ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ & ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ

Π Τ Υ Χ Ι Α Κ Η Ε Ρ Γ Α Σ Ι Α

Πολλαπλή Παλινδρόμηση με Έμφαση στο Πρόβλημα της Ετεροσκεδαστικότητας

ΚΟΝΤΟΥΔΑΚΗΣ ΖΩΗΣ

ΜΑΡΤΖΑΚΛΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ

ΧΟΡΤΑΤΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ ΧΡΗΣΤΟΣ

ΕΙΣΗΓΗΤΗΣ

ΜΕΓΑΡΙΤΗΣ ΑΘΑΝΑΣΙΟΣ

Μ Ε Σ Ο Λ Ο Γ Γ Ι 2 0 1 5

ΤΕΧΝΟΛΟΓΙΚΟ ΕΚΠΑΙΔΕΥΤΙΚΟ ΙΔΡΥΜΑ ΔΥΤΙΚΗΣ ΕΛΛΑΔΟΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ & ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ ΚΑΙ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗΣ

Π Τ Υ Χ Ι Α Κ Η Ε Ρ Γ Α Σ Ι Α

Πολλαπλή Παλινδρόμηση με Έμφαση στο πρόβλημα της Ετεροσκεδαστικότητας

ΚΟΝΤΟΥΔΑΚΗΣ ΖΩΗΣ (Α.Μ. 14570)

ΜΑΡΤΖΑΚΛΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ (Α.Μ. 14640)

ΧΟΡΤΑΤΟΣ ΚΩΝΣΤΑΝΤΙΝΟΣ ΧΡΗΣΤΟΣ (Α.Μ. 14853)

ΕΙΣΗΓΗΤΗΣ

ΜΕΓΑΡΙΤΗΣ ΑΘΑΝΑΣΙΟΣ

Μ Ε Σ Ο Λ Ο Γ Γ Ι 2 0 1 5

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Περίληψη

Η *Απλή Γραμμική Παλινδρόμηση* αφορά έναν πληθυσμό όπου εξετάζουμε δύο χαρακτηριστικά- μεταβλητές Y & X και πιο συγκεκριμένα, μας ενδιαφέρει να δούμε πώς διαμορφώνονται οι τιμές της εξαρτημένης μεταβλητής Y όταν μεταβάλλονται οι τιμές της ανεξάρτητης μεταβλητής X.

Στην περίπτωση που διερευνούμε τρεις ή περισσότερες μεταβλητές με απώτερο σκοπό να διαπιστώσουμε κατά κύριο λόγο το πώς μεταβάλλονται οι τιμές της εξαρτημένης μεταβλητής Y όταν γνωρίζουμε τις τιμές των ανεξαρτήτων μεταβλητών X_1, X_2, …, X_n και κατά δεύτερον να εξετάσουμε τον βαθμό συσχέτισης ανάμεσα σε αυτές τις μεταβλητές.

Η σχέση η οποία συνδέει την εξαρτημένη μεταβλητής Y με τις περισσότερες από δύο ανεξάρτητες μεταβλητές καλείται *Πολλαπλή Παλινδρόμηση*. Η *Απλή Παλινδρόμηση* αναπαριστάνεται στο χώρο με μια ευθεία ή με μια άλλη καμπύλη ενώ αντίθετα η *Πολλαπλή Παλινδρόμηση* απεικονίζεται με μια επιφάνεια (επίπεδο) ή μια υπερεπιφάνεια.

Η Πολλαπλή Γραμμική Παλινδρόμηση εφαρμόζεται στην περίπτωση που διαθέτουμε ένα πλήθος ανεξαρτήτων μεταβλητών X_i, οι οποίες εμφανίζουν έντονες κατά προτίμηση γραμμικές συσχετίσεις με την εξαρτημένη μεταβλητή Y. Όσο πιο ισχυρές είναι αυτές οι συσχετίσεις τόσο πιο καλή εφαρμογή βρίσκει η Πολλαπλή Γραμμική Παλινδρόμηση μέσω της μεθόδου των Ελαχίστων Τετραγώνων και προσαρμόζεται έτσι η εκτιμώμενη ευθεία παλινδρόμησης στα δεδομένα. Η μέθοδος των Ελαχίστων Τετραγώνων μας βοηθά να εκτιμήσουμε τους συντελεστές παλινδρόμησης $\hat{a}, \hat{\beta}$, να υπολογίσουμε τα σφάλματα ή αποκλίσεις ε_i, να κάνουμε ελέγχους υποθέσεων αναφορικά με τις παραμέτρους της παλινδρόμησης και παράλληλα να προβλέψουμε τιμές της εξαρτημένης μεταβλητής Y.

Αν όμως υφίστανται έντονες συσχετίσεις και ανάμεσα στις ανεξάρτητες μεταβλητές X_i , κατάσταση η οποία δεν είναι επιθυμητή ή οι προϋποθέσεις του γραμμικού μοντέλου δεν πληρούνται, τότε συνήθως καταφεύγουμε στην χρήση κατάλληλων μετασχηματισμών για να διορθωθεί το πρόβλημα και να προχωρήσουμε στην υλοποίηση της μεθόδου και στην στατιστική συμπερασματολογία.

Οι προϋποθέσεις που αφορούν το γραμμικό μοντέλο είναι :1) Η κανονικότητα των σφαλμάτων ε_i, 2) η ανεξαρτησία των σφαλμάτων ε_i, 3) η γραμμικότητα μεταξύ των ανεξαρτήτων μεταβλητών X_i και της εξαρτημένης μεταβλητής Y και 4) η Ομοσκεδαστικότητα των σφαλμάτων ε_i.

Όταν τα σφάλματα ε_i δεν έχουν ίσες διακυμάνσεις προκύπτει το πρόβλημα της Ετεροσκεδαστικότητας, το οποίο μπορούμε να διαπιστώσουμε αν ισχύει διαγραμματικά εξετάζοντας τα σφάλματα ή «υπόλοιπα» του μοντέλου. Η Ομοσκεδαστικότητα είναι μία από τις βασικότερες υποθέσεις του γραμμικού μοντέλου και η μη ικανοποίηση της έχει αντίκτυπο στην σωστή εφαρμογή της Πολλαπλής Παλινδρόμησης. Όπως αναφέραμε και προηγουμένως, η λύση στο πρόβλημα της Ετεροσκεδαστικότητας και γενικά σε κάθε απόκλιση από τις υποθέσεις της Παλινδρόμησης είναι ο μετασχηματισμός των μεταβλητών.

ΠΡΟΛΟΓΟΣ

Το θέμα που πραγματεύεται η παρούσα πτυχιακή εργασία εστιάζεται στην *Πολλαπλή Παλινδρόμηση (Multiple Regression).*

Η παλινδρόμηση είναι μια ευρέως χρησιμοποιούμενη στατιστική τεχνική μοντελοποίησης για την έρευνα της συσχέτισης μεταξύ μίας εξαρτώμενης μεταβλητής και μιας ή περισσότερων ανεξάρτητων μεταβλητών. Χρησιμοποιείται με σκοπό την εκχώρηση δεδομένων σε μία πραγματική μεταβλητή πρόβλεψης, όπως ισχύει και στην περίπτωση της κατηγοριοποίησης όταν είναι διακριτή, αλλιώς καλείται παλινδρόμηση αν η μεταβλητή είναι συνεχής.

Η παλινδρόμηση προϋποθέτει ότι τα σχετικά δεδομένα ταιριάζουν με μερικά γνωστά είδη συνάρτησης και μετά καθορίζει την καλύτερη συνάρτηση αυτού του είδους που μοντελοποιεί τα δεδομένα που έχουν δοθεί. Αποτέλεσμα της παλινδρόμησης όταν χρησιμοποιείται ως τεχνική εξόρυξης δεδομένων, αποτελεί ένα μοντέλο που χρησιμοποιείται αργότερα για να προβλέψει τις τιμές της κατηγορίας για τα νέα δεδομένα. Τέτοια παραδείγματα εφαρμογής της παλινδρόμησης αποτελεί η πρόβλεψη της ζήτησης για ένα νέο προϊόν ή υπηρεσία συναρτήσει των δαπανών διαφήμισης ή ο υπολογισμός της ταχύτητας του ανέμου σε σχέση με την θερμοκρασία, την υγρασία και την ατμοσφαιρική πίεση του περιβάλλοντος.

Συμπερασματικά, θα πραγματοποιηθεί μια *θεωρητική επισκόπηση* της μεθόδου της *Πολλαπλής Παλινδρόμησης* δίνοντας έμφαση στο πρόβλημα της *Ετεροσκεδαστικότητας, της μη ισότητας δηλαδή των διακυμάνσεων* των τυχαίων μεταβλητών .

ΕΙΣΑΓΩΓΗ

Ως Στατιστική ορίζεται η επιστήμη η οποία ασχολείται με την συλλογή, την ανάλυση και την ερμηνεία δεδομένων. Εναλλακτικά μπορούμε να πούμε ότι Στατιστική είναι η προσπάθεια εξαγωγής συμπερασμάτων κάτω από συνθήκες αβεβαιότητας. Συμπερασματικά, η Στατιστική περιλαμβάνει τόσο τις μεθόδους συλλογής και επεξεργασίας στοιχείων, όσο και τις μεθόδους ανάλυσης και μελέτης τους, ανακαλύπτοντας με αυτόν τον τρόπο τις σχέσεις που υφίστανται ανάμεσα στα διάφορα φαινόμενα και διατυπώνοντας συμπεράσματα που είναι χρήσιμα για την λήψη ορθών αποφάσεων.

Τα βασικά στάδια που ακολουθούνται για την εξέταση των ιδιοτήτων των διαφόρων στοιχείων μιας πολυπληθούς ομάδας είναι τα εξής:

a) Η συγκέντρωση των στατιστικών στοιχείων που είναι αναγκαία για την μελέτη του προβλήματος που θέλουμε να διερευνήσουμε.
b) Η συστηματική επεξεργασία και παρουσίαση των στατιστικών στοιχείων υπό την μορφή στατιστικών πινάκων και διαγραμμάτων.

c) Η ανάλυση των στοιχείων αυτών και η εξαγωγή συμπερασματολογίας για την λήψη ορθών αποφάσεων.

Η λέξη «Στατιστική» προέρχεται από την λατινική λέξη «status» και η οποία σημαλινει κράτος. Αρχικά η Στατιστική δήλωνε την συλλογή στοιχείων για κρατικούς σκοπούς και ανάγκες. Κατά το πέρασμα του χρόνου, η Στατιστική θα ξεφύγει από τον περιγραφικό της χαρακτήρα με την ανάπτυξη ενός νέου κλάδου, του λογισμού των Πιθανοτήτων.

Η Στατιστική βρίσκει εφαρμογές σε όλες σχεδόν τις εκφάνσεις της ανθρώπινης δραστηριότητας. Ενδεικτικά ορισμένα πεδία εφαρμογής της είναι η Δημογραφία, η Ιατρική, η Αστρονομία, η Γεωργία, η Βιομηχανία, η Διοίκηση Επιχειρήσεων, η Οικονομία κ.α.

Πιο ειδικά, στην περίπτωση των διπαραμετρικών κατανομών, όπου μελετάμε συγχρόνως δύο μεταβλητές, ενδιαφερόμαστε να εξακριβώσουμε την ύπαρξη ή όχι συσχέτισης μεταξύ τους, δηλαδή να διαπιστώσουμε αν οι τιμές της μιας μεταβλητής επηρεάζονται από τις τιμές της άλλης και να διακριβώσουμε τον τρόπο συσχέτισης τους. Αυτά στην Απλή Γραμμική Παλινδρόμηση όπου έχουμε μια εξαρτημένη μεταβλητή Υ και μια ανεξάρτητη μεταβλητή Χ. Όταν το πλήθος των ανεξαρτήτων μεταβλητών είναι μεγαλύτερο ή ίσο του δύο, τότε μιλάμε για την Πολλαπλή Παλινδρόμηση, που αποτελεί το αντικείμενο της συγκεκριμένης πτυχιακής εργασίας.

- *Κεφάλαιο 1: Ανάλυση Παλινδρόμησης*

▪ *1.1: Το Γενικό Γραμμικό Μοντέλο*

Βασικός στόχος της μελέτης των Γραμμικών Μοντέλων είναι η πρόβλεψη μιας μεταβλητής Υ με βάση τα στοιχεία που διαθέτουμε για ένα σύνολο άλλων μεταβλητών $X_{0,}X_1$……, $X_{\rho-1}$. Η ***Υ αναφέρεται σαν εξαρτημένη μεταβλητή***, ενώ οι $X_{0,}X_1$……, $X_{\rho-1}$ σαν ***ανεξάρτητες ή προβλέπουσες***. Οι εφαρμογές των Γραμμικών Μοντέλων είναι πολλές και ανάγονται σχεδόν σε όλες τις επιστήμες.

Στο κλασσικό Γραμμικό Μοντέλο υποθέτουμε ότι η Υ αποτελείται από ένα *γραμμικό κομμάτι* που περιέχει τα $X_{0,}X_1$……, $X_{\rho-1}$ και ένα *τυχαίο σφάλμα ε* που αντανακλά τα διάφορα σφάλματα των μετρήσεων καθώς επίσης και επιδράσεις από άλλες προβλέπουσες μεταβλητές που δεν περιλήφθηκαν στο μοντέλο. Οι τιμές που παίρνουν τα x_i θεωρούνται *καθορισμένες (μη τυχαίες)* ενώ αντίθετα τα σφάλματα ε και επομένως και τα παρατηρούμενα Υ θεωρούνται σαν *τυχαίες μεταβλητές* που ικανοποιούν κάποιες στατιστικές υποθέσεις.

Η γενική μορφή του γραμμικού μοντέλου είναι η

$$Y=\beta_0+\beta_1X_1+\ldots+\beta_{\rho-1}X_{\rho-1}+\varepsilon$$

όπου ο όρος ‘’γραμμικό’’ αναφέρεται στις άγνωστες παραμέτρους $\beta_0,\ldots,\beta_{\rho-1}$.

Μια μέθοδος εύρεσης εκτιμητριών για τις παραμέτρους β_i είναι η λεγόμενη *Μέθοδος Ελαχίστων Τετραγώνων*, σύμφωνα με την οποία επιδιώκουμε να ελαχιστοποιήσουμε το άθροισμα των τετραγώνων των σφαλμάτων.

▪ *1.2: Απλή Γραμμική Παλινδρόμηση*

Η απλούστερη περίπτωσης παλινδρόμησης είναι η *Απλή Γραμμική Παλινδρόμηση* όπου υπάρχει μόνο μια ανεξάρτητη μεταβλητή Χ και η εξαρτημένη μεταβλητή Υ μπορεί να προσεγγιστεί ικανοποιητικά από μια γραμμική συνάρτηση του Χ.
Ας υποθέσουμε γενικά ότι έχουμε n ζεύγη παρατηρήσεων (x_i,y_i) i=1,2…,n και ότι αναζητούμε προσέγγιση της μορφής:

$$Y_i=\beta_{0+}\beta_1 X_i+\varepsilon_i$$

όπου τα ε_i θα παριστάνουν τις αποκλίσεις της πραγματικής τιμής από την προηγούμενη. Είναι φανερό ότι η εκλογή των παραμέτρων β_0 και β_1 θα πρέπει να γίνει κατά τέτοιο τρόπο ώστε να ελαχιστοποιηθούν οι ποσότητες ε_i. Για να αποφευχθεί το πρόβλημα των θετικών – αρνητικών αποκλίσεων αρχικά υψώνουμε τα ε_i στο τετράγωνο και στη συνέχεια αθροίζουμε φτάνοντας έτσι στο άθροισμα τετραγώνων το οποίο προσπαθούμε να ελαχιστοποιήσουμε. Εφαρμόζοντας την παραπάνω μέθοδο καταλήγουμε στα $\widehat{\beta}_0$ και $\widehat{\beta}_1$ που είναι γνωστά με την ονομασία *εκτιμήτριες ελάχιστων τετραγώνων* και δίνονται από τους ακόλουθους τύπους:

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

$$\widehat{\beta}_0 = \frac{1}{n}\left(\sum y_i - \widehat{\beta}_1 \sum x_i\right) = \overline{y} - \widehat{\beta}_1 \overline{x}$$

όπου

$$S_{xy} = \sum \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$$

$$S_{xx} = \sum \left(x_i - \overline{x}\right)^2$$

ή καλύτερα υπολογιστικά

$$S_{xy} = \sum x_i y_i - n\overline{x}\overline{y}$$

$$S_{xx} = \sum x_i^2 - n\overline{x}^2$$

- *Κεφάλαιο 2: Πολλαπλή Παλινδρόμηση*

▪ *2.1: Γενική Εισαγωγή στην Πολλαπλή Παλινδρόμηση*

Σε πολλά πρακτικά προβλήματα είναι απαραίτητο να χρησιμοποιήσουμε δύο και περισσότερες ανεξάρτητες μεταβλητές προκειμένου να ερμηνεύσουμε με μεγαλύτερη ακρίβεια ένα φυσικό φαινόμενο ώστε να βγάλουμε σωστότερα συμπεράσματα. Για παράδειγμα προκείμενου να χρησιμοποιηθεί ένα μοντέλο παλινδρόμησης για να προβλεφθεί η ζήτηση ενός προϊόντος μιας εταιρίας σε 25 διαφορετικές πόλεις είναι ίσως σκόπιμο να χρησιμοποιηθούν κοινωνικοοικονομικές μεταβλητές (μέσο οικογενειακό εισόδημα, μόρφωση του αρχηγού της οικογένειας και μέσος αριθμός και χρόνος εκπαίδευσης), δημογραφικές μεταβλητές (μέσο μέγεθος οικογενειών, ποσοστό συνταξιούχων) και περιβαλλοντικές μεταβλητές (μέση ημερήσια θερμοκρασία, δείκτης ατμοσφαιρικής ρύπανσης).

Μοντέλα παλινδρόμησης που περιέχουν δύο ή περισσότερες ανεξάρτητες μεταβλητές ονομάζονται ***Μοντέλα Πολλαπλής Παλινδρόμησης (multiple regression models***).

▪ *2.2: Μοντέλο με Δύο ανεξάρτητες μεταβλητές*

Το μοντέλο με δύο ανεξάρτητες μεταβλητές είναι η φυσική επέκταση της απλής ευθείας παλινδρόμησης ώστε να μελετώνται δύο ανεξάρτητες μεταβλητές X_1 και X_2. Έτσι θα έχουμε :

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad i= 1,2,\ldots,n \qquad (1)$$

όπου:

-Y_i είναι η τιμή της εξαρτημένης μεταβλητής στην i παρατήρηση.

-X_{i1} και X_{i2} είναι τιμές των ανεξάρτητων μεταβλητών X_1 και X_2 στην i παρατήρηση οι οποίες υποτίθεται ότι είναι γνωστές σταθερές.

-α, β_1 και β_2 είναι οι παράμετροι του μοντέλου.

- Τα ε_i είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την Κανονική κατανομή $N(0,\sigma^2)$.

Η ***συνάρτηση παλινδρόμησης (regression function)*** ή αλλιώς ***συνάρτηση ανταπόκρισης (response function)*** του μοντέλου (1) είναι:

$$E\left(Y \middle| x_1, x_2\right) = a + \beta_1 x_1 + \beta_2 x_2$$

Η συνάρτηση αυτή ονομάζεται αρκετές φορές ***επιφάνεια παλινδρόμησης (regression surface) ή επιφάνεια ανταπόκρισης (response surface)***. Οι παράμετροι της πολλαπλής παλινδρόμησης έχουν ερμηνείες ανάλογες με αυτές της γραμμικής παλινδρόμησης. Έτσι στην επιφάνεια παλινδρόμησης :

- Το α αντιστοιχεί στο σημείο τομής του άξονα του Υ από την επιφάνεια (επίπεδο) παλινδρόμησης.
- Το β_1 δείχνει την μεταβολή της Ε(Υ) όταν το x_1 αυξάνει κατά μια μονάδα ενώ το x_2 παραμένει σταθερό
- Το β_2 δείχνει την μεταβολή της Ε(Υ) όταν x_2 αυξάνει κατά μια μονάδα ενώ το x_1 παραμένει σταθερό.

2.3: Μοντέλο με k ανεξάρτητες μεταβλητές

Το μοντέλο παλινδρόμησης με k ανεξάρτητες μεταβλητές $X_1,X_2,\ldots X_\kappa$ θα έχει τη μορφή

$$Y_i = \alpha + \beta 1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_\kappa X_\kappa + \varepsilon_i \qquad i=1,2,\ldots,n \qquad (1.2)$$

όπου

-Y_i είναι η τιμή της εξαρτημένης (dependent) ή μεταβλητής απόκρισης (response) για την i παρατήρηση.

-X_{i1}, $X_{i2,\ldots,}X_{ik}$ είναι οι τιμές των ανεξάρτητων (indepedent) ή «προβλεπουσών» (predictor) μεταβλητών στην i παρατήρηση (υποτίθενται γνωστές σταθερές).

- β_0, $\beta_1,\ldots,\beta_k$ είναι (k+1) άγνωστες παράμετροι (συντελεστές παλινδρόμησης) που ζητείται να εκτιμηθούν.

-Τα ε_i είναι ανεξάρτητα σφάλματα και ακολουθούν την Κανονική κατανομή $N(0,\sigma^2)$.

Και στην περίπτωση αυτή α είναι η Ε(Υ) για $X_1 = X_2 = \ldots X_\kappa = 0$ ενώ το β_i (i= 1,2,…k) δείχνει την μεταβολή της Ε(Υ) όταν η μεταβλητή X_i αυξηθεί κατά μια μονάδα ενώ όλες οι άλλες ανεξάρτητες μεταβλητές παραμένουν σταθερές.

