Κεφάλαιο 4: Διαγνωστικά στο Μοντέλο Απλής Γραμμικής Παλινδρόμησης

Το μοντέλο γραμμικής παλινδρόμησης με κανονικά σφάλματα

Επανεξετάζουμε το μοντέλο γραμμικής παλινδρόμησης : $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ [4-1]

Y_i είναι η τιμή της παρατηρούμενης απόκρισης στην i δοκιμή (εισαγωγή δεδομένων).

β_0 είναι η σταθερά παράμετρος και β_1 είναι η παράμετρος κλίσης.

X_i είναι η τιμή της ανεξάρτητης μεταβλητής στην i δοκιμή (εισαγωγή δεδομένων).

ε_i είναι i.i.d., ανεξάρτητα $N\left(0,\ \sigma^2\right)$

Θα κάνουμε ανάλυση καταλοίπων για την εξέταση έξι σημαντικών τύπων αποκλίσεων από το παραπάνω μοντέλο:

1. Η συνάρτηση παλινδρόμησης δεν είναι γραμμική.
2. Η παρουσία ακραίων τιμών.
3. Οι όροι σφάλματος δεν έχουν σταθερή διακύμανση.
4. Οι όροι σφάλματος δεν είναι ανεξάρτητοι.
5. Οι όροι σφάλματος δεν ακολουθούν την κανονική κατανομή.
6. Μία ή περισσότερες σημαντικές μεταβλητές έχουν παραλειφθεί.

Τα Υπόλειπα

Το υπόλοιπο είναι η διαφορά της παρατηρούμενης τιμής (δεδομένων) Y_i μείον την προσαρμοσμένη (προβλεπόμενη) τιμή $\hat{Y}_i$.

$e_i = Y_i - \hat{Y}_i$ [4-2]

Στο μοντέλο [4–1] οι όροι σφάλματος ε_i θεωρούνται ανεξάρτητες τυχαίες μεταβλητές με μέσο όρο 0 και σταθερή διακύμανση σ^2. Εάν το μοντέλο ταιριάζει στα δεδομένα, τότε τα παρατηρούμενα (δεδομένα) κατάλοιπα e_i θα πρέπει να αντικατοπτρίζουν αυτές τις ιδιότητες για τα θεωρητικά κατάλοιπα ε_i. Αυτή είναι η βάση της ανάλυσης καταλοίπων.

Ο μέσος όρος των n καταλοίπων δίνεται από:

$$\bar{e} = \frac{\sum e_i}{n} \quad [4\text{-}3]$$

Η διακύμανση των n καταλοίπων δίνεται από:

$$V\left(e_i\right) = \frac{\sum\left(e_i - \bar{e}\right)^2}{n-2} \quad [4\text{-}4]$$

και από τότε $\overline{e} = 0 : V\left(e_i\right) = \frac{\sum e_i^2}{n-2}$ [4-5]

Χρησιμοποιώντας [3-9], $\hat{V}\left(e_i\right) = \frac{SSE}{n-2} = MSE$ [4-6]

Έτσι, εάν το μοντέλο που καθορίζεται από[4-1], τότε το MSE είναι ένας αμερόληπτος εκτιμητής του της διασποράς των σφαλμάτων σ^2 που είναι η διακύμανση των όρων σφάλματος ε_i.

θα χρησιμοποιήσουμε e^* για να συμβολίσουμε τα studenedized κατάλοιπα: Έτσι,

$$e_i^* = \frac{e_i - \overline{e}}{\sqrt{MSE}} \quad [4\text{-}7]$$

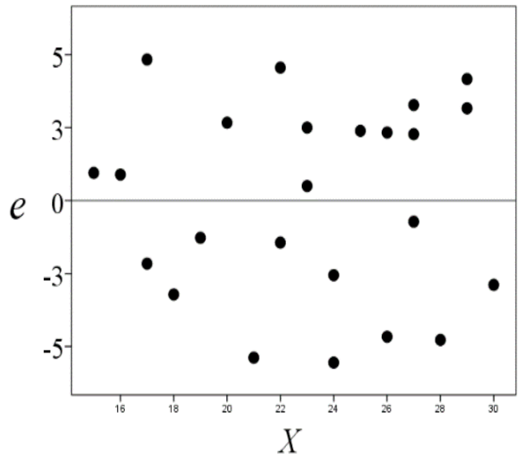
και από τότε $\overline{e} = 0 : e_i^* = \frac{e_i}{\sqrt{MSE}}$ [4-8]

