze → General Linear Model → Repeated Measures**, προκειμένου να εμφανιστεί το παράθυρο της **Εικόνας 6.23**. Στο μικρό κουτί με την ένδειξη **Within-Subject Factor Name:** δίνουμε ένα όνομα στον παράγοντα για τον οποίο ενδιαφερόμαστε (έστω "Week"). Στο κουτί με την ένδειξη **Number of Levels:** βάζουμε τον αριθμό των επαναλαμβανόμενων μετρήσεων που έχουμε, όπου στην περίπτωσή μας έχουμε μετρήσεις για 4 εβδομάδες. Τέλος, στο κουτί με την ένδειξη **Measure Name:** μπορούμε να δώσουμε ένα όνομα στις μετρήσεις, το οποίο να περιγράφει σε τι αναφέρονται. Στο παράδειγμά μας δώσαμε το όνομα "Sales" (δηλαδή, πωλήσεις).



Εικόνα 6.23 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Στο παράθυρο της **Εικόνας 6.23**, πατώντας **Add** και στις δύο επιλογές και στη συνέχεια **Define** θα μεταφερθούμε στο παράθυρο της **Εικόνας 6.24**. Στο δεξιό μέρος του συγκεκριμένου παραθύρου, στο λευκό κουτί με την ένδειξη **Within-Subjects Variables (Week):** μπορούμε να περάσουμε τα τέσσερα δείγματα (στήλες) που έχουμε στη διάθεσή μας. Επιλέγοντας **Plots**, μπορούμε να κατασκευάσουμε ένα διάγραμμα με τους μέσους των δειγμάτων. Επιλέγοντας **Options** (επιλογή που είναι παρόμοια με αυτή του παραθύρου της **Εικόνας 6.4** που είδαμε στην ανάλυση διακύμανσης) μπορούμε να επιλέξουμε να εμφανιστούν κάποια περιγραφικά μέτρα για τα τέσσερα δείγματά μας, καθώς και οι εκτιμήσεις των παραμέτρων. Στη συνέχεια, μπορούμε, αν θέλουμε, να επιλέξουμε να αποθηκεύσουμε τα τυποποιημένα κατάλοιπα, προκειμένου να διεξάγουμε τον έλεγχο κανονικότητας. Αφού επιλέξουμε ότι επιθυμούμε, επιστρέφουμε στο παράθυρο της **Εικόνας 6.24** και πατάμε **OK**, προκειμένου να εμφανιστούν τα αποτελέσματα. Στο Output του SPSS θα εμφανιστεί μια σειρά από πίνακες, αλλά εμείς θα εστιάσουμε στους πίνακες που αφορούν τον έλεγχο σφαιρικότητας του Mauchly **(Πίνακας 6.8)** και τους ελέγχους για την ισότητα των μέσων **(Πίνακας 6.9)**.



Εικόνα 6.24 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Στον **Πίνακα 6.8** παρουσιάζεται ο έλεγχος σφαιρικότητας (**Sphericity**) του Mauchly, οποίος ελέγχει μία λίγο διαφορετική υπόθεση σχετικά με τις διακυμάνσεις των καταλοίπων. Καλό είναι να μην απορρίπτεται η συγκεκριμένη υπόθεση. Στο παράδειγμά μας, η *p*-value για τον έλεγχο αυτό είναι 0,112 > 0,05, οπότε η μηδενική υπόθεση δεν μπορεί να απορριφθεί. Ο **Πίνακας 6.9** παρουσιάζει 4 διαφορετικούς ελέγχους. Ο πρώτος έλεγχος υποθέτει ότι η υπόθεση της σφαιρικότητας ισχύει (όπως στο παράδειγμά μας). Αν, όμως, η υπόθεση αυτή παραβιάζεται, τότε εστιάζουμε στους τρεις επόμενους ελέγχους και ιδιαίτερα σε αυτούς των Greenhouse-Geisser και των Huynh-Feldt.

Σχετικά με την υπόθεση της κανονικότητας των καταλοίπων, μπορούμε να αποθηκεύσουμε τα κατάλοιπα, όπως και προηγουμένως, και να διεξάγουμε τον συγκεκριμένο έλεγχο. Όπως προκύπτει, η υπόθεση της ισότητας των μέσων απορρίπτεται σε επίπεδο σημαντικότητας $a = 5\%$. Δηλαδή, κάποιος ή κάποιοι μέσοι διαφέρουν μεταξύ τους. Στην περίπτωση αυτή μπορούμε να διεξάγουμε τον έλεγχο *t* για όλα τα δυνατά ζεύγη παρατηρήσεων. Όμως, αυτό δεν είναι σωστό, καθώς η διακύμανση δεν είναι η ίδια για όλα τα δείγματα (όπως έχουμε υποθέσει).

Πίνακας 6.8 *Έλεγχος σφαιρικότητας του Mauchly.*

Mauchly's Test of Sphericity[a]

Measure: Sales

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b] Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Week	.398	8.957	5	.112	.622	.744	.333

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept
Within Subjects Design: Week

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Πίνακας 6.9 *Έλεγχοι για την ισότητα των μέσων.*

Tests of Within-Subjects Effects						
Measure: Sales						
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Week	Sphericity Assumed	991.500	3	330.500	127.561	**.000**
	Greenhouse-Geisser	991.500	1.865	531.709	127.561	**.000**
	Huynh-Feldt	991.500	2.231	444.504	127.561	**.000**
	Lower-bound	991.500	1.000	991.500	127.561	**.000**
Error (Week)	Sphericity Assumed	85.500	33	2.591		
	Greenhouse-Geisser	85.500	20.512	4.168		
	Huynh-Feldt	85.500	24.536	3.485		
	Lower-bound	85.500	11.000	7.773		

Θα πρέπει να επισημάνουμε στο σημείο αυτό ότι το **Eviews** δεν περιέχει την τεχνική *Repeated Measures ANOVA*, καθώς και τους ελέγχους που συμπεριλαμβάνονται σε αυτή (Mauchly's Test of Sphericity, Greenhouse-Geisser, Huynh-Feldt και Lower-bound), προκειμένου να αναλυθεί η διακύμανση για εξαρτημένα δείγματα.

Βιβλιογραφία

Ελληνόγλωσση

Δαφέρμος, Β. (2002). *Επαναληπτικές Στατιστικές Μετρήσεις στις Κοινωνικές Επιστήμες*. Αθήνα: Εκδόσεις Leader Books.

Παυλόπουλος, Β. (2008). *Μοντέλα Ανάλυσης Διακύμανσης*. Αθήνα: Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών.

Τσαγρής, Μ. (2006). *Ανάλυση Διακύμανσης στο SPSS* (Πανεπιστημιακές Σημειώσεις).

Ξενόγλωσση

Montgomery, D.C. (2001). *Design and Analysis of Experiments* (5th ed.). John and Wiley and Sons Inc.

Wooldridge, J.M. (2013). *Introductory Econometrics: A Modern Approach* (5th ed.). Mason, Ohio: South-Western, Cengage Learning.

Κεφάλαιο 7 Λογιστική παλινδρόμηση και διαχωριστική ανάλυση

Σύνοψη
Το κεφάλαιο αυτό επικεντρώνεται σε προχωρημένες μεθόδους της στατιστικής, όπως είναι η μη γραμμική παλινδρόμηση, η λογιστική παλινδρόμηση για την περίπτωση μιας δίτιμης εξαρτημένης μεταβλητής, η καμπύλη λειτουργικού χαρακτηριστικού δείκτη και η διαχωριστική ανάλυση. Οι στόχοι του κεφαλαίου είναι να μπορεί ο χρήστης του SPSS και του Eviews να κατανοεί τις προχωρημένες αυτές τεχνικές και να τις εφαρμόζει ορθά στα συγκεκριμένα προγράμματα.

Προαπαιτούμενη γνώση
Απατούνται γνώσεις προχωρημένης στατιστικής.

7.1 Προχωρημένη απλή παλινδρόμηση (μία ανεξάρτητη μεταβλητή)

Στην ενότητα αυτή θα αναλύσουμε την απλή παλινδρόμηση σε πιο προχωρημένο επίπεδο. Θα πρέπει, επίσης, να υπενθυμίσουμε ότι εν γένει στη γραμμική παλινδρόμηση υποθέτουμε γραμμική σχέση μεταξύ της ανεξάρτητης και της εξαρτημένης μεταβλητής.

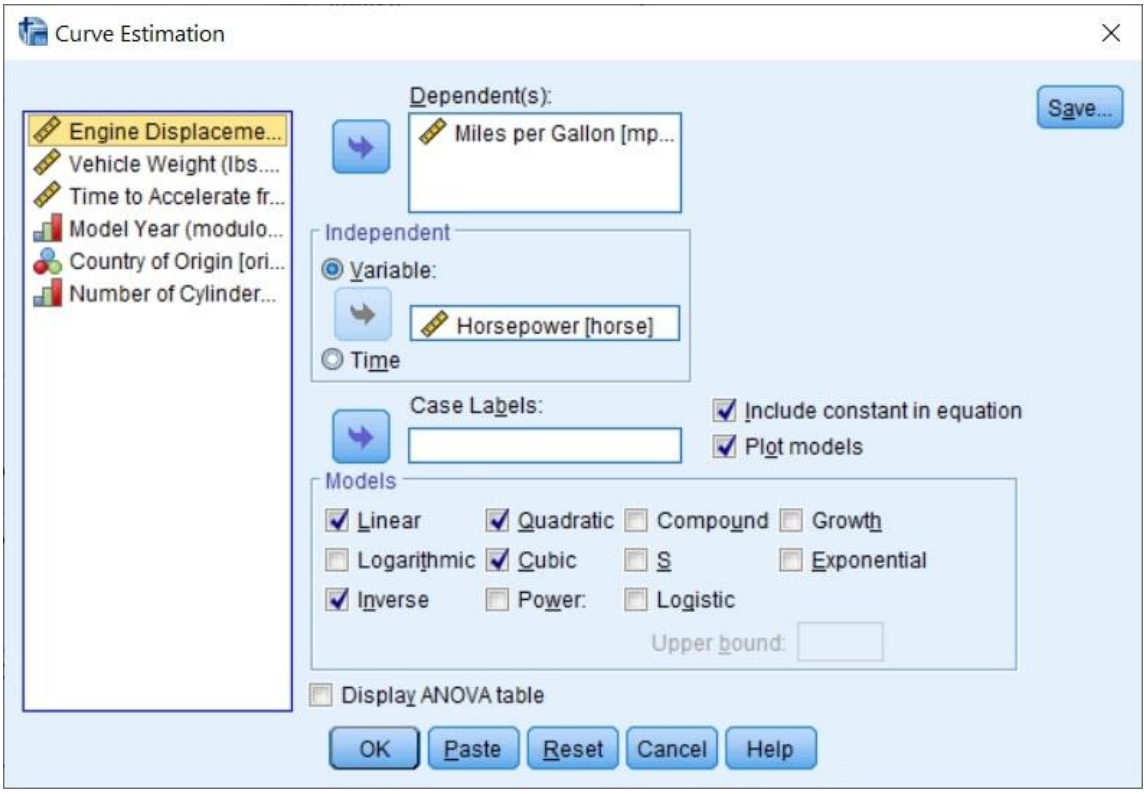
- Στο **SPSS:** Προκειμένου να εξετάσουμε την απλή παλινδρόμηση σε πιο προχωρημένο επίπεδο, θα χρησιμοποιήσουμε ένα καινούριο σετ δεδομένων, το οποίο αναφέρεται σε αυτοκίνητα **(cars.sav).** Πιο συγκεκριμένα, θα εστιάσουμε στη σχέση μεταξύ της κατανάλωσης βενζίνης (μίλια ανά γαλόνι – *miles per gallon)* και της ιπποδύναμης (horsepower). Όπως φαίνεται και από το διάγραμμα διασποράς **(Διάγραμμα 7.1)**, του οποίου την κατασκευή θα εξηγήσουμε στη συνέχεια, η σχέση μεταξύ κατανάλωσης και ιπποδύναμης προφανώς και δεν είναι γραμμική.

Προκειμένου, όμως, να γίνει πιο κατανοητή η ανάλυση της προχωρημένης παλινδρόμησης, εφαρμόζουμε την ακόλουθη διαδικασία. Επιλέγουμε **Analyze → Regression → Curve Estimation**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 7.1**. Στο παράθυρο αυτό έχουμε περάσει τις μεταβλητές που μας ενδιαφέρουν στα δεξιά κουτιά με τις ενδείξεις Dependent(s) (εξαρτημένη) και Independent Variable: (ανεξάρτητη). Επίσης, έχουμε «τσεκάρει» τις επιλογές **Quadratic**, **Cubic**, **Inverse** και **Display ANOVA table**. Έστω, λοιπόν, ότι με $\boldsymbol{y}$ συμβολίζουμε την εξαρτημένη μεταβλητή και με $\boldsymbol{x}$ την ανεξάρτητη. Ο **Πίνακας 7.1** παρουσιάζει τα υποδείγματα που έχουμε επιλέξει.

Πίνακας 7.1 *Μοντέλα και μαθηματική έκφρασή τους.*

Μοντέλο	Μαθηματική έκφραση
Γραμμικό (Linear)	$\boldsymbol{y = \alpha + \beta x}$
Τετραγωνικό (Quadratic)	$\boldsymbol{y = \alpha + \beta x + \gamma x^2}$
Κυβικό (Cubic)	$\boldsymbol{y = \alpha + \beta x + \gamma x^2 + \delta x^3}$
Αντίστροφο (Inverse)	$\boldsymbol{\frac{1}{y} = \alpha + \beta x}$ **ή** $\boldsymbol{y = \alpha + \beta x}$

Πατώντας **ΟΚ** θα παρουσιαστούν τα αποτελέσματα στο Output του SPSS, με τη μορφή μιας σειράς πινάκων, καθώς και του **Διαγράμματος 7.1**. Όπως προκύπτει από το συγκεκριμένο διάγραμμα διασποράς, η ευθεία γραμμή δεν φαίνεται να ταιριάζει και στα δεδομένα μας. Αντιθέτως, οι καμπύλες γραμμές φαίνεται να ταιριάζουν περισσότερο. Ο **Πίνακας 7.2** παρουσιάζει τους συντελεστές προσδιορισμού, καθώς και τις προσαρμοσμένες τιμές των συντελεστών αυτών που λαμβάνουν υπόψη τον αριθμό των παρατηρήσεων και των παραμέτρων του υποδείγματος.

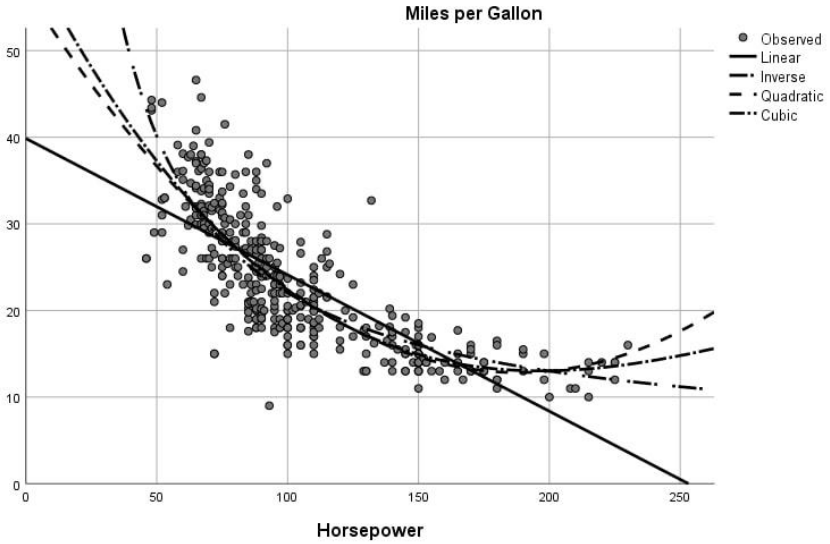


Εικόνα 7.1 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Πίνακας 7.2 *Συντελεστές προσδιορισμού των μοντέλων.*

Μοντέλο	R	R Square	Adjusted R Square	Std. Error of the Estimate
Γραμμικό	.771	.595	.594	4.974
Αντίστροφο	.812	.659	.658	4.562
Τετραγωνικό	.824	.679	.677	4.437
Κυβικό	.824	.679	.677	4.436

Στην ανάλυσή μας θα εστιάσουμε στις τιμές των προσαρμοσμένων συντελεστών προσδιορισμού του κάθε υποδείγματος. Υπενθυμίζουμε ότι ο συντελεστής προσδιορισμού είναι το ποσοστό της μεταβλητότητας (διακύμανσης) της εξαρτημένης μεταβλητής που ερμηνεύεται από την ανεξάρτητη μεταβλητή και εν γένει από το υπόδειγμα. Όπως φαίνεται από τον **Πίνακα 7.2**, ο προσαρμοσμένος συντελεστής προσδιορισμού έχει τη χαμηλότερη τιμή στο γραμμικό υπόδειγμα, ενώ η τιμή του είναι η ίδια στο τετραγωνικό και στο κυβικό υπόδειγμα. Το τελευταίο συμβαίνει, καθώς με βάση τα αποτελέσματα του SPSS (τα οποία δεν παρουσιάζονται στην ενότητα αυτή για λόγους συντομίας), ο κυβικός όρος δεν είναι στατιστικά σημαντικός. Δηλαδή, είτε έχουμε το τετραγωνικό υπόδειγμα είτε έχουμε το κυβικό υπόδειγμα, δεν αλλάζει κάτι. Οπότε, με βάση τα αποτελέσματα αυτά προτιμάμε το τετραγωνικό υπόδειγμα. Αν, όμως, η διαφορά μεταξύ του γραμμικού και του τετραγωνικού υποδείγματος ήταν μικρή (για παράδειγμα, αν από το 0,594 πηγαίναμε στο 0,6) τότε ίσως δεν θα ήταν σκόπιμο να επιλέξουμε το τετραγωνικό υπόδειγμα. Ωστόσο, αυτός ο τρόπος επιλογής υποδείγματος είναι πολύ απλός. Υπάρχουν πολύ πιο αποτελεσματικοί τρόποι, οι οποίοι όμως δεν θα αναλυθούν, καθώς αυτό ξεπερνάει τους σκοπούς του παρόντος βιβλίου.

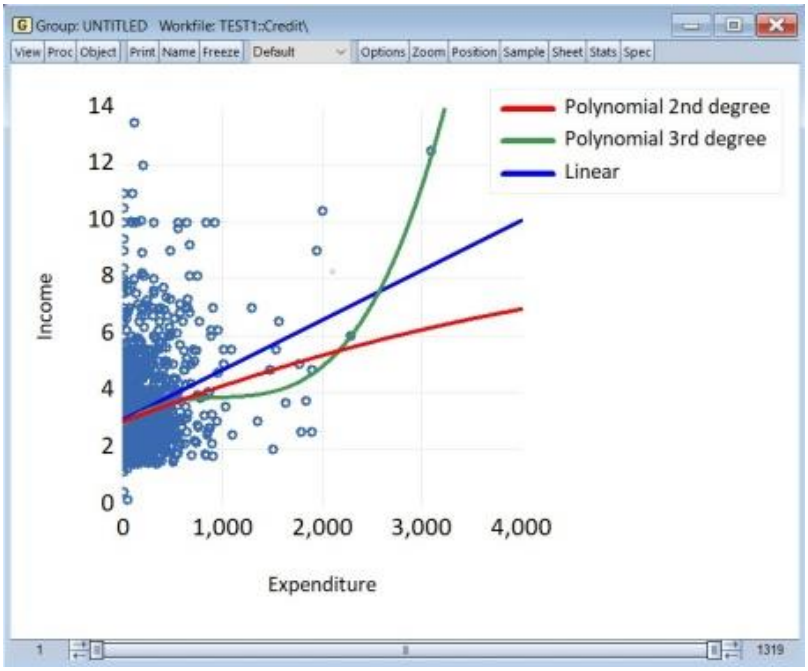


Διάγραμμα 7.1 *Διάγραμμα διασποράς με τις γραμμές των διάφορων μοντέλων.*

- Στο **Eviews**: Όπως αναφέρθηκε παραπάνω, η απλή παλινδρόμηση που εκτιμήσαμε στην ενότητα 5.3 του κεφαλαίου 5, χρησιμοποιώντας τη μεταβλητή "expenditure" ως εξαρτημένη μεταβλητή $\boldsymbol{y}$ και τη μεταβλητή "income" ως ανεξάρτητη μεταβλητή $\boldsymbol{x}$ (eq01), υποθέτει ότι η σχέση μεταξύ των δύο αυτών μεταβλητών είναι **γραμμική** (**linear**): $\boldsymbol{y = \alpha + \beta_1 x.}$ Όμως, από το διάγραμμα διασποράς της **Εικόνας 5.6** δεν προκύπτει κάποιο ασφαλές συμπέρασμα σχετικά με τη συγκεκριμένη υπόθεση. Συνεπώς, θα πρέπει να διερευνήσουμε τη μορφή της συγκεκριμένης σχέσης και να εκτιμήσουμε πιθανά εναλλακτικά υποδείγματα, όπως για παράδειγμα:
- **Πολυωνυμικό 2ου βαθμού (quadratic):** $y = \alpha + \beta_1 x + \beta_2 x^2$
- **Πολυωνυμικό 3ου βαθμού (cubic):** $y = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

Αρχικά, μπορούμε να κατασκευάσουμε ένα διάγραμμα διασποράς στο οποίο θα περιλαμβάνονται τρεις εναλλακτικές γραμμές παλινδρόμησης, ανάλογα με τη μορφή της ανεξάρτητης μεταβλητής $\boldsymbol{x}$ ("income"): **γραμμική**, **πολυωνυμική 2ου βαθμού** και **πολυωνυμική 3ου βαθμού**. Ακολουθούμε ακριβώς την ίδια διαδικασία που περιγράψαμε στην ενότητα 5.2, με μόνη διαφορά ότι στην επιλογή "Fit lines:" επιλέγουμε **Regression Line**. Στη συνέχεια, πατάμε το κουμπί "Options" που βρίσκεται ακριβώς δίπλα και εμφανίζεται το παράθυρο της **Εικόνας 7.2**. Στο πεδίο "Added Elements" έχει προστεθεί το στοιχείο "Regression Line", το οποίο αφορά την απλή γραμμική παλινδρόμηση. Στο πεδίο "Specification" μας δίνεται η δυνατότητα να μετασχηματίσουμε τόσο την εξαρτημένη μεταβλητή $\boldsymbol{y}$ όσο και την ανεξάρτητη μεταβλητή $\boldsymbol{x}$. Αφήνουμε την εξαρτημένη μεταβλητή $\boldsymbol{y}$ ως έχει, επιλέγοντας **None** στην κατηγορία "**Y transformations:**". Για να μετασχηματίσουμε τη μεταβλητή $\boldsymbol{x}$ σε πολυωνυμική μορφή 2ου βαθμού, στην κατηγορία "**X transformations:**" επιλέγουμε **Polynomial** και δίπλα γράφουμε **2**. Επίσης, για να είναι ανθεκτικά (ή εύρωστα) τα αποτελέσματα της συγκεκριμένης γραμμής παλινδρόμησης, «τσεκάρουμε» το κουτί **Robustness Iterations** και επιλέγουμε τον αριθμό των επαναλήψεων που θέλουμε να πραγματοποιήσει το Eviews (έστω **1000**). Στη συνέχεια, πατάμε το κουμπί "Add" που βρίσκεται κάτω από το πεδίο "Added Elements" και επιλέγουμε **Regression Line** στο παράθυρο "Add Element" που θα εμφανιστεί. Πατώντας **OK**, εμφανίζεται ένα δεύτερο στοιχείο στο πεδίο "Added Elements", το οποίο και πάλι ονομάζεται "Regression Line", αλλά αυτή τη φορά αφορά την παλινδρόμηση όπου η μεταβλητή $\boldsymbol{x}$ έχει πολυωνυμική μορφή 2ου βαθμού. Ακολουθούμε ακριβώς την ίδια διαδικασία, προκειμένου να μετασχηματίσουμε τη μεταβλητή $\boldsymbol{x}$ σε πολυωνυμική μορφή 3ου βαθμού, γράφοντας όμως **3** δίπλα στην επιλογή **Polynomial**. Πλέον, στο πεδίο "Added Elements" υπάρχουν τρία στοιχεία με το όνομα "Regression Line", τα οποία αφορούν τις τρεις παλινδρομήσεις: **γραμμική**, **πολυωνυμική 2ου βαθμού** και **πολυωνυμική 3ου βαθμού**. Πατώντας **OK** και ξανά **OK** εμφανίζεται το διάγραμμα διασποράς **(Εικόνα 7.3)**, το οποίο περιλαμβάνει αυτές τις τρεις γραμμές παλινδρόμησης.

