ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

«ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΜΕ ΕΦΑΡΜΟΓΕΣ ΣΤΟ ΣΤΑΤΙΣΤΙΚΟ ΠΑΚΕΤΟ R»

ΠΑΝΑΓΙΩΤΟΠΟΥΛΟΥ ΜΑΡΙΑ – ΕΛΕΥΘΕΡΙΑ

Τριμελής Επιτροπή: **Φουσκάκης Δημήτρης (Επιβλέπων Καθηγητής)**

Παπανικολάου Βασίλης

Σπηλιώτης Ιωάννης

Αθήνα, Δεκέμβριος 2013

ΕΥΧΑΡΙΣΤΙΕΣ

Αρχικά, θα ήθελα να ευχαριστήσω θερμά τον Επίκουρο Καθηγητή του Ε.Μ.Π., κ. Φουσκάκη Δημήτρη για την πολύτιμη βοήθεια και καθοδήγησή του καθ' όλη τη διάρκεια της εκπόνησης της διπλωματικής μου εργασίας.

Επιπλέον, δεν θα μπορούσα να παραλείψω τους γονείς και την αδερφή μου, καθώς και τους πολύ καλούς μου φίλους για την στήριξη και την βοήθειά τους κατά τη διάρκεια της εργασίας μου.

Μαριέλλα Παναγιωτοπούλου

Αθήνα, Δεκέμβριος 2013

ΠΕΡΙΛΗΨΗ

Σε πολλές επιστήμες είναι επιτακτική η ανάγκη δημιουργίας στατιστικών μοντέλων, δηλαδή μοντέλων που να αποδίδουν και να περιγράφουν τη σχέση εξάρτησης μεταξύ μεταβλητών. Φυσικά, αναφερόμαστε σε μια σχέση στοχαστική, δηλαδή υπάρχει ένα ποσοστό αβεβαιότητας το οποίο ενσωματώνεται στη σχέση μεταξύ των μεταβλητών με την έννοια των «τυχαίων σφαλμάτων». Για το σκοπό αυτό έχουν αναπτυχθεί τα μοντέλα παλινδρόμησης, τα οποία αποτελούν βασικό εργαλείο για την ανάλυση δεδομένων, που προκύπτουν από την μελέτη στοχαστικών φαινομένων.

Στην παρούσα διπλωματική αναλύεται η θεωρία των γενικευμένων γραμμικών μοντέλων παλινδρόμησης και γίνεται εφαρμογή σε πραγματικά δεδομένα με τη βοήθεια του στατιστικού πακέτου της R.

Στο πρώτο κεφάλαιο, γίνεται μία εισαγωγή στα γενικευμένα γραμμικά μοντέλα και παρουσιάζεται η δομή ενός γενικευμένου γραμμικού μοντέλου, καθώς και κάποια εισαγωγικά στοιχεία για την Εκθετική Οικογένεια Κατανομών. Επιπλέον, παρουσιάζεται η διαδικασία προσαρμογής του μοντέλου με τη μέθοδο μέγιστης πιθανοφάνειας και με την Quasi-πιθανοφάνεια και αναπτύσσονται οι απαιτούμενοι έλεγχοι ισχύος προϋποθέσεων στα εν λόγω μοντέλα. Εν κατακλείδι, παρουσιάζονται κριτήρια επιλογής και καταλληλότητας μοντέλων, καθώς και διαγνωστικές μέθοδοι.

Στο δεύτερο κεφάλαιο, αναλύονται κάποιες κατηγορίες μοντέλων για δίτιμα ή διωνυμικά δεδομένα. Ειδικότερα, παρουσιάζεται εκτενώς το μοντέλο της λογιστικής παλινδρόμησης και ακολούθως το μοντέλο probit και το μοντέλο complementary log-log.

Στο τρίτο κεφάλαιο, παρουσιάζεται η δομή και η θεωρία μοντέλων τύπου λογιστικής παλινδρόμησης για κατηγορικές μεταβλητές απόκρισης με περισσότερες από δύο κατηγορίες. Αρχικά, παρουσιάζεται το πολυωνυμικό μοντέλο για κατηγορικές μεταβλητές που δεν υποδηλώνουν κάποια διάταξη και στη συνέχεια αναπτύσσονται κάποια μοντέλα για μεταβλητές διάταξης. Αυτά είναι το λογιστικό μοντέλο των διαδοχικών κατηγοριών, το λογιστικό μοντέλο των συνεχιζόμενων λόγων και το μοντέλο των αναλογικών συμπληρωματικών πιθανοτήτων.

Στο τέταρτο καφάλαιο, παρουσιάζονται κατάλληλα μοντέλα για την περίπτωση των διακριτών δεδομένων. Δίνεται έμφαση στο μοντέλο της παλινδρόμησης Poisson και γίνεται αναφορά στο μοντέλο της αρνητικής διωνυμικής κατανομής.

Στο πέμπτο κεφάλαιο, γίνεται μία εισαγωγή στο στατιστικό πακέτο R και αναλύονται οι εντολές και οι συναρτήσεις της R, που χρησιμοποιούνται για την ανάλυση των γενικευμένων γραμμικών μοντέλων. Στο έκτο κεφάλαιο, με τη βοήθεια της R, παρουσιάζονται κάποιες εφαρμογές μοντέλων παλινδρόμησης με πραγματικά δεδομένα. Συγκεκριμένα, αναλύεται μία εφαρμογή στην παλινδρόμηση Poisson και γίνεται μία σύγκριση με το μοντέλο της αρνητικής διωνυμικής κατανομής και μία εφαρμογή με δίτιμη μεταβλητή απόκρισης.

ABSTRACT

At a wide variety of scientific disciplines, it is imperative to create statistical models, i.e. models to deliver and describe the relationship of dependence between variables. As a result, we refer to a stochastic relationship, i.e. there is a percentage of uncertainty which is incorporated into the relationship between the variables with the meaning of "random errors". For this reason, regression models have been developed, as an essential tool for the analysis of data, resulting from the study of stochastic phenomena.

In this dissertation, the theory of generalized linear regression models is analyzed and applications are provided using real data with the statistical package R.

The first chapter contains an introduction of generalized linear models and its basic structure, as well as some introductory information for the Exponential Family of Distributions. In addition, the goodness of model fit is access using the maximum likelihood methods and the Quasi-likelihood methods and the necessary statistical hypothesis tests are analyzed. In conclusion, diagnostic methods are presented.

The second chapter deals with models for binary or binomial data. In particular, the logistic regression model and furthermore the probit model and complementary log-log model are extensively presented.

The third chapter, discusses the structure and the theory of several types of logistic regression for multi-categorical response variables. Initially, the multinomial model for nominal responses is presented and then some models for ordinal responses are demonstrated. These are the adjacent categories logit model, the continuation ratio logit model and the proportional odds model.

In the fourth chapter, appropriate models in the case of discrete data are discussed. Emphasis is given for the Poisson regression model and the Negative Binomial model.

The fifth chapter provides some applications of generalized linear models to the statistical package R and analyze the commands and functions in R, used for the analysis of these models. The sixth chapter provides applications to the models discussed earlier using real data sets and R. In particular, using count response data both the Poisson and the Negative Binomial models are fitted and their fit is been compared. Furthermore, an application to binary response data using different link functions is discussed and a comparison of these different models is made.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΕΦΑΛΑΙΟ 1

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ (GLM)

1.1 Εισαγωγή

Πολλές φορές στην καθημερινότητά μας, αλλά και σε πάρα πολλούς τομείς της επιστήμης και της τεχνολογίας καλούμαστε να εξετάσουμε ένα χαρακτηριστικό ενός συνόλου, το οποίο στα μαθηματικά καλούμε μεταβλητή. Φυσικά, τις περισσότερες φορές το χαρακτηριστικό αυτό επηρεάζεται από διάφορους παράγοντες, δηλαδή η τιμή της προς εξέταση μεταβλητής σχετίζεται με την τιμή κάποιων άλλων μεταβλητών. Η σχέση μεταξύ δύο ή περισσότερων μεταβλητών καλείται μοντέλο και αποτελεί μια πιο απλή αναπαράσταση της πραγματικότητας.

Τα μοντέλα διακρίνονται σε ντετερμινιστικά και στοχαστικά. Τα ντετερμινιστικά είναι εκείνα, των οποίων τα αποτελέσματα είναι γνωστά από πριν, ενώ στα στοχαστικά μοντέλα υπεισέρχεται μια σχέση αβεβαιότητας. Τα μοντέλα, που περιλαμβάνουν ένα στοχαστικό μέρος καλούνται στατιστικά μοντέλα.

Η σχέση μεταξύ δύο τυχαίων μεταβλητών X και Y στην πιο απλή μορφή της μπορεί να εκφραστεί μέσω ενός μοντέλου παλινδρόμησης, το οποίο καλείται απλό γραμμικό μοντέλο και δίνεται από τη σχέση:

$$\mathrm{Y} = \beta_0 + \beta_1 x + \varepsilon\,, \qquad (1.1)$$

όπου x είναι η παρατηρούμενη τιμή της επεξηγηματικής τ.μ. X και ε το τυχαίο σφάλμα με μέση τιμή $\mathrm{E}(\varepsilon)=0$. Η τυχαία μεταβλητή Y καλείται εξαρτημένη ή αποκριτική μεταβλητή (*dependent or response variable*) και η μεταβλητή X καλείται ανεξάρτητη ή επεξηγηματική μεταβλητή (*independent or predictor*).

Το παραπάνω μοντέλο εκφράζει τη σχέση της μεταβλητής απόκρισης Y με μία επεξηγηματική μεταβλητή X. Γενικεύοντας το παραπάνω μοντέλο με χρήση περισσότερων επεξηγηματικών μεταβλητών $\mathrm{X}_1, \mathrm{X}_2, ..., \mathrm{X}_k$ έχουμε το πολλαπλό γραμμικό μοντέλο που εκφράζεται από τη σχέση

$$\mathrm{Y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon\,, \quad (1.2)$$

όπου $x_1, x_2, ..., x_k$ οι παρατηρούμενες τιμές των επεξηγηματικών μεταβλητών και ε το τυχαίο σφάλμα με μέση τιμή $\mathrm{E}(\varepsilon) = 0$.
Έστω ότι διαθέτουμε τυχαίο δείγμα $(\mathrm{Y}_i, \mathrm{X}_{i1}, ..., \mathrm{X}_{ik})$, $i = 1, .., n$ και έστω $(y_i, x_{i1}, ..., x_{ik})$ οι αντίστοιχες παρατηρούμενες τιμές (δεδομένα).

Το γενικό γραμμικό μοντέλο δίνεται από τη σχέση

$$\mathrm{Y_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_\kappa x_{ik} + \varepsilon_i \text{, (1.3)}$$

όπου

- Y_i, $i = 1, 2, \ldots, n$ είναι οι ανεξάρτητες και ισόνομες μεταβλητές απόκρισης.
- x_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, k$ οι παρατηρούμενες τιμές των ανεξάρτητων ή επεξηγηματικών μεταβλητών X_j του μοντέλου για την i – οστή παρατήρηση.
- $\beta_0, \beta_1, \ldots, \beta_k$ οι άγνωστες παράμετροι ή συντελεστές του μοντέλου, τις οποίες θέλουμε να εκτιμήσουμε από τα δεδομένα μας και
- ε_i, $i = 1, 2, \cdots, n$ τα ανεξάρτητα και ισόνομα τυχαία σφάλματα με μηδενική μέση τιμή.

Χρησιμοποιώντας τα δεδομένα που διαθέτουμε η σχέση (1.3) υπό μορφή πινάκων μπορεί να γραφεί

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \text{, (1.4)}$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

όπου $\mathbf{X}$ είναι ένας $n \times p$ πίνακας $(p = k + 1)$, ο οποίος καλείται πίνακας σχεδιασμού, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^{\mathrm{T}}$ το $p \times 1$ διάνυσμα των άγνωστων παραμέτρων και $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^{\mathrm{T}}$ το $n \times 1$ διάνυσμα των τυχαίων σφαλμάτων.

Τα τυχαία σφάλματα είναι πολύ βασικό κομμάτι για το γενικό γραμμικό μοντέλο και θα πρέπει να πληρούν τις παρακάτω προϋποθέσεις:

- $\mathrm{E}(\varepsilon_i) = 0$, για κάθε i.
- $V(\varepsilon_i) = \sigma^2$, για κάθε i (υπόθεση ομοσκεδαστικότητας).
- $Cov(\varepsilon_i, \varepsilon_j) = 0$, για $i \neq j$, δηλαδή τα τυχαία σφάλματα είναι ασυσχέτιστα.
- το $n-$διάστατο διάνυσμα $\boldsymbol{\varepsilon}$ ακολουθεί την πολυδιάστατη Κανονική κατανομή, δηλαδή $\boldsymbol{\varepsilon} \sim \mathrm{N_n}(\mathbf{0}, \boldsymbol{\sigma}^2 I)$.

Από τα παραπάνω η αναμενόμενη τιμή της μεταβλητής απόκρισης δίνεται από τη σχέση

$$\mathrm{E}\left(\mathrm{Y} \mid \mathbf{X} = \boldsymbol{x}\right) = \mathbf{X}\boldsymbol{\beta}\,. \quad (1.5)$$

Η σχέση (1.5) καλείται κανονικό γραμμικό μοντέλο, καθώς η αναμενόμενη μέση τιμή της τυχαίας μεταβλητής Y συνδέεται γραμμικά με τις επεξηγηματικές μεταβλητές του μοντέλου. Πολλές φορές το $\mathbf{X}\boldsymbol{\beta}$ ονομάζεται συστηματικό ή μη-στοχαστικό μέρος του μοντέλου (το στοχαστικό μέρος του μοντέλου είναι η υπόθεση ότι $\mathrm{Y}_i \mid \boldsymbol{x} \sim \mathrm{N}\left(\mu_i, \sigma^2\right)$).

Η Στατιστική ανάλυση, που αναπτύσσεται γύρω από το γραμμικό μοντέλο βασίζεται στο ότι πληρούνται οι προϋποθέσεις για το μοντέλο αυτό, δηλαδή ότι τα σφάλματα είναι ασυσχέτιστες τυχαίες μεταβλητές και ακολουθούν την κανονική κατανομή. Φυσικά, άμεσα προκύπτει ότι και η μεταβλητή $\mathrm{Y} \mid \boldsymbol{x}$ θα ακολουθεί την Κανονική κατανομή. Στην πραγματικότητα, δεν είναι πάντα εφικτό να πληρούνται οι προϋποθέσεις του μοντέλου. Για παράδειγμα, μπορεί η εξαρτημένη μεταβλητή να είναι μια διακριτή μεταβλητή, και πιο συγκεκριμένα ποσοτική, όπως ο αριθμός τροχαίων ατυχημάτων, ή μια δίτιμη μεταβλητή, όπως η εμφάνιση ή μη ενός τύπου καρκίνου σε έναν ασθενή κ.λ.π. Επιπλέον, υπάρχουν περιπτώσεις, όπου η μεταβλητή απόκρισης είναι συνεχής, αλλά είναι λογικό να θεωρήσουμε ότι ακολουθεί την Κανονική κατανομή, όπως για παράδειγμα η διάρκεια ζωής ενός εξαρτήματος κ.λ.π. Σ' αυτές τις περιπτώσεις υπάρχει η ανάγκη ανάπτυξης μοντέλων τύπου παλινδρόμησης, αφού οι μεταβλητές αυτές συχνά εξαρτώνται από άλλες μεταβλητές (επεξηγηματικές μεταβλητές). Έτσι, λοιπόν αναπτύχθηκε μια ευρύτερη κλάση μοντέλων, τα «Γενικευμένα Γραμμικά Μοντέλα», που αποτελούν επεκτάσεις του γραμμικού μοντέλου. Τα μοντέλα αυτά προτάθηκαν από τους John Nelder και Robert Wedderburn (1972) και αποτελούν μια πολύ σημαντική κλάση μοντέλων για μεταβλητές απόκρισης με κατανομές από την Εκθετική Οικογένεια Κατανομών (ΕΟΚ).

1.2 Η Δομή του Μοντέλου

Θα παρουσιάσουμε τα βασικά μέρη ενός γενικευμένου γραμμικού μοντέλου. Θεωρούμε ανεξάρτητες τυχαίες μεταβλητές $\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_n$, όπου κάθε μια από αυτές ακολουθεί την ίδια κατανομή από την Εκθετική Οικογένεια Κατανομών και εξαρτάται από μια παράμετρο θ_i. Επιπλέον, θεωρούμε τον $n \times p$ πίνακα σχεδιασμού $\mathbf{X} = \left[\boldsymbol{x}_1^{\mathrm{T}}, \ldots, \boldsymbol{x}_n^{\mathrm{T}}\right]$ με $\boldsymbol{x}_i^{\mathrm{T}}$ να είναι η i – οστή στήλη των παρατηρούμενων τιμών των

επεξηγηματικών μεταβλητών και έστω $\boldsymbol{\beta}^{\mathrm{T}} = [\beta_0, \beta_1, \cdots, \beta_k]$ το διάνυσμα των άγνωστων παραμέτρων του μοντέλου $(p = k+1)$.

Η συνάρτηση πυκνότητας πιθανότητας (για συνεχή μεταβλητή) ή η συνάρτηση μάζας πιθανότητας (για διακριτή μεταβλητή) για κάθε Y_i στην κανονική μορφή δίνεται από τη σχέση

$$f(y_i;\theta_i) = \exp\left[y_i b(\theta_i) + c(\theta_i) + d(y_i)\right]. \quad (1.6)$$

Συνεπώς, η από κοινού συνάρτηση πυκνότητας ή μάζας πιθανότητας είναι της μορφής

$$f(y_1, \cdots, y_n; \theta_1, \cdots, \theta_n) = \exp\left[\sum_{i=1}^{n} y_i b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)\right]. \quad (1.7)$$

Θεωρούμε ότι οι επεξηγηματικές μεταβλητές του μοντέλου συνδέονται γραμμικά με μία συνάρτηση της αναμενόμενης μέσης τιμής της μεταβλητής απόκρισης, σχηματίζοντας έτσι μία σχέση, η οποία καλείται γραμμική προβλέπουσα (*linear predictor*) $\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. Οι ανεξάρτητες μεταβλητές σχετίζονται με την αναμενόμενη τιμή της Y_i μέσω ενός κατάλληλου μετασχηματισμού $\mathrm{g}(\cdot)$ της μ_i, όπου $\mathrm{E}(\mathrm{Y}_i) = \mu_i$, δηλαδή $g(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. Επισημαίνουμε εδώ ότι η αναμενόμενη τιμή της μεταβλητής απόκρισης και η γραμμική προβλέπουσα δεν έχουν εν γένει γραμμική σχέση. Ωστόσο, αυτό επιτυγχάνεται μέσω της συνάρτησης $g(\cdot)$.

Η συνάρτηση $g(\cdot)$ είναι μια 1-1, μονότονη και διαφορίσιμη συνάρτηση και καλείται συνάρτηση σύνδεσης (*link function*).

Με βάση τα παραπάνω διακρίνουμε τα τρία βασικά μέρη ενός γενικευμένου γραμμικού μοντέλου:

1. ***Στοχαστικό μέρος***

   Οι τυχαίες μεταβλητές $\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_n$ θα πρέπει να ακολουθούν την ίδια κατανομή από την Εκθετική Οικογένεια Κατανομών.

2. ***Συστηματικό μέρος***

   Οι επεξηγηματικές μεταβλητές (*predictor*) συνδέονται γραμμικά με μία συνάρτηση της αναμενόμενης μέσης τιμής της μεταβλητής απόκρισης σχηματίζοντας τη γραμμική προβλέπουσα (*linear predictor*) $\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$.

3. *Συνάρτηση σύνδεσης*

Η γραμμική προβλέπουσα στη γενική περίπτωση σχετίζεται με την αναμενόμενη τιμή της μεταβλητής απόκρισης μ_i μέσω της συνάρτησης σύνδεσης $g(\cdot)$, η οποία αποτελεί ένα μετασχηματισμό της μ_i, ώστε να επιτευχθεί η γραμμική δομή του μοντέλου. Ισχύει, λοιπόν η σχέση

$$\eta_i = g\left(\mu_i\right) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} \ .$$

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

1. $\mathrm{Y}_i \mid \mathbf{X} = \boldsymbol{x} \sim$ ***Κανονική κατανομή***

$$\eta_i = g\left(\mu_i\right) = \mu_i \ (\textit{identity link}).$$

2. $\mathrm{Y}_i \mid \mathbf{X} = \boldsymbol{x} \sim$ ***Κατανομή Bernoulli ή Διωνυμική***

$$\eta_i = g\left(\mu_i\right) = \log\left(\frac{\mu_i}{1-\mu_i}\right) \ (\textit{logit link}).$$

3. $\mathrm{Y}_i \mid \mathbf{X} = \boldsymbol{x} \sim$ ***Κατανομή Poisson***

$$\eta_i = g\left(\mu_i\right) = \log\left(\mu_i\right) \ (\textit{log link}).$$

4. $\mathrm{Y}_i \mid \mathbf{X} = \boldsymbol{x} \sim$ ***Κατανομή Gamma ή Εκθετική***

$$\eta_i = g\left(\mu_i\right) = \frac{1}{\mu_i} \ (\textit{reciprocal link}).$$

Με τις παραπάνω επιλογές ισχύει ότι $\eta_i = g\left(\mu_i\right) = \theta_i$ και σε αυτή την περίπτωση η συνάρτηση σύνδεσης καλείται «κανονική» συνάρτηση σύνδεσης (*canonical link*).

Στον Πίνακα 1.1 αναγράφονται συγκεντρωτικά οι κανονικές συναρτήσεις σύνδεσης για τις συνηθέστερες επιλογές κατανομών, που σχετίζονται με τα Γενικευμένα Γραμμικά Μοντέλα.

Κατανομή	**Κανονική συνάρτηση σύνδεσης**
Κανονική	$\eta_i = g\left(\mu_i\right) = \mu_i$ (*identity link*)

Διωνυμική	$\eta_i = g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$ $(logit\ link)$
Poisson	$\eta_i = g(\mu_i) = \log(\mu_i)$ $(log\ link)$
Εκθετική	$\eta_i = g(\mu_i) = \frac{1}{\mu_i}$ $(reciprocal\ link)$
Γάμμα	$\eta_i = g(\mu_i) = \frac{1}{\mu_i}$ $(reciprocal\ link)$

Πίνακας 1: *Κανονικές συναρτήσεις σύνδεσης για Γενικευμένα Γραμμικά Μοντέλα.*

Φυσικά, υπάρχουν και άλλες συναρτήσεις σύνδεσης, 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 η Φ είναι η συνάρτηση κατανομής της τυποποιημένης Κανονικής κατανομής.

2. ***Complementary log-log link***

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

3. ***Power family link***

$$\eta_i = \begin{cases} \mu_i^{\lambda} & ,\lambda \neq 0 \\ \log(\mu_i) & ,\lambda = 0 \end{cases}.$$

Οι συναρτήσεις σύνδεσης logit, probit και complementary log-log χρησιμοποιούνται για διωνυμικά (ή δίτιμα) δεδομένα. Οι μετασχηματισμοί logit και probit είναι σχεδόν πανομοιότυποι, αλλά προτιμάται ο logit, γιατί έχει ευκολότερη και άμεση ερμηνεία όντας ο λογάριθμος των συμπληρωματικών ή σχετικών πιθανοτήτων (*odds*). Το μοντέλο complementary log-log συμπίπτει με τα μοντέλα probit και logit για τιμές κοντά στο $p = 0.5$, ενώ διαφέρει για τιμές κοντά στο μηδέν ή στο ένα. Για δεδομένα, που εκφράζουν αριθμό συμβάντων σε ορισμένο χρονικό ή χωρικό διάστημα (τα οποία συνήθως θεωρούμε ότι προέρχονται από την κατανομή Poisson) χρησιμοποιείται η συνάρτηση log.

Η επιλογή της συνάρτησης σύνδεσης μπορεί να θεωρηθεί ισοδύναμη με την επιλογή μετασχηματισμού της αναμενόμενης μέσης τιμής της εξαρτημένης μεταβλητής. Η συνάρτηση σύνδεσης ουσιαστικά άρει τους περιορισμούς σχετικά με το διάστημα, που λαμβάνει τιμές η αναμενόμενη μεταβλητή απόκρισης και εκμεταλλεύεται την κατανομή αυτής. Έτσι, ακατάλληλη επιλογή συνάρτησης σύνδεσης μπορεί να προκαλέσει προβλήματα στην προσαρμογή του μοντέλου, και κατά συνέπεια στο γενικευμένο γραμμικό μοντέλο.

1.3 Εκθετική Οικογένεια Κατανομών και Γενικευμένα Γραμμικά Μοντέλα

Μια ευρύτατη οικογένεια κατανομών με εξαιρετικά χρήσιμες ιδιότητες άμεσα συνυφασμένη με την φιλοσοφία των γενικευμένων γραμμικών μοντέλων είναι η Εκθετική Οικογένεια Κατανομών (Ε.Ο.Κ.). Ανάλογα με τον αριθμό των παραμέτρων που φέρει διακρίνουμε τη μονοπαραμετρική και πολυπαραμετρική Εκθετική Οικογένεια.

1.3.1 Εκθετική Οικογένεια Κατανομών

Θεωρούμε ένα σύνολο ανεξάρτητων τυχαίων μεταβλητών $\mathrm{Y} = (\mathrm{Y}_1, \mathrm{Y}_2, ..., \mathrm{Y}_n)$, όπου κάθε μία από αυτές προέρχεται από μια κατανομή της Εκθετικής Οικογένειας Κατανομών. Τότε κάθε Y_i θα έχει συνάρτηση πυκνότητας πιθανότητας (για συνεχή Y_i) ή συνάρτηση μάζας πιθανότητας (για διακριτή Y_i) της μορφής

$$f(y_i;\theta_i,\varphi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\alpha(\varphi)} + c(y_i,\varphi)\right\}, \quad (1.8)$$

όπου οι συναρτήσεις $\alpha(\cdot), b(\cdot), c(\cdot)$ είναι γνωστές, y_i $(i = 1,2,..,n)$ το $i-$οστό αποτέλεσμα μιας τυχαίας δειγματοληψίας από μία κατανομή της Ε.Ο.Κ. Η θ_i ονομάζεται φυσική ή κανονική παράμετρος θέσης (*natural or canonical location parameter*) και η φ παράμετρος κλίμακας ή μεταβλητότητας (*scale or dispersion parameter*). Αν θεωρήσουμε την παράμετρο φ γνωστή, τότε η συνάρτηση $\alpha(\varphi)$ μπορεί να εκφραστεί ως $\alpha\varphi$, με α μια σταθερά.

Ανάλογα με τον αριθμό των παραμέτρων που φέρει η κατανομή ανήκει στη μονοπαραμετρική ή στη πολυπαραμετρική Εκθετική Οικογένεια. Πολλές από τις γνωστές κατανομές ανήκουν στη μονοπαραμετρική ή στην πολυμαραμετρική

Εκθετική Οικογένεια. Εύκολα μπορεί να διαπιστώσει κανείς ότι στην εν λόγω οικογένεια ανήκουν: (α) από τις διακριτές κατανομές, η κατανομή Bernoulli, η διωνυμική, η Poisson κλπ. και (β) από τις συνεχείς, η εκθετική, η Κανονική, η Γάμμα κλπ. κατανομές.

Παραθέτουμε κάποια παραδείγματα κατανομών από την Εκθετική Οικογένεια Κατανομών.

Κανονική Κατανομή

Η συνάρτηση πυκνότητας πιθανότητας είναι

$$f\left(y_i;\mu_i,\sigma^2\right)=\frac{1}{\left(2\pi\sigma^2\right)^{1/2}}\exp\left[-\frac{1}{2\sigma^2}\left(y_i-\mu_i\right)^2\right],$$

Η εν λόγω κατανομή ανήκει στην Ε.Ο.Κ και η παραπάνω σχέση γράφεται στη μορφή (1.8) ως εξής

$$f(y_i;\mu_i,\sigma^2)=\exp\left\{\left[y_i\mu_i-\frac{\mu_i^2}{2}\right]\frac{1}{\sigma^2}-\frac{y_i^2}{2\sigma^2}-\frac{1}{2}\log\left(2\pi\sigma^2\right)\right\},$$

όπου $\theta_i=\mu_i$, $b(\theta_i)=\mu_i^2/2$, $\alpha(\varphi)=\varphi$, $\varphi=\sigma^2$ και $c(y_i;\varphi)=-\frac{1}{2}\left\{\frac{y_i^2}{\sigma^2}+\log(2\pi\sigma^2)\right\}$.

Η παράμετρος θέσης είναι η μ_i και η παράμετρος κλίμακας η σ^2.

Η Κανονική κατανομή χρησιμοποιείται για πραγματικά δεδομένα όπου πιστεύουμε ότι προέρχονται από μια συμμετρική κατανομή και αποτελούν τις τιμές συνεχών τυχαίων μεταβλητών. Χρησιμοποιείται ευρύτατα για τρείς κυρίως λόγους. Αρχικά, παρατηρείται ότι πολλά τυχαία φαινόμενα περιγράφονται ικανοποιητικά με την εν λόγω κατανομή, όπως για παράδειγμα το ύψος ή τα επίπεδα αρτηριακής πίεσης των ανθρώπων κλπ. Επιπλέον, ακόμη και αν τα δεδομένα δεν ακολουθούν Κανονική κατανομή, σύμφωνα με το Κεντρικό Οριακό Θεώρημα, η μέση τιμή ή το ολικό μιας ακολουθίας ανεξάρτητων και ισόνομων τυχαίων μεταβλητών ακολουθεί προσεγγιστικά Κανονική κατανομή. Τέλος, υπάρχει μια ευρεία στατιστική θεωρία που βασίζεται στην Κανονική κατανομή. Συνεπώς, αν μια συνεχής τυχαία μεταβλητή δεν ακολουθεί Κανονική κατανομή είναι σύνηθες πριν γίνει έλεγχος για πιθανή άλλη γνωστή κατανομή να προσπαθήσουμε να διακρίνουμε κάποιο μετασχηματισμό της

μεταβλητής αυτής, όπως $\mathrm{Y}' = \log \mathrm{Y}$ ή $\mathrm{Y}' = \sqrt{\mathrm{Y}}$, που τις μετατρέπει σε Κανονικές μεταβλητές.

Poisson κατανομή

Η συνάρτηση μάζας πιθανότητας της διακριτής τυχαίας μεταβλητής $\mathrm{Y}_i \sim \mathrm{Poisson}(\theta_i)$ είναι

$$f(y_i;\theta_i) = \frac{\theta_i^{y} e^{-\theta_i}}{y_i!},$$

όπου η τυχαία μεταβλητή Y_i παίρνει τιμές 0, 1, 2, …

Η κατανομή Poisson ανήκει στην Ε.Ο.Κ., αφού η συνάρτηση μάζας πιθανότητας μπορεί να γραφεί στη μορφή

$$f(y_i;\mu_i) = \exp[y_i \log \mu_i - \mu_i - \log(y_i!)],$$

όπου $\theta_i = \log \mu_i$, $b(\theta_i) = \exp(\theta_i)$ και $c(y_i, \varphi) = -\log(y_i!)$. Η παράμετρος θέσης είναι η μ_i, ενώ η παράμετρος κλίμακας είναι σταθερή και ίση με 1, $\varphi = 1$.

Η κατανομή Poisson χρησιμοποιείται για να περιγράψει τον αριθμό των πραγματοποιήσεων ενός συμβάντος ή γεγονότος σε μια συγκεκριμένη χρονική περίοδο όπου τα συμβάντα θεωρούμε ότι είναι ανεξάρτητα. Κάποια παραδείγματα είναι: ο αριθμός των ορθογραφικών λαθών σε μια σελίδα μιας εφημερίδας, ο αριθμός των περιστατικών υγείας ενός ατόμου κ.λ.π. Αν μια τυχαία μεταβλητή ακολουθεί τη κατανομή $\mathrm{Poisson}(\theta_i)$ τότε η μέση τιμή και η διασπορά αυτής είναι ίσες και μάλιστα ισχύει $\mu_i = \sigma_i^2 = \theta_i$. Πραγματικά δεδομένα, που ενδεχομένως ακολουθούν την εν λόγω κατανομή συνήθως έχουν μεγαλύτερη διασπορά από την θεωρητική αυτή τιμή. Αυτό αποδίδεται στο φαινόμενο της υπερσκέδασης ή υπερμεταβλητότητας (*overdispersion*), και έτσι το μοντέλο, που θα χρησιμοποιήσουμε πρέπει να αντιμετωπίζει αυτό το πρόβλημα, όπως θα δούμε και σε επόμενο κεφάλαιο.

1.3.2 Μέση τιμή και Διασπορά

Χρησιμοποιώντας τις γνωστές ιδιότητες μιας συνάρτησης πιθανότητας μπορούμε να έχουμε κάποιες χρήσιμες εκφράσεις για την μέση τιμή και τη διασπορά της τυχαίας μεταβλητής Y, λαμβάνοντας υπόψη φυσικά ότι η κατανομή της ανήκει στην Εκθετική Οικογένεια κατανομών. Από τον ορισμό της συνάρτησης πυκνότητας πιθανότητας μιας μεταβλητής γνωρίζουμε ότι το εμβαδόν που περικλείεται στην καμπύλη είναι ίσο με ένα, δηλ.

$$\int f\left(y;\theta,\varphi\right)dy = 1\,,\ (1.9)$$

όπου η ολοκλήρωση γίνεται στο διάστημα που ορίζεται η συνάρτηση $f(y;\theta,\varphi)$ (η παραπάνω σχέση ισχύει για συνεχείς τυχαίες μεταβλητές. Αν μια τ.μ. Y είναι διακριτή, τότε η ολοκλήρωση αντικαθίσταται από άθροισμα στο πεδίο ορισμού της $f(y;\theta,\varphi)$).

Παραγωγίζοντας την παραπάνω σχέση ως προς θ κατά μέλη έχουμε ότι

$$\frac{\partial}{\partial\theta}\int f\left(y;\theta,\varphi\right)dy = \frac{\partial}{\partial\theta}\cdot 1 = 0.\ (1.10)$$

Και τελικά προκύπτει ότι

$$\int \frac{\partial f\left(y;\theta,\varphi\right)}{\partial\theta}dy = 0.\quad (1.11)$$

Ομοίως, παραγωγίζοντας δεύτερη φορά ως προς θ τη σχέση (1.11) έχουμε ότι

$$\int \frac{\partial^2 f\left(y;\theta,\varphi\right)}{\partial\theta^2}dy = 0.\quad (1.12)$$

Χρησιμοποιώντας τις παραπάνω σχέσεις για κατανομές, που ανήκουν στην Εκθετική Οικογένεια Κατανομών καταλήγουμε στα εξής συμπεράσματα:

Η συνάρτηση πιθανότητας στη γενική της μορφή γράφεται

$$f(y;\theta,\varphi) = \exp\left\{\frac{y\theta - b(\theta)}{\alpha(\varphi)} + c(y,\varphi)\right\}.\ (1.13)$$

Συνεπώς, παραγωγίζοντας την παραπάνω σχέση προκύπτει ότι

$$\frac{\partial f\left(y;\theta,\varphi\right)}{\partial\theta} = \left[\frac{y - b'(\theta)}{\alpha(\varphi)}\right] f\left(y;\theta,\varphi\right).\ (1.14)$$

Από τη σχέση (1.11), προκύπτει ότι

$$\int\left[\frac{y - b'(\theta)}{\alpha(\varphi)}\right] f\left(y;\theta,\varphi\right)dy = 0 \Rightarrow \mathrm{E}\left(\mathrm{Y}\right) - b'\left(\theta\right) = 0 \Rightarrow \mathrm{E}\left(\mathrm{Y}\right) = b'(\theta).\ (1.15)$$

Κατά τον ίδιο τρόπο παραγωγίζοντας τη σχέση (1.14) ως προς θ και λαμβάνοντας υπόψη τη σχέση (1.12) έχουμε ότι

$$\int \frac{-b''(\theta)}{\alpha(\varphi)} f\left(y;\theta,\varphi\right) + \left[\frac{y - b'(\theta)}{\alpha(\varphi)}\right]^2 f\left(y;\theta,\varphi\right)dy = 0\,.\ (1.16)$$

Αντικαθιστώντας τον όρο $b'(\theta)$ από τη σχέση (1.15) προκύπτει ότι

$$\frac{-b''(\theta)}{\alpha(\varphi)}\int f\left(y;\theta,\varphi\right)dy+\frac{1}{\alpha^2(\varphi)}\int\left(y-\mathrm{E}(\mathrm{Y})\right)^2 f\left(y;\theta,\varphi\right)dy=0\Rightarrow\frac{1}{\alpha^2(\varphi)}V(\mathrm{Y})=\frac{b''(\theta)}{\alpha(\varphi)}$$
$$\Rightarrow V(\mathrm{Y})=b''(\theta)\alpha(\varphi). \quad (1.17)$$

Η τελευταία σχέση μπορεί να γραφεί και ως $V(\mathrm{Y})=V(\mu)\alpha(\varphi)$, όπου $V(\mu)=b''(\theta)$ καλείται συνάρτηση διασποράς ή μεταβλητότητας (*variance function*) και αποτελεί μια έκφραση της εξάρτησης της διασποράς της τυχαίας μεταβλητής Y με την μέση τιμή αυτής.

Επιπλέον, είναι χρήσιμο για την περαιτέρω ανάλυση των γενικευμένων γραμμικών μοντέλων να βρούμε μια έκφραση για τη μέση τιμή και τη διασπορά των παραγώγων του λογαρίθμου της συνάρτησης πιθανοφάνειας ως προς θ, γνωστών και ως συναρτήσεων score.

Η συνάρτηση πιθανοφάνειας, θεωρώντας ότι έχουμε μια μόνο παρατήρηση y θα ταυτίζεται με τη συνάρτηση πιθανότητας. Άρα, θα είναι

$$L(y;\theta,\varphi)=\exp\left\{\frac{y\theta-b(\theta)}{\alpha(\varphi)}+c(y,\varphi)\right\}.$$

Συνεπώς, ο λογάριθμος της συνάρτησης πιθανότητας θα είναι

$$l\left(\theta;y,\varphi\right)=\frac{y\theta-b(\theta)}{\alpha(\varphi)}+c(y,\varphi). \quad (1.18)$$

Παραγωγίζοντας την παραπάνω σχέση ως προς θ προκύπτει ότι

$$U=U\left(\theta,\varphi;y\right)=\frac{y-b'(\theta)}{\alpha(\varphi)}. \quad (1.19)$$

Η συνάρτηση U αποτελεί μια στατιστική συνάρτηση (*score statistic*) και αφού εξαρτάται από την τυχαία μεταβλητή Υ μπορεί να θεωρηθεί ως μια τυχαία μεταβλητή.

Έχει νόημα, λοιπόν να βρούμε μια σχέση για την μέση τιμή και τη διασπορά της μεταβλητής αυτής. Η αναμενόμενη τιμή της θα είναι

$$\mathrm{E}\left(U\right)=\frac{1}{\alpha(\varphi)}\left(\mathrm{E}\left(\mathrm{Y}\right)-b'\left(\theta\right)\right). \quad (1.20)$$

Αντικαθιστώντας την έκφραση για την μέση τιμή της Υ από τη σχέση (1.15) προκύπτει ότι

$$\mathrm{E}\left(U\right) = 0 \ . \ (1.21)$$

Από ιδιότητες της συνάρτησης διασποράς μιας τυχαίας μεταβλητής προκύπτει ότι

$$V\left(U\right) = \frac{1}{\alpha^2(\varphi)} V\left[\mathrm{Y}\right]. \ (1.22)$$

Χρησιμοποιώντας τη σχέση (1.17) έχουμε ότι

$$J = V\left(U\right) = \frac{1}{\alpha^2(\varphi)} b''(\theta)\alpha(\varphi) = \frac{b''(\theta)}{\alpha(\varphi)}. \ (1.23)$$

Η διασπορά της τυχαίας μεταβλητής U καλείται πληροφορία (*information*) και μπορούμε να τη συμβολίσουμε με J , όπως φαίνεται παραπάνω.

Στον Πίνακα 2 παρουσιάζονται συγκεντρωτικά οι ιδιότητες των πιο συνηθισμένων κατανομών από την Εκθετική Οικογένεια Κατανομών.

