ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΓΙΑ ΔΕΔΟΜΕΝΑ ΑΣΦΑΛΙΣΤΙΚΩΝ ΑΠΟΖΗΜΙΩΣΕΩΝ

Γεώργιος Χριστοδουλάκης

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

Σεπτέμβριος 2018

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σ χ ο λ ή Χ ρ η μ α τ ο ο ι κ ο ν ο μ ι κ ή ς κ α ι Σ τ α τ ι σ τ ι κ ή ς

Τ μ ή μ α Σ τ α τ ι σ τ ι κ ή ς κ α ι Α σ φ α λ ι σ τ ι κ ή ς Ε π ι σ τ ή μ η ς

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΓΙΑ ΔΕΔΟΜΕΝΑ ΑΣΦΑΛΙΣΤΙΚΩΝ ΑΠΟΖΗΜΙΩΣΕΩΝ

Γεώργιος Χριστοδουλάκης

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

Σεπτέμβριος 2018

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 υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Κ. Πολίτης (Επιβλέπων)
- Χ. Ευαγελάρας
- Γ. Βεροπούλου

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

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

Linear regression models for insurance data

By

George Christodoulakis

MSc Dissertation

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

Piraeus, Greece

September 2018

Στους γονείς μου

Πέτρο και Χρυσάνθη

Ευχαριστίες

Θα ήθελα να ευχαριστήσω όσους συντέλεσαν στην ολοκλήρωση της παρούσας Διπλωματικής εργασίας. Αρχικά θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα αναπληρωτή καθηγητή κ. Πολίτη Κωνσταντίνο για την γνώση, την καθοδήγηση και τις πολύτιμες συμβουλές που μου προσέφερε καθ'όλη τη διάρκεια της συγγραφής της Διπλωματικής εργασίας. Επίσης θα ήθελα να ευχαριστήσω τα μέλη της τριμελούς επιτροπής επίκουρο καθηγητή κ. Χ. Ευαγγελάρα και αναπληρώτρια καθηγήτρια κα Γ. Βεροπούλου για τη συμμετοχή τους στην εξεταστική επιτροπή της παρούσας Διπλωματικής εργασίας.

Τέλος, θα ήθελα να ευχαριστήσω θερμά τα μέλη της οικογένειας μου που με στηρίζουν σε όλα τα χρόνια της ακαδημαϊκής μου πορείας, τους φίλους και τους συμφοιτητές μου για την ηθική συμπαράστασή τους.

Περίληψη

Σε πολλές στατιστικές εφαρμογές, συναντάμε το πρόβλημα της μελέτης της σχέσης δύο ή περισσότερων τυχαίων μεταβλητών. Η σχέση αυτή μπορεί να μελετηθεί με κατάλληλα μοντέλα παλινδρόμησης, τα οποία χρησιμοποιούνται ευρέως σήμερα στην διοίκηση των επιχειρήσεων, στην οικονομία , στην υγεία, στο ασφαλιστικό κλάδο, στη βιολογία και τις κοινωνικές επιστήμες. Στη στατιστική, η ανάλυση παλινδρόμησης είναι μια διαδικασία για την ποσοτική εκτίμηση και μελέτη της σχέσης μεταξύ διαφόρων μεταβλητών. Περιέχει πολλές τεχνικές για την μοντελοποίηση και την ανάλυση των μεταβλητών αυτών, ενώ συνήθως επικεντρώνεται στη γραμμική σχέση μεταξύ μιας εξαρτημένης μεταβλητής και μίας ή περισσότερων ανεξάρτητων μεταβλητών.

Στη συνέχεια, στην εκτός από την περιγραφή της παραπάνω μεθόδου, στην παρούσα διπλωματική εργασία παρουσιάζονται ορισμοί και τεχνικές της θεωρίας κινδύνου για την μελέτη ασφαλιστικών αποζημιώσεων, καθώς και για την επιλογή του σημείου αντασφάλισης που πρέπει να επιλέξει μια ασφαλιστική εταιρεία. Η συνεχής και σωστή αξιολόγηση του κινδύνου στον ασφαλιστικό κλάδο, ώστε να υπάρχουν επαρκή αποθέματα, έχει τις ρίζες της στην δεκαετία του 1970, ενώ αναπροσδιορίστηκε στα μέσα της δεκαετίας του 1990.

Abstract

In many statistical applications, we encounter the problem of studying the relationship of two or more random variables. This relationship can be studied with appropriate regression models, which are widely used today in business administration, economy, health, insurance, biology and social sciences. In statistics, regression analysis is a process for quantifying and studying the relationship between different variables. It contains many techniques for the modeling and analysis of these variables, while it usually focuses on the linear relationship between a dependent variable and one or more independent variables.

Then, besides the description of the above method, this diploma thesis presents definitions and techniques of risk theory for the study of insurance indemnities, as well as analyses the cut-off point of reinsurance that an insurance company could choose, in order to «protect its funds». Continuous and accurate risk assessment in the insurance industry to ensure that there are sufficient reserves, is rooted in the 1970s and redefined in the mid-1990s.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Σκοπός της εργασίας

Η συλλογή και η ανάλυση των δεδομένων ανέκαθεν αποτελούσαν σημαντικό στοιχείο των επιχειρήσεων. Στις μέρες μας όμως, αποτελούν τον πυρήνα τους. Με την ανάπτυξη νέων τεχνολογιών, όπως τα μέσα κοινωνικής δικτύωσης, οι φορητές συσκευές, η συλλογή και η δυνατότητα για επεξεργασία μεγάλων (big data)

δεδομένων, η ανάπτυξη μεθόδων machine learning, τα δεδομένα απέκτησαν μία εντελώς νέα διάσταση. Οι μεγάλες εταιρείες, όπως για παράδειγμα οι ασφαλιστικές, είναι σε θέση με τις κατάλληλες τεχνικές ανάλυσης των δεδομένων που συλλέγουν, να αποφασίσουν σωστότερα, να προβλέψουν κινδύνους και να βαδίσουν με μεγαλύτερη ασφάλεια σε ένα δύσκολο χώρο όπως αυτός του ασφαλιστικού κλάδου. Οι ασφαλιστικές εταιρείες, όπως όλες οι επιχειρήσεις, για να αναπτυχθούν έχουν ανάγκη να νιώθουν ασφαλείς, απέναντι σε επικείμενες δυσμενείς αλλαγές που μπορεί να πραγματοποιηθούν. Ένας τρόπος για να αντισταθμίσουν τους κινδύνους που διατρέχουν ιδιαίτερα τα άτομα αλλά και οι επιχειρήσεις, είναι η αγορά ασφαλιστικής κάλυψης ή αντασφαλιστικής αντίστοιχα αν μιλάμε για ασφαλιστικές εταιρείες. Οι ασφαλιστικές και οι αντασφαλιστικές εταιρείες και πιο συγκεκριμένα τα αναλογιστικά τμήματά τους, μελετούν κατά κύριο λόγο τις αποζημιώσεις που μπορεί να προκύψουν. Με αυτό τον τρόπο θα είναι σε θέση να είναι συνεπείς προς τις υποχρεώσεις που έχουν αναλάβει έναντι των ασφαλισμένων τους. Η διαδικασία αξιολόγησης των κινδύνων από τη μεριά των ασφαλιστικών εταιρειών, αποτελεί ζήτημα ύψιστης σημασίας, και προϋποθέτει την σωστή τιμολόγηση των κινδύνων που αναλαμβάνουν καθώς και αποτελεσματική διαχείριση των αποθεμάτων, έτσι ώστε να είναι φερέγγυες απέναντι στους ασφαλισμένους και συνεπείς στις εκτιμήσεις που ζητούν οι εποπτικές αρχές.

1.2 Διάρθρωση της εργασίας

Η παρούσα διπλωματική εργασία καλείται να περιγράψει και να εφαρμόσει μεθόδους μοντέλων γραμμικής παλινδρόμησης καθώς επίσης, με την βοήθεια της λογιστικής παλινδρόμησης να μελετήσει τις ζημιές της ασφαλιστικής επιχείρησης που θα προέλθουν από διεκδικήσεις των ασφαλισμένων. Η ανάλυση θα πραγματοποιηθεί σε δύο σετ δεδομένων. Συγκεκριμένα, στο δεύτερο κεφάλαιο της εργασίας θα γίνει

περιγραφή και ανάπτυξη της θεωρίας της γραμμικής παλινδρόμησης, αναπτύσσοντας τις προϋποθέσεις αλλά και κάποιες μεθόδους προκειμένου να μπορεί να εφαρμοστεί. Το τρίτο κεφάλαιο, περιγράφει την μέθοδο της λογιστικής παλινδρόμησης η οποία χρησιμοποιείται ευρέως για την μελέτη των αποζημιώσεων. Στο κεφάλαιο αυτό γίνεται επίσης εισαγωγή στην έννοια της αντασφάλισης χαρτοφυλακίου όπου περιγράφονται οι διάφοροι τύποι και οι μορφές αντασφάλισης που επικρατούν. Τέλος, στο κεφάλαιο 4 θα γίνει προσπάθεια να εφαρμοστούν, τα όσα θεωρητικά αναπτύχθηκαν στα κεφάλαια 2 και 3. Συγκεκριμένα, στην ενότητα 4.3, με εφαρμογή μεθόδων παλινδρόμησης, γίνεται ανάλυση σε πραγματικά δεδομένα, που αφορά στην κατανάλωση του καυσίμου των οχημάτων, και αναπτύσσεται το γραμμικό μοντέλο παλινδρόμησης για την πρόβλεψη αυτής. Τέλος, στην ενότητα 4.4.3 εφαρμόζοντας την αντασφάλιση υπερβάλλοντος ζημίας, πραγματοποιείται ανάλυση για την εύρεση του ποσού των αποζημιώσεων πάνω από το οποίο η ασφαλιστική εταιρεία έχει συμφέρον να αντασφαλιστεί.

Κεφάλαιο 2

Εισαγωγή στην Γραμμική Παλινδρόμηση

Η παλινδρόμηση είναι η μελέτη της σχέσης μεταξύ δύο ή περισσότερων τυχαίων μεταβλητών. Σε κάποιες περιπτώσεις μπορεί να ερμηνεύσει κανείς πώς μεταβάλλεται μία μεταβλητή (απόκρισης) καθώς οι τιμές μιας άλλης (προβλέπουσας) μεταβλητής αλλάζουν. Σε πολλές περιπτώσεις κρίνεται απαραίτητη η ανάλυση δεδομένων προκειμένου να προκύψουν απαντήσεις στα ερωτήματα που μας ενδιαφέρουν. Στην αναλογιστική επιστήμη, η γραμμική παλινδρόμηση είναι χρήσιμη για την απάντηση ερωτημάτων όπως: σε μία ασφαλιστική εταιρία επηρεάζεται το μέγεθος της αποζημίωσης του ασφαλισμένου από την ηλικία ή το φύλο του οδηγού; Το είδος του αυτοκινήτου επηρεάζει το μέγεθος της ασφαλιστικής αποζημίωσης; Η

ηλικία του οδηγού επηρεάζει την πιθανότητα ατυχήματος και το μέγεθος της ασφαλιστικής αποζημίωσης;

Επομένως συνήθως υπάρχει μία μεταβλητή η οποία πρέπει να μελετηθεί (π.χ ασφαλιστικές αποζημιώσεις, πλήθος αποζημιώσεων) και αρκετές μεταβλητές οι οποίες ενδεχομένως να ερμηνεύουν αυτήν την μεταβλητή. Στην περίπτωση που η μεταβλητή αυτή συνδέεται γραμμικά (ή μη) με τις μεταβλητές που την επεξηγούν, τότε μπορεί να εφαρμοστούν διάφορες μέθοδοι όπως είναι αυτή της γραμμικής παλινδρόμησης.

2.1 Απλή γραμμική παλινδρόμηση

Στην ενότητα αυτή παρουσιάζεται το απλούστερο μοντέλο παλινδρόμησης, αυτό της απλής γραμμικής παλινδρόμησης, με τη βοήθεια ενός παραδείγματος από την ασφαλιστική επιστήμη και ιδιαίτερα από το χώρο της ασφάλισης οχημάτων.

Ο **Error! Unknown switch argument.** έχει δύο στήλες. Πρώτης στήλη (Y) αφορά στο μέγεθος της ασφαλιστικής αποζημίωσης και η δεύτερη στήλη (X) είναι το βάρος του οχήματος.

Η ασφαλιστική εταιρία θα ήθελε να γνωρίζει τι θα συμβεί στο μέγεθος των ασφαλιστικών αποζημιώσεων Y, ανάλογα με το βάρος του οχήματος.

Με τη βοήθεια της μεθόδου παλινδρόμησης, υπό τις κατάλληλες προϋποθέσεις, η παραπάνω πρόβλεψη είναι εφικτή, θεωρώντας ότι η σχέση μεταξύ των μεταβλητών X και Y είναι γραμμική.

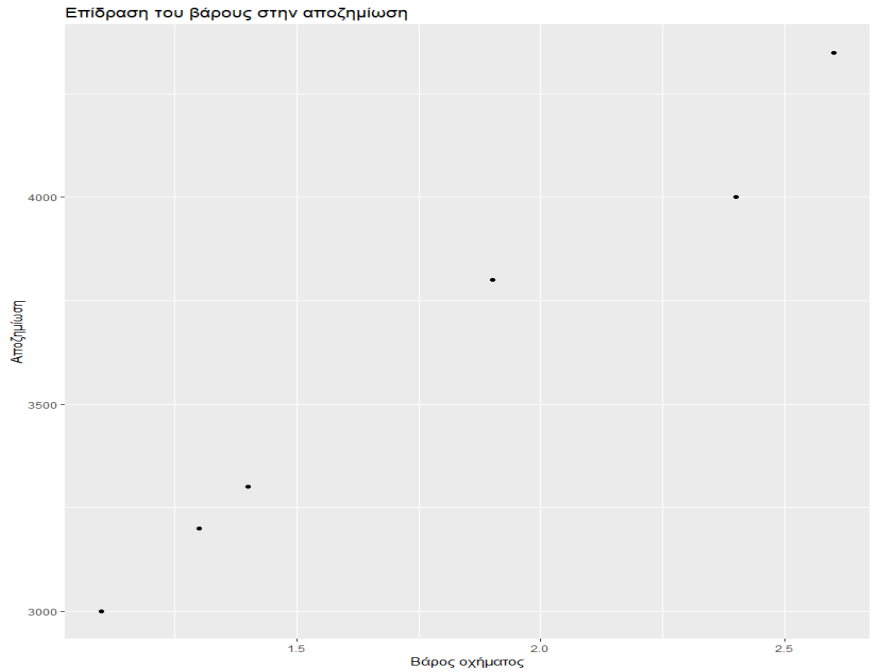
Το απαραίτητο πρώτο βήμα πριν γίνει χρήση της μεθόδου παλινδρόμησης είναι η δημιουργία ενός διαγράμματος διασποράς προκειμένου να αποτυπωθεί μία εικόνα των δεδομένων και να αξιολογηθεί αν η μέθοδος της παλινδρόμησης κρίνεται επαρκής. Κάνοντας ένα διάγραμμα διασποράς (όπως είναι το **Error! Unknown switch**

argument.) των τιμών X και Y του παρακάτω πίνακα προκύπτει το ακόλουθο γράφημα.

Αποζημίωση	**Βάρος αυτοκινήτου (tn)**
3000	1.1
3200	1.3
3300	1.4
4000	2.4
3800	1.9
4350	2.6

Πίνακας 2.1: Παράδειγμα

Στο επόμενο διάγραμμα, στον άξονα των τετμημένων βρίσκεται το βάρος του οχήματος X και στον άξονα των τεταγμένων τα αντίστοιχα ποσά ασφαλιστικών αποζημιώσεων. Σκοπός της παλινδρόμησης είναι να αντιληφθούμε πώς αλλάζουν οι τιμές της μεταβλητής Y καθώς οι τιμές της μεταβλητής X μεταβάλλονται. Το διάγραμμα διασποράς μπορεί να δώσει μια γρήγορη πρώτη εντύπωση.

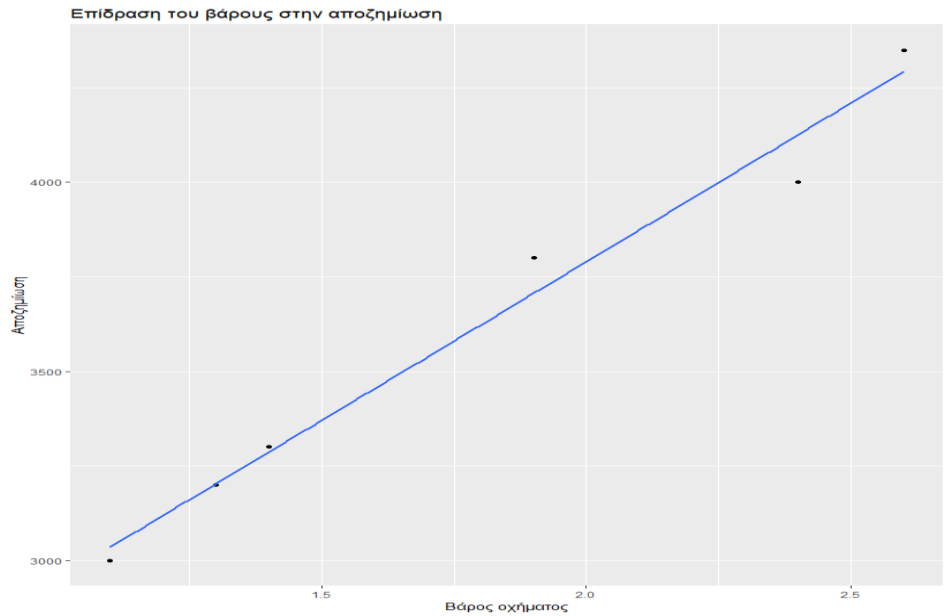


Γράφημα 2.1: Διάγραμμα διασποράς

Όπως προκύπτει από το γράφημα αυτό μάλλον υπάρχει γραμμική σχέση ανάμεσα στη μεταβλητή του βάρους και της αποζημίωσης.

