ΤΕΙ ΠΑΤΡΩΝ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ
ΕΠΙΧΕΙΡΗΜΑΤΙΚΟΥ ΣΧΕΔΙΑΣΜΟΥ
ΚΑΙ ΠΛΗΡΟΦΟΡΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΒΗΜΑΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΣΕ ΟΙΚΟΝΟΜΙΚΑ ΔΕΔΟΜΕΝΑ

Επιβλέπων Καθηγητής: κ. ΓΕΩΡΓΙΟΥ ΒΑΣΙΛΕΙΟΣ

ΔΕΛΗΓΙΑΝΝΗΣ ΔΗΜΗΤΡΙΟΣ ΑΜ : 564

ΠΑΤΡΑ 2008

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ, ΕΙΚΟΝΩΝ, ΠΙΝΑΚΩΝ ΚΑΙ ΔΙΑΓΡΑΜΜΑΤΩΝ

Σχήμα **Σελ.**

Πίνακας

Εικόνα

Διάγραμμα

ΕΙΣΑΓΩΓΗ

Συχνά δύο ή περισσότερες ποσοτικές μεταβλητές εξετάζονται μαζί, με την ελπίδα να προσδιοριστεί η σχέση που υποτίθεται ότι υπάρχει μεταξύ τους. Για παράδειγμα, θα μπορούσε να εξεταστεί η επίδραση που έχει η θερμοκρασία στα πειραματικά αποτελέσματα μιας χημικής διαδικασίας.
Πολλές φορές επίσης, δύο ή περισσότερες ποσοτικές μεταβλητές εξετάζονται με στόχο την πρόβλεψη της μίας από τις άλλες. Για παράδειγμα, γνωρίζοντας τη σχέση μεταξύ εξόδων που δαπανώνται για διαφήμιση και εσόδων από πωλήσεις κάποιου προϊόντος, θα ήταν χρήσιμο να προβλεφθούν, κατά πιθανότητα, τα έσοδα όταν για τη διαφήμιση του προϊόντος δαπανηθεί κάποιο γνωστό ποσό .
Η στατιστική μεθοδολογία που χρησιμοποιεί τη σχέση μεταξύ δύο ή περισσότερων ποσοτικών μεταβλητών έτσι ώστε η μία να μπορεί να προβλεφθεί από την άλλη ή τις άλλες καλείται Ανάλυση Παλινδρόμησης (Regression Analysis) .Τον όρο παλινδρόμηση χρησιμοποίησε για πρώτη φορά ο F. Galton το 1900 όταν μελετώντας τη σχέση μεταξύ του ύψους των γονέων και αυτού των παιδιών τους, παρατήρησε ένα είδος επαναφοράς (παλινδρόμησης) του ύψους των παιδιών στο ύψος των γονέων τους.
Η Ανάλυση Παλινδρόμησης αποτελεί έναν από τους σημαντικότερους κλάδους της Στατιστικής με ευρείες εφαρμογές σε όλες τις σύγχρονες επιστήμες από τις Θεωρητικές μέχρι τις Επιχειρηματικές, Οικονομικές, Βιολογικές, Κοινωνικές κ.λ.π. Επιπλέον περιέχει πολλά και ενδιαφέροντα θεωρητικά θέματα.
Με τη φράση ανάλυση παλινδρόμησης εννοούμε διάφορες γραφικές και αναλυτικές μεθόδους που σκοπό έχουν την αναζήτηση σχέσεων μεταξύ μιας μεταβλητής, την οποία ονομάζουμε εξαρτημένη μεταβλητή (dependent variable) και μιας άλλης ή άλλων μεταβλητών, τις οποίες ονομάζουμε ανεξάρτητες μεταβλητές (independent variables). Όταν μια τέτοια σχέση (μοντέλο) βρεθεί, τότε μπορούμε να χρησιμοποιήσουμε το μοντέλο αυτό για να κάνουμε προβλέψεις, για να βρούμε ποιες από τις ανεξάρτητες μεταβλητές επηρεάζουν περισσότερο την εξαρτημένη μεταβλητή, ή να ελέγξουμε διάφορες υποθέσεις.
Στο σημείο αυτό θα πρέπει να τονίσουμε την διαφορά μεταξύ της ανάλυσης παλινδρόμησης (regression analysis) και της ανάλυσης συσχετίσεως (correlation analysis). Στην ανάλυση παλινδρόμησης η σχέση, την οποία αναφέραμε προηγουμένως, είναι προς μία κατεύθυνση μόνο. Δηλαδή αγνοεί την πιθανή επίδραση της εξαρτημένης μεταβλητής στις ανεξάρτητες μεταβλητές. Σε μερικές περιπτώσεις, κυρίως σε εργαστηριακά πειράματα, ο ερευνητής μπορεί να προκαθορίσει τις τιμές των ανεξάρτητων μεταβλητών και στη συνέχεια να παρατηρήσει το μέγεθος της εξαρτημένης μεταβλητής. Στην περίπτωση αυτή είναι φανερό ότι η σχέση είναι προς μία μόνο κατεύθυνση. Σε πολλές περιπτώσεις όμως ο ερευνητής παρατηρεί συγχρόνως όλες τις μεταβλητές. Η ανάλυση παλινδρόμησης, για τέτοια δεδομένα, μπορεί και πάλι να χρησιμοποιηθεί όταν ο σκοπός της μελέτης μας είναι να μελετήσουμε την μεταβολή μιας εξ αυτών σε σχέση με τις υπόλοιπες.
Η ανάλυση συσχετίσεως χρησιμοποιείται όταν κάποιος θέλει να μελετήσει την ταυτόχρονη μεταβλητότητα ενός συνόλου μεταβλητών. Στην ανάλυση συσχετίσεως οι σχέσεις μεταξύ των διάφορων μεταβλητών δεν είναι, γενικώς, μονοκατευθυντικές. Ένα απλό παράδειγμα όπου η χρήση της ανάλυσης συσχετίσεως είναι αναγκαία είναι το εξής. Ας υποθέσουμε ότι κάποιος ερευνητής ενδιαφέρεται να μελετήσει την ταυτόχρονη μεταβολή (σαν συνάρτηση της ηλικίας) του ύψους και του βάρους κάποιου πληθυσμού. Ο σκοπός, στην περίπτωση αυτή, δεν είναι να

περιγράψουμε πως μεταβάλλεται το ύψος συναρτήσει του βάρους ή αντίστροφα, αλλά να μελετήσουμε την από κοινού μεταβολή του ύψους και του βάρους, στον εν λόγω πληθυσμό, από τα νήπια μέχρι τους ηλικιωμένους. Από την άλλη πλευρά, εάν θέλουμε να προβλέψουμε το ύψος από το βάρος, θα πρέπει να χρησιμοποιήσουμε μεθόδους ανάλυσης παλινδρόμησης.

Έχει αναφερθεί το πρόβλημα της παλινδρόμησης, σε σχέση με τυχαίες μεταβλητές των οποίων γνωρίζαμε την από κοινού κατανομή. Στην πράξη όμως υπάρχουν προβλήματα όπου η από κοινού κατανομή των τυχαίων μεταβλητών δεν είναι γνωστή και κατά συνέπεια η παλινδρόμηση της μιας εξ αυτών ως προς τις άλλες δεν μπορεί να βρεθεί. Στην περίπτωση αυτή, με βάση τα δεδομένα ενός τυχαίου δείγματος, πρέπει να εκτιμήσουμε την εν λόγω παλινδρόμηση.

Στη συνέχεια θα ασχοληθούμε με μια ειδική κατηγορία μοντέλων παλινδρόμησης, τα οποία όμως έχουν αρκετά ευρύ φάσμα εφαρμογών, τα μοντέλα γραμμικής παλινδρόμησης.

ΚΕΦΑΛΑΙΟ 1 ΓΕΝΙΚΑ[1]

Κατά την μελέτη φαινομένων όπου αναλύονται συγχρόνως περισσότερες της μιας τυχαίες μεταβλητές, εκείνο που διερευνάται κατ' αρχή είναι η ύπαρξη ή όχι αλληλοεπίδρασης / αλληλεξάρτησης μεταξύ τους.
Σε δείγμα n περιπτώσεων (παρατηρήσεων) όπου μελετούνται ταυτοχρόνως δύο τυχαίες μεταβλητές X και Y, εξετάζεται η συμπεριφορά του πλήθους των τιμών y_j, ή y, που εμφανίζονται σε ζεύγη (x_i,y_i) ή (x,y) για κάθε ξεχωριστή συγκεκριμένη τιμή x_i ή x. Δηλαδή, για κάθε τιμή y_i εμφανίζεται ένα σύνολο τιμών $x_j= \{x_1, x_2,…., x_j\}$, υπό μορφή (x_1, y_i), (x_2, y_i),….,(x_j, y_i), και όχι μια μοναδική τιμή x_j.
Το στατιστικό μέγεθος που δείχνει τη σχέση μεταξύ των δύο τυχαίων μεταβλητών X και Y και συγχρόνως μετράει την ένταση αυτής της σχέσης είναι ο συντελεστής συσχέτισης ρ_{xy} ή ο συντελεστής θεωρητικής συσχέτισης ρ

$\rho_{xy} = \sigma_{xy} / \sigma_x*\sigma_y$

Αποδεικνύεται ότι ο συντελεστής συσχέτισης ρ_{xy} ικανοποιεί την ανισότητα

$-1 \le \rho_{xy} \le +1$
Από την παραπάνω σχέση φαίνεται:

I. $0 < \rho \le +1$, οι δύο τυχαίες μεταβλητές X και Y χαρακτηρίζονται ως θετικά συσχετισμένες
II. $-1 \le \rho < 0$, οι δύο τυχαίες μεταβλητές X και Y χαρακτηρίζονται ως αρνητικά συσχετισμένες
III. $\rho = 0$, οι δύο τυχαίες μεταβλητές X και Y χαρακτηρίζονται ως ασυσχέτιστες μεταξύ τους.

Όταν $\rho = 0$, δηλαδή όταν οι X και Y είναι ασυσχέτιστες, δεν σημαίνει ότι οι X και Y είναι και στοχαστικά ανεξάρτητες. Στην περίπτωση αυτής της στοχαστικής εξάρτησης, οι δύο μεταβλητές συνδέονται με κάποια μη γραμμική σχέση.
Όπως φαίνεται από την $\rho_{xy} = \sigma_{xy} / \sigma_x*\sigma_y$, ο παρονομαστής είναι πάντα ο ίδιος. Κατά συνέπεια, η τιμή του ρ_{xy} εξαρτάται από το μέγεθος του $\sigma_{xy,}$ το οποίο είναι άθροισμα (ολοκλήρωμα) των γινομένων $(X – \mu_X)*(Y – \mu_Y)$. Η μέγιστη τιμή του σ_{xy} επιτυγχάνεται όταν η σειρά κατάταξης των τιμών της X, για το σύνολο των περιπτώσεων, είναι η ίδια με τη σειρά κατάταξης των τιμών της Y. Δηλαδή όταν μεγάλες τιμές της X ($x > \mu_X$) εμφανίζονται συχνότερα (σε ζεύγη) με μεγάλες τιμές της Y ($y > \mu_Y$) και αντίστοιχα μικρές τιμές της X συσχετίζονται συχνότερα (σε ζεύγη) με μικρές τιμές της Y . Τότε οι όροι $(x – \mu_X)$ και $(y – \mu_Y)$ εμφανίζονται συχνότερα, είτε και οι δύο θετικοί είτε και οι δύο αρνητικοί. Κατά συνέπεια το γινόμενο $(x – \mu_X)*(y – \mu_Y)$ εμφανίζεται συχνότερα θετικό και η συσχέτιση ρ_{xy} έχει θετική τιμή. Ανάλογα όταν μεγάλες τιμές της X ($x > \mu_X$) εμφανίζονται συχνότερα με μικρές τιμές της Y ($y < \mu_Y$), ενώ μικρές τιμές της X ($x < \mu_X$) εμφανίζονται συχνότερα με μεγάλες τιμές της Y ($y > \mu_Y$), τότε το γινόμενο $(x – \mu_X)*(y – \mu_Y)$ εμφανίζεται συχνότερα αρνητικό και η συσχέτιση ρ_{xy} έχει αρνητική τιμή.
Παραδείγματα τυχαίων μεταβλητών που παρουσιάζουν τα τρία είδη συσχέτισης :

[1] Βλέπε http : //web.auth.gr/e-topo/TOMEIS_INDEX/TOMEASB/Lafazani/Give/kef10_2_Palindr_sysxet.pdf Σελ.400

- Το ύψος και το βάρος των ανθρώπων είναι δύο θετικά συσχετισμένες τυχαίες μεταβλητές. Συνήθως, οι υψηλότεροι άνθρωποι είναι και βαρύτεροι.
- Η απασχόληση και η ανεργία σε μια γεωγραφική περιοχή είναι δύο αρνητικά συσχετισμένες τυχαίες μεταβλητές. Όπου υπάρχει μεγάλο ποσοστό απασχόλησης, το ποσοστό ανεργίας είναι χαμηλό.
- Το ύψος και το μηνιαίο εισόδημα των ανθρώπων είναι δύο ασυσχέτιστες τυχαίες μεταβλητές. Με κανένα τρόπο δεν επηρεάζει η μια μεταβλητή τη άλλη.

1.1 Σχέσεις Μεταξύ Μεταβλητών

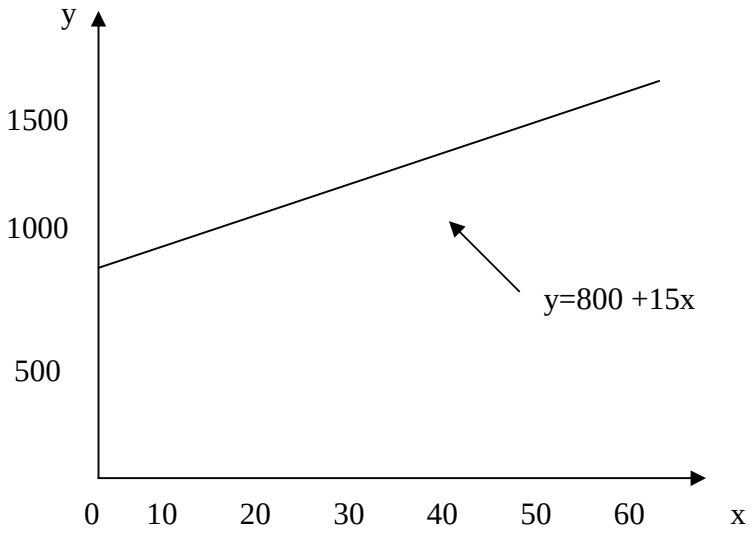
Η έννοια της σχέσης μεταξύ δύο μεταβλητών, όπως π.χ. μεταξύ εσόδων και εξόδων μιας οικογένειας ή ηλικίας και βάρους ενός παιδιού θεωρείται γνωστή. Οι σχέσεις μεταξύ μεταβλητών διακρίνονται σε συναρτησιακές και στατιστικές ανάλογα με το αν οι μεταβλητές είναι μη στοχαστικές ή στοχαστικές (τυχαίες) .

Συναρτησιακή σχέση μεταξύ δύο μεταβλητών

Η συναρτησιακή σχέση μεταξύ δύο μεταβλητών εκφράζεται με κάποιο μαθηματικό τύπο. Αν x είναι η ανεξάρτητη μεταβλητή και y η εξαρτημένη, μια συναρτησιακή σχέση είναι του τύπου y=f(x). Για δεδομένη τιμή του x η συνάρτηση f μας δίνει την αντίστοιχη τιμή του y .

Παράδειγμα

Μια εταιρεία ενοικιάζει τα ποδήλατα της με βάση τη σχέση y=800 +15x όπου, y είναι το ποσό (σε €) που ο πελάτης θα πληρώσει για να διανύσει μια απόσταση x (σε km) . Δύο άτομα που πρόσφατα νοίκιασαν ποδήλατα από την εν λόγω εταιρεία, πλήρωσαν 1100 € και 1565 € για αποστάσεις 20 km και 51 km αντίστοιχα. Η γραφική παράσταση της συναρτησιακής σχέσης που χρησιμοποιεί η εταιρεία όπως και τα δύο ζεύγη (xi, yi) των παρατηρήσεων (δεδομένων) που αναφέραμε δίνονται στο Σχήμα 1 . Είναι φανερό ότι τα σημεία που αντιστοιχούν στις παρατηρήσεις είναι σημεία καμπύλης, γραφικής παράστασης της συναρτησιακής σχέσης, που εδώ είναι ευθεία γραμμή .



Σχήμα 1 Παράδειγμα συναρτησιακής σχέσης

Στατιστική σχέση μεταξύ δύο μεταβλητών

Η στατιστική σχέση διαφέρει από την συναρτησιακή στο ότι δεν συνιστά μια τέλεια σχέση. Γενικά, οι παρατηρήσεις σε μια στατιστική σχέση δεν αποτελούν σημεία της καμπύλης που δίνεται από τη σχέση.

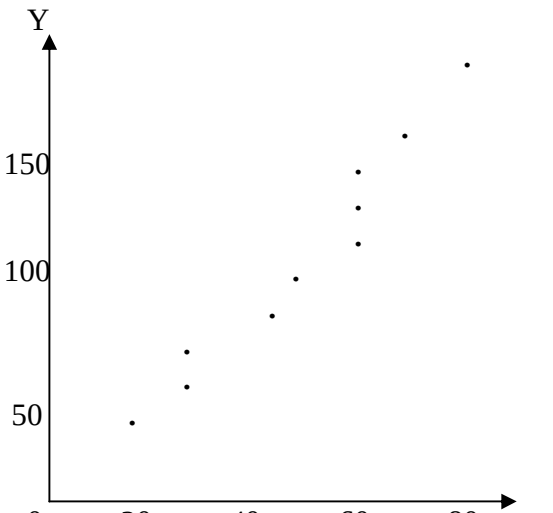
Παράδειγμα

Η Ελληνική Βιομηχανία Όπλων (ΕΒΟ) απασχολεί μέρος του εργατικού δυναμικού της κατασκευάζοντας κάνες ενός τύπου όπλων. Η απασχόληση σε ώρες εργασίας ανά μήνα διαφέρει από μήνα σε μήνα ανάλογα με το μέγεθος της παραγγελίας που πρέπει να εκτελεστούν κάθε μήνα.
Στον Πίνακα 1 δίνονται οι παραγγελίες προς την ΕΒΟ των δέκα τελευταίων μηνών μαζί με τις αντίστοιχες ώρες εργασίας. Τα δεδομένα του πίνακα που αποτελούν παρατηρήσεις δύο μεταβλητών παριστάνονται γραφικά στο Σχήμα α. Ως εξαρτημένη μεταβλητή (Υ) θεωρούμε τις ώρες εργασίας και ως ανεξάρτητη (Χ) το μέγεθος της παραγγελίας. Το Σχήμα β προκύπτει από το Σχήμα α χαράσσοντας την πλησιέστερη ευθεία προς τα δεδομένα σημεία. Εδώ φαίνεται καθαρά ότι κάποια σχέση πρέπει να υπάρχει μεταξύ μεγέθους παραγγελίας και των ωρών εργασίας που απαιτούνται για την εκτέλεση της από το γεγονός ότι για μεγαλύτερο αριθμό παραγγελιών υπάρχει η ένδειξη ότι απαιτούνται περισσότερες ώρες εργασίας. Όμως η σχέση δεν

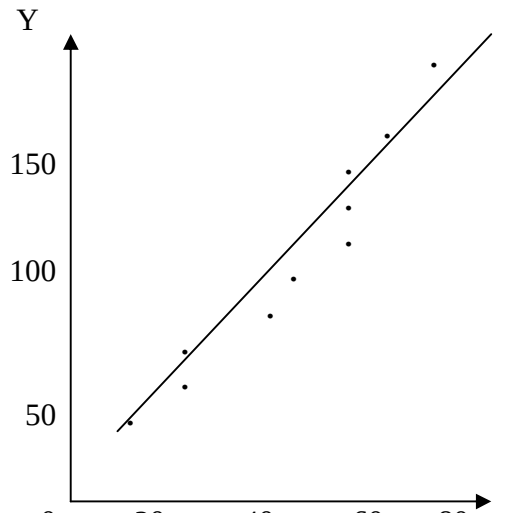
είναι απόλυτη . Βλέπουμε να υπάρχει μια διασπορά των σημείων που δείχνει ότι μέρος της μεταβλητότητας των ωρών εργασίας δε συνδέεται με το μέγεθος της παραγγελίας . Για παράδειγμα, το μέγεθος παραγγελίας του $1^{ου}$ και του $8^{ου}$ μήνα ήταν το ίδιο (30) όμως για την εκτέλεση απαιτήθηκε διαφορετικός αριθμός ωρών εργασίας. Λόγω της διασποράς των σημείων που παρατηρείται σε μια στατιστική σχέση, ένα σχήμα σαν το α λέγεται διάγραμμα διασποράς . Στη στατιστική ορολογία κάθε σημείο ενός διαγράμματος αποτελεί μία παρατήρηση. Η ευθεία γραμμή που δίνεται στο σχήμα β εκφράζει την στατιστική σχέση μεταξύ ωρών εργασίας και μεγέθους παραγγελίας. Η ευθεία αυτή παριστά γραφικά τη γενική τάση με την οποία οι ώρες εργασίας επηρεάζονται από το μέγεθος της παραγγελίας. Παρατηρούμε ότι τα περισσότερα σημεία δεν αποτελούν σημεία της ευθείας γραμμής. Αυτή η διασπορά των σημείων γύρω από την γραμμή αντιπροσωπεύει το μέρος της μεταβλητότητας των ωρών εργασίας που δε συνδέεται με το μέγεθος παραγγελίας και που συνήθως θεωρείται ότι είναι τυχαίας προέλευσης.
Οι στατιστικές σχέσεις, αν και δεν παρέχουν την ακρίβεια μίας συναρτησιακής, αποτελούν ένα καλύτερο τρόπο μελέτης ενός φαινομένου ή προβλήματος δεδομένου ότι τα φαινόμενα διέπονται από νόμους τύχης, περιέχουν αβεβαιότητες κλπ. που δε μπορούν να εκφραστούν σε μια συναρτησιακή σχέση που πάντα εξιδανικεύει την πραγματικότητα και συνήθως είναι άγνωστη. Για αυτό οι στατιστικές σχέσεις έχουν εφαρμογές σε όλες σχεδόν τις επιστήμες.

Μήνας Παραγωγής I	Μέγεθος Παραγγελίας xi	Ώρες Εργασίας yi
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132

Πίνακας 1 Παραγγελίες ΕΒΟ



Σχήμα α Διάγραμμα Διασποράς



Σχήμα β Στατιστική σχέση

1.2 ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ[2]

Για όλους μας η λέξη ανάλυση έχει μια κάποια σημασία, είτε αυτή είναι αφηρημένη (π.χ. ανάλυση ενός γεγονότος ή ανάλυση μιας υπόθεσης κτλ) είτε έχει πιο συγκεκριμένη μορφή (π.χ. χημική ανάλυση, ιατρική ανάλυση κτλ.).
Σε οποιαδήποτε όμως μορφή της η λέξη ανάλυση δηλώνει τη βαθύτερη κατανόηση κάποιου αντικειμένου και κατά συνέπεια την πιθανή εξαγωγή συμπερασμάτων. Σ' αυτές τις γενικές γραμμές κινείται και η λεγόμενη στατιστική ανάλυση. Πιο συγκεκριμένα το αντικείμενο της στατιστικής ανάλυσης είναι να ανακαλύψει τι συμπεράσματα μπορούν να εξαχθούν, από ένα σύνολο δεδομένων, και να παρουσιάσει τα συμπεράσματα αυτά κατά τρόπο i) απλό και σαφή και ii) συνεπή και ακριβή.

Η στατιστική ανάλυση ή απλώς ανάλυση περιλαμβάνει τα εξής στάδια:

1)Αρχική επεξεργασία των δεδομένων. Το στάδιο αυτό περιλαμβάνει: i) τον υπολογισμό διαφόρων συγκεντρωτικών ποσοτήτων, (π.χ. μέση τιμή, διακύμανση) για κάθε μεταβλητή που υπεισέρχεται στο πρόβλημα. ii) Την γραφή των δεδομένων κατά τρόπο κατάλληλο για την λεπτομερή ανάλυση που θα ακολουθήσει και iii) τον ποιοτικό έλεγχο των δεδομένων. Με την φράση ποιοτικό έλεγχο των δεδομένων εννοούμε α) τον έλεγχο, οπτικό ή αυτόματο, των δεδομένων για την εύρεση τιμών οι οποίες δεν είναι λογικά συνεπείς με το υπόλοιπο σύνολο των δεδομένων (π.χ. μία ή περισσότερες τιμές οι οποίες είναι αρκετά μεγαλύτερες ή μικρότερες από

[2] Βλέπε Βιβ. Γραμμικά Μοντέλα Κεφάλαιο 1 Σελ. 1 έως 5

το κύριο σώμα των δεδομένων) ή τιμές οι οποίες είναι εκτός του εύρους που περιμέναμε για τις διάφορες μεταβλητές (π.χ για μια μεταβλητή βρίσκουμε αρνητική τιμή ενώ είναι γνωστό ότι η μεταβλητή αυτή μπορεί να πάρει μόνο θετικές τιμές, β) διάφορες γραφικές παραστάσεις, ανά δύο, των μεταβλητών προς εξακρίβωση της σχέσης μεταξύ τους, γ) έλεγχος της μεθόδου ή των μεθόδων με τις οποίες συγκεντρώθηκαν τα δεδομένα. Ο έλεγχος αυτός γίνεται για την εξακρίβωση του κατά πόσο οι παρατηρήσεις μας είναι σωστές ή περιέχουν κάποιο σφάλμα και δ) έλεγχος για την ύπαρξη χαμένων παρατηρήσεων (missing values), δηλαδή παρατηρήσεων που για διάφορους λόγους δεν έχουν καταγραφεί στα προς εξέταση δεδομένα.
2) Κύρια ανάλυση, η οποία πρόκειται να αποτελέσει και τη βάση για την εξαγωγή συμπερασμάτων, και
3) Παρουσίαση των συμπερασμάτων κατά τρόπο ακριβή, συνεπή και κατανοητό. Συνήθως η ερμηνεία των δεδομένων περιλαμβάνει και κάποιο ή πολλά στοιχεία υποκειμενικότητας.

Τα προηγούμενα στάδια μας βοηθούν στην εύρεση του στατιστικού μοντέλου. Ένα μοντέλο, γενικώς, είναι η μαθηματική προσέγγιση σε κάποιο φαινόμενο ή μηχανισμό που παράγει τα προς ανάλυση δεδομένα. Η κύρια διαφορά μεταξύ ενός στατιστικού (στοχαστικού) μοντέλου και ενός αιτιοκρατικού μοντέλου είναι ότι τα συμπεράσματα που διατυπώνονται στην πρώτη περίπτωση λαμβάνουν υπ' όψιν τους την αβεβαιότητα, τα σφάλματα ή την τυχαιότητα, που υπάρχει στο υπό μελέτη φυσικό φαινόμενο.
Για την εύρεση ενός στατιστικού μοντέλου δηλαδή για τα στάδια (1) και (2) υπάρχουν διάφορες μέθοδοι και τεχνικές. Εμείς εδώ θα ασχοληθούμε με τα μοντέλα της γραμμικής παλινδρόμησης. Κύριο ρόλο για την μελέτη των μοντέλων αυτών, αλλά και πιο προχωρημένων στατιστικών μεθόδων, παίζει η μέθοδος των ελαχίστων τετραγώνων.
Στη μαθηματική του μορφή το πρόβλημα της παλινδρόμησης έχει ως εξής: δοθέντων δύο τυχαίων μεταβλητών (τ.μ.) Y και X να ευρεθεί η $E(Y|X=x)$. Με άλλα λόγια ζητάμε να βρούμε την υπό συνθήκη αναμενόμενη τιμή της τ.μ. Y για δοθείσα τιμή x της τ.μ. X. Η σχέση που μας δίνει η $E(Y|X=x)$ παριστάνει ένα στατιστικό μοντέλο και ονομάζεται απλή παλινδρόμηση της Y ως προς X. Για περισσότερες από δύο τ.μ. η γενίκευση του παραπάνω προβλήματος είναι ο προσδιορισμός της $E(Y|X_1=x_1,X_2=x_2,...,X_\kappa=x_\kappa)$. Και στην περίπτωση αυτή η σχέση που μας δίνει η $E(Y|X_1=x_1,X_2=x_2,....,X_\kappa=x_\kappa)$ παριστάνει ένα στατιστικό μοντέλο και καλείται πολλαπλή παλινδρόμηση.
Μολονότι το πρόβλημα της απλής ή πολλαπλής παλινδρόμησης στη θεωρητική του μορφή έχει λύση, στην πράξη τα πράγματα είναι διαφορετικά. Ή η από κοινού κατανομή των μεταβλητών που μελετάμε δεν είναι γνωστή ή όταν είναι γνωστή περιλαμβάνει άγνωστες παραμέτρους. Στην τελευταία περίπτωση η χρήση δεδομένων, για την εκτίμηση των άγνωστων παραμέτρων, είναι αναγκαία.

Θα κλείσουμε με μερικά γενικά σχόλια για τα στατιστικά μοντέλα.
i) Για κάθε στατιστικό μοντέλο υπάρχουν τουλάχιστον δύο διαφορετικές φυσικές ερμηνείες. Στην πρώτη υπάρχει ένας καλώς ορισμένος πληθυσμός μονάδων, ατόμων, αντικειμένων κτλ. από τον οποίο εκλέγουμε τυχαία ένα δείγμα. Στην περίπτωση αυτή οι ιδιότητες του μοντέλου μας, το οποίο σημειωτέον βρήκαμε βασισμένοι στο εν λόγω δείγμα, ισχύουν για όλο τον πληθυσμό. Μια πιο συνηθισμένη περίπτωση είναι αυτή που οι παρατηρήσεις μας (το δείγμα) προέρχονται από κάποιο σύστημα το οποίο υπόκειται σε τυχαίες διακυμάνσεις. Στην περίπτωση αυτή, όπου η λέξη πληθυσμός χρησιμοποιείται τελείως υποθετικά, το στατιστικό μοντέλο αντανακλά κυρίως τις ιδιότητες του δείγματος. Ακόμη καλύτερα το στατιστικό μοντέλο μας περιγράφει τι θα συνέβαινε

αν οι παρατηρήσεις συνεχιζόταν επ' άπειρο κάτω από τις ίδιες συνθήκες. Επειδή η διατήρηση των ίδιων συνθηκών είναι μια δύσκολη ή αδύνατη υπόθεση γι' αυτό η εγκυρότητα του μοντέλου μας θα πρέπει να δοκιμασθεί και κάτω από διαφορετικές συνθήκες ή να συγκριθεί με παρόμοια αποτελέσματα σχετικών πειραμάτων.
ii) Ένα στατιστικό μοντέλο συνήθως περιλαμβάνει άγνωστες παραμέτρους. Οι παράμετροι αυτοί παίζουν ένα βασικό ρόλο. Ένα από τα κύρια προβλήματα της στατιστικής είναι η χρησιμοποίηση των δεδομένων κατά τον καλύτερο δυνατό τρόπο για την εκτίμηση των εν λόγω παραμέτρων. Σε κάθε συγκεκριμένο πρόβλημα μερικές από τις παραμέτρους είναι αμέσου ενδιαφέροντος και μερικές εμμέσου ενδιαφέροντος .
iii) Το στατιστικό μοντέλο είναι αρκετές φορές προσωρινό. Σε μερικές περιπτώσεις, θεωρητικά αποτελέσματα ή προηγούμενη εμπειρία καθιστούν την εύρεση του μοντέλου ένα εύκολο έργο. Σε άλλες περιπτώσεις όμως, όπου τα δεδομένα είναι πιο σύνθετα και η εμπειρία ή τα θεωρητικά αποτελέσματα δεν υπάρχουν, η εύρεση ενός κατάλληλου μοντέλου είναι έργο αρκετά δύσκολο. Στην περίπτωση αυτή ξεκινάμε με ένα κάποιο μοντέλο, που μας φαίνεται λογικό. Στην συνέχεια ελέγχουμε την ορθότητα ή μη του εν λόγω μοντέλου. Αν, το μοντέλο μας, δεν είναι σωστό το τροποποιούμε ανάλογα και ξαναρχίζουμε από την αρχή.
iv) Το στατιστικό μοντέλο που θα καταλήξουμε θα πρέπει να είναι όσο γίνεται πιο απλό (π.χ. να περιλαμβάνει μικρό αριθμό αγνώστων παραμέτρων) ενώ συγχρόνως θα πρέπει να ερμηνεύει ικανοποιητικά τα πειραματικά δεδομένα.

1.3 ΓΕΝΙΚΟ ΓΡΑΜΜΙΚΟ ΥΠΟΔΕΙΓΜΑ Ή ΜΟΝΤΕΛΟ[3]

Κατά τη μελέτη διαφόρων πραγματικών φαινομένων, είτε οικονομικού, είτε κοινωνικού, είτε δημογραφικού, είτε γεωγραφικού χαρακτήρα, διερευνάται η ταυτόχρονη επίδραση διαφόρων παραγόντων στη διαμόρφωση του θεωρουμένου φαινομένου. Εάν οι διάφοροι παράγοντες θεωρηθούν ως τυχαίες μεταβλητές $X_1, X_2, \ldots, X_N$ και η μορφή του διαμορφούμενου φαινομένου στον συγκεκριμένο χωρόχρονο ως τυχαία μεταβλητή Y, τότε η διερεύνηση του φαινομένου ανάγεται στο πρόβλημα της ανάλυσης της σχέσης μεταξύ των θεωρούμενων n + 1 μεταβλητών. Αυτή η σχέση δημιουργεί συγχρόνως ένα υπόδειγμα ή μοντέλο του πραγματικού φαινομένου που μελετάται.
Ως βασική προϋπόθεση σε τέτοιου είδους προβλήματα τίθεται η γραμμικότητα, ως προς τις παραμέτρους των μοντέλων. Αυτό σημαίνει ότι το μαθηματικό μοντέλο που προσδιορίζεται δεν θα εκφράζει πάντα την πραγματική σχέση μεταξύ των μεταβλητών, δηλαδή δεν θα ανταποκρίνεται πάντοτε στην πραγματικότητα. Θα υπάρχει μια διαφορά μεταξύ της παρατηρούμενης τιμής της Y και αυτής που παρέχεται από το γραμμικό μοντέλο. Αυτή η διαφορά θεωρείται ως μια άλλη τυχαία μεταβλητή E που ονομάζεται σφάλμα (ή υπόλοιπο ή κατάλοιπο) και οφείλεται σε μια σειρά από παραμέτρους και μεταβλητές που δεν ελήφθησαν υπόψη κατά τη μελέτη του φαινομένου. Έτσι για κάθε περίπτωση / παρατήρηση κατά τον χρόνο t το μαθηματικό μοντέλο θα είναι της μορφής :

$Y_t = z\,(X_{t1}, X_{t2}, \ldots, X_{tn}) + E_t$

Τα μοντέλα αυτού του είδους ονομάζονται στοχαστικά σε αντίθεση με τα προσδιοριστικά μοντέλα, σύμφωνα με τα οποία εφόσον η παρατηρούμενη μεταβλητή Y δεν υπόκειται σε

[3] Βλέπε http : //web.auth.gr/e-topo/TOMEIS_INDEX/TOMEASB/Lafazani/Give/kef10_2_Palindr_sysxet.pdf Σελ. 404 έως 406.

σφάλματα, είναι δυνατή η ακριβής πρόβλεψη των τιμών της Υ από τις τιμές της Χ. Στην πραγματικότητα όμως, δεν μπορεί να προβλεφθεί ακριβώς η τιμή της Υ.
Όταν στο στοχαστικό μοντέλο οι μεταβλητές είναι οποιαδήποτε μορφής τότε ονομάζεται γενικό γραμμικό υπόδειγμα ή μοντέλο και δίνεται από τη σχέση :

$Y = b_0 + b_1X_1 + b_2X_2 + \ldots. + b_nX_n + E, \qquad n > 1$

όπου :
$X_1, X_2, \ldots, X_n$: ανεξάρτητες ή ελεγχόμενες μεταβλητές που η μέτρησή τους δεν υπόκειται σε σφάλματα
$b_0, b_1, b_2, \ldots, b_n$: άγνωστες παράμετροι
Ε : τυχαία σφάλματα που ακολουθούν την κανονική κατανομή, δηλαδή $E[e] = 0$

Ειδικές περιπτώσεις του γενικού γραμμικού μοντέλου αποτελούν :

I. Το μοντέλο της Απλής Παλινδρόμησης : Η ποσοτική μεταβλητή Υ συνδέεται μόνο με μια άλλη ποσοτική μεταβλητή Χ.
II. Το μοντέλο της Πολλαπλής Παλινδρόμησης : Η ποσοτική μεταβλητή Υ συνδέεται με ένα πλήθος n ποσοτικών μεταβλητών, $X_1, X_2, \ldots, X_n$.
III. Το μοντέλο της Ανάλυσης Διακύμανσης : Οι μεταβλητές $X_1, X_2, \ldots, X_n$ είναι ή θεωρούνται ποιοτικές.
IV. Το μοντέλο της Ανάλυσης Συνδιακύμανσης : Οι μεταβλητές $X_1, X_2, \ldots, X_n$ είναι οποιουδήποτε χαρακτήρα, άλλες ποιοτικές και άλλες ποσοτικές.

Το μοντέλο παλινδρόμησης αποτελεί την τυπική περίπτωση έκφρασης δύο βασικών αρχών μιας στατιστικής σχέσης: (i) της διασποράς των παρατηρήσεων γύρω από την καμπύλη της στατιστικής σχέσης και (ii) της τάσης της εξαρτημένης μεταβλητής y να μεταβάλλεται με την ανεξάρτητη ή τις ανεξάρτητες μεταβλητές σύμφωνα με κάποιο συστηματικό τρόπο. Τα στοιχεία αυτά, περιέχονται σε ένα μοντέλο παλινδρόμησης αν υποτεθεί ότι (i) Οι τιμές της εξαρτημένης μεταβλητής y που περιέχονται στα δεδομένα αποτελούν τυχαίο δείγμα από έναν πληθυσμό με κάποια κατανομή για κάθε διακεκριμένη τιμή της ανεξάρτητης μεταβλητής x. (ii) Οι μέσες τιμές των κατανομών αυτών μεταβάλλονται με κάποιο συστηματικό τρόπο.

ΚΕΦΑΛΑΙΟ 2 ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Με την απλή γραμμική παλινδρόμηση, όπως έχει αναφερθεί, αναλύεται η σχέση μεταξύ δύο ποσοτικών μεταβλητών των οποίων οι τιμές διατίθενται από ένα σύνολο παρατηρήσεων. Η μία μεταβλητή που θεωρείται ως ανεξάρτητη, συμβολίζεται με Χ και η άλλη που θεωρείται ως εξαρτημένη, συμβολίζεται με Υ. Π.χ., Χ={ο αριθμός των παιδιών μιας οικογένειας}, Υ={η δαπάνη για αγορά γάλακτος}.

Στα προβλήματα παλινδρόμησης ακολουθούνται δύο τελείως διαφορετικές μεταξύ τους διαδικασίες. Η πρώτη αποβλέπει στην αντίληψη της αιτίας που επιτρέπει στη μεταβλητή Χ να επιδρά στη μεταβλητή Υ, δηλαδή διερευνάται εάν και τι είδους σχέση αναπτύσσεται μεταξύ των δύο μεταβλητών. Η δεύτερη αποβλέπει στην πρόβλεψη τιμής για την εξαρτημένη μεταβλητή Υ μιας περίπτωσης από την τιμή που έχει η ανεξάρτητη μεταβλητή Χ για αυτήν. Στο παράδειγμα που αναφέρθηκε φαίνεται ότι οι δύο μεταβλητές συσχετίζονται, και μάλιστα θετικά (όσο περισσότερα είναι τα παιδιά τόσο μεγαλύτερη θα είναι η δαπάνη). Το είδος της σχέσης μεταξύ των μεταβλητών αποσαφηνίζεται με το διάγραμμα διασποράς των ζευγών (x,y) του συνόλου των διατιθεμένων περιπτώσεων.

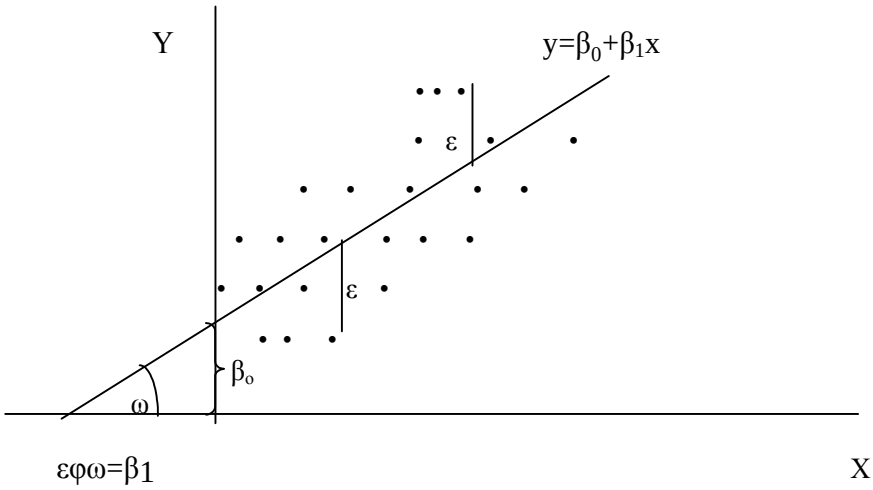
2.1 Το απλό γραμμικό μοντέλο[4]

Η μορφή του μοντέλου της απλής γραμμικής παλινδρόμησης είναι η

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad (\ i = 1,.., n\ ), \qquad (2.1)$$

όπου n είναι το μέγεθος του δείγματος, δηλαδή ο αριθμός των ζευγών (x_i, y_i) που έχουμε στη διάθεσή μας, y είναι η εξαρτημένη μεταβλητή, x είναι η ανεξάρτητη και β_0, β_1 είναι άγνωστες, αλλά σταθερές, παράμετροι. Ειδικά β_0 είναι η τομή της ευθείας με τον άξονα των Υ, ενώ β_1 είναι η κλίση της ίδιας ευθείας. Τέλος ε είναι το σφάλμα το οποίο μας δείχνει πόσο μακριά από την ευθεία βρίσκεται κάποια παρατήρηση όπως φαίνεται και στο Σχήμα 2.

[4] Βλέπε Βιβλ. Γραμμικά Μοντέλα Κεφάλαιο 2.2 Σελ. 8 έως 17



Σχήμα 2

Η εξίσωση (2.1), αποτελεί το μοντέλο μας για το συγκεκριμένο δείγμα και εκφράζει αυτό που εμείς πιστεύουμε για την σχέση που υπάρχει μεταξύ των y και x.

2.2 Εκτίμηση των παραμέτρων

Για την εκτίμηση των παραμέτρων β_0 και β_1 του μοντέλου (2.1) έχουν προταθεί κατά καιρούς, διάφορες μέθοδοι. Εμείς εδώ θα χρησιμοποιήσουμε την μέθοδο των ελαχίστων τετραγώνων. Σύμφωνα με τη μέθοδο αυτή οι εκτιμητές των παραμέτρων εκλέγονται κατά τέτοιο τρόπο, ώστε το άθροισμα τετραγώνων των σφαλμάτων ε_i (i=1,2,...,n) να γίνεται ελάχιστο. Αν δηλαδή θέσουμε:

$S = \sum\varepsilon^2_j$ για i =1 έως n τότε, από την (2.1) έχουμε $S = \sum (y_i -\beta_0-\beta_1x_i)^2$ για i =1 έως n.

Οι εκτιμητές ελαχίστων τετραγώνων β_0, β_1 των παραμέτρων β_0, β_1 θα προκύψουν από τις εξισώσεις

$$\frac{\partial S}{\partial \beta_o}=0 \quad \Rightarrow \quad \sum_{i=1}^{n}(y_i-\hat{\beta}_o-\hat{\beta}_1x_i)=0 \ ,$$

$$\frac{\partial S}{\partial \beta_1}=0 \quad \Rightarrow \quad \sum_{i=1}^{n}x_i(y_i-\hat{\beta}_o-\hat{\beta}_1x_i)=0 \ ,$$

όπου έχουμε αντικαταστήσει (β_0,β_1) με $(\hat{\beta}_0, \hat{\beta}_1)$. Από τις τελευταίες εξισώσεις παίρνουμε

$$\hat{\beta}_o n+\hat{\beta}_1\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i \qquad (2.2\alpha)$$

$$\hat{\beta}_o \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \qquad (2.2\beta)$$

Οι εξισώσεις (2.2α, β) ονομάζονται *κανονικές εξισώσεις* και η λύση τους μας δίνει

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \Big/ n}{\sum_{i=1}^{n} x_j^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \qquad (2.3)$$

και $$\hat{\beta}_o = \overline{y} - \hat{\beta}_1 \overline{x} \qquad (2.4)$$

όπου $\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ και $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$.

'Εχοντας τους εκτιμητές β_0, β_1 η εκτιμώμενη εξίσωση της απλής γραμμικής παλινδρόμησης είναι η

$$\hat{y}_i = \hat{\beta}_o + \hat{\beta}_1 x_i \quad \Leftrightarrow \quad \hat{y}_i = \overline{y} + \hat{\beta}_1 (x_i - \overline{x}) \qquad (2.5)$$

Η εξίσωση (2.5) είναι το μοντέλο που ζητάμε να βρούμε.