Κατανομή	$a(\theta)$	$b(\theta)$	$c(y,\varphi)$	$\mu(\theta) = \mathrm{E}\left(\mathrm{Y}\right)$	$V(\mathrm{Y})$	$V(\mu)$
Poisson (θ)	1	e^{θ}	$-\log(y!)$	e^{θ}	μ	μ
Διωνυμική (n, θ)	1	$n\log(1+e^{\theta})$	$\log\binom{n}{y}$	$ne^{\theta}/(1+e^{\theta})$	$n\pi(1-\pi)$	$n\mu(1-\mu)$
Κανονική (θ, σ^2)	σ^2	$\theta^2/2$	$-\frac{1}{2}\frac{y^2}{\varphi}$ $-\frac{1}{2}\log(2\pi\varphi)$	θ	σ^2	σ^2

Πίνακας 2: *Ιδιότητες ορισμένων κατανομών της Εκθετικής Οικογένειας Κατανομών.*

1.4 Εκτιμητική

1.4.1 Εισαγωγή

Για να προσαρμόσουμε το πιθανό μοντέλο στα δεδομένα μας, έτσι ώστε να έχουμε μια εκτίμηση για την αναμενόμενη τιμή της μεταβλητής Y, θα πρέπει να εκτιμήσουμε τις παραμέτρους του μοντέλου με βάση τις παρατηρήσεις μας. Στα γενικευμένα γραμμικά μοντέλα, λοιπόν προσαρμόζοντας το κατάλληλο μοντέλο στα δεδομένα μας μπορούμε να έχουμε Εκτίμηση κατά Σημείο ή κατά Διάστημα των παραμέτρων της κατανομής της μεταβλητής απόκρισης χρησιμοποιώντας μεθόδους, που βασίζονται στη Μέθοδο Μέγιστης Πιθανοφάνειας. Συνήθως, η Εκτιμήτρια Μέγιστης Πιθανοφάνειας (ΕΜΠ) υπάρχει και μπορούμε να βρούμε εύκολα μια μαθηματική της έκφραση. Ωστόσο, υπάρχουν περιπτώσεις, όπου για την εύρεση των εκτιμητριών οδηγούμαστε σε ένα μη γραμμικό σύστημα εξισώσεων και για την επίλυσή του χρειάζεται πολλές φορές η εφαρμογή αριθμητικών μεθόδων, όπως η μέθοδος Newton-Raphson.

Παραθέτουμε ένα απλό παράδειγμα εκτίμησης παραμέτρων με τη μέθοδο μέγιστης πιθανοφάνειας κάνοντας και μια εισαγωγή στην επαναληπτική μέθοδο N-R.

1.4.2 Παράδειγμα Δεδομένων Διάρκειας Ζωής

Τα δεδομένα του Πίνακα 3 απεικονίζουν τη διάρκεια ζωής 25 ρουλεμάν ίδιου τύπου τα οποία δοκιμάζονται με το ίδιο φορτίο. Η πλέον κατάλληλη κατανομή, για να περιγράψουμε δεδομένα διάρκειας ζωής είναι η κατανομή Weibull με παραμέτρους α και λ .

Η συνάρτηση πυκνότητας πιθανότητας της εν λόγω κατανομής είναι

$$f\left(y;\lambda,\alpha\right)=\frac{\lambda y^{\lambda-1}}{\alpha^{\lambda}}\exp\left\{-\left(\frac{y}{\alpha}\right)^{\lambda}\right\}. \quad (1.24)$$

όπου $y>0,$ καθώς εκφράζει διάρκεια ζωής, λ είναι η παράμετρος σχήματος και α είναι η παράμετρος κλίμακας.

Επιπλέον, η συνάρτηση αξιοπιστίας είναι

$$S\left(y;\lambda,\alpha\right)=\mathrm{P}(\mathrm{Y}>y)=\exp\left\{-\left(\frac{y}{\alpha}\right)^{\lambda}\right\}. \quad (1.25)$$

Πίνακας 3: *Ο αριθμός των περιστροφών των 25 ρουλεμάν (εκκατομύρια)*

17.88	98.64	67.80*	45.60	173.40*
54.12	33.00	105.84*	68.64	51.84
93.12	67.80	42.12	128.04	68.88*
28.92	105.12	67.80*	48.48	51.96
55.56	41.52	127.92	68.64*	84.12

Ας θεωρήσουμε ότι οι παρατηρήσεις μας χωρίζονται στα ακόλουθα υποσύνολα

U = {μη αποκομμένες παρατηρήσεις}

C = {αποκομμένες παρατηρήσεις}

Τότε, η συνάρτηση πιθανοφάνειας δίνεται ως

$$L(y_1, y_2, \cdots, y_n; \lambda, \alpha) = \prod_{i \in U} \frac{\lambda y_i^{\lambda-1}}{\alpha^{\lambda}} \exp\left\{-\left(\frac{y_i}{\alpha}\right)^{\lambda}\right\} * \prod_{i \in C} \exp\left\{-\left(\frac{y_i}{\alpha}\right)^{\lambda}\right\}.$$

Ο λογάριθμος της πιθανοφάνειας είναι

$$l(\lambda, \alpha) = \sum_{i/u} \log f(y_i) + \sum_{i \in C} \log S(y_i) = \kappa \log \lambda - \kappa\lambda \log \alpha + (\lambda - 1) \sum_{i \in U} \log y_i - \sum_{i=1}^{n} \left(\frac{y_i}{\alpha}\right)^{\lambda},$$

όπου κ το πλήθος των μη-αποκομμένων παρατηρήσεων του υποσυνόλου U και $n = 25$ το μέγεθος του δείγματος.

Οι πρώτες παράγωγοι ως προς α και λ είναι

$$\frac{\partial l}{\partial \alpha} = -\frac{\kappa\lambda}{\alpha} + \lambda\alpha^{-\lambda-1} \sum_{i=1}^{n} y_i^{\lambda} = 0 \Rightarrow \frac{\kappa\hat{\lambda}}{\hat{\alpha}} = \frac{\hat{\lambda} \sum_{i=1}^{n} y_i^{\hat{\lambda}}}{\hat{\alpha}^{\hat{\lambda}+1}} \Rightarrow \hat{\alpha} = \left\{\frac{\sum_{i=1}^{n} y_i^{\hat{\lambda}}}{\kappa}\right\}^{1/\hat{\lambda}} \quad (1.26)$$

$$\frac{\partial l}{\partial \lambda} = \frac{\kappa}{\lambda} - \kappa \log \alpha + \sum_{i \in U} \log y_i - \sum_{i=1}^{n} \left(\frac{y_i}{\alpha}\right)^{\lambda} \log\left(\frac{y_i}{\alpha}\right)$$

Αντικαθιστώντας το $\hat{\alpha}$ στην εξίσωση $\partial l / \partial \lambda = 0$, προκύπτει

$$\frac{\sum_{i=1}^{n} y_i^{\hat{\lambda}} \log(y_i)}{\sum_{i=1}^{n} y_i^{\hat{\lambda}}} - \frac{1}{\hat{\lambda}} - \frac{\sum_{i \in U} \log(y_i)}{\kappa} = 0, \quad (1.27)$$

η οποία επιλύεται ως προς την $\hat{\lambda}$ μόνο με αριθμητικές μεθόδους.

Για τον προσδιορισμό των εκτιμητριών για τις παραμέτρους α και λ της κατανομής Weibull θα χρησιμοποιήσουμε αριθμητικές μεθόδους και συγκεκριμένα τη μέθοδο Newton-Raphson.

Σκοπός μας είναι η εύρεση της τιμής $\hat{\lambda}$, που μηδενίζει την συνάρτηση $g\left(\hat{\lambda}\right)=\frac{\sum_{i=1}^{n} y_i^{\hat{\lambda}} \log\left(y_i\right)}{\sum_{i=1}^{n} y_i^{\hat{\lambda}}}-\frac{1}{\hat{\lambda}}-\frac{\sum_{i\in U} \log\left(y_i\right)}{\kappa}$.

Η κλίση της συνάρτησης $g\left(\hat{\lambda}\right)$ στο σημείο $\hat{\lambda}^{(m-1)}$ δίνεται από τη σχέση:

$$\left[\frac{dg}{d\hat{\lambda}}\right]_{\hat{\lambda}=\hat{\lambda}^{(m-1)}}=g'\left(\hat{\lambda}^{(m-1)}\right)=\frac{g\left(\hat{\lambda}^{(m)}\right)-g\left(\hat{\lambda}^{(m-1)}\right)}{\hat{\lambda}^{(m)}-\hat{\lambda}^{(m-1)}},(1.28)$$

όπου η απόσταση $\hat{\lambda}^{(m)}-\hat{\lambda}^{(m-1)}=h$ με h πολύ μικρό. Αν θεωρήσουμε ότι στο $(m)-$βήμα η $\hat{\lambda}^{(m)}$ αποτελεί λύση της εξίσωσης $\mathrm{g}(\hat{\lambda})=0 \Rightarrow \mathrm{g}\left(\hat{\lambda}^{(m)}\right)=0$. Τότε η (1.28) γράφεται

$$\hat{\lambda}^{(m)}=\hat{\lambda}^{(m-1)}-\frac{g\left(\hat{\lambda}^{(m-1)}\right)}{g'\left(\hat{\lambda}^{(m-1)}\right)}\cdot(1.29)$$

Ξεκινώντας, λοιπόν από μια αρχική τιμή $\hat{\lambda}^{(1)}$ και χρησιμοποιώντας τη σχέση (1.29) εκτελούμε τον αλγόριθμο Newton-Raphson για κατάλληλο αριθμό βημάτων m έως ότου η μέθοδος συγκλίνει.

Η πρώτη παράγωγος της συνάρτησης $g\left(\hat{\lambda}\right)$ ως προς $\hat{\lambda}$ θα είναι

$$g'\left(\hat{\lambda}\right)=\frac{\sum_{i=1}^{n} y_i^{\hat{\lambda}} \log\left(y_i\right)^2}{\sum_{i=1}^{n} y_i^{\hat{\lambda}}}-\frac{\sum_{i=1}^{n} y_i^{\hat{\lambda}} \log\left(y_i\right)^2}{\sum_{i=1}^{n} y_i^{\hat{\lambda}}}+\frac{1}{\hat{\lambda}^2}.$$

Με βάση τα δεδομένα και δοκιμάζοντας διάφορες τιμές για την παράμετρο $\hat{\lambda}$ έχουμε π.χ $g(1)<0$,$g(1.5)<0,\ldots$. Έτσι, επιλέγουμε μια αρχική τιμή για την $\hat{\lambda}^{(1)}$ ίση με 1.9. Από την τιμή αυτή η μέθοδος συγκλίνει γρήγορα στην εκτίμηση $\hat{\lambda}=1.9467$. Αντικαθιστώντας στη σχέση (1.26) έχουμε την εκτίμηση για την παράμετρο κλίμακας $\hat{\alpha}$=91.6383.

1.4.3 Μέθοδος Μέγιστης Πιθανοφάνειας

Θεωρούμε ανεξάρτητες τυχαίες μεταβλητές, έστω $\mathrm{Y}_1,\mathrm{Y}_2,\ldots,\mathrm{Y}_n$, ώστε να ικανοποιούν τις προϋποθέσεις ενός γενικευμένου γραμμικού μοντέλου με $\mathrm{E}\left(\mathrm{Y}_i\right)=\mu_i$ και συνάρτηση σύνδεσης $\eta_i=g\left(\mu_i\right)=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$.

Οι παρατηρούμενες τιμές των επεξηγηματικών μεταβλητών $\boldsymbol{x}_i$ και οι αντίστοιχοι συντελεστές $\boldsymbol{\beta}$ ορίζονται ως εξής: $\boldsymbol{x}_i^{\mathrm{T}} = \left(x_{i0}, x_{i1}, \cdots, x_{ik}\right)$ και $\boldsymbol{\beta}^{\mathrm{T}} = \left(\beta_0, \beta_1, \ldots, \beta_k\right)$.

Επιπλέον, οι εκφράσεις για την μέση τιμή και τη διασπορά για κάθε μια από τις τ.μ Y_i δίνονται από τις σχέσεις

$$\mathrm{E}(\mathrm{Y}_i) = \mu_i = b'(\theta_i). \quad (1.30)$$

$$V\left(\mathrm{Y}_i\right) = \alpha(\varphi) b''(\theta_i). \quad (1.31)$$

Η συνάρτηση πιθανοφάνειας ενός δείγματος τιμών $y_1, y_2, \cdots, y_n$ των τ.μ. $\mathrm{Y}_1, \mathrm{Y}_2, \cdots, \mathrm{Y}_n$ δίνεται από τη σχέση

$$L = \prod_{i=1}^{n} f\left(y_i; \theta_i, \varphi\right) = \prod_{i=1}^{n} \exp\left[\frac{y_i\theta_i - b(\theta_i)}{\alpha(\varphi)} + c(y_i, \varphi)\right]. \quad (1.32)$$

Ο λογάριθμος της συνάρτησης πιθανοφάνειας δίνεται από τη σχέση

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \left\{\frac{y_i\theta_i - b(\theta_i)}{\alpha(\varphi)} + c(y_i, \varphi)\right\}. \quad (1.33)$$

Άρα, παραγωγίζοντας την σχέση (1.33) ως προς τις παραμέτρους του μοντέλου χρησιμοποιώντας τον κανόνα της αλυσίδας καταλήγουμε στην παρακάτω σχέση

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n}\left[\frac{\partial l_i}{\partial \beta_j}\right] = \sum_{i=1}^{n}\left[\frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_j}\right], \; j = 0, 2, \ldots, k, \quad (1.34)$$

Αναλύοντας ξεχωριστά κάθε όρο του δεξιού μέλους της σχέσης (1.34) θα έχουμε μια απλούστερη και πιο εύχρηστη σχέση για κάθε μία από τις παραπάνω συναρτήσεις. Γνωρίζουμε ότι για κάθε Y_i, ο λογάριθμος της πιθανοφάνειας είναι

$$l_i = \frac{y_i\theta_i - b(\theta_i)}{\alpha(\varphi)} + c(y_i, \varphi). \quad (1.35)$$

Παραγωγίζοντας την σχέση (1.35) ως προς θ_i και χρησιμοποιώντας την σχέση (1.30) προκύπτει ότι

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - \mu_i}{\alpha(\varphi)}. \quad (1.36)$$

Επιπλέον είναι άμεσο ότι ο όρος $\dfrac{\partial \theta_i}{\partial \mu_i}$ εκφράζεται ως εξής

$$\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1}. \quad (1.37)$$

Παραγωγίζοντας την σχέση (1.30) ως προς θ_i προκύπτει

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i). \quad (1.38)$$

Αντικαθιστώντας τον όρο $b''(\theta_i)$ από τη σχέση (1.31) η σχέση (1.37) γράφεται

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{\alpha(\varphi)}{V(\mathrm{Y}_i)}. \quad (1.39)$$

Τέλος, γνωρίζουμε ότι

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}. \quad (1.40)$$

Συνεπώς,

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

Θέτοντας $U_j = 0\,, j = 0,\cdots,k$ έχουμε ένα σύστημα $(k+1)$ εξισώσεων με $(k+1)$ άγνωστες παραμέτρους του μοντέλου. Αυτές οι εξισώσεις καλούνται εξισώσεις score της μέγιστης πιθανοφάνειας.

Στη συνέχεια, μπορούμε να εκτιμήσουμε τα β_j με τη μέθοδο Newton-Raphson μέσω του επαναληπτικού σχήματος

$$\boldsymbol{b}^{(m)} = \boldsymbol{b}^{(m-1)} - \frac{\boldsymbol{U}^{(m-1)}}{\boldsymbol{U}'^{(m-1)}}\,, \quad (1.41)$$

όπου $\boldsymbol{b}^{(m)}$ η νέα εκτίμηση στο $m-$οστό βήμα της επαναληπτικής διαδικασίας, η οποία υπολογίζεται ως συνάρτηση των εκτιμήσεων των συντελεστών $\boldsymbol{b}^{(m-1)}$ του προηγούμενου $(m-1)$ βήματος, των τιμών των συναρτήσεων score $\boldsymbol{U}^{(m-1)} = [U_0, U_1, ..., U_k]^{\mathrm{T}}$ υπολογισμένες στο $(m-1)$ βήμα, καθώς και του $p \times p$ πίνακα $\boldsymbol{U}'^{(m-1)}$ των δευτέρων μερικών παραγώγων με στοιχείο στην j γραμμή και r στήλη $\frac{\partial^2 l}{\partial b_j \partial b_r}$ υπολογισμένο στο $(m-1)$ βήμα.

Συνήθως, προσεγγίζουμε τον πίνακα $\boldsymbol{U}'$ με τον πίνακα πληροφορίας $\boldsymbol{J}(\boldsymbol{b})$, του οποίου τα στοιχεία είναι οι αρνητικές αναμενόμενες τιμές των δευτέρων μερικών παραγώγων του λογαρίθμου της συνάρτησης πιθανοφάνειας.

Γνωρίζοντας ότι ο πίνακας πληροφορίας εκφράζεται ως:

$$J_{jr} = E[U_j U_r] = \sum_{i=1}^{n} \frac{\mathrm{E}\left[(y_i - \mu_i)^2\right] x_{ij} x_{ir}}{[V(\mathrm{Y}_i)]^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2, \; j, r = 0, 1, ..., k \quad (1.42)$$

Οπότε μπορούμε να χρησιμοποιήσουμε τον ακόλουθο επαναληπτικό τύπο:

$$\boldsymbol{b}^{(m)} = \boldsymbol{b}^{(m-1)} + \left(\boldsymbol{J}^{(m-1)}\right)^{-1} \boldsymbol{U}^{(m-1)}\,, \quad (1.43)$$

όπου $\left(\boldsymbol{J}^{(m-1)}\right)^{-1}$ είναι ο αντίστροφος του πίνακα πληροφορίας και $\boldsymbol{U}^{(m-1)}$ είναι το διάνυσμα των στοιχείων $\boldsymbol{U} = \left(U_0, U_1, \cdots, U_k\right)$ υπολογισμένα κατά την $\left(m-1\right)-$ επανάληψη.

Συνεπώς, πολλαπλασιάζοντας και τις δύο πλευρές της σχέσης (1.43) με την τρέχουσα εκτίμηση του $\boldsymbol{J}^{(m-1)}$, προκύπτει η ακόλουθη σχέση κατά το $m-$οστό βήμα $\boldsymbol{J}^{(m-1)}\boldsymbol{b}^{(m)} = \boldsymbol{J}^{(m-1)}\boldsymbol{b}^{(m-1)} + \boldsymbol{U}^{(m-1)}$. (1.44)

Από την (1.42) ο πίνακας $\boldsymbol{J}$ μπορεί να γραφεί ως

$$\boldsymbol{J} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}, (1.45)$$

όπου $\boldsymbol{W} = diag\left(w_{ii}\right)$ και $w_{ii} = \dfrac{1}{V\left(\mathrm{Y}_i\right)}\left(\dfrac{\partial \mu_i}{\partial \eta_i}\right)^2 \; i = 1, 2, ..., n$.

Συνεπώς, $\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(m\text{-}1)}\boldsymbol{X}\right]\boldsymbol{b}^{(m)} = \left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(m\text{-}1)}\boldsymbol{X}\right]\boldsymbol{b}^{(m-1)} + \boldsymbol{U}^{(m\text{-}1)} = \left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{(m\text{-}1)}\mathbf{z}^{(m\text{-}1)}\right]$ με z_i το $i-$οστό στοιχείο του $n-$διάστατου διανύσματος $\mathbf{z}^{(m-1)}$, το οποίο δίνεται από τη σχέση

$$z_i = \sum_{r=0}^{k} x_{ir} b_r^{(m-1)} + \left(y_i - \mu_i\right)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \text{ με } \mu_i, \eta_i \text{ υπολογισμένα στο } \boldsymbol{b} = \boldsymbol{b}^{(m-1)}.$$

Άρα, ο επαναληπτικός τύπος (1.44) μπορεί να γραφεί ως $\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{W}}^{(m\text{-}1)}\boldsymbol{X}\boldsymbol{b}^{(m)} = \boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{W}}^{(m\text{-}1)}\mathbf{z}^{(m\text{-}1)}$, (1.46)

όπου $\hat{\boldsymbol{W}}$ συμβολίζουμε τον αντίστοιχο πίνακα μετά την αντικατάσταση των $\boldsymbol{\beta}$ από τις εκτιμήτριες $\boldsymbol{b}$.

Η τελική σχέση (1.46) μας θυμίζει τη μορφή των κανονικών εξισώσεων των σταθμισμένων ελαχίστων τετραγώνων του γενικού γραμμικού μοντέλου, που επιλύεται με επαναληπτικές μεθόδους. Για το λόγο αυτό, η μέθοδος εκτίμησης ονομάζεται μέθοδος επαναληπτικών σταθμισμένων ελαχίστων τετραγώνων (*iteratively weighted least squares*). Πολλά στατιστικά πακέτα, που περιλαμβάνουν συναρτήσεις για την προσαρμογή των γενικευμένων γραμμικών μοντέλων χρησιμοποιούν τον παραπάνω αλγόριθμο για την εκτίμηση των παραμέτρων. Ο αλγόριθμος τερματίζει για το διάνυσμα $\boldsymbol{b}^{(m)}$ για το οποίο ισχύει το εξής κριτήριο:

$$\left|\frac{b_i^{(m-1)} - b_i^m}{b_i^m}\right| < \delta \; , i = 1, 2, \cdots, k \, ,$$

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

1.4.4 Quasi-Πιθανοφάνεια

Σύμφωνα, με τη θεωρία, που αναπτύξαμε παραπάνω χρησιμοποιούμε τη μέθοδο μέγιστης πιθανοφάνειας για την εκτίμηση των παραμέτρων $\beta_0, \beta_1, \cdots, \beta_k$ ενός γενικευμένου γραμμικού μοντέλου. Βέβαια, η μέθοδος μέγιστης πιθανοφάνειας προϋποθέτει ότι γνωρίζουμε την κατανομή της τυχαίας μεταβλητής Y_i. Ωστόσο, σε πολλά προβλήματα στατιστικής μπορούμε να αντλήσουμε κάποιες πληροφορίες από το δείγμα μας για την κατανομή της τυχαίας μεταβλητής Y_i, αλλά δεν είναι πάντα εφικτό να ορίσουμε την συναρτησιακή της μορφή. Συνεπώς, στα πλαίσια αυτής της παρατήρησης προτάθηκε από τον Wedderburn (1974) η μέθοδος της quasi-πιθανοφάνειας (*quasi-likelihood*). Αποδεικνύεται ότι η μέθοδος έχει πολλές από τις σημαντικές ιδιότητες της συνηθισμένης μεθόδου της πιθανοφάνειας και σε αυτό οφείλεται το όνομά της.

Για να εφαρμοστεί η μέθοδος αρκεί να γνωρίζουμε τις δύο πρώτες ροπές της κατανομής της Y_i, δηλαδή

$$\mu = \mathrm{E}(\mathrm{Y}_i)$$
$$\text{και } \alpha(\varphi)V(\mu_i) = \varphi V(\mu_i) = V(\mathrm{Y}_i),$$

όπου θεωρήσαμε ότι $\alpha(\varphi) = \varphi$. Επιπλέον, θεωρούμε τις παρατηρήσεις ανεξάρτητες μεταξύ τους.

Οι συναρτήσεις score είναι:

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{(y_i - \mu_i)}{V(\mathrm{Y}_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right]$$

Και μπορούν να γραφούν σε μορφή πινάκων ως εξής

$$\boldsymbol{U} = \frac{1}{\varphi} \boldsymbol{D}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}), (1.47)$$

όπου ο πίνακας $\boldsymbol{D}$ έχει στοιχεία $d_{ij} = \frac{\partial \mu_i}{\partial \beta_j}$ και $\boldsymbol{V}^{-1}$ ο αντίστροφος πίνακας του πίνακα $\boldsymbol{V} = diag(V(\mu_1), ..., V(\mu_n))$.

Λύνοντας το σύστημα $\boldsymbol{D}^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0}$ λαμβάνουμε την εκτιμήτρια quasi-πιθανοφάνειας για το διάνυσμα $\boldsymbol{\beta}$ των παραμέτρων του μοντέλου.

1.5 Έλεγχοι Υποθέσεων

1.5.1 Εισαγωγή

Μετά την εκτίμηση των παραμέτρων του μοντέλου με κατάλληλες μεθόδους, δηλαδή την προσαρμογή του μοντέλου, όπως αναλύσαμε και παραπάνω θα πρέπει να εξετάσουμε την καταλληλότητα του μοντέλου. Να ερευνήσουμε δηλαδή τη σχέση των δεδομένων, που έχουμε παρατηρήσει με τις τιμές που προβλέπονται μέσω του μοντέλου μας.

Συχνά πραγματοποιούμε στατιστικούς ελέγχους σχετικά με τις παραμέτρους του μοντέλου, έτσι ώστε να αποφανθούμε ποιές μεταβλητές είναι οι στατιστικά σημαντικές για το μοντέλο μας. Αυτό επιτυγχάνεται με τον έλεγχο Wald, καθώς και με την ελεγχοσυνάρτηση deviance, όπως θα δούμε παρακάτω.

Επιπλέον, είναι εξίσου σημαντικό να προβούμε σε συγκρίσεις μοντέλων, ώστε να αποφανθούμε ποιο είναι το καλύτερο μοντέλο για τα δεδομένα μας και αυτό γίνεται με τη βοήθεια της ελεγχοσυνάρτησης deviance.

1.5.2 Έλεγχος Wald

Από τη θεωρία της μεθόδου μέγιστης πιθανοφάνειας, για μεγάλα δείγματα, η εκτιμήτρια μέγιστης πιθανοφάνειας $\boldsymbol{b}$ για την παράμετρο $\boldsymbol{\beta}$ ακολουθεί ασυμπτωτικά την πολυμετάβλητη Κανονική κατανομή, δηλαδή

$$\boldsymbol{b} \dot{\sim} \mathrm{N}_p\left(\boldsymbol{\beta}, \hat{\mathrm{V}}(\boldsymbol{b})\right),\ p = k+1,$$

όπου $\hat{\mathrm{V}}(\boldsymbol{b}) = \boldsymbol{J}^{-1}(\boldsymbol{b})$ η εκτιμήτρια του πίνακα διασποράς – συνδιασποράς της ασυμπτωτικής κατανομής της $\boldsymbol{b}$ και $\boldsymbol{J}(\boldsymbol{b})$ ο παρατηρούμενος πίνακας πληροφορίας με στοιχεία $J_{jr} = \sum_{i=1}^{n} \frac{\mathrm{E}\left[(y_i - \mu_i)^2\right] x_{ij} x_{ir}}{[Var(\mathrm{Y}_i)]^2} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2,\ j, r = 0, 1, \ldots, k.$

Σύμφωνα με τη θεωρία της ΕΜΠ για τα γενικευμένα γραμμικά μοντέλα, ο παρατηρούμενος πίνακας πληροφορίας είναι

$$\boldsymbol{J}(\boldsymbol{b}) = \mathbf{X}^{\mathrm{T}} \hat{\mathbf{W}} \mathbf{X}.$$

Άρα, οι εκτιμήτριες μέγιστης πιθανοφάνειας $\boldsymbol{b}$ ακολουθούν ασυμπτωτικά την Κανονική κατανομή με μέση τιμή $\boldsymbol{\beta}$ και πίνακα διασποράς - συνδιασποράς $\mathbf{X}^{\mathrm{T}} \hat{\mathbf{W}} \mathbf{X}$, δηλαδή $\boldsymbol{b} \dot{\sim} \mathrm{N}_p\left(\boldsymbol{\beta}, \left(\mathbf{X}^{\mathrm{T}} \hat{\mathbf{W}} \mathbf{X}\right)^{-1}\right),\ p = k+1.$

Συνεπώς, για τον έλεγχο Wald μιας μηδενικής υπόθεσης, έστω $H_0 : \beta_j = 0$ με εναλλακτική $H_1 : \beta_j \neq 0$ χρησιμοποιούμε την ελεγχοσυνάρτηση Wald που ακολουθεί προσεγγιστικά την Κανονική κατανομή

$$Z = \frac{b_j - \beta_j}{\left(J^{-1}(\boldsymbol{b})_{jj}\right)^{\frac{1}{2}}} \dot{\sim} N(0,1) \ ,$$

όπου $J^{-1}(\boldsymbol{b})_{jj}$ το $j-$ οστό διαγώνιο στοιχείο του αντιστρόφου του παρατηρούμενου πίνακα πληροφορίας $\boldsymbol{J}\left(\boldsymbol{b}\right)$.

Στην περίπτωση, που δεν απορρίψουμε την μηδενική υπόθεση, δηλαδή θεωρήσουμε ότι $\beta_j = 0$, τότε η μεταβλητή X_j δεν συμβάλλει σημαντικά στην πρόβλεψη της Y και μπορεί να αφαιρεθεί από το μοντέλο.

Η παραπάνω στατιστική συνάρτηση Z μπορεί να χρησιμοποιηθεί και για την κατασκευή ενός $100(1-a)\%-$διαστήματος εμπιστοσύνης για την παράμετρο β_j $b_j \pm z_{a/2} se(b_j)$, όπου με $z_{a/2}$ συμβολίζουμε το άνω $100(a/2)-$ποσοστιαίο σημείο της $N(0,1)$ κατανομής και $se\left(b_j\right) = \left(\hat{V}\left(b_j\right)\right)^{1/2} = J^{-1}(\boldsymbol{b})_{jj}^{\ 1/2}$.

1.6 Έλεγχοι Καλής Προσαρμογής

Ένα πολύ σημαντικό θέμα στην ανάλυση ενός μοντέλου είναι η αποτίμηση της προσαρμογής του μοντέλου στα δεδομένα, δηλαδή κατά πόσο το προσαρμοσμένο μοντέλο αποκλίνει ή όχι από τα δεδομένα που έχουμε παρατηρήσει. Αυτό επιτυγχάνεται με τους ελέγχους καλής προσαρμογής. Στα πλαίσια των γενικευμένων γραμμικών μοντέλων θα αναλύσουμε την τεχνική του λόγου των πιθανοφανειών για τον έλεγχο υποθέσεων με τις παραμέτρους του μοντέλου.

1.6.1 Ελεγχοσυνάρτηση του Λόγου των Πιθανοφανειών

Ένας τρόπος να ελέγξουμε την καταλληλότητα ενός μοντέλου, έστω M_0 είναι να το συγκρίνουμε με ένα πιο γενικό μοντέλο, έστω M_S, το οποίο θα περιλαμβάνει το μέγιστο αριθμό παραμέτρων που μπορούν να εκτιμηθούν. Αυτό καλείται κορεσμένο μοντέλο (*saturated model*).

Αν θεωρήσουμε n παρατηρήσεις y_i $i = 1,2,...,n$, έτσι ώστε η γραμμική προβλέπουσα (*linear predictor*) $\boldsymbol{x}_i^T \boldsymbol{\beta}$ να λαμβάνει διαφορετική τιμή για κάθε μία από αυτές, τότε το μοντέλο M_S θα έχει αριθμό παραμέτρων ίσο με n (δηλαδή ίσο με

το μέγεθος του δείγματος). Το μοντέλο αυτό καλείται επίσης μέγιστο μοντέλο (*maximal or full model*). Ωστόσο, αν κάποιες παρατηρήσεις έχουν την ίδια γραμμική προβλέπουσα, τότε ο μέγιστος αριθμός παραμέτρων, που μπορεί να εκτιμηθούν για το κορεσμένο μοντέλο αντιστοιχεί στον αριθμό των διακεκριμένων γραμμικών προβλεπουσών, δηλαδή $p_S \le n$. Γενικότερα, αν θεωρήσουμε p_S τον μέγιστο αριθμό παραμέτρων προς εκτίμηση, $\boldsymbol{\beta}_S$ το διάνυσμα των παραμέτρων για το κορεσμένο μοντέλο, καθώς και $\boldsymbol{b}_S$ το διάνυσμα των εκτιμητριών μέγιστης πιθανοφάνειας και αντίστοιχα για το μοντέλο M_0 θα είναι $p_0 \le p_S \le n$ ο αριθμός των παραμέτρων του μοντέλου και $\boldsymbol{\beta}_0$ και $\boldsymbol{b}_0$ το διάνυσμα των συντελεστών και των εκτιμώμενων συντελεστών για το μοντέλο M_0.

Για τα γενικευμένα γραμμικά μοντέλα, τα δύο μοντέλα θα πρέπει να ακολουθούν την ίδια κατανομή από την Εκθετική Οικογένεια Κατανομών και να έχουν την ίδια συνάρτηση σύνδεσης.

Για να ελέγξουμε λοιπόν την προσαρμογή του υποψήφιου μοντέλου συγκρίνουμε τις μεγιστοποιημένες συναρτήσεις πιθανοφάνειας αυτών.

Συνεπώς, για το μοντέλο M_0 έχουμε ότι

$$\max_{\boldsymbol{\theta}\in \mathrm{M}_0} L(\boldsymbol{\theta};\boldsymbol{y}) = L(\boldsymbol{b}_0;\boldsymbol{y}).$$

Για το κορεσμένο μοντέλο M_S

$$\max_{\boldsymbol{\theta}\in \mathrm{M}_s} L(\boldsymbol{\theta};\boldsymbol{y}) = L(\boldsymbol{b}_s;\boldsymbol{y}).$$

Η συνάρτηση Πιθανοφάνειας για το κορεσμένο μοντέλο υπολογισμένη για τα $\boldsymbol{b}_S, L(\boldsymbol{b}_s;\boldsymbol{y})$ θα είναι σίγουρα μεγαλύτερη από κάθε άλλη συνάρτηση πιθανοφάνειας για τις συγκεκριμένες παρατηρήσεις, καθώς αποτελεί την καλύτερη περιγραφή που μπορούμε να έχουμε στα δεδομένα μας. Θεωρώντας $L(\boldsymbol{b}_0;\boldsymbol{y})$ να είναι η συνάρτηση πιθανοφάνειας για το μοντέλο M_0, δηλαδή υπολογισμένη για τα $\boldsymbol{b}_0$, τότε η ελεγχοσυνάρτηση του λόγου των πιθανοφανειών (*likelihood ratio test*)

$$\lambda = \frac{L(\boldsymbol{b}_s;\boldsymbol{y})}{L(\boldsymbol{b}_0;\boldsymbol{y})}\ ,\ 0 \le \lambda \le 1$$

είναι ένας τρόπος αποτίμησης της καταλληλότητας του μοντέλου. Ωστόσο, συνηθέστερα χρησιμοποιείται το διπλάσιο του λογαρίθμου του λόγου των πιθανοφανειών, δηλαδή

$$D = 2\log\lambda = 2\left\{\log L(\boldsymbol{b_s};\boldsymbol{y}) - \log L(\boldsymbol{b}_0;\boldsymbol{y})\right\} = 2(l_s - l_0)$$

όπου l_s ο λογάριθμος του $L(\boldsymbol{b}_s;\boldsymbol{y})$ και αντίστοιχα l_0 ο λογάριθμος του $L(\boldsymbol{b}_0;\boldsymbol{y})$. Από τη θεωρία πιθανοτήτων γνωρίζουμε ότι η ελεγχοσυνάρτηση D ακολουθεί προσεγγιστικά την X^2 κατανομή με d βαθμούς ελευθερίας, δηλαδή $2\log\lambda = 2\left[l_s - l_0\right] \dot{\sim} \mathrm{X}_d^2$ ασυμπτωτικά, όπου $d = p_s - p_0$ η διαφορά των διαστάσεων των παραμετρικών χώρων.

Η παραπάνω ελεγχοσυνάρτηση καλείται deviance και ο όρος αυτός προτάθηκε από τους Nelder και Wedderburn. Αν η D λαμβάνει μεγάλες τιμές, τότε σημαίνει ότι το απλούστερο μοντέλο M_0 δεν εκφράζει επαρκώς τα δεδομένα, συγκριτικά με το κορεσμένο μοντέλο M_s. Αντίθετα, μικρές τιμές της deviance δηλώνουν ότι το υποψήφιο μοντέλο μας έχει ικανοποιητική προσαρμογή στα δεδομένα μας.

1.6.2 Ελεγχοσυνάρτηση Deviance για το Κανονικό Γραμμικό Μοντέλο.

Θεωρούμε το μοντέλο $\mathrm{Y}_i \sim \mathrm{N}(\mu_i,\sigma^2)$, $i=1,2,...,n$, όπου Y_i ανεξάρτητες τυχαίες μεταβλητές. Η δομή του μοντέλου εκφράζεται από τη σχέση

$\mathrm{E}(\mathrm{Y}_i) = \mu_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, $i=1,2,...,n$.

Ο λογάριθμος της συνάρτησης πιθανοφάνειας της κανονικής κατανομής είναι

$$l(\boldsymbol{\beta};\boldsymbol{y}) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2 - \frac{1}{2}n\log(2\pi\sigma^2).$$

Αν θεωρήσουμε η_i διακεκριμένες γραμμικές προβλέπουσες θα έχουμε n το πλήθος παραμέτρους για το κορεσμένο μοντέλο. Σύμφωνα με τη μέθοδο μέγιστης πιθανοφάνειας η εκτιμήτρια για κάθε μ_i θα είναι $\hat{\mu}_i = y_i$. Συνεπώς, η μέγιστη τιμή του λογαρίθμου της πιθανοφάνειας για το κορεσμένο μοντέλο είναι

$$l(\boldsymbol{b}_{\mathrm{s}};\boldsymbol{y}) = -\frac{1}{2}n\log(2\pi\sigma^2).$$

Για το μοντέλο, που μας ενδιαφέρει με αριθμό παραμέτρων, έστω $p_0 < n$ η εκτιμήτρια για κάθε μ_i θα είναι $\hat{\mu}_i = \hat{y}_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}$. Άρα, η μέγιστη τιμή του λογαρίθμου της πιθανοφάνειας για το μοντέλο θα είναι

$$l(\boldsymbol{b};\boldsymbol{y}) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b})^2 - \frac{1}{2}n\log(2\pi\sigma^2).$$

Η ελεγχοσυνάρτηση deviance είναι:

$$D = 2[l(\boldsymbol{b}_{\mathrm{s}};\boldsymbol{y}) - l(\boldsymbol{b};\boldsymbol{y})] = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i-\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}\right)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i-\hat{\mu}_i\right)^2.$$

Από τη θεωρία του γενικού γραμμικού μοντέλου γνωρίζουμε ότι η εκτιμήτρια $\boldsymbol{b}$ για το διάνυσμα των παραμέτρων $\boldsymbol{\beta}$ δίνεται από τη σχέση $\boldsymbol{b} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$.

1.7 Σύγκριση Μοντέλων με Χρήση της Deviance

Συνήθως, στην ανάλυση δεδομένων σκοπός μας είναι να βρούμε το καλύτερο μοντέλο μεταξύ άλλων, δηλαδή εκείνο που περιγράφει καλύτερα τα δεδομένα μας και όχι να ελέγξουμε την προσαρμογή ενός συγκεκριμένου μοντέλου. Επομένως, θεωρούμε δύο μοντέλα M_0 και M_1, τα οποία σύμφωνα με τη θεωρία των γενικευμένων γραμμικών μοντέλων θα πρέπει να πληρούν τις παρακάτω προϋποθέσεις:

- Οι τυχαίες μεταβλητές Y_i να ακολουθούν την ίδια Κατανομή από την Εκθετική Οικογένεια Κατανομών.
- Τα μοντέλα να έχουν την ίδια συνάρτηση σύνδεσης.
- Το σύνολο των επεξηγηματικών μεταβλητών του μοντέλου M_0 να αποτελεί υποσύνολο του συνόλου των επεξηγηματικών μεταβλητών του μοντέλου M_1.

Με άλλα λόγια, πρόκειται για τη σύγκριση δύο εμφωλευμένων ή ιεραρχικών μοντέλων (*nested or hierarchical models*) ως προς την προσαρμογή τους στα δεδομένα που διαθέτουμε.

Έστω, η μηδενική υπόθεση

$$\mathrm{H}_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{p_0} \end{bmatrix},$$

η οποία αντιστοιχεί στο μοντέλο M_0 και εναλλακτική την

$$\mathrm{H}_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{p_1} \end{bmatrix},$$

η οποία αντιστοιχεί στο μοντέλο M_1.

Από την παραπάνω θεωρία γνωρίζουμε ότι η ελεγχοσυνάρτηση deviance για κάθε ένα από τα παραπάνω μοντέλα είναι :

$$D_0 = 2\left(l_s - l_0\right) \text{ και } D_1 = 2\left(l_s - l_1\right).$$

Συνεπώς, για τον έλεγχο της μηδενικής υπόθεσης H_0 έναντι της εναλλακτικής H_1 χρησιμοποιούμε τη διαφορά των συναρτήσεων deviance

$$\Delta D = D_0 - D_1 = 2\left[l\left(\boldsymbol{b}_s;\boldsymbol{y}\right) - l\left(\boldsymbol{b}_0;\boldsymbol{y}\right)\right] - 2\left[l\left(\boldsymbol{b}_s;\boldsymbol{y}\right) - l\left(\boldsymbol{b}_1;\boldsymbol{y}\right)\right] = 2\left[l\left(\boldsymbol{b}_1;\boldsymbol{y}\right) - l\left(\boldsymbol{b}_0;\boldsymbol{y}\right)\right].$$

Αν τα δύο μοντέλα έχουν καλή προσαρμογή στα δεδομένα, τότε $D_0 \dot{\sim} \mathrm{X}^2\left(p_s - p_0\right)$ και $D_1 \dot{\sim} \mathrm{X}^2\left(p_s - p_1\right)$, άρα $\Delta D \dot{\sim} \mathrm{X}^2\left(d\right)$, όπου $d = p_1 - p_0$. Η μεταβολή, λοιπόν στη deviance μεταξύ δύο εμφωλευμένων μοντέλων μπορεί να ελεγχθεί χρησιμοποιώντας την κατανομή X^2. Για μεγάλες τιμές της διαφοράς ΔD, η μηδενική υπόθεση H_0 απορρίπτεται και έτσι δεχόμαστε το μοντέλο M_1. Συνεπώς, η διαφορά ΔD θα είναι πάντα θετική, καθώς το μοντέλο M_0 θα έχει μεγαλύτερη απόκλιση από το κορεσμένο μοντέλο συγκριτικά με το μοντέλο M_1 που θα έχει μικρότερη απόκλιση, αφού περιέχει περισσότερες μεταβλητές.

