ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

Μέθοδοι επιλογής βέλτιστου συνόλου ανεξάρτητων μεταβλητών σε μοντέλα γραμμικής παλινδρόμησης

Βασιλική Μ. Τροχοπούλου

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς

2009

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην από 7/5/2007 συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική.

Τα μέλη της Επιτροπής ήταν:
Καθηγητής Μ. Κούτρας (Επιβλέπων)
Αναπληρωτής Καθηγητής Κ. Τσίμπος
Λέκτορας Γ. Βερροπούλου

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN APPLIED STATISTICS

Methods for selecting an optimal set of independent variables in linear regression models

Vasiliki M. Trochopoulou

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfillment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece

2009

Στους γονείς μου

Ευχαριστίες

Θέλω να ευχαριστήσω τον επιβλέποντα καθηγητή μου κύριο Μάρκο Κούτρα για την υπομονή που έδειξε μέχρι και την ολοκλήρωση της εργασίας μου.

ΠΕΡΙΛΗΨΗ

Στη γραμμική παλινδρόμηση, ένα από τα πλέον σημαντικά προβλήματα είναι η επιλογή ενός υποσυνόλου από τις ανεξάρτητες μεταβλητές που είναι διαθέσιμες κάθε φορά, έτσι ώστε αφενός μεν να υπάρχει εξοικονόμηση κόστους κατά την πρόβλεψη της εξαρτημένης μεταβλητής, αφετέρου δε να μην προκύπτει μεγάλη απώλεια στην αποτελεσματικότητα του μοντέλου πρόβλεψης. Το πρόβλημα αυτό έχει πολλές πρακτικές εφαρμογές σε διάφορους τομείς, όπως κοινωνικές επιστήμες, οικονομία, μετεωρολογικά φαινόμενα και πολλές άλλες.

Στη βιβλιογραφία έχει προταθεί ένα μεγάλο πλήθος από μεθόδους και κριτήρια βελτιστότητας. Στην παρούσα διπλωματική εργασία παρουσιάζουμε τις κυριότερες τεχνικές οι οποίες οδηγούν στην επιλογή του βέλτιστου συνόλου ανεξάρτητων μεταβλητών για την πρόβλεψη μιας εξαρτημένης μεταβλητής μέσω ενός γραμμικού μοντέλου.

Αρχικά παρουσιάζεται η μέθοδος της «εξέτασης όλων των δυνατών μοντέλων» (δηλαδή όλων των δυνατών συνδυασμών από ανεξάρτητες μεταβλητές), δίνοντας αναλυτικά τα διάφορα κριτήρια που έχουν προταθεί για τον εντοπισμό του βέλτιστου μοντέλου. Κάθε ένα από τα κριτήρια εφαρμόζεται σε ένα δείγμα 30 παρατηρήσεων οι οποίες ελήφθησαν από μια βάση με δεδομένα κατανάλωσης διαφόρων τύπων αυτοκινήτων με διαφορετικές παραμέτρους λειτουργίας.

Στη συνέχεια παρουσιάζονται οι λεγόμενες επαναληπτικές μέθοδοι επιλογής μεταβλητών. Με αυτές δημιουργείται μια αλληλουχία γραμμικών μοντέλων εισάγοντας ή διαγράφοντας κάθε φορά μια ανεξάρτητη μεταβλητή μέχρις ότου να φτάσουν σε κάποιο σημείο όπου ικανοποιείται ένα κριτήριο διακοπής.

Στο τελευταίο μέρος δίνονται διάφορα παραδείγματα συνόλων δεδομένων που έχουν χρησιμοποιηθεί στη βιβλιογραφία για τη σύγκριση μεθόδων εντοπισμού του βέλτιστου συνόλου ανεξάρτητων μεταβλητών σε ένα γραμμικό μοντέλο και γίνεται σύγκριση των μεθόδων που παρουσιάστηκαν στα πλαίσια της παρούσας διπλωματικής.

ABSTRACT

In linear regression, one of the most important problems is the selection of a subset of independent variables that are available, so that on one hand the prediction of the dependent variable is cost effective, and on the other hand the efficiency loss experienced in the prediction model is as small as possible. This problem has many practical applications in various fields, such as social sciences, economics, meteorological phenomena and many others.

In the literature a variety of methods and optimality criteria have been suggested for selecting the appropriate set of variables. In this dissertation we present the most popular techniques which lead to the selection of the optimal set of independent variables, in order to predict a dependent variable through a linear model.

Firstly, we present the "examination of all possible models" method (i.e. of all possible combinations of independent variables), describing in detail various criteria that have been proposed to identify the optimal model. Each one of the criteria is applied to a sample of 30 observations, which have been extracted from a data base containing the gas consumption (mileage) of several types of cars with different operating parameters.

Next, we present the so-called iterative variable selection methods. Through these a sequence of linear models is created by introducing or deleting at each step an independent variable until a stopping criterion is met.

In the last section we give several examples of data sets used in the bibliography to compare the methods of establishing an optimal set of independent variables in a linear model and present several results related to the methods described in this thesis.

ΠΕΡΙΕΧΟΜΕΝΑ

Κατάλογος πινάκων

Κατάλογος Σχημάτων

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Μοντέλα παλινδρόμησης και οι χρήσεις τους

Το αντικείμενο στο οποίο στηρίζεται η εργασία αυτή είναι η περιγραφή των μεθόδων αντιμετώπισης ενός από τα πλέον σημαντικά προβλήματα της γραμμικής παλινδρόμησης: της επιλογής ενός υποσυνόλου από ανεξάρτητες μεταβλητές, έτσι ώστε αφενός μεν να υπάρχει εξοικονόμηση κόστους κατά την πρόβλεψη της εξαρτημένης μεταβλητής αφετέρου και να μην προκύπτει μεγάλη απώλεια στην αποτελεσματικότητα του μοντέλου πρόβλεψης.

Θα ξεκινήσουμε παρουσιάζοντας μερικά εισαγωγικά στοιχεία για τη γραμμική παλινδρόμηση. Στη γραμμική παλινδρόμηση διακρίνουμε το *απλό γραμμικό μοντέλο* το οποίο ποσοτικοποιεί τη σχέση δύο συνεχών μεταβλητών X και Y υπό τη μορφή ενός γραμμικού υποδείγματος στο οποίο οι τιμές της μιας μεταβλητής προβλέπονται με βάση τις τιμές της άλλης. Αν οι τιμές της μεταβλητής Y προβλέπονται με βάση τις τιμές της X, τότε η Y ονομάζεται **εξαρτημένη μεταβλητή** και η μεταβλητή X ονομάζεται **ανεξάρτητη**. Η σχέση με την οποία συνδέονται οι μεταβλητές X και Y είναι της μορφής $Y \cong \beta_0 + \beta_1 X$ και παριστάνεται με μία ευθεία, όπου β_0 και β_1 είναι οι παράμετροι της σχέσης. Ειδικότερα, ο συντελεστής β_0 είναι ο σταθερός όρος δηλαδή το σημείο από όπου ξεκινά η ευθεία που προσπαθούμε να φέρουμε ανάμεσα από τα σημεία των συντεταγμένων των δύο μεταβλητών. Ο συντελεστής β_1 αντιπροσωπεύει την κλίση , δηλαδή δείχνει τη μεταβολή της εξαρτημένης μεταβλητής Y για τη μεταβολή της X κατά μια μονάδα. Αν η κλίση β_1 είναι θετική

τότε έχουμε αύξηση της Y όταν αυξάνεται η X , ενώ αντίθετα αν η κλίση β_1 είναι αρνητική τότε θα έχουμε μείωση της Y.

Αν η μεταβλητή Y δεν υπόκειται σε σφάλματα, οπότε για κάθε συγκεκριμένη τιμή της ανεξάρτητης μεταβλητής X μπορούμε να προβλέψουμε ακριβώς την τιμή της μεταβλητής Y, το μοντέλο καλείται προσδιοριστικό. Στην πραγματικότητα όμως δεν μπορούμε να προβλέψουμε την τιμή της μεταβλητής Y με απόλυτη ακρίβεια. Ένα μοντέλο που δίνει τη δυνατότητα στη μεταβλητή Y να μη βρίσκεται ακριβώς πάνω στην ευθεία $Y = \beta_0 + \beta_1 X$ είναι το ακόλουθο

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

όπου η ποσότητα ε θεωρείται ως ένα τυχαίο σφάλμα και παριστάνει τη διαφορά της παρατηρούμενης τιμής για τη μεταβλητή Y, από τη θεωρητική τιμή $\beta_0 + \beta_1 X$, για δοσμένη τιμή της μεταβλητής X.

Η εκτίμηση της παλινδρόμησης που βασίζεται σε περισσότερες από μια ανεξάρτητες μεταβλητές είναι γνωστή ως πολλαπλή παλινδρόμηση και αποτελεί επέκταση της απλής παλινδρόμησης. Στο υπόδειγμα αυτό η εξαρτημένη μεταβλητή Y συνδέεται γραμμικά με τις ανεξάρτητες μεταβλητές $X_1,..,X_{p-1}$ μέσω της σχέσης

$$Y \cong \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1}$$

Στο υπόδειγμα πολλαπλής παλινδρόμησης η εξαρτημένη μεταβλητή Y εξαρτάται ταυτόχρονα από μια σειρά μεταβλητών, η ατομική επίδραση κάθε μεταβλητής μπορεί να καθοριστεί από τους αντίστοιχους συντελεστές παλινδρόμησης $\beta_1, \beta_2,.., \beta_k$ των k ανεξάρτητων μεταβλητών . Ειδικότερα, ο συντελεστής β_0 είναι ο σταθερός όρος, δηλαδή η τιμή της εξαρτημένης μεταβλητής Y όταν όλες οι ανεξάρτητες μεταβλητές παίρνουν την τιμή μηδέν, ενώ ο συντελεστής β_i εκφράζει τη μεταβολή της εξαρτημένης μεταβλητής Y για μια μονάδα αύξησης της ανεξάρτητης μεταβλητής X_i, εφόσον οι τιμές των άλλων ανεξάρτητων μεταβλητών παραμένουν σταθερές.

Η γραμμική σχέση $Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1}$ θα ίσχυε αν δεν υπήρχαν σφάλματα στις μετρήσεις ούτε άλλοι παράγοντες, εκτός από αυτούς που αντιπροσωπεύονται από τις μεταβλητές $X_{i1},..,X_{i,p-1}$, οι οποίοι να επιδρούν κατά

κάποιο τρόπο στη διαμόρφωση των τιμών της μεταβλητής *Y*. Για το λόγο αυτό, όταν αναφερόμαστε στη σχέση που υπάρχει μεταξύ των τιμών των μεταβλητών, κάνουμε χρήση μιας επιπλέον μεταβλητής όπως και στο απλό γραμμικό υπόδειγμα, που καλείται σφάλμα και συμβολίζεται με το *ε*. Στην περίπτωση αυτή το γραμμικό μοντέλο παίρνει τη μορφή

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

όπου ε_i είναι οι αποκλίσεις των παρατηρούμενων τιμών της εξαρτημένης μεταβλητής *Y* από την τιμή που προβλέπεται από τις τιμές των ανεξάρτητων μεταβλητών $X_{i1},..,X_{i,p-1}$ μέσω του γραμμικού συνδυασμού

$$\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

Παραπάνω αναφέραμε τη μορφή του απλού και του πολλαπλού γραμμικού μοντέλου καθώς και κάποια συμπεράσματα για αυτά τα μοντέλα, στη συνέχεια θα ασχοληθούμε με την διαδικασία κατασκευής ενός γραμμικού μοντέλου. Η διαδικασία αυτή περιλαμβάνει τέσσερα στάδια :

1. Τη συλλογή και την προετοιμασία των δεδομένων
2. Τη μείωση του πλήθους των ανεξάρτητων μεταβλητών X_i
3. Την επιλογή του μοντέλου χωρίς περιττές μεταβλητές
4. Τον τελικό έλεγχο και την έγκριση του μοντέλου

Θα αναλύσουμε κάθε ένα από τα παραπάνω στάδια με τη σειρά.

Ο τρόπος συλλογής δεδομένων για τη διαδικασία κατασκευής γραμμικού μοντέλου ποικίλει ανάλογα με το είδος της μελέτης στην οποία ανήκουν τα δεδομένα. Παραθέτουμε τέσσερα βασικά είδη μελέτης.

- **Ελεγχόμενα πειράματα**: Σε πειράματα τέτοιου είδους οι ανεξάρτητες μεταβλητές χρησιμοποιούνται σε κάθε μια πειραματική μονάδα και στη συνέχεια παρατηρούνται οι τιμές που προκύπτουν για την εξαρτημένη

  μεταβλητή. Για παράδειγμα, ένα ελεγχόμενο πείραμα μπορεί να αναφέρεται στην επίδραση που ασκεί το *μέγεθος* της παρουσίασης ενός γραφήματος καθώς και *ο χρόνος* που χρειάζεται για να αναλυθεί πάνω στην *ακρίβεια* με την οποία εκτελείται η ανάλυση της παρουσίασης. Στη περίπτωση αυτή η εξαρτημένη μεταβλητή αναφέρεται στην ακρίβεια της ανάλυσης της παρουσίασης και οι

ανεξάρτητες μεταβλητές παίρνουν τις τιμές τους από το *μέγεθος* της παρουσίασης και το *χρόνο* που χρειάζεται για την ανάλυση του. Στα πειράματα αυτά οι ανεξάρτητες μεταβλητές καλούνται παράγοντες ή ελεγχόμενες μεταβλητές.

- **Ελεγχόμενα πειράματα με συμπληρωματικές μεταβλητές**: Στη προσπάθεια μείωσης των σφαλμάτων σε ένα γραμμικό μοντέλο γίνεται χρήση κάποιων συμπληρωματικών μεταβλητών. Οι μεταβλητές αυτές δεν είναι πάντα εύκολο να ενσωματωθούν στο γραμμικό μοντέλο. Στο προηγούμενο παράδειγμα θα μπορούσαμε να χρησιμοποιήσουμε ως συμπληρωματική μεταβλητή τα χρόνια σπουδών, κάτι που θα μπορούσε να επηρεάσει την εξαρτημένη μεταβλητή που εκφράζεται από την *ακρίβεια* με την οποία η ανάλυση της παρουσίασης εκτελείται καθώς θα έδινε πιο συγκεκριμένα αποτελέσματα κατά την ανάλυση της σχέσης των δύο ανεξαρτήτων μεταβλητών με την εξαρτημένη.

- **Μελέτες που προέρχονται από παρατηρήσεις**: Οι μελέτες αυτές προορίζονται να επιβεβαιώσουν ή όχι τις υποθέσεις που προέκυψαν σε προηγούμενες μελέτες ή στοιχεία. Τα δεδομένα για τις μελέτες αυτές συλλέγονται από ανεξάρτητες μεταβλητές για τις προηγούμενες μελέτες έχουν δείξει ότι επηρεάζουν την εξαρτημένη μεταβλητή καθώς και από των οποίων θέλουμε να διερευνήσουμε την επιρροή στην εξαρτημένη μεταβλητή. Οι μεταβλητές που προέκυψαν σε προηγούμενες μελέτες καλούνται **αρχικές μεταβλητές** ενώ οι μεταβλητές που απεικονίζουν την υπάρχουσα γνώση καλούνται **μεταβλητές ελέγχου**. Οι μεταβλητές ελέγχου χρησιμοποιούνται λόγω των επιρροών τους στην εξαρτημένη μεταβλητή. Για παράδειγμα σε μια μελέτη για την επιρροή της βιταμίνης Ε στην εμφάνιση ενός ορισμένου είδους καρκίνου, γνωστοί παράγοντες κινδύνου, όπως η ηλικία, το φύλο, και η φυλή, θα θεωρούνται ως μεταβλητές ελέγχου και η ποσότητα της βιταμίνης Ε που λαμβάνεται καθημερινά θεωρείται ως αρχική μεταβλητή. Η εξαρτημένη μεταβλητή είναι η εμφάνιση ενός συγκεκριμένου είδους καρκίνου κατά την διάρκεια λήψης της βιταμίνης Ε.

- **Ελεγχόμενες μελέτες που βασίζονται στις παρατηρήσεις**: Σε διάφορους τομείς όπως οι επιστήμες υγείας ή οι κοινωνικές επιστήμες δεν μπορούν να πραγματοποιηθούν ελεγχόμενα πειράματα δεδομένου ότι οι ανεξάρτητες μεταβλητές δεν είναι άμεσα μετρήσιμες. Αυτό έχει ως αποτέλεσμα μελέτες που

έχουν ως αντικείμενο κάποιον από αυτούς τους τομείς να κατατάσσονται σε μια κατηγορία που καλείται *Ελεγχόμενες μελέτες που βασίζονται στις παρατηρήσεις* όπου η μόνη διέξοδος για τους ερευνητές είναι να προσπαθήσουν να εντοπίσουν άλλες μεταβλητές που θα μπορούσαν πιθανώς να σχετίζονται με την εξαρτημένη μεταβλητή χρησιμοποιώντας κάποια μελέτη. Προφανώς, ένα τέτοιο σύνολο ενδεχομένως χρήσιμων ανεξάρτητων μεταβλητών μπορεί να είναι μεγάλο, κάποιες όμως από αυτές τις μεταβλητές μπορούν να απορριφθούν. Μια ανεξάρτητη μεταβλητή υπάρχει πιθανότητα να μην είναι θεμελιώδης για ένα πρόβλημα ή να υπόκειται σε μεγάλα λάθη μέτρησης, οπότε δεν είναι χρήσιμη. Θα μπορούσε όμως αποτελεσματικά να αναπαράγει μια άλλη μεταβλητή από το σύνολο. Οι ανεξάρτητες μεταβλητές που δεν μπορούν να μετρηθούν μπορούν είτε να διαγραφούν είτε να αντικατασταθούν από μεταβλητές που σχετίζονται αρκετά με τις αρχικές.

Αφού συλλεχθούν τα δεδομένα ακολουθείται μια διαδικασία ελέγχου των δεδομένων με χρήση διαγραμμάτων με τα οποία μπορεί να γίνει αντιληπτό κάθε λανθασμένο δεδομένο καθώς και οι πιθανές ακραίες τιμές. Οι δυσκολίες με τα λανθασμένα δεδομένα παρουσιάζονται κυρίως στα μεγάλα σύνολα δεδομένων. Όπου είναι δυνατόν, ο ερευνητής πρέπει να επιτηρήσει και να ελέγξει προσεκτικά τη διαδικασία συλλογής δεδομένων για να μειώσει την πιθανότητα καταγραφής λανθασμένων στοιχείων.

Η διαδικασία μείωσης των ανεξάρτητων μεταβλητών ποικίλει ανάλογα με το είδος της μελέτης που πραγματοποιείται. Θα αναφέρουμε ξεχωριστά τους τρόπους με τους οποίους μπορούν να μειωθούν οι μεταβλητές σε κάθε ένα από τα παραπάνω είδη μελέτης.

[1] **Ελεγχόμενα πειράματα**: Στα πειράματα αυτά δεν είναι ιδιαίτερα σημαντικό το θέμα τις μείωσης των μεταβλητών. Αφού επιλεχθούν οι μεταβλητές και δημιουργηθεί το μοντέλο ο ερευνητής είναι σε θέση να παρατηρήσει και να μελετήσει τη φύση και το μέγεθος των επιδράσεων που δέχεται η εξαρτημένη μεταβλητή από όλες τις ανεξάρτητες που έχουν επιλεγεί.

[2] **Ελεγχόμενα πειράματα με συμπληρωματικές μεταβλητές**: Στα πειράματα αυτά θα ήταν χρήσιμο να μειωθούν κάποιες από τις συμπληρωματικές

μεταβλητές για τις οποίες δεν μπορούμε να είμαστε απόλυτα σίγουροι ότι βοηθούν στη μείωση των σφαλμάτων. Το σύνολο των συμπληρωματικών μεταβλητών που χρησιμοποιούνται σε ένα γραμμικό μοντέλο είναι συνήθως μικρό, επομένως δε δημιουργείται κάποιο πρόβλημα εάν κάποιες ή όλες οι συμπληρωματικές μεταβλητές απορριφθούν από το μοντέλο. Προφανώς θα απορρίπτονται οι μεταβλητές εκείνες οι οποίες δεν επηρεάζουν σημαντικά την τιμή της εξαρτημένης μεταβλητής.

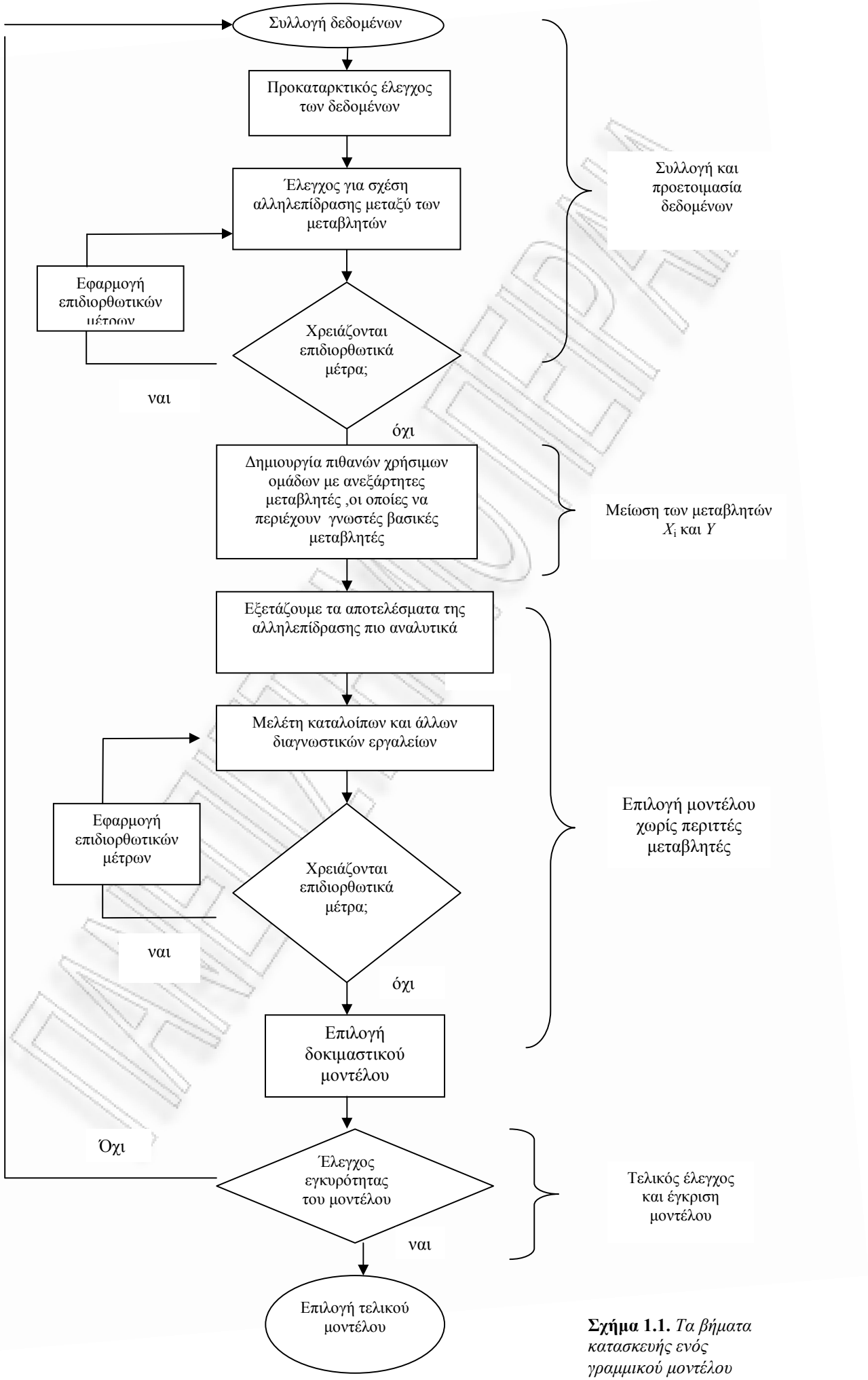
[3] **Μελέτες που προέρχονται από παρατηρήσεις**: Σε μελέτες τέτοιου είδους δεν μπορεί να γίνει καμία μείωση μεταβλητών. Οι **μεταβλητές ελέγχου** οι οποίες έχουν επιλεχθεί με βάση μια προηγούμενη γνώση θα πρέπει να διατηρηθούν με σκοπό την σύγκριση τους με πιο πρόσφατες μελέτες ακόμα και αν αυτές δεν οδηγούν στην μείωση του σφάλματος που παρουσιάζουν οι μεταβλητές. Οι **αρχικές μεταβλητές** είναι αυτές που επηρεάζουν την εξαρτημένη μεταβλητή και για το λόγο αυτό πρέπει να υπάρχουν στο γραμμικό μοντέλο.

[4] **Ελεγχόμενες μελέτες που βασίζονται στις παρατηρήσεις**: Στις μελέτες τέτοιου είδους είναι απαραίτητη η μείωση των ανεξάρτητων μεταβλητών καθώς η παρουσία τους στο γραμμικό μοντέλο προκαλεί την αύξηση του δείγματος των συντελεστών του γραμμικού μοντέλου, μειώνει την περιγραφική ικανότητα του μοντέλου και δυσκολεύει την ικανότητα πρόβλεψης. Η ικανότητα πρόβλεψης του μοντέλου δυσχεραίνεται όταν οι ανεξάρτητες μεταβλητές που διατηρούνται στο μοντέλο δεν σχετίζονται με την εξαρτημένη μεταβλητή.

Το επόμενο στάδιο μετά την μείωση των ανεξάρτητων μεταβλητών είναι η δημιουργία γραμμικών μοντέλων και η επιλογή ενός «καλού» μοντέλου. Η επιλογή του «καλού» γραμμικού μοντέλου εξαρτάται άμεσα από τα κριτήρια που χρησιμοποιούνται κάθε φορά. Για παράδειγμα, ένα επιλεγμένο γραμμικό μοντέλο μπορεί να επηρεάζεται αρκετά από μια μεμονωμένη περίπτωση ενώ κάποιο άλλο να μην επηρεάζεται ή να παρουσιάζει μεγαλύτερη συσχέτιση μεταξύ των σφαλμάτων σε αντίθεση με κάποιο άλλο γραμμικό μοντέλο. Ανεξάρτητα της επιλογής κάποιου μοντέλου ως «καλό» θα πρέπει να ελεγχθούν και τα υπόλοιπα. Μετά από ένα λεπτομερή έλεγχο και τη χρήση διαφόρων διορθωτικών μέσων, όπως οι μετασχηματισμοί, καταλήγουμε στο τελικό μοντέλο το οποίο θα χρησιμοποιηθεί.

Το επόμενο στάδιο είναι η επικύρωση του μοντέλου, αναφέρεται στην σταθερότητα και την λογική των συντελεστών του γραμμικού μοντέλου, στην ικανότητα των λειτουργιών του μοντέλου και στην ικανότητα να γενικευτούν τα συμπεράσματα που προέρχονται από την ανάλυση του γραμμικού μοντέλου. Η επικύρωση είναι ένα χρήσιμο και απαραίτητο μέρος στη διαδικασία κατασκευής ενός γραμμικού μοντέλου.

