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

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ ΣΤΗΝ

ΑΝΑΛΟΓΙΣΤΙΚΗ ΕΠΙΣΤΗΜΗ ΚΑΙ ΔΙΟΙΚΗΤΙΚΗ ΚΙΝΔΥΝΟΥ

ΤΙΜΟΛΟΓΗΣΗ ΝΟΣΟΚΟΜΕΙΑΚΩΝ ΠΡΟΓΡΑΜΜΑΤΩΝ ΜΕ ΤΗΝ ΧΡΗΣΗ ΓΕΝΙΚΕΥΜΕΝΩΝ ΓΡΑΜΜΙΚΩΝ ΜΟΝΤΕΛΩΝ

Αγρίτη Βασιλική

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και

Ασφαλιστικής Επιστήμης του Πανεπιστημίου

Πειραιώς ως μέρος των απαιτήσεων για την

απόκτηση του Μεταπτυχιακού Διπλώματος

Ειδίκευσης στην Αναλογιστική Επιστήμη και Διοικητική Κινδύνου.

Πειραιάς

Δεκέμβιος 2017

Στην οικογένεια μου,

Μαρία, Γιάννης και Νικολία

Περίληψη

Σκοπός αυτής της διπλωματικής εργασίας είναι η μελέτη των γενικευμένων γραμμικών μοντέλων εφαρμόζοντας αυτα σε αληθινά δεδομένα. Για την περάτωση αυτης της διπλωματικής εργασίας χρησιμοποιήθηκαν ασφαλιστικά δεδομένα μετά απο συλλογή τους και ανάλυσή τους απο τους Piet de Jong και Gillian Z. Heller για τις ανάγκες του βιβλίου Generalized Linear Models for Insurance Data. Αυτά τα δεδομένα αναφέρθηκαν στην Αυστραλία κατα τη χρονική περίοδο απο τον Ιούλιο του 1989 έως το τέλος του 1999.

Η εργασία αυτή χωρίζεται σε δύο μέρη. Στο πρώτο μέρος παρουσιάζονται τα γενικευμένα γραμμικά μοντέλα και η θεωρία τους για την κατανόηση και την περαιτέρω εφαρμογη τους στην συνέχεια. Εστιάζουμε ιδιαίτερα στις μεταβλητές απόκρισης που είναι διακριτές και αναφέρονται σε δίτιμα δεδομένα (binary data).

Στο δεύτερο μέρος γίνεται η μελέτη των δεδομένων χρησιμοποιώντας το στατιστικό πακέτο λογισμικού R. Για τις ανάγκες της ανάλυσης θα επιχειρηθεί να επιλεγεί το βέλτιστο μοντέλο εφαρμόζοντας την λογιστική παλινδρόμηση στα δεδομένα, όπου η μεταβλητή απόκρισης ‘Αποζημίωση’, μετασχηματίστηκε σε μία δίτιμη μεταβλητή. Ως επεξηγηματικές μεταβλητές θα θεωρηθούν:

- ο βαθμός τραυματισμού, όπως κωδικοποιήθηκε, σε τρίτιμη μεταβλητή με επίπεδα, ‘χαμηλός’, ‘σοβαρός’ και ‘θάνατος’,
- η κατηγορική μεταβλητή νομική εκπροσώπηση και
- η μεταβλητη καθυστέρησης διακανονισμού.

Αρχικά θα εξετασθεί η σημαντικότητα της εισαγωγής του πρώτου όρου-μεταβλητής, με χρήση ελέγχου X^2, του κριτηρίου AIC και του κριτηρίου BIC για να διερευνηθεί κατά πόσο οι νέες μεταβλητές βελτιώνουν την εκτίμηση. Εν συνεχεία και αφού καταλήξουμε στην επιλογή του βέλτιστου μοντέλου με την παραπάνω διαδικασια, θα επιχειρηθεί να εξεταστεί εάν η επιλογή του παραπάνω υποδείγματος είναι ικανοποιητική, χρησιμοποιώντας την μέθοδο Stepwise και Backwards selection με βάση τα κριτήρια AIC και BIC. Τέλος, γίνεται η ερμηνεία των επιλεχθέντων μοντέλων και συγκρίνονται με τα αποτελέσματα που παρουσίασαν οι de Jong και Heller.

Abstract

The purpose of this thesis is to study the theory of generalized linear models by applying them to real data. To complete this thesis, insurance data were used after being collected and analyzed by Piet de Jong and Gillian Z. Heller for the need of the book Generalized Linear Models for Insurance Data. These data were reported in Australia during the period from July 1989 to the end of 1999.

This thesis is divided into two parts. In the first part, we present the generalized linear models and their theoretical background in order to understand them and apply them. We mainly focus on response variables that are distinct and refer to binary data.

In the second part we analyze the data using the statistical software package R. For the analysis needs, we will try to select the optimal model by applying the logistic regression to the data, where the response variable 'Claims' was transformed into a two-tier variable. As explanatory variables will be considered:

- the degree of injury, as encoded, to a categorical variable with levels, 'low', 'severe' and 'death',
- the categorical variable legal representation and
- the variable settlement delay.

Initially, the importance of entering the first variable will be tested, using X^2 test, AIC criterion and BIC criterion, in order to examine whether the new variables improve the estimation. Subsequently, and after ending up to the optimal model selection with the above process, we will attempt to examine whether the selection of the above model is appropriate, using the Stepwise and Backwards method based on the AIC and BIC criteria. Finally, the interpretations of the selected models are made and compared with the results presented by de Jong and Heller.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Εισαγωγή

Τα γενικευμένα γραμμικά μοντέλα αναπτύχθηκαν από τους John Nelder και Robert Wedderburn το 1972. Το μεγαλύτερο μέρος της θεωρίας των γενικευμένων γραμμικών μοντέλων δεν αποτελεί κάτι καινούργιο στο χώρο της στατιστικής. Αποτελεί 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. Ακόμα, η συγκεκριμένη ομαδοποίηση δημιούργησε προϋποθέσεις για περαιτέρω μελέτη νέων τεχνικών και σε συνδυασμό με την χρήση υπολογιστικών προγραμμάτων μπορούμε να μελετήσουμε δύσκολα προβλήματα που δε μπορούσαμε να μελετήσουμε πριν τη χρήση των γενικευμένων γραμμικών μοντέλων.

Τα Γενικευμένα Γραμμικά Μοντέλα (Generalized Linear Models) είναι φυσική γενίκευση των κλασσικών γραμμικών μοντέλων. Τα Γενικευμένα Γραμμικά Μοντέλα περιλαμβάνουν σαν ειδική περίπτωση την γραμμική παλινδρόμηση, την ανάλυση διασποράς, τα logit και probit μοντέλα, τα λογαριθμογραμμικά και τα πολυωνυμικά μοντέλα, καθώς και κάποια μοντέλα της ανάλυσης επιβίωσης. Αποδεικνύεται ότι αυτά τα μοντέλα μοιράζονται κάποιες κοινές ιδιότητες, καθώς και ότι έχουν κοινή μέθοδο εκτίμησης παραμέτρων. Οι κοινές αυτές ιδιότητες μας επιτρέπουν να μελετήσουμε μέσω των Γενικευμένων Γραμμικών Μοντέλων (ΓΓΜ) μία ευρεία ομάδα στατιστικών μοντέλων παρά το καθένα από αυτά ξεχωριστά.

Το κλασικό μοντέλο παλινδρόμησης με k ερμηνευτικές μεταβλητές έχει την εξής μορφή:

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i$$

ή ισοδύναμα η παραπάνω σχέση γράφεται

$$E(Y_i) = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}.$$

Το κλασικό μοντέλο παλινδρόμησης εκφράζει τη σχέση ανάμεσα στη μέση απόκλιση και τις ερμηνευτικές μεταβλητές. Βασική υπόθεση στο μοντέλο είναι ότι

$$\varepsilon_i \sim N(0, \sigma^2)$$

και τα ε_i, που συμβολίζουν τα σφάλματα, είναι ανεξάρτητα ανά δύο. Συνεπώς, τα Y_i ακολουθούν την κανονική κατανομή με σταθερή διακύμανση.

Στην πράξη, κάποια από τις δύο υποθέσεις μπορεί να μην ισχύει και γι' αυτό αναζητήθηκαν εναλλακτικές μέθοδοι.

Το 1972 οι Nelder & Wedderburn παρουσίασαν μία ενοποιημένη θεωρία για γραμμικά μοντέλα, που δεν απαιτεί την υπόθεση της κανονικότητας για τη μεταβλητή απόκρισης. Σύμφωνα με αυτήν,

- Τα γραμμικά μοντέλα μπορούν να μελετηθούν ενιαία κάτω από την υπόθεση ότι η κατανομή της μεταβλητής απόκρισης ανήκει στην εκθετική οικογένεια κατανομών.
- Για όλες τις κατανομές μέσα στην οικογένεια αυτή, οι εκτιμητές μεγίστης πιθανοφάνειας (ε.μ.π.) των παραμέτρων του μοντέλου μπορούν να βρεθούν με τον ίδιο αλγόριθμο.

1.2 Μονοπαραμετρική Εκθετική Οικογένεια Κατανομών

Έστω μία τυχαία μεταβλητή Y της οποίας η συνάρτηση πυκνότητας πιθανότητας εξαρτάται από μία παράμετρο θ. Συμφωνα με την Annete J. Dobson (Dobson, 2010) η κατανομή ανήκει στην εκθετική οικογένεια, αν μπορεί να γραφεί στην ακόλουθη μορφή:

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)},\ y \in Y, \theta \in \Theta, \quad (1)$$

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

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

Από την τελευταία σχέση παρατηρούμε την συμμετρία μεταξύ των y και θ. Επίσης, επισημαίνεται ότι το στήριγμα της συνάρτησης πυκνότητας πιθανότητας, δηλαδή το σύνολο $S = \{y: f(y;\Theta) > 0\}$ πρέπει να είναι ανεξάρτητο από την παράμετρο θ.

Στην περίπτωση όπου $a(y) = y$ μπορούμε να πούμε ότι η κατανομή είναι σε κανονική μορφή και το $b(\theta)$ ονομάζεται φυσική παράμετρος της κατανομής.

Σημείωση: Η κανονική κατανομή , η κατανομή Poisson και η Διωνυμική κατανομή είναι κάποιες από τις πολύ γνωστές κατανομές που ανήκουν στην εκθετική οικογένεια.

1.2.1 Ιδιότητες της Εκθετικής Οικογένειας Κατανομών

Στο σημείο αυτό θα δώσουμε κάποιες εκφράσεις για τη μέση τιμή και τη διασπορά του $a(y)$ όπως αυτή δώθηκε στην μορφή της εκθετικής κατανομής (1). Η διαδικασία που θα ακολουθήσουμε είναι όμοια με αυτή που εφαρμόζεται για τον υπολογισμό κάθε συνάρτησης πυκνότητας πιθανότητας.

Από τον ορισμό της συνάρτησης πιθανότητας, ολοκληρώνοντας για όλες τις τιμές του y και θεωρώντας ότι η τυχαία μεταβλητή Y είναι συνεχής έχουμε

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

Παραγωγίζοντας και τα δύο μέλη της παραπάνω σχέσης ως προς θ, θα καταλήξουμε στη σχέση:

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

Ισοδύναμα, παραγωγίζοντας δύο φορές και τα δύο μέλη της ίδιας σχέσης ως προς θ, θα καταλήξουμε στη σχέση:

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

Στη συνέχεια από τη γενική μορφή που δώσαμε για τις κατανομές που ανήκουν στην εκθετική κατανομή

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

και αφού παραγωγίσουμε ώστε να έχουμε

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

προκύπτει ότι

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

Η παραπάνω σχέση μπορεί να απλοποιηθεί και να γραφεί ως εξής

$$b'(\theta)E[\alpha(Y)] + c(\theta) = 0.$$

Ή ισοδύναμα, έχουμε

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

Ακολουθώντας παρόμοια διαδικασία καταλήγουμε στην ακόλουθη σχέση για την $Var[\alpha(Y)]$

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

Εκτός των παραπάνω, χρειάζεται επίσης να υπολογίσουμε εκφράσεις για την μέση τιμή και τη 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 :

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

Παραγωγίζοντας την παραπάνω συνάρτηση ως προς θ προκύπτει ότι

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

Η συνάρτηση $u(\theta; y)$ που ορίσαμε ως $u(\theta; y) = \frac{dl(\theta;y)}{d\theta}$ ονομάζεται score συνάρτησης και καθώς εξαρτάται από το y, μπορεί να θεωρηθεί ως τυχαία μεταβλητή και ορίζεται ακολούθως

$$u(\theta; Y) = U = \alpha(Y)b'(\theta) + c'(\theta),$$

και η μέση τιμή αυτής είναι

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

Η διασπορά της U ονομάζεται πληροφορία και θα τη συμβολίζουμε με I. Η πληροφορία λοιπόν δίνεται από τη σχέση

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

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

1.3 Γενικευμένα γραμμικά μοντέλα

Για τις ανάγκες μίας στατιστικής μελέτης είναι απαραίτητη πολλές φορές η πρόβλεψη των τιμών μίας μεταβλητής, η οποία ονομάζεται μεταβλητή απόκρισης, μέσω κάποιων γνωστών μεταβλητών που ονομάζονται επεξηγηματικές μεταβλητές. Στατιστικό μοντέλο ονομάζεται η δημιουργία μίας μαθηματικής σχέσης μεταξύ αυτών των μεταβλητών.

Για την πρόβλεψη της μεταβλητής απόκρισης ακολουθείται η παρακάτω διαδικασία:

1. Αρχικά, γίνεται ο προσδιορισμός του μοντέλου. Για να προσδιοριστεί ένα μοντέλο χρειάζεται μία εξίσωση που συνδέει τη μεταβλητή απόκρισης με την επεξηγηματική μεταβλητή και η κατανομή που ακολουθεί η κατανομή απόκρισης.

2. Γίνεται η εκτίμηση των παραμέτρων που χρησιμοποιούνται στο μοντέλο.
3. Δημιουργούμε διαστήματα εμπιστοσύνης και κάνουμε ελέγχους υποθέσεων για τις παραμέτρους του μοντέλου.
4. Ερμηνεύουμε τις τιμές των αποτελεσμάτων και ελέγχουμε την επάρκεια του μοντέλου. Η επάρκεια του μοντέλου μας αναφέρεται στο πόσο καλά ερμηνεύονται τα δεδομένα από το μοντέλο μας.

Σε ένα στατιστικό μοντέλο, επιθυμούμε η μεταβλητή απόκρισης δοθέντων των επεξηγηματικών μεταβλητών να ακολουθεί την κανονική κατανομή. Σε πολλές περιπτώσεις, η παραπάνω υπόθεση δεν ισχύει καθώς η μεταβλητή απόκρισης μπορεί για παράδειγμα να παίρνει τις τιμές 0 ή 1, που συμβολίζουν την αποτυχία και την επιτυχία αντίστοιχα, και για το λόγο αυτό θεωρούμε 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 κατανομών. (Για περισσότερες πληροφορίες σχετικά με την εκθετική οικογένεια κατανομών ανατρέξτε σε προηγούμενες παραγράφους.)

Σε ένα γενικευμένο γραμμικό μοντέλο θεωρούμε ένα σύνολο ανεξάρτητων τυχαίων μεταβλητών $Y_1, \dots, Y_N$ όπου κάθε μία από τις οποίες ακολουθεί μία κατανομή που ανήκει στην εκθετική οικογένεια κατανομών. Αυτές οι ανεξάρτητες τυχαίες μεταβλητές έχουν τις εξής 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 ίδια μορφής, δηλαδή όλες ανήκουν στην ίδια κατανομή.

   Η από κοινού συνάρτηση πυκνότητας πιθανότητας των $Y_1, \dots, Y_N$ είναι :

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

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

Για τον προσδιορισμό του μοντέλου, συνήθως ενδιαφερόμαστε για ένα σύνολο παραμέτρων $\mathbf{b} = (b_1, \dots, b_p)$, όπου $p < N$. Υποθέτουμε ότι $E(Y_i) = \mu_i$, όπου μ_i είναι συνάρτηση του θ_i.

