ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
΄΄ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΕΠΙΧΕΙΡΗΣΙΑΚΗ ΕΡΕΥΝΑ΄΄

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανδρομάχη Σκουφά

Επιβλέπουσα: Φ. Κολυβά-Μαχαίρα
Επικ.Καθηγήτρια Α.Π.Θ

Θεσσαλονίκη Νοέμβριος 2008

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
΄΄ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΕΠΙΧΕΙΡΗΣΙΑΚΗ ΕΡΕΥΝΑ΄΄

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανδρομάχη Σκουφά

Επιβλέπουσα: Φ. Κολυβά-Μαχαίρα
Επικ.Καθηγήτρια Α.Π.Θ

Θεσσαλονίκη Νοέμβριος 2008

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
΄΄ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΕΠΙΧΕΙΡΗΣΙΑΚΗ ΕΡΕΥΝΑ΄΄

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανδρομάχη Σκουφά

Επιβλέπουσα: Φ. Κολυβά-Μαχαίρα
Επικ.Καθηγήτρια Α.Π.Θ

Εγκρίθηκε από την τριμελή επιτροπή τον Νοέμβριο 2008

.....	.....	.....
Φ. Κολυβά-Μαχαίρα	Μωυσιάδης Πολυχρόνης	Μπόρα-Σέντα Ευθυμία
Επικ.Καθηγήτρια Α.Π.Θ	Καθηγητής Α.Π.Θ	Επικ.Καθηγήτρια Α.Π.Θ

Θεσσαλονίκη Νοέμβριος 2008

«Σε εκείνους που οφείλω τα πάντα,
στους γονείς μου
και
στον Γιώργο *»*

ΠΕΡΙΕΧΟΜΕΝΑ

1. Εισαγωγή

Η Λογιστική παλινδρόμηση είναι μία μέθοδος πολυπαραγοντικής στατιστικής ανάλυσης που χρησιμοποιεί ένα σύνολο ανεξαρτήτων μεταβλητών για την διερεύνηση της κίνησης μιας κατηγορικής εξαρτημένης μεταβλητής.
Η Λογιστική παλινδρόμηση είναι χρήσιμη σε καταστάσεις στις οποίες επιθυμούμε την πρόβλεψη ύπαρξης ή της απουσίας ενός χαρακτηριστικού ή ενός συμβάντος. Η πρόβλεψη αυτή βασίζεται στην κατασκευή ενός γραμμικού μοντέλου και συγκεκριμένα στον προσδιορισμό των τιμών που παίρνουν οι συντελεστές ενός συνόλου ανεξαρτήτων μεταβλητών που χρησιμοποιούνται ως μεταβλητές πρόβλεψης.
Η πρόβλεψη κατα πόσο θα συμβεί η όχι κάποιο γεγονός και η αναγνώριση των μεταβλητών των αναγκαιών για την πρόβλεψη είναι θέμα συχνού προβληματισμού.
Για την αντιμετώπιση τέτοιων προβλημάτων το υπόδειγμα της γραμμικής παλινδρόμησης δεν είναι κατάλληλο για την εκτίμηση των τιμών της εξαρτημένης μεταβλητής από τις τιμές των ανεξάρτητων.
Σε μια τέτοια περίπτωση χρησιμοποιώντας την τιμή 1 για το ενδεχόμενο της ‘’επιτυχίας’’ (την πραγματοποίηση,δηλαδή,του γεγονότος) και την τιμή 0 για το ενδεχόμενο της ‘’αποτυχίας’’, ο υπολογισμός της μέσης τιμής της εξαρτημένης δίτιμης μεταβλητής, ουσιαστικά ορίζει την αναλογία p, των επιτυχιών στο σύνολο των δυνατών τιμών της.
Όπως με την βοήθεια του μοντέλου της γραμμικής παλινδρόμησης, εκτιμάται η μέση τιμή της συνεχούς μεταβλητής Y, για ένα συγκεκριμένο σύνολο τιμών των ανεξάρτητων μεταβλητών , έτσι μπορεί να εκτιμηθεί και η πιθανότητα p της επιτυχίας μιας δίτιμης μεταβλητής (η μέση τιμή της δηλαδή) για ένα σύνολο τιμών μιας ή περισσότερων ανεξάρτητων μεταβλητών.
Η τεχνική που χρησιμοποιείται σε αυτές τις περιπτώσεις ονομάζεται λογιστική παλινδρόμηση(logistic regression).
Εκτός από την πρόβλεψη ένα μοντέλο λογιστικής παλινδρόμησης δίνει την δυνατότητα να εκτιμήσουμε την επίδραση κάθε ανεξάρτητης μεταβλητής στην διαμόρφωση των τιμών της εξαρτημένης μεταβλητής.
Στην λογιστική παλινδρόμηση, σε αντίθεση με την πολλαπλή γραμμική παλινδρόμηση είναι δυνατόν να χρησιμοποιηθουν ως εξαρτημένες μεταβλητές εκτός από αναλογικές αριθμητικές μεταβλητές(ratio scales) και κατηγορικές μεταβλητές(nominal scale).
Η πιο διαδεδομένη βιβλιογραφικά έκφραση της λογιστικής παλινδρόμησης είναι:

$$Ln(odds)=a+b_1x_1+b_2x_2+\ldots..+b_kx_k$$

Το δεξί μέλος της εξίσωσης δημιουργείται από ένα γραμμικό συνδυασμό των ανεξάρτητών μεταβλητών που συμμετέχουν στο μοντέλο παλινδρόμησης. Το αριστερό μέλος περιέχει τις τιμές της εξαρτημένης μεταβλητής με την μορφή του λογαρίθμου των odds(απόδοση), δηλαδή του λογαρίθμου της σχέσης:

$$odds=\frac{prob}{(1-prob)}$$

Τα odds εναλλακτικά ονομάζονται logit και ο όρος prob εκφράζει την πιθανότητα του συμβάντος του γεγονότος.

Οι συντελεστές των ανεξαρτήτων μεταβλητών στην εξίσωση λογιστικής παλινδρόμησης εκτιμούνται με βάση της μεθόδου Μέγιστης Πιθανοφάνειας. Σύμφωνα με την μέθοδο αυτή η τιμή των συντελεστών των ανεξαρτήτων μεταβλητών είναι αυτή που κάνει τις παρατηρηθείσες τιμές της εξαρτημένης μεταβλητής πιο πιθανές, βάσει του συνόλου των ανεξαρτήτων μεταβλητών.

2. ***Εκτιμητές μέγιστής πιθανοφάνειας***

Ας είναι $y=(Y_1,.....,Y_n)$ ένα τυχαίο ανεξάρτητο διάνυσμα με συνάρτηση πυκνότητας πιθανότητας

$f_i(y_i;\boldsymbol{\theta})$, $\theta\in\Omega\subset R^r$ και $\omega\subset\Omega$.

2.1 Η log-likelihood συνάρτηση

Σαν συνάρτηση πιθανοφάνειας (Likelihood function) ορίζουμε την από κοινού κατανομή των τυχαίων μεταβλητών $Y_1,.....,Y_n$ όταν η κατανομή θεωρείται συνάρτηση της παραμέτρου **θ**, δηλαδή:

$$f(\mathbf{y};\boldsymbol{\theta})=\prod_{i}^{n}[f_i(y_i;\theta)=L(\theta;y)$$

Συχνά χρησιμοποιούμε τον λογάριθμο της συνάρτησης πιθανοφάνειας και την ονομάζουμε log-likelihood συνάρτηση η οποία είναι:

$$\text{Log}=\sum_{i}^{n}\log f_i(y_i;\boldsymbol{\theta})$$

Ένας απλός τρόπος για να εκτιμήσουμε την παραμέτρο **θ** είναι να μεγιστοποιήσουμε την συνάρτηση πιθανοφάνειας (maximize) επιλέγοντας τις πιο πιθανοφανείς τιμές των παραμέτρων προκειμένου να οδηγήσουν στα παρατηρούμενα αποτελέσματα.

Τελικά ορίζουμε ως **εκτιμητή μέγιστης πιθανοφάνειας** $\hat{\theta}$ της παραμέτρου **θ** την:

Log≥ Log για κάθε **θ**

2.1.1 Το διάνυσμα των score

Η πρώτη παράγωγος της log-likelihood συνάρτησης ονομάζεται Fisher's score συνάρτηση και συμβολίζεται ως:

$$u(\boldsymbol{\theta})=\frac{\partial \mathbf{Log}\, L\left(\boldsymbol{\theta};\mathbf{y}\right)}{\partial \boldsymbol{\theta}}$$

Αν η log-likelihood συνάρτηση είναι κοίλη μπορούμε να βρούμε τους εκτιμητές μέγιστης πιθανοφάνειας μηδενίζοντας το διάνυσμα των score, λύνοντας δηλαδή τις εξισώσεις:

$$u(\hat{\boldsymbol{\theta}})=\mathbf{0}$$

2.1.2. Πίνακας Πληροφορίας(Information Matrix)

Το διάνυσμα των score έχει κάποιες ενδιαφέρουσες στατιστικές ιδιότητες. Η μέση τιμή του είναι ίση με μηδέν, δηλαδή:

$$E[u(\boldsymbol{\theta})]=0$$

και ο πίνακάς διακυμάνσεων-συνδιακυμάνσεων δίνεται από τον Information Matrix:

$$var[u(\boldsymbol{\theta})]= E[u(\boldsymbol{\theta})u'(\boldsymbol{\theta})]=I(\boldsymbol{\theta})$$

Κάτω από συνθήκες ομαλότητας ο πίνακας πληροφορίας(information matrix) ορίζεται και ως η αρνητική μέση τιμή της δεύτερης παραγώγου της log- likehood συνάρτησης, δηλαδή ως:

$$I(\boldsymbol{\theta})=-E\left[\frac{\partial^2 \mathbf{Log}\, L\left(\boldsymbol{\theta};\mathbf{y}\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right]$$

2.1.3 Newton-Raphson και Fisher Scoring

Ο υπολογισμός των εκτιμητών μέγιστής πιθανοφάνειας συχνά απαιτεί επαναληπτικές διαδικασίες. Ας θεωρήσουμε ότι θέλουμε να εκτιμήσουμε με το διάνυσμα των score τους εκτιμητές μέγιστης πιθανοφάνειας $\hat{\boldsymbol{\theta}}$ γύρω από μία αρχική τιμή $\boldsymbol{\theta}_0$ χρησιμοποιώντας σειρά taylor πρώτης τάξης, δηλαδή:

$$u(\hat{\boldsymbol{\theta}})]\approx u(\boldsymbol{\theta}_0)+\frac{\partial \mathbf{u}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$$

Ας είναι Η ο εσσιανός πίνακας ή αλλιώς ο πίνακας των δεύτερων παραγώγων της log-likelihood συνάρτησης:

$$H(\boldsymbol{\theta})=\frac{\partial^2 logL}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}=\frac{\partial \mathbf{u}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

Θέτωντας $u(\hat{\boldsymbol{\theta}})=\mathbf{0}$ και λύνοντας ως προς $\hat{\boldsymbol{\theta}}$ έχουμε:

$$\bar{\boldsymbol{\theta}}=\boldsymbol{\theta_0}-H^{-1}(\boldsymbol{\theta_0})\, u(\boldsymbol{\theta}_0)\ (1)$$

αυτό το αποτέλεσμα αποτελεί την βάση μιας επαναληπτικής διαδικασίας προσσεγγιστικής λύσης για τους εκτιμητές μέγιστης πιθανοφάνειας γνωστή ως τεχνική Newton-Raphson. Δοσμένης μιας αρχικής τιμής χρησιμοποιούμε την (1) εξίσωση για να λάβουμε μια βελτιωμένη εκτίμηση και επαναλαμβάνουμε μέχρις ώτου οι διαφορές ανάμεσα στις διαδοχικές εκτιμήσεις να τείνουν στο μηδεν.
Μία εναλλακτική διαδικασία που προτάθηκε από τον Fisher είναι να αντικαταστήσουμε τον εσσιανό πίνακα με την μέση τίμή του δηλαδή με τον information matrix I:

$$\bar{\boldsymbol{\theta}}=\boldsymbol{\theta_0}-I^{-1}(\boldsymbol{\theta_0})\, u(\boldsymbol{\theta}_0)$$

γνωστό ως Fisher Scoring.

2.2 Η κατανομή της τυχαίας μεταβλητής $-2\log\lambda_n$

Θεώρημα

Υποθέτουμε ότι ισχύουν οι συνθήκες ομαλότητας, ότι το μέγεθος n του δείγματος είναι αρκετά μεγάλο και ότι υπάρχει μονοσήμαντα ορισμένος εκτιμητής μέγιστης πιθανοφάνειας $\bar{\theta}_n$ του θ. Τότε για τον έλεγχο της υπόθεσης H_0: θЄω=$\{\theta_0\}$, η ασυμπτωτική κατανομή της στατιστικής

συνάρτησης $-2\log\lambda_n$ είναι η , όταν η H_0 είναι αληθής, δηλαδή

$$P(-2\log\lambda_n\leq x)\xrightarrow[n\to\infty]{} P(\chi_1^2\leq x)$$

Επειδή κατά την απόδειξη θα χρησιμοποιήσουμε τις συνθήκες ομαλότητας τις αναφέρουμε παρακάτω:

Συνθήκες ομαλότητας

1) Για όλα τα x Є R-N, όπου P_θ (χ Є N)=0 και για κάθε θ Є Ω υπάρχουν οι παράγωγοι

$\frac{\partial}{\partial\theta}\log f(x;\theta)$, $\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)$, $\frac{\partial^3}{\partial\theta^3}\log f(x;\theta)$, $\theta \in \Omega$

2) Υπάρχει μετρήσιμη συνάρτηση $H:R\to R^+$

α) $\left|\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\right| < H(x)$ για κάθε $\theta \in \Omega$ και

β) $E_\theta H(x) < M < \infty$ και το M είναι ανεξάρτητο του θ

3) $E_\theta[\frac{\partial}{\partial\theta}\log f(x;\theta)]=0$ για κάθε $\theta \in \Omega$

4) Για κάθε $\theta \in \Omega$ ισχύει:

α) $E_\theta[\frac{\partial}{\partial\theta}\log f(x;\theta)]^2 = -E_\theta[\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)] = I(\theta)$ και

β) $0 < I(\theta) < \infty$

όπου $I(\theta)$ ο πίνακας οληροφορίας.

Απόδειξη Θεωρήματος

Αν θέσουμε $\psi(x;\theta)= \log f(x;\theta)$,

$\frac{\partial}{\partial\theta}$

$\Psi'(x;\theta)=$ $\psi(x;\theta)=\varphi(x;\theta)$,

$\Psi''(x;\theta)=\varphi'(x;\theta)$ και $\psi'''(x;\theta)=\varphi''(x;\theta)$.

Αναπτύσσουμε σε σειρά Taylor τη συνάρτηση $\psi(x;\theta)$ στην περιοχή του θ, με όρους μέχρι τρίτης τάξης. Έτσι έχουμε:

$\psi(x;\theta_0)= \psi(x;\theta)+(\theta_0-\theta)\,\psi'(x;\theta)+ \quad (\theta_0-\theta)^2\,\psi'' (x;\theta)+ \quad (\theta_0-\theta)^3\,\psi'''(x;\theta^*)=$

$= \psi(x;\theta)+(\theta_0-\theta)\,\varphi(x;\theta))+ \quad (\theta_0-\theta)^2\,\varphi'(x;\theta))+ \quad (\theta_0-\theta)^3\,\varphi''(x;\theta^*)$

Όπου θ^* κατάλληλη τιμή του θ μεταξύ θ και θ_0.

Αν αθροίσουμε τις n παραπάνω σχέσεις για τις δάφορες τιμές του x_i, i=1,2,….,n, θα έχουμε:

$$\sum_{i=1}^{n}\psi(x_i;\theta_0) = \sum_{i=1}^{n}\psi(x_i;\theta) + (\theta_0-\theta)+ \quad (\theta_0-\theta)^2\sum_{i=1}^{n}\varphi'(x_i;\theta) +$$

$$+ \quad (\theta_0-\theta)^3\sum_{i=1}^{n}\varphi''(x_i;\theta^*) \quad (\text{σχέση } 1)$$

Αλλά από τη συνθήκη ομαλότητας (2) είναι:

$|\varphi''(x_i;\theta^*)| < H(x_i)$,

άρα υπάρχει κατάλληλος λ_n^* :

$|\lambda_n^*| \le 1$ ώστε $\Sigma| | = \lambda_n^* \Sigma\, H(x_i)$, και η σχέση 1 γίνεται:

$$\sum_{i=1}^{n}\psi(x_i;\theta_0) = \sum_{i=1}^{n}\psi(x_i;\theta) + (\theta_0-\theta)+ \quad (\theta_0-\theta)^2\sum_{i=1}^{n}\varphi'(x_i;\theta) +$$

$$+ \quad (\theta_0-\theta)^3\lambda_n^*\,\Sigma\, H(x_i) \quad (\text{σχέση2})$$

Αν είναι $\hat{\theta}_n$ ο εκτιμητής μέγιστης πιθανοφάνειας για το θ, τότε $\hat{\theta}_n \xrightarrow{P_\theta} \theta$. Άρα για κάθε ε>0 υπάρχει N=N(ε) ώστε για κάθε n≥N $P(|\theta_n - \theta| < \varepsilon) > 1-\varepsilon$.

Αν στη σχέση2 αντικαταστήσουμε το θ με το $\hat{\theta}_n$ έχουμε:

$$\sum_{i=1}^{n}\psi(x_i;\theta_0) = \sum_{i=1}^{n}\psi(x_i;\hat{\theta}_n) + (\theta_0-\hat{\theta}_n)^2\sum_{i=1}^{n}\varphi'(x_i;\hat{\theta}_n) +$$

$$(\theta_0-\hat{\theta}_n)^3\,\lambda_n\sum_{i=1}^{n}H(xi)$$

Διότι $\sum_{i=1}^{n}\varphi(x_i;\hat{\theta}_n)=0$, αφού $\hat{\theta}_n$ είναι εκτιμητής μέγιστης πιθανοφάνειας του θ και λ_n

η αντίστοιχη τιμή του λ_n^* όταν θ=$\hat{\theta}_n$.

Επειδή $\psi(x_i;\theta)$=logf(x_i;θ) προκύπτει ότι

$$-2\log\lambda_n=-2[\log L(\mathbf{X}/\theta_0)-\log L(\mathbf{x}/\hat{\theta}_n)]=-2[\sum_{i=1}^{n}\psi(x_i;\theta_0) - \Sigma_{\perp}(i=1)^{1}n\,\psi(x_i;\theta^{*}_{i}n)\,]$$

$$=-[\sqrt{n}(\hat{\theta}_n-\theta_0)]\frac{1}{n}\sum_{i=1}^{n}\varphi'(x_i;\hat{\theta}_n) + [\sqrt{n}(\hat{\theta}_n-\theta_0)]^2(\hat{\theta}_n-\theta_0)](\text{σχέση3})$$

Επειδή όμως $\sqrt{n}(\hat{\theta}_n-\theta_0)$

$\sqrt{n}(\hat{\theta}_n-\theta_0)$ όπου X~N(0,1).

Τότε όμως $[\sqrt{n}(\hat{\theta}_n-\theta_0)]^2$ όπου η $X^2\sim X_1^2$, άρα

$[\sqrt{n}(\hat{\theta}_n-\theta_0)]^2$(σχέση4)

Αν δεχτούμε ότι) (σχέση5)

Τότε από τα θεωρήματα σύγκλισης προκύπτει ότι

$$-[\sqrt{n}(\hat{\theta}_n-\theta_0)]^2\frac{1}{n}\sum_{i=1}^{n}\varphi'(x_i;\hat{\theta}_n)\xrightarrow{\kappa.N}X_1^2 \qquad (\text{σχέση6})$$

Επειδή ακόμη $\hat{\theta}_n$ - $\theta_0 \xrightarrow{P_{\theta_0}} 0$,

H(x) και $|\tilde{\lambda}_n|<1$ η σχέση4 συνεπάγεται ότι

$[\sqrt{n}(\hat{\theta}_n - \theta_0)]^2(\hat{\theta}_n - \theta_0)]$ $\tilde{\lambda}_n \frac{1}{n}\sum_{i=1}^{n} H(xi) \xrightarrow{\kappa.N}$

0, άρα

$[\sqrt{n}(\hat{\theta}_n - \theta_0)]^2(\hat{\theta}_n - \theta_0)]$

0 (σχέση7)

Από τις σχέσεις 3,6,7 προκύπτει ότι $P_{\theta_0}(-2\log\lambda_n \le x) \xrightarrow{n} P(\mathbf{X}_1^2)$,$X \epsilon R$

Για να ολοκληρωθεί η απόδειξη πρέπει να αποδείξουμε ότι ο ισχυρισμός της σχέσης 5 είναι αληθής.

