Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

Διπλωματική Εργασία

Γενικευμένα Γραμμικά Μοντέλα με Χρήση του Στατιστικού Πακέτου R

Νταϊλιάνας Χρίστος

Επιτροπή καθηγητών: Γ. Κοκολάκης (Υπεύθυνος καθηγητής)
Φ. Βόντα
Χ. Καρώνη

Αθήνα, Μάρτιος 2012

ΠΕΡΙΛΗΨΗ

Σκοπός αυτής της διπλωματικής εργασίας είναι η μελέτη των γενικευμένων γραμμικών μοντέλων. Η εργασία αυτή χωρίζεται σε δύο μέρη. Στο πρώτο μέρος παρουσιάζονται τα γενικευμένα γραμμικά μοντέλα και στο δεύτερο μέρος δίνονται παραδείγματα χρήσης των γενικευμένων γραμμικών μοντέλων χρησιμοποιώντας το στατιστικό πακέτο λογισμικού R.

Μετά από μια σύντομη εισαγωγή στο απλό γραμμικό μοντέλο και σε βασικές έννοιες της στατιστικής, γίνεται αναφορά στα γενικευμένα γραμμικά μοντέλα. Περιγράφονται οι μέθοδοι εκτίμησης των παραμέτρων που χρησιμοποιούνται στα πλαίσια των γενικευμένων γραμμικών μοντέλων και οι μέθοδοι ελέγχου της καταλληλότητας του μοντέλου. Έπειτα αναλύονται μέθοδοι για την αντιμετώπιση των κατηγορικών δεδομένων με τη βοήθεια γενικευμένων γραμμικών μοντέλων. Έμφαση δίνεται σε δίτιμες μεταβλητές και τη λογιστική παλινδρόμηση, καθώς και την Poisson παλινδρόμηση.

Στο δεύτερο μέρος αναφέρονται κάποια γενικά στοιχεία για το στατιστικό πακέτο R και στη συνέχεια γίνεται περιγραφή εντολών που χρησιμοποιούνται για γενικευμένα γραμμικά μοντέλα. Στο τελευταίο τμήμα της εργασίας αυτής παρουσιάζονται εφαρμογές και παραδείγματα γενικευμένων γραμμικών μοντέλων μέσω της γλώσσας R.

ABSTRACT

The goal of this project is the study of generalized linear models. The project is divided in two parts. In the first part the generalized linear models are presented, while in the second part examples of the use of generalized linear models using the statistical software R are given.

After a short introduction to the linear model and basic statistical principles, the generalized linear models are introduced. Methods of estimation on generalized linear models and goodness of fit statistics are described. Afterwards, methods for categorical data using generalized linear models are analyzed. Mainly, we work with binary variables and logistic regression, as well as Poisson regression.

In the second part, some general principles on the R language are mentioned and then commands that are used in generalized linear models are described. In the final part of the project applications and examples of generalized linear models using the R language are presented.

Περιεχόμενα

1 ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

1.1 Εισαγωγή

Τα γενικευμένα γραμμικά μοντέλα αναπτύχθηκαν από τους John Nelder και Robert Wedderburn το 1972. Σημαντική είναι και η συμβολή του Peter McCullagh ο οποίος είναι ο συγγραφέας μαζί με τον John Nelder του βιβλίου Generalized Linear Models του 1983. Το μεγαλύτερο μέρος της θεωρίας των γενικευμένων γραμμικών μοντέλων δεν αποτελεί κάτι καινούργιο στο χώρο της στατιστικής. Αποτελούν μία σύνδεση και επέκταση γνωστών μοντέλων παλινδρόμησης τα οποία εμφανίζουν κοινές ιδιότητες και έχουν κοινή μέθοδο εκτίμησης παραμέτρων, ωστόσο τα κοινά χαρακτηριστικά των εννοιών που μελετώνται μας οδηγούν στην ομαδοποίηση των τεχνικών και δημιουργούν ένα σύνολο, αυτό των γενικευμένων γραμμικών μοντέλων, όπου μπορούμε να μελετήσουμε τις κοινές αυτές ιδιότητες ως μία ομάδα στατιστικών μοντέλων. Επίσης, αυτή η ομαδοποίηση δημιούργησε προϋποθέσεις για περαιτέρω μελέτη νέων τεχνικών για την αντιμετώπιση διάφορων θεμάτων και σε συνδυασμό με τη χρήση υπολογιστών μπορούμε να μελετήσουμε προβλήματα τα οποία δεν μπορούσαμε πριν την χρήση γενικευμένων γραμμικών μοντέλων, αλλά και να προχωρήσουμε σε δύσκολους υπολογισμούς.

1.2 Μονοπαραμετρική Εκθετική Οικογένεια Κατανομών

Έστω μία τυχαία μεταβλητή Y της οποίας η συνάρτηση πυκνότητας πιθανότητας εξαρτάται από μία παράμετρο θ. Η κατανομή ανήκει στην εκθετική οικογένεια, αν μπορεί να γραφεί στη μορφή:

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)},\ y\epsilon Y, \theta\epsilon\Theta,$$

ή ισοδύναμα με $s(y) = exp[d(y)]$ και $t(\theta) = exp[c(\theta)]$ μπορούμε να γράψουμε την παραπάνω σχέση στη μορφή:

$$f(y;\theta) = exp[a(y)b(\theta) + c(\theta) + d(y)],\ y\epsilon Y, \theta\epsilon\Theta.$$

Στην τελευταία σχέση παρατηρούμε την συμμετρία μεταξύ των y και θ.

Τονίζεται ότι το στήριγμα της συνάρτησης πυκνότητας πιθανότητας, δηλαδή το σύνολο $S = \{y : f(y;\theta) > 0\}$ πρέπει να είναι ανεξάρτητο από την παράμετρο θ.

Αν ισχύει $a(y) = y$, τότε μπορούμε να πούμε ότι η κατανομή είναι σε κανονική μορφή και το $b(\theta)$ ονομάζεται φυσική παράμετρος της κατανομής.

1.2.1 Παραδείγματα Εκθετικής Οικογένειας Κατανομών

Πολλές γνωστές κατανομές, ανήκουν στην εκθετική οικογένεια. Η Κανονική κατανομή, η κατανομή Poisson και η Διωνυμική κατανομή είναι κάποιες κατανομές που όπως φαίνεται παρακάτω, ανήκουν στην εκθετική οικογένεια, αφού μπορούν να γραφούν στην κανονική μορφή.

• Η Κανονική κατανομή με γνωστή διασπορά σ^2 έχει συνάρτηση πυκνότητας πιθανότητας:

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

Η συνάρτηση της Κανονικής κατανομής μπορεί να γραφτεί:

$$f(y;\mu,\sigma) = e^{\frac{y\mu-\frac{\mu^2}{2}}{\sigma^2}-\frac{1}{2}[\frac{y^2}{\sigma^2}+ln(2\pi\sigma^2)]}.$$

Ανήκει δηλαδή στην εκθετική οικογένεια κατανομών, με:

$$\theta = \mu,$$

$$b(\theta) = \frac{\mu^2}{2},$$

$$\alpha(\phi) = \phi,$$

$$\phi = \sigma^2.$$

Το στήριγμα είναι το σύνολο των πραγματικών αριθμών, άρα το y δεν εξαρτάται από το μ.

• Η κατανομή Poisson έχει συνάρτηση πυκνότητας πιθανότητας:

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!},\ y\epsilon Z_+,$$

και μπορεί να γραφτεί στη μορφή:

$$f(y;\mu) = e^{y*ln(\mu)-\mu-ln(y!)}.$$

Δηλαδή ανήκει στην εκθετική οικογένεια κατανομών με:

$$\theta = ln(\mu),$$

$$\mu = e^{\theta},$$

$$b(\theta) = \mu,$$

$$c(y, \theta) = -ln(y!).$$

Το στήριγμα είναι το σύνολο των πραγματικών αριθμών, οπότε το y δεν εξαρτάται από το $ln(\mu)$.

• Η Διωνυμική κατανομή με παραμέτρους n και p έχει συνάρτηση πυκνότητας πιθανότητας:

$$f(y; n, p) = \begin{pmatrix} n \\ y \end{pmatrix} p^y (1-p)^{n-y},$$

και μπορεί να γραφτεί στη μορφή:

$$f(y; n, p) = exp[y * ln(\tfrac{p}{1-p}) + n * ln(1-p) + ln \begin{pmatrix} n \\ y \end{pmatrix}].$$

Δηλαδή ανήκει στην εκθετική οικογένεια κατανομών με:

$$\theta = ln(\tfrac{p}{1-p}),$$

$$c(y, \phi) = ln \begin{pmatrix} n \\ y \end{pmatrix}.$$

Οι τιμές που μπορεί να πάρει το y στην Διωνυμική κατανομή είναι ακέραιες και ισχύει $0 \leq y \leq n$, οπότε συμπεραίνουμε πως δεν εξαρτάται από το $ln(\frac{p}{1-p})$.

1.2.2 Ιδιότητες της Εκθετικής Οικογένειας Κατανομών

Όπως αναφέραμε, μία κατανομή ανήκει στην Εκθετική Οικογένεια Κατανομών αν μπορεί να γραφεί στη μορφή:

$$f(y; \theta) = s(y)t(\theta)e^{a(y)b(\theta)},\ y \epsilon Y, \theta \epsilon \Theta.$$

Χρειαζόμαστε εκφράσεις για την μέση τιμή και την διασπορά του $a(Y)$ [2]. Για να τις βρούμε, θα χρησιμοποιήσουμε τα ακόλουθα αποτελέσματα, που εφαρμόζονται για κάθε συνάρτηση πυκνότητας πιθανότητας. Από τον ορισμό της συνάρτησης πιθανότητας, ολοκληρώνοντας για όλες τις τιμές του y, θεωρώντας ότι η τυχαία μεταβλητή Y είναι συνεχής έχουμε,

$$\int_Y f(y; \theta) dy = 1.$$

Αν παραγωγίσουμε τα δύο μέλη ως προς θ, θα καταλήξουμε στη σχέση:

$$\int_Y \tfrac{df(y;\theta)}{d\theta} dy = 0.$$

Ισοδύναμα, αν παραγωγίσουμε δύο φορές τα δύο μέλη ως προς θ, θα καταλήξουμε στη σχέση:

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

Αυτά τα αποτελέσματα μπορούμε να τα χρησιμοποιήσουμε για κατανομές που ανήκουν στην εκθετική οικογένεια. Από τη γενική μορφή που είναι:

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

αφού παραγωγίσουμε ώστε να έχουμε:

$$\frac{df(y;\theta)}{d\theta} = [a(y)b^{'}(\theta) + c^{'}(\theta)]f(y;\theta),$$

προκύπτει η σχέση:

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

Αυτή η έκφραση, μπορεί να απλοποιηθεί στην σχέση:

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

Ισοδύναμα, καταλήγουμε στη σχέση:

$$E[a(Y)] = -c^{'}(\theta)/b^{'}(\theta).$$

Με ανάλογο τρόπο θα δουλέψουμε για τον υπολογισμό της διασποράς $var[a(Y)]$:

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

Ο δεύτερος όρος του δεξιού μέλους είναι ίσος με:

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

Έπειτα, φτάνουμε στην σχέση:

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

Τελικά θα καταλήξουμε στην σχέση:

$$Var[a(Y)] = \frac{b^{''}(\theta)c^{'}(\theta) - c^{''}(\theta)b^{'}(\theta)}{[b^{'}(\theta)]^3}.$$

Χρειαζόμαστε επίσης εκφράσεις για την μέση τιμή και την διασπορά των παραγώγων της λογαριθμικής συνάρτησης πιθανοφάνειας. Ο λογάριθμος της συνάρτησης πιθανοφάνειας για την εκθετική οικογένεια κατανομών, θεωρώντας ότι έχουμε μία μόνο παρατήρηση y, είναι:

$$l(\theta; y) = a(y)b(\theta) + c(\theta) + d(y).$$

Η παράγωγος αυτής της συνάρτησης ως προς θ είναι:

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

Η συνάρτηση $u(\theta; y)$ που ορίσαμε ως $u(\theta; y) = dl(\theta; y)/d\theta$ ονομάζεται score συνάρτηση και καθώς εξαρτάται από το y, μπορεί να θεωρηθεί ως τυχαία μεταβλητή που είναι:

$$u(\theta; Y) = U = a(Y)b^{'}(\theta) + c^{'}(\theta),$$

και η μέση τιμή αυτής είναι:

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

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

Η διασπορά της U ονομάζεται πληροφορία και θα τη συμβολίζουμε με $\mathcal{I}$. Η πληροφορία λοιπόν είναι:

$$\mathcal{I} = Var(U) = [b^{'}(\theta)^2]Var[a(Y)] \Leftrightarrow$$

$$\mathcal{I} = Var(U) = \frac{b^{''}(\theta)c^{'}(\theta)}{b^{'}(\theta)} - c^{''}(\theta).$$

Η στατιστική συνάρτηση score χρησιμοποιείται για συμπερασματολογία σχετικά με τις τιμές παραμέτρων στα γενικευμένα γραμμικά μοντέλα.

Μία ακόμα ιδιότητα της συνάρτησης U είναι η ακόλουθη:

$$var(U) = E(U^2) = E(-U^{'}).$$

Η πρώτη ισότητα προκύπτει από την ιδιότητα $var(U) = E(U^2) - [E(U)]^2$ με $E(U) = 0$, όπως δείξαμε προηγουμένως. Για να δείξουμε πως προκύπτει η δεύτερη ισότητα, παραγωγίζουμε το U ως προς θ κι έχουμε:

$$U^{'} = \frac{dU}{d\theta} = a(Y)b^{''}(\theta) + c^{''}(\theta).$$

Η μέση τιμή της παραγώγου της συνάρτησης U είναι:

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

$$= b^{''}(\theta)[-\frac{c^{'}(\theta)}{b^{'}(\theta)}] + c^{''}(\theta)$$

$$= -Var(U) = -\mathcal{I}.$$

Η τελευταία σχέση προκύπτει κάνοντας αντικατάσταση την ισότητα:

$$\mathcal{I} = Var(U) = \frac{b^{''}(\theta)c^{'}(\theta)}{b^{'}(\theta)} - c^{''}(\theta).$$

1.3 Στατιστικό μοντέλο

Σε μία στατιστική μελέτη, πολλές φορές χρειάζεται η πρόβλεψη μιας μεταβλητής, η οποία ονομάζεται μεταβλητή απόκρισης, μέσω κάποιων γνωστών μεταβλητών που ονομάζονται επεξηγηματικές μεταβλητές. Στατιστικό μοντέλο ονομάζεται η δημιουργία μιας μαθηματικής σχέσης μεταξύ αυτών των μεταβλητών. Η διαδικασία αυτής της μελέτης έχει τα σημαντικά βήματα που περιγράφονται παρακάτω.

Το πρώτο βήμα είναι ο προσδιορισμός του μοντέλου. Ένα μοντέλο προσδιορίζεται από δύο μέρη:

α) Μία εξίσωση που συνδέει τη μεταβλητή απόκρισης με την επεξηγηματική μεταβλητή και

β) Την κατανομή που ακολουθεί η μεταβλητή απόκρισης.

Απαραίτητη είναι στη συνέχεια η εκτίμηση των παραμέτρων που χρησιμοποιούνται στο μοντέλο. Στη συνέχεια δημιουργούμε διαστήματα εμπιστοσύνης και κάνουμε ελέγχους υποθέσεων για τις παραμέτρους του μοντέλου. Επίσης ερμηνεύουμε τις τιμές των αποτελεσμάτων και ελέγχουμε την επάρκεια του μοντέλου, δηλαδή πόσο καλά ερμηνεύονται τα δεδομένα από το μοντέλο μας.

1.4 Γενικευμένα γραμμικά μοντέλα

Σημαντικό ρόλο στα στατιστικά μοντέλα παίζει η Κανονική κατανομή, την οποία θέλουμε να ακολουθεί η μεταβλητή απόκρισης δοθέντων των επεξηγηματικών μεταβλητών. Πολλές φορές αυτό δεν συμβαίνει, καθώς η μεταβλητή απόκρισης μπορεί για παράδειγμα να παίρνει τις τιμές 0 ή 1, δηλαδή αποτυχία ή επιτυχία. Μπορούμε να θεωρήσουμε ότι οι μεταβλητές απόκρισης μπορούν να προέρχονται από μία γενικότερη οικογένεια κατανομών. Στην περίπτωση των γενικευμένων γραμμικών μοντέλων η μεταβλητή Y δοθείσης της τιμής X ακολουθεί κατανομές που ανήκουν στην εκθετική οικογένεια [2], [7], [3].

Ένα γενικευμένο γραμμικό μοντέλο ορίζεται από ένα σύνολο ανεξάρτητων τυχαίων μεταβλητών $Y_1, ..., Y_N$ καθεμία από τις οποίες ακολουθεί μία κατανομή που ανήκει στην εκθετική οικογένεια με τις ακόλουθες ιδιότητες:

1. Η κατανομή που ακολουθεί το κάθε Y_i έχει την κανονική μορφή και εξαρτάται από μία μόνο παράμετρο θ_i. Τα θ_i δεν είναι απαραίτητο να είναι όλα ίδια,

$$f(y_i;\theta_i) = exp[y_i b_i(\theta_i) + c_i(\theta_i) + d_i(y_i)].$$

2. Οι κατανομές από όλα τα Y_i είναι της ίδιας μορφής (για παράδειγμα όλες ανήκουν στην Κανονική, ή όλες στην Διωνυμική κατανομή).

Έτσι, η από κοινού συνάρτηση πυκνότητας πιθανότητας των $Y_1, ..., Y_N$ είναι:

$$f(y_1, ..., y_N; \theta_1, ..., \theta_N) = \prod_{i=1}^{N} exp[y_i b(\theta_i) + c(\theta_i) + d(y_i)]$$

$$= exp[\sum_{i=1}^{N} y_i b(\theta_i) + \sum_{i=1}^{N} c(\theta_i) + \sum_{i=1}^{N} d(y_i)].$$

Για τον προσδιορισμό του μοντέλου, συνήθως ενδιαφερόμαστε για ένα σύνολο παραμέτρων $\mathbf{b} = (b_1, ..., b_p)$, όπου $p < N$. Υποθέτουμε ότι $E(Y_i) = \mu_i$ όπου μ_i είναι κάποια συνάρτηση του θ_i. Για ένα γενικευμένο γραμμικό μοντέλο υπάρχει μετασχηματισμός του μ_i, τέτοιος ώστε [2],[7]:

$$g(\mu_i) = \mathbf{x}_i^T \mathbf{b}.$$

Στην εξίσωση αυτή, η g θεωρείται μονότονη και διαφορίσιμη συνάρτηση που λέγεται συνάρτηση σύνδεσης και το **x** το οποίο είναι ένα διάνυσμα επεξηγηματικών μεταβλητών:

$$\mathbf{x}_i = \begin{bmatrix} x_{i1} \\ . \\ . \\ x_{ip} \end{bmatrix},$$

έτσι,

$$\mathbf{x}_i^T = \begin{bmatrix} x_{i1} & . & . & x_{ip} \end{bmatrix},$$

και το **b** είναι ένα p x 1 διάνυσμα με παραμέτρους:

$$\mathbf{b} = \begin{bmatrix} b_1 \\ . \\ . \\ b_p \end{bmatrix}.$$

1.5 Συνάρτηση σύνδεσης

Έστω ένα διάνυσμα $y = (y_1, ..., y_N)$ από N στοιχεία μιας τυχαίας μεταβλητής Υ της οποίας οι συνιστώσες είναι ανεξάρτητα κατανεμημένες τυχαίες μεταβλητές με μέσους $\boldsymbol{\mu} = (\mu_1, ..., \mu_N)$. Ο καθορισμός του $\boldsymbol{\mu}$ γίνεται με έναν μικρό αριθμό άγνωστων παραμέτρων $b_1, ..., b_p$, $(p < N)$. Οι μεταβλητές $x_1, x_2, ..., x_p$ δημιουργούν μια γραμμική πρόβλεψη, η οποία είναι $\eta = \sum_1^p x_j b_j$. Στο γενικό γραμμικό μοντέλο έχουμε $\mu = \eta$, δηλαδή έχουμε την ταυτοτική συνάρτηση ως συνάρτηση σύνδεσης.

Στα γενικευμένα γραμμικά μοντέλα υπάρχει συνάρτηση g και ένα σύνολο παραμέτρων $\mathbf{b} = (b_1, ..., b_p)$, $(p < N)$ τέτοια ώστε ένας γραμμικός συνδυασμός των $\mathbf{b}$ να είναι ίσος με τη συνάρτηση της αναμενόμενης τιμής μ_i των Y_i, δηλαδή:

$$g(\mu_i) = \mathbf{x}_i^T \mathbf{b}.$$

Συνεπώς ένα γενικευμένο γραμμικό μοντέλο αποτελείται από τις εξής συνιστώσες:

• $Y_1, ..., Y_N$ με κατανομή από την εκθετική οικογένεια,

• παραμέτρους: $\mathbf{b} = \begin{bmatrix} b_1 \\ . \\ . \\ b_p \end{bmatrix}$,

• επεξηγηματικές μεταβλητές: $\mathbf{X} = \begin{bmatrix} \mathbf{X}_1^T \\ . \\ . \\ \mathbf{X}_N^T \end{bmatrix} = \begin{bmatrix} X_{11} & . & . & X_{1p} \\ . & . & & . \\ . & & . & . \\ X_{N1} & . & . & X_{Np} \end{bmatrix}$,

• μία γνήσια μονότονη και διαφορίσιμη συνάρτηση σύνδεσης g τέτοια ώστε:

$$g(\mu_i) = \mathbf{x}_i^T \mathbf{b} \quad , \quad \text{όπου } \mu_i = E(Y_i).$$

Η συνάρτηση σύνδεσης (link function) συσχετίζει τη γραμμική παράμετρο με την αναμενόμενη τιμή μ της μεταβλητής απόκρισης y. Στα κλασσικά γραμμικά μοντέλα η μέση τιμή μ ταυτίζεται με τη γραμμική πρόβλεψη. Τότε είναι εύλογο ότι η ταυτοτική συνάρτηση σύνδεσης μπορεί να πάρει οποιαδήποτε πραγματική τιμή. Σε περιπτώσεις όπου έχουμε διακριτές τιμές και η κατανομή που ακολουθούν είναι η Poisson κατανομή, πρέπει να ισχύει $\mu > 0$. Μπορεί στην περίπτωση αυτή η γραμμική παράμετρος η να είναι αρνητικό τη στιγμή που το μ δεν είναι. Μοντέλα με τέτοιου είδους μεταβλητές, εκφράζονται με τη λογαριθμική συνάρτηση σύνδεσης, $\eta = log(\mu)$, ή $\mu = e^{\eta}$ ώστε να υπάρχει γραμμική σχέση. Το μ

πρέπει να είναι θετικό σε αυτές τις περιπτώσεις.