Εικόνα 7.2 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*



Εικόνα 7.3 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Το Eviews (σε αντίθεση με το SPSS) δεν παρέχει τη δυνατότητα ταυτόχρονης εκτίμησης των τριών διαφορετικών υποδειγμάτων. Συνεπώς, το καθένα από αυτά θα πρέπει να εκτιμηθεί ξεχωριστά. Τα αποτελέσματα της εκτίμησης για το γραμμικό υπόδειγμα (eq01 στο Eviews workfile) τα έχουμε ήδη παρουσιάσει στην ενότητα 5.3 του κεφαλαίου 5 **(Εικόνα 5.11)**. Για την εκτίμηση καθενός από τα άλλα δύο υποδείγματα, «ανοίγουμε» την eq01, επιλέγουμε **Proc → Specify/Estimate**, με αποτέλεσμα να εμφανιστεί το παράθυρο "Equation Estimation". Στο tab "Specification", στην επιλογή "Equation specification" γράφουμε:

- Για το **πολυωνυμικό 2ου βαθμού** υπόδειγμα:

  expenditure income income^2 c

  ή expenditure=c(1)+c(2)*income+c(3)*income^2

- Για το **πολυωνυμικό 3ου βαθμού** υπόδειγμα:

expenditure income income^2 income^3 c

ή expenditure=c(1)+c(2)*income+c(3)*income^2+c(4)*income^3

Σε καθεμία από τις περιπτώσεις αυτές, στην επιλογή "Estimation settings" επιλέγουμε **LS – Least Squares (NLS and ARMA)** στο **Method:** και **1 1319** στο **Sample:**. Στον **Πίνακα 7.3** παρουσιάζονται ο συντελεστής προσδιορισμού R^2, ο διορθωμένος συντελεστής προσδιορισμού $\bar{R}^2$, καθώς και το τυπικό σφάλμα της παλινδρόμησης (S.E), για καθένα από τα τρία εναλλακτικά υποδείγματα. Υπενθυμίζουμε και στο σημείο αυτό ότι ο συντελεστής προσδιορισμού είναι το ποσοστό της μεταβλητότητας της εξαρτημένης μεταβλητής που εξηγείται από την παλινδρόμηση.

Πίνακας 7.3 *Συντελεστές προσδιορισμού για τα εναλλακτικά υποδείγματα.*

Υπόδειγμα	R^2	$\bar{R}^2$	S.E.
Γραμμικό	0.079019	0.078320	261.3415
Πολυωνυμικό 2ου βαθμού	0.079630	0.078231	261.3541
Πολυωνυμικό 3ου βαθμού	0.080925	0.078828	261.2694

Όπως προκύπτει από τον συγκεκριμένο πίνακα, οι συντελεστές προσδιορισμού (κανονικός και διορθωμένος) έχουν πολύ χαμηλές τιμές σε όλα τα εναλλακτικά υποδείγματα. Επίσης, οι τιμές των συντελεστών αυτών δεν εμφανίζουν μεγάλες διαφορές μεταξύ των διαφορετικών υποδειγμάτων. Ο λόγος είναι ότι όπως προκύπτει από τις εκτιμήσεις (οι οποίες δεν παρουσιάζονται στο σημείο αυτό για λόγους συντομίας), ο τετραγωνικός όρος δεν είναι στατιστικά σημαντικός στο πολυωνυμικό υπόδειγμα 2ου βαθμού, ενώ τόσο ο τετραγωνικός όσο και ο κυβικός όρος δεν είναι στατιστικά σημαντικοί στο πολυωνυμικό υπόδειγμα 3ου βαθμού. Συνεπώς, επιλέγουμε το γραμμικό υπόδειγμα, παρόλο που και αυτό έχει πολύ χαμηλούς συντελεστές προσδιορισμού. Όμως, όπως αναφέραμε και προηγουμένως, υπάρχουν πιο αποτελεσματικοί τρόποι για την επιλογή του κατάλληλου υποδείγματος, οι οποίοι δεν θα αναλυθούν προς το παρόν.

7.2 Λογιστική παλινδρόμηση για δίτιμη εξαρτημένη μεταβλητή στο SPSS

Έστω, λοιπόν, ότι θέλουμε να διεξάγουμε μια παλινδρόμηση (απλή ή πολλαπλή), στην οποία η εξαρτημένη μεταβλητή παίρνει μόνο δύο τιμές, 0 ή 1 (δηλαδή, ναι ή όχι, επιτυχία ή αποτυχία κλπ.). Στην περίπτωση αυτή, η κλασική παλινδρόμηση που αναλύσαμε στο κεφάλαιο 5 δεν μπορεί να εφαρμοστεί. Ο λόγος είναι ότι η εξαρτημένη μεταβλητή δεν ακολουθεί την κανονική κατανομή, ενώ δεν είναι ούτε συνεχής μεταβλητή. Οπότε, θα εφαρμόσουμε τη λεγόμενη λογιστική παλινδρόμηση. Έστω ότι $\boldsymbol{y}$ είναι η δίτιμη εξαρτημένη μεταβλητή ($\boldsymbol{y = 0}$ ή $\boldsymbol{y = 1}$) και $\boldsymbol{x_1}$ και $\boldsymbol{x_2}$ δύο ανεξάρτητες μεταβλητές. Το υπόδειγμα της λογιστικής παλινδρόμησης γράφεται ως εξής

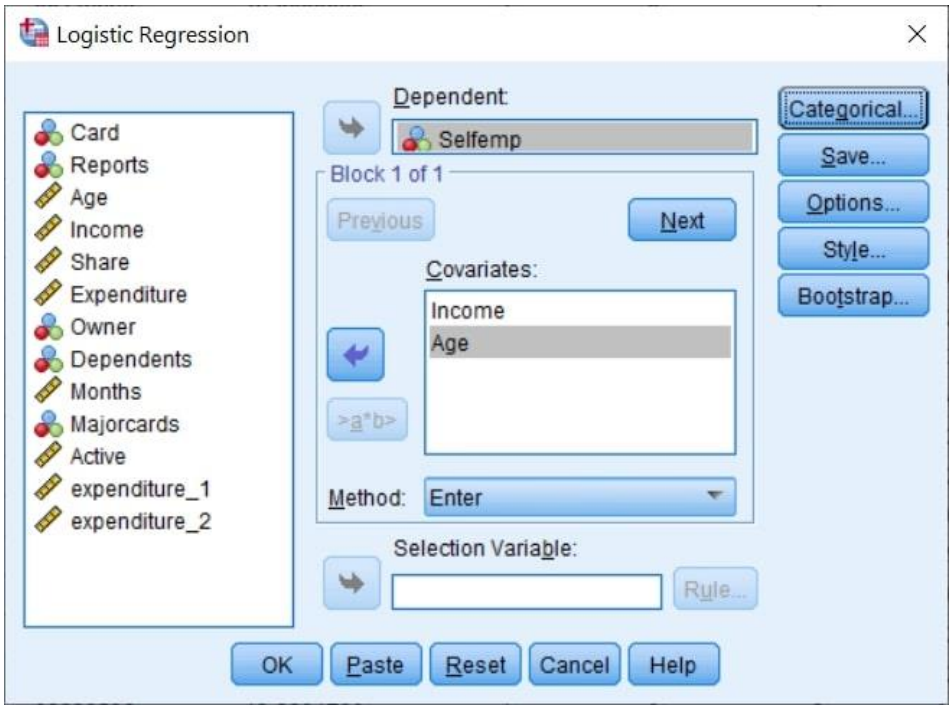
$$\mathbf{\log \frac{y}{1-y} = a + \beta_1 x_1 + \beta_2 x_2} \quad \text{ή} \quad \mathbf{y = \frac{e^{a+\beta_1 x_1+\beta_2 x_2}}{1+e^{a+\beta_1 x_1+\beta_2 x_2}}}$$

Οπότε, όταν χρησιμοποιήσουμε τον λογάριθμο, το υπόδειγμα γίνεται γραμμικό. Διαφορετικά, η εξαρτημένη μεταβλητή $\boldsymbol{y}$ θα σχετίζεται μέσω της εκθετικής συνάρτησης με τις ανεξάρτητες μεταβλητές και άρα όχι γραμμικά. Θα πρέπει να επισημάνουμε στο σημείο αυτό ότι ο λογαριθμικός μετασχηματισμός που έχουμε πραγματοποιήσει ονομάζεται logit (λογιστικός μετασχηματισμός), για αυτό και logistic regression ή λογιστική παλινδρόμηση. Προφανώς, η ονομασία οφείλεται στη λογιστική κατανομή και δεν έχει καμία σχέση με λογιστές ή logistics!

Γενικά, με τη λογιστική παλινδρόμηση εκτιμάμε την πιθανότητα της επιτυχίας, δηλαδή την πιθανότητα η εξαρτημένη μεταβλητή να πάρει την τιμή 1. Επίσης, χρησιμοποιείται ο αυτόματος κανόνας ότι, αν η εκτιμώμενη πιθανότητα είναι μεγαλύτερη ή ίση του 0,5, τότε η εκτιμώμενη τιμή της εξαρτημένης

μεταβλητής είναι 1. Παρακάτω, θα εξετάσουμε πώς μπορούμε αλλάξουμε αυτόν τον κανόνα με τη χρήση της καμπύλης ROC.

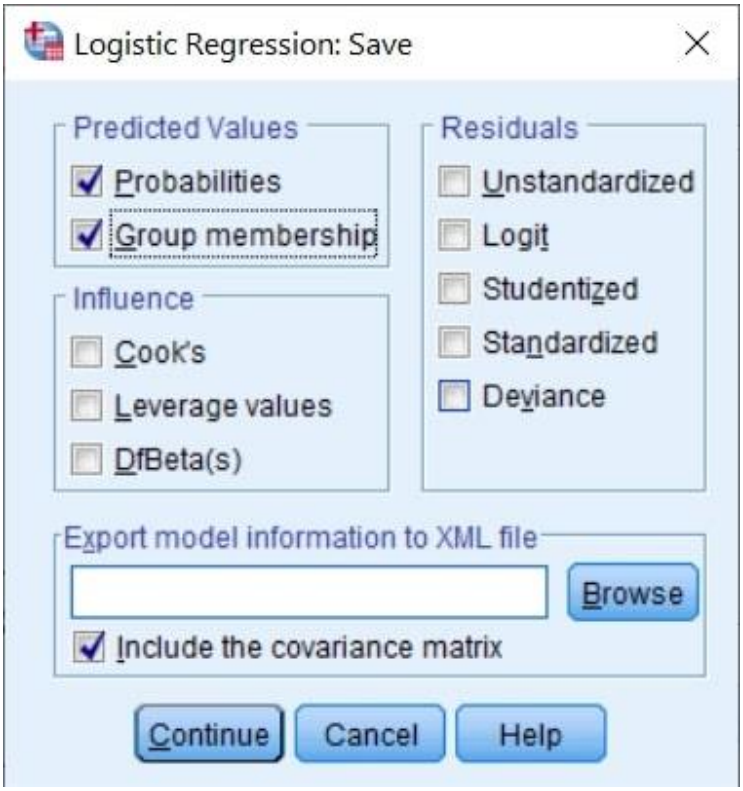
Ας εξετάσουμε, λοιπόν, ένα παράδειγμα λογιστικής παλινδρόμησης χρησιμοποιώντας τα δεδομένα των πιστωτικών καρτών (**credit.sav**). Έστω ότι θέλουμε να ελέγξουμε αν το είδος της απασχόλησης επηρεάζεται από το εισόδημα και την ηλικία, καθώς και αν μπορούμε να προβλέψουμε το είδος της απασχόλησης με βάση τις δύο αυτές μεταβλητές. Επιλέγουμε **Analyze → Regression → Binary Logistic**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 7.4**.



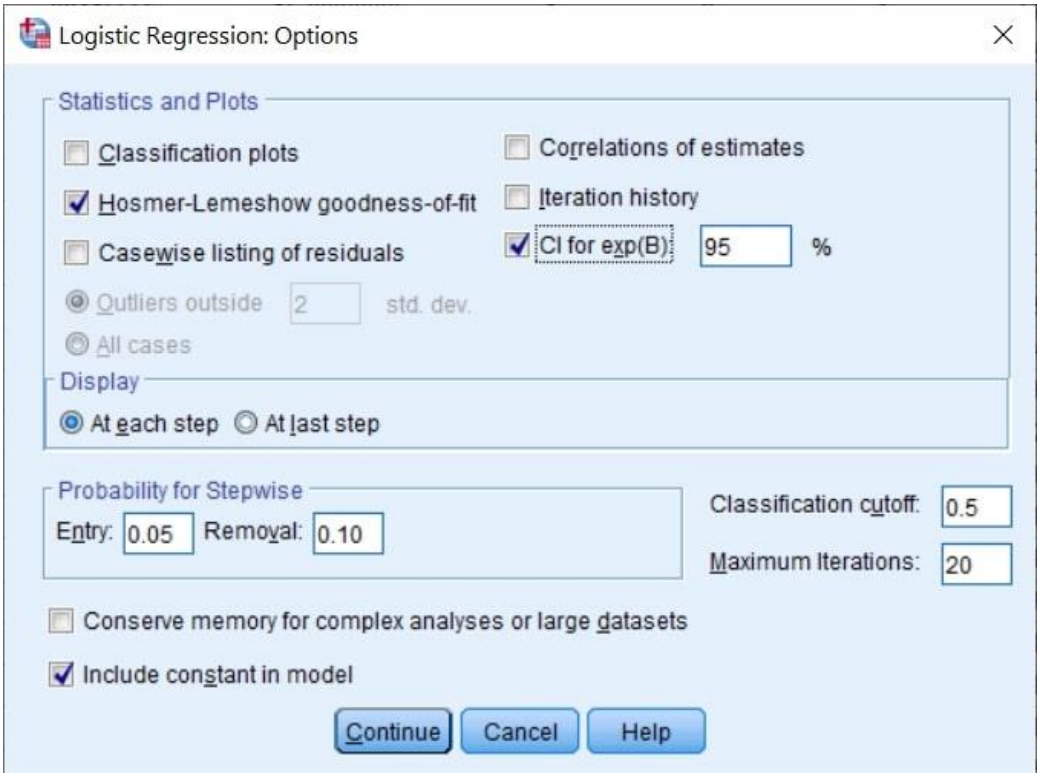
Εικόνα 7.4 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Όπως φαίνεται, έχουμε περάσει την εξαρτημένη δίτιμη μεταβλητή στο λευκό κουτί με την ένδειξη **Dependent:** και τις ανεξάρτητες μεταβλητές στο λευκό κουτί με την ένδειξη **Covariates:**. Πατώντας **Save** στο παράθυρο της **Εικόνας 7.4** μεταφερόμαστε στο παράθυρο της **Εικόνας 7.5**, όπου θα «τσεκάρουμε» τα κουτιά **Probabilities** και **Group membership**. Στη συνέχεια, πατάμε **Continue** και επιστρέφουμε στο παράθυρο της **Εικόνας 7.4**.

Πατώντας **Options** στο παράθυρο της **Εικόνας 7.4** μεταφερόμαστε στο παράθυρο της **Εικόνας 7.6**. Στο παράθυρο αυτό «τσεκάρουμε» τα κουτιά **Hosmer-Lemeshow goodness-of-fit** και **CI for exp(B)**. Αν στο παράθυρο της **Εικόνας 7.4** θέλουμε να επιλέξουμε τον αλγόριθμο bootstrap (κάτι που δεν έχουμε κάνει στην παρούσα ανάλυση), θα πρέπει να έχουμε υπόψη μας ότι δεν θα αποθηκευτούν οι εκτιμώμενες πιθανότητες και οι εκτιμώμενες τιμές της εξαρτημένης μεταβλητής. Όμως, θα λάβουμε τις εκτιμήσεις μεροληψίας των εκτιμημένων παραμέτρων, καθώς και τα 95% διαστήματα εμπιστοσύνης των παραμέτρων αυτών.



Εικόνα 7.5 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Εικόνα 7.6 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Οπότε, πατώντας **OK** στο παράθυρο της **Εικόνας 7.4** θα εμφανιστεί μια σειρά από διάφορους πίνακες στο Output του SPSS. Εμείς θα επικεντρωθούμε στους πίνακες που εμφανίζονται κάτω από την ένδειξη

Block 1: Method = Enter

Αρχικά, στον **Πίνακα 7.4** παρουσιάζεται ένας έλεγχος για τη στατιστική σημαντικότητα όλων των μεταβλητών ταυτόχρονα. Η αντίστοιχη *p*-value είναι μικρότερη του 0,01, γεγονός που υποδηλώνει ότι ένας τουλάχιστον εκτιμημένος συντελεστής είναι διαφορετικός από το μηδέν. Εύκολα μπορεί να αντιληφθεί κανείς ότι ο συγκεκριμένος έλεγχος είναι αντίστοιχος του *F*-test της πολλαπλής γραμμικής παλινδρόμησης.

Πίνακας 7.4 *Έλεγχος στατιστικής σημαντικής σημαντικότητας του μοντέλου συνολικά.*

Omnibus Tests of Model Coefficients				
		Chi-square	df	Sig.
Step 1	Step	19.926	2	.000
	Block	19.926	2	.000
	Model	19.926	2	.000

Ο **Πίνακας 7.5** παρουσιάζει δύο ψευδο-R^2 συντελεστές. Και στην περίπτωση της λογιστικής παλινδρόμησης, οι τιμές τους κυμαίνονται από 0 έως 1, ενώ υψηλές τιμές τους δείχνουν καλή προσαρμογή του υποδείγματος. Στο παράδειγμά μας, οι συγκεκριμένες τιμές είναι εξαιρετικά χαμηλές.

Πίνακας 7.5 *Ψευδο-R^2 συντελεστές.*

Model Summary			
Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	642.272[a]	.015	.038
a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.			

Στον **Πίνακα 7.6** παρουσιάζεται το αποτέλεσμα του ελέγχου καλής προσαρμογής των Hosmer και Lemeshow, όπου η αντίστοιχη *p*-value είναι 0,177. Γενικά, τιμές της *p*-value μεγαλύτερες του 0,05 είναι επιθυμητές στον συγκεκριμένο έλεγχο. Οπότε, ο έλεγχος καλής προσαρμογής δεν απορρίπτεται και άρα το υπόδειγμά μας προσαρμόζεται στα δεδομένα στατιστικά επαρκώς.

Πίνακας 7.6 *Έλεγχος καλής προσαρμογής των Hosmer και Lemeshow.*

Hosmer and Lemeshow Test			
Step	Chi-square	df	Sig.
1	11.461	8	.177

Ο **Πίνακας 7.7** είναι ένας πίνακας επαλήθευσης. Θυμηθείτε ότι στο παράθυρο της **Εικόνας 7.5** επιλέξαμε να αποθηκευτούν οι εκτιμημένες τιμές της εξαρτημένης μεταβλητής. Οπότε, το SPSS δημιούργησε έναν πίνακα διπλής εισόδου, με βάση τις παρατηρούμενες και τις εκτιμημένες τιμές. Παρατηρούμε, λοιπόν, ότι από τους 1319 συμμετέχοντες στο δείγμα μας, όλοι προβλέφθηκαν ως μη αυτοαπασχολούμενοι. Οπότε, με βάση το νούμερο αυτό το συνολικό ποσοστό σωστής κατάταξης είναι 93,1%. Καθώς, όμως, αυτό είναι το πραγματικό ποσοστό των μη αυτοαπασχολούμενων στο δείγμα, το υπόδειγμα δεν φαίνεται να ήταν πολύ αποτελεσματικό στην πρόβλεψη του καθεστώτος απασχόλησης (το οποίο θα εξεταστεί στην ενότητα 7.3 με την καμπύλη ROC).