'Ενα σπουδαίο ρόλο στην μελέτη της γραμμικής παλινδρόμησης, γενικώς, παίζουν τα υπόλοιπα (residuals) $e_i = y_i - \hat{y}_i \quad (i = 1, 2, \ldots, n)$. Από την (2.5) παίρνουμε ότι

$$y_i - \hat{y}_i = (y_i - \overline{y}) - \hat{\beta}_1 (x_i - \overline{x}) \quad \Rightarrow \quad \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} (y_i - \overline{y}) - \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \overline{x}) = 0 \quad .$$

Δηλαδή το άθροισμα των υπολοίπων ισούται με μηδέν. Η ιδιότητα αυτή των υπολοίπων ισχύει, για κάθε μοντέλο γραμμικής παλινδρόμησης και οφείλεται στην ύπαρξη του σταθερού όρου β_0 στο μοντέλο. Η παράλειψη του β_0 από ένα μοντέλο συνεπάγεται ότι η μέση τιμή της εξαρτημένης μεταβλητής γίνεται μηδέν όταν, οι τιμές όλων των ανεξάρτητων μεταβλητών γίνουν μηδέν. Αυτό όμως είναι μια πολύ ισχυρή υπόθεση, η οποία συνήθως δεν ισχύει. Στο σημείο αυτό θα πρέπει να τονισθεί, ότι η απαλοιφή του β_0, σε ένα μοντέλο γραμμικής παλινδρόμησης, είναι πάντοτε δυνατή αν κεντράρουμε τα δεδομένα. Για παράδειγμα το μοντέλο (2.1) μπορεί να γραφτεί σαν

$y_i - \overline{y} = \beta_o + \beta_1 \overline{x} - \overline{y} + \beta_1 (x_i - \overline{x}) + \varepsilon_i, \ (i = 1, 2, \ldots, n)$ ή $y_i' = \beta_o' + \beta_1 x_i' + \varepsilon_i$, όπου $y_i' = y_i - \overline{y}$, $\beta_o' = \beta_o + \beta_1 \overline{x} - \overline{y}$ και $x_i' = x_i - \overline{x}$. Ο εκτιμητής ελαχίστων τετραγώνων του β_1 θα δίνεται από την (2.3), ενώ από την (2.4) παίρνουμε $\hat{\beta}_o' = \overline{y}' - \hat{\beta}_1' \overline{x}' = 0$ επειδή $\overline{x}' = \overline{y}' = 0$, για οποιοδήποτε τιμή του β_1. Συνεπώς, επειδή η τελευταία σχέση ισχύει πάντοτε, το μοντέλο (2.1) μπορεί να γραφεί στην μορφή

$$y_i - \bar{y} = \beta_1 (x_i - \bar{x}) + \varepsilon_i, \quad (i = 1,2,...,n).$$

2.3 Ανάλυση διακύμανσης

Είναι γνωστό ότι ένα από τα μέτρα μεταβλητότητας ενός συνόλου δεδομένων είναι και η (δειγματική) διακύμανση. Κατά συνέπεια το άθροισμα τετραγώνων $\sum_{i=1}^{n} (y_i - \bar{y})^2$ εκφράζει την (ολική) μεταβλητότητα των παρατηρήσεων y_i (i=1,2,...,n). Την μεταβλητότητα αυτή μπορούμε να την χωρίσουμε σε δύο επιμέρους αθροίσματα τετραγώνων. Για το σκοπό αυτό θεωρούμε την ταυτότητα: $y_i - \hat{y}_i = y_i - \bar{y} - (\hat{y}_i - \bar{y})$.Παίρνοντας τα τετράγωνα αμφοτέρων των μελών και αθροίζοντας για i = 1 έως n έχουμε:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 - 2\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{y}) \quad . \qquad (2.6)$$

Αλλά $\sum_{i=1}^{n} (y_i - \bar{y})(\hat{y}_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$, λόγω της (2.5)

$= \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$, λόγω της (2.3)

$= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$, λόγω της (2.5)

'Αρα η (2.6) γράφεται σαν

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (2.7)$$

Στην τελευταία σχέση, επειδή $\sum \hat{y}_i = \sum y_i$, η ποσότητα $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ είναι το άθροισμα τετραγώνων που οφείλεται στην παλινδρόμηση, ενώ $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ είναι το άθροισμα τετραγώνων των υπολοίπων . Συνεπώς η σχέση (2.7) μπορεί να γραφεί με λόγια ως εξής:

Ολική μεταβλητότητα των y	=	**Άθροισμα τετραγώνων παλινδρόμησης**	+	**Άθροισμα τετραγώνων υπολοίπων**

ή συμβολικά $SS_{tot} = SS_{reg} + SS_{res}$. (2.8)

Σε κάθε άθροισμα τετραγώνων αντιστοιχούμε έναν αριθμό ο οποίος ονομάζεται βαθμός ή βαθμοί

ελευθερίας. Ο αριθμός αυτός μας δείχνει τον αριθμό των ανεξάρτητων πληροφοριών, των σχετιζόμενων με τις ανεξάρτητες τιμές y_1, y_2, ..., y_n, οι οποίες είναι αναγκαίες για τον υπολογισμό του εν λόγω αθροίσματος τετραγώνων.

Για παράδειγμα, για τον υπολογισμό του $\sum_{i=1}^{n}(y_i-\bar{y})^2$ χρειαζόμαστε (n-1) ανεξάρτητες πληροφορίες, επειδή από τις τιμές $y_1-\bar{y}$, $y_2-\bar{y}$, …,$y_n-\bar{y}$ μόνο οι (n-1) είναι ανεξάρτητες λόγω του ότι $\sum_{i=1}^{n}(y_i-\bar{y})=0$. Επίσης, το άθροισμα τετραγώνων που οφείλεται στην παλινδρόμηση μπορεί να υπολογιστεί από μία μόνο συνάρτηση των y_1, y_2, ..., y_n, την $\hat{\beta}_1$, επειδή $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2=\hat{\beta}_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2$.

Συνεπώς το άθροισμα αυτό έχει ένα βαθμό ελευθερίας. Κατά συνέπεια, αφαιρώντας, το άθροισμα τετραγώνων των υπολοίπων έχει (n-2) βαθμούς ελευθερίας. Δηλαδή, κατ' αντιστοιχία με την (2.8), έχουμε χωρίσει τους βαθμούς ελευθερίας ως

$$(n-1) = (1) + (n-2) \qquad (2.9)$$

βαθμοί ελευθερίας SS_{tot} = αριθμός παρατηρήσεων (μέγεθος δείγματος) -1.
βαθμοί ελευθερίας SS_{reg} = αριθμός ανεξάρτητων μεταβλητών στο μοντέλο.
βαθμοί ελευθερίας SS_{res} = αριθμός παρατηρήσεων - αριθμός παραμέτρων μοντέλου (συμπεριλαμβανομένου του β_0).

Από την σχέση (2.9) γίνεται φανερό ότι όταν ξέρουμε τους βαθμούς ελευθερίας για οποιαδήποτε δύο από τα παραπάνω αθροίσματα, μπορούμε να βρούμε τους βαθμούς ελευθερίας για το τρίτο άθροισμα τετραγώνων.

2.4 Συντελεστής προσδιορισμού, R^2

Αν στη σχέση (2.8) διαιρέσουμε τα μέλη με το SS_{tot} παίρνουμε

$SS_{reg} / SS_{tot} = 1 – SS_{res} / SS_{tot}$.

Το αριστερό σκέλος στην παραπάνω ισότητα είναι μεταξύ 0 και 1 και εκφράζει το ποσοστό της μεταβλητότητας της μεταβλητής Y, που εξηγείται από την παλινδρόμηση, δηλαδή από την ανεξάρτητη μεταβλητή X. Το ποσοστό αυτό το ονομάζουμε συντελεστή προσδιορισμού, και το συμβολίζουμε με R^2. Δηλαδή

$$R^2 = SS_{reg} / SS_{tot} = 1 – SS_{res} / SS_{tot} \qquad (2.10)$$

Το R^2, όπως το ορίσαμε, έχει μερικές "καλές" ιδιότητες

I. Μπορεί να υπολογισθεί εύκολα .
II. Είναι καθαρός αριθμός. Δεν εξαρτάται δηλαδή από τις μονάδες των μετρήσεων.
III. Μπορεί να γενικευθεί εύκολα στον πολλαπλό συντελεστή συσχέτισης .

Η χρήση του συντελεστή προσδιορισμού R^2 για την μέτρηση της ερμηνευτικής ικανότητας του υποδείγματος έχει ένα μειονέκτημα. Αν το δείγμα των παρατηρήσεων είναι αρκετά μικρό, τότε αυξάνοντας τον αριθμό των ερμηνευτικών μεταβλητών η τιμή του συντελεστή R^2 μπορεί να αυξηθεί σημαντικά και, μάλιστα, να τείνει προς τη μονάδα εικονικά. Αυτό μπορεί να συμβαίνει χωρίς αναγκαστικά το υπόδειγμα να έχει μεγάλη ερμηνευτική ικανότητα και οφείλεται στη μείωση των βαθμών ελευθερίας στην εκτίμηση του υποδείγματος, που συνεπάγεται η αύξηση των παραμέτρων του υποδείγματος όταν ο αριθμός των παρατηρήσεων του δείγματος είναι σχετικά μικρός και δεν αυξάνει με τον αριθμό των παραμέτρων. Από τον ορισμό του συντελεστή R^2, μπορούμε να διαπιστώσουμε ότι μια μείωση των βαθμών ελευθερίας μπορεί να προκαλέσει σημαντική μείωση του RSS σχετικά με το TSS, που συνεπάγεται ότι ο συντελεστής R^2 να αυξηθεί σημαντικά. Για την αποφυγή του προβλήματος αυτού, έχει προταθεί η χρήση του διορθωμένου (προσαρμοσμένου) συντελεστή R^2 (adjusted-R^2), λαμβάνοντας υπόψη τους βαθμούς ελευθερίας.

2.5 Ιδιότητες των εκτιμητών ελαχίστων τετραγώνων

Από την παράγραφο 2.3 γίνεται φανερό ότι για την εύρεση των εκτιμητών ελαχίστων τετραγώνων δεν χρειάζεται καμία απολύτως υπόθεση πιθανοθεωρητικής φύσης. Για να μελετήσουμε όμως τις ιδιότητες των εν λόγω εκτιμητών χρειαζόμαστε τις παρακάτω βασικές υποθέσεις, σχετικά με το μοντέλο (2.1).

1) Τα σφάλματα ε_i είναι τυχαίες μεταβλητές με μέση τιμή μηδέν και κοινή διακύμανση σ^2. Δηλαδή $E(\varepsilon_i) = 0$ και $Var(\varepsilon_i|x_i) = \sigma^2$ για κάθε i =1, 2,..., n. Η διακύμανση σ^2 θεωρείται σταθερή αλλά άγνωστη.

2) Τα σφάλματα ε_i είναι ασυσχέτιστα μεταξύ τους, δηλαδή $Cov(\varepsilon_i,\varepsilon_j)=0$ για κάθε $i \neq j=1, 2,..., n$. Είναι γνωστό ότι η έννοια της συσχέτισης αναφέρεται σε δύο τυχαίες μεταβλητές. Η συσχέτιση αφορά την ίδια μεταβλητή, τα σφάλματα, για διαφορετικές παρατηρήσεις. Η συσχέτιση αυτού του τύπου είναι γνωστή ως *αυτοσυσχέτιση*.

Σαν συνέπεια των υποθέσεων (1) και (2) έχουμε, από την (2.1), ότι

$$E(y_i) = \beta_0 + \beta_1 x_i,\ Var(y_i) = \sigma^2$$

καθώς επίσης και ότι οι μετρήσεις y_i και y_j, $i \neq j$, είναι ασυσχέτιστες.

Οι παραπάνω υποθέσεις μας χρειάζονται για να υπολογίσουμε την αναμενόμενη τιμή και την διακύμανση των εκτιμητών $\hat{\beta}_o$ και $\hat{\beta}_1$.

Μια τρίτη υπόθεση η οποία χρειάζεται τόσο για τον έλεγχο υποθέσεων σχετικά με τα β_0 και β_1, αλλά και για τον έλεγχο άλλων υποθέσεων είναι η ακόλουθη:

3) Τα σφάλματα ε_i έχουν κανονική κατανομή. Από την πρώτη των υποθέσεων παίρνουμε ότι $\varepsilon_i \sim N(0, \sigma^2)$.

Με την προσθήκη της (3) τα σφάλματα ε_i, ε_j και κατά συνέπεια τα y_i και y_j γίνονται ανεξάρτητα. Με την βοήθεια των υποθέσεων (1) και (2) και από τις σχέσεις (2.3) και (2.4) μπορούμε να αποδείξουμε τα εξής θεωρήματα:

Θεώρημα 1- Για το μοντέλο (2.1) οι εκτιμητές ελαχίστων τετραγώνων $\hat{\beta}_o$, $\hat{\beta}_1$ είναι αμερόληπτοι εκτιμητές. Δηλαδή

$$E\left(\hat{\beta}_o\right)=\beta_0 \text{ και } E\left(\hat{\beta}_1\right)=\beta_1$$

Θεώρημα 2- Οι διακυμάνσεις των εκτιμητών $\hat{\beta}_o$ και $\hat{\beta}_1$ δίνονται από τις σχέσεις:

$$Var(\hat{\beta}_1)=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad \text{και} \quad Var(\hat{\beta}_0)=\frac{\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\frac{\sigma^2}{n} \qquad (2.11)$$

Θεώρημα 3- Η κατανομή των εκτιμητών ελαχίστων τετραγώνων $\hat{\beta}_o$ και $\hat{\beta}_1$ είναι η κανονική. (Οι μέσες τιμές και οι διακυμάνσεις των κανονικών κατανομών δίνονται από τα θεωρήματα 1 και 2 αντίστοιχα).

2.6 Εκτίμηση του σ^2

Στην παράγραφο 2.5 είδαμε ότι μία από τις υποθέσεις μας για τα σφάλματα ήταν η $Var(\varepsilon_i)=\sigma^2$, (i=1, 2,..., n). Επειδή το σ^2 είναι η $E(\varepsilon^2_i)$ θα πρέπει να περιμένουμε ότι ο εκτιμητής του, $\hat{\sigma}^2$, θα δίνεται από την αναμενόμενη τιμή του SS_{res}. Πράγματι, με τη βοήθεια της πρώτης και δεύτερης υπόθεσης της παραγράφου 2.5 μπορούμε να αποδείξουμε το

Θεώρημα 4- Το μέσο τετράγωνο των υπολοίπων είναι ένας αμερόληπτος εκτιμητής του σ^2. Δηλ.

$\hat{\sigma}^2=MS_{res}$ με $E(MS_{res}) = \sigma^2$

Από το προηγούμενο θεώρημα γίνεται αμέσως φανερό ότι ο εκτιμητής $\hat{\sigma}^2$ εξαρτάται από το μοντέλο που προσαρμόζουμε στα συγκεκριμένα δεδομένα και συνεπώς από την ορθότητα ή μη του εν λόγω μοντέλου. Επειδή δε κάθε μοντέλο συνοδεύεται από τις υποθέσεις (1) και (2) της 2.5 συνεπάγεται ότι όταν έστω και μία από αυτές δεν ισχύει τότε $\hat{\sigma}^2$ γίνεται ένας "κακός" εκτιμητής του σ^2 . Για παράδειγμα όταν τα σφάλματα ε δεν είναι μόνο τυχαίες μεταβλητές αλλά εμπεριέχουν και σταθερές ποσότητες τότε ο $\hat{\sigma}^2$ υπερεκτιμά το σ^2. Στην περίπτωση αυτή λέμε ότι έχουμε έλλειψη προσαρμογής (lack of fit). Το ιδανικό θα ήταν να μπορούσαμε να βρούμε ένα εκτιμητή του σ^2, ο οποίος να μην εξαρτάται από το προσαρμοζόμενο μοντέλο. Γενικώς ένας τέτοιος εκτιμητής μπορεί να επιτευχθεί όταν έχουμε επαναλήψεις στα δεδομένα μας, δηλαδή όταν για την ίδια τιμή της X έχουμε πολλές τιμές της Y ή όταν έχουμε εκ των προτέρων κάποια πληροφορία για την διακύμανση σ^2.
Στην πραγματικότητα ο εκτιμητής $\hat{\sigma}^2=MS_{res}$ είναι ένας αμερόληπτος εκτιμητής της διακύμανσης των παρατηρήσεων περί την παλινδρόμηση, την οποία συμβολίζουμε με σ^2_{YX}. Το σ^2 μπορεί να είναι ή να μην είναι ίσο με το σ^2_{YX}. Εάν το μοντέλο που προσαρμόζουμε είναι το

σωστό τότε: $\sigma^2 = \sigma^2_{YX}$

Κάνοντας χρήση της τρίτης υπόθεσης της 2.5 μπορούμε να αποδείξουμε ότι

$$(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2} \qquad (2.12)$$

2.7 Διαστήματα εμπιστοσύνης και έλεγχος υποθέσεων

Όταν τα σφάλματα ε_i στο μοντέλο (2.1) ακολουθούν την κανονική κατανομή τότε τόσο οι εκτιμητές $\hat{\beta}_o$ και $\hat{\beta}_1$ των παραμέτρων β_0 και β_1, όσο και οι προβλεπόμενες τιμές $\hat{y}_i$ ακολουθούν κανονική κατανομή. Αυτό ισχύει επειδή όλες αυτές οι ποσότητες είναι γραμμικές συναρτήσεις των παρατηρήσεων y_i και συνεπώς των ε_i. Το γεγονός αυτό μας επιτρέπει την εύρεση διαστημάτων εμπιστοσύνης και τον έλεγχο υποθέσεων για τις ποσότητες.

Διάστημα εμπιστοσύνης και έλεγχος υποθέσεων για το β_0.

Από το Θεώρημα 3 σε συνδυασμό με τα Θεωρήματα 1 και 2 παίρνουμε ότι

$$\hat{\beta}_o \sim N\left(\beta_o, \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}\sigma^2\right).$$

Εκτιμώντας το σ^2 με το $\hat{\sigma}^2$ και κάνοντας χρήση της σχέσης (2.12) έχουμε ότι ένα 100(1-α)% διάστημα εμπιστοσύνης για το β_0 είναι το

$$\left(\hat{\beta}_o - t_{\alpha/2;\ n-2}\sqrt{\hat{Var}(\hat{\beta}_o)}\ ,\ \hat{\beta}_o + t_{\alpha/2;\ n-2}\sqrt{\hat{Var}(\hat{\beta}_o)}\right) \qquad (2.13)$$

όπου $\hat{Var}(\hat{\beta}_0)$ είναι η εκτιμώμενη διακύμανση του $\hat{\beta}_o$, α είναι το επίπεδο σημαντικότητας και $t_{\alpha/2,\ n-2}$ είναι το α / 2 εκατοστιαίο σημείο της t_{n-2}-κατανομής δηλαδή το σημείο εκείνο για το οποίο $P(t_{n-2} \geq t_{\alpha/2,\ n-2})$.
Για τον έλεγχο της υπόθεσης

$H_0 : \beta_0 = \beta^*_0$ ως προς την $H_a : \beta_0 \neq \beta^*_0$

όπου β^*_0 είναι μια οποιαδήποτε τιμή του β_0, χρησιμοποιούμε το στατιστικό

$$t = \frac{\hat{\beta}_o - \beta^*_o}{\sqrt{\hat{Var}(\hat{\beta}_o)}} \qquad (2.14)$$

και απορρίπτουμε την H_0 αν $|t| \geq t_{\alpha/2,\, n-2}$.

Διάστημα εμπιστοσύνης και έλεγχος υποθέσεων για το β_1

Όπως και προηγουμένως έτσι και τώρα από το Θεώρημα 3 παίρνουμε ότι

$$\hat{\beta}_1 \sim N\left(\beta_1,\ \sigma^2 \Big/ \sum_{i=1}^{n} (x_i - \overline{x})^2 \right) .$$

Άρα ένα 100(1-α)% διάστημα εμπιστοσύνης για το β_1 είναι το

$$\left(\hat{\beta}_1 - t_{\alpha/2;\ n-2} \sqrt{\hat{Var}(\hat{\beta}_1)} \ ,\ \ \hat{\beta}_1 + t_{\alpha/2;\ n-2} \sqrt{\hat{Var}(\hat{\beta}_1)} \right). \quad (2.15)$$

Για τον έλεγχο της υπόθεσης

$$H_0 : \beta_1 = \beta^*_1 \text{ ως προς την } H_a : \beta_1 \neq \beta^*_1$$

όπου β^*_1 είναι μια οποιαδήποτε τιμή του β_1, χρησιμοποιούμε το στατιστικό

$$t = \frac{\hat{\beta}_1 - \beta^*_1}{\sqrt{\hat{Var}(\hat{\beta}_1)}}$$

και απορρίπτουμε την H_0 αν $|t| \geq t_{\alpha/2,\, n-2}$.

2.8 Συσχέτιση μεταξύ X και Y[5]

Όπως έχουμε αναφέρει το μοντέλο της γραμμικής παλινδρόμησης είναι μία σχέση προς μία μόνο κατεύθυνση. Σε μια τέτοια σχέση οι ανεξάρτητες μεταβλητές υποτίθεται ότι είναι σταθερές και χωρίς σφάλμα. Η υπόθεση ότι οι ανεξάρτητες μεταβλητές είναι σταθερές μπορεί να "χαλαρώσει" αρκεί να καθορίσουμε εμείς ποια είναι η εξαρτημένη και ποιες οι ανεξάρτητες μεταβλητές. Και η άλλη υπόθεση ότι οι ανεξάρτητες μεταβλητές μετρούνται χωρίς σφάλμα μπορεί να "χαλαρώσει", όμως η ανάλυση τότε δυσκολεύει και γι' αυτό δεν θα ασχοληθούμε με την περίπτωση αυτή.
Στην περίπτωση λοιπόν που και η X, εκτός από την Y, είναι τυχαία μεταβλητή τότε το μοντέλο μας $y = \beta_0 + \beta_1 x + \varepsilon$ και όσα έχουμε αναφέρει μέχρι τώρα σχετικά με αυτό ισχύουν με την προϋπόθεση ότι θέλουμε να μελετήσουμε την επίδραση της τυχαίας μεταβλητής X πάνω στην τυχαία μεταβλητή Y. Αν ρ(X, Y) είναι ο συντελεστής συσχετίσεως μεταξύ των τυχαίων μεταβλητών X και Y τότε ένας εκτιμητής αυτού, βασισμένος σε ένα τυχαίο δείγμα $(x_1,y_1),...,(x_n,y_n)$ μεγέθους n, είναι ο δειγματικός συντελεστής συσχετίσεως του Pearson

[5] Βλέπε Βιβλ. Γραμμικά Μοντέλα Κεφάλαιο 2.2 Σελ. 19 έως 21.

$$r(x, y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad . \quad (2.16)$$

Συγκρίνοντας την (2.3) με την (2.16) παίρνουμε

$$\hat{\beta}_1=\left\{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right\}^{1/2} r(x, y), \qquad \text{ή} \qquad \hat{\beta}_1=\frac{S_X}{S_Y}r(x, y)$$

όπου $(n-1)S_X^2=\sum_{i=1}^{n}(x_i-\bar{x})^2$ και $(n-1)S_Y^2=\sum_{i=1}^{n}(y_i-\bar{y})^2$.

Ακόμη μπορούμε να δούμε ότι για την απλή γραμμική παλινδρόμηση (και μόνο γι' αυτή)

$$R^2 = r^2.$$

Για τον έλεγχο της υπόθεσης $H_0 : \rho = 0$ ως προς μία από τις εναλλακτικές $H_0 : \rho>0$, $H_\alpha : \rho<0$, $H_\alpha : \rho \neq 0$ χρησιμοποιούμε το στατιστικό

$$t = r\frac{\sqrt{n-2}}{\sqrt{1-r^2}},$$

του οποίου η κατανομή είναι η t_{n-2} όταν η από κοινού κατανομή των X και Y είναι η διδιάστατη κανονική . Δηλαδή η υπόθεση H_0 απορρίπτεται όταν $|t| \geq t_{\alpha/2, n-2}$. Για τον έλεγχο της $H_0 : \rho = 0$ η υπόθεση της διδιάστατης κανονικής μπορεί να αντικατασταθεί από την υπόθεση τουλάχιστον μία από τις Y και X να ακολουθεί κανονική κατανομή.

2.9 Ερμηνεία της $\hat{y}=\hat{\beta}_o+\hat{\beta}_1 x$

Όπως έχουμε αναφέρει στα προηγούμενα, μετά τον υπολογισμό των εκτιμητών ελαχίστων τετραγώνων $\hat{\beta}_o$ και $\hat{\beta}_1$ παραμέτρων β_0 και β_1 στο μοντέλο $y = \beta_0 + \beta_1 x + \varepsilon$, φθάνουμε στο εκτιμώμενο μοντέλο $\hat{y}=\hat{\beta}_o+\hat{\beta}_1 x$. Υποθέτοντας ότι το εκτιμώμενο μοντέλο είναι "σωστό" μπορούμε να πούμε ότι το $\hat{\beta}_1$ μας μετράει την μεταβολή της y όταν η x αυξάνει κατά μια μονάδα.

2.10 Έλλειψη προσαρμογής[6]

Αναφέραμε ότι ένα στατιστικό μοντέλο είναι γενικώς προσωρινό. Ακόμη, είναι προφανές, ότι σκοπός της ανάλυσης είναι η εύρεση ενός σωστού μοντέλου. Σε μερικές περιπτώσεις μπορούμε να ελέγξουμε αν ένα μοντέλο είναι σωστό ή όχι. Πριν όμως δώσουμε την τεχνική για τον έλεγχο αυτό, ας δούμε ποιές είναι οι συνέπειες ενός μη σωστού μοντέλου. Ξέρουμε ότι το υπόλοιπο για $X=x_i$ είναι $e_i = y - \hat{y}_i$. Αν τώρα $E(Y_i)=v_i$ είναι η τιμή της $E(Y_i)$ την οποία παίρνουμε από το σωστό μοντέλο, οποιοδήποτε και αν είναι αυτό, για $X=x_i$, τότε μπορούμε να γράψουμε:

$$\begin{aligned} e_i &= y_i - \hat{y}_i = (y_i - \hat{y}_i) - E(y_i - \hat{y}_i) + E(y_i - \hat{y}_i) \\ &= \{(y_i - \hat{y}_i) - [v_i - E(\hat{y}_i)]\} + v_i - E(\hat{y}_i) \\ &= \quad q_i \quad + \quad B_i \end{aligned}$$

όπου $q_i = \{(y_i - \hat{y}_i) - (v_i - E(\hat{y}_i))\}$ και $B_i = v_i - E(\hat{y}_i)$.

Η ποσότητα B_i είναι το σφάλμα μεροληψίας για $X=x_i$. Αν το μοντέλο είναι σωστό τότε $E(\hat{y}_i)=v_i$ και συνεπώς $B_i=0$. Αν το μοντέλο δεν είναι σωστό τότε $E(\hat{y}_i) \neq v_i$ και συνεπώς $B_{ij} \neq 0$. Στην περίπτωση αυτή η τιμή του B_{ij} εξαρτάται από το σωστό μοντέλο και την τιμή x_i.

Από τον ορισμό των q_i εύκολα μπορούμε να δούμε ότι τα q_i είναι οι τιμές μιας τυχαίας μεταβλητής με μέση τιμή μηδέν, και αυτό ισχύει ανεξάρτητα αν το μοντέλο είναι σωστό ή όχι δηλαδή ανεξάρτητα αν $E(\hat{y}_i)=v_i$ ή όχι. Ακόμη, μπορούμε να δείξουμε ότι τα q_i είναι συσχετισμένα μεταξύ τους και ότι

$$E(\sum q^2_i) = (n-2)*\sigma^2 \quad \text{για } i=1 \text{ εως } n$$

Συνεπώς $E(MS_{res}) = \sigma^2$, αν το μοντέλο μας είναι σωστό ή $E(MS_{res}) = \sigma^2 + 1/(n-2)*\sum B^2_i$ για i=1 εως n αν το μοντέλο μας δεν είναι σωστό.

Όταν λοιπόν το μοντέλο μας δεν είναι σωστό τότε (i) τα υπόλοιπα μπορούν να εκφρασθούν σαν άθροισμα δύο ποσοτήτων μιας τυχαίας (q_i) και μιας μη τυχαίας ή συστηματικής (B_i) και (ii) το MS_{res} δεν είναι ένας αμερόληπτος εκτιμητής για το σ^2. Για τον έλεγχο της ορθότητας του μοντέλου διακρίνουμε δύο περιπτώσεις.

Η πρώτη είναι η περίπτωση όπου ένας εκτιμητής $\hat{\sigma}^2_1$, του σ^2, είναι γνωστός εκ των προτέρων.

Όταν αυτό συμβαίνει, τότε με ένα F-τεστ μπορούμε να ελέγξουμε την υπόθεση $H_o: \sigma^2 = \hat{\sigma}^2_1$. Η απόρριψη της H_0 συνεπάγεται την μη ορθότητα του μοντέλου μας, ενώ η μη απόρριψή της σημαίνει ότι δεν έχουμε αρκετές ενδείξεις για να πούμε ότι το μοντέλο μας δεν είναι σωστό.

Η δεύτερη είναι η περίπτωση όπου στα δεδομένα μας έχουμε επαναληπτικές μετρήσεις. Με τη φράση επαναληπτικές μετρήσεις εννοούμε ότι στα δεδομένα μας, για μία τουλάχιστον τιμή της X, υπάρχουν δύο ή περισσότερες τιμές της Y. Κάνοντας χρήση αυτών των επαναλαμβανόμενων μετρήσεων μπορούμε να βρούμε έναν εκτιμητή του σ^2. Ένας τέτοιος εκτιμητής λέμε ότι παριστάνει το γνήσιο σφάλμα, διότι όταν οι τιμές της X είναι ταυτόσημες για δύο παρατηρήσεις, τότε μόνο τυχαίες διακυμάνσεις μπορούν να επηρεάσουν το αποτέλεσμα και να δώσουν διαφορές μεταξύ των εν λόγω παρατηρήσεων. Αυτές οι διαφορές μας δίνουν ένα εκτιμητή του σ^2 ο οποίος είναι αρκετά καλύτερος από οποιονδήποτε άλλο εκτιμητή. Γι' αυτό ακριβώς το λόγο θα πρέπει, όταν σχεδιάζουμε ένα πείραμα, να το σχεδιάζουμε κατά τέτοιο τρόπο ώστε να εμφανίζονται επαναληπτικές μετρήσεις.

---

[6] Βλέπε Βιβλ. Γραμμικά Μοντέλα Κεφάλαιο 2.2 Σελ. 23 έως 25.

ΚΕΦΑΛΑΙΟ 3 ΠΟΛΛΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ[7]

Στην παρούσα παράγραφο θα αναφέρουμε μερικά αποτελέσματα για την πολλαπλή γραμμική παλινδρόμηση. Τα αποτελέσματα αυτά, στην ουσία, είναι γενικεύσεις των αποτελεσμάτων της απλής γραμμικής παλινδρόμησης .

3.1 Περιγραφή των δεδομένων και του μοντέλου

Στην πολλαπλή γραμμική παλινδρόμηση έχουμε μία εξαρτημένη μεταβλητή, την y, και p ανεξάρτητες μεταβλητές, τις x_1, x_2, ..., x_p. Συνεπώς η μορφή των δεδομένων για n παρατηρήσεις παίρνει τη μορφή ενός n x (p+1) πίνακα ως εξής:

Μορφή δεδομένων για την πολλαπλή γραμμική παλινδρόμηση							
Αριθμός Παρατήρησης		Τιμές					
		y	x_1	x_2	x_3	…	x_p
1		y_1	x_{11}	x_{12}	x_{13}	…	x_{1p}
2		y_2	x_{21}	x_{22}	x_{23}	…	x_{2p}
3		y_3	x_{31}	x_{32}	x_{33}	…	x_{3p}
.		.	.	.	.	.	.
.		.	.	.	.	.	.
.		.	.	.	.	.	.
n		y_n	x_{n1}	x_{n2}	x_{n3}	…	x_{np}

Πίνακας 2

όπου x_{ij} είναι η τιμή της j (j=1,2,...,p) μεταβλητής για την i (i=1,2,...,n) παρατήρηση. Το γραμμικό μοντέλο που συνδέει την εξαρτημένη με τις ανεξάρτητες μεταβλητές είναι το

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots. + \beta_p x_{ip} + \varepsilon_i, \quad (i=1,2,...,n) \quad (3.1)$$

όπου τα β_0, β_1, β_2, ..., β_p ονομάζονται μερικοί συντελεστές παλινδρόμησης, και είναι άγνωστες,

[7] Βλέπε Βιβλ. Γραμμικά Μοντέλα Κεφάλαιο 2.3 Σελ. 30 έως 32.

αλλά σταθερές παράμετροι προς εκτίμηση. Τα ε_i (i=1, 2, ..., n) είναι όπως και στην απλή γραμμική παλινδρόμηση. Στο σημείο αυτό θα πρέπει να τονίσουμε ότι η γραμμικότητα του μοντέλου (3.1) νοείται ως προς τις άγνωστες παραμέτρους. Αν p=1 τότε η (3.1) παίρνει την μορφή της (2.1) ενώ αν p=2 τότε η (3.1) παριστάνει ένα διδιάστατο επίπεδο στον τρισδιάστατο χώρο (y, x_1, x_2).

3.2 Εκτίμηση των παραμέτρων

Οι εκτιμητές ελαχίστων τετραγώνων $\hat{\beta}_0$, $\hat{\beta}_1$,..., $\hat{\beta}_p$ των παραμέτρων β_0, β_1, ..., β_p θα προκύψουν από την ελαχιστοποίηση της παράστασης

$$S=\sum_{i=1}^{n}\varepsilon_i^2=\sum_{i=1}^{n}(y_i-\beta_o-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip})^2$$

ως προς β_0, β_1, ..., β_p. Από τις εξισώσεις $\partial S/\partial\beta_i = 0$ (i=0,1,2,...,p) προκύπτει ότι οι $\hat{\beta}_0$, $\hat{\beta}_1$,..., $\hat{\beta}_p$ ικανοποιούν τις κανονικές εξισώσεις

$$\begin{aligned}
S_{11}\hat{\beta}_1+S_{12}\hat{\beta}_2+\cdots+S_{ip}\hat{\beta}_p&=S_{Y1}\\
S_{12}\hat{\beta}_1+S_{22}\hat{\beta}_2+\cdots+S_{2p}\hat{\beta}_p&=S_{Y2}\\
&\cdots\cdots\cdots\cdots\\
S_{1p}\hat{\beta}_1+S_{2p}\hat{\beta}+\cdots+S_{pp}\hat{\beta}_p&=S_{Yp}
\end{aligned}$$

όπου $S_{ij}=\sum_{k=1}^{n}(x_{ki}-\bar{x}_i)(x_{kj}-\bar{x}_j),\quad i,\ j=1,\ 2,\ \ldots,\ p$

$$S_{Yi}=\sum_{k=1}^{n}(y_k-\bar{y})(x_{ki}-\bar{x}_i)\qquad i=1,\ 2,\ \ldots,\ p,$$

$$\bar{x}_i=\sum_{k=1}^{n}x_{ki}/n\ ,\qquad \bar{y}=\sum_{k=1}^{n}y_k/n$$

και $\hat{\beta}_o=\bar{y}-\hat{\beta}_1\bar{x}_1-\hat{\beta}_2\bar{x}_2-\cdots-\hat{\beta}_p\bar{x}_p$. (3.2)

Υποθέτοντας ότι το σύστημα των εξισώσεων έχει λύση, μπορούμε να βρούμε τα $\hat{\beta}_0$, $\hat{\beta}_1$,..., $\hat{\beta}_p$, και συνεπώς η εκτιμώμενη παλινδρόμηση θα είναι

$\hat{y}_i=\hat{\beta}_o+\hat{\beta}_1x_{i1}+\hat{\beta}_2x_{i2}+\cdots+\hat{\beta}_px_{ip}$. (3.3)

Τέλος, τα υπόλοιπα e_i ορίζονται όπως και στην απλή γραμμική παλινδρόμηση δηλαδή

$e_i = y_i - \hat{y}_i$.

3.3 Ιδιότητες των εκτιμητών[8] $\hat{\boldsymbol{\beta}}$

Εάν υποθέσουμε ότι στο μοντέλο (3.1) τα σφάλματα ε_i είναι ασυσχέτιστες τυχαίες μεταβλητές με $E(\varepsilon_i) = 0$ και $Var(\varepsilon_i) = \sigma^2$ σταθερή αλλά άγνωστη τότε

1) $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$

2) $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

3) Αν $\mathbf{x_0} = (1, x_{o1}, x_{o2}, ..., x_{op})$ τότε η προβλεπόμενη τιμή του y, $\hat{y}_0$ στο σημείο $\mathbf{x_0}$ θα είναι

$\hat{y}_o = \mathbf{x}_o'\hat{\boldsymbol{\beta}}$ και $\mathrm{Var}(\hat{y}_o)=\sigma^2\mathbf{x}_o'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_o$

4) Ένας αμερόληπτος εκτιμητής του σ^2 είναι το $\hat{\sigma}^2 = MS_{res}$ δηλ.

$$\mathbf{E(\hat{\sigma}^2)=E(MS_{res})=\sigma^2} \ .$$

5) Το τετράγωνο του πολλαπλού συντελεστή συσχετίσεως R ορίζεται σαν

$$R^2 = SS_{reg} / SS_{tot} = 1 - SS_{res} / SS_{tot}.$$

Από τον ορισμό του R^2 προκύπτει ότι $0 \le R^2 \le 1$. Η τιμή του R^2 μας δίνει το ποσοστό της ολικής μεταβλητότητας που εξηγείται από την ύπαρξη της παλινδρόμησης. Εδώ θα πρέπει να τονισθεί ότι ενώ "μικρές" τιμές του R^2 δείχνουν γενικώς ένα "φτωχό" μοντέλο αντίθετα "μεγάλες" τιμές δεν συνεπάγονται κατ' ανάγκην ένα "σωστό" μοντέλο. (Η τιμή του R^2 μπορεί πάντοτε να γίνει ίση με τη μονάδα.)

Αν για τα σφάλματα ε_i υποθέσουμε ακόμη ότι $\varepsilon_i \sim N(0,\sigma^2)$, τότε για τους εκτιμητές $\hat{\boldsymbol{\beta}}$ μπορούμε να αποδείξουμε:

1) Οι εκτιμητές ελαχίστων τετραγώνων είναι αμερόληπτοι εκτιμητές των παραμέτρων του μοντέλου με ελάχιστη διακύμανση.

2) Οι εκτιμητές ελαχίστων τετραγώνων $\hat{\boldsymbol{\beta}}$ ακολουθούν μία p+1-διάστατη κανονική κατανομή με μέση τιμή $\boldsymbol{\beta}$ και πίνακα διακυμάνσεων-συνδιακυμάνσεων $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ δηλ.

$\hat{\boldsymbol{\beta}} \sim \mathrm{N}_{p+1}(\boldsymbol{\beta},\ \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.

3) Οι εκτιμητές ελαχίστων τετραγώνων συμπίπτουν με τους εκτιμητές μεγίστης πιθανοφάνειας.

4) Η τυχαία μεταβλητή $w = SS_{res} / \sigma^2$ έχει την χ^2-κατανομή με n-p-1 βαθμούς ελευθερίας.

5) Οι τυχαίες μεταβλητές $\hat{\boldsymbol{\beta}}$ και $\hat{\sigma}^2 = MS_{res}$ είναι ανεξάρτητες μεταξύ τους.

6) Ο έλεγχος της υπόθεσης $H_o : \beta_i = b_i$ (i=0,1,2,...,p), όπου b_i είναι μια γνωστή τιμή, γίνεται με το στατιστικό,

$$\mathbf{t}= \frac{\hat{\boldsymbol{\beta}}_i - \mathbf{b}_i}{\hat{\sigma}\sqrt{\mathbf{c}_{i+1,\ i+1}}},$$

όπου $c_{i+1,\ i+1}$ είναι το (i+1,i+1) στοιχείο του πίνακα $(\mathbf{X}'\mathbf{X})^{-1}$, και $\hat{\sigma}=\sqrt{\hat{\sigma}^2}=\sqrt{MS_{res}}$.

Αν $|t| \ge t_{\alpha/2,\ n-p-1}$ τότε H_0 απορρίπτεται.

7) Ο έλεγχος της υπόθεσης $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$, ότι δηλ. δεν υπάρχει παλινδρόμηση γίνεται με το F-τεστ, $F = MS_{reg} / MS_{res}$, και η H_0 απορρίπτεται όταν $F \ge F_{\alpha,\ p,n-p-1}$.

---

[8] Βλέπε Βιβλ. Γραμμικά Μοντέλα Κεφάλαιο 2.3 Σελ. 33 έως 39.

8) Ένα 100(1-α) % διάστημα εμπιστοσύνης για το διάνυσμα **β** των παραμέτρων είναι το

$$(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \le (p+1)MS_{res}F_{a;\ p+1,\ n-p-1} \quad .$$

9) Ο έλεγχος για την έλλειψη προσαρμογής του μοντέλου μας γίνεται με το F_{lof} -τεστ, $F_{lof} = MS_{lof} / MS_{pe}$.

10) Αν $\hat{y}_o$ είναι η πρόβλεψή μας στο σημείο $\mathbf{x_0}$ τότε ένα 100(1-α) % διάστημα εμπιστοσύνης για την $E(y_0)$ είναι το $\hat{y} \pm t_{\alpha/2;\ n-p-1}\sqrt{\hat{\sigma}^2 x_o'(X'X)^{-1}x_o}$.

3.4 Έλεγχος υποθέσεων

Εκτός από τον έλεγχο υποθέσεων, για τις παραμέτρους β_i (i=0,1,2,...,p) του μοντέλου (3.1), που είδαμε διάφορες άλλες υποθέσεις μπορούν να προκύψουν κατά την ανάλυση δεδομένων με τη βοήθεια γραμμικών μοντέλων. Οι πιο συχνές υποθέσεις για έλεγχο είναι:

1) Ενα υποσύνολο των β_i (i=0,1,2,...,p) ισούται με μηδέν.
2) Ενα υποσύνολο των β_i (i=0,1,2,...,p) είναι ίσα μεταξύ τους.
3) Ενα υποσύνολο των β_i (i=0,1,2,...,p) ικανοποιεί κάποια γραμμική σχέση.

Όλες οι παραπάνω υποθέσεις και τυχόν άλλες μπορούν να ελεγχθούν με τη βοήθεια μιας γενικευμένης θεωρίας η οποία είναι γνωστή ως αξίωμα του έξτρα αθροίσματος τετραγώνων.

Το μοντέλο θα αναφέρεται στο εξής σαν το πλήρες μοντέλο (ΠΜ). Η προς έλεγχο υπόθεση H_0 γενικώς, δίνει τιμές σε κάποιες από τις παραμέτρους. Αν αντικαταστήσουμε τις τιμές αυτές στο πλήρες μοντέλο, τότε λαμβάνουμε ένα καινούριο μοντέλο, το οποίο είναι το μοντέλο όταν η υπόθεση H_0 ισχύει και το συμβολίζουμε ως H_0 μοντέλο. Ο αριθμός των προς εκτίμηση παραμέτρων στο H_0 μοντέλο είναι μικρότερος από τον αριθμό των προς εκτίμηση παραμέτρων στο πλήρες μοντέλο.

Έστω ότι $\hat{y}_i$ και $\hat{y}_i^*$ είναι αντιστοίχως οι προβλεπόμενες τιμές στο πλήρες μοντέλο και στο H_0 μοντέλο. Τότε το άθροισμα τετραγώνων των υπολοίπων για το πλήρες μοντέλο θα είναι

$$SS_{res}(\Pi M) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

ενώ το αντίστοιχο άθροισμα για το μοντέλο που προκύπτει όταν η H_0 ισχύει είναι

$$SS_{res}(H_oM) = \sum_{i=1}^{n}(y_i - \hat{y}_i^*)^2 \quad .$$

Στο πλήρες μοντέλο (3.1) έχουμε p+1 παραμέτρους. Αν υποθέσουμε ότι στο μοντέλο που προκύπτει όταν η H_0 ισχύει έχουμε και διαφορετικές παραμέτρους, τότε μπορούμε να δείξουμε ότι

$$w_1 = \{SS_{res}(H_0M) - SS_{res}(\Pi M)\}/\sigma^2 \sim \chi^2_{p+1-\kappa}.$$

Τέλος από την ιδιότητα (4) καθώς και από το γεγονός ότι το $w=SS_{res}(\Pi M)/\sigma^2$ είναι ανεξάρτητο από το w_1,ο έλεγχος της H_0 μπορεί να γίνει με το πηλίκο,

$$F = \frac{(SS_{res}(H_oM) - SS_{res}(\Pi M))/(p+1-k)}{SS_{res}(\Pi M)/(n-p-1)} \quad . \qquad (3.4)$$

Αν $F \ge F_{\alpha,\ p+1-k,\ n-p-1}$ τότε η προς έλεγχο υπόθεση H_0 απορρίπτεται.

Η εξαγωγή του F-τεστ βασίζεται εξ ολοκλήρου στην υπόθεση ότι $\varepsilon_i \sim N(0,\sigma^2)$. Στην περίπτωση που η υπόθεση αυτή δεν ισχύει τότε μια ένδειξη για την ορθότητα ή μη της προς έλεγχο υπόθεσης H_0 μπορεί να γίνει με τα δύο παρακάτω κριτήρια:

Κριτήριο 1) *Το τετράγωνο του συντελεστή πολλαπλής συσχετίσεως, R^2.*

Ξέρουμε ότι $R^2 = SS_{reg} / SS_{tot}$. Επίσης όσο μεγαλύτερη είναι η τιμή του R^2 τόσο καλύτερα το προσαρμοζόμενο μοντέλο εξηγεί τα δεδομένα. Μπορούμε λοιπόν να συγκρίνουμε τις τιμές του R^2 για το H_0 μοντέλο και για το πλήρες μοντέλο. Αν η αύξηση του R^2 από το H_0 μοντέλο στο πλήρες μοντέλο κρίνεται σαν αξιόλογη τότε η προς έλεγχο υπόθεση απορρίπτεται.

Κριτήριο 2) *Η τετραγωνική ρίζα* $\hat{\sigma} = \sqrt{MS_{res}}$

Το μέσο τετράγωνο των υπολοίπων είναι ένας αμερόληπτος εκτιμητής του σ^2. Συνεπώς όσο μικρότερο είναι το $\hat{\sigma}$ τόσο το καλύτερο, με την έννοια ότι τόσο καλύτερες θα είναι οι προβλέψεις μας. Αν λοιπόν η ελάττωση της τιμής του $\hat{\sigma}$ πηγαίνοντας από το H_0 μοντέλο στο πλήρες μοντέλο κρίνεται σαν αξιόλογη, τότε η προς έλεγχο υπόθεση H_0 απορρίπτεται.

3.5 Ερμηνεία των εκτιμητών $\hat{\boldsymbol{\beta}}$

Όπως και στην απλή γραμμική παλινδρόμηση ο εκτιμητής $\hat{\beta}_i$ του β_i (i=1,2,...,p) μας εκφράζει την μεταβολή του Y στην μοναδιαία μεταβολή της τιμής της X_i όταν όλες οι άλλες ανεξάρτητες μεταβλητές είναι σταθερές.
Το πρόσημο + ή - πριν από τον εκτιμητή μας δείχνει την διεύθυνση της παραπάνω μεταβολής.

3.6. Μερικός συντελεστής συσχετίσεως

Η ανάλυση μερικής συσχέτισης διαχειρίζεται τα δεδομένα με τέτοιο τρόπο ώστε να είναι δυνατός ο προσδιορισμός της σχέσης μεταξύ της εξαρτημένης μεταβλητής Y (αποτέλεσμα y) και μιας ανεξάρτητης μεταβλητής X (επίδραση x), με την υπόθεση ότι δεν υπάρχει άλλη μεταβλητή που να επιδρά και αυτή συγχρόνως (και στην εξαρτημένη και στην ανεξάρτητη). Υποτίθεται, δηλαδή, ότι εάν υπάρχει άλλη μεταβλητή αυτή δεν επιδρά αλλά θεωρείται σταθερή ή ανενεργή (μη επιδρούσα).