Ένα σημαντικό ερώτημα σε πάρα πολλά προβλήματα, σχεδόν κάθε είδους, όπως παραγωγής (βιομηχανική, αγροτική κ.λπ.), εκπαίδευσης (μαθητών, στελεχών, στρατιωτών κλπ.) πρόβλεψης (εκλογές, καιρός κλπ.) χωροθέτησης, βελτιστοποίησης και άλλων, είναι αν μπορούμε να εκτιμήσουμε ή να προβλέψουμε την τιμή μιας ή περισσοτέρων ‘’μεταβλητών’’ κάτω από ορισμένες συνθήκες. Οι δοσμένες συνθήκες περιγράφονται και αυτές από μεταβλητές, οι τιμές των τιμών είναι δυνατό να ελεγχθούν από τον ερευνητή. Έτσι για παράδειγμα η μεταβλητή Υ που ζητούμε να εκτιμηθεί ή να προβλεφθεί, μπορεί να παριστάνει ’’ζήτηση κάποιου προϊόντος στην αγορά’’, ‘’παραγωγή κάποιου γεωργικού προϊόντος’’, ‘’απόδοση μαθητού’’, αύξηση ποσοστού σε εκλογές’’, κλπ. Ενώ οι μεταβλητές X_i που περιγράφουν τις συνθήκες και που μπορούν να ελεγχθούν, μπορεί να παραστάνουν ‘’τιμή πώλησης προϊόντος’’, ‘’συσκευασία’’, ‘’κόστος διαφήμισης’’, ‘’ταχύτητα διανομής’’, ‘’ποικιλία’’, ‘’λίπανση’’, ‘’θερμοκρασία’’, ‘’είδος διδασκαλίας’’, ‘’φύλο’’, και πολλά άλλα. Για την εύρεση του μοντέλου εκτίμησης ή πρόβλεψης χρησιμοποιούνται δεδομένα που έχουν προκύψει από μια σειρά n παρατηρήσεων και που συχνά δίνονται με τη μορφή του παρακάτω πίνακα:

$$\begin{bmatrix} x_{11} & x_{21} & x_{31} & \dots & x_{k1} & y_1 \\ x_{12} & x_{22} & x_{32} & \dots & x_{k2} & y_2 \\ .. & .. & .. & \dots & .. & .. \\ x_{1n} & x_{2n} & x_{3n} & \dots & x_{kn} & y_n \end{bmatrix} \quad (1.1)$$

Οι γραμμές του πίνακα παριστάνουν τις παρατηρήσεις, ενώ οι στήλες δίνουν τις τιμές των αντίστοιχων μεταβλητών για κάθε παρατήρηση. Η μορφή του μοντέλου πρόβλεψης μπορεί να είναι οποιαδήποτε, εδώ όμως θα ασχοληθούμε μόνο με το γραμμικό μοντέλο. Το γενικό γραμμικό μοντέλο είναι το:

Για τις μεταβλητές X_1, X_2,…X_k η ονομασία «ανεξάρτητες» δε σημαίνει ότι είναι πράγματι ανεξάρτητες. Μπορεί για παράδειγμα να ισχύει $X^2=X_1^2$ ή $X_3=X_1+X_2$. Ο λόγος στον οποίο οφείλεται αυτή η ονομασία, είναι ότι το ζητούμενο συνήθως είναι, πως οι τιμές αυτών των μεταβλητών επηρεάζουν τις τιμές της εξαρτημένης μεταβλητής και μπορούν να ελέγχονται από τον ερευνητή. Από το τελευταίο αυτό προκύπτει και η ονομασία «προβλέπουσες» μεταβλητές. Στην πράξη πολλές φορές οι μεταβλητές εναλλάσσουν ρόλους. Μια μεταβλητή, πχ που στο πρώτο μέρος μιας μελέτης είναι εξαρτημένη, μπορεί στο δεύτερο μέρος της μελέτης να είναι ποσοτικές, να περιγράφουν δηλ. μετρήσιμα μεγέθη. Το σφάλμα ε, περιέχει κάθε απόκλιση της πραγματικής κατάστασης από το μοντέλο. Έτσι εκτός από τα πιθανά σφάλματα μετρήσεων, περιέχει επίσης και σφάλματα προσαρμογής, που οφείλονται είτε σε παράλειψη μεταβλητών είτε σε χρήση μεταβλητών που δε σχετίζονται με την Y. Η δυνατότητα των προβλεπουσών μεταβλητών να συσχετίζονται μεταξύ τους διευρύνει τις περιπτώσεις εφαρμογής του μοντέλου (1.2).

▪ *2.4: Μέθοδος των ελαχίστων Τετραγώνων: Εκτίμηση των παραμέτρων του γραμμικού μοντέλου*

1. *- Μοντέλο Δύο μεταβλητών*

Η εκτιμήτρια της επιφάνειας παλινδρόμησης

$$E\left(Y \middle| x_1, x_2\right) = \mu_{Y|x_1,x_2} = a + \beta_1 x_1 + \beta_2 x_2$$

θα είναι η επιφάνεια

$$\hat{\mu}_{Y|x_1,x_2} = \hat{a} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

Οι τιμές a,b_1,b_2 των εκτιμητριών $\hat{a}$, $\hat{\beta}_1$ και $\hat{\beta}_2$ προκύπτουν κατά τα γνωστά με την μέθοδο των ελαχίστων τετραγώνων.

Το σύστημα των κανονικών εξισώσεων για τα a,b_1,b_2 είναι

$$\sum y_i = \nu \cdot a + \left(\sum x_{i1}\right)b_1 + \left(\sum x_{i2}\right)b_2$$

$$\sum x_{i1} y_i = \left(\sum x_{i1}\right)a + \left(\sum x_{i1}^2\right)b_1 + \left(\sum x_{i1}x_{i2}\right)b_2$$

$$\sum x_{i2} y_i = \left(\sum x_{i2}\right)a + \left(\sum x_{i2}^2\right)b_2 + \left(\sum x_{i1}x_{i2}\right)b_2$$

Η επίλυση των κανονικών αυτών εξισώσεων λόγω της πολυπλοκότητας των πράξεων γίνεται στον υπολογιστή.

Εκτός από την προσαρμογή με την μέθοδο των ελαχίστων τετραγώνων κάποιου μοντέλου σε μια σειρά από δεδομένα υπάρχει το πρόβλημα στην πολλαπλή παλινδρόμηση το κατά πόσον μερικοί από τους όρους $\beta_i x_i$ στο μοντέλο έχουν σημαντική συνεισφορά στην εξήγηση της διακύμανσης που παρατηρείται στην εξαρτημένη μεταβλητή Y_i. Η πολλαπλή παλινδρόμηση παρέχει τη στατιστική συμπερασματολογία για τον καθορισμό του κατά πόσο μια μεταβλητή είναι σημαντική με έλεγχο της μηδενικής υπόθεσης $H_0 : \beta_i = 0$ έναντι της εναλλακτικής $H_1 : \beta_i \neq 0$, i= 1,2,…, k. Αν η H_0 δεν απορριφθεί για κάποια τιμή του i συμπεραίνουμε ότι δεν υπάρχουν στοιχεία ικανά να μας πείσουν ότι η μεταβλητή έχει συνεισφορά σημαντική στο μοντέλο. Στην περίπτωση αυτή ο όρος $\beta_i x_i$ διαγράφεται από το μοντέλο. Στην περίπτωση αυτή ο όρος $\beta_i x_i$ διαγράφεται από το μοντέλο απλοποιώντας έτσι τη διαδικασία.

Πρέπει να σημειώσουμε ότι οι έλεγχοι υποθέσεων είναι μέθοδοι που βοηθούν τον ερευνητή να καθορίσει τη σημαντικότητα μεταβλητών του μοντέλου. Θα πρέπει όμως να τονισθεί ότι η απόφαση για το κατά πόσον μια μεταβλητή θα πρέπει να περιληφθεί στο μοντέλο ή όχι δεν θα πρέπει να ληφθεί με αποκλειστικό κριτήριο τον προηγηθέντα έλεγχο υποθέσεων. Οποιαδήποτε πρόσθετη πληροφορία είναι διαθέσιμη στον ερευνητή η οποία μπορεί να θεωρηθεί από αυτόν περισσότερο πειστική από ότι ο έλεγχος υποθέσεων δεν θα πρέπει να αγνοείται. Ο έλεγχος της υποθέσεως που προαναφέραμε στηρίζεται στη στατιστική συνάρτηση

$$T = \frac{\widehat{\beta}_i}{S_{\widehat{\beta}_i}}$$

όπου $\widehat{\beta}_i$ είναι η εκτιμήτρια ελαχίστων τετραγώνων του συντελεστή β_i της μεταβλητής X_i στο γενικό γραμμικό μοντέλο και $S_{\widehat{\beta}_i}$ είναι η εκτιμούμενη τυπική απόκλιση της εκτιμήτριας $\widehat{\beta}_i$. Όπως συνήθως η τιμή της στατιστικής συνάρτησης T συγκρίνεται με τα ποσοστιαία σημεία της κατανομής t με n-k-1 βαθμούς ελευθερίας. Οι υπολογισμοί αυτοί γίνονται συνήθως στους υπολογιστές.

- Μοντέλο k μεταβλητών

Έστω ότι σε κάποιο πρόβλημα έχουμε τα δεδομένα (1.2). Τότε όπως αναφέρθηκε, θα ικανοποιείται το σύστημα (1.3) δηλ. το

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + ... + \beta_k x_{k1} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + ... + \beta_k x_{k2} + \varepsilon_2 \\ &..... \\ y_n &= \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + ... + \beta_k x_{kn} + \varepsilon_n \end{aligned} \qquad (1.4)$$

Το ζητούμενο είναι να λυθεί το σύστημα (1.4) ως προς τα β_i με τον «καλύτερο» δυνατόν τρόπο δηλ. με έναν τρόπο που να ελαχιστοποιεί όσο είναι δυνατόν τα σφάλματα. Υπάρχουν πολλά κριτήρια ελαχιστοποίησης των σφαλμάτων, ένα από τα οποία είναι αυτό που απαιτεί το άθροισμα των τετραγώνων των σφαλμάτων $\sum_{i=1}^{n} \varepsilon_i^2$ να γίνεται ελάχιστο. Η μέθοδος που χρησιμοποιείται για την εκτίμηση των β_i μ' αυτό το κριτήριο λέγεται «***μέθοδος ελαχίστων τετραγώνων***». Οι εκτιμήσεις των β_i θα συμβολίζονται με $\widehat{\beta}_i$ και μ' αυτές το μοντέλο παίρνει τη μορφή:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + ... + \widehat{\beta}_k X_k \qquad (1.5)$$

Η τιμή $\widehat{Y}$ είναι η εκτίμηση της πραγματικής τιμής τη Y, όταν δίνονται οι τιμές X_1, X_2,…,X_k και διαφέρει απ' αυτήν κατά ένα σφάλμα. Αν οι X_j πάρουν τις τιμές X_{ji} που δίνονται στην i γραμμή του πίνακα (1.1), τότε η $\widehat{Y}$ συμβολίζεται με $\widehat{Y}_i$ και ισχύει $\varepsilon_i = Y_i - \widehat{Y}_i$

όπου $\sum_{i=1}^{n} \varepsilon_i^2$ είναι ελάχιστο. Η διαφορά $Y_i - \widehat{Y}$ λέγεται ***υπόλοιπο (residual***).

1. Η διαφορά των δύο τιμών Y και $\widehat{Y}_i$ διαιρεμένη με την εκτιμηθείσα τυπική απόκλιση των σφαλμάτων λέγεται **τυποποιημένο υπόλοιπο (standardized residual).**

▪ *2.5 Μέθοδος σταθμισμένων ελαχίστων τετραγώνων*

Όπως είδαμε υπάρχουν περιπτώσεις όπου η βασική προϋπόθεση της σταθερότητας της διασποράς δεν ικανοποιείται. Μπορούμε στη γενική περίπτωση να υποθέσουμε ότι ισχύει

$$\underset{\square}{Y} = X \underset{\square}{\beta} + \underset{\square}{\varepsilon} \qquad (3.18)$$

όπου

$$E\left(\varepsilon\right)=0\,,\;\; V\left(\varepsilon\right)=V\sigma^2 \qquad (3.19)$$

ή αν ενδιαφερόμαστε για ελέγχους υποθέσεων (*F – τεστ*) ότι

$$\varepsilon \sim N(0,V\sigma^2) \qquad (3.20)$$

▪ *2.6: Έλεγχοι υποθέσεων- Ένας μερικός έλεγχος του μοντέλου*

Είδαμε ότι κάτω από τις συνθήκες κανονικότητας $\varepsilon \sim N\left(0,\sigma^2 I_n\right)$ και με την βοήθεια του στατιστικού F=MSR/MSE μπορούμε να ελέγχουμε συνολικά το μοντέλο (1.2). Αν ο λόγος F είναι σημαντικός, το μοντέλο είναι «καλό», ενώ αν είναι ασήμαντος, τότε είτε η Y δεν εξαρτάται καθόλου από τις $X_1 \ldots X_K$ ή η σχέση εξάρτησης είναι διαφορετική από την (1.2).

Μέχρι τώρα μιλήσαμε για δύο μεθόδους ελέγχου υποθέσεων στην πολλαπλή παλινδρόμηση. Ο ένας αναφερόταν στον έλεγχο των συγκεκριμένων όρων του μοντέλου (όταν ελέγχεται η υπόθεση $H_0 : \beta_1 = \ldots \beta_k = 0$) που περιλαμβάνει όλους τους όρους του γενικού γραμμικού μοντέλου. Υπάρχει και μια τρίτη στατιστική μεθοδολογία που βρίσκεται στο ενδιάμεσο των δυο προαναφερθεισών μεθοδολογιών. Η μεθοδολογία αυτή επιτρέπει τον ταυτόχρονο έλεγχο ενός αριθμού από τους όρους του μοντέλου χωρίς ταυτόχρονα να απαιτεί να ελεγχθούν όλοι οι όροι του μοντέλου. Η μεθοδολογία αυτή είναι χρήσιμη όταν ο ερευνητής ξέρει ότι κάποιοι από τους όρους πρέπει οπωσδήποτε να χρησιμοποιηθούν αλλά είναι βέβαιος για έναν αριθμό από τους υπόλοιπους όρους του μοντέλου και θεωρεί ότι χρειάζεται έναν έλεγχο για να αποφασίσει για όλους του υπόλοιπους όρους ταυτόχρονα. Πιο συγκεκριμένα αν β_1, β_2,…, β_q είναι οι συντελεστές των όρων για τους οποίους είμαστε βέβαιοι ότι θα πρέπει να περιλαμβάνονται στο μοντέλο και $\beta_{q+1} = \beta_{q+2} = \ldots = \beta_k = 0$.

Με την υπόθεση αυτή ελέγχουμε ουσιαστικά το κατά πόσον το πλήρες μοντέλο

$$E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q + \beta_{q+1} x_{q+1} + \ldots + \beta_k X_k$$

είναι κατάλληλο για να περιγράψει τα δεδομένα. Το άθροισμα των τετραγώνων των λαθών του μοντέλου αυτού συμβολίζεται με SSE_1 και οι βαθμοί ελευθερίας για το λάθος συμβολίζονται με DF_1. Το περιορισμένο μοντέλο που μας ενδιαφέρει να εξετάσουμε είναι

$$E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_q X_q,\ q < k$$

Το άθροισμα των τετραγώνων των λαθών για το περιορισμένο αυτό μοντέλο συμβολίζεται με SSE_2 και οι βαθμοί ελευθερίας του λάθους συμβολίζεται με DF_2. Η στατιστική συνάρτηση F που χρησιμοποιείται για τον έλεγχο αυτής της υπόθεσης είναι η

$$F = \frac{\left(SSE_2 - SSE_1\right)/\left(DF_2 - DF_1\right)}{SSE_1 / DF_1}$$

Η τιμή της στατιστικής αυτής συνάρτησης συγκρίνεται με τα εκατοστιαία σημεία της κατανομής F που δίνονται στους αντίστοιχους πίνακες με (DF_2 _ DF_1) βαθμούς ελευθερίας του αριθμητή και DF_1 βαθμούς ελευθερίας για τον παρονομαστή.

▪ *2.7: Συντελεστής Προσδιορισμού- Μερικός συντελεστής Προσδιορισμού – Συσχέτισης*

Η σχέση SST=SSR+SSE εκφράζει ότι η συνολική διασπορά γύρω από το μέσο όρο αναλύεται σε δύο μέρη. Στο SSR που εξηγείται από τη γραμμική σχέση και στο SSE, που οφείλεται είτε σε σφάλματα είτε σε άλλους απροσδιόριστους παράγοντες. Ένα μοντέλο, επομένως θα είναι τόσο περισσότερο κατάλληλο για τα δεδομένα μας, όσο το SSE είναι πιο μικρό συγκρινόμενο με το SSR.

Το τελευταίο οδηγεί στον παρακάτω ορισμό. Η ποσότητα

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (1.43)$$

λέγεται **συντελεστής προσδιορισμού** (***Coefficient of determination***). Είναι ένας αριθμός μεταξύ 0 και 1, και μάλιστα όσο πλησιάζει το 1 τόσο η προσαρμογή είναι καλύτερη ενώ όσο πλησιάζει το 0 γίνεται χειρότερη. Ο αριθμός $\%R^2$ εκφράζει το ποσοστό της συνολικής μεταβλητότητας που εξηγείται από το γραμμικό μοντέλο.

Ο συντελεστής προσδιορισμού R^2 μπορεί να χρησιμοποιηθεί και για τον έλεγχο της υπόθεσης $H_0 : \beta_1=\beta_2=\ldots\beta_\kappa=0$ αντί του F. Πράγματι μπορεί εύκολα να διαπιστωθεί ότι

$$F = \frac{n-k-1}{k} \cdot \frac{R^2}{1-R^2} \quad \text{ή}$$

$$R^2 = \frac{kF/(n-k-1)}{1+kF/(n-k-1)}$$

Όταν το δείγμα είναι μικρό σε σύγκριση με τον αριθμό των μεταβλητών, ένας «καλύτερος» συντελεστής προσδιορισμού υπολογίζεται με αντικατάσταση στη δεύτερη από τις σχέσεις (1.43) των αθροισμάτων τετραγώνων με μέσα τετράγωνα. Ο συντελεστής που προκύπτει συμβολίζεται με $\bar{R}^2$ και λέγεται **«διορθωμένος συντελεστής προσδιορισμού»** (adjusted coefficient of determination), και ισχύει

$$\bar{R}^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = 1 - \frac{s^2}{\operatorname{var} Y} \qquad (1.45)$$

Ο συμβολισμός $\bar{R}^2$,δε σημαίνει ότι ο διορθωμένος συντελεστής προσδιορισμού είναι πάντα θετικός αριθμός. Πράγματι μπορεί το $\bar{R}^2$ να πάρει και αρνητικές τιμές αρκεί το s^2 να είναι μεγαλύτερο του VarY, πράγμα που συμβαίνει κάποιες φορές στην πράξη. Η σχέση μεταξύ των δύο συντελεστών είναι

$$1-\bar{R}^2=\frac{n-1}{n-k-1}\left(1-R^2\right)$$

Η τετραγωνική ρίζα του R^2 λέγεται **συντελεστής πολλαπλής συσχέτισης** (***coefficient of multiple correlation***) και συμβολίζεται με R. Πολλοί ερευνητές προκειμένου να αναφερθούν στην προσαρμογή του μοντέλου, προτιμούν να χρησιμοποιούν τον συντελεστή R αντί του R^2. Ο λόγος γίνεται φανερός αν σκεφθούμε ότι επειδή $0<R^2<1$ ισχύει $R>R^2$, η φυσική όμως σημασία του R και ιδιαίτερα του προσήμου του είναι δύσκολο να εξηγηθεί.

Ας θεωρήσουμε πάλι το πλήρες μοντέλο

$Y = \beta_0 + \beta_1X_1 + \beta_2\chi_2 + \ldots + \beta_kX_k + \varepsilon$

και το περιορισμένο που προκύπτει από αυτό με την απαλοιφή της μεταβλητής X_i

$Y = \beta_0 + \beta_1X_1 + \ldots + \beta_{i-1}X_{i+1} + \beta_{i-1}X_{i-1} + \beta_{t+1}X_{i+1} + \beta_kX_k + \varepsilon.$

Αν SSE και $SSE_{(i)}$ είναι το άθροισμα τετραγώνων των αποκλίσεων που μένουν ανεξήγητα από την προσαρμογή του πλήρους και του περιορισμένου μοντέλου αντίστοιχα, τότε προφανώς η διαφορά $SSE_{(i)}$ - SSE δίνει το μέρος της ανεξήγητης διασποράς που εξηγεί η μεταβλητή X_i και λέγεται ***μερικός συντελεστής προσδιορισμού (coefficient of partial determination)*** και θα το συμβολίσουμε με $r^2_{Y_i;12\ldots(i-1)(i+1)\ldots k}$ δηλαδή:

$$r^2_{Y_i;12\ldots(i-1)(i+1)\ldots k}=\frac{SSE_{(i)}-SSE}{SSE_{(i)}}$$

Ο συντελεστής αυτός συμβολίζεται επίσης και με $r^2_{Y_i;k}$ όπου ο υποδείκτης k σημαίνει ότι οι προβλέπουσες μεταβλητές μαζί με την X_i είναι σε πλήθος k. Ο συμβολισμός αυτός δεν δείχνει ποιες από τις μεταβλητές είναι στο μοντέλο, πράγμα που δημιουργεί μερικές φορές σύγχυση, ιδιαίτερα όταν μελετούμε ταυτόχρονα πολλά μοντέλα με το ίδιο πλήθος μεταβλητών. Η θετική τετραγωνική ρίζα του μερικού συντελεστή προσδιορισμού συμβολίζεται $r_{Y_i;12\ldots(i-1)(i+1)\ldots k}$ και λέγεται ***μερικός συντελεστής συσχέτισης (partial correlation coefficient)***.

▪ *2.8: Συντελεστής Μερικής Συσχέτισης (Partial Correlation Coefficient)*

Στην απλή παλινδρόμηση η σχέση μεταξύ Y και X_1 μπορεί να μετρηθεί από το δειγματικό συντελεστή συσχέτισης r_{YX_1}. Ομοίως η ένταση της σχέσης της X_1 με το Y, λαμβάνονται υπόψη την επίδραση της X_2 στην X_1, μπορεί να συνοψισθεί από το *δειγματικό συντελεστή συσχέτισης* των καταλοίπων της Παλινδρόμησης του Y στην X_2. Ο τελευταίος αυτός συντελεστής συσχέτισης ονομάζεται ***συντελεστής μερικής συσχέτισης*** και συμβολίζεται με $r_{YX_1|X_2}$. Πολλές φορές χρησιμοποιείται ο όρος μερικός συντελεστής συσχέτισης μεταξύ του Y και του X_1 προσαρμοζόμενος για το X_2 (partial correlation between Y and X_1 adjusted for X_2).

▪ *2.9: Πρόβλεψη και παρεμβολή*

Έστω ότι για τις μεταβλητές Y, X_1, $X_{2,...,}$ X_k έχουμε κάνει ένα σύνολο παρατηρήσεων που δίνονται με ένα πίνακα της μορφής (1.1). Ενδιαφερόμαστε να εκτιμήσουμε την τιμή Y_0 της Y, όταν οι τιμές των X_i είναι δοσμένες π.χ. $X_{10,}$ X_{20},…,X_k αντίστοιχα. Μια τέτοια πρόβλεψη της τιμής της Y, είναι αξιόπιστη όταν οι δοσμένες τιμές των X_i είναι μέσα στην περιοχή που «καλύπτεται» από τα δεδομένα. Αν k=1, η πρόβλεψη σ' αυτή την περίπτωση λέγεται ***παρεμβολή*** (***interpolation***). Για k > 1 χρησιμοποιούμε γενικά τον όρο ***πρόβλεψη*** (***prediction - forecasting***).

Με την υπόθεση ότι το μοντέλο που προσαρμόζεται στα δεδομένα είναι $Y = X\beta + \varepsilon$ και ότι τα σφάλματα ικανοποιούν τις συνθήκες $E(\varepsilon) = 0$ & $V(\varepsilon) = \sigma^2 I_n$, εκτιμούμε το διάνυσμα των παραμέτρων $\hat{\beta}$.

