ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Η ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΓΙΑ ΓΕΓΟΝΟΤΑ ΜΕ ΧΑΜΗΛΗ ΣΥΧΝΟΤΗΤΑ ΕΜΦΑΝΙΣΗΣ

Αθανάσιος Γ. Μπαγατέλας

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Ιούνιος 2020

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Επίκουρος Καθηγητής Τζαβελάς Γεώργιος (Επιβλέπων)
- Επίκουρος Καθηγητής Ευαγγελάρας Χαράλαμπος
- Αναπληρωτής Καθηγητής Πολίτης Κωνσταντίνος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

LOGISTIC REGRESSION IN RARE EVENTS DATA

By

Athanasios G. Bagatelas

MSc Dissertation

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

Piraeus, Greece
June 2020

Στους γονείς μου

Γεώργιο και Αικατερίνη

Ευχαριστίες

Ξεκινώντας, θα ήθελα να ευχαριστήσω τους γονείς μου για όλη την στήριξη που μου παρείχαν όλα αυτά τα χρόνια σε όλο αυτό το δύσβατο δρόμο της γνώσης. Από τις ευχαριστίες δεν θα μπορούσε να λείπει ο επιβλέπων Καθηγητής τις διπλωματικής μου εργασίας Γεώργιος Τζαβελάς που ήταν πάντοτε εκεί να με καθοδηγεί και να με στηρίζει σε κάθε δυσκολία που συνάντησα. Κλείνοντας, θα ήθελα να ευχαριστήσω το σύνολο των καθηγητών του ΠΜΣ Εφαρμοσμένης Στατιστικής για την αρίστη συνεργασία που είχαμε καθ' όλη την διάρκεια του μεταπτυχιακού προγράμματος καθώς επίσης και για τις γνώσεις που μου μεταλαμπάδευσαν.

Περίληψη

Στην εργασία αυτή εξετάζεται το στατιστικό πρόβλημα της εκτίμησης των σπάνιων γεγονότων με την κλασσική λογιστική παλινδρόμηση. Σπάνια γεγονότα θεωρούνται αυτά που έχουν συχνότητα εμφάνισης μικρότερη του 5%. Αρχικά, έγινε περιγραφή των βασικών χαρακτηριστικών και των ιδιοτήτων της λογιστικής παλινδρόμησης. Στην συνέχεια αναπτύχθηκαν τα δυο βασικά προβλήματα που συναντάμε σε τέτοιου είδους δεδομένα, τα οποία είναι η ελλιπής στρατηγικές που υπάρχουν για την συλλογή τέτοιων δεδομένων καθώς επίσης και η δυσκολία στο να εξηγηθούν και να προβλέψουν. Έπειτα καταγράφθηκαν αναλυτικά οι διορθώσεις που πρότειναν οι King Gary και Langche Zeng (2001) και με την βοήθεια προσομοιώσεων έγινε σύγκριση των μεθόδων και διαπιστώθηκε πως οι διορθώσεις αυτές βελτιώνουν αρκετά την ακρίβεια των εκτιμήσεων. Τέλος, χρησιμοποιώντας τις παραπάνω μεθόδους σε ένα πραγματικό σετ δεδομένων που αναφέρεται στον σακχαρώδη διαβήτη έγινε ξεκάθαρη η χρησιμότητα τους και σε πραγματικά δεδομένα.

Abstract

The purpose of this paper is to describe the statistical problem of estimating rare events by Logistic Regression. Rare events are the events that occur with low frequency (less than 5%). First, the basic concepts of the logistic regression are described and the main problems of the statistical analysis of such data are analyzed. These problems are the inefficient common used strategies for collecting data with rare events as well as the difficulty in explaining and predicting. Next the correction methods proposed by King Gary and Langche Zeng (2001) are explained in details and their efficiency in reducing the biased are compared with the help of simulated data. Finally, a real dataset that refers to diabetes were used in order to be more clear that the corrections improve the estimations.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Εισαγωγή

Στην παρούσα εργασία θα αντιμετωπίσουμε το στατιστικό πρόβλημα ανάλυσης δεδομένων χρησιμοποιώντας την λογιστική παλινδρόμηση σε σπάνια γεγονότα με δίτιμη εξαρτημένη μεταβλητή, όπου δηλαδή υπάρχουν πολύ λιγότερα γεγονότα με τιμή 1 από ότι με τιμή 0. Πριν όμως αναφερθούμε στο πρόβλημα αυτό και στις λύσεις του θα αναπτύξουμε στο πρώτο κεφάλαιο γενικά την μέθοδο της λογιστικής παλινδρόμησης αναφέροντας βασικά χαρακτηριστικά και ιδιότητες της.

Δεδομένα με σπάνια γεγονότα τα συναντάμε για παράδειγμα σε μολύνσεις από σπάνιες ασθένειες, πολέμους, πραξικοπήματα κ.α. Αυτού του είδους τα δεδομένα έχουν δυσκολία στο να εξηγηθούν και να προβλεφθούν. Τα προβλήματα που προκύπτουν έχουν πολλές πηγές. Στην παρούσα εργασία θα ασχοληθούμε με τις δύο πιο κύριες. Αυτές είναι το πρόβλημα της υποεκτίμησης των σπάνιων γεγονότων από την λογιστική παλινδρόμηση καθώς και η υπερβολικά ανεπαρκής στρατηγικές συλλογής δεδομένων τέτοιου είδους.

Στο λογιστικό μοντέλο όπου μοντελοποιούμε την δίτιμη εξαρτημένη μεταβλητή αρκετές υποθέσεις του κλασσικού μοντέλου όπως ομοσκεδαστικότητα, γραμμικότητα, κανονικότητα παραβιάζονται. Όταν τα δεδομένα αναφέρονται σε σπάνια γεγονότα τότε αυτές οι παραβιάσεις έχουν σημαντικές συνέπειες στην ανάλυση. Για παράδειγμα οι συντελεστές της λογιστικής παλινδρόμησης είναι μεροληπτικοί σε μικρά δείγματα αλλά αυτό που πρέπει να αναφέρουμε είναι ότι σε σπάνια γεγονότα η μεροληψία αυτή μπορεί να είναι ουσιαστικής σημασίας καταλήγοντας σε πολύ μικρές εκτιμήσεις των γεγονότων. Ένα άλλο πρόβλημα το οποίο παραβλέπετε είναι ότι σχεδόν όλοι οι γνωστοί μέθοδοι υπολογισμού των πιθανοτήτων των γεγονότων στο λογιστικό μοντέλο δεν είναι κατάλληλες για τα σπάνια γεγονότα με αποτέλεσμα να υπάρχουν λάθη που οδηγούν στην ίδια κατεύθυνση με την μεροληψία των συντελεστών δηλαδή στην υποεκτίμηση των γεγονότων. Για τα παραπάνω προβλήματα στην παρούσα εργασία θα αναπτύξουμε τις δύο διορθώσεις των εκτιμήσεων *(prior correction, weighting)* που πρότειναν οι King Gary και Langche Zeng (2001). Επίσης θα αναφέρουμε τις μεθόδους που πρότειναν οι King Gary και Langche Zeng (2001) καθώς και ο Firth (1993) για την μείωση της μεροληψίας στον υπολογισμό των παραμέτρων β_i, καθώς επίσης και τον υπολογισμό της πιθανότητας π_i με διόρθωση που πρότειναν οι King Gary και Langche Zeng (2001).

Μια δεύτερη πηγή δυσκολιών στην ανάλυση των σπάνιων γεγονότων είναι η συλλογή των δεδομένων. Υπάρχει πάντα μια διαμάχη ανάμεσα στο να συλλέξουμε περισσότερες

παρατηρήσεις και στο να συμπεριλάβουμε περισσότερες ή “καλύτερες” επεξηγηματικές μεταβλητές. Στα σπάνια γεγονότα το να συλλέξουμε δεδομένα χωρίς την ύπαρξη γεγονότων με τιμή $Y = 1$ οδηγεί στο να επιλέγουμε πολύ μεγάλα σύνολα δεδομένων με ελάχιστες επεξηγηματικές μεταβλητές. Αυτή είναι μια λογική επιλογή δεδομένων αλλά επειδή αρκετές φορές η ύπαρξη τόσο μεγάλων συνόλων δεδομένων δεν είναι εφικτή πολλοί ερευνητές οδηγήθηκαν στο να βρούνε πιο αποδοτικές στρατηγικές συλλογής δεδομένων, κάποιες από τις οποίες θα αναφέρουμε στην παρούσα εργασία στο κεφάλαιο 2. Πιο συγκεκριμένα θα αναπτύξουμε τις μεθόδους case-cohort και δικτυωτή δειγματοληψία. Η κύρια ιδέα της πρώτης είναι να επιλέγουμε όλα τα διαθέσιμα γεγονότα $(Y = 1)$ και να επιλέγουμε με τυχαία δειγματοληψία τα μη γεγονότα $(Y = 0)$. Η δεύτερη μέθοδος μας προτείνει να χρησιμοποιούμαι έναν τροποποιημένο τρόπο καταμέτρησης των γεγονότων ο οποίος οδηγεί σε εκτιμήσεις με μικρότερο σφάλμα δειγματοληψίας σε σχέση με τον συνηθισμένο τρόπο καταμέτρησης.

Αφού λοιπόν αναπτύξουμε τα παραπάνω, στο κεφάλαιο 5 θα προσομοιώσουμε δεδομένα για λογιστική παλινδρόμηση. Έπειτα τρέχοντας τα μοντέλα που προτείναμε θα παρατηρήσουμε ότι μας δίνουν καλύτερη εκτίμηση με μικρότερο σφάλμα σε σχέση με το μοντέλο της κλασσικής παλινδρόμησης όταν τα δεδομένα μας είναι μικρά και έχουν σπάνια γεγονότα.

Ο κύριος προορισμός λοιπόν αυτής της εργασίας είναι να ενσωματώσουμε αυτούς τους τύπους διορθώσεων τους οποίους και θα μελετήσουμε ξεχωριστά και να τους εφαρμόσουμε σε ένα σετ δεδομένων με πραγματικά στοιχεία έτσι ώστε για να γίνει πιο κατανοητή η χρησιμότητά τους.

Κ Ε Φ Α Λ Α Ι Ο 1

ΒΑΣΙΚΕΣ ΑΡΧΕΣ ΛΟΓΙΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΙΣΗΣ

Στο παρακάτω κεφάλαιο θα κάνουμε μια εισαγωγή στην λογιστική παλινδρόμηση και θα αναφέρουμε βασικές έννοιες και χαρακτηριστικά της.

1.1 Εισαγωγή στην Λογιστική Παλινδρόμηση

Η λογιστική παλινδρόμηση είναι ένα μη γραμμικό μοντέλο το οποίο έχει αναπτυχθεί σε μεγάλο βαθμό τις τελευταίες δεκαετίες. Χρησιμοποιείται σε πολλά πεδία όπως στην βιοστατιστική, στην χρηματοοικονομική ανάλυση, στην κοινωνιολογία και σε άλλα πολλά. Στο λογιστικό μοντέλο τα σφάλματα δεν υπακούν στην κανονική κατανομή. Επίσης η εξαρτημένη μεταβλητή είναι συνήθως δίτιμη με δυνατές τιμές $Y = 1$ ή $Y = 0$ όπου αναφέρονται στην ύπαρξη ή όχι ενός γεγονότος.

Το λογιστικό μοντέλο ορίζεται ως εξής :

$$Y_i = E(Y_i) + \varepsilon_i \ ,$$

όπου Y_i είναι ανεξάρτητη τυχαία μεταβλητή Bernoulli και ισχύει ότι

$$E(Y_i) = \pi_i = \frac{e^{\beta_0 + x_i * \beta_i}}{1 + e^{\beta_0 + x_i * \beta_i}}$$

Στην παρακάτω παράγραφο θα εξηγήσουμε πως βγαίνει η $E(Y_i)$.

Μπορούμε να γράψουμε ότι:

$$P(Y_i = 1) = \pi_i$$

$$P(Y_i = 0) = 1 - \pi_i$$

για την πιθανότητα να συμβεί και να μην συμβεί το γεγονός αντίστοιχα.

Χρησιμοποιώντας τώρα τον ορισμό της αναμενόμενης τιμής έχουμε ότι

$$E(Y_i) = 1 * \pi_i + 0 * (1 - \pi_i) = \pi_i$$

Που σημαίνει ότι η αναμενόμενη τιμή του Y_i είναι η πιθανότητα να συμβεί το γεγονός ($Y_i = 1$).

Σε αυτήν την περίπτωση για να συνδέσουμε την πιθανότητα π_i με την γραμμική έκφραση

$$\beta_0 + \sum_{i=1}^{k} x_i * \beta_i$$

χρησιμοποιούμε μια συνάρτηση g(π) με πεδίο ορισμού το διάστημα [0,1] και πεδίο τιμών το διάστημα (-∞,+∞). Υπάρχουν αρκετές τέτοιες συναρτήσεις σύνδεσης όμως οι παρακάτω τρεις είναι οι πιο βασικές:

Λογιστική Συνάρτηση

$$g(\pi) = log\frac{\pi}{1-\pi}$$

Probit ή αντίστροφη κανονική συνάρτηση

$$g(\pi) = \Phi^{-1}(\pi)$$

Συμπληρωματική log-log συνάρτηση

$$g(\pi) = \log\{-\log(1-\pi)\}$$

Στις περισσότερες των περιπτώσεων όμως χρησιμοποιείτε ο λογιστική συνάρτηση γιατί είναι εύκολο να ερμηνευτεί καθώς ισούται με τον λογάριθμο της σχετική πιθανότητας (odds) να συμβεί το γεγονός. Η σχετική πιθανότητα ενός γεγονότος ισούται με την πιθανότητά να συμβεί προς την πιθανότητα να μην συμβεί το γεγονός.

Έτσι λοιπόν στο λογιστικό μοντέλο έχουμε ότι:

$$log\frac{\pi_i}{1-\pi_i} = \beta_0 + x_i * \beta_i \Leftrightarrow$$

$$\frac{\pi_i}{1-\pi_i} = e^{\beta_0 + x_i*\beta_i} \Leftrightarrow$$

$$\pi_i = \frac{e^{\beta_0 + x_i*\beta_i}}{1+ e^{\beta_0 + x_i*\beta_i}} \Leftrightarrow$$

$$E(Y_i) = \frac{e^{\beta_0 + x_i*\beta_i}}{1 + e^{\beta_0 + x_i*\beta_i}}.$$

Μια ισοδύναμη της τελευταίας σχέσης είναι η σχέση:

$$E(Y_i) = \frac{1}{1 + e^{-\beta_0 - x_i*\beta_i}}$$

1.2 Εκτίμηση παραμέτρων με την μέθοδο της μέγιστης πιθανοφάνειας

Στο κλασσικό μοντέλο παλινδρόμησης μελετάμε την σχέση μιας εξαρτημένης μεταβλητής (Response Y) με μια ομάδα ανεξάρτητων μεταβλητών (explanatory X_i) θεωρώντας κάποιες υποθέσεις, όπως για παράδειγμα ότι η τυχαία μεταβλητή Y ακολουθεί κανονική κατανομή. Στην λογιστική παλινδρόμηση μοντελοποιούμε την δίτιμη τυχαία μεταβλητή Y (Binary) και αρκετές υποθέσεις του κλασσικού μοντέλου όπως ομοσκεδαστικότητα, γραμμικότητα, κανονικότητα παραβιάζονται. Συνεπώς η κλασσική μέθοδος ελαχίστων τετραγώνων για την εκτίμηση των συντελεστών (OLS) είναι ανεπαρκής και καταφεύγουμε σε άλλες μεθόδους. Η πιο διαδεδομένη και ευρέως εφαρμόσιμη είναι η εκτίμηση μέγιστης πιθανοφάνειας που δεν εξαρτάται από τις παραπάνω παραβιάσεις.

Η μέθοδος μέγιστης πιθανοφάνειας οφείλεται στον Fisher και είναι μια τεχνική κατασκευής εκτιμητών για μια παράμετρο θ. Στην λογιστική παλινδρόμηση όπου η Y~B(n,π), 0<π<1 $n \in N$, η παράμετρος θ είναι η $\theta = \log(\frac{\pi}{1-\pi})$. Ο υπολογισμός του εκτιμητή μέγιστης πιθανοφάνειας πραγματοποιείται ακολουθώντας δύο βήματα. Στο πρώτο βήμα φτιάχνουμε την συνάρτηση πιθανοφάνειας. Για να γίνει αυτό θα πρέπει πρώτα να προσδιορίσουμε ένα μοντέλο, δηλαδή αυτό σημαίνει να διαλέξουμε μια κατανομή πιθανότητας (Binomial στην λογιστική παλινδρόμηση) για την εξαρτημένη μεταβλητή και στην συνέχεια θα πρέπει να επιλέξουμε μια συνάρτηση σύνδεσης που να σχετίζει τις παραμέτρους της κατανομής με τις επεξηγηματικές μεταβλητές. Στην περίπτωσή του λογιστικού μοντέλου η ανεξάρτητη μεταβλητή ακολουθεί διωνυμική κατανομή με μια δοκιμή (δοκιμή Bernoulli) και παράμετρο π_i η οποία θεωρείται ότι εξαρτάται από τις επεξηγηματικές μεταβλητές. Τέλος υποθέτουμε ότι οι παρατηρήσεις μας είναι ανεξάρτητες μεταξύ των ατόμων.

Στο δεύτερο βήμα μεγιστοποιούμε την συνάρτηση πιθανοφάνειας ως προς τις άγνωστες παραμέτρους. Αυτό απαιτεί μια επαναληπτική υπολογιστική μέθοδο και συνεπώς αυτό σημαίνει ότι χρειάζονται διαδοχικές προσεγγίσεις τις λύσης. Τέτοιες μέθοδοι προσέγγισης είναι υπολογιστικά απαιτητικές και έχουν αναπτυχθεί μόνο τις τελευταίες δεκαετίες που έχουν αναπτυχθεί και οι ηλεκτρονικοί υπολογιστές σε μεγάλο βαθμό.