Με την μέθοδο γραμμικής παλινδρόμησης η ευθεία της μορφής

$Y = \beta_0 + \beta_1 x$ που προκύπτει, φαίνεται στο **Error! Unknown switch argument.**.2 .



Γράφημα 2.2: Διάγραμμα διασποράς και ευθεία ελαχίστων τετραγώνων

2.2 Υποθέσεις παλινδρόμησης

Το μοντέλο στο οποίο έγινε αναφορά στην προηγούμενη ενότητα έχει την γενική μορφή (Κούτρας 2015)

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \textbf{(1.1)}$$

και αποτελεί ένα στοχαστικό μοντέλο για το οποίο θα γίνουν οι παρακάτω υποθέσεις:

- Οι ποσότητες β_0 και β_1 είναι άγνωστες παράμετροι.
- Όπως αναφέρθηκε παραπάνω, η μεταβλητή ε_iτου μοντέλου $Y = \beta_0 + \beta_1 X + \varepsilon_i$ είναι τυχαία μεταβλητή και άρα η εξαρτημένη μεταβλητή Y θα είναι και αυτή τυχαία μεταβλητή. Οι τιμές της εξαρτημένης μεταβλητής κατά την i επανάληψη θα συμβολίζονται με y_i.
- Η προβλέπουσα μεταβλητή X είναι γνωστό διάνυσμα τιμών (ή πίνακας τιμών) και οι τιμές που μπορεί να λάβει έχουν προκύψει από κάποια δειγματοληψία. Οι τιμές της ανεξάρτητης μεταβλητής κατά την i επανάληψη συμβολίζονται με x_i.
- Τα τυχαία σφάλματα ε_i θα έχουν μέση τιμή $E[\varepsilon_i] = 0$ και διασπορά $V[\varepsilon_i] = \sigma^2$.
- Τέλος, τα σφάλματα τα οποία αντιστοιχούν σε διαφορετικές επαναλήψεις θα είναι μεταξύ τους ασυσχέτιστα και άρα θα ισχύει

$$\mathrm{Cov}[\varepsilon_i, \varepsilon_j] = \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases}.$$

Επομένως, το απλό γραμμικό μοντέλο περιλαμβάνει τις δύο παρακάτω σχέσεις της αναμενόμενης τιμής E[Y] και της διασποράς V[Y] της μεταβλητής Y:

$$E[Y_i] = E[\beta_0 + \beta_1 x_i + \varepsilon_i] = \beta_0 + \beta_1 X_i ,$$

$$V[Y_i] = \sigma^2$$

Επειδή η διασπορά σ^2 παίρνει θετικές τιμές, η παρατηρούμενη τιμή της i-οστής παρατήρησης y_i της εξαρτημένης μεταβλητής Y θα διαφέρει από την αναμενόμενη τιμή E[Y] της παραπάνω σχέσης. Για να ληφθεί υπόψη αυτή η διαφορά μεταξύ της αναμενόμενης τιμής και των παρατηρούμενων δεδομένων, χρησιμοποιείται μια ποσότητα που ονομάζεται κατάλοιπο του υποδείγματος και συμβολίζεται με $\hat{\varepsilon}_i$.

Τα κατάλοιπα γραφικά, είναι οι κατακόρυφες αποστάσεις της ευθείας από τα ζεύγη (X_t,Y_t) και η πιο σημαντική ιδιότητα της ευθείας παλινδρόμησης είναι ότι το άθροισμα των τετραγώνων των καταλοίπων έχει ελαχιστοποιηθεί.

Με τη βοήθεια της παραπάνω ιδιότητας προκύπτει ο κατάλληλος προσδιορισμός των παραμέτρων β_0 και β_1, έτσι ώστε να ελαχιστοποιούνται οι κατακόρυφες αποστάσεις

$\varepsilon_i = y_i - (\beta_0 + \beta_1 x)$

Για τον καθορισμό των παραμέτρων της βέλτιστης ευθείας παλινδρόμησης η μέθοδος που έχει επικρατήσει είναι η ελαχιστοποίηση της ποσότητας

$$\sum_i^{\nu} \varepsilon_i^2 = \sum_i^{\nu} [y_i - (\beta_0 + \beta_1 x_i)]^2.$$

Η ποσότητα $\sum_i^{\nu} \varepsilon_i^2$ είναι το άθροισμα τετραγώνων των σφαλμάτων (SSE), με την ελαχιστοποίηση της οποίας προκύπτουν οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0$ και $\hat{\beta}_1$, των παραμέτρων β_0 και β_1 αντίστοιχα.

Τα σφάλματα ε_i είναι μη παρατηρούμενες ποσότητες και επομένως αυτό που στην πράξη μπορούμε να ελαχιστοποιήσουμε με την βοήθεια των παραπάνω εκτιμήσεων είναι οι εκτιμήσεις των σφαλμάτων, που ονομάζονται κατάλοιπα όπως ήδη αναφέραμε,

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

Η παραπάνω μέθοδος είναι γνωστή ως μέθοδος ελαχίστων τετραγώνων και εμφανίστηκε για πρώτη φορά σε έντυπη μορφή από τον Legendre το 1805 (wikipedia.org).

Με την εκτίμηση των παραμέτρων β_0 και β_1 είναι δυνατό να υπολογιστεί η προσαρμοσμένη τιμή y_i, όταν η μεταβλητή X παίρνει την τιμή x_i. Η προσαρμοσμένη τιμή y_i ισούται με $\hat{y}_i = E[\hat{Y}] = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Οι εκτιμήτριες ελαχίστων τετραγώνων $\hat{\beta}_0$ και $\hat{\beta}_1$ είναι αμερόληπτες εκτιμήτριες και έχουν την μικρότερη διακύμανση από όλες τις άλλες αμερόληπτες

εκτιμήτριες των παραμέτρων β_0 και β_1 της σχέσης (1.1), με την προϋπόθεση ότι υπάρχει γραμμική σχέση ανάμεσα στη μεταβλητή απόκρισης και τις προβλέπουσες μεταβλητές (Κούτρας 2015). Αυτό είναι γνωστό στην διεθνή βιβλιογραφία με τον όρο BLUE που σημαίνει Best Linear Unbiased Estimators.

2.3 Συντελεστής προσδιορισμού

Είναι φανερό ότι για να έχει το παραπάνω μοντέλο καλή προβλεπτική ικανότητα θα πρέπει η ποσότητα SSE = $\sum_i^v \varepsilon_i^2$ να παίρνει μικρές τιμές, μιας και αποτελεί το άθροισμα των τετραγώνων των καταλοίπων και άρα το σύνολο των λανθασμένων προβλέψεων του μοντέλου.

Η ποσότητα ε_i είναι τυχαία μεταβλητή, καθώς οφείλεται σε τυχαίους παράγοντες οι οποίοι δεν μπορούν να προβλεφθούν. Επομένως το SSE είναι το άθροισμα των τετραγώνων των καταλοίπων που οφείλεται σε τυχαίους παράγοντες και δεν μπορεί να ερμηνευθεί.

Υπάρχει όμως και η μεταβλητότητα του μοντέλου η οποία μπορεί να ερμηνευθεί και συμβολίζεται με SSR. Το άθροισμα SSR+SSE δίνει την συνολική μεταβλητότητα του μοντέλου η οποία συμβολίζεται με SSTOπου σημαίνει ότι

$$SSTO = SSR+SSE.$$

Με βάση την παραπάνω ανάλυση, ένα κριτήριο για το κατά πόσο η ευθεία $Y=\beta_0 +\beta_1 X+\varepsilon_i$ που έχει παραχθεί με τη μέθοδο ελαχίστων τετραγώνων έχει ικανοποιητική προβλεπτική ικανότητα, είναι το μέγεθος του αθροίσματος τετραγώνων SSE. Ένα αντικειμενικό κριτήριο για τον προσδιορισμό της ποιότητας

του μοντέλου μας, είναι ο δείκτης R^2 ο οποίος ονομάζεται συντελεστής προσδιορισμού και δίνεται από τον τύπο

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

Ισχύει ότι $0 \leq R^2 \leq 1$ και όσο πιο κοντά βρίσκεται αυτός ο δείκτης στη μονάδα τόσο πιο καλή πρόβλεψη προκύπτει από το μοντέλο. Αντίθετα, αν ο δείκτης βρίσκεται κοντά στο μηδέν, το μοντέλο δίνει εκτιμήσεις με μεγάλα σφάλματα και συνεπώς έχει χαμηλή προβλεπτική ικανότητα (Draperand Smith, 1998).

2.4 Κανονικό γραμμικό μοντέλο

Το στοχαστικό μοντέλο της μορφής (1.1), για το οποίο ισχύουν οι παραπάνω πέντε υποθέσεις, με την επιπλέον υπόθεση ότι τα σφάλματα ε_i ακολουθούν την κανονική κατανομή $N(0,\sigma^2)$ ονομάζεται κανονικό γραμμικό μοντέλο. Στην περίπτωση που τα σφάλματα δεν ακολουθούν την κανονική κατανομή, αλλά ακολουθούν την κατανομή Poisson ή την Διωνυμική κατανομή, άλλες μέθοδοι έχουν αναπτυχθεί. Για την τελευταία περίπτωση γίνεται λεπτομερής περιγραφή σε επόμενη ενότητα.

Συμπερασματικά, αφού τα σφάλματα τα οποία αντιστοιχούν σε διαφορετικές περιπτώσεις είναι μεταξύ τους ασυσχέτιστα και ακολουθούν την κανονική κατανομή, θα είναι και ανεξάρτητα. Επομένως, η τιμή που παίρνει ένα σφάλμα σε κάποια επανάληψη, δεν δίνει καμία πληροφορία για την τιμή του σφάλματος σε κάποια επόμενη περίπτωση.

2.5 Κανονικό μοντέλο Πολλαπλής γραμμικής παλινδρόμησης

Σε προηγούμενη ενότητα έγινε μελέτη στην περίπτωση που οι τιμές μιας προβλέπουσας μεταβλητής X επηρεάζουν τις τιμές μιας μεταβλητής απόκρισης Y. Στην ενότητα αυτή θα αναλυθεί η περίπτωση που υπάρχουν δύο ή περισσότερες

ανεξάρτητες μεταβλητές X_i i=1,2,3,… ,n οι τιμές των οποίων επηρεάζουν τις τιμές μιας μεταβλητής απόκρισης Y. Η κεντρική ιδέα πίσω από την προσθήκη περισσότερων από μία ανεξάρτητων μεταβλητών, σε ένα απλό μοντέλο γραμμικής παλινδρόμησης, είναι να ερμηνευθεί το "κομμάτι" της μεταβλητής Y, το οποίο μένει ανεξήγητο από μία μόνο μεταβλητή.

Έστω ότι αυτή τη φορά πρέπει να εξεταστεί με ποιό τρόπο μεταβάλλεται το μέγεθος των ασφαλιστικών αποζημιώσεων από την ηλικία του οδηγού (X_1), το φύλο του οδηγού (X_2) και το είδος του αυτοκινήτου (X_3). Σε αυτήν την περίπτωση θα είναι διαθέσιμες τρείς ανεξάρτητες μεταβλητές, οι τιμές των οποίων επηρεάζουν το μέγεθος της ασφαλιστικής αποζημίωσης.

Το μοντέλο πολλαπλής παλινδρόμησης περιγράφεται παραπάνω αποτελεί την γενική περίπτωση, όπου κάποιες μεταβλητές είναι συνεχείς ενώ κάποιες άλλες κατηγορικές, και θα έχει την εξής μορφή:

$$Y_i=\beta_0 +\beta_1 x_{i1}+ \beta_2 x_{i2}+ \beta_3 x_{i3}+\varepsilon_i ,$$

όπου η μεταβλητή x_{i1} εκφράζει την ηλικία του οδηγού στην i επανάληψη του πειράματος, η μεταβλητή x_{i2} εκφράζει το φύλο του οδηγού στην i επανάληψη και τέλος η μεταβλητή x_{i3} συμβολίζει το είδος του οχήματος κατά την i επανάληψη.

Όπως και στο απλό στατιστικό γραμμικό μοντέλο για τον προσδιορισμό των παραμέτρων β_0, β_1,…, β_v είναι απαραίτητο να ελαχιστοποιηθούν οι κατακόρυφες αποστάσεις

$$\varepsilon_i = y_i – (\beta_0+\beta_1 x_{i1}+ \beta_2 x_{i2}+…+ \beta_v x_{ip}+\varepsilon_i \ ), \ i=1,2,..,n$$

Για τις κατακόρυφες αυτές αποστάσεις ε_i θα χρησιμοποιηθούν όπως αναφέρθηκε και παραπάνω οι όροι αποκλίσεις ή σφάλματα, ενώ για τις εκτιμήσεις των σφαλμάτων $\hat{\varepsilon}_i = y_i – \hat{y}_i$ ο όρος κατάλοιπα (residuals).

Για τον καθορισμό της βέλτιστης ευθείας παλινδρόμησης η μέθοδος που έχει επικρατήσει είναι η μέθοδος ελαχίστων τετραγώνων στην οποία έγινε αναφορά και στην προηγούμενη ενότητα, κατά την οποία ελαχιστοποιείται η παρακάτω ποσότητα:

$$\text{SSE} = \sum_{i}^{n} \varepsilon_i^2 = \sum_{i}^{n} \left[y_i - \left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_v x_{ip} \right) \right]^2$$

Συμβολίζοντας τους εκτιμητές που προκύπτουν με $\hat{\beta}_i$ η ευθεία $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_v x_{ip}$, i =1,2,3,… ,n καλείται ευθεία ελαχίστων τετραγώνων.

Για το παραπάνω μοντέλο πολλαπλής παλινδρόμησης ισχύουν οι παρακάτω υποθέσεις:

- Οι τιμές των ανεξάρτητων μεταβλητών συμβολίζονται με $x_{i1}, x_{i2}, \ldots, x_{ip}$, i= 1,2,..,n και είναι γνωστοί αριθμοί.

- Οι ποσότητες $\beta_0, \beta_1, \ldots, \beta_p$ είναι άγνωστες παράμετροι.

- Τα ε_i i=1,2,…,n ονομάζονται τυχαία σφάλματα και ακολουθούν την κανονική κατανομή $N(0,\sigma^2)$.
- Τα σφάλματα που αντιστοιχούν σε διαφορετικές επαναλήψεις του πειράματος είναι ασυσχέτιστα και αφού ακολουθούν την κανονική κατανομή θα είναι και ανεξάρτητα.
- Η εξαρτημένη μεταβλητή Y_i , είναι τυχαία μεταβλητή που ακολουθεί την κανονική κατανομή (αφού τα ε_i είναι τυχαίες μεταβλητές που ακολουθούν την $N(0,\sigma^2)$ και οι τιμές της συμβολίζονται με y_i.

(Κούτρας 2015)

Ένα μοντέλο το οποίο ικανοποιεί τις παραπάνω υποθέσεις ονομάζεται κανονικό μοντέλο πολλαπλής παλινδρόμησης και η γενική του μορφή είναι

$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_v x_{ip} + \varepsilon_i$ i= 1,2,…,n.

2.6 Φυσική ερμηνεία των παραμέτρων

Η παράμετρος $\hat{\beta}_0$ εκφράζει την αναμενόμενη τιμή της μεταβλητής απόκρισης Y_i κατά την i επανάληψη του πειράματος, όταν όλες οι επεξηγηματικές μεταβλητές έχουν μηδενική τιμή.

Η παράμετρος $\hat{\beta}_i$ εκφράζει την μεταβολή της μεταβλητής απόκρισης Y_i όταν η τιμή της ανεξάρτητης μεταβλητής X_i μεταβληθεί κατά μία μονάδα και η τιμή των άλλων μεταβλητών παραμείνει σταθερή, όπου i=1,2,..,p.

