



ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ
ΤΟΜΕΑΣ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΕΠΙΧΕΙΡΗΣΙΑΚΗΣ ΕΡΕΥΝΑΣ

Θεόδωρος Εβρένογλου

ΜΕΙΩΣΗ ΤΗΣ ΜΕΡΟΛΗΨΙΑΣ ΤΟΥ ΕΚΤΙΜΗΤΗ ΜΕΓΙΣΤΗΣ
ΠΙΘΑΝΟΦΑΝΕΙΑΣ ΣΕ ΜΟΝΤΕΛΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Μεταπτυχιακή Διατριβή

Ιωάννινα, 2019

Η παρούσα Μεταπτυχιακή Διατριβή εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική και Επιχειρησιακή Έρευνα που απονέμει το Τμήμα Μαθηματικών του Πανεπιστημίου Ιωαννίνων.

Εγκρίθηκε την 20/05/2019 από την εξεταστική επιτροπή:

Ονοματεπώνυμο	Βαθμίδα
Δημήτριος Μαυρίδης	Επίκουρος Καθηγητής
Σωτήριος Λουκάς	Καθηγητής
Απόστολος Μπατσίδης	Επίκουρος Καθηγητής

ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ

“Δηλώνω υπεύθυνα ότι η παρούσα διατριβή εκπονήθηκε κάτω από τους διεθνείς ηθικούς και ακαδημαϊκούς κανόνες δεοντολογίας και προστασίας της πνευματικής ιδιοκτησίας. Σύμφωνα με τους κανόνες αυτούς, δεν έχω προβεί σε ιδιοποίηση ξένου επιστημονικού έργου και έχω πλήρως αναφέρει τις πηγές που χρησιμοποίησα στην εργασία αυτή.”

Θεόδωρος Εβρένογλου

Στη μνήμη του πατέρα μου και του παππού μου

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα πρωτίστως να ευχαριστήσω τον κ. Δημήτριο Μαυρίδη ο οποίος με προθυμία ανέλαβε το ρόλο του επιβλέποντα της παρούσας διατριβής και στάθηκε ουσιαστικός αρωγός με τις συμβουλές και την καθοδήγησή του κατά την διάρκεια της συγγραφής της. Παράλληλα ευχαριστώ τον κ. Σωτήριο Λουκά και τον κ. Απόστολο Μπατσίδα που δέχθηκαν να βρίσκονται στην τριμελή επιτροπή αξιολόγησης της διατριβής. Επιπλέον, ευχαριστώ όλα τα μέλη του Τομέα Στατιστικής και Επιχειρησιακής Έρευνας του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων για την πολύτιμη καθοδήγηση και τις γνώσεις που πρόθυμα μου έδωσαν τόσο στα προπτυχιακά όσο και στα μεταπτυχιακά μου χρόνια. Επίσης, θα ήθελα να ευχαριστήσω τους φίλους και συναδέλφους μου μεταπτυχιακούς φοιτητές Γεώργιο Καρακατσούλη, Βασίλειο Μουλόπουλο και Αθανάσιο Καραμανίδα για τις συμβουλές που μου πρόσφεραν κατά τη διάρκεια της συγγραφής της εργασίας. Τέλος, εκφράζω ένα μεγάλο ευχαριστώ σε όλη την οικογένειά μου και ιδιαίτερα στην μητέρα μου Μαρίνα και τον αδερφό μου Γιάννη. Χωρίς τη στήριξή τους δεν θα είχα καταφέρει να φτάσω ως εδώ.

Περιεχόμενα

Σελίδα

1	Γενικευμένα Γραμμικά Μοντέλα	3
1.1	Το Γενικευμένο Γραμμικό Μοντέλο	4
1.2	Η βασική δομή του μοντέλου	6
1.3	Συναρτήσεις Σύνδεσης	9
1.4	Εκτίμηση των παραμέτρων	12
1.4.1	Η μέθοδος Newton-Raphson	15
1.4.2	Ο αλγόριθμος του Fisher	15
1.5	Εφαρμογή του αλγορίθμου του Fisher στη λογιστική παλινδρόμηση	18
2	Ανάλυση Επιβίωσης	23
2.1	Βασικές συναρτήσεις	24
2.2	Λογοκριμένα δεδομένα	30
2.3	Εκτίμηση της συνάρτησης επιβίωσης	34
2.4	Σύγκριση καμπυλών επιβίωσης	39
2.4.1	Logrank Test	42
2.5	Το μοντέλο παλινδρόμησης του Cox	44
2.5.1	Προσαρμογή του μοντέλου του Cox	47
2.6	Οι υποθέσεις του μοντέλου του Cox και ο έλεγχος τους	54

3 Μείωση της μεροληψίας	61
3.1 Τα σύμβολα $\mathcal{O}$, o , $\mathcal{O}_p$, o_p	63
3.2 Cumulants	67
3.3 Index Notation	70
3.4 Μέθοδος μέγιστης πιθανοφάνειας σε όρους index notation	74
3.5 Αντιμετώπιση της μεροληψίας του εκτιμητή μέγιστης πιθανοφάνειας	79
3.5.1 Διόρθωση της μεροληψίας	81
3.5.2 Η μέθοδος των Cordeiro & McCullagh	88
3.5.3 Μείωση της μεροληψίας	90
3.5.4 Η μέθοδος μείωσης της μεροληψίας του Firth	92
3.6 Το πρόβλημα των διαχωρισμένων δεδομένων	101
3.7 Εφαρμογή της μεθόδου του Firth	111
3.7.1 Εφαρμογή στη λογιστική παλινδρόμηση	111
3.7.2 Εφαρμογή στο μοντέλο του Cox	119
Βιβλιογραφία	136

ΠΡΟΛΟΓΟΣ

Η μέθοδος της μέγιστης πιθανοφάνειας αποτελεί ίσως την πιο διαδεδομένη μορφή εκτίμησης παραμέτρων σε μοντέλα παλινδρόμησης τα οποία συχνά χρησιμοποιούνται για την μοντελοποίηση προβλημάτων που βρίσκουν εφαρμογή στην καθημερινότητα. Τόσο το συγκεκριμένο μεθοδολογικό πλαίσιο που πρέπει να ακολουθηθεί για τον υπολογισμό του, όσο και ευνοϊκές ασυμπτωτικές του ιδιότητες δικαιολογούν τη δημοφιλία του συγκεκριμένου εκτιμητή. Ωστόσο, σε περιπτώσεις όπου δεν έχουμε στη διάθεσή μας κάποιο μεγάλο δείγμα η μέθοδος μπορεί να οδηγήσει σε παραπλανητικά συμπεράσματα. Ιδιαίτερα, σε τέτοιες περιπτώσεις ένα μεγάλο ποσό μεροληψίας εισάγεται στον εκτιμητή μέγιστης πιθανοφάνειας και έτσι βλάπτεται η ακρίβειά του. Η παρούσα εργασία οργανώνεται ως εξής:

- Στα κεφάλαια 1 & 2 γίνεται μια σύντομη εισαγωγή σε περιπτώσεις διάφορων μοντέλων που στόχο έχουν να μοντελοποιήσουν πραγματικά προβλήματα. Πιο συγκεκριμένα στο κεφάλαιο 1 δίνεται μια σύντομη εισαγωγή στα γενικευμένα γραμμικά μοντέλα (generalized linear models) τα οποία προκύπτουν ως επέκταση της κλασικής θεωρίας του γραμμικού μοντέλου. Στο

κεφάλαιο 2 γίνεται εισαγωγή στα μοντέλα ανάλυσης επιβίωσης τα οποία επεκτείνουν την θεωρία των γενικευμένων γραμμικών μοντέλων προκειμένου να επιτευχθεί η μοντελοποίηση προβλημάτων στα οποία παρουσιάζονται λογοκριμένοι χρόνοι επιβίωσης. Το μοντέλο στο οποίο δίνεται ιδιαίτερη βαρύτητα είναι το μοντέλο ανάλογων κινδύνων του Cox.

- Στο 3ο κεφάλαιο της εργασίας παρουσιάζεται μια ανασκόπηση των μεθόδων για την μείωση της μεροληψίας του εκτιμητή μέγιστης πιθανοφάνειας. Οι μέθοδοι αυτές είναι εξαιρετικά σημαντικές αφού σε περιπτώσεις μοντέλων που επιχειρούν να μοντελοποιήσουν πραγματικά προβλήματα και πιο συγκεκριμένα βιοϊατρικά προβλήματα, η βελτιστοποίηση της ακρίβειας των αποτελεσμάτων αποκτά μια επιπλέον βαρύτητα. Εκτός, από τις μεθόδους ιδιαίτερη σημασία δίνεται και στην παρουσίαση του μαθηματικού υπόβαθρου που χρειάζεται για την κατανόηση των μεθόδων. Η κεντρική μέθοδος που παρουσιάζεται είναι αυτή που πρότεινε ο Firth (Firth,1993) και η οποία λόγω των ιδιοτήτων του νέου εκτιμητή έχει αποκτήσει ιδιαίτερη δημοφιλία σε σχέση με τις υπόλοιπες μεθόδους. Τέλος, προκειμένου να γίνει πιο ξεκάθαρη η δραστηριότητα και η αποτελεσματικότητα αυτής της μεθόδου δίνονται τα αποτελέσματα από τις αναλύσεις δύο πραγματικών συνόλων δεδομένων.

Κεφάλαιο 1

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Οι πιθανότητες και η στατιστική αφορούν κυρίως τη μελέτη της αβεβαιότητας. Τα στατιστικά μοντέλα μας βοηθούν να κατανοήσουμε μια τυχαία διαδικασία για την οποία έχουν συλλεχθεί κάποιες παρατηρήσεις. Η κατανόηση της τυχαίας διαδικασίας είναι από μόνη της κάτι το ενδιαφέρον, επιπλέον όμως μέσω ενός μοντέλου μας επιτρέπεται να κάνουμε προβλέψεις και ίσως, το πιο σημαντικό, να λαμβάνουμε αποφάσεις. Επιπλέον, πρέπει να τονιστεί πως στο πλαίσιο της διαδικασίας μοντελοποίησης τα συμπεράσματά μας βασίζονται μόνο σε ένα, ενδεχομένως μικρό, δείγμα παρατηρήσεων και επομένως υπόκεινται σε σφάλματα. Για την συγγραφή αυτού του κεφαλαίου βασιστήκαμε κυρίως στα βιβλία των Agresti (2015, 2003), McCullagh and Nelder (1989), ενώ για κάποια επιπρόσθετα στοιχεία χρησιμοποιήσαμε τα βιβλία των Ntzoufras (2011) και Keith Knight (1999). Για περισσότερες λεπτομέρειες σχετικά με αυτά τα βιβλία παραπέμπουμε στο παράρτημα των αναφορών που βρίσκεται στο τέλος της μεταπτυχιακής διατριβής.

1.1 Το Γενικευμένο Γραμμικό Μοντέλο

Έστω $Y_1, Y_2, \dots, Y_n$, οι ανεξάρτητες παρατηρήσεις μίας τυχαίας μεταβλητής Y που ονομάζουμε μεταβλητή απόκρισης ή αλλιώς εξαρτημένη μεταβλητή (dependent variable). Όπως γνωρίζουμε το γραμμικό μοντέλο δίνεται από τη σχέση:

$$Y = X\beta + \epsilon \quad (1.1)$$

και γράφεται με τη μορφή πινάκων με τον εξής τρόπο

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

όπου $Y = (Y_1, Y_2, \dots, Y_n)^T$ είναι η στήλη των παρατηρήσεων της εξαρτημένης μεταβλητής (dependent variable), ο πίνακας X είναι ο $n \times p$ πίνακας των ανεξάρτητων μεταβλητών $X_1, X_2, \dots, X_p$. Ο πίνακας αυτός από αυτό το σημείο και για όλη την υπόλοιπη εργασία θα υποτίθεται ως ένας πίνακας πλήρους βαθμίδας (full rank). Συμβολίζουμε ως $\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$ τη στήλη που περιέχει τις παραμέτρους του γραμμικού μοντέλου (regression parameters) που θεωρούνται άγνωστες και πρέπει να εκτιμηθούν, ενώ τέλος με $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$ συμβολίζουμε τη στήλη των τυχαίων σφαλμάτων (random errors) στα οποία υπόκειται το μοντέλο μας. Τα τυχαία σφάλματα παίζουν καθοριστικό ρόλο στο κανονικό γραμμικό μοντέλο και πρέπει να πληρούν τη βασική υπόθεση πως:

$$\epsilon \sim N_n(0, \sigma^2 I_n).$$

η οποία υποδηλώνει τα εξής:

$$E(\epsilon_i) = 0, \forall i$$

$$Var(\epsilon_i) = \sigma^2, \forall i$$

$$Cov(\epsilon_i, \epsilon_j) = 0, \forall i, j \text{ με } i \neq j.$$

Από τις υποθέσεις για τα τυχαία σφάλματα αλλά και την αρχική μορφή του μοντέλου (1.1) εύκολα προκύπτει ότι: $E[Y|X] = X\beta$. Από αυτή τη σχέση εξάγουμε το συμπέρασμα πως η μέση τιμή (mean value) της εξαρτημένης μεταβλητής Y συνδέεται γραμμικά με τις επεξηγηματικές μεταβλητές.

Πολλές φορές όμως πρέπει να αντιμετωπίσουμε προβλήματα στα οποία οι υποθέσεις που παρουσιάστηκαν προηγουμένως δεν είναι εφικτό να πληρούνται. Μια περίπτωση στην οποία οι υποθέσεις του προηγούμενου μοντέλου δεν είναι ρεαλιστικές είναι η περίπτωση στην οποία η εξαρτημένη μεταβλητή είναι μια διακριτή μεταβλητή η οποία, για παράδειγμα, μοντελοποιεί την εμφάνιση ή όχι μιας ασθένειας. Μια άλλη περίπτωση είναι εκείνη στην οποία η εξαρτημένη μεταβλητή Y είναι συνεχής αλλά μοντελοποιεί χρόνο, τότε και πάλι η υπόθεση της κανονικότητας είναι πολύ πιθανό να μην ικανοποιείται. Παρόλο, που οι υποθέσεις δεν πληρούνται και η θεωρία του κανονικού γραμμικού μοντέλου δεν μπορεί να εφαρμοστεί, η ανάγκη για την κατασκευή ενός μοντέλου παραμένει διότι η εξάρτηση μίας μεταβλητής από κάποιες άλλες επεξηγηματικές μπορεί να υφίσταται.

Το πρόβλημα το οποίο παρουσιάστηκε προηγουμένως απασχόλησε την επιστημονική κοινότητα για χρόνια ώσπου το 1972 οι Nelder και Wedderburn στην εργασία τους με τίτλο "Generalized Linear Models" (1972) γενίκευσαν την ιδέα

του κανονικού γραμμικού μοντέλου με τη βοήθεια μίας ευρύτερης οικογένειας κατανομών η οποία ονομάζεται Εκθετική Οικογένεια Κατανομών.

1.2 Η βασική δομή του μοντέλου

Ορισμός 1.2.1. Λέμε ότι η κατανομή μιας τυχαίας μεταβλητής Y , με ανεξάρτητες παρατηρήσεις $(y_1, y_2, \dots, y_n)$ ανήκει στην εκθετική οικογένεια κατανομών με παράμετρο διασποράς (exponential dispersion family of distributions) αν η συνάρτηση πυκνότητας πιθανότητας, ή η συνάρτηση πιθανότητας, γράφεται στη μορφή

$$f_Y(y_i; \theta_i, \phi) = \exp \left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right), \quad i = 1, 2, \dots, n \quad (1.2)$$

όπου το θ_i ονομάζεται φυσική ή κανονική παράμετρος θέσης (natural or canonical location parameter), η ϕ είναι παράμετρος κλίμακας ή μεταβλητοτητας (scale or dispersion parameter). Συχνά έχουμε ότι $a(\phi) = 1$ και $c(y_i, \phi) = c(y_i)$, σε αυτή την περίπτωση προκύπτει η φυσική εκθετική οικογένεια κατανομών (natural exponential family of distributions) που έχει τη μορφή

$$f_Y(y_i; \theta_i) = h(y_i) \exp(y_i \theta_i - b(\theta_i)), \quad i = 1, 2, \dots, n.$$

Τέλος, τα $b(\theta_i)$ και $c(y_i, \phi)$ είναι γνωστές συναρτήσεις, ενώ σε όλα τα μοντέλα που θα δούμε παρακάτω ισχύει ότι $a(\phi) = \frac{\phi}{w_i}$, για $\phi > 0$, όπου τα w_i είναι γνωστά εκ των προτέρων βάρη, συνήθως ίσα με 1.

Θεώρημα 1.2.2. Έστω Y μια τυχαία μεταβλητή με την κατανομή της Y να ανήκει στην εκθετική οικογένεια κατανομών, δηλαδή η συνάρτηση πυκνότητας πιθανότητας ή η συνάρτηση πιθανότητας είναι της μορφής (1.2) τότε ισχύουν τα

παρακάτω:

$$1. E(y_i) = b'(\theta_i) \quad (1.3)$$

$$2. Var(y_i) = b''(\theta_i)a(\phi_i) \quad (1.4)$$

Απόδειξη. Η απόδειξη θα γίνει για τη συνεχή περίπτωση και υπό την ισχύ των συνθηκών ομαλότητας (regularity conditions) που αναφέρονται στο κεφάλαιο 3 της εργασίας.

Αφού, η f είναι συνάρτηση πυκνότητας πιθανότητας θα ισχύει ότι

$$\int f_Y(y_i; \theta_i, \phi) dy_i = 1$$

όπου η ολοκλήρωση γίνεται υπεράνω του διαστήματος που ορίζεται η συνάρτηση. Παραγωγίζοντας την προηγούμενη σχέση δύο φορές παίρνουμε τα εξής αποτελέσματα:

$$\frac{d}{d\theta_i} \int f_Y(y_i; \theta_i, \phi) dy_i = \frac{d}{d\theta_i} 1 \Rightarrow \int \frac{d}{d\theta_i} (f_Y(y_i; \theta_i, \phi)) dy_i = 0 \quad (1.5)$$

$$\Rightarrow \int \frac{d^2}{d\theta_i^2} f_Y(y_i; \theta_i, \phi) dy_i = 0 \quad (1.6)$$

Όμως η f έχει τη μορφή (1.2), επομένως και λόγω της (1.5) έχουμε ότι:

$$\begin{aligned} & \int \frac{d}{d\theta_i} \left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right) f_Y(y_i; \theta_i, \phi) dy_i = 0 \\ & \Rightarrow \int \left(\frac{y_i - b'(\theta_i)}{a(\phi)} \right) f_Y(y_i; \theta_i, \phi) dy_i = 0 \quad (*) \\ & \Rightarrow E(y_i) - b'(\theta_i) = 0 \Rightarrow E(y_i) = b'(\theta_i) \end{aligned}$$

Παραγωγίζοντας την σχέση (*) έχουμε ότι:

$$\begin{aligned} & \int \left[\frac{-b''(\theta_i)}{a(\phi)} f_Y(y_i; \theta_i, \phi) + \left(\frac{y_i - b'(\theta_i)}{a(\phi)} \right)^2 f_Y(y_i; \theta_i, \phi) \right] dy_i = 0 \\ & \Rightarrow \frac{b''(\theta_i)}{a(\phi)} = \frac{1}{(a(\phi))^2} Var(y_i) \Rightarrow Var(y_i) = b''(\theta_i)a(\phi) \end{aligned}$$

□

Για να μπορέσουμε να κάνουμε πιο απλή τη μετάβαση από το κανονικό γραμμικό μοντέλο στο γενικευμένο γραμμικό μοντέλο θα χωρίσουμε το κανονικό μοντέλο σε 3 μέρη και θα εξηγήσουμε πως αυτά γενικεύονται (βλ. Agresti, 2003). Τα μέρη αυτά είναι:

1. Το στοχαστικό μέρος (stochastic part) του κανονικού μοντέλου όπου η εξαρτημένη μεταβλητή $Y \sim N_n(X\beta, \sigma^2)$.
2. Το συστηματικό μέρος όπου βρίσκεται η γραμμική προβλέπουσα (linear predictor)

$$\eta = X\beta \quad (1.7)$$

3. Η σύνδεση στοχαστικού και συστηματικού μέρους, η οποία επιτυγχάνεται μέσω της μέσης τιμής της τυχαίας μεταβλητής Y , αφού

$$\mu_i = \eta_i = \sum_{j=1}^p \beta_j X_{ij}, \quad i = 1, 2, \dots, n \quad (1.8)$$

Οι προηγούμενες έννοιες για το κανονικό γραμμικό μοντέλο μπορούν να γενικευτούν ως εξής:

1. Στο στοχαστικό μέρος η τυχαία μεταβλητή Y ακολουθεί οποιαδήποτε κατανομή ανήκει στην εκθετική οικογένεια κατανομών.
2. Στο συστηματικό μέρος οι επεξηγηματικές μεταβλητές συνδέονται γραμμικά με μία συνάρτηση, $\eta(\cdot)$ που ονομάζεται γραμμική προβλέπουσα (linear predictor)

$$\eta_i = \sum_{j=1}^p \beta_j X_{ij}, \quad i = 1, 2, \dots, n \quad (1.9)$$

3. Η σύνδεση στοχαστικού και συστηματικού μέρους επιτυγχάνεται μέσω μιας συνεχώς διαφορίσιμης και μονότονης συνάρτησης της πληθυσμιακής μέσης τιμής μ , που ονομάζεται συνάρτηση σύνδεσης (link function) και συνδέει το στοχαστικό και το συστηματικό μέρος, ώστε αν $g(\cdot)$ είναι η συνάρτηση σύνδεσης θα ισχύει ότι:

$$g(\mu) = \eta = X\beta \quad (1.10)$$

Εκτός λοιπόν από τη γενίκευση στο κομμάτι της κατανομής της εξαρτημένης μεταβλητής Y η άλλη βασική διαφορά του κανονικού γραμμικού μοντέλου και του γενικευμένου γραμμικού μοντέλου βρίσκεται στον τρόπο σύνδεσης στοχαστικού και συστηματικού μέρους. Στο κανονικό γραμμικό μοντέλο το συστηματικό μέρος συνδέεται άμεσα με τη μέση τιμή, ενώ στο γενικευμένο γραμμικό μοντέλο το συστηματικό μέρος συνδέεται με τη μέση τιμή της Y μέσω μίας συνάρτησης της. Η συνάρτηση $g(\cdot)$ μέσω της οποίας η μέση τιμή της εξαρτημένης μεταβλητής Y συνδέεται με το συστηματικό μέρος του μοντέλου ονομάζεται **συνάρτηση σύνδεσης (link function)** και υποτίθεται ως μια συνεχώς διαφορίσιμη και 1-1 συνάρτηση (βλ. Ntzoufras, 2011).

1.3 Συναρτήσεις Σύνδεσης

Η συνάρτηση σύνδεσης (link function) είναι μια μονότονη και συνεχώς διαφορίσιμη συνάρτηση η οποία χρησιμοποιείται για να συνδέσει την γραμμική προβλέπουσα (linear predictor), η , με τη μέση τιμή της εξαρτημένης μεταβλητής Y . Συνήθως, δεν υπάρχει κάποιος περιορισμός στον τρόπο που θα ορίσουμε τέτοιες συναρτήσεις όμως είθισται να εστιάζουμε στη μέση τιμή της κατανομής επειδή τα μέτρα κεντρικής θέσης είναι συνήθως αυτά στα οποία επικεντρώνεται το ενδιαφέρον μας. Στην κλασική περίπτωση των κανονικών γραμμικών μοντέλων η

μέση τιμή και η γραμμική προβλέπουσα ταυτίζονται, αφού και οι δύο μπορούν να πάρουν τιμές στην πραγματική ευθεία, έτσι για το μοντέλο αυτό ισχύει ότι:

$$g(\mu_i) = \mu_i = \eta_i = \sum_{j=1}^p \beta_j X_{ij}, i = 1, 2, \dots, n$$

Επομένως, γίνεται αντιληπτό πως μια επιθυμητή ιδιότητα που πρέπει να έχει μια συνάρτηση σύνδεσης είναι να απεικονίζει τις τιμές που λαμβάνει η μέση τιμή στην πραγματική ευθεία, στην οποία λαμβάνει τιμές η γραμμική προβλέπουσα. Για παράδειγμα, εάν $Y \sim \text{Binomial}(1, \pi)$ τότε $E(Y) = \mu = \pi$. Σε αυτήν την περίπτωση η χρήση της ταυτοτικής απεικόνισης, $g(\pi) = \pi$, για τη σύνδεση της γραμμικής προβλέπουσας με τη μέση τιμή δεν είναι επιτρεπτή γιατί δεν πληροί τη σχέση

$$\pi = \sum_{j=1}^p \beta_j X_{ij}, i = 1, 2, \dots, n \quad (1.11)$$

διότι το αριστερό μέλος της ισότητας (1.11) λαμβάνει τιμές στο διάστημα $[0, 1]$, ενώ το δεξί μέλος λαμβάνει τιμές σε όλη την πραγματική ευθεία. Έτσι είναι απαραίτητη η εύρεση μίας συνάρτησης σύνδεσης η οποία θα απεικονίζει το διάστημα $[0, 1]$ στο $\mathbb{R}$ και θα είναι συνάρτηση της μέσης τιμής π . Η συνάρτηση σύνδεσης (link function) που χρησιμοποιείται συνήθως σε αυτήν την περίπτωση είναι η logit συνάρτηση, ή αλλιώς logit μετασχηματισμός και δίνεται από τη σχέση $g(\pi) = \text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$, όπου $g: [0, 1] \mapsto \mathbb{R}$. Το μοντέλο

$$\text{logit}(\pi) = \sum_{j=1}^p \beta_j X_{ij}, i = 1, 2, \dots, n \quad (1.12)$$

είναι γνωστό ως μοντέλο λογιστικής παλινδρόμης (logistic regression model).

Παρατηρήσεις:

- (1) Η $\text{Binomial}(1, \pi) \equiv \text{Bernoulli}(\pi)$ ανήκει στην εκθετική οικογένεια κατανομών.

Απόδειξη. Πράγματι, αν $Y \sim \text{Binomial}(1, \pi)$

$$\Rightarrow f_Y(y; \pi) = \pi^y (1 - \pi)^{1-y} = \exp \left(y \log \left(\frac{\pi}{1 - \pi} \right) + \log (1 - \pi) \right)$$

Έτσι για $\theta = \log \left(\frac{\pi}{1 - \pi} \right)$, $b(\theta) = \log (1 - \pi) = -\log(1 + e^\theta)$, $\alpha(\phi) = 1$ και $c(y, \phi) = 0$ έχουμε ότι η Bernoulli (π) ανήκει στην εκθετική οικογένεια κατανομών. Παρατηρούμε ότι σε αυτήν την περίπτωση η παράμετρος διασποράς είναι ίση με 1. $\square$

- (2) Με παρόμοιο τρόπο μπορεί να αποδειχθεί πως και άλλες κατανομές, όπως η Κανονική κατανομή, η κατανομή Poisson, η κατανομή Γάμμα και η εκθετική κατανομή ανήκουν στην εκθετική οικογένεια.
- (3) Στην απόδειξη πως η κατανομή Bernoulli ανήκει στην εκθετική οικογένεια κατανομών παρατηρούμε πως η κανονική παράμετρος θ ταυτίζεται με την συνάρτηση σύνδεσης $g(\pi)$ που δόθηκε προηγουμένως για το μοντέλο της λογιστικής παλινδρόμησης. Σε αυτήν την περίπτωση η συνάρτηση σύνδεσης ονομάζεται **κανονική συνάρτηση σύνδεσης (canonical link function)**.
- (4) Εφόσον ο μετασχηματισμός $g(\cdot)$ έχει υποτεθεί ως ένας 1-1 μετασχηματισμός μπορούμε να γράψουμε ότι:

$$g(\mu_i) = \eta_i \Leftrightarrow \mu_i = g^{-1}(\eta_i) = g^{-1}(X\beta) \quad (1.13)$$

Με αυτήν την παρατήρηση σκοπός είναι να τονιστεί πως παρόλο που το γενικευμένο γραμμικό μοντέλο εκφράζεται ως μια συνάρτηση της μέσης τιμής είναι πάντα εφικτό να λυθεί ως προς τη μέση τιμή μ .