Αν, τώρα συμβολίσουμε με SST, το συνολικό άθροισμα τετραγώνων των αποκλίσεων γύρω από το μέσο όρο και με SSR το μέρος αυτού του αθροίσματος που απομένει από το SST μετά την αφαίρεση του SSE, τότε μπορούμε να πούμε ότι SSR είναι το μέρος της συνολικής διασποράς που εξηγείται από την παλινδρόμηση. Θα έχουμε επομένως: SST = SSR + SSE

Αν επομένως υποθέσουμε ότι όλες οι παράμετροι $\beta_1,\beta_2,\ldots\beta\kappa$ είναι ίσες με 0,τότε θα είναι και το SSR ίσο με 0, και κατά συνέπεια και ο λόγος

$$F = \frac{MSR}{MSE} = \frac{SSR / k}{SSE / (n-k-1)}$$

Ο λόγος F κατανέμεται με κατανομή $F_{k,n-k-1}$, πράγμα που οδηγεί στον παρακάτω έλεγχο:

Η υπόθεση H_0 : $\beta_1=\beta_2=\ldots=\beta_\kappa=0$ έναντι της εναλλακτικής της

H_1 : ένα τουλάχιστο από τα $\beta_1,\ldots\beta_\kappa \neq 0$,

ελέγχεται συγκρίνοντας την τιμή του λόγου F με την κρίσιμη τιμή $F_{K,n-k-1,\alpha}$ για

επίπεδο σημαντικότητας α. Τότε αν:

$F > F_{K,n-k-1,\alpha}$ απορρίπτουμε την H_0,σε ε.σ. α

$F < F_{K,n-k-1,\alpha}$ δεν απορρίπτουμε την H_0

Η μη απόρριψη της H_0 στην τελευταία περίπτωση, σημαίνει ότι από τα δεδομένα του πειράματος δεν προκύπτουν επαρκείς ενδείξεις για την ύπαρξη γραμμικής σχέσης μεταξύ των μεταβλητών Y και $X_1,X_{2,...}X_{K..}$ Αυτό βέβαια δεν αποκλείει την ύπαρξη κάποιας άλλης σχέσης, π.χ. πολυωνυμικής, εκθετικής κλπ. Μεταξύ των ίδιων μεταβλητών.

▪ *2.10: Ορθογωνιότητα (Orthogonality)*

Δύο μεταβλητές X_1 και X_2 λέγονται ***ορθογώνιες (orthogonal)*** αν η παλινδρόμηση της Y επί της X_1, προσαρμοσμένης για την X_2 ταυτίζεται με την παλινδρόμηση της Y επί της X_1 αγνοώντας την X_2. Η ευχάριστη αυτή κατάσταση συμβαίνει όταν ο δειγματικός συντελεστής συσχέτισης των X_1 και X_2 είναι ακριβώς μηδέν. Όταν η X_1 και η X_2 είναι ορθογώνιες, η επίδραση κάθε μιας από τις μεταβλητές είναι σαφώς καθορισμένη. Για το λόγο αυτό, όποτε είναι δυνατό τα πειράματα σχεδιάζονται με τρόπο ώστε οι μεταβλητές να είναι ορθογώνιες.

▪ *2.11: Σφάλμα προσαρμογής- Επαναλαμβανόμενες μετρήσεις*

Το άθροισμα τετραγώνων SSE οφείλεται, όπως αναφέρθηκε και προηγούμενα, είτε σε σφάλματα είτε σε άλλους παράγοντες. Δεν είναι εύκολο, στη γενική περίπτωση να ξεχωρίσουμε τα σφάλματα από το SSE. Αν όμως στα δεδομένα μας υπάρχουν επαναλαμβανόμενες μετρήσεις, τότε μπορούμε να υπολογίσουμε το μέρος του SSE που οφείλεται μόνο σε σφάλματα. Αν αυτό το μέρος είναι μεγάλο σε σύγκριση με το μέρος του SSE που οφείλεται και σε άλλους παράγοντες, τότε δεν μπορούμε να αμφισβητήσουμε το μοντέλο. Αν όμως το μέρος που οφείλεται σε σφάλματα είναι ασήμαντο σε σχέση με το άλλο μέρος, θα σημαίνει ότι το μοντέλο θέλει αναθεώρηση. Ένας έλεγχος των διαγραμμάτων διασποράς είναι στην περίπτωση αυτή χρήσιμος για την προσαρμογή καταλληλότερου μοντέλου. Ας υποθέσουμε λοιπόν ότι υπάρχουν επαναλαμβανόμενες μετρήσεις .Τότε στο σημείο πχ. P_1 για το οποίο $X_1=x_{11}$, $X_2=x_{21}$…, $X_K=x_{K1}$, θα έχουμε αντί μιας πολλές n_1 έστω παρατηρήσεις τις οποίες μπορούμε να συμβολίσουμε Y_{11}, Y_{12},…, Y_{1n}, $n_1 \geq 1$. Όμοια στο σημείο P_2 για το οποίο $X_1 = x_{12,}$ $X_2 = x_{22}$,…,$X_k = x_{k2}$ θα έχουμε n_2 παρατηρήσεις τις Y_{21} , Y_{22},…,$Y_{2n2,}$ $n_2 \geq 1$. Συνεχίζοντας ανάλογα βρίσκουμε ότι οι n παρατηρήσεις που θα έχουμε συνολικά, ταξινομούνται σε m ομάδες με n_j παρατηρήσεις η κάθε μια όπου $n_1 + n_2 + \ldots + n_m = n$. Μπορούμε επομένως συνοπτικά να γράψουμε τις n παρατηρήσεις όπως παρακάτω:

$$Y_{11}, Y_{12}, ..., Y_{1n_1} (X_1 = x_{11}, X_2 = x_{21}, ..., X_k = x_{k1})$$
$$Y_{21}, Y_{21}, ..., Y_{2n_2} (X_1 = x_{12}, X_2 = x_{22}, ..., X_k = x_{k2})$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$Y_{m1}, Y_{m2}, ..., Y_{m,n_m} (X_1 = x_{1m}, X_2 = x_{2m}, ..., X_k = x_{km})$$

Το μέρος του SSE που οφείλεται σε καθαρά σφάλματα, μόνο από τις n_1 παρατηρήσεις στο σημείο P_1 θα είναι το άθροισμα των τετραγώνων των αποκλίσεων γύρω από το μέσο όρο $\overline{Y}_1$, των παρατηρήσεων αυτών. Θα είναι δηλαδή

$$\sum_{s=1}^{n_1} \left(Y_{ls} - \overline{Y}_1 \right)^2 = \sum_{s=1}^{n_1} Y_{ls}^2 - n_1 \overline{Y}_1^2 = (n_1 - 1) s_{Yl}^2$$

όπου s_{Y_l} η τυπική απόκλιση των n_1 πρώτων παρατηρήσεων της Y.

Το συνολικό επομένως καθαρό σφάλμα θα είναι το άθροισμα $SS_e = \sum_{r=1}^{m} \sum_{s=1}^{n_r} \left(Y_{rs} - \overline{Y}_r \right)^2$

που έχει $n_e = \sum_{r=1}^{m} n_r - m$ βαθμούς ελευθερίας.

Το μέσο άθροισμα τετραγώνων που οφείλεται σε καθαρά σφάλματα θα είναι

$$s_e^2 = MS_e = \frac{\sum_{r=1}^{m} \sum_{s=1}^{n_r} \left(Y_{rs} - \overline{Y}_r \right)^2}{\sum_{r=1}^{m} n_r - m}$$

που είναι ένας εκτιμητής του σ^2, οποιοσδήποτε και αν είναι το μοντέλο, που προσαρμόζουμε στα δεδομένα μας.

Ας συμβολίσουμε τώρα MS_L το μέσο άθροισμα τετραγώνων που προκύπτει αν αφαιρέσουμε το καθαρό σφάλμα από το SSE και διαιρέσουμε με $n_L = (n-k-1)-n_e$. Τότε μπορεί να δειχτεί ότι ο λόγος $F = \frac{MS_L}{s_e^2} = \frac{\left(SSE - SS_e \right) / n_L}{s_e^2}$

ακολουθεί κατανομή F με n_L και n_e βαθμούς ελευθερίας.

Αν ο λόγος F είναι σημαντικός, τότε το μέρος του SSR που οφείλεται σε μη ελέγξιμους παράγοντες είναι σημαντικά μεγαλύτερο από το μέρος που οφείλεται σε καθαρά σφάλματα. Στην περίπτωση αυτή το s_e^2 είναι καλύτερη εκτίμηση του σ^2 απ' ότι το MSE. Μια πρώτη αντιμετώπιση είναι να προσθέσουμε δευτεροβάθμιους όρους στο μοντέλο και να ελέγξουμε την προσαρμογή του νέου μοντέλου. Αν ο λόγος F είναι ασήμαντος, τότε η ακρίβεια του μοντέλου δεν αμφισβητείται. Στην περίπτωση αυτή εκτιμούν το σ^2 πάλι από το MSE, αν και αυτή η εκτίμηση δεν είναι πολύ διαφορετική από την s_e^2.

▪ *2.12: Το πρόβλημα της Πολυσυγγραμμικότητας (Multicollinearity)*

Η ερμηνεία ενός προβλήματος με την χρησιμοποίηση της μεθόδου αναλύσεως της πολλαπλής παλινδρόμησης επιτυγχάνεται καλύτερα όταν *οι ανεξάρτητες μεταβλητές που αποτελούν το μοντέλο είναι μεταξύ τους ασυσχέτιστες*. Όταν υφίστανται *έντονες συσχετίσεις* μεταξύ των μεταβλητών είναι δύσκολο, αν όχι αδύνατο, να αξιολογηθεί η ουσιαστική προσφορά μιας συγκεκριμένης ανεξάρτητης μεταβλητής επί της εξαρτημένης που οφείλεται αποκλειστικά στη συγκεκριμένη ανεξάρτητη μεταβλητή. Όταν οι ανεξάρτητες μεταβλητές δεν είναι ορθογώνιες μεταξύ τους είναι ενδεχόμενο οι εκτιμούμενοι συντελεστές παλινδρόμησης να είναι εξαιρετικά ασταθείς και οι τιμές τους να υφίστανται δραματικές αλλαγές όταν κάποια νέα μεταβλητή προστίθενται ή απομακρύνεται ή όταν μικρές μεταβολές στα δεδομένα του προβλήματος. ***Η κατάσταση η οποία δημιουργείται όταν υπάρχουν ισχυρές συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών στην πολλαπλή παλινδρόμηση ονομάζεται πολυσυγγραμμικότητα (multicollinearity).*** Στις περιπτώσεις που το πρόβλημα αυτό υφίσταται θα πρέπει κανείς να είναι ιδιαίτερα προσεκτικός στην ερμηνεία όλων των εκτιμητριών που προκύπτουν από το μοντέλο αυτό. Υπάρχει μια σειρά από προειδοποιητικές ενδείξεις που αν ο ερευνητής τις προσέξει είναι δυνατόν να αντιληφθεί ότι υπάρχει πολυσυγγραμμικότητα. Η πιο σημαντική από αυτές είναι ο ***πίνακας των συντελεστών συσχετίσεως (Correlation Matrix)*** των ανεξάρτητων μεταβλητών. Αν στον πίνακα αυτόν υπάρχουν μεγάλες θετικές ή αρνητικές τιμές θα έχουμε μια ένδειξη ότι οι αντίστοιχες ανεξάρτητες μεταβλητές που χρησιμοποιούνται στο μοντέλο έχουν μεταξύ τους ισχυρό βαθμό συσχέτισης. Το στατιστικό συμπέρασμα που προκύπτει στις περιπτώσεις αυτές είναι ότι κάποιες από τις μεταβλητές συνεισφέρουν ελάχιστα ή καθόλου, στην πρόβλεψη της εξαρτημένης μεταβλητής οπότε και θα πρέπει να απομακρυνθούν από το μοντέλο. Εάν, παρ' όλα αυτά, ο ερευνητής είναι βέβαιος ότι ο καθαρισμός των ανεξάρτητων μεταβλητών έγινε σωστά και θα πρέπει να εξετάσει δύο άλλες ενδείξεις για το κατά πόσον υπάρχει πολυσυγγραμμικότητα:

- Εάν τα πρόσημα ορισμένων συντελεστών στην παλινδρόμηση είναι αντίθετα από αυτά που θα περίμενε κανείς λόγω της φύσης του προβλήματος και
- Εάν σημαντικοί συντελεστές της παλινδρόμησης εμφανίζονται να έχουν μεγάλες τιμές στις τυπικές αποκλίσεις τους

Οποιαδήποτε από τις δύο αυτές ενδείξεις θα πρέπει να προβληματίσει τον ερευνητή και να τον οδηγήσει σε μια σοβαρή έρευνα για το κατά πόσον υφίστανται πολυσυγγραμμικότητα. Ο καθορισμός εκείνων των γραμμικών συνδυασμών των παραμέτρων β που μπορούν να

εκτιμηθούν με ακρίβεια είναι εξαιρετικά δύσκολο στην περίπτωση που υφίστανται πολυσυγγραμμικότητα. Παρότι, εν γένει, δεν είναι δυνατόν να εξαλειφθεί τελείως το πρόβλημα αυτό υπάρχει μια διαδικασία με την οποία ο ερευνητής μπορεί να εργασθεί με ένα μοντέλο που προκύπτει από το αρχικό μετασχηματισμό των αρχικών μεταβλητών σε ένα σύνολο άλλων μεταβλητών που είναι ασυσχέτιστες μεταξύ τους. ***Η μεθοδολογία αυτή ονομάζεται ανάλυση κυρίων συνιστωσών (principal component analysis***). Η τεχνική αυτή είναι μια πάρα πολύ ισχυρή τεχνική τόσο στον εντοπισμό της πολυσυγγραμμικότητας όσο και ως μεθοδολογία που οδηγεί στον καθορισμό εκείνων των γραμμικών συνδυασμών των συντελεστών παλινδρόμησης που μπορούν να εκτιμηθούν με ακρίβεια. Μια άλλη προσέγγιση είναι να υπολογισθούν οι k συντελεστές προσδιορισμού των παλινδρομήσεων κάθε μιας από τις ανεξάρτητες μεταβλητές στις υπόλοιπες k-1 ανεξάρτητες μεταβλητές. Εκείνες οι μεταβλητές που εμφανίζουν υψηλό συντελεστή προσδιορισμού θα πρέπει να θεωρηθεί ότι είναι συγγραμμικές με τουλάχιστον μια από τις υπόλοιπες μεταβλητές. Στη συνέχεια θα πρέπει να υπολογισθεί η ομάδα εκείνων των μεταβλητών που έχουν υψηλή πολυσυγγραμμικότητα και μια ή περισσότερες από αυτές τις μεταβλητές μέσα στη συγκεκριμένη ομάδα θα πρέπει να απομακρυνθούν πριν προχωρήσει κανείς σε ανάλυση παλινδρόμησης για την αρχική εξαρτημένη μεταβλητή.

Συμπερασματικά, με τον όρο ***πολυσυγγραμμικότητα (multicollinearity***), εννοούμε την ύπαρξη μιας ανεξάρτητης μεταβλητής X_j που είναι γραμμικά συσχετισμένη με μια άλλη ανεξάρτητη μεταβλητή ή με ένα γραμμικό συνδυασμό άλλων ανεξάρτητων μεταβλητών. Η πολυσυγγραμμικότητα είναι αρκετά συχνό φαινόμενο, ιδιαίτερα όταν τα δεδομένα προέρχονται από κοινωνικές ή οικονομικές μελέτες και είναι από τις κυριότερες αιτίες για την εξαγωγή λαθεμένων συμπερασμάτων στην πολλαπλή γραμμική παλινδρόμηση. Η ύπαρξη πολυσυγγραμμικότητας συνεπάγεται την αύξηση των τυπικών σφαλμάτων των συντελεστών παλινδρόμησης. Μάλιστα αν υπάρχει πλήρης ή τέλεια πολυσυγγραμμικότητα αν δηλ. $X_j = \lambda_0 + \sum_{i \neq j} \lambda_i X_i$ τότε ο πίνακας σχεδιασμού X έχει βαθμό μικρότερο του k+1 δηλ. είναι ιδιάζων και επομένως δεν μπορούν να βρεθούν συντελεστές παλινδρόμησης. Όσο περισσότερο προσεγγίζεται η παραπάνω ακραία περίπτωση τόσο περισσότερο προσεγγίζεται η παραπάνω ακραία περίπτωση, τόσο περισσότερες υπολογιστικές δυσκολίες δημιουργούνται. Η απαλοιφή της μεταβλητής X_j δεν λύνει πάντα το πρόβλημα διότι σε αυτήν την περίπτωση οι συντελεστές παλινδρόμησης των υπολοίπων μεταβλητών δεν εκτιμούνται αμερόληπτα. Πράγματι ας θεωρήσουμε το μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad (2.1)$$

όπου οι μεταβλητές X_1, X_2 είναι γραμμικά συσχετισμένες. Αν η γραμμική συσχέτιση είναι πλήρης δηλ. αν $X_2 = \lambda + \mu X_1$ τότε όπως είναι γνωστό $r_{X_1, X_2} = \pm 1$ αλλιώς αν η γραμμική συσχέτιση είναι σχεδόν πλήρης τότε $\left| r_{X_1, X_2} \right|$ είναι περισσότερο ή λιγότερο κοντά στο +1.

Εν συνεχεία μπορούμε να δείξουμε ότι ισχύει

$$s(\hat{\beta}_1) = \sqrt{\frac{\sum \varepsilon_i^2}{n-3}} \left(\sum \left(X_{1i} - \bar{X}_1 \right)^2 \left(1 - r^2_{X_1, X_2} \right) \right)^{-1/2}$$

$$s(\hat{\beta}_2) = \sqrt{\frac{\sum \varepsilon_i^2}{n-3}} \left(\sum \left(X_{2i} - \bar{X}_2 \right)^2 \left(1 - r^2_{X_1, X_2} \right) \right)^{-1/2}$$

Αν $r_{X_1, X_2 =} \pm 1$ τότε $s(\hat{\beta}_1)$, $s(\hat{\beta}_2)$ απειρίζονται, ενώ όσο περισσότερο πλησιάζει το r_{X_1, X_2} την τιμή 1 τόσο περισσότερο μεγαλώνουν οι τυπικές αποκλίσεις των $\hat{\beta}_1$, $\hat{\beta}_2$. Το ίδιο συμβαίνει και όταν υπάρχουν περισσότερες μεταβλητές. Μπορεί όμως σε αυτήν την περίπτωση να έχουμε ακόμη και πλήρη πολυσυγγραμμικότητα, π.χ. $X_2 = X_3 + X_4$ χωρίς να έχουμε μεγάλες ανά δύο συσχετίσεις των προβλεπουσών μεταβλητών. Έτσι ο έλεγχος και μόνο του πίνακα των συσχετίσεων δεν αποκαλύπτει πάντα την ύπαρξη πολυσυγγραμμικότητας.

Ας υποθέσουμε τώρα ότι επειδή r_{X_1, X_2} είναι αρκετά μεγάλο, απαλείφουμε τη μεταβλητή X_2, θεωρώντας ως καταλληλότερο το μοντέλο $\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1$. Τότε μπορεί να δειχτεί ότι

$$E\hat{b}_1 = \beta_1 + \beta_2 \frac{s_{X_1}}{S_{X_2}} r_{X_1 X_2}$$

δηλαδή για να είναι ο $\hat{b}_1$ αμερόληπτος εκτιμητής του β_1 θα πρέπει είτε $\beta_2 = 0$ είτε $r_{X_1 X_2} = 0$.

▪ *2.13: Αμφικλινής Παλινδρόμηση (Ridge Regression)*

Μια εναλλακτική μέθοδος για την αντιμετώπιση της πολυγραμμικότητας είναι μέθοδος της ***αμφικλινούς παλινδρόμησης (ridge regression***). Η μέθοδος αυτή προτάθηκε από το Hoerl (1962) και αναλύθηκε με λεπτομερή τρόπο από τους Hoerl & Kennard (1970). Η διαδικασία που προτείνεται με τη μέθοδο αυτή έχει ως στόχο να ξεπεράσει τη δυσκολία που δημιουργείται από την ύπαρξη συσχέτισης μεταξύ των ανεξάρτητων μεταβλητών. Σύμφωνα με τν μέθοδο αυτή προστίθεται μια σταθερά Θ στα στοιχεία της διαγωνίου του πίνακα. Z΄Z όπου:

$$Z = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ x_{m1} & x_{m2} & \dots & x_{mk} \end{pmatrix}$$

(Z είναι δηλαδή ο πίνακας που έχει ως στήλες του τις παρατηρήσεις των ανεξάρτητων μεταβλητών $X_1,X_2,\dots,X_{k)}$.

Οι εκτιμήτριες των συντελεστών παλινδρόμησης της μεθόδου αυτής δίνονται από τον τύπο: $\widehat{\beta}^{*} = \left(Z'Z + kI\right)^{-1} Z'Y$

Οι εκτιμητές που προκύπτουν από την μέθοδο αυτή είναι πάντοτε μεροληπτικοί αλλά οι διασπορές τους είναι αισθητά μικρότερες από αυτές των εκτιμητριών ελαχίστων τετραγώνων. Οι λόγοι που οι εκτιμήτριες τετραγώνων χρησιμοποιούνται περισσότερο στις μεθόδους παλινδρόμησης οφείλονται στο ότι είναι απλές όσον αφορά τον υπολογισμό τους, γεωμετρικά αισθητές και με δεδομένο ότι κάποιες υποθέσεις ισχύουν με απόλυτο τρόπο είναι βέλτιστες για μια σειρά από κριτήρια. Οι εκτιμήτριες ελαχίστων τετραγώνων έχουν χρησιμοποιηθεί με επιτυχία για περισσότερα από 160 χρόνια. Τα τελευταία χρόνια, κυρίως λόγω της πληθώρας των ηλεκτρονικών υπολογιστών διαφόρων μορφών που είναι διαθέσιμοι, έχουν αναπτυχθεί και χρησιμοποιούνται μια σειρά από άλλες εκτιμήτριες κυρίως για να καλύψουν κάποιες συγκεκριμένες αδυναμίες της μεθόδου ελαχίστων τετραγώνων. (Για παράδειγμα οι εκτιμήτριες που αναφέραμε στη μέθοδο της αμφικλινούς παλινδρόμησης αναπτύχθηκαν για να ξεπερασθεί το πρόβλημα που δημιουργείται στις εκτιμήτριες της παλινδρόμησης από το πρόβλημα της πολυσυγγραμμικότητας)

▪ *2.14: Όριο ανοχής*

Ο χαρακτηρισμός μιας ανεξάρτητης μεταβλητής X_i ως σημαντικής για την πρόβλεψη των τιμών της Y, εξαρτάται όπως προκύπτει από τα προηγούμενα όχι μόνο από την εκτίμηση $\widehat{\beta}_i$ του συντελεστού παλινδρόμησης της X_i, αλλά και από το τυπικό σφάλμα με το οποίο εκτιμάται ο συντελεστής $\widehat{\beta}_i$ δηλ, από το $s\left(\widehat{\beta}_i\right)$. Συντελεστές με μεγάλα τυπικά σφάλματα

θεωρούνται μη αξιόπιστοι και μπορούν να παίρνουν πολύ διαφορετικές τιμές για διαφορετικά δείγματα.