Πίνακας 7.7 *Σχέση εκτιμημένων και παρατηρούμενων τιμών.*

Classification Table[a]					
			Predicted		
			Selfemp		Percentage
	Observed		No	Yes	Correct
Step 1	Selfemp	**No**	1228	0	100.0
		Yes	91	0	.0
	Overall Percentage				93.1
a. The cut value is .500					

Στον **Πίνακα 7.8** παρουσιάζονται οι εκτιμήσεις των συντελεστών της παλινδρόμησης. Η στήλη **B** αναφέρεται στην επίδραση των μεταβλητών πάνω στο logit της πιθανότητας κάποιος να είναι αυτοαπασχολούμενος, ενώ η στήλη **Exp(B)** στα odds αυτής της πιθανότητας, που ορίζονται ως $\frac{P(y=1)}{1-P(y=1)}$. Αν οι εκτιμημένοι συντελεστές είναι θετικοί και άρα το εκθετικό τους είναι μεγαλύτερο της μονάδας, αυτό σημαίνει ότι και οι δύο μεταβλητές επιδρούν θετικά στην πιθανότητα κάποιος να είναι ελεύθερος επαγγελματίας. Θα πρέπει να υπενθυμίσουμε στο σημείο αυτό ότι το υπόδειγμά μας είναι ένα γενικευμένο γραμμικό υπόδειγμα, υπό την έννοια ότι, ενώ η εξαρτημένη μεταβλητή δεν συνδέεται γραμμικά με τις παραμέτρους των ανεξάρτητων μεταβλητών, η γραμμική αυτή σχέση προκύπτει μόλις χρησιμοποιηθεί ο λογιστικός μετασχηματισμός. Όμως, η γραμμική αυτή σχέση είναι μεταξύ της *μετασχηματισμένης* εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών.

Πίνακας 7.8 *Εκτιμημένες παράμετροι του μοντέλου της λογιστικής παλινδρόμησης.*

Variables in the Equation									
								95% C.I.for EXP(B)	
		B	S.E.	Wald	Df	Sig.	Exp(B)	Lower	Upper
Step 1[a]	Income	**.156**	.054	8.463	1	**.004**	**1.169**	1.052	1.298
	Age	**.027**	.010	6.605	1	**.010**	**1.027**	1.006	1.048
	Constant	-4.101	.390	110.467	1	.000	.017		
a. Variable(s) entered on step 1: Income, Age.									

7.3 Καμπύλη ROC στο SPSS

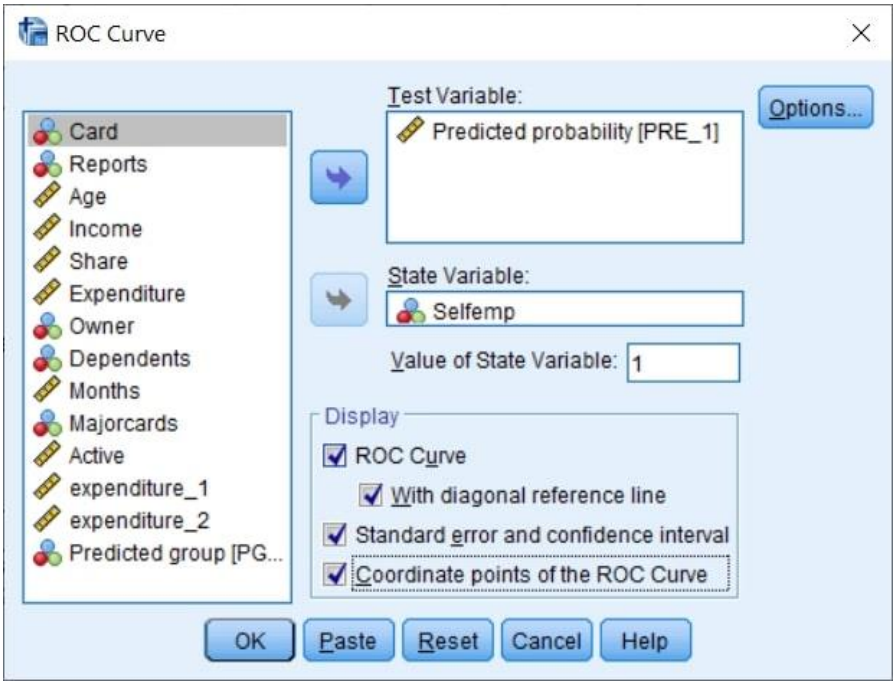
Στο παράθυρο της **Εικόνας 7.7** θα περάσουμε στο πάνω κουτί με την ένδειξη **Test Variable:** τη στήλη με τις εκτιμημένες πιθανότητες από το υπόδειγμά μας, ενώ στο κάτω κουτί με την ένδειξη **State Variable:** θα περάσουμε την παρατηρούμενη τιμή της εξαρτημένης μεταβλητής. Επίσης, θα ορίσουμε ότι η τιμή 1 είναι η τιμή ενδιαφέροντος και θα «τσεκάρουμε» όλες τις διαθέσιμες επιλογές. Στη συνέχεια, πατώντας **Continue** θα εμφανιστούν κάποιοι πίνακες (**Πίνακες 7.9** και **7.10**), καθώς και ένα διάγραμμα **(Διάγραμμα 7.2)**.

Όμως, πριν προχωρήσουμε, θα πρέπει να αναφέρουμε κάποια πράγματα για την καμπύλη ROC (*Receiver Operating Curve* - Λειτουργική Χαρακτηριστική Καμπύλη) και την AUC (*Area Under the Curve* - Επιφάνεια Κάτω από την Καμπύλη). Η ευαισθησία (*sensitivity*) και η ειδικότητα (*specificity*) ορίζονται ως εξής:

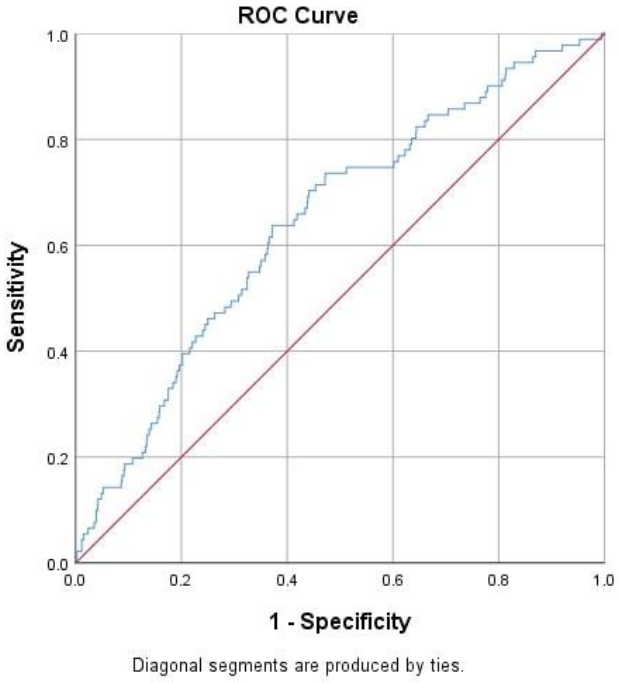
$$\text{Ευαισθησία } = \frac{\text{TP}}{\text{TP+FN}} \text{ και Ειδικότητα} = \frac{\text{TN}}{\text{TN+FP}}$$

Τα παραπάνω ακρωνύμια ερμηνεύονται ως εξής: TP (true positive) = αληθής πρόβλεψη επιτυχίας, FN (false negative) = ψευδής πρόβλεψη αποτυχίας, FP (false positive) = ψευδής πρόβλεψη επιτυχίας και TN (true negative) = αληθής πρόβλεψη αποτυχίας. Στο παράδειγμά μας, η επιτυχία είναι όταν το άτομο είναι ελεύθερος επαγγελματίας. Οπότε, η ευαισθησία είναι η πιθανότητα θετικής ανίχνευσης της επιτυχίας, όταν αυτή υπάρχει. Η ειδικότητα είναι η πιθανότητα σωστής ανίχνευσης της αποτυχίας, όταν αυτή υπάρχει. Αυτό που επιθυμούμε είναι οι δύο αυτές τιμές να είναι ταυτόχρονα υψηλές ή η τιμή της ευαισθησίας να είναι

υψηλή και η τιμή *1 - ειδικότητα* να είναι χαμηλή. Οπότε, στην καμπύλη ROC **(Διάγραμμα 7.2)** επιθυμούμε να βρισκόμαστε ψηλά στον κατακόρυφο άξονα και αριστερά στον οριζόντιο.



Εικόνα 7.7 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Διάγραμμα 7.2 *Καμπύλη ROC.*

Πίνακας 7.9 *Εκτιμημένη τιμή του AUC.*

Area Under the Curve				
Test Result Variable(s): Predicted probability				
			Asymptotic 95% Confidence Interval	
Area	Std. Error[a]	Asymptotic Sig.[b]	Lower Bound	Upper Bound
.644	.029	.000	.587	.702
The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.				
a. Under the nonparametric assumption				
b. Null hypothesis: true area = 0.5				

Πίνακας 7.10 *Κάποιες ενδεικτικές τιμές της καμπύλης ROC.*

Coordinates of the Curve		
Test Result Variable(s): Predicted probability		
Positive if Greater Than or Equal To[a]	Sensitivity	1 - Specificity
.0000000	1.000	1.000
.0227355	1.000	.999
.0252589	1.000	.998
.0262882	1.000	.998
.0265714	1.000	.997
.0279346	1.000	.996
.0297365	1.000	.995
.0319232	1.000	.994
.0337406	.989	.994
.0341873	.989	.993
.0344357	.989	.993
.0347071	.989	.992
.0349613	.989	.991
.0352428	.989	.990
.0354350	.989	.989
The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group.		
a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.		

7.4 Λογιστική παλινδρόμηση για δίτιμη εξαρτημένη μεταβλητή στο Eviews

Όπως αναλύθηκε παραπάνω, η γενική μορφή ενός υποδείγματος λογιστικής παλινδρόμησης (logistic regression – **logit**) με δύο ανεξάρτητες μεταβλητές είναι:

$$\ln\frac{\mathbf{P(y=1)}}{\mathbf{1-P(y=1)}} = \boldsymbol{\alpha} + \boldsymbol{\beta_1 x_1} + \boldsymbol{\beta_2 x_2} \quad \text{ή} \quad \mathbf{P(y=1)} = \frac{\mathbf{e^{\alpha+\beta_1 x_1+\beta_2 x_2}}}{\mathbf{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2}}},$$

όπου η εξαρτημένη μεταβλητή $\boldsymbol{y}$ παίρνει τις τιμές 0 και 1. Το υπόδειγμα είναι ένα γενικευμένο γραμμικό, καθώς η πρώτη σχέση, η οποία προκύπτει από τον *μετασχηματισμό της εξαρτημένης μεταβλητής,* είναι μια ευθεία γραμμή, ενώ η δεύτερη δεν είναι. Επίσης, η εκτίμηση ενός τέτοιου υποδείγματος γίνεται με τη μέθοδο της **μεγίστης πιθανοφάνειας** (**maximum likelihood**).

Ας δούμε, λοιπόν, ένα διαφορετικό παράδειγμα σε σχέση με το SPSS, εξετάζοντας αν το να είναι ένα άτομο ιδιοκτήτης της τρέχουσας κατοικίας του (μεταβλητή "owner" που παίρνει την τιμή 1 αν το άτομο είναι ιδιοκτήτης, και 0 αν δεν είναι) επηρεάζεται από το εισόδημά του (μεταβλητή "income"), καθώς και από την ηλικία του (μεταβλητή "age"). Μπορούμε, εναλλακτικά:

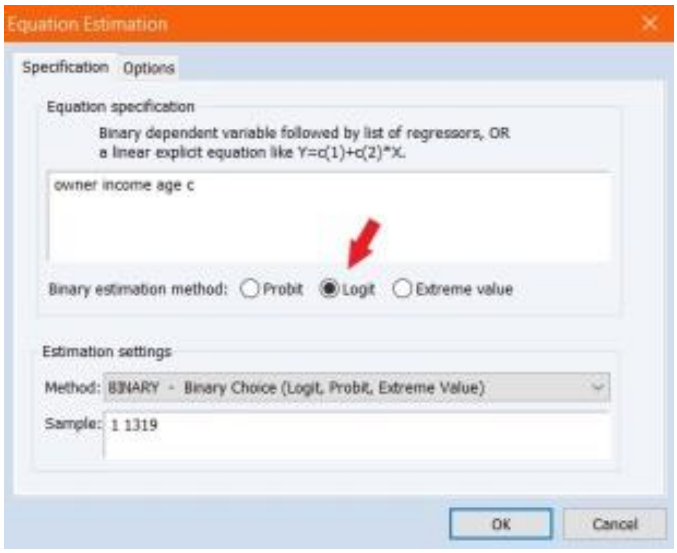
- Να «μαρκάρουμε» τις 3 αυτές μεταβλητές στο Eviews workfile και με δεξί κλικ να επιλέξουμε **Open → as Equation**.
- Να ανοίξουμε τις 3 μεταβλητές ως Group και στη συνέχεια να επιλέξουμε **Proc → Make Equation**.

Και στις δύο περιπτώσεις, θα εμφανιστεί το παράθυρο "Equation Estimation", όπου στο "Estimation settings" στην επιλογή **Method:** θα επιλέξουμε **BINARY – Binary Choice (Logit, Probit, Extreme Value)**.

Στο τροποποιημένο πλέον παράθυρο "Equation Estimation" **(Εικόνα 7.8)**, στο tab "Specification", στην επιλογή "Equation specification" εμφανίζονται με τη σειρά οι μεταβλητές μας, με πρώτη την εξαρτημένη (εναλλακτικά, μπορούμε να γράψουμε την εξίσωση με την κανονική της μορφή, δηλαδή **owner=c(1)+c(2)*income+c(3)*age**). Ακριβώς από κάτω εμφανίζεται η επιλογή "**Binary estimation method:**", όπου επιλέγουμε **Logit** (όπως υποδεικνύεται από το κόκκινο βέλος). Επίσης, διατηρούμε το **1 1319** στην επιλογή **Sample:** και τις προεπιλογές του Eviews στο tab "Options" (με τις οποίες δεν θα ασχοληθούμε προς το παρόν). Πατώντας **OK** εμφανίζονται τα αποτελέσματα της εκτιμημένης παλινδρόμησης (με τη μέθοδο της μεγίστης πιθανοφάνειας), καθώς και κάποια βασικά διαγνωστικά μέτρα **(Εικόνα 7.9)**. Τη συγκεκριμένη εκτίμηση την αποθηκεύσαμε στο Eviews workfile με το όνομα eq05. Εναλλακτικά, ένας γρηγορότερος τρόπος εκτίμησης της συγκεκριμένης παλινδρόμησης είναι να γράψουμε στο **Command line**:

equation eq05.logit owner income age c

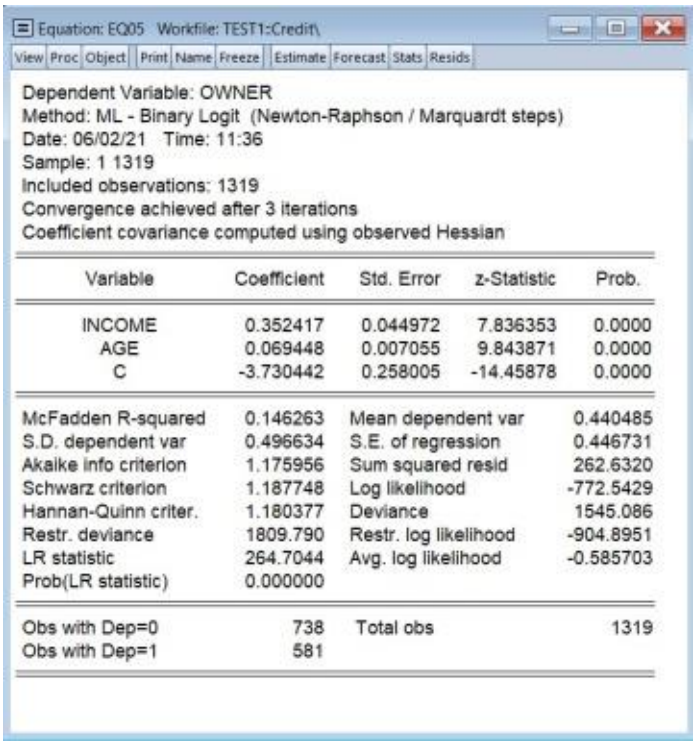
και να πατήσουμε **Enter**. Θα πάρουμε και πάλι τα αποτελέσματα της **Εικόνας 7.9**, ενώ η εξίσωσή μας θα αποθηκευτεί αυτόματα στο Eviews workfile ως eq05.



Εικόνα 7.8 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Στο πάνω μέρος των αποτελεσμάτων αναφέρονται η εξαρτημένη μεταβλητή, η μέθοδος εκτίμησης και το δείγμα που χρησιμοποιήθηκε για την εκτίμηση. Στη συνέχεια, παρουσιάζονται οι εκτιμημένοι συντελεστές για τις μεταβλητές "income" και "age", καθώς και τον σταθερό όρο, τα τυπικά τους σφάλματα, οι *z*-στατιστικές τους (καθώς το logit υπόδειγμα εκτιμήθηκε με τη μέθοδο της μεγίστης πιθανοφάνειας και όχι με αυτή των ελαχίστων τετραγώνων), καθώς και οι αντίστοιχες *p*-values. Όπως και στο SPSS, οι

συγκεκριμένοι συντελεστές αναφέρονται στον λογιστικό μετασχηματισμό της εξαρτημένης μεταβλητής "owner" και όχι στην ίδια την εξαρτημένη μεταβλητή. Από τις *z*-στατιστικές του μετασχηματισμένου υποδείγματος, προκύπτει ότι οι εκτιμημένοι συντελεστές των μεταβλητών "income" και "age" είναι στατιστικά σημαντικοί σε επίπεδο σημαντικότητας $a = 5\%$, καθώς οι αντίστοιχες *p*-values είναι μικρότερες του 0,05. Επίσης, έχουν θετική επίδραση στον λογιστικό μετασχηματισμό και της μεταβλητής "owner", καθώς και στην ίδια τη μεταβλητή γενικότερα. Με άλλα λόγια, όσο αυξάνονται το ετήσιο εισόδημα και η ηλικία ενός ατόμου τόσο αυξάνονται και οι πιθανότητες να είναι το άτομο αυτό ιδιοκτήτης της τρέχουσας κατοικίας του.



Equation: EQ05 Workfile: TEST1:Credit\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: OWNER
Method: ML - Binary Logit (Newton-Raphson / Marquardt steps)
Date: 06/02/21 Time: 11:36
Sample: 1 1319
Included observations: 1319
Convergence achieved after 3 iterations
Coefficient covariance computed using observed Hessian

Variable	Coefficient	Std. Error	z-Statistic	Prob.
INCOME	0.352417	0.044972	7.836353	0.0000
AGE	0.069448	0.007055	9.843871	0.0000
C	-3.730442	0.258005	-14.45878	0.0000

McFadden R-squared	0.146263	Mean dependent var	0.440485
S.D. dependent var	0.496634	S.E. of regression	0.446731
Akaike info criterion	1.175956	Sum squared resid	262.6320
Schwarz criterion	1.187748	Log likelihood	-772.5429
Hannan-Quinn criter.	1.180377	Deviance	1545.086
Restr. deviance	1809.790	Restr. log likelihood	-904.8951
LR statistic	264.7044	Avg. log likelihood	-0.585703
Prob(LR statistic)	0.000000		

Obs with Dep=0	738	Total obs	1319
Obs with Dep=1	581		

Εικόνα 7.9 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Οπότε, το εκτιμημένο logit υπόδειγμα θα είναι:

$$\ln\frac{\widehat{\mathbf{P(Owner=1)}}}{\mathbf{1-P(Owner=1)}} = \mathbf{-3.730442 + (0.352417 \times Income) + (0.069448 \times Age)},$$

ή

$$\widehat{\mathbf{P(Owner = 1)}} = \frac{\mathbf{e^{-3.730442+(0.352417\times Income)+(0.069448\times Age)}}}{\mathbf{1 + e^{-3.730442+(0.352417\times Income)+(0.069448\times Age)}}}.$$

Οι συντελεστές που αναφέρονται στην εξαρτημένη μεταβλητή απευθείας δεν υπολογίζονται από το Eviews, αλλά μπορούν να υπολογιστούν πολύ εύκολα, υπολογίζοντας το εκθετικό για κάθε εκτιμημένο συντελεστή ($e^{\widehat{\beta}_i}$). Οπότε, ο συντελεστής του ετήσιου εισοδήματος θα είναι $e^{\widehat{\beta}_1} = \mathbf{1,422502}$, της ηλικίας $e^{\widehat{\beta}_2} = \mathbf{1,071916,}$ ενώ ο σταθερός όρος θα είναι $e^{\widehat{\alpha}} = \mathbf{0,023982}$. Οι συγκεκριμένοι συντελεστές εκφράζουν τα **odds ratios** $\left(\frac{P(Y=1)}{P(Y=0)}\right)$. Δηλαδή, για την περίπτωση του εισοδήματος, αν το άτομο Α έχει ετήσιο εισόδημα 9 (δηλαδή, \$90.000) και το άτομο Β ετήσιο εισόδημα 10 (δηλαδή, \$100.000), τότε τα odds του ατόμου Β να είναι ιδιοκτήτης της τρέχουσας κατοικίας του είναι **1,422502** φορές τα odds του ατόμου Α να είναι ιδιοκτήτης της τρέχουσας κατοικίας του. Αντίστοιχα, για την ηλικία, αν το άτομο Α είναι 30 ετών και το άτομο Β 31 ετών, τότε τα odds του ατόμου Β να είναι ιδιοκτήτης της τρέχουσας κατοικίας του είναι **1,071916** φορές τα odds του ατόμου Α να είναι ιδιοκτήτης της τρέχουσας κατοικίας του.