Ας υποθέσουμε ότι σε κάποια έρευνα ενδιαφερόμαστε για τις τιμές των μεταβλητών: X_1 = βάρος, X_2 = ύψος και X_3 = ηλικία. Είναι σε όλους γνωστό ότι τόσο το ύψος, όσο και το βάρος αυξάνουν όσο η ηλικία αυξάνει. Συνεπώς, ο υπολογισμός της τιμής του δειγματικού συντελεστή συσχέτισης του Pearson μεταξύ των μεταβλητών βάρος και ύψος δεν θα ανταποκρίνεται στην πραγματικότητα. Για να βρούμε μια πιο αξιόπιστη τιμή για την συσχέτιση των δύο αυτών μεταβλητών θα πρέπει οι τιμές της τρίτης μεταβλητής, της ηλικίας, να είναι σταθερές (ή να μην διαφέρουν πολύ μεταξύ τους). Η επιλογή ενός τέτοιου δείγματος αφ' ενός μεν δεν είναι πάντοτε δυνατή, αφ' ετέρου δε δεν είναι απαραίτητο. Για το λόγο αυτό εισάγεται η έννοια του μερικού συντελεστού συσχετίσεως, ο οποίος για την περίπτωση μας ορίζεται ως

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{1 - r_{13}^2} \quad \sqrt{1 - r_{23}^2}} \qquad (3.5)$$

όπου $r_{ij} = r(x_i, x_j)$ (i,j,=1,2,3; $i \neq j$) και $r_{12,3}$ είναι ο μερικός συντελεστής συσχετίσεως μεταξύ των τυχαίων μεταβλητών X_1 και X_2 όταν η X_3 παραμένει σταθερή. Τύποι ανάλογοι με τον (3.5) ισχύουν και για τους $r_{13.2}$ και $r_{23.1}$. Γενικότερα, εάν έχουμε $X_1,X_2,...,X_p$ τυχαίες μεταβλητές, ο μερικός συντελεστής συσχετίσεως μεταξύ των μεταβλητών X_i, X_j, όταν οι υπόλοιπες μεταβλητές παραμένουν σταθερές, ισούται με τον συντελεστή συσχετίσεως των υπολοίπων, της παλινδρόμησης της x_i ως προς τις υπόλοιπες (εκτός της x_j) ανεξάρτητες μεταβλητές, και των υπολοίπων, της παλινδρόμησης της x_j ως προς τις υπόλοιπες (εκτός της x_i) ανεξάρτητες μεταβλητές.

3.7 Υποθέσεις για τις ανεξάρτητες μεταβλητές $X_1, X_2,..., X_p$

Μέχρι τώρα, τόσο για την απλή γραμμική παλινδρόμηση όσο και για την πολλαπλή, δεν έχουμε κάνει καμία άλλη υπόθεση εκτός των υποθέσεων για τα σφάλματα ε_i
Υπάρχουν όμως δύο υποθέσεις οι οποίες αφορούν τις ανεξάρτητες μεταβλητές $X_1,X_2,...,X_p$.
Υπόθεση 1) Οι ανεξάρτητες μεταβλητές δεν είναι τυχαίες μεταβλητές. Δηλαδή οι τιμές τους είναι σταθερές ή μπορούν να προκαθορισθούν.
Υπόθεση 2) Οι τιμές των ανεξάρτητων μεταβλητών δεν υπόκεινται σε σφάλμα.

Οι υποθέσεις αυτές μολονότι δεν επηρεάζουν τόσο πολύ τα μαθηματικά αποτελέσματα, εν τούτοις επηρεάζουν τα συμπεράσματα τα οποία προκύπτουν από την ανάλυση παλινδρόμησης.
Η πρώτη υπόθεση ικανοποιείται όταν ο πειραματιστής μπορεί να προκαθορίσει τις τιμές των X_i (i=1,2,...,p) σε κάποια επιθυμητά επίπεδα. Γίνεται φανερό ότι η υπόθεση αυτή δύσκολα ικανοποιείται ιδίως όταν ο αριθμός των ανεξάρτητων μεταβλητών είναι μεγάλος. Η μη ικανοποίηση της υπόθεσης 1 έχει σαν αποτέλεσμα τον περιορισμό της εγκυρότητας των αποτελεσμάτων μας. Πιο συγκεκριμένα, στην περίπτωση αυτή, τα αποτελέσματά μας ισχύουν μόνο για τα εν λόγω δεδομένα.
Η δεύτερη, από τις παραπάνω υποθέσεις, είναι πολύ δύσκολο (αν όχι αδύνατο) να ικανοποιηθεί. Η ύπαρξη σφάλματος στις μετρήσεις μας θα επηρεάσει την διακύμανση των υπολοίπων, τον πολλαπλό συντελεστή συσχετίσεως, R, καθώς και τους εκτιμητές των παραμέτρων του μοντέλου. Το ακριβές ποσοστό αυτής της επίδρασης εξαρτάται από πολλούς παράγοντες, οι σπουδαιότεροι από τους οποίους είναι (i) η τυπική απόκλιση του σφάλματος των μετρήσεων και (ii) η μορφή της συσχέτισης μεταξύ τους. Το αποτέλεσμα της ύπαρξης σφάλματος στις μετρήσεις είναι η αύξηση της διακύμανσης των υπολοίπων και η ελάττωση της τιμής του πολλαπλού συντελεστή συσχέτισης. Η επίδραση του εν λόγω σφάλματος στους εκτιμητές των παραμέτρων του μοντέλου είναι πιο δύσκολο να προσδιορισθεί καθ' ότι ο εκτιμητής π.χ. του β_i (i=1,2,...,p) επηρεάζεται όχι μόνο από το σφάλμα των μετρήσεων στην μεταβλητή X_i, αλλά και από το σφάλμα των μετρήσεων σε όλες τις ανεξάρτητες μεταβλητές. Αν η διακύμανση του σφάλματος των μετρήσεων είναι μικρότερη από την διακύμανση των σφαλμάτων ε του μοντέλου, τότε η ύπαρξη σφάλματος στις μετρήσεις είναι ασήμαντη. Ακόμη όμως και σ' αυτή την περίπτωση θα πρέπει να είμαστε προσεκτικοί όταν ερμηνεύουμε τους συντελεστές στο εκτιμώμενο μοντέλο.

ΚΕΦΑΛΑΙΟ 4 ΑΝΑΛΥΣΗ ΤΩΝ ΥΠΟΛΟΙΠΩΝ[9]

Η ανάλυση των υπολοίπων είναι ένα από τα βασικότερα εργαλεία όχι μόνο για την γραμμική παλινδρόμηση αλλά γενικότερα για κάθε μοντέλο παλινδρόμησης (γραμμικής ή μη) και ανάλυσης διακύμανσης.

Η σπουδαιότητα των υπολοίπων έγκειται στη δυνατότητα που μας παρέχουν να ελέγχουμε αφ' ενός μεν τις τρεις υποθέσεις για τα σφάλματα ε_i δηλαδή i) $Cov(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$ ii) $Var(\varepsilon_i|x_i) = \sigma^2$ σταθερή αλλά άγνωστη, iii) $\varepsilon_i \sim N(0,\sigma^2)$, αφ' ετέρου δε την ορθότητα ή μη του χρησιμοποιούμενου μοντέλου.

Για το μοντέλο της πολλαπλής γραμμικής παλινδρόμησης, ως γνωστόν, τα υπόλοιπα ορίζονται σαν $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$. Συνεπώς

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{I_n} - \mathbf{P})\mathbf{y} \quad (4.1)$$

όπου $\mathbf{P}=\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Αν τώρα στην τελευταία σχέση αντικαταστήσουμε $\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ παίρνουμε

$$\mathbf{e} = (\mathbf{I_n}-\mathbf{P})\boldsymbol{\varepsilon} . \quad (4.2)$$

Οι σχέσεις (4.1) και (4.2) μας δείχνουν ότι τα υπόλοιπα μπορούν να εκφρασθούν σαν γραμμικές σχέσεις είτε των παρατηρήσεων y_i είτε των άγνωστων σφαλμάτων ε_i. Συνεπώς μπορούν να χρησιμοποιηθούν για τον έλεγχο της ορθότητας ή μη του μοντέλου, και για τον έλεγχο των υποθέσεων για τα σφάλματα.

Από την σχέση (4.2) παίρνουμε ότι $E(\mathbf{e}) = \mathbf{0}$ και $Var(\mathbf{e}) = (\mathbf{I_n}-\mathbf{P})\sigma^2$. Άρα η διακύμανση ενός συγκεκριμένου υπολοίπου, e_i θα είναι $Var(e_i)=(1-p_{ii})\sigma^2$, ενώ η συνδιακύμανση μεταξύ δύο οποιονδήποτε υπολοίπων, e_i και e_j, θα είναι $Cov(e_i, e_j)= -p_{ij}\sigma^2$, όπου p_{ij} είναι το (i,j) στοιχείο του πίνακα **P**.

Εκτός από τα υπόλοιπα $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, τα οποία θα τα ονομάσουμε αρχικά υπόλοιπα ή απλώς υπόλοιπα, υπάρχουν ακόμη άλλοι δύο τύποι υπολοίπων.

i) Τα τυποποιημένα υπόλοιπα (standardized residuals) σε απομίμηση της τυπικής κανονικής κατανομής. Ως γνωστόν αν $X \sim N(0,\sigma^2)$ τότε $X/\sigma \sim N(0,1)$. Συνεπώς τα τυποποιημένα υπόλοιπα ορίζονται σαν $e_{si} = e_i/\hat{\sigma}^2$, όπου $\hat{\sigma}^2=MS_{res}$. Από την τελευταία σχέση γίνεται φανερό ότι τα τυποποιημένα υπόλοιπα δεν ακολουθούν τυπική κανονική κατανομή, επειδή ο παρανομαστής στη σχέση ορισμού τους δεν είναι η τυπική απόκλιση του e_i.

ii) Τα μαθητικοποιημένα υπόλοιπα (Studentized residuals) είναι τα αρχικά υπόλοιπα διαιρεμένα με την εκτιμώμενη τυπική τους απόκλιση. Επειδή όταν $\varepsilon_i \sim N(0,\sigma^2)$, τότε $e_i \sim N(0,Var(e_i))$ με $Var(e_i) = (1-p_{ii})\sigma^2$, τα μαθητικοποιημένα υπόλοιπα ορίζονται σαν

$$\mathbf{t_i} = \frac{\mathbf{e_i}}{\hat{\sigma}\sqrt{\mathbf{1-p_{ii}}}} .$$

[9] Βλέπε Βιβλ. Γραμμικά Μοντέλα Κεφάλαιο 2.4 Σελ. 44 έως 49.

Τόσο τα μαθητικοποιημένα υπόλοιπα, όσο και τα τυποποιημένα δεν ακολουθούν τυπική κανονική κατανομή, αλλά την t κατανομή με n-p-1 βαθμούς ελευθερίας.

Από τη σχέση (4.1) παίρνουμε :

$$\text{Cov}(\mathbf{e}, \hat{\mathbf{y}}) = \text{Cov}[(\mathbf{I_n}-\mathbf{P})\mathbf{y}, \mathbf{X}(\hat{\boldsymbol{\beta}})]$$

$$= \text{Cov}[(\mathbf{I_n}-\mathbf{P})\mathbf{y}, \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]$$

$$= \text{Cov}[(\mathbf{I_n}-\mathbf{P})\mathbf{y}, \mathbf{P}\mathbf{y}]$$

$$= (\mathbf{I_n}-\mathbf{P})\text{Var}(\mathbf{y})\mathbf{P} = \sigma^2(\mathbf{I_n}-\mathbf{P})\mathbf{P} = \sigma^2(\mathbf{P}-\mathbf{P}^2) = \mathbf{0}.$$

Βλέπουμε δηλαδή ότι τα e_i και $\hat{y}_i$ είναι ασυσχέτιστες τυχαίες μεταβλητές. Η ιδιότητα αυτή θα παίξει βασικό ρόλο στην ανάλυση των υπολοίπων.

Για την ανάλυση των υπολοίπων χρησιμοποιούμε τόσο γραφικές όσο και στατιστικές μεθόδους.

Από τις δύο αυτές μεθόδους οι διάφορες γραφικές παραστάσεις των υπολοίπων αποτελούν το βασικότερο εργαλείο για την διάγνωση λαθών στην ανάλυση δεδομένων.

Από τη σχέση $V(\mathbf{e}) = (\mathbf{I_n}-\mathbf{P})\boldsymbol{\varepsilon}$ βλέπουμε ότι τα υπόλοιπα **e** είναι συσχετισμένες τυχαίες μεταβλητές. Η συσχέτιση που υπάρχει μεταξύ των υπολοίπων δεν επηρεάζει τις γραφικές παραστάσεις των υπολοίπων, εκτός αν ο λόγος (n-p-1)/n δηλ. (βαθμοί ελευθερίας των υπολοίπων)/(αριθμός υπολοίπων) είναι πολύ μικρός.

4.1 Έλεγχος των υποθέσεων για τα σφάλματα

Στην παράγραφο αυτή θα αναπτύξουμε στατιστικές μεθόδους για τον έλεγχο της ορθότητας ή μη των τριών υποθέσεων που αναφέρονται στα σφάλματα ε_i ενός μοντέλου. Λόγω των σχέσεων (4.1) και (4.2) ο έλεγχος των εν λόγω υποθέσεων στηρίζεται στην μελέτη των υπολοίπων είτε με τη βοήθεια γραφικών μεθόδων είτε με τη βοήθεια στατιστικών τεστ.

Έλεγχος ύπαρξης ασυσχέτιστων σφαλμάτων

Η υπόθεση ότι τα σφάλματα είναι ασυσχέτιστα σημαίνει ότι η τιμή του σφάλματος ε_i, για την i-στη παρατήρηση, δεν εξαρτάται από την τιμή του σφάλματος για οποιαδήποτε άλλη παρατήρηση. Επειδή η συσχέτιση, στην παρούσα περίπτωση, αναφέρεται σε μία μεταβλητή, την ε, αντί για δύο, όπως γνωρίζουμε, γι' αυτό είναι καλύτερα να χρησιμοποιούμε τον όρο *αυτοσυσχέτιση* αντί του όρου συσχέτιση.

Βασικό στοιχείο για τον έλεγχο ύπαρξης αυτοσυσχέτισης είναι να γνωρίζουμε την χρονολογική σειρά των παρατηρήσεών μας. Στην περίπτωση αυτή μπορούμε να αντικαταστήσουμε τον δείκτη i τόσο στο μοντέλο της απλής γραμμικής παλινδρόμησης, όσο και σε αυτό της πολλαπλής, με τον δείκτη t. Έτσι π. χ. αντί για $y_i=\beta_0+\beta_1x_i+\varepsilon_i$ θα έχουμε $y_t=\beta_0+\beta_1x_t+\varepsilon_t$.

Τεστ των ροών : Με δεδομένο ότι η χρονολογική σειρά των παρατηρήσεων είναι γνωστή, τότε ο έλεγχος ότι τα σφάλματα είναι ασυσχέτιστα γίνεται με το *τεστ των ροών*. Το τεστ των ροών συνίσταται στην εξέταση της διάταξης των προσήμων (+ ή -) των υπολοίπων.

Ας υποθέσουμε ότι έχουμε n υπόλοιπα εκ των οποίων n_1 είναι θετικά, $n_2 = n-n_1$ αρνητικά και έστω ότι η χρονολογική σειρά των παρατηρήσεων μας έδωσε την εξής ακολουθία:

+ + - + - - - + + - + + + .

Για να εξετάσουμε αν η ακολουθία αυτή παρουσιάζει κάποια συστηματική απόκλιση κάνουμε το εξής: μετράμε των αριθμό u των "ροών" που εμφανίζονται στην ακολουθία, όπου σαν ροή νοείται κάθε συνεχής ακολουθία ομοίων συμβόλων, όπως φαίνεται παρακάτω (κάθε ακολουθία μέσα σε παρένθεση είναι μία ροή)

(+ +) (-) (+) (- - -) (+ +) (-) (+ + +) .

Στην προκειμένη περίπτωση n=13, n_1=8, n_2=5 και u=7. Προφανώς το u είναι μία τυχαία μεταβλητή. Η μέση τιμή και διακύμανση της u είναι

$\mu = 2n_1*n_2 / n_1+n_2 +1$ και $\sigma^2 = 2n_1n_2*(2n_1n_2-n_1-n_2) / (n_1+n_2)^2*(n_1+n_2-1)$.

Το επόμενο βήμα είναι να εξετάσουμε κατά πόσο η ακολουθία των προσήμων των υπολοίπων είναι ή δεν είναι τυχαία, ούτως ώστε να συμπεράνουμε ότι τα σφάλματα είναι ασυσχέτιστα ή συσχετισμένα. Η εξέταση αυτή για μικρά n_1 και n_2 (π.χ. $n_1 \leq 20$ και $n_2 \leq 20$) γίνεται με την βοήθεια πινάκων ενώ για μεγάλα n_1 και n_2 (π.χ. $n_1>20$ και $n_2>20$) κάνουμε χρήση του τύπου

$$Z = u-\mu + ½ \ / \sigma .$$

Η κατανομή του παραπάνω τύπου είναι, ασυμπτωτικά, η N(0,1). Συνεπώς η υπόθεση ότι τα σφάλματα είναι ασυσχέτιστα απορρίπτεται όταν $|Z|\geq Z_{\alpha/2}$ όπου $Z_{\alpha/2}$ είναι το α/2 εκατοστιαίο σημείο της τυπικής κανονικής κατανομής.

Τεστ των Durbin – Watson : Ένα από τα πιο γνωστά τεστ για τον έλεγχο ύπαρξης ή μη αυτοσυσχέτισης μεταξύ των σφαλμάτων ε του μοντέλου μας είναι το τεστ των Durbin – Watson. Σύμφωνα με αυτό υποθέτουμε ότι τα σφάλματα ε στο μοντέλο μας είναι της μορφής

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t,$$

όπου ρ ($0\leq\rho\leq1$) είναι ένας συντελεστής προς εκτίμηση και u_t είναι μια τυχαία μεταβλητή για την οποία υποθέτουμε ότι ακολουθεί την $N(0,\sigma^2_u)$ κατανομή. Η αυτοσυσχέτιση αυτή ονομάζετε αυτοσυσχέτιση $1^{ου}$ βαθμού.

Ο έλεγχος της υπόθεσης H_0:ρ=0 ως προς την $H_α$:ρ>0 γίνεται με το στατιστικό των Durbin-Watson

$$\mathbf{d} = \frac{\sum_{t=2}^{n} (\mathbf{e_t} - \mathbf{e_{t-1}})^2}{\sum_{t=1}^{n} \mathbf{e_t^2}}$$

όπου e_t και e_{t-1} είναι, αντίστοιχα, τα υπόλοιπα στην χρονική στιγμή t και t-1.Η αποδοχή ή η απόρριψη της H_0 γίνεται με την βοήθεια πινάκων. Έτσι η H_0 απορρίπτεται αν $d < d_L$, ενώ η H_0 δεν μπορεί να απορριφθεί αν $d > d_U$. Σε περίπτωση που $d_L<d<d_U$, τότε το τεστ είναι αναποφάσιστο.

Αν και αρνητική συσχέτιση είναι δύσκολο να συναντήσουμε στην πράξη, εν τούτοις ο έλεγχος της υπόθεσης H_0:ρ=0 ως προς την $H_α$:ρ<0 γίνεται με την βοήθεια του στατιστικού d* =4-d .

Ένας εκτιμητής του ρ δίνεται από την σχέση

$$\hat{\rho} = \frac{\sum_{t=2}^{n} \mathbf{e_t e_{t-1}}}{\sum_{t=1}^{n} \mathbf{e_t^2}} .$$

Είναι εύκολο να δειχθεί ότι $d=2(1-\hat{\rho})$. Συνεπώς μια τιμή του d κοντά στο δύο (άρα το ρ είναι κοντά στο μηδέν) συνεπάγεται ότι δεν υπάρχει θετική αυτοσυσχέτιση $1^{ου}$ βαθμού μεταξύ των σφαλμάτων του μοντέλου μας. Αντίθετα μια d κοντά στο μηδέν (άρα το ρ είναι κοντά στο ένα) συνεπάγεται ότι υπάρχει θετική αυτοσυσχέτιση 1ου βαθμού μεταξύ των σφαλμάτων του μοντέλου μας.

Έλεγχος σταθερής διακύμανσης.[10]

Σε πολλά πρακτικά προβλήματα η υπόθεση ότι η διακύμανση είναι σταθερή για κάθε παρατήρηση (δηλ. $Var(\varepsilon_i)=\sigma^2$) δεν ισχύει. Κατά κανόνα η $Var(\varepsilon_i)$ εξαρτάται από το μέγεθος μιας άλλης μεταβλητής, συχνά της ανεξάρτητης.

Ο τύπος

$$T_{21} = \sum_{i=1}^{n} e_i^2 \hat{y}_i$$

μας παρέχει την δυνατότητα ελέγχου της υπόθεσης ότι η διακύμανση, σ^2, των σφαλμάτων, είναι σταθερή. Η αριθμητική τιμή του τύπου συγκρίνεται με τιμές από ειδικούς Πίνακες.

Έλεγχος κανονικότητας των σφαλμάτων.

Στα μοντέλα γραμμικής παλινδρόμησης μια από τις υποθέσεις που κάνουμε για τα σφάλματα είναι ότι ακολουθούν κανονική κατανομή. Η υπόθεση αυτή είναι απαραίτητη για την εύρεση διαστημάτων εμπιστοσύνης και για την εγκυρότητα των F και t τεστ.

Από την σχέση, $\mathbf{e}=(\mathbf{I}-\mathbf{P})\boldsymbol{\varepsilon}$ παίρνουμε ότι

$e_i = \varepsilon_i - \sum p_{ij}*\varepsilon_j$, για j =1 εως n.

Η σχέση αυτή μας λέει ότι όταν η τιμή του n-p (p = αριθμ. ανεξάρτητων μεταβλητών στο μοντέλο +1) είναι μικρή και μερικά από τα p_{ij} παίρνουν μεγάλες τιμές, τότε τα υπόλοιπα e_i τείνουν να συμπεριφέρονται όπως συμπεριφέρεται ο όρος $\sum p_{ij}*\varepsilon_j$, για j =1 εως n. Αυτό έχει σαν αποτέλεσμα τα υπόλοιπα να συμπεριφέρονται σαν κανονικές μεταβλητές ακόμη και όταν τα σφάλματα ε_i δεν ακολουθούν κανονική κατανομή. Συνεπώς, στην προκειμένη περίπτωση, οποιοδήποτε τεστ για τον έλεγχο της κανονικότητας των σφαλμάτων περιμένουμε να μας δίνει θετικά αποτελέσματα, και γι᾽ αυτό θα πρέπει να τα χρησιμοποιούμε με προσοχή. Οποιοδήποτε γνωστό τεστ π.χ. χ^2-κριτήριο προσαρμογής, κύρτωση, λοξότητα, μπορεί να χρησιμοποιηθεί για τον έλεγχο της κανονικότητας των σφαλμάτων με την προϋπόθεση ότι η τιμή του n-p είναι αρκετά μεγάλη. Η αξιοπιστία όμως αυτών των στατιστικών δεν κρίνεται ως ικανοποιητική (έχουν μικρή ισχύ).

Ένα πολύ γνωστό και αξιόπιστο τεστ είναι αυτό των Shapiro-Wilk (1965). Το τεστ αυτό είναι και το πιο συχνά χρησιμοποιούμενο στην πράξη.

Στο σημείο αυτό θα πρέπει να τονισθεί ότι, όλα τα γνωστά στατιστικά τεστ για τον έλεγχο της κανονικής κατανομής απαιτούν ανεξάρτητα δείγματα. Συνεπώς η χρήση αυτών για τον έλεγχο της κανονικότητας των υπολοίπων θα πρέπει να γίνεται με προσοχή επειδή όπως έχουμε αναφέρει τα υπόλοιπα δεν είναι ασυσχέτιστα. Διάφορες μελέτες προσομοίωσης που έχουν γίνει (π.χ. White και MacDonald (1980), Cook και Weisberg (1982), Pierce και Gray (1982)) έδειξαν ότι η συμπεριφορά των τεστ αυτών, όταν εφαρμόζονται στα υπόλοιπα μιας παλινδρόμησης, είναι ικανοποιητική, όσον αφορά το επίπεδο σημαντικότητας του τεστ, όταν n=20 και έχουμε τέσσερες έως έξι ανεξάρτητες μεταβλητές στο μοντέλο μας, ή n=40 και υπάρχουν οκτώ ανεξάρτητες μεταβλητές στο μοντέλο μας.

Αν η γραφική παράσταση των υπολοίπων δεν είναι ευθεία σε "ικανοποιητικό" βαθμό ή αν το στατιστικό τεστ μας παρέχει ενδείξεις ότι η υπόθεση της κανονικότητας των σφαλμάτων δεν ισχύει τότε δεν δεχόμαστε την εν λόγω υπόθεση.

[10] Βλέπε Βιβλ. Γραμμικά Μοντέλα Κεφάλαιο 2.4 Σελ. 49 έως 52.

4.2 Έλεγχος ορθότητας του μοντέλου[11]

Ξέρουμε πως μπορούμε να ελέγξουμε, με την βοήθεια ενός F-τεστ, την ορθότητα ή μη του μοντέλου μας, στην περίπτωση που στα δεδομένα μας υπάρχουν επαναληπτικές μετρήσεις. Η χρήση όμως του παραπάνω τεστ θα πρέπει να γίνεται με προσοχή καθ' όσον η εγκυρότητά του εξαρτάται από την κατανομή των σφαλμάτων του μοντέλου μας. Ακόμη όμως και στην περίπτωση που το F-τεστ ισχύει και είναι σημαντικό η μόνη πληροφορία που έχουμε είναι ότι το μοντέλο μας δεν είναι σωστό. Με άλλα λόγια το F-τεστ για τον έλεγχο προσαρμογής δεν μας λέει σε ποιο "σημείο" το μοντέλο μας είναι λάθος.

Έλεγχος ορθότητας του μοντέλου

Όλες οι σχέσεις μεταξύ μιας εξαρτημένης μεταβλητής και ενός αριθμού ανεξαρτήτων μεταβλητών δεν είναι κατ' ανάγκη γραμμικές. Αντίθετα θα πρέπει να περιμένουμε ότι, στην πράξη, μία γραμμική σχέση είναι μάλλον η εξαίρεση παρά ο κανόνας. Εν τούτοις η σπουδαιότητα των γραμμικών μοντέλων είναι πιο μεγάλη από ότι φαίνεται. Για παράδειγμα, ενώ η συναρτησιακή σχέση ίσως να μην είναι γραμμική σε όλο το εύρος των τιμών των ανεξάρτητων μεταβλητών, μπορεί να είναι γραμμική ή περίπου γραμμική σε υποδιαστήματα των τιμών των ανεξάρτητων μεταβλητών. Ακόμη, ένας κατάλληλος μετασχηματισμός των δεδομένων μπορεί να μετατρέψει ένα θεωρητικά μη γραμμικό μοντέλο σε γραμμικό.

Γίνεται φανερό ότι η γραφική παράσταση της εξαρτημένης Y ως προς την ανεξάρτητη μεταβλητή X, πρέπει να είναι ένα από τα πρώτα και βασικά βήματα για την ανάλυση δεδομένων με τη βοήθεια της απλής γραμμικής παλινδρόμησης. Αντίθετα, για τα μοντέλα της πολλαπλής γραμμικής παλινδρόμησης, οι γραφικές παραστάσεις της εξαρτημένης μεταβλητής Y ως προς κάθε μία από τις ανεξάρτητες μεταβλητές δεν είναι και τόσο σημαντικές, εκτός αν οι ανεξάρτητες μεταβλητές είναι στατιστικώς ανεξάρτητες μεταξύ τους. Η αδυναμία αυτή στα μοντέλα πολλαπλής γραμμικής παλινδρόμησης αντιμετωπίζεται με την εισαγωγή της έννοιας των μερικών υπολοίπων και τις γραφικές παραστάσεις αυτών. μοντέλο.

Εισαγωγή νέων όρων στο μοντέλο μας

Το ότι το μοντέλο μας περνάει τους έλεγχους γραμμικότητας δεν σημαίνει ότι είναι και σωστό. Πολλές φορές παρίσταται ανάγκη εισαγωγής νέων όρων στο μοντέλο μας, οι οποίοι μπορεί να είναι συναρτήσεις μιας ή περισσότερων ανεξάρτητων μεταβλητών που ήδη υπάρχουν στο μοντέλο μας, ή μιας ανεξάρτητης μεταβλητής που δεν υπάρχει στο μοντέλο (π.χ. ο χρόνος).

4.3 Ακραίες παρατηρήσεις

Γνώρισμα των ακραίων παρατηρήσεων είναι η μη προσαρμογή τους σε κάποιο μοντέλο που φαίνεται να προσαρμόζεται το κύριο σώμα των παρατηρήσεων. Συνεπώς η k- παρατήρηση

[11] Βλέπε Βιβλ. Γραμμικά Μοντέλα Κεφάλαιο 2.4 Σελ. 57 έως 62.

χαρακτηρίζεται σαν ακραία αν η $E(Y_\kappa)$ είναι σημαντικά διαφορετική από τη $\mathbf{x_\kappa\beta}$ ενώ για τις υπόλοιπες παρατηρήσεις έχουμε ότι $E(Y_i)\approx\mathbf{x_i\beta}$, $(i \neq k)$. Αυτό σημαίνει ότι το υπόλοιπο $\mathbf{e_k = \hat{y}_k - x_k\hat{\beta}}$, το οποίο είναι το δειγματικό ανάλογο του σφάλματος $\varepsilon_\kappa = y_\kappa - \mathbf{x_\kappa\beta}$, θα τείνει να διαφέρει αρκετά από το μηδέν. Κατά συνέπεια σαν υποψήφιες ακραίες παρατηρήσεις θα μπορούσαν να χαρακτηρισθούν παρατηρήσεις y_κ για τις οποίες η $|e_\kappa|$ είναι πολύ μεγάλη. Εν τούτοις επειδή τα υπόλοιπα e_i $(i=1,2,\ldots,n)$ είναι από μόνες τους τυχαίες μεταβλητές, μερικά από αυτά μπορούν να πάρουν μεγάλες τιμές χωρίς να είναι κατ' ανάγκην ακραίες τιμές, ενώ μια γνήσια ακραία παρατήρηση μπορεί να μην αντιστοιχεί στο υπόλοιπο με την μεγαλύτερη απόλυτη τιμή. Η ύπαρξη ακραίων παρατηρήσεων στα δεδομένα μας απαιτεί προσοχή στην ανάλυση τους για να προσδιορισθούν οι λόγοι που δίνουν τις μεγάλες αποκλίσεις μεταξύ των παρατηρούμενων (Y_i) και προβλεπόμενων $(\hat{y}_i)$ τιμών της εξαρτημένης μεταβλητής.

Αν οι ακραίες παρατηρήσεις είναι αποτέλεσμα λανθασμένων μετρήσεων ή κακής καταγραφής, τότε μπορούμε να παραλείψουμε τις εν λόγω παρατηρήσεις από το σύνολο των δεδομένων μας. Σε αντίθετη περίπτωση οι ακραίες παρατηρήσεις όχι μόνο δεν θα πρέπει να παραλειφθούν αλλά θα πρέπει να εξετασθούν πιο προσεκτικά και με μεγαλύτερη σχολαστικότητα διότι μπορεί να μας δίνουν πληροφορία, για το υπό εξέταση φαινόμενο ή πείραμα, που δεν μπορούν να μας δώσουν οι υπόλοιπες παρατηρήσεις.

Για την ανίχνευση και προσδιορισμό των ακραίων παρατηρήσεων χρησιμοποιούνται τόσο γραφικές όσο και στατιστικές μέθοδοι.

Η ύπαρξη ακραίων παρατηρήσεων στα δεδομένα μας περιπλέκει το πρόβλημα της ανάλυσης των εν λόγω δεδομένων και καθιστά τις τυχόν διορθώσεις μια αρκετά περίπλοκη υπόθεση. Γενικώς το πρώτο βήμα είναι να επαναλάβουμε την ανάλυση αφού πρώτα διαγράψουμε τις ακραίες παρατηρήσεις από το σύνολο των δεδομένων μας. Αν τα συμπεράσματά μας δεν αλλάζουν σημαντικά, τότε μπορούμε να αγνοήσουμε τις ακραίες παρατηρήσεις. Η παραπάνω πορεία γίνεται πιο αξιόλογη όταν ξέρουμε ότι το μοντέλο μας είναι σωστό. Αν όμως το μοντέλο μας δεν είναι σωστό, τότε η ύπαρξη ακραίων παρατηρήσεων μπορεί να σημαίνει ανάγκη για μετασχηματισμό των παρατηρήσεων.

Σαν συμπέρασμα από τα παραπάνω βγαίνει το ακόλουθο, στην περίπτωση που έχουμε ακραίες παρατηρήσεις. Είτε επαναλαμβάνουμε την διαδικασία προσαρμογής των δεδομένων σε κάποιο μοντέλο, αφού πρώτα έχουμε διαγράψει τις ακραίες παρατηρήσεις, ή κάνουμε κατάλληλο μετασχηματισμό σε κάποιες μεταβλητές και επαναλαμβάνουμε την ανάλυση χρησιμοποιώντας ολόκληρο το σύνολο των δεδομένων.

4.4 Επηρεάζουσες παρατηρήσεις[12]

Ενώ οι ακραίες τιμές ορίζονται σε σχέση με το προσαρμοζόμενο μοντέλο οι επηρεάζουσες παρατηρήσεις ορίζονται σε σχέση με τα αρχικά δεδομένα.

Πιο συγκεκριμένα, σε ένα σύνολο δεδομένων είναι πιθανόν να υπάρχει μία ή περισσότερες παρατηρήσεις οι οποίες έχουν σημαντική επίδραση στις τιμές των εκτιμητών των παραμέτρων του μοντέλου μας. Ένας τρόπος να μετρήσουμε την επίδραση αυτή είναι να υπολογίσουμε τις τιμές των εκτιμητών με και χωρίς την παρουσία των εν λόγω παρατηρήσεων.

Μία από τις απλούστερες γραφικές μεθόδους είναι η γραφική παράσταση των αρχικών δεδομένων, για την απλή γραμμική παλινδρόμηση, ή γενικότερα των μερικών υπολοίπων ως προς κάθε μία από τις ανεξάρτητες μεταβλητές, για την περίπτωση της πολλαπλής γραμμικής

[12] Βλέπε Βιβλ. Γραμμικά Μοντέλα Κεφάλαιο 2.4 Σελ. 64 έως 66.

παλινδρόμησης. Από μία τέτοια γραφική παράσταση μπορούμε να δούμε, συνήθως εύκολα, ποιες παρατηρήσεις είναι επηρεάζουσες ή τουλάχιστον υποψήφιες επηρεάζουσες.
Για τον στατιστικό έλεγχο της ύπαρξης επηρεαζουσών παρατηρήσεων υπάρχουν διάφοροι μέθοδοι. Μία από αυτές λόγω του ότι είναι απλή και περισσότερο χρησιμοποιούμενη είναι η μέθοδος του Cook (Cook’s distance).
Η ύπαρξη επηρεαζουσών παρατηρήσεων δεν σημαίνει ότι τα αποτελέσματα της ανάλυσης της παλινδρόμησης δεν είναι σωστά. Εν τούτοις είναι σημαντικό για τον ερευνητή να είναι σε θέση να προσδιορίσει τις επηρεάζουσες παρατηρήσεις και να προσπαθήσει να δώσει λογικές εξηγήσεις για την παρουσία των μέσα στο πλαίσιο του υπό μελέτη προβλήματος.

ΚΕΦΑΛΑΙΟ 5 ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΣΤΗΝ ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ[13]

5.1 ΠΟΛΥΣΥΓΓΡΑΜΜΙΚΟΤΗΤΑ

Με τον όρο πολυσυγγραμμικότητα εννοείται η ύπαρξη μιας ανεξάρτητης μεταβλητής X_J η οποία συσχετίζεται γραμμικά με μια άλλη ανεξάρτητη μεταβλητή ή με ένα γραμμικό συνδυασμό άλλων ανεξάρτητων μεταβλητών.
Είναι ένα φαινόμενο που απαντάται κυρίως σε μεταβλητές του κοινωνικού και οικονομικού χώρου. Αποτελεί μια από τις κυριότερες αιτίες για την εξαγωγή λανθασμένων συμπερασμάτων κατά την εφαρμογή μιας πολλαπλής γραμμικής παλινδρόμησης.
Η ύπαρξη πολυσυγγραμμικότητας συνεπάγεται την αύξηση των τυπικών σφαλμάτων (αποκλίσεων) των συντελεστών παλινδρόμησης. Είναι δυνατό να υπάρχει πλήρης πολυσυγγραμμικότητα στο πλήθος των ανεξάρτητων (προβλεπουσών) μεταβλητών X_i δηλαδή

$$X_j = \lambda_0 + \Sigma \lambda_i X_i \qquad i \neq j$$

χωρίς να υπάρχουν μεγάλες ανά δύο συσχετίσεις. Αυτό σημαίνει ότι ο έλεγχος μόνο του πίνακα συσχέτισης δεν αποκαλύπτει πάντα την ύπαρξη πολυσυγγραμμικότητας.

5.2 ΟΡΙΟ ΑΝΟΧΗΣ

Ο όρος αυτός είναι συνώνυμος αυτού της πολυσυγγραμμικότητας. Μια ανεξάρτητη μεταβλητή X_i θεωρείται σημαντική σε ένα μοντέλο πολλαπλής παλινδρόμησης όταν ο εκτιμούμενος συντελεστής παλινδρόμησης της bi, έχει μεγάλη τιμή. Δεν αρκεί όμως μόνο αυτό. Η σημαντικότητά της εξαρτάται και από το τυπικό σφάλμα με το οποίο εκτιμάται ο b_i. Συντελεστές με μεγάλα τυπικά σφάλματα θεωρούνται μη αξιόπιστοι και μπορούν να παίρνουν πολύ διαφορετικές τιμές για διάφορα δείγματα.
Όταν οι μεταβλητές X_i εμφανίζουν πολυσυγγραμμικότητα αυτό έχει ως αποτέλεσμα, οι συντελεστές παλινδρόμησης που εκτιμούνται να εμφανίζουν, με τη σειρά τους, μεγάλα τυπικά σφάλματα.
Υποθέτουμε ότι
$X_i = b_0 + b_1X_1 + b_2X_2 + \ldots + b_{i-1}X_{i-1} + b_{i+1}X_{i+1} + \ldots. + b_KX_K + E$
είναι το μοντέλο που θεωρεί τη μεταβλητή X_i ως εξαρτημένη από τις υπόλοιπες ανεξάρτητες. Τότε :

- η ποσότητα σ^2_i εκφράζει τη διακύμανση της μεταβλητής X_i (ή τη διακύμανση του παραπάνω μοντέλου), και
- η ποσότητα R^2_i εκφράζει τον συντελεστή προσδιορισμού της X_i (ή τον συντελεστή προσδιορισμού του παραπάνω μοντέλου). Η ποσότητα αυτή πλησιάζει τη μονάδα εάν υπάρχει μεγάλη πολυσυγγραμμικότητα.

Έστω η σχέση

$$\sigma^2(b_i) = \sigma^2 / (1- R^2_i)*(n-1)*\sigma^2_i$$

[13] Βλέπε http : //web.auth.gr/e-topo/TOMEIS_INDEX/TOMEASB/Lafazani/Give/kef10_2_Palindr_sysxet.pdf Σελ. 475 έως 478.

όπου σ^2 είναι η διακύμανση των σφαλμάτων.
Η ποσότητα 1- $R^2_i \rightarrow 0$, όταν $R^2_i \rightarrow 1$, και επομένως το τυπικό σφάλμα (η τυπική απόκλιση) του συντελεστή b_i που υπολογίζεται από την παραπάνω σχέση, αυξάνεται. Μάλιστα για σταθερό μέγεθος δείγματος και σταθερή διακύμανση σφαλμάτων το τυπικό σφάλμα αυξάνεται τόσο περισσότερο όσο μικρότερη είναι η ποσότητα 1- R^2_i . Όμως εάν η ποσότητα αυτή γίνει αρκετά μικρή είναι ενδεχόμενο να υπάρξουν και υπολογιστικά προβλήματα. Έτσι ορίζεται μια οριακή τιμή έτσι ώστε εάν η ποσότητα 1- R^2_i είναι μικρότερη από αυτήν τότε αποκλείεται η μεταβλητή X_i από το μοντέλο παλινδρόμησης. Η οριακή αυτή τιμή ονομάζεται όριο ανοχής.

5.3 ΕΤΕΡΟΣΚΕΔΑΣΤΙΚΟΤΗΤΑ

Κατά την ανάλυση παλινδρόμησης θεωρούμε ότι τα σφάλματα έχουν σταθερή διακύμανση. Σε πολλά όμως φαινόμενα αυτή η προϋπόθεση δεν ισχύει. Εάν για παράδειγμα η τυχαία εξαρτημένη μεταβλητή Y ακολουθεί κατανομή Poisson ή Διωνυμική, τότε η διακύμανση των τιμών της Y και επομένως και των υπολοίπων θα εξαρτάται από τη μέση τιμή και συνεπώς δεν θα είναι σταθερή. Λέγεται τότε ότι τα δεδομένα εμφανίζουν ετεροσκεδαστικότητα. Ένας τρόπος για να διαπιστωθεί το φαινόμενο της ετεροσκεδαστικότητας είναι η εξέταση της μορφής των υπολοίπων μετα την εφαρμογή του μοντέλου.

5.4 ΑΥΤΟΣΥΣΧΕΤΙΣΗ

Μια από τις υποθέσεις του μοντέλου της πολλαπλής παλινδρόμησης είναι η ανεξαρτησία των υπολοίπων e_i. Η ανεξαρτησία αυτή συνδέεται με το πρόβλημα της αυτοσυσχέτισης. Εάν, για παράδειγμα, το μοντέλο της παλινδρόμησης αναφέρεται σε δεδομένα που έχουν συλλεχθεί σε διαφορετικές χρονικές περιόδου, είναι πιθανό η τιμή της εξαρτημένης μεταβλητής Y της περιόδου k να σχετίζεται με τη δική της τιμή της περιόδου k-1. Τότε εμφανίζεται το πρόβλημα της αυτοσυσχέτισης. Ο δείκτης ο οποίος πληροφορεί για τον βαθμό αυτοσυσχέτισης ή για το βαθμό ανεξαρτησίας των υπολοίπων είναι το στατιστικό Durbin-Watson. Οι τιμές αυτού του στατιστικού κυμαίνονται στο διάστημα [0,4].

- Τιμές του στατιστικού γύρω στο 2 → δεν υπάρχει σχέση μεταξύ διαδοχικών υπολοίπων,
- Τιμές του στατιστικού κοντά στο 0 → τα διαδοχικά υπόλοιπα συσχετίζονται θετικά,
- Τιμές του στατιστικού κοντά στο 4 → υπάρχει ισχυρή αρνητική σχέση μεταξύ των υπολοίπων.

Στην πράξη, οι τιμές του στατιστικού στο διάστημα [1.5,2.5] δεν δημιουργούν πρόβλημα.

ΚΕΦΑΛΑΙΟ 6 ΕΠΙΛΟΓΗ ΑΝΕΞΑΡΤΗΤΩΝ ΜΕΤΑΒΛΗΤΩΝ[14]

Ένα από τα προβλήματα, στην πολλαπλή γραμμική παλινδρόμηση, είναι και αυτό της *επιλογής ανεξάρτητων μεταβλητών*. Το πρόβλημα προκύπτει όταν ο αριθμός των ανεξάρτητων μεταβλητών, που έχουμε προς μελέτη, είναι αρκετά μεγάλος (π.χ. >20) και ενδιαφερόμαστε να επιλέξουμε ένα υποσύνολο από αυτές. Αλλά και όταν ο αριθμός των ανεξάρτητων μεταβλητών δεν είναι αρκετά μεγάλος αντιμετωπίζουμε ανάλογο πρόβλημα εάν το ενδιαφέρον μας συγκεντρώνεται στην επιλογή των πιο "σημαντικών" μεταβλητών. Ένα μοντέλο με μικρότερο αριθμό ανεξάρτητων μεταβλητών, και συνεπώς και παραμέτρων, είναι πιο επιθυμητό, διότι είναι πιο οικονομικό και εύκολο στην ερμηνεία του.

Η μεθοδολογία, για την επιλογή ανεξάρτητων μεταβλητών προϋποθέτει ότι 1^{o}) η μορφή (π.χ. X, 1/X, X^2, logX κ.τ.λ.) με την οποία κάθε ανεξάρτητη μεταβλητή εισέρχεται στο μοντέλο είναι γνωστή, και 2^{o}) το μοντέλο μας είναι σωστό. Στην πράξη πρώτα γίνεται η επιλογή των μεταβλητών που θα "εισέλθουν" στο μοντέλο και στην συνέχεια βρίσκεται η μορφή της κάθε μεταβλητής στο μοντέλο καθώς και ο έλεγχος της ορθότητάς του.

Σε πολλές πρακτικές εφαρμογές της ανάλυσης παλινδρόμησης, το σύνολο μεταβλητών που περιλαμβάνονται στο μοντέλο παλινδρόμησης δεν προκαθορίζεται και είναι συχνά το πρώτο μέρος της ανάλυσης για να επιλέξει αυτές τις μεταβλητές. Φυσικά, υπάρχουν μερικές περιπτώσεις όταν θεωρητικές ή άλλες εκτιμήσεις καθορίζουν τις μεταβλητές που περιλαμβάνονται στην εξίσωση-εκεί το πρόβλημα της επιλογής μεταβλητής δεν προκύπτει. Αλλά σε καταστάσεις όπου δεν υπάρχει καμία ευδιάκριτη θεωρία, το πρόβλημα για την επιλογή μεταβλητών σε μια εξίσωση παλινδρόμησης γίνεται σημαντικό. Υποθέστε ότι έχουμε μια αποκριτική μεταβλητή και ένα σύνολο k προβλεπουσών μεταβλητών X_1, X_2, ….X_K : και επιθυμούμε να καθιερώσουμε μια εξίσωση γραμμικής παλινδρόμησης για αυτήν την συγκεκριμένη αποκριτική μεταβλητή σε σχέση με τις προβλέπουσες μεταβλητές. Θέλουμε να καθορίσουμε ή να επιλέξουμε το καλύτερο (σημαντικότερο ή εγκυρότερο) υποσύνολο των k προβλεπουσών μεταβλητών και το αντίστοιχο καλύτερο-κατάλληλο μοντέλο παλινδρόμησης για την περιγραφή της σχέσης μεταξύ του Y και των X. Τι ακριβώς εννοούμε λέγοντας το «καλύτερο» εξαρτάται εν μέρει από το γενικό στόχο μας στη διαμόρφωση του μοντέλου.