Σημειώνεται ότι ενώ η χρήση της κατανομής X^2 δικαιολογείται μόνο ασυμπτωτικά (για μεγάλα δείγματα), η προσέγγιση είναι πολύ καλύτερη όταν αξιολογούμε μεταβολές των τιμών της ελεγχοσυνάρτησης deviance από ότι όταν εξετάζουμε μεμονωμένα τις τιμές αυτές.

1.8 Διαγνωστικοί Έλεγχοι

Για τον προσδιορισμό της καταλληλότητας ενός γενικευμένου γραμμικού μοντέλου χρησιμοποιούνται διάφορες διαγνωστικές μέθοδοι ανάλογες με αυτές που χρησιμοποιούνται στο γραμμικό μοντέλο παλινδρόμησης. Αυτές βασίζονται στα υπόλοιπα.

1.8.1 Υπόλοιπα

Τα υπόλοιπα αποτελούν ένα μέτρο της απόκλισης των προσαρμοσμένων τιμών της μεταβλητής απόκρισης $\hat{\mu}_i$ και των παρατηρήσεων αυτής y_i. Βέβαια, θα πρέπει να σημειώσουμε ότι υπάρχουν διαφοροποιήσεις σε σχέση με την περίπτωση του γραμμικού μοντέλου, καθώς η διασπορά συνήθως δεν είναι σταθερή. Στη γραμμική παλινδρόμηση χρησιμοποιούνται τα συνήθη υπόλοιπα $y_i - \hat{\mu}_i$, έτσι ώστε να εξεταστεί αν παραβιάζεται η υπόθεση της ομοσκεδαστικότητας. Ωστόσο, στα γενικευμένα γραμμικά μοντέλα, τα συνήθη υπόλοιπα $y_i - \hat{\mu}_i, i = 1,...,n$ δεν είναι

συγκρίσιμα, διότι οι διασπορές τους είναι άνισες. Για το λόγο αυτό, στα πλαίσια της θεωρίας των γενικευμένων γραμμικών μοντέλων χρησιμοποιούνται κάποια άλλα είδη υπολοίπων. Τα πιο συνηθισμένα είναι τα υπόλοιπα deviance, τα υπόλοιπα Pearson και τα υπόλοιπα της πιθανοφάνειας.

- **Υπόλοιπα deviance**

Το υπόλοιπο deviance ορίζεται ως η τετραγωνική ρίζα της συνεισφοράς της i – οστής παρατήρησης στην ελεγχοσυνάρτηση deviance D και δίνεται από τη σχέση

$$\varepsilon_i^D = sign(y_i - \hat{\mu}_i)\sqrt{d_i(y_i, \hat{\mu}_i)}\,,\ i = 1,2,...,n\,,\ (1.48)$$

όπου $d_i(y_i, \hat{\mu}_i)$ είναι η τιμή της ελεγχοσυνάρτησης deviance υπολογισμένη στην i – οστή παρατήρηση. Η συνάρτηση $sign(y_i - \hat{\mu}_i)$ υπολογίζει το πρόσημο της διαφοράς $y_i - \hat{\mu}_i$, δηλαδή δίνει θετικό πρόσημο στην ποσότητα $\sqrt{d_i(y_i, \hat{\mu}_i)}$, αν $y_i > \hat{\mu}_i$ και αρνητικό αν $y_i < \hat{\mu}_i$, αντίστοιχα. Το άθροισμα τετραγώνων των υπολοίπων deviance ε_i^D, $\sum_{i=1}^{n}\left(\varepsilon_i^D\right)^2$ ισούται με την ελεγχοσυνάρτηση D.

Επιπλέον, ορίζουμε τα τυποποιημένα υπόλοιπα deviance ως

$$\varepsilon_i^{DS} = \frac{\varepsilon_i^D}{\sqrt{1 - \hat{h}_{ii}}}\,,\ (1.49)$$

όπου $\hat{h}_{ii}$ είναι το i – οστό διαγώνιο στοιχείο του πίνακα $\hat{\mathbf{H}}$ (*hat matrix*)

$$\hat{\mathbf{H}} = \hat{\mathbf{W}}^{1/2}\mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\hat{\mathbf{W}}\mathbf{X}\right)\mathbf{X}^{\mathbf{T}}\hat{\mathbf{W}}^{1/2}\,,\ (1.50)$$

όπου $\hat{\mathbf{W}}$ είναι ο σταθμισμένος πίνακας του προσαρμοσμένου μοντέλου με στοιχεία $\hat{\mathbf{W}} = diag\left(\hat{w}_{ii}\right)$ και $w_{ii} = \frac{1}{\hat{V}\left(\mathrm{Y}_i\right)}\left(\frac{\partial \hat{\mu}_i}{\partial \eta_i}\right)^2 = \frac{1}{\hat{V}\left(\mathrm{Y}_i\right)(g'(\hat{\mu}_1))^2}\ i = 1,2,...,n$.

- **Υπόλοιπα Pearson**

Η γενική μορφή των υπολοίπων Pearson είναι

$$\varepsilon_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}},\ i = 1,2,...,n\,,\ (1.51)$$

όπου $V(\hat{\mu}) = b''(\hat{\theta})$ είναι η συνάρτηση διασποράς. Παρατηρούμε ότι το άθροισμα τετραγώνων των υπολοίπων Pearson ισούται με την ελεγχοσυνάρτηση X^2 του Pearson.

Τα τυποποιημένα υπόλοιπα Pearson ορίζονται ως

$$\varepsilon_i^{PS} = \frac{\varepsilon_i^P}{\sqrt{1-\hat{h}_{ii}}}, i = 1,2,...,n \text{ , (1.52)}$$

όπου $\hat{h}_{ii}$ αποτελεί το $i-$ οστό διαγώνιο στοιχείο του πίνακα $\hat{\mathbf{H}}$, όπως ορίζεται από τη σχέση (1.50).

- **Υπόλοιπα της πιθανοφάνειας**

Τα υπόλοιπα πιθανοφάνειας (likelihood residuals) ορίζονται ως:

$$\varepsilon_i^L = sign(y_i - \hat{\mu}_i)\sqrt{\hat{h}_{ii}(\varepsilon_i^{PS})^2 + (1-\hat{h}_{ii})(\varepsilon_i^{DS})^2}, i = 1,2,...,n. \text{ (1.53)}$$

1.8.2 Χρήση των Υπολοίπων

Τα υπόλοιπα, όπως και στο πολλαπλό γραμμικό μοντέλο, χρησιμοποιούνται κυρίως για τον γραφικό έλεγχο της καταλληλότητας του μοντέλου. Δημιουργώντας κατάλληλα γραφήματα των υπολοίπων έναντι διαφόρων δεικτών μπορούμε να ελέγξουμε τις προϋποθέσεις ενός γενικευμένου γραμμικού μοντέλου.

Ο πιο απλός τρόπος είναι να αναπαραστήσουμε γραφικά τα υπόλοιπα (*index plot*) ως προς τη σειρά των παρατηρήσεων στο αρχείο δεδομένων. Η παρουσία ασυνήθιστα μεγάλων υπολοίπων υποδεικνύει ότι το μοντέλο ίσως δεν είναι ικανοποιητικό λόγω της παρουσίας έκτροπων παρατηρήσεων. Στην περίπτωση, που η σειρά των παρατηρήσεων έχει κάποιο νόημα με βάση το δείγμα μας, δηλαδή για παράδειγμα αν οι παρατηρήσεις δίνονται σε χρονική σειρά, το ίδιο γράφημα μπορεί να δείξει την παρουσία συσχέτισης των υπολοίπων. Συγκεκριμένα με αυτό το γράφημα ελέγχουμε ουσιαστικά την υπόθεση της ανεξαρτησίας των τυχαίων μεταβλητών απόκρισης Y_i $(i = 1,...,n)$. Αν τα υπόλοιπα δεν παρουσιάζουν κάποια ιδιαίτερη συμπεριφορά και κατανέμονται τυχαία γύρω από το μηδέν δεχόμαστε ότι ευσταθεί η υπόθεση της ανεξαρτησίας των μεταβλητών απόκρισης.

Επίσης, τα γραφήματα των υπολοίπων σε σχέση με τις προσαρμοσμένες τιμές $\hat{\mu}_i$ ή με την γραμμική προβλέπουσα $\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}$ ή σε σχέση με κάθε επεξηγηματική μεταβλητή χωριστά, μπορούν να αποβούν χρήσιμα στην εξέταση της ορθότητας του συστηματικού μέρους του μοντέλου. Ειδικότερα, τα γραφήματα αυτά αποτελούν διαγράμματα διασποράς και δεν θα πρέπει να υποδηλώνουν μια συστηματική συμπεριφορά, δηλαδή η δεσμευμένη κατανομή των υπολοίπων δεν θα πρέπει να μεταβάλλεται σε σχέση με τις εκτιμώμενες τιμές ή σε σχέση με την κάθε επεξηγηματική μεταβλητή του μοντέλου. Αν μια επεξηγηματική μεταβλητή είναι

κατηγορική, τότε αντί για γράφημα των υπολοίπων σε σχέση με αυτή τη μεταβλητή δημιουργούμε μία σειρά από θηκογράμματα των υπολοίπων, ένα για κάθε κατηγορία. Τα «κουτιά» θα πρέπει να έχουν περίπου τον ίδιο μέσο και το ίδιο διάνοιγμα, ώστε να είναι το μοντέλο ικανοποιητικό για να περιγράψει τα δεδομένα μας. Σε περίπτωση λοιπόν που διαπιστώσουμε την ύπαρξη συστηματικών χαρακτηριστικών, αυτό θα οφείλεται στο ότι δεν ευσταθεί μία ή περισσότερες προϋποθέσεις του μοντέλου και πιθανώς θα πρέπει να συμπεριλάβουμε νέες επεξηγηματικές μεταβλητές στο μοντέλο ή να μετασχηματίσουμε μια ήδη υπάρχουσα επεξηγηματική μεταβλητή. Όλες αυτές οι γραφικές παραστάσεις χρησιμεύουν, εκτός των άλλων και στον εντοπισμό έκτροπων ή άτυπων σημείων (*outliers*) στα δεδομένα.

Τέλος, ένα άλλο χρήσιμο γραφικό εργαλείο για τα υπόλοιπα είναι το διάγραμμα Q-Q (*Q-Q plot*), δηλαδή το γράφημα των ποσοστημορίων, που αντιστοιχούν στα υπόλοιπα σε σχέση με τα ποσοστημόρια, που αντιστοιχούν στην Κανονική κατανομή. Το γράφημα αυτό χρησιμοποιείται περισσότερο σαν ένδειξη της καλής προσαρμογής του μοντέλου (κυρίως ως προς τον εντοπισμό έκτροπων παρατηρήσεων) παρά σαν ένδειξη της κανονικότητας των υπολοίπων.

1.8.3 Σημεία Επιρροής

Πολλές φορές, υπάρχουν παρατηρήσεις στα δεδομένα, που έχουν μεγάλη επιρροή στις εκτιμήτριες των παραμέτρων των μοντέλων, και κατά συνέπεια στην προσαρμογή του μοντέλου. Συγκεκριμένα, μικρές αλλαγές αυτών των παρατηρήσεων ή και αφαίρεση αυτών μπορεί να επηρεάσει δραστικά την προσαρμογή του μοντέλου. Αυτές οι παρατηρήσεις καλούνται σημεία επιρροής. Γενικώς, ο εντοπισμός και η εξέταση αυτών πραγματοποιούνται μέσω διαφόρων διαγνωστικών μέτρων και γραφημάτων.

- **Απόσταση του Cook**

Ένα πολύ χρήσιμο μέτρο για τον εντοπισμό σημείων επιρροής είναι η απόσταση Cook. Πιο συγκεκριμένα, εκφράζει την αλλαγή στην εκτίμηση των παραμέτρων του μοντέλου, όταν μια συγκεκριμένη παρατήρηση αφαιρεθεί.

Η στατιστική συνάρτηση του Cook δίνεται από τη σχέση

$$CD_i = \frac{1}{p}\left(\boldsymbol{b}_{(i)} - \boldsymbol{b}\right)^T J(\boldsymbol{b})(\boldsymbol{b}_{(i)} - \boldsymbol{b}), \quad (1.54)$$

όπου p ο αριθμός των παραμέτρων του μοντέλου, $\boldsymbol{b}_{(i)}$ το διάνυσμα των εκτιμητριών των συντελεστών του μοντέλου, όταν παραλείπεται η $i-$ παρατήρηση και $\boldsymbol{b}$ το διάνυσμα εκτιμητριών των συντελεστών του μοντέλου, όταν χρησιμοποιείται όλο το δείγμα. Επίσης, $J(\boldsymbol{b}) = \boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}$ είναι η παρατηρούμενη πληροφορία κατά Fisher.

Η συνάρτηση CD_i μπορεί να προσεγγιστεί από την πιο απλή μορφή

$$CD_i \approx \frac{\hat{h}_{ii}(\varepsilon_{\mathrm{i}}^{\mathrm{PS}})^2}{p(1-\hat{h}_{ii})}, \quad (1.55)$$

όπου $\hat{h}_{ii}$ το $i-$οστό διαγώνιο στοιχείο του πίνακα $\hat{\mathbf{H}}$ (*hat matrix*) $\hat{\mathbf{H}} = \hat{\mathbf{W}}^{1/2}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\hat{\mathbf{W}}\mathbf{X}\right)\mathbf{X}^{\mathrm{T}}\hat{\mathbf{W}}^{1/2}$.

Η στατιστική συνάρτηση του Cook, CD_i αποτελεί έναν συνδυασμό των υπολοίπων $\varepsilon_{\mathrm{i}}^{\mathrm{PS}}$ και των τιμών $\hat{h}_{ii}$.

Επιπλέον, υπάρχει και η αποκαλούμενη τροποποιημένη στατιστική συνάρτηση του Cook, που ορίζεται ως

$$C_i = \left|\varepsilon_i^L\right|\sqrt{\frac{(n-p)\hat{h}_{ii}}{p(1-\hat{h}_{ii})}}, \; i = 1,2,...,n.$$

Η στατιστική συνάρτηση C_i προτιμάται έναντι της κλασσικής απόστασης Cook όσον αφορά στον εντοπισμό σημείων αυξημένης επιρροής.

- **Συνάρτηση «Δέλτα-Βήτα» (*delta-beta*)**

Συχνά επιθυμούμε να εξετάσουμε την επιρροή κάθε παρατήρησης χωριστά στην εκτίμηση της καθεμίας παραμέτρου β_j, διότι ενδέχεται κάποιες παρατηρήσεις να μην εντοπίζονται με την απόσταση Cook, καθώς πιθανώς δεν επηρεάζουν το σύνολο των παραμέτρων . Για το σκοπό αυτό χρησιμοποιούμε τη στατιστική συνάρτηση «Δέλτα-Βήτα» η οποία ορίζεται από τη σχέση

$$\Delta_i b_j = \frac{(\mathbf{X}^{\mathrm{T}}\hat{\mathbf{W}}\mathbf{X})^{-1}_{j+1}\boldsymbol{x}_i(y_i - \hat{\mu}_i)}{(1-\hat{h}_{ii})se(b_j)}, \; j = 0,1,...,k, \; i = 1,2,...,n,$$

όπου $se(b_j)$ το τυπικό σφάλμα της εκτιμήτριας b_j, $(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}_{j+1}$ είναι η $(j+1)-$οστή παρατήρηση του πίνακα διασποράς – συνδιασποράς των εκτιμημένων παραμέτρων και $\boldsymbol{x}_i$ είναι το διάνυσμα των τιμών των επεξηγηματικών μεταβλητών της $i-$ οστής παρατήρησης.

Η στατιστική αυτή συνάρτηση εκφράζει τη μεταβολή της τιμής της εκτιμήτριας b_j, αν αφαιρέσουμε την i παρατήρηση. Μεγάλες τιμές της συνάρτησης $\Delta_i b_j$ υποδηλώνουν ποιες παρατηρήσεις επηρεάζουν την εκτίμηση του υπό εξέταση συντελεστή β_j.

Γενικότερα, η εξέταση των παρατηρήσεων επιρροής στην εκτίμηση του μοντέλου μπορεί να γίνει με γραφήματα των υπολοίπων πιθανοφάνειας ή deviance ως προς τις τιμές $\hat{h}_{ii}$ ή ακόμη και με γραφήματα των υπολοίπων πιθανοφάνειας ή των τιμών $\hat{h}_{ii}$ έναντι του αριθμού των παρατηρήσεων.

1.9 Επιλογή Κατάλληλου Μοντέλου

Για την επιλογή του κατάλληλου μοντέλου, αλλά και για την σύγκριση διαφορετικών μοντέλων ως προς τη σπουδαιότητά τους χρησιμοποιούνται τα μέτρα καταλληλότητας. Πρόκειται για κάποιες αριθμητικές ποσότητες, που χρησιμοποιούνται για την αξιολόγηση ενός μοντέλου, αλλά κυρίως για την επιλογή του βέλτιστου μοντέλου μεταξύ άλλων. Θα παρουσιάσουμε κάποια από αυτά και συγκεκριμένα τα κριτήρια AIC και BIC και τους συντελεστές προσαρμογής ή συντελεστές προσδιορισμού (*coefficient of fit or coefficient of determination*).

1.9.1 Δείκτες Καλής Προσαρμογής AIC και BIC

- **Akaike's information criterion (*AIC*)**

Το AIC αποτελεί ένα κριτήριο επιλογής του βέλτιστου μοντέλου με το όσο το δυνατόν μικρότερο αριθμό παραμέτρων. Ορίζεται από τη σχέση

$$AIC = 2d - 2\log L \text{ , } (1.56)$$

όπου L η μεγιστοποιημένη τιμή της συνάρτησης πιθανοφάνειας για το εκτιμημένο μοντέλο και d ο αριθμός παραμέτρων του μοντέλου.

Συγκρίνοντας όλα τα υποψήφια μοντέλα με βάση το παραπάνω κριτήριο φαίνεται να είναι βέλτιστο εκείνο με το μικρότερο AIC. Η εισαγωγή περισσότερων παραμέτρων στο μοντέλο αυξάνει την προσαρμογή του, ανεξάρτητα αν είναι στατιστικά σημαντικές ή όχι, καθώς αυξάνει ο όρος $\log L$ με την επιπλέον πρόσθεση μεταβλητών. Ωστόσο, αυξάνεται και ο πρώτος όρος του AIC, το d , δηλαδή ο αριθμός των μεταβλητών και μειώνεται ο δεύτερος όρος του AIC. Τελικά, η εισαγωγή επιπλέον παραμέτρων στο μοντέλο μειώνει την τιμή του AIC στην περίπτωση που η προσαρμογή του μοντέλου βελτιώνεται. Η ποσότητα $2d$ καλείται ποινή (*penalty*).

- **Bayesian information criterion (*BIC*)**

Το κριτήριο BIC προτάθηκε από τον Schwarz (1978) και ορίζεται από τη σχέση

$$BIC = d \log n - 2 \log L \ , \ (1.57)$$

όπου L η μεγιστοποιημένη τιμή της συνάρτησης πιθανοφάνειας για το εκτιμημένο μοντέλο και d αριθμός παραμέτρων του μοντέλου και n ο αριθμός των παρατηρήσεων.

Αποτελεί ένα ακόμη κριτήριο επιλογής βέλτιστου μοντέλου ανάμεσα σε μοντέλα με διαφορετικό αριθμό παραμέτρων, όχι απαραίτητα εμφωλευμένων. Η λογική και η χρήση του είναι όμοια με το κριτήριο AIC. Ωστόσο, η διαφορά τους έγκειται στο γεγονός ότι στην περίπτωση του BIC η εισαγωγή επιπρόσθετων παραμέτρων αποθαρρύνεται σε μεγαλύτερο βαθμό από το AIC.

Σύμφωνα με την κλίμακα του Raftery μπορούμε να κρίνουμε κατά πόσο ένα μοντέλο έχει καλύτερη προσαρμογή στα δεδομένα σε σχέση με ένα άλλο λαμβάνοντας την απόλυτη τιμή της διαφοράς των τιμών, που λαμβάνουν τα κριτήρια BIC για δύο μοντέλα (*Hardin and Hilbe*, 2007). Η κλίμακα αυτή αναγράφεται στον παρακάτω πίνακα.

Διαφορά των τιμών BIC	**Ένδειξη**
$0-2$	Ασθενής (Weak)
$2-8$	Θετική (Positive)
$6-10$	Ισχυρή (Strong)
>10	Πολύ Ισχυρή (Very Strong)

Πίνακας 4: *Κλίμακα τιμών του κριτηρίου BIC.*

1.9.2 Συντελεστές Προσδιορισμού

Γενικότερα, οι συντελεστές προσδιορισμού εκφράζουν το ποσοστό μεταβλητότητας της μεταβλητής απόκρισης, που εξηγείται από το μοντέλο που προσαρμόζεται σε σχέση με ένα άλλο ιεραρχικό μοντέλο ή με ένα μοντέλο, που περιλαμβάνει μόνο το σταθερό όρο. Για αυτό το λόγο δεν εξετάζουν την προσαρμογή ενός μοντέλου, αλλά αποτελούν έναν δείκτη για την επιλογή του κατάλληλου μοντέλου μεταξύ δύο εμφωλευμένων μοντέλων.

Στην περίπτωση του απλού γραμμικού μοντέλου έτσι και στα γενικευμένα γραμμικά χρησιμοποιούμε κατάλληλους δείκτες R^2 για τον προσδιορισμό του ποσοστού της μεταβλητότητας, που περιγράφεται από το μοντέλο.

Μια έκφραση για το συντελεστή προσδιορισμού R^2 για το πολλαπλό γραμμικό μοντέλο είναι η

$R^2 = \frac{SSR}{SST}$, δηλαδή ο λόγος του αθροίσματος τετραγώνων λόγω της παλινδρόμησης προς το ολικό άθροισμα τετραγώνων.

Επειδή, το SST προκύπτει από το γραμμικό μοντέλο, που περιέχει μόνο ένα σταθερό όρο, ένα φυσικό ανάλογο του R^2 είναι το ψευδο- R^2 του McFadden (1974)

$$R_L^2 = 1 - \frac{l(\boldsymbol{b})}{l_0},$$

όπου $l(\boldsymbol{b})$ είναι η μεγιστοποιημένη λογαριθμοποιημένη πιθανοφάνεια για το μοντέλο, που μας ενδιαφέρει και l_0 είναι η μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας για το μοντέλο, που περιέχει μόνο το σταθερό όρο. Η μέγιστη τιμή του συντελεστή R_L^2 είναι 1 και προκύπτει, όταν προσαρμόζουμε ένα κορεσμένο μοντέλο.

Η τιμή αυτού του κριτηρίου είναι μηδέν για ένα μοντέλο, που δεν περιέχει επεξηγηματικές μεταβλητές και αυξάνεται με την εισαγωγή μεταβλητών.

Ένας άλλος δείκτης προσδιορισμού, που αποτελεί τροποποίηση του παραπάνω δείκτη και μπορεί να λάβει την μέγιστη τιμή 1 ακόμη και στην περίπτωση ενός άλλου μοντέλου από το κορεσμένο μοντέλο είναι ο δείκτης ψευδο- R^2

$$R_D^2 = \frac{l(\boldsymbol{b}) - l_0}{l_S - l_0},\ 0 \le R_D^2 \le 1,$$

όπου $l(\boldsymbol{b})$ είναι η μεγιστοποιημένη λογαριθμοποιημένη πιθανοφάνεια για το μοντέλο που μας ενδιαφέρει, l_0 είναι η μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας για το μοντέλο, που περιέχει μόνο το σταθερό όρο και l_S η μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας για το κερεσμένο μοντέλο.

Παρατηρούμε ότι ο δείκτης R_D^2 μπορεί να εκφραστεί χρησιμοποιώντας τις στατιστικές συναρτήσεις deviance για τα δύο μοντέλα. Έτσι, η παραπάνω σχέση μπορεί να γραφεί ως εξής

$$R_D^2 = \frac{D_0 - D_M}{D_0},$$

όπου D_0, D_M οι τιμές της ελεγχοσυνάρτησης deviance των δύο μοντέλων, το ένα με το σταθερό όρο και το άλλο με τις επεξηγηματικές μεταβλητές, αντίστοιχα.

Αποδεικνύεται ότι $R_L^2 = \left\{1 - \frac{ls}{l_0}\right\} R_D^2$.

Ένας ακόμη συντελεστής προσδιορισμού είναι ο ψευδο- R_M^2 , ο οποίος ορίζεται από τη σχέση

$$R_M^2 = 1 - \left(\frac{L_0}{L_1}\right)^{2/n},$$

όπου L_0, L_1 οι μεγιστοποιημένες συναρτήσεις πιθανοφάνειας για δύο εμφωλευμένα μοντέλα, το ένα μόνο με το σταθερό όρο και το άλλο το υποψήφιο μοντέλο.

Τέλος, ένας ακόμη συντελεστής προσδιορισμού είναι ο διορθωμένος ή προσαρμοσμένος δείκτης

$$R_N^2 = \frac{R_M^2}{\max R_M^2}.$$

Ο δείκτης αυτός είναι κατάλληλος για τα διακριτά μοντέλα, αφού ισχύει ότι μεγιστοποιημένες συναρτήσεις πιθανοφάνειας για δύο εμφωλευμένα μοντέλα, L_0, L_1 θα είναι κάτω της μονάδας, $L_0 < L_1 < 1$.

Υπάρχουν αρκετές προτάσεις και τροποποιήσεις για μέτρα τύπου R^2 στα πλαίσια των γενικευμένων γραμμικών μοντέλων. Ωστόσο, η χρήση τους δεν ενδείκνυται γενικώς για τα γενικευμένα γραμμικά μοντέλα, παρά μόνο για την περίπτωση του απλού γραμμικού μοντέλου, αφού δεν είναι κατάλληλα για την σύγκριση μοντέλων.

ΚΕΦΑΛΑΙΟ 2

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΓΙΑ ΔΙΩΝΥΜΙΚΑ ΔΕΔΟΜΕΝΑ

2.1 Δίτιμες Μεταβλητές Απόκρισης

Σε πολλές εφαρμογές η εξαρτημένη μεταβλητή παίρνει δύο μόνο τιμές 0 και 1, οι οποίες αντιστοιχούν σε δύο ενδεχόμενα, τα δύο πιθανά αποτελέσματα μιας διαδικασίας ή ενός «πειράματος». Για παράδειγμα, το αν ο ασθενής έχει καρκίνο ή όχι, το αν ο άνεργος βρίσκει δουλειά ή όχι. Οι τιμές 0 και 1 της μεταβλητής αποτελούν μια αυθαίρετη κωδικοποίηση των δύο ενδεχομένων, που εκφράζουν «επιτυχία» και «αποτυχία» με πιθανότητες π και $1 - \pi$, αντίστοιχα. Σε αυτό το κεφάλαιο, λοιπόν θα αναλύσουμε τη θεωρία των γενικευμένων γραμμικών μοντέλων για δίτιμες μεταβλητές απόκρισης.

Έστω μια δίτιμη τυχαία μεταβλητή Z:

$$Z = \begin{cases} 1, & \text{με πιθανότητα } \pi \\ 0, & \text{με πιθανότητα } 1-\pi \end{cases}.$$

Προφανώς τότε η τυχαία μεταβλητή Z ακολουθεί την κατανομή Bernoulli, δηλ. $Z \sim B(\pi)$ με μέση τιμή

$$E(Z) = 1(\pi) + 0(1 - \pi) = \pi \quad (2.1)$$

και διασπορά $V(Z) = E\{Z - E(Z)\}^2 = (1-\pi)^2 \pi + (0-\pi)^2 (1-\pi) = \pi(1-\pi)$. (2.2)

Σημειώνουμε ότι η μέση τιμή και η διασπορά της τ.μ Z εξαρτώνται από την πιθανότητα εμφάνισης επιτυχίας π. Συνεπώς, σ' αυτήν την περίπτωση η υπόθεση του γραμμικού μοντέλου φαίνεται να μην ευσταθεί, καθώς η διασπορά της μεταβλητής απόκρισης δεν είναι σταθερή.

Επεκτείνοντας σε μια σειρά από n ανεξάρτητες δοκιμές "Bernoulli", δηλαδή αν υπάρχουν n ανεξάρτητες τέτοιες τυχαίες μεταβλητές $Z_1, Z_2, \ldots, Z_n$ με $P(Z_i = 1) = \pi_i$, για $i = 1, 2, \ldots, n$ με συνάρτηση πιθανότητας για κάθε i – παρατήρηση

$$f(z_i) = \pi_i^{z_i}(1-\pi_i)^{1-z_i} \quad ,i = 1,2,\cdots,n \quad (2.3)$$

τότε η από κοινού συνάρτηση μάζας πιθανότητας ενός τυχαίου δείγματος $Z = (Z_1, Z_2, \ldots, Z_n)$ θα είναι

$$L(\pi_1,\cdots,\pi_n;z_1,\cdots,z_n) = \prod_{i=1}^{n} f(z_i) = \prod_{i=1}^{n} \pi_i^{z_i}(1-\pi_i)^{1-z_i} = \exp\left[\sum_{i=1}^{n} z_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} \ln(1-\pi_i)\right],$$
(2.4)

η οποία παρατηρούμε ότι είναι μέλος της Εκθετικής Οικογένειας Κατανομών.

Επιπλέον, υπό την υπόθεση ότι η πιθανότητα επιτυχίας π είναι ίδια σε κάθε δοκιμή μπορούμε να ορίσουμε την τυχαία μεταβλητή

$$Y = \sum_{i=1}^{n} Z_i.$$

Τότε η Y εκφράζει τον αριθμό επιτυχιών στις n ανεξάρτητες δοκιμές Bernoulli. Η τυχαία μεταβλητή Y ακολουθεί την διωνυμική κατανομή, δηλ. $Y \sim binomial(n,\pi)$ με συνάρτηση μάζας πιθανότητας

$$P(Y = y) = f(y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \; y = 0,1,2,\ldots,n \; . \quad (2.5)$$

Η μέση τιμή της Y είναι $E(Y) = n\pi$ και η διασπορά $V(Y) = n\pi(1-\pi)$. Στην ειδική περίπτωση, που $n = 1$, έχουμε **δυαδικά δεδομένα**, αλλιώς **διωνυμικά**.

Στη γενική περίπτωση, όπου έχουμε N ανεξάρτητες τυχαίες μεταβλητές $Y_1, Y_2, \ldots, Y_N$ κάθε μια από αυτές να απεικονίζει τον αριθμό επιτυχιών σε N διαφορετικές υποομάδες του πληθυσμού και $Y_i \sim binomial(n_i, \pi_i)$, τότε η συνάρτηση πιθανοφάνειας ενός δείγματος τιμών $y_1, y_2, \ldots, y_N$ με μέσες τιμές $E(y_i) = \mu_i = n_i\pi_i$ θα είναι

$$L(\pi_1,\cdots,\pi_n;y_1,y_2,\ldots,y_N) = \prod_{i=1}^{N}\binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i} \; . \quad (2.6)$$

2.2 Ο Μετασχηματισμός Logit

Το επόμενο βήμα για την επιλογή ενός κατάλληλου μοντέλου για τα δεδομένα μας, αφού έχουμε προσδιορίσει την κατανομή της μεταβλητής απόκρισης, δηλαδή το στοχαστικό μέρος του μοντέλου, είναι ο προσδιορισμός του συστηματικού μέρους του μοντέλου. Δεδομένου ότι η αναμενόμενη τιμή μιας δίτιμης μεταβλητής Y_i είναι

π_i, θέλουμε οι πιθανότητες π_i, $i = 0,1,...,n$ να εξαρτώνται από ένα διάνυσμα επεξηγηματικών μεταβλητών $x_i^{\mathrm{T}} = [x_{i0}, x_{i1}, ..., x_{ik}]$.

Η πιο απλή περίπτωση είναι να επιλέξουμε μια γραμμική σχέση μεταξύ των πιθανοτήτων και των επεξηγηματικών μεταβλητών, η οποία εκφράζεται από τη σχέση

$$\pi_i = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}, \qquad (2.7)$$

όπου $\boldsymbol{\beta}$ το διάνυσμα των παραμέτρων του μοντέλου. Το παραπάνω μοντέλο καλείται γραμμικό μοντέλο παλινδρόμησης. Φυσικά, η παραπάνω επιλογή επιφέρει βασικά προβλήματα. Ένα από αυτά είναι ότι, ενώ το π_i εκφράζει πιθανότητα και λαμβάνει τιμές στο διάστημα $[0,1]$ φαίνεται να εξισώνεται με τη σχέση $\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta}$, η οποία μπορεί να πάρει οποιαδήποτε πραγματική τιμή, αφού δεν έχουμε κανένα περιορισμό ως προς τις επεξηγηματικές μεταβλητές του μοντέλου.

Για να αποφύγουμε αυτό το πρόβλημα χρησιμοποιούμε το μετασχηματισμό logit της πιθανότητας π_i, που δίνεται από την σχέση

$$\eta_i = g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right). \quad (2.8)$$

Ο λόγος $\pi_i / 1-\pi_i$, που περικλείεται στον λογάριθμο της παραπάνω συνάρτησης καλείται συμπληρωματικές (ή σχετικές) πιθανότητες (*odds*). Πολλές φορές ο μετασχηματισμός logit (σχέση (2.8)) καλείται και λογάριθμος των συμπληρωματικών πιθανοτήτων (*log-odds*).

Ο μετασχηματισμός *logit* μπορεί να εφαρμοστεί και για τα διωνυμικά δεδομένα όπου η αναμενόμενη μέση τιμή είναι ίση με $n_i\pi_i$. Τότε, θα έχουμε την εξής σχέση

$$\mathrm{logit}(n_i\pi_i) = \log\frac{n_i\pi_i}{n_i - n_i\pi_i} = \log\frac{\pi_i}{1-\pi_i}.$$

Παρατηρούμε, λοιπόν ότι και για δυαδικά δεδομένα και για διωνυμικά το αποτέλεσμα είναι το ίδιο.

Ο αντίστροφος μετασχηματισμός του *logit* καλείται *antilogit* και μας δίνει τη δυνατότητα να εκτιμήσουμε την πιθανότητα «επιτυχίας».

Άρα, $\pi_i = \mathrm{logit}^{-1}(\eta_i) = \dfrac{e^{\eta_i}}{1+e^{\eta_i}}$.

Σύμφωνα με τη θεωρία των γενικευμένων γραμμικών μοντέλων είμαστε σε θέση να ορίσουμε το λογιστικό μοντέλο, υποθέτοντας ότι ο μετασχηματισμός logit της πιθανότητας π_i συνδέεται με τις επεξηγηματικές μεταβλητές του μοντέλου μέσω μιας γραμμικής σχέσης.

Παρατήρηση

Σημειώνουμε ότι η επιλογή ενός γραμμικού μοντέλου για να περιγράψει τα δεδομένα μας, επιφέρει και άλλα προβλήματα. Αν υποθέσουμε ότι το μοντέλο θα έχει τη μορφή (1.4), τότε δεδομένου ότι η Y_i είναι μία δίτιμη τυχαία μεταβλητή παρατηρούμε ότι τα σφάλματα ε_i μπορούν να πάρουν μόνο δύο πιθανές τιμές. Αυτές είναι οι εξής

αν $\mathrm{Y}_i = 1,$ τότε $\varepsilon_i = 1 - \pi_i$
και αν $\mathrm{Y}_i = 0,$ τότε $\varepsilon_i = -\pi_i$.

Συνεπώς, η κατανομή των σφαλμάτων είναι απίθανο να είναι κανονική. Τέλος, η διασπορά των σφαλμάτων δεν είναι σταθερή όπως επιβάλουν οι προϋποθέσεις ενός γραμμικού μοντέλου παλινδρόμησης, αφού

$$V(\varepsilon_i) = V(\mathrm{Y}_i) = \mathrm{E}\left\{\mathrm{Y}_i - \mathrm{E}\left(\mathrm{Y}_i\right)\right\}^2 = \left(1-\pi_i\right)^2 \pi_i + \left(0-\pi_i\right)^2 \left(1-\pi_i\right) = \pi_i\left(1-\pi_i\right).$$

Εν κατακλείδι, η επιλογή του γραμμικού μοντέλου για δυαδικά δεδομένα (ή διωνυμικά δεδομένα) δεν ευσταθεί. Συνεπώς, δημιουργείται η ανάγκη δημιουργίας μοντέλων κατάλληλων για την περιγραφή τέτοιου είδους δεδομένων. Ένα από αυτά είναι το λογιστικό μοντέλο.

2.3 Διωνυμικά Δεδομένα και Συναρτήσεις Σύνδεσης

Για διωνυμικά δεδομένα (μεταβλητές απόκρισης) υπάρχουν και άλλες επιλογές μοντέλων, και κατά συνέπεια συναρτήσεων σύνδεσης.

- Ένα από αυτά είναι το μοντέλο probit, το οποίο έχει συνάρτηση σύνδεσης

$$g\left(\pi_i\right) = \Phi^{-1}\left(\pi_i\right),$$

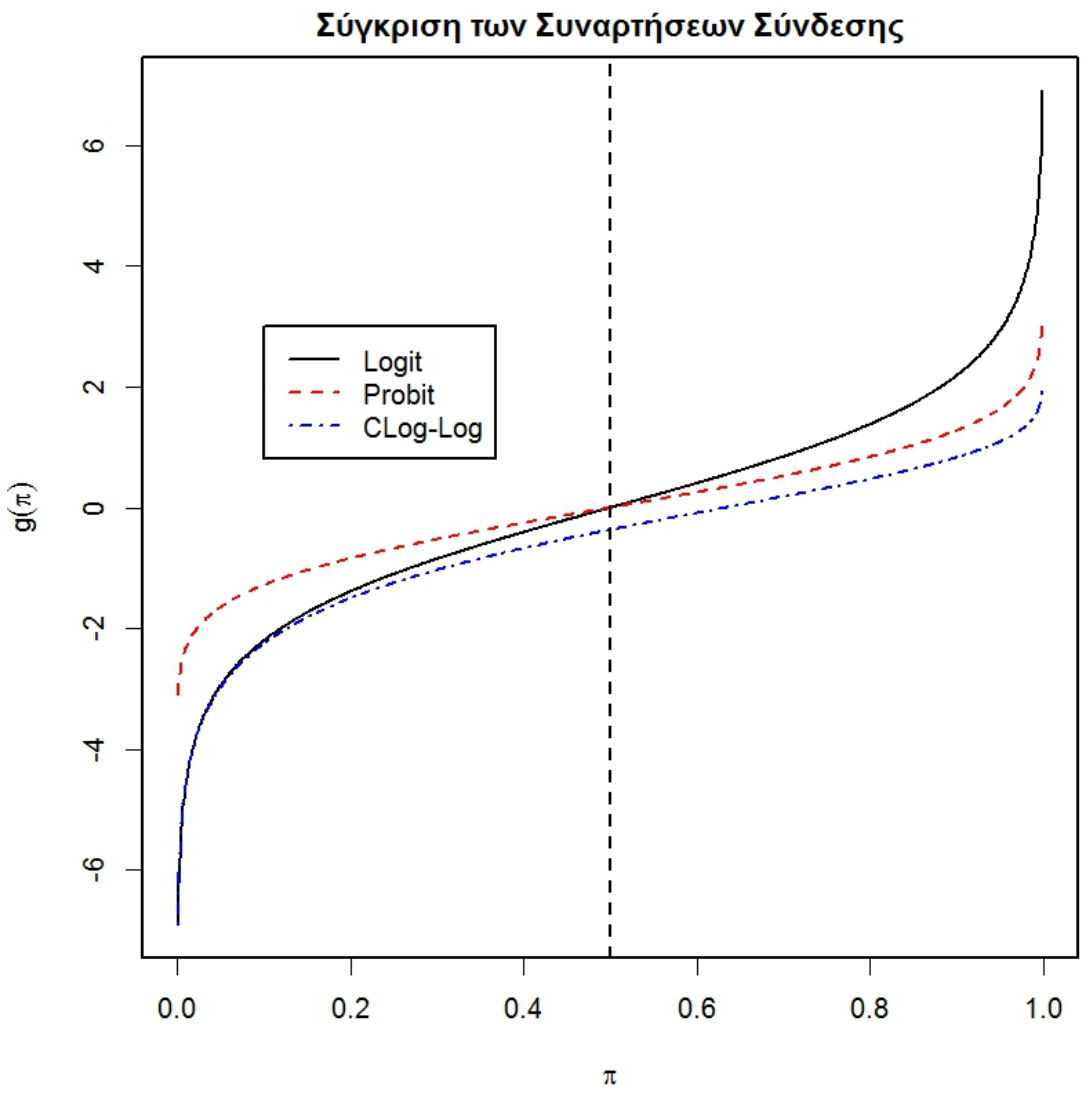
όπου με Φ συμβολίζουμε τη συνάρτηση κατανομής της Κανονικής κατανομής $\mathrm{N}\left(0,1\right)$.