Η ύπαρξη πολυσυγγραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών έχει ως αποτέλεσμα την εκτίμηση συντελεστών παλινδρόμησης με μεγάλα τυπικά σφάλματα. Πράγματι μπορεί να δειχτεί ότι $s^2\left(\hat{\beta}_i\right)=\frac{s^2}{(1-R_i^2)(N-1)s_i^2}$

όπου s^2 είναι η διασπορά των σφαλμάτων στο μοντέλο $Y=X\beta+\varepsilon$, s_i^2 είναι η διασπορά της μεταβλητής X_i και R_i^2 ο συντελεστής προσδιορισμού του μοντέλου $X_i=X_{(i)}\beta_{(i)}+\varepsilon$ που θεωρεί τη μεταβλητή X_i ως εξαρτημένη από τις υπόλοιπες ανεξάρτητες μεταβλητές. Με X_i και β_i συμβολίσαμε όπως και προηγούμενα τους πίνακες που προκύπτουν από τους X και β΄ με τη διαγραφή της στήλης τους που αντιστοιχεί στη μεταβλητή X_i. Η ποσότητα R_i^2 πλησιάζει την μονάδα αν υπάρχει μεγάλη πολυσυγγραμμικότητα στα δεδομένα αν δηλ. η μεταβλητή είναι περίπου γραμμικός συνδυασμός των υπόλοιπων ανεξάρτητων μεταβλητών. Στη περίπτωση αυτή είναι φανερό ότι η ποσότητα $1\text{-}R_i^2$ πλησιάζει το μηδέν και επομένως το τυπικό σφάλμα του συντελεστή $\hat{\beta}_i$ αυξάνει. Μάλιστα για σταθερό μέγεθος δείγματος και σταθερή διασπορά σφαλμάτων το τυπικό σφάλμα αυξάνει τόσο περισσότερο όσο μικρότερο γίνεται το $1-R_i^2$. Αν όμως το $1\text{-}R_i^2$ γίνει αρκετά μικρό τότε είναι ενδεχόμενο να έχουμε και υπολογιστικά προβλήματα. Έτσι ορίζουμε ανάλογα με την αριθμητική που χρησιμοποιούμε μια οριακή τιμή που λέγεται ***όριο ανοχής (tolerance bound***) και είναι τέτοια ώστε αν το $1\text{-}R_i^2$ γίνει μικρότερο από το όριο ανοχής να αποκλείσουμε την μεταβλητή X_i από το μοντέλο παλινδρόμησης. Το όριο ανοχής που δέχεται αυτόματα το SPSS στις νεώτερες παραλλαγές του είναι T=0.0001 ενώ σε παλιότερες ήταν 0.01. Υπάρχει τρόπος όμως να ορίσουμε ως όριο ανοχής και κάποια άλλη μεγαλύτερη τιμή ώστε να αποκλείσουμε μεταβλητές που έχουν μεγάλη πολυσυγγραμμικότητα. Στο πρόγραμμα REGRESSION SPSS η τιμή $1\text{-}R_i^2$ αναφέρεται ως ανοχή (TOLERANCE) της μεταβλητής X_i και η τιμή αυτή συγκρίνεται με το όριο ανοχής. Αν η «ανοχή της X_i» είναι μικρότερη του ορίου ανοχής T η μεταβλητή X_i δεν μπαίνει στο μοντέλο ακόμη και αν η συμμετοχή της θα εξηγούσε μεγάλο μέρος της ανεξήγητης διασποράς. Τέλος είναι δυνατό η προσθήκη μιας μεταβλητής στο μοντέλο με επιτρεπή ανοχή να ελαττώσει τόσο πολύ την ανοχή κάποιας άλλης μεταβλητής που υπήρχε ήδη στο μοντέλο και έτσι να δημιουργηθεί πρόβλημα. Γι' αυτό το SPSS για κάθε μεταβλητή που δεν είναι στο μοντέλο υπολογίζει όχι μόνο την δική της ανοχή που θα είχε αν συμμετείχε στο μοντέλο αλλά και όλως των άλλων μεταβλητών που είναι στο μοντέλο. Τυπώνει τέλος ως ελάχιστη ανοχή (MIN TOLER) τη μικρότερη από τις τιμές αυτές.

- *Κεφάλαιο 3: Η εξέταση των Υπολοίπων*

▪ *3.1: Εισαγωγή*

Τα υπόλοιπα e_i ορίζονται ως οι n διαφορετικές $\varepsilon_i = Y_i - \widehat{Y}_i$, i =1,2,……., n όπου Y_i είναι μια παρατήρηση και $\widehat{Y}_i$ είναι η αντίστοιχη προσαρμοσμένη τιμή που προσδιορίζεται από την προσαρμοσμένη εξίσωση παλινδρόμησης.

Από τον ορισμό αυτό μπορούμε να δούμε ότι τα υπόλοιπα e_i είναι οι διαφορές μεταξύ αυτού που πραγματικά παρατηρείται και αυτού που προβλέπεται από την εξίσωση παλινδρόμησης, δηλαδή είναι η ποσότητα την οποία η εξίσωση της παλινδρόμησης δεν είναι ικανή να ερμηνευτεί. Έτσι μπορούμε να θεωρήσουμε τα e ως τα παρατηρούμενα σφάλματα αν το μοντέλο είναι σωστό. Για να εκτελέσουμε τώρα την ανάλυση της παλινδρόμησης κάνουμε συγκεκριμένες υποθέσεις για τα σφάλματα, οι συνήθεις υποθέσεις είναι ότι τα σφάλματα είναι ανεξάρτητα, έχουν μέση τιμή μηδέν, σταθερή διασπορά σ^2 και ακολουθούν μια κανονική κατανομή. Η τελευταία υπόθεση απαιτείται για να κάνουμε F – ελέγχους. Έτσι αν το μοντέλο προσαρμογής είναι σωστό, τα υπόλοιπα θα πρέπει να παρουσιάζουν τάσεις που τείνουν να επιβεβαιώσουν τις υποθέσεις που κάναμε ή το λιγότερο δεν θα πρέπει να παρουσιάζουν μια άρνηση των υποθέσεων. Την τελευταία αυτή ιδέα θα πρέπει να τη διατηρούμε στο μυαλό μας όταν εξετάζουμε τα υπόλοιπα και οφείλουμε να διερωτηθούμε: ‘’Δείχνουν τα υπόλοιπα ότι οι υποθέσεις μας είναι λανθασμένες’’. Αφού εξετάσουμε τα υπόλοιπα τότε θα είμαστε σε θέση να συμπεράνουμε (1) ότι οι υποθέσεις φαίνεται ότι παραβιάζονται (με τρόπο που μπορεί να προσδιοριστεί) (2) ότι οι υποθέσεις δε φαίνεται να παραβιάζονται. Σημειώνουμε ότι η περίπτωση (2) δεν σημαίνει ότι καταλήγουμε στο συμπέρασμα ότι οι υποθέσεις είναι σωστές, απλά σημαίνει ότι με βάση τα διαθέσιμα δεδομένα δεν έχουμε λόγο να πούμε ότι οι υποθέσεις δεν είναι σωστές. Το ίδιο ισχύει και στις περιπτώσεις ελέγχου υποθέσεων όπου είτε απορρίπτουμε είτε δεν απορρίπτουμε (πάρα αποδεχόμαστε) τη μηδενική υπόθεση. Τώρα θα δώσουμε τρόπους εξέτασης των υπολοίπων για να ελέγξουμε το μοντέλο. Όλοι αυτοί οι τρόποι είναι γραφικοί, εύκολα εφαρμόσιμοι και συνήθως πολύ αποκαλυπτικοί όταν οι υποθέσεις παραβιάζονται. Οι βασικοί τρόποι για να παραστήσουμε γραφικά τα υπόλοιπα e είναι

1. Συνολικά.
2. Σε χρονική ακολουθία, αν είναι γνωστή η σειρά.
3. Ως προς τις προσαρμοσμένες τιμές $\widehat{Y}_i$.
4. Ως προς τις ανεξάρτητες μεταβλητές X_{ij} , για j = 1,2,…..k
5. Με κάθε τρόπο που είναι λογικός για το ειδικότερο πρόβλημα που εξετάζεται.

▪ *3.2: Συνολικό Διάγραμμα*

Αν το μοντέλο μας είναι σωστό τα υπόλοιπα αυτά θα πρέπει να μοιάζουν με παρατηρήσεις από μια κανονική κατανομή με μέση τιμή μηδέν.

Πρώτα σημειώνουμε ότι η μέση τιμή των υπολοίπων είναι μηδέν, αλλά αυτό αναγκαστικά συμβαίνει σε κάθε μοντέλο παλινδρόμησης που περιλαμβάνει ένα σταθερό όρο β_0. Αυτό εύκολα φαίνεται από την πρώτη κανονική εξίσωση που προκύπτει διαφορίζοντας το

άθροισμα τετραγώνων των σφαλμάτων ως προς β_0. Αν το μοντέλο προσαρμογής είναι $E(Y) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$, η εξίσωση μπορεί να γραφεί ως

$$-2\sum\left(Y_i - b_0 - b_1 X_{1i} - ... - b_k X_{ki}\right) = 0$$

όπου το άθροισμα λαμβάνεται για i=1,2,……..,n. Η εξίσωση αυτή ανάγεται στην

$$\sum\left(Y_i - \hat{Y}_i\right) = 0$$

Επομένως ισχύει

$$\sum e_i = \frac{\sum e_i}{n} = 0$$

Μια εναλλακτική διαδικασία είναι να κατασκευάσουμε είτε ένα ***κανονικό διάγραμμα (normal plot)*** είτε ένα ***ημικανονικό διάγραμμα (half normal plot***) των υπολοίπων σε χαρτί τυπικής πιθανότητας. Τα σημεία θα πέφτουν κατά προσέγγιση πάνω σε μια ευθεία γραμμή. Ωστόσο, για μια ακόμα φορά, απαιτούνται κανόνες αξιολόγησης του διαγράμματος.

Όταν ο αριθμός των υπολοίπων είναι πολύ μεγάλος, το συνολικό διάγραμμα μπορεί να γίνει σε μορφή *ιστογράμματος* αντί σε μορφή *στικτού διαγράμματος*. Σ αυτή την περίπτωση, το *κανονικό ή το ημικανονικό διάγραμμα* μπορεί να κατασκευαστεί μόνο από μια συλλογή των μικρότερων παρατηρήσεων, για παράδειγμα με (έστω, π.χ.) 200 παρατηρήσεις, στο ημικανονικό διάγραμμα μπορούμε να χρησιμοποιήσουμε τη $10^η$ μικρότερη, $20^η$ μικρότερη κτλ μέχρι (έστω, π.χ.)την $180^η$ και κατόπιν να κατασκευάσουμε το διάγραμμα όλων όσων απομένουν στις οποίες είναι πιθανό να παρατηρηθεί η εκτός-ουράς της κατανομής συμπεριφορά.

Η Μορφή ‘’Τυπικής Κανονικής Απόκλισης’’ των Υπολοίπων

Συνήθως υποθέτουμε ότι $\varepsilon \sim N\left(0, \sigma^2\right)$, έτσι ώστε $e_i / \sigma \sim N(0,1)$. Τώρα αν το μοντέλο είναι σωστό, το μέσο τετράγωνο υπολοίπων

$$s^2 = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{(n-p)} = \frac{\sum_{i=1}^{n} e_i^2}{(n-p)}$$

εκτιμά τη σ^2. (Αν αγνοήσουμε το σφάλμα στρογγυλοποίησης τότε $\bar{e} = \sum e_i / n = 0$). Η ποσότητα e_i/s συχνά ονομάζεται ***μορφή της τυπικής κανονικής απόκλισης*** (***unit normal deviate form)*** του υπολοίπου e_i. Τα e_i/s, i=1,2,…….n, μπορούν να εξεταστούν σε ένα συνολικό διάγραμμα για να δούμε αν η υπόθεση ε_i/σ – N(0,1) είναι λανθασμένη. Επειδή το 95% μιας N(0,1) κατανομής βρίσκεται μεταξύ των ορίων (-1.96, 1,96) μπορούμε να αναμένουμε (κατά προσέγγιση) ότι χονδρικά το 95% των ei/s είναι μεταξύ των ορίων (-2, 2).

Μερικές φορές είναι κατάλληλο να εξετάσουμε τα υπόλοιπα με αυτό τον εναλλακτικό τρόπο, για παράδειγμα να ελέγξουμε για *απομονωμένα σημεία (outliers*).

▪ *3.3: Διάγραμμα σε Χρονική Ακολουθία*

Το Σχήμα αυτό είναι ενδεικτικό ότι μια μακροχρόνια χρονική επίδραση (a long-term time effect) δεν επηρεάζει τα δεδομένα (ή αν τα επηρεάζει, η επίδραση έχει κάπως εξηγηθεί από μια μεταβλητή X η οποία επίσης υπόκειται σε μια χρονική επίδραση).

1) Η διασπορά δεν είναι σταθερή αλλά αυξάνεται με το χρόνο, που σημαίνει ότι μια ανάλυση σταθμισμένων ελαχίστων τετραγώνων θα έπρεπε να είχε χρησιμοποιηθεί.
2) Ένας γραμμικός ως προς το χρόνο όρος θα έπρεπε να είχε περιληφθεί στο μοντέλο.
3) Γραμμικοί και τετραγωνικοί ως προς το χρόνο όροι θα έπρεπε να είχαν περιληφθεί στο μοντέλο.

Φυσικά, μπορεί να εμφανιστούν συνδυασμοί και παρεκκλίσεις αυτών των ατελειών.

▪ *3.4: Διάγραμμα ως προς* $\widehat{Y_i}$

Εδώ τα διαγράμματα θα υποδείκνυαν:

1) *Μη σταθερή διασπορά* όπως υποτέθηκε ανάγκη για σταθμισμένα ελάχιστα τετράγωνα ή μετασχηματισμό των παρατηρήσεων Y_i πριν κάνουμε μια ανάλυση παλινδρόμησης.
2) *Σφάλμα στην ανάλυση*. Η παρέκκλιση (departure) από την εξίσωση προσαρμογής είναι συστηματική (αρνητικά υπόλοιπα αντιστοιχούν σε χαμηλές τιμές της Y ενώ θετικά υπόλοιπα σε υψηλές τιμές της Y). Το αποτέλεσμα μπορεί επίσης να οφείλεται στη λανθασμένη παράλειψη του όρου β_0 από το μοντέλο.
3) *Ανεπαρκές μοντέλο* – ανάγκη για επιπλέον όρους στο μοντέλο (π.χ. τετραγωνικοί όροι ή όροι σταυρωτών γινομένων) ή ανάγκη για ένα μετασχηματισμό των παρατηρήσεων Y_i πριν από την ανάλυση.

Το διάγραμμα των υπολοίπων $\varepsilon_i = Y_i - \widehat{Y_i}$ ως προς $\widehat{Y_i}$ και όχι ως προς τα Y_i, στο συνήθη γραμμικό μοντέλο το κατασκευάζουμε επειδή τα e και Y συνήθως είναι συσχετισμένα ενώ τα e και $\widehat{Y}$ δεν είναι. Ένας τρόπος για να το δούμε αυτό είναι να θεωρήσουμε τα διαγράμματα των e_i ως προς (i) τα Y_i και (ii) τα $\widehat{Y_i}$ και να βρούμε την κλίση μιας γραμμής ελαχίστων τετραγώνων διαμέσου των σημείων. Η κλίση αυτή θα είναι περίπτωση (i) $1-R^2$ και στην (ii) 0.

▪ *3.5: Διάγραμμα ως προς τις Ανεξάρτητες Μεταβλητές Xji, i=1,2,…….n*

Η μορφή αυτών των διαγραμμάτων είναι ίδια με εκείνη των διαγραμμάτων ως προς Y_i, με τη διαφορά ότι τώρα χρησιμοποιούμε (αντί των τιμών που αντιστοιχούν στα Y_i) τις τιμές

των αντίστοιχων X_{ji} και συγκεκριμένα τις X_{j1}, X_{j2},……..X_{jn}. Για μια φορά ακόμα η συνολική εικόνα μιας οριζόντιας ζώνης των υπολοίπων εκλαμβάνεται ως ικανοποιητική. Οι ανωμαλίες υποδεικνύουν:

1) Μη σταθερή διασπορά, ανάγκη για σταθμισμένα ελάχιστα τετράγωνα ή για ένα μετασχηματισμό των Y.
2) Σφάλμα στους υπολογισμούς, η γραμμική επίδραση των X_j δεν εξαλείφεται.
3) Ανάγκη για επιπλέον όρους για παράδειγμα, ένας τετραγωνικός όρος του X_j στο μοντέλο ή ένας μετασχηματισμός των Y.

Σε μικρά προβλήματα παλινδρόμησης τα οποία περιλαμβάνουν μόνο δυο ή τρεις ανεξάρτητες μεταβλητές X, είναι δυνατό να κατασκευάσουμε ένα διάγραμμα του χώρου των δυο ή τριών διαστάσεων στον οποίο βρίσκονται τα σημεία των δεδομένων. Σε μια τέτοια περίπτωση μπορούμε να κατασκευάσουμε ένα διάγραμμα των σημείων στα οποία λαμβάνονται οι παρατηρήσεις κα να γράψουμε τα υπόλοιπα δίπλα σ αυτά τα σημεία. Όπου αυτό είναι δυνατό, το διάγραμμα συχνά δίνει μια καλή οπτική κατανόηση της περίπτωσης. Όταν έχουμε περισσότερες από X – μεταβλητές είναι δυνατό να κατασκευάσουμε τέτοια διαγράμματα για υποσύνολα των μεταβλητών και αυτός ο τρόπος είναι μερικές φορές ο κατάλληλος.

▪ *3.6: Απομονωμένες Τιμές*

Απομονωμένη ή ακραία τιμή (outlier) μεταξύ των υπολοίπων είναι εκείνη που κατ' απόλυτη τιμή είναι αρκετά μεγαλύτερη από τις υπόλοιπες και ίσως βρίσκεται σε απόσταση τριών ή τεσσάρων τύπων αποκλίσεων από τη μέση τιμή των υπολοίπων. Η απομονωμένη τιμή αποτελεί ιδιομορφία και υποδεικνύει ένα σημείο των δεδομένων που δεν είναι καθόλου αντιπροσωπευτικό όπως τα άλλα δεδομένα και θα πρέπει να εξεταστεί ιδιαίτερα προσεκτικά για να δούμε αν η αιτία της ιδιομορφίας του μπορεί να προσδιοριστεί.

▪ *3.7: Σειριακή Συσχέτιση Υπολοίπων*

Στην παλινδρόμηση, συνήθως υποθέτουμε ότι τα σφάλματα παρατήρησης είναι κατά ζεύγη ασυσχέτιστα. Αν αυτή η υπόθεση είναι ουσιαστικά λανθασμένη, τότε αναμένουμε ότι τα διαγράμματα των υπολοίπων σε χρονική σειρά ή σε κάποια άλλη λογική σειρά που ορίζεται από τις πρακτικές περιστάσεις, θα μας βοηθήσουν να το ανακαλύψουμε. Υπάρχουν, βέβαια, διάφοροι τρόποι με τους οποίους τα σφάλματα μπορεί να σχετίζονται. Ένας συνηθισμένος τρόπος είναι να είναι ***σειριακά συσχετισμένα (serially correlated)***, δηλαδή, οι συσχετίσεις μεταξύ σφαλμάτων που έχουν s βήματα είναι πάντοτε ίδιες. Γι αυτή τη συσχέτιση θα χρησιμοποιήσουμε το συμβολισμό ρ_s, s=1,2,…….

Ειδικότερα, αν τα υπόλοιπα παρουσιάζουν *τοπική θετική σειριακή συσχέτιση (locale positive serial correlation*), τότε διαδοχικά υπόλοιπα σε χρονική ακολουθία τείνουν να είναι περισσότερο ανόμοια απ' ότι τα μη διαδοχικά. Η συσχέτιση μεταξύ υπολοίπων που απέχουν ένα (δυο ή τρία…) βήμα(τα) ονομάζεται ***σειριακή συσχέτιση υστέρησης*** με βήμα -1 (ή 2 ή 3,....) (lag-1 (or 2 or 3…..) serial correlation). Εμπειρικά η σειριακή συσχέτιση υστέρησης με

βήμα-1 μπορεί να εξεταστεί κατασκευάζοντας το διάγραμμα κάθε υπολοίπου εκτός του πρώτου έναντι του υπολοίπου που προηγείται.

Χρήση Σταθμισμένων Ελαχίστων Τετραγώνων για Σειριακά Συσχετισμένα Δεδομένα

Μια βασική μέθοδος ανάλυσης που μπορεί να χρησιμοποιηθεί αν υπάρχουν σειριακές συσχετίσεις στα υπόλοιπα, είναι τα ***σταθμισμένα ελάχιστα τετράγωνα***.

Δυο Έλεγχοι για Σειριακή Συσχέτιση

Δυο αρκετά γνωστοί τρόποι για τον έλεγχο υποδειγμάτων σειριακής συσχέτισης στα υπόλοιπα είναι ο ***έλεγχος ροών*** και ο ***έλεγχος των Durbin – Watson***.

▪ *3.8: Εξέταση Ροών στο Διάγραμμα Χρονικής Ακολουθίας των Υπολοίπων*

Όταν είναι γνωστή η χρονική ακολουθία ενός συνόλου υπολοίπων, μερικές φορές παρατηρούνται ομάδες με θετικά ή αρνητικά υπόλοιπα κατά ένα ασυνήθιστο τρόπο.

▪ *3.9: Έλεγχος των Durbin-Watson για ένα Συγκεκριμένο Τύπο Σειριακής Συσχέτισης*

Ένας δημοφιλής έλεγχος για τον εντοπισμό ενός συγκεκριμένου τύπου σειριακής συσχέτισης είναι ο ονομαζόμενος έλεγχος των ***Durbin-Watson***.

Συνήθως υποθέτουμε ότι τα σφάλματα e_u είναι ανεξάρτητες $N(0, \sigma^2)$ μεταβλητές έτσι ώστε όλες οι σειριακές συσχετίσεις είναι $\rho_s = 0$. Θα ελέγξουμε αυτή τη μηδενική υπόθεση H_0: όλα τα $\rho_s = 0$ με τον έλεγχο των Durbin-Watson έναντι της εναλλακτικής H_1: $\rho_s=\rho^s$.

Μια από τις υποθέσεις του μοντέλου $\underset{\square}{Y} = X\,\underset{\square}{\beta} + \underset{\square}{\varepsilon}$, όπου $\underset{\square}{\varepsilon} = \left(\varepsilon_1, \varepsilon_2, ..., \varepsilon_n\right)'$, είναι η ανεξαρτησία των τυχαίων μεταβλητών ε_i. Ένα είδος πιθανής εξάρτησης των μεταβλητών ε_i είναι της μορφής

$$\varepsilon_k = \rho\varepsilon_{k-1} + z_k \qquad (3.24)$$

όπου $z_k \square N(0, \sigma^2)$ και z_k ανεξάρτητη από τις ε_{k-1}, ε_{k-2}, ……. και τις z_{k-1}, z_{k-2} κλπ. Τότε ο συντελεστής συσχέτισης των σφαλμάτων ε_t και ε_{t+s} θα είναι $\rho_s = \rho^s$. Ένα στατιστικό για τον έλεγχο της υπόθεσης H_0:ρ=0 με εναλλακτική την H_1:$\rho_s=\rho^s$ είναι το στατιστικό Durbin – Watson που ορίζεται με τον παρακάτω τρόπο.

Έστω ε_i το σφάλμα της i-στής παρατήρησης μετά την προσαρμογή του γραμμικού μοντέλου. Υπολογίζουμε την ποσότητα

$$d = \frac{\sum_{k}^{n} \left(\varepsilon_k - \varepsilon_{k-1} \right)^2}{\sum_{k=1}^{n} \varepsilon_k^2} \qquad (3.25)$$

και βρίσκουμε από κατάλληλους πίνακες τις τιμές dL και du. Για τον έλεγχο της υπόθεσης H_0 : $\rho = 0$ διακρίνουμε τις περιπτώσεις:

- Εναλλακτική H_1 : $\rho > 0$

  $d < d_L$ → απορρίπτεται η H_0 σε ε.σ. α

  $d > du$ → δεν απορρίπτεται η H_0

  $d_L \leq d \leq d_u$ → δεν προκύπτει αξιόπιστο συμπέρασμα.

- Εναλλακτική H_1 : $\rho < 0$

  Όπως το 1 με (4 – d) αντί του d.