Πριν παρουσιάσουμε τις μεθόδους επιλογής ανεξάρτητων μεταβλητών θα αναφέρουμε, συνοπτικά, ποια είναι τα αποτελέσματα της μη "σωστής" διαγραφής μεταβλητών από ένα μοντέλο γραμμικής παλινδρόμησης. Πιο συγκεκριμένα ας υποθέσουμε ότι έχουμε ένα μοντέλο με p ανεξάρτητες μεταβλητές από το οποίο διαγράφουμε τις p-q (q<p) μεταβλητές (με άλλα λόγια επιλέγουμε q μεταβλητές σαν μεταβλητές για το μοντέλο μας).

Περίπτωση 1 (Υπο-προσαρμογή) Αν το "σωστό" μοντέλο περιλαμβάνει περισσότερες από q ανεξάρτητες μεταβλητές τότε μιλάμε για *υπο-προσαρμογή*. Στην περίπτωση αυτή σε όλες τις εκτιμήσεις μας (μερικοί συντελεστές παλινδρόμησης, διακύμανση, πρόβλεψη κ.ά.) εμπεριέχεται ένα σταθερό ποσό μεροληψίας. Αυτό που κάνουμε, όταν διαγράφουμε σημαντικές ανεξάρτητες μεταβλητές από το μοντέλο μας, είναι να προσθέτουμε την μεταβλητότητα που ερμηνεύεται από αυτές στο άθροισμα τετραγώνων των υπολοίπων. Αυτό έχει σαν αποτέλεσμα την αύξηση του αντίστοιχου μέσου τετραγώνου.

[14] Βλέπε Βιβλ. Γραμμικά Μοντέλα Κεφάλαιο 2.5 Σελ. 67 έως 68.

Περίπτωση 2 (Υπερ-προσαρμογή) Αν το "σωστό" μοντέλο περιλαμβάνει λιγότερες από q ανεξάρτητες μεταβλητές τότε μιλάμε για *υπερ-προσαρμογή*. Στην περίπτωση αυτή οι διακυμάνσεις των εκτιμητών μας και των προβλέψεων είναι μεγαλύτερες από ότι στο "σωστό" μοντέλο. Πιο συγκεκριμένα έστω ότι $\hat{\beta}_o, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k$ (k<q) είναι οι εκτιμητές ελαχίστων τετραγώνων για το "σωστό" μοντέλο και $\hat{\beta}_o^*, \hat{\beta}_1^*, \hat{\beta}_2^*, \cdots, \hat{\beta}_q^*$ οι αντίστοιχοι εκτιμητές για το μοντέλο με τις q επιλεγμένες μεταβλητές. Τότε

$$\mathrm{Var}(\hat{\beta}_i^*) \geq \mathrm{Var}(\hat{\beta}_i) \ (i=0, 1, 2, \cdots, k)$$

Ανάλογη σχέση ισχύει και για την διακύμανση των προβλέψεων για τα δύο μοντέλα.
Από τα προηγούμενα γίνεται φανερό ότι ένα πολύ απλό μοντέλο μπορεί να μας δίνει μη αμερόληπτους εκτιμητές, ενώ ένα πιο σύνθετο μοντέλο μπορεί να μας δίνει μεγάλες διακυμάνσεις. Συνεπώς, στις περισσότερες περιπτώσεις, η επιλογή του "σωστού" μοντέλου είναι ένας συμβιβασμός μεταξύ των δύο αυτών καταστάσεων.

Κίνδυνοι χρησιμοποιώντας μη-πειραματικά δεδομένα

Όταν κάνουμε υπολογισμούς παλινδρόμησης σε μη-πειραματικά δεδομένα που προκύπτουν από συνεχόμενες διαδικασίες και όχι από ένα σχεδιασμένο πείραμα, μπορεί να προκύψουν μερικά επικίνδυνα ενδεχόμενα όπως : α) Τα σφάλματα στο μοντέλο μπορεί να μην είναι τυχαία αλλά μπορεί να προκύπτουν από την κοινή επίδραση διαφόρων μεταβλητών που δεν περιλαμβάνονται στην εξίσωση της παλινδρόμησης, ή δεν έχουν καν μετρηθεί. β) Μπορεί να εισαχθεί μεροληψία. γ) Συνεπώς, η εξίσωση πρόβλεψης γίνεται αναξιόπιστη οφειλόμενη στη κοινή αλληλεπίδραση και τη σύγχυση των αποτελεσμάτων. δ) Το εύρος τιμών γίνεται μη-έγκυρο για την πρόβλεψη. ε) Δημιουργούνται μεγάλες συσχετίσεις μεταξύ προβλεπουσών μεταβλητών.
Ένα προσεκτικά σχεδιασμένο πείραμα μπορεί να εξαλείψει όλες τις αμφιβολίες που αναφέραμε παραπάνω.

6.1 Βήματα για την επιλογή του καλύτερου μοντέλου παλινδρόμησης[15]:

1. Καθορίστε τους πιθανούς όρους που περιλαμβάνονται στο μοντέλο.
2. Καθορίστε ένα κριτήριο για την επιλογή ενός μοντέλου.
3. Καθορίστε μια στρατηγική για την επιλογή μεταβλητών.
4. Αξιολογήστε το μοντέλο που επιλέγεται.

Βήμα 1 : Καθορίστε τους πιθανούς όρους που περιλαμβάνονται στο μοντέλο.

Σε εφαρμοσμένους τομείς πολλές ανεξάρτητες μεταβλητές (όπως το μελλοντικό εισόδημα) δεν είναι άμεσα μετρήσιμες. Υπό τέτοιους όρους, οι ερευνητές αναγκάζονται συχνά να ερευνήσουν για τις πιθανές ανεξάρτητες μεταβλητές που θα μπορούσαν πιθανά να αφορούν την εξαρτημένη ύπο μελέτη μεταβλητή. Επομένως, σε πρώτη φάση, προσπαθούμε να βρούμε τους πιθανούς όρους ή τις προβλέπουσες μεταβλητές ή τις λειτουργίες αυτών που θα μπορούσαν να περιληφθούν στο μοντέλο σε οποιοδήποτε σημείο στο στάδιο ανάπτυξης των μοντέλων. Δεδομένου αυτών των όρων, πρέπει να εξετάσουμε μερικά πράγματα όπως: εάν υπάρχει οποιαδήποτε αλληλεπίδραση, ή πρόβλημα πολυσυγραμμικότητας, ή πολυωνυμικοί όροι ή μερικοί άλλοι μετασχηματισμένοι όροι που απαιτούνται. Κατά συνέπεια, το πρόβλημα της επιλογής μεταβλητής και η λειτουργική προδιαγραφή της εξίσωσης συνδέονται μεταξύ τους. Οι ερωτήσεις που πρέπει να απαντηθούν διατυπώνοντας την εξίσωση παλινδρόμησης είναι : Ποιες μεταβλητές πρέπει να περιληφθούν, και με ποια μορφή θα πρέπει να περιληφθούν; Αν και ιδανικά τα δύο προβλήματα (επιλογή μεταβλητής και λειτουργική προδιαγραφή) πρέπει να λυθούν ταυτόχρονα, για απλότητα θα προτείνουμε να αντιμετωπιστούν ξεχωριστά. Καθορίζουμε αρχικά τις μεταβλητές που θα περιληφθούν στην εξίσωση και μετά από αυτό ερευνάμε την ακριβή μορφή στην οποία η μεταβλητή εισάγεται. Αυτή η προσέγγιση είναι ακριβώς μια απλοποίηση- αλλά καθιστά το πρόβλημα της επιλογής μεταβλητής ανιχνεύσιμο. Για ευκολία, υποθέστε ότι Z_1, Z_2, …Z_R, όλες οι συναρτήσεις μιας ή περισσοτέρων μεταβλητών X, που αντιπροσωπεύουν το πλήρες σύνολο των μεταβλητών από το οποίο η εξίσωση πρόκειται να επιλεχθεί και ότι το σύνολο περιλαμβάνει οποιεσδήποτε συναρτήσεις όπως τετράγωνα, σταυρωτά γινόμενα, λογαρίθμους, αντίστροφες και δυνάμεις, που πιστεύεται ότι είναι επιθυμητές και απαραίτητες. Έχοντας ένα υποσύνολο αυτών των προβλεπουσών μεταβλητών (ή τις συναρτήσεις τους) μπορούμε να δημιουργήσουμε όλα τα πιθανά μοντέλα. Εντούτοις, για χάρη της απλότητας αναφέρουμε τα μοντέλα που περιλαμβάνουν τις απλές προβλέπουσες μεταβλητές X_1, X_2, ….X_K μόνο για τώρα- αλλά οι ίδιες τεχνικές που αναφέρονται παρακάτω μπορούν να εφαρμοστούν στους συναρτησιακούς όρους Z_1, Z_2, …Z_R που προαναφέραμε.

[15] Βλέπε http : //www.angelfire.com/ab5/get5/selreg.pdf Σελ. 4 έως 5.

Βήμα 2 : Καθορίστε ένα κριτήριο για την επιλογή ενός μοντέλου.

Ένα σημαντικό και κρίσιμο βήμα στην επιλογή του καλύτερου μοντέλου είναι να διευκρινιστεί το κριτήριο επιλογής. Ένα κριτήριο επιλογής είναι ένας δείκτης που μπορεί να υπολογιστεί για κάθε υποψήφιο μοντέλο και να χρησιμοποιηθεί για να συγκρίνει τα μοντέλα. Κατά συνέπεια, λαμβάνοντας υπόψη ένα ιδιαίτερο κριτήριο επιλογής, τα υποψήφια μοντέλα μπορούν να διαταχθούν από το καλύτερο ως το χειρότερο. Αυτό βοηθά αυτόματα τη διαδικασία επιλογής του καλύτερου μοντέλου. Αυτή η συγκεκριμένη διαδικασία επιλογής κριτηρίου μπορεί να μην βρει το καλύτερο μοντέλο. Εν τούτοις, η χρησιμοποίηση ενός ειδικού κριτηρίου επιλογής μπορεί ουσιαστικά να μειώσει την εργασία που περιλαμβάνεται για την εύρεση ενός καλού μοντέλου. Προφανώς, το κριτήριο επιλογής πρέπει να αφορά το στόχο της ανάλυσης. Πολλά κριτήρια επιλογής για την επιλογή του καλύτερου μοντέλου έχουν προταθεί. Στη διαδικασία επιλογής μιας εξίσωσης συνήθως εμπλέκονται δύο αντιτιθέμενα κριτήρια[16] :

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 δύο αυτών ακραίων περιπτώσεων είναι αυτό που συνήθως ονομάζεται *επιλογή της καλύτερης εξίσωσης παλινδρόμησης*. Ωστόσο οι μέθοδοι επιλογής ανεξάρτητων μεταβλητών δεν είναι απαραίτητο να καταλήγουν στο ίδιο μοντέλο. Για να επιλέξουμε μεταξύ διαφορετικών μοντέλων μπορούμε να εφαρμόσουμε τόσο ‘υποκειμενικά’, όσο και ‘αντικειμενικά’ κριτήρια.

Στα υποκειμενικά κριτήρια συγκαταλέγονται α) οικονομία μοντέλου, β) πραγματικότητα και γ) συμφωνία μεθόδων επιλογής. Ας δούμε καθένα από τα κριτήρια αυτά χωριστά.

Οικονομία μοντέλου. Σύμφωνα με το κριτήριο αυτό επιλέγουμε εκείνο το μοντέλο που έχει τον μικρότερο αριθμό ανεξάρτητων μεταβλητών.

Πραγματικότητα. Ας υποθέσουμε ότι από προηγούμενες μελέτες ή από εμπειρία είναι γνωστό ότι ένας ή περισσότεροι παράγοντες θα πρέπει να συμπεριλαμβάνονται (ή να μην συμπεριλαμβάνονται) στο μοντέλο. Στην περίπτωση αυτή επιλέγουμε εκείνο το μοντέλο το οποίο περιέχει τους περισσότερους (ή τους λιγότερους) από τους παράγοντες αυτούς.

Συμφωνία μεθόδων. Σύμφωνα με το κριτήριο αυτό επιλέγουμε εκείνο το μοντέλο στο οποίο συμφωνούν οι περισσότερες από τις μεθόδους επιλογής. Σε μερικές περιπτώσεις τέτοιο μοντέλο μπορεί να μην υπάρχει.

Στα αντικειμενικά κριτήρια συγκαταλέγονται : 1) Το τετράγωνο του πολλαπλού συντελεστή προσδιορισμού $R^{2,}$ 2) Ο προσαρμοσμένος συντελεστής πολλαπλού προσδιορισμού R^2_{adj}, 3) Η στατιστική C_p του Mallows, 4) Το μέσο τετράγωνο υπολοίπων (MSR), 5) Το κριτήριο

[16] Βλέπε Βιβλ. Εφαρμοσμένη ανάλυση παλινδρόμησης Σελ. 349.

πληροφοριών του Akaike (AIC), 6) Το Hannan και Quinn κριτήριο, και 7) Το κριτήριο του Schwarz. Ας δούμε καθένα από τα κριτήρια αυτά χωριστά.

Κριτήριο 1 : Το τετράγωνο του πολλαπλού συντελεστή προσδιορισμού R^2[17]

Ο συντελεστής πολλαπλού προσδιορισμού, R^2 είναι μια αναλογία τετραγώνων :
$R^2 = 1 - SS_{res} / SS_{tot}$
όπου ο παρανομαστής είναι σταθερός για όλες τις πιθανές παλινδρομήσεις.
Για να χρησιμοποιούμε το R^2 θα πρέπει α) τα προς επιλογή μοντέλα να έχουν τον ίδιο αριθμό παραμέτρων και β) να μην έχουμε μετασχηματίσει την εξαρτημένη μεταβλητή. Αν οι προϋποθέσεις αυτές ικανοποιούνται, τότε επιλέγουμε εκείνο το μοντέλο με την μεγαλύτερη τιμή για το R^2. Το κριτήριο αυτό είναι χρήσιμο όταν σκοπός του μοντέλου μας είναι να περιγράψει την σχέση μεταξύ εξαρτημένης μεταβλητής και ανεξάρτητων μεταβλητών. Δυστυχώς το R^2 παρέχει ένα ανεπαρκές κριτήριο για την επιλογή υποσυνόλων. Το R^2 ποικίλει αντιστρόφως με το SS_{res}, αλλά ξέρουμε ότι το SS_{res} δεν μπορεί ποτέ να αυξηθεί δεδομένου ότι οι πρόσθετες ανεξάρτητες μεταβλητές συμπεριλαμβάνονται στο μοντέλο. Κατά συνέπεια το R^2 θα είναι μέγιστο όταν περιλαμβάνονται όλες οι πιθανές μεταβλητές X στην εξίσωση παλινδρόμησης. Ο λόγος χρησιμοποίησης του R^2 δεν γίνεται για να μεγιστοποιήσουμε το R^2, μάλλον η πρόθεση είναι να βρεθεί το σημείο όπου η προσθήκη περισσότερων X δεν είναι σημαντική επειδή οδηγεί σε μια πολύ μικρή αύξηση του R^2. Το R^2 δείχνει ότι υπάρχουν p παράμετροι σε k= p-1 προβλέπουσες μεταβλητές στην εξίσωση παλινδρόμησης στην οποία το R^2 είναι βασισμένο. Το R^2 δεν λαμβάνει υπόψη τον αριθμό παραμέτρων στο μοντέλο. Εάν οι προβλέπουσες μεταβλητές επιλέγονται τυχαία από κάποια κατανομή (ας πούμε κανονική) τότε η χρήση του R^2 μπορεί να είναι ικανοποιητική. Εάν οι επεξηγηματικές μεταβλητές καθορίζονται και ελέγχονται, τότε το R^2 απλά απεικονίζει την ελεγχόμενη μεταβολή στις επεξηγηματικές μεταβλητές. Για την απλή γραμμική παλινδρόμηση με μια δεδομένη κλίση, η πολλαπλή συσχέτιση μπορεί να αυξηθεί ή να μειωθεί με την αύξηση ή τη μείωση της αλλαγής των επεξηγηματικών μεταβλητών.
Το R^2 δεν είναι ο κατάλληλος δείκτης για να συγκρίνουμε ένα μοντέλο με q μεταβλητές με ένα μοντέλο με $p < q$ μεταβλητές, επειδή το R^2 αυξάνεται πάντα όταν μια νέα επεξηγηματική μεταβλητή εισέλθει στο μοντέλο και δεν μπορούμε να βγάλουμε ένα σωστό συμπέρασμα για τη σημαντικότητα της μεταβλητής αυτής.

Κριτήριο 2 : Ο προσαρμοσμένος συντελεστής πολλαπλού προσδιορισμού R^2_{adj}

$R^2_{Adj} = 1 - SS_{res}/(n-p) / SS_{tot}/(n-1) = 1 - s^2*(n-1)/SS_{tot}$
Ο προσαρμοσμένος συντελεστής πολλαπλού προσδιορισμού λαμβάνει υπόψη τους βαθμούς ελευθερίας των αθροισμάτων τετραγώνων που εμπλέκονται σε αυτόν. Εάν η σταθερά δεν συμπεριλαμβάνεται στο μοντέλο υποσυνόλου, τότε το (n-1) αντικαθίσταται από το n. Το R^2_{Adj} αυξάνεται εάν και μόνο εάν το $SS_{res}/(n-p)$ μειώνεται δεδομένου ότι το $SS_{tot}/(n-1)$ καθορίζεται για το Y. Μεγάλες τιμές του R^2_{Adj} οδηγούν σε καλά ή επιθυμητά μοντέλα. Το R^2_{Adj} δεν μπορεί να είναι ποτέ αρνητικό όπως το R^2. Επίσης το R^2_{Adj} αποτελεί ένα «ποινικοποιημένο μέτρο καλής

[17] Βλέπε http : //www.angelfire.com/ab5/get5/selreg.pdf Σελ. 5 έως 7.

προσαρμογής» (penalized measure of goodness of fit), αφού υπάρχει πιθανότητα αν προσθέσουμε μια μεταβλητή στο μοντέλο να μειωθεί το R^2_{Adj}, οπότε να συμπεράνουμε ότι αυτή η μεταβλητή δεν είναι απαραίτητη για το μοντέλο. Αν όμως αυξηθεί το R^2_{Adj} τότε η μεταβλητή αυτή είναι σημαντική για το μοντέλο και πρέπει να την συμπεριλάβουμε σε αυτό.

Κριτήριο 3 : Στατιστική C_p του Mallows.

Οι προβλεπόμενες μεταβλητές που λαμβάνονται από μια εξίσωση παλινδρόμησης βασισμένη σε ένα υποσύνολο μεταβλητών έχουν γενικά μεροληψία. Για να κρίνουμε την απόδοση μιας εξίσωσης πρέπει να λάβουμε υπόψη το μέσο τετράγωνο των υπολοίπων της προβλεπόμενης μεταβλητής παρά τη διασπορά. Ο C.L. Mallows πρότεινε μια στατιστική της μορφής :
$C_p = RSS / s^2 – (n-2p)$
η οποία λαμβάνει υπόψη και την μεροληψία καθώς και την διασπορά, όπου RSS 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 X και θεωρείται να είναι μια αξιόπιστη αμερόληπτη εκτίμηση της διασποράς σφάλματος σ^2.
Συνδέεται στενά με το R^2 και προφανώς με το R^2_{adj} και έχει άμεση σχέση με τη μερική F-στατιστική. Επίσης, $E(C_p) = p$ όταν δεν υπάρχει καμία μεροληψία στο προσαρμοσμένο μοντέλο. Η απόκλιση του C_p από το p μπορεί να χρησιμοποιηθεί ως μέτρο της μεροληψίας και τα υποσύνολα των μεταβλητών που «παράγουν» τιμές του C_p που είναι κοντά στο p είναι τα επιθυμητά υποσύνολα. Ο υπολογισμός του C_p υποθέτει αμερόληπτους εκτιμητές που μπορούν, όχι πάντα, να είναι αληθινοί. Ένα μειονέκτημα του C_p είναι ότι φαίνεται να είναι απαραίτητο να αξιολογηθεί το C_p για όλα (ή τα περισσότερα) από τα πιθανά υποσύνολα για να επιτρέψει την ερμηνεία. Επίσης μερικές φορές η επιλογή μπορεί να μην είναι σαφής. 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.

Με 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.

Σύμφωνα με το Mallow (1973) όλα τα μοντέλα με $C_p \approx p+1$ θα πρέπει να θεωρούνται σαν

υποψήφια για περαιτέρω μελέτη. Στην περίπτωση που η επιλογή γίνεται μεταξύ μοντέλων με ίδιο

αριθμό παραμέτρων, τότε επιλέγουμε εκείνο το μοντέλο για το οποίο $C_p \leq p+1$. Ο Hocking (1976) πρότεινε δύο άλλα κριτήρια για την επιλογή του καλύτερου μοντέλου με την βοήθεια του δείκτη του Mallow. Το πρώτο είναι το $C_p \leq p+1$, εάν σκοπός του μοντέλου είναι η πρόβλεψη, και το δεύτερο το $C_p \leq 2(p+1)-p$, εάν σκοπός του μοντέλου είναι απλώς η εκτίμηση των παραμέτρων. Ο Mallows πρότεινε ότι τα καλά μοντέλα θα έχουν αρνητικό ή μικρό C_p-p.

Κριτήριο 4 : Μέσο τετράγωνο υπολοίπων (MSR).

Με μια εξίσωση p-μεταβλητών, το MSR ορίζεται ως εξής :
$(MSR)_p = (SSE)_p /n-p$
Μεταξύ δύο εξισώσεων, συνήθως προτιμάται αυτή με το μικρότερο MSR. Συνδέεται με το R^2 και με το R^2_{adj} και έχει σχέση με τη μερική F-στατιστική. Αυτό το κριτήριο δεν χρειάζεται να επιλέξει το πλήρες μοντέλο δεδομένου ότι και το SS_{res} και το (n-p) θα μειωθεί καθώς το p αυξάνεται και έτσι το MSR_p μπορεί να μειωθεί ή να αυξηθεί. Ωστόσο, στις περισσότερες περιπτώσεις, αυτό το κριτήριο ευνοεί μεγάλα υποσύνολα παρά μικρότερα και είναι χρήσιμο όταν σκοπός του μοντέλου μας είναι η πρόβλεψη.

Κριτήριο 5 : Κριτήριο πληροφοριών του Akaike (AIC)[18].

Η διαδικασία AIC χρησιμοποιείται για να αξιολογήσει πόσο καλά το υποψήφιο μοντέλο προσεγγίζει το αληθινό μοντέλο με την αξιολόγηση της διαφοράς μεταξύ των προβλέψεων του διανύσματος y στα πλαίσια του αληθινού μοντέλου και του υποψηφίου μοντέλου χρησιμοποιώντας την απόσταση Kullback-Leibler. Η απόσταση Kullback-Leibler είναι η απόσταση μεταξύ της αληθινής πυκνότητας και της κατ' εκτίμηση πυκνότητας για κάθε μοντέλο. Το κριτήριο είναι :

$$\mathbf{AIC} = \ln|\hat{\boldsymbol{\Sigma}}_{\mathbf{p}}| + \frac{2pq + q(q+1)}{n}, \text{όπου } \hat{\boldsymbol{\Sigma}}_{\mathbf{p}} = \mathbf{Y}'[\mathbf{I} - \mathbf{X}_{\mathbf{P}}(\mathbf{X}'_{\mathbf{P}}\mathbf{X}_{\mathbf{P}})^{-1}\mathbf{X}'_{\mathbf{P}}]\mathbf{Y}.$$

Το μοντέλο που προβλέπει καλύτερα το Y εξ αδιαιρέτου, με αυτήν την διαδικασία, είναι αυτό που έχει την ελάχιστη τιμή AIC.

Η διορθωμένη μορφή του κριτηρίου πληροφοριών του Akaike.

Το κριτήριο πληροφοριών του Akaike ενδέχεται να οδηγήσει σε μικρά δείγματα. Κατά συνέπεια, προτάθηκε μια διορθωμένη έκδοση του AIC, η οποία είναι :

$$\mathbf{AIC_C} = \ln|\hat{\boldsymbol{\Sigma}}_{\mathbf{p}}| + \frac{(n+p)q}{n-p-q-1}$$

Το καλύτερο υποσύνολο του x, με αυτήν την διαδικασία, είναι αυτό που έχει την ελάχιστη τιμή AIC_C.

[18] Βλέπε http : //www.jstatsoft.org/v07/il2/paper Σελ. 11

Κριτήριο 6 : Hannan και Quinn.

Αν και το κριτήριο πληροφοριών HQ που παρουσιάστηκε από τους Hannan και Quinn προοριζόταν για χρήση με τα αυτοπαλινδρομικά πρότυπα, μπορεί επίσης να εφαρμοστεί στα μοντέλα παλινδρόμησης. Το κριτήριο είναι :

$$\mathbf{HQ} = \ln|\tilde{\Sigma}_{\mathbf{p}}| + \frac{2ln(ln(n))pq}{n}$$

Το καλύτερο μοντέλο είναι το μοντέλο που αντιστοιχεί στην ελάχιστη τιμή HQ.

Η διορθωμένη μορφή του κριτηρίου πληροφοριών των Hannan και Quinn.

Το κριτήριο πληροφοριών των Hannan και Quinn συνήθως πλεονάζει όταν εφαρμόζεται σε μικρά δείγματα (McQuarrie & Tsai). Επομένως, οι McQuarrie & Tsai πρότειναν μια διορθωμένη έκδοση του, η οποία είναι :

$$\mathbf{HQ}_C = \ln|\tilde{\Sigma}_p^2| + \frac{2ln(ln(n))pq}{n-p-q-1}$$

Ομοίως, η διαδικασία προσδιορίζει το καλύτερο υποσύνολο του x που παράγει τη μικρότερη τιμή.

Κριτήριο 7 : Κριτήριο του Schwarz

Η συνάρτηση υπολογίζει το Μπευζιανό κριτήριο πληροφοριών του Schwarz για κάθε μοντέλο χρησιμοποιώντας την απόσταση Kullback-Leibler, η οποία μπορεί να χρησιμοποιηθεί για να προσδιορίσει το καλύτερο μοντέλο. Το κριτήριο είναι :

$$\mathbf{BIC} = \ln|\tilde{\Sigma}_p^2| + \frac{ln(n)p}{n}$$

Το καλύτερο μοντέλο από τη διαδικασία είναι το μοντέλο που αντιστοιχεί στην ελάχιστη τιμή. Δεν τείνει να επιλέγει μοντέλα με μεγάλο αριθμό παραμέτρων (οπότε και δεν οδηγεί σε υπερ-παραμετροποίηση (over-fitting)) και μπορεί να χρησιμοποιηθεί για τη σύγκριση οποιονδήποτε μοντέλων (όχι μόνο restricted και unrestricted) αρκεί όμως να έχουν την ίδια εξαρτημένη μεταβλητή.

Βήμα 3 : Καθορίστε μια στρατηγική για την επιλογή μεταβλητών[19].

Το βασικό βήμα για την επιλογή του καλύτερου μοντέλου είναι να καθοριστεί η στρατηγική για την επιλογή μεταβλητών. Μια τέτοια στρατηγική λαμβάνει υπόψη τον καθορισμό για το πόσες μεταβλητές και επίσης ποιες συγκεκριμένες μεταβλητές πρέπει να είναι στο τελικό μοντέλο. Παραδοσιακά τέτοιες στρατηγικές έχουν εστιάσει στην απόφαση εάν μια μεταβλητή πρέπει να προστεθεί σε ένα μοντέλο ή εάν μια μεταβλητή πρέπει να διαγραφεί από ένα μοντέλο. Δεν

[19] Βλέπε http : //www.angelfire.com/ab5/get5/selreg.pdf Σελ. 8 έως 9.

υπάρχει καμία στατιστική διαδικασία ή στρατηγική για την επιλογή μεταβλητών. Μερικές δημοφιλείς στρατηγικές είναι :

1. Όλες οι δυνατές παλινδρομήσεις,
2. Διαδικασία της προς τα εμπρός επιλογής,
3. Διαδικασία της προς τα πίσω απαλοιφής,
4. Διαδικασία της παλινδρόμησης κατά βήματα.

Όμως δεν οδηγούν όλες απαραιτήτως στην ίδια λύση όταν εφαρμόζονται στο ίδιο πρόβλημα (αν και για πολλά προβλήματα, θα μας δώσουν την ίδια απάντηση). Στην πραγματικότητα καμία από τις διαδικασίες επιλογής μεταβλητών που αναφέρθηκαν παραπάνω δεν εγγυάται ότι θα επιλέξει την καλύτερη εξίσωση παλινδρόμησης για ένα δεδομένο σύνολο δεδομένων- αυτό οφείλεται στο γεγονός ότι δεν υπάρχει συνήθως μια «ενιαία» καλύτερη εξίσωση- αλλά μάλλον θα επιλέξει διάφορες εξίσου καλές με αυτές.

Στρατηγική 1 : Όλες οι δυνατές παλινδρομήσεις.

Με τον όρο "όλες τις δυνατές παλινδρομήσεις" εννοούμε ότι, σε ένα πρόβλημα με p ανεξάρτητες μεταβλητές, εκτελούμε πρώτα όλες τις δυνατές παλινδρομήσεις με μια ανεξάρτητη μεταβλητή, στη συνέχεια με δύο ανεξάρτητες μεταβλητές, με τρείς ανεξάρτητες μεταβλητές κ.ο.κ έως ότου εξαντλήσουμε και τις p ανεξάρτητες μεταβλητές. Η μέθοδος αυτή είναι η 'καλύτερη' με την προϋπόθεση ότι ο αριθμός των ανεξάρτητων μεταβλητών δεν είναι αρκετά μεγάλος. Πρακτικά η διαδικασία όλων των δυνατών παλινδρομήσεων προτιμάται πέρα από οποιαδήποτε άλλη στρατηγική επιλογής μεταβλητών. Εγγυάται ότι θα βρει το καλύτερο μοντέλο. Αυτή η διαδικασία είναι πολύ άμεση και εφαρμόσιμη εξίσου καλά και για τα γραμμικά και μη γραμμικά δεδομένα. Η στρατηγική αρχικά απαιτεί την προσαρμογή όλων των δυνατών εξισώσεων παλινδρόμησης οι οποίες περιλαμβάνουν τη X_0 και οποιοδήποτε αριθμό από τις μεταβλητές X_1, X_2, …, X_K (όπου έχουμε προσθέσει μια εικονική μεταβλητή X_0 =1 στο σύνολο των X ως συνήθως). Επειδή κάθε X_i μπορεί να είναι ή μπορεί να μην είναι (δύο δυνατότητες) στην εξίσωση της παλινδρόμησης, και αυτό ισχύει για κάθε X_i, i=1,2,…,k (k το πλήθος X), υπάρχουν συνολικά 2^k εξισώσεις (ο όρος X_0 είναι πάντοτε στην εξίσωση). Αν k=10, τότε ένας ασυνήθιστα μεγάλος αριθμός, 2^k = 1024 εξισώσεων πρέπει να εξεταστεί. Κάθε εξίσωση παλινδρόμησης αξιολογείται σύμφωνα με κάποιο κριτήριο. Τα τρία κριτήρια που χρησιμοποιούνται πιο συχνά είναι :

1. Η τιμή του R^2 που επιτυγχάνεται από την προσαρμογή ελαχίστων τετραγώνων.
2. Η τιμή του s^2, το μέσο τετράγωνο υπολοίπων.
3. Η στατιστική C_p.

Η επιλογή για το ποια εξίσωση είναι καλύτερο να χρησιμοποιηθεί γίνεται έπειτα με την αξιολόγηση των προαναφερθέντων μοντέλων. Κατά τη χρησιμοποίηση αυτής της μεθόδου, προσδιορίζονται οι πιο ελπιδοφόρες χρησιμοποιώντας αυτά τα κριτήρια και έπειτα αναλύονται προσεκτικά με την εξέταση των υπολοίπων για παρεκτρεπόμενες τιμές, αυτοσυσχέτιση, ή την

ανάγκη για μετασχηματισμούς πριν αποφασίσουμε σχετικά με το τελικό μοντέλο. Τα διάφορα υποσύνολα που ερευνώνται μπορεί να προτείνουν ερμηνείες των στοιχείων τα οποία ίσως να έχουν αγνοηθεί σε μια πιο περιορισμένη προσέγγιση επιλογής μεταβλητής. Αυτή η μέθοδος δίνει σαφώς σε έναν αναλυτή τον μέγιστο διαθέσιμο αριθμό πληροφοριών σχετικά με τη φύση των σχέσεων μεταξύ του Y και του συνόλου των X. Είναι η μόνη μέθοδος που εγγυάται να βρει το μοντέλο που έχει τα πλέον προτιμητέα κριτήρια (υπό την έννοια ότι οποιοδήποτε κριτήριο επιλογής θα βελτιστοποιηθεί αριθμητικά για το συγκεκριμένο δείγμα που είναι υπό μελέτη). Εντούτοις, φυσικά, η χρήση αυτής της στρατηγικής δεν εγγυάται την εύρεση του ορθού μοντέλου και τέτοια συμπεράσματα ενδέχεται να ποικίλουν από δείγμα σε δείγμα, ακόμα και αν όλα τα δείγματα επιλέγονται από τον ίδιο πληθυσμό. Κατά συνέπεια, η επιλογή του καλύτερου μοντέλου μπορεί να ποικίλει από δείγμα σε δείγμα. Στην πραγματικότητα, σε πολλές περιπτώσεις, διάφοροι λογικοί υποψήφιοι για το καλύτερο μοντέλο μπορούν να βρεθούν με διαφορετικά κριτήρια επιλογής που προτείνουν διαφορετικά καλύτερα μοντέλα. Επίσης, αυτή η στρατηγική δεν χρησιμοποιείται πάντα επειδή το απαραίτητο ποσό υπολογισμού γίνεται ‘μη πρακτικό’ όταν ο αριθμός των μεταβλητών που εξετάστηκε στο βήμα 1 είναι μεγάλος. Ενώ σημαίνει ότι ο ερευνητής έχει ‘εξετάσει όλες τις δυνατές παλινδρομήσεις’, σημαίνει επίσης ότι έχει εξετάσει έναν μεγάλο αριθμό εξισώσεων παλινδρόμησης τις οποίες η ευφυής σκέψη συχνά θα απέρριπτε. Ο χρόνος που απαιτείται για τον υπολογισμό των κριτηρίων σε όλα τα δυνατά μοντέλα είναι πολύ μεγάλος.

Έχουν προταθεί ορισμένες συντομότερες μέθοδοι (μια εκ των οποίων είναι η παλινδρόμηση καλύτερων υποσυνόλων που αναφέρουμε παρακάτω) οι οποίες δεν περιλαμβάνουν τον υπολογισμό ολόκληρου του συνόλου των εξισώσεων αναζητώντας τα επιθυμητά υποσύνολα. Αλλά με έναν μεγάλο αριθμό μεταβλητών, αυτές οι μέθοδοι πάλι περιλαμβάνουν ένα μη αμελητέο ποσό υπολογισμού. Επομένως, πολλές εναλλακτικές μέθοδοι έχουν προταθεί ως υπολογιστικά εφικτές για να προσεγγίσουν την διαδικασία ‘όλων των δυνατών παλινδρομήσεων’. Αν και αυτές οι μέθοδοι δεν είναι εγγυημένες ως προς την εύρεση του καλύτερου μοντέλου, μπορούν (με προσεκτική χρήση) να σταχυολογήσουν ουσιαστικά όλες τις πληροφορίες από τα στοιχεία που απαιτούνται για να επιλέξουν το καλύτερο μοντέλο.

Το βασικό μειονέκτημα της μεθόδου αυτής είναι ότι όσο μεγαλύτερος είναι ο αριθμός των ανεξάρτητων μεταβλητών, τόσο περισσότερο δυσκολεύουν οι απαραίτητοι υπολογισμοί και τα αποτελέσματα που παίρνουμε, από έναν υπολογιστή, είναι δύσκολο να διαβασθούν. Για παράδειγμα για p=6 υπάρχουν 2^6-1=63 γραμμικές παλινδρομήσεις, ενώ για p=10 υπάρχουν 2^{10}-1=1023 γραμμικές παλινδρομήσεις. Για την αντιμετώπιση του προβλήματος αυτού έχουν κατά καιρούς προταθεί διάφορες μέθοδοι. Γενικά, η ανάλυση όλων των εξισώσεων παλινδρόμησης είναι εντελώς αδικαιολόγητη. Αν και αυτή η ανάλυση σημαίνει ότι ο στατιστικός έχει “εξετάσει όλες τις δυνατότητες” επίσης σημαίνει ότι ο στατιστικός έχει εξετάσει ένα μεγάλο αριθμό εξισώσεων παλινδρόμησης τις οποίες με έξυπνες σκέψεις συχνά απορρίπτει από χέρι. Ο υπολογιστικός χρόνος που χρησιμοποιείται αποτελεί σπατάλη και η απαιτούμενη φυσική προσπάθεια για την εξέταση όλων των αποτελεσμάτων από τον υπολογιστή είναι τεράστια όταν εξετάζονται πολλές μεταβλητές. Είναι συνεπώς προτιμότερο να χρησιμοποιηθεί κάποια διαδικασία επιλογής που θα περιορίσει αυτή την εργασία.

Παλινδρόμηση «καλύτερων υποσυνόλων»[20].

Μια εναλλακτική λύση εκτέλεσης όλων των παλινδρομήσεων είναι να χρησιμοποιηθεί ένα πρόγραμμα που παρέχει μια λίστα των καλύτερων εξισώσεων Ε (ο πειραματιστής επιλέγει τις Ε) με μια προβλέπουσα μεταβλητή, με δύο, με τρείς …. και ούτω καθεξής μέσω της εξέτασης κάποιου επιλεγμένου εκ των προτέρων κριτηρίου ή κριτηρίων, όχι με την εξέταση όλων των εξισώσεων 2^K (επιλογή που είναι αυθαίρετη, έτσι ώστε να μην είναι αυθαίρετες οι εξισώσεις που πρέπει να περιληφθούν στον κατάλογο), με την β_0 να περιλαμβάνεται σε όλες αυτές τις εξισώσεις. Το πιθανό μειονέκτημα αυτής της διαδικασίας είναι ότι τείνει να παρέχει στις εξισώσεις πάρα πολλές συμπεριλαμβανόμενες προβλέπουσες μεταβλητές.

Στρατηγική 2 : Διαδικασία της προς τα εμπρός επιλογής **(Forward Selection)[21]**.

Για την υλοποίηση της μεθόδου αυτής ακολουθούμε τα επόμενα βήματα.
Βήμα 1: Ξεκινάμε από ένα μοντέλο το οποίο περιέχει μόνο τον σταθερό όρο.
Βήμα 2: Εκτελούμε όλες της γραμμικές παλινδρομήσεις της εξαρτημένης μεταβλητής με κάθε μία από της p ανεξάρτητες μεταβλητές. (Δηλαδή προσαρμόζουμε το μοντέλο $y=\beta_o+\beta_1 x_j$ (j=1, 2, . . .,p)). Για κάθε μία από αυτές τις παλινδρομήσεις έχουμε μια κρίσιμη πιθανότητα (p-value) για το F-τεστ που ελέγχει την H_o:$\beta_1=0$. Η ανεξάρτητη μεταβλητή με την μικρότερη κρίσιμη πιθανότητα εισέρχεται στο μοντέλο. Ας ονομάσουμε την μεταβλητή αυτή χ_1. (Προφανώς η μεταβλητή αυτή μπορεί να είναι μια οποιαδήποτε από τις ανεξάρτητες μεταβλητές $X_1, X_2, . . .,X_p$. Δηλαδή $\chi_1 = X_k$ (k=1, 2, . . .,p)).
Βήμα 3: Στην συνέχεια στο μοντέλο $y=\beta_o+\beta_1 x_1$ προσθέτουμε διαδοχικά μία κάθε φορά από τις υπόλοιπες p-1 ανεξάρτητες μεταβλητές. (Δηλαδή προσαρμόζουμε το μοντέλο $y=\beta_o+\beta_1\chi_1+\beta_2 x_j$ (j=1, 2, . . .,k-1, k+1,. . .,p). Για κάθε μία από αυτές τις παλινδρομήσεις έχουμε μια κρίσιμη πιθανότητα (p-value) για το F-τεστ που ελέγχει την H_o:$\beta_2=0$. Η ανεξάρτητη μεταβλητή με την μικρότερη κρίσιμη πιθανότητα είναι η νέα μεταβλητή που εισέρχεται στο μοντέλο. Ας ονομάσουμε την μεταβλητή αυτή χ_2. (Προφανώς η μεταβλητή αυτή μπορεί να είναι μια οποιαδήποτε από τις ανεξάρτητες μεταβλητές $X_1, X_2, . . .,X_{k-1}, X_{k+1}, . . .,X_P$. Δηλαδή $\chi_2 = X_K$ (k= 1, 2, . . ., k-1, k+1, . . ., p)).
Είναι προφανές ότι αν η διαδικασία αυτή συνεχισθεί χωρίς κανένα περιορισμό τότε το τελικό μοντέλο θα είναι αυτό που περιλαμβάνει όλες τις ανεξάρτητες μεταβλητές. Για τον σκοπό αυτό θέτουμε μια μέγιστη τιμή, p_{max}, για την κρίσιμη πιθανότητα p. Έτσι, για να εισέλθει μια ανεξάρτητη μεταβλητή στο μοντέλο θα πρέπει $p \leq p_{max}$. Μια συνηθισμένη τιμή για το p_{max} είναι η 0,25.

[20] Βλέπε http : //www.angelfire.com/ab5/get5/selreg.pdf Σελ. 19.
[21] Βλέπε Βιβλ. Γραμμικά Μοντέλα Κεφάλαιο 2.5 Σελ. 68 έως 69.

Στρατηγική 3 : Διαδικασία της προς τα πίσω απαλοιφής **(Backward Elimination)**[22].

Η μέθοδος αυτή ακολουθεί την αντίθετη διαδικασία από αυτή της προηγούμενης μεθόδου. 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. Πιο συγκεκριμένα η υλοποίηση της μεθόδου αυτής γίνεται ακολουθώντας τα επόμενα βήματα.

Βήμα 1: Ξεκινάμε προσαρμόζοντας το μοντέλο εκείνο που περιλαμβάνει όλες τις ανεξάρτητες μεταβλητές. Δηλαδή το μοντέλο $y=\beta_0+\beta_1x_1+\beta_2x_2+\propto+\beta_px_p+\varepsilon$. Στην συνέχεια υπολογίζουμε, για κάθε X_j (j=1, 2, . . .,6), τις τιμές των F-τεστ (ισοδύναμα των t-τεστ), που ελέγχουν τις υποθέσεις $H_{0j}:\beta_j=0$ (j=1, 2, . . .,p), και τις αντίστοιχες κρίσιμες πιθανότητες. Διαγράφουμε, από το μοντέλο μας, την μεταβλητή X_j για την οποία ο έλεγχος της $H_{0j}:\beta_j=0$ έδωσε την μεγαλύτερη κρίσιμη πιθανότητα. Συνεπώς, στην φάση αυτή, το μοντέλο μας είναι το $y=\beta_o+\beta_1x_1+….+\beta_{j-1}x_{j-1}+\beta_{j+1}x_{j+1}+….+\beta_px_p+\varepsilon$.

Βήμα 2: Προσαρμόζουμε το μοντέλο στο βήμα 1 και υπολογίζουμε, για τις μεταβλητές στο μοντέλο μας, τις τιμές των F-τεστ (ισοδύναμα των t-τεστ), που ελέγχουν τις υποθέσεις $H_{0k}:\beta_k=0$ (k=1, . . .,j-1, j+1, . . .,p), και τις αντίστοιχες κρίσιμες πιθανότητες. Διαγράφουμε, από το μοντέλο μας, την μεταβλητή X_k για την οποία ο έλεγχος της $H_{0k}:\beta_k=0$ έδωσε την μεγαλύτερη κρίσιμη πιθανότητα. Συνεπώς, στην φάση αυτή, το μοντέλο μας είναι το $y=\beta_0+\beta_1x_1+….+\beta_{k-1}x_{k-1}+\beta_{k+1}x_{k+1}+….+\beta_{j-1}x_{j-1}+\beta_{j+1}x_{j+1}+….+\beta_px_p+\varepsilon$.

Γίνεται φανερό ότι αν η διαδικασία αυτή συνεχισθεί, χωρίς κανένα περιορισμό, τότε το τελικό μοντέλο θα περιλαμβάνει μόνο τον σταθερό όρο β_0. Για τον σκοπό αυτό θέτουμε μια ελάχιστη τιμή, p_{MIN}, για την κρίσιμη πιθανότητα p. Έτσι, για να διαγραφεί μια ανεξάρτητη μεταβλητή από το μοντέλο θα πρέπει $p \geq p_{MIN}$. Μια συνηθισμένη τιμή για το p_{MIN} είναι η 0,10.

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 απαιτείται για την εφαρμογή της, συγκριτικά με τη μέθοδο "όλων των δυνατών παλινδρομήσεων". Ωστόσο, αν τα δεδομένα δίνουν ένα Χ΄Χ πίνακα που είναι σχεδόν ιδιάζων τότε η υπερπροσαρμοσμένη εξίσωση μπορεί να μην έχει νόημα λόγω των σφαλμάτων στρογγυλοποίησης. Όμως το πρόβλημα αυτό δεν είναι συνήθως σοβαρό αν χρησιμοποιήσουμε σύγχρονα προγράμματα υπολογιστή για την αντιστροφή πίνακα. Πρέπει να αναγνωρίζουμε ότι αν και μια μεταβλητή απαλειφτεί κατά τη διαδικασία, τότε παραμένει για πάντα εκτός παλινδρόμησης. Επομένως όλα τα εναλλακτικά μοντέλα που περιλαμβάνουν μεταβλητές που έχουν απαλειφτεί δεν είναι πλέον διαθέσιμα για εξέταση.

[22] Βλέπε Βιβλ. Γραμμικά Μοντέλα Κεφάλαιο 2.5 Σελ. 70 έως 71.

Στρατηγική 4 : Διαδικασία της παλινδρόμησης κατά βήματα **(Stepwise Selection)**[23].

Κοινό χαρακτηριστικό και των δύο προηγούμενων μεθόδων είναι ότι αν μια μεταβλητή εισέλθει στο μοντέλο ή διαγραφεί από αυτό, τότε θα μείνει, αντίστοιχα, μονίμως εντός ή εκτός του μοντέλου. Η μικτή μέθοδος επιλογής ανεξάρτητων μεταβλητών διορθώνει το πρόβλημα αυτό. Αν έχουμε στη διάθεση μας ένα μεγάλο αριθμό ανεξάρτητων μεταβλητών είναι πιθανόν κάποιες από αυτές να είναι περιττές, δηλαδή να μην έχουν να προσφέρουν επιπλέον πληροφορία για την μεταβλητότητα της Y, όταν ήδη έχουμε θεωρήσει την εξάρτηση της Y από κάποιες άλλες μεταβλητές. Υπάρχουν διάφορες μέθοδοι για την επιλογή των «σημαντικότερων» μεταβλητών και η πιο διαδεδομένη είναι η τεχνική της βηματικής παλινδρόμησης (stepwise regression technique).
Η διαδικασία αυτής της τεχνικής αρχίζει με το απλό σταθερό μοντέλο $y_i = \beta_0 + \varepsilon_i$. Σε κάθε βήμα προστίθεται μια ανεξάρτητη μεταβλητή στο μοντέλο μόνο αν αυτή η μεταβλητή προσφέρει σημαντική πληροφορία για την Y, επιπρόσθετα στην πληροφορία που παρέχουν οι ανεξάρτητες μεταβλητές που είναι ήδη στο μοντέλο από το προηγούμενο βήμα. Η διαδικασία αυτή επιτυγχάνεται με τους εξής στατιστικούς ελέγχους :

1. Αν κάποια από τις μεταβλητές που δεν ήταν στο μοντέλο πρέπει να προστεθεί σε αυτές που ήδη έχουν συμπεριληφθεί.
2. Αν κάποια από τις μεταβλητές που ήταν στο μοντέλο του προηγούμενου βήματος πρέπει να παραμείνει στο νέο μοντέλο, στο οποίο έχει συμπεριληφθεί μια νέα μεταβλητή.