Τα μοντέλα probit έχουν ευρεία εφαρμογή σε τομείς βιολογικών και κοινωνικών επιστημών, όπου υπάρχει φυσική ερμηνεία για το μοντέλο.

- Επιπλέον, μια άλλη επιλογή συνάρτησης σύνδεσης είναι η λεγόμενη συνάρτηση complementary log-log, η οποία ορίζεται ως εξής:

$$g(\pi_i) = \log\left[-\log(1-\pi_i)\right].$$

Το παραπάνω μοντέλο χρησιμοποιείται κυρίως σε περιπτώσεις, όπου η πιθανότητα να συμβεί ένα γεγονός είναι πολύ μικρή ή πολύ μεγάλη. Ειδάλλως, αν η πιθανότητα είναι κοντά στο $1/2$ είναι ισοδύναμο με το μοντέλο probit ή το λογιστικό μοντέλο.
Οι δύο παραπάνω συναρτήσεις σύνδεσης μαζί με την logit είναι οι πιο διαδεδομένες για διωνυμικά δεδομένα.
Όλες οι συναρτήσεις σύνδεσης φαίνονται στο Γράφημα 2.1. Παρατηρούμε ότι όλες οι συναρτήσεις σύνδεσης είναι συνεχείς και αύξουσες. Οι συναρτήσεις logit και probit φαίνεται να παρουσιάζουν αρκετές ομοιότητες και συγκεκριμένα για πιθανότητα ίση με 1/2 ταυτίζονται. Μία από αυτές είναι η συμμετρικότητα, που παρουσιάζουν ως προς την τιμή $\pi = 1/2$. Αντιθέτως, η συνδέουσα συνάρτηση complementary log-log δεν είναι συμμετρική. Ένας λόγος, που δεν χρησιμοποιείται τόσο συχνά είναι ότι για πιθανότητες μικρότερες του 1/2 φαίνεται να μην έχει τόσο «καλή συμπεριφορά». Παρατηρούμε ότι για μικρές τιμές του π, η συνάρτηση complementary log-log σχεδόν ταυτίζεται με τη λογιστική συνάρτηση, ενώ όσο η πιθανότητα πλησιάζει τη μονάδα φαίνεται να τείνει προς το ∞ πιο αργά συγκριτικά με τις άλλες συναρτήσεις σύνδεσης.



Γράφημα 2.1: *Σύγκριση των συναρτήσεων σύνδεσης σε σχέση με τις τιμές, που μπορεί να πάρει η πιθανότητα π για διωνυμικά δεδομένα.*

Παρακάτω θα παρουσιάσουμε και τα τρία μοντέλα παλινδρόμησης, που αναφέραμε παραπάνω ξεκινώντας από το πιο διαδεδομένο, το λογιστικό μοντέλο.

2.4 Λογιστική Παλινδρόμηση

2.4.1 Εισαγωγή

Θεωρούμε n ανεξάρτητες τυχαίες μεταβλητές $Y_1, Y_2, ..., Y_n$, έτσι ώστε η $i-$οστή παρατήρηση να είναι η πραγμάτωση μιας τυχαίας μεταβλητής Y_i, όπου η Y_i ακολουθεί την διωνυμική κατανομή

$Y_i \sim binomial(n_i, \pi_i)$ (2.9)

με παραμέτρους n_i να είναι ο αριθμός δοκιμών «Bernoulli» της i μονάδας του δείγματος και π_i η αντίστοιχη πιθανότητα επιτυχίας. Έστω $n_i = 1$ για όλα τα i. Αυτό προσδιορίζει το στοχαστικό μέρος του μοντέλου.

Εν συνεχεία, αν υποθέσουμε ότι ο μετασχηματισμός logit των π_i είναι μια γραμμική συνάρτηση των παραμέτρων του μοντέλου

$$\operatorname{logit}(\pi_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} \quad , (2.10)$$

όπου $\boldsymbol{x}_i$ το διάνυσμα των επεξηγηματικών μεταβλητών του μοντέλου και $\boldsymbol{\beta}$ το διάνυσμα των παραμέτρων του μοντέλου παλινδρόμησης.

Το μοντέλο, λοιπόν που διαμορφώνεται από τις σχέσεις (2.9) και (2.10) αποτελεί ένα γενικευμένο γραμμικό μοντέλο με συνάρτηση σύνδεσης logit και δίτιμη μεταβλητή απόκρισης.

2.4.2 Εκτίμηση Συντελεστών

Σύμφωνα, με τη θεωρία των γενικευμένων γραμμικών μοντέλων, η προσαρμογή του μοντέλου στα δεδομένα μας γίνεται με τη μέθοδο μέγιστης πιθανοφάνειας. Η συνάρτηση πιθανοφάνειας γράφεται ως

$$L(\boldsymbol{\pi};\boldsymbol{y}) = \prod_{i=1}^{n}\binom{n_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{n_i-y_i}. \quad (2.11)$$

Οι εκτιμήσεις μέγιστης πιθανοφάνειας των παραμέτρων $\boldsymbol{\beta}$, οπότε και των πιθανοτήτων $\pi_i = g^{-1}(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})$, προκύπτουν από τη μεγιστοποίηση της λογαριθμικής συνάρτησης πιθανοφάνειας

$$\begin{aligned} l = \log L(\boldsymbol{\pi};\boldsymbol{y}) &= \sum_{i=1}^{n}\left\{\log\binom{n_i}{y_i} + y_i\log\pi_i + (n_i - y_i)\log(1-\pi_i)\right\} \\ &= \sum_{i=1}^{n}\left\{\log\binom{n_i}{y_i} + y_i\log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i\log(1-\pi_i)\right\} \qquad (2.12) \\ &= \sum_{i=1}^{n}\left\{\log\binom{n_i}{y_i} + y_i\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - n_i\log\left(1+e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}\right)\right\}. \end{aligned}$$

Παραγωγίζοντας την σχέση (2.12) ως προς τις παραμέτρους του μοντέλου β_j έχουμε

$$\frac{\partial \log L(\boldsymbol{\pi};\boldsymbol{y})}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} n_i x_{ij} e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}} (1+e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}})^{-1}, j = 0,1,...,k$$

$$= \sum_{i=1}^{n} \left\{ y_i - n_i e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}} (1+e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}})^{-1} \right\} x_{ij}$$

$$= \sum_{i=1}^{n} \left(y_i - n_i \pi_i \right) x_{ij}.$$

Συνεπώς, οι εκτιμήτριες μέγιστης πιθανοφάνειας των β_j προκύπτουν από τη λύση των εξισώσεων score

$$\sum_{i=1}^{n} \left(y_i - n_i \pi_i \right) x_{ij} = 0 \Rightarrow \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{\mu}) = \boldsymbol{0},$$

όπου $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}$ με $\mu_i = n_i \pi_i,\ i = 1,2,...,n$ η μέση τιμή για κάθε συνιστώσα του τυχαίου δείγματος.

Εξισώνοντας τότε τις μερικές παραγώγους με μηδέν παίρνουμε τις εκτιμήσεις των $b_0, b_1, ..., b_k$ των παραμέτρων του μοντέλου $\beta_0, \beta_1, ..., \beta_k$. Δημιουργείται, λοιπόν ένα σύστημα από $p = k+1$ μη γραμμικές εξισώσεις και $k+1$ αγνώστους, το οποίο λύνεται μόνο με επαναληπτικές μεθόδους.

2.4.3 Ερμηνεία των Συντελεστών

Οι παράμετροι της λογιστικής παλινδρόμησης μπορούν να εκφραστούν μέσα από τις συμπληρωματικές πιθανοτήτες, των λεγόμενων $odds$. Στην περίπτωση που έχουμε διωνυμικά δεδομένα, ο λόγος των συμπληρωματικών πιθανοτήτων ορίζεται ως η πιθανότητα πραγματοποίησης του γεγονότος «επιτυχία» $(\mathrm{Y} = 1)$ σε σχέση με την πιθανότητα πραγματοποίησης του γεγονότος «αποτυχία» $(\mathrm{Y} = 0)$. Αυτό μαθηματικά εκφράζεται από τη σχέση

$$odds = \frac{\pi}{1-\pi}.$$

Μπορεί επίσης να ερμηνευτεί ως η ποσότητα που πρέπει να πολλαπλασιαστεί η σχετική πιθανότητα αποτυχίας, ώστε να υπολογιστεί η πιθανότητα επιτυχίας. Για παράδειγμα, αν η συμπληρωματική πιθανότητα επιτυχίας πάρει την τιμή 2, δηλαδή $odds = 2$, τότε αυτό ερμηνεύεται ως εξής: η πιθανότητα η τυχαία μεταβλητή Y να πάρει την τιμή 1 είναι 2 φορές μεγαλύτερη από την πιθανότητα η Y να πάρει την τιμή

0. Ωστόσο, αν $odds = 1$, τότε παρατηρούμε ότι η πιθανότητα «επιτυχίας» είναι ίση με την πιθανότητα «αποτυχίας», δηλ. ίση με $1/2$. Επιπλέον, αν $odds = 0.6$ αυτό μπορεί να ερμηνευτεί με δύο τρόπους. Δηλαδή ή ότι η πιθανότητα επιτυχίας είναι ίση με το 60% της πιθανότητας αποτυχίας ή ότι η πιθανότητα επιτυχίας είναι κατά 40% μικρότερη από την πιθανότητα αποτυχίας.

Συνοψίζοντας, αν θεωρήσουμε $\alpha = odds = \dfrac{\pi}{1-\pi}$, τότε αν $\alpha > 1$, η πιθανότητα επιτυχίας είναι $(\alpha - 1)100\%$ φορές μεγαλύτερη από την πιθανότητα αποτυχίας. Τέλος, αν $\alpha < 1$ τότε η πιθανότητα επιτυχίας είναι $(1-\alpha)100\%$ φορές μικρότερη από την πιθανότητα αποτυχίας.

Χρησιμοποιώντας τις συμπληρωματικές πιθανότητες το λογιστικό μοντέλο μπορεί να γραφεί ως

$$\mathrm{Y}_i \sim binomial\left(\frac{odds_i}{1+odds_i}, n_i\right), \log(odds_i) = \beta_0 + \sum_{j=0}^{k} \beta_j x_{ij} = \boldsymbol{x}_i^{\mathrm{T}} \beta .$$

Φυσικά, αν οι παράμετροι έχουν εκτιμηθεί, τότε $odds_i = e^{\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{b}}$. Συνεπώς, σύμφωνα με την ερμηνεία των συμπληρωματικών πιθανοτήτων προκύπτει ότι η ποσότητα e^{b_j} είναι ο παράγοντας επί τον οποίο πολλαπλασιάζεται η σχετική πιθανότητα πραγματοποίησης του ενδεχομένου «επιτυχία», όταν η ανεξάρτητη μεταβλητή x_{ij} αυξηθεί κατά μία μονάδα, με δεδομένο ότι οι υπόλοιπες ανεξάρτητες μεταβλητές παραμένουν σταθερές.

2.4.4 Διαστήματα Εμπιστοσύνης

Μπορούμε να κατασκευάσουμε διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου με βάση τη στατιστική συνάρτηση Wald. Ένα $100(1-a)\%-$διάστημα εμπιστοσύνης για την παράμετρο β_j είναι το $b_j \pm z_{a/2} se(b_j)$, $j = 0,1,...,k$. Με τη βοήθεια ενός διαστήματος εμπιστοσύνης για τη παράμετρο του μοντέλου β_j μπορούμε να κατασκευάσουμε ένα διάστημα εμπιστοσύνης για το λόγο των συμπληρωματικών πιθανοτήτων. Άρα, ένα $100(1-a)\%-$διάστημα εμπιστοσύνης για την εκτίμηση των συμπληρωματικών πιθανοτήτων είναι $\exp\left(b_j \pm z_{a/2} se(b_j)\right)$.

Επιπλέον, μπορούμε να δημιουργήσουμε $100(1-a)\%-$διαστήματα εμπιστοσύνης για την γραμμική προβλέπουσα του μοντέλου δοσμένων των τιμών των επεξηγηματικών μεταβλητών του μοντέλου. Αν θεωρήσουμε το σύνολο τιμών των επεξηγηματικών

μεταβλητών να είναι $\boldsymbol{x}_0^{\mathrm{T}} = [1, x_{01}, x_{02}, ..., x_{0k}]$, η εκτιμήτρια της διασποράς της γραμμικής προβλέπουσας θα είναι

$V(\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{b}) = \boldsymbol{x}_0^{\mathrm{T}} V(\boldsymbol{b}) \boldsymbol{x}_0 = \boldsymbol{x}_0^{\mathrm{T}} (\boldsymbol{X}^{\mathrm{T}} \hat{\boldsymbol{V}} \boldsymbol{X})^{-1} \boldsymbol{x}_0$. (2.13)

Συνεπώς, ένα $100(1-a)\%-$διάστημα εμπιστοσύνης για τη γραμμική προβλέπουσα θα είναι

$\boldsymbol{x}_0\boldsymbol{b} - z_{a/2}\sqrt{Var(\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{b})} \le \boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{b} \le \boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{b} + z_{a/2}\sqrt{Var(\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{b})}$. (2.14)

Το διάστημα εμπιστοσύνης για τη γραμμική προβλέπουσα μας βοηθά να κατασκευάσουμε ένα διάστημα εμπιστοσύνης για την πιθανότητα εμφάνισης επιτυχίας π_0, δοσμένου του σύνολου τιμών $\boldsymbol{x}_0^{\mathrm{T}} = [1, x_{01}, x_{02}, ..., x_{0k}]$.

Αν θεωρήσουμε

$$L(\boldsymbol{x}_0) = \boldsymbol{x}_0\boldsymbol{b} - z_{a/2}\sqrt{Var(\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{b})}$$

και

$$\mathrm{U}(\boldsymbol{x}_0) = \boldsymbol{x}_0\boldsymbol{b} + z_{a/2}\sqrt{Var(\boldsymbol{x}_0^{\mathrm{T}}\boldsymbol{b})}$$

το άνω και κάτω όριο του διαστήματος εμπιστοσύνης για τη γραμμική προβλέπουσα από τη σχέση (2.14), τότε ένα $100(1-a)\%-$διάστημα εμπιστοσύνης για την εκτίμηση της πιθανότητας εμφάνισης επιτυχίας π_0 θα είναι

$$\frac{\exp[\mathrm{L}(\boldsymbol{x}_0)]}{1+\exp[\mathrm{L}(\boldsymbol{x}_0)]} \le \pi_0 \le \frac{\exp[\mathrm{U}(\boldsymbol{x}_0)]}{1+\exp[\mathrm{U}(\boldsymbol{x}_0)]}.$$

2.4.5 Κριτήρια Καλής Προσαρμογής

Ελεγχοσυνάρτηση deviance

Έστω $\mathrm{Y}_1, \mathrm{Y}_2, ..., \mathrm{Y}_n$ ανεξάρτητες τυχαίες μεταβλητές όπου κάθε μία από αυτές προέρχεται από την διωνυμική κατανομή, δηλαδή $\mathrm{Y}_i \sim binomial(n_i, \pi_i)$. Η μέση τιμή της Y_i είναι ίση με $n_i\pi_i$, δηλαδή $\mathrm{E}(\mathrm{Y}_i) = \mu_i = n_i\pi_i$, $\mu_i > 0$.

Η ελεγχοσυνάρτηση deviance ορίζεται από τη σχέση

$$D = 2\log\frac{L\{\textit{κορεσμένο μοντέλο}\}}{L\{\textit{υποψήφιο μοντέλο}\}}. (2.15)$$

Για το κορεσμένο μοντέλο με $p_S = n$ παραμέτρους ισχύει ότι $\tilde{\mu}_i = y_i$, δηλαδή οι προβλεπόμενες μέσες τιμές μ_i εκτιμώνται από τις παρατηρούμενες τιμές y_i των τυχαίων μεταβλητών Y_i.

Έτσι, η μεγιστοποιημένη τιμή του λογαρίθμου της συνάρτησης της πιθανοφάνειας υπό την υπόθεση H_S του κορεσμένου μοντέλου θα είναι

$$l_S = \log L_S = \sum_{i=1}^{n} \left\{ \log \binom{n_i}{y_i} + y_i \log \tilde{\pi}_i + (n_i - y_i) \log(1 - \tilde{\pi}_i) \right\}$$, όπου $\tilde{\pi}_i = y_i / n_i$.

Για το υποψήφιο μοντέλο με αριθμό παραμέτρων p_0 μικρότερο του n ισχύει ότι $\hat{\mu}_i = n_i \hat{\pi}_i$, όπου $\hat{\pi}_i = \dfrac{e^{\mathbf{x}_i^T \boldsymbol{b}}}{1 + e^{\mathbf{x}_i^T \boldsymbol{b}}}$ οι πιθανότητες απόκρισης που προκύπτουν από την προσαρμογή του μοντέλου.

Η μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας υπό την υπόθεση H_0 του υποψήφιου μοντέλου θα είναι

$$l_0 = \log L_0 = \sum_{i=1}^{n} \left\{ \log \binom{n_i}{y_i} + y_i \log \hat{\pi}_i + \left(n_i - y_i\right) \log\left(1 - \hat{\pi}_i\right) \right\}.$$

Η ελεγχοσυνάρτηση deviance στην περίπτωση της λογιστικής παλινδρόμησης ορίζεται ως

$$\begin{aligned}
D &= D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\left\{ \tilde{l}_S - \hat{l}_0 \right\} \\
&= 2\sum_{i=1}^{n} \left\{ \log \binom{n_i}{y_i} + y_i \log \tilde{\pi}_i + (n_i - y_i) \log(1 - \tilde{\pi}_i) - \log \binom{n_i}{y_i} - y_i \log \hat{\pi}_i - \left(n_i - y_i\right) \log\left(1 - \hat{\pi}_i\right) \right\} \\
&= 2\sum_{i=1}^{n} \left\{ y_i \log\left(\frac{\tilde{\pi}_i}{\hat{\pi}_i} \right) + (n_i - y_i) \log\left(\frac{1 - \tilde{\pi}_i}{1 - \hat{\pi}_i} \right) \right\} \\
&= 2\sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i} \right) + \left(n_i - y_i\right) \log \frac{n_i - y_i}{n_i - \hat{\mu}_i} \right].
\end{aligned}$$

Όταν, το μοντέλο της λογιστικής παλινδρόμησης περιγράφει ικανοποιητικά τα δεδομένα μας και το δείγμα τιμών που διαθέτουμε είναι μεγάλο, η ελεγχοσυνάρτηση deviance ασυμπτωτικά ακολουθεί τη X^2 κατανομή με βαθμούς ελευθερίας ίσους με τη διαφορά των παραμέτρων του κορεσμένου μοντέλου με το υποψήφιο μοντέλο. Παρ'όλα αυτά στην περίπτωση των δυαδικών δεδομένων η ελεγχοσυνάρτηση deviance δεν προτιμάται για την καλή προσαρμογή του μοντέλου, αφού εξαρτάται μόνο από τις προσαρμοσμένες τιμές $\hat{\mu}_i$ και όχι άμεσα από τις παρατηρήσεις y_i και δεν ακολουθεί προσεγγιστικά την X^2 κατανομή.

Pearson

Για τη μελέτη της καταλληλότητας του μοντέλου εκτός από την ελεγχοσυνάρτηση deviance χρησιμοποιείται και η ελεγχοσυνάρτηση X^2 του Pearson, που δίνεται από τη σχέση

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}. \quad (2.16)$$

Όταν το προσαρμοσμένο μοντέλο αποτελεί μια καλή περιγραφή για τα δεδομένα μας, οι δύο ελεγχοσυναρτήσεις X^2 του Pearson και deviance είναι ασυμπτωτικά ισοδύναμες και ακολουθούν την ίδια κατανομή X^2. Η επιλογή μεταξύ των δύο στατιστικών συναρτήσεων εξαρτάται από το ποια από τις δύο προσεγγίζεται ικανοποιητικά από την X^2 κατανομή. Ωστόσο, όταν έχουμε μη ομαδοποιημένα δεδομένα (δίτιμα δεδομένα, δηλαδή $n_i = 1$, για κάθε i) ή συνεχείς (ή σχεδόν συνεχείς) επεξηγηματικές μεταβλητές η ελεγχοσυνάρτηση X^2 του Pearson δεν ακολουθεί προσεγγιστικά την X^2 κατανομή.

Hosmer-Lemeshow

Ένα άλλο μέτρο καταλληλότητας του μοντέλου είναι ο έλεγχος των Hosmer και Lemeshow (1980). Ο έλεγχος αυτός διαφοροποιείται από τις ελεγχοσυναρτήσεις του Pearson και deviance, καθώς μπορεί να εφαρμοστεί σε μη ομαδοποιημένα δεδομένα. Για να υπολογιστεί ο έλεγχος αυτός, τα δεδομένα αρχικά διατάσσονται σε αύξουσα σειρά, σύμφωνα με την τιμή της εκτιμημένης πιθανότητας $\hat{\pi}_i$ για την i – οστή ομάδα και στη συνέχεια χωρίζονται σε ομάδες, οι οποίες συνήθως δεν ξεπερνάνε τις 10 σε αριθμό.

Η ελεγχοσυνάρτηση Hosmer Lemeshow δίνεται από τον τύπο

$$X^2_{HL} = \sum_{i=1}^{g} \frac{(o_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)}, \quad (2.17)$$

όπου m_i ο αριθμός των παρατηρήσεων για την i – οστή από τις g ομάδες, o_i είναι ο συνολικός αριθμός των επιτυχιών στην i – οστή ομάδα, e_i ο συνολικός αναμενόμενος αριθμός επιτυχιών στην i – οστή ομάδα και $\hat{\pi}_i = e_i / m_i$ η μέση πιθανότητα επιτυχίας της i – οστής ομάδας.

Ουσιαστικά, αποτελεί έναν έλεγχο X^2 του Pearson για έναν $g \times 2$ πίνακα συνάφειας και ακολουθεί προσεγγιστικά την X^2 κατανομή με $(g-2)$ βαθμούς ελευθερίας. Έχει δειχθεί ότι δεν αποτελεί ένα καλό μέτρο αξιολόγησης προσαρμογής για τη λογιστική παλινδρόμηση σε περιπτώσεις όπου υπάρχουν κατηγορικές και συνεχείς επεξηγηματικές μεταβλητές. Αντιθέτως, όταν οι επεξηγηματικές μεταβλητές είναι

συνεχείς είναι ένα προτιμητέο μέτρο για τον έλεγχο της προσαρμογής του λογιστικού μοντέλου.

2.4.6 Κριτήρια Επιλογής Μοντέλου

Στη λογιστική παλινδρόμηση, όπως και σε όλα τα γενικευμένα γραμμικά μοντέλα μπορεί να χρησιμοποιηθούν τα κριτήρια AIC και BIC, καθώς και διάφοροι συντελεστές προσδιορισμού για την επιλογή του βέλτιστου μοντέλου. Τα κριτήρια αυτά για την επιλογή του βέλτιστου μοντέλου παρουσιάζονται παρακάτω.

Κριτήριο AIC

Στη λογιστική παλινδρόμηση έχει τη μορφή

$$AIC = -2\left[\sum_{i=1}^{n}\log\binom{n_i}{y_i} + y_i \log\hat{\pi}_i + (n_i - y_i)\log(1-\hat{\pi}_i)\right] + 2p \text{ , (2.18)}$$

όπου $\hat{\pi}_i = \dfrac{e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}}}{1+e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}}}$.

Κριτήριο BIC

Στη λογιστική παλινδρόμηση θα έχει τη μορφή

$$BIC = -2\left[\sum_{i=1}^{n}\log\binom{n_i}{y_i} + y_i \log\hat{\pi}_i + (n_i - y_i)\log(1-\hat{\pi}_i)\right] + p\log n \text{ , (2.19)}$$

όπου $\hat{\pi}_i = \dfrac{e^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{b}}}{1+e^{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{b}}}$.

Κριτήρια R^2

Όπως στο γενικό γραμμικό μοντέλο έτσι και στη λογιστική παλινδρόμηση υπάρχουν κάποιοι δείκτες προσδιορισμού αντίστοιχοι με το γνωστό συντελεστή προσδιορισμού R^2. Γενικά, έχει παρατηρηθεί ότι κανένα από αυτά τα μέτρα δεν αποτελούν ικανοποιητικά μέτρα προσαρμογής. Ωστόσο, χρησιμοποιούνται κυρίως για την σύγκριση της προσαρμογής μοντέλων στα ίδια δεδομένα.

Ψευδο-R^2

Ο ψευδο-R^2 συντελεστής είναι ένα μέτρο που δεν ενδείκνυται γενικά για χρήση, καθώς δεν είναι εύκολη η ερμηνεία του σύμφωνα με τους Mittlobock και Schemper (1996). Ωστόσο, χρησιμοποιείται σε πληθώρα στατιστικών πακέτων και διαθέτει διαφορετικά ονόματα σε κάθε ένα από αυτά (ψευδο-R^2 στο STATA, Cox-snell-R^2 στο SPSS κλπ.). Βασίζεται στο λογάριθμο μέγιστης πιθανοφάνειας του μοντέλου, που περιλαμβάνει μόνο το σταθερό όρο και του μοντέλου, που περιλαμβάνει όλες τις επεξηγηματικές μεταβλητές. Για αυτό το λόγο δίνεται από τον τύπο

$$R_L^2 = \frac{l_0 - l_1}{l_0} = 1 - \frac{l_1}{l_0}, \quad (2.20)$$

όπου l_1 είναι η μεγιστοποιημένη τιμή του λογαρίθμου της συνάρτησης πιθανοφάνειας του υποψήφιου μοντέλου, δηλαδή του μοντέλου, που περιλαμβάνει τις k επεξηγηματικές μεταβλητές και l_0 είναι η μεγιστοποιημένη τιμή του λογαρίθμου της συνάρτησης πιθανοφάνειας για το μοντέλο μόνο με το σταθερό όρο.

Ο παραπάνω συντελεστής παίρνει τη μέγιστη τιμή του, δηλαδή $R_L^2 = 1$ όταν προσαρμόζουμε το κορεσμένο μοντέλο (συνήθως με αριθμό παραμέτρων ίσο με το μέγεθος του δείγματος n). Σε περιπτώσεις όπου το υποψήφιο μοντέλο έχει αριθμό παραμέτρων μικρότερο από n, ο ψευδο-R^2 συντελεστής θα έχει τιμή μικρότερη του 1. Συχνά, χρησιμοποιούμε τον διορθωμένο ψευδο-R^2 συντελεστής του R_L^2,

$$R_{LS}^2 = \frac{R_L^2}{\max R_L^2} = \frac{l_0 - l_1}{l_0 - l_S}, \quad (2.21)$$

όπου l_S η μεγιστοποιημένη τιμή του λογαρίθμου της συνάρτησης πιθανοφάνειας για το κορεσμένο μοντέλο και l_0 είναι η μεγιστοποιημένη τιμή του λογαρίθμου της συνάρτησης πιθανοφάνειας για το μοντέλο μόνο με το σταθερό όρο.

2.4.7 Διαγνωστικοί Έλεγχοι

Υπόλοιπα

Το μοντέλο της λογιστική παλινδρόμησης ανήκει στην κλάση των γενικευμένων γραμμικών μοντέλων. Συνεπώς, τα υπόλοιπα υπολογίζονται από τους γενικούς τύπους, που έχουμε αναφέρει στην παράγραφο 1.8.1.

- **Υπόλοιπα deviance**

Τα υπόλοιπα deviance ορίζονται από τη σχέση

$$\varepsilon_i^D = sign(y_i - \hat{\mu}_i)\sqrt{d_i(y_i, \hat{\mu}_i)}\ , (2.22)$$

όπου $d_i(y_i, \hat{\mu}_i) = 2\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (n_i - y_i)\log\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right],\ i = 0,1,...,n$.

Τα τυποποιημένα υπόλοιπα Deviance ορίζονται μέσω της σχέσης

$$\varepsilon_i^{DS} = \frac{\varepsilon_i^D}{\sqrt{1-\hat{h}_{ii}}}\ , (2.23)$$

όπου $\hat{h}_{ii}$ το διαγώνιο στοιχείο του $n \times n$ πίνακα $\hat{\mathbf{H}}$(*hat matrix*), που δίνεται από τη σχέση

$$\hat{\mathbf{H}} = \hat{\mathbf{W}}^{1/2}\mathbf{X}\left(\mathbf{X}^{\mathbf{T}}\hat{\mathbf{W}}\mathbf{X}\right)\mathbf{X}^{\mathbf{T}}\hat{\mathbf{W}}^{1/2}\ ,$$

όπου $\mathbf{X}$ ο $n \times p$ πίνακας σχεδιασμού και $\mathbf{W}$ ο σταθμισμένος πίνακας πληροφορίας του προσαρμοσμένου μοντέλου, ο οποίος είναι ένας $n \times n$ διαγώνιος πίνακας με στοιχεία $w_{ii} = n_i\hat{\pi}_i(1-\hat{\pi}_i),\ i = 0,1,...,n$, δηλαδή την εκτιμημένη διασπορά της απόκρισης Y_i.

- **Υπόλοιπα Pearson**

Τα υπόλοιπα Pearson, δεδομένου ότι στην περίπτωση της λογιστικής παλινδρόμησης ισχύει ότι $var(\hat{\mu}_i) = n_i\hat{\pi}_i(1-\hat{\pi}_i)$, δίνονται από τον τύπο

$$\varepsilon_i^P = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}.\quad (2.24)$$

Τα τυποποιημένα υπόλοιπα Pearson ορίζονται μέσω της σχέσης

$$\varepsilon_i^{PS} = \frac{\varepsilon_i^P}{\sqrt{1-\hat{h}_{ii}}}.\quad (2.25)$$

όπου $\hat{h}_{ii}$ το διαγώνιο στοιχείο του $n \times n$ πίνακα $\hat{\mathbf{H}}$ (*hat matrix*) ,όπως αναφέραμε και παραπάνω.

- **Υπόλοιπα πιθανοφάνειας**

Τα υπόλοιπα πιθανοφάνειας ορίζονται από τη σχέση

$$\varepsilon_i^L = sign(y_i - \hat{\mu}_i)\sqrt{\hat{h}_{ii}(\varepsilon_i^{PS})^2 + (1-\hat{h}_{ii})(\varepsilon_i^{DS})^2},\ i = 1,2,...,n \quad (2.26)$$

και αποτελούν σταθμισμένο συνδυασμό των παραπάνω υπολοίπων.

Γενικά, τα τυποποιημένα υπόλοιπα Pearson και deviance προσεγγιστικά ακολουθούν την τυποποιημένη Κανονική κατανομή $N(0,1)$ δεδομένου φυσικά ότι ο αριθμός δοκιμών για την i μονάδα του δείγματος, n_i δεν είναι μικρός. Μπορεί να γίνει έλεγχος της υπόθεσης αυτής και γραφικά με κατάλληλα γραφήματα της Κανονικής κατανομής (*Q-Q plots*).

2.4.8 Το Πρόβλημα της Υπερμεταβλητότητας για Διωνυμικά Δεδομένα

Παρ' όλο, που το μοντέλο της λογιστικής παλινδρόμησης είναι το πιο διαδεδομένο για δίτιμα ή διωνυμικά δεδομένα, πολλές φορές η προσαρμογή του φαίνεται να μην είναι ικανοποιητική. Σ' αυτήν την περίπτωση εμφανίζεται το πρόβλημα της υπερμεταβλητότητας (*overdispersion*) ή της επιπλέον διωνυμικής μεταβλητότητας (*extra binomial variation*), όπως καλείται μερικές φορές. Αυτό σημαίνει ότι μετά την προσαρμογή του μοντέλου τα δεδομένα παρουσιάζουν μεγαλύτερη διασπορά από τις θεωρητικές $Var(Y_i) = n_i\pi_i(1-\pi_i)$. Μία ένδειξη της παρουσίας της υπερμεταβλητότητας είναι όταν η τιμή της ελεγχοσυνάρτησης $deviance/df$ είναι μεγαλύτερη της μονάδας, όπου df είναι οι βαθμοί ελευθερίας της X^2 κατανομής που ακολουθεί η ελεγχοσυνάρτηση deviance. Ωστόσο, ο Lindsey (1999) θεωρεί καλύτερο κριτήριο για την ύπαρξη υπερμεταβλητότητας, η τιμή της στατιστικής συνάρτησης $deviance$ να είναι τουλάχιστον διπλάσια των βαθμών ελευθερίας $df(=n-p)$, το οποίο ισοδυναμεί με τη χρήση του κριτηρίου AIC.

Μερικοί λόγοι, που οδηγούν σε αυτό το φαινόμενο είναι οι εξής:

1. Η επιλογή της διωνυμικής κατανομής δεν είναι η κατάλληλη για να περιγράψει τα δεδομένα.
2. Η επιλογή της συνάρτησης σύνδεσης δεν είναι σωστή. Ενδεχομένως η επιλογή μιας άλλης συνδέουσας συνάρτησης, όπως η probit ή η complementary log-log να είναι καταλληλότερη για το μοντέλο.
3. Δεν έχει καθοριστεί σωστά η γραμμική προβλέπουσα του μοντέλου. Πιθανότατα, να μην έχουν συμπεριληφθεί ενδεχόμενες αλληλεπιδράσεις μεταξύ των επεξηγηματικών μεταβλητών ή κάποιες στατιστικά σημαντικές

επεξηγηματικές μεταβλητές να είναι εκτός μοντέλου ή να χρειάζεται κάποιος μετασχηματισμός κάποιας εκ των μεταβλητών, όπως για παράδειγμα η λογαριθμική συνάρτηση.

4. Ενδέχεται οι τυχαίες μεταβλητές Y_i να μην είναι ανεξάρτητες.
5. Υπάρχει ένα ή περισσότερα έκτροπα σημεία (*outliers*) στα δεδομένα.

Σε περιπτώσεις, όπου η υπερμεταβλητότητα οφείλεται στο ότι δεν έχει επιλεχθεί η κατάλληλη κατανομή τότε θα χρησιμοποιήσουμε κάποια άλλη κατανομή κατάλληλη να περιγράψει τα δεδομένα και να επιτρέψει μια επιπλέον μεταβλητότητα σε αυτή συγκριτικά με την εν λόγω κατανομή. Στην περίπτωση των διωνυμικών δεδομένων μπορεί να χρησιμοποιηθεί η κατανομή βήτα-διωνυμική, αντί της διωνυμικής. Η κατανομή βήτα-διωνυμική είναι μια επέκταση της διωνυμικής κατανομής χρησιμοποιώντας και την κατανομή Βήτα. Η κατανομή αυτή όμως δεν ανήκει στην εκθετική οικογένεια και επομένως δεν εκτιμώνται οι παράμετροι του μοντέλου με τη μέθοδο μέγιστης πιθανοφάνειας στα πλαίσια των γενικευμένων γραμμικών μοντέλων, αλλά με τη μέθοδο της Quasi-πιθανοφάνειας, που αναπτύξαμε και σε προηγούμενο κεφάλαιο.

Παρατηρούμε ότι αν θεωρήσουμε ότι η πιθανότητα απόκρισης π_i είναι μια τυχαία μεταβλητή με μέση τιμή $\mathrm{E}(\pi_i) = p_i$ και διασπορά $V(\pi_i) = \psi p_i(1-p_i)$, όπου $\psi > 0$ μια άγνωστη παράμετρος κλίμακας και η τυχαία μεταβλητή $\mathrm{Y}_i \mid \pi_i$ ακολουθεί την διωνυμική κατανομή $binomial(n_i, \pi_i)$, τότε

$\mathrm{E}\left\{\mathrm{E}(\mathrm{Y}_i \mid \pi_i)\right\} = n_i \mathrm{E}(\pi_i) = n_i p_i$ και

$$
\begin{aligned}
V(\mathrm{Y}_i) &= \mathrm{E}\left\{V(\mathrm{Y}_i \mid \pi_i)\right\} + V\left\{\mathrm{E}(\mathrm{Y}_i \mid \pi_i)\right\} \\
&= n_i \mathrm{E}\left\{\pi_i(1-\pi_i)\right\} + V(n_i \pi_i) \\
&= n_i \left\{\mathrm{E}(\pi_i) - \mathrm{E}(\pi_i^2)\right\} + n_i^2 \psi p_i(1-p_i) \\
&= n_i \left\{p_i - (\psi p_i(1-p_i) + p_i^2)\right\} + n_i^2 \psi p_i(1-p_i) \\
&= n_i(n_i-1)\psi p_i(1-p_i) + n_i p_i(1-p_i) \\
&= n_i p_i(1-p_i)[1 + (n_i-1)\psi].
\end{aligned}
$$

Συνεπώς, $V(\mathrm{Y}_i) = n_i p_i(1-p_i)[1+(n_i-1)\psi] = n_i p_i(1-p_i)\varphi$, δηλαδή πρόκειται για τη διωνυμική διασπορά πολλαπλασιασμένη με έναν παράγοντα φ. Στην περίπτωση που $\varphi = 1$ (δηλαδή $n_i = 1$ ή $\psi = 0$) η διασπορά θα είναι ίση ακριβώς με τη διωνυμική διασπορά.

ΚΕΦΑΛΑΙΟ 3

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΓΙΑ ΚΑΤΗΓΟΡΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ Ή ΜΕΤΑΒΛΗΤΕΣ ΔΙΑΤΑΞΗΣ

3.1 Λογιστική Παλινδρόμηση για Κατηγορικές Μεταβλητές Απόκρισης

Μια γενίκευση της λογιστικής παλινδρόμησης είναι η περίπτωση, που η μεταβλητή απόκρισης είναι κατηγορική μεταβλητή με περισσότερες από δύο κατηγορίες (*nominal*). Ένα παράδειγμα μιας κατηγορικής μεταβλητής θα μπορούσε να είναι η επιλογή κατάλληλης θεραπείας σε σχέση με τρείς θεραπείες, που εφαρμόστηκαν σε καρκινοπαθείς ασθενείς. Στην περίπτωση αυτή η μεταβλητή απόκρισης έχει τρείς κατηγορίες, μία για κάθε θεραπεία. Ο McFadden (1974) πρότεινε μια τροποποίηση του λογιστικού μοντέλου, το μοντέλο διακριτής επιλογής (*discrete choice model*). Ο όρος αυτός χρησιμοποιείται κυρίως στην οικονομετρία και στην επιχειρησιακή έρευνα. Ωστόσο, στις ιατρικές επιστήμες έχει επικρατήσει ο όρος πολυωνυμικό μοντέλο λογιστικής παλινδρόμησης (*multinomial or polytomous or polychotomous*). Η ανάλυση της στατιστικής μοντελοποίησης των κατηγορικών δεδομένων βασίζεται στην πολυωνυμική κατανομή.

3.2 Πολυωνυμική Κατανομή

Θεωρούμε μια κατηγορική τυχαία μεταβλητή Y με J κατηγορίες. Επιπλέον, ορίζουμε $\pi_j = \Pr(\mathrm{Y} = j)$ να είναι η πιθανότητα η παρατηρούμενη τιμή y της τυχαίας μεταβλητής Y να βρίσκεται στην $j-$οστή κατηγορία ($j = 1, 2, ..., J$). Θεωρούμε επίσης ότι $\sum_{j=1}^{J} \pi_j = 1$. Αν θεωρήσουμε n ανεξάρτητες παρατηρήσεις της τυχαίας μεταβλητή Y και συμβολίσουμε y_1 τον αριθμό παρατηρήσεων στην 1^{η} κατηγορία, y_2 τον αριθμό παρατηρήσεων στην 2^{η} κατηγορία κλπ, τότε τα δεδομένα μπορούν να ανπαρασταθούν από το παρακάτω διάνυσμα

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_J \end{bmatrix}, \text{ με } \sum_{j=1}^{J} y_j = n .$$

Μπορούμε να θεωρήσουμε τότε ότι τα παραπάνω δεδομένα προέρχονται από την πολυωνυμική κατανομή

$$f(\boldsymbol{y}\,|\,n)=\frac{n!}{y_1!\,y_2!\cdots y_J!}\pi_1^{y_1}\cdots\pi_J^{y_J}\,.\,(3.1)$$

Στην ειδική περίπτωση, όπου θεωρούμε ότι έχουμε 2 κατηγορίες, δηλαδή $J = 2$ έχουμε ότι $\pi_2 = 1-\pi_1$ και $y_2 = n - y_1$ και προκύπτει η συνάρτηση μάζας πιθανότητας της διωνυμικής κατανομής.