Παραθέτουμε στην επόμενη σελίδα ένα διάγραμμα με τα βήματα κατασκευής του γραμμικού μοντέλου.



Σχήμα 1.1. *Τα βήματα κατασκευής ενός γραμμικού μοντέλου*

1.2 Η χρησιμότητα επιλογής ενός υποσυνόλου εξαρτημένων μεταβλητών για την εκτίμηση της ανεξάρτητης μεταβλητής

Υπάρχουν περιπτώσεις που καθιστούν αναγκαία την επιλογή ενός μικρού υποσυνόλου από ένα μεγαλύτερο σύνολο μεταβλητών οι οποίες χρησιμοποιούνται για την πρόβλεψη μιας εξαρτημένης μεταβλητής. Για παράδειγμα, η διαδικασία πρόβλεψης της εξαρτημένης μεταβλητής Y μπορεί να θεωρείται οικονομικά ασύμφορη και αρκετά χρονοβόρα εφόσον χρησιμοποιηθούν όλες οι τιμές που έχουμε για τις ανεξάρτητες μεταβλητές, για το λόγο αυτό αναζητούμε ένα μικρό υποσύνολο μεταβλητών που θα μπορεί με αρκετή ακρίβεια να προβλέψει την τιμή της μεταβλητής Y. Η επιλογή αυτού του συνόλου θα μπορούσε προφανώς στη συνέχεια να χρησιμοποιηθεί για κάποια μελλοντική πρόβλεψη εφόσον τα δεδομένα είναι αντιπροσωπευτικά των συνθηκών υπό τις οποίες θα γίνει η πρόβλεψη. Από την άλλη πλευρά, στην προσπάθεια μας να κατανοήσουμε την επίδραση που έχει η μια μεταβλητή πάνω στη άλλη, ιδιαίτερα όταν τα δεδομένα που έχουμε συλλέξει είναι μέσω παρατήρησης ή μέσω έρευνας και όχι δεδομένα που προήλθαν μέσω πειραμάτων, μπορεί να είναι επιθυμητό να μην χρησιμοποιήσουμε πολλές μεταβλητές οι οποίες έχουν κάποια ισχυρή επίδραση μεταξύ τους.

Κάποιες φορές τα δεδομένα για την πρόβλεψη που θέλουμε να κάνουμε έχουν ήδη συλλεχθεί για προηγούμενες προβλέψεις ή για άλλους σκοπούς με αποτέλεσμα να μην υπάρχει επιπλέον κόστος εάν συμπεριληφθούν στο μοντέλο πρόβλεψης. Αυτό συμβαίνει συχνά με τα μετεωρολογικά δεδομένα ή σε έρευνες που κάνει η κυβέρνηση για οικονομικές προβλέψεις. Σε άλλες περιπτώσεις μπορεί να υπάρχει σημαντικό επιπλέον κόστος εάν χρησιμοποιηθούν όλα τα δεδομένα που έχει ως αποτέλεσμα την εξέταση του κόστους σε σχέση με την ακρίβεια των προβλέψεων.

Σε γενικές γραμμές στη συνέχεια θα θεωρούμε ότι όλες οι μεταβλητές μπορούν να εισαχθούν καθώς και να διαγραφούν από το γραμμικό μοντέλο, αν αυτό κρίνεται αναγκαίο. Φυσικά αυτό δεν συμβαίνει πάντοτε στη πράξη, ενώ πολλές φορές χρησιμοποιούμε νέες μεταβλητές που κατασκευάζονται από τις ήδη υπάρχουσες, για παράδειγμα τα τετράγωνα των μεταβλητών ή τις αλληλεπιδράσεις μεταξύ δύο ή περισσοτέρων μεταβλητών.

1.3 Το πλήθος των μεταβλητών που πρέπει να υπάρχουν σε ένα γραμμικό μοντέλο

Το ιδανικό για την πρόβλεψη της εξαρτημένης μεταβλητής Y μέσω ενός γραμμικού μοντέλου είναι να περιλαμβάνει όλες εκείνες τις ανεξάρτητες μεταβλητές που επηρεάζουν την μεταβλητή Y, κάτι τέτοιο όμως, όπως αναφέρθηκε προηγουμένως, θα πρέπει να εξεταστεί διεξοδικά και να αποφασισθεί αν είναι αναγκαίο από πρακτική άποψη.

Ας υποθέσουμε ότι η εξαρτημένη μεταβλητή Y συνδέεται γραμμικά με τις ανεξάρτητες μεταβλητές X_1 , X_2 , … , X_k μέσω της σχέσης

$$Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \varepsilon$$

όπου τα κατάλοιπα ε έχουν μέση τιμή μηδέν και διασπορά σ^2. Οι συντελεστές $\beta_0, \beta_1, .., \beta_k$ είναι συνήθως άγνωστοι και για την εκτίμησή τους χρησιμοποιούμε την μέθοδο των ελάχιστων τετραγώνων. Οι εκτιμήσεις των συντελεστών παλινδρόμησης οι οποίες συμβολίζονται συνήθως με b(ή με $\hat{\beta}$), δίνονται από τον τύπο

$$b = (X'X)^{-1} X'y$$

όπου

$$b' = (b_0, b_1, ..., b_k)\,,$$

X είναι ένα πίνακας $n \times (k+1)$ και y είναι ένα διάνυσμα μήκους n που περιέχει τις τιμές που έχουν παρατηρηθεί και οι οποίες στην συνέχεια θα εκτιμηθούν. Αποδεικνύεται ότι, για τη διασπορά της σημειακής εκτιμήτριας $\hat{Y} = x'b$ όπου $x = (1, x_1, ..., x_k)$, $b' = (b_0, b_1, ..., b_k)$ ισχύει

$$\text{var}(x'b) = \sigma^2 x'(X'X)^{-1} x\,.$$

Εφαρμόζοντας παραγοντοποίηση Cholesky στον πίνακα $(X'X)^{-1}$ μπορούμε να γράψουμε

$$(X'X)^{-1} = R^{-1}(R^{-1})'$$

όπου R είναι ένας άνω τριγωνικός $(k+1)\times(k+1)$ πίνακας. Επομένως

$$\text{var}(x'b) = \sigma^2 (x'R^{-1})(x'R^{-1})'$$

και έτσι η διακύμανση της εκτιμώμενης τιμής $x'b$ του Y παίρνει τη πιο βολική μορφή ενός αθροίσματος τετραγώνων (των στοιχείων του διανύσματος $x'R^{-1}$).

Προκειμένου να εξετάσουμε την πρόβλεψη της μεταβλητής Y χρησιμοποιώντας μόνο τις πρώτες p ανεξάρτητες μεταβλητές ($p < k$) θεωρούμε τον πίνακα

$$X = (X_A, X_B)$$

όπου X_A είναι ένας πίνακας που περιέχει τις πρώτες p+1 στήλες του X και ο πίνακας X_B περιέχει τις υπόλοιπες $k-p$ στήλες.

Αν εφαρμόσουμε την παραγοντοποίηση Cholesky θα έχουμε

$$X'_A X_A = R'_A R_A$$

όπου R_A είναι ένας πίνακας που περιέχει τις πρώτες p+1 γραμμές και στήλες του άνω τριγωνικού πίνακα R. Θεωρούμε στη συνέχεια ένα διάνυσμα x_A που περιέχει τα πρώτα p+1 στοιχεία του $x = (1, x_1, ..., x_k)$ και εφαρμόζοντας τη μέθοδο των ελαχίστων τετραγώνων στο μοντέλο που χρησιμοποιεί τις p πρώτες μόνο μεταβλητές θα έχουμε τη σχέση

$$\text{var}(x'_A b_A) = \sigma^2 (x'_A R_A^{-1})(x'_A R_A^{-1})'$$

η οποία αποτελεί την διακύμανση των νέων προβλεπόμενων τιμών της μεταβλητής Y. Επομένως

$$\text{var}(x'b) \geq \text{var}(x'_A b_A),$$

δηλαδή παρατηρούμε πως, στα γραμμικά μοντέλα που οι παράμετροι έχουν εκτιμηθεί με τη μέθοδο των ελαχίστων τετραγώνων, η διακύμανση των προβλεπόμενων τιμών αυξάνεται καθώς αυξάνεται ο αριθμός των μεταβλητών που χρησιμοποιείται για την πρόβλεψη.

Η παραπάνω σχέση της διακύμανσης με το πλήθος των μεταβλητών είναι κάπως δύσκολο να κατανοηθεί αφού μάλιστα θα μπορούσαμε ως συνέπεια να σκεφθούμε ότι η καλύτερη πρόβλεψη γίνεται στην ακραία περίπτωση που έχουμε ένα

μοντέλο χωρίς μεταβλητές! Ωστόσο, αν για παράδειγμα η τιμή πρόβλεψης για τη μεταβλητή Y είναι πάντα η ίδια ανεξάρτητα από τις τιμές των μεταβλητών X τότε ναι μεν η διακύμανση για τις προβλέψεις των τιμών της μεταβλητής Y είναι ίση με το μηδέν αλλά παράλληλα θα αυξηθεί υπερβολικά η μεροληψία.

Εάν το πραγματικό μοντέλο δίνεται από τη σχέση $Y = \beta_0 + \sum_{i=1}^{k} \beta_i X_i + \varepsilon$, τότε

$$b_A = (X'_A X_A)^{-1} X'_A y$$

οπότε

$$E(b_A) = (X'_A X_A)^{-1} X'_A X\beta = (X'_A X_A)^{-1} X'_A (X_A, X_B)\beta =$$
$$= (X'_A X_A)^{-1} (X'_A X_A, X'_A X_B)\beta = \beta_A + (X'_A X_A)^{-1} X'_A X_B \beta_B$$

όπου β_A, β_B είναι διανύσματα που περιέχουν αντίστοιχα τα πρώτα $(p+1)$ και τα τελευταία $(k-p)$ στοιχεία του διανύσματος β. Η παραπάνω σχέση δίνει την μεροληψία των πρώτων $(p+1)$ συντελεστών παλινδρόμησης που προκύπτουν από την παράλειψη των τελευταίων $(k-p)$ μεταβλητών. Η μεροληψία για την εκτίμηση της μεταβλητής Y για δοσμένο x είναι

$$x'\beta - E(x'_A b_A) = x'_A \beta_A + x'_B \beta_B - x'_A \beta_A + x'_A (X'_A X_A)^{-1} X'_B X_B \beta_B =$$
$$= \{x'_B - x'_A (X'_A X_A)^{-1} X'_A X_B\}\beta_B .$$

Η προσθήκη μεταβλητών στο μοντέλο έχει ως αποτέλεσμα τη μείωσης της μεροληψίας σε σχέση με την αύξηση της διακύμανσης. Εάν εισάγουμε στο μοντέλο μια μεταβλητή που δεν έχει χρησιμοποιηθεί μέχρι στιγμής, τότε θα παρατηρήσουμε αύξηση της διακύμανσης. Όμως αν η προσθήκη κάποιας μεταβλητής δημιουργεί μεταβολή στη μεροληψία τότε η αύξηση της τιμής που παρουσιάζει η εκτίμηση της διακύμανσης μπορεί να υπερβαίνει το όφελος από την μείωση της μεροληψίας. Το ερώτημα που προκύπτει είναι πως θα μπορέσουμε να εξισορροπήσουμε τη σχέση ανάμεσα στη διακύμανση και τη μεροληψία. Το πρόβλημα αρχίζει να δημιουργείται από τη στιγμή που το μοντέλο δεν έχει επιλεγεί ανεξάρτητα από τα δεδομένα τα οποία εισάγονται σε αυτό.

Σημειώνουμε πως η εισαγωγή επιπλέον μεταβλητών δεν συνεπάγεται απαραίτητα και την μείωση της μεροληψίας, αυτό συνδέεται άμεσα με την επιλογή του συνόλου μεταβλητών που θα χρησιμοποιηθούν για την πρόβλεψη.

Οι διάφορες μέθοδοι επιλογής βέλτιστου συνόλου ανεξάρτητων μεταβλητών βασίζονται στις παραπάνω παρατηρήσεις για την επιλογή του τελικού μοντέλου.

1.4 Επιλογή βέλτιστου μοντέλου με χρήση στατιστικών προγραμμάτων

Η εφαρμογή μιας διαδικασίας επιλογής του καλύτερου υποσυνόλου ανεξάρτητων μεταβλητών σε ένα γραμμικό μοντέλο έχει συνήθως νόημα εάν έχουμε στη διάθεσή μας μεγάλο αριθμό ανεξάρτητων μεταβλητών. Στη περίπτωση αυτή είναι σχεδόν απαραίτητο να καταφύγουμε στη χρήση στατιστικών πακέτων αφού ο όγκος των αριθμητικών πράξεων είναι υπερβολικά μεγάλος για να γίνει με το χέρι ή με υπολογιστές χειρός (κομπιουτεράκια).

Σήμερα όλα σχεδόν τα κλασσικά στατικά πακέτα (SPSS, MINITAB, Statgraphics, SAS, S-plus, R) έχουν ενσωματωμένες διαδικασίες αυτόματης επιλογής βέλτιστου συνόλου ανεξάρτητων μεταβλητών για γραμμικά μοντέλα. Για τα παραδείγματα που θα επεξεργαστούμε αναλυτικά στη συνέχεια θα χρησιμοποιήσουμε το στατιστικό πακέτα SPSS επισημαίνοντας ότι οι διαφορές αν ήθελε κανείς να καταφύγει σε κάποιο άλλο πακέτο θα ήταν πολύ μικρές.

Η άμεση διάθεση στατιστικών πακέτων σε ηλεκτρονικούς υπολογιστές με δυνατότητα αυτόματης επιλογής βέλτιστων γραμμικών μοντέλων ενθαρρύνει την «τυφλή» χρήση μεθόδων επιλογής, με αποτέλεσμα το τελικό μοντέλο να μην είναι τις πιο πολλές φορές ούτε λογικό ούτε αποτελεσματικό. Η υψηλή ταχύτητα των υπολογιστών σε συνδυασμό με τη χρήση αποτελεσματικών αλγορίθμων σημαίνει ότι είναι εφικτό να βρεθεί ένα υποσύνολο μεταβλητών, από ένα σύνολο των 10 έως και 150 μεταβλητών, το οποίο θεωρείται ικανό για την πρόβλεψη του γραμμικού μοντέλου σύμφωνα με τη μέθοδο των ελαχίστων τετραγώνων. Μια από τις πιο

σημαντικές ευκολίες που θα πρέπει να προσφέρει στον χρήστη ένα στατιστικό πακέτο είναι η δημιουργία μιας οικογένειας εναλλακτικών μοντέλων ώστε ο τελευταίος να μπορεί να αποφασίσει σε ποιο ταιριάζουν επαρκώς τα δεδομένα του.

Συνήθως, παρά το μεγάλο αριθμό των μεταβλητών και συνεπώς το μεγάλο αριθμό των πιθανών γραμμικών μοντέλων, το καλύτερο υποσύνολο που επιλέγεται για τη πρόβλεψη μπορεί να μη δίνει την καλύτερη προσαρμογή στα δεδομένα του δείγματος. Στη πράξη, τα υποσύνολα που επιλέγονται θα πρέπει να εξετάζονται λεπτομερώς πριν χρησιμοποιηθούν για την πρόβλεψη του γραμμικού μοντέλου.

Μια αποτελεσματική «ενδιάμεση» λύση είναι να επιλέξουμε αρχικά και να εφαρμόσουμε (πάντα με τη χρήση στατιστικού πακέτου) μια μέθοδο επιλογής υποσυνόλου μεταβλητών (ως κριτήριο επιλογής θα μπορούσαμε να θεωρήσουμε για παράδειγμα το μικρό υπολογιστικό κόστος) και στη συνέχεια τις υπόλοιπες, έτσι ώστε θα μπορούμε να διακρίνουμε τις κυρίαρχες μεταβλητές στο μοντέλο μας με κάθε μέθοδο. Στη συνέχεια μπορούμε να συνδυάσουμε όλες αυτές τις μεταβλητές για να καταλήξουμε στο τελικό υποσύνολο μεταβλητών που θα χρησιμοποιήσουμε.

Εάν με την πρώτη μέθοδο που θα υλοποιήσουμε παρατηρήσουμε μικρή μείωση στην τιμή του αθροίσματος των τετραγώνων των υπολοίπων χρησιμοποιώντας κάποιες από τις μεταβλητές αντί για όλες τις διαθέσιμες μεταβλητές τότε θα προσπαθήσουμε να περιορίσουμε το σύνολο στο μικρό αυτό αριθμό μεταβλητών ή σε ακόμα μικρότερο αριθμό ελέγχοντας ξανά τις μεταβλητές.

Η «παραδοσιακή» προσέγγιση όσον αφορά την κατασκευή ενός γραμμικού μοντέλου έχει σημειώσει κάποια πρόοδο στο σχεδιασμό και τη συσχέτιση της εξαρτημένης με τις ανεξάρτητες μεταβλητές. Κάτι τέτοιο επιτυγχάνεται με τη χρήση αλγορίθμων ή διάφορων άλλων μετασχηματισμών που έχουν ως σκοπό να προσεγγίσουν με ακρίβεια τη γραμμικότητα και την ομοιογένεια της διακύμανσης. Εν συνεχεία επιλέγονται οι σημαντικότερες από τις μεταβλητές που υπάρχουν στα δεδομένα και εισάγονται στο μοντέλο. Η διαδικασία αυτή είναι εφικτή όταν τα δεδομένα προέρχονται από καλά σχεδιασμένο πείραμα, ενώ δύσκολα εφαρμόζονται σε δεδομένα παρατήρησης από την πράξη όπου οι ανεξάρτητες μεταβλητές είναι υψηλά συσχετισμένες.

1.5 Περιεχόμενα της διπλωματικής

Η παρούσα διπλωματική εργασία χωρίζεται σε τέσσερις κύριες ενότητες-Κεφάλαια.

Στο Κεφάλαιο 1 δόθηκαν κάποια εισαγωγικά στοιχεία για τη γραμμική παλινδρόμηση και έγινε αναφορά στο πρόβλημα της επιλογής ενός βέλτιστου συνόλου ανεξάρτητων μεταβλητών έτσι ώστε αφενός μεν να υπάρχει εξοικονόμηση κόστους κατά την πρόβλεψη της εξαρτημένης μεταβλητής, αφετέρου δε να μην προκύπτει μεγάλη απώλεια στην αποτελεσματικότητα του μοντέλου πρόβλεψης..

Στο Κεφάλαιο 2 παρουσιάζεται η μέθοδος της εξέτασης όλων των δυνατών γραμμικών μοντέλων και δίνεται ένα συγκεκριμένο παράδειγμα από πραγαμτικά δεδομένα που θα χρησιμοποιηθεί στη συνέχεια για την επιλογή βέλτιστου συνόλου μεταβλητών. Στα δεδομένα αυτά εφαρμόζονται τα διάφορα κριτήρια που έχουν προταθεί για την υλοποίηση της μεθόδου «εξέτασης όλων των δυνατών μοντέλων» και πιο συγκεκριμένα το κριτήριο R^2, το κριτήριο *SSE*, το κριτήριο R^2_{adj}, το κριτήριο *MSE*, το κριτήριο C_p του Mallows, το κριτήριο *PRESS*, το κριτήριο *AIC*, το κριτήριο *BIC* και το κριτήριο S_p.

Στο Κεφάλαιο 3 περιγράφονται και εφαρμόζονται επαναληπτικές μέθοδοι με της οποίες δημιουργούνται αυτόματα υποσύνολα ανεξάρτητων μεταβλητών που προσεγγίζουν ή συμπίπτουν με το βέλτιστο μοντέλου. Ειδικότερα, παρουσιάζεται η μέθοδος Forward Ranking, η μέθοδος Backward Ranking, η μέθοδος Stepwise Regression, η μέθοδος Backward Elimination, η μέθοδος Forward Selection, η μέθοδος Forward Procedure, η μέθοδος της διαδοχικής αντικατάστασης μεταβλητών και τέλος η μέθοδος της αντικατάστασης δύο μεταβλητών.

Στο τελευταίο κεφάλαιο (Κεφάλαιο 4) δίνονται ορισμένα παραδείγματα από τη διεθνή βιβλιογραφία στα οποία γίνεται σύγκριση και μελέτη της αποτελεσματικότητας των μεθόδων επιλογής βέλτιστου συνόλου ανεξάρτητων μεταβλητών σε ένα γραμμικό μοντέλο.

ΚΕΦΑΛΑΙΟ 2

Η μέθοδος της εξέτασης όλων των δυνατών μοντέλων

2.1 Κριτήρια βελτιστότητας

Σκοπός της μεθόδου «εξέτασης όλων των δυνατών μοντέλων» είναι να μελετηθούν όλα τα δυνατά γραμμικά μοντέλα που μπορούν να δημιουργηθούν με χρήση των διαθέσιμων μεταβλητών (δηλαδή μοντέλα με καμία, μία, δύο, κλπ ανεξάρτητες μεταβλητές) και με τη χρήση κατάλληλων κριτηρίων, που θα παρουσιάσουμε παρακάτω, να προσδιορίσουμε αρχικά μια μικρή ομάδα γραμμικών μοντέλων που θεωρούνται «καλά» και, τέλος, να καταλήξουμε στο βέλτιστο γραμμικό μοντέλο.

Στη διαδικασία επιλογής ενός γραμμικού μοντέλου συνήθως εμπλέκονται δύο αντιτιθέμενα κριτήρια. Από τη μία μεριά για την κατασκευή ενός γραμμικού μοντέλου με σκοπό την πρόβλεψη είναι χρήσιμο να έχουμε όσο το δυνατόν περισσότερες ανεξάρτητες μεταβλητές έτσι ώστε οι προσαρμοσμένες τιμές (εκτιμήσεις) να είναι αξιόπιστες. Από την άλλη μεριά όμως επειδή η συγκέντρωση πολλών ανεξάρτητων μεταβλητών και η επακόλουθη επεξεργασία τους κοστίζει, θα θέλαμε το γραμμικό μοντέλο να περιλαμβάνει όσο το δυνατό λιγότερες μεταβλητές.

Ο συμβιβασμός μεταξύ αυτών των δύο ακραίων περιπτώσεων είναι αυτό που συνήθως ονομάζεται επιλογή του *καλύτερου γραμμικού μοντέλου με τη χρήση γραμμικής παλινδρόμησης*. Για να πετύχουμε το καλύτερο μοντέλο δεν υπάρχει μια

μοναδική στατιστική διαδικασία. Παρακάτω θα περιγράψουμε διάφορες διαδικασίες που έχουν προταθεί, είναι γεγονός πως όταν όλες αυτές οι διαδικασίες εφαρμοστούν στο ίδιο πρόβλημα δεν οδηγούν αναγκαστικά στην επιλογή του ίδιου γραμμικού μοντέλου.

Στο κεφάλαιο αυτό θα μελετήσουμε όλους τους δυνατούς συνδυασμούς των ανεξάρτητων μεταβλητών, με σκοπό την επιλογή του βέλτιστου γραμμικού μοντέλου, εφαρμόζοντας τα εξής κριτήρια:

α. Συντελεστής προσδιορισμού R_p^2

β. Τροποποιημένος συντελεστής προσδιορισμού R_{adj}^2

γ. Error sum of squares SSE_p

δ. mean square error MSE_p

ε. C_p του Mallows

ζ. Mean square error of prediction $PRESS_p$

η. Akaike's information criterion AIC

θ. Bayesian information criterion BIC.

Τα κριτήρια αυτά επηρεάζονται άμεσα από τον αριθμό των παραμέτρων του γραμμικού μοντέλου. Για κάποια από αυτά αυξάνεται η τιμή τους καθώς εισάγεται νέα μεταβλητή στο μοντέλο (δηλαδή όταν αυξάνεται ο αριθμός των παραμέτρων β_i) για κάποια άλλα μειώνεται η τιμή του κριτηρίου, υπάρχουν όμως και κάποια κριτήρια που η τιμή τους απλά επηρεάζεται από τον αριθμό τον παραμέτρων χωρίς να υπάρχει κάποια σχέση μονοτονίας που να τα συνδέει.

Τα κριτήρια θα αναλυθούν διεξοδικά στις επόμενες παραγράφους. Στην επόμενη παράγραφο θα δοθεί ένα παράδειγμα το οποίο θα χρησιμοποιηθεί στη συνέχεια για να παρουσιαστεί αναλυτικά η εφαρμογή καθενός κριτηρίου. Για την επεξεργασία των δεδομένων θα χρησιμοποιήσουμε το στατιστικό πρόγραμμα *SPSS.*

2.2 Ένα παράδειγμα

Για την καλύτερη κατανόηση της διαδικασίας της επιλογής βέλτιστου μοντέλου θα χρησιμοποιήσουμε ένα παράδειγμα που ανήκει στη κατηγορία **ελεγχόμενες μελέτες που βασίζονται στις παρατηρήσεις**. Το παράδειγμα αναφέρεται στη πρόβλεψη της εξαρτημένης μεταβλητής

Y: απόσταση (σε μίλια) που διανύει ένα αυτοκίνητο με ένα γαλόνι βενζίνης (μίλια ανά γαλόνι, mpg).

Θα εξετάσουμε διάφορους παράγοντες που ίσως επιδρούν στην εξαρτημένη μεταβλητή *Y*, πιο συγκεκριμένα τους επόμενους πέντε:

X_1: πλήθος κυλίνδρων

X_2: κυβικά (κυβικές ίντσες, 1cubic inch=16,39ml)

X_3: ίπποι

X_4: βάρος αυτοκινήτου(lbs, 1lb=0,454gr)

X_5: μέση επιτάχυνση (μίλια ανά sec)

Εξετάσαμε ένα δείγμα 30 αυτοκινήτων, το οποίο επιλέχθηκε τυχαία από τη βάση δεδομένων «1983 ASA Data Exposition dataset» η οποία περιέχει τιμές για 406 διαφορετικά αυτοκίνητα. Η βάση αυτή δημιουργήθηκε από τους Ernesto Ramos and David Donoho και βρίσκεται στην ηλεκτρονική σελίδα:

http://lib.stat.cmu.edu/datasets/cars.data.

Οι τιμές για τις ανεξάρτητες μεταβλητές που θα εξετάσουμε καθώς και για την εξαρτημένη μεταβλητή παρουσιάζονται στον πίνακα της επόμενης σελίδας.

Χρησιμοποιώντας ως ανεξάρτητες μεταβλητές τις X_1, X_2, X_3, X_4, X_5 μπορούν να διαμορφωθούν συνολικά $2^5 - 1 = 31$ διαφορετικά μοντέλα στα οποία και θα εφαρμόσουμε διάφορα κριτήρια ώστε να καταλήξουμε στο βέλτιστο μοντέλο. Σε αυτά θα πρέπει να προστεθεί και ένα (τετριμμένο) γραμμικό μοντέλο της μορφής $Y_i = \beta_0 + \varepsilon_i$ στο οποίο δεν χρησιμοποιείται καμία μεταβλητή.