Στην **Εικόνα 7.9**, κάτω από τις εκτιμήσεις των συντελεστών παρουσιάζονται τα ακόλουθα διαγνωστικά μέτρα της εκτίμησης του logit υποδείγματος:

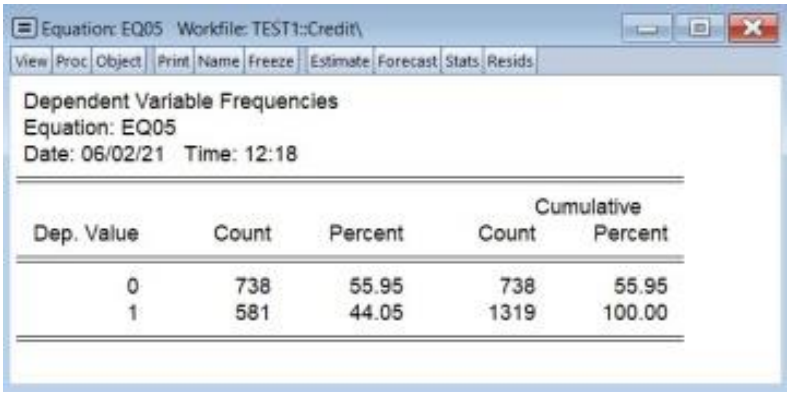
- **McFadden R-squared**: Αποτελεί έναν δείκτη μεγίστης πιθανοφάνειας, ο οποίος παίρνει τιμές μεταξύ 0 και 1 και είναι ανάλογος του συντελεστής προσδιορισμού R^2 της γραμμικής παλινδρόμησης. Ουσιαστικά, ο δείκτης αυτός είναι ένα *pseudo*-R^2.
- **Mean dependent var**: Είναι ο μέσος της εξαρτημένης μεταβλητής.
- **S.D dependent var**: Είναι η τυπική απόκλιση (standard deviation) της εξαρτημένης μεταβλητής.
- **S.E of regression**: Είναι το τυπικό σφάλμα (standard error) της παλινδρόμησης.
- **Akaike info criterion, Schwarz criterion και Hannan-Quinn criter**: Όπως και στην περίπτωση της γραμμικής παλινδρόμησης, τα τρία αυτά κριτήρια αφορούν την επιλογή του κατάλληλου υποδείγματος. Επιλέγουμε το υπόδειγμα εκείνο στο οποίο οι τιμές των κριτηρίων αυτών παίρνουν τη χαμηλότερη τιμή.
- **Sum squared resid**: Είναι το άθροισμα τετραγώνων των καταλοίπων.
- **Log likelihood**: Είναι η μεγιστοποιημένη τιμή της συνάρτησης log likelihood.
- **Restr. log likelihood**: Είναι η μεγιστοποιημένη τιμή της συνάρτησης log likelihood, όταν όλοι οι συντελεστές κλίσης έχουν τεθεί ίσοι με το μηδέν.
- **Avg. log likelihood**: Είναι η τιμή της συνάρτησης log likelihood διαιρεμένη με τον αριθμό των παρατηρήσεων του δείγματος.
- **Deviance**: Αποτελεί ένα μέτρο που υπολογίζει την απόκλιση του εκτιμημένου logit υποδείγματος από ένα τέλειο υπόδειγμα ("saturated model"), ενώ το **Restr. deviance** είναι η τιμή του ίδιου μέτρου, όταν όλοι οι συντελεστές κλίσης έχουν τεθεί ίσοι με το μηδέν.
- **LR statistic**: Είναι η τιμή της Likelihood ratio στατιστικής για τη μηδενική υπόθεση ότι όλοι οι συντελεστές κλίσης (εκτός του σταθερού όρου) είναι στατιστικά ίσοι με μηδέν. Η εναλλακτική υπόθεση στην περίπτωση αυτή είναι ότι έστω ένας από τους συντελεστές αυτούς δεν είναι στατιστικά μηδέν. Η **Prob(LR statistic)** είναι η αντίστοιχη *p*-value για τον έλεγχο αυτό. Καθώς και η συγκεκριμένη *p*-value είναι μικρότερη από το 0,05, αυτό σημαίνει ότι η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας $a = 5\%$. Αυτό είναι αναμενόμενο, καθώς από τις *z*-στατιστικές προκύπτει ότι οι εκτιμημένοι συντελεστές των μεταβλητών "income" και "age" είναι στατιστικά σημαντικοί για επίπεδο σημαντικότητας $a = 5\%$.

Τέλος, στο κάτω μέρος της **Εικόνας 7.9** εμφανίζονται ο αριθμός των παρατηρήσεων, όπου η εξαρτημένη μεταβλητή "owner" παίρνει την τιμή 0 (**Obs with Dep=0**), ο αριθμός των παρατηρήσεων, όπου παίρνει την τιμή 1 (**Obs with Dep=1**), καθώς και το σύνολό τους (**Total obs**).

Όπως και στην περίπτωση της γραμμικής παλινδρόμησης, έχοντας ανοικτή την eq05 και επιλέγοντας **View** (είτε από το μενού του Eviews είτε από το toolbar της eq05), μπορούμε να υπολογίσουμε τα αντίστοιχα διαγνωστικά μέτρα για το εκτιμημένο logit υπόδειγμα και να κάνουμε έλεγχο κανονικότητας των καταλοίπων. Επίσης, στο logit υπόδειγμα μπορούμε να δημιουργήσουμε τα κανονικά (ordinary), τα τυποποιημένα (standardized) ή τα γενικευμένα (generalized) κατάλοιπα, επιλέγοντας **Proc → Make Residual Series** (είτε από το μενού του Eviews είτε από το toolbar της eq05) και κάνοντας την αντίστοιχη επιλογή στο παράθυρο "Make Residuals" που εμφανίζεται **(Εικόνα 5.12)**.

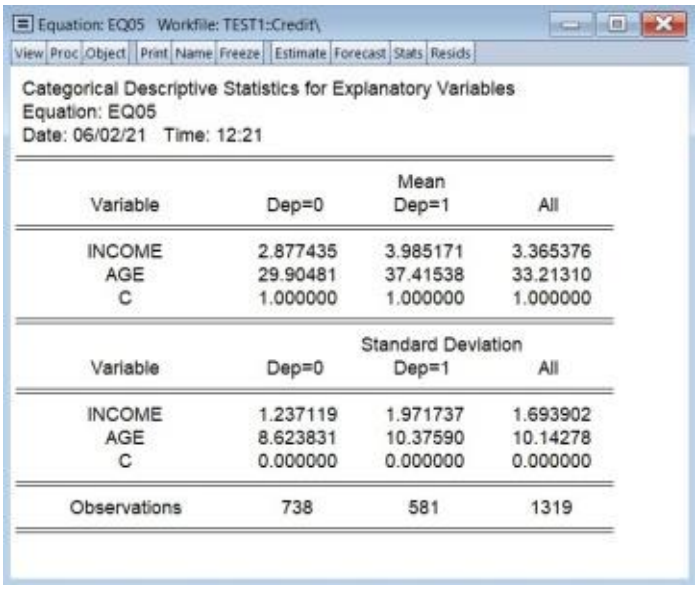
Στο logit υπόδειγμα, το Eviews περιλαμβάνει μια σειρά από πολύ βασικά διαγνωστικά μέτρα και στατιστικούς ελέγχους. Έχοντας ανοικτή την eq05, επιλέγουμε **View** (είτε από το μενού του Eviews είτε από το toolbar της eq05) και στη συνέχεια:

- **Dependent Variable Frequencies**: Εμφανίζεται ένας πίνακας που παρουσιάζει τη συχνότητα και τη σωρευτική (**Cumulative**) συχνότητα όπου η εξαρτημένη μεταβλητή "owner" παίρνει τις τιμές 0 και 1 **(Εικόνα 7.10)**.



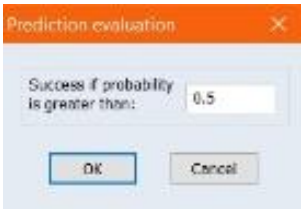
Εικόνα 7.10 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

- **Categorical Regressor Stats**: Εμφανίζεται ένας πίνακας, ο οποίος παρουσιάζει τα βασικά περιγραφικά μέτρα (μέσο και τυπική απόκλιση) για καθεμία από τις ανεξάρτητες μεταβλητές **(Εικόνα 7.11)**. Τα μέτρα αυτά έχουν υπολογιστεί για το σύνολο του δείγματος, καθώς και για τα μέρη του δείγματος όπου η εξαρτημένη μεταβλητή "owner" παίρνει την τιμή 0 (**Dep=0**) και την τιμή 1 (**Dep=1**).



Εικόνα 7.11 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

- **Expectation-Prediction Evaluation**: Η συγκεκριμένη επιλογή παρουσιάζει το πόσο «σωστά» έχει εκτιμήσει το logit υπόδειγμα την πιθανότητα η εξαρτημένη μεταβλητή να πάρει την τιμή 1. Ο κανόνας που συνήθως χρησιμοποιείται, όπως φαίνεται και από το παράθυρο που εμφανίζεται στην **Εικόνα 7.12**, είναι ότι, αν η πιθανότητα είναι πάνω από 0,5, τότε η εκτιμημένη τιμή της εξαρτημένης μεταβλητής θα είναι 1. Φυσικά, μπορούμε να αλλάξουμε τον κανόνα αυτό και να θέσουμε την πιθανότητα που επιθυμούμε. Πατώντας **OK**, εμφανίζονται τα αποτελέσματα **(Εικόνα 7.13)**, στα οποία υπάρχουν δύο πίνακες. Τα αποτελέσματα που μας ενδιαφέρουν παρουσιάζονται στον πίνακα που βρίσκεται στο πάνω μέρος (ο πίνακας που βρίσκεται στο κάτω μέρος παρουσιάζει τα αντίστοιχα αποτελέσματα χρησιμοποιώντας αναμενόμενες τιμές αντί για πιθανότητες).



Εικόνα 7.12 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Equation: EQ05 Workfile: TEST1::Credit\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Expectation-Prediction Evaluation for Binary Specification
Equation: EQ05
Date: 06/03/21 Time: 14:11
Success cutoff: C = 0.5

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
P(Dep=1)<=C	595	253	848	738	581	1319
P(Dep=1)>C	143	328	471	0	0	0
Total	738	581	1319	738	581	1319
Correct	595	328	923	738	0	738
% Correct	80.62	56.45	69.98	100.00	0.00	55.95
% Incorrect	19.38	43.55	30.02	0.00	100.00	44.05
Total Gain*	-19.38	56.45	14.03			
PercentGain**	NA	56.45	31.84			

	Estimated Equation			Constant Probability		
	Dep=0	Dep=1	Total	Dep=0	Dep=1	Total
E(# of Dep=0)	474.56	263.44	738.00	412.92	325.08	738.00
E(# of Dep=1)	263.44	317.56	581.00	325.08	255.92	581.00
Total	738.00	581.00	1319.00	738.00	581.00	1319.00
Correct	474.56	317.56	792.12	412.92	255.92	668.84
% Correct	64.30	54.66	60.05	55.95	44.05	50.71
% Incorrect	35.70	45.34	39.95	44.05	55.95	49.29
Total Gain*	8.35	10.61	9.35			
PercentGain**	18.96	18.96	18.96			

*Change in "% Correct" from default (constant probability) specification
**Percent of incorrect (default) prediction corrected by equation

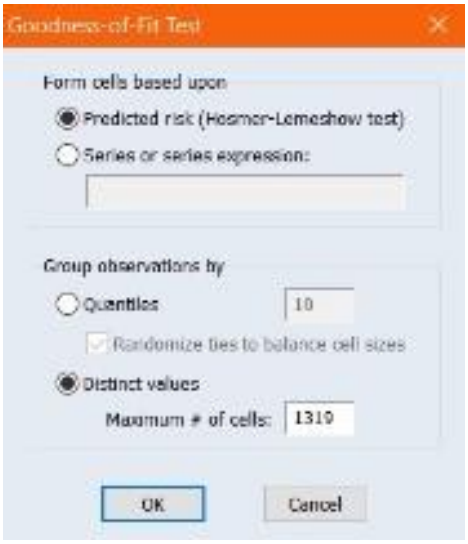
Εικόνα 7.13 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

Στο αριστερό τμήμα του πίνακα που βρίσκεται στο πάνω μέρος, ταξινομούνται οι παρατηρήσεις που έχουν predicted probabilities μικρότερες ή μεγαλύτερες από το cutoff value **'C'** (το οποίο έχουμε θέσει στο 0,5). Για $\boldsymbol{P \leq C = 0,05}$, όταν η εξαρτημένη μεταβλητή "owner" παίρνει την τιμή 0 (**Dep=0**) οι παρατηρήσεις είναι 595, και όταν παίρνει την τιμή 1 (**Dep=1**) είναι 253. Αντίστοιχα για $\boldsymbol{P > C = 0,05}$, οι παρατηρήσεις είναι 143 όταν **Dep=0**, και 328 όταν **Dep=1**. Η «σωστή» **("Correct")** ταξινόμηση προκύπτει όταν η predicted probability είναι μικρότερη ή ίση του 0,5 και η παρατηρούμενη παίρνει την τιμή 0 (Dep=0), ή όταν η predicted probability είναι μεγαλύτερη του 0,5 και η παρατηρούμενη παίρνει την τιμή 1 (Dep=1). Οπότε, 595 παρατηρήσεις για Dep=0 και 328 παρατηρήσεις για Dep=1 έχουν ταξινομηθεί «σωστά» από το εκτιμημένο υπόδειγμα. Στη συνέχεια, παρουσιάζεται το ποσοστό των παρατηρήσεων που έχει ταξινομηθεί «σωστά»: 80,62% για Dep=0, 56,45% για Dep=1 και 69,98% συνολικά, και το ποσοστό των παρατηρήσεων που δεν έχει ταξινομηθεί «σωστά»: 19,38% για Dep=0, 43,55% για Dep=1 και 30,02% συνολικά. Όπως έχει ήδη αναφερθεί, το ποσοστό των παρατηρήσεων Dep=1 που έχει προβλεφθεί «σωστά» ονομάζεται *«ευαισθησία» (sensitivity)*, ενώ το ποσοστό των παρατηρήσεων Dep=0 που έχει προβλεφθεί «σωστά» ονομάζεται *ειδικότητα (specificity).* Όπως προκύπτει από τα αποτελέσματα, το sensitivity στο υπόδειγμά μας είναι 56,45% και το specificity είναι 80,62%.

Στο δεξί τμήμα του πίνακα που βρίσκεται στο πάνω μέρος, παρουσιάζεται μια αντίστοιχη ταξινόμηση των παρατηρήσεων, όταν το υπόδειγμα έχει εκτιμηθεί με μόνο τον σταθερό όρο ως ανεξάρτητη μεταβλητή και έχει σταθερή πιθανότητα (υπόδειγμα "constant probability"). Οι δύο τελευταίες ενδείξεις του πίνακα που βρίσκεται στο πάνω μέρος **("Total Gain" και "Percent Gain")** αφορούν την προβλεπτική ικανότητα του

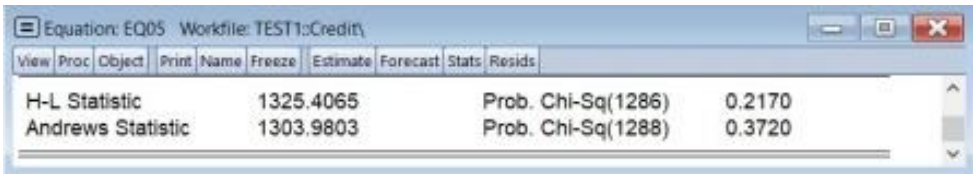
υποδείγματος. Το "Total Gain" δείχνει πόσο βελτιώνονται οι προβλέψεις από το εκτιμημένο logit υπόδειγμα σε σχέση με το υπόδειγμα "constant probability": -19,38% για Dep=0, 56,45% για Dep=1 και 14,03% συνολικά. Το "Percent Gain" υπολογίζει τη συγκεκριμένη βελτίωση ως ποσοστό της «μη-σωστής» ταξινόμησης του υποδείγματος "constant probability": $31{,}84\%\ (= 14{,}03\%/44{,}05\%)$. Τέλος, θα πρέπει να επισημάνουμε στο σημείο αυτό πως το Eviews δεν παρέχει τη δυνατότητα εξαγωγής των καμπυλών ROC και AUC.

- **Goodness-of-Fit Test (Hosmer-Lemeshow)**: Η συγκεκριμένη επιλογή μας επιτρέπει να ελέγξουμε την «καλή προσαρμογή» (Goodness-of-Fit) του υποδείγματος με τη χρήση των στατιστικών ελέγχου του **Andrews** και των **Hosmer-Lemeshow**. Θα πρέπει να επισημάνουμε στο σημείο αυτό πως και οι δύο αυτές στατιστικές κατανέμονται ασυμπτωτικά ως χ^2. Ο έλεγχος των Hosmer-Lemeshow ομαδοποιεί τις παρατηρήσεις υποθέτοντας ότι η predicted probability είναι Dep=1, ενώ ο έλεγχος του Andrews είναι πιο γενικός και ομαδοποιεί τις παρατηρήσεις υποθέτοντας οποιαδήποτε μορφή των μεταβλητών. Στο παράθυρο που εμφανίζεται **(Εικόνα 7.14)**, επιλέγουμε **Predicted risk (Hosmer-Lemeshow test**), καθώς και τον τρόπο ομαδοποίησης των παρατηρήσεων της εξαρτημένης μεταβλητής. Όταν η μεταβλητή που θέλουμε να ομαδοποιήσουμε παίρνει πολλές τιμές, επιλέγουμε σε πόσα **Quantiles** θέλουμε να ομαδοποιηθεί και «τσεκάρουμε» το κουτί "Randomize ties to balance cell sizes". Αντίθετα, όταν η μεταβλητή που θέλουμε να ομαδοποιήσουμε παίρνει σχετικά λίγες τιμές, όπως στο παράδειγμά μας, όπου η εξαρτημένη μεταβλητή "owner" παίρνει τις τιμές 0 και 1, επιλέγουμε **Distinct values**. Επίσης, στο πεδίο **Maximum # of cells:** γράφουμε 1319 που είναι ο αριθμός των παρατηρήσεών μας. Πατώντας OK εμφανίζονται τα αποτελέσματα, όπου στην τελευταία στήλη παρουσιάζεται η συνεισφορά κάθε παρατήρησης στη συνολική στατιστική ελέγχου των Hosmer-Lemeshow (H-L). Στο κάτω μέρος των αποτελεσμάτων εμφανίζονται οι τιμές των στατιστικών ελέγχου Hosmer-Lemeshow και Andrews, μαζί με τις αντίστοιχες *p*-values **(Εικόνα 7.15)**.[5] Όπως φαίνεται από τα αποτελέσματα αυτά, και οι δύο p-values είναι μεγαλύτερες του 0,05, γεγονός που σημαίνει ότι η μηδενική υπόθεση της «καλής προσαρμογής» του υποδείγματος δεν μπορεί να απορριφθεί σε επίπεδο στατιστικής σημαντικότητας $a = 5\%$. Οπότε, μπορούμε να συμπεράνουμε πως το εκτιμημένο logit υπόδειγμα προσαρμόζει στατιστικά επαρκώς τις παρατηρήσεις των μεταβλητών "owner", "income" και "age".



Εικόνα 7.14 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

[5] Καθώς ο συγκεκριμένος πίνακας είναι πάρα πολύ μεγάλος σε έκταση και καλύπτει αρκετές σελίδες, εμείς παρουσιάζουμε μόνο το κάτω μέρος του, το οποίο άλλωστε είναι και αυτό που έχει σημασία στην ανάλυσή μας.



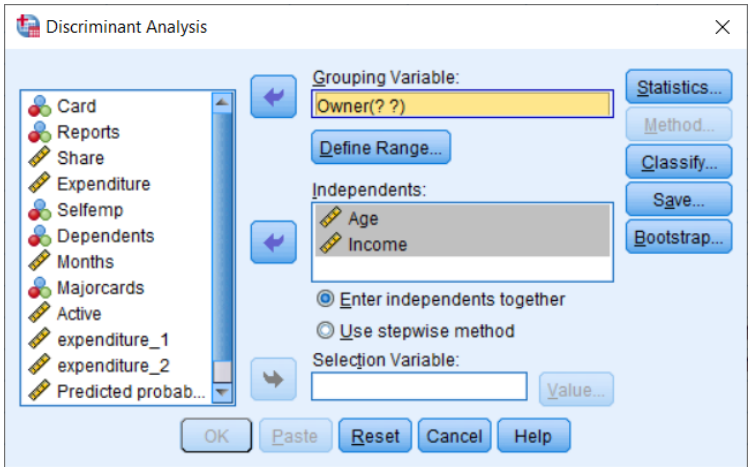
Εικόνα 7.15 *Reprint Courtesy of IHS Markit, © 2020 IHS Global Inc.*

7.5 Διαχωριστική ανάλυση

Στη συγκεκριμένη ενότητα θα αναλύσουμε τη λεγόμενη διαχωριστική ανάλυση (discriminant analysis) ή διακριτική ανάλυση, όπως είναι περισσότερο γνωστή. Αρχικά, ας υποθέσουμε ότι έχουμε μετρήσεις από μία ή περισσότερες μεταβλητές, ενώ, επίσης, γνωρίζουμε και τις διάφορες ομάδες του δείγματος (στο παράδειγμα της λογιστικής παλινδρόμησης που εκτιμήσαμε παραπάνω με τα δεδομένα των πιστωτικών καρτών (**credit.sav**), γνωρίζαμε αν το κάθε άτομο που συμμετείχε στο δείγμα ήταν ιδιοκτήτης σπιτιού, καθώς και το εισόδημα και την ηλικία του)). Σκοπός της διαχωριστικής ανάλυσης είναι να διαχωρίσουμε τους ιδιοκτήτες σπιτιών από τους μη ιδιοκτήτες με βάση τις μεταβλητές που έχουμε στη διάθεσή μας. Θα πρέπει να τονίσουμε στο σημείο αυτό ότι οι μεταβλητές που θα χρησιμοποιηθούν θα πρέπει οπωσδήποτε να είναι αυστηρά ποσοτικές (δηλαδή αριθμητικές), και πιο συγκεκριμένα να είναι συνεχείς ή έστω διακριτές με πολλές όμως τιμές. Δηλαδή, δεν μπορούν να χρησιμοποιηθούν μεταβλητές που αφορούν την ομάδα αίματος, τις πολιτικές πεποιθήσεις και, γενικά, κατηγορικές μεταβλητές.