Για ένα γενικευμένο γραμμικό μοντέλο υπάρχει μετασχηματισμός του μ_i, τέτοιος ώστε:

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

Στην παραπάνω εξίσωση θεωρούμε ότι η g είναι μία μονότονη και διαφορίσιμη συνάρτηση η οποία ονομάζεται συνάρτηση σύνδεσης και το $\mathbf{x}$ το οποίο είναι ένα διάνυσμα επεξηγηματικών μεταβλητών και ορίζεται ως εξής:

$$\mathbf{x_i^T} = [x_{i1} \ \dots \ x_{ip}]$$

και το b είναι ένα $\mathrm{p} \times 1$ διάνυσμα με παραμέτρους

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

1.3.1 Συναρτήσεις σύνδεσης (link functions)

Έστω ένα διάνυσμα $y = (y_1, \dots, y_N)$ το οποίο αποτελείται από N το πλήθος στοιχεία μίας τυχαίας μεταβλητής Y της οποίας οι συνιστώσες είναι ανεξάρτητα κατανεμημένες τυχαίες μεταβλητές με μέσο $\boldsymbol{\mu} = (\mu_1, \dots, \mu_N)$. Ο καθορισμός του $\boldsymbol{\mu}$ γίνεται με έναν μικρό αριθμό αγνώστων παραμέτρων $b_1, \dots, b_p$, όπου $p < N$. Οι μεταβλητές $x_1, \dots, x_p$ δημιουργούν μια γραμμική πρόβλεψη, η οποία είναι η $\eta = \sum_{j=1}^{p} x_j b_j$.

Στο κλασσικο μοντέλο παλινδρόμησης έχουμε $\mu = \eta$, δηλαδή έχουμε την ταυτοτική συνάρτηση ως συνάρτηση σύνδεσης.

Συνοψίζοντας, μπορούμε να πούμε ότι ένα γενικευμένο γραμμικό μοντέλο αποτελείται από τις ακόλουθες συνιστώσες:

1. Ένα σύνολο ανεξάρτητων τυχαίων μεταβλητών $Y_1, \dots, Y_N$ με κατανομή από την εκθετική οικογένεια.
2. Ένα $\mathrm{p} \times 1$ διάνυσμα με παραμέτρους

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

3. Επεξηγηματικές μεταβλητές

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1^T \\ \vdots \\ \boldsymbol{X}_N^T \end{bmatrix}.$$

4. Μία γνήσια μονότονη και διαφορίσιμη συνάρτηση σύνδεσης g τέτοια ώστε

$$g(\mu_i) = \mathbf{x_i^T}\,\mathbf{b}, \qquad \text{με} \quad \mu_i = E(Y_i).$$

Η συνάρτηση σύνδεσης, link function, συσχετίζει τη γραμμική παράμετρο με την αναμενόμενη τιμή μ της μεταβλητής απόκρισης y. Στα κλασικά γραμμικά μοντέλα η μέση τιμή μ ταυτίζεται με τη γραμμική πρόβλεψη. Επομένως, είναι φανερό 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, πρέπει να ισχύει $\mu > 0$. Μοντέλα με τέτοιου είδους μεταβλητές, εκφράζονται με τη λογαριθμική συνάρτηση σύνδεσης, $\eta = \log(\mu)$, ώστε να υπάρχει γραμμική σχέση. Το μ σε αυτές τις περιπτώσεις πρέπει να είναι θετικό.

Για την περίπτωση της διωνυμικής κατανομής, θεωρούμε τρεις βασικές συναρτήσεις σύνδεσης, οι οποίες είναι οι ακόλουθες:

- logit: $\eta = \log\left(\frac{\mu}{1-\mu}\right)$
- 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, ανά περίπτωση, έχουν την ακόλουθη μορφή:

- Κανονική: $\eta = \mu$.
- Γάμμα: $\eta = \mu^{-1}$.
- Poisson: $\eta = \log(\mu)$.
- Διωνυμική: $\eta = \log\left(\frac{p}{1-\mu}\right)$.

Σε ένα Γενικευμένο Γραμμικό Μοντέλο (ΓΓΜ), η συνάρτηση g καλείται συνάρτηση σύνδεσης και συνδέει το στοχαστικό τμήμα του μοντέλου με το μη στοχαστικό τμήμα. Το στοχαστικό τμήμα του μοντέλου είναι η μέση τιμή της τυχαίας μεταβλητής (τ.μ.) Υ ενώ ως μη στοχαστικό τμήμα του μοντέλου νοείται ο γραμμικός συνδυασμός των ερμηνευτικών μεταβλητών X_j.

Πιο συγκεκριμένα, έστω ότι συμβολίζουμε με $\mu_i = E(Y_i)$ τη μέση τιμή της μεταβλητής απόκρισης. Υποθέτουμε ότι αυτή εξαρτάται από τις τιμές των X_j για $j = 1,2,..,k$.

Στη συνέχεια, θεωρούμε τη γραμμική συνάρτηση πρόβλεψης

$$\eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}$$

όπου X_{ij} είναι η τιμή της μεταβλητής X_j για την παρατήρηση i.

Επομένως, η συνάρτηση σύνδεσης συνδέει τη μέση τιμή της μεταβλητής απόκρισης με τη παραπάνω συνάρτηση πρόβλεψης ως εξής:

$$\eta_i = g(\mu_i) = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}.$$

Μία ειδική περίπτωση συνάρτησης σύνδεσης ορίζεται από την ακόλουθη σχέση

$$g = (b')^{-1}$$

η οποία αποτελεί την αντίστροφη συνάρτηση της παραγώγου της b. Στην περίπτωση αυτή η g ονομάζεται κανονική συνάρτηση σύνδεσης (canonical link function).

Σύμφωνα με τα παραπάνω παρατηρούμε ότι για την κανονική κατανομή η b' είναι η ταυτοτική συνάρτηση επομένως και η κανονική συνάρτηση σύνδεσης είναι η ταυτοτική συνάρτηση.

1.3.2 Περίπτωση Δίτιμων δεδομένων

Πολλές φορές συναντώνται περιπτώσεις κατά τις οποίες η μεταβλητή απόκρισης είναι διακριτή για το μοντέλο το οποιο εξετάζεται. Ενα απλό παράδειγμα διακριτής κατανομής από την εκθετική οικογένεια είναι η διωνυμική κατανομή.

Μια διάκριση που μπορεί να γίνει για ενα γενικευμένο γραμμικό μοντέλο, είναι όταν

- Τα δεδομένα να είναι ομαδοποιημένα. Στην περίπτωση αυτή αναφερόμαστε συνήθως σε διωνυμικά δεδομένα (binomial data).
- Τα δεδομένα να μην είναι ομαδοποιημένα. Στην περίπτωση αυτή γνωρίζουμε για κάθε άτομο στο δείγμα την τιμή της απόκρισης (0=αποτυχία, 1= επιτυχία), επομένως αναφερόμαστε σε δίτιμα δεδομένα (binary data).

Για δίτιμα δεδομένα, η κατανομή πιθανότητας των Y_i είναι

$$P(Y_i = y_i) = {p_i}^{y_i}(1 - p_i)^{1-y_i}, y_i = 0,1.$$

Αν θεωρήσουμε ότι προσεγγιστικά η μεταβλητή Y_i για το άτομο i στο δείγμα ακολουθεί κατανομή $N(p_i, \sigma^2)$, τότε

$$p_i = E(Y_i) = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}.$$

με αποτέλεσμα να μπορούν να εκτιμηθούν οι παράμετροι β_i με μεθόδους συνήθους γραμμικής παλινδρόμησης.

Διάφορα προβλήματα όμως προκύπτουν με την παραπάνω προσέγγιση:

- Δεν καθίσταται δυνατό να χρησιμοποιηθεί η κανονική προσέγγιση στη διωνυμική κατανομή $B_i(n, p_i)$ όταν $n = 1$.
- Η διακύμανση της Y_i είναι $p_i \cdot (1 - p_i)$ και άρα δεν είναι σταθερή.
- Η εκτιμώμενη τιμή

$$\hat{p}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j X_{ij}$$

μπορεί να μην ανήκει στο διάστημα $[0,1]$ όπως θα έπρεπε. (Πολίτης Κ, 2014.)

Συνεπώς, θα χρησιμοποιηθεί σα μεταβλητή απόκρισης ένας μετασχηματισμό της μέσης τιμής της μεταβλητής Y,

$$\eta_i = g(p_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

όπου η g είναι μία συνάρτηση που απεικονίζει το διάστημα $[0,1]$ στην πραγματική ευθεία, έτσι ώστε

$$g^{-1}(\eta_i) = p_i \in [0,1].$$

Επομένως, η συνάρτηση g είναι η συνάρτηση σύνδεσης.

Για δίτιμα ή διωνυμικά δεδομένα χρησιμοποιούνται οι ακόλουθες τρεις συναρτήσεις σύνδεσης:

1. Logit :

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

2. Probit:

$$\eta_i = Probit(p_i) = \Phi^{-1}(p_i),$$

όπου Φ είναι η αθροιστική συνάρτηση κατανομής της τυποποιημένης κανονικής.

3. Complementary log-log ή cloglog :

$$\eta_i = log[-log(1-p_i)].$$

<u>Παρατηρήσεις:</u>

- Αν χρησιμοποιείται η συνάρτηση logit ως συνάρτηση σύνδεσης τη συνάρτηση logit τότε το μοντέλο στο οποίο αναφερόμαστε είναι αυτό της λογιστικής παλινδρόμησης.
- Τα μοντέλα με τις συναρτήσεις σύνδεσης logit και probit δίνουν αρκετά παρόμοια αποτελέσματα, τα οποία με μία αλλαγή κλίμακας μπορούν να γίνουν σχεδόν ταυτόσημα.

- Οι συναρτήσεις logit και probit είναι και οι δύο συμμετρικές για p και $1-p$, ενώ η συνάρτηση complementary log-log (c log-log) δεν είναι. Εφόσον οι συναρτήσεις είναι συμμετρικές δεν έχει σημασία ποιο από τα δύο αποτελέσματα της δίτιμης μεταβλητής θεωρούμε ως «επιτυχία» ή «αποτυχία».
- Η συνάρτηση σύνδεσης c log-log χρησιμοποιείται κυρίως σε περιπτώσεις όπου η πιθανότητα να συμβεί ένα γεγονός είναι πολύ μικρή ή πολύ μεγάλη.
- Οι τρεις αυτές συναρτήσεις είναι οι πιο διαδεδομένες που χρησιμοποιούνται στην πράξη. Οι συναρτήσεις που αναφέραμε είναι όλες συνεχείς και αύξουσες στο διάστημα (0,1).
- Για μικρές τιμές της πιθανότητας επιτυχίας p , η συνάρτηση complementary log-log link δίνει παρόμοια αποτελέσματα με τη συνάρτηση logit διότι στην περίπτωση αυτή

$$log(1-p) \approx -p$$

και συνεπώς

$$log[-log(1-p)] \approx log(p)$$

οπότε

$$logit(p) \approx log(p) + p \approx log(p).$$

Η επιλογή των τριών συναρτήσεων σύνδεσης δεν είναι τυχαία. Στη παράγραφο που ακολουθεί δίνεται η φυσική και μαθηματική ερμηνεία για την επιλογή των συναρτήσεων αυτών.

1.3.3 Η συνάρτηση σύνδεσης logit

Για την περίπτωση συγκεκριμένα της συνάρτηση σύνδεσης logit θεωρούμε ότι η κατανομή της U είναι η λογιστική κατανομή.

Έστω μία τ.μ. Y. Θεωρούμε ότι η Y έχει τη λογιστική κατανομή με παραμέτρους μ και s^2 όταν η αθροιστική της συνάρτηση γράφεται στην ακόλουθη μορφή

$$P(Y \leq y) = \frac{1}{1 + e^{-\frac{y-\mu}{s}}}, \qquad -\infty < y < \infty.$$

όπου η παράμετρος μ ισούται με τη μέση τιμή της κατανομής.

Επομένως, θεωρώντας ότι για τη συγκεκριμένη παρατήρηση i η μέση τιμή ισούται με

$$\mu_i = \beta_0 + \beta_1 x_i$$

και ότι $s = 1$, για όλες τις παρατηρήσεις, παίρνουμε

$$p_i = P(Y_i = 1) = P(U_i > 0)$$

$$= 1 - \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$$= \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}.$$

Αντίστοιχα,

$$1 - p_i = P(Y_i = 0) = P(U_i \leq 0)$$

$$= \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}}.$$

Διαιρώντας τις δύο σχέσεις που προέκυψαν κατά μέλη προκύπτει ότι

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i \, .$$

Από την τελευταία σχέση παρατηρούμε ότι προκύπτει η συνάρτηση logit σαν μία συνάρτηση που συνδέει την πιθανότητα επιτυχίας με την τιμή της ερμηνευτικής μεταβλητής.

1.4 Εκτίμηση παραμέτρων

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

Στα γραμμικά μοντέλα, για την εκτίμηση των παραμέτρων $b_1, \ldots, b_p$ μπορούν να χρησιμοποιηθούν δύο μέθοδοι, η μέθοδος μέγιστης πιθανοφάνειας ή η μέθοδος ελαχίστων τετραγώνων. Υπάρχει και μία παραλλαγή της μεθόδου των ελαχίστων τετραγώνων η οποία ονομάζεται μέθοδος των σταθμισμένων ελαχίστων τετραγώνων.

Οι εκτιμήσεις συνήθως υπολογίζονται αριθμητικά από επαναληπτικές διαδικασίες οι οποίες είναι πολύ κοντά με τη μέθοδο εκτίμησης των σταθμισμένων ελαχίστων τετραγώνων.

1.4.1 Μέθοδος Μέγιστης Πιθανοφάνειας

Έστω οι τυχαίες μεταβλητές $Y_1, \dots, Y_N$ με από κοινού συνάρτηση πυκνότητας πιθανότητας $f(y, \theta)$ η οποία εξαρτάται από το διάνυσμα των παραμέτρων $\theta = [\theta_1, \dots, \theta_p]^T$.

Έστω Θ οι δυνατές τιμές του διανύσματος των παραμέτρων. Ο εκτιμητής μέγιστης πιθανοφάνειας του θ είναι η τιμή $\hat{\theta}$ η οποία μεγιστοποιεί τη συνάρτηση πιθανοφάνειας

$$L(\hat{\theta}, y) = \sup\{L(\theta, y)\}, \theta \in \Theta.$$

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

$$l(\hat{\theta}, y) = \sup\{l(\theta, y)\}, \theta \in \Theta.$$

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

$$\frac{\partial l(\theta, y)}{\partial \theta_j} = 0, \qquad j = 1, \dots, p.$$

Παρατηρήσεις:

1. Είναι απαραίτητο να ελεγχθεί ότι οι λύσεις των παραπάνω εξισώσεων αντιστοιχούν σε μέγιστες τιμές του $l(\theta, y)$. Ο έλεγχος αυτός επιτυγχάνεται όταν ο πίνακας των δεύτερων παραγώγων,

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

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

2. Πρέπει να ελέγξουμε για την ύπαρξη τιμών της παραμέτρου θ στα άκρα του χώρου παραμέτρων Θ που να δίνουν τοπικά μέγιστα για τη συνάρτηση $l(\theta, y)$. Εάν υπάρχουν, η τιμή του $\hat{\theta}$ που αντιστοιχεί στο μεγαλύτερο από όλα τα μέγιστα, είναι ο εκτιμητής μέγιστης πιθανοφάνειας.
3. Οι εκτιμητές μέγιστης πιθανοφάνειας έχουν την ακόλουθη σημαντική ιδιότητα. Αν $g(\theta)$ είναι μία συνάρτηση της παραμέτρου θ, τότε ο εκτιμητής μέγιστης πιθανοφάνειας του $g(\theta)$ είναι $g(\hat{\theta})$. Συνεπώς μπορούμε να εκτιμήσουμε τη μέγιστη πιθανοφάνεια με οποιαδήποτε συνάρτηση των παραμέτρων είναι πιο εύκολη στη χρήση και στη συνέχεια χρησιμοποιώντας την προαναφερθείσα ιδιότητα εξάγουμε εκτιμήσεις για τις ζητούμενες παραμέτρους.

4. Ιδιότητες που χαρακτηρίζουν τους εκτιμητές μέγιστης πιθανοφάνειας είναι μεταξύ άλλων η συνέπεια, η επάρκεια και η ασυμπτωτική ικανότητα.

