ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΑΝΑΛΟΓΙΣΤΙΚΗ ΕΠΙΣΤΗΜΗ ΚΑΙ ΔΙΟΙΚΗΤΙΚΗ ΚΙΝΔΥΝΟΥ**

ΑΝΑΛΟΓΙΣΤΙΚΑ ΜΟΝΤΕΛΑ ΤΙΜΟΛΟΓΗΣΗΣ ΚΑΙ ΑΝΑΛΥΣΗ ΑΣΦΑΛΙΣΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕ ΧΡΗΣΗ ΤΟΥ ΠΑΚΕΤΟΥ R

ΠΕΤΡΟΥ ΜΑΡΙΑ-ΕΛΕΥΘΕΡΙΑ

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος στη Αναλογιστική Επιστήμη και Διοικητική Κινδύνου

Πειραιάς,
Ιούνιος 2020

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ **ΑΝΑΛΟΓΙΣΤΙΚΗ ΕΠΙΣΤΗΜΗ ΚΑΙ ΔΙΟΙΚΗΤΙΚΗ ΚΙΝΔΥΝΟΥ**

ΑΝΑΛΟΓΙΣΤΙΚΑ ΜΟΝΤΕΛΑ ΤΙΜΟΛΟΓΗΣΗΣ ΚΑΙ ΑΝΑΛΥΣΗ ΑΣΦΑΛΙΣΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕ ΧΡΗΣΗ ΤΟΥ ΠΑΚΕΤΟΥ R

ΠΕΤΡΟΥ ΜΑΡΙΑ-ΕΛΕΥΘΕΡΙΑ

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος στη Αναλογιστική Επιστήμη και Διοικητική Κινδύνου

Πειραιάς,
Ιούνιος 2020

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 Αναλογιστική Επιστήμη και Διοικητική Κινδύνου.

Τα μέλη της Επιτροπής ήταν:

- Αναπληρωτής Καθηγητής: Ψαρράκος Γεώργιος (Επιβλέπων)
- Αναπληρωτής Καθηγητής: Πολίτης Κωνσταντίνος
- Επίκουρος Καθηγητής: Πιτσέλης Γεώργιος

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
ACTUARIAL SCIENCE AND RISK MANAGEMENT

ACTUARIAL PRICING MODELS AND ANALYSIS OF INSURANCE DATA USING THE R PACKET

By

Petrou Maria-Eleftheria

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 Actuarial Science and Risk Management

Piraeus, Greece
June 2020

Στους γονείς μου

Κωνσταντίνο και Ευθυμία

Ευχαριστίες

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον Αναπληρωτή Καθηγητή του τμήματος Ασφαλιστικής και Στατιστικής Επιστήμης, κ. Ψαρράκο Γεώργιο για την πολύτιμη βοήθεια και καθοδήγησή του καθ' όλη την διάρκεια εκπόνησης της διπλωματικής μου εργασίας, όπως επίσης τα μέλη της τριμελούς επιτροπής κ. Πιτσέλη Γεώργιο και κ. Πολίτη Κωνσταντίνο.

Επιπλέον, ένα μεγάλο ευχαριστώ στους γονείς μου και τους φίλους μου για την στήριξη που μου προσέφεραν και την υπομονή που έδειξαν κατά την διάρκεια της εργασίας μου.

Μαρία Πέτρου

Αθήνα, Ιούνιος 2020

Περίληψη

Σε πολλές επιστήμες αντιμετωπίζουμε την ανάγκη δημιουργίας στατιστικών μοντέλων, δηλαδή μοντέλων που να μελετάνε και να περιγράφουν την σχέση μεταξύ δύο ή περισσότερων μεταβλητών. Σκοπός είναι η πρόβλεψη των τιμών της μίας μεταβλητής μέσω των υπολοίπων. Το ίδιο συμβαίνει και στην ασφαλιστική επιστήμη όπου με την βοήθεια ιστορικών δεδομένων μπορούμε να προβλέψουμε ζημιές, αποθέματα, τιμές ασφαλίστρων κ.λ.π. Η παρούσα διπλωματική επικεντρώνεται στην τιμολόγηση ενός ασφαλιστικού προϊόντος με βάση κάποια χαρακτηριστικά του ασφαλισμένου μέσω γενικευμένων γραμμικών μοντέλων και την ανάλυση των δεδομένων με τα οποία χτίζεται το μοντέλο.

Στο πρώτο κεφάλαιο αναλύουμε το απλό γραμμικό μοντέλο και στην συνέχεια επεκτείνουμε τα αποτελέσματά μας στο πολλαπλό. Στο δεύτερο κεφάλαιο γίνεται μια εισαγωγή στο γενικευμένο γραμμικό μοντέλο και παρουσιάζεται η δομή αυτού, ο τρόπος εκτίμησης των παραμέτρων του και οι διαγνωστικοί έλεγχοι που χρειάζονται να γίνουν επάνω στα αποτελέσματα μας. Τέλος, στο τρίτο κεφάλαιο παρουσιάζονται εφαρμογές των γενικευμένων γραμμικών μοντέλων στην ασφάλιση αυτοκινήτου. Πιο συγκεκριμένα, παρουσιάζονται τεχνικές υπολογισμού του καθαρού ασφαλίστρου (pure premium) που θα χρεώναμε έναν ασφαλισμένο με βάση τα ατομικά χαρακτηριστικά αυτού και προτείνεται ένα κατάλληλο μοντέλο. Για τους υπολογισμούς μας χρησιμοποιούμε το στατιστικό πακέτο R, το οποίο χρησιμοποιείται ευρέως για την στατιστική ανάλυση δεδομένων καθώς και των γενικευμένων γραμμικών μοντέλων.

Abstract

In many sciences we encounter the necessity of creating statistical models, namely the models which observe and describe a mathematical relationship between one or more variables. Accordingly, in insurance science we can predict damages, loss reserving, insurance rates etc., with the help of historical data. This thesis investigates the pricing process of an insurance product based on certain features of the customer, using generalized linear models, and the data analysis with which the models are being created.

In the first chapter, we analyze the simple linear model and later on we expand our results to multiple. In the second chapter there is an introduction into the generalized linear model and a presentation of its structure, the way of assessment of the parameters and the diagnostic checks which are required to be performed on our results. Ultimately, in the third and final chapter, there are implementations of the generalized linear models in car insurance. More precisely, there is a presentation of techniques of calculating the pure premium, with which we would charge a customer, based on their personal features and we suggest a suitable model. For our calculations we use R, which is widely used for statistical data analysis, as well as for fitting the generalized linear models.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

Γραμμικά Μοντέλα Παλινδρόμησης

1.1 Μοντέλα Παλινδρόμησης

Σε αρκετές περιπτώσεις επίλυσης προβλημάτων μας ενδιαφέρει η ταυτόχρονη μελέτη δύο ή περισσότερων μεταβλητών για να προσδιορίσουμε με ποιον τρόπο οι μεταβλητές αυτές σχετίζονται μεταξύ τους. Η ανάλυση παλινδρόμησης είναι μία στατιστική τεχνική η οποία εξετάζει την σχέση μεταξύ των υπό εξέταση μεταβλητών. Οι εφαρμογές της ανάλυσης παλινδρόμησης είναι αμέτρητες και συναντώνται σχεδόν σε κάθε επιστήμη όπως είναι η ιατρική, η μηχανική, η οικονομία, η βιολογία και η ασφαλιστική (βλέπε Raymond et al. (2010)). Για παράδειγμα στην ασφάλιση αυτοκινήτου μας ενδιαφέρει να εξετάσουμε την σχέση ανάμεσα στο ασφάλιστρο που θα χρεώσουμε έναν ασφαλισμένο με βάση κάποια χαρακτηριστικά του όπως είναι για παράδειγμα η ηλικία του, το φύλο του, το αυτοκίνητό του ή ακόμα και τα χρόνια που είναι ενεργός οδηγός. Για διαφορετικά χαρακτηριστικά θα μπορούσαμε να χρεώσουμε περισσότερο ή λιγότερο κάποιον, αφού ένας έμπειρος οδηγός υποθετικά αποτελεί μικρότερο κίνδυνο για την ασφαλιστική εταιρία από έναν πιο νέο. Στο κεφάλαιο αυτό θα εξετάσουμε το απλό γραμμικό μοντέλο το οποίο αποτελείται από δύο μεταβλητές. Πιο συγκεκριμένα, θα μελετήσουμε την μεταβλητή Y την οποία μας ενδιαφέρει να εξετάσουμε και την μεταβλητή X μέσω της οποίας θα μπορέσουμε να καταλήξουμε σε αποτελέσματα για την υπό εξέταση μεταβλητή μας. Ο όρος γραμμικό αναφέρεται στην σχέση με την οποία συναντάμε την εξαρτημένη μεταβλητή Y και την ανεξάρτητη X και όχι στους συντελεστές τους οποίους θα χρησιμοποιήσουμε για την εκτίμηση. Στην συνέχεια θα επεκτείνουμε τα συμπεράσματα μας στο πολλαπλό γραμμικό μοντέλο και θα δούμε πως αυτά αποτυπώνονται μέσω της γλώσσας προγραμματισμού R. Στο παρόν κεφάλαιο η βιβλιογραφία που χρησιμοποιήθηκε είναι από των Montgomery et al. (1982), του Boland J. Philip (2007) και του Rencher & Schaalje (2007).

1.2 Απλό Γραμμικό Μοντέλο Παλινδρόμησης

Ένα απλό γραμμικό μοντέλο παλινδρόμησης δίνεται από την παρακάτω μορφή (βλέπε Montgomery et al. (1982)) :

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

όπου

- Y είναι η εξαρτημένη μεταβλητή του μοντέλου, ή μεταβλητή απόκρισης (response variable). Πρόκειται για την μεταβλητή η οποία μας ενδιαφέρει να εξετάσουμε.
- X είναι η ανεξάρτητη μεταβλητή του μοντέλου, ή επεξηγηματική μεταβλητή.
- β_0 είναι ο σταθερός όρος του μοντέλου (intercept) ο οποίος είναι άγνωστος και θα εκτιμηθεί στην συνέχεια.
- β_1 είναι η κλίση του μοντέλου ή ευθείας παλινδρόμησης
- ε είναι τα τυχαία σφάλματα τα οποία προκύπτουν συγκρίνοντας τις πραγματικές τιμές του μοντέλου και τις εκτιμώμενες τιμές αυτών.

Για τα τυχαία σφάλματα υποθέτουμε ότι ακολουθούν κανονική κατανομή με μέση τιμή 0 και διακύμανση σ^2, δηλ. $\varepsilon \sim N(0, \sigma^2)$. Επίσης είναι ασυσχέτιστα, δηλαδή η τιμή του ενός δεν εξαρτάται από την τιμή κάποιου άλλου. Υποθέτουμε επίσης ότι για κάθε τιμή της μεταβλητής X, η Y ακολουθεί κάποια κατανομή. Στην περίπτωση του απλού γραμμικού μοντέλου υποθέτουμε ότι η μεταβλητή Y ακολουθεί την κανονική κατανομή με μέση τιμή $\beta_0 + \beta_1 X$ και διακύμανση σ^2 ($Y \sim N(\beta_0 + \beta_1 X\ , \sigma^2)$. Αυτό προκύπτει από το παρακάτω:

$$E(Y) = \beta_0 + \beta_1 X,$$

και

$$Var(Y) = Var(\beta_0 + \beta_1 X + \varepsilon) = \sigma^2$$

Άρα η μέση τιμή του Y έχει γραμμική σχέση με την μεταβλητή X παρόλο που η διακύμανση του Y δεν εξαρτάται από την τιμή του X. Επιπλέον επειδή τα σφάλματα είναι ασυσχέτιστα, οι μεταβλητές απόκρισης είναι επίσης ασυσχέτιστες. Οι παράμετροι β_0, β_1 λέγονται και συντελεστές παλινδρόμησης (regression coefficients). Η κλίση β_1 δείχνει την αναμενόμενη αλλαγή της μέσης τιμής του Y όταν το X αυξάνεται κατά μία μονάδα. Εάν το εύρος των τιμών του X περιλαμβάνει το 0, τότε το β_0 είναι η μέση τιμή του Y για $X = 0$.

1.3 Μέθοδος Ελαχίστων Τετραγώνων

Οι παράμετροι β_0, β_1 είναι άγνωστοι και θα εκτιμηθούν από το τυχαίο δείγμα παρατηρήσεων. Έστω ότι έχουμε n ζευγάρια δεδομένων $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$. Για το τυχαίο δείγμα, το απλό γραμμικό μοντέλο γίνεται:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1,2,\ldots,n\,.$$

Η μέθοδος που θα χρησιμοποιήσουμε για την εκτίμηση των παραμέτρων βασίζεται στην ελαχιστοποίηση των τετραγωνικών σφαλμάτων ε_i και ονομάζεται μέθοδος των ελαχίστων τετραγώνων. Η ποσότητα που χρειάζεται να ελαχιστοποιήσουμε είναι η παρακάτω:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

και αυτό επιτυγχάνεται με μερική παραγώγιση ως προς β_0, β_1 και εξισώνοντας με μηδέν. Πιο συγκεκριμένα:

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0 \Leftrightarrow -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0,$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0 \Leftrightarrow -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0,$$

όπου $\hat{\beta}_0, \hat{\beta}_1$ είναι οι εκτιμητές ελαχίστων τετραγώνων των τιμών β_0, β_1. Στην συνέχεια θα σχηματίσουμε το παρακάτω σύστημα εξισώσεων:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} {x_i}^2 = \sum_{i=1}^{n} x_i y_i \end{cases}. \tag{1.1}$$

Οι εξισώσεις (1.1) ονομάζονται κανονικές εξισώσεις και η λύση τους είναι:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

με

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}, \qquad \bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}.$$

Το προσαρμοσμένο μοντέλο είναι το:

$$E(y_i) = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad i = 1,2,\ldots,n\,.$$

Η διαφορά μεταξύ της πραγματικής τιμής y_i και της εκτιμώμενης τιμής $\hat{y}_i$ ονομάζεται υπόλοιπο ή κατάλοιπο (residual) και για την $i-$ εκτίμηση γράφεται ως

$$\varepsilon_\iota = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \qquad i = 1,2,\ldots,n.$$

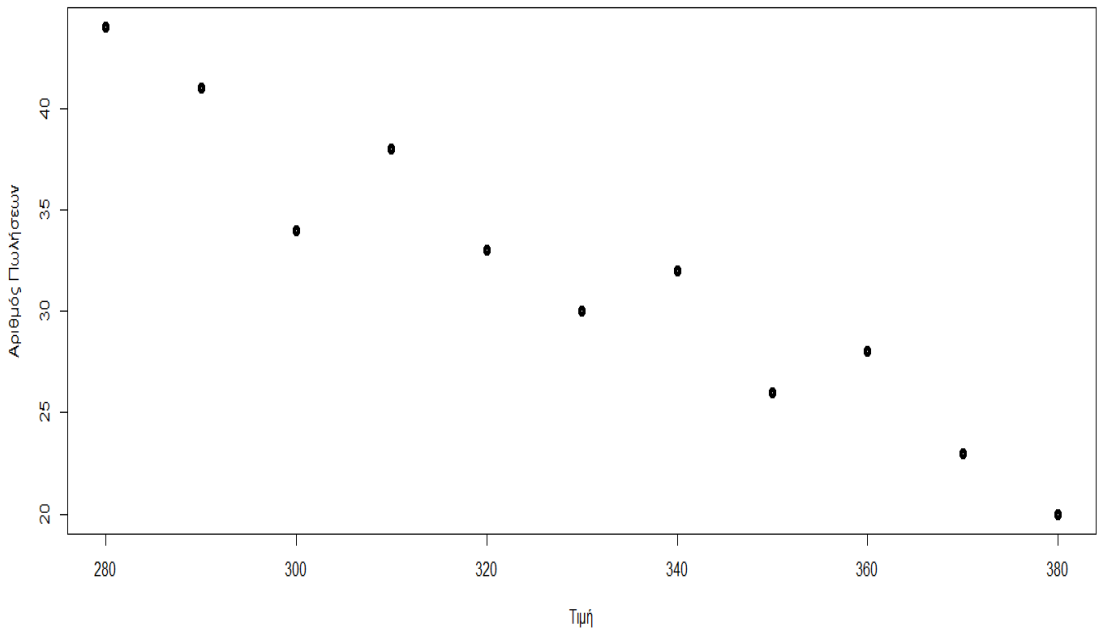
Στη συνέχεια δίνουμε ένα Παράδειγμα γραμμικού μοντέλου σε ασφαλιστικά δεδομένα με σκοπό την μελέτη επίδρασης της τιμής ενός ασφαλιστικού προϊόντος στον αριθμό των πωλήσεων.

Παράδειγμα 1*: Μια ασφαλιστική επιχείρηση μελετάει την επίδραση της τιμής ενός ασφαλιστικού προϊόντος στον αριθμό των πωλήσεων. Η μελέτη γίνεται πανελλαδικά και η τιμή του προϊόντος αλλάζει από πόλη σε πόλη. Για τον λόγο αυτό η επιχείρηση συνέλλεξε τα ακόλουθα δεδομένα.*

i −Παρατήρηση	Τιμή (x_i)	Αριθμός πωλήσεων (y_i)
1	280	44
2	290	41
3	300	34
4	310	38
5	320	33
6	330	30
7	340	32
8	350	26
9	360	28
10	370	23
11	380	20

Πίνακας 1.1. Δεδομένα Παραδείγματος 1

Στο παρακάτω σχήμα βλέπουμε των αριθμό των πωλήσεων του προϊόντος συναρτήσει της τιμής του.



Γράφημα 1.1: Διάγραμμα των σημείων για τα δεδομένα του Παραδείγματος 1.

Για την εκτίμηση των παραμέτρων υπολογίζουμε τις ποσότητες:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

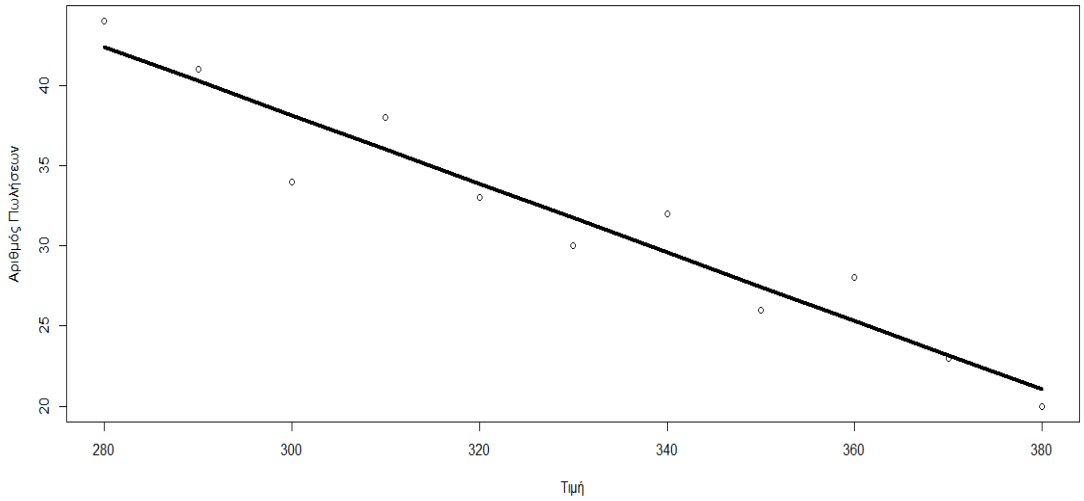
με $\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$ =300 και $S_{xx} = \sum_{i=1}^{n} (x_i - 300)^2 = 11000, S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x}) y_i = -2350.$

.

Επομένως $$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-2350}{11000} = -0.2136364,$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 102.227$$

και η ευθεία παλινδρόμησης είναι η $\hat{y}_i = 102.227 - 0.2136364 x_i$, την οποία προσαρμόζουμε στο διάγραμμα των σημείων.



Γράφημα 1.2: Διάγραμμα των σημείων για τα δεδομένα του παραδείγματος 1 και ευθεία παλινδρόμησης.

Στον παρακάτω πίνακα βλέπουμε τις παρατηρούμενες τιμές y_i, τις εκτιμώμενες τιμές $\hat{y}_i$ και τα κατάλοιπα ε_i.

Αριθμός πωλήσεων (y_i)	***Εκτιμώμενες τιμές πωλήσεων*** $(\hat{y}_i)$	***Κατάλοιπα*** (ε_ι)
44	*42.4*	*2.4*
41	*40.3*	*0.7*
34	*38.1*	*-4.1*
38	*36*	*2*
33	*33.8*	*-0.8*
30	*31.7*	*-1.7*
32	*29.5*	*2.5*
26	*27.4*	*-1.4*
28	*25.3*	*2.7*
23	*23.1*	*0.1*
20	*21.04*	*1.4*

Πίνακας 1.2: Οι τιμές των μεταβλητών $y_i, \hat{y}_i, \varepsilon_\iota$ του παραδείγματος 1.

1.4 Ιδιότητες των Εκτιμητών Ελαχίστων Τετραγώνων $\hat{\beta}_0, \hat{\beta}_1$

Οι εκτιμητές ελαχίστων τετραγώνων $\hat{\beta}_0, \hat{\beta}_1$ έχουν κάποιες πολύ σημαντικές ιδιότητες. Πρώτον παρατηρούμε από τις εξισώσεις S_{xy}, S_{xx} ότι τα $\hat{\beta}_0, \hat{\beta}_1$είναι γραμμικοί συνδυασμοί των παρατηρήσεων y_i. Για παράδειγμα:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i,$$

με $c_i = (x_i - \bar{x})/S_{xx}$ για $i = 1,2, \dots n$.

Δεύτερον οι εκτιμητές ελαχίστων τετραγώνων $\hat{\beta}_0, \hat{\beta}_1$ είναι αμερόληπτοι εκτιμητές των παραμέτρων του μοντέλου β_0, β_1, δηλαδή ισχύει $E(\hat{\beta}_0) = \beta_0$ και $E(\hat{\beta}_1) = \beta_1$. Θα αποδείξουμε το παραπάνω για το $\hat{\beta}_1$ και αντίστοιχα αποδεικνύεται για το $\hat{\beta}_0$.

$$E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i E(y_i)$$
$$= \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i\, x_i,$$

με $\sum_{i=1}^{n} c_i = 0$ και $\sum_{i=1}^{n} c_i\, x_i = 1$.

Συνεπάγεται $E(\hat{\beta}_1) = \beta_1$. Η διακύμανση του β_1 είναι

$$Var(\hat{\beta}_1) = Var\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} {c_i}^2 Var(y_i), \tag{1.2}$$

επειδή οι παρατηρήσεις $\hat{y}_i$ είναι ασυσχέτιστες και άρα η διακύμανση του αθροίσματος είναι τελικά το άθροισμα των διακυμάνσεων. Επιπλέον από την υπόθεση $Var(y_i) = \sigma^2$ η (1.2) γίνεται:

$$Var(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} {c_i}^2 = \sigma^2 \frac{\sum_{i=1}^{n}(\chi_i - \bar{x})^2}{{S_{xx}}^2} = \frac{\sigma^2}{S_{xx}}.$$

Η διακύμανση του β_0 είναι

$$Var(\hat{\beta}_0) = Var(\bar{y} - \hat{\beta}_1 \bar{x})$$

$$= Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x} Cov(\bar{y}, \hat{\beta}_1),$$

με $Var(\bar{y}) = \sigma^2/n$ και $Cov(\bar{y}, \hat{\beta}_1) = 0$ (η απόδειξη παραλείπεται).
Επομένως,

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{S_{xx}} = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right).$$

Ακόμα ένα σημαντικό αποτέλεσμα που αφορά την ποιότητα των εκτιμητών ελαχίστων τετραγώνων $\hat{\beta}_0$, $\hat{\beta}_1$ είναι το θεώρημα του Gauss-Markov το οποίο αποδεικνύει ότι για κάθε μοντέλο παλινδρόμησης στο οποίο ισχύουν οι υποθέσεις $E(\varepsilon_i) = 0$ και $Var(\varepsilon_i) = \sigma^2$ για τα τυχαία ασυσχέτιστα σφάλματα, τότε οι εκτιμητές ελαχίστων τετραγώνων είναι αμερόληπτοι και έχουν ελάχιστη διακύμανση σε σύγκριση με όλους τους άλλους αμερόληπτους εκτιμητές οι οποίοι είναι γραμμικοί συνδυασμοί των y_i. Αποδεικνύεται ότι εκτιμητές ελαχίστων τετραγώνων είναι οι καλύτεροι αμερόληπτοι εκτιμητές ως προς το ότι έχουν την ελάχιστη διακύμανση που θα μπορούσαν να έχουν. Παρακάτω θα δούμε κάποιες άλλες πολύ χρήσιμες ιδιότητες του προσαρμοσμένου μοντέλου παλινδρόμησης.

1. Το άθροισμα των καταλοίπων e_i για κάθε μοντέλο παλινδρόμησης που περιέχει τον σταθερό όρο β_0 είναι πάντα μηδέν (0).
$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0. \qquad (1.3)$$
2. Το άθροισμα των παρατηρούμενων τιμών y_i ισούται με το άθροισμα των προβλεπόμενων τιμών $\hat{y}_i$.
$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i.$$
3. Η ευθεία ελαχίστων τετραγώνων περνάει πάντα από το σημείο $(\bar{y}, \bar{x})$.

1.5 Εκτίμηση της Διακύμανσης σ^2

Πέρα από την εκτίμηση των β_0, β_1 χρειαζόμαστε μια εκτίμηση για το σ^2. Για τον υπολογισμό του σ^2 θα χρειαστούμε την παρακάτω ποσότητα την οποία ονομάζουμε τετραγωνικό άθροισμα των σφαλμάτων (error sum of squares) ή υπόλοιπο μεταβλητότητας (residual variation) και συμβολίζουμε με SSE.

$$SSE = \sum_{i=1}^{n}(e_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \qquad (1.4)$$

Αντικαθιστώντας $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ η (1.4) γίνεται:

$$SSE = \sum_{i=1}^{n} {y_i}^2 - n\bar{y} - \hat{\beta}_1 S_{xy}. \tag{1.5}$$

Θέτουμε $\sum_{i=1}^{n} {y_i}^2 - n\bar{y} = SST$ οπότε η (1.5) γίνεται:

$$SSE = SST - \hat{\beta}_1 S_{xy}.$$

Το τετραγωνικό άθροισμα των σφαλμάτων έχει $n-2$ βαθμούς ελευθερίας, επειδή 2 βαθμοί ελευθερίας έχουν χρησιμοποιηθεί για την εκτίμηση των β_0, β_1. Επίσης ισχύει ότι $E(SSE) = (n-2)\sigma^2$, άρα ένας αμερόληπτος εκτιμητής για το σ^2 είναι:

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = MSE. \tag{1.6}$$

Η ποσότητα MSE ονομάζεται μέσο τετραγωνικό σφάλμα (residual mean square). Η τετραγωνική ρίζα του $\hat{\sigma}^2$ ονομάζεται το τυπικό σφάλμα της παλινδρόμησης (standard error of regression). Επειδή το $\hat{\sigma}^2$ εξαρτάται από το SSE, οποιαδήποτε παραβίαση των υποθέσεων μπορεί να επιδράσει σε μεγάλο βαθμό στην τιμή του $\hat{\sigma}^2$.

Συνεχίζοντας με το *Παράδειγμα 1* θα υπολογίσουμε την εκτίμηση της διακύμανσης $\hat{\sigma}^2 = \frac{SSE}{n-2}$. Από την (1.5) προκύπτει ότι:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 44.13636,$$

επομένως $\hat{\sigma}^2 = \frac{44.13636}{2} = 4.90404$.

1.6 Έλεγχοι Υποθέσεων στο Απλό Γραμμικό Μοντέλο

Όταν κατασκευάζουμε ένα μοντέλο μας ενδιαφέρει να κάνουμε ελέγχους υποθέσεων και να κατασκευάζουμε διαστήματα εμπιστοσύνης που αφορούν τις παραμέτρους του. Για να γίνουν οι έλεγχοι υποθέσεων χρειάζεται να κάνουμε την υπόθεση ότι τα κατάλοιπα e_i ακολουθούν την κανονική κατανομή με μέση τιμή 0 και διακύμανση σ^2, καθώς και ότι είναι ασυσχέτιστα.

1.6.1 t-test

Ας υποθέσουμε ότι θέλουμε να εξετάσουμε αν η μεταβλητή β_1 είναι ίση με μια σταθερά έστω β_{10}. Ορίζουμε ως μηδενική υπόθεση H_o την $\beta_1 = \beta_{10}$ και ως εναλλακτική αυτής την $H_1: \beta_1 \neq \beta_{10}$. Αφού τα κατάλοιπα e_i ακολουθούν την κανονική κατανομή με μέση τιμή 0 και διακύμανση σ^2 και είναι ασυσχέτιστα, οι παρατηρήσεις y_i ακολουθούν την κανονική κατανομή με μέση τιμή $\beta_0 + \beta_1 x_i$ και διακύμανση σ^2. Θυμίζουμε ότι ο εκτιμητής ελαχίστων τετραγώνων $\hat{\beta}_1$ είναι γραμμικός συνδυασμός των παρατηρήσεων y_i, άρα ο $\hat{\beta}_1$ ακολουθεί κανονική κατανομή με μέση τιμή β_1 και διακύμανση σ^2/S_{xx}. Άρα η στατιστική συνάρτηση ελέγχου

$$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2/S_{xx}}} \ ,$$

ακολουθεί την τυπική κανονική κατανομή $Z_0 \sim N(0,1)$ αν ισχύει η μηδενική υπόθεση. Αν το σ^2 ήταν γνωστό θα μπορούσαμε να χρησιμοποιήσουμε την στατιστική συνάρτηση Z_0 για να ελέγξουμε την υπόθεσή μας. Όπως είδαμε, συνήθως το σ^2 είναι άγνωστο και το εκτιμάμε από το μοντέλο μας. Από την σχέση (1.6) έχουμε ότι

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

και έχουμε αποδείξει ότι το $\hat{\sigma}^2$ είναι αμερόληπτος εκτιμητής του σ^2. Αντικαθιστώντας όπου $\sigma^2 = \hat{\sigma}^2 = MSR$ προκύπτει ότι η στατιστική συνάρτηση ελέγχου είναι η παρακάτω:

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MSE/S_{xx}}}, \qquad (1.7)$$

η οποία ακολουθεί την κατανομή t_{n-2} αν ισχύει η μηδενική υπόθεση. Απορρίπτουμε την μηδενική υπόθεση H_o αν ισχύει,

$$|t_0| > t_{\alpha/2,n-2},$$

όπου α είναι το άνω ποσοστιαίο σημείο της κατανομής t_{n-2}, με n να είναι ο αριθμός των παρατηρήσεων.

Εναλλακτικά μπορούμε να χρησιμοποιήσουμε το κριτήριο του p-value. Στην στατιστική το $p - value$ (probability value) είναι η πιθανότητα να υπάρχουν τιμές τουλάχιστον όσο ακραίες όσο οι παρατηρούμενες όταν η μηδενική υπόθεση είναι αληθής. Απορρίπτουμε την μηδενική υπόθεση όταν $p \leq \alpha$ όπου α είναι το επιθυμητό επίπεδο σημαντικότητας.

Ο παρονομαστής της εξίσωσης (1.7), ονομάζεται και εκτιμώμενο τυπικό λάθος της μεταβλητής (estimated standard error) και συμβολίζεται ως

$$se(\hat{\beta}_1) = \sqrt{MSE/S_{xx}}\,.$$

Επομένως η (1.7) γράφεται και ως

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{se(\hat{\beta}_1)}\,.$$

Με παρόμοιο τρόπο υπολογίζουμε την μηδενική υπόθεση για την παράμετρο β_0. Για να ελέγξουμε την:

$$H_0: \beta_0 = \beta_{10}, \qquad H_1: \beta_0 \neq \beta_{10},$$

χρησιμοποιούμε την στατιστική συνάρτηση

$$t_0 = \frac{\hat{\beta}_0 - \beta_{10}}{\sqrt{MSE(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}} = \frac{\hat{\beta}_0 - \beta_{10}}{se(\hat{\beta}_0)}\,.$$

Απορρίπτουμε την μηδενική υπόθεση $H_0: \beta_0 = \beta_{10}$, εάν $|t_0| > t_{\alpha/2,n-2}$.
Μια ειδική περίπτωση των υποθέσεων είναι όταν η μηδενική υπόθεση είναι η παρακάτω:

$$H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0.$$

Εξετάζουμε δηλαδή αν υπάρχει όντως γραμμική εξάρτηση μεταξύ του x και του y. Σε περίπτωση μη απόρριψης της μηδενικής υπόθεσης οδηγούμαστε στο συμπέρασμα ότι δεν υπάρχει γραμμική σχέση μεταξύ των x, y. Ο τρόπος αξιολόγησης της μηδενικής υπόθεσης μπορεί να γίνει με δύο τρόπους. Ο πρώτος τρόπος είναι μέσω της στατιστικής συνάρτησης t για

$$\beta_{10} = 0,$$

δηλαδή εάν

$$|t_0| = \left|\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}\right| > t_{\alpha/2,n-2},$$

τότε απορρίπτουμε την μηδενική υπόθεση.

Ο δεύτερος τρόπος είναι μέσω της ανάλυσης της διακύμανσης (ANOVA). Η ανάλυση της διακύμανσης βασίζεται στον διαμοιρασμό της συνολικής μεταβλητότητας του μοντέλου στην μεταβλητή απόκρισης y_i. Αρχικά θεωρούμε ότι:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)\,. \tag{1.8}$$

Υψώνοντας και τα δύο μέλη της εξίσωσης (1.8) στο τετράγωνο και αθροίζοντας για όλες τις παρατηρήσεις n έχουμε:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})\,(y_i - \hat{y}_i), \tag{1.9}$$

με

$$2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})\,(y_i - \hat{y}_i) = 2\sum_{i=1}^{n}\hat{y}_i(y_i - \hat{y}_i) - 2\bar{y}\sum_{i=1}^{n}(y_i - \hat{y}_i) = 2\sum_{i=1}^{n}\hat{y}_i e_i - 2y\sum_{i=1}^{n} e_i = 0.$$

Από την πρόταση (1.3) δείξαμε ότι $\sum_{i=1}^{n} e_i = 0$. Επομένως, η (1.9) γίνεται

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

η οποία γράφεται συμβολικά ως:

$$SST = SSR + SSE$$

όπου

- $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ λέγεται άθροισμα τετραγώνων παλινδρόμησης (regression sum of squares) και δείχνει την απόκλιση της προβλεπόμενης τιμής από την μέση τιμή της μεταβλητής απόκρισης.
- $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ λέγεται άθροισμα τετραγώνων των σφαλμάτων (error sum of squares) ή υπόλοιπο μεταβλητότητας (residual variation).
- $SST = SSR + SSE$ λέγεται ολική μεταβλητότητα (total variation).

Οι βαθμοί ελευθερίας για κάθε όρο αναλύονται ως εξής: Η ολική μεταβλητότητα SST έχει $df_T = n - 1$ βαθμούς ελευθερίας καθώς ένας βαθμός έχει χαθεί για τον υπολογισμό της μέσης τιμής $\bar{y}$. Ο όρος SSR έχει $df_R = 1$ βαθμό ελευθερίας γιατί εξαρτάται ακριβώς από μία παράμετρο την $\hat{\beta}_1$. Τέλος ο όρος SSE έχει $df_E = n - 2$ βαθμούς ελευθερίας αφού όπως είδαμε για τον υπολογισμό του $y_i - \hat{y}_i$ χρειαστήκαμε τους εκτιμητές ελαχίστων τετραγώνων $\hat{\beta}_0, \hat{\beta}_1$.

Για να ελέγξουμε την μηδενική υπόθεση $H_0: \beta_1 = 0$, θα χρησιμοποιήσουμε την στατιστική συνάρτηση ελέγχου διακυμάνσεων F με:

$$F_o = \frac{\frac{SSR}{df_R}}{\frac{SSE}{df_E}} = \frac{\frac{SSR}{1}}{\frac{SSE}{n-2}} = \frac{MSR}{MSE},$$

η οποία ακολουθεί την $F_{1,n-2}$ κατανομή. Απορρίπτουμε την μηδενική υπόθεση H_0, εάν $F_o > F_{1,n-2}$.

Συνεχίζοντας με το *Παράδειγμα 1* θα ελέγξουμε την υπόθεση $H_0: \hat{\beta}_1 = 0$ με την χρήση της στατιστικής συνάρτηση t_0 καθώς και με την ανάλυση της διακύμανσης F_o.

Η τιμή της στατιστικής συνάρτησης t_0 είναι ίση με:

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MSE/S_{xx}}} = \frac{-0.2136364}{\sqrt{4.90404/11000}} = -10.118.$$

Επιπλέον για $\alpha = 0.01$, $t_{\alpha/2,n-2} = t_{0.05,9}$ =1.833113.
Προκύπτει ότι $|t_0| > t_{0.25,9}$, επομένως απορρίπτουμε την $H_0: \hat{\beta}_1 = 0$ σε επίπεδο σημαντικότητας $\alpha = 0.05$ και συμπεραίνουμε ότι υπάρχει γραμμική εξάρτηση μεταξύ των μεταβλητών χ, y.

Η τιμή της στατιστικής συνάρτηση F_o είναι ίση με:

$$F_o = \frac{MSR}{MSE} = \frac{502.0456}{4.90404} = 102.3739.$$

Επιπλέον για $\alpha = 0.05$, $F_{1,n-2} = F_{1,9}$ =3.245089e-06.
Προκύπτει ότι $F_o > F_{1,9}$,επομένως απορρίπτουμε την $H_0: \hat{\beta}_1 = 0$ σε επίπεδο σημαντικότητας $\alpha = 0.05$ και συμπεραίνουμε ότι υπάρχει γραμμική εξάρτηση μεταξύ των μεταβλητών x, y. Παρατηρούμε ότι και με τις δύο προσεγγίσεις φτάσαμε στο ίδιο συμπέρασμα.