Μη γραμμικότητα της συνάρτησης παλινδρόμησης

Η πρώτη απόκλιση του μοντέλου παλινδρόμησης στο[4-1]θα μπορούσε να προκληθεί από τη μη γραμμική σχέση μεταξύ της προβλεπόμενης μεταβλητής Y_i και της μεταβλητή πρόβλεψης X_i. Συνήθως μπορεί να ανιχνεύσουμε μη γραμμικότητα χρησιμοποιώντας το διάγραμμα διασποράς των δεδομένων της προβλεπόμενης μεταβλητής Y_i έναντι της μεταβλητής πρόβλεψης X_i.

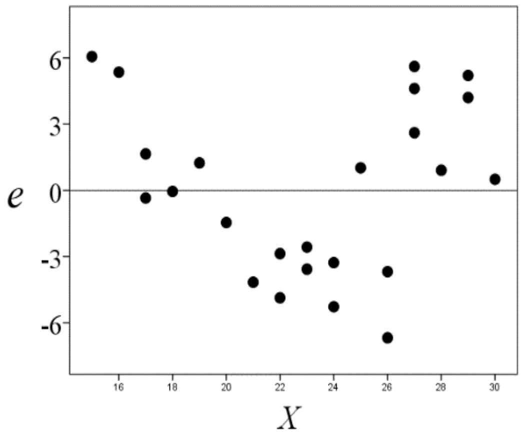
Ωστόσο, είναι καλύτερο να χρησιμοποιήσετε το διάγραμμα διασποράς των καταλοίπων e_i έναντι της μεταβλητής πρόβλεψης X_i. Εξετάσαμε τρία δείγματα ζευγαρωμένων δεδομένων $\left(X_i, Y_i\right)$. Για κάθε δείγμα υπολογίσαμε την εκτιμώμενη εξίσωση παλινδρόμησης $\hat{Y} = b_0 + b_1 X$. Υπολογίσαμε επίσης τα κατάλοιπα $e_i = Y_i - \hat{Y}_i$. Για κάθε δείγμα παρουσιάζουμε τα διαγράμματα διασποράς του e_i εναντίον του X_i.

Εικόνα 4-1



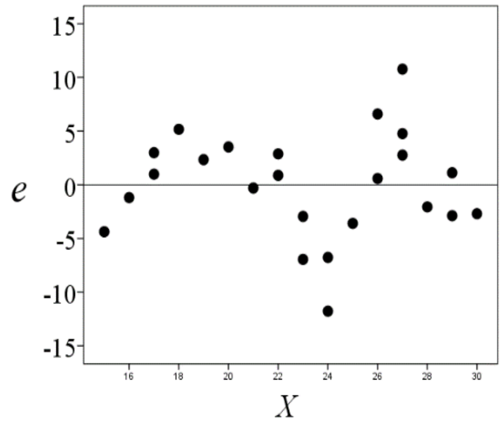
Τα σημεία στο στο **Σχήμα 4-1** των e_i έναντι των X_i είναι **τυχαία διασκορπισμένα (κατανεμημένα)** γύρω από την οριζόντια γραμμή $e = 0$. Αυτό αποκαλύπτει ότι ένα μοντέλο γραμμικής παλινδρόμησης είναι κατάλληλο για αυτό το δείγμα. Επομένως, **δεν υπάρχει απόκλιση από τη γραμμικότητα**.