- Εναλλακτική H_1 : $\rho \neq 0$ (δίπλευρο τεστ)

  $d < d_L$ ή $4 – d < d_L$ → απορρίπτεται η H_0 σε ε.σ. 2α

  $d > d_u$ και $4 – d > d_u$ → δεν απορρίπτεται η H_0

Σε κάθε άλλη περίπτωση, δεν προκύπτει αξιόπιστο συμπέρασμα.

Η μεροληψία όπως και η ετεροσκεδαστικότητα είναι δυο περιπτώσεις όπου οι συνθήκες κανονικότητας δεν ισχύουν και φυσικά δεν είναι οι μοναδικές. Πολλοί ερευνητές πρότειναν μεθόδους αντιμετώπισης τέτοιων «προβληματικών» καταστάσεων, πολλές από τις οποίες υπάρχουν και στα στατιστικά πακέτα που κυκλοφορούν. Τέτοιες είναι η ***Αμφικλινής παλινδρόμηση (Ridge regression***), η ***παλινδρόμηση των κύριων συνιστωσών (Principal component regression***), των ***ιδιοτιμών (latent root regression***) και άλλες.

- *Κεφάλαιο 4: Μετασχηματισμοί Μεταβλητών σε περιπτώσεις απόκλισης από τις υποθέσεις*

▪ *4.1 Εισαγωγή*

Οι υποθέσεις που απαιτούμε να ικανοποιούνται, ώστε να προσαρμόσουμε με ικανοποιητική αξιοπιστία το μοντέλο παλινδρόμησης $Y = X\beta + \varepsilon$ είναι:

1. Συνθήκες κανονικότητας των σφαλμάτων, δηλαδή $\varepsilon \sim N(0, I\sigma^2)$.

2. Οι μεταβλητές Xi να είναι ποσοτικές (μετρήσιμες).

Πολλές φορές όμως στην πράξη *οι μεταβλητές Xi είναι ποιοτικές ή απλά κατηγοριοποιημένες,* οπότε δεν ισχύει η δεύτερη προϋπόθεση. Άλλες φορές τα *σφάλματα δεν ακολουθούν κανονική κατανομή* (ούτε βέβαια και η Y), ή ακολουθούν μεν *κανονική κατανομή αλλά με διαφορετικές διασπορές*. Τέλος μπορεί τα σφάλματα να *μην είναι ασυσχέτιστα μεταξύ τους*. Στις τελευταίες περιπτώσεις δεν ικανοποιείται η πρώτη προϋπόθεση.

Μία μέθοδος για την αντιμετώπιση τέτοιων καταστάσεων είναι η ***αναζήτηση μετασχηματισμών με τη βοήθεια των οποίων επιτυγχάνεται η άρση των αποκλίσεων***.

Στο σημείο αυτό θα περιγράψουμε μετασχηματισμούς για ειδικές περιπτώσεις.

▪ *4.2: Μετασχηματισμοί στο Γραμμικό Μοντέλο*

Οι βασικές ιδέες της ανάλυσης παλινδρόμησης μπορούν να χρησιμοποιηθούν αποτελεσματικά και σε μερικές περιπτώσεις όπου έχουμε *μη γραμμική σχέση* μεταξύ x και y.

Πολλές φορές κατά τη μελέτη της σχέσης μεταξύ δύο μεταβλητών x και y, διαπιστώνουμε ότι αν και υπάρχει εξάρτηση μεταξύ x και y εν τούτοις απέχει πολύ από το να είναι γραμμική. Οι στατιστικές μέθοδοι για τη μελέτη μη γραμμικών μοντέλων είναι πολύ πιο δύσκολες από τις αντίστοιχες γραμμικές. Σε μερικές όμως περιπτώσεις είναι δυνατό να γίνουν στα x ή/ και στα y *κατάλληλοι μετασχηματισμοί* ώστε η σχέση που προκύπτει να είναι περίπου γραμμική. Τότε κάνοντας χρήση των προηγούμενων και αντιστρέφοντας στη συνέχεια τις μετασχηματισμένες μεταβλητές μπορούμε να πάρουμε τα ζητούμενα συμπεράσματα για τις αρχικές ποσότητες.

Παρακάτω δίνουμε μερικά ***Μη Γραμμικά Μοντέλα*** και τους αντίστοιχους *μετασχηματισμούς Γραμμικοποίησης*. Επίσης δίνουμε την ορολογία των συγκεκριμένων μοντέλων όπως την συναντάμε στα περισσότερα στατιστικά πακέτα.

1. ***Πολλαπλασιαστικό μοντέλο (Multiplicative model)***

   Αρχικό μοντέλο : $y = a_0 \cdot a_1^x \cdot \varepsilon$

   Μετασχηματισμός : y΄ = lny, x΄ = x, ε΄ = lnε

   Μετασχηματισμένο μοντέλο : $y' = \beta_0 + \beta_1 \cdot x' + \varepsilon'$ όπου $\beta_0 = \ln a_o, \beta_1 = \ln \alpha_1$

2. ***Εκθετικό μοντέλο (exponential model)***

   Αρχικό μοντέλο : $y = \exp(\alpha_0 + \alpha_1 \cdot x)$

   Μετασχηματισμός : y΄ = lny, x΄ = x, έ = ε

   Μετασχηματισμένο μοντέλο : $y' = \beta_0 + \beta_1 \cdot x' + \varepsilon'$ όπου $\beta_0 = \alpha_0, \beta_1 = \alpha_1$

3. ***Αντίστροφο μοντέλο (Reciprocal model)***

   Αρχικό μοντέλο : $1/y = \alpha_0 + \alpha_1 \cdot x$

   Μετασχηματισμός : y΄ = 1/y, x΄ = x, ε΄ = ε

   Μετασχηματισμένο μοντέλο : $y' = \beta_0 + \beta_1 \cdot x' + \varepsilon'$ όπου $\beta_0 = \alpha_0, \beta_1 = \alpha_1$

▪ *4.3: Ετεροσκεδαστικότητα*

Μια από τις βασικές προϋποθέσεις για την ανάλυση της παλινδρόμησης είναι η σταθερότητα της διασποράς των σφαλμάτων. Σε πολλές περιπτώσεις όμως αυτή η προϋπόθεση δεν ισχύει. Αν για παράδειγμα η μεταβλητή Υ είναι μια τυχαία μεταβλητή με κατανομή Poisson ή Διωνυμική, η διασπορά των τιμών της Υ άρα και των υπολοίπων θα εξαρτάται από τη μέση τιμή και επομένως δεν θα είναι σταθερή. Λέμε, τότε, ότι τα δεδομένα έχουν ***Ετεροσκεδαστικότητα (Heteroscedasticity***) και ένας τρόπος να το ανακαλύψουμε είναι να εξετάσουμε τα υπόλοιπα μετά τη μορφή του μοντέλου. Για το σκοπό αυτό κάνουμε τη γραφική παράσταση $\left(\widehat{Y}_i, e_i / s\right)$, όπου Yi η προβλεπόμενη τιμή στην παρατήρηση i και $e_i / s = \left(Y_i - \widehat{Y}_i\right) / s$ το τυποποιημένο υπόλοιπο στην παρατήρηση αυτή.

Αν η γραφική παράσταση έχει τη μορφή παράλληλων ευθειών που περικλείουν όλα τα σημεία τότε δεν έχουμε ετεροσκεδαστικότητα. Αν όμως έχουμε κάποια άλλη μορφή, τότε υπάρχει πρόβλημα. Π.χ. μια μορφή που δείχνει μια αύξηση της διασποράς με την αύξηση του Υ, ή μια μορφή που δείχνει ότι υπάρχει συστηματικό λάθος (αρνητικά υπόλοιπα σε χαμηλές τιμές του Υ και θετικά στις υψηλές) ή ακόμα και μια μορφή που δείχνει ότι πιθανόν χρειάζονται και άλλοι όροι στο μοντέλο. Ανάλογη εικόνα δίνει και η γραφική παράσταση (X_i, e_i/s).

Ας υποθέσουμε ότι έχουμε την περίπτωση που δείχνει μια αύξηση της διασποράς με την αύξηση του Υ και μάλιστα ότι συμβαίνει

$$Var(\varepsilon_i) = k^2 x_i^2, k > 0 \qquad (3.13)$$

όπου x_i, η προβλέπουσα μεταβλητή και ε_i τα υπόλοιπα μετά την προσαρμογή του μοντέλου $Y = \beta_0 + \beta_1 X + \varepsilon$. Αν θεωρήσουμε το μοντέλο

$$\frac{Y}{X} = \beta_1 + \beta_0 \frac{1}{X} + \frac{\varepsilon}{X} \qquad (3.14)$$

Τότε προφανώς θα ισχύει $Var\left(\frac{\varepsilon_i}{x_i}\right)=k^2=$ σταθερά και επομένως θα ισχύει η *βασική προϋπόθεση της σταθερότητας της διασποράς*. Όμως το μοντέλο (3.14) γράφεται

$$W = a_0 + a_1 Z + e$$

όπου

$$W=Y/X,\ Z=1/X,\ \alpha_0=\beta_1,\ \alpha_1=\beta_0,\ e=\varepsilon/X \quad (3.15)$$

είναι δηλαδή ένα γραμμικό μοντέλο που ικανοποιεί τις συνθήκες κανονικότητας.

Εργαζόμενοι με το νέο μοντέλο εκτιμούμε τους συντελεστές α_i και από αυτούς τους β_i.

▪ *4.4: Πολυπλοκότερα Μοντέλα- Εισαγωγή*

Στη συνέχεια θα δοθούν παραδείγματα πολυπλοκότερων μοντέλων. Μερικά απ αυτά τα μοντέλα περιλαμβάνουν μετασχηματισμούς των μεταβλητών και τη χρήση των εικονικών (ή βωβών) μεταβλητών.

Ο γενικότερος τύπος γραμμικού μοντέλου ως προς τις μεταβλητές $X_1, X_2,\ldots, X_\kappa$ μπορεί να γραφεί με τη μορφή

$$Y = \beta_0 + Z_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_\rho Z_\rho + \varepsilon \quad (5.0.1)$$

Η μεταβλητή Z_0 είναι μια εικονική μεταβλητή που η τιμή της είναι πάντοτε ίση με τη μονάδα, δηλαδή $Z_0 = 1$ και για αυτό γενικά θα την παραλείπουμε.

- *4.5: Πολυωνυμικά Μοντέλα Διαφόρων Τάξεων ως προς Xj*

Μοντέλα Πρώτης – Τάξης

1.Αν στην εξίσωση (5.0.1) θέσουμε p = 1 και $Z_1 = X$, παίρνουμε το απλό μοντέλο πρώτης τάξης με μία προβλέπουσα μεταβλητή:

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (5.1.1)$$

2. Αν στην εξίσωση (5.0.1) θέσουμε p =κ και $Z_j = X_j$, παίρνουμε ένα μοντέλο πρώτης – τάξης με κ προβλέπουσες μεταβλητές:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon \qquad (5.1.2)$$

Μοντέλα Δεύτερης – Τάξης

1.Αν στην εξίσωση (5.0.1) θέσουμε p = 2, $Z_1 = X$, $Z_2 = X^2$, και $\beta_2 = \beta_{11}$, παίρνουμε ένα μοντέλο δεύτερης - τάξης με μια προβλέπουσα μεταβλητή:

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon \qquad (5.1.3)$$

2.Αν στην εξίσωση (5.0.1) θέσουμε p = 5, $Z_1 = X_1$, $Z_2 = X_2$, $Z_3 = X_1^2$, $Z_4 = X_2^2$, $Z_5 = X_1X_2$, $\beta_3 = \beta_{11}$, $\beta_4 = \beta_{22}$, $\beta_5 = \beta_{12}$, παίρνουμε ένα μοντέλο δεύτερης – τάξης με δύο προβλέπουσες μεταβλητές:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + ... + \beta_{12} X_1 X_2 + \varepsilon \qquad (5.1.4)$$

Μοντέλα Τρίτης – Τάξης

1.Αν στην εξίσωση (5.0.1) θέσουμε p = 3, $Z_1 = X$, $Z_2 = X^2$, $Z_3 = X^3$, $\beta_2 = \beta_{11}$ και $\beta_3 = \beta_{111}$, παίρνουμε ένα μοντέλο τρίτης – τάξης με μια προβλέπουσα μεταβλητή:

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \beta_{111} X^3 + \varepsilon \qquad (5.1.5)$$

2.Αν στην εξίσωση (5.0.1) θέσουμε p = 9 και γίνει η κατάλληλη ταυτοποίηση των β_i Z_i το μοντέλο (5.0.1) μπορεί να παριστά ένα μοντέλο τρίτης – τάξης με δυο προβλέπουσες μεταβλητές που δίνεται από την εξίσωση:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{12} X_1 X_2 + \beta_{22} X_2^2 + \beta_{111} X^3 + \beta_{112} X_1^2 X_2 + \beta_{122} X_1 X_2^2 + \beta_{222} X_2^3 + \varepsilon$$
(5.1.6)

Μετασχηματισμοί

Αν ένα μοντέλο δεύτερης – τάξης δεν είναι επαρκές, τότε μπορεί να είναι επαρκές ένα μοντέλο τρίτης – τάξης. Ωστόσο δεν θα πρέπει μηχανικά κάθε φορά να προσθέτουμε όρους μεγαλύτερης τάξης. Πολλές φορές είναι αποτελεσματικότερο να διερευνήσουμε τις επιδράσεις που προκύπτουν από άλλους μετασχηματισμούς των ανεξάρτητων μεταβλητών, ή από μετασχηματισμούς της εξαρτημένης μεταβλητής, ή και από τα δύο. Το ίδιο ισχύει επίσης όταν πρέπει να αποφασίσουμε για ένα μοντέλο πρώτης – τάξης έναντι ενός μοντέλου δεύτερης – τάξης. Για παράδειγμα, αν είναι κατάλληλη η προσαρμογή μιας ευθείας γραμμής της εξαρτημένης logY στη X, τότε μια τέτοια μορφή μοντέλου είναι προτιμότερη από μια τετραγωνική μορφή της Y στη X, υποθέτοντας ότι η συμπεριφορά των υπολοίπων καθιστά δυνατή την προσαρμογή και στις δύο περιπτώσεις.

4.6: Μοντέλα Με Μετασχηματισμούς Διαφορετικούς Από Τους Μετασχηματισμούς Ακέραιων Δυνάμεων

Μοντέλα που Προκύπτουν Μετασχηματίζοντας Μόνο τις Xj

Ο ΑΝΤΙΣΤΡΟΦΟΣ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΣ. Αν στην εξίσωση (5.0.1) πάρουμε p=2, $Z_1 = 1/X_1$, $Z_2 = 1/X_2$, τότε έχουμε το μοντέλο

$$Y = \beta_0 + \beta_1\left(\frac{1}{X_1}\right) + \beta_2\left(\frac{1}{X_2}\right) + \varepsilon \qquad (5.2.1)$$

Ο ΑΛΓΟΡΙΘΜΙΚΟΣ ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΣ. Παίρνοντας p=2, $Z_1=\ln X_1$, $Z_2=\ln X_2$, η εξίσωση (5.1.0) γίνεται

$$Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \varepsilon \qquad (5.2.2)$$

Ο ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΣ ΤΕΤΡΑΓΩΝΙΚΗΣ ΡΙΖΑΣ. Για παράδειγμα,

$$Y = \beta_0 + \beta_1 X_1^{1/2} + \beta_2 X_2^{1/2} + \varepsilon \qquad (5.2.3)$$

Είναι φανερό ότι υπάρχουν πολλοί τέτοιοι δυνατοί μετασχηματισμοί, και μπορούμε να θεωρήσουμε μοντέλα που περιλαμβάνουν λίγους ή πολλούς τέτοιους όρους. Βέβαια, διάφοροι και διαφορετικοί μεταξύ τους μετασχηματισμοί μπορούν να εμφανιστούν στο ίδιο μοντέλο. Συχνά η επιλογή του είδους του μετασχηματισμού, αν πρόκειται να γίνει οποιοσδήποτε μετασχηματισμός, είναι δύσκολη απόφαση. Η επιλογή θα πρέπει να γίνει στη βάση προηγούμενης γνώσης σχετικά με τις μεταβλητές που μελετώνται. Ο σκοπός που

κάνουμε τέτοιους μετασχηματισμούς είναι για να μπορούμε να χρησιμοποιήσουμε ένα μοντέλο παλινδρόμησης απλής μορφής ως προς τις μετασχηματισμένες μεταβλητές, παρά ένα πολυπλοκότερο μοντέλο των αρχικών μεταβλητών.

Περίληψη

Οι μετασχηματισμοί που αναφέρθηκαν σ αυτή την ενότητα είναι μερικοί μεταξύ πολλών μετασχηματισμών που χρησιμοποιήθηκαν πρόσφατα για το μετασχηματισμό πολύπλοκων μοντέλων σε γραμμικά. Όταν, όπως υποθέτουμε εδώ, οι προβλέπουσες μεταβλητές δεν υπόκεινται σε σφάλμα, τότε δεν υπάρχουν προβλήματα στον μετασχηματισμό τους. Ωστόσο, για μετασχηματισμούς στην εξαρτημένη μεταβλητή, Y, πρέπει να είμαστε ιδιαίτερα προσεκτικοί να ελέγξουμε ότι οι υποθέσεις των ελαχίστων τετραγώνων (ανεξάρτητα σφάλματα, $N(0,\sigma^2)$) εξακολουθούν να ισχύουν κάνοντας το μετασχηματισμό. Μπορούμε πολλές φορές να αποφύγουμε το μετασχηματισμό της εξαρτημένης μεταβλητής ψάχνοντας για κατάλληλους μετασχηματισμούς των X.

4.7: Οικογένειες Μετασχηματισμών

Μετασχηματισμοί της Εξαρτημένης Μεταβλητής

Μια χρήσιμη οικογένεια μετασχηματισμών της (απαραίτητα θετικής) εξαρτημένης μεταβλητής Y δίνεται από τους μετασχηματισμούς δύναμης

$$W = \begin{cases} \left(Y^{\lambda} - 1\right)/\lambda, \lambda \neq 0 \\ \ln Y, \lambda = 0.1 \end{cases} \qquad (5.3.1)$$

Η συνεχής αυτή οικογένεια μετασχηματισμών εξαρτάται από μια απλή παράμετρο λ. Τα δεδομένα μπορούν να χρησιμοποιηθούν για να εκτιμηθεί αυτή η παράμετρος, καθώς επίσης και το διάνυσμα των παραμέτρων β στο μοντέλο προσαρμογής, έστω το,

$$W = X\beta + \varepsilon \qquad (5.3.2)$$

όπου $W = (W_1, W_2, \ldots\ldots W_n)'$. Υπάρχουν δύο κύριοι τρόποι για την εκτίμηση του λ. Ο ένας τρόπος είναι να χρησιμοποιηθεί η ***μέθοδος μέγιστης πιθανοφάνειας*** κάτω από την υπόθεση ότι $\varepsilon \sim N(0, I\sigma^2)$ για την κατάλληλη επιλογή του λ.

Η Σπουδαιότητα του Ελέγχου των Υπολοίπων

Οι μετασχηματισμοί στην εξαρτημένη μεταβλητή επηρεάζουν την κατανομή των σφαλμάτων. Η υπόθεσή μας είναι ότι, μετά τον μετασχηματισμό, τα σφάλματα στη μετασχηματισμένη μεταβλητή θα είναι $N(0, I\sigma^2)$. Επομένως είναι σημαντικό να εξετάσουμε τα υπόλοιπα από το μοντέλο που τελικά προσαρμόστηκε, για να δούμε αν αυτές οι υποθέσεις φαίνεται να μην ισχύουν.

Γενικά, όταν κάνουμε ένα μετασχηματισμό είναι αδύνατο να συνδέσουμε τις παραμέτρους του μοντέλου που χρησιμοποιείται για τα μετασχηματισμένα δεδομένα με τις παραμέτρους στο μοντέλο που αρχικά προτάθηκε για τα μη-μετασχηματισμένα δεδομένα. Συνήθως δεν υπάρχει μαθηματική ισοδυναμία εκτός ίσως κατά μια προσέγγιση άποψη μέσω μιας ανάπτυξης σε σειρά Taylor.

4.8: Η Χρήση «Εικονικών» Μεταβλητών στην Πολλαπλή Παλινδρόμηση

Η Γενική Έννοια των ‘’Εικονικών’’ Μεταβλητών

Οι μεταβλητές που μελετήσαμε στις εξισώσεις παλινδρόμησης συνήθως μπορούν να πάρουν τιμές σε κάποιο συνεχές διάστημα. Ωστόσο, μερικές φορές πρέπει να εισάγουμε ένα παράγοντα που έχει ένα ή περισσότερες διακριτές στάθμες ή επίπεδα. Για παράδειγμα, μπορεί να έχουμε δεδομένα από τρεις μηχανές ή από δυο εργοστάσια ή από έξι χειριστές. Σε μια τέτοια περίπτωση δεν μπορούμε να θεωρήσουμε μια συνεχή κλίμακα για τη μεταβλητή ‘’μηχανή’’ ή τη μεταβλητή ‘’εργοστάσιο ή τη μεταβλητή ‘’χειριστής’’. Αυτό που πρέπει να κάνουμε είναι να καθορίσουμε γι αυτές τις μεταβλητές κάποιες στάθμες έτσι ώστε να λαμβάνεται υπόψη το γεγονός ότι οι διάφορες μηχανές ή εργοστάσια ή χειριστές μπορεί να έχουν ξεχωριστές ντετερμινιστικές (deterministic) επιδράσεις στην εξαρτημένη μεταβλητή. Μεταβλητές αυτού του τύπου συνήθως ονομάζονται εικονικές ή βωβές μεταβλητές (dummy variables). Οι μεταβλητές αυτές συνήθως (αλλά όχι πάντοτε) δε συσχετίζονται με οποιεσδήποτε φυσικές στάθμες που μπορεί να υπάρχουν στους παράγοντες.

Ένα παράδειγμα εικονικής μεταβλητής αποτελεί η μεταβλητή X_0 (της οποίας η τιμή είναι πάντοτε η μονάδα) που αντιστοιχεί στον όρο β_0 σε ένα μοντέλο παλινδρόμησης. Η παρουσία της μεταβλητής X_0 δεν είναι αναγκαία, αλλά ωστόσο πολλές φορές διευκολύνει το συμβολισμό. Άλλες εικονικές μεταβλητές κάνουν κάτι περισσότερο από το να διευκολύνουν απλά το συμβολισμό.

Χρονικές Τάσεις στα Δεδομένα

Σε πολλές πρακτικές περιπτώσεις, εμφανίζονται χρονικές τάσεις στις τιμές της εξαρτημένης μεταβλητής. Μερικές φορές η τάση αποτελεί και το μοναδικό παράγοντα που επηρεάζει την εξαρτημένη μεταβλητή και μερικές φορές εκτός των επιδράσεων που προκαλούν άλλες ανεξάρτητες μεταβλητές εμφανίζεται και η επίδραση της χρονικής τάσης. Γενικά μιλώντας, μπορούμε να λάβουμε υπόψη μας τη χρονική τάση χρησιμοποιώντας μία ή περισσότερες κατάλληλα ορισμένες εικονικές μεταβλητές. Προσθέτουμε τότε κατάλληλους όρους αυτών των εικονικών μεταβλητών στο υπόλοιπο μοντέλο που προκύπτει από άλλες

ανεξάρτητες μεταβλητές και στη συνέχεια προσαρμόζεται ολόκληρο το μοντέλο μ ένα τρόπο όμοιο μ εκείνον που διευκρινίσαμε στο παραπάνω παράδειγμα της μπλοκ-μεταβλητής. Αν και η συζήτησή μας εστιάζεται κυρίως στις χρονικές τάσεις, πρέπει να θυμόμαστε ότι άλλες παράμετροι που είναι κατάλληλες για το πρόβλημα που μελετάμε, πρέπει να εκτιμώνται ταυτόχρονα.