1.6.2 Συντελεστής Προσδιορισμού R^2

Η ποσότητα

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST},$$

ονομάζεται συντελεστής προσδιορισμού R^2 (coefficient of determination) και εκφράζει το ποσοστό της συνολικής μεταβλητότητας των y_i που εξηγείται από το μοντέλο. Παίρνει τιμές στο διάστημα [0,1]. Όταν όλες οι προβλεπόμενες τιμές $\hat{y}_i$ συμπίπτουν με τις πραγματικές τιμές y_i, δηλαδή $\hat{y}_i = y_i$ θα έχουμε $SSE = 0$ και $R^2 = 1$ ενώ όταν η κλίση $\hat{\beta}_1$=0 θα έχουμε $R^2 = 0$.
Για τα δεδομένα του *Παραδείγματος 1* ο συντελεστής προσδιορισμού R^2 είναι ίσος με:

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE} = \frac{502.0456}{502.0456 + 44.13636} = 0.9191,$$

το οποίο σημαίνει ότι 92% της συνολικής μεταβλητότητας των y_i εξηγείται από το μοντέλο.
Ο συντελεστής προσδιορισμού R^2 θα πρέπει να χρησιμοποιείται με προσοχή, γιατί παρόλο που μπορεί να έχουμε μεγάλο R^2, αυτό δεν θα σημαίνει απαραίτητα ότι το επιλεγμένο μοντέλο παλινδρόμησης είναι το κατάλληλο για την εκτίμηση των δεδομένων μας.

Αν λάβουμε υπόψη μας το μέγεθος του μοντέλου ορίζουμε τον προσαρμοσμένο συντελεστή προσδιορισμού ο οποίος στην γενική του μορφή έχει τον παρακάτω τύπο:

$$R^2_{adj} = 1 - (1 - R)\frac{n-1}{n-p-1}.$$

Για τα δεδομένα του *Παραδείγματος 1* ο προσαρμοσμένος συντελεστής προσδιορισμού R^2_{adj} είναι ίσος με:

$$R^2_{adj} = 1 - (1 - 0.9191)\frac{10}{8} = 0.9.$$

Συνοψίζοντας θα παρουσιάσουμε τα παραπάνω αποτελέσματα μέσω της R και στην συνέχεια θα επεκταθούμε στο πολλαπλό γραμμικό μοντέλο.

```
> ##εισαγωγη των σημείων χ,y σε διανύσματα

> x<-c(280,290,300,310,320,330,340,350,360,370,380)
> y<-c(44,41,34,38,33,30,32,26,28,23,20);
> data<-data.frame(x,y)

> ##συνάρτηση lm για τον υπολογισμό του απλού γραμμικού μοντέλου

> slm<-lm(y~x,data=data)
> summary(slm)

Call:
lm(formula = y ~ x, data = data)

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

```
-4.1364 -1.2500 -0.1818  1.7955  2.6818

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.22727    6.99970   14.61 1.42e-07 ***
x            -0.21364    0.02111  -10.12 3.25e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.215 on 9 degrees of freedom
Multiple R-squared:  0.9192,  Adjusted R-squared:  0.9102
F-statistic: 102.4 on 1 and 9 DF,  p-value: 3.245e-06
```

- Η στήλη **`Estimate`** περιέχει τις εκτιμήσεις των συντελεστών της παλινδρόμησης $\hat{\beta}_o, \hat{\beta}_1$.
- Η στήλη **Std. Error** περιέχει τα αντίστοιχα τυπικά σφάλματα των συντελεστών της παλινδρόμησης.
- Η στήλη **t value** περιέχει την τιμή την τιμή του στατιστικού ελέγχου T για τον έλεγχο της μηδενικής υπόθεσης $H_0: \hat{\beta}_1 = 0$.
- Η στήλη **Pr(>|t|)** περιέχει τις $p - value$ τιμές των 2 ουρών. Για να πάρουμε την τιμή της μίας διαιρούμε με 2.
- Η γραμμή **Multiple R-squared** περιέχει την τιμή του συντελεστή προσδιορισμού R^2, ενώ **Adjusted R-squared** περιέχει την τιμή του προσαρμοσμένου συντελεστή προσδιορισμού R^2_{adj}.
- Η γραμμή **F-Statistic** περιέχει την τιμή του στατιστικού ελέγχου F για τον έλεγχο της μηδενικής υπόθεσης $H_0: \hat{\beta}_1 = 0$ και την τιμή $p - value$ για αυτή την τιμή.
- Η γραμμή **Signif. Codes** μας δείχνει πόσο σίγουροι μπορούμε να είμαστε ότι ο αντίστοιχος συντελεστής παλινδρόμησης είναι σημαντικός για την εξαρτημένη μεταβλητή. Παρατηρούμε ότι τα αποτελέσματά μας συμπίπτουν με αυτά που προέκυψαν μέσω της R. Παρακάτω λαμβάνουμε και τις προβλεπόμενες τιμές $\hat{y}_i$ του μοντέλου.

Με την παρακάτω εντολή μπορούμε να πάρουμε τις προβλεπόμενες τιμές $\hat{y}_i$ του μοντέλου.

```
> slm$fitted.values
       1        2        3        4        5        6
42.40909 40.27273 38.13636 36.00000 33.86364 31.72727
       7        8        9       10       11
29.59091 27.45455 25.31818 23.18182 21.04545
```

1.7 Πολλαπλό Γραμμικό Μοντέλο

Ένα γραμμικό μοντέλο το οποίο περιλαμβάνει παραπάνω από έναν συντελεστή παλινδρόμησης ονομάζεται πολλαπλό γραμμικό μοντέλο. Στα επόμενα κεφάλαια θα δούμε πως το απλό γραμμικό μοντέλο επεκτείνεται στο πολλαπλό και αποτελεί ουσιαστικά ειδική περίπτωση αυτού.

Ένα μοντέλο της παρακάτω μορφής:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_i, \qquad (1.10)$$

ονομάζεται πολλαπλό γραμμικό μοντέλο παλινδρόμησης με k επεξηγηματικές μεταβλητές $x_i, i = 1,2, \dots k$ και $k+1$ συντελεστές παλινδρόμησης $\beta_j, j = 0,1, .. k$. Υποθέτουμε, επίσης ότι ε_i είναι τα τυχαία σφάλματα για τα οποία υποθέτουμε ότι ακολουθούν κανονική κατανομή με μέση τιμή 0 και διακύμανση σ^2, δηλ. $\varepsilon_i \sim N(0, \sigma^2)$. Επίσης είναι ασυσχέτιστα, δηλαδή η τιμή του ενός δεν εξαρτάται από την τιμή του άλλου. Οι συντελεστές παλινδρόμησης β_j αναπαριστούν την αναμενόμενη αλλαγή της τιμής της μεταβλητής απόκρισης y για μία μονάδα αύξησης ή μείωσης της μεταβλητής x_i με την προϋπόθεση ότι οι υπόλοιπες μεταβλητές $x_i,\ i \neq j$ μένουν σταθερές. Τα μοντέλα πολλαπλής γραμμικής παλινδρόμησης χρησιμοποιούνται σαν εμπειρικά μοντέλα πρόβλεψης. Γενικά, κάθε μοντέλο παλινδρόμησης το οποίο είναι γραμμικό ως προς τις παραμέτρους είναι ένα γραμμικό μοντέλο παλινδρόμησης. Έτσι τα παρακάτω μοντέλα μπορούν να επιλυθούν κατά τον ίδιο τρόπο και με τις ίδιες τεχνικές του πολλαπλού γραμμικού μοντέλου παλινδρόμησης αρκεί να γίνει η κατάλληλη παραμετροποίηση. Για παράδειγμα:

- $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$, θέτοντας $x = x_1$, $x^2 = x_2$, $x^3 = x_3$ καταλήγουμε στο πολλαπλό γραμμικό μοντέλο της μορφής $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$.

1.8 Εκτίμηση των Παραμέτρων στο Πολλαπλό Γραμμικό Μοντέλο

Όπως και στο απλό γραμμικό μοντέλο έτσι και στο πολλαπλό η εκτίμηση των παραμέτρων θα γίνει μέσω της μεθόδου ελαχίστων τετραγώνων. Η γεωμετρική ερμηνεία της μεθόδου ελαχίστων τετραγώνων στο πολλαπλό γραμμικό μοντέλο είναι να ελαχιστοποιηθεί η επιφάνεια που περνάει από όλα τα σημεία y_i.

1.8.1 Μέθοδος Ελαχίστων Τετραγώνων

Η μέθοδος των ελαχίστων τετραγώνων μπορεί να χρησιμοποιηθεί για την εκτίμηση των συντελεστών παλινδρόμησης της (1.10). Ας υποθέσουμε ότι έχουμε $n > k$ παρατηρήσεις, και ότι y_i είναι η τιμή της μεταβλητής απόκρισης για την $i -$ παρατήρηση. Επίσης x_{ij} είναι η τιμή της επεξηγηματική μεταβλητής x_j για την $i -$ παρατήρηση. Τότε τα δεδομένα μας θα έχουν την παρακάτω μορφή:

Παρατήρηση i	y	x_1	x_2	...	x_k
1	y_1	x_{11}	x_{12}	...	x_{1k}
2	y_2	x_{21}	x_{22}	...	x_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	...	$\vdots$
n	y_n	x_{n1}	x_{n2}	...	x_{nk}

Πίνακας 1.3: Ανάλυση των συνιστωσών ενός πολλαπλού γραμμικού μοντέλου

Επομένως μπορούμε να γράψουμε το πολλαπλό γραμμικό μοντέλο της μορφής (1.10) στην παρακάτω μορφή:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

$$= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, \quad i = 1,2\ldots,n.$$

Η συνάρτηση ελαχίστων τετραγώνων είναι η

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2.$$

Παραγωγίζοντας ως προς $\beta_0, \beta_1, \ldots, \beta_k$ την συνάρτηση S και εξισώνοντας την με μηδέν δημιουργούμε ένα σύστημα $\kappa + 1$ εξισώσεων το οποίο θα επιλύσουμε ταυτόχρονα για να βρούμε τους εκτιμητές ελαχίστων τετραγώνων $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. Πιο συγκεκριμένα έχουμε:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0$$

και

$$\frac{\partial S}{\partial \beta_j} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0\,, j = 1,2,..k.$$

Απλοποιώντας τις δύο παραπάνω εξισώσεις καταλήγουμε στις παρακάτω κανονικές εξισώσεις ελαχίστων τετραγώνων.

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} {x_{i1}}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} = \sum_{i=1}^{n} x_{i1} y_i$$

$$\vdots$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} {x_{ik}}^2 = \sum_{i=1}^{n} x_{ik} y_i,$$

και σε μορφή πινάκων έχουμε:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon},$$

όπου,

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Η συνάρτηση ελαχίστων τετραγώνων σε μορφή πινάκων γίνεται

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i{}^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$$
$$= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{2\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta}.$$

και οι εκτιμητές ελαχίστων τετραγώνων πρέπει να ικανοποιούν την παρακάτω εξίσωση:

$$\frac{\partial S}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X\beta} = 0$$
$$\Longrightarrow \boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}. \qquad (1.11)$$

Οι εξισώσεις (1.11) είναι κανονικές εξισώσεις ελαχίστων τετραγώνων. Για την επίλυσή τους πολλαπλασιάζουμε και τα δύο μέλη της (1.11) με τον αντίστροφο πίνακα του $\boldsymbol{X}'\boldsymbol{X}$, οπότε η (1.11) γίνεται

$$(\boldsymbol{X}'\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}'\boldsymbol{y}$$
$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}'\boldsymbol{y},$$

οι οποίοι είναι οι εκτιμητές ελαχίστων τετραγώνων του πολλαπλού γραμμικού μοντέλου, με την προϋπόθεση ότι ο αντίστροφος πίνακας υπάρχει. Ο αντίστροφος πίνακας $(\boldsymbol{X}'\boldsymbol{X})^{-\boldsymbol{1}}$ υπάρχει πάντα αν οι επεξηγηματικές μεταβλητές $\boldsymbol{X}$ είναι γραμμικώς ανεξάρτητες, δηλαδή αν καμία στήλη του $\boldsymbol{X}$ δεν είναι γραμμικός συνδυασμός κάποια άλλης στήλης του $\boldsymbol{X}$.

Το προσαρμοσμένο μοντέλο γραμμικής παλινδρόμησης είναι το

$$E(\widehat{\boldsymbol{y}}) = \widehat{\boldsymbol{y}} = \boldsymbol{x}'\widehat{\boldsymbol{\beta}} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j\, x_j.$$

Το διάνυσμα των προσαρμοσμένων τιμών $\hat{y}_{\boldsymbol{i}}$ είναι το

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{Hy},$$

Ο διαγώνιος πίνακας $\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}'$ ονομάζεται και πίνακας hat (hat matrix). Ο πίνακας hat και οι ιδιότητές του παίζουν σημαντικό ρόλο στην ανάλυση παλινδρόμησης.
Η διαφορά μεταξύ της παρατηρούμενης τιμής y_i και της προσαρμοσμένης τιμής $\hat{y}_i$ ονομάζεται κατάλοιπο, $e_i = y_i - \hat{y}_i$. Τα κατάλοιπα μπορούν να γραφτούν σε μορφή πινάκων

$$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}}$$
$$\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{Hy} = (\boldsymbol{1} - \boldsymbol{H})\boldsymbol{y},$$

1.8.2 Ιδιότητες των Εκτιμητών Ελαχίστων Τετραγώνων

Αποδεικνύουμε ότι εκτιμητές ελαχίστων τετραγώνων $\widehat{\boldsymbol{\beta}}$ είναι αμερόληπτοι εκτιμητές των παραμέτρων του μοντέλου $\boldsymbol{\beta}$, δηλαδή ισχύει $E\left(\widehat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$.

$$E\left(\widehat{\boldsymbol{\beta}}\right) = E[(\boldsymbol{X}'\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}'\boldsymbol{y}] = E[((\boldsymbol{X}'\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}'(\boldsymbol{X\beta} + \boldsymbol{\varepsilon})]$$
$$= E[(\boldsymbol{X}'\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}'\boldsymbol{X\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}'\boldsymbol{\varepsilon}] = \boldsymbol{\beta},$$

αφού $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ και $(\boldsymbol{X}'\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}$.

1.9 Εκτίμηση της Διακύμανσης σ^2

Παρόμοια με το απλό γραμμικό μοντέλο θα χρειαστούμε έναν εκτιμητή για την διακύμανση σ^2. Για τον υπολογισμό του σ^2 θα χρειαστούμε το τετραγωνικό άθροισμα των σφαλμάτων (error sum of squares) ή υπόλοιπο μεταβλητότητας (residual variation) SSE.

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \boldsymbol{e}'\boldsymbol{e}, \qquad (1.12)$$

Αντικαθιστώντας στην (1.12) όπου $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ η (1.12) γίνεται

$$\begin{aligned} SSE &= \left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right) \\ &= \boldsymbol{y}'\boldsymbol{y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} + \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}. \end{aligned} \qquad (1.13)$$

Αντικαθιστώντας όπου $\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$ η (1.13) γίνεται:

$$SSE = \boldsymbol{y}'\boldsymbol{y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}.$$

Αποδεικνύεται ότι το τετραγωνικό άθροισμα των σφαλμάτων έχει $n - p$ βαθμούς ελευθερίας, αφού έχουμε εκτιμήσει p παραμέτρους στο γραμμικό μοντέλο. Το μέσο τετραγωνικό σφάλμα (residual mean square) MSE ισούται με:

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

και αποδεικνύεται ότι $E(MSE) = \sigma^2$, επομένως είναι ένας αμερόληπτος εκτιμητής της διακύμανσης ο οποίος δίνεται από την παρακάτω σχέση

$$\hat{\sigma}^2 = MSE.$$

1.10 Έλεγχοι Υποθέσεων στο Πολλαπλό Γραμμικό Μοντέλο

Όπως είδαμε και στο απλό γραμμικό μοντέλο έτσι και στο πολλαπλό όταν κατασκευάζουμε ένα μοντέλο μας ενδιαφέρει να κάνουμε ελέγχους υποθέσεων και να κατασκευάζουμε διαστήματα εμπιστοσύνης που αφορούν τις παραμέτρους του. Για να γίνουν οι έλεγχοι υποθέσεων χρειάζεται να κάνουμε την υπόθεση ότι τα κατάλοιπα e_i ακολουθούν την κανονική κατανομή με μέση τιμή 0 και διακύμανση σ^2, καθώς και ότι είναι ασυσχέτιστα.

1.10.1 Επάρκεια του Μοντέλου

Αφού έχουμε εκτιμήσει τις παραμέτρους του επιλεγμένου μοντέλου, το πρώτο πράγμα που χρειάζεται να εξετάσουμε είναι αν υπάρχει γραμμική σχέση μεταξύ της μεταβλητής απόκρισης y και όλων των επεξηγηματικών μεταβλητών $x_1, x_2, \dots, x_k$. Με αυτή την διαδικασία εξετάζουμε ουσιαστικά την επάρκεια του μοντέλου μας (model adequacy). Οι υποθέσεις που κάνουμε είναι οι παρακάτω

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_\kappa = 0$$

$$H_1: \beta_j \neq 0 \text{ για τουλάχιστον ένα j.}$$

Αν απορρίψουμε την μηδενική υπόθεση σημαίνει ότι τουλάχιστον μία επεξηγηματική μεταβλητή $x_1, x_2, \dots, x_k$ συνεισφέρει σημαντικά στο μοντέλο.

Το τεστ που θα χρησιμοποιήσουμε είναι μία γενίκευση του F τεστ που είδαμε στο απλό γραμμικό μοντέλο. Η στατιστική συνάρτηση ελέγχου είναι η

$$F_0 = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE},$$

η οποία ακολουθεί την $F_{k,n-k-1}$ κατανομή. Απορρίπτουμε την μηδενική υπόθεση H_0 εάν ισχύει ότι:

$$F_0 > F_{k,n-k-1},$$

Για τον υπολογισμό του SSR χρησιμοποιούμε τον παρακάτω τύπο

$$SSR = \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \frac{\left(\sum_{i=1}^{y} y_i\right)^{\mathbf{2}}}{n},$$

ή εάν γνωρίζουμε την ολική μεταβλητότητα SST, τότε $SSR = SST - SSE$, όπου

$$SST = \boldsymbol{y}'\boldsymbol{y} - \frac{\left(\sum_{i=1}^{y} y_i\right)^{\mathbf{2}}}{n}.$$

Ένας εναλλακτικός τρόπος για να εξετάσουμε την επάρκεια του μοντέλου είναι μέσω του συντελεστή προσδιορισμού R^2και του προσαρμοσμένου συντελεστή προσδιορισμού R^2_{adj}. Θυμίζουμε ότι $R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR+SSE}$ $\kappa\alpha\iota$ $R^2_{adj} = 1 - \frac{SSE(n-p)}{SST/(n-1)}$.

Γνωρίζουμε ότι για κάθε επιπλέον επεξηγηματική μεταβλητή που θα προσθέσουμε στο μοντέλο μας ο συντελεστής προσδιορισμού R^2 θα αυξηθεί, ακόμα και αν η νέα μεταβλητή δεν συμβάλλει σημαντικά στο υπό εξέταση μοντέλο. Για αυτό τον λόγο πολλοί προτιμούν τον προσαρμοσμένο δείκτη R^2_{adj}. Ο λόγος είναι ο εξής: ο προσαρμοσμένος συντελεστής προσδιορισμού θα αυξηθεί μόνο αν η νέα μεταβλητή συνεισφέρει σημαντικά στο μοντέλο. Αυτό εξηγείται μαθηματικά. Αφού ο παρονομαστής μένει πάντα σταθερός όσες μεταβλητές και να προσθέσουμε στο μοντέλο, μας ενδιαφέρουν τελικά μόνο εκείνες οι μεταβλητές οι οποίες μειώνουν την τιμή του αριθμητή, δηλαδή του SSE. Για εκείνες τις μεταβλητές και μόνο ο R^2_{adj} θα αυξηθεί.

1.10.2 t-τεστ

Αφού έχουμε καταλήξει στο ότι τουλάχιστον μία επεξηγηματική μεταβλητή χρειάζεται στο μοντέλο μας, μένει να καταλήξουμε στο πόσες ακόμα χρειαζόμαστε και ποιες

θα είναι αυτές. Με την, προσθήκη κάθε νέας επεξηγηματικής μεταβλητής στο μοντέλο έχουμε αύξηση της τιμής συνολικής διακύμανσης SST και μείωση του αθροίσματος των τετραγώνων SSR. Χρειάζεται να αποφασίσουμε εάν η αύξηση του SST είναι αρκετή ώστε να μας εγγυηθεί ότι η νέα μεταβλητή είναι σημαντική για το μοντέλο. Γενικά χρησιμοποιούμε μόνο όσες μεταβλητές χρειαζόμαστε καθώς παραπάνω πληροφορία μπορεί να επιδράσει σημαντικά στην επάρκεια του μοντέλου.
Οι υποθέσεις με τις οποίες θα ελέγξουμε την σημαντικότητα μίας νέας μεταβλητής είναι οι παρακάτω

$$H_0: \beta_j = 0, \quad H_1: \beta_j \neq 0.$$

Εάν δεν απορρίψουμε την μηδενική υπόθεση για τον συντελεστή β_j, τότε μπορούμε να υποθέσουμε ότι η επεξηγηματική μεταβλητή x_j μπορεί να μην συμπεριληφθεί στο μοντέλο. Η στατιστική συνάρτηση ελέγχου είναι η

$$t_0 = \frac{\widehat{\beta_j}}{\sqrt{\hat{\sigma}^2 C_{jj}}} = \frac{\widehat{\beta_j}}{se(\widehat{\beta_j})},$$

όπου C_{jj} είναι το διαγώνιο στοιχείο του πίνακα $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ που αντιστοιχεί στον συντελεστή $\widehat{\beta_j}$. Απορρίπτουμε την μηδενική υπόθεση H_0 εάν ισχύει ότι

$$|t_0| > t_{\frac{\alpha}{2}, \nu-\kappa-1}.$$

Σημειώνουμε ότι το παραπάνω τεστ εξετάζει την συνεισφορά του j-συντελεστή παλινδρόμησης δεδομένου ότι υπάρχουν οι $j-1$ συντελεστές στο μοντέλο.

Παρακάτω θα δούμε ένα παράδειγμα πολλαπλού μοντέλου γραμμικής παλινδρόμησης από το βιβλίο του Boland (2007).

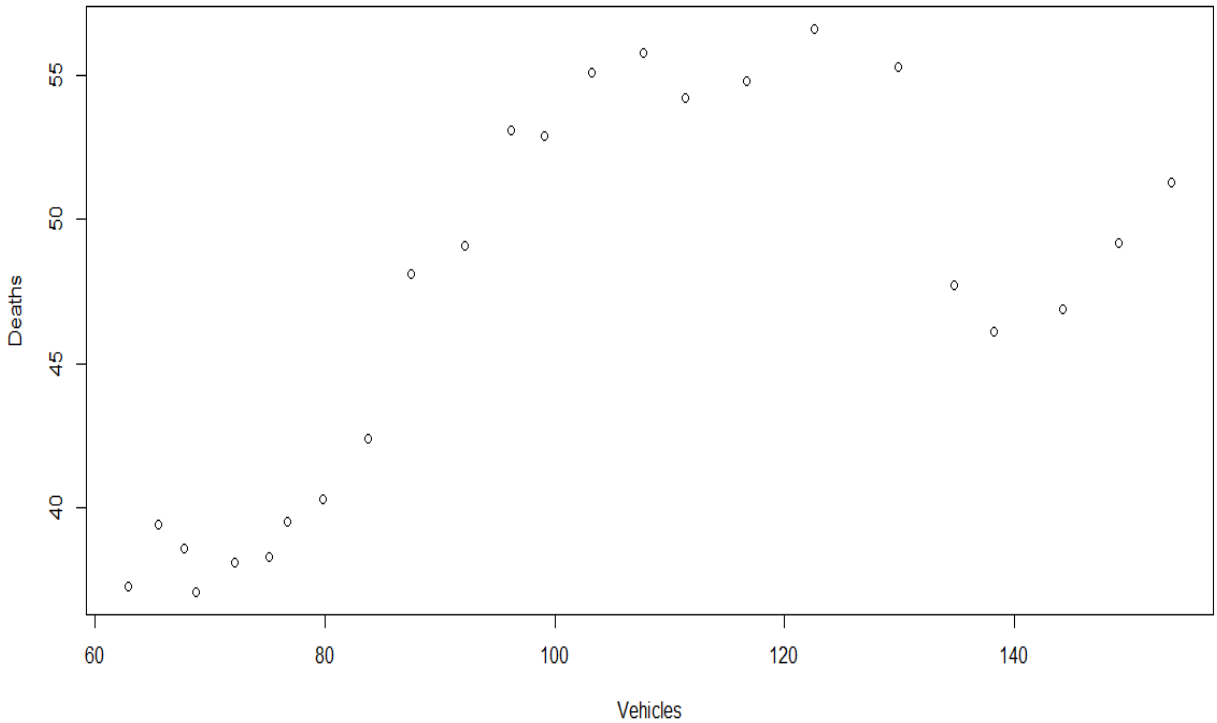
Παράδειγμα 1.2: *Στον παρακάτω πίνακα έχουμε καταγράψει τους ετήσιους θανάτους από τροχαία και τον αριθμό των οχημάτων που ενεπλάκησαν στα τροχαία για κάθε έτος. Το μοντέλο που επιλέξαμε για την προσομοίωση των παρακάτω δεδομένων είναι το*

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

i –Έτος	*Θάνατοι* (y_i)	*Αριθμός οχημάτων* (x_i)
1	*37.3*	*62.9*
2	*39.4*	*65.5*
3	*38.6*	*67.8*
4	*37.1*	*68.8*
5	*38.1*	*72.2*
6	*38.3*	*75.2*
7	*39.5*	*76.7*
8	*40.3*	*79.8*
9	*42.4*	*83.8*
10	*48.1*	*87.5*

11	*49.1*	*92.2*
12	*53.1*	*96.2*
13	*52.9*	*99.1*
14	*55.1*	*103.2*
15	*55.8*	*107.7*
16	*54.2*	*111.4*
17	*54.8*	*116.7*
18	*56.6*	*122.6*
19	*55.3*	*129.9*
20	*47.7*	*134.8*
21	*46.1*	*138.2*
22	*46.9*	*144.2*
23	*49.2*	*149.1*
24	*51.3*	*153.7*

Πίνακας 1.4: Δεδομένα παραδείγματος 1.2



Γράφημα 1.3: Διάγραμμα των σημείων για τα δεδομένα του παραδείγματος 1.2.

Όπως είδαμε οι εκτιμητές ελαχίστων τετραγώνων προκύπτουν από την λύση της εξίσωσης

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \ ,$$

όπου

$$\boldsymbol{X} = \begin{bmatrix} 1 & 62.9 & 3956.41 \\ 1 & 65.5 & 4290.25 \\ 1 & 67.8 & 4596.84 \\ 1 & 68.8 & 4733.44 \\ 1 & 72.2 & 5212.84 \\ 1 & 75.2 & 5655.04 \\ 1 & 76.7 & 5882.89 \\ 1 & 79.8 & 6368.04 \\ 1 & 83.8 & 7022.44 \\ 1 & 87.5 & 7656.25 \\ 1 & 92.2 & 8500.84 \\ 1 & 96.2 & 9254.44 \\ 1 & 99.1 & 9820.81 \\ 1 & 103.2 & 10650.24 \\ 1 & 107.7 & 11599.29 \\ 1 & 111.4 & 12409.96 \\ 1 & 116.7 & 13618.89 \\ 1 & 122.6 & 15030.76 \\ 1 & 129.9 & 16874.01 \\ 1 & 134.8 & 18171.04 \\ 1 & 138.2 & 19099.24 \\ 1 & 144.2 & 20793.64 \\ 1 & 149.1 & 22230.81 \\ 1 & 153.7 & 23623.69 \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} 37.3 \\ 39.4 \\ 38.6 \\ 37.1 \\ 38.1 \\ 38.3 \\ 39.5 \\ 40.3 \\ 42.4 \\ 48.1 \\ 49.1 \\ 53.1 \\ 52.9 \\ 55.1 \\ 55.8 \\ 54.2 \\ 54.8 \\ 56.5 \\ 55.3 \\ 47.4 \\ 46.1 \\ 46.9 \\ 49.2 \\ 51.3 \end{bmatrix}.$$

Ο πίνακας $\boldsymbol{X}'\boldsymbol{X}$ είναι ο

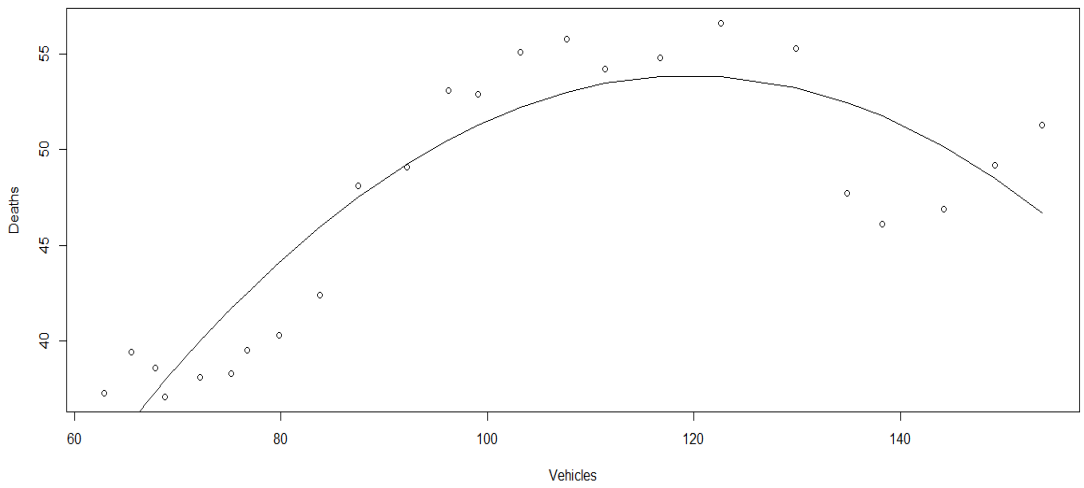
$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 24 & 2439.2 & 267052.1 \\ 2439.2 & 2607052.1 & 31221744.8 \\ 267052.1 & 31221744.8 & 3851938543.4 \end{bmatrix}$$

και οι εκτιμητές ελαχίστων τετραγώνων είναι οι

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} -34.415654325 \\ 1.476488182 \\ -0.006172108 \end{bmatrix}.$$

Άρα το πολλαπλό γραμμικό μοντέλο είναι το

$$\hat{y}_i = -34.415654325 + 1.476488182x_i - 0.006172108x_i^2.$$



Γράφημα 1.4: Διάγραμμα των σημείων για τα δεδομένα του παραδείγματος 1.2 και η εξίσωση ελαχίστων τετραγώνων .

Η εκτίμηση της διακύμανσης είναι

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-p} = \frac{\boldsymbol{y}'\boldsymbol{y} - \widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y}}{n-p} = \frac{203.4132}{24-3} = 9.686343.$$

Για να εξετάσουμε την επάρκεια του μοντέλου θα χρησιμοποιήσουμε την συνάρτηση ελέγχου

$$F_0 = \frac{SSR/k}{SSE/(n-k-1)} = \frac{\left(\widehat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{y} - \frac{\left(\sum_{i=1}^{y} y_i\right)^2}{n}\right)/2}{SSE/(24-3)} = \frac{456.62}{9.686343} = 47.1406.$$

Επίσης $F_{k,n-k-1} = F_{2,21} = 1.71727e-08$.
Προκύπτει ότι $F_0 > F_{2,21}$, επομένως απορρίπτουμε την μηδενική υπόθεση H_o. Για να ελέγξουμε την σημαντικότητα του συντελεστή $\hat{\beta}_2$ θα χρησιμοποιήσουμε την στατιστική συνάρτηση ελέγχου

$$t_0 = \frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2 C_{22}}} = -6.548564,$$

όπου C_{jj} είναι το διαγώνιο στοιχείο του πίνακα $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ που αντιστοιχεί στον συντελεστή $\hat{\beta}_2$. Επίσης για επίπεδο σημαντικότητας $\alpha = 0.01$ έχουμε ότι

$$t_{\frac{\alpha}{2},\nu-\kappa-1} = t_{\frac{0.01}{2},24-2-1} = 1.720743.$$

Προκύπτει ότι $|t_0| > t_{\frac{\alpha}{2},\nu-\kappa-1}$ απορρίπτουμε την μηδενική υπόθεση για $\alpha = 0.01$ και καταλήγουμε στο συμπέρασμα ότι ο εκτιμητής $\hat{\beta}_2$ είναι στατιστικά σημαντικός για το μοντέλο μας.
Ο συντελεστής προσδιορισμού είναι

$$R^2 = \frac{SSR}{SSE + SSR} = \frac{913.24}{913.24 + 203.4132} = 0.8178367,$$

και ο προσαρμοσμένος συντελεστής προσδιορισμού είναι

$$R^2_{adj} = 1 - (1 - R)\frac{n-1}{n-p-1} = 1 - (1 - 0.8178367)\frac{23}{21} = 0.8004878.$$

Τέλος θα παρουσιάσουμε τα παραπάνω αποτελέσματα μέσω της R.

```
> ##εισαγωγή των σημείων x,y σε διανύσματα
x<-c(62.9,65.5,67.8,68.8,72.2,75.2,76.7,79.8,83.8,87.5,92.2,96.2,99.1,103.2
,107.7,111.4,116.7,122.6,129.9,134.8,138.2,144.2,149.1,153.7)
y<-c(37.3,39.4,38.6,37.1,38.1,38.3,39.5,40.3,42.4,48.1,49.1 ,53.1,52.9,55.1
,55.8,54.2,54.8,56.6,55.3,47.7,46.1,46.9,49.2,51.3)
> data<-data.frame(x,y)

> ##συνάρτηση lm για τον υπολογισμό του πολλαπλού γραμμικού μοντέλου

> data<-data.frame(x,y)
> x2<-(x^2)
> mlm<-lm(formula= y~x+(x2),data=data)
> summary(mlm)

Call:
lm(formula = y ~ x + (x2), data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.6524 -3.0792  0.7063  2.6401  4.5874 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.442e+01  1.020e+01  -3.373  0.00288 ** 
x            1.476e+00  2.021e-01   7.306 3.42e-07 ***
x2          -6.172e-03  9.425e-04  -6.549 1.74e-06 ***
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.112 on 21 degrees of freedom
Multiple R-squared:  0.8178,	Adjusted R-squared:  0.8005 
F-statistic: 47.14 on 2 and 21 DF,  p-value: 1.717e-08
```

- Η στήλη **Estimate** περιέχει τις εκτιμήσεις των συντελεστών της παλινδρόμησης $\hat{\beta}_o, \hat{\beta}_1, \hat{\beta}_2$.
- Η στήλη **Std. Error** περιέχει τα αντίστοιχα τυπικά σφάλματα των συντελεστών της παλινδρόμησης.
- Η στήλη **t value** περιέχει την τιμή την τιμή του στατιστικού ελέγχου T για τον έλεγχο της μηδενικής υπόθεσης $H_0: \hat{\beta}_j = 0$.
- Η στήλη **Pr(>|t|)** περιέχει τις $p - value$ τιμές των 2 ουρών. Για να πάρουμε την τιμή της μίας διαιρούμε με 2.
- Η γραμμή **Multiple R-squared** περιέχει την τιμή του συντελεστή προσδιορισμού R^2, ενώ **Adjusted R-squared** περιέχει την τιμή του προσαρμοσμένου συντελεστή προσδιορισμού R^2_{adj}.
- Η γραμμή **F-Statistic** περιέχει την τιμή του στατιστικού ελέγχου F για τον έλεγχο της μηδενικής υπόθεσης $H_0: \hat{\beta}_1 = \hat{\beta}_2 = \hat{\beta}_3 = 0$ και την τιμή $p - value$ για αυτή την τιμή.
- Η γραμμή **Signif. Codes** μας δείχνει πόσο σίγουροι μπορούμε να είμαστε ότι ο αντίστοιχος συντελεστής παλινδρόμησης είναι σημαντικός για την εξαρτημένη

μεταβλητή. Τα αστεράκια χρησιμοποιούνται για μία γρήγορη κατάταξη μεταξύ των επεξηγηματικών μεταβλητών που χρησιμοποιούμε στο μοντέλο μας. Παρατηρούμε ότι τα αποτελέσματά μας συμφωνούν. Κάποιες επιπλέον συναρτήσεις για να πάρουμε μεμονωμένα στοιχεία του μοντέλου είναι οι παρακάτω:

Η συνάρτηση **coefficients( )** μας δίνει τους εκτιμητές ελαχίστων τετραγώνων της παλινδρόμησης $\hat{\beta}_o, \hat{\beta}_1, \hat{\beta}_2$ οι οποίοι είναι αντίστοιχα -34.415654325, 1.476488182, -0.006172108.

```
> coefficients(mlm)
  (Intercept)            x           x2
-34.415654325   1.476488182  -0.006172108
```

Η συνάρτηση **fitted( )** μας δίνει τις προβλεπόμενες τιμές $\hat{y}_i$ οι οποίες προκύπτουν αν αντικατ αστήσουμε στην ευθεία παλινδρόμησης τα δεδομένα μας x_i.