Στο τέλος η διαδικασία της βηματικής παλινδρόμησης δίνει ένα σύνολο από ανεξάρτητα χρήσιμες μεταβλητές για την εξήγηση της Y.

Για την υλοποίηση της μεθόδου αυτής ακολουθούμε τα επόμενα βήματα.
Βήμα 1: Ξεκινάμε όπως και στην μέθοδο της προσθήκης ανεξάρτητων μεταβλητών. Δηλαδή από το μοντέλο το οποίο περιέχει μόνο τον σταθερό όρο.
Βήμα 2: Εκτελούμε όλες της γραμμικές παλινδρομήσεις της εξαρτημένης μεταβλητής με κάθε μία από της p ανεξάρτητες μεταβλητές. (Δηλαδή προσαρμόζουμε το μοντέλο $y=\beta_0+\beta_1 x_j$ (j=1, 2, . . .,p)). Για κάθε μία από αυτές τις παλινδρομήσεις έχουμε μια κρίσιμη πιθανότητα (p-value) για το F-τεστ που ελέγχει την $H_0:\beta_1=0$. Η ανεξάρτητη μεταβλητή με την μικρότερη κρίσιμη πιθανότητα εισέρχεται στο μοντέλο. Ας ονομάσουμε την μεταβλητή αυτή χ_1. (Προφανώς η μεταβλητή αυτή μπορεί να είναι μια οποιαδήποτε από τις ανεξάρτητες μεταβλητές $X_1, X_2, . . .,X_p$. Δηλαδή $\chi_1 = X_\kappa$ (k=1, 2, . . .,p)).
Βήμα 3: Στην συνέχεια στο μοντέλο $y=\beta_0+\beta_1\chi_1$ προσθέτουμε διαδοχικά μία κάθε φορά από τις υπόλοιπες p-1 ανεξάρτητες μεταβλητές. (Δηλαδή προσαρμόζουμε το μοντέλο $y=\beta_0+\beta_1\chi_1+\beta_2 x_j$ (j=1, 2, . . .,k-1, k+1,. . .,p). Για κάθε μία από αυτές τις παλινδρομήσεις έχουμε μια κρίσιμη πιθανότητα (p-value) για το F-τεστ που ελέγχει την $H_0:\beta_2=0$. Η ανεξάρτητη μεταβλητή με την μικρότερη κρίσιμη πιθανότητα είναι η νέα μεταβλητή που εισέρχεται στο μοντέλο. Ας ονομάσουμε την μεταβλητή αυτή χ_2. (Προφανώς η μεταβλητή αυτή μπορεί να είναι μια οποιαδήποτε από τις ανεξάρτητες μεταβλητές $X_1, X_2, . . .,X_{\kappa-1}, X_{\kappa+1}, . .,X_p$. Δηλαδή $\chi_2 = X_\kappa$ (κ= 1, 2, . . ., k-1, k+1, . . ., p)).
Βήμα 4: Έχοντας το μοντέλο $y=\beta_0+\beta_1\chi_1+\beta_2\chi_2$ εξετάζουμε αν κάποια από τις μεταβλητές x_1 και x_2 μπορεί να διαγραφεί από το μοντέλο μας. Ο τρόπος ελέγχου είναι ο ίδιος με αυτόν που

[23] Βλέπε Βιβλ. Γραμμικά Μοντέλα Κεφάλαιο 2.5 Σελ. 73.

διαγράφουμε μεταβλητές στην μέθοδο διαγραφής μεταβλητών.
Η διαδικασία αυτή συνεχίζεται έως ότου καμία νέα μεταβλητή δεν μπορεί να εισέλθει στο μοντέλο μας, αλλά και καμία μεταβλητή δεν μπορεί να διαγραφεί από αυτό. Η απόφαση για το αν μια μεταβλητή μπορεί να εισέλθει στο μοντέλο ή να διαγραφεί από αυτό συγκρίνοντας τις κατάλληλες κρίσιμες πιθανότητες με τις προκαθορισμένες τιμές των p_{MIN} και p_{MAX}, αντίστοιχα. Ένα καλό ζευγάρι τιμών για τα p_{MIN} και p_{MAX} είναι p_{MIN}=0,10 και p_{MAX}=0,25. Το συγκεκριμένο ζευγάρι τιμών δεν είναι το μοναδικό. Για οποιοδήποτε όμως ζευγάρι τιμών θα πρέπει να ισχύει $p_{MAX} \geq p_{MIN}$.
Η μέθοδος αυτή είναι η καλύτερη από τις μεθόδους επιλογής μεταβλητών που αναφέραμε. 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. Όπως με όλες τις μεθόδους που αναφέραμε παραπάνω, έτσι και εδώ απαιτείται μια λογική αιτιολόγηση στην αρχική επιλογή των μεταβλητών και στην κριτική εξέταση του μοντέλου μέσω της εξέτασης των υπολοίπων. Είναι εύκολο να βασιστούμε αρκετά στην αυτόματη επιλογή που γίνεται από τον υπολογιστή.
Τόσο η μέθοδος της προς τα πίσω απαλοιφής και η προς τα εμπρός επιλογής όσο και η μέθοδος της παλινδρόμησης κατά βήματα παρουσιάζουν κάποια δυσκολία η οποία δεν είναι προφανής με την πρώτη ματιά. Στην κατά βήματα διαδικασία, για παράδειγμα, ο μερικός F-έλεγχος που γίνεται στην μεταβλητή εισόδου χρησιμοποιεί τη μεγαλύτερη μερική F-τιμή από όλες εκείνες τις μερικές F-τιμές που μπορεί να έχουν επιλεγεί, για μεταβλητές που δεν είναι στην παλινδρόμηση σε εκείνο το στάδιο. Η σωστή δειγματική κατανομή και η κατανομή ελέγχου της μηδενικής υπόθεσης δεν είναι η συνήθης F-κατανομή όπως υποθέτουμε, και είναι πολύ δύσκολο να βρεθεί, εκτός από ορισμένες απλές περιπτώσεις. Για παράδειγμα, μελέτες έχουν δείξει, ότι, σε μερικές περιπτώσεις όπου ένας F-έλεγχος για εισαγωγή μεταβλητής έγινε σε επίπεδο σημαντικότητας α, η κατάλληλη πιθανότητα ήταν qα όπου σε εκείνο το στάδιο υπήρχαν q υποψήφιες μεταβλητές για να μπουν στην παλινδρόμηση. Για αυτό το πρόβλημα μια δυνατότητα είναι να βρούμε τα σωστά επίπεδα πιθανότητας για κάθε δοθείσα περίπτωση, ενώ μια άλλη δυνατότητα είναι να χρησιμοποιήσουμε μια διαφορετική στατιστική ελέγχου αντί του μερικού F. Οι δυνατότητες αυτές έχουν εξεταστεί σε πρόσφατες ερευνητικές εργασίες αλλά το συνολικό πρόβλημα δεν έχει κατάλληλα αντιμετωπιστεί σε έκταση που θα μπορούσε να δικαιολογήσει την τροποποίηση των διαδικασιών. Μέχρι να αντιμετωπιστεί αυτό το πρόβλημα, μπορούμε να χρησιμοποιούμε τις διαδικασίες όπως δίνονται, να μην μας απασχολεί το πρόβλημα των πραγματικών επιπέδων πιθανότητας και απλά να θεωρούμε τις διαδικασίες ότι κάνουν μια σειρά εσωτερικών συγκρίσεων που θα δίνουν ότι φαίνεται να είναι το πιο χρήσιμο σύνολο προβλεπουσών μεταβλητών.

Όλες οι προηγούμενες μέθοδοι μπορούν να τροποποιηθούν κατάλληλα εάν υπάρχουν πληροφορίες ότι κάποια ή κάποιες από τις ανεξάρτητες μεταβλητές θα πρέπει οπωσδήποτε να συμπεριληφθούν στο μοντέλο ή αντίθετα να διαγραφούν από αυτό.

Επιλογή ‘κανόνων διακοπής’ για τις διαδικασίες βηματικής παλινδρόμησης[24].

Η επιλογή F-to-enter και F-to-remove θα καθορίσει κατά ένα μεγάλο μέρος το χαρακτήρα της διαδικασίας βηματικής παλινδρόμησης. Είναι συνετό να θέσουμε :

α) F-to-enter μεγαλύτερη από την F-to-remove ή

β) F-to-enter μικρότερη από την F-to-remove

για να παρασχεθεί ‘προστασία’ στις προβλέπουσες μεταβλητές που έχουν ήδη αναγνωρισθεί στην εξίσωση. Ειδάλλως, κάποιος μερικές φορές μπορεί να απορρίψει τις προβλέπουσες μεταβλητές που μόλις αναγνωρίστηκαν. Επίσης, είναι δυνατό να τεθούν και οι δύο ίσες. Μια άλλη δημοφιλής επιλογή είναι η F-to-enter και η F-to-remove να είναι ίσες με 4 το οποίο αντιστοιχεί κατά προσέγγιση στο επίπεδο 5% της F-κατανομής. Για θεωρητικούς λόγους, μερικοί προτείνουν τη χρήση του σημείου 25% της F-κατανομής ως F-enter και το σημείο 10% της κατάλληλης F-κατανομής για F-remove. Η επιλογή των τιμών για τις F-enter και F-remove αποτελεί κατά ένα μεγάλο μέρος θέμα προσωπικής προτίμησης του αναλυτή, και συχνά λαμβάνονται σημαντικές πρωτοβουλίες σε αυτή την περιοχή.

[24] Βλέπε http : //www.angelfire.com/ab5/get5/selreg.pdf Σελ. 29 έως 33.

Πλεονεκτήματα της βηματικής παλινδρόμησης.
Μόνο οι μεταβλητές που είναι σε σημαντικό βαθμό γραμμικά σχετικές με το y, συμπεριλαμβάνονται στο μοντέλο. Ο αναλυτής μπορεί να επιλέξει τα κατώτατα επίπεδα σημασίας όσον αφορά στο συνυπολογισμό ή στην αφαίρεση. Αιτιολογεί την επίδραση της προσθήκης νέων ανεξάρτητων μεταβλητών σε ολόκληρο το μοντέλο παρά την επίδραση στο y μεμονωμένα και είναι υπολογιστικά αποδοτικό.

Μειονεκτήματα της βηματικής παλινδρόμησης.
Πολλά t-test (ή F-test μιας μεταβλητής) θα έχουν γίνει, και έτσι υπάρχει μια υψηλή πιθανότητα ότι τουλάχιστον μια ανεξάρτητη μεταβλητή θα έχει συμπεριληφθεί όταν δεν θα έπρεπε. Κάποια πολυσυγραμμικότητα μπορεί να παραμείνει. Συνήθως μόνο οι όροι πρώτης τάξης εξετάζονται για το μοντέλο, δεδομένου ότι ο αριθμός των όρων υψηλότερης τάξης που είναι πιθανοί υποψήφιοι για συνυπολογισμό αυξάνεται γρήγορα με τον αριθμό των ανεξάρτητων μεταβλητών. Αυτό μπορεί να οδηγήσει σε μερικούς σημαντικούς συσχετισμούς μεταξύ του y και των όρων υψηλότερης τάξης που δεν δοκιμάζονται ποτέ. Συνεπώς, οποιοιδήποτε όροι υψηλότερης τάξης που υπάρχει υποψία ότι συσχετίζονται σημαντικά με το y πρέπει να λαμβάνονται υπόψη στη βηματική ανάλυση.

Μια σημείωση για τη βηματική παλινδρόμηση.
Η βηματική παλινδρόμηση μπορεί μόνο να χρησιμεύσει ως ένα μερικό εργαλείο για την διαλογή μεταβλητών που εξετάζονται ήδη για συνυπολογισμό. Πρέπει πάντα να προηγείται από μια διαδικασία επιλογής βασισμένη στις βασικές αρχές και την κρίση ειδικών. Πρέπει επίσης να ακολουθηθεί από μια ανάλυση 'όλων των πιθανοτήτων' χρησιμοποιώντας τα αποτελέσματα της βηματικής, και τελικά από αναλύσεις στις οποίες συμπεριλαμβάνονται όλοι οι σχετικοί όροι υψηλότερης τάξης (πολυωνυμικοί και αλληλεπίδρασης κ.τ.λ.). Η βηματική μπορεί πάλι να χρησιμοποιηθεί σε αυτή τη φάση.

Επιλογή του μοντέλου.
Ενώ οι διαδικασίες που συζητήθηκαν δεν επιλέγουν απαραιτήτως το απόλυτο καλύτερο μοντέλο, επιλέγουν συνήθως ένα αποδεκτό. Εντούτοις, εναλλακτικές διαδικασίες έχουν προταθεί στις προσπάθειες να βελτιωθεί η επιλογή του μοντέλου. Μια πρόταση ήταν: Τρέξτε τη διαδικασία βηματικής παλινδρόμησης με δεδομένο επίπεδο για αποδοχή και απόρριψη. Όταν η διαδικασία επιλογής σταματήσει, καθορίστε τον αριθμό μεταβλητών στο τελικό επιλεγμένο μοντέλο. Χρησιμοποιώντας αυτόν τον αριθμό μεταβλητών, για παράδειγμα το q, κάντε όλα τα πιθανά σύνολα μεταβλητών του q από τις αρχικές μεταβλητές του k και επιλέξτε το καλύτερο σύνολο : εντούτοις, το προστιθέμενο πλεονέκτημα (χρησιμοποιώντας το αποτέλεσμα της διαδικασίας βηματικής παλινδρόμησης ως δεδομένο) αυτής της διαδικασίας είναι δευτερεύον.

Βήμα 4 : Αξιολογήστε το μοντέλο που επιλέγεται.

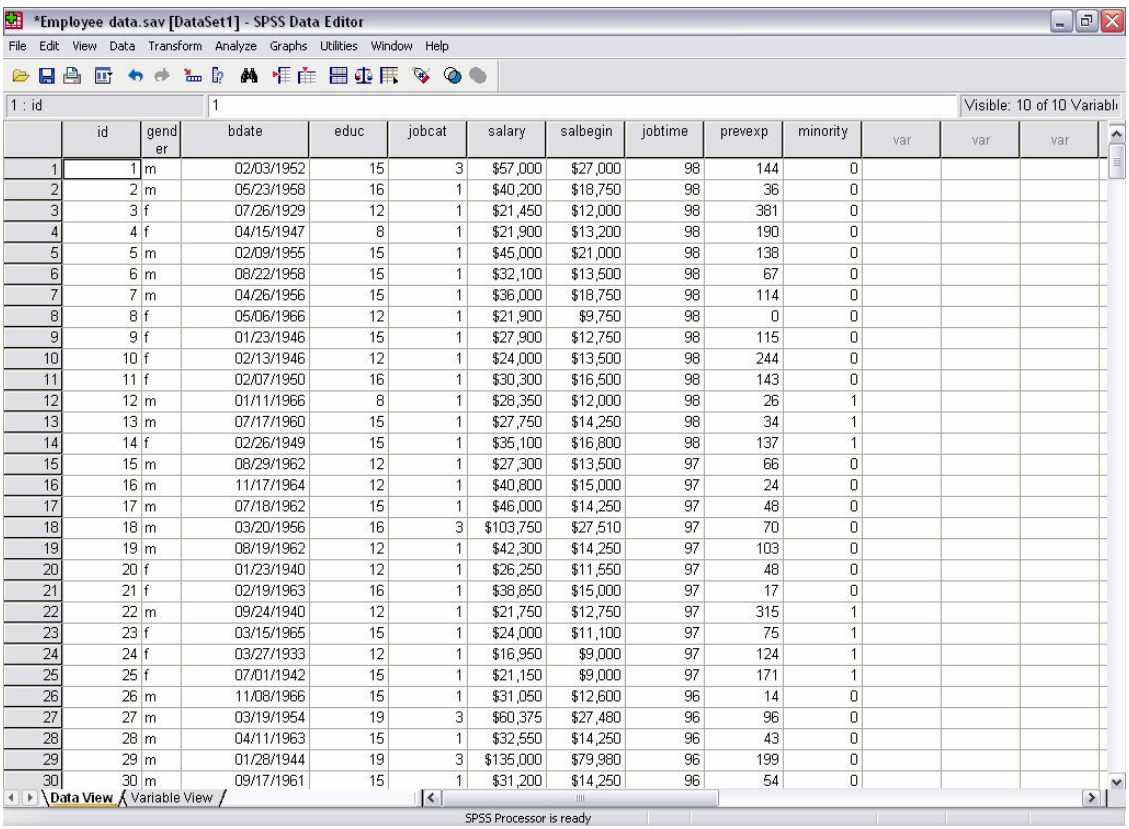
Έχοντας διευκρινίσει τους πιθανούς όρους ή τις μεταβλητές που θα περιληφθούν στο μοντέλο, το κριτήριο για την επιλογή ενός μοντέλου και τη στρατηγική για την εφαρμογή του κριτηρίου, πρέπει να διεξάγουμε την ανάλυση σύμφωνα με το πρόγραμμα για να λάβουμε το απαιτούμενο μοντέλο μας. Η καταλληλότητα του μοντέλου που επιλέγεται πρέπει βεβαίως να εξεταστεί με τις συνηθισμένες διαγνωστικές μεθόδους παλινδρόμησης για να καταδείξει ότι το μοντέλο που επιλέγεται είναι λογικό για τα διαθέσιμα δεδομένα.

ΚΕΦΑΛΑΙΟ 7 ΕΦΑΡΜΟΓΗ ΣΤΟ SPSS ΜΕ ΟΙΚΟΝΟΜΙΚΑ ΔΕΔΟΜΕΝΑ

Θα μελετήσουμε το αρχείο employee data.sav το οποίο βρίσκεται στο στατιστικό πακέτο SPSS. Ο φάκελος Employee data περιέχει 474 παρατηρήσεις από άτομα διαφορετικών ηλικιών, φύλου και εκπαιδευτικής καταρτίσεως. Όπως φαίνεται παρακάτω στο Data Editor οι αρχικές μας μεταβλητές είναι :

a. id : Κώδικας εργαζομένου από 1 έως 474
b. gender : Φύλλο (male ή female)
c. bdate : Ημερομηνία γεννήσεως
d. educ : Τα έτη εκπαίδευσης
e. jobcat : Η κατηγορία εργασίας
f. salary : Ο τωρινός μισθός
g. salbegin : Ο αρχικός μισθός
h. jobtime : Η περίοδος που βρίσκεται άνεργος σε μήνες
i. prevexp : Η προϋπηρεσία σε μήνες
j. minority : Αν ανήκει σε κάποια μειονότητα

Υποθέτοντας ότι ο μισθός εξαρτάται από το φύλλο, την ημερομηνία γεννήσεως, την κατηγορία εργασίας, τον αρχικό μισθό, την περίοδο που βρίσκεται άνεργος, την προϋπηρεσία αλλά και από το αν ανήκει σε κάποια μειονότητα, θέλουμε να κατασκευάσουμε και να ελέγξουμε το αντίστοιχο μοντέλο παλινδρόμισης.


*Employee data.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : id | 1 | Visible: 10 of 10 Variabl

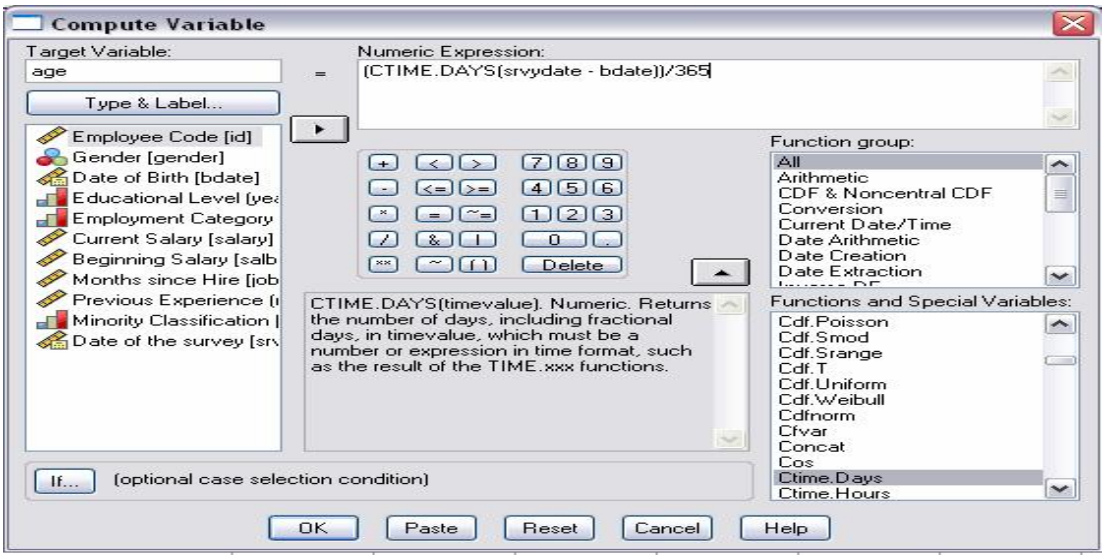
	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	var	var	var
1	1	m	02/03/1952	15	3	$57,000	$27,000	98	144	0			
2	2	m	05/23/1958	16	1	$40,200	$18,750	98	36	0			
3	3	f	07/26/1929	12	1	$21,450	$12,000	98	381	0			
4	4	f	04/15/1947	8	1	$21,900	$13,200	98	190	0			
5	5	m	02/09/1955	15	1	$45,000	$21,000	98	138	0			
6	6	m	08/22/1958	15	1	$32,100	$13,500	98	67	0			
7	7	m	04/26/1956	15	1	$36,000	$18,750	98	114	0			
8	8	f	05/06/1966	12	1	$21,900	$9,750	98	0	0			
9	9	f	01/23/1946	15	1	$27,900	$12,750	98	115	0			
10	10	f	02/13/1946	12	1	$24,000	$13,500	98	244	0			
11	11	f	02/07/1950	16	1	$30,300	$16,500	98	143	0			
12	12	m	01/11/1966	8	1	$28,350	$12,000	98	26	1			
13	13	m	07/17/1960	15	1	$27,750	$14,250	98	34	1			
14	14	f	02/26/1949	15	1	$35,100	$16,800	98	137	1			
15	15	m	08/29/1962	12	1	$27,300	$13,500	97	66	0			
16	16	m	11/17/1964	12	1	$40,800	$15,000	97	24	0			
17	17	m	07/18/1962	15	1	$46,000	$14,250	97	48	0			
18	18	m	03/20/1956	16	3	$103,750	$27,510	97	70	0			
19	19	m	08/19/1962	12	1	$42,300	$14,250	97	103	0			
20	20	f	01/23/1940	12	1	$26,250	$11,550	97	48	0			
21	21	f	02/19/1963	16	1	$38,850	$15,000	97	17	0			
22	22	m	09/24/1940	12	1	$21,750	$12,750	97	315	1			
23	23	f	03/15/1965	15	1	$24,000	$11,100	97	75	1			
24	24	f	03/27/1933	12	1	$16,950	$9,000	97	124	1			
25	25	f	07/01/1942	15	1	$21,150	$9,000	97	171	1			
26	26	m	11/08/1966	15	1	$31,050	$12,600	96	14	0			
27	27	m	03/19/1954	19	3	$60,375	$27,480	96	96	0			
28	28	m	04/11/1963	15	1	$32,550	$14,250	96	43	0			
29	29	m	01/28/1944	19	3	$135,000	$79,980	96	199	0			
30	30	m	09/17/1961	15	1	$31,200	$14,250	96	54	0			

Data View | Variable View

SPSS Processor is ready

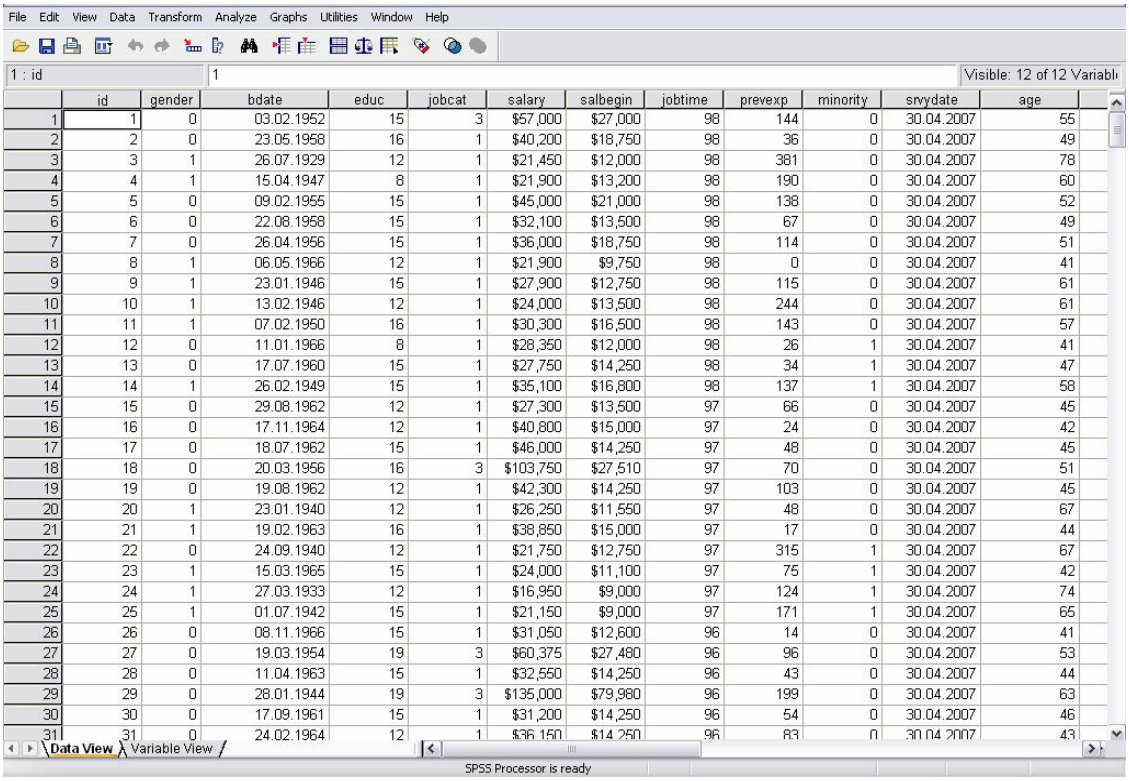
Εικόνα 1

Η μεταβλητή gender παρουσιάζει μια περίεργη συμπεριφορά. Είναι μια μη-ποιοτική μεταβλητή και επειδή είναι μια μεταβλητή η οποία είναι *type=string* όπου έχει τιμές m για male και f για female την μετατρέπουμε σε *type=numeric(1,0)*, όπου για την τιμή male βάζουμε 0 και για την τιμή female βάζουμε 1.
Επίσης για να υπολογίσουμε την ηλικία των εργαζομένων θα δημιουργήσουμε μια νέα μεταβλητή, η οποία θα μας δείχνει την σημερινή ημερομηνία και στη συνέχεια θα την αφαιρέσουμε από την ημερομηνία γεννήσεως δηλαδή την μεταβλητή bdate για να βρούμε την ηλικία.
Τη νέα αυτή μεταβλητή την οποία θα την ονομάσουμε age και θα είναι type=numeric θα την υπολογίσουμε χρησιμοποιώντας τη διαδικασία *Transform/Compute* variables. Στη φόρμα που άνοιξε κάνουμε το εξής :



Εικόνα 2

Στο παράθυρο Target Variable βάζουμε το όνομα της νέας μεταβλητής, δηλαδή age, στο παράθυρο Numeric Expression επιλέγουμε την συνάρτηση CTIME.DAYS και μέσα στην παρένθεση βάζουμε srvydate-bdate. Έπειτα το διαιρούμε με το 365 και πατάμε OK.
Όπως βλέπουμε στο παρακάτω παράθυρο στο Data Editor υπάρχουν δυο καινούργιες μεταβλητές : η μεταβλητή srvydate και η μεταβλητή age.

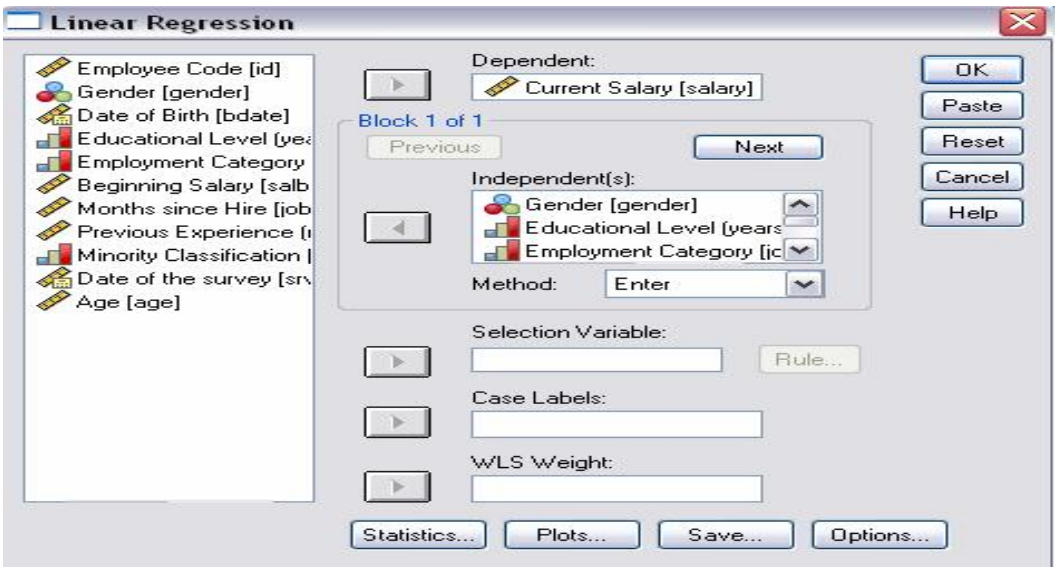


	id	gender	bdate	educ	jobcat	salary	salbegin	jobtime	prevexp	minority	srvydate	age
1	1	0	03.02.1952	15	3	$57,000	$27,000	98	144	0	30.04.2007	55
2	2	0	23.05.1958	16	1	$40,200	$18,750	98	36	0	30.04.2007	49
3	3	1	26.07.1929	12	1	$21,450	$12,000	98	381	0	30.04.2007	78
4	4	1	15.04.1947	8	1	$21,900	$13,200	98	190	0	30.04.2007	60
5	5	0	09.02.1955	15	1	$45,000	$21,000	98	138	0	30.04.2007	52
6	6	0	22.08.1958	15	1	$32,100	$13,500	98	67	0	30.04.2007	49
7	7	0	26.04.1956	15	1	$36,000	$18,750	98	114	0	30.04.2007	51
8	8	1	06.05.1966	12	1	$21,900	$9,750	98	0	0	30.04.2007	41
9	9	1	23.01.1946	15	1	$27,900	$12,750	98	115	0	30.04.2007	61
10	10	1	13.02.1946	12	1	$24,000	$13,500	98	244	0	30.04.2007	61
11	11	1	07.02.1950	16	1	$30,300	$16,500	98	143	0	30.04.2007	57
12	12	0	11.01.1966	8	1	$28,350	$12,000	98	26	1	30.04.2007	41
13	13	0	17.07.1960	15	1	$27,750	$14,250	98	34	1	30.04.2007	47
14	14	1	26.02.1949	15	1	$35,100	$16,800	98	137	1	30.04.2007	58
15	15	0	29.08.1962	12	1	$27,300	$13,500	97	66	0	30.04.2007	45
16	16	0	17.11.1964	12	1	$40,800	$15,000	97	24	0	30.04.2007	42
17	17	0	18.07.1962	15	1	$46,000	$14,250	97	48	0	30.04.2007	45
18	18	0	20.03.1956	16	3	$103,750	$27,510	97	70	0	30.04.2007	51
19	19	0	19.08.1962	12	1	$42,300	$14,250	97	103	0	30.04.2007	45
20	20	1	23.01.1940	12	1	$26,250	$11,550	97	48	0	30.04.2007	67
21	21	1	19.02.1963	16	1	$38,850	$15,000	97	17	0	30.04.2007	44
22	22	0	24.09.1940	12	1	$21,750	$12,750	97	315	1	30.04.2007	67
23	23	1	15.03.1965	15	1	$24,000	$11,100	97	75	1	30.04.2007	42
24	24	1	27.03.1933	12	1	$16,950	$9,000	97	124	1	30.04.2007	74
25	25	1	01.07.1942	15	1	$21,150	$9,000	97	171	1	30.04.2007	65
26	26	0	08.11.1966	15	1	$31,050	$12,600	96	14	0	30.04.2007	41
27	27	0	19.03.1954	19	3	$60,375	$27,480	96	96	0	30.04.2007	53
28	28	0	11.04.1963	15	1	$32,550	$14,250	96	43	0	30.04.2007	44
29	29	0	28.01.1944	19	3	$135,000	$79,980	96	199	0	30.04.2007	63
30	30	0	17.09.1961	15	1	$31,200	$14,250	96	54	0	30.04.2007	46
31	31	0	24.02.1964	12	1	$36,150	$14,250	96	83	0	30.04.2007	43

Εικόνα 3

7.1 ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΗ ΜΕΘΟΔΟ ENTER

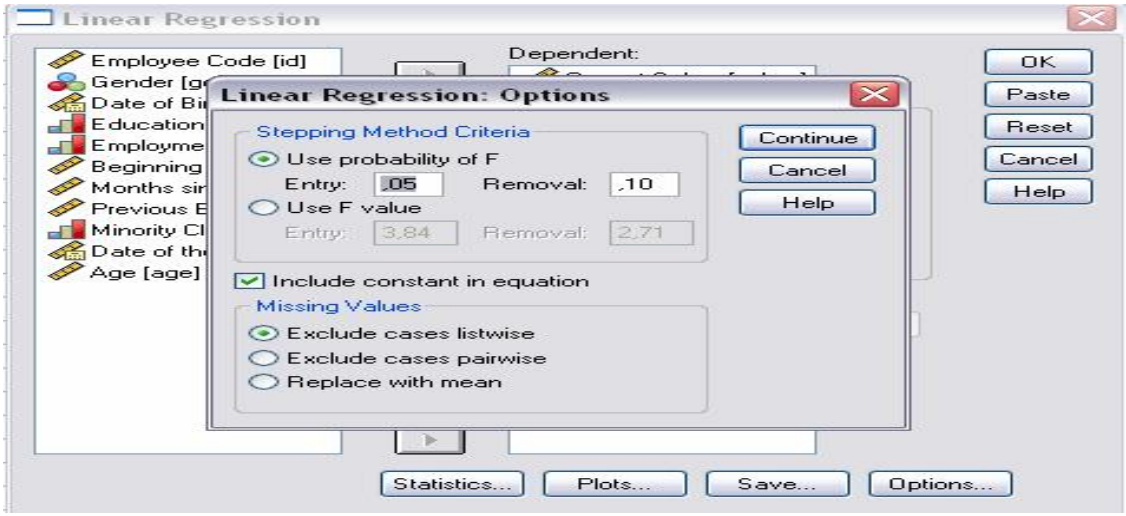
Χρησιμοποιώντας τη διαδικασία *Analyze/Regression/Linear* προκύπτει η παρακάτω φόρμα :



Εικόνα 4

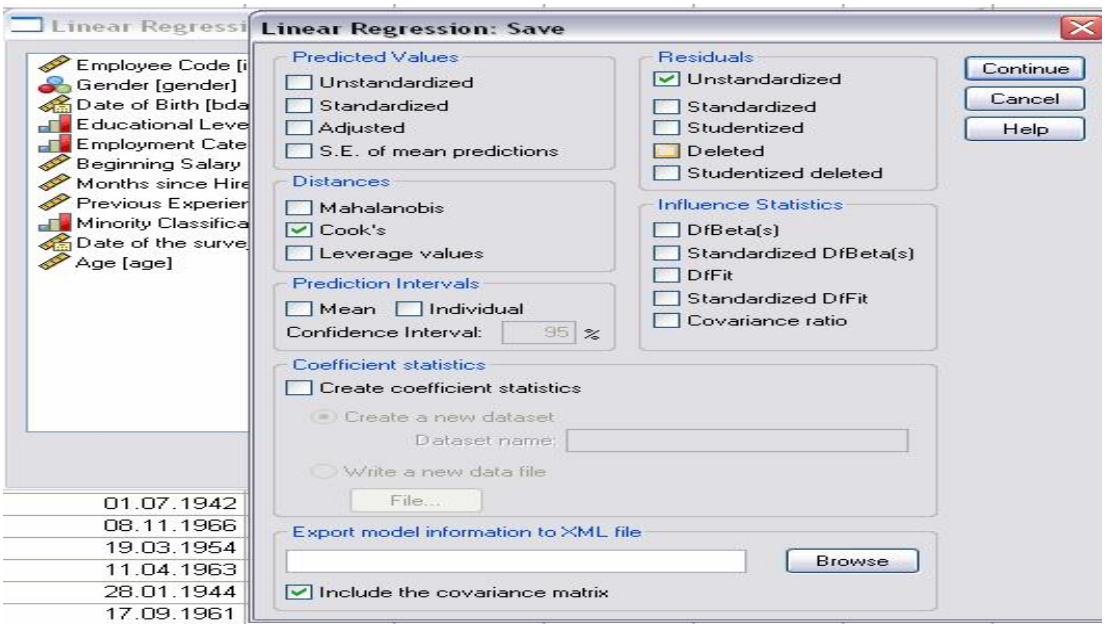
Στο παράθυρο **Dependent** μεταφέρουμε την μεταβλητή **Current Salary**. Στο παράθυρο

Independent(s) μεταφέρουμε τις μεταβλητές **Gender, Educational Level, Employment Category, Beginning Salary, Months since Hire, Previous Experience, Minority Classification** και **Age**. Στο παράθυρο **Method** επιλέγουμε την επιλογή **Enter**.



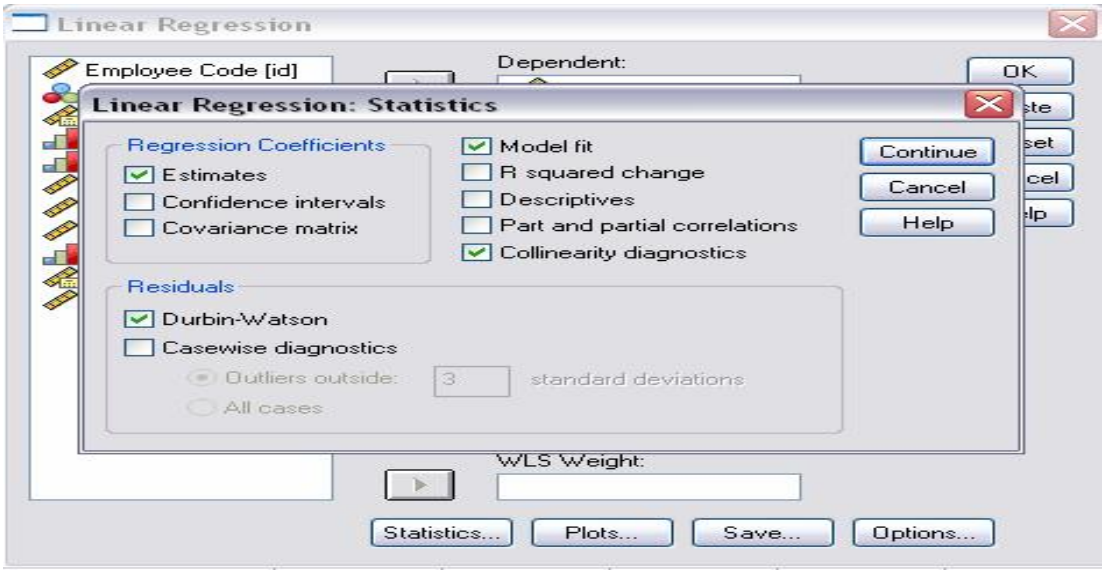
Εικόνα 5

Στην επιλογή **Options** έχουμε τα κριτήρια εισόδου-εξόδου των μεταβλητών στο μοντέλο με **F-entry** 0,05 και **F-removal** 0,10. Η ένδειξη **Include constant in equation** τσεκάρεται για να πάρουμε το σταθερό όρο του μοντέλου της παλινδρόμησης και στη συνέχεια **Continue**.



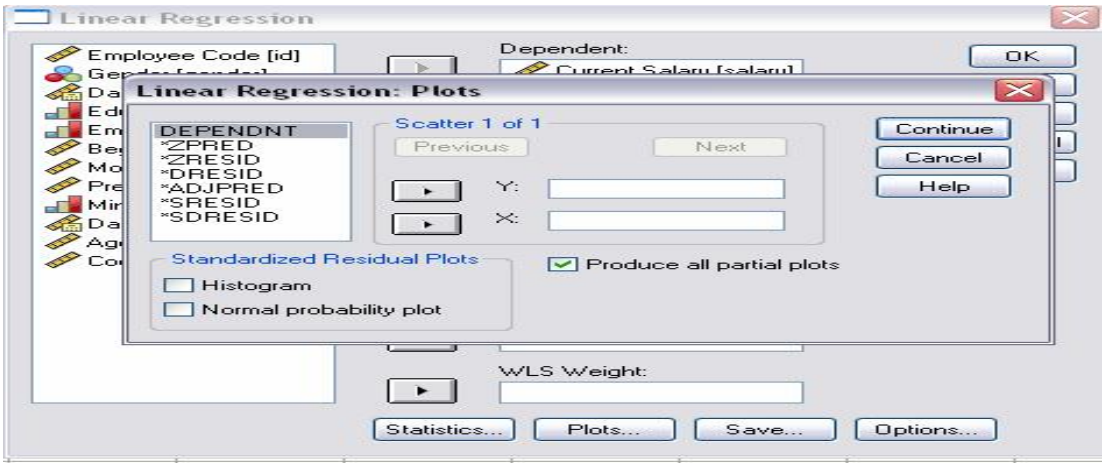
Εικόνα 6

Στην επιλογή **Save** από την περιοχή **Distances** τσεκάρουμε την επιλογή **Cook's** και στη συνέχεια **Continue**.



Εικόνα 7

Στην επιλογή **Statistics** εμφανίζεται η φόρμα *Linear Regression Statistics*. Από την περιοχή **Regression Coefficients** τσεκάρουμε τις επιλογές **Estimates, Model fit** και **Collinearrity diagnostics**. Από την περιοχή **Residuals** τσεκάρουμε την επιλογή **Durbin-Watson** και στη συνέχεια **Continue**.



Εικόνα 8

Στην επιλογή **Plots** τσεκάρουμε την επιλογή **Produce all partial plots** και στη συνέχεια **Continue**.
Επιστρέφουμε στη φόρμα *Linear Regression* και πατάμε **OK**.

7.1.1 ΕΜΦΑΝΙΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΣΤΟ OUTPUT

Regression

Variables Entered/Removed(b)

Model	Variables Entered	Variables Removed	Method
1	Age, Beginning Salary, Months since Hire, Minority Classification, Gender, Educational Level (years), Employment Category, Previous Experience (months)(a)	.	Enter

a All requested variables entered.
b Dependent Variable: Current Salary

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,919(a)	,844	,842	$6,804.678	1,842

a Predictors: (Constant), Age, Beginning Salary, Months since Hire, Minority Classification, Gender, Educational Level (years), Employment Category, Previous Experience (months)
b Dependent Variable: Current Salary

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	116431326192,544	8	14553915774,068	314,315	,000(a)
	Residual	21484887290,331	464	46303636,402		
	Total	137916213482,875	472			

a Predictors: (Constant), Age, Beginning Salary, Months since Hire, Minority Classification, Gender, Educational Level (years), Employment Category, Previous Experience (months)
b Dependent Variable: Current Salary

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta	Tolerance		VIF	Std. Error
1	(Constant)	-9201,795	3585,516		-2,566	,011		
	Gender	-1821,259	774,466	-,053	-2,352	,019	,658	1,520
	Educational Level (years)	455,592	154,114	,077	2,956	,003	,496	2,016
	Employment Category	5753,386	622,689	,260	9,240	,000	,423	2,367
	Beginning Salary	1,329	,070	,612	18,873	,000	,319	3,135
	Months since Hire	154,589	31,651	,091	4,884	,000	,970	1,031
	Previous Experience (months)	-14,939	5,484	-,091	-2,724	,007	,298	3,359
	Minority Classification	-970,413	784,623	-,024	-1,237	,217	,927	1,079
	Age	-65,494	47,502	-,045	-1,379	,169	,313	3,196

a Dependent Variable: Current Salary

Residuals Statistics(a)

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	$13,619.49	$130,700.26	$34,418.45	$15,705.940	473
Std. Predicted Value	-1,324	6,130	,000	1,000	473
Standard Error of Predicted Value	560,933	3525,955	903,403	255,033	473
Adjusted Predicted Value	$13,573.80	$129,122.05	$34,414.07	$15,692.686	473
Residual	-$23,208.023	$46,345.547	$.000	$6,746.764	473
Std. Residual	-3,411	6,811	,000	,991	473
Stud. Residual	-3,539	6,840	,000	1,002	473
Deleted Residual	-$24,994.711	$46,749.957	$4.373	$6,892.353	473
Stud. Deleted Residual	-3,584	7,206	,002	1,015	473
Mahal. Distance	2,209	125,732	7,983	7,293	473
Cook's Distance	,000	,107	,002	,009	473
Centered Leverage Value	,005	,266	,017	,015	473

a Dependent Variable: Current Salary

Στον πίνακα **Model Summary** βρίσκουμε τα κριτήρια $\mathbf{R^2}$ και $\mathbf{R^2_{adj}}$ με τιμές που είναι αρκετά ικανοποιητικές, $\mathbf{R^2}$=0,850 και $\mathbf{R^2_{adj}}$=0,822.
Το κριτήριο **MSR**(μέσο τετράγωνο υπολοίπων)= 46303636,402 και το κριτήριο $\mathbf{C_p}$=464 βγαίνει από τον τύπο $RSS/s^2 – (n – 2p)$.
Στον πίνακα **ANOVA** εξετάζουμε την στατιστική σημαντικότητα όλου του μοντέλου. Αυτό γίνεται κάνοντας τον έλεγχο :
$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$ *vs* $H_1: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 \neq 0$
Στη στήλη *Sig* το *p_value* είναι 0 δηλαδή < 0,05, άρα απορρίπτουμε την H_0 και αποδεχόμαστε ότι το μοντέλο μας είναι στατιστικά σημαντικό. Να σημειώσουμε ότι τα β_1, β_2, …, β_8 είναι εκτιμητές.
Στον πίνακα **Coefficients** εξετάζουμε αν οι μεταβλητές μας είναι στατιστικά σημαντικές ή όχι. Κάνουμε τον έλεγχο $H_0 : \hat{\beta}_i = 0$ *VS* $H_1 : \hat{\beta}_i \neq 0$.Βλέπουμε ότι οι μεταβλητές minority και age έχουν p_values 0,217 και 0,169 αντίστοιχα οι οποίες είναι > 0,05, άρα απορρίπτονται από το μοντέλο. Στο παρακάτω Output έχουμε τα αποτελέσματα χωρίς αυτές τις δυο μεταβλητές.