Για την περίπτωση της διωνυμικής κατανομής, θεωρούμε τρείς βασικές συναρτήσεις σύνδεσης [7].

Αυτές είναι:

• logit: $\eta = log(\frac{\mu}{1-\mu})$.

• probit: $\eta = \Phi^{-1}(\mu)$, όπου Φ είναι η συνάρτηση κατανομής της Κανονικής κατανομής $N(0,1)$.

• complementary log-log: $\eta = log(-log(1-\mu))$.

Όταν $\eta = \theta$, όπου θ είναι η κανονική παράμετρος, οι κανονικές συναρτήσεις σύνδεσης είναι για τις παρακάτω κατανομές:

• Κανονική: $\eta = \mu$.

• Poisson : $\eta = log(\mu)$.

• Διωνυμική: $\eta = log(\frac{p}{1-\mu})$.

• Γάμμα: $\eta = \mu^{-1}$.

1.5.1 Παραδείγματα

Ιστορική Γλωσσολογία

Ας θεωρήσουμε μία γλώσσα, η οποία προέρχεται από μία άλλη γλώσσα. Ένα απλό μοντέλο για την αλλαγή στο λεξιλόγιο είναι, αν οι γλώσσες, μεταβάλλονται σε χρόνο t, τότε η πιθανότητα να εμφανίζουν συγγενικές λέξεις, για μία συγκεκριμένη έννοια, είναι $e^{-\theta t}$, όπου θ είναι μία παράμετρος, κατά προσέγγιση η ίδια για πολλές κοινές έννοιες. Σε μία μελέτη N διαφορετικών εννοιών, υποθέτουμε ότι ένας γλωσσολόγος κρίνει αν οι λέξεις των δύο γλωσσών έχουν συγγένεια ή όχι για κάποια έννοια [2].

Μπορούμε να αναπτύξουμε ένα γενικευμένο γραμμικό μοντέλο για να περιγράψουμε αυτή την κατάσταση.

Ορίζουμε τις τιμές $Y_1, ..., Y_N$ ως εξής:

$$Y_i = \begin{cases} 1 & \text{αν} \quad \text{οι δύο γλώσσες εμφανίζουν συγγενικές λέξεις για μία έννοια i} \\ 0 & \text{αν} \quad \text{οι λέξεις δεν σχετίζονται} \end{cases} .$$

Τότε:

$$P(Y_i = 1) = e^{-\theta t}, \theta \geq 0$$

και

$$P(Y_i = 0) = 1 - e^{-\theta t}, \theta \geq 0$$

για $\theta = [0, \infty]$ επειδή αναφερόμαστε σε πιθανότητες και οι τιμές των $P(Y_i)$ ανήκουν στο διάστημα $[0, 1]$.

Αυτή είναι μία περίπτωση διωνυμικής κατανομής (n, π) με $n = 1$ και $E(Y_i) = \pi = e^{-\theta t}$. Στην περίπτωση αυτή ως συνάρτηση σύνδεσης g χρησιμοποιήσαμε τη λογαριθμική:

$$g(\pi) = log(\pi) = -\theta t \ ,$$

έτσι ώστε $g[E(Y)]$ να είναι γραμμική ως προς την παράμετρο θ. Σε αυτήν την περίπτωση,

$$\mathbf{x}_i = [-t] \text{ και } \mathbf{b} = [\theta].$$

Ποσοστά θνησιμότητας

Για μεγάλο πληθυσμό, η πιθανότητα θανάτου ενός τυχαίου ατόμου σε κάποιο χρόνο είναι μικρή. Αν θεωρήσουμε ότι θάνατοι από κάποια μη μεταδοτική ασθένεια είναι ανεξάρτητα γεγονότα, τότε ο αριθμός θανάτων $\mathbf{Y}$ σε έναν πληθυσμό, μπορεί να μοντελοποιηθεί, μέσω της κατανομής Poisson:

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

όπου $y = 0, 1, 2, ...$ και $\mu = E(Y)$ είναι ο αναμενόμενος αριθμός θανάτων σε μία χρονική περίοδο, για παράδειγμα ένα έτος.

Το μ εξαρτάται από τον πληθυσμό, την χρονική περίοδο που παρακολουθούμε και διάφορα χαρακτηριστικά των μελών του πληθυσμού, όπως ηλικία, φύλο και ιατρικό ιστορικό.

Για παράδειγμα, μπορούμε να μοντελοποιήσουμε αυτή τη μελέτη ως:

$$E(Y) = \mu = n\lambda(\mathbf{x}^T\mathbf{b}),$$

όπου n είναι το μέγεθος πληθυσμού και $\lambda(\mathbf{x}^T\mathbf{b})$ το ποσοστό ανά 100.000 ατόμων ανά χρόνο.

Στον πίνακα 1 φαίνονται μετρήσεις για τις ηλικίες, τους θανάτους, τον πληθυσμό και η αναλογία των θανάτων ανά τον πληθυσμό.

Ηλικίες	Θάνατοι y_i	Πληθυσμός n_i	Αναλογία $y_i/n_i * 10000$
30-34	1	17,742	5.6
35-39	5	16,554	30.2
40-44	5	16,059	31.1
45-49	12	13,083	91.7
50-54	25	10,784	231.8
55-59	38	9,645	394.0
60-64	54	10,706	504.4
65-69	65	9,933	654.4

Πίνακας 1: *Αριθμός θανάτων ανδρών, από καρδιολογικά αίτια και το μέγεθος του πληθυσμού σε ομάδες ηλικιών, ανά 5 χρόνια.*

Ένα πιθανό μοντέλο είναι: $E(Y_i) = \mu_i = n_i e^{\theta_i} \quad , \quad Y_i \sim \text{Poisson}(\mu_i)$,

όπου:
$i = 1$ για την ομάδα ηλικίας 30-34 χρονών,
$i = 2$ για την ομάδα ηλικίας 35-39 χρονών,
$i = 3$ για την ομάδα ηλικίας 40-44 χρονών,
$i = 4$ για την ομάδα ηλικίας 45-49 χρονών,
$i = 5$ για την ομάδα ηλικίας 50-54 χρονών,
$i = 6$ για την ομάδα ηλικίας 55-59 χρονών,
$i = 7$ για την ομάδα ηλικίας 60-64 χρονών,
$i = 8$ για την ομάδα ηλικίας 65-69 χρονών.

Μπορούμε να δημιουργήσουμε το γενικευμένο γραμμικό μοντέλο χρησιμοποιώντας τη λογαριθμική συνάρτηση σύνδεσης:

$$g(\mu_i) = log(\mu_i) = log(n_i) + \theta_i,$$

η οποία έχει τη γραμμική συνιστώσα $\mathbf{x}_i^T\mathbf{b}$, με:

$$\mathbf{x}_i^T = \begin{bmatrix} log(n_i) & i \end{bmatrix} \text{ και } \mathbf{b} = \begin{bmatrix} 1 \\ \theta \end{bmatrix}.$$

[2]

2 ΕΚΤΙΜΗΤΙΚΗ

Έχοντας επιλέξει ένα συγκεκριμένο μοντέλο, είναι απαραίτητο στη συνέχεια να εκτιμήσουμε τις παραμέτρους και να υπολογίσουμε τις προβλεπόμενες τιμές.

Στα στατιστικά μοντέλα, για την εκτίμηση των παραμέτρων $b_1, ..., b_p$ χρησιμοποιείται η μέθοδος μέγιστης πιθανοφάνειας ή η μέθοδος ελαχίστων τετραγώνων.

2.1 Μέθοδος Μέγιστης Πιθανοφάνειας

Έστω $Y_1, ..., Y_N$ τυχαίες μεταβλητές με από κοινού συνάρτηση πυκνότητας πιθανότητας $f(\mathbf{y}, \boldsymbol{\theta})$ η οποία εξαρτάται από το διάνυσμα των παραμέτρων:

$$\boldsymbol{\theta} = [\theta_1, ..., \theta_p]^T.$$

Έστω Θ οι δυνατές τιμές του διανύσματος των παραμέτρων. Ο εκτιμητής μέγιστης πιθανοφάνειας του $\boldsymbol{\theta}$ είναι η τιμή $\hat{\boldsymbol{\theta}}$ η οποία μεγιστοποιεί τη συνάρτηση πιθανοφάνειας, $L(\hat{\boldsymbol{\theta}}, \mathbf{y}) = sup\{L(\boldsymbol{\theta}, \mathbf{y})\}$, $\theta \epsilon \Theta$.

Ισοδύναμα, μπορούμε να υπολογίσουμε την τιμή $\hat{\boldsymbol{\theta}}$ που μεγιστοποιεί το λογάριθμο της συνάρτησης πιθανοφάνειας, δηλαδή:

$$l(\hat{\boldsymbol{\theta}}, \mathbf{y}) = sup\{l(\boldsymbol{\theta}, \mathbf{y})\}, \theta \epsilon \Theta.$$

Ο εκτιμητής $\hat{\boldsymbol{\theta}}$ λαμβάνεται με διαφόριση της λογαριθμικής συνάρτησης πιθανοφάνειας για κάθε συνιστώσα του $\boldsymbol{\theta}$ και λύνοντας τις εξισώσεις:

$$\frac{\partial}{\partial \theta_j} l(\boldsymbol{\theta}, \mathbf{y}) = 0 \quad , \quad j = 1, ..., p.$$

Θα πρέπει να ελεγχθεί ότι οι λύσεις αντιστοιχούν σε μέγιστες τιμές του $l(\boldsymbol{\theta}, \mathbf{y})$. Αυτό επιτυγχάνεται όταν ο πίνακας των δευτέρων παραγώγων,

$$\frac{\partial^2 l(\boldsymbol{\theta}, \mathbf{y})}{\partial \theta_j \partial \theta_k},$$

για την τιμή $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ είναι αρνητικά ορισμένος.

2.2 Μέθοδος Ελαχίστων Τετραγώνων

Έστω $Y_1, ..., Y_N$ τυχαίες μεταβλητές με μέσες τιμές:

$$E(Y_i) = \mu_i \quad , \quad i = 1, ..., N.$$

Ας υποθέσουμε ότι τα μ_i είναι συναρτήσεις των παραμέτρων

$$\mathbf{b} = \begin{bmatrix} b_1 \\ . \\ . \\ b_p \end{bmatrix}, \quad (p < N),$$

που πρέπει να εκτιμηθούν.

Η μέθοδος ελαχίστων τετραγώνων χρησιμοποιείται για την εύρεση εκτιμητή $\hat{\mathbf{b}}$ που ελαχιστοποιεί το άθροισμα τετραγώνων των όρων των σφαλμάτων ε_i.

Με άλλα λόγια, θέλουμε να ελαχιστοποιήσουμε την ποσότητα:

$$S = \sum_{i=1}^{N} [Y_i - \mu_i(b)]^2,$$

δηλαδή:

$$S = \sum_{i=1}^{N} \varepsilon_i^2.$$

Οι εκτιμητές $\hat{b}$ βρίσκονται από παραγώγιση του S ως προς κάθε συνιστώσα b_j του **b** και λύνοντας στη συνέχεια τις εξισώσεις:

$$\frac{dS}{db_j} \quad , \quad j = 1, ..., p.$$

Είναι απαραίτητο να ελέγξουμε ότι οι λύσεις αντιστοιχούν σε ελάχιστα και γι' αυτό θα πρέπει ο πίνακας των δευτέρων παραγώγων να είναι θετικά ορισμένος.

2.3 Εκτίμηση στο Γενικευμένο Γραμμικό Μοντέλο

Οι μέθοδοι που χρησιμοποιούνται στα γενικευμένα γραμμικά μοντέλα βασίζονται στη μέθοδο Newton-Raphson και στη μέθοδο των score, την οποία θα εξηγήσουμε παρακάτω [2].

Θεωρούμε $Y_1, ..., Y_N$ ανεξάρτητες τυχαίες μεταβλητές. Θέλουμε να εκτιμήσουμε τις παραμέτρους **b** που σχετίζονται με τα Y_i μέσω των σχέσεων:

$$E(y_i) = \mu_i \text{ και } g(\mu_i) = \mathbf{x}_i^T \mathbf{b}.$$

Για κάθε Y_i, η λογαριθμική συνάρτηση πιθανοφάνειας είναι:

$$l_i = \frac{y_i\theta_i - b(\theta_i)}{\alpha(\Phi_i) + (y_i, \Phi_i)}.$$

Για να βρούμε την εκτιμήτρια μέγιστης πιθανοφάνειας για την παράμετρο b_j, χρειαζόμαστε τη:

$$\frac{\partial l}{\partial b_j} = U_j = \sum_{i=1}^{N} [\frac{\partial l_i}{\partial b_i}] = \sum_{i=1}^{N} [\frac{\partial l_i}{\partial_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial b_i}].$$

Θα υπολογίσουμε τον κάθε όρο του δεξιού μέλους ξεχωριστά:

- $\frac{\partial l_i}{\partial_i} = b'(\theta_i)(y_i - \mu_i)$,
- $\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}}$, με: $\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) var(Y_i)$,
- $\frac{\partial \mu_i}{\partial b_i} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial b_i} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$.

Τελικά καταλήγουμε στη σχέση,

$$U_j = \sum_{i=1}^{N} [\frac{y_i - \mu_i}{var(Y_i)} x_{ij} (\frac{\partial \mu_i}{\partial \eta_i})].$$

Η συνάρτηση $U = [U_1, ..., U_N]$ ονομάζεται score συνάρτηση.

Ο πίνακας διασποράς-συνδιασποράς των U_i έχει όρους $\mathcal{I}_{jk} = E[U_j U_k]$ που σχηματίζουν τον πίνακα:

$$\mathcal{I}_{jk} = E(\sum_{i=1}^{N} [\frac{Y_i - \mu_i}{var(Y_i)} x_{ij} (\frac{\partial \mu_i}{\partial \eta_i})] \sum_{l=1}^{N} [\frac{Y_l - \mu_l}{var(Y_l)} x_{lk} (\frac{\partial \mu_l}{\partial \eta_l})]) = \sum_{i=1}^{N} \frac{E[(Y_i - \mu_i)^2] x_{ij} x_{ik}}{[var(Y_i)]^2} (\frac{\partial \mu_i}{\partial \eta_i})^2,$$

όπου προκύπτει $\mathcal{I}_{jk} = \sum_{i=1}^{N} \frac{x_{ij} x_{ik}}{var(Y_i)} (\frac{\partial \mu_i}{\partial \eta_i})^2$. Ο πίνακας $\mathcal{I}_{jk}$ ονομάζεται πίνακας πληροφορίας. Από εδώ, μπορούμε να γράψουμε:

$$\mathcal{I} = \mathbf{X}^T \mathbf{W} \mathbf{X},$$

όπου **W** είναι ένας διαγώνιος N x N πίνακας, με στοιχεία $w_{ii} = \frac{1}{var(Y_i)} (\frac{\partial \mu_i}{\partial \eta_i})^2$.

Η μέθοδος Newton-Raphson δίνει την m-οστή προσέγγιση από τη σχέση:

$$b^{(m)} = b^{(m-1)} - [\frac{\partial^2 l}{\partial b_j b_k}]^{-1}_{b=b^{(m-1)}} U^{(m-1)}.$$

Ο πίνακας πληροφορίας $\mathcal{I} = E[UU^T]$, έχει τα στοιχεία:

$$\mathcal{I}_{jk} = E[U_j U_k] = E[\frac{\partial l}{\partial b_j} \frac{\partial l}{\partial b_k}] = E[\frac{\partial^2 l}{\partial b_j b_k}].$$

Από τα παραπάνω προκύπτει:

$$b^{(m)} = b^{(m-1)} + [\mathcal{I}^{(m-1)}]^{-1} U^{(m-1)}.$$

Λαμβάνουμε λοιπόν την εξίσωση:

$$\mathcal{I}^{(m-1)} b^{(m)} = \mathcal{I}^{(m-1)} b^{(m-1)} + U^{(m-1)},$$

και τελικά καταλήγουμε στην εξίσωση:

$$\mathbf{X}^T \mathbf{W} \mathbf{X} \mathbf{b}^{(m)} = \mathbf{X}^T \mathbf{W} \mathbf{z} \, .$$

Η προηγούμενη εξίσωση έχει την ίδια μορφή με τις κανονικές εξισώσεις των γενικευμένων γραμμικών μοντέλων που προκύπτουν από τα σταθμισμένα ελάχιστα τετράγωνα με τη διαφορά του ότι πρέπει να λυθούν με μια επαναληπτική μέθοδο επειδή τα **z** και **W** εξαρτώνται σε γενικές γραμμές από το **b**. Δηλαδή οι εκτιμήτριες μέγιστης πιθανοφάνειας των γενικευμένων γραμμικών μοντέλων προκύπτουν από μια επαναληπτική διαδικασία σταθμισμένων ελαχίστων τετραγώνων.

Στην περίπτωση των γενικευμένων γραμμικών μοντέλων, η διαδικασία της εκτίμησης των παραμέτρων, ακολουθείται από τη διαδικασία προσδιορισμού της προσαρμογής του μοντέλου. Δηλαδή είναι σημαντικό να ερευνήσουμε τη σχέση των δεδομένων που έχουμε παρατηρήσει, με τις τιμές που προβλέπονται μέσω του μοντέλου μας.

2.4 Μελέτη καταλληλότητας του μοντέλου

Αφού δημιουργήσουμε ένα μοντέλο, θα πρέπει να εξετάσουμε την ικανότητα του μοντέλου μας να περιγράψει τα δεδομένα που έχουμε. Είναι σημαντική δηλαδή η αξιολόγηση της σημαντικότητας των μεταβλητών στο μοντέλο. Θέλουμε να ερευνήσουμε ποιες μεταβλητές X_i είναι σημαντικές.

Για να επιλέξουμε τις στατιστικά σημαντικές μεταβλητές σε ένα μοντέλο, πραγματοποιούμε στατιστικούς ελέγχους. Αυτοί οι έλεγχοι καταλληλότητας, θα μας βοηθήσουν να ερευνήσουμε ποιες μεταβλητές δεν συνεισφέρουν στο μοντέλο. Αυτό θα το ελέγξουμε μέσω ελέγχων που αφορούν τους συντελεστές **b**.

Για να ερευνήσουμε την επάρκεια ενός μοντέλου, μπορεί να χρησιμοποιηθεί η δειγματική κατανομή της στατιστικής συνάρτησης D, όπως θα δούμε στη συνέχεια. Αυτό γίνεται εκτιμώντας το D από τα δεδομένα και συγκρίνοντας την τιμή με την κατάλληλη χ^2_{N-p} κατανομή.

2.4.1 Στατιστική συνάρτηση deviance

Η στατιστική συνάρτηση deviance, που επίσης ονομάζεται στατιστική συνάρτηση αναλογίας λογαριθμικής πιθανοφάνειας είναι:

$$D = 2[l(\hat{\mathbf{b}}_{max}; \mathbf{y}) - l(\hat{\mathbf{b}}; \mathbf{y})].$$

Το αποτέλεσμα του μετασχηματισμού της deviance είναι:

$$D = 2[l(\hat{\mathbf{b}}_{max}; \mathbf{y} - l(\mathbf{b}_{max}; \mathbf{y})] - 2[l(\hat{\mathbf{b}}; \mathbf{y} - l(\mathbf{b}; \mathbf{y})] + 2[l(\mathbf{b}_{max}; \mathbf{y} - l(\mathbf{b}; \mathbf{y})].$$

Ο πρώτος όρος, ακολουθεί την κατανομή χ^2_m, όπου m είναι ο αριθμός παραμέτρων του πλήρους μοντέλου. Ο δεύτερος όρος ακολουθεί την κατανομή χ^2_p όπου p είναι ο αριθμός παραμέτρων στο μοντέλο που μας ενδιαφέρει. Ο τρίτος όρος είναι μια θετική σταθερά η οποία θα είναι κοντά στο μηδέν, αν το μοντέλο που μας ενδιαφέρει περιγράφει τα δεδομένα σχεδόν τόσο καλά, όσο το πλήρες μοντέλο. Το πλήρες μοντέλο είναι ένα γενικευμένο γραμμικό μοντέλο που χρησιμοποιεί την ίδια κατανομή και έχει την ίδια συνάρτηση σύνδεσης με το μοντέλο που μας ενδιαφέρει. Ο αριθμός των παραμέτρων στο πλήρες μοντέλο είναι ίσος με τον αριθμό των παρατηρήσεων N. Όταν υπάρχει καλή προσαρμογή στο μοντέλο, ισχύει:

$$D \sim \chi^2_{m-p}.$$

Αν το μοντέλο είναι κατάλληλο, τότε θα πρέπει η τιμή του D να είναι κοντά στο μέσο της κατανομής. Αν λοιπόν ένα μοντέλο με p παραμέτρους περιγράφει καλά ένα σύνολο από N παρατηρήσεις, έτσι ώστε $D \sim \chi^2_{N-p}$ τότε θα πρέπει $D \simeq N - p$. Για κάποιες κατανομές όπως η Poisson, η τιμή του D μπορεί να υπολογιστεί απ'ευθείας από τις προσαρμοσμένες τιμές και να συγκριθεί με τους βαθμούς ελευθερίας για να εκτιμηθεί η καλή προσαρμογή.

Για να κάνουμε έλεγχο υποθέσεων, χρησιμοποιώντας την στατιστική συνάρτηση deviance, θα ορίσουμε ένα μοντέλο για κάθε υπόθεση και θα συγκρίνουμε τις στατιστικές συναρτήσεις της καλής προσαρμογής για τα συγκεκριμένα μοντέλα.

Τα μοντέλα που θα συγκρίνουμε πρέπει να έχουν την ίδια κατανομή και την ίδια συνάρτηση σύνδεσης. Έτσι θα έχουμε τη μηδενική υπόθεση H_0 και την εναλλακτική H_1, όπου:

$$H_0 : \mathbf{b}_0 = \begin{bmatrix} b_1 \\ . \\ . \\ b_q \end{bmatrix} \quad \text{vs} \quad H_1 : \mathbf{b}_1 = \begin{bmatrix} b_1 \\ . \\ . \\ b_p \end{bmatrix}.$$

Κάνουμε τον έλεγχο H_0 vs H_1 χρησιμοποιώντας τη διαφορά των συναρτήσεων deviance:

$$\Delta D = D_0 - D_1.$$

Η συνάρτηση deviance ορίζεται ως εξής:

$$D = 2[l(\mathbf{b}_{max}; \mathbf{y}) - l(\mathbf{b}; \mathbf{y})] \quad , \quad \text{οπότε:}$$

$$\Delta D = D_0 - D_1 = 2[l(\mathbf{b}_{max}; \mathbf{y}) - l(\mathbf{b}_0; \mathbf{y})] - 2[l(\mathbf{b}_{max}; \mathbf{y}) - l(\mathbf{b}_1; \mathbf{y})] = 2[l(\mathbf{b}_1; \mathbf{y}) - l(\mathbf{b}_0; \mathbf{y})].$$

Αν τα μοντέλα περιγράφουν καλά τα δεδομένα, τότε:

$$D_0 \sim \chi^2_{N-q} \quad \text{και} \quad D_1 \sim \chi^2_{N-p}.$$

Στην περίπτωση αυτή ισχύει, $\Delta D \sim \chi^2_{p-q}$. Αν η D ακολουθεί την κατανομή χ^2_{p-q}, τότε θα επιλέγουμε το μοντέλο H_0 γιατί είναι πιο απλό.