Στον επόμενο πίνακα δίνονται οι συναρτήσεις σύνδεσης που χρησιμοποιούνται ανάλογα με το είδος της κατανομής που ακολουθεί η εξαρτημένη μεταβλητή

Y . Σημειώνουμε πως σε αυτές τις περιπτώσεις οι συναρτήσεις σύνδεσης είναι κανονικές.

Κατανομή	Συνάρτηση Σύνδεσης
$Y_i \sim \text{Κανονική}$	$g(\mu_i) = \mu_i$, (identity link)
$Y_i \sim \text{Poisson}$	$g(\mu_i) = \log(\mu_i)$, (log link)
$Y_i \sim \text{Binomial}(n_i, \pi_i)$	$g(\mu_i) = \log(\frac{\mu_i}{1-\mu_i})$, (logit link)
$Y_i \sim \text{Gamma}$	$g(\mu_i) = -\frac{1}{\mu_i}$, (negative inverse)
$Y_i \sim \text{Εκθετική}$	$g(\mu_i) = -\frac{1}{\mu_i}$, (negative inverse)

1.4 Εκτίμηση των παραμέτρων

Προκειμένου να προσαρμόσουμε ένα μοντέλο στα δεδομένα μας θα πρέπει να εκτιμήσουμε τις άγνωστες παραμέτρους β_j , $j = 1, 2, \dots, p$. Στα γενικευμένα γραμμικά μοντέλα για την εκτίμηση των παραμέτρων χρησιμοποιούνται μέθοδοι που βασίζονται στην μέγιστη πιθανοφάνεια. Ένα περίγραμμα της μεθόδου δίνεται παρακάτω.

Ορισμός 1.4.1. Έστω $Y_1, Y_2, \dots, Y_n$ τυχαίο δείγμα από έναν πληθυσμό με συνάρτηση πιθανότητας $p(y; \theta)$ ή συνάρτηση πυκνότητας πιθανότητας $f(y; \theta)$ τότε η συνάρτηση

$$L(y; \theta) = \prod_{i=1}^n p_{Y_i}(y_i; \theta) \quad \text{ή} \quad L(y; \theta) = \prod_{i=1}^n f_{Y_i}(y_i; \theta) \quad (1.14)$$

που θεωρείται ως συνάρτηση της παραμέτρου θ , ονομάζεται συνάρτηση πιθανοφάνειας.

Παρατήρηση:

- (1) Η συνάρτηση πιθανοφάνειας συμβολίζει την πιθανότητα πραγματοποίησης του τυχαίου δείγματος $\{Y_1, Y_2, \dots, Y_n\}$. Επομένως ο πιο ενδεδειγμένος τρόπος για την εκτίμηση της άγνωστης παραμέτρου θ είναι να προσπαθήσουμε να βρούμε εκείνο το θ για το οποίο μεγιστοποιείται η συνάρτηση πιθανοφάνειας (ή ισοδύναμα ο λογάριθμος της συνάρτησης πιθανοφάνειας).

Ορισμός 1.4.2. Έστω $L(y; \theta)$ η συνάρτηση πιθανοφάνειας ενός τυχαίου δείγματος $Y_1, Y_2, \dots, Y_n$. Ο εκτιμητής $\hat{\theta} = \hat{\theta}(Y_1, Y_2, \dots, Y_n)$ λέγεται εκτιμητής μέγιστης πιθανοφάνειας του θ , εάν

$$L(y; \hat{\theta}) = \max_{\theta \in \Theta} L(y; \theta) \quad (1.15)$$

Συνεπώς ο εκτιμητής μέγιστης πιθανοφάνειας προκύπτει ως λύση ενός προβλήματος μεγιστοποίησης. Όταν ο εκτιμητής υπάρχει είναι πιθανό να μπορούμε να τον προσδιορίσουμε με αναλυτικές μεθόδους, σε κλειστή μορφή. Ωστόσο, δεν είναι λίγες και οι περιπτώσεις όπου το πρόβλημα μεγιστοποίησης δεν λύνεται αναλυτικά, σε αυτές τις περιπτώσεις ο εκτιμητής προσδιορίζεται με χρήση αριθμητικών μεθόδων κάποιες από τις οποίες θα περιγραφούν στη συνέχεια. Αν η κατανομή της Y ανήκει στην εκθετική οικογένεια κατανομών τότε γράφεται στη μορφή της σχέσης (1.2). Επομένως, η συνάρτηση πιθανοφάνειας δίνεται από τη σχέση:

$$L(y_i; \theta_i, \phi) = \prod_{i=1}^n f_Y(y_i, \theta_i, \phi) = \prod_{i=1}^n \exp \left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right) \quad (1.16)$$

ενώ ο λογάριθμος της συνάρτησης πιθανοφάνειας από τη σχέση:

$$\log(L(y_i; \theta_i, \phi)) = l(y_i; \theta_i, \phi) = \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \quad (1.17)$$

Έτσι, παραγωγίζοντας την (1.17) ως προς τις άγνωστες παραμέτρους β_j και με τη βοήθεια του κανόνα της αλυσίδας (chain rule) παίρνουμε τη σχέση:

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}, \quad j = 1, 2, \dots, p \quad (1.18)$$

Η συνάρτηση της πρώτης παραγώγου του λογαρίθμου της πιθανοφάνειας ονομάζεται συνάρτηση score και συμβολίζεται με U_j . Οι μερικές παράγωγοι της προηγούμενης σχέσης γράφονται ως εξής (βλ. Agresti, 2015):

- $\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{\alpha(\phi)}$.
- $\frac{\partial \eta_i}{\partial \beta_j} = X_{ij}$.
- $\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i} \right)^{-1} = (b''(\theta_i))^{-1} = \left(\frac{Var(Y_i)}{\alpha(\phi)} \right)^{-1} = \frac{\alpha(\phi)}{Var(Y_i)}$.

Επομένως, η προηγούμενη σχέση γράφεται:

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^n \frac{y_i - \mu_i}{Var(Y_i)} X_{ij} \frac{\partial \mu_i}{\partial \eta_i}, \quad j = 1, 2, \dots, p. \quad (1.19)$$

Οι παράμετροι β_j , $j = 1, 2, \dots, p$ εμφανίζονται στην προηγούμενη εξίσωση μέσω της σχέσης: $\mu_i = g^{-1} \left(\sum_{j=1}^p \beta_j X_{ij} \right)$. Λύνοντας, το σύστημα εξισώσεων:

$$U_j = 0, \quad j = 1, 2, \dots, p. \quad (1.20)$$

βρίσκουμε τον εκτιμητή μέγιστης πιθανοφάνειας. Πολλές φορές όμως το προηγούμενο σύστημα εξισώσεων δεν μπορεί να μας δώσει λύση σε κλειστή αναλυτική μορφή, συνεπώς για την εύρεση του εκτιμητή θα πρέπει να στραφούμε σε αριθμητικές μεθόδους.

1.4.1 Η μέθοδος Newton-Raphson

Η μέθοδος Newton-Raphson σε κάθε βήμα προσεγγίζει στο σημείο $\hat{\beta}^{(\kappa)}$ την συνάρτηση πιθανοφάνειας χρησιμοποιώντας τις παραγώγους δεύτερης τάξης, δηλαδή μια τετραγωνική προσέγγιση. Η βασική εξίσωση μετάβασης από μια προσέγγιση στην επόμενη δίνεται από τη σχέση:

$$\begin{aligned}\hat{\beta}^{(\kappa)} &= \hat{\beta}^{(\kappa-1)} - \left[\frac{\partial^2 l}{\partial \beta_j \partial \beta_r} \right]_{\beta=\hat{\beta}^{(\kappa-1)}}^{-1} U^{(\kappa-1)} \\ &= \hat{\beta}^{(\kappa-1)} - (H^{(\kappa-1)})^{-1} U^{(\kappa-1)}\end{aligned}\quad (1.21)$$

όπου H είναι ο $p \times p$ εσσιανός πίνακας που έχει ως στοιχεία του τις μερικές παραγώγους δεύτερης τάξης του λογαρίθμου της συνάρτησης πιθανοφάνειας υπολογισμένες στο σημείο $\beta = \hat{\beta}^{(\kappa-1)}$, ενώ U είναι το $p \times 1$ διάνυσμα score που έχει ως στοιχεία τις μερικές παραγώγους πρώτης τάξης των παραμέτρων τις παλινδρόμησης υπολογισμένες στο σημείο $\beta = \hat{\beta}^{(\kappa-1)}$.

Ο αλγόριθμος τερματίζει όταν η διαφορά των εκτιμώμενων παραμέτρων από ένα βήμα στο επόμενο γίνει επαρκώς μικρή, δηλαδή όταν ισχύει η σχέση:

$$\|\hat{\beta}^{(\kappa)} - \hat{\beta}^{(\kappa-1)}\| < \epsilon \quad (1.22)$$

όπου το ϵ είναι μια προκαθορισμένη ακρίβεια που βάζουμε αυθαίρετα. Μια συνηθισμένη τιμή για το ϵ είναι 10^{-6} .

1.4.2 Ο αλγόριθμος του Fisher

Μια τροποποίηση του αλγορίθμου Newton-Raphson δίνεται μέσω του αλγορίθμου του Fisher, στον οποίο ο εσσιανός πίνακας H που κάνει χρήση της παρατηρούμενης πληροφορίας, αντικαθίσταται από τον πίνακα της μέσης τιμής του πίνακα

H. Ο πίνακας της μέσης τιμής του εσσιανού ονομάζεται πίνακας πληροφορίας του Fisher και κάνει χρήση της αναμενόμενης πληροφορίας.

Το τυπικό jr-στοιχείο του πίνακα πληροφορίας δίνεται από τη σχέση

$$I_{jr} = E(U_j U_r) = \sum_{i=1}^n \frac{E(y_i - \mu_i)^2 X_{ij} X_{ir}}{(Var(Y_i))^2} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 = \sum_{i=1}^n \frac{X_{ij} X_{ir}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

$$j, r = 1, 2, \dots, p \quad (1.23)$$

Ειδικότερα, επειδή βρισκόμαστε στο περιβάλλον της εκθετικής οικογένειας κατανομών, γνωρίζουμε ότι (Cox & Hinkley, 1974):

$$E \left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_r} \right) = -E \left[\left(\frac{\partial l}{\partial \beta_j} \right) \left(\frac{\partial l}{\partial \beta_r} \right) \right] = E(U_j U_r) = I_{jr}$$

$$\Rightarrow -E \left(\frac{\partial^2 l}{\partial \beta_j \partial \beta_r} \right) = \sum_{i=1}^n \frac{X_{ij} X_{ir}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 = I_{jr} \quad (1.24)$$

Η γενική μορφή του πίνακα πληροφορίας, σύμφωνα με τη μορφή του τυπικού του στοιχείου δίνεται από τη σχέση:

$$I = X^T W X \quad (1.25)$$

όπου X είναι ο πίνακας της παλινδρόμησης και $W = diag(w_{ii}) = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$, $i = 1, 2, \dots, n$.

Από τα προηγούμενα συμπεραίνουμε πως η βασική σχέση μετάβασης από ένα βήμα στο επόμενο για τον αλγόριθμο του Fisher δίνεται από την εξίσωση:

$$\hat{\beta}^{(\kappa)} = \hat{\beta}^{(\kappa-1)} + (I^{(\kappa-1)})^{-1} U^{(\kappa-1)}$$

$$\Rightarrow I^{(\kappa-1)} \hat{\beta}^{(\kappa)} = I^{(\kappa-1)} \hat{\beta}^{(\kappa-1)} + U^{(\kappa-1)}$$

$$\Rightarrow X^T W^{(\kappa-1)} X \hat{\beta}^{(\kappa)} = X^T W^{(\kappa-1)} X \hat{\beta}^{(\kappa-1)} + U^{(\kappa-1)}$$

$$\Rightarrow X^T W^{(\kappa-1)} X \hat{\beta}^{(\kappa)} = X^T W^{(\kappa-1)} z^{(\kappa-1)} \quad (1.26)$$

όπου το i -οστό στοιχείο του πίνακα z δίνεται από τη σχέση

$$z_i = \sum_{j=1}^p X_{ij} \hat{\beta}_j + (y_i - \hat{\mu}_i) \frac{\partial \eta_i}{\partial \mu_i}. \quad (1.27)$$

Έτσι σε κάθε βήμα του αλγορίθμου ο βελτιωμένος εκτιμητής μέγιστης πιθανοφάνειας δίνεται από τη σχέση:

$$\hat{\beta}^{(\kappa)} = (X^T W^{(\kappa-1)} X)^{-1} X^T W^{(\kappa-1)} z^{(\kappa-1)} \quad (1.28)$$

Ένα κριτήριο τερματισμού του αλγορίθμου είναι το:

$$\|\hat{\beta}^{(\kappa)} - \hat{\beta}^{(\kappa-1)}\| < \epsilon$$

όπου το ϵ είναι μια προκαθορισμένη ακρίβεια του αλγορίθμου συνήθως $\epsilon = 10^{-6}$. Ο αλγόριθμος του Fisher ονομάζεται συχνά και ως μέθοδος των σταθμισμένων ελαχίστων τετραγώνων (iteratively weighted least squares).

Οι διαφορές μεταξύ του αλγορίθμου Newton-Raphson και του αλγορίθμου του Fisher είναι λεπτές αλλά και σημαντικές δύο από τις κυριότερες δίνονται παρακάτω (βλ. Knight, 2000):

- Όταν και οι δύο αλγόριθμοι συγκλίνουν ο αλγόριθμος Newton-Raphson είναι συνήθως γρηγορότερος.
- Η ακτίνα σύγκλισης του αλγορίθμου του Fisher είναι συνήθως μεγαλύτερη από την αντίστοιχη του αλγορίθμου Newton-Raphson. Αυτός είναι και ο λόγος που η επιλογή του αρχικού σημείου δεν είναι τόσο σημαντική στον αλγόριθμο του Fisher όσο στον αλγόριθμο Newton-Raphson.

1.5 Εφαρμογή του αλγορίθμου του Fisher στη λογιστική παλινδρόμηση

Θα εφαρμοστεί ο αλγόριθμος του Fisher για την περίπτωση όπου $Y_i \sim \text{Binomial}(n_i, \pi_i)$. Γνωρίζουμε πως σε αυτήν την περίπτωση: $E(Y_i) = n_i \pi_i$ ενώ ως συνάρτηση σύνδεσης χρησιμοποιείται η:

$$\begin{aligned} \text{logit}(\pi(x)) &= \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) \\ \Leftrightarrow \pi(x) &= \frac{\exp\left(\sum_{j=1}^p b_j X_{ij}\right)}{1 + \exp\left(\sum_{j=1}^p b_j X_{ij}\right)} \end{aligned}$$

Επομένως, το μοντέλο της λογιστικής παλινδρόμησης έχει τη μορφή

$$\text{logit}(\pi(x_i)) = \sum_{j=1}^p \beta_j X_{ij}$$

Είδαμε, πως στα γενικευμένα γραμμικά μοντέλα για την εκτίμηση των παραμέτρων της παλινδρόμησης θα πρέπει να χρησιμοποιήσουμε τη μέθοδο της μέγιστης πιθανοφάνειας. Ο λογάριθμος της συνάρτησης πιθανοφάνειας δίνεται από τη σχέση:

$$\begin{aligned} l &= \log(p_Y(y_i, \pi_i)) = \sum_{i=1}^n \left(\log\binom{n_i}{y_i} + y_i \log(\pi_i) + (n_i - y_i) \log(1 - \pi_i) \right) \\ &= \sum_{i=1}^n \log\binom{n_i}{y_i} + \sum_{j=1}^p \left(\sum_{i=1}^n y_i X_{ij} \right) \beta_j - \sum_{i=1}^n n_i \log\left(1 + \exp\left(\sum_{j=1}^p \beta_j X_{ij} \right) \right) \end{aligned}$$

Επομένως,

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^n (y_i - n_i \pi(x_i)) X_{ij}$$

Ο εκτιμητής μέγιστης πιθανοφάνειας προκύπτει ως επίλυση του συστήματος

εξισώσεων

$$\frac{\partial l}{\partial \beta_j} = 0, j = 1, 2, \dots, p$$

Δεδομένης, της μορφής του $\pi(x_i)$ συμπεραίνουμε πως το σύστημα αυτό είναι αδύνατο να λυθεί σε κλειστή μορφή και έτσι πρέπει να προσεγγίσουμε τη λύση του με αριθμητικές μεθόδους.

Η εξίσωση που μας δίνει τους βελτιωμένους εκτιμητές μέγιστης πιθανοφάνειας σε κάθε βήμα του αλγορίθμου το Fisher είδαμε πως δίνεται από τη σχέση:

$$\hat{\beta}^{(\kappa)} = (X^T W^{(\kappa-1)} X)^{-1} X^T W^{(\kappa-1)} z^{(\kappa-1)},$$

με το τυπικό i -οστό στοιχείο του πίνακα z να είναι το:

$$z_i = \sum_{j=1}^p X_{ij} \hat{\beta}_j + (y_i - \hat{\mu}_i) \frac{\partial \eta_i}{\partial \mu_i}$$

$$\Rightarrow z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i) \frac{\partial \eta_i}{\partial \mu_i}$$

Τότε όμως, $\eta_i = \log \left(\frac{\pi_i}{1 - \pi_i} \right) \stackrel{\mu_i = n_i \pi_i}{=} \log \left(\frac{\mu_i / n_i}{1 - \mu_i / n_i} \right) = \log \left(\frac{\mu_i}{n_i - \mu_i} \right)$

Έτσι, $\frac{\partial \eta_i}{\partial \mu_i} = \frac{n_i}{\mu_i (n_i - \mu_i)}$

Άρα το τυπικό στοιχείο του πίνακα z είναι το:

$$z_i = \hat{\eta}_i + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i (n_i - \hat{\mu}_i)},$$

ενώ τα στοιχεία w_{ii} του διαγώνιου πίνακα W γνωρίζουμε πως δίνονται από τη σχέση

$$w_{ii} = \frac{\hat{\mu}_i (n_i - \hat{\mu}_i)}{n_i}$$

Από τις προηγούμενες σχέσεις είμαστε σε θέση να υπολογίζουμε σε κάθε βήμα τον βελτιωμένο εκτιμητή μέγιστης πιθανοφάνειας:

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

Το κριτήριο τερματισμού του αλγορίθμου είναι:

$$\|\hat{\beta}^{(\kappa)} - \hat{\beta}^{(\kappa-1)}\| < \epsilon,$$

όπου το ϵ είναι μια προκαθορισμένη ακρίβεια του αλγορίθμου συνήθως $\epsilon = 10^{-6}$.

Η περίπτωση της κανονικής συνάρτησης σύνδεσης

Στην περίπτωση κατά την οποία η συνάρτηση σύνδεσης που έχουμε επιλέξει είναι κανονική (canonical link function) οι αλγόριθμοι που αναφέρθηκαν προηγουμένως συμπίπτουν (βλ. Agresti, 2003). Έστω λοιπόν ότι έχουμε κανονική συνάρτηση σύνδεσης, τότε

$$\eta_i = \theta_i = \sum_{j=1}^p \beta_j X_{ij}.$$

Τότε όμως,

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i). \quad (1.29)$$

Έτσι τελικά,

$$\frac{\partial l_i}{\partial \beta_j} = \frac{y_i - \mu_i}{\text{Var}(Y_i)} X_{ij} \frac{\partial \mu_i}{\partial \eta_i} = \frac{y_i - \mu_i}{\text{Var}(Y_i)} X_{ij} b''(\theta_i) = \frac{y_i - \mu_i}{\alpha(\phi)} X_{ij} \quad (1.30)$$

$$\Rightarrow \frac{\partial^2 l_i}{\partial \beta_j \partial \beta_h} = -\frac{X_{ij}}{\alpha(\phi)} \left(\frac{\partial \mu_i}{\partial \beta_h} \right). \quad (1.31)$$

Παρατηρούμε λοιπόν ότι στην περίπτωση της κανονικής συνάρτησης σύνδεσης η δεύτερη παράγωγος του λογαρίθμου της συνάρτησης πιθανοφάνειας δεν εξαρτάται από τα δεδομένα y_i . Συνεπώς,

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_h} = E \left(\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_h} \right), \quad 1 \leq j, h \leq p. \quad (1.32)$$

Από την προηγούμενη ισότητα είναι φανερό ότι στην περίπτωση της κανονικής συνάρτησης σύνδεσης ισχύει ότι: $H = -I$, όπου H είναι ο εσσιανός πίνακας (hessian matrix) του λογαρίθμου της συνάρτησης πιθανοφάνειας και I ο πίνακας πληροφορίας (information matrix) του Fisher. Λόγω της ταύτισης αυτών των δυο πινάκων έπεται πως και οι αλγόριθμοι Newton-Raphson και Fisher θα ταυτίζονται για την περίπτωση της κανονικής συνάρτησης σύνδεσης. Τέλος, ιδιαίτερη μορφή έχουν και τα στοιχεία του πίνακα πληροφορίας του Fisher στην περίπτωση της κανονικής συνάρτησης σύνδεσης. Το τυπικό στοιχείο του πίνακα δίνεται από τη σχέση

$$\begin{aligned} I_{jr} &= \sum_{i=1}^n \frac{X_{ij} X_{ir}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 = \sum_i^n \frac{X_{ij} X_{ir}}{b''(\theta_i) \frac{\phi}{p_i}} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \\ &= \frac{1}{\phi} \sum_{i=1}^n \frac{p_i X_{ij} X_{ir}}{b''(\theta_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2. \end{aligned} \quad (1.33)$$

Έτσι γενικά ο πίνακας πληροφορίας του Fisher μπορεί να γραφεί στη μορφή

$$I = \frac{1}{\phi} X^T W X \quad (1.34)$$

όπου το τυπικό στοιχείο του $W = \text{diag}(w_{ii})$ δίνεται από τη σχέση

$$w_{ii} = \frac{p_i}{b''(\theta_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2. \quad (1.35)$$

Στην περίπτωση της κανονικής συνάρτησης σύνδεσης γνωρίζουμε ότι $\left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 = (b''(\theta_i))^2$. Αν επιπλέον θεωρήσουμε τα εκ των προτέρων βάρη p_i ίσα με 1, τότε το τυπικό στοιχείο του πίνακα W δίνεται από τη σχέση:

$$w_{ii} = b''(\theta_i) = \frac{1}{\phi} Var(Y_i). \quad (1.36)$$

Κεφάλαιο 2

ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ

Η ανάλυση επιβίωσης (survival analysis) είναι ένας από τους παλαιότερους στατιστικούς κλάδους με ρίζες στη δημογραφία και στην αναλογιστική επιστήμη, κατά τον 17^ο αιώνα. Στη σημερινή εποχή η ανάλυση επιβίωσης αποτελεί ταυτόσημη έννοια με αυτήν της ανάλυσης του χρονικού διαστήματος που προηγείται της καταγραφής ενός συμβάντος. Ως όρος χρησιμοποιείται κατά κύριο λόγο σε βιοϊατρικές επιστήμες, όπου το ενδιαφέρον επικεντρώνεται στην παρατήρηση του χρόνου ζωής είτε των ασθενών είτε των πειραματόζων.

Η ανάπτυξη των επιχειρησιακών ερευνών κατέδειξε ότι υπάρχουν και πολλά άλλα προβλήματα και μοντέλα διαφορετικής υφής, στα οποία η ανάλυση επιβίωσης μπορεί να εφαρμοστεί και να προσφέρει λύσεις. Ακόμα περισσότερο με την ταχύτατη εξέλιξη των ηλεκτρονικών υπολογιστών, η εφαρμογή και η ανάπτυξη της ανάλυσης επιβίωσης γίνεται ολοένα και μεγαλύτερη, καθώς παράγεται μεγάλος αριθμός λογισμικών πακέτων για το σκοπό αυτό.

Γίνεται εύκολα αντιληπτό ότι η περιγραφή και η μοντελοποίηση του χρόνου ζωής αποτελεί το κύριο ζητούμενο σε πολλές εκφάνσεις της ανθρώπινης ζωής και

δραστηριότητας. Ως εκ τούτου οι μεθοδολογίες που αναπτύσσονται στην ανάλυση επιβίωσης αποτελούν πολύτιμο εργαλείο για ερευνητές διαφόρων επιστημονικών κλάδων. Για τη συγγραφή αυτού του κεφαλαίου βασιστήκαμε κυρίως στα βιβλία των: Collett (2015), Kleinbaum and Klein (2010) και Kalbfleisch and Prentice (2011).

2.1 Βασικές συναρτήσεις

Ξεκινάμε περιγράφοντας το είδος του προβλήματος που θεραπεύει η ανάλυση επιβίωσης, η οποία γενικά είναι μία συλλογή στατιστικών διαδικασιών για την ανάλυση δεδομένων όπου η μεταβλητή ενδιαφέροντος είναι **ο χρόνος μέχρι την εμφάνιση ενός γεγονότος**. Επί της ουσίας, ενδιαφερόμαστε για την εμφάνιση ή όχι ενός γεγονότος, αλλά κυρίως για το χρόνο που περνάει κάθε άτομο χωρίς την εμφάνιση του υπό μελέτη γεγονότος.

Με τον όρο *χρόνος* εννοούμε τα χρόνια, τους μήνες, τις εβδομάδες, τις ημέρες ή γενικότερα το χρονικό διάστημα από την έναρξη της παρακολούθησης ενός ατόμου μέχρι και την εμφάνιση του γεγονότος.

Παραδείγματα γεγονότων αποτελούν είτε ο θάνατος, είτε το χρονικό διάστημα ύφεσης μίας ασθένειας, είτε ο χρόνος από την αποφυλάκιση ενός ατόμου μέχρι την επόμενη σύλληψη του και γενικότερα οποιαδήποτε καθορισμένη εμπειρία ενδιαφέροντος μπορεί να βιώσει ένα άτομο.

Στην ανάλυση επιβίωσης συνήθως η χρονική μεταβλητή αναφέρεται ως *χρόνος επιβίωσης*, διότι μας δίνει το χρόνο που ένα άτομο έχει επιβιώσει κατά τη διάρκεια κάποιας παρακολούθησης. Επιπλέον, είναι σύνηθες ο όρος *γεγονός* να αντικαθίσταται από τον όρο *αποτυχία*, επειδή το γεγονός το οποίο συχνά μελετάμε είναι

ο θάνατος ή κάποια άλλη αρνητική για ένα άτομο εμπειρία.

Όπως αναφέρθηκε προηγουμένως, θεμελιώδη ρόλο στην ανάλυση επιβίωσης έχει η τυχαία μεταβλητή που περιγράφει το χρόνο επιβίωσης ενός ατόμου. Η μεταβλητή αυτή θα συμβολίζεται με το γράμμα T . Εφόσον η μεταβλητή T παριστά χρόνο έπεται άμεσα ότι μπορεί να πάρει οποιαδήποτε τιμή μεγαλύτερη ή ίση του μηδενός. Συμβολίζουμε με το γράμμα t μια συγκεκριμένη τιμή της μεταβλητής T . Για παράδειγμα, εάν μας ενδιαφέρει να αξιολογήσουμε εάν ένα άτομο επιβιώνει για περισσότερο από 5 χρόνια μετά την υποβολή σε θεραπεία καρκίνου, τότε το t είναι ίσο με 5.