1.4.2 Μέθοδος Ελαχίστων Τετραγώνων

Έστω οι τυχαίες μεταβλητές $Y_1, \dots, Y_N$ με μέσες τιμές

$$\mu_i = E(Y_i), \qquad \text{για } i = 1, \dots, N.$$

Υποθέτουμε ότι τα μ_i είναι συναρτήσεις των παραμέτρων

$$b = \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix},$$

που πρέπει να εκτιμηθούν.

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

Πιο συγκεκριμένα, θέλουμε να ελαχιστοποιήσουμε την ποσότητα

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

επομένως,

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

Για να βρούμε τους εκτιμητές $\hat{b}$ παραγωγίζουμε το S ως προς κάθε συνιστώσα b_j του b και λύνοντας τις εξισώσεις που ακολουθούν:

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

Σημείωση: Οι λύσεις που προκύπτουν πρέπει να ελέγξουμε ότι αντιστοιχούν σε ελάχιστα. Δηλαδή θα πρέπει ο πίνακας των δεύτερων παραγώγων να είναι θετικά ορισμένος.

Παρατηρήσεις:

1. Σημαντική διαφορά μεταξύ των μεθόδων των ελαχίστων τετραγώνων και της μέγιστης πιθανοφάνειας είναι ότι τα ελάχιστα τετράγωνα μπορεί να χρησιμοποιηθούν χωρίς να

κάνουμε υποθέσεις για τις κατανομές των μεταβλητών απόκρισης Y_i. Αρκεί ο καθορισμός των αναμενόμενων τιμών και πιθανόν για την δομή της διακύμανσης – συνδιακύμανσης.

2. Στη μέθοδο μέγιστης πιθανοφάνειας χρειάζεται να καθορίσουμε την από κοινού πυκνότητα πιθανότητας των Y_i.
3. Για να εξάγουμε την δειγματική κατανομή των εκτιμητών ελαχίστων τετραγώνων χρειαζόμαστε επιπροσθέτως προϋποθέσεις για τα Y_i.

1.5 Στατιστική συνάρτηση απόκλισης (Deviance)

Η απόκλιση αποτελεί γενίκευση της έννοιας του αθροίσματος των τετραγώνων των καταλοίπων (Residual SS) και προκύπτει από τον έλεγχο του λόγου πιθανοφανειών (Likelihood Ratio Test, LRT). (Πολίτης Κ., 2014)

Έστω ότι έχουμε δύο εμφωλευμένα μοντέλα (nested models), M_1 και M_2, δηλαδή το σύνολο των επεξηγηματικών μεταβλητών του M_1 είναι υποσύνολο αυτών του M_2.

Έστω επίσης ότι οι συναρτήσεις πιθανοφάνειας των δύο μοντέλων είναι $L(M_1)$ και $L(M_2)$ αντίστοιχα.

Ορίζουμε στη συνέχεια

$$l(M_1) = \log L(M_1) \text{ και } l(M_2) = log\ L(M_2).$$

Τότε η στατιστική ελέγχου για τον έλεγχο του λόγου πιθανοφανειών ορίζεται από τη σχέση

$$-2\log\frac{L(M_1)}{L(M_2)} = -2\big(l(M_1) - l(M_2)\big) \quad \textbf{(*)}$$

Η απόκλιση του μοντέλου M_1 ορίζεται ως η τιμή της ποσότητας στην **(*)** όταν το μοντέλο M_2 είναι το κορεσμένο μοντέλο, δηλαδή όταν το μοντέλο M_2 έχει τόσες παραμέτρους όσα και τα δεδομένα.

Για κανονικά δεδομένα, η απόκλιση συμπίπτει με το Residual SS, δηλαδή

$$\text{SSE} = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2.$$

Για μη κανονικές αποκρίσεις, η απόκλιση είναι ο λογάριθμος της πιθανοφάνειας του προσαρμοσμένου μοντέλου, αφού

$$\text{D} = -2[\text{l(model)} - \text{l(saturated)}]$$

και για το κορεσμένο μοντέλο, ισχύει

$$\mathrm{l}(\mathrm{saturated}) = 0.$$

Γενικά όσο πιο μικρή είναι η απόκλιση ενός μοντέλου, τόσο πιο κοντά είναι στο κορεσμένο μοντέλο, κατι το οποίο είναι ένδειξη καλής προσαρμογής. Για αυτό το λόγο η απόκλιση θα μπορούσε να χρησιμοποιηθεί ως μία στατιστική συνάρτηση ελέγχου για την υπόθεση για το προσαρμοσμένο μοντέλο δε διαφέρει σημαντικά από το κορεσμένο μοντέλο. Ωστόσο, η κατανομή της ποσότητας D στη γενική περίπτωση δεν είναι γνωστή.

Υπάρχει όμως η δυνατότητα να εξεταστεί αν υπάρχει στατιστικά σημαντική διαφορά μεταξύ δύο εμφωλευμένων μοντέλων. Αυτό μπορεί να πραγματοποιηθεί χρησιμοποιώντας τον έλεγχο του λόγου πιθανοφανειών ως κριτήριο τη διαφορά στην απόκλιση ανάμεσα στα δύο μοντέλα.

Έστω ότι το μοντέλο M_1έχει απόκλιση D_1 και βαθμούς ελευθερίας df_1 , ενώ το μοντέλο M_2 έχει απόκλιση D_2 και βαθμούς ελευθερίας df_2.

Επομένως, παρατηρούμε ότι η διαφορά

$$D_1 - \mathrm{D}_2 = -2[\mathrm{l}(\mathrm{reduced\ model}) - \mathrm{l}(\mathrm{full\ model})]$$

ακολουθεί προσεγγιστικά την κατανομή χ_p^2, όπου $p = \mathrm{df}_1 - df_2$.

Παρατηρούμε ότι πρόκειται για τον έλεγχο του λόγου πιθανοφανειών, αφού

$$D_1 - \mathrm{D}_2 = -2\left[\log\frac{L(reduced\ model)}{L(full\ model)}\right]$$

όπου L(.) είναι η συνάρτηση πιθανοφάνειας.

Σημείωση:

- Στα Γενικευμένα Γραμμικά Μοντέλα, η σειρά εισαγωγής των μεταβλητών στο μοντέλο επηρεάζει τη σημαντικότητα αυτών. Επομένως, είναι απαραίτητο να γνωρίζουμε ποιες μεταβλητές υπάρχουν ήδη στο μοντέλο.

1.6 Έλεγχος της ολικής επάρκειας ενός μοντέλου

Η επάρκεια ενός μοντέλου είναι η ικανότητα του μοντέλου να περιγράψει τα δεδομένα που μελετάμε. Ένα μοντέλο που περιλαμβάνει τόσες παραμέτρους όσα είναι και τα δεδομένα περιγράφει τα δεδομένα επακριβώς άρα είναι ένα επαρκές μοντέλο. Το μοντέλο αυτό δε μπορεί να χρησιμοποιηθεί για πρόβλεψη, αλλά μπορεί να μας είναι χρήσιμο για την αξιολόγηση άλλων μοντέλων συγκρίνοντας τα με αυτό.

Αν υποθέσουμε ότι θέλουμε να εκτιμήσουμε την επάρκεια ενός μοντέλου που περιγράφει ένα σύνολο δεδομένων που μελετάμε. Η εκτίμηση γίνεται συγκρίνοντας την πιθανοφάνεια του μοντέλου που μελετάμε με την πιθανοφάνεια για το κορεσμένο (saturated) μοντέλο.

Το κορεσμένο μοντέλο μπορεί να θεωρηθεί ότι μας εφοδιάζει με μία πλήρη περιγραφή των δεδομένων και όπως αναφέρθηκε είναι ένα επαρκές μοντέλο. Με το κορεσμένο μοντέλο έχουμε τη μεγαλύτερη πολυπλοκότητα. Χρησιμοποιώντας ένα μοντέλο με ένα μικρότερο αριθμό παραμέτρων μπορούμε να επιτύχουμε την επιθυμητή προσαρμογή και συγχρόνως δεν συμπεριλαμβάνουμε παραμέτρους που δε χρειαζόμαστε.

Οι συναρτήσεις πιθανοφάνειας υπολογίζονται στον εκτιμητή μέγιστης πιθανοφάνειας b_{max} και b αντίστοιχα και λαμβάνουμε $L(b_{max}; y)$ και $L(b; y)$ αντίστοιχα. Αν το μοντέλο που μας ενδιαφέρει περιγράφει τα δεδομένα ικανοποιητικά, τότε $L(b; y)$ πρέπει να είναι κοντά στο $L(b_{max}; y)$. Σε αντίθετη περίπτωση, δηλαδή αν το μοντέλο δεν περιγράφει ικανοποιητικά τα δεδομένα, τότε το $L(b; y)$ πρέπει να είναι μικρότερο από το $L(b_{max}; y)$. Αυτό μας οδηγεί στη χρήση του Γενικευμένου λόγου πιθανοφάνειας

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

ή ισοδύναμα το λογάριθμο της ακόλουθης σχέσης

$$\log \lambda = \log\big(L(b_{max}; y)\big) - \log\big(L(b; y)\big) = l(b_{max}; y) - l\big((b; y)\big)$$

σαν μέτρο καλής προσαρμογής του μοντέλου.

Μεγάλες τιμές του $\log \lambda$ είναι ένδειξη μη καλής προσαρμογής του μοντέλου. Για να βρούμε την κριτική περιοχή του $\log \lambda$ πρέπει να βρούμε τη δειγματική κατανομή του.

1.6.1 Έλεγχος της ολικής επάρκειας ενός μοντέλου για δίτιμα δεδομένα

Υπάρχουν διάφοροι μέθοδοι με τις οποίες μπορούμε να ελέγξουμε την ολική επάρκεια ενός μοντέλου για δίτιμα δεδομένα. Από τις υπάρχουσες μεθόδους δεν φαίνεται κάποια να δίνει πάντα ικανοποιητικά αποτελέσματα. Όπως αναφέραμε και παραπάνω, για δίτιμα δεδομένα η κατανομή της απόκλισης D δεν είναι γνωστή, ούτε προσεγγιστικά.

Ο συνηθέστερος έλεγχος που χρησιμοποιείται είναι ο έλεγχος των Hosmer-Lemeshow (HL).

Η διαδικασία που ακολουθείται είναι η εξής:

1. Διατάσσουμε τις παρατηρήσεις ανάλογα με την προβλεπόμενη πιθανότητα επιτυχίας.

2. Χωρίζουμε τις διατεταγμένες παρατηρήσεις σε g ομάδες, με ίσο περίπου αριθμό παρατηρήσεων, και για κάθε μία από αυτές καταγράφουμε τον αριθμό επιτυχιών και αποτυχιών, σχηματίζοντας έτσι έναν πίνακα $g \times 2$.

3. Η στατιστική συνάρτηση των HL, X_{HL} είναι το X^2 του Pearson για τον παραπάνω πίνακα.
   Κάτω από την μηδενική υπόθεση, H_0: οι παρατηρηθείσες τιμές της Y δε διαφέρουν από τις εκτιμώμενες τιμές. Η συνάρτηση X_{HL} ακολουθεί την κατανομή χ^2_{g-2}.

Απόρριψη της H_0 δηλώνει ότι το μοντέλο μας είναι ανεπαρκές για το συγκεκριμένο ε.σ. του ελέγχου.

1.6.2 Εκτίμηση στο Γενικευμένο Γραμμικό Μοντέλο

Στα γενικευμένα γραμμικά μοντέλα, οι μέθοδοι που χρησιμοποιούνται για την εκτίμηση βασίζονται σε δύο μεθόδους, στη μέθοδο Newton-Raphson και στη μέθοδο των score.

Έστω $Y_1, \dots, Y_N$ ανεξάρτητες τυχαίες μεταβλητές. Σκοπός μας είναι να εκτιμήσουμε τις παραμέτρους b που σχετίζονται με τα Y_i μέσω των σχέσεων που ακολουθούν

$$\mu_i = E(Y_i)$$

και

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

Για κάθε Y_i, η λογαριθμική συνάρτηση πιθανοφάνειας είναι :

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

Σύμφωνα με την Annete J. Dobson (Dobson, 2010), για τον υπολογισμό της εκτιμήτριας μέγιστης πιθανοφάνειας για την παράμετρο b_j, χρειαζόμαστε την ακόλουθη σχέση:

$$\frac{\partial \mathrm{l}}{\partial \mathrm{b_j}} = \mathrm{U_j} = \sum_{\mathrm{i=1}}^{\mathrm{N}} \left[\frac{\partial \mathrm{l_i}}{\partial \mathrm{b_i}}\right] = \sum_{\mathrm{i=1}}^{\mathrm{N}} \left[\frac{\partial \mathrm{l_i}}{\partial \theta_\mathrm{i}} \frac{\partial \theta_\mathrm{i}}{\partial \mu_\mathrm{i}} \frac{\partial \mu_\mathrm{i}}{\partial \mathrm{b_i}}\right]$$

όπου

- $\frac{\partial \mathrm{l_i}}{\partial \theta_\mathrm{i}} = \mathrm{b}'(\theta_\mathrm{i})(\mathrm{y_i} - \mu_\mathrm{i})$
- $\frac{\partial \theta_\mathrm{i}}{\partial \mu_\mathrm{i}} = \frac{1}{\frac{\partial \mu_\mathrm{i}}{\partial \theta_\mathrm{i}}}$, όπου $\frac{\partial \mu_\mathrm{i}}{\partial \theta_\mathrm{i}} = \mathrm{b}''(\theta_\mathrm{i})\mathrm{Var}(\mathrm{Y_i})$

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

Επομένως, η παραπάνω σχέση γράφεται στην ακόλουθη μορφή:

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

Η συνάρτηση $U = (U_1, \dots, U_N)$ που προκύπτει ονομάζεται score συνάρτηση.

Ο πίνακας διασποράς – συνδιασποράς των U_i αποτελείται από τους όρους $I_{jk} = E[U_j U_k]$ που σχηματίζουν τον πίνακα:

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

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

οπότε προκύπτει ότι

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

Ο πίνακας I_{jk} που προκύπτει ονομάζεται πίνακας πληροφορίας.

Ο παραπάνω πίνακας γράφεται επίσης στη μορφή:

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

όπου $\mathbf{W}$ είναι ένας διαγώνιος $N \times N$ πίνακας, τα στοιχεία του οποίου είναι τα

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

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

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

Ο πίνακας πληροφορίας $I = E[UU^T]$ αποτελείται από τα στοιχεία

$$I_{jk} = E\left[U_j U_k\right] = E\left[\frac{\partial l}{\partial b_j}\frac{\partial l}{\partial b_k}\right] = E\left[\frac{\partial^2 l}{\partial b_j b_k}\right].$$

Σύμφωνα με τις σχέσεις που προέκυψαν παραπάνω προκύπτει ότι

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

Επομένως,

$$I^{(m-1)} b^{(m)} = I^{(m-1)} b^{(m-1)} + U^{(m-1)}.$$

απ' όπου προκύπτει ότι

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

Παρατηρούμε ότι η σχέση που προέκυψε έχει την ίδια μορφή με τις κανονικές εξισώσεις των γενικευμένων γραμμικών μοντέλων που προκύπτουν από τα σταθμισμένα ελάχιστα τετράγωνα με μόνη διαφορά ότι πρέπει να λυθούν με μία επαναληπτική μέθοδο επειδή τα **z** και **W**εξαρτώνται από το b. Αυτό σημαίνει ότι οι εκτιμήτριες μέγιστης πιθανοφάνειας των γενικευμένων γραμμικών μοντέλων προκύπτουν από μία επαναληπτική διαδικασία σταθμισμένων ελαχίστων τετραγώνων.

1.6.3 Εκτίμηση των παραμέτρων στη λογιστική παλινδρόμηση με δίτιμα δεδομένα

Θεωρούμε το μοντέλο με k ερμηνευτικές μεταβλητές,

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Με τη μέθοδο μέγιστης πιθανοφάνειας, το διάνυσμα των παραμέτρων

$$\boldsymbol{\beta}' = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$$

προκύπτει μεγιστοποιώντας τη συνάρτηση πιθανοφάνειας, ή ισοδύναμα το λογάριθμο της, δηλαδή επιλύοντας το σύστημα

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$$

ως προς β_i, για $i = 1,2,\ldots,k$.