Η πολυωνυμική κατανομή δεν ανήκει στην εκθετική οικογένεια κατανομών.

3.3 Το Πολυωνυμικό Μοντέλο

Θεωρούμε μια κατηγορική μεταβλητή απόκρισης Y με J κατηγορίες. Θα πρέπει να θεωρήσουμε μια κατηγορία αναφοράς, έστω την πρώτη κατηγορία. Τότε θα πρέπει να δημιουργήσουμε $J-1$ συναρτήσεις logit, ώστε να έχουμε μια ερμηνεία των odds μεταξύ όλων των κατηγοριών σε σχέση με την κατηγορία αναφοράς.

Συνεπώς, λαμβάνουμε τις συναρτήσεις

$$g_j = \log\left(\frac{\pi_j(\boldsymbol{x})}{\pi_1(\boldsymbol{x})}\right) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j,\ \ j=1,...,J\ ,\ (3.2)$$

όπου $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_j$ αποτελεί τη γραμμική προβλέπουσα και $\boldsymbol{x}^{\mathrm{T}} = [x_0, x_1, ..., x_{\mathrm{k}}]$ το διάνυσμα των $p = k+1$ επεξηγηματικών μεταβλητών του μοντέλου.

Οι συναρτήσεις g_j χρησιμοποιούνται για την εκτίμηση των συντελεστών του μοντέλου.

Αν η εκτιμήτρια για κάθε διάνυσμα $\boldsymbol{\beta}_j$ των συντελεστών του μοντέλου είναι ένα διάνυσμα $\boldsymbol{b}_j$, τότε από τη σχέση (3.2) λαμβάνουμε τις εκτιμώμενες πιθανότητες $\hat{\pi}_j$

$$\hat{\pi}_j = \hat{\pi}_1 \exp(\boldsymbol{x}^T\boldsymbol{b}_j),\ \ j=2,...,J\ ,$$

όπου $\hat{\pi}_1 = \dfrac{1}{1+\sum_{j=2}^{J}\exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{b}_j)}$, αφού ισχύει η σχέση $\sum_{j=1}^{J}\pi_j = 1$.

Συνεπώς, οι εκτιμημένες πιθανότητες $\hat{\pi}_j$ δίνονται από τη σχέση

$$\hat{\pi}_j = \frac{\exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{b}_j)}{1+\sum_{j=2}^{J}\exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{b}_j)},\ j=2,...,J\ .$$

3.3.1 Ερμηνεία των Συντελεστών του Μοντέλου

Για λόγους απλότητας, όσον αφορά την ερμηνεία των συντελεστών ενός λογιστικού μοντέλου παλινδρόμησης με κατηγορική μεταβλητή απόκρισης θα θεωρήσουμε ως

επεξηγηματική μεταβλητή μια δίτιμη μεταβλητή. Έστω, λοιπόν ότι μελετάμε ένα πρόβλημα όπου η μεταβλητή απόκρισης Y με J κατηγορίες και η επεξηγηματική μεταβλητή X, η οποία λαμβάνει δύο τιμές 0 και 1, ανάλογα με το αν ο παράγοντας υπάρχει ή όχι.

Ο λόγος των συμπληρωματικών πιθανοτήτων (*odds ratio*) της απόκρισης $\mathrm{Y}=j$ σε σχέση με την απόκριση για την κατηγορία αναφοράς $\mathrm{Y}=1$ δίνεται από τη σχέση

$$OR_j = \frac{P(\mathrm{Y}=j\,|\,\mathrm{X}=1)\,/\,P(\mathrm{Y}=1\,|\,\mathrm{X}=1)}{P(\mathrm{Y}=j\,|\,\mathrm{X}=0)\,/\,P(\mathrm{Y}=1\,|\,\mathrm{X}=0)} = \frac{\pi_{j1}}{\pi_{11}} \Big/ \frac{\pi_{j0}}{\pi_{10}},\ j=2,\ldots,J\,.$$

όπου π_{j1} και π_{j0} είναι οι πιθανότητες της απόκρισης $\mathrm{Y}=j$ σε σχέση με το αν η επεξηγηματική τιμή λαμβάνει την τιμή 0 ή 1.

Το μοντέλο έχει τη μορφή $\log\frac{\pi_j}{\pi_1} = \beta_{0j} + \beta_{1j}x$.

Στην περίπτωση, που $\mathrm{X} = 0$, τότε έχουμε $\log\frac{\pi_{j0}}{\pi_{10}} = \beta_{0j}$, ενώ στην περίπτωση, που $\mathrm{X} = 1$, τότε ο λογάριθμος των odds θα είναι $\log\frac{\pi_{j1}}{\pi_{11}} = \beta_{0j} + \beta_{1j}$.

Συνεπώς, ο λογάριθμος του λόγου των σχετικών πιθανοτήτων θα είναι

$$\log(OR_j) = \log\frac{\pi_{ja}}{\pi_{1a}} - \log\frac{\pi_{jb}}{\pi_{1b}} = \beta_{1j}\ .$$

Τέλος, ο λόγος των συμπληρωματικών πιθανοτήτων δίνεται από τη σχέση $OR_j = \exp(\beta_{1j})$.

3.4 Λογιστική Παλινδρόμηση για Μεταβλητές Διάταξης

Υπάρχουν περιπτώσεις, όπου η μεταβλητή απόκρισης Y δεν είναι απλά μια κατηγορική μεταβλητή, αλλά μια μεταβλητή διάταξης. Θα μπορούσαμε να χρησιμοποιήσουμε το πολυωνυμικό λογιστικό μοντέλο για την προσαρμογή των δεδομένων μας, καθώς η απόκριση αποτελεί μια κατηγορική μεταβλητή. Ωστόσο, κάτι τέτοιο θα οδηγούσε σε λάθος συμπεράσματα όσον αφορά την ερμηνεία των συντελεστών του μοντέλου, καθώς δεν θα είχαμε λάβει υπόψη την διάταξη των κατηγοριών της μεταβλητής απόκρισης. Για την στατιστική μοντελοποίηση αυτών των μεταβλητών αναπτύσσονται κάποια άλλα μοντέλα λογιστικής παλινδρόμησης, τα

οποία λαμβάνουν υπόψη τη σειρά των κατηγοριών της μεταβλητής απόκρισης. Κάποια από αυτά είναι το λογιστικό μοντέλο των διαδοχικών κατηγοριών (*adjacent categories logit model*), το λογιστικό μοντέλο των συνεχιζόμενων λόγων (*continuation ratio logit model*) και το μοντέλο των διαδοχικών συμπληρωματικών πιθανοτήτων (*proportional odds model*). Ωστόσο, θα πρέπει να είμαστε προσεκτικοί στην επιλογή του μοντέλου, καθώς η ερμηνεία των log-odds διαφέρει μεταξύ αυτών. Εμείς θα πρέπει να αποφασίσουμε ποιο μοντέλο έχει νόημα για τα δεδομένα μας.

3.4.1 Το Μοντέλο των Διαδοχικών Συμπληρωματικών Πιθανοτήτων

Σε ορισμένες περιπτώσεις η μεταβλητή απόκρισης μπορεί να είναι μια συνεχής μεταβλητή, η οποία είναι δύσκολο να μετρηθεί, ώστε να έχει νόημα. Για παράδειγμα, έστω ότι παίρνουμε τιμές για κάποιο δείκτη για μια ασθένεια και μας ενδιαφέρει η κατηγοριοποίηση αυτού όσον αφορά τη σοβαρότητα της ασθένειας, δηλαδή αν ο ασθενής «δεν νοσεί», αν έχει «ήπια μορφή της νόσου» και αν «νοσεί σοβαρά». Συνεπώς, θέλουμε να συγκρίνουμε την πιθανότητα να έχουμε μικρότερες ή ίσες αποκρίσεις από κάποια συγκεκριμένη τιμή, που έχουμε ορίσει ως κατηγορία, $\mathrm{Y} \le j$ με την πιθανότητα να έχουμε μεγαλύτερη απόκριση, $\mathrm{Y} > j$. Έτσι, λαμβάνουμε J κατηγορίες για την μεταβλητή απόκρισης με αντίστοιχες πιθανότητες $\pi_1, \pi_2, ..., \pi_J$

($\sum_{j=1}^{J} \pi_j = 1$).

Συνεπώς, το μοντέλο των διαδοχικών συμπληρωματικών πιθανοτήτων ορίζεται από τη σχέση

$$\log\left[\frac{P(\mathrm{Y} \le j \mid \boldsymbol{x})}{P(\mathrm{Y} > j \mid \boldsymbol{x})}\right] = \beta_{0j} + \beta_1 x_1 + ... + \beta_{p-1} x_{p-1} \text{ , για } j = 1, ..., J \quad (3.3)$$

όπου $\boldsymbol{x} = [1, x_1, x_2, ..., x_\kappa]$ οι παρατηρούμενες τιμές των επεξηγηματικών μεταβλητών του μοντέλου.

3.4.2 Το Λογιστικό Μοντέλο των Διαδοχικών Κατηγοριών

Υπάρχουν περιπτώσεις, που θέλουμε να εξετάσουμε την κάθε απόκριση με την αμέσως μεγαλύτερη. Για αυτή την περίπτωση κατάλληλο είναι το μοντέλο των διαδοχικών κατηγοριών και ορίζεται από τη σχέση

$$\log\left(\frac{P(\mathrm{Y} = j \mid \boldsymbol{x})}{P(\mathrm{Y} = j+1 \mid \boldsymbol{x})}\right) = \log\frac{\pi_j}{\pi_{j+1}} = \boldsymbol{x}^T \boldsymbol{\beta}_j \text{ , } j = 1, ..., J \quad (3.4)$$

Το μοντέλο αυτό απλοποιείται λαμβάνοντας τη μορφή

$$\log\frac{\pi_j}{\pi_{j+1}} = \beta_{0j} + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}, \text{ για } j = 1,\ldots,J,$$

όπου η επίδραση της κάθε επεξηγηματικής μεταβλητής θεωρείται ίδια για κάθε $\frac{\pi_j}{\pi_{j+1}}$, $j = 1,\ldots,J$.

3.4.3 Το Λογιστικό Μοντέλο των Συνεχιζόμενων Λόγων

Ας υποθέσουμε ότι θέλουμε να συγκρίνουμε την κάθε απόκριση με όλες τις αποκρίσεις, που είναι μικρότερες από αυτή, δηλαδή να κάνουμε σύγκριση της $\mathrm{Y} = j$ με την $\mathrm{Y} < j$, για $j = 1,\ldots,J$. Σε αυτές τις περιπτώσεις χρησιμοποιείται το λογιστικό μοντέλο των συνεχιζόμενων λόγων και ορίζεται ως εξής

$$\log\left(\frac{P(\mathrm{Y} = j \mid \boldsymbol{x})}{P(\mathrm{Y} < j \mid \boldsymbol{x})}\right) = \boldsymbol{x}^T \boldsymbol{\beta}_j, \text{ για } j = 1,\ldots,J. \quad (3.5)$$

Το μοντέλο αυτό είναι προτιμότερο από το μοντέλο των διαδοχικών συμπληρωματικών πιθανότήτων, όσον αφορά την ερμηνεία των πιθανοτήτων των διακεκριμένων κατηγοριών $\pi_1, \pi_2, \ldots, \pi_J$.

ΚΕΦΑΛΑΙΟ 4

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΓΙΑ ΑΠΑΡΙΘΜΗΤΑ ΔΕΔΟΜΕΝΑ

4.1 Παλινδρόμηση Poisson

Το μοντέλο παλινδρόμησης Poisson εφαρμόζεται σε περιπτώσεις, όπου η μεταβλητή απόκρισης είναι μεταβλητή, που εκφράζει τον αριθμό συμβάντων σε συγκεκριμένο χωρικό ή χρονικό διάστημα. Για παράδειγμα μπορεί να είναι αριθμός ατυχημάτων ανά βδομάδα σε μια διασταύρωση, αριθμός επισκέψεων στο γιατρό ανά μήνα, κλπ. Μία κατάλληλη κατανομή για την περιγραφή του αριθμού των «επιτυχιών» (πραγματοποίησης ενός γεγονότος) σε ένα ορισμένο χρονικό διάστημα είναι η κατανομή Poisson.

Ας θεωρήσουμε ένα δείγμα από n παρατηρήσεις $y_1, y_2, ..., y_n$, οι οποίες αποτελούν την πραγμάτωση των τυχαίων ανεξάρτητων μεταβλητών Y_i, όπου $\mathrm{Y}_i \sim Poisson(\mu_i)$ και θέλουμε σύμφωνα με τη θεωρία των γενικευμένων γραμμικών μοντέλων η μέση τιμή της τυχαίας μεταβλητή Y_i να είναι μια γραμμική συνάρτηση των επεξηγηματικών μεταβλητών του μοντέλου $\boldsymbol{x}_i$, δηλαδή $\mu_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. Ωστόσο, ο περιορισμός $\mu_i > 0$, που επιβάλλεται στην κατανομή Poisson δεν ικανοποιείται καθώς ο όρος $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ μπορεί να λάβει οποιαδήποτε πραγματική τιμή, ακόμη και αρνητική. Συνεπώς, υπάρχει η ανάγκη ορισμού μιας κατάλληλης συνάρτησης σύνδεσης, ενός μετασχηματισμού δηλαδή της παραμέτρου μ_i, ώστε να εξασφαλιστεί ο παραπάνω περιορισμός. Η πιο συνηθισμένη επιλογή είναι η συνδετική συνάρτηση $g(\mu) = \log(\mu)$. Το μοντέλο παλινδρόμησης Poisson καλείται και μοντέλο Poisson log-linear, λόγω της συνάρτησης σύνδεσής του. Η δομή του είναι η ακόλουθη

$$\mathrm{Y}_i \overset{iid}{\sim} Poisson(\mu_i)\,,\ \log(\mu_i) = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\,,\ i = 1, 2, ..., n\,.$$

4.1.1 Εκτίμηση των Συντελεστών

Η εκτίμηση των παραμέτρων του μοντέλου γίνεται με τη μέθοδο μέγιστης πιθανοφάνειας . Η συνάρτηση πιθανοφάνειας L δίνεται από τη σχέση

$$L(\boldsymbol{y}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i\,!}.$$

Ο λογάριθμος της συνάρτησης πιθανοφάνειας είναι

$$l = \log L(y,\beta) = \log \prod_{i=1}^{n} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \sum_{i=1}^{n}\{-\mu_i + y_i \log \mu_i - \log(y_i!)\}$$ και

αντικαθιστώντας την έκφραση $\mu_i = e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}}$ λαμβάνει τη μορφή

$$l = \log L(\boldsymbol{y},\boldsymbol{\beta}) = \sum_{i=1}^{n}\left\{-e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}} + y_i \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} - \log(y_i!)\right\}.$$

Παραγωγίζοντας την παραπάνω σχέση ως προς τις άγνωστες παραμέτρους του μοντέλου θα έχουμε ότι

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n}\left\{-x_{ij}e^{\boldsymbol{x}_i^T\boldsymbol{\beta}} + y_i x_{ij}\right\} = \sum_{i=1}^{n}\left\{x_{ij}\left(y_i - e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}\right)\right\}, j = 0,1,...,k.$$

Συνεπώς, θέτοντας τις παραπάνω σχέσεις ίσες με 0 προκύπτει ένα σύστημα $p = k+1$ εξισώσεων με $k+1$ αγνώστους. Επιλύοντας το σύστημα

$$\sum_{i=1}^{n}\left\{x_{ij}\left(y_i - e^{\boldsymbol{x}_i^T\boldsymbol{b}}\right)\right\} = 0, j = 0,1,2,...,k \quad (4.1)$$

βρίσκουμε τις εκτιμήτριες των παραμέτρων β_j, έστω b_j.

Σε μορφή πινάκων η παραπάνω σχέση μπορεί να γραφεί ως

$\mathbf{X}^{\mathrm{T}}(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) = \mathbf{0}$, όπου $\hat{\mu}_i^{\mathrm{T}} = (\hat{\mu}_1, \hat{\mu}_2, ..., \hat{\mu}_{\mathrm{n}})$ με $\hat{\mu}_i = e^{\boldsymbol{x}_i^T\boldsymbol{b}}$,

η οποία λύνεται χρησιμοποιώντας επαναληπτικές μεθόδους, καθώς είναι μη γραμμική.

4.1.2 Ερμηνεία των Συντελεστών

Στο μοντέλο Poisson, ο λογάριθμος της αναμενόμενης τιμής της τυχαίας μεταβλητής Y συνδέεται με μια γραμμική σχέση των επεξηγηματικών μεταβλητών του μοντέλου και αυτό έχει σαν αποτέλεσμα μια εκθετική επίδραση κάθε επεξηγηματικής μεταβλητής X_j στη μεταβολή της αναμενόμενης τιμής της Y.

Αν θεωρήσουμε το πιο απλό παράδειγμα μοντέλου με μια μόνο επεξηγηματική μεταβλητή, τότε η αναμενόμενη τιμή της Y μπορεί να εκφραστεί από την παρακάτω σχέση

$\log \mu_i = \beta_0 + \beta_1 x_i \Leftrightarrow \mu_i = e^{\beta_0}e^{\beta_1 x_i}$.

Παρατηρούμε ότι η ποσότητα e^{β_0} εκφράζει την αναμενόμενη τιμή της Y, όταν η επεξηγηματική μεταβλητή X_i είναι ίση με 0. Η ερμηνεία του συντελεστή β_1 διαφέρει από αυτή του γραμμικού μοντέλου. Για την ερμηνεία του λοιπόν αρχικά θα ορίσουμε ως $\mu(x) = E(\mathrm{Y} \mid \mathrm{X} = x)$ να είναι η αναμενόμενη μέση τιμή της Y δεδομένου ότι η

επεξηγηματική μεταβλητή του μοντέλου είναι σταθερή. Τότε αν η μεταβλητή X αυξηθεί κατά μια μονάδα, δηλαδή $\mathrm{X} = x + 1$ θα έχουμε το εξής αποτέλεσμα

$$\log(\mu(x+1)) - \log(\mu(x)) = \beta_1 \Leftrightarrow \mu(x+1) = e^{\beta_1}\mu(x).$$

Συνεπώς, η αναμενόμενη τιμή της Y για μια μονάδα αύξησης της επεξηγηματικής μεταβλητής X θα είναι ίση με e^{β_1} φορές την $\mu(x) = E(\mathrm{Y} \mid \mathrm{X} = x)$.

Γενικεύοντας για περισσότερες από μια επεξηγηματικές μεταβλητές, έστω $j = 1,...,k$ μεταβλητές μπορούμε πούμε ότι κάθε μια από τις ποσότητες e^{β_j}, εκφράζει την αναμενόμενη πολλαπλασιαστική μεταβολή της Y για μια μονάδα αύξησης της αντίστοιχης επεξηγηματικής μεταβλητής του μοντέλου X_j, δεδομένου ότι οι υπόλοιπες επεξηγηματικές μεταβλητές παραμένουν σταθερές.

Στην περίπτωση που η επεξηγηματική μεταβλητή X είναι κατηγορική με K κατηγορίες, τότε η γραμμική προβλέπουσα εκφράζεται ως μια γραμμική συνάρτηση $(K-1)$ στο πλήθος εικονικών μεταβλητών ή ψευδομεταβλητών (*dummy variables*) οι οποίες συμβολίζονται με D_j, $j = 2,...,K$. Θεωρούμε την πρώτη κατηγορία ως την κατηγορία αναφοράς και κατά συνέπεια θέτουμε τον συντελεστή αυτής ίσο με 0, $\beta_1 = 0$.

Τότε, το μοντέλο λαμβάνει την παρακάτω έκφραση

$$\log(\mu_i) = \beta_0 + \sum_{j=2}^{K} \beta_j D_{ij} \Leftrightarrow \mu_i = e^{\beta_0} \exp\left(\sum_{j=2}^{K} \beta_j D_{ij}\right) = B_0 \prod_{j=2}^{K} B_j^{D_{ij}},$$

όπου $B_j = e^{\beta_j}$ για $j \in \{0, 2, 3, \ldots, K\}$.

Όταν η $i-$ παρατήρηση ανήκει στην πρώτη κατηγορία, τότε $\mu_i = e^{\beta_0} = B_0$, ενώ αν η $i-$ παρατήρηση ανήκει σε κάποια από τις j κατηγορίες ($\mathrm{X}_i = j$), τότε
$\mathrm{E}(\mathrm{Y} \mid \mathrm{X} = j) = B_0 B_j = B_j\, \mathrm{E}(\mathrm{Y} \mid \mathrm{X} = 1)$.

Συνεπώς, μπορούμε να πούμε ότι η ποσότητα $B_j = e^{\beta_j}$ είναι η σχετική μεταβολή της αναμενόμενης τιμής της Y, όταν μια παρατήρηση ανήκει στην $j-$οστή κατηγορία της ανεξάρτητης μεταβλητής X συγκριτικά με την κατηγορία αναφοράς.

Στην περίπτωση της πολλαπλής παλινδρόμησης Poisson η ερμηνεία των συντελεστών είναι η ίδια ακριβώς, απλά προκειμένου να δώσουμε μια ερμηνεία για την αναμενόμενη τιμή μεταβλητής απόκρισης θα πρέπει σε κάθε αλλαγή κάθε επεξηγηματικής μεταβλητής οι υπόλοιπες να παραμένουν σταθερές.

4.1.3 Έλεγχος των Συντελεστών και Διαστήματα Εμπιστοσύνης

Οι εκτιμήτριες της μέγιστης πιθανοφάνειας ακολουθούν ασυμπτωτικά την Κανονική κατανομή

$$\mathrm{Z} = \frac{b_j - \beta_j}{\left(\boldsymbol{J}^{-1}(\boldsymbol{b})_{jj}\right)^{\frac{1}{2}}} \dot{\sim} \mathrm{N}(0,1),\ j = 0,1,...,k, \quad (4.2)$$

όπου $\boldsymbol{J}^{-1}(\boldsymbol{b})$ ο αντίστροφος πίνακας του παρατηρούμενου πίνακα πληροφορίας $\boldsymbol{J}(\boldsymbol{b}) = \mathbf{X}^{\mathrm{T}}\hat{\mathbf{W}}\mathbf{X}$. Ο πίνακας $\hat{\mathbf{W}}$ είναι ένας διαγώνιος πίνακας $\hat{\mathbf{W}} = diag(e^{\boldsymbol{x}_i^T \boldsymbol{b}})$.
Το παραπάνω αποτέλεσμα (4.2) μπορεί να χρησιμοποιηθεί και για την κατασκευή ενός $100(1-a)\%-$διαστήματος εμπιστοσύνης για την παράμετρο β_j, όπως $b_j \pm z_{a/2} se(b_j)$, καθώς και του $100(1-a)\%-$διαστήματος εμπιστοσύνης $\exp\{b_j \pm z_{a/2} se(b_j)\}$ για την e^{β_j}. Με $z_{a/2}$ συμβολίζουμε το άνω $100(a/2)$ ποσοστιαίο σημείο της $N(0,1)$ κατανομής και με $se(b_j)$ συμβολίζουμε το τυπικό σφάλμα του συντελεστή b_j και ορίζεται από τη σχέση
$se(b_j) = \left(J^{-1}(\boldsymbol{b})_{jj}\right)^{\frac{1}{2}}$, όπου $J^{-1}(\boldsymbol{b})_{jj}$ συμβολίζουμε το $j-$οστό διαγώνιο στοιχείο του πίνακα $\boldsymbol{J}^{-1}(\boldsymbol{b})$.

4.1.4 Κριτήρια Καλής Προσαρμογής

Ελεγχοσυνάρτηση deviance

Αν θεωρήσουμε $\hat{\mu}_i = e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}}$ να είναι η εκτιμήτρια μέγιστης πιθανοφάνειας για τα μ_i στο υποψήφιο μοντέλο και $\tilde{\mu}_i = y_i$ η εκτιμήτρια μέγιστης πιθανοφάνειας για το κορεσμένο μοντέλο, τότε η ελεγχοσυνάρτηση deviance δίνεται από τη σχέση

$$D = 2\sum_{i=1}^{n}\{y_i \log(y_i) - y_i - \log(y_i!) - y_i \log(\hat{\mu}_i) + \hat{\mu}_i + \log(y_i!)\} \Leftrightarrow$$

$$D = 2\sum_{i=1}^{n}\{y_i \log(y_i) - y_i - y_i \log(\hat{\mu}_i) + \hat{\mu}_i\} \Leftrightarrow$$

$$D = 2\left\{\sum_{i=1}^{n} y_i \log(y_i/\hat{\mu}_i) - \sum_{i=1}^{n}(y_i - \hat{\mu}_i)\right\}.$$

Η παραπάνω σχέση απλοποιείται ως $D = 2\sum_{i=1}^{n} y_i \log(y_i/\hat{\mu}_i)$, αφού ο όρος $\sum_{i=1}^{n}(y_i - \hat{\mu}_i)$ μηδενίζεται λόγω της σχέσης (4.1).

Pearson

Η στατιστική συνάρτηση X^2 του Pearson δίνεται από τη σχέση

$$X^2 = \sum_{i=1}^{n} \frac{\left(y_i - \hat{\mu}_i\right)^2}{\hat{\mu}_i}.$$

Οι ελεγχοσυναρτήσεις deviance και του Pearson είναι ασυμπτωτικά ισοδύναμες και ακολουθούν και οι δύο την κατανομή X^2, όταν το προσαρμοσμένο μοντέλο είναι ορθό. Φυσικά, η deviance έχει το πλεονέκτημα ότι μπορεί να χρησιμοποιηθεί και για την σύγκριση μεταξύ δύο ιεραρχικών μοντέλων.

4.1.5 Κριτήρια Επιλογής Μοντέλου

Όπως και στη γενική περίπτωση των γενικευμένων γραμμικών μοντέλων, έτσι και στη παλινδρόμηση Poisson μπορούν να χρησιμοποιηθούν τα κριτήρια AIC και BIC για την επιλογή του βέλτιστου μοντέλου.

Κριτήριο AIC

Για την περίπτωση της παλινδρόμησης Poisson το κριτήριο AIC δίνεται από τη σχέση

$$AIC = -2\sum_{i=1}^{n}\left\{-\hat{\mu}_i + y_i \log \hat{\mu}_i - \log(y_i!)\right\} + 2p.$$

Κριτήριο BIC

Για την περίπτωση της παλινδρόμησης Poisson το κριτήριο BIC δίνεται από τη σχέση

$$BIC = -2\sum_{i=1}^{n}\left\{-\hat{\mu}_i + y_i \log \hat{\mu}_i - \log(y_i!)\right\} + p \log n.$$

4.1.6 Διαγνωστικοί Έλεγχοι

Υπόλοιπα

Στη παλινδρόμηση Poisson χρησιμοποιούνται κυρίως τα υπόλοιπα Pearson και τα υπόλοιπα deviance, τα οποία θα ορίσουμε για το συγκεκριμένο μοντέλο.

Υπόλοιπα Pearson

Τα υπόλοιπα Pearson για την περίπτωση της Poisson παλινδρόμησης ορίζονται από τη σχέση

$$\varepsilon_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}, \text{ όπου } \hat{\mu}_i = e^{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b}}$$

Το άθροισμα τετραγώνων αυτών των υπολοίπων αποτελεί τον γνωστό έλεγχο καλής προσαρμογής X^2 του Pearson,

$$\mathrm{X}^2 = \sum_{i=1}^{n}\left(\frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}\right)^2 = \sum_{i=1}^{n}\frac{\left(y_i - \hat{\mu}_i\right)^2}{\hat{\mu}_i}.$$

Υπόλοιπα deviance

Τα υπόλοιπα deviance ορίζονται ως

$$\varepsilon_i^D = sign(y_i - \hat{\mu}_i)\sqrt{d_i(y_i, \hat{\mu}_i)},$$

όπου $d_i(y_i, \hat{\mu}_i) = 2\{y_i \log(y_i / \hat{\mu}_i) - (y_i - \hat{\mu}_i)\}$ αποτελεί τη συμβολή της $i-$οστής παρατήρησης στην ελεγχοσυνάρτηση deviance

$$D = 2\left\{\sum_{i=1}^{n} y_i \log(y_i / \hat{\mu}_i) - \sum_{i=1}^{n}\left(y_i - \hat{\mu}_i\right)\right\}.$$

Υπάρχουν και διάφοροι άλλοι διαγνωστικοί έλεγχοι, τους οποίους αναπτύξαμε στο 1^0 Κεφάλαιο, οι οποίοι μπορούν να χρησιμοποιηθούν και στην περίπτωση της παλινδρόμησης Poisson.

4.1.7 Το Πρόβλημα της Υπερμεταβλητότητας για το Μοντέλο Poisson

Παρ' όλο, που το μοντέλο της παλινδρόμησης Poisson είναι το πιο διαδεδομένο για απαριθμητά δεδομένα, πολλές φορές η προσαρμογή του φαίνεται να μην είναι ικανοποιητική. Σ' αυτήν την περίπτωση εμφανίζεται το πρόβλημα της υπερμεταβλητότητας (*overdispersion*). Αυτό σημαίνει ότι μετά την προσαρμογή του μοντέλου τα δεδομένα παρουσιάζουν μεγαλύτερη διασπορά από τις θεωρητικές $Var(\mathrm{Y}_i) = \mu_i = \mathrm{E}(\mathrm{Y}_i)$. Μία ένδειξη της παρουσίας της υπερμεταβλητότητας είναι όταν η τιμή της ελεγχοσυνάρτησης $deviance / df$ είναι μεγαλύτερη της μονάδας. Ωστόσο, ο Lindsey (1999) θεωρεί ως καλύτερο κριτήριο για την ύπαρξη υπερμεταβλητότητας, η τιμή της στατιστικής συνάρτησης $deviance$ να είναι τουλάχιστον διπλάσια των βαθμών ελευθερίας $df(=n-p)$, το οποίο ισοδυναμεί με τη χρήση του κριτηρίου AIC.

Μερικοί λόγοι, που οδηγούν σε αυτό το φαινόμενο είναι οι εξής:

1. Η επιλογή της κατανομής Poisson δεν είναι η κατάλληλη για να περιγράψει τα δεδομένα.
2. Δεν έχει καθοριστεί σωστά η γραμμική προβλέπουσα του μοντέλου. Πιθανότατα, να μην έχουν συμπεριληφθεί ενδεχόμενες αλληλεπιδράσεις μεταξύ των επεξηγηματικών μεταβλητών ή κάποιες στατιστικά σημαντικές επεξηγηματικές μεταβλητές να είναι εκτός μοντέλου ή να χρειάζεται κάποιος μετασχηματισμός κάποιας εκ των μεταβλητών, όπως για παράδειγμα η λογαριθμική συνάρτηση.
3. Ενδέχεται οι τυχαίες μεταβλητές Y_i να μην είναι ανεξάρτητες.
4. Υπάρχει ένα ή περισσότερα έκτροπα σημεία (*outliers*) στα δεδομένα.

Σε περιπτώσεις, όπου η υπερμεταβλητότητα οφείλεται στο ότι δεν έχει επιλεχθεί η κατάλληλη κατανομή τότε θα χρησιμοποιήσουμε κάποια άλλη κατανομή κατάλληλη να περιγράψει τα δεδομένα και να επιτρέψει επιπλέον μεταβλητότητα. Στην περίπτωση των απαριθμητών δεδομένων μπορεί να χρησιμοποιηθεί η αρνητική διωνυμική κατανομή, η οποία επιτρέπει μεγαλύτερη διακύμανση από την κατανομή Poisson. Επειδή, η Poisson κατανομή αποτελεί ειδική περίπτωση της αρνητικής διωνυμικής, μπορεί να χρησιμοποιηθεί και ένας έλεγχος του λόγου της πιθανοφάνειας για την επιλογή μεταξύ των δύο. Επιπλέον, για την προσαρμογή του μοντέλου χρησιμοποιούνται οι μέθοδοι, που έχουμε αναπτύξει στα πλαίσια των γενικευμένων γραμμικών μοντέλων, καθώς η αρνητική διωνυμική κατανομή ανήκει στην εκθετική οικογένεια κατανομών.

Ωστόσο, αν η επιλογή της κατανομής Poisson είναι σωστή και το μοντέλο έχει καθοριστεί ορθά, τότε το φαινόμενο της υπερμεταβλητότητας οφείλεται σε άλλες πιθανές παραμέτρους, όπως αναφέραμε παραπάνω (δηλαδή υπάρχουν έκτροπα σημεία ή οι τυχαίες μεταβλητές Y_i δεν είναι ανεξάρτητες μεταξύ τους). Σε αυτήν την περίπτωση η αυξημένη μεταβλητότητα μπορεί να μοντελοποιηθεί μέσω της σχέσης $V(\mathrm{Y}_i) = \varphi V(\mu_i)$, όπου $\mu_i = \mathrm{E}(\mathrm{Y}_i)$, $V(\mu_i) = \mu_i$ και φ η παράμετρος μεταβλητότητας.

Η παράμετρος μεταβλητότητας φ εκτιμάται από την ελεγχοσυνάρτηση $deviance/df$, δηλαδή με τον λόγο της ελεγχοσυνάρτησης deviance και των βαθμών ελευθερίας της X^2 κατανομής που ακολουθεί η deviance ή χρησιμοποιώντας την ελεγχοσυνάρτηση καλής προσαρμογής του Pearson στη θέση της ελεγχοσυνάρτησης deviance, δηλαδή

$$\tilde{\varphi} = \frac{\mathrm{X}^2}{d} \Rightarrow \tilde{\varphi} = \frac{1}{n-k-1}\sum_{i=1}^{n}\frac{(y_i - \hat{\mu}_i)^2}{\mathrm{V}(\hat{\mu}_i)},$$

όπου οι τιμές $\hat{\mu}_i$ προκύπτουν από την προσαρμογή του μοντέλου με όλες τις k διαθέσιμες επεξηγηματικές μεταβλητές του μοντέλου και $\mathrm{V}(\hat{\mu}_i)$ είναι η συνάρτηση διασποράς. Στην περίπτωση της Κανονικής κατανομής $\tilde{\varphi} = S^2$, η αμερόληπτη εκτιμήτρια της σ^2, αφού $\mathrm{V}(\hat{\mu}_i) = 1$.

4.2 Το Μοντέλο της Αρνητικής Διωνυμικής Κατανομής

Έχουμε δείξει ότι μια κατάλληλη κατανομή για την περιγραφή απαριθμητών δεδομένων είναι η κατανομή Poisson. Ωστόσο, για την περίπτωση της κατανομής Poisson γνωρίζουμε ότι η διασπορά θα πρέπει να είναι ίση με τη μέση τιμή της μεταβλητής. Πολλές φορές, όταν εφαρμοστεί το μοντέλο Poisson στα δεδομένα η υπόθεση αυτή δεν ευσταθεί, καθώς παρατηρείται το φαινόμενο της υπερμεταβλητότητας. Σ' αυτή την περίπτωση μπορούμε να χρησιμοποιήσουμε μια άλλη κατανομή, ώστε να καλυφθεί αυτή η επιπλέον μεταβλητότητα. Μια κατάλληλη κατανομή είναι η αρνητική διωνυμική κατανομή και επιτρέπει μεγαλύτερη διακύμανση από την κατανομή Poisson.

4.2.1 Αρνητική Διωνυμική Κατανομή

Μια διακριτή τυχαία μεταβλητή Y λέμε ότι ακολουθεί την αρνητική διωνυμική κατανομή (NB), δηλαδή $\mathrm{Y} \sim \mathrm{NB}(\pi, r)$, αν η συνάρτηση μάζας πιθανότητας δίνεται από τη σχέση

$$f(y;\pi,r) = \frac{\Gamma(y+r)}{y!\Gamma(r)}\pi^r(1-\pi)^y, \quad (4.3)$$

όπου $y = 0, 1, ..., r$ και $r = 1, 2, \ldots,$ $0 \le \pi \le 1$.

Η μέση τιμή και η διασπορά αυτής της κατανομής δίνονται από τις σχέσεις

$$\mathrm{E(Y)} = r(1-\pi)/\pi \quad (4.4)$$

$$V(\mathrm{Y}) = r(1-\pi)\pi^2 . (4.5)$$

Συχνά, η παράμετρος $r = \mathrm{N} \in \mathbb{Z}$. Σε αυτές τις περιπτώσεις, η κατανομή (4.3) χρησιμοποιείται για να περιγράψει τον αριθμό δοκιμών Bernoulli, που πρέπει να γίνουν έως ότου επιτευχθούν N επιτυχίες, με πιθανότητα επιτυχίας π. Γενικεύοντας, αν θεωρήσουμε ότι η παράμετρος $r > 0$, τότε η κατανομή καλείται κατανομή Poyla. Η κατανομή Poyla χρησιμοποιείται για την μοντελοποίηση δεδομένων (απαριθμητών), που παρουσιάζουν αυξημένη μεταβλητότητα, καθώς ο δείκτης διασποράς DI (*dispersion index*) δίνεται από τη σχέση

$$DI = \frac{V(\mathrm{Y})}{E(\mathrm{Y})} = \frac{1}{\pi}.$$

Το μοντέλο της αρνητικής διωνυμικής κατανομής μπορεί να εκφραστεί ως μία μίξη της κατανομής Poisson και Gamma και έτσι θα έχει τη μορφή

$\mathrm{Y} \mid u \sim Poisson(\mu u)$ και $u \sim Gamma(r, r)$,

όπου $\mu = r(1-\pi)/\pi$.

Με την παραμετροποίηση $\mu = r(1-\pi)/\pi$ η μέση τιμή της Y θα είναι $\mathrm{E}(\mathrm{Y}) = \mu$ και η διασπορά της $V(\mathrm{Y}) = \mu(\mu + r)$ σύμφωνα με τις σχέσεις (4.4), (4.5). Συνεπώς, η περιθώρια κατανομή της Y θα είναι

$$f(y) = \int_0^\infty f(y \mid u) f(u) = \frac{\Gamma(y+r)}{y!\Gamma(\mathrm{r})} \left(\frac{r}{r+\mu} \right)^r \left(\frac{\mu}{r+\mu} \right)^y,$$

η οποία είναι η αρνητική διωνυμική κατανομή με παραμέτρους $r/(r+\mu)$ και r.

ΚΕΦΑΛΑΙΟ 5

ΤΟ ΣΤΑΤΙΣΤΙΚΟ ΠΑΚΕΤΟ R ΚΑΙ Η ΧΡΗΣΗ ΤΟΥ

5.1 Γενικά για το Στατιστικό Πακέτο R

Η μελέτη των γενικευμένων γραμμικών μοντέλων απαιτεί πολλές φορές πολύπλοκους υπολογισμούς και εξειδικευμένα γραφήματα. Ωστόσο, η χρήση διαφόρων στατιστικών πακέτων σε υπολογιστή, όπως το Minitab, το SPSS, η R κ.λ.π. μας βοηθάει να βγάλουμε συμπεράσματα γρήγορα και εύκολα κατά τη μελέτη πολλών μοντέλων.

Για την κατανόηση της θεωρίας και την μελέτη των γενικευμένων γραμμικών μοντέλων θα χρησιμοποιήσουμε το στατιστικό πακέτο R, που αποτελεί ένα από τα πιο διαδεδομένα στατιστικά πακέτα στο χώρο της Στατιστικής.

Πρόκειται για μια γλώσσα προγραμματισμού και ένα ολοκληρωμένο περιβάλλον εργασίας, που χρησιμεύει κυρίως για ανάλυση δεδομένων και εφαρμογή διαφόρων «κλασσικών» και σύγχρονων στατιστικών τεχνικών.

Η R μας δίνει τη δυνατότητα να δημιουργήσουμε απλό κώδικα, δηλαδή συναρτήσεις κατάλληλες για διάφορους στατιστικούς υπολογισμούς. Μπορεί να χρησιμοποιηθεί είτε με κατευθείαν εντολές, είτε με συναρτήσεις, που ο χρήστης προγραμματίζει για την επίλυση των πολύπλοκων στατιστικών προβλημάτων. Αυτή είναι και η βασική διαφοροποίηση που εμφανίζει συγκριτικά με άλλα στατιστικά πακέτα. Επίσης, ο χρήστης μπορεί να χρησιμοποιήσει και έτοιμα προγράμματα, που είναι ενσωματωμένα μέσα σε βιβλιοθήκες. Η ποικιλία τέτοιων προγραμμάτων είναι τεράστια.