Στην συνέχεια θα αναπτύξουμε τα βήματα που αναφέραμε παραπάνω πιο αναλυτικά. Αρχικά θα κάνουμε κάποιες υποθέσεις για να φτιάξουμε την συνάρτηση πιθανοφάνειας. Υποθέτουμε ότι έχουμε δεδομένα για n ανεξάρτητες παρατηρήσεις. Επίσης για κάθε παρατήρηση i έχουμε το Y_i και το X_i, όπου Y_i είναι μια τυχαία μεταβλητή με δυνατές τιμές 1 ή 0 και

$$X_i = [1, X_{i1}, \dots, X_{in}]'$$

είναι ένας πίνακας στήλη που περιέχει τις επεξηγηματικές μεταβλητές μαζί με την μονάδα που αντιπροσωπεύει τον σταθερό όρο. Εμείς εδώ για διευκόλυνση θα τον γράφουμε X_i και όχι σαν πίνακα. Υποθέτουμε επίσης ότι η πιθανότητα π_i είναι η πιθανότητα το Y_i να είναι 1

$$\Pr(Y_i = 1) = \pi_i$$

ακόμα θεωρούμε ότι τα δεδομένα μας προέρχονται από λογιστικό μοντέλο. Έτσι η πιθανότητα π_i ισούται με

$$\pi_i = \frac{1}{1 + \exp(-\beta * X_i)}$$

Θα κατασκευάσουμε τώρα λοιπόν την συνάρτηση πιθανοφάνειας, η οποία θα εκφράζει την πιθανότητα των παρατηρούμενων δεδομένων συναρτήσει των άγνωστων παραμέτρων. Η πιθανότητα αυτή μπορεί να γραφτεί ως εξής:

$$L = \Pr(Y_1, Y_2, \dots, Y_n) =>$$
$$L = \Pr(Y_1) * \Pr(Y_2) * \dots \Pr(Y_n)$$
$$= \prod_{i=1}^{n} \Pr(Y_i)$$

Αφού η Y ακολουθεί διωνυμική κατανομή ξέρουμε ότι $\Pr(Y_i = 1) = \pi_i$ και $\Pr(Y_i = 0) = 1 - \pi_i$. Συνδυάζοντας τις δυο προηγούμενες σχέσεις μπορούμε να έχουμε την πιθανότητα $\Pr(Y_i)$ για κάθε τιμή i, έτσι λοιπόν καταλήγουμε στην ακόλουθη συνάρτηση

$$\Pr(Y_i) = \pi_i^{\,Y_i}(1 - \pi_i)^{\,1-Y_i}$$

Όπου για $Y = 1$ μας δίνει $\Pr(Y_i = 1) = \pi_i$ και για $Y = 0$ μας δίνει $\Pr(Y_i = 0) = 1 - \pi_i$. Τώρα μέσω της παραπάνω σχέσης η συνάρτηση πιθανοφάνειας παίρνει την ακόλουθη μορφή:

$$L = \prod_{i=1}^{n} \Pr(Y_i)$$
$$= \prod_{i=1}^{n} (\pi_i^{\,Y_i} * (1 - \pi_i)^{\,1-Y_i})$$
$$= \prod_{i=1}^{n} (\frac{\pi_i}{1-\pi_i})^{Y_i} * (1 - \pi_i)$$

Στην συνέχεια θα λογαριθμήσουμε και τις δύο πλευρές διότι είναι πιο εύκολο να μεγιστοποιήσουμε τον λογάριθμο log(L) της L αντί για την ίδια την L. Αυτό μας επιτρέπεται καθώς ο λογάριθμος είναι γνησίως αύξουσα συνάρτηση και οι δύο συναρτήσεις (logL,L) μεγιστοποιούνται στο ίδιο ακριβώς σημείο. Έτσι λοιπόν καταλήγουμε στην ακόλουθη συνάρτηση:

$$logL = \sum_{i=1}^{n} Y_i * \log(\frac{\pi_i}{1 - \pi_i}) + \sum_{i=1}^{n} \log(1 - \pi_i)$$

Τέλος αντικαθιστώντας στην παραπάνω συνάρτηση την σχέση

$$\pi_i = \frac{1}{1 + \exp(-\beta * \mathrm{X}_i)} \tag{1.1}$$

καταλήγουμε στην συνάρτηση πιθανοφάνειας :

$$logL = \sum_{i=1}^{n} \beta * X_i * Y_i - \sum_{i=1}^{n} log(1 + exp(\beta * X_i)) \tag{1.2}$$

Έχοντας βρει τώρα την συνάρτηση πιθανοφάνειας θα προχωρήσουμε στο βήμα δύο, δηλαδή θα πρέπει να βρούμε τα $\boldsymbol{\beta}$ που μεγιστοποιούν την (1.2). Υπάρχουν αρκετοί τρόποι μεγιστοποίησης, η πιο γνωστή προσέγγιση είναι να βρούμε την παράγωγο της ως προς β και στην συνέχεια να την θέσουμε ίση με μηδέν. Ακολουθώντας αυτήν την διαδικασία καταλήγουμε στο παρακάτω αποτέλεσμα:

$$\frac{\partial logL}{\partial \beta} = \sum_{i=1}^{n} X_i * Y_i - \sum_{i=1}^{n} X_i * (1 + \exp(-\beta * X_i))^{-1} =>$$

$$\frac{\partial logL}{\partial \beta} = \sum_{i=1}^{n} X_i * Y_i - \sum_{i=1}^{n} X_i * \hat{Y}_i = 0 \tag{1.3}$$

Όπου

$$\widehat{Y_i} = \frac{1}{1 + \exp(-\beta * \mathrm{X_i})}$$

Από την (1.1) μπορεί να θεωρηθεί ότι η $\widehat{Y_i}$ είναι η εκτιμώμενη τιμή του Y για δοσμένο X_i. Επειδή το **β** είναι πίνακας στήλη η παραπάνω συνάρτηση είναι στην πραγματικότητα ένα σύστημα k+1 εξισώσεων, ένα για κάθε β. Εκτός από σπάνιες περιπτώσεις που η λύση είναι εύκολο να βρεθεί χρησιμοποιούνται επαναληπτικές μέθοδοι που έχουν διαδοχικές προσεγγίσεις μέχρι η λύση να συγκλίνει στην πραγματική τιμή. Υπάρχουν και εδώ αρκετές μέθοδοι υπολογισμού που δίνουν την ίδια λύση, διαφέρουν όμως στην ταχύτητα σύγκλισης η οποία επηρεάζεται από τις αρχικές τιμές.

Η πιο διαδεδομένες μέθοδοι είναι η Newton-Raphson και η Fisher Scoring. Η Newton-Raphson μέθοδος χρησιμοποιεί τον ακόλουθο τύπο:

$$\beta_{j+1} = \beta_j - H^{-1}(\beta_j)U(\beta_j)$$

όπου U(β) είναι ένα διάνυσμα με τις πρώτες παραγώγους της logL ως προς β, και πιο συγκεκριμένα

$$\mathrm{U}(\beta) = \frac{\partial logL}{\partial \beta} = \sum_i X_i * y_i - \sum_i X_i * \hat{y}_i$$

Επίσης ο H(β) είναι ένας πίνακας με τις δεύτερες παραγώγους της logL ως προς β και ονομάζεται Εσσιανός πίνακας έχοντας τον ακόλουθο τύπο:

$$\mathrm{H}(\beta) = \frac{\partial^2 logL}{\partial \beta \partial \acute{\beta}} = \sum_i X_i * \acute{X}_i * \hat{y}_i(1 - \hat{y}_i)$$

Η μέθοδος Fisher Scoring που πρότεινε ο Fisher ήρθε για να ξεπεράσει το πρόβλημα της σύγκλισης που είχε κάποιες φορές η μέθοδος Newton-Raphson. Ουσιαστικά αντικατέστησε τον Εσσιανό πίνακα -H(β) με τον πίνακα πληροφορίας I(β). Ο πίνακας αυτός ισούται με την αναμενόμενη αρνητική τιμή των δεύτερων παραγώγων του logL, δηλαδή ισχύει ότι

$$I(\beta) = -E(\nabla^2 logL)$$

Έτσι λοιπόν ο τύπος της μεθόδου Fisher Scoring είναι ο ακόλουθος:

$$\beta_{j+1} = \beta_j + I^{-1}(\beta_j)U(\beta_j)$$

Σχετικά με αυτές τις δυο μεθόδους θα λέγαμε ότι:

- Η μέθοδος Newton-Raphson συγκλίνει πιο γρήγορα στην λύση όμως υπάρχουν και περιπτώσεις που αποτυγχάνει να βρει λύση. Τέτοιες περιπτώσεις συναντάμε όταν η συνάρτηση δεν είναι ομαλή ή όταν περιέχει περισσότερα από ένα τοπικά ακρότατα.
- Η μέθοδος Fisher Scoring αντίθετα είναι πιο αξιόπιστη και συγκλίνει σε περιπτώσεις που η Newton-Raphson δεν μπορεί, έτσι συνεπώς προτείνετε στις περισσότερες των περιπτώσεων.
- Επιπλέον η Fisher Scoring μέθοδος δεν εξαρτάται από τα δεδομένα μας και έτσι υπολογιστικά είναι πιο εύκολη.

1.3 Έλεγχος σημαντικότητας των συντελεστών

Αφού εκτιμήσαμε τους συντελεστές με την μέθοδο μέγιστης πιθανοφάνειας στην συνέχεια ελέγχουμε αν αυτοί οι συντελεστές είναι σημαντικοί. Υπάρχει ο έλεγχος λόγου πιθανοφανειών (Alan Agresti 2007) που χρησιμοποιεί την στατιστική συνάρτηση:

$$\boldsymbol{G^2 = Deviance(without\ variables)\ -\ Deviance(with\ variables)}$$

και συγκρίνει το μοντέλο με τις μεταβλητές και αυτό με το σταθερό όρο μόνο. Βέβαια μπορούμε να συγκρίνουμε και ένα μοντέλο που περιέχει κάποιες μεταβλητές με ένα μοντέλο με περισσότερες για να αποφανθούμε εάν οι επιπρόσθετες είναι σημαντικές και πρέπει να συμπεριληφθούν στο μοντέλο. Η απόκλιση ***(Deviance)*** που πρώτοι ανέφεραν οι Nelder και Wedderburn (1972) είναι η ακόλουθη ποσότητα

$$\boldsymbol{D = -2log\left(\frac{likelihood\ fitted\ model}{likelihood\ saturated\ model}\right)}$$

Όπου κορεσμένο (saturated) μοντέλο είναι αυτό που περιέχει τόσες παραμέτρους όσες είναι τα δεδομένα μας. Στην λογιστική παλινδρόμηση όπου η ανεξάρτητη μεταβλητή είναι δίτιμη (1 ή 0) η πιθανοφάνεια του κορεσμένου μοντέλου ισούται με 1. Πράγματι αυτό ισχύει γιατί στην επίλυση του συστήματος (1.3) ο αριθμός των εξισώσεων είναι όσες και ο αριθμός των αγνώστων και έτσι έχουμε τέλεια προσαρμογή.

Συνεπώς η απόκλιση στο λογιστικό μοντέλο είναι:

$$\boldsymbol{D = -2log(likelihood\ fitted\ model)}$$

Συνεπώς η στατιστική συνάρτηση G^2 στην λογιστική παλινδρόμηση παίρνει την ακόλουθη μορφή:

$$G^2 = 2(logL_{with\ variables} - logL_{without\ variables})$$

Τέλος για να αποφανθούμε για την σημαντικότητα των μεταβλητών συγκρίνουμε την στατιστική συνάρτηση G^2 με το κατάλληλο ποσοστιαίο σημείο της

$$x^2_{DF=[df(without\ variables)-df(with\ variables)]}$$

και αν

$$G^2 > x^2_{DF}$$

τότε λέμε ότι τουλάχιστον μια μεταβλητή είναι στατιστικά σημαντική. Κάνοντας έλεγχο ενός μοντέλου με το μοντέλο με τον σταθερό όρο η στατιστική συνάρτηση G^2 παίζει τον ίδιο ρόλο που παίζει και η F στην γραμμική παλινδρόμηση.

Υπάρχει επίσης και ο έλεγχος του Wald που χρησιμοποιεί την στατιστική συνάρτηση

$$\mathrm{W} = \frac{\widehat{\beta_\iota}}{SE(\widehat{\beta_\iota})}$$

Από τις ασυμπτωτικές ιδιότητες των ΕΜΠ έχουμε ότι

- $\widehat{\beta_\iota} \sim N\left(\beta_i\ , \sigma^2(\hat{\beta}_i)\right)$
- Ο $\widehat{\beta_\iota}$ συνεπής εκτιμητής του β_i

Από το Θεώρημα του Slutsky (1925) έχουμε ότι η στατιστική συνάρτηση W είναι ασυμπτωτικά κανονική. Επομένως το αντίστοιχο p-value υπολογίζεται ως εξής P(|Z|>W). Ο έλεγχος του Wald μας απαντάει στο ερώτημα εάν η παράμετρος β_i είναι στατιστικά σημαντική. Βέβαια όταν η μεταβλητή είναι κατηγορική δεν μας δίνει αξιόπιστα αποτελέσματα γ' αυτό θα πρέπει να εκτελούμε έλεγχο λόγου πιθανοφανειών ανάμεσα στο μοντέλο με την συγκεκριμένη μεταβλητή και σε αυτό χωρίς αυτήν.

1.4 Μέτρα προσαρμογής ενός μοντέλου

Στην συνήθη γραμμική παλινδρόμηση ως δείκτη ποιότητας του μοντέλου έχουμε το ευρέως γνωστό R^2 και το προσαρμοσμένο R^2, όπου παίρνει τιμές στο διάστημα 0 έως 1

$$R^2_{adj} = 1 - \frac{SSE/(n-m)}{SST/(n-1)}$$

Τιμές του κοντά στην μονάδα μας φανερώνουν καλή προσαρμογή του μοντέλου στα δεδομένα ενώ αντίθετα μικρές τιμές του μας δείχνουν κακή προσαρμογή. Το R^2 ερμηνεύεται ως το ποσοστό της συνολικής διακύμανσης που εξηγείτε από το μοντέλο.

Για το λογιστικό μοντέλο έχουν προταθεί αρκετές παραλλαγές του R^2, εμείς θα αναφέρουμε τις πιο βασικές. Αυτές είναι του McFadden, McFadden Adjusted (1974), του Cox and Snell (1989), του (1991) και του Efron (1978) που είναι το κλασσικό R^2. Αυτές οι παραλλαγές φαίνονται αναλυτικά στον παρακάτω πίνακα:

• McFadden	$R^2 = 1 - \frac{logL_{\mathrm{M}}}{logL_{null}}$
• McFadden Adjusted	$R^2 = 1 - \frac{logL_{\mathrm{M}} - m}{logL_{null}}$, m το πλήθος των παραμέτρων
• Cox and Snell	$R^2 = 1 - (\frac{logL_{null}}{logL_{\mathrm{M}}})^{2/n}$, n πλήθος των παρατηρήσεων
• Nagelkerke	$R^2 = 1 - \frac{(\frac{\mathrm{Lnull}}{\mathrm{L_M}})^{2/n}}{1 - Lnull^{2/n}}$
• Efron	$R^2 = 1 - \frac{\sum (y_i - \hat{\pi}_i)^2}{\sum (y_i - \bar{y})^2}$, το κλασσικό R^2

Πίνακας 1.1: R^2 για Λογιστική Παλινδρόμηση

Όπου $\hat{\pi}_i$ είναι οι εκτιμημένη τιμή της πιθανότητας πi μέσω του μοντέλου. Επίσης $logL_{\mathrm{M}}$ είναι ο λογάριθμός της μέγιστης πιθανοφάνειας του μοντέλου M και $logLnull$ είναι ο λογάριθμός της μέγιστης πιθανοφάνειας του μοντέλου που δεν χρησιμοποιεί επεξηγηματικές μεταβλητές. Συνήθως προτείνεται ως καλύτερο το Cox and Snell και τιμές του μεταξύ 0.2 και 0.4 δείχνουν καλή προσαρμογή. Αυτή η εξίσωση μπορεί να χρησιμοποιηθεί για κάθε μοντέλο παλινδρόμησης που εκτιμάται με την μέθοδο της μέγιστης πιθανοφάνειας.

1.5 Προβλεπτική Ισχύς ενός μοντέλου: Καμπύλες ROC

Σ' αυτή την παράγραφο θα αναφέρουμε της καμπύλες ROC *(David W. Hosmer, Jr., Stanley Lemeshow 2001)* που μας περιγράφουν το πόσο καλά ταξινομεί το μοντέλο μας την εξαρτημένη μεταβλητή Y (0 ή 1) για δοθέντες τιμές ανεξαρτήτων μεταβλητών. Πριν αναλύσουμε τις καμπύλες αυτές είναι χρήσιμο να αναφέρουμε δύο μέτρα όπου βοηθούν στην κατανόηση και στην δημιουργία τους. Τα δύο μέτρα αυτά είναι η ευαισθησία (sensitivity) και η ειδικότητα (specificity). Αυτές οι έννοιες προέρχονται από την Βιοστατιστική και ορίζονται ως εξής:

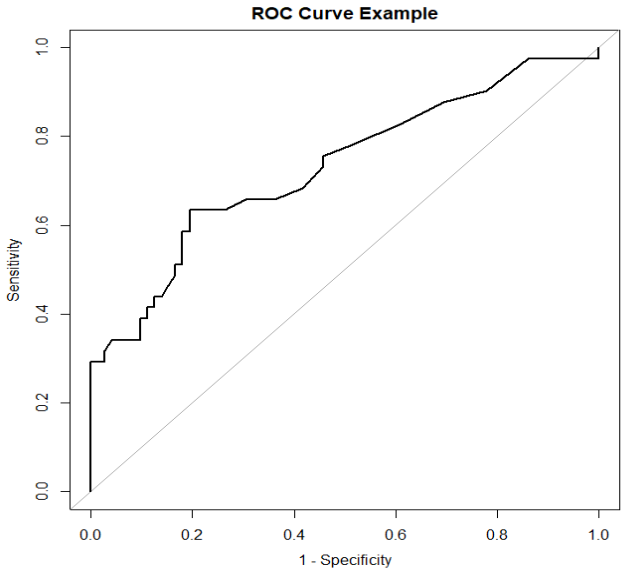
$$\text{Ευαισθησία } = P(\hat{Y} = 1 | Y = 1)$$
$$\text{Ειδικότητα } = P(\hat{Y} = 0 | Y = 0)$$

Οι ποσότητες αυτές όμως εξαρτώνται από την τιμή ενός κατωφλιού. Η δημοφιλέστερη μέθοδος για την επιλογή του κατωφλιού είναι η μέθοδος Youden (1950) που χρησιμοποιεί τον παρακάτω δείκτη:

J = ευαισθησία – (1-ειδικότητα)

Αφού λοιπόν έχουμε υπολογίσει τις τιμές αυτές (ευαισθησία, ειδικότητα) για τα διάφορα κατώφλια επιλέγουμε αυτό που μεγιστοποιεί την ποσότητα J.

Οι καμπύλες ROC δημιουργούνται χρησιμοποιώντας τις τιμές sensitivity, 1- specificity. Μια τέτοια καμπύλη έχει την παρακάτω μορφή:



Γράφημα 1.1: Παράδειγμα καμπύλης ROC

Η περιοχή κάτω από την καμπύλη που ονομάζεται AUC *(Area Under the Curve)* παίρνει τιμές από 0 έως 1 και μας δείχνει την ικανότητα του μοντέλου να διαχωρίζει τα δεδομένα και να δίνει τιμή $Y = 1$ ή $Y = 0$ για την εξαρτημένη μεταβλητή. Ένας γενικός κανόνας για να αξιολογούμε την ταξινόμηση είναι ο ακόλουθος:

- Αν $\text{AUC} = 0.5$, έχουμε κακή ταξινόμηση (ταξινομεί με πιθανότητα 50%, δηλαδή σαν μια ρίψη ενός νομίσματος)
- Αν $0.5 < \text{AUC} < 0.7$, έχουμε αρκετά φτωχή ταξινόμηση
- Αν $0.7 \leq \text{AUC} < 0.8$, έχουμε μια αποδεχτή ταξινόμηση
- Αν $0.8 \leq \text{AUC} < 0.9$, έχουμε πολύ καλή ταξινόμηση
- Αν $\text{AUC} \geq 0.9$, έχουμε τέλεια ταξινόμηση

Η διαγώνιος είναι η καμπύλη για το μοντέλο που ταξινομεί με πιθανότητα 50%, ανεξάρτητα από τις τιμές των ανεξαρτήτων μεταβλητών. Επίσης στην πράξη είναι εξαιρετικά ασυνήθιστο να πετύχουμε $\text{AUC} \geq 0.9$. Εδώ θα πρέπει να αναφέρουμε ότι μπορεί ένα μοντέλο να μην έχει καλή προσαρμογή αλλά να έχει καλή ταξινόμηση γ' αυτό προτείνεται να ελέγχουμε και τα δύο χαρακτηρίστηκα.

Προφανώς όσο πιο ψηλά είναι η καμπύλη τόσο καλύτερη προβλεπτική ικανότητα έχει το μοντέλο καθώς η περιοχή κάτω από την καμπύλη *(Area Under the Curve)* είναι μεγαλύτερη.

Κ Ε Φ Α Λ Α Ι Ο 2

ΜΕΘΟΔΟΙ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΓΙΑ ΣΠΑΝΙΑ ΓΕΓΟΝΟΤΑ

Στο παρακάτω κεφάλαιο θα αναφέρουμε τα προβλήματα που συναντάμε στα σπάνια γεγονότα και τρόπους δειγματοληψίας που έχουν προταθεί.

2.1 Βασικά προβλήματα στα σπάνια γεγονότα

Αρχικά θα πρέπει να ορίσουμε πότε ένα γεγονός θεωρείται σπάνιο.

Ορισμός: Ένα γεγονός θεωρείται σπάνιο όταν το ποσοστό καταγραφής του στο δείγμα είναι μικρότερο του 5%.