```
> fitted(mlm)
        1        2        3        4        5        6
34.03606 35.81444 37.31805 37.95143 40.01258 41.71274
        7        8        9       10       11       12
42.52116 44.10388 45.97080 47.52186 49.24846 50.50311
       13       14       15       16       17       18
51.28923 52.22350 53.01006 53.46952 53.83326 53.83033
       19       20       21       22       23       24
53.23196 52.46134 51.75245 50.15336 48.51778 46.71262
```

Η συνάρτηση **residuals( )** μας δίνει τα κατάλοιπα ε_i του μοντέλου μας. Την διαφορά δηλαδή της πραγματικής τιμής των y_i με την εκτιμώμενη τιμή αυτών $\hat{y}_i$

```
> residuals(mlm)
         1          2          3          4          5
 3.2639355  3.5855626  1.2819462 -0.8514321 -1.9125836
         6          7          8          9         10
-3.4127422 -3.0211597 -3.8038752 -3.5708007  0.5781365
        11         12         13         14         15
-0.1484577  2.5968898  1.6107706  2.8765001  2.7899419
        16         17         18         19         20
 0.7304781  0.9667366  2.7696698  2.0680432 -4.7613403
        21         22         23         24
-5.6524499 -3.2533601  0.6822156  4.5873750
```

Συνοψίζοντας, είδαμε την γενική μορφή ενός γραμμικού μοντέλου παλινδρόμησης. Οι υποθέσεις που έγιναν για την χρήση του μοντέλου συνοψίζονται παρακάτω:

- Γραμμικότητα: Θεωρούμε πως υπάρχει γραμμική σχέση μεταξύ των συντελεστών παλινδρόμησης και της μεταβλητής απόκρισης καθώς επίσης ότι έχουν χρησιμοποιηθεί όλοι οι απαραίτητοι για το μοντέλο συντελεστές παλινδρόμησης.
- Σταθερή διακύμανση: Οι μεταβλητές απόκρισης y_i έχουν σταθερή διακύμανση και είναι ανεξάρτητες μεταξύ τους.
- Τα τυχαία σφάλματα ε_i έχουν μέση τιμή 0 και διασπορά σ^2 και είναι ασυσχέτιστα μεταξύ τους.
- Οι μεταβλητές απόκρισης y_i ακολουθούν την κανονική κατανομή.

Στο επόμενο κεφάλαιο θα αναλύσουμε τα γενικευμένα γραμμικά μοντέλα τα οποία αποτελούν επέκταση του απλού και του πολλαπλού μοντέλου παλινδρόμησης.

ΚΕΦΑΛΑΙΟ 2

Γενικευμένα Γραμμικά Μοντέλα

2.1 Η Μέθοδος Μέγιστης Πιθανοφάνειας

Στο προηγούμενο κεφάλαιο μελετήσαμε το απλό γραμμικό μοντέλο για το οποίο υποθέσαμε ότι έχει σταθερή διακύμανση, η οποία συνήθως προέρχεται από μία κανονική κατανομή. Υπάρχουν πολλοί τύποι μεταβλητών για τους οποίους αυτή η υπόθεση δεν μπορεί να γίνει γι' αυτό χρειαζόμαστε άλλες μεθόδους εκτίμησης των παραμέτρων. Για παράδειγμα η μεταβλητή απόκρισης μπορεί να είναι μια αναλογία η οποία κυμαίνεται μεταξύ του 0 και του 1 (επιτυχία, αποτυχία). Για τέτοιους τύπους μεταβλητών συνήθως υποθέτουμε την διωνυμική κατανομή. Επίσης η μεταβλητή απόκρισης μπορεί να είναι μια διακριτή τυχαία μεταβλητή η οποία εκφράζει πλήθος (π.χ. ατυχημάτων, κινδύνων κλπ.). Αυτού του τύπου τυχαίες μεταβλητές είναι όπως είπαμε διακριτές και μη αρνητικές οπότε η κανονική κατανομή δεν μας καλύπτει. Μια πιο ταιριαστή κατανομή για να μοντελοποιήσουμε τέτοιου είδους δεδομένα είναι συνήθως η κατανομή Poisson. Τέλος μπορεί η μεταβλητή απόκρισης να είναι συνεχής και θετική. Σε τέτοιες περιπτώσεις μπορούμε να χρησιμοποιήσουμε την Γάμμα κατανομή ή την αντίστροφη Γκαουσιανή.

Για την εκτίμηση των παραμέτρων του απλού γραμμικού μοντέλου χρησιμοποιήσαμε την μέθοδο των ελαχίστων τετραγώνων. Η μέθοδος των ελαχίστων τετραγώνων είναι κατάλληλη μέθοδος αν θέλουμε να μοντελοποιήσουμε δεδομένα για τα οποία έχουμε υποθέσει ότι προέρχονται από την κανονική κατανομή. Στην συνέχεια θα αναπτύξουμε μια πιο γενική μέθοδο εκτίμησης των παραμέτρων την μέθοδο της μέγιστης πιθανοφάνειας της οποίας ειδική περίπτωση είναι η μέθοδος ελαχίστων τετραγώνων. Η μέθοδος αυτή είναι κατάλληλη για την εκτίμηση των παραμέτρων μοντέλων με δεδομένα που δεν προέρχονται από την κανονική κατανομή και προέρχονται από άλλες κατανομές όπως για παράδειγμα είναι η Poisson, η διωνυμική, η εκθετική και η Γάμμα κατανομή.

Η μέθοδος της μέγιστης πιθανοφάνειας απαιτεί να γνωρίζουμε εκ των προτέρων την κατανομή που ακολουθούν τα δεδομένα μας άρα και την συνάρτηση πυκνότητας πιθανότητας σ.π.π. (για συνεχή κατανομή) ή τη συνάρτηση πιθανότητας σ.π. (για διακριτή κατανομή). Έχει ως στόχο να επιλέξει αυτές τις τιμές για τις άγνωστες παραμέτρους οι οποίες μεγιστοποιούν την σ.π.π. ή την σ.π. των παρατηρούμενων τιμών. Δηλαδή καθορίζουμε τις άγνωστες παραμέτρους με τέτοιο τρόπο ώστε να μεγιστοποιείται η πιθανότητα εμφάνισης του συγκεκριμένου δείγματος. Σε ειδικές περιπτώσεις η μέθοδος χρησιμοποιήθηκε από τον Gauss, αλλά και νωρίτερα από τον Laplace, όμως ως γενική μέθοδος εκτίμησης προτάθηκε, ονομάστηκε και καθιερώθηκε από τον Fisher σε μια σειρά εργασιών του (1912, 1922, 1925, 1934), όπου μελέτησε τις ιδιότητές της. Η μέθοδος της μέγιστης πιθανοφάνειας έχει πλέον ταυτιστεί με το όνομα του Fisher, μπορεί να εφαρμοστεί εύκολα σε πάρα πολλά προβλήματα εκτίμησης, ερμηνεύεται διαισθητικά πολύ απλά και γενικά παράγει καλούς εκτιμητές ειδικά για μεγάλο μέγεθος δείγματος. Στο υπόλοιπο του κεφαλαίου η βιβλιογραφία που χρησιμοποιείται είναι από την Dobson (2002), και από τους McCullagh & Nelder (1989).

Ορισμός 2.1: Έστω $y_1, y_2, y_3, \dots, y_n$ είναι ένα τυχαίο δείγμα ανεξάρτητων παρατηρήσεων που προέρχονται από μία κατανομή με σ.π.π. ή σ.π. $f(y;\theta)$. Η από κοινού σ.π.π. ή σ.π. των $y_1, y_2, y_3, \dots, y_n$, η οποία είναι συνάρτηση της παραμέτρου θ, ονομάζεται συνάρτηση πιθανοφάνειας του τυχαίου δείγματος και δίνεται από την σχέση

$$L(\theta) = f(y_1, y_2, \dots, y_n; \theta) = \prod_{i=1}^{n} f(y_i; \theta). \tag{2.1}$$

Όπως αναφέραμε στην εισαγωγή του δευτέρου κεφαλαίου, η μέθοδος μέγιστης πιθανοφάνειας συνίσταται στην μεγιστοποίηση της συνάρτησης πιθανοφάνειας $L(\theta)$ ως προς την παράμετρο θ.

Ορισμός 2.2: Ο εκτιμητής μέγιστης πιθανοφάνειας (Ε.Μ.Π.) της παραμέτρου θ ορίζεται ως η τιμή του θ που μεγιστοποιεί ολικά την συνάρτηση πιθανοφάνειας

$$L(\theta) = \prod_{i=1}^{n} f(y_i; \theta).$$

Το ολικό μέγιστο μπορεί να επιτυγχάνεται σε μία τιμή του θ, ή σε περισσότερες από μία τιμές ή να μην υπάρχει. Ανάλογα ο Ε.Μ.Π. μπορεί να είναι μοναδικός ή να υπάρχουν πολλοί ή να μην υπάρχει. Εάν η συνάρτηση $L(\theta)$ παραγωγίζεται ως προς θ, το μέγιστο υπάρχει και επιτυγχάνεται σε εσωτερικό σημείο της συνάρτησης τότε μπορεί να βρεθεί με παραγώγιση. Σε αυτές τις περιπτώσεις λόγω της μορφής της $L(\theta)$ είναι συχνά πιο εύκολο να μεγιστοποιήσουμε τον (νεπέριο) λογάριθμο $l(\theta) = \log L(\theta)$. Προφανώς, κάθε τιμή του θ που μεγιστοποιεί την συνάρτηση $l(\theta) = \log\ L(\theta)$ επίσης μεγιστοποιεί και την συνάρτηση $L(\theta)$, γιατί ο λογάριθμος είναι γνησίως αύξουσα συνάρτηση. Από τα παραπάνω ο λογάριθμος της συνάρτησης πιθανοφάνειας έχει την παρακάτω μορφή

$$\begin{aligned} l(\theta) &= \log L(\theta) \\ &= \log f(y_1, y_2, \dots., y_n; \theta) \\ &= \log \prod_{i=1}^{n} f(y_i; \theta) \\ &= \sum_{i=1}^{n} \log f(y_i; \theta). \end{aligned} \tag{2.2}$$

Για να βρούμε την μέγιστη τιμή της παραμέτρου θ παραγωγίζουμε την (2.2) ως προς θ και την θέτουμε ίση με μηδέν (0). Η (2.2) γίνεται δηλαδή,

$$\frac{\partial}{\partial\theta} l(\theta) = \sum_{i=1}^{n} \frac{\frac{\partial}{\partial\theta} f(y_i; \theta)}{f(y_i; \theta)} = 0, \tag{2.3}$$

η οποία αναφέρεται ως εξίσωση πιθανοφάνειας. Στην περίπτωση που το θ είναι διανυσματική παράμετρος $\theta = (\theta_1, \theta_2, \dots, \theta_r)$ και το μέγιστο μπορεί να βρεθεί με παραγώγιση, ο ε.μ.π. είναι λύση ως προς $\theta_1, \theta_2, \dots, \theta_r$ του συστήματος των εξισώσεων πιθανοφάνειας

$$\sum_{i=1}^{n} \frac{\frac{\partial}{\partial \theta_j} f(y_i;\theta)}{f(y_i;\theta)} = 0. \tag{2.4}$$

Λύνοντας την (2.3) ή τις (2.4) πρέπει περαιτέρω να διαπιστώνεται ότι η λύση αντιστοιχεί σε ολικό μέγιστο. Ανάλογα με την μορφή της (2.1) οι εξισώσεις (2.3), (2.4) μπορούν να επιλυθούν είτε σε κλειστή μορφή ως συναρτήσεις του y_i είτε σε ανοιχτή μορφή με την χρήση αριθμητικών μεθόδων επίλυσης. Παρακάτω δίνουμε μερικά παραδείγματα υπολογισμού εκτιμητών μέγιστης πιθανοφάνειας.

Παράδειγμα 2.1: Έστω $y_1, y_2, \dots, y_n$ είναι ένα δείγμα που προέρχεται από έναν πληθυσμό από την εκθετική κατανομή με παράμετρο $\theta > 0$. Η σ.π.π. της εκθετικής κατανομής δίνεται από

$$f(y;\theta) = \theta e^{-\theta y}.$$

Η συνάρτηση πιθανοφάνειας είναι

$$L(\theta) = f(y_1, y_2, \dots, y_n;\theta) = \prod_{i=1}^{n} f(y_i;\theta) = \theta e^{-\theta y_1} \theta e^{-\theta y_2} \dots \theta e^{-\theta y_n} = \theta^n e^{-\theta(\sum_{=1}^{n} y_i)}$$

και ο λογάριθμος της συνάρτησης πιθανοφάνειας θα είναι

$$\log L(\theta) = \log \theta^n e^{-\theta(\sum_{=1}^{n} y_i)} = n\log\theta - \theta\left(\sum_{i=1}^{n} y_i\right).$$

Στην συνέχεια παραγωγίζοντας ως προς θ και θέτοντας την παράγωγο ίση με μηδέν βρίσκουμε τον ε.μ.π. της παραμέτρου θ τον οποίο θα συμβολίζουμε από εδώ και στο εξής με $\hat{\theta}$.

$$\frac{\partial}{\partial\theta} l(\theta) = 0 \Leftrightarrow \frac{n}{\theta} - \left(\sum_{i=1}^{n} y_i\right) = 0 \Leftrightarrow \frac{n}{\theta} = \sum_{i=1}^{n} y_i \Leftrightarrow \hat{\theta} = \frac{n}{\sum_{i=1}^{n} y_i},$$

ο οποίος αποτελεί ολικό μέγιστο, αφού $\frac{\partial^2}{\partial\theta} l(\theta) = -\frac{n}{\theta^2} < 0$.

Όπως γνωρίζουμε η μέση τιμή και η διακύμανση της εκθετικής κατανομής με παράμετρο θ είναι $E[Y] = \frac{1}{\theta}$ και $Var[Y] = \frac{1}{\theta^2}$ αντίστοιχα. Αντικαθιστώντας με τον Ε.Μ.Π που βρήκαμε μπορούμε να έχουμε μια εκτίμηση της μέσης τιμής και της διακύμανσης του τυχαίου δείγματος.

Παράδειγμα 2.2: Έστω $y_1, y_2, \dots, y_n$ είναι ένα δείγμα που προέρχεται από έναν πληθυσμό από την κανονική κατανομή $N\sim(\mu, \sigma^2)$. Διακρίνουμε 3 περιπτώσεις:

<u>1η περίπτωση:</u> σ^2 γνωστό, $\mu = \theta$ άγνωστο με $\theta \in \mathbb{R}$. Τότε η σ.π.π. της κανονικής κατανομής δίνεται από:

$$f(y;\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\theta)^2}.$$

Η συνάρτηση πιθανοφάνειας είναι

$$L(\theta) = f(y_1, y_2, \dots, y_n;\theta) = \prod_{i=1}^{n} f(y_i;\theta) = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2}$$

και ο λογάριθμος της συνάρτησης πιθανοφάνειας θα είναι

$$\log L(\theta) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2.$$

Στην συνέχεια παραγωγίζοντας ως προς θ και θέτοντας την παράγωγο ίση με μηδέν βρίσκουμε τον ε.μ.π. της παραμέτρου θ

$$\frac{\partial}{\partial\theta} l(\theta) = 0 \Leftrightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \theta) = 0 \Leftrightarrow \sum_{i=1}^{n} y_i - n\theta = 0 \Leftrightarrow \hat{\theta} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y},$$

ο οποίος αποτελεί ολικό μέγιστο, αφού $\frac{\partial^2}{\partial\theta} l(\theta) = -\frac{n}{\sigma^2} < 0$.

<u>2η περίπτωση:</u> μ γνωστό, $\sigma^2 = \theta$ άγνωστο με $\theta \in (0, \infty)$. Τότε η σ.π.π. της κανονικής κατανομής δίνεται από

$$f(y;\theta) = \frac{1}{\theta^{1/2}\sqrt{2\pi}} e^{-\frac{1}{2\theta}(y-\mu)^2},$$

επομένως έχουμε

$$L(\theta) = (y_1, y_2, \dots, y_n; \theta) = \prod_{i=1}^{n} f(y_i;\theta) = \frac{1}{\theta^{n/2}(2\pi)^{n/2}} e^{-\frac{1}{2\theta}\sum_{i=1}^{n}(y_i-\mu)^2}$$

και ο λογάριθμος της συνάρτησης πιθανοφάνειας θα είναι

$$\log L(\theta) = -\frac{n}{2}\log\theta - \frac{n}{2}\log 2\pi - \frac{1}{2\theta}\sum_{i=1}^{n}(y_i - \mu)^2.$$

Στην συνέχεια παραγωγίζοντας ως προς θ και θέτοντας την παράγωγο ίση με μηδέν βρίσκουμε τον ε.μ.π. της παραμέτρου θ

$$\frac{\partial}{\partial\theta} l(\theta) = 0 \Leftrightarrow -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}(y_i - \mu)^2 = 0 \Leftrightarrow \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2,$$

ο οποίος αποτελεί ολικό μέγιστο, αφού $\frac{\partial^2}{\partial\theta} l(\theta) = -\frac{n}{2\theta^2} < 0$.

<u>3η περίπτωση:</u> μ, σ^2 άγνωστα, οπότε θ=(μ, σ^2) με $\theta \in \mathbb{R} \times (0, \infty)$. Τότε η σ.π.π. της κανονικής κατανομής δίνεται από:

$$f(y;\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2},$$

επομένως έχουμε

$$L(\theta) = (y_1, y_2, \dots, y_n; \theta) = \prod_{i=1}^{n} f(y_i;\theta) = \frac{1}{\sigma^n(2\pi)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}$$

και ο λογάριθμος της συνάρτησης πιθανοφάνειας θα είναι

$$\log L(\theta) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

Στην συνέχεια παραγωγίζοντας ως προς μ, σ^2 και θέτοντας την παράγωγο ίση με μηδέν βρίσκουμε

$$\frac{\partial}{\partial\sigma^2} l(\theta) = 0 \Leftrightarrow -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i - \mu)^2 = 0,$$

$$\frac{\partial}{\partial\mu} l(\theta) = 0 \Leftrightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu) = 0.$$

Τότε λύνοντας το σύστημα των εξισώσεων έχουμε

$$\mu = \bar{y},$$
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

2.2 Συνάρτηση Score

Εάν έχουμε μόνο μία παράμετρο προς εκτίμηση έστω θ, η παράγωγος του λογαρίθμου της συνάρτησης πιθανοφάνειας ονομάζεται συνάρτηση score, (Dunn (2018)) ορίζεται ως $U(\theta) = \frac{dl}{d\theta}$, και η εξίσωση $U(\hat{\theta}) = 0$ που λύνουμε για την εύρεση του $\hat{\theta}$ ονομάζεται score εξίσωση. Για τις score συναρτήσεις ισχύουν τα παρακάτω:

- $E(U(\theta)) = 0.$
- $Var(U(\theta)) = E(U(\theta)^2).$

Ορίζουμε την ποσότητα

$$J(\theta) = -\frac{d^2 l(y;\theta)}{d\theta^2} = -\frac{dU(\theta)}{d\theta},$$

να είναι η παράγωγος της συνάρτησης score.

Ορισμός 2.3 Η ποσότητα $I(\theta) = E[J(\theta)]$ ονομάζεται πληροφορία Fisher. Η πληροφορία Fisher μετράει την ποσότητα της πληροφορίας που περιέχει η παρατήρηση y για την παράμετρο θ. Η διασπορά του εκτιμητή μέγιστης πιθανοφάνειας προσεγγιστικά θα είναι ίση με

$$Var(\hat{\theta}) = \frac{1}{I(\theta)}.$$

2.3 Μέθοδος Μέγιστης Πιθανοφάνειας στα Μοντέλα Παλινδρόμησης

Τα μοντέλα παλινδρόμησης έχουν σαν γενική μορφή την εξής

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip},$$

όπου $\eta_i = g(\mu_i)$ για κάποια γνωστή συνάρτηση g και $\mu_i = E[y_i]$, όπου y_i είναι οι μεταβλητές απόκρισης του μοντέλου μας και φ θα ονομάζεται η παράμετρος κλίμακας (dispersion parameter) η οποία θα δηλώνει την διασπορά των y_i.

Για τα μοντέλα παλινδρόμησης ο λογάριθμος της συνάρτησης πιθανοφάνειας είναι ο

$$l(\beta_0, \beta_1, \dots, \beta_p; y) = \sum_{i=1}^{n} \log f\ (y_i;\ \mu_i\ , \varphi).$$

Οι συναρτήσεις score είναι της μορφής

$$U(\beta_j) = \frac{\partial l(\beta_0, \beta_1, \dots, \beta_p; y)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{f(y_i; \mu_i, \varphi)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j}.$$

Για κάθε άγνωστη παράμετρο παλινδρόμησης β_j αντιστοιχεί μία συνάρτηση score.

Παράδειγμα 2.2 Έστω ότι έχουμε το μοντέλο παλινδρόμησης της παρακάτω μορφής

$$\log \frac{\mu}{1-\mu} = \eta = \beta_0 + \beta_1 x,$$

με y_i να ακολουθούν την κατανομή Bernoulli με πιθανότητα επιτυχίας μ_i, η οποία έχει συνάρτηση πιθανότητας $f(y_i; \mu_i) = {\mu_i}^{y_i}(1-\mu_i)^{1-y_i}$. Οι παράμετροι προς εκτίμηση είναι οι β_0, β_1, άρα χρειαζόμαστε 2 συναρτήσεις score τις $U(\beta_0), U(\beta_1)$ Αρχικά λύνουμε ως προς μ και έχουμε $\mu = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$. Συνεπώς, οι μερικές παράγωγοι της μ ως προς τις παραμέτρους β_0 και β_1 είναι $\frac{\partial \mu}{\partial \beta_0} = \mu(1-\mu)$ και $\frac{\partial \mu}{\partial \beta_1} = \mu(1-\mu)x$. Αν θεωρήσουμε πως έχουμε μόνο μία παρατήρηση y τότε οι συναρτήσεις score θα είναι:

$$U(\beta_0) = \frac{\partial \log\, f(y;\mu)}{\partial \beta_0} = \frac{d\log\, f(y;\mu)}{d\beta_0} \times \frac{\partial \mu}{\partial \beta_0} = y - \mu,$$

$$U(\beta_1) = \frac{\partial \log\, f(y;\mu)}{\partial \beta_1} = \frac{d\log\, f(y;\mu)}{d\beta_0} \times \frac{\partial \mu}{\partial \beta_1} = (y-\mu)x.$$

Επομένως, οι score εξισώσεις θα είναι:

$$U(\hat{\beta}_0) = \sum_{i=1}^{n} y_i - \hat{\mu}_i = 0,$$

$$U(\hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)x_i = 0,$$

όπου $\log\left\{\frac{\hat{\mu}_i}{1-\hat{\mu}_i}\right\} = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Λύνοντας ταυτόχρονα τις παραπάνω εξισώσεις βρίσκουμε τα $\hat{\beta}_0, \hat{\beta}_1$.

2.4 Εισαγωγή στα Γενικευμένα Γραμμικά Μοντέλα

Τα Γενικευμένα Γραμμικά Μοντέλα (Generalized Linear Models) προτάθηκαν από τους John Nelder και Robert Weddeburn το 1972 ως ένας τρόπος ενοποίησης άλλων στατιστικών μοντέλων όπως για παράδειγμα τις Λογιστικής Παλινδρόμησης, της Γραμμικής Παλινδρόμησης, της παλινδρόμησης Poisson και άλλων. Είναι πολύ διαδεδομένα στην στατιστική γιατί επιτρέπουν την ανάλυση τόσο ποιοτικών όσο και ποσοτικών μεταβλητών, καθώς και συνεχών αλλά και διακριτών μεταβλητών. Θεωρούνται μια φυσική επέκταση των απλών γραμμικών μοντέλων που επιτρέπει στην μέση τιμή ενός πληθυσμού να εξαρτάται από μία γραμμική παράμετρο πρόβλεψης (linear predictor) μέσα από μία μη γραμμική συνάρτηση σύνδεσης (link function). Επίσης επιτρέπει στην κατανομή της εξαρτημένης μεταβλητής ή μεταβλητή απόκρισης (response variable) να προέρχεται από οποιαδήποτε κατανομή της

εκθετικής οικογένειας κατανομών (exponential family). Η εκθετική οικογένεια κατανομών περιλαμβάνεις τις πιο κοινές κατανομές όπως είναι η κανονική, η διωνυμική, η κατανομή Poisson και άλλες.
Προτού περιγράψουμε την μορφή των γενικευμένων γραμμικών μοντέλων ας δούμε τα 3 διαδικασίες τις οποίες θα πρέπει να λάβουμε υπόψη:

- Επιλογή μοντέλου
- Εκτίμηση παραμέτρων
- Πρόβλεψη μελλοντικών τιμών (McCullagh and Nelder, 1989)

Επιλογή μοντέλου
Το πρόβλημα της επιλογής του καταλληλότερου μοντέλου συνοψίζεται στο εξής ερώτημα. Ποιες επεξηγηματικές μεταβλητές θα πρέπει να προσθέσουμε στο συστημικό όρο του μοντέλου μας έτσι ώστε οι προβλεπόμενες τιμές μας να είναι όσο γίνεται πιο κοντά στις πραγματικές τιμές και ταυτόχρονα να αποφεύγεται η πολυπλοκότητα στον προσδιορισμό του μοντέλου μας; Προφανώς όσο πιο πολλές οι επεξηγηματικές μεταβλητές x_i τόσο πιο κοντά είναι τα αποτελέσματα στα πραγματικά. Πρέπει όμως να υπάρξει μια ισορροπία μεταξύ του κόστους του να προσθέσουμε μια επιπλέον επεξηγηματική μεταβλητή x_i το οποίο κόστος αντιπροσωπεύεται μέσω της πολυπλοκότητας.
Εκτίμηση παραμέτρων
Έστω ότι έχουμε επιλέξει το κατάλληλο μοντέλο. Το επόμενο που χρειάζεται είναι να εκτιμήσουμε τις παραμέτρους αυτού. Αυτό θα γίνει μέσω της μεθόδου της μέγιστης πιθανοφάνειας που αναπτύξαμε παραπάνω
Πρόβλεψη μελλοντικών τιμών
Με την προϋπόθεση ότι το μοντέλο που επιλέξαμε είναι σωστό και ότι η διαδικασία με την οποία παράχθηκαν τα δεδομένα που χρησιμοποιήσαμε για την κατασκευή του μοντέλου παραμένει σταθερή χρειάζεται να χρησιμοποιήσουμε κάποια μέτρα ακριβείας του μοντέλου έτσι ώστε να μπορούμε να προβλέψουμε μελλοντικές τιμές είτε για διαφορετικά δεδομένα στον ίδιο χρόνο είτε για τα ίδια σε διαφορετικό χρόνο.
Από το απλό στο γενικευμένο
Τα Γενικευμένα Γραμμικά Μοντέλα είναι μια επέκταση των κλασσικών γραμμικών μοντέλων. Για να απλοποιήσουμε την μετάβαση από το απλό γραμμικό μοντέλο στο γενικευμένο γραμμικό μοντέλο ας δούμε τα παρακάτω:
Όπως αναλύσαμε στο Κεφάλαιο 1 για το απλό γραμμικό μοντέλο είχαμε τα εξής :

Ο τυχαίος όρος ή μεταβλητή απόκρισης Y ακολουθεί κανονική κατανομή με μέση τιμή μ και σταθερή διακύμανση σ^2,δηλαδή

$$Y \sim N(\mu, \sigma^2)$$

Ο όρος πρόβλεψης (linear predictor) είναι γραμμικής μορφής ως προς τους συντελεστές β_i της παλινδρόμησης. Πιο συγκεκριμένα είναι της μορφής:

$$\eta = \sum_{i=1}^{n} x_i \ \beta_i$$

Η σχέση μεταξύ του τυχαίου όρου και του συστημικού όρου είναι της μορφής

$$\mu = \eta$$

Μέσω αυτής της γενίκευσης βλέπουμε ένα νέο σύμβολο το οποίο θα χρησιμοποιούμε από εδώ και πέρα για τον όρο πρόβλεψης. Για το απλό γραμμικό μοντέλο βλέπουμε ότι το η και το μ είναι ταυτόσημα. Στην πραγματικότητα όμως ορίζουμε ως

$$\eta_i = g(\mu_i),$$

όπου g ονομάζεται η συνάρτηση σύνδεσης (link function).
Τα Γενικευμένα Γραμμικά Μοντέλα μας επιτρέπουν δύο επεκτάσεις.

1. Η κατανομή του τυχαίου όρου Y μπορεί να προέρχεται από οποιαδήποτε κατανομή της εκθετικής οικογένειας κατανομών.
2. Η συνάρτηση σύνδεσης g μπορεί να είναι οποιαδήποτε μονότονη, παραγωγίσιμη συνάρτηση.

2.4.1 Εκθετική Οικογένεια Κατανομών (ΕΟΚ)

Ας θεωρήσουμε μια τυχαία μεταβλητή Y, για την οποία η σ.π.π. ή η σ.π εξαρτάται από μία παράμετρο θ. Θα λέμε ότι η κατανομή ανήκει στην εκθετική οικογένεια κατανομών (ΕΟΚ) αν μπορεί να γραφεί στην μορφή

$$f(y;\theta) = s(y)t(\theta)exp\{a(y)b(\theta)\}, \tag{2.5}$$

όπου $a(y), b(\theta), s(y), t(\theta)$ είναι γνωστές συναρτήσεις. Η (2.5) μπορεί να γραφεί εναλλακτικά ως

$$f(y;\theta) = exp\{a(y)b(\theta) + c(\theta) + d(y)\}, \tag{2.6}$$

όπου $s(y) = exp\ (d(y))$ και $t(\theta) = exp\ (c(\theta))$. Αν $a(y) = y$ λέμε ότι η κατανομή βρίσκεται στην κανονική της μορφή. Το $b(\theta)$ πολλές φορές ονομάζεται ως η φυσική παράμετρος της κατανομής.

1. **Κανονική κατανομή. $Y \sim N(\mu, \sigma^2)$**

Η συνάρτηση πυκνότητας της κανονικής κατανομής είναι της μορφής ,

$$f(y;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}, \quad -\infty < y < \infty,$$

όπου η μέση τιμή μ είναι η παράμετρος που μας ενδιαφέρει να εκτιμήσουμε και σ^2 είναι η διακύμανση την οποία θεωρούμε γνωστή. Η παραπάνω σχέση μπορεί να γραφεί ως

$$f(y;\mu) = exp\left\{-\frac{y^2}{\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\}$$

και είναι της μορφής (2.6) με $b(\mu) = \mu/\sigma^2$, η φυσική παράμετρος και $c(\mu) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$ και $d(y) = -\frac{y^2}{\sigma^2}$

Η κανονική κατανομή χρησιμοποιείται για την μοντελοποίηση συνεχών μεταβλητών, τα οποία έχουν συμμετρική κατανομή. Οι 3 κύριες περιπτώσεις στις οποίες την συναντάμε είναι οι εξής:

1. Για να περιγράψουμε φυσικά φαινόμενα όπως είναι για παράδειγμα το ύψος, ή η πίεση του αίματος στους ανθρώπους.
2. Ακόμα και αν τα δεδομένα δεν είναι κανονικά κατανεμημένα, το μέσο ή το συνολικό ενός τυχαίου δείγματος προσεγγιστικά θα ακολουθεί την κανονική κατανομή. Αυτό το αποτέλεσμα εξηγείται από το Κεντρικό Οριακό Θεώρημα.

3. Εξαιτίας του ότι έχει αναπτυχθεί πολύ μεγάλη θεωρία και στατιστικά αποτελέσματα για την κανονική κατανομή, ακόμα και αν η τυχαία συνεχής μεταβλητή y δεν ακολουθεί την κανονική κατανομή αξίζει συχνά να προσπαθήσουμε μέσω μετασχηματισμού αυτής να την κανονικοποιήσουμε. Αυτό μπορεί να γίνει με διάφορους μετασχηματισμούς όπως για παράδειγμα θέτοντας $y' = logy$ ή $y' = \sqrt{y}$. Σκοπός είναι η τυχαία μεταβλητή y' να ακολουθεί την κανονική κατανομή.

2. Κατανομή Poisson, $\boldsymbol{Y \sim P(\mu)}$

Η συνάρτηση πιθανότητας της κατανομής Poisson είναι η

$$f(y;\theta) = \frac{e^{-\theta}\theta^{y}}{y!}\ , y = 0,1,2, \ldots,$$

η οποία μπορεί να γραφεί ως:

$$f(y;\theta) = \exp(y\log\theta - \theta - \log y!).$$

Επειδή $a(y) = y$ λέμε ότι είναι στην κανονική της μορφή και ότι η φυσική παράμετρος είναι $b(\theta) = ln\theta$. Η κατανομή Poisson χρησιμοποιείται για την μοντελοποίηση μεταβλητών που μετράνε αριθμό πραγματοποίησης κάποιου γεγονότος με την προϋπόθεση, ότι η πιθανότητα πραγματοποίησης του γεγονότος είναι μικρή και ότι τα γεγονότα συμβαίνουν ανεξάρτητα το ένα από το άλλο. Παραδείγματα στα οποία χρησιμοποιούμε την κατανομή Poisson είναι όταν θέλουμε να μετρήσουμε τον αριθμό των ατόμων που επισκέπτονται ένα κατάστημα ή τον αριθμό των ορθογραφικών λαθών σε μια σελίδα ή τον αριθμό των ασθενειών που μπορεί να καταγραφούν σε ένα άτομο μια συγκεκριμένη χρονική περίοδο.

3. Διωνυμική κατανομή, $\boldsymbol{Y \sim binomial(n, \pi)}$

Ας υποθέσουμε ότι έχουμε μια σειρά τυχαίων γεγονότων, τα οποία θα ονομάσουμε δοκιμές τα οποία έχουν 2 πιθανά αποτελέσματα , επιτυχία ή αποτυχία. Η τυχαία μεταβλητή Y θα αναπαριστά τον αριθμό των επιτυχιών σε n ανεξάρτητες δοκιμές για τα οποία η πιθανότητα επιτυχίας θα είναι ίδια και ίση με π σε κάθε δοκιμή. Τότε η τυχαία μεταβλητή Y ακολουθεί την διωνυμική κατανομή με συνάρτηση πιθανότητας

$$f(y;\pi) = \binom{n}{y}\pi^{y}(1-\pi)^{\,n-y}, y = 0,1,2, \ldots, n.$$

Η παράμετρος που μας ενδιαφέρει είναι η πιθανότητα επιτυχίας π θεωρώντας γνωστό το n. Η συνάρτηση πιθανότητας γράφεται ως

$$f(y;\pi) = \exp\,(y\log\pi - y\log(1-\pi) + nlog(1-\pi) + \log\binom{n}{y},$$

η οποία είναι της μορφής (2.6) με $b(\pi) = \log(\pi) - \log(1-\pi) = \log\left(\frac{\pi}{1-\pi}\right)$. Η διωνυμική κατανομή χρησιμοποιείται συνήθως για παρατηρήσεις που έχουν μόνο 2 πιθανές εκβάσεις. Για παράδειγμα ο αριθμός των υποψηφίων που περνάνε ένα τεστ, ή ο αριθμός των επιζώντων από κάποια ασθένεια.

2.4.2 Ιδιότητες των Κατανομών που ανήκουν στην Ε.Ο.Κ

Από τον ορισμό της συνάρτησης πυκνότητας πιθανότητας ισχύει ότι το ολοκλήρωμα αυτής με όρια της τιμές για τις οποίες ορίζεται η τυχαία μεταβλητή y είναι ίσο με 1. Δηλαδή

$$\int f(y;\theta)\, dy = 1. \tag{2.7}$$

Προφανώς, αν η τυχαία μεταβλητή είναι διακριτή αντικαθιστούμε το ολοκλήρωμα με άθροισμα. Παραγωγίζοντας και τις 2 πλευρές της (2.7) ως προς θ λαμβάνουμε:

$$\frac{d}{d\theta}\int f(y;\theta)\, dy = \frac{d}{d\theta}1 = 0. \tag{2.8}$$

Αλλάζοντας την σειρά της παραγώγισης και της ολοκλήρωσης η (2.8) γίνεται

$$\int \frac{df(y;\theta)}{d\theta} dy = 0. \tag{2.9}$$

Με τον ίδιο τρόπο αν παραγωγίσουμε την (2.8) δεύτερη φορά ως προς θ και αλλάξουμε την σειρά της παραγώγισης και της ολοκλήρωσης όπως κάναμε στην (2.9) λαμβάνουμε:

$$\int \frac{d^2 f(y;\theta)}{d\theta^2} dy = 0. \tag{2.10}$$