Η διαχωριστική ανάλυση έχει δύο βασικές κατηγορίες, τη γραμμική (linear) και την τετραγωνική (quadratic) διαχωριστική ανάλυση. Στη γραμμική διαχωριστική ανάλυση χρησιμοποιούμε γραμμές ή «τοίχους», προκειμένου να διαχωρίσουμε τις ομάδες (οι οποίες θα πρέπει να είναι ήδη γνωστές, καθώς διαφορετικά δεν μπορούμε να την κάνουμε αυτήν την ανάλυση). Η τετραγωνική διαχωριστική ανάλυση μας επιτρέπει να διαχωρίσουμε τις ομάδες χρησιμοποιώντας καμπύλες γραμμές ή «καμπύλους τοίχους» και όχι ευθείες γραμμές. Στην ενότητα αυτή θα επικεντρωθούμε μόνο στη γραμμική περίπτωση, καθώς η δεύτερη περίπτωση (τετραγωνική διαχωριστική ανάλυση) δεν προσφέρεται από το SPSS. Θα πρέπει, επίσης, να σημειώσουμε πως το Eviews δεν περιλαμβάνει ούτε τη γραμμική ούτε την τετραγωνική διαχωριστική ανάλυση.

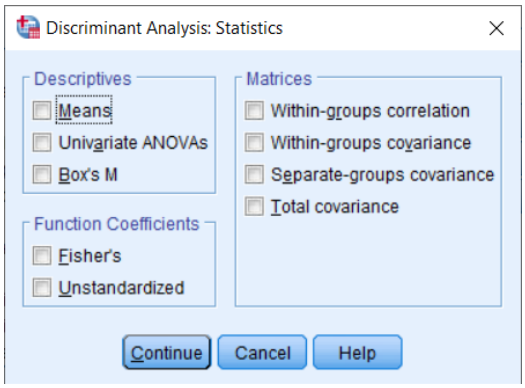
Προκειμένου να πραγματοποιήσουμε τη διαχωριστική ανάλυση στο SPSS, επιλέγουμε **Analyze → Classify → Discriminant**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 7.16**.



Εικόνα 7.16 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

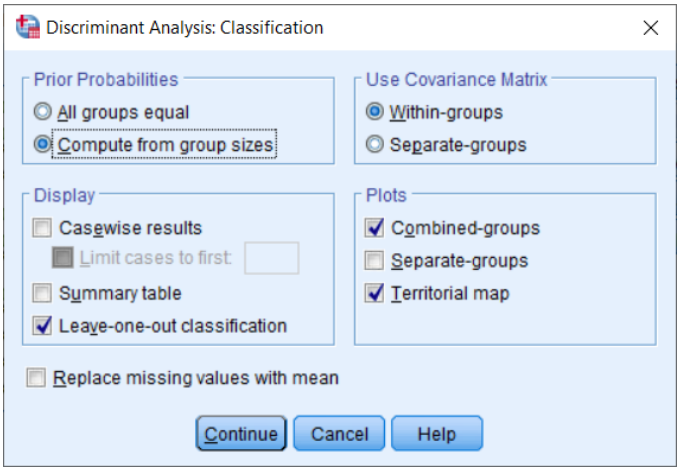
Η μεταβλητή που δηλώνει την ομάδα (ιδιοκτήτης του σπιτιού ή όχι) θα περαστεί στο κουτί με την ένδειξη **Grouping Variable:**. Στη συνέχεια, επιλέγουμε **Define Range**, προκειμένου να δηλώσουμε πόσες ομάδες

έχουμε. Στο παράδειγμά μας έχουμε 2 ομάδες, οι οποίες είναι αριθμημένες με 0 και 1, οπότε στη συγκεκριμένη ένδειξη θα συμπληρώσουμε 0 και 1. Όπως φαίνεται στην **Εικόνα 7.16**, οι συνεχείς μεταβλητές έχουν περαστεί στο κουτί με την ένδειξη **Independents:**. Ακριβώς από κάτω μας δίνεται η επιλογή να χρησιμοποιήσουμε όλες τις μεταβλητές στην ανάλυσή μας ή να επιτρέψουμε στο SPSS να επιλέξει αυτό ποιες μεταβλητές θα πρέπει να χρησιμοποιηθούν (κάτι αντίστοιχο αναλύσαμε και στην πολλαπλή παλινδρόμηση). Στην ανάλυσή μας θα το αφήσουμε ως έχει και θα επιλέξουμε **Statistics**, προκειμένου να εμφανιστεί το παράθυρο της **Εικόνας 7.17** (προφανώς, ο χρήστης του SPSS μπορεί να εμβαθύνει περισσότερο, αν το επιθυμεί).



Εικόνα 7.17 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Μπορούμε, επίσης, αν το επιθυμούμε, να «τσεκάρουμε» την επιλογή **Means**, προκειμένου να εμφανιστούν οι μέσοι των μεταβλητών για κάθε ομάδα ξεχωριστά. Αν στο παράθυρο της **Εικόνας 7.16** πατήσουμε **Classify**, τότε θα εμφανιστεί το παράθυρο της **Εικόνας 7.18**, στο οποίο θα «τσεκάρουμε» τις επιλογές **Compute from group sizes**, **Leave-one-out cross validation**, **Combined-groups** και **Territorial map**.

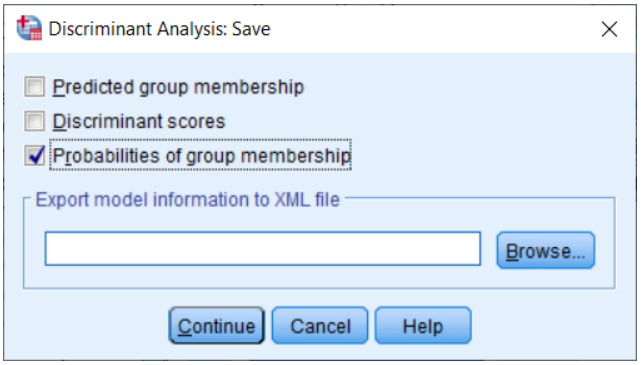


Εικόνα 7.18 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Στη συνέχεια, πατάμε **Continue** και επιστρέφουμε στο παράθυρο της **Εικόνας 7.16**, όπου θα επιλέξουμε **Save** και θα μεταβούμε στο παράθυρο της **Εικόνας 7.19**. Στο τελευταίο αυτό παράθυρο θα «τσεκάρουμε» τις επιλογές **Predicted group membership** και **Probabilities of group membership**. Ο αλγόριθμος bootstrap δεν θα χρησιμοποιηθεί στο σημείο αυτό, καθώς η χρήση του δεν πραγματοποιεί την ανάλυση που επιθυμούμε.

Έχοντας τελειώσει με τις διάφορες επιλογές, πατάμε **OK** στο παράθυρο της **Εικόνας 7.16**, προκειμένου να προκύψουν τα αποτελέσματά μας στο Output του SPSS. Θα εμφανιστούν διάφορα στατιστικά, όπως οι συντελεστές του πρώτου ιδιοδιανύσματος που χρησιμοποιήθηκε, προκειμένου να γίνει ο διαχωρισμός των δύο ομάδων, ο έλεγχος του Wilks για την ισότητα των μέσων των διανυσμάτων των δύο

ομάδων, καθώς και ο **Πίνακας 7.11**. Δυστυχώς, επειδή έχουμε μόνο δύο ομάδες παρατηρήσεων, δεν θα εμφανιστούν αυτόματα κάποια διαγράμματα. Οπότε, θα πρέπει να τα κατασκευάσουμε με τον τρόπο που έχουμε ήδη περιγράψει.



Εικόνα 7.19 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Ο **Πίνακας 7.11** είναι χωρισμένος σε δύο μέρη. Ας εξετάσουμε πρώτα το πάνω μέρος του (**Original**). Ο συγκεκριμένος αλγόριθμος παίρνει καθεμία παρατήρηση ξεχωριστά και με βάση τις μεταβλητές του υποδείγματος την κατατάσσει σε μία ομάδα. Αυτό το κάνει για όλες τις παρατηρήσεις και στο τέλος δημιουργεί τον συγκεκριμένο πίνακα. Στην πρώτη γραμμή εμφανίζονται οι τιμές 600 και 138, δηλαδή 738 στο σύνολο. Αυτό σημαίνει ότι από τους 738 μη ιδιοκτήτες σπιτιών, ο αλγόριθμος κατέταξε σωστά τους 600. Οι 138 θεωρήθηκαν, λανθασμένα, ιδιοκτήτες σπιτιών. Στη δεύτερη γραμμή έχουμε 581 ιδιοκτήτες σπιτιών, από τους οποίους οι 318 σωστά θεωρήθηκαν ιδιοκτήτες, ενώ οι υπόλοιποι όχι. Οπότε, έχουμε $138 + 263 = 401$ άτομα που καταχωρίστηκαν σε λάθος ομάδα. Αν μπορούσαμε να διεξάγουμε τετραγωνική διαχωριστική ανάλυση, αυτά τα λάθη θα ήταν ενδεχομένως μικρότερα. Συνεπώς, έχουμε 600 και 318 άτομα (στη διαγώνιο) τα οποία κατατάχθηκαν σωστά. Αυτό σημαίνει ότι 918 από τα 1319 άτομα (δηλαδή, το 69,6% του δείγματος) κατατάχθηκαν σωστά.

Πίνακας 7.11 *Αποτελέσματα επαλήθευσης ομάδων.*

			Predicted Group Membership		
		Owner	No	Yes	Total
Original	Count	No	600	138	738
		Yes	263	318	581
	%	No	81.3	18.7	100.0
		Yes	45.3	54.7	100.0
Cross-validated[b]	Count	No	597	141	738
		Yes	263	318	581
	%	No	80.9	19.1	100.0
		Yes	45.3	54.7	100.0

a. 69.6% of original grouped cases correctly classified.
b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.
c. 69.4% of cross-validated grouped cases correctly classified.

Στο κάτω μέρος του **Πίνακα 7.11** οι τιμές είναι λίγο διαφορετικές. Ο συγκεκριμένος αλγόριθμος πηγαίνει στην πρώτη παρατήρηση και την αφαιρεί. Στη συνέχεια, πραγματοποιεί τη διαχωριστική ανάλυση με βάση τις υπόλοιπες παρατηρήσεις και προβλέπει την ομάδα αυτής της τιμής. Αυτό γίνεται για όλες τις παρατηρήσεις. Η συγκεκριμένη μέθοδος επαλήθευσης του αλγόριθμου είναι πιο σωστή, γιατί δεν έχει συμμετοχή στον αλγόριθμο η κάθε παρατήρηση της οποίας η ομάδα προβλέπεται. Οπότε, το συνολικό εκτιμώμενο ποσοστό σωστής κατάταξης (69,49% στην περίπτωση αυτή) είναι πιο σωστό, πιο αμερόληπτο, πιο αντικειμενικό και πιο αντιπροσωπευτικό.

Βιβλιογραφία

Ελληνόγλωσση

Αδαμίδη, Ε. (2012). *Καμπύλες Λειτουργικού Χαρακτηριστικού Δέκτη και Στατιστική Ανάλυση Πραγματικών Ιατρικών Δεδομένων* (Διπλωματική Εργασία). Αθήνα: Εθνικό Μετσόβιο Πολυτεχνείο.

Ντζούφρας, Ι. (2005). *Εισαγωγή στη Βιοστατιστική και την Επιδημιολογία*. Αθήνα: Οικονομικό Πανεπιστήμιο Αθηνών.

Ξενόγλωσση

Johnson, R.A., & Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). New Jersey: Prentice Hall.

Kleinbaum, D.G., & Klein, M. (2002). *Logistic Regression: A Self-learning Text* (2nd ed.). New York: Springer.

Κεφάλαιο 8 Αξιοπιστία ερωτηματολογίου και παραγοντική ανάλυση

Σύνοψη

Το τελευταίο κεφάλαιο του βιβλίου επικεντρώνεται σε ζητήματα που αφορούν περισσότερο τις κοινωνικές επιστήμες, όπως είναι ο βαθμός ταύτισης δύο κατηγορικών μεταβλητών, η αξιοπιστία ενός ερωτηματολογίου, καθώς και η παραγοντική ανάλυση. Το κεφάλαιο αυτό αφορά αποκλειστικά τη χρήση του SPSS (καθώς οι τεχνικές που θα αναλυθούν δεν περιλαμβάνονται στο Eviews) και έχει ως στόχους να μπορεί ο χρήστης του συγκεκριμένου προγράμματος να κατανοεί τις τεχνικές αυτές και να είναι σε θέση να τις εφαρμόζει ορθά στο SPSS.

Προαπαιτούμενη γνώση

Απατούνται γνώσεις προχωρημένης στατιστικής.

8.1 Αξιοπιστία ή βαθμός ταύτισης δύο κατηγορικών μεταβλητών (kappa του Cohen)

Έστω, λοιπόν, ότι έχουμε στη διάθεσή μας δεδομένα που αφορούν κατατάξεις που προέρχονται από δύο κριτές (για παράδειγμα, βαθμολογίες σε σχολικό επίπεδο ή αποτελέσματα δύο μεθόδων) και επιθυμούμε να ερευνήσουμε τον βαθμό συμφωνίας των δύο κριτών ή των δύο μεθόδων που ακολούθησαν. Προκειμένου να διεξάγουμε τη συγκεκριμένη ανάλυση, θα πρέπει υπολογίσουμε τον συντελεστή kappa του Cohen[6] (1960).

Στο παράδειγμά μας, έχουμε τις 15 απαντήσεις των δύο κριτών για ένα συγκεκριμένο θέμα (δηλαδή, τεχνητά δεδομένα τα οποία είναι διαθέσιμα στο αρχείο **Cohen.sav**). Οι απαντήσεις που έδωσαν οι 2 κριτές είναι A, B ή C. Οπότε, θα ακολουθήσουμε σχεδόν την ίδια διαδικασία που εφαρμόσαμε για τον χ^2 έλεγχο ανεξαρτησίας (**Analyze → Descriptive Statistics → Crosstabs**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 4.20**). Θα φτιάξουμε δύο μεταβλητές, όπου η μία θα περιλαμβάνει τις απαντήσεις του πρώτου κριτή και η άλλη του δεύτερου, και θα τις περάσουμε στα κουτιά με τις ενδείξεις **Row(s):** και **Column(s):**. Στη συνέχεια, θα επιλέξουμε **Statistics**, προκειμένου να μεταβούμε στο παράθυρο της **Εικόνας 4.22**. Στο παράθυρο αυτό θα «τσεκάρουμε» την επιλογή **Kappa** που βρίσκεται στο κάτω δεξιό μέρος. Πατώντας **Continue** και στη συνέχεια **OK** προκύπτουν τα αποτελέσματα **(Πίνακας 8.1)**, στα οποία φαίνεται ότι οι δύο κριτές συμφωνούν στις $3 + 3 + 3 = 9$ από τις 15 απαντήσεις τους.

Πίνακας 8.1 *Πίνακας διπλής εισόδου με τις απαντήσεις των δύο κριτών.*

Rater 1 * Rater 2 Crosstabulation					
Count		Rater 2			
		A	B	C	Total
Rater 1	A	3	1	1	5
	B	1	3	1	5
	C	1	1	3	5
Total		5	5	5	15

Ο συντελεστής kappa του Cohen παίρνει τιμές από 0 (δηλαδή, απόλυτη ασυμφωνία) μέχρι 1 (δηλαδή, απόλυτη συμφωνία), αλλά μπορεί να πάρει και αρνητικές τιμές. Το τελευταίο υποδηλώνει αντίθετες

[6] Η Wikipedia παρουσιάζει τον τρόπο υπολογισμού του συντελεστή αυτού: http://en.wikipedia.org/wiki/Cohen's_kappa

απόψεις μεταξύ των δύο κριτών. Η τιμή του συντελεστή kappa στο παράδειγμά μας είναι ίση με 0,400, ενώ η αντίστοιχη *p*-value του ελέγχου της μηδενικής υπόθεσης ότι η τιμή του συντελεστή kappa είναι ίση με το μηδέν είναι ίση με 0,028. Αυτό σημαίνει ότι ο βαθμός συμφωνίας μεταξύ των δύο κριτών είναι στατιστικά σημαντικός, όμως δεν είναι τόσο υψηλός όσο ίσως να επιθυμούμε.

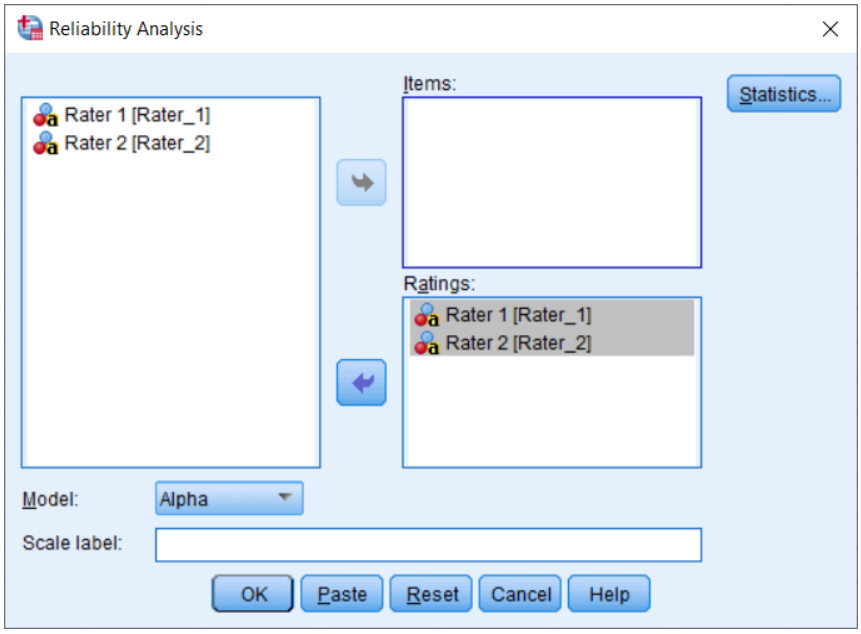
Πίνακας 8.2 *Συντελεστής κάππα του Cohen.*

Symmetric Measures					
		Value	Asymptotic Standard Error[a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	.400	.190	2.191	.028
N of Valid Cases		15			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.

8.2 Αξιοπιστία ή βαθμός ταύτισης περισσότερων από δύο κατηγορικών μεταβλητών (kappa του Fleiss)

Στην περίπτωση που έχουμε περισσότερες από δύο κατηγορικές μεταβλητές, συνήθως υπολογίζουμε τον συντελεστή kappa του Fleiss (1981). Στην ενότητα αυτή θα χρησιμοποιήσουμε τα ίδια δεδομένα με προηγουμένως, οπότε οι συντελεστές kappa του Cohen και του Fleiss προφανώς θα ταυτίζονται. Επιλέγουμε **Analyze→Scale→Reliability Analysis**, προκειμένου να εμφανιστεί το παράθυρο της **Εικόνας 8.1**. Στο παράθυρο αυτό, θα επιλέξουμε τις μεταβλητές που επιθυμούμε και θα τις περάσουμε στο κάτω δεξιά κουτί με την ένδειξη **Ratings:**.

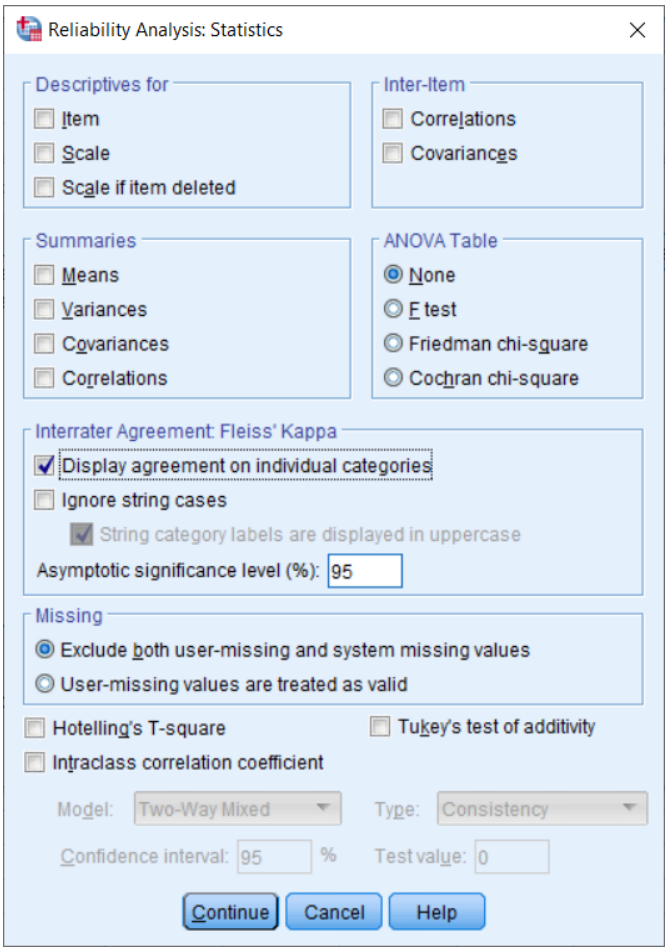


Εικόνα 8.1 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Επιλέγοντας **Statistics…** θα εμφανιστεί το παράθυρο της **Εικόνας 8.2**, στο οποίο έχουμε «τσεκάρει» την επιλογή **Display agreement on individual categories**. Πατάμε **Continue** και επιστρέφουμε στο παράθυρο της **Εικόνας 8.1**, όπου πατώντας **OK** θα εμφανιστούν δύο πίνακες. Ο **Πίνακας 8.3** είναι ο πρώτος από τους δύο πίνακες και αυτός που μας ενδιαφέρει περισσότερο.

Στον **Πίνακα 8.3** παρουσιάζονται ο συντελεστής kappa, το τυπικό του σφάλμα, η τιμή της ελεγχοσυνάρτησης, καθώς και η αντίστοιχη *p*-value για τον έλεγχο της μηδενικής υπόθεσης ότι ο συντελεστής kappa είναι μηδέν. Επίσης, εμφανίζεται και το 95% διάστημα εμπιστοσύνης για την

πραγματική τιμή του συγκεκριμένου συντελεστή. Όπως φαίνεται, υπάρχουν κάποιες επιπλέον πληροφορίες σε σχέση με την προηγούμενη μέθοδο.



Εικόνα 8.2 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Πίνακας 8.3 *Συντελεστής κάππα του Cohen και το διάστημα εμπιστοσύνης του.*

Overall Agreementa,b						
		Asymptotic			Asymptotic 95% Confidence Interval	
	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Agreement	.400	.183	2.191	.028	.389	.411
a. Sample data contains 15 effective subjects and 2 raters.						
b. Rating category values are case sensitive.						