Αν αναπτύξουμε σε σειρά taylor τη συνάρτηση $\varphi'(x;\hat{\theta}_n)$ έχουμε

$$\varphi'(x;\hat{\theta}_n) = \varphi'(x;\hat{\theta}_0) + (\hat{\theta}_n - \theta_0)$$

Όπου θ^* κατάλληλη τιμή μεταξύ θ_0 και $\hat{\theta}_n$.

Αθροίζοντας έχουμε

$$\frac{1}{n}\sum_{i=1}^{n}\varphi'(x_i;\tilde{\theta}_n)$$ -

$$\frac{1}{n}\sum_{i=1}^{n}\varphi'(x_i;\theta_o)$$

$=(\theta_n-\theta_0)$

Όπου λ_n^* κατάλληλη τιμή $0<|\lambda_n^*|<1$

Επειδή όμως όπως προαναφέρθηκε $\tilde{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0$ θα είναι

$-I(\theta_0)$

$0<|\lambda_n^*|<1$ και

Οι παραπάνω συγκλισεις συνεπάγονται την αλήθεια της σχέσης5.

Το παραπάνω θεώρημα γενικεύεται δηλαδή αν το Ω είναι ένα r-διάστατο ανοικτό υποσύνολο του R^r r>1, και η υπόθεση H_0:**θ**Єω, όπου ω ένα m-διάστατο θποσύνολο του Ω (m<r), τότε εφόσον ισχύουν οι συνθληκες ομαλότητας και η H_0 είναι αληθής, η κατανομή της στατιστικής

συνάρτησης $-2\log\lambda_n$ είναι η δηλαδή

$P_\theta(-2\log\lambda_n\leq\chi)\rightarrow P(\mathbf{X^2_{r-m}}\leq x), x\epsilon R$

2.3 Έλεγχοι Υποθέσεων

2.3.1 Wald test

Κάτω από την υπόθεση ότι ισχύουν οι συνθήκες ομαλότητας, οι εκτιμητές μέγιστης πιθανοφάνειας $\hat{\theta}$ για μεγάλα δείγματά ακολουθούν την κανονική κατανομή με μέση τιμή την πραγματική τιμή του θ και με πίνακα διακυνάνσεων ίσο με τον αντίστοφο του πίνακα πληροφορίας(information matrix**)** I^{-1}.

$\hat{\theta} \sim N_p(\theta, I^{-1}(\theta))$

Υπό την συνθήκη

$H_0: \theta=\theta_0,$

Η τετραγωνική μορφή:

$W=(\hat{\theta} -\theta_0)' var^{-1}(\hat{\theta})(\hat{\theta} -\theta_0)$

Για μεγάλα δείγματα τείνει ασυμπτωτικά στην X^2-κατανομή με p βαθμούς ελευθερίας.
Συνήθως χρησιμοποιούμε την τετραγωνική ρίζα του στατιστικού του Wald και ορίζουμε τον λόγο:

$$Z=\frac{\hat{\theta}_j}{\sqrt{var\left(\hat{\theta}_j\right)}}$$

ως z-στατιστικό.

2.3.2 Score tests

Όταν ισχύουν οι συνθήκες ομαλότητας το διάνυσμα των score ακολουθεί ασυμπτωτικά την κανονική κατανομή με μέση τιμή μηδέν και πίνακα διακυμάνσεων ίσο με τον πίνακα πληροφορίας.

$U(\theta) \sim N_p(0, I(\theta))$

Υπό την συνθήκη

H_0: $\theta=\theta_0$,

η τετραγωνική μορφή:

$Q=u(\theta_0)'I^{-1}(\theta_0)\, u(\theta_0)$

τείνει για μεγάλα δείγματα στην X^2 κατανομή με p βαθμούς ελευθερίας.

2.3.3 Likelihood ratio test

Ας υποθέσουμε ότι έχουμε δύο μοντέλα ω_1 και ω_2 τέτοια ώστε $\omega_1 \subset \omega_2$. Ας υποθέσουμε για

παράδειγμα ότι το ω_1 είναι το απλούστερο μόντέλο που προέκυψε από το ω_2 θέτωντας κάποιες παραμέτρους του ω_2 ίσες με μηδέν. Θέλουμε να ελέξουμε την υπόθεση ότι πράγματι οι συμτελεστές αυτοί είναι μηδέν.
Η βασική ιδέα είναι να συγκρίνουμε τις μέγιστές τιμές πιθανοφάνειας των δύο αυτών μοντέλων.
Η μέγιστη πιθανοφάνεια για το μοντέλο ω_1 είναι:

$$) = L(\hat{\theta}_{\omega_1},y)$$

Όπου $\hat{\theta}_{\omega_1}$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας της παραμέτρου θ για το μοντέλο ω_1

Η μέγιστη πιθανοφάνεια για το μοντέλο ω_1 είναι:

$$)=L(\hat{\theta}_{\omega_2},y)$$

Όπου $\hat{\theta}_{\omega_2}$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας της παραμέτρου θ για το μοντέλο ω_2

Ο λόγος αυτών των δύο ποσοτήτων

$$\lambda=\frac{\mathbf{L}\left(\hat{\theta}_{\omega_1},\mathbf{y}\right)}{\mathbf{L}\left(\hat{\theta}_{\omega_2},\mathbf{y}\right)}$$

παίρνει τιμές απο 0 (η συνάρτηση πιθανοφάνειας δεν παίρνει αρνητικές τιμές) ως 1.

Τιμές κοντά στο μηδέν δηλώνουν ότι το απλούστερο μοντέλο δεν γίνεται αποδεκτό ενώ τιμές κοντά στο 1 δηλώνουν ότι το απλούστερο μοντέλο είναι σχεδόν τόσο καλό όσο και το πιο σύνθετο μοντέλο.
Κάτω από συνθήκες ομαλότητας η ποσότητ -2logλ τείνει για αρκετά μεγάλα δείγματα στην X^2 κατανομή με βαθμούς ελευθερίας ίσους με την διαφορά στον αριθμό των παραμέτρων ανάμεσα στα δύο μοντέλα.Δηλαδη:

$$-2\log\lambda = 2\log L[\hat{\theta}_i(\omega_i 2), y] - 2\log L[\hat{\theta}_i(\omega_i 1), y] \rightarrow X^2_v$$

όπου $v=\dim(\omega_2)-\dim(\omega_1)$

Ας σημειώσουμε ότι ότι ο υπολογισμός του likelihood ratio test απαιτεί την προσαρμογή δύο μοντέλων, το wald test απαιτεί τουλάχιστον ένα μοντέλο ενώ το score test δεν απαιτεί κάποιο μοντέλο.

3. Το Μοντέλο της Λογιστικής Παλινδρόμησης

Αν προσπαθήσουμε να εκφράσουμε την πιθανότητα επιτυχίας p, μιας δίτιμης μεταβλητής Y με την βοήθεια ενός απλού γραμμικού μοντέλου :

$$p = b_0 + b_1 x$$

όπου x οι τιμές μιας ανεξάρτητης μεταβλητής X

το κύριο πρόβλημα που θα αντιμετωπίσουμε είναι ότι, αν και οι τιμές της p θεωρητικά δεν μπορούν να βρίσκονται εκτός του διαστήματος [0,1], οι τιμές της ποσότητας b_0+b_1x κυμαίνονται σε όλο το εύρος των πραγματικών αριθμών. Μια πρώτη σκέψη για την αντιμετώπιση αυτού του προβλήματος είναι να αντικαταστήσουμε στο μοντέλο την πιθανότητα p του γεγονότος της επιτυχίας με την σχετική πιθανότητα επιτυχίας. Δηλαδη με τον λόγο της πιθανότητας του γεγονότος της επιτυχίας προς την πιθανότητα του γεγονότος της αποτυχίας

$$\frac{p}{1-p}$$

Ο λόγος αυτός αν και θεωρητικά παίρνει τιμές μέχρι και το $+\infty$, δεν παίρνει αρνητικές τιμές.

Επομένως και πάλι ένα γραμμικό μοντέλο της μορφής

$$\frac{p}{1-p} = b_0 + b_1 x$$

δεν είναι επαρκές για την εκτίμηση της πιθανότητας p.

Αν όμως αντί του λόγου

$$\frac{p}{1-p}$$

χρησιμοποιηθεί ο φυσικός λογάριθμος

$$\ln\left[p/(1-p)\right]$$

τότε οι τιμές του μετασχηματισμένου λόγου, οι οποίες κυμαίνονται στο διάστημα (-∞,+∞) μπορούν να εκτιμήθούν με την βοήθεια ενός γραμμικού μοντέλου της μορφής

$$\ln\left[p/(1-p)\right] = b_0+b_1x$$

Η συνάρτηση **ln[p/ 1-p]** η οποία συνδέει την πιθανότητα της επιτυχίας p με την ανεξαρτητη μεταβλητή X, στην ορολογία των λογαριθμικών μοντέλων ονομάζεται logit της p και συμβολίζεται με logit(p). Δηλαδή

$$\text{Logit}(p)= \ln\left[p/(1-p)\right] = b_0+b_1x$$

Αντιλογαριθμίζοντας τα δύο μέλη της εξίσωσης:

$$\ln\left[p/(1-p)\right] = b_0+b_1x$$

προκύπτει:

$$\frac{p}{1-p} = e^{b_0+b_1x}$$

Και αν θέσουμε $z= b_0+b_1x$ παίρνουμε:

$$\frac{p}{1-p} = e^{z}$$

Επιλύοντας την τελευταία εξίσωση ως προς p, παίρνουμε:

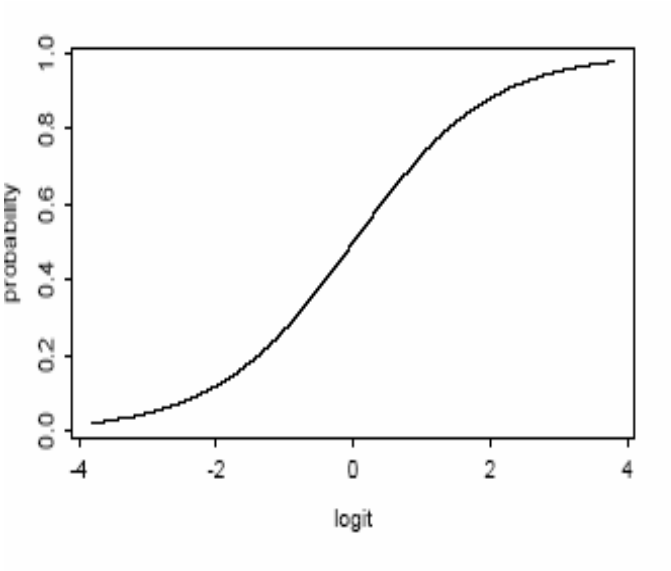
$$p= \frac{e^{z}}{1+e^{z}}$$

Η τελευταία εξίσωση (διατυπώνεται επίσης με την μορφή $P=1/1+e^{-z}$) αποτελεί την εκτίμηση της πιθανότητας της επιτυχίας p της δίτιμης μεταβλητής Y.
Το θεωρητικό διάγραμμα της συνάρτησης

f(z)=

η οποία εκτιμά την p είναι σιγμοειδής (σχήμα1) ενώ οι τιμές της κυμαίνονται στο διάστημα [0,1] εφόσον η z παίρνει τιμές στο διάστημα [-∞,+∞] και είναι αυτό στο οποίο κατατείνουν οι αθροιστικές πιθανότητες κανονικής κατανομής.

Σχήμα1



Η συνάρτηση f(z)= είναι επομένως κατάλληλη να χρησιμοποιηθεί ως μοντέλο για την εκτίμηση μιας πιθανότητας.
Από το διάγραμμα γίνεται φανερό ότι η σχέση της ανεξάρτητής μεταβλητής X και της πιθανότητας πραγματοποίησης του γεγονότος είναι μη γραμμική και οι εκτιμητές πιθανότητας θα βρίσκονται μεταξύ 0 και 1 ανεξαρτήτως της τιμής του Z.

Το γραμμικό μοντέλο που χρησιμοποιήθηκε για την εκτίμηση του λογαρίθμου της σχετικής πιθανότητας της επιτυχίας της δίτιμης μεταβλητής Y διευρύνεται και στην περίπτωση των περισσότερων της μίας ανεξαρτήτων μεταβλητών. Σε αυτήν την περίπτωση θέτωντας

$$Z=b_0+b_1x_1+b_2x_2+\ldots+b_kx_k$$

όπου $x_1,x_2,\ldots,x_k$ είναι οι τιμές των ανεξάρτητων μεταβλητών $X_1,X_2,\ldots,X_k$ προκύπτει

$$\ln[p/(1-p)] = b_0+b_1x_1+b_2x_2+\ldots+b_kx_k$$

ή ισοδύναμα:

$$\frac{p}{1-p}=e^{b_0+b_1x_1+b_2x_2+\ldots+b_kx_k}$$

Ενώ κατά την κλασσική γραμμική παλινδρόμηση η εκτίμηση των παραμέτρων $b_0,b_1,b_2,\ldots b_k$ γίνεται με την μέθοδο των ελαχίστων τετραγώνων κατά την λογιστική παλινδρόμηση η εκτίμηση των παραμέτρων γίνεται με την μέθοδο της μέγιστης πιθανοφάνειας(maximum likelihood) δηλαδή επιλέγονται οι πιο πιθανοφανείς τιμές των παραμέτρων προκειμένου να οδηγήσουν στα παρατηρούμενα αποτελέσματα.

4. Εκτίμηση καλής προσαρμογής του Μοντέλου

Υπάρχουν πολλοί τρόποι για την εκτίμηση της καλής προσαρμογής του μοντέλου λογιστικής παλινδρόμησης για τα δεδομένα καθενός προβλήματος. Μεταξύ αυτών είναι, η σύγκριση του αριθμού των προβλεπόμενων με τις εμπειρικές (παρατηρούμενες) παρατηρήσεις σε πίνακα ταξινόμησης, το ιστόγραμμα των εκτιμώμενων πιθανοτήτων, η διερεύνηση της πιθανοφάνειας των αποτελεσμάτων , το στατιστικό Z^2 της καλής προσαρμογής, το υπόδειγμα του X^2 και ο έλεγχος των Hosmer και Lameshow της ποσοστιαίας κατανομής των παρατηρήσεων σε ομάδες.

4.1. Ταξινόμηση των παρατηρήσεων

Μέτρο εκτίμησης καλής προσαρμογής του μοντέλου για τα δεδομένα συγκεκριμένου παραδείγματος είναι η εκτίμηση του βαθμού ορθής ταξινόμησης.Δηλαδή το ποσοστό των ορθών ταξινομημένων παρατηρήσεων με την σύγκριση του αριθμού των προβλεπόμενων ως προς τις εμπειρικές παρατηρήσεις στις ομάδες της εξαρτημένης μεταβλητής, όπως ακριβώς λειτουργεί ο συντελεστής R^2, για το μοντέλο της κλασσικής παλινδρόμησης.

Πρόκειται ουσιαστικά για την κατασκευή ενός πίνακα συνάφειας στον οποίο οι παρατηρήσεις ταξινομούνται ως προς την πραγματοποίηση του γεγονότος με βάση τα δειγματικά δεδομένα(observed outcome) και ως προς την πραγματοποίηση του γεγονότος με βάση τις εκτιμήσεις του μοντέλου (predicted outcome). Για να εκτιμηθεί –σύμφωνα με το μοντέλο- ότι το γεγονός θα συμβεί σε μια παρατήρηση, πρέπει η εκτιμώμενη πιθανότητα πραγματοποίησης του γεγονότος, για την παρατήρηση, να είναι μεγαλύτερη ή ίση με 0.50.

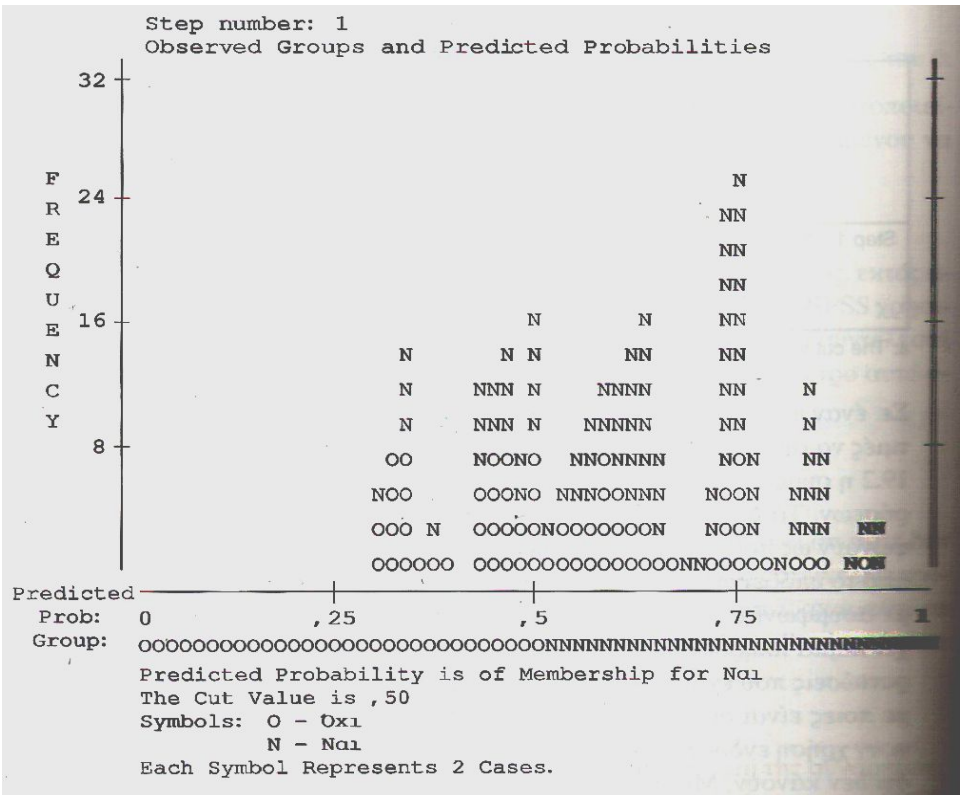
Πίνακας ταξινόμησης των Παρατηρήσεων

predicted observed	Οχι	Ναι	Percentage correct
Οχι	52	58	47.3
Ναι	38	158	76,7
Overall percentage			64,8

Σε έναν πίνακά ταξινόμησης θα πρέπει οι παρατηρούμενες και οι εκτιμώμενες τιμές να συμφωνούν κατά το δυνατόν περισσότερο. Στον παραπάνω πίνακα η συμφωνία αυτή προσσεγγίζει το 65% περίπου των παρατηρήσεων.
Τα διαγώνια κελιά του πίνακα περιέχουν τις παρατηρήσεις που συμφωνούν οι παρατηρούμενες και οι εκτιμώμενες τιμές ενώ τα εκτός διαγωνίου κελιά περιέχουν τις παρατηρήσεις που δεν συμφωνούν οι παρατηρούμενες και οι εκτιμώμενες τιμές.
Το πρόβλημα με τον πίνακα ταξινόμησης είναι ότι δεν μας πληροφορεί για το μέγεθος των εκτιμώμενων πιθανοτήτων στις περιπτώσεις που έχουμε λανθασμένες εκτιμήσεις. Ο πίνακας ταξινόμησης δεν δείχνει την κατανομή των εκτιμώμενων πιθανοτήτων στις δύο ομάδες αλλά κατά πόσο η πιθανότητα για κάθε ομάδα είναι μεγαλύτερη ή μικρότερη του 50%.