Τα παραπάνω αποτελέσματα θα τα χρησιμοποιήσουμε για συναρτήσεις που ανήκουν στην εκθετική οικογένεια κατανομών. Παραγωγίζοντας την

$$f(y;\theta) = exp\{a(Y)b(\theta) + c(\theta) + d(y)\},$$

έχουμε

$$\frac{df(y;\theta)}{d\theta} = [a(Y)b'(\theta) + c'(\theta)]f(y;\theta). \tag{2.11}$$

Ολοκληρώνοντας το 2ο μέλος της (2.11) έχουμε

$$\int [a(Y)b'(\theta) + c'(\theta)]f(y;\theta)dy = 0. \tag{2.12}$$

Λαμβάνοντας υπόψη ότι

$$\int a(y)\, f(y;\theta)dy = E[a(Y)]$$

και

$$\int c'(\theta)\, f(y;\theta)dy = c'(\theta),$$

η (2.12) γίνεται

$$b'(\theta)E[a(Y)] + c'(\theta) = 0. \tag{2.13}$$

Λύνοντας ως προς $E[a(y)]$ την (2.13) έχουμε ότι

$$E[a(Y)] = -\frac{c'(\theta)}{b'(\theta)}. \tag{2.14}$$

Κατά τον ίδιο τρόπο θα βρούμε και την διακύμανση $Var[a(y)]$ μέσω της δεύτερης παραγώγου. Παραγωγίζοντας την (2.11) δεύτερη φορά έχουμε,

$$\frac{d^2 f(y;\theta)}{d\theta^2} = [a(Y)b''(\theta) + c''(\theta)]f(y;\theta) + [a(Y)b'(\theta) + c'(\theta)]^2 f(y;\theta)dy. \qquad (2.15)$$

Ο δεύτερος όρος της (2.15) μπορεί να γραφεί μέσω της (2.14) με την εξής μορφή:

$$b'(\theta)^2\{a(y) - E[a(Y)]\}^2 f(y;\theta).$$

Χρησιμοποιώντας την σχέση (2.10) και το ορισμό της διακύμανσης της $a(Y)$ που δίνεται από

$$\int\{a(y) - E[a(Y)]\}^2 f(y;\theta)dy = Var[a(Y)],$$

έχουμε,

$$\int \frac{d^2 f(y;\theta)}{d\theta^2} dy = b''(\theta)E[a(Y)] + c''(\theta) + b'(\theta)^2 Var[a(Y)] = 0. \qquad (2.16)$$

Λύνοντας ως προς $Var[a(Y)]$ και αντικαθιστώντας $E[a(Y)] = -\frac{c'(\theta)}{b'(\theta)}$ η (2.16) γίνεται

$$Var[a(Y)] = [b''(\theta)c'(\theta) - c''(\theta)b'(\theta)]/b'(\theta)^3. \qquad (2.17)$$

Οι εξισώσεις (2.14) και (2.17) μπορούν να χρησιμοποιηθούν για να πάρουμε την μέση τιμή και την διακύμανση της κανονικής, διωνυμικής και Poisson κατανομής τις οποίες εξετάσαμε παραπάνω. Η ποσότητα $Var[a(Y)] = Var(Y)$, αφού $a(Y) = Y$, είναι μια συνάρτηση μεταξύ της μεταβλητής θ (άρα του μέσου της κατανομής), η οποία συμβολίζεται $V(\theta)$ και μιας συνάρτησης $\alpha(\varphi)$. Η συνάρτηση $V(\theta)$ ονομάζεται συνάρτηση διακύμανσης (variance function). Η συνάρτηση $\alpha(\varphi)$ είναι συνήθως της μορφής:

$$a(\varphi) = \varphi,$$

όπου φ συμβολίζεται και ως σ^2 και ονομάζεται παράμετρος διασποράς (dispersion parameter). Επομένως $Var(y) = V(\theta)a(\varphi)$.

Στον παρακάτω πίνακα βλέπουμε τις τιμές του $V(\theta), a(\varphi)$ για κάποιες γνωστές κατανομές.

Κατανομή	$V(\theta)$	$a(\varphi)$
Κανονική	*1*	σ^2
Poisson	θ	*1*
Διωνυμική	$\theta(1-\theta)$	*1*
Γάμμα	θ^2	$1/a$

Πίνακας 2.1: Τιμές της συνάρτησης διακύμανσης και της παραμέτρου διασποράς για κάποιες γνωστές κατανομές.

Οι παραπάνω τιμές μπορούν να αλλάξουν. Αυτό εξαρτάται από τον τρόπο παραμετροποίησης κάθε κατανομής. Στην συνέχεια θα βρούμε αντίστοιχες εκφράσεις για την μέση τιμή και την διακύμανση χρησιμοποιώντας την λογαριθμική συνάρτηση. Η λογαριθμική συνάρτηση της

$$f(y;\theta) = exp\{a(y)b(\theta) + c(\theta) + d(y)\},$$

είναι $$l(y;\theta) = \{a(y)b(\theta) + c(\theta) + d(y)\}. \tag{2.18}$$

Η παράγωγος της (2.18) ως προς θ είναι

$$U(y;\theta) = \frac{dl(y;\theta)}{d\theta} = a(y)b'(\theta) + c'(\theta).$$

Η συνάρτηση U είναι η συνάρτηση score που είδαμε σε προηγούμενο κεφάλαιο. Αφού εξαρτάται από το y μπορούμε να θεωρήσουμε ότι πρόκειται για μια τυχαία μεταβλητή

$$U = a(Y)b'(\theta) + c'(\theta). \tag{2.19}$$

Η αναμενόμενη τιμή της (2.19) θα είναι

$$E[U] = E[a(Y)]b'(\theta) + c'(\theta) \tag{2.20}$$

και αντικαθιστώντας $E[a(y)] = -\frac{c'(\theta)}{b'(\theta)}$ η (2.20) γίνεται

$$E[U] = -\frac{c'(\theta)}{b'(\theta)}b'(\theta) + c'(\theta) = 0.$$

Η διακύμανση της U συμβολίζεται ως J και ισούται με

$$J = var(U) = b'(\theta)^2 Var[a(Y)]\,. \tag{2.21}$$

Αντικαθιστώντας $Var[a(Y)] = [b''(\theta)c'(\theta) - c''(\theta)b'(\theta)]/b'(\theta)^3$ η (2.21) γίνεται

$$J = var(U) = (b''(\theta)c'(\theta)/b'(\theta)) - c''(\theta).$$

2.4.3 Συνάρτηση Σύνδεσης

Στα γενικευμένα γραμμικά μοντέλα ορίζουμε μεταξύ του μέσου των εξαρτημένων μεταβλητών y_i, $E(Y_i) = \mu_i$ και των ανεξάρτητων μεταβλητών $x_1, x_2, \dots x_n$ μια συνάρτηση σύνδεσης έτσι ώστε

$$g(E(Y_i)) = g(\mu_i) = x_i^T\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + .. \beta_n x_n = \eta_i\ ,$$

όπου

- $g(\cdot)$ είναι μια μονότονη, παραγωγίσιμη συνάρτηση η οποία λέγεται συνάρτηση σύνδεσης,
- x_i είναι ένα $p \times 1$ διάνυσμα επεξηγηματικών μεταβλητών,
$$x_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}, \quad \text{άρα } x_i^T = [x_{i1} \quad \vdots \quad x_{ip}]$$
- και β είναι ένα $p \times 1$ διάνυσμα παραμέτρων
$$\beta_i = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}.$$

Υπάρχουν πολλές επιλογές για την συνάρτηση σύνδεσης. Κάποιες από αυτές είναι οι εξής:

- $Y_i \sim$Κανονική κατανομή τότε $\eta_i = g(\mu_i) = \mu_i$ (identity link),
- $Y_i \sim$Bernoulli κατανομή ή Διωνυμική τότε $\eta_i = g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$ (logit link),
- $Y_i \sim$Poisson τότε $\eta_i = g(\mu_i) = \log \mu_i$ (log link),
- $Y_i \sim$Γάμμα κατανομή ή εκθετική τοτε $\eta_i = g(\mu_i) = \frac{1}{\mu_i}$ (reciprocal link).

Για τις παραπάνω συναρτήσεις σύνδεσης ισχύει ότι $\eta_i = g(\mu_i) = \theta_i$ και εξαιτίας αυτού ονομάζονται και κανονικές συναρτήσεις σύνδεσης. Φυσικά υπάρχουν και άλλες συναρτήσεις που χρησιμοποιούνται στα γενικευμένα γραμμικά μοντέλα οι οποίες δεν είναι κανονικές όπως:

1. **Probit link**

$$\eta_i = g(\mu_i) = \Phi^{-1}[\mu_i],$$

όπου Φ είναι η συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής.

2. **Complementary log-log link**

$$\eta_i = g(\mu_i) = log\{\log[1-\mu_i]\}.$$

Οι συναρτήσεις σύνδεσης logit, probit και complementary log-log link χρησιμοποιούνται για διωνυμικά ή δίτιμα δεδομένα. Προτιμάται περισσότερο ο μετασχηματισμός logit έναντι του probit, γιατί έχει ευκολότερη και άμεση ερμηνεία αφού πρόκειται για τον λογάριθμο των συμπληρωματικών ή σχετικών πιθανοτήτων (odds). Το μοντέλο complementary log-log συμπίπτει με τα μοντέλα probit και logit για τιμές που βρίσκονται κοντά στο $p = 0.5$ ενώ διαφέρει για τιμές κοντά στο 0 και στο 1.

2.5 Εκτίμηση Παραμέτρων

Σε αυτήν την ενότητα θα αναπτύξουμε τις μεθόδους με τις οποίες θα μπορέσουμε να εκτιμήσουμε τις άγνωστες παραμέτρους σε ένα γενικευμένο γραμμικό μοντέλο. Οι μέθοδοι βασίζονται στην μέθοδο μέγιστης πιθανοφάνειας. Παρόλο που σε ειδικές περιπτώσεις μπορούμε να βρούμε κλειστούς τύπους για τον υπολογισμό των παραμέτρων, τις περισσότερες φορές χρειαζόμαστε αριθμητικές μεθόδους οι οποίες συνήθως είναι επαναληπτικές και βασίζονται στον αλγόριθμο των Newton-Raphson.

2.5.1 Εκτίμηση μέσω της Μεθόδου Μέγιστης Πιθανοφάνειας

Υποθέτουμε ότι έχουμε $Y_1, Y_2, \dots, Y_n$ τυχαίες μεταβλητές οι οποίες ικανοποιούν τις ιδιότητες ενός γενικευμένου γραμμικού μοντέλου. Σκοπός μας είναι να εκτιμήσουμε τις παραμέτρους β_i για τις οποίες ισχύει $E(Y_i) = \mu_i$ και $g(\mu_i) = \boldsymbol{x_i^T \beta}$. Για κάθε Y_i η λογαριθμική συνάρτηση πιθανοφάνειας είναι

$$l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i),$$

με

- $E(Y_i) = \mu_i = -\frac{c'(\theta_i)}{b'(\theta_i)}$,
- $Var(Y_i) = \frac{[b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)]}{[b'(\theta_i)]^3}$,
- $g(\mu_i) = \boldsymbol{x_i^T \beta}$.

Η λογαριθμική συνάρτηση πιθανοφάνειας για όλα τα Y_i είναι

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i). \tag{2.22}$$

Παραγωγίζοντας ως προς β_j την (2.22) θα πάρουμε τους εκτιμητές μέγιστης πιθανοφάνειας των β_j, χρησιμοποιώντας τον κανόνα της αλυσίδας.

$$\frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}\right]. \tag{2.23}$$

Θα αναλύσουμε κάθε όρο της (2.23) ξεχωριστά.

- $\frac{\partial l_i}{\partial \theta_i} = y_i b'(\theta_i) + c'^{(\theta_i)} = b'(\theta_i)(y_i - \mu_i)$,
- $\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}}$,

  με $\frac{\partial \mu_i}{\partial \theta_i} = \frac{-c''(\theta_i)}{b'(\theta_i)} + \frac{c'(\theta_i)}{[b'(\theta_i)]^2} = b'(\theta_i)Var(Y_i)$

- $\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$.

Από τα παραπάνω η (2.23) γίνεται τελικά:

$$U_j = \sum_{i=1}^{n} \left[\frac{y_i - \mu_i}{Var(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right]. \tag{2.24}$$

Ο πίνακας διακυμάνσεων – συνδιακυμάνσεων των U_j είναι

$$J_{ik} = E[U_j U_k]. \tag{2.25}$$

Ο πίνακας J_{ik} λέγεται και πίνακας πληροφορίας J. Αναλύοντας την (2.25) έχουμε:

$$J_{ik} = E\left\{\sum_{i=1}^{n}\left[\frac{Y_i - \mu_i}{Var[Y_i]} x_{ij}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right]\sum_{l=1}^{n}\left[\frac{Y_l - \mu_l}{Var(Y_i)} x_{lk}\left(\frac{\partial \mu_l}{\partial \eta_l}\right)\right]\right\}$$

$$= \sum_{l=1}^{n} \frac{E[(Y_i - \mu_i)^2] x_{ij} x_{ik}}{[Var(Y_i)]^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2, \tag{2.26}$$

με $E[(Y_i - \mu_i)(Y_l - \mu_l)] = 0$ για $i \neq l$, αφού τα Y_i είναι ανεξάρτητα μεταξύ τους. Χρησιμοποιώντας $E[(Y_i - \mu_i)^2] = Var[Y_i]$ η (2.26) γίνεται:

$$J_{ik} = \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 .$$

Η επαναληπτική μέθοδος Newton-Raphson ικανοποιεί την παρακάτω εξίσωση.

$$b^{(m)} = b^{(m-1)} + [J^{(m-1)}]^{-1} U^{(m-1)}, \tag{2.27}$$

όπου

- $b^{(m)}$ είναι το διάνυσμα των εκτιμήσεων των παραμέτρων $\beta_1, \beta_2, .., \beta_p$ στην m επανάληψη.
- $[J^{(m-1)}]^{-1}$ είναι ο αντίστροφος πίνακας του πίνακα πληροφορίας που δίνεται στην εξίσωση (2.26) για την $m-1$ επανάληψη.
- $U^{(m-1)}$ είναι το διάνυσμα των στοιχείων που δίνεται στην εξίσωση (2.24) για την $m-1$ επανάληψη.

Πολλαπλασιάζοντας και τα δύο μέλη της (2.27) με $J^{(m-1)}$ έχουμε:

$$J^{(m-1)}\, b^{(m)} = J^{(m-1)} b^{(m-1)} + U^{(m-1)} . \tag{2.28}$$

Η J μπορεί να γραφεί και ως $J = X^T W X$, όπου W είναι ο $N \times N$ διαγώνιος πίνακας με στοιχεία

$$w_{ii} = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 . \tag{2.29}$$

Αναλύοντας το δεξί μέρος της (2.28) παίρνουμε

$$\sum_{k=1}^{p} \sum_{i=1}^{n} \frac{x_{ij}x_{ik}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 b_k^{(m-1)} + \sum_{i=1}^{n} \frac{(Y_i - \mu_i)x_{ij}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right),$$

για την επανάληψη $(m-1)$. Θέτοντας

$$z = z_i = \sum_{k=1}^{p} x_{ik} b_k^{(m-1)} + (Y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i}\right), \tag{2.30}$$

η (2.28) γράφεται ως

$$X^T W X b^{(m)} = X^T W z. \tag{2.31}$$

Η (2.31) λύνεται επαναληπτικά καθώς τα W, z εξαρτώνται από το b. Άρα για να βρούμε τους εκτιμητές μέγιστης πιθανοφάνειας στα γενικευμένα γραμμικά μοντέλα, χρησιμοποιούμε μια επαναληπτική μέθοδο ελαχίστων τετραγώνων με βάρη (iterative weighted least squares procedure) (Charnes et al. 1976). Ο παραπάνω αλγόριθμος τερματίζει για το διάνυσμα $b^{(m)}$ για το οποίο ισχύει:

$$\left| \frac{b^{(m)} - b^{(m-1)}}{b^{(m)}} \right| < \delta,$$

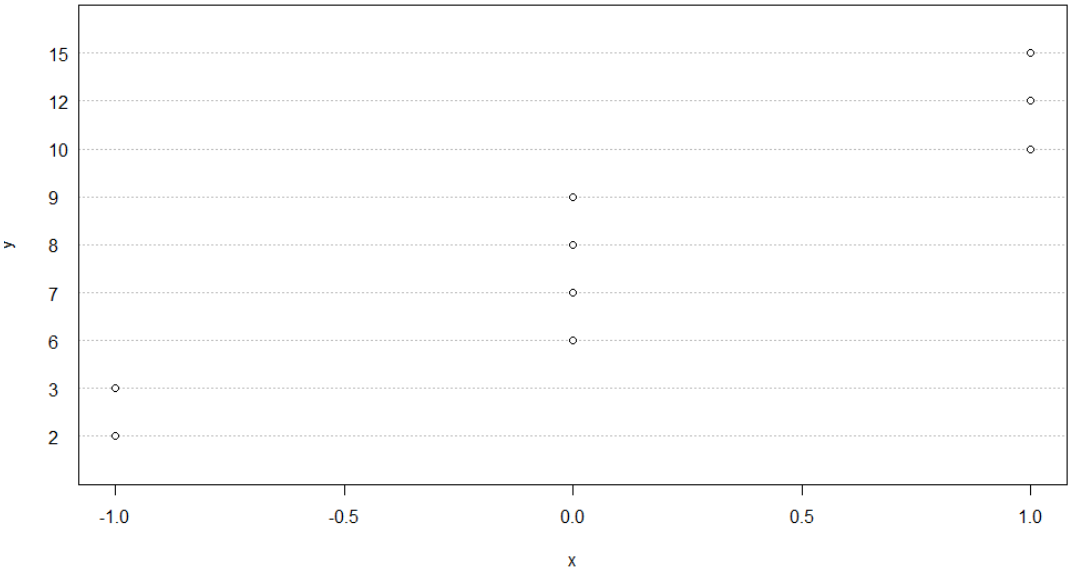
όπου δ μια πολύ μικρή ποσότητα της τάξεως του 10^{-6}.

Τα περισσότερα στατιστικά πακέτα, τα οποία περιλαμβάνουν μεθόδους εκτίμησης πάνω στα γενικευμένα γραμμικά μοντέλα χρησιμοποιούν επαναληπτικούς αλγορίθμους οι οποίοι βασίζονται στην εξίσωση (2.31). Στο πρώτο βήμα χρησιμοποιούμε μια αρχική τιμή για το $b^{(0)}$ για να βρεθούν τα W, z. Στην συνέχεια μέσω της εξίσωσης (2.31) και επαναλαμβάνοντας την ίδια διαδικασία βρίσκουμε προσεγγιστικές τιμές για τα $b^{(1)}, b^{(2)}, \dots, b^{(m)}$ Όταν η διαφορά μεταξύ δύο διαδοχικών τιμών $b^{(m)}, b^{(m-1)}$ είναι αρκετά μικρή τότε η τιμή $b^{(m)}$χρησιμοποιείται ως τελικός εκτιμητής. Στην συνέχεια παρουσιάζεται ένα παράδειγμα από το βιβλίο του Dobson, όπου αναλύεται η παραπάνω επαναληπτική διαδικασία για δεδομένα που προέρχονται από την κατανομή Poisson.

Παράδειγμα 2.1. Έστω ότι έχουμε Y_i τυχαίες μεταβλητές οι οποίες ακολουθούν την κατανομή Poisson και την επεξηγηματική μεταβλητή x_i. Στον παρακάτω πίνακα φαίνονται τα δεδομένα για τις δύο μεταβλητές.

Y_i	X_i
2	-1
3	-1
6	0
7	0
8	0
9	0
10	1
12	1
15	1

Πίνακας 2.2: Δεδομένα της παλινδρόμησης Poisson



Γράφημα 2.2 : Διάγραμμα δεδομένων Παραδείγματος 2.1.

Για την Poisson κατανομή ισχύει ότι η μέση τιμή είναι ίση με την διακύμανση. Δηλαδή $E(Y_i) = Var(Y_i)$. Αρχικά θεωρούμε το παρακάτω γραμμικό μοντέλο.

$$E(Y_i) = \mu_i = \beta_0 + \beta_1 x_i = x_i^T \beta,$$

όπου

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \text{ και } x_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix},$$

για $i = 1,2, \dots, n$. Άρα θεωρούμε ότι η συνάρτηση σύνδεσης $g(\mu_i) = \mu_i = x_i^T \beta = \eta_i$ είναι η ταυτοτική συνάρτηση. Τότε

$$\frac{\partial \mu_i}{\partial \eta_i} = 1$$

και η εξίσωση (2.29) γίνεται $w_{ii} = \frac{1}{Var(Y_i)} = \frac{1}{\beta_0 + \beta_1 x_i}$. Αντικαθιστώντας στην (2.30) όπου

$$z = z_i = \sum_{k=1}^{p} x_{ik} b_k^{(m-1)} + (Y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right),$$

την εκτίμηση $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$ λαμβάνουμε $z_i = \beta_0 + \beta_1 x_i + (Y_i - \beta_0 - \beta_1 x_i) = Y_i$. Επίσης

$$J = X^T W X = \begin{bmatrix} \sum_{i=1}^{n} \frac{1}{\beta_0 + \beta_1 x_i} & \sum_{i=1}^{n} \frac{x_i}{\beta_0 + \beta_1 x_i} \\ \sum_{i=1}^{n} \frac{x_i}{\beta_0 + \beta_1 x_i} & \sum_{i=1}^{n} \frac{{x_i}^2}{\beta_0 + \beta_1 x_i} \end{bmatrix}$$

και

$$X^T W z = \begin{bmatrix} \sum_{i=1}^{n} \frac{Y_i}{\beta_0 + \beta_1 x_i} \\ \sum_{i=1}^{n} \frac{Y_i x_i}{\beta_0 + \beta_1 x_i} \end{bmatrix}.$$

Στην συνέχεια βρίσκουμε τους εκτιμητές μέγιστης πιθανοφάνειας μέσω της επαναληπτικής σχέσης

$$(X^T W X)^{m-1} \beta^m = X^T W z^{m-1}.$$

Για τα δεδομένα μας έχουμε $n = 9$ και

$$y = z = \begin{bmatrix} 2 \\ 3 \\ \vdots \\ 15 \end{bmatrix} \text{ και } X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_9 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & -1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}.$$

Θέτοντας σαν αρχικές τιμές $\beta_0^{(1)} = 7$ και $\beta_1^{(1)} = 5$ βρίσκουμε την δεύτερη προσέγγιση για τους δύο εκτιμητές.

$$(X^TWX)^{(1)} = \begin{bmatrix} 1.82429 & -0.75 \\ -0.75 & 1.25 \end{bmatrix} \text{ και } (X^TWz)^{(1)} = \begin{bmatrix} 9.869048 \\ 0.583333 \end{bmatrix}.$$

Άρα
$$\beta_0^{(2)} = \left[(X^TWX)^{(1)}\right]^{-1}(X^TWz)^{(1)}$$

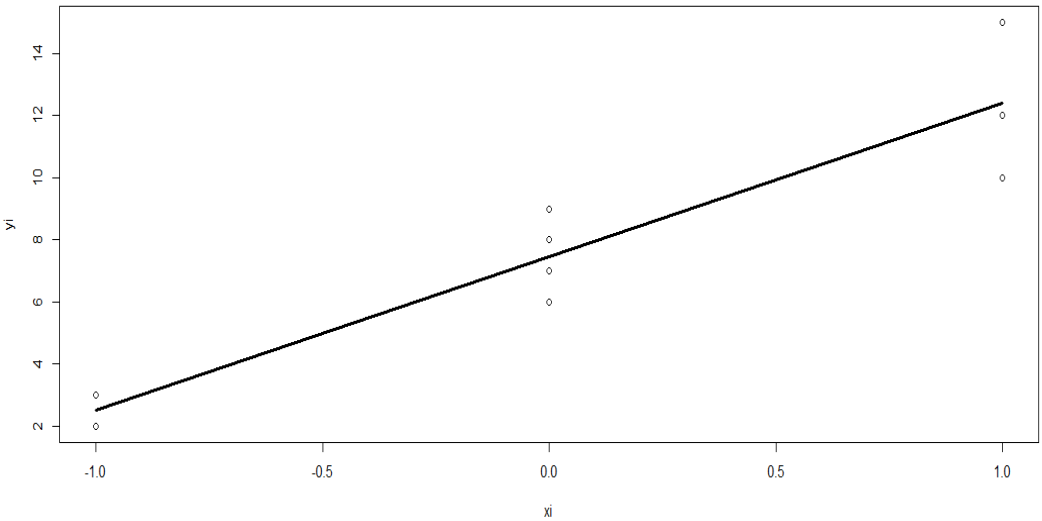
$$= \begin{bmatrix} 0.729167 & 0.4375 \\ 0.4375 & 1.0625 \end{bmatrix} \begin{bmatrix} 9.869048 \\ 0.583333 \end{bmatrix} = \begin{bmatrix} 7.4514 \\ 4.9375 \end{bmatrix}.$$

Η επαναληπτική μέθοδος συνεχίζεται μέχρι η διαφορά δύο διαδοχικών εκτιμήσεων να είναι πολύ μικρή. Τα αποτελέσματα σε κάθε επανάληψη φαίνονται στον παρακάτω πίνακα.

$\boldsymbol{m}$	$\boldsymbol{\beta_0^m}$	$\boldsymbol{\beta_1^m}$
1	*7*	*5*
2	*7.4514*	*4.9353*
3	*7.45163*	*4.93531*
4	*7.45163*	*4.9353*

Πίνακας 2.2 Τιμές των εκτιμητών μέγιστης πιθανοφάνειας σε κάθε επανάληψη.

Οι εκτιμητές μέγιστης πιθανοφάνειας είναι $\hat{\beta}_0 = 7.45163$ και $\hat{\beta}_1 = 4.9353$, άρα $E(Y_i) = \mu_i = 7{,}45163 + 4.9353x_i$.



Γράφημα 2.6.2 : Προσαρμογή μοντέλου Poisson στα δεδομένα του Παραδείγματος 2.1.

2.5.2 Quasi – Πιθανοφάνεια

Όπως είδαμε προηγουμένως, η εκτίμηση των παραμέτρων $\beta_0, \beta_1, \dots \beta_n$ του γενικευμένου γραμμικού μοντέλου γίνεται μέσω της μεθόδου μέγιστης πιθανοφάνειας. Η μέθοδος της μέγιστης πιθανοφάνειας προϋποθέτει να γνωρίζουμε εκ των προτέρων την κατανομή της μεταβλητής απόκρισης Y_i. Στην στατιστική αυτό δεν είναι πάντα εφικτό για αυτό τον λόγο προτάθηκε από τον Wedderburn (1974), η μέθοδος της quasi-πιθανοφάνειας (quasi – likelihood). Για να εφαρμοστεί η μέθοδος της quasi-πιθανοφάνειας δεν χρειάζεται να γνωρίζουμε την κατανομή της μεταβλητής απόκρισης Y_i παρά μόνο τις δύο πρώτες ροπές

αυτής, δηλαδή την ροπή πρώτης τάξης $E(Y_i)$ και την ροπή δεύτερης τάξης $E(Y_i^2)$. Γνωρίζοντας τις 2 ροπές αυτομάτως γνωρίζουμε και την διακύμανση αφού $Var(Y_i) = E[(Y_i - E(Y_i))^2]$. Επίσης, θεωρούμε ότι οι παρατηρήσεις είναι ανεξάρτητες μεταξύ τους, Τότε οι συναρτήσεις score θα είναι

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{y_i - \mu_i}{Var(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right].$$

2.6 Επιλογή Μοντέλου

Τα κυριότερα εργαλεία της στατιστικής συμπερασματολογίας για την εξαγωγή συμπερασμάτων είναι δύο: τα διαστήματα εμπιστοσύνης και τα τεστ υποθέσεων. Τα διαστήματα εμπιστοσύνης χρησιμοποιούνται ως ένα μέτρο για το πόσο κοντά είναι ο εκτιμητής μας στην πραγματική τιμή της μεταβλητής. Τα τεστ υποθέσεων χρησιμοποιούνται ως μέτρο σύγκρισης μεταξύ δύο μοντέλων έτσι ώστε να καταλήξουμε σε αυτό το οποίο προσομοιώνει καλύτερα τα δεδομένα μας. Στα γενικευμένα γραμμικά μοντέλα, τα δύο μοντέλα θα πρέπει να ακολουθούν την ίδια κατανομή και να έχουν την ίδια συνάρτηση σύνδεσης, αλλά το γραμμικό κομμάτι του ενός θα πρέπει να έχει παραπάνω παραμέτρους από το αντίστοιχο γραμμικό κομμάτι του άλλου. Το πιο απλό μοντέλο με τις λιγότερες παραμέτρους θα αντιστοιχεί στην μηδενική υπόθεση H_0 και θα πρέπει να είναι μια ειδική περίπτωση του πιο γενικού μοντέλου, ενώ το πιο γενικό μοντέλο θα αντιστοιχεί στην εναλλακτική υπόθεση H_1. Αν το πιο απλό μοντέλο προσομοιώνει τα δεδομένα κατά τον ίδιο τρόπο με το πιο σύνθετο τότε προτιμούμε το απλό και απορρίπτουμε την H_1 . Αν το πιο σύνθετο μοντέλο προσομοιώνει τα δεδομένα στατιστικώς σημαντικότερα από το πιο απλό τότε προτιμούμε το σύνθετο και απορρίπτουμε την H_0. Ως κριτήριο για να καταλήξουμε στην απόρριψη ή μη κάποιας υπόθεσης χρησιμοποιούμε την μέγιστη τιμή της συνάρτησης πιθανοφάνειας ή της λογαριθμικής συνάρτησης πιθανοφάνειας, την μικρότερη τιμή της μεθόδου ελαχίστων τετραγώνων, την συνάρτηση ελέγχου Wald και άλλα.

2.6.1 Έλεγχος Wald

Από την θεωρία της μεθόδου μέγιστης πιθανοφάνειας, γνωρίζουμε ότι ο εκτιμητής μέγιστης πιθανοφάνειας b για την παράμετρο β ακολουθεί ασυμπτωτικά την πολυμεταβλητή κανονική κατανομή, δηλαδή:

$$b \sim N_p\left(\beta, \hat{V}(b)\right), p = n + 1,$$

όπου

- $\hat{V}(b) = J^{-1}(b)$,
- $J^{-1}(b)$, είναι ο αντίστροφος του πίνακα πληροφορίας $J(b)$ με στοιχεία

$$J_{ik} = \sum_{l=1}^{n} \frac{E[(Y_i - \mu_i)^2] x_{ij} x_{ik}}{[Var(Y_i)]^2} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2.$$

Ο παρατηρούμενος πίνακας πληροφορίας για τα γενικευμένα γραμμικά μοντέλα είναι

$$J(b) = X^T W X.$$

Συνεπώς, οι εκτιμητές μέγιστης πιθανοφάνειας b ακολουθούν τελικά την πολυμεταβλητή κατανομή με τις εξής παραμέτρους, $b \sim N_p(\beta, (X^T W X)^{-1}), p = n + 1$
Ως κριτήριο ελέγχου της μηδενικής υπόθεσης $H_0: \beta_j = 0$ έναντι της εναλλακτικής $H_1: \beta_j \neq 0$ χρησιμοποιούμε την συνάρτηση ελέγχου Wald που ακολουθεί προσεγγιστικά την τυπική κανονική κατανομή

$$Z = \frac{b_j - \beta_j}{(J^{-1}(b)_{ii})^{1/2}} \sim N(0,1).$$

Στην περίπτωση, που δεν απορρίψουμε την μηδενική υπόθεση, δηλαδή αν θεωρήσουμε ότι $\beta_j = 0$, τότε λέμε ότι η μεταβλητή X_j δεν συμβάλλει σημαντικά στο μοντέλο και μπορούμε να την παραβλέψουμε.

2.6.2 Επιλογή Μοντέλου και Απόκλιση

Ένας τρόπος για να επιλέξουμε το καταλληλότερο μοντέλο είναι να το συγκρίνουμε με ένα πιο γενικό μοντέλο το οποίο περιέχει όλες τις δυνατές παραμέτρους. Το μοντέλο αυτό ονομάζεται πλήρες ή κορεσμένο μοντέλο. Τα δύο μοντέλα θα πρέπει να ακολουθούν την ίδια κατανομή και να έχουν την ίδια συνάρτηση σύνδεσης. Ας υποθέσουμε ότι m είναι ο αριθμός παραμέτρων του κορεσμένου μοντέλου. Τότε θέτουμε ως β_{max} το διάνυσμα των παραμέτρων της παλινδρόμησης και b_{max} το διάνυσμα των εκτιμήσεων αυτών. Η μέγιστη τιμή της συνάρτησης πιθανοφάνειας για το κορεσμένο μοντέλο στο σημείο b_{max} συμβολίζεται $L(b_{max}; y)$. Η αντίστοιχη μέγιστη τιμή για το μοντέλο που μας ενδιαφέρει θα συμβολίζεται $L(b; y)$. Τότε ορίζουμε ως λόγο των πιθανοφανειών την παρακάτω ποσότητα

$$\lambda = \frac{L(b_{max}; y)}{L(b; y)}. \tag{2.32}$$

Η (2.32) χρησιμοποιείται ως ένα κριτήριο για την αξιολόγηση του μοντέλου μας. Αν αντί για την συνάρτηση πιθανοφάνειας χρησιμοποιήσουμε την λογαριθμική συνάρτηση πιθανοφάνειας, τότε η (2.32) εκφράζεται ως εξής

$$\log \lambda = l(b_{max}; y) - l(b; y). \tag{2.33}$$

Προτιμάμε να χρησιμοποιούμε την (2.33) για τους ελέγχους μας έναντι της (2.32). Μεγάλες τιμές του $\log \lambda$ υποδεικνύουν ότι το μοντέλο που μας ενδιαφέρει είναι μια κακή επιλογή μοντέλου για να περιγράψουμε τα δεδομένα μας σε σχέση με το κορεσμένο. Η απόκλιση του κορεσμένου μοντέλου από το μοντέλο που εξετάζουμε ορίζεται ως

$$\begin{aligned} D &= 2[l(b_{max}; y) - l(b; y)] \\ &= 2[l(b_{max}; y) - l(\beta_{max}; y)] - 2[l(b; y) - l(\beta; y)] + 2[l(\beta_{max}; y) - l(\beta; y)]. \end{aligned}$$

Η παραπάνω συνάρτηση ελέγχου προτάθηκε από τους Nelder και Wedderburn (1972). Αν η D λαμβάνει μεγάλες τιμές, τότε το απλούστερο μοντέλο δεν εκφράζει επαρκώς τα δεδομένα

συγκριτικά με το πλήρες ή κορεσμένο. Αντίθετα μικρές τιμές της D δηλώνουν ότι το υποψήφιο μοντέλο μας έχει ικανοποιητική προσαρμογή στα δεδομένα μας.
Τέλος, από την θεωρία πιθανοτήτων γνωρίζουμε ότι η συνάρτηση ελέγχου D ακολουθεί προσεγγιστικά την X^2 κατανομή με d βαθμούς ελευθερίας, όπου d η διαφορά του αριθμού των παραμέτρων μεταξύ των δύο μοντέλων.

Απόκλιση στο διωνυμικό μοντέλο
Αν οι μεταβλητές απόκρισης $Y_1, Y_2, .., Y_n$ είναι ανεξάρτητες και ακολουθούν την διωνυμική κατανομή, δηλ. $Y_i \sim binomial(n_i, p_i)$. Τότε η λογαριθμική συνάρτηση πιθανοφάνειας είναι η:

$$l(\beta; y) = \sum_{i=1}^{n} \left[y_i \log \pi_i - y_i \log(1-\pi_i) + n_i \log(1-\pi_i) + \log \binom{n_i}{y_i} \right].$$

Για το κορεσμένο μοντέλο τα p_i είναι όλα διαφορετικά άρα $\beta = [\pi_1, \pi_2, .., \pi_n]^T$. Οι εκτιμητές μέγιστης πιθανοφάνειας είναι $\hat{\pi}_i = \frac{y_i}{n}$ και η μέγιστη τιμή της λογαριθμικής συνάρτησης πιθανοφάνειας είναι

$$l(b_{max}; y) = \sum_{i=1}^{n} \left[y_i \log \frac{y_i}{n} - y_i \log\left(1 - \frac{y_i}{n}\right) + n_i \log\left(1 - \frac{y_i}{n}\right) + \log \binom{n_i}{y_i} \right].$$

Για οποιοδήποτε άλλο μοντέλο με $p < n$ παραμέτρους, θέτουμε $\hat{\pi}_i$ τους εκτιμητές μέγιστης πιθανοφάνειας και ως $\hat{y}_i = n\hat{\pi}_i$ τις τιμές που ικανοποιούν το μοντέλο μας. Τότε η λογαριθμική συνάρτηση πιθανοφάνειας για αυτές τις τιμές είναι

$$l(b; y) = \sum_{i=1}^{n} \left[y_i log \frac{\hat{y}_i}{n} - y_i \log\left(1 - \frac{\hat{y}_i}{n}\right) + n_i \log\left(1 - \frac{\hat{y}_i}{n}\right) + \log \binom{n_i}{y_i} \right].$$

η απόκλιση είναι ίση με

$$\begin{aligned} D &= 2[l(b_{max}; y) - l(b; y)] \\ &= 2 \sum_{i=1}^{n} \left[y_i \log \frac{y_i}{\hat{y}_i} + (n_i - y_i) \log \left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \right]. \end{aligned}$$

Απόκλιση στο μοντέλο Poisson
Αν οι μεταβλητές απόκρισης $Y_1, Y_2, .., Y_n$ είναι ανεξάρτητες και ακολουθούν την κατανομή Poisson, δηλ. $Y_i \sim Poisson(\lambda_i)$ Τότε η λογαριθμική συνάρτηση πιθανοφάνειας είναι η

$$l(\beta; y) = \sum_{i=1}^{n} y_i \log \lambda_i - \sum_{i=1}^{n} \lambda_i - \sum_{i=1}^{n} \log y_i!.$$

Για το κορεσμένο μοντέλο τα λ_i είναι όλα διαφορετικά άρα $\beta = [\lambda_1, \lambda_2, .., \lambda_n]^T$. Οι εκτιμητές μέγιστης πιθανοφάνειας είναι $\hat{\lambda}_i = y_i$ και η μέγιστη τιμή της λογαριθμικής συνάρτησης πιθανοφάνειας είναι

$$l(b_{max}; y) = \sum_{i=1}^{n} y_i \log y_i - \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \log y_i!.$$

Ας υποθέσουμε ότι έχουμε ένα άλλο μοντέλο με $p < n$ παραμέτρους. Στην κατανομή Poisson ισχύει $E[y_i] = \lambda_i$, άρα οι εκτιμώμενες τιμές $\hat{y}_i = \hat{\lambda}_i$. Τότε σε αυτή την περίπτωση η μέγιστη τιμή της λογαριθμικής συνάρτησης πιθανοφάνειας είναι η

$$l(b; y) = \sum_{i=1}^{n} y_i \log \hat{y}_i - \sum_{i=1}^{n} \hat{y}_i - \sum_{i=1}^{n} \log y_i!.$$

Επομένως η απόκλιση είναι ίση με

$$D = 2[l(b_{max}; y) - l(b; y)]$$
$$= 2\left[\sum_{i=1}^{n} y_i \log\frac{y_i}{\hat{y}_i} - \sum_{i=1}^{n}(y_i - \hat{y}_i)\right].$$

Για τα περισσότερα κορεσμένα μοντέλα συνήθως ισχύει $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$. Για αυτές τις περιπτώσεις γράφουμε

$$D = 2\sum_{i=1}^{n} o_i \log\left(\frac{o_i}{e_i}\right),$$

όπου

- o_i είναι οι παρατηρούμενες-πραγματικές τιμές y_i,
- e_i είναι οι εκτιμώμενες τιμές $\hat{y}_i$.

Η τιμή της συνάρτησης ελέγχου D μπορεί να υπολογιστεί από τα δεδομένα και μπορεί να συγκριθεί με την τιμή της κατανομής x^2με $N - p$ βαθμούς ελευθερίας, ως ένα κριτήριο καταλληλόλητας του μοντέλου μας, καθώς ισχύει ότι προσεγγιστικά $D \sim x^2(N - p)$. Πιο συγκεκριμένα επιστρέφοντας στο Παράδειγμα 2.1, οι προσαρμοσμένες τιμές είναι:

$$\hat{y}_i = b_0 + b_1 x_i$$

x_i	y_i	$\hat{y}_i$	$y_i \log\frac{y_i}{\hat{y}_i}$
-1	*2*	*2.51633*	*-0.4593086*
-1	*3*	*2.51633*	*0.52743239*
0	*6*	*7.45163*	*-1.3000399*
0	*7*	*7.45163*	*-0.4376585*
0	*8*	*7.45163*	*0.56806991*
0	*9*	*7.45163*	*1.69912599*
1	*10*	*12.38693*	*-2.140567*
1	*12*	*12.38693*	*-0.380822*
1	*15*	*12.38693*	*2.8711247*
Total	***72***	***72***	***0.94735***

Πίνακας 2.3 : Αποτελέσματα του Παραδείγματος 2.1 για το μοντέλο Poisson

με $b_0 = 7.45163$ και $b_1 = 4.9353$. Η τιμή της συνάρτησης ελέγχου είναι $D = 2 \times (0.94735 - 0) = 1.8947$ το οποίο είναι μικρό σε σχέση με τους $N - p = 9 - 2 = 7$ βαθμούς ελευθερίας. Χρησιμοποιώντας την παρακάτω συνάρτηση στην R μπορούμε να υπολογίσουμε την πιθανότητα της τιμής D, για την οποία υποθέτουμε ότι ακολουθεί την x^2 κατανομή με 7 βαθμούς ελευθερίας.