Πίνακας 2.1: *Δεδομένα του παραδείγματος κατανάλωσης βενζίνης*

α/α	Y	X_1	X_2	X_3	X_4	X_5
1	18	8	307	130	3504	12
2	15	8	350	165	3693	11,5
3	18	8	318	150	3436	11
4	16	8	304	150	3433	12
5	17	8	302	140	3449	10,5
6	15	8	429	198	4341	10
7	14	8	454	220	4354	9
8	14	8	440	215	4312	8,5
9	14	8	455	225	4425	10
10	15	8	390	190	3850	8,5
11	15	8	383	170	3563	10
12	14	8	340	160	3609	8
13	15	8	400	150	3761	9,5
14	14	8	455	225	3086	10
15	24	4	113	95	2372	15
16	22	6	198	95	2833	15,5
17	18	6	199	97	2774	15,5
18	21	6	200	85	2587	16
19	27	4	97	88	2130	14,5
20	26	4	97	46	1835	20,5
21	25	4	110	87	2672	17,5
22	24	4	107	90	2430	14,5
23	25	4	104	95	2375	17,5
24	26	4	121	113	2234	12,5
25	21	6	199	90	2648	15
26	10	8	360	215	4615	14
27	10	8	307	200	4376	15
28	11	8	318	210	4382	13,5
29	9	8	304	193	4732	18,5
30	27	4	97	88	2130	14,5

Πιο συγκεκριμένα παρατηρούμε ότι υπάρχουν

$$\binom{5}{1} = 5$$

γραμμικά μοντέλα με μια μόνο μεταβλητή, της μορφής

$$Y_i = \beta_0 + \beta_1 X_{ij} + \varepsilon_i .$$

Όμοια υπάρχουν

$$\binom{5}{2} = 10$$

γραμμικά μοντέλα που περιέχουν δύο από τις πέντε ανεξάρτητες μεταβλητές, της μορφής

$$Y_i = \beta_0 + \beta_1 X_{ij} + \beta_2 X_{ik} + \varepsilon_i$$

όπου

$$i \in \{1,...,30\} \text{ και } j,k \in \{1,2,3,4,5\},\ j \neq k\,.$$

Με συνδυασμούς τριών ανεξάρτητων μεταβλητών μπορούν να δημιουργηθούν και πάλι δέκα γραμμικά μοντέλα, της μορφής

$$Y_i = \beta_0 + \beta_1 X_{ij} + \beta_2 X_{ik} + \beta_3 X_{iz} + \varepsilon_i$$

όπου

$$i \in \{1,...,30\}\ ,\ j,k,z \in \{1,2,...,5\} \text{ και } j \neq k,\quad k \neq z$$

αφού

$$\binom{5}{3} = \binom{5}{2} = 10$$

Συνδυάζοντας τέσσερις ανεξάρτητες μεταβλητές θα πάρουμε

$$\binom{5}{4} = 5$$

γραμμικά μοντέλα, της μορφής

$$Y_i = \beta_0 + \beta_1 X_{ij} + \beta_2 X_{ik} + \beta_3 X_{iz} + \beta_4 X_{im} + \varepsilon_i$$

όπου

$i \in \{1,...,30\}\ ,\ j,k,z,m \in \{1,2,...,5\}$ και $j \neq k,\quad j \neq z,\quad j \neq m,\quad k \neq z,\quad k \neq m,\quad z \neq m,\quad .$

Τέλος υπάρχει ένα μόνο μοντέλο με όλες τις μεταβλητές, το πλήρες μοντέλο, που θα έχει τη μορφή

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \varepsilon_i$$

Στις επόμενες παραγράφους αυτού του κεφαλαίου θα περιγράψουμε αναλυτικά τα κριτήρια επιλογής βέλτιστου γραμμικού μοντέλου και για κάθε ένα θα εφαρμόσουμε τη μέθοδο στα δεδομένα του Πίνακα 2.1.

2.3 Το κριτήριο R^2

Σύμφωνα με το κριτήριο R^2 ένα γραμμικό μοντέλο θα έπρεπε να θεωρείται βέλτιστο εάν δίνει στο συντελεστή προσδιορισμού R^2 τη μεγαλύτερη δυνατή τιμή. Επειδή όμως το R^2 αυξάνει συνεχώς, όσο αυξάνει το πλήθος των ανεξάρτητων μεταβλητών που υπάρχουν στο γραμμικό μοντέλο θα προέκυπτε κάθε φορά ως βέλτιστο το πλήρες μοντέλο.

Μία τέτοια επιλογή είναι χρονοβόρα όσον αφορά τη μελέτη του μοντέλου και δεν έχει καμία πρακτική αξία, για το λόγο αυτό εφαρμόζουμε το κριτήριο R^2 σε όλα τα διαφορετικά υποσύνολα, ανάλογα με το πλήθος των μεταβλητών δηλαδή ανά μια, δύο, τρεις, τέσσερις κτλ ανεξάρτητες μεταβλητές και τέλος στο πλήρες μοντέλο. Σε κάθε ένα υποσύνολο αναζητούμε τη μεγαλύτερη τιμή για το κριτήριο R^2. Στη συνέχεια συγκρίνουμε τις μέγιστες τιμές και εξετάζουμε τις μεταβολές που γίνονται στη τιμή του κριτηρίου από το ένα υποσύνολο στο άλλο. Εάν η προσθήκη μιας μεταβλητής στο γραμμικό μοντέλο δεν επιφέρει σημαντική αύξηση στη τιμή του R^2 τότε θεωρούμε ως βέλτιστο μοντέλο εκείνο που αφενός δίνει μέγιστη τιμή στο κριτήριο R^2, στο υποσύνολο που ανήκει, και αφετέρου διατηρεί ένα μικρό αριθμό μεταβλητών.

Αν εφαρμόσουμε το κριτήριο σε όλα τα γραμμικά μοντέλα που μπορούν να προκύψουν για τα δεδομένα του Παραδείγματος της Παραγράφου 2.2 παίρνουμε τις για το κριτήριο R^2 τις τιμές που παρουσιάζονται στον Πίνακα 2.2. Για κάθε συγκεκριμένο πλήθος ανεξάρτητων μεταβλητών έχει σημειωθεί με έντονα γράμματα η παρατηρηθείσα μεγαλύτερη τιμή για το κριτήριο R^2.

Πίνακας 2.2: *Το κριτήριο R^2 για τα δεδομένα του παραδείγματος κατανάλωσης βενζίνης*

Πλήθος μεταβλητών	Μεταβλητές	Κριτήριο R^2
1	X_1	0,845
	X_2	0,731
	X_3	0,781
	X_4	**0,879**
	X_5	0,216
2	X_1X_2	0,845
	X_1X_3	0,839
	X_1X_4	**0,922**
	X_1X_5	0,878
	X_2X_3	0,799
	X_2X_4	0,892
	X_2X_5	0,818
	X_3X_4	0,886
	X_3X_5	0,797
	X_4X_5	0,879
3	$X_1X_2X_3$	0,924
	$X_1X_2X_4$	0,925
	$X_1X_2X_5$	0,894
	$X_1X_3X_4$	0,924
	$X_1X_3X_5$	**0,947**
	$X_1X_4X_5$	0,938
	$X_2X_3X_4$	0,892
	$X_2X_3X_5$	0,860
	$X_3X_4X_5$	0,888
	$X_2X_4X_5$	0,907
4	$X_1X_2X_3X_4$	0,938
	$X_1X_2X_3X_5$	0,950
	$X_2X_3X_3X_4$	0,908
	$X_1X_3X_4X_5$	**0,953**
	$X_1X_2X_4X_5$	0,938
5	$X_1X_2X_3X_4X_5$	**0,955**

Παρατηρούμε πως στο υποσύνολο το οποίο αποτελείται από τα γραμμικά μοντέλα που περιέχουν μόνο μία ανεξάρτητη μεταβλητή, το κριτήριο R^2 μεγιστοποιείται με τιμή 0,879 για το μοντέλο

$$Y = \beta_0 + \beta_1 X_4 + \varepsilon .$$

Στο υποσύνολο που αποτελείται από γραμμικά μοντέλα που περιέχουν δύο ανεξάρτητες μεταβλητές, το κριτήριο R^2 μεγιστοποιείται με τιμή 0,922 για το μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_4 + \varepsilon .$$

Για το υποσύνολο που αποτελείται από γραμμικά μοντέλα που περιέχουν τρεις ανεξάρτητες μεταβλητές, το κριτήριο R^2 μεγιστοποιείται με τιμή 0,947 για το μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$$

Το κριτήριο R^2 μεγιστοποιείται με τιμή 0,953 στο υποσύνολο των γραμμικών μοντέλων με τέσσερις ανεξάρτητες μεταβλητές για το μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon$$

και τέλος η τιμή του R^2 για το πλήρες μοντέλο είναι ίση με 0,955.

Στον παρακάτω πίνακα παρουσιάζονται οι μέγιστες τιμές του κριτηρίου R^2 για κάθε διαφορετικό πλήθος μεταβλητών καθώς και η μεταβολή στη τιμή του κάθε φορά που προστίθεται μια νέα μεταβλητή.

Πλήθος μεταβλητών	Μεταβλητές	R^2	Μεταβολή
1	X_4	0,879	
2	X_1X_4	0,922	0,043
3	$X_1X_3X_5$	0,947	0,025
4	$X_1X_3X_4X_5$	0,953	0,006
5	$X_1X_2X_3X_4X_5$	0,955	0,002

Για την επιλογή του βέλτιστου μοντέλου διαλέγουμε εκείνο για το οποίο η τιμή του κριτηρίου R^2 δεν αυξάνεται σημαντικά με την προσθήκη νέας μεταβλητής στο μοντέλο. Για το λόγο αυτό ως βέλτιστο γραμμικό μοντέλο θα μπορούσαμε να θεωρήσουμε το μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon \ .$$

2.4 Το κριτήριο *SSE*

Σύμφωνα με το κριτήριο *SSE (error sum of squares),* βέλτιστο μοντέλο θα έπρεπε να θεωρείται το μοντέλο που δίνει τη μικρότερη τιμή για το *SSE*. Επειδή όμως το *SSE* μειώνεται συνεχώς, όσο αυξάνει το πλήθος των ανεξάρτητων μεταβλητών που υπάρχουν στο γραμμικό μοντέλο (αυτό ισχύει γιατί η τιμή του κριτηρίου *SSE* μεταβάλλεται αντίστροφα ως προς τη μονοτονία από την αντίστοιχη τιμή του κριτηρίου R^2, αφού τα δύο κριτήρια συνδέονται με την σχέση

$$R^2 = 1 - \frac{SSE}{SSTO}$$

όπου η ποσότητα *SSTO* παραμένει σταθερή για όλα τα γραμμικά μοντέλα) θα προέκυπτε και πάλι ως βέλτιστο το πλήρες μοντέλο.

Για τον παραπάνω λόγο θα καταφύγουμε στη διαδικασία που ακολουθήσαμε και στην επιλογή του βέλτιστου μοντέλου με το κριτήριο R^2, δηλαδή την εξέταση του κριτηρίου *SSE* σε κάθε ένα υποσύνολο που δημιουργείται ανάλογα με το πλήθος των μεταβλητών δηλαδή ανά μια, δύο, τρεις κτλ ανεξάρτητες μεταβλητές και τέλος στο πλήρες μοντέλο. Σε κάθε ένα υποσύνολο θα αναζητούμε τη μικρότερη τιμή για το κριτήριο *SSE*. Στη συνέχεια θα συγκρίνουμε τις τιμές του κριτηρίου για κάθε υποσύνολο, με στόχο να βρούμε εκείνο το γραμμικό μοντέλο(βέλτιστο) που δίνει μικρή τιμή για το *SSE* αλλά παράλληλα διατηρεί και ένα μικρό αριθμό ανεξάρτητων μεταβλητών. Πιο συγκεκριμένα αν η προσθήκη μιας μεταβλητής στο γραμμικό μοντέλο δεν επιφέρει σημαντική μείωση στη τιμή του *SSE* θα προτιμήσουμε το μοντέλο με τις λιγότερες ανεξάρτητες μεταβλητές.

Εφαρμόζοντας το κριτήριο σε όλα τα γραμμικά μοντέλα που μπορούν να προκύψουν για τα δεδομένα του παραδείγματος της Παραγράφου 2.2 παίρνουμε τις επόμενες για το κριτήριο *SSE* τις τιμές που παρουσιάζονται στον Πίνακα 2.3. Για κάθε συγκεκριμένο πλήθος ανεξάρτητων μεταβλητών έχει σημειωθεί με έντονα γράμματα η παρατηρηθείσα μεγαλύτερη τιμή για το κριτήριο *SSE*.

Πίνακας 2.3: *Το κριτήριο SSE για τα δεδομένα του παραδείγματος κατανάλωσης βενζίνης*

Πλήθος μεταβλητών	Μοντέλο	Κριτήριο *SSE*
1	X_1	136,412
	X_2	237,012
	X_3	193,103
	X_4	**106,476**
	X_5	691,789
2	X_1X_2	136,39
	X_1X_3	94,63
	X_1X_4	**68,73**
	X_1X_5	107,985
	X_2X_3	177,49
	X_2X_4	95,558
	X_2X_5	160,518
	X_3X_4	100,3
	X_3X_5	179,03
	X_4X_5	106,471
3	$X_1X_2X_3$	67,427
	$X_1X_2X_4$	65,945
	$X_1X_2X_5$	93,435
	$X_1X_3X_4$	66,911
	$X_1X_3X_5$	**46,327**
	$X_1X_4X_5$	55.049
	$X_2X_3X_4$	95,156
	$X_2X_3X_5$	123,258
	$X_3X_4X_5$	98,356
	$X_2X_4X_5$	82,023
4	$X_1X_2X_3X_4$	54,384
	$X_1X_2X_3X_5$	43,817
	$X_2X_3X_4X_5$	80,888
	$X_1X_3X_4X_5$	**41,73**
	$X_1X_2X_4X_5$	54,254
5	$X_1X_2X_3X_4X_5$	**39,97**

Παρατηρούμε πως στο υποσύνολο το οποίο αποτελείται από τα γραμμικά μοντέλα που περιέχουν μόνο μία ανεξάρτητη μεταβλητή, το κριτήριο *SSE* ελαχιστοποιείται με τιμή 106,476 για το μοντέλο

$$Y = \beta_0 + \beta_1 X_4 + \varepsilon$$

Στο υποσύνολο που αποτελείται από γραμμικά μοντέλα που περιέχουν δύο ανεξάρτητες μεταβλητές, το κριτήριο *SSE* ελαχιστοποιείται με τιμή 68,730 για το μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_4 + \varepsilon$$

Για το υποσύνολο που αποτελείται από γραμμικά μοντέλα που περιέχουν τρεις ανεξάρτητες μεταβλητές, το κριτήριο SSE ελαχιστοποιείται με τιμή 46,327 για το μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon \ .$$

Το κριτήριο *SSE* ελαχιστοποιείται με τιμή 41,730 στο υποσύνολο των γραμμικών μοντέλων με τέσσερις ανεξάρτητες μεταβλητές για το μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon$$

και τέλος η τιμή του SSE για το πλήρες μοντέλο είναι 39,970.

Στο παρακάτω πίνακα παρουσιάζονται οι ελάχιστες τιμές του κριτηρίου *SSE* για κάθε συγκεκριμένο πλήθος μεταβλητών καθώς και η μεταβολή στη τιμή του κάθε φορά που προστίθεται μια νέα μεταβλητή.

Πλήθος μεταβλητών	Μοντέλο	*SSE*	Μεταβολή
1	X_4	106,476	
2	X_1X_4	68,730	37,746
3	$X_1X_3X_5$	46,327	22,403
4	$X_1X_3X_4X_5$	41,730	4,597
5	$X_1X_2X_3X_4X_5$	39,970	1,76

Για την επιλογή του βέλτιστου μοντέλου διαλέγουμε εκείνο για το οποίο η τιμή του κριτηρίου SSE δεν μειώνεται σημαντικά με την προσθήκη νέας μεταβλητής στο μοντέλο. Για το λόγο αυτό ως βέλτιστο γραμμικό μοντέλο θα μπορούσαμε και πάλι να θεωρήσουμε το μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon \ .$$

2.5 Το κριτήριο R^2_{adj}

Σύμφωνα με το κριτήριο R^2_{adj}, βέλτιστο μοντέλο θεωρείται το μοντέλο που δίνει τη μέγιστη τιμή στον τροποποιημένο συντελεστή προσδιορισμού R^2_{adj} ο οποίος ορίζεται από τον τύπο

$$R^2_{adj} = 1 - \frac{\frac{MSE}{n-p}}{\frac{SSTO}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO}$$

Το κριτήριο R^2_{adj} συνδέεται με το κριτήριο R^2 μέσω της σχέση:

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSE}{SSTO} == 1 - \frac{n-1}{n-p}(1-R^2)$$

όπου n:πλήθος παρατηρήσεων και p:πλήθος παραμέτρων.

Το κριτήριο R^2_{adj} δε συνδέεται με κάποια σχέση αναλογίας με την προσθήκη των μεταβλητών στο γραμμικό μοντέλο, δηλαδή δε μεταβάλλεται μονότονα όταν προσθέτουμε συνεχώς νέες μεταβλητές στο μοντέλο που χρησιμοποιούμε. Επομένως φαίνεται λογικό να δεχόμαστε ως βέλτιστο μοντέλο εκείνο που δίνει τη μέγιστη τιμή για το κριτήριο R^2_{adj}.

Εφαρμόζουμε το κριτήριο σε όλα τα γραμμικά μοντέλα που μπορούν να προκύψουν για τα δεδομένα του παραδείγματος της Παραγράφου 2.2 και παίρνουμε για το κριτήριο R^2_{adj} τις τιμές που παρουσιάζονται στον Πίνακα 2.4. Για κάθε συγκεκριμένο πλήθος ανεξάρτητων μεταβλητών έχει σημειωθεί με έντονα γράμματα η παρατηρηθείσα μεγαλύτερη τιμή για το κριτήριο R^2_{adj}.

Πίνακας 2.4: *Το κριτήριο R^2_{adj} για τα δεδομένα του παραδείγματος κατανάλωσης βενζίνης*

Πλήθος μεταβλητών	Μοντέλο	Κριτήριο R^2_{adj}
1	X_1	0,84
	X_2	0,722
	X_3	0,773
	X_4	**0,875**
	X_5	0,188
2	X_1X_2	0,834
	X_1X_3	0,885
	X_1X_4	**0,916**
	X_1X_5	0,868
	X_2X_3	0,784
	X_2X_4	0,884
	X_2X_5	0,805
	X_3X_4	0,878
	X_3X_5	0,782
	X_4X_5	0,87
3	$X_1X_2X_3$	0,915
	$X_1X_2X_4$	0,917
	$X_1X_2X_5$	0,882
	$X_1X_3X_4$	0,924
	$X_1X_3X_5$	**0,941**
	$X_1X_4X_5$	0,93
	$X_2X_3X_4$	0,88
	$X_2X_3X_5$	0,844
	$X_3X_4X_5$	0,876
	$X_2X_4X_5$	0,896
4	$X_1X_2X_3X_4$	9,928
	$X_1X_2X_3X_5$	0,942
	$X_2X_3X_4X_5$	0,908
	$X_1X_3X_4X_5$	**0,945**
	$X_1X_2X_4X_5$	0,929
5	$X_1X_2X_3X_4X_5$	**0,945**

Παρατηρούμε ότι η τιμή που μεγιστοποιεί το κριτήριο R^2_{adj} είναι 0,945 και την δίνουν δύο μοντέλα, το πλήρες και το μοντέλο που περιέχει τις μεταβλητές X_1, X_3, X_4, X_5. Θα θεωρήσουμε ως βέλτιστο γραμμικό μοντέλο εκείνο με τις λιγότερες ανεξάρτητες μεταβλητές, εφόσον δίνουν την ίδια τιμή, δηλαδή το μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon .$$

2.6 Το κριτήριο *MSE*

Σύμφωνα με το κριτήριο *MSE (mean square error)*, βέλτιστο θεωρείται το μοντέλο που δίνει την ελάχιστη τιμή του *MSE*. Το κριτήριο *MSE* δίνει πληροφορίες ισοδύναμες με το κριτήριο R^2_{adj} σε κάθε γραμμικό μοντέλο και συνδέεται με το κριτήριο *SSE* με τη σχέση

$$MSE = \frac{SSE}{n-p} .$$

Το *MSE* επηρεάζεται από την προσθήκη νέων ανεξάρτητων μεταβλητών στο μοντέλο για το λόγο αυτό αναζητούμε εκείνο το γραμμικό μοντέλο που δίνει τη μικρότερη τιμή για το κριτήριο *MSE* και ταυτόχρονα διατηρεί ένα μικρό αριθμό ανεξάρτητων μεταβλητών. Η διαδικασία που θα ακολουθήσουμε για την επιλογή του βέλτιστου μοντέλου είναι όμοια με τη διαδικασία που ακολουθήθηκε για τα κριτήρια R^2_{adj}, θα εξετάσουμε δηλαδή το κριτήριο *MSE* σε κάθε ένα υποσύνολο γραμμικών μοντέλων που δημιουργείται ανάλογα με το πλήθος των ανεξάρτητων μεταβλητών.

Εφαρμόζοντας το κριτήριο *MSE* σε όλα τα γραμμικά μοντέλα που μπορούν να προκύψουν για τα δεδομένα του παραδείγματος της Παραγράφου 2.2 παίρνουμε τις επόμενες για το κριτήριο τις τιμές που παρουσιάζονται στον Πίνακα 2.5. Για κάθε συγκεκριμένο πλήθος ανεξάρτητων μεταβλητών έχει σημειωθεί με έντονα γράμματα η παρατηρηθείσα μικρότερη τιμή για το κριτήριο *MSE*.

Πίνακας 2.5: *Το κριτήριο MSE για τα δεδομένα του παραδείγματος κατανάλωσης βενζίνης*

Πλήθος μεταβλητών	Μοντέλο	Κριτήριο *MSE*
1	X_1	4,872
	X_2	8,465
	X_3	6,897
	X_4	**3,803**
	X_5	24,707
2	X_1X_2	5,051
	X_1X_3	3,505
	X_1X_4	**2,546**
	X_1X_5	3,999
	X_2X_3	6,574
	X_2X_4	3,539
	X_2X_5	5,945
	X_3X_4	3,715
	X_3X_5	6,631
	X_4X_5	3,943
3	$X_1X_2X_3$	2,593
	$X_1X_2X_4$	2,536
	$X_1X_2X_5$	3,594
	$X_1X_3X_4$	2,574
	$X_1X_3X_5$	**1,785**
	$X_1X_4X_5$	2,117
	$X_2X_3X_4$	3,66
	$X_2X_3X_5$	4,741
	$X_3X_4X_5$	3,155
	$X_2X_4X_5$	3,155
4	$X_1X_2X_3X_4$	2,175
	$X_1X_2X_3X_5$	1,753
	$X_2X_3X_4X_5$	3,236
	$X_1X_3X_4X_5$	**1,669**
	$X_1X_2X_4X_5$	2,17
5	$X_1X_2X_3X_4X_5$	**1,665**

Παρατηρούμε πως στο υποσύνολο το οποίο αποτελείται από τα γραμμικά μοντέλα που περιέχουν μόνο μία ανεξάρτητη μεταβλητή, το κριτήριο *MSE* ελαχιστοποιείται με τιμή 3,803 για το μοντέλο

$$Y = \beta_0 + \beta_1 X_4 + \varepsilon .$$

Στο υποσύνολο που αποτελείται από γραμμικά μοντέλα που περιέχουν δύο ανεξάρτητες μεταβλητές, το κριτήριο *MSE* ελαχιστοποιείται με τιμή 2,546 για το μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_4 + \varepsilon$$

Για το υποσύνολο που αποτελείται από γραμμικά μοντέλα που περιέχουν τρεις ανεξάρτητες μεταβλητές, το κριτήριο *MSE* ελαχιστοποιείται με τιμή 1,785 για το μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon .$$

Το κριτήριο *MSE* ελαχιστοποιείται με τιμή 1,669 στο υποσύνολο των γραμμικών μοντέλων με τέσσερις ανεξάρτητες μεταβλητές για το μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon$$

και τέλος η τιμή του *MSE* για το πλήρες μοντέλο είναι 1,665.

Στο παρακάτω πίνακα παρουσιάζονται οι ελάχιστες τιμές του κριτηρίου *MSE* για κάθε ένα υποσύνολο.

Πλήθος μεταβλητών	Μοντέλο	Κριτήριο *MSE*
1	X_4	3,803
2	X_1X_4	2,546
3	$X_1X_3X_5$	1,785
4	$X_1X_3X_4X_5$	1,669
5	$X_1X_2X_3X_4X_5$	1,665

Για την επιλογή του βέλτιστου μοντέλου διαλέγουμε εκείνο για το οποίο η τιμή του κριτηρίου *MSE* γίνεται ελάχιστη. Για το λόγο αυτό ως βέλτιστο γραμμικό μοντέλο θα θεωρήσουμε το πλήρες μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon \ .$$

Σημειώνεται ότι, λόγω της πολύ μικρής διαφοράς ανάμεσα στις τιμές του κριτηρίου όταν έχουμε 5 μεταβλητές και όταν έχουμε τις (καλύτερες) 4 ή ακόμη και 3 μεταβλητές, αν θα θέλαμε να κάνουμε οικονομία στις μετρήσεις (ή στις πράξεις που απαιτούνται για τους υπολογισμούς) θα μπορούσαμε επίσης να προτείνουμε τη χρήση του μοντέλου

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon \ .$$

2.7 Το κριτήριο C_p του Mallows

Ένα άλλο κριτήριο επιλογής βέλτιστου μοντέλου είναι το C_p, το οποίο αρχικά προτάθηκε από τον C. L. Mallows (1973) και έχει τη μορφή

$$C_p = \frac{SSE_p}{MSE(X_1,\ldots,X_{p-1})} - (n-2p)$$

όπου n το πλήθος παρατηρήσεων και p το πλήθος παραμέτρων του μοντέλου συμπεριλαμβανομένου του β_0. Όπως έδειξε ο R. W. Kennard (1971), το κριτήριο C_p συνδέεται άμεσα με τα κριτήρια R^2_{adj} και R^2.

Αν ένα γραμμικό μοντέλο με p παραμέτρους είναι επαρκές, δηλαδή δεν παρουσιάζει έλλειψη προσαρμογής, τότε

$$E(SSE) = (n-p)\sigma^2 \ .$$

Υποθέτοντας ότι $E(MSE) = \sigma^2$, κατά προσέγγιση ισχύει, και ότι ο λόγος SSE / MSE έχει αναμενόμενη τιμή

$$\frac{(n-p)\sigma^2}{\sigma^2} = n - p\,,$$

οπότε κατά προσέγγιση θα ισχύει

$$E(C_p) = p$$

για ένα επαρκές μοντέλο.