4.2.Ιστόγραμμα Εκτιμώμενων Πιθανοτήτων

Όπως είπαμε και πριν το πρόβλημα με τον πίνακα ταξινόμησης είναι ότι δεν μας δίνει πληροφοριες για το μέγεθος των εκτιμώμενων πιθανοτήτων στις περιπτώσεις που έχουμε λανθασμένες εκτιμήσεις. Για παράδειγμα δεν γνωρίζουμε ποιές είναι οι εκτιμώμενες πιθανότητες για τα 38 άτομα που ενώ η παρατηρούμενη τιμή είναι «Ναι», από το μοντέλο εκτιμάται «Όχι». Μια ερώτηση που πρέπει να απαντηθεί είναι αν οι εκτιμώμενες πιθανότητες για τα άτομα αυτά είναι κοντά στο όριο του 0,50 ή απέχουν σημαντικά από αυτό (γεγονός που θα μείωνε ακόμη περισσότερο την αξιοπιστία του μοντέλου).
Ερωτήματα τέτοια μπορούν να απαντηθούν από το ιστόγραμμα εκτιμώμενων πιθανοτήτων. Το ιστόγραμμα εκτιμωμένων πιθανοτήτων για τα δεδομένα που χρησιμοποιήσαμε και στον πίνακα ταξινόμησης είναι το παρακάτω:



Ο οριζόντιος άξονας του διαγράμματος ταξινόμησης αντιστοιχεί στις εκτιμώμενες από το μοντέλο πιθανότητες. Επάνω στον άξονα τοποθετούνται οι παρατηρήσεις της ανάλυσης υπό μορφή ιστογράμματος. Οι παρατηρήσεις οι οποίες έχουν εκτιμώμενη πιθανότητα πραγματοποίησης του γεγονότος μεγαλύτερη του 0.5 τοποθετούνται στο δεξιό μέλος του άξονα, ενώ οι παρατηρήσεις με πιθανότητα μικρότερη του 0.5 τοποθετούνται στο αριστερό μέλος του άξονα. Έπίσης κάθε παρατήρηση ορίζεται από το αρχικό διάγραμμα της κατηγορίας που αντιστοιχεί στην πραγματοποίηση ή μη του γεγονότος, έτσι όπως προκύπτει από τα δειγματικά δεδομένα.

Όσο στο ιστόγραμμα των εκτιμώμενων πιθανοτήτων διακρίνονται με μεγαλύτερη ευκρίνεια οι δύο ομάδες παρατηρήσεων , δηλαδή οι παρατηρήσεις ως προς το γεγονός που συνέβη βρίσκονται στο δεξιό τμήμα του 0.5 και ως προς το γεγονός που δεν συνέβη βρίσκονται στο αριστερό του 0.5 τόσο το μοντέλο είναι καλά προσαρμοσμένο στα δεδομένα του συγκεκριμένου παραδείγματος.

Στο ιστόγραμμα του παραδείγματος μας εμφανίζεται ένας αρκετά μεγάλος αριθμός παρατηρήσεων με ορθά εκτιμώμενες πιθανότητες επιτυχίας σημαντικά μεγαλύτερες του 0.5 (αρκετά από αυτές μεγαλύτερες του 0.7). Αντίθετα ο αριθμός των παρατηρήσεων που έχουν ορθά εκτιμώμενες πιθανότητες αποτυχίας είναι αρκετά μικρότερος και με τιμές των αντίστοιχων πιθανοτήτων να μην διαφοροποιούνται σημαντικά από το 0.5. Επιπλέον το σύνολο αυτών των παρατηρήσεων περιλαμβάνεται στο διάστημα (0.25,0.5) των εκτιμώμενων πιθανοτήτων.

4.3 Διερεύνηση της πιθανοφάνειας των αποτελεσμάτων

Εκτός από την διαπίστωση του κατά πόσο ικανοποιητικά το μοντέλο της λογιστικής παλινδρομησης ταξινομεί τις παρατηρήσεις μπορεί να αναζητηθεί κατά πόσο είναι «πιθανά» τα αποτελέσματα του μοντέλου με βάση τις δεδομένες παραμέτρους.

Η πιθανότητα των παρατηρούμενων αποτελεσμάτων είναι γνωστή ως «πιθανοφάνεια» και επειδή η πιθανοφάνεια έιναι μικρός αριθμός (μικρότερος της μονάδας) είναι συνήθες να χρησιμοποιείται το διπλάσιο του λογαρίθμου της πιθανοφάνειας (-2lnL) ως μέτρο καλής προσαρμογής του εκτιμώμενου μοντέλου στα δεδομένα μας.

Όσο μεγαλύτερη είναι η τιμή της συνάρτησης της πιθανοφάνειας ή όσο μικρότερη είναι η τιμή της συνάρτησης λογαριθμοπιθανοφάνειας -2lnL, τόσο καλύτερη είναι η προσαρμογή του μοντέλου στα δειγματικά δεδομένα. Δηλαδή όσο μεγαλύτερη είναι η πιθανότητα τα διαθέσιμα δειγματικά δεδομένα να προέρχονται από έναν πληθυσμό με παραμετρους τις εκτιμώμενες από το μοντέλο, τόσο καλύτερη η προσαρμογή του λογιστικού μοντέλου.

Θα λέγαμε ότι το μοντέλο είναι απόλυτα προσαρμοσμένο όταν η πιθανοφάνεια είναι 1 και η τιμή -2lnL μηδενική.

Επειδή όμως η τιμή του L και αντίστοιχα και η τιμή του -2lnL εξαρτώνται κάθε φορά από το μέγεθος του δείγματος, προκειμένου οι δύο αυτές ποσότητες να χρησιμοποιηθούν για την αξιολόγηση ενός λογιστικού μοντέλου, θα πρέπει να συγκρίνονται με τις αντίστοιχες ποσότητες ενός άλλου απλούστερου μοντέλου (baseline model) που εκτιμάται για τα ίδια δειγματικά δεδομένα. Η σύγκριση αυτή μπορεί να γίνει με την βοήθεια του λόγου των μέγιστων τιμών της συνάρτησης πιθανοφάνειας γαι τα δύο συγκρινόμενα μοντέλα.

Αν για παράδειγμα πρόκειται να αξιολογηθεί η προσαρμογή του μοντέλου

$\text{logit}(p)=b_0+b_1x_1+b_2x_2$ (με μέγιστη πιθανοφάνεια L_2)

σε σχέση με τα απλούστερα μοντέλα

$\text{logit}(p)= b_0+b_1x_1$ (με μέγιστη πιθανοφάνεια L_1) και

$\text{logit}(p)= b_0$ (με μέγιστη πιθανοφάνεια L_0)

αυτό γίνεται με την βοήθεια του λόγου

$-2\ln\left(\frac{L_1}{L_2}\right)=-2\ln L_1-(-2\ln L_2)$ για την πρώτη σύγριση

και του λόγου:

$-2\ln\left(\frac{L_0}{L_2}\right)=-2\ln L_0-(-2\ln L_2)$ για την δεύτερη σύγκριση

Υπό την προυπόθεση ότι ισχύουν οι μηδενικές υποθέσεις:

H_0: $\beta_2=0$ για την πρώτη σύγκριση

και

H_0: $\beta_1=\beta_2=0$ για την δεύτερη σύγκριση.

Οι λόγοι που προκύπτουν ακολουθούν την κατανομή X^2, με βαθμούς ελευθερίας 1 και 2 αντίστοιχα. Γενικά οι βαθμοί ελευθερίας του λόγου των τιμών της συνάρτησης πιθανοφάνειας είναι ίσοι με την διαφορά του αριθμού των ανεξάρτητων μεταβλητών που περιλαμβάνονται στα δύο συκρινόμενα κάθε φορά μοντέλα.

4.4.Το μοντέλο X^2

Στην περίπτωση που αξιολογείται συνολικά η προσαρμογή του μοντέλου

$logit(p)=b_0+b_1x_1+b_2x_2+….+b_kx_k$

στα δειγματικά δεδομένα χρησιμοποιείται ο λόγος

$-2\ln(\frac{L_0}{L_f})=-2\ln L_0-(-2\ln L_f)$

όπου L_f η τιμή της συνάρτησης πιθανοφάνειας για το πλήρες μοντέλο και L_0 η τιμή της συνάρτησης πιθανοφάνειας για το μοντέλο που περιλαμβάνει μόνο το σταθερό όρο (σημειώνεται ότι εάν δεν περιλαμβάνεται σταθερός όρος στο μοντέλο χρησιμοποιείται για σύγκριση η πιθανοφάνεια για το μοντέλο χωρίς καμιά μεταβλητή).

Η ποσότητα $-2\ln(\frac{L_0}{L_f})$ ακολουθεί την κατανομή X^2 με κ-βαθμούς ελευθερίας.
Με άλλα λόγια, το μοντέλο X^2- με το αντίστοιχο επίπεδο σημαντικότητας - ελέγχει την μηδενική υπόθεση για τους συντελεστές $\beta_1,\beta_2,...,\beta_κ$

H_0: $\beta_1=\beta_2=....=\beta_κ=0$

Η απόρριψη της μηδενικής υπόθεσης οδηγεί στην ενναλακτική

H_A: τουλάχιστον ένας συντελεστής β_i ,i=1,2,…,κ είναι διάφορος του 0.

Ο έλεγχος αυτός είναι ισοδύναμός με τον έλεγχο F που πραγματοποιείται σε ένα μοντέλο κλασικής πολλαπλής γραμμικής παλινδρόμησης προκειμένου να αξιολογηθεί η προσαρμογή του στα δειγματικά δεδομένα.

4.5. Έλεγχος βελτίωσης της τιμής X^2

Η βελτιωμένη τιμή του X^2 είναι η μεταβολή της τιμής -2lnL μεταξύ διαδοχικών σταδίων στη δημιουργία του μοντέλου. Ελέγχει την μηδενική υπόθεση ότι οι συντελεστές παλινδρόμησης των μεταβλητών που εισέρχονται στο μοντέλο στο τελευταίο στάδιο είναι μηδενικοί. Η τιμή X^2 είναι η ίδια με την βελτιωμένη τιμή του εάν θεωρήσουμε τα μοντέλα με το σταθερό όρο και το πλήρες μοντέλο(με όλες τις μεταβλητες). Εάν αντιθέτως, θεωρήσουμε περισσότερα των δύο αυτών υποδειγμάτων, κάνοντας π.χ επιλογή μεταβλητών, οι τιμές των X^2 θα διαφέρουν.
Ο έλεγχος είναι ανάλογος με εκείνον της μεταβολής F στην κλασσική πολλαπλή γραμμική παλινδρόμηση.

4.6. Στατιστικό Z^2 της καλής προσαρμογής

Ένα άλλο μέτρο καλής προσαρμογής του μοντέλου είναι το στατιστικό Z^2 το οποίο συγκρίνει τις παρατηρούμενες πιθανότητες με εκείνες που προβλέπονται από το μοντέλο.

Η τιμή του στατιστικού Z^2 δίνεται από την σχέση

$$Z^2 = \frac{\Sigma(\text{υπόλοιπο}_i)2}{Pi(1 - Pi)}$$

Όπου το υπόλοιπο είναι η διαφορά μεταξύ της παρατηρούμενης τιμής Y_i και της προβλεπόμενης P_i.

4.7. Έλεγχος των Hosmer και Lemeshow

Οι Hosmer και Lemeshow χρησιμοποίησαν το στατιστικό C για τον έλεγχο της καλής προσαρμογής των δεδομένων (ποσοστιαίας κατανομής των παρατηρήσεων σε ομάδες). Το στατιστικό C θεωρείται ότι κατανέμεται με βάση την κατανομή X^2 του Pearson.
Τιμή του X^2 που αντιστοιχεί σε επίπεδο σημαντικότητας a>0.05 δηλώνει ότι το μοντέλο της λογιστικής παλινδρόμησης είναι καλά προσαρμοσμένο στα δεδομένα.

4.8. Ο συντελεστής προσδιορισμού R^2 των Cox και Snell

Για την αξιολόγηση της προσαρμογής του λογιστικού μοντέλου εκτός από όλα τα προηγούμενα χρησιμοποιείται και ένα επιπλέον μέτρο καλής προσαρμογής, αντίστοιχο με τον συντελεστή προσδιορισμού R^2 που χρησιμοποιείται στην πολλαπλή γραμμική παλινδρόμηση. Το μέτρο αυτό ονομάζεται R^2 των Cox και Snell και ισούται με

$$R^2 = 1 - \left[\frac{L_0}{L_f}\right]^{2/n}$$

όπου n το μέγεθος του δείγματος.

Το πρόβλημα με τον συγκεκριμένο συντελεστή προσδιορισμού είναι ότι ποτέ δεν καταλήγει να πάρει μέγιστη τιμή το 1. Ο Nagelkerke το 1991 πρότεινε μια τροποποίηση του συντελεστή των Cox και Snell προκειμένου να ξεπεράσει το συγκεκριμένο πρόβλημα. Ο συντελεστής που πρότεινε είναι ο:

$$\widetilde{R^2} = R^2/R^2_{max} \in (0,1)$$

όπου $R^2_{max} = 1 - [L_0]^{2/n}$

5. Έλεγχος και Ερμηνεία των Συντελεστών Παλινδρόμησης

5.1. Ερμηνεία των Συντελεστών Παλινδρόμησης

Η ερμηνεία των συντελεστών του μοντέλου της γραμμικής παλινδρόμησης είναι απλή και ορίζεται ευθέως. Ο κάθε ένας από τους συντελεστές $b_1,b_2,\ldots,b_k$ της εξίσωσης

$$\hat{y}=b_0+b_1x_1+b_2x_2+\ldots+b_kx_k$$

εκφράζει την μέση μεταβολή της εξαρτημένης μεταβλητής Υ για μία μονάδα αύξησης της αντίστοιχης ανεξάρτητης μεταβλητής X_i, όταν οι τιμές των άλλων ανεξάρτητων μεταβλητών παραμένουν σταθερές.

Στη λογιστική παλινδρόμηση η ερμηνεία των συντελεστών μπορεί να γίνει σε εναρμόνιση με τα παραπάνω εφόσον προηγουμένως γίνει κάποιος αναγκαίος μετασχηματισμός. Ο μετασχηματισμός αυτός αφορά την απόδοση της εξίσωσης της λογιστικής παλινδρόμησης με το λογάριθμο του λόγου πιθανότητας πραγματοποίησης του γεγονότος προς την πιθανότητα μη πραγματοποίησης του γεγονότος. Ουσιαστικά είναι η μορφή της γραμμικής συνάρτησης με την οποία μπορεί να εκτιμηθεί ο λογάριθμός της σχετικής πιθανότητας, δηλαδή:

$$\mathrm{Ln}\left[\frac{p}{1-p}\right]=b_0+b_1x_1+b_2x_2+\ldots+b_kx_k$$

Με τον παραπάνω μετασχηματισμό προκύπτει η ερμηνεία των συντελεστών του λογιστικού μοντέλου με τρόπο ανάλογο με αυτόν της γραμμικής παλινδρόμησης.

Κάθε ένας από τους συντελεστές $b_1,b_2,\ldots,b_k$ μπορεί να ερμηνευθεί ως τη μεταβολή του λογαρίθμου της σχετικής πιθανότητας $\mathrm{Ln}\left[\frac{p}{1-p}\right]$ από την κατά μονάδα μεταβολή της ανεξάρτητης μεταβλητής X_i, εφόσον οι τιμές των υπολοίπων ανεξάρτητων μεταβλητών παραμένουν σταθερές.

Επειδή είναι όμως πιο ενδιαφέρον οι συντελεστές να μπορούν να ερμηνεύουν την μεταβολή στο λόγο των πιθανοτήτων –και όχι στο λογάριθμο της σχετικής πιθανότητας-από την κατα μονάδα μεταβολή της ανεξάρτητης μεταβλητής X_i, είναι προτιμότερο η εξίσωση

$$\mathrm{Ln}\left[\frac{p}{1-p}\right]=b_0+b_1x_1+b_2x_2+\ldots+b_kx_k$$

να γραφεί με αντιλογαρίθμιση των δύο μερών της ως

$$\frac{p}{1-p}=e^{b_0+b_1x_1+b_2x_2+\ldots+b_kx_k}=e^{b_0}e^{b_1x_1}e^{b_2x_2}\ldots\ldots\ldots e^{b_kx_k}$$

Από την τελευταία εξίσωση προκύπτει ότι ο παράγοντας e^{b_i} είναι ο παράγοντας επί τον οποίο πολλαπλασιάζεται η σχετική πιθανότητα πραγματοποίησης του γεγονότος, όταν η ανεξάρτητη μεταβλητή X_i, αυξηθεί κατά μία μονάδα και εφόσον οι υπόλοιπες μεταβλητές παραμένουν σταθερές.

Αν το b_i είναι θετικό, ο παράγοντας e^{b_i} είναι μεγαλύτερος από την μονάδα, γεγονός που σημαίνει ότι η σχετική πιθανότητα αυξάνει. Αν το b_i είναι αρνητικό, ο παράγοντας e^{b_i} είναι μικρότερος της μονάδας, δηλαδή η σχετική πιθανότητα (δηλαδή η η εξαρτημένη μεταβλητη

$\frac{p}{1-p}$) μειώνεται. Τέλος όταν το b_i είναι μηδέν, ο παράγοντας e^{b_i} γίνεται ίσος με την μονάδα, που σημαίνει ότι δεν συμβαίνει καμιά μεταβολή στην σχετική πιθανότητα.

5.2. Έλεγχοι για τους συντελεστές του λογιστικού μοντέλου

Ο προσδιορισμός των συντελεστών του λογιστικού μοντέλου γίνονται με την βοήθεια της μεθόδου εκτίμησης μέγιστης πιθανοφάνειας, μιας διαδικασίας η οποία εκτιμά τις πληθυσμιακές παραμέτρους του μοντέλου με κριτήριο τη μεγιστοποίηση της πιθανότητας τα διαθέσιμα δειγματικά δεδομένα να έχουν παραχθεί από τις εκτιμώμενες παραμέρους.
Εκτός των συντελεστών του δειγματικού μοντέλου $b_1,b_2,\ldots,b_k$, η μέθοδος εκτίμησης μέγιστης πιθανοφάνειας εκτιμά και τα αντίστοιχα τυπικά σφάλματα αυτών.

Είναι γνωστό ότι ένας οποιοσδήποτε συντελεστής b_i, του λογιστικού μοντέλου, για μεγάλα δείγματα ακολουθεί κατα προσέγγιση την κανονική κατανομή με μέση τιμή τον αντίστοιχο πληθυσμιακό συντελεστή β_i, και εκτιμώμενο τυπικό σφάλμα $se(b_i)$. Επομένως όταν το μέγεθος του δείγματος είναι αρκετά μεγάλο το πηλίκο:

$$\frac{b_i - \beta_i}{se(b_i)}$$

ακολουθεί κατά προσέγγιση την τυπική κανονική κατανομή.

Ο έλεγχος της μηδενικής υπόθεσης ότι ο συντελεστής παλινδρόμησης είναι μηδενικός δηλαδή

H_0: $\beta_i=0$ (ή ισοδύναμα $e^{\beta_i}=1$)

έναντι της ενναλακτικής

H_A: $\beta_i\neq0$ (ή ισοδύναμα $e^{\beta_i}\neq1$)

βασίζεται στο στατιστικό του Wald και στο επίπεδο στατιστικής σημαντικότητας του, το οποίο ακολουθεί X^2 κατανομή.

Αν μια μεταβλητή έχει ένα βαθμό ελευθερίας το στατιστικό του Wald είναι το τετράγωνο του λόγου του συντελεστή προς το τυπικό σφάλμα αυτού:

$$z=[\qquad ^2$$

όπου όταν ισχύει η μηδενική υπόθεση H_0: $\beta_i=0$, ακολουθεί X^2 κατανομή με 1 βαθμό ελευθερίας. Για κατηγορικές μεταβλητές το στατιστικό του Wald έχει βαθμούς ελευθερίας κατά ένα λιγότερους από τον αριθμό των κατηγοριών.
Η εγκυρότητα του στατιστικού του Wald απαιτεί την εφαρμογή του σε μεγάλα δείγματα, ώστε να διασφαλίζεται και η αξιοπιστία των ασυμπτωτικών προσεγγίσεων από τις οποίες προκύπτει η δειγματοληπτική κατανομή του.
Το στατιστικό του Wald έχει ένα βασικό μειονέκτημα. Με την αύξηση της απόλυτης τιμής του συντελεστή παλινδρόμησης, το εκτιμώμενο τυπικό σφάλμα γίνεται επίσης μεγάλο που οδηγεί σε πολύ μικρή τιμή του στατιστικού Wald, με αποτέλεσμα η υπόθεση ότι η τιμή του συντελεστή παλινδρόμησης είναι μηδενική να απορρίπτεται ενώ θα έπρεπε να γίνεται αποδεκτή.
Συμπερασματικά όταν η τιμή του συντελεστή παλινδρόμησης είναι μεγάλη, δεν πρέπει να στηριζόμαστε στο στατιστικό του Wald για τον έλεγχο του αντίστοιχου συντελεστή παλινδρόμησης. Σε αυτές τις περιπτώσεις είναι προτιμότερο να χρησιμοποιείται για τον

έλεγχο ενός συντελεστή του μοντέλου ο λόγος των μέγιστων τιμών της συνάρτησης πιθανοφάνειας (likelihood ratio statistic).