Στην περίπτωση της λογιστικής παλινδρόμησης που μελετάμε, καταγράφουμε πόσες περιπτώσεις έχουμε $Y = 1$ και $Y = 0$ στο σύνολο των δεδομένων μας. Αν κάποιο από αυτά τα δύο ποσοστά είναι μικρότερο του 5% τότε θεωρείται πως το αντίστοιχο γεγονός μας είναι σπάνιο και θα πρέπει να εργαστούμε με διαφορετικό τρόπο απ' ότι συνήθως. Τα δύο κύρια προβλήματα που συναντάμε στα σπάνια γεγονότα είναι το πρόβλημα της υποεκτίμησης των γεγονότων με την κλασσική μέθοδο της λογιστικής παλινδρόμησης όπου το μοντέλο υποεκτιμά την πιθανότητα να συμβεί το ενδεχόμενο $Y = 1$ και κατά συνέπεια υπερεκτιμά την πιθανότητα να συμβεί το ενδεχόμενο $Y = 0$. Το άλλο σημαντικό πρόβλημα που συναντάμε είναι η δυσκολία της συλλογής δεδομένων τέτοιου είδους τα οποία να είναι κατάλληλα προς στατιστική ανάλυση. Οι βασικές στρατηγικές δειγματοληψίας είναι δυο *(Paul S. Levy, Stanley Lemeshow 2013)*:

- Η τυχαία δειγματοληψία. Στην μέθοδο αυτή θα πρέπει όλες οι μονάδες του πληθυσμού να έχουν την ίδια πιθανότητα επιλογής στο δείγμα. Η μέθοδος αυτή είναι μάλλον ιδεατή γιατί απαιτεί την καταγραφή του πληθυσμού και επιλογή δείγματος επιλέγοντας n μονάδες από την λίστα του πληθυσμού N με τυχαίο τρόπο.
- Η εξωγενής στρωματοποιημένη δειγματοληψία. Σύμφωνα με αυτή προσδιορίζουμε τις μεταβλητές οι οποίες είναι σημαντικές και σχετίζονται με την μεταβλητή Y. Στρωματοποιούμε τον πληθυσμό ανάλογα με τις μεταβλητές αυτές και στη συνέχεια από κάθε στρώμα παίρνουμε ένα απλό τυχαίο δείγμα. Το αποτέλεσμα είναι εκτιμητές με μικρότερη διασπορά σε σχέση με την απλή τυχαία δειγματοληψία.

Στην επιδημιολογία και οι δύο αυτές στρατηγικές είναι γνωστές με το όνομα διαστρωματοποιημένες (cohort) μελέτες και δεν είναι κατάλληλες για μελέτη σπανίων δεδομένων. Στις επόμενες παραγράφους θα αναφέρουμε κάποιους τρόπους δειγματοληψίας που έχουν προταθεί σε δεδομένα με σπάνια γεγονότα.

2.2 Δειγματοληψία case-cohort

Όταν τα δεδομένα είναι σπάνια όπως αναφέραμε και στην προηγούμενη παράγραφο θα πρέπει να καταφύγουμε σε άλλους τρόπους δειγματοληψίας. Ένας τρόπος για να εξοικονομήσουμε κόστος είναι να επιλέξουμε τα δεδομένα μας τυχαία μέσα στις κατηγορίες της εξαρτημένης μεταβλητής Y. Τέτοιου τύπου σχεδιασμοί στην επιδημιολογία ονομάζονται case-control. Η βασική ιδέα αυτού του σχεδιασμού (case-cohort) είναι να επιλέγουμε τα δεδομένα μας από συλλογές δεδομένων και πιο συγκεκριμένα για $Y = 1$ (cases) επιλέγουμε όλα τα διαθέσιμα δεδομένα ή κάνουμε τυχαία επιλογή και για $Y = 0$ (controls) επιλέγουμε με τυχαία δειγματοληψία. Τέτοιου είδους σχεδιασμοί εμπεριέχουν εκ τον προτέρων πληροφορία σχετικά με τα πληθυσμιακά ποσοστά των γεγονότων ($Y = 1$). Τέτοιες μελέτες έχουν ως αφετηρία την συλλογή ορισμένων μεταβλητών από μια μεγάλη cohort μελέτη και στην συνέχεια επιλέγουμε όλα τα δεδομένα με τιμή $Y = 1$ και κάνουμε τυχαία δειγματοληψία για τα δεδομένα με τιμή $Y = 0$. Θα πρέπει να αναφέρουμε επίσης ότι τέτοιες μελέτες είναι κατάλληλες εάν θέλουμε να προσθέσουμε μεταβλητές που έχουν υψηλό κόστος.

Κάνοντας συλλογή των δεδομένων μας με αυτόν τον τρόπο θα πρέπει να προσέξουμε κάποια σημεία που μπορούν να φέρουν σύγχυση. Για τις μεθόδους prior-correction και weighting τις οποίες θα αναφέρουμε στα επόμενα κεφάλαια απαιτείται δειγματοληψία τέτοια ώστε οι παρατηρήσεις $Y = 1$ και $Y = 0$ να είναι ανεξάρτητες και τυχαίες. Αυτές τις συνθήκες τις ικανοποιούν οι σχεδιασμοί που αναφέραμε στην προηγούμενη παράγραφο, ενώ σε αντίθετη περίπτωση απαιτούνται διαφορετικές μέθοδοι ανάλυσης. Ένα δεύτερο σημείο που θα πρέπει να προσέξουμε είναι όταν επιλέγουμε την μεταβλητή Y θα πρέπει να προσέχουμε να μην την επιλέγουμε σε διαφορικές ανεξάρτητες μεταβλητές X. Για να γίνει πιο κατανοητό θα αναφέρουμε το ακόλουθο παράδειγμα. Έστω ότι επιλέγουμε όλα τα άτομα από ένα συγκεκριμένο νοσοκομείο που πάσχουν από μια σπάνια ασθένεια ($Y = 1$) και επίσης επιλέγουμε ένα τυχαίο δείγμα υγιών ατόμων από το σύνολο της χώρας ($Y = 0$). Το σημείο που θα πρέπει να προσεχθεί είναι ότι εδώ δημιουργείται μεροληψία καθώς τα άτομα που έχουν την ασθένεια έχουν κάνει τα σωστά διαγνωστικά τεστ και έχουν λάβει την σωστή φαρμακευτική

αγωγή σε αντίθεση με τα υγιή άτομα. Ένας τρόπος για να αποφύγουμε την μεροληψία είναι να επιλέξουμε τα υγιή άτομα που έχουν κάνει το ίδιο διαγνωστικό τεστ και έχει βγει αρνητικό. Ένα άλλο σημείο στο οποίο θα πρέπει να δώσουμε προσοχή είναι στην βέλτιστη αναλογία ανάμεσα στο να συλλέξουμε περισσότερα δεδομένα και στο να συλλέξουμε περισσότερες ή "καλύτερες" επεξηγηματικές μεταβλητές διότι διαφέρει από μελέτη σε μελέτη και συνεπώς εμπεριέχει την κρίση του εκάστοτε αναλυτή. Εμείς στην παρούσα παράγραφο θα αναφέρουμε κάποια στατιστικά αποτελέσματα και διαδικασίες που έχουν προτείνει οι King Gary και Langche Zeng (2001) στην εργασία τους για την βέλτιστη αναλογία. Αρχικά όταν τα γεγονότα $Y = 1$ και τα μη-γεγονότα $Y = 0$ είναι εύκολο να συλλεχθούν και υπάρχουν πολλές παρατηρήσεις για το καθένα τότε ένα δείγμα με ίσο αριθμό παρατηρήσεων είναι ιδανικό. Αυτό είναι ένα αρκετά καλό δείγμα αλλά σε σχεδόν όλες τις περιπτώσεις που έχουμε σπάνια γεγονότα ο αριθμός των παρατηρήσεων για $Y = 1$ είναι περιορισμένος και κατά συνέπεια ως λύση συλλέγουμε όλα τα διαθέσιμα δεδομένα ή ένα μεγάλο ποσοστό αυτών ($Y = 1$). Στην συνέχεια η κρίσιμη απόφαση που θα πρέπει να πάρουμε είναι ο αριθμός των μη-γεγονότων που θα συλλέξουμε. Εάν η συλλογή τους δεν απαιτεί μεγάλο κόστος τότε θα πρέπει να πάρουμε όσο περισσότερα δεδομένα μπορούμε καθώς όπως γνωρίζουμε όσο περισσότερες παρατηρήσεις έχουμε τόσο πιο ακριβής είναι η ανάλυση μας. Όμως ένας καλός κανόνας που θα πρέπει να ακολουθούμε είναι να μην συλλέγουμε περισσότερες από δύο έως πέντε φορές περισσότερα $Y = 0$ απ' ότι $Y = 1$ διότι θα έχουμε το πρόβλημα τις υποεκτίμησης των γεγονότων.

Τέλος μια χρήσιμη στρατηγική συλλογής των δεδομένων είναι η ακολουθιακή. Πιο συγκεκριμένα αρχικά συλλέγουμε ίδιο αριθμό γεγονότων ($Y = 1$) και μη-γεγονότων ($Y = 0$) και αν τα τυπικά σφάλματα είναι μικρά και τα διαστήματα εμπιστοσύνης είναι μικρού μήκους τότε σταματάμε. Εάν δεν συμβεί αυτό τότε συνεχίζουμε να συλλέγουμε παρατηρήσεις $Y = 0$ με τυχαίο τρόπο και σταματάμε όταν τα διαστήματα εμπιστοσύνης γίνουν μικρά ανάλογα πάντα με τους σκοπούς της εκάστοτε μελέτης.

2.3 Δικτυωτή δειγματοληψία (network sampling)

Αυτή την διαδικασία δειγματοληψίας την προτείναν στον βιβλίο τους οι Paul S. Levy, Stanley Lemeshow (Sampling of Populations: Methods and Applications). Αυτός ο τρόπος δειγματοληψίας λοιπόν προτείνει να χρησιμοποιούμαι έναν τροποποιημένο τρόπο καταμέτρησης ο οποίος οδηγεί σε εκτιμήσεις με μικρότερο σφάλμα δειγματοληψίας σε σχέση με τον συνηθισμένο τρόπο καταμέτρησης.

Ένας τρόπος καταμέτρησης είναι η διαδικασία με την οποία κατανέμονται οι μονάδες. Για παράδειγμα εάν θέλουμε να μετρήσουμε τις γεννήσεις σε ένα συγκεκριμένο πληθυσμό σε μια συγκεκριμένη χρονική περίοδο ένας συνηθισμένος τρόπος καταμέτρησης θα ήταν να προσμετράτε κάθε γέννηση στο σπίτι των γονέων. Ένας ενναλακτικός τρόπος θα ήταν η κάθε γέννηση να προσμετράτε και στο σπίτι των γονέων αλλά και των παππούδων. Αυτός ο τρόπος καταμέτρησης επιτρέπει την κάθε μονάδα να προσμετράτε σε περισσότερες από μια λίστες. Ένας τρόπος καταμέτρησης που αντιστοιχεί μια μονάδα σε μια μόνο λίστα λέγεται συμβατικός τρόπος *(conventional counting rule)* ενώ αυτός που επιτρέπει την αντιστοίχιση σε περισσότερες από μια λίστες λέγεται πολλαπλός τρόπος καταμέτρησης *(multiplicity counting rule)*. Οι σχεδιασμοί που χρησιμοποιούν πολλαπλούς τρόπους καταμέτρησης ονομάζονται δικτυωτοί τρόποι δειγματοληψίας και έχουν αναπτυχθεί κυρίως τις τελευταίες δύο δεκαετίες στις επιστήμες υγείας σε περιπτώσεις που υπάρχουν σπάνια γεγονότα.

Για να γίνει περισσότερο κατανοητό θα αναφέρουμε το ακόλουθο παράδειγμα. Έστω ότι μια χώρα έχει 100 νοσοκομεία και θέλουμε να διεξάγουμε μια έρευνα σε 10 νοσοκομεία με σκοπό να εκτιμήσουμε τον συνολικό αριθμό των ατόμων που λαμβάνουν θεραπεία για μια ασθένεια Α σε μια συγκεκριμένη χρονική περίοδο, συνεπώς το σπάνιο γεγονός εδώ είναι ο αριθμός των ατόμων που έλαβαν θεραπεία. Εάν οι ασθενείς λάμβαναν μόνο από ένα νοσοκομείο θεραπεία θα ήταν μια απλή διαδικασία, εδώ όμως ένας ασθενής μπορεί να λάβει θεραπεία σε περισσότερα από ένα. Έστω ότι έχουμε 3 ασθενείς που έλαβαν θεραπεία. Με συγκεκριμένη χρονική σειρά ο πρώτος πήγε στα νοσοκομεία (4,1,2,3), ο δεύτερος στα (4,5) και ο τρίτος μόνο στο 6. Εάν χρησιμοποιούσαμε την τυπική καταμέτρηση οι ασθενείς 1 και 2 θα καταχωρούνταν μόνο από το πρώτο νοσοκομείο που έλαβαν θεραπεία, δηλαδή από το 4. Εάν θεωρήσουμε X_i τον αριθμό των ασθενών που έλαβαν θεραπεία από το i νοσοκομείο θα έχουμε τις ακόλουθες τιμές $X_4=2$, $X_6=1$ και όλα τα υπόλοιπα $X_i=0$.

Για ένα απλό δείγμα με n=10 νοσοκομεία με συμβατικό τρόπο καταμέτρησης έχουμε ότι η εκτίμηση $\tilde{X} = \left(\frac{100}{10}\right) * X$. Οι δυνατές περιπτώσεις είναι:

- Να μην έχουμε σε κανένα από το 10 νοσοκομεία που επιλέξαμε κάποια εγγραφή ασθενή (δηλαδή να επιλέξουμε κάποια τα 98 που έμειναν)
- Στα 10 νοσοκομεία που επιλέξαμε να είναι μόνο το ένα με την μια εγγραφή (X_6=1)
- Στα 10 νοσοκομεία που επιλέξαμε να είναι μόνο το ένα με τις δύο εγγραφές (X_4=2)
- Στα 10 νοσοκομεία που επιλέξαμε να βρίσκονται και τα δύο νοσοκομεία που έχουν καταγράψει ασθενή (X_6=1, X_4=2)

Απαριθμώντας όλα τα δείγματα καταλήγουμε τον παρακάτω πίνακα:

$\tilde{X}$	**Relative Freq**
0	0.8091
10	0.0909
20	0.0909
30	0.0091

Όπου *Relative Freq* είναι η σχετική πιθανότητα τα 10 νοσοκομεία που επιλέξαμε να βρίσκονται στην κάθε περίπτωση.

Στην συνέχεια θα υπολογίσουμε την μέση τιμή $E(\tilde{X})$ και το τυπικό σφάλμα $SE(\tilde{X})$ με την βοήθεια των τύπων που πρότειναν οι Paul S. Levy, Stanley Lemeshow στο βιβλίο τους Sampling of Populations (2008). Οι τύποι λοιπόν είναι οι ακόλουθοι

$$E(\tilde{X}) = \sum_{i=1}^{k} (\tilde{X} * \pi_i)$$

$$Var(\tilde{X}) = \sum_{i=1}^{k} [\tilde{X} - E(\tilde{X})]^2 * \pi_i$$

όπου k είναι ο αριθμός των δυνατών περιπτώσεων.

Έτσι λοιπόν καταλήγουμε στις τιμές $E(\tilde{X})$=3, $SE(\tilde{X})$=6.681. Εδώ παρατηρούμε λοιπόν πως στο 80,91% των δειγμάτων δεν έχουν καταγραφή ασθενή.

Από την άλλη εάν χρησιμοποιήσουμε πολλαπλούς τρόπους καταμέτρησης θα πρέπει να χρησιμοποιήσουμε τον ακόλουθο τύπο:

$$X^* = \sum_{j=1}^{m} \frac{\delta_{ij}}{S_j} \quad \text{, όπου} \begin{cases} m = o\ συνολικός\ αριθμός\ ασθενών\ στο\ δείγμα \\ \delta_{ij} = \begin{cases} 1\text{ , } αν\ o\ ασθενής\ j\ πήγε\ στο\ i\ νοσοκομείο \\ 0, αλλίως \end{cases} \\ S_j = o\ αριθμός\ των\ ασθενών\ που\ πήγαν\ στο\ j\ νοσοκομείο \end{cases}$$

Έτσι η εκτίμηση εδώ είναι $\tilde{X}_{\text{mult}} = \left(\frac{N}{n}\right) * X^*$. Υπολογίζοντας τώρα τα X^* για το προηγούμενο παράδειγμα έχουμε:

$$X_1^* = \frac{1}{4}, X_2^* = \frac{1}{4}, X_3^* = \frac{1}{4}, X_4^* = \frac{3}{4}, X_5^* = \frac{1}{2}, X_6^* = 1$$

Στην συνέχεια υπολογίζοντας το $\tilde{X}_{\text{mult}}$ και την *Relative Freq* με την ίδια λογική με προηγουμένως καταλήγουμε τον παρακάτω πίνακα:

$\tilde{X}_{\text{mult}}$	Relative Freq
0	0.5223
2.5	0.1843
5	0.0807
7.5	0.08132
10	0.08251
12.5	0.03076
15	0.01003
17.5	0.00839
20	0.00192
22.5	0.00074
25	0.00019
27.5	0.00001
30	0.000003

Χρησιμοποιώντας και εδώ τους τύπους για την μέση τιμή και την διασπορά παίρνουμε τις ακόλουθες τιμές $E(\tilde{X}_{\text{mult}})$= 3, $SE(\tilde{X}_{\text{mult}}) = 4.13$.

Ως συμπέρασμα μπορούμε να πούμε ότι η παραπάνω μέθοδος είναι πιο αξιόπιστη καθώς αυξάνει την απόδοση του δείγματος μικραίνοντας το τυπικό σφάλμα. Αξίζει να αναφέρουμε εδώ πώς χρησιμοποιώντας τον πολλαπλό τρόπο καταμέτρησης το 52,22% όλων των δειγμάτων n=10 δεν θα έχουν ασθενείς σε αντίθεση με το συμβατικό τρόπο καταμέτρησης που μας δίνει 80,91%.

Κ Ε Φ Α Λ Α Ι Ο 3

ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ ΣΤΗΝ ΠΕΡΙΠΤΩΣΗ ΣΠΑΝΙΩΝ ΓΕΓΟΝΟΤΩΝ

Σχεδιασμοί στους οποίους επιλέγουμε το Y, μπορεί να είναι συνεπής μόνο κάτω από κάποιες διορθώσεις στην εκτίμηση του Y. Στο παρακάτω κεφάλαιο λοιπόν θα αναφέρουμε τις δύο διορθώσεις των εκτιμήσεων (prior correction, weighting). Αυτές τις διορθώσεις τις πρότειναν οι King Gary και Langche Zeng (2001).

3.1 Prior correction

Αρχικά θα πρέπει να αναφέρουμε ότι η μέθοδος Prior Correction απαιτεί τον υπολογισμό του γνωστού εκτιμητή μέγιστης πιθανοφάνειας και στην συνέχεια γίνεται διόρθωση αυτού με την εκ το προτέρων γνώση δυο τιμών. Αυτές οι τιμές είναι το ποσοστό των γεγονότων ($Y = 1$) στον συνολικό πληθυσμό (τ) και το ποσοστό των γεγονότων ($Y = 1$) στο δείγμα μας ($\bar{Y}$), είτε από case-cohort δείγμα είτε από άλλες δειγματοληψίες. Στην συνέχεια θα παρουσιάσουμε την μέθοδο σε τέσσερα βήματα ξεκινώντας από την γενική περίπτωση και καταλήγοντας στην λογιστική παλινδρόμηση που είναι ο στόχος μας. Αναπτύσσοντας σταδιακά την μέθοδο σε τέσσερα βήματα θα γίνει έτσι πιο κατανοητή στον αναγνώστη. Στην γενική περίπτωση η εφαρμογή της μεθόδου είναι αρκετά δύσκολη και πολλές φορές ανέφικτη σε αντίθεση με το λογιστικό μοντέλο που η εφαρμογή της είναι εύκολη, συνεπής και αποτελεσματική.