Επομένως, τα «βέλτιστα» γραμμικά μοντέλα θα παρουσιάζονται ως σημεία, στο διάγραμμα του C_p ως προς p, που βρίσκονται κοντά στη διαγώνια γραμμή

$$C_p = p\,.$$

Τα γραμμικά μοντέλα που θα παρουσιάζουν έλλειψη προσαρμογής, δηλαδή μεροληπτικά γραμμικά μοντέλα, θα δίνουν σημεία πάνω από τη γραμμή $C_p = p$.

Η τιμή του κριτηρίου C_p αυξάνεται καθώς εισάγεται μια νέα ανεξάρτητη μεταβλητή στο γραμμικό μοντέλο. Χρησιμοποιώντας το κριτήριο C_p επιδιώκουμε να προσδιορίσουμε εκείνα τα γραμμικά μοντέλα για τα οποία η τιμή του κριτηρίου είναι μικρή και είναι κοντά στην τιμή του p. Τα σύνολα με μικρή τιμή του C_p έχουν μικρή τιμή και για το μέσο τετραγωνικό σφάλμα και όταν η τιμή του C_p είναι κοντά στο p τότε η μεροληψία για το μοντέλο είναι μικρή.

Μπορεί κάποιες φορές να καταλήξουμε σε γραμμικό μοντέλο με μικρή τιμή του C_p ενώ παράλληλα να εμφανίζεται και σημαντική μεροληψία. Σε αυτή την περίπτωση θα προτιμήσουμε ένα γραμμικό μοντέλο που δίνει μεγαλύτερη τιμή στο κριτήριο C_p αντί ενός που έχει μεροληψία.

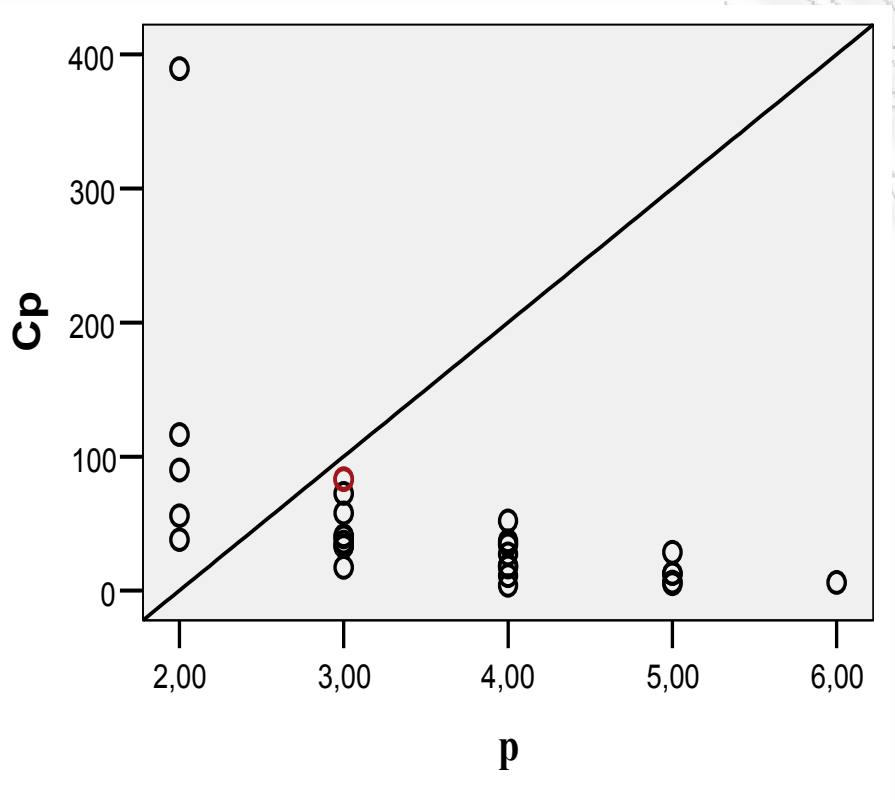
Εφαρμόζουμε το κριτήριο σε όλα τα γραμμικά μοντέλα που μπορούν να προκύψουν για τα δεδομένα του παραδείγματος της Παραγράφου 2.2 και παίρνουμε για το κριτήριο C_p τις τιμές που παρουσιάζονται στον Πίνακα 2.6. Για κάθε συγκεκριμένο πλήθος ανεξάρτητων μεταβλητών έχει σημειωθεί με έντονα γράμματα η παρατηρηθείσα καλύτερη τιμή για το κριτήριο δηλαδή η τιμή που είναι κοντά στην τιμή του p

Πίνακας 2.6: *Το κριτήριο C_p για τα δεδομένα του παραδείγματος κατανάλωσης βενζίνης*

Πλήθος μεταβλητών	p	Μοντέλο	Κριτήριο C_p
1	2	X_1	55,93
		X_2	116,35
		X_3	89,98
		X_4	**37,95**
		X_5	389,49
2	3	X_1X_2	57,92
		X_1X_3	32,83
		X_1X_4	**17,28**
		X_1X_5	40,86
		X_2X_3	82,6
		X_2X_4	33,39
		X_2X_5	72,41
		X_3X_4	36,24
		X_3X_5	83,53
		X_4X_5	39,95
3	4	$X_1X_2X_3$	18,5
		$X_1X_2X_4$	17,61
		$X_1X_2X_5$	34,12
		$X_1X_3X_4$	18,19
		$X_1X_3X_5$	**3,82**
		$X_1X_4X_5$	11,06
		$X_2X_3X_4$	35,15
		$X_2X_3X_5$	52,03
		$X_3X_4X_5$	37,07
		$X_2X_4X_5$	27,26
4	5	$X_1X_2X_3X_4$	12,66
		$X_1X_2X_3X_5$	6,32
		$X_2X_3X_4X_5$	28,58
		$X_1X_3X_4X_5$	**5,06**
		$X_1X_2X_4X_5$	12,58
5	6	$X_1X_2X_3X_4X_5$	**6,01**

Για να προσδιορίσουμε γραφικά το βέλτιστο μοντέλο θα μπορούσαμε επίσης να μελετήσουμε και το επόμενο διάγραμμα του C_p έναντι του p.

Σχήμα 2.1 *Διάγραμμα των τιμών* C_p *ως προς* p *παραδείγματος κατανάλωσης βενζίνης*



Με βάση το παραπάνω πίνακα τιμών του κριτηρίου C_p και σε συνδυασμό με το διάγραμμα του C_p ως προς p ως βέλτιστο μοντέλο θα θεωρήσουμε το μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$$

εφόσον δίνει μικρή τιμή για το C_p και ταυτόχρονα η τιμή του είναι σχεδόν ίση με το πλήθος των παραμέτρων του μοντέλου, δηλαδή

$$C_p \approx p \Rightarrow 3{,}82 \approx 4 .$$

Κάτι τέτοιο θα μπορούσαμε να παρατηρήσουμε και από το Σχήμα 2.1 εφόσον το σημείο που αντιστοιχεί στο μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$$

βρίσκεται πιο κοντά στη διαγώνιο γραμμή

$$C_p = p .$$

2.8 Το κριτήριο $PRESS_p$

Το κριτήριο $PRESS_p$ *(prediction sum of squares)* προτάθηκε από τον Allen (1971) και χρησιμοποιείται με σκοπό να ελεγχθεί η καταλληλότητα των ανεξάρτητων μεταβλητών στο να προβλέψουν την τιμή της εξαρτημένης μεταβλητής.

Η διαδικασία εφαρμογής του κριτηρίου ξεκινάει με τη διαγραφή του πρώτου συνόλου παρατηρήσεων για την εξηρτημένη και τις ανεξάρτητες μεταβλητές και προσαρμόζοντας όλα τα δυνατά γραμμικά μοντέλα στις υπόλοιπες παρατηρήσεις. Στη συνέχεια χρησιμοποιούμε το κάθε γραμμικό μοντέλο για να προβλέψουμε το Y_1 από το $\hat{Y}_{1p}$ βρίσκοντας έτσι ένα εκτιμώμενο σφάλμα $Y_1 - \hat{Y}_{1p}$ για όλα τα δυνατά γραμμικά μοντέλα.

Την διαδικασία αυτή την επαναλαμβάνουμε διαγράφοντας τώρα το δεύτερο σύνολο παρατηρήσεων για την εξηρτημένη και τις ανεξάρτητες μεταβλητές ώστε να πάρουμε τις τιμές του εκτιμώμενου σφάλματος $Y_2 - \hat{Y}_{2p}$, διαγράφουμε την τρίτη παρατήρηση για να πάρουμε τις τιμές του εκτιμώμενου σφάλματος $Y_3 - \hat{Y}_{3p}$ και συνεχίζουμε έως ότου να πάρουμε τόσες διαγραφές όσες και οι παρατηρήσεις μας.

Εφόσον πάρουμε όλες τις τιμές για τα εκτιμώμενα σφάλματα υπολογίζουμε για κάθε γραμμικό μοντέλο το άθροισμα τετραγώνων των εκτιμώμενων σφαλμάτων

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_{ip})^2 .$$

Αφού τελειώσει η διαδικασία εφαρμογής του κριτηρίου $PRESS_p$ θα επιλέγουμε ως βέλτιστο γραμμικό μοντέλο εκείνο που δίνει την μικρότερη τιμή για το άθροισμα τετραγώνων των εκτιμώμενων σφαλμάτων αλλά ταυτόχρονα να μην περιλαμβάνει πολλές μεταβλητές.

Εφαρμόζουμε το κριτήριο σε όλα τα γραμμικά μοντέλα που μπορούν να προκύψουν για τα δεδομένα του παραδείγματος της Παραγράφου 2.2 παίρνουμε για το κριτήριο $PRESS_p$ τις τιμές που παρουσιάζονται στον Πίνακα 2.7. Για κάθε συγκεκριμένο πλήθος ανεξάρτητων μεταβλητών έχει σημειωθεί με έντονα γράμματα η παρατηρηθείσα μικρότερη τιμή για το κριτήριο.

Πίνακας 2.7: *Το κριτήριο $PRESS_p$ για τα δεδομένα του παραδείγματος κατανάλωσης βενζίνης*

Πλήθος μεταβλητών	p	Μοντέλο	Κριτήριο $PRESS_p$
1	2	X_1	136,39
		X_2	237,05
		X_3	139,08
		X_4	**106,49**
		X_5	691,81
2	3	X_1X_2	136,39
		X_1X_3	94,61
		X_1X_4	**68,73**
		X_1X_5	107,96
		X_2X_3	177,44
		X_2X_4	95,53
		X_2X_5	160,56
		X_3X_4	100,32
		X_3X_5	179,03
		X_4X_5	166,46
3	4	$X_1X_2X_3$	67,39
		$X_1X_2X_4$	65,93
		$X_1X_2X_5$	93,42
		$X_1X_3X_4$	66,91
		$X_1X_3X_5$	**46,36**
		$X_1X_4X_5$	55,05
		$X_2X_3X_4$	95,14
		$X_2X_3X_5$	123,25
		$X_3X_4X_5$	98,34
		$X_2X_4X_5$	82,04
4	5	$X_1X_2X_3X_4$	54,37
		$X_1X_2X_3X_5$	43,85
		$X_2X_3X_4X_5$	80,86
		$X_1X_3X_4X_5$	**41,72**
		$X_1X_2X_4X_5$	54,25
5	6	$X_1X_2X_3X_4X_5$	**39,95**

Παρατηρούμε πως στο υποσύνολο το οποίο αποτελείται από τα γραμμικά μοντέλα που περιέχουν μόνο μία ανεξάρτητη μεταβλητή, το κριτήριο $PRESS_p$ ελαχιστοποιείται με τιμή 106,49 για το μοντέλο

$$Y = \beta_0 + \beta_1 X_4 + \varepsilon .$$

Στο υποσύνολο που αποτελείται από γραμμικά μοντέλα που περιέχουν δύο ανεξάρτητες μεταβλητές, το κριτήριο $PRESS_p$ ελαχιστοποιείται με τιμή 68,73 για το μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_4 + \varepsilon .$$

Για το υποσύνολο που αποτελείται από γραμμικά μοντέλα που περιέχουν τρεις ανεξάρτητες μεταβλητές, το κριτήριο $PRESS_p$ ελαχιστοποιείται με τιμή 46,36 για το μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon .$$

Το κριτήριο $PRESS_p$ ελαχιστοποιείται με τιμή 41,72 στο υποσύνολο των γραμμικών μοντέλων με τέσσερις ανεξάρτητες μεταβλητές για το μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon$$

και τέλος η τιμή του $PRESS_p$ για το πλήρες μοντέλο είναι 39,95.

Στο παρακάτω πίνακα παρουσιάζονται οι ελάχιστες τιμές του κριτηρίου $PRESS_p$ για κάθε πλήθος ανεξάρτητων μεταβλητών καθώς και η μεταβολή στη τιμή του κάθε φορά που προστίθεται μια νέα μεταβλητή.

Πλήθος μεταβλητών	Μοντέλο	$PRESS_p$	Μεταβολή
1	X_4	106,49	
2	X_1X_4	68,73	37,76
3	$X_1X_3X_5$	46,36	22,37
4	$X_1X_3X_4X_5$	41,72	4,64
5	$X_1X_2X_3X_4X_5$	39,95	1,77

Για την επιλογή του βέλτιστου μοντέλου διαλέγουμε εκείνο για το οποίο η τιμή του κριτηρίου $PRESS_p$ δε μειώνεται σημαντικά με την προσθήκη νέας μεταβλητής στο μοντέλο. Για το λόγο αυτό ως βέλτιστο γραμμικό μοντέλο θα θεωρήσουμε το μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon \ .$$

2.9 Το κριτήριο *AIC*

Το κριτήριο *AIC (Akaike's information criterion)* αναπτύχθηκε από τον Akaike (1971) και καθιερώθηκε με την ονομασία Akaike's information criterion. Αποτελεί ένα ακόμη τρόπο σύμφωνα με τον οποίο μπορούμε να ελέγξουμε την καταλληλότητα των ανεξάρτητων μεταβλητών στο γραμμικό μοντέλο. Για την εύρεση του βέλτιστου γραμμικού μοντέλου επιλέγουμε εκείνο το μοντέλο που ελαχιστοποιεί το *AIC*.

Η τιμή του κριτηρίου *AIC* υπολογίζεται από το τύπο

$$AIC = n \ln(\frac{SSE}{n}) + 2p$$

όπου p ο αριθμός των παραμέτρων του γραμμικού μοντέλο. Με βάση το τύπο του κριτηρίου *AIC* παρατηρούμε πως όταν αυξάνεται το πλήθος των p παραμέτρων τότε η τιμή του *SSE* μειώνεται ενώ παράλληλα αυξάνεται ο τελευταίος προσθετέος του κριτηρίου. Πρακτικά έχει παρατηρηθεί ότι η τιμή του *AIC* μειώνεται με την εισαγωγή παραμέτρων στο μοντέλο έως κάποια παράμετρο που η εισαγωγή της θα οδηγήσει σε αύξηση του *AIC*.

Εφαρμόζουμε στη συνέχεια το κριτήριο σε όλα τα γραμμικά μοντέλα που μπορούν να προκύψουν για τα δεδομένα του παραδείγματος της Παραγράφου 2.2 και παίρνουμε για το κριτήριο *AIC* τις τιμές που παρουσιάζονται στον Πίνακα 2.8. Για κάθε συγκεκριμένο πλήθος ανεξάρτητων μεταβλητών έχει σημειωθεί με έντονα γράμματα η παρατηρηθείσα μικρότερη τιμή για το κριτήριο.

Πίνακας 2.8: *Το κριτήριο AIC για τα δεδομένα του παραδείγματος κατανάλωσης βενζίνης*

Πλήθος μεταβλητών	Μοντέλο	Κριτήριο *AIC*
1	X_1	49,43
	X_2	66,01
	X_3	59,86
	X_4	**42**
	X_5	100,14
2	X_1X_2	51,43
	X_1X_3	40,46
	X_1X_4	**30,87**
	X_1X_5	44,42
	X_2X_3	59,33
	X_2X_4	40,76
	X_2X_5	56,32
	X_3X_4	42,21
	X_3X_5	59,59
	X_4X_5	44
3	$X_1X_2X_3$	32,3
	$X_1X_2X_4$	31,36
	$X_1X_2X_5$	42,08
	$X_1X_3X_4$	32,06
	$X_1X_3X_5$	**21,04**
	$X_1X_4X_5$	233,44
	$X_2X_3X_4$	42,63
	$X_2X_3X_5$	50,39
	$X_3X_4X_5$	43,62
	$X_2X_4X_5$	38,17
4	$X_1X_2X_3X_4$	27,85
	$X_1X_2X_3X_5$	21,36
	$X_2X_3X_4X_5$	39,76
	$X_1X_3X_4X_5$	**19,9**
	$X_1X_2X_4X_5$	27,77
5	$X_1X_2X_3X_4X_5$	**20,61**

Στο παρακάτω πίνακα παρουσιάζονται οι ελάχιστες τιμές του κριτηρίου *AIC* για κάθε πλήθος ανεξάρτητων μεταβλητών καθώς και η μεταβολή στη τιμή του κάθε φορά που προστίθεται μια νέα μεταβλητή.

Πλήθος μεταβλητών	Μοντέλο	*AIC*
1	X_4	42
2	X_1X_4	30,87
3	$X_1X_3X_5$	21,04
4	$X_1X_3X_4X_5$	19,9
5	$X_1X_2X_3X_4X_5$	20,61

Παρατηρούμε ότι η τιμή που ελαχιστοποιεί το κριτήριο *AIC* είναι 19,9 και την δίνει το γραμμικό μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon$$

το οποίο και θα θεωρηθεί ως βέλτιστο σύμφωνα με την εφαρμογή αυτού του κριτηρίου.

2.10 Το κριτήριο *BIC*

Το κριτήριο *BIC (Bayesian information criterion)*ανατήχθηκε από τον Schwarz (1978). Το *BIC* χρησιμοποιείται συνήθως για τον έλεγχο της καταλληλότητας των ανεξάρτητων μεταβλητών σε ένα γραμμικό μοντέλο και την εύρεση του βέλτιστου μοντέλου όταν αυτό αποτελείται από πολλές παραμέτρους, σε αντίθεση με το κριτήριο *AIC* που περιορίζεται στην εύρεση του βέλτιστου μοντέλου όταν ο αριθμός των παραμέτρων είναι μικρός.

Η τιμή του κριτηρίου *BIC* υπολογίζεται από το τύπο

$$BIC = n\ln(\frac{SSE}{n}) + 2(p+2)q - 2q^2$$

όπου p ο αριθμός των παραμέτρων,

$$q = \frac{\hat{\sigma}^2}{\frac{SSE}{n}}$$

και $\hat{\sigma}^2$ είναι η εκτίμηση της διασποράς όταν το μοντέλο περιέχει όλες τις ανεξάρτητες μεταβλητές.

Για την επιλογή του βέλτιστου μοντέλου σύμφωνα με το κριτήριο αυτό επιλέγουμε το μοντέλο που ελαχιστοποιεί τν τιμή του *BIC*. Τα κριτήρια *AIC* και *BIC* χρησιμοποιούνται σχεδόν με τον ίδιο τρόπο για την επιλογή του βέλτιστου μοντέλου με τη μόνη διαφορά, η οποία μας προτρέπει στη χρήση του *BIC*, ότι το *AIC* δεν αποδείχθηκε ποτέ συνεπές αφού είχε την τάση να υπερεκτιμάει τα γραμμικά μοντέλο όσον αφορά την τελική επιλογή.

Εφαρμόζουμε το κριτήριο σε όλα τα γραμμικά μοντέλα που μπορούν να προκύψουν για τα δεδομένα του παραδείγματος της Παραγράφου 2.2 και παίρνουμε για το κριτήριο *BIC* τις τιμές που παρουσιάζονται στον Πίνακα 2.9. Για κάθε συγκεκριμένο πλήθος ανεξάρτητων μεταβλητών έχει σημειωθεί με έντονα γράμματα η παρατηρηθείσα μικρότερη τιμή για το κριτήριο *BIC*.

Παρατηρούμε ότι η τιμή που ελαχιστοποιεί το κριτήριο *BIC* είναι 26,64 και την δίνει το γραμμικό μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$$

το οποίο και θα θεωρηθεί ως βέλτιστο σύμφωνα με τον τρόπο εφαρμογής αυτού του κριτηρίου.

Πίνακας 2.9: *Το κριτήριο BIC για τα δεδομένα του παραδείγματος κατανάλωσης βενζίνης*

Πλήθος μεταβλητών	p	Μοντέλο	Κριτήριο *BIC*
1	2	X_1	52,24
		X_2	68,81
		X_3	62,66
		X_4	**44,8**
		X_5	100,94
2	3	X_1X_2	55,63
		X_1X_3	44,67
		X_1X_4	**35,07**
		X_1X_5	48,63
		X_2X_3	63,54
		X_2X_4	44,96
		X_2X_5	60,52
		X_3X_4	46,41
		X_3X_5	63,79
		X_4X_5	48,2
3	4	$X_1X_2X_3$	37,9
		$X_1X_2X_4$	37,23
		$X_1X_2X_5$	47,69
		$X_1X_3X_4$	37,67
		$X_1X_3X_5$	**26,64**
		$X_1X_4X_5$	239,05
		$X_2X_3X_4$	48,23
		$X_2X_3X_5$	56
		$X_3X_4X_5$	49,23
		$X_2X_4X_5$	43,78
4	5	$X_1X_2X_3X_4$	34,85
		$X_1X_2X_3X_5$	28,37
		$X_2X_3X_4X_5$	46,76
		$X_1X_3X_4X_5$	**26,91**
		$X_1X_2X_4X_5$	34,78
5	6	$X_1X_2X_3X_4X_5$	**29,02**

2.11 Το κριτήριο S_p

Η διαδικασία επιλογής βέλτιστου μοντέλου με τη χρήση του κριτηρίου S_p βασίζεται στην ελαχιστοποίηση του αναμενόμενου μέσου τετραγωνικού σφάλματος των προβλέψεων (mean square error of prediction) ο οποίος συμβολίζεται με *MSEP* και δίνεται από τον τύπο

$$MSEP(\hat{Y}_p) = \sum_{Y} (Y - \hat{Y}_p)^2$$

όπου Y και $\hat{Y}$ είναι η παρατηρηθείσα και η προβλεπόμενη τιμή για την εξαρτημένη μεταβλητή.

Αποδεικνύεται ότι το *MSEP* δίνεται και από τη σχέση

$$MSEP(\hat{Y}) = \frac{\sigma^2}{n}(1 + n + T)$$

όπου n είναι το πλήθος των παρατηρήσεων, σ^2 η διακύμανση των υπολοίπων και

$$T = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)}$$

η γνωστή στατιστική συνάρτηση t του συντελεστή β_i της μεταβλητής X_i ($\hat{\beta}_i$ είναι η εκτιμήτρια ελαχίστων τετραγώνων του συντελεστή β_i της μεταβλητής X_i και $s(\hat{\beta}_i)$ η εκτιμούμενη τυπική απόκλιση της εκτιμήτριας $\hat{\beta}_i$).

Η αναμενόμενη τιμή του *MSEP* για όλα τα γραμμικά μοντέλα υπολογίζεται από μια νέα παράμετρο ε_p που δίνεται από το τύπο

$$\varepsilon_p = \frac{\sigma_p^2}{n}(1 + n + \frac{p(n+1)}{n-p-2})$$

και εκτιμάται από την παράμετρο E_p που υπολογίζεται από τη σχέση

$$E_p = \frac{SSE_p}{n(n-p)}(1 + n + \frac{p(n+1)}{n-p-2}),$$

όπου SSE_p είναι το άθροισμα των τετραγωνικών σφαλμάτων. Η διαδικασία ελαχιστοποίησης του E_p, για μερικά τυχαία πειράματα και για δεδομένη τιμή του n, ανάγεται τελικά στη ελαχιστοποίηση του S_p που δίνεται από τον τύπο

Πίνακας 2.10: *Το κριτήριο S_p για τα δεδομένα του παραδείγματος κατανάλωσης βενζίνης*

Πλήθος μεταβλητών	p	Μοντέλο	Κριτήριο S_p
1	2	X_1	0,187
		X_2	0,326
		X_3	0,265
		X_4	**0,146**
		X_5	0,95
2	3	X_1X_2	0,202
		X_1X_3	0,14
		X_1X_4	**0,102**
		X_1X_5	0,16
		X_2X_3	0,263
		X_2X_4	0,142
		X_2X_5	0,238
		X_3X_4	0,149
		X_3X_5	0,265
		X_4X_5	0,158
3	4	$X_1X_2X_3$	0,108
		$X_1X_2X_4$	0,106
		$X_1X_2X_5$	0,15
		$X_1X_3X_4$	0,107
		$X_1X_3X_5$	**0,074**
		$X_1X_4X_5$	0,088
		$X_2X_3X_4$	0,152
		$X_2X_3X_5$	0,198
		$X_3X_4X_5$	0,158
		$X_2X_4X_5$	0,131
4	5	$X_1X_2X_3X_4$	0,095
		$X_1X_2X_3X_5$	0,076
		$X_2X_3X_4X_5$	0,141
		$X_1X_3X_4X_5$	**0,073**
		$X_1X_2X_4X_5$	0,094
5	6	$X_1X_2X_3X_4X_5$	**0,076**

$$S_p = \frac{SSE_p}{(n-p)(n-p-2)}.$$

Εφαρμόζουμε το κριτήριο σε όλα τα γραμμικά μοντέλα που μπορούν να προκύψουν για τα δεδομένα του παραδείγματος της Παραγράφου 2.2 και παίρνουμε για το κριτήριο S_p τις τιμές που παρουσιάζονται στον Πίνακα 2.10. Για κάθε συγκεκριμένο πλήθος ανεξάρτητων μεταβλητών έχει σημειωθεί με έντονα γράμματα η παρατηρηθείσα μικρότερη τιμή για το κριτήριο S_p.

Στο παρακάτω πίνακα παρουσιάζονται οι ελάχιστες τιμές του κριτηρίου S_p για κάθε ένα υποσύνολο καθώς και η μεταβολή στη τιμή του κάθε φορά που προστίθεται μια νέα μεταβλητή.

Πίνακας 2.10: Το κριτήριο S_p για τα δεδομένα του παραδείγματος κατανάλωσης βενζίνης

Πλήθος μεταβλητών	Μοντέλο	S_p	Μεταβολή
1	X_4	0,146	
2	X_1X_4	0,102	0,044
3	$X_1X_3X_5$	0,074	0,028
4	$X_1X_3X_4X_5$	0,073	0,001
5	$X_1X_2X_3X_4X_5$	0,076	0,003

Για την επιλογή του βέλτιστου μοντέλου διαλέγουμε εκείνο για το οποίο η τιμή του κριτηρίου S_p δεν μειώνεται σημαντικά με την προσθήκη νέας μεταβλητής στο μοντέλο. Για το λόγο αυτό ως βέλτιστο γραμμικό μοντέλο θα θεωρήσουμε το μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon .$$

2.12 Ανακεφαλαίωση

Σκοπός του Κεφαλαίου 2 είναι να εξεταστούν όλα τα δυνατά γραμμικά μοντέλα και να επιλέξουμε από αυτά το βέλτιστο γραμμικό μοντέλο με την εφαρμογή διαφόρων κριτηρίων.