Για διωνυμικά (v, p) δεδομένα, έχουμε δει ότι η συνάρτηση πιθανότητας γράφεται στη μορφή

$$f_Y(y; p) = \exp\left[y \log\frac{p}{1-p} + v\log(1-p) + \log\binom{v}{y}\right],$$

άρα για $v = 1$, κατανομή Bernoulli, παίρνουμε

$$f_Y(y; p) = \exp\left[y \log \frac{p}{1-p} + \log(1-p)\right] c(y),$$

όπου η συνάρτηση c δεν εξαρτάται από την παράμετρο p.

Επομένως, αν έχουμε n παρατηρήσεις $y_1, y_2, \dots, y_n$ από την κατανομή $\mathrm{B_i}(1,\ p_i)$ αντίστοιχα, η πιθανοφάνεια του δείγματος θα είναι

$$L_n(\mathbf{p}; y) = \exp\left[\sum_{i=1}^{n} y_i \ \log \frac{p_i}{1-p_i} + \sum_{i=1}^{n} log(1-p_i)\right] c_n(y)$$

όπου

$$\mathbf{p} = (p_1, \dots, p_n)', \qquad \boldsymbol{y} = (y_1, \dots, y_n)'$$

και η συνάρτηση c_n δεν εξαρτάται από τις άγνωστες παραμέτρους p_i.

Ο λογάριθμος της συνάρτησης πιθανοφάνειας γράφεται

$$l(\boldsymbol{p}; y) = \sum_{i=1}^{n} y_i \ \log \frac{p_i}{1-p_i} + \sum_{i=1}^{n} \log\ (1-p_i)$$

$$= \sum_{i=1}^{n} (y_i \log p_i + \ (1-y_i) \log(1-p_i))$$

Χρησιμοποιώντας τη σχέση

$$g(p_i) = \ \log\left(\frac{p_i}{1-p_i}\right) = \sum_{j=0}^{k} \beta_j x_{ij}$$

όπου έχουμε θέσει για ευκολία στο συμβολισμό $x_{0,j} = 1,\ \forall j$, οπότε μετά από πράξεις προκύπτει ότι

$$\frac{\partial l}{\partial \beta_r} = \sum_{i=1}^{n} y_i \ x_{ir} - \sum_{i=1}^{n} p_i \ x_{ir} = \sum_{i=1}^{n} (y_i - p_i) \ x_{ir}.$$

Θέτοντας

$$\frac{\partial l}{\partial \beta_r} = 0 \quad \text{για } r = 0,1,2, \dots, k$$

καταλήγουμε σε ένα σύστημα $(k+1)$ εξισώσεων ως προς τις παραμέτρους β_r.

1.6.4 Έλεγχος επάρκειας ενός ΓΓΜ για δίτιμα δεδομένα

Γνωρίζουμε ότι η απόκλιση ενός ΓΓΜ με δίτιμες αποκρίσεις δεν ακολουθεί την κατανομή χ^2. Συνηθέστερος έλεγχος για την εξεταζόμενη περίπτωση είναι αυτός των Hosmer-Lemeshow, όπως αναφέρθηκε.

Ένα χρήσιμο μέσο αξιολόγησης της προσαρμογής ενός μοντέλου είναι ένας πίνακας ταξινόμησης, όπου ακολουθείται η εξής διαδικασία (Πολίτης Κ., 2014):

1. Υπολογίζουμε όλες τις εκτιμώμενες τιμές $\hat{p}_i$ για $i = 1,2, \dots, n$.
2. Επιλέγουμε μία πιθανότητα p_0 ως κατώφλι ή σημείο αποκοπής (cutoff point).
   Για $i = 1,2, \dots, n$, αν
   - $\hat{p}_i \geq p_0$, τότε θεωρούμε ότι για την παρατήρηση i το υπόδειγμα προβλέπει «επιτυχία» $(\hat{Y}_i = 1)$.
   - $\hat{p}_i < p_0$, τότε θεωρούμε ότι για την παρατήρηση i το υπόδειγμα προβλέπει «αποτυχία» $(\hat{Y}_i = 0)$.
3. Ο πίνακας ταξινόμησης είναι ένας πίνακας δύο διαστάσεων που μας δίνει τις συχνότητες για τις επιτυχίες και τις αποτυχίες ανάμεσα στις παρατηρηθείσες και τις εκτιμώμενες τιμές.

Εκτιμώμενο αποτέλεσμα	**Παρατηρούμενο αποτέλεσμα**		
		Επιτυχία	Αποτυχία
	Επιτυχία (πάνω από την τιμή p_0)	$\boldsymbol{a}$	$\boldsymbol{b}$
	Αποτυχία (κάνω από την τιμή p_0)	$\boldsymbol{c}$	$\boldsymbol{d}$

όπου $a + b + c + d = n$.

Μπορούμε να αξιολογήσουμε την ευαισθησία (severity) και την ειδικότητα (specificity) του μοντέλου ως εξής:

$$\text{Ευαισθησία} = \frac{a}{a + c}\text{ , (ποσοστό επιτυχιών που "ταξινομούνται" σωστά)}$$

και

$$\text{Ειδικότητα} = \frac{d}{b+d}\text{ , (ποσοστό αποτυχιών που "ταξινομούνται" σωστά).}$$

1.6.5 Διαγνωστικές μέθοδοι : Κατάλοιπα

Η επάρκεια ενός μοντέλου εξετάζεται με τη βοήθεια μιας στατιστικής συνάρτησης καλής προσαρμογής. Οι συναρτήσεις αυτές δίνουν ένα γενικό μέτρο για την καλή προσαρμογή και δε δίνουν πληροφορία για τη μορφή του μοντέλου. Η εξερεύνηση των ιδιαιτεροτήτων ενός μοντέλου γίνεται με τη χρήση των καταλοίπων.

Για το μοντέλο έχουμε ότι

$$\frac{(Y_i - \mu_i)}{\sigma} \sim N(0,1).$$

Τα κατάλοιπα που αντιστοιχούν στα Y_i ορίζονται ως $(y_i - \hat{\mu}_i)$ όπου $\hat{\mu}_i$ είναι οι προσαρμοσμένες τιμές που υπολογίζονται με την εκτίμηση της μέγιστης πιθανοφάνειας b.

Εάν το $\hat{\sigma}$ είναι μία εκτίμηση του σ τότε ορίζονται τα τυποποιημένα κατάλοιπα r_i και δίνονται από τη σχέση

$$r_i = \frac{(Y_i - \mu_i)}{\hat{\sigma}}.$$

Είναι απαραίτητο να εξεταστεί η κανονικότητα των τυποποιημένων καταλοίπων και η ύπαρξη συσχέτισης μεταξύ των r_i. Εάν τα κατάλοιπα είναι περίπου ασυσχέτιστα θα ακολουθούν κατά προσέγγιση την κατανομή $N(0,1)$.

Στα γενικευμένα γραμμικά μοντέλα ορίζονται τυποποιημένα κατάλοιπα με πολλούς διαφορετικούς τρόπους. Ένας τρόπος να ορίσουμε τα τυποποιημένα κατάλοιπα είναι να γενικεύσουμε τον ορισμό των τυποποιημένων καταλοίπων για κανονικά μοντέλα και να προκύψει ο ορισμός των τυποποιημένων μοντέλων για γενικευμένα γραμμικά μοντέλα.

Έστω ότι $s_i = \sqrt{var(\mu_i)}$ η εκτιμηθείσα τιμή για την τυπική απόκλιση των προσαρμοσμένων τιμών $\hat{\mu}_i$. Επομένως, τα τυποποιημένα κατάλοιπα για γενικευμένα γραμμικά μοντέλα ορίζονται ως εξής:

$$r_i = \frac{(y_i - \hat{\mu}_i)}{s_i}.$$

Υπάρχουν τέσσερα είδη καταλοίπων σε ένα γενικευμένο γραμμικό μοντέλο:

1. Deviance residuals
2. Response residuals
3. Pearson residuals

4. Working residuals

Για διαγνωστικούς σκοπούς, χρησιμοποιούμε κυρίως είτε τα κατάλοιπα απόκλισης (deviance residuals) ή τα κατάλοιπα του Pearson.

Στη απλή παλινδρόμηση, τα κατάλοιπα είναι οι διαφορές ανάμεσα στις παρατηρηθείσες και τις προβλεπόμενες τιμές για κάθε παρατήρηση. Σε ένα γενικευμένο γραμμικό μοντέλο, τα κατάλοιπα αυτά ονομάζονται κατάλοιπα απόκρισης (response residuals).

Για ένα μοντέλο με δίτιμα δεδομένα, αυτά είναι

$$e_i = y_i - \hat{y}_i = y_i - \hat{p}_i,$$

αφού η προβλεπόμενη τιμή για κάθε παρατήρηση είναι η εκτιμώμενη πιθανότητα επιτυχίας.

Τα κατάλοιπα απόκλισης είναι οι τετραγωνικές ρίζες των προσθετέων στο άθροισμα

$$-2 \sum \left[y_i \log\left(\frac{\hat{p}_i}{y_i}\right) + (1 - y_i) \log\left(\frac{1 - \hat{p}_i}{1 - y_i}\right) \right],$$

όπου σε κάθε τετραγωνική ρίζα βάζουμε το ίδιο πρόσημο με αυτό που υπάρχει στο αντίστοιχο response residual.

Τέλος, τα κατάλοιπα του Pearson προκύπτουν από τα κατάλοιπα απόκρισης μετά από τυποποίηση και δίνονται από την ακόλουθη σχέση

$$e_i = \frac{(y_i - \hat{p}_i)}{\sqrt{\hat{p}_i(1 - \hat{p}_i)}}.$$

Για μεγάλο μέγεθος δείγματος, αυτά έχουν προσεγγιστικά μέση τιμή μηδέν και διακύμανση ίση με 1, δηλαδή

$$\mathrm{E}(e_i) \approx 0 \quad \text{και} \quad \mathrm{Var}\,(e_i) \approx 1.$$

1.7 Κριτήτια AIC και BIC

Σε αρκετές περιπτώσεις αντιμετοπίζουμε το πρόβλημα της επιλογής των επεξηγηματικών μεταβλητών που θα χρησιμοποιηθούν στο τελικό μοντελο. Αν διαθέτουμε p επεξηγηματικές μεταβλητές, τότε υπάρχουν 2^p πιθανά μοντέλα που προκύπτουν από διαφορετικούς συνδυασμούς των επεξηγηματικών μεταβλητών. Αυτά τα μοντέλα θα πρέπει να αξιολογηθούν με βάση κάποιο κριτήριο και αυτό που χρησιμοποιείται συνήθως είναι το κριτήριο ποινικοποιημένης πιθανοφάνειας (π.χ AIC ή BIC). Η επιλογή του μοντέλου πραγματοποιείται για εκείνο το οποίο ελαχιστοποιεί την τιμή του κριτηρίου. Τα κριτήρια αποτελούνται απο δύο

μέρη, όπου το ένα μέρος ανταμείβει τα μοντέλα με καλή προβλεπτική ικανότητα, ενώ το άλλο μέρος δίνει «ποινή» στα μοντέλα με πολλές επεξηγηματικές μεταβλητές.

Η τιμή του κριτηρίου AIC για ένα μοντέλο m δίνεται απο τον τύπο:

$$AIC(m) = -2\log L(m) + 2d_m \ ,$$

όπου L(m) ειναι η μεγιστοποιημένη τιμή της πιθανοφάνειας του μοντέλου m και d_m είναι ο αριθμός των αγνώστων παραμέτρων του μοντέλου m. Στην διαδικασία επιλογής του μοντέλου επιλέγεται αυτό το οποίο ελαχιστοποιεί την «απόκλιση» μεταξύ του προσαρμοσμένου και του «πραγματικού» μοντέλου.

Το κριτήριο BIC χρησιμοποιείται συχνά στην Μπεϋζιανή Στατιστική και μοιάζει αρκετά με το κριτήριο AIC, με την διαφορά οτι «τιμωρεί» τα μοντέλα με πολλές επεξηγηματικές μεταβλητές όταν το μέγεθος του δείγματος είναι μεγαλύτερο του 7.39. ($e^2 = 7.39$)

Η τιμή του κριτηρίου BIC για ενα μοντέλο m δίνεται από τον τύπο:

$$BIC(m) = -2\log L(m) + d_m \log n \ ,$$

όπου L(m) ειναι η μεγιστοποιημένη τιμη της πιθανοφάνειας του μοντέλου m και d_m είναι ο αριθμός των αγνώστων παραμέτρων του μοντέλου m και n ο αριθμός του δείγματος. (Φουσκακης, 2013).

ΚΕΦΑΛΑΙΟ 2

ΠΕΡΙΓΡΑΦΗ ΚΑΙ ΑΝΑΛΥΣΗ ΣΕ ΑΣΦΑΛΙΣΤΙΚΑ ΔΕΔΟΜΕΝΑ

2.1 Γενικά

Η μελέτη των γενικευμένων γραμμικών μοντέλων, πολλές φορές μας οδηγεί σε πολύπλοκους υπολογισμούς. Η χρήση διάφορων στατιστικών πακέτων σε υπολογιστή, μας βοηθάει να κάνουμε 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.

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

2.1.1 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 επιλογή 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. 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 logit, θα κάνουμε την εξής παρακάτω διαδικασία: Θα καταχωρήσουμε στην R τα δεδομένα και θα δημιουργήσουμε ένα πλαίσιο όπου το ονομάζουμε, έστω ”data.frame”. Η εντολή που θα δώσουμε είναι:

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

2.2 Προβλήματα με τα δεδομένα και Μεροληψία

Τα σύνολα ασφαλιστικών δεδομένων είναι συνήθως πολύ μεγάλα, και προβλήματα όπως το να λείπουν τιμές (συχνά υποδεικνύονται με κενό ή μηδέν) κάνουν συχνά την εμφάνισή τους, κάτι το οποίο πρέπει να επιλυθεί πριν από τη στατιστική μοντελοποίηση. Προβλήματα προκύπτουν συχνά επειδή εκείνοι που συλλέγουν ή εισάγουν πληροφορίες δεν εκτιμούν τις στατιστικές χρήσεις στις οποίες τα δεδομένα θα τεθούν. Μια στατιστική ανάλυση είναι, ιδανικά, "αμερόληπτη", δηλαδη τα αποτελέσματα δεν ευνοούν καποιο συμπέρασμα. Οι «μεροληψιες» προκύπτουν με πολλούς τρόπους:

- Τα αποτελέσματα συχνά λογοκρίνονται. Για παράδειγμα, κατά τη μελέτη της μέσης διάρκειας ζωής από εκείνους που γεννήθηκαν το 1950, οι θάνατοι από αυτή την ομάδα εμφανίστηκαν μόνο για νεώτερες ηλικίες: δεν έχουν γίνει ακόμα παρατηρήσεις σχετικά με τις ζωές του επιζήσαντων. Αυτό είναι ένα προφανές παράδειγμα λογοκρισίας. Για τα προσωπικά δεδομένα τραυματισμού, το μέσο όρο και τυπική απόκλιση των ποσών των απαιτήσεων καταγραφής είναι σχεδιασμένα κατά μήνας ατυχήματος, δηλαδή ο μήνας κατά τον οποίο το ατύχημα συνέβη. Φαίνεται ότι το μέσο ποσό της απαίτησης καταγραφής μειώνεται με το χρόνο.Ωστόσο, αυτή η εμφάνιση είναι παραπλανητική. Οι μεταγενέστεροι μήνες ατυχημάτων έχουν πολλές ανεξόφλητες απαιτησεις, οι οποίες είναι

συνήθως υψηλές επειδή εμπλέκονται μεγαλύτερα ποσά και είναι πιο αμφιλεγόμενα. Έτσι το χαρακτηριστικό της πτώσης των ποσών τον απαιτήσεων με την πάροδο του χρόνου είναι συνέπεια της μεροληπτικής δειγματοληψίας, με τις μεγαλύτερες απαιτήσεις να εχουν λογοκριθεί από το δείγμα. Οποιοδήποτε μοντέλο για αυτά τα δεδομένα είναι πιθανό να είναι παραπλανητικο εκτός αν αντιμετωπιστει αυτή η μεροληπτική δειγματοληψία.

2.3 Μεταβλητή απόκρισης με Δίτιμα δεδομενα

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.