Ορισμός 2.1.1. Η **αθροιστική συνάρτηση κατανομής** της τυχαίας μεταβλητής T συμβολίζεται με F_T και είναι μια πραγματική συνάρτηση που ορίζεται ως εξής: $F_T(t) \equiv F(t) = P(T \leq t), \forall t \geq 0$.

Ο προηγούμενος ορισμός ορίζει γενικά την έννοια της αθροιστικής συνάρτησης κατανομής. Σε όρους της ανάλυσης επιβίωσης η αθροιστική συνάρτηση κατανομής της τυχαίας μεταβλητής T , που παριστά χρόνο επιβίωσης, εκφράζει την πιθανότητα αποτυχίας μέχρι τη χρονική στιγμή t . Στη συνέχεια δίνονται μερικές ιδιότητες της αθροιστικής συνάρτησης κατανομής.

Ιδιότητες:

- 1) Από τον ορισμό προκύπτει ότι το σύνολο τιμών της $F(\cdot)$ είναι το διάστημα $[0, 1]$.
- 2) Η $F(t)$ είναι αύξουσα και συνεχής συνάρτηση του t .
- 3) $F(0) = 0$ και $\lim_{t \rightarrow +\infty} F(t) = 1$.

Ορισμός 2.1.2. Έστω T μία τυχαία μεταβλητή. Η T λέγεται συνεχής αν υπάρχει μια μη αρνητική, ολοκληρώσιμη συνάρτηση f_T ορισμένη στο σύνολο των πραγματικών αριθμών $\mathbb{R}$, τέτοια ώστε: $P(T \in B) = \int_B f_T(t)dt, B \subseteq \mathbb{R}$.

Η συνάρτηση f_T ονομάζεται **συνάρτηση πυκνότητας πιθανότητας** της τυχαίας μεταβλητής T .

Ιδιότητες:

$$1) f_T(t) \equiv f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} = F'(t), \forall t \geq 0.$$

$$2) f(t) \geq 0, \forall t \geq 0.$$

$$3) \int_0^{+\infty} f_T(t) dt = 1.$$

Σε όρους ανάλυσης επιβίωσης και για μικρό Δt η ποσότητα $f(t) \Delta t$ εκφράζει προσεγγιστικά την πιθανότητα αποτυχίας στο διάστημα $[t, t + \Delta t)$

Ορισμός 2.1.3. Η **συνάρτηση επιβίωσης** της τυχαίας μεταβλητής T συμβολίζεται ως S_T και μας δίνει την πιθανότητα η τυχαία μεταβλητή T να πάρει τιμές μεγαλύτερες του t , δηλαδή

$$S_T(t) \equiv S(t) = P(T > t) = 1 - F(t), \forall t \geq 0 \quad (2.1)$$

Άμεσα προκύπτει από τον ορισμό ότι η συνάρτηση επιβίωσης δίνει την πιθανότητα ότι ένα άτομο επιβιώνει περισσότερο από κάποιο συγκεκριμένο χρόνο t .

Ιδιότητες:

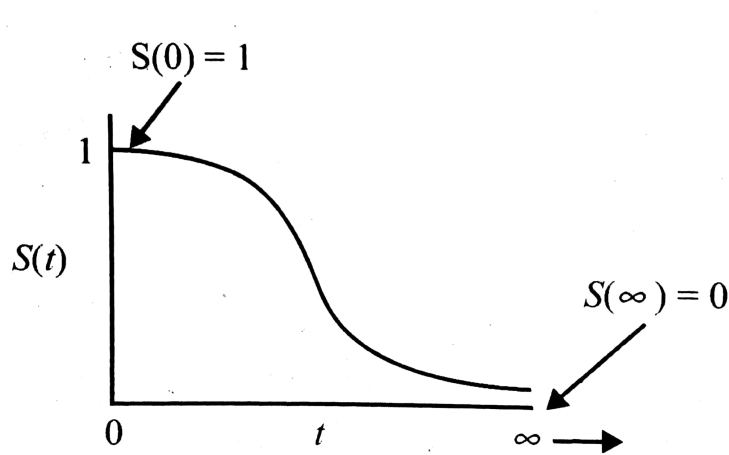
1) Η συνάρτηση επιβίωσης είναι μία φθίνουσα συνάρτηση του t .

2) Για $t = 0$, έχουμε ότι $S(0) = 1$. Η πρακτική σημασία αυτής της ιδιότητας είναι πως στην αρχή της έρευνας και αφού κανείς δεν έχει εμφανίσει το γεγονός η πιθανότητα επιβίωσης είναι ίση με 1.

3) $\lim_{t \rightarrow +\infty} S(t) = 0$. Η πρακτική σημασία της συγκεκριμένης ιδιότητας είναι πως εάν αυξήσουμε χωρίς όριο το χρόνο τότε είναι φυσιολογικό να μην επιζήσει κανείς και ως εκ τούτου η πιθανότητα επιβίωσης είναι ίση με το 0.

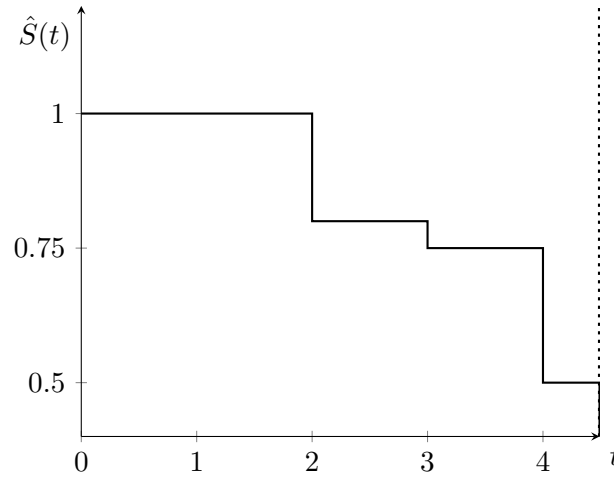
Οι προηγούμενες θεωρητικές ιδιότητες της συνάρτησης επιβίωσης δίνονται παρακάτω μέσω ενός γραφήματος.

Σχήμα 2.1: Θεωρητικές ιδιότητες συνάρτησης επιβίωσης.



Στην πράξη, όταν χρησιμοποιούμε πραγματικά δεδομένα τα γραφήματα που δημιουργούνται έχουν τη μορφή μιας "βηματικής συνάρτησης" (step function) και όχι λείες καμπύλες (βλ. Kleinbaum and Klein, 2010). Ο λόγος για τον οποίο συμβαίνει αυτό είναι διότι στην πράξη παρατηρούμε τα γεγονότα σε μια διακριτή χρονική κλίμακα (ημέρες, εβδομάδες, κ.τ.λ.). Επομένως, είναι πρακτικά αδύνατο να έχουμε την ακριβή μορφή της καμπύλης επιβίωσης και για αυτό το λόγο θέλουμε να βρούμε μία εκτίμησή της. Επιπρόσθετα, επειδή η περίοδος μελέτης δεν είναι ποτέ απεριόριστη είναι πιθανό κάποια άτομα που βρίσκονται στη μελέτη να μην εμφανίσουν το γεγονός. Η καμπύλη της εκτιμώμενης συνάρτησης επιβίωσης, $\hat{S}(t)$, που παίρνουμε σε πραγματικά σύνολα δεδομένων έχει τη μορφή που δίνεται στο επόμενο γράφημα.

Σχήμα 2.2: Μορφή Καμπύλης επιβίωσης σε πραγματικά σύνολα δεδομένων



Ορισμός 2.1.4. Η **συνάρτηση κινδύνου** της τυχαίας μεταβλητής T εκφράζει το στιγμιαίο ρυθμό αποτυχίας τη χρονική στιγμή t , δοθείσης της επιβίωσης μέχρι τη χρονική στιγμή t . Συμβολίζεται με $h_T(t)$ και δίνεται από τη μαθηματική έκφραση:

$$h_T(t) \equiv h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t) | T \geq t}{\Delta t} = \frac{f(t)}{S(t)}, t \geq 0. \quad (2.2)$$

Ιδιότητες:

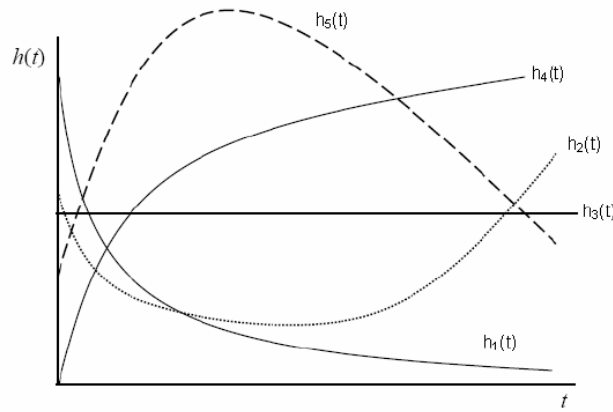
- 1) $h(t) \geq 0$, για κάθε $t \geq 0$.
- 2) Η συνάρτηση $h(\cdot)$ δεν είναι άνω φραγμένη.

Σε όρους ανάλυσης επιβίωσης για μικρό Δt η ποσότητα $h(t) \Delta t$ εκφράζει προσεγγιστικά την πιθανότητα θανάτου στο διάστημα $[t, t + \Delta t)$, δοθείσης της επιβίωσης μέχρι τη χρονική στιγμή t .

Όπως και η συνάρτηση επιβίωσης έτσι και η συνάρτηση κινδύνου μπορεί να παρασταθεί γραφικά για διάφορες τιμές του t . Το επόμενο γράφημα δίνει κάποιες

μορφές συναρτήσεων κινδύνου. Σε αντίθεση με τη συνάρτηση επιβίωσης, το γράφημα της h_T δεν είναι απαραίτητο να ξεκινάει από το 1 και να πηγαίνει προς το 0, αλλά μπορεί να ξεκινάει από οπουδήποτε και να πηγαίνει είτε προς τα πάνω, είτε προς τα κάτω, προς οποιαδήποτε κατεύθυνση στην πάροδο του χρόνου.

Σχήμα 2.3: Συναρτήσεις κινδύνου.



Οι δύο τελευταίες συναρτήσεις που δόθηκαν, δηλαδή η συνάρτηση επιβίωσης και η συνάρτηση κινδύνου παίζουν κομβικό ρόλο στην ανάλυση επιβίωσης. Ανεξάρτητα, όμως από το ποιά από τις δύο συναρτήσεις είμαστε σε θέση να γνωρίζουμε άμεσα υπάρχουν ορισμένες σχέσεις οι οποίες συνδέουν αυτές τις δύο συναρτήσεις. Έτσι στην ουσία εάν γνωρίζουμε την συνάρτηση κινδύνου είμαστε σε θέση μέσω αυτών των σχέσεων να υπολογίσουμε τη συνάρτηση επιβίωσης και αντίστροφα. Οι σχέσεις αυτές είναι οι εξής:

1. $S(t) = e^{-\int_0^t h(u)du} = \exp\{-\int_0^t h(u)du\}.$
2. $h(t) = -\frac{d \log S_T(t)}{dt}.$

Ορισμός 2.1.5. Η συνάρτηση που ορίζεται μέσω του ολοκληρώματος $\int_0^t h(u)du$ ονομάζεται *αθροιστική συνάρτηση κινδύνου* (cumulative hazard function) της τυχαίας μεταβλητής T και συμβολίζεται με $H_T(t)$ επομένως ισχύει ότι:

$$H_T(t) \equiv H(t) = \int_0^t h(u)du \quad (2.3)$$

και εκφράζει την πιθανότητα αποτυχίας την χρονική στιγμή t δοθείσης της επιβίωσης μέχρι αυτή τη χρονική στιγμή.

Για αυτή τη συνάρτηση ισχύει ότι:

- $H(0) = 0$.
- $\lim_{t \rightarrow +\infty} H(t) = +\infty$.

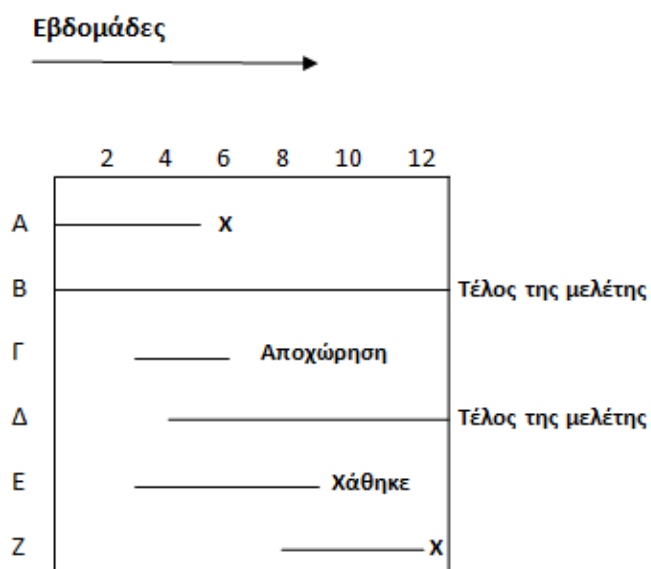
2.2 Λογοκριμένα δεδομένα

Στην ανάλυση επιβίωσης τις περισσότερες φορές εμφανίζεται το πρόβλημα σύμφωνα με το οποίο έχουμε κάποια πληροφορία σχετικά με το χρόνο επιβίωσης ενός ατόμου όμως δεν τον γνωρίζουμε ακριβώς. Σε αυτήν την περίπτωση λέμε ότι τα δεδομένα μας είναι **λογοκριμένα (censored)**. Οι κύριοι λόγοι για τους οποίους μπορεί να εμφανίζονται τέτοια δεδομένα είναι τρεις:

1. Ένα άτομο εισέρχεται στη μελέτη, αλλά μέχρι το τέλος της δεν έχει εμφανίσει το υπό μελέτη χαρακτηριστικό.
2. Ένα άτομο αποχωρεί από την έρευνα λόγω θανάτου και ο θάνατος δεν είναι το υπό μελέτη χαρακτηριστικό.
3. Ένα άτομο εισέρχεται στην έρευνα όμως στην πορεία χάνεται με αποτέλεσμα να σταματήσει η παρακολούθησή του.

Αυτές οι τρεις περιπτώσεις για τις οποίες μπορεί να προκύψουν λογοκριμένα δεδομένα δίνονται γραφικά παρακάτω.

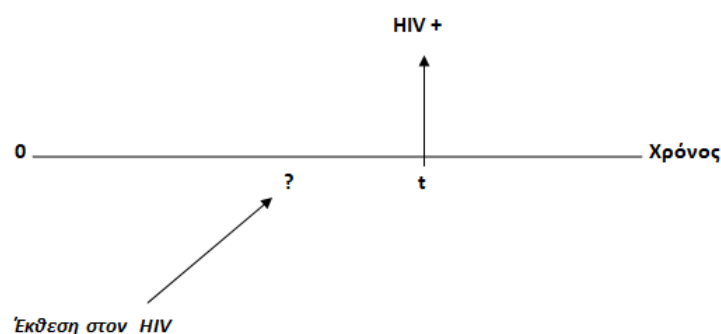
Σχήμα 2.4



Παρατηρούμε στο προηγούμενο γράφημα ότι τα άτομα των οποίων ο χρόνος επιβίωσης έχει λογοκριθεί εμφανίζουν το γεγονός σε μεταγενέστερο χρόνο από αυτόν της συλλογής του δείγματος, δηλαδή όταν εμφανίσουν το υπό μελέτη γεγονός η παρακολούθησή τους θα έχει σταματήσει. Τέτοιες περιπτώσεις λογοκριμένων δεδομένων ονομάζονται **δεξιά λογοκριμένα (right censored)**

Εκτός από την περίπτωση των δεξιά λογοκριμένων δεδομένων υπάρχει και η περίπτωση των **αριστερά λογοκριμένων δεδομένων (left censored data)**. Σε αυτήν την περίπτωση ο πραγματικός χρόνος επιβίωσης ενός ατόμου είναι μικρότερος ή ίσος από τον παρατηρούμενο. Ένα παράδειγμα αριστερά λογοκριμένων δεδομένων δίνεται στο παρακάτω σχήμα.

Σχήμα 2.5: Παράδειγμα αριστερά λογικριμένων δεδομένων

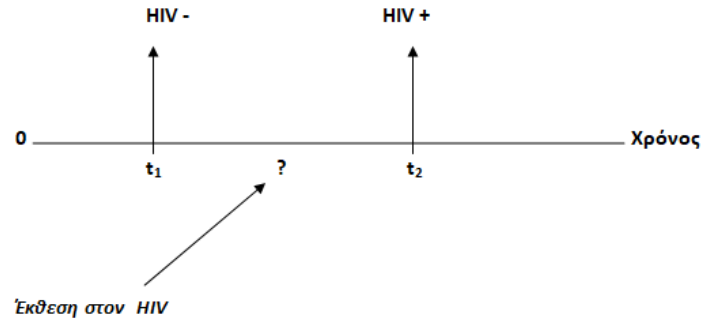


Βλέπουμε ότι ένα άτομο εκτέθηκε στον ιο HIV μια συγκεκριμένη χρονική στιγμή, όμως η διάγνωση της ασθένειας έγινε σε μεταγενέστερο χρόνο. Έτσι εάν υποθέσουμε ότι το γεγονός είναι η έκθεση HIV μπορούμε να εξάγουμε το συμπέρασμα ότι ο πραγματικός χρόνος επιβίωσης είναι άγνωστος και μικρότερος από τον παρατηρούμενο.

Τέλος, πολλές φορές τα δεδομένα ανάλυσης επιβίωσης μπορεί να εμφανίζουν **διαστηματική λογοκρισία (interval censored)**. Αυτές οι περιπτώσεις λογοκρισίας εμφανίζονται όταν ο πραγματικός χρόνος επιβίωσης βρίσκεται ανάμεσα σε δύο γνωστές χρονικές στιγμές, αλλά δεν τον γνωρίζουμε ακριβώς. Το επόμενο σχήμα δίνει ένα παράδειγμα διαστηματικής λογοκρισίας. Βλέπουμε πως ένα άτομο τη χρονική στιγμή t_1 έκανε το test για τον ιο HIV και αυτό ήταν αρνητικό, τη χρονική στιγμή t_2 επανέλαβε το ίδιο test και αυτό ήταν θετικό με αποτέλεσμα να γνωρίζουμε μόνο ότι το άτομο αυτό έχει εκτεθεί στον ιο στο χρονικό διάστημα (t_1, t_2) , αλλά δεν γνωρίζουμε την ακριβή χρονική στιγμή.

Επιπλέον, είναι χρήσιμο να επισημάνουμε πως οι βασικοί μηχανισμοί από τους οποίους προκαλείται η λογοκρισία (censoring) είναι δύο και δίνονται παρακάτω:

Σχήμα 2.6: Παράδειγμα διαστηματικής λογοκρισίας



1. Λογοκρισία τύπου I (*Type I censoring*)

Η παρακολούθηση των μονάδων γίνεται για ένα συγκεκριμένο χρονικό διάστημα. Έτσι, στο τέλος της έρευνας γνωρίζουμε την ακριβή διάρκεια ζωής μόνο εάν αυτή είναι μικρότερη από το προκαθορισμένο χρονικό διάστημα.

2. Λογοκρισία τύπου II (*Type II censoring*)

Η παρακολούθηση του πειράματος σταματάει όταν πραγματοποιηθεί ένας συγκεκριμένος αριθμός αποτυχιών-θανάτων. Αντίθετα, με τη λογοκρισία τύπου I, εδώ η διάρκεια παρακολούθησης είναι τυχαία.

Μία πολύ σημαντική υπόθεση που θα κάνουμε για το φαινόμενο της λογοκρισίας (*censoring*) είναι αυτή της ανεξαρτησίας (*independent censoring*). Σύμφωνα με αυτήν την υπόθεση στο εσωτερικό κάθε ομάδας παρακολούθησης ο ρυθμός εμφάνισης αποτυχιών-θανάτων (*failure rate*) για τα άτομα με λογοκριμένους χρόνους επιβίωσης είναι ίσος με το ρυθμό εμφάνισης αποτυχιών-θανάτων για τα άτομα που παραμένουν στο σύνολο κινδύνου και δεν έχουν λογοκριμένο χρόνο επιβίωσης. Όλες οι μέθοδοι που θα αναπτυχθούν σε αυτό το κεφάλαιο, υποθέτουν την ικανοποίηση της συνθήκης της ανεξαρτησίας.

2.3 Εκτίμηση της συνάρτησης επιβίωσης

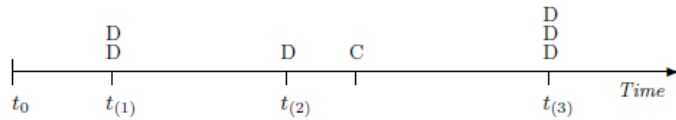
Ένα από τα αρχικά βήματα στην ανάλυση επιβίωσης είναι η παρουσίαση αριθμητικών ή και γραφικών αποτελεσμάτων που βοηθούν στην περιγραφή του χρόνου επιβίωσης των ατόμων τα οποία λαμβάνουν μέρος σε μια μελέτη. Η πιο ευρέως διαδεδομένη συνάρτηση, από αυτές που δόθηκαν προηγουμένως, είναι η συνάρτηση επιβίωσης. Έτσι προκύπτει η ανάγκη για την εκτίμηση της συγκεκριμένης συνάρτησης αλλά και του γραφήματός της. Επιπρόσθετα, όπως είδαμε κεντρικό ρόλο διαδραματίζει η λογοκρισία (censoring) που εμφανίζουν οι χρόνοι επιβίωσης. Επομένως, με βάση τα προηγούμενα συμπεραίνουμε πως η μέθοδος εκτίμησης της συνάρτησης επιβίωσης θα πρέπει να λαμβάνει υπόψιν της το φαινόμενο της λογοκρισίας (censoring). Μια τέτοια μέθοδος είναι η μη-παραμετρική μέθοδος εκτίμησης της συνάρτησης επιβίωσης που προτάθηκε από τους Kaplan-Meier (1958). Προκειμένου να εφαρμοστεί η συγκεκριμένη μέθοδος είναι απαραίτητες οι εξής υποθέσεις:

1. Τα άτομα που έχουν λογοκριμένο χρόνο επιβίωσης έχουν την ίδια πιθανότητα επιβίωσης με τα άτομα που συνεχίζουν την παρακολούθηση.
2. Οι πιθανότητες επιβίωσης είναι οι ίδιες για τα άτομα που εισήλθαν στην έρευνα σε διαφορετικές χρονικές στιγμές.
3. Το γεγονός που μελετάται συμβαίνει σε μια καθορισμένη χρονική στιγμή.

Καθοριστικό ρόλο στην εκτίμηση με τη μέθοδο Kaplan-Meier έχουν τα διαστήματα τα οποία δημιουργούμε με τέτοιο τρόπο ώστε κάθε διάστημα να ξεκινάει σε μια χρονική στιγμή θανάτου-αποτυχίας και να περιέχει μόνο μια τέτοια χρονική στιγμή. Για να γίνει πιο ξεκάθαρος ο τρόπος κατασκευής αυτών των

διαστημάτων αρχικά θεωρούμε ένα απλουστευμένο παράδειγμα με τρεις χρονικές στιγμές θανάτου-αποτυχίας και στη συνέχεια θα γενικεύσουμε αυτήν την ιδέα (βλ. Collett, 2015). Θεωρούμε λοιπόν τις παρατηρούμενες χρονικές στιγμές θανάτων-αποτυχιών t_1, t_2, t_3 τις οποίες και διατάσσουμε σε αύξουσα σειρά ώστε: $t_{(1)} < t_{(2)} < t_{(3)}$, ενώ με c συμβολίζουμε το λογοκριμένο χρόνο επιβίωσης. Επομένως, τα διαστήματα που θα κατασκευαστούν ξεκινούν τις χρονικές στιγμές $t_{(1)}, t_{(2)}$ και $t_{(3)}$ και περιλαμβάνουν μόνο μια χρονική στιγμή θανάτου-αποτυχίας, ανεξάρτητα από τον αριθμό των θανάτων-αποτυχιών που συμβαίνουν εκείνη τη χρονική στιγμή. Παρακάτω δίνεται το σχήμα που δείχνει τον τρόπο που κατασκευάζονται τα διαστήματα στην περίπτωση που περιγράψαμε προηγουμένως.

Σχήμα 2.7: Κατασκευή χρονικών διαστημάτων



Απο το προηγούμενο σχήμα βλέπουμε πως η έναρξη της μελέτης γίνεται τη χρονική στιγμή t_0 και έτσι η αρχική περίοδος ξεκινάει από το t_0 και τελειώνει πριν τη χρονική στιγμή $t_{(1)}$, τη χρονική στιγμή αυτή παρατηρούνται οι πρώτοι θάνατοι-αποτυχίες και έτσι το πρώτο κατασκευασμένο διάστημα ξεκινάει τότε και τελειώνει ακριβώς πριν την επόμενη χρονική στιγμή θανάτου-αποτυχίας που είναι η $t_{(2)}$, δηλαδή είναι το $[t_{(1)}, t_{(2)})$. Όμοια το δεύτερο διάστημα είναι το $[t_{(2)}, t_{(3)})$ και περιέχει τη χρονική στιγμή θανάτου-αποτυχίας $t_{(2)}$ καθώς και τη χρονική στιγμή λογοκρισίας, c . Τέλος, υπάρχει και το χρονικό διάστημα που ξεκινάει τη χρονική στιγμή $t_{(3)}$ που περιέχει την μεγαλύτερη χρονική στιγμή στην οποία συμβαίνουν οι τελευταίοι θάνατοι-αποτυχίες. Αφού είδαμε πως κατασκευάζονται τα χρονικά διαστήματα είμαστε πλέον σε θέση να γενικεύσουμε την προηγούμενη κατασκευή.