Εικόνα 4-2



Τα σημεία **στο Σχήμα 4-2** των e_i έναντι των X_i δεν είναι τυχαία διασκορπισμένα (κατανεμημένα) γύρω από την οριζόντια γραμμή $e = 0$. Στην πραγματικότητα**, ακολουθούν ένα μοτίβο** που μοιάζει περισσότερο με μια κοίλη παραβολή. Αυτό αποκαλύπτει ότι ένα μοντέλο γραμμικής παλινδρόμησης δεν είναι κατάλληλο για αυτό το δείγμα. Έτσι, υπάρχει **απόκλιση από τη γραμμικότητα**.

Εικόνα 4-3

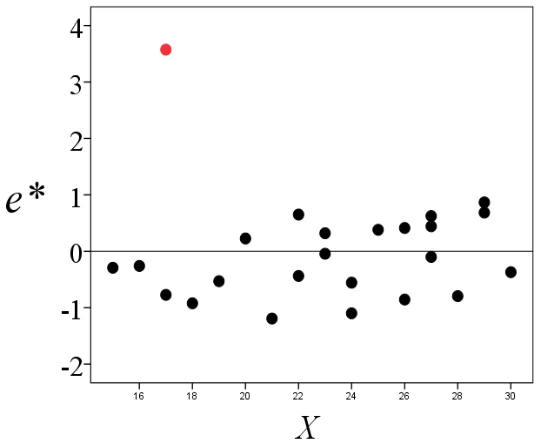


Τα σημεία στο στο **Σχήμα 4-3** των e_i έναντι των X_i δεν είναι τυχαία διασκορπισμένα (κατανεμημένα) γύρω από την οριζόντια γραμμή $e = 0$. Στην πραγματικότητα, **ακολουθούν ένα μοτίβο** που μοιάζει περισσότερο με κυβικό πολυώνυμο. Αυτό αποκαλύπτει ότι ένα μοντέλο γραμμικής παλινδρόμησης δεν είναι κατάλληλο για αυτό το δείγμα. Έτσι, υπάρχει **απόκλιση από τη γραμμικότητα**.

Παρουσία ακραίων παρατηρήσεων

Τα ακραία σημεία είναι ακραίες παρατηρήσεις. Το διάγραμμα διασποράς των studentized καταλοίπων που δίνονται στην [4-8] βοηθούν στον εντοπισμό ακραίων τιμών που απέχουν πολλές τυπικές αποκλίσεις από το μηδέν. Εντοπίζουμε τις ακραίες τιμές από το **διάγραμμα διασποράς του** e_i^* **εναντίον του** X_i**.** Μια παρατήρηση χαρακτηρίζεται ως ακραία τιμή εάν βρίσκεται πάνω από 2,5 τυπικές αποκλίσεις από το μέσο studentized κατάλοιπο που είναι μηδέν. Από το Σχήμα 4-4 παρατηρούμε ότι υπάρχει μια ακραία τιμή που απέχει περίπου 3,6 τυπικές αποκλίσεις από το μηδέν.

Εικόνα 4-4

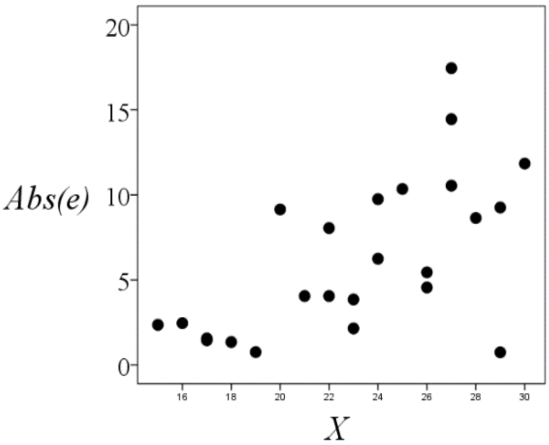


Σε μοντέλο[4-1], αναφέρεται ότι: ε_i είναι ανεξάρτητοι $N\left(0,\ \sigma^2\right)$. Έτσι θα εξετάσουμε πώς να προσδιορίσουμε τις αποκλίσεις από την ανεξαρτησία, την κανονικότητα και τη μη σταθερή διακύμανση.