3.1.1 Γενική Περίπτωση

Ξεκινώντας θα χρειαστεί να κάνουμε κάποιες υποθέσεις. Ας υποθέσουμε λοιπόν ότι X,Y είναι τυχαίες μεταβλητές με συναρτήσεις πυκνότητας P(X,Y) που αντιστοιχούν στο πλήρες δείγμα μας από μια case-cohort δειγματοληψία. Επίσης υποθέτουμε ότι οι x,y είναι τυχαίες μεταβλητές με αντίστοιχες συναρτήσεις πυκνότητας P(x,y) που αντιστοιχούν σε ένα μικρότερο κομμάτι του δείγματος το οποίο περιέχει όλα τα γεγονότα ($Y = 1$) και ένα μέρος των μη γεγονότων ($Y = 0$) το οποίο πάρθηκε με τυχαία επιλογή από τα X,Y. Με τον τρόπο που ορίσαμε το μικρότερο δείγμα η συνάρτηση πυκνότητας P(x,y) έχει ορισθεί έτσι ώστε να ισχύει η ακόλουθη ισότητα:

$$P(x|y) = P(X|Y) \tag{3.1}$$

παρόλο που οι αντίστοιχες περιθώριες πιθανότητες P(x), P(y), P(y|x) δεν είναι υποχρεωτικά ίσες με τις αντίστοιχες P(X), P(Y), P(Y|X). Σκοπός μας τώρα λοιπόν είναι να μπορέσουμε αποφανθούμε σχετικά με την τιμή της πιθανότητας P(Y|X). Από τον ορισμό της δεσμευμένης πιθανότητας έχουμε την παρακάτω ισότητα:

$$P(Y|X) = \frac{P(X\,Y)}{P(X)} =>$$

$$P(Y|X) = P(X|Y)\frac{P(Y)}{P(X)} \overset{(3.1)}{\Longrightarrow}$$

$$P(Y|X) = P(x|y) * \frac{P(Y)}{P(X)} =>$$

$$P(Y|X) = P(y|x)\left[\frac{P(Y)*P(x)}{P(y)*P(X)}\right] \tag{3.2}$$

Ο ισχυρισμός που κάνουμε σε αυτό το σημείο είναι ότι μπορούμε να εκτιμήσουμε την P(Y|X) με ένα ανεξάρτητο και τυχαίο δείγμα που προέρχεται από P(X|Y) ή P(Y|X) είτε από P(x|y) ή P(y|x) πολλαπλασιάζοντας την εκτίμηση με την τελευταία τιμή της σχέσης (3.2). Ο κύριος στόχος μας λοιπόν σε αυτό το σημείο είναι να εκτιμήσουμε την P(Y|X). Για να προχωρήσουμε στην απόδειξη θα υποθέσουμε ότι D, d είναι τυχαία δείγματα από τις P(X|Y) και P(x|y) μεγέθους n. Όταν λοιπόν το n τείνει στο άπειρο έχουμε :

$$P(Y|X, D) \;=\; P(X|Y, D) * \frac{P(Y, D)}{P(X, D)}$$

όπου αυτή η σχέση συγκλίνει στην

$$P(X|Y) * \frac{P(Y)}{P(X)} \;= P(Y|X)$$

Αντίθετα όμως η σχέση

$$P(y|x, d) \;=\; P(x|y, d) * \frac{P(y,d)}{P(x,d)} \tag{3.3}$$

συγκλίνει στην

$$P(x|y) * \frac{P(y)}{P(x)} \;=\; P(y|x)$$

αλλά δεν συγκλίνει στην P(Y|X) που μας ενδιαφέρει, καθώς από υπόθεση δεν είναι υποχρεωτικά ίσες οι πιθανότητες P(y|x) = P(Y|X).

Στην συνέχεια θα θεωρήσουμε της παρακάτω ποσότητες:

- $A_y = \frac{P(Y|d)}{P(y|d)}$

  που είναι μια συνάρτηση του y

- $\mathrm{B} = \frac{P(x|d)}{P(X|d)} = [\sum_{all\ y} P(y|x)A_y]^{-1}$

που είναι ένας σταθερός παράγοντας κανονικοποίησης.

Πολλαπλασιάζοντας τώρα την σχέση (3.3) με A_yB και από τις δύο πλευρές έχω

$$\mathrm{A_yB} * \mathrm{P(y|x,d)} = \mathrm{P(x|y,d)} * \frac{P(y,d)}{P(x,d)} * \mathrm{A_yB}$$

$$= \mathrm{P(x|y,d)} * \frac{P(y,d)}{P(x,d)} * \frac{P(Y|d)}{P(y|d)} * \frac{P(x|d)}{P(X|d)}$$

$$= \mathrm{P(x|y,d)} * \frac{P(Y|d)}{P(X|d)}$$

όπου αυτό το γινόμενο συγκλίνει στην P(X|Y)* $\frac{P(Y)}{P(X)}$ = P(Y|X) αφού η P(x|y,d) συγκλίνει στην P(x|y)= P(X|Y), η P(Y|d) συγκλίνει στην P(Y) και η P(X|d) συγκλίνει στην P(X). Συνεπώς η διορθωμένη συνάρτηση της υποδειγματοληψίας είναι συνεπής για την κατανομή που μας ενδιαφέρει αφού η A_yB*P(y|x,d) συγκλίνει στην P(Y|X). Αυτό συμβαίνει για κάθε δειγματοληψία στην οποία επιλέγετε το Y ή το Y|X ή το X αλλά όχι το X|Y.

Ανακεφαλαιώνοντας ουσιαστικά σταθμίζουμε την πιθανότητά P(y|x,d) με τους λόγους των πιθανοτήτων P(x), P(X), P(Y), P(y) που είναι η σχέση αυτών των πιθανοτήτων ανάμεσα στο πλήρες δείγμα (γενικό πληθυσμό) και στο μικρότερο δείγμα που έχουμε. Έτσι ουσιαστικά εκτιμώντας την P(y|x,d) πολλαπλασιασμένη με τους λόγους A_yB μπορούμε να αποφανθούμε για την P(Y|X) *(αφού A_y B*P(y|x,d) συγκλίνει στην P(Y|X))*.

3.1.2 Πεπερασμένα μοντέλα με διακριτές επιλογές (Finite Discrete Models)

Μοντέλα με διακριτές επιλογές θεωρούνται αυτά όπου η εξαρτημένη μεταβλητή Y παίρνει διακριτές τιμές (π.χ. Y=1,2,3…), τέτοια μοντέλα για παράδειγμα είναι το πολυωνυμικό και το λογιστικό. Έχουν ως συνάρτηση πιθανότητας την P(Y=j|X), για j=1,2… για πεπερασμένο αριθμό. Θεωρούμε ότι η

$$P(Y = j|D) = \tau_j$$

είναι γνωστή από τον συνολικό πληθυσμό και επίσης ότι η

$$P(y = j|d) = \bar{y}_j$$

είναι γνωστή από το δείγμα μας ή ότι έστω μπορεί να εκτιμηθεί από το δείγμα μας. Σε αυτή την περίπτωση οι παράγοντες διόρθωσης είναι οι ακόλουθοι

$$\mathrm{A_j} = \frac{\tau_j}{\bar{y}_j}\,,\ \mathrm{B}^{-1} = \sum_{j=1}^{J} P(y=j|x,d)\frac{\tau_j}{\bar{y}_j}$$

Συνεπώς αν πολλαπλασιάσουμε με τους παράγοντες διόρθωσης καταλήγουμε στην παρακάτω σχέση:

$$P(y=j|x,d)A_jB = \frac{P(y=j|x,d)\ *\ \frac{\tau_j}{\bar{y}_j}}{\sum_{k=1}^{J} P(y=k|x,d) * \frac{\tau_k}{\bar{y}_k}}$$

η οποία συγκλίνει στην πιθανότητα $P(Y=j|X)$ που μας ενδιέφερε. Οπότε καταλήξαμε ότι αν χρησιμοποιήσουμε τους παράγοντες $\mathrm{A_j}$, B μπορούμε να εκτιμήσουμε την $P(Y=j|X)$.

3.1.3 Δίτιμα μοντέλα

Δίτιμα μοντέλα είναι ουσιαστικά μια υποκατηγορία των πεπερασμένων μοντέλων με διακριτές επιλογές. Η διαφορά εδώ είναι η εξαρτημένη μεταβλητή Y παίρνει μόνο τις τιμές 0,1. Σε αυτά τα μοντέλα λοιπόν έχουμε ότι $\mathrm{P(Y=1)} = \tau$, $\mathrm{P(y=1)} = \bar{y}$. Οι παράγοντες διόρθωσης εδώ είναι οι ακόλουθοι

$$\mathrm{A_1} = \frac{\tau}{\bar{y}}\,,\ \mathrm{A_0} = \frac{(1-\tau)}{(1-\bar{y})}$$

$$\mathrm{B^{-1}} = P(y=1|x,d) * \frac{\tau}{\bar{y}} + [1 - P(y=1|x,d)] * \frac{(1-\tau)}{(1-\bar{y})}.$$

Πολλαπλασιάζοντας πάλι με τους παράγοντες διόρθωσης έχουμε την ακόλουθη σχέση:

$$P(y=1|x,d) * A_1\mathrm{B} = \frac{P(y=1|x,d) * \frac{\tau}{\bar{y}}}{P(y=1|x,d) * \frac{\tau}{\bar{y}} + [1 - P(y=1|x,d)] * \frac{(1-\tau)}{(1-\bar{y})}}$$

$$= \left[1 + \left(\frac{1}{P(y=1|x,d)} - 1\right)\left(\frac{1-\tau}{\tau}\right) * \left(\frac{\bar{y}}{1-\bar{y}}\right)\right]^{-1}$$

3.1.4 Λογιστική παλινδρόμηση

Το λογιστικό μοντέλο είναι αυτό που μας ενδιαφέρει και είναι μια υποκατηγορία των δίτιμων μοντέλων που αναφέραμε στην προηγούμενη παράγραφο. Έχει ως συνάρτηση πιθανότητας την

$$P(y = 1|x, d) = \frac{1}{1 + e^{-x_i\beta}}$$

Πολλαπλασιάζοντας και εδώ με τους παράγοντες διόρθωσης A_1B έχουμε την ακόλουθη σχέση

$$P(y = 1|x, d) * A_1B = \left[1 + e^{-x_i\beta + \ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]}\right]^{-1}$$

Από την οποία βλέπουμε ότι δεν χρειάζεται να αλλάξει ο εκτιμητής μέγιστης πιθανοφάνειας για τον συντελεστή β_1 παρά μόνο για τον σταθερό όρο. Έτσι λοιπόν στο λογιστικό μοντέλο κάτω από τους δειγματοληπτικούς σχεδιασμούς που αναφέραμε ο εκτιμητής μέγιστης πιθανοφάνειας για τον $\hat{\beta}_1$ είναι στατιστικά συνεπής στην εκτίμηση του β_1 ενώ ο εκτιμητής μέγιστης πιθανοφάνειας του $\hat{\beta}_0$ χρειάζεται διόρθωση και πιο συγκεκριμένα χρειάζεται την αφαίρεση της ποσότητας

$$\ln[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)]$$

έτσι ο τελικός εκτιμητής είναι ο

$$\hat{\beta}_0 - \ln[(\frac{1-\tau}{\tau})(\frac{\bar{y}}{1-\bar{y}})]$$

Οι αναλυτές συνήθως δεν ενδιαφέρονται για τους συντελεστές β αλλά για την πιθανότητα

$$P(y = 1|\beta) = \pi_i = \frac{1}{1 + e^{-x_i\beta}}$$

η οποία όμως απαιτεί καλή εκτίμηση των β_1, β_0 . Το μεγάλο πλεονέκτημα αυτής της μεθόδου είναι ότι μπορεί εύκολα να εφαρμοστεί καθώς χρειάζεται τον κλασσικό υπολογισμό των εκτιμητών μέγιστης πιθανοφάνειας για τους συντελεστές β_1, β_0 που μπορεί να γίνει από σχεδόν όλα τα στατιστικά πακέτα και στην συνέχεια να εφαρμοστεί η διόρθωση στον σταθερό όρο. Εάν η εκ των προτέρων γνώση των ποσοστών των γεγονότων και οι ανεξάρτητες μεταβλητές είναι ακριβείς τότε οι εκτιμήσεις είναι συνεπής. Εάν όμως το μοντέλο δεν έχει προσδιοριστεί καλά οι εκτιμήσεις δεν είναι τόσο αποτελεσματικές όσο αυτές της μεθόδου weighting την οποία θα μελετήσουμε στην επόμενη ενότητα.

3.2 Στάθμιση (Weighting)

Στην παρακάτω παράγραφο θα αναπτύξουμε την μέθοδο Weighting για τον υπολογισμό των συντελεστών β_i. Η διαδικασία αυτή έχει ως στόχο να σταθμίσουμε τα δεδομένα με τέτοιο τρόπο ώστε να εξισορροπήσουμε τις διαφορές των ποσοστών των γεγονότων ($Y = 1$) ανάμεσα στον συνολικό πληθυσμό (τ) και στο δείγμα μας ($\bar{y}$). Αυτή η διαδικασία την οποία πρώτοι ανέφεραν οι Manski και Lerman (1977) απαιτεί την μεγιστοποίηση μιας σταθμισμένης εξίσωσης της πιθανοφάνειας και όχι της κλασσικής εξίσωσης. Δηλαδή αντί να μεγιστοποιήσουμε την πιθανοφάνεια

$$\ln L(\beta|y) = -\sum_{i=1}^{n} \ln(1 + e^{(1-2y_i)x_i*\beta})$$

θα πρέπει να μεγιστοποιήσουμε την ακόλουθη σταθμισμένη εξίσωση:

$$\ln L_w(\beta|y) = W_1 \sum_{\{Y_i=1\}} \ln(\pi_i) + W_0 \sum_{\{Y_i=0\}} \ln(1-\pi_i) \ \rightarrow$$

$$\ln L_w(\beta|y) = -\sum_{i=1}^{n} [w_i * \ln(1 + e^{(1-2y_i)x_i*\beta}] \qquad (3.4)$$

όπου τα βάρη είναι

$$W_1 = \frac{\tau}{\bar{y}}, \ W_0 = \frac{(1-\tau)}{(1-\bar{y})}, \ W_i = W_1 * Y_i + W_0 * (1 - Y_i)$$

Παρατηρώντας την εξίσωση (3.4) βλέπουμε ότι είναι η κλασσική εξίσωση πιθανοφάνειας πολλαπλασιασμένη με ένα βάρος. Πράγμα το οποίο διευκολύνει τον υπολογισμό της καθώς σχεδόν όλα τα στατιστικά πακέτα μπορούν να την υπολογίσουν τρέχοντας ένα λογιστικό μοντέλο με βάρος ίσο με w_i. Η μέθοδος στάθμισης (Weighting) μπορεί να είναι πιο ακριβής από την εκ των προτέρων διόρθωση (Prior correction) όταν το δείγμα μας είναι μεγάλο και όταν το ποσοστό των γεγονότων (Y=1) δεν είναι καλά προσδιορισμένο. Αντίθετα η την εκ των προτέρων διόρθωση είναι πιο ακριβής μέθοδος όταν το δείγμα μας είναι μικρό. Γενικά όπως καταλαβαίνουμε η μέθοδος στάθμισης είναι προτιμότερη τις περισσότερες φορές καθώς είναι δύσκολο να προσδιοριστεί με ακρίβεια το ποσοστό των γεγονότων. Όμως υπάρχουν ακόμα δύο ζητήματα τα οποία περιορίζουν την εφαρμογή της. Το πρώτο είναι η μεγάλη μεροληψία που υπάρχει στον συνήθη υπολογισμό των τυπικών σφαλμάτων και το δεύτερο είναι ότι στα σπάνια γεγονότα ή στα πεπερασμένα δείγματα οι διορθώσεις δουλεύουν χωρίς τροποποιήσεις στην μέθοδο prior correction. Στο επόμενο κεφάλαιο θα αναπτύξουμε κάποιες διορθώσεις για τα παραπάνω δύο προβλήματα.

Κ Ε Φ Α Λ Α Ι Ο 4

ΤΡΟΠΟΙ ΑΝΤΙΜΕΤΩΠΙΣΗΣ ΤΗΣ ΜΕΡΟΛΗΨΙΑΣ

Στο παρακάτω κεφάλαιο θα αναπτύξουμε τις μεθόδους που πρότεινε ο Firth (1993) και οι King Gary και Langche Zeng (2001) στην εργασία τους για να υπολογίσουμε ακριβέστερα την $P(Y = 1|\hat{\beta})$. Στα σπάνια γεγονότα ή στα πεπερασμένα δείγματα εμφανίζονται δύο βασικά προβλήματα. Το πρώτο είναι η μεροληψία που υπάρχει στην εκτίμηση $\hat{\beta}$ και το δεύτερο πρόβλημα είναι ότι και εάν καταφέρουμε να αφαιρέσουμε την μεροληψία από την εκτίμηση $\hat{\beta}$ η πιθανότητα $P(Y = 1|\hat{\beta})$ θα συνεχίσει να μην είναι αμερόληπτη όπως θα δούμε παρακάτω.

4.1 Υπολογισμός των παραμέτρων β_i μειώνοντας την μεροληψία

Γνωρίζουμε από την βιβλιογραφία ότι η συνηθισμένη εκτίμηση με την μέθοδο μέγιστης πιθανοφάνειας των παραμέτρων β_i στο λογιστικό μοντέλο είναι συνεπής αλλά επίσης είναι και μεροληπτική σε πεπερασμένα δείγματα και ότι υπάρχουν μέθοδοι που δίνουν πιο συνεπή και λιγότερο μεροληπτικούς εκτιμητές. Η μεροληψία στα σπάνια γεγονότα μεγεθύνετε σε ακόμα μεγαλύτερο βαθμό και πολλοί συγγραφείς έχουν προτείνει ότι ένα δείγμα με 200 παρατηρήσεις και πάνω μειώνει σημαντικά την μεροληψία. Έχουν προταθεί επίσης τρείς μέθοδοι για την διόρθωση της μεροληψίας, η πρώτη είναι η Exact λογιστική παλινδρόμηση η οποία είναι εφαρμόσιμη σε μικρά δείγματα (n<200) με όλες τις ανεξάρτητες μεταβλητές να είναι κατηγορικές και λίγες σε πλήθος. Η δεύτερη είναι η Penalize εκτίμηση της μέγιστης πιθανοφάνειας που έχει προταθεί από τον Firth το 1993 και η τρίτη είναι η μέθοδος που προτείναν οι King and Zeng (2001) στην εργασία τους παίρνοντας και αυτοί ως αφετηρία την προσέγγιση των McCullagh και Nelder (1989) εστιάζοντας την στα σπάνια γεγονότα. Στην παρούσα εργασία θα αναπτύξουμε τις δύο τελευταίες μεθόδους αρχίζοντας με την μέθοδο των King and Zeng (2001) στην επόμενη παράγραφο.

Η μέθοδος αυτή μπορεί να χρησιμοποιηθεί στο λογιστικό μοντέλο με την εκ των προτέρων διόρθωση καθώς επίσης και στο μοντέλο με την σταθμισμένη διόρθωση. Απαιτεί δύο βήματα, στο πρώτο υπολογίζουμε την μεροληψία του εκτιμητή $\hat{\beta}$ με τον τύπο

$$bias(\hat{\beta}) = (X^T W X)^{-1} * (X^T W \xi)$$

Όπου

$$\xi_i = 0.5 Q_{ii}[(1 + W_1)\hat{\pi}_i - W_1$$

Q_{ii} είναι τα διαγώνια στοιχεία της έκφρασης

$$Q = X(X^T WX)^{-1} X^T$$

και επίσης

$$W = diag\{\hat{\pi}_i(1-\hat{\pi}_i)W_i\}$$

Ο τύπος αυτός είναι εύκολο να υπολογιστεί καθώς αρκεί να τρέξουμε μια σταθμισμένη παλινδρόμηση ελαχίστων τετραγώνων με ανεξάρτητη μεταβλητή την X, την ξ ως εξαρτημένη μεταβλητή και W ως το μέτρο στάθμισης.