Έστω λοιπόν n άτομα με παρατηρούμενους χρόνους επιβίωσης $t_1, t_2, \dots, t_n$. Με βάση τα προηγούμενα γίνεται αντιληπτό πως οι χρόνοι επιβίωσης των n ατόμων μπορεί να είναι λογοκριμένοι, ενώ είναι πιθανό κάποια άτομα να εμφανίζουν το υπό μελέτη χαρακτηριστικό την ίδια χρονική στιγμή. Έτσι μπορούμε να πούμε χωρίς βλάβη της γενικότητας πως υπάρχουν r διακεκριμένες χρονικές στιγμές θανάτου με $r \leq n$. Όπως και στην πιο απλή περίπτωση έτσι και τώρα διατάσσουμε αυτούς τους χρόνους σε αύξουσα σειρά ώστε: $t_{(1)} < t_{(2)} < \dots < t_{(r)}$. Συμβολίζουμε με n_j τον αριθμό των ατόμων που είναι ζωντανοί ακριβώς πριν τη χρονική στιγμή $t_{(j)}$, $j = 1, 2, \dots, r$ και με d_j τον αριθμό των ατόμων που πεθαίνουν τη χρονική στιγμή $t_{(j)}$. Επομένως, στο χρονικό διάστημα $(t_{(j)} - \delta, t_{(j)}]$, όπου δ μια απειροελάχιστη χρονική στιγμή, περιλαμβάνεται μια χρονική στιγμή θανάτου. Έτσι η πιθανότητα θανάτου σε αυτό το διάστημα εκτιμάται από την ποσότητα: $\frac{d_j}{n_j}$, ενώ η πιθανότητα επιβίωσης από την ποσότητα: $\frac{n_j - d_j}{n_j}$. Από τον τρόπο με τον οποίο κατασκευάζονται τα διαστήματα το χρονικό διάστημα $(t_{(j)}, t_{(j+1)} - \delta]$ δεν περιέχει καμία χρονική στιγμή θανάτου-αποτυχίας, άρα η πιθανότητα επιβίωσης σε αυτό το χρονικό διάστημα είναι 1. Τελικά, η από κοινού πιθανότητα επιβίωσης από τη χρονική στιγμή $t_{(j)} - \delta$ στη χρονική στιγμή $t_{(j)}$ και από τη χρονική στιγμή $t_{(j)}$ ως τη χρονική στιγμή $t_{(j+1)} - \delta$ είναι $\frac{n_j - d_j}{n_j}$. Δεδομένου τώρα ότι το δ είναι μια απειροελάχιστη ποσότητα, μπορούμε να πούμε πως η προηγούμενη ποσότητα είναι ουσιαστικά η πιθανότητα επιβίωσης στο διάστημα $[t_{(j)}, t_{(j+1)})$. Τέλος, υπό την υπόθεση πως οι αποτυχίες-θάνατοι στο δείγμα είναι ανεξάρτητοι μεταξύ τους, έχουμε ότι η εκτιμώμενη συνάρτηση επιβίωσης σε μια χρονική στιγμή t που βρίσκεται στο κ -κατασκευασμένο χρονικό διάστημα, $\kappa = 1, 2, \dots, r$ θα είναι η εκτιμώμενη πιθανότητα επιβίωσης μετά τη χρονική στιγμή $t_{(k)}$, δηλαδή η πιθανότητα επιβίωσης στο διάστημα $[t_{(k)}, t_{(k+1)})$, καθώς και σε όλα τα προηγούμενα διαστήματα. Επομένως, με βάση αυτά ορίστηκε από τους Kaplan and Meier

(1958) η εκτίμηση της συνάρτησης επιβίωσης να είναι η ποσότητα:

$$\hat{S}(t) = \prod_{j=1}^k \frac{n_j - d_j}{n_j} \quad (2.4)$$

για $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, 2, \dots, r$. Σημειώνουμε ότι:

- $\hat{S}(t) = 1$, για $t < t_{(1)}$.
- $\hat{S}(t) \rightarrow +\infty$, για $t \geq t_{(r+1)}$.

Ο εκτιμητής $\hat{S}(t)$ της συνάρτησης επιβίωσης ονομάζεται εκτιμητής Kaplan-Meier.

Το τυπικό σφάλμα του εκτιμητή Kaplan-Meier

Μια συνηθισμένη διαδικασία στη στατιστική μετά την εύρεση ενός εκτιμητή είναι η εύρεση της σχέσης από την οποία προκύπτει η διακύμανση και το τυπικό σφάλμα του συγκεκριμένου εκτιμητή. Με βάση τον εκτιμητή που δόθηκε προηγουμένως για την συνάρτηση επιβίωσης μπορούμε τώρα να γράψουμε ότι:

$$\hat{S}(t) = \prod_{j=1}^k \hat{p}_j, \quad k = 1, 2, \dots, r$$

όπου $\hat{p}_j = \frac{n_j - d_j}{n_j}$, τότε

$$\log(\hat{S}(t)) = \sum_{j=1}^k \log(\hat{p}_j) \xrightarrow{\text{ανεξαρτησία}} \text{Var}(\log(\hat{S}_t)) = \sum_{j=1}^k \text{Var}(\log(\hat{p}_j)).$$

Παρατηρούμε επιπλέον πως ο αριθμός των ατόμων που επιβιώνουν στο διάστημα που ξεκινάει την $t_{(j)}$ χρονική στιγμή ακολουθεί διωνυμική κατανομή με παραμέτρους n_j και p_j . Ο αριθμός των ατόμων που επιβιώνουν σε αυτό το διάστημα

είναι $n_j - d_j$ επομένως,

$$\begin{aligned} \text{Var}(n_j - d_j) &= n_j p_j (1 - p_j) \Rightarrow \frac{1}{n_j^2} \text{Var}(n_j - d_j) = \frac{p_j(1 - p_j)}{n_j} \\ \Rightarrow \text{Var}(\hat{p}_j) &= \frac{p_j(1 - p_j)}{n_j} \end{aligned}$$

Έτσι η διακύμανση του $\hat{p}_j$ εκτιμάται από την ποσότητα:

$$\widehat{\text{Var}}(\hat{p}_j) = \frac{\hat{p}_j(1 - \hat{p}_j)}{n_j} \quad (2.5)$$

Προκειμένου να υπολογίσουμε την διακύμανση της ποσότητας $\log(\hat{p}_j)$ θα χρησιμοποιηθεί ένα γενικό συμπέρασμα για την προσέγγιση της διακύμανσης μιας συνάρτησης μιας τυχαίας μεταβλητής σύμφωνα με το οποίο η διακύμανση της $g(X)$, όπου X μια τυχαία μεταβλητή, δίνεται από τη σχέση

$$\text{Var}(g(X)) \approx \left(\frac{dg(X)}{dX} \right)^2 \text{Var}(X). \quad (2.6)$$

Η μέθοδος αυτή είναι γνωστή ως δέλτα μέθοδος (*delta method*) (Kelley, 1928).

Με βάση αυτά

$$\text{Var}(\log(\hat{p}_j)) \approx \frac{1}{\hat{p}_j^2} \text{Var}(\hat{p}_j)$$

Έτσι η διακύμανση της ποσότητας $\log(\hat{p}_j)$ μπορεί να εκτιμηθεί από τη σχέση

$$\text{Var}(\log(\hat{p}_j)) \approx \frac{d_j}{n_j(n_j - d_j)}$$

$$\text{Επομένως, } \text{Var}(\log(\hat{S}_t)) = \sum_{j=1}^k \text{Var}(\log(\hat{p}_j)) \approx \sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)}.$$

Όμως και πάλι από τη μέθοδο δέλτα (delta method) έχουμε ότι:

$$Var(\log(\hat{S}(t))) = \frac{1}{\hat{S}(t)^2} Var(\hat{S}(t)) \Rightarrow Var(\hat{S}(t)) = \hat{S}(t)^2 \sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)} \quad (2.7)$$

ενώ επιπλέον,

$$se(\hat{S}(t)) = \hat{S}(t) \left[\sum_{j=1}^k \frac{d_j}{n_j(n_j - d_j)} \right]^{1/2}, \quad t_{(k)} \leq t < t_{(k+1)} \quad (2.8)$$

Οι προηγούμενες σχέσεις για την διακύμανση και το τυπικό σφάλμα της εκτιμήτριας Kaplan-Meier είναι γνωστές ως τύποι του Greenwood (Greenwood, 1926).

2.4 Συγκριση καμπυλών επιβίωσης

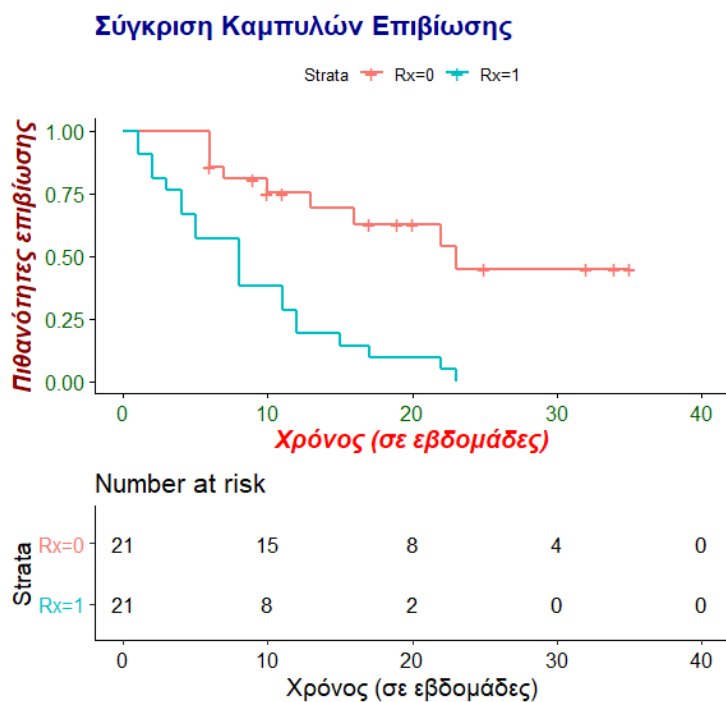
Ο πιο εύκολος τρόπος για να συγκρίνουμε τους χρόνους επιβίωσης (survival times) των ατόμων που βρίσκονται σε δύο group είναι να σχεδιάσουμε τις αντίστοιχες εκτιμώμενες καμπύλες επιβίωσης, σε ένα κοινό γράφημα. Το γράφημα αυτό μπορεί να μας δώσει αρκετές πληροφορίες σχετικά με τους χρόνους επιβίωσης. Ένα τέτοιο παράδειγμα δίνεται στη συνέχεια. Το σύνολο δεδομένων δόθηκε πρώτη φορά στην εργασία των Freinreich et al. (1963) και περιλαμβάνει συνολικά 42 παρατηρήσεις. Έτσι, λοιπόν στα 42 άτομα που έπασχαν από λευχαιμία δημιουργήθηκαν δυο group θεραπειών. Στο ένα group χορηγήθηκε ένα συμβατικό φάρμακο (treatment group), ενώ στο άλλο group χορηγήθηκε μια placebo θεραπεία. Οι ερευνητές ήθελαν να μελετήσουν το χρόνο κατά τον οποίο η ασθένεια βρισκόταν σε ύφεση. Έτσι ως αποτυχία θεωρείται η διακοπή του χρονικού διαστήματος ύφεσης που συνεπάγεται το θάνατο του ατόμου. Το συγκεκριμένο σύνολο δεδομένων θα χρησιμοποιηθεί για όλες τις πρακτικές

εφαρμογές των θεωρητικών συμπερασμάτων αυτού του κεφαλαίου. Παρακάτω δίνουμε μια περιγραφή των μεταβλητών που υπάρχουν στο σύνολο δεδομένων. Όλες οι στατιστικές αναλύσεις που θα παρουσιαστούν σε αυτό και σε επόμενα κεφάλαια έγιναν με τη βοήθεια του στατιστικού πακέτου της **R**.

- **ID**: αύξοντας αριθμός ατόμου
- **time**: χρόνος επιβίωσης (ύφεση της ασθένειας) σε εβδομάδες
- **status**: μεταβλητή που υποδεικνύει τη λογοκρισία (censoring)
(0=censored, 1=αποτυχία (failure))
- **Rx**: μεταβλητή που υποδεικνύει το είδος της θεραπείας
(0=συμβατική θεραπεία, 1=placebo)
- **logWBC**: συνεχής μεταβλητή που υποδεικνύει την αριθμητική τιμή που προέκυψε από τη μέτρηση των λευκών αιμοσφαιρίων σε λογαριθμική κλίμακα.

Στο σχήμα 2.8 δίνονται οι καμπύλες Kaplan-Meier σε ένα κοινό γράφημα, ενώ ο δεύτερος πίνακας δίνει το σύνολο ρίσκου σε κάθε χρονική στιγμή της έρευνας. Βλέπουμε ότι η καμπύλη επιβίωσης για το group των ατόμων στα οποία εφαρμόστηκε η συμβατική θεραπεία ($Rx=0$) είναι υψηλότερη σε σχέση με αυτή των ατόμων που έλαβαν θεραπεία placebo ($Rx=1$). Έτσι, προκύπτει ότι τα άτομα που έλαβαν συμβατική θεραπεία έχουν καλύτερη πρόγνωση επιβίωσης σε σχέση με τα άτομα που ανήκαν στο placebo group.

Σχήμα 2.8: Σύγκριση καμπύλων επιβίωσης, όπου με + παριστάνονται οι λογοκριμένες χρονικές στιγμές



Υπάρχουν δύο λόγοι για τους οποίους είναι πιθανό να παρατηρηθούν διαφορές σε καμπύλες επιβίωσης, όπως αυτές στο προηγούμενο παράδειγμα. Ο πρώτος λόγος για τον οποίο μπορούμε να παρατηρήσουμε διαφορές είναι ότι όντως υπάρχουν αυτές και οι χρόνοι επιβίωσης στα επιμέρους group διαφέρουν. Ο δεύτερος λόγος για τον οποίο μπορεί να παρατηρήσουμε διαφορές είναι η τυχαίότητα η οποία διέπει τα παρατηρούμενα αποτελέσματα, έτσι είναι πιθανό να παρατηρούνται διαφορές ενώ στην πραγματικότητα δεν υπάρχουν. Έτσι προκειμένου να ενισχυθούν τα συμπεράσματά μας κρίνεται απαραίτητη η δημιουργία ενός στατιστικού test για να ελέγχεται η υπόθεση της ταύτισης ή όχι των καμπυλών επιβίωσης. Ένας τέτοιος στατιστικός έλεγχος μπορεί να πραγματοποιηθεί με την εφαρμογή

του *logrank test* που προτάθηκε στη βιβλιογραφία από τους Mantel and Haenszel (1959).

2.4.1 Logrank Test

Προκειμένου να κατασκευαστεί ένα στατιστικό test για τη σύγκριση δυο καμπυλών επιβίωσης θεωρούμε ξεχωριστούς τους χρόνους θανάτου στα δύο group. Έστω ότι αυτά ονομάζονται Group I και Group II. Υποθέτουμε επίσης ότι υπάρχουν r ξεχωριστές στιγμές θανάτου οι οποίες είναι και πάλι διατεταγμένες. Επιπλέον, υποθέτουμε ότι τη χρονική στιγμή $t_{(j)}$ στο Group I συμβαίνουν συνολικά d_{1j} γεγονότα, ενώ στο Group II για την ίδια χρονική στιγμή συμβαίνουν d_{2j} γεγονότα. Τέλος, συμβολίζουμε με n_{1j} το συνολικό αριθμό ατόμων που βρίσκονται στο Group I ακριβώς πριν τη χρονική στιγμή $t_{(j)}$ και n_{2j} το συνολικό αριθμό ατόμων που βρίσκονται στο Group II ακριβώς πριν τη χρονική στιγμή $t_{(j)}$. Συνολικά, τη χρονική στιγμή $t_{(j)}$ συμβαίνουν και στα δύο group $d_j = d_{1j} + d_{2j}$ γεγονότα, ενώ πριν τη χρονική στιγμή αυτή και στα δύο group βρίσκονται $n_j = n_{1j} + n_{2j}$ άτομα.

Θεωρούμε τη μηδενική υπόθεση πως δεν υπάρχει στατιστικά σημαντική διαφορά ανάμεσα στις καμπύλες επιβίωσης των δύο group. Ένας τρόπος για να εξετάσουμε την μηδενική υπόθεση είναι να θεωρήσουμε τη διαφορά του παρατηρούμενου και του αναμενόμενου αριθμού ατόμων που εμφανίζουν το υπό μελέτη γεγονός σε κάθε χρονική στιγμή. Έστω O_{1j} η τυχαία μεταβλητή που παριστά το πλήθος των θανάτων στο Group I. Τότε υπό την μηδενική υπόθεση η τυχαία μεταβλητή θα περιγράφεται ικανοποιητικά από την υπεργεωμετρική κατανομή,

δηλαδή $O_{1j} \sim Hg(d_j, n_j - d_j, n_{1j})$. Έτσι,

$$P(O_{1j} = d_{1j}) = \frac{\binom{d_j}{d_{1j}} \binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}} \quad (2.9)$$

Εύκολα προκύπτει ότι ο αναμενόμενος αριθμός ατόμων που εμφανίζουν το υπό μελέτη γεγονός στο Group I τη χρονική στιγμή $t_{(j)}$ είναι $E(O_{1j}) = \frac{n_{1j}d_j}{n_j} = e_{1j}$. Προκειμένου να συνδυάσουμε όλες τις χρονικές στιγμές θανάτου και έτσι να έχουμε ένα συνολικό μέτρο για την απόκλιση μεταξύ παρατηρούμενων και προβλεπόμενων τιμών θεωρούμε την ποσότητα

$$U_L = \sum_{j=1}^r (O_{1j} - e_{1j}) \quad (2.10)$$

Από τον τρόπο που ορίστηκε η προηγούμενη ποσότητα είναι ξεκάθαρο πως $E(U_L) = 0$, ενώ επιπλέον λόγω της ανεξαρτησίας των χρονικών στιγμών εμφάνισης του υπο μελέτη γεγονότος θα ισχύει ότι:

$$Var(U_L) = \sum_{j=1}^r Var(O_{1j}) = \frac{n_{1j}n_{2j}(n_j - d_j)}{n_j^2(n_j - 1)} \quad (2.11)$$

Επομένως, λόγω του κεντρικού οριακού θεωρήματος θα ισχύει ότι η ποσότητα $W_L = \frac{U_L}{\sqrt{V_L}} \sim N(0, 1) \Rightarrow \frac{U_L^2}{V_L} \sim X_1^2$. Το στατιστικό test που κατασκευάστηκε χρησιμοποιείται για την εφαρμογή του log-rank test. Μεγάλες τιμές του στατιστικού αποτελούν ένδειξη για απόρριψη της μηδενικής υπόθεσης. Επιπρόσθετα, αφού $W \sim X_1^2$ η τιμή p-value που συνδέεται με το στατιστικό έλεγχο μπορεί να βρεθεί από τους πίνακες της X^2 κατανομής.

Στον επόμενο πίνακα δίνονται τα αποτελέσματα που προέκυψαν από την εφαρμογή του logrank test για το σύνολο δεδομένων που περιγράφηκε προηγουμένως. Η προσαρμογή αυτή μας έδειξε ότι η υπόθεση πως οι καμπύλες επιβίωσης ταυτίζονται απορρίπτεται σε επίπεδο σημαντικότητας 5%. Το συμπέρασμα επιβεβαιώνει

αυτό που παρατηρήσαμε και στην αντίστοιχη γραφική σύγκριση των καμπυλών επιβίωσης.

	N	Observed	Expected	$(O-E)^2/V$
$Rx = 0$	21	9	19.3	16.8
$Rx = 1$	21	21	10.7	16.8

Chisq=16,8 on 1 degrees of freedom p-value=4e-05

2.5 Το μοντέλο παλινδρόμησης του Cox

Οι μη-παραμετρικές μέθοδοι που παρουσιάστηκαν στις προηγούμενες ενότητες είναι πολύ χρήσιμες όταν θέλουμε να συγκρίνουμε τις καμπύλες επιβίωσης σε δύο group. Ωστόσο, στις περισσότερες μελέτες υπάρχουν καταγεγραμμένες περισσότερες από μια μεταβλητές οι οποίες μπορούν να επηρεάσουν το χρόνο επιβίωσης και ονομάζονται ανεξάρτητες. Οι μεταβλητές αυτές είναι απαραίτητο να ληφθούν υπόψη και κάτι τέτοιο επιτυγχάνεται μέσω της κατασκευής ενός μοντέλου παλινδρόμησης. Η κλασική θεωρία των μοντέλων αδυνατεί να μοντελοποιήσει την περίπτωση των δεδομένων επιβίωσης, αφού σε αυτήν την περίπτωση είναι αναγκαία η κατασκευή ενός μοντέλου το οποίο θα λαμβάνει υπόψη και την λογοκρισία (censoring) που πιθανότητα υπάρχει στους χρόνους επιβίωσης. Ένα πολύ γνωστό μοντέλο παλινδρόμησης που μπορεί να μοντελοποιεί ικανοποιητικά τέτοιες περιπτώσεις είναι το μοντέλο παλινδρόμησης του Cox (Cox regression model) που προτάθηκε από τον Cox (1972).

Μια θεμελιώδης συνάρτηση για την μοντελοποίηση προβλημάτων ανάλυσης επιβίωσης είναι η συνάρτηση κινδύνου (hazard function) η οποία ορίστηκε προγενέστερα. Ως εκ τούτου, οι τιμές αυτής της συνάρτησης είναι αυτές που

προσπαθεί να εκτιμήσει το μοντέλο του Cox με τη χρήση του κατάλληλου συνδυασμού από επεξηγηματικές μεταβλητές. Επιπλέον, έχουν δοθεί σχέσεις που συνδέουν τη συνάρτηση κινδύνου με τις υπόλοιπες θεμελιώδεις συναρτήσεις της ανάλυσης επιβίωσης και έτσι η εύρεση εκτιμώμενων τιμών για αυτήν συνεπάγεται και την εύρεση εκτιμώμενων τιμών για τις υπόλοιπες συναρτήσεις, υπό την επίδραση της προσαρμογής του μοντέλου.

Έστω λοιπόν ότι ο κίνδυνος θανάτου (hazard of death) σε μια συγκεκριμένη χρονική στιγμή εξαρτάται από τις τιμές $x_1, x_2, \dots, x_p$ των p επεξηγηματικών μεταβλητών $X_1, X_2, \dots, X_p$. Βασική προϋπόθεση είναι οι μεταβλητές αυτές να έχουν καταγραφεί στην αρχή της μελέτης, διαφορετικά θα πρέπει να χρησιμοποιηθεί η επέκταση του μοντέλου του Cox για χρονοεξαρτώμενες μεταβλητές (Cox regression model with time dependent variables). Η επέκταση αυτή δεν αφορά άμεσα τους σκοπούς αυτής της μεταπτυχιακής διατριβής και έτσι δεν θα γίνει κάποια περαιτέρω αναφορά σε αυτή. Συμβολίζουμε με x το διάνυσμα που περιέχει τις τιμές των ανεξάρτητων μεταβλητών, δηλαδή $x = (x_1, x_2, \dots, x_p)^T$. Θεωρούμε επίσης την συνάρτηση $h_0(t)$, που είναι η συνάρτηση κινδύνου για ένα άτομο που είναι τέτοιο ώστε όλες οι τιμές των επεξηγηματικών μεταβλητών που συνθέτουν το διάνυσμα x να είναι 0. Η συνάρτηση αυτή είναι γνωστή ως *αναφορική συνάρτηση κινδύνου* (baseline hazard function). Επομένως, η συνάρτηση κινδύνου για το i -οστό άτομο γράφεται ως:

$$h_i(t) = \psi(x_i)h_0(t) \quad (2.12)$$

όπου $\psi(x_i)$ είναι μια συνάρτηση του x_i , που είναι το διάνυσμα των επεξηγηματικών μεταβλητών (explanatory variables) για το i -οστό άτομο με συνιστώσες $x_{1i}, x_{2i}, \dots, x_{pi}$. Παρατηρούμε ότι ο σχετικός κίνδυνος $\psi(x_i)$ δεν μπορεί να είναι αρνητικός. Επομένως, η επιλογή της μορφής αυτής της συνάρτησης περιορίζεται στις μη-αρνητικές. Έτσι λοιπόν ο Cox (1972) θεώρησε πως μια λογική επιλογή

για την συνάρτηση $\psi(x_i)$ είναι η

$$\psi(x_i) = \exp\{\eta_i\}$$

όπου η_i είναι η γραμμική προβλέπουσα (linear predictor) , δηλαδή

$$\eta_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} = \beta^T X, \beta = (\beta_1, \beta_2, \dots, \beta_p)^T.$$

Συχνά στη βιβλιογραφία που σχετίζεται με την ανάλυση επιβίωσης η γραμμική προβλέπουσα η_i ονομάζεται και προγνωστικός δείκτης (prognostic index).

Έτσι η μορφή του μοντέλου που εισήχθη από τον Cox (1972) είναι

$$h_i(t) = \exp\{\beta^T x_i\} h_0(t). \quad (2.13)$$

Εύκολα από το μοντέλο (2.13) μπορούμε να μεταβούμε στο μοντέλο

$$\log \left(\frac{h_i(t)}{h_0(t)} \right) = \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} \quad (2.14)$$

και έτσι από το μοντέλο του Cox μπορούμε να μεταβούμε σε ένα γραμμικό μοντέλο με εξαρτημένη μεταβλητή το λογάριθμο του λόγου κινδύνου (hazard ratio) $\frac{h_i(t)}{h_0(t)}$. Στο μοντέλο του Cox παρατηρείται για τη συνάρτηση κινδύνου η εξής ιδιότητα: *οι συναρτήσεις κινδύνου των ατόμων είναι ανάλογες μεταξύ τους και ο λόγος των συναρτήσεων κινδύνου δύο ατόμων είναι ανεξάρτητος του χρόνου.* Η αναλογικότητα μεταξύ των συναρτήσεων κινδύνου είναι μια από τις βασικές υποθέσεις του μοντέλου του Cox και για αυτό το λόγο το μοντέλο αυτό είναι επίσης γνωστό ως μοντέλο των ανάλογων κινδύνων του Cox (Cox proportional hazard model).

Βασικός λόγος για τον οποίο το μοντέλο του Cox χρησιμοποιείται σε πολλές εφαρμογές, ιδιαίτερα στον κλάδο της βιοστατιστικής, αποτελεί το γεγονός ότι για την κατασκευή του δεν γίνεται κάποια συγκεκριμένη υπόθεση σχετικά με την κατανομή της αναφορικής συνάρτησης κινδύνου (baseline function) $h_0(t)$, έτσι

το μοντέλο αυτό είναι ένα ημι-παραμετρικό μοντέλο (semi-parametric model). Αξίζει να σημειωθεί ότι στη βιβλιογραφία έχουν προταθεί μοντέλα στα οποία η αναφορική συνάρτηση κινδύνου ακολουθεί κάποια συγκεκριμένη κατανομή. Αυτά τα μοντέλα ονομάζονται παραμετρικά μοντέλα ανάλογων κινδύνων (parametric proportional hazards models). Ωστόσο, σε πραγματικά προβλήματα οι περιπτώσεις στις οποίες οι ερευνητές είναι σε θέση να γνωρίζουν την κατανομή που περιγράφει το χρόνο επιβίωσης είναι περιορισμένες. Έτσι η δημοφιλία του μοντέλου του Cox πηγάζει από το γεγονός ότι μας δίνει τη δυνατότητα να κάνουμε μια ασφαλή επιλογή μοντέλου χωρίς να ανησυχούμε για το εάν έχουμε κάνει τη σωστή επιλογή της κατανομής, κάτι που είναι απαραίτητο στις περιπτώσεις των παραμετρικών μοντέλων.