Μη σταθερή διακύμανση

Όταν η διακύμανση των όρων σφάλματος δεν είναι σταθερή στο σ^2 το **διάγραμμα διασποράς της απόλυτης τιμής των** e_i **έναντι των** X_i δείχνει ένα μοτίβο όπου η απόλυτη τιμή των e_i αυξάνεται όσο η τιμή του X αυξάνεται (βλ. Εικόνα 4-5 παρακάτω). Το μοτίβο μοιάζει με τρίγωνο που υποδηλώνει μη σταθερότητα της διακύμανσης των όρων σφάλματος.

Εικόνα 4-5



Εκτός από την οπτική εκτίμηση της μη σταθερότητας του διαγράμματος διασποράς των καταλοίπων e_i έναντι των σημείων X, μπορούν να χρησιμοποιηθούν δύο έλεγχοι υποθέσεων: το modify Levene test και/ή ο έλεγχος Breusch-Pagan.

Το modify Levene test

Το modify Levene test είναι ένας στατιστικός έλεγχος για μη σταθερή διακύμανση που δεν προϋποθέτει την κανονικότητα των όρων σφάλματος. Οι υποθέσεις είναι οι εξής:

H_0 : the variance of the error terms is constant; H_1 : the variance of the error terms is *not* constant

Αρχικά χωρίζουμε τις τιμές της μεταβλητής πρόβλεψης X σε δύο ομάδες: τις σχετικά χαμηλές τιμές των X που τοποθετούνται στην πρώτη ομάδα και τις σχετικά υψηλές τιμές των X που κατατάσσονται στη δεύτερη ομάδα. Συμβολίζουμε το i-οστό κατάλοιπο της πρώτης ομάδας με e_{i1} και το i-οστό υπόλοιπο της δεύτερης ομάδας με e_{i2}. Επίσης χρησιμοποιούμε n_1 και n_2 για να συμβολίσουμε τα μεγέθη του δείγματος κάθε ομάδας, όπου $n = n_1 + n_2$. θα χρησιμοποιήσουμε $\tilde{e}_1$ και $\tilde{e}_2$ για να συμβολίσουμε τις διάμεσους των καταλοίπων σε κάθε ομάδα. Επίσης, έστω $d_{i1} = |e_{i1} - \tilde{e}_1|$ και $d_{i2} = |e_{i2} - \tilde{e}_2|$ οι απόλυτες αποκλίσεις των καταλοίπων γύρω από τη διάμεσο της ομάδας τους.

Η στατιστική συνάρτηση είναι: $t_L^* = \dfrac{\bar{d}_1 - \bar{d}_2}{\sqrt{s^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ [4-10]

όπου, s^2 είναι η εκτιμώμενη διακύμανση: $s^2 = \dfrac{\sum_{i=1}^{n_1}\left(d_{i1} - \bar{d}_1\right) + \sum_{i=1}^{n_2}\left(d_{i2} - \bar{d}_2\right)}{n-2}$ [4-11]

Εάν οι όροι σφάλματος έχουν σταθερή διακύμανση, τότε t_L^* ακολουθεί την κατανομή t με n – 2 βαθμούς ελευθερίας. Ως εκ τούτου, απορρίπτουμε την H_0 αν $|t^*| > t\left(1-\frac{a}{2}, n-2\right)$ οπότε αποφασίζουμε ότι η διακύμανση των όρων σφάλματος δεν είναι σταθερή.

Το τεστ Breusch-Pagan

Αυτή ο έλεγχος υποθέτει ότι η διακύμανση κάθε όρου σφάλματος e_i, που συμβολίζεται με σ_i^2 σχετίζεται με τη μεταβλητή X με τον ακόλουθο τρόπο: $\log_e \sigma_i^2 = \gamma_0 + \gamma_1 X_i$. [4-12]

Η σχέση [4-12] μοιάζει με ένα μοντέλο γραμμικής παλινδρόμησης, επομένως η σταθερότητα της διακύμανσης του σφάλματος ισοδυναμεί με το να είναι η τιμή της κλίσης ίση με μηδέν, δηλαδή $\gamma_1 = 0$.