2.4.2 Υπόλοιπα

Τα υπόλοιπα είναι οι διαφορές των παρατηρήσεων y_i με τις τιμές των εκτιμημένων τιμών $\hat{y}_i$: $\hat{e}_i = y_i - \hat{y}_i$. Μπορούμε να χρησιμοποιήσουμε τα υπόλοιπα για να ελέγξουμε τα χαρακτηριστικά των μοντέλων. Τα υπόλοιπα μας δίνουν πληροφορίες για την καταλληλότητα του μοντέλου. Αυτό ισχύει γενικά. Στα γενικευμένα γραμμικά μοντέλα ορίζουμε τα εξής υπόλοιπα [7]:

• Υπόλοιπα deviance, ή υπόλοιπα deviance:

$$r_D = sign(y - \mu)\sqrt{d_i} \quad , \quad \sum_{i=1}^{N} r_D^2 = D.$$

Σε ένα γενικευμένο γραμμικό μοντέλο, κάθε μονάδα, έχει μια τιμή d_i, έτσι ώστε:

$$\sum_{i=1}^{N} d_i = D.$$

Για παράδειγμα, για την Poisson κατανομή έχουμε:

$$r_D = sign(y - \mu)[2(log(\tfrac{y}{\mu}) - y + \mu)]^2.$$

• Υπόλοιπα Pearson:

$$r_P = \frac{y_i - \mu_i}{\sqrt{var(y_i)}} \quad , \quad i = 1, ..., N.$$

Ένα μειονέκτημα των υπολοίπων Pearson είναι ότι η κατανομή του r_P για μη-Κανονικές κατανομές, είναι δύσχρηστα και δεν έχουν παρόμοιες ιδιότητες με αυτές των υπολοίπων της Κανονικής κατανομής.

• Υπόλοιπα Anscobe:

Ο Anscobe πρότεινε να οριστεί ένα υπόλοιπο χρησιμοποιώντας μια συνάρτηση $A(y)$, όπου η $A(\bullet)$, θα έκανε την κατανομή του $A(Y)$ όσο το δυνατόν πιο κοντά στην Κανονική. Η συνάρτηση $A(\bullet)$ δίνεται από:

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

Για την Poisson κατανομή: $\int \frac{d\mu}{\mu^{\frac{1}{3}}} = \frac{3}{2}\mu^{\frac{2}{3}}$.

Έτσι τα Anscobe υπόλοιπα, είναι:

$$r_A = \frac{\frac{3}{2}(y^{\frac{2}{3}} - \mu^{\frac{2}{3}})}{\mu^{\frac{1}{6}}}.$$

Για την Γάμμα κατανομή τα υπόλοιπα Anscobe παίρνουν τη μορφή:

$$r_A = \frac{3(y^{\frac{1}{3}} - \mu^{\frac{1}{3}})}{\mu^{\frac{1}{3}}}.$$

3 ΠΟΛΛΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

3.1 Εισαγωγή

Σε αυτό το κεφάλαιο, αναφερόμαστε σε μοντέλα της μορφής:

$$E(Y_i) = \mu_i = \mathbf{x}_i^T \mathbf{b} \quad , \quad Y_i \sim N(\mu_i, \sigma^2),$$

όπου οι $Y_1, ..., Y_N$ είναι ανεξάρτητες τυχαίες μεταβλητές. Οι μεταβλητές Y_i ακολουθούν την Κανονική κατανομή [2]. Τα γενικά γραμμικά μοντέλα, είναι μία περίπτωση των γενικευμένων γραμμικών μοντέλων.

Ένα μοντέλο που έχει τουλάχιστον δύο ανεξάρτητες επεξηγηματικές μεταβλητές ονομάζεται μοντέλο πολλαπλής γραμμικής παλινδρόμησης. Η πολλαπλή γραμμική παλινδρόμηση, γράφεται ισοδύναμα:

$$\mathbf{y} = \mathbf{Xb} + \boldsymbol{\varepsilon}.$$

Όπου:

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ . \\ . \\ Y_N \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ . \\ . \\ \mathbf{x}_N^T \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_1 \\ . \\ . \\ b_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ . \\ . \\ \varepsilon_N \end{bmatrix}.$$

Τα $\boldsymbol{\varepsilon}$ είναι ανεξάρτητα και ακολουθούν την Κανονική κατανομή $N(0, \sigma^2)$.

3.2 Γενικό γραμμικό μοντέλο

Ο όρος γενικό γραμμικό μοντέλο, χρησιμοποιείται για μοντέλα, των οποίων οι τιμές απόκρισης ακολουθούν την Κανονική κατανομή και μπορούμε να έχουμε οποιοδήποτε συνδυασμό κατηγορικών, ή συνεχών επεξηγηματικών μεταβλητών.

Ας θεωρήσουμε ότι έχουμε p επεξηγηματικές μεταβλητές $X = (X_1, \ldots, X_p)$ οι οποίες συνδέονται γραμμικά με την ποσοτική μεταβλητή απόκρισης Y.

Αυτή είναι η περίπτωση του γενικού γραμμικού μοντέλου:

$$Y = b_0 + b_1 X_1 + ... + b_p X_p + \varepsilon \, , \varepsilon \sim N(0, \sigma^2) \Rightarrow$$

$$E(Y|X_1, ..., X_p) = b_0 + b_1 X_1 + ... + b_p X_p.$$

• Η παράμετρος b_0 εκφράζει την μέση τιμή της τυχαίας μεταβλητής Y όταν όλα τα $X_j, j = 1, ..., p$ είναι μηδέν.

• Η παράμετρος $b_j, j = 1, ..., p$ εκφράζει την αναμενόμενη μεταβολή της τιμής της Y, όταν η X_j αυξηθεί κατά μία μονάδα και οι υπόλοιπες $X_k, k \neq j$, παραμείνουν σταθερές.

Οι προϋποθέσεις του γενικού γραμμικού μοντέλου είναι:
1.Γραμμικότητα
2.Κανονικότητα σφαλμάτων
3.Ομοσκεδαστικότητα
4.Ανεξαρτησία Σφαλμάτων

Το γενικό ή πολλαπλό γραμμικό μοντέλο παλινδρόμησης, μπορεί να γραφεί σε μορφή πινάκων ως:

$$\mathbf{Y} = \mathbf{Xb} + \boldsymbol{\varepsilon},$$

όπου:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_{11} & . & . & X_{1p} \\ . & X_{21} & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ 1 & X_{n1} & . & . & X_{np} \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ . \\ . \\ b_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{bmatrix}.$$

3.3 Εκτίμηση παραμέτρων

3.3.1 Μέθοδος Μέγιστης Πιθανοφάνειας

Ο εκτιμητής μέγιστης πιθανοφάνειας δίνεται από τη σχέση:

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},$$

υπό την προϋπόθεση ότι ο $\mathbf{X}^T\mathbf{X}$ δεν είναι ο μοναδιαίος πίνακας. Για τα μοντέλα πολλαπλής παλινδρόμησης, ο πίνακας των σταθερών $\mathbf{X}$ θα πρέπει να έχει ανεξάρτητες στήλες, ώστε ο πίνακας $\mathbf{X}^T\mathbf{X}$ να μην είναι μοναδιαίος. Καθώς $E(\hat{\mathbf{b}}) = \mathbf{b}$, ο εκτιμητής είναι αμερόληπτος και έχει πίνακα διακύμανσης - συνδιακύμανσης:

$$\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1} = \mathcal{I}^{-1}.$$

Μπορεί να δειχθεί ότι:

$$\hat{\sigma}^2 = \tfrac{1}{N-p}(\mathbf{y} - \mathbf{Xb})^T(\mathbf{y} - \mathbf{Xb}),$$

είναι αμερόληπτος εκτιμητής του σ^2 και μπορεί να χρησιμοποιηθεί για να εκτιμήσουμε το $\mathcal{I}$ κι έτσι να προβούμε σε συμπερασματολογία για το $\hat{\mathbf{b}}$.

3.3.2 Μέθοδος Ελαχίστων Τετραγώνων

Με τη μέθοδο ελαχίστων τετραγώνων οι εκτιμήτριες είναι:

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Επίσης,

$$S^2_{y|X_1,\ldots,X_p} = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Οι οποίες είναι τυχαίες μεταβλητές. Οπότε καταλήγουμε στο μοντέλο:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{b}}.$$

Το $\hat{\mathbf{Y}}$ καλείται προβλεπόμενη τιμή και είναι η αναμενόμενη που θα πάρει η τυχαία μεταβλητή **Y**, όταν **X=x**.

Αντίστοιχα, όπως και στο απλό γραμμικό μοντέλο ενδιαφερόμαστε για τους ελέγχους υποθέσεων σχετικά με τα b_0 και b_j, $j = 1, \ldots p$. Επίσης ενδιαφέρον έχει και το F-test το οποίο ελέγχει τη μηδενική υπόθεση $H_0 = b_1 = b_2, ..., b_p = 0$ με εναλλακτική ότι τουλάχιστον ένα από τα b_j δεν είναι 0.

3.4 Deviance για το πολλαπλό γραμμικό μοντέλο

Η deviance, υπολογίζεται ως:

$$D = 2[l(b_{max}; y) - l(b; y)].$$

Οπότε:

$$\begin{aligned} D &= \tfrac{1}{\sigma^2}(\mathbf{y} - \mathbf{Xb})^T(\mathbf{y} - \mathbf{Xb}) \\ &= \tfrac{1}{\sigma^2}(\mathbf{y}^T\mathbf{y} - 2\mathbf{b}^T\mathbf{X}^T\mathbf{y} + \mathbf{b}^T\mathbf{X}^T\mathbf{Xb}) \\ &= \tfrac{1}{\sigma^2}(\mathbf{y}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{y}), \end{aligned}$$

επειδή $\mathbf{X}^T\mathbf{Xb} = \mathbf{X}^T\mathbf{y}$ αφού, $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.

Αν το μοντέλο είναι σωστό, τότε η deviance θα ακολουθεί την χ-τετράγωνο κατανομή με N-p βαθμούς ελευθερίας.

3.5 Έλεγχος Υποθέσεων

Ας υποθέσουμε την μηδενική υπόθεση H_0 και την υπόθεση H_1 ως εξής:

$$H_0 : b_1 = 0 \quad \text{και} \quad H_1 : b_1 \neq 0,$$

$$H_0 : b_2 = 0 \quad \text{και} \quad H_1 : b_2 \neq 0,$$

$$\begin{array}{ccc} . & . & . \\ . & . & . \\ . & . & . \end{array}$$

$$H_0 : b_p = 0 \quad \text{και} \quad H_1 : b_p \neq 0,$$

Έστω ότι τα $\mathbf{X}_0$ και $\mathbf{X}_1$ συμβολίζουν τους πίνακες σχεδιασμού, $\hat{\mathbf{b}}_0$ και $\hat{\mathbf{b}}_1$ τους εκτιμητές μέγιστης πιθανοφάνειας και $\mathbf{D}_0$ και $\mathbf{D}_1$ τις αποκλίσεις.

Κάνουμε τον έλεγχο H_0 vs H_1 χρησιμοποιώντας:

$$\begin{aligned} \Delta D = D_o - D_1 &= \tfrac{1}{\sigma^2}(\mathbf{y}^T\mathbf{y} - \mathbf{b}_0^T\mathbf{X}_0^T\mathbf{y}) - \tfrac{1}{\sigma^2}(\mathbf{y}^T\mathbf{y} - \mathbf{b}_1^T\mathbf{X}_1^T\mathbf{y}) \\ &= \tfrac{1}{\sigma^2}(\mathbf{b}_1^T\mathbf{X}_1^T\mathbf{y} - \mathbf{b}_0^T\mathbf{X}_0^T\mathbf{y}). \end{aligned}$$

Αν το μοντέλο μας αντιστοιχεί στην H_1, είναι πιθανό να προσαρμόζεται στα δεδομένα καλα, έτσι υποθέτουμε ότι η D_1 ακολουθεί την κεντρική χ-τετράγωνο κατανομή με N-p βαθμούς ελευθερίας. Από την άλλη, η D_0 μπορεί να ακολουθεί την μη κεντρική χ-τετράγωνο κατανομή με N-q βαθμούς ελευθερίας αν η H_0 δεν είναι σωστή.

Όταν το σ^2 είναι άγνωστο, χρησιμοποιούμε το πηλίκο:

$$F = \frac{\frac{D_0 - D_1}{p-q}}{\frac{D_1}{N-p}} = \frac{\frac{\sum[y_i - \hat{\mu}_i(0)]^2 - \sum[y_i - \hat{\mu}_i(1)]^2}{p-q}}{\frac{\sum[y_i - \hat{\mu}_i(1)]^2}{N-p}}.$$

Το F το υπολογίζουμε από τις τιμές των δεδομένων.

Ο έλεγχος:

$$F = \tfrac{D_0 - D_1}{p-q} \Big/ \tfrac{D_1}{N-p} = \frac{\mathbf{b}_0^T\mathbf{X}_0^T\mathbf{y} - \mathbf{b}_1^T\mathbf{X}_1^T\mathbf{y}}{p-q} \Big/ \frac{\mathbf{y}^T\mathbf{y} - \mathbf{b}_1^T\mathbf{X}_1^T\mathbf{y}}{N-p},$$

θα ακολουθεί την κεντρική κατανομή $F(p-q, N-p)$, αν η μηδενική υπόθεση είναι σωστή ή η F σε άλλη περίπτωση θα έχει μη κεντρική κατανομή. Οπότε, τιμές για το F που είναι μεγάλες, συγκριτικά με την κατανομή $F(p-q, N-p)$, μας δίνουν στοιχεία για να απορρίψουμε τη μηδενική υπόθεση.

Παρουσιάζουμε αυτόν τον έλεγχο υποθέσεων μέσω του πίνακα ανάλυσης διασποράς, όπως φαίνεται στον πίνακα 2 [2].

Πηγή διακύμανσης	Βαθμοί ελευθερίας	Άθροισμα τετραγώνων τετραγώνων	Μέσο άθροισμα τετραγώνων
Μοντέλο με b_0	q	$\mathbf{b}_0^T\mathbf{X}_0^T\mathbf{y}$	
Βελτίωση από μοντέλο με b_1	p-q	$\mathbf{b}_1^T\mathbf{X}_1^T\mathbf{y} - \mathbf{b}_0^T\mathbf{X}_0^T\mathbf{y}$	$\frac{\mathbf{b}_1^T\mathbf{X}_1^T\mathbf{y} - \mathbf{b}_0^T\mathbf{X}_0^T\mathbf{y}}{p-q}$
Υπόλοιπα	N-p	$\mathbf{y}^T\mathbf{y} - \mathbf{b}_1^T\mathbf{X}_1^T\mathbf{y}$	$\frac{\mathbf{y}^T\mathbf{y} - \mathbf{b}_1^T\mathbf{X}_1^T\mathbf{y}}{N-p}$
Σύνολο	N	$\mathbf{y}^T\mathbf{y}$	

Πίνακας 2: *Πίνακας ανάλυσης διασποράς.*

3.6 Συντελεστής προσδιορισμού, R^2

Ένα συνηθισμένο μέτρο για την καλή προσαρμογή ενός μοντέλου πολλαπλής παλινδρόμησης, είναι η σύγκριση με ένα πιο απλό μοντέλο χρησιμοποιώντας τη μέθοδο ελαχίστων τετραγώνων.

Για την σχέση $\mathbf{y} = \mathbf{Xb} + \boldsymbol{\varepsilon}$ όπου τα $\boldsymbol{\varepsilon}$ είναι ανεξάρτητα,η μέση τιμή είναι $E(\varepsilon_i) = 0$ και η διασπορά είναι $Var(\varepsilon_i) = \sigma^2$ για όλα τα i.

Το κριτήριο των ελαχίστων τετραγώνων είναι:

$$S = \sum_{i=1}^{N} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{Xb})^T(\mathbf{Y} - \mathbf{Xb}).$$

Όπως έχουμε δει, η εκτίμηση ελαχίστων τετραγώνων είναι:

$$\hat{\mathbf{b}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Έτσι, η ελάχιστη τιμή του S είναι:

$$\hat{S} = (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}) = \mathbf{y}^T\mathbf{y} - \mathbf{b}^T\mathbf{X}^T\mathbf{y}.$$

Το ελάχιστο μοντέλο είναι:

$$E(Y_i) = \mu.$$

Για το ελάχιστο μοντέλο είναι: $E(y) = \mathbf{Xb}$ με: $\mathbf{b} = [\mu]$, $\mathbf{X} = 1$.

Οπότε: $\mathbf{X}^T\mathbf{X} = N$ και $\mathbf{X}^T\mathbf{y} = \sum y_i$, έτσι: $\mathbf{b} = \hat{\mu} = \bar{y}$.

Στην περίπτωση αυτή,η τιμή του S είναι:

$$\hat{S}_0 = \mathbf{y}^T\mathbf{y} - N\bar{y}^2 = \sum(y_i - \bar{y})^2.$$

Έτσι, το $\hat{S}_0$ είναι ανάλογο με τη διακύμανση των παρατηρήσεων και είναι η μεγαλύτερη, ή "χειρότερη δυνατή" τιμή του S.

Ο συντελεστής προσδιορισμού R^2 μπορεί να εξηγηθεί ως το ποσοστό της συνολικής μεταβλητότητας στα δεδομένα που εξηγείται από το μοντέλο. Ο συντελεστής προσδιορισμού δηλαδή, εκφράζει το πόσο καλά σχετίζονται οι επεξηγηματικές μεταβλητές με την απόκριση και ορίζεται ως:

$$R^2 = \frac{\hat{S}_0 - \hat{S}}{\hat{S}_0} = \frac{\mathbf{b}^T\mathbf{X}^T\mathbf{y} - N\bar{y}^2}{\mathbf{y}^T\mathbf{y} - N\bar{y}^2}.$$

Οι τιμές του συντελεστή προσδιορισμού είναι από 0 έως 1. Η τετραγωνική ρίζα του R^2 ονομάζεται συντελεστής πολλαπλής συνδιακύμανσης.

Παρά το γεγονός ότι είναι πολύ διαδεδομένος για τον έλεγχο της προσαρμογής του μοντέλου, ο συντελεστής προσδιορισμού έχει κάποιους περιορισμούς.Δεν είναι εύκολος ο προσδιορισμός της δειγματικής κατανομής του R^2. Επίσης, η τιμή του δεν μεταβάλλεται με την αλλαγή του αριθμού των παραμέτρων που χρησιμοποιούνται στο προσαρμοσμένο μοντέλο.

3.7 Υπόλοιπα

Για να ελέγξουμε την προσαρμογή του μοντέλου, μπορούμε να μελετήσουμε τα υπόλοιπα. Δημιουργώντας γραφήματα υπολοίπων, ελέγχουμε αν το μοντέλο που καθορίσαμε, είναι σωστό.

Τα υπόλοιπα ε_i ακολουθούν την Κανονική κατανομή με μέσο 0 και διακύμανση σ^2 και είναι ανεξάρτητα.Τα υπόλοιπα ορίζονται ως εξής:

$$\hat{\varepsilon_i} = y_i - x_i^T\hat{b}.$$

Με το γράφημα υπολοίπων, μπορούμε να ελέγξουμε την κανονικότητα. Δημιουργούμε γραφήματα των τυποποιημένων υπολοίπων, με τις προσαρμοσμένες τιμές και με κάθε μία από τις επεξηγηματικές μεταβλητές.

4 ΔΙΤΙΜΕΣ ΜΕΤΑΒΛΗΤΕΣ ΚΑΙ ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

4.1 Δίτιμες μεταβλητές

Θεωρούμε ένα πείραμα στο οποίο η απόκριση Y μπορεί να πάρει μόνο μία από δύο δυνατές τιμές. Για παράδειγμα, οι δύο κατηγορίες είναι επιτυχία ή αποτυχία, ή παρών-απών. Συμβολίζουμε την επιτυχία με 1 και την αποτυχία με 0. Τότε η Y ακολουθεί την κατανομή Bernoulli με παράμετρο π.

Η Y τότε είναι δίτιμη μεταβλητή, δηλαδή:

$$Y = \begin{cases} 1 & \text{αν το αποτέλεσμα είναι επιτυχία} \\ 0 & \text{αν το αποτέλεσμα είναι αποτυχία} \end{cases} .$$

Αν π είναι η πιθανότητα επιτυχίας, τότε :

$$P(Y = 1) = \pi \text{ και } P(Y = 0) = 1 - \pi.$$

Η συνάρτηση μάζας πιθανότητας της τυχαίας μεταβλητής Y είναι:

$$f(y; \pi) = P(Y = y) = \pi^y (1 - \pi)^{1-y}, y \epsilon \{0, 1\}.$$

Αν έχουμε $Y_i, i = 1, ..., n$ τυχαίες δίτιμες μεταβλητές, όπου είναι $Y_i \sim Bernoulli(\pi_i)$, τότε η από κοινού συνάρτηση μάζας πιθανότητας των Y_i είναι:

$$f(\mathbf{y}; \boldsymbol{\pi}) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i} = exp[\sum_{i=1}^{n} y_i log(\frac{\pi_i}{1 - \pi_i}) + \sum_{i=1}^{n} log(1 - \pi_i)], \quad (1)$$

όπου $\boldsymbol{\pi} = [\pi_1, ..., \pi_n]^T$ και $\mathbf{y} = [y_1, ..., y_n]^T$.

Αν τα π_j είναι όλα ίσα μεταξύ τους, μπορούμε να ορίσουμε τη συνάρτηση $Y = \sum_{i=1}^{n} Y_i$, όπου το Y παριστάνει τον αριθμό των επιτυχιών σε n ανεξάρτητες προσπάθειες και ακολουθεί τη διωνυμική κατανομή, με παραμέτρους n, π και συνάρτηση μάζας πιθανότητας:

$$P(Y = y) = \begin{pmatrix} n \\ y \end{pmatrix} \pi^y (1 - \pi)^{n-y} \quad , \quad \text{όπου } y = 0, 1, ..., n.$$

Αν $Y_i \sim b(n_i, \pi_i)$ τότε η λογαριθμική συνάρτηση πιθανοφάνειας είναι:

$$l(\pi_1, ..., \pi_N; y_1, ..., y_N) = \sum_{i=1}^{N} [y_i log(\frac{\pi_i}{1 - \pi_i}) + n_i log(1 - \pi_i) + log \begin{pmatrix} n_i \\ y_i \end{pmatrix}]. \quad (2)$$

4.2 Γενικευμένα Γραμμικά Μοντέλα με Δίτιμες Μεταβλητές

Έστω N ανεξάρτητες μεταβλητές Y_i για τις οποίες ισχύει: $Y_i \sim b(n_i, \pi_i)$. Θέλουμε να μελετήσουμε τη σχέση μεταξύ της πιθανότητας επιτυχιών και της επεξηγηματικής μεταβλητής [7].