ΜΙΑ ΧΡΟΝΙΚΗ ΤΑΣΗ. Όταν στα δεδομένα υπάρχει μια χρονική τάση, τότε χρειάζεται να λάβουμε υπόψη μια απλή εικονική μεταβλητή.

ΔΥΟ ΧΡΟΝΙΚΕΣ ΤΑΣΕΙΣ. Όταν υπάρχουν δυο χρονικές τάσεις, θα πρέπει να προσδιορίσουμε μια εικονική μεταβλητή για κάθε μια τάση. Το πρόβλημα διαιρείται σε δύο κύρια επίπεδα πολυπλοκότητας: (1) όταν είναι γνωστό ποια δεδομένα βρίσκονται σε κάθε τάση, και (2) όταν αυτό δεν είναι γνωστό.

Στο επίπεδο (1) όταν είναι γνωστό ποια δεδομένα βρίσκονται σε κάθε τάση. Υποθέτουμε, για παράδειγμα, ότι και οι δύο χρονικές τάσεις είναι και οι δυο ευθείες γραμμές. Τότε σε αυτό το επίπεδο μπορούμε να συζητήσουμε δυο ακόμα επίπεδα: (1α) όταν η τετμημένη της τομής των δυο γραμμών μπορεί να θεωρηθεί ότι είναι μια συγκεκριμένη τιμή στην οποία υπάρχουν μια ή περισσότερες παρατηρήσεις και στο (1β) όταν η τετμημένη της τομής των δυο γραμμών είναι άγνωστη.

Κεφάλαιο 5: Επιλογή της «καλύτερης» εξίσωσης Παλινδρόμησης

5.1: Εισαγωγή

Στην παρούσα ενότητα θα ασχοληθούμε μόνο με τη χρήση ειδικών στατιστικών μεθόδων για την επιλογή μεταβλητών στην παλινδρόμηση. Υποθέτουμε ότι θέλουμε να κατασκευάσουμε μια εξίσωση γραμμικής παλινδρόμησης για μια συγκεκριμένη εξαρτημένη μεταβλητή Y σε σχέση με τις βασικές «ανεξάρτητες» ή προβλέπουσες μεταβλητές X_1, X_2, …….$X_κ$. Υποθέτουμε επιπλέον ότι οι Z_1, Z_2,……Z_r, είναι συναρτήσεις μίας ή περισσότερων μεταβλητών X, που αντιπροσωπεύουν το πλήρες σύνολο των μεταβλητών από τις οποίες πρόκειται να επιλεγεί η εξίσωση και ότι αυτό το σύνολο περιλαμβάνει οποιεσδήποτε συναρτήσεις, όπως τετράγωνα, σταυρωτά γινόμενα, λογάριθμους, αντίστροφες και δυνάμεις που πιστεύεται ότι είναι επιθυμητές και αναγκαίες. Στη διαδικασία επιλογής μιας εξίσωσης συνήθως εμπλέκονται δυο αντίθετα κριτήρια:

1. Για την κατασκευή μιας εξίσωσης χρήσιμης για σκοπούς πρόβλεψης θα πρέπει το μοντέλο μας να περιλαμβάνει όσο το δυνατό περισσότερες μεταβλητές Z έτσι ώστε οι προσαρμοσμένες τιμές (εκτιμήσεις) να είναι αξιόπιστες.

2. Επειδή η συγκέντρωση πληροφοριών για ένα μεγάλο αριθμό Z και η επακόλουθη επεξεργασία τους κοστίζουν, θα θέλαμε η εξίσωση να περιλαμβάνει όσο το δυνατό λιγότερες Z.

Ο συμβιβασμός μεταξύ των δυο αυτών ακραίων περιπτώσεων είναι αυτό που συνήθως ονομάζεται επιλογή της καλύτερης εξίσωσης παλινδρόμησης. Για να πετύχουμε την καλύτερη εξίσωση δεν υπάρχει μια και μοναδική στατιστική διαδικασία. Αν γνωρίζαμε το μέγεθος του σ^2 (την πραγματική τυχαία διασπορά των παρατηρήσεων) για οποιοδήποτε καλά ορισμένο πρόβλημα, η επιλογή καλύτερης εξίσωσης παλινδρόμησης θα ήταν ευκολότερη διαδικασία.

Κάθε εξίσωση παλινδρόμησης αξιολογείται σύμφωνα με κάποιο κριτήριο, τα τρία κριτήρια που θα είναι:

1. Η τιμή του $\boldsymbol{R^2}$ που επιτυγχάνεται από την προσαρμογή ελαχίστων τετραγώνων.
2. Η τιμή της $\boldsymbol{s^2}$, το μέσο τετράγωνο των υπολοίπων
3. Η στατιστική ***Cp***.

Η Χρήση του Μέσου Τετραγώνου Υπολοίπων s^2

Αν σ' ένα μεγάλο πρόβλημα γίνουν όλες οι παλινδρομήσεις, η αξιολόγηση του μέσου μεγέθους του μέσου τετραγώνου των υπολοίπων καθώς αυξάνεται ο αριθμός των μεταβλητών στην παλινδρόμηση, υποδεικνύει το βέλτιστο σημείο αποκοπής για τον αριθμό των μεταβλητών, στην παλινδρόμηση.

Η προσαρμογή εξισώσεων παλινδρόμησης που περιλαμβάνουν περισσότερες προβλέπουσες μεταβλητές από όσες χρειάζονται για να πάρουμε μια ικανοποιητική προσαρμογή στα δεδομένα ονομάζεται ***υπερπροσαρμογή (overfitting***). Προσθέτοντας όλο και περισσότερες προβλέπουσες μεταβλητές σε ένα ήδη υπερπροσαρμοσμένο μοντέλο, το μέσο τετράγωνο υπολοίπων θα τείνει να σταθεροποιηθεί και να προσεγγίζει την πραγματική τιμή της σ^2 καθώς ο αριθμός των μεταβλητών αυξάνεται, υπό την προϋπόθεση ότι όλες οι σημαντικές μεταβλητές έχουν περιληφθεί στο μοντέλο και ο αριθμός των παρατηρήσεων υπερβαίνει κατά πολύ τον αριθμό των μεταβλητών στην προσαρμοσμένη εξίσωση.

Η Χρήση της Στατιστικής C_p του Mallows

Μια εναλλακτική στατιστική η οποία έγινε αρκετά δημοφιλής τα τελευταία χρόνια είναι η στατιστική $\boldsymbol{C_p}$, η οποία αρχικά προστέθηκε από τον C. L. Mallows. Η στατιστική αυτή έχει τη μορφή

$$C_p = RSS_p/s^2 - (n - 2p) \qquad (6.1.1)$$

όπου RSS_p είναι το άθροισμα τετραγώνων των υπολοίπων από ένα μοντέλο που περιλαμβάνει p παραμέτρους, p είναι ο αριθμός των παραμέτρων στο μοντέλο συμπεριλαμβανομένου του β_0 και s^2 είναι το μέσο τετράγωνο υπολοίπων από την μεγαλύτερη προτεινόμενη εξίσωση που περιλαμβάνει όλες τις Z και θεωρείται να είναι μια αξιόπιστη αμερόληπτη εκτίμηση της διασποράς σφάλματος σ^2. Όπως έδειξε ο R. W. Kennard, η C_p συνδέεται στενά με το προσαρμοσμένο R^2, που συμβολίζεται με R^2_a, και επίσης συνδέεται με τη στατιστική R^2. Επειδή επίσης υποθέτουμε ότι $E(s^2) = \sigma^2$, κατά προσέγγιση ισχύει, ότι ο

λόγος RSS_p/s^2 έχει αναμενόμενη τιμή $(n-p)\sigma^2/\sigma^2=n-p$, οπότε κατά προσέγγιση επίσης ισχύει $E(C_p) = p$ για ένα επαρκές μοντέλο. Έπειτα ότι ένα διάγραμμα της C_p ως προς p θα αναδείξει τα «επαρκή» μοντέλα ως σημεία που βρίσκονται τελείως κοντά στη διαγώνιο γραμμή $C_p = p$. Εξισώσεις με αξιόλογη έλλειψη προσαρμογής, δηλαδή, μεροληπτικές εξισώσεις , θα δίνουν σημεία πάνω (συχνά πολύ πάνω) από τη γραμμή $C_p = p$. Λόγω της τυχαίας μεταβλητότητας, σημεία που αντιπροσωπεύουν εξισώσεις καλής προσαρμογής μπορεί επίσης να βρίσκονται κάτω από τη γραμμή $C_p = p$. Το πραγματικό ύψος C_p κάθε σημείου του διαγράμματος είναι επίσης σημαντικό επειδή (αυτό μπορεί να δειχτεί) είναι μια εκτίμηση του ολικού αθροίσματος τετραγώνων των αποκλίσεων (διασπορά συν σφάλμα μεροληψίας) του προσαρμοσμένου μοντέλου από το πραγματικό αλλά άγνωστο μοντέλο. Καθώς προστίθενται όροι στο μοντέλο για να ελαττωθεί το RSS_p, η C_p συνήθως αυξάνεται. Το «βέλτιστο» μοντέλο επιλέγεται αφού εξετάσουμε το διάγραμμα της C_p. Μπορούμε να ψάξουμε για μια παλινδρόμηση με μικρή τιμή της C_p περίπου ίση με p. Όταν η επιλογή δεν είναι ξεκάθαρη, τότε είναι θέμα προσωπικής κρίσης αν κάποιος προτιμήσει:

1. Μια μεροληπτική εξίσωση που δεν αντιπροσωπεύει τα πραγματικά δεδομένα τόσο καλά, επειδή έχει μεγαλύτερο RSS_p (έτσι ώστε $C_p > p$) αλλά έχει μια μικρότερη εκτίμηση C_p της συνολικής μεταβλητότητας (διασπορά σφάλματος συν σφάλμα μεροληψίας) από το πραγματικό αλλά άγνωστο μοντέλο, ή
2. Μια εξίσωση με περισσότερες παραμέτρους που προσαρμόζει καλύτερα τα πραγματικά δεδομένα (δηλαδή είναι $C_p = p$) αλλά έχει μια μεγαλύτερη συνολική μεταβλητότητα (διασπορά σφάλματος συν σφάλμα μεροληψίας) από το αληθινό αλλά άγνωστο μοντέλο.

Μ' άλλα λόγια, το μικρότερο μοντέλο έχει τη μικρότερη C_p τιμή, αλλά η τιμή C_p του μεγαλύτερου μοντέλου (το οποίο έχει μεγαλύτερη p τιμή) είναι πλησιέστερα στην τιμή του p.

Κατά συνέπεια, ένα μεγάλο πρόβλημα στην παλινδρόμηση είναι, το να αποφασίσουμε, ποιες από τις μεταβλητές X_1, X_2, …….Χκ για τις οποίες υπάρχει υπόνοια ότι επηρεάζουν την μεταβολή της Y, πρέπει να μπουν στο μοντέλο. Πως δηλαδή θα βρούμε το καλύτερο μοντέλο που προσαρμόζεται στα δεδομένα μας. Απ' όσα αναφέρθηκαν στα προηγούμενα προκύπτει ότι ένα μοντέλο είναι περισσότερο αξιόπιστο όσο περισσότερες μεταβλητές Xi χρησιμοποιεί για πρόβλεψη, αφού ο συντελεστής προσδιορισμού αυξάνει με το k ή ισοδύναμα η διασπορά των σφαλμάτων ελαττώνεται. Από την άλλη μεριά όμως με την αύξηση του k αυξάνεται η διασπορά των συντελεστών παλινδρόμησης αφού αυξάνει η πολυγγραμμικότητα ή ελαττώνεται η ανοχή και επίσης αυξάνει το κόστος της συλλογής των δεδομένων.

Δυστυχώς δεν υπάρχει διαδικασία που να οδηγεί σε ένα μοναδικά ορισμένο καλύτερο μοντέλο. Έτσι η τελική επιλογή του μοντέλου γίνεται με την ευθύνη του ερευνητή, ο οποίος εκτός από στατιστική πρέπει να γνωρίζει καλά και το συγκεκριμένο πρόβλημα.

Ένας από τους τρόπους για την επιλογή του καλύτερου μοντέλου είναι να θεωρήσουμε όλα τα δυνατά μοντέλα και να τα συγκρίνουμε μεταξύ τους. Ως κριτήρια επιλογής μπορούν να χρησιμοποιηθούν τα γνωστά στατιστικά R^2 ή s^2, είτε άλλα στατιστικά όπως το C_p του Mallows. Αναφερθήκαμε προηγουμένως σε ένα κριτήριο το οποίο τελευταία εφαρμόζεται πολύ συχνά. Το κριτήριο αυτό βοηθά στον καθορισμό ενός *''optimal'' υποσυνόλου ανεξάρτητων μεταβλητών* και ονομάζεται ***κριτήριο C_p*** και έχει προταθεί από τον *Mallows* (1973). Το κριτήριο αυτό έχει ως στόχο να ισορροπήσει το πλήθος του αριθμού των μεταβλητών που περιλαμβάνονται στο μοντέλο σε σχέση με το αποτέλεσμα που έχει να

παραλειφθεί κάποια σημαντική μεταβλητή και συνδέεται άμεσα με τον συντελεστή προσδιορισμού.

Σε κάθε περίπτωση τα στατιστικά αυτά πρέπει να υπολογιστούν για όλα τα δυνατά μοντέλα. Επομένως ένα πρώτο πρόβλημα είναι το πώς να διατάξουμε τα μοντέλα ώστε ο υπολογισμός αυτών των στατιστικών να γίνει ευκολότερος και ακριβέστερος. Είναι φανερό ότι, αν δύο από μοντέλα διαφέρουν μεταξύ τους ως προς μια μόνο μεταβλητή, τα αντίστοιχα στατιστικά θα σχετίζονται με ευκολότερες σχέσεις απ' ότι αν τα μοντέλα διέφεραν σε περισσότερες μεταβλητές. Ο Garside το 1965 πρότεινε για κάθε k μια διάταξη των μοντέλων, που να ικανοποιεί την προηγούμενη απαίτηση. Αν k=3 μπορούμε να διατάξουμε τα 8 δυνατά μοντέλα με τον παρακάτω τρόπο

( )-(1)-(12)-(2)-(23)-(123)-(13)-(3)

όπου (κλ.….) παριστάνει το μοντέλο $EY = \beta_0 + \beta_1 X_k + \beta_2 X_\lambda + \ldots$ και ( ) το μοντέλο $EY = \beta_0$. Όμοια αν k=4 τα 16 μοντέλα διατάσσονται:

( )-(1)-(12)-(2)-(23)-(123)-(13)-(3)-(34)-(134)-(1234)-(234)-(24)-(124)-(14)-(4).

Ο Furnival όμως το 1971 έδωσε μια μέθοδο υπολογισμού των στατιστικών ενός μοντέλου από τα προηγούμενα, όταν αυτά είναι διαταγμένα με δυαδικό τρόπο, π.χ. για k=3,

( )-(1)-(2)-(12)-(3)-(13)-(23)-(123)

Αφού με κάποιο τρόπο βρεθούν τα στατιστικά για όλα τα δυνατά μοντέλα, η επιλογή του καλύτερου θα γίνει με κάποιο από τα κριτήρια.

5.2: Επιλογή Μοντέλου με τη μέθοδο του Αποκλεισμού Μεταβλητών (Backward Elimination Procedure)

Στην στατιστική βιβλιογραφία υπάρχουν πολλές μέθοδοι για τον καθορισμό του καλύτερου υποσυνόλου από ένα σύνολο μεταβλητών που είναι υποψήφιες νε περιληφθούν σε ένα μοντέλο πολλαπλής παλινδρόμησης. Μια από τις μεθόδους αυτές που θα εξηγήσουμε αργότερα ξεκινά με μοντέλο που δεν έχει καμία μεταβλητή και στη συνέχεια προσθέτει κάθε φορά από μια μεταβλητή που έχει σημαντική συνεισφορά στο μοντέλο. Η μέθοδος αυτή ονομάζεται συνήθως ***μέθοδος προοδευτικής προσθήκης μεταβλητών (forward procedure)***. Η μέθοδος που θα συζητήσουμε εδώ χρησιμοποιεί την ακριβώς αντίθετη λογική: ξεκινά περιλαμβάνοντας όλες τις μεταβλητές στο μοντέλο και σε κάθε βήμα αποκλείει μια μεταβλητή που δεν έχει σημαντική συνεισφορά σε αυτό. Η μέθοδος αυτή ονομάζεται ***μέθοδος αποκλεισμού μεταβλητών (backward elimination procedure***). Η μέθοδος αποκλεισμού μεταβλητών αποκλείει σε κάθε βήμα την μεταβλητή εκείνη που έχει το μεγαλύτερο *παρατηρούμενο επίπεδο σημαντικότητας (P-VALUE*) με την προϋπόθεση ότι το παρατηρούμενο αυτό επίπεδο σημαντικότητας υπερβαίνει το επίπεδο σημαντικότητας α που έχουμε προκαθορίσει.

Θα πρέπει να επισημάνουμε εδώ ότι ενδεχομένως μια διαφορετική μέθοδος επιλογής μεταβλητών για την πολλαπλή παλινδρόμηση θα καταλήξει σε διαφορετικό σύνολο μεταβλητών που θα πρέπει να περιληφθούν στο μοντέλο. Εν γένει όμως υπάρχει όμως

υπάρχει αρκετή συμφωνία μεταξύ των μεταβλητών του τελικού μοντέλου και αυτών που εμφανίζονται ως υποψήφιες για αποκλεισμό στο πρώτο στάδιο της μεθόδου αποκλεισμού μεταβλητών που περιγράψαμε. Αυτό συμβαίνει γιατί ο αποκλεισμός μιας μεταβλητής μεταβάλλει τα παρατηρούμενα επίπεδα σημαντικότητας που αντιστοιχούν στις υπόλοιπες μεταβλητές. Εξάλλου από τη στιγμή που ο ερευνητής έχει πρόσβαση σε υπολογιστή είναι λογικό και φρόνιμο να αποκλείει μα μεταβλητή σε κάθε βήμα.

5.3: Η Πιθανότητα λάθους πρώτου είδους

(Probality of a type 1 error)

Σε κάθε βήμα στην ***μέθοδο αποκλεισμού μεταβλητών (backward elimination procedure)*** γίνεται ένας έλεγχος υποθέσεως σε κάποιο συγκεκριμένο και προκαθορισμένο επίπεδο σημαντικότητας α. Η τιμή του α υποτίθεται ότι εκπροσωπεί την πιθανότητα να κάνουμε λάθος πρώτου είδους (να απορρίψουμε δηλαδή την μηδενική υπόθεση ενώ η υπόθεση αυτή είναι σωστή). Παρ' όλα αυτά η τιμή του α είναι πράγματι η πιθανότητα να κάνουμε λάθος πρώτου είδους μόνο στο πρώτο βήμα και μόνον αν οι υποθέσεις για την κατανομή των ε_i ισχύουν.

5.4: Διάστημα εμπιστοσύνης για τον μέσο $\mu_{Y/X}$

Όπως και στην απλή γραμμική παλινδρόμηση το 100(1-α)% διάστημα εμπιστοσύνης του μέσου $\mu_{Y|X}$ δίνεται από τον τύπο $\widehat{\mu}_{Y|x} \pm t_{n-k-1,1-a/2} S_{\widehat{\mu}_{Y|x}}$

όπου $\widehat{\mu}_{Y|x}$ είναι μια σημειακή εκτίμηση του $\mu_{Y|X}$ και υπολογίζεται με αντικατάσταση των γνωστών τιμών των μεταβλητών των $X_1,\ldots, X_\kappa$ στην εκτιμήτρια ελάχιστων τετραγώνων της εξίσωσης παλινδρόμησης. Το τυπικό λάθος του $\mu_{Y|X}$ είναι δύσκολο να υπολογισθεί στην πολλαπλή παλινδρόμηση και για το λόγο αυτό συνήθως χρησιμοποιούνται οι τιμές που δίνει ο υπολογιστής.

5.5: Διάστημα πρόβλεψης για το Y

Όπως έχουμε δει μια σημαντική χρησιμότητα της θεωρίας παλινδρόμησης είναι η πρόβλεψη αγνώστων τιμών της εξαρτημένης μεταβλητής Y για δοθείσες τιμές των ανεξάρτητων μεταβλητών.

Το 100(1-α)% διάστημα πρόβλεψης για την τιμή Y με την χρησιμοποίηση της πολλαπλής παλινδρόμησης για συγκεκριμένες τιμές ανεξάρτητων μεταβλητών $X_1, X_2,\ldots X_k$ δίνεται από τον τύπο

$$\widehat{Y} \pm t_{n-k-1,1-a/2} S_{\widehat{Y}}$$

Επειδή και στην περίπτωση αυτή ο υπολογισμός του τυπικού λάθους $S_{\bar{Y}}$ είναι δύσκολος χρησιμοποιούνται στατιστικά πακέτα για τον προσδιορισμό του διαστήματος πρόβλεψης. Το στοιχείο που χρειάζεται για τον καθορισμό του είναι το μοντέλο που θα χρησιμοποιηθεί για να δώσει την συγκεκριμένη πρόβλεψη. Το μοντέλο αυτό δεν είναι βέβαια υποχρεωτικά αυτό το οποίο προκύπτει ως τελευταίο βήμα στη μέθοδο αποκλεισμού μεταβλητών.

5.6: Επιλογή Μοντέλου με τη μέθοδο της προοδευτικής προσθήκης Μεταβλητών (Forward Procedure)

Η ***μέθοδος της προοδευτικής προσθήκης μεταβλητών*** χρησιμοποιεί την ακριβώς αντίθετη λογική από αυτήν που αναπτύξαμε στην προηγούμενη ενότητα. Εδώ το μοντέλο αναπτύσσεται με την προσθήκη κάθε φορά μιας ανεξάρτητης μεταβλητής. Το πλεονέκτημα στη μέθοδο αυτή είναι ότι οι πίνακες που πρέπει να αντιστραφούν θα είναι εν γένει, μικρότεροι από αυτούς που χρησιμοποιούμε στην μέθοδο αποκλεισμού μεταβλητών. Η μέθοδος όμως αυτή παρουσιάζει μειονεκτήματα σε σχέση με την προηγούμενη από την ιδιαιτερότητα που παρατηρείται λόγω των μεγαλύτερων τιμών των συντελεστών συσχέτισης των εξαρτημένων μεταβλητών με τις ανεξάρτητες. Ένα άλλο μειονέκτημα είναι ότι υπάρχει η δυνατότητα μόνο προσθήκης ανεξάρτητων μεταβλητών οι οποίες από τη στιγμή που θα ενσωματωθούν στη διαδικασία, διατηρούνται έστω και αν η τιμή της στατιστικής συνάρτησης T του συντελεστή τους παλινδρόμησης σε κάποιο βήμα δεν υπερβαίνει το κρίσιμο επίπεδο.

Η μέθοδος επιλογής μοντέλου με την προσθήκη μεταβλητών ακολουθεί τα εξής βήματα:

1^{ο}: Από ένα σύνολο ανεξάρτητων μεταβλητών X_1, X_2,… που είναι υποψήφιες να περιληφθούν στο μοντέλο διαλέγουμε την μεταβλητή X_1 που έχει τον μεγαλύτερο συντελεστή συσχέτισης με την εξαρτημένη μεταβλητή Y.