Ο λόγος αυτός είναι η ποσότητα:

$$-2\ln\left(\frac{L_1}{L_2}\right)=-2\ln L_1-(-2\ln L_2)$$

όπου L_1 είναι η μέγιστη τιμή της συνάρτησης πιθανοφάνειας όταν ο συντελεστής β_i (δηλαδή η αντίστοιχη ανεξάρτητη μεταβλητή), για τον οποίο γίνεται ο έλεγχος, δεν περιλαμβάνεται στο μοντέλο και L_2, η μέγιστη τιμή της συνάρτησης πιθανοφάνειας, όταν ο συντελεστής β_i, περιλαμβάνεται στο μοντέλο. Η κατανομή του λόγου των μέγιστων τιμών της συνάρτησης πιθανοφάνειας όταν ισχύει η μηδενική υπόθεση H_0: $\beta_i=0$ ακολουθεί την κατανομή X^2 με 1 βαθμό ελευθερίας.
Η συνάρτηση -2lnL, η οποία προκύπτει μετασχηματίζοντας λογαριθμικά την L, ονομάζεται συνάρτηση λογαριθμο-πιθανοφάνειας (log likelihood function).
Επειδή η μέγιστη τιμή της συνάρτησης πιθανοφάνειας αυξάνει όσο αυξάνει και ο αριθμός των ανεξάρτητων μεταβλητών του μοντέλου ισχύει γενικά

$-\infty < L_1 < L_2$ ή $0 < L_1/L_2 < 1$

λογαριθμίζοντας τα δύο μέλη της ανισότητας προκύπτει

$$-\infty < \ln\left(\frac{L_1}{L_2}\right) < 0 \quad \text{ή} \quad 0 < 2\ln\left(\frac{L_1}{L_2}\right) < +\infty$$

Όσο μεγαλύτερη είναι επομένως η τιμή L_2 σε σχέση με την τιμή L_1 τόσο καλύτερη η προσαρμογή του πιο σύνθετου μοντέλου στα δειγματικά δεδομένα και επομένως τόσο μεγαλύτερη η τιμή του λόγου $-2\ln\left(\frac{L_1}{L_2}\right)$. Κατά συνέπεια τόσο αυξάνει και η πιθανότητα να βρίσκεται ο λόγος $-2\ln\left(\frac{L_1}{L_2}\right)$ στο άνω άκρο της κατανομής X^2.
Δηλαδή υποθέτωντας ότι η τιμή ενός συντελεστή β_i του μοντέλου είναι ίσος με 0 (H_0: $\beta_i=0$), αναμένουμε ο λόγος $\frac{L_1}{L_2}$ να είναι ίσος με την μονάδα ή ισοδύναμα ο λόγος $-2\ln\left(\frac{L_1}{L_2}\right)$ να είναι ίσος με το 0. Αν η τιμή του λόγου $-2\ln\left(\frac{L_1}{L_2}\right)$ είναι πολύ μεγαλύτερη του 0,

ώστε να ορίζεται στη δεξία ουρά της κατανομής X^2 μια περιοχή εμβαδού μικρότερου από 0.5, τότε μηδενική υπόθεση (H_0: $\beta_i=0$) απορρίπτεται και γίνεται δεκτή η εναλλακτική υπόθεση (H_A: $\beta_i\neq0$).

6. Διαστήματα εμπιστοσύνης για τους συντελεστές του λογαριθμικού μοντέλου

Εκτός των ελέγχων που μπορούν να γίνουν για τους συντελεστές παλινδρόμησης του λογιστικού μοντέλου, μπορούν να κατασκευαστούν και τα αντίστοιχα διαστήματα εμπιστοσύνης. Έτσι το 95% διάστημα εμπιστοσύνης για το συντελεστή β_i είναι

$$(b_i - 1.96 * se(b_i) , b_i + 1.96 * se(b_i))$$

Εφόσον η κατανομή του b_i είναι κατά προσέγγιση κανονική.

Διαστήματα εμπιστοσύνης μπορούν να κατασκευαστούν και για τις ποσότητες e^{β_i} όπου ι=1,2,..,κ με αντιλογαρίθμηση των προηγούμενων ορίων. Δηλαδή το 95% διάστημα εμπιστοσύνης για τον αντιλογάριθμο του β_1, e^{β_i} είναι

$$(e^{[b_i - 1.96*se(b_i)]} , e^{[b_i = 1.96*se(b_i)]}).$$

7. Μερική Συσχέτιση

Όπως στην πολλαπλή γραμμική παλινρόμηση έτσι και στην λογιστική παλινδρόμηση είναι δύσκολο να καθοριστεί η συμβολή καθεμιάς από τις ανεξάρτητες μεταβλητές επί της εξαρτημένης, και αυτό γιατί η συμβολή αυτή εξαρτάται από την παρουσία των λοιπών μεταβλητών στο μοντέλο.
Το πρόβλημα γίνεται εξαιρετικά σημαντικό όταν οι ανεξάρτητες μεταβλητές σχετίζονται σε μεγάλο βαθμό μεταξύ τους.
Στατιστικό που χρησιμοποιείται για τις μερικές στσχετίσεις καθεμιάς των ανεξάρτητων μεταβλητών με την εξαρτημένη είναι το στατιστικό R, το οποίο παίρνει τιμές μεταξύ -1 και +1. Θετική τιμή του στατιστικού R δηλώνει ότι με την μεταβολή της ανεξάρτητης μεταβλητής αυξάνεται και η πιθανόφάνεια να συμβεί το γεγονός, ενώ αρνητική τιμή του στατιστικού R δηλώνει το αντίθετο.
Μικρές τιμές του στατιστικού R δηλώνουν ότι η ανεξάρτητη μεταβλητή έχει μικρή συμμετοχή στο μοντέλο.
Η τιμή του στατιστικού R δίνεται από την μαθηματική σχέση:

$$R=\pm\sqrt{((\text{στατιστικό του Wald} - 2k)/[-2\ln L]_0)}$$

όπου k είναι οι βαθμοί ελευθερίας για την δεδομένη μεταβλητή και $-2\ln L_0$ είναι το διπλάσιο του λογάριθμου της πιθανοφάνειας του μοντέλου που περιλαμβάνει μόνο τον σταθερό όρο (ή του μοντέλου χωρίς μεταβλητές όταν δεν υπάχει σταθερός όρος).
Να σημειωθεί ότι στην παραπάνω εξίσωση αν το στατιστικό του Wald είναι μικρότερο του -2k ο αριθμητής του κλάσματος και συνεπώς η τιμή του στατιστικού R γίνεται μηδενική. Τέλος το πρόσημο του στατιστικού R είναι ταυτόσημο με αυτό του αντίστοιχου συνελεστή παλινδρόμησης.

8. Προσθήκη στο Μοντέλο Όρων Αλληλεπίδρασης

Όπως στην γραμμική έτσι και στην λογιστική παλινδρόμηση μπορούμε να συμπεριλάβουμε στο μοντέλο μας όρους που εκφάζουν την αλληλεπίδραση των μεταβλητών. Η αλληλεπίδραση μπορεί να αφορά όχι μόνο συνεχείς αλλά και κατηγορικές μεταβλητές. Στην δεύτερη περίπτωση η αλληλεπίδραση της κατηγορικής μεταβλητής με άλλη ή άλλες είναι το γινόμενο των τιμών των συνεχών μεταβλητών επι τις νέες μεταβλητές, τις ψευδομεταβλητές, που έχουν ήδη δημιουργηθεί ή των ψευδομεταβλητών μεταξύ τους.
Παράδειγμα για διμερή κατηγορική μεταβλητή δημιουργούνται δύο ψευδομεταβλητές με τιμές τους αριθμούς 0 και 1 , για τριμερή κατηγορική μεταβλητη δημιουργούνται τρεις ψευδομεταβλητές , καθεμια να έχει τιμή 1 και οι άλλες τιμή 0, έτσι που η αλληλεπίδραση να εκφράζεται ως το γινόμενο των τιμών των άλλων μεταβλητών επί τις τιμές των ψευδομεταβλητών.

9. Επιλογή των Ανεξάρτητων Μεταβλητών

Στην λογιστική παλινδρόμηση, όπως και στις άλλες πολυμεταβλητές στατιστικές ενδιαφερόμαστε να αναγνωρίσουμε ομάδες ανεξάρτητων μεταβλητών που είναι καλύτερες για την πρόβλεψη της εξαρτημένης μεταβλητής. Δεν υπάρχει ο «άριστος» αλγόριθμος για την «άριστη» επιλογή των μεταβλητών και διαφορετικοί αλγόριθμοι επιλογής μπορούν να οδηγήσουν σε διαφορετικά μοντέλα.
Επομένως καλό είναι να εξετάζουμε διάφορα δυνατά μοντέλα και να επιλέγουμε εκείνο με βάση την ερμηνεία των αποτελεσμάτων, την οικονομία του και την δυνατότητα στην απόκτηση των δεδομένων των μεταβλητών.
Δεν πρέπει να ξεχνάμε ότι κάθεμια επιλογή αφορά δεδομένα του συγκεκριμένου δείγματος και δεν εγγυάται ότι για ένα άλλο δείγμα από τον ίδιο, με τον προηγούμενο δείγμα, πληθυσμό θα προκύψει το ίδιο μοντέλο.
Υπάρχουν τρεις κυρίως μέθοδοι επιλογής των ανεξάρτητων μεταβλητών, εκείνη κατά την οποία εισέρχονται ταυτόχρονα όλες οι μεταβλητές, η προοδευτική κατά στάδια επιλογή και η προς τα πίσω κατά στάδια απάλειψη μεταβλητών.
Σε όλες τις περιπτώσεις τα κριτήρια που πρέπει να ικανοποιούνται αφορούν και την εισαγωγή αλλά και την απομάκρυνση των μεταβλητών από την εξίσωση παλινδρόμησης. Δηλαδή μια μεταβλητή εισέρχεται στο μοντέλο εφόσον ικανοποιεί αρχικά το κριτήριο εισαγωγής και στη συνέχεια και το κριτήριο παραμονής.
Το κριτήριο εισαγωγής (score statistic) είναι κοινό για όλες τις μεθόδους επιλογής ενώ η απομάκρυνση μιας μεταβλητής από το μοντέλο μπορεί να γίνεται ή με το κριτήριο του Wald ή με το λόγο των μέγιστων τιμών της συνάρτησης πιθανοφάνειας (likelihood ratio test) ή τέλος με την βοήθεια του conditional likelihood ratio test μιας παραλλαγής του λόγου των τιμών της συνάρτησης πιθανοφάνειας όταν οι συντελεστές του μοντέλου εκτιμώνται υπό συνθήκες.
Τέλος όλες οι μεταβλητές που χρησιμοποιούνται για την αντιπροσώπευση της ίδιας κατηγορικής μεταβλητής εισέρχονται ή απομακρύνονται από το μοντέλο ταυτοχρόνως.

9.1. Κριτήρια επιλογής Μεταβλητών

A) Score Statistic

Το score statistic υπολογίζεται για κάθε μεταβλητή που δεν είναι στο μοντέλο για να καθορίσει εάν η μεταβλητή πρέπει να χρησιμοποιηθεί στο μοντέλο.

Ας υποθέσουμε ότι υπάρχουν k_1 μεταβλητές οι $r_1,r_2,\ldots,r_{k1}$ στο μοντέλο και k_2 μεταβλητές οι $\gamma_1,\gamma_2,\ldots,\gamma_{k2}$ έξω από το μοντέλο.

Το score statistic για την μεταβλητή γ_i ορίζεται ως:

$\mathbf{S}_i=(L^*_{\gamma i})^2 B_{22,i}$, εάν η γ_i είναι μία μη κατηγορική μεταβλητή.

Εάν η γ_i είναι μία κατηγορική μεταβλητή με m κατηγορίες μετατρέπεται σε ένα ψευδοδιάνυσμα (m-1) διάστασης. Συμβολίζουμε αυτές τις νέες m-1 μεταβλητές ως

$$\vec{\gamma 1},\ldots..,\vec{\gamma}_{i+m-2}$$

Το score statistic για την γ_i είναι τότε:

$$\mathbf{S}_i=(L^*\vec{\gamma}_i)' B_{22,i} L^*\vec{\gamma}$$

όπου

$(L^*\vec{\gamma})' = (L^*\vec{\gamma}_i)',\ldots., L^*\vec{\gamma}_{i+m-2})')$ και ο (m-1)*(m-2) πίνακας $B_{22,i}$ είναι

$B_{22,I} = (A_{22,i} - A_{21.i} * A_{11}^{-1} * A_{12.i})^{-1}$, με

$A_{11}= \mathbf{a}'\vec{\gamma}\,\mathbf{a}$

$A_{12}= \mathbf{a}'\vec{\gamma}\,\boldsymbol{\gamma}_i$

$A_{22}= \boldsymbol{\gamma}_i'\vec{\gamma}\,\boldsymbol{\gamma}_i$, όπου **a** είναι ο πίνακας για τις μεταβλητές $r_1,r_2,\ldots,r_{k1}$ και $\boldsymbol{\gamma}_i$ είναι ο

πίνακας με τις ψετδομεταβλητές $\vec{\gamma 1},\ldots..,\vec{\gamma}_{i+m-2}$

Σημειώνουμε ότι ο πίνακας **a** περιέχει μία στήλη από μονάδες εκτός και αν ο σταθερός όρος έχει εξαιρεθεί. Βασιζόμενοι στην εκτίμηση μέγιστης πιθανοφάνειας για τις παραμέτρους στο μοντέλο,ορίζεται ως $\vec{\gamma}$, $\vec{\gamma}$ =diag{$\pi_1(1-\pi_1),\ldots., \pi_n(1-\pi_n)$}.

Η ασυμπτωτική κατανομή των score statistic είναι X^2 με βαθμούς ελευθερίας ίσους με τον αριθμό των μεταβλητών που συμπερριλήφθηκάν.

Τέλος σημειώνουμε ότι:

α) Αν το μοντέλο είναι στην αρχή και δεν περιέχεται καμία μεταβλητή ο $B_{22,i}$ ορίζεται από τον

$A^{-1}_{22,i}$ και $\tilde{V}$ = I_n

β) Αν $B_{22,i}$ δεν είναι θετικά ορισμένος το score statistic και τα κατάλοιπα X^2 statistic είναι μηδέν.

B) <u>Wald statistic</u>

Το στατιστικό του Wald υπολογίζεται για τις μεταβλητές που έχουν εισέλθει στο μοντέλο ώστε να προσδιορίσει ποιές μεταβλητές πρέπει να μετακινηθούν.

Εάν η X_i μεταβλητή δεν είναι κατηγορική το στατιστικό του Wald ορίζεται ως:

$$Wald_i = \frac{\hat{\beta}_i^2}{\hat{\sigma}^2_{\hat{\beta}_i}}$$

Εάν η μεταβλητή όμως X_i είναι κατηγορική το στατιστικό του Wald υπολογίζεται ως εξής:

Ας είναι $\hat{\beta}_i$ το διάνυσμα που περιέχει τις εκτιμήσεις μέγιστης πιθανοφάνειας για τις (m-1) ψευδομεταβλητές και C ο ασυμπτωτικός πίνακας συσχετίσεων για το β_i .

$\hat{\beta}_i$

Το στατιστικό του Wald τότε είναι:

$Wald_i = \hat{\beta}_i' C^{-1} \hat{\beta}_i$

Η ασυμπτωτική κατανομή του Wald statistic ακολουθεί X^2 κατανομή με βαθμούς ελευθερίας ίσους με τον αριθμό των εκτιμώμενων παραμέρων.

Γ) Likelihood Ratio (LR) Statistic

Το LR στατιστικό χρησιμοποιείται για να προσδιορίσει αν μια μεταβλητή πρέπει να μετακινηθεί από μοντέλο. Θεωρούμε ότι υπάρχουν r_1 μεταβλητές στο μοντέλο που ελέγχουμε το οποίο ονομάζουμε και ως πλήρες μοντέλο. Υπολογίζουμε την τιμή της συνάρτησης μέγιστης πιθανοφάνειας του πλήρους μοντέλου. Για κάθε μία μεταβλητή που μετακινείται κάθε φορά από το πλήρες μοντέλο υπολογίζεται και η συνάρτηση μέγιστης πιθανοφάνειας του περιορισμένου μοντέλου που δημιουργείται κάθε φορά.
Το LR στατιστικό ορίζεται ως

LR=$-2\ln$ =

=-2lnl(περιορισμένο μοντέλο)-(-2lnl(πλήρες μοντέλο))

Το LR στατιστικό ασυμπτωτικά ακολουθεί X^2 κατανομή με βαθμούς ελευθερίας ίσους με την διαφορά των εκτιμώμενων παραμέτρων στα δύο μοντέλα.

Δ. Conditional Statistic

Το υπο συνθήκη στατιστικό υπολογίζεται επίσης για κάθε μεταβλητή του μοντέλου. Η φόρμουλα για το υπόσυνθήκη στατιστικό είναι η ίδια με το LR στατιστικό με την διαφορά ότι οι εκτιμητές των παραμέτρων για κάθε περιορισμένο μοντέλο είναι υπο συνθήκη εκτιμητές και όχι εκτιμητές μέγιστης πιθανοφάνειας.
Οι υπο συνθήκη εκτιμητές ορίζονται ως εξής:
Ας είναι $\hat{\beta}=(\hat{\beta}_1, \hat{\beta}_2,,,,,, \hat{\beta}_{r_1})$ οι εκτιμητές μέγιστης πιθανιφάνειας για τις r_1 μεταβλητές που βρίσκονται στο μοντέλο και ας είναι C ο ασυμπτωτικός πίνακας συσχετίσεων

για το . Αν η μεταβλητή X_i μετακινηθεί από το μοντέλο, ο υπό συνθήκη εκτιμητής για τις παραμέτρους που έμειναν στο μοντέλο δεδομένου του $\hat{\beta}$ είναι :

$$\hat{\beta}_{(i)} = \hat{\beta}_{(i)} - c_{12}^{(i)} (c_{22}^{(i)})^{-1} \hat{\beta}_{(i)}$$

όπου

$\hat{\beta}_i$ είναι ο εκτιμητής μέγιστης πιθανοφάνειας της παραμέτρου της X_i μεταβλητής και
$\hat{\beta}_{(i)}$ είναι το $\hat{\beta}$ χωρίς την συνιστώσα $\hat{\beta}_i$, $c_{12}^{(i)}$ είναι η συσχέτιση μεταξύ $\hat{\beta}_{(i)}$ και $\hat{\beta}_i$

Και

$c_{22}^{(i)}$ είναι η συσχέτιση του $\hat{\beta}_i$

Τότε το υπό συνθήκη στατιστικό υπολογίζεται από την παρακάτω σχέση:

-2[L($\beta^*_{1((i))}$ − L(πλήρες μοντέλο)]

όπου

L($\tilde{\beta}_{(i)}$) είναι η log likelihood συνάρτηση.

9.2. Μέθοδοι Επιλογής των ανεξάρτητων μεταβλητών ενός μοντέλου

Α. Ταυτόχρονη Είσοδος Μεταβλητών

Η μέθοδος της ταυτόχρονης εισόδου των μεταβλητών είναι διαδικασία εισόδου ομάδας ανεξαρτήτων μεταβλητών χωρίς την εφαρμογή κάποιου κριτηρίου όπως γίνεται με τις μεθόδους επιλογής.
Κατά την μέθοδο αυτή οι ανεξάρτητες μεταβλητές εισέρχονται ως δέσμη στην εξίσωση παλινδρόμησης σε ένα στάδιο.