8.3 Αξιοπιστία ενός ερωτηματολογίου (alpha του Cronbach)

Όταν πραγματοποιούμε ανάλυση μέσω ερωτηματολογίων είναι απαραίτητη η μέτρηση της αξιοπιστίας τους. Η συγκεκριμένη μέτρηση πραγματοποιείται χρησιμοποιώντας το alpha του Cronbach (1951) *(Cronbach's alpha).* Η τιμή του συντελεστή αξιοπιστίας που προκύπτει από τη συγκεκριμένη μεθοδολογία κυμαίνεται από 0 έως 1. Ο συντελεστής alpha, όπως ο συντελεστής διχοτομικής αξιοπιστίας (split-half coefficient), καθώς και μερικοί άλλοι, μετρούν ουσιαστικά την εσωτερική συνέπεια και όχι την αξιοπιστία ενός ερωτηματολογίου.

Γενικά, το alpha του Cronbach προτιμάται αντί του συντελεστή διχοτομικής αξιοπιστίας. Ο συντελεστής του ημίκλαστου, όπως αλλιώς λέγεται ο συντελεστής διχοτομικής αξιοπιστίας, χωρίζει τυχαία

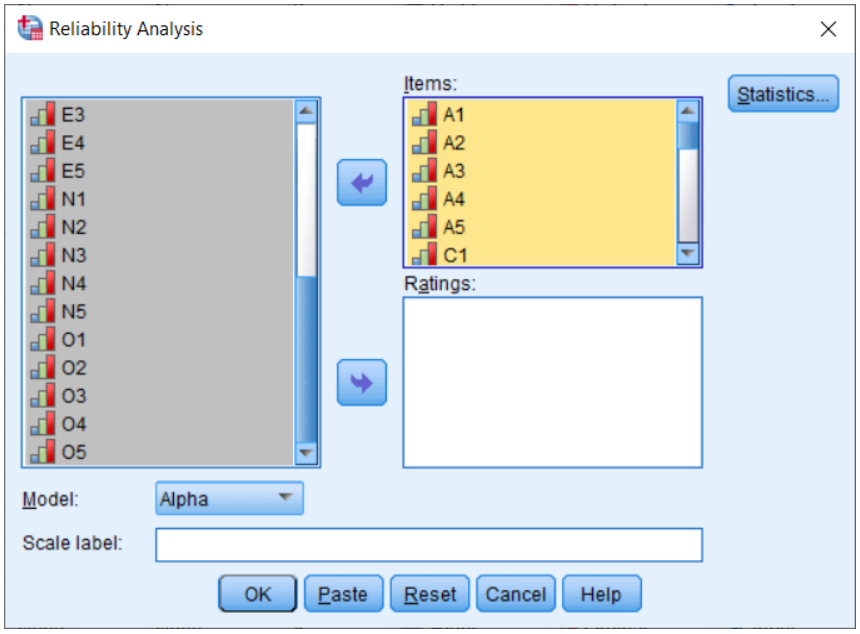
τις ερωτήσεις ενός ερωτηματολογίου στα δύο και στη συνέχεια υπολογίζει τον συντελεστή συσχέτισης μεταξύ των δύο σκορ που έχουν προκύψει από τα δύο «μισά» μέρη του ερωτηματολογίου. Επίσης, πρέπει οπωσδήποτε να χρησιμοποιηθεί και η φόρμουλα διόρθωσης των Spearman-Brown. Το πρόβλημα με τον διαχωρισμό των ερωτήσεων ενός ερωτηματολογίου στα δύο είναι ότι ο συγκεκριμένος συντελεστής βασίζεται σε έναν μόνο διαχωρισμό του ερωτηματολογίου. Όμως, αρκεί να σκεφτεί κανείς ότι για ένα ερωτηματολόγιο 20 ερωτήσεων υπάρχουν 184.756 δυνατοί τρόποι προκειμένου να «χωριστούν» οι ερωτήσεις στα δύο και συνεπώς, μπορούν να υπολογιστούν 184.756 διαφορετικοί συντελεστές διχοτομικής αξιοπιστίας. Ο συντελεστής alpha του Cronbach αντιμετωπίζει το συγκεκριμένο πρόβλημα. Μαθηματικά, η τιμή του alpha ισούται με τον μέσο όρο όλων αυτών των διαφορετικών συντελεστών διχοτομικής αξιοπιστίας. Επιπλέον, οι συντελεστές alpha δεν χρειάζονται διόρθωση μέσω της φόρμουλας των Spearman-Brown.

Η ανάλυση του Charter (2003) επιβεβαιώνει ότι η συχνότητα χρησιμοποίησης του διχοτομικού συντελεστή αξιοπιστίας τείνει να μειώνεται με την πάροδο των ετών, ενώ ο συντελεστής alpha του Cronbach ολοένα και κερδίζει έδαφος. Είναι απαραίτητο να αναφερθεί στο σημείο αυτό ότι, αν οι τιμές των συντελεστών αξιοπιστίας είναι μεγαλύτερες του 0,70, αυτό σημαίνει ότι το ερωτηματολόγιο είναι αξιόπιστο.

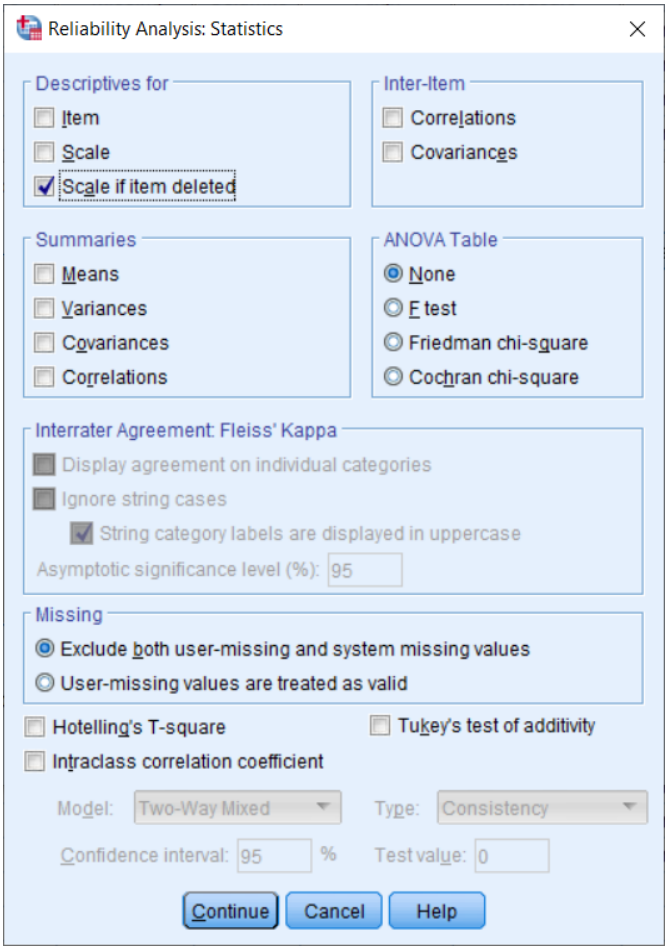
Στο παράδειγμά μας, χρησιμοποιούμε τις απαντήσεις από ένα ερωτηματολόγιο αξιολόγησης προσωπικότητας, το οποίο αποτελείται από 25 ερωτήσεις[7] και το οποίο μοιράστηκε σε 2.800 άτομα. Τα δεδομένα είναι διαθέσιμα στο αρχείο **bfi.sav**. Οι δυνατές απαντήσεις ήταν σε διατακτική κλίμακα, από το 1 έως και το 6. Προκειμένου να μετρήσουμε την αξιοπιστία του συγκεκριμένου ερωτηματολογίου χρησιμοποιώντας το SPSS, επιλέγουμε **Analyze → Scale → Reliability Analysis**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 8.3**, το οποίο είναι ίδιο με αυτό της **Εικόνας 8.1**. Στο παράθυρο αυτό, θα περάσουμε τις απαντήσεις στο δεξί κουτί με την ένδειξη **Items:**, ενώ στην επιλογή **Model** είναι προεπιλεγμένο το **Alpha** (με τον τρόπο αυτό το SPSS υπολογίζει τον συντελεστή αξιοπιστίας alpha του Cronbach). Στη συνέχεια, πατώντας **Statistics** θα εμφανιστεί το παράθυρο της **Εικόνας 8.4.** Στο παράθυρο αυτό έχουμε «τσεκάρει» το κουτάκι **Scale if item deleted**, προκειμένου το SPSS να υπολογίσει την αξιοπιστία του ερωτηματολογίου εξαιρώντας μία ερώτηση κάθε φορά. Αυτό αποτελεί μία ένδειξη για το αν θα πρέπει μία συγκεκριμένη ερώτηση να παραμείνει στο ερωτηματολόγιο ή να απαλειφθεί από αυτό. Αν, για παράδειγμα, η αξιοπιστία του ερωτηματολογίου αυξάνεται χωρίς τη συγκεκριμένη ερώτηση, αυτό αποτελεί μία ένδειξη για περαιτέρω μελέτη της ερώτησης αυτής. Βεβαίως, ισχύει και το αντίστροφο. Αν, δηλαδή, η αξιοπιστία του ερωτηματολογίου μειώνεται χωρίς τη συγκεκριμένη ερώτηση, αυτό αποτελεί μία ένδειξη καταλληλότητας της ερώτησης αυτής. Τέλος, πατώντας **Continue** στο παράθυρο της **Εικόνας 8.4** και στη συνέχεια **OK**, προκύπτουν τα αποτελέσματα στο Output του SPSS **(Πίνακες 8.4** και **8.5)**.

Όπως φαίνεται στον **Πίνακα 8.4**, η τιμή του συντελεστή alpha είναι ίση με 0,525. Αναφέραμε προηγουμένως ότι, για να θεωρηθεί ένα ερωτηματολόγιο αξιόπιστο, θα πρέπει η τιμή του συντελεστή alpha να είναι μεγαλύτερη από το 0,70. Επίσης, σε κάποιες επιστήμες, όπως είναι η ιατρική, η επιθυμητή αξιοπιστία πρέπει να είναι μεγαλύτερη από το 0,90 ή ακόμα και το 0,95. Ο **Πίνακας 8.5** παρουσιάζει τους συντελεστές alpha δίπλα από κάθε ερώτηση, οι οποίοι έχουν υπολογιστεί έχοντας αφαιρεθεί αυτή η ερώτηση. Παρατηρούμε ότι, όταν αφαιρεθεί η πρώτη ερώτηση (A1), η αξιοπιστία αυξάνεται στο 0,538, ενώ, αν αφαιρεθεί η δεύτερη ερώτηση (A2), η αξιοπιστία μειώνεται στο 0,511. Αυτό σημαίνει ότι η προσθήκη της ερώτησης A1 μειώνει την αξιοπιστία του ερωτηματολογίου, ενώ η προσθήκη της ερώτησης A2 αυξάνει την αξιοπιστία του.

[7]Τα δεδομένα συλλέχθηκαν από τους Revelle, Wilt & Rosenthal (2010) για το project SAPA (https://www.sapa-project.org/).



Εικόνα 8.3 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Εικόνα 8.4 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Πίνακας 8.4 *Συντελεστής άλφα του Cronbach.*

Reliability Statistics	
Cronbach's Alpha	N of Items
.525	25

Πίνακας 8.5: *Αξιοπιστία του ερωτηματολογίου, όταν αφαιρούνται ερωτήσεις.*

Item-Total Statistics				
	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	**Cronbach's Alpha if Item Deleted**
A1	91.80	99.605	-.013	**.538**
A2	89.41	95.790	.175	**.511**
A3	89.61	94.351	.202	**.507**
A4	89.52	96.538	.085	**.524**
A5	89.66	96.675	.118	**.519**
C1	89.68	97.353	.096	**.521**
C2	89.83	94.931	.177	**.510**
C3	89.90	98.322	.048	**.528**
C4	91.65	97.238	.077	**.525**
C5	90.90	96.890	.052	**.531**
E1	91.23	101.403	-.086	**.554**
E2	91.05	98.983	-.011	**.541**
E3	90.22	94.676	.180	**.510**
E4	89.80	97.818	.043	**.530**
E5	89.81	96.134	.125	**.518**
N1	91.26	87.024	.399	**.471**
N2	90.69	87.807	.385	**.474**
N3	90.98	85.880	.433	**.464**
N4	91.00	89.552	.310	**.487**
N5	91.23	87.928	.350	**.479**
O1	89.39	97.612	.105	**.520**
O2	91.52	95.189	.120	**.519**
O3	89.75	97.090	.113	**.519**
O4	89.28	95.074	.203	**.508**
O5	91.74	98.659	.031	**.531**

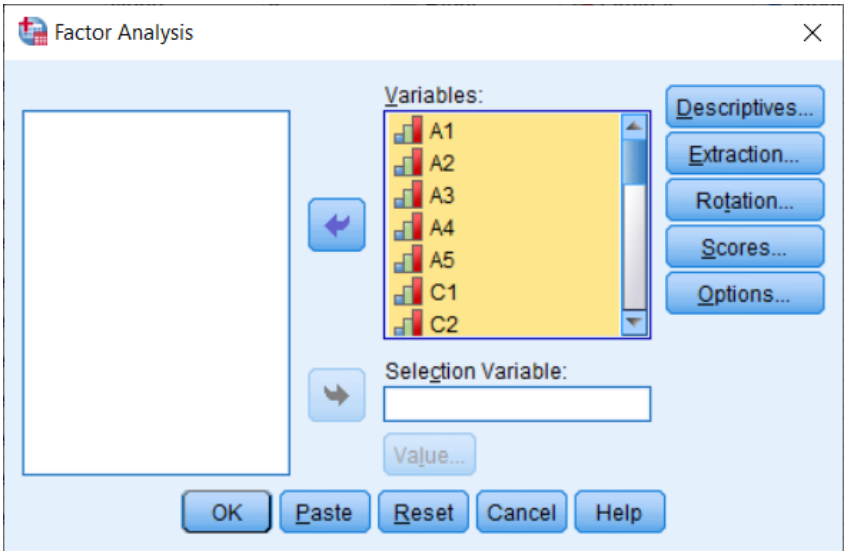
8.4 Παραγοντική ανάλυση σε ένα ερωτηματολόγιο

Ο Charles Spearman το 1904 ήταν ο πρώτος ψυχολόγος που μελέτησε την παραγοντική ανάλυση. Μας παρείχε πολλές λεπτομέρειες για τη μέθοδο που ακολούθησε, η οποία, όμως, ήταν επικεντρωμένη σε υποδείγματα με έναν παράγοντα. Ο Spearman ανακάλυψε ότι τα αποτελέσματα κάποιων μαθητών σε διάφορα (φαινομενικά ανεξάρτητα) αντικείμενα είχαν θετική συσχέτιση μεταξύ τους. Αυτό τον οδήγησε στο να διαμορφώσει την πεποίθηση ότι μία γενική νοητική ικανότητα αποτελούσε τη βάση και οδηγούσε στον σχηματισμό της ανθρώπινης γνωστικής συμπεριφοράς.

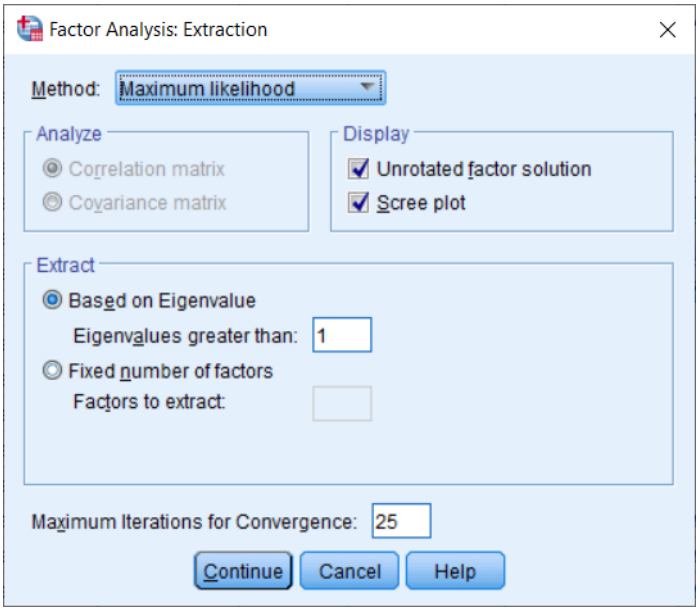
Ακριβώς αυτή είναι και η λογική της παραγοντικής ανάλυσης. Δηλαδή, να εξάγει έναν αριθμό λανθανουσών (μη παρατηρηθεισών) μεταβλητών, οι οποίες ονομάζονται παράγοντες. Οι παράγοντες αυτοί αποτελούν τις μεταβλητές που βρίσκονται πίσω από τις παρατηρηθείσες μεταβλητές και είναι εκείνοι που μας βοηθούν να καταλάβουμε τις παρατηρηθείσες μεταβλητές μέσω της ομαδοποίησής τους. Οι παράγοντες που θα εξαχθούν θα πρέπει να είναι σε θέση να αναπαράξουν τον πίνακα συσχέτισης, ο οποίος υπολογίζεται από τις παρατηρηθείσες μεταβλητές. Στην πράξη βέβαια αυτό δεν συμβαίνει, με αποτέλεσμα να είμαστε ικανοποιημένοι, όταν οι αποκλίσεις τους είναι μικρές.

Στην ανάλυσή μας θα χρησιμοποιήσουμε τα ίδια δεδομένα με τις προηγούμενες ενότητες του κεφαλαίου αυτού (**bfi.sav**), προκειμένου να αναλύσουμε λίγο καλύτερα την παραγοντική ανάλυση. Οι 25 ερωτήσεις ομαδοποιούνται σε 5 υποτιθέμενους παράγοντες: Agreeableness, Conscientiousness, Extraversion, Neuroticism και Openness. Οπότε, διεξάγοντας την παραγοντική ανάλυση αναμένουμε να εξαχθούν οι 5 αυτοί παράγοντες, οι οποίοι θα έχουν την κατάλληλη βαρύτητα ή θα συσχετίζονται κατάλληλα με τις 5 ερωτήσεις που τους αφορούν.

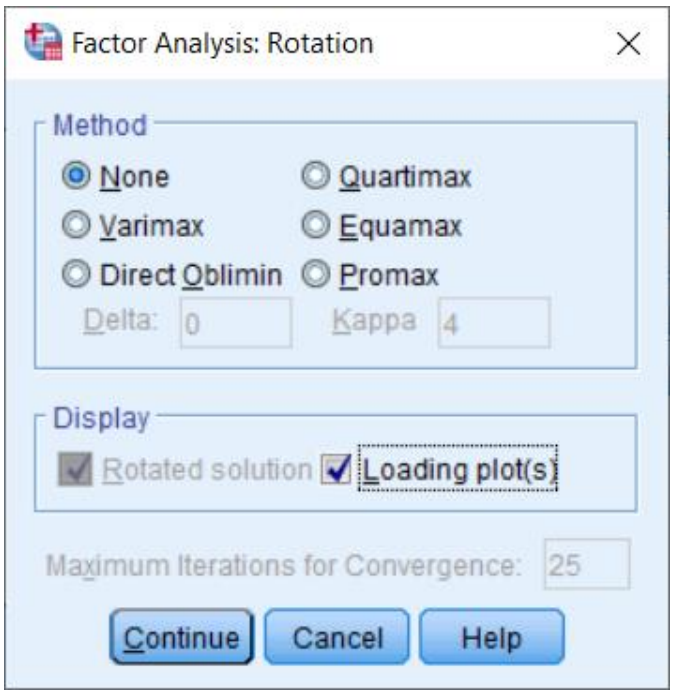
Προκειμένου να διεξάγουμε την παραγοντική ανάλυση, επιλέγουμε **Analyze → Dimensionality Reduction → Factor…**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 8.5**. Στο παράθυρο αυτό θα περάσουμε όλες τις ερωτήσεις στο δεξιό κουτί με την ένδειξη **Variables:**. Στη συνέχεια, πατώντας **Extraction** θα εμφανιστεί το παράθυρο της **Εικόνας 8.6**, όπου στην επιλογή **Method** θα επιλέξουμε **Maximum likelihood**, ενώ θα «τσεκάρουμε» και την επιλογή **scree plot**. Πατώντας **Continue** θα επιστρέψουμε στο παράθυρο της **Εικόνας 8.5**, στο οποίο θα επιλέξουμε **Rotation**, με αποτέλεσμα να εμφανιστεί το παράθυρο της **Εικόνας 8.7**. Στο τελευταίο αυτό παράθυρο μπορούμε να «τσεκάρουμε» την επιλογή **Loading plot(s)**. Επίσης, στο μενού **Method** δεν έχουμε επιλέξει κάτι (**None**). Οι επιλογές αυτές αφορούν μεθόδους περιστροφής των παραγόντων και θα εξηγήσουμε στη συνέχεια τι σημαίνουν. Και πάλι, πατώντας **Continue** επιστρέφουμε στο παράθυρο της **Εικόνας 8.5**, όπου επιλέγοντας **Scores…** θα εμφανιστεί το παράθυρο της **Εικόνας 8.8** στο οποίο και πάλι δεν κάνουμε κάποια επιλογή. Τέλος, πατώντας ξανά **Continue** επιστρέφουμε στο παράθυρο της **Εικόνας 8.5** και επιλέγουμε **Options**, με αποτέλεσμα να προκύψει το παράθυρο της **Εικόνας 8.9**. Στο τελευταίο αυτό παράθυρο «τσεκάρουμε» το κουτί **Supress small coefficients**. Πατώντας **Continue** επιστρέφουμε για άλλη μία φορά στο παράθυρο της **Εικόνας 8.5**, όπου πατώντας **OK** εμφανίζονται στο Output του SPSS μία σειρά από πίνακες **(Πίνακες 8.6 - 8.9)** και γραφήματα **(Διαγράμματα 8.1** και **8.2)**.