Η R μας εφοδιάζει μεταξύ άλλων με:

- αποτελεσματική διαχείριση και αποθήκευση δεδομένων
- γραφικές λειτουργίες για την ανάλυση δεδομένων και την απεικόνιση είτε σε υπολογιστή είτε σε έντυπη μορφή
- εργαλεία για χειρισμό πινάκων πολλών διαστάσεων
- μια απλή και αποτελεσματική γλώσσα προγραμματισμού, η οποία καλείται «S».
- όλα τα απαραίτητα εργαλεία για τη δημιουργία συναρτήσεων

Επιπλέον, αποτελεί ελεύθερο λογισμικό και διατίθεται στην ιστοσελίδα http://www.r-project.org ή από ένα από τα άλλα πρότυπα (*mirrors*) του CRAN (*Comprehensive R Archive*) http://cran.r-project.org, το οποίο είναι ένα δίκτυο διανομής της R σε πολλά μέρη του κόσμου μέσω Διαδικτύου.
Η R υποστηρίζει πολλές πλατφόρμες και λειτουργικά, όπως Linux, Mac OS και Windows.

5.2 Γενικευμένα Γραμμικά Μοντέλα με Χρήση της R

Για την προσαρμογή ενός γενικευμένου γραμμικού μοντέλου στην R χρησιμοποιούμε τη συνάρτηση **glm()**, η οποία έχει τη μορφή:

```
glm(formula, family = gaussian, data, weights, subset,
    na.action, start = NULL, etastart, mustart, offset,
    control = list(...), model = TRUE, method = "glm.fit",
    x = FALSE, y = TRUE, contrasts = NULL, ...)
```

Η συνάρτηση glm() βρίσκεται στην βιβλιοθήκη stats της R και μπορεί να γίνει χρήση αυτής εκτελώντας στην επιφάνεια εργασίας της R την εντολή *library(stats)*.

5.2.1 Παράμετροι της Συνάρτησης glm()

Περιλαμβάνει κάποιες παραμέτρους, οι οποίες λαμβάνουν ορίσματα από το χρήστη. Μόλις χρησιμοποιήσουμε την συνάρτηση λοιπόν στην R θα μας επιστραφεί ένα αντικείμενο (*object*) της κλάσης glm().
Παρακάτω δίνουμε μια σύντομη περιγραφή των παραμέτρων εισόδου της συνάρτησης glm().

formula: είναι μια συμβολική αναπαράσταση του μοντέλου, που θέλουμε να προσαρμόσουμε. Η δομή της είναι η ίδια με την περίπτωση του απλού γραμμικού μοντέλου, σύμφωνα με τη συνάρτηση lm(). Στην περίπτωση των γενικευμένων γραμμικών μοντέλων εκφράζουμε σε συμβολική μορφή τη γραμμική προβλέπουσα του μοντέλου. Για παράδειγμα, για το μοντέλο (Y ~ κατανομή από την εκθετική οικογένεια κατανομών με γραμμική προβλέπουσα $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$) η παράμετρος formula θα έχει τη μορφή $y \sim x_1 + x_2$, όπου στα διανύσματα y, x_1, x_2 έχουμε αποθηκεύσει τις τιμές της μεταβλητής απόκρισης Y και των επεξηγηματικών μεταβλητών X_1, X_2 του μοντέλου.

Κάποια άλλα σύμβολα, που χρησιμοποιούνται στην παράμετρο formula εκτός του '+' είναι τα εξής:

- $a : b$ αν υπάρχει συσχέτιση μεταξύ των μεταβλητών a και b .
- $a * b$ είναι ισοδύναμη με την παραπάνω έκφραση.
- $-$ μπορεί να χρησιμοποιηθεί για να αφαιρέσουμε κάποιον όρο από το μοντέλο.
- 1 μπορεί να χρησιμοποιηθεί για να αφαιρέσουμε το σταθερό όρο από το μοντέλο (π.χ. $y \sim x_1 + x_2 - 1$).
- 0 όπως και η παραπάνω έκφραση μπορεί να χρησιμοποιηθεί για να αφαιρέσουμε το σταθερό όρο από το μοντέλο (π.χ. $y \sim x_1 + x_2 + 0$ ή $y \sim 0 + x_1 + x_2$).

family: είναι μια συνάρτηση, η οποία περιλαμβάνει δύο παραμέτρους. Την κατανομή της μεταβλητής απόκρισης και την συνάρτηση σύνδεσης.

Η παράμετρος *family* μπορεί να πάρει τις ακόλουθες τιμές χαρακτήρων:

binomial (*logit, probit, log, complementary log-log*)

gaussian (*identity, log, inverse*)

Gamma (*identity, inverse, log*)

inverse.gaussian ($1/\mu^2$, *identity, inverse, log*)

poisson (*identity, log, square root*)

quasi (*logit, probit, complementary log-log, identity, inverse, log,* $1/\mu^2$, *square root*)

Στις παρενθέσεις αναφέρονται όλες οι πιθανές συναρτήσεις σύνδεσης για κάθε μια από τις παραπάνω κατανομές. Αν δεν οριστεί συνάρτηση σύνδεσης η R για κάθε μία από αυτές χρησιμοποιεί την κανονική συνάρτηση σύνδεσης, δηλαδή

binomial (*logit*)

gaussian (*identity*)

Gamma (*inverse*)

inverse.gaussian ($1/\mu^2$)

poisson (*log*)

quasi (*identity, variance = "constant"*)

quasibinomial (*logit*)

quasipoisson (*log*)

Η R είναι "*case-sensitive*", δηλαδή κάνει διαχωρισμό μεταξύ μικρών και κεφαλαιών χαρακτήρων. Συνεπώς, αν θέλουμε η τιμή της παραμέτρου *family* να είναι η Gamma θα πρέπει να γραφεί έτσι ακριβώς και όχι gamma.

Για να προσαρμόσουμε ένα μοντέλο στην R με διαφορετικές συναρτήσεις σύνδεσης από τις κανονικές θα πρέπει απλά να εισάγουμε την συνάρτηση σύνδεσης που

θέλουμε στην παράμετρο *family*. Για παράδειγμα, αν θέλουμε να προσαρμόσουμε στα δεδομένα μας το μοντέλο probit θα χρησιμοποιήσουμε την παρακάτω εντολή.

```
probit.glm<-glm(y~x,family=binomial(link="probit"))
```

Στην περίπτωση της λογιστικής παλινδρόμησης η εντολή θα είναι

```
logit.glm<-glm(y~x,family=binomial(link= "logit")) ή

logit.glm<-glm(y~x,family=binomial())
```

καθώς η προκαθορισμένη τιμή για την διωνυμική κατανομή είναι η συνάρτηση σύνδεσης logit.

data:είναι τα δεδομένα μας. Αφορούν τις παρατηρήσεις όλων των μεταβλητών του προβλήματος (ή του πειράματος).

offset:είναι μια συνιστώσα της γραμμικής προβλέπουσας, η οποία είναι γνωστή εξ' αρχής και για αυτό το λόγο δεν χρειάζεται να εκτιμηθεί κάποιος συντελεστής για αυτή από τα δεδομένα. Προσδιορίζοντας, λοιπόν την παράμετρο *offset* στη συνάρτηση glm θεωρείται σταθερός όρος κατά την προσαρμογή του μοντέλου.

Οι παράμετροι *weights, subset, start, etastart* και *mustart* αφορούν την μέθοδο εκτίμησης των παραμέτρων του μοντέλου. Όπως έχουμε αναλύσει σε προηγούμενο κεφάλαιο στα γενικευμένα γραμμικά μοντέλα χρησιμοποιείται η μέθοδος επαναληπτικών σταθμισμένων ελαχίστων τετραγώνων (IWST), η οποία, σε κάθε επανάληψη $(m+1)$, χρησιμοποιεί τον επαναληπτικό τύπο $\boldsymbol{b}^{(m+1)} = \left(\mathbf{X}^{\mathrm{T}}\hat{\mathbf{W}}^{(m)}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\hat{\mathbf{W}}^{(m)}\mathbf{z}^{(m)}$, όπου $\mathbf{z}^{(m)}$ είναι διάνυσμα μεγέθους n με στοιχεία $z_i^{(m)} = \eta_i^{(m)} + (y_i - \mu_i^{(m)})\left(g'(\mu_i^{(m)})\right)^2$ και $\hat{\mathbf{W}}^{(m)} = diag(w_{ii}^{(m)})$ διαγώνιος πίνακας με στοιχεία $w_{ii}^{(m)} = \dfrac{1}{V(\mu_i^{(m)})(g'(\mu_i^{(m)}))^2}$.

Επισημαίνουμε ότι για γενικευμένα γραμμικά μοντέλα με κανονικές συναρτήσεις σύνδεσης ο παραπάνω αλγόριθμος είναι ουσιαστικά η επαναληπτική διαδικασία Newton-Raphson.

Ο αλγόριθμος για την μέθοδο IWST περιλαμβάνει τα παρακάτω βήματα:

1. Αρχικά, δίνουμε κάποιες αρχικές τιμές για τα μ_i ($\mu_i^{(0)}$).
2. Υπολογίζουμε τις τιμές για τα $w_{ii}^{(0)}$.
3. Υπολογίζουμε τα $z_i^{(0)}$.

4. Υπολογίζουμε τις εκτιμήσεις των παραμέτρων $\boldsymbol{b}^{(1)}$ με βάση των επαναληπτικό τύπο που δόθηκε παραπάνω.
5. Επαναλαμβάνουμε τα βήματα 2, 3 και 4 έως ότου ο αλγόριθμος συγκλίνει.

weights: εάν το επιθυμούμε δίνουμε ένα διάνυσμα των αρχικών τιμών w_{ii}, δηλαδή το $w_{ii}^{(0)}$.

***subset*:** εφόσον το επιθυμούμε για την προσαρμογή του μοντέλου να χρησιμοποιήσουμε ένα υποσύνολο των τιμών της μεταβλητής απόκρισης.

Οι τρείς παράμετροι, που παρουσιάζονται παρακάτω ουσιαστικά είναι ένας τρόπος να ορίσει ο χρήστης από ποιο σημείο θα ξεκινήσει ο αλγόριθμος.

***start*:** ένα διάνυσμα αρχικών τιμών για τις παραμέτρους του μοντέλου.

***etastart*:** ένα διάνυσμα αρχικών τιμών για τη γραμμική προβλέπουσα του μοντέλου.

mustart: ένα διάνυσμα αρχικών τιμών για την μέση τιμή της μεταβλητής απόκρισης.

method: προσδιορίζει την μέθοδο, που χρησιμοποιείται για τον υπολογισμό των εκτιμητριών των παραμέτρων του μοντέλου. Με την προκαθορισμένη τιμή *method="glm.fit"* η R χρησιμοποιεί την μέθοδο επαναληπτικών σταθμισμένων ελαχίστων τετραγώνων. Αν ορίσουμε ως *method="model.frame"* θα επιστραφεί στην R ένα πλαίσιο δεδομένων με τις παρατηρούμενες τιμές των μεταβλητών του μοντέλου.

5.2.2 Το Αντικείμενο της Κλάσης glm()

Ένα αντικείμενο της κλάσης glm() είναι μια λίστα, η οποία περιλαμβάνει πολλά επιμέρους αποτελέσματα.

coefficients(object): ένα διάνυσμα, που περιλαμβάνει τους εκτιμημένους συντελεστές του μοντέλου. Η R καλεί *intercept* το σταθερό όρο και δίνει κατάλληλα ονόματα για τις υπόλοιπες εκτιμημένες τιμές των συντελεστών ανάλογα με τα ονόματα που έχουμε δώσει στις επεξηγηματικές μεταβλητές του μοντέλου.

residuals(object, type=" "): η συνάρτηση αυτή απιστρέφει το διάνυσμα των υπολοίπων του μοντέλου. Η παράμετρος *type* καθορίζει ποια υπόλοιπα θα επιστραφούν, π.χ. για τα υπόλοιπα Pearson η συνάρτηση θα γραφεί *residuals(object, type="pearson")*, ενώ για τα υπόλοιπα deviance θα γραφεί *residuals(object, type="deviance")*, το οποίο είναι ισοδύναμο με το *residuals(object)*.

fitted.values(object): οι παρατηρούμενες προσαρμοσμένες μέσες τιμές, που προκύπτουν μετασχηματίζοντας τη γραμμική προβλέπουσα με την αντίστροφη συνάρτηση σύνδεσης.

predict(object, type=" ",se.fit=False): η συνάρτηση αυτή επιστρέφει τις παρατηρούμενες προσαρμοσμένες μέσες τιμές υπολογισμένες στην κλίμακα της συνάρτησης σύνδεσης αν δεν ορίσουμε την παράμετρο *type* στην R. Οι εκφράσεις *predict(object, type="link")* και *predict(object)* είναι ισοδύναμες. Επιπλέον, δίνοντας στην παράμετρο type την τιμή response η συνάρτηση αυτή επιστρέφει τις παρατηρούμενες μέσες τιμές, μετασχηματίζοντας την γραμμική προβλέπουσα με την αντίστροφη συνάρτηση σύνδεσης. Η έκφραση λοιπόν *predict(object, type="response")* είναι ισοδύναμη με την *fitted.values(object)*. Στην περίπτωση, που δώσουμε την τιμή TRUE στην παράμετρο *se.fit*, θα επιστραφεί στην R και ένα διάνυσμα με τις εκτιμώμενες τιμές των τυπικών σφαλμάτων των προβλέψεων.

linear.predictors(object): οι τιμές της γραμμικής προβλέπουσας για κάθε παρατήρηση χρησιμοποιώντας τους εκτιμητές των συντελεστών.

AIC(object): η τιμή για το κριτήριο AIC.

BIC(object): η τιμή για το κριτήριο BIC.

deviance(object): είναι η ελεγχοσυνάρτηση deviance για το μοντέλο, που προσαρμόσαμε.

null.deviance(object): είναι η ελεγχοσυνάρτηση deviance για το μοντέλο, που περιλαμβάνει μόνο το σταθερό όρο.

Μια περιγραφή του μοντέλου μας μπορούμε να έχουμε με την συνάρτηση ***summary(object)***. Συγκεκριμένα, λαμβάνουμε κάποια περιγραφικά χαρακτηριστικά για τα υπόλοιπα *deviance*, τις εκτιμήσεις και τα τυπικά σφάλματα των συντελεστών του μοντέλου, $p-$τιμές για τους ελέγχους Wald των συντελεστών και τέλος τις τιμές των ελεγχοσυναρτήσεων *deviance* και *null deviance*.

Με την συνάρτηση ***plot(object)*** λαμβάνουμε γραφικές παραστάσεις και διαγράμματα Q-Q, καθώς και άλλα γραφήματα των υπολοίπων για τον εντοπισμό τυχόν σημείων επιρροής.

Τέλος, με τη συνάρτηση ***anova(object, test="Chi")*** λαμβάνουμε τον πίνακα ανάλυσης διασποράς του μοντέλου. Με τη βοήθεια της παραπάνω συνάρτησης, όπως θα δούμε και στο επόμενο κεφάλαιο, μπορούμε να συγκρίνουμε δύο εμφωλευμένα μοντέλα (έστω object1 και objetc2) ως προς την προσαρμογή τους. Η εντολή που πρέπει να χρησιμοποιήσουμε είναι η ***anova(object1, object2, test="Chi")***.

ΚΕΦΑΛΑΙΟ 6

ΕΦΑΡΜΟΓΕΣ ΜΕ ΧΡΗΣΗ ΤΟΥ ΣΤΑΤΙΣΤΙΚΟΥ ΠΑΚΕΤΟΥ R

6.1 Εφαρμογή στην Παλινδρόμηση Poisson

6.1.1 Περιγραφή των Δεδομένων

Στην εφαρμογή αυτή χρησιμοποιούμε στοιχεία από τις καταστροφές, που προκλήθηκαν σε συγκεκριμένους τύπους αεροσκαφών του Ναυτικού των Ηνωμένων Πολιτειών (*Montgomery, 2006*). Έχουμε συλλέξει δείγμα από 30 δοκιμαστικές αποστολές, οι οποίες πραγματοποιήθηκαν κατά τη διάρκεια του πολέμου του Βιετνάμ και χρησιμοποιήθηκαν δύο τύποι αεροσκαφών. Τα δεδομένα μας αποτελούνται από 4 μεταβλητές.

damage: είναι ο αριθμός των τμημάτων του αεροσκάφους, που σημειώθηκαν καταστροφές κατά τη διάρκεια των δοκιμαστικών αποστολών.

type: είναι μια δίτιμη μεταβλητή, που αναφέρεται στον τύπο του αεροσκάφους, που χρησιμοποιήθηκε σε κάθε αποστολή (0 για το A4, 1 για το A6).

bombload: είναι το φορτίο της βόμβας (σε τόνους).

airexp: η εμπειρία του πληρώματος (συνολικός αριθμός μηνών).

Σε αυτή την εφαρμογή μπορούμε να χρησιμοποιήσουμε την κατανομή Poisson για τον αριθμό καταστροφών (*damage*), που προκλήθηκαν στα αεροσκάφη μετά από κάθε αποστολή. Συνεπώς, θεωρώντας ότι η μεταβλητή απόκρισης είναι η μεταβλητή *damage* μπορούμε να προσαρμόσουμε ένα μοντέλο της παλινδρόμησης Poisson στα δεδομένα μας με επεξηγηματικές μεταβλητές τις *type*, *bombload*, *airexp*.

6.1.2 Εισαγωγή των Δεδομένων στην R

Έχουμε δημιουργήσει ένα αρχείο excel τύπου text (*Tab delimited; txt*), το οποίο περιλαμβάνει 4 στήλες με τα δεδομένα μας και το καλούμε aircraft_damage_data.txt. Για να εισάγουμε τα δεδομένα μας στην R χρησιμοποιώντας την συνάρτηση *read.table*().

```
>aircraft<-read.table("c:\\aircraft_damage_data.txt",header=TRUE)
```

Η παράμετρος header παίρνει μια λογική τιμή TRUE ή FALSE και αναφέρεται στο αν η πρώτη γραμμή του αρχείου περιλαμβάνει τα ονόματα των μεταβλητών. Πολλές φορές η πρώτη γραμμή στο αρχείο μας περιλαμβάνει τα ονόματα των μεταβλητών, για παράδειγμα και γι' αυτό θα πρέπει να ορίσουμε την παράμετρο αυτή ως TRUE.

Συνεπώς τα δεδομένα μας παίρνουν την εξής μορφή

```
> aircraft
   damage type bombload airexp
1       0    0       4    91.5
2       1    0       4    84.0
3       0    0       4    76.5
4       0    0       5    69.0
5       0    0       5    61.5
6       0    0       5    80.0
7       1    0       6    72.5
8       0    0       6    65.0
9       0    0       6    57.5
10      2    0       7    50.0
11      1    0       7   103.0
12      1    0       7    95.5
13      1    0       8    88.0
14      1    0       8    80.5
15      2    0       8    73.0
16      3    1       7   116.1
17      1    1       7   100.6
18      1    1       7    85.0
19      1    1      10    69.4
20      2    1      10    53.9
21      0    1      10   112.3
22      1    1      12    96.7
23      1    1      12    81.1
24      2    1      12    65.6
25      5    1       8    50.0
26      1    1       8   120.0
27      1    1       8   104.4
28      5    1      14    88.9
29      5    1      14    73.7
30      7    1      14    57.8
```

Ωστόσο, στα δεδομένα μας έχουμε μια κατηγορική μεταβλητή, η οποία αναφέρεται στον τύπο του αεροσκάφους (*type*). Για να ορίσουμε την μεταβλητή *type* ως κατηγορική χρησιμοποιούμε την συνάρτηση *as.factor*().

```
>attach(aircraft) #Με την εντολή αυτή μπορούμε να λαμβάνουμε τις μεταβλητές του πίνακα δεδομένων μας απλά γράφοντας το όνομα τους.

>type<-as.factor(type) #Η R ορίζει αυτόματα την πρώτη κατηγορία ως κατηγορία αναφοράς. Στη συγκεκριμένη περίπτωση η κατηγορία αναφοράς είναι η type = 0 .
```

6.1.3 Περιγραφή των Μεταβλητών

Αρχικά θα παρουσιάσουμε κάποια περιγραφικά χαρακτηριστικά για τις μεταβλητές του μοντέλου.

Χρησιμοποιώντας την εντολή summary() λαμβάνουμε κάποια αριθμητικά μέτρα της μεταβλητής *damage*.

```
>summary(damage)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  0.000   0.250   1.000   1.533   2.000   7.000
```

Παρατηρούμε ότι η μέγιστη και η ελάχιστη τιμή στο δείγμα μας για την μεταβλητή *damage* είναι 7 και 0, αντίστοιχα. Επιπλέον, η δειγματική διάμεσος είναι 1 και ο δειγματικός μέσος είναι 1.533. Τέλος, το 1° και το 3° τεταρτημόριο λαμβάνουν τις τιμές 0.250 και 2, αντίστοιχα.

Επιπλέον, η δειγματική διασπορά δίνεται με την παρακάτω εντολή

```
>var(damage)
[1] 3.154023
```

Και οι σχετικές συχνότητες για κάθε τιμή της μεταβλητής damage

Η μεταβλητή *type* είναι μια δίτιμη μεταβλητή. Συνεπώς έχει νόημα να δούμε τις συχνότητες αλλά και τις σχετικές συχνότητες για κάθε κατηγορία.
Στη R εκτελούμε τις παρακάτω εντολές

```
> table(type)
type
 0   1
15 15
> prop.table(table(type))
type
0   1
0.5 0.5
```

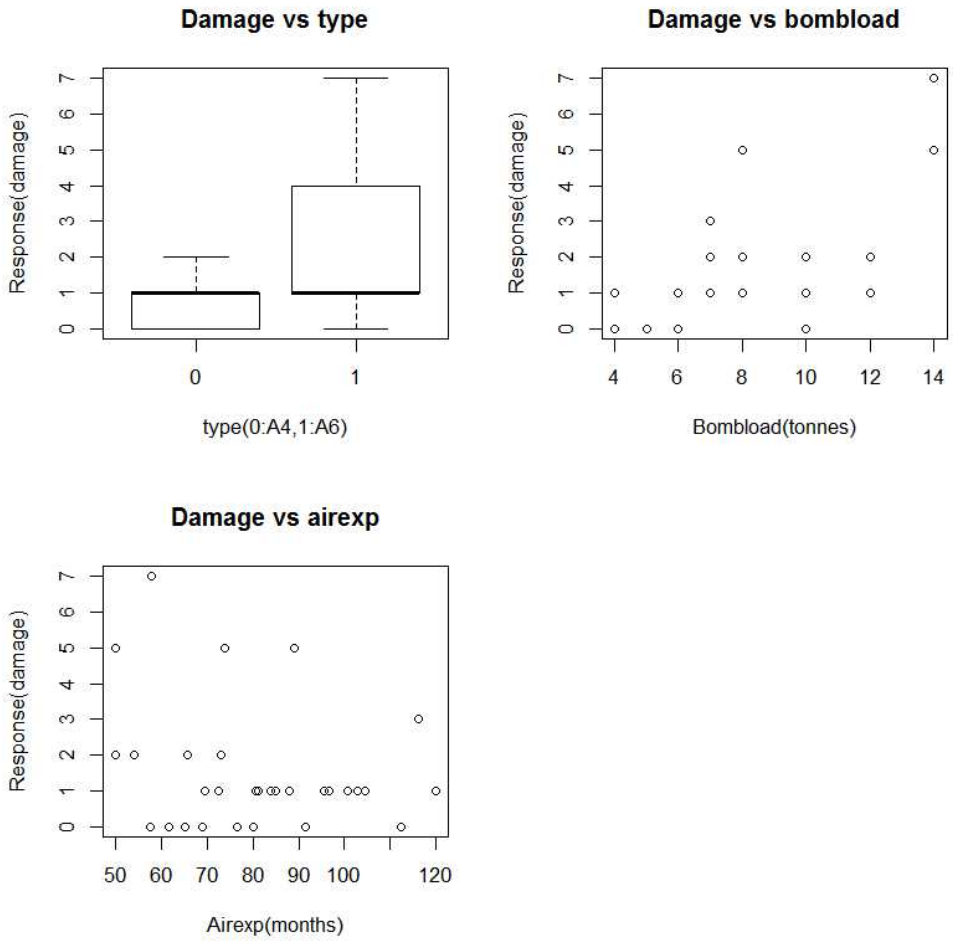
Παρατηρούμε λοιπόν ότι ο αριθμός των αεροσκαφών τύπου $A6 = 1$ είναι ίσος με τον αριθμό αεροσκαφών τύπου $A4 = 0$ στο δείγμα μας.

Για τις μεταβλητές *bombload* και *airexp* εκτελούμε τις παρακάτω εντολές

```
>summary(bombload)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

    4.0     6.0       7.5         8.1      10.0    14.0

>summary(airexp)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.

  50.00   66.45    80.25     80.77   94.50  120.00
```

Πριν την προσαρμογή του μοντέλου θα ελέγξουμε γραφικά τη σχέση κάθε επεξηγηματικής μεταβλητής με την μεταβλητή απόκρισης (*damage*).
Στην R εκτελούμε τις παρακάτω εντολές

```
>par(mfrow=c(2,2))
>plot(damage~type,ylab="Response(damage)",xlab="type(0:A4,1:A6)",main="Dama
ge vs type")
>plot(damage~bombload,ylab="Response(damage)",xlab="Bombload(tonnes)",main
="Damage vs bombload")
>plot(damage~airexp,ylab="Response(damage)",xlab="Airexp(months)",main="Dam
age vs airexp")
```



Γράφημα 6.1: *Περιγραφή της μεταβλητής απόκρισης σε σχέση με την κάθε επεξηγηματική μεταβλητή.*

Από το παραπάνω γράφημα παρατηρούμε ότι η μεταβλητή *damage* σχετίζεται με όλες τις επεξηγηματικές μεταβλητές. Για να δούμε πως διαφοροποιούνται οι τιμές της μεταβλητής *damage* ως προς τις δύο κατηγορίες της επεξηγηματικής μεταβλητής *type* δημιουργούμε δύο θηκογράμματα, ενώ για τις υπόλοιπες επεξηγηματικές μεταβλητές χρησιμοποιούμε διαγράμματα διασποράς.

Επίσης πριν την προσαρμογή του μοντέλου ελέγχουμε τη συσχέτιση μεταξύ όλων των μεταβλητών με αριθμητικούς και γραφικούς τρόπους. Σημειώνουμε εδώ ότι για τον υπολογισμό της συσχέτισης μεταξύ μίας δίτιμης και μίας συνεχής μεταβλητής χρησιμοποιούμε τη σημειακή δισειριακή συσχέτιση. Στην R ο συντελεστής αυτός δίνεται με την ίδια εντολή που χρησιμοποιείται για τον συντελεστή συσχέστισης του Pearson.

Οι συντελεστές συσχέτισης των μεταβλητών δίνονται στην R με την εντολή

```
> cor.aircraft<-cor(aircraft).
```

```
> cor.aircraft
            damage      type    bombload      airexp
damage    1.0000000 0.4963423  0.67608073 -0.23337694
type      0.4963423 1.0000000  0.71237621  0.22322455
bombload  0.6760807 0.7123762  1.00000000 -0.03241923
airexp   -0.2333769 0.2232245 -0.03241923  1.00000000
```

Αποτελέσματα 6.1

Τα αποτελέσματα συνοψίζονται στον παρακάτω πίνακα.

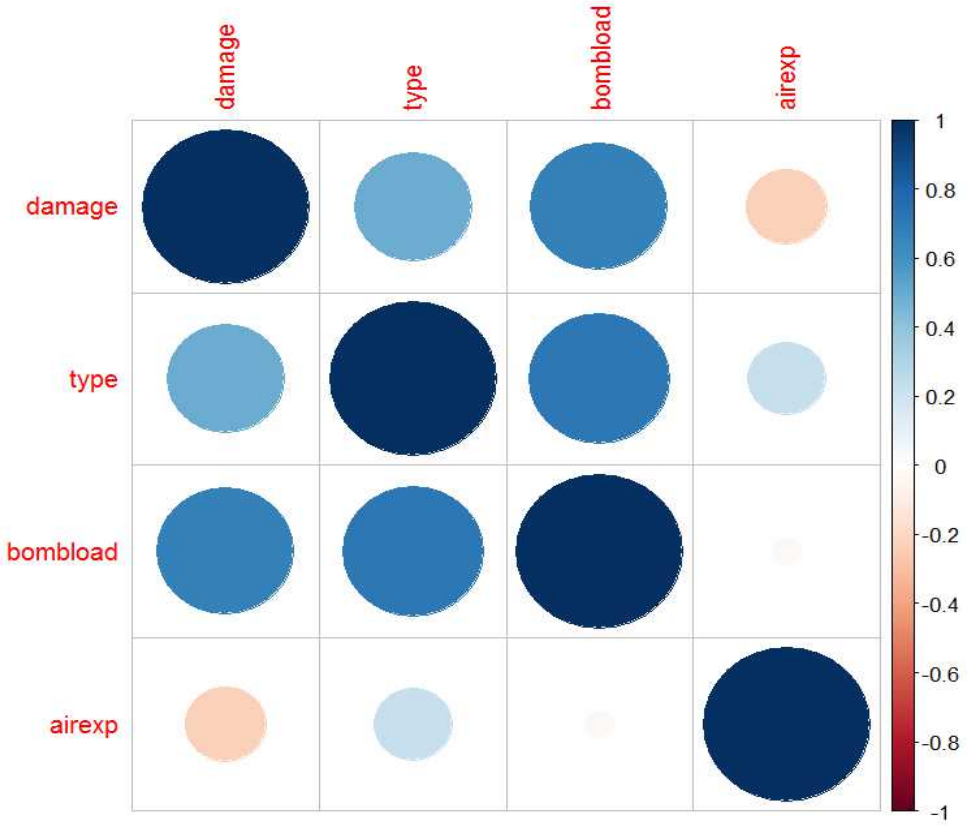
	damage	***type***	***bombload***	***airexp***
damage	1	0.496	0.676	-0.233
type	0.496	1	0.712	0.223
bombload	0.676	0.712	1	-0.032
airexp	-0.233	0.223	-0.032	1

Πίνακας 6.1: *Συσχέτιση μεταξύ όλων των μεταβλητών του μοντέλου.*

Επιπλέον, βλέπουμε γραφικά τη συσχέτιση μεταξύ των μεταβλητών με την παρακάτω σχηματική αναπαράσταση.

Αυτό επιτυγχάνεται στην R με τις παρακάτω εντολές

```
>library(graphics)
>library(corrplot)
>corrplot(cor.aircraft)
```



Γράφημα 6.2: *Γραφική αναπαράσταση της συσχέτισης μεταξύ των μεταβλητών.*

Από τον Πίνακα 5.1, αλλά και από το Γράφημα 5.2 παρατηρούμε το συντελεστή συσχέτισης, που υπάρχει μεταξύ των μεταβλητών. Συγκεκριμένα, όσο πιο μεγάλος και πιο σκούρος είναι ο κύκλος τόσο μεγαλύτερο συντελεστή συσχέτισης υποδηλώνει. Πιο συγκεκριμένα, οι μεταβλητές, που παρατηρείται να έχουν μεγάλο συντελεστή συσχέτισης είναι η *damage* με την *bombload* και η *type* με την *bombload*. Υπάρχει έντονη συσχέτιση μεταξύ των μεταβλητών *type* και *bombload* και το γεγονός αυτό μας οδηγεί στο συμπέρασμα της πολυσυγγραμικότητας για το μοντέλο μας.

6.1.4 Προσαρμογή του Μοντέλου

Η προσαρμογή του μοντέλου της παλινδρόμησης Poisson με συνάρτηση σύνδεσης *log* γίνεται με την εκτέλεση των ακόλουθων εντολών στην R

```
> air.model<-glm(damage~type+bombload+airexp,family="poisson",data=aircraft)
```

Με την συνάρτηση *summary*() λαμβάνουμε κάποια περιγραφικά χαρακτηριστικά για τα υπόλοιπα deviance και πιο συγκεκριμένα την ελάχιστη τιμή, τη μέγιστη τιμή, τη διάμεσο και το πρώτο και τρίτο τεταρτημόριο. Στη συνέχεια παρουσιάζονται οι εκτιμήσεις των παραμέτρων του μοντέλου, τα τυπικά σφάλματα τους, οι τιμές των ελέγχων Wald για τους συντελεστές του μοντέλου και οι αντίστοιχες $p-$τιμές των ελέγχων. Τέλος, παρουσιάζονται κάποιοι δείκτες καλής προσαρμογής για το μοντέλο μας. Πιο συγκεκριμένα, η τιμή της στατιστικής συνάρτησης deviance για το μοντέλο που προσαρμόζουμε τα δεδομένα μας (*Residual deviance*) και για το μοντέλο, που περιλαμβάνει μόνο το σταθερό όρο (*Null deviance*), καθώς και η τιμή του δείκτη AIC.

```
>summary(air.model)

Call:
glm(formula = damage ~ type + bombload + airexp, family = "poisson",
    data = aircraft1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6418  -1.0064  -0.0180   0.5581   1.9094

Coefficients:
              Estimate   Std. Error  z value  Pr(>|z|)
(Intercept) -0.406023   0.877489   -0.463   0.6436
type         0.568772   0.504372    1.128    0.2595
bombload     0.165425   0.067541    2.449    0.0143 *
airexp      -0.013522   0.008281   -1.633   0.1025
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 53.883  on 29  degrees of freedom
Residual deviance: 25.953  on 26  degrees of freedom
```

```
AIC: 87.649
Number of Fisher Scoring iterations: 5
```

Αποτελέσματα 6.2

Επιπλέον με την συνάρτηση *confint*() λαμβάνουμε 95% διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου.

```
> confint(air.model)
Waiting for profiling to be done…
                 2.5 %       97.5 %
(Intercept) -2.19824432 1.253906807
type        -0.42621497 1.567016666
bombload     0.03526245 0.301488598
airexp      -0.02998744 0.002670296
```

Τα αποτελέσματα για τα διαστήματα εμπιστοσύνης συνοψίζονται στον παρακάτω πίνακα.

	2.5%	**97.5%**	**Εκτιμήσεις** b_j	$exp(b_j)$
(Intercept)	-2.19824432	1.253906807	-0.406023	0.666295
type	-0.42621497	1.567016666	0.568772	1.766097
bombload	0.03526245	0.301488598	0.165425	1.179894
airexp	-0.02998744	0.002670296	-0.013522	0.986569

Πίνακας 6.2: *95% διαστήματα εμπιστοσύνης και εκτιμήτριες μέγιστης πιθανοφάνειας για τις παραμέτρους του μοντέλου.*

Από τον παραπάνω πίνακα των 95% διαστημάτων εμπιστοσύνης για τις παραμέτρους του μοντέλου παρατηρούμε ότι μόνο το διάστημα εμπιστοσύνης για τον συντελεστή της μεταβλητής *bombload* δεν περιλαμβάνει το μηδέν, γεγονός που σηματοδοτεί ότι η μεταβλητή αυτή είναι στατιστικά σημαντική για το μοντέλο, άρα και για την πρόβλεψη της μεταβλητής *damage*.

Στα ίδια συμπεράσματα καταλήγουμε, όπως αναμενόταν, και με βάση τα Αποτελέσματα 6.2. Η τιμή του ελέγχου Wald για τον συντελεστή β_2 είναι 2.449 και η $p-$τιμή (= 0.0143) αυτού είναι μικρότερη από 0.05, το οποίο σημαίνει ότι ο αριθμός των καταστροφών *damage* πράγματι σχετίζεται με το φορτίο της βόμβας, που χρησιμοποιήθηκε, υπό την προϋπόθεση ότι οι λοιπές επεξηγηματικές μεταβλητές παραμένουν σταθερές.

Το προσαρμοσμένο μοντέλο αποτελεί ένα καλό εναλλακτικό μοντέλο έναντι του κορεσμένου μοντέλου, όπως μπορούμε να διαπιστώσουμε από την $p-$τιμή του ελέγχου για την σημαντικότητά του ($p-\text{τιμή} = P(\mathrm{X}^2 > 25.953) = 0.4656818$). Αυτό αποτελεί μια ένδειξη καλής προσαρμογής του μοντέλου στα δεδομένα. Η τιμή αυτή δίνεται από την R με την εντολή

```
>1-pchisq(air.model$deviance,air.model$df.residual)
[1] 0.4656818
```

Συμπεραίνουμε, λοιπόν ότι παρ' όλο που οι επεξηγηματικές μεταβλητές δεν είναι όλες στατιστικά σημαντικές, σύμφωνα με τον έλεγχο Wald, η παρουσία όλων στο μοντέλο φαίνεται να βελτιώνει την προσαρμογή του. Αυτό φαίνεται και από την τιμή της ελεγχοσυνάρτησης *deviance* για το μοντέλο που περιλαμβάνει μόνο το σταθερό όρο, που είναι 53.883 συγκριτικά με την τιμή της ελεγχοσυνάρτησης *deviance* για το υποψήφιο μοντέλο, που είναι αισθητά μικρότερη και ίση με 25.953, όπως φαίνεται στα Αποτελέσματα 6.2, δηλαδή η πρόσθεση των τριών μεταβλητών μειώνει αισθητά την τιμή της *deviance*.

Η σύγκριση των δύο εμφωλευμένων μοντέλων μπορεί να γίνει με χρήση μιας στατιστικής συνάρτησης *deviance* για σύγκριση δύο μοντέλων. Πραγματοποιούμε δηλαδή έναν έλεγχο με μηδενική υπόθεση H_0: ισχύει το μοντέλο M_0 (όπου M_0 το μοντέλο που περιλαμβάνει μόνο το σταθερό όρο) με εναλλακτική H_1: ισχύει το μοντέλο M_1 (όπου M_1 το υποψήφιο μοντέλο) με τη βοήθεια μίας στατιστικής συνάρτησης $D0 - D1 \dot{\sim} \mathrm{X}^2_d$, όπου d η διαφορά των παραμέτρων των δύο μοντέλων, όπου στην περίπτωσή μας είναι 3. Στην R αυτό επιτυγχάνεται με τις παρακάτω εντολές

```
>ddev<-air.model$null.deviance-air.model$deviance

>df<-air.model$df.null-air.model$df.residual

>pvalue<-1-pchisq(ddev,df)
```

```
>data.frame(ddev,df,pvalue) # Δημιουργούμε ένα πλαίσιο δεδομένων, που περιλαμβάνει τις παραπάνω τιμές, που υπολογίσαμε στην R

Ddev      df    pvalue

1 27.9299   3   3.757192e-06
```

Παρατηρούμε ότι η $p-$τιμή του ελέγχου αυτού είναι 3.757192e-06, η οποία είναι πολύ μικρή, γεγονός που μας οδηγεί στο συμπέρασμα να απορρίψουμε τη μηδενική υπόθεση, δηλαδή το υποψήφιο μοντέλο είναι προτιμότερο από το μοντέλο, που περιλαμβάνει μόνο το σταθερό όρο.

6.1.5 Υπόλοιπα και Διαγνωστικοί Έλεγχοι

Η θεωρία, που έχουμε αναφέρει για το μοντέλο παλινδρόμησης Poisson προϋποθέτει ότι η διασπορά και η μέση τιμή της τυχαίας μεταβλητής Y είναι ίσες. Γι' αυτό πριν ελέγξουμε την καλή προσαρμογή του μοντέλου στα δεδομένα με κατάλληλους διαγνωστικούς ελέγχους θα πρέπει να ελέγξουμε αν υπάρχει το φαινόμενο της υπερμεταβλητότητας ή υπομεταβλητότητας στα δεδομένα. Αυτό θα το εξετάσουμε με δύο τρόπους. Αρχικά, υπολογίζοντας την τιμή της ελεγχοσυνάρτησης $\varphi = deviance / df$, όπου $deviance$ είναι η τιμή της ελεγχοσυνάρτησης $deviance$ για το προσαρμοσμένο μοντέλο και df οι βαθμοί ελευθερίας της X^2 κατανομής, που ακολουθεί η ελεγχοσυνάρτηση $deviance$. Όπως έχουμε αναφέρει και στη θεωρία αν η τιμή φ είναι μεγαλύτερη του 1 τότε έχουμε ενδείξεις για υπερμεταβλητότητα, ενώ αν είναι μικρότερη του 1 έχουμε ενδείξεις για υπομεταβλητότητα στα δεδομένα.

Η ελεγχοσυνάρτηση φ υπολογίζεται στην R με την παρακάτω εντολή

```
> deviance(air.model)/air.model$df.residual

[1] 0.9981985
```

Επιπλέον, ο δεύτερος τρόπος εξέτασης του φαινομένου της υπο/υπερδιασποράς στα δεδομένα επιτυγχάνεται χρησιμοποιώντας έναν έλεγχο υποθέσεων (*Cameron και Trivedi*, *1990*). Θα ελέγξουμε την υπόθεση ότι $E(Y) = V(Y) = \mu$ ως την μηδενική υπόθεση με εναλλακτική ότι η διασπορά της τυχαίας μεταβλητής Y δίνεται από τη σχέση $V(Y) = \mu + c \cdot f(\mu)$, όπου $f(\cdot)$ μία μονότονη συνάρτηση. Ο έλεγχος αυτός ισοδυναμεί με το να ελέγξουμε την υπόθεση $H_0 : c = 0$ με εναλλακτική $H_1 : c \neq 0$.