2.5.1 Προσαρμογή του μοντέλου του Cox

Εφόσον καταλήξαμε στη μορφή που θα έχει το μοντέλο, το επόμενο βήμα αφορά την εκτίμηση των παραμέτρων β της παλινδρόμησης. Η μέθοδος που χρησιμοποιείται για την εκτίμηση είναι και πάλι η μέθοδος της μέγιστης πιθανοφάνειας. Έστω λοιπόν ότι υπάρχουν n άτομα και r διακριτοί χρόνοι αποτυχιών-θανάτων, ενώ υπάρχουν συνολικά $n - r$ δεξιά λογοκριμένες χρονικές στιγμές επιβίωσης. Οι r χρονικές στιγμές θεωρούνται και πάλι διατεταγμένες, δηλαδή $t_{(1)} < t_{(2)} < \dots < t_{(r)}$. Το σύνολο των ατόμων που δεν έχουν εμφανίσει το υπό μελέτη χαρακτηριστικό και δεν έχουν λογοκριμένο χρόνο επιβίωσης ακριβώς πριν τη χρονική στιγμή $t_{(j)}$ συμβολίζεται με $R_{(t_{(j)})}$ και ονομάζεται ως σύνολο ρίσκου (risk set). Ο Cox (1972) απέδειξε πως η κατάλληλη συνάρτηση πιθανοφάνειας

δίνεται από τη σχέση:

$$L(\beta) = \prod_{j=1}^r \frac{\exp\{\beta^T x_{(j)}\}}{\sum_{l \in R(t_{(j)})} \exp\{\beta^T x_l\}} \quad (2.15)$$

όπου $x_{(j)}$ είναι το διάνυσμα των ανεξάρτητων μεταβλητών για το άτομο που εμφανίζει το γεγονός την χρονική στιγμή $t_{(j)}$. Στον παρονομαστή της συνάρτησης πιθανοφάνειας βρίσκεται το άθροισμα των τιμών της $\exp\{\beta^T x\}$ υπεράνω όλων των ατόμων που βρίσκονται στο σύνολο κινδύνου την χρονική στιγμή $t_{(j)}$. Παρατηρούμε επιπλέον ότι το γινόμενο υπολογίζεται για τις χρονικές στιγμές αποτυχίας που είναι πλήρως καταγεγραμμένες και όχι για λογοκριμένες χρονικές στιγμές. Η πληροφορία για τα άτομα με λογοκριμένο χρόνο επιβίωσης υπεισέρχεται στη συνάρτηση πιθανοφάνειας μέσω του παρονομαστή στον οποίο βρίσκεται το άθροισμα υπεράνω των συνόλων κινδύνου τα οποία προκύπτουν πριν από την λογοκριμένη χρονική στιγμή.

Η συνάρτηση πιθανοφάνειας (2.15) δεν είναι η κλασική συνάρτηση πιθανοφάνειας αφού το γινόμενο δεν υπολογίζεται υπεράνω όλων των χρονικών στιγμών αλλά μονό για τις χρονικές στιγμές που υπάρχει παρατηρούμενο γεγονός. Η συνάρτηση αυτή είναι γνωστή στη βιβλιογραφία ως *μερική συνάρτηση πιθανοφάνειας* (*partial likelihood function*) και προτάθηκε και αυτή από τον Cox (1972). Η συνάρτηση μερικής πιθανοφάνειας εξαρτάται μόνο από τη σειρά των χρονικών στιγμών αποτυχιών-θανάτων και επομένως τα συμπεράσματα για την επίδραση των επεξηγηματικών μεταβλητών στην συνάρτηση κινδύνου εξαρτώνται μόνο από τους ταξινομημένους χρόνους επιβίωσης. Τέλος, σημειώνουμε πως ο εκτιμητής προκύπτει κανονικά από τη μεγιστοποίηση της συνάρτησης μερικής πιθανοφάνειας, ενώ σχετικά με την μεγιστοποίηση αντιμετωπίζουμε την συνάρτησης μερικής πιθανοφάνειας όπως την αντίστοιχη κλασική συνάρτηση. Στην ουσία συμπεριφερόμαστε σε αυτήν τη συνάρτηση σαν να είναι μια συνηθισμένη συνάρ-

τηση πιθανοφάνειας, ενώ ως επαναληπτική διαδικασία μπορεί χρησιμοποιηθεί ο αλγόριθμος Newton-Raphson που περιγράφηκε στο προηγούμενο κεφάλαιο. Ο Cox απέδειξε σε εργασία του (Cox, 1975) πως ο εκτιμητής που προκύπτει από αυτή τη μεγιστοποίηση διατηρεί όλες τις ασυμπτωτικές ιδιότητες που έχει ο κλασικός εκτιμητής μέγιστης πιθανοφάνειας. Στις ασυμπτωτικές ιδιότητες του εκτιμητή μέγιστης πιθανοφάνειας θα γίνει αναφορά στο επόμενο κεφάλαιο.

Κατασκευή της συνάρτησης πιθανοφάνειας

Όπως ήδη αναφέρθηκε, στο μοντέλο του Cox ο κίνδυνος θανάτου στη χρονική στιγμή t για το i -οστό άτομο δίνεται από τη σχέση:

$$h_i(t) = \exp\{\beta^T x_i\} h_0(t), \quad i = 1, 2, \dots, n$$

Το βασικό επιχείρημα για την κατασκευή της συνάρτησης πιθανοφάνειας αυτού του μοντέλου είναι πως τα διαστήματα ανάμεσα σε διαδοχικές χρονικές στιγμές θανάτων, δεν περιέχουν καμία πληροφορία σχετικά με την επίδραση των επεξηγηματικών μεταβλητών στη συνάρτηση κινδύνου (hazard function). Τα μη-πληροφοριακά διαστήματα προκύπτουν αφού η αναφορική συνάρτηση κινδύνου (baseline hazard function) είναι πιθανό να μηδενίζεται σε αυτά τα διαστήματα που δεν περιλαμβάνουν αποτυχίες-θανάτους. Έτσι, έχει νόημα να θεωρήσουμε την πιθανότητα το i -άτομο να εμφανίσει το γεγονός την $t_{(j)}$ χρονική στιγμή δθέντος ότι η $t_{(j)}$ στιγμή είναι μια από τις r παρατηρούμενες χρονικές στιγμές θανάτων, δηλαδή την:

$$P(\text{το άτομο με μεταβλητές } x_{(j)} \text{ πεθαίνει την στιγμή } t_{(j)} \mid \text{ένας θάνατος την } t_{(j)}).$$

Από τον ορισμό της δεσμευμένης πιθανότητας η προηγούμενη πιθανότητα ισούται με

$$\frac{P(\text{το άτομο με μεταβλητές } x_{(j)} \text{ πεθαίνει την στιγμή } t_{(j)})}{P(\text{ένας θάνατος την } t_{(j)})}$$

Λόγω της ανεξαρτησίας ο παρονομαστής του προηγούμενου κλάσματος είναι το άθροισμα των πιθανοτήτων θανάτου την $t_{(j)}$ χρονική στιγμή για όλα τα άτομα που βρίσκονται στο σύνολο κινδύνου εκείνη την χρονική στιγμή. Έστω l τα άτομα αυτά, τότε η τελευταία σχέση γράφεται ως:

$$\frac{P(\text{το άτομο με μεταβλητές } x_{(j)} \text{ πεθαίνει την στιγμή } t_{(j)})}{\sum_{l \in R(t_{(j)})} P(\text{το άτομο } l \text{ πεθαίνει την στιγμή } t_{(j)})}$$

Αν τώρα στην τελευταία σχέση αντικαταστήσουμε την πιθανότητα θανάτου την χρονική στιγμή $t_{(j)}$ με την πιθανότητα θανάτου στο διάστημα $(t_{(j)}, t_{(j)} + \delta t)$ και διαιρέσουμε αριθμητή και παρονομαστή με δt , τότε παίρνοντας το όριο για $\delta t \rightarrow 0$ η προηγούμενη σχέση γράφεται ως:

$$\frac{h_i(t_j)}{\sum_{l \in R(t_{(j)})} h_l(t_{(j)})} = \frac{h_0 \exp\{\beta^T x_{(j)}\}}{\sum_{l \in R(t_{(j)})} h_0 \exp\{\beta^T x_l\}} = \frac{\exp\{\beta^T x_{(j)}\}}{\sum_{l \in R(t_{(j)})} \exp\{\beta^T x_l\}}$$

Τέλος, παίρνοντας το γινόμενο υπεράνω των r χρονικών στιγμών θανάτων έχουμε την συνάρτηση πιθανοφάνειας (2.15)

$$L(\beta) = \prod_{j=1}^r \frac{\exp\{\beta^T x_{(j)}\}}{\sum_{l \in R(t_{(j)})} \exp\{\beta^T x_l\}}$$

Εκτίμηση της αναφορικής συνάρτησης κινδύνου

Η εκτιμώμενη συνάρτηση κινδύνου δίνεται από τη σχέση

$$\hat{h}_i(t) = \exp\{\hat{\beta}^T x_i\} \hat{h}_0(t). \quad (2.16)$$

Ένας εκτιμητής της αναφορικής συνάρτησης κινδύνου δόθηκε από τους Kalbfleisch and Prentice (1973). Σύμφωνα με την εργασία τους η εκτιμώμενη αναφορική συνάρτηση κινδύνου την $t_{(j)}$ χρονική στιγμή δίνεται από τη σχέση: $\hat{h}_0(t_{(j)}) = 1 - \hat{q}_j$, όπου q_j είναι η λύση της εξίσωσης

$$\sum_{l \in D(t_{(j)})} \frac{\exp\{\beta^T x_l\}}{1 - \hat{q}_j \exp\{\hat{\beta}^T x_l\}} = \sum_{l \in R(t_{(j)})} \exp\{\hat{\beta}^T x_l\}, \quad j = 1, 2, \dots, r$$

όπου $D(t_{(j)})$ είναι το σύνολο των d_j ατόμων που εμφανίζουν το γεγονός την $t_{(j)}$ χρονική στιγμή. Η λύση της εξίσωσης είναι η

$$\hat{q}_j = \left(1 - \frac{\exp\{\hat{\beta}^T x_{(j)}\}}{\sum_{l \in R(t_{(j)})} \exp\{\hat{\beta}^T x_l\}} \right)^{\exp\{-\hat{\beta}^T x_l\}} \quad j = 1, 2, \dots, r \quad (2.17)$$

όπου $x_{(j)}$ είναι το διάνυσμα των επεξηγηματικών μεταβλητών για το άτομο που εμφανίζει το γεγονός την $t_{(j)}$ χρονική στιγμή.

Η ποσότητα $\hat{q}_j$ μπορεί να χρησιμοποιηθεί για την εκτίμηση της πιθανότητας ένα άτομο να επιβιώνει στο διάστημα από την $t_{(j)}$ χρονική στιγμή μέχρι την $t_{(j+1)}$ χρονική στιγμή. Επομένως, η αναφορική συνάρτηση επιβίωσης (baseline survival function) εκτιμάται από τη σχέση:

$$\hat{S}_0(t) = \prod_{j=1}^k \hat{q}_j, \quad j = 1, 2, \dots, r \quad (2.18)$$

για $t_{(k)} \leq t < t_{(k+1)}$, $k = 1, 2, \dots, r - 1$.

Επιπλέον, αφού γνωρίζουμε ότι γενικά ισχύει ότι $H(t) = -\log(S(t))$, μπορούμε να έχουμε μια εκτίμηση για την αναφορική αθροιστική συνάρτηση κινδύνου (baseline cumulative hazard function):

$$\hat{H}_0(t) = -\log(\hat{S}_0(t)) = -\sum_{j=1}^k \log \hat{q}_j. \quad (2.19)$$

Παίρνοντας το ολοκλήρωμα στην σχέση (2.16) έχουμε ότι:

$$\hat{H}_i(t) = \exp\{\hat{\beta}^T x_i\} \hat{H}_0(t), \quad i = 1, 2, \dots, n, \quad (2.20)$$

ενώ εφαρμόζοντας τη σχέση που συνδέει την συνάρτηση επιβίωσης με την αναφορική αθροιστική συνάρτηση κινδύνου έχουμε ότι:

$$\hat{S}_i(t) = \left(\hat{S}_0(t) \right)^{\exp\{\hat{\beta}^T x_i\}}, \quad i = 1, 2, \dots, n \quad (2.21)$$

για $t_{(k)} \leq t < t_{(k+1)}$.

Στη συνέχεια παρουσιάζονται τα αποτελέσματα που προέκυψαν από την προσαρμογή του μοντέλου του Cox για το σύνολο δεδομένων που ήδη χρησιμοποιήθηκε προηγουμένως. Οι ανεξάρτητες μεταβλητές είναι η κατηγορική μεταβλητή Rx που υποδεικνύει το είδος της θεραπείας και η συνεχής logWBC την αριθμητική τιμή που προέκυψε από τη μέτρηση των λευκών αιμοσφαιρίων, σε λογαριθμική κλίμακα. Τα κύρια αποτελέσματα που μας έδωσε αυτή η προσαρμογή είναι πως:

1. Η αύξηση της μεταβλητής logWBC κατά μια μονάδα αυξάνει κατά 1.386 φορές τον λογάριθμο της συνάρτησης κινδύνου.
2. Το να ανήκει κάποιος στην κατηγορία (1), δηλαδή να λαμβάνει placebo θεραπεία αυξάνει κατά 4 φορές τον κίνδυνο να παραμείνει σε ύφεση για μεγαλύτερο χρονικό διάστημα.
3. Το συγκεκριμένο μοντέλο εξηγεί το 67.1% της συνολικής μεταβλητότητας.

Πίνακας 2.1

<i>Προσαρμογή του μοντέλου του Cox</i>	
Rx (1)	1.386***
exp(Rx) (1)	3.999
se(Rx) (1)	(0.425)
z	3.26
p	0.0011
95% conf. interval	(1.739, 9.195)
logWBC	1.691***
exp(logWBC)	5.424
se(coef)	(0.336)
z	5.03
p	4.8e-07
95% conf. interval	(2.808, 10.478)
Observations	42
R ²	0.671
Max. Possible R ²	0.988
Log Likelihood	-69.828
Wald Test	33.600*** (df = 2)
LR Test	46.712*** (df = 2)
Score (Logrank) Test	46.068*** (df = 2)

Note:

* p<0.1; **p<0.05; *** p<0.01

2.6 Οι υποθέσεις του μοντέλου του Cox και ο έλεγχος τους

Για να ολοκληρωθεί η διαδικασία κατασκευής ενός στατιστικού μοντέλου και να έχουν νόημα τα αποτελέσματα που στη συνέχεια προκύπτουν από την προσαρμογή του, θα πρέπει να ελέγχουμε την ισχύ των υποθέσεων από τις οποίες διέπεται το εκάστοτε μοντέλο. Το μοντέλο του Cox που παρουσιάστηκε προηγουμένως διέπεται από 3 υποθέσεις.

1. Δεν υπάρχουν ακραίες τιμές (outliers).
2. Η σχέση μεταξύ του λογαρίθμου της συνάρτησης κινδύνου και των ανεξάρτητων μεταβλητών είναι γραμμική.
3. Η αναλογικότητα των κινδύνων ισχύει.

Η τελευταία υπόθεση αποτελεί ίσως και την πιο γνωστή υπόθεση του μοντέλου του Cox. Υπό αυτήν την υπόθεση, για την συνάρτηση επιβίωσης ενός ατόμου με διάνυσμα ανεξάρτητων μεταβλητών x_1 ισχύει ότι:

$$S(t, x_1) = [S_0(t)]^{\exp\{\beta^T x_1\}}.$$

Από την τελευταία σχέση εύκολα προκύπτει ότι:

$$\log[-\log(S(t, x_1))] = \beta^T x_1 + \log[-\log(S_0(t))]$$

Έστω τώρα ένα άλλο άτομο με διάνυσμα ανεξάρτητων μεταβλητών x_2 , τότε πάλι υπό την υπόθεση της αναλογικότητας των κινδύνων για αυτό το άτομο θα ισχύει ότι:

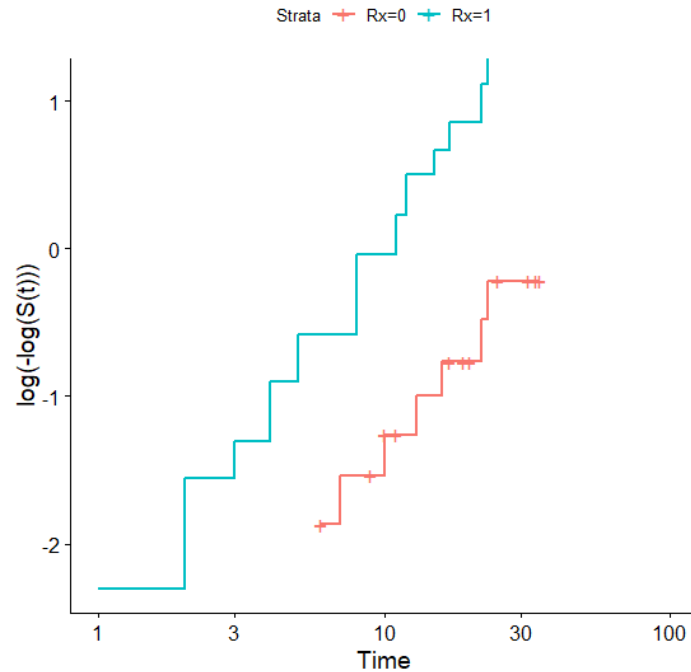
$$\log[-\log(S(t, x_2))] = \beta^T x_2 + \log[-\log(S_0(t))].$$

Αφαιρώντας τις δύο σχέσεις μεταξύ τους προκύπτει ότι:

$$\log[-\log(S(t, x_1))] - \log[-\log(S(t, x_2))] = \beta^T(x_1 - x_2) \quad (2.22)$$

Παρατηρούμε ότι στην σχέση (2.22) η αναφορική συνάρτηση έχει απλοποιηθεί, έτσι η διαφορά των log-log καμπυλών επιβίωσης δεν εξαρτάται από το χρόνο. Η σχέση αυτή προέκυψε υπό την υπόθεση της αναλογικότητας των κινδύνων, έτσι υπό αυτήν την υπόθεση οι γραφικές παραστάσεις των log-log καμπυλών επιβίωσης ως προς το χρόνο, είναι παράλληλες και απέχουν μια σταθερή απόσταση $b^T(x_1 - x_2)$. Αυτός είναι ένας πρώτος τρόπος ελέγχου για την υπόθεση της αναλογικότητας των κινδύνων. Από την εφαρμογή αυτού του κριτηρίου στο σύνολο δεδομένων που χρησιμοποιήθηκε σε αυτό το κεφάλαιο προέκυψε για την κατηγορική μεταβλητή Rx το επόμενο γράφημα.

Σχήμα 2.9: log-log καμπύλες για τον έλεγχο της αναλογικότητας των κινδύνων



Στο προηγούμενο γράφημα δεν παρουσιάζεται κάποια τομή μεταξύ των δύο καμπυλών, ενώ επιπλέον η μεταξύ τους απόσταση φαίνεται να είναι σταθερή. Ως εκ τούτου, συμπεραίνουμε πως η υπόθεση της αναλογικότητας των κινδύνων πληρείται για την μεταβλητή Rx. Βασικό μειονέκτημα της μεθόδου είναι η αδυναμία παρουσίασης αποτελεσμάτων για τις συνεχείς μεταβλητές. Στην περίπτωση κατά την οποία θέλουμε να χρησιμοποιήσουμε το συγκεκριμένο κριτήριο θα πρέπει πρώτα να κατηγοριοποιήσουμε την συνεχή μεταβλητή και στη συνέχεια να σχεδιάσουμε τις αντίστοιχες καμπύλες για κάθε κατηγορία. Η κατηγοριοποίηση όμως που μπορεί να γίνει δεν είναι μονοσήμαντα ορισμένη, συνεπώς είναι πιθανό για διαφορετικές κατηγοριοποιήσεις της ίδιας μεταβλητής να πάρουμε διαφορετικά γραφήματα και πιθανότατα και διαφορετικά συμπεράσματα σχετικά με την υπόθεση των αναλογικών κινδύνων.

Ένας άλλος τρόπος και ίσως ο πιο διαδεδομένος για τη διερεύνηση της υπόθεσης των αναλογικών κινδύνων (proportional hazards assumption) είναι η γραφική παράσταση των υπολοίπων Schoenfeld συναρτήσει του χρόνου. Ο Schoenfeld (1982) πρότεινε αυτά τα υπόλοιπα για το μοντέλο αναλογικών κινδύνων του Cox. Η διαφορά αυτών των υπολοίπων με άλλες περιπτώσεις είναι πως για αυτά υπολογίζεται ένα ξεχωριστό υπόλοιπο για κάθε άτομο για κάθε μεταβλητή. Δηλαδή, εάν έχουμε p επεξηγηματικές μεταβλητές, τότε για κάθε άτομο υπολογίζονται p Schoenfeld υπόλοιπα (residuals). Εάν το γράφημα των Schoenfeld υπολοίπων ως προς το χρόνο έχει τυχαία μορφή τότε η υπόθεση των αναλογικών κινδύνων θα πληρείται. Αν X_j είναι η μεταβλητή για την οποία θέλουμε να υπολογίσουμε το i -οστό Schoenfeld υπόλοιπο, τότε αυτό δίνεται από τη σχέση:

$$r_{S_{ji}} = \delta_i \{x_{ji} - \hat{a}_{ji}\} \quad (2.23)$$

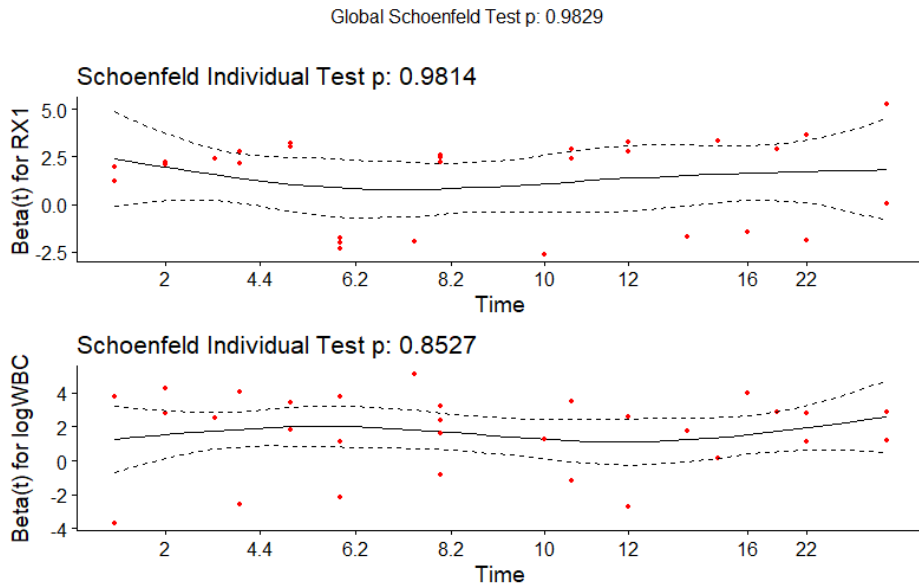
όπου x_{ji} είναι η τιμή της j -επεξηγηματικής μεταβλητής $j = 1, 2, \dots, p$ για το

i -οστό άτομο της έρευνας. Επιπλέον

$$\hat{a}_{ji} = \frac{\sum_{l \in R(t_{(i)})} x_{jl} \exp\{\hat{\beta}^T x_l\}}{\sum_{l \in R(t_{(i)})} \exp\{\hat{\beta}^T x_l\}} \quad (2.24)$$

όπου $R(t_i)$ είναι το σύνολο κινδύνου (risk set) για τα άτομα που έχουν επιβιώσει μέχρι τη χρονική στιγμή ακριβώς πριν την t_i και $\delta_i = 0$ αν ο χρόνος t_i είναι λογοκριμένος, ενώ $\delta_i = 1$ αν ο χρόνος t_i είναι πλήρως καταγεγραμμένος. Από τις γραφικές παραστάσεις των Schoenfeld υπολοίπων ως προς το χρόνο για το σύνολο δεδομένων αυτού του κεφαλαίου, προέκυψαν τα παρακάτω γραφήματα.

Σχήμα 2.10: Γράφημα των Schoenfeld υπολοίπων ως προς το χρόνο.



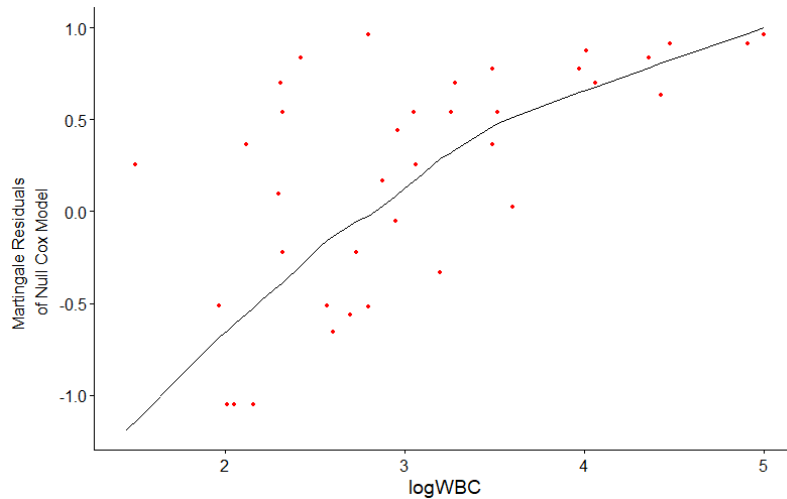
Από τα προηγούμενα γραφήματα δεν παρατηρούμε κάποιο μοτίβο στον τρόπο κατανομής των υπολοίπων και έτσι συμπεραίνουμε πως η υπόθεση της αναλογικότητας των κινδύνων (proportional hazards assumption) πληρείται.

Τα υπόλοιπα martingale (martingale residuals) χρησιμοποιούνται για την εύρεση της συναρτησιακής μορφής μιας μεταβλητής που πρόκειται να εισαχθεί στο μοντέλο του Cox. Η πιο απλή μορφή είναι να σχεδιάσουμε τα υπόλοιπα martingale από ένα μηδενικό μοντέλο (null model) δηλαδή από ένα μοντέλο με $\hat{\beta} = 0$. Προφανώς, αυτά τα γραφήματα έχουν νόημα για τις συνεχείς μεταβλητές, αφού η γραμμικότητα δεν είναι κάτι που μας απασχολεί για την περίπτωση των κατηγορικών μεταβλητών. Τα υπόλοιπα martingale ορίζονται από τη σχέση:

$$\hat{r}_{M_i} = \delta_i - \hat{r}_{c_i}, \quad i = 1, 2, \dots, n \quad (2.25)$$

όπου $\hat{r}_{c_i} = \hat{H}_0(t_i) \exp\{\hat{\beta}^T x_i\}$. Οι ποσότητες r_{c_i} ονομάζονται υπόλοιπα Cox-Snell, για το i -οστό άτομο. Στο μοντέλο του Cox για το οποίο παρουσιάζουμε τους ελέγχους υπάρχει μια συνεχής μεταβλητή που είναι η logWBC. Στη συνέχεια δίνεται το γράφημα των martingale υπολοίπων για το μηδενικό μοντέλο ως προς τη συνεχή αυτή μεταβλητή. Από το γράφημα παρατηρούμε πως η υπόθεση της γραμμικότητας δεν παραβιάζεται.

Σχήμα 2.11: Martingale υπόλοιπα για τον έλεγχο της γραμμικότητας.