Μία σημαντική επισήμανση είναι ότι το ποσοστό της συνολικής μεταβλητότητας που ερμηνεύει η κάθε επεξηγηματική μεταβλητή αλλάζει ανάλογα με την σειρά εισαγωγής της στο μοντέλο. Δηλαδή, άλλο ποσοστό μεταβλητότητας θα ερμηνεύει η ανεξάρτητη μεταβλητή X_1 αν εισαχθεί σε ένα μοντέλο που δεν περιέχει άλλες ανεξάρτητες μεταβλητές και άλλο ποσοστό θα ερμηνεύει αν εισαχθεί σε ένα μοντέλο στο οποίο υπάρχει ήδη μια ακόμα ανεξάρτητη μεταβλητή. Επίσης μια μεταβλητή η οποία κρίνεται αρχικά ως μη χρήσιμη στο μοντέλο, και άρα δεν συμπεριλαμβάνεται σε αυτό, μπορεί τελικά με την είσοδο άλλων επεξηγηματικών μεταβλητών, να εισαχθεί στο μοντέλο. Αυτό οφείλεται στην αλληλεπίδραση που μπορεί να υπάρχει ανάμεσα στις ανεξάρτητες μεταβλητές.(Κούτρας 2015)

2.6.1 Επεξηγηματικές μεταβλητές

- Μετασχηματισμοί των επεξηγηματικών μεταβλητών είναι συχνά απαραίτητοι, προκειμένου να τηρούνται οι υποθέσεις της πολλαπλής γραμμικής παλινδρόμησης. Με αυτόν τον τρόπο δίνεται η δυνατότητα να χρησιμοποιηθούν συχνότερα και άρα να γενικευτεί η μέθοδος αυτή.
- Σε περίπτωση που γραφική παράσταση της αναμενόμενης μέσης τιμής της εξαρτημένης μεταβλητής είναι καμπύλη, η ανεξάρτητη μεταβλητή μπορεί να υψωθεί σε κάποια δύναμη και έτσι δημιουργείται ένας πολυωνυμικός μετασχηματισμός. Άλλοι συνήθεις μετασχηματισμοί είναι ο λογάριθμος (ή διπλός λογάριθμος) και η τετραγωνική ρίζα της μεταβλητής απόκρισης.
- Ο συνδυασμός των επεξηγηματικών μεταβλητών είναι πολλές φορές χρήσιμος. Για παράδειγμα, για τον υπολογισμό του δείκτη μάζας σώματος κάποιου, διαιρούμε των ύψος του (μεταβλητή X_1) με το τετράγωνο της μάζας

του (X_2^2). Τα γινόμενα μεταξύ των ανεξάρτητων μεταβλητών ονομάζονται αλληλεπιδράσεις και αρκετές φορές κρίνονται σημαντικά, για την πρόβλεψη της προβλέπουσας μεταβλητής, και επομένως εισάγονται στο μοντέλο.

- Ψευδομεταβλητές (Dummy variables) και παράγοντες συχνά χρησιμοποιούνται στα μοντέλα παλινδρόμησης. Ένας παράγοντας ο οποίος έχει μόνο δύο επίπεδα μπορεί να εισαχθεί στο μοντέλο ως μια δείκτρια μεταβλητή I_n η οποία παίρνει συνήθως τιμές μηδέν και ένα, υποδεικνύοντας έτσι ποια κατηγορία του παράγοντα είναι παρούσα και ποια όχι. Κατηγορικές μεταβλητές με περισσότερα από δύο επίπεδα, απαιτούν αρκετές ψευδομεταβλητές για να αναπαρασταθούν (Hardy Melissa 1993).

2.6.2 Πολυσυγγραμικότητα (VIF)

Ο δείκτης VIF (Variance Inflation Factor) χρησιμοποιείται για να εντοπιστεί η ενδεχόμενη πολυσυγγραμικότητα μεταξύ των επεξηγηματικών μεταβλητών, σε ένα μοντέλο πολλαπλής παλινδρόμησης (Belsley D.A, Kuh E. and Welsch R.E. 1980). Ο δείκτης αυτός εκτιμά, πόση από την συνολική διακύμανση του μοντέλου, οφείλεται στην υψηλή συσχέτιση (συγγραμμικότητα) που παρουσιάζεται συχνά, μεταξύ των ανεξάρτητων μεταβλητών.

Η υψηλή τιμή του δείκτη VIF υποδηλώνει αυξημένη διακύμανση στην εκτίμηση των συντελεστών παλινδρόμησης λόγω πολυσυγγραμμικότητας μεταξύ των επεξηγηματικών μεταβλητών και επομένως επηρεάζει τα αποτελέσματα της παλινδρόμησης (StewartG.W. 1987).

Ενδείξεις πολυσυγγραμμικότητας μπορεί να παρατηρηθούν σε περιπτώσεις όπου μικρές αλλαγές στα δεδομένα επιφέρουν μεγάλες αλλαγές στις εκτιμήσεις των παραμέτρων, οι συντελεστές παλινδρόμησης έχουν μεγάλα τυπικά σφάλματα και χαμηλά επίπεδα σημαντικότητας παρότι είναι από κοινού σημαντικά και ο συντελεστής προσδιορισμού R^2 είναι αρκετά υψηλός και τέλος σε περιπτώσεις που οι

συντελεστές παλινδρόμησης δίνουν ερμηνείες που εμφανώς δεν αντικατοπτρίζουν την πραγματικότητα (WilliamH. Greene 2002).

2.7 Έλεγχοι υποθέσεων

Στην παρακάτω ενότητα θα γίνει αναφορά στους σημαντικότερους ελέγχους υποθέσεων που μπορούν να πραγματοποιηθούν σε ένα μοντέλο παλινδρόμησης.

Έλεγχος σημαντικότητας πολλαπλής παλινδρόμησης

H_0: $\beta_1=\beta_2=\ldots=\beta_p=0$

H_1: τουλάχιστον ένα από τα $\beta_i \neq 0$, i= 1,2,…,n

Οι παραπάνω υποθέσεις σχετίζονται με την σημαντικότητα του μοντέλου παλινδρόμησης. Η αποτυχία να απορριφθεί η μηδενική υπόθεση υποδεικνύει πως δεν υπάρχει γραμμική σχέση ανάμεσα στην εξαρτημένη και τις ανεξάρτητες μεταβλητές. Σε περίπτωση όμως που απορριφθεί η μηδενική υπόθεση, αυτό σημαίνει ότι τουλάχιστον ένα από τα x_i έχει στατιστικά σημαντική γραμμική σχέση με την y.

Σε ένα κανονικό μοντέλο γραμμικής παλινδρόμησης αν ισχύει η παραπάνω μηδενική υπόθεση, η τυχαία μεταβλητή $\frac{MSR}{MSE}$ ακολουθεί την κατανομή F με n και n-p βαθμούς ελευθερίας.

Με MSR = $\frac{SSR}{p}$ συμβολίζεται το μέσο άθροισμα τετραγώνων της παλινδρόμησης ενώ με MSE = $\frac{SSE}{n-p-1}$ συμβολίζεται το μέσο τετραγωνικό σφάλμα του μοντέλου.

Ο κανόνας απόφασης για τον έλεγχο του παραπάνω ελέγχου σε επίπεδο σημαντικότητας α είναι ο εξής:

Αν ισχύει $F>F_{p-1,n-p}(\alpha)$ απορρίπτεται η H_0.

Αν ισχύει $F \leq F_{p-1,n-p}(\alpha)$ δεν απορρίπτεται η H_0,

όπου $F = \frac{MSR}{MSE}$

Σε αντίθεση με τον παραπάνω πολλαπλό έλεγχο, μεμονωμένοι έλεγχοι για την κάθε παράμετρο του μοντέλου είναι εφικτοί.

T-test

Το t-testστην ανάλυση παλινδρόμησης χρησιμοποιείται για να προσδιοριστεί αν οι διαθέσιμες επεξηγηματικές μεταβλητές είναι στατιστικά σημαντικές και επομένως αν πρέπει να συμπεριληφθούν στο τελικό μοντέλο παλινδρόμησης. Ο έλεγχος υπόθεσης που διεξάγεται είναι ο εξής:

$H_0 : \beta_i = 0$

$H_1 : \beta_i \neq 0$

Ο κανόνας απόφασης για τον παραπάνω έλεγχο υπόθεσης στο κανονικό μοντέλο πολλαπλής παλινδρόμησης σε επίπεδο σημαντικότητας α είναι ο εξής:

- Αν ισχύει $|T|>t_v(\alpha/2)$ τότε απορρίπτεται η H_0.
- Αν ισχύει $|T| \leq t_v(\alpha/2)$ τότε δεν απορρίπτεται η H_0,

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

Το παραπάνω τεστ εκτιμάει την συνεισφορά κάποιας μεταβλητής, ενώ ταυτόχρονα κάποιες άλλες μεταβλητές περιλαμβάνονται ενδεχομένως στο μοντέλο (Draperand Smith, 1998).

2.8 Μετασχηματισμοί Κανονικότητας και Γραμμικότητας

Πολύ συχνά κάποιες από τις παραπάνω υποθέσεις του υποδείγματος παλινδρόμησης παραβιάζονται. Σε αυτές τις περιπτώσεις είναι απαραίτητοι οι μετασχηματισμοί των μεταβλητών απόκρισης ή ακόμα και των επεξηγηματικών μεταβλητών, προκειμένου να ικανοποιηθούν οι υποθέσεις κανονικότητας και ομοσκεδαστικότητας των σφαλμάτων και της γραμμικής σχέσης ανάμεσα στις «ανεξάρτητες» και τις «εξαρτημένες» μεταβλητές. Όταν μία επεξηγηματική μεταβλητή έχει ισχυρή ασυμμετρία, η γραμμική σχέση με την εξαρτημένη μεταβλητή Y είναι σχεδόν απίθανη. Με κατάλληλους μετασχηματισμούς σε τέτοιου τύπου μεταβλητές είναι δυνατή μία κανονικοποίηση της κατανομής τους ή δημιουργία συμμετρίας της κατανομής τους με αποτέλεσμα συχνά να δημιουργούν μία πιο ισχυρή γραμμική σχέση ανάμεσα στις δύο μεταβλητές (De Jong and Heller, 2008).

Οι πιο συνηθισμένοι μετασχηματισμοί των μεταβλητών είναι η τετραγωνική ρίζα, ο λογάριθμος και η ύψωση σε κάποια δύναμη της εκάστοτε μεταβλητής. Ειδικότερα με την μέθοδο Box-Cox είναι δυνατή η εύρεση του σωστού μετασχηματισμού που ταιριάζει ανάλογα με την περίπτωση (Box and Cox, 1964).

Στο παρελθόν τέτοιου τύπου μετασχηματισμοί ήταν πολύ σημαντικοί, πριν να γίνει διαθέσιμη η μέθοδος των γενικευμένων γραμμικών μοντέλων. Τότε, ήταν απαραίτητη η εύρεση κατάλληλου μετασχηματισμού προκειμένου να ικανοποιηθεί η υπόθεση της κανονικότητας και να γίνει η χρήση της κανονικής γραμμικής παλινδρόμησης. Με τη χρήση γενικευμένων γραμμικών μοντέλων, που αναλύεται στο επόμενο κεφάλαιο, τα δεδομένα μοντελοποιούνται δίχως κάποιος μετασχηματισμός να είναι απαραίτητος (Πολίτης 2015).

Κεφάλαιο 3

Λογιστική Παλινδρόμηση

3.1 Εισαγωγή στη λογιστική παλινδρόμηση

Σε κάποιες περιπτώσεις παλινδρόμησης, η υπόθεση της κανονικότητας της μεταβλητής απόκρισης παραβιάζεται. Σε περίπτωση που δεν είναι δυνατή η εύρεση κάποιου μετασχηματισμού για την κανονικοποίηση της μεταβλητής απόκρισης, δεν μπορεί να χρησιμοποιηθεί η μέθοδος γραμμικής παλινδρόμησης που αναπτύχθηκε στο προηγούμενο κεφάλαιο.

Στο κεφάλαιο αυτό θα γίνει η περιγραφή της μεθόδου με την οποία μπορούμε να αναλύσουμε τέτοιου είδους δεδομένα.

Αρχικά, εξετάζεται αν με κατάλληλους μετασχηματισμούς της εξαρτημένης μεταβλητής, μπορεί εκείνη να προσεγγιστεί ικανοποιητικά από την κανονική κατανομή. Συνήθεις μετασχηματισμοί είναι ο λογάριθμος της μεταβλητής απόκρισης, η ύψωση σε κάποια δύναμη κτλ.

Σε περίπτωση που βρεθεί κατάλληλος μετασχηματισμός της μεταβλητής απόκρισης ώστε να προσεγγίζεται ικανοποιητικά από την κανονική κατανομή, χρησιμοποιείται η μέθοδος γραμμικής παλινδρόμησης. Πολύ συχνά όμως υπάρχουν περιπτώσεις που η κατανομή της εξαρτημένης μεταβλητής Y_i δεν επιδέχεται τέτοιους μετασχηματισμούς ώστε να έχει τις επιθυμητές ιδιότητες. Χαρακτηριστικό παράδειγμα είναι όταν η μεταβλητή απόκρισης είναι δίτιμη. Σε αυτήν την περίπτωση οι τιμές που παίρνει η μεταβλητή είναι δύο, για παράδειγμα 0 και 1. Συνήθως με 0 συμβολίζεται η απουσία ενός χαρακτηριστικού που μας ενδιαφέρει και με 1 συμβολίζεται η παρουσία του χαρακτηριστικού αυτού (Πολίτης 2015). Παράδειγμα

τέτοιας μεταβλητής αποτελεί η παροχή ασφαλιστικής αποζημίωσης ή όχι, σε κάποιο περιστατικό ατυχήματος.

Στην παραπάνω περίπτωση, η μεταβλητή απόκρισης ακολουθεί την κατανομή Bernoulli(1,p), όπου p είναι η πιθανότητα επιτυχίας (δηλαδή η παρουσία του χαρακτηριστικού). Όπως είναι γνωστό από την θεωρία πιθανοτήτων στην κατανομή αυτή η μέση τιμή ισούται με $E[Y_I] = p_i$.

3.1.1 Το μοντέλο της λογιστικής παλινδρόμησης

Επομένως αν προσαρμόσουμε μια γραμμική σχέση της παραπάνω μέσης τιμής Y_i με τις ανεξάρτητες X_i=1,2,…,p προκύπτει η παρακάτω σχέση:

$p_i = E[Y_i] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$

Εκτός από το πρόβλημα αδυναμίας προσέγγισης της κατανομής Bernoulli από την κανονική κατανομή, εμφανίζεται και το πρόβλημα της ετεροσκεδαστικότητας. Η διακύμανση σε αυτήν την περίπτωση εξαρτάται από την πιθανότητα επιτυχίας p_i, η οποία δεν είναι σταθερή και ισούται με p_i $(1-p_i)$. Τέλος, ένα πρόβλημα το οποίο ανακύπτει και δεν επιτρέπει να χρησιμοποιηθεί το παραπάνω μοντέλο σε αυτήν την μορφή είναι το γεγονός ότι το σύνολο τιμών της $p_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ ενδέχεται να μην είναι το [0,1], όπως θα έπρεπε.

Για τους παραπάνω λόγους χρησιμοποιείται μία συνάρτηση g: [0,1] $\rightarrow \mathbb{R}$ της μέσης τιμής της μεταβλητής απόκρισης Y, η οποία ονομάζεται συνάρτηση σύνδεσης. Η συνάρτηση αυτή μετασχηματίζει την μεταβλητή και την απεικονίζει στο διάστημα [0,1]. Αυτή η συνάρτηση είναι “1-1” και γνησίως μονότονη και άρα αντιστρέψιμη.

$\eta_i = g(p_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \Rightarrow g^{-1}(\eta_i) = p_i \in [0,1]$

Η πιο συνηθισμένη συνάρτηση σύνδεσης που χρησιμοποιείται για δίτιμα δεδομένα είναι η Logit

$\eta_i = \text{logit}(p_i) = \log\left[\frac{p_i}{1-p_i}\right]$

Ο λόγος $\frac{p_i}{1-p_i}$ ονομάζεται σχετική πιθανότητα (odds) του ενδεχομένου που μας ενδιαφέρει ("επιτυχία"). Για παράδειγμα, αν η πιθανότητα επιτυχίας είναι 0,25, η σχετική πιθανότητα επιτυχίας είναι $\frac{0.25}{0.75} = \frac{1}{3}$, δηλαδή μία επιτυχία σε κάθε 3 αποτυχίες.

Επομένως το μοντέλο λογιστικής παλινδρόμησης χρησιμοποιώντας σαν συνάρτησης σύνδεσης την Logit (λογάριθμος της σχετικής πιθανότητας επιτυχίας) θα έχει τη μορφή

$\log\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ip}$

3.2 Ερμηνεία παραμέτρων λογιστικής παλινδρόμησης

Η παραπάνω σχέση γράφεται ισοδύναμα στην μορφή

$$\frac{p_i}{1-p_i} = e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p \beta x_{ip}} = e^{\beta_0} e^{\beta_1 x_{i1}} e^{\beta_2 x_{i2}} \ldots e^{\beta_p x_{ip}}$$

Επομένως αυξάνοντας της επεξηγηματική μεταβλητή x_1 κατά μία μονάδα πολλαπλασιάζεται η σχετική πιθανότητα επιτυχίας με e^{β_1}**.** Αν το $\beta < 0$ το αποτέλεσμα της αύξησης του x_i είναι η ελάττωση της σχετικής πιθανότητας. Αντίστροφα, αν $\beta > 0$ τότε η αύξηση του x_i επιφέρει και αύξηση της σχετικής πιθανότητας. Τέλος, αν $\beta = 0$ τότε $e^{\beta} = 1$ και δεν υπάρχει καμία επιρροή στην τιμή της σχετικής πιθανότητας.