Επομένως, ελέγχουμε : $H_0 : \gamma_1 = 0$ κατά $H_1 : \gamma_1 \neq 0$

Ο έλεγχος πραγματοποιείται με γραμμική παλινδρόμηση των τετραγώνων των καταλοίπων e_i^2 ως η προβλεπόμενη μεταβλητή και X_i ως προγνωστική μεταβλητή. Λαμβάνουμε το άθροισμα παλινδρόμησης των τετραγώνων αυτού του μοντέλου που συμβολίζεται με SSR^*. Λαμβάνουμε επίσης το άθροισμα σφάλματος των τετραγώνων σφάλματος της παλινδρόμησης του Y στο X, που συμβολίζεται με SSE. Το στατιστικό τεστ είναι ένα στατιστικό τεστ Chi-Square ως εξής:

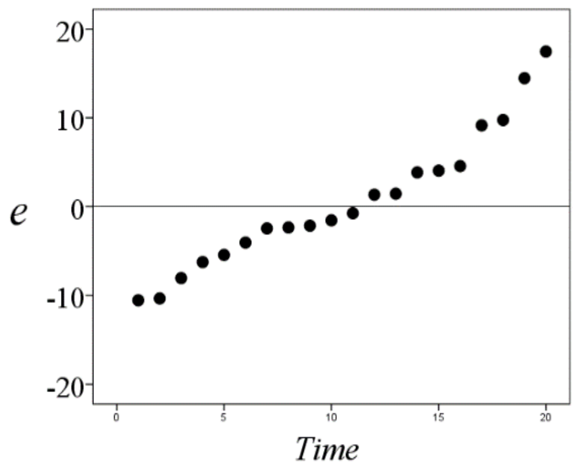
$$\chi^2_{BP} = \frac{SSR^*}{2} \div \left(\frac{SSE}{n}\right)^2 \qquad [4\text{-}13]$$

Αν H_0 ισχύει τότε χ^2_{BP} ακολουθεί την κατανομή Chi-Square με έναν βαθμό ελευθερίας. Τέλος, απορρίπτουμε H_0 αν η τιμή της στατιστικής συνάρτησης είναι μεγαλύτερη από την κρίσιμη τιμή: δηλαδή απορρίπτουμε H_0 αν $\chi^2_{BP} > \chi^2(1-\alpha,\ 1)$

Μη ανεξαρτησία των Όρων Σφάλματος

Κάθε φορά που λαμβάνονται δεδομένα σε μια χρονική ακολουθία ή κάποιον άλλο τύπο ακολουθίας, μια γραφική παράσταση ακολουθίας των καταλοίπων αποκαλύπτει εάν υπάρχει οποιαδήποτε συσχέτιση μεταξύ όρων σφάλματος που βρίσκονται κοντά ο ένας στον άλλο. Το Σχήμα 4-6 δείχνει ένα παράδειγμα μη ανεξαρτησίας που παρουσιάζεται σε μια γραφική παράσταση διασποράς των καταλοίπων e έναντι του χρόνου. Απεικονίζει ένα γραμμικό εφέ τάσης που σχετίζεται με το χρόνο.

Εικόνα 4-6



Μη κανονικότητα των Όρων Σφάλματος

Ο καλύτερος τρόπος για να ελέγξετε την κανονικότητα των όρων σφάλματος είναι με την κανονική γραφική παράσταση πιθανότητας στην οποία τα υπολείμματα σχεδιάζονται με την αναμενόμενη τιμή τους. Μια γραφική παράσταση κανονικής πιθανότητας που είναι σχεδόν γραμμική υποδηλώνει συμφωνία με την κανονικότητα, ενώ μια γραφική παράσταση που αποκλίνει ουσιαστικά από τη γραμμικότητα υποδηλώνει ότι η κατανομή των όρων σφάλματος δεν είναι κανονική.