```
> pvalue<-pchisq(1.8947,df=7,lower.tail=FALSE)
> pvalue

[1] 0.9654378
```

Η μηδενική υπόθεση μας (όπως θα δούμε και στο επόμενο κεφάλαιο) είναι ότι το μοντέλο που έχουμε επιλέξει είναι το κατάλληλο μοντέλο. Η τιμή του p-value

υποδεικνύει ότι δεν μπορούμε να απορρίψουμε την μηδενική υπόθεση και αυτό αποτελεί ένδειξη καλής προσαρμογής των δεδομένων μας στο μοντέλο.

Στον παρακάτω πίνακα βλέπουμε την μορφή που παίρνει η συνάρτηση ελέγχου D για κάποιες γνωστές κατανομές.

Κατανομή	***D***
Κανονική	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
Poisson	$2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{\hat{y}_i}\right) - (y_i - \hat{y}_i)\right\}$
Διωνυμική	$2\sum_{i=1}^{n}\left\{y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log[(n_i - y_i)/(n_i - \hat{y}_i)]\right\}$
Γάμμα	$2\sum_{i=1}^{n}\left\{-\log\left(\frac{y_i}{\hat{y}_i}\right) + (y_i - \hat{y}_i)/\hat{y}_i\right\}$
Αντίστροφη Gaussian	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(\hat{y}_i{}^2 y_i)$

Πίνακας 2.4: Τιμή της συνάρτησης D για κάποιες γνωστές κατανομές

2.6.3 Σύγκριση μοντέλων με αποκλίσεις

Έστω ότι έχουμε δύο μοντέλα τα οποία έχουν την ίδια συνάρτηση σύνδεσης και την ίδια συνάρτηση κατανομής με M_0 να είναι ένα μοντέλο με q μεταβλητές και M_1 να ένα πιο γενικό μοντέλο M_1 με $p > q$ μεταβλητές. Μας ενδιαφέρει να επιλέξουμε το μοντέλο εκείνο που δίνει καλύτερη προσαρμογή στα δεδομένα μας και ταυτόχρονα να αποφύγουμε την περιττή πολυπλοκότητα. Για το μοντέλο M_0 έχουμε την μηδενική υπόθεση

$$H_0: \beta = M_0 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix}$$

και για το μοντέλο M_1 έχουμε την εναλλακτική της μηδενικής την

$$H_1: \beta = M_1 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},$$

με $q < p < N$. Για τον έλεγχο της H_0 έναντι της H_1 θα ελέγξουμε την διαφορά των αποκλίσεων των μοντέλων M_0 και M_1. Ορίζουμε ως

$$\begin{aligned} \triangle D &= D_0 - D_1 \\ &= 2[l(b_{max}; y) - l(b_0; y)] - 2[l(b_{max}; y) - l(b_1; y)] \\ &= 2[l(b_1; y) - l(b_0; y)]. \end{aligned}$$

Εάν και τα δύο μοντέλα περιγράφουν καλά τα δεδομένα μας τότε $D_0 \sim x^2(N - q)$ και $D_1 \sim x^2(N - p)$. Τότε η $\triangle D \sim x^2(p - q)$.

Για επίπεδο σημαντικότητας α (π.χ. $\alpha = 5\%$) απορρίπτουμε την μηδενική υπόθεση H_0 αν $\triangle D = D_0 - D_1 > x^2_{1-a}(p-q)$.

2.6.4 Συνάρτηση Ελέγχου Pearson

Μία ακόμη συνάρτηση ελέγχου που χρησιμοποιούμε πέρα από της απόκλισης είναι η στατιστική συνάρτηση ελέγχου X^2 του Pearson στην γενική της μορφή η οποία είναι η παρακάτω

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{V(\hat{y}_i)},$$

όπου $V(\hat{y}_i)$ είναι η εκτιμώμενη τιμή της συνάρτησης διακύμανσης για την κατανομή που μας ενδιαφέρει. Για την κανονική κατανομή $X^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, ενώ για την Poisson $X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i}$ είναι η στατιστική συνάρτηση ελέγχου X^2.

2.6.5 Δείκτες Καλής Προσαρμογής AIC και BIC

Για την επιλογή του κατάλληλου μοντέλου, αλλά και για την σύγκριση διαφορετικών μοντέλων ως προς την σημαντικότητά τους χρησιμοποιούνται τα μέτρα καταλληλότητας. Πρόκειται για αριθμητικές ποσότητες οι οποίες χρησιμοποιούνται για την αξιολόγηση ενός μοντέλου, καθώς και για την επιλογή του βέλτιστου μοντέλου. Θα παρουσιάσουμε κάποια τέτοια κριτήρια όπως είναι τα κριτήρια AIC και BIC.
Το AIC (Akaike' s information criterion) αποτελεί ένα κριτήριο επιλογής του βέλτιστου μοντέλου το οποίο όμως θα έχει όσο το δυνατόν λιγότερες παραμέτρους. Όπως έχουμε αναφέρει το βέλτιστο μοντέλο για να προσομοιώσει τα δεδομένα μας θα είναι αυτό που θα έχει όσο το δυνατόν λιγότερες παραμέτρους, καθώς έτσι μειώνεται η πολυπλοκότητα. Ορίζεται ως

$$AIC = 2d - 2\log L,$$

όπου

- L η μέγιστη τιμή της συνάρτησης πιθανοφάνειας για το εκτιμώμενο μοντέλο,
- d ο αριθμός παραμέτρων του μοντέλου μας.

Αν συγκρίνουμε όλα τα υποψήφια μοντέλα με το κριτήριο του AIC θα προτιμήσουμε αυτό με το μικρότερο δείκτ. Όσο εισάγουμε νέες παραμέτρους στο μοντέλο μας αυξάνεται η προσαρμογή του στα δεδομένα μας, ανεξάρτητα αν είναι στατιστικώς σημαντικές ή όχι, καθώς αυξάνεται ο όρος $logL$. Ταυτόχρονα αυξάνεται και ο όρος d, δηλαδή ο αριθμός των μεταβλητών, αλλά τελικά έχουμε μείωση του AIC.
Το κριτήριο BIC (Bayesian information criterion) προτάθηκε από τον Schwarz (1978) και ορίζεται από την σχέση:

$$BIC = d\log n - 2\log L,$$

όπου

- L η μέγιστη τιμή της συνάρτησης πιθανοφάνειας για το εκτιμώμενο μοντέλο,
- d ο αριθμός παραμέτρων του μοντέλου μας,
- n ο αριθμός των παρατηρήσεων.

Η λογική και η χρήση του είναι παρόμοια με του AIC, μόνο που στην περίπτωση του κριτηρίου του BIC η εισαγωγή επιπρόσθετων παραμέτρων αποθαρρύνεται ακόμη περισσότερο σε σχέση με του AIC.
Στο παρακάτω σχήμα παρουσιάζεται η κλίμακα του Raftery, από την οποία μπορούμε να κρίνουμε κατά πόσο ένα μοντέλο έχει καλύτερη προσαρμογή στα δεδομένα μας σε σχέση με ένα άλλο. Το κριτήριο που θα χρησιμοποιήσουμε είναι η απόλυτη διαφορά των τιμών BIC των 2 μοντέλων.

Διαφορά των τιμών BIC	**Ένδειξη**
0-2	Ασθενής
2-8	Θετική
6-10	Ισχυρή
>10	Πολύ ισχυρή

Πίνακας 2.5 : Κριτήριο Raftery.

2.6.6 Κατάλοιπα

Τα **κατάλοιπα (residuals)** αποτελούν ένα μέτρο της απόκλισης μεταξύ των προσαρμοσμένων τιμών της μεταβλητής απόκρισης και των πραγματικών τιμών αυτής. Στα γενικευμένα γραμμικά μοντέλα υπάρχει διαφοροποίηση στον τρόπο υπολογισμού των καταλοίπων σε σχέση με το απλό γραμμικό μοντέλο καθώς η διασπορά δεν είναι πάντα σταθερή. Στο απλό γραμμικό μοντέλο χρησιμοποιούνται τα κατάλοιπα $e_i = y_i - \hat{y}_i, i = 1,2,..n$. Στα γενικευμένα γραμμικά μοντέλα τα κατάλοιπα ε_i δεν μπορούν να χρησιμοποιηθούν, γιατί οι διασπορές τους είναι άνισες. Εξαιτίας αυτού στα γενικευμένα γραμμικά μοντέλα χρησιμοποιούνται τα κατάλοιπα Pearson, Anscombe και Deviance τα οποία θα δούμε αναλυτικά παρακάτω. (McCullagh & Nelder (1989)).

Κατάλοιπα Pearson
Τα κατάλοιπα Pearson ορίζονται ως εξής:

$$e_i^p = \frac{y_i - \mu_i}{\sqrt{V(\mu_i)}}, i = 1,2,..n,$$

όπου $V(\mu_i)$ είναι η συνάρτηση διακύμανσης. Το όνομα των καταλοίπων προκύπτει από το γεγονός ότι για την κατανομή Poisson τα κατάλοιπα Pearson είναι η τετραγωνική ρίζα της συνάρτησης ελέγχου Pearson, για την οποία δείξαμε ότι είναι η στατιστική συνάρτηση ελέγχου X^2, άρα για την Poisson κατανομή ισχύει ότι $\sum_{i=1}^{n}(e_i^p)^2 = X^2$.

Κατάλοιπα Anscombe
Ένα μειονέκτημα των καταλοίπων Pearson είναι ότι η κατανομή των e_i^p για μη κανονικές κατανομές είναι μη συμμετρικές (skewed), οπότε δεν έχουν τις ίδιες ιδιότητες με τα κατάλοιπα των κανονικών κατανομών. Ο Anscombe πρότεινε να αντικατασταθεί ο όρος y_i με μία συνάρτηση $A(y_i)$, όπου $A(\cdot)$ είναι μια συνάρτηση η οποία κανονικοποιεί την κατανομή της

$A(Y)$. Ο Wedderburn (1972) έδειξε ότι για τις λογαριθμικές συναρτήσεις πιθανοφάνειας των γενικευμένων γραμμικών μοντέλων η συνάρτηση $A(\cdot)$ δίνεται από τον παρακάτω τύπο:

$$A(\cdot) = \int d\mu / V^{\frac{1}{3}}(\mu).$$

Επομένως για την κατανομή Poisson έχουμε ότι

$$A(\mu) = \int \frac{1}{\mu^{1/3}} d\mu = \frac{3}{2}\mu^{2/3}.$$

Η παραμετροποίηση που έγινε για την κανονικοποίηση της συνάρτηση κατανομής μέσω της συνάρτησης $A(\cdot)$ για την κατανομή Poisson δεν σταθεροποιεί ταυτόχρονα την διακύμανση. Για να επιτευχθεί αυτό αντικαθιστούμε τον όρο $\sqrt{V(\mu_i)}$ με τον κανονικοποιημένο όρο $\frac{A\prime(\mu)}{\sqrt{V(\mu_i)}}$. Για την κατανομή Poisson $\frac{A\prime(\mu)}{\sqrt{V(\mu_i)}} = \mu^{1/6}$ και τα κατάλοιπα Anscombe υπολογίζονται ως:

$$e_i^A = \frac{\frac{3}{2}\left(y_i^{\frac{2}{3}} - \mu_i^{\frac{2}{3}}\right)}{\mu_i^{1/6}}, i = 1,2,..n.$$

Για την κατανομή γάμμα τα κατάλοιπα Anscombe είναι της μορφής:

$$e_i^A = \frac{3\left(y_i^{\frac{1}{3}} - \mu_i^{\frac{1}{3}}\right)}{\mu_i^{1/3}}, i = 1,2,..n.$$

Κατάλοιπα Deviance

Τα κατάλοιπα deviance ορίζονται ως εξής:

$$e_i^d = sign(y_i - \mu_i)\sqrt{d_i(y_i, \mu_{i_i})}, \;\; i = 1,2,..n,$$

όπου $d_i(y_i, \mu_i)$ είναι η τιμή της συνάρτησης ελέγχου deviance D, για την $i-$οστή παρατήρηση. Το άθροισμα των τετραγώνων των καταλοίπων deviance ισούται με την συνάρτηση ελέγχου D, δηλαδή $\sum_{i=1}^{n} e_i^d = D$. Άρα για την κατανομή Poisson τα κατάλοιπα deviance θα είναι ίσα με:

$$e_i^d = sign(y_i - \mu_i)\{2(y_i\,(\log y_i/\mu_i) - y_i + \mu_i\}^{1/2}.$$

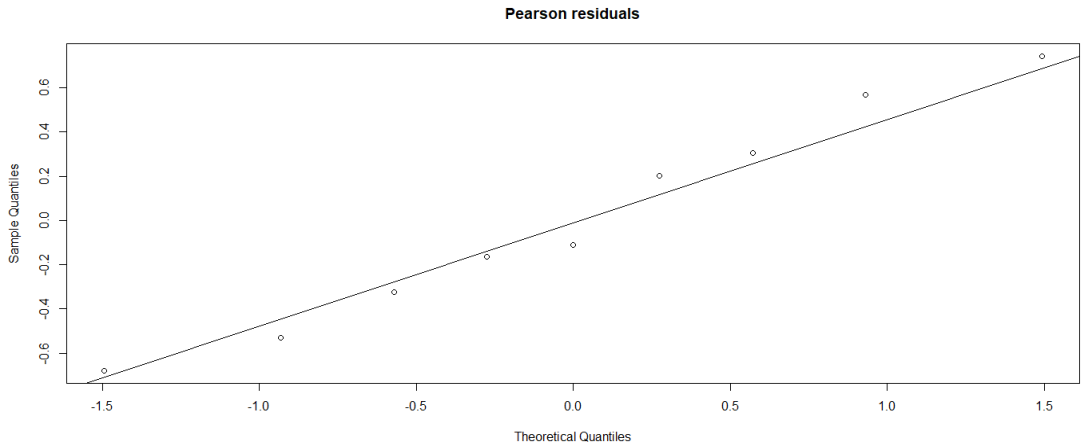
Παρόλο που τα κατάλοιπα Anscombe και τα κατάλοιπα Deviance φαίνεται να έχουν διαφορετικές συναρτήσεις για τον τρόπο υπολογισμού τους, οι τιμές που παίρνουν για δοσμένα y_i, μ_i είναι συνήθως πολύ κοντά.

Στον παρακάτω πίνακα φαίνονται τα κατάλοιπα Pearson, Anscombe, deviance του Παραδείγματος 2.1.

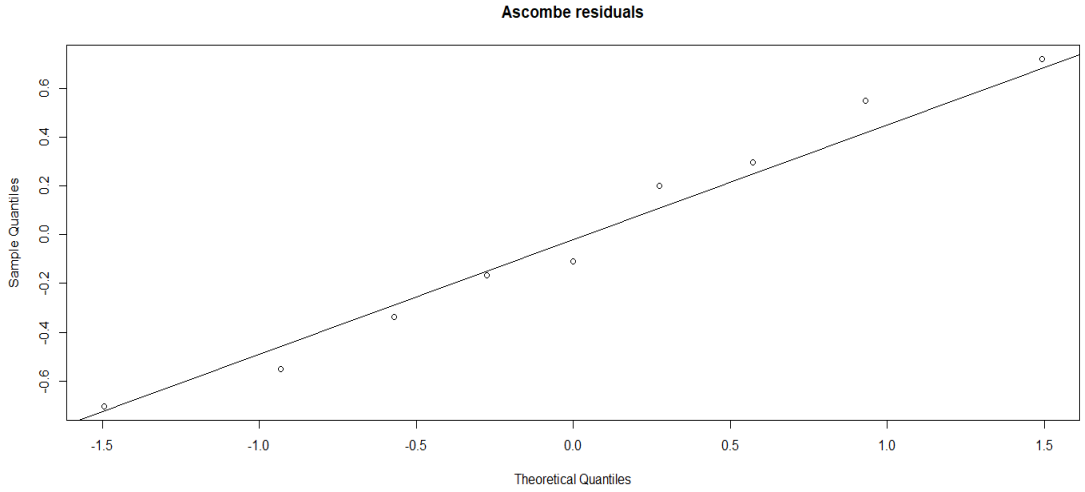
y_i	$\widehat{y}_i = \mu_i$	***Person residuals***	***Anscombe residuals***	***Deviance Residuals***
2	*2.51633*	*-0.32549*	*-0.3377821*	*-0.3377022*
3	*2.51633*	*0.30491*	*0.2958892*	*0.2958459*
6	*7.45163*	*-0.53178*	*-0.5507341*	*-0.5506179*
7	*7.45163*	*-0.16545*	*-0.1671642*	*-0.1671612*
8	*7.45163*	*0.20089*	*0.1984987*	*0.1984940*
9	*7.45163*	*0.56722*	*0.5491939*	*0.5491011*
10	*12.38693*	*-0.6782*	*-0.7020878*	*-0.7019431*
12	*12.38693*	*-0.10994*	*-0.1105191*	*-0.1105186*
15	*12.38693*	*0.74245*	*0.7185326*	*0.7184076*
72	***72***	***0.00461***	***-0.1061729***	***-0.1060944***

Πίνακας 2.6: Κατάλοιπα Pearson, Anscombe, deviance για το Παράδειγμα 1.

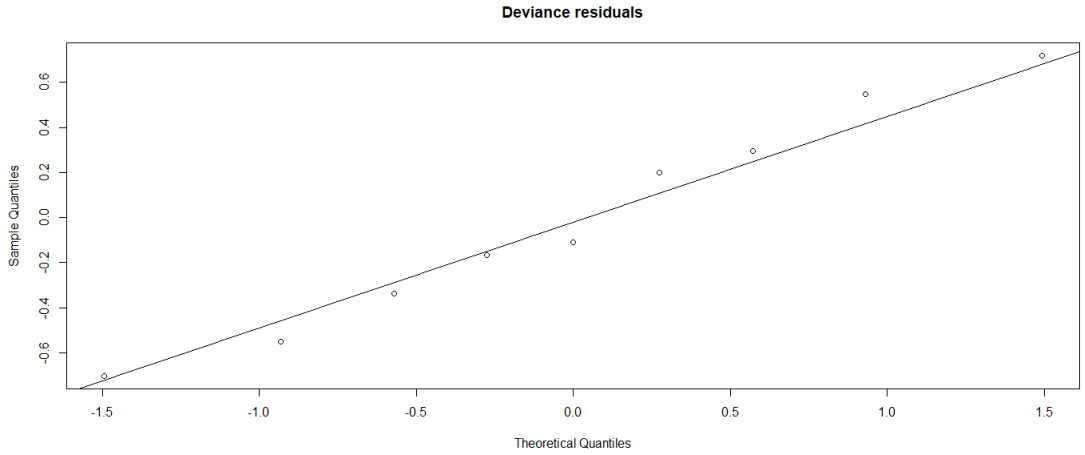
Παρακάτω βλέπουμε τα Q-Q γραφήματα των καταλοίπων. Παρατηρούμε ότι και στα 3 γραφήματα τα σημεία σχηματίζουν μια σχετικά καλά ορισμένη ευθεία που υποδηλώνει την μη ύπαρξη άτυπων σημείων και ενδεχομένως καλή προσαρμογή του μοντέλου μας.



Γράφημα 2.7.1: Διάγραμμα Q-Q των καταλοίπων Pearson

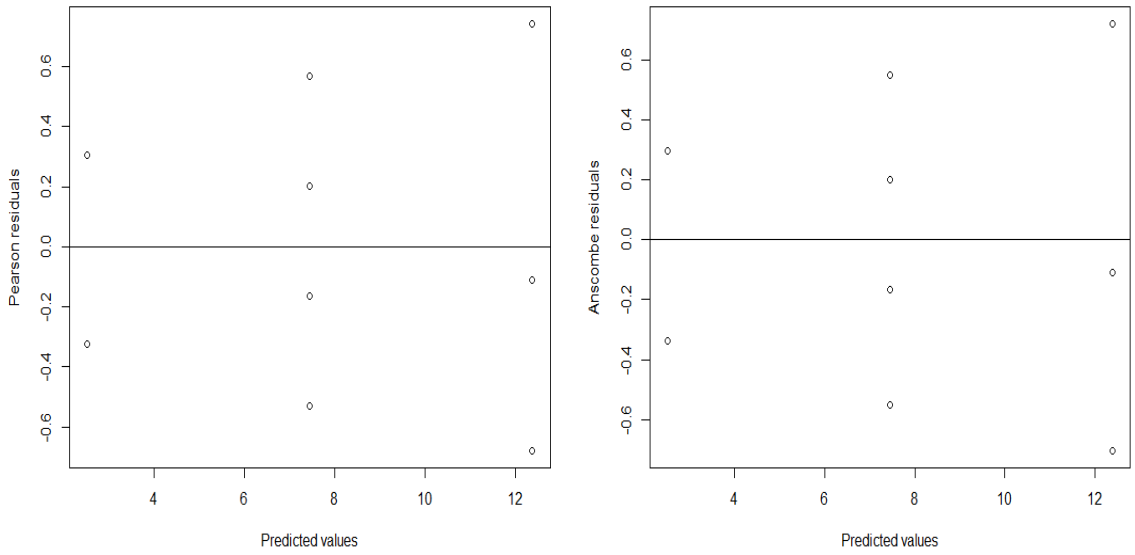


Γράφημα 2.7.2: Διάγραμμα Q-Q των καταλοίπων Anscombe

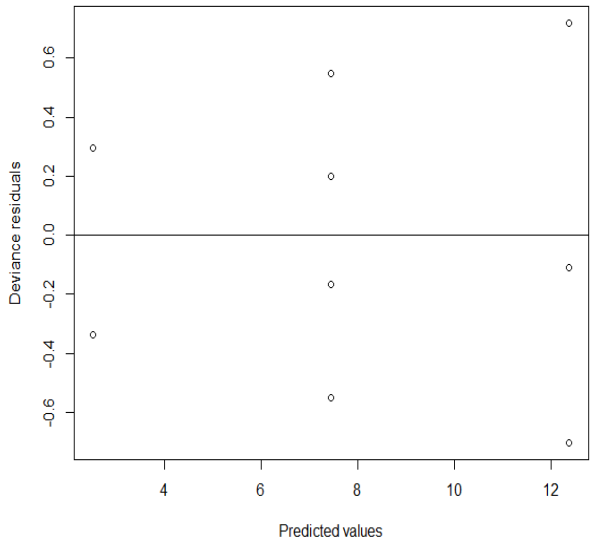


Γράφημα 2.7.3: Διάγραμμα Q-Q των καταλοίπων Deviance

Παρακάτω βλέπουμε τα γραφήματα των προβλεπόμενων τιμών έναντι των καταλοίπων. Παρατηρούμε πως τα κατάλοιπα είναι συμμετρικά κατανεμημένα ως προς την ευθεία $y = 0$ κάτι που υποδηλώνει ότι είναι κανονικά κατανεμημένα.



Γράφημα 2.7.4: Προβλεπόμενες τιμές $\hat{y}_i$ έναντι των καταλοίπων Pearson και Anscombe



Γράφημα 2.7.5: Προβλεπόμενες τιμές $\hat{y}_i$ έναντι των καταλοίπων Deviance

Στην συνέχεια βλέπουμε κάποια αποτελέσματα του Παραδείγματος 1 και πως αποτυπώνονται στην R.

```
> example1<-glm(yi~x1,data=w,family=poisson(link="identity"))
> summary(example1)

Call:
glm(formula = yi ~ x1, family = poisson(link = "identity"), data = w)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7019  -0.3377  -0.1105   0.2958   0.7184
```

```
Coefficients:
            Estimate Std. Error  z value Pr(>|z|)
(Intercept)   7.4516     0.8841   8.428  < 2e-16 ***
x1            4.9353     1.0892   4.531 5.86e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 18.4206  on 8  degrees of freedom
Residual deviance:  1.8947  on 7  degrees of freedom
AIC: 40.008

Number of Fisher Scoring iterations: 3
```

Μέσω της συνάρτησης glm (·) προσαρμόζουμε τα δεδομένα μας στο μοντέλο που έχουμε επιλέξει. Η εντολή summary(·) μας δίνει τις παρακάτω πληροφορίες:

Deviance Residuals: Παρουσιάζονται κάποιες πληροφορίες σχετικά με τα κατάλοιπα deviance, συγκεκριμένα η μέγιστη και η ελάχιστη τιμή αυτών καθώς και οι τιμές που βρίσκονται στο πρώτο και το τρίτο τεταρτημόριο.

Coefficients: Παρουσιάζονται κάποιες πληροφορίες σχετικά με τους συντελεστές της παλινδρόμησης που έχουμε εκτιμήσει. Συγκεκριμένα:

- Η στήλη **Estimate** περιέχει τις εκτιμήσεις των συντελεστών της παλινδρόμησης β_0, β_1.
- Η στήλη **Std. Error** περιέχει τα τυπικά σφάλματα των συντελεστών της παλινδρόμησης
- Η στήλη **z-value** περιέχει τις τιμές των ελέγχων **Wald**
- Η στήλη **Pr(>|z|)** περιέχει τις αντίστοιχες τιμές των ελέγχων

(Dispersion parameter for poisson family taken to be 1): Τιμή της παραμέτρου κλίμακας φ

Null Deviance: Η τιμή της συνάρτησης ελέγχου D για το μοντέλο που περιλαμβάνει μόνο τον σταθερό όρο.

Residual Deviance: Η τιμή της συνάρτησης ελέγχου D για το μοντέλο που περιλαμβάνει τους όρους που έχουμε προσαρμόσει τα δεδομένα μας.

AIC: Η τιμή του δείκτη AIC.

Number of Fischer Scoring iterations: Αριθμός των επαναλήψεων που χρειάστηκε ο αναδρομικός αλγόριθμος για να τερματίσει.

Με τις παρακάτω εντολές μπορούμε να πάρουμε τις προβλέψεις που προκύπτουν από το προσαρμοσμένο μοντέλο μας.

```
> predictions<-fitted.values(example1,type="response")
> data.frame(predictions)