Αν μία επεξηγηματική μεταβλητή έχει m επίπεδα, τότε υποθέτοντας ότι το επίπεδο m είναι το επίπεδο αναφοράς, το μοντέλο θα έχει τη μορφή

$$\log\left[\frac{p_i}{1-p_i}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{k-1} x_{m-1}$$

όπου x_j είναι δείκτριες μεταβλητές με $x_j = 1$ αν η x είναι στο επίπεδο j και 0 αλλιώς, j = 1,2,…,m-1.

3.3 Διωνυμική κατανομή

Αν όλες οι επεξηγηματικές μεταβλητές είναι κατηγορικές και η μεταβλητή απόκρισης είναι δίτιμη τότε τα δεδομένα μπορούν να ομαδοποιηθούν. Σε αυτήν την περίπτωση η μεταβλητή απόκρισης θα ακολουθεί την Διωνυμική κατανομή $Y_i \sim Bin(n_i, p_i)$, όπου n_i είναι το πλήθος των ατόμων σε κάθε ομάδα i και p_i η πιθανότητα επιτυχίας. Επομένως προκύπτει ένα γενικευμένο γραμμικό μοντέλο στο οποίο η ανεξάρτητη μεταβλητή Y_i αντιπροσωπεύει το πλήθος των επιτυχιών στις n_i δοκιμές. Βασική υπόθεση είναι οι Y_i να είναι ανεξάρτητες ανά δύο τυχαίες μεταβλητές.

Στην πολλαπλό μοντέλο παλινδρόμησης, η συνάρτηση της μέσης τιμής και η συνάρτηση διακύμανσης, έχουν εντελώς διαφορετικές παραμέτρους. Αυτό όμως δεν ισχύει στην λογιστική παλινδρόμηση. Η τιμή του p_i προσδιορίζει και την συνάρτηση μέσης τιμής και την συνάρτηση διακύμανσης. Επομένως, χρειάζεται να εκτιμήσουμε την p_i (Πολίτης 2015).

3.4 Έλεγχος καλής προσαρμογής

Η καλή προσαρμογή ενός μοντέλου στα δεδομένα είναι ένα φυσικό ερώτημα για όλα τα στατιστικά μοντέλα. Οι βασικές αρχές ελέγχου σημαντικότητας, επιλογής μοντέλου και διαγνωστικών ελέγχων είναι όμοιες με αυτές των γραμμικών μοντέλων. Παρ' όλα αυτά υπάρχουν διαφορές σε κάποιες τεχνικές λεπτομέρειες.

Ένας τρόπος ελέγχου καλής προσαρμογής του μοντέλου είναι η σύγκρισή του με το κορεσμένο μοντέλο.

Σε αντίθεση με την κατανομή Bernoulli η Διωνυμική κατανομή προσεγγίζεται από την Κανονική όταν το πλήθος n των ανεξάρτητων δοκιμών τείνει στο άπειρο (με βάση το Κεντρικό Οριακό Θεώρημα) και άρα μπορούμε να εφαρμόσουμε έλεγχο

καλής προσαρμογής με βάση την απόκλιση ενός μόνο μοντέλου. Οι υποθέσεις του ελέγχου είναι οι ακόλουθες:

H_0 : Το μοντέλο δεν διαφέρει σημαντικά από το κορεσμένο (επομένως είναι επαρκές)

H_1 : Το μοντέλο διαφέρει σημαντικά από το κορεσμένο

- ✓ Σημειώνουμε ότι το κορεσμένο μοντέλο είναι αυτό που έχει τόσες παραμέτρους, όσες και το πλήθος των δεδομένων.
- ✓ Πλήρες μοντέλο είναι το μοντέλο που χρησιμοποιεί όλες τις διαθέσιμες μεταβλητές.

3.5 Η Λογιστική παλινδρόμηση στην αντασφάλιση χαρτοφυλακίου

Εισαγωγή

Η αντασφάλιση είναι μια οικονομική συναλλαγή κατά την οποία μέρος του κινδύνου μεταφέρεται από τη μία ασφαλιστική εταιρεία (Πρωτασφαλιστής) σε μια αντασφαλιστική εταιρεία (Αντασφαλιστής) με κάποιο οικονομικό αντάλλαγμα. Οι όροι της αντασφάλισης καθορίζονται και κατοχυρώνονται μέσα από μια αντασφαλιστική σύμβαση η οποία υπογράφεται και από τις δύο πλευρές. Κατά την διαδικασία αυτή ο αντασφαλιστής, με αντάλλαγμα κάποιο ασφάλιστρο συμφωνεί να αποζημιώσει τον Πρωτασφαλιστή, στο ακέραιο ή για κάποιο μέρος της ασφαλιστικής ευθύνης που έχει αναλάβει λόγω της σύμβασης που έχει εκδοθεί. Οι πιο συνήθεις λόγοι αντασφάλισης είναι η ελάφρυνση των κινδύνων από το χαρτοφυλάκιο, η απαλλαγή από υποχρέωση τήρησης αποθεματικών και η μεταφορά τεχνογνωσίας (Rodolfo Wehrhahn 2009). Στις μέρες μας, με την έλευση του Solvency II, η ανάγκη αντασφάλισης για τις ασφαλιστικές εταιρείες, είναι πιο επιτακτική από ποτέ.

3.5.1 Τύποι αντασφάλισης

Τα δύο κύρια συμβόλαια αντασφάλισης είναι τα «προαιρετικά» συμβόλαια (facultative reinsurance) και τα συμβόλαια που καθορίζονται με κάποια συνθήκη (treaty reinsurance).

Στα προαιρετικά συμβόλαια, η ασφαλιστική εταιρεία διαπραγματεύεται ένα συμβόλαιο για κάθε ασφαλιστική κατηγορία που επιθυμεί να αντασφαλίσει. Είναι ιδιαιτέρως χρήσιμα σε περιπτώσεις που χρειάζεται να καλύψει πολύ μεγάλους κινδύνους, τους οποίους δεν είναι πρόθυμη ή δεν είναι σε θέση να καλύψει μόνη της, ή κινδύνους που είναι αρκετά ασυνήθιστοι. Στην περίπτωση αυτή δεν υπάρχει κάποια προκαθορισμένη σύμβαση μεταξύ της ασφαλιστικής και της αντασφαλιστικής εταιρείας και οι όροι καθορίζονται ανάλογα την περίπτωση.

Η προαιρετική μέθοδος συνήθως είναι αρκετά χρονοβόρα και μεγάλη σε κόστος με αποτέλεσμα η μέθοδος της αντασφάλισης με σύμβαση να είναι αυτή που χρησιμοποιείται πιο συχνά. Στην περίπτωση αυτή οι δύο εταιρείες καθορίζουν εκ των προτέρων τους όρους και τις προϋποθέσεις της αντασφάλισης. Ο αντασφαλιστής δεν είναι σε θέση να εξετάζει την κάθε περίπτωση μεμονωμένα και να αποφασίζει αλλά αντιθέτως είναι υποχρεωμένος να τηρήσει το συμβόλαιο που έχει συναφθεί και να καλύψει την ασφαλιστική εταιρεία για τον κίνδυνο που έχει συμφωνηθεί (Outreville,1998).

3.5.2 Μορφές αντασφάλισης

Οι δύο κύριες μορφές αντασφάλισης είναι η αναλογική (proportional) και η μη-αναλογική (non-proportional) αντασφάλιση.

Στην αναλογική αντασφάλιση υπάρχει η ίδια αναλογία στην εκχώρηση κινδύνου από τον πρωτασφαλιστή και τον αντασφαλιστή, στην εκχώρηση των ασφαλίστρων και στην αποζημίωση της ζημίας. Στην μορφή αυτή αντασφάλισης ο αντασφαλιστής επιστρέφει στον πρωτασφαλιστή ποσοστό από την προμήθεια (ceding comission) επί του ασφαλίστρου που λαμβάνει, έτσι ώστε ο πρωτασφαλιστής να καλύψει το κόστος εγγραφής στο χαρτοφυλάκιο.

Κάθε τύπος αντασφάλισης δίχως τα παραπάνω χαρακτηριστικά καλείται μη-αναλογική αντασφάλιση. Την μορφή αυτή επιλέγουν συνήθως ασφαλιστικές εταιρείες που στοχεύουν κατά κύριο λόγο στην προστασία του χαρτοφυλακίου τους από πολλαπλούς κινδύνους που μπορεί να οδηγήσουν σε υπερβολικά μεγάλες αποζημιώσεις (Wehrhahn, 2009).

3.5.3 Αντασφάλιση υπερβάλλοντος ζημίας (Excess of Loss)

Η πιο συνηθισμένη μορφή μη αναλογικής αντασφάλισης είναι αυτή της υπερβάλλοντος ζημίας. Στην περίπτωση αυτή τα ενδιαφερόμενα μέρη συνάπτουν μία σύμβαση κατά την οποία ο αντασφαλιστής θα πληρώσει όταν μία ζημιά υπερβαίνει ένα συγκεκριμένο όριο (σημείο υπέρβασης) για ένα συγκεκριμένο κίνδυνο ή περισσότερους κινδύνους που μπορεί να επέλθουν από ένα γεγονός. Στο πλαίσιο αυτής της σύμβασης είναι δυνατό να υπάρχουν περισσότερα από ένα σημεία υπέρβασης, εφόσον το ένα δεν καλύπτει το άλλο (Carter 1983).

Ακολουθεί ένα παράδειγμα σύμβασης με πολλαπλά σημεία υπέρβασης και στην συνέχεια οι αποζημιώσεις οι οποίες προκύπτουν.

Σημείο	Ποσό

υπέρβασης	αντασφάλισης
50.000 €	200.000 €
200.000 €	500.000 €
500.000 €	1.000.000 €

Πίνακας 3.1: Παράδειγμα αντασφάλισης

Στην περίπτωση αυτή η αποζημιώσεις θα είναι οι ακόλουθες:

Ποσό ζημιάς	Πρωτασφαλιστής	Αντασφαλιστής		
		1ο επίπεδο	2ο επίπεδο	3ο επίπεδο
50.000 €	50.000 €	0 €	0 €	0 €
100.000 €	50.000 €	50.000 €	0 €	0 €
300.000 €	50.000 €	200.000 €	50.000 €	0 €
900.000 €	50.000 €	200.000 €	500.000 €	150.000 €

Πίνακας 3.2 : παράδειγμα αντασφαλιστικών αποζημιώσεων

Η πρώτη εφαρμογή σύμβασης υπερβάλλοντος ζημίας ευρέως αποδίδεται στον Cuthbert Heath στις αρχές του 20^{ου} αιώνα. (**https://blog.willis.com/2013/12/dealing-with-the-big-stuff-excess-of-loss-reinsurance/**)

Στην εργασία αυτή θα γίνει χρήση της αντασφάλισης υπερβάλλοντος ζημίας σε δεδομένα ασφαλιστικών αποζημιώσεων στο κεφάλαιο που ακολουθεί.

Κεφάλαιο 4

Εφαρμογή μεθόδου γραμμικής παλινδρόμησης

4.1 Στατιστικό πακέτο R

Η R είναι μια γλώσσα προγραμματισμού που παρέχει στον χρήστη την δυνατότητα να κάνει μεταξύ άλλων υπολογιστική στατιστική και γραφήματα. Έχει πολλές ομοιότητες με την S γλώσσα η οποία αναπτύχθηκε στα εργαστήρια Bell από τον John Chambers. Μπορεί να θεωρηθεί μία διαφορετική εφαρμογή της S,αν και υπάρχουν σημαντικές διαφορές, ο κώδικας της γλώσσας Sμπορεί να εφαρμοστεί στην R.

Η R στην σημερινή της μορφή είναι ένα στατιστικό πακέτο το οποίο αποτελεί αποτέλεσμα συνεργασίας και συλλογικής προσπάθειας καθώς δίνει την δυνατότητα σε όλους να κάνουν βελτιώσεις στον πηγαίο κώδικα της R και να τις δημοσιεύουν. Παρέχει την δυνατότητα στο χρήστη να εφαρμόσει μια μεγάλη ποικιλία στατιστικών (γραμμικά και μη γραμμικά μοντέλα, στατιστικά test, πολυμεταβλητή ανάλυση κτλ) και γραφικών τεχνικών (ιστόγραμμα, qqplot, pie chart κτλ). Επιπρόσθετα αυτό έχει σαν αποτέλεσμα να έχουν γίνει πολλές βελτιώσεις από τότε που δημιουργήθηκε. Οι αρχικοί δημιουργοί ήταν οι Robert Gentleman και Ross Ihaka (https://www.r-project.org/contributors.html).

4.1.1 Περιβάλλον R

Η Reίναι ένα λογισμικό για την επεξεργασία, τον υπολογισμό και τη γραφική απεικόνιση δεδομένων. Για τα παραπάνω, περιέχει πληθώρα δυνατοτήτων στο χρήστη όπως:

- Αποτελεσματική διαχείριση και αποθήκευση δεδομένων,
- Μια μεγάλη, συνεκτική και ολοκληρωμένη συλλογή εργαλείων για την ανάλυση δεδομένων,
- Εργαλεία για την γραφική απεικόνιση των δεδομένων,
- Μια καλά αναπτυγμένη και αποτελεσματική γλώσσα προγραμματισμού

Επίσης, επιτρέπει στον χρήστη να αλληλοεπιδρά με άλλες γλώσσες όπως Java, C/C++, Python και με άλλα στατιστικά πακέτα όπως SAS, SPSS, Minitabκα.

Η R μπορεί να εμπλουτιστεί πολύ εύκολα μέσω πακέτων. Διαθέτει πάνω από 5000 πακέτα τα οποία είναι διαθέσιμα online μέσω του CRAN. Αυτό έχει σαν αποτέλεσμα να χρησιμοποιείται σε πολλούς επιστημονικούς τομείς (οικονομία, αστρονομία, χημεία, ιατρική κτλ) και μεγάλες εταιρείες, όπως για παράδειγμα Google, LinkedIn, Facebook (https://www.r-project.org/about.html).

4.2 Περιγραφή των δεδομένων

Για τη μελέτη των βασικών μοντέλων που παρουσιάστηκαν στο θεωρητικό κομμάτι της εν λόγω εργασίας θα χρησιμοποιηθεί ένα σετ δεδομένων που είναι διαθέσιμο στην R και είναι το mtcars.

Όλοι οι κώδικες που θα χρησιμοποιηθούν για την ανάλυση, βρίσκονται στο Παράρτημα Β.

Παρουσιάζονται αρχικά κάποιες αρχικές γραμμές από τα δεδομένα που έχουμε στη διάθεσή μας:

	MPG	CYL	DISP	HP	DRAT	WT	QSEC	VS	AM	GEAR	CARB
MAZDA RX4	21	6	160	110	3,9	2,62	16,46	0	1	4	4
MAZDA RX4 WAG	21	6	160	110	3,9	2,875	17,02	0	1	4	4
DATSUN 710	22,8	4	108	93	3,85	2,32	18,61	1	1	4	1
HORNET 4 DRIVE	21,4	6	258	110	3,08	3,215	19,44	1	0	3	1
HORNET SPORTABOUT	18,7	8	360	175	3,15	3,44	17,02	0	0	3	2
VALIANT	18,1	6	225	105	2,76	3,46	20,22	1	0	3	1

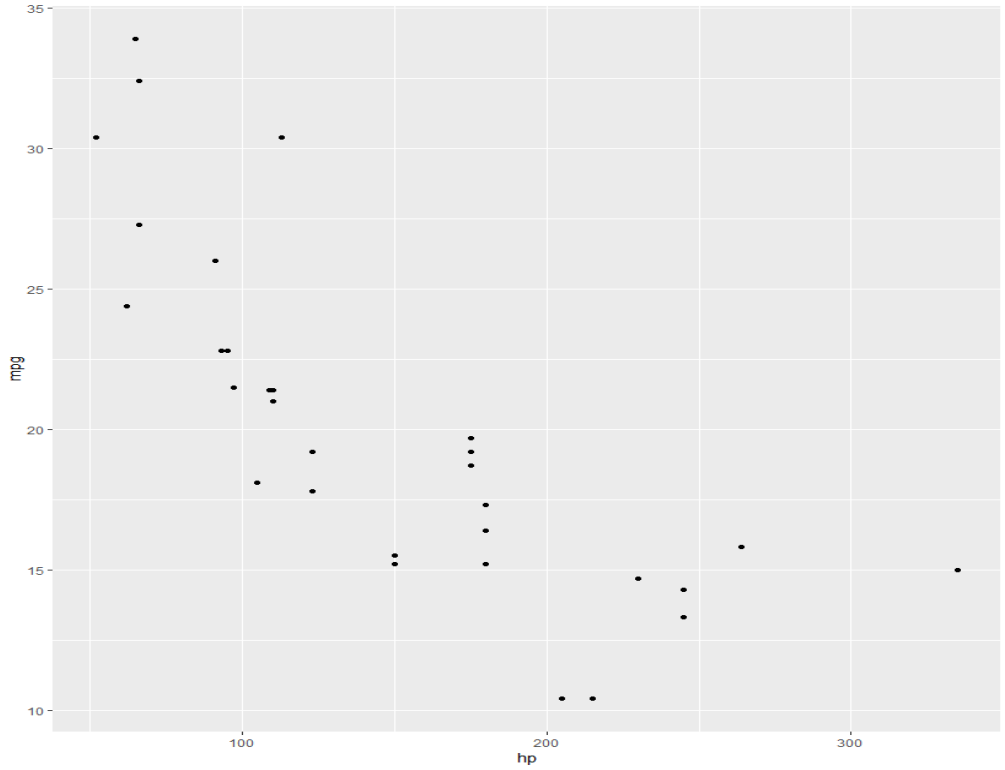
Πίνακας 4.1: Δεδομένα αυτοκινήτων (mtcars)

Αξίζει να παρατηρηθεί ότι κάθε γραμμή από τα δεδομένα αντιπροσωπεύει ένα μοντέλο οχήματος, το οποίο φαίνεται στην πρώτη στήλη του παραπάνω πίνακα. Συνεπώς κάθε γραμμή αντιπροσωπεύει ένα όχημα αποκλειστικά με τα αντίστοιχα χαρακτηριστικά του όπως είναι για παράδειγμα τα μίλια ανά γαλόνι (MPG) (δηλαδή η επάρκεια καυσίμου), ο αριθμός των κυλίνδρων του οχήματος (CYL)το εκτόπισμα του αυτοκινήτου (DISP), οι ίπποι (HP), ολόγος οπίσθιου άξονα(DRAT), το βάρος (WT), χρόνος 0-100 (QSEC),τρόπος μετάδοσης κίνησης (AM), αν είναι δηλαδή όχημα με αυτόματο κιβώτιο ταχυτήτων ή με μηχανικό, το πλήθος εμπρόσθιων τροχών(GEAR) και το πλήθος καρμπιρατέρ (CARB).