Μοντελοποιούμε τις πιθανότητες π_i, ως:

$$g(\pi_i) = \eta_i = \mathbf{x}_i^T \mathbf{b},$$

όπου $\mathbf{x}_i$ το διάνυσμα επεξηγηματικών μεταβλητών, το **b** το διάνυσμα των παραμέτρων και g η συνάρτηση σύνδεσης. Οι τιμές του π_i ανήκουν στο διάστημα $[0, 1]$ επειδή είναι πιθανότητες. Αυτό μας περιορίζει σχετικά με τα $\mathbf{x}_i$ και **b**.

Για να αποφύγουμε αυτό το πρόβλημα που προκύπτει από τη μοντελοποίηση του **π**, επιλέγουμε κατάλληλη τη συνάρτηση σύνδεσης για να παίρνουμε τιμές για εκτίμηση του π μέσα στο διάστημα $[0, 1]$. Για τη διωνυμική κατανομή, επιλέγουμε μία από τις επόμενες συναρτήσεις σύνδεσης:

• logit: $g(\pi_i) = log(\frac{\pi_i}{1-\pi_i})$,

οπότε θα είναι: $\pi_i = \frac{e^{n_i}}{1+e^{n_i}}$.

Το μοντέλο που έχει συνάρτηση σύνδεσης τη logit, χρησιμοποιείται πολύ σε δίτιμα δεδομένα.

• probit: $g(\pi_i) = \Phi^{-1}(\pi_i)$, όπου με Φ συμβολίζεται η συνάρτηση κατανομής της Κανονικής κατανομής και Φ^{-1} είναι η αντίστροφή της,

οπότε $\pi_i = \Phi(n_i)$.

• complementary log-log ή cloglog: $g(\pi_i) = log(-log(1 - \pi_i))$,

οπότε: $\pi_i = 1 - exp(-e^{n_i})$.

Χρησιμοποιούμε την cloglog συνάρτηση σύνδεσης, κυρίως σε περιπτώσεις όπου η πιθανότητα να συμβεί ένα γεγονός είναι πολύ μικρή ή πολύ μεγάλη.

Αυτές οι τρείς συναρτήσεις είναι οι πιο διαδεδομένες που χρησιμοποιούνται στην πράξη. Υπάρχουν αρκετές άλλες συναρτήσεις σύνδεσης $g(\pi)$. Οι συναρτήσεις που αναφέραμε είναι όλες συνεχείς και αύξουσες στο διάστημα (0,1).

Μία πολύ σημαντική χρήση των μοντέλων για διωνυμικά δεδομένα ήταν η μελέτη αποτελεσμάτων σε πειράματα όπως αριθμός πειραματόζωων που σκοτώθηκαν από μία τοξική ουσία.

Τα Probit μοντέλα χρησιμοποιούνται σε διάφορους τομείς βιολογικών και κοινωνικών επιστημών, όπου υπάρχει φυσική ερμηνεία του μοντέλου. Τα Probit μοντέλα, ήταν από τα αρχικά μοντέλα που χρησιμοποιήθηκαν γι' αυτές τις αναλύσεις.

Έχουμε λοιπόν:

$$\pi = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} exp[-\frac{1}{2}(\frac{s-\mu}{\sigma})^2]ds = \Phi(\frac{x-\mu}{\sigma}),$$

όπου το Φ δηλώνει την συνάρτηση κατανομής της κανονικής κατανομής N(0,1).

Οπότε:

$$\Phi^{-1}(\pi) = b_1 + b_2 x,$$

όπου $b_1 = -\frac{\mu}{\sigma}$ και $b_2 = \frac{1}{\sigma}$ και η συνάρτηση σύνδεσης g είναι η αντίστροφη συνάρτηση της αθροιστικής κατανομής Φ^{-1}.

Για παράδειγμα $x = \mu$ ονομάζεται η μέση θανατηφόρα δόση, επειδή αναφέρεται στη δόση μιας ουσίας που χρειάζεται για να σκοτώσει τα μισά ζώα ενός πειράματος.

4.3 Λογιστική παλινδρόμηση

Σε κάποιες περιπτώσεις, η μεταβλητή απόκρισης είναι διακριτή. Μπορεί να μην ακολουθεί την Κανονική κατανομή και μπορεί να παίρνει περισσότερες από δύο τιμές.

Σε τέτοιες περιπτώσεις θέλουμε να βρούμε το καλύτερο μοντέλο που περιγράφει τη σχέση μεταξύ μεταβλητής απόκρισης και επεξηγηματικών μεταβλητών. Μία μέθοδος για τέτοια στατιστική ανάλυση είναι η Λογιστική Παλινδρόμηση [2].

Στη Λογιστική Παλινδρόμηση η μεταβλητή απόκρισης ακολουθεί την κατανομή Bernoulli ή την διωνυμική κατανομή.

Σε περίπτωση όπου η μεταβλητή απόκρισης είναι μία τυχαία μεταβλητή από την κατανομή Bernoulli έχουμε:

$$E(y_i) = \pi_i = P(x_i),$$

και

$$Var(y_i) = \pi_i(1 - \pi_i).$$

P_i είναι η πιθανότητα στη Bernoulli.

Σύμφωνα με το μοντέλο $Y_i = b_0 + b_1 x_i + \varepsilon_i$, έχουμε:

$$E(Y_i) = b_0 + b_1 X_i,$$

άρα:

$$b_0 + b_j X_i = \pi_i.$$

Επίσης,

$$P(Y_i = 1) = \pi_i \quad , \quad P(Y_i = 0) = 1 - \pi_i.$$

Άρα η μέση τιμή $E(Y_i) = b_0 + b_1 X_i$ είναι η πιθανότητα ότι η $Y_i = 1$.

Αν υποθέσουμε ότι $y = 1$ επιτυχία και $y = 0$ αποτυχία με πιθανότητα π και $1 - \pi$ αντίστοιχα, τότε έχουμε:

$$E(Y_i) = \frac{exp(b_0 + b_1 X_i)}{1 + exp(b_0 + b_1 X_i)}.$$

Σε περίπτωση που οι X_i παρατηρήσεις είναι τυχαίες, η $E(Y_i)$ είναι υποθετική μέση τιμή.

Για να εκτιμήσουμε τις παραμέτρους, θα χρησιμοποιήσουμε τη μέθοδο μέγιστης πιθανοφάνειας. Η συνάρτηση πιθανοφάνειας είναι:

$$L(b) = \prod_{i=1}^{n} \begin{pmatrix} n_i \\ y_i \end{pmatrix} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i},$$

όπου το π_i είναι:

$$\pi_i = \frac{e^{b_0 + b_1 X_{1i} + \ldots + b_k X_{ki}}}{1 + e^{b_0 + b_1 X_{1i} + \ldots + b_k X_{ki}}}.$$

Αυτό που ψάχνουμε, είναι να βρούμε τα $\hat{b}_0, \hat{b}_1, ..., \hat{b}_k$ που μεγιστοποιούν το $L(b)$.

4.3.1 Γενικό λογιστικό μοντέλο

Το γενικό λογιστικό μοντέλο, έχει τη μορφή:

$$logit\pi_i = log(\tfrac{\pi_i}{1-\pi_i}) = \mathbf{x}_i^T\mathbf{b},$$

όπου το $\mathbf{x}_i$ είναι ένα διάνυσμα με συνεχείς τιμές, που αντιστοιχούν σε συμμεταβλητές και εικονικές μεταβλητές. Το **b**, είναι το διάνυσμα παραμέτρων.

Αυτό το μοντέλο, είναι αρκετά διαδεδομένο για ανάλυση δεδομένων όπου περιέχονται δίτιμες και διωνυμικές αποκρίσεις και διάφορες επεξηγηματικές μεταβλητές. Μας εφοδιάζει με σημαντική τεχνική, ανάλογη της πολλαπλής παλινδρόμησης.

Οι εκτιμήσεις μέγιστης πιθανοφάνειας των παραμέτρων **b**, οπότε και των πιθανοτήτων: $\pi_i = g(\mathbf{x}_i^T\mathbf{b})$ προκύπτουν από τη μεγιστοποίηση της λογαριθμικής συνάρτησης πιθανοφάνειας:

$$l(\pi; y) = \textstyle\sum_{i-1}^{N}\{y_i log(\pi_i) + (n_i - y_i)log(1-\pi_i) + log\begin{pmatrix} n_i \\ y_i \end{pmatrix}\}.$$

4.3.2 Κριτήρια καλής προσαρμογής

Αντί να χρησιμοποιήσουμε τη μέθοδο εκτίμησης μέγιστης πιθανοφάνειας, θα μπορούσαμε να εκτιμήσουμε τις παραμέτρους, ελαχιστοποιώντας το σταθμισμένων άθροισμα τετραγώνων.

$$S_w = \textstyle\sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)}.$$

Καθώς $E(Y_i) = n_i\pi_i$ και $Var(Y_i) = n_i\pi_i(1-\pi_i)$.

Αυτό είναι ισοδύναμο με το να ελαχιστοποιήσουμε τη στατιστική συνάρτηση χ-τετράγωνο του Pearson,

$$X^2 = \textstyle\sum \frac{(o_i - e_i)^2}{e_i},$$

όπου o είναι οι συχνότητες που παρατηρούμε και e οι αναμενόμενες συχνότητες.

$$X^2 = \textstyle\sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i} + \sum_{i=1}^{N} \frac{[(n_i - y_i) - n_i(1-\pi_i)]^2}{n_i(1-\pi_i)}$$

$$= \textstyle\sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1-\pi_i)}(1 - \pi_i + \pi_i) = S_w.$$

Όταν το X^2 παίρνει τιμές στις εκτιμημένες αναμενόμενες συχνότητες, η στατιστική συνάρτηση είναι:

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)},$$

η οποία είναι ασυμπτωτικά ισοδύναμη με την deviance:

$$D = 2\sum_{i=1}^{N}\{y_i log\left(\begin{array}{c} y_i \\ \hat{y}_i \end{array}\right) + (n_i - y_i)log(\tfrac{n_i - y_i}{n_i - \hat{y}_i})\}.$$

Η απόδειξη της σχέσης μεταξύ X^2 και D χρησιμοποιεί το ανάπτυγμα της σειράς του Taylor του $s * log(s/t)$ για $s = t$.

$$s * log(\tfrac{s}{t}) = (s - t) + \tfrac{1}{2}\tfrac{(s-t)^2}{t} + ... \quad ,$$

έτσι έχουμε,

$$D = 2\sum_{i=1}^{N}\{(y_i - n_i\hat{\pi}_i) + \tfrac{1}{2}\tfrac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + (n_i - y_i - n_i - n_i\hat{\pi}_i) + \tfrac{1}{2}\tfrac{(n_i - y_i - n_i - n_i\hat{\pi}_i)^2}{n_i - n_i\hat{\pi}_i} + ...\},$$

$$D \simeq \sum_{i=1}^{N} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)} = X^2.$$

Αυτά ισχύουν υπό την προϋπόθεση ότι το μοντέλο είναι σωστό. Είναι $D \sim \chi^2_{(N-p)}$ και προσεγγιστικά, $X^2 \sim \chi^2_{(N-p)}$.

Επιλέγουμε μεταξύ των D, X^2 σύμφωνα με την επάρκεια προσέγγισης στην κατανομή χ^2_{N-p}. Υπάρχουν κάποια στοιχεία για να προτείνουμε ότι η X^2 είναι καλύτερη από την D, επειδή η D επηρεάζεται υπερβολικά από πολύ μικρές συχνότητες. Ωστόσο και οι δύο προσεγγίσεις, είναι αρκετά φτωχές αν οι αναμενόμενες συχνότητες είναι πολύ μικρές, για παράδειγμα, κάτω από το 1.

4.3.3 Στατιστική συνάρτηση deviance

Για να ελέγξουμε την καλή προσαρμογή του μοντέλου, χρησιμοποιούμε την στατιστική συνάρτηση deviance:

$$D = 2[l(\hat{\pi}_{max}; y) - l(\hat{\pi}; y)].$$

Αν οι μεταβλητές απόκρισης $Y_1, ...Y_N$ είναι ανεξάρτητες και ακολουθούν τη διωνυμική κατανομή,τότε, η λογαριθμική συνάρτηση πιθανοφάνειας, είναι:

$$l(\mathbf{b}, \mathbf{y}) = \sum_{i=1}^{T}\{y_i log\pi_i - y_i log(1 - \pi_i) + n_i log(1 - \pi_i) + log\left(\begin{array}{c} n_i \\ y_i \end{array}\right)\}.$$

Τα π_i είναι όλα διαφορετικά. Έτσι, $\mathbf{b} = [\pi_i, ..., \pi_N]^T$. Οι εκτιμητές μέγιστης πιθανοφάνειας είναι: $\hat{\pi}_i = y_i/n_i$, ώστε η μέγιστη τιμή της λογαριθμικής συνάρτησης πιθανοφάνειας, είναι:

$$l(\mathbf{b}_{max};\mathbf{y}) = \sum[y_i log\left(\begin{array}{c} y_i \\ n_i \end{array}\right)] - y_i log(\tfrac{n_i - y_i}{n_i}) + n_i log(\tfrac{n_i - y_i}{n_i}) + log\left(\begin{array}{c} n_i \\ y_i \end{array}\right).$$

Για οποιοδήποτε άλλο μοντέλο με $p < N$ παραμέτρους, ας υποθέσουμε ότι $\hat{\pi}_i$ δηλώνει την εκτιμήτρια μέγιστης πιθανοφάνειας για τις πιθανότητες και $\hat{y}_i = n_i\hat{\pi}_i$ δηλώνει τις προσαρμοσμένες τιμές. Τότε, η λογαριθμική συνάρτηση πιθανοφάνειας, για αυτές τις τιμές, είναι:

$$l(\mathbf{b};\mathbf{y}) = \sum[y_i log\left(\begin{array}{c} \hat{y}_i \\ n_i \end{array}\right)] - y_i log(\tfrac{n_i - \hat{y}_i}{n_i}) + n_i log(\tfrac{n_i - \hat{y}_i}{n_i}) + log\left(\begin{array}{c} n_i \\ y_i \end{array}\right),$$

οπότε, η deviance είναι:

$$\begin{aligned} D &= 2[l(\mathbf{b}_{max};\mathbf{y}) - l(\mathbf{b};\mathbf{y})] \\ &= 2\sum_{i=1}^{N}\{y_i log\left(\begin{array}{c} y_i \\ \hat{y}_i \end{array}\right) + (n_i - y_i) log(\tfrac{n_i - y_i}{n_i - \hat{y}_i})\}, \end{aligned}$$

και θα έχει τη μορφή:

$$D = 2\sum olog\tfrac{o}{e},$$

όπου το o, παριστάνει τις συχνότητες που παρατηρούμε y_i και $(n_i - y_i)$ και το e δείχνει τις αναμενόμενες συχνότητες που παρατηρούμε, $n_i\hat{\pi}_i$ και $(n_i - n_i\hat{\pi}_i)$.

Το D δεν περιέχει το σ^2, σε αντίθεση με την περίπτωση όπου η απόκριση ακολουθεί Κανονική κατανομή, οπότε, μπορεί να χρησιμοποιηθεί κατ' ευθείαν για τον έλεγχο καλής προσαρμογής, όπως και για ελέγχους υποθέσεων, με την προσέγγιση:

$$D \sim \chi^2_{N-p},$$

όπου p ο αριθμός των παραμέτρων που εκτιμήθηκαν και N ο αριθμός συμμεταβλητών.

4.3.4 Στατιστική συνάρτηση score

Έστω ότι η μεταβλητή απόκρισης ακολουθεί τη διωνυμική κατανομή B(n,b). Η b είναι η παράμετρος της κατανομής.

Μία αριθμητική μέθοδος υπολογισμού εκτιμητριών, είναι η μέθοδος των score. Η λογαριθμική συνάρτηση πιθανοφάνειας για τη διωνυμική κατανομή είναι:

$$l = ylog(b) + (n - y)log(1 - b) + log\left(\begin{array}{c} n \\ y \end{array}\right).$$

Η συνάρτηση score είναι:

$$U = \frac{y-nb}{b(1-b)}.$$

Αυτό προκύπτει επειδή:

$$U = \frac{\partial l}{\partial b},$$

αλλά,

$$\frac{\partial l}{\partial b} = \frac{y}{b} - \frac{n-y}{1-b} = \frac{y-nb}{b(1-b)}.$$

Ισχύει ότι:

$$E(U) = 0,$$

και

$$Var(U) = \frac{n}{b(1-b)}.$$

Η πληροφορία είναι:

$$\mathcal{I} = Var(U) = \frac{n}{b(1-b)},$$

οπότε προκύπτει ότι:

$$U^T \mathcal{I} U = \frac{(y-nb)^2}{nb(1-b)}.$$

4.4 Υπόλοιπα

Για τη λογιστική παλινδρόμηση υπάρχουν δύο βασικοί τύποι υπολοίπων που αντιστοιχούν στον έλεγχο καλής προσαρμογής του μοντέλου. Αν υπάρχουν m διαφορετικές παρατηρήσεις, τότε μπορούν να εκτιμηθούν m υπόλοιπα. Με Y_k συμβολίσουμε τον αριθμό επιτυχιών, n_k τον αριθμό επαναλήψεων και $\hat{\pi}_k$ το ποσοστό που εκτιμήθηκε.

4.4.1 Υπόλοιπα Pearson

Τα υπόλοιπα Pearson, ή χ-τετράγωνο είναι:

$$X_k = \frac{(y_k - n_k\hat{\pi}_k)}{\sqrt{n_k\hat{\pi}_k}(1-\hat{\pi}_k)},\ k = 1, ..., m.$$

Είναι:

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

Από την προηγούμενη σχέση, $\sum_{i=1}^{m} X_k^2 = X^2$, ο Pearson χ-τετράγωνο έλεγχος προσαρμογής.
Τα υπόλοιπα Pearson είναι:

$$r_{Pk} = \frac{X_k}{\sqrt{1-h_k}},$$

όπου h_k είναι η μόχλευση, που λαμβάνεται από τον πίνακα hat matrix. Ο πίνακας hat matrix είναι ο πίνακας $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Η τιμή h_{ii}, το i-στο στοιχείο της διαγωνίου του πίνακα hat matrix ονομάζεται μόχλευση (leverage) της i-στης παρατήρησης. Μία παρατήρηση με μεγάλη μόχλευση μπορεί να δημιουργήσει σημαντική διαφορά στην προσαρμογή του μοντέλου.

4.4.2 Υπόλοιπα deviance

Με παρόμοιο τρόπο ορίζουμε τα υπόλοιπα deviance.

$$d_k = sign(y_k - n_k\hat{\pi}_k)\{2[y_k log(\frac{y_k}{n_k\hat{\pi}_k}) + (n_k - y_k)log(\frac{n_k - y_k}{n_k - n_k\hat{\pi}_k})]\}^{1/2},$$

όπου ο όρος $sign(y_k - n_k\hat{\pi}_k)$ μας εξασφαλίζει ότι το d_k έχει το ίδιο πρόσημο με X_k.

Όπως έχουμε γράψει στα προηγούμενα, η deviance είναι:

$$\begin{aligned} D &= 2[l(\mathbf{b}_{max}; \mathbf{y}) - l(\mathbf{b}; \mathbf{y})] \\ &= 2\sum_{i=1}^{N}[y_i log\left(\begin{array}{c} y_i \\ \hat{y}_i \end{array}\right) + (n_i - y_i)log(\tfrac{n_i - y_i}{n_i - \hat{y}_i})]. \end{aligned}$$

Από αυτή τη σχέση, έχουμε την deviance, $\sum_{k=1}^{m} d_k^2 = D$.

Επίσης, τα τυποποιημένα υπόλοιπα deviance ορίζονται ως:

$$r_{Dk} = \frac{d_k}{\sqrt{1-h_k}}.$$

Τα υπόλοιπα αυτά, μπορούμε να τα χρησιμοποιήσουμε για να ελέγξουμε την καταλληλότητα ενός μοντέλου. Για παράδειγμα, θα έπρεπε να σχεδιαστούν ως προς κάθε επεξηγηματική μεταβλητή του μοντέλου για να ελέγξουμε αν η υπόθεση της γραμμικότητας ισχύει και να σχεδιαστούν ως προς άλλες πιθανές επεξηγηματικές μεταβλητές που δεν ανήκουν στο μοντέλο μας. Τα τυποποιημένα υπόλοιπα ακολουθούν την Κανονική κατανομή με μέσο 0 και τυπική απόκλιση 1.

5 POISSON ΠΑΛΙΝΔΡΟΜΗΣΗ

5.1 Κατανομή Poisson

Η συνάρτηση πιθανότητας της κατανομής Poisson είναι:

$$f(y) = \frac{e^{-\mu}\mu^y}{y!}.$$

Η παράμετρος μ είναι θετική. Η μέση τιμή είναι:

$$E(y) = \mu,$$

και η διασπορά είναι:

$$Var(y) = \mu.$$

5.2 Poisson παλινδρόμηση

Στην Poisson Παλινδρόμηση οι μεταβλητές απόκρισης είναι τιμές που ακολουθούν την Poisson κατανομή. Πολλές φορές σε εφαρμογές παρατηρούμε δεδομένα συχνοτήτων. Η κατανομή Poisson είναι πολύ χρήσιμη στην ανάλυση τέτοιων δεδομένων. Έστω ότι η μεταβλητή απόκρισης y_i μπορεί να πάρει τιμές από τις $0, 1, 2, ..., n$ και μία κατανομή που είναι κατάλληλη για να περιγράψει τα δεδομένα μας είναι η κατανομή Poisson. Είναι $E(y) = \mu$ και $Var(y) = \mu$. Η μέση τιμή είναι ίση με την διασπορά.

Η συνάρτηση σύνδεσης που συνδέει τη μέση τιμή $E(y_i)$ με τις επεξηγηματικές μεταβλητές είναι:

$$g(\mu_i) = n_i = b_0 + b_1x_1 + ... + b_kx_k = \mathbf{x}_i^T\mathbf{b}.$$

Υποθέτουμε ότι $E(Y_i) = \mu_i = n_i\theta_i$.