Από θεωρητική άποψη, η εξέταση όλων των δυνατών γραμμικών μοντέλων είναι καλύτερη αφού δίνει στον ερευνητή τη δυνατότητα να μελετήσει όλες τις δυνατές περιπτώσεις με αποτέλεσμα να είναι σε θέση να επιλέξει με μεγαλύτερη ακρίβεια το καλύτερο υποσύνολο μεταβλητών. Πρακτικά όμως ο υπολογιστικός χρόνος που χρησιμοποιείται για την προσαρμογή όλων των γραμμικών μοντέλων αποτελεί σπατάλη και η απαιτούμενη πλήρης φυσική προσπάθεια για την εξέταση όλων των αποτελεσμάτων από τον υπολογιστή είναι τεράστια όταν εξετάζονται όλες οι μεταβλητές. Είναι συνεπώς προτιμότερο να χρησιμοποιηθεί κάποια διαδικασία επιλογής που θα περιορίσει τη διαδικασία αυτή.

Στις παραγράφους που προηγήθηκαν εφαρμόσαμε τα κριτήρια επιλογής του βέλτιστου μοντέλου, στα δεδομένα του παραδείγματος της Παραγράφου 2.2. Είναι αξιοπρόσεκτο ότι δεν έδωσαν όλα τα κριτήρια ως βέλτιστο γραμμικό μοντέλο το ίδιο μοντέλο. Τα συνολικά αποτελέσματα από την εφαρμογή των κριτηρίων για κάθε γραμμικό μοντέλο παρουσιάζονται στον Πίνακα 2.11

Η εφαρμογή του κριτηρίου R^2 στα δεδομένα του παραδείγματος της Παραγράφου 2.2 μας δίνει ως βέλτιστο μοντέλο το

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$$

Όταν στο μοντέλο έχει εισαχθεί μια μόνο μεταβλητή, η μεγαλύτερη τιμή του κριτηρίου δίνεται από τη μεταβλητή X_4. Η μεταβλητή X_4 παραμένει στα μοντέλα που θεωρούνται «καλά», για κάθε κατηγορία υποσυνόλου μεταβλητών, εκτός από το γραμμικό μοντέλο που αποτελείται από τρεις μεταβλητές και ορίζεται ως βέλτιστο.

Θα μπορούσε επίσης να είχε επιλεγεί το μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_4 + \beta_3 X_5 + \varepsilon$$

με αντίστοιχη τιμή του συντελεστή προσδιορισμού

Πίνακας 2.11: *Τα αποτελέσματα των κριτηρίων για τα δεδομένα του παραδείγματος κατανάλωσης βενζίνης*

Πλήθος Μεταβλητών	Μοντέλο	R^2	SSE	R^2_{adj}	MSE	C_p	$PRESS_p$	AIC	BIC	S_p
1	X_1	0,845	136,412	0,84	4,872	55,93	136,39	49,43	52,24	0,187
	X_2	0,731	237,012	0,722	8,465	116,35	237,05	66,01	68,81	0,326
	X_3	0,781	193,103	0,773	6,897	89,98	139,08	59,86	62,66	0,265
	X_4	**0,879**	**106,48**	0,875	**3,803**	37,95	**106,49**	**42**	**44,8**	**0,146**
	X_5	0,216	691,789	0,188	24,707	389,49	691,81	100,14	100,94	0,95
2	X_1X_2	0,845	136,39	0,834	5,051	57,92	136,39	51,43	55,63	0,202
	X_1X_3	0,839	94,63	0,885	3,505	32,83	94,61	40,46	44,67	0,14
	X_1X_4	**0,922**	**68,73**	0,916	**2,546**	17,28	**68,73**	**30,87**	**35,07**	**0,102**
	X_1X_5	0,878	107,985	0,868	3,999	40,86	107,96	44,42	48,63	0,16
	X_2X_3	0,799	177,49	0,784	6,574	82,6	177,44	59,33	63,54	0,263
	X_2X_4	0,892	95,558	0,884	3,539	33,39	95,53	40,76	44,96	0,142
	X_2X_5	0,818	160,518	0,805	5,945	72,41	160,56	56,32	60,52	0,238
	X_3X_4	0,886	100,3	0,878	3,715	36,24	100,32	42,21	46,41	0,149
	X_3X_5	0,797	179,03	0,782	6,631	83,53	179,03	59,59	63,79	0,265
	X_4X_5	0,879	106,471	0,87	3,943	39,95	166,46	44	48,2	0,158
3	$X_1X_2X_3$	0,924	67,427	0,915	2,593	18,5	67,39	32,3	37,9	0,108
	$X_1X_2X_4$	0,925	65,945	0,917	2,536	17,61	65,93	31,36	37,23	0,106
	$X_1X_2X_5$	0,894	93,435	0,882	3,594	34,12	93,42	42,08	47,69	0,15
	$X_1X_3X_4$	0,924	66,911	0,924	2,574	18,19	66,91	32,06	37,67	0,107
	$X_1X_3X_5$	**0,947**	**46,327**	0,941	**1,785**	3,82	**46,36**	**21,04**	**26,64**	**0,074**
	$X_1X_4X_5$	0,938	55.049	0,93	2,117	11,06	55,05	233,44	239,05	0,088
	$X_2X_3X_4$	0,892	95,156	0,88	3,66	35,15	95,14	42,63	48,23	0,152
	$X_2X_3X_5$	0,860	123,258	0,844	4,741	52,03	123,25	50,39	56	0,198
	$X_3X_4X_5$	0,888	98,356	0,876	3,155	37,07	98,34	43,62	49,23	0,158
	$X_2X_4X_5$	0,907	82,023	0,896	3,155	27,26	82,04	38,17	43,78	0,131
4	$X_1X_2X_3X_4$	0,938	54,384	9,928	2,175	12,66	54,37	27,85	34,85	0,095
	$X_1X_2X_3X_5$	0,950	43,817	0,942	1,753	6,32	43,85	21,36	28,37	0,076
	$X_2X_3X_4X_5$	0,908	80,888	0,908	3,236	28,58	80,86	39,76	46,76	0,141
	$X_1X_3X_4X_5$	**0,953**	**41,73**	**0,945**	**1,669**	5,06	**41,72**	**19,9**	26,91	**0,073**
	$X_1X_2X_4X_5$	0,938	54,254	0,929	2,17	12,58	54,25	27,77	34,78	0,094
5	$X_1X_2X_3X_4X_5$	**0,955**	**39,97**	0,945	**1,665**	6,01	**39,95**	**20,61**	**29,02**	**0,076**

$$R^2 = 0{,}938$$

έτσι ώστε να επιβεβαιώνεται η επιλογή του μοντέλου με τη μεταβλητή X_4, αλλά το μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$$

δίνει μεγαλύτερη τιμή στο R^2 διατηρώντας το ίδιο πλήθος μεταβλητών.

Η εφαρμογή του κριτηρίου *SSE* στα δεδομένα του παραδείγματος της Παραγράφου 2.2 μας δίνει ως βέλτιστο μοντέλο το

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$$

Είναι αναμενόμενο τα δύο αυτά κριτήρια να καταλήγουν στο ίδιο βέλτιστο μοντέλο εφόσον συνδέονται με σχέση αναλογίας ως προς την μονοτονία τους όταν αυξάνουμε τον αριθμό των χρησιμοποιούμενων μεταβλητών. Και σε αυτό κριτήριο η εμφάνιση της μεταβλητής X_4 στα «καλά» γραμμικά μοντέλα είναι αισθητή αλλά δεν δίνει τη καλύτερη τιμή στο *SSE* για μοντέλα τριών μεταβλητών.

Η επιλογή του βέλτιστου μοντέλου με βάση το κριτήριο του τροποποιημένου συντελεστή προσδιορισμού R^2_{adj} είναι πιο εύκολη αφού δέχεται ως βέλτιστο εκείνο που μεγιστοποιεί την τιμή του. Το μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon$$

αποτελεί το βέλτιστο μοντέλο για τα δεδομένα του παραδείγματος της Παραγράφου 2.2 με τιμή του τροποποιημένου συντελεστή προσδιορισμού

$$R^2_{adj} = 0{,}945.$$

Η εφαρμογή του κριτηρίου *MSE* στα δεδομένα του παραδείγματος της Παραγράφου 2.2 μας δίνει ως βέλτιστο μοντέλο το πλήρες μοντέλο ή, αν θα θέλαμε να κάνουμε οικονομία στις μετρήσεις, το μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon \ .$$

Παρατηρούμε πως και αυτό το κριτήριο δημιουργεί ένα προβληματισμό ως προς την επιλογή του βέλτιστου ανάμεσα στα γραμμικά μοντέλα

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon \quad \text{και} \quad Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_4 + \beta_3 X_5 + \varepsilon$$

αφού από τη μία το δεύτερο μοντέλο περιέχει τη μεταβλητή X_4 που εμφανίζεται σε όλα τα «καλά» υποσύνολα μεταβλητών και από την άλλη το πρώτο μοντέλο δίνει την μικρότερη τιμή του κριτηρίου σε συνδυασμό με το πλήθος των μεταβλητών.

Η εφαρμογή του κριτηρίου C_p στα δεδομένα του παραδείγματος της Παραγράφου 2.2 δίνει ως καλύτερα μοντέλα τα εξής:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon \text{ και } Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon$$

Και τα δύο μοντέλα, για την κατηγορία μεταβλητών που ανήκουν, έχουν μικρή τιμή για το κριτήριο C_p και σχεδόν ίση με το αντίστοιχο αριθμό παραμέτρων του μοντέλου. Όταν η επιλογή του βέλτιστου μοντέλου δεν είναι ξεκάθαρη τότε είναι θέμα προσωπικής κρίσης αν κάποιος θα προτιμήσει:

α. μια μεροληπτική εξίσωση που δεν αντιπροσωπεύει τα πραγματικά δεδομένα τόσο καλά, επειδή έχει μεγαλύτερο SSE_p (έτσι ώστε $C_p > p$) αλλά έχει μικρότερη εκτίμηση C_p της συνολικής μεταβλητότητας από το πραγματικό αλλά άγνωστο μοντέλο

β. μια εξίσωση με περισσότερες παραμέτρους που προσαρμόζει καλύτερα τα πραγματικά δεδομένα (δηλαδή $C_p = p$) αλλά έχει μια μεγαλύτερη συνολική μεταβλητότητα από το αληθινό αλλά άγνωστο μοντέλο. Με άλλα λόγια, το μικρότερο μοντέλο έχει τη μικρότερη τιμή C_p, αλλά η τιμή C_p του μεγαλύτερου είναι πλησιέστερα στη τιμή του p.

στη περίπτωση θα πρέπει να επιλέξουμε ως βέλτιστο γραμμικό μοντέλο το

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon \ .$$

Η εφαρμογή του κριτηρίου $PRESS_p$ στα δεδομένα του παραδείγματος της Παραγράφου 2.2 μας δίνει ως βέλτιστο μοντέλο το

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$$

Στη περίπτωση αυτή μια προφανής επιλογή είναι το μοντέλο που περιέχει τις μεταβλητές X_1 και X_4 το οποίο έχει μικρή τιμή για το άθροισμα τετραγώνων εκτιμώμενων σφαλμάτων που είναι ίση με 68,73. Υπάρχουν και μικρότερες τιμές

αθροίσματος τετραγώνων εκτιμώμενων σφαλμάτων αλλά απαιτούν τη χρήση τριών ανεξάρτητων μεταβλητών ενώ το κέρδος είναι συγκριτικά μικρό.

Η τιμή του αθροίσματος τετραγώνων εκτιμώμενων σφαλμάτων μειώνεται καθώς αυξάνονται οι μεταβλητές, αλλά καθώς μεταβαίνουμε από τις τρείς μεταβλητές στις τέσσερις η μεταβολή της τιμής του κριτηρίου είναι μικρή και επομένως επιλέγουμε ως βέλτιστο το μοντέλο με τις τρείς μεταβλητές.

Το κριτήριο $PRESS_p$ έχει το πλεονέκτημα ότι δίνει αρκετή λεπτομερή πληροφορία σχετικά με τη σταθερότητα των διαφόρων προσαρμοσμένων μοντέλων στο χώρο των δεδομένων. Ένα μεγάλο μειονέκτημα όμως είναι ότι απαιτεί τεράστιους υπολογισμούς. Επίσης δεν υπάρχουν σαφείς κανόνες για την επιλογή του μοντέλου.

Η επιλογή του βέλτιστου γραμμικού μοντέλου με βάση την εφαρμογή των κριτηρίων *AIC* και *BIC* απαιτεί την ελαχιστοποίηση τους. Σύμφωνα με το κριτήριο *AIC* βέλτιστο θεωρείται το μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon$$

το οποίο δίνει την μικρότερη τιμή 19,9 για το κριτήριο.

Σύμφωνα με το κριτήριο *BIC* βέλτιστο θεωρείται το μοντέλο

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$$

το οποίο δίνει την μικρότερη τιμή 26,64 για το κριτήριο .

Η εφαρμογή του κριτηρίου S_p στα δεδομένα του παραδείγματος της Παραγράφου 2.2 μας δίνει ως βέλτιστο μοντέλο το

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$$

Η επιλογή του μοντέλου έχει γίνει με βάση την ελαχιστοποίηση του κριτηρίου, παρατηρούμε πως η ελάχιστη τιμή 0,073 δίνεται όταν στο μοντέλο υπάρχουν τέσσερις μεταβλητές, η τιμή αυτή όμως δε διαφέρει σημαντικά με την ελάχιστη τιμή που παίρνουμε από το μοντέλο με τις μεταβλητές X_1, X_3 και X_5 το οποίο και θα θεωρήσουμε βέλτιστο.

Στον Πίνακα 2.12 παρουσιάζονται τα βέλτιστα μοντέλα που προέκυψαν από την εφαρμογή του κάθε κριτηρίου.

Πίνακας 2.12: Τα *βέλτιστα μοντέλα των κριτηρίων για τα δεδομένα του παραδείγματος κατανάλωσης βενζίνης*

Κριτήρια	Βέλτιστο γραμμικό μοντέλο
R^2	$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$
SSE	$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$
R^2_{adj}	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon$
MSE	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$
C_p	$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$
$PRESS_p$	$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$
AIC	$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon$
BIC	$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$
S_p	$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon$

ΚΕΦΑΛΑΙΟ 3

Επαναληπτικές μέθοδοι

3.1 Εισαγωγικά

Στις περιπτώσεις όπου το σύνολο των ανεξάρτητων μεταβλητών X είναι αρκετά μεγάλο, η εύρεση ενός «καλού» υποσυνόλου μεταβλητών με βάση τα κριτήρια του Κεφαλαίου 2 ίσως να μην είναι εφικτή, δεδομένου ότι τα γραμμικά μοντέλα που πρέπει να δημιουργηθούν είναι πάρα πολλά και ο απαιτούμενος χρόνος ακόμα και με ηλεκτρονικό υπολογιστή είναι απαγορευτικός.

Για το λόγο αυτό αναζητούμε διαδικασίες οι οποίες μπορούν να μειώσουν τον αριθμό των ανεξάρτητων μεταβλητών και έτσι η δημιουργία ενός «καλού» υποσυνόλου, που θα χρησιμοποιηθεί για την επιλογή του βέλτιστου γραμμικού μοντέλου, να είναι εφικτή. Αυτές οι διαδικασίες καλούνται επαναληπτικές μέθοδοι. Σκοπός τους είναι να δημιουργήσουν μια αλληλουχία γραμμικών μοντέλων εισάγοντας ή διαγράφοντας κάθε φορά μια ανεξάρτητη μεταβλητή με τελικά αποτέλεσμα τη δημιουργία είτε του βέλτιστου γραμμικού μοντέλου είτε ενός μοντέλου που βρίσκεται «πολύ κοντά» στο βέλτιστο.

Μια σημαντική διαφορά ανάμεσα στις επαναληπτικές μεθόδους και τα κριτήρια που αναφέραμε στο Κεφάλαιο 2 κατά την εξέταση όλων των δυνατών μοντέλων, είναι πως οι επαναληπτικές μέθοδοι αναζητούν ένα υποσύνολο μεταβλητών με σκοπό να ορίσουν το βέλτιστο γραμμικό μοντέλο, ενώ στο Κεφάλαιο 2 μελετήσαμε κριτήρια

που πρέπει να εφαρμοστούν σε κάθε γραμμικό μοντέλο έως ότου καταλήξουν στο βέλτιστο.

Από την άλλη πλευρά με την εξέταση όλων των δυνατών μοντέλων διάφορα γραμμικά μοντέλα μπορούν να θεωρηθούν ως «καλά» και από αυτά να ορίσουμε το βέλτιστο, σε αντίθεση με τις επαναληπτικές μεθόδους που καταλήγουν σε ένα μόνο μοντέλο. Το τελευταίο αποτελεί και μια βασική αδυναμία των επαναληπτικών μεθόδων αφού σε κάποιες περιπτώσεις η λανθασμένη επιλογή της αλληλουχίας υποσυνόλων μεταβλητών μπορεί να οδηγήσει στη δημιουργία ενός «φτωχού», για εκτίμηση, γραμμικού μοντέλου καθώς και να αποκλείσει κάποια άλλα μοντέλα που ενδεχομένως να είναι «καλά».

Ο καλύτερος τρόπος για την εύρεση του βέλτιστου γραμμικού μοντέλου, όταν το πλήθος των ανεξάρτητων μεταβλητών είναι αρκετά μεγάλο, είναι η επιλογή ενός μικρότερου υποσυνόλου μεταβλητών χρησιμοποιώντας τις επαναληπτικές μεθόδους και στη συνέχεια η εφαρμογή στο υποσύνολο αυτό καθενός από τα κριτήρια που αναφέραμε στο Κεφάλαιο 2 για τον εντοπισμό του βέλτιστου μοντέλου.

3.2 Η μέθοδος Forward Ranking

Η μέθοδος αυτή ταξινομεί τις ανεξάρτητες μεταβλητές κατά φθίνουσα σειρά όσον αφόρα την σπουδαιότητά τους. Ως πιο σημαντική ανεξάρτητη μεταβλητή θεωρείται εκείνη που δίνει τη μικρότερη τιμή για το κριτήριο

$$F_s = \frac{\frac{SS_{k-p}}{k-p}}{MSE_k}$$

όπου

$$SS_{k-p} = SSR_k - SSR_p$$

και k ο αριθμός των ανεξάρτητων μεταβλητών που χρησιμοποιούνται. Η διαδικασία εφαρμογής του κριτηρίου είναι η εξής:

Βήμα 1. Υπολογίζουμε την τιμή του κριτηρίου F_s για κάθε μία ανεξάρτητη μεταβλητή και επιλέγουμε εκείνη που δίνει τη μικρότερη τιμή για το F_s. Επειδή η τιμή του κριτηρίου F_s για κάθε μια μεταβλητή επηρεάζεται από την τιμή του SS_{k-p}, δεδομένου ότι το MSE_k παίρνει μια μόνο τιμή (αυτή που προκύπτει από το πλήρες μοντέλο), θα μπορούσαμε να επιλέξουμε ως σημαντική μεταβλητή εκείνη που ελαχιστοποιεί το SS_{k-p} ή ισοδύναμα μεγιστοποιεί το SSR_1 (Regression Sum of Squares) για μια μεταβλητή.

Βήμα 2. Ελέγχουμε τις υπόλοιπες μεταβλητές που παρέμειναν σε συνδυασμό με την ανεξάρτητη μεταβλητή που από το Βήμα 1 θεωρείται σημαντική και αναζητούμε το ζευγάρι των ανεξάρτητων μεταβλητών που μας δίνει τη μεγαλύτερη τιμή για το SSR_2 (Regression Sum of Squares για δύο μεταβλητές). Η μεταβλητή που μας δίνει το επιθυμητό αποτέλεσμα είναι η δεύτερη πιο σημαντική μετά από εκείνη που επιλέχτηκε κατά την εκτέλεση του Βήματος 1.

Βήμα 3. Επαναλαμβάνουμε την ίδια διαδικασία αναζητώντας τη μεταβλητή που σε συνδυασμό με τις ανεξάρτητες μεταβλητές του Βήματος 2 δίνει τη μεγαλύτερη τιμή για το SSR_3 (Regression Sum of Squares για τρεις μεταβλητές). Αυτή θεωρείται ως η τρίτη πιο σημαντική ανεξάρτητη μεταβλητή. Η διαδικασία ταξινόμησης των ανεξάρτητων μεταβλητών ανάλογα με την σπουδαιότητά τους ολοκληρώνεται με την εξέταση και των k ανεξάρτητων μεταβλητών.

Η τιμή του κριτηρίου που έχει καθοριστεί σε κάθε ένα από τα παραπάνω βήματα, κατά την ολοκλήρωση του βήματος, συγκρίνεται με την τιμή του $F_{k-p,n-k}(a)$ συνήθως για α=0,05.

Εάν η τιμή του F_s σε κάποιο βήμα θεωρηθεί σημαντική, δηλαδή ικανοποιεί τη σχέση

$$F_s > F_{k-p,n-k}(a)$$

τότε η αντίστοιχη μεταβλητή που έχει επιλεχθεί στο βήμα αυτό παραμένει στο μοντέλο.

Η διαδικασία επιλογής των ανεξάρτητων μεταβλητών συνεχίζεται μέχρι να φτάσουμε στο βήμα εκείνο για το οποίο η τιμή του κριτηρίου F_s δεν ικανοποιεί τη σχέση

$$F_s > F_{k-p,n-k}(a).$$

Οι ανεξάρτητες μεταβλητές που έχουν παραμείνει στο γραμμικό μοντέλο μετά τον υπολογισμό των τιμών του κριτηρίου F_s σε κάθε βήμα είναι αυτές που ορίζουν το βέλτιστο γραμμικό μοντέλο.

Εφαρμόζουμε στη συνέχεια τη μέθοδο *Forward Ranking* για τα δεδομένα του παραδείγματος της Παραγράφου 2.2.

Βήμα 1. Υπολογίζουμε την τιμή του κριτηρίου

$$F_s = \frac{\frac{SS_{k-p}}{k-p}}{MSE_k}$$

για την ανεξάρτητη μεταβλητή που μεγιστοποιεί το. Οι τιμές του SSR_1 δίνονται στο παρακάτω πίνακα:

Μοντέλο	*SSR*
X_1	745,588
X_2	644,988
X_3	688,897
X_4	775,524
X_5	190,211

Παρατηρούμε ότι η τιμή που μεγιστοποιεί το κριτήριο SSR_1 είναι 775,524 και τη δίνει η μεταβλητή X_4. Υπολογίζουμε το F_s για την μεταβλητή αυτή και έχουμε F_s=9,98. Η τιμή του κριτηρίου F_s συγκρίνεται με την τιμή του ποσοστιαίου σημείου $F_{4,26}(0,05) = 2,74$, και αφού ισχύει

$$F_s > F_{4,26}(0{,}05)$$

η μεταβλητή X_4 παραμένει στο γραμμικό μοντέλο.

Βήμα 2. Υπολογίζουμε την τιμή του κριτηρίου

$$F_s = \frac{\frac{SS_{k-p}}{k-p}}{MSE_k}$$

για την ανεξάρτητη μεταβλητή που μεγιστοποιεί το SSR_2. Οι τιμές του SSR_2 δίνονται στο παρακάτω πίνακα.

Μοντέλο	*SSR*
X_1X_4	813,27
X_2X_4	786,442
X_3X_4	781,7
X_4X_5	775,529

Παρατηρούμε ότι η τιμή που μεγιστοποιεί το κριτήριο SSR_2 είναι 813,27 και τη δίνουν οι μεταβλητές X_1, X_4. Υπολογίζουμε το F_s για τις μεταβλητές αυτές και έχουμε F_s=5,757. Συγκρίνοντας την τιμή του κριτηρίου F_s με την τιμή του ποσοστιαίου σημείου $F_{3,26}(0{,}05) = 2{,}98$ βρίσκουμε

$$F_s > F_{3,26}(0{,}05)$$

επομένως η μεταβλητή X_1 παραμένει στο γραμμικό μοντέλο.

Βήμα 3. Υπολογίζουμε την τιμή του κριτηρίου

$$F_s = \frac{\frac{SS_{k-p}}{k-p}}{MSE_k}$$

για την ανεξάρτητη μεταβλητή που μεγιστοποιεί το SSR_3. Οι τιμές του SSR_3 δίνονται στο παρακάτω πίνακα.

Μοντέλο	*SSR*
$X_1X_2X_4$	816,055
$X_1X_3X_4$	815,089
$X_1X_4X_5$	826,951

Παρατηρούμε ότι η τιμή που μεγιστοποιεί το κριτήριο SSR_3 είναι 826,951 και τη δίνουν οι μεταβλητές X_1, X_4 και X_5. Υπολογίζουμε το F_s για τις μεταβλητές αυτές και έχουμε F_s=4,53.

Η τιμή του κριτηρίου F_s συγκρίνεται με την τιμή του ποσοστιαίου σημείου $F_{2,26}(0,05) = 3,37$ και αφού ισχύει

$$F_s > F_{2,26}(0,05),$$

η μεταβλητή X_5 παραμένει στο γραμμικό μοντέλο.

Βήμα 4. Υπολογίζουμε την τιμή του κριτηρίου

$$F_s = \frac{\frac{SS_{k-p}}{k-p}}{MSE_k}$$

για την ανεξάρτητη μεταβλητή που μεγιστοποιεί το SSR_4. Οι τιμές του SSR_4 δίνονται στο παρακάτω πίνακα.

Μοντέλο	*SSR*
$X_1X_2X_4X_5$	827,746
$X_1X_3X_4X_5$	840,27

Παρατηρούμε ότι η τιμή που μεγιστοποιεί το κριτήριο SSR_4 είναι 840,27και τη δίνουν οι μεταβλητές X_1, X_3, X_4, και X_5, υπολογίζουμε το F_s για τις μεταβλητές αυτές και έχουμε F_s=1,06. Η τιμή του κριτηρίου F_s

συγκρίνεται με το τιμή του ποσοστιαίου σημείου $F_{1,26}(0,05) = 4,23$. Τώρα όμως ισχύει

$$F_s < F_{1,26}(0,05)$$

οπότε η μεταβλητή X_5 απορρίπτεται από το γραμμικό μοντέλο.

Στο σημείο αυτό τερματίζεται η διαδικασία *Forward Ranking* προτείνοντας ως βέλτιστο γραμμικό μοντέλο το

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_4 + \beta_3 X_5 + \varepsilon \ .$$

3.3 Η μέθοδος Backward Ranking

Η μέθοδος αυτή θα μπορούσαμε να πούμε ότι λειτουργεί αντίστροφα από τη μέθοδο *Forward Ranking*, δηλαδή στη περίπτωση αυτή ξεκινάμε με το πλήρες μοντέλο (αυτό που περιέχει όλες τις διαθέσιμες μεταβλητές) και απορρίπτουμε κάποιες ανεξάρτητες μεταβλητές, σύμφωνα με τη διαδικασία που θα αναφέρουμε παρακάτω, με στόχο τη δημιουργία του βέλτιστου μοντέλου.