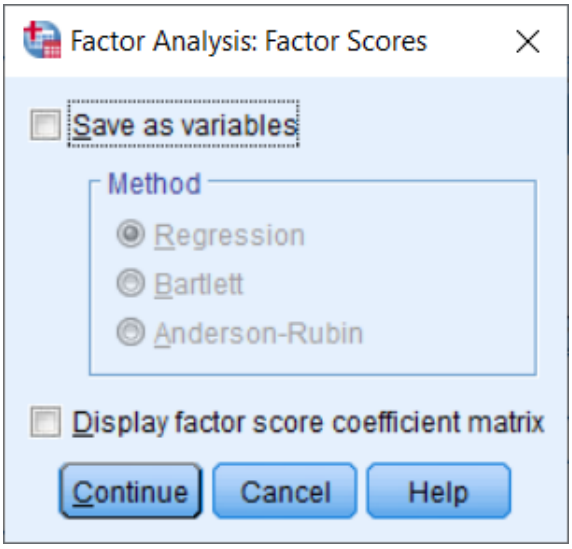
Εικόνα 8.5 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



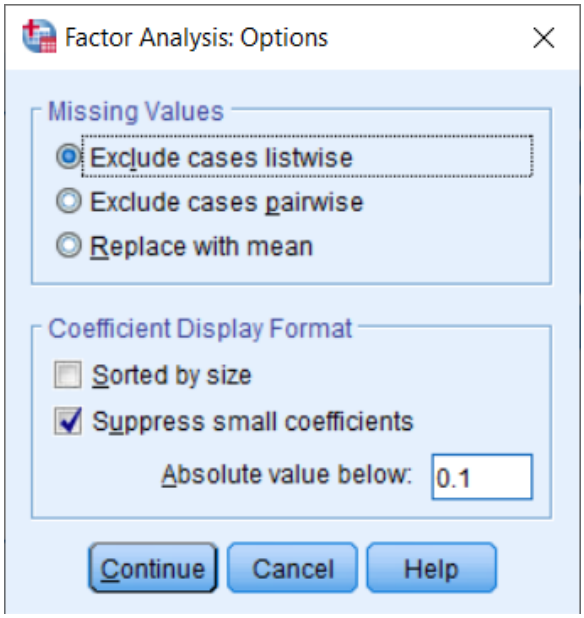
Εικόνα 8.6 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Εικόνα 8.7 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Εικόνα 8.8 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*



Εικόνα 8.9 *Reprint Courtesy of International Business Machines Corporation, © International Business Machines Corporation.*

Ο πρώτος από τους πίνακες των αποτελεσμάτων **(Πίνακας 8.6)** παρουσιάζει τα λεγόμενα Communalities των ερωτήσεων, δηλαδή, το ποσοστό της διακύμανσης της κάθε ερώτησης που μπορεί να εξηγηθεί από τους εξαχθέντες παράγοντες. Το Communality θυμίζει κάπως ως έννοια τον συντελεστή προσδιορισμού (R^2) που αναλύσαμε στο κεφάλαιο 5 που αφορούσε τη γραμμική παλινδρόμηση.

Ο δεύτερος από τους πίνακες των αποτελεσμάτων **(Πίνακας 8.7)** εμφανίζει τις ιδιοτιμές όλων των παραγόντων. Ωστόσο, στις 3 τελευταίες στήλες φαίνονται τιμές μόνο για τους 6 πρώτους παράγοντες. Θα πρέπει να υπενθυμίσουμε στο σημείο αυτό πως στο παράθυρο της **Εικόνας 8.6** είχαμε την επιλογή να ορίσουμε τον αριθμό των παραγόντων που θέλουμε να εξαχθούν. Η επιλογή μας ήταν να εξαχθούν όσοι παράγοντες θα είχαν ιδιοτιμή μεγαλύτερη του 1. Οπότε, από τα αποτελέσματα του **Πίνακα 8.7** προκύπτει ότι 6 εξαχθέντες παράγοντες εξηγούν σωρευτικά σχεδόν το 45% της συνολικής διακύμανσης των

μεταβλητών (ερωτήσεων). Ο συγκεκριμένος πίνακας παρουσιάζει, επίσης, το ποσοστό του κάθε παράγοντα ξεχωριστά.

Στο πρώτο γράφημα των αποτελεσμάτων **(Διάγραμμα 8.1)** απεικονίζονται οι ιδιοτιμές των παραγόντων. Ένα τέτοιο διάγραμμα είναι πολλές φορές χρήσιμο, καθώς μας βοηθάει γραφικά να αποφασίσουμε πόσους παράγοντες χρειαζόμαστε. Όπως φαίνεται στο συγκεκριμένο διάγραμμα, σχηματίζεται μία καμπύλη με αρνητική κλίση που από κάποιο σημείο και μετά γίνεται ευθεία (plateau) ή αλλιώς αποκτά μία σταθερότητα. Το σημείο, στο οποίο γίνεται η αλλαγή της κλίσης (και όπου σχηματίζεται ένας αγκώνας), είναι το σημείο που σταματάμε, καθώς σε αυτό βρίσκεται ο αριθμός των παραγόντων που χρειαζόμαστε.

Στον τρίτο από τους πίνακες των αποτελεσμάτων **(Πίνακας 8.8)** παρουσιάζονται οι συντελεστές βαρύτητας των ερωτήσεων για κάθε παράγοντα. Οι συντελεστές αυτοί δείχνουν τη βαρύτητα που έχει η κάθε ερώτηση στον κάθε παράγοντα. Παρατηρήστε ότι σε ορισμένα κελιά του **Πίνακα 8.8** δεν εμφανίζονται τιμές. Ο λόγος είναι ότι οι τιμές αυτές είναι μικρότερες του 0,1. Θυμηθείτε ότι στο παράθυρο της **Εικόνας 8.9** επιλέξαμε να μην εμφανιστούν οι τιμές που έχουν απόλυτη τιμή μικρότερη του 0,1. Αν στο παράθυρο αυτό επιλέγαμε μεγαλύτερη τιμή, τότε θα εμφανιζόντουσαν ακόμα λιγότερες τιμές στον **Πίνακα 8.8**. Με βάση, λοιπόν, τις τιμές που εμφανίζονται στον συγκεκριμένο πίνακα, θα προσπαθήσουμε να ερμηνεύσουμε τους παράγοντες. Καθώς γνωρίζουμε ότι οι 25 ερωτήσεις του ερωτηματολογίου ομαδοποιούνται σε 5 ομάδες, αναμένουμε μεγάλες (κατά απόλυτη τιμή) τιμές σε κάποιες ερωτήσεις για κάθε παράγοντα. Οπότε και η ομαδοποίηση θα είναι ξεκάθαρη (θα επιστρέψουμε, όμως, στο σημείο αυτό λίγο αργότερα).

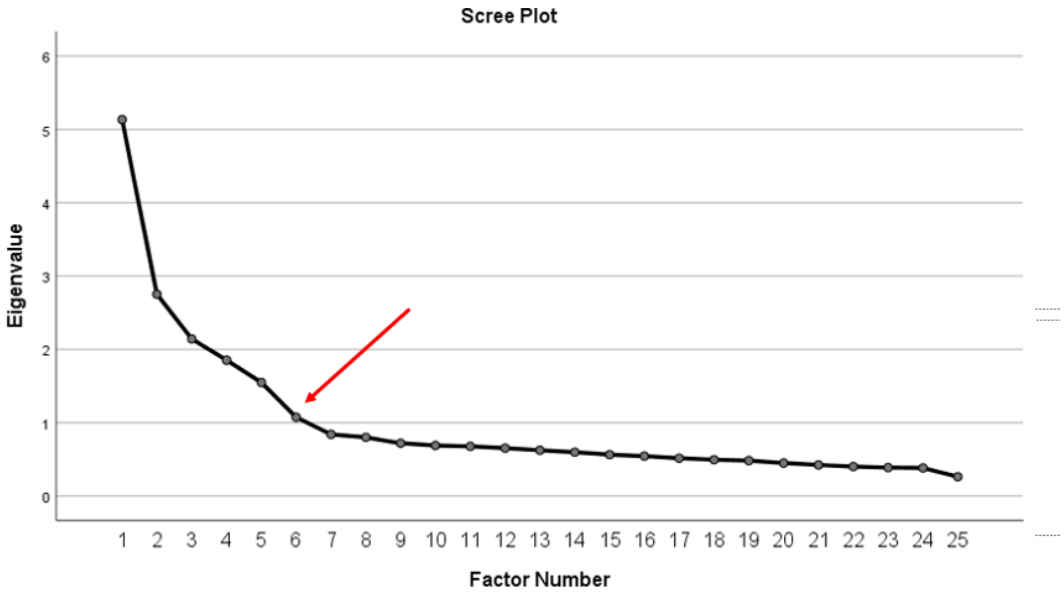
Πίνακας 8.6 *Communalities των ερωτήσεων.*

Communalities		
	Initial	Extraction
A1	.201	.325
A2	.393	.518
A3	.433	.527
A4	.270	.304
A5	.418	.484
C1	.294	.363
C2	.358	.502
C3	.255	.315
C4	.415	.576
C5	.376	.434
E1	.332	.387
E2	.458	.547
E3	.398	.478
E4	.472	.568
E5	.367	.401
N1	.590	.727
N2	.572	.698
N3	.476	.520
N4	.439	.496
N5	.318	.348
O1	.256	.337
O2	.214	.297
O3	.353	.490
O4	.178	.244
O5	.232	.360
Extraction Method: Maximum Likelihood.		

Ο τέταρτος από τους πίνακες των αποτελεσμάτων **(Πίνακας 8.9)** εμφανίζει το αποτέλεσμα του ελέγχου καλής προσαρμογής των 6 εξαχθέντων παραγόντων, δηλαδή κατά πόσο ο πίνακας συσχέτισης που σχηματίζεται από τους 6 εξαχθέντες παράγοντες αποκλίνει από τον παρατηρηθέντα πίνακα συσχέτισης. Όπως φαίνεται στον **Πίνακα 8.9**, η *p*-value (Sig.) είναι ίση με το μηδέν. Αυτό σημαίνει ότι οι 6 παράγοντες δεν είναι αρκετοί και άρα το υπόδειγμα της παραγοντικής ανάλυσης δεν έχει κάνει καλή προσαρμογή. Επιπλέον, το δεύτερο γράφημα των αποτελεσμάτων (**Διάγραμμα 8.2**) απεικονίζει στον τριδιάστατο χώρο τους συντελεστές των 3 πρώτων παραγόντων.

Πίνακας 8.7 *Ιδιοτιμές του κάθε παράγοντα.*

Total Variance Explained						
	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.134	20.537	20.537	4.480	**17.919**	17.919
2	2.752	11.008	31.545	2.385	**9.541**	27.460
3	2.143	8.571	40.116	1.576	**6.302**	33.762
4	1.852	7.409	47.525	1.255	**5.018**	38.780
5	1.548	6.193	53.718	1.015	**4.059**	42.839
6	1.074	4.294	58.012	.536	**2.144**	**44.983**
7	.840	3.358	61.370			
8	.799	3.197	64.567			
9	.719	2.876	67.443			
10	.688	2.752	70.195			
11	.676	2.705	72.901			
12	.652	2.607	75.508			
13	.623	2.493	78.001			
14	.597	2.386	80.387			
15	.563	2.252	82.640			
16	.543	2.173	84.813			
17	.515	2.058	86.871			
18	.495	1.978	88.849			
19	.483	1.931	90.779			
20	.449	1.796	92.575			
21	.423	1.693	94.269			
22	.401	1.603	95.871			
23	.388	1.551	97.422			
24	.382	1.527	98.950			
25	.263	1.050	100.000			
Extraction Method: Maximum Likelihood.						



Διάγραμμα 8.1 *Scree plot των παραγόντων.*

Αν τώρα στο παράθυρο της **Εικόνας 8.6** επιλέξουμε ως μέθοδο εξαγωγής των παραγόντων τη μέθοδο **Principal components**, στο παράθυρο της **Εικόνας 8.7** επιλέξουμε την περιστροφή **Varimax** και στο παράθυρο της **Εικόνας 8.9** επιλέξουμε να μην εμφανίζονται οι τιμές των συντελεστών που είναι χαμηλότερες του 0,45 (κατά απόλυτη τιμή), τότε θα προκύψει ο **Πίνακας 8.10.** Στον συγκεκριμένο πίνακα φαίνεται πιο καθαρά ότι 3 διαφορετικές ομάδες ερωτήσεων έχουν μεγάλη βαρύτητα στους τρεις πρώτους παράγοντες. Οι ερωτήσεις N1-N5 έχουν μεγάλη βαρύτητα στον πρώτο παράγοντα, οι ερωτήσεις C1-C5 στον δεύτερο παράγοντα και οι ερωτήσεις A1-A5 στον τρίτο παράγοντα. Επίσης, 4 από τις 5 ερωτήσεις της ομάδας E έχουν μεγάλη βαρύτητα στον τέταρτο παράγοντα, ενώ οι ερωτήσεις της ομάδας O έχουν μεγάλη βαρύτητα στον πέμπτο και στον έκτο παράγοντα.

Θα πρέπει να επισημάνουμε στο σημείο αυτό πως η άσκηση αυτή είχε σκοπό να δείξει τη σημαντικότητα των παραθύρων των **Εικόνων 8.6, 8.7** και **8.9**. Η επιλογή (α) της μεθόδου εξαγωγής των παραγόντων, (β) της μεθόδου περιστροφής, και (γ) του ορίου κάτω από το οποίο δεν θα εμφανίζονται οι τιμές των συντελεστών, έχει ιδιαίτερη σημασία, προκειμένου να ερμηνεύσουμε τους παράγοντες με τον καλύτερο δυνατό τρόπο.

Πίνακας 8.8 *Πίνακας συντελεστών των παραγόντων για κάθε ερώτηση.*

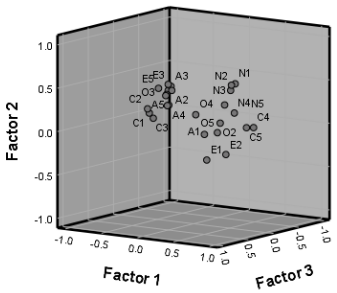
Factor Matrixa						
	Factor					
	1	2	3	4	5	6
A1	.238		.123		-.376	.331
A2	-.407	.363	-.152	.141	.363	-.212
A3	-.465	.395	-.220	.120	.300	
A4	-.389	.207		.273	.173	
A5	-.546	.289	-.245		.175	
C1	-.293	.202	.459			.155
C2	-.272	.262	.523	.159		.227
C3	-.283	.134	.394	.233		
C4	.455		-.541	-.164		.216
C5	.488		-.354	-.231	.126	
E1	.356	-.305	.246		.278	.161
E2	.583	-.227	.190		.336	
E3	-.448	.431	-.138	-.200		.172
E4	-.554	.328	-.289	.130	-.183	.143
E5	-.411	.425	.108		-.189	
N1	.604	.569			-.168	
N2	.587	.559			-.118	-.136
N3	.526	.482				
N4	.584	.228			.293	
N5	.415	.298		.172	.205	.123
O1	-.271	.243	.133	-.414		.123
O2	.195		-.251	.407		.171
O3	-.331	.344		-.495		
O4	.107	.173	.109	-.283	.329	
O5	.180		-.218	.459		.236

Extraction Method: Maximum Likelihood.
a. 6 factors extracted. 4 iterations required.

Πίνακας 8.9 χ^2 *έλεγχος καλής προσαρμογής των παραγόντων που εξήχθησαν.*

Goodness-of-fit Test		
Chi-Square	df	Sig.
896.699	165	**.000**

Factor Plot



Διάγραμμα 8.2 *Τριδιάστατο γράφημα των 3 πρώτων παραγόντων.*

Πίνακας 8.10 *Πίνακας συντελεστών των περιστραμμένων παραγόντων για κάθε ερώτηση.*

Rotated Component Matrixa						
	Component					
	1	2	3	4	5	6
A1			-.661			
A2			.749			
A3			.710			
A4			.548			
A5			.584			
C1		.653				
C2		.738				
C3		.678				
C4		-.689				
C5		-.625				
E1				.729		
E2				.728		
E3					.579	
E4				-.582		
E5				-.512		
N1	.837					
N2	.835					
N3	.795					
N4	.617					
N5	.608					
O1					.689	
O2						.662
O3					.663	
O4						
O5						.704

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 7 iterations.

Βιβλιογραφία

Ξενόγλωσση

Charter, R.A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of General Psychology*, *130*(3), 290-304. https://doi.org/10.1080/00221300309601160

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. https://doi.org/10.1007/BF02310555

Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions* (2nd ed.). New York: John Wiley and Sons Inc.

Johnson, R.A., & Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). New Jersey: Prentice Hall.

Revelle, W., Wilt, J., & Rosenthal, A. (2010) Individual differences in cognition: new methods for examining the personality-cognition link. In A. Gruszka, G. Matthews & B. Szymura (Eds.), *Handbook of individual differences in cognition: attention, memory and executive control* (pp. 27-49). New York: Springer Science. https://doi.org/10.1007/978-1-4419-1210-7_2

Ενδεικτική Βιβλιογραφία

Ελληνόγλωσση

Αδαμίδη, Ε. (2012). *Καμπύλες Λειτουργικού Χαρακτηριστικού Δέκτη και Στατιστική Ανάλυση Πραγματικών Ιατρικών Δεδομένων* (Διπλωματική Εργασία). Αθήνα: Εθνικό Μετσόβιο Πολυτεχνείο.

Βιτωράτου, Σ., & Λύκου, Α. (2017). *Σημειώσεις Σεμιναρίου SPSS* Αθήνα: Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών.

Δαφέρμος, Β. (2002). *Επαναληπτικές Στατιστικές Μετρήσεις στις Κοινωνικές Επιστήμες*. Αθήνα: Εκδόσεις Leader Books.

Δαφέρμος, Β. (2005). *Κοινωνική Στατιστική με το SPSS*. Θεσσαλονίκη: Εκδόσεις Ζήτη.

Καλαματιανού, Α. (2003). *Κοινωνική Στατιστική, Μέθοδοι Μονοδιάστατης Ανάλυσης*. Αθήνα: Εκδόσεις Παπαζήση.

Μπερσίμης, Σ. (2007). *Διδακτικές Σημειώσεις Προγράμματος Επιμόρφωσης στην Ιατρική Στατιστική*. Πειραιάς: Πανεπιστήμιο Πειραιώς.

Ντζούφρας, Ι. (2005). *Εισαγωγή στη Βιοστατιστική και την Επιδημιολογία*. Αθήνα: Οικονομικό Πανεπιστήμιο Αθηνών.

Ξεκαλάκη, Ε. (2001). *Μη Παραμετρική Στατιστική*. Αθήνα: Εκδόσεις Μπένου.

Ξεκαλάκη, Ε. (2004). *Τεχνικές Δειγματοληψίας*. Αθήνα: Εκδόσεις Οικονομικού Πανεπιστημίου Αθηνών.

Πανάρετος, Ι., & Ξεκαλάκη, Ε. (2003). *Εισαγωγή στη Στατιστική Σκέψη. Περιγραφική Στατιστική* (Τόμ. Ι). Αθήνα: Εκδόσεις Ι. Πανάρετος.

Πανάρετος, Ι., & Ξεκαλάκη, Ε. (2003). *Εισαγωγή στη Στατιστική Σκέψη. Εισαγωγή στις Πιθανότητες και στη Στατιστική Συμπερασματολογία* (Τόμος ΙΙ). Αθήνα: Εκδόσεις Ι. Πανάρετος.

Πανάρετος, Ι. (2001). *Γραμμικά Μοντέλα με Έμφαση στις Εφαρμογές*. Αθήνα: Οικονομικό Πανεπιστήμιο Αθηνών.

Παυλόπουλος, Β. (2008). *Μοντέλα Ανάλυσης Διακύμανσης*. Αθήνα: Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών.

Ρούσσος, Π., & Ευσταθίου, Γ. (2008). *Σύντομο Εγχειρίδιο SPSS 16.0* Αθήνα: Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών.

Τσαγρής, Μ. (2006). *Ανάλυση Διακύμανσης στο SPSS* (Πανεπιστημιακές Σημειώσεις).

Χαλικιάς, Ι. (2003). *Στατιστική: Μέθοδοι Ανάλυσης για Επιχειρηματικές Αποφάσεις* (2η έκδ.). Αθήνα: Εκδόσεις Rosili.

Ψιλούτσικου, Μ. (2005). *Σημειώσεις για το Μάθημα Ποσοτικές Μέθοδοι ΙΙ*. Αθήνα: Οικονομικό Πανεπιστήμιο Αθηνών.

Ξενόγλωσση

Bartholomew, D.J., Steele, F., & Moustaki, I. (2008). *Analysis of Multivariate Social Science Data*. New York: Chapman and Hall/CRC.

Draper, N.R., & Smith, H. (1981). *Applied Regression Analysis* (2nd ed.). New York: John Wiley.

Edwards, A. (1948). Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika*, *13*(3), 185-187. https://doi.org/10.1007/BF02289261

Efron, B., & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall/CRC.

Johnson, R.A., & Wichern, D.W. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). New Jersey: Prentice Hall.

Kleinbaum, D.G., & Klein, M. (2002). *Logistic Regression: A Self-learning Text* (2nd ed.). New York: Springer.

Kuritz, S.J., Landis, J.R., & Koch, G.G. (1988). A general overview of Mantel-Haenszel methods: Applications and recent developments. *Annual Review of Public Health*, *9*(1), 123-160. https://doi.org/10.1146/annurev.pu.09.050188.001011

Maxwell, A.E. (1970). Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry*, *116*(535), 651-655. https://doi.org/10.1192/bjp.116.535.651

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153-157. https://doi.org/10.1007/BF02295996

Montgomery, D.C. (2001). *Design and Analysis of Experiments* (5th ed.). John and Wiley and Sons Inc.

Spearman, Ch. (1904). General intelligence objectively determined and measured. *American Journal of Psychology, 15*(2), 201-293. https://doi.org/10.2307/1412107

Stuart, A.A. (1955). A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika*, *42*(3/4), 412-416. https://doi.org/10.1093/biomet/42.3-4.412

Wooldridge, J.M. (2013). *Introductory Econometrics: A Modern Approach* (5th ed.). Mason, Ohio: South-Western, Cengage Learning.