Β. Προοδευτική Κατά Στάδια Επιλογή Μεταβλητών

Η προοδευτική κατά στάδια επιλογή (forward selections) των ανεξάρτητων μεταβλητών στην λογιστική παλινρόμηση ακολουθεί την ίδια διαδικασία με εκείνη που ακολουθείται και στην πολλαπλή γραμμική παλινδρόμηση.
Η αρχή γίνεται με μοντέλο που περιλαμβάνει μόνο το σταθερό όρο και σε κάθε στάδιο εισέρχεται η μεταβλητή με το μικρότερο επίπεδο σημαντικότητας για το αντίστοιχο κριτήριο εισαγωγής score statistic, με την προυπόθεση ότι δεν υπερβαίνει μία δεδομένη οριακή τιμή (π.χ 0.05).
Οι μεταβλητές που έχουν εισέλθει ελέγχονται αν ικανοποιούν κάποι κριτήριο απομάκρυνσης. Εάν ως κριτήριο απομάκρυνσης χρησιμοποιείται το στατιστικό του Wald, εξετάζονται όλα τα στατιστικά wald των μεταβλητών που έχουν εισέλθει στο μοντέλο και απομακρύνεται η μεταβλητή με το μεγαλύτερο επίπεδο στατιστικής σημαντικότητας του στατιστικού του Wald, με την προυπόθεση ότι δεν θα υπερβαίνει μία δεδομένη οριακή τιμή
(π.χ 0.01).
Εφόσον δεν υπάρχουν στο μοντέλο μεταβλητές που να ικανοποιούν κριτήριο απομάκρυνσης, η διαδικασία συνεχίζεται με την είσοδο νέας μεταβλητής.
Εάν αντιθέτως κάποια μεταβλητή έχει επιλεγεί να απομακρυνθεί και το μοντέλο καταλήγει σε κάποιο προηγούμενο, η επιλογή μεταβλητών παύει, αλλιώς εκτιμάται το μοντέλο χωρίς την μεταβλητή αυτή και στη συνέχεια εξετάζονται για απομάκρυνση οι άλλες μεταβλητές.
Η διαδικασία αυτή συνεχίζεται μέχρις ότου δεν μπορούν πλέον να απομακρυνθούν από το μοντέλο μεταβλητές και ξαναρχίζει η διαδικασία εξέτασης για είσοδο μεταβλητών.
Η όλη διαδικασία συνεχίζεται μέχρις ότου είτε καταλήξουμε σε κάποιο προηγούμενο μοντέλο είτε δεν υπάρχουν μεταβλητές που να ικανοποιούν κριτήρια εισόδου και εξόδου.

Αντί του στατιστικού του Wald, ένα ακόμη κριτήριο απομάκρυνσης μεταβλητών είναι ο έλεγχος του λόγου πιθανοφάνειας. Σύμφωνα με αυτόν τον έλεγχο το μοντέλο εκτιμάται ύστερα από την απομάκρυνση διαδοχικά κάθεμιας μεταβλητής και ελέγχου της μεταβολής που δημιουργήθηκε στο λόγο της πιθανοφάνειας.
Ο έλεγχος του λόγου πιθανοφάνειας, για την μηδενική υπόθεση ότι οι συντελεστές παλινδρόμησης των απομακρυσμένων μεταβλητών είναι μηδενικοί, γίνεται με την διαίρεση της τιμής της πιθανοφάνειας του μοντέλου χωρίς την μεταβλητή προς την τιμή της πιθανοφάνειας του μοντέλου αυτου που περιλαμβάνει την μεταβλητή.
Εάν η μηδενική υπόθεση είναι αληθής και το μέγεθος του δείγματός ικανοποιητικά μεγάλο το στατιστικό -2lnL ακολουθεί X^2 κατανομή με k βαθμούς ελευθερίας , όπου k είναι η διαφορά μεταξύ του αριθμού των όρων του πλήρους μοντέλου της λογιστικής παλινδρόμησης από το μή πλήρες.
Τέλος εκτός από το στατιστικό του Wald και του λόγου της πιθανοφάνειας ως κριτήριο απομάκρυνσης μεταβλητών μπορεί να χρησιμοποιηθεί και ο έλεγχος με υπό
συνθήκη στατιστικό.
Ο έλεγχος αυτός, όπως ο έλεγχος του λόγου πιθανοφάνειας,βασίζεται στην διαφορά της πιθανοφάνειας μεταξύ πλήρους και μη πλήρους μοντέλων, με την διαφορά ότι ο έλεγχος με το υπό συνθήκη κριτήριο γίνεται με βάση τους δεδομένους εκτιμητές παραμέτρων.
Η τιμή του στατιστικού του υπολοίπου X^2 ελέγχει την μηδενική υπόθεση ότι οι συντελεστές των εκτός του μοντέλου μεταβλητών είναι μηδενικοί.
Εάν το επίπεδο στατιστικής σημαντικότητας είναι μικρό, δηλαδή αν πρέπει να απορριφθεί η μηδενική υπόθεση ότι όλοι οι συντελεστές είναι μηδενικοί είναι λογικό να συνεχίσουμε την διαδικασία με επιλογή μεταβλητών.
Εάν δεν μπορούμε να απορρίψουμε την υπόθεση ότι οι συντελεστες είναι μηδενικόι, η επιλογή μεταβλητών διακόπτεται.
Ο έλεγχος με κριτήριο υπό συνθήκη στατιστικού δεν είναι ιδιαίτερης υπολογιστικής δυσκολίας αφού δεν απαιτεί επανεκτίμηση του μοντέλου με την απουσία καθε μιας από τις μεταβλητές.

Γ. Προς τα Πίσω Κατά Στάδια Απάλειψη Μεταβλητών

Η προς τα πίσω κατα στάδια απάλειψη μεταβλητών στο μοντέλο (backward elimination), αντίθετα με την προοδευτική επιλογή μεταβλητών που ξεκινά χωρίς την παρουσία μεταβλητών στο μοντέλο, ξεκινά με όλες τις μεταβλητές στο μοντέλο και σταδιακά τις απομακρύνει με βάση κάποιο κριτήριο(κριτήριο εξόδου).
Όπως στην προοδευτική επιλογή έτσι και στην προς τα πίσω, σε κάθε στάδιο οι μεταβλητές ελέγχονται για είσοδο με το Score statistic και ελέγχονται για έξοδο με το στατιστικό του Wald ή του λόγου της πιθανοφάνειας ή του υπό συνθήκη στατιστικού.

10.Διαγνωστικά Καταλληλότητας του Μοντέλου

Όποτε δημιουργείται ένα μοντέλο πρέπει να ελέγχεται η αποτελεσματικότητα του μοντέλου αυτού να εκφράσει τα δεδομένα του. Στην λογιστική παλινδρόμηση για τον έλεγχο αυτό χρησιμοποιούνται διάφορα διαγνωστικά στατιστικά μέτρα και διαγράμματα.

10.1. Διαγνωστικές μέθοδοι

Α. Έλεγχος υπολοίπων

Το υπόλοιπο (ε_i) αποτελεί μέτρο της διαφοράς μεταξύ παρατηρούμενης και προβλεπόμενης πιθανότητας ενός ενδεχομένου.
Το τυποποιημένο υπόλοιπο Z που δίνεται από την σχέση

$$Z = \frac{\varepsilon_i}{\sqrt{P(1-P_i)}}$$

Θεωρείται στοιχείο του στατιστικού χ^2 της καλής προσαρμογής. Εάν το μέγεθος του δείγματος είναι μεγάλο τα τυποποιημένα υπόλοιπα ακολουθούν κατά προσέγγιση την κανονική κατανομή με μέσο όρο μηδέν και τυπική απόκλιση 1.
Ένα άλλο μέτρο διαφοράς είναι η απόκλιση παρατήρησης. Η απόκλιση συγκρίνει την προβλεπόμενη πιθανότητα που έχει μια παρατήρηση να είναι στη σωστή ομάδα, προς την άριστη πρόβλεψη που είναι ίση με 1 και υπολογίζεται από την σχέση:

$$\text{Απόκλιση} = -\sqrt{-2\log \hat{A}}$$

όπου Α είναι η προβλεπόμενη πιθανότητα για την εν λόγω ομάδα, ενώ το αρνητικό πρόσημο μπροστά από την τετραγωνική ρίζα τοποθετείται όταν το ενδεχόμενο για την συγκεκριμένη παρατήρηση δεν έχει συμβεί.

Όταν η τιμή της απόκλισης είναι μεγάλη φαίνεται ότι το μοντέλο δεν έιναι καλά προσαρμοσμένο στην παρατήρηση. Εάν το μέγεθος του δέιγματος έναι μεγάλο η απόκλιση ακολουθεί κατά προσέγγιση την κανονική κατανομή.
Το τυποποιημένο κατά Student υπόλοιπο παρατήρησης είναι η μεταβολή στην απόκλιση του μοντέλου έπειτα από τον αποκλισμό της παρατήρησης από το μοντέλο. Το τυποποιημένο κατά Student κατάλοιπο εξαρτάται από την απόσταση των τιμών των παρατηρήσεων των ανεξαρτήτων μεταβλητών από τους μέσους όρους των μεταβλητών αυτών. Εάν έχουμε διαφοροποιήσεις μεταξύ της απόλκλισης και του τυποποιημένου κατά Student υπόλοιπου τότε έχουμε και ασυνήθεις τιμές παρατηρήσεων.
Τέλος το υπόλοιπο των τυπικών μονάδων, εκφρασμένων σε λογαριθμική μορφή λόγου πιθανοτήτων, είναι το υπόλοιπο του μοντέλου που δίνεται από την σχέση:

$$(\text{logit } \varepsilon)_i = \frac{\varepsilon_i}{P_i(1-P_i)}$$

Β. <u>Έλεγχος ασυνήθων παρατηρήσεων</u>

Η συμβολή των τιμών των ανεξάρτητων μεταβλητών επί των προβλεπόμενων τιμών της εξαρτημένης μεταβλητής και ο εντοπισμός «ύποπτων» τιμών παρατηρήσεων γίνεται με την χρησιμοποίηση των παρακάτω στατιστικών μέτρων:

Το στατιστικό της τιμής μόχλευσης της παρατήρησης χρησιμοποιείται συχνά για τον εντοπισμό των παρατηρήσεων εκείνων που ασκούν μεγάλη επίδραση επί των προβλεπόμενων τιμών. Οι τιμές της μόχλευσης εξαρτώνται τόσο από τις τιμές της εξαρτημένης μεταβλητής όσο από τις τιμές των ανεξάρτητων μεταβλητών. Οι τιμές μόχλευσης των παρατηρήσεων είναι τα διαγώνια στοιχεία πίνακα Α που περιγράφουν την επίδραση της τιμής της εξαρτημένης μεταβλητής στην πρόβλεψη της τελικής τιμής της, κατά την σχέση:

$$\hat{Y} = AY$$

Οι τιμές μόχλευσης κυμαίνονται μεταξύ 0 και 1, με μέση τιμή k/n, όπου k είναι ο αριθμός των εκτιμώμενων παραμέτρων στο μοντέλο (συμπεριλαμβανόμενου και του σταθερού όρου) και n το μέγεθος του δείγματος.Τιμή ίση με μηδέν δηλώνει έλλειψη επίδρασης των παρατηρήσεων στην καλή προσαρμογή του μοντέλου λογιστικής παλινδρόμησης, ενώ τιμή ίση με 1 δηλώνει ακριβώς το αντίθετο. Είναι σκόπιμο όλες οι παρατηρήσεις να ασκούν την ίδια επίδραση, δηλαδή όλες οι τιμές μοχλευσης να είναι γύρω στην μέση τιμή k/n. Στην πράξη πρέπει να εξετάζονται για τυχόν απομάκρυνση από το μοντέλο λογιστικής παλινδρόμησης οι παρατηρήσεις με τιμές μόχλευσης μεγαλύτερες από την 2k/n. Μπορούμε να χρησιμοποιούμε ως σημείο αποδοχής και απόρριψης των τιμών μόχλευσης το 3k/n στις περιπτώσεις όπου k>6 και (n-k)>12. Πιό πρακτικά, τιμές μόχλευσης μεγαλύτερες του 0.5 πρέπει να αποφεύγονται, ενώ τιμές μικρότερες του 0.2 θεωρούνται ασφαλείς και τιμές μεταξύ 0.2 και 0.5 μπορεί να εμπεριέχουν κάποιο κίνδυνο.

Η απόσταση Cook είναι μέτρο επίδρασης της τιμής κάπιας παρατήρησης και προσδιορίζει το βαθμό της επίδρασης με την απομάκρυνση της παρατήρησης επί του υπολοίπου τόσο της ίδιας της παρατήρησης όσο και των καταλόιπων των υπολοίπων παρατηρήσεων. Η απόσταση Cook εξαρτάται από το τυποποιημένο υπόλοιπο για την δεδομένη παρατήρηση καθώς και το στατιστικό της τιμής μόχλευσης της παρατήρησης. Η απόσταση Cook δίνεται από την σχέση:

$$C_i = \frac{Z_i^2 * h_i}{(1 - h_i)^2}$$

Όπου z_i είναι το τυποποιημένο κατάλοιπο και h_i το στατιστικό της τιμής μόχλευσης.

Τέλος, άλλο ένα διαγνωστικό μέτρο είναι η μεταβολή των συντελεστών του μοντέλου της λογιστικής παλινδρόμησης, ύστερα από την απομάκρυνση μιας παρατήρησης. Το διαγνωστικό αυτό μέτρο το συμβολίζουμε ως DfBeta. Ο υπολογισμός της μεταβολής γίνεται για κάθε συντελεστή,συμπεριλαμβανομένου του σταθερού όρου. Έτσι η μεταβολή του συντελεστή β_i, όταν απομακρυνθεί η παρατήρηση i, είναι:

$$\text{DfBeta}=\beta_i - \beta_j^{(i)}$$

όπου β_i είναι η τιμή του συντελεστή όταν η παρατήρηση i απομακρυνθεί από το μοντέλο. Μεγάλες τιμές μεταβολής μας οδηγούν στην αναγνώριση παρατηρήσεων που πρέπει να επανεξεταστούν.

10.2.Διαγνωστικά Διαγράμματα

Με βάση τις τιμές των παραπάνω διαγνωστικών μέτρων μπορούμε, όπου είναι δυνατόν, να αποκτήσουμε διαγράμματα κανονικών κατανομών, όπως είναι το διάγραμμα των υπολοίπων απόκλισης από την κανονικότητα, το διάγραμμα της κανονικότητας κατανομής των αποκλίσεων με βάση τις πιθανότητες,το διάγραμμα κανονικότητας των τυποποιημένων υπολοίπων z, απεικονισμένων ως προς τη σειρά των παρατηρήσεων, και το διάγραμμα διασποράς των μεταβολών συντελεστή παλινδρόμησης απεικονισμένων ως προς τη σειρά των παρατηρήσεων.
Το διάγραμμα των υπολοίπων απόκλισης από την κανονικότητα βασίζεται στις αποκλίσεις μεταξύ των παρατηρούμενων και των θεωρητικών τιμών. Σύμφωνα με αυτό το διάγραμμα η συγκέντρωση των σημείων σε μία ζώνη εκατέρωθεν ευθείας γραμμής στο ύψος της μηδενικής τιμής δηλώνει την κανονικότητα της κατανομής των τιμών της εξαρτημένης μεταβλητής.
Στο διάγραμμα κανονικότητας της κατανομής των αποκλίσεων με βάση τις πιθανότητες εάν τα σημεία βρίσκονται σε ευθεία γραμμή η κατανομή των τιμών της εξαρτημένης μεταβλητής θα είναι κανονική.
Το διάγραμμα των τυποποιημένων υπολοίπων z απεικονισμένων ως προς τη σειρά των παρατηρήσεων μας βοηθά στον εντοπισμό παρατηρήσεων με υψηλές τιμές τυποποιημένων υπολοίπων.
Το διάγραμμα των τιμών μόχλευσης μας φανερώνει τις παρατηρήσεις με τις υψηλότερες τιμές του στατιστικού μόχλευσης.
Τέλος το διάγραμμα διασποράς των μεταβολών συντελεστή παλινδρόμησης μας βοηθά στον εντοπισμό των παρατηρήσεων που ασκούν σημαντική επίδραση κατά την εκτίμηση του συντελεστή παλινδρόμησης.

11. Εφαρμογή Λογιστικής παλινδρόμησης

Σε δείγμα 172 ασθενών τέθηκε το ερώτημα της γενικής ικανοποίησης από την χρήση κινητής μερικής οδοντοστοιχίας. Θα χρησιμοποιήσουμε τη μέθοδο λογιστικής παλινδρόμησης για την δημιουργία ενός μοντέλου πρόβλεψης της πιθανότητας γενικής ικανοποίησης, με εξαρτημένη μεταβλητή την γενική ικανοποίηση (ναι, όχι) και ανεξάρτητες μεταβλητές τις:

A_1 Φύλο (άνδρας,γυναίκα)
A_2 Ηλικία
A_3 Οικογενειάκη Κατάσταση (έγγαμος,άγαμος, Χήρος)
A_4 Παιδιά (όχι,ναι)
A_5 Μορφωτικό επίπεδο (δημοτικό, γυμνάσιο, λύκειο,τεχνικές σχολές, πανεπιστήμιο)
A_6 Οικονομική Κατάσταση (<500, '5000-1000', '1000-1500', '1500-3000', >3000)
A_7 Επάγγελμα (δημόσιος υπάλληλος,ιδιωτικός υπάλληλος, ελεύθερος επαγγελματίας, συνταξιούχος, άνεργος)
A_8 Συστηματική λήψη φαρμάκων (όχι, ναι)
A_9 Φορέας περίθλαψης (Οδοντιατρική σχολή, οδοντιατρική φρουρά, γενικό νοσοκομείο)
B_2 Προηγούμενη εμπειρία (όχι,ναι)
B_3 Παλαιότητα μερικής οδοντοστοιχίας (2002,2003,2004,2005,2006)
B_{16} Συχνότητα καθαρισμού την ημέρα (καμιά,'1','2','3','≥4)
$Γ_1$ Γνάθος (άνω, κάτω)
$Γ_2$ Κατηγορία νωδότητας κατά Kennendy ('1','2','3','4')
$Γ_3$ Ύπαρξη δεττερεύουσας νωδότητας (όχι,ναι)
$Γ_4$ Είδος ανταγωνιστών (όλα τα φυσικά δόντια,φυσικά δόντια και γέφυρες,σχεδόν όλα τα φυσικά δόντια, μερική οδοντοστοιχία, απώλεια οπίσθιων και όχι αντικατάσταση, ολική οδοντοστοιχία)
$Γ_5$ Αντικατάσταση πρόσθιων δοντιών (όχι, ναι)
$Γ_6$ Ακεραιότητα μερικής οδοντοστοιχίας (ακέραιη, θραύση αγκίστρου, θραύση εφαπτήρα, θραύση τεχνητών δοντιών, θραύση πτερυγίου, εκτεταμένες αποτριβές δοντιών)
$Γ_7$ Αριθμός δοντιών που αντικαθιστά η μερική οδοντοστοιχία

Δεδομένα

α/α	Μεταβλητές																			
	Y	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	B_2	B_3	B_{16}	$Γ_1$	$Γ_2$	$Γ_3$	$Γ_4$	$Γ_5$	$Γ_6$	$Γ_7$
1	2	1	70	1	1	5	3	4	0	2	1	3	1	1	2	1	2	0	1	6
2	2	1	58	1	1	4	3	3	0	1	1	3	1	2	1	0	6	0	4	5
3	2	2	64	2	0	1	2	4	0	1	1	4	2	2	2	0	6	0	1	3
4	1	1	52	1	1	2	3	2	0	1	1	4	0	2	1	1	2	0	1	5
5	2	2	58	1	1	1	3	1	0	1	1	3	2	1	1	0	3	0	1	6
6	1	2	61	1	1	3	1	5	1	1	1	4	1	2	2	1	3	0	1	5
7	2	2	64	1	1	1	3	4	0	1	1	3	3	2	2	1	2	0	1	5
8	2	2	61	1	1	1	3	5	0	1	0	3	2	2	1	0	2	0	1	6
9	1	2	53	1	1	1	3	5	0	1	1	4	2	2	2	1	4	0	1	3
10	2	2	53	1	1	1	3	5	0	1	1	4	2	1	3	1	4	1	1	8
11	1	2	36	2	0	2	1	5	0	1	1	4	1	2	2	1	2	0	1	5
12	1	1	70	1	1	5	3	4	1	2	1	3	1	1	2	1	4	1	1	8
13	2	1	70	1	1	5	3	4	1	2	0	3	1	2	2	1	4	1	6	5
14	2	1	78	1	1	5	4	4	1	2	0	3	2	2	1	1	6	1	1	11
15	2	2	72	3	1	1	3	5	1	2	0	5	2	2	2	0	4	1	1	6
16	1	2	66	1	1	2	3	5	1	2	1	3	4	2	1	0	2	0	6	7
17	2	1	70	3	1	5	4	4	1	2	1	5	3	1	3	1	4	1	1	5
18	2	1	70	3	1	5	4	4	1	2	0	5	3	2	1	1	4	0	1	5
19	1	1	65	1	1	5	3	1	1	2	0	4	2	1	1	0	1	0	1	6
20	2	2	81	3	1	2	2	5	1	2	0	3	1	1	3	1	4	1	1	9
21	2	2	81	3	1	2	2	5	1	2	0	4	1	2	1	0	4	0	1	8
22	2	1	82	1	1	5	4	4	1	2	0	1	2	1	1	1	6	1	1	6
23	2	1	73	1	1	5	3	4	0	2	1	4	2	2	1	0	6	0	1	8
24	2	1	70	1	1	5	4	4	1	2	1	1	3	1	1	1	4	1	4	10
25	2	1	70	1	1	5	4	4	1	2	1	1	3	2	2	1	4	1	6	8