  predictions
1    2.516332
2    2.516332
3    7.451633
4    7.451633
5    7.451633
6    7.451633
7   12.386934
8   12.386934
9   12.386934
```

Αφού αναλύσαμε τα γενικευμένα γραμμικά μοντέλα και τις ιδιότητες αυτών καθώς και το πώς αποτελούν επέκταση των απλών γραμμικών μοντέλων που μελετήσαμε στο Κεφάλαιο 1, στο επόμενο κεφάλαιο θα δούμε κάποιες εφαρμογές αυτών στον ασφαλιστικό τομέα και θα αναλύσουμε τα αποτελέσματα των εκάστοτε μοντέλων.

ΚΕΦΑΛΑΙΟ 3

Μοντέλο Συλλογικού Κινδύνου στην Γενική Ασφάλιση

3.1 Τιμολόγηση στην Ασφάλεια Αυτοκινήτου

Ατυχήματα συμβαίνουν κάθε μέρα. Κάποια είναι πιο σοβαρά από άλλα. Δεν μπορούμε να αποτρέψουμε τα ατυχήματα, από το να συμβούν αλλά μπορούμε να προστατεύσουμε τους εαυτούς μας σε περίπτωση οικονομικής ζημιάς εξαιτίας τους. Αυτό γίνεται μέσω της ασφάλισης. Ως καταναλωτές, προτιμάμε ασφάλειες οι οποίες καλύπτουν όσο το δυνατόν περισσότερα με όσο το δυνατόν χαμηλότερο κόστος. Ως ασφαλιστές, θέλουμε να έχουμε κέρδος και ταυτόχρονα να αποκτήσουμε περισσότερους πελάτες. Αν η τιμή του ασφαλίστρου είναι πολύ υψηλή, οι καταναλωτές θα προτιμήσουν ασφαλιστικές εταιρίες που παρέχουν τις ίδιες υπηρεσίες με φθηνότερο ασφάλιστρο. Αν η τιμή είναι πολύ χαμηλή, ο ασφαλιστής έχει μεγαλύτερο κίνδυνο να καλύψει χωρίς το ανάλογο αποθεματικό οπότε κινδυνεύει ακόμα και με χρεωκοπία. Άτομα τα οποία έχουν εμπλακεί στο παρελθόν σε πολλά ατυχήματα πληρώνουν ακριβότερο ασφάλιστρο λόγω υψηλότερου κινδύνου. Επιπλέον, όταν το ασφάλιστρο είναι πολύ χαμηλό ρισκάρεις σαν ασφαλιστική να προσελκύσεις πελάτες υψηλού ρίσκου, αφού η τιμή που θα πλήρωναν σε άλλες ασφαλιστικές εταιρίες θα ήταν πιο υψηλή. Εξαιτίας των παραπάνω συμπεραίνουμε ότι ο τρόπος κοστολόγησης ενός ασφαλισμένου είναι πολύ σημαντικός για μια ασφαλιστική εταιρία.

Υπάρχουν αρκετοί παράγοντες τους οποίους χρειάζεται να λάβουμε υπόψιν κατά την τιμολόγηση του ασφαλίστρου. Τα δεδομένα που έχουμε για τους καταναλωτές για παράδειγμα, αλλά και η γνώση γύρω από την αγορά στην οποία απευθυνόμαστε. Τα γενικευμένα γραμμικά μοντέλα χρησιμοποιούνται ευρέως στον ασφαλιστικό τομέα και βασίζονται σε αρκετές υποθέσεις μία εκ των οποίων είναι η ανεξαρτησία των δεδομένων. Σε πολλά στατιστικά προβλήματα όμως αυτή η υπόθεση μπορεί να μην ισχύει.

Η ασφάλεια χωρίζεται σε δύο βασικές κατηγορίες. Η πρώτη λέγεται ασφάλεια ζωής και αφορά τον κίνδυνο που καλύπτει η ασφαλιστική σχετικά με την ζωή. Σε αυτή την κατηγορία υπάγονται οι ασφάλειες θανάτου, συνταξιοδότησης, αναπηρίας κλπ. Η δεύτερη κατηγορία λέγεται γενική ασφάλιση και αφορά τον κίνδυνο που καλύπτει η ασφαλιστική σχετικά με υλικές ζημιές. Οι μέθοδοι που χρησιμοποιούνται ανάλογα με την κατηγορία της ασφάλισης διαφέρουν μεταξύ τους. Στο κεφάλαιο αυτό θα εξετάσουμε πως μπορούμε να υπολογίσουμε το καθαρό ασφάλιστρο με βάση κάποια χαρακτηριστικά του ασφαλισμένου για ένα χαρτοφυλάκιο αυτοκινήτων. Τα δεδομένα είναι πραγματικά και προέρχονται από μια Γαλλική ασφαλιστική εταιρία. Τέλος τα παραδείγματα οι εφαρμογές και το υπόλοιπο κεφάλαιο προέρχεται από το βιβλίο του Charpentier (2014) και συγκεκριμένα από το Κεφάλαιο 14 με τίτλο General Insurance Pricing.

3.2 Το Μοντέλο Συλλογικού Κινδύνου στην Γενική Ασφάλιση

Μια βασική αρχή για τον υπολογισμό του ασφαλίστρου είναι η αρχή της αναμενόμενης τιμής. Ο υπολογισμός του ασφαλίστρου συμπεριλαμβανομένου του κινδύνου δίνεται από τον παρακάτω τύπο:

$$\pi(S) = (1 + a)E(S), \tag{3.1}$$

όπου

- $a > 0$ είναι μια σταθερά.
- S είναι μια τυχαία μεταβλητή των (ετησίων) ζημιών.

Ας υποθέσουμε ότι N_t είναι μια τυχαία μεταβλητή, η οποία εκφράζει τον αριθμό των απαιτήσεων (claims) για μια περίοδο $[0, t]$ και Y_i είναι το ποσό για την $i -$ απαίτηση. Τότε η συνολική ζημιά για την περίοδο $[0, t]$ θα είναι:

$$S_t = \sum_{i=1}^{N_t} Y_i,$$

με $S_t = 0$ αν $N_t = 0$.
Εάν στην (3.1) θέσουμε όπου $a = 0$, τότε το ασφάλιστρο ονομάζεται καθαρό ασφάλιστρο (pure premium). Για να μπορέσουμε να υπολογίσουμε το καθαρό ασφάλιστρο θα πρέπει πρώτα να υπολογίσουμε το $E(S)$, όπου $S = S_1$ είναι η ετήσια συνολική ζημία. Αν $N_1 = N$ και $Y_1, Y_2, \dots, Y_n, \dots$ είναι ανεξάρτητα μεταξύ τους και αν οι ζημιές Y_i είναι ανεξάρτητες και ομοιόμορφα κατανεμημένες τότε:

$$\pi = E(S) = E(N)E(Y). \tag{3.2}$$

Από την σχέση (3.2) φαίνεται ότι το ετήσιο καθαρό ασφάλιστρο είναι το γινόμενο δύο όρων:

- $E(N)$ αναμενόμενη τιμή των ετησίων απαιτήσεων.
- $E(Y)$ αναμενόμενη τιμή των ατομικών αποζημιώσεων.

Ένας ακόμα παράγοντας που θα πρέπει να λάβουμε υπόψιν για τον υπολογισμό του ασφαλίστρου είναι η ετερογένεια. Στην συνέχεια θα δούμε ένα παράδειγμα υπολογισμού του ασφαλίστρου για δύο ασφαλιστικές εταιρίες οι οποίες ακολουθούν διαφορετική πολιτική τιμολόγησης. Ας υποθέσουμε ότι έχουμε μια δίτιμη τυχαία μεταβλητή Z η οποία παίρνει τις παρακάτω τιμές:

- Υψηλός κίνδυνος με πιθανότητα 50%.
- Χαμηλός Κίνδυνος με πιθανότητα 50%.

Επιπλέον, η τυχαία μεταβλητή N ακολουθεί διωνυμική κατανομή με πιθανότητα επιτυχίας είτε 10% είτε 20%, ανάλογα με την τιμή που θα πάρει η τυχαία μεταβλητή Z και η ζημιά Y είναι σταθερή με τιμή $Y = 100$. Τότε υπάρχουν δύο πιθανές τιμές για την τιμή του ασφαλίστρου οι οποίες φαίνονται παρακάτω:

- Η ασφαλιστική εταιρία να χρεώσει τους ασφαλισμένους της το ίδιο ποσό ασφαλίστρου, $\pi = E(S) = E(N)E(Y) = \frac{10+20}{100} * \frac{1}{2} * 100 = 15.$
- Η ασφαλιστική εταιρία να χρεώσει τους ασφαλισμένους της διαφορετικό ποσό ασφαλίστρου ανάλογα με την κατηγορία κινδύνου στην οποία ανήκουν. Για ασφαλισμένους χαμηλού κινδύνου το ποσό του ασφαλίστρου θα είναι

$\pi = E(S) = E(N)E(Y) = \frac{10}{100} * 100 = 10$, ενώ για ασφαλισμένους υψηλού κινδύνου το αντίστοιχο ποσό θα είναι $\pi = E(S) = E(N)E(Y) = \frac{20}{100} * 100 = 20$.

Υποθέτοντας ότι έχουμε δύο διαφορετικές ασφαλιστικές εταιρίες οι οποίες ακολουθούν διαφορετική πολιτική τιμολόγησης, τότε οι ασφαλισμένοι χαμηλότερου κινδύνου θα προτιμήσουν την δεύτερη εταιρία που χρεώνει ανάλογα τον κίνδυνο σε αντίθεση με τους ασφαλισμένους υψηλότερου κινδύνου οι οποίοι θα προτιμήσουν την πρώτη εταιρία η οποία χρεώνει 15 για αναμενόμενο κίνδυνο των 20. Από οικονομικής πλευράς φαίνεται πως η ασφαλιστική εταιρία η οποία δεν λαμβάνει υπόψιν τον κίνδυνο, δηλαδή την ετερογένεια μεταξύ των ασφαλισμένων δεν θα επιβιώσει σε μια ανταγωνιστική αγορά. Άρα, αν Z είναι η παρατηρούμενη μεταβλητή ετερογένειας, τότε η ασφαλιστική εταιρία θα πρέπει να χρεώνει:

$$\pi(z) = E(S|Z = z) = E(N|Z = z)E(Y|Z = z).$$

Το πρόβλημα που συχνά συναντάμε είναι το εξής. Δεν υπάρχουν δεδομένα τα οποία να μας επιτρέπουν να έχουμε ακριβή πρόβλεψη της ετερογένειας. Αντιθέτως, η ασφαλιστική εταιρία μπορεί να διαθέτει πληροφορίες σχετικά με τον ασφαλισμένο οι οποίες συνοψίζονται σε ένα διάνυσμα $\boldsymbol{X}$. Σε αυτή την περίπτωση, όπου έχουμε μερική εικόνα της ετερογένειας, η ασφαλιστική εταιρία θα πρέπει να χρεώνει:

$$\pi(x) = E(S|\boldsymbol{X} = \boldsymbol{x}) = E(N|\boldsymbol{X} = \boldsymbol{x})E(Y|\boldsymbol{X} = \boldsymbol{x}).$$

Σκοπός του κεφαλαίου θα είναι να προτείνουμε μοντέλα πρόβλεψης με την βοήθεια των γενικευμένων γραμμικών μοντέλων που είδαμε στο Κεφάλαιο 2, τα οποία θα μας βοηθήσουν να εκτιμήσουμε τις ποσότητες $E(N|\boldsymbol{X} = \boldsymbol{x})$, $E(Y|\boldsymbol{X} = \boldsymbol{x})$, όπου:

- $E(N|\boldsymbol{X} = \boldsymbol{x})$ είναι οι μέσες ετήσιες απαιτήσεις ενός ασφαλισμένου με χαρακτηριστικά x.
- $E(Y|\boldsymbol{X} = \boldsymbol{x})$ είναι μέσες ετήσιες αποζημιώσεις για έναν ασφαλισμένο με χαρακτηριστικά x.

Τα δεδομένα που θα χρησιμοποιήσουμε προέρχονται από μια Γαλλική ασφαλιστική εταιρία και περιέχουν πληροφορίες σχετικά με τους πελάτες και τα συμβόλαια για ένα χαρτοφυλάκιο αυτοκινήτων. Το πακέτο το οποίο χρειάζεται να εγκαταστήσουμε στην R, προκειμένου να έχουμε πρόσβαση στα δεδομένα ονομάζεται ***CASdatasets*** και έχει δημιουργηθεί με σκοπό να καλύψει τις ανάγκες του βιβλίου ***Computational Actuarial Science with R του Arthur Charpientier*** (βλέπε Charpientier, 2014) το οποίο θα ακολουθήσουμε στο παρόν κεφάλαιο. Μέσω των εντολών του Παραρτήματος Β εγκαθιστούμε το πακέτο που περιέχει τα δεδομένα που θα χρησιμοποιήσουμε στο υπόλοιπο του Κεφαλαίου 3.

Στην συνέχεια θα αναλύσουμε τα δύο σύνολα δεδομένων (datasets) που θα χρησιμοποιήσουμε στους υπολογισμούς μας.
Το πρώτο ονομάζεται **freMTPLfreq** και περιέχει δεδομένα σχετικά με κάποια χαρακτηριστικά του πελάτη όπως ηλικία, περιοχή κλπ. καθώς και πληροφορίες που αφορούν το συμβόλαιο. Θα το αναθέσουμε σε μια μεταβλητή με το όνομα Contracts και θα δούμε λίγο αναλυτικότερα τις στήλες από τις οποίες αποτελείται.
Η μεταβλητή Contracts περιέχει 413169 παρατηρήσεις από 10 στήλες των οποίων τα ονόματα είναι:

- **PolicyId,** μοναδικός αριθμός συμβολαίου
- **ClaimNB**, αριθμός απαιτήσεων κατά την περίοδο έκθεσης στον κίνδυνο.
- **Exposure**, έκθεση στον κίνδυνο (σε χρόνια).
- **Power**, δύναμη του αυτοκινήτου (ταξινομημένο με βάση κάποια κλίμακα).
- **CarAge**, ηλικία του αυτοκινήτου σε χρόνια.
- **DriverAge**, ηλικία του οδηγού σε χρόνια (18 και άνω).
- **Brand**, μάρκα του αυτοκινήτου.
- **Gas**, βενζίνη (Diesel ή Regular).
- **Region**, περιοχή της Γαλλίας.
- **Density**, πυκνότητα των κατοίκων στην περιοχή (αριθμός κατοίκων ανά τετραγωνικό μέτρο) που ζει ο οδηγός του αυτοκινήτου.

Το δεύτερο ονομάζεται **freMTPLSev** και περιέχει δεδομένα σχετικά με τις απαιτήσεις των ασφαλισμένων για την ίδια ασφαλιστική εταιρία. Θα το αναθέσουμε σε μια μεταβλητή με το όνομα Claims και θα δούμε λίγο αναλυτικότερα τις στήλες από τις οποίες αποτελείται.

Η μεταβλητή Claims περιέχει 16181 παρατηρήσεις από 2 στήλες των οποίων τα ονόματα είναι:

- **PolicyID**, μοναδικός αριθμός συμβολαίου (ίδιος με αυτό του Contracts έτσι ώστε να συνδέονται μεταξύ τους),
- **ClaimAmount**, κόστος της απαίτησης.

Τέλος με τις θα χωρίσουμε σε διαστήματα κάποιες μεταβλητές καθώς θα μας διευκολύνει στην συνέχεια στους υπολογισμούς μας. (βλ. Παράρτημα Β)

3.3 Αριθμός των Απαιτήσεων

Ο αριθμός των απαιτήσεων ακολουθεί διακριτή κατανομή και είναι πάντα θετικός αριθμός. Οι πιο συνηθισμένες κατανομές τις οποίες υποθέτουμε για τον αριθμό των απαιτήσεων είναι μεταξύ άλλων η κατανομή Poisson και η αρνητική διωνυμική.

3.3.1 Παλινδρόμηση Poisson

Ας υποθέσουμε ότι οι απαιτήσεις για έναν ασφαλισμένο, ακολουθούν κατανομή Poisson με παράμετρο $\lambda > 0$ και ότι ο αριθμός των απαιτήσεων για ένα χρονικό διάστημα $[t, t+h]$ ακολουθούν την κατανομή Poisson με παράμετρο λh, με $h > 0$. Σκοπός μας είναι να συνδέσουμε την ετήσια συχνότητα με την παρατηρούμενη συχνότητα για μία συγκεκριμένη περίοδο έκθεσης στον κίνδυνο. Για i-ασφαλισμένο ισχύουν τα παρακάτω:

- Ο ετήσιος αριθμός απαιτήσεων, N_i για την περίοδο $[0,1]$ είναι συνήθως μία μη παρατηρούμενη μεταβλητή.

- Ο πραγματικός αριθμός απαιτήσεων για τις οποίες έχουμε δεδομένα στην βάση, Y_i για την περίοδο $[0, E_i]$, όπου E_i η έκθεση στον κίνδυνο, είναι μία παρατηρούμενη μεταβλητή.

Τότε η μέση τιμή και η δειγματική διακύμανση της τυχαίας μεταβλητής N χωρίς να χρησιμοποιήσουμε κάποια επεξηγηματική μεταβλητή θα είναι αντίστοιχα:

- $m = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} E_i}$,
- $s^2 = \frac{\sum_{i=1}^{n} [Y_i - mE_i]^2}{\sum_{i=1}^{n} E_i}$,

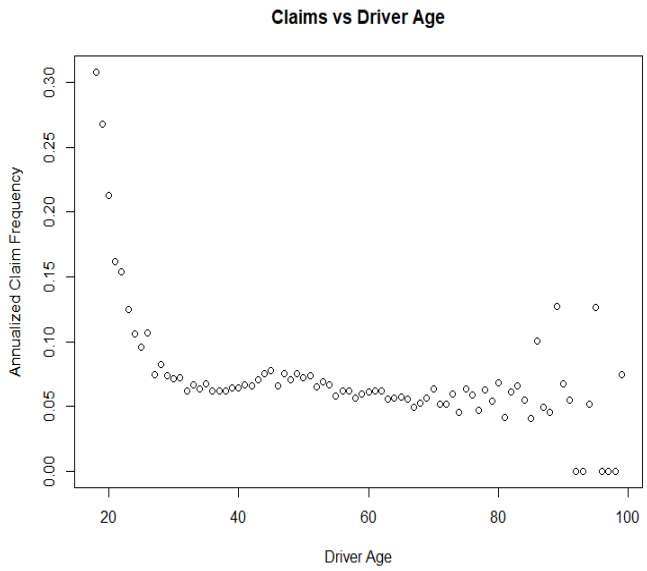
οι τιμές των οποίων υπολογίζονται μέσω R (βλ. Παράρτημα Β) και είναι ίσες με

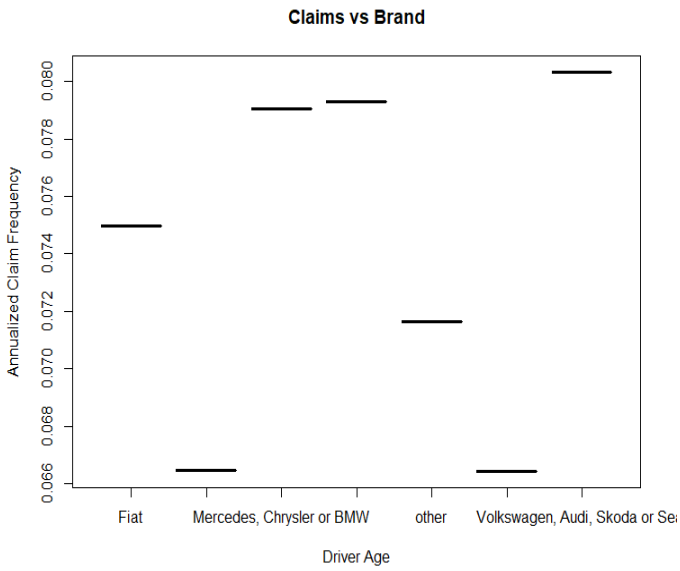
$\boldsymbol{m}$	*0.06979859*
$\boldsymbol{s^2}$	*0.07396742*

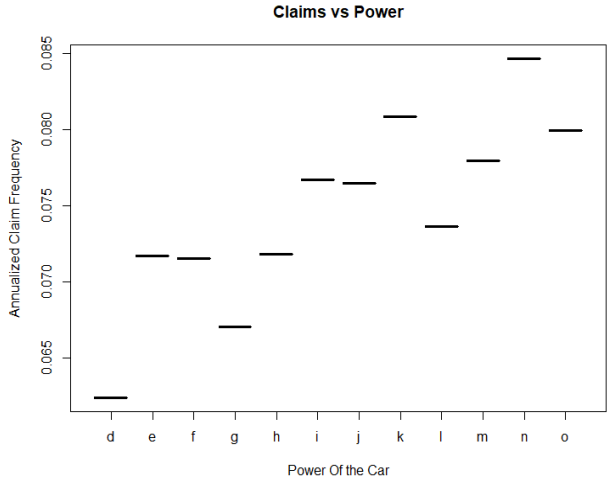
Πίνακας 3.1: Αποτελέσματα δειγματικής μέσης τιμής και διακύμανσης.

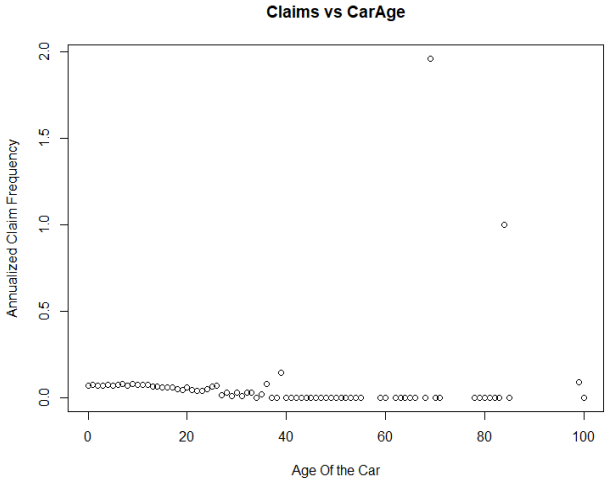
Όπως γνωρίζουμε για την Poisson κατανομή οι δύο ποσότητες θα πρέπει να είναι ίσες.

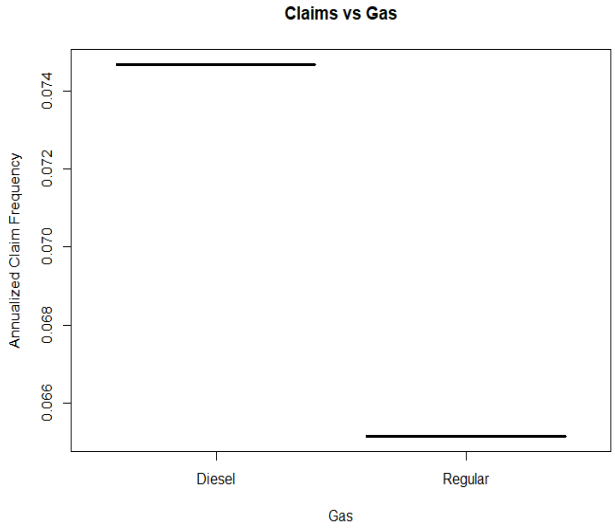
Στην συνέχεια και με την βοήθεια sql εντολών μέσω της R μπορούμε να μελετήσουμε την σχέση μεταξύ της μεταβλητής απόκρισης Y_i, η οποία για εμάς είναι ο αριθμός των ετήσιων απαιτήσεων, και της κάθε μίας επεξηγηματικής μεταβλητής ξεχωριστά. Επίσης θα συνδέσουμε την παρατηρούμενη έκθεση στον κίνδυνο με την ετήσια. Οι εντολές και τα αποτελέσματα φαίνονται στο Παράρτημα Β. Παρακάτω βλέπουμε τις αντίστοιχες γραφικές απεικονίσεις.

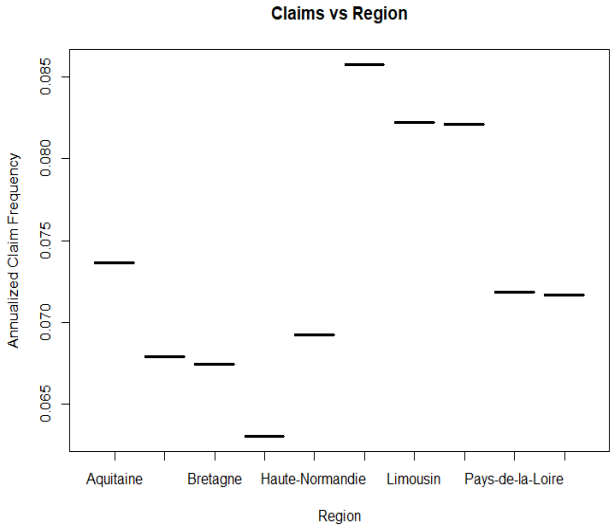


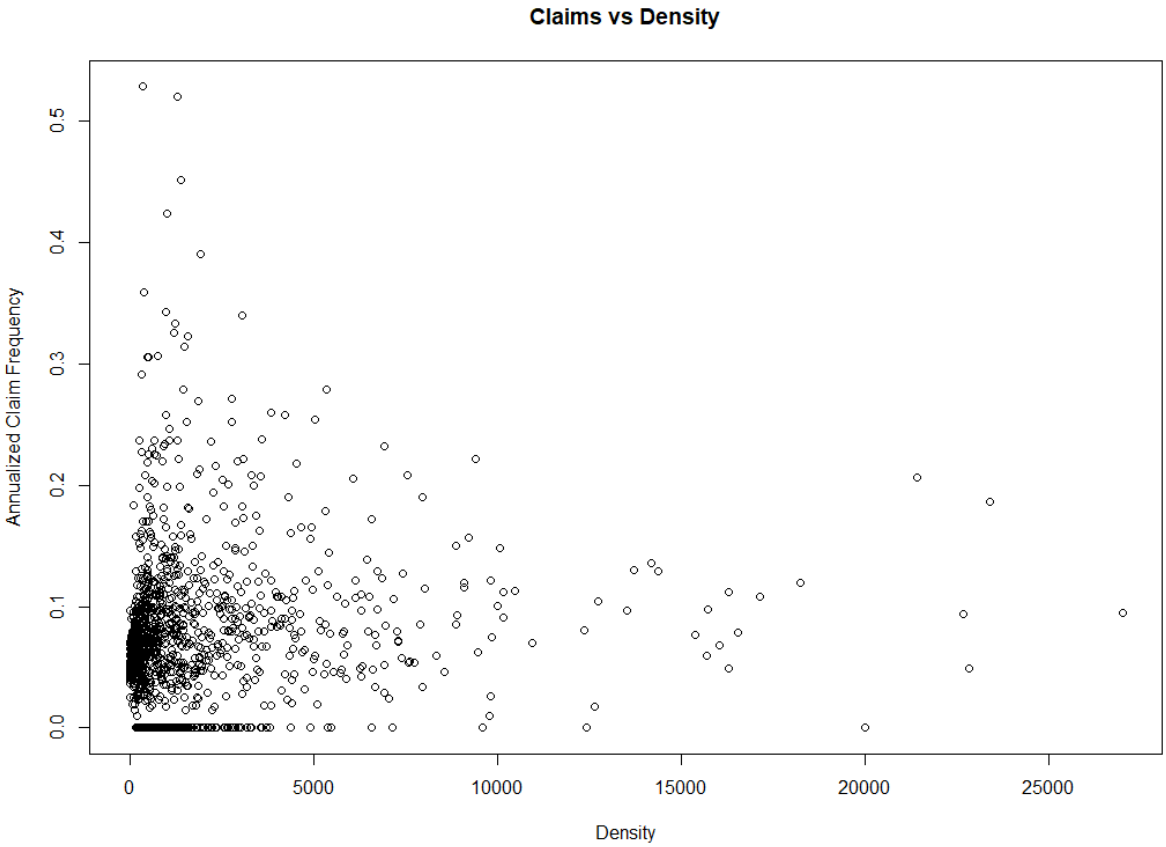



Claims vs Power
Annualized Claim Frequency
0.065
0.070
0.075
0.080
0.085
d
e
f
g
h
i
j
k
l
m
n
o
Power Of the Car


Claims vs CarAge
Annualized Claim Frequency
0.0
0.5
1.0
1.5
2.0
0
20
40
60
80
100
Age Of the Car


Claims vs Gas
Annualized Claim Frequency
0.066
0.068
0.070
0.072
0.074
Diesel
Regular
Gas


Claims vs Region
Annualized Claim Frequency
0.065
0.070
0.075
0.080
0.085
Aquitaine
Bretagne
Haute-Normandie
Limousin
Pays-de-la-Loire
Region



Γράφημα 3.1: Περιγραφή της μεταβλητής απόκρισης ως προς κάθε επεξηγηματική μεταβλητή ξεχωριστά.

Ας υποθέσουμε ότι ο ετήσιος αριθμός των απαιτήσεων για τον i −ασφαλισμένο είναι N_i και ότι $N_i \sim P(\lambda)$, δηλαδή όλοι ασφαλισμένοι έχουν τον ίδιο αριθμό αναμενόμενων απαιτήσεων. Αν παρατηρούσαμε τον i −ασφαλισμένο για μια περίοδο E_i τότε ο αριθμός των απαιτήσεων Y_i ακολουθεί κατανομή Poisson με παράμετρο λE_i, δηλ. $Y_i \sim P(\lambda E_i)$. Για την εκτίμηση του λ θα χρησιμοποιήσουμε την μέθοδο μέγιστης πιθανοφάνειας που αναπτύξαμε στο Κεφάλαιο 2. Η συνάρτηση πιθανοφάνειας είναι

$$L(\lambda, Y, E) = \prod_{i=1}^{n} \frac{e^{-\lambda E_i}[E_i]^{Y_i}}{Y_i!},$$

άρα η λογαριθμική συνάρτηση πιθανοφάνειας είναι

$$\log\ L(\lambda, Y, E) = -\lambda \sum_{i=1}^{n} E_i + \sum_{i=1}^{n} Y_i \log[\lambda E_i] - \log\left(\prod_{i=1}^{n} Y_i!\right).$$

Η παράγωγος ως προς λ είναι

$$\frac{\partial}{\partial \lambda} \log L(\lambda, Y, E) = 0 \Rightarrow -\sum_{i=1}^{n} E_i + \frac{1}{\lambda} \sum_{i=1}^{n} Y_i = 0,$$

άρα για $\lambda = \hat{\lambda}$

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} E_i} = \sum_{i=1}^{n} \omega_i \frac{Y_i}{E_i},$$

με $\omega_i = \frac{E_i}{\sum_{i=1}^{n} E_i}$.

Όπως αναφέραμε στην εισαγωγή, είναι πιο λογικό να θεωρήσουμε ότι το λ εξαρτάται από τον ασφαλισμένο. Υποθέτουμε ότι, $N_i \sim P(\lambda_i)$, με N να είναι ο ετήσιος αριθμός απαιτήσεων τον οποίο θέλουμε να προβλέψουμε. Η τυχαία μεταβλητή N_i δεν είναι μια παρατηρούμενη τυχαία μεταβλητή, άρα δεν έχουμε επαρκή δεδομένα για αυτήν. Παρ' όλα αυτά ο αριθμός των απαιτήσεων Y_i κατά την περίοδο έκθεσης στον κίνδυνο E_i μπορεί να παρατηρηθεί και υπάρχουν επαρκή δεδομένα τα οποία θα χρησιμοποιήσουμε για την πρόβλεψη της μεταβλητής N_i. Οπότε η $N_i \sim P(\lambda_i)$, μπορεί να γραφτεί ισοδύναμα ως $Y_i \sim P(\lambda_i E_i)$. Χρησιμοποιώντας την λογαριθμική συνάρτηση σύνδεσης έχουμε τα παρακάτω

$$g(\lambda_i) = \log(\lambda_i) = x_i^T \beta.$$

Επομένως,

$$\lambda_i = e^{x_i^T \beta} \text{ και } Y_i \sim P(e^{x_i^T \beta + \log(E_i)}).$$

Η έκθεση στον κίνδυνο χρησιμοποιείται ως μια επεξηγηματική μεταβλητή της οποίας όμως ο συντελεστής δεν χρειάζεται να εκτιμηθεί και θεωρούμε ότι είναι ίσος με 1. Αυτός είναι και ο αντισταθμιστικός (offset) παράγοντας της παλινδρόμησης. Άρα για να προβλέψουμε την τυχαία μεταβλητή N, δημιουργούμε ένα μοντέλο παλινδρόμησης με την μεταβλητή πρόβλεψης να είναι ο παρατηρούμενος αριθμός των απαιτήσεων Y_i, και ο αντισταθμιστικός παράγοντας να είναι ο λογάριθμος της έκθεσης στον κίνδυνο E_i. Για $\lambda_i = e^{x_i^T \beta}$ η λογαριθμική συνάρτηση πιθανοφάνειας γράφεται ως

$$\log L(\beta; Y, E) = \sum_{i=1}^{n} [Y_i \log(\lambda_i E_i) - [\lambda_i E_i] - \log(Y_i!)]$$

$$= \sum_{i=1}^{n} Y_i [x_i^T \beta + \log(E_i)] - \exp[x_i^T \beta + \log(E_i)] - \log(Y_i!).$$

Οι εκτιμητές μέγιστης πιθανοφάνειας μπορούν να βρεθούν μέσω της μεθόδου Fisher Scoring που αναπτύξαμε στο δεύτερο κεφάλαιο. Εμείς για τους υπολογισμούς μας θα χρησιμοποιήσουμε την συνάρτηση glm(·) της R. Ο γενικός τρόπος ορισμού της συνάρτησης για το μοντέλο μας είναι ο παρακάτω:

```
> glm(Y~X1+X2+X3+offset(E),family=poisson(link="log"))
```

Ορίζουμε την κατανομή που ακολουθεί η τυχαία μεταβλητή Y μέσω του ορίσματος `family=poisson` και την συνάρτηση σύνδεσης μέσω του ορίσματος `link="log"`. Για παράδειγμα αν θέλουμε να προβλέψουμε την μεταβλητή ClaimNB μέσω των επεξηγηματικών μεταβλητών Gas, DriverAge και Density καταλήγουμε στα παρακάτω αποτελέσματα για τους συντελεστές παλινδρόμησης (βλ. Παράρτημα Β) με την ηλικία του οδηγού να είναι χωρισμένη στα διαστήματα (22,26], (26,42], (42,74], [74, Inf) και με την πυκνότητα των κατοίκων να είναι χωρισμένη στα διαστήματα (40,200], (200,500], (500,74], [74, Inf).

Συντ/στες Παλινδρόμησης	Εκτίμηση συντελεστών			
β_0	-1.86471			
β_{gas}	-0.20598			
$\beta_{DriverAge}$	-0.61606	-1.07765	-1.10706	
$\beta_{Density}$	0.18473	0.31822	0.52694	0.63717

Πίνακας 3.2: Αποτελέσματα μοντέλου Poisson $\log Y_i = \beta_0 + \beta_{gas} + \beta_{DriverAge} + \beta_{Density} + \exp E$.

Ας υποθέσουμε ότι έχουμε μία κατηγορική επεξηγηματική μεταβλητή, για παράδειγμα το είδος της βενζίνης (το οποίο είναι είτε Diesel είτε κανονική). Ο αριθμός των απαιτήσεων ανάλογα το είδος της βενζίνης είναι ο παρακάτω

Diesel	Regular
8446	7735

Πίνακας 3.3: Αριθμός των απαιτήσεων ανάλογα το είδος της βενζίνης.

και η συνολική έκθεση στον κίνδυνο είναι ίση με

Diesel	Regular
113104.8	118719.4

Πίνακας 3.4: Συνολική έκθεση στον κίνδυνο ανάλογα το είδος της βενζίνης.

Οπότε διαιρώντας τον αριθμό των απαιτήσεων με την συνολική έκθεση στον κίνδυνο μπορούμε να υπολογίσουμε την ετήσια συχνότητα των απαιτήσεων είναι ανάλογα το είδος της βενζίνης.

Diesel	Regular
0.07467412	0.06515364

Πίνακας 3.5: Ετήσια συχνότητα των απαιτήσεων ανάλογα το είδος της βενζίνης.

Αν χρησιμοποιήσουμε την παλινδρόμηση Poisson όπως είδαμε προηγουμένως χωρίς όμως να έχουμε τον σταθερό όρο β_o θα πάρουμε τα παρακάτω αποτελέσματα για τους συντελεστές.

Συντ/στες Παλινδρόμησης	Εκτίμηση συντελεστών
β_{Diesel}	-2.59462
$\beta_{Regulas}$	-2.73101

Πίνακας 3.6: Αποτελέσματα μοντέλου Poisson $\log Y_i = \beta_{gas} X_1$.

Επομένως το γενικευμένο γραμμικό μοντέλο έχει την παρακάτω μορφή:

$$\log \mu_i = \beta_{gas} X_1 = \begin{cases} -2.59462, \alpha\nu\ X_1 = Diesel \\ -2.73101, \alpha\nu\ X_1 = Regular \end{cases}$$

$$\Longrightarrow \mu_i = e^{\beta_{gas} X_1} = \begin{cases} e^{-2.59462}, \alpha\nu\ X_1 = Diesel \\ e^{-2.7310}, \alpha\nu\ X_1 = Regular \end{cases}$$

Τα ίδια αποτελέσματα λαμβάνουμε μέσω της R μέσω της παρακάτω εντολής. Παρατηρούμε ότι οι προβλεπόμενες τιμές είναι ίσες με την ετήσια συχνότητα των απαιτήσεων ανά είδος βενζίνης.

```
> exp(coefficients(regpoislog))
  x1Diesel  x1Regular

0.07467412 0.06515364
```

Ενσωματώνοντας και τον σταθερό όρο β_oστο μοντέλο τα αποτελέσματα διαμορφώνονται ως εξής

Συντ/στες Παλινδρόμησης	Εκτίμηση συντελεστών
β_0	*-2.59462*
$\beta_{Regulas}$	*-0.13639*

Πίνακας 3.7: Αποτελέσματα μοντέλου Poisson $\log Y_i = \beta_0 + \beta_{gas}X_1$

Συνεπώς, το γενικευμένο γραμμικό μοντέλο έχει την παρακάτω μορφή:

$$\log \mu_i = \beta_0 + \beta_{gas}X_1 = \begin{cases} -2.59462, \alpha\nu\ X_1 = Diesel \\ -2.59462 - 0.13639 = -2.73101, \alpha\nu\ X_1 = Regular \end{cases}$$

$$\Longrightarrow \mu_i = e^{\beta_0 + \beta_{gas}X_1} = \begin{cases} e^{-2.59462}, \alpha\nu\ X_1 = Diesel \\ e^{-2.73101}, \alpha\nu\ X_1 = Regular \end{cases}$$

Τα ίδια αποτελέσματα λαμβάνουμε μέσω της R μέσω της παρακάτω εντολής.

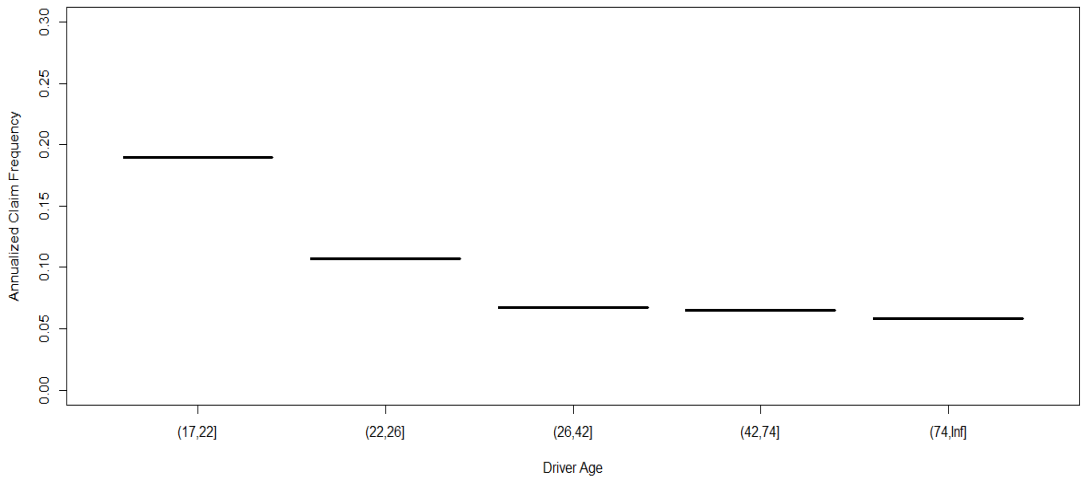
```
> exp(coefficients(regpoislog1))
(Intercept)   x1Regular
 0.07467412  0.87250624

> prod(exp(coefficients(regpoislog1)))
[1] 0.06515364
```

Από τα παραπάνω αποτελέσματα παρατηρούμε ότι η συχνότητα των απαιτήσεων για τα αμάξια που κινούνται με Diesel είναι 0.0746 ενώ για τα αμάξια που κινούνται με κανονική βενζίνη είναι 0.87250624 τις τιμής αναφοράς, δηλαδή 0.06515364. Παρατηρούμε ότι και με τα δύο μοντέλα οι τιμές μας συμπίπτουν.

Παραπάνω είδαμε κάποια παραδείγματα μοντέλων Poisson για ποιοτικά δεδομένα. Στα ασφαλιστικά δεδομένα όμως συναντάμε και συνεχή δεδομένα, όπως για παράδειγμα είναι το βάρος του αυτοκινήτου, η τιμή του αυτοκινήτου, η ηλικία του οδηγού κλπ. Αν πάρουμε για παράδειγμα την ηλικία του οδηγού μπορούμε να διαπιστώσουμε ότι συνήθως οι πιο έμπειροι οδηγοί μειώνουν τον δείκτη ατυχημάτων, ενώ ταυτόχρονα ο δείκτης ατυχημάτων ανεβαίνει όταν οι οδηγοί είναι ηλικιωμένοι. Για να ενσωματώσουμε συνεχείς μεταβλητές στο μοντέλο μας φτιάχνουμε διαφορετικά διαστήματα για την κάθε μεταβλητή, στα οποία θα βρίσκεται όλο το εύρος των τιμών της και για κάθε διάστημα υπολογίζουμε διαφορετικό συντελεστή. Παρακάτω βλέπουμε τα αποτελέσματα ενός μοντέλου Poisson στο οποίο χωρίζουμε την επεξηγηματική μεταβλητή DriverAge σε διαστήματα των (17,22], (22,26], (26,42], (42,74], (74,Inf). Τέλος χρησιμοποιούμε την λογαριθμική συνάρτηση σύνδεσης. Τα αποτελέσματα φαίνονται παρακάτω και ο αντίστοιχος κώδικας που χρησιμοποιήθηκε στην R βρίσκεται στο Παράρτημα Β.

Συντ/στες Παλινδρόμησης	Εκτίμηση συντελεστών			
β_0	-1.66337			
$\beta_{DriverAge}$	-0.56935	-1.04009	-1.06454	-1.17659

Πίνακας 3.8: Αποτελέσματα μοντέλου Poisson $\log Y_i = \beta_0 + \beta_{DriverAge} X_1$



Γράφημα 3.2: Γραφική αναπαράσταση του γενικευμένου γραμμικού μοντέλου παλινδρόμησης Poisson με την ηλικία ως διάστημα.

Στο επόμενο παράδειγμα βλέπουμε πως θα διαμορφωνόταν το μοντέλο μας αν δεν χωρίζαμε σε διαστήματα την ηλικία του οδηγού και θεωρούσαμε πως για κάθε ηλικία ο συντελεστής παλινδρόμησης είναι ίδιος. Τα αποτελέσματα φαίνονται παρακάτω.

Συντ/στες Παλινδρόμησης	Εκτίμηση συντελεστών
β_0	-2.1513378
$\beta_{DriverAge}$	-0.0111060

Πίνακας 3.9: Αποτελέσματα μοντέλου Poisson $\log Y_i = \beta_0 + \beta_{DriverAge} X_1$.

	Μοντέλο Α	Μοντέλο Β
AIC	136001	136467
Null Deviance	105613	105613
Residual Deviance	104734	105206

Πίνακας 3.10: Δείκτες AIC, Null Deviance, Residual Deviance για τα δύο μοντέλα.

Παρατηρούμε ότι οι δείκτες Null Deviance, Residual Deviance και AIC του πρώτου μοντέλου είναι μικρότεροι από του δεύτερου άρα προτιμούμε το πρώτο μοντέλο (βλ. Παράρτημα Β). Το δεύτερο μοντέλο δεν λαμβάνει υπόψιν την ηλικιακή ομάδα που βρίσκεται ο οδηγός. Τα αποτελέσματα φαίνονται και γραφικά. (Γράφημα 3.2, Γράφημα 3.3).

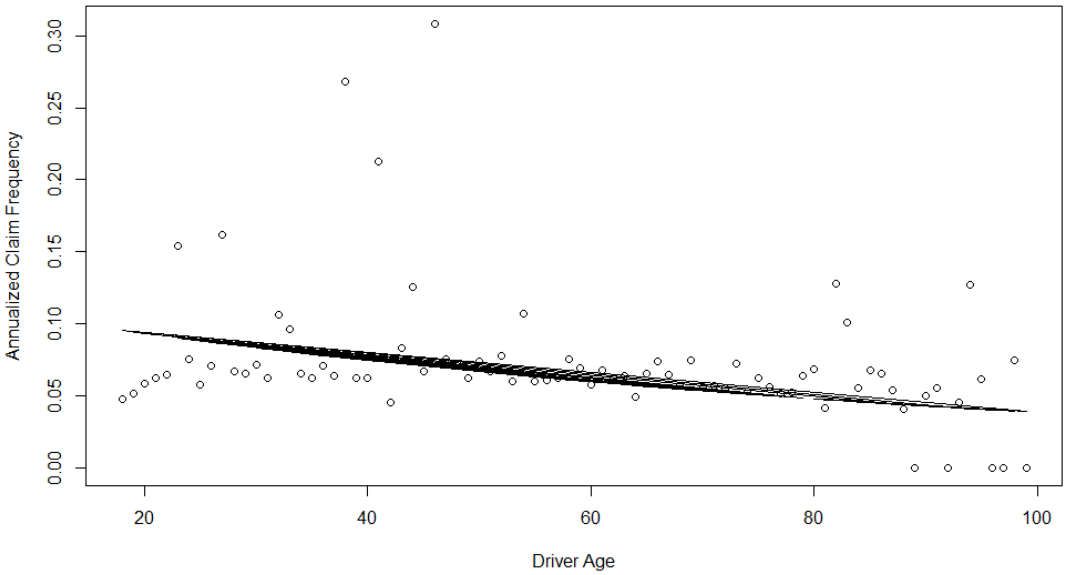
Να σημειώσουμε εδώ ότι αν θελήσουμε να πάρουμε τις εκτιμώμενες τιμές για το δεύτερο μοντέλο θα παρατηρήσουμε ότι έχουμε διαφορετικές εκτιμήσεις για την ίδια ηλικία καθώς υπολογίζεται ο αντισταθμιστικός παράγοντας offset(log(Exposure)) με βάση τα

δεδομένα, ο οποίος δεν είναι σταθερός και ίσος με ένα. Για να το αποτρέψουμε αυτό χρησιμοποιούμε τις παρακάτω εντολές για να πάρουμε τις εκτιμώμενες τιμές για κάθε ηλικία.

```
> newdb<-data.frame(DriverAge=18:99,Exposure=1)
> pred.poisson<-predict(AgedDriver3,newdata = newdb,type="response",se=TRUE
)
```

Με την παρακάτω εντολή μπορούμε να δούμε την ευθεία παλινδρόμησης για το μοντέλο Poisson AgedDriver3.

```
> plot(18:99,pred.poisson$fit,ylim=c(0,.3),type="l",xlab="Age of the driver
",ylab="Annualized Claim Frequency")
```



Γράφημα 3.3: Γραφική αναπαράσταση του γενικευμένου γραμμικού μοντέλου παλινδρόμησης Poisson $E(Y) = e^{-2.1513378-(0.0111060*X)+\log E}$.

3.3.2 Αρνητικό Διωνυμικό Μοντέλο

Όπως αναφέραμε και στην εισαγωγή, ο αριθμός των απαιτήσεων πέρα από την κατανομή Poisson μπορεί να ακολουθεί αρνητική διωνυμική κατανομή. Στο μοντέλο Poisson η μέση τιμή είναι ίση με την διακύμανση κάτι το οποίο δεν ισχύει πάντα για τα δεδομένα μας. Σε αυτές τις περιπτώσεις καταφεύγουμε στην αρνητική διωνυμική κατανομή η οποία έχει δύο παραμέτρους και μπορεί να προσαρμοστεί καλύτερα στα δεδομένα μας από ότι η Poisson.

Αν υποθέσουμε ότι $Y/\Theta \sim P(\lambda\Theta)$ ακολουθεί την αρνητική διωνυμική κατανομή και Θ ακολουθεί την Γάμα κατανομή με ίδια παράμετρο κλίμακας και σχήματος $a > 0$, τέτοιες ώστε $E(\Theta) = 1$. Τότε η τυχαία μεταβλητή Y ακολουθεί την αρνητική διωνυμική κατανομή με συνάρτηση πιθανότητας

$$P(Y = y) = \frac{\Gamma(y + a^{-1})}{\Gamma(y + 1)\Gamma(a^{-1})}\left(\frac{1}{1 + \lambda/\alpha}\right)^{a^{-1}}\left(1 - \frac{1}{1 + \lambda/\alpha}\right)^{y}, \forall y \in \mathbb{N}.$$

Θέτουμε $r = a^{-1}$ και $p = (1 + a\lambda)^{-1}$ τότε η παραπάνω γίνεται

$$f(y) = P(Y = y) = \binom{y}{y + r - 1} p^r (1 - p)^y, \forall y \in \mathbb{N},$$

το οποίο γράφεται ως

$$f(y) = \exp\left[ylog(1 - p) + rlogp + log\binom{y}{y + r - 1}\right], \forall y \in \mathbb{N}$$

και ανήκει στην εκθετική οικογένεια κατανομών με $\alpha(y) = y, b(p) = \log(1 - p)$, $c(p) = rlog(p)$ και γνωστό r. Η μέση τιμή είναι:

$$E(Y) = \frac{-c'(p)}{b'(\theta)} = \frac{r(1 - p)}{p} = \lambda$$

και η διακύμανση

$$Var(Y) = [b''(p)c'(p) - c''(p)b'(p)]/b'(p)^3 \quad = \frac{r(1-p)}{p^2}.$$

Η διακύμανση μπορεί να γραφεί και ως

$$Var(Y) = \frac{1}{p}E(Y) = [1 + a\lambda]\lambda.$$

Για $E(Y) = \lambda = \mu$ και $Var(Y) = [1 + a\lambda]\lambda$ έχουμε το Αρνητικό Διωνυμικό Μοντέλο Παλινδρόμησης Τύπου 2 (NB2), (βλέπε Hilbe (2011)).

Η κανονική συνάρτηση σύνδεσης, η οποία είναι τέτοια ώστε

$$g(\lambda) = \theta = \log(1 - p) = \log\left(1 - \frac{1}{1 + a\lambda}\right) = \log\left(\frac{a\lambda}{1 + \alpha\lambda}\right),$$

είναι η:

$$g(\mu) = \log\left(\frac{a\mu}{1 + \alpha\mu}\right) = \log(\alpha\mu) - \log(1 + a\mu). \tag{3.3}$$

Η συνάρτηση που χρησιμοποιείται στην R για αρνητικά διωνυμικά μοντέλα τύπου 2 ονομάζεται glm.nb(·) και βρίσκεται στο πακέτο **MASS**.