Ο έλεγχος αυτός επιτυγχάνεται στην R με την συνάρτηση *dispersiontest()*.

```
> dispersiontest(air.model)
Overdispersion test
data:  air.model
z = -1.1134, p-value = 0.8672
```

```
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
 0.7883916
```

Παρατηρούμε λοιπόν ότι η $p-$τιμή του ελέγχου αυτού είναι 0.8672 οπότε δεν έχουμε ενδείξεις για να απορρίψουμε την μηδενική υπόθεση πράγμα που σημαίνει ότι δεν υπάρχει μεγάλη διασπορά στα δεδομένα.

Τέλος, θα πρέπει να εξετάσουμε τις υποθέσεις του μοντέλου και γενικά την προσαρμογή του μοντέλου με τη βοήθεια των υπολοίπων και κατάλληλων διαγνωστικών ελέγχων, έτσι ώστε να διαπιστώσουμε αν είναι κατάλληλο το μοντέλο για την περιγραφή των δεδομένων μας.

Θα χρησιμοποιήσουμε τα υπόλοιπα *deviance* και τα υπόλοιπα Pearson, τα οποία υπολογίζονται στην R με τις ακόλουθες εντολές.

```
> res_pearson<-residuals(air.model,type="pearson")
> res_deviance<-residuals(air.model)
```

Αν και είναι προφανές ότι τα υπόλοιπα deviance και τα υπόλοιπα Pearson δεν κατανέμονται σύμφωνα με την Κανονική κατανομή, συνηθίζεται να γίνεται ο γραφικός έλεγχος καταλληλότητας της Κανονικής κατανομής μέσω ενός διαγράμματος Q-Q για τον εντοπισμό απόμακρων ή άτυπων παρατηρήσεων.

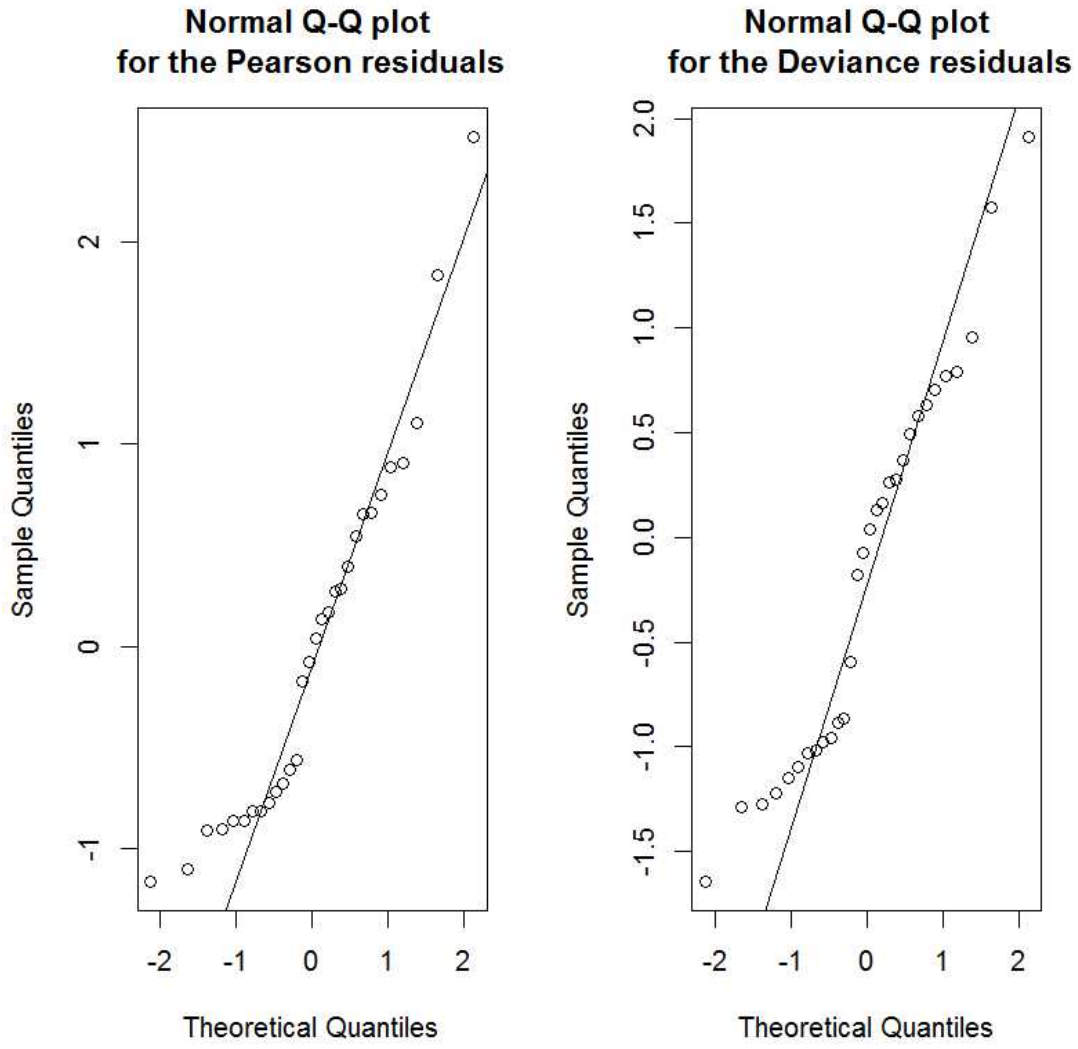
Το γράφημα αυτό για τα υπόλοιπα Pearson λαμβάνεται από την R με τις εντολές

```
>qqnorm(res_pearson)
>qqline(res_pearson)
```

Αντίστοιχα, το γράφημα για τα υπόλοιπα *deviance* λαμβάνεται από την R με τις εντολές

```
>qqnorm(res_deviance)
>qqline(res_deviance)
```

Τα δύο γραφήματα απεικονίζονται παρακάτω.



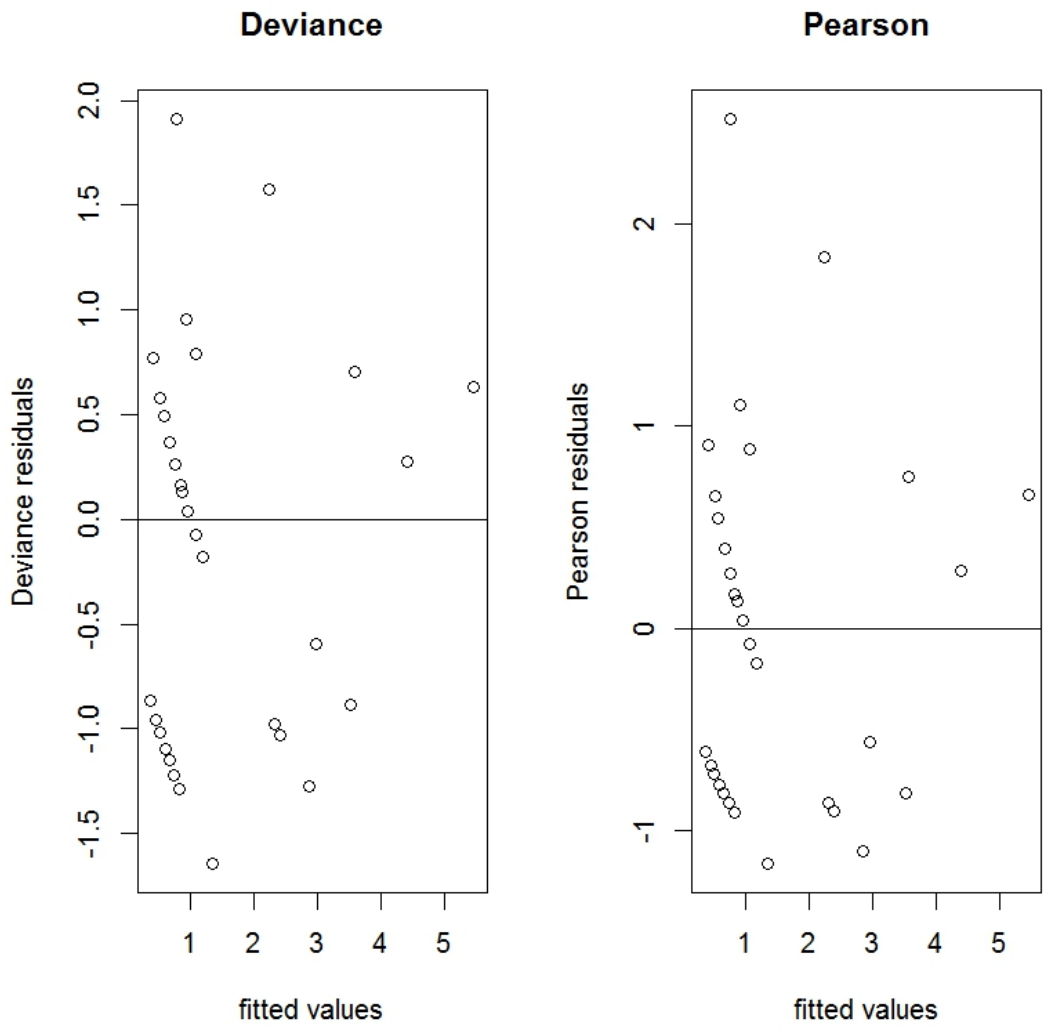
Γράφημα 6.3: Διαγράμματα Q-Q *των υπολοίπων Pearson και deviance.*

Παρατηρούμε και στα δύο γραφήματα ότι τα σημεία σχηματίζουν μια σχετικά καλά ορισμένη ευθεία, γεγονός που υποδηλώνει τη μη ύπαρξη άτυπων σημείων και ενδεχομένως την καλή προσαρμογή του μοντέλου, με εξαίρεση 2 παρατηρήσεις κάτω αριστερά και στα δύο γραφήματα.

Επιπλέον, εξετάζουμε τα γραφήματα των υπολοίπων *deviance* και Pearson σε σχέση με τις προσαρμοσμένες τιμές $\hat{y}_i = \hat{\mu}_i = \exp(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{b})$, τα οποία λαμβάνουμε από την R με τις ακόλουθες εντολές.

```
par(mfrow=c(1,2))
plot(fitted.values(air.model),res_deviance,xlab="fitted values",ylab="Deviance
residuals",main="Deviance")
abline(h=0)
```

```
plot(fitted.values(air.model),res_pearson,xlab="fitted values",ylab="Pearson
residuals",main="Pearson")
abline(h=0)
```



Γράφημα 6.4: *Υπόλοιπα deviance και Pearson σε σχέση με τις προσαρμοσμένες τιμές* $\hat{\mu}_i$.

Από το παραπάνω γράφημα παρατηρούμε ότι τα υπόλοιπα συμπεριφέρονται "αρκετά" τυχαία.

6.1.6 Επιλογή Μεταβλητών

Για να επιλέξουμε το βέλτιστο μοντέλο, δηλαδή το μοντέλο, που περιλαμβάνει τον κατάλληλο αριθμό επεξηγηματικών μεταβλητών, ώστε να αποτελεί την καλύτερη προσαρμογή στα δεδομένα μας, θα προσαρμόσουμε όλα τα μοντέλα, που θα μπορούσαμε να έχουμε με συνδυασμό των τριών επεξηγηματικών μεταβλητών. Αυτά

είναι επτά στον αριθμό (συμπεριλαμβανομένου του μοντέλου χωρίς καμία επεξηγηματική μεταβλητή). Η επιλογή του βέλτιστου μοντέλου θα γίνει με βάση την ελεγχοσυνάρτηση *deviance*, καθώς και με τις τιμές των κριτηρίων AIC και BIC. Προσαρμόζουμε όλα τα πιθανά μοντέλα στην R με τις παρακάτω εντολές

```
>x1.model<-glm(damage~type,family="poisson",data=aircraft)
>x2.model<-glm(damage~bombload,family="poisson",data=aircraft)
>x3.model<-glm(damage~airexp,family="poisson",data=aircraft)
>x1x2.model<-glm(damage~type+bombload,family="poisson",data=aircraft)
>x1x3.model<-glm(damage~type+airexp,family="poisson",data=aircraft)
>x2x3.model<-glm(damage~bombload+airexp,family="poisson",data=aircraft)
>x1x2x3.model<-
glm(damage~type+bombload+airexp,family="poisson",data=aircraft)
```

Στη συνέχεια, κάνουμε συγκρίσεις των παραπάνω μοντέλων με το μοντέλο, που περιλαμβάνει και τις τρεις επεξηγηματικές μεταβλητές με τη βοήθεια της ελεγχοσυνάρτησης *deviance* χρησιμοποιώντας την συνάρτηση *anova*() της R. Εκτελούμε τις παρακάτω εντολές

```
> anova(x1.model,x1x2x3.model,test="Chisq")
> anova(x2.model,x1x2x3.model,test="Chisq")
> anova(x3.model,x1x2x3.model,test="Chisq")
> anova(x1x2.model,x1x2x3.model,test="Chisq")
> anova(x1x3.model,x1x2x3.model,test="Chisq")
> anova(x2x3.model,x1x2x3.model,test="Chisq")
```

Λαμβάνουμε τα παρακάτω αποτελέσματα εκτελώντας τις παραπάνω εντολές

```
> anova(x1.model,x1x2x3.model,test="Chisq")

Analysis of Deviance Table

Model 1: damage ~ type

Model 2: damage ~ type + bombload + airexp

  Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1        28     38.283

2        26     25.953  2    12.33 0.002101 **

---

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

Αποτελέσματα 6.3

```
>  anova(x2.model,x1x2x3.model,test="Chisq")
Analysis of Deviance Table

Model 1: damage ~ bombload

Model 2: damage ~ type + bombload + airexp

  Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1        28     29.206

2        26     25.953  2   3.2527   0.1966
```

Αποτελέσματα 6.4

```
>anova(x3.model,x1x2x3.model,test="Chisq")
Analysis of Deviance Table

Model 1: damage ~ airexp

Model 2: damage ~ type + bombload + airexp

  Resid. Df Resid. Dev  Df   Deviance  Pr(>Chi)

1        28     50.537

2        26     25.953       2       24.584   4.588e-06 ***

---

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

Αποτελέσματα 6.5

```
> anova(x1x2.model,x1x2x3.model,test="Chisq")
Analysis of Deviance Table

Model 1: damage ~ type + bombload

Model 2: damage ~ type + bombload + airexp

  Resid. Df Resid. Dev Df  Deviance Pr(>Chi)

1        27     28.634

2        26     25.953    1         2.6812   0.1015
```

Αποτελέσματα 6.6

```
> anova(x1x3.model,x1x2x3.model,test="Chisq")
Analysis of Deviance Table

Model 1: damage ~ type + airexp

Model 2: damage ~ type + bombload + airexp

  Resid. Df   Resid. Dev Df Deviance Pr(>Chi)

1        27     32.192

2        26     25.953       1       6.2386   0.0125 *

---

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

Αποτελέσματα 6.7

```
>anova(x2x3.model,x1x2x3.model,test="Chisq")
Analysis of Deviance Table

Model 1: damage ~ bombload + airexp

Model 2: damage ~ type + bombload + airexp

  Resid.Df  Resid. Dev   Df    Deviance  Pr(>Chi)

1        27     27.220

2        26     25.953          1        1.2667    0.2604
```

Αποτελέσματα 6.8

```
>AIC(x1x2x3.model,x1x2.model,x1x3.model,x2x3.model,x1.model,x2.model,x3.mo
del)
             df       AIC

x1x2x3.model  4  87.64922

x1x2.model      3  88.33037

x1x3.model      3  91.88781

x2x3.model      3  86.91589

x1.model         2  95.97952

x2.model         2  86.90196
```

```
x3.model      2 108.23330
>BIC(x1x2x3.model,x1x2.model,x1x3.model,x2x3.model,x1.model,x2.model,x3.model)
              df      BIC
x1x2x3.model  4  93.25401
x1x2.model    3  92.53396
x1x3.model    3  96.09140
x2x3.model    3  91.11948
x1.model      2  98.78192
x2.model      2  89.70435
x3.model      2 111.03570
```

Συγκεντρωτικά τα αποτελέσματα των παραπάνω πινάκων ανάλυσης της *deviance* για κάθε ένα από τα εναλλακτικά μοντέλα σε σχέση με το μοντέλο, που περιλαμβάνει και τις τρείς επεξηγηματικές μεταβλητές παρουσιάζονται στον παρακάτω πίνακα συμπεραλβανομένων και των τιμών των κριτηρίων AIC και BIC.

Μοντέλο (Model)	**Deviance**	**Deviance(Full)-Deviance(Model)**	**p-value**	**AIC**	**BIC**
x1x2x3.model (Full)	25.953			87.64922	93.25401
x1x2.model	28.634	2.6812	0.1015	88.33037	92.53396
x1x3.model	32.192	6.2386	0.0125	91.88781	96.09140
x2x3.model	27.220	1.2667	0.2604	86.91589	91.11948
x1.model	38.283	12.33	0.002101	95.97952	98.78192
x2.model	29.206	3.2527	0.1966	86.90196	89.70435
x3.model	50.537	24.584	< 0.001	108.23330	111.03570

Πίνακας 6.3: *Αποτελέσματα του πίνακα ανάλυσης της Deviance, καθώς και οι τιμές των κριτηρίων AIC και BIC για όλα τα μοντέλα.*

Παρατηρούμε, λοιπόν από τα αποτελέσματα του Πίνακα 6.3 ότι το μοντέλο, που περιλαμβάνει και τις τρείς επεξηγηματικές μεταβλητές είναι προτιμότερο από όλα τα άλλα πιθανά μοντέλα, που μπορούμε να φτιάξουμε με κάθε δυνατό συνδυασμό των επεξηγηματικών μεταβλητών σύμφωνα με την τιμή της ελεγχοσυνάρτησης *deviance*, καθώς και με τους ελέγχους AIC και BIC, παρ' όλο που το μοντέλο μας περιλαμβάνει

μια μη στατιστικά σημαντική μεταβλητή, την *airexp* και έναν "σχετικά" σημαντικό παράγοντα, την μεταβλητή *type*. Αυτό μπορεί να οφείλεται και στο φαινόμενο της πολυσυγγραμικότητας (*multicolinearity*), που παρουσιάζεται στις επεξηγηματικές μεταβλητές.

6.2 Σύγκριση του Μοντέλου Poisson με το Μοντέλο της Αρνητικής Διωνυμικής Κατανομής

Σκοπός της παρακάτω ανάλυσης είναι να προσαρμόσουμε στα δεδομένα μας όπως παρουσιάζονται στην Παράγραφο 6.1 το μοντέλο της αρνητικής διωνυμικής κατανομής και να κάνουμε μια σύγκριση με το μοντέλο Poisson. Θα προσαρμόσουμε, λοιπόν το αρνητικό διωνυμικό μοντέλο και θα συγκρίνουμε τα δύο μοντέλα με βάση τα κριτήρια AIC, BIC και την στατιστική συνάρτηση *deviance*.

6.2.1 Προσαρμογή του Μοντέλου

Για την προσαρμογή του αρνητικού διωνυμικού μοντέλου χρησιμοποιούμε τη συνάρτηση glm.nb(), η οποία βρίσκεται στη βιβλιοθήκη MASS της R.
Εκτελώντας την παρακάτω εντολή στην R προσαρμόζουμε το μοντέλο στα δεδομένα μας.

```
>air1.model<-glm.nb(damage~type+bombload+airexp,data=aircraft,link=log)
```

Και με τη συνάρτηση *summary()*, όπως και προηγουμένως λαμβάνουμε κάποια στοιχεία για το μοντέλο μας.

```
>summary(air1.model)

Call:

glm.nb(formula = damage ~ type + bombload + airexp, data = aircraft,

    link = log, init.theta = 20198.97302)

Deviance Residuals:

     Min       1Q   Median       3Q      Max

-1.64180  -1.00637  -0.01803   0.55807   1.90929

Coefficients:

                  Estimate Std. Error z value Pr(>|z|)

(Intercept)  -0.406038   0.877536  -0.463   0.6436
```

```
type          0.568792   0.504391   1.128   0.2595
bombload      0.165422   0.067545   2.449   0.0143 *
airexp       -0.013522   0.008281  -1.633   0.1025
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for Negative Binomial(20198.97) family taken to be 1)
    Null deviance: 53.879  on 29  degrees of freedom
Residual deviance: 25.951  on 26  degrees of freedom
AIC: 89.65
Number of Fisher Scoring iterations: 1
              Theta:  20199
          Std. Err.:  474216
Warning while fitting theta: iteration limit reached
 2 x log-likelihood:  -79.65
```

Αποτελέσματα 6.9

Από τα Αποτελέσματα 6.9 προκύπτει ότι οι εκτιμήσεις των συντελεστών για το μοντέλο της αρνητικής διωνυμικής κατανομής είναι παρόμοιες με αυτές για το Poisson μοντέλο. Τα δύο μοντέλα είναι εμφωλευμένα με το Poisson να είναι ειδική περίπτωση του μοντέλου της αρνητικής διωνυμικής κατανομής. Έτσι, μπορούμε να χρησιμοποιήσουμε την ελεγχοσυνάρτηση του λόγου των πιθανοφανειών (*loglikelihood ratio test*) για να συγκρίνουμε τα δύο μοντέλα, η οποία όπως έχουμε αναφέρει και στη θεωρία δίνεται από τη σχέση $LR = 2(l_1 - l_0)$ και ακολουθεί ασυμπτωτικά την X^2 κατανομή, δηλαδή $LR \dot{\sim} \mathrm{X}^2_d$. Οι μεγιστοποιημένες συναρτήσεις πιθανοφάνειας (σε λογαριθμική κλίμακα) l_0 και l_1 υπολογίζονται υπό την μηδενική υπόθεση H_0: ισχύει το μοντέλο Poisson με την εναλλακτική H_1: ισχύει το μοντέλο της αρνητικής διωνυμικής κατανομής.

Συνεπώς, μπορούμε να λάβουμε την $p-$τιμή αυτού του ελέγχου στην R με την παρακάτω εντολή.

```
>pchisq(2*(logLik(air1.model)-logLik(air.model)),df=1,lower.tail=FALSE)
'log Lik.' 1 (df=5)
```

Συνεπώς, η $p-$τιμή του ελέγχου αυτού υποδεικνύει ότι το Poisson μοντέλο είναι προτιμότερο από το μοντέλο της αρνητικής διωνυμικής κατανομής για την περιγραφή των συγκεκριμένων δεδομένων, καθώς δεν έχουμε ενδείξεις ώστε να απορρίψουμε τη μηδενική υπόθεση.
Θα πρέπει να είμαστε προσεκτικοί όσον αφορά το μοντέλο της αρνητικής διωνυμικής κατανομής, καθώς η προσαρμογή δεν πληρεί τα προεπιλεγμένα όρια σύγκλισης.
Ωστόσο, αφού η προσαρμογή των δύο μοντέλων είναι παραπλήσια μπορούμε να συγκρίνουμε την τιμή του δείκτη AIC για τα δύο μοντέλα. Το AIC για το Poisson μοντέλο είναι μικρότερο (87.649) συγκριτικά με αυτό του μοντέλου της αρνητικής διωνυμικής κατανομής (89.65).
Επιπλέον, μπορούμε να υπολογίσουμε και τον δείκτη BIC για τα δύο μοντέλα με την συνάρτηση BIC().

Συνεπώς, για το Poisson μοντέλο θα έχουμε

```
> BIC(air.model)
[1] 93.25401
```

ενώ για το μοντέλο της αρνητικής διωνυμικής κατανομής θα έχουμε το εξής αποτέλεσμα

```
> BIC(air1.model)
[1] 96.65575
```

Παρατηρούμε από τα παραπάνω αποτελέσματα ότι και η σύγκριση του δείκτη BIC μας οδηγεί στο συμπέρασμα ότι το μοντέλο Poisson είναι προτιμότερο.
Επιπλέον, με τη βοήθεια της R εκτελώντας τις παρακάτω εντολές λαμβάνουμε τις εκτιμήσεις και 95% διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου για τα δύο μοντέλα.

- Για το μοντέλο Poisson

```
> est<-cbind(estimate=coef(air.model),confint(air.model))
Waiting for profiling to be done...
> est
              estimate       2.5 %      97.5 %
(Intercept) -0.40602269 -2.19824432 1.253906807
```

```
type        0.56877242 -0.42621497 1.567016666
bombload    0.16542540  0.03526245 0.301488598
airexp     -0.01352232 -0.02998744 0.002670296
```

Επιπλέον, μπορούμε να υπολογίσουμε τα ίδια ακριβώς πράγματα για τον όρο $\exp(\beta_j)$ εκτελώντας την παρακάτω εντολή στην R.

```
> exp(est)
            estimate     2.5 %   97.5 %
(Intercept) 0.6662950 0.1109979 3.504006
type        1.7660977 0.6529760 4.792330
bombload    1.1798949 1.0358915 1.351870
airexp      0.9865687 0.9704577 1.002674
```

- Για το μοντέλο της αρνητικής διωνυμικής κατανομής

Εκτελώντας τις παρακάτω εντολές στην R λαμβάνουμε τα ίδια ακριβώς αποτελέσματα για το μοντέλο της αρνητικής διωνυμικής κατανομής με συνάρτηση σύνδεσης *log*.

```
> est1<-cbind(Estimate=coef(air1.model),confint(air1.model))
Waiting for profiling to be done...
> est1
             Estimate       2.5 %      97.5 %
(Intercept) -0.40603830 -2.19830587 1.253923323
type         0.56879150 -0.42621848 1.567059862
bombload     0.16542155  0.03525339 0.301488878
airexp      -0.01352182 -0.02998761 0.002671456
```

```
> exp(est1)
            Estimate     2.5 %   97.5 %
```

(Intercept)	0.6662847	0.1109910	3.504064
type	1.7661314	0.6529737	4.792537
bombload	1.1798904	1.0358822	1.351870
airexp	0.9865692	0.9704576	1.002675

Παρακάτω παρουσιάζουμε συγκεντρωτικά κάποια αποτελέσματα και για τα δύο μοντέλα, που προσαρμόστηκαν στα δεδομένα.

	Poisson model	**Negative Binomial model**
	Εκτιμήσεις (95% διαστήματα εμπιστοσύνης)	**Εκτιμήσεις (95% διαστήματα εμπιστοσύνης)**
β_0	-0.406 (0.111, 3.504)	-0.406 (-2.198, 1.254)
β_1	0.569 (-0.426, 1.567)	0.569 (-0.426, 1.567)
β_2	0.165 (0.035, 0.301)	0.165 (0.035, 0.301)
β_3	-0.014 (-0.03, 0.003)	-0.014 (-0.03, 0.003)
$exp(\beta_0)$	0.666 (0.111, 3.504)	0.666 (0.111, 3.504)
$exp(\beta_1)$	1.766 (0.653, 4.792)	1.766 (0.653, 4.793)
$exp(\beta_2)$	1.18 (1.036, 1.352)	1.18 (1.036, 1.352)
$exp(\beta_3)$	0.987 (0.97, 1.003)	0.987 (0.971, 1.003)
AIC	87.649	89.65
BIC	93.25401	96.65575

Πίνακας 6.4: *Εκτιμήσεις (95% διαστήματα εμπιστοσύνης) για τις παραμέτρους και των δύο μοντέλων, καθώς και για τα* $exp(\beta_j)$.

6.2.2 Συμπεράσματα

Το Poisson μοντέλο προτιμάται από το μοντέλο της αρνητικής διωνυμικής κατανομής, καθώς όπως έχουμε ήδη ελέγξει δεν υπάρχει υπερμεταβλητότητα στα δεδομένα.

6.3 Σύγκριση του Μοντέλου Poisson με το Μοντέλο Poisson Πληθωρισμού στο Μηδέν

Το μοντέλο Poisson πληθωρισμού στο μηδέν (*zero-inflated*) χρησιμοποιείται κυρίως όταν σε απαριθμητά δεδομένα παρουσιάζεται ένας σχετικά μεγάλος αριθμός μηδενικών συγκριτικά με το δείγμα.

Για ένα διάνυσμα μεταβλητών απόκρισης $Y = (Y_1, ..., Y_n)^T$ το εν λόγω μοντέλο ορίζεται ως εξής:

$$Y_i \sim \begin{cases} 0, \text{ με πιθανότητα } p_i \\ Poisson(\lambda_i), \text{ με πιθανότητα } 1-p_i \end{cases},$$

όπου οι παράμετροι $\boldsymbol{p} = (p_1,...,p_n)^{\mathrm{T}}$ και $\boldsymbol{\lambda} = (\lambda_1,...,\lambda_n)^{\mathrm{T}}$ μοντελοποιούνται σύμφωνα με τη θεωρία των γενικευμένων γραμμικών μοντέλων χρησιμοποιώντας τις συναρτήσεις σύνδεσης *logit* και *log*, αντίστοιχα. Ειδικότερα, δημιουργούνται δύο μοντέλα $\log(\boldsymbol{\lambda}) = \mathbf{B}\boldsymbol{\beta}$ και $\mathrm{logit}(\boldsymbol{p}) = \mathbf{G}\boldsymbol{\gamma}$, όπου $\mathbf{B}$ και $\mathbf{G}$ οι αντίστοιχοι πίνακες σχεδιασμού. Για ευκολία παρακάτω θεωρούμε ότι ο πίνακας σχεδιασμού για το logit μοντέλο περιέχει μόνο μία στήλη (στο μοντέλο έχουμε μόνο τη σταθερά), δηλαδή θεωρούμε στο μοντέλο ότι όλα τα μηδενικά έχουν την ίδια πιθανότητα να περιέχονται στη "μηδενική συνιστώσα" του μοντέλου.

Παρατηρούμε ότι το 27% των παρατηρήσεων της μεταβλητής απόκρισης damage λαμβάνουν την τιμή 0. Συνεπώς, με βάση τα παραπάνω θα συγκρίνουμε το μοντέλο Poisson με το μοντέλο Poisson πληθωρισμού στο μηδέν.
Προσαρμόζουμε το μοντέλο Poisson πληθωρισμού στο μηδέν με την συνάρτηση *zeroinfl*().

```
> mod1<-zeroinfl(damage~type+bombload+airexp|1,data=aircraft)
```

Μπορούμε να το συγκρίνουμε με το μοντέλο Poisson, που έχουμε προσαρμόσει και το έχουμε ονομάσει *air.model*.
Για την σύγκριση των δύο (μη εμφωλευμένων) μοντέλων θα χρησιμοποιήσουμε τον έλεγχο Vuong (1989). Δύο μοντέλα έχουν "παρόμοια" προσαρμογή στα δεδομένα, αν η διαφορά μεταξύ των συναρτήσεων πιθανοφάνειάς τους (σε λογαριθμική κλίμακα) είναι ίση με μηδέν. Συνεπώς, η μηδενική υπόθεση του ελέγχου θα είναι

$$H_0 : E_0\left[\log\frac{f(Y_i \mid X_i;\boldsymbol{\beta})}{g(Y_i \mid Z_i;\boldsymbol{\gamma})}\right] = 0 \text{ , (με εναλλακτική ότι δεν ισχύει η ισότητα),}$$

όπου $f(Y_i \mid X_i;\boldsymbol{\beta})$, $g(Y_i \mid Z_i;\boldsymbol{\gamma})$ οι δεσμευμένες συναρτήσεις πιθανότητας για τις μεταβλητές $Y_i \mid X_i$ και $Y_i \mid Z_i$, αντίστοιχα, X_i και Z_i είναι οι πίνακες σχεδιασμού των δύο μοντέλων και $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ είναι οι παράμετροι των δύο μοντέλων. Για τον παραπάνω έλεγχο χρησιμοποιούμε την ελεγχοσυνάρτηση

$$\frac{1}{n} LR(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}),$$

όπου $LR(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})$ είναι ο λόγος των δύο πιθανοφανειών και $\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}$ οι εκτιμήτριες μέγιστης πιθανοφάνειας για τις παραμέτρους των δύο μοντέλων. Αποδεικνύεται ότι κάτω από την μηδενική υπόθεση,

$$\frac{LR(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}})}{(\sqrt{n})\hat{\omega}_n} \dot{\sim} N(0,1),$$

όπου $\hat{\omega}_n = \frac{1}{n}\sum_{i=1}^{n}\left[\log\frac{\mathrm{f}(\mathrm{Y_i}\,|\,\mathrm{X_i};\hat{\beta}_\mathrm{n})}{\mathrm{g}(\mathrm{Y_i}\,|\,\mathrm{Z_i};\hat{\gamma}_\mathrm{n})}\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\log\frac{\mathrm{f}(\mathrm{Y_i}\,|\,\mathrm{X_i};\hat{\beta}_\mathrm{n})}{\mathrm{g}(\mathrm{Y_i}\,|\,\mathrm{Z_i};\hat{\gamma}_\mathrm{n})}\right]^2.$

Ο έλεγχος αυτός γίνεται στην R χρησιμοποιώντας την συνάρτηση *vuong()*.

```
> vuong(air.model,mod1)
Vuong Non-Nested Hypothesis Test-Statistic: 0.03511112
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
in this case:
model1 > model2, with p-value 0.4859956
```

Από τα παραπάνω καταλήγουμε ότι το μοντέλο παλινδρόμησης Poisson δεν διαφέρει σημαντικά από το μοντέλο Poisson πληθωρισμού στο μηδέν.

6.4 Εκτίμηση Συντελεστών στο Μοντέλο Poisson

Από τις προηγούμενες συγκρίσεις και ελέγχους καλής προσαρμογής καταλήγουμε ότι το μοντέλο Poisson είναι προτιμότερο για την περιγραφή των δεδομένων. Συνεπώς, θα παρουσιάσουμε την ερμηνεία των συντελεστών του μοντέλου σύμφωνα με τα Αποτελέσματα 6.2.

Το μοντέλο Poisson για τα δεδομένα μας έχει την ακόλουθη δομή

$$damage_i \sim Poisson(\mu_i)$$
$$\log(\mu_i) = \beta_0 + \beta_1 type_i + \beta_2 bombload_i + \beta_3 airexp_i$$
$$i = 1,2,...,30.$$

Το παρατηρούμενο προσαρμοσμένο μοντέλο λαμβάνει τη μορφή

$$\log(\hat{\mu}_i) = b_0 + b_1 type_i + b_2 bombload_i + b_3 airexp_i$$
$$\log(\hat{\mu}_i) = -0.406 + 0.569 type_i + 0.165 bombload_i - 0.014 airexp_i$$

ή

$$\hat{\mu}_i = \exp(-0.406 + 0.569 type_i + 0.165 bombload_i - 0.014 airexp_i)$$
$$\hat{\mu}_i = \exp(-0.406) \times \exp(0.569 type_i) \times \exp(0.165 bombload_i) \times \exp(-0.014 airexp_i).$$

Από τις εκτιμήσεις των συντελεστών $\beta_0, \beta_1, \beta_2, \beta_3$ μπορούμε να βγάλουμε κάποια συμπεράσματα στην περίπτωση, που το μοντέλο που προσαρμόσαμε είναι το κατάλληλο. Συνεπώς, καταλήγουμε στα εξής συμπεράσματα:

Ο αναμενόμενος αριθμός καταστροφών, που προκλήθηκαν στον τύπο αεροσκάφους A6 (*type=1*) είναι $\exp(0.569) = 1.77$ φορές μεγαλύτερος από τον αντίστοιχο αριθμό καταστροφών του αεροσκάφους A4 (*type=0*) όταν και τα δύο αεροσκάφη κατείχαν το ίδιο φορτίο βόμβας και ήταν επανδρωμένα με πλήρωμα με ίδια εμπειρία.

Κάθε επιπλέον τόνος, που χρησιμοποιήθηκε για το φορτίο της βόμβας φαίνεται να αυξάνει τον αναμενόμενο αριθμό καταστροφών, που προκλήθηκαν κατά $(\exp(0.165) - 1) \times 100\% \approx (1.18 - 1) \times 100\% = 18\%$ υπό την προϋπόθεση ότι τα αεροσκάφη είναι του ίδιου τύπου (A4 ή A6) και είναι επανδρωμένα με πλήρωμα με την ίδια εμπειρία.

Κάθε επιπλέον μήνας εμπειρίας του πληρώματος μειώνει την αναμενόμενη τιμή των καταστροφών των αεροσκαφών κατά ποσοστό $(1 - \exp(-0.014)) \times 100\% \approx (1 - 0.99) \times 100\% = 1\%$ υπό την προϋπόθεση ότι τα αεροσκάφη είναι του ίδιου τύπου (A4 ή A6) και κατέχουν το ίδιο φορτίο βόμβας.

6.4.1 Προβλέψεις Διαφόρων Σεναρίων

Στη εν λόγω παράγραφο θα δημιουργήσουμε διάφορες "ρεαλιστικές" τιμές για τις επεξηγηματικές μεταβλητές και θα εκτιμήσουμε με βάση το μοντέλο Poisson, που είδαμε παραπάνω την αναμενόμενη τιμή της μεταβλητής απόκρισης.

Δημιουργούμε δύο διανύσματα για τις επεξηγηματικές μεταβλητές *bombload* και *airexp*, όπως φαίνεται παρακάτω. Για το σκοπό αυτό χρησιμοποιούνται οι συναρτήσεις *min()*, *max*(), οι οποίες επιστρέφουν την ελάχιστη και τη μέγιστη τιμή και τις συναρτήσεις *mean*() και *median*(), οι οποίες επιστρέφουν τη μέση τιμή και τη διάμεσο.

Τα διανύσματα τα ονομάζουμε *bomb* και *air*.

```
bomb<-
c(min(aircraft1$bombload),mean(aircraft1$bombload),median(aircraft1$bombload),
max(aircraft1$bombload))

> bomb

[1]  4.0  8.1  7.5 14.0

 air<-
c(max(aircraft1$airexp),mean(aircraft1$airexp),median(aircraft1$airexp),min(aircraft
1$airexp))

> air

[1] 120.00000  80.76667  80.25000  50.00000

data<-cbind(bomb,air)
```

Εισάγουμε σε ένα διάνυσμα τις εκτιμήσεις των μεταβλητών του μοντέλου *air.model* και το καλούμε *coef*.