2.3.1 Περιγραφή Μεταβλητών

Για την περάτωση αυτης της διπλωματικής εργασίας χρησιμοποιήθηκαν ασφαλιστικά δεδομένα με 22036 απαιτήσεις (N=22036) μετά απο συλλογή και ανάλυση τους απο τους Piet de Jong και Gillian Z. Heller για τις ανάγκες του βιβλίου Generalized Linear Models for Insurance Data. Αυτές οι απαιτήσεις αναφέρθηκαν στην Αυστραλία κατα τη χρονική περίοδο απο τον Ιούλιο του 1989 εως το τέλος του 1999. Αποζημιώσεις με μηδενική πληρωμή εξαιρέθηκαν.
Οι μεταβλητές που χρησιμοποιήθηκαν για το μοντέλο και την ανάλυση του είναι οι ακόλουθες:

- Το ποσό των απαιτήσεων (Claims) το οποίο είναι μια συνεχής μεταβλητή. Οι συνεχείς μεταβλητές ονομάζονται επίσης μεταβλητές "διαστήματος" για να δείξουν ότι μπορούν να πάρουν τιμές οπουδήποτε σε ένα διάστημα της πραγματικής γραμμής. Στην περίπτωσή μας η μεταβλητή αυτή έχει βαριά δεξιά ουρά. Υπάρχει, δηλαδή, ένας μικρός

αριθμός πολύ μεγάλων απαιτήσεων πέραν των 50.000 δολαρίων και μεγαλυτερη απαίτηση είναι περίπου 4,5 εκατομμύρια δολάρια. Για αυτόν τον λόγο η μεταβλητή των «Απαιτήσεων» θα μετατραπεί σε μια δίτιμη μεταβλητή, όπου

$$Claims = \begin{cases} 0, & claims < 50000 \\ 1, & claims \geq 50000 \end{cases}$$

Δηλαδή ορίζουμε:

- 0: Χαμηλή Απαίτηση (Low Claim)
- 1: Υψηλή Απαήτηση (High Claim)

- Η νομική εκπροσώπηση (Legal Representation) είναι μια κατηγορική μεταβλητή με δύο επίπεδα "όχι" ή "ναι". Οι μεταβλητές που λαμβάνουν μόνο δύο πιθανές τιμές συχνά κωδικοποιούνται "0" και "1" και ονομάζονται επίσης δυαδικές, δείκτες ή μεταβλητές Bernoulli. Οι δυαδικές μεταβλητές υποδηλώνουν την παρουσία ή την απουσία ενός χαρακτηριστικού ή την εμφάνιση ή μη εμφάνισης ενός γεγονότος.
- Οι τραυματισμοι (Injury) είναι επίσης μια κατηγορική μεταβλητή, ονομάζεται επίσης και ποιοτική. Η μεταβλητή έχει τρεις τιμές που αντιστοιχούν σε διαφορετικά επίπεδα σωματικής βλάβης. Το επίπεδο 1 υποδεικνύει το χαμηλότερο επίπεδο τραυματισμού, το 2 το υψηλό επίπεδο τραυματισμού, ενώ το 3^{ο} επιπεδο υποδεικνύει το θάνατο. Το επίπεδο 9 αντιστοιχεί σε ένα "άγνωστο" ή μη καταγεγραμμένο επίπεδο τραυματισμού και ως εκ τούτου πιθανώς δεν υποδεικνύει σωματική βλάβη. Οι κατηγορικές μεταβλητές γενικά αναλαμβάνουν ένα από ένα διακριτό σύνολο τιμών που είναι ονομαστικής φύσης και δεν χρειάζεται να είναι διατάξιμος.
- Τελος, έχουμε την μεταβλητη καθυστέρησης διακανονισμού (Settlement Delay). Αυτη ειναι μια συνεχής μεταβλητή, η οποία στην πράξη είναι ένας αριθμός ημερών.

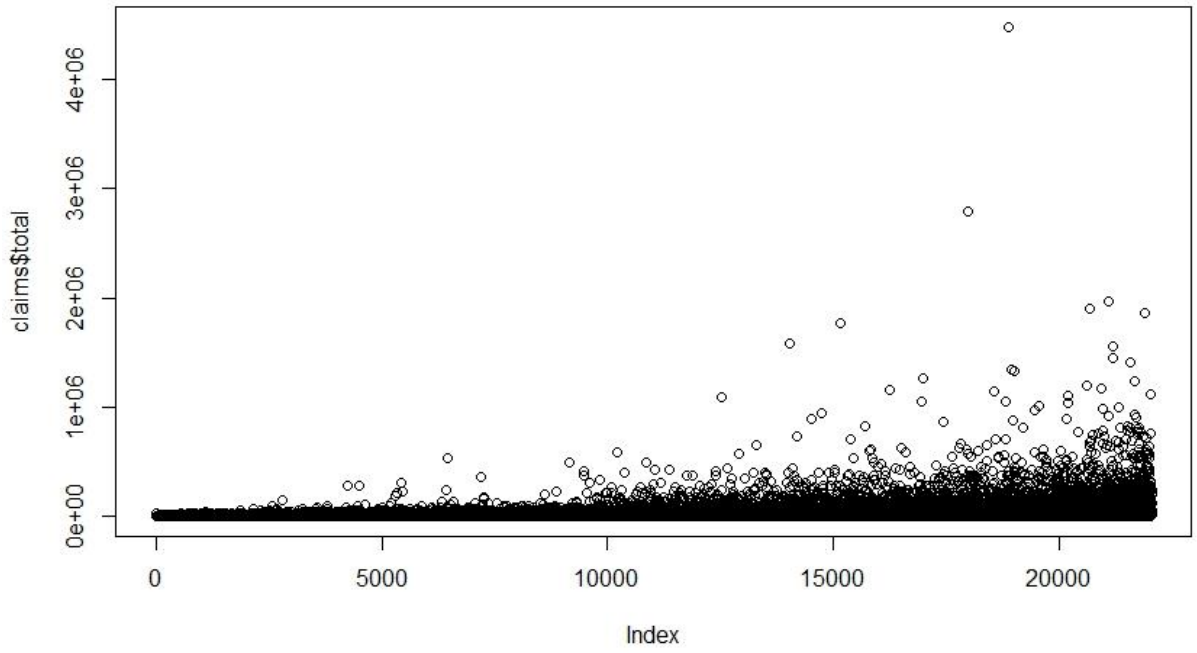
2.3.2 Περιγραφικά στατιστικά των Δεδομενων

Η αρχικη μεταβλητη που δηλωνει το ποσο των απαιτήσεων (Claims) είναι μια συνεχης μεταβλητη αποτελουμενη απο 22036 παρατηρησεις (n= 22036).

Πίνακας 2.1: Απαιτησεις (Claims)

Min.	10
1st Quartile	6,297
Median	13,854
Mean	38,367
3rd Quartile	35,123
Max	4,485,797

Παρατηρούμε οτι στα δεδομένα μας υπάρχει μεγάλο εύρος (range) αφου η μικρότερη παρατήρηση είναι 10 (Min. = 10) ενω η μεγαλύτερη ειναι 4,485,797 (Max = 4,485,797). Ταυτόχρονα, η παρατήρηση που ειναι μεγαλύτερη η ίση με το 75% των παρατηρήσεων (3rd Quartile) είναι 35,123. Η τιμή αυτή βρίσκεται «μακριά» απο την μέγιστη (Max) ενώ είναι και μικρότερη απο τον μέσο (Mean=38,367)



Σύμφωνα με τα παραπάνω και με το διάγραμα των αποζημειώσεων παρατηρείται ότι η μεταβλητή αυτή έχει βαριά δεξιά ουρά. Για αυτόν τον λόγο, αλλά και για να υπάρχουν αρκετές παρατηρήσεις για τις υψηλές απαιτήσεις, η μεταβλητή των «Απαιτήσεων» θα μετατραπεί σε μια δίτιμη μεταβλητή με κατώφλι τα 50.000 δολάρια.

Οπου η Χαμηλή Απαίτηση περιλαμβάνει 18100 παρατηρήσεις και η Υψηλή Απαίτηση εμπεριέχει τις εναπομείνασες 3936 παρατηρήσεις. (Πίνακας 2.2)

Πίνακας 2.2: Δίτιμη Μεταβλητή (Claims)

Claims	Frequency	Percent
Low Claim	18100	82%
High Claim	3936	18%
Total	22036	100%

Στους Πίνακες 2.3 και 2.4 δίνονται οι συχνότητες για τις μεταβλητές Νομική εκπροσώπηση (Legal representation) και Τραυματισμός (Injury) αντιστοιχα.

Πίνακας 2.3: Legal Representation

Legal Representation	Frequency	Percent
No	8008	36%
Yes	14028	64%
Total	22036	100%

Πίνακας 2.4: Injury

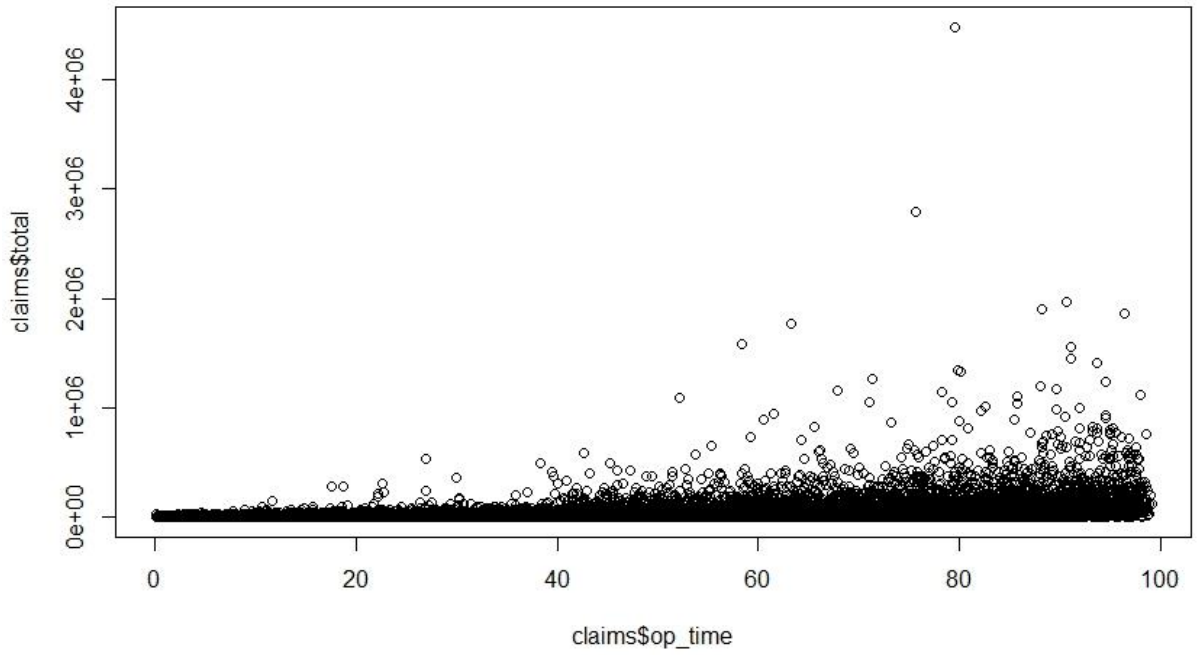
Injury	Frequency	Percent
Plain Injury	20218	92%
Severe Injury	194	1%
Death	258	1%
NA's	1366	6%
Total	22036	100%

Παρατηρούμε από τα δεδομένα ότι δύο στους τρεις (64%) επέλεξαν να εχουν νομική εκπροσώπηση, ενώ το υπόλοιπο 36% δεν ζήτησε νομική υποστήριξη.

Όσον αφορά τους τραυματισμούς, παρατηρείται οτι το μεγαλύτερο μερος αυτών (92%) είναι ελαφράς μορφής. Από ενα μικρό ποσοστό (2%) μοιράζονται αντίστοιχα οι σοβαροί τραυματισμοί και οι θάνατοι. Επίσης, για το 6% των παρατηρήσεων δεν εχει καταγραφεί κάποιου είδους τραυματισμός.

Πίνακας 2.5: Operational time

Min.	0.10
1st Quartile	23
Median	45.90
Mean	46.33
3rd Quartile	69.30
Max	99.10



Τέλος, για την μεταβλητή καθυστέρησης διακανονισμού (Operational time) παρατηρούμε οτι στα δεδομένα μας υπάρχει μεγάλο εύρος (range) αφου η μικροτερη παρατηρηση ειναι 0.10 (Min. = 0.10) ενω η μεγαλυτερη ειναι 99.10 (Max = 99.10). Επιπροσθέτως, σύμφωνα με το παραπάνω διάγραμμα φαίνεται να υπάρχει θετική συσχέτιση ανάμεσα στην μεταβλητη καθυστέρησης διακανονισμού και τις απαιτήσεις.

2.4 Ανάλυση Λογιστικής Παλινδρόμησης

Για τις ανάγκες της ανάλυσης θα επιχειρηθεί να προσαρμοστεί ένα μοντέλο λογιστικής παλινδρόμησης στα δεδομένα, όπου η μεταβλητή απόκρισης 'Αποζημίωση', μετασχηματίστηκε σε μία δίτιμη μεταβλητή, θεωρώντας ως υψηλές αποζημιώσεις, αυτές που υπερβαίνουν τις

50,000 Δολάρια Αυστραλίας. Ως επεξηγηματικές μεταβλητές θα θεωρηθούν ο βαθμός τραυματισμού *(inj_cat)*, όπως κωδικοποιήθηκε, σε τρίτιμη μεταβλητή με επίπεδα, 'χαμηλός', 'σοβαρός' και 'θάνατος', η νομική εκπροσώπηση *(legrep)* και η μεταβλητή 'operational time' *(op_time)*.

Αρχικά, θα επιχειρηθεί η διερεύνηση του βέλτιστου μοντέλου, χωρίς να επιχειρηθεί ερμηνεία των παραμέτρων, θεωρώντας ως μοντέλο βάσης για τις συγκρίσεις, το μηδενικό μοντέλο (Null Model), ή μοντέλο με τη σταθερά, όπως ορίζεται παρακάτω:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + e_i \, .$$

Θα εκτιμηθούν τρία εναλλακτικά μοντέλα και θα εξετασθεί η σημαντικότητα της εισαγωγής του πρώτου όρου-μεταβλητής, με χρήση ελέγχου X^2, του κριτήριου AIC και του κριτήριου BIC για να διερευνηθεί κατά πόσο οι νέες μεταβλητές βελτιώνουν την εκτίμηση. Συγκεκριμένα, μοντέλα που θα προσαρμοστούν στα δεδομένα παραθέτονται παρακάτω:

1. $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 inj\_cat_i + e_i \, ,$
2. $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1' legrep_i + e_i \, ,$
3. $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1'' op\_time_i + e_i \, .$

Κατά τη διαδικασία σύγκρισης των μοντέλων με ένα όρο σε σχέση με το σταθερό, παρατηρείται ότι και στις τρείς περιπτώσεις ο σταθερός όρος αλλά και οι όροι που εισέρχονται στο μοντέλο είναι στατιστικά σημαντικοί σε επίπεδο στατιστικής σημαντικότητας 1% (Πίνακας 2.6), υπονοώντας ότι οι τρείς μεταβλητές επεξηγούν ένα στατιστικά σημαντικό κομμάτι της επεξηγηματικής μεταβλητής.

Πίνακας 2.6: Εκτίμηση Παραμέτρων ανά μοντέλο
(Μοντέλο σταθερού όρου έναντι μοντέλων με έναν όρο)

Μοντέλο Λογιστικής Παλινδρόμησης	Term	Estimate	Std.Error	z value	Pr(>|z|)
total_cat ~ 1	(Intercept)	-1.48	0.02	-82.70	<2e-16 ***
	(Intercept)	-1.51	0.02	-82.56	< 2e-16***
	inj_cat severe injury	1.40	0.14	9.68	< 2e-16***
	inj_cat death	0.56	0.14	3.98	6.79e-05***
	(Intercept)	-1.66	0.03	-52.01	< 2e-16***
	legrepyes	0.26	0.04	6.84	7.68e-12 ***
	(Intercept)	-5.00	0.07	-66.96	<2e-16***
	op_time	0.06	0.00	56.48	<2e-16***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Συνεχίζοντας στη διαδικασία σύγκρισης για την επιλογή του βέλτιστου μοντέλου με έναν όρο, τα τρία κριτήρια συγκλίνουν στο ίδιο συμπέρασμα. Το μοντέλο με τον όρο *op_time*, βελτιώνει την εκτίμηση σε σχέση με το σταθερό μοντέλο, πετυχαίνοντας τη μεγιστη διαφορά της Deviance σε σχέση με το μοντέλο σταθερού όρου, αλλά και τη μικρότερη τιμή του AIC και BIC κριτηρίου.