Στην περίπτωση αυτή ο αναλυτής θέλει να μελετήσει την κατανάλωση καυσίμου του οχήματος σε σχέση με τα επιμέρους χαρακτηριστικά του, προκειμένου να βοηθηθεί στην κατάλληλη αγορά οχήματος. Επομένως εξετάζει την συσχέτιση που

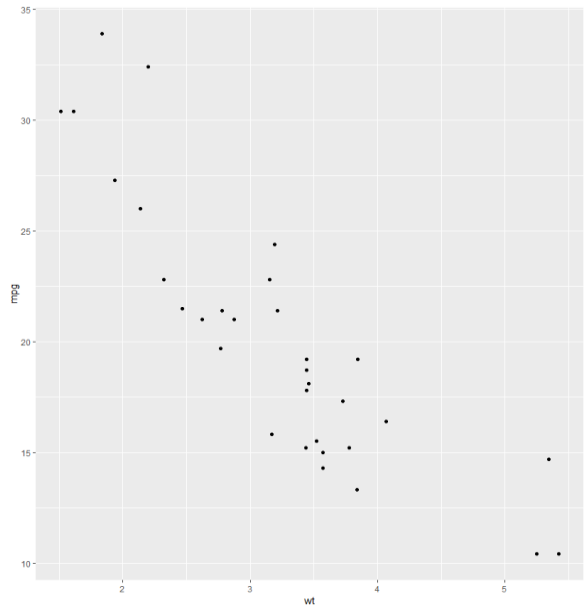
μπορεί να υπάρχει στην κατανάλωση καυσίμου με μεταβλητές όπως το βάρος, η επιτάχυνση, το πλήθος κυλίνδρων του οχήματος, ο μηχανισμός εναλλαγής ταχυτήτων κα.

Ξεκινώντας την ανάλυση των δεδομένων ακολουθούν κάποια περιγραφικά στοιχεία, ούτως ώστε να υπάρχει καλύτερη εικόνα για τα δεδομένα και συγκεκριμένα για να διερευνηθεί η ύπαρξη κάποιων μοτίβων που μπορεί να υπάρχουν. Συνήθως, στην ανάλυση παλινδρόμησης ένα διάγραμμα διασποράς είναι ένα πολύ αποτελεσματικό εργαλείο. Για το λόγο αυτό παρουσιάζουμε τα ακόλουθα γραφήματα σύμφωνα με τα οποία μπορούν να προκύψουν χρήσιμα συμπεράσματα:

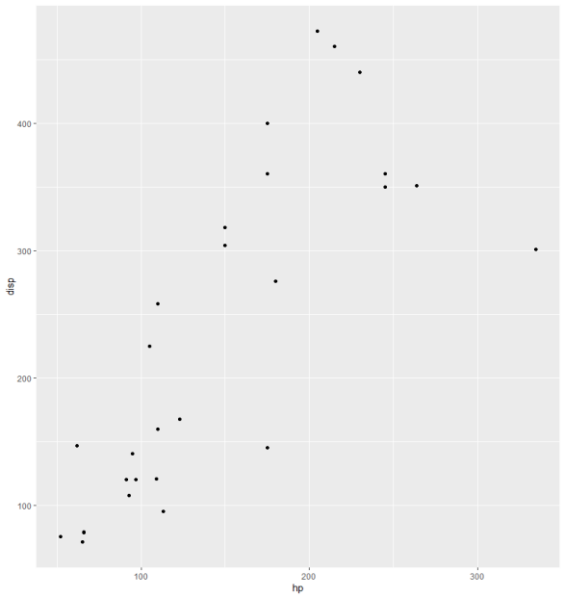


Γράφημα 4.1

Από το παραπάνω διάγραμμα διασποράς (Γράφημα 4.1) προκύπτει ότι τα οχήματα με υψηλούς ίππους, διανύουν λιγότερα μίλια ανά γαλόνι καυσίμου, και άρα έχουν υψηλότερη κατανάλωση, σε σχέση με τα οχήματα με χαμηλούς ίππους.



Γράφημα4.2 : διάγρ. διασποράς wg-mpg



Γράφημα4.3: διάγρ. Διασποράς hp-mpg

Το γράφημα 4.2 παρουσιάζει το διάγραμμα διασποράς των μεταβλητών του βάρους και της μεταβλητής mpg, όπου παρατηρείται ότι όσο αυξάνει το βάρος, τόσο μειώνεται και η απόσταση που διανύει κάποιο όχημα με την ίδια ποσότητα καυσίμου.

Τέλος, στο γράφημα 4.3 απεικονίζεται διάγραμμα διασποράς των μεταβλητών hp και disp, όπου προκύπτει το συμπέρασμα ότι τα οχήματα με περισσότερους ίππους έχουν μεγαλύτερο εκτόπισμα από τα οχήματα με λιγότερους ίππους.

Αντίστοιχα παρουσιάζουμε κάποια περιγραφικά στατιστικά για τα δεδομένα:

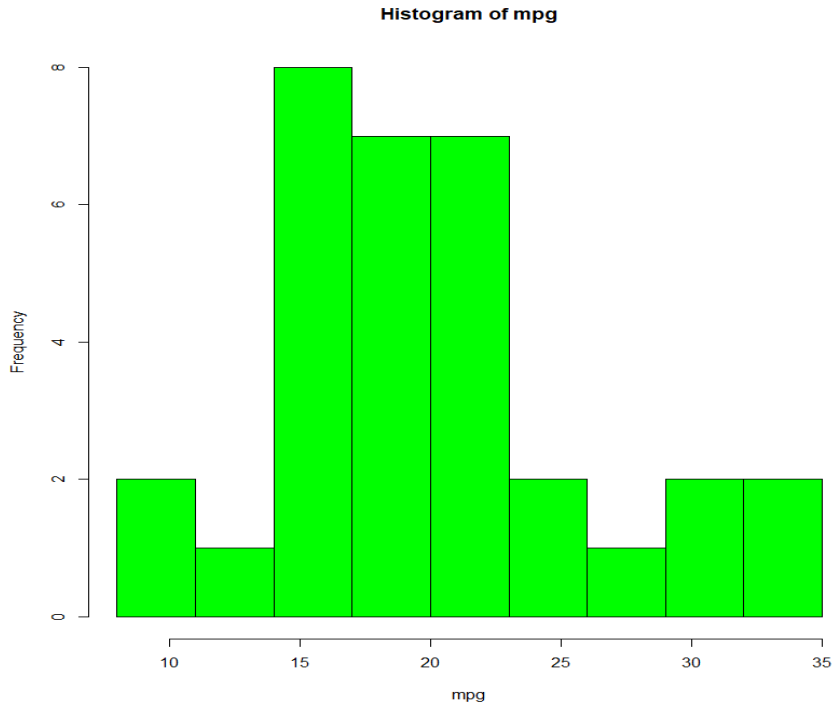
mpg	cyl	disp	hp	drat	wt
Min. :10.40	Min. :4.000	Min. : 71.1	Min. : 52.0	Min. :2.760	Min. :1.513
1st Qu.:15.43	1st Qu.:4.000	1st Qu.:120.8	1st Qu.: 96.5	1st Qu.:3.080	1st Qu.:2.581
Median :19.20	Median :6.000	Median :196.3	Median :123.0	Median :3.695	Median :3.325
Mean :20.09	Mean :6.188	Mean :230.7	Mean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

Πίνακας 4.2 : Περιγραφικά στατιστικά δεδομένων (mtcars)

Αξίζει να σημειωθεί ότι οι μεταβλητές παρουσιάζουν σημαντικές διαφορές στις τιμές τους. Η μεταβλητή μέτρησης της κατανάλωσης του οχήματος (mpg) παίρνει τιμές από 10,40 έως 33,90 επομένως υπάρχουν κατηγορίες οχημάτων με εντελώς διαφορετικό επίπεδο κατανάλωσης.

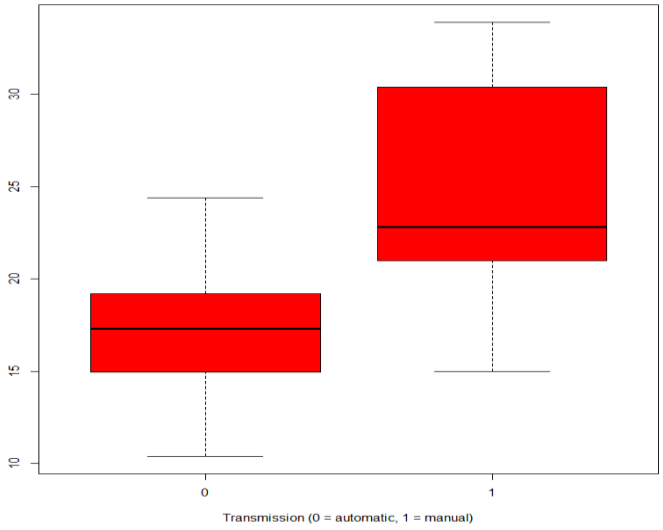
Η σημαντική αυτή διαφορά στην κατανάλωση ενδεχομένως να οφείλεται στο μεγάλο εύρος τιμών που παρατηρείται στην ιπποδύναμη των οχημάτων (52hp-335hp), στις διαφορές που παρουσιάζονται στο βάρος των οχημάτων (1513kg-5.425kg) ή στον τρόπο μετάδοσης κίνησης (am), αν δηλαδή το αυτοκίνητο είναι αυτόματο ή με μηχανικό κιβώτιο ταχυτήτων.

Παρουσιάζονται επίσης κάποια γραφήματα που αφορούν την μεταβλητή mpg.



Γράφημα4.4 Ιστόγραμμα συχνοτήτων mpg

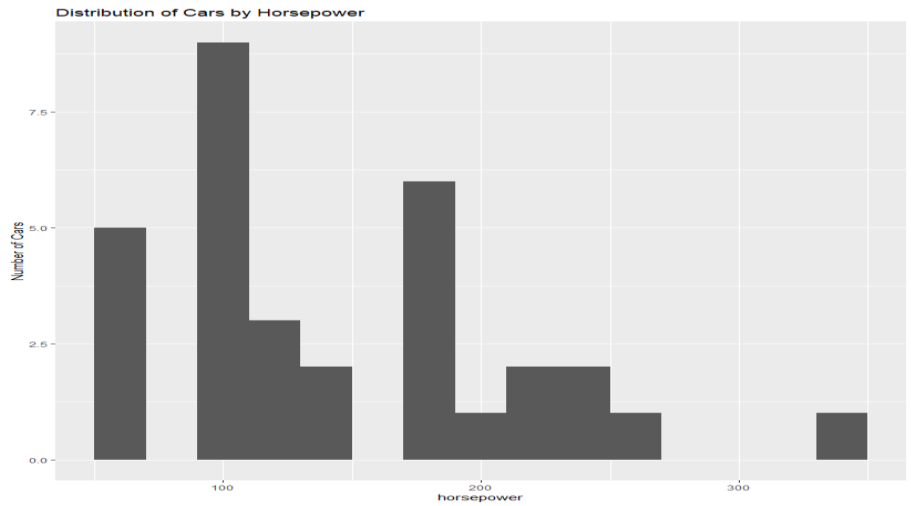
Παρατηρούμε ότι η μεταβλητή mpgπαρουσιάζει αισθητά μεγαλύτερη συχνότητα μεταξύ των τιμών 15 και 23, σε σχέση με τις άλλες κλάσεις στο διάγραμμα.



Γράφημα4.5: Θηκόγραμμα mpg ανά κατηγορία am

Από το παραπάνω θηκόγραμμα είναι φανερό ότι τα οχήματα με αυτόματο σύστημα εναλλαγής ταχυτήτων έχουν σαφώς μεγαλύτερη κατανάλωση από τα μηχανικά.

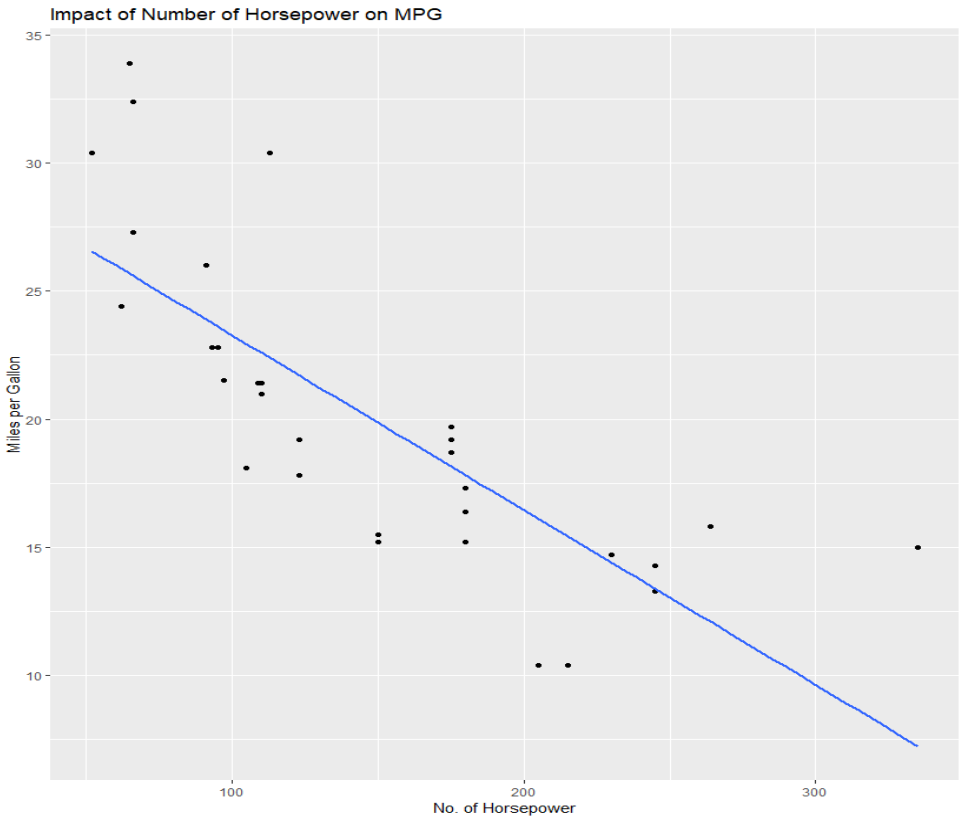
Στην συνέχεια ακολουθεί το ιστόγραμμα συχνοτήτων, που απεικονίζει το πλήθος των αυτοκινήτων σε κάθε κατηγορία ιπποδύναμης. Όπως φαίνεται παρακάτω η κατηγορία των 100 ίππων είναι αυτή που συγκεντρώνει τα περισσότερα οχήματα ενώ ακολουθεί η κατηγορία με τους 180 ίππους.