2^{ο}: Υπολογίζουμε την τιμή της στατιστικής συνάρτησης T για τον έλεγχο της υπόθεσης H_0 : $\beta_1 = 0$ ότι δηλαδή το Y δεν σχετίζεται με το X_1 στο απλό γραμμικό μοντέλο. Εάν με βάση τα στοιχεία του δείγματος, η H_0 δεν απορριφθεί σε κάποιο συγκεκριμένο επίπεδο σημαντικότητας θα πρέπει να οδηγηθούμε στο συμπέρασμα ότι καμιά από τις ανεξάρτητες μεταβλητές δεν πρέπει να συμπεριληφθεί στο μοντέλο και επομένως σταματάμε την διαδικασία στο σημείο αυτό. Αν $|T|$ υπερβαίνει κάποια κρίσιμη τιμή $t_{n-2,1-a/2}$ προχωρούμε σε επιλογή μιας δεύτερης μεταβλητής.

3^{ο}: Για να επιλέξουμε τη δεύτερη μεταβλητή X_2 υπολογίζουμε τους *συντελεστές μερικής συσχέτισης (partial correlations)* του Y με κάθε μια από τις υπόλοιπες υποψήφιες ανεξάρτητες μεταβλητές διατηρώντας ταυτόχρονα την μεταβλητή X_1 σταθερά. Η X_2 θα είναι η μεταβλητή με τον μεγαλύτερο απόλυτο συντελεστή μερικής συσχέτισης . Στη συνέχεια, για να ελέγξουμε την υπόθεση:

$$H_0 : \beta_2 = 0$$

Στο μοντέλο γραμμικής παλινδρόμησης με δύο ανεξάρτητες μεταβλητές X_1 και X_2 υπολογίζουμε την τιμή της αντίστοιχης στατιστικής συνάρτησης T. Αν $|T| \geq t_{n-3,1-a/2}$ για κάποιο επίπεδο σημαντικότητας α θα πρέπει να σταματήσουμε τη

διαδικασία επιλογής και να θεωρήσουμε ότι το μοντέλο θα πρέπει να περιέχει μόνο την ανεξάρτητη μεταβλητή X_1. Αν η απόλυτη τιμή του Τ υπερβαίνει την κρίσιμη τιμή προχωρούμε στην επιλογή της τρίτης μεταβλητής.

$4^{ο}$: Προχωρούμε με τον ίδιο τρόπο υπολογίζοντας και πάλι διαδοχικά συντελεστές μερικής συσχέτισης μεγαλύτερης τάξης σε κάθε βήμα προσθέτοντας στο μοντέλο τη μεταβλητή με την μεγαλύτερη απόλυτη τιμή συντελεστή μερικής συσχέτισης με το Υ διατηρώντας όλες τις μεταβλητές που έχουν προστεθεί μέχρι εκείνη τη στιγμή στο μοντέλο σταθερές. Σε κάθε βήμα υπολογίζεται ο συντελεστής συσχέτισης του Υ με τις επιλεγείσες μεταβλητές όπως επίσης και η τιμή της στατιστικής συνάρτησης F για τον έλεγχο της συνολικής υπόθεσης ότι όλοι οι συντελεστές παλινδρόμησης είναι μηδέν. Με τον τρόπο αυτό είναι δυνατόν να παρακολουθούμε τη βελτίωση που παρατηρείται με τη χρησιμοποίηση των διαφορετικών υποσυνόλων ανεξάρτητων μεταβλητών στην εξήγηση της συνολικής διασποράς. Η διαδικασία θα πρέπει να σταματάει όταν η τιμή της στατιστικής συνάρτησης $|T|$ για την ανεξάρτητη μεταβλητή που επελέγη τελευταία δεν υπερβαίνει κάποιο προκαθορισμένο κρίσιμο επίπεδο. Το τελικό σύνολο ανεξάρτητων μεταβλητών που επιλέγεται για να χρησιμοποιηθεί στο μοντέλο είναι αυτό που αποτελείται από όλες τις μεταβλητές εκτός της τελευταίας που έχει τιμή της στατιστικής συνάρτησης $|T|$ μη στατιστικά σημαντική. Φυσικά αναμένεται ότι η τιμή της στατιστικής συνάρτησης F για τον έλεγχο της συνολικής υπόθεσης για μηδενικούς συντελεστές παλινδρόμησης για τις μεταβλητές που έχουν επιλεγεί δεν θα ξεπερνά το συγκεκριμένο κρίσιμο επίπεδο που έχει επιλεγεί.

5.7:Μέθοδος της Βηματικής Παλινδρόμησης (Stepwise Regression)

Η μέθοδος της ***βηματικής παλινδρόμησης (stepwise regression***) είναι μια άλλη μέθοδος επιλογής ενός ‘’καλού’’ υποσυνόλου ανεξάρτητων μεταβλητών. Η μέθοδος αυτή είναι παρόμοια με την ***μέθοδο της προοδευτικής προσθήκης μεταβλητών***. Η διαφορά των δύο μεθόδων έγκειται στο γεγονός ότι για κάθε διαδοχικό βήμα η υπόθεση

$$H_0 : \beta_j = 0$$

ελέγχεται για όλες τις δυνατές ανεξάρτητες μεταβλητές ώστε να αποκλείονται εκείνες για τις οποίες οι τιμές της στατιστικής $|T_j|$ είναι μικρότερες από ένα προκαθορισμένο κρίσιμο επίπεδο. Η επόμενη μεταβλητή προστίθεται στο υποσύνολο με την ίδια διαδικασία χρησιμοποίησης του *κριτηρίου του μεγίστου συντελεστή συσχέτισης* όπως στη μέθοδο της προοδευτικής προσθήκης μεταβλητών. Αυτή η βηματική επιλογή συνεχίζεται μέχρις ότου φθάσουμε σε ένα υποσύνολο για το οποίο καμιά από τις μεταβλητές που περιέχει το υποσύνολο αυτό δεν έχουν τιμή για τη στατιστική συνάρτηση $|T_j|$ μικρότερη από κάποια συγκεκριμένη κρίσιμη τιμή γης μεταβλητής t και δεν υπάρχουν άλλες μεταβλητές που θα πρέπει να αξιολογηθούν για να περιληφθούν στο μοντέλο. Όπως είναι φανερό η διαδικασία αυτή είναι πολύπλοκη, χρειάζεται πολλούς υπολογισμούς και για το λόγο αυτό γίνεται μόνο με στατιστικά πακέτα. Όλα σχεδόν τα στατιστικά πακέτα που κυκλοφορούν προβλέπουν τη δυνατότητα εφαρμογής της μεθόδου βηματικής παλινδρόμησης.

5.8: Σύγκριση της μεθόδου της βηματικής παλινδρόμησης με την μέθοδο αποκλεισμού μεταβλητών.

Τόσο η ***μέθοδος αποκλεισμού μεταβλητών*** όσο και η ***μέθοδος της βηματικής παλινδρόμησης*** είναι διαδικασίες που χρησιμοποιούνται στη Στατιστική ώστε να καθορισθεί ένα υποσύνολο ανεξάρτητων μεταβλητών που είναι χρήσιμο στην πρόβλεψη της εξαρτημένης μεταβλητής. Με δεδομένο ότι οι δύο αυτές μέθοδοι είναι διαφορετικές (όπως επίσης είναι διαφορετικές από την μέθοδο της προοδευτικής προσθήκης μεταβλητών που επίσης εξετάσαμε) είναι φυσικό ότι δεν μπορεί να περιμένει κανείς να δώσουν τα ίδια ακριβώς υποσύνολα μεταβλητών. Παρ’ όλα αυτά όμως , εν γένει τα δύο αυτά υποσύνολα δεν διαφέρουν σημαντικά.

▪ *5.9: Σταδιακή επιλογή μεταβλητών*

Με τις σημερινές δυνατότητες των υπολογιστών, η μέθοδος της επιλογής του <<καλύτερου μοντέλου>> μεταξύ όλων των 2k δυνατών μοντέλων μπορεί να εφαρμοστεί μέχρι και για $k \leq 12$. Μάλιστα με κατάλληλο προγραμματισμό ο χρόνος μπορεί να μειωθεί σημαντικά, όχι όμως και ο χώρος. Έτσι για μεγάλες τιμές του k είναι απαραίτητο αλλά και για τις μικρότερες βολικότερο, να περιοριστεί ο όγκος των υπολογισμών. Διάφορες μέθοδοι έχουν προταθεί, εμείς όμως θα αναπτύξουμε στα παρακάτω τρείς από αυτές, την **«προς τα εμπρός επιλογή»** (FORWARD SELECTION), την **«προς τα πίσω απαλειφή»** (BACKWARD ELIMINATION), και την **«βήμα-προς βήμα παλινδρόμηση»** (STERWISE REGRESSION).

Στην ***προς τα εμπρός επιλογή*** ξεκινάμε τη διαδικασία εύρεσης του καλύτερου μοντέλου από το μοντέλο θέσης $Y = \beta_0 + \varepsilon$. Προχωράμε, προσθέτοντας στη συνέχεια μια μεταβλητή, κάθε φορά εκείνη, που εξηγεί το μεγαλύτερο ποσοστό της ανεξήγητης διασποράς. Προσθέτουμε επομένως σε κάθε βήμα εκείνη τη μεταβλητή που η συμμετοχή της δίνει στο μοντέλο τον απόλυτα μεγαλύτερο μερικό συντελεστή συσχέτισης. Επειδή, όμως η προσθήκη της νέας μεταβλητής στο μοντέλο έχει νόημα μόνο εφόσον ο συντελεστής παλινδρόμησής της είναι σημαντικά διαφορετικός του μηδέν, γι' αυτό συγκρίνουμε σε κάθε βήμα το λόγο F της μεταβλητής που πρόκειται να μπει στο μοντέλο με κάποια κρίσιμη τιμή FIN. Αν ισχύει

$$F > FIN$$

η μεταβλητή προστίθενται στο μοντέλο. Η κρίσιμη αυτή τιμή μπορεί να είναι η $F_{1,n-1,a}$ όπου α το επίπεδο σημαντικότητας που μας ενδιαφέρει.

Το SPSS ως F_{IN} παίρνει την τιμή 3.84 (δηλαδή το $F_{1,N,0.05}$ για $N > 1000$), υπάρχει όμως δυνατότητα να ορίσουμε οποιαδήποτε άλλη τιμή ως F_{IN}. Ακόμη επειδή προκύπτει ότι αύξηση του τετραγώνου του μερικού συντελεστή συσχέτισης συνεπάγεται αύξηση του F, μπορούμε να συνοψίσουμε ως εξής: σε κάθε βήμα της ***«προς τα εμπρός επιλογής»***, υπολογίζονται οι λόγοι F για όλες τις μεταβλητές που δεν έχουν μπει ακόμα στο μοντέλο. Αν ο μεγαλύτερος από τους λόγους αυτούς είναι μεγαλύτερος και του F_{IN}, η αντίστοιχη μεταβλητή προστίθενται στο μοντέλο και συνεχίζουμε στο επόμενο βήμα. Αν όλοι οι λόγοι είναι μικρότεροι του F_{IN}, η διαδικασία σταματά και το τελευταίο μοντέλο θεωρείται ως το «καλύτερο».

Στην «***προς τα πίσω απαλοιφή***» ακολουθούμε την ακριβώς αντίστροφη πορεία. Ξεκινάμε από το πλήρες μοντέλο, απαλείφοντας σε κάθε βήμα μια μεταβλητή. Η μεταβλητή αυτή πρέπει να έχει το μικρότερο λόγο F (άρα και τον απόλυτα μικρότερο συντελεστή μερικής συσχέτισης) και να ικανοποιεί τη σχέση

$$F < F_{OUT}$$

Η κρίσιμη τιμή Fout όπως και η F_{IN} να είναι $F_{1,n-1,a}$ το SPSS ως Fout παίρνει την τιμή 2.71, που είναι η τιμή για $F_{1,N,0.10}$, $N > 1000$. Η διαδικασία σταματά όταν έχουν απαλειφθεί όλες οι μεταβλητές ή όταν όλες όσες έχουν μείνει στο μοντέλο έχουν λόγο $F \geq$ Fout.

Αξίζει να σημειωθεί ότι ακόμη κι αν θεωρήσουμε F_{IN} = Fout οι δυο μέθοδοι που αναφέρθηκαν δεν καταλήγουν πάντα στο ίδιο μοντέλο.

Η τρίτη μέθοδος «***βήμα προς βήμα» ή «βηματική***» παλινδρόμηση είναι μια βελτιωμένη παραλλαγή της «***προς τα εμπρός επιλογής***». Ξεκινάμε και εδώ από το απλό μοντέλο θέσης και σε κάθε βήμα προσθέτουμε μια μεταβλητή στο μοντέλο. Πριν όμως προχωρήσουμε στο επόμενο βήμα ελέγχουμε τις μεταβλητές που ήδη είναι στο μοντέλο με μέθοδο ανάλογη της «***προς τα πίσω απαλοιφής***». Συγκεκριμένα, σε κάθε βήμα, υπολογίζονται όλοι οι λόγοι F, που αντιστοιχούν στις μεταβλητές που δεν συμμετέχουν στο μοντέλο, και που θα προέκυπταν αν μια κάθε φορά απ' αυτές συμμετείχε επί πλέον στο μοντέλο.

Αν όλοι οι λόγοι είναι μικρότεροι μιας τιμής F_{IN} η διαδικασία σταματά, αλλιώς προσθέτουμε εκείνη τη μεταβλητή που έχει μεγαλύτερο F και ικανοποιεί την $F > F_{IN}$. Στη συνέχεια υπολογίζουμε τους λόγους F των υπολοίπων μεταβλητών που ήταν ήδη μέσα στο μοντέλο, όπως αυτοί διαμορφώνονται μετά την προσθήκη της νέας μεταβλητής. Αν για κάποιες απ' αυτές συμβεί ο λόγος F να ικανοποιεί τη σχέση $F <$ Fout, τότε οι μεταβλητές αυτές απαλείφονται από το μοντέλο και προχωράμε στο επόμενο βήμα. Είναι φανερό ότι για

να μην είναι ατέρμονη η διαδικασία θα πρέπει Fout < F_{IN}. Πάντως, ανάλογα με τον καθορισμό των τιμών F_{IN}, Fout η μέθοδος μπορεί να καταλήγει σε διαφορετικά μοντέλα.

Το μειονέκτημα των μεθόδων σταδιακής επιλογής είναι ότι δεν μπορούν να εκτιμήσουν το «αμέσως καλύτερο» μοντέλο, παρά μόνο το «καλύτερο». Έτσι δεν έχει ο ερευνητής τη δυνατότητα, αξιοποιώντας τις γνώσεις του ως προς τη φυσική σημασία των μεταβλητών, να επιλέξει μεταξύ «κοντινών» μοντέλων.

Όσον αφορά τις μεθόδους σταδιακής επιλογής μπορούμε να κάνουμε τις εξής παρατηρήσεις:

(1) Η σειρά με την οποία μπαίνουν οι μεταβλητές στο μοντέλο όταν εργαζόμαστε με ***FORWARD ή STERWISE*** μέθοδο δεν δείχνει τη σημαντικότητα των μεταβλητών. Το ανάλογο ισχύει και για την μέθοδο ***BACKWARD***.

(2) Μία μεταβλητή που μπήκε κάποτε στο μοντέλο, μπορεί αργότερα να απαλειφθεί κα να ξαναμπεί, χωρίς αυτό να σημαίνει τίποτε το ουσιαστικό για τη σημαντικότητα της μεταβλητής. Και οι τρεις μέθοδοι υπολογίζουν το λόγο F που εξαρτάται από το μερικό συντελεστή συσχέτισης, ο οποίος με τη σειρά του εξαρτάται από το πόσες και ποιες μεταβλητές είναι ήδη στο μοντέλο.

Στην επόμενη ενότητα θα συνοψίζουμε τις τρεις μεθόδους σταδιακής επιλογής μεταβλητών.

5.10: Η Διαδικασία της Προς τα Πίσω Απαλοιφής

Η διαδικασία της ***προς τα πίσω απαλοιφής*** είναι οικονομικότερη από τη διαδικασία «όλων των παλινδρομήσεων» υπό την έννοια ότι προσπαθεί να εξετάσει μόνο τις «καλύτερες» παλινδρομήσεις που περιλαμβάνουν ένα συγκεκριμένο αριθμό μεταβλητών. Τα βασικά βήματα σ' αυτή τη διαδικασία είναι τα εξής:

1. Υπολογίζεται μια εξίσωση παλινδρόμησης που περιλαμβάνει όλες τις μεταβλητές.
2. Υπολογίζεται η τιμή του μερικού F-ελέγχου για κάθε προβλέπουσα μεταβλητή η οποία θεωρείται ως να ήταν η τελευταία μεταβλητή που εισήχθη στην εξίσωση της παλινδρόμησης.
3. Η μικρότερη τιμή του μερικού F-ελέγχου, έστω F_L, συγκρίνεται με ένα προεπιλεγμένο επίπεδο σημαντικότητας, έστω F_0.
    a. Αν $F_L < F_0$, τότε αφαιρείται από το μοντέλο η μεταβλητή Z_L, στην οποία οφείλεται η τιμή F_L, από την παλινδρόμηση και ξανά υπολογίζεται η εξίσωση παλινδρόμησης με τις απομένουσες μεταβλητές: επαναλαμβάνεται το στάδιο (2).
    b. Αν $F_L > F_0$, τότε θεωρούμε (δεχόμαστε) την εξίσωση της παλινδρόμησης όπως υπολογίστηκε και σταματάμε.

Η διαδικασία της προς τα πίσω απαλοιφής είναι μια ικανοποιητική διαδικασία, ειδικότερα για στατιστικούς που θέλουν να βλέπουν όλες τις μεταβλητές στην εξίσωση ώστε «να μην χάσουν τίποτα». Είναι αρκετά οικονομικότερη σε ότι αφορά τον υπολογιστικό χρόνο και το ανθρώπινο δυναμικό που απαιτούνται για την εφαρμογή της, συγκριτικά με τη μέθοδο «όλων των δυνατών παλινδρομήσεων». Κανείς πρέπει να αναγνωρίσει ότι άπαξ και μια μεταβλητή απαλειφθεί κατά τη διαδικασία, παραμένει για πάντα εκτός παλινδρόμησης. Επομένως όλα τα εναλλακτικά μοντέλα που περιλαμβάνουν μεταβλητές που έχουν απαλειφθεί δεν είναι πλέον διαθέσιμα για εξέταση.

5.11: Η Διαδικασία της παλινδρόμησης κατά Βήματα

Η μέθοδος της ***προς τα πίσω απαλοιφής*** ξεκινάει με τη μεγαλύτερη εξίσωση παλινδρόμησης, χρησιμοποιώντας όλες τις μεταβλητές και ακολούθως ελαττώνει τον αριθμό των μεταβλητών στην εξίσωση έως ότου φτάσει σε μια απόφαση για την εξίσωση που θα χρησιμοποιηθεί. Η διαδικασία ***επιλογής μεταβλητών κατά βήματα*** αφορά στην προσπάθεια να πετύχουμε το ίδιο με το προηγούμενο αποτέλεσμα αλλά δουλεύοντας από την αντίθετη κατεύθυνση, δηλαδή, εισάγοντας μεταβλητές έως ότου φτάσουμε σε μια ικανοποιητική εξίσωση παλινδρόμησης. Η σειρά εισαγωγής των μεταβλητών καθορίζεται χρησιμοποιώντας το *συντελεστή μερικής συσχέτισης* ως ένα μέτρο της σπουδαιότητας των μεταβλητών που δεν έχουν ακόμα εισαχθεί στην εξίσωση. Η βασική διαδικασία είναι η εξής. Πρώτα επιλέγεται εκείνη η Z που συσχετίζεται περισσότερο με την Y (ας υποθέσουμε ότι η Z_l) και βρίσκουμε τη γραμμική εξίσωση παλινδρόμησης πρώτης – τάξης $Y = f(Z_l)$. Ελέγχεται αν η μεταβλητή αυτή είναι σημαντική. Αν δεν είναι, η διαδικασία σταματάει και υιοθετείται το $Y = \bar{Y}$ ως το καλύτερο μοντέλο, διαφορετικά ψάχνουμε για τη δεύτερη προβλέπουσα μεταβλητή που θα εισαχθεί στην παλινδρόμηση. Εξετάζονται οι συντελεστές μερικής συσχέτισης για όλες τις

προβλέπουσες μεταβλητές που σ' αυτό το στάδιο δεν περιλαμβάνονται στην εξίσωση παλινδρόμησης, δηλαδή της Z_j, $j \neq 1$ με την Y, δηλαδή, οι Y και Z_j είναι οι δύο προσαρμοσμένες για τη σχέση τους της ευθείας γραμμής με τη Z_j, και η συσχέτιση μεταξύ αυτών των προσαρμοσμένων μεταβλητών υπολογίζεται για όλα τα $j \neq 1$. Από μαθηματικής πλευράς αυτοί είναι ισοδύναμοι με το να βρούμε τις συσχετίσεις μεταξύ (1) των υπολοίπων από την παλινδρόμηση $Y = f(Z_l)$ και (2) των υπολοίπων από καθεμία από τις j παλινδρομήσεις $Z_j = f_j (Z_l)$ (τις οποίες πράγματι δεν έχουμε εκτελέσει). Ακολούθως επιλέγεται, η Z_j που έχει το μεγαλύτερο συντελεστή μερικής συσχέτισης με την Y, (ας υποθέσουμε ότι είναι η Z_2) και προσαρμόζεται μια δεύτερη εξίσωση παλινδρόμησης $Y = f(Z_1, Z_2)$. Η συνολική παλινδρόμηση ελέγχεται για τη σημαντικότητά της, σημειώνεται η βελτίωση στην τιμή του R^2, εξετάζονται οι μερικές F-τιμές και για τις δυο μεταβλητές που είναι τώρα στην εξίσωση (όχι δηλαδή μόνο γι' αυτή που μόλις μπήκε στην εξίσωση). Η μικρότερη από τις δύο μερικές F-τιμές συγκρίνεται με ένα κατάλληλο ποσοστιαίο σημείο F και η αντίστοιχη προβλέπουσα μεταβλητή παραμένει στην εξίσωση ή απορρίπτεται σύμφωνα με το αν ο έλεγχος είναι σημαντικός ή όχι. Η διαδικασία αυτή, δηλαδή ο έλεγχος για τη «λιγότερο χρήσιμη προβλέπουσα μεταβλητή που μόλις εισήχθη στην εξίσωση» γίνεται σε κάθε στάδιο της διαδικασίας κατά βήματα. Μια μεταβλητή που μπορεί να ήταν η καλύτερη για να μπει στην εξίσωση σε ένα προηγούμενο στάδιο, μπορεί σ' ένα επόμενο στάδιο να είναι περιττή λόγω της σχέσης της με άλλες μεταβλητές που είναι τώρα στην παλινδρόμηση. Για να ελεγχθεί αυτό, υπολογίζεται το μερικό F κριτήριο για κάθε μεταβλητή στην παλινδρόμηση σε κάθε στάδιο των υπολογισμών και η μικρότερη από αυτές τις μερικές F-τιμές (η οποία μπορεί να σχετίζεται με την τελευταία μεταβλητή που μπήκε στην παλινδρόμηση ή με μια προηγούμενη) συγκρίνεται τότε με ένα προεπιλεγμένο ποσοστιαίο σημείο, της κατάλληλης F-κατανομής. Με τον τρόπο αυτό κρίνεται η συνεισφορά, σ' εκείνη τη φάση, της λιγότερο αξιόλογης μεταβλητής στην παλινδρόμηση, θεωρώντας την ως να ήταν η τελευταία μεταβλητή που εισήχθη στην παλινδρόμηση, ανεξάρτητα του πραγματικού σημείου εισόδου στο μοντέλο. Αν η μεταβλητή που ελέγχεται δίνει μια μη σημαντική συνεισφορά, τότε αφαιρείται από το μοντέλο κα η κατάλληλη προσαρμοσμένη εξίσωση παλινδρόμησης υπολογίζεται τότε για όλες τις υπόλοιπες μεταβλητές που παραμένουν ακόμα στο μοντέλο. Τότε ελέγχεται η καλύτερη από τις μεταβλητές που δεν είναι ακόμα στο μοντέλο (δηλαδή εκείνη της οποίας η μερική συσχέτιση με την Y, δοθέντων των προβλεπουσών μεταβλητών που είναι ήδη στην εξίσωση, είναι μεγαλύτερη) για να δούμε αν ξεπερνά τον μερικό F-έλεγχο εισόδου. Αν το ξεπερνά, η μεταβλητή εισάγεται στην παλινδρόμηση και επιστρέφουμε στον έλεγχο όλων των μερικών F για τις μεταβλητές που είναι στην παλινδρόμηση. Αν δεν το ξεπερνά, τότε επιχειρείται μια επιπλέον απαλοιφή μεταβλητής. Τελικά (εκτός κι αν οι α-τιμές εισόδου και εξόδου δεν έχουν επιλεγεί σωστά για να δώσουν ένα κυκλικό αποτέλεσμα), όταν καμία μεταβλητή στην τρέχουσα εξίσωση δεν μπορεί να επαλειφθεί και η επόμενη καλύτερη υποψήφια μεταβλητή δεν μπορεί να πάρει τη θέση της στην εξίσωση, η διαδικασία σταματά. Καθώς εισέρχεται κάθε μεταβλητή στην παλινδρόμηση, συνήθως καταγράφεται και εμφανίζεται στα αποτελέσματα του ηλεκτρονικού υπολογιστή η επίδρασή της στην τιμή του R^2.