Regression

Variables Entered/Removed(b)

Model	Variables Entered	Variables Removed	Method
1	Previous Experience (months), Months since Hire, Beginning Salary, Gender, Educational Level (years), Employment Category(a)	.	Enter

a All requested variables entered.
b Dependent Variable: Current Salary

Model Summary(b)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,918(a)	,843	,841	$6,809.086	1,850

a Predictors: (Constant), Previous Experience (months), Months since Hire, Beginning Salary, Gender, Educational Level (years), Employment Category
b Dependent Variable: Current Salary

Στον πίνακα **Model Summary** μπορούμε να δούμε κατά πόσο οι ανεξάρτητες μεταβλητές μπορούν να εξηγήσουν τη μεταβολή της εξαρτημένης μεταβλητής. Αυτό το βλέπουμε από το $\mathbf{R^2_{adj}}$ το οποίο είναι 0,841. Σε σχέση με το προηγούμενο μοντέλο δεν άλλαξε σημαντικά, μετά την αφαίρεση των δυο μεταβλητών. Όπως επίσης και το $\mathbf{R^2}$.
Επίσης με το τεστ του *Durbin-Watson* μπορούμε να κάνουμε τον έλεγχο αυτοσυσχέτισης για το μοντέλο μας. Στην συγκεκριμένη περίπτωση ο δείκτης είναι 1,850 και είναι πολύ καλός, δηλαδή δεν υπάρχει αυτοσυσχέτιση.

ANOVA(b)

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	116264670118,587	6	19377445019,765	417,945	,000(a)
	Residual	21651825317,752	467	46363651,644		
	Total	137916495436,340	473			

a Predictors: (Constant), Previous Experience (months), Months since Hire, Beginning Salary, Gender, Educational Level (years), Employment Category
b Dependent Variable: Current Salary

Στον πίνακα **ANOVA** βρίσκουμε αν είναι στατιστικά σημαντικό το μοντέλο μας. Κάνουμε τον έλεγχο :
$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$ VS $H_1 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 \neq 0$
Τα $\beta_1, \beta_2, \ldots, \beta_6$ είναι εκτιμητές.
Στη στήλη Sig το p_value είναι 0 που είναι < 0,05, δηλαδή το μοντέλο μας είναι στατιστικά σημαντικό.
Έχουμε το κριτήριο **MSR**(μέσο τετράγωνο υπολοίπων) = 46363651,644 και το κριτήριο **Cp** που μπορούμε να το υπολογίσουμε και είναι $\mathbf{C_p} = 467$.

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Tolerance	VIF
1	(Constant)	-11817,136	3193,726		-3,700	,000		
	Gender	-2005,047	728,208	-,059	-2,753	,006	,744	1,345
	Educational Level (years)	471,879	153,656	,080	3,071	,002	,499	2,005
	Employment Category	5802,331	620,945	,263	9,344	,000	,425	2,352
	Beginning Salary	1,328	,070	,612	19,030	,000	,325	3,079
	Months since Hire	148,524	31,325	,088	4,741	,000	,987	1,013
	Previous Experience (months)	-21,446	3,300	-,131	-6,500	,000	,823	1,215

a Dependent Variable: Current Salary

Στον πίνακα **Coefficients** κάνουμε έλεγχο για τη στατιστική σημαντικότητα των μεταβλητών. Ο έλεγχος είναι :
$H_0 : \hat{\boldsymbol{\beta}}_i = 0$ VS $H_1 : \hat{\boldsymbol{\beta}}_i \neq 0$

Η στήλη Sig μας δίνει τα p_value όπου για όλες τις μεταβλητές είναι < 0,05. Άρα όλες οι μεταβλητές είναι στατιστικά σημαντικές.

Η στήλη Collinearity Statistics με τους δείκτες VIF και Tolerance μας δείχνει αν υπάρχει πολυσυγραμμικότητα στο μοντέλο. Για τον δείκτη Tolerance όλες οι τιμές είναι > 0,2 και για τον δείκτη VIF όλες οι τιμές είναι < 5, άρα δεν υπάρχει ένδειξη πολυσυγραμμικότητας.

Residuals Statistics(a)

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	$12,941.10	$130,789.61	$34,419.57	$15,678.096	474
Std. Predicted Value	-1,370	6,147	,000	1,000	474
Standard Error of Predicted Value	481,450	3518,579	789,466	248,130	474
Adjusted Predicted Value	$12,830.26	$129,255.73	$34,414.28	$15,668.857	474
Residual	-$22,924.045	$46,564.902	$.000	$6,765.762	474
Std. Residual	-3,367	6,839	,000	,994	474
Stud. Residual	-3,492	6,864	,000	1,003	474
Deleted Residual	-$24,661.359	$46,916.438	$5.289	$6,896.929	474
Stud. Deleted Residual	-3,535	7,232	,002	1,016	474
Mahal. Distance	1,367	125,306	5,987	6,964	474
Cook's Distance	,000	,132	,003	,010	474
Centered Leverage Value	,003	,265	,013	,015	474

a Dependent Variable: Current Salary

Στον πίνακα **Residuals Statistics** κάνουμε έλεγχο για ύπαρξη ή όχι ετεροσκεδαστικότητας.Συγκρίνω το μέσο του Cook’s Distance με το 4/(n-k-1), δηλαδή 0,003 με 0,00856531,επειδή 0,003 < 0,00856531 δεν υπάρχει ετεροσκεδαστικότητα στο μοντέλο.

7.2 ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ ΜΕ ΤΗ ΜΕΘΟΔΟ STEPWISE

Θα ακολουθήσουμε την ίδια διαδικασία που αναφέραμε στο Κεφάλαιο 7.1, στο παράθυρο Dependent θα μεταφέρουμε την μεταβλητή Current salary, στο παράθυρο Independent(s) θα εισάγουμε όλες τις μεταβλητές και αντί για τη μέθοδο Enter επιλέγουμε τη μέθοδο Stepwise. Στην Εικόνα 4 στα Statistics, Save και Options αφήνουμε τις επιλογές που επιλέξαμε. Έτσι έχουμε τους επόμενους πίνακες στο Output.

7.2.1 ΕΜΦΑΝΙΣΗ ΤΩΝ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΣΤΟ OUTPUT

Regression

Variables Entered/Removed(a)

Model	Variables Entered	Variables Removed	Method
1	Beginning salary	,	Stepwise (Criteria: Probability-of-F-to-enter <=,050, Probability-of-F-to-remove >=,100).
2	Employment category	,	Stepwise (Criteria: Probability-of-F-to-enter <=,050, Probability-of-F-to-remove >=,100).
3	Previous experience (months)	,	Stepwise (Criteria: Probability-of-F-to-enter <=,050, Probability-of-F-to-remove >=,100).
4	Months since hire	,	Stepwise (Criteria: Probability-of-F-to-enter <=,050, Probability-of-F-to-remove >=,100).
5	Educational level (years)	,	Stepwise (Criteria: Probability-of-F-to-enter <=,050, Probability-of-F-to-remove >=,100).
6	Gender	,	Stepwise (Criteria: Probability-of-F-to-enter <=,050, Probability-of-F-to-remove >=,100).

a Dependent Variable: Current salary

Η μέθοδος Stepwise επιλέγει τις καλύτερες μεταβλητές για το μοντέλο μας. Αν κοιτάξουμε τον παραπάνω πίνακα θα δούμε ότι οι πιο σημαντικές μεταβλητές για το μοντέλο μας είναι οι εξής: Beginning salary, Employment category, Previous experience, Months since hire, Educational level και Gender.

Model Summary(g)

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,880a	,775	,774	$8,119.79	
2	,898b	,806	,805	$7,548.01	
3	,909c	,827	,826	$7,133.58	
4	,914d	,836	,835	$6,947.63	
5	,917e	,840	,839	$6,863.87	
6	,918f	,843	,841	$6,815.68	1,850

a Predictors: (Constant), Beginning salary
b Predictors: (Constant), Beginning salary, Employment category
c Predictors: (Constant), Beginning salary, Employment category, Previous experience (months)
d Predictors: (Constant), Beginning salary, Employment category, Previous experience (months), Months since hire
e Predictors: (Constant), Beginning salary, Employment category, Previous experience (months), Months since hire, Educational level (years)
f Predictors: (Constant), Beginning salary, Employment category, Previous experience (months), Months since hire, Educational level (years), Gender
g Dependent Variable: Current salary

Στον πίνακα **Model Summary** έχουμε τα κριτήρια $\mathbf{R^2}$ και $\mathbf{R^2_{adj}}$. Στο μοντέλο με τις 6 καλύτερες μεταβλητές το $\mathbf{R^2}$ = 0,843 και το $\mathbf{R^2_{adj}}$ = 0,841, τα οποία είναι και τα δύο ικανοποιητικά. Ο δείκτης **Durbin-Watson** είναι σε καλό επίπεδο και είναι ίσο με 1,850, δηλαδή δεν υπάρχει αυτοσυσχέτιση.
Επίσης να αναφέρουμε ότι και η μέθοδος **Enter** αλλά και η μέθοδος **Stepwise** επέλεξαν τις ίδιες μεταβλητές για το καλύτερο μοντέλο και τα $\mathbf{R^2}$, $\mathbf{R^2_{adj}}$ και ο δείκτης **Durbin-Watson** είναι ίδια.

ANOVA(g)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	106862706669,340	1	106862706669,340	1620,826	,000a
	Residual	31053506813,535	471	65931012,343		
	Total	137916213482,875	472			
2	Regression	111139190822,232	2	55569595411,116	975,378	,000b
	Residual	26777022660,643	470	56972388,640		
	Total	137916213482,875	472			
3	Regression	114049771385,160	3	38016590461,720	747,065	,000c
	Residual	23866442097,715	469	50887936,242		
	Total	137916213482,875	472			
4	Regression	115326042679,617	4	28831510669,904	597,302	,000d
	Residual	22590170803,259	468	48269595,733		
	Total	137916213482,875	472			
5	Regression	115914580758,552	5	23182916151,710	492,074	,000e
	Residual	22001632724,323	467	47112703,906		
	Total	137916213482,875	472			
6	Regression	116268896465,578	6	19378149410,930	417,152	,000f
	Residual	21647317017,297	466	46453469,994		
	Total	137916213482,875	472			

Στον πίνακα **ANOVA** βρίσκουμε αν είναι στατιστικά σημαντικό το μοντέλο μας. Κοιτάμε την τελευταία γραμμή. Κάνουμε τον έλεγχο :

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$ VS $H_1 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 \neq 0$

Τα $\beta_1, \beta_2, \ldots, \beta_6$ είναι εκτιμητές.

Στη στήλη Sig το p_value είναι 0 που είναι < 0,05, δηλαδή το μοντέλο μας είναι στατιστικά σημαντικό.

Έχουμε το κριτήριο **MSR**(μέσο τετράγωνο υπολοίπων) = 46453469,994 και το κριτήριο **Cp** που μπορούμε να το υπολογίσουμε και είναι $\mathbf{C_p}$ = 466.

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	1929,517	889,168		2,170	,031		
	Beginning salary	1,910	,047	,880	40,259	,000	1,000	1,000
2	(Constant)	1038,773	832,923		1,247	,213		
	Beginning salary	1,469	,067	,677	21,829	,000	,429	2,330
	Employment category	5937,464	685,314	,269	8,664	,000	,429	2,330
3	(Constant)	3043,572	830,627		3,664	,000		
	Beginning salary	1,468	,064	,677	23,080	,000	,429	2,330
	Employment category	6146,203	648,274	,278	9,481	,000	,428	2,334
	Previous experience (months)	-23,768	3,143	-,146	-7,563	,000	,996	1,004
4	(Constant)	-10293,847	2717,031		-3,789	,000		
	Beginning salary	1,479	,062	,681	23,852	,000	,429	2,332
	Employment category	6059,134	631,603	,274	9,593	,000	,428	2,336
	Previous experience (months)	-23,791	3,061	-,146	-7,773	,000	,996	1,004
	Months since hire	163,747	31,845	,096	5,142	,000	,999	1,001
5	(Constant)	-15014,828	2998,243		-5,008	,000		
	Beginning salary	1,366	,069	,629	19,773	,000	,337	2,966
	Employment category	5852,463	626,722	,265	9,338	,000	,424	2,356
	Previous experience (months)	-19,552	3,253	-,120	-6,010	,000	,861	1,162
	Months since hire	154,277	31,575	,091	4,886	,000	,992	1,008
	Educational level (years)	540,886	153,034	,091	3,534	,000	,512	1,954
6	(Constant)	-11762,219	3201,675		-3,674	,000		
	Beginning salary	1,329	,070	,612	19,014	,000	,325	3,079
	Employment category	5790,385	622,728	,262	9,298	,000	,424	2,359
	Previous experience (months)	-21,455	3,303	-,131	-6,496	,000	,823	1,215
	Months since hire	147,805	31,441	,087	4,701	,000	,986	1,014
	Educational level (years)	473,577	153,901	,080	3,077	,002	,499	2,004
	Gender	-2015,028	729,617	-,059	-2,762	,006	,744	1,345

a Dependent Variable: Current salary

Στον πίνακα **Coefficients** κάνουμε έλεγχο για τη στατιστική σημαντικότητα των μεταβλητών. Ο έλεγχος είναι :

$H_0 : \hat{\beta}_i = 0$ VS $H_1 : \hat{\beta}_i \neq 0$

Η στήλη Sig μας δίνει τα p_value όπου για όλες τις μεταβλητές είναι < 0,05. Άρα όλες οι μεταβλητές είναι στατιστικά σημαντικές.
Η στήλη Collinearity Statistics με τους δείκτες VIF και Tolerance μας δείχνει αν υπάρχει πολυσυγραμμικότητα στο μοντέλο. Για τον δείκτη Tolerance όλες οι τιμές είναι > 0,2 και για τον δείκτη VIF όλες οι τιμές είναι < 5, άρα δεν υπάρχει ένδειξη πολυσυγραμμικότητας.

Excluded Variables(g)

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	Age	-,136(a)	-6,471	,000	-,286	1,000	1,000	1,000
	Educational level (years)	,173(a)	6,385	,000	,283	,599	1,669	,599
	Gender	-,061(a)	-2,509	,012	-,115	,792	1,263	,792
	Employment category	,269(a)	8,664	,000	,371	,429	2,330	,429
	Months since hire	,101(a)	4,707	,000	,212	1,000	1,000	1,000
	Minority classification	-,040(a)	-1,809	,071	-,083	,975	1,025	,975
	Previous experience (months)	-,138(a)	-6,571	,000	-,290	,998	1,002	,998
2	Age	-,140(b)	-7,273	,000	-,318	,999	1,001	,429
	Educational level (years)	,157(b)	6,212	,000	,276	,596	1,678	,348
	Gender	-,050(b)	-2,186	,029	-,100	,789	1,267	,396
	Months since hire	,096(b)	4,834	,000	,218	,999	1,001	,429
	Minority classification	-,033(b)	-1,601	,110	-,074	,974	1,027	,427
	Previous experience (months)	-,146(b)	-7,563	,000	-,330	,996	1,004	,428
3	Age	-,067(c)	-2,068	,039	-,095	,354	2,823	,353
	Educational level (years)	,102(c)	3,870	,000	,176	,516	1,940	,339
	Gender	-,078(c)	-3,614	,000	-,165	,769	1,301	,395
	Months since hire	,096(c)	5,142	,000	,231	,999	1,001	,428
	Minority classification	-,010(c)	-,521	,602	-,024	,950	1,052	,427
4	Age	-,081(d)	-2,578	,010	-,118	,352	2,844	,351
	Educational level (years)	,091(d)	3,534	,000	,161	,512	1,954	,337
	Gender	-,069(d)	-3,261	,001	-,149	,763	1,311	,393
	Minority classification	-,015(d)	-,778	,437	-,036	,948	1,055	,426
5	Age	-,068(e)	-2,184	,029	-,101	,346	2,888	,337
	Gender	-,059(e)	-2,762	,006	-,127	,744	1,345	,325
	Minority classification	-,016(e)	-,841	,401	-,039	,948	1,055	,335
6	Age	-,046(f)	-1,406	,160	-,065	,313	3,194	,299
	Minority classification	-,024(f)	-1,266	,206	-,059	,927	1,078	,322

a Predictors in the Model: (Constant), Beginning salary
b Predictors in the Model: (Constant), Beginning salary, Employment category
c Predictors in the Model: (Constant), Beginning salary, Employment category, Previous experience (months)
d Predictors in the Model: (Constant), Beginning salary, Employment category, Previous experience (months), Months since hire
e Predictors in the Model: (Constant), Beginning salary, Employment category, Previous experience (months), Months since hire, Educational level (years)
f Predictors in the Model: (Constant), Beginning salary, Employment category, Previous experience (months), Months since hire, Educational level (years), Gender
g Dependent Variable: Current salary

Στον παραπάνω πίνακα βλέπουμε στην τελευταία γραμμή ότι οι μεταβλητές **Age** και **Minority classification** δεν είναι στατιστικά σημαντικές για το μοντέλο μας και έτσι η μέθοδος Stepwise τις απορρίπτει αυτόματα από το μοντέλο.

Casewise Diagnostics(a)

Case Number	Std. Residual	Current salary
18	6,041	$103,750
32	3,610	$110,625
103	3,504	$97,000
106	3,595	$91,250
205	-3,364	$66,750
218	6,830	$80,000
274	4,437	$83,750
446	3,111	$100,000
454	3,708	$90,625

a Dependent Variable: Current salary

Residuals Statistics(a)

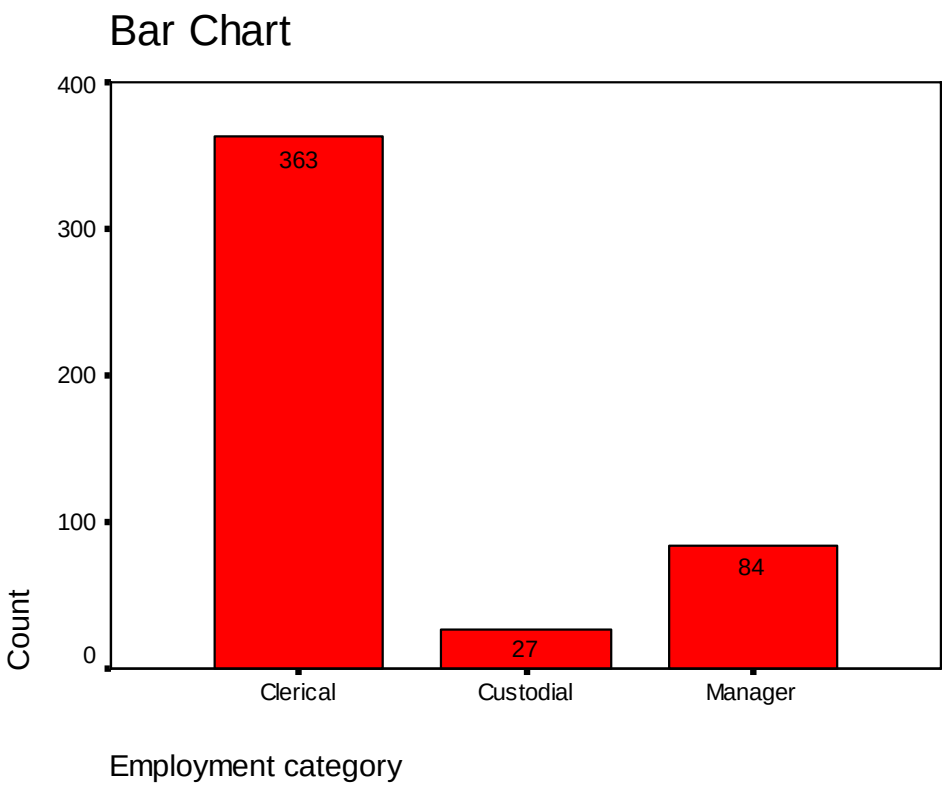
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	$12,925.00	$130,804.26	$34,424.08	$15,678.86	474
Std. Predicted Value	-1,369	6,141	,000	,999	474
Standard Error of Predicted Value	$481.99	$3,522.30	$791.14	$248.35	474
Adjusted Predicted Value	$12,813.45	$129,275.34	$34,418.78	$15,669.72	474
Residual	-$22,930.74	$46,549.01	-$4.51	$6,765.77	474
Std. Residual	-3,364	6,830	-,001	,993	474
Stud. Residual	-3,490	6,856	,000	1,002	474
Deleted Residual	-$24,668.83	$46,903.07	$.79	$6,897.12	474
Stud. Deleted Residual	-3,532	7,222	,002	1,015	474
Mahal. Distance	1,363	125,062	5,987	6,951	474
Cook's Distance	,000	,132	,003	,010	474
Centered Leverage Value	,003	,265	,013	,015	474

a Dependent Variable: Current salary

Στον πίνακα **Residuals Statistics** κάνουμε έλεγχο για ύπαρξη ή όχι ετεροσκεδαστικότητας. Συγκρίνω το μέσο του Cook's Distance με το 4/(n-k-1), δηλαδή 0,003 με 0,00856531,επειδή 0,003 < 0,00856531 δεν υπάρχει ετεροσκεδαστικότητα στο μοντέλο.

7.3 ΓΡΑΦΙΚΗ ΠΑΡΟΥΣΙΑΣΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Για να δημιουργήσουμε ένα **Bar Chart** επιλέγουμε από το μενού **SPSS Data Editor** το εξής : **Graphs/ Bar**. Στο παράθυρο **Bar Charts** επιλέγουμε το *Simple* σαν τύπο και στο **Data in Chart Are** επιλέγουμε το *Summaries for groups of cases* και στη συνέχεια κάνουμε κλικ στο **Define**.
Θα εμφανιστεί το παράθυρο **Define Simple Bar** όπου στη περιοχή **Bars Represent** επιλέγουμε το *N (Number) of cases* και στη περιοχή **Category Axis** μεταφέρουμε την μεταβλητή ***Jobcat***. Στη συνέχεια πατάμε *OK* και εμφανίζεται στο Output το παρακάτω διάγραμμα :

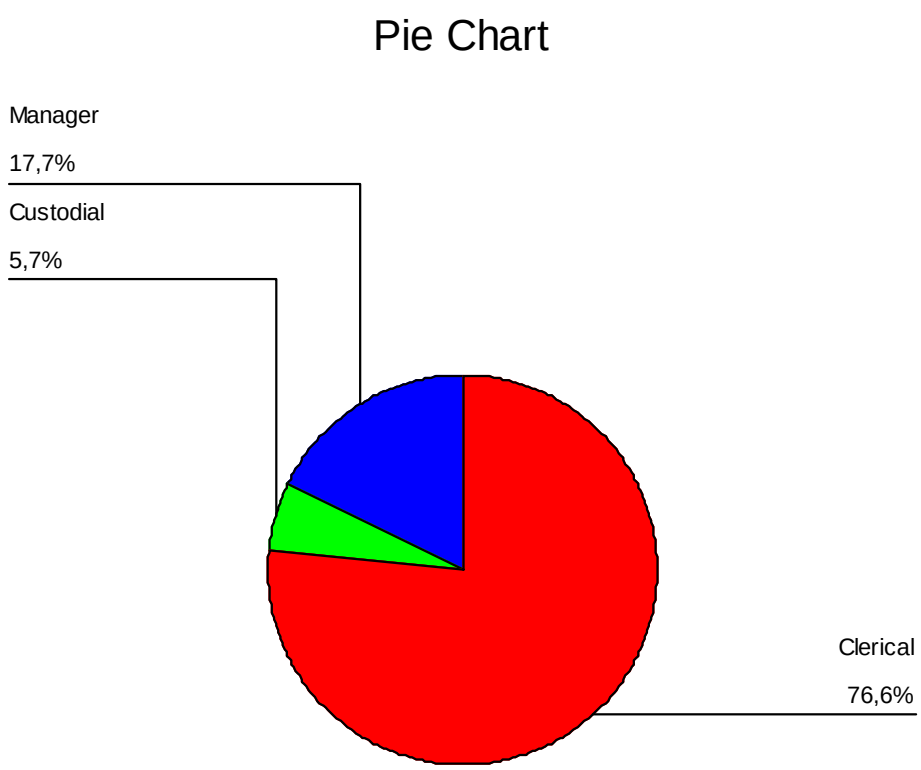


Για να βρούμε πόσοι *Clerical*, *Custodial* και *Manager* αντιστοιχούν σε κάθε στήλη κάνουμε διπλό κλικ στο διάγραμμα. Στο παράθυρο που θα εμφανιστεί επιλέγουμε από το μενού **Chart Edit** το εξής : **Format** / **Bar Label Style**. Στο παράθυρο **Bar Label Style** επιλέγουμε την εικόνα *Standard* και στη συνέχεια πατάμε *Apply All* και *Close.*

Για να βάλουμε τίτλο στο διάγραμμα, από το μενού **Chart Edit**or επιλέγουμε **Chart** / **Title**. Στο *Title 1 textbox* γράφουμε ***Bar Chart*** και πατάμε *OK*.

Από το παραπάνω διάγραμμα παρατηρούμε ότι υπάρχουν 363 εργαζόμενοι στη κατηγορία *Clerical*, 27 εργαζόμενοι στη κατηγορία *Custodial* και 84 εργαζόμενοι στη κατηγορία *Manager*.

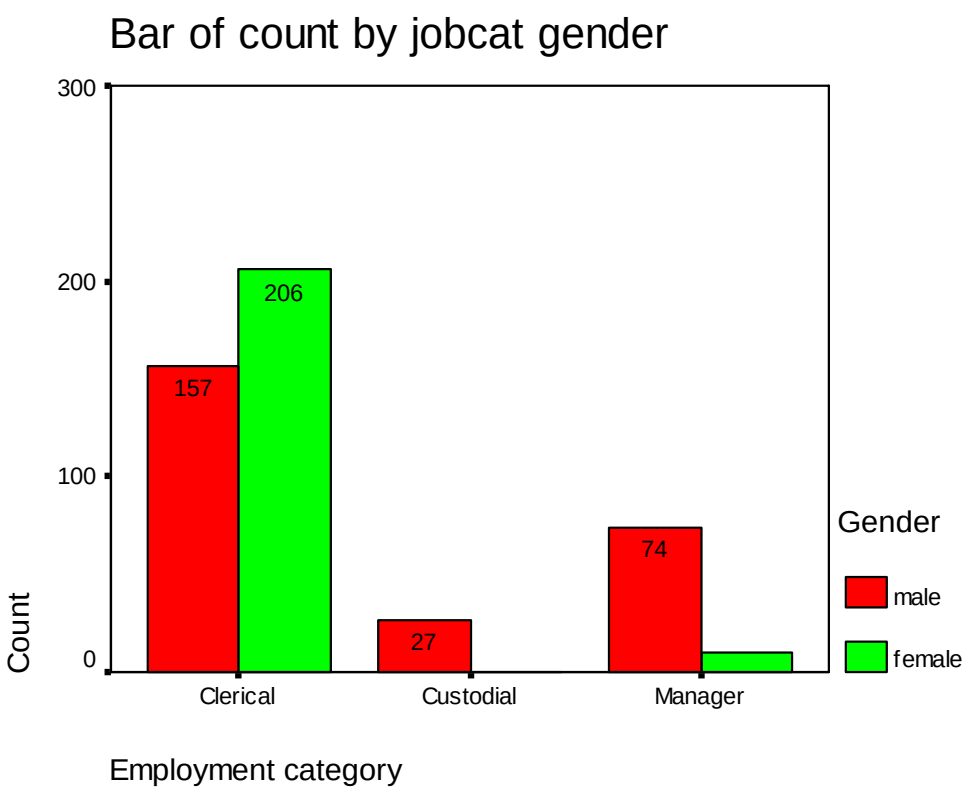
Για να δημιουργήσουμε ένα **Pie Chart** επιλέγουμε από το μενού **SPSS Data Editor** το εξής : **Graphs** / **Pie**.
Επιλέγουμε την εικόνα *Simple* και πατάμε **Define**. Στο παράθυρο που θα ανοίξει στην περιοχή **Slices Represent** επιλέγουμε το *% of cases* και στη περιοχή **Define Slices by** μεταφέρουμε την μεταβλητή ***Jobcat*** από τη λίστα. Στη συνέχεια πατάμε *OK* και εμφανίζεται το παρακάτω διάγραμμα.



Για να βρούμε τι ποσοστό έχει η κάθε κατηγορία κάνουμε διπλό κλικ στο διάγραμμα και από το μενού **Chart Editor** επιλέγουμε : **Chart** / **Options**. Κάνουμε κλικ στο *Percents* και πατάμε *OK*.
Βλέπουμε ότι το 76,6% των εργαζομένων ανήκει στη κατηγορία *Clerical*, το 17,7% ανήκει στη κατηγορία *Manager* και το 5,7% ανήκει στη κατηγορία *Custodial*.

Για να εξετάσουμε περισσότερες μεταβλητές από το μενού **SPSS Data Editor** επιλέγουμε **Graphs** / **Bar**, στη συνέχεια επιλέγουμε την εικόνα *Clustered* και πατάμε **Define**. Στη περιοχή **Category Axis** μεταφέρουμε την μεταβλητή ***Jobcat*** από τη λίστα και στη περιοχή **Define**

Clusters by μεταφέρουμε την μεταβλητή ***Gender***. Στη περιοχή **Bar Represent** αφήνουμε την επιλογή *N (Number) of cases* και πατάμε *OK*.



Παρατηρούμε ότι : Στη κατηγορία *Clerical* υπάρχουν 157 άνδρες και 206 γυναίκες, στη κατηγορία *Custodial* υπάρχουν 27 άνδρες και στη κατηγορία *Manager* υπάρχουν 74 άνδρες και 10 γυναίκες. Επίσης παρατηρούμε ότι στη κατηγορία *Custodial* υπάρχουν μόνο άνδρες και στη κατηγορία *Manager* υπάρχουν πολύ περισσότεροι άνδρες ***managers*** παρά γυναίκες ***managers***.
Αν θέλουμε να δημιουργήσουμε και ένα πίνακα με τα στοιχεία του διαγράμματος επιλέγουμε από το μενού : **Analyze / Descriptive Statistics / Crosstabs**. Στη περιοχή **Row(s)** μεταφέρουμε την μεταβλητή ***Gender*** και στη περιοχή **Column(s)** μεταφέρουμε την μεταβλητή ***Jobcat***. Αν πατήσουμε *OK* τότε στο Output θα έχουμε τους παρακάτω πίνακες :

Case Processing Summary

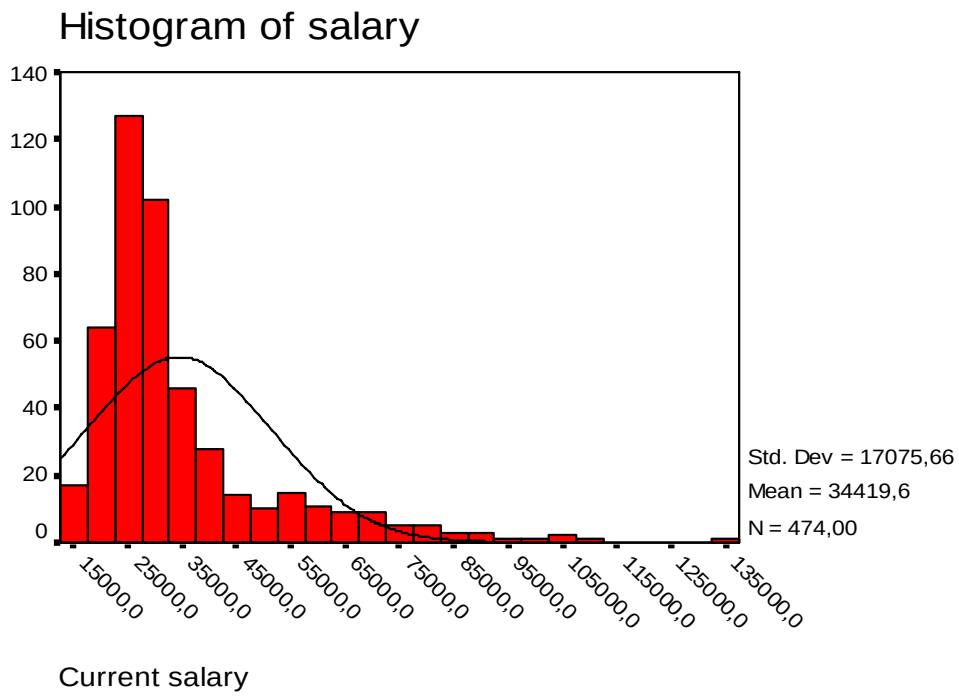
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Gender * Employment category	474	100,0%	0	,0%	474	100,0%

Gender * Employment category Crosstabulation

Count

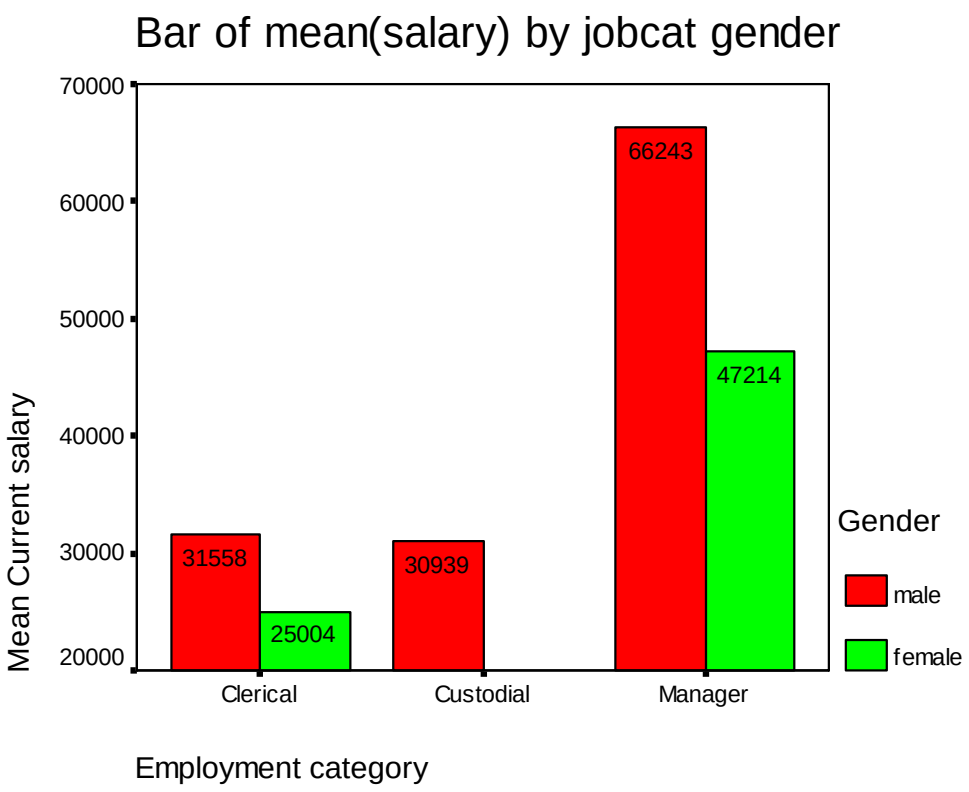
		Employment category			Total
		Clerical	Custodial	Manager	Clerical
Gender	male	157	27	74	258
	female	206		10	216
Total		363	27	84	474

Για να δημιουργήσουμε ένα ιστόγραμμα για την μεταβλητή ***Salary*** επιλέγουμε από το μενού **Graphs/Histogram**. Στη περιοχή **Variable list** μεταφέρουμε την μεταβλητή ***Salary***, επιλέγουμε το **Display normal curve** και στη συνέχεια πατάμε *OK*.



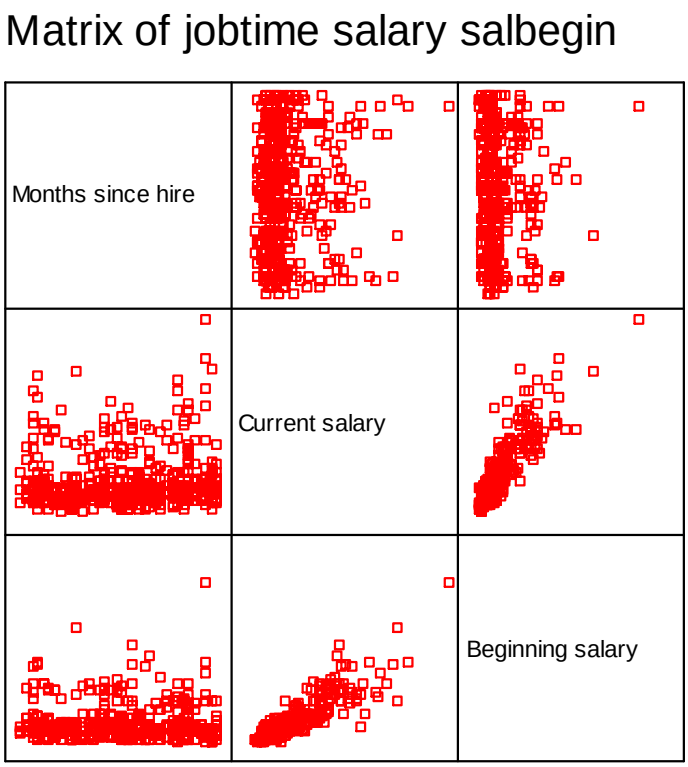
Παρατηρούμε ότι τα δεδομένα μας ομαδοποιούνται σε διαστήματα. Οι αριθμοί κάτω από τις στήλες υποδεικνύουν τη μέση τιμή για κάθε διάστημα. Για παράδειγμα η μέση τιμή του διαστήματος μεταξύ 12500 και 17500 είναι 15000.
Κάθε στήλη αντιπροσωπεύει τον αριθμό των κελιών. Καμία στήλη δεν είναι τυπωμένη για τα διαστήματα που δεν έχουν καμία τιμή. Επίσης παρατηρούμε ότι η κατανομή δεν είναι συμμετρική. Η αμερόληπτη τυπική απόκλιση είναι **s** = 17075.66 (το SPSS υπολογίζει μόνο την αμερόληπτη τυπική απόκλιση) και ο μέσος είναι **M** = 34419.6.
Με την ίδια διαδικασία μπορούμε να κάνουμε ιστόγραμμα και για τις υπόλοιπες μεταβλητές.

Το παρακάτω διάγραμμα μπορούμε να το υπολογίσουμε επιλέγοντας από το μενού **SPSS Data Editor** το εξής : **Graphs / Bar**. Επιλέγουμε την εικόνα *Simple* και στην περιοχή **Data in Chart Are** αφήνουμε το *Summaries for groups of cases* και πατάμε *OK*. Στο παράθυρο που θα ανοίξει στην περιοχή **Category Axis** μεταφέρουμε την μεταβλητή ***Jobcat*** από τη λίστα, στη περιοχή **Bars Represent** επιλέγουμε το *Others summary function* και μεταφέρουμε την μεταβλητή ***Salary*** από τη λίστα. Στη συνέχεια πατάμε *OK*.



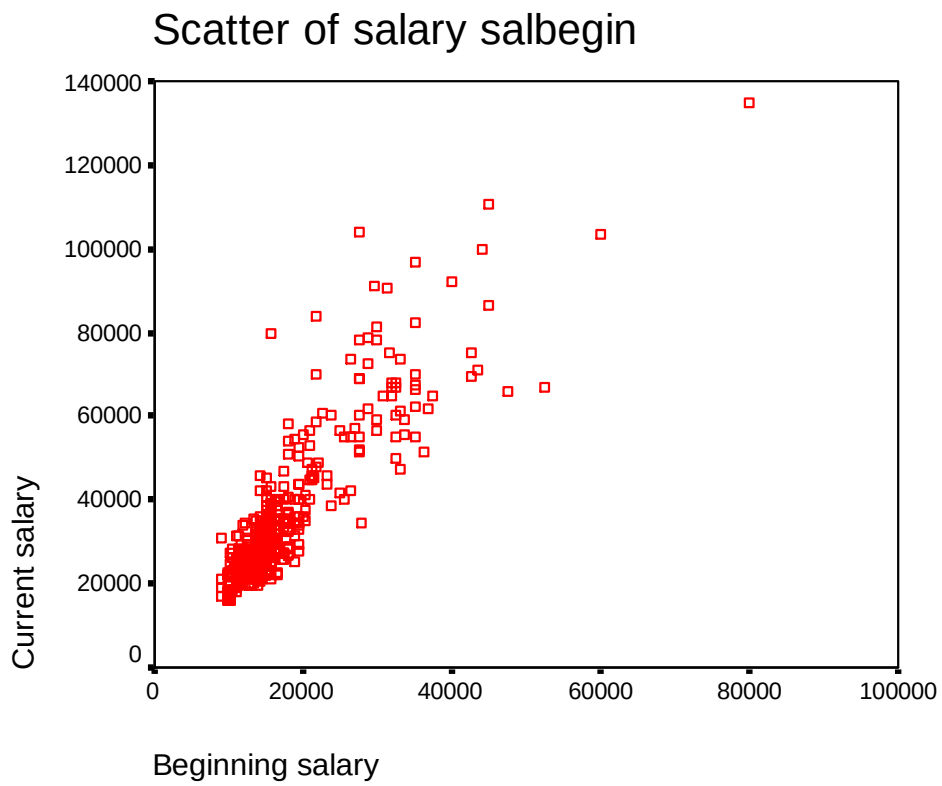
Παρατηρούμε ότι στη κατηγορία *Clerical* ο μέσος μισθός για τους άνδρες είναι 31558$ και για τις γυναίκες είναι 25004$. Στη κατηγορία *Custodial* ο μέσος μισθός για τους άνδρες είναι 30939$ και όπως αναφέραμε παραπάνω δεν υπάρχουν γυναίκες σε αυτή την κατηγορία. Στη κατηγορία *Manager* ο μέσος μισθός για τους άνδρες είναι 66243$ και για τις γυναίκες είναι 47214$.

Ο μέσος μισθός για την κατηγορία *Manager* και για τους άνδρες και για τις γυναίκες είναι σχεδόν ο διπλάσιος από τις άλλες κατηγορίες. Επίσης ο μέσος μισθός των ανδρών στις κατηγορίες *Clerical* και *Manager* είναι παραπάνω από αυτόν των γυναικών.

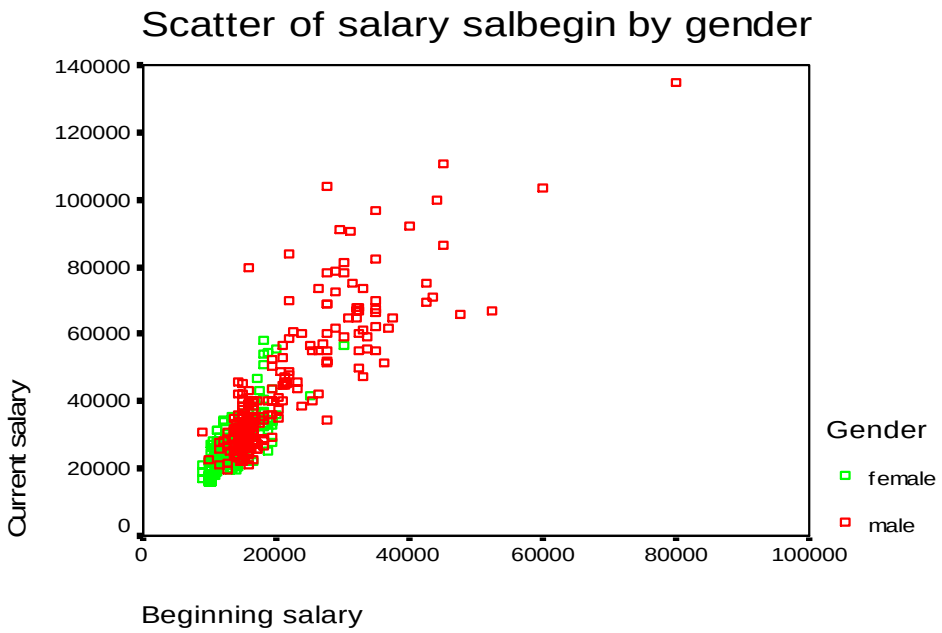


Στο παραπάνω διάγραμμα μπορούμε να δούμε τη σχέση μεταξύ των μεταβλητών ***Jobtime***, ***Salary*** και ***Salbegin***. Μπορούμε να το υπολογίσουμε αν από το μενού επιλέξουμε **Graphs / Scatter** και στο παράθυρο **Scatterplot** επιλέξουμε την εικόνα *Matrix*.

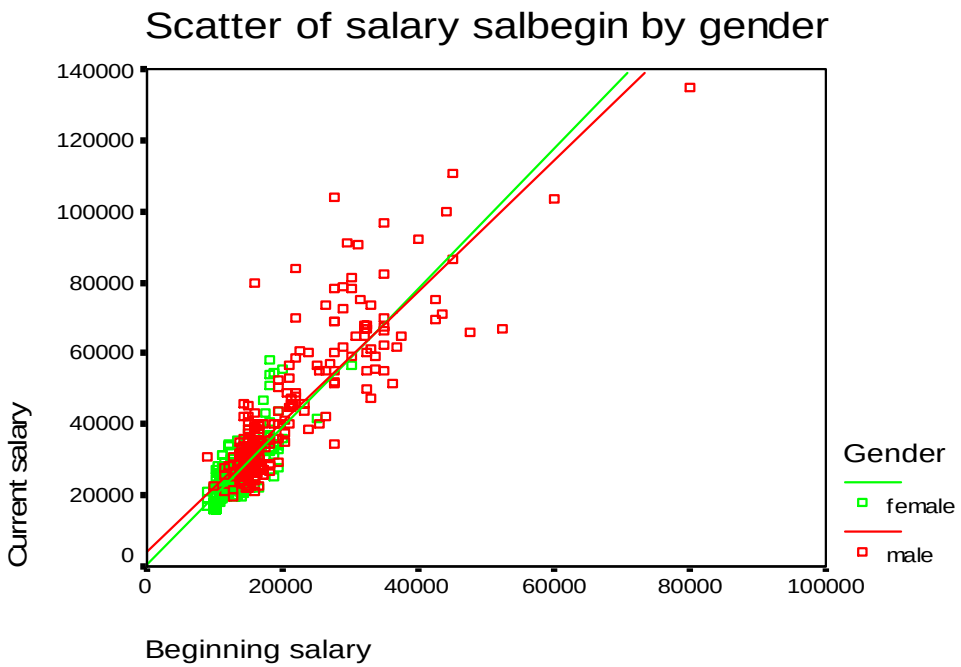
Αν θέλουμε να υπολογίσουμε τη σχέση μεταξύ δυο μόνο μεταβλητών τότε στο παράθυρο **Scatterplot** επιλέγουμε την εικόνα *Simple*. Στο παράθυρο **Simple Scatterlot** στη περιοχή **Y Axis** μεταφέρουμε την μεταβλητή ***Salary*** και στη περιοχή **X Axis** μεταφέρουμε την μεταβλητή ***Salbegin***. Αν πατήσουμε *OK* θα έχουμε το παρακάτω διάγραμμα.