Για παράδειγμα ας υποθέσουμε ότι Y_i είναι ο αριθμός σε αιτήσεις αποζημίωσης ασφαλίσεων για ένα τύπο αυτοκινήτου. Αυτό θα εξαρτάται από τον αριθμό αυτοκινήτων αυτού του τύπου που έχουν ασφαλιστεί, n_i και άλλες μεταβλητές που επηρεάζουν το θ_i, όπως η ηλικία αυτοκινήτου ή η περιοχή που χρησιμοποιήθηκαν τα αυτοκίνητα αυτά.

Για να αναλύσουμε δεδομένα τέτοιου είδους, μπορούμε να χρησιμοποιήσουμε το μοντέλο:

$$\theta_i = e^{x_i^T b},$$

οπότε το γενικευμένο γραμμικό μοντέλο εκφράζεται ως:

$$E(Y_i) = \mu_i = n_i e^{x_i^T b}, Y_i \sim Poisson(\mu_i).$$

Για παράδειγμα, όπου το $\mathbf{x}_i$ είναι 0, ή 1, το ποσοστιαίο πηλίκο είναι:

$$RR = \frac{E(Y_i|X_i=1)}{E(Y_i|X_i=0)} = e^b.$$

Η εκτίμηση της παραμέτρου b γίνεται μέσω της θεωρίας πιθανοφάνειας για γενικευμένα γραμμικά μοντέλα.

Αν $\hat{b}$ είναι η εκτιμήτρια μέγιστης πιθανοφάνειας, τότε μπορούμε να ελέγξουμε τις υποθέσεις με στατιστικό έλεγχο Wald, έλεγχο score και έλεγχο αναλογίας πιθανοφάνειας.

Οι αναμενόμενες τιμές δίνονται από:

$$\hat{Y}_i = \hat{\mu}_i = n_i e^{x_i \hat{b}}.$$

Στην περίπτωση όπου οι μεταβλητές απόκρισης $Y_1, ..., Y_N$ είναι ανεξάρτητες και $Y_i \sim Poisson(\lambda_i)$, η λογαριθμική συνάρτηση πιθανοφάνειας είναι:

$$l(\boldsymbol{\beta}; \mathbf{y}) = \sum y_i log(\lambda_i) - \sum \lambda_i - \sum log(y_i!).$$

Για το κορεσμένο μοντέλο, τα λ_i είναι όλα διαφορετικά, οπότε: $\boldsymbol{\beta} = [\lambda_1, ..., \lambda_N]^T$. Οι εκτιμήτριες μέγιστης πιθανοφάνειας είναι $\hat{\lambda}_i = y_i$ και έτσι η μέγιστη τιμή της λογαριθμικής συνάρτησης πιθανοφάνειας είναι:

$$l(\mathbf{b}_{max}; \mathbf{y}) = \sum y_i log(y_i) - \sum y_i - \sum log(y_i!) \ .$$

5.3 Μέθοδος μέγιστης πιθανοφάνειας

Ο προσδιορισμός των παραμέτρων γίνεται με την μέθοδο μέγιστης πιθανοφάνειας. Για το μοντέλο Poisson η συνάρτηση πιθανοφάνειας είναι:

$$L(\mathbf{y}, \mathbf{b}) = \prod_{i=1}^n f_i(y_i) = \prod_{i=1}^n \frac{e^{-\mu_i} \mu^{y_i}}{y_i!} = \frac{\prod_{i=1}^n \mu_i^{y_i} exp(-\sum_{i=1}^n \mu_i)}{\prod_{i=1}^n y_i!} \ .$$

Για τις μεταβλητές απόκρισης $y_1, ..., y_n$ έχουμε:

$$ln\{L(\mathbf{y}; \mathbf{b})\} = log \prod_{i=1}^n (\frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}) = \sum_{i=1}^n [y_i \mathbf{x}_i \mathbf{b} - e^{\mathbf{x}_i \mathbf{b}} - ln(y_i!)] \ .$$

Παραγωγίζοντας την παραπάνω σχέση και θέτοντάς τη ίση με μηδέν, έχουμε:

$$\frac{\partial}{\partial b}[ln\{L(\mathbf{y}; \mathbf{b})\}] = 0,$$

$$\sum_{i=1}^n [y_i \mathbf{x}_i - exp(\mathbf{x}_i b)\mathbf{x}_i] = 0 \Rightarrow \sum_{i=1}^n (y_i - \mu_i)\mathbf{x}_i = 0.$$

Σε μορφή πινάκων, η σχέση αυτή γράφεται ως:

$$\mathbf{X}(\mathbf{y} - \boldsymbol{\mu}) = 0.$$

5.4 Μελέτη καταλληλότητας μοντέλου

5.4.1 Deviance για ένα μοντέλο Poisson

Υποθέτουμε ότι το μοντέλο που μας ενδιαφέρει, έχει $p < N$ παραμέτρους. Ο εκτιμητής μέγιστης πιθανοφάνειας **b** μπορεί να υπολογίσει εκτιμήτριες $\hat{\lambda}_i$ και αφού οι εκτιμώμενες τιμές είναι $\hat{y}_i = \hat{\lambda}_i$, επειδή $E(Y_i) = \lambda_i$. Η μέγιστη τιμή της λογαριθμικής συνάρτησης πιθανοφάνειας σε αυτή την περίπτωση, είναι:

$$l(\mathbf{b}; \mathbf{y}) = \sum y_i log(\hat{y}_i) - \sum \hat{y}_i - \sum log(y_i!),$$

οπότε, η deviance είναι:

$$\begin{aligned} D &= 2[l(\mathbf{b}_{max}; \mathbf{y}) - l(\mathbf{b}; \mathbf{y})] \\ &= 2[\sum y_i log(y_i/\hat{y}_i) - \sum(y_i - \hat{y}_i)]. \end{aligned}$$

Για τα περισσότερα μοντέλα, μπορεί να δειχθεί ότι $\sum y_i = \sum \hat{y}_i$. Οπότε η D μπορεί να γραφεί στη μορφή:

$$D = 2 \sum o_i * log(o_i/e_i),$$

όπου το o_i χρησιμοποιείται για να δηλώσουμε την τιμή y_i και το e_i χρησιμοποιείται για να δηλώσουμε την αναμενόμενη τιμή $\hat{y}_i$.

Η τιμή της D, μπορεί να υπολογιστεί από τα δεδομένα στην περίπτωσή μας. Αυτή η τιμή μπορεί να συγκριθεί με την κατανομή $\chi^2(N - p)$.

5.4.2 Υπόλοιπα

Καθώς $Var(Y_i) = E(Y_i)$ για την κατανομή Poisson, το τυπικό σφάλμα του Y_i εκτιμάται ως $\sqrt{e_i}$, έτσι τα υπόλοιπα Pearson είναι:

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}},$$

όπου το o_i δηλώνει τις τιμές του y_i που παρατηρούμε.

Για την κατανομή Poisson, τα υπόλοιπα που δίνονται από την προηγούμενη σχέση και ο έλεγχος καλής προσαρμογής χ-τετράγωνο, σχετίζονται ως εξής:

$$X^2 = \sum r_i^2 = \sum \frac{(o_i - e_i)^2}{e_i}.$$

Η απόκλιση για ένα μοντέλο Poisson, είναι όπως δείξαμε στην προηγούμενη ενότητα. Μπορεί να γραφτεί στη μορφή:

$$D = 2[\sum o_i log(o_i/e_i) - \sum(o_i - e_i)].$$

Για τα περισσότερα μοντέλα, μπορεί να δειχθεί ότι $\sum o_i = \sum e_i$. Οπότε η D μπορεί να γραφεί στη μορφή:

$$D = 2 \sum o_i log(o_i/e_i),$$

και τα υπόλοιπα deviance:

$$d_i = sign(o_i - e_i)\sqrt{2[o_i log(\tfrac{o_i}{e_i}) - (o_i - e_i)]}, i = 1, ..., N,$$

Έτσι ώστε $D = \sum d_i^2$.

Όταν η X^2 παίρνει τιμές στις αναμενόμενες συχνότητες, τότε η στατιστική συνάρτηση X^2 είναι ασυμπτωτικά ισοδύναμη με την D. Χρησιμοποιώντας τις σειρές Taylor έχουμε:

$$o * log(\tfrac{o}{e}) = (o - e) + \tfrac{1}{2}\tfrac{(o-e)^2}{e} + ...$$

Έτσι, προσεγγιστικά έχουμε:

$$D = 2 \sum [(o_i - e_i) + \tfrac{1}{2}\tfrac{(o_i - e_i)^2}{e_i} - (o_i - e_i)]$$

$$= \sum \tfrac{(o_i - e_i)^2}{e_i} = X^2.$$

Τα X^2 και D μπορούν να χρησιμοποιηθούν απ' ευθείας για να ελέγξουμε την καλή προσαρμογή, καθώς μπορούν να υπολογιστούν από τα δεδομένα του προσαρμοσμένου μοντέλου. Μπορούν να συγκριθούν με την κεντρική X^2 κατανομή με N-p βαθμούς ελευθερίας, όπου p είναι ο αριθμός των παραμέτρων που έχουν εκτιμηθεί. Η X^2 κατανομή, είναι συχνά καλύτερη από την D, επειδή η D επηρεάζεται από πολύ μικρές συχνότητες.

5.5 Λογαριθμικά - γραμμικά μοντέλα

Για μοντέλα που βασίζονται στην κατανομή Poisson με τύπο:

$$f(\mathbf{y}; \boldsymbol{\mu}) = \prod_{i=1}^{N} \mu_i^{y_i} e^{-\mu_i} y_i! \ ,$$

όπου $\boldsymbol{\mu}$ είναι ένα διάνυσμα των μ_i.

Οι αναμενόμενες συχνότητες δίνονται από τον τύπο: $E(Y_i) = \mu_i$, $i = 1, ...N$.

Λόγω της συνάρτησης πιθανότητας της αναμενόμενης τιμής, παίρνουμε σαν συνάρτηση σύνδεσης τη λογαριθμική συνάρτηση:

$$log\{E(Y_i)\} = \mathbf{x}_i^T \mathbf{b}.$$

Ο όρος λογαριθμικό - γραμμικό μοντέλο χρησιμοποιείται για να περιγράψει μοντέλα τέτοιου είδους [2].

Αν δυο μεταβλητές είναι ανεξάρτητες, η από κοινού πιθανότητα είναι $\theta_{jk} = \theta_{j.}\theta_{.k}$, όπου $\theta_{j.}$ και $\theta_{.k}$ είναι οι περιθώριες πιθανότητες με $\sum_j \theta_(j.) = 1$, $\sum_k \theta_(.k) = 1$.

5.5.1 Υπερδιασπορά

Υπερδιασπορά, συμβαίνει όταν η διασπορά είναι μεγαλύτερη από τη μέση τιμή. Παρόλο του ότι στην Poisson κατανομή είναι ίσες.
Η αρνητική Διωνυμική κατανομή μας δίνει ένα εναλλακτικό μοντέλο με

$$Var(Y_i) = \phi E(Y_i),$$

όπου $\phi > 1$ είναι μία παράμετρος που μπορεί να εκτιμηθεί. Η υπερδιασπορά μπορεί να συμβεί λόγω της έλλειψης ανεξαρτησίας μεταξύ των παρατηρήσεων.

6 ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΜΕ ΧΡΗΣΗ ΤΟΥ ΣΤΑΤΙΣΤΙΚΟΥ ΠΑΚΕΤΟΥ R

6.1 Γενικά

Η μελέτη των γενικευμένων γραμμικών μοντέλων, πολλές φορές μας οδηγεί σε πολύπλοκους υπολογισμούς. Η χρήση διάφορων στατιστικών πακέτων σε υπολογιστή, μας βοηθάει να λύσουμε πιο εύκολα και γρήγορα, διάφορους υπολογισμούς που προκύπτουν κατά τη στατιστική μελέτη πολλών μοντέλων.

Το στατιστικό πακέτο R είναι ένα αρκετά διαδεδομένο πακέτο. Είναι αρκετά ευέλικτο και χρησιμοποιείται κατά κόρον στον χώρο της στατιστικής. Μας δίνει τη δυνατότητα, να πραγματοποιήσουμε διάφορους στατιστικούς υπολογισμούς γράφοντας απλό κώδικα.

6.2 Χρήση της R

Η γλώσσα προγραμματισμού R είναι πολύ χρήσιμη στην εφαρμογή σύγχρονων στατιστικών τεχνικών. Η γλώσσα R μπορεί να αποκτηθεί ελεύθερα μέσω της ιστοσελίδας: http://www.r-projects.org.

Μπορεί να χρησιμοποιηθεί με κατευθείαν εντολές που υπάρχουν. Επίσης υπάρχουν προγράμματα που μπορούν να χρησιμοποιηθούν για επίλυση πολύπλοκων στατιστικών προβλημάτων.

Όπως έχουμε αναφέρει και πιο πριν, στα γενικευμένα γραμμικά μοντέλα χρησιμοποιούνται πολλές τεχνικές και στατιστικές μέθοδοι, των οποίων οι υπολογισμοί έχουν μεγάλο όγκο, όμως η R είναι ένα χρήσιμο εργαλείο, όπου μπορούμε να το χρησιμοποιήσουμε για να προσεγγίσουμε τους υπολογισμούς μας.

Η R μπορεί να χειριστεί διανύσματα, πίνακες, πλαίσια δεδομένων και συναρτήσεις [11].

6.2.1 Απλό Γραμμικό Μοντέλο στην R

Για να προσαρμόσουμε ένα μοντέλο απλής γραμμικής παλινδρόμησης και να ελέγξουμε κατά πόσο η επεξηγηματική μεταβλητή X επηρεάζει την Y, αρχικά δημιουργούμε ένα διάγραμμα διασποράς για να ελέγξουμε αν η γραμμική συνάρτηση φαίνεται να είναι η κατάλληλη. Αυτό γίνεται με την εντολή plot(x,y), όπου x και y είναι οι τιμές του τυχαίου δείγματος για τις μεταβλητές X και Y αντίστοιχα.

Με το διάγραμμα θα ελέγξουμε αν είναι λογική η υπόθεση της γραμμικότητας.

Με την εντολή $lm(y \sim x)$ προσαρμόζουμε το γραμμικό μοντέλο με y τα δεδομένα της μεταβλητής απόκρισης και x τα δεδομένα για την επεξηγηματική μεταβλητή.

Η R θα μας επιστρέψει τις τιμές $\hat{b}_0$ και $\hat{b}_1$.

Η εντολή summary() θα μας δώσει μια περιεκτική περίληψη για τα αποτελέσματα της ανάλυσης, όπως:

• Περιγραφικούς δείκτες υπολοίπων.
• P-τιμή για τον έλεγχο του b_0 και του b_1.
• Διορθωμένο συντελεστή προσδιορισμού.
• Στοιχεία για τον F-έλεγχο.

6.3 Γενικευμένα Γραμμικά Μοντέλα στην R

Στην R η συνάρτηση που προσαρμόζει ένα γενικευμένο γραμμικό μοντέλο είναι η glm κι έχει τη μορφή:

glm(formula, family, data).

Με τον όρο formula δείχνουμε τις μεταβλητές απόκρισης και τις επεξηγηματικές μεταβλητές στο γενικευμένο γραμμικό μοντέλο που θέλουμε να προσαρμόσουμε στην R.

Στην υπόδειξη family δηλώνουμε την κατανομή που ακολουθούν οι παρατηρήσεις της μεταβλητής απόκρισης. Σημειώνουμε επίσης και το είδος της συνάρτησης σύνδεσης που θέλουμε, στην περίπτωση που μελετάμε.

Στον όρο data θα δηλώσουμε το πλαίσιο δεδομένων, με το όνομα που έχουμε δώσει, αφού καταχωρήσαμε τις τιμές των παρατηρήσεών μας στην R [5].

Έτσι, για παράδειγμα αν θέλουμε να προσαρμόσουμε ένα γενικευμένο γραμμικό μοντέλο το οποίο έχει μεταβλητή απόκρισης Y της οποίας οι παρατηρήσεις ακολουθούν τη διωνυμική κατανομή και επεξηγηματική μεταβλητή X, χρησιμοποιώντας ως συνάρτηση σύνδεσης την probit, θα κάνουμε την εξής διαδικασία:

Θα καταχωρήσουμε στην R τα δεδομένα και θα δημιουργήσουμε ένα πλαίσιο όπου το ονομάζουμε, έστω ”data.frame”. Η εντολή που θα δώσουμε είναι:

glm1<-glm(Y∼x,family=binomial(link=probit),data=data.frame).

6.3.1 Μεταβλητές απόκρισης ακολουθούν την Κανονική κατανομη

Σε περίπτωση που οι τιμές της μεταβλητής απόκρισης ακολουθούν την κανονική κατανομή, μπορούμε να προσαρμόσουμε το μοντέλο στη γλώσσα R με τη χρήση της συνάρτησης lm [4]. Τα γενικά γραμμικά μοντέλα είναι μία ειδική περίπτωση των γενικευμένων γραμμικών μοντέλων, όπου τα στοιχεία της μεταβλητής απόκρισης κατανέμονται κανονικά. Οι επεξηγηματικές μεταβλητές μπορούν να είναι συνδυασμός κατηγορικών και συνεχών μεταβλητών. Στην περίπτωση που θα χρησιμοποιήσουμε την εντολή lm, θα γράψουμε στην R $lm(y \sim x_1 + x_2)$, για ένα παράδειγμα, όπου είναι x_1 και x_2 οι επεξηγηματικές μεταβλητές.

Μπορούμε επίσης να χρησιμοποιήσουμε και την εντολή glm, η οποία θα μας δώσει τα ίδια αποτελέσματα με την lm στην περίπτωση όπου τα δεδομένα της μεταβλητής απόκρισης προέρχονται από την Κανονική κατανομή. Για αυτή την περίπτωση, θα γράψουμε στην R:

$$glm(y \sim x_1 + x_2, family = gaussian).$$

Η R θα μας επιστρέψει τους συντελεστές του μοντέλου μας, τους βαθμούς ελευθερίας και την απόκλιση.

Στη συνέχεια, με την χρήση της εντολής summary( ), μπορούμε να δούμε διάφορες πληροφορίες που μας δίνει η R, για το συγκεκριμένο μοντέλο, όπως P-τιμές, τις εκτιμήσεις των παραμέτρων και δείκτες για την τιμή υπολοίπων απόκλισης. Αντίστοιχα, χρησιμοποιώντας την εντολή summary( ) στην περίπτωση αυτή, αν είχαμε δουλέψει με το μοντέλο που προσαρμόζει η εντολή lm στην R, θα είχαμε τον συντελεστή προσδιορισμού, F-έλεγχο και το διορθωμένο συντελεστή προσδιορισμού για το μοντέλο μας.

6.3.2 Μεταβλητή απόκρισης με δεδομένα που ακολουθούν τη Διωνυμική και Bernoulli κατανομή

Πολύ συχνά, η μεταβλητή απόκρισης παίρνει τις τιμές 0 και 1. Το 0 είναι ο αριθμός που λαμβάνουμε ως αποτυχία και το 1 ο αριθμός που λαμβάνουμε ως επιτυχία. Στην περίπτωση αυτή, η μεταβλητή απόκρισης έχει δεδομένα που ακολουθούν τη Bernoulli κατανομή.

Η εντολή που θα δώσουμε στην R για να προσαρμόσει το γενικευμένο γραμμικό μοντέλο σε αυτή την περίπτωση, είναι η glm. Στην περίπτωση αυτή θα έχουμε family=binomial. Αν δεν συμπληρώσουμε τίποτα μετά από το family=binomial, η

R, θα θεωρήσει ότι η συνάρτηση σύνδεσης στην περίπτωσή μας είναι η λογιστική.

Σε διαφορετική περίπτωση, αν θα θέλαμε να εξετάσουμε το μοντέλο μας για την περίπτωση όπου η συνάρτηση σύνδεσης είναι η κανονική, θα μπορούσαμε να συμπληρώσουμε:

family=binomial(link=probit).

Επίσης θα μπορούσαμε να εξετάσουμε για link=cloglog.

Θα μπορούσαμε σε κάποιο πρόβλημα να δουλέψουμε πάνω σε κάποια μοντέλα με διαφορετικές συναρτήσεις σύνδεσης και έπειτα, από την τιμή του D που θα βρούμε μέσω της R, μετά την χρήση της εντολής glm, να δούμε ποιο μοντέλο είναι αυτό που περιγράφει καλύτερα τα δεδομένα μας.

6.3.3 Μεταβλητή απόκρισης με δεδομένα που ακολουθούν την κατανομή Poisson

Σε περιπτώσεις όπου η μεταβλητή απόκρισης ακολουθεί την Poisson κατανομή, στην R θα χρησιμοποιούμε τον όρο family=poisson στην εντολή glm για να δείξουμε ότι το γενικευμένο γραμμικό μοντέλο που μας ενδιαφέρει, αφορά την παλινδρόμηση Poisson.

Όπως και στα προηγούμενα, μπορούμε να επιλέξουμε ποιά συνάρτηση σύνδεσης θέλουμε να χρησιμοποιήσουμε στο μοντέλο. Αυτό γίνεται αντίστοιχα, όπως περιγράψαμε στην περίπτωση που είχαμε δεδομένα της διωνυμικής κατανομής.

Για παράδειγμα, αν δηλώσουμε:

family=poisson(link=identity),

η R θα μας επιστρέψει αποτελέσματα για μοντέλο Poisson με συνάρτηση σύνδεσης την ταυτοτική.

Πολλές φορές μπορεί να υπάρχουν δεδομένα, που είναι μέρος της παραμέτρου b_1. Στην R, για να ξεχωρήσουμε αυτόν τον όρο από τις υπόλοιπες επεξηγηματικές μεταβλητές, θα χρησιμοποιήσουμε την εντολή offset. Αυτό γίνεται, για να μην ληφθεί από την R σαν επεξηγηματική μεταβλητή [4].

6.4 Παραδείγματα Γενικευμένων Γραμμικών Μοντέλων στην Γλώσσα Προγραμματισμού R

Σε αυτό το κομμάτι της διπλωματικής εργασίας θα δούμε κάποιες από τις εντολές που αναφέραμε στις προηγούμενες παραγράφους, πως λειτουργούν στη γλώσσα R και θα δούμε τα αποτελέσματα που μας επιστρέφει η R. Επίσης θα συνδέσουμε τα αποτελέσματα αυτά με τη θεωρία των γενικευμένων γραμμικών μοντέλων μέσω κάποιων παραδειγμάτων.

6.4.1 Θνησιμότητα σκαθαριών

Ο πίνακας που ακολουθεί δείχνει τον αριθμό σκαθαριών που πέθαναν μετά από πεντάωρη έκθεση σε τοξικό αέριο, σε διάφορες συγκεντρώσεις. (Τα δεδομένα χρησιμοποιήθηκαν από τον Chester Bliss το 1935) [2].