```
> library(MASS)
> glm.nb(Y~X1+X2+X3+offset(log(E)))
```

Παρατηρούμε ότι σε αυτή την μορφή δεν έχουμε δώσει τιμή ούτε για την συνάρτηση σύνδεσης, αλλά ούτε και για την κατανομή των δεδομένων μας. Επίσης στην προκειμένη η μεταβλητή a είναι άγνωστη και η εκτίμηση της γίνεται μέσω της εντολής **summary( )**. Στο Παράρτημα Β μπορούμε να δούμε ένα παράδειγμα αρνητικού διωνυμικού μοντέλου παλινδρόμησης με άγνωστη τιμή για το a.

Η παράμετρος α βρίσκεται στην μεταβλητή Theta και είναι ίση με $\alpha = 0.879$. Επίσης μπορούμε να βελτιστοποιήσουμε το μοντέλο μην λάμβάνοντας υπόψιν κάποιες επεξηγηματικές μεταβλητές οι οποίες δεν είναι στατιστικά σημαντικές (π.χ. CarAge(1,4)). Αξίζει να αναφέρουμε πως η κανονική συνάρτηση σύνδεσης (3.3) δεν είναι αποδεκτή επιλογή στην R για την αρνητική διωνυμική, οπότε καταφεύγουμε συνήθως στην λογαριθμική συνάρτηση σύνδεσης log.

Στην περίπτωση όπου έχουμε γνωστό α, χρησιμοποιούμε την συνάρτηση glm(·) ως εξής

```
> glm(Y~X1+X2+X3+offset(log(E)),family=negative.binomiale(1))
```

Σε αυτή την περίπτωση θεωρούμε ότι $\alpha = 1$, $Var(Y) = \lambda + \lambda^2$ και

$$g(\mu) = \log\left(\frac{\mu}{1+\mu}\right).$$

3.4 Ατομικές Απαιτήσεις και Προτεινόμενα Μοντέλα

Στην εισαγωγή του Κεφαλαίου 3, είδαμε ότι το καθαρό ασφάλιστρο το οποίο αφορά έναν ασφαλισμένο με χαρακτηριστικά x είναι ίσο με $\pi(x) = E(N|\boldsymbol{X} = \boldsymbol{x})E(Y|\boldsymbol{X} = \boldsymbol{x})$. Στις προηγούμενες παραγράφους δόθηκαν κάποια μοντέλα για τον υπολογισμό του όρου $E(N|\boldsymbol{X} = \boldsymbol{x})$. Σκοπός των παρακάτω παραγράφων είναι να προτείνουμε μοντέλα για τον υπολογισμό του $E(Y|\boldsymbol{X} = \boldsymbol{x})$, της ατομικής απαίτησης δηλαδή για τον κάθε ασφαλισμένο με βάση τα χαρακτηριστικά του. Τα μοντέλα που θα χρησιμοποιήσουμε και εδώ είναι τα γενικευμένα γραμμικά μοντέλα και τα δεδομένα μας προέρχονται όπως είπαμε και προηγουμένως από μία Γαλλική ασφαλιστική εταιρία, περιέχονται στην μεταβλητή Claims και αποτελούνται από τις παρακάτω συνιστώσες:

- **PolicyID**, μοναδικός αριθμός συμβολαίου (ίδιος με αυτό του Contracts έτσι ώστε να συνδέονται μεταξύ τους)
- **ClaimAmount**, κόστος της απαίτησης

```
> tail(CLAIMS)
      PolicyID ClaimAmount
16176   303133         769
16177   302759          61
16178   299443        1831
16179   303389        4183
16180   304313         566
16181   206241        2156
```

Χρησιμοποιώντας την εντολή merge(·) μπορούμε να ενσωματώσουμε τα δεδομένα της μεταβλητής Claims σε αυτά της μεταβλητής Contracts, κάτι το οποίο θα μας βοηθήσει αργότερα στους υπολογισμούς μας.

```
> claims<- merge(CLAIMS,CONTRACTS)
> claims.f<-merge(CLAIMS,CONTRACTS.f)
```

Επειδή οι ατομικές απαιτήσεις είναι πάντα ένας θετικός αριθμός θα χρησιμοποιήσουμε συνεχείς κατανομές από την εκθετική οικογένεια κατανομών οι οποίες έχουν ως σύνολο τιμών τιμές στον $\mathbb{R}_+$.

3.4.1 Μοντέλο Γάμμα

Μία συνεχής τυχαία μεταβλητή Y ακολουθεί την κατανομή Γάμμα αν έχει συνάρτηση πυκνότητας (Charpentier (2014, p.499)):

$$f(y) = \frac{1}{y\Gamma(\varphi^{-1})}\left(\frac{y}{\mu\varphi}\right)^{\varphi^{-1}} e^{\frac{-y}{\mu\varphi}}, \forall y \in \mathbb{R}_+,$$

με $\varphi > 0$ να είναι η παράμετρος θέσης (shape parameter), $\mu > 0$ να είναι η παράμετρος κλίμακας (scale parameter) και μπορεί να γραφεί ως

$$f(y) = exp\left[\frac{\frac{y}{\mu} - (-\log(\mu))}{-\varphi} + \frac{1-\varphi}{\varphi}\log y - \frac{\log\varphi}{\varphi} - \log\Gamma(\varphi^{-1})\right], \forall y \in \mathbb{R}_{+,}$$

η οποία ανήκει στην εκθετική οικογένεια κατανομών για $\alpha(y) = y$, $b(\mu) = \frac{1}{-\varphi\mu}$, $c(\mu) = \frac{\log\mu}{-\varphi}$ και γνωστό φ. Η μέση τιμή είναι

$$E(Y) = \frac{-c'(\mu)}{b'(\mu)} = \frac{\frac{1}{\varphi\mu}}{\frac{1}{\varphi\mu^2}} = \frac{\varphi\mu^2}{\varphi\mu} = \mu.$$

Παρατηρούμε ότι η μέση τιμή του Y εξαρτάται μόνο από την παράμετρο κλίμακας μ. Η διακύμανση είναι αντίστοιχα:

$$Var(Y) = [b''(\mu)c'(\mu) - c''(\mu)b'(\mu)]/b'(\mu)^3 \ = \varphi\mu^2 = \varphi E(Y)^2.$$

Ορίζουμε ως συντελεστή μεταβλητότητας (coefficient of variation) μιας τυχαίας μεταβλητής Yκαι συμβολίζουμε ως CV, την ποσότητα:

$$CV = \frac{\sqrt{Var(Y)}}{E(Y)} = \sqrt{\varphi}.$$

Ο συντελεστής μεταβλητότητας είναι ένας καθαρός αριθμός απαλλαγμένος από μονάδες μέτρησης της μεταβλητής και είναι ένας δείκτης που εκφράζει τη διασπορά των τιμών σε σχέση με το μέσο. Παρατηρούμε ότι ανεξάρτητα από το πόσο μικρή ή μεγάλη είναι η μέση τιμή μ ο συντελεστής μεταβλητότητας παραμένει σταθερός αφού το φ είναι γνωστό. Όταν έχουμε τυχαίες μεταβλητές οι οποίες είναι θετικές και ασύμμετρες προς τα δεξιά (right skewed) και έχουν σταθερό συντελεστή μεταβλητότητας τότε αυτό αποτελεί μια ένδειξη ότι προέρχονται από την Γάμμα κατανομή.

Η κανονική συνάρτηση σύνδεσης για το μοντέλο Γάμμα είναι η $g(\mu_i) = -\frac{1}{\mu_i}$, η συνάρτηση όμως που τα περισσότερα πακέτα στατιστικής χρησιμοποιούν είναι η αντίστροφη θετική συνάρτηση σύνδεσης $\eta_i = g(\mu_i) = \frac{1}{\mu_i}$ για την οποία ισχύει ο περιορισμός $\mu_i > 0$, άρα και $\eta_i > 0$. Άλλες πιθανές επιλογές για την συνάρτηση σύνδεσης είναι οι παρακάτω:

- $g(\mu_i) = \log\mu_i$ (log-link).
- $g(\mu_i) = \mu_i$ με $\mu_i > 0$ (ταυτοτική).
- $g(\mu_i) = {\mu_i}^{\alpha}$.

Η απόκλιση είναι ίση με $D(y, \hat{\mu}) = -2\sum_{i=1}^{n}\left(\log\frac{y_i}{\hat{\mu}_i}\right)$. Όταν έχουμε μικρές τιμές για τον συντελεστή μεταβλητότητας $CV(CV < 0.7)$, τότε η κατανομή Γάμμα μπορεί να προσεγγιστεί από την λογαριθμοκανονική κατανομή.

3.4.2 Το Λογαριθμοκανονικό Μοντέλο

Η τυχαία μεταβλητή Y ακολουθεί την λογαριθμοκανονική κατανομή αν έχει συνάρτηση πυκνότητας

$$f(y) = \frac{1}{y\sqrt{2\pi\sigma^2}} e^{\left\{\frac{-(\ln y-\mu)^2}{2\sigma^2}\right\}}, \forall y \in \mathbb{R}_+,$$

η οποία δεν ανήκει στην εκθετική οικογένεια κατανομών (άρα δεν μπορούμε να χρησιμοποιή σουμε γενικευμένα γραμμικά μοντέλα για την μεταβλητή απόκρισης Y). Η μέση τιμή της μεταβλητής Y είναι $E(Y) = e^{\mu+\frac{\sigma^2}{2}}$. Επίσης ισχύει το εξής: αν $\log Y$ ακολουθεί την κανονική κατανομή (η οποία ανήκει στην ΕΟΚ) τότε η Y ακολουθεί την λογαριθμοκανονική κατανομή.

3.4.3 Σύγκριση Μοντέλου Γάμμα με το Λογαριθμοκανονικό

Έστω ότι τα Y_i έχουν προσαρμοστεί σε ένα μοντέλο Γάμμα. Τότε θεωρώντας την λογαριθμική συνάρτηση σύνδεσης και χρησιμοποιώντας την σειρά Taylor δεύτερης τάξης ο όρος $\log Y_i$ αναλύεται ως εξής:

$$\log Y_i \sim \log \mu_\iota + \frac{1}{\mu_i}[Y_i - \mu_i] - \frac{1}{2\mu_i^2}[Y_i - \mu_i]^2.$$

Εάν το φ είναι μικρό τότε,

$$E(\log Y_i) \sim \log \mu_i - \frac{1}{2\mu_i^2} Var[Y_i] = \log \mu_i - \frac{\varphi\mu_i^2}{2\mu_i^2} = \log \mu_i - \frac{1}{2}\varphi.$$

Επίσης, αποδεικνύεται ότι $Var(\log Y_i) \sim \varphi$. Με βάση τα παραπάνω και θεωρώντας την λογαριθμική συνάρτηση σύνδεσης για τις τυχαίες μεταβλητές Y_i οι οποίες έχουν διακύμανση $\varphi\mu_i^2$, όπου $\mu_i = e^{x_i^T}\beta$, τότε

$$E(\log Y_i) \sim x_i^T\beta - \frac{1}{2}\varphi$$

και

$$Var(\log Y_i) \sim \varphi.$$

Έστω τώρα ότι τα Y_i έχουν προσαρμοστεί σε ένα λογαριθμοκανονικό μοντέλο της μορφής:

$$\log Y_i = x_i^T a + \varepsilon_\iota,$$

όπου ε_ι τα τυχαία σφάλματα, με $\varepsilon_\iota \sim N(0, \sigma^2)$ και $Var(\log Y_i) = \sigma^2$.

Εάν ο συντελεστής μεταβλητότητας είναι μικρός τότε, οι εκτιμήσεις των συντελεστών των δύο μοντέλων β, α θα πρέπει να είναι σχετικά κοντά εκτός από τον σταθερό όρο. Παρακάτω βλέπουμε τα αποτελέσματα ενός πολλαπλού λογαριθμοκανονικού μοντέλου παλινδρόμησης με επεξηγηματικές μεταβλητές να είναι η ηλικία του αμαξιού χωρισμένη σε διάστημα των (0,15], (15, Inf) και το είδος της βενζίνης. Επίσης έχουμε περιορίσει τις απαιτήσεις μας να είναι μικρότερες των 15000.

Συντ/στες Παλινδρόμησης	Εκτίμηση log συντελεστών	Εκτίμηση συντελεστών
β_0	*7.22023*	*1366.8078397*
$\beta_{CarAge(15,Inf)}$	*-0.04508*	*0.9559250*
$\beta_{GasRegular}$	*-0.02203*	*0.9782092*

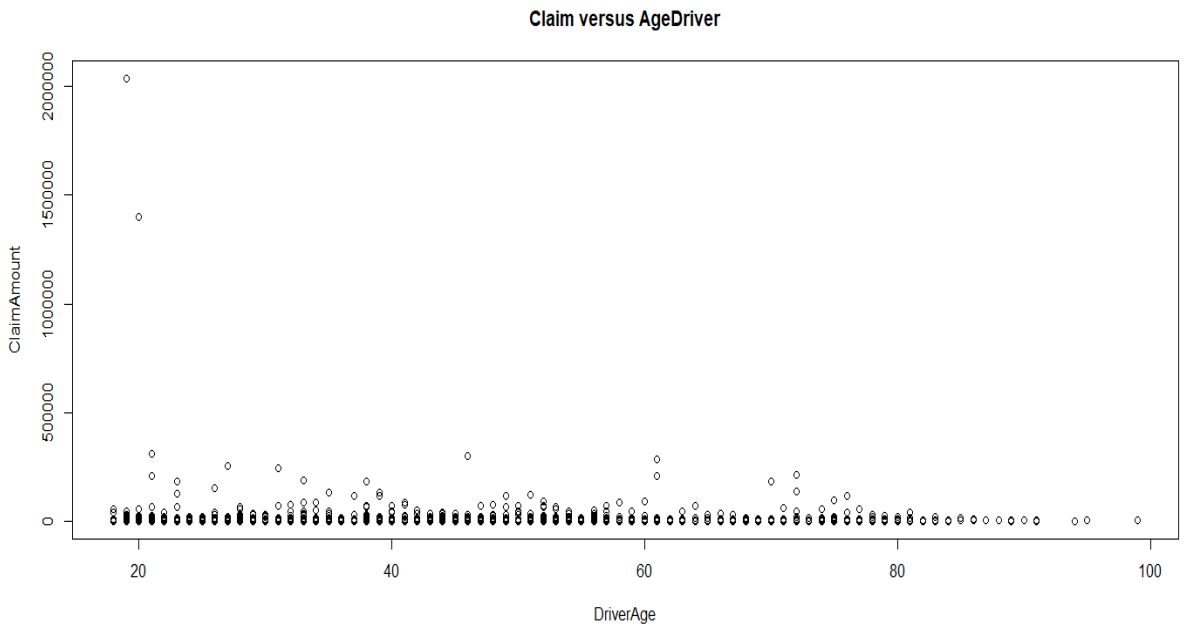
Πίνακας 3.11: Αποτελέσματα πολλαπλού λογαριθμοκανονικού μοντέλου για απαιτήσεις μικρότερες των 15000.

Στην συνέχεια παρουσιάζουμε τα αποτελέσματα που προκύπτουν από ένα αντίστοιχο γενικευμένο μοντέλο γάμμα για τις ίδιες επεξηγηματικές μεταβλητές.

Συντ/στες Παλινδρόμησης	Εκτίμηση log συντελεστών	Εκτίμηση συντελεστών
β_0	*6.79747*	*895.5777014*
$\beta_{CarAge(15,Inf)}$	*0.01670*	*1.0168369*
$\beta_{GasRegular}$	*-0.02591*	*0.9744271*

Πίνακας 3.12: Αποτελέσματα μοντέλου γάμμα για απαιτήσεις μικρότερες των 15000.

.

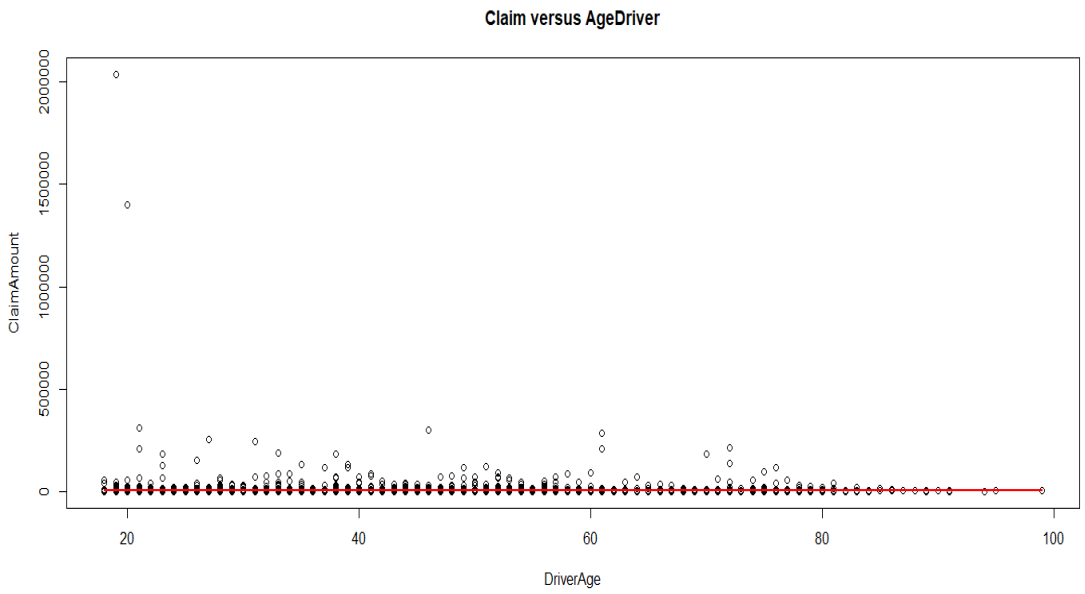


Γράφημα 3.4: Γραφική αναπαράσταση της μεταβλητής απόκρισης και της επεξηγηματικής μεταβλητής ηλικία του οδηγού.

Από το Γράφημα 3.4 παρατηρούμε ότι οι εκτιμήσεις είναι σχετικά κοντά για τα δύο μοντέλα πέρα από τον σταθερό όρο. Αν το ποσό των απαιτήσεων δεν είναι πολύ μεγάλο, τότε το λογαριθμικό και το γάμμα μοντέλο είναι σχετικά κοντά. Παρακάτω θα εξετάσουμε τα δύο μοντέλα χρησιμοποιώντας ως επεξηγηματική μεταβλητή την ηλικία του οδηγού (ως συνεχή μεταβλητή) και θα δούμε πως συμπεριφέρονται για μεγάλες απαιτήσεις. Θυμίζουμε ότι στα προηγούμενα παραδείγματα περιορίσαμε τις απαιτήσεις να είναι μικρότερες από 15000. Για τους συντελεστές του λογαριθμοκανονικού μοντέλου παίρνουμε τα παρακάτω αποτελέσματα.

Συντ/στες Παλινδρόμησης	Εκτίμηση log συντελεστών	Εκτίμηση συντελεστών
β_0	*6.7361807*	*842.33743*
$\beta_{DriverAge}$	*0.0020374*	*1.00204*

Πίνακας 3.13: Αποτελέσματα πολλαπλού λογαριθμοκανονικού μοντέλου για όλες τις απαιτήσεις.



Γράφημα 3.5: Γραφική αναπαράσταση του λογαριθμικού γραμμικού μοντέλου παλινδρόμησης $E(Y) = e^{6.7361807+(0.0020374*x)+\frac{\sigma^2}{2}}$.

Το σ είναι ίσο με:

```
> sigma<-summary(logn.reg)$sigma
> sigma
[1] 1.114849
```

Για το λογαριθμοκανονικό μοντέλο ισχύει ότι οι εκτιμώμενες τιμές είναι ίσες με $\hat{y} = e^{x_i^T a+(\sigma^2/2)}$, όπου σ^2 να είναι μια εκτίμηση της διακύμανσης. Για το μοντέλο γάμμα παίρνουμε τα παρακάτω αποτελέσματα.

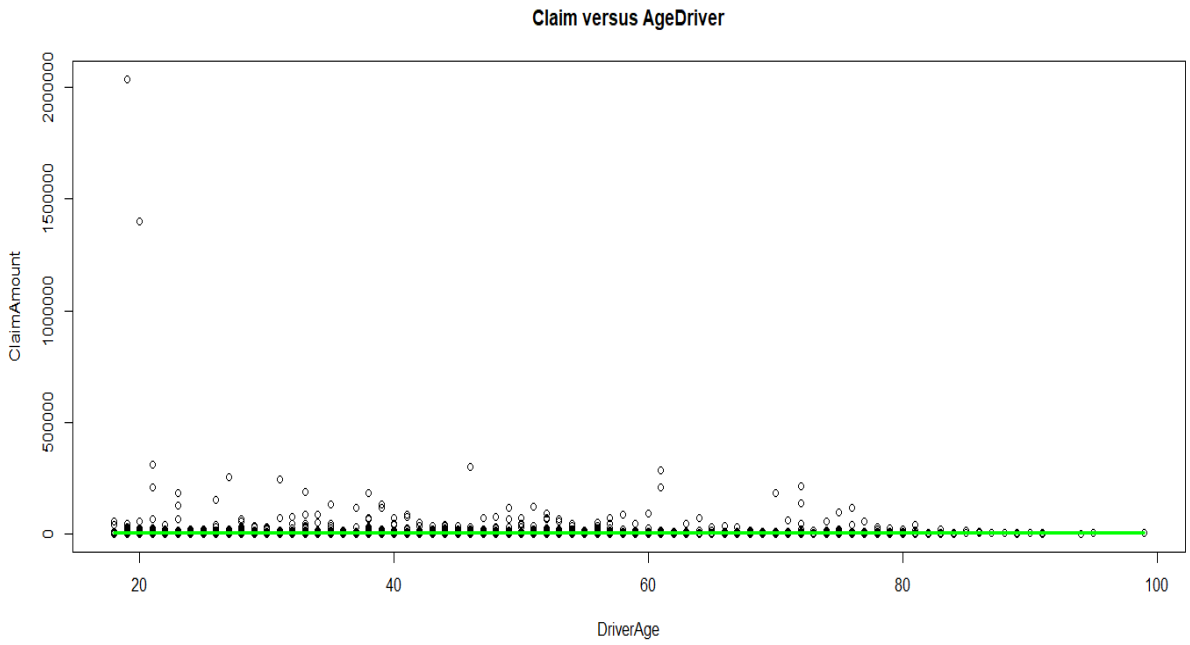
Συντ/στες Παλινδρόμησης	Εκτίμηση log συντελεστών	Εκτίμηση συντελεστών
β_0	*8.095074*	*3278.2803782*
$\beta_{DriverAge}$	*-0.009926*	*0.9901227*

Πίνακας 3.14: Αποτελέσματα γενικευμένου γραμμικού μοντέλου γάμμα για όλες τις απαιτήσεις

Παρατηρούμε ότι για τα δύο μοντέλα οι συντελεστές παλινδρόμησης έχουν αντίθετα πρόσημα. Στον πίνακα 3.1 του Παραρτήματος Α βλέπουμε τις εκτιμώμενες τιμές για κάθε μοντέλο στην κάθε ηλικία. Τα παραπάνω αποτελέσματα προκύπτουν από τους κώδικες του Παραρτήματος Β.

Παρατηρούμε ότι στις μικρές ηλικίες το μοντέλο Γάμμα φαίνεται να προσαρμόζεται καλύτερα καθώς το ποσό των απαιτήσεων είναι πιο μεγάλο σε σχέση με άλλες ηλικιακές ομάδες, ενώ το αντίθετο συμβαίνει με το λογαριθμοκανονικό μοντέλο όπου στις μικρές ηλικίες το ποσό των απαιτήσεων είναι μικρότερο απ' ότι σε άλλες ηλικιακές ομάδες. Επίσης

για τις μεγάλες ηλικίες φαίνεται ότι το λογαριθμοκανονικό μοντέλο προσαρμόζεται καλύτερα. Πιο συγκεκριμένα συμπεριφέρεται αντίθετα από ότι το γάμμα και το προβλεπόμενο ποσό απαιτήσεων είναι πιο υψηλό σε σχέση με το αντίστοιχο του γάμμα μοντέλου. Θεωρούμε λογικό ότι πολύ μικρές ηλικίες άρα άπειροι οδηγοί και πολύ μεγάλες ηλικίες είναι σε ομάδες με μεγαλύτερο κίνδυνο για τις ασφαλιστικές. Συμπεραίνουμε πως για μεγάλα ποσά απαιτήσεων (large claims) χρειαζόμαστε έναν νέο τρόπο προσέγγισης έτσι ώστε η τιμολόγηση του ασφαλισμένου να είναι λογική και δίκαιη. (Beirland & Teugels (1992))



Γράφημα 3.6: Γραφική αναπαράσταση του γάμμα μοντέλου παλινδρόμησης $E(Y) = e^{8.095074-(0.009926*X)}$.

3.4.4 Μοντέλα για Μεγάλες Απαιτήσεις

Στην θεωρία πιθανοτήτων ισχύει ότι $E(Y) = E(E(Y|Z))$ για οποιοδήποτε Z. Άρα για μία πολυωνυμική τυχαία μεταβλητή Z η οποία παίρνει τιμές z_i ισχύει:

$$E(Y) = \sum_i E(Y|Z = z_i)P(Z = z_i).$$

Άρα για κάθε δεσμευμένη πιθανότητα $E(Y|X)$ ισχύει:

$$E(Y|X) = \sum_i E(Y|Z = z_i, \boldsymbol{X})P(Z = z_i|\boldsymbol{X}).$$

Θεωρούμε ότι το Z αναπαριστά κάποια πληροφορία για το μέγεθος της απαίτησης και ανήκει είτε στο διάστημα $\{Y > s\}$ είτε στο $\{Y \leq s\}$ για κάποιο (υψηλό) ποσό s. Επομένως,

$$E(Y|X) = E(Y|X, Y \leq s)P(Y \leq s|X) + E(Y|X, Y > s)P(Y > s|X),$$

όπου

- $A = E(Y|X, Y \leq s)$ είναι το μέσο κόστος των μικρών απαιτήσεων (απαιτήσεις που δεν ξεπερνούν την τιμή s).

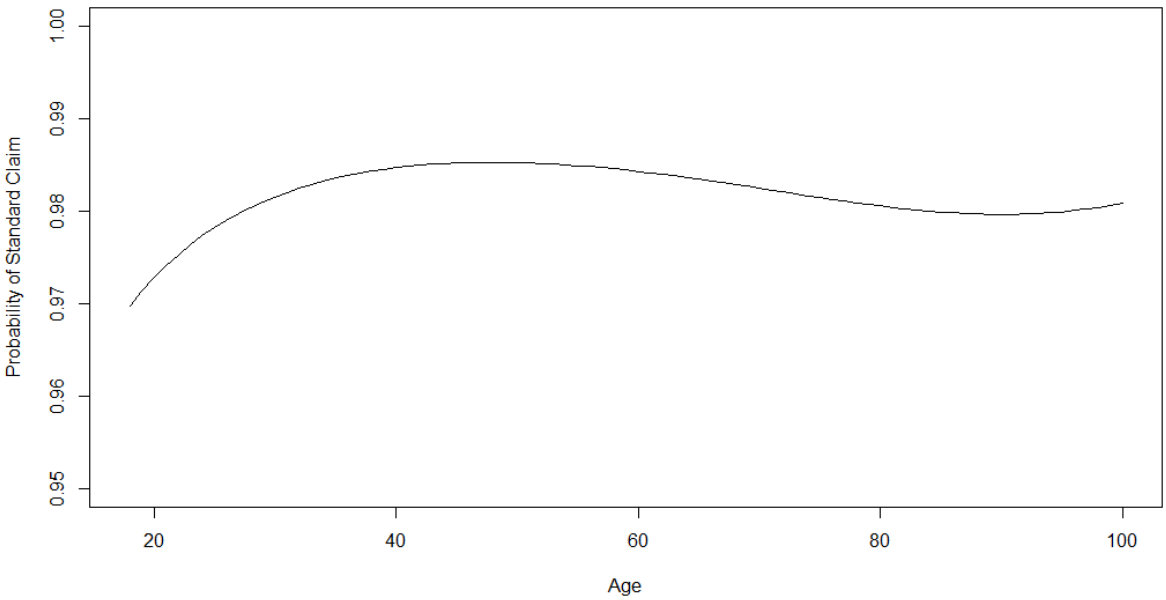
- $B = E(Y|X, Y > s)$ είναι το μέσο κόστος των μεγάλων απαιτήσεων (απαιτήσεις που ξεπερνούν την τιμή s).
- $C = P(Y \leq s|X), P(Y > s|X)$ είναι η πιθανότητα η απαίτηση να είναι μικρή ή μεγάλη αντίστοιχα.

Για την πρόβλεψη του όρου C μπορούμε να χρησιμοποιήσουμε λογιστική παλινδρόμηση (logistic regression) αφού πρόκειται για δύο πιθανά ενδεχόμενα. Για τους όρους A και B θα τρέξουμε δύο διαφορετικές παλινδρομήσεις αφού τα δεδομένα μας είναι ομαδοποιημένα. Θεωρούμε ότι το όριο s είναι ίσο με 10000 και το φτάνουν περίπου το 2% των απαιτήσεων.

```
> s<-10000
> claims$Standard<-(claims$ClaimAmount<s)
> mean(CLAIMS$Standard)
[1] 0.982943
```

Μέσω λογιστικής παλινδρόμησης θα υπολογίσουμε την πιθανότητα ότι η απαίτηση θα είναι μικρότερη του s.

```
> library(splines)
> age<-seq(18,100)
> regC<-glm(Standard~bs(DriverAge),data=claims,family=binomial (link="logit
"))
> pred_values<-predict(regC,newdata=data.frame(DriverAge=age),type="respons
e",se=TRUE)
```



Γράφημα 3.7: Η πιθανότητα να έχουμε μικρή απαίτηση, δεδομένου ότι υπάρχει απαίτηση, ως συνάρτηση της ηλικίας του οδηγού.

Ο παρακάτω κώδικας δημιουργεί το γράφημα 3.7.

```
> plot(age,pred_values$fit,ylim=c(0.95,1),type="l",colous="black",xlab="Age
",ylab="Probability of Standard Claim")
```

Για να μοντελοποιήσουμε τις μικρές και τις μεγάλες απαιτήσεις θα χρησιμοποιήσουμε δύο γάμμα μοντέλα για την προσαρμογή των δεδομένων μας.

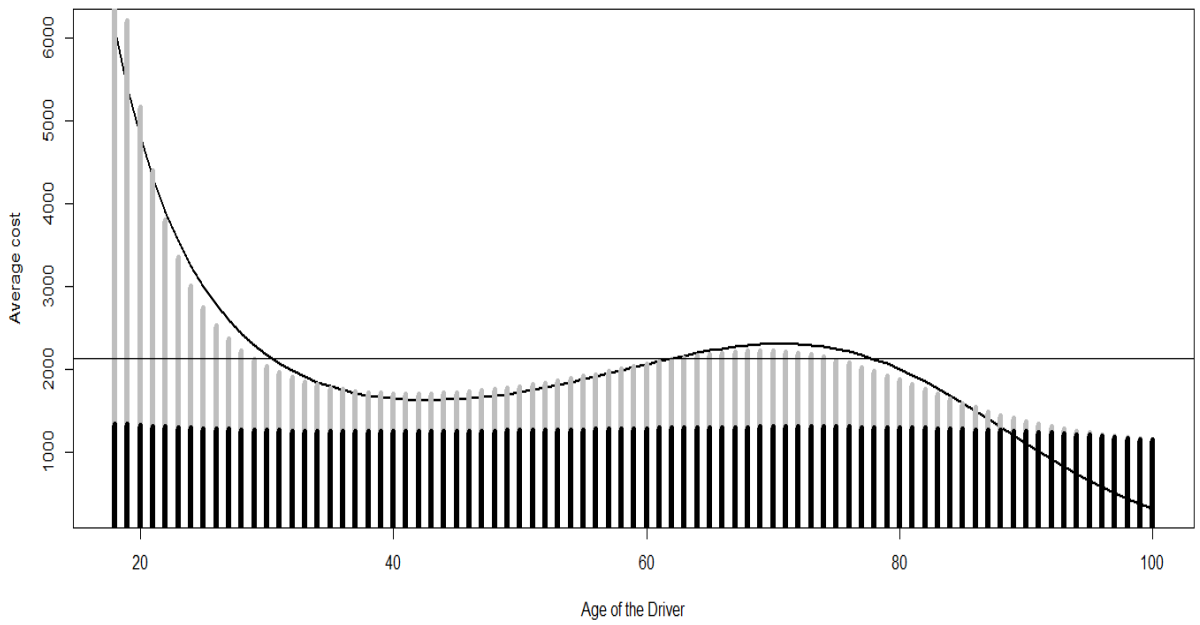
```
> indexstandard<-which(claims$ClaimAmount<s)
> mean(claims$ClaimAmount[indexstandard])
[1] 1280.085
> mean(claims$ClaimAmount[-indexstandard])
[1] 51106.23
```

```
> regA<-glm(ClaimAmount~bs(DriverAge),data=claims[indexstandard,],family=Ga
mma(link="log"))
> ypA<-predict(regA,newdata=data.frame(DriverAge=age),type="response")
> regB<-glm(ClaimAmount~bs(DriverAge),data=claims[-indexstandard,],family=G
amma(link="log"))
> ypB<-predict(regB,newdata=data.frame(DriverAge=age),type="response")
```

Για να κάνουμε την σύγκριση μεταξύ των μοντέλων θα προσαρμόσουμε όλα τα δεδομένα μας σε ένα μοντέλο γάμμα.

```
> reg=glm(ClaimAmount~bs(DriverAge),data=claims,family=Gamma(link="log"))
> yp<-predict(reg,newdata=data.frame(DriverAge=age),type="response")
> ypC<-predict(regC,newdata=data.frame(DriverAge=age),type="response")
```

```
> plot(age,yp,type="l",lwd=2,ylab="Average cost",xlab="Age of the Driver")
> lines(age,ypC*ypA+(1-ypC)*ypB,type="h",col="grey",lwd=6)
> lines(age,ypC*ypA,type="h",col="black",lwd=6)
> abline(h=mean(claims$ClaimAmount,col="orange",lty=5))
```



Γράφημα 3.8: Μέσο κόστος απαιτήσεων σε συνάρτηση με την ηλικία του οδηγού, για $s = 10000$.

Η οριζόντια γραμμή του γραφήματος 3.8 είναι το μέσο κόστος των απαιτήσεων. Η μαύρη γραμμή είναι οι προβλέψεις οι οποίες αφορούν τα μη ομαδοποιημένα δεδομένα reg. Η μαύρη επιφάνεια αφορά απαιτήσεις οι οποίες δεν ξεπερνούν το όριο s ενώ η γκρι επιφάνεια αφορά απαιτήσεις που ξεπερνούν το όριο s.

3.4.5.Η Επιλογή του Κατάλληλου Μοντέλου

Παραπάνω είδαμε διαφορετικά μοντέλα στα οποία προσαρμόσαμε τα δεδομένα μας ώστε να φτάσουμε στον τελικό στόχο μας ο οποίος είναι ο υπολογισμός του ασφαλίστρου $\pi = E(S) = E(N)E(Y)$. Για τον υπολογισμό του όρου $E(N)$ που ήταν ο αναμενόμενος αριθμός των απαιτήσεων για τον κάθε ασφαλισμένο με βάση κάποια χαρακτηριστικά του, χρησιμοποιήσαμε το μοντέλο Poisson και το διωνυμικό μοντέλο, ενώ για τον όρο $E(Y)$ που ήταν το αναμενόμενο κόστος των απαιτήσεων για κάθε ασφαλισμένο, χρησιμοποιήσαμε το λογαριθμοκανονικό μοντέλο, το μοντέλο γάμμα, καθώς και ένα πιο γενικό μοντέλο στο οποίο ομαδοποιήσαμε τα δεδομένα μας σε μικρές και μεγάλες απαιτήσεις. Το ερώτημα είναι πιο μοντέλο είναι το καλύτερο και πως φτάνω σε αυτό.

Παρακάτω θα κατασκευάσουμε ένα μοντέλο το οποίο θα αποτελείται μόνο από στατιστικά σημαντικές επεξηγηματικές μεταβλητές για να προβλέψουμε την ετήσια συχνότητα ατυχημάτων. Θα το εξετάσουμε μέσω Poisson και αρνητικής διωνυμικής παλινδρόμησης και θα επιλέξουμε εκείνο με τους χαμηλότερους δείκτες AIC, Null Deviance, και Residual Deviance. Οι επεξηγηματικές μεταβλητές που θα χρησιμοποιήσουμε είναι η ηλικία του οδηγού, η ηλικία του αμαξιού την οποία θα την χωρίσουμε σε διαστήματα, η πυκνότητα των κατοίκων ανά περιοχή, η δύναμη του αυτοκινήτου η οποία θα χωριστεί επίσης σε 3 τύπους (DEF, GH, other) και ο τύπος βενζίνης. Παρακάτω παραθέτουμε τις εντολές με τις οποίες θα γίνουν τα παραπάνω στην R.