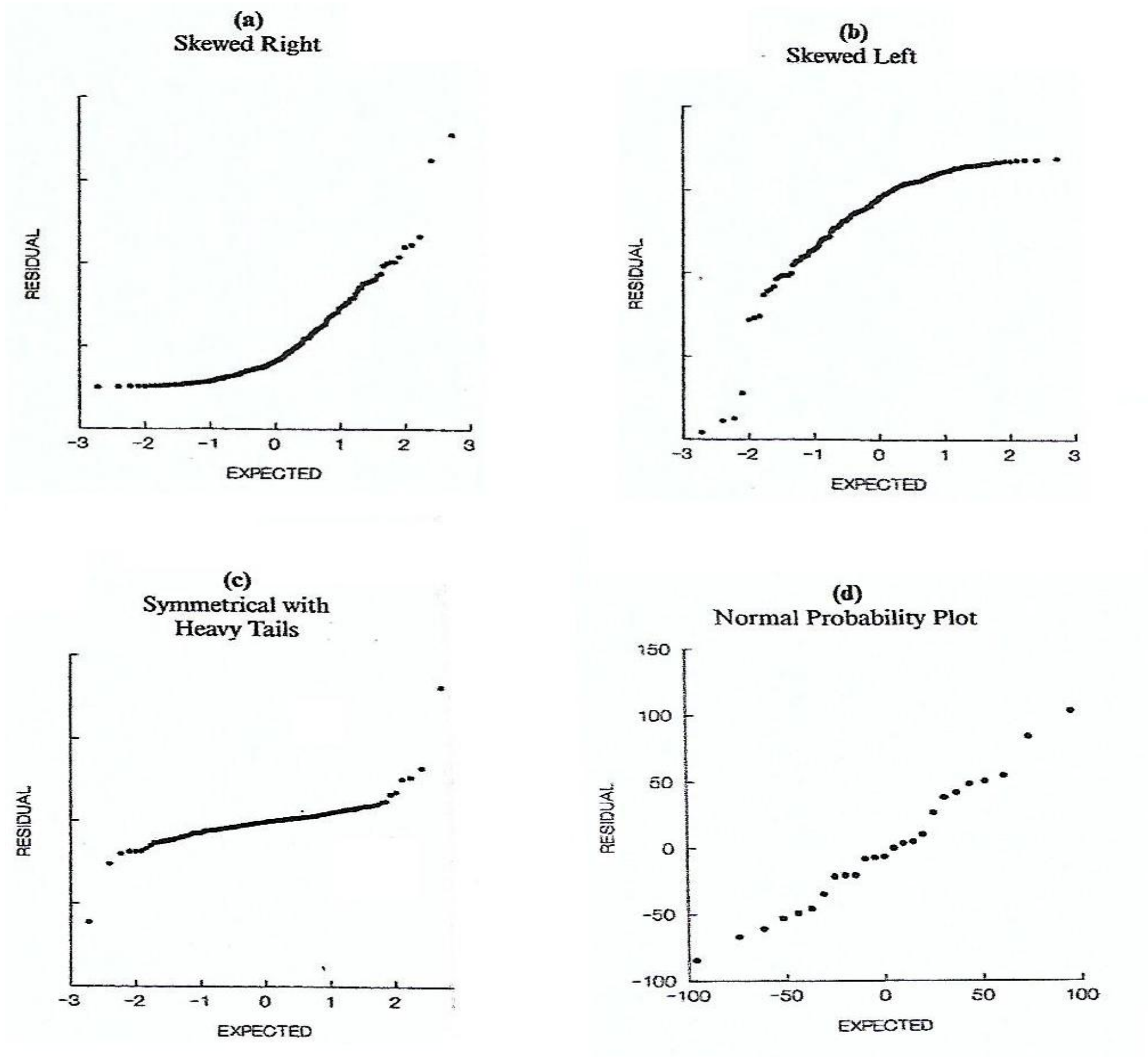
Η αναμενόμενη τιμή των όρων σφάλματος είναι μηδέν και η απόκλιση των όρων σφάλματος εκτιμάται από το MSE. Η θεωρία έχει δείξει ότι μια καλή προσέγγιση της αναμενόμενης τιμής της kth μικρότερης (διατεταγμένης) παρατήρησης σε ένα τυχαίο δείγμα μεγέθους n είναι:

$$\sqrt{MSE}\left[z\left(\frac{k-0.375}{n+0.25}\right)\right] \qquad [4\text{-}9]$$