Γράφημα4.6

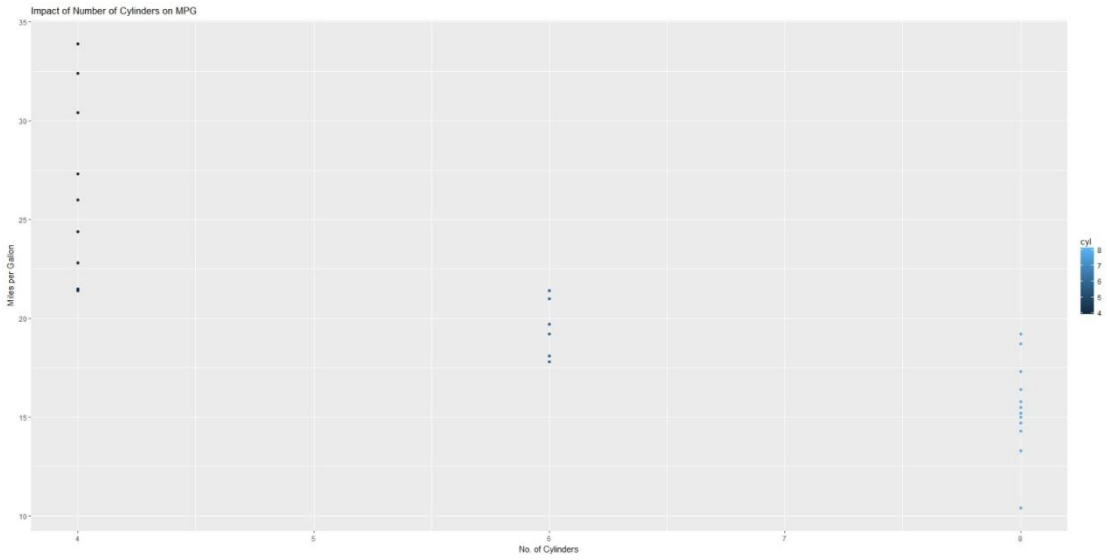
Αξίζει να σημειωθεί ότι τα κενά στο παραπάνω ιστόγραμμα οφείλονται στο μικρό πλήθος των δεδομένων που αναλύονται, καθώς όπως προκύπτει δεν υπάρχουν οχήματα στις συγκεκριμένες κατηγορίες ίππων.

Παρακάτω απεικονίζεται το διάγραμμα διασποράς της κατανάλωσης του οχήματος σε σχέση με το πλήθος των ίππων που διαθέτει το αυτοκίνητο. Στο γράφημα αυτό έχει σχεδιαστεί και η ευθεία ελαχίστων τετραγώνων η οποία μας βοηθάει να συμπεράνουμε ότι υπάρχει αρνητική σχέση μεταξύ των δύο μεταβλητών, η οποία όμως δεν φαίνεται να είναι ισχυρή.



Γράφημα4.7: ευθεία ελ. τετραγώνων hp-mpg

Επιπρόσθετα από το παραπάνω γράφημα προκύπτει το συμπέρασμα ότι όσο αυξάνονται οι ίπποι του οχήματος τόσο μικρότερη απόσταση διανύει το όχημα με την ίδια ποσότητα καυσίμου.



Γράφημα4.8

Τέλος, από το παραπάνω διάγραμμα γίνεται έντονα αντιληπτό ότι τα οχήματα με 8 κυλίνδρους έχουν την μεγαλύτερη κατανάλωση ενώ ακολουθούν τα αυτοκίνητα με 6 κυλίνδρους και τέλος τα αυτοκίνητα με 4 κυλίνδρους.

4.3 Γραμμική Παλινδρόμηση

Στην συνέχεια γίνεται χρήση του μοντέλου ανάλυσης διακύμανσης, αφού η επεξηγηματική μεταβλητή που εισάγεται αρχικά είναι κατηγορική, στο οποίο η μεταβλητή απόκρισης είναι η κατανάλωση καυσίμου και επεξηγηματική μεταβλητή η μεταβλητή amη οποία έχει δύο επίπεδα, αυτόματο – μηχανικό κιβώτιο ταχυτήτων.

Estimate Std. Error	t value	Pr(>|t|)
(Intercept) 17.151.12	15.25	0,00
factor(am)manual 7.24 1.76	4.11	0,00

Πίνακας 4.3

One way anova model

Η ερμηνεία των αποτελεσμάτων του παραπάνω μοντέλου ανάλυσης διακύμανσης (ANOVA) είναι όμοια με του μοντέλου της απλής γραμμικής παλινδρόμησης. Όπως παρατηρούμε παραπάνω η μεταβλητή am κρίνεται σημαντική σε κάθε επίπεδο σημαντικότητας αφού p-value < 0,01.

Στην συνέχεια παρουσιάζεται το πλήρες μοντέλο το οποίο περιλαμβάνει σαν επεξηγηματικές μεταβλητές όλες τις διαθέσιμες μεταβλητές των δεδομένων.

Μοντέλο πολυμεταβλητής γραμμικής παλινδρόμησης

	Estimate	***Std. Error***	***t value***	***Pr(>|t|)***
(Intercept)	12,30	18,72	0,66	0,52
cyl	-0,11	1,05	-0,11	0,92
disp	0,01	0,02	0,75	0,46
hp	-0,02	0,02	-0,99	0,33
drat	0,79	1,64	0,48	0,64
wt	-3,72	1,89	-1,96	0,06
qsec	0,82	0,73	1,12	0,27
vfactor(vs)1	0,32	2,10	0,15	0,88
factor(am)1	2,52	2,06	1,23	0,23
gear	0,66	1,49	0,44	0,67
carb	-0,20	0,83	-0,24	0,81

Πίνακας4.4

Από τα παραπάνω αποτελέσματα παρατηρείται ότι στο στάδιο της απλής γραμμικής παλινδρόμησης, όπου η μεταβλητή am είναι η μοναδική επεξηγηματική μεταβλητή, η μεταβλητή αυτή είναι στατιστικά σημαντική. Με την είσοδο επιπλέον επεξηγηματικών μεταβλητών η am δεν κρίνεται πλέον σημαντική, επομένως ο αναλυτής καταλήγει στο συμπέρασμα ότι πιθανόν να υπάρχει γραμμική εξάρτηση μεταξύ της μεταβλητής με κάποιες από τις υπόλοιπες μεταβλητές του μοντέλου.

4.3.1 Έλεγχος πολυσυγγραμμικότητας

Στην συνέχεια ακολουθεί ο έλεγχος πολυσυγγραμμικότητας των δεδομένων. Μεταβλητές οι οποίες είναι γραμμικά εξαρτημένες με κάποιες άλλες θα πρέπει να αφαιρεθούν από το μοντέλο, καθώς επηρεάζουν αρνητικά τα αποτελέσματα.

	VIF
CYL	15.373833
DISP	21.620241
HP	9.832037
DRAT	3.374620
WT	15.164887
QSEC	7.527958
FACTOR(VS)	4.965873
FACTOR(AM)	4.648487
GEAR	5.357452
CARB	7.908747

Πίνακας4.5

Οι μεταβλητές CYL, DISP και WT έχουν εξαιρετικά υψηλές τιμές, καθώς οι τιμές για το VIF που είναι μεγαλύτερες από 10 θεωρούνται μεγάλες. Θα πρέπει επίσης να δώσουμε προσοχή στις τιμές VIF μεταξύ 5 και 10. Σε αυτό το σημείο μπορεί να εξεταστεί το ενδεχόμενο να αφήσουμε μόνο μία από αυτές τις μεταβλητές στο μοντέλο.

Για την εύρεση του βέλτιστου μοντέλου θα εκτελεστεί η μέθοδος stepwise regression. Τα αποτελέσματα είναι τα ακόλουθα.

Stepwise selection method

	Estimate	***Std. Error***	***t value***	***Pr(>|t|)***

(intercept)	9,618	6,960	1,382	0,178
wt	-3,917	0,711	-5,507	0,000
qsec	1,226	0,289	4,247	0,000
factor(am)1	2,936	1,411	2,081	0,047

Πίνακας4.6

Η τιμή του συντελεστή προσαρμογής για το παραπάνω μοντέλο είναι περίπου $R^2 = 0.85$

Όπως φαίνεται επομένως, το αποτέλεσμα είναι συνεπές με το μοντέλο της stepwise παλινδρόμησης και η προσθήκη οποιασδήποτε επιπλέον μεταβλητής εκτός από τα wt, am και qsec θα αυξήσει δραματικά την παραλλαγή στο μοντέλο και η τιμή p αμέσως θα γίνει ασήμαντη.

Ένας ακόμα τρόπος για να καταλήξουμε στο βέλτιστο μοντέλο παλινδρόμησης είναι δημιουργία πολλαπλών μοντέλων στα οποία κάθε φορά εισέρχεται μία νέα μεταβλητή.

```
fit1 <- lm(mpg ~ factor(am), data = mtcars)
fit2 <- lm(mpg ~ factor(am) + wt, data = mtcars)
fit3 <- lm(mpg ~ factor(am) + wt + qsec, data = mtcars)
fit4 <- lm(mpg ~ factor(am) + wt + qsec + hp, data = mtcars)
fit5 <- lm(mpg ~ factor(am) + wt + qsec + hp + drat, data = mtcars)
```

Στη συνέχεια χρησιμοποιώντας ανάλυση διακύμανσης για το μοντέλο fit5 που περιέχει όλες τις μεταβλητές προκύπτουν τα ακόλουθα αποτελέσματα.

	Df	***Sum Sq***	***Mean Sq***	***F value***	***Pr(>F)***

factor(am)	1	405,15	405,15	66,402	0,00
wt	1	442,58	442,58	72,536	0,00
qsec	1	109,03	109,03	17,870	0,00
hp	1	9,22	9,22	1,511	0,23
drat	1	1,43	1,43	0,234	0,63
Residuals	26	158,64	6,10		

Πίνακας4.7

Συνεπώς, η εισαγωγή των μεταβλητών wt, am και qsec κρίνεται στατιστικά σημαντική επιβεβαιώνοντας τα αποτελέσματα της stepwiseμεθόδου.

4.3.2 Τελικό μοντέλο

Το τελικό μοντέλο στο οποίο καταλήγουμε και ο αντίστοιχος πίνακας των συντελεστών δίνονται παρακάτω.

	Estimate	***Std. Error***	***t value***	***Pr(>|t|)***
(Intercept)	9,618	6,960	1,382	0,178
wt	-3,917	0,711	-5,507	0,000
qsec	1,226	0,289	4,247	0,000
factor(am)1	2,936	1,411	2,081	0,047

Πίνακας4.8

$$Mpg = 9,618 – 3,917*wt + 1,226*qsec + 2,936*am$$

Μπορεί να παρατηρηθεί ότι όλες οι μεταβλητές είναι στατιστικά σημαντικές. Το μοντέλο αυτό εξηγεί περίπου το 85% της διακύμανσης σε μίλια ανά γαλόνι (mpg). Τώρα, όταν διαβάζουμε τον συντελεστή για το am, λέμε ότι, κατά μέσο όρο, τα μηχανικά κιβώτια ταχυτήτων διανύουν περίπου 2,94 (MPG) περισσότερα μίλια ανά γαλόνι καυσίμου, από τα αυτόματα κιβώτια ταχυτήτων. Ωστόσο, αυτή η επίδραση ήταν πολύ υψηλότερη όταν δεν ρυθμίσαμε το βάρος και το qsec.

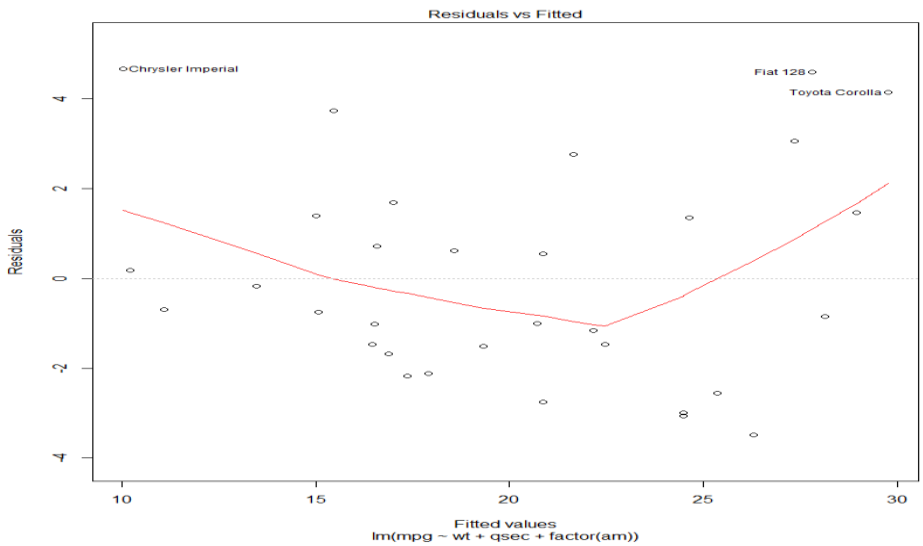
4.3.3 Διαγνωστικές Μέθοδοι για το Μοντέλο Παλινδρόμησης

Εκτελώντας ακόμη μία φορά τον έλεγχο πολυσυγγραμμικότητας για το τελικό μοντέλο, παρατηρούμε ότι όλες οι τιμές του δείκτη VIFείναι αρκετά χαμηλές (<5) και επομένως δεν υπάρχει γραμμική εξάρτηση μεταξύ των μεταβλητών.

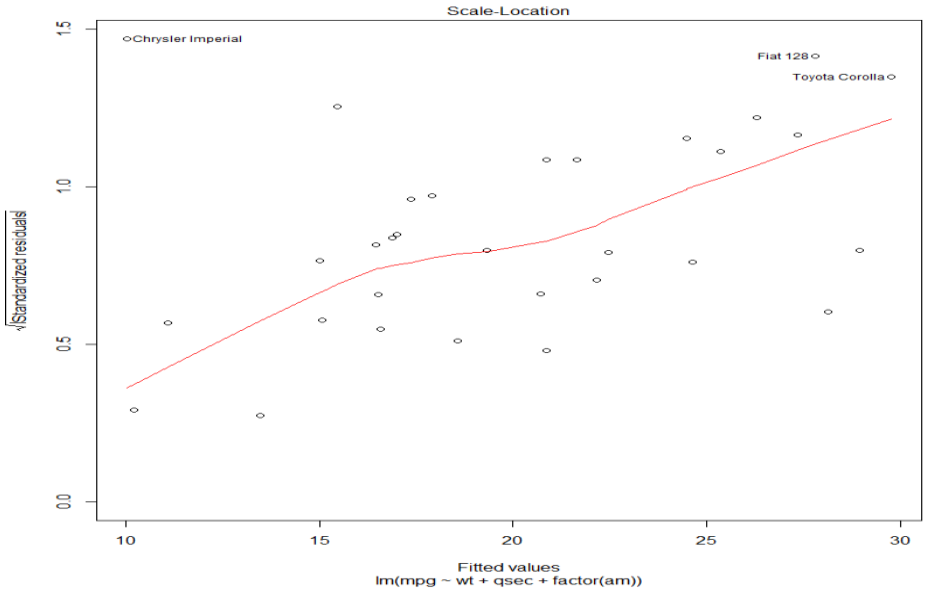
	Vif
wt	2,48
qsec	1,36
factor(am)	2,54

Πίνακας4.9

Σχεδιάζοντας τα κατάλοιπα έναντι των προσαρμοσμένων τιμών, ψάχνουμε για οποιοδήποτε είδος μοτίβου. Το ίδιο πράγμα με τις προσαρμοσμένες τιμές έναντι των τυποποιημένων, όπου σχεδιάζεται μια συνάρτηση των τυποποιημένων καταλοίπων. Τα παρακάτω διαγράμματα δείχνουν ότι δεν υπάρχουν συγκεκριμένα μοτίβα στα κατάλοιπα.

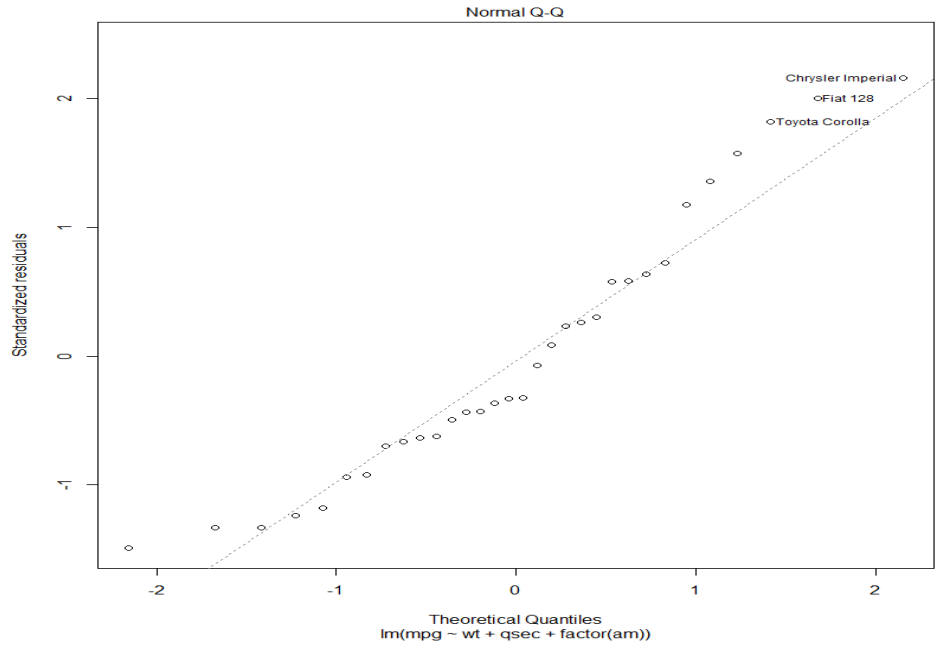


Γράφημα4.9 : διάγραμμα διασποράς καταλοίπων



Γράφημα 4.10: διάγραμμα διασποράς τυποποιημένων καταλοίπων

Το Q-Q plot, προσπαθεί να ελέγξει εικονικά την κανονικότητα των σφαλμάτων, σχεδιάζοντας τα θεωρητικά μεγέθη της κανονικής κατανομής από τα τυποποιημένα κατάλοιπα:



Γράφημα4.11 : διάγραμμα κανονικότητας σφαλμάτων

Όπως προκύπτει από το παραπάνω γράφημα τα σφάλματα φαίνεται να ακολουθούν την κανονική κατανομή.