Η μέθοδος *Backward Ranking* ταξινομεί τις ανεξάρτητες μεταβλητές με αύξουσα σειρά όσον αφορά την σπουδαιότητά του. Τα βήματα της διαδικασίας εφαρμογής της μεθόδου είναι τα εξής:

Βήμα 1. Υπολογίζουμε την τιμή του κριτηρίου F_s για κάθε μία ανεξάρτητη μεταβλητή του πλήρους μοντέλου και επιλέγουμε εκείνη που δίνει τη μεγαλύτερη τιμή για το F_s.

Επειδή η τιμή του κριτηρίου F_s για κάθε μια μεταβλητή επηρεάζεται από τη τιμή του SS_{k-1} (δεδομένου ότι το MSE_k παίρνει μια μόνο τιμή, αυτή που προκύπτει από το πλήρες μοντέλο) θα μπορούσαμε να επιλέξουμε ως λιγότερο σημαντική μεταβλητή εκείνη η οποία ελαχιστοποιεί το SSR_1 (Regression Sum of Squares για μια μεταβλητή).

Βήμα 2. Ελέγχουμε τις υπόλοιπες μεταβλητές που παρέμειναν σε συνδυασμό με την ανεξάρτητη μεταβλητή που από το Βήμα 1. θεωρείται λιγότερο σημαντική και αναζητούμε το ζευγάρι των ανεξάρτητων μεταβλητών που μας δίνει την μικρότερη τιμή για το SSR_2 (Regression Sum of Squares για δύο μεταβλητές). Η μεταβλητή που μας δίνει το επιθυμητό αποτέλεσμα είναι η δεύτερη λιγότερο σημαντική μετά από εκείνη του Βήματος 1.

Η διαδικασία ταξινόμησης των ανεξάρτητων μεταβλητών ανάλογα με τη σπουδαιότητά τους ολοκληρώνεται με την εξέταση και των k ανεξάρτητων μεταβλητών, με αποτέλεσμα να έχουν τοποθετηθεί οι ανεξάρτητες μεταβλητές σε αύξουσα σειρά.

Η τιμή του κριτηρίου που έχει καθοριστεί σε κάθε ένα από τα παραπάνω βήματα, κατά την ολοκλήρωση του βήματος, συγκρίνεται με την τιμή του $F_{k-p,n-k}(a)$ συνήθως για a=0,05. Εάν η τιμή του F_s σε κάθε βήμα θεωρηθεί μη σημαντική, δηλαδή δεν ικανοποιεί τη σχέση

$$F_s > F_{k-p,n-k}(a)$$

τότε η αντίστοιχη μεταβλητή που έχει επιλεχθεί στο βήμα αυτό απορρίπτεται από το γραμμικό μοντέλο.

Η διαδικασία επιλογής των ανεξάρτητων μεταβλητών συνεχίζεται μέχρι να φτάσουμε στο βήμα εκείνο για το οποίο η τιμή του κριτηρίου F_s ικανοποιεί τη σχέση

$$F_s > F_{k-p,n-k}(a)\,.$$

Η μεταβλητή που αντιστοιχεί στο βήμα αυτό καθώς και οι υπόλοιπες μεταβλητές που θεωρούνται πιο σημαντικές συντελούν το βέλτιστο γραμμικό μοντέλο.

Εφαρμόζουμε τώρα τη μέθοδο *Backward Ranking* για τα δεδομένα του Παραδείγματος της Παραγράφου 2.2:

Βήμα 1. Υπολογίζουμε την τιμή του κριτηρίου F_s για την ανεξάρτητη μεταβλητή που ελαχιστοποιεί το SSR_1. Η μεταβλητή που δίνει το ελάχιστο SSR_1 είναι η X_5, όπου με $SS_1(X_5) = 190{,}211$ ενώ η αντίστοιχη τιμή για το κριτήριο F είναι ίση με

$$F_s = 114{,}21.$$

Η τιμή του F_s ικανοποιεί τη σχέση

$$F_s > F_{1,26}(0{,}05) = 4{,}23$$

με αποτέλεσμα να παραμένει η X_5 στο γραμμικό μοντέλο αλλά και ταυτόχρονα να τερματίζεται η διαδικασία επιλογής βέλτιστου μοντέλου με τη μέθοδο *Backward Ranking*

Επομένως με τη διαδικασία *Backward Ranking* προτείνεται ως βέλτιστο γραμμικό μοντέλο το

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon .$$

3.4 Η μέθοδος Stepwise Regression

Σύμφωνα με τη μέθοδο Stepwise Regression, το βέλτιστο γραμμικό μοντέλο αναπτύσσεται με την προσθήκη μιας ανεξάρτητης μεταβλητής κάθε φορά (εφόσον θεωρείται ότι είναι σημαντική για την πρόβλεψη της εξαρτημένης) αλλά και τον παράλληλο έλεγχο κατά πόσο το διαμορφωμένο σε κάθε φάση μοντέλο χρειάζεται όλες τις εισαχθείσες μεταβλητές.

Η εισαγωγή της κάθε μεταβλητής γίνεται εφόσον ικανοποιείται κάποιο κριτήριο επιλογής της για εισαγωγή, ενώ αντίστοιχο κριτήριο εφαρμόζεται για να αποφασισθεί αν κάποια ήδη εισηγμένη μεταβλητή θα πρέπει να αφαιρεθεί από το μοντέλο (γιατί κάποιες άλλες μεταβλητές μπορούν να καλύψουν την «απουσία» της). Η διαδικασία εφαρμογής της μεθόδου Stepwise Regression σε μορφή βημάτων είναι η εξής:

Βήμα 1. Μια ανεξάρτητη μεταβλητή εισάγεται στο μοντέλο εφόσον ικανοποιεί το κριτήριο

$$F_k^* = \frac{MSR(X_k)}{MSE(X_k)} > F_{1,n-p}(a)$$

Αν δεν υπάρχουν μεταβλητές που να ικανοποιούν το κριτήριο αυτό τότε δε μπορούμε να προχωρήσουμε, αν όμως υπάρχουν κάποιες μεταβλητές τότε επιλέγουμε, για να εισαχθεί στο μοντέλο, τη μεταβλητή εκείνη με τη μεγαλύτερη τιμή του F_k^*.

Ας θεωρήσουμε πως η μεταβλητή X_1 επιλέγεται πρώτη για να εισαχθεί στο μοντέλο, δηλαδή δημιουργείται αρχικά το μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon .$$

Βήμα 2. Προσαρμόζουμε όλα τα μοντέλα που έχουν την μεταβλητή X_1 και άλλη μία, δηλαδή μοντέλα της μορφής

$$Y = \beta_0 + \beta_1 X_1 + \beta_k X_k + \varepsilon \ , \ k = 2,..., p-1 .$$

Η επιλογή των μεταβλητών γίνεται εφόσον ικανοποιείται το κριτήριο

$$F_{k,1}{}^* = \frac{MSR(X_k / X_1)}{MSE(X_k, X_1)} > F_{1,n-p}(a)$$

Αν δεν υπάρχουν μεταβλητές που να ικανοποιούν το κριτήριο αυτό τότε δεν μπορούμε να προχωρήσουμε. Αν όμως υπάρχουν κάποιες μεταβλητές τότε επιλέγω να εισάγω στο μοντέλο τη μεταβλητή με τη μεγαλύτερη τιμή του $F_{k,1}{}^*$. Ας θεωρήσουμε ότι η μεταβλητή X_2 μεγιστοποιεί τη τιμή του κριτηρίου $F_{k,1}{}^*$, οπότε εισάγεται στο μοντέλο και έτσι παίρνει τη μορφή

$$Y = \beta_0 + \beta_1 X_1 + + \beta_2 X_2 + \varepsilon .$$

Βήμα 3. Στο στάδιο αυτό ελέγχουμε αν μπορούμε να απορρίψουμε από το μοντέλο μεταβλητές που είναι άχρηστες μετά την εισαγωγή της τελευταίας μεταβλητής που μπήκε στο μοντέλο. Η επιλογή τέτοιων μεταβλητών γίνεται με βάση τη σχέση

$$F_{k,2}{}^{**} = \frac{MSR(X_k / X_2)}{MSE(X_k, X_2)} < F_{1,n-p}(a)$$

Αν δεν υπάρχουν τέτοιες μεταβλητές τότε δε βγάζουμε καμία μεταβλητή από το γραμμικό μοντέλο. Αν υπάρχουν επιλέγουμε εκείνη που ελαχιστοποιεί το $F_{k,2}^{**}$ και τη βγάζουμε από το μοντέλο.

Βήμα 4. Επαναλαμβάνουμε τα Βήματα 2 και 3 μέχρις ότου να μην μπορούμε να εισάγουμε καμιά νέα μεταβλητή ούτε να εξάγουμε κάποια μεταβλητή που βρίσκεται ήδη στο μοντέλο.

Εφαρμόζουμε στη συνέχεια τη μέθοδο *Stepwise Regression* στα δεδομένα του Παραδείγματος της Παραγράφου 2.2. Θ θεωρήσουμε ως τιμή εισαγωγής των ανεξάρτητων μεταβλητών στα γραμμικό μοντέλο το $F_{enter} = 3{,}84$ και τιμή εξαγωγής $F_{remove} = 2{,}710$.

Βήμα 1. Υπολογίζουμε τη τιμή του κριτηρίου

$$F_k^* = \frac{MSR(X_k)}{MSE(X_k)} > F_{1,n-p}(a)$$

για κάθε μια μεταβλητή και έχουμε τα εξής αποτελέσματα:

F_1^*	153,040
F_2^*	76,197
F_3^*	99,890
F_4^*	203,940
F_5^*	7,699

Με βάση τα τιμές αυτές, επιλέγουμε να εισάγουμε στο μοντέλο τη μεταβλητή X_4 η οποία ικανοποιεί το κριτήριο αλλά ταυτόχρονα μας δίνει και τη μεγαλύτερη τιμή του $F_k^* = \frac{MSR(X_k)}{MSE(X_k)}$.

Βήμα 2. Υπολογίζουμε τη τιμή του κριτηρίου

$$F_{k,1}{}^{*} = \frac{MSR(X_k / X_1)}{MSE(X_k, X_1)}$$

για τις υπόλοιπες μεταβλητές δεδομένου ότι η μεταβλητή X_4 υπάρχει στο μοντέλο και έχουμε τα εξής αποτελέσματα:

$F_{1,4}{}^{*}$	14,82
$F_{2,4}{}^{*}$	3,08
$F_{3,4}{}^{*}$	1,662
$F_{5,4}{}^{*}$	0,0012

επιλέγουμε να εισάγουμε στο μοντέλο τη μεταβλητή X_1 η οποία ικανοποιεί την ανισότητα

$$F_{k,1}{}^{*} = \frac{MSR(X_k / X_1)}{MSE(X_k, X_1)} > F_{1,n-p}(a)$$

αλλά ταυτόχρονα μας δίνει και τη μεγαλύτερη τιμή του αριστερού μέλους.

Βήμα 3. Μετά την εισαγωγή της δεύτερης μεταβλητής X_1 στο μοντέλο ελέγχουμε εάν εξακολουθεί η αρχική X_4 να είναι σημαντική για το μοντέλο. Υπολογίζουμε τη τιμή της ποσότητας

$$F_{4,1}{}^{**} = \frac{MSR(X_4 / X_1)}{MSE(X_4, X_1)}$$

και αφού έχουμε

$$F_{4,1}{}^{**} = 26,58 > 2,710$$

παραμένουν στο γραμμικό μοντέλο και οι δύο μεταβλητές.

Βήμα 4. Επιστρέφουμε στο Βήμα 2 και υπολογίζουμε τη τιμή του κριτηρίου για κάθε μια μεταβλητή δεδομένου ότι οι μεταβλητές X_4 και X_1 υπάρχουν στο μοντέλο. Οι τιμές που προκύπτουν για το κριτήριο είναι οι εξής:

$F_{2,4.1}^{*}$	1,089
$F_{3,4.1}^{*}$	0,70
$F_{5,4.1}^{*}$	6,45

επιλέγουμε να εισάγουμε στο μοντέλο τη μεταβλητή X_5 η οποία ικανοποιεί το κριτήριο αλλά ταυτόχρονα μας δίνει και τη μεγαλύτερη τιμή του. Έχοντας εισάγει άλλη μια μεταβλητή στο μοντέλο θα πρέπει να εξετάσουμε εάν οι προηγούμενες εξακολουθούν να είναι σημαντικές. Για το λόγο αυτό εφαρμόζουμε το Βήμα 3.και έχουμε τα εξής αποτελέσματα για το αντίστοιχο κριτήριο:

$F_{5,1}^{*}$	145,98
$F_{5,4}^{*}$	585,318

εφόσον οι τιμές του κριτηρίου και για τις δύο μεταβλητές είναι μεγαλύτερες από την δοσμένη τιμή εξαγωγής μεταβλητών $F_{remove} = 2{,}710$, παραμένουν στο μοντέλο και οι τρείς ανεξάρτητες μεταβλητές X_4, X_1 και X_5.

Βήμα 5. Επιστρέφουμε στο Βήμα 2 και υπολογίζονατς το αντίστοιχο κριτήριο, οι τιμές που προκύπτουν είναι οι εξής:

$F_{2,1.4.5}^{*}$	0,366
$F_{3,1.4.5}^{*}$	8,004

Εισάγουμε στο μοντέλο τη μεταβλητή X_3 η οποία ικανοποιεί το κριτήριο. Στη συνέχεια θα ελέγξουμε την σημαντικότητα των μεταβλητών που ήδη υπάρχουν στο μοντέλο δεδομένης της εισαγωγής της X_3. Οι τιμές που προκύπτουν είναι οι εξής:

$F_{1,3}^*$	28,09
$F_{4,3}^*$	1,662
$F_{5,3}^*$	2,12

Παρατηρούμε ότι οι μεταβλητές X_4 και X_5 ικανοποιούν το κριτήριο του Βήματος 3 αλλά θα επιλέξουμε να απορρίψουμε από το μοντέλο εκείνη που ελαχιστοποιεί την τιμή του, δηλαδή τη μεταβλητή X_4.

Η διαδικασία της μεθόδου *Stepwise Regression* σταματάει εφόσον δεν υπάρχει κάποια άλλη μεταβλητή που μπορεί να εισαχθεί στο μοντέλο.

Επομένως το βέλτιστο μοντέλο που προκύπτει με εφαρμογή μεθόδου *Stepwise Regression* είναι το

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon \ .$$

Όπως είναι φανερό η διαδικασία *Stepwise Regression* είναι αρκετά πολύπλοκη και χρειάζεται πολλούς υπολογισμούς. Για το λόγο αυτό υλοποιείται συνήθως (ιδιαίτερα όταν έχουμε μεγάλο πλήθος δεδομένων και μεταβλητών) μόνο με στατιστικά πακέτα.

Με τη χρήση του πακέτου SPSS, για το παράδειγμα της Παραγράφου 2.2, προκύπτει ο παρακάτω πίνακας που μας δίνει απευθείας όλα τας βήματα της διαδικασίας, δηλαδή τις μεταβλητές που εισάγονται ή απορρίπτονται από το γραμμικό μοντέλο.

Πίνακας 3.1: *SPSS output για τη μέθοδο Stepwise Regression*

Variables Entered/Removed(a)

Model	Variables Entered	Variables Removed
1	x4	.
2	x1	.
3	x5	.
4	x3	.
5	.	x4

a Dependent Variable: y

3.5 Η μέθοδος Backward Elimination

Η μέθοδος Backward Elimination θα μπορούσαμε να πούμε ότι είναι η αντίστροφη της μεθόδου *Stepwise Regression*, δηλαδή αρχίζουμε από το πλήρες μοντέλο

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} + \varepsilon$$

και στη συνέχεια απορρίπτουμε μια-μια τις ανεξάρτητες μεταβλητές μέχρι να καταλήξουμε στο βέλτιστο μοντέλο. Η κάθε μία μεταβλητή απορρίπτεται με βάση το κριτήριο

$$F_k^{**} = \frac{MSR(X_k / X_1,..., X_{p-1})}{MSE(X_1,..., X_{p-1})} < F_{1,n-p}(a)$$

και εφόσον έχει τη μικρότερη τιμή για το F_k^{**}. Στη συνέχεια επαναλαμβάνουμε το παραπάνω βήμα και η διαδικασία σταματάει όταν δεν μπορεί να απορριφθεί καμία μεταβλητή.

Εφαρμόζουμε την μέθοδο *Backward Elimination* για το παράδειγμα της Παραγράφου 2.2. Θα χρησιμοποιήσουμε ως τιμή με βάση την οποία απορρίπτονται οι μεταβλητές από γραμμικό μοντέλο την

$$F_{remove} = 2,710 .$$

Οι τιμές που προκύπτουν κατά το πρώτο στάδιο εξέτασης των ανεξάρτητων μεταβλητών για το κριτήριο F_k^{**} είναι οι εξής:

F_1^{**}	24,57
F_2^{**}	1,057
F_3^{**}	8,57
F_4^{**}	2,31
F_5^{**}	8,65

Παρατηρούμε πως οι μεταβλητές X_4 και X_2 ικανοποιούν το κριτήριο

$$F_k^{**} < F_{1,n-p}(a)$$

αλλά εκείνη που ελαχιστοποιεί το F_k^{**} με αποτέλεσμα να απορρίπτεται από το γραμμικό μοντέλο είναι η X_2.

Ελέγχουμε τώρα το γραμμικό μοντέλο που περιέχει τις υπόλοιπες ανεξάρτητες μεταβλητές για να δούμε αν σε αυτό το στάδιο απορρίπτεται κάποια άλλη αυτό. Οι τιμές που προκύπτουν για την ποσότητα F_k^{**} κατά το δεύτερο στάδιο εξέτασης των ανεξάρτητων μεταβλητών είναι οι εξής:

F_1^{**}	33,93
F_3^{**}	7,98
F_4^{**}	2,70
F_5^{**}	15,087

Παρατηρούμε πως μόνο η μεταβλητή X_4 ικανοποιεί το κριτήριο

$$F_k^{**} < F_{1,n-p}(a)$$

επομένως είναι εκείνη που απορρίπτεται από το γραμμικό μοντέλο.

Στη συνέχεια ελέγχουμε εάν οι ανεξάρτητες μεταβλητές X_1, X_3 και X_5 που έχουν παραμείνει στο γραμμικό μοντέλο ικανοποιούν το κριτήριο $F_k^{**} < F_{1,n-p}(a)$. Οι τιμές που προκύπτουν για την ποσότητα F_k^{**} κατά το τρίτο στάδιο είναι οι εξής:

F_1^{**}	74,46
F_3^{**}	34,6
F_5^{**}	27,10

Παρατηρούμε πως καμία από τις μεταβλητές που υπάρχουν στο γραμμικό μοντέλο δεν ικανοποιεί το κριτήριο F_k^{**} επομένως παραμένουν όλες στο μοντέλο και ορίζουν ως βέλτιστο γραμμικό μοντέλο το

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_5 + \varepsilon \ .$$

Και σε αυτή τη περίπτωση, όπως στη μέθοδο *Stepwise Regression*, η διαδικασία επιλογής του βέλτιστου μοντέλου απαιτεί πολλούς υπολογισμούς για το λόγο αυτό συνήθως χρειάζεται να κάνουμε χρήση στατιστικών πακέτων.

Με τη χρήση του πακέτου SPSS, για το παράδειγμα της Παραγράφου 2.2, προκύπτει ο παρακάτω πίνακας που μας δίνει απευθείας τις μεταβλητές που απορρίπτονται από το γραμμικό μοντέλο.

Πίνακας 3.2: *SPSS output για τη μέθοδο Backward Elimination*

Variables Entered/Removed(b)

Model	Variables Entered	Variables Removed	Method
1	x5, x4, x1, x3, x2(a)	.	Enter
2	.	x2	Backward (criterion: Probability of F-to-remove >= ,100).
3	.	x4	Backward (criterion: Probability of F-to-remove >= ,100).

a All requested variables entered.

b Dependent Variable: y

3.6 Η μέθοδος Forward Selection

Η μέθοδος Forward Selection αποτελεί μια πιο απλοποιημένη μορφή της μεθόδου *Stepwise Regression*, αφού η διαδικασία που ακολουθούμε για την επιλογή

του βέλτιστου μοντέλου είναι ίδια παραλείποντας μόνο το βήμα που ελέγχει την σημαντικότητα των ήδη υπαρχόντων ανεξάρτητων μεταβλητών στο μοντέλο σε σχέση με εκείνη που έχει εισαχθεί τελευταία.

Κατά την εφαρμογή της μεθόδου *Forward Selection* στο παράδειγμα της Παραγράφου 2.2 θα καταλήξουμε στην εισαγωγή των ίδιων ανεξάρτητων μεταβλητών στο γραμμικό μοντέλο χωρίς όμως να απορρίψουμε από το μοντέλο την μεταβλητή X_4. Επομένως το βέλτιστο γραμμικό μοντέλο για τη μέθοδο αυτή θεωρείται το

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon .$$

3.7 Η μέθοδος Forward Procedure

Σύμφωνα με τη μέθοδο Procedure (προοδευτική προσθήκη μεταβλητών), το βέλτιστο γραμμικό μοντέλο αναπτύσσεται με την προσθήκη κάθε φορά μιας ανεξάρτητης μεταβλητής. Η μέθοδος αυτή παρουσιάζει ένα μειονέκτημα σε σχέση με τις προηγούμενες στο ότι, με τη μέθοδο αυτή κάθε φορά που εισάγεται στο μοντέλο μια μεταβλητή δεν έχουμε τη δυνατότητα να την απορρίψουμε σε κάποιο παρακάτω βήμα τις διαδικασίας.

Δίνονται στη συνέχεια αναλυτικά τα βήματα για την επιλογή βέλτιστου γραμμικού μοντέλου σύμφωνα με τη μέθοδο *Forward Procedure*.

Βήμα 1. Από το σύνολο των ανεξάρτητων μεταβλητών που είναι υποψήφιες να περιληφθούν στο μοντέλο διαλέγουμε εκείνη που έχει το μεγαλύτερο συντελεστή συσχέτισης με την εξαρτημένη μεταβλητή *Y*.

Βήμα 2. Υπολογίζουμε την τιμή της στατιστικής συνάρτησης

$$T = \frac{\hat{\beta}_k}{s(\hat{\beta}_k)}$$

για την ανεξάρτητη μεταβλητή που έχουμε επιλέξει στο Βήμα 1.Εάν ισχύει

$$|T| > t_{n-p}(a/2)$$

τότε εισάγεται η μεταβλητή αυτή στο μοντέλο.

Βήμα 3. Για να επιλέξουμε τη δεύτερη ανεξάρτητη μεταβλητή υπολογίζουμε το συντελεστή συσχέτισης της εξαρτημένης μεταβλητή Y με κάθε μια από τις υπόλοιπες ανεξάρτητες μεταβλητές και επιλέγουμε εκείνη με τη μεγαλύτερη τιμή διατηρώντας ταυτόχρονα τη μεταβλητή από το Βήμα 1.

Βήμα 4. Επιστρέφουμε στο Βήμα 2 υπολογίζοντας την τιμή της στατιστικής συνάρτησης

$$T = \frac{\hat{\beta}_k}{s(\hat{\beta}_k)}$$

για την ανεξάρτητη μεταβλητή που έχουμε επιλέξει από το Βήμα 2. Η διαδικασία τερματίζεται όταν ικανοποιηθεί η σχέση

$$|T| > t_{n-p}(a/2)$$

σε κάποιο βήμα.

Εφαρμόζουμε τώρα τη μέθοδο *Forward Procedure* για το παράδειγμα της Παραγράφου 2.2.

Βήμα 1. Επιλέγουμε την μεταβλητή που παρουσιάζει τη μεγαλύτερη συσχέτιση με την εξαρτημένη μεταβλητή Y, σύμφωνα με τον πίνακα συσχετίσεων που δίνεται στην επόμενη σελίδα που ελήφθη από το στατιστικό πακέτο SPSS, θα επιλέξουμε ως πρώτη μεταβλητή για αν εισαχθεί στο μοντέλο την X_5.

Βήμα 2. Η τιμή της της στατιστικής συνάρτησης

$$T = \frac{\hat{\beta}_k}{s(\hat{\beta}_k)}$$

για την ανεξάρτητη μεταβλητή X_5 είναι ίση με 2,775. Αφού

$$t_{30-6}(0{,}05/2) = 1{,}71$$

για τη μεταβλητή X_5 ισχύει

$$|2{,}775| > 1{,}71,$$

επομένως η ανεξάρτητη μεταβλητή X_5 παραμένει στο μοντέλο.

Πίνακας 3.3: *Πίνακας συσχετίσεων για τα δεδομένα του παραδείγματος της Παραγράφου* 2.2

Correlations

		y	x1	x2	x3	x4	x5
y	Pearson Correlation	1	-,919(**)	-,855(**)	-,884(**)	-,938(**)	,464(**)
	Sig. (2-tailed)		,000	,000	,000	,000	,010
	N	30	30	30	30	30	30
x1	Pearson Correlation	-,919(**)	1	,928(**)	,829(**)	,873(**)	-,653(**)
	Sig. (2-tailed)	,000		,000	,000	,000	,000
	N	30	30	30	30	30	30
x2	Pearson Correlation	-,855(**)	,928(**)	1	,903(**)	,849(**)	-,765(**)
	Sig. (2-tailed)	,000	,000		,000	,000	,000
	N	30	30	30	30	30	30
x3	Pearson Correlation	-,884(**)	,829(**)	,903(**)	1	,904(**)	-,636(**)
	Sig. (2-tailed)	,000	,000	,000		,000	,000
	N	30	30	30	30	30	30
x4	Pearson Correlation	-,938(**)	,873(**)	,849(**)	,904(**)	1	-,493(**)
	Sig. (2-tailed)	,000	,000	,000	,000		,006
	N	30	30	30	30	30	30
x5	Pearson Correlation	,464(**)	-,653(**)	-,765(**)	-,636(**)	-,493(**)	1
	Sig. (2-tailed)	,010	,000	,000	,000	,006	
	N	30	30	30	30	30	30

Βήμα 3. Η τιμή της στατιστικής συνάρτησης

$$T = \frac{\widehat{\beta}_k}{s(\widehat{\beta}_k)}$$

για την ανεξάρτητη μεταβλητή X_5 είναι ίση με 2,775. Για την μεταβλητή X_5 ισχύει

$$|2{,}775| > 1{,}71 = t_{30-6}(0{,}05/2),$$

επομένως η ανεξάρτητη μεταβλητή παραμένει στο μοντέλο.

Βήμα 4. Επιλέγουμε την μεταβλητή που παρουσιάζει τη μεγαλύτερη συσχέτιση με την εξαρτημένη μεταβλητή *Y,* εκτός από την μεταβλητή X_5 που έχει ήδη επιλεχθεί. Η μεταβλητή X_2 μας δίνει τη μεγαλύτερη τιμή συσχέτισης με την *Y* με τιμή -0,855.