Στην ακόλουθη παράγραφο θα δείξουμε πως η μεροληψία αυτή μπορεί να υπολογισθεί για κάθε γενικευμένο γραμμικό μοντέλο που έχει την κανονική συνάρτηση σύνδεσης όπως συμβαίνει και στο λογιστικό μοντέλο με την διωνυμική κατανομή. Όπως ήδη αναφέραμε υπολογίζεται μέσω του τύπου

$$bias(\hat{\beta}) = (X^T WX)^{-1} * (X^T W\xi)$$

όπου ο πρώτος όρος $(\mathrm{X}^T\ \mathrm{WX})^{-1}$ είναι ο πίνακας πληροφορίας Fisher και το ξ_i είναι ίσο με

$$\xi_i = -0{,}5(\frac{\mu_i{}'}{\mu_i{}''})\ Q_{ii}\ ,$$

με μ_i είναι η αντίστροφη συνάρτηση σύνδεσης σχετίζοντας την $\mu_i = E(Y_i)$ με την $n_i = x_i\beta$. Επίσης $\mu_i{}'$, $\mu_i{}''$ είναι η πρώτη και η δεύτερη παράγωγος της μ_i ως προς n_i . Ακόμη Q_{ii} είναι τα διαγώνια στοιχεία του $\mathrm{X}(\mathrm{X}^T\ \mathrm{WX})^{-1} * \mathrm{X}^T$. Το κλειδί στην παραγώγιση μας είναι ότι η σταθμισμένη συνάρτηση πιθανοφάνειας μπορεί εύκολα να γίνει ισοδύναμη με την μη σταθμισμένη συνάρτηση αλλάζοντας την συνάρτηση πιθανότητας σε

$$\Pr(Y_i) = {\pi_i}^{W_1 Y_i}(1-\pi_i)^{W_0(1-Y_i)}$$

Έτσι γίνεται

$$\mu_i = E(Y_i) = [\frac{1}{1+e^{-n_i}}]^{W_1} \equiv\ {\pi_i}^{W_1}$$

Και κατά συνέπεια έχουμε:

$\mu_i{}' = W_1{\pi_i}^{W_1}(1-\pi_i)$

$\mu_i{}'' = W_1{\pi_i}^{W_1}(1-\pi_i)[W_1 - (1+W_1)\pi_i]$

$\xi_i = 0{,}5Q_{ii}[(1+W_1)\pi_i - W_1]$

Στην συνέχεια παραγωγίζοντας δύο φορές την σταθμισμένη συνάρτηση πιθανοφάνειας

$$-E\left(\frac{\partial^2 logL_w(\beta|y)}{\partial\beta_i\partial\beta_k}\right) = \sum_{\iota=1}^{n}\pi_i(1-\pi_i)X_j W_i X_k^T = \{\mathrm{X}^T\ \mathrm{WX}\}_{j,k}$$

παίρνουμε το $W = diag\{\hat{\pi_\iota}(1-\hat{\pi_\iota})W_i\}$.

Αφού έχουμε τώρα υπολογίσει την μεροληψία ακολουθεί το επόμενο βήμα στο οποίο χρησιμοποιούμε τον ακόλουθο τύπο για τον υπολογισμό του εκτιμητή του β με διόρθωση

$$\tilde{\beta} = \hat{\beta} - bias(\hat{\beta})$$

από εδώ και πέρα θα συμβολίζουμε με $\tilde{\beta}$ την διορθωμένη εκτίμηση και με $\hat{\beta}$ την εκτίμηση με την μέθοδο της μέγιστης πιθανοφάνειας. Επίσης ο πίνακας διακύμανσης της εκτίμησης $\tilde{\beta}$ βρίσκεται με τον τύπο

$$V(\tilde{\beta}) = \left[\frac{n}{n+k}\right]^2 V(\hat{\beta})$$

Ακόμη θα πρέπει να αναφέρουμε ότι εάν ισχύουν οι παρακάτω δύο ανισότητες

$$\left[\frac{n}{n+k}\right]^2 < 1 \text{ και } V(\tilde{\beta}) < V(\hat{\beta})$$

τότε μειώνοντας την μεροληψία της εκτίμησης του β μειώνεται ταυτόχρονα και η διακύμανση.

4.2 Εκτιμητές μέγιστης πιθανοφάνειας με ποινή (PMLE Firth)

Η προσαρμογή του εκτιμητή μέγιστης πιθανοφάνειας με ποινή είναι μια γενική μέθοδος που εμπεριέχει διάφορες διορθώσεις, εμείς εδώ θα αναφέρουμε την διόρθωση που πρότεινε ο Firth (1993). Αυτή η μέθοδος βοηθάει στο να διορθωθεί η μεροληψία σε μικρά δείγματα και μπορεί να εφαρμοστεί σε διάφορα μοντέλα όπως στο λογιστικό μοντέλο, στην παλινδρόμηση Poisson καθώς επίσης και στην Cox παλινδρόμηση. Ας θεωρήσουμε αρχικά ένα μοντέλο με εξαρτημένη μεταβλητή Y και ένα σύνολο από ανεξάρτητες X συμπεριλαμβανομένου και του σταθερού όρου. Έστω τώρα λοιπόν X ένας πίνακας που περιέχει τα παρατηρούμενα x και Y ένα διάνυσμα που περιέχει τα παρατηρούμενα y. Ο Firth πρότεινε ως εναλλακτική συνάρτηση πιθανοφάνειας την ακόλουθη σχέση:

$$logL^*(\beta) = logL(\beta) + \frac{1}{2}\log det(I(\beta))$$

όπου $logL(\beta)$ είναι ο γνωστός λογάριθμός της συνάρτησης πιθανοφάνειας.

Η διόρθωση ουσιαστικά αναφέρεται στο δεύτερο όρο $\frac{1}{2}logdet(I(\beta))$ που είναι το μισό του λογαρίθμου της ορίζουσας του παρατηρούμενου πίνακα πληροφορίας $I(\beta)$. Στο λογιστικό μοντέλο συγκεκριμένα ο πίνακας ισούται με $I(\beta) = (\acute{X}WX)$ όπου W είναι ένας διαγώνιος πίνακας με στοιχεία $diag[\pi_i(1-\pi_i)]$.

Η μέθοδος αυτή μπορεί να εκτιμήσει πάντα χωρίς προβλήματα και ο πίνακας W μεγιστοποιείτε όταν $\pi_i = \frac{1}{2}$.

Πράγματι κάθε όρος $\pi_i(1-\pi_i)$ της ορίζουσας μεγιστοποιείται για $\pi_i = \frac{1}{2}$ αφού η παράγωγος είναι $-2\pi_i + 1$ και η δεύτερη παράγωγος -2<0.

Επίσης θα πρέπει να αναφέρουμε ότι χάρη στον Heinze (2013) και τους συνεργάτες του η μέθοδος αυτή έχει ενσωματωθεί στο στατιστικό πακέτο R μέσο της βιβλιοθήκης logistf την οποία θα χρησιμοποιήσουμε και εμείς για να την εφαρμόσουμε σε επόμενο κεφάλαιο.

4.3 Υπολογισμός της πιθανότητας π_i

Στην παρακάτω παράγραφο θα ασχοληθούμε με την εκτίμηση της πιθανότητας

$$\pi_i = \frac{1}{1 + e^{-x_i * \beta}}$$

που είναι η ποσότητα που ενδιαφέρει όλους τους ερευνητές. Έχοντας υπολογίσει την εκτίμηση με διόρθωση των παραμέτρων β ($\tilde{\beta}$) η οποία έχει λιγότερη μεροληψία και μικρότερη διακύμανση εφόσον ισχύουν οι ανισότητες που αναφέραμε προηγουμένως τότε κατά συνέπεια θα έχει και μικρότερο MSE σφάλμα. Χρησιμοποιώντας την τώρα λοιπόν θα υπολογίσουμε την ακόλουθη πιθανότητα

$$\tilde{\pi}_0 = P\left(Y_0 = 1 \middle| \tilde{\beta}\right) = \frac{1}{1+e^{-x_0 * \tilde{\beta}}} \qquad (4.0)$$

που είναι προτιμότερη από την αντίστοιχη πιθανότητα που χρησιμοποιεί την $\hat{\beta}$. Παρόλα αυτά η εκτίμηση $\tilde{\pi}$ και πάλι δεν είναι ιδανική διότι αγνοεί την αβεβαιότητα της εκτίμησης του $\tilde{\beta}$. Σε αρκετές περιπτώσεις αγνοώντας την αβεβαιότητα η εκτίμηση δεν επηρεάζεται και αλλάζει μόνο η τυπική απόκλιση πράγμα το οποίο δεν συμβαίνει στα σπάνια γεγονότα που μελετάμε εμείς. Αυτό έχει σαν αποτέλεσμα πολύ μικρές τιμές στην εκτίμηση των σπανίων γεγονότων.

Για να γίνει πιο κατανοητό ας υποθέσουμε ότι έχουμε μια συνεχή τυχαία μεταβλητή την οποία την μετατρέπουμε σε δίτιμη χρησιμοποιώντας ένα κατώφλι. Εάν λοιπόν αγνοήσουμε την αβεβαιότητα της εκτίμησης του β αυτό έχει ως αποτέλεσμα να οδηγηθούμε σε κατανομή με μικρή διακύμανση και κατά συνέπεια πιο στενές ουρές. Υπολογίζοντας τώρα την πιθανότητα δεξιά του κατωφλιού μέσω του εμβαδού θα μας δώσει μικρές τιμές ενώ σε αντίθετη περίπτωση λαμβάνοντας υπόψη την αβεβαιότητα της εκτίμησης οι ουρές θα γίνουν πιο φαρδιές και κατά συνέπεια και η εκτίμηση της πιθανότητας μεγαλύτερη. Έτσι λοιπόν για να υπολογίσουμε την εκτίμηση της πιθανότητας $\tilde{\pi}$ θα χρησιμοποιήσουμε τον ακόλουθο τύπο που πρότειναν οι King Gary και Langche Zeng (2001):

$$P(Y_i = 1) = \int P(Yi = 1|\beta^*)\, P(\beta^*) d\beta^* \qquad (4.1)$$

Όπου β^* είναι μια ψευδομεταβλητή που την χρησιμοποιούμε για την εκτίμηση της αβεβαιότητας. Η εκτίμηση αυτή μπορεί να θεωρηθεί και ως συχνότητα της δειγματικής κατανομής του $\tilde{\beta}$ και συνεπώς η σχέση (4.1) είναι η αναμενόμενη ποσότητα $E(P(Y_i = 1|\tilde{\beta})$ η οποία είναι η εκτίμηση της πιθανότητας π.

Για τον υπολογισμό της σχέσης (4.1) έχουν προταθεί δύο προσεγγίσεις. Η πρώτη απαιτεί προσομοιώσεις. Παίρνουμε λοιπόν ένα τυχαίο β από το P(β) και υπολογίζουμε την πιθανότητα π και αυτό το επαναλαμβάνουμε αρκετές φορές και υπολογίζουμε τον μέσο όρο. Ο αριθμός των επαναλήψεων είναι ανάλογος με την ακρίβεια της εκτίμησης που θέλουμε.

Ο δεύτερος τρόπος είναι γενικά εύκολος. Απαιτεί περισσότερο υπολογιστικό κόστος αλλά βοηθάει στην εκ βαθέως κατανόηση της διόρθωσης. Αυτή η μέθοδος μας λέει ότι εξίσωση (4.1) μπορεί να προσεγγισθεί με την σχέση:

$$P(Y_i = 1) = \tilde{\pi}_i + C_i \tag{4.2}$$

Όπου

$$C_i = (0.5 - \tilde{\pi}_i)\tilde{\pi}_i(1 - \tilde{\pi}_i)X_0 V(\tilde{\beta})X_0^T$$

Πιο αναλυτικά αν αναπτύξουμε την σειρά Taylor μέχρι τη δεύτερη τάξη για την π_0 έχουμε:

$$P(Y_0 = 1) = \tilde{\pi}_0 + \left[\frac{\partial \pi_0}{\partial \beta}\right]_{\beta=\tilde{\beta}} (\beta - \tilde{\beta}) + \frac{1}{2}(\beta - \tilde{\beta})' \left[\frac{\partial^2 \pi_0}{\partial \beta \partial \acute{\beta}}\right]_{\beta=\tilde{\beta}} (\beta - \tilde{\beta})$$

Χρησιμοποιώντας τώρα τις παρακάτω παραγωγίσεις:

$$\left[\frac{\partial \pi_0}{\partial \beta}\right]_{\beta=\tilde{\beta}} \overset{(4.0)}{\Longrightarrow} \frac{-1(e^{-x_0*\tilde{\beta}} * (-x_0))}{(1 + e^{-x_0*\tilde{\beta}})^2} = \frac{-x_0 * e^{-x_0*\tilde{\beta}}}{(1 + e^{-x_0*\tilde{\beta}})^2} = x_0 * (1 - \pi_0) * (\pi_0)$$

$$\frac{\partial^2 \pi_0}{\partial \beta \partial \acute{\beta}} = \frac{-{x_0}^2 * e^{-x_0*\tilde{\beta}}}{(1 + e^{-x_0*\tilde{\beta}})^2} + \frac{2{x_0}^2 * e^{-2x_0*\tilde{\beta}}}{(1 + e^{-x_0*\tilde{\beta}})^3} = -{x_0}^2(1 - \pi_0)\pi_0 + 2{x_0}^2((1 - \pi_0)^2 * \pi_0)$$

ο δεύτερος όρος γίνεται

$$\tilde{\pi}_0(1 - \tilde{\pi}_0)X_0(\beta - \tilde{\beta})$$

και ο τρίτος

$$(0.5 - \tilde{\pi}_0)\tilde{\pi}_0(1 - \tilde{\pi}_0)X_0 D\acute{x}_0$$

με D να είναι ένας $k * k$ πίνακας όπου για k, j το στοιχείο του ισούται με $(\beta_k - \widetilde{\beta_k})\left(\beta_j - \widetilde{\beta_j}\right)'$. Παίρνοντας τώρα την αναμενόμενη τιμή της πιθανότητας π0 και χρησιμοποιώντας ότι $b = E\left(\beta - \tilde{\beta}\right)$ και ότι ο πίνακας διακυμάνσεων είναι $V\left(\tilde{\beta}\right)$έχουμε:

$$P(Y_0 = 1) = E\left(\frac{1}{1 + e^{-x_0 * \beta}}\right) =$$

$$\tilde{\pi}_0 + \tilde{\pi}_0(1 - \tilde{\pi}_0)X_0 b + (0.5 - \tilde{\pi}_0)(\tilde{\pi}_0 - (\tilde{\pi}_0)^2)X_0[V\left(\tilde{\beta}\right) + bb^T]\acute{X}_0 \qquad (4.3)$$

και καθώς το b τείνει στο μηδέν η παραπάνω σχέση γίνεται ίση με την (4.2) εξίσωση.

ΚΕΦΑΛΑΙΟ 5

ΣΥΓΚΡΙΣΗ ΤΩΝ ΜΕΘΟΔΩΝ ΜΕ ΤΗ ΒΟΗΘΕΙΑ ΠΡΟΣΟΜΟΙΩΜΕΝΩΝ ΔΕΙΓΜΑΤΩΝ

Στο παρακάτω κεφάλαιο θα προσομοιώσουμε δεδομένα μεγέθους n για λογιστική παλινδρόμηση με σταθερό όρο και μια ανεξάρτητη μεταβλητή και στην συνέχεια θα εκτελέσουμε λογιστική παλινδρόμηση για να ελέγξουμε την ακρίβεια των εκτιμήσεων καθώς μειώνεται ο αριθμός των παρατηρήσεων και ο αριθμός των γεγονότων. Έπειτα θα τρέξουμε τα μοντέλα με τις διορθώσεις για δούμε πώς βελτιώνονται οι εκτιμήσεις.

5.1 Προσομοιώσεις δεδομένων για λογιστική παλινδρόμηση

Θα προσομοιώσουμε δεδομένα χρησιμοποιώντας το στατιστικό πακέτο R τρέχοντας τον παρακάτω κώδικα που βρήκαμε στο ακόλουθο link:
https://stats.stackexchange.com/questions/46523/how-to-simulate-artificial-data-for-logistic-regression/46525
Κάνοντας σε αυτόν κάποιες αλλαγές έτσι ώστε να τον προσαρμόσουμε στα ζητούμενα μας δηλαδή στο να παράγει δεδομένα με σταθερό όρο και μια μόνο ανεξάρτητη μεταβλητή καταλήξαμε στον ακόλουθο κώδικας:

```
set.seed(1990)
n.obs= …
b0= …
b1= …
x1 = rnorm(n.obs, mean=0, sd=1)            # create continuous variable
z = b0 + b1*x1                             # linear combination
pr = 1/(1+exp(-z))
y = rbinom(length(x1),1,pr)
df = data.frame(y=y,x1=x1)
summary(glm( y~x1,data=df, family="binomial"))
```

Ας εξηγήσουμε τώρα λοιπόν τι κάνει γραμμή προς γραμμή έτσι ώστε να γίνει πιο κατανοητή η διαδικασία της προσομοίωσης. Στην πρώτη θέτουμε έναν συγκεκριμένο αριθμό στην συνάρτηση set.seed έτσι ώστε να παράγει κάθε φορά τους ίδιους αριθμούς με τον ίδιο αριθμό

παρατηρήσεων και με τους ίδιους συντελεστές β_0, β_1. Στις επόμενες τρείς θέτουμε των αριθμό των παρατηρήσεων που θέλουμε και τους συντελεστές β_0, β_1 αντίστοιχα. Στην επόμενη παράγει αριθμούς από τυπική κανονική κατανομή με μέγεθος παρατηρήσεων n, όσους δηλαδή του θέσαμε. Στις γραμμές 6,7,8 φτιάχνει έναν γραμμικό συνδυασμό των συντελεστών και της ανεξάρτητης μεταβλητής x_1 και μετά φτιάχνει για κάθε παρατήρηση (πραγματικό αριθμό) μια πιθανότητα χρησιμοποιώντας την inverse logit συνάρτηση και στην συνέχεια παράγει αριθμούς από την διωνυμική κατανομή μεγέθους n με πιθανότητα αυτήν που δημιούργησε πριν. Τέλος στις δύο τελευταίες γραμμές δημιουργεί τον πίνακα με τα δεδομένα μας και εκτελεί την λογιστική παλινδρόμηση.

Στο δικό μας πείραμα θα χρησιμοποιήσουμε μέγεθος δείγματος n ίσο με 5000, 1000, 500, 250 και 100 και για κάθε n θα προσομοιώσουμε δεδομένα έτσι ώστε ο αριθμός των γεγονότων να είναι 50%, 10%, 5% και 1%. Για να πετύχουμε των αριθμό των γεγονότων που θέλουμε θα χρησιμοποιήσουμε τους παρακάτω συνδυασμούς β_0 και β_1:

- Για αριθμό γεγονότων ίσο με 50% θα πάρω $\beta_0=0$ και $\beta_1=2$
- Για αριθμό γεγονότων ίσο με 10% θα πάρω $\beta_0=-3.3$ και $\beta_1=2$
- Για αριθμό γεγονότων ίσο με 5% θα πάρω $\beta_0=-4.3$ και $\beta_1=2$
- Για αριθμό γεγονότων ίσο με 1% θα πάρω $\beta_0=-6.6$ και $\beta_1=2$

Έτσι λοιπόν καταλήγουμε στον παρακάτω σχεδιασμό προσομοίωσης:

	n				
p	**5000**	**1000**	**500**	**250**	**100**
0.5	2500	500	250	125	50
0.1	500	100	50	25	10
0.05	250	50	25	13	5
0.01	50	10	5		

Πίνακας 5.1: Σχεδιασμός προσομοίωσης

Όπου επίσης έχουμε τους ακόλουθους γραμμικούς σχεδιασμούς:

$$z_{0.5} = 2x_1 \text{ , } x_1 \sim N(0,1)$$

$$z_{0.1} = -3.3+2x_1 \text{ , } x_1 \sim N(0,1)$$

$$z_{0.05} = -4.3+2x_1 \text{ , } x_1 \sim N(0,1)$$

$$z_{0.01} = -6.6+2x_1 \text{ , } x_1 \sim N(0,1)$$

Αφού τρέξουμε τον κώδικα για καθένα από τους παραπάνω συνδυασμούς θα πάρουμε τις παρακάτω εκτιμήσεις του καθενός μοντέλου:

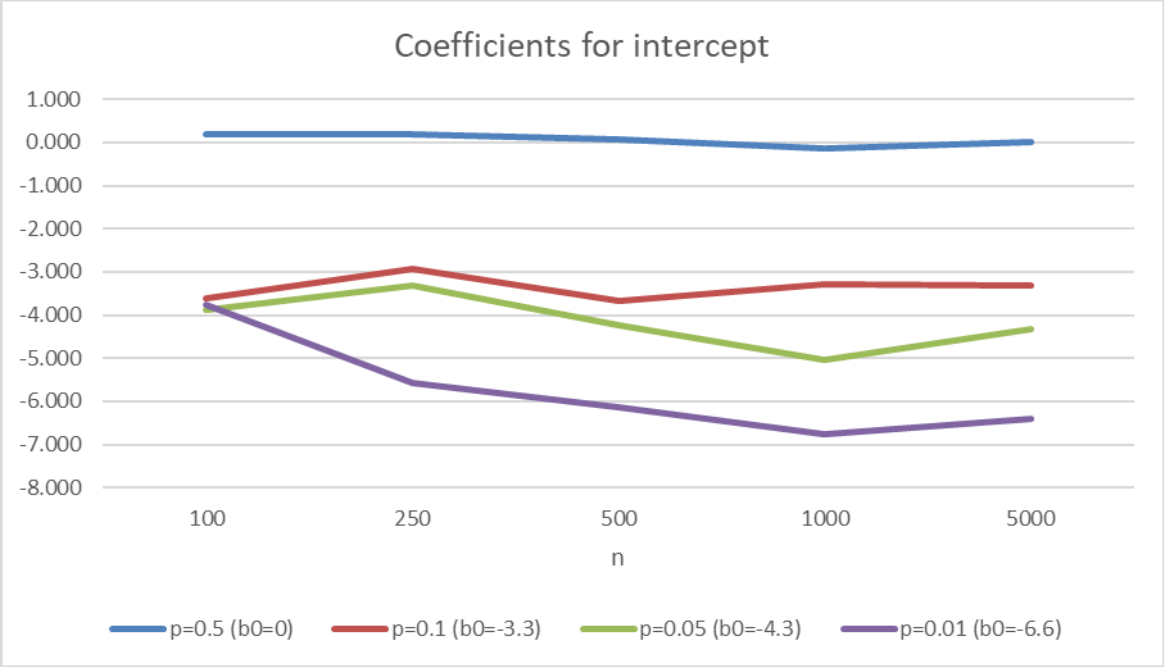
Coefficients for intercept					
	n				
p (β_0)	**100**	**250**	**500**	**1000**	**5000**
0.5 (β0=0)	0.191	0.205	0.083	-0.150	0.010
0.1 (β0=-3.3)	-3.618	-2.942	-3.664	-3.293	-3.325
0.05 (β0=-4.3)	-3.893	-3.323	-4.238	-5.041	-4.319
0.01 (β0=-6.6)	-3.772	-5.557	-6.126	-6.764	-6.414

Πίνακας 5.2: Εκτιμήσεις του β_0 με Λογιστική Παλινδρόμηση

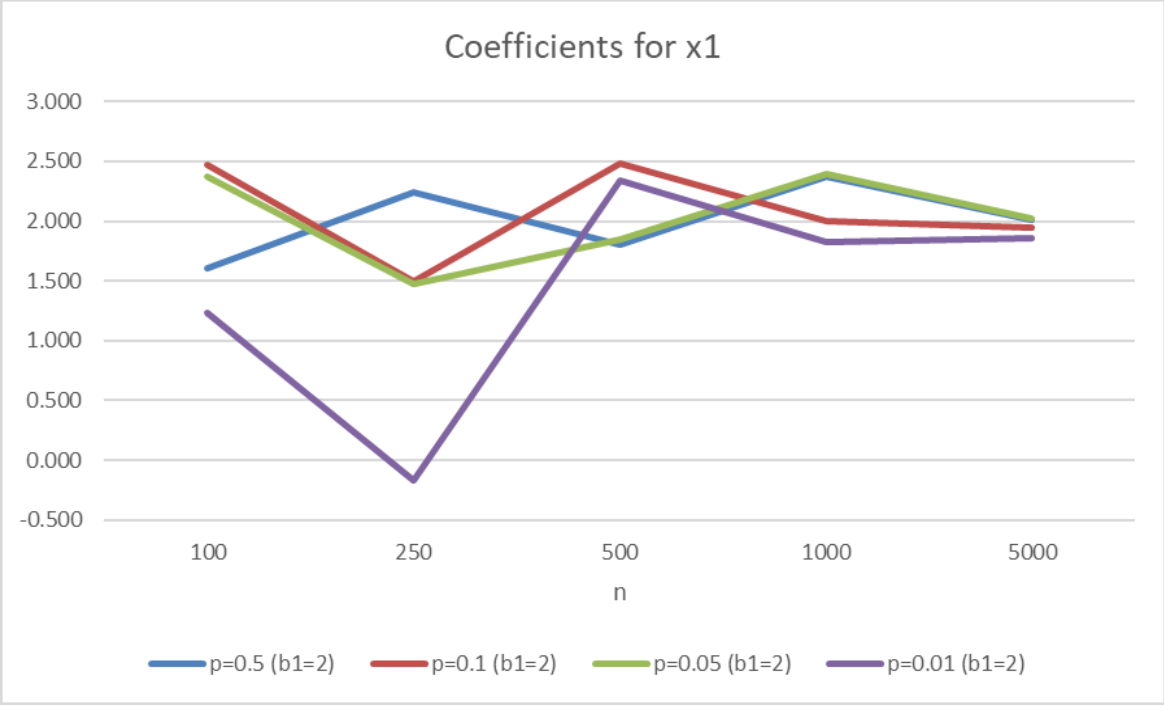
Coefficients for x_1					
	n				
p (β_1=2)	**100**	**250**	**500**	**1000**	**5000**
0.5	1.608	2.242	1.798	2.375	2.015
0.1	2.474	1.493	2.480	1.996	1.944
0.05	2.370	1.476	1.852	2.394	2.020
0.01	1.231	-0.171	2.344	1.829	1.863

Πίνακας 5.3: Εκτιμήσεις του β_1 με Λογιστική Παλινδρόμηση

Με μια πρώτη ματιά παρατηρούμαι ότι καθώς μικραίνει το δείγμα μας η εκτίμηση των συντελεστών απομακρύνετε από την πραγματική τιμή που έχουμε ορίσει, όπως επίσης και όταν το πλήθος των γεγονότων στο δείγμα μας μικραίνει και πάλι η εκτίμηση απομακρύνετε από την πραγματική τιμή. Για να γίνουν πιο κατανοητές οι παρατηρήσεις που κάναμε θα απεικονίσουμε τους δύο παραπάνω πίνακες με την βοήθεια του Excel στα δύο παρακάτω σχήματα:



Γράφημα 5.1: Γράφημα για τις εκτιμήσεις του β_0



Γράφημα 5.2: Γράφημα για τις εκτιμήσεις του β_1

Παρατηρώντας τώρα και τα γραφήματα γίνεται πιο εμφανές το πρόβλημα της εκτίμησης τον συντελεστών με την κλασσική λογιστική παλινδρόμηση όταν μικραίνει το δείγμα και έχουμε σπάνια γεγονότα. Συνεπώς οι διορθώσεις που αναφέραμε είναι ένα πολύ χρήσιμο εργαλείο στα χέρια των ερευνητών έτσι ώστε να έχουν ακριβείς εκτιμήσεις.

Τελικά όμως όπως έχουμε αναφέρει και στα προηγούμενα κεφάλαια αυτό που ενδιαφέρει περισσότερο τους ερευνητές είναι η πιθανότητα να συμβεί το γεγονός γι' αυτό τον λόγο στην επόμενη παράγραφο θα υπολογίσουμε την πιθανότητα

$$P(y = 1|x_1 = 1) = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$$

χρησιμοποιώντας τους συντελεστές που εκτίμησε με το μοντέλο και τους πραγματικούς συντελεστές που θέσαμε στην προσομοίωση έτσι ώστε να δούμε κατά πόσο η πιθανότητα της εκτίμησης είναι κοντά με την πραγματική. Υπολογίζοντας τώρα τις πιθανότητες για n ίσο με 5000, 1000, 500, 250 και 100 και ο αριθμός των γεγονότων να είναι 50%, 10%, 5%, 1% για κάθε n καταλήγουμε στους παρακάτω πίνακες:

Pr(Y=1|x1=1) για n=100 και p=0.01				
	β_0	β_1	$Z_{0.01}$	Pr
True	-6.6	2	-4.60	1.00%
glm	-3.772	1.231	-2.54	7.31%

Pr(Y=1|x1=1) για n=250 και p=0.01				
	β_0	β_1	$Z_{0.01}$	Pr
True	-6.6	2	-4.60	1.00%
glm	-5.557	-0.171	-5.73	0.32%

Pr(Y=1|x1=1) για n=500 και p=0.01				
	β_0	β_1	$Z_{0.01}$	Pr
True	-6.6	2	-4.60	1.00%
glm	-6.126	2.344	-3.78	2.23%

Pr(Y=1|x1=1) για n=1000 και p=0.01				
	β_0	β_1	$Z_{0.01}$	Pr
True	-6.6	2	-4.60	1.00%
glm	-6.764	1.829	-4.93	0.71%

Pr(Y=1|x1=1) για n=5000 και p=0.01				
	β_0	β_1	$Z_{0.01}$	Pr
True	-6.6	2	-4.60	1.00%
glm	-6.414	1.863	-4.55	1.04%

Πίνακας 5.4: Εκτιμήσεις των πιθανοτήτων για αριθμό γεγονότων 1% και για διάφορα n

Pr(Y=1|x1=1) για n=100 και p=0.05				
	β_0	β_1	$Z_{0.05}$	Pr
True	-4.3	2	-2.30	9.11%
glm	-3.893	2.370	-1.52	17.90%

Pr(Y=1|x1=1) για n=250 και p=0.05				
	β_0	β_1	$Z_{0.05}$	Pr
True	-4.3	2	-2.30	9.11%
glm	-3.323	1.4762	-1.85	13.62%

Pr(Y=1|x1=1) για n=500 και p=0.05				
	β_0	β_1	$Z_{0.05}$	Pr
True	-4.3	2	-2.30	9.11%
glm	-4.2379	1.8517	-2.39	8.42%

Pr(Y=1|x1=1) για n=1000 και p=0.05				
	β_0	β_1	$Z_{0.05}$	Pr
True	-4.3	2	-2.30	9.11%
glm	-5.041	2.394	-2.65	6.61%

Pr(Y=1|x1=1) για n=5000 και p=0.05				
	β_0	β_1	$Z_{0.05}$	Pr
True	-4.3	2	-2.30	9.11%
glm	-4.319	2.020	-2.30	9.13%

Πίνακας 5.5: Εκτιμήσεις των πιθανοτήτων για αριθμό γεγονότων 5% και για διάφορα n

Pr(Y=1|x1=1) για n=100 και p=0.1				
	β_0	β_1	$Z_{0.1}$	Pr
True	-3.3	2	-1.30	21.42%
glm	-3.618	2.474	-1.14	24.17%

Pr(Y=1|x1=1) για n=250 και p=0.1				
	β_0	β_1	$Z_{0.1}$	Pr
True	-3.3	2	-1.30	21.42%
glm	-2.942	1.493	-1.45	19.01%

Pr(Y=1|x1=1) για n=500 και p=0.1				
	β_0	β_1	$Z_{0.1}$	Pr
True	-3.3	2	-1.30	21.42%
glm	-3.664	2.480	-1.18	23.45%

Pr(Y=1|x1=1) για n=1000 και p=0.1				
	β_0	β_1	$Z_{0.1}$	Pr
True	-3.3	2	-1.30	21.42%
glm	-3.293	1.996	-1.30	21.48%

Pr(Y=1|x1=1) για n=5000 και p=0.1				
	β_0	β_1	$Z_{0.1}$	Pr
True	-3.3	2	-1.30	21.42%
glm	-3.325	1.944	-1.38	20.10%

Πίνακας 5.6: Εκτιμήσεις των πιθανοτήτων για αριθμό γεγονότων 10% και για διάφορα n

Pr(Y=1|x1=1) για n=100 και p=0.5				
	β_0	β_1	$Z_{0.5}$	Pr
True	0	2	2.00	88.08%
glm	0.191	1.608	1.80	85.81%

Pr(Y=1|x1=1) για n=250 και p=0.5				
	β_0	β_1	$Z_{0.5}$	Pr
True	0	2	2.00	88.08%
glm	0.205	2.2423	2.45	92.03%

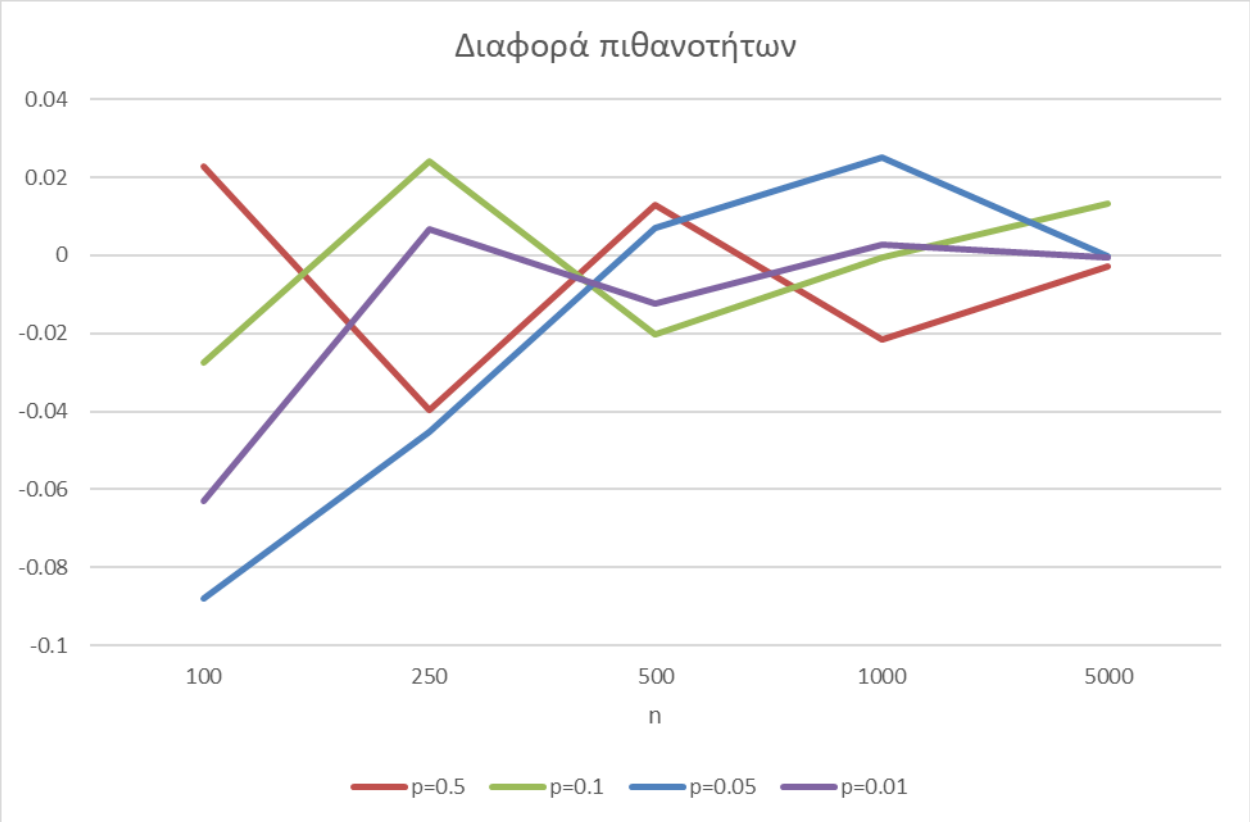
Pr(Y=1|x1=1) για n=500 και p=0.5				
	β_0	β_1	$Z_{0.5}$	Pr
True	0	2	2.00	88.08%
glm	0.0831	1.7978	1.88	86.77%

Pr(Y=1|x1=1) για n=1000 και p=0.5				
	β_0	β_1	$Z_{0.5}$	Pr
True	0	2	2.00	88.08%
glm	-0.150	2.375	2.23	90.25%

Pr(Y=1|x1=1) για n=5000 και p=0.5				
	β_0	β_1	$Z_{0.5}$	Pr
True	0	2	2.00	88.08%
glm	0.010	2.015	2.03	88.35%

Πίνακας 5.7: Εκτιμήσεις των πιθανοτήτων για αριθμό γεγονότων 50% και για διάφορα n

Θα παρουσιάσουμε τώρα σε γράφημα τις διαφορές των πραγματικών πιθανοτήτων από αυτών που εκτιμήσαμε έτσι ώστε να γίνει το πιο κατανοητό το πρόβλημα της εκτίμησης των γεγονότων όταν έχουμε μικρό δείγμα με σπάνια γεγονότα. Το γράφημα είναι τα ακόλουθο:



Γράφημα 5.3: Γράφημα με την διάφορα των πραγματικών πιθανοτήτων από αυτών που εκτιμήσαμε

Και σε αυτό το σχήμα παρατηρούμε πώς με την κλασσική λογιστική παλινδρόμηση οι εκτιμήσεις για τα γεγονότα απέχουν αρκετά από την πραγματική τιμή και πως μόνο όταν έχουμε πολλά δεδομένα πράγμα αδύνατον για τα σπάνια γεγονότα οι εκτιμήσεις είναι ακριβείς.

5.2 Σύγκριση των μεθόδων σε σχέση με την ακρίβεια των εκτιμήσεων

Στην παρακάτω παράγραφο θα εκτιμήσουμε τους συντελεστές με την μέθοδο που πρότειναν οι King Gary και Langche Zeng (2001) καθώς επίσης και με την μέθοδο μέγιστης πιθανοφάνειας με ποινή που πρότεινε ο Firth (1993). Θα εκτελέσουμε 1000 προσομοιώσεις και θα συγκρίνουμε τα αποτελέσματα των δύο αυτών εκτιμήσεων με την μέθοδο της κλασσικής λογιστικής παλινδρόμησης.

Για να εκτιμήσουμε τους συντελεστές με την μέθοδο της μέγιστης πιθανοφάνειας με ποινή του Firth θα χρησιμοποιήσουμε το στατιστικό πακέτο R και την βιβλιοθήκη logistf. Μπορεί κανείς να βρει πληροφορίες για την συγκεκριμένη βιβλιοθήκη στον παρακάτω σύνδεσμο:

https://cran.r-project.org/web/packages/logistf/index.html

Αξίζει τώρα να αναφέρουμε κάποιες βασικές πληροφορίες σχετικά με την βιβλιοθήκη. Οι συντάκτες αυτής είναι οι Georg Heinze, Meinhard Ploner, Daniela Dunkler, Harry Southworth. Ο σκοπός της δημιουργίας της είναι να προσαρμόζει μοντέλα λογιστικής παλινδρόμησης χρησιμοποιώντας την μέθοδο του Firth που είναι ισοδύναμη με την μέθοδο της ποινής του Jeffrey. Η κύρια συνάρτηση αυτής της βιβλιοθήκης είναι η “logistf” όπου προσαρμόζει το ζητούμενο μοντέλο και έχει την ακόλουθη γενική μορφή:

logistf(formula = attr(data, "formula"), data = sys.parent(), pl = TRUE, alpha = 0.05, control, plcontrol, firth = TRUE, init, weights, plconf = NULL, dataout = TRUE, ...)

Τα βασικά στοιχεία που θα πρέπει να συμπληρώσουμε είναι η *formula* στην οποία γράφουμε τον τύπο της παλινδρόμησης που θέλουμε να εκτελέσουμε, στην θέση *data* τοποθετούμε το σετ δεδομένων που έχουμε, στην θέση *pl* καθορίζουμε αν τα διαστήματα εμπιστοσύνης και οι δοκιμές θα πρέπει να βασίζονται στην πιθανοφάνεια με ποινή ή όχι, στην θέση alpha γράφουμε το επίπεδο σημαντικότητας που θέλουμε (ως προκαθορισμένο είναι το 0.05) καθώς επίσης θα πρέπει να ορίσουμε στην θέση *firth* αν θέλουμε να εκτελέσει την μέθοδο του Firth (TRUE) ή την κλασσική εκτίμηση με την μέθοδο της μέγιστης πιθανοφάνειας (FALSE).

Για να εκτιμήσουμε τους συντελεστές με την μέθοδο που πρότειναν οι King Gary και Langche Zeng (2001) θα χρησιμοποιήσουμε την βιβλιοθήκη Zelig που υπάρχει διαθέσιμη για το στατιστικό πακέτο R. Μπορεί κανείς να βρει πληροφορίες για την συγκεκριμένη βιβλιοθήκη στον παρακάτω σύνδεσμο:

https://cran.r-project.org/web/packages/Zelig/index.html

Οι συντάκτες αυτής είναι οι Christine Choirat, Christopher Gandrud, James Honaker, Kosuke Imai, Gary King, Olivia Lau, IQSS Harvard University. Κύριος σκοπός αυτής είναι να φέρει μαζί κοινά στατιστικά μοντέλα σε μια ενιαία βιβλιοθήκη. Ένα από αυτά τα μοντέλα είναι και αυτό που μας ενδιαφέρει. Η βασική συνάρτηση αυτής της βιβλιοθήκης είναι η "Zelig" και για την λογιστική παλινδρόμηση για σπάνια γεγονότα έχει την ακόλουθη γενική μορφή:

```
zelig(Y ~ X1 + X2 + ... , model = "relogit", tau = NULL, case.control = c("prior", "weighting"), bias.correct = TRUE, data = mydata, ...)
```

Τα βασικά στοιχεία που θα πρέπει να συμπληρώσουμε εδώ είναι το πρώτο στο οποίο γράφουμε τον τύπο της παλινδρόμησης που θέλουμε να εκτελέσουμε, επίσης στην θέση data τοποθετούμε το σετ δεδομένων που έχουμε, στην θέση model γράφουμε *relogit* καθώς το μοντέλο που θέλουμε να τρέξουμε είναι λογιστική παλινδρόμηση για σπάνια γεγονότα, στην θέση *tau* γράφουμε το ποσοστό των γεγονότων στον συνολικό πληθυσμό, στην θέση *case.control* συμπληρώνουμε την μέθοδο που θέλουμε να χρησιμοποιήσουμε *prior* ή *weighting* καθώς επίσης συμπληρώνουμε με TRUE ή FALSE στην θέση *bias.correct* αν θέλουμε να γίνει διόρθωση έτσι ώστε να μειωθεί η μεροληψία.