Λεξικό Στατιστικών Όρων

A

acceptance region	περιοχή αποδοχής
acceptance sampling	δειγματοληψία αποδοχής
accessibility sampling	δειγματοληψία προσιτότητας
additive model	προσθετικό μοντέλο
Algorithm	αλγόριθμος
Alpha	άλφα
alternative hypothesis	εναλλακτική υπόθεση
analysis of covariance	ανάλυση συνδιακύμανσης
analysis of variance	ανάλυση διακύμανσης
antithetic random variables	αντίθετες ή αντιθετικές τυχαίες μεταβλητές
Approximation	προσέγγιση
area sampling	δειγματοληψία κατά περιοχές
arithmetic mean	αριθμητικός μέσος
Assumption	υπόθεση
Asymptotic	ασυμπτωτικός
Autocorrelation	αυτοσυσχέτιση
autocorrelation function	συνάρτηση αυτοσυσχέτισης
Average	μέσος όρος
Axis	άξονας

B

backward elimination procedure	μέθοδος αποκλεισμού μεταβλητών
bar chart	ραβδόγραμμα
Bayes factor	παράγοντας του Bayes
Bayes' theorem	θεώρημα του Bayes
Bayesian inference	συμπερασματολογία κατά Bayes
Bayesian information criterion	κριτήριο πληροφορίας του Bayes
Bayesian statistics	μπεϋζιανή στατιστική ή στατιστική κατά Bayes
beta	βήτα
beta distribution	κατανομή βήτα
beta function	συνάρτηση βήτα
beta-binomial distribution	κατανομή βήτα-διωνυμική
Bias	μεροληψία
bimodal distribution	δικόρυφη κατανομή
binary data	δυαδικά δεδομένα
binomial coefficient	διωνυμικός συντελεστής
binomial distribution	διωνυμική κατανομή
binomial test	διωνυμικός έλεγχος
Biostatistics	βιοστατιστική
bivariate distribution	διμεταβλητή κατανομή
box plot	διάγραμμα πλαισίου απολήξεων ή θηκόγραμμα
branching processes	κλαδωτές ανελίξεις
Buffon's needle	βελόνα του Buffon

C

canonical correlation analysis	ανάλυση κανονικών συσχετίσεων
capability index	δείκτης ικανότητας
capture-recapture methods	μέθοδοι σύλληψης και επανασύλληψης
categorical variable	κατηγορική μεταβλητή
categorical data analysis	ανάλυση κατηγορικών δεδομένων
causality	αιτιότητα
censored data	λογοκριμένα δεδομένα
census	απογραφή
Central Limit Theorem	Κεντρικό Οριακό Θεώρημα
characteristic function	χαρακτηριστική συνάρτηση
chi-square distribution	χι-τετράγωνο κατανομή
classification	κατάταξη
clinical trials	κλινικές δοκιμές
cluster analysis	ανάλυση σε ομάδες
clustered bar chart	ομαδοποιημένο ραβδόγραμμα
clustered sampling	δειγματοληψία κατά ομάδες
coefficient	συντελεστής
coefficient of determination	συντελεστής προσδιορισμού
coefficient of variation	συντελεστής μεταβλητότητας
cohort studies	μελέτες κοόρτης
collinearity	συγγραμμικότητα
compound Poisson process	σύνθετη διαδικασία Poisson
concordant pair	αρμονικό ζεύγος
conditional distribution	δεσμευμένη κατανομή
conditional expectation	δεσμευμένη αναμονή
conditional probability	δεσμευμένη πιθανότητα
confidence band	ζώνη εμπιστοσύνης
confidence interval	διάστημα εμπιστοσύνης
confirmatory factor analysis	επιβεβαιωτική παραγοντική ανάλυση
confounding factor	συγχυτικός παράγοντας
conjugate distribution	συζυγής κατανομή
consistency	συνέπεια
contingency coefficient	συντελεστής συνάφειας
contingency table	πίνακας συνάφειας
continuity correction	διόρθωση συνέχειας
continuous distribution	συνεχής κατανομή
continuous random variable	συνεχής τυχαία μεταβλητή
continuous variable	συνεχής μεταβλητή
contrasts	διαφορές
control chart	διάγραμμα ελέγχου
control-cases studies	μελέτες μαρτύρων-ασθενών
convolution	συνέλιξη
correlation	συσχέτιση
correlation coefficient	συντελεστής συσχέτισης
correlation matrix	πίνακας συσχετίσεων
correlogram	κορελόγραμμα
correspondence analysis	ανάλυση αντιστοιχιών
covariance	συνδιακύμανση
covariate	συμμεταβλητή

coverage probability	πιθανότητα κάλυψης
credibility interval	διάστημα αξιοπιστίας
credibility region	περιοχή αξιοπιστίας
critical value	κριτική τιμή
cross-sectional studies	διατμηματικές μελέτες
cumulative distribution function	αθροιστική συνάρτηση κατανομής
curve	καμπύλη

D

Data	δεδομένα
data analysis	ανάλυση δεδομένων
data mining	εξόρυξη δεδομένων
Deciles	δεκατημόρια
degree of association	βαθμός συνάφειας ή σύνδεσης
degrees of freedom	βαθμοί ελευθερίας
Demography	δημογραφία
Dendrogram	δενδρόγραμμα
dependent variable	εξαρτημένη μεταβλητή
descriptive statistics	περιγραφική στατιστική
design matrix	πίνακας σχεδιασμού
design of experiments	σχεδιασμός πειραμάτων
diffusion process	πρότυπα διάχυσης
discordant pair	δυσαρμονικό ζεύγος
discrete distribution	διακριτή κατανομή
discrete random variable	διακριτή τυχαία μεταβλητή
discriminant analysis	διαχωριστική ή διακριτική ανάλυση
Dispersion	διασπορά
Distribution	κατανομή
dot plot	σημειόγραμμα
dummy variable	ψευδομεταβλητή

E

efficient estimator	αποτελεσματικός εκτιμητής
efficiency	αποτελεσματικότητα
epidemiology	επιδημιολογία
Ergodic Theorem	Εργοδικό Θεώρημα
ergodicity	εργοδικότητα
error	λάθος, σφάλμα
estimation	εκτίμηση
estimator	εκτιμητής
estimated value	εκτιμημένη τιμή
event	γεγονός
exact test	ακριβής έλεγχος
expectation	αναμονή
expected value	αναμενόμενη τιμή
experimental units	πειραματικές μονάδες
explanatory variable	επεξηγηματική μεταβλητή
exponential distribution	εκθετική κατανομή
exponential family	εκθετική οικογένεια

F

factor	παράγοντας
factor analysis	παραγοντική ανάλυση
factor loadings	επιβαρύνσεις των παραγόντων
factorial design	παραγοντικός σχεδιασμός
false negative case	ψευδής θετική περίπτωση
false positive case	ψευδής αρνητική περίπτωση
finite	πεπερασμένος
first quartile	πρώτο τεταρτημόριο
fixed effects	σταθερές επιδράσεις
fixed effects model	μοντέλο σταθερών επιδράσεων
follow-up studies	μελέτες παρακολούθησης
forecasting methods	μέθοδοι προβλέψεων
forward procedure	μέθοδος προοδευτικής προσθήκης μεταβλητών
frame	πλαίσιο
frequency	συχνότητα
frequency polygon	πολύγωνο συχνοτήτων

G

game theory	θεωρία παιγνίων
gamma distribution	κατανομή γάμμα
gamma function	συνάρτηση γάμμα
general linear model	γενικό γραμμικό μοντέλο
generalized linear models	γενικευμένα γραμμικά μοντέλα
geometric distribution	γεωμετρική κατανομή
geometric mean	γεωμετρικός μέσος
goodness of fit test	έλεγχος καλής προσαρμογής
graph	γράφημα

H

harmonic mean	αρμονικός μέσος
hazard function	συνάρτηση κινδύνου
hazard rate	ρυθμός κινδύνου
hazard rate function	συνάρτηση βαθμού κινδύνου
heterogeneity	ετερογένεια
heteroscedasticity	ετεροσκεδαστικότητα
hierarchical model	ιεραρχικό μοντέλο
highest posterior density	περιοχή υψίστης a-posteriori πυκνότητας
histogram	ιστόγραμμα
homogeneity	ομοιογένεια
homoscedasticity	ομοσκεδαστικότητα
hypergeometric distribution	υπεργεωμετρική κατανομή
hypothesis testing	έλεγχος υπόθεσης

I

independent variable	ανεξάρτητη μεταβλητή
index number	αριθμοδείκτης
inertia	αδράνεια
infinite	άπειρος

information	μέτρο πληροφορίας
information matrix	πίνακας πληροφορίας
informative prior distribution	πληροφοριακή prior κατανομή
interaction	αλληλεπίδραση
interquartile range	ενδοτεταρτημοριακό εύρος
interval estimation	εκτίμηση σε διάστημα
interval scale	κλίμακα διαστήματος
irreducible Markov chain	αδιαχώριστη αλυσίδα Markov

J

joint distribution	από κοινού κατανομή
judgemental or purposive sampling	δειγματοληψία κρίσης ή σκοπιμότητας

K

kurtosis	κύρτωση

L

lack of memory	έλλειψη μνήμης
latent variables	λανθάνουσες μεταβλητές
latin squares	λατινικά τετράγωνα
law of large numbers	νόμος των μεγάλων αριθμών
least squares estimators	εκτιμητές ελαχίστων τετραγώνων
leptokurtic distribution	λεπτόκυρτη κατανομή
linear model	γραμμικό μοντέλο
location parameter	παράμετρος θέσης
logistic model	λογιστικό μοντέλο
logistic regression	λογιστική παλινδρόμηση
logistics distribution	λογιστική κατανομή
loglinear model	λογαριθμικό μοντέλο
lognormal distribution	λογαριθμοκανονική κατανομή
longitudinal data analysis	ανάλυση διαμήκων δεδομένων
longitudinal studies	διαμήκεις μελέτες
loss function	συνάρτηση απώλειας

M

main effects	κύριες επιδράσεις
marginal distribution	περιθώρια κατανομή
market research	έρευνα αγοράς
Markov chains	αλυσίδες Markov ή Μαρκοβιανές αλυσίδες
maximum	μέγιστο
maximum likelihood	μέγιστη πιθανοφάνεια
mean	μέσος
mean absolute deviation	μέση απόλυτη απόκλιση
mean square error	μέσο τετραγωνικό σφάλμα
measure theory	θεωρία μέτρου
measures of association	μέτρα συνάφειας
median	διάμεσος
mesokurtic distribution	μεσόκυρτη κατανομή
minimum	ελάχιστο
missing values	ελλείπουσες τιμές

mixed effects model	μοντέλο μεικτών επιδράσεων
mode	κορυφή
model	υπόδειγμα
moment	ροπή
moment generating function	ροπογεννήτρια συνάρτηση
monotone regression	μονότονη παλινδρόμηση
mortality (death) rate	ρυθμός θνησιμότητας
moving average	κινητός μέσος
multicollinearity	πολυσυγγραμμικότητα
multidimensional scaling techniques	πολυδιάστατες τεχνικές κλιμακοποίησης
multilevel models	πολυεπίπεδα μοντέλα
multinomial distribution	πολυωνυμική κατανομή
multiple comparisons	πολλαπλοί έλεγχοι
multiple correspondence analysis	πολλαπλή ανάλυση αντιστοιχιών
multiple linear regression	πολλαπλή γραμμική παλινδρόμηση
multivariate	πολυμεταβλητός
multivariate analysis of variance	πολυμεταβλητή ανάλυση διακύμανσης
multivariate case	πολυμεταβλητή περίπτωση

N

negative binomial distribution	αρνητική διωνυμική κατανομή
negative confounder	αρνητικός συγχυτικός παράγοντας
negative predicted value	αρνητική προβλεπτική τιμή
nested models	φωλιασμένα μοντέλα
nominal scale	ονομαστική κλίμακα
nominal variable	ονομαστική μεταβλητή
non-informative prior distribution	μη πληροφοριακή prior κατανομή
non-linear model	μη-γραμμικό μοντέλο
nonparametric statistics	μη παραμετρική στατιστική
normal distribution	κανονική κατανομή
nuisance factor	ενοχλητικός παράγοντας
null hypothesis	μηδενική υπόθεση

O

observations	παρατηρήσεις
observed significance level (*p*-value)	παρατηρηθέν επίπεδο σημαντικότητας (*p*-τιμή)
odds ratio	λόγος συμπληρωματικών πιθανοτήτων ή κλάσμα λόγου πιθανοτήτων
one-sided test	μονόπλευρος έλεγχος
one-way ANOVA	ανάλυση διακύμανσης κατά έναν παράγοντα
operating characteristic curve	χαρακτηριστική λειτουργική καμπύλη
opinion poll	σφυγμομέτρηση κοινής γνώμης
ordered data	διατεταγμένα δεδομένα
ordinal scale	κλίμακα διάταξης
ordinary least squares method	μέθοδος ελαχίστων τετραγώνων
orthogonal contrasts	ορθογώνιες διαφορές
outliers	ακραίες τιμές
overparameterization	υπερπαραμετροποίηση

P

	παράμετρος

parameter	
parametric statistics	παραμετρική στατιστική
partial correlation coefficient	μερικός συντελεστής συσχέτισης
permutation	αναδιάταξη
pie chart	κυκλικό διάγραμμα ή διάγραμμα πίτας
pivotal quantity	αντιστρεπτή ποσότητα
platykurtic distribution	πλατύκυρτη κατανομή
plot	γράφημα
point estimation	σημειακή εκτίμηση
point process	σημειακή διαδικασία
population	πληθυσμός
population characteristic	χαρακτηριστικό του πληθυσμού
positive confounder	θετικός συγχυτικός παράγοντας
positive predicted value	θετική προβλεπτική τιμή
posterior or a-posteriori distribution	posterior κατανομή
predicted value	προβλεφθείσα τιμή
prediction interval	διάστημα πρόβλεψης
prevalence	επιπολασμός
principal component analysis	ανάλυση κύριων συνιστωσών
principal coordinate analysis	ανάλυση κύριων συντεταγμένων
prior or a-priori distribution	prior κατανομή
probability	πιθανότητα
probability density function	συνάρτηση πυκνότητας πιθανότητας
probability generating function	πιθανογεννήτρια συνάρτηση
probability sampling	δειγματοληψία κατά πιθανότητα
prospective studies	προοπτικές μελέτες

Q

qualitative	ποιοτικός
quantitative	ποσοτικός
queues	ουρές
quota sampling	δειγματοληψία με προκαθορισμένα ποσοστά

R

random effects	τυχαίες επιδράσεις
random effects model	μοντέλο τυχαίων επιδράσεων
random factor	τυχαίος παράγοντας
random processes	τυχαίες διαδικασίες
random sample	τυχαίο δείγμα
random variable	τυχαία μεταβλητή
random walk	τυχαίος περίπατος
randomization	τυχαιοποίηση
randomized complete block design	τυχαιοποιημένοι πλήρως σχεδιασμοί κατά μπλοκ
range	εύρος
ranks	τάξεις μεγέθους
ratio scale	κλίμακα λόγου
recurrent state	επαναληπτική κατάσταση
regression analysis	ανάλυση παλινδρόμησης
regression coefficients	συντελεστές παλινδρόμησης
regressors	παλινδρομητές

rejection region	περιοχή απόρριψης
rejection sampling	δειγματοληψία απόρριψης
relative risk	σχετικός κίνδυνος
reliability	αξιοπιστία
reliability coefficient	συντελεστής αξιοπιστίας
reliability function	συνάρτηση αξιοπιστίας
renewal process	ανανεωτική διαδικασία
renewal theory	θεωρία ανανέωσης
repeated measures	επαναλαμβανόμενες μετρήσεις
replication	επαναληψιμότητα
residuals	κατάλοιπα
response surface	επιφάνεια απόκρισης
response variable	απαντητική μεταβλητή
retrospective studies	αναδρομικές μελέτες
ridge regression	αμφικλινής παλινδρόμηση
robust regression	εύρωστη παλινδρόμηση
runs test	έλεγχος ροών

S

sample	δείγμα
sample surveys	δειγματοληπτικές έρευνες
sampling distribution	δειγματική κατανομή
sampling frame	δειγματοληπτικό πλαίσιο
sampling mean	δειγματικός μέσος
sampling techniques	δειγματοληπτικές τεχνικές
sampling theory	θεωρία δειγματοληψίας
sampling units	δειγματοληπτικές μονάδες
saturated (full) model	κορεσμένο μοντέλο
scale	κλίμακα
scale parameter	παράμετρος κλίμακας
scatter plot	διάγραμμα διασποράς
seasonality	εποχικότητα
semiparametric	ημιπαραμετρικός
sensitivity analysis	ανάλυση ευαισθησίας
serial correlation coefficient	σειριακός συντελεστής συσχέτισης
shape parameter	παράμετρος σχήματος
sign test	προσημικός έλεγχος
significance level	επίπεδο σημαντικότητας
simple linear regression	απλή γραμμική παλινδρόμηση
simple random sampling	απλή τυχαία δειγματοληψία
simulation	προσομοίωση
size effect	επίδραση μεγέθους
skewed distribution	ασύμμετρη κατανομή
skewness	ασυμμετρία
specification limits	όρια προδιαγραφών
specificity	ειδικότητα
sphericity	σφαιρικότητα
split-half reliability	αξιοπιστία ημίκλαστου
stable process	σταθερή διεργασία
standard deviation	τυπική απόκλιση
standard error	τυπικό σφάλμα

standardization	τυποποίηση
standardized values	τυποποιημένες τιμές
state space	χώρος καταστάσεων
stationary distribution	στάσιμη κατανομή
stationary process	στάσιμη διαδικασία
statistical	στατιστικός (επίθετο)
statistical inference	στατιστική συμπερασματολογία
statistically significant	στατιστικά σημαντικός
statistician	στατιστικός (ο/η, ιδιότητα)
statistics	στατιστική
stem-and-leaf plot	διάγραμμα μίσχου-φύλλου
stepwise regression	βηματική παλινδρόμηση
stochastic models	στοχαστικά μοντέλα
stochastic processes	στοχαστικές ανελίξεις
stratified analysis	στρωματοποιημένη ανάλυση
stratified randomization	στρωματοποιημένη τυχαιοποίηση
stratified sampling	στρωματοποιημένη δειγματοληψία
study population	υπό μελέτη πληθυσμός
subjective probability	υποκειμενική πιθανότητα
sufficiency	επάρκεια
sum	άθροισμα
sum of products	άθροισμα γινομένων
sum of squares	άθροισμα τετραγώνων
survival analysis	ανάλυση επιβίωσης
survival function	συνάρτηση επιβίωσης
symmetric distribution	συμμετρική κατανομή
systematic sampling	συστηματική δειγματοληψία

T

target population	αντικειμενικός πληθυσμός
third quartile	τρίτο τεταρτημόριο
time series	χρονολογικές σειρές
transformation	μετασχηματισμός
transient state	μεταβατική κατάσταση
transition matrix	πίνακας μετάβασης
transition probability	πιθανότητα μετάβασης
trend	τάση
trimmed mean	περικομμένος μέσος
two-sided test	αμφίπλευρος έλεγχος
two-way ANOVA	ANOVA κατά δύο παράγοντες

U

unbiased estimator	αμερόληπτος εκτιμητής
uniform distribution	ομοιόμορφη κατανομή
unimodal distribution	μονοκόρυφη κατανομή
univariate case	μονομεταβλητή περίπτωση
universe	ολότητα
unstable process	μη σταθερή διεργασία

V

variability	μεταβλητότητα
variable	μεταβλητή
variance	διακύμανση
variance-covariance matrix	πίνακας διακύμανσης-συνδιακύμανσης

W

waiting time	χρόνος αναμονής
weighted least squares method	μέθοδος σταθμισμένων ελαχίστων τετραγώνων
weighted mean	σταθμισμένος μέσος

Ο σκοπός του βιβλίου είναι να βοηθήσει φοιτητές, κυρίως προπτυχιακούς, σε θέματα επιλογής και διεξαγωγής στατιστικών αναλύσεων με τη χρήση δύο στατιστικών προγραμμάτων, του SPSS και του EVIEWS. Καλύπτεται η απαραίτητη θεωρία και παρέχεται βήμα-βήμα ο τρόπος υλοποίησης διάφορων στατιστικών αναλύσεων. Το υλικό καλύπτει με έναν περιεκτικό τρόπο εισαγωγικές μεν, θεμελιώδεις δε, αρχές και έννοιες της στατιστικής, οι οποίες θα είναι χρήσιμες για έναν προπτυχιακό φοιτητή. Ξεκινώντας από μια πολύ σύντομη εισαγωγή στη στατιστική, το βιβλίο συνεχίζει με μία περιγραφή των προαναφερθέντων στατιστικών προγραμμάτων. Έπειτα, εν συντομία παρουσιάζονται αρκετά στατιστικά εργαλεία, στη θεωρία, και ο τρόπος υλοποίησής τους στα δύο προγράμματα. Το φάσμα των εργαλείων είναι αρκετά ευρύ για φοιτητές που θέλουν να εισαχθούν στην επιστήμη της στατιστικής, αλλά και της οικονομετρίας. Για παράδειγμα, παρουσιάζονται διαγραμματική απεικόνιση μεταβλητών, παραμετρικοί και μη παραμετρικοί έλεγχοι υποθέσεων (κανονικότητας, συσχέτισης, εξάρτησης, μέσων όρων κλπ). Έμφαση δίνεται στην παλινδρόμηση, απλή και πολλαπλή, αλλά αναφέρονται και πιο προηγμένες τεχνικές, όπως η λογιστική παλινδρόμηση και η διαχωριστική ανάλυση. Για τη διευκόλυνση και εξοικείωση των φοιτητών με την αγγλική ορολογία παρέχεται ένα αγγλο-ελληνικό λεξικό στατιστικών όρων στο τέλος του βιβλίου.

Το παρόν σύγγραμμα δημιουργήθηκε στο πλαίσιο του Έργου ΚΑΛΛΙΠΟΣ+	
Χρηματοδότης	Υπουργείο Παιδείας και Θρησκευμάτων, Προγράμματα ΠΔΕ, ΕΠΑ 2020-2025
Φορέας υλοποίησης	ΕΛΚΕ ΕΜΠ
Φορέας λειτουργίας	ΣΕΑΒ/Παράρτημα ΕΜΠ/Μονάδα Εκδόσεων
Διάρκεια 2ης Φάσης	2020-2023
Σκοπός	Η δημιουργία ακαδημαϊκών ψηφιακών συγγραμμάτων ανοικτής πρόσβασης (περισσότερων από 700) • Προπτυχιακών και μεταπτυχιακών εγχειριδίων • Μονογραφιών • Μεταφράσεων ανοικτών textbooks • Βιβλιογραφικών Οδηγών
Επιστημονικά Υπεύθυνος	Νικόλαος Μήτρου, Καθηγητής ΣΗΜΜΥ ΕΜΠ

ISBN: 978-618-5667-05-4	**DOI:** http://dx.doi.org/10.57713/kallipos-68

Το παρόν σύγγραμμα χρηματοδοτήθηκε από το Πρόγραμμα Δημοσίων Επενδύσεων του Υπουργείου Παιδείας.