Βήμα 5. Εκτελούμε το Βήμα 2 για την μεταβλητή X_2. Έχουμε

$$T = \frac{\widehat{\beta}_2}{s(\widehat{\beta}_2)} = -8{,}729$$

και ελέγχουμε εάν μπορούμε να εισάγουμε την X_2 στο γραμμικό μοντέλο. Αφού

$$t_{30-6}(0{,}05/2) = 1{,}71$$

και ισχύει

$$|-8{,}729| > t_{30-6}(0{,}05/2)$$

οι ανεξάρτητες μεταβλητές X_2 και X_5 θα πρέπει να διατηρηθούν στο γραμμικό μοντέλο.

Βήμα 6. Επιλέγουμε τη μεταβλητή που παρουσιάζει τη μεγαλύτερη συσχέτιση με την εξαρτημένη μεταβλητή *Y,*εκτός από τις μεταβλητές X_2 και X_5 που έχουν ήδη επιλεγεί. Η μεταβλητή X_3 μας δίνει τη μεγαλύτερη τιμή συσχέτισης με την *Y* με τιμή -0,884.

Βήμα 7. Εκτελούμε και πάλι το Βήμα 2 για την μεταβλητή X_3. Αφού

$$T = \frac{\widehat{\beta}_3}{s(\widehat{\beta}_3)} = -9{,}995$$

θα ισχύει ξανά η ανισότητα

$$|T| = |-9{,}995| > t_{30-6}(0{,}05/2),$$

επομένως οι ανεξάρτητες μεταβλητές X_3, X_2 και X_5 περιέχονται στο γραμμικό μοντέλο.

Βήμα 8. Επιλέγουμε την μεταβλητή που παρουσιάζει τη μεγαλύτερη συσχέτιση με την εξαρτημένη μεταβλητή *Y,*εκτός από τις μεταβλητές X_3, X_2 και X_5 που

έχουν ήδη επιλεγεί. Η μεταβλητή X_1 μας δίνει τη μεγαλύτερη τιμή συσχέτισης με την Y με τιμή -0,919.

Βήμα 9. Εκτελούμε το Βήμα 2 για την μεταβλητή X_1 για την οποία βρίσκουμε

$$T = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} = -12{,}371.$$

Ελέγχοντας εάν μπορούμε να εισάγουμε την X_1 στο γραμμικό μοντέλο, βρίσκουμε $|-12{,}371| > t_{30-6}(0{,}05/2)$ όπου $t_{30-6}(0{,}05/2) = 1{,}71$. Επομένως οι ανεξάρτητες μεταβλητές X_1, X_3, X_2 και X_5 θα περιέχονται στο γραμμικό μοντέλο που αναζητούμε.

Βήμα 10. Επιλέγουμε και την τελευταία ανεξάρτητη μεταβλητή X_4 και ελέγχουμε εάν μπορεί να εισαχθεί και αυτή στο γραμμικό μοντέλο. Για την X_4 έχουμε

$$T = \frac{\hat{\beta}_4}{s(\hat{\beta}_4)} = -14{,}281$$

και παρατηρούμε πως και για αυτή τη μεταβλητή ικανοποιείται η σχέση

$$|T| > t_{n-p}(a/2)$$

αφού

$$|-14{,}281| > t_{30-6}(0{,}05/2).$$

Σύμφωνα λοιπόν με τη μέθοδο *Forward Procedure* θα θεωρήσουμε ως βέλτιστο μοντέλο το πλήρες μοντέλο.

3.8 Η μέθοδος της διαδοχικής αντικατάστασης μεταβλητών

Σε αυτή την παράγραφο θα εξετάσουμε πως μπορούμε να αντικαταστήσουμε μεταβλητές που έχουν επιλεγεί, με κάποια άλλη που θα δίνει μικρότερη τιμή για το *SSR*. Η μέθοδος της διαδοχικής αντικατάστασης παρουσιάζεται αναλυτικά στο βιβλίο Miller (2002).

Ας υποθέσουμε, για παράδειγμα, ότι έχουμε 26 μεταβλητές, που συμβολίζονται με τα γράμματα του αγγλικού αλφαβήτου, και σε κάποιο στάδιο έχουμε επιλέξει ένα σύνολο από τέσσερις μεταβλητές, έστω τις

A, B, C, D

από τις οποίες θέλουμε να αντικαταστήσουμε την μεταβλητή *A*.

Λογικά, θα υπάρχουν πολλές μεταβλητές, από τις υπόλοιπες 22, που θα μπορούσαν να δώσουν μικρή τιμή για το *SSR* όταν βρίσκονται στο ίδιο μοντέλο με τις μεταβλητές *B, C, D*. Ας υποθέσουμε ότι η μεταβλητή που δίνει την ελάχιστη τιμή για το *SSR* είναι η *M*, την οποία και θα αντικαταστήσουμε με την *A* έχοντας πλέον μοντέλο με τις μεταβλητές *M, B, C, D*.

Στη συνέχεια θα μπορούσαν να αντικατασταθούν και οι μεταβλητές *, B, C, D* καθώς και η *M*. Αν σε κάποια από αυτές τις αντικαταστάσεις δε θα βρεθεί μεταβλητή που θα ελαχιστοποιεί την τιμή του *SSR*, συνεχίζουμε την διαδικασία με την επόμενη μεταβλητή. Μερικές φορές οι μεταβλητές που έχουν αντικατασταθεί μπορεί να εμφανιστούν ξανά στο μοντέλο. Η διαδικασία σταματάει όταν η αντικατάσταση μιας μεταβλητής δεν επιφέρει μείωση στην τιμή του *SSR*.

Ας υποθέσουμε ότι το μοντέλο που δίνει την μικρότερη τιμή στο *SSR* αποτελείται από τις μεταβλητές *B, E, S, T*. Εάν ξεκινούσαμε από το μοντέλο με τις μεταβλητές *P, E, S, T* τότε θα καταλήγαμε στο επιθυμητό αποτέλεσμα με μόνο μια αντικατάσταση. Κάτι τέτοιο τις περισσότερες φορές δεν είναι εφικτό δεδομένου του μεγάλου αριθμού των μοντέλων που αποτελούνται από τέσσερις μεταβλητές. Ακόμα και αν σταθούμε τυχεροί και εντοπίσουμε ένα μοντέλο, με τέσσερις ανεξάρτητες μεταβλητές, που δίνει την ελάχιστη τιμή για το *SSR* και πάλι δεν θα είμαστε σίγουροι πως πρόκειται για το καλύτερο μοντέλο με αποτέλεσμα να καταφύγουμε σε άλλες μεθόδους ελέγχου.

Η ακόλουθη τροποποίηση αυξάνει την πιθανότητα εύρεσης ενός κατάλληλου γραμμικού μοντέλου. Υποθέτουμε ότι ξεκινάμε με το μοντέλο που περιέχει τις μεταβλητές *A, B, C, D* και δίνει τιμή για το *SSR* 100. Στο σημείο αυτό αναζητούμε μια μεταβλητή που θα αντικαταστήσει την μεταβλητή *A*, χωρίς όμως να προβούμε στην αντικατάσταση αναζητούμε μια αντικαταστάτρια μεταβλητή για την *B*. Θεωρούμε πως η αντικατάσταση της μεταβλητής *B* από την *M* είναι αυτή που έδωσε

την μικρότερη τιμή για το *SSR* και έτσι συνεχίζουμε την διαδικασία για την αντικατάσταση των υπόλοιπων τριών μεταβλητών (εκτός της *M*).

Η διαδοχική αντικατάσταση μεταβλητών απαιτεί περισσότερους υπολογισμούς από ότι η μέθοδος *forward selection* ή η μέθοδος *stepwise regression* αλλά εξακολουθεί να εφαρμόζεται σε προβλήματα με αρκετές εκατοντάδες μεταβλητές όταν απαιτούνται υποσύνολά της, ας πούμε μέχρι 20-30 μεταβλητές,. Μια εναλλακτική τεχνική που μπορεί να χρησιμοποιηθεί όταν το πλήθος των μεταβλητών είναι αρκετά μεγάλο είναι να επιλεγεί τυχαία στην αρχή ένα γραμμικό μοντέλο και στη συνέχεια να εφαρμοστεί σε αυτό η διαδοχική αντικατάσταση μεταβλητών.

3.9 Η μέθοδος της αντικατάστασης ζευγών μεταβλητών

Η αντικατάσταση δύο μεταβλητών από το σύνολο που έχουμε επιλέξει για την εκτίμηση του γραμμικού μοντέλου, με μία άλλη μειώνει το αριθμό των μεταβλητών του υποσυνόλου κατά ένα και με διαδοχικές εφαρμογές της τεχνικής αυτής μπορούμε να οδηγηθούμε στην εύρεση ενός καλύτερου υποσυνόλου. Η μέθοδος της αντικατάστασης ζευγών μεταβλητών παρουσιάζεται στο βιβλίο του Miller (2002).

Για υποσύνολα μεγέθους p μεταβλητών υπάρχουν $p(p-1)/2$ ζεύγη μεταβλητών που μπορούν να λαμβάνονται υπόψη για αντικατάσταση. Μια συστηματική διαδικασία αντικατάστασης ζευγών μεταβλητών περιγράφεται στα παρακάτω βήματα.

Βήμα 1. Δημιουργούμε ένα τυχαίο υποσύνολο $p/2$ μεταβλητών

Βήμα 2. Εξετάζουμε όλες τις πιθανές αντικαταστάσεις των ζευγών μεταβλητών στο ήδη υπάρχον υποσύνολο. Αυτό περιλαμβάνει την αντικατάσταση μόνο ενός ζεύγους μεταβλητών.

Βήμα 3. Βρίσκουμε την καλύτερη αντικατάσταση του κάθε ζεύγους μεταβλητών, δηλαδή εκείνη που ελαχιστοποιεί την τιμή του *SSR* και μετά την αντικατάσταση επιστρέφουμε στο Βήμα 2.

Βήμα 4. Εάν επαναλάβουμε τα Βήματα 2, 3 και δεν έχουμε δημιουργήσει ένα γραμμικό μοντέλο με περισσότερες μεταβλητές από το αρχικό, τότε εισάγουμε στο μοντέλο μια μεταβλητή και επιστρέφουμε στο Βήμα 2.

Βήμα 5. Επαναλαμβάνουμε την διαδικασία έως ότου το γραμμικό μοντέλο που θα προκύψει να είναι το επιθυμητό.

ΚΕΦΑΛΑΙΟ 4

Σύγκριση των επαναληπτικών μεθόδων

4.1 Εισαγωγικά

Στο Κεφάλαιο αυτό δίνονται ορισμένα παραδείγματα από τη διεθνή βιβλιογραφία στα οποία γίνεται μελέτη της αποτελεσματικότητας των επαναληπτικών μεθόδων επιλογής βέλτιστου συνόλου ανεξάρτητων μεταβλητών σε ένα γραμμικό μοντέλο σε συγκεκριμένα πραγματικά δεδομένα από διάφορες επιστημονικές περιοχές.

Από τις μελέτες αυτές καθίσταται φανερό ότι οι μέθοδοι που θεωρούνται «οικονομικές», όπως forward selection, backward elimination και η μέθοδος διαδοχικής αντικατάστασης μεταβλητών, συνήθως δεν προσαρμόζονται καλά με αποτέλεσμα τα οδηγούν σε λανθασμένη επιλογή υποσυνόλων. Από αυτές τις τρείς μεθόδους εκείνη που προσαρμόζεται καλύτερα στο γραμμικό μοντέλο είναι η μέθοδος της διαδοχικής αντικατάστασης μεταβλητών. Μια πιο εκτεταμένη σύγκριση των μεθόδων έχει γίνει από τον Berk (1978b) ο οποίος εξέτασε δημοσιευμένα παραδείγματα όπου είχαν περισσότερες παρατηρήσεις από μεταβλητές. Σε κάποια από αυτά παρατήρησε πως οι μέθοδοι forward selection και backward elimination προσαρμόστηκαν καλύτερα στα γραμμικά μοντέλα, ανεξάρτητα από τον αριθμό των μεταβλητών. Στις περισσότερες περιπτώσεις οι διαφορές στη τιμή του *SSE* ανάμεσα στα υποσύνολα που έχουν επιλεχθεί από τις παραπάνω μεθόδους είναι μικρές.

4.2 Πρόβλεψη βροχοπτώσεων

Τα δεδομένα του παραδείγματος που θα παρουσιάσουμε στην παράγραφο αυτή προέρχονται από την εργασία του Biondini et al. (1977). Πρόκειται για ένα πείραμα που αναφέρεται στο πως η παρουσία των σύννεφων στον ουρανό επηρεάζει τον καθημερινό αριθμό βροχοπτώσεων, ο οποίος αποτελεί την εξαρτημένη μεταβλητή του πειράματος.

Για την πρόβλεψη των βροχοπτώσεων τις μέρες που τα σύννεφα στον ουρανό είναι πυκνά απαιτείται μια εξίσωση η οποία δημιουργείται από ήδη υπάρχουσες εξισώσεις με παρατηρήσεις από μέρες που δεν υπήρχαν σύννεφα στον ουρανό. Εάν τα πυκνά σύννεφα στον ουρανό επηρεάζουν τις βροχοπτώσεις τότε τα αποτελέσματα των προβλέψεων για τις μέρες αυτές θα είναι εντελώς διαφορετικά από τις παρατηρήσεις που έχουν συλλεχθεί.

Στο συγκεκριμένο πείραμα έχουμε στη διάθεσή μας 5 μεταβλητές, οι οποίες μπορούν να φτάσουν στις 20 αν χρησιμοποιήσουμε όχι μόνο τους γραμμικούς όρους, αλλά και τους αντίστοιχους δευτεροβάθμιους και τέλος όρους αλληλεπίδρασης μεταξύ των μεταβλητών.

Στο πείραμα είχαν συλλεχθεί 58 παρατηρήσεις, κάποιες από αυτές όμως αφορούσαν μέρες που υπήρχαν σύννεφα στον ουρανό. Ο συγγραφέας αποφάσισε να κατηγοριοποιήσει επίσης τις ημέρες σύμφωνα με ραντάρ ήχων.

Ο Πίνακας 4.1 περιλαμβάνει τα δεδομένα των 5 μεταβλητών που χρησιμοποιήθηκαν και τις τιμές της εξαρτημένης μεταβλητής η οποία αναφέρεται στο ύψος των βροχοπτώσεων που περιλαμβάνονται στη μεγαλύτερη υποδιαίρεση της ταξινόμησης των στοιχείων και ανέρχονται στις 14 παρατηρήσεις.

Πίνακας 4.1: *Πίνακας δεδομένων του παραδείγματος της πρόβλεψης βροχοπτώσεων*

Ημερομηνία	X_1	X_2	X_3	X_4	X_5	Y
1 Ιουλίου 1971	2,0	0,041	2,70	2	12	0,32
15 Ιουλίου 1971	3,0	0,100	3,40	1	8	1,18
17 Ιουλίου 1971	3,0	0,607	3,60	1	12	1,93
9 Αυγούστου 1973	23,0	0,058	3,60	2	8	2,67
9 Σεπτεμβρίου 1973	1,0	0,026	3,55	0	10	0,16
25 Ιουνίου 1975	5,3	0,526	4,35	2	6	6,11
9 Ιουλίου 1975	4,6	0,307	2,30	1	8	0,47
16 Ιουλίου 1975	4,9	0,194	3,35	0	12	4,56
18 Ιουλίου 1975	12,1	0,751	4,85	2	8	6,35
24 Ιουλίου 1975	6,8	0,796	3,95	0	10	5,74
30 Ιουλίου 2009	11,3	0,398	4,00	0	12	4,45
16 Αυγούστου 1975	2,2	0,230	3,80	0	8	1,16
28 Αυγούστου 1975	2,6	0,136	3,15	0	12	0,82
12 Σεπτεμβρίου 1975	7,4	0,168	4,65	0	10	0,28

Στον Πίνακα 4.2 παρουσιάζονται τα αποτελέσματα που προέκυψαν με τις μεθόδους *Forward selection* και διαδοχικής αντικατάστασης μεταβλητών. Πιο συγκεκριμένα δίνονται τα βέλτιστα μοντέλα που έδωσε η κάθε μέθοδος καθώς επίσης και η τιμή του αθροίσματος τετραγώνων των υπολοίπων *SSE* για κάθε ένα από αυτά τα μοντέλα.

Σημειώνουμε ότι οι μεταβλητές με ονομασία X_6 έως X_{20} αναφέρονται αντίστοιχα στους παρακάτω τετραγωνικούς όρους και όρους αλληλεπίδρασης:

$$X_1^2, X_2^2, X_3^2, X_4^2, X_5^2,$$

$$X_1X_2, X_1X_3, X_1X_4, X_1X_5, X_2X_3, X_2X_4, X_2X_5, X_3X_4, X_3X_5, X_4X_5.$$

Πίνακας 4.2: *Τιμές του SSE για τα δεδομένα του παραδείγματος της πρόβλεψης βροχοπτώσεων (οι αριθμοί στις παρενθέσεις υποδεικνύουν τις επιλεγμένες μεταβλητές του γραμμικού μοντέλου)*

Πλήθος μεταβλητών	**Forward selection**	**Μέθοδος διαδοχικής αντικατάστασης μεταβλητών**
σταθερός όρος	72,29	72,29
	-	-
1	26,87	26,87
	(15)	(15)
2	21,56	21,56
	(14,15)	(14,15)
3	19,49	19,49
	(14,15,17)	(14,15,17)
4	11,98	11,98
	(12,14,15,17)	(12,14,15,17)
5	9,05	8,7
	(6,12,14,15,17)	(1,12,15,17,19)

Στο παράδειγμα αυτό, με τη χρήση της μεθόδου διαδοχικής αντικατάστασης έχει βρεθεί μόνο ένα υποσύνολο μεταβλητών που θεωρείται καλύτερο από εκείνα τα υποσύνολα που επιλέχθηκαν με την χρήση της μεθόδου *Forward selection*.

Παρατηρούμε πως καμία από τις δύο μεθόδους δεν κατέληξε σε υποσύνολο που να περιέχει τις μεταβλητές $X_{9,}X_{17}$ και X_{20} παρά το γεγονός πως ένα τέτοιο υποσύνολο θεωρείται πως προσαρμόζεται καλύτερα στο γραμμικό μοντέλο, σε αντίθεση με το υποσύνολο με μεταβλητές $X_{14,}X_{15,}X_{17}$ που εμφανίζεται και στις δύο μεθόδους.

Στην πραγματικότητα τα πέντε καλύτερα υποσύνολα μεταβλητών που αποτελούνται από τρεις μεταβλητές προσαρμόζονται πιο καλά στο γραμμικό μοντέλο από κάθε άλλο υποσύνολο μεταβλητών που έχει επιλεγεί είτε με την μέθοδο *Forward selection* είτε με τη μέθοδο της διαδοχικής αντικατάστασης μεταβλητών. Τα αποτελέσματα των καλύτερων υποσυνόλων παρουσιάζονται στο Πίνακα 4.3.

Υπενθυμίζεται ότι οι μεταβλητές με αρίθμηση από το 11 και άνω είναι όροι αλληλεπίδρασης. Παρατηρούμε, και με τη μέθοδο *Forward selection* και με τη μέθοδο της διαδοχικής αντικατάστασης μεταβλητών, για πρώτη φορά επιλέγεται μια «καθαρή» μεταβλητή (δηλαδή όχι αλληλεπίδραση) όταν φτάσουμε σε μοντέλα που χρησιμοποιούν πέντε μεταβλητές. Η πρώτη μεταβλητή που εισέρχεται σε αυτά είναι η X_1 ή το τετράγωνό της X_1^2. Αντίθετα με την μέθοδο της εξέτασης όλων των συνδυασμών, στα βέλτιστα μοντέλα 3 μεταβλητών αρχίζουν να εμφανίζονται «καθαρές» μεταβλητές.

Στη περίπτωση που χρησιμοποιήσουμε την μέθοδο *stepwise regression* θα πρέπει να ελέγξουμε τις τιμές που θα πάρουμε για τα F_d και F_e . Εάν η τιμή του F_e είναι μεγαλύτερη από 2,71 τότε η μέθοδος σταματάει με την εισαγωγή μόνο μίας μεταβλητής στο μοντέλο. Εάν η τιμή του F_e είναι μεταξύ των 1,07 και 2,71 τότε επιλέγουμε το μοντέλο με τις μεταβλητές X_{14} και X_{15}. Εάν η τιμή του F_e είναι μικρότερη από 1,07 τότε η διαδικασία επιλογής γραμμικού μοντέλου είναι ίδια με την εφαρμογή της μεθόδου *Forward selection.* Για να βρούμε το καλύτερο υποσύνολο μεταβλητών που αποτελείται από τρείς μεταβλητές θα πρέπει να κάνουμε κάποια διαγραφή. Για παράδειγμα, μετά την επιλογή του μοντέλου με μεταβλητές X_{14}, X_{15} , X_{17} εάν τιμή του F_d είναι μεγαλύτερη από 2,74 τότε απαιτείται η διαγραφή της μεταβλητής X_{14}. Γενικά μια τέτοια διαγραφή πραγματοποιείται όταν $F_d > F_e$ και η διαδικασία επαναλαμβάνεται κυκλικά.

Πίνακας 4.3: *Τιμές του SSE για τα πέντε καλύτερα υποσύνολα (με* 1-5 *μεταβλητές) στα δεδομένα του παραδείγματος της πρόβλεψης βροχοπτώσεων*

Πλήθος μεταβλητών	***SSE***	**Μεταβλητές**
1	26,87	X_{15}
	27,2	X_{11}
	32,18	X_2
	34,01	X_7
	42,99	X_{17}
2	21,56	$X_{15}X_{14}$
	21,81	$X_{15}X_1$
	22,29	$X_{15}X_{12}$
	22,73	$X_{15}X_6$
	23,98	$X_{15}X_{13}$
3	12,61	$X_9X_{17}X_{20}$
	15,56	$X_2X_9X_{20}$
	16,12	$X_9X_{15}X_{20}$
	16,29	$X_5X_{10}X_{11}$
	17,24	$X_7X_9X_{20}$
4	11,49	$X_9X_{10}X_{17}X_{20}$
	11.63	$X_5X_9X_{17}X_{20}$
	11,77	$X_8X_9X_{17}X_{20}$
	11,85	$X_5X_{10}X_{11}X_{16}$
	11,97	$X_2X_9X_{10}X_{20}$
5	6,61	$X_1X_2X_6X_{12}X_{15}$
	8,12	$X_9X_{12}X_{14}X_{15}X_{20}$
	8,44	$X_1X_2X_{12}X_{13}X_{15}$
	8,7	$X_1X_{12}X_{15}X_{17}X_{19}$
	8,82	$X_1X_3X_6X_8X_{13}$

Άλλη μια σχετικά γρήγορη διαδικασία που μπορεί να χρησιμοποιηθεί για την εύρεση βέλτιστου γραμμικού μοντέλου είναι η χρήση ενός αλγορίθμου αντικατάστασης μεταβλητών, ξεκινώντας με την επιλογή τυχαίων υποσυνόλων μεταβλητών. Χρησιμοποιώντας εκατό τυχαία υποσύνολα των τριών και τεσσάρων μεταβλητών έχουμε στο Πίνακα 4.4 τα μοντέλα που επιλέχθηκαν καθώς και τις συχνότητες του κάθε υποσυνόλου.

Πίνακας 4.4: *Τιμές του SSE και της συχνότητας εμφάνισης για μοντέλα που επιλέχθηκαν με χρήση του αλγορίθμου αντικατάστασης μεταβλητών, ξεκινώντας με τυχαία υποσύνολα μεταβλητών*

Υποσύνολα 3 μεταβλητών	***SSE***	**Συχνότητα**	**Υποσύνολα 4 μεταβλητών**	***SSE***	**Συχνότητα**
$X_9X_{17}X_{20}$	12,61	41	$X_9X_{10}X_{17}X_{20}$	11,49	41
$X_5X_{10}X_{11}$	16,29	3	$X_5X_{10}X_{11}X_{16}$	11,85	3
$X_{14}X_{15}X_{17}$	19,49	52	$X_{12}X_{14}X_{15}X_{17}$	11,98	52
$X_3X_8X_{11}$	20,34	4	$X_3X_8X_{11}X_{19}$	17,70	4

Το υποσύνολο που εμφανίζει τη μεγαλύτερη συχνότητα κατά την εφαρμογή αυτής της διαδικασίας δεν αποτελεί απαραίτητα και το βέλτιστο, στη περίπτωσή μας όμως τυχαίνει το βέλτιστο υποσύνολο και των τριών αλλά και των τεσσάρων μεταβλητών να παρουσιάζει τη μεγαλύτερη συχνότητα.

Μέχρι τώρα έχουν εξεταστεί τα πέντε καλύτερα υποσύνολα μεταβλητών σε κάθε κατηγορία ανάλογα με τον αριθμό των μεταβλητών (βλέπε Πίνακα 4.3). Παρατηρούμε πως η διαφορά ποιότητας (σε σχέση με το άθροισμα τετραγώνων των υπολοίπων *SSE*) ανάμεσα στα υποσύνολα 3 και 4 μεταβλητών που δίνουν την καλύτερη προσαρμογή είναι πού μικρή.

Κλείνοντας, θα κάνουμε ορισμένες μικρές παρατηρήσεις σχετικά με τα αποτελέσματα που πήραμε παραπάνω. Παρατηρούμε ότι υπάρχει μια μεγάλη πτώση της τιμής του *SSE* καθώς εισάγεται στο μοντέλο η πρώτη μεταβλητή, πιο συγκεκριμένα η $X_{15} = X_2X_3$ ή η $X_{11} = X_1X_2$. Στη συνέχεια παρατηρείται σταδιακή πτώση της βελτίωσης του *SSE* με την εισαγωγή και των νέων μεταβλητών. Αυτό δείχνει διαισθητικά πως το καλύτερο που θα μπορούσε να γίνει είναι να χρησιμοποιήσουμε μια από αυτές τις δύο μεταβλητές για την πρόβλεψη του ύψους των βροχοπτώσεων.

4.3 Πρόβλεψη χρήσης ατμού σε βιομηχανική μονάδα

Τα δεδομένα του παραδείγματος προέρχονται από την βάση δεδομένων STEAM και χρησιμοποιούνται στο βιβλίο των Draper and Smith (1981). Πρόκειται για ένα πείραμα που αναφέρεται στη μηνιαία χρήση ατμού σε μια βιομηχανική μονάδα, η οποία αποτελεί την εξαρτημένη μεταβλητή του πειράματος.

Σε αυτό το πείραμα έχουμε στη διάθεσή μας 9 μεταβλητές οι οποίες επηρεάζουν την μηνιαία χρήση ατμού στη συγκεκριμένη βιομηχανική μονάδα όπως η αποθήκευση πραγματικού λιπαρού οξέως (X_1), η ακατέργαστη γλυκερίνη που παράγεται (X_2), η μέση ταχύτητα ανέμου (σε μέτρα ανά ώρα) (X_3), οι ημερολογιακές ημέρες ανά μήνα (X_4), οι λειτουργικές ημέρες ανά μήνα (X_5), οι ημέρες με θερμοκρασία κάτω των 32 βαθμών κελσίου (X_6), η μέση ατμοσφαιρική θερμοκρασία (X_7), η μέση τετραγωνική ταχύτητα ανέμου (X_8) και ο αριθμός εκκινήσεων (X_9).