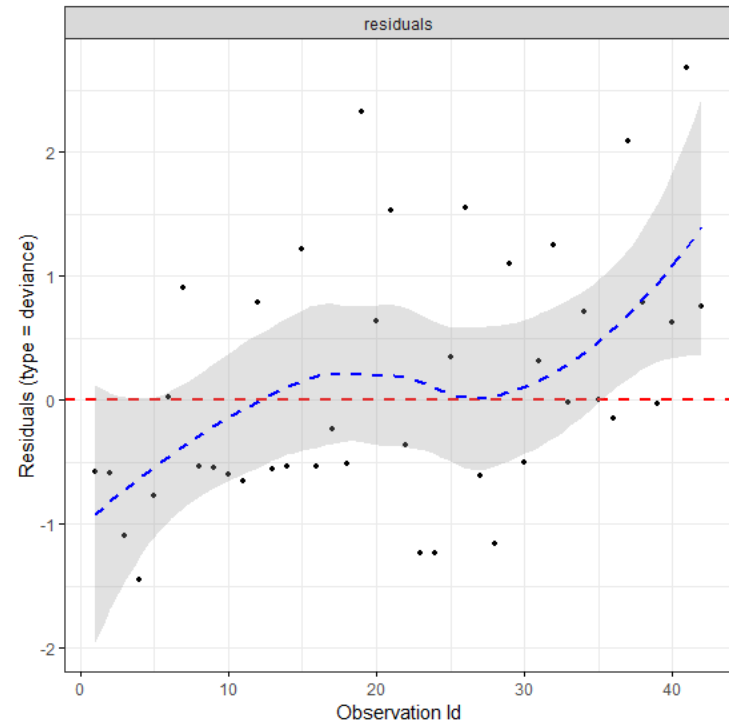


Τέλος, ο έλεγχος για την ύπαρξη ακραίων τιμών γίνεται με τη βοήθεια των deviance υπολοίπων (deviance residuals), Therneau et al. (1990). Τα υπόλοιπα αυτά ορίζονται από τη σχέση

$$r_{D_i} = \text{sgn}(r_{M_i})[-2\{r_{M_i} + \delta_i \log(\delta_i - r_{M_i})\}]^{1/2} \quad (2.26)$$

όπου $\text{sgn}(\cdot)$ η συνάρτηση πρόσημο η οποία παίρνει την τιμή 1 όταν το όρισμα της είναι θετικό και το -1 όταν το όρισμα της είναι αρνητικό, ενώ με r_{M_i} συμβολίζουμε τα martingale υπόλοιπα. Η συμμετρία του γραφήματος αυτών των υπολοίπων γύρω από το 0, υποδεικνύει πως δεν υπάρχουν ακραίες τιμές. Το γράφημα που δημιουργήθηκε με βάση το σύνολο δεδομένων που χρησιμοποιήσαμε, δίνεται παρακάτω και από αυτό συμπεραίνουμε πως δεν υπάρχουν ακραίες τιμές.

Σχήμα 2.12: Deviance υπόλοιπα για τον έλεγχο των ακραίων τιμών.



Κεφάλαιο 3

ΜΕΙΩΣΗ ΤΗΣ ΜΕΡΟΛΗΨΙΑΣ

Είδαμε στα προηγούμενα κεφάλαια ότι η μέθοδος της πιθανοφάνειας έχει κεντρικό ρόλο στην προσαρμογή ενός μοντέλου αφού με αυτή τη μέθοδο εκτιμούνται οι παράμετροι της παλινδρόμησης β_j , $j = 1, 2, \dots, p$. Η μέθοδος της μεγιστοποίησης της συνάρτησης πιθανοφάνειας είναι η πλέον διαδεδομένη μέθοδος εκτίμησης παραμέτρων λόγω της ευκολίας με την οποία αυτή μπορεί να μεγιστοποιηθεί. Επιπλέον, ένας εξίσου σημαντικός λόγος για την ευρεία χρήση της μεθόδου είναι οι ευνοϊκές ιδιότητες που διαθέτει ο εκτιμητής μέγιστης πιθανοφάνειας σε περιπτώσεις μεγάλων δειγμάτων. Μερικές από αυτές τις ιδιότητες είναι η ασυμπτωτική κανονικότητα, η συνέπεια και η αμεροληψία. Ωστόσο όλες οι ασυμπτωτικές ιδιότητες που αναφέρθηκαν ισχύουν κάτω από κάποιες συνθήκες, οι οποίες είναι γνωστές ως συνθήκες ομαλότητας (regularity conditions) και δίνονται παρακάτω (Cox and Hinkley, 1974)

1. Ο παραμετρικός χώρος έχει πεπερασμένη διάσταση, είναι κλειστός και συμπαγής, ενώ η πραγματική τιμή της υπό εκτίμηση παραμέτρου βρίσκεται στο εσωτερικό του παραμετρικού χώρου. Επιπλέον, ο παραμετρικός χώρος δεν πρέπει να εξαρτάται από τον δειγματικό χώρο.
2. Οι κατανομές πιθανότητας που ορίζονται από δύο διαφορετικά σημεία του παραμετρικού χώρου, είναι διαφορετικές.
3. Η συνάρτηση πιθανοφάνειας είναι σχεδόν σίγουρα (almost surely) τρεις φορές παραγωγίσιμη σε μία γειτονιά της πραγματικής τιμής της παραμέτρου. Επιπρόσθετα σε μία γειτονιά της άγνωστης παραμέτρου η ποσότητα

$$n^{-1} \left| \frac{\partial^3 l(\beta)}{\partial \beta_s \partial \beta_t \partial \beta_u} \right|, \quad 1 \leq s, t, u \leq p \quad (3.1)$$

είναι άνω φραγμένη από μια συνάρτηση της τυχαίας μεταβλητής Y της οποίας η αναμενόμενη τιμή υποθέτουμε ότι υπάρχει και είναι πεπερασμένη.

4. Για το διάνυσμα score (score vector) ισχύει ότι

$$E(U(\beta)U(\beta)^t) = E\left(-\frac{\partial}{\partial \theta} U^t(\beta)\right) = I(\beta) \quad (3.2)$$

όπου $I(\beta)$ είναι ο πίνακας πληροφορίας του Fisher, ο οποίος είναι θετικά ορισμένος και πεπερασμένος στη γειτονιά της πραγματικής τιμής της υπό εκτίμηση παραμέτρου β .

Ορισμός 3.0.1. Μια ακολουθία $\{X_n\}$ τυχαίων μεταβλητών λέμε ότι συγχλίνει κατά κατανομή σε μια τυχαία μεταβλητή X και συμβολίζουμε $X_n \xrightarrow{d} X$ αν,

$$\lim F_n(x) = F(x)$$

για κάθε $x \in \mathbb{R}$ στο οποίο η F είναι συνεχής, όπου F_n και F είναι οι αθροιστικές συναρτήσεις κατανομών των τυχαίων μεταβλητών X_n και X αντίστοιχα.

Όλες οι προηγούμενες υποθέσεις και ο ορισμός της σύγκλισης κατά κατανομή είναι αναγκαία προκειμένου να διατυπώσουμε ένα από τα πιο βασικά θεωρήματα της ασυμπτωτικής στατιστικής που είναι το κεντρικό οριακό θεώρημα.

Θεώρημα 3.0.2 (Κεντρικό οριακό θεώρημα). Έστω $X_1, X_2, \dots, X_p$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές και έστω ότι οι συνθήκες ομαλότητας ικανοποιούνται, τότε με πιθανότητα που τείνει στο 1 και για $n \rightarrow +\infty$ υπάρχει ακολουθία λύσεων $\hat{\beta}_j, j = 1, 2, \dots, p$ της συνάρτησης πιθανοφάνειας τέτοια ώστε

$$\sqrt{n}(\hat{\beta}_j - \beta) \xrightarrow{d} N(0, I(\beta)^{-1}) \quad (3.3)$$

Ένα από τα βασικά συμπεράσματα που μας δίνει το κεντρικό οριακό θεώρημα είναι πως ο εκτιμητής μέγιστης πιθανοφάνειας είναι ασυμπτωτικά αμερόληπτος. Δυστυχώς, η ιδιότητα της αμεροληψίας είναι μόνο μια ασυμπτωτική ιδιότητα, δηλαδή ισχύει σε περιπτώσεις μεγάλων δειγμάτων τα οποία στην πράξη δεν είναι πάντοτε διαθέσιμα. Η μεροληψία του εκτιμητή μέγιστης πιθανοφάνειας θεωρείται από πολλούς ως παθογένεια της μεθόδου την οποία πολλοί ερευνητές προσπαθούν να διορθώσουν επινοώντας διάφορους τρόπους αντιμετώπισης αυτού του προβλήματος. Προτού όμως δούμε κάποιες από αυτές τις μεθόδους κρίνεται σκόπιμο να δοθούν κάποιες βασικές έννοιες που θα χρησιμοποιηθούν στην παρούσα μεταπτυχιακή διατριβή από εδώ και στο εξής και έχουν κεντρικό ρόλο στις μεθόδους αντιμετώπισης του προβλήματος της μεροληψίας του εκτιμητή μέγιστης πιθανοφάνειας.

3.1 Τα σύμβολα $\mathcal{O}$, o , $\mathcal{O}_p$, o_p

Η ουσία των ασυμπτωτικών μεθόδων βρίσκεται στην έννοια της προσέγγισης. Μπορούμε να προσεγγίζουμε συναρτήσεις τυχαίων μεταβλητών, μέσες τιμές, δια-

κύμανσεις, συνδιακυμάνσεις κ.α. Βέβαια, μια προσεκτική χρήση αυτών των συμβόλων απαιτεί να εστιάσουμε την προσοχή μας στην ακρίβεια ή αλλιώς την τάξη της προσέγγισης. Ένας τρόπος ώστε να μπορούμε να καταγράφουμε την τάξη μιας προσέγγισης είναι η χρήση των συμβόλων του "μεγάλου" $\mathcal{O}$ και του "μικρού" o . Έτσι χρησιμοποιούμε τα σύμβολα $\mathcal{O}$ και o για μη-στοχαστικές (ή αλλιώς ντετερμινιστικές) μεταβλητές και τα σύμβολα $\mathcal{O}_p$ και o_p για στοχαστικές μεταβλητές. Στη συνέχεια δίνονται οι ορισμοί και από τις βασικές ιδιότητες αυτών των εννοιών με τρόπο παρόμοιο που αυτές περιγράφονται από τους Bishop et al. (1974).

Ορισμός 3.1.1. Έστω δύο ακολουθίες $a_n, n \in \mathbb{N}$ και $b_n, n \in \mathbb{N}$. Θα γράφουμε ότι $a_n = \mathcal{O}(b_n)$ αν και μόνο αν $\exists \epsilon > 0$ και ένας θετικός ακέραιος $N(\epsilon)$ τέτοιος ώστε αν $n \geq N(\epsilon) \Rightarrow \left| \frac{a_n}{b_n} \right| < \epsilon$. Συμβολικά:

$$a_n = \mathcal{O}(b_n) \Leftrightarrow (\exists \epsilon > 0) \wedge (\exists N(\epsilon) \in \mathbb{N}) : \forall n \geq N(\epsilon) \Rightarrow \left| \frac{a_n}{b_n} \right| < \epsilon \quad (3.4)$$

Παρατήρηση: Από τον προηγούμενο ορισμό υπονοείται ότι $a_n = \mathcal{O}(b_n)$ τότε ο λόγος $\left| \frac{a_n}{b_n} \right|$ είναι φραγμένος για μεγάλο n .

Ορισμός 3.1.2. Έστω δύο ακολουθίες $a_n, n \in \mathbb{N}$ και $b_n, n \in \mathbb{N}$. Θα λέμε ότι η $a_n = o(b_n)$ αν ο λόγος $\left| \frac{a_n}{b_n} \right| \xrightarrow{n \rightarrow +\infty} 0$. Συμβολικά:

$$a_n = o(b_n) \Leftrightarrow (\forall \epsilon > 0) (\exists n(\epsilon) \in \mathbb{N}) (\forall n \in \mathbb{N}) \text{ με } n \geq n(\epsilon) \Rightarrow \left| \frac{a_n}{b_n} \right| < \epsilon \quad (3.5)$$

Παρατήρηση:

1. Η ιδέα πίσω από τους δύο ορισμούς είναι η σύγκριση της τάξης μεγέθους (order of magnitude) μιας ακολουθίας $a_n, n \in \mathbb{N}$ σε σχέση με μια άλλη ακολουθία $b_n, n \in \mathbb{N}$. Στις περισσότερες εφαρμογές η ακολουθία που βρίσκεται

εκτός το $\mathcal{O}(\cdot)$ ή του $o(\cdot)$ είναι η ακολουθία ενδιαφέροντος, ενώ η ακολουθία που βρίσκεται εντός του $\mathcal{O}(\cdot)$ ή του $o(\cdot)$ είναι μια γνωστή ακολουθία σύγκρισης, συνήθως τέτοιες ακολουθίες επιλέγονται να έχουν τη μορφή $b_n = n^a$, για $a \in \mathbb{R}$ και για $n \in \mathbb{N}$.

2. Η ερμηνεία της σχέσης $a_n = \mathcal{O}(b_n)$ είναι πως η ακολουθία $a_n, n \in \mathbb{N}$ έχει την ίδια ή μικρότερη τάξη σε σχέση με την $b_n, n \in \mathbb{N}$, ενώ η ερμηνεία της σχέσης $a_n = o(b_n)$ είναι πως η ακολουθία $a_n, n \in \mathbb{N}$ έχει γνήσια μικρότερη τάξη σε σχέση με την ακολουθία $b_n, n \in \mathbb{N}$

Ορισμός 3.1.3. Έστω δύο ακολουθίες $X_n, n \in \mathbb{N}$ και $a_n, n \in \mathbb{N}$. Θα λέμε ότι:

$$X_n = \mathcal{O}_p(a_n) \Leftrightarrow (\forall \epsilon > 0) (\exists n(\epsilon) > 0) \wedge (\exists K(\epsilon) \in \mathbb{N}) \text{ τέτοια ώστε}$$

$$\text{αν } n \geq n(\epsilon) \Rightarrow P\left(\frac{|X_n|}{|a_n|} \leq K(\epsilon)\right) \geq 1 - \epsilon \quad (3.6)$$

Ορισμός 3.1.4. Έστω $X_n, n \in \mathbb{N}$ και $a_n, n \in \mathbb{N}$ δύο ακολουθίες πραγματικών αριθμών, θα λέμε ότι:

$$X_n = o_p(a_n) \Leftrightarrow (\forall \epsilon > 0) (\forall \delta > 0) (\exists n(\delta, \epsilon) \in \mathbb{N}) \text{ τέτοιο ώστε αν}$$

$$n \geq n(\delta, \epsilon) \Rightarrow P\left(\frac{|X_n|}{|a_n|} < \delta\right) \geq 1 - \epsilon \quad (3.7)$$

Παρατήρηση:

1. Η ερμηνεία και αυτών των ορισμών είναι αντίστοιχη με την ερμηνεία των συμβόλων $\mathcal{O}$ και o , δηλαδή όταν γράφουμε $X_n = \mathcal{O}_p(a_n)$ τότε η τάξη της $X_n, n \in \mathbb{N}$ είναι ίδια ή μικρότερη από την τάξη της $a_n, n \in \mathbb{N}$ κατά πιθανότητα, ενώ όταν $X_n = o_p(a_n)$ τότε η τάξη της $X_n, n \in \mathbb{N}$ είναι μικρότερη από την τάξη της $a_n, n \in \mathbb{N}$ κατά πιθανότητα.

Στον επόμενο πίνακα δίνονται κάποιες ιδιότητες των στοχαστικών και των ντε-τερμινιστικών συμβόλων τάξης. Αν $a, b \in \mathbb{R}$ και $k = \max\{a, b\}$ τότε,

Γινόμενα	Αθροίσματα	Συνθέσεις
$o(n^a) o(n^b) = o(n^{a+b})$	$o(n^a) + o(n^b) = o(n^k)$	$\mathcal{O}_p(\mathcal{O}(n^a)) = \mathcal{O}_p(n^a)$
$\mathcal{O}(n^a) \mathcal{O}(n^b) = \mathcal{O}(n^{a+b})$	$\mathcal{O}(n^a) + \mathcal{O}(n^b) = \mathcal{O}(n^k)$	$o(\mathcal{O}_p(n^a)) = o_p(n^a)$
$\mathcal{O}(n^a) o(n^b) = o(n^{a+b})$	$\mathcal{O}(n^a) + o(n^b) = \mathcal{O}(n^k)$	$o_p(\mathcal{O}_p(n^a)) = o_p(n^a)$
$o_p(n^a) o_p(n^b) = o_p(n^{a+b})$	$o_p(n^a) + o_p(n^b) = o_p(n^k)$	-
$\mathcal{O}_p(n^a) \mathcal{O}_p(n^b) = \mathcal{O}_p(n^{a+b})$	$\mathcal{O}_p(n^a) + \mathcal{O}_p(n^b) = \mathcal{O}_p(n^k)$	-
$\mathcal{O}_p(n^a) o_p(n^b) = o_p(n^{a+b})$	$\mathcal{O}_p(n^a) + o_p(n^b) = \mathcal{O}_p(n^k)$	-
$o_p(n^a) o(n^b) = o_p(n^{a+b})$	-	-
$\mathcal{O}_p(n^a) o(n^b) = o_p(n^{a+b})$	-	-
$\mathcal{O}(n^a) o_p(n^b) = o_p(n^{a+b})$	-	-
$\mathcal{O}(n^a) \mathcal{O}_p(n^b) = \mathcal{O}_p(n^{a+b})$	-	-

Στη συνέχεια δίνεται ένας χρήσιμος τρόπος σύνδεσης μεταξύ $\mathcal{O}_p$ και o_p .

Θεώρημα 3.1.5. (*Kosmidis, PhD thesis, 2007*)

Αν $X_n = \mathcal{O}_p(n^{-a})$, όπου $a > 0$, τότε $X_n = o_p(n^{-a+t})$, για κάθε $t > 0$.

Απόδειξη. Έστω ότι $X_n = \mathcal{O}_p(n^{-a})$, τότε έχουμε ότι $(\forall \epsilon > 0) (\exists n(\epsilon) > 0) \wedge (\exists K(\epsilon) \in \mathbb{N})$ τέτοια ώστε αν $n \geq n(\epsilon) \Rightarrow P(n^a |X_n| \leq K(\epsilon)) \geq 1 - \epsilon$ ή ισοδύναμα αν $n \geq n(\epsilon) \Rightarrow P(n^{a-t} |X_n| \leq n^{-t} K(\epsilon)) \geq 1 - \epsilon$, για $t > 0$. Σταθεροποιούμε ένα $\epsilon > 0$ τότε $(\forall n \geq n(\epsilon)) (\exists \delta > n^{-t} K(\epsilon))$:

$$1 - \epsilon \leq P(n^{a-t} |X_n| \leq n^{-t} K(\epsilon)) \leq P(n^{a-t} |X_n| < \delta) \quad (3.8)$$

Θεωρούμε το σύνολο $\mathcal{D}_n = \{\delta : \delta > n^{-t} K(\epsilon) | t > 0, \epsilon > 0\}$ και έτσι $\mathcal{D}_n \subset \mathbb{R}^+$.

Επιπλέον, αν $\overline{\mathcal{D}_n} = \mathbb{R}^+ - \mathcal{D}_n$, το συμπλήρωμα του $\mathcal{D}_n$ και για κάθε υποσύνολο A

του $\mathbb{R}^+$ ορίζουμε το μήκος του A ως: $\text{len}A = \sup A - \inf A$. Έτσι, έχουμε ότι $\text{len}\overline{\mathcal{D}_n} = n^{-t}K(\epsilon)$.

Η σχέση (3.8) ικανοποιείται για κάθε $\delta \in \mathcal{D}_{n(\epsilon)}$, όμως όχι απαραίτητα για κάθε $\delta \in \overline{\mathcal{D}_{n(\epsilon)}}$ επίσης παρατηρούμε ότι $\mathcal{D}_\infty \equiv \mathbb{R}^+$ και $\overline{\mathcal{D}_\infty} \equiv \emptyset$, όπου $\emptyset$ το κενό σύνολο. Άρα, η συνάρτηση $\overline{\mathcal{D}_n}$ είναι μια γνήσιως φθίνουσα συνάρτηση του n . Έτσι η εγκυρότητα της (3.8) μπορεί να εξασφαλιστεί για όλα τα $\delta \in \mathbb{R}^+$ που είναι τέτοια ώστε $n \geq n^*(\delta, \epsilon)$, όπου το $n^*(\delta, \epsilon)$ είναι ένας θετικός ακέραιος επαρκώς μεγαλύτερος από το $n(\epsilon)$.

Τελικά, για κάθε $\delta \in \mathbb{R}^+$ υπάρχει $n^*(\delta, \epsilon) > n(\epsilon)$ τέτοιο ώστε αν $n > n^*(\delta, \epsilon) \Rightarrow P(n^{a-t}|X_n| < \delta) \geq 1 - \epsilon$, για $t > 0$. Λόγω της τυχαίότητας με την οποία επιλέξαμε το σταθεροποιημένο $\epsilon > 0$ η τελευταία σχέση ισχύει $\forall \epsilon > 0$. Έτσι από τον ορισμό 3.2.4 έχουμε τελικά ότι $X_n = o_p(n^{-a+t})$, για κάθε $t > 0$. $\square$

3.2 Cumulants

Οι cumulants είναι μια έννοια άμεσα συνδεδεμένη με αυτή των ροπών (moments), αφού όπως θα δειχθεί στη συνέχεια οι cumulants ορίζονται με βάση την ροπογεννήτρια συνάρτηση. Παρόλα αυτά είναι προτιμότερο πολλές φορές να εργαζόμαστε με τους cumulants παρά με τις ροπές, δύο λόγοι για τους οποίους συμβαίνει αυτό είναι (βλ. McCullagh, 2018):

1. Οι περισσότεροι στατιστικοί υπολογισμοί είναι ευκολότεροι όταν χρησιμοποιούμε cumulants παρά όταν χρησιμοποιούμε ροπές.
2. Για ανεξάρτητες τυχαίες μεταβλητές οι cumulants του αθροίσματος είναι το άθροισμα των cumulants.

Ορισμός 3.2.1. Έστω Y μια μονοδιάστατη τυχαία μεταβλητή με πυκνότητα $f_Y(y)$. Η r τάξης ροπή περί το μηδέν της τυχαίας μεταβλητής Y ορίζεται να είναι η ποσότητα $\mu_r = E(Y^r)$, $r = 1, 2, \dots$. Επιπλέον, η ποσότητα $M_Y(t) = E(e^{tY})$, $t \in \mathbb{R}$ ονομάζεται ροπογεννήτρια συνάρτηση.

Παρατήρηση:

1. Αν Y είναι μια τυχαία μεταβλητή ώστε $M_Y(t) < +\infty$ για $|t| < t_0$, όπου t_0 μια θετική σταθερά, τότε η ροπογεννήτρια συνάρτηση της τυχαίας μεταβλητής Y μπορεί να αναπτυχθεί με τη βοήθεια δυναμοσειράς ως εξής:

$$\begin{aligned} M_Y(t) &= E(e^{tY}) = E(1 + Yt + Y^2 \frac{t^2}{2!} + Y^3 \frac{t^3}{3!} + \dots) \\ &= 1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \dots \end{aligned} \quad (3.9)$$

Από την προηγούμενη σχέση γίνεται αντιληπτό πως οι ροπές r τάξης περί το μηδέν μπορούν να προσδιοριστούν από την ακόλουθη σχέση:

$$\mu_r = \left. \frac{d^r}{dt^r} M_Y(t) \right|_{t=0} \quad (3.10)$$

Ορισμός 3.2.2. Έστω Y μια τυχαία μεταβλητή για την οποία η ροπογεννήτρια συνάρτηση $M_Y(t)$ είναι πεπερασμένη για $|t| < t_0$, τότε ορίζουμε ως cumulant-γεννήτρια συνάρτηση την συνάρτηση $K_Y(t)$ που έχει τη μορφή:

$$K_Y(t) = \log(M_Y(t)) \quad (3.11)$$

Παρατήρηση:

1. Εφόσον η cumulant-γεννήτρια συνάρτηση ορίζεται με βάση τον λογάριθμο της αντίστοιχης ροπογεννήτριας, συμπεραίνουμε πως το θεώρημα του μονοσημάντου των ροπογεννητριών μπορεί να επεκταθεί και ως θεώρημα του μονοσημάντου της cumulant-γεννήτριας συνάρτησης.

2. Η cumulant-γεννήτρια συνάρτηση μπορεί να αναπτυχθεί σε μια δυναμοσειρά που έχει τη μορφή $K_Y(t) = k_1 t + k_2 \frac{t^2}{2!} + k_3 \frac{t^3}{3!} + k_4 \frac{t^4}{4!} + \dots$. Προκειμένου να προσδιορίσουμε τους πρώτους 4 όρους k_1, k_2, k_3, k_4 εργαζόμαστε ως εξής: Γνωρίζουμε ότι

$$K_Y(t) = \log(M_Y(t)) \Rightarrow K_Y(t) = \log(1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \dots).$$

Γενικά ισχύει ότι

$$\log(1 + z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \frac{z^4}{4} + \dots \quad (3.12)$$

Θα εφαρμόσουμε την σχέση (3.12) για να προσδιορίσουμε τους cumulants μέχρι και 4ης τάξης συναρτήσει των ροπών περί το μηδέν μ_r , ο r τάξης cumulant θα συμβολίζεται με k_r . Επιπλέον, προκύπτει δύναμη του t μεγαλύτερη του 4 θα αγνοείται. Επομένως,

$$\begin{aligned} K_Y(t) &= \log(M_Y(t)) = \log(1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \dots) = \\ &= \left(\mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \mu_4 \frac{t^4}{4!} \right) - \frac{1}{2} \left(\mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} \right)^2 \\ &\quad + \frac{1}{3} \left(\mu_1 t + \mu_2 \frac{t^2}{2!} \right)^3 - \frac{1}{4} \mu_1^4 t^4 \\ &= \left(\mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \mu_4 \frac{t^4}{4!} \right) - \frac{1}{2} \left(\mu_1^2 t^2 + \frac{1}{4} \mu_2^2 t^4 + \mu_1 \mu_2 t^3 + \frac{1}{3} \mu_1 \mu_3 t^4 \right) \\ &\quad + \frac{1}{3} \left(\mu_1^3 t^3 + \frac{3}{2} \mu_1^2 \mu_2 t^4 \right) - \frac{1}{4} \mu_1^4 t^4 \\ &= \mu_1 t + (\mu_2 - \mu_1^2) \frac{t^2}{2!} + (\mu_3 - 3\mu_1 \mu_2 + 2\mu_1^3) \frac{t^3}{3!} + (\mu_4 - 3\mu_2^2 - 4\mu_1 \mu_3 + 12\mu_1^2 \mu_2 - 6\mu_1^4) \frac{t^4}{4!} \end{aligned}$$

Έτσι, από την τελευταία σχέση έχουμε ότι:

$$k_1 = \mu_1 \quad (3.13)$$

$$k_2 = \mu_2 - \mu_1^2 = \text{Var}(Y) \quad (3.14)$$

$$k_3 = \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 = E(Y - \mu_1)^3 \quad (3.15)$$

$$k_4 = \mu_4 - 3\mu_2^2 - 4\mu_1\mu_3 + 12\mu_1^2\mu_2 - 6\mu_1^4 \quad (3.16)$$

Οι προηγούμενες σχέσεις είναι γνωστές στη βιβλιογραφία ως σχέσεις του Bartlett (Bartlett, 1953).