26	2	2	66	1	1	3	3	5	0	2	1	4	3	2	1	0	2	0	1	5
27	2	2	56	1	1	4	4	3	0	2	1	5	3	2	1	0	3	1	1	7
28	1	2	68	3	1	3	2	5	1	2	0	4	2	1	2	1	4	1	1	8
29	2	2	68	3	1	3	2	5	1	2	0	5	2	2	1	0	4	0	1	8
30	2	1	54	1	1	5	4	1	0	2	1	2	1	2	2	1	2	0	1	8
31	2	2	61	1	1	3	4	1	0	2	0	5	2	1	1	0	2	1	1	10
32	2	1	68	1	1	5	4	4	0	2	0	3	1	2	2	1	3	1	1	7
33	2	2	68	1	1	1	3	5	1	2	0	4	3	1	1	0	4	0	1	8
34	2	2	68	1	1	1	3	5	1	2	1	1	3	2	2	1	4	0	1	3
35	2	1	67	1	1	5	3	4	1	2	1	1	1	2	1	1	2	1	1	5
36	2	1	65	1	1	5	3	1	0	2	0	1	3	1	2	1	2	0	6	7
37	1	1	76	1	1	5	3	4	1	2	1	5	2	1	1	0	4	0	1	8
38	1	2	73	1	1	1	2	5	1	2	1	5	0	2	1	0	2	0	1	7
39	1	1	51	1	1	5	3	1	0	2	1	5	1	1	1	0	2	0	1	8
40	2	1	76	1	1	5	3	4	1	2	1	1	1	1	2	1	4	0	1	5
41	2	1	76	1	1	5	3	4	1	2	0	5	1	2	2	0	4	0	1	4
42	2	2	78	3	1	3	3	5	1	2	0	3	2	2	1	0	5	0	1	6
43	2	2	65	1	1	1	4	5	1	2	1	3	3	2	3	1	2	1	1	7
44	1	1	73	1	1	5	4	4	1	2	0	3	0	2	1	0	1	0	2	7
45	2	2	64	1	1	2	3	5	1	2	1	1	3	1	2	1	2	0	1	4
46	2	1	71	1	1	5	3	4	1	2	1	1	2	1	2	1	4	1	1	11
47	2	1	71	1	1	5	3	4	1	2	1	1	2	2	1	1	4	1	6	7
48	2	1	72	1	1	5	3	4	0	2	1	3	3	1	2	0	1	0	1	4
49	2	2	73	3	1	1	3	5	1	2	1	1	3	1	1	0	4	0	1	6
50	2	2	73	3	1	1	3	5	1	2	0	1	3	2	2	0	4	0	1	3
51	2	1	76	1	1	5	3	4	1	2	0	4	3	2	1	0	5	0	1	7
52	2	1	70	1	1	5	3	4	1	2	1	4	4	2	1	0	6	0	1	6
53	1	2	75	3	1	1	2	5	1	1	0	4	2	2	1	0	5	0	1	6
54	1	2	60	1	1	1	2	5	0	1	1	4	4	1	1	0	2	0	1	8
55	2	1	59	1	1	5	4	1	1	1	1	4	1	2	2	0	2	0	1	3
56	1	1	73	1	1	1	2	4	0	1	1	3	3	1	1	0	4	0	2	7
57	1	1	73	1	1	1	2	4	0	1	1	3	3	2	2	1	4	0	1	4
58	2	1	64	1	1	1	2	1	0	1	1	3	2	2	1	0	6	0	1	8
59	1	2	51	1	1	2	3	5	0	1	1	4	0	1	2	0	4	0	1	4
60	2	2	51	1	1	2	3	5	0	1	1	4	3	2	1	0	4	0	1	6
61	1	1	62	1	1	5	3	1	1	1	1	3	0	1	1	0	3	1	2	7
62	2	2	70	1	1	1	1	4	1	1	0	3	2	1	2	0	4	0	1	4
63	2	2	70	1	1	1	1	4	1	1	0	3	2	2	2	0	4	0	1	4
64	2	1	71	1	1	3	2	4	0	1	0	3	1	1	1	0	4	0	1	8
65	1	1	71	1	1	3	2	4	0	1	0	3	1	2	1	1	4	1	2	11
66	2	1	49	1	1	4	3	2	0	1	1	4	2	1	1	0	3	1	2	10
67	1	1	53	1	1	4	2	2	0	1	1	3	0	1	3	1	2	0	1	4
68	1	1	62	2	0	1	2	3	1	1	1	3	0	2	1	0	3	0	1	6
69	1	1	68	1	1	5	4	4	0	1	1	4	2	2	1	0	3	0	1	2
70	2	1	80	1	1	4	3	4	1	3	1	5	2	1	1	0	4	0	1	4

71	2	1	80	1	1	4	3	4	1	3	0	2	2	2	2	1	4	1	1	11
72	2	1	59	1	1	4	3	1	1	3	1	4	3	1	1	0	2	0	1	7
73	1	1	59	1	1	4	3	4	0	3	1	3	2	2	1	0	2	0	1	8
74	2	1	75	1	1	4	3	4	1	3	1	4	1	2	3	1	3	0	3	4
75	1	1	55	1	1	4	3	4	0	3	1	3	2	1	2	1	5	1	1	9
76	1	1	75	1	1	5	4	4	1	3	0	1	2	1	2	1	3	1	1	10
77	2	2	52	1	0	3	3	1	0	3	0	5	2	2	1	0	2	0	1	7
78	2	1	62	1	1	5	3	4	0	3	1	3	2	1	3	1	4	1	2	7
79	2	1	62	1	1	5	3	4	0	3	1	3	2	2	3	1	4	0	2	5
80	2	1	59	1	1	5	4	4	0	3	1	4	2	1	1	0	2	0	1	8
81	2	1	72	3	1	5	4	4	1	3	1	4	3	1	1	0	1	0	1	5
82	2	1	70	1	1	4	3	4	1	3	1	3	4	1	1	0	4	0	2	6
83	2	1	70	1	1	4	3	4	1	3	1	1	4	2	1	0	4	0	2	7
84	2	1	73	1	1	5	4	4	1	3	1	5	1	2	3	0	1	0	1	3
85	2	1	69	1	1	5	4	4	1	3	0	4	2	1	1	1	6	1	1	10
86	1	1	66	1	1	5	4	4	1	3	1	4	3	2	3	1	2	0	1	4
87	2	1	76	1	1	3	3	4	1	3	1	4	2	2	1	0	6	0	1	8
88	2	1	65	1	1	4	3	4	1	3	1	2	2	2	3	1	2	1	1	8
89	2	1	67	1	1	5	4	4	1	3	1	3	1	2	2	0	2	0	1	2
90	2	1	72	1	1	5	4	4	1	3	1	3	3	1	1	0	4	0	1	6
91	2	1	72	1	1	5	4	4	1	3	1	3	3	2	3	1	4	1	2	7
92	1	1	77	1	0	5	4	4	1	3	0	4	3	2	1	1	6	1	1	9
93	1	1	51	1	1	4	3	1	0	3	1	4	1	1	1	0	2	0	1	5
94	2	1	77	1	1	4	3	4	1	3	1	2	3	1	1	1	4	0	1	8
95	2	1	77	1	1	4	3	4	1	3	1	2	3	2	1	1	4	1	1	5
96	2	1	82	1	1	5	3	4	0	3	1	2	2	1	1	1	4	1	1	12
97	2	1	82	1	1	5	3	4	0	3	1	2	2	2	2	1	4	1	1	10
98	2	1	53	1	1	4	3	1	1	3	1	1	2	2	2	1	2	0	3	3
99	2	1	67	1	1	4	4	4	1	3	1	3	3	1	1	1	4	1	1	7
100	2	1	67	1	1	4	4	4	1	3	0	3	3	2	2	1	4	1	1	9
101	2	1	66	1	1	5	4	4	1	3	1	2	2	2	2	1	2	0	1	3
102	2	1	67	1	1	4	3	4	0	3	1	3	2	2	1	1	6	1	1	8
103	2	1	57	1	1	5	4	4	1	3	1	3	1	1	2	0	4	0	1	4
104	2	1	57	1	1	5	4	4	1	3	1	4	1	2	1	0	4	0	1	5
105	2	1	44	1	1	4	3	1	1	3	1	2	2	2	2	1	1	0	1	3
106	2	1	75	1	1	5	4	4	1	3	1	2	4	1	2	1	2	0	2	7
107	2	1	69	1	1	5	4	4	1	3	0	3	2	2	1	0	2	1	1	8
108	2	1	74	1	1	5	4	4	1	3	0	1	4	2	1	0	6	0	2	6
109	2	1	62	1	1	5	4	4	0	3	1	3	2	1	3	1	2	1	1	9
110	2	2	59	1	1	5	4	1	0	3	1	3	3	2	1	0	2	0	1	8
111	2	1	74	1	1	5	4	4	0	3	1	5	3	2	3	1	6	1	1	7
112	2	1	70	1	1	4	4	4	1	3	1	3	1	1	1	1	4	0	1	7
113	2	1	70	1	1	4	4	4	1	3	1	3	1	2	1	1	4	1	1	7
114	2	1	57	1	1	5	4	1	1	3	1	4	4	2	2	1	2	1	1	10
115	2	1	61	1	1	5	4	4	0	3	1	3	2	1	1	0	2	0	1	5
116	1	1	72	1	1	5	4	4	1	3	0	5	3	1	2	0	4	0	1	3
117	1	1	72	1	1	5	4	4	1	3	1	1	3	2	2	0	4	0	1	3
118	2	1	67	1	1	4	4	4	1	3	0	4	2	1	1	0	4	1	1	9
119	2	1	67	1	1	4	4	4	1	3	1	4	2	2	1	0	4	0	1	8
120	2	1	61	1	1	4	3	4	1	3	1	5	1	1	1	0	4	0	1	7

121	2	1	61	1	1	4	3	4	1	3	1	5	1	2	3	1	4	0	1	6
122	1	1	55	1	1	3	3	4	1	3	0	5	2	2	2	1	6	0	1	6
123	2	1	69	1	1	2	3	2	1	1	1	4	4	1	1	1	4	1	1	14
124	1	1	69	1	1	2	3	2	1	1	0	4	4	2	1	0	4	1	1	9
125	1	1	65	1	1	1	2	4	0	1	1	3	1	1	2	1	4	1	1	5
126	1	1	65	1	1	1	2	4	0	1	1	3	1	2	2	1	4	0	1	4
127	1	2	50	1	1	2	1	5	1	1	1	4	.	2	1	0	2	0	1	6
128	1	2	60	1	1	1	1	4	1	1	1	3	1	1	2	1	4	0	1	4
129	2	2	60	1	1	1	1	4	1	1	1	3	1	2	1	1	4	1	1	8
130	2	1	73	1	1	3	4	3	0	1	1	4	1	2	1	1	1	1	1	9
131	2	2	72	1	1	1	2	5	1	1	1	3	3	2	1	1	2	1	1	9
132	2	1	40	2	0	5	2	3	1	1	1	4	2	1	3	1	2	1	1	11
133	2	1	72	1	1	5	3	4	1	1	1	1	1	1	1	0	4	0	1	7
134	1	1	72	1	1	5	3	4	1	1	1	4	2	2	1	0	4	0	1	5
135	2	2	71	1	1	1	1	5	1	1	1	4	4	2	2	1	2	0	1	4
136	2	2	66	1	1	2	2	4	1	1	1	1	3	2	1	0	6	0	2	6
137	1	1	60	1	1	1	2	1	0	1	1	1	0	2	1	0	6	0	1	5
138	2	2	55	1	1	1	1	5	0	1	1	1	4	2	2	1	2	0	2	6
139	2	2	58	1	1	1	1	5	1	1	1	3	3	2	3	1	3	1	1	6
140	2	1	74	1	1	1	2	4	1	1	1	4	1	2	3	1	4	1	1	3
141	2	1	46	1	1	1	2	2	0	1	1	4	1	1	1	0	4	0	1	5
142	2	1	46	1	1	1	2	2	0	1	1	4	1	2	3	1	4	0	5	6
143	2	2	50	1	1	3	1	5	0	1	1	4	2	2	2	1	2	0	1	3
144	1	1	69	1	1	1	1	4	0	1	1	3	0	2	3	1	3	0	1	5
145	2	1	44	1	1	5	4	1	0	1	1	3	3	1	1	0	3	0	1	6
146	1	2	43	1	1	1	2	5	0	1	1	3	3	2	1	0	1	0	1	4
147	2	2	51	1	1	3	4	3	0	1	1	3	2	2	1	0	2	0	1	3
148	2	1	75	1	0	3	4	4	1	1	1	5	2	2	1	1	4	1	1	5
149	.	2	67	1	1	5	3	4	1	1	.	.	.	1	1	0	4	0	1	5
150	2	2	67	1	1	5	3	4	1	1	1	1	3	2	1	0	4	0	1	4
151	2	2	69	1	1	3	2	5	1	2	0	5	2	2	3	1	2	0	1	6
152	2	1	73	1	1	5	3	4	0	2	1	1	1	1	1	0	6	0	2	8
153	2	2	52	1	1	2	2	5	1	2	1	1	1	1	1	0	2	0	1	8
154	2	1	68	1	1	5	4	4	1	2	1	5	2	2	1	0	2	0	1	7
155	2	2	60	1	1	1	2	5	0	2	1	1	1	1	2	0	4	0	1	4
156	1	2	60	1	1	1	2	5	0	2	1	1	1	2	1	0	4	0	1	6
157	2	2	75	3	1	5	3	4	1	2	1	2	2	1	1	0	5	0	1	8
158	2	2	74	3	1	3	3	5	1	2	0	5	2	1	2	1	6	1	1	10
159	2	2	75	3	1	1	2	4	0	2	1	1	1	1	1	0	4	0	6	5
160	2	2	75	3	1	1	2	4	0	2	1	3	1	2	1	0	4	0	1	5
161	2	2	72	1	1	1	3	5	1	2	0	4	2	1	1	0	2	1	1	10
162	2	2	59	3	1	1	3	5	1	2	0	5	3	2	1	1	2	1	1	8
163	2	1	63	1	1	4	3	4	0	2	0	5	1	2	3	1	6	1	1	8
164	2	2	59	1	1	1	2	4	1	1	1	4	2	2	1	0	2	0	1	7
165	2	1	71	1	1	4	2	4	1	1	0	5	3	2	1	0	2	0	1	8

166	1	2	75	1	1	1	2	4	1	1	1	3	1	1	1	0	4	0	1	8
167	1	2	75	1	1	1	2	4	1	1	1	4	1	2	1	0	4	0	1	4
168	1	2	71	1	1	1	2	5	1	1	1	3	2	2	2	0	2	0	1	4
169	2	1	71	1	1	1	2	4	1	1	1	1	3	1	3	1	4	0	3	5
170	2	1	71	1	1	1	2	4	1	1	1	3	3	2	3	1	4	0	1	5
171	2	2	60	1	1	1	2	5	0	1	1	4	.	2	1	0	2	0	1	4
172	1	1	71	1	1	4	2	4	1	1	0	4	3	2	1	0	2	0	1	8

Τα αποτελέσματα που πήραμε από την ανάλυση λογιστικής παλινδρόμησης εμφανίζονται παρακάτω:

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	169	98,3
	Missing Cases	3	1,7
	Total	172	100,0
Unselected Cases		0	,0
Total		172	100,0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
1,00	0
2,00	1

		Frequency	Parameter coding				
			(1)	(2)	(3)	(4)	(5)
ΕΞΕΤΑΣΗ Μ.Ο.	ΑΚΕΡΑΙΗ	142	1,000	,000	,000	,000	,000
	ΘΡΑΥΣΗ ΑΓΚΙΣΤΡΟΥ	15	,000	1,000	,000	,000	,000
	ΘΡΑΥΣΗ ΕΦΑΠΤΗΡΑ	3	,000	,000	1,000	,000	,000
	ΘΡΑΥΣΗ ΤΕΧΝΗΤΩΝ ΔΟΝΤΙΩΝ	2	,000	,000	,000	1,000	,000
	ΘΡΑΥΣΗ ΠΤΕΡΥΓΙΟΥ	1	,000	,000	,000	,000	1,000
	ΕΚΤΕΤΑΜΕΝΕΣ ΑΠΟΤΡΙΒΕΣ ΔΟΝΤΙΩΝ	6	,000	,000	,000	,000	,000
ΠΟΙΑ Η ΚΑΤΑΣΤΑΣΗ ΤΩΝ ΑΝΤΑΓΩΝΙΣΤΩΝ;	ΟΛΑ ΤΑ ΦΥΣΙΚΑ ΔΟΝΤΙΑ	8	1,000	,000	,000	,000	,000
	ΦΥΣΙΚΑ ΔΟΝΤΙΑ ΚΑΙ ΓΕΦΥΡΕΣ	49	,000	1,000	,000	,000	,000
	ΣΧΕΔΟΝ ΟΛΑ ΤΑ ΦΥΣΙΚΑ ΔΟΝΤΙΑ ΚΑΙ ΟΧΙ ΑΝΤΙΚΑΤΑΣΤΑΣΗ	13	,000	,000	1,000	,000	,000
	ΑΥΤΩΝ ΠΟΥ ΧΑΘΗΚΑΝ Μ.Ο.	75	,000	,000	,000	1,000	,000
	ΑΠΩΛΕΙΑ ΟΠΙΣΘΙΩΝ ΚΑΙ ΟΧΙ ΑΝΤΙΚΑΤΑΣΤΑΣΗ	5	,000	,000	,000	,000	1,000
	ΟΛΙΚΗ ΟΔΟΝΤΟΣΤΟΙΧΙΑ	19	,000	,000	,000	,000	,000
ΠΟΣΕΣ ΦΟΡΕΣ ΚΑΘΑΡΙΖΕΤΕ ΤΗ Μ.Ο ΚΑΤΑ ΤΗ ΔΙΑΡΚΕΙΑ ΤΗΣ ΗΜΕΡΑΣ.;	ΚΑΜΙΑ	9	1,000	,000	,000	,000	
	1	45	,000	1,000	,000	,000	
	2	59	,000	,000	1,000	,000	
	3	44	,000	,000	,000	1,000	
	>4	12	,000	,000	,000	,000	
ΜΟΡΦΩΤΙΚΟ ΕΠΙΠΕΔΟ	ΔΗΜΟΤΙΚΟ	47	1,000	,000	,000	,000	
	ΓΥΜΝΑΣΙΟ	12	,000	1,000	,000	,000	
	ΛΥΚΕΙΟ	17	,000	,000	1,000	,000	
	ΤΕΧΝΙΚΕΣ ΣΧΟΛΕΣ	30	,000	,000	,000	1,000	
	ΠΑΝΕΠΙΣΤΗΜΙΟ	63	,000	,000	,000	,000	
ΕΠΑΓΓΕΛΜΑ	ΔΗΜΟΣΙΟΣ ΥΠΑΛΛΗΛΟΣ	18	1,000	,000	,000	,000	
	ΙΔΙΩΤΙΚΟΣ ΥΠΑΛΛΗΛΟΣ	7	,000	1,000	,000	,000	
	ΕΛΕΥΘΕΡΟΣ ΕΠΑΓΓΕΛΜΑΤΙΑΣ	6	,000	,000	1,000	,000	
	ΣΥΝΤΑΞΙΟΥΧΟΣ	100	,000	,000	,000	1,000	
	ΑΝΕΡΓΟΣ/ΑΝΕΠΑΓΓΕΛΤΟΣ	38	,000	,000	,000	,000	
ΟΙΚΟΝΟΜΙΚΗ ΚΑΤΑΣΤΑΣΗ	<500	11	1,000	,000	,000		
	500-1000	39	,000	1,000	,000		
	1000-1500	71	,000	,000	1,000		
	1500-3000	48	,000	,000	,000		
ΟΙΚΟΓΕΝΕΙΑΚΗ ΚΑΤΑΣΤΑΣΗ	ΕΓΓΑΜΟΣ/Η	148	1,000	,000			
	ΑΓΑΜΟΣ/Η	4	,000	1,000			
	ΧΗΡΟΣ/Α	17	,000	,000			
ΚΑΤΑΣΚΕΥΗ ΣΕ	ΟΔΟΝΤΙΑΤΡΙΚΗ ΣΧΟΛΗ	61	1,000	,000			
	ΟΔΟΝΤΙΑΤΡΕΙΟ ΦΡΟΥΡΑΣ	55	,000	1,000			