όπου, $z(A)$ δηλώνει το (A)-100 εκατοστημόριο της τυπικής κανονικής κατανομής.

Μερικά παραδείγματα γραφικών κανονικών πιθανοτήτων δίνονται παρακάτω:

Εικόνα 4-7



Το Σχήμα 4-7(δ) παρουσιάζει μια κανονική γραφική παράσταση πιθανότητας η οποία είναι κοντά σε μια ευθεία γραμμή και η οποία επαληθεύει την κανονικότητα των όρων σφάλματος. Τα υπόλοιπα στοιχεία δείχνουν αποκλίσεις από την κανονικότητα των όρων σφάλματος. Το Σχήμα 14-7(α) αποκαλύπτει μια κατανομή των όρων σφάλματος η οποία είναι λοξή προς τα δεξιά (positively skewed), ενώ η Εικόνα 14-7(β) αποκαλύπτει μια κατανομή των όρων σφάλματος η οποία είναι λοξή προς τα αριστερά (negatively skewed). Το Σχήμα 14-7(γ) αποκαλύπτει μια κατανομή των όρων σφάλματος που είναι συμμετρική με βαριές ουρές. Θα πρέπει να δηλωθεί ότι οι άλλες αποκλίσεις μπορεί να διαταράξουν το κανονικό διάγραμμα πιθανοτήτων επομένως, είναι μια καλή στρατηγική να εξετάσουμε στο τέλος την απόκλιση από την κανονικότητα.

Τεστ συσχέτισης για κανονικότητα

Εκτός από την οπτική αξιολόγηση της κατά προσέγγιση γραμμικότητας των σημείων που απεικονίζονται σε μια κανονική γραφική παράσταση πιθανοτήτων, μπορεί να πραγματοποιηθεί ένας έλεγχος για την κανονικότητα των όρων σφάλματος με τον υπολογισμό του συντελεστή συσχέτισης μεταξύ των καταλοίπων e_i έναντι της αναμενόμενης τιμής υπό την προϋπόθεση της κανονικότητας. Αυτό ονομάζεται Correlation Test for Normality. Η κρίσιμη τιμή για μια τέτοια δοκιμή δίνεται από τον πίνακα που ετοίμασαν οι Looney και Gulledge για διάφορες τιμές του μεγέθους δείγματος n. Η κανονικότητα απορρίπτεται εάν η κρίσιμη τιμή είναι μεγαλύτερη από τον συντελεστή συσχέτισης που υπολογίζεται.

Περίληψη

Συνοψίζουμε τα 4 από τα 6 σημεία απόκλισης από το γραμμικός μοντέλο που αναφέρεται στη σελίδα 1 ως εξής:

Απόκλιση από τη Γραμμικότητα

Γράφημα των καταλοίπων e έναντι του X. Δεν υπάρχει απόκλιση από τη γραμμικότητα εάν τα σημεία είναι ομοιόμορφα κατανεμημένα πάνω και κάτω από την οριζόντια γραμμή $e = 0$. Διαφορετικά, εάν υπάρχει ένα μοτίβο των σημείων γύρω από αυτή τη γραμμή, τότε υπάρχει απόκλιση από τη γραμμικότητα.