3. Στην περίπτωση που γνωρίζουμε την cumulant-γεννήτρια συνάρτηση μπορούμε να προσδιορίσουμε τους cumulants r τάξης από τη σχέση:

$$k_r = k_r(Y) = \left. \frac{d^r}{dt^r} K_Y(t) \right|_{t=0} \quad (3.17)$$

4. Έστω $Y_1, Y_2, \dots, Y_n$ ανεξάρτητες και ισόνομες τυχαίες μεταβλητές και $S_n = \sum_{i=1}^n Y_i$, τότε:

$$M_{S_n}(t) = E(e^{t(Y_1+Y_2+\dots+Y_n)}) = (M_{Y_1}(t))^n \quad (3.18)$$

$$K_{S_n}(t) = \log(M_{S_n}(t)) = n \log(M_{Y_1}(t)) = nK_{Y_1}(t) \quad (3.19)$$

3.3 Index Notation

Ο όρος index notation χρησιμοποιείται για να περιγράψει ένα σύνολο κανόνων και συμβάσεων που διευκολύνουν τον συμβολισμό και την περιγραφή πολυδιάστατων αλγεβρικών κατασκευών. Η σύμβαση που μας επιτρέπει να τοποθετήσουμε τους δείκτες των στοιχείων πινάκων τόσο ως άνω δείκτες (upper indices) όσο και ως κάτω δείκτες (lower indices) είναι η βάση για τη συγκεκριμένη θεωρία. Για παράδειγμα αν οι δείκτες r, s, t έχουν εύρος $\{1, 2, \dots, p\}$ οι συμβολισμοί i_{rs}, i^{rs}, a_r^s

υποδηλώνουν στοιχεία ενός $p \times p$ πίνακα, ενώ ο συμβολισμός $l_{rst}, b^{rst}, b_{rs}^t, b_r^{st}$ θα υποδηλώνει στοιχεία πινάκων με διαστάσεις $p \times p \times p$ (three way arrays). Ένας άνω δείκτης ονομάζεται ανταλλοιώτος (contravariant index), ενώ ένας κάτω δείκτης ονομάζεται συναλλοιώτος (covariant index). Επιπρόσθετα κεντρικό ρόλο έχει η σύμβαση άθροισης του Einstein (Einstein, 1916) η οποία διατυπώνεται ως εξής: Όταν ο ίδιος δείκτης εμφανίζεται δύο φορές μια φορά στο ανταλλοίωτο μέρος μιας ποσότητας a και μια φορά στο συναλλοίωτο μέρος μιας ποσότητας b τότε υποδεικνύεται άθροιση ως προς αυτό το δείκτη. Για παράδειγμα,

$$a^r b_r := \sum_{r=1}^p a^r b_r$$

Η σημασία της χρήσης δεικτών έγκειται στο γεγονός ότι μας επιτρέπει να απλοποιήσουμε παραστάσεις πολυδιάστατων ποσοτήτων, οι οποίες μπορούν έτσι να αντιμετωπιστούν όπως οι αντίστοιχες μονοδιάστατες ποσότητες. Για παράδειγμα έστω $x = [x^r]$ και $y = [y^r]$ δύο διανύσματα στήλες στον $\mathbb{R}^p$ και $A = [a_{ij}]$ ένας $p \times p$ πίνακας τότε:

$$\begin{aligned} x \cdot y &= x^r y^r \\ x^T A x &= a_{ij} x^i x^j \end{aligned}$$

Παρόλο που στα προηγούμενα παραδείγματα η χρήση της θεωρίας index notation φαίνεται περιττή, αφού δεν είναι δύσκολη η αναπαράσταση των ποσοτήτων με χρήση πινάκων παρατηρούμε ότι οποιαδήποτε μετάθεση των στοιχείων στις πίνακικές μορφές δεν θα ήταν επιτρεπτή, διότι τότε θα παραβιαζόταν η βασική θεωρία του πολλαπλασιασμού πινάκων. Αντίθετα αυτό δε συμβαίνει στις εκφράσεις που έχουν γραφεί με τη βοήθεια δεικτών αφού σε αυτές επιδρά η σύμβαση άθροισης. Βέβαια υπάρχουν και περιπτώσεις στις οποίες είναι αδύνατο να δώσουμε μια συμπαγή μορφή για ορισμένες περιπτώσεις πινάκων με τη χρήση του κλασικού

συμβολισμού. Θεωρούμε για παράδειγμα την κυβική μορφή που εκφράζεται με τη χρήση index notation ως:

$$y = a_{ijk}x^i x^j x^k$$

όπου $A = [a_{ijk}]$ ένας $p \times p \times p$ πίνακας. Η συγκεκριμένη κυβική μορφή δεν μπορεί να εκφραστεί απευθείας με χρήση πινάκων, μια πιθανή γραφή του δίνεται στη συνέχεια. Θεωρούμε την ποσότητα $z = [x^T A_k x]$, όπου $A_k = [a_{ijk}]$, k σταθερό και θέτουμε $y = z^T x$. Τότε η ποσότητα

$$y = \sum_{k=1}^p (x^T A_k x) x^k$$

αποτελεί έναν τρόπο για τη γραφή της κυβικής μορφής που εκφράστηκε προηγουμένως με χρήση index notation. Από αυτό το παράδειγμα γίνεται αντιληπτό πως σε ποσότητες μεγαλύτερων διαστάσεων η γραφή με χρήση του κλασικού συμβολισμού των πινάκων είναι σχεδόν αδύνατη.

Παρατήρηση:

1. Σε γινόμενα στοιχείων πινάκων όπου εφαρμόζεται η συνθήκη άθροισης του Einstein οι επαναλαμβανόμενοι δείκτες ονομάζονται dummy indices, ενώ αυτοί που εμφανίζονται μόνο μια φορά ονομάζονται free indices.
2. Αν όλοι οι δείκτες έχουν εύρος $\{1, 2, \dots, p\}$ τότε το γινόμενο των στοιχείων των πινάκων δίνει έναν πίνακα διάστασης p^h , όπου h ο αριθμός των free indices.

Παράδειγμα: (*Kosmidis, PhD thesis, 2007*) Θεωρούμε τα διανύσματα δύο τυχαίων μεταβλητών $X = [X^r]$ και $Y = [Y^r]$ που ανήκουν στον $\mathcal{R}^q$ και έχουν πίνακες διακυμάνσεων-συνδιακυμάνσεων $\Sigma_X = \text{Var}(X)$ και $\Sigma_Y = \text{Var}(Y)$

αντίστοιχα. Επιπλέον, $\Sigma_{XY} = [Cov(X_r, Y_s)]$. Θεωρούμε τους γραμμικούς μετασχηματισμούς $g(X) = AX$ και $h(Y) = BY$ με A και B να είναι δύο $k \times q$ πίνακες γνωστών σταθερών ποσοτήτων. Γνωρίζουμε ότι η τυχαία μεταβλητή $Z = g(X) + h(Y)$ έχει πίνακα διακυμάνσεων-συνδιακυμάνσεων

$$\Sigma = A\Sigma_X A^T + A\Sigma_{XY} B^T + B\Sigma_{XY}^T A^T + B\Sigma_Y B^T.$$

Ο προηγούμενος πίνακας γράφεται σε μορφή index notation ως

$$\sigma^{rs} = a_t^r a_u^s \sigma_X^{tu} + a_t^r b_u^s \sigma_{XY}^{tu} + b_t^r a_u^s \sigma_{XY}^{tu} + b_t^r b_u^s \sigma_Y^{tu}$$

όπου $A=[a_t^r]$ και $B=[b_t^r]$. Παρατηρούμε ότι και πάλι καμία μετάθεση στα στοιχεία των διάφορων γινομένων που εμφανίζονται στην σχέση που έχει γραφεί με τη βοήθεια index notation δε θα έχει επίδραση στο τελικό αποτέλεσμα. Στην προηγούμενη σχέση βλέπουμε ότι οι δείκτες που εμφανίζονται μόνο μια φορά σε κάθε όρο είναι οι δείκτες r, s και συνεπώς αυτοί αποτελούν τους free indices, ενώ οι επαναλαμβανόμενοι δείκτες είναι οι t, u και συνεπώς αυτοί αποτελούν τους dummy indices. Τέλος, παρακάτω δίνεται ένας κανόνας για ευκολότερη χρήση των free και dummy indices.

1. Από τη στιγμή που θα έχουμε επιλέξει τα γράμματα που θα χρησιμοποιηθούν ως free indices στο ένα μέλος της εξίσωσης, τα γράμματα που θα χρησιμοποιηθούν ως free indices στο άλλο μέλος της εξίσωσης δεν μπορούν να είναι διαφορετικά.
2. Για αθροίσματα γινομένων πινάκων, κάθε dummy index μπορεί να εμφανιστεί μόνο δύο φορές σε κάθε αθροιστέο, μια φορά στο ανταλλοίωτο μέρος μιας ποσότητας και μια φορά στο συναλλοίωτο μέρος μιας άλλης ποσότητας.

Ορισμός 3.3.1. Το σύμβολο δ_s^r καλείται ως το δέλτα του Kronecker και ορίζεται με τον εξής τρόπο:

$$\delta_s^r = \begin{cases} 1, & \text{αν } r = s \\ 0, & \text{αν } r \neq s \end{cases}$$

Παρατήρηση:

1. Από τον ορισμό του δέλτα του Kronecker είναι προφανές ότι $[\delta_s^r] = \mathcal{I}_p$, όπου $\mathcal{I}_p$ είναι ο $p \times p$ μοναδιαίος πίνακας. Επιπλέον, λόγω της σύμβασης άθροισης του Einstein:

$$\delta_s^r x^s = x^r, \quad (3.20)$$

ενώ εαν συμβολίσουμε ως $[i^{rs}]$ τον αντίστροφο πίνακα του $[i_{rs}]$ τότε ισχύει ότι:

$$i^{rs} i_{st} = \delta_t^r, \quad r, t = 1, 2, \dots, p \quad (3.21)$$

3.4 Μέθοδος μέγιστης πιθανοφάνειας σε όρους index notation

Έστω $l(\beta)$ ο λογάριθμος της συνάρτησης πιθανοφάνειας που προέκυψε από ένα στατιστικό μοντέλο με πυκνότητα $f_Y(y; \beta)$, $\beta \in \Theta \subseteq R^p$. Τότε συμβολίζουμε την μερική παράγωγο m τάξης της $l(\beta)$ ως:

$$U_{R_m}(\beta) = U_{r_1 r_2, \dots, r_m}(\beta) = \frac{\partial^m l(\beta)}{\partial \beta_{r_1} \partial \beta_{r_2}, \dots, \partial \beta_{r_m}} \quad (3.22)$$

Παρατήρηση:

1. Συχνά είναι χρήσιμο να γράφουμε πίνακες της μορφής $a_{r_1 r_2 \dots r_m}$ ή $b^{s_1 s_2 \dots s_n}$ χρησιμοποιώντας έναν πιο συμπαγή συμβολισμό. Έτσι θεωρούμε την πεπερασμένη ακολουθία των m δεικτών $R_m = \{r_1, r_2, \dots, r_m\}$ την οποία ονομάζουμε ως πολυδείκτη (multiindex). Αν θεωρήσουμε με παρόμοιο τρόπο έναν ακόμη πολυδείκτη $S_n = \{s_1, s_2, \dots, s_n\}$ τότε μπορούμε να γράψουμε ότι:

$$a_{R_m} = a_{r_1 r_2 \dots r_m} \quad (3.23)$$

$$b^{S_n} = b^{s_1 s_2 \dots s_n} \quad (3.24)$$

$$c_{R_m}^{S_n} = c_{r_1 r_2 \dots r_m}^{s_1 s_2 \dots s_n} \quad (3.25)$$

Με τη βοήθεια του προηγούμενου συμβολισμού μπορούμε να ορίσουμε τις null αναμενόμενες τιμές (null expected values), ή ισοδύναμα τις από κοινού null ροπές (joint null moments) των μερικών παραγώγων του λογαρίθμου της συνάρτησης πιθανοφάνειας ως εξής:

$$v_{R_m} \equiv v_{R_m}(\beta) = E\{U_{R_m}(\beta)\} \quad (3.26)$$

$$v_{R_m, S_n} \equiv v_{R_m, S_n}(\beta) = E\{U_{R_m}(\beta), U_{S_n}(\beta)\} \quad (3.27)$$

$$v_{R_m, S_n, T_q} \equiv v_{R_m, S_n, T_q}(\beta) = E\{U_{R_m}(\beta), U_{S_n}(\beta), U_{T_q}(\beta)\} \quad (3.28)$$

Η λέξη "null" μπαίνει για να υποδηλώσει ότι οι διαδικασίες της παραγωγίσης και της αναμενόμενης τιμής γίνονται για την ίδια τιμή της παραμέτρου β . Επιπλέον, κάτω από τις συνθήκες ομαλότητας και αφού $v_{R_m} = E\{U_{R_m}(\beta)\}$ για $m = 1 \Rightarrow v_{R_1} = v_r = E(U_r(\beta)) = 0$, άρα $v_r = 0$. Η παραγωγή της τελευταίας σχέσης μπορεί να δώσει μερικά ακόμη χρήσιμα συμπεράσματα. Πρώτα όμως πρέπει να δοθεί η μέθοδος με την οποία παραγωγίζονται τέτοιες ποσότητες. Σε αυτήν

την κατεύθυνση, καθοριστικό ρόλο έχει το θεώρημα του Skovgaard (Skovgaard 1986) το οποίο διατυπώνεται ως εξής:

Θεώρημα 3.4.1. Έστω $v_{R_a, S_b, T_c, \dots} = E(U_{R_a}, U_{S_b}, U_{T_c}, \dots; \beta) = E(D_1, D_2, \dots, D_m; \beta)$

τότε:

$$\frac{\partial v_{R_a, S_b, T_c, \dots}}{\partial \beta^u} = \sum_i E(D_1, D_2, \dots, \frac{\partial D_i}{\partial \beta^u}, \dots, D_m; \beta) + E(D_1, D_2, \dots, D_m, U_u; \beta) \quad \square$$

Τότε ξεκινώντας από τη σχέση:

$$v_r = 0$$

και παραγωγίζοντας την ως προς β^s έχουμε ότι: $\frac{\partial v_r}{\partial \beta^s} = v_{rs} + v_{r,s}$. Πράγματι,

$$\begin{aligned} \frac{\partial}{\partial \beta^s} v_r &= \frac{\partial}{\partial \beta^s} E\{U_r(\beta)\} = \frac{\partial}{\partial \beta^s} \int U_r(\beta) \exp\{\log(f_Y(y; \theta))\} d\mu^* \\ &= \int \frac{\partial}{\partial \beta^s} U_r(\beta) \exp\{\log(f_Y(y; \theta))\} d\mu^* \\ &= \int U_{rs}(\beta) \exp\{\log(f_Y(y; \theta))\} d\mu^* + \int U_r(\beta) U_s(\beta) \exp\{\log(f_Y(y; \theta))\} d\mu^* \\ &= E(U_{rs}(\beta)) + E(U_r(\beta) U_s(\beta)) = v_{rs} + v_{r,s} \end{aligned}$$

Έτσι, αφού $v_r = 0 \Rightarrow \frac{\partial v_r}{\partial \beta^s} = v_{rs} + v_{r,s} = 0$. Όμοια από την παραγωγή της τελευταίας σχέσης ως προς β^s έχουμε: $\frac{\partial}{\partial \beta^t} (v_{rs} + v_{r,s}) = v_{rst} + v_{rs,t} + v_{r,s,t} = 0$. Πράγματι,

$$\frac{\partial (v_{rs} + v_{r,s})}{\partial \beta^t} = 0 \Rightarrow \frac{\partial}{\partial \beta^t} v_{rs} + \frac{\partial}{\partial \beta^t} v_{r,s} = 0$$

Τότε όμως,

$$\begin{aligned} \frac{\partial}{\partial \beta^t} v_{rs} &= \frac{\partial}{\partial \beta^t} \int U_{rs}(\beta) \exp(\log(f_Y(y; \theta))) d\mu^* = \int \frac{\partial}{\partial \beta^t} U_{rs}(\beta) \exp(\log(f_Y(y; \theta))) d\mu^* \\ &= \int U_{rst}(\beta) \exp(\log(f_Y(y; \theta))) d\mu^* + \int U_{rs}(\beta) U_t(\beta) \exp(\log(f_Y(y; \theta))) d\mu^* \\ &= v_{rst} + v_{rs,t} \end{aligned}$$

ενώ,

$$\begin{aligned} \frac{\partial}{\partial \beta^t} v_{r,s} &= \frac{\partial}{\partial \beta^t} \int U_r(\beta) U_s(\beta) \exp(\log(f_Y(y; \theta))) d\mu^* = \int \frac{\partial}{\partial \beta^t} U_r(\beta) U_s(\beta) \exp(\log(f_Y(y; \theta))) d\mu^* \\ &= \int U_{rt}(\beta) U_t(\beta) \exp(\log(f_Y(y; \theta))) d\mu^* + \int U_r(\beta) U_{st}(\beta) \exp(\log(f_Y(y; \theta))) d\mu^* \\ &\quad + \int U_r(\beta) U_s(\beta) U_t(\beta) \exp(\log(f_Y(y; \theta))) d\mu^* = v_{rt,s} + v_{r,st} + v_{r,s,t} \end{aligned}$$

Τελικά,

$$\begin{aligned} \partial(v_{rs} + v_{r,s}) &= 0 \Rightarrow \frac{\partial}{\partial \beta^t} v_{rs} + \frac{\partial}{\partial \beta^t} v_{r,s} = 0 \\ \Rightarrow v_{rst} + v_{rs,t} + v_{rt,s} + v_{r,st} + v_{r,s,t} &= 0 \end{aligned}$$

όμως κάτω από τις συνθήκες ομαλότητας: $v_{rs,t} = v_{rt,s} = v_{r,st}$ έτσι:

$$v_{rs,t} + v_{rt,s} + v_{r,st} = v_{rs,t}[3] \text{ και } \acute{\alpha}\rho\alpha$$

$$v_{rst} + v_{rs,t}[3] + v_{r,s,t} = 0$$

Οι τρεις σχέσεις που αποδείχθηκαν είναι:

$$v_r = 0 \tag{3.29}$$

$$v_{rs} + v_{r,s} = 0 \tag{3.30}$$

$$v_{rst} + v_{rs,t}[3] + v_{r,s,t} = 0 \tag{3.31}$$

οι σχέσεις αυτές είναι γνωστές ως σχέσεις του Bartlett (Bartlett, 1953)

Κάποιες φορές, είναι προτιμότερο να δουλεύουμε με τους null cumulants αντί για τις null ροπές (null moments), για το συμβολισμό των null cumulants και προκειμένου να μη δημιουργηθεί περαιτέρω σύγχυση σχετικά με το συμβολισμό θα δεσμεύσουμε το γράμμα k . Ορίζουμε τους null cumulants ανά παρατήρηση (null cumulants *per observation*) ως: $k_{r,s} = n^{-1}E(U_r(\beta)U_s(\beta))$, $k_{r,st} = n^{-1}E(U_r(\beta)U_{st}(\beta))$, $k_{r,s,t} = n^{-1}E(U_r(\beta)U_s(\beta)U_t(\beta))$ και ούτω καθεξής.

Έτσι οι σχέσεις του Bartlett μπορούν να γραφούν και για τους null cumulants ανά παρατήρηση ως εξής:

$$k_r = 0 \quad (3.32)$$

$$k_{rs} + k_{r,s} = 0 \quad (3.33)$$

$$k_{rst} + k_{rs,t}[3] + k_{r,s,t} = 0 \quad (3.34)$$

Τέλος, για να ξεχωρίζουμε το στοχαστικό από το ντετερμινιστικό μέρος των μερικών παραγώγων του λογαρίθμου της συνάρτησης πιθανοφάνειας ορίζουμε τις κεντραρισμένες τυχαίες μεταβλητές.

$$H_{R_m} = U_{R_m} - v_{R_m} \quad (3.35)$$

Αυτές οι ποσότητες εμφανίζονται σε αναπτύγματα της πιθανοφάνειας και είναι χρήσιμες για τον εντοπισμό των ασυμπτωτικών τάξεων διάφορων όρων. Στη συνέχεια δίνεται ο πίνακας με τις ασυμπτωτικές τάξεις των κύριων ποσοτήτων που εμφανίζονται στη συνέχεια της εργασίας, όπως αυτός υπάρχει στο βιβλίο των Pace and Salvan (1997).

Ποσότητα	Ασυμπτωτική τάξη
U_r	$\mathcal{O}_p(n^{1/2})$
U_{rs}	$\mathcal{O}_p(n)$
$v_{r,s}$	$\mathcal{O}(n)$
$v^{r,s}$	$\mathcal{O}(n^{-1})$
v_{rst}	$\mathcal{O}(n)$
U_{rst}	$\mathcal{O}_p(n)$
$v_{rs,t}$	$\mathcal{O}(n)$
$E(U_r U_s H_{tu})$	$\mathcal{O}(n)$
$E(U_r U_s U_t H_{uv})$	$\mathcal{O}(n^2)$
$v^{r,s} U_r U_s$	$\mathcal{O}_p(1)$
$U_{R_m}, m \geq 2$	$\mathcal{O}_p(n)$
$v_{R_m}, m \geq 2$	$\mathcal{O}(n)$
$H_{R_m} = U_{R_m} - v_{R_m}$	$\mathcal{O}_p(n^{1/2})$

3.5 Αντιμετώπιση της μεροληψίας του εκτιμητή μέγιστης πιθανοφάνειας

Όπως έχει ήδη αναφερθεί η μέθοδος της μέγιστης πιθανοφάνειας αποτελεί μια από τις πιο διαδεδομένες στατιστικές μεθόδους και αποτελεί τη βασική μέθοδο για την εκτίμηση των παραμέτρων της παλινδρόμησης. Επιπλέον, ο εκτιμητής αυτός έχει επιθυμητές ασυμπτωτικές ιδιότητες που όμως ισχύουν κάτω από ορισμένες συνθήκες. Βασικό μειονέκτημα της μεθόδου αποτελεί το γεγονός ότι ο εκτιμητής αυτός είναι αμερόληπτος μόνο σε περιπτώσεις μεγάλων δειγμάτων. Αντίθετα, όταν δεν υπάρχει μεγάλο δείγμα ο εκτιμητής αυτός είναι μεροληπτικός. Επιπρόσθετα, οι King and Zeng (2001) σημειώνουν πως είναι πιθανό να

προκληθεί μεροληψία στα αποτελέσματα των εκτιμήσεων και σε περιπτώσεις όπου το δείγμα δεν είναι μικρό όμως είναι σπάνια η παρατήρηση του υπό μελέτη γεγονότος. Το συγκεκριμένο πρόβλημα παρουσιάζεται πιο αναλυτικά στην εργασία Altman et al. (2016). Σχετικά, με την περίπτωση της μεροληψίας σε μικρά δείγματα έχουν προταθεί διάφορα πιθανά μεγέθη δείγματος έτσι ώστε να μπορούμε να χαρακτηρίσουμε ένα δείγμα ως μικρό. Για παράδειγμα στην περίπτωση της λογιστικής παλινδρόμησης ένα μέγεθος δείγματος μικρότερο του 200 χαρακτηρίζεται ως μικρό από πολλούς συγγραφείς. Σχετικά, όμως με τη μεροληψία που προκαλείται στις εκτιμήσεις από το φαινόμενο των σπάνιων γεγονότων δεν υπάρχει κάποιο ξεκάθαρο μέτρο για τον προσδιορισμό του μεγέθους δείγματος. Η μεροληψία αυτού του είδους προκαλείται από την έλλειψη παρατηρούμενων γεγονότων, ενώ από την άλλη πλευρά η παρατήρηση των μη-γεγονότων υπερτερεί για κάποιες δεκάδες ή ακόμη και εκατοντάδες παρατηρήσεις. Ένα εμπειρικό κριτήριο για τον χαρακτηρισμό του υπό μελέτη γεγονότος ως σπάνιου είναι το κριτήριο των γεγονότων ανά μεταβλητή (events per variable) το οποίο συχνά συναντάται στη βιβλιογραφία με τον συμβολισμό EPV. Η συγκεκριμένη ποσότητα υπολογίζεται από τη σχέση:

$$EPV = \frac{\text{Αριθμός παρατηρούμενων γεγονότων}}{\text{Συνολικός αριθμός ανεξάρτητων μεταβλητών}}$$

Σε εργασία των Pavlou et al. (2016) τονίζεται πως τιμές αυτού του δείκτη μικρότερες από 10 υποδηλώνουν πως η μέθοδος της μέγιστης πιθανοφάνειας δεν δίνει αξιόπιστα αποτελέσματα και συνεπώς είναι απαραίτητο να προβούμε σε κάποια τροποποίηση, ενώ τιμές του δείκτη μεταξύ του 10 και του 20 υποδηλώνουν πως θα ήταν συνετό να κάνουμε κάποια τροποποίηση στον εκτιμητή. Το συγκεκριμένο κριτήριο είναι αρκετά εμπειρικό και κατακρίνεται από κάποιους ερευνητές όπως οι Van Smeden et al. (2018). Ακολούθως το ενδιαφέρον επικεντρώνεται στην παρουσίαση μεθόδων που μπορούν να εφαρμοστούν ώστε να αντιμετωπιστεί το

πρόβλημα της μεροληψίας και εφαρμόζονται είτε αν το δείγμα είναι μικρό είτε αν το γεγονός είναι σπάνιο.

Από το ανάπτυγμα της μεροληψίας το οποίο δίνεται αναλυτικά σε επόμενη ενότητα βλέπουμε ότι ο κυρίαρχος όρος της μεροληψίας έχει ασυμπτωτική τάξη $\mathcal{O}(n^{-1})$. Ορισμένοι ερευνητές συχνά παραβλέπουν το συγκεκριμένο πρόβλημα, ωστόσο κάποιες φορές είναι πιθανό η μεροληψία του εκτιμητή να είναι μεγάλη με αποτέλεσμα οι εκτιμήσεις να μην είναι αξιόπιστες. Η σημασία του εκτιμητή μέγιστης πιθανοφάνειας ανάγκασε τους ερευνητές να προσπαθήσουν να διορθώσουν το πρόβλημα της μεροληψίας. Οι μέθοδοι αντιμετώπισης του προβλήματος μπορούν να ομαδοποιηθούν σε δύο κατηγορίες:

1. Στις μεθόδους διόρθωσης της μεροληψίας.
2. Στις μεθόδους μείωσης της μεροληψίας.

3.5.1 Διόρθωση της μεροληψίας

Οι μέθοδοι διόρθωσης της μεροληψίας βασίζονται στα εξής δύο βήματα, η διαδικασία αυτή προτάθηκε από τον Efron (1975):

1. Υπολόγισε τον πρώτο όρο του ασυμπτωτικού αναπτύγματος της μεροληψίας, δηλαδή τον $\mathcal{O}(n^{-1})$ όρο.
2. Αφαίρεσε τον $\mathcal{O}(n^{-1})$ όρο από τους αρχικούς εκτιμητές μέγιστης πιθανοφάνειας, $\hat{\beta}$.

Όπως γίνεται αντιληπτό βασικό ρόλο σε μια μέθοδο διόρθωσης έχει ο υπολογισμός του πρώτου όρου του αναπτύγματος της μεροληψίας, μια έκφραση για αυτό τον όρο δίνεται στην εργασία των Cox and Snell (1968). Η μέθοδος υπολογισμού

δίνεται στη συνέχεια με παρόμοιο τρόπο που αυτή παρουσιάζεται στο βιβλίο των Pace and Salvani (1997).