Στο Πίνακα 4.5 παρουσιάζονται τα αποτελέσματα από την προσαρμογή των μεθόδων forward selection, backward selection και από τη μέθοδο της διαδοχικής αντικατάστασης μεταβλητών. Γνωρίζουμε πως η μέθοδος backward elimination προσαρμόζεται εφόσον το πλήθος των παρατηρήσεων είναι μεγαλύτερο από το πλήθος των μεταβλητών. Στην περίπτωση αυτή ικανοποιείται η παραπάνω προϋπόθεση με αποτέλεσμα η προσαρμογή της μεθόδου backward elimination να είναι εφικτή.

Οι μέθοδοι που εξετάζουμε στο συγκεκριμένο παράδειγμα έχουν όλες προσαρμοστεί αρκετά καλά και κυρίως η μέθοδος της διαδοχικής αντικατάστασης η οποία κατάφερε να βρει τα καλύτερα υποσύνολα μεταβλητών για κάθε μέγεθος του γραμμικού μοντέλου. Παρατηρούμε πως η τιμή του *SSE* είναι σχεδόν σταθερή για τα γραμμικά μοντέλα που αποτελούνται από τρείς ή τέσσερις μεταβλητές, πράγμα το οποίο δείχνει ότι το βέλτιστο μοντέλο θα αποτελείται το πολύ από τρείς μεταβλητές. Η τιμή του *SSE* για το πλήρες γραμμικό μοντέλο είναι 4,87 και έχει 15 βαθμούς ελευθερίας. Διαιρώντας την τιμή του *SSE* με τους βαθμούς ελευθερίας βρίσκουμε την εκτίμηση της διακύμανσης των καταλοίπων ίση με 0,32, όσο είναι και η διαφορά του *SSE* μεταβαίνοντας από το υποσύνολων των τριών μεταβλητών σε τέσσερις καθώς και από το υποσύνολο των τεσσάρων μεταβλητών σε πέντε.

Πίνακας 4.5: *Τιμές του SSE για τα μοντέλα του παραδείγματος της πρόβλεψης χρήσης ατμού σε βιομηχανική μονάδα (οι αριθμοί στις παρενθέσεις υποδεικνύουν τις επιλεγμένες μεταβλητές του γραμμικού μοντέλου)*

Πλήθος μεταβλητών	Forward selection	Backward elimination	Μέθοδος διαδοχικής αντικατάστασης μεταβλητών
σταθερός όρος	63,82	63,82	63,82
1	18,22	18,22	18,22
	(7)	(7)	(7)
2	8,93	8,93	8,93
	(1,7)	(1,7)	(1,7)
3	7,68	7,68	7,34
	(1,5,7)	(1,5,7)	(4,5,7)
4	6,8	6,93	6,8
	(1,4,5,7)	(1,5,7,9)	(1,4,5,7)
5	6,46	6,54	6,41
	(1,4,5,7,9)	(1,5,7,8,9)	(1,2,5,7,9)

Στο Πίνακα 4.6 παρουσιάζονται τα πέντε καλύτερα υποσύνολα που αποτελούνται από μία, δύο και τρείς μεταβλητές. Παρατηρούμε πως

i. στην περίπτωση χρήσης μιας μεταβλητής για το γραμμικό μοντέλο, η μεταβλητή X_7 που επιλέχτηκε ως η καλύτερη και από τις τρεις μεθόδους έχει μεγάλη διαφορά ως προς την τιμή του *SSE* από οποιαδήποτε άλλη επιλογή μιας μεταβλητής.

ii. στην περίπτωση χρήσης δύο μεταβλητών για το γραμμικό μοντέλο, υπάρχουν τουλάχιστον 3 καλοί συνδυασμοί, σχεδόν ισοδύναμοι ως προς την τιμή του *SSE*.

iii. στην περίπτωση χρήσης τριών μεταβλητών για το γραμμικό μοντέλο, υπάρχουν τουλάχιστον 5 καλοί συνδυασμοί, σχεδόν ισοδύναμοι ως προς την τιμή του *SSE*.

Από μια προσεκτική ανάλυση των αποτελεσμάτων, οι Draper and Smith (1981) διαπίστωσαν ότι, αν προσαρμόσουμε τη μέθοδο ελαχίστων τετραγώνων στις παραμέτρους του γραμμικού μοντέλου θα προκύψει μικρή μεροληψία όσον αφορά τα μοντέλα με μία ή δύο μεταβλητές, ενώ εάν επιλεχθεί μοντέλο με τρείς μεταβλητές η μεροληψία θα είναι μεγαλύτερη.

Πίνακας 4.6: *Τα πέντε καλύτερα υποσύνολα μεταβλητών του παραδείγματος της πρόβλεψης χρήσης ατμού σε βιομηχανική μονάδα*

Πλήθος μεταβλητών	*SSE*	Μεταβλητές
1	18,22	X_7
	37,62	X_6
	45,47	X_5
	49,46	X_3
	53,88	X_8
2	8,93	X_1, X_7
	9,63	X_5, X_7
	9,78	X_2, X_7
	15,6	X_4, X_7
	15,99	X_7, X_9
3	7,34	X_4, X_5, X_7
	7,68	X_1, X_5, X_7
	8,61	X_1, X_7, X_9
	8,69	X_1, X_4, X_7
	8,71	X_5, X_7, X_8

4.4 Πρόβλεψη του αριθμού αυτοκτονιών

Τα δεδομένα του παραδείγματος αυτού προέρχονται από τη βάση δεδομένων DETPOIT και χρησιμοποιήθηκαν από τους Gunst and Mason (1980). Αναφέρονται στον ετήσιο αριθμό αυτοκτονιών για τα έτη 1961 έως 1973 στο Detroit των Ηνωμένων Πολιτειών.

Εδώ έχουμε 13 παρατηρήσεις και 11 διαθέσιμες μεταβλητές με αποτέλεσμα να υπάρχει μόνο ένας βαθμός ελευθερίας για την εκτίμηση των υπολοίπων εάν το μοντέλο περιλαμβάνει και το σταθερό όρο. Στο Πίνακα 4.7 παρουσιάζονται τα αποτελέσματα της εφαρμογής των μεθόδων forward selection, backward elimination και της μεθόδου της διαδοχικής αντικατάστασης μεταβλητών.

Πίνακας 4.7: *Τιμές του SSE για τα μοντέλα του παραδείγματος της πρόβλεψης του αριθμού αυτοκτονιών (οι αριθμοί στις παρενθέσεις υποδεικνύουν τις επιλεγμένες μεταβλητές του γραμμικού μοντέλου)*

Πλήθος μεταβλητών	Forward selection	Backward elimination	Μέθοδος διαδοχικής αντικατάστασης μεταβλητών
σταθερός όρος	3221,8	3222,8	3223,8
1	200	680,4	200
	(6)	(11)	(6)
2	33,83	134	33,83
	(4,6)	(4,11)	(4,6)
3	21,19	23,51	21,19
	(4,6,10)	(3,4,11)	(4,6,10)
4	13,32	10,67	13,32
	(1,4,6,10)	(3,4,8,11)	(1,4,6,10)
5	8,2	8,89	2,62
	(1,2,4,6,10)	(3,4,7,8,11)	(1,2,4,9,11)
6	2,38	6,91	1,37
	(1,2,4,6,10,11)	(3,4,7,8,9,11)	(1,2,4,6,7,11)

Σε αυτό το παράδειγμα καμία από τις «οικονομικές» μεθόδους δεν προσαρμόζεται καλά, ειδικά στην επιλογή των καλύτερων υποσυνόλων που αποτελούνται από τρείς ή τέσσερις μεταβλητές. Για τα μεγαλύτερα υποσύνολα η εφαρμογή της μεθόδου διαδοχικής αντικατάστασης είναι επιτυχής.

Η μέθοδος backward elimination παραλείπει τη μεταβλητή X_6 από τα περισσότερα υποσύνολα που έχει θεωρήσει ως καλύτερα και διατηρεί σε κάποια από αυτά τις μεταβλητές X_3, X_8 και X_{10} έως το τελευταίο στάδιο της εφαρμογής της μεθόδου.

Σύμφωνα με τους Gunst and Mason (1980), οι αποκαλούμενες «οικονομικές» μέθοδοι για την επιλογή μεταβλητών δεν προσαρμόζονται καλά στα δεδομένα μας εάν ο λόγος του πλήθους των παρατηρήσεων προς το πλήθος των μεταβλητών είναι μικρότερο ή ίσο από τη μονάδα. Σε αυτές τις περιπτώσεις τα καλύτερα υποσύνολα p μεταβλητών δεν περιέχουν απαραίτητα τα καλύτερα υποσύνολα p-1 μεταβλητών. Πολλές φορές μάλιστα, τα υποσύνολα αυτά δεν έχουν κοινές μεταβλητές με αποτέλεσμα οι μέθοδοι που προσθέτουν ή αφαιρούν μια μεταβλητή σε κάθε βήμα να μην μπορούν να εντοπίσουν τα βέλτιστα υποσύνολα προβλεπουσών μεταβλητών ή να δυσκολεύονται να φτάσουν σε αυτά.

Ένα αξιοσημείωτο χαρακτηριστικό σε αυτό το σύνολο δεδομένων είναι πως η πρώτη μεταβλητή, X_6, που επιλέγεται από την μέθοδο forward selection για να εισαχθεί στο μοντέλο είναι συγχρόνως η πρώτη μεταβλητή που επιλέγεται από την μέθοδο backward elimination να απορριφθεί από το μοντέλο.

Στο Πίνακα 4.8 παρουσιάζονται τα πέντε καλύτερα υποσύνολα για κάθε πλήθος μεταβλητών. Τα υποσύνολα των τριών μεταβλητών παρουσιάζουν ένα ιδιαίτερο ενδιαφέρον αφού το υποσύνολο που αποτελείται από τις μεταβλητές X_2, X_4, και X_{11} είναι πολύ καλύτερο από κάθε άλλο στην κατηγορία αυτή αλλά κανένα υποσύνολο που να αποτελείται από τις μεταβλητές αυτές (μεμονωμένα) ή σε συνδυασμούς ανά δύο δεν προκύπτει ως το καλύτερο υποσύνολο στη κατηγορία που ανήκει. Η μεταβλητή X_2 έχει την μικτότερη συσχέτιση 0,21, σε απόλυτη τιμή, με την εξαρτημένη μεταβλητή. Η επόμενη μικρότερη τιμή συσχέτισης με την εξαρτημένη μεταβλητή είναι 0,55και οι υπόλοιπες τιμές δεν ξεπερνούν την 0,9.

Οι τιμές του *SSE* για τα υποσύνολα των τριών μεταβλητών παρουσιάζονται στον Πίνακα 4.9.

Πίνακας 4.8: *Τα πέντε καλύτερα υποσύνολα μεταβλητών του παραδείγματος της πρόβλεψης του αριθμού αυτοκτονιών*

Πλήθος μεταβλητών	*SSE*	Μεταβλητές
1	200	X_6
	227,4	X_1
	264,6	X_9
	277,7	X_8
	298,3	X_7
2	33,83	X_4, X_6
	44,77	X_2, X_7
	54,45	X_1, X_9
	55,49	X_5, X_6
	62,46	X_3, X_8
3	6,77	X_2, X_4, X_{11}
	21,19	X_4, X_6, X_{10}
	23,05	X_1, X_4, X_6
	23,51	$X_3, X_4 X_{11}$
	25,04	X_4, X_6, X_{11}
4	3,79	X_2, X_4, X_6, X_{11}
	4,58	X_1, X_2, X_4, X_{11}
	5,24	X_2, X_4, X_7, X_{11}
	5,41	X_2, X_4, X_9, X_{11}
	6,38	X_1, X_4, X_8, X_{11}
5	2,62	$X_1, X_2, X_4, X_9, X_{11}$
	2,64	$X_1, X_2, X_4, X_6, X_{11}$
	2,75	$X_1, X_2, X_4, X_7, X_{11}$
	2,8	$X_2, X_4, X_6, X_7, X_{11}$
	3,12	$X_2, X_4, X_6, X_9, X_{11}$

Πίνακας 4.9: *Τιμές του SSE για τους συνδυασμούς των μεταβλητών X_2, X_4, X_{11} του παραδείγματος της πρόβλεψης του αριθμού αυτοκτονιών*

Μεταβλητές	*SSE*
X_2	3080
X_4	1522
X_{11}	680
X_2, X_4	1158
$X_{2,} X_{11}$	652
$X_{4,} X_{11}$	134

4.5 Πρόβλεψη χρόνου ζωής

Τα δεδομένα που χρησιμοποιούνται στην παρούσα παράγραφο προέρχονται από το σύνολο δεδομένων POLLUTE (βλέπε Gunst and Mason (1980)). Η εξαρτημένη μεταβλητή αντιστοιχεί στον ρυθμό θνησιμότητας (mortality rate) ανά 100000 κατοίκους για άτομα που διαμένουν σε 60 διαφορετικές περιοχές των Ηνωμένων Πολιτειών Αμερικής

Οι ανεξάρτητες μεταβλητές αντιστοιχούν σε κοινωνικοοικονομικά στοιχεία που αφορούν τους κατοίκους της περιοχής καθώς επίσης και σε μετεωρολογικά δεδομένα και σε δεδομένα σχετικά με το ύψος της ρύπανσης της περιοχής.

Ο Πίνακας 4.10 παρουσιάζει τα σύνολα των μεταβλητών που επιλέχθηκαν με εφαρμογή των μεθόδων forward selection, backward elimination και της μεθόδου της διαδοχικής αντικατάστασης μεταβλητών.

Στο παράδειγμα αυτό το πλήθος των παρατηρήσεων είναι αρκετά μεγαλύτερο από το πλήθος των μεταβλητών. Η μέθοδος της διαδοχικής αντικατάστασης κατέληξε στα καλύτερα υποσύνολα μεταβλητών για κάθε πλήθος μεταβλητών ενώ η μέθοδος forward selection δεν προσαρμόζεται καλά μόνο όταν αναζητούμε γραμμικό μοντέλο που αποτελείται από τέσσερις μεταβλητές. Στην περίπτωση αυτή το σύνολο στο οποίο καταλήγει η μέθοδος forward selection είναι το δεύτερο καλύτερο στη κατηγορία των τεσσάρων μεταβλητών.

Πίνακας 4.10: *Τιμές του SSE για τα μοντέλα του παραδείγματος της πρόβλεψης του χρόνου ζωής (οι αριθμοί στις παρενθέσεις υποδεικνύουν τις επιλεγμένες μεταβλητές του γραμμικού μοντέλου)*

Πλήθος μεταβλητών	Forward selection	Backward elimination	Μέθοδος διαδοχικής αντικατάστασης μεταβλητών
σταθερός όρος	228308	228309	228310
1	133695	133695	133695
	(9)	(9)	(9)
2	99841	127803	99841
	(6,9)	(9,12)	(6,9)
3	82389	91777	82389
	(2,6,9)	(9,12,13)	(2,6,9)
4	72250	78009	69154
	(2,6,9,14)	(6,9,12,13)	(1,2,9,14)
5	64634	69136	64634
	(1,2,6,9,14)	(2,6,9,12,13)	(1,2,6,9,14)
6	60539	64712	60539
	(1,2,3,6,9,14)	(2,5,6,9,12,13)	(1,2,3,6,9,14)

Τέλος η μέθοδος backward elimination δεν καταφέρνει να προσαρμοστεί καλά δεδομένου ότι έχει απορρίψει τη μεταβλητή X_{14} από το υποσύνολο που αποτελείται από δέκα μεταβλητές καθώς και από κάθε άλλο μικρότερο υποσύνολο.

Σε αυτή τη περίπτωση έχουμε μια πολύ καλή εκτίμηση της διακύμανσης των υπολοίπων. Η τιμή του *SSE* για το πλήρες μοντέλο είναι 53680 με 44 βαθμούς ελευθερίας οπότε η τιμή που προκύπτει για την εκτίμηση της διακύμανσης των υπολοίπων είναι 1220.

Η μείωση της τιμής του *SSE*, μεταβαίνοντας από το υποσύνολο των έξι μεταβλητών στο υποσύνολο των επτά μεταβλητών, είναι μικρότερη από το διπλάσιο της διακύμανσης των υπολοίπων γεγονός που υποδεικνύει πως το καλύτερο υποσύνολο για πρόβλεψη δε πρέπει να ξεπερνάει τις έξι μεταβλητές. Η μείωση της τιμής του *SSE*, μεταβαίνοντας από το υποσύνολο των πέντε μεταβλητών στο υποσύνολο των έξι μεταβλητών, δεν είναι σημαντική αφού είναι μικρότερη από το τετραπλάσιο της διακύμανσης των υπολοίπων.

Στο Πίνακα 4.11 παρουσιάζονται τα πέντε καλύτερα μοντέλα για κάθε κατηγορία μεταβλητών. Μια παρατήρηση για το παράδειγμα αυτό είναι ότι οι τιμές του *SSE* για τα υποσύνολα του ίδιου πλήθους μεταβλητών φαίνονται να είναι αρκετά κοντά μεταξύ τους κάτι που θα μπορούσε να οδηγήσει στο συμπέρασμα ότι δεν είναι σαφής η υπεροχή κάποιου γραμμικού μοντέλου έναντι των υπολοίπων. Ωστόσο αν συγκρίνουμε τις διαφορές των τιμών του *SSE* με την εκτίμηση της διακύμανσης των υπολοίπων (η οποία βρέθηκε παραπάνω ότι είναι ίση με 1220) θα οδηγηθούμε στο συμπέρασμα ότι μόνο ένα μικρός αριθμός υποσυνόλων έχουν σχετικά κοντινές τιμές για το *SSE*.

Είναι αξιοσημείωτο ότι στα καλύτερα σύνολα κάθε κατηγορίας που δίνονται στον Πίνακα 4.11, συμμετέχει συνεχώς η μεταβλητή X_9 η οποία αντιστοιχεί στο ποσοστό του πληθυσμού οι οποίοι είναι μη λευκοί (έγχρωμοι).

Κάποιες άλλες μεταβλητές που εμφανίζονται αρκετές φορές στα βέλτιστα υποσύνολα μεταβλητών είναι οι ετήσιες βροχοπτώσεις (X_1), η μέση θερμοκρασία τον μήνα Ιανουάριο (X_1), η μέση θερμοκρασία τον μήνα Ιούλιο (X_{13}), η διάμεσος των ετών εκπαίδευσης (X_6) καθώς και η συγκέντρωση του διοξειδίου του θείου (X_{14}).

Πίνακας 4.11: *Τα πέντε καλύτερα υποσύνολα μεταβλητών του παραδείγματος της πρόβλεψης του χρόνου ζωής*

Πλήθος μεταβλητών	*SSE*	Μεταβλητές
1	133695	X_9
	168696	X_6
	169041	X_1
	186716	X_7
	186896	X_{14}
2	99841	X_6, X_9
	103859	X_2, X_9
	109203	X_9, X_{14}
	112259	X_4, X_9
	115541	X_9, X_{10}
3	82389	X_2, X_6, X_9
	83335	X_1, X_9, X_{14}
	85242	X_6, X_9, X_{14}
	88543	X_2, X_9, X_{14}
	88920	X_6, X_9, X_{11}
4	69154	X_1, X_2, X_9, X_{14}
	72250	X_2, X_6, X_9, X_{14}
	74666	X_2, X_5, X_6, X_9
	76230	X_2, X_6, X_8, X_9
	76276	X_1, X_6, X_9, X_{14}
5	64634	$X_1, X_2, X_6, X_9, X_{14}$
	65660	$X_1, X_2, X_3, X_9, X_{14}$
	66555	$X_1, X_2, X_8, X_9, X_{14}$
	66837	$X_1, X_2, X_9, X_{10}, X_{14}$
	67622	$X_2, X_4, X_6, X_9, X_{14}$

Για να διερευνηθεί καλύτερα η σημασία της μεταβλητής X_9 προσδιορίστηκαν τα καλύτερα μοντέλα που προκύπτουν χωρίς χρήσης της μεταβλητή X_9. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 4.12 από όπου καθίσταται φανερό ότι όλα είναι κατώτερης ποιότητας από τα αντίστοιχα μοντέλα που περιέχουν τη μεταβλητή X_9..

Πίνακας 4.12: *Τα καλύτερα υποσύνολα ανά κατηγορία χωρίς την μεταβλητή X_9 του παραδείγματος της πρόβλεψης του χρόνου ζωής*

Μεταβλητές	*SSE*
X_6	168696
X_1, X_{14}	115749
X_1, X_4, X_{14}	102479
X_1, X_4, X_7, X_{14}	92370
$X_1, X_2, X_4, X_{11}, X_{14}$	87440
$X_1, X_2, X_3, X_4, X_{11}, X_{14}$	81846

ΣΥΜΠΕΡΑΣΜΑΤΑ

Στα πλαίσια της παρούσας διπλωματικής εργασίας έγινε παρουσίαση του θεωρητικού υποβάθρου, των μεθοδολογιών και την εφαρμογής αυτών στο πρόβλημα της εύρεσης του βέλτιστου γραμμικού μοντέλου όταν διαθέτουμε μεγάλο αριθμό ανεξάρτητων μεταβλητών και επιθυμούμε να προβλέψουμε μια εξαρτημένη με τη μέθοδο της παλινδρόμησης.

Παρουσιάστηκε αναλυτικά πώς εφαρμόζεται η μέθοδος «εξέτασης όλων των δυνατών μοντέλων» σε ένα τυχαίο δείγμα 30 μεταβλητών οι οποίες προέρχονται από μια βάση με δεδομένα κατανάλωσης διαφόρων τύπων αυτοκινήτων με διαφορετικές παραμέτρους λειτουργίας. Πιο συγκεκριμένα έγινε εφαρμογή με τα πιο συνηθισμένα κριτήρια που έχουν προταθεί στη διεθνή βιβλιογραφία για την υλοποίηση της μεθόδου, δηλαδή το κριτήριο R^2, το κριτήριο *SSE*, το κριτήριο R_{adj}^{p}, το κριτήριο *MSE*, το κριτήριο C_p του Mallows, το κριτήριο $PRESS_p$, το κριτήριο *AIC*, το κριτήριο *BIC* και το κριτήριο S_p.

Σε μια περαιτέρω προσέγγιση του προβλήματος, παρουσιάστηκαν και εφαρμόστηκαν οι κυριότερες επαναληπτικές μέθοδοι με τις οποίες δημιουργούνται αυτόματα υποσύνολα ανεξάρτητων μεταβλητών που είναι βέλτιστα ή βρίσκονται «κοντά» στο βέλτιστο μοντέλο. Ειδικότερα, εφαρμόστηκαν η μέθοδος Forward Ranking, η μέθοδος Backward Ranking, η μέθοδος Stepwise Regression, η μέθοδος Backward Elimination, η μέθοδος Forward Selection, η μέθοδος Forward Procedure, η μέθοδος της διαδοχικής αντικατάστασης μεταβλητών και η μέθοδος της αντικατάστασης δύο μεταβλητών και παρουσιάστηκαν τα αντίστοιχα αποτελέσματα.

Τέλος μέσω συγκεκριμένων πραγματικών δεδομένων έγινε σύγκριση των διαδικασιών αυτόματης επιλογής (επαναληπτικών μεθόδων) ως προς τη δυνατότητα ανίχνευσης του βέλτιστου μοντέλου. Τα αποτελέσματα που προέκυψαν από την προσαρμογή των μεθόδων στα πραγματικά δεδομένα, όπως ήταν αναμενόμενο, δεν ήταν τα ίδια. Δε θα μπορούσαμε όμως με βεβαιότητα να πούμε πιο από τα γραμμικά μοντέλα που έχουν προταθεί από κάθε μία μέθοδο είναι το καλύτερο. Η βασική διαφορά των δύο μεθόδων, που τις καθιστά και μη συγκρίσιμες, είναι πως με την μέθοδο «εξέτασης όλων των δυνατών μοντέλων» ο ερευνητής είναι σε θέση να

εξετάσει όλα τα γραμμικά μοντέλα που προκύπτουν και αφού ορίσει κάποια από αυτά ως «καλά» να καταλήξει στο βέλτιστο. Από την άλλη μεριά οι επαναληπτικές μέθοδοι εφαρμόζουν μια ακολουθία βημάτων, εισάγοντας ή διαγράφοντας μεταβλητές, με σκοπό την εύρεση του βέλτιστου μοντέλου. Με βάση τις μεθόδους αυτές ο ερευνητής δεν γνωρίζει τα ενδιάμεσα στάδια έως την επιλογή του βέλτιστου, επομένως δεν είναι σε θέση να διακρίνει μια λανθασμένη ή μια «φτωχή» επιλογή μοντέλου για την πρόβλεψη την εξαρτημένης μεταβλητής.

Βιβλιογραφία

[1] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19,** 716–723.

[2] Allen, D. M. (1971). The prediction sum of squares as a criterion for selection predictor variables, *Technical Report* No.23, Department of Statistics, University of Kentuky.

[3] Berk, K.N. (1978b), *Comparing subset regression procedures*, *Technometrics*, **20**, 1-6.

[4] Biondini, R., Simpson, J. and Woodley, W.(1977), *Empirical predictors for natural and seeded rainfalls in the Florida Area Cumulus Experiment (FACE)*, 1970-1975, *J. Appl. Meteor.*, **16**, 585-594.

[5] Draper, N. R. and Smith, H. (1981), *Applied Regression Analysis*, 2nd Edition, Wiley, New York.

[6] Gunst, R.F. and Mason, R.L (1980), *Regression analysis and its application*, Marcel Dekker, New York.

[7] Hocking, R. R. and Leslie, R. N. (1967). Selection of the best Subset in regression analysis, *Technometrics,* **9**, 531-540.

[8] Kennard, R. W. (1971). A note on the C_p Statistic. *Technometrics,* **13**, 899-900, (corrections: Kennard, R. W. (1973), *Technometrics*, **15**, 657).

[9] Kim, S-S. (1998). All possible subset regressions using the triangular decomposition, *Journal of Statistical Computation and Simulation*, **65**, 81-94.

[10] Mallows, C. L. (1973). Some comments on C_p, *Technometrics*, **37**, 362-372.

[11] Miller, A.J (2002), *Selection of subsets of regression variables*, 2nd Edition, Chapman and Hall, New York.

[12] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics,* **6**, 461-466.

[13] Thompson, M. L. (1978). Selection of variables in multiple regression, part I: a review and evaluation, *International Statistical Review*, **46**, 1-19.

[14] Thompson, M. L. (1978). Selection of variables in multiple regression, part II: Chosen Procedures, Computations and Examples, *International Statistical Review*, **46**, 129-146.

[15] Weakliem, D. L. (1999). A critique of the Bayesian information criterion for model selection, *Sociological Methods & Research*, **27**, 359-397.