4.4 Insurance Data

Στη συνέχεια της ανάλυσης στην παρούσα εργασία, θα γίνει χρήση του πακέτου Insurance data το οποίο είναι διαθέσιμο για λήψη μέσω του ιστότοπου CRAN. Περιλαμβάνει δεδομένα από το συμβούλιο Ασφαλιστικών Ερευνών (IRC) της Αμερικής. Τα δεδομένα συλλέχθηκαν το 2004 και το 2005 και περιέχουν πληροφορίες σχετικά με τα δημογραφικά χαρακτηριστικά του ατόμου που διεκδικεί κάποια αποζημίωση, την ύπαρξη clm (0=όχι, 1=ναι) και το ύψος της οικονομικής αποζημίωσης claimcst0 (σε $), την αξία του οχήματος veh_value (σε 10.000$), το είδος του οχήματος veh_body, το εκτόπισμα του οχήματος disp, την ηλικία του οχήματος (1=youngest, 2, 3,4), το φύλο του οδηγού (M=male, F=female), περιοχή κατοικίας του οδηγού (A, B,.., F) και την κατηγορία της ηλικίας του οδηγού (1=youngest,2,…6).

Η πλήρης μελέτη που πραγματοποιήθηκε περιέχει διεκδικήσεις αποζημιώσεων από 68.856 άτομα, βασισμένες σε τριάντα δύο ασφαλιστές.

4.4.1 Περιγραφική ανάλυση

Για την κατανόηση των δεδομένων, την επιλογή του τρόπου ανάλυσης και την ορθότερη εξαγωγή συμπερασμάτων κατά την στατιστική ανάλυση που θα ακολουθήσει, θα πρέπει αρχικά να παρουσιαστούν μερικά περιγραφικά στοιχεία.

Ακολουθεί ο πίνακας με την μέση τιμή αποζημιώσεων ανά τύπο οχήματος.

Veh_body	mean_claims
BUS	1.485 €
CONVT	2.296 €
COUPE	2.761 €

HBACK	2.048 €
HDTOP	2.268 €
MCARA	762 €
MIBUS	2.700 €
PANVN	2.147 €
RDSTR	685 €
SEDAN	1.817 €
STNWG	2.015 €
TRUCK	2.662 €
UTE	2.297 €

Πίνακας4.10

Παρατηρούμε ότι η μεγαλύτερη μέση αποζημίωση παρουσιάζεται σε οχήματα τύπου COUPE ενώ η μικρότερη σε οχήματα τύπου RDSTR.

Στη συνέχεια ακολουθούν οι πίνακες με την μέση τιμή αποζημίωσης ανά ηλικία οχήματος και φύλο.

veh_age	mean_claims
1	1.885 €
2	1.975 €
3	1.996 €
4	2.169 €

gender	mean claims
F	1.854 €
M	2.230 €

Πίνακες 4.11(μ.τ. ανά ηλικία) & 4.12(μ.τ. ανά φύλο)

Παρατηρούμε ότι τα παλαιότερα αυτοκίνητα καθώς και οι άντρες έχουν υψηλότερες τιμές αποζημιώσεων.

Στον παρακάτω πίνακα παρουσιάζονται μερικά από τα πιο χαρακτηριστικά περιγραφικά στοιχεία των αποζημιώσεων, έχοντας αφαιρέσει από τα δεδομένα τις μηδενικές τιμές της μεταβλητής claimcst0.

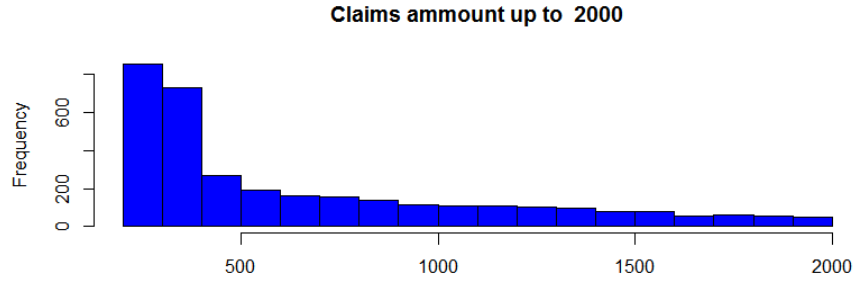
Αποζημιώσεις

Min.	200
1st Qu.	353.8
Median	761.6
Mean	2014.4
3rd Qu.	2091.4
Max.	55922.1

Πίνακας4.13

Παρατηρούμε ότι η μέση τιμή των δεδομένων είναι 2014.4, η ελάχιστη τιμή είναι 200 ενώ η μέγιστη 55.922,1 . Επομένως τα δεδομένα παρουσιάζουν ένα τεράστιο εύρος τιμών.

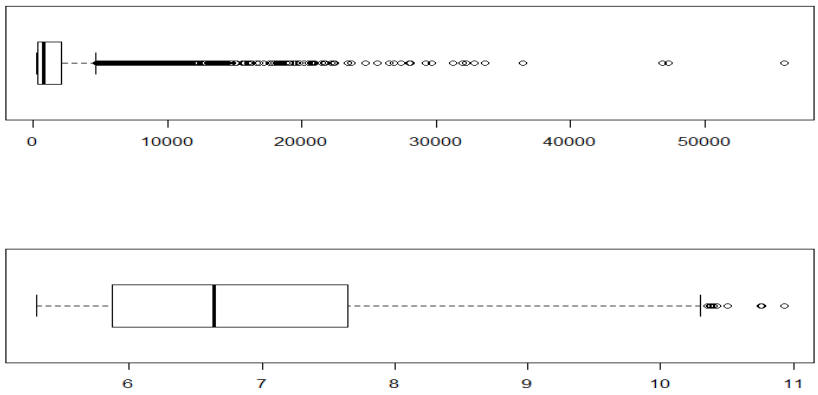
Στη συνέχεια θα ήταν χρήσιμο ένα ιστόγραμμα συχνοτήτων για να φανεί γραφικά αν η κατανομή του πλήθους των αποζημιώσεων, μέχρι 2000$, για τα δεδομένα θυμίζει κάποια γνωστή κατανομή. Η τιμή 2000$ επελέγη, καθώς όπως προκύπτει από τον πίνακα 4.13 η τιμή αυτή βρίσκεται κοντά στον μέσο όρο, καθώς επίσης και στην τιμή του τρίτου ποσοστημορίου των αποζημιώσεων.



Γράφημα4.12 : Ιστόγραμμα αποζημιώσεων έως 2000

Όπως φαίνεται στο παραπάνω ιστόγραμμα οι τιμές των θετικών ασφαλιστικών αποζημιώσεων δεν ακολουθούν κάποια γνωστή κατανομή.

Στη συνέχεια, ακολουθεί το διάγραμμα boxplot για τις πραγματικές θετικές τιμές και για τον λογάριθμο των τιμών των αποζημιώσεων, το οποίο δίνει μια ακόμη εικόνα για την κατανομή των δεδομένων.



Γράφημα4.13: θηκόγραμμα αποζημιώσεων

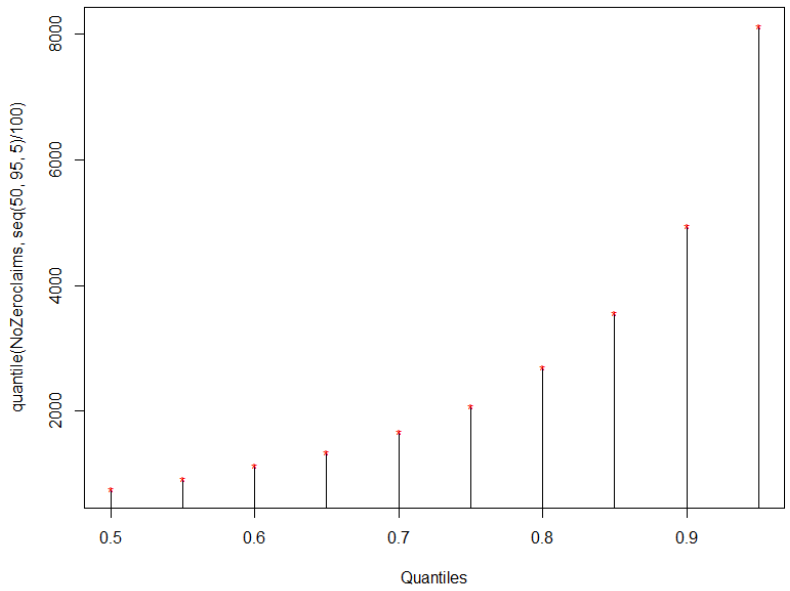
Από το θηκόγραμμα στις θετικές τιμές των ασφαλιστικών αποζημιώσεων, προκύπτει ότι οι περισσότερες από τις αποζημιώσεις παίρνουν τιμές μικρότερες από 5000€, καθώς επίσης υπάρχουν πολλές ακραίες υψηλές τιμές (βαριά δεξιά ουρά).

Το θηκόγραμμα για τον λογάριθμο των αποζημιώσεων οδηγεί σε πιο χρήσιμα αποτελέσματα καθώς έχει σαφώς μικρότερη δεξιά ουρά και απεικονίζει καλύτερα τα δεδομένα.

50%	55%	60%	65%	70%	75%	80%	85%	90%	95%
761.565	930.8815	1146.372	1362.2315	1679.271	2091.425	2714.7322	3565.1805	4952.976	8124.6605

Πίνακας4.14 : Ποσοστημόρια θετικών αποζημιώσεων

Από τον παραπάνω πίνακα με τα ποσοστημόρια των τιμών των αποζημιώσεων, καθώς και από το γράφημα που ακολουθεί, παρατηρείται ότι το 50% των αποζημιώσεων είναι μεγαλύτερες από 762€ και το 95% των αποζημιώσεων δεν υπερβαίνουν τα 8.126€.



Γράφημα4.14

Στη συνέχεια παρουσιάζεται ο πίνακας με την μέση τιμή, την διάμεσο, την τυπική απόκλιση την ελάχιστη και μέγιστη τιμή αποζημιώσεων ανά περιοχή και φύλο.

gender	area	mean_claims	median_claims	sd_claims	min_claims	max_claims
F	A	1786.	729.	3187.	200.	47297.
F	B	1725.	760.	2572.	200.	18291.
F	C	1889.	720.	2980.	200.	28048.
F	D	1901.	783.	3160.	200.	23405.
F	E	1829.	705.	2526.	200.	14199.
F	F	2303.	793.	4107.	200.	36502.
M	A	2066.	868.	4096.	200.	55922.
M	B	2030.	793.	3416.	200.	33642.
M	C	2230.	756.	3922.	200.	31244.
M	D	1741.	703.	2811.	200.	22405.
M	E	2792.	867.	5000.	200.	32196.
M	F	3612.	1108.	7038.	200.	46868.

Πίνακας4.15: περιγραφικά στατιστικά mpg ανά φύλο και περιοχή

Όπως φαίνεται παραπάνω για τις γυναίκες αλλά και για τους άντρες, οι υψηλότερες κατά μέσο όρο αποζημιώσεις και ταυτόχρονα με την μεγαλύτερη τυπική απόκλιση, παρουσιάζονται στην περιοχή F.

Επιπλέον, αξίζει να σημειωθεί ότι, τόσο για τους άντρες όσο και για τις γυναίκες, η περιοχή A παρουσιάζει αρκετά υψηλή τυπική απόκλιση στις παρατηρήσεις. Επίσης στην περιοχή αυτή, παρουσιάζεται και η μέγιστη τιμή αποζημιώσεων, και για τα δύο φύλα. Έτσι παρότι οι μέσες τιμές αποζημιώσεων στην περιοχή αυτή δεν είναι ιδιαιτέρως υψηλές, υπάρχει ο κίνδυνος ακραίων τιμών και επομένως πρέπει να ληφθεί υπόψη από την ασφαλιστική εταιρεία.

4.4.2 Λογιστική παλινδρόμηση

Στη συνέχεια με την χρήση της μεθόδου της λογιστικής παλινδρόμησης η ασφαλιστική εταιρεία επιθυμεί να εκτιμήσει την πιθανότητα ύπαρξης ασφαλιστικής αποζημίωσης με βάση κάποια χαρακτηριστικά.

Ακολουθούν τα αποτελέσματα λογιστικής παλινδρόμησης με μεταβλητή απόκρισης την μεταβλητή clm η οποία εκφράζει την ύπαρξη ασφαλιστικής αποζημίωσης (0=όχι, 1=ναι) και επεξηγηματικές μεταβλητές τις μεταβλητές veh_value (αξία οχήματος σε $10.000s), το τετράγωνο της veh_value, veh_age (ηλικία του οχήματος), gender (φύλο οδηγού), area (περιοχή διαμονής: A,B,C,D,E,F), και τέλος την agecat (κατηγορία ηλικίας οδηγού :1=youngest,2,3,4,5,6).

```
Call:
glm(formula = clm ~ veh_value + I(veh_value^2) + veh_age + gender +
    area + agecat, family = binomial, data = na.omit(car))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4790  -0.3940  -0.3707  -0.3472   3.1711
```

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.729916   0.100536 -27.154  < 2e-16 ***
veh_value       0.242279   0.042289   5.729 1.01e-08 ***
I(veh_value^2) -0.027733   0.006328  -4.383 1.17e-05 ***
veh_age         0.026540   0.018547   1.431   0.1524
genderM        -0.034435   0.031439  -1.095   0.2734
areaB           0.092845   0.045938   2.021   0.0433 *
areaC           0.036096   0.041864   0.862   0.3886
areaD          -0.115818   0.056185  -2.061   0.0393 *
areaE          -0.054056   0.061643  -0.877   0.3805
areaF           0.048069   0.071287   0.674   0.5001
agecat         -0.078023   0.010887  -7.166 7.70e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33767  on 67855  degrees of freedom
Residual deviance: 33639  on 67845  degrees of freedom
AIC: 33661

Number of Fisher Scoring iterations: 6
```

Από τα παραπάνω αποτελέσματα προκύπτει ότι δεν είναι όλες οι μεταβλητές στατιστικά σημαντικές αφού οι τιμές p-value σε κάποιες περιπτώσεις είναι αρκετά υψηλές.

Με τη χρήση της εντολής steptο μοντέλο το οποίο περιέχει μόνο τις στατιστικά σημαντικές μεταβλητές είναι το εξής:

```
Call:
glm(formula = clm ~ veh_value + I(veh_value^2) + agecat, family
= binomial,
    data = na.omit(car))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4478  -0.3930  -0.3716  -0.3488   3.1185

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.598438   0.058460 -44.448  < 2e-16 ***
veh_value       0.204228   0.035560   5.743 9.29e-09 ***
I(veh_value^2) -0.024289   0.005854  -4.149 3.34e-05 ***
agecat         -0.080451   0.010805  -7.446 9.63e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33767  on 67855  degrees of freedom
Residual deviance: 33657  on 67852  degrees of freedom
AIC: 33665
```