	ΓΕΝΙΚΟ ΝΟΣΟΚΟΜΕΙΟ ΦΡΟΥΡΑΣ	53	,000	,000			
ΚΑΤΗΓΟΡΙΟΠΟΙΗΣΗ ΚΑΤΑ ΚΕΝΝΕDY	I	95	1,000	,000			
	II	50	,000	1,000			
	III	24	,000	,000			
ΓΝΑΘΟΣ ΜΕ Μ.Ο.	ΑΝΩ	66	1,000				
	ΚΑΤΩ	103	,000				
ΕΙΝΑΙ Η ΣΥΓΚΕΚΡΙΜΕΝΗ Μ.Ο. Η ΠΡΩΤΗ ΠΟΥ ΦΟΡΑΤΕ;	ΟΧΙ	44	1,000				
	ΝΑΙ	125	,000				
ΠΑΙΔΙΑ	ΟΧΙ	7	1,000				
	ΝΑΙ	162	,000				
ΑΝΤΙΚΑΘΙΣΤΑ Η Μ.Ο. ΠΡΟΣΘΙΑ ΔΟΝΤΙΑ;	ΟΧΙ	114	1,000				
	ΝΑΙ	55	,000				
ΥΠΑΡΞΗ ΔΕΥΤΕΡΕΥΟΥΣΑΣ ΝΩΔΟΤΗΤΑΣ(ΥΠΟ ΚΑΤΗΓΟΡΙΑ)	ΟΧΙ	88	1,000				
	ΝΑΙ	81	,000				
ΣΥΣΤΗΜΑΤΙΚΗ ΛΗΨΗ ΦΑΡΜΑΚΩΝ	ΟΧΙ	56	1,000				
	ΝΑΙ	113	,000				
ΧΡΟΝΙΑ	ΟΧΙ	62	1,000				
ΝΟΣΗΜΑΤΑ	ΝΑΙ	107	,000				
ΦΥΛΟ	ΑΡΡΕΝ	111	1,000				
	ΘΗΛΥ	58	,000				

Block 0: Beginning Block

Iteration History(a,b,c)

Iteration		-2 Log likelihood	Coefficients Constant
Step 0	1	194,060	,959
	2	193,819	1,042
	3	193,819	1,044
	4	193,819	1,044

a Constant is included in the model.
b Initial -2 Log Likelihood: 193,819
c Estimation terminated at iteration number 4 because parameter estimates changed by less than ,001.

Classification Table[a,b]

Observed			Predicted		
			Γενική Ικανοποίηση 2		Percentage Correct
			1,00	2,00	
Step 0	Γενική Ικανοποίηση 2	1,00	0	44	,0
		2,00	0	125	100,0
	Overall Percentage				74,0

a. Constant is included in the model.

b. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	1,044	,175	35,480	1	,000	2,841

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	A1(1)	,001	1	,970
		A2	2,008	1	,156
		A3	3,066	2	,216
		A3(1)	,608	1	,436
		A3(2)	1,222	1	,269
		A4(1)	1,073	1	,300
		A5	6,406	4	,171
		A5(1)	3,473	1	,062
		A5(2)	1,639	1	,200
		A5(3)	,062	1	,804
		A5(4)	1,663	1	,197
		A6		3	,005
		A6(1)	,652	1	,419
		A6(2)	10,656	1	,001
		A6(3)	1,532	1	,216
		A7	2,415	4	,660
		A7(1)	,032	1	,859
		A7(2)	1,073	1	,300
		A7(3)	,284	1	,594
		A7(4)	1,172	1	,279
		A8_1(1)	1,969	1	,161
		A8_2(1)	2,710	1	,100
		A9	13,644	2	,001
		A9(1)	13,638	1	,000
		A9(2)	3,961	1	,047
		Г1(1)	,004	1	,947
		Г2	3,211	2	,201
		Г2(1)	,009	1	,925
		Г2(2)	1,312	1	,252
		Г3(1)	1,175	1	,278
		Г4	5,043	5	,411
		Г4(1)	,573	1	,449
		Г4(2)	,086	1	,770
		Г4(3)	2,960	1	,085
		Г4(4)	,290	1	,590
		Г4(5)	,522	1	,470
		Г5(1)	3,961	1	,047
		Г6	2,539	5	,771
		Г6(1)	,943	1	,332
		Г6(2)	,003	1	,953
		Г6(3)	1,075	1	,300
		Г6(4)	,712	1	,399
		Г6(5)	,354	1	,552
		Г7	2,307	1	,129
		B3	2,338	1	,126
		B2_1(1)	,033	1	,856
		B2_2	,228	1	,633

	B16	27,988	4	,000
	B16(1)	27,006	1	,000
	B16(2)	,013	1	,910
	B16(3)	1,527	1	,216
	B16(4)	1,905	1	,167
Overall Statistics		65,914	43	,014

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	80,483	41	,000
	Block	80,483	41	,000
	Model	80,483	41	,000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	113,336(a)	,379	,555

a Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	10,823	8	,212

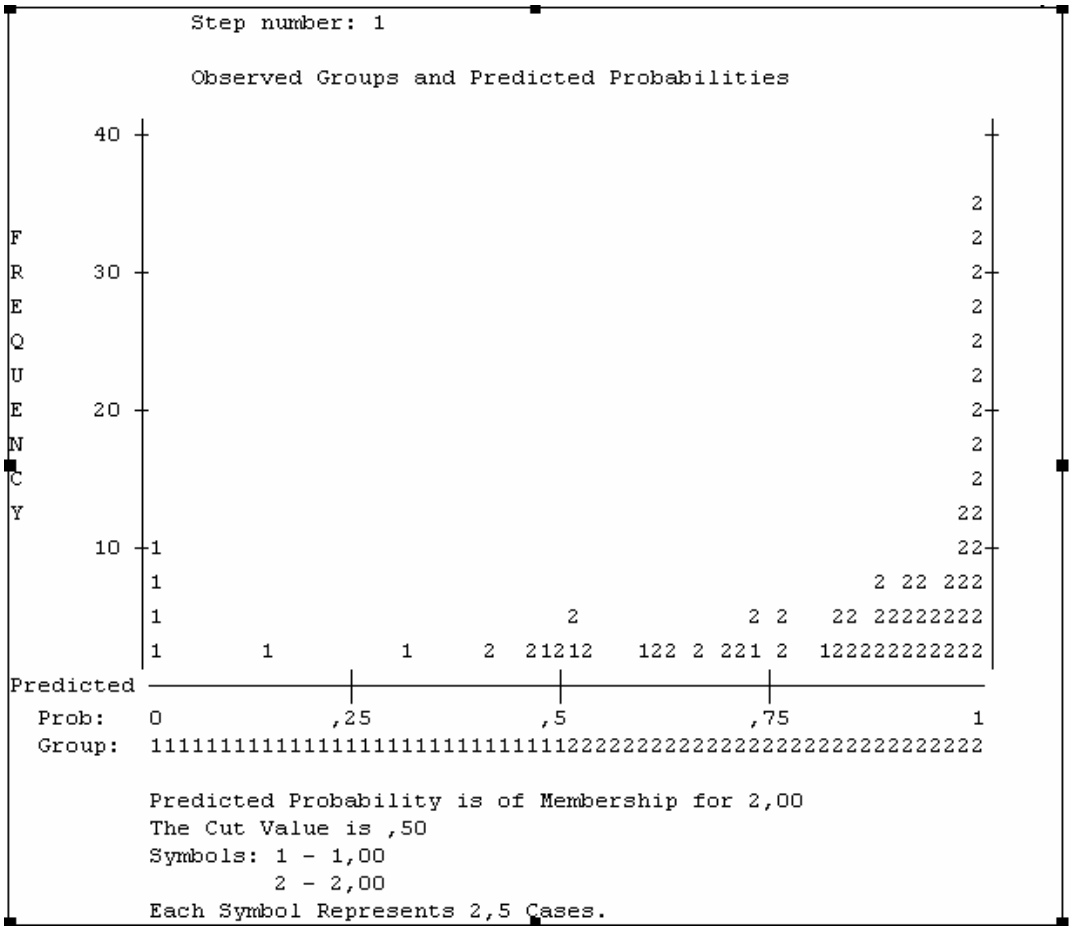
Contingency Table for Hosmer and Lemeshow Test

		Γενική Ικανοποίηση 2 = 1,00		Γενική Ικανοποίηση 2 = 2,00		
		Observed	Expected	Observed	Expected	Total
Step 1	1	17	15,727	0	1,273	17
	2	10	10,118	7	6,882	17
	3	6	7,365	11	9,635	17
	4	3	4,816	14	12,184	17
	5	5	2,715	12	14,285	17
	6	1	1,686	16	15,314	17
	7	1	,990	16	16,010	17
	8	1	,454	16	16,546	17
	9	0	,124	17	16,876	17
	10	0	,007	16	15,993	16

Πίνακας Ταξινόμηση[a]

			Predicted		
			Γενική Ικανοποίηση		Percentage
	Observed		όχι	ναι	Correct
Step 1	Γενική Ικανοποίηση	όχι	27	17	61,4
	2	ναι	8	117	93,6
	Overall Percentage				85,2

a. The cut value is ,500



Στον πρώτο πίνακα των αποτελεσμάτων δίνεται ο αριθμός των έγκυρων παρατηρήσεων της ανάλυσης καθώς και ο αριθμός των παρατηρήσεων με missing values. Στον επόμενο πίνακα δίνεται η εσωτερική κωδικοποίηση των κατηγοριών της εξαρτημένης μεταβλητής κατά την ανάλυση. Ο μικρότερος κωδικός της εξαρτημένης μεταβλητής επανακωδικοποιείται εσωτερικά με την τιμή 0 ενώ ομεγαλύτερος με την τιμή 1 (δηλαδή με την τιμή που υποδηλώνει την πραγματοποίηση του γεγονότος). Στην περίπτωση αυτή, η πραγματοποίηση του γεγονότος (δηλαδή η απάντηση *ναι* στην γενική ικανοποίηση) είναι κωδικοποιημένη στο αρχείο των δεδομένων με την τιμή 2 άρα επανακοδικοποιείται με την τιμή 1. Αντίστοιχα, η μη πραγματοποίηση του γεγονότος (η απάντηση *όχι* στην ερώτηση γενικής ικανοποίησης) επανακωδικοποιείται με την τιμή 0. Στον τρίτο πίνακα των αποτελεσμάτων δίνεται η κωδικοποίηση των κατηγορικών ανεξάρτητων μεταβλητών.Ως κατηγορία αναφοράς στις κατηγορικές μεταβλητές επιλέξαμε την κατηγορία με τον μικρότερο κωδικό.
Οι τρεις πίνακες υπό τον τίτλο Block0: Beginning Block αφορούν το αρχικό μοντέλο της ανάλυσης το οποίο αποτελείται μόνο από το σταθερό όρο χωρίς άλλη ανεξάρτητη μεταβλητήστην εξίσωση παλινδρόμησης. Δίνεται ο πίνακας ταξινόμησης, η τιμή του σταθερού όρου και οι αντίστοιχοι έλεγχοι, καθώς και η αξιολόγηση των μεταβλητών που δεν έχουν εισέλθει ακόμα στο μοντέλο. Δηλαδή η σημαντικότητα κάθε μιας από τις ανεξάρτητες μεταβλητές αν έμπαινε μόνη της στο μοντέλο μαζί με το σταθερό όρο. Το κριτήριο βάση του οποίου γίνεται ο έλεγχος (Score) δίνει την βαρύτητα κάθε ανεξάρτητητης μεταβλητής στην πρόγνωση των τιμών της εξαρτημένης. Με μία πρώτη ματιά στις τιμές του Score αλλά και των αντίστοιχων πιθανοτήτων (sig), προκύπτει ότι τη μεγαλύτερη βαρύτητα στην πρόγνωση της εξαρτημένης μεταβλητής έχει η συχνότητα καθαρισμού, ο φορέας περίθαλψης και η οικονομική κατάστση των ασθενών.

Οι επόμενοι πίνακες μετά τον τίτλο Block1: Method=Enter αφορούν την μορφή του τελικού μοντέλου, καθώς και την αξιολόγησή του.

11.1.Συντελεστές Παραμέτρων

Οι συντελεστές παραμέτρων του μοντέλου της λογιστικής παλινδρόμησης που υπολογίστηκαν με την χρήση της τεχνικής της ταυτόχρονης εισόδου των ανεξάρτητων μεταβλητών (ως μπλόκ) στην εξίσωση παλινδρόμησης, οι επαγωγικοί έλεγχοι και τα διαστήματα εμπιστοσύνης τους, ύστερα από 20 επαναλήψεις δίνονται στον πίνακα

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95,0% C.I.for EXP(B)	
								Lower	Upper
Step 1	A1(1)	-1,352	1,355	,996	1	,318	,259	,018	3,680
	A2	,051	,044	1,332	1	,249	1,052	,965	1,146
	A3			2,822	2	,244			
	A3(1)	-1,342	1,472	,832	1	,362	,261	,015	4,675
	A3(2)	2,134	3,395	,395	1	,530	8,445	,011	6549,754
	A4(1)	-2,883	1,934	2,222	1	,136	,056	,001	2,479
	A5			6,289	4	,179			
	A5(1)	1,188	1,480	,644	1	,422	3,279	,180	59,599
	A5(2)	-,157	1,593	,010	1	,921	,855	,038	19,402
	A5(3)	1,918	1,426	1,809	1	,179	6,805	,416	111,251
	A5(4)	2,496	1,124	4,935	1	,026	12,133	1,342	109,739
	A6			10,966	3	,012			
	A6(1)	-2,454	1,810	1,838	1	,175	,086	,002	2,985
	A6(2)	-5,572	1,827	9,305	1	,002	,004	,000	,136
	A6(3)	-2,485	1,212	4,203	1	,040	,083	,008	,897
	A7			3,260	4	,515			
	A7(1)	1,662	1,378	1,455	1	,228	5,269	,354	78,407
	A7(2)	3,637	2,089	3,032	1	,082	37,981	,633	2278,580
	A7(3)	19,771	8269,424	,000	1	,998	4E+008	,000	.
	A7(4)	,911	1,177	,600	1	,439	2,488	,248	24,974
	A8_1(1)	,717	2,256	,101	1	,751	2,048	,025	170,368
	A8_2(1)	-,492	2,284	,046	1	,830	,612	,007	53,739
	A9			4,954	2	,084			
	A9(1)	1,262	1,145	1,215	1	,270	3,534	,374	33,363
	A9(2)	2,313	1,040	4,945	1	,026	10,103	1,316	77,579
	Γ1(1)	-,429	,626	,470	1	,493	,651	,191	2,220
	Γ2			7,153	2	,028			
	Γ2(1)	-3,073	1,524	4,069	1	,044	,046	,002	,917
	Γ2(2)	-3,589	1,351	7,058	1	,008	,028	,002	,390
	Γ3(1)	,623	,928	,451	1	,502	1,865	,303	11,503
	Γ4			7,038	5	,218			
	Γ4(1)	-1,942	1,703	1,300	1	,254	,143	,005	4,041
	Γ4(2)	-1,236	1,344	,846	1	,358	,291	,021	4,045
	Γ4(3)	-4,133	1,883	4,817	1	,028	,016	,000	,643
	Γ4(4)	-1,743	1,316	1,756	1	,185	,175	,013	2,305
	Γ4(5)	-2,824	1,770	2,545	1	,111	,059	,002	1,907
	Γ5(1)	-,286	,989	,084	1	,772	,751	,108	5,217
	Γ6			,345	5	,997			
	Γ6(1)	,645	1,541	,175	1	,676	1,905	,093	39,018
	Γ6(2)	1,061	1,812	,342	1	,558	2,888	,083	100,745
	Γ6(3)	19,584	22538,056	,000	1	,999	3E+008	,000	.
	Γ6(4)	15,149	23489,019	,000	1	,999	3793248	,000	.
	Γ6(5)	19,248	40192,970	,000	1	1,000	2E+008	,000	.
	Γ7	,075	,204	,135	1	,713	1,078	,722	1,609
	B3	-,751	,308	5,956	1	,015	,472	,258	,863
	B2_1(1)	-2,212	1,688	1,717	1	,190	,110	,004	2,992
	B2_2	1,616	1,369	1,393	1	,238	5,033	,344	73,653
	B16			,859	4	,930			
	B16(1)	-32,960	12408,531	,000	1	,998	,000	,000	.
	B16(2)	1,129	1,391	,658	1	,417	3,091	,202	47,225
	B16(3)	,827	1,266	,427	1	,513	2,287	,191	27,331
	B16(4)	1,115	1,287	,751	1	,386	3,050	,245	38,011
	Constant	3,418	4,947	,477	1	,490	30,508		

a. Variable(s) entered on step 1: A1, A2, A3, A4, A5, A6, A7, A8_1, A8_2, A9, Γ1, Γ2, Γ3, Γ4, Γ5, Γ6, Γ7, B3, B2_1, B2_2, B16.

Με βάση το στατιστικό του Wald, σημαντική επίδραση στη διαμόρφωση των τιμών της εξαρτημένης μεταβλητής έχουν οι μεταβλητές:

A_6: Οικονομική κατάσταση
$Γ_2$: Κατηγορία Νωδότητας κατά Kenedy
B_3:Παλαιότητα μερικής οδοντοστοιχίας

11.2. Εκτίμηση της καλής Προσαρμογής του Μοντέλου

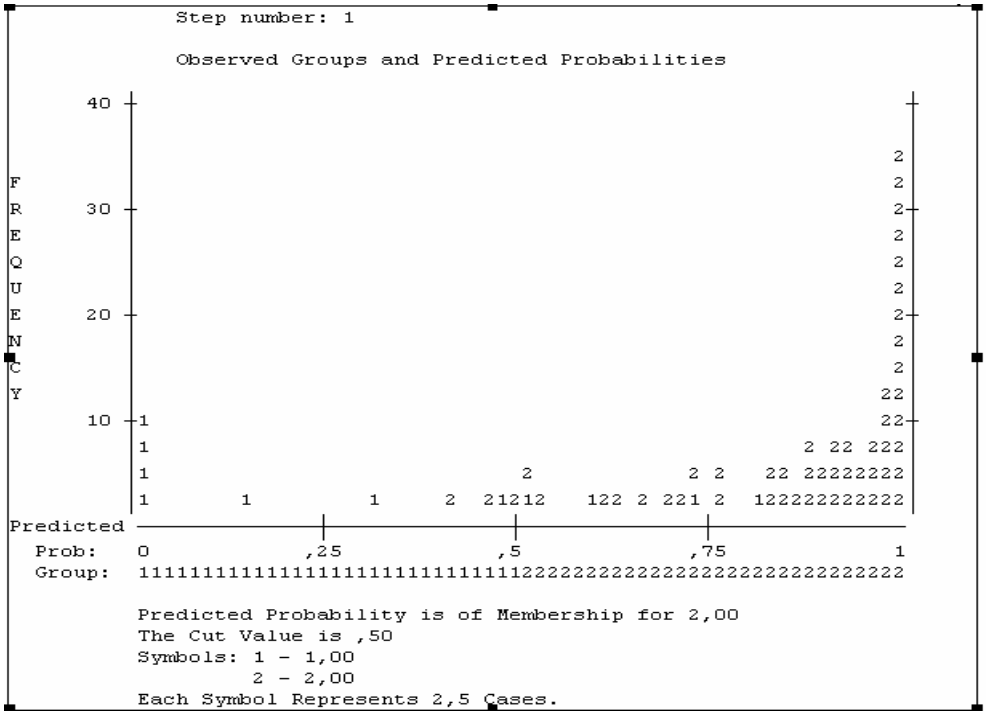
Σύμφωνα με τον πίνακα ταξιμόμησης των παρατηρήσεων ως προς την γενική ικανοποιήση των ασθενών, διαπιστώνεται ότι 27 ασθενείς (ποσοστό 61.4%) που δεν είναι ικανοποιημένοι προβλέθηκαν ορθώς από το μοντέλο να μην είναι ικανοποιημένοι. Ομοίως 118 από τους ασθενείς (ποσοστό 94.4%) που είναι ικανοποιημένοι προβλέθηκαν ορθώς από το μοντέλο να είναι ικανοποιημένοι. Μόνο 24(14%) ασθενείς ταξινομήθηκαν λανθασμένα, γεγονός που οδηγεί στο συμπέρασμα ότι το μοντέλο λογιστικής παλινδρόμησης είναι πολύ καλά προσαρμοσμένο στα δεδομένα(συνολική ορθή πρόβλεψη 86%).