Πίνακας 2.7: Πίνακας Ανάλυσης Διακύμανσης (Deviance) με χρήση ελέγχου Chi-sq
(Μοντέλο σταθερού όρου έναντι μοντέλων με έναν όρο)

Μοντέλο Λογιστικής Παλινδρόμησης	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)	AIC	BIC
total_cat ~ 1	20,669	19,821				19,823	19,831
total_cat ~ inj_cat	20,667	19,722	2	99	< 2.2e-16 ***	19,728	19,752
total_cat ~ legrep	20,668	19,773	1	48	4.623e-12 ***	19,777	19,793
total_cat ~ op_time	20,668	14,700	1	5,121	< 2.2e-16 ***	14,704	14,720
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1							

Θεωρώντας ως βέλτιστο μοντέλο, αυτό στο οποίο έχει εισέλθει η operational time μεταβλητή, θα συνεχιστεί η εισαγωγή όρων, έτσι ώστε να διερευνηθεί η περαιτέρω βελτίωση της εκτίμησης σε σχέση με το νέο μοντέλο βάσης προς σύγκριση.

Με την ίδια λογική εισάγοντας αρχικά σαν δεύτερο όρο την μεταβλητή inj_cat και την μεταβλητή legrep, η παλινδρόμηση είναι ξανά στατιστικά σημαντική σε επίπεδο σηματικότητας 5%, δεδομένου ότι στο μοντέλο έχει εισέλθει η μεταβλητή operational time (Πίνακας 2.8).

Πίνακας 2.8: Εκτίμηση Παραμέτρων ανά μοντέλο
(Μοντέλο με operational time έναντι μοντέλων με δύο όρους)

Μοντέλο Λογιστικής Παλινδρόμησης	Term	Estimate	Std.Error	z value	Pr(>|z|)
	(Intercept)	-5.00	0.07	-66.96	<2e-16***
	op_time	0.06	0.00	56.48	<2e-16***
	(Intercept)	-5.01	0.07	-66.97	< 2e-16***
	op_time	0.06	0.00	56.23	< 2e-16***
	inj_cat severe injury	0.45	0.16	2.75	0.005.98**
	inj_cat death	0.88	0.18	4.83	1.37e-06***
	(Intercept)	-5.17	0.08	-63.59	< 2e-16***
	op_time	0.06	0.00	56.41	< 2e-16***
	legrep yes	0.25	0.04	5.65	1.62ε-08***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Το μοντέλο το οποίο φαίνεται να βελτιώνει την εκτίμηση με βάση τον ελέγχο X^2, είναι αυτό που στο οποίο ο δεύτερος όρος είναι η νομική εκπροσώπηση. Το συγκεκριμένο μοντέλο φαίνεται να είναι αυτό που πετυχαίνει και την μικρότερη τιμή των κριτηρίων AIC και BIC (Πίνακας 2.9).

Πίνακας 2.9: Πίνακας Ανάλυσης Απόκλισης (Deviance) με χρήση ελέγχου Chi-sq
(Μοντέλο με operational time έναντι μοντέλων με δύο όρους)

Μοντέλο Λογιστικής Παλινδρόμησης	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)	AIC	BIC
total_cat ~ op_time	20,668	14,700				14,704	14,720
total_cat ~ op_time + inj_cat	20,666	14,671	2	29	3.9e-07***	14,679	14,710
total_cat ~ op_time + legrep	20,667	14,668	1	32	1.33e-08***	14,674	14,698
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1							

Τέλος, εισάγοντας και τον τρίτο όρο (inj_cat) στο μοντέλο , παρατηρείται ότι οι συντελεστές είναι στατιστικά σημαντικοί σε επίπεδο σημαντικότητας 5% για την μεταβλητή 'τραυματισμός' (Πίνακας 2.10). Η εισαγωγή του τρίτου όρου φαίνεται να βελτιώνει περαιτέρω την εκτίμηση κάτι που υποστηρίζεται και από τον έλεγχο X^2 αλλά και με την χρήση κριτηρίων AIC και BIC (Πίνακας 2.11).

Πίνακας 2.10: Εκτίμηση Παραμέτρων ανά μοντέλο (Μοντέλο με τρείς όρους)

Μοντέλο Λογιστικής Παλινδρόμησης	Term	Estimate	Std.Error	z value	Pr(>|z|)
	(Intercept)	-5.17	0.08	-63.59	< 2e-16***
	op_time	0.06	0.00	56.41	< 2e-16***
	legrep yes	0.25	0.04	5.65	1.62ε-08***
	(Intercept)	-5.18	0.08	-63.61	< 2e-16***
	op_time	0.06	0.00	56.15	< 2e-16***
	legrep yes	0.26	0.04	5.87	4.39ε-09***
	inj_cat severe injury	0.48	0.16	2.93	0.00343**
	inj_cat death	0.91	0.18	5.02	5.31ε-07***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					

Πίνακας 2.11: Πίνακας Ανάλυσης Απόκλισης (Deviance) με χρήση ελέγχου Chi-sq (Μοντέλο με τρείς όρους)

Μοντέλο Λογιστισκής Παλινδρόμησης	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)	AIC	BIC
total_cat ~ op_time + legrep	20,667	14,668				14,674	14,698
total_cat ~ op_time + legrep + inj_cat	20,665	14,636	2	32	1.0793-07***	14,646	14,685
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1							

Επιχειρώντας να μελετηθεί η στατιστική σημαντικότητα, όρων αλληλεπίδρασης στο μοντέλο, εκτιμήθηκαν τα τρία παρακάτω υποδείγματα:

1. $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \beta_0 + \beta_1 op\_time_i + \beta_2 legrep_i + \beta_3 inj\_cat_i + \beta_4 op\_time\_legrep_{i_i}$,
2. $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \beta_0 + \beta_1 op\_time_i + \beta_2 legrep_i + \beta_3 inj\_cat_i + \beta_4^{'} op\_time\_inj\_cat_i$,
3. $\log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \beta_0 + \beta_1 op\_time_i + \beta_2 legrep_i + \beta_3 inj\_cat_i + \beta_4^{''} legrep\_inj\_cat_i$,

Οι έλεγχοι χ^2 απορρίπτουν την υπόθεση της βελτίωσης της εκτίμησης σε σύγκριση με το μοντέλο με τους τρείς όρους σε επίπεδο στατιστικής σημαντικότητς 5%, στα δύο από τα τρία μοντέλα, προκρίνοντας την εισαγωγή του όρου αλληλεπίδρασης 'Βαθμός Ατυχήματος*Νομική Εκπροσώπηση' Το συγκεκριμένο μοντέλο φαίνεται να μειώνει περαιτέρω την τιμή του κριτηρίου AIC αλλά όχι την αντίστοιχη του κριτηρίου BIC η οποία παρουσιάζεται ελάχιστα αυξημένη. (Πίνακας 2.12).

Πίνακας 2.12: Πίνακας Ανάλυσης Διακύμανσης (Deviance) με χρήση ελέγχου Chi-sq (Μοντέλα με αλλλεπίδραση)

Μοντέλο Λογιστισκής Παλινδρόμησης	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)	AIC	BIC
total_cat ~ op_time + legrep + inj_cat	20,665	14,636				14,646	14,685
total_cat ~ op_time * legrep + inj_cat	20,664	14,634	1	1.3318	0.2485	14,646	14,694
total_cat ~ op_time * inj_cat + legrep	20,663	14,632	2	3.2914	0.1929	14,646	14,702
total_cat ~ legrep * inj_cat + op_time	20,663	14616	2	19.672	5.348e-05 ***	14,630	14,686

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Παρατηρείται ότι η εκτίμηση των συντελεστών της παλινδρόμησης για τις μεταβλητές νομική εκπροσώπηση και βαθμός ατυχήματος διαφοροποιούνται. Ο αντίστοιχος συντελεστής ομώς για το επίπεδο 'Βαθμός Ατυχήματος' = Σοβαρό φαίνεται να είναι μη στατιστικά σημαντικός σε επίπεδο στατιστικής σημαντικότηας 5%. Αντίθετα, ο όρος της αλληλεπίδρασης είναι στατιστικά σημαντικός (Πίνακας 2.13).

Πίνακας 2.13: Εκτίμηση Παραμέτρων ανά μοντέλο
(Μοντέλο με αλληλεπίδραση 'Βαθμός Ατυχήματος*Νομική Εκπροσώπηση')

Μοντέλο Λογιστικής Παλινδρόμησης	Term	Estimate	Std.Error	z value	Pr(>|z|)
	(Intercept)	-5.16	0.08	-63.29	0.00
	op_time	0.06	0.00	56.14	0.00
	inj_cat severe injury	-0.27	0.26	-1.02	0.307
	inj_cat death	0.49	0.28	1.76	<0.05
	Legrep yes	0.23	0.05	5.00	0.00
	inj_cat severe injury:legrep yes	1.34	0.35	3.86	0.00
	inj_cat death:legrep yes	0.76	0.37	2.05	<0.05

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Σαν επόμενο βήμα, εισάγοντας έναν δεύτερο όρο αλληλεπίδρασης στο μοντέλο, ο έλεγχος X^2 απορρίπτει και στις δύο περιπτώσεις την υπόθεση βελτίωσης της εκτίμησης συγκριτικά με το μοντέλο σύγκρισης, υπονοώντας ότι το μοντέλο με την αλληλεπίδραση'Βαθμός Ατυχήματος*Νομική Εκπροσώπηση' ειναι επαρκές. Επιπρόσθετα, παρατηρείται ότι οι τιμές για τα κριτήρια AIC και BIC δεν μειώνονται περαιτέρω (Πίνακας 2.14).

Πίνακας 2.14: Πίνακας Ανάλυσης Διακύμανσης (Deviance) με χρήση ελέγχου Chi-sq (Μοντέλα με δύο αλληλεπιδράσεις)

Μοντέλο Λογιστισκής Παλινδρόμησης	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)	AIC	BIC
total_cat ~ legrep * inj_cat + op_time	20,663	14,616				14,630	14,686
total_cat ~ op_time * legrep + inj_cat * legrep	20,662	14,615	1	1.0087	0.3152	14,631	14,695
total_cat ~ op_time * inj_cat + inj_cat * legrep	20,661	14,614	2	2.4379	0.2955	14,631	14,703

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Συνεπώς, με την μέθοδο των διαδοχικών ελέγχων X^2 ένταξης μεταβλητών στο υπόδειγμα λογιστικής παλινδρόμησης και το κριτήριο AIC, το βέλτιστο μοντέλο είναι εκείνο που περιέχει τις τρείς κύριες μεταβλητές και έναν όρο αλληλεπίδρασης, όπως περιγράφεται παρακάτω:

$$\textbf{Μοντέλο 1}: \log\left(\frac{p_i}{1-p_i}\right) = -5.18 + 0.06 * \left[op_{time_i}\right] +$$

$$+0.26 * [legrep_i = yes] - 0.27 * \left[inj_{cat_i} = severe\right] +$$

$$+0.49 * [inj_{cat_i} = death] + 1.34 * [inj_{cat_i} = severe * legrep_i = yes] +$$

$$+0.76 * [inj_{cat_i} = death * legrep_i = yes] + e_i$$

Αντίθετα, με τη χρήση του κριτηρίου BIC επαρκές είναι το μοντέλο χωρίς την αλληλεπίδραση:

$$\textbf{Μοντέλο 2}: \log\left(\frac{p_i}{1-p_i}\right) = -5.18 + 0.06 * \left[op_{time_i}\right] + 0.26 * [legrep_i = yes] +$$

$$+0.48 * \left[inj_{cat_i} = severe\right] + 0.91 * [inj_{cat_i} = death] + e_i$$

2.5 Επιλογή Μοντέλου με την Μέθοδο Stepwise και Backward

Επιχειρώντας να εξεταστεί εάν η επιλογή του παραπάνω υποδείγματος είναι συστηματική, θα επιχειρηθεί η δερεύνηση βέλτιστου μοντέλου χρησιμοποιοώντας την ρουτίνα stepAIC() για τις τρεις παρακάτω μεθόδους:

- Stepwise Regression με μοντέλο εκκίνησης το υπόδειγμα σταθερού όρου.
- Stepwise Regression με μοντέλο εκκίνησης το κορεσμένο υπόδειγμα.
- Backward Elimination με μοντέλο εκκίνησης το κορεσμένο υπόδειγμα.

Από την ανάλυση προκύπτει ότι χρησιμοποιώντας σαν κριτήριο επιλογής το BIC, το βέλτιστο μοντέλο δεν συμπίπτει με το μοντέλο που επιλέχθηκε από την ανάλυση που παριγράφηκε παραπάνω. Ανεξάρτητα από τον αλγόριθμο επιλογής μεταβλητών (Backward/Stepwise) το μοντέλο που ελαχιστοποιεί το κριτήριο BIC είναι εκείνο που περιέχει τις τρείς κύριες μεταβλητές χωρίς αλληλεπίδραση.

Αντίθετα, με χρήση του κριτηρίου AIC, οι αλγόριθμοι τείνουν να καταλήγουν σε μεγαλύτερα μοντέλα. Συγκεκριμένα, η διαδικασία Stepwise Regression με μοντέλο εκκίνησης το μοντέλο σταθερού όρου επιλέγει ως βέλτιστο το μοντέλο με την αλληλεπίδραση 'Βαθμός Ατυχήματος*Νομική Εκπροσώπηση', ενώ οι διαδικασίες Backward Elimination και Stepwise Regression με μοντέλο εκκίνησης το κορεσμένο, καταλήγει στο μοντέλο με όλες τις αλληλεπιδράσεις (Πίνακες 2.15 & 2.16).

Πίνακας 2.15: Επιλογή Βέλτιστου Υποδείγματος με χρήση κριτηρίων AIC και BIC (Regression Selection)
Μοντέλο Εκκίνησης: Υπόδιεγμα Σταθερού Όρου

Model (AIC Criterion)	Df	Deviance	AIC	Model (BIC Criterion)	Df	Deviance	BIC
total_cat ~ 1			19,823	total_cat ~ 1			19,831
op_time	1	14,700	14,704	op_time	1	14,700	14,720
legrep	1	14,668	14,674	legrep	1	14,668	14,698
inj_cat	2	14,636	14,646	inj_cat	2	14,636	14,685
inj_cat:legrep	2	14,616	14,630	**<none>**		14,636	14,685
<none>		14,616	14,630	inj_cat:legrep	2	14,616	14,686
op_time:legrep	1	14,615	14,631	op_time:legrep	1	14,634	14,694
op_time:inj_cat	2	14,614	14,632	op_time:inj_cat	2	14,632	14,702

Πίνακας 2.16: Επιλογή Βέλτιστου Υποδείγματος με χρήση κριτηρίων AIC και BIC (Backward Elimination & Stepwise Regression)
Μοντέλο Εκκίνηση: Κορεσμένο Υπόδειγμα

Model (AIC Criterion)	Df	Deviance	AIC
total_cat ~ op_time * legrep * inj_cat			14,632
<none>		14,608	14,632
op_time:legrep:inj_cat	2	14,613	14,633

Model (BIC Criterion)	Df	Deviance	BIC
total_cat ~ op_time * legrep * inj_cat			14,727
op_time:legrep:inj_cat	2	14,613	14,712
op_time:inj_cat	2	14,615	14,694
op_time:legrep	1	14,616	14,686
legrep:inj_cat	2	14,636	14,685
<none>		14,636	14,685
inj_cat:legrep	2	14,616	14,686
op_time:legrep	1	14,634	14,694
inj_cat	2	14,668	14,698
op_time:inj_cat	2	14,632	14,702
legrep	1	14,671	14,710
op_time	1	19,670	19,710

Σε αυτό το σημείο πρέπει να αναφερθεί οτι ο όρος <none> τον οποίο παρατηρούμε στα πίνακες 2.15 και 2.16, χρησιμοποιείται ως ενα είδος ορίου για την επιλογή του μοντέλου. Πιο συγκεκριμένα μέχρι να συναντήσουμε τον όρο <none>, παρατηρούμε οτι τα κριτήρια AIC και BIC μειώνονται. Αυτό έχει ως αποτέλεσμα να μας οδηγήσει στην επιλογή των μεταβλητών που βρίσκονται πάνω απο τον όρο <none> για το βέλτιστο μοντέλο. Για τις μεταβλητές που βρίσκονται κάτω απο τον όρο <none>, παρατηρείται ότι τα κριτήρια AIC και BIC αυξάνονται και αυτό καθιστά αδύνατη την επιλογή τους στο μοντέλο.