Τώρα θα προσομοιώσουμε δεδομένα με μέγεθος παρατηρήσεων n=250 και συντελεστές β_0= -4.3 και β_1=2 έτσι ώστε να έχουμε μικρό αριθμό γεγονότων (κοντά στο 5%). Για να το καταφέρουμε αυτό θα χρησιμοποιήσουμε τον κώδικα που να αναφέραμε στην προηγούμενη παράγραφο με κάποιες διαφορές. Θα τον εκτελέσουμε λοιπόν μέσα σε έναν επαναληπτικό βρόγχο 1000 φορές όπου σε κάθε επανάληψη θα εκτελούμε και τα τρία μοντέλα με σκοπό να παίρνουμε τους συντελεστές. Στην συνέχεια θα υπολογίσουμε το μέσω τετραγωνικό σφάλμα αυτών με τον ακόλουθο τύπο:

$$E(\widehat{\beta_\iota} - \beta_i)^2$$

Για i=0,1 καθώς έχουμε μόνο σταθερό όρο και μια ανεξάρτητη μεταβλητή. Στις επόμενες παραγράφους θα παραθέσουμε τα αποτελέσματα αφού τρέξουμε τον κώδικα.

Στον παρακάτω πίνακα βλέπουμε την μέση τιμή των συντελεστών μετά από την εκτέλεση κώδικα:

	Means	
	(Intercept)	**X_1**
MLE	-4.53	2.15
PMLE	-4.32	2.03
Zelig	-4.31	2.02

Πίνακας 5.8: Μέση τιμή των συντελεστών μετά από την εκτέλεση του κώδικα

Όπου παρατηρούμε πως οι εκτιμήσεις με τις μεθόδους που αναφέραμε είναι αρκετά πιο κοντά στις πραγματικές τιμές πράγμα το οποίο κάνει τις εκτιμήσεις πιο ακριβείς. Θα υπολογίσουμε επίσης την πιθανότητα

$$P(y = 1|x_1 = 1) = \frac{e^{\beta_0+\beta_1 x_1}}{1+e^{\beta_0+\beta_1 x_1}}$$

Χρησιμοποιώντας τους παραπάνω μέσους συντελεστές. Καταλήγουμε λοιπόν στις ακόλουθες τιμές:

- **MLE**: 8.4%
- **PMLE**: 9.2%
- **Zelig**: 9.2%

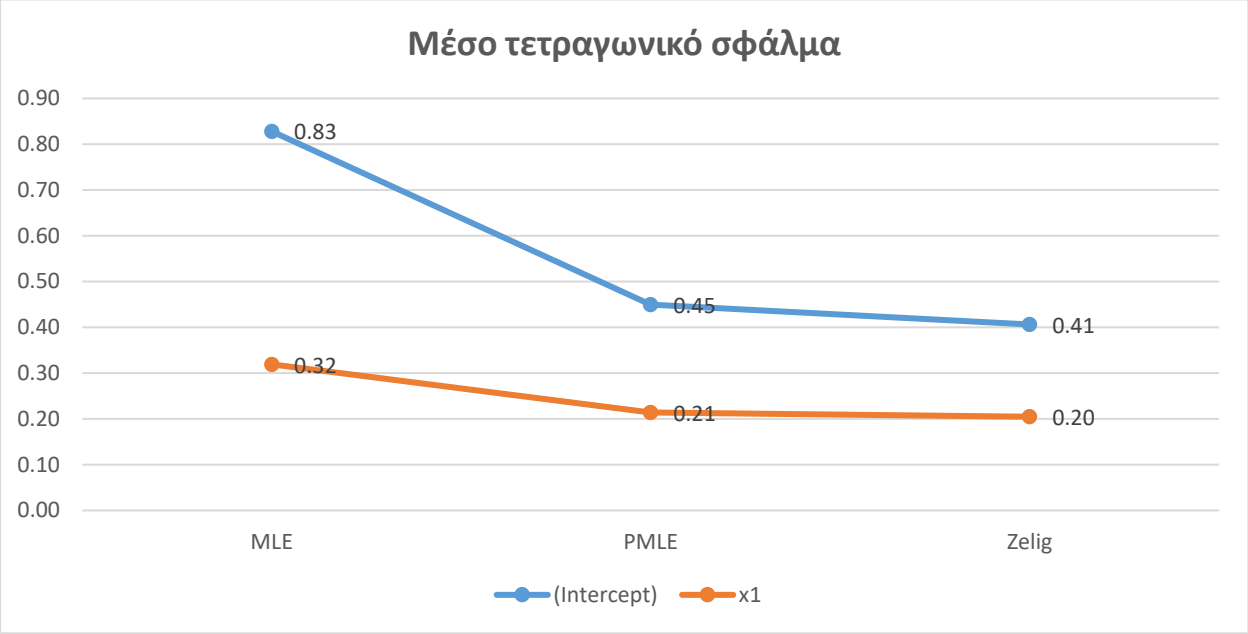
Παρατηρώντας λίγο τις τιμές αυτές μπορούμε να πούμε ότι η εκτίμηση με την κλασσική μέθοδο της μέγιστης πιθανοφάνειας όπως έχουμε αναφέρει υποεκτιμά την πιθανότητα εμφάνισης του γεγονότος ενώ οι άλλες δύο εκτιμήσεις είναι πιο κοντά στην πραγματική τιμή *(9.11%)*.

Στον επόμενο πίνακα βλέπουμε το μέσο τετραγωνικό σφάλμα για τους συντελεστές για καθεμιά απ' τις μεθόδους:

	ΜΤΣ	
	(Intercept)	**x1**
MLE	0.83	0.32
PMLE	0.45	0.21
Zelig	0.41	0.20

Πίνακας 5.9: Μέσο τετραγωνικό σφάλμα για τους συντελεστές

Θα παρουσιάσουμε τώρα σε γράφημα τα μέσα τετραγωνικά σφάλματα έτσι ώστε να γίνει το πιο εμφανές η διαφορά που υπάρχει ανάμεσα στις μεθόδους.



Γράφημα 5.4: Γράφημα με το Μέσο τετραγωνικό σφάλμα για κάθε μέθοδο

Εξετάζοντας τον πίνακα και το σχήμα καταλαβαίνουμε ότι οι εκτιμήσεις των δύο μεθόδων με διορθώσεις όπως είχαμε πει και στην προηγούμενη παράγραφο είναι καλύτερες από την κλασσική λογιστική παλινδρόμηση καθώς έχουν μικρότερο σφάλμα.

Ο κώδικας της παραπάνω διαδικασίας βρίσκεται στο παράρτημα Α.

ΚΕΦΑΛΑΙΟ 6

ΕΦΑΡΜΟΓΗ ΤΩΝ ΜΕΘΟΔΩΝ ΣΕ ΠΡΑΓΜΑΤΙΚΑ ΔΕΔΟΜΕΝΑ

Στο παρακάτω κεφάλαιο θα εφαρμόσουμε τις μεθόδους που έχουμε αναφέρει σε ένα σετ δεδομένων με πραγματικά στοιχεία με σκοπό να ελέγξουμε την ορθότητα τους.

6.1 Περιγραφή του συνόλου των δεδομένων μας

Το σύνολο δεδομένων που θα χρησιμοποιήσουμε αναφέρεται στον σακχαρώδη διαβήτη. Καλό θα είναι για αρχή να αναφέρουμε δύο λόγια για αυτήν την ασθένεια ώστε να γνωρίζουμε για τα δεδομένα που θα αναλύσουμε. Ο σακχαρώδης διαβήτης λοιπόν είναι μια μεταβολική ασθένεια η οποία χαρακτηρίζεται από αύξηση της συγκέντρωσης του σακχάρου στο αίμα και διαταραχή του μεταβολισμού της γλυκόζης, είτε ως αποτέλεσμα ελαττωμένης έκκρισης ινσουλίνης είτε λόγω ελάττωσης της ευαισθησίας των κυττάρων του σώματος στην ινσουλίνη. Οι κύριοι τύποι σακχαρώδους διαβήτη είναι

- ο διαβήτης τύπου 1,
- ο διαβήτης τύπου 2
- ο διαβήτης της κύησης.

Ο σακχαρώδης διαβήτης θεωρείται χρόνια ασθένεια και μπορεί να προκαλέσει μια σειρά σοβαρών επιπλοκών, όπως καρδιαγγειακή νόσο, χρόνια νεφρική ανεπάρκεια, βλάβες του αμφιβληστροειδούς, βλάβες των νεύρων, στυτική δυσλειτουργία κ.ά. Πρωτεύοντα ρόλο στη θεραπεία του σακχαρώδους διαβήτη παίζει η χορήγηση ινσουλίνης, οπού συνήθως χορηγείται ενέσιμη. Σύμφωνα με έρευνες το ποσοστό των ατόμων με διαβήτη στην Ελλάδα είναι περίπου 8%. Το σύνολο δεδομένων περιέχει στοιχεία από δύο έρευνες, την ATTICA και την CARDIO. Εμείς θα κρατήσουμε τα στοιχεία μόνο της πρώτης διότι η δεύτερη έχει λίγες παρατηρήσεις και σε πολλές από αυτές υπάρχουν ελλειπούσες τιμές. Το σετ δεδομένων λοιπόν εμπεριέχει ενενήντα έξι μεταβλητές, πολλές από τις οποίες βέβαια έχουν αρκετές ελλειπούσες τιμές όπου θα αναφέρουμε σε επόμενη παράγραφο τον τρόπο με τον οποίο θα εργαστούμε σχετικά με αυτές. Η δίτιμη εξαρτημένη που θέλουμε να εκτιμήσουμε είναι η ***dm***, η οποία έχει ως δυνατές τιμές τις εξής *"No diabetes", "diabetes "*. Ποιο αναλυτικά οι μεταβλητές που περιέχει το σετ δεδομένων μας είναι οι ακόλουθες:

code	dm	PROTEIN	total_meat	waist_transformed
entry	glucose	FAT	Tsoftdrinks	educat_transformed
age	TC	pc_mufa	Cola_drinks	year_smo_transformed
sex	hchol	pc_pufa	Soft_drinks	sbp_transformed
weight	TGL	pc_sfa	Soft_drinks_light	dbp_transformed
height	HDL_C	mufa_sfa	CVD_events	TC_transformed
bmi	LDL_C	MedDietScore	Time	TGL_transformed
waist	TAC	meddietgroup	HTN_incidence	HDL_C_transformed
hip	crp	tfish	HCHOL_incidence	LDL_C_transformed
educat	ZDRS	tnuts	DM_incidence	crp_transformed
family	STAI	tlegumes	ourea	MedDietScore_transformed
SES	fastfood	tdairy	creatinine	ourea_transformed
PhysAct_level	oliveoil	tfruit	creat_clearence	glucose_transformed
smoking_current	seedoil	tveggies	CCr_classes	
cig_day	butter	tpotatoe	bmi_coded	
year_smo	margarini	tcereals	bmi.coded_abbreviated	
ETS	coffee_drinking	tsweets	age_transformed	
htn	coffee_quant	tredmeat	weight_transformed	
sbp	energy	tegg	bmi_transformed	
dbp	CHO	tpoultry	hip_transformed	

Πίνακας 6.1: Οι μεταβλητές του σετ δεδομένων μας (KON.ATTICA Sakxaro1)

6.2 Επιλογή μεταβλητών και εκτέλεση των μοντέλων

Παρατηρούμε πως υπάρχουν στο τέλος πολλές μεταβλητές μετασχηματισμένες αλλά εμείς δεν θα ασχοληθούμε με αυτές. Για τις υπόλοιπες αρχικά θα τρέξουμε έναν κώδικα στο πακέτο R για να υπολογίσουμε τις ελλειπούσες τιμές. Αφού τον εκτελέσουμε θα πάρουμε τους παρακάτω πίνακες:

Variables	***NAs***
DataSet	0
code	0
entry	0
age	0
sex	0
family	0
PhysAct_level	0
smoking_current	0
dm	0
hchol	0
oliveoil	0
MedDietScore	0
meddietgroup	0

Variables	***NAs***
crp	108
HDL_C	111
TGL	123
creatinine	152
creat_clearence	156
CCr_classes	156
ourea	158
LDL_C	207
seedoil	217
HTN_incidence	309
DM_incidence	360
SES	390
ETS	393

Variables	***NAs***
tfruit	650
tveggies	650
tpotatoe	650
tcereals	650
tsweets	650
tegg	650
tpoultry	650
total_meat	650
tsoftdrinks	650
Cola_drinks	650
Soft_drinks	650
Soft_drinks_light	650
ZDRS	731

CVD_events	0	cig_day	459	STAI	737
Time	0	year_smo	465	TAC	817
weight	5	margarini	484	age_transformed	0
butter	5	HCHOL_incidence	607	weight_transformed	0
height	13	energy	608	bmi_transformed	0
bmi	13	fastfood	619	hip_transformed	0
educat	17	tredmeat	632	waist_transformed	0
TC	31	tfish	634	educat_transformed	0
bmi_coded	31	CHO	650	year_smo_transformed	0
bmi.coded_abbreviated	31	PROTEIN	650	sbp_transformed	0
glucose	32	FAT	650	dbp_transformed	0
coffee_drinking	32	pc_mufa	650	TC_transformed	0
coffee_quant	32	pc_pufa	650	TGL_transformed	0
waist	62	pc_sfa	650	HDL_C_transformed	0
hip	64	mufa_sfa	650	LDL_C_transformed	0
htn	68	tnuts	650	crp_transformed	0
sbp	74	tlegumes	650	MedDietScore_transformed	0
dbp	74	tdairy	650	ourea_transformed	0
				glucose_transformed	0

Πίνακας 6.2: Ο αριθμός των ελλειπούσων τιμών για κάθε μεταβλητή

Εμείς θα κρατήσουμε τις μεταβλητές που βρίσκονται στην πρώτη στήλη καθώς οι επόμενες δύο περιέχουν τις μετασχηματισμένες μεταβλητές και αυτές με μεγάλο αριθμό ελλειπούσων τιμών *(>100)*. Στην συνέχεια όπου υπάρχουν ελλειπούσες τιμές θα αφαιρέσουμε ολόκληρες τις γραμμές. Έτσι λοιπόν θα καταλήξουμε σε ένα σετ δεδομένων με 789 γραμμές και 31 στήλες. Οι στήλες *DataSet, code, entry* μας περιγράφουν το σετ δεδομένων μας και δεν είναι χρήσιμες στην ανάλυση. Επίσης η μεταβλητή *glucose* έρχεται μετά από εξέταση και υψηλές τιμές της μας δείχνει αν κάποιος πάσχει από σακχαρώδη διαβήτη. Δεν θα ασχοληθούμε και με αυτήν την μεταβλητή καθώς μας ενδιαφέρει αν κάποιες συνήθειες των ασθενών οδηγούν σε σακχαρώδη διαβήτη ή όχι.

Τώρα λοιπόν θα κάνουμε ελέγχους για το αν οι υπόλοιπες μεταβλητές επηρεάζουν την εξαρτημένη μεταβλητή"dm"**.** Για να ελέγξουμε τις μη διατάξιμες κατηγορικές μεταβλητές θα χρησιμοποιήσουμε τον έλεγχο X^2 του Pearson μέσω της συνάρτησης **chisq.test()** της R. Στην συνέχεια αν κάποια μεταβλητή αποδειχθεί σημαντική θα υπολογίσουμε τον συντελεστή Φ του Pearson για να ελέγξουμε την ένταση της συσχέτισης μέσω της συνάρτησης **Phi()** που υπάρχει διαθέσιμη στην βιβλιοθήκη DescTools. Βέβαια σε κάποιες μεταβλητές ο έλεγχος X^2 του Pearson ίσως να μην μας δίνει σωστά αποτελέσματα γ' αυτό σε αυτές τις περιπτώσεις θα χρησιμοποιήσουμε τον ακριβή έλεγχο του Fisher μέσω της συνάρτησης **fisher.test()**. Κάνοντας λοιπόν τους αντίστοιχούς ελέγχους καταλήγουμε στον παρακάτω πίνακα:

Variables	*Chi.Squared Pearson*	*Fisher's Test*	*Φ Pearson Coefs*
sex	0.01	-	0.09
family	-	0.00	0.23
PhysAct_level	-	0.57	0.04
smoking_current	0.04	-	0.08
htn	0.00	-	0.21
hchol	0.14	-	0.06
oliveoil	-	0.16	0.06
butter	0.60	-	0.02
coffee_drinking	0.50	-	0.03
CVD_events	-	0.00	0.21
coffee_quant	-	0.13	0.08

Πίνακας 6.3: Έλεγχος συσχέτισης των μη διατάξιμων κατηγορικών μεταβλητών με την εξαρτημένη μεταβλητή μας

Για να ελέγξουμε τις συνεχείς μεταβλητές θα χρησιμοποιήσουμε τον έλεγχο του Wilcoxon μέσω της συνάρτησης **wilcox.test()** της R. Στην συνέχεια αν κάποια μεταβλητή αποδειχθεί σημαντική θα χρησιμοποιήσουμε την συνάρτηση **biserial.cor()** που υπάρχει διαθέσιμη στην βιβλιοθήκη *ltm* για υπολογίσουμε την point-biserial συσχέτιση. Αυτή υπολογίζει την συσχέτιση ανάμεσα σε μια δίτιμη και μια συνεχής μεταβλητή. Κάνοντας λοιπόν τους αντίστοιχούς ελέγχους καταλήγουμε στον παρακάτω πίνακα:

Variables	Wilcoxon test	Biserial Coefs
age	0.000	-0.299
weight	0.000	-0.123
height	0.653	0.016
bmi	0.000	-0.156
hip	0.007	-0.084
educat	0.000	0.196
sbp	0.000	-0.259
dbp	0.000	-0.121
Time	0.259	0.020
TC	0.058	-0.072
MedDietScore	0.000	0.208

Πίνακας 6.4: Έλεγχος συσχέτισης των συνεχών μεταβλητών με την εξαρτημένη μεταβλητή μας

Για να ελέγξουμε τις διατάξιμες κατηγορικές μεταβλητές θα χρησιμοποιήσουμε τον *"linear-by-linear association"* έλεγχο μέσω της συνάρτησης **lbl_test()** που υπάρχει διαθέσιμη στην βιβλιοθήκη coin. Στην συνέχεια αν κάποια μεταβλητή αποδειχθεί σημαντική θα υπολογίσουμε τον συντελεστή rho του Spearman μέσω της συνάρτηση **cor.test()**. Κάνοντας λοιπόν τους αντίστοιχούς ελέγχους καταλήγουμε στον παρακάτω πίνακα:

Variables	Linear by Linear Association test	Spearman's rho
meddietgroup	0.00	-0.26
coffee_quant	0.81	0.00
bmi.coded_abbreviated	0.00	0.12
bmi_coded	0.16	0.07

Πίνακας 6.5: Έλεγχος συσχέτισης των διατάξιμων κατηγορικών μεταβλητών με την εξαρτημένη μεταβλητή μας

Στην συνέχεια θα σπάσουμε το σετ δεδομένων μας σε training και τεστ. Το πρώτο θα περιέχει 150 παρατηρήσεις έτσι ώστε να είναι ένα μικρό δείγμα με σπάνια γεγονότα και το δεύτερο θα περιέχει 50 παρατηρήσεις με σκοπό να ελέγξουμε τα αποτελέσματα των μοντέλων. Χρησιμοποιώντας λοιπόν το training σετ θα τρέξουμε το κλασσικό μοντέλο λογιστικής παλινδρόμησης και αυτό που πρότειναν οι King Gary και Langche Zeng (2001) και θα συγκρίνουμε τα αποτελέσματα. Ως ανεξάρτητες μεταβλητές θα κρατήσουμε τις δύο μεταβλητές με τη μεγαλύτερη συσχέτιση, οι οποίες είναι **“age”**, **“sbp”.**

Αρχικά θα ελέγξουμε το ποσοστό των γεγονότων στο training σετ μας. Το ποσοστό αυτό είναι 2.67% πράγμα το οποίο σημαίνει ότι έχουμε δεδομένα με σπάνια γεγονότα και μπορούμε να εργαστούμε με τις μεθόδους που έχουμε αναφέρει. Τρέχοντας τώρα λοιπόν την απλή λογιστική παλινδρόμηση παίρνουμε το παρακάτω output:

```
glm(formula = dm ~ sbp + age, family = "binomial", data = dataset,
   maxit = 50)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.80643  -0.20966  -0.13328  -0.08828   2.86179

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.34841    3.83779  -2.436   0.0149 *
sbp             0.00681    0.02992   0.228   0.8199
age             0.09330    0.04499   2.074   0.0381 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 36.887  on 149  degrees of freedom
Residual deviance: 30.114  on 147  degrees of freedom
AIC: 36.114

Number of Fisher Scoring iterations: 7
```

Στην συνέχεια αφού τρέξουμε το μοντέλο των King Gary και Langche Zeng θα πάρουμε το ακόλουθο output:

```
Call:
z5$zelig(formula = dm ~ sbp + age, maxit = 50, tau = c(0.08),
   bias.correct = TRUE, data = dataset)

Deviance Residuals:
   Min      1Q   Median      3Q     Max
-1.7387  -0.6907  -0.4509  -0.3510   1.8348

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -7.24611    3.83779  -1.888    0.059
sbp              0.02284    0.02992   0.763    0.445
age              0.05956    0.04499   1.324    0.186

(Dispersion parameter for binomial family taken to be 1)

   Null deviance: 36.887  on 149  degrees of freedom
Residual deviance: 30.114  on 147  degrees of freedom
AIC: 36.114

Number of Fisher Scoring iterations: 7

Next step: Use 'setx' method
```

Θα πρέπει εδώ να αναφέρουμε πώς στο παραπάνω μοντέλο κάναμε διόρθωση μεροληψίας και συμπληρώσαμε με 0,08 την τιμή tau που είναι το πραγματικό ποσοστό των γεγονότων στο

πληθυσμό στην Ελλάδα. Αυτήν την τιμή την γνωρίζουμε μέσα από έρευνες όπως αναφέραμε και στον πρόλογο της παραγράφου.