Δόση x_i	Αριθμός σκαθαριών n_i	Σκαθάρια που πέθαναν y_i
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Πίνακας 3: *Δεδομένα θνησιμότητας σκαθαριών.*

Στην R αρχικά θα καταχωρήσουμε τις τιμές του παραπάνω πίνακα ως διανύσματα, χρησιμοποιώντας τις παρακάτω εντολές:

dose<-c(1.6907,1.7242,1.7552,1.7842,1.8113,1.8369,1.8610,1.8839),

beetles<-c(59,60,62,56,63,59,62,60),

killed<-c(6,13,18,28,52,53,61,60).

Για να χρησιμοποιήσουμε τα δεδομένα μας και για να προσαρμόσουμε το μοντέλο, θα δημιουργήσουμε ένα πλαίσιο δεδομένων για τις τιμές αυτές. Αυτό γίνεται με την εντολή που ακολουθεί:

plaisio<-data.frame(dose,beetles,killed).

Επειδή στη μελέτη αυτού του προβλήματος μας ενδιαφέρει η θνησιμότητα των σκαθαριών, η μεταβλητή απόκρισης θα είναι μεταξύ δύο τιμών: "σκαθάρια που έζησαν" και "σκαθάρια που πέθαναν". Έχουμε δηλαδή, ένα παράδειγμα γενικευμένου γραμμικού μοντέλου με δίτιμη μεταβλητή. Οπότε θα ορίσουμε την απόκριση Y με τον ακόλουθο τρόπο:

plaisio$Y<-cbind(plaisio$killed,plaisio$beetles-plaisio$killed).

Η R θα μας επιστρέψει ένα πλαίσιο από δύο ενωμένα διανύσματα από τα οποία το ένα είναι ο αριθμός των σκαθαριών που πέθαναν και το άλλο είναι ο αριθμός των σκαθαριών που επιβίωσαν.

Τα αποτελέσματα που επιστρέφει η R για τις δυο τελευταίες εντολές, είναι:

```
> plaisio
    dose beetles killed
1 1.6907      59      6
2 1.7242      60     13
3 1.7552      62     18
4 1.7842      56     28
5 1.8113      63     52
6 1.8369      59     53
7 1.8610      62     61
8 1.8839      60     60
```

Πίνακας 4: *Πλαίσιο δεδομένων για τις τιμές dose, beetles, killed.*

```
> plaisio$Y
     [,1] [,2]
[1,]    6   53
[2,]   13   47
[3,]   18   44
[4,]   28   28
[5,]   52   11
[6,]   53    6
[7,]   61    1
[8,]   60    0
```

Πίνακας 5: *Πλαίσιο δεδομένων για τις τιμές της απόκρισης.*

Οι τιμές της μεταβλητής απόκρισης προέρχονται από την κατανομή Bernoulli. Θα προσαρμόσουμε το λογιστικό μοντέλο:

$$\pi_i = \frac{exp(b_1+b_2x_i)}{1+exp(b_1+b_2x_i)},$$

οπότε:

$$log(\tfrac{\pi_i}{1-\pi_i}) = b_1 + b_2 x_i,$$

και

$$log(1-\pi_i) = -log[1 + exp(b_1 + b_2 x_i)].$$

Προκύπτει οπότε η λογαριθμική συνάρτηση πιθανοφάνειας:

$$l = \textstyle\sum_{i=1}^{N} [y_i(b_1 + b_2 x_i) - n_i log[1 + exp(b_1 + b_2 x_i)] + log \begin{pmatrix} n_i \\ y_i \end{pmatrix}].$$

Τα score για b_1 και b_2 είναι:

$$U_1 = \tfrac{\partial l}{\partial b_1} = \textstyle\sum (y_i - n_i [\tfrac{exp(b_1+b_2 x_i)}{1+exp(b_1+b_2 x_i)}]) = \sum (y_i - n_i \pi_i),$$

$$U_2 = \tfrac{\partial l}{\partial b_2} = \textstyle\sum (y_i x_i - n_i x_i [\tfrac{exp(b_1+b_2 x_i)}{1+exp(b_1+b_2 x_i)}]) = \sum x_i (y_i - n_i \pi_i),$$

και ο πίνακας πληροφορίας είναι:

$$\mathcal{I} = \begin{bmatrix} \sum n_i \pi_i (1-\pi_i) & \sum n_i x_i \pi_i (1-\pi_i) \\ \sum n_i x_i \pi_i (1-\pi_i) & \sum n_i x_i^2 \pi_i (1-\pi_i) \end{bmatrix}.$$

Οι εκτιμήτριες μέγιστης πιθανοφάνειας, βρίσκονται λύνοντας την επαναληπτική εξίσωση:

$$\mathcal{I}^{(m-1)} \mathbf{b}^m = \mathcal{I}^{(m-1)} \mathbf{b}^{(m-1)} + \mathbf{U}^{(m-1)},$$

όπου (m), η m-οστή προσέγγιση και **b** το διάνυσμα των εκτιμήσεων.

Ξεκινώντας από τις τιμές $b_1^{(0)} = 0$ και $b_2^{(0)} = 0$ κάνουμε διαδοχικές εκτιμήσεις, όπου φαίνονται στον παρακάτω πίνακα. Ο πίνακας επίσης δείχνει την αύξηση των τιμών της λογαριθμικής συνάρτησης πιθανοφάνειας παραλείποντας τον σταθερό όρο $log \begin{pmatrix} n_i \\ y_i \end{pmatrix}$. Οι προσαρμοσμένες τιμές $\hat{y}_i = n_i \hat{\pi}_i$ φαίνονται για κάθε βήμα.

Για την τελική προσέγγιση, ο εκτιμώμενος πίνακας διακύμανσης - συνδιακύμανσης για το **b**, $[\mathcal{I}(\mathbf{b})^{-1}]$, φαίνεται επίσης στον Πίνακα 6, όπως και η deviance.

Η deviance είναι $D = 2\sum_{i=1}^{N} [y_i log(\frac{y_i}{\hat{y}_i}) + (n_i - y_i) log(\frac{n-y_i}{n-\hat{y}_i})]$.

Οι εκτιμήσεις με τα τυπικά τους σφάλματα είναι:

$b_1 = -60.72$ και τυπικό σφάλμα=$\sqrt{26.840} = 5.18$ και

	Αρχική Εκτίμηση	Προσέγγιση πρώτη	δεύτερη	έκτη
b_1	0	-37.856	-53.853	-60.717
b_2	0	21.337	30.384	34.270
log πιθανοφάνεια	-333.404	-200.010	-187.274	-186.235

Παρατηρήσεις			Προσαρμοσμένες τιμές		
y_1	6	29.5	8.505	4.543	3.458
y_2	13	30.0	15.366	11.254	9.842
y_3	18	31.0	24.808	23.058	22.451
y_4	28	28.0	30.983	32.947	33.898
y_5	52	31.5	43.362	48.197	50.096
y_6	53	29.5	46.741	51.705	53.291
y_7	61	31.0	53.595	58.061	59.222
y_8	60	30.0	54.734	58.036	58.743

$$[\mathcal{I}(\hat{\mathbf{b}})]^{-1} = \begin{bmatrix} 26.840 & -15.082 \\ -15.082 & 8.481 \end{bmatrix}, \quad \text{D=11.23}$$

Πίνακας 6: *Προσαρμόζοντας ένα λογιστικό γραμμικό μοντέλο στα δεδομένα θνησιμότητας σκαθαριών.*

$b_2 = 34.72$ και τυπικό σφάλμα=$\sqrt{8.481} = 2.91$.

Επίσης, $D = 11.23$.

Για να προσαρμόσουμε στην R το μοντέλο μας θα χρησιμοποιήσουμε την παρακάτω εντολή:

montelo<-glm(Y∼dose,family=binomial,data=plaisio).

Τα αποτελέσματα που μας δίνει η R φαίνονται στον Πίνακα 7.

Χρήσιμα αποτελέσματα μπορούμε να έχουμε και με την εντολή summary(montelo), όπως φαίνεται στον Πίνακα 8.

Αν το μοντέλο μας δίνει καλή περιγραφή των δεδομένων, η απόκλιση θα πρέπει να ακολουθεί την κατανομή χ^2_6, επειδή έχουμε $N = 8$ παρατηρήσεις και $p = 2$

```
> montelo

Call:  glm(formula = Y ~ dose, family = binomial, data = plaisio)

Coefficients:
(Intercept)         dose
     -60.72        34.27

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:      284.2
Residual Deviance: 11.23        AIC: 41.43
```

Πίνακας 7: *Αποτελέσματα της R για το γενικευμένο γραμμικό μοντέλο με όνομα montelo.*

```
> summary(montelo)

Call:
glm(formula = Y ~ dose, family = binomial, data = plaisio)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5941  -0.3944   0.8329   1.2592   1.5940

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -60.717      5.181  -11.72   <2e-16 ***
dose          34.270      2.912   11.77   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 284.202  on 7  degrees of freedom
Residual deviance:  11.232  on 6  degrees of freedom
AIC: 41.43

Number of Fisher Scoring iterations: 4
```

Πίνακας 8: *Αποτελέσματα για την εντολή summary για το γενικευμένο γραμμικό μοντέλο με όνομα montelo.*

παραμέτρους.

Το πάνω 5% ποσοστιαίο σημείο της κατανομής χ^2_6 είναι 12.59, όμως η τιμή της D που βρίσκουμε είναι πολύ μεγαλύτερη από την αναμενόμενη, οπότε καταλαβαίνουμε ότι το μοντέλο δεν προσαρμόζεται καλά στα δεδομένα.

Μπορούμε να χρησιμοποιήσουμε διάφορες συναρτήσεις σύνδεσης. Τα αποτελέσματα φαίνονται στον Πίνακα 9.

Μεταξύ των μοντέλων, αυτό των ακραίων τιμών, φαίνεται να προσαρμόζεται καλύτερα στα δεδομένα μας.

Στην R, αν δεν δηλώσουμε ποιά συνάρτηση σύνδεσης θέλουμε να χρησιμο-

Y	Logistic	Probit	Extreme value
6	3.46	3.36	5.59
13	9.84	10.72	11.28
18	22.45	23.48	20.95
28	33.90	33.82	30.37
52	50.10	49.62	47.78
53	53.29	53.32	54.14
61	59.22	59.66	61.11
60	58.74	59.23	59.95
D	11.23	10.12	3.45

Πίνακας 9: *Σύγκριση διαφορετικών μοντέλων για τα δεδομένα θνησιμότητας σκαθαριών. Επίσης φαίνονται οι τιμές των deviance για το κάθε μοντέλο.*

ποιηθεί, αυτή που χρησιμοποιείται αυτόματα, είναι η logit.

Θα χρησιμοποιήσουμε τώρα στην R τις συναρτήσεις σύνδεσης probit και cloglog. Αυτό θα το κατορθώσουμε με τις επόμενες εντολές, όπου βλέπουμε και τα αποτελέσματα που μας δίνει η R:

Για τη συνάρτηση σύνδεσης probit:

montelo2<-glm(Y∼dose,family=binomial(link=probit),data=plaisio):

```
> montelo2

Call:  glm(formula = Y ~ dose, family = binomial(link = probit), data = plaisio)

Coefficients:
(Intercept)         dose
     -34.94        19.73

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:      284.2
Residual Deviance: 10.12        AIC: 40.32
```

Πίνακας 10: *Αποτελέσματα της R για το μοντέλο με συνάρτηση σύνδεσης τη probit.*

Για τη συνάρτηση σύνδεσης cloglog:

montelo3<-glm(Y∼dose,family=binomial(link=cloglog),data=plaisio):

```
> montelo3

Call:  glm(formula = Y ~ dose, family = binomial(link = cloglog), data = plaisio)

Coefficients:
(Intercept)         dose
     -39.57        22.04

Degrees of Freedom: 7 Total (i.e. Null);  6 Residual
Null Deviance:      284.2
Residual Deviance: 3.446        AIC: 33.64
```

Πίνακας 11: *Αποτελέσματα της R για το μοντέλο με συνάρτηση σύνδεσης τη cloglog.*

6.4.2 Παράδειγμα δεδομένων Poisson παλινδρόμησης

Στον πίνακα που ακολουθεί, υπάρχουν παρατηρήσεις σχετικά με τις επιδόσεις φοιτητών που χωρίζονται σε 3 τμήματα, στη γραπτή εξέταση ενός μαθήματος σε διαδοχικά εξάμηνα για κάποιο διάστημα. Συγκεκριμένα, οι παρατηρήσεις y είναι ο αριθμός φοιτητών που έγραψαν άριστα και το x μας δείχνει από ποιό τμήμα είναι ο κάθε φοιτητής. Για το πρώτο τμήμα, θέτουμε x=-1, για το δεύτερο, x=0 και για το τρίτο, x=1 [2].

y_i	2	3	6	7	8	9	10	12	15
x_i	-1	-1	0	0	0	0	1	1	1

Πίνακας 12: *Δεδομένα για το παράδειγμα Poisson παλινδρόμησης.*

Θα καταχωρήσουμε στην R τα δεδομένα μας δημιουργώντας αρχικά δύο διανύσματα.

y<-c(2,3,6,7,8,9,10,12,15),

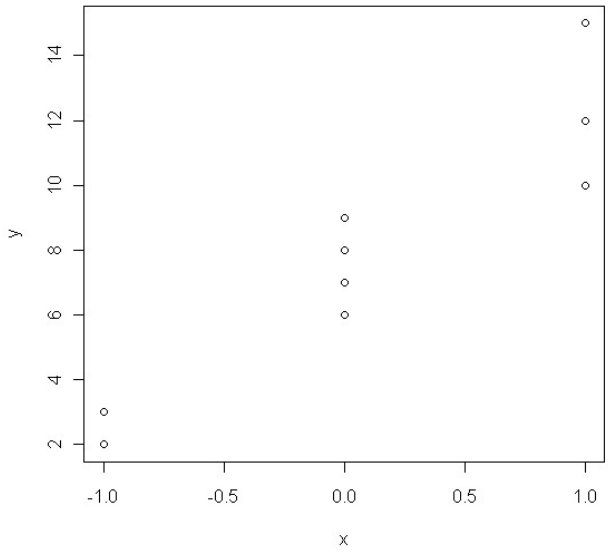
και

x<-c(-1,-1,0,0,0,0,1,1,1).

Έπειτα, θα δημιουργήσουμε με τη βοήθεια της R ένα διάγραμμα για αυτές τις τιμές.

Αυτό θα γίνει με την εντολή plot(x,y) και φαίνεται στο Σχήμα 1 που ακολουθεί.

Από το γράφημα παρατηρούμε ότι η μεταβλητότητα αυξάνεται με το Y. Μπορούμε να υποθέσουμε ότι οι τιμές της μεταβλητής απόκρισης Y είναι τυχαίες με-



Σχήμα 1: *Διάγραμμα τιμών του παραδείγματος Poisson παλινδρόμησης.*

ταβλητές που ακολουθούν την κατανομή Poisson.

Στην Poisson κατανομή η αναμενόμενη τιμή και η διασπορά του Y είναι ίσες.

$$E(Y_i) = Var(Y_i).$$

Ας μοντελοποιήσουμε τη σχέση μεταξύ του Y_i και του x_i με την γραμμική σχέση:

$$E[Y_i] = \mu_i = b_1 + b_2 x_i = \mathbf{x}_i^T \mathbf{b},$$

$$\text{όπου:} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \quad \text{και} \quad \mathbf{x}_i = \begin{bmatrix} 1 \\ x_i \end{bmatrix} \quad , \quad \text{για i=1,...,N.}$$

Μπορούμε λοιπόν να πάρουμε για συνάρτηση σύνδεσης $g(\mu_i)$ την ταυτοτική συνάρτηση:

$$g(\mu_i) = \mu_i = \mathbf{x}_i^T \mathbf{b} = \eta_i.$$

Επομένως έχουμε $\frac{\partial \mu_i}{\partial \eta_i} = 1$, η οποία απλοποιεί τις εξισώσεις:

$$w_{ii} = \frac{1}{Var(Y_i)} (\frac{\partial \mu_i}{\partial \eta_i})^2 \quad \text{και} \quad z_i = \sum_{k=1}^{p} x_{ik} b_k^{m-1} + (y_i - \mu_i)(\frac{\partial \eta_i}{\partial \mu_i}).$$

Επειδή εδώ έχουμε $E(Y_i) = Var(Y_i)$ και λόγω της $w_{ii} = \frac{1}{Var(Y_i)} (\frac{\partial \mu_i}{\partial \eta_i})^2$, θα έχουμε:

$$w_{ii} = \frac{1}{Var(Y_i)} = \frac{1}{b_1+b_2x_i}.$$

Χρησιμοποιώντας την εκτίμηση $\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$, η εξίσωση:

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

γίνεται:

$$z_i = b_1 + b_2 X_i + (y_i - b_{1b2} x_i) = y_i.$$

Επίσης:

$$\mathcal{I} = \mathbf{X}^T\mathbf{W}\mathbf{X} = \begin{bmatrix} \sum_{i=1}^{N} \frac{1}{b_1+b_2x_i} & \sum_{i=1}^{N} \frac{x_i}{b_1+b_2x_i} \\ \sum_{i=1}^{N} \frac{x_i}{b_1+b_2x_i} & \sum_{i=1}^{N} \frac{x_i^2}{b_1+b_2x_i} \end{bmatrix},$$

και

$$\mathbf{X}^T\mathbf{W}\mathbf{z} = \begin{bmatrix} \sum_{i=1}^{N} \frac{y_i}{b_1+b_2x_i} \\ \sum_{i=1}^{N} \frac{x_iy_i}{b_1+b_2x_i} \end{bmatrix}.$$

Οι εκτιμητές μέγιστης πιθανοφάνειας βρίσκονται από τις επαναληπτικές εξισώσεις:

$$(\mathbf{X}^T\mathbf{W}\mathbf{X})^{(m-1)}\mathbf{b}^{(m)} = \mathbf{X}^T\mathbf{W}\mathbf{z}^{(m-1)},$$

όπου η τιμή $\mathbf{b}^{(m-1)}$ δηλώνει τη (m-1) επανάληψη της μεθόδου Newton-Raphson.

Για να δημιουργήσουμε στην R το μοντέλο μας, αρχικά θα φτιάξουμε ένα πλαίσιο δεδομένων με τις τιμές των x,y. Αυτό θα γίνει όπως φαίνεται στον Πίνακα 13.

```
> plaisio<-data.frame(y,x)
> plaisio
   y  x
1  2 -1
2  3 -1
3  6  0
4  7  0
5  8  0
6  9  0
7 10  1
8 12  1
9 15  1
```

Πίνακας 13: *Πλαίσιο δεδομένων με τα διανύσματα των τιμών των x,y.*

Στη συνέχεια θα κάνουμε εκτιμήσεις, χρησιμοποιώντας την εντολή glm:

montelo.glm<-glm(y∼x,family=poisson(link=identity),data=plaisio),

όπου θέσαμε ως family=poisson(link=identity), επειδή στο παράδειγμά μας έχουμε πάρει την ταυτοτική συνάρτηση σύνδεσης.

Τα αποτελέσματα που μας δίνει η R φαίνονται στον Πίνακα 14.

```
> montelo.glm

Call:  glm(formula = y ~ x, family = poisson(link = identity), data = plaisio)

Coefficients:
(Intercept)            x
      7.452        4.935

Degrees of Freedom: 8 Total (i.e. Null);  7 Residual
Null Deviance:      18.42
Residual Deviance: 1.895        AIC: 40.01
```

Πίνακας 14: *Αποτελέσματα της R για το γενικευμένο γραμμικό μοντέλο που ορίσαμε.*

Για τα δεδομένα μας έχουμε N=9.

$$\mathbf{y} = \mathbf{z} = \begin{bmatrix} 2 \\ 3 \\ . \\ . \\ 15 \end{bmatrix} \text{ και}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ . \\ . \\ \mathbf{x}_9 \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ 1 & -1 \\ . & . \\ . & . \\ 1 & 1 \end{bmatrix}.$$

Από τη γραφική παράσταση που σχεδιάσαμε με τη βοήθεια της R νωρίτερα, προκύπτουν οι αρχικές εκτιμήσεις $b_1^{(1)} = 7$ και $b_2^{(1)} = 5$.

Έχουμε:

$$(\mathbf{X}^T\mathbf{W}\mathbf{X})^{(1)} = \begin{bmatrix} 1.821429 & -0.75 \\ -0.75 & 1.25 \end{bmatrix},$$

$$(\mathbf{X}^T\mathbf{W}\mathbf{z})^{(1)} = \begin{bmatrix} 9.869048 \\ 0.583333 \end{bmatrix},$$

οπότε:

$$\mathbf{b}^{(2)} = [(\mathbf{X}^T\mathbf{W}\mathbf{X})^{(1)}]^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{z})^{(1)} =$$

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

Η επαναληπτική διαδικασία, μπορεί να συνεχιστεί μέχρι να συγκλίνει. Τα αποτελέσματα φαίνονται στον Πίνακα 15.

m	1	2	3	4
$b_1^{(m)}$	7	7.45139	7.45163	7.45163
$b_2^{(m)}$	5	4.93750	4.93531	4.93530

Πίνακας 15: *Διαδοχικές προσεγγίσεις για τους συντελεστές παλινδρόμησης του παραδείγματος Poisson παλινδρόμησης.*

Οι εκτιμήτριες μέγιστης πιθανοφάνειας είναι: $\hat{b}_1 = 7.45163$ και $\hat{b}_2 = 4.93530$, μετά από 4 βήματα.

Για τις τιμές αυτές, ο αντίστροφος του πίνακα πληροφορίας, είναι:

$$\mathcal{I}^{-1} = \begin{bmatrix} 0.7817 & 0.4166 \\ 0.4166 & 1.1863 \end{bmatrix}.$$

Αυτός είναι ο πίνακας διασποράς-συνδιασποράς για το $\hat{b}$.

Στην R αρχικά θα υπολογίσουμε για τα δεδομένα μας τον πίνακα πληροφορίας. Είναι λοιπόν:

```
> inf_mat<-matrix((function(x,y)c(sum(1/(7.4516+4.93535*x)),sum(x/(7.4516+4.93535*x))

,sum(x/(7.4516+4.93535*x)),sum(x^2/(7.4516+4.93535*x))))(x,y),ncol=2)
```

Το αποτέλεσμα της R όταν ζητήσουμε να μας επιστρέψει τον πίνακα πληροφορίας φαίνονται στον Πίνακα 16.