Είναι κοινός τόπος ότι η μέθοδος αυτή είναι καλύτερη από τις μεθόδους επιλογής μεταβλητών που συζητήσαμε και γενικά συνιστάται η χρήση τους. Είναι οικονομικότερη σε ότι αφορά τη χρήση του ηλεκτρονικού υπολογιστή από τις άλλες μεθόδους που αναφέραμε και αποφεύγει να χρησιμοποιεί περισσότερα Χ από όσα είναι απαραίτητα για τη βελτίωση της εξίσωσης σε κάθε φάση. Ωστόσο, πολύ εύκολα μπορεί να γίνει κατάχρηση της διαδικασίας παλινδρόμησης κατά βήματα από ένα «ανώριμο» στατιστικό. Έτσι κι εδώ

απαιτείται μια λογική αιτιολόγηση στην αρχική επιλογή των μεταβλητών και στην κριτική εξέταση του μοντέλου μέσω της εξέτασης των υπολοίπων.

5.12 Η Διαδικασία Παλινδρόμησης κατά Στάδια

Η μέθοδος αυτή δεν δίνει πραγματική λύση ελαχίστων τετραγώνων για τις μεταβλητές που περιλαμβάνονται στην τελική εξίσωση. Η βασική ιδέα είναι η εξής. Αφότου προσαρμοστεί μια εξίσωση παλινδρόμησης για τη μεταβλητή Χ που είναι η πιο συσχετισμένη με την Υ, βρίσκουμε τα υπόλοιπα $Y_i - \widehat{Y}_i$. Τα υπόλοιπα αυτά θεωρούνται τώρα ως τιμές μιας εξαρτημένης μεταβλητής και παλινδρομούνται ξανά ως προς μια Χ (από εκείνες που απομένουν) η οποία σχετίζεται περισσότερο με αυτή τη νέα εξαρτημένη μεταβλητή. Η διαδικασία συνεχίζεται μέχρι το στάδιο που επιθυμούμε. Οι νέες Χ μεταβλητές δεν προσαρμόζονται για τις προηγούμενες Χ μεταβλητές. Επειδή σε κάθε στάδιο έχουμε

Εξαρτημένη μεταβλητή = Προσαρμοσμένη Εξαρτημένη μεταβλητή + (Εξαρτημένη μεταβλητή – Προσαρμοσμένη Εξαρτημένη μεταβλητή)

οι εξισώσεις παλινδρόμησης μπορούν να αντικαθίστανται οπισθοδρομικά (δηλαδή προς τα πίσω) στάδιο – στάδιο έως ότου πετύχουμε την τελική κατά στάδια εξίσωση, η οποία δεν είναι λύση ελαχίστων τετραγώνων για τις μεταβλητές που περιλαμβάνονται.

Ενώ η διαδικασία αυτή δίνει πάντοτε μικρότερη ακρίβεια, δηλαδή, μεγαλύτερο μέσο τετράγωνο υπολοίπων, απ' ότι η διαδικασία των ελαχίστων τετραγώνων, έχει το εξής πλεονέκτημα. Μας δίνει τη δυνατότητα να επιλέξουμε μια πρώτη μεταβλητή για λόγους διαφορετικούς από εκείνους της συσχέτισής της με την Υ. Για παράδειγμα, δοθέντος ενός συνόλου μεταβλητών Χ με μεγάλη συσχέτιση μεταξύ τους, οι διαδικασίες επιλογής θα επιλέξουν πρώτα εκείνη τη Χ που έχει τη μεγαλύτερη συσχέτιση με την Υ και έστω ότι είναι η X_i. Στο επόμενο στάδιο, όλες οι άλλες Χ μεταβλητές που δεν είναι στην παλινδρόμηση προσαρμόζονται ανάλογα με τη συσχέτισή τους με τη X_1. Επομένως, αν η X_2 έχει μεγάλη συσχέτιση με τη X_1, τότε αυτή μπορεί να απορριφθεί ως πιθανή μεταβλητή. Ωστόσο η X_2 μπορεί να είναι ακριβώς η μεταβλητή που ο πειραματιστής θέλει να χρησιμοποιεί, για παράδειγμα, η X_2 μπορεί να είναι άμεσα κάτω από τον έλεγχο του πειραματιστή ενώ η X_1 μπορεί να μην είναι. Για παράδειγμα, σε μια περίπτωση μάρκετινγκ, η X_2 μπορεί να αντιστοιχεί στη δαπάνη που διαθέτει κάποιος για διαφήμιση, ενώ η X_1 μπορεί να είναι οι ανταγωνιστικές δαπάνες για διαφήμιση. Επομένως, ο πειραματιστής θα μπορούσε να χρησιμοποιεί ένα μοντέλο παλινδρόμησης $Y = f(X_2)$ και κατόπιν να συνεχίσει το πρόβλημα χρησιμοποιώντας τα υπόλοιπα (από αυτή την προσαρμογή) ως την εξαρτημένη μεταβλητή σε περαιτέρω παλινδρομήσεις.

Επιπλέον, υπάρχουν περιπτώσεις όπου μια τάση στα δεδομένα θα έπρεπε να εξαλειφθεί πριν προσπαθήσουμε να γράψουμε μια εξίσωση πρόβλεψης. Οι οικονομολόγοι συχνά διορθώνουν τα δεδομένα για τάσεις ή εποχικότητα και κατόπιν προχωρούν στην ανάλυση των αποκλίσεων που προκύπτουν με διαδικασίες ελαχίστων τετραγώνων.

- *Κεφάλαιο 6: Γραμμικό Μοντέλο και Στατιστικά Πακέτα*

6.1: Εισαγωγή

Η ***ανάλυση Πολλαπλής Παλινδρόμησης*** με πλήθος ανεξάρτητων μεταβλητών x>2 μπορεί εύκολα να γίνει σε οποιοδήποτε υπολογιστικό σύστημα αφού σήμερα είναι σχεδόν όλα εφοδιασμένα με κάποιο ή κάποια στατιστικά πακέτα (*Statgraphics, SPSS, MINITAB, BDMP,SAS* κλπ.) που διαθέτουν έτοιμα προγράμματα ανάλυσης του γενικού Γραμμικού Μοντέλου.

Το βασικό μενού που δίνει την δυνατότητα να επιλέξεις από τις x, (x>2), ανεξάρτητες μεταβλητές αυτές που πράγματι είναι απαραίτητες για την ερμηνεία της Y είναι η ***κλιμακωτή ή βήμα προς βήμα (Stepwise) παλινδρόμηση***. Τις εξής τρείς μεθόδους, χρησιμοποιεί το παραπάνω μενού για την επιλογή του ‘’καλύτερου μοντέλου’’ μεταξύ όλων των 2^x δυνατών μοντέλων όπου x το πλήθος των ανεξάρτητων μεταβλητών που είναι ‘’υποψήφιες’’ για να ερμηνεύσουν την εξαρτημένη μεταβλητή Y:

1. ***Την ‘’προς τα εμπρός επιλογή’’ (forward selection)***
2. ***Την ‘’προς τα πίσω απάλειψη’’ (backward elimination) και***
3. ***Την ‘’κλιμακωτή (βήμα προς βήμα) παλινδρόμηση που είναι συνδυασμός των 1 και 2.***

Στην ***προς τα εμπρός επιλογή*** ξεκινάμε τη διαδικασία εύρεσης του καλύτερου μοντέλου από το μοντέλο θέσης $Y= \beta_0 + \varepsilon$. Προχωρούμε, προσθέτοντας στη συνέχεια μια μεταβλητή, κάθε φορά εκείνη που εξηγεί το μεγαλύτερο ποσοστό της ανεξήγητης διασποράς. Προσθέτουμε επομένως σε κάθε βήμα εκείνη τη μεταβλητή που η συμμετοχή της στο μοντέλο δίνει τον απόλυτα μεγαλύτερο μερικό συντελεστή συσχέτισης. Επειδή όμως η προσθήκη της νέας μεταβλητής στο μοντέλο έχει νόημα μόνο εφόσον ο *συντελεστής παλινδρόμησης* της είναι σημαντικά διαφορετικός από το μηδέν, γι’ αυτό συγκρίνουμε σε κάθε βήμα το λόγο F της μεταβλητής που πρόκειται να μπει στο μοντέλο με κάποια κρίσιμη τιμή F_{IN}. Αν ισχύει

$$F> F_{IN}$$

η μεταβλητή προστίθεται στο μοντέλο. Η κρίσιμη αυτή τιμή μπορεί να είναι $F_{1,n-1,a,}$ όπου *α το επίπεδο σημαντικότητας* που μας ενδιαφέρει. Υπάρχει όμως η δυνατότητα να ορίσουμε οποιαδήποτε άλλη τιμή ως $F_{IN.}$ Ακόμη, επειδή προκύπτει ότι αύξηση του τετραγώνου του μερικού συντελεστή συσχέτισης συνεπάγεται αύξηση του F, μπορούμε να συνοψίσουμε ως εξής: Σε κάθε βήμα της προς τα εμπρός επιλογής, υπολογίζονται οι λόγοι F για όλες της μεταβλητές που δεν έχουν μπει ακόμη στο μοντέλο. Αν ο μεγαλύτερος από τους λόγους αυτούς είναι μεγαλύτερος και του $F_{IN,}$ η διαδικασία σταματά και το τελευταίο μοντέλο θεωρείται ως το ‘’καλύτερο’’.

Στην ***προς τα πίσω απαλοιφή*** ακολουθούμε ακριβώς την αντίστροφη πορεία. Ξεκινούμε από το πλήρες μοντέλο, απαλείφοντας σε κάθε βήμα μια μεταβλητή. Η μεταβλητή αυτή πρέπει να έχει το μικρότερο λόγο F (άρα και τον απόλυτα μικρότερο συντελεστή μερικής συσχέτισης) και να ικανοποιεί τη σχέση

$$F > F_{OUT}$$

Η κρίσιμη τιμή F_{OUT}, όπως και η $F_{IN,}$ να είναι η $F_{1,n-1,a.}$ Η διαδικασία σταματά όταν έχουν απαλειφθεί όλες οι μεταβλητές ή όταν όλες όσες έχουν μείνει στο μοντέλο έχουν λόγο $F \geq F_{OUT}$. Αξίζει να σημειωθεί ότι ακόμη κι να θεωρήσουμε $F_{IN}=F_{OUT}$ οι δύο μέθοδοι που αναφέρθηκαν δεν καταλήγουν πάντα στο ίδιο μοντέλο.

Η τρίτη μέθοδος της ***''βήμα προς βήμα'' παλινδρόμησης*** είναι βελτίωση των δύο προηγούμενων. Ξεκινούμε και εδώ από το απλό μοντέλο θέσης και σε κάθε βήμα προσθέτουμε μια μεταβλητή στο μοντέλο. Πριν όμως προχωρήσουμε στο επόμενο βήμα ελέγχουμε τις μεταβλητές που ήδη είναι στο μοντέλο με μέθοδο ανάλογης της ***''προς τα πίσω απαλοιφής''***. Συγκεκριμένα, σε κάθε βήμα, υπολογίζονται πρώτα όλοι οι λόγοι F, που αντιστοιχούν στις μεταβλητές που δεν συμμετέχουν στο μοντέλο, και που θα προέκυπταν αν μια κάθε φορά απ' αυτές συμμετείχε επιπλέον στο μοντέλο.

Αν όλοι οι λόγοι είναι μικρότεροι μιας τιμής F_{IN} η διαδικασία σταματά, αλλιώς προσθέτουμε εκείνη τη μεταβλητή που έχει το μεγαλύτερο F, και ικανοποιεί την $F > F_{IN.}$ Στη συνέχεια υπολογίζουμε τους λόγους F των υπόλοιπων μεταβλητών που ήταν ήδη μέσα στο μοντέλο όπως αυτοί διαμορφώνονται μετά την προσθήκη της νέας μεταβλητής, Αν για κάποιες από αυτές συμβεί ο λόγος F να ικανοποιεί τη σχέση $F < F_{OUT,}$ τότε οι μεταβλητές αυτές απαλείφονται από το μοντέλο και προχωράμε στο επόμενο βήμα. Είναι φανερό ότι για να μην είναι ατέρμονη η διαδικασία θα πρέπει $F_{OUT} < F_{IN.}$ Πάντως, *ανάλογα με τον καθορισμό των τιμών $F_{IN,}$ F_{OUT} η μέθοδος μπορεί να καταλήγει σε διαφορετικά μοντέλα.*

Τα ελάχιστα στοιχεία τα οποία, ένα οποιοδήποτε στατιστικό πακέτο δίνει στο χρήστη είναι:

2. Οι εκτιμήσεις ελάχιστων τετραγώνων $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{\rho-1}$ (coefficients).

3. Η αμερόληπτη εκτιμήτρια S^2 της διασποράς σ^2 ή ισοδύναμα η εκτιμήτρια s της τυπικής απόκλισης σ (standard deviation (error) of regression line).

4. Οι εκτιμήσεις $s\left(\hat{\beta}_i^i\right)$ των τυπικών αποκλίσεων των εκτιμητριών ελαχίστων τετραγώνων $\hat{\beta}_i$ (standard deviation (error) of coefficients)

5. Τα πηλίκα t (t ratios ή t values)

$$\frac{\hat{\beta}_j}{s\left(\hat{\beta}_j\right)}$$

που χρησιμοποιούνται για τον έλεγχο της $H_{0:}\, \beta_j = 0$ κατά της $H_{1:}\, \beta_{j \neq 0,}$ j=0,1,2,…, ρ-1.

6. Το ***συντελεστή προσδιορισμού R^2 (coefficient of Determination, R – squared, multiple R- squared)*** και τον ***προσαρτημένο συντελεστή προσδιορισμού R^2 (Adj.***)

7. Τον πίνακα ανάλυσης διασποράς (Analysis of variance, Anova Table).

8. Την εκτίμηση του πίνακα διακυμάνσεων συνδιακυμάνσεων $\hat{\beta}, s^2\left(\hat{\beta}\right)$

9. Τον πίνακα συσχέτισης των ανεξάρτητων μεταβλητών.

10. Τις προσαρμοσμένες τιμές $\hat{y}_i = \sum_{j=0}^{\rho-1} x_{ij}\hat{\beta}_j, i = 1,2,...,\nu$

11. Τα κατάλοιπα $\hat{\varepsilon}_i = y_i - \hat{y}_i, i = 1,2,...,\nu$

12. Την τιμή του *Durbin – Watson στατιστικού* για τον έλεγχο της αυτοσυσχέτισης των κατάλοιπων (βασική προϋπόθεσης της ανάλυσης παλινδρόμησης).

6.2: Ερμηνεία του πίνακα Πολλαπλής Παλινδρόμησης από το SPSS

Η *τιμή F* (F VALUE) μετρά το βαθμό συμφωνίας των δεδομένων με την μηδενική υπόθεση ότι όλοι οι συντελεστές β είναι ίσοι με το μηδέν. Η τιμή αυτή είναι ο λόγος των δύο μέσων τετραγώνων (MEAN SQUARE).

F=MODEL MEAN SQUARE/ ERROR MEAN SQUARE

Οι βαθμοί ελευθερίας στον αριθμητή και στον παρονομαστή είναι οι βαθμοί ελευθερίας του μοντέλου (MODEL DF) και οι βαθμοί ελευθερίας του σφάλματος (ERROR DF). Το *παρατηρούμενο επίπεδο σημαντικότητας (κρίσιμο επίπεδο, p- value)*, δηλαδή η πιθανότητα να καταλήξουμε σε μια τιμή της στατιστικής συνάρτησης F τόσο μεγάλη ή μεγαλύτερη από την παρατηρηθείσα τιμή, όταν η H_0 ισχύει,(PR>F).
Οι βαθμοί ελευθερίας του μοντέλου (MODEL DF) αναφέρεται πάντοτε στον αριθμό των παραμέτρων β του μοντέλου. Ο συνολικός αριθμός βαθμών ελευθερίας (TOTAL DF) αντιστοιχεί στον αριθμό παρατηρήσεων μείων 1, ενώ οι βαθμοί ελευθερίας του λάθους (ERROR DF) είναι η διαφορά μεταξύ των δύο προαναφερθέντων βαθμών ελευθερίας.

MODEL DF = k
ERROR DF = n-k-1
TOTAL DF = n-1

Τα αθροίσματα των τετραγώνων καθορίζονται με τον ίδιο τρόπο όπως στην απλή γραμμική παλινδρόμηση. Δηλαδή:

MODEL SUM OF SQUARES = SSTr = $\sum\left(\hat{Y}_i - \bar{Y}\right)^2$

ERROR SUM OF SQUARES = SSE = $\sum\left(Y_i - \widehat{Y}_i\right)^2$

TOTAL SUM OF SQUARES = SSTR = $\sum\left(Y_i - \overline{Y}\right)^2$

Προφανώς και πάλι, το συνολικό άθροισμα τετραγώνων προκύπτει ως το άθροισμα των δύο άλλων αθροισμάτων τετραγώνων. Τα μέσα τετράγωνα (MEAN SQUARE) είναι, όπως και προηγουμένως, η τιμή του συντελεστή προσδιορισμού του μοντέλου δηλαδή το ποσοστό της συνολικής διακύμανσης του Y που μπορεί να εξηγηθεί (να αποδοθεί) από το μοντέλο γραμμικής παλινδρόμησης που χρησιμοποιήθηκε

r^2=MODEL SS/TOTAL SS $\equiv \frac{SST_r}{SST}$ =1-ERROR SS/TOTAL SS $\equiv 1 - \frac{SSE}{SST}$

Η ετικέτα ESTIMATE δίνει τις τιμές b_i των στατιστικών συναρτήσεων (εκτιμητριών) $\widehat{\beta}_i$ ενώ η ετικέτα STD ERROR OF ESTIMATE δίνει τιμές των αντίστοιχων εκτιμητριών των αντίστοιχων αποκλίσεων $S_{\widehat{\beta}_i}$. Η τιμή της στατιστικής συνάρτησης T που αντιστοιχεί στο μοντέλο παλινδρόμησης που χρησιμοποιήθηκε εμφανίζεται στην ετικέτα T FOR H_0. Τέλος το παρατηρούμενο επίπεδο σημαντικότητας ή, αλλιώς ελάχιστο επίπεδο σημαντικότητας (P VALUE) που αντιστοιχεί στον έλεγχο $H_0 : b_i = 0$ έναντι της εναλλακτικής $H_1 : b_i \neq 0$ εμφανίζεται στην ετικέτα PR > $|T|$. Επομένως αν για παράδειγμα χρησιμοποιήσουμε ως επίπεδο σημαντικότητας την τιμή 0.05 οποιαδήποτε τιμή του παρατηρούμενου επιπέδου σημαντικότητας (P-VALUE) μεγαλύτερη από 0.05 είναι υποψήφια για να αποκλεισθεί από το γενικό γραμμικό μοντέλο που χρησιμοποιείται.

ΕΠΙΛΟΓΟΣ

Αναφερθήκαμε εκτενώς στην *Πολλαπλή Παλινδρόμηση (Multiple Regression)* εξετάζοντας όλες τις παραμέτρους της όπως το Μοντέλο, τη Μέθοδο των Ελαχίστων Τετραγώνων, τους συντελεστές προσδιορισμού και μερικής συσχέτισης. Επιπλέον κάναμε μνεία στην εξέταση των σφαλμάτων ε_i , είδαμε την επιλογή της «καλύτερης» εξίσωσης παλινδρόμησης μέσω των μεθόδων Backward, Forward και Stepwise. Εν τέλει, μιλήσαμε για τους μετασχηματισμούς στην περίπτωση που υφίστανται αποκλίσεις από τις υποθέσεις της Παλινδρόμησης, περίπτωση στην οποία υπάγεται και το πρόβλημα της Ετεροσκεδαστικότητας που είναι ένα αρκετά συχνό φαινόμενο.

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΕΛΛΗΝΙΚΕΣ ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ (REFERENCES)

Καρλής, Δημ. (2005). *Πολυμεταβλητή Στατιστική Ανάλυση*. Αθήνα: Εκδόσεις Αθ. Σταμούλης.

Κιόχος, Π. (1993). *Περιγραφική Στατιστική*. Αθήνα: Εκδόσεις Interbooks.

Πανάρετος, Ι. (1997). *Γραμμικά μοντέλα με έμφαση στις εφαρμογές (3η έκδ)*. Αθήνα.

Πανάρετος, Ι. & Ξεκαλάκη Ε. (1993). *Εισαγωγή στην στατιστική σκέψη- Τόμος 1 (Περιγραφική Στατιστική)*. Αθήνα.

Σφακιανάκης, Μ. (2000). *Υπολογιστική Στατιστική*. Αθήνα.

Μπόρα- Σέντα Ε. & Μωϋσιάδης Χ. (1997). *Εφαρμοσμένη Στατιστική (2η έκδ.).* Θεσσαλονίκη.

ΞΕΝΕΣ ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ (REFERENCES)

Howitt, D., & Cramer, D. (2006). *Στατιστική με το SPSS 13* (3η έκδ.). Εκδόσεις Κλειδάριθμος.

Draper Norman & Smith Harry (1997). *Εφαρμοσμένη Ανάλυση Παλινδρόμησης (2η έκδ.*) Αθήνα.

ΗΛΕΚΤΡΟΝΙΚΕΣ ΠΗΓΕΣ (ELECTRONIC SOURCES)

http://en.wikipedia.org/wiki/Main_Page (Wikipedia)