```
> CONTRACTS.f$CarAge<-cut(CONTRACTS$CarAge,c(0,15,Inf),include.lowest=TRUE)
> levels(CONTRACTS.f$CarAge)
[1] "[0,15]"   "(15,Inf]"
> CONTRACTS.f$Power<-factor(1*(CONTRACTS.f$Power%in%letters[4:6])+2*(CONTRA
CTS.f$Power%in%letters[7:8]),labels=c("other","DEF","GH"))
> levels(CONTRACTS.f$Power)

[1] "other" "DEF"   "GH"

## μοντέλο Poisson
> freg<-formula(ClaimNb~DriverAge+CarAge+Density+Power+Gas+offset(log(Expos
ure)))
> regp<-glm(freg,data=CONTRACTS.f,family=poisson(link="log"))
> summary(regp)

Call:
glm(formula = freg, family = poisson(link = "log"), data = CONTRACTS.f)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7831  -0.3393  -0.2655  -0.1488   6.5234

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.69700    0.04644 -36.541  < 2e-16 ***
DriverAge(22,26]      -0.62563    0.04609 -13.575  < 2e-16 ***
DriverAge(26,42]      -1.09463    0.03648 -30.005  < 2e-16 ***
DriverAge(42,74]      -1.09399    0.03561 -30.723  < 2e-16 ***
DriverAge(74,Inf]     -1.10953    0.05188 -21.384  < 2e-16 ***
```

```
CarAge(15,Inf]        -0.22701    0.03072  -7.389 1.48e-13 ***
Density(40,200]        0.17985    0.02676   6.720 1.81e-11 ***
Density(200,500]       0.31060    0.02968  10.467  < 2e-16 ***
Density(500,4.5e+03]   0.51489    0.02597  19.823  < 2e-16 ***
Density(4.5e+03,Inf]   0.60570    0.03497  17.321  < 2e-16 ***
PowerDEF              -0.15403    0.02407  -6.399 1.57e-10 ***
PowerGH               -0.13689    0.02633  -5.200 2.00e-07 ***
GasRegular            -0.19784    0.01619 -12.217  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 105613  on 413168  degrees of freedom
Residual deviance: 103891  on 413156  degrees of freedom
AIC: 135174

Number of Fisher Scoring iterations: 6

## Αρνητικό Διωνυμικό Μοντέλο

> regnegative<-glm(freg,family=negative.binomial(theta=1,link="log"),data=C
ONTRACTS.f)
> summary(regnegative)
Call:
glm(formula = freg, family = negative.binomial(1), data = CONTRACTS.f)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7386  -0.3358  -0.2638  -0.1486   5.8713

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -1.67620    0.06249 -26.822  < 2e-16 ***
DriverAge(22,26]      -0.63903    0.06227 -10.262  < 2e-16 ***
DriverAge(26,42]      -1.11289    0.04950 -22.482  < 2e-16 ***
DriverAge(42,74]      -1.11204    0.04837 -22.991  < 2e-16 ***
DriverAge(74,Inf]     -1.13078    0.06971 -16.221  < 2e-16 ***
CarAge(15,Inf]        -0.22525    0.04064  -5.542 2.99e-08 ***
Density(40,200]        0.18039    0.03543   5.091 3.55e-07 ***
Density(200,500]       0.31169    0.03935   7.920 2.38e-15 ***
Density(500,4.5e+03]   0.51652    0.03444  14.996  < 2e-16 ***
Density(4.5e+03,Inf]   0.60572    0.04648  13.032  < 2e-16 ***
PowerDEF              -0.15205    0.03207  -4.741 2.13e-06 ***
PowerGH               -0.13600    0.03506  -3.879 0.000105 ***
GasRegular            -0.19827    0.02155  -9.200  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(1) family taken to be 1.681248)

    Null deviance: 92939  on 413168  degrees of freedom
Residual deviance: 91292  on 413156  degrees of freedom
AIC: 134803

Number of Fisher Scoring iterations: 6
```

Αν είχαμε να επιλέξουμε ανάμεσα στα δύο μοντέλα θα προτιμούσαμε το αρνητικό διωνυμικό μοντέλο καθώς όλοι οι δείκτες του είναι μικρότεροι σε σχέση με το μοντέλο Poisson. Παρατηρούμε επίσης πως και στα δύο μοντέλα όλες οι επεξηγηματικές μεταβλητές είναι στατιστικά σημαντικές (***) και δεν θα μπορούσαμε να απορρίψουμε καμία από το μοντέλο που κατασκευάσαμε. Τέλος θα δούμε την μορφή που παίρνουν τα δύο γενικευμένα γραμμικά μοντέλα.

1. **Μοντέλο Poisson**

$$\log \mu_i = \beta_0 + \beta_{DriverAge}X_1 + \beta_{CarAge}X_2 + \beta_{Density}X_3 + \beta_{Power}X_4 + \beta_{Gas}X_5 + \log E_i.$$

2. **Αρνητικό Διωνυμικό Μοντέλο**

$$\log \mu_i = \beta_{DriverAge}X_1 + \beta_{CarAge}X_2 + \beta_{Density}X_3 + \beta_{Power}X_4 + \beta_{Gas}X_5 + \log E_i.$$

Επομένως η ετήσια συχνότητα ατυχημάτων ενός οδηγού ηλικίας 43 ετών, με αμάξι ηλικίας 5 ετών που καταναλώνει βενζίνη Diesel, ο οποίος ζει σε περιοχή πυκνότητας 500, με δύναμη αυτοκινήτου τύπου GH θα είναι ίση με

$$\mu_i = e^{-1.697-1.09399+0.31060-0.13689+\log 1} = 0.07300112,$$

για το μοντέλο Poisson και

$$\mu_i = e^{-1.67620-1.11204+0.31169-0.13600+\log 1} = 0.07335,$$

για το αρνητικό διωνυμικό μοντέλο.
Για το ποσό της ατομικής ζημιάς θα χρησιμοποιήσουμε το μοντέλο που αναλύσαμε στο κεφάλαιο 3.4.4. Θυμίζουμε ότι οι προβλεπόμενες τιμές για το μοντέλο θα είναι:

$$E(Y) = E(Y|X, Y \leq s)P(Y \leq s|X) + E(Y|X, Y > s)P(Y > s|X),$$

όπου

- $A = E(Y|X, Y \leq s)$ είναι το μέσο κόστος των μικρών απαιτήσεων (απαιτήσεις που δεν ξεπερνούν την τιμή s).
- $B = E(Y|X, Y > s)$ είναι το μέσο κόστος των μεγάλων απαιτήσεων (απαιτήσεις που ξεπερνούν την τιμή s).
- $C = P(Y \leq s|X), P(Y > s|X)$ είναι η πιθανότητα η απαίτηση να είναι μικρή ή μεγάλη αντίστοιχα.

Άρα συνεχίζοντας τον κώδικά μας στην R και αφού έχουμε όλες τις συνιστώσες που χρειαζόμαστε υπολογίζουμε τις προβλεπόμενες τιμές οι οποίες φαίνονται στον Πίνακα 3.2 του Παραρτήματος.

```
> ClaimAMount<-data.frame(age,Predictions=ypA*pred_values$fit+(1-pred_valu
es$fit)*ypB)
```

Επομένως, αν στόχος μας ήταν να υπολογίσουμε το ετήσιο ασφάλιστρο που θα χρεώναμε έναν ασφαλισμένο 43 ετών, με αμάξι ηλικίας 5 ετών που καταναλώνει βενζίνη Diesel, ο οποίος ζει σε περιοχή πυκνότητας 500, με δύναμη αυτοκινήτου τύπου GH το αποτέλεσμα θα ήταν

$$\pi = 0.07300112 * 1691.771 = 123.50117778352\ \nu.\mu.$$

με το μοντέλο Poisson και

$$\pi = 0.07335 * 1691.771 = 124.09140285\ \nu.\mu,$$

με το αρνητικό διωνυμικό μοντέλο.

Συνοψίζοντας, καταλήγουμε στο συμπέρασμα πως σαν ερευνητές δεν μπορούμε εκ των προτέρων να γνωρίζουμε το βέλτιστο μοντέλο για να προσαρμόσουμε τα δεδομένα μας, καθώς δεν υπάρχει βέλτιστο μοντέλο αλλά διαφορετικές στρατηγικές κτισίματος μοντέλων που οδηγούν στο πιο κατάλληλο από όλα τα υποψήφια μοντέλα. Αυτό το οποίο μπορούμε να κάνουμε είναι να βλέπουμε τις διαφορές ανάμεσα σε δείκτες που αναφέραμε (AIC, BIC, Απόκλιση) μεταξύ των προτεινόμενων μοντέλων και να προσπαθούμε να φτιάχνουμε όσο το δυνατόν καταλληλότερα μοντέλα με βάση αυτούς τους δείκτες. Είναι πολύ σημαντική η επιλογή της κατανομής για την εξαρτημένη μεταβλητή, της συνάρτησης σύνδεσης η οποία μετασχηματίζει τον γραμμικό όρο καθώς και των επεξηγηματικών μεταβλητών καθώς με βάση αυτές χτίζεται το μοντέλο με το οποίο θα προβλέπουμε τις τιμές των εκάστοτε εξαρτημένων μεταβλητών. Προφανώς, η εμπειρία του ερευνητή παίζει καθοριστικό ρόλο καθώς μπορεί να χρειαστεί να λάβει υπόψιν του άγνωστους παράγοντες, όπως είναι η αγορά στην οποία απευθύνεται, ο χρόνος στον οποίο τα ιστορικά δεδομένα, τα οποία θα χρησιμοποιήσει για την ανάλυσή του προήλθαν, οι οικονομικές συνθήκες κλπ. Υπάρχουν βέβαια και περιπτώσεις τις οποίες ούτε ο πιο έμπειρος ερευνητής μπορεί να προβλέψει. Ας πάρουμε για παράδειγμα τις πρόσφατες επιπτώσεις λόγω της πανδημίας του COVID – 19 στην ασφάλιση. Μεγάλο μέρος των ασφαλιστικών εσόδων χάθηκε, εξαιτίας της προσωρινής διακοπής των αερολιμένων και των εμπορικών δρομολογίων των πλοίων, καθώς μεγάλο μέρος των εταιριών διέκοψε την ασφάλιση των οχημάτων τους λόγω έλλειψης εσόδων την τρέχουσα χρονική στιγμή. Αν κάποιος ερευνητής χρησιμοποιούσε δεδομένα από την τρέχουσα χρονική περίοδο, ίσως να έπρεπε να τα απορρίψει καθώς δεν αποτελούν αντιπροσωπευτική εικόνα της συνήθους πραγματικότητας.

Σκοπός της παρούσας διπλωματικής ήταν η κατανόηση του απλού γραμμικού μοντέλου και της επέκτασης αυτού που είναι το γενικευμένο, καθώς και η παρουσίαση διαφόρων μοντέλων με σκοπό τον υπολογισμό του ατομικού ασφαλίστρου για ασφαλισμένους με διαφορετικά χαρακτηριστικά. Για τους υπολογισμούς μας χρησιμοποιήσαμε την γλώσσα προγραμματισμού R, η οποία είναι πολύ διαδεδομένη για σκοπούς ανάλυσης δεδομένων και κατασκευής μοντέλων πρόβλεψης. Τα βήματα που ακολουθήσαμε σε κάθε εφαρμογή μπορούν να εφαρμοστούν από τον κάθε αναγνώστη ως μορφή άσκησης και δημιουργίας νέων μοντέλων με βάση τις ατομικές του ανάγκες.

ΠΑΡΑΡΤΗΜΑ

Α. Πίνακες

Ηλικία	*Λογαριθμοκανονικό*	*Γάμμα*	*Ηλικία*	*Λογαριθμοκανονικό*	*Γάμμα*
18	*1626.684*	*2741.878*	*57*	*1761.215*	*1861.741*
19	*1630.002*	*2714.796*	*58*	*1764.807*	*1843.352*
20	*1633.326*	*2687.981*	*59*	*1768.406*	*1825.145*
21	*1636.658*	*2661.431*	*60*	*1772.013*	*1807.117*
22	*1639.996*	*2635.143*	*61*	*1775.627*	*1789.268*
23	*1643.340*	*2609.115*	*62*	*1779.248*	*1771.594*
24	*1646.692*	*2583.344*	*63*	*1782.877*	*1754.096*
25	*1650.050*	*2557.827*	*64*	*1786.513*	*1736.770*
26	*1653.416*	*2532.563*	*65*	*1790.157*	*1719.615*
27	*1656.788*	*2507.548*	*66*	*1793.808*	*1702.630*
28	*1660.167*	*2482.780*	*67*	*1797.466*	*1685.813*
29	*1663.553*	*2458.257*	*68*	*1801.132*	*1669.161*
30	*1666.946*	*2433.976*	*69*	*1804.806*	*1652.675*
31	*1670.345*	*2409.934*	*70*	*1808.486*	*1636.351*
32	*1673.752*	*2386.131*	*71*	*1812.175*	*1620.188*
33	*1677.166*	*2362.562*	*72*	*1815.871*	*1604.185*
34	*1680.586*	*2339.226*	*73*	*1819.574*	*1588.340*
35	*1684.014*	*2316.121*	*74*	*1823.285*	*1572.651*
36	*1687.448*	*2293.244*	*75*	*1827.004*	*1557.117*
37	*1690.890*	*2270.593*	*76*	*1830.730*	*1541.737*
38	*1694.339*	*2248.165*	*77*	*1834.464*	*1526.509*
39	*1697.794*	*2225.959*	*78*	*1838.205*	*1511.431*
40	*1701.257*	*2203.973*	*79*	*1841.954*	*1496.502*
41	*1704.727*	*2182.203*	*80*	*1845.711*	*1481.721*
42	*1708.203*	*2160.649*	*81*	*1849.475*	*1467.085*
43	*1711.687*	*2139.308*	*82*	*1853.247*	*1452.595*
44	*1715.178*	*2118.177*	*83*	*1857.027*	*1438.247*
45	*1718.676*	*2097.255*	*84*	*1860.815*	*1424.041*
46	*1722.182*	*2076.540*	*85*	*1864.610*	*1409.975*
47	*1725.694*	*2056.029*	*86*	*1868.413*	*1396.048*
48	*1729.214*	*2035.721*	*87*	*1872.223*	*1382.259*
49	*1732.740*	*2015.614*	*88*	*1876.042*	*1368.606*
50	*1736.274*	*1995.705*	*89*	*1879.868*	*1355.088*
51	*1739.816*	*1975.992*	*90*	*1883.702*	*1341.703*
52	*1743.364*	*1956.475*	*91*	*1887.544*	*1328.451*
53	*1746.919*	*1937.150*	*94*	*1899.116*	*1289.474*
54	*1750.482*	*1918.016*	*95*	*1902.989*	*1276.737*
55	*1754.052*	*1899.071*	*99*	*1918.562*	*1227.036*
56	*1757.630*	*1880.314*			

Πίνακας 3.1:Προβλεπόμενες τιμές σε κάθε ηλικία για κάθε μοντέλο.

Ηλικία	*Ποσό*	*Ηλικία*	*Ποσό*	*Ηλικία*	*Ποσό*
18	*7619.651*	*47*	*1727.959*	*76*	*2059.506*
19	*6199.263*	*48*	*1742.783*	*77*	*2014.567*
20	*5155.679*	*49*	*1759.725*	*78*	*1965.304*
21	*4378.610*	*50*	*1778.694*	*79*	*1912.554*
22	*3792.560*	*51*	*1799.589*	*80*	*1857.247*
23	*3345.194*	*52*	*1822.296*	*81*	*1800.366*
24	*2999.787*	*53*	*1846.680*	*82*	*1742.903*
25	*2730.263*	*54*	*1872.582*	*83*	*1685.827*
26	*2517.895*	*55*	*1899.811*	*84*	*1630.032*
27	*2349.087*	*56*	*1928.138*	*85*	*1576.313*
28	*2213.864*	*57*	*1957.292*	*86*	*1525.329*
29	*2104.834*	*58*	*1986.960*	*87*	*1477.588*
30	*2016.467*	*59*	*2016.774*	*88*	*1433.433*
31	*1944.588*	*60*	*2046.321*	*89*	*1393.042*
32	*1886.019*	*61*	*2075.137*	*90*	*1356.440*
33	*1838.322*	*62*	*2102.709*	*91*	*1323.511*
34	*1799.611*	*63*	*2128.484*	*92*	*1294.024*
35	*1768.418*	*64*	*2151.877*	*93*	*1267.659*
36	*1743.597*	*65*	*2172.280*	*94*	*1244.034*
37	*1724.243*	*66*	*2189.079*	*95*	*1222.736*
38	*1709.645*	*67*	*2201.669*	*96*	*1203.343*
39	*1699.236*	*68*	*2209.480*	*97*	*1185.446*
40	*1692.567*	*69*	*2211.994*	*98*	*1168.664*
41	*1689.281*	*70*	*2208.769*	*99*	*1152.659*
42	*1689.092*	*71*	*2199.469*	*100*	*1137.137*
43	*1691.771*	*72*	*2183.876*		
44	*1697.135*	*73*	*2161.917*		
45	*1705.034*	*74*	*2133.676*		
46	*1715.344*	*75*	*2099.400*		

Πίνακας 3.2:Προβλεπόμενες τιμές σε κάθε ηλικία για το μοντέλο της Παραγράφου 3.4.5

Β. Κώδικας R

Εγκατάσταση πακέτου με τα δεδομένα που χρησιμοποιούνται στο Κεφάλαιο 3.

```
> install.packages("CASdatasets", repos = "http://dutangc.free.fr/pub/RRepos/", type = "source")
> library(CASdatasets)
> data(freMTPLfreq)
> CONTRACTS<-data.frame(freMTPLfreq)
> names(CONTRACTS)

[1] "PolicyID"  "ClaimNb"   "Exposure"  "Power"
[5] "CarAge"    "DriverAge" "Brand"     "Gas"
[9] "Region"    "Density"


> data(freMTPLsev)
> CLAIMS<-data.frame(freMTPLsev)
> names(CLAIMS)

[1] "PolicyID"    "ClaimAmount"
```

Μεταβλητές τις οποίες χωρίζουμε σε διαστήματα για διευκόλυνση στους υπολογισμούς μας.

```
> CONTRACTS.f<-CONTRACTS
> CONTRACTS.f$DriverAge<-cut(CONTRACTS$DriverAge,c(17,22,26,42,74,Inf))
> CONTRACTS.f$CarAge<-cut(CONTRACTS$CarAge,c(0,1,4,15,Inf))
> CONTRACTS.f$Density<-cut(CONTRACTS$Density,c(0,40,200,500,4500,Inf,include.lowest=TRUE))
```

Μέση τιμή και Δειγματική Διακύμανση της τυχαίας μεταβλητής N για την παλινδρόμηση Poisson.

```
> vY<-CONTRACTS.f$ClaimNb
> vE<-CONTRACTS.f$Exposure
> m<-sum(vY)/sum(vE)
> v<-sum((vY-m*vE)^2)/sum(vE)
> cat("average =",m,"variance =",v)

average = 0.06979859 variance = 0.07396742
```

Σχέση μεταξύ παρατηρούμενης ετήσιας συχνότητας και της κάθε μίας επεξηγηματικής μεταβλητής ξεχωριστά. (Power, CarAge). Κατά τον ίδιο τρόπο μελετάμε και τις υπόλοιπες.

```
> power<- sqldf("Select Power,sum(Exposure) as Exposure,sum(ClaimNb) as Number_of_Claims,sum(ClaimNb)/sum(Exposure) as Annualized_Freq from CONTRACTS group by Power");power

Power   Exposure    Number_of_Claims Annualized_Freq
  d    37820.9222          2359        0.06237288
  e    44635.3284          3201        0.07171449
  f    55884.9546          3997        0.07152193
  g    51657.1308          3464        0.06705754
  h    13920.7818          1000        0.07183505
  i     9412.4368           722        0.07670702
  j     9279.9139           710        0.07650933
  k     4700.5367           380        0.08084183
  l     2200.6932           162        0.07361317
  m      975.0504            76        0.07794469
```

```
  n       661.2747            56        0.08468492
  o       675.1442            54        0.07998292
```

```
> carage<- sqldf("Select CarAge,sum(Exposure) as Exposure,sum(ClaimNb) as N
umber_of_Claims,sum(ClaimNb)/sum(Exposure) as Annualized_Freq from CONTRACT
S group by CarAge");carage
```

CarAge	Exposure	Number_of_Claims	Annualized_Freq
0	8783.562517	620	0.070586393
1	18163.628724	1314	0.072342373
2	17377.688680	1234	0.071010594
3	15836.778999	1101	0.069521713
4	14983.823589	1086	0.072478162
5	14459.124920	1020	0.070543688
6	13804.441612	1033	0.074830988
7	12923.110511	1004	0.077690274
8	13095.985553	943	0.072006799
9	12689.403084	987	0.077781436
10	13835.462630	1058	0.076470157
11	12034.917540	890	0.073951483
12	11946.280868	874	0.073160845
13	11116.345446	699	0.062880378
14	9965.393580	639	0.064121903
15	8777.941396	513	0.058441949
16	6277.666179	380	0.060532050
17	4675.312136	279	0.059675160
18	3483.719160	173	0.049659571
19	2308.435369	99	0.042886191
20	1478.996867	87	0.058823654
21	963.485930	42	0.043591711
22	626.359278	25	0.039913195
23	412.220951	16	0.038814136
24	303.003699	15	0.049504346
25	219.079156	14	0.063903843
26	160.313684	11	0.068615478
27	150.308212	2	0.013305993
28	141.018219	4	0.028365129
29	133.394628	1	0.007496554
30	124.990000	4	0.032002560
31	101.260000	1	0.009875568
32	71.790000	2	0.027859033
33	62.560000	2	0.031969309
34	53.808212	0	0.000000000
35	46.730000	1	0.021399529
36	37.498197	3	0.080003847
37	24.090000	0	0.000000000
38	17.940000	0	0.000000000
39	13.920000	2	0.143678161
40	7.390000	0	0.000000000
41	6.000000	0	0.000000000
42	6.870000	0	0.000000000
43	7.830000	0	0.000000000
44	9.940000	0	0.000000000
45	10.650000	0	0.000000000
46	12.880000	0	0.000000000
47	13.160000	0	0.000000000
48	9.350000	0	0.000000000
49	5.920000	0	0.000000000
50	4.512732	0	0.000000000
51	4.000000	0	0.000000000
52	3.410000	0	0.000000000
53	0.630000	0	0.000000000
54	1.220000	0	0.000000000
55	0.160000	0	0.000000000
59	1.150000	0	0.000000000
60	0.100000	0	0.000000000
62	0.750000	0	0.000000000

```
 63          1.000000          0              0.000000000
 64          0.700000          0              0.000000000
 65          0.990000          0              0.000000000
 66          0.990000          0              0.000000000
 68          0.040000          0              0.000000000
 69          0.510000          1              1.960784314
 70          0.300000          0              0.000000000
 71          0.220000          0              0.000000000
 78          1.000000          0              0.000000000
 79          1.000000          0              0.000000000
 80          1.000000          0              0.000000000
 81          1.000000          0              0.000000000
 82          1.000000          0              0.000000000
 83          0.540000          0              0.000000000
 84          1.000000          1              1.000000000
 85          0.970000          0              0.000000000
 99         11.005464          1              0.090863952
100         13.190000          0              0.000000000
```

Παράδειγμα παλινδρόμησης Poisson με μεταβλητή απόκρισης την μεταβλητή ClaimNB και μ ε επεξηγηματικές μεταβλητές τις Gas, DriverAge και Density.

```
> reg=glm(ClaimNb~Gas+DriverAge+Density+offset(log(Exposure)),data=CONTRACT
S.f,family=poisson(link="log"))
> summary(reg)


Call:
glm(formula = ClaimNb ~ Gas + DriverAge + Density + offset(log(Exposure)),
    family = poisson(link = "log"), data = CONTRACTS.f)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7655  -0.3385  -0.2669  -0.1488   6.5202

Coefficients:
                      Estimate Std.Error z value Pr(>|z|)
(Intercept)           -1.86471  0.04047 -46.079  < 2e-16 ***
GasRegular            -0.20598  0.01603 -12.846  < 2e-16 ***
DriverAge(22,26]      -0.61606  0.04608 -13.370  < 2e-16 ***
DriverAge(26,42]      -1.07967  0.03640 -29.657  < 2e-16 ***
DriverAge(42,74]      -1.07765  0.03549 -30.362  < 2e-16 ***
DriverAge(74,Inf]     -1.10706  0.05188 -21.338  < 2e-16 ***
Density(40,200]        0.18473  0.02675   6.905 5.02e-12 ***
Density(200,500]       0.31822  0.02966  10.730  < 2e-16 ***
Density(500,4.5e+03]   0.52694  0.02593  20.320  < 2e-16 ***
Density(4.5e+03,Inf]   0.63717  0.03482  18.300  < 2e-16 ***

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

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 105613  on 413168  degrees of freedom
Residual deviance: 103986  on 413159  degrees of freedom
AIC: 135263

Number of Fisher Scoring iterations: 6
```

Αριθμός των απαιτήσεων ανάλογα το είδος της βενζίνης, συνολική έκθεση στον κίνδυνο, ετήσια συχνότητα των απαιτήσεων ανάλογα το είδος της βενζίνης.

```
> vY<- CONTRACTS.f$ClaimNb
> vE<- CONTRACTS.f$Exposure
> x1<- CONTRACTS.f$Gas
> tapply(vY, x1, sum)

Diesel    Regular
 8446      7735

> tapply(vE,x1,sum)

Diesel      Regular
113104.8    118719.4

> tapply(vY, x1, sum)/tapply(vE,x1,sum)

Diesel         Regular
0.07467412     0.06515364
```

Παλινδρόμηση Poisson με το είδος της βενζίνης ως επεξηγηματική μεταβλητή, χωρίς τον σταθερό όρο με χρήση της λογαριθμικής συνάρτηση σύνδεσης.

```
> regpoislog<- glm(vY~0+x1+offset(log(vE)),data=df,family=poisson(link="log
"))
> summary(regpoislog)

Call:
glm(formula = vY ~ 0 + x1 + offset(log(vE)), family = poisson(link = "log")
,
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5092  -0.3610  -0.2653  -0.1488   6.5858

Coefficients:
          Estimate Std. Error z value Pr(>|z|)
x1Diesel  -2.59462    0.01088  -238.5   <2e-16 ***
x1Regular -2.73101    0.01137  -240.2   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 450747  on 413169  degrees of freedom
Residual deviance: 105537  on 413167  degrees of freedom
AIC: 136799

Number of Fisher Scoring iterations: 6
```

Παλινδρόμηση Poisson με το είδος της βενζίνης ως επεξηγηματική μεταβλητή, με τον σταθερό όρο με χρήση της λογαριθμικής συνάρτηση σύνδεσης.

```
> regpoislog1 <- glm(vY~x1+offset(log(vE)),data=df,family=poisson(link="log
"))
> summary(regpoislog1)

Call:
glm(formula = vY ~ x1 + offset(log(vE)), family = poisson(link = "log"),
```

```
    data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5092  -0.3610  -0.2653  -0.1488   6.5858

Coefficients:
            Estimate Std. Error  z value Pr(>|z|)
(Intercept) -2.59462    0.01088 -238.454   <2e-16 ***
x1Regular   -0.13639    0.01574   -8.666   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 105613  on 413168  degrees of freedom
Residual deviance: 105537  on 413167  degrees of freedom
AIC: 136799

Number of Fisher Scoring iterations: 6
```

Ηλικία του οδηγού ως διάστημα

```
> AgeD<-cut(CONTRACTS$DriverAge,c(17,22,26,42,74,Inf));AgeD

Levels: (17,22] (22,26] (26,42] (42,74] (74,Inf]
```

Γενικευμένο γραμμικό μοντέλο Poisson με την ηλικία του οδηγού ως διάστημα με χρήση της λογαριθμικής συνάρτησης σύνδεσης.

```
> Agedmodel<-glm(ClaimNb~AgeD+offset(log(Exposure)),family=poisson(link="lo
g"),data=CONTRACTS.f)
> summary(Agedmodel)

Call:
glm(formula = ClaimNb ~ AgeD + offset(log(Exposure)), family = poisson,
    data = CONTRACTS.f)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6218  -0.3615  -0.2632  -0.1491   6.5690

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.66337    0.03365  -49.43   <2e-16 ***
AgeD(22,26]  -0.56935    0.04602  -12.37   <2e-16 ***
AgeD(26,42]  -1.04009    0.03628  -28.67   <2e-16 ***
AgeD(42,74]  -1.06454    0.03542  -30.05   <2e-16 ***
AgeD(74,Inf] -1.17659    0.05177  -22.73   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 105613  on 413168  degrees of freedom
Residual deviance: 104734  on 413164  degrees of freedom
AIC: 136001

Number of Fisher Scoring iterations: 6
```

Γενικευμένο γραμμικό μοντέλο Poisson με την ηλικία του οδηγού ως επεξηγηματική μεταβλητή με χρήση της λογαριθμικής συνάρτησης σύνδεσης. (η ηλικία του οδηγού δεν είναι χωρισμένη σε διαστήματα).

```
> AgedDriver3<-glm(ClaimNb~DriverAge+offset(log(Exposure)),family=poisson,d
ata=CONTRACTS)
> summary(AgedDriver3)

Call:
glm(formula = ClaimNb ~ DriverAge + offset(log(Exposure)), family = poisson
,
    data = CONTRACTS)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5523  -0.3510  -0.2678  -0.1504   6.4415

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.1513378  0.0262347  -82.00   <2e-16 ***
DriverAge   -0.0111060  0.0005579  -19.91   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 105613  on 413168  degrees of freedom
Residual deviance: 105206  on 413167  degrees of freedom
AIC: 136467

Number of Fisher Scoring iterations: 6
```

Γενικευμένο αρνητικό διωνυμικό μοντέλο με άγνωστη τιμή για το α. (συντελεστή μεταβλητότητας)

```
> freg<-formula(ClaimNb~DriverAge+CarAge+Density+Brand+Power+Gas+offset(log
(Exposure)))
> regNB2<-glm.nb(freg,data=CONTRACTS.f)
> summary(regNB2)

Call:
glm.nb(formula = freg, data = CONTRACTS.f, init.theta = 0.8790117269,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7570  -0.3377  -0.2697  -0.1577   5.9407

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -1.859424 0.068359 -27.201  < 2e-16 ***
DriverAge(22,26]                 -0.625005 0.049019 -12.750  < 2e-16 ***
DriverAge(26,42]                 -1.106626 0.038988 -28.384  < 2e-16 ***
DriverAge(42,74]                 -1.096200 0.038195 -28.700  < 2e-16 ***
DriverAge(74,Inf]                -1.121930 0.055027 -20.389  < 2e-16 ***
CarAge(1,4]                      -0.008798 0.033934  -0.259 0.795434
CarAge(4,15]                     -0.006337 0.032477  -0.195 0.845303
CarAge(15,Inf]                   -0.243862 0.043501  -5.606 2.07e-08 ***
Density(40,200]                   0.182414 0.027813   6.559 5.43e-11 ***
Density(200,500]                  0.313545 0.030959  10.128  < 2e-16 ***
Density(500,4.5e+03]              0.530294 0.027253  19.458  < 2e-16 ***
Density(4.5e+03,Inf]              0.668306 0.038110  17.536  < 2e-16 ***
BrandJapanese (except Nissan) o -0.198113 0.048045  -4.123 3.73e-05 ***
```

```
BrandMercedes, Chrysler or BMW  -0.003178 0.056001  -0.057 0.954741
BrandOpel, General Motors or Fo  0.049850 0.046911   1.063 0.287941
Brandother                      -0.061290 0.065344  -0.938 0.348263
BrandRenault, Nissan or Citroen -0.072295 0.040995  -1.764 0.077813 .
BrandVolkswagen, Audi, Skoda or  0.009161 0.048229   0.190 0.849350
Powere                           0.074481 0.029421   2.532 0.011357 *
Powerf                           0.101107 0.028727   3.520 0.000432 ***
Powerg                           0.083786 0.028584   2.931 0.003377 **
Powerh                           0.107364 0.041570   2.583 0.009802 **
Poweri                           0.215363 0.045449   4.739 2.15e-06 ***
Powerj                           0.203057 0.046386   4.378 1.20e-05 ***
Powerk                           0.293959 0.059832   4.913 8.97e-07 ***
Powerl                           0.176923 0.089053   1.987 0.046955 *
Powerm                           0.232771 0.126194   1.845 0.065103 .
Powern                           0.280968 0.143869   1.953 0.050826 .
Powero                           0.220574 0.155418   1.419 0.155831
GasRegular                      -0.170893 0.018112  -9.435  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(0.879) family taken to be 1)

    Null deviance: 87697  on 383119  degrees of freedom
Residual deviance: 86006  on 383090  degrees of freedom
  (30049 observations deleted due to missingness)
AIC: 129131

Number of Fisher Scoring iterations: 1
```

Πολλαπλό λογαριθμοκανονικό μοντέλο παλινδρόμησης με επεξηγηματικές μεταβλητές να είναι η ηλικία του αμαξιού χωρισμένη σε διάστημα των (0,15], (15, Inf) και το είδος της βενζίνης για απαιτήσεις <15000.

```
> lognormal<-lm(log(ClaimAmount)~CarAge+Gas,data=claims.f[claims$ClaimAmoun
t<15000,])
> summary(lognormal)

Call:
lm(formula = log(ClaimAmount) ~ CarAge + Gas, data = claims.f[claims$ClaimA
mount < 15000, ])

Residuals:
    Min      1Q  Median      3Q     Max
-6.1043 -0.2612  0.2646  0.3367  2.8428

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.79747    0.01154 589.062   <2e-16 ***
CarAge(15,Inf]  0.01670    0.03204   0.521    0.602
GasRegular     -0.02591    0.01660  -1.561    0.119
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.045 on 16003 degrees of freedom
Multiple R-squared:  0.0001614,       Adjusted R-squared:  3.643e-05
F-statistic: 1.292 on 2 and 16003 DF,  p-value: 0.2749
```

Οι συντελεστές του λογαριθμικού μοντέλου.

```
> exp(coefficients(lognormal))

(Intercept)     CarAge(15,Inf]     GasRegular
 895.5777014      1.0168369         0.9744271
```

Γενικευμένο γραμμικό μοντέλο γάμμα με επεξηγηματικές μεταβλητές να είναι η ηλικία του αμαξιού χωρισμένη σε διάστημα των (0,15], (15, Inf) και το είδος της βενζίνης για απαιτήσεις < 15000.

```
> gamma<-glm(ClaimAmount~CarAge+Gas,family=Gamma(link="log"),data=claims.f[
claims$ClaimAmount<15000,])
> summary(gamma)

Call:
glm(formula = ClaimAmount ~ CarAge + Gas, family = Gamma(link = "log"),
    data = claims.f[claims$ClaimAmount < 15000, ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3252  -0.6126  -0.1515  -0.0788   3.9639

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
 (Intercept)     7.22023    0.01204 599.783   <2e-16 ***
CarAge(15,Inf) -0.04508    0.03343  -1.349    0.178
GasRegular     -0.02203    0.01732  -1.272    0.203
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 1.187882)

    Null deviance: 13455  on 16005  degrees of freedom
Residual deviance: 13450  on 16003  degrees of freedom
AIC: 262086

Number of Fisher Scoring iterations: 5
```

Οι συντελεστές του γάμμα μοντέλου.

```
> exp(coefficients(gamma))

(Intercept)      CarAge(15,Inf) GasRegular
1366.8078397      0.9559250      0.9782092
```

Πολλαπλό λογαριθμοκανονικό μοντέλο παλινδρόμησης με επεξηγηματικές μεταβλητές να είναι η ηλικία του οδηγού χωρίς περιορισμό των απαιτήσεων.

```
> logn.reg<-lm(log(ClaimAmount)~DriverAge,data=claims)
> summary(logn.reg)

Call:
lm(formula = log(ClaimAmount) ~ DriverAge, data = claims)

Residuals:
    Min      1Q  Median      3Q     Max
-6.1612 -0.2802  0.2204  0.3081  7.7520

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.7361807  0.0277458 242.782  < 2e-16 ***
DriverAge   0.0020374  0.0005868   3.472 0.000518 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.115 on 16179 degrees of freedom
Multiple R-squared:  0.0007446,       Adjusted R-squared:  0.0006828
```

```
F-statistic: 12.06 on 1 and 16179 DF,  p-value: 0.0005177
```

Οι συντελεστές του λογαριθμικού μοντέλου.

```
exp(coefficients(logn.reg))
(Intercept)   DriverAge
  842.33743     1.00204
```

Γενικευμένο μοντέλο παλινδρόμησης γάμμα με επεξηγηματικές μεταβλητές να είναι η ηλικία του οδηγού χωρίς περιορισμό των απαιτήσεων.

```
> gamma.reg<-glm(ClaimAmount~DriverAge,family=Gamma(link="log"),data=claims
)
> summary(gamma.reg)

Call:
glm(formula = ClaimAmount ~ DriverAge, family = Gamma(link = "log"),
    data = claims)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-3.520  -0.926  -0.565  -0.365  38.540

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.095074   0.203485  39.782   <2e-16 ***
DriverAge   -0.009926   0.004304  -2.307   0.0211 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 66.85011)

    Null deviance: 27064  on 16180  degrees of freedom
Residual deviance: 26607  on 16179  degrees of freedom
AIC: 279095

Number of Fisher Scoring iterations: 9
```

Οι συντελεστές του μοντέλου γάμμα.

```
> exp(coefficients(gamma.reg))
 (Intercept)    DriverAge
3278.2803782    0.9901227
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

Beirlant, J. & Teygels, J. (1992). Modelling large claims in non-life insurance. *Insurance: Mathematics and Economics,* **11**, 17-29

Boland, P.J. (2007). *Statistical and Probabilistic Methods in Actuarial Science,* Chapman & Hall / CRC. London.

Charnes, A., Frome, E.L. and Yu, P.L. (1976). The equivalence of generalized least squares and maximum likelihood estimates in the exponential family. *Journal of the American Statistical Association, **71,*** 169-71

Charpentier, A. (2014). *Computational Actuarial Science with R,* Chapman & Hall / CRC. London.

Dobson, A.J. (2002). *An Introduction to Generalized Linear Models.* Second Edition, Chapman & Hall / CRC. London.

Dunn, P.K & Smyth, G.K. (2018). *Generalized Linear Models with Examples in R.* Springer Texts in Satistics. Rencher.

Fisher, R.A (1912). On the absolute criterion for fitting frequency curves. *Messenger of Mathematics,* **41**, 155-160

Fisher, R.A (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society,* **222**, 309-368.

Fisher, R.A (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society,* **22**, 700-725.

Fisher, R.A (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society A,* **144**, 285-307.

Hilbe, J. (2011). *Negative Binomial Regression,* Wiley, New York.

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models.* Second Edition, Chapman & Hall / CRC. London.

Montgomery, D.C., Peck, E.A. & Vining, G.G. (1982), *Introduction to Linear Regression Analysis Fifth Edition*, John Wiley & Sons. New York.

Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, Series A,* **135**, 370-384.

Raymond H.M., Montgomery D.C., Vining G.G., Robinson T.J. (2010). *Generalized Linear Models With Applications in Engineering and the Sciences*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Rencher, A.C & Schaalje, G.B. (2007). *Linear Models in Statistics.* Second Edition, John Wiley Montgomery & Sons. New York.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics,* **6**, 461-464.

Wedderburn, R.W.M. (1974). Quasi - Likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrica*, **61**, 439-447.