Classification Table[a]

Observed			Predicted		
			Γενική Ικανοποίηση 2		Percentage Correct
			1,00	2,00	
Step 1	Γενική Ικανοποίηση 2	1,00	27	17	61,4
		2,00	7	118	94,4
	Overall Percentage				85,8

a. The cut value is ,500

Από το διάγραμμα απεικόνισης των εκτιμώμενων πιθανοτήτων διαπιστώνεται ότι το μοντέλο της λογιστικής παλινδρόμησης διαχωρίζει επιτυχώς τις δύο ομάδες των ασθενών
(κωδικός 1: μη ικανοποιημένοι, κωδικός 2: ικανοποιημένοι), τοποθεντώντας κάθε μία στα άκρα του διαγράμματος (εμφανίζουν υψηλή πιθανότητα). Εξαίρεση αποτελούν 5 παρατηρήσεις που ενώ οι ασθενείς δηλώνουν ικανοποιημένοι εντάσσονται σε εκείνους που δεν είναι ικανοποιημένοι και 10 ασθενείς οι οποίοι ενώ δηλώνουν μη ικανοποιημένοι εντάσονται στους ικανοποιημένους. Όμως η εκτιμώμενη πιθανότητα και για αυτές τις περιπτώσεις είναι σχετικά χαμηλή.



Η αξιολόγηση της καλής προσαρμογής του μοντέλου λογιστικής παλινδρόμησης στα δειγματικά δεδομένα γίνεται επίσης και με τον λόγο των μέγιστών τιμών της συνάρτησης πιθανοφάνειας (likelihood ratio statistic) για το πλήρες μοντέλο (L_F) και το μοντέλο που περιλαμβάνει μόνο το σταθερό όρο (L_0). Η τιμή του λόγου είναι:

$$-2\ln\left(\frac{L_0}{L_F}\right) = 82{,}098 \text{ (Model Chi-square)}$$

Ενώ η πιθανότητα να προκύψει μια τόσο μεγάλη τιμή για την κατανομή X^2 με 44 βαθμούς ελευθερίας είναι Sig.< 0.0005. Επομένως η μηδενική υπόθεση H_0: $\beta_1=\beta_2=...=\beta_{44}=0$ απορρίπτεται. Δηλαδή οι ανεξάρτητες μεταβλητές συμβάλλουν σημαντικά στην πρόγνωση των τιμών της εξαρτημένης μεταβλητής.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	82,098	43	,000
	Block	82,098	43	,000
	Model	82,098	43	,000

Στον επόμενο πίνακα (Model Summary) δίνεται η τιμή της συνάρτησης λογαριθμο-πιθανοφάνειας (-2Log likelihood=111.721) για το τελικό μοντέλο μαζι με το συντελεστή προσδιορισμού των Cox & Snell (0.385) και το συντελεστή προσδιορισμού του Nagelkerke (0.564). Περίπου δηλαδή 56% της μεταβλητότητας της εξαρτημένης μεταβλητής ερμηνεύονται από τις ανεξάτητες μεταβλητές του μοντέλου.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	111,721(a)	,385	,564

11.3.Επιλογή των ανεξάρτητων μεταβλητών

Το μοντέλο λογιστικής παλινδρόμησης, για το παράδειγμα που χρησιμοποιήθηκε, εφαρμόστηκε θεωρώντας ότι οι ανεξάρτητες μεταβλητές εισήλθαν στο μοντέλο ταυτόχρονα.
Στη συνέχεια εφαρμόζουμε την μέθοδο της προοδευτική επιλογής με κριτήριο απομάκρυνσης μεταβλητών το κριτήριο του λόγου πιθανοφάνειας.

Variables in the Equation

							95,0% C.I.for EXP(B)	
	B	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
A6(οικονομική Κατάσταση)			11,154	3	,011			
A6(<500)	--,563	,901	,392	1	,532	,569	,097	3,325
A6(500-1000)	-1,848	,607	9,260	1	,002	,158	,048	,518
A6(1000-1500)	-,403	,565	,508	1	,476	,668	,221	2,022
Γ2(κατηγορία Νοδώτητας Kenedy			8,319	2	,016			
Γ2(1)	-2,171	1,091	3,958	1	,047	,114	,013	,968
Γ2(2)	-3,046	1,137	7,178	1	,007	,048	,005	,441
B3(Παλαιότητα)	-,448	,187	5,742	1	,017	,639	,443	,922
B16(Συχνότητα Καθαρισμού)			,449	4	,978			
B16(1)	-23,006	12190,993	,000	1	,998	,000	,000	.
B16(2)	,167	,817	,042	1	,838	1,182	,238	5,857
B16(3)	,447	,806	,307	1	,579	1,564	,322	7,595
B16(>4)	,273	,845	,104	1	,747	1,314	,251	6,888
Constant	5,479	1,557	12,386	1	,000	239,559		

Συνεπώς η εξίσωση παλινδρόμησης θα είναι:

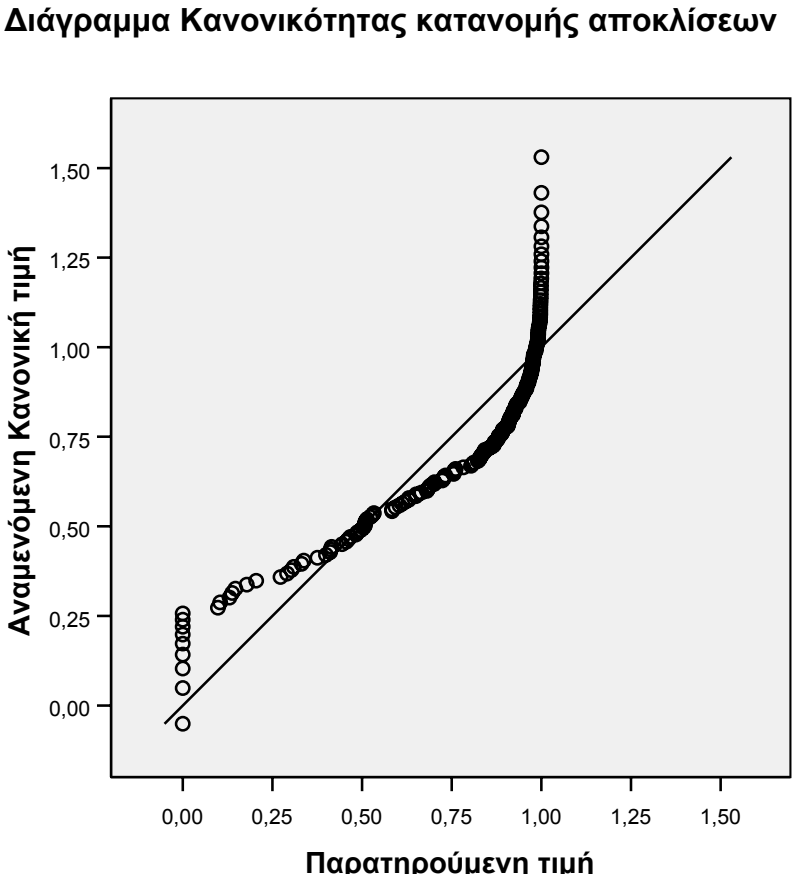
Z=5.479-0.563A6(1)-1.848A6(2)–0.403A6(3)–2.171Γ2(1)–3.046Γ2(2)–0.448B3

- 23.006B16(1)+0.167B16(2)+0.447B16(3)+0.273B16(4)

$$P=\frac{e^{Z}}{1+e^{Z}}$$

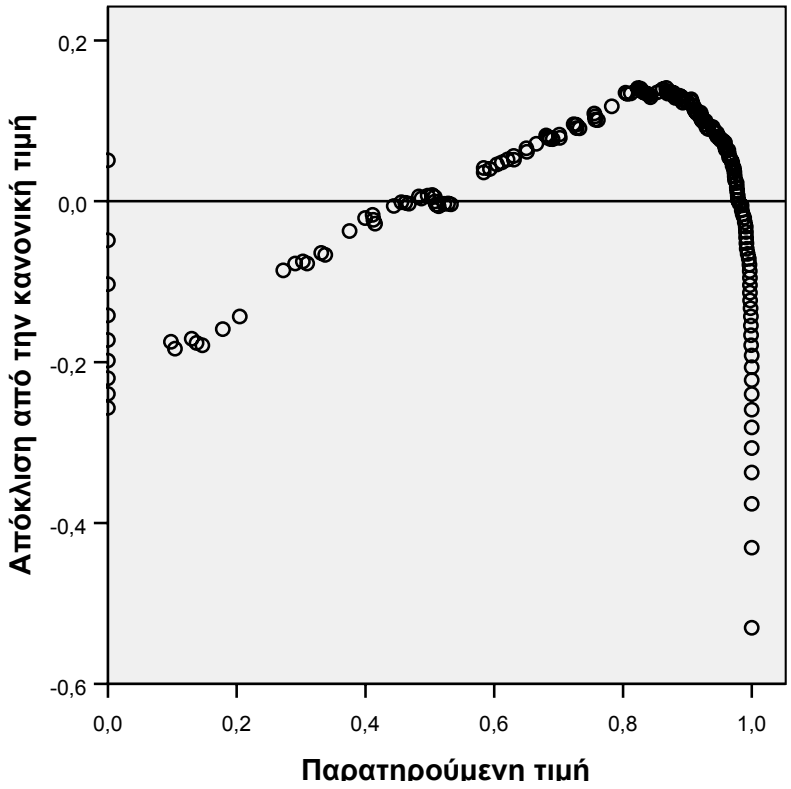
Όπου Z η μεταβλητή της σχέσης
Ο τελευταίος αυτός πίνακας είναι και ο πλέον σημαντικός διότι μας δίνει τους συντελεστές του τελικού μοντέλου μαζί και με τα διαστήματα εμπιστοσύνης αυτών. Σύμφωνα με το κριτήριο του Wald, σημαντική επίδραση επίδραση στην διανόρφωση των τιμών της εξαρτημένης μεταβλητής έχουν οι μεταβλητές Α6 ΄΄Οικονομική Κατάσταση΄΄, Γ2 ΄΄Κατηγορία Νοδώτητας κατά Kenedy΄΄ και η Β3 ΄΄Παλαιότητα Οδοντοστοιχίας". Σύμφωνα με τους εκτιμώμενους συντελεστές για μια μονάδα αύξησης της μεταβλητής οικονομικής κατάστασης έχουμε μείωση της σχετικής πιθανότητας ικανοποίησης περίπου κατά 85%.

11.4. Διαγνωστικά Διαγράμματα του Μοντέλου

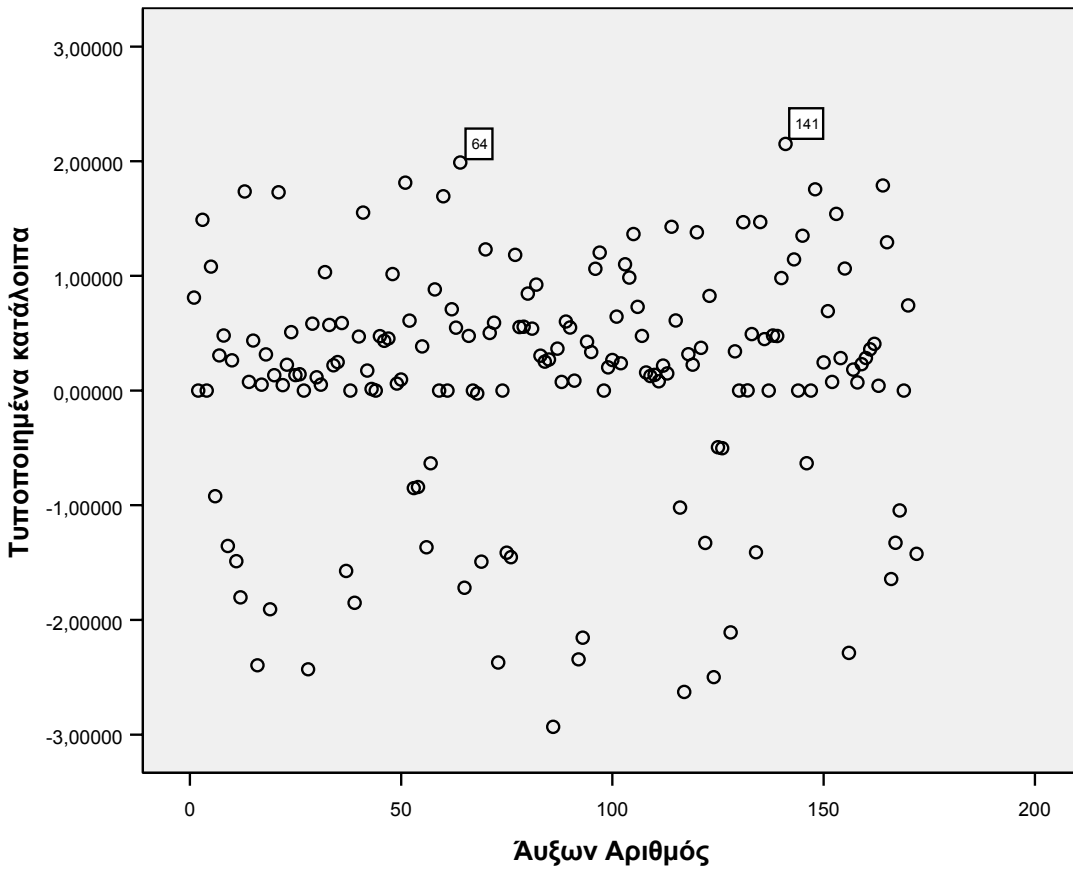


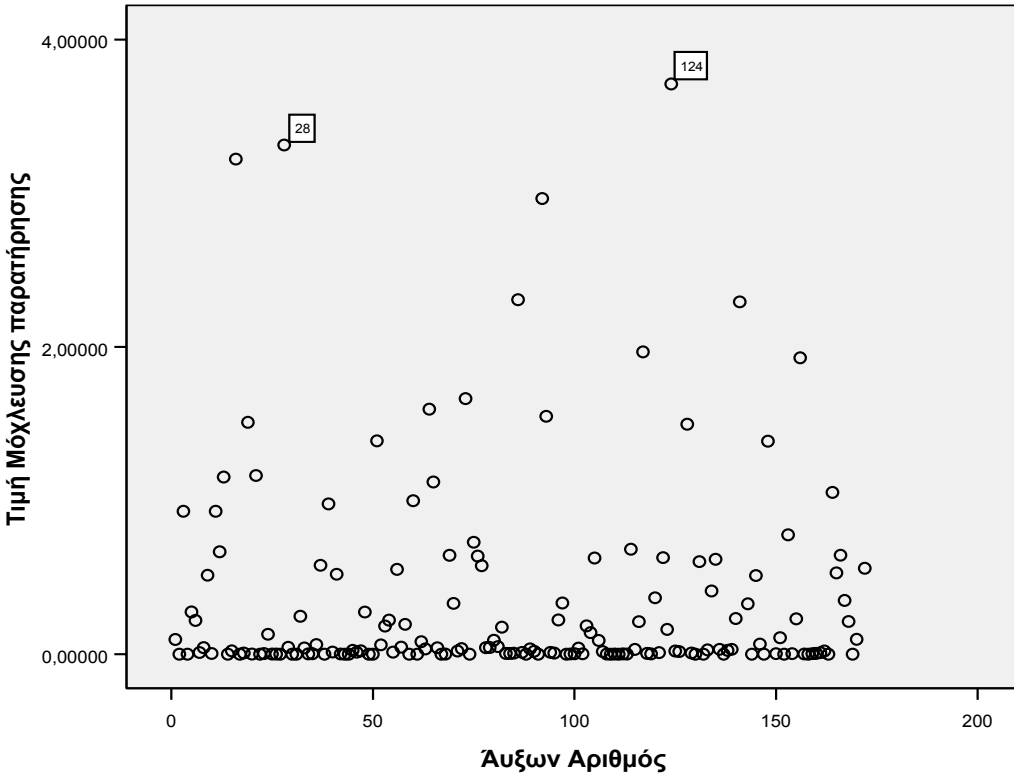
Το παραπάνω διάγραμμα δείχνει την σε ικανοποιητικό βαθμό κανονικότητα της κατανομής των τιμών της εξαρτημένης μεταβλητής ανεξαιρέσουμε κάποιες ακραίες τιμές.

Διάγραμμα καταλοίπων απόκλισης από την κανονικότητα



Ομοίως το παραπάνω διάγραμμα των υπολοίπων απόκλισης από την κανονικότητα δείχνει την συγκέντρωση των σημείων σε μια ζώνη εκατέρωθεν ευθείας γραμμής στο ύψος της μηδενικής κατανομής, γεγονός που υποστηρίζει την κανονικότητα της κατανομής των τιμών της εξαρτημένης μεταβλητής αν εξαιράσουμε τις ακραίες τιμές.


Τυποποιημένα κατάλοιπα
3,00000
2,00000
1,00000
0,00000
-1,00000
-2,00000
-3,00000
64
141
0
50
100
150
200
Άυξων Αριθμός


Τιμή Μόχλευσης παρατήρησης
4,00000
2,00000
0,00000
124
28
0
50
100
150
200
Άυξων Αριθμός

Το διάγραμμα των τυποποιημένων καταλοίπων σε τιμές z απεικονισμένων ως προς τη σειρά των παρατηρήσεων, δείχνει ότι οι παρατηρήσεις με αριθμό 64 και 141 έχουν σημαντικά υψηλότερες τιμές τυποποιημένων καταλοίπων , ενώ το διάγραμμα των τιμών μόλευσης των παρατηρήσεων εντοπίζει την 28 και 124 παρατήρηση ως τις υψηλότερες του παραπάνω στατιστικού.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Christensen, R. 1997. Log-Linear Models and Logistic Regression, 2nd edition, New York: Springer-Verlac.

Cox D.R. and Snell E.J.(1989) Analysis of Binary Data, 2nd editin. Chapman and Hall, London

Γναρδέλλης χαράλαμπος. Ανάλυση Δεδομένων με το SPSS 14.0 for Windows. Εκδόσεις Παπαζήση. Αθήνα 2006

Hauck, W.W. and A. Donner. 1997. Wald's test aw applied to hypotheses in logit analysis. Journal of the American Statistical Association,72:851-853

Hosmer,D.W. and S. Lemeshow. 1989. Applied logistic Regression. New York: John Wiley and Sons

Menard, scott W. Applied logistic Regression analysis. 2nd ed. Sage Publications, London.

SPSS. 1999. SPSS Regression Models. Chicago: SPSS Inc.

SPSS Inc (2005) SPSS Base 14.0 User's Guide inc, Chicago

Φ. Κολυβά- Μαχαίρα. Μαθηματική στατιστική. Τόμος 1. Εκδόσεις Ζήτη Θεσσαλονίκη 1998

Φ. Κολυβά- Μαχαίρα. Μαθηματική στατιστική. Τόμος 2. Θεσσαλονίκη 1985