Χρησιμοποιώντας την εντολή solve, η R θα μας επιστρέψει τον αντίστροφο του πίνακα πληροφορίας, όπως φαίνεται στον Πίνακα 17.

```
> inf_mat
           [,1]       [,2]
[1,]  1.5738214 -0.5526432
[2,] -0.5526432  1.0370240
```

Πίνακας 16: *Αποτελέσματα της R για τον πίνακα πληροφορίας.*

```
> jantistrofos<-solve(inf_mat)
> jantistrofos
          [,1]      [,2]
[1,] 0.7816709 0.4165623
[2,] 0.4165623 1.1862892
```

Πίνακας 17: *Αποτελέσματα της R για τον αντίστροφο του πίνακα πληροφορίας.*

Έτσι, για παράδειγμα, ένα 95% διάστημα εμπιστοσύνης για το b_2, είναι:

$$4.9353 \pm 1.96\sqrt{1.1863}, \quad \text{ή} \quad (2.80,7.07).$$

Μία σημαντική εντολή είναι η summary. Δίνοντας αυτή την εντολή, μπορούμε να δούμε σημαντικές πληροφορίες για το μοντέλο μας:

```
> summary(montelo.glm)

Call:
glm(formula = y ~ x, family = poisson(link = identity), data = plaisio)

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

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

(Dispersion parameter for poisson family taken to be 1)

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

Number of Fisher Scoring iterations: 3
```

Πίνακας 18: *Αποτελέσματα της R για την εντολή summary.*

Η deviance είναι:

$$\begin{aligned} D &= 2[l(\mathbf{b}_{max};\mathbf{y}) - l(\mathbf{b};\mathbf{y})] \\ &= 2[\sum y_i log(y_i/\hat{y}_i) - \sum(y_i - \hat{y}_i)]. \end{aligned}$$

Για τα περισσότερα μοντέλα, μπορεί να δειχθεί ότι $\sum y_i = \sum \hat{y}_i$. Οπότε η D μπορεί να γραφεί στη μορφή:

$$D = 2 \sum o_i log(o_i/e_i),$$

όπου o_i χρησιμοποιείται για να δηλώσουμε την τιμή y_i και το e_i χρησιμοποιείται για να δηλώσουμε την αναμενόμενη τιμή $\hat{y}_i$.

Η τιμή της D, μπορεί να υπολογιστεί από τα δεδομένα στην περίπτωσή μας. Αυτή η τιμή μπορεί να συγκριθεί με την κατανομή χ^2_{N-p}. Συγκεκριμένα για το παράδειγμά μας:

Στον πίνακα που 19, βλέπουμε τις αναμενόμενες τιμές.

$$\hat{y}_i = b_1 + b_2 x_i,$$

όπου έχουμε: $b_1 = 7.45163$ και $b_2 = 4.93530$.

x_i	y_i	$\hat{y}_i$	$y_i log(y_i/\hat{y}_i)$
-1	2	2.51633	-0.45931
-1	3	2.51633	0.52743
0	6	7.45163	-1.30004
0	7	7.45163	-0.43766
0	8	7.45163	0.56807
0	9	7.45163	1.69913
1	10	12.38693	-2.14057
1	12	12.38693	-0.38082
1	15	12.38693	2.87112
Σύνολο	72	72	0.94735

Πίνακας 19: *Αποτελέσματα για το μοντέλο Poisson.*

Η τιμή της deviance είναι: $D = 2 \bullet (0.94735 - 0) = 1.8947$, η οποία έχει μικρή τιμή σε σχέση με την χ^2 κατανομή με βαθμούς ελευθερίας N-p=9-2=7. Για την ακρίβεια, η απόκλιση είναι κάτω από το 5% της ουράς της κατανομής χ^2_7 υποδεικνύοντας ότι το μοντέλο προσαρμόζεται καλά στα δεδομένα.

Θα εξετάσουμε την καταλληλότητα του μοντέλου, αναλύοντας τα υπόλοιπα. Θα υπολογίσουμε τα υπόλοιπα Pearson και τα υπόλοιπα Deviance. Τέλος, θα δημιουργήσουμε ιστογράμματα και γραφήματα για να ελέγξουμε την κανονικότητα των υπολοίπων.

Για να βρούμε τα υπόλοιπα Pearson, θα δώσουμε στην R την εντολή:
> res.pearson<-residuals(montelo.glm,type=”pearson”).
Τα αποτελέσματα που παίρνουμε είναι τα εξής:

```
> res.pearson
         1          2          3          4          5          6          7
-0.3254956  0.3049041 -0.5317790 -0.1654475  0.2008840  0.5672155 -0.6782012
         8          9
-0.1099399  0.7424520
```

Πίνακας 20: *Υπόλοιπα Pearson για το μοντέλο.*

Για να βρούμε τα υπόλοιπα Deviance, θα δώσουμε στην R την εντολή:
> res.deviance<-residuals(montelo.glm,type=”deviance”).
Τα αποτελέσματα που παίρνουμε είναι τα εξής:

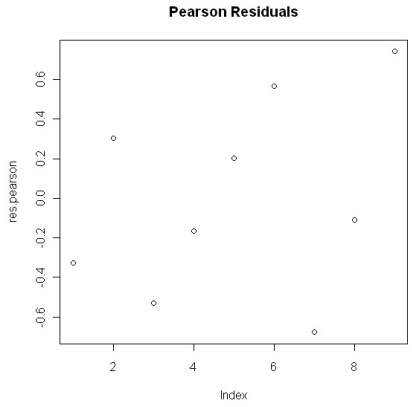
```
> res.deviance
         1          2          3          4          5          6          7
-0.3377034  0.2958445 -0.5506190 -0.1671624  0.1984928  0.5490999 -0.7019443
         8          9
-0.1105199  0.7184063
```

Πίνακας 21: *Υπόλοιπα Deviance για το μοντέλο.*

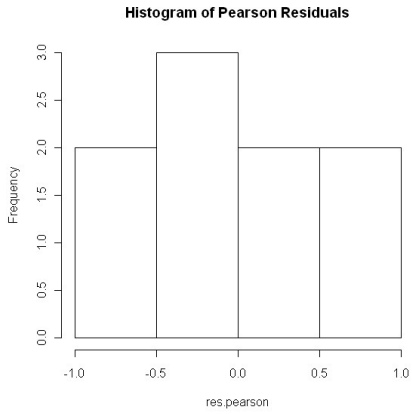
Τα γραφήματα και τα ιστογράμματα για τα υπόλοιπα Pearson και τα υπόλοιπα Deviance αντίστοιχα θα δημιουργηθούν ως εξής:

> plot(res.pearson,main=”Pearson Residuals”),
> hist(res.pearson,main=”Histogram of Pearson Residuals”),
> plot(res.deviance,main=”Deviance Residuals”),
> hist(res.deviance,main=”Histogram of Deviance Residuals”).

Τα αποτελέσματα που μας δίνει η R φαίνονται στα Σχήματα 2, 3, 4, 5 αντίστοιχα.

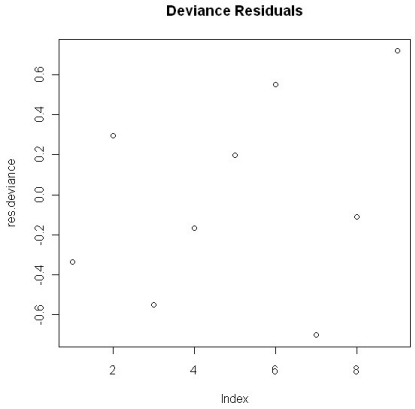


Σχήμα 2: *Γράφημα για τα υπόλοιπα Pearson.*

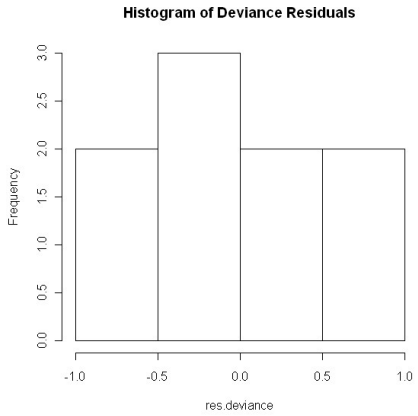


Σχήμα 3: *Ιστόγραμμα για τα υπόλοιπα Pearson.*

Από το γράφημα για τα υπόλοιπα Pearson το οποίο φαίνεται στο Σχήμα 2 παρατηρούμε ότι τα υπόλοιπα, σε σχέση με τη σειρά των δεδομένων, δεν παρουσιάζουν κάποια σχέση και τα υπόλοιπα συμπεριφέρονται τυχαία. Έτσι καταλήγουμε στο συμπέρασμα ότι η υπόθεση ανεξαρτησίας των σφαλμάτων είναι λογική. Στο ιστόγραμμα των υπολοίπων Pearson που ακολουθεί, φαίνεται πως κατανέμονται με κανονικό τρόπο, οπότε ισχύει η κανονικότητα των υπολοίπων.



Σχήμα 4: *Γράφημα για τα υπόλοιπα Deviance.*



Σχήμα 5: *Ιστόγραμμα για τα υπόλοιπα Deviance.*

Από το γράφημα υπολοίπων Deviance που φαίνεται στο Σχήμα 4, παρατηρούμε ότι έχουμε ανεξαρτησία και από το ιστόγραμμα που φαίνεται στο Σχήμα 5 συμπεραίνουμε ότι τα υπόλοιπα Deviance κατανέμονται κανονικά.

6.4.3 Μελέτη σε Βρετανούς γιατρούς

Στον Πίνακα 22, φαίνονται δεδομένα που αφορούν μία πολύ γνωστή μελέτη από τον Sir Richard Doll και τους συνεργάτες του το 1951. Σε αυτή την έρευνα, όλοι οι Βρετανοί γιατροί ρωτήθηκαν μέσω ενός σύντομου ερωτηματολογίου σχετικά με το αν καπνίζουν ή όχι. Αφού συγκεντρώθηκαν πληροφορίες σχετικά με τον αριθμό θανάτων από καρδιολογικά αίτια μετά από 10 χρόνια, φαίνονται τα αποτελέσματα στον Πίνακα 22. Επίσης φαίνεται ο αριθμός ατόμων με τις ηλικίες τους στη διάρκεια της ανάλυσης [2], [1].

Καπνιστές:	Ναι		Όχι	
Ηλικίες	Θάνατοι	Άτομα	Θάνατοι	Άτομα
35 - 44	32	52407	2	18790
45 - 54	104	43248	12	10673
55 - 64	206	28612	28	5710
65 - 74	186	12663	28	2585
75 - 84	102	5317	31	1462

Πίνακας 22: *Θάνατοι από στεφανιαία νόσο μετά από 10 χρόνια, μεταξύ ανδρών, Βρετανών γιατρών, σε κατηγορία κατά ηλικία και κατά το αν καπνίζουν ή όχι το 1951.*

Θα καταχωρήσουμε στην R τα δεδομένα μας με τις εντολές:

age<-c(”35-44”,”45-54”,”55-64”,”65-74”,”75-84”,”35-44”,”45-54”,”55-64”,”65-74”,”75-84”),

deaths<-c(32,104,206,186,102,2,12,28,28,31),

personyears<-c(52407,43248,28612,12663,5317,18790,10673,5710,2585,1462),

smokers<-rep(c(”yes”,”no”),each=5).

Στη συνέχεια θα δημιουργήσουμε ένα πλαίσιο δεδομένων με τα στοιχεία αυτά:

plaisio<-data.frame(age,deaths,personyears,smokers).

Το αποτέλεσμα που λαμβάνουμε φαίνονται στον Πίνακα 23.

	age	deaths	personyears	smokers
1	35-44	32	52407	yes
2	45-54	104	43248	yes
3	55-64	206	28612	yes
4	65-74	186	12663	yes
5	75-84	102	5317	yes
6	35-44	2	18790	no
7	45-54	12	10673	no
8	55-64	28	5710	no
9	65-74	28	2585	no
10	75-84	31	1462	no

Πίνακας 23: *Πλαίσιο δεδομένων που δημιουργήσαμε για τα δεδομένα που αφορούν το παράδειγμα Βρετανών γιατρών.*

Τα ερωτήματα που μας ενδιαφέρουν είναι:

1. Τα ποσοστά θανάτων είναι πιο ψηλά στους καπνιστές απ' ότι στους μη-καπνιστές;

2. Αν ναι, πόσο πιο ψηλά;

3. Υπάρχει διαφοροποίηση που σχετίζεται με την ηλικία;

Με τη βοήθεια της R θα σχεδιάσουμε τα γράφημα θανάτων, ανά 100 χιλιάδες άτομα για καπνιστές και μη-καπνιστές και για κάθε ηλικία. Αυτό θα γίνει με τις εντολές:

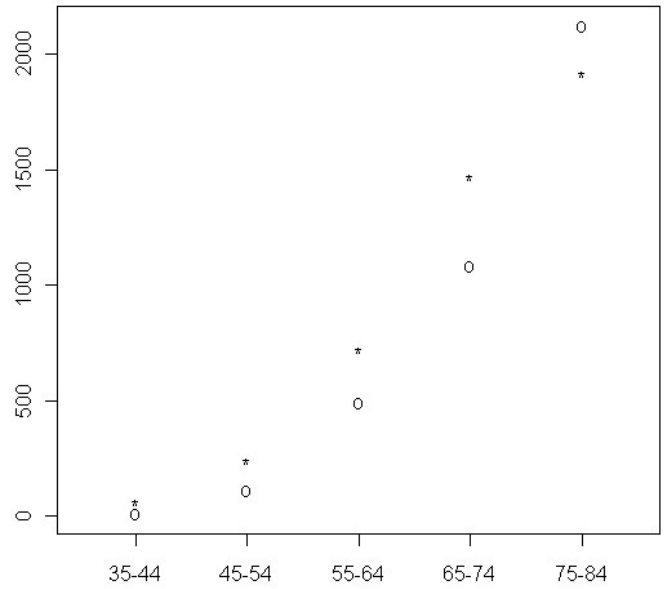
```
plot(plaisio$age,plaisio$deaths*100000/personyears,lty=0)

points(plaisio$age[smokers==”yes”],
plaisio$deaths[smokers==”yes”]*100000/personyears[smokers==”yes”],pch=”*”)

points(plaisio$age[smokers==”no”],
plaisio$deaths[smokers==”no”]*100000/personyears[smokers==”no”],pch=”o”).
```

Στο γράφημα να σχεδιαστούν με αστεράκια οι τιμές για καπνιστές και κύκλους οι τιμές για τους μη-καπνιστές.

Το αποτέλεσμα που δίνει η R φαίνεται στο Σχήμα 6.



Σχήμα 6: *Διάγραμμα που δείχνει την αναλογία θανάτων προς 100.000 άτομα ανά έτος για καπνιστές και μη-καπνιστές, όπου με αστεράκια οι τιμές για καπνιστές και κύκλους οι τιμές για τους μη-καπνιστές.*

Όπως βλέπουμε στο γράφημα, τα ποσοστά θανάτου στους καπνιστές, είναι γενικά μεγαλύτερα από ότι είναι στους μη-καπνιστές. Η διαφορά δεν αυξάνεται υπερβολικά με την αύξηση της ηλικίας. Διάφορα μοντέλα μπορούν να περιγράψουν τα δεδομένα μας. Ένα μοντέλο, είναι αυτό της μορφής:

$log(deaths_i) = log(population_i) + b_1 + b_2 smoke_i + b_3 agecat_i + b_4 agesq_i + b_5 smkage_i,$

όπου το i δηλώνει από ποιο διάστημα ηλικίας είναι και αν είναι καπνιστής ή όχι, δηλαδή:

i=1,...,5 για ηλικίες 35-44,...,75-84 και καπνιστές και

i=6,...,10 για ηλικίες 35-44,...,75-84 και μη-καπνιστές.

Ο όρος $deaths_i$ δηλώνει την απόκριση, ο όρος $population_i$ δηλώνει τον αριθμό γιατρών που κινδυνεύουν στην κάθε ομάδα i.

Για τους υπόλοιπους όρους:

$smoke_i$: είναι ίσο με 1 για καπνίζοντες και ίσο με 0 για μη-καπνίζοντες,

$agecat_i$: παίρνει τις τιμές 1,...,5 για ομάδες ηλικίας 35-44,...,75-84,

$agesq_i$: είναι το τετράγωνο του $agecat_i$ και

$smkage_i$: είναι ίσο με το $agecat_i$ για καπνιστές και 0 για μη-καπνιστές ώστε να μας δώσει ένα ποσοστό της διαφοράς με την αύξηση ηλικίας.

Αντίστοιχα στην R θα δημιουργήσουμε το μοντέλο ως εξής:

personyears<-log(personyears),

smokers<-ifelse(plaisio$smokers==”yes”,1,0),

age<-as.numeric(plaisio$age),

agesq<-age^2,

smkage<-ifelse(smokers==0,age,0),

Από τις τιμές που ορίσαμε, οι μεταβλητές: smokers, age, agesq,και smkage, είναι οι επεξηγηματικές μεταβλητές, των οποίων θέλουμε να εκτιμήσουμε τις παραμέτρους. Αντιθέτως, οι τιμές που έχουμε για personyears είναι ένας γνωστός επιπρόσθετος σταθερός όρος στην παράμετρο b_1. Στην R, για να ξεχωρήσουμε αυτόν τον όρο από τις υπόλοιπες επεξηγηματικές μεταβλητές, θα χρησιμοποιήσουμε την εντολή offset. Πιο συγκεκριμένα, τον όρο personyears θα τον ορίσουμε στη R ως: offset=personyears, αφού προηγουμένως έχουμε θέσει ότι το νέο personyears είναι το log(personyears), όπως φαίνεται και από τις προηγούμενες εντολές.

Θα δημιουργήσουμε το μοντέλο στην R, όπως φαίνεται στον Πίνακα 24, πληκτρολογώντας την εντολή:
montelo<-glm(deaths∼smokers+age+agesq+smkage,family=poisson,
offset=personyears).

```
> montelo

Call:  glm(formula = deaths ~ smokers + age + agesq + smkage, family = poisson,      offset = personyears)

Coefficients:
(Intercept)      smokers          age        agesq       smkage
   -10.7918       1.4410       2.3765      -0.1977      -0.3075

Degrees of Freedom: 9 Total (i.e. Null);  5 Residual
Null Deviance:      935.1
Residual Deviance: 1.635        AIC: 66.7
```

Πίνακας 24: *Αποτελέσματα της R για το γενικευμένο γραμμικό μοντέλο.*

Σημαντικά στοιχεία για το μοντέλο μας, θα μας δώσει η εντολή summary(montelo), όπως φαίνονται στον Πίνακα 25.

```
> summary(montelo)

Call:
glm(formula = deaths ~ smokers + age + agesq + smkage, family = poisson,
    offset = personyears)

Deviance Residuals:
       1         2         3         4         5         6         7         8
 0.43820  -0.27329  -0.15265   0.23393  -0.05700  -0.83049   0.13404   0.64107
       9        10
-0.41058  -0.01275

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.79176    0.45008 -23.978  < 2e-16 ***
smokers       1.44097    0.37220   3.872 0.000108 ***
age           2.37648    0.20795  11.428  < 2e-16 ***
agesq        -0.19768    0.02737  -7.223 5.08e-13 ***
smkage       -0.30755    0.09704  -3.169 0.001528 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 935.0673  on 9  degrees of freedom
Residual deviance:   1.6354  on 5  degrees of freedom
AIC: 66.703

Number of Fisher Scoring iterations: 4
```

Πίνακας 25: *Αποτελέσματα της R για την εντολή summary στο γενικευμένο γραμμικό μοντέλο.*

Ο Πίνακας 26 που ακολουθεί μας δείχνει τις εκτιμήσεις παραμέτρων στη μορφή αναλογιών $e^{\hat{b}_j}$.

Όρος	agecat	agesq	smoke	smkage
$\hat{b}$	2.376	-0.198	1.441	-0.308
$s.e.(\hat{b})$	0.208	0.027	0.372	0.097
Wald statistic	11.43	-7.22	3.87	-3.17
p-value	<0.001	<0.001	<0.001	<0.002
Rate ratio	10.77	0.82	4.22	0.74
95% δ.ε.	7.2, 16.2	0.78, 0.87	2.04, 8.76	0.61, 0.89

Πίνακας 26: *Εκτιμήσεις παραμέτρων που βρέθηκαν προσαρμόζοντας το μοντέλο στα δεδομένα.*

Ο στατιστικός έλεγχος Wald ελέγχει αν $b_j = 0$, όλες οι p-τιμές είναι πολύ μικρές και το 95% διάστημα εμπιστοσύνης για $e^{\hat{b}_j}$ δείχνει ότι όλοι οι όροι είναι σημαντικοί για το μοντέλο. Οι εκτιμήσεις δείχνουν ότι η επικινδυνότητα για θάνατο, από καρδιολογικά αίτια ήταν 4 φορές μεγαλύτερος για καπνιστές, παρά σε μη-καπνιστές, όπως δείχνει το rate ratio.

Ηλικία	καπνιστής	Θάνατοι	αναμενόμενοι θάνατοι	Pearson υπόλοιπα	υπόλοιπα απόκλισης
1	1	32	29.58	0.444	0.438
2	1	104	106.81	-0.272	-0.273
3	1	206	208.20	-0.152	-0.153
4	1	186	182.83	0.235	0.234
5	1	102	102.58	-0.057	-0.057
1	0	2	3.41	-0.766	-0.830
2	0	12	11.54	0.135	0.134
3	0	28	27.74	0.655	0.641
4	0	28	30.23	-0.405	-0.411
5	0	31	31.07	-0.013	-0.013

Πίνακας 27: *Παρατηρήσεις και εκτιμήσεις για αριθμούς θανάτων και τα υπόλοιπα του μοντέλου.*

Ο πίνακας 27 δείχνει τα ότι το μοντέλο προσαρμόζεται πολύ καλά στα δεδομένα μας. Ο αναμενόμενος αριθμός θανάτων είναι σχεδόν ίσος με τον αριθμό θανάτων που παρατηρούμε κι έτσι τα υπόλοιπα Pearson και τα υπόλοιπα απόκλισης είναι αρκετά μικρά.

Το άθροισμα τετραγώνων για τα υπόλοιπα Pearson και deviance είναι αντίστοιχα 1.550 και 1.635.

6.4.4 Δίαιτα υδατανθράκων

Τα δεδομένα του Πίνακα 28 που ακολουθεί δείχνουν τιμές συνολικών θερμίδων που λαμβάνονται από σύνθετους υδατάνθρακες, για είκοσι άνδρες διαβητικούς που χρησιμοποιούν ινσουλίνη και έκαναν δίαιτα υψηλών υδατανθράκων για έξι μήνες. Τα αποτελέσματα, πιστεύεται ότι σχετίζονται με την ηλικία, το βάρος του σώματος και άλλες συνιστώσες της έρευνας όπως το ποσοστό θερμίδων ως πρωτεΐνες. Αυτές οι μεταβλητές, χρησιμοποιούνται ως επεξηγηματικές μεταβλητές [2].