```
coef<-air.model$coef
```

Θέλουμε να υπολογίσουμε για κάποια συγκεκριμένα σενάρια την αναμενόμενη τιμή καταστροφών και για τους δύο τύπους αεροσκαφών (A4 και A6).

Θα δημιουργήσουμε, λοιπόν την παρακάτω συνάρτηση, την οποία καλούμε *senarios*, η οποία δέχεται ως ορίσματα τον διδιάστατο πίνακα *data* και το διάνυσμα *coef* και επιστρέφει μια λίστα, το πρώτο στοιχείο της οποίας περιλαμβάνει τις αναμενόμενες τιμές για τα 4 πιθανά σενάρια για τον τύπο A4, ενώ το δεύτερο στοιχείο αυτής περιλαμβάνει τα αντίστοιχα σενάρια για τον τύπο A6.

Η συνάρτηση στην R είναι

```
senarios<-function(data,coef){

y<-seq(1:4)

z<-seq(1:4)

for(i in 1:4){

y[i]<-exp(coef[1]+coef[3]*data[i,1]+coef[4]*data[i,2])

z[i]<-y[i]*exp(coef[2])

}
```

```
p<-list(y,z)

return(p)

}
```

Εκτελώντας την παραπάνω συνάρτηση στην R

```
> sen<-senarios(data,coef)

> sen

[[1]]

[1] 0.2548779 0.8536482 0.7784094 3.4341602



[[2]]

[1] 0.4501479 1.5076550 1.3747732 6.0651781
```

Στον παρακάτω πίνακα φαίνονται τα αποτελέσματα για τα 4 σενάρια.

Σενάρια	**Φορτίο βομβών (τόνους)**	**Εμπειρία (μήνες)**	**Αναμενόμενος αριθμός καταστροφών για το Α4**	**Αναμενόμενος αριθμός καταστροφών για το Α6**
Καλύτερο (Minimum)	4.0	120.00	0.255	0.450
Τυπικό (Mean)	8.1	80.77	0.854	1.507
Median	7.5	80.25	0.778	1.374
Χειρότερο (Maximum)	14.0	50.00	3.434	6.065

Πίνακας 6.5: *Σημειακές εκτιμήσεις των αναμενόμενων καταστροφών των δύο τύπων αεροσκαφών για τα διάφορα σενάρια.*

Αναλύοντας, λοιπόν τα αποτελέσματα του παραπάνω πίνακα θα παρατηρήσουμε ότι για μια τυπική αποστολή (*mean*) με το αεροσκάφος Α4 ο αριθμός περιοχών, που θα καταστραφούν αναμένεται να είναι 0.85, ενώ σε μια αντίστοιχη αποστολή με το αεροσκάφος Α6 θα είναι 1.5. Επιπλέον, για το χειρότερο σενάριο (*maximum*), που μπορούμε να έχουμε, δηλαδή τον μέγιστο φορτίο βομβών ίσο με 14 τόνους και συνολικά στο πλήρωμα την ελάχιστη εμπειρία ίση με 50 μήνες, ο αναμενόμενος αριθμός καταστροφών στο αεροσκάφος Α4 θα είναι 3.43, ενώ στο αεροσκάφος Α6 θα είναι 6.06.

6.5 Εφαρμογή σε Δίτιμα Δεδομένα με Διαφορετικές Συναρτήσεις Σύνδεσης

6.5.1 Περιγραφή της Εφαρμογής

Στην εφαρμογή αυτή χρησιμοποιούμε στοιχεία μιας ιατρικής μονάδας με 54 ασθενείς μέσης ηλικίας (*Agresti, 1990*). Οι ασθενείς υποβλήθηκαν σε μια ψυχιατρική εξέταση για να διαπιστωθεί αν υπάρχουν συμπτώματα γεροντικής άνοιας. Στα πλαίσια αυτής της εξέτασης, έγιναν μετρήσεις όσον αφορά το βαθμό ευφυΐας των ασθενών σύμφωνα με την κλίμακα ευφυΐας ενηλίκων Wechsler, το οποίο καλείται WAIS (*Wechsler Adult Intelligence Scale*). Σκοπός της έρευνας αυτής ήταν να διαπιστωθεί αν υπάρχει γεροντική άνοια και ποιά είναι η συσχέτιση με το βαθμό της κλίμακας WAIS.

Η μεταβλητή απόκρισης δείχνει αν ο ασθενής πάσχει από γεροντική άνοια ή όχι. Συνεπώς, η μεταβλητή απόκρισης του προβλήματος ακολουθεί τη διωνυμική κατανομή για την ειδική περίπτωση όπου η παράμετρος $n = 1$ ή αλλιώς την κατανομή Bernoulli. Γι' αυτό το λόγο θα προσαρμόσουμε το μοντέλο της λογιστικής παλινδρόμησης στα δεδομένα μας και εν συνεχεία θα το συγκρίνουμε με τα μοντέλα *probit* και *complementary log-log*.

Οι συναρτήσεις σύνδεσης, δηλαδή που χρησιμοποιούμε είναι:

- *logit*: $g(\mu) = \text{logit}(\mu) = \log\frac{\mu}{1-\mu}$.
- *probit*: $g(\mu) = \Phi^{-1}(\mu)$.
- *cloglog* (*complementary log-log link*): $g(\mu) = \log(-\log(1-\mu))$.

Τα δεδομένα φαίνονται στον παρακάτω πίνακα.

Μεταβλητή	**Περιγραφή**	**Κωδικοποίηση/Μονάδα μέτρησης**
Senility	Ύπαρξη γεροντικής άνοιας	1 = ΝΑΙ 0 = ΟΧΙ
wais	Βαθμός ευφυΐας	Μία κλίμακα, που παίρνει τιμές από $0,...,20$

Πίνακας 6.6: *Περιγραφή των μεταβλητών.*

6.5.2 Εισαγωγή των Δεδομένων στην R

Τα δεδομένα είναι διαθέσιμα στο δικτυακό τόπο

http://statathens.aueb.gr/~jbn/winbugs_book/files/chap02/ex3/chap02_ex3_wais_data.dat.

Συνεπώς, για την εισαγωγή των δεδομένων στην R θα χρησιμοποιήσουμε την παρακάτω εντολή, ώστε να δημιουργήσουμε ένα πλαίσιο δεδομένων το οποίο καλούμε *data*.

```
>data<-read.table("http://stat-athens.aueb.gr/~jbn/winbugs_book/files/chap02/ex3/chap02_ex3_wais_data.dat",header=TRUE)
```

Χρησιμοποιούμε την συνάρτηση *attach*(), έτσι ώστε να λαμβάνουμε τις τιμές των μεταβλητών του πλαισίου δεδομένων μας απλά γράφοντας το όνομά τους.

```
>attach(data)
```

Η μεταβλητή *senility* είναι μια κατηγορική μεταβλητή, γι' αυτό θα πρέπει να την ορίσουμε στην R ως κατηγορική. Αυτό επιτυγχάνεται στην R με την εντολή

```
>senility<-as.factor(senility)
```

6.5.3 Περιγραφή των Δεδομένων

Κάποια περιγραφικά χαρακτηριστικά για τις μεταβλητές του προβλήματος δίνονται στην R χρησιμοποιώντας τη συνάρτηση *summary*().

```
> summary(senility)
 0  1
40 14
> summary(wais)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.00    9.00   11.00   11.57   14.00   20.00
```

Παρατηρούμε ότι η ελάχιστη τιμή και η μέγιστη τιμή για την μεταβλητή *wais* είναι 4.00 και 20.00, αντίστοιχα. Επιπλέον, η μέση τιμή είναι 11.57, η διάμεσος είναι 11.00 και τέλος οι τιμές για το πρώτο και τρίτο τεταρτημόριο, που είναι 9.00 και 14.00, αντίστοιχα. Όσον αφορά τη δίτιμη μεταβλητή απόκρισης λαμβάνουμε τη συχνότητα των παρατηρήσεων για τις δύο κατηγορίες στο δείγμα μας και πιο συγκεκριμένα για

την κατηγορία 1 (ύπαρξη γεροντικής άνοιας) η συχνότητα είναι 40 και για την κατηγορία 0 (μη ύπαρξη γεροντικής άνοιας) η συχνότητα είναι 14 .

6.5.4 Προσαρμογή των Διαφορετικών Μοντέλων και Συμπεράσματα

Για την προσαρμογή των μοντέλων παλινδρόμησης θα χρησιμοποιήσουμε τη συνάρτηση glm() της R.

- **Μοντέλο λογιστικής παλινδρόμησης**

Η προσαρμογή του μοντέλου γίνεται εκτελώντας τις παρακάτω εντολές

```
logit.m<-glm(senility~wais,family=binomial,data=data)
summary(logit.m)
```

```
Call:
glm(formula = senility ~ wais, family = binomial, data = data1)


Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.6702  -0.7402  -0.4749   0.5200   2.1157
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.4040     1.1918   2.017  0.04369 *
wais         -0.3235     0.1140  -2.838  0.00453 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


(Dispersion parameter for binomial family taken to be 1)
   Null deviance: 61.806  on 53  degrees of freedom
Residual deviance: 51.017  on 52  degrees of freedom
AIC: 55.017
```

```
Number of Fisher Scoring iterations: 5
```

Αποτελέσματα 6.10

- **Μοντέλο Probit**

Η προσαρμογή του μοντέλου γίνεται εκτελώντας τις παρακάτω εντολές

```
> probit.m<-glm(senility~wais,family=binomial(link="probit"),data=data)
> summary(probit.m)
Call:
glm(formula = senility ~ wais, family = binomial(link = "probit"),
    data = data1)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6344  -0.7543  -0.4743   0.5527   2.1170
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.38616    0.68528   2.023  0.04310 *
wais        -0.18801    0.06301  -2.984  0.00284 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61.806  on 53  degrees of freedom
Residual deviance: 50.984  on 52  degrees of freedom
AIC: 54.984
Number of Fisher Scoring iterations: 5
```

Αποτελέσματα 6.11

Το προσαρμοσμένο μοντέλο θα είναι

$g(\hat{\pi}) = 1.38616 - 0.18801 wais$

Όσον αφορά την ερμηνεία των συντελεστών του μοντέλου αναφέρουμε ότι η τιμή του z-score (δηλαδή η εκτιμώμενη πιθανότητα της γεροντικής άνοιας) μειώνεται κατά 0.18801 για αύξηση της επεξηγηματικής μεταβλητής *wais* κατά μία μονάδα. Επιπλέον η αναμενόμενη πιθανότητα εμφάνισης της ασθένειας όταν η επεξηγηματική μεταβλητή *wais* είναι ίση με μηδέν είναι Φ(1.38616) = 0.917151, όπου Φ() είναι η συνάρτηση κατανομής πιθανότητας της τυποποιημένης κανονικής κατανομής.

- **Διωνυμικό μοντέλο με συνάρτηση σύνδεσης complementary log-log**

Η προσαρμογή του μοντέλου γίνεται εκτελώντας τις παρακάτω εντολές

```
> complog.m<-glm(senility~wais,family=binomial(link="cloglog"),data=data)

> summary(complog.m)

Call:
glm(formula = senility ~ wais, family = binomial(link = "cloglog"),

    data = data1)

Deviance Residuals:

    Min       1Q   Median       3Q      Max

-1.7474  -0.7265  -0.4988   0.4662   2.0717

Coefficients:

            Estimate Std. Error z value Pr(>|z|)

(Intercept)  1.42612    0.81686   1.746   0.0808 .

wais        -0.25075    0.08389  -2.989   0.0028 **

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

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 61.806  on 53  degrees of freedom

Residual deviance: 51.311  on 52  degrees of freedom
AIC: 55.311
Number of Fisher Scoring iterations: 6
```

Αποτελέσματα 6.12

Από τα Αποτελέσματα 6.12 το μοντέλο μας θα έχει τη μορφή

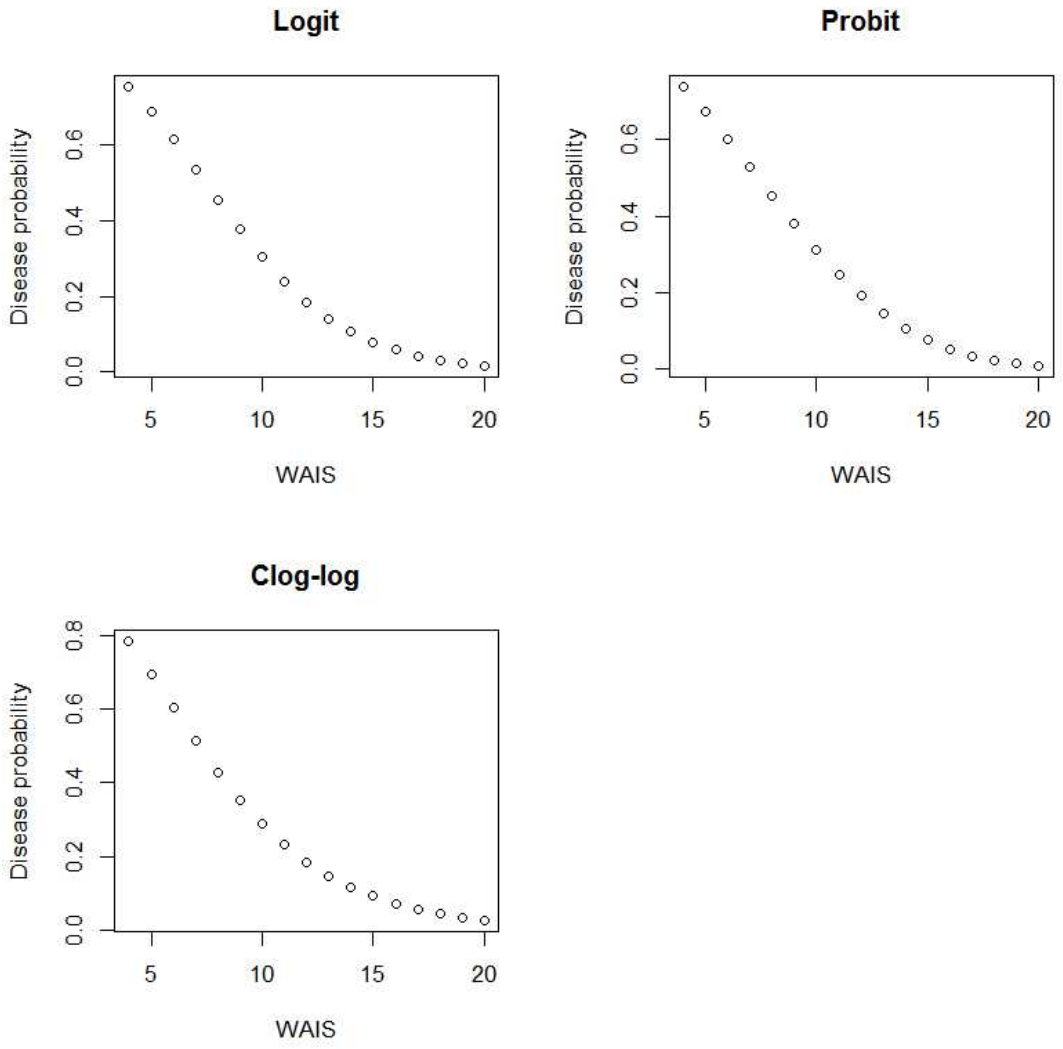
$$\log(-\log(1-\hat{\pi}_i)) = b_0 + b_1 wais \Rightarrow \log(-\log(1-\hat{\pi}_i)) = 1.42612 - 0.25075 wais.$$

Συνεπώς, η αναμενόμενη πιθανότητα εμφάνισης της ασθένειας όταν η επεξηγηματική μεταβλητή *wais* είναι ίση με μηδέν είναι

$$1-\exp(-\exp(b_0)) = 1-\exp(-\exp(1.42612)) = 0.9844317.$$

Επιπλέον, αν η τιμή (έστω x) της μεταβλητής *wais* αυξηθεί κατά 1 μονάδα, η αναμενόμενη πιθανότητα εμφάνισης της ασθένειας θα γίνει

$$1-[1-\pi(x)]^{\exp(-0.25075)} = 1-[1-\pi(x)]^{2.177586}, \text{ όπου } \log(-\log(1-\pi(x))) = b_0 + b_1 x.$$



Γράφημα 6.6: *Γραφική παράσταση της εκτιμώμενης πιθανότητας εμφάνισης της ασθένειας σε σχέση με την μεταβλητή WAIS.*

Από την παραπάνω γραφική παράσταση διακρίνουμε ότι και τα τρία γενικευμένα γραμμικά μοντέλα προσδίδουν μια αρνητική σχέση μεταξύ του WAIS σκορ και της εμφάνισης γεροντικής άνοιας.

Παραμέτροι	**Logit**	**Probit**	**Clog-log**
b_0	2.4040	1.38616	1.42612
b_1	-0.3235	-0.18801	-0.25075
AIC	55.017	54.984	55.311
Deviance	51.017	50.984	51.311

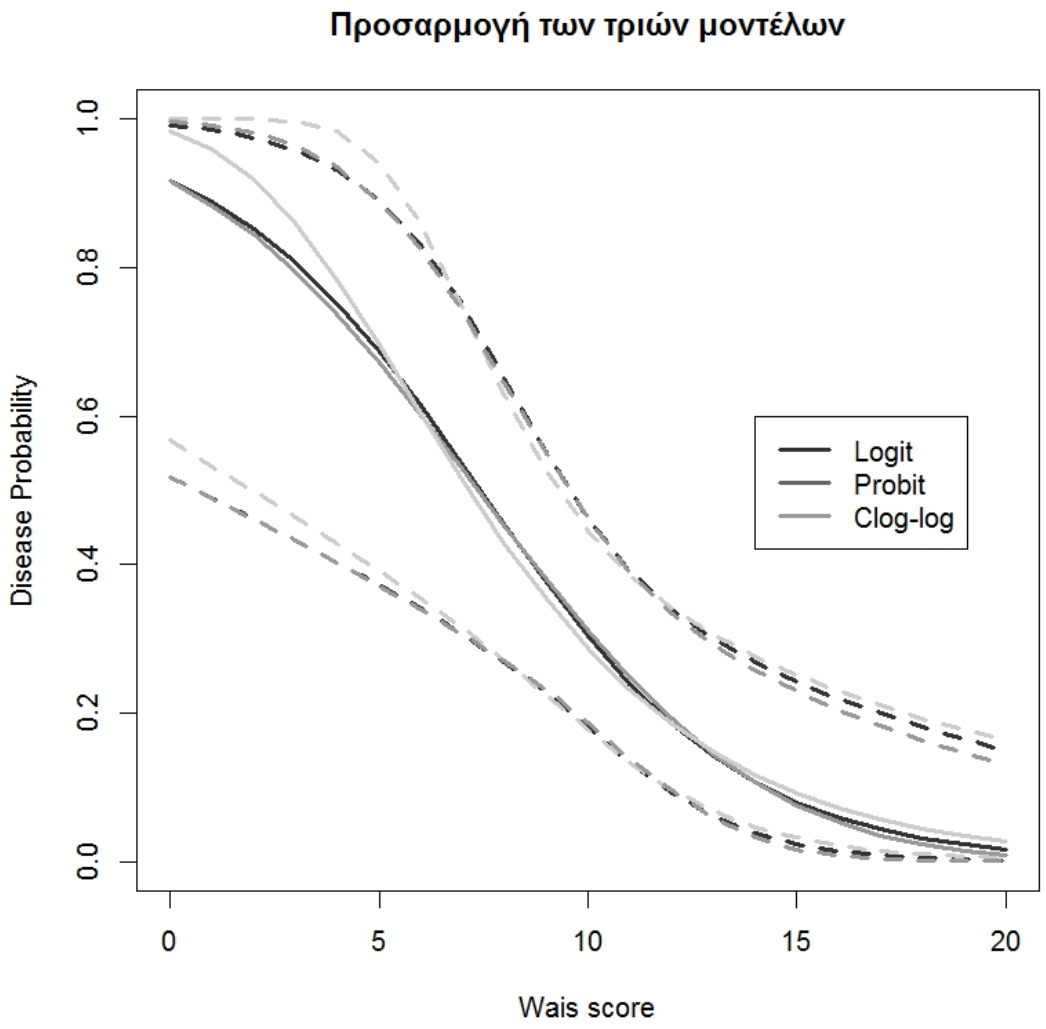
Πίνακας 6.7: *Εκτιμήσεις των συντελεστών, δείκτης AIC και ελεγχοσυνάρτηση deviance για κάθε ένα από τα τρία μοντέλα.*

Σύμφωνα με την τιμή της *deviance* μπορούμε να πούμε ότι το μοντέλο *probit* έχει καλύτερη προσαρμογή στα δεδομένα μας συγκριτικά με τα άλλα δύο μοντέλα, αφού λαμβάνει την τιμή 50.984, η οποία είναι η μικρότερη συγκριτικά με τα άλλα δύο μοντέλα. Στο ίδιο συμπέρασμα καταλήγουμε συγκρίνοντας τα μοντέλα μας με βάση το κριτήριο AIC, καθώς για το μοντέλο *probit* λαμβάνει την μικρότερη τιμή, 54.984 συγκριτικά με τα άλλα δύο μοντέλα.

Επιπλέον, χρησιμοποιώντας τον παρακάτω κώδικα στην R δημιουργούμε τα γραφήματα προσαρμογής των τριών μοντέλων σε σχέση με τις τιμές, που μπορεί να πάρει η επεξηγηματική μεταβλητή WAIS, καθώς και τα αντίστοιχα 95% διαστήματα εμπιστοσύνης των πιθανοτήτων επιτυχίας.

```
wais<-c(0:20)
senility<-rep(1,21)
new<-data.frame(wais,senility)
data2<-rbind(data,new)

logit1.m<-glm(senility~wais,family=binomial,data=data2[1:54,])
probit1.m<-glm(senility~wais,family=binomial(link="probit"),data=data2[1:54,])
complog1.m<-glm(senility~wais,family=binomial(link="cloglog"),data=data2[1:54,])

lyhat<-predict(logit1.m,newdata=as.data.frame(data2[55:75,1]),type="response")
pyhat<-predict(probit1.m,newdata=as.data.frame(data2[55:75,1]),type="response")
cyhat<-predict(complog1.m,newdata=as.data.frame(data2[55:75,1]),type="response")

par(cex.main=1.1,lwd=3)
color<-gray(1:4/5)
```

```
plot(data2[55:75,1],lyhat,ylim=c(0,1),xlim=c(0,20),type="l",xlab="Wais
score",ylab="Disease Probability", main="Προσαρμογή των τριών
μοντέλων",col=color[1],lwd=3)

legend(14,0.6,lty=c(1,1,1),col=color,lwd=3,legend=c("Probit","Clog-log","Logit"))

lines(data2[55:75,1],pyhat,lwd=3,col=color[3])
lines(data2[55:75,1],cyhat,lwd=3,col=color[4])

lzhat<-predict(logit1.m,newdata=as.data.frame(data2[55:75,1]),se.fit=TRUE)
lzupper<-lzhat$fit+1.96*lzhat$se.fit
lzlower<-lzhat$fit-1.96*lzhat$se.fit
lyupper<-exp(lzupper)/(1+exp(lzupper))
lylower<-exp(lzlower)/(1+exp(lzlower))

lines(data2[55:75,1],lyupper,lty=2,lwd=3,col=color[1])
lines(data2[55:75,1],lylower,lty=2,lwd=3,col=color[1])

czhat<-predict(complog1.m,newdata=as.data.frame(data2[55:75,1]),se.fit=TRUE)
czupper<-czhat$fit+1.96*czhat$se.fit
czlower<-czhat$fit-1.96*czhat$se.fit
cyupper<-(1-exp(-exp(czupper)))
cylower<-(1-exp(-exp(czlower)))

lines(data2[55:75,1],cyupper,lty=2,lwd=3,col=color[4])
lines(data2[55:75,1],cylower,lty=2,lwd=3,col=color[4])

pzhat<-predict(probit1.m,newdata=as.data.frame(data2[55:75,1]),se.fit=TRUE)
pzupper<-pzhat$fit+1.96*pzhat$se.fit
pzlower<-pzhat$fit-1.96*pzhat$se.fit
pyupper<-pnorm(pzupper,0,1)
pylower<-pnorm(pzlower,0,1)

lines(data2[55:75,1],pyupper,lty=2,lwd=3,col=color[3])
lines(data2[55:75,1],pylower,lty=2,lwd=3,col=color[3])
```



Γράφημα 6.7: *Προσαρμογή των τριών γενικευμένων γραμμικών μοντέλων. Παρουσιάζονται οι εκτιμώμενες πιθανότητες επιτυχίας, καθώς και τα αντίστοιχα 95% διαστήματα εμπιστοσύνης.*

Στο Γράφημα 6.7 παρατηρούμε ότι τα μοντέλα *probit* και *logit* έχουν (σχεδόν) όμοια προσαρμογή. Οι εκτιμήσεις των πιθανοτήτων να εμφανιστεί γεροντική άνοια είναι περίπου ίδιες για τα δύο μοντέλα. Αντιθέτως, το μοντέλο *complementary log-log* φαίνεται να παρουσιάζει κάποιες αποκλίσεις συγκριτικά με τα άλλα δύο μοντέλα κυρίως για μεγάλες τιμές της πιθανότητας εμφάνισης της ασθένειας. Θυμίζουμε, ότι με βάση το κριτήριο επιλογής μοντέλου AIC το *probit* μοντέλο φαίνεται να έχει ελαφρώς καλύτερη προσαρμογή από το λογιστικό μοντέλο, ενώ με βάση τη γραφική αναπαράσταση της προσαρμογής των μοντέλων διακρίνουμε ότι η προβλεπτική συμπεριφορά των δύο μοντέλων σχεδόν ταυτίζεται. Συνεπώς, μπορούμε τελικά να επιλέξουμε το λογιστικό μοντέλο, καθώς η ερμηνεία των συντελεστών αυτού είναι

πιο εύκολη και έχει το πλεονέκτημα ότι συνδέεται με τις συμπληρωματικές πιθανότητες, γεγονός που καθιστά την ερμηνεία των συντελεστών πιο άμεση, ειδικότερα σε πιο πολύπλοκα μοντέλα.

6.5.5 Το Λογιστικό Μοντέλο Παλινδρόμησης

Η προσαρμογή του μοντέλου λογιστικής παλινδρόμησης δίνει τα παρακάτω αποτελέσματα

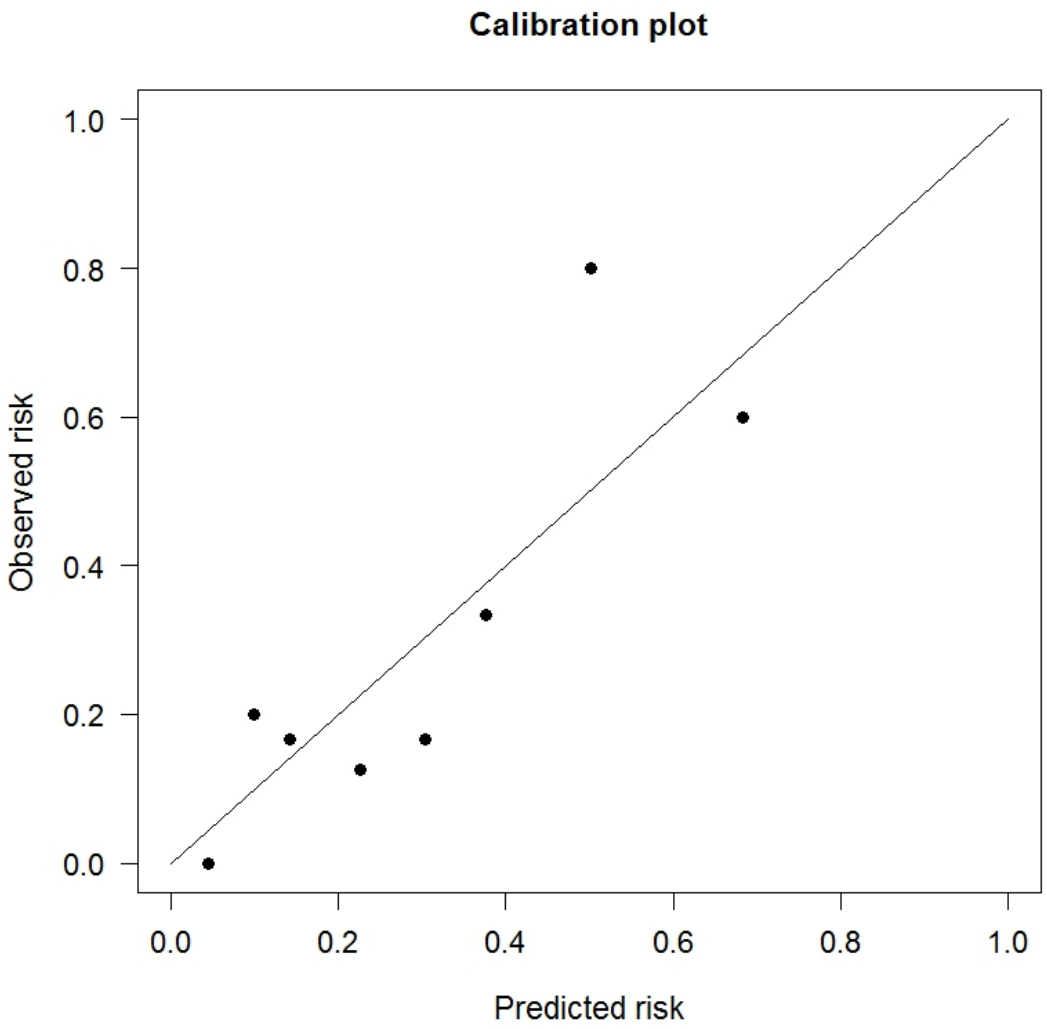
```
> logit.m<-glm(senility~wais,family=binomial,data=data)
> summary(logit.m)
Call:
glm(formula = senility ~ wais, family = binomial, data = data)
Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.6702  -0.7402  -0.4749   0.5200   2.1157
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.4040     1.1918   2.017  0.04369 *
wais         -0.3235     0.1140  -2.838  0.00453 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)
   Null deviance: 61.806  on 53  degrees of freedom
Residual deviance: 51.017  on 52  degrees of freedom
AIC: 55.017
Number of Fisher Scoring iterations: 5
```

Αποτελέσματα 6.13

Από τα Αποτελέσματα 6.13 παρατηρούμε ότι η επεξηγηματική μεταβλητή *wais* είναι στατιστικά σημαντική για το μοντέλο, αφού η $p-$τιμή του ελέγχου Wald για τον συντελεστή της είναι ίση με $0.00453 < 0.001$.

Για να ελέγξουμε την καλή προσαρμογή του μοντέλου θα εφαρμόσουμε τον έλεγχο καλής προσαρμογής των Hosmer-Lemeshow. Αυτό στην R επιτυγχάνεται με την συνάρτηση *plotCalibration()*.

```
predRisk<-predRisk(logit.m)

rangeaxis<-c(0,1)

plotCalibration(data,2,predRisk,groups=9)

$Table_HLtest

                    total meanpred meanobs predicted observed

[0.0168,0.0795)     8    0.044   0.000      0.35        0

[0.0795,0.1416)    10    0.099   0.200      0.99        2

0.1416              6    0.142   0.167      0.85        1

[0.1857,0.3034)     8    0.226   0.125      1.81        1

0.3034              6    0.303   0.167      1.82        1

0.3757              6    0.376   0.333      2.25        2

[0.4541,0.6137)     5    0.502   0.800      2.51        4

[0.6137,0.7521]     5    0.684   0.600      3.42        3

$Chi_square

[1] 4.522

$df

[1] 7

$p_value

[1] 0.7181
```



Γράφημα 6.8: *Διάγραμμα διασποράς των παρατηρούμενων ποσοστών επιτυχίας στις επιλεγμένες ομάδες ως προς τις εκτιμώμενες με βάση το μοντέλο λογιστικής παλινδρόμησης.*

Επισημαίνουμε ότι τα Predicted risk είναι οι εκτιμώμενες πιθανότητες επιτυχίας $\hat{\pi} = \frac{e_i}{m_i}$ στην ομάδα i και τα Observed risk είναι οι σχετικές συχνότητες εμφάνισης "επιτυχίας" στην ομάδα i και ορίζονται από τη σχέση $\tilde{\pi}_i = \frac{o_i}{m_i}$.

Παρατηρούμε ότι οι τιμές είναι "σχετικά" κοντά στην ευθεία, γεγονός που υποδηλώνει ότι το μοντέλο της λογιστικής παλινδρόμησης περιγράφει ικανοποιητικά τα δεδομένα μας.

Η προσαρμοσμένη εξίσωση παλινδρόμησης για το λογιστικό μοντέλο θα είναι

$$\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = 2.4040 - 0.3235 wais \Rightarrow \frac{\hat{\pi}}{1-\hat{\pi}} = \exp(2.4040 - 0.3235 wais).$$

Η αναμενόμενη συμπληρωματική (ή σχετική) πιθανότητα (odds) εμφάνισης γεροντικής άνοιας όταν το wais σκορ είναι μηδέν είναι $\exp(b_0) = \exp(2.4040) = 11.067$. Συνεπώς, όταν το wais σκορ είναι μηδέν η πιθανότητα εμφάνισης γεροντικής άνοιας είναι $\exp(2.40)/(1+\exp(2.40)) = 0.92$. Επιπλέον, αύξηση του wais σκορ κατά μία μονάδα αναμένεται να επιφέρει μείωση στον λογάριθμο της συμπληρωματικής πιθανότητας κατά 0.3235, ή ισοδύναμα αναμένεται να επιφέρει $1-\exp(-0.3235) = 1-0.7236 = 0.2764 \times 100\% = 27.64\%$ μείωση της σχετικής πιθανότητας εμφάνισης γεροντικής άνοιας.

Ολοκληρώνοντας την εφαρμογή παρουσιάζουμε τους διαγνωστικούς ελέγχους για το λογιστικό μοντέλο, οι οποίοι βασίζονται στα υπόλοιπα.

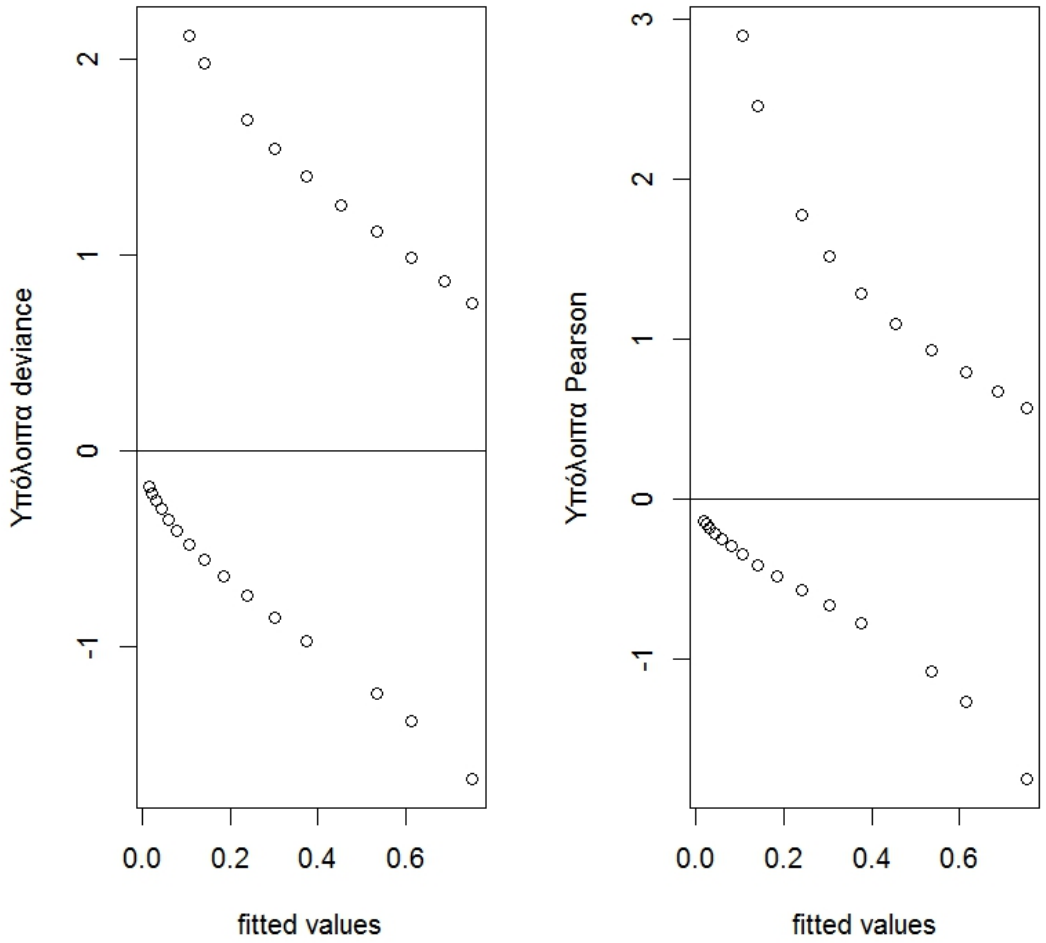
Αρχικά, υπολογίζουμε τα υπόλοιπα *deviance*, Pearson στην R με τις παρακάτω εντολές

```
res.deviance<-residuals(logit.m)

res.pearson<-residuals(logit.m,type="pearson")
```

Επιπλέον, υπολογίζουμε τις προσαρμοσμένες τιμές $\hat{\mu}_i = \hat{\pi}_i$.

```
fit<-fitted.values(logit.m)
```

```
par(mfrow=c(1,2))

plot(fit,res.deviance,xlab="fitted values",ylab="Υπόλοιπα deviance")
abline(h=0)

plot(fit,res.pearson,xlab="fitted values",ylab="Υπόλοιπα Pearson")
abline(h=0)
```



Γράφημα 6.9: *Τα διαγράμματα των υπολοίπων deviance και Pearson σε σχέση με τις εκτιμώμενες πιθανότητες εμφάνισης της ασθένειας.*

Από τα παραπάνω διαγράμματα παρατηρούμε ότι δεν μπορούμε να αντλήσουμε πληροφορίες για το μοντέλο μας λόγω του ότι τα δεδομένα μας είναι δίτιμα.

ΚΕΦΑΛΑΙΟ 7

ΕΠΙΛΟΓΟΣ

Η εργασία αυτή είχε ως σκοπό την εισαγωγή στα γενικευμένα γραμμικά μοντέλα και τον τρόπο προσαρμογής αυτών με τη χρήση του στατιστικού πακέτου R. Αρχικά, αναπτύχθηκε η βασική θεωρία και η δομή των γενικευμένων γραμμικών μοντέλων. Εν συνεχεία παρουσιάστηκαν αναλυτικά κάποιες μορφές μοντέλων, όπως το μοντέλο της λογιστικής παλινδρόμησης, το μοντέλο της παλινδρόμησης Poisson, το μοντέλο της αρνητικής διωνυμικής κατανομής, καθώς και κάποια μοντέλα για αποκριτικές μεταβλητές διάταξης. Τέλος, έγιναν κάποιες εφαρμογές σε κάποια από τα μοντέλα, που αναπτύχθηκαν στην διπλωματική με τη βοήθεια της R. Με τη μορφή παραδειγμάτων, παρουσιάστηκαν οι εντολές, που χρησιμοποιήθηκαν και αναλύθηκαν τα αποτελέσματα για την διεξαγωγή συμπερασμάτων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Α) Διεθνής Βιβλιογραφία

Agresti, A. (1990). *Categorial Data Analysis*. Wiley-Interscience. New York.

Cameron, A.C. & Trivedi, P.K. (1990). Regression-based tests for Overdispersion in the Poisson model. *Journal of Econometrics*, **6**, 347-364.

Crawley, M.J. (2007). *The R book.* John Wiley & Sons. New York.

Dobson, A.J. (2002). *An Introduction to Generalized Linear Models*. Second edition. Chapman & Hall / CRC. London.

Gill, J. (2001). *Generalized Linear Models: A Unified Approach*. Sage University Papers Series on Quantitative Applications in the Social Sciences. Sage. New York.

Hardin, J.W. & Hilbe, J.M. (2007). *Generalized Linear Models and Extensions*. Second edition. Stata Press.

Hosmer, D.W. & Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics: Theory and Methods*, **9**, 1043-1069.

Hosmer, D.W. & Lemeshow, S. (2000). *Applied Logistic Regression*. Second edition. John Wiley & Sons. New York.

Lindsey, J.K. (1997). *Applying Generalized Linear Models*. Springer – Verlag. New York.

Lindsey, J.K. (1999). On the use of Corrections for Overdispersion. *Applied Statistics*, **48**, 553-561.

Maindonald, J. & Braun, J. (2003). *Data Analysis and Graphics Using R – an Example – based Approach*. Cambridge University Press. London.

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*. Second edition. Chapman & Hall / CRC. London.

McCulloch, C.E., Searle, S.R. (2001). *Generalized, Linear and Mixed Models*. John Wiley & Sons. New York.

McFadden, D. (1974). *Conditional Logit Analysis of Qualitative Choice Behavior*. Academic Press. New York.

Mittlböck, M. & Schemper, M. (1996). Explained Variation for Logistic Regression. *Statistics in Medicine*, **15**, 1987-1997.

Montgomery, D., Peck, E. & Vining, G.G. (2006). *Introduction to Linear Regression Analysis*. Fourth edition. Wiley. Hoboken. New Jersey.

Myers, R.H., Montgomery, D.C., Vining, G.G. & Robinson, T.J. (2010). *Generalized Linear Models With Applications in Engineering and the Sciences*. Second 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.

Ntzoufras, I. (2008). *Bayesian Modeling Using Winbungs*. John Wiley & Sons. New York.

Olsson, U. (2002). *Generalized Linear Models: An Applied Approach*. Studentlitteratur. Lund.

Raftery, A. (1995). Bayesian Model Selection in Social Research. *Sociological Methology*, **25**, 111-163.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6,** 461-464.

Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*. **57**, 307-333.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439-447.

Β) Ελληνική Βιβλιογραφία

Καρώνη, Χ. (2009). *Μοντέλα Αξιοπιστίας και Επιβίωσης*. Συμεών. Αθήνα.

Κοκολάκης, Γ. & Φουσκάκης, Δ. (2009). *Στατιστική Θεωρία και Εφαρμογές*. Συμεών. Αθήνα.

Οικονόμου, Π. & Καρώνη, Χ. (2010). *Στατιστικά Μοντέλα Παλινδρόμησης*. Συμεών. Αθήνα.

Φουσκάκης, Δ. (2009). Παρουσίαση στο μάθημα Ανάλυση δεδομένων με Η/Υ-ΣΕΜΦΕ, (http://www.math.ntua.gr/~fouskakis/, τελευταία πρόσβαση στις 1/11/2013).

Γ) Ιστοσελίδες

http://biomet.oxfordjournals.org/content/68/1/13.short

http://cran.r-project.org/web/packages/HSAUR/

http://statmath.wu.ac.at/courses/heather_turner/

http://www.ats.ucla.edu/stat/r/dae/logit.htm

http://www.ats.ucla.edu/stat/r/dae/mlogit.htm

http://www.ats.ucla.edu/stat/r/dae/nbreg.htm

http://www.ats.ucla.edu/stat/r/dae/poissonreg.htm

http://www.ats.ucla.edu/stat/r/dae/probit.htm

http://www.ats.ucla.edu/stat/r/dae/zipoisson.htm