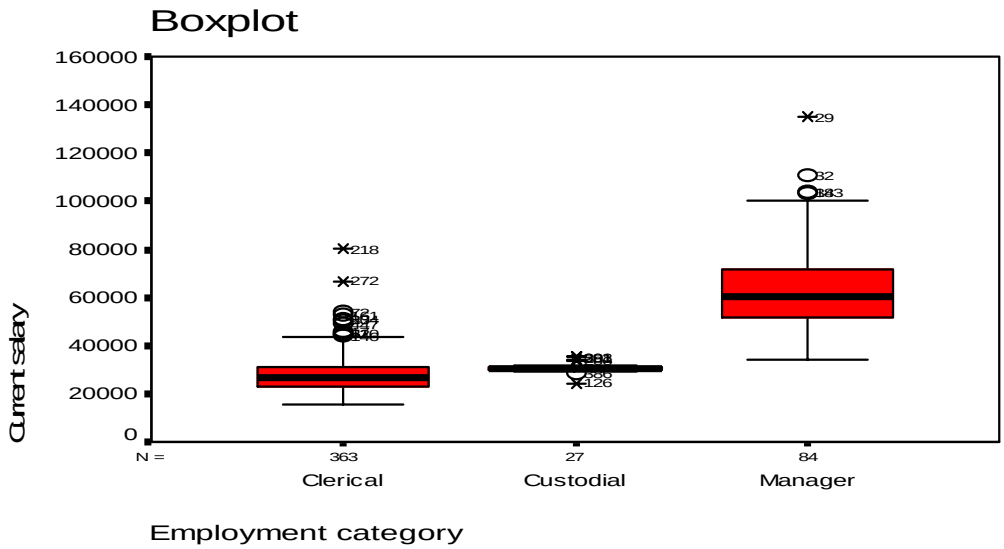
Επίσης μπορούμε να υπολογίσουμε τη σχέση μεταξύ των μεταβλητών ***Salary*** και ***Salbegin*** κατά άνδρες και γυναίκες. Αυτό μπορεί να γίνει αν στο παράθυρο **Simple Scatterplot** και στη περιοχή **Set Markers by** μεταφέρουμε την μεταβλητή ***Gender*** και στη συνέχεια πατάμε *OK*.



Αν κάνουμε διπλό κλικ στο διάγραμμα *Scatter of salary salbegin by gender* και επιλέξουμε από το μενού **Chart / Options** θα μας βγάλει το παράθυρο **Scatterplot Options**. Με την επιλογή **Fit Line** μπορούμε να δημιουργήσουμε μια γραμμή παλινδρόμησης στο διάγραμμα, κάνοντας κλικ στο *Fit Options* επιλέγουμε
την εικόνα *Linear regression* και στη συνέχεια πατάμε *OK*.

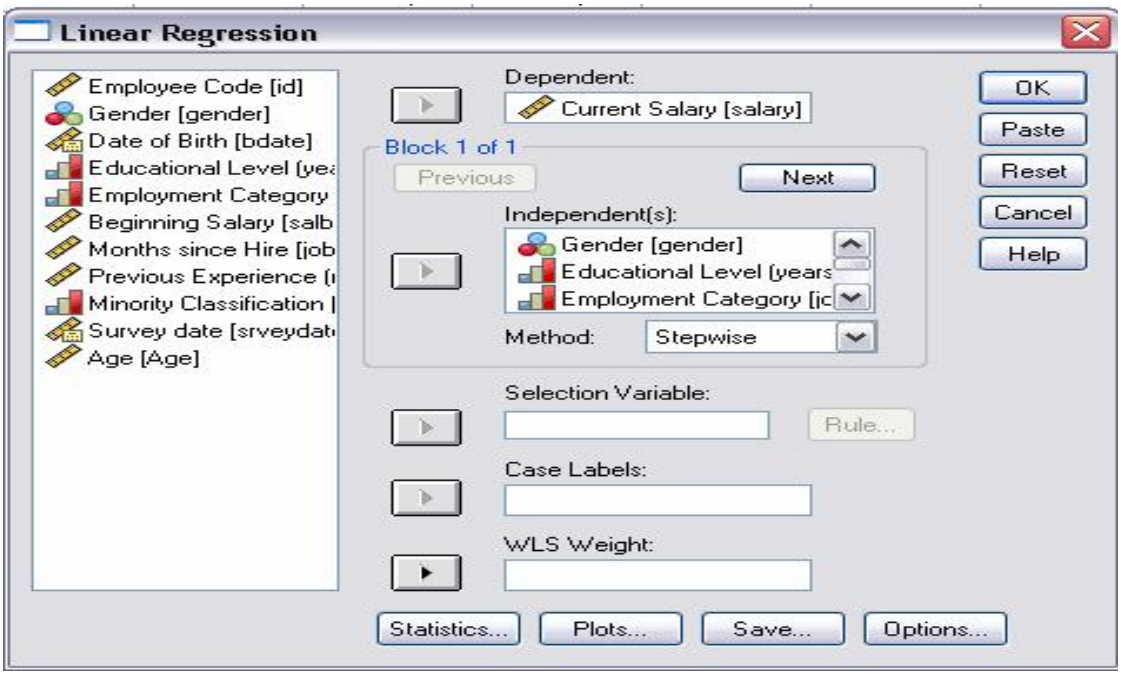


Από το μενού **SPSS Data Editor** επιλέγουμε **Graphs / Interactive / Boxplot**. Στο παράθυρο **Create Boxplot** μεταφέρουμε την μεταβλητή ***Salary*** στο **Y Axis** και την μεταβλητή ***Jobcat*** στο **X Axis**. Στη συνέχεια πατάμε *OK* και έχουμε το παρακάτω διάγραμμα.



Από το μήκος του κιβωτίου, μπορούμε να καθορίσουμε τη μεταβλητότητα. Όσο μεγαλύτερο το κιβώτιο, τόσο μεγαλύτερη η διάδοση των στοιχείων. Εδώ τη μεγαλύτερη μεταβλητότητα την έχει η κατηγορία *Manager*. Η οριζόντια γραμμή μέσα στο κιβώτιο αντιπροσωπεύει τη διάμεσο.

Από το μενού **SPSS Data Editor** επιλέγουμε **Analyze/Regression/Linear**. Όπως φαίνεται στο παρακάτω παράθυρο μεταφέρουμε την μεταβλητή *Current Salary* στο παράθυρο **Dependen**t και τις μεταβλητές *Gender, Educational Level, Employment Category, Beginning Salary, Months since Hire, Previous Experience, Minority Classification* και *Age* στο παράθυρο **Independent**. Στο παράθυρο **Method** επιλέγουμε *Stepwise*. Στην επιλογή **Plots** τσεκάρουμε το *Histogram* και το *Normal probability plot* από την περιοχή *Standardized Residual Plots*. Στην επιλογή **Save** επιλέγουμε τα *Unstandardized residuals*. Στη συνέχεια πατάμε Continue και OK.

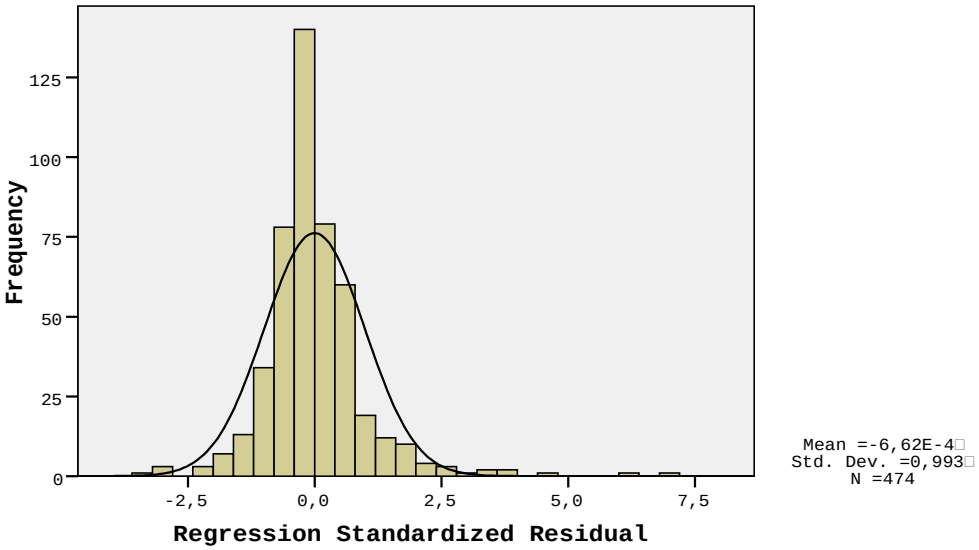


Εικόνα 9

Στα παρακάτω διαγράμματα μπορούμε να διαπιστώσουμε ότι τα υπόλοιπα ακολουθούν την κανονική κατανομή. Εφόσον οι τιμές των υπολοίπων είναι διάσπαρτες γύρω από τον άξονα του μηδενός σε μια ζώνη που πάνω –κάτω έχει το ίδιο πλάτος, κρίνουμε ότι το διάγραμμα υπολοίπων είναι ικανοποιητικό και δεν μας δείχνει κάποιο σοβαρό πρόβλημα του μοντέλου. Ανησυχούμε λιγάκι για το γεγονός ότι τα υπόλοιπα εμφανίζονται να ανοίγουν λίγο (σαν βεντάλια), όμως δεν μπορούμε να κρίνουμε κατά πόσο αυτός ο ετεροσκεδασμός είναι άξιος περαιτέρω μελέτης.

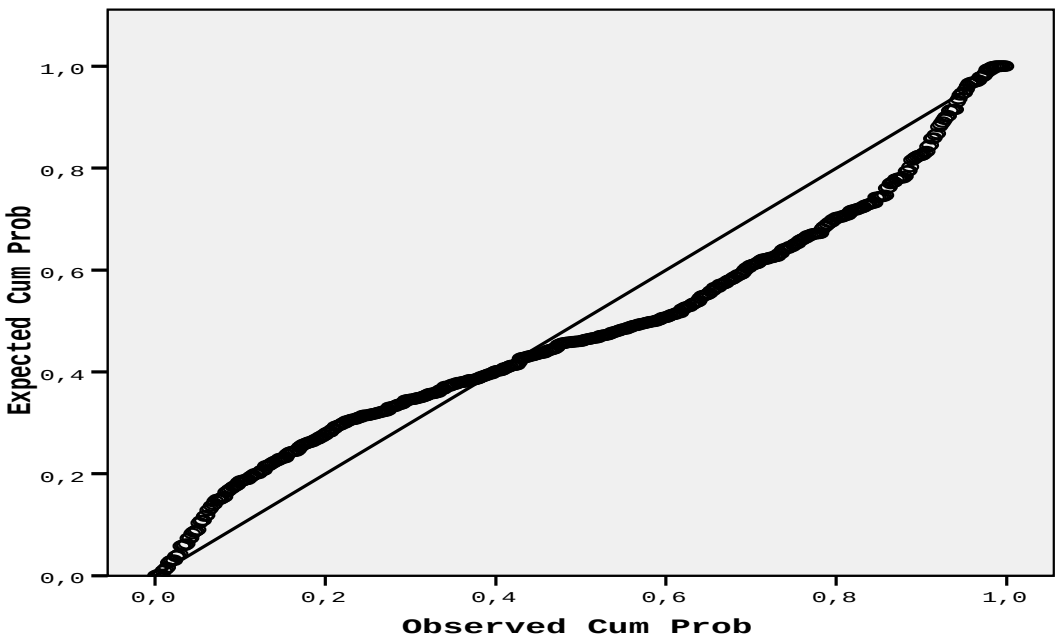
Histogram

Dependent Variable: Current Salary



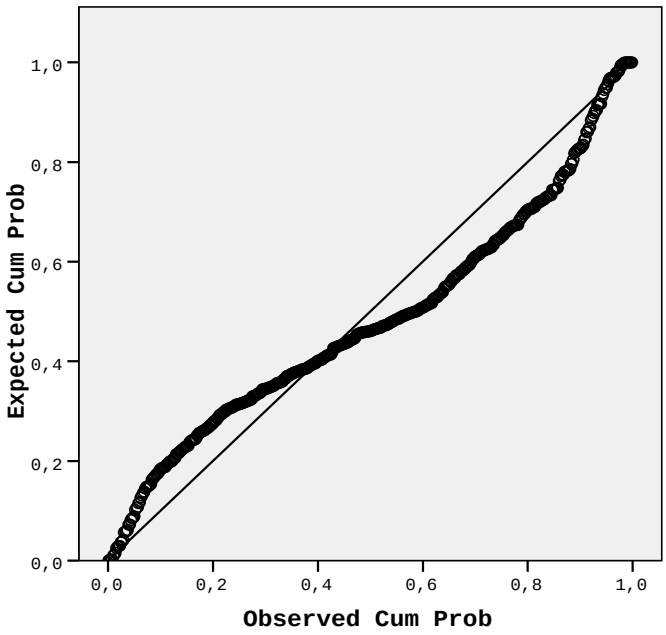
Normal P-P Plot of Regression Standardized Residual

Dependent Variable: Current Salary

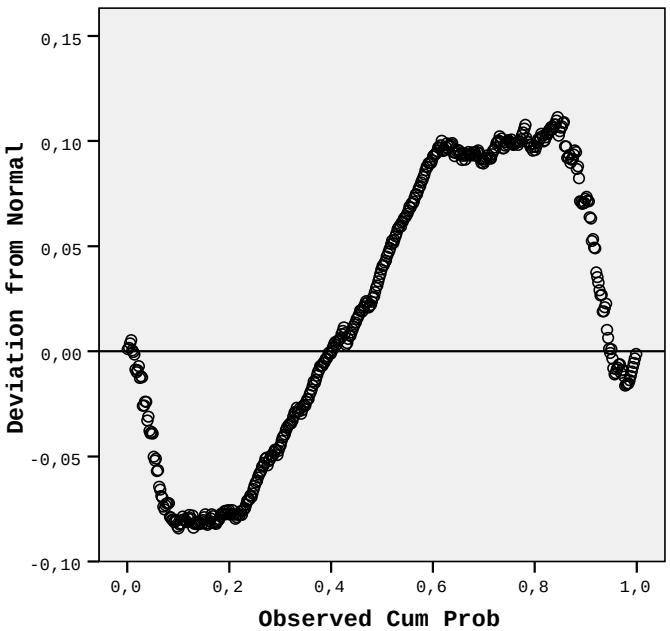


Από το μενού **SPSS Data Editor** επιλέγουμε **Analyze/ Descriptive Statistics/ P-P Plots** και **Q-Q Plots**. Στο παράθυρο **Variables** μεταφέρουμε την μεταβλητή *Unstanstardized residual* και στην περιοχή **Test Distribution** αφήνουμε την επιλογή *Normal*. Στην συνέχεια πατάμε OK.

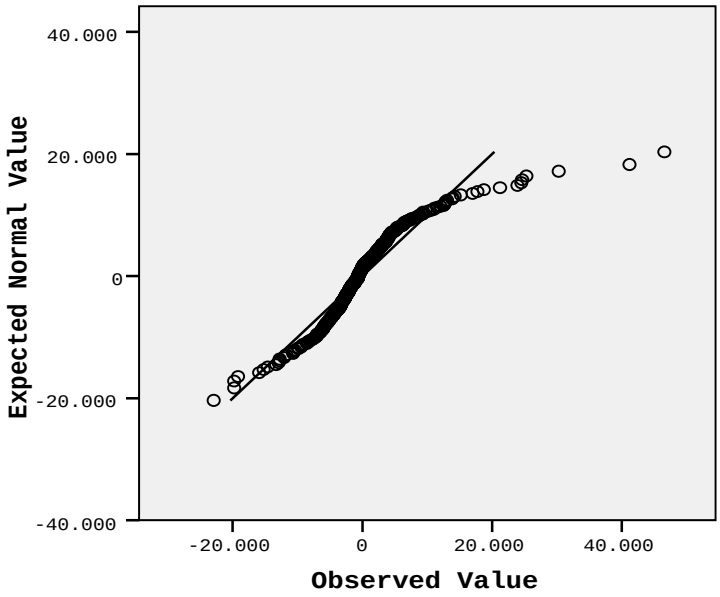
Normal P-P Plot of Unstandardized Residual



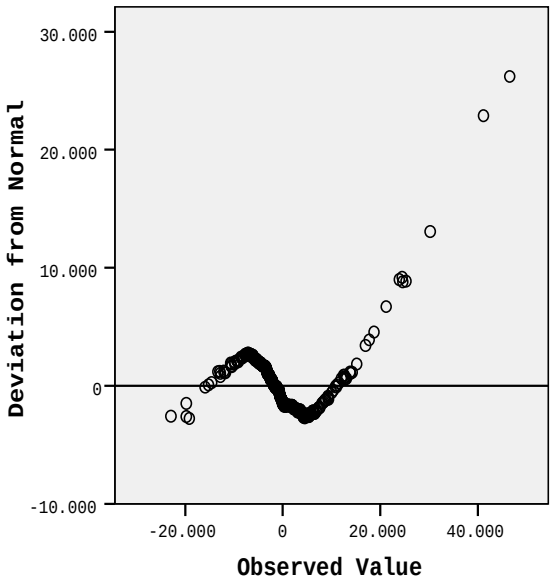
Detrended Normal P-P Plot of Unstandardized Residual



Normal Q-Q Plot of Unstandardized Residual



Detrended Normal Q-Q Plot of Unstandardized Residual



ΚΕΦΑΛΑΙΟ 8 ΕΦΑΡΜΟΓΗ ΣΤΟ R ΜΕ ΟΙΚΟΝΟΜΙΚΑ ΔΕΔΟΜΕΝΑ

8.1 Παράδειγμα freeny

R version 2.6.0 (2007-10-03)
Copyright (C) 2007 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> freeny

	y	lag.quarterly.revenue	price.index	income.level	market.potential
1962.25	8.79236	8.79636	4.70997	5.82110	12.9699
1962.5	8.79137	8.79236	4.70217	5.82558	12.9733
1962.75	8.81486	8.79137	4.68944	5.83112	12.9774
1963	8.81301	8.81486	4.68558	5.84046	12.9806
1963.25	8.90751	8.81301	4.64019	5.85036	12.9831
1963.5	8.93673	8.90751	4.62553	5.86464	12.9854
1963.75	8.96161	8.93673	4.61991	5.87769	12.9900
1964	8.96044	8.96161	4.61654	5.89763	12.9943
1964.25	9.00868	8.96044	4.61407	5.92574	12.9992
1964.5	9.03049	9.00868	4.60766	5.94232	13.0033
1964.75	9.06906	9.03049	4.60227	5.95365	13.0099
1965	9.05871	9.06906	4.58960	5.96120	13.0159
1965.25	9.10698	9.05871	4.57592	5.97805	13.0212
1965.5	9.12685	9.10698	4.58661	6.00377	13.0265
1965.75	9.17096	9.12685	4.57997	6.02829	13.0351
1966	9.18665	9.17096	4.57176	6.03475	13.0429
1966.25	9.23823	9.18665	4.56104	6.03906	13.0497
1966.5	9.26487	9.23823	4.54906	6.05046	13.0551
1966.75	9.28436	9.26487	4.53957	6.05563	13.0634
1967	9.31378	9.28436	4.51018	6.06093	13.0693
1967.25	9.35025	9.31378	4.50352	6.07103	13.0737
1967.5	9.35835	9.35025	4.49360	6.08018	13.0770
1967.75	9.39767	9.35835	4.46505	6.08858	13.0849
1968	9.42150	9.39767	4.44924	6.10199	13.0918
1968.25	9.44223	9.42150	4.43966	6.11207	13.0950
1968.5	9.48721	9.44223	4.42025	6.11596	13.0984
1968.75	9.52374	9.48721	4.41060	6.12129	13.1089
1969	9.53980	9.52374	4.41151	6.12200	13.1169
1969.25	9.58123	9.53980	4.39810	6.13119	13.1222
1969.5	9.60048	9.58123	4.38513	6.14705	13.1266
1969.75	9.64496	9.60048	4.37320	6.15336	13.1356
1970	9.64390	9.64496	4.32770	6.15627	13.1415
1970.25	9.69405	9.64390	4.32023	6.16274	13.1444
1970.5	9.69958	9.69405	4.30909	6.17369	13.1459
1970.75	9.68683	9.69958	4.30909	6.16135	13.1520

```
1971    9.71774              9.68683     4.30552      6.18231          13.1593
1971.25 9.74924              9.71774     4.29627      6.18768          13.1579
1971.5  9.77536              9.74924     4.27839      6.19377          13.1625
1971.75 9.79424              9.77536     4.27789      6.20030          13.1664
```

> summary(freeny)

```
       y          lag.quarterly.revenue  price.index     income.level
 Min.   :8.791   Min.   :8.791          Min.   :4.278   Min.   :5.821
 1st Qu.:9.045   1st Qu.:9.020          1st Qu.:4.392   1st Qu.:5.948
 Median :9.314   Median :9.284          Median :4.510   Median :6.061
 Mean   :9.306   Mean   :9.281          Mean   :4.496   Mean   :6.039
 3rd Qu.:9.591   3rd Qu.:9.561          3rd Qu.:4.605   3rd Qu.:6.139
 Max.   :9.794   Max.   :9.775          Max.   :4.710   Max.   :6.200

market.potential
 Min.   :12.97
 1st Qu.:13.01
 Median :13.07
 Mean   :13.07
 3rd Qu.:13.12
 Max.   :13.17
```

Όπου lag.quarterly.revenue είναι το τριμηνιαίο εισόδημα, price.index είναι ο δείχτης τιμών, income.level είναι το επίπεδο απολαβής, market.pontetial είναι το δυναμικό αγοράς και y είναι το εισόδημα.

> z <- lm(y~lag.quarterly.revenue + price.index + income.level + market.potential,data=freeny)
Ορίζουμε στο αντικείμενο z το μοντέλο με όλες τις μεταβλητές, ως εξαρτημένη την y και ως ανεξάρτητες τις μεταβλητές lag.quarterly.revenue, price.index, income.level και την market.potential. Πλέον έχουμε το μοντέλο μας στο αντικείμενο z.

> summary(z)
Η συνάρτηση summary μας δίνει μια αναλυτική περίληψη των αποτελεσμάτων της ανάλυσης παλινδρόμησης.

Call:
lm(formula = y ~ lag.quarterly.revenue + price.index + income.level + market.potential, data = freeny)

```
Residuals:
       Min          1Q      Median          3Q         Max
-0.0259426 -0.0101033  0.0003824  0.0103236  0.0267124

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           -10.4726     6.0217  -1.739   0.0911 .
lag.quarterly.revenue   0.1239     0.1424   0.870   0.3904
price.index            -0.7542     0.1607  -4.693 4.28e-05 ***
income.level            0.7675     0.1339   5.730 1.93e-06 ***
```

```
market.potential        1.3306     0.5093   2.613   0.0133 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.01473 on 34 degrees of freedom
Multiple R-Squared: 0.9981,     Adjusted R-squared: 0.9978
F-statistic:  4354 on 4 and 34 DF,  p-value: < 2.2e-16
```

> step(z,data=freeny)

Start: AIC=-324.36
y ~ lag.quarterly.revenue + price.index + income.level + market.potential

```
                         Df Sum of Sq     RSS     AIC
- lag.quarterly.revenue  1 0.0001642    0.01 -325.50
<none>                                  0.01 -324.36
- market.potential       1 0.0014805    0.01 -319.22
- price.index            1 0.0047767    0.01 -306.88
- income.level           1      0.01    0.01 -299.99
```

Step: AIC=-325.5
y ~ price.index + income.level + market.potential

```
                   Df Sum of Sq     RSS     AIC
<none>                           0.01 -325.50
- market.potential  1  0.004017    0.01 -310.84
- price.index       1      0.01    0.02 -297.40
- income.level      1      0.02    0.02 -283.59
```

Call:
lm(formula = y ~ price.index + income.level + market.potential, data = freeny)

Coefficients:

(Intercept)	price.index	income.level	market.potential
-13.3101	-0.8349	0.8456	1.6273

Η συνάρτηση step(z,data=freeny) διαλέγει ένα βέλτιστο μοντέλο προσθέτοντας ή αφαιρώντας παράγοντες διατηρώντας την ιεραρχία σύμφωνα με τη μεγαλύτερη τιμή του AIC από την βηματική stepwise παλινδρόμηση. Το μοντέλο μας στην αρχή έχει ως εξαρτημένη μεταβλητή την y και ανεξάρτητες τις lag.quarterly.revenue, price.index, income.level και market.potential.
Έχουμε τη μεγαλύτερη τιμή του AIC για το μοντέλο μας όπου ισούται με -324,36, όμως η μεταβλητή lag.quarterly.revenue έχει AIC = -325,50 σε σχέση με τις υπόλοιπες ανεξάρτητες, επομένως η συνάρτηση step αφαιρεί την μεταβλητή από το μοντέλο επειδή έχει μικρότερη τιμή.
Στο επόμενο βήμα το μοντέλο μας έχει ως εξαρτημένη μεταβλητή την y και ανεξάρτητες τις price.index, income.level και market.potential. Η τιμή του AIC για αυτό το μοντέλο είναι ίσο με

-325,50 και παρατηρούμε ότι όλες οι μεταβλητές έχουνε μεγαλύτερες τιμές, οπότε η συνάρτηση step επιλέγει σαν βέλτιστο μοντέλο το μοντέλο με τις μεταβλητές price.index, income.level και market.potential.

Στη συνέχεια ορίζουμε με a τη συνάρτηση step(z,data=freeny) και με την συνάρτηση summary(a) μπορούμε να δούμε αναλυτικά τα αποτελέσματα της βηματικής παλινδρόμησης.

```
> a <- step(z,data=freeny)
```

```
> summary(a)

Call:
lm(formula = y ~ price.index + income.level + market.potential,
    data = freeny)

Residuals:
       Min          1Q      Median          3Q         Max
-0.0273061 -0.0090031  0.0007218  0.0111354  0.0270294

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -13.31014    5.04423  -2.639 0.012339 *
price.index       -0.83488    0.13084  -6.381 2.44e-07 ***
income.level       0.84556    0.09904   8.538 4.47e-10 ***
market.potential   1.62735    0.37682   4.319 0.000123 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01468 on 35 degrees of freedom
Multiple R-Squared: 0.998,      Adjusted R-squared: 0.9978
F-statistic:  5846 on 3 and 35 DF,  p-value: < 2.2e-16
```

Από τα παραπάνω αποτελέσματα μπορούμε να δούμε πληροφορίες για τα κριτήρια R^2 και R^2_{Adj}, δηλαδή τον συντελεστή προσδιορισμού και τον προσαρμοσμένο συντελεστή προσδιορισμού της βηματικής παλινδρόμησης αντίστοιχα.
Το R^2 ισούται με 0,998 και το R^2_{Adj} με 0,9978. Οι υψηλές τιμές των συντελεστών δείχνουν ότι το μοντέλο ερμηνεύει πολύ καλά τις μεταβολές της εξαρτημένης μεταβλητής με τη βοήθεια των ανεξάρτητων. Το 99,8% της μεταβλητότητας που υπάρχει ερμηνεύεται από το μοντέλο. Επίσης το βέλτιστο μοντέλο που προκύπτει είναι :

y = -13,31014 – 0,83488*price.index + 0,84556*income.level + 1,62735*market.potential

όπου y είναι η εξαρτημένη μεταβλητή, οι μεταβλητές price.index, income.level και market.potential είναι οι ανεξάρτητες μεταβλητές του μοντέλου μας, το -13,31014 είναι ο σταθερός μας όρος και η ερμηνεία του είναι ότι δεδομένου ότι οι άλλες μεταβλητές είναι σταθερές τα δεδομένα ξεκινούν από το -13,31014. Ο μερικός συντελεστής της μεταβλητής price.index μας δείχνει ότι δεδομένου ότι όλες οι άλλες μεταβλητές είναι σταθερές αν αυξηθεί η price.index κατά 1 μονάδα τότε η εξαρτημένη μεταβλητή y θα μειωθεί κατά 0,83488. Το ίδιο συμβαίνει και για τους άλλους συντελεστές των μεταβλητών αντίστοιχα.

```
> extractAIC(a, k=2)
[1]    4.0000 -325.4971
```

Με την συνάρτηση extractAIC(a, k=2) παίρνουμε πληροφορίες για το κριτήριο του Akaike(AIC). Η τιμή του είναι -325,4971 η οποία είναι η ελάχιστη τιμή για το μοντέλο το οποίο περιέχει την σταθερά του μοντέλου και τις ανεξάρτητες μεταβλητές price.index, income.level και market.potential. Άρα το μοντέλο αυτό είναι το βέλτιστο. Το ίδιο ισχύει και για το κριτήριο του Schwarz το οποίο δίνεται από την παρακάτω συνάρτηση extractAIC(a, k=log(nrow(a$model))) και η τιμή του είναι -318,8428.

```
> extractAIC(a, k=log(nrow(a$model)))
[1]    4.0000 -318.8428
```

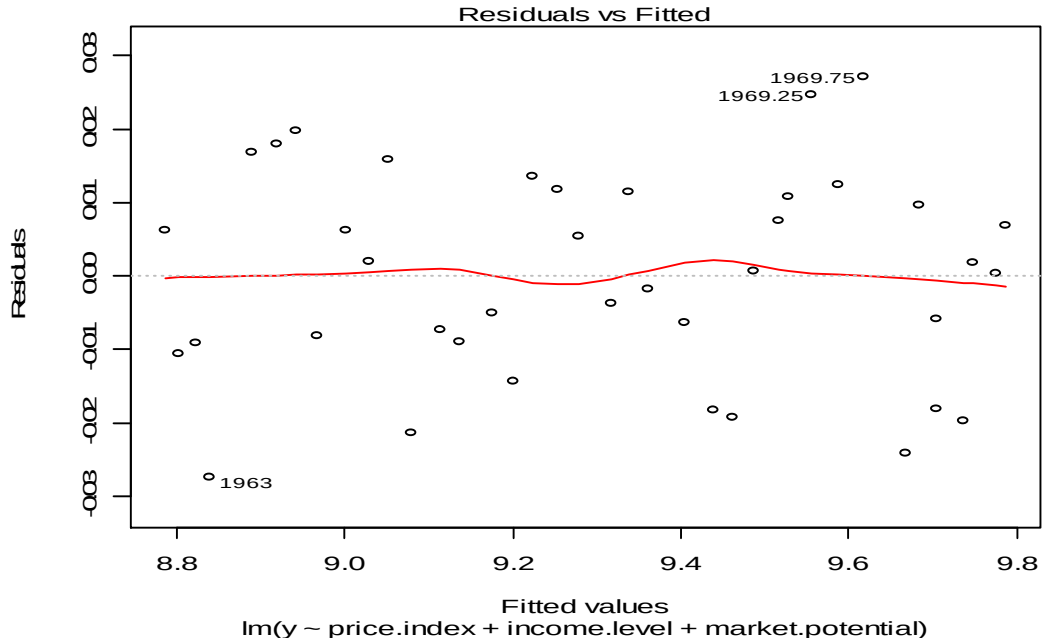
```
> rss=sum((a$residual)^2)
> sigma.full=summary(z)$sigma
> Cp=rss/sigma.full^2-nrow(a$model) + 2*ncol(a$model)
> Cp
[1] 3.756788
```

Σύμφωνα με τη στατιστική C_p του Mallows τα μοντέλα θα έχουν αρνητικό ή μικρό C_p – p, όπου p είναι ο αριθμός των παραμέτρων στο μοντέλο συμπεριλαμβανομένου του β_0, δηλαδή της σταθεράς. Εδώ το C_p είναι ίσο με 3,756788 και επειδή C_p – p = -1,756788 το μοντέλο μας είναι καλό.

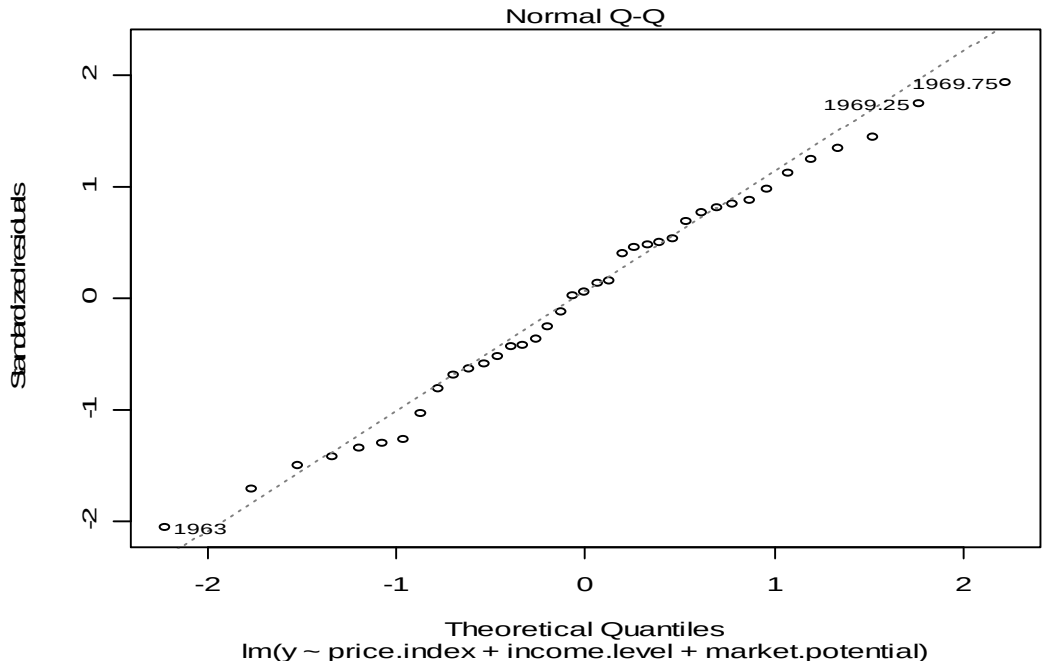
```
> msr <- deviance(a)/(39-4)
> msr
[1] 0.0002154044
```

Με τη συνάρτηση msr <- deviance(a)/(n-p) παίρνουμε πληροφορία για το μέσο τετράγωνο υπολοίπων που είναι ίσο με 0,000215044. Το deviance(a) είναι το άθροισμα των τετραγώνων των υπολοίπων για τη συνάρτηση step(z, data=freeny). Το n είναι ο αριθμός των δεδομένων μας και p είναι ο αριθμός των παραμέτρων στο μοντέλο μας. Αυτό το κριτήριο όμως δεν χρειάζεται να επιλέξει το πλήρες μοντέλο.

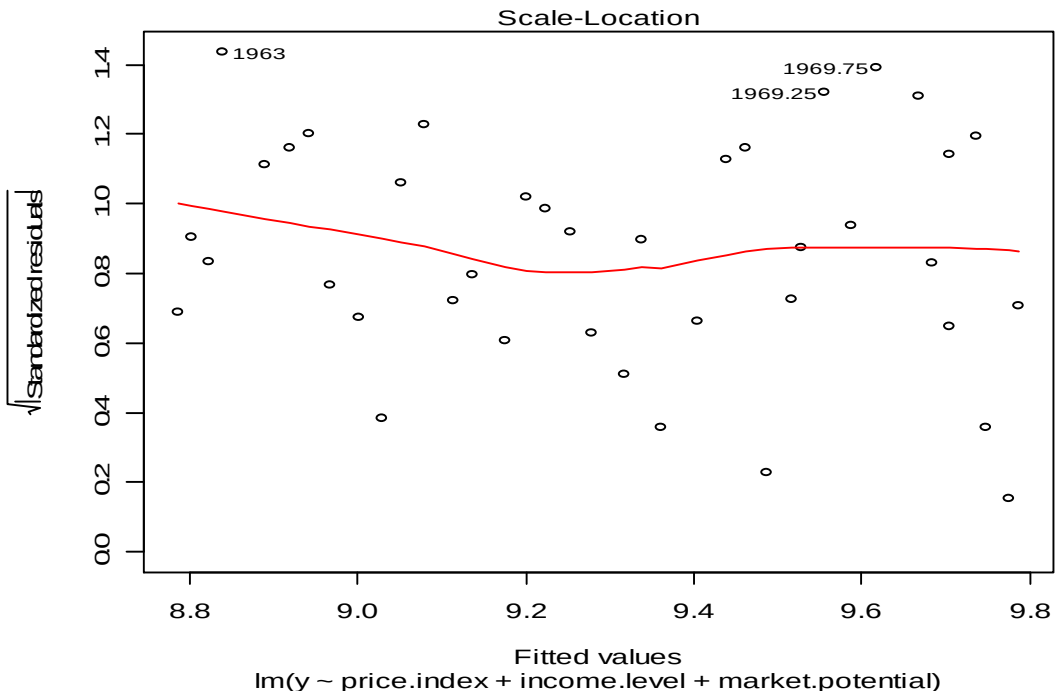
```
> plot(a)
```



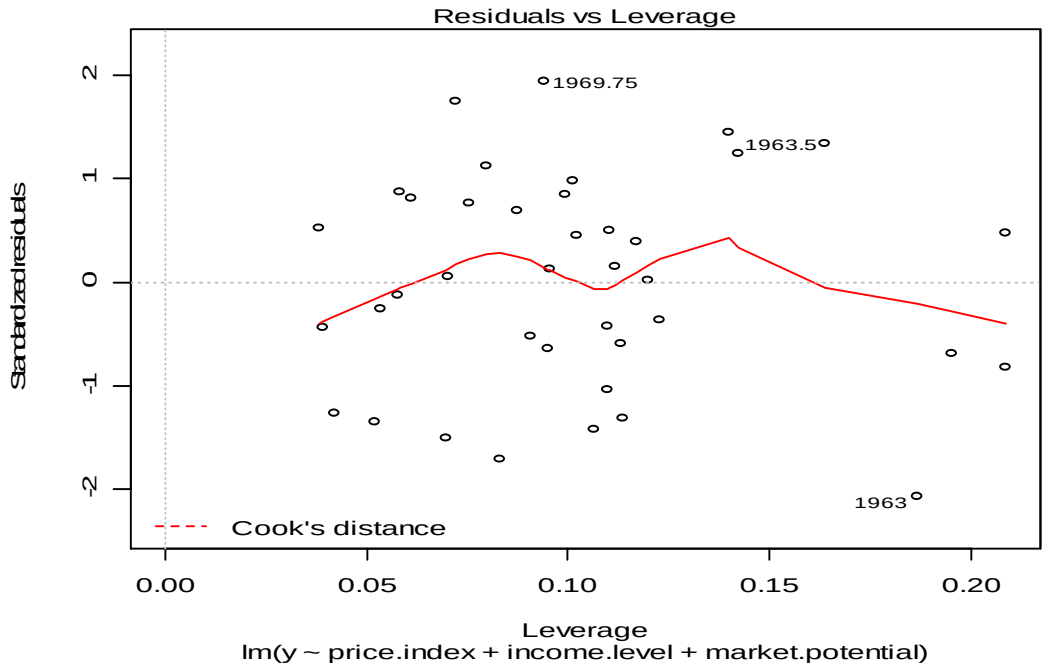
Από το παραπάνω διάγραμμα φαίνεται ότι τα υπόλοιπα είναι ομοιόμορφα κατανεμημένα και δεν παρουσιάζουν ανομοιογένεια των διασπορών. Επίσης η γραφική παράσταση των υπολοίπων δείχνει μια τυχαία τοποθέτηση τους γύρω από τη γραμμή που αντιστοιχεί σε υπόλοιπο ``0``.



Από το παραπάνω διάγραμμα φαίνεται ότι τα υπόλοιπα ακολουθούν την κανονική κατανομή. Δεν υπάρχει ένδειξη παραβίασης της υπόθεσης της κανονικότητας των υπολοίπων, αφού είναι συγκεντρωμένα γύρω από μια ευθεία.



Από το παραπάνω διάγραμμα παρατηρούμε ότι η διασπορά των υπολοίπων είναι σταθερή κατά μήκος των προσαρμοσμένων τιμών της παλινδρόμησης. Επίσης τα υψηλότερα σημεία είναι τα υπόλοιπα με τις μεγαλύτερες τιμές.



Στο παραπάνω διάγραμμα εξετάζουμε αν έχουμε απομονωμένες τιμές (outliers). Όπως φαίνεται υπάρχουν δυο τέτοιες τιμές όμως δεν μπορούμε να τις απορρίψουμε γιατί μπορεί να περιλαμβάνουν πληροφορίες που δεν υπάρχουν στα άλλα δεδομένα διότι προκύπτουν από ένα

ασυνήθιστο συνδυασμό περιστάσεων που έχουν ζωτικό ενδιαφέρον και απαιτείται συνεπώς περαιτέρω διερεύνηση. Έτσι είναι καλύτερα για το μοντέλο μας να μην υπάρχουν εξ' αρχής στα δεδομένα.

8.2 Παράδειγμα mtcars

R version 2.6.0 (2007-10-03)
Copyright (C) 2007 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> mtcars

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2

```
Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

> summary(mtcars)

```
      mpg             cyl             disp             hp       
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
 Median :19.20   Median :6.000   Median :196.3   Median :123.0  
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
      drat             wt             qsec             vs        
 Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
 1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
 Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
 Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
 3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000  
 Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000  
       am              gear            carb      
 Min.   :0.0000   Min.   :3.000   Min.   :1.000  
 1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000  
 Median :0.0000   Median :4.000   Median :2.000  
 Mean   :0.4062   Mean   :3.688   Mean   :2.812  
 3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000  
 Max.   :1.0000   Max.   :5.000   Max.   :8.000  
```

Όπου mpg είναι ο αριθμός μιλίων με ένα γαλόνι βενζίνης, cyl είναι ο αριθμός των κυλίνδρων, disp είναι η μετακίνηση, hp είναι η ιπποδύναμη, drat το μήκος του άξονα, wt είναι το βάρος σε λίβρες, qsec είναι η επιτάχυνση, am είναι αν το αυτοκίνητο είναι αυτόματο ή χειροκίνητο, gear είναι ο αριθμός των εξαρτημάτων και carb είναι ο αριθμός των καρμπιρατέρ.

> c <- lm(mpg~cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb,data=mtcars)
Ορίζουμε στο αντικείμενο c το μοντέλο με όλες τις μεταβλητές, ως εξαρτημένη την mpg και ως ανεξάρτητες τις cyl, disp, hp, drat, wt, qsec, vs, am, gear, carb. Πλέον έχουμε το μοντέλο μας στο αντικείμενο c.

> summary(c)
Η συνάρτηση summary μας δίνει μια αναλυτική περίληψη των αποτελεσμάτων της ανάλυσης παλινδρόμησης.

Call:
lm(formula = mpg ~ cyl + disp + hp + drat + wt + qsec + vs +
    am + gear + carb, data = mtcars)

```
Residuals:
    Min      1Q  Median      3Q     Max 
```

```
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-Squared: 0.869,      Adjusted R-squared: 0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

> step(c,data=mtcars)

```
Start:  AIC=70.9
mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq     RSS     AIC
- cyl   1     0.080 147.574  68.915
- vs    1     0.160 147.655  68.932
- carb  1     0.407 147.901  68.986
- gear  1     1.353 148.847  69.190
- drat  1     1.627 149.121  69.249
- disp  1     3.917 151.411  69.736
- hp    1     6.840 154.334  70.348
- qsec  1     8.864 156.359  70.765
<none>              147.494  70.898
- am    1    10.547 158.041  71.108
- wt    1    27.014 174.509  74.280

Step:  AIC=68.92
mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb

       Df Sum of Sq     RSS     AIC
- vs    1     0.269 147.843  66.973
- carb  1     0.520 148.094  67.028
- gear  1     1.821 149.395  67.308
```

```
- drat  1     1.983 149.557  67.342
- disp  1     3.901 151.475  67.750
- hp    1     7.363 154.937  68.473
<none>              147.574  68.915
- qsec  1    10.093 157.668  69.032
- am    1    11.836 159.410  69.384
- wt    1    27.028 174.602  72.297

Step:  AIC=66.97
mpg ~ disp + hp + drat + wt + qsec + am + gear + carb

       Df Sum of Sq     RSS     AIC
- carb  1     0.685 148.528  65.121
- gear  1     2.144 149.987  65.434
- drat  1     2.214 150.057  65.449
- disp  1     3.647 151.489  65.753
- hp    1     7.106 154.949  66.475
<none>              147.843  66.973
- am    1    11.569 159.412  67.384
- qsec  1    15.683 163.526  68.200
- wt    1    27.380 175.223  70.410

Step:  AIC=65.12
mpg ~ disp + hp + drat + wt + qsec + am + gear

       Df Sum of Sq     RSS     AIC
- gear  1     1.565 150.093  63.457
- drat  1     1.932 150.460  63.535
<none>              148.528  65.121
- disp  1    10.110 158.639  65.229
- am    1    12.323 160.852  65.672
- hp    1    14.826 163.354  66.166
- qsec  1    26.408 174.936  68.358
- wt    1    69.127 217.655  75.350

Step:  AIC=63.46
mpg ~ disp + hp + drat + wt + qsec + am

       Df Sum of Sq     RSS     AIC
- drat  1     3.345 153.438  62.162
- disp  1     8.545 158.639  63.229
<none>              150.093  63.457
- hp    1    13.285 163.378  64.171
- am    1    20.036 170.129  65.466
- qsec  1    25.574 175.668  66.491
- wt    1    67.572 217.665  73.351
Step:  AIC=62.16
mpg ~ disp + hp + wt + qsec + am
```

```
       Df Sum of Sq     RSS     AIC
- disp  1     6.629 160.066  61.515
<none>              153.438  62.162
- hp    1    12.572 166.010  62.682
- qsec  1    26.470 179.908  65.255
- am    1    32.198 185.635  66.258
- wt    1    69.043 222.481  72.051

Step:  AIC=61.52
mpg ~ hp + wt + qsec + am

       Df Sum of Sq     RSS     AIC
- hp    1     9.219 169.286  61.307
<none>              160.066  61.515
- qsec  1    20.225 180.291  63.323
- am    1    25.993 186.059  64.331
- wt    1    78.494 238.560  72.284

Step:  AIC=61.31
mpg ~ wt + qsec + am

       Df Sum of Sq    RSS    AIC
<none>              169.29  61.31
- am    1     26.18 195.46  63.91
- qsec  1    109.03 278.32  75.22
- wt    1    183.35 352.63  82.79

Call:
lm(formula = mpg ~ wt + qsec + am, data = mtcars)

Coefficients:
(Intercept)           wt         qsec           am
      9.618       -3.917        1.226        2.936
```

Η συνάρτηση step(c,data=mtcars) διαλέγει ένα βέλτιστο μοντέλο προσθέτοντας ή αφαιρώντας παράγοντες διατηρώντας την ιεραρχία σύμφωνα με τη μεγαλύτερη τιμή του AIC από την βηματική stepwise παλινδρόμηση.
Παρατηρούμε ότι από την αρχή μέχρι το σημείο όπου η συνάρτηση step επιλέγει το βέλτιστο μοντέλο υπάρχουν, σε κάθε βήμα, μεταβλητές οι οποίες έχουν μικρότερη τιμή AIC από τη τιμή του AIC της βηματικής παλινδρόμησης. Έτσι σε κάθε βήμα αφαιρείται η μεταβλητή η οποία έχει την μικρότερη τιμή AIC από όλες τις άλλες. Στο τελευταίο βήμα όπου έχουμε AIC = 61,31 έχουμε και το βέλτιστο μοντέλο αφού οι τρεις μεταβλητές που έμειναν έχουν μεγαλύτερη τιμή AIC.