Αφού υπολογίσαμε τα coefficients τώρα θα ελέγξουμε μέσω του τεστ σετ την ακρίβεια των εκτιμήσεων. Αφού τρέξουμε τους αντίστοιχους κώδικες καταλήγουμε στους παρακάτω δύο πίνακες:

GLM Model

	PreddictValues	
actual_values	FALSE	TRUE
0	47	0
1	3	0

Zelig Model

	PreddictValues	
actual_values	FALSE	TRUE
0	46	1
1	0	3

Πίνακας 6.6: Πίνακες ταξινομήσεων του τεστ σετ GLM Model vs Zelig Model

Όπου παρατηρούμε πως το απλό λογιστικό μοντέλο δεν μπορεί να κάνει διαχωρισμό και προβλέπει όλες τις παρατηρήσεις ως *"FALSE"* (δηλαδή 0) ενώ τρεις από αυτές είναι *"TRUE"* (δηλαδή 1). Το αντίστοιχο Zelig μοντέλο προβλέπει καλύτερα την εξαρτημένη μεταβλητή μας *("dm")* καθώς προβλέπει σωστά τα 3 *"TRUE"* και κάνει μόνο μια λάθος προβλέψει σε ένα *"FALSE"*. Εν κατακλείδι είδαμε και εδώ πώς το μοντέλο με τις διορθώσεις εκτιμά καλύτερα όταν έχουμε μικρό δείγμα με σπάνια γεγονότα ξεπερνώντας το πρόβλημα της υποεκτίμησης των σπάνιων γεγονότων που υπάρχει στην απλή λογιστική παλινδρόμηση.

Ο κώδικας της παραπάνω διαδικασίας βρίσκεται στο παράρτημα Β.

Κ Ε Φ Α Λ Α Ι Ο 7

ΣΥΖΗΤΗΣΗ ΣΥΜΠΕΡΑΣΜΑΤΩΝ

Όπως έχει ήδη αναφερθεί, ο κύριος στόχος αυτής της εργασίας είναι η αναφορά του στατιστικού προβλήματος της εκτίμησης των σπάνιων γεγονότων με την κλασσική λογιστική παλινδρόμηση και τρόποι αντιμετώπισης του. Για την επίλυση λοιπόν αυτού του προβλήματος καταγράψαμε αναλυτικά τις διορθώσεις που προτείναν οι King Gary και Langche Zeng (2001) καθώς επίσης και την διόρθωση που πρότεινε ο Firth (1993). Και οι δύο αυτές μέθοδοι απαιτούν τον υπολογισμό του κλασσικού εκτιμητή μέγιστης πιθανοφάνειας και στην συνέχεια χρησιμοποιούν διορθώσεις πάνω σε αυτόν. Γενικά οι μέθοδοι αυτοί είναι εύκολο να εφαρμοστούν και προσφέρουν σημαντικές βελτιώσεις. Ως μόνο αρνητικό για την μέθοδο των King Gary και Langche Zeng θα μπορούσαμε να αναφέρουμε πως είναι απαραίτητη η γνώση του ποσοστού των γεγονότων του συνολικού πληθυσμού. Όπως είδαμε στο κεφάλαιο 5 μέσω των προσομοιώσεων οι δύο αυτές μέθοδοί έχουν μικρότερο σφάλμα σε σχέση με την κλασσική λογιστική παλινδρόμηση πράγμα που τις κάνει πιο αποτελεσματικές. Ακόμη θα πρέπει να αναφέρουμε πως και στην εφαρμογή που κάναμε σε πραγματικά δεδομένα οι εκτιμήσεις βελτιώθηκαν αισθητά. Όπως είδαμε η κλασσική λογιστική παλινδρόμηση μας ταξινομούσε όλες τις παρατηρήσεις του τεστ σετ δεδομένων ως “0” όπως φαίνεται και στον πίνακα 6.6. Αντίθετα οι μέθοδος των King Gary και Langche Zeng που χρησιμοποιήσαμε ταξινομεί ακριβέστερα τις παρατηρήσεις και έχουμε μεγαλύτερο αριθμό σωστών ταξινομήσεων (βλέπε πίνακα 6.6). Κρίνοντας από όλα τα παραπάνω θα μπορούσαμε να προτείνουμε ανεπιφύλακτα τις μεθόδους σε όποιον αναλυτή είχε να αντιμετωπίσει δεδομένα με σπάνια γεγονότα.

ΠΑΡΑΡΤΗΜΑΤΑ

Παράρτημα Α: Κώδικας για την σύγκριση των μεθόδων μέσω προσομοιώσεων

```
library(data.table)
library(logistf)
library(Zelig)
number.of.obs = 250         # Set the number of observation you want
#########################################################################
mse_intercept = function(intercept){
  thb =((intercept-(-4.3))^2)
  print(sum(thb)/1000)
}
mse_x1 = function(x1_model){
  thb2 =((x1_model-2)^2)
  print(sum(thb2) /1000)
}
p.hat = function(b_0,b_1,x) {
  pHAT = (exp(b_0+b_1*x)/(1+exp(b_0+b_1*x)))
  return(pHAT)
}
#########################################################################
coefs = NULL
Std_Error_MLE = NULL
Phat_MLE = NULL
coefs_PMLE = NULL
Std_Error_PMLE = NULL
Phat_PMLE = NULL
coefs_Zelig = NULL
Std_Error_Zelig = NULL
Phat_Zelig = NULL
```

```
for(i in 1:1000) {
n.obs = number.of.obs
b0 = -4.3
b1 = 2
x1 = rnorm(n.obs)          # create continuous variable
z = b0 + b1*x1             # linear combination
pr = 1/(1+exp(-z))
y = rbinom(length(x1),1,pr)
df = data.frame(y=y,x1=x1)

##########  MLE
model1 = glm( y~x1,data=df,family="binomial",maxit = 50)
###### Calculate the Phat
phat0_MLE = p.hat(model1$coefficients[1],model1$coefficients[2],1)
Phat_MLE = rbind(Phat_MLE,phat0_MLE)

## Get the outputs ###
coefs = rbind(coefs,model1$coefficients)
sd_mle = summary(model1)$coefficients[, 2]
Std_Error_MLE = rbind(Std_Error_MLE,sd_mle)

##########  PMLE
fit_PMLE = logistf(y~x1, data=df, pl = TRUE, firth = TRUE)

###### Calculate the Phat ######
phat0_PMLE = p.hat(fit_PMLE$coefficients[1],fit_PMLE$coefficients[2],1)
Phat_PMLE = rbind(Phat_PMLE,phat0_PMLE)

### Get the outputs ###
coefs_PMLE = rbind(coefs_PMLE,fit_PMLE$coefficients)
Std_Error_PMLE = rbind(Std_Error_PMLE,sqrt(diag(vcov(fit_PMLE))))

##########  Zelig
```

```
  fit.zelig = zelig(y~x1, model = "relogit", tau = NULL,
                    bias.correct = TRUE,
                    data = df, maxit = 50)

###### Calculate the Phat ######
phat0_Zelig = p.hat( coefficients(fit.zelig)[1], coefficients(fit.zelig)[2],1)
Phat_Zelig = rbind(Phat_Zelig,phat0_Zelig)

### Get the outputs ####
  coefs_Zelig = rbind(coefs_Zelig,coefficients(fit.zelig))
  Std_Error_Zelig = rbind(Std_Error_Zelig,get_se(fit.zelig))
}

#### Export the outputs from MLE

Phat_MLE = as.data.table(Phat_MLE)
results_MLE = as.data.table(coefs)
Std_Error_MLE = as.data.table(Std_Error_MLE)

############## Calculate the MSE and export it ##############
mse_intercept_MLE = mse_intercept(results_MLE$`(Intercept)`)
mse_x1_MLE = mse_x1(results_MLE$x1)

mse_MLE = cbind(mse_intercept_MLE,mse_x1_MLE)
mse_MLE = as.data.frame(mse_MLE)

fwrite(mse_MLE,"mse_MLE.csv")
fwrite(Phat_MLE,"Phat_MLE.csv")
fwrite(Std_Error_MLE,"Std_Error_MLE.csv")
fwrite(results_MLE,"Coefs_MLE.csv")

#### Export the outputs from PMLE
```

```
Phat_PMLE = as.data.table(Phat_PMLE)
results_PMLE = as.data.table(coefs_PMLE)
Std_Error_PMLE = as.data.table(Std_Error_PMLE)

############## Calculate the MSE and export it ##############
mse_intercept_PMLE = mse_intercept(results_PMLE$`(Intercept)`)
mse_x1_PMLE = mse_x1(results_PMLE$x1)

mse_PMLE = cbind(mse_intercept_PMLE,mse_x1_PMLE)
mse_PMLE = as.data.frame(mse_PMLE)

fwrite(mse_PMLE,"mse_PMLE.csv")
fwrite(Phat_PMLE,"Phat_PMLE.csv")
fwrite(results_PMLE,"Coefs_PMLE.csv")
fwrite(Std_Error_PMLE,"Std_Error_PMLE.csv")

#### Export the outputs from Zelig

Phat_Zelig=as.data.table(Phat_Zelig)
results_Zelig=as.data.table(coefs_Zelig)
Std_Error_Zelig=as.data.table(Std_Error_Zelig)

############## Calculate the MSE and export it ##############
mse_intercept_Zelig = mse_intercept(results_Zelig$`(Intercept)`)
mse_x1_Zelig = mse_x1(results_Zelig$x1)

mse_Zelig = cbind(mse_intercept_Zelig,mse_x1_Zelig)
mse_Zelig = as.data.frame(mse_Zelig)

fwrite(mse_Zelig,"mse_Zelig.csv")
fwrite(Phat_Zelig,"Phat_Zelig.csv")
fwrite(results_Zelig,"Coefs_Zelig.csv")
fwrite(Std_Error_Zelig,"Std_Error_Zelig.csv")
```

Παράρτημα Β: Κώδικας για την εφαρμογή σε πραγματικά δεδομένα

```
library(data.table)
library(logistf)
library(Zelig)
#############################################################################
df=fread("KON.ATTICA Sakxaro1.txt", na.strings = "")
df <- subset(df, select = -c(DataSet))
df <- subset(df, select = -c(code))
df <- subset(df, select = -c(entry ))
#############################################################################
sapply(df, function(x) sum(is.na(x)))  ### count the NAs from all columns
df = na.exclude(df)
df$sex = as.factor(df$sex)
df$family = as.factor(df$family)
df$PhysAct_level = as.factor(df$PhysAct_level)
df$smoking_current = as.factor(df$smoking_current)
df$htn = as.factor(df$htn)
df$hchol = as.factor(df$hchol)
df$oliveoil = as.factor(df$oliveoil)
df$butter = as.factor(df$butter)
df$coffee_drinking = as.factor(df$coffee_drinking)
df$CVD_events = as.factor(df$CVD_events)

df$coffee_quant = factor(df$coffee_quant, order = TRUE)
df$meddietgroup = factor(df$meddietgroup, order = TRUE)
df$bmi.coded_abbreviated = factor(df$bmi.coded_abbreviated, order = TRUE)
df$bmi_coded = factor(df$bmi_coded, order = TRUE)

mean(df$dm)
######################### Testing the association between the Variables
chisq.test(df$dm,df$sex)
chisq.test(df$dm,df$family)
```

```
chisq.test(df$dm,df$PhysAct_level)
chisq.test(df$dm,df$smoking_current)
chisq.test(df$dm,df$htn)
chisq.test(df$dm,df$hchol)
chisq.test(df$dm,df$oliveoil)
chisq.test(df$dm,df$butter)
chisq.test(df$dm,df$coffee_drinking)
chisq.test(df$dm,df$CVD_events)

######## Check the intensity of correlation for all the non-ordinal categorical variables
library(DescTools)
### We should Run the below rows for all the non-ordinal categorical variables
t=table(df$dm,df$coffee_quant)
Phi(t, digits = 2)

#### Excact Fisher's Test
fisher.test(df$dm,df$oliveoil)
fisher.test(df$dm,df$family)
fisher.test(df$dm,df$PhysAct_level)
fisher.test(df$dm,df$coffee_quant)
fisher.test(df$dm,df$CVD_events)

#########################################################################
wilcox.test(df$age ~ df$dm, alternative = "two.sided")
wilcox.test(df$weight ~ df$dm, alternative = "two.sided")
wilcox.test(df$height ~ df$dm, alternative = "two.sided")
wilcox.test(df$bmi ~ df$dm, alternative = "two.sided")
wilcox.test(df$hip ~ df$dm, alternative = "two.sided")
wilcox.test(df$educat ~ df$dm, alternative = "two.sided")
wilcox.test(df$sbp ~ df$dm, alternative = "two.sided")
wilcox.test(df$dbp ~ df$dm, alternative = "two.sided")
wilcox.test(df$Time ~ df$dm, alternative = "two.sided")
wilcox.test(df$TC ~ df$dm, alternative = "two.sided")
```

```
wilcox.test(df$MedDietScore ~ df$dm, alternative = "two.sided")

######## ######## Check the intensity of correlation for all the continues variables
library(ltm)
### We should Run the below row for all the continues variables
biserial.cor(df$MedDietScore,df$dm)

############### Linear by Linear Association ###############
library(coin)
dm_with_meddietgroup = table(df$meddietgroup,df$dm)
lbl_test(dm_with_meddietgroup)

dm_with_coffee_quant = table(df$coffee_quant,df$dm)
lbl_test(dm_with_coffee_quant)

dm_with_bmi.coded_abbreviated = table(df$bmi.coded_abbreviated,df$dm)
lbl_test(dm_with_bmi.coded_abbreviated)

dm_with_bmi_coded = table(df$bmi_coded,df$dm)
lbl_test(dm_with_bmi_coded)

### We should Run the below rows for all the ordinal categorical variables
cor.test(df$dm, as.numeric(df$bmi_coded),
        method = c("spearman"),
        exact = NULL, conf.level = 0.95)

#############################################################################
################# Create a train and a test set
set.seed(3062)
train_ind = sample(seq_len(nrow(df)), size = 150)
train = df[train_ind, ]
test = df[-train_ind, ]
testF_id = sample(seq_len(nrow(test)), size = 50)
```

```
testF = df[testF_id, ]
mean(testF$dm)
test = testF
dataset = train
mean(dataset$dm)

#############################################################################
##
model1 = glm(dm~sbp+age
         ,data=dataset,family="binomial",maxit = 50)

summary(model1)
aaa = predict(model1, test, type="response")

confusion_Matrix = table(actual_values=test$dm,PreddictValues=aaa>0.5)
print(confusion_Matrix)

#############################################################################
##
fit.zelig = zelig(dm~sbp+age
                  ,model = "relogit",
                  tau = c(0.08),
                  bias.correct = TRUE,
                  data = dataset, maxit = 50)
summary(fit.zelig)

aaa_Zelig = predict(fit.zelig, test, type="response")

confusion_Matrix_Zelig=table(actual_values=test$dm,PreddictValues=as.numeric(unlist(aaa
_Zelig)) >0.5)
print(confusion_Matrix_Zelig)
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

1. Αντζουλάκος, Δ. (2013), Ανάλυση δεδομένων με τη Χρήση Στατιστικών Πακέτων – Εισαγωγή στο R
2. Ηλιόπουλος, Γ. (2017), Γενικευμένα Γραμμικά Μοντέλα, πανεπιστημιακές σημειώσεις
3. Κούτρας, Μ. (2012), Πολυμεταβλητή Ανάλυση, πανεπιστημιακές σημειώσεις
4. Κούτρας, Μ. , Ευαγγελάρας Χ. (2010), Ανάλυση Παλινδρόμησης, εκδόσεις Αθ. Σταμούλης
5. Τζαβελάς, Γ. (2018), Βιοστατιστική και Στατιστικές Μέθοδοι στην Επιδημιολογία, πανεπιστημιακές σημειώσεις
6. Τζαβελάς, Γ. (2018), Γενικευμένα Γραμμικά Μοντέλα, πανεπιστημιακές σημειώσεις
7. Πολίτης, Κ. (2018), Βιοστατιστική και Στατιστικές Μέθοδοι στην Επιδημιολογία, πανεπιστημιακές σημειώσεις

Ξένη

1. Agresti A. (2007), An Introduction to Categorical Data Analysis, Wiley, New York
2. Alison Paul D. (1999), Logistic Regression Using the SAS System
3. Cox, D.R. and Snell, E.J. , (1989) Analysis of Binary Data (2nd ed.) Chapman and Hall
4. Efron, B. (1978), Regression and ANOVA with zero-one data: measures of residualvariation, Journal of the American Statistical Association 73, 113-121
5. Firth, D. (1993), Bias Reduction of Maximum Likelihood Estimates. Biometrika 80 (1): 27–38
6. Hosmer David W., Jr., Lemeshow S. (2001), Applied Logistic Regression
7. King G., Zeng L. (2001), Logistic Regression in Rare Events Data
8. Levy Paul S., Lemeshow S. (2013), Sampling of Populations: Methods and Applications
9. Heinz L. (University of Linz, Austria), The Problem of Modeling Rare Events in ML-based Logistic Regression

10. McFadden, D. (1974), Conditional logit analysis of qualitative choice behaviour, in: P.Zarembka (ed.), Frontiers in Econometrics, Academic Press, New York, 105-142
11. McCullagh P. , Nelder John A. (1989), Generalized Linear Models, Second Edition
12. Nagelkerke, N.J.D. (1991) A note on a general definition of the coefficient of determination. Biometrika 78: 691-692
13. Westphal C. (2013), Logistic Regression for Extremely Rare Events: The Case of School Shootings
14. Youden, W.J. (1950), Index for rating diagnostic tests - Cancer

Διαδικτυακές Ιστοσελίδες

1. https://en.wikipedia.org/wiki/Main_Page
2. https://stackoverflow.com/
3. https://github.com/
4. https://www.r-bloggers.com/example-8-15-firth-logistic-regression/
5. https://www.quora.com/