```
Number of Fisher Scoring iterations: 6
```

Επομένως το μοντέλο έχει τη μορφή

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -2{,}60 + 0{,}20*\text{veh\_value} - 0{,}02*\text{I(veh\_value\^2)} - 0{,}08*\text{agecat}$$

ή ισοδύναμα,

$$\frac{\hat{p}}{1-\hat{p}} = e^{-2{,}60} e^{0{,}20\ \text{veh\_value}}\ e^{-\ 0{,}02\ \text{I(veh\_value\^2)}}\ e^{-\ 0.08\ \text{agecat}}$$

Σύμφωνα με τα παραπάνω μοντέλο, η αύξηση της τιμής ενός οχήματος κατά 10.000$ αυξάνει πολλαπλασιαστικά την εκτιμώμενη σχετική πιθανότητα εμφάνισης ασφαλιστικής αποζημίωσης κατά $e^{0,20} = 1,221$φορές.

Ομοίως η αύξηση της μεταβλητής agecat κατά ένα επίπεδο μειώνει πολλαπλασιαστικά την εκτιμώμενη σχετική πιθανότητα εμφάνισης αποζημίωσης κατά $e^{-\ 0,08} = 0.923$φορές.

4.4.3 Αντασφάλιση Excess of loss

Στην συνέχεια ακολουθεί η ανάλυση για την εύρεση του ποσού των αποζημιώσεων πάνω από το οποίο η ασφαλιστική εταιρεία έχει συμφέρον να αντασφαλιστεί.

Μελετώντας τα δεδομένα στα οποία οι αποζημιώσεις έχουν θετική τιμή προκύπτει ότι το a = 75% των αποζημιώσεων έχουν ύψος έως d = 2.091,4€.

Πιο συγκεκριμένα, από τις συνολικά 4.624 περιπτώσεις όπου δίδεται αποζημίωση μόνο οι 1.156 (25%) ξεπερνούν σε μέγεθος τα 2.091€.

Ας υποθέσουμε τώρα ότι από το σημείο αυτό (d) και πέρα, η ασφαλιστική εταιρεία αποφασίζει να συνάψει σύμβαση αντασφάλισης, καθώς αντιλαμβάνεται ότι πρόκειται για ακραίες περιπτώσεις αποζημιώσεων όπου τα κόστη μπορούν προκαλέσουν ζημιά στην ρευστότητά της. Επομένως για ενδεχόμενη αποζημίωση x, η ασφαλιστική εταιρεία θα πληρώνει το σύνολο της ζημιάς αν αυτή είναι μικρότερη ή ίση από το σημείο d, ενώ η αποζημιώσεις που ξεπερνούν το όριο αυτό θα καλύπτονται από την αντασφαλιστική εταιρεία.

$$Y=\begin{cases} x, & x \le d \\ d, & x > d \end{cases}$$

Στη συνέχεια, δημιουργώντας την νέα κατηγορική μεταβλητή paynopay η οποία παίρνει δύο τιμές, την τιμή 1 όταν το ποσό της αποζημίωσης είναι μικρότερο από 2.091,4€ και την τιμή 0 όταν το ποσό της αποζημίωσης ξεπεράσει το σημείο αυτό, μπορεί να προκύψει ένα νέο μοντέλο λογιστικής παλινδρόμησης.

Σύμφωνα με το μοντέλο αυτό το οποίο έχει σαν μεταβλητή απόκρισης την δίτιμη μεταβλητή paynopay, η ασφαλιστική εταιρεία θα είναι σε θέση να γνωρίζει ποιες μεταβλητές επηρεάζουν σημαντικά την χρήση αντασφάλισης στις αποζημιώσεις που θα προκύψουν.

Ακολουθεί το μοντέλο το οποίο περιέχει μόνο τις ανεξάρτητες μεταβλητές gender, agecat, valuecat και veh_age οι οποίες κρίνονται στατιστικά σημαντικές.

```
Call:
glm(formula = payNOpay ~ gender + agecat + valuecat + veh_age,
    family = binomial, data = na.omit(car4))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9469   0.4078   0.7107   0.7906   0.9492

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        1.21574    0.13465   9.029  < 2e-16 ***
genderM           -0.12479    0.06971  -1.790   0.0734 .
```

```
agecat                 0.10118    0.02436   4.153 3.28e-05 ***
valuecat(2.5,5]        0.16790    0.09761   1.720   0.0854 .
valuecat(5,7.5]       -0.11601    0.25439  -0.456   0.6484
valuecat(7.5,10]      -0.10429    1.16231  -0.090   0.9285
valuecat(10,12.5]    -13.68775  229.39979  -0.060   0.9524
valuecat(12.5,100]   -13.97353  324.74370  -0.043   0.9657
veh_age               -0.15709    0.03576  -4.392 1.12e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5200.5  on 4623  degrees of freedom
Residual deviance: 5141.4  on 4615  degrees of freedom
AIC: 5159.4

Number of Fisher Scoring iterations: 11
```

Τέλος, πραγματοποιώντας διαδοχικές παλινδρομήσεις για διάφορες τιμές του ποσοστού a που η ασφαλιστική ενδέχεται να επιλέξει ως όριο (κατώφλι), παρουσιάζεται ο πίνακας με τις μεταβλητές που κρίνονται στατιστικά σημαντικές (σε επίπεδο σημαντικότητας 5%). Υπενθυμίζεται εδώ ότι το αείναι ουσιαστικά το (κάτω) ποσοστημόριο της κατανομής για τα μεγέθη των αποζημιώσεων.

a	d	veh_body	veh_age	gender	area	agecat	valuecat
40%	500€						
45%	622,49€						
50%	761,57€		☻				
55%	930,88€		☻			☻	
60%	1.146,4€		☻			☻	
65%	1.362,2€		☻			☻	
70%	1.679,3€		☻			☻	
75%	2.091,4€		☻	☻		☻	☻
80%	2.714,7€		☻	☻	☻	☻	
85%	3.565,2€		☻	☻	☻	☻	☻
90%	4.953€		☻	☻	☻	☻	☻

Για παράδειγμα αν η ασφαλιστική εταιρεία επιθυμεί να καλύψει αποζημιώσεις πάνω από τα 1679,3€, δηλαδή το 30% των αποζημιώσεών της, τότε στατιστικά

σημαντικές είναι οι μεταβλητές ηλικία του οχήματος και η κατηγορία της ηλικίας του οδηγού του οχήματος. Ακολουθούν τα αποτελέσματα του παραπάνω παραδείγματος.

```
Call:
glm(formula = payNOpay ~ veh_age + agecat, family = binomial,
    data = na.omit(car4))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7949  -1.4469   0.7987   0.8740   1.0126

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.89882    0.11428   7.865 3.69e-15 ***
veh_age     -0.15117    0.03099  -4.878 1.07e-06 ***
agecat       0.10675    0.02299   4.643 3.43e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5648.9  on 4623  degrees of freedom
Residual deviance: 5604.5  on 4621  degrees of freedom
AIC: 5610.5

Number of Fisher Scoring iterations: 4
```

Όπως προκύπτει από τα αντίστοιχα p-valueοι δύο μεταβλητές veh_age και agecat είναι στατιστικά σημαντικές (με πολύ μικρή τιμή για το p-value), επομένως θεωρούνται εξαιρετικά χρήσιμες για την πρόβλεψη χρήσης αντασφάλισης της ασφαλιστικής εταιρείας.

Αξίζει να σημειωθεί ότι η μεταβλητή της ηλικίας του οχήματος veh_age και η κατηγορία της ηλικίας του οδηγού agecat κρίνονται στατιστικά σημαντικές για κάθε ποσοστό a $\geq$ 55%.

Τέλος παρατηρούμε ότι όσο αυξάνεται το ποσοστό a τόσο περισσότερες μεταβλητές κρίνονται σημαντικές για την πρόβλεψη χρήσης αντασφάλισης από την ασφαλιστική εταιρεία.

Παράρτημα Α

Στο σημείο αυτό εξετάζεται η καταλληλότητα των δεδομένων insurance data για την εφαρμογή κλασσικής παλινδρόμησης. Αρχικά εξετάζεται αν οι θετικές τιμές των ασφαλιστικών αποζημιώσεων της μεταβλητής claimcst0 ακολουθούν την κανονική κατανομή, με τον έλεγχο Shapiro-Wilk, τα αποτελέσματα του οποίου ακολουθούν παρακάτω.

Οι υποθέσεις του ελέγχου είναι οι ακόλουθες:

H_0 : Το δεδομένα ακολουθούν την κανονική κατανομή

H_1 : Τα δεδομένα δεν ακολουθούν την κανονική κατανομή

```
shapiro.test(ALLclaims$claimcst0)

      Shapiro-Wilk normality test

data:  ALLclaims$claimcst0
W = 0.51246, p-value < 2.2e-16
```

Από τα παραπάνω αποτελέσματα είναι εμφανές ότι οι τιμές των θετικών ασφαλιστικών αποζημιώσεων δεν ακολουθούν την κανονική κατανομή.

Στη συνέχεια εξετάζεται, έπειτα από κατάλληλους μετασχηματισμούς στις τιμές των θετικών αποζημιώσεων, η προσαρμογή των τιμών των αποζημιώσεων στην κανονική κατανομή.

- Έλεγχος κανονικότητας στον λογάριθμο της μεταβλητής claimcst0

```
shapiro.test(log(ALLclaims$claimcst0))

      Shapiro-Wilk normality test

data:  log(ALLclaims$claimcst0)
W = 0.94058, p-value < 2.2e-16
```

- Έλεγχος κανονικότητας με διπλό λογάριθμο στη μεταβλητή claimcst0

```
Shapiro-Wilk normality test

data:  log(log(ALLclaims$claimcst0))
W = 0.95249, p-value < 2.2e-16
```

- Έλεγχος κανονικότητας στον αντίστροφο της μεταβλητής claimcst0

```
shapiro.test(1/ALLclaims$claimcst0)

      Shapiro-Wilk normality test

data:  1/ALLclaims$claimcst0
W = 0.85169, p-value < 2.2e-16
```

- Έλεγχος κανονικότητας στο τετράγωνο της μεταβλητής claimcst0

```
shapiro.test(ALLclaims$claimcst0^2)

      Shapiro-Wilk normality test

data:  ALLclaims$claimcst0^2
W = 0.14917, p-value < 2.2e-16
```

Παρατηρούμε ότι σε κανέναν από τους παραπάνω ελέγχους τα δεδομένα δεν προσαρμόζονται στην κανονική κατανομή και επομένως η εφαρμογή της μεθόδου γραμμικής παλινδρόμησης δεν μπορεί να εφαρμοστεί

Παράρτημα Β

Παρακάτω παρατίθεται συνοπτικά ο κώδικας στην R που χρησιμοποιήθηκε στα δύο σετ δεδομένων, για την δημιουργία γραφημάτων, την ανάλυση, την εφαρμογή γραμμικής παλινδρόμησης (ενότητα 4.3) και τη λογιστική παλινδρόμηση (ενότητα 4.4.2) με χρήση της οποίας έγινε και η εφαρμογή της μεθόδου αντασφάλισης (ενότητα 4.4.3)

Δεδομένα mtcars

Κώδικας1 (Περιγραφικά στατιστικά και γραφήματα)

```
summary(mtcars)
boxplot(mpg ~ am, data = mtcars, xlab = "Transmission (0 = automatic, 1 = manual)",
col  = 2)
with(mtcars, hist(mpg, breaks= seq(8,36,3), col="green"))
ggplot(mtcars, aes(mpg)) + geom_histogram(binwidth = 4) + xlab('Miles per Gallon')
+ ylab('Number of Cars') +  ggtitle('Distribution of Cars by Mileage')
```

```
ggplot(mtcars, aes(hp, mpg)) + geom_point() +  geom_smooth(method = "lm", se = FALSE) + ylab("Miles per Gallon") + xlab("No. of Horsepower") + ggtitle("Impact of Number of Horsepower on MPG")
ggplot(mtcars,aes(x=hp, y=mpg))+geom_point()
ggplot(mtcars,aes(x=wt, y=mpg))+geom_point()
ggplot(mtcars,aes(x=cyl, y=mpg))+geom_point()
ggplot(mtcars,aes(x=am, y=mpg))+geom_point()
ggplot(mtcars,aes(x=hp, y=disp))+geom_point()
```

Κώδικας 2 (Απλή γραμμική παλινδρόμηση)

```
mtcars$amfactor <- factor(mtcars$am, labels = c("automatic", "manual"))
summary(lm(mpg ~ amfactor, data = mtcars))
```

Κώδικας 3 (Πολλαπλή γραμμική παλινδρόμηση)

```
Summary(lm(mpg  ~  cyl+disp+hp+drat+wt+qsec+factor(vs)+factor(am)+gear+carb, data = mtcars))
fit <- lm(mpg ~ cyl+disp+hp+drat+wt+qsec+factor(vs)+factor(am)+gear+carb, data = mtcars)
step <- stepAIC(fit, direction="both", trace=FALSE)
summary(step)
fit1 <- lm(mpg ~ factor(am), data = mtcars)
fit2 <- lm(mpg ~ factor(am)+wt, data = mtcars)
fit3 <- lm(mpg ~ factor(am)+wt+qsec, data = mtcars)
fit4 <- lm(mpg ~ factor(am)+wt+qsec+hp, data = mtcars)
fit5 <- lm(mpg ~ factor(am)+wt+qsec+hp+drat, data = mtcars)
```

```
anova(fit5)

finalfit <- lm(mpg ~ wt+qsec+factor(am), data = mtcars)

qqPlot(finalfit, main="Normal Q-Q plot")
```

Insurance Data

Κώδικας 1 (Γραφήματα θετικών αποζημιώσεων)

```
library(readr)

library(dplyr)

library( hflights)

head(car)

summary(car)

NoZeroclaims = (  car  %>%

                  filter(claimcst0 > 0 )    %>%

                  select(claimcst0)

NoZeroclaims = as.matrix(NoZeroclaims)

orio = 2000

par(mfrow = c(2,1))

hist(NoZeroclaims[NoZeroclaims < orio]   , main = paste("Αποζημιώσεις μέχρι ",

orio), xlab = "", col = "blue" )

boxplot(NoZeroclaims)
```

```
boxplot(NoZeroclaims[NoZeroclaims < orio])
quantile(NoZeroclaims, seq(50,95,5)/100)
```

Κώδικας 2 (επιλογή μεταβλητών και ανάλυση θετικών αποζημιώσεων)

```
claims = select(car, claimcst0)
ALLclaims = (  car  %>%
                filter(claimcst0 > 0 )    %>%
                select(claimcst0, veh_body, veh_age, gender, area, agecat)
)
ALLclaims  %>% group_by(gender,area) %>%
   summarise(mean_claims = mean(claimcst0, na.rm = T),
             median_claims = median(claimcst0, na.rm = T),
             sd_claims =  sd(claimcst0, na.rm = T),
             min_claims = min(claimcst0),
             max_claims =  max(claimcst0)
```

Κώδικας 3 (λογιστική παλινδρόμηση)

```
model <- glm(clm ~ veh_value + I(veh_value^2) + veh_age + gender + area + agecat ,
family=binomial, data=na.omit(car))
summary(model)
model1<- step(model)
model1 <- glm(clm ~ veh_value + I(veh_value^2) + agecat, family=binomial,
data=na.omit(car))
summary(model1)
```

Κώδικας 4 (Excess of Loss a=75%)

```
car3 = car2 %>% filter(car2$claimcst0>0)
 a= 0.75
d = quantile(car3$claimcst0, a)
###δημιουργία μεταβλητής paynopay
payNOpay = as.factor(ifelse(car3$claimcst0<d,1,0))
model3 <- glm(payNOpay ~ veh_body + veh_age + gender + area +agecat +
valuecat, family=binomial, data=na.omit(car4))
 summary(model3)
model4<- step(model3)
 summary(model4)
```

Βιβλιογραφία

Α. Ελληνική

1. Κούτρας, Μ. (2015) *Ανάλυση Παλινδρόμησης και Ανάλυση Διακύμανσης*. Σημειώσεις ΠΜΣ «Εφαρμοσμένη Στατιστική», Τμήμα Στατιστικής και Ασφ. Επιστήμης, Πανεπιστήμιο Πειραιώς.

2. Πολίτης, Κ. (2015) *Γενικευμένα Γραμμικά Μοντέλα*. Σημειώσεις ΠΜΣ «Εφαρμοσμένη Στατιστική», Τμήμα Στατιστικής και Ασφ. Επιστήμης, Πανεπιστήμιο Πειραιώς.

Β. Ξενόγλωσση

1. Belsley D.A, Kuh E. and Welsch R.E. (1980) *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley, New Work.

2. Box, G.E.P and Cox D.R (1964), *An Analysis of transformations,* Journal of the Royal Stastical Society, New York

3. Draper, N. R, and Smith, H. (1998) *Applied regression analysis,* 3rd edition, Wiley, New York.

4. Greene W. H., (2002), *Econometric Analysis 5th edition*

file:///C:/Users/user/Downloads/William_H_Greene-Econometric_Analysis-Prentice%20(1).pdf

5. Hardy, M. & NetLibrary (1993), *Regression with dummy variables*, Sage Publications, Newbury Park.

6. Outreville, J. F. (1998), *Theory and practice of insurance*, Kluwer Academic Publishers, London.

7. De Jong, P. and Heller G.Z. (2008), *Generalized linear models for insurance data*, Cambridge University Press.

8. Diacon, S.R., and Carter. R. L. (1988), *Reinsurance 2nd edition*, John Murray, London.

9. Stewart G.W., (1987), *Collinearity and Least Square Regression*, Institute of Mathematical Statistiscs https://www.jstor.org/stable/2245615?seq=1#page_scan_tab_contents

10. Wehrhahn R., (2009), *Introduction to reinsurance the world bank* http://siteresources.worldbank.org/Introduction_to_Reinsurance.pdf

Διαδίκτυο

1.https://blog.willis.com/2013/12/dealing-with-the-big-stuff-excess-of-loss-reinsurance/

2. https://www.r-project.org/contributors.html

3. https://www.r-project.org/about.html

4. https://www.Wikipedia.org