Στη συνέχεια ορίζουμε με d τη συνάρτηση step(c,data=mtcars) και με την συνάρτηση summary(d) μπορούμε να δούμε αναλυτικά τα αποτελέσματα της βηματικής παλινδρόμησης.

```
> d <- step(c,data=mtcars)

> summary(d)

Call:
lm(formula = mpg ~ wt + qsec + am, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4811 -1.5555 -0.7257  1.4110  4.6610

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6178     6.9596   1.382 0.177915
wt           -3.9165     0.7112  -5.507 6.95e-06 ***
qsec          1.2259     0.2887   4.247 0.000216 ***
am            2.9358     1.4109   2.081 0.046716 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.459 on 28 degrees of freedom
Multiple R-Squared: 0.8497,     Adjusted R-squared: 0.8336
F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.210e-11
```

Από τα παραπάνω αποτελέσματα μπορούμε να δούμε πληροφορίες για τα κριτήρια R^2 και R^2_{Adj}, δηλαδή τον συντελεστή προσδιορισμού και τον προσαρμοσμένο συντελεστή προσδιορισμού της βηματικής παλινδρόμησης αντίστοιχα.
Το R^2 ισούται με 0,8497 και το R^2_{Adj} με 0,8336. Οι υψηλές τιμές των συντελεστών δείχνουν ότι το μοντέλο ερμηνεύει πολύ καλά τις μεταβολές της εξαρτημένης μεταβλητής με τη βοήθεια των ανεξάρτητων. Το 85% της μεταβλητότητας που υπάρχει ερμηνεύεται από το μοντέλο. Επίσης το μοντέλο που προκύπτει είναι :

Mpg = 9,6178 -3,9165*wt + 1,2259*qsec + 2,9358*am

όπου mpg είναι η εξαρτημένη μεταβλητή, οι μεταβλητές wt, qsec και am είναι οι ανεξάρτητες μεταβλητές του μοντέλου μας, το 9,6178 είναι ο σταθερός μας όρος και η ερμηνεία του είναι ότι δεδομένου ότι όλες οι άλλες μεταβλητές είναι σταθερές τα δεδομένα ξεκινούν από το 9,6178. Ο μερικός συντελεστής της μεταβλητής wt μας δείχνει ότι δεδομένου ότι όλες οι άλλες μεταβλητές είναι σταθερές αν αυξηθεί η wt κατά 1 μονάδα τότε η εξαρτημένη μεταβλητή mpg θα μειωθεί κατά 3,9165. Το ίδιο συμβαίνει και με τις άλλες μεταβλητές αντίστοιχα.

```
> extractAIC(d, k=2)
[1]   4.0000 61.3073
```

Με την συνάρτηση extractAIC(a, k=2) παίρνουμε πληροφορίες για το κριτήριο του Akaike(AIC). Η τιμή του είναι 61,3073 η οποία είναι η ελάχιστη τιμή για το μοντέλο το οποίο περιέχει την σταθερά του μοντέλου και τις ανεξάρτητες μεταβλητές wt, qsec και am. Άρα το μοντέλο αυτό είναι το βέλτιστο. Το ίδιο ισχύει και για το κριτήριο του Schwarz το οποίο δίνεται

από την παρακάτω συνάρτηση extractAIC(a, k=log(nrow(a$model))) και η τιμή του είναι 67,17025.

```
> extractAIC(d, k=log(nrow(d$model)))
[1]  4.00000 67.17025
```

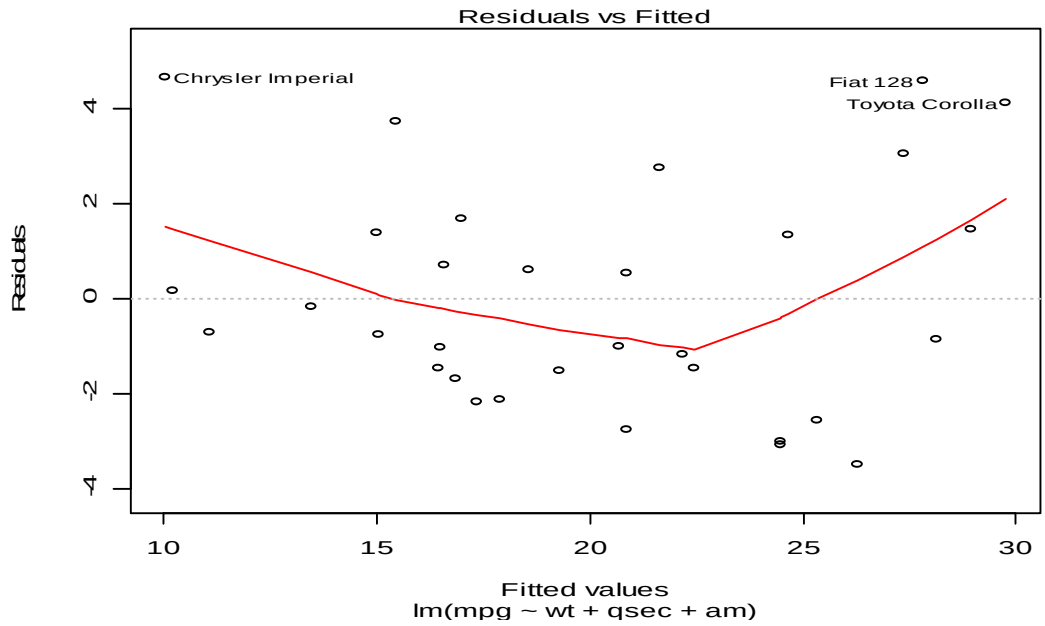
```
> rss=sum((d$residual)^2)
> sigma.full=summary(c)$sigma
> Cp=rss/sigma.full^2-nrow(d$model) + 2*ncol(d$model)
> Cp
[1] 0.1026357
```

Σύμφωνα με τη στατιστική C_p του Mallows τα μοντέλα θα έχουν αρνητικό ή μικρό C_p – p, όπου p είναι ο αριθμός των παραμέτρων στο μοντέλο συμπεριλαμβανομένου του β_0, δηλαδή της σταθεράς. Εδώ το C_p είναι ίσο με 0,1026357 και επειδή C_p – p = -3,8973643 το μοντέλο μας είναι καλό.

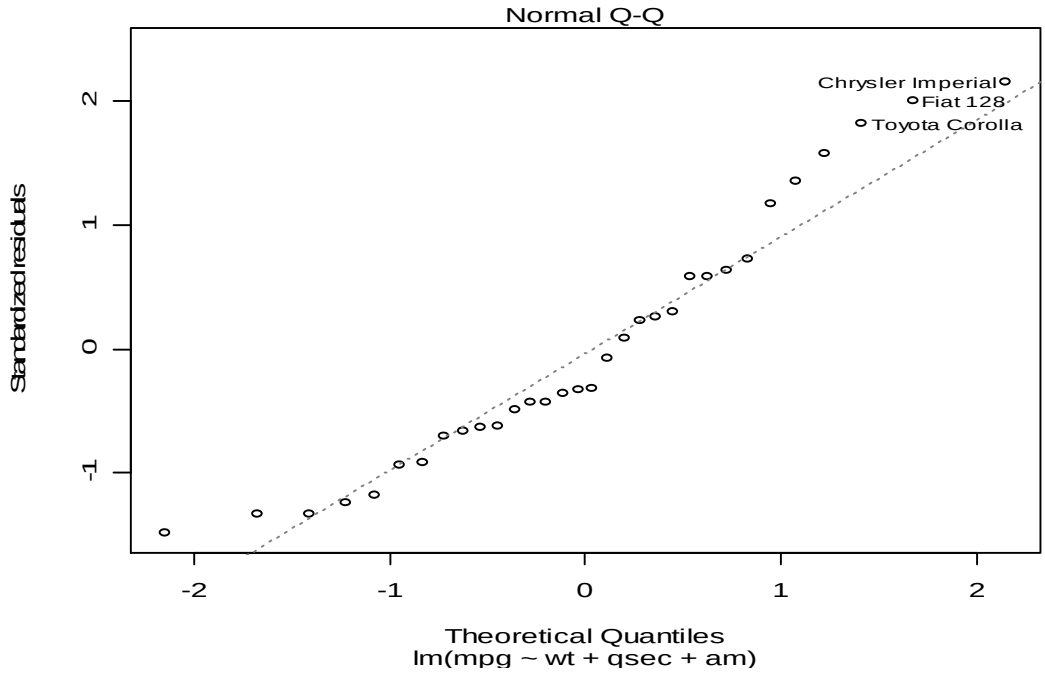
```
> msr<-deviance(d)/(32-4)
> msr
[1] 6.045926
```

Με τη συνάρτηση msr <- deviance(d)/(n-p) παίρνουμε πληροφορία για το μέσο τετράγωνο υπολοίπων που είναι ίσο με 6,045926. Το deviance(d) είναι το άθροισμα των τετραγώνων των υπολοίπων για τη συνάρτηση step(c, data=mtcars). Το n είναι ο αριθμός των δεδομένων μας και p είναι ο αριθμός των παραμέτρων στο μοντέλο μας. Αυτό το κριτήριο όμως δεν χρειάζεται να επιλέξει το πλήρες μοντέλο.

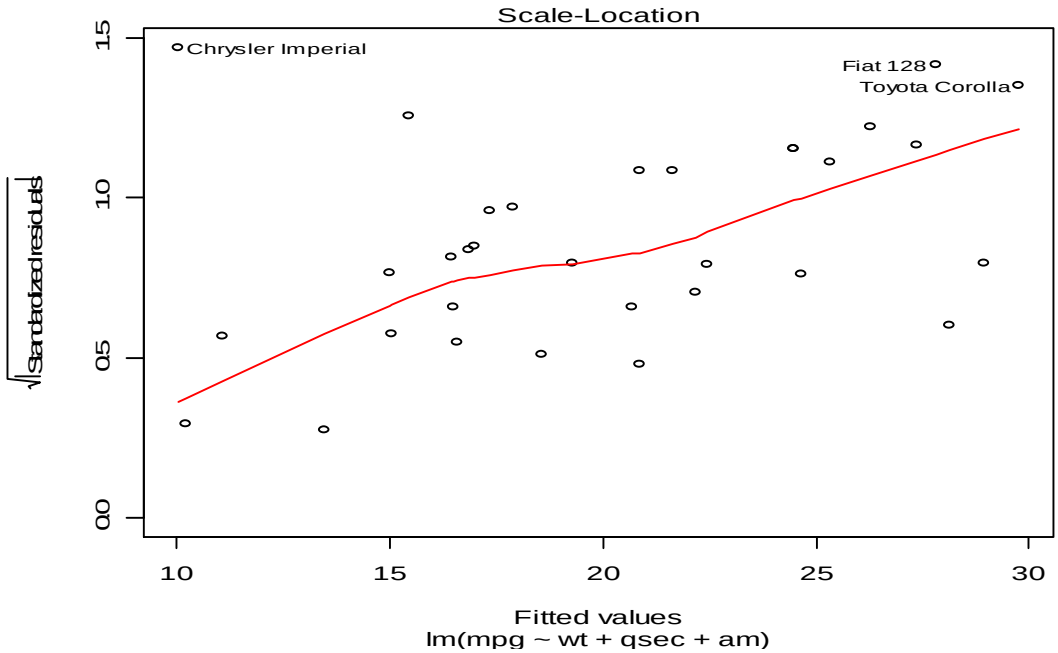
```
> plot(d)
```



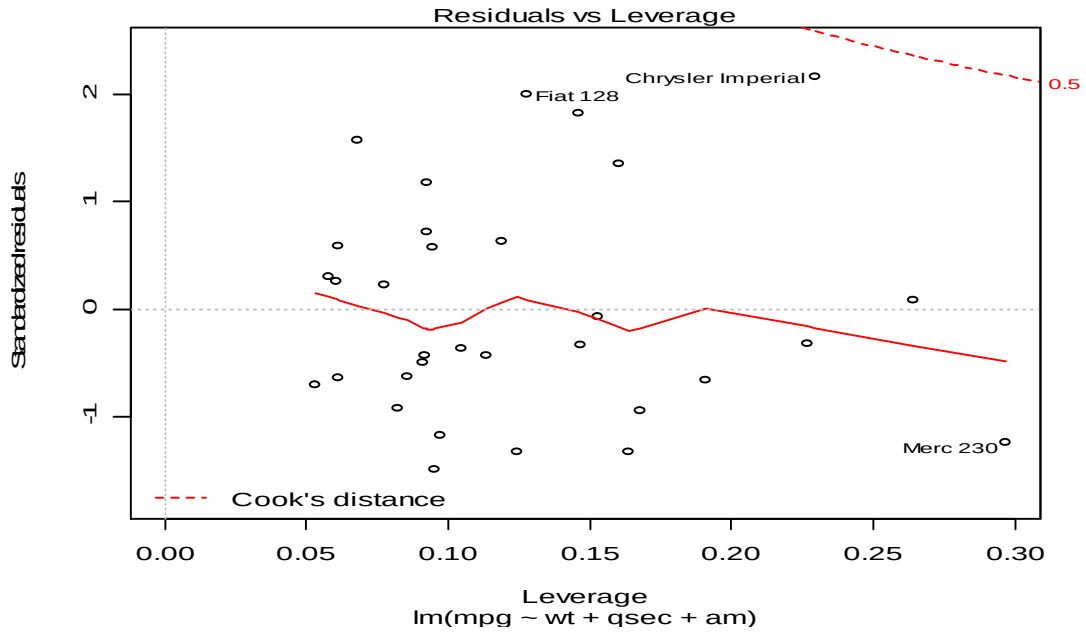
Από το παραπάνω διάγραμμα φαίνεται ότι τα υπόλοιπα είναι ομοιόμορφα κατανεμημένα και δεν παρουσιάζουν ανομοιογένεια των διασπορών. Επίσης η γραφική παράσταση των υπολοίπων δείχνει μια τυχαία τοποθέτηση τους γύρω από τη γραμμή που αντιστοιχεί σε υπόλοιπο ``0``.



Από το παραπάνω διάγραμμα φαίνεται ότι τα υπόλοιπα ακολουθούν την κανονική κατανομή. Δεν υπάρχει ένδειξη παραβίασης της υπόθεσης της κανονικότητας των υπολοίπων, αφού είναι συγκεντρωμένα γύρω από μια ευθεία, αν και μερικές μεταβλητές ξεφεύγουν για λίγο από την ευθεία.



Από το παραπάνω διάγραμμα παρατηρούμε ότι η διασπορά των υπολοίπων είναι σταθερή κατά μήκος των προσαρμοσμένων τιμών της παλινδρόμησης. Επίσης τα υψηλότερα σημεία είναι τα υπόλοιπα με τις μεγαλύτερες τιμές.



Στο παραπάνω διάγραμμα εξετάζουμε αν έχουμε απομονωμένες τιμές (outliers). Όπως φαίνεται υπάρχουν τρεις τέτοιες τιμές όμως δεν μπορούμε να τις απορρίψουμε γιατί μπορεί να περιλαμβάνουν πληροφορίες που δεν υπάρχουν στα άλλα δεδομένα διότι προκύπτουν από ένα ασυνήθιστο συνδυασμό περιστάσεων που έχουν ζωτικό ενδιαφέρον και απαιτείται συνεπώς περαιτέρω διερεύνηση. Έτσι είναι καλύτερα για το μοντέλο μας να μην υπάρχουν εξ' αρχής στα δεδομένα.

8.3 Παράδειγμα swiss

R version 2.6.0 (2007-10-03)
Copyright (C) 2007 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.
Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> swiss

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Courtelary	80.2	17.0	15	12	9.96	22.2
Delemont	83.1	45.1	6	9	84.84	22.2
Franches-Mnt	92.5	39.7	5	5	93.40	20.2
Moutier	85.8	36.5	12	7	33.77	20.3
Neuveville	76.9	43.5	17	15	5.16	20.6
Porrentruy	76.1	35.3	9	7	90.57	26.6
Broye	83.8	70.2	16	7	92.85	23.6
Glane	92.4	67.8	14	8	97.16	24.9
Gruyere	82.4	53.3	12	7	97.67	21.0
Sarine	82.9	45.2	16	13	91.38	24.4
Veveyse	87.1	64.5	14	6	98.61	24.5
Aigle	64.1	62.0	21	12	8.52	16.5
Aubonne	66.9	67.5	14	7	2.27	19.1
Avenches	68.9	60.7	19	12	4.43	22.7
Cossonay	61.7	69.3	22	5	2.82	18.7
Echallens	68.3	72.6	18	2	24.20	21.2
Grandson	71.7	34.0	17	8	3.30	20.0
Lausanne	55.7	19.4	26	28	12.11	20.2
La Vallee	54.3	15.2	31	20	2.15	10.8
Lavaux	65.1	73.0	19	9	2.84	20.0
Morges	65.5	59.8	22	10	5.23	18.0
Moudon	65.0	55.1	14	3	4.52	22.4
Nyone	56.6	50.9	22	12	15.14	16.7
Orbe	57.4	54.1	20	6	4.20	15.3
Oron	72.5	71.2	12	1	2.40	21.0
Payerne	74.2	58.1	14	8	5.23	23.8
Paysd'enhaut	72.0	63.5	6	3	2.56	18.0
Rolle	60.5	60.8	16	10	7.72	16.3
Vevey	58.3	26.8	25	19	18.46	20.9
Yverdon	65.4	49.5	15	8	6.10	22.5
Conthey	75.5	85.9	3	2	99.71	15.1
Entremont	69.3	84.9	7	6	99.68	19.8
Herens	77.3	89.7	5	2	100.00	18.3
Martigwy	70.5	78.2	12	6	98.96	19.4
Monthey	79.4	64.9	7	3	98.22	20.2
St Maurice	65.0	75.9	9	9	99.06	17.8
Sierre	92.2	84.6	3	3	99.46	16.3
Sion	79.3	63.1	13	13	96.83	18.1
Boudry	70.4	38.4	26	12	5.62	20.3
La Chauxdfnd	65.7	7.7	29	11	13.79	20.5
Le Locle	72.7	16.7	22	13	11.22	18.9
Neuchatel	64.4	17.6	35	32	16.92	23.0
Val de Ruz	77.6	37.6	15	7	4.97	20.0
ValdeTravers	67.6	18.7	25	7	8.65	19.5
V. De Geneve	35.0	1.2	37	53	42.34	18.0
Rive Droite	44.7	46.6	16	29	50.43	18.2
Rive Gauche	42.8	27.7	22	29	58.33	19.3

> summary(swiss)

```
   Fertility      Agriculture     Examination      Education    
 Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00  
 1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00  
 Median :70.40   Median :54.10   Median :16.00   Median : 8.00  
 Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98  
 3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00  
 Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00  
```

```
   Catholic       Infant.Mortality
 Min.   :  2.150   Min.   :10.80
 1st Qu.:  5.195   1st Qu.:18.15
 Median : 15.140   Median :20.00
 Mean   : 41.144   Mean   :19.94
 3rd Qu.: 93.125   3rd Qu.:21.70
 Max.   :100.000   Max.   :26.60
```

> x <- lm(Fertility~Agriculture + Examination + Education + Catholic + Infant.Mortality,data=swiss)
Ορίζουμε στο αντικείμενο x το μοντέλα με όλες τις μεταβλητές, ως εξαρτημένη την Fertility και ως ανεξάρτητες τις Agriculture, Examination, Education, Catholic, Infant.Mortality. Πλέον έχουμε το μοντέλο μας στο αντικείμενο x.

> summary(x)

Η συνάρτηση summary μας δίνει μια αναλυτική περίληψη των αποτελεσμάτων της ανάλυσης παλινδρόμησης.

```
Call:
lm(formula = Fertility ~ Agriculture + Examination + Education +
    Catholic + Infant.Mortality, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 **
Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-Squared: 0.7067,     Adjusted R-squared: 0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10


> step(x,data=swiss)

Start:  AIC=190.69
Fertility ~ Agriculture + Examination + Education + Catholic +
    Infant.Mortality

                   Df Sum of Sq    RSS    AIC
- Examination       1      53.0 2158.1  189.9
<none>                         2105.0  190.7
- Agriculture       1     307.7 2412.8  195.1
- Infant.Mortality  1     408.8 2513.8  197.0
- Catholic          1     447.7 2552.8  197.8
- Education         1    1162.6 3267.6  209.4

Step:  AIC=189.86
Fertility ~ Agriculture + Education + Catholic + Infant.Mortality

                   Df Sum of Sq    RSS    AIC
<none>                         2158.1  189.9
- Agriculture       1     264.2 2422.2  193.3
- Infant.Mortality  1     409.8 2567.9  196.0
- Catholic          1     956.6 3114.6  205.1
- Education         1    2250.0 4408.0  221.4
```

Call:
lm(formula = Fertility ~ Agriculture + Education + Catholic + Infant.Mortality, data = swiss)

Coefficients:

(Intercept)	Agriculture	Education	Catholic	Infant.Mortality
62.1013	-0.1546	-0.9803	0.1247	1.0784

Η συνάρτηση step(x,data=swiss) διαλέγει ένα βέλτιστο μοντέλο προσθέτοντας ή αφαιρώντας παράγοντες διατηρώντας την ιεραρχία σύμφωνα με τη μεγαλύτερη τιμή του AIC από την βηματική stepwise παλινδρόμηση. Το μοντέλο μας στην αρχή έχει ως εξαρτημένη μεταβλητή την Fertility και ανεξάρτητες τις Agriculture, Examination, Education, Catholic και Infant.Mortality.
Έχουμε τη μεγαλύτερη τιμή του AIC για το πλήρες μοντέλο μας όπου ισούται με 190,69, όμως η μεταβλητή Examination έχει AIC = 189,9 σε σχέση με τις υπόλοιπες ανεξάρτητες, επομένως η συνάρτηση step αφαιρεί την μεταβλητή από το μοντέλο επειδή έχει μικρότερη τιμή.
Στο επόμενο βήμα το μοντέλο μας έχει ως εξαρτημένη μεταβλητή την Fertility και ανεξάρτητες τις Agriculture,Education, Catholic και Infant.Mortality. Η τιμή του AIC για αυτό το μοντέλο είναι ίσο με 189,86 και παρατηρούμε ότι όλες οι μεταβλητές έχουνε μεγαλύτερες τιμές, οπότε η συνάρτηση step επιλέγει σαν βέλτιστο μοντέλο το μοντέλο με τις μεταβλητές Agriculture, Infant.Mortality, Catholic, Education.

Στη συνέχεια ορίζουμε με s τη συνάρτηση step(x,data=swiss) και με την συνάρτηση summary(s) μπορούμε να δούμε μια αναλυτική περίληψη των αποτελεσμάτων της ανάλυσης παλινδρόμησης.

```
> s <- step(x)

> summary(s)

Call:
lm(formula = Fertility ~ Agriculture + Education + Catholic +
   Infant.Mortality, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-14.6765  -6.0522   0.7514   3.1664  16.1422

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
Agriculture      -0.15462    0.06819  -2.267  0.02857 *
Education        -0.98026    0.14814  -6.617 5.14e-08 ***
Catholic          0.12467    0.02889   4.315 9.50e-05 ***
Infant.Mortality  1.07844    0.38187   2.824  0.00722 **
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.168 on 42 degrees of freedom
Multiple R-Squared: 0.6993,     Adjusted R-squared: 0.6707
F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10
```

Από τα παραπάνω αποτελέσματα μπορούμε να δούμε πληροφορίες για τα κριτήρια R^2 και R^2_{Adj}, δηλαδή τον συντελεστή προσδιορισμού και τον προσαρμοσμένο συντελεστή προσδιορισμού αντίστοιχα.
Το R^2 ισούται με 0,6993 και το R^2_{Adj} με 0,6707. Οι υψηλές τιμές των συντελεστών δείχνουν ότι το μοντέλο ερμηνεύει πολύ καλά τις μεταβολές της εξαρτημένης μεταβλητής με τη βοήθεια των ανεξάρτητων. Το 70% της μεταβλητότητας που υπάρχει ερμηνεύεται από το μοντέλο. Επίσης το μοντέλο που προκύπτει είναι :

Fertility = 62,10131 - 0,15462*Agriculture - 0,98026*Education + 0,12467*Catholic + 1,07844*Infant.Mortality

όπου Fertility είναι η εξαρτημένη μεταβλητή, οι μεταβλητές Agriculture, Education, Catholic και Infant.Mortality είναι οι ανεξάρτητες μεταβλητές του μοντέλου μας, το 62,10131 είναι ο σταθερός μας όρος και η ερμηνεία του είναι ότι δεδομένου ότι οι άλλες μεταβλητές είναι σταθερές τα δεδομένα ξεκινούν από το 62,10131. Ο μερικός συντελεστής της μεταβλητής Agriculture μας δείχνει ότι δεδομένου ότι όλες οι άλλες μεταβλητές είναι σταθερές αν αυξηθεί η Agriculture κατά 1 μονάδα τότε η εξαρτημένη μεταβλητή Fertility θα μειωθεί κατά 0,15462. Το ίδιο συμβαίνει και για τους άλλους συντελεστές των μεταβλητών αντίστοιχα.

```
> extractAIC(s, k=2)
[1]   5.0000 189.8606
```

Με την συνάρτηση extractAIC(a, k=2) παίρνουμε πληροφορίες για το κριτήριο του Akaike(AIC). Η τιμή του είναι 189,8606 η οποία είναι η ελάχιστη τιμή για το μοντέλο το οποίο περιέχει την σταθερά του μοντέλου και τις ανεξάρτητες μεταβλητές Agriculture, Education, Catholic και Infant.Mortality. Άρα το μοντέλο αυτό είναι το βέλτιστο. Το ίδιο ισχύει και για το κριτήριο του Schwarz το οποίο δίνεται από την παρακάτω συνάρτηση extractAIC(a, k=log(nrow(a$model))) και η τιμή του είναι 199,1114.

```
> extractAIC(s, k=log(nrow(s$model)))
[1]   5.0000 199.1114
```

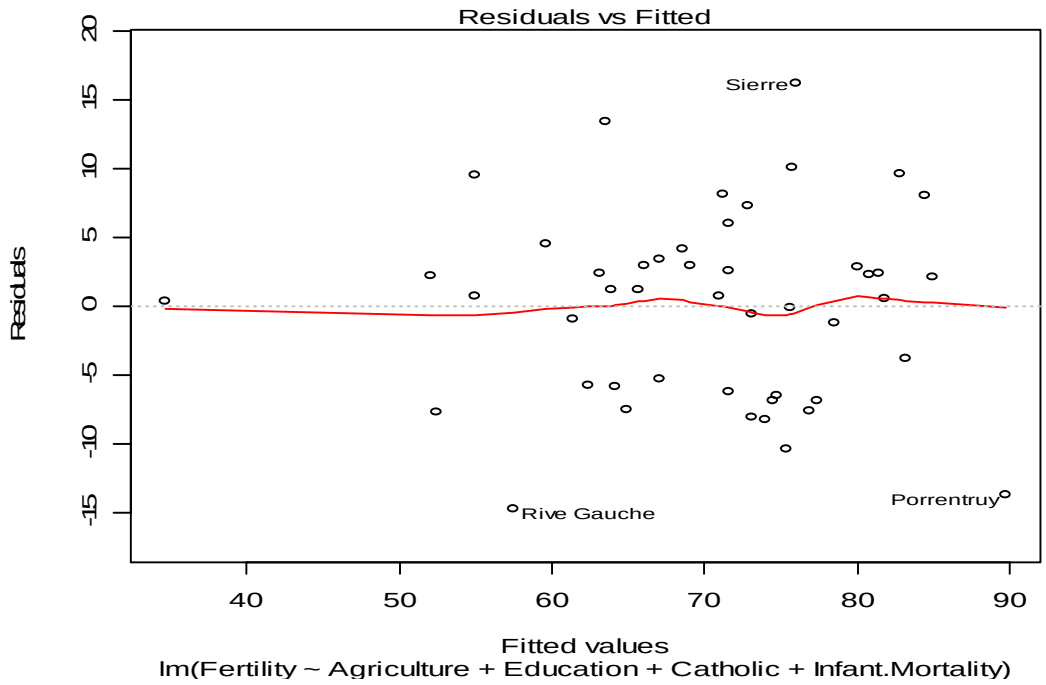
```
> rss=sum((s$residual)^2)
> sigma.full=summary(x)$sigma
> Cp=rss/sigma.full^2-nrow(s$model) + 2*ncol(s$model)
> Cp
[1] 5.0328
```

Σύμφωνα με τη στατιστική C_p του Mallows τα μοντέλα θα έχουν αρνητικό ή μικρό C_p – p, όπου p είναι ο αριθμός των παραμέτρων στο μοντέλο συμπεριλαμβανομένου του β_0, δηλαδή της σταθεράς. Εδώ το C_p είναι ίσο με 5,0328 και επειδή C_p – p = 0,0328 το μοντέλο μας είναι καλό.

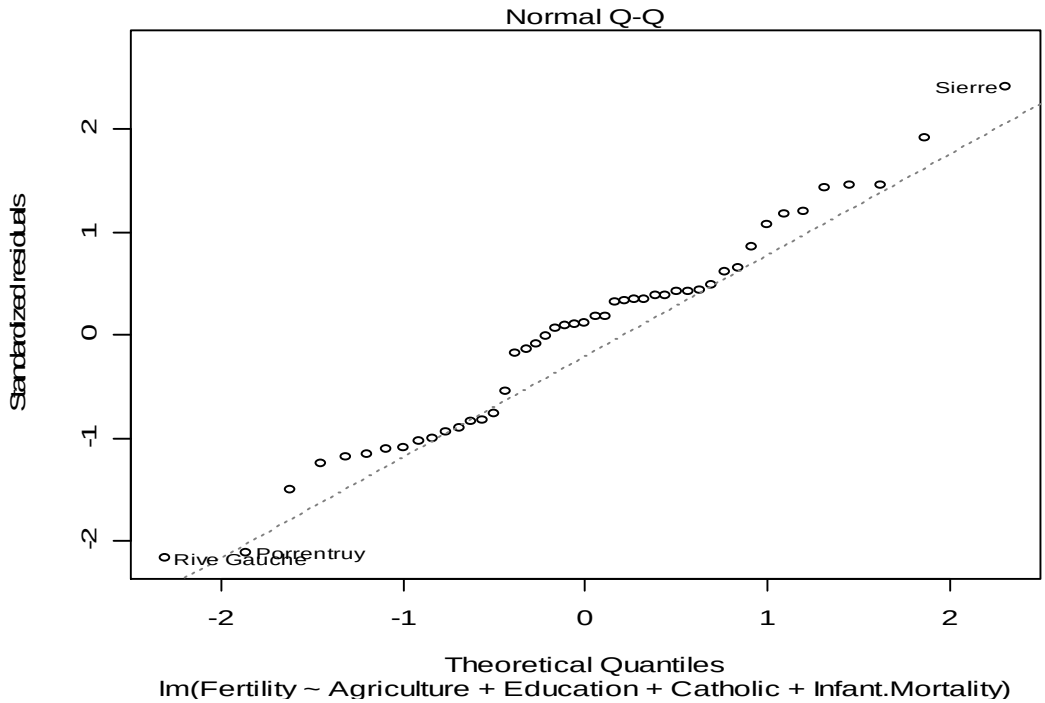
```
> msr <- deviance(s)/(47-5)
> msr
[1] 51.38261
```

Με τη συνάρτηση msr <- deviance(s)/(n-p) παίρνουμε πληροφορία για το μέσο τετράγωνο υπολοίπων που είναι ίσο με 51,38261. Το deviance(s) είναι το άθροισμα των τετραγώνων των υπολοίπων για τη συνάρτηση step(x, data=swiss). Το n είναι ο αριθμός των δεδομένων μας και p είναι ο αριθμός των παραμέτρων στο μοντέλο μας. Αυτό το κριτήριο όμως δεν χρειάζεται να επιλέξει το πλήρες μοντέλο.

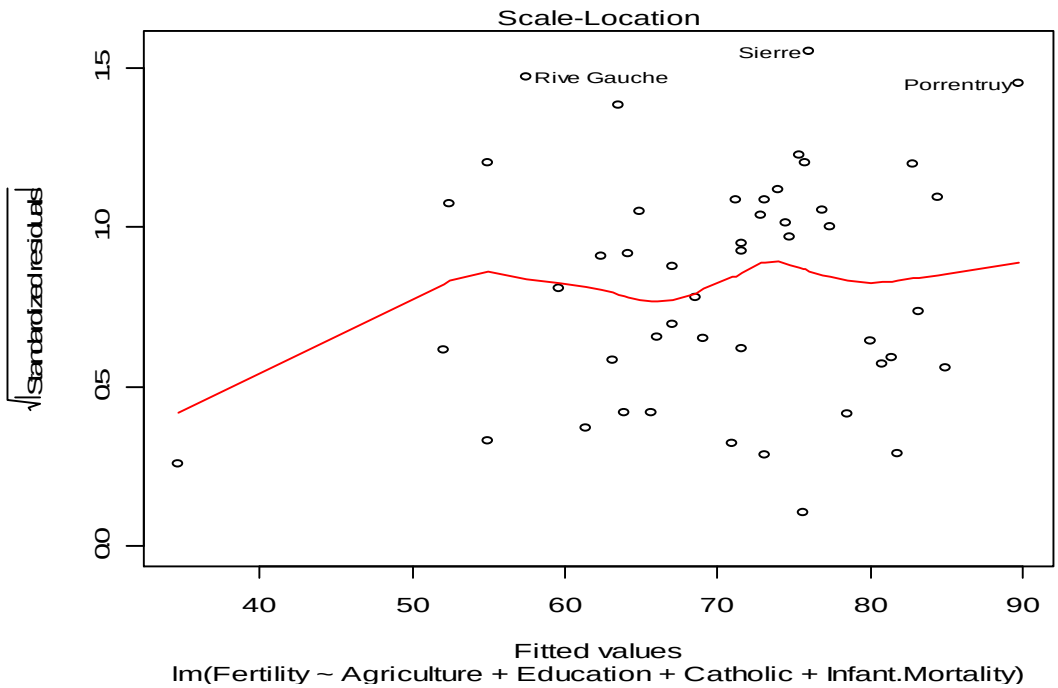
```
> plot(s)
```



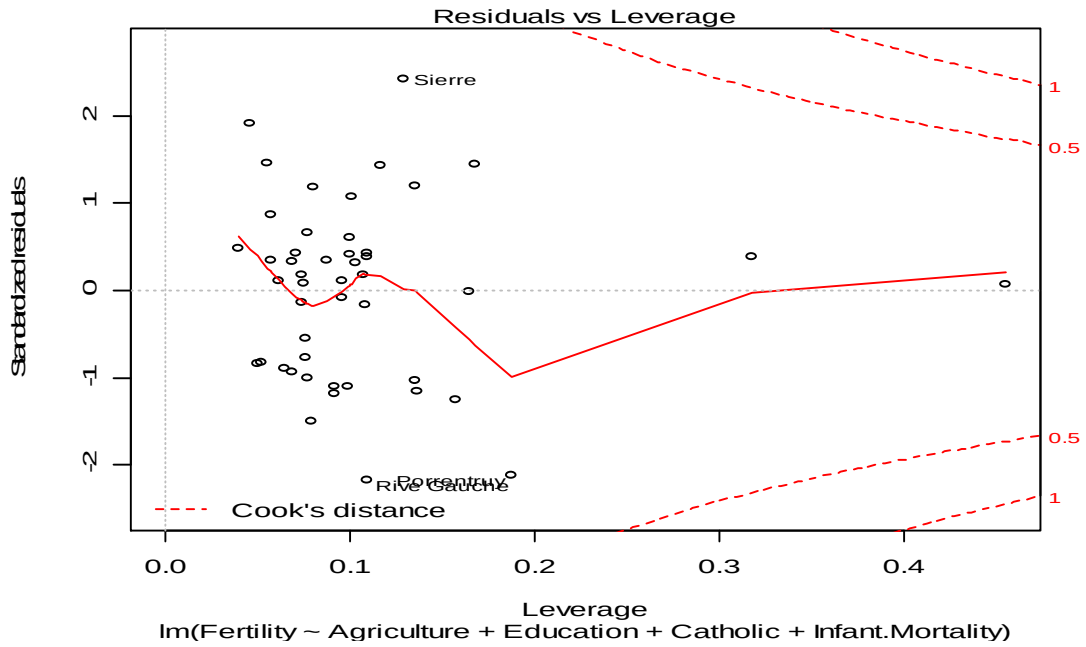
Από το παραπάνω διάγραμμα φαίνεται ότι τα υπόλοιπα είναι ομοιόμορφα κατανεμημένα και δεν παρουσιάζουν ανομοιογένεια των διασπορών. Επίσης η γραφική παράσταση των υπολοίπων δείχνει μια τυχαία τοποθέτηση τους γύρω από τη γραμμή που αντιστοιχεί σε υπόλοιπο ``0``.



Από το παραπάνω διάγραμμα φαίνεται ότι τα υπόλοιπα ακολουθούν την κανονική κατανομή. Δεν υπάρχει ένδειξη παραβίασης της υπόθεσης της κανονικότητας των υπολοίπων, αφού είναι συγκεντρωμένα γύρω από μια ευθεία.



Από το παραπάνω διάγραμμα παρατηρούμε ότι η διασπορά των υπολοίπων είναι σταθερή κατά μήκος των προσαρμοσμένων τιμών της παλινδρόμησης. Επίσης τα υψηλότερα σημεία είναι τα υπόλοιπα με τις μεγαλύτερες τιμές.



Στο παραπάνω διάγραμμα εξετάζουμε αν έχουμε απομονωμένες τιμές (outliers). Όπως φαίνεται υπάρχουν δυο τέτοιες τιμές όμως δεν μπορούμε να τις απορρίψουμε γιατί μπορεί να περιλαμβάνουν πληροφορίες που δεν υπάρχουν στα άλλα δεδομένα διότι προκύπτουν από ένα ασυνήθιστο συνδυασμό περιστάσεων που έχουν ζωτικό ενδιαφέρον και απαιτείται συνεπώς περαιτέρω διερεύνηση. Έτσι είναι καλύτερα για το μοντέλο μας να μην υπάρχουν εξ' αρχής στα δεδομένα.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα πτυχιακή ασχοληθήκαμε με τις μεθόδους επιλογής ανεξάρτητων μεταβλητών και με τα κριτήρια για την επιλογή του βέλτιστου μοντέλου. Η καλύτερη μέθοδος επιλογής μεταβλητών είναι η βηματική παλινδρόμηση γιατί είναι οικονομικότερη σε ότι αφορά τον υπολογιστικό χρόνο και το ανθρώπινο δυναμικό που απαιτούνται για την εφαρμογή της και αποφεύγει να χρησιμοποιεί περισσότερες ανεξάρτητες μεταβλητές (x) από όσα είναι απαραίτητα για την βελτίωση της εξίσωσης σε κάθε φάση. Τα κριτήρια για την επιλογή ενός μοντέλου με τα οποία ασχολήθηκα είναι τα εξής : R^2, R^2_{Adj}, C_p, MSR, AIC, HQ, BIC.
Η εφαρμογή της βηματικής παλινδρόμησης και των κριτηρίων στα προγράμματα SPSS και R-project με οικονομικά δεδομένα μας έδειξε ότι τελικά η μέθοδος της βηματικής παλινδρόμησης επιλέγει συγκεκριμένες ανεξάρτητες μεταβλητές που πρέπει να είναι στο βέλτιστο μοντέλο. Επίσης τα κριτήρια, κάθε ένα ξεχωριστά, επιλέγουν ένα βέλτιστο μοντέλο και μπορούν να χρησιμοποιηθούν για τη σύγκριση των μοντέλων της βηματικής παλινδρόμησης. Το R^2 δεν είναι ο κατάλληλος δείκτης για να συγκρίνουμε 2 μοντέλα, επειδή το R^2 αυξάνεται πάντα όταν μια νέα επεξηγηματική μεταβλητή εισέλθει στο μοντέλο και δεν μπορούμε να βγάλουμε ένα σωστό συμπέρασμα για τη σημαντικότητα της μεταβλητής αυτής. Το R^2_{Adj} αποτελεί ένα «ποινικοποιημένο» (penalized) μέτρο καλής εφαρμογής αφού υπάρχει πιθανότητα αν προσθέσουμε μια μεταβλητή στο μοντέλο να μειωθεί το R^2_{Adj}, οπότε να συμπεράνουμε ότι αυτή η μεταβλητή δεν είναι απαραίτητη για το μοντέλο. Αν όμως αυξηθεί το R^2_{Adj} τότε η μεταβλητή αυτή είναι σημαντική για το μοντέλο και πρέπει να την συμπεριλάβουμε σε αυτό. Σύμφωνα με το

Mallow (1973) όλα τα μοντέλα με $C_p \approx p+1$ θα πρέπει να θεωρούνται σαν υποψήφια για

περαιτέρω μελέτη. Στην περίπτωση που η επιλογή γίνεται μεταξύ μοντέλων με ίδιο αριθμό παραμέτρων, τότε επιλέγουμε εκείνο το μοντέλο για το οποίο $C_p \leq p+1$. Ο Hocking (1976) πρότεινε δύο άλλα κριτήρια για την επιλογή του καλύτερου μοντέλου με την βοήθεια του δείκτη του Mallow. Το πρώτο είναι το $C_p \leq p+1$, εάν σκοπός του μοντέλου είναι η πρόβλεψη, και το δεύτερο το $C_p \leq 2(p+1)-p$, εάν σκοπός του μοντέλου είναι απλώς η εκτίμηση των παραμέτρων. Ο Mallows πρότεινε ότι τα καλά μοντέλα θα έχουν αρνητικό ή μικρό C_p-p. Το κριτήριο MSR δεν χρειάζεται να επιλέξει το πλήρες μοντέλο δεδομένου ότι και το SS_{res} και το (n-p) θα μειωθεί καθώς το p αυξάνεται και έτσι το MSR_p μπορεί να μειωθεί ή να αυξηθεί. Ωστόσο, στις περισσότερες περιπτώσεις, αυτό το κριτήριο ευνοεί μεγάλα υποσύνολα παρά μικρότερα και είναι χρήσιμο όταν σκοπός του μοντέλου μας είναι η πρόβλεψη. Το κριτήριο του Akaike, του Schwarz και του Hannan και Quinn είναι να επιβάλουν κάποια ποινή για κάθε μοντέλο με περισσότερες παραμέτρους γιατί αλλιώς η πιθανοφάνεια είναι λογικό να αυξάνει προσθέτοντας παραμέτρους, επομένως αυτή η ποινή αποζημιώνει για τις παραπανίσιες παραμέτρους. Το κριτήριο του Schwarz λαμβάνει υπόψη του στην ποινή αυτή τόσο τον αριθμό των παραπανίσιων παραμέτρων αλλά και το μέγεθος του δείγματος κάτι το οποίο δεν συμβαίνει στις περιπτώσεις του AIC και του HQ.
Έχουν αναφερθεί αρκετά μειονεκτήματα για τις πολυβηματικές διαδικασίες. Ένα από αυτά είναι ότι από τις μεθόδους Forward selection, Backward elimination, Stepwise selection δεν εξασφαλίζει ότι το υποσύνολο από ένα συγκεκριμένο πλήθος μεταβλητών εμφανίζεται προς σύγκριση. Επίσης, το γεγονός ότι προτείνεται μια σειρά ελέγχων σημαντικότητας για τις μεταβλητές ενδέχεται να είναι παραπλανητικό, αφού για παράδειγμα υπάρχει πιθανότητα η

πρώτη μεταβλητή που εισέρχεται στο μοντέλο (στην Forward selection μέθοδο) να μην είναι απαραίτητη όταν εισέλθουν κάποιες άλλες μεταβλητές στο μοντέλο ή μπορεί η πρώτη μεταβλητή που διαγράφεται στην Backward elimination μέθοδο να είναι η πρώτη που εισέρχεται στην Forward selection μέθοδο. Οι Gorman και Toman (1966) παρατηρούν ότι είναι σπάνιο να υπάρχει ένα μόνο βέλτιστο υποσύνολο, αλλά ότι συνήθως υπάρχουν διάφορα εξίσου καλά υποσύνολα. Στα προγράμματα SPSS και R-project επιλέξαμε ένα μοντέλο το οποίο θεωρείται το καλύτερο ως προς την ικανότητά του να προσαρμοστεί στα δεδομένα. Επίσης το μοντέλο δεν παρουσιάζει κανένα απολύτως πρόβλημα στα κατάλοιπα του, τα οποία είναι κανονικά δεν έχουν αυτοσυσχέτιση αλλά ούτε και ετεροσκεδαστικότητα.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Βιβλία

1. Γραμμικά Μοντέλα : Καρακώστας Κωνσταντίνος, Πανεπιστήμιο Ιωαννίνων.
2. Εφαρμοσμένη Ανάλυση Παλινδρόμησης : Χατζηκωνσταντινίδης Ευστάθιος, Καλαματιανού Αγλαία, Εκδόσεις Παπαζήση.
3. Applied Linear Statistical Models : John Neter, Michael H. Kutner, William Wasserman, Christopher J. Nachtsheim.

Ηλεκτρονικές Σελίδες

1. Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης :
   - http : //web.auth.gr/e-topo/TOMEIS_INDEX/TOMEASB/Lafazani/Give/kef10_2_Palindr_sysxet.pdf
   - http://users.auth.gr/~dkugiu/Teach/CivilTransport/ApliedStats Chapters.pdf
2. Ehsan’s Statistical pages :
   - http : //www.angelfire.com/ab5/get5/selreg.pdf
3. Journal of Statistical Software :
   - http : //www.jstatsoft.org/v07/il2/paper
4. Illinois Quantitative Graduates :
   - http ://www.iqgrads.net/jtemplin/teaching/psyc791s07/psyc791s07_02.pdf

Προγράμματα που χρησιμοποιηθήκανε

1. Στατιστική με το SPSS έκδοση 15 για Windows.
2. R version 2.6.0, The R Foundation for Statistical Computing.