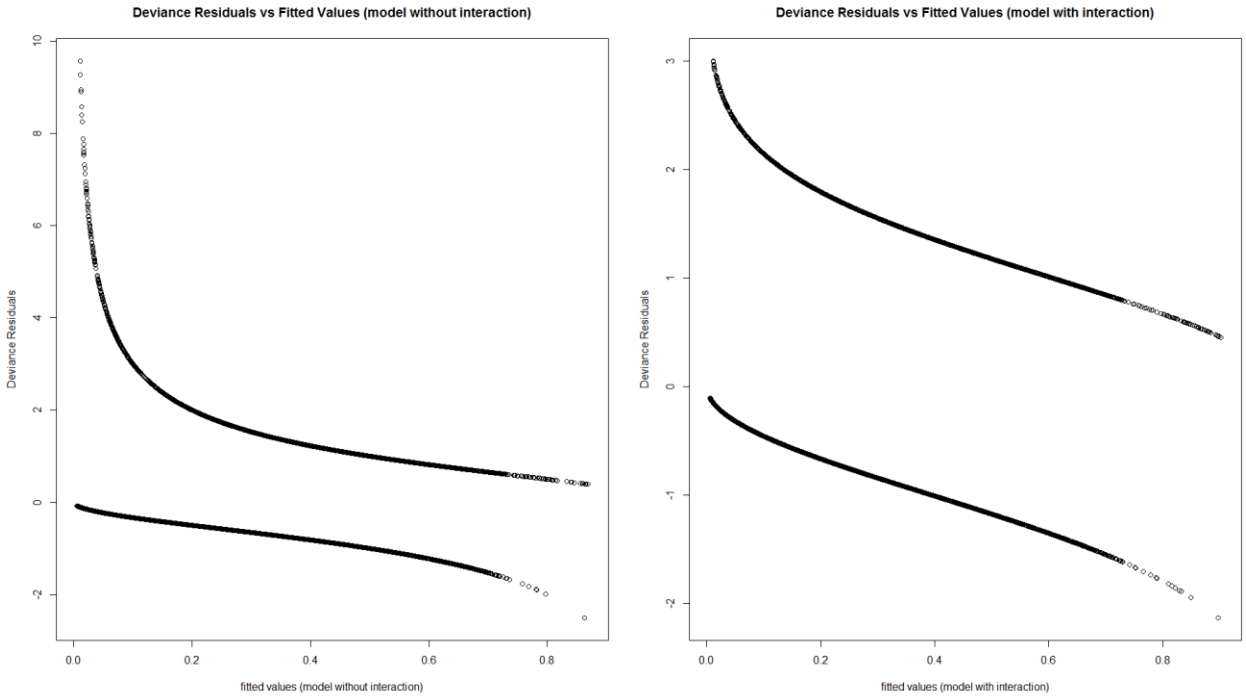
Λόγω του γεγονότος ότι μέθοδοι επιλογής μεταβλητών καταλήγουν σε διαφορετικά υποδείγματα, θα επιχειρηθεί η περαιτέρω διερεύνηση βέλτιστου υποδείγματος με τη χρήση του ελέγχου καλής προσαρμογής Hosmer Lemeshow (Πίνακας 2.17).

Πίνακας 2.17: Έλεγχος Καλής Προσαρμογής Hosmer Lemeshow

Μοντέλο Λογιστικής Παλινδρόμησης	Chi-sq	df	Pr(>Chi)
total_cat ~ op_time + legrep + inj_cat	80.94	8	0.00
total_cat ~ legrep * inj_cat + op_time	72.95	8	0.00

Παρατηρείται ότι και για τα δύο μοντέλα που εκτιμήθηκαν, ο έλεγχος καλης προσαρμογής απορρίπτει την μηδενική υπόθεση, υπονοώντας ότι σε επίπεδο σημαντικότητας 5% ότι οι προβλεπόμενες τιμές διαφέρουν σημαντικά από τις παρατηρούμενες.

Επιχειρώντας να εξετάσουμε την συμπεριφορά των καταλοίπων στο διάγραμμα των καταλοίπων Deviance έναντι των προβλεπόμενων τιμών παρατηρείται ότι οι εκτιμήσεις που συνδέονται με το επίπεδο αποζημίωσης «Μεγάλη», παρουσιάζουν μεγάλες τιμές κάτι το οποίο ενδεχομένως να συνδέεται με την έντονη ασυμμετρία της μεταβλητής απόκρισης.



Αναφορικά με την ερμηνεία των παραμέτρων, ξεκινώντας από το υπόδειγμα χωρίς την αλληλεπίδραση,

$$\textbf{Μοντέλο 2}: \log\left(\frac{p_i}{1-p_i}\right) = -5.18 + \left[0.06 * op_{time_i}\right] + 0.26 * [legrep_i = yes] +$$

$$+0.48 * \left[inj_{cat_i} = severe\right] + 0.91 * [inj_{cat_i} = death] + e_i$$

για άτομα με μηδενική τιμή της μεταβλητής operational time, χωρίς νομική υποστήριξη και απλού ατυχήματος, η πιθανότητα να λάβει μεγάλη αποζημίωση είναι μειωμένη κατά $e^{-5.18} = 0.005$ φορές συγκριτικά με την πιθανότητα να λάβει μικρή αποζημίωση. Σύμφωνα, λοιπόν με το εκτιμηθέν μοντέλο η πιθανότητα των παραπάνω ασφαλισμένων να λάβουν μικρή αποζημίωση, εκτιμάται σε $1 - \frac{e^{-5.18}}{1+e^{-5.18}} = 99.5\%$. Αύξηση της μεταβλητής Operational Time κατά μία μονάδα, εκτιμάται ότι αυξάνει την σχετική πιθανότητα να λάβει μεγάλη αποζημίωση κατά $e^{0.06} = 1.06$ φορές ή 6% για συγκεκριμένα επίπεδα ατυχήματος και νομικής υποστήριξης. Θεωρώντας μοναδιαία αύξηση της μεταβλητής Operational time η ερμηνεία της παραμέτρου μπορεί να υπολογιστεί μαθηματικά θεωρώντας τη διαφορά των δύο εξισώσεων. Ο υπολογισμός παρατίθεται παρακάτω:

$$\log\left(\frac{p_i}{1-p_i}|optime = 0\right) - \log\left(\frac{p_i}{1-p_i}|optime = 1\right) = 0.06 \leftrightarrow$$

$$\left(\frac{p_i}{1-p_i}|optime = 0\right) = e^{0.06} * \left(\frac{p_i}{1-p_i}|optime = 1\right)$$

Αντίστοιχα, η σχετική πιθανότητα όταν ο ασφαλισμένος έχει νομική υποστήριξη είναι κατά 30% μεγαλύτερη συγκριτικά με τον ασφαλισμένο οποίος δεν έχει νομική υποστήριξη, για συκερκιμένο τύπο ατυχήματος και σταθερό Operational Time. Τέλος, αναφορικά με το επίπεδο ατυχήματος, η σχετική πιθανότητα ενός ασφαλισμένου να λάβει μεγάλη αποζημίωση δεδομένου ενός σοβαρού ατυχήματος, εκτιμάται ότι είναι αυξημένη κατά 60% συγκριτικά με τις περιπτώσεις ασφαλισμένων με απλό ατύχημα, ενώ σε περίπτωση θανάτου η αντίστοιχη σχετική πιθανότητα λήψης μεγάλης αποζημίωσης, εκτιμάται ότι είναι κατά 2.5 φορές περίπου μεγαλύτερη συγκριτικά με τις περιπτώσεις απλού ατυχήματος, για δεδομένη νομική υποστήριξη και σταθερή τιμή operational time.

Στο μοντέλο με την αλληλεπίδραση,

$$\textbf{Μοντέλο 1}\text{: } \log\left(\frac{p_i}{1-p_i}\right) = -5.18 + 0.06 * [op_{time_i}] +$$

$$+0.26 * [legrep_i = yes] - 0.27 * [inj_{cat_i} = severe] +$$

$$+0.49 * \left[inj_{cat_i} = death\right] + 1.34 * \left[inj_{cat_i} = severe * legrep_i = yes\right] +$$

$$+0.76 * [inj_{cat_i} = death * legrep_i = yes] + e_i$$

παρατηρείται πως ο σταθερός όρος και η εκτίμηση της παραμέτρου της μεταβλητής operational time δεν άλλαξε. Αναφορικά με την ερμηνεία των κατηγορικών όρων, 'Νομική Υποστήριξη' και 'Ατύχημα', η ερμηνεία αλλάζει λόγω της σημαντικότητας του όρου της αλληλεπίδρασης. Συνεπώς, θεωρώντας τις τιμές για τα διαφορετικά επίπεδα των μεταβλητών οι εξισώσεις μεταβάλονται όπως παρακάτω:

Για τους ασφαλισμένους με σταθερό operational time , σοβαρό ατύχημα και χωρίς νομική υποστήριξη

$$\log\left(\frac{p_i}{1-p_i}|optime = 0, inj = severe, legrep = no\right) =$$

$$-5.18 + 0.26 * [legrep = 0] - 0.54 * [injcat = 2] + 1.34 * [legrep = 0 * injcat = 2]$$

Για τους ασφαλισμένους με σταθερό operational time , σοβαρό ατύχημα και νομική υποστήριξη

$$\log\left(\frac{p_i}{1-p_i}|optime = 0, inj = severe, legrep = yes\right) =$$

$$-5.18 + 0.26 * [legrep = 1] - 0.54 * [injcat = 2] + 1.34 * [legrep = 1 * injcat = 2]$$

Για τους ασφαλισμένους με σταθερό operational time, θανατηφόρο ατύχημα αλλά χωρίς νομική υποστήριξη

$$\log\left(\frac{p_i}{1-p_i}|optime = 0, inj = death, legrep = no\right) =$$

$$-5.18 + 0.26 * [legrep = 0] + 0.49 * [injcat = 3] + 0.76 * [legrep = 0 * injcat = 3]$$

Για τους ασφαλισμένους με σταθερό operational time ,θανατηφόρο ατύχημα και νομική υποστήριξη

$$\log\left(\frac{p_i}{1-p_i}|optime = 0, inj = death, legrep = yes\right) =$$

$$-5.18 + 0.26 * [legrep = 1] + 0.49 * [injcat = 3] + 0.76 * [legrep = 1 * injcat = 3]$$

Σύνφωνα με το μοντέλο, η σχετική πιθανότητα οι ασφαλισμένοι με νομική υποστήριξη και σοβαρό ατύχημα να λάβουν μεγάλη αποζημίωση, είναι κατά 2.88 φορές μεγαλύτερη συγκριτικά με τους ασφαλισμένους η οποίοι δεν εκπροσωπούνται νομικά, ενώ η σχετική πιθανότητα των ασφαλισμένων με θανατηφόρο ατύχημα και νομική εκπροσώπηση εκτιμάται κατά 4.5 φορές μεγαλύτερη, συγκριτικά με τους αντίστοιχους που δεν εκπροσωπούνται νομικά.

ΚΕΦΑΛΑΙΟ 3

ΣΥΜΠΕΡΑΣΜΑΤΑ

Σκοπός αυτής της διπλωματικής εργασίας είναι η μελέτη των γενικευμένων γραμμικών μοντέλων εφαρμόζοντας αυτα σε αληθινά ασφαλιστικά δεδόμενα. Για την πραγματοποίηση αυτού χρησιμοποιήθηκαν ασφαλιστικά δεδομένα τα οποία μελέτησαν οι Piet de Jong και Gillian Z. Heller για τις ανάγκες του βιβλίου Generalized Linear Models for Insurance Data. Για την επίτευξη του σκοπού αυτού χρησιμοποιήθηκε το στατιστικό πακέτο R, μέσω του οποίου έγινε η επεξεργασια και η ανάλυση των δεδομένων για την επιλογή του καταλληλότερου μοντέλου.

Για τις ανάγκες της ανάλυσης επιχειρίθηκε να επιλεγεί το βέλτιστο μοντέλο εφαρμόζοντας την λογιστική παλινδρόμηση στα δεδομένα, όπου η μεταβλητή απόκρισης 'Αποζημίωση', μετασχηματίστηκε σε μία δίτιμη μεταβλητή. Ως επεξηγηματικές μεταβλητές θα θεωρηθούν:

- ο βαθμός τραυματισμού, όπως κωδικοποιήθηκε, σε τρίτιμη μεταβλητή με επίπεδα, 'χαμηλός', 'σοβαρός' και 'θάνατος',
- η κατηγορική μεταβλητή νομική εκπροσώπηση και
- η συνεχής μεταβλητη καθυστέρησης διακανονισμού.

Αρχικά θα εξετάστηκε η σημαντικότητα της εισαγωγής του πρώτου όρου-μεταβλητής, με χρήση ελέγχου X^2, του κριτήριου AIC και του κριτήριου BIC για να διερευνηθεί κατά πόσο οι νέες μεταβλητές βελτιώνουν την εκτίμηση. Για την πραγματοποιήση του παραπάνω χρησιμοποιήθηκαν οι παρακάτω εντολές στην γλώσσα προγραμματισμού R.

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

Summary()

anova(,test='Chisq')

AIC() και BIC()

Κατά την διαδικασία επιλογής του βέλτιστου μοντέλου παρατηρούμε οτι ο όρος *operational time (*καθυστέρηση διακανονισμού), βελτιώνει την εκτίμηση σε σχέση με το σταθερό μοντέλο, πετυχαίνοντας τη μεγιστη διαφορά της Deviance σε σχέση με το μοντέλο σταθερού όρου, αλλά και τη μικρότερη τιμή του AIC και BIC κριτηρίου. Εν συνεχεία, για την περαιτέρω βελτιστοποίηση του μοντέλου, εισέρχεται ως δεύτερος όρος η μεταβλητή Legal representation (νομική εκπροσώπηση). Τέλος, εισαγωγή του τρίτου όρου injury (βαθμός τραυματισμού), φαίνεται να βελτιώνει περαιτέρω την εκτίμηση. Επιπροσθέτως, μελετήθηκε αν κάποια αλληλεπίδραση απο τις παραπάνω μεταβλητές βελτιώνει περαιτέρω το μοντέλο, προκρίνοντας την εισαγωγή του όρου αλληλεπίδρασης 'Βαθμός Ατυχήματος*Νομική Εκπροσώπηση', θέτοντας όμως το επίπεδο 'Βαθμός Ατυχήματος' = Σοβαρό μη στατιστικά σημαντικό.