Παρουσία Outliers

Γράφημα των καταλοίπων e^* έναντι του X. Δεν υπάρχουν ακραίες τιμές εάν δεν υπάρχουν σημεία που απέχουν περισσότερο από 2,5 τυπικές αποκλίσεις από την οριζόντια γραμμή $e^* = 0$. Σημειώστε ότι ο αριθμός 2,5 είναι αφηρημένος. Γενικά αν υπάρχει σημείο πάνω από 2 τυπικές αποκλίσεις από τη γραμμή $e^* = 0$ τότε χρήζει περεταίρω έρευνας.

Μη σταθερότητα διασποράς των καταλοίπων

Σχεδιάζουμε την απόλυτη τιμή των καταλοίπων έναντι X, δηλαδή $|e|$ έναντι X. Για ευκολία θα γράψουμε $Abs(e)$ αντί για $|e|$. Υπάρχει απόκλιση από τη σταθερή διακύμανση εάν υπάρχει ένα μοτίβο στην παραπάνω γραφική παράσταση που ανοίγει από αριστερά προς τα δεξιά και μοιάζει με τρίγωνο. Επιπλέον, υπάρχουν δύο στατιστικά τεστ υποθέσεων που μπορούμε να εκτελέσουμε: το τροποποιημένο τεστ Levene και το τεστ Breusch-Pagan.

Απόκλιση από την κανονικότητα των όρων σφάλματος

Σχεδιάζουμε τους όρους σφάλματος σε σχέση με την αναμενόμενη τιμή τους υπό την προϋπόθεση της κανονικότητας. Δεν υπάρχει απόκλιση από την κανονικότητα εάν τα σημεία βρίσκονται κοντά στην ευθεία. Διαφορετικά έχουμε μια απόκλιση από την κανονικότητα, οπότε το μοτίβο της γραφικής παράστασης μπορεί να μας οδηγήσει σε τρεις πιθανότητες αναχώρησης σχετικά με την κατανομή των όρων σφάλματος: μια θετικά λοξή κατανομή, μια αρνητικά λοξή κατανομή ή μια συμμετρική κατανομή με βαριές ουρές (βλ. Εικόνα 4-7) στη σελίδα 4. Επιπλέον, μπορούμε να εκτελέσουμε μια κανονικότητα.

To modified Levene test

(a) H_0 : the variance of the error terms is constant; H_1 : the variance of the error terms is *not* constant

(b) Στατιστική συνάρτηση: $t_L^* = \dfrac{\bar{d}_1 - \bar{d}_2}{\sqrt{s^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

(c) Κρίσιμη τιμή: $|t^*| > t\left(1 - \dfrac{a}{2}, n-2\right)$

(d) Ολοκληρώστε το τεστ: Απόρριψη H_0 αν $|t^*| > t\left(1 - \dfrac{a}{2}, n-2\right)$

(e) Αν απορρίψουμε H_0 συμπεραίνουμε την απόκλιση από τη σταθερή διακύμανση των όρων σφάλματος.

Το τεστ Breusch-Pagan

(a) H_0 : the variance of the error terms is constant; H_1 : the variance of the error terms is *not* constant

(b) Στατιστική συνάρτηση: $\chi^2_{BP} = \dfrac{SSR^*}{2} \div \left(\dfrac{SSE}{n}\right)^2$

(c) Κρίσιμη τιμή: $\chi^2(1-\alpha,\ 1)$

(d) Ολοκληρώστε το τεστ: Απόρριψη H_0 αν $\chi^2_{BP} > \chi^2(1-\alpha,\ 1)$

(e) Αν απορρίψουμε H_0 συμπεραίνουμε την απόκλιση από τη σταθερή διακύμανση των όρων σφάλματος.

Το τεστ συσχέτισης για κανονικότητα

(a) H_0 : the error terms ~N H_1 : the error terms do not ~N

(b) Στατιστική συνάρτηση: $r(e,\ EXPe)$

(c) Κρίσιμη τιμή (C.V.): από τον πίνακα Looney-Gulledge

(d) Συμπέρασμα: Απόρριψη H_0 αν $r(e,\ EXPe) <$ C.V.

(e) Αν απορρίψουμε H_0 συμπεραίνουμε την απόκλιση από την κανονικότητα των όρων σφάλματος.