Υδατάνθρακες y	Ηλικία x_i	Βάρος x_2	Πρωτεΐνη x_3
33	33	100	14
40	47	92	15
37	49	135	18
27	35	144	12
30	46	140	15
43	52	101	15
34	62	95	14
48	23	101	17
30	32	98	15
38	42	105	14
50	31	108	17
51	61	85	19
30	63	130	19
36	40	127	20
41	50	109	15
42	64	107	16
46	56	117	18
24	61	100	13
35	48	118	18
37	28	102	14

Πίνακας 28: *Υδατάνθρακες, ηλικία, σχετικό βάρος και πρωτεΐνη για 20 άνδρες διαβητικούς.*

Θα γράψουμε τα δεδομένα μας στην R ως διανύσματα:

```
y<-c(33,40,37,27,30,43,34,48,30,38,50,51,30,36,41,42,46,24,35,37)
```

x1<-c(33,47,49,35,46,52,62,23,32,42,31,61,63,40,50,64,56,61,48,28)

x2<-c(100,92,135,144,140,101,95,101,98,105,108,85,130,127,109,107,117,100,118,102)

x3<-c(14,15,18,12,15,15,14,17,15,14,17,19,19,20,15,16,18,13,18,14).

Θέλουμε να εξετάσουμε αν οι μεταβλητές μας παρουσιάζουν κάποια συσχέτιση μεταξύ τους. Θα χρησιμοποιήσουμε τον δειγματικό συντελεστή συσχέτισης. Ο δειγματικός συντελεστής συσχέτισης εκτιμά το βαθμό στον οποίο οι μεταβλητές είναι γραμμικά συσχετισμένες. Το δειγματικό συντελεστή συσχέτισης τον χρησιμοποιούμε στην R με τη βοήθεια της εντολής cor(x,y). Έτσι για τις μεταβλητές x1, x2, x3 χρησιμοποιώντας την εντολή cor στην R θα πάρουμε τα εξής αποτελέσματα:

```
> cor(x1,x2)
[1] -0.04287871
> cor(x1,x3)
[1] 0.1934031
> cor(x2,x3)
[1] 0.1472595
```

Πίνακας 29: *Αποτελέσματα της R για τον συντελεστή συσχέτισης.*

Το παράδειγμά μας, είναι μία περίπτωση πολλαπλής παλινδρόμησης. Θα ξεκινήσουμε προσαρμόζοντας το μοντέλο:

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

στο οποίο οι υδρογονάνθρακες, συνδέονται γραμμικά με την ηλικία,το βάρος και την πρωτεΐνη.

Στην R η εντολή με την οποία θα δημιουργήσουμε το συγκεκριμένο μοντέλο, είναι η diaitamodel<-lm(y∼x1+x2+x3), η οποία παρουσιάζεται στον Πίνακα 30. Στον Πίνακα 31 παρουσιάζονται τα υπόλοιπα και στον Πίνακα 32 που ακολουθεί παρουσιάζονται τα μερικά υπόλοιπα για το μοντέλο που δημιουργήσαμε.

Στην περίπτωση αυτή:

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ . \\ . \\ Y_N \end{bmatrix},$$

```
> diaitamodel<-lm(y~x1+x2+x3)
> diaitamodel

Call:
lm(formula = y ~ x1 + x2 + x3)

Coefficients:
(Intercept)           x1           x2           x3
    36.9601      -0.1137      -0.2280       1.9577
```

Πίνακας 30: *Αποτελέσματα της R για το γενικευμένο μοντέλο με όνομα diaitamodel.*

```
> residuals(diaitamodel)
          1           2           3           4           5
-4.814975971 -0.005358449  1.153603108  3.360565803  0.825798562
          6           7           8           9          10
 5.615179589 -1.658448448  3.403140114 -10.342399622  2.348198044
         11          12          13          14          15
 7.908672497  3.159138722 -7.352727284 -6.609048135  5.211965771
         16          17          18          19          20
 5.389687464  6.845025089 -8.674325424 -4.836368399 -0.927323029
```

Πίνακας 31: *Αποτελέσματα της R για τα υπόλοιπα στο γενικευμένο μοντέλο με όνομα diaitamodel.*

```
> residuals(diaitamodel,"partial")
          x1          x2          x3
1  -3.3201319  -2.375190  -8.5346299
2  -0.1019834   4.258566  -1.7672998
3   0.8296255  -4.387219   5.2647995
4   4.6280572  -4.232412  -4.2745132
5   0.8428500  -5.855110  -0.9361428
6   4.9501729   7.826948   3.8532383
7  -3.4602187   1.921424  -5.3781023
8   6.0347478   5.614909   5.5566239
9  -8.7338792  -7.446579 -12.1043409
10  2.8199549   3.647897  -1.3714558
11  9.6308693   8.524319  10.0621563
12  1.4710448   9.019185   9.2280477
13 -9.2681739 -11.753462  -1.2838183
14 -5.9099385 -10.325731   1.4175734
15  4.7743118   5.599595   3.4500245
16  3.3605645   6.233352   5.5854587
17  5.7253130   5.408516  10.9562215
18 -10.3624193  -6.234540 -14.3516919
19  -5.0466697  -6.500895  -0.7251720
20  1.1359028   1.056428  -4.6469769
attr(,"constant")
[1] 37.6
```

Πίνακας 32: *Αποτελέσματα της R για τα μερικά υπόλοιπα στο γενικευμένο μοντέλο με όνομα diaitamodel.*

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ . & . & . & . \\ . & . & . & . \\ 1 & x_{N1} & x_{N2} & x_{N3} \end{bmatrix} \text{ και}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ . \\ . \\ \beta_3 \end{bmatrix}.$$

Για τα δεδομένα αυτά:

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 752 \\ 34596 \\ 82270 \\ 12105 \end{bmatrix} \text{ και}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 20 & 923 & 2214 & 318 \\ 923 & 45697 & 102003 & 14780 \\ 2214 & 102003 & 250346 & 35306 \\ 318 & 14780 & 35306 & 5150 \end{bmatrix}.$$

Οπότε, η λύση της $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}$, με προσέγγιση τεσσάρων δεκαδικών ψηφίων είναι:

$$\mathbf{b} = \begin{bmatrix} 36.9601 \\ -0.1137 \\ -0.2280 \\ 1.9577 \end{bmatrix} \text{ και}$$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 4.8158 & -0.0113 & -0.0188 & -0.1362 \\ -0.0113 & 0.0003 & 0.0000 & -0.0004 \\ -0.0188 & 0.0000 & 0.0002 & -0.0002 \\ -0.1362 & -0.0004 & -0.0002 & 0.0114 \end{bmatrix}.$$

Επίσης, $\mathbf{y}^T\mathbf{y} = 29368$, $N\bar{y}^2 = 28275.2$ και $\mathbf{b}^T\mathbf{X}^T\mathbf{y} = 28800.337$. Χρησιμοποιώντας την εξίσωση $\hat{\sigma}^2 = \frac{1}{N-p}(\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b})$, για να προσεγγίσουμε έναν αμερόληπτο εκτιμητή για το σ^2, θα βρούμε: $\hat{\sigma}^2 = 35.479$ και έτσι, έχουμε τα τυπικά σφάλματα για τα στοιχεία του **b** τα οποία φαίνονται στον Πίνακα 33 που ακολουθεί.

Η R μας βοηθάει να βρούμε τα τυπικά σφάλματα αλλά και άλλες μετρήσεις, όπως το συντελεστή προσδιορισμού και P-τιμές για τα b_i. Αυτό μπορούμε να το πετύχουμε με την εντολή summary.

Πρέπει να ελέγξουμε την κανονικότητα των σφαλμάτων, την ομοσκεδαστικότητα και την ανεξαρτησία των υπολοίπων.
Θα ελέγξουμε την κανονικότητα των σφαλμάτων δίνοντας στην R τις εντολές:

Όρος	Εκτίμηση	Τυπικό σφάλμα
Σταθερά	36.960	13.071
Συντελεστής για ηλικία	-0.114	0.109
Συντελεστής για βάρος	-0.228	0.083
Συντελεστής για πρωτεΐνη	1.958	0.635

Πίνακας 33: *Εκτιμήσεις για το μοντέλο.*

```
> summary(diaitamodel)

Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3424  -4.8203   0.9897   3.8553   7.9087

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.96006   13.07128   2.828  0.01213 *
x1          -0.11368    0.10933  -1.040  0.31389
x2          -0.22802    0.08329  -2.738  0.01460 *
x3           1.95771    0.63489   3.084  0.00712 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.956 on 16 degrees of freedom
Multiple R-squared: 0.4805,     Adjusted R-squared: 0.3831
F-statistic: 4.934 on 3 and 16 DF,  p-value: 0.01297
```

Πίνακας 34: *Αποτελέσματα της R για την εντολή summary.*

```
> confint(diaitamodel)
                2.5 %      97.5 %
(Intercept)  9.2501740 64.66993787
x1          -0.3454360  0.11808330
x2          -0.4045820 -0.05145268
x3           0.6117998  3.30362531
```

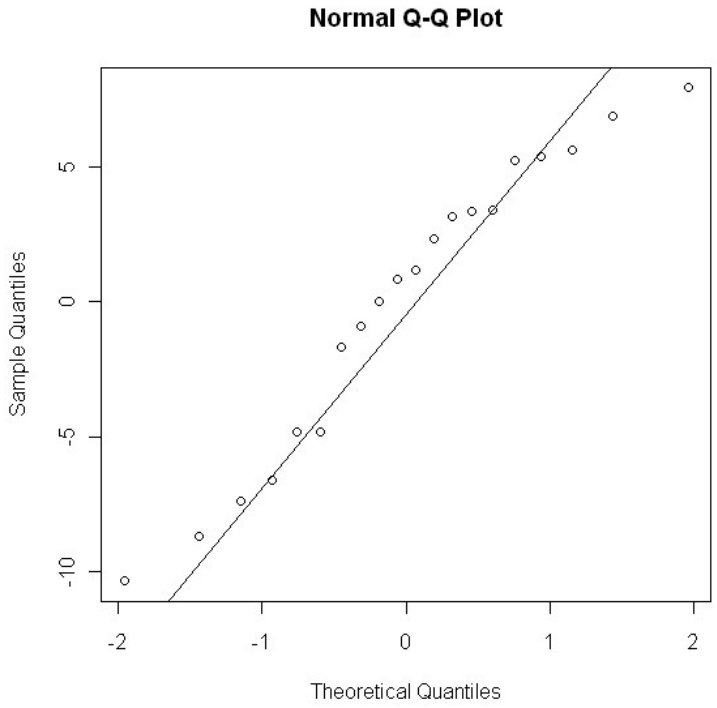
Πίνακας 35: *Συμμετρικό 95% Δ.Ε. για τις παραμέτρους.*

```
> predict(diaitamodel)
       1        2        3        4        5        6        7        8
37.81498 40.00536 35.84640 23.63943 29.17420 37.38482 35.65845 44.59686
       9       10       11       12       13       14       15       16
40.34240 35.65180 42.09133 47.84086 37.35273 42.60905 35.78803 36.61031
      17       18       19       20
39.15497 32.67433 39.83637 37.92732
```

Πίνακας 36: *Προβλεπόμενες τιμές για κάθε* x_i.

> qqnorm(residuals(diaitamodel)) και

> qqline(residuals(diaitamodel)).
Το αποτελέσμα που παίρνουμε φαίνονται στο Σχήμα 7:



Σχήμα 7: *Έλεγχος για κανονικότητα σφαλμάτων.*

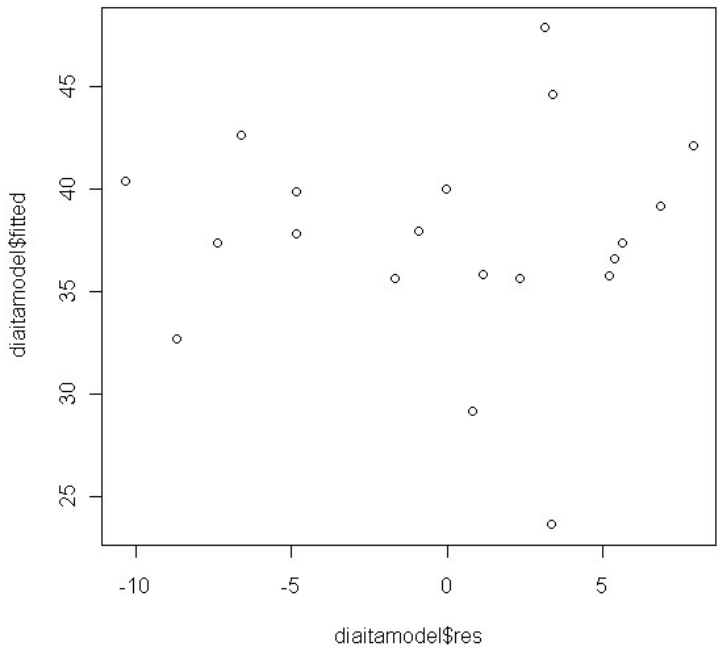
Στη συνέχεια θα ελέγξουμε την ομοσκεδαστικότητα. Αυτό θα το πετύχουμε με χρήση της εντολής > plot(diaitamodel$res,diaitamodel$fitted). Τα αποτελέσματα φαίνονται στο Σχήμα 8.

Τέλος, θα ελέγξουμε την ανεξαρτησία των σφαλμάτων, όπως φαίνεται στο Σχήμα 9, χρησιμοποιώντας την εντολή:
> plot(diaitamodel$res).

Από τους ελέγχους που πραγματοποιήσαμε, είδαμε ότι ισχύουν οι προϋποθέσεις του γενικού γραμμικού μοντέλου.

Για να εξηγήσουμε την χρήση της deviance, θα κάνουμε έλεγχο υποθέσεων, H_0 ότι η μεταβλητή απόκρισης δεν εξαρτάται από την ηλικία. Ο έλεγχος είναι δηλαδή:

$$H_0 : b_1 = 0 \quad \text{vs} \quad H_1 : b_1 \neq 0.$$



Σχήμα 8: *Έλεγχος για ομοσκεδαστικότητα.*

Στο Σχήμα 8 φαίνεται η γραφική παράσταση των υπολοίπων συναρτήσει των προβλεπόμενων τιμών. Τα ζεύγη αυτά στην συγκεκριμένη περίπτωση δεν εμφανίζουν κάποιο συστηματικό τρόπο συμπεριφοράς, οπότε μπορούμε να θεωρήσουμε ότι ισχύει η προϋπόθεση της ομοσκεδαστικότητας.

Το μοντέλο της μηδενικής υπόθεσης είναι το:

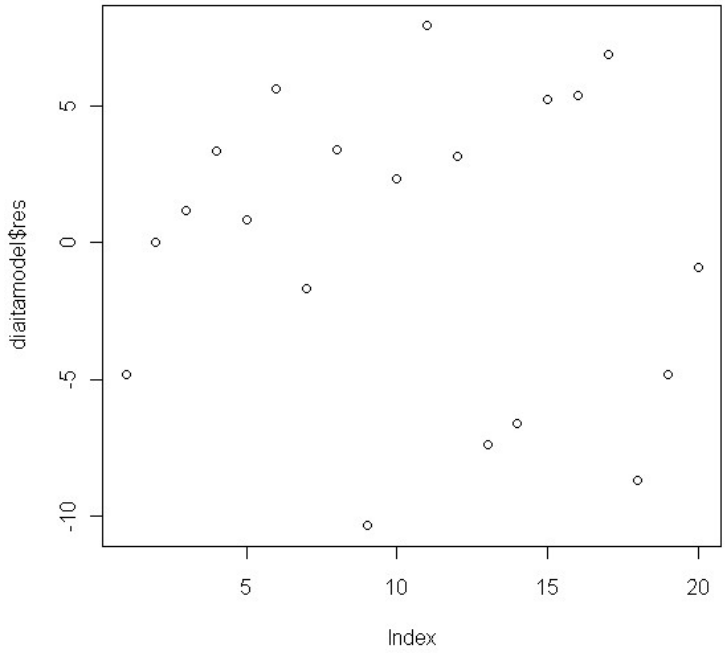
$$E(Y_i) = b_0 + b_2 x_{i2} + b_3 x_{i3}.$$

Ο νέος πίνακας για το $\mathbf{X}$ είναι ο πίνακας $\mathbf{X}$, αν αφαιρέσουμε την δεύτερη στήλη. Έτσι έχουμε:

$$\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 752 \\ 82270 \\ 12105 \end{bmatrix} \text{ και}$$

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 20 & 2214 & 318 \\ 2214 & 250346 & 35306 \\ 318 & 35306 & 5150 \end{bmatrix}.$$

Οπότε, η λύση της $\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y}$ είναι:



Σχήμα 9: *Έλεγχος για ανεξαρτησία των σφαλμάτων.*

$$\mathbf{b} = \begin{bmatrix} 33.130 \\ -0.222 \\ 1.824 \end{bmatrix}.$$

Για το μοντέλο $E(Y_i) = b_0 + b_2 x_{i2} + b_3 x_{i3}$, $\mathbf{b}^T \mathbf{X}^T \mathbf{y} = 28761.978$.

Στο Σχήμα 9 παρατηρούμε από τη γραφική παράσταση ότι τα υπόλοιπα συμπεριφέρονται με τυχαίο τρόπο, οπότε έχουμε ανεξαρτησία σφαλμάτων.

Στον Πίνακα 37 που ακολουθεί, παρουσιάζεται μία περίληψη του ελέγχου. Η τιμή $F = 38.36/35.48 = 1.08$ δεν είναι σημαντική, αν συγκριθεί με την F(1,16) κατανομή, έτσι, τα δεδομένα μας, δεν μας δίνουν στοιχεία, ώστε να απορρίψουμε την μηδενική υπόθεση. Έτσι μπορούμε να πούμε ότι η μεταβλητή απόκρισης δεν φαίνεται να εξαρτάται από την ηλικία. Ο Πίνακας 37 είναι ο πίνακας ανάλυσης διασποράς, για τα δύο μοντέλα που ελέγχουμε.

Παρατηρούμε ότι οι παράμετροι που εκτιμούμε για κάθε μοντέλο, διαφέρουν. Για παράδειγμα, ο συντελεστής της πρωτεΐνης είναι 1.958 για το μοντέλο που έχει τον όρο για την ηλικία, όπως φαίνεται από τον Πίνακα 34, αλλά 1.824 όταν παραλείπεται ο όρος της ηλικίας όπως μπορούμε να δούμε στον Πίνακα 38. Αυτό είναι ένα παράδειγμα έλλειψης της ορθογωνιότητας.

Πηγή διακύμανσης	Βαθμοί ελευθερίας	Άθροισμα τετραγώνων	Μέσο άθροισμα τετραγώνων
Μοντέλο H_0	3	28761.978	
Βελτίωση από μοντέλο H_1	1	38.359	38.36
Υπόλοιπα	16	567.663	35.48
Σύνολο	20	29368.000	

Πίνακας 37: *Πίνακας ανάλυσης διασποράς.*

Θα ελέγξουμε με τη βοήθεια της R αν η μεταβλητή απόκρισης εξαρτάται από την ηλικία. Έχουμε ορίσει προηγουμένως το μοντέλο μας με την κατάλληλη εντολή στην R. Θα ορίσουμε και το άλλο μοντέλο ως:

diaitamodel_h0<-lm(y∼x2+x3),

όπου παραλείψαμε το x1 σε σχέση με το προηγούμενο μοντέλο ($H_0 : b_1 = 0$).

Τα αποτελέσματα που παίρνουμε είναι:

```
> diaitamodel_h0<-lm(y~x2+x3)
> diaitamodel_h0

Call:
lm(formula = y ~ x2 + x3)

Coefficients:
(Intercept)           x2           x3
    33.1303      -0.2216       1.8243
```

Πίνακας 38: *Αποτελέσματα για το μοντέλο, όπου δεν έχει την ηλικία.*

Θα δώσουμε την εντολή για να δημιουργήσει η R τον πίνακα ανάλυσης διακύμανσης:

```
> elegxos<-anova(diaitamodel_h0,diaitamodel)
> elegxos
Analysis of Variance Table

Model 1: y ~ x2 + x3
Model 2: y ~ x1 + x2 + x3
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     17 606.02
2     16 567.66  1    38.359 1.0812 0.3139
```

Πίνακας 39: *Δημιουργία πίνακα ανάλυσης διακύμανσης στη R.*

Από τα αποτελέσματα που μας επιστρέφει η R, μπορούμε να δούμε το άθροισμα τετραγώνων υπολοίπων και την τιμή F μεταξύ άλλων και επιβεβαιώνουμε ότι η μεταβλητή απόκρισης δεν εξαρτάται από την ηλικία.

ΕΠΙΛΟΓΟΣ

Η εργασία αυτή είχε ως σκοπό την εισαγωγή στα γενικευμένα γραμμικά μοντέλα και την χρήση του στατιστικού πακέτου R. Αφού αναπτύχθηκε η βασική θεωρία γύρω από τα γενικευμένα γραμμικά μοντέλα, παρουσιάστηκαν εντολές του στατιστικού πακέτου R γύρω από τα γενικευμένα γραμμικά μοντέλα. Τέλος, έγιναν κάποιες εφαρμογές πολλαπλής παλινδρόμησης, παλινδρόμησης Poisson και λογιστικής παλινδρόμησης στην R, όπου είδαμε στην πράξη πώς χρησιμοποιούνται οι εντολές αυτές και πως ερμηνεύονται τα αποτελέσματα που μας επιστρέφει η R.

Αναφορές

[1] Breslow Day (1987). Statistical Methods in Cancer Research. Vol 2. The Design and Analysis of Cohort Studies. (IARC Scientific Publications No. 82), Lyon, IARC.

[2] Dobson A. J. (2002). An Introduction To Generalized Linear Models. Second Edition. Chapman & Hall/CRC.

[3] Dunteman G. H. & Moon-Ho R. Ho (2006). An Introduction To generalised Linear Models. Sage Publications.

[4] Everitt B. S. and Hothorn T. (2010). A Handbook of Statistical Analyses Using R. Second Edition. Chapman & Hall / CRC.

[5] Faraway J. J. (2006). Extending The Linear Model with R. Chapman & Hall / CRC.

[6] Hosmer D. W. and Lemeshov S. (2000). Applied Logistic Regression. Second Edition. John Wiley & Sons, INC.

[7] McCullagh P. and Nelder J. A. (1989). Generalized Linear Models. Second Edition. Chapman and Hall.

[8] McCullogh C. E. (2001). Generalized, Linear, and Mixed Models. John Wiley & Sons, INC.

[9] Venables W. N., Smith D. M. and the R Development Core Team (1999). An Introduction to R.

[10] Κοκολάκης Γ. & Φουσκάκης Δ. (2009). Στατιστική Θεωρία & Εφαρμογές. Εκδόσεις Συμεών.

[11] Φουσκάκης Δ. Σημειώσεις του Μαθήματος Ανάλυση Δεδομένων με Η/Υ.