Συνεπώς, με την μέθοδο των διαδοχικών ελέγχων X^2 ένταξης μεταβλητών στο υπόδειγμα λογιστικής παλινδρόμησης και το κριτήριο AIC, το βέλτιστο μοντέλο είναι εκείνο που περιέχει τις τρείς κύριες μεταβλητές και έναν όρο αλληλεπίδρασης, όπως περιγράφεται παρακάτω:

$$\textbf{Μοντέλο 1}\colon \log\left(\frac{p_i}{1-p_i}\right) = -5.18 + 0.06 * \left[op_{time_i}\right] +$$

$$+0.26 * [legrep_i = yes] - 0.27 * \left[inj_{cat_i} = severe\right] +$$

$$+0.49 * [inj_{cat_i} = death] + 1.34 * [inj_{cat_i} = severe * legrep_i = yes] +$$

$$+0.76 * [inj_{cat_i} = death * legrep_i = yes] + e_i$$

Αντίθετα, με τη χρήση του κριτηρίου BIC επαρκές είναι το μοντέλο χωρίς την αλληλεπίδραση:

$$\textbf{Μοντέλο 2}\colon \log\left(\frac{p_i}{1-p_i}\right) = -5.18 + 0.06 * \left[op_{time_i}\right] + 0.26 * [legrep_i = yes] +$$

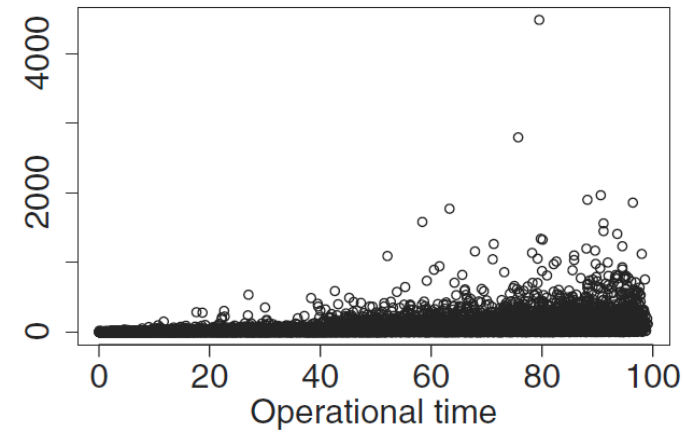
$$+0.48 * \left[inj_{cat_i} = severe\right] + 0.91 * [inj_{cat_i} = death] + e_i$$

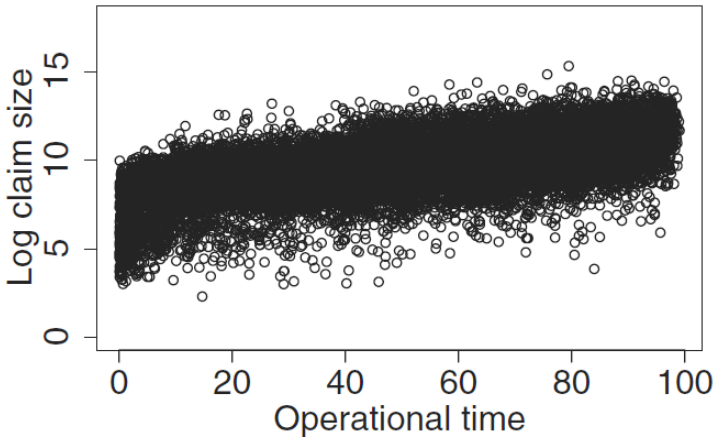
Στη συνέχεια, εξετάστηκε εάν η επιλογή του παραπάνω υποδείγματος είναι συστηματική και για αυτό το λόγο επιχειρήθηκε η δερεύνηση βέλτιστου μοντέλου χρησιμοποιοώντας την εντολή stepAIC() για τις μεθόδους Stepwise και Backward. Χρησιμοποιώντας σαν κριτήριο

επιλογής το BIC, το βέλτιστο μοντέλο δεν συμπίπτει με το μοντέλο που επιλέχθηκε από την ανάλυση που παριγράφηκε παραπάνω αφού το μοντέλο που ελαχιστοποιεί το κριτήριο BIC είναι εκείνο που περιέχει τις τρείς κύριες μεταβλητές χωρίς αλληλεπίδραση.

Αντίθετα, με χρήση του κριτηρίου AIC, καταλήγουμε σε μεγαλύτερα μοντέλα. Συγκεκριμένα, η διαδικασία με μοντέλο εκκίνησης το μοντέλο σταθερού όρου επιλέγει ως βέλτιστο το μοντέλο με την αλληλεπίδραση 'Βαθμός Ατυχήματος*Νομική Εκπροσώπηση', ενώ οι διαδικασίες το μοντέλο εκκίνησης το κορεσμένο, καταλήγει στο μοντέλο με όλες τις αλληλεπιδράσεις.

Εν αντιθέσει, στην ανάλυση τους οι Piet de Jong και Gillian Z. Heller στο βιβλίο τους χρησιμοποιούν την Γαμμα κατανομή για την μεταβλητή των αποζημειώσεων αφού δεν την μετατρέπουν σε μεταβλητή με δίτιμα δεδομενα. Ετσι εχουν μια συνεχή μεταβλητή απόκρισης την οποία στην συνέχεια την λογαριθμίζουν.Ο στόχος του μετασχηματισμού αυτου είναι να καταστεί πιο ευκολότερα επιδεκτικές τις μεταβλητές οι στατιστική ανάλυση.





Αποδεικνύεουν ότι η σχέση μεταξύ των αποζημειώσεων και της καθυστέρησης διακανονισμού είναι κατά προσέγγιση γραμμική. Επομένως καταλήγουν σε ένα μοντέλο με επεξηγηματικές μεταβλητές νομική εκπροσώπηση, καθυστέρηση διακανονισμού καθως και την αλληλεπίδραση τους το οποίο έχει την μορφή:

$$y \sim G(\mu, \nu), \qquad lm\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

όπου x_1= καθυστέρηση διακανονισμού

x_2= νομική εκπροσώπηση

$x_1 x_2$= η αλληλεπίδραση των μεταβλητών καθυστέρηση διακανονισμού και νομική εκπροσώπηση.

ΒΙΒΛΙΟΓΡΑΦΙΑ

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

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

Piet de Jong and Gillian Z. Heller (2008). Generalized Linear Models for Insurance Data. Cambridge University Press

Κωσταντίνος Πολίτης, Σημειώσεις Μαθήματος, Γενικευμένα γραμμικά μοντέλα, 2014 (Πανεπιστημιο Πειραιώς)

Γεώργιος Τζαβελλάς, Σημειώσεις στα Γενικευμένα Γραμμικά Μοντέλα (Πανεπιστημιο Πειραιώς)

Δημήτριος Φουσκάκης, (2013). Ανάλυση Δεδομένων με Χρήση της R, Εκδόσεις Τσότρας. Αθήνα

Κωνσταντίνος Φωκιανός & Χαράλαμπος Χαραλάμπους. Εισαγωγή στην R, Τμήμα Μαθηματικών & Στατιστικής. Πανεπιστήμιο Κύπρου. 2η Έκδοση: Ιανουάριος 2010

ΠΑΡΑΡΤΗΜΑ

Σύνταξη για την προσαρμογή ενός μοντέλου στην R

Η γενική σύνταξη για την προσαρμογή ενός ΓΓΜ είναι
glm(formula, family, data, weights, control)
όπου
formula : δίνει τη μαθηματική περιγραφή του υποδείγματος, δηλαδή την εξαρτημένη και τις ερμηνευτικές μεταβλητές που έχουμε επιλέξει π.χ. $p{\sim}x1 + x2 + x3$

family : δηλώνει την οικογένεια κατανομών από την οποία προέρχεται η κατανομή της Y . Για ένα διωνυμικό μοντέλο, η προεπιλεγμένη συνάρτηση είναι η logit. Διαφορετικά, αν θέλουμε να χρησιμοποιήσουμε π.χ. τη συνάρτηση Probit, η σύνταξη είναι family=binomial(link=probit).

Data: είναι ο πίνακας δεδομένων (data frame) που περιέχει όλες τις μεταβλητές που χρησιμοποιούμε.

Σχετική πιθανότητα (odds)

Γενικά, η σχετική πιθανότητα ενός ενδεχομένου Α ορίζεται ως ο λόγος

$$\frac{P(A)}{1-P(A)} = \frac{P(A)}{P(A^C)},$$

όπου $P(A)$ δηλώνει την πιθανότητα να συμβεί το ενδεχόμενο A.

Τιμή της σχετικής πιθανότητας μεγαλύτερη του 1 δηλώνει ότι το ενδεχόμενο στον αριθμητή είναι πιο πιθανό να συμβεί από αυτό στον παρονομαστή.

Η έννοια της σχετικής πιθανότητας είναι σημαντική για την ερμηνεία των παραμέτρων σε ένα μοντέλο λογιστικής παλινδρόμησης.

Βασικός λόγος που η συνάρτηση logit προτιμάται σε σχέση με άλλες συναρτήσεις σύνδεσης σε λογιστικά μοντέλα είναι η εύκολη διαισθητική ερμηνεία των αποτελεσμάτων με βάση τη σχετική πιθανότητα.

Γνωρίζουμε ότι

$$logit(p) = log\left(\frac{p}{1-p}\right),$$

άρα η συνάρτηση logit αναφέρεται στο λογάριθμο της σχετικής πιθανότητας του ενδεχομένου που μας ενδιαφέρει.

Σημείωση:
Υπάρχει 1-1 αντιστοιχία ανάμεσα στην πιθανότητα επιτυχίας και το λογάριθμο της σχετικής πιθανότητας αφού

$$p = \frac{odds}{1 + odds}.$$

Για παράδειγμα, σε ένα μοντέλο με δύο μεταβλητές η σχέση

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

ή ισοδύναμα

$$\hat{p} = \frac{e^{(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2)}}{1 + e^{(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2)}}.$$

Από την παραπάνω σχέση γίνεται φανερό ότι η μοναδιαία αύξηση της τιμής της μεταβλητής X_1 προκαλεί πολλαπλασιαστική αύξηση της σχετικής πιθανότητας επιτυχίας κατά $exp(\hat{\beta}_1)$, στην περίπτωση όπου η μεταβλητή X_2 είναι σταθερή.

Εντολες στην R

```
install.packages ("data.table")

library(data.table)

getwd()


claims<-read.csv("persinj.csv",header=TRUE,sep=",")

View(claims)

claims$total_cat[claims$total< 50000] <- 0

claims$total_cat[claims$total>=50000] <- 1


claims$total_cat_factor<-factor(claims$total_cat,labels=c("low","high"))

claims$legrep<-factor(claims$legrep, labels=c("no","yes"))
```

```
injury<-subset(claims,select=c(inj1,inj2,inj3,inj4,inj5))
View(injury)
injury$max_inj<- apply(injury, 1, function(x)max(x[!is.na(x)]))

injury$max_inj_cat[injury$max_inj<=4]<-1
injury$max_inj_cat[injury$max_inj==5]<-2
injury$max_inj_cat[injury$max_inj==6]<-3
injury$max_inj_cat<-factor(injury$max_inj_cat,labels=c("plain injury","severe injury","death"))

claims<-cbind(claims,injury$max_inj_cat)
colnames(claims)[colnames(claims) == 'injury$max_inj_cat'] <- 'inj_cat'

claims_final<-subset(claims,select=c(total_cat,inj_cat,legrep,op_time))
claims_final<-na.omit(claims_final)

claims.NULL<-glm(total_cat~1,data=claims_final,family=binomial)

claims_inj<-glm(total_cat~inj_cat,data=claims_final,family=binomial)
claims_leg<-glm(total_cat~legrep,data=claims_final,family=binomial)
claims_op<-glm(total_cat~op_time,data=claims_final,family=binomial)

summary(claims.NULL)
summary(claims_inj)
summary(claims_leg)
summary(claims_op)
```

```
anova(claims.NULL,claims_inj,test='Chisq')
anova(claims.NULL,claims_leg,test='Chisq')
anova(claims.NULL,claims_op,test='Chisq')

claims_op_inj<-glm(total_cat~op_time+inj_cat,data=claims_final,family=binomial)
claims_op_leg<-glm(total_cat~op_time+legrep,data=claims_final,family=binomial)

summary(claims_op_inj)
summary(claims_op_leg)

anova(claims_op,claims_op_inj,test='Chisq')
anova(claims_op,claims_op_leg,test='Chisq')

AIC(claims_op_inj)
AIC(claims_op_leg)

BIC(claims_op_inj)
BIC(claims_op_leg)

  claims_op_leg_inj<-
glm(total_cat~op_time+legrep+inj_cat,data=claims_final,family=binomial)
  summary(claims_op_leg_inj)
 anova(claims_op_leg,claims_op_leg_inj,test='Chisq')
AIC(claims_op_leg_inj)
BIC(claims_op_leg_inj)



claims_opleg<-glm(total_cat~op_time*legrep+inj_cat,data=claims_final,family=binomial)
```

```
claims_opinj<-glm(total_cat~op_time*inj_cat+legrep,data=claims_final,family=binomial)
claims_leginj1<-glm(total_cat~legrep*inj_cat+op_time,data=claims_final,family=binomial)

summary(claims_opleg)
summary(claims_opinj)
summary(claims_leginj1)

anova(claims_op_leg_inj,claims_opleg,test='Chisq')
anova(claims_op_leg_inj,claims_opinj,test='Chisq')
anova(claims_op_leg_inj,claims_leginj1,test='Chisq')

AIC(claims_opleg)
AIC(claims_opinj)
AIC(claims_leginj1)

BIC(claims_opleg)
BIC(claims_opinj)
BIC(claims_leginj1)

summary(claims)
logtotal<-log(claims$total)
plot(claims$inj_cat,claims$total_cat_factor)
plot(op_time)
hist(claims$inj_cat)
summary(claims$total)
plot(claims$total)
pie(claims$legrep)
pie(claims$inj_cat)
```

```
LegalRepresentation<-table(claims$legrep)

pie(LegalRepresentation)


injury_cat<-table(claims$inj_cat)

pie(injury_cat)


summary(claims_leginj1)

claims_leginj_opleg<-
glm(total_cat~op_time*legrep+inj_cat*legrep,data=claims_final,family=binomial)

claims_leginj_opinj<-
glm(total_cat~op_time*inj_cat+inj_cat*legrep,data=claims_final,family=binomial)


anova(claims_leginj,claims_leginj_opleg,test='Chisq')

anova(claims_leginj,claims_leginj_opinj,test='Chisq')


AIC(claims_leginj_opleg)

AIC(claims_leginj_opinj)


BIC(claims_leginj_opleg)

BIC(claims_leginj_opinj)


modelNULL<-glm(total_cat~1,data=claims_final,family=binomial)

modelFULL<-glm(total_cat~.,data=claims_final,family=binomial)


modelAIC_CS<-stepAIC(modelNULL,direction='both',scope=list(upper=
~op_time*inj_cat*legrep,lower= ~1),trace=1)

modelBIC_CS<-
stepAIC(modelNULL,k=log(nrow(claims_final)),direction='both',scope=list(upper=
~op_time*inj_cat*legrep,lower= ~1),trace=1)
```

```
modelAIC_FS<-stepAIC(modelFULL,direction='both',scope=list(upper=
~op_time*inj_cat*legrep,lower= ~1),trace=1)

modelBIC_FS<-
stepAIC(modelFULL,k=log(nrow(claims_final)),direction='both',scope=list(upper=
~op_time*inj_cat*legrep,lower= ~1),trace=1)


modelAIC_FB<-stepAIC(modelFULL,direction='backward',scope=list(upper=
~op_time*inj_cat*legrep,lower= ~1),trace=1)

modelBIC_FB<-
stepAIC(modelFULL,k=log(nrow(claims_final)),direction='backward',scope=list(upper=
~op_time*inj_cat*legrep,lower= ~1),trace=1)


m1<-glm(total_cat~op_time+legrep+inj_cat,data=claims_final,family=binomial)

m2<-glm(total_cat~op_time+inj_cat*legrep,data=claims_final,family=binomial)

install.packages("ResourceSelection")

library(ResourceSelection)

hoslem.test(m1$y,fitted(m1),g=10)

hoslem.test(m2$y,fitted(m2),g=10)


res.m1<-resid(m1,type="deviance")

fit.m1<-fitted(m1)


res.m1<-resid(m1,type="deviance")

fit.m2<-fitted(m2)


par(mfrow=c(1,2))

plot(fit.m1,res.m1,main="Deviance Residuals vs Fitted Values (model without
interaction)",xlab="fitted values (model without interaction)",ylab="Deviance Residuals")

plot(fit.m2,res.m2,main="Deviance Residuals vs Fitted Values (model with
interaction)",xlab="fitted values (model with interaction)",ylab="Deviance Residuals")
```

```
summary(m1.c)
summary(m2.c)

hoslem.test(m1$y,fitted(m1),g=10)
hoslem.test(m2$y,fitted(m2),g=10)

claims_final$total_cat_factor<-factor(claims_final$total_cat,labels=c("low","high"))

claims_final$optime_cat[claims_final$op_time<median(claims_final$op_time) ] <- 0
claims_final$optime_cat[claims_final$op_time>=median(claims_final$op_time) ]  <- 1

claims_final$optime_cat<-factor(claims_final$optime_cat,labels=c("low","high"))

library("MASS", lib.loc="C:/Program Files/R/R-3.4.1/library")

tbl1 <- table(claims_final$total_cat_factor, claims_final$inj_cat)
tbl1
chisq.test(tbl1)

tbl2 <- table(claims_final$total_cat_factor, claims_final$legrep)
tbl2
chisq.test(tbl2)


tbl3<- table(claims_final$total_cat_factor, claims_final$optime_cat)
tbl3
chisq.test(tbl3)
```