Έστω $\hat{\beta}$ ο εκτιμητής μέγιστης πιθανοφάνειας του β . Τότε από το ανάπτυγμα Taylor της συνάρτησης score υπολογισμένη στο $\hat{\beta}$ έχουμε ότι:

$$\begin{aligned} U_r(\hat{\beta}) &= U_r(\beta) + U_{rs}(\beta) (\hat{\beta} - \beta)^s + \frac{1}{2} U_{rst}(\beta) (\hat{\beta} - \beta)^{st} + \frac{1}{6} U_{rstu}(\beta) (\hat{\beta} - \beta)^{stu} + \mathcal{O}_p(n^{-1}) \\ \Rightarrow 0 &= U_r(\beta) + U_{rs}(\beta) (\hat{\beta} - \beta)^s + \frac{1}{2} U_{rst}(\beta) (\hat{\beta} - \beta)^{st} + \frac{1}{6} U_{rstu}(\beta) (\hat{\beta} - \beta)^{stu} + \mathcal{O}_p(n^{-1}) \end{aligned}$$

Χρησιμοποιώντας το μετασχηματισμό (3.35) $\Rightarrow U_{R_m} = H_{R_m} + v_{R_m}$ έχουμε ότι:

$$\begin{aligned} 0 &= U_r(\beta) + (\hat{\beta} - \beta)^s (v_{rs} + H_{rs}) + \frac{1}{2} (v_{rst} + H_{rst}) (\hat{\beta} - \beta)^{st} + \\ &+ \frac{1}{6} (v_{rstu} + H_{rstu}) (\hat{\beta} - \beta)^{stu} + \mathcal{O}_p(n^{-1}) \\ v_{rs} &\stackrel{= -v_{r,s}}{\Rightarrow} 0 = U_r(\beta) + (\hat{\beta} - \beta)^s (-v_{r,s} + H_{rs}) + \frac{1}{2} (v_{rst} + H_{rst}) (\hat{\beta} - \beta)^{st} + \\ &+ \frac{1}{6} (v_{rstu} + H_{rstu}) (\hat{\beta} - \beta)^{stu} + \mathcal{O}_p(n^{-1}) \\ \Rightarrow (\hat{\beta} - \beta)^s v_{r,s} &= U_r(\beta) + (\hat{\beta} - \beta)^s H_{rs} + \frac{1}{2} (\hat{\beta} - \beta)^{st} v_{rst} + \frac{1}{2} (\hat{\beta} - \beta)^{st} H_{rst} + \\ &+ \frac{1}{6} (\hat{\beta} - \beta)^{stu} v_{rstu} + \mathcal{O}_p(n^{-1}) \end{aligned}$$

Για τον υπολογισμό των ασυμπτωτικών τάξεων των όρων που εμφανίζονται στο ανάπτυγμα και με χρήση του προηγούμενου πίνακα έχουμε ότι:

$$U_r = \mathcal{O}_p(n^{1/2}) \quad (3.36)$$

$$(\hat{\beta} - \beta)^s v_{r,s} = \mathcal{O}_p(n^{-1/2}) \cdot \mathcal{O}_p(n) = \mathcal{O}_p(n^{1/2}) \quad (3.37)$$

$$(\hat{\beta} - \beta)^s H_{rs} = \mathcal{O}_p(n^{-1/2}) \cdot \mathcal{O}_p(n^{1/2}) = \mathcal{O}_p(1) \quad (3.38)$$

$$(\hat{\beta} - \beta)^{st} v_{rst} = \mathcal{O}_p(n^{-1}) \cdot \mathcal{O}_p(n) = \mathcal{O}_p(1) \quad (3.39)$$

$$(\hat{\beta} - \beta)^{st} H_{rst} = \mathcal{O}_p(n^{-1}) \cdot \mathcal{O}_p(n^{1/2}) = \mathcal{O}_p(n^{-1/2}) \quad (3.40)$$

$$(\hat{\beta} - \beta)^{stu} v_{rstu} = \mathcal{O}_p(n^{-3/2}) \cdot \mathcal{O}_p(n) = \mathcal{O}_p(n^{-1/2}) \quad (3.41)$$

Τέλος, ο όρος $(\hat{\beta} - \beta)^{stu} H_{rstu} = \mathcal{O}_p(n^{-3/2}) \cdot \mathcal{O}_p(n^{1/2}) = \mathcal{O}_p(n^{-1})$ και επομένως ενσωματώνεται στο γενικό όρο $\mathcal{O}_p(n^{-1})$ του αναπτύγματος. Παρατηρούμε ότι στο προηγούμενο ανάπτυγμα η πτώση στις ασυμπτωτικές τάξεις γίνεται κατά $\mathcal{O}_p(n^{-1/2})$. Έτσι προκειμένου να αποφύγουμε τον ξεχωριστό υπολογισμό των ασυμπτωτικών τάξεων, όπως κάναμε προηγουμένως, θα χρησιμοποιούμε το σύμβολο $\dot{+}$ κάθε φορά που οι επερχόμενοι όροι αυτού του συμβόλου έχουν ασυμπτωτική τάξη μικρότερη κατά $\mathcal{O}_p(n^{-1/2})$ σε σχέση με τους προηγούμενους όρους και το σύμβολο $\ddot{+}$ κάθε φορά που οι επερχόμενοι όροι αυτού του συμβόλου έχουν ασυμπτωτική τάξη μικρότερη κατά $\mathcal{O}_p(n^{-1})$ σε σχέση με τους προηγούμενους όρους. Έτσι με τη χρήση αυτού του συμβολισμού το προηγούμενο ανάπτυγμα γράφεται τώρα ως εξής:

$$\begin{aligned} (\hat{\beta} - \beta)^s v_{r,s} &= U_r(\beta) \dot{+} (\hat{\beta} - \beta)^s H_{rs} + \frac{1}{2}(\hat{\beta} - \beta)^{st} v_{rst} \dot{+} \frac{1}{2}(\hat{\beta} - \beta)^{st} H_{rst} + \\ &+ \frac{1}{6}(\hat{\beta} - \beta)^{stu} v_{rstu} \dot{+} \mathcal{O}_p(n^{-1}) \\ \Rightarrow (\hat{\beta} - \beta)^s v_{r,s} v^{r,\tilde{r}} &= U_r(\beta) v^{r,\tilde{r}} \dot{+} (\hat{\beta} - \beta)^s H_{rs} v^{r,\tilde{r}} + \frac{1}{2}(\hat{\beta} - \beta)^{st} v_{rst} v^{r,\tilde{r}} \dot{+} \\ &\dot{+} \frac{1}{2}(\hat{\beta} - \beta)^{st} H_{rst} v^{r,\tilde{r}} + \frac{1}{6}(\hat{\beta} - \beta)^{stu} v_{rstu} v^{r,\tilde{r}} \dot{+} \mathcal{O}_p(n^{-2}) \end{aligned}$$

Λόγω της σύμβασης άθροισης του Einstein και με χρήση του δέλτα του Kronecker οι ποσότητες της προηγούμενης σχέσης γράφονται:

$$(\hat{\beta} - \beta)^s v_{r,s} v^{r,\tilde{r}} = (\hat{\beta} - \beta)^{\tilde{r}} = \mathcal{O}_p(n^{-1/2}) \quad (3.42)$$

$$U_r(\beta) v^{r,\tilde{r}} = U^{\tilde{r}}(\beta) = \mathcal{O}_p(n^{1/2}) \cdot \mathcal{O}_p(n^{-1}) = \mathcal{O}_p(n^{-1/2}) \quad (3.43)$$

$$H_{rs} v^{r,\tilde{r}} = H_s^{\tilde{r}} = \mathcal{O}_p(n^{1/2}) \cdot \mathcal{O}_p(n^{-1}) = \mathcal{O}_p(n^{-1/2}) \quad (3.44)$$

$$v_{rst} v^{r,\tilde{r}} = v_{st}^{\tilde{r}} = \mathcal{O}_p(1) \quad (3.45)$$

$$H_{rst} v^{r,\tilde{r}} = H_{st}^{\tilde{r}} = \mathcal{O}_p(n^{-1/2}) \quad (3.46)$$

$$v_{rstu} v^{r,\tilde{r}} = v_{stu}^{\tilde{r}} = \mathcal{O}_p(1) \quad (3.47)$$

Έτσι η τελευταία σχέση γράφεται ως εξής:

$$(\hat{\beta} - \beta)^{\tilde{r}} = U^{\tilde{r}}(\beta) + H_s^{\tilde{r}}(\hat{\beta} - \beta)^s + \frac{1}{2}v_{st}^{\tilde{r}}(\hat{\beta} - \beta)^{st} + \frac{1}{2}H_{st}^{\tilde{r}}(\hat{\beta} - \beta)^{st} + \\ + \frac{1}{6}v_{stu}^{\tilde{r}}(\hat{\beta} - \beta)^{stu} + \mathcal{O}_p(n^{-2})$$

Προκειμένου να επαναφέρουμε το προηγούμενο ανάπτυγμα σε όρους που περιλαμβάνουν άμεσα τις μερικές παραγώγους του λογαρίθμου της συνάρτησης πιθανοφάνειας εφαρμόζεται ο αλγόριθμος πίσω αντικατάστασης (backwards substitution or iterative substitution) για αναλυτική περιγραφή του αλγορίθμου παραπέμπουμε στους Pace and Salvani (1997). Επιπλέον, χρησιμοποιούμε έναν πιο συμπαγή συμβολισμό για τα διαφορικά του λογαρίθμου της συνάρτησης πιθανοφάνειας, δηλαδή όπου $U^{st}(\beta) = U^s(\beta)U^t(\beta)$ και ούτω καθεξής. Έτσι, έχουμε ότι:

$$(\hat{\beta} - \beta)^{\tilde{r}} = U^{\tilde{r}}(\beta) + \frac{1}{2}v_{st}^{\tilde{r}}U^{st}(\beta) + H_s^{\tilde{r}}U^s(\beta) + \frac{1}{6}(v_{stu}^{\tilde{r}} + 3v_{sw}^{\tilde{r}}v_{tu}^w)U^{stu}(\beta) \\ + \frac{1}{2}(H_{st}^{\tilde{r}} + 2v_{sw}^{\tilde{r}}H_t^w + H_w^{\tilde{r}}v_{st}^w)U^{st}(\beta) + H_w^{\tilde{r}}H_s^wU^s(\beta) + \mathcal{O}_p(n^{-2})$$

Για ευκολία ως προς τον συμβολισμό επιλέγουμε ως άνω δείκτη του προηγούμενου αναπτύγματος το r αντί για το $\tilde{r}$ και τότε η προηγούμενη σχέση γράφεται ισοδύναμα ως:

$$\overbrace{(\hat{\beta} - \beta)^r}^{\mathcal{O}_p(n^{-1/2})} = \overbrace{U^r(\beta) + \frac{1}{2}v_{st}^rU^{st}(\beta) + H_s^rU^s(\beta)}^{\mathcal{O}_p(n^{-1})} + \frac{1}{6}(v_{stu}^r + 3v_{sw}^rv_{tu}^w)U^{stu}(\beta) \\ + \frac{1}{2}(H_{st}^r + 2v_{sw}^rH_t^w + H_w^rv_{st}^w)U^{st}(\beta) + H_w^rH_s^wU^s(\beta) + \mathcal{O}_p(n^{-2})$$

Παίρνοντας τις αναμενόμενες τιμές στο ανάπτυγμα της διαφοράς που δόθηκε προηγουμένως μπορούμε να πάρουμε το ανάπτυγμα της μεροληψίας, όπου μας ενδιαφέρει να βρούμε την έκφραση για τον $\mathcal{O}(n^{-1})$ όρο. Έτσι, έχουμε ότι:

$$E(\hat{\beta} - \beta)^r = \frac{1}{2}v_{st}^rE(U^{st}(\beta)) + E(H_s^rU^r(\beta)) + \mathcal{O}(n^{-2})$$

Η ασυμπτωτική τάξη των $\mathcal{O}_p(n^{-3/2})$ όρων του αναπτύγματος της διαφοράς γίνεται $\mathcal{O}(n^{-2})$ υπό την επίδραση της μέσης τιμής ενδεικτικά αποδεικνύουμε εδώ για τη μέση τιμή του πρώτου κατά σειρά όρου που εμφανίζεται στο προηγούμενο ανάπτυγμα της διαφοράς. Πράγματι,

$$\begin{aligned} E(v_{stu}^r U^{stu}(\beta)) &= E(v_{stu}^{r,\tilde{r}} v^{s,\tilde{s}} U^s(\beta) U^t(\beta) U^u(\beta)) = v_{stu}^{r,\tilde{r}} E(U^s(\beta) U^t(\beta) U^u(\beta)) \\ &= v_{stu}^{r,\tilde{r}} E\left(U_s(\beta) v^{s,\tilde{s}} U_t(\beta) v^{t,\tilde{t}} U_u(\beta) v^{u,\tilde{u}}\right) = \underbrace{v_{stu}^{r,\tilde{r}}}_{\mathcal{O}(n)} \underbrace{v^{s,\tilde{s}} v^{t,\tilde{t}} v^{u,\tilde{u}}}_{\mathcal{O}(n^{-4})} \underbrace{v_{s,t,u}^{s,\tilde{s},\tilde{t},\tilde{u}}}_{\mathcal{O}(n)} = \mathcal{O}(n^{-2}) \end{aligned}$$

Σχετικά τώρα με τους όρους $\mathcal{O}(n^{-1})$ όρους του προηγούμενου αναπτύγματος:

$$\begin{aligned} E(U^{st}(\beta)) &= E(U^s(\beta) U^t(\beta)) = E(v^{s,\tilde{s}} U_s(\beta) v^{t,\tilde{t}} U_t(\beta)) = \\ &= v^{s,\tilde{s}} v^{t,\tilde{t}} E(U_s(\beta) U_t(\beta)) = v^{s,\tilde{s}} v^{t,\tilde{t}} v_{s,t}^{s,\tilde{s}} = v^{s,t} \end{aligned} \quad (3.48)$$

$$\begin{aligned} E(H_s^r U^s(\beta)) &= E(v^{r,\tilde{r}} H_{rs} v^{s,\tilde{s}} U_s(\beta)) = v^{r,\tilde{r}} v^{s,\tilde{s}} E((U_{rs}(\beta) + v_{rs}) U_s(\beta)) = \\ &= v^{r,\tilde{r}} v^{s,\tilde{s}} E(U_{rs}(\beta) U_s(\beta) + v^{r,\tilde{r}} v^{s,\tilde{s}} v_{rs} E(U_s(\beta))) = v^{r,\tilde{r}} v^{s,\tilde{s}} E(U_{rs}(\beta) U_s(\beta)) + \\ &+ v^{r,\tilde{r}} v^{s,\tilde{s}} v_{rs} \cdot 0 = v^{r,\tilde{r}} v^{s,\tilde{s}} v_{rs,s} \end{aligned} \quad (3.49)$$

Όπως έχουμε δει και νωρίτερα η ευκολία που μας δίνει η χρήση του index notation μας επιτρέπει να αλλάζουμε τους συμβολισμούς χωρίς να μεταβάλλεται το αποτέλεσμα, με μοναδική προϋπόθεση να είμαστε προσεκτικοί ως προς τη σύμβαση άθροισης του Einstein έτσι για να "ελαφρύνουμε" το συμβολισμό μπορούμε να γράψουμε ότι: $v^{r,\tilde{r}} v^{s,\tilde{s}} v_{rs,s} = v^{r,s} v^{t,u} v_{st,u}$, αφού και στις δύο πλευρές της ισότητας έχουμε μια ποσότητα που έχει free index το r. Έτσι τελικά, $E(H_s^r U^s(\beta)) = v^{r,s} v^{t,u} v_{st,u}$. Τότε,

$$E(\hat{\beta} - \beta)^r = \frac{1}{2} v_{st}^r v^{s,t} + v^{r,s} v^{t,u} v_{st,u} + \mathcal{O}(n^{-2})$$

Όμως, $v_{st}^r v^{s,t} = v^{r,\tilde{r}} v_{st}^{s,t} = v^{r,s} v^{t,u} v_{stu}$.

Έτσι, έχουμε ότι

$$\begin{aligned} E(\hat{\beta} - \beta)^r &= \frac{1}{2} v^{r,s} v^{t,u} v_{stu} + v^{r,s} v^{t,u} v_{st,u} + \mathcal{O}(n^{-2}) \\ \Rightarrow E(\hat{\beta} - \beta)^r &= \frac{1}{2} v^{r,s} v^{t,u} (v_{stu} + 2 v_{st,u}) + \mathcal{O}(n^{-2}) \end{aligned} \quad (3.50)$$

Έχουμε λοιπόν υπολογίσει τον πρώτο όρο του αναπτύγματος της μεροληψίας, με τη βοήθεια των από κοινού null moments μπορούμε τώρα με τη βοήθεια του ορισμού των null cumulants *ανά παρατήρηση* και των σχέσεων του Bartlett να γράψουμε την προηγούμενη σχέση σε όρους null cumulants *ανά παρατήρηση*. Έτσι αφού, $k_{r,s} = \frac{1}{n} E(U_r U_s) = \frac{1}{n} v_{r,s} \Rightarrow v^{r,s} = \frac{1}{n} k^{r,s}$, εδώ αξίζει να σημειωθεί πως η ποσότητα $k_{r,s}$ είναι γνωστή στη βιβλιογραφία ως πίνακας πληροφορίας του Fisher *ανά παρατήρηση*. Επιπλέον, από τις σχέσεις του Bartlett έχουμε ότι:

$$\begin{aligned} v_{rst} + v_{rs,t}[3] + v_{r,s,t} &= 0 \Rightarrow v_{rst} + v_{rs,t}[2] + v_{r,st} + v_{r,s,t} = 0 \\ \Rightarrow v_{rst} + v_{rs,t}[2] &= -v_{r,st} - v_{r,s,t} \end{aligned}$$

Ενώ, $v_{rs,t} = n k_{rs,t}$ και $v_{r,s,t} = n k_{r,s,t}$ Άρα:

$$\begin{aligned} E(\hat{\beta} - \beta)^r &= \frac{1}{2} v^{r,s} v^{t,u} (v_{stu} + 2 v_{st,u}) + \mathcal{O}(n^{-2}) \\ \Rightarrow E(\hat{\beta} - \beta)^r &= \frac{1}{2} v^{r,s} v^{t,u} (-v_{s,tu} - v_{s,t,u}) + \mathcal{O}(n^{-2}) \\ \Rightarrow E(\hat{\beta} - \beta)^r &= \frac{1}{2} \frac{1}{n} k^{r,s} \frac{1}{n} k^{t,u} (-n k_{s,tu} - n k_{s,t,u}) + \mathcal{O}(n^{-2}) \\ \Rightarrow E(\hat{\beta} - \beta)^r &= -\frac{1}{n} k^{r,s} k^{t,u} \left(\frac{k_{s,tu} + k_{s,t,u}}{2} \right) + \mathcal{O}(n^{-2}) \end{aligned} \quad (3.51)$$

Η τελευταία σχέση αποτελεί τη σχέση που έδωσαν οι Cox and Snell (1968) για τον $\mathcal{O}(n^{-1})$ όρο της μεροληψίας του εκτιμητή μέγιστης πιθανοφάνειας. Είναι φανερό από την τελευταία σχέση πως η αναμενόμενη τιμή του εκτιμητή μέγιστης

πιθανοφάνειας $\hat{\beta}$ με β την αληθινή τιμή της παραμέτρου μπορεί να γραφεί ως:

$$E(\hat{\beta}) = \beta + \frac{b_1(\beta)}{n} + \frac{b_2(\beta)}{n^2} + \dots \quad (3.52)$$

όπου n είναι το μέγεθος του δείγματος και b_t , $t = 1, 2, \dots$ είναι $\mathcal{O}(1)$ συναρτήσεις της πραγματικής τιμής του β . Έτσι σύμφωνα με το σκεπτικό της μεθόδου διόρθωσης της μεροληψίας από τη σχέση:

$$E(\hat{\beta}) - \frac{b_1(\beta)}{n} = \beta + \frac{b_2(\beta)}{n^2} + \dots \quad (3.53)$$

προκύπτει ο βελτιωμένος ως προς τη μεροληψία εκτιμητής

$$\hat{\beta}_{BC} = \hat{\beta} - \frac{b_1(\hat{\beta})}{n} \quad (3.54)$$

Οι Anderson and Richardson (1979) και Schaefer (1983) βασιζόμενοι στη σχέση των Cox and Snell υπολόγισαν τον πρώτο όρο της μεροληψίας στην περίπτωση της λογιστικής παλινδρόμησης και έδωσαν τους βελτιωμένους εκτιμητές. Οι δύο μελέτες καταλήγουν στο συμπέρασμα πως στο μοντέλο της λογιστικής παλινδρόμησης μια τέτοια διόρθωση είναι αναγκαία σε περιπτώσεις μικρών δειγμάτων, αφού τότε η μεροληψία του εκτιμητή μέγιστης πιθανοφάνειας είναι μεγάλη και επηρεάζει την αξιοπιστία των αποτελεσμάτων. Οι Cordeiro-McCullagh (1991) επέκτειναν αυτά τα αποτελέσματα και διόρθωσαν τη μεροληψία του εκτιμητή μέγιστης πιθανοφάνειας για την οικογένεια των γενικευμένων γραμμικών μοντέλων. Στη συνέχεια δίνεται η προσέγγιση των Cordeiro & McCullagh (1991) για την περίπτωση των γενικευμένων γραμμικών μοντέλων με κανονική συνάρτηση σύνδεσης (canonical link function) όπου η παράμετρος διασποράς (dispersion parameter) θεωρείται γνωστή και ίση με 1.

3.5.2 Η μέθοδος των Cordeiro & McCullagh

Σύμφωνα με το αποτέλεσμα των Cox and Snell (1968) ισχύει ότι

$$E(\hat{\beta} - \beta)^r \approx -\frac{1}{2}v^{r,s}v^{t,u}(v_{s,tu} + v_{s,t,u})$$

όμως στην περίπτωση όπου έχουμε κανονική συνάρτηση σύνδεσης ισχύει ότι: $v_{s,tu} = 0$. Πράγματι,

$$v_{s,tu} = E(U_r(\beta)U_{st}(\beta)) = E\left(\frac{\partial l}{\partial \beta^r} \frac{\partial^2 l}{\partial \beta^s \partial \beta^t}\right)$$

όμως στην περίπτωση της κανονικής συνάρτησης σύνδεσης η ποσότητα $\frac{\partial^2 l}{\partial \beta^s \partial \beta^t}$ δεν εξαρτάται από τα δεδομένα Y_i επομένως μπορεί να "βγει" από την αναμενόμενη τιμή. Ετσι, ύστερα από αυτή τη διαπίστωση έχουμε πλέον ότι:

$$v_{s,tu} = E(U_r(\beta)U_{st}(\beta)) = E\left(\frac{\partial l}{\partial \beta^r} \frac{\partial^2 l}{\partial \beta^s \partial \beta^t}\right) = \frac{\partial^2 l}{\partial \beta^s \partial \beta^t} E\left(\frac{\partial l}{\partial \beta^r}\right) = 0,$$

αφού όμως $E(U_r(\beta)) = 0$. Άρα, έχουμε ότι:

$$E(\hat{\beta} - \beta)^r \approx -\frac{1}{2}v^{r,s}v^{t,u}v_{s,t,u} \quad (3.55)$$

όπου $v^{r,s}$ ο αντίστροφος του πίνακα πληροφορίας του Fisher, δηλαδή σε μορφή πινάκων ο πίνακας $(X^T W X)^{-1}$ με $W = Cov(Y)$ για την περίπτωση που μελετάμε. Εάν, θεωρήσουμε ότι το τυχόν στοιχείο του πίνακα παλινδρόμησης γράφεται σε index notation μορφή ως x_r^i , δηλαδή $X = [x_r^i]$ και αφού στην περίπτωση της κανονικής συνάρτησης σύνδεσης έχουμε ότι: $\frac{\partial l}{\partial \beta^r} = \frac{(y_i - \mu_i)x_r^i}{\alpha(\phi)}$. Επιπλέον, υποθέσαμε ότι $a(\phi) = 1 \Rightarrow \frac{\partial l}{\partial \beta^r} = (Y_i - \mu_i)x_r^i$. Τότε

$$v_{s,t,u} = E(U_s(\beta)U_t(\beta)U_u(\beta)) = E\left(\frac{\partial l}{\partial \beta^s} \frac{\partial l}{\partial \beta^t} \frac{\partial l}{\partial \beta^u}\right) = \sum_{i=1}^n x_s^i x_t^i x_u^i E(Y_i - \mu_i)^3$$

και αφού, λόγω της (3.15), $E(Y_i - \mu_i)^3 = k_{3i}$, όπου k_{3i} ο τρίτης τάξης cumulant της μεταβλητής απόκρισης Y_i έχουμε τελικά ότι

$$v_{s,t,u} = \sum_{i=1}^n x_s^i x_t^i x_u^i k_{3i} \quad (3.56)$$

Επιπλέον, αν θεωρήσουμε ότι $b_s = -\frac{1}{2} v_{s,t,u} v^{t,u} = -\frac{1}{2} \sum_{i=1}^n x_s^i k_{3i} x_t^i x_u^i v^{t,u}$ και αφού $v^{t,u}$ είναι ο αντίστροφος του πίνακα πληροφορίας του Fisher η ποσότητα $x_t^i x_u^i v^{t,u}$ μπορεί να γραφεί σε μορφή πινάκων ως ο $n \times n$ πίνακας:

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

επομένως η ποσότητα $x_t^i x_u^i v^{t,u}$ είναι το i -στοιχείο του πίνακα Q και έτσι η σχέση για το b_s γράφεται τώρα ως:

$$b_s = -\frac{1}{2} \sum_{i=1}^n x_s^i k_{2i} Q_{ii} \frac{k_{3i}}{k_{2i}} \quad (3.57)$$

ενώ επιπλέον γνωρίζουμε ότι $k_2 = \text{Cov}(Y) = W$. Τότε η ποσότητα $x_s^i k_{2i} Q_{ii}$ είναι το i -στοιχείο του πίνακα $X^T W \xi$, όπου $\xi_i = -\frac{1}{2} Q_{ii} \frac{k_{3i}}{k_{2i}}$. Έτσι, το b_s έχει πινακική μορφή: $X^T W \xi$. Επανερχόμαστε τώρα στην αρχική σχέση (3.54):

$$E(\hat{\beta} - \beta)^r = b^r = -\frac{1}{2} v^{r,s} v^{t,u} v_{s,t,u} = v^{r,s} b_s$$

Παρατηρούμε ότι η πινακική μορφή της τελευταίας σχέσης είναι:

$$b = E(\hat{\beta} - \beta) = (X^T W X)^{-1} X^T W \xi \quad (3.58)$$

Η μορφή της μεροληψίας που προέκυψε είναι γνωστή στη βιβλιογραφία, αφού εμφανίζεται στη θεωρία των κανονικών γραμμικών μοντέλων. Πιο συγκεκριμένα η ποσότητα αυτή ονομάζεται επανασταθμισμένος εκτιμητής ελαχίστων τετραγώνων (weighted least squares estimator). Έτσι το βασικό συμπέρασμα

που έδωσε η προσέγγιση των Cordeiro & McCullagh (1991) είναι ότι μπορούμε να υπολογίσουμε το διάνυσμα της μεροληψίας προσαρμόζοντας ένα κανονικό μοντέλο γραμμικής παλινδρόμησης με εξαρτημένη μεταβλητή ξ και ανεξάρτητη το διάνυσμα X , ενώ το W αποτελεί το διάνυσμα βάρους (weight vector). Στην παρουσίαση της μεθόδου θεωρήσαμε για ευκολία ότι $\alpha(\phi) = 1$. Στην περίπτωση όπου $\alpha(\phi) \neq 1$ η μόνη διαφορά που θα παρατηρηθεί βρίσκεται στο τυπικό στοιχείο του W το οποίο αντί για $\text{Var}(Y_i)$ γίνεται $\frac{\text{Var}(Y_i)}{(\alpha(\phi))^2}$. Έτσι η μοναδική ουσιαστική υπόθεση που κάναμε αφορά την κανονική συνάρτηση σύνδεσης (canonical link function). Τα αποτελέσματα μπορούν να γενικευτούν και σε περιπτώσεις μη-κανονικής συνάρτησης σύνδεσης (non canonical link function) όμως στην παρούσα εργασία ασχολούμαστε μόνο με τις περιπτώσεις κανονικής συνάρτησης σύνδεσης που αποτελούν και τις πιο δημοφιλείς σε στατιστικές εφαρμογές.

3.5.3 Μείωση της μεροληψίας

Η βασική διαφορά μεταξύ των μεθόδων μείωσης της μεροληψίας και διόρθωσης της είναι πως οι μέθοδοι μείωσης της μεροληψίας δεν εξαρτώνται άμεσα από τους αρχικούς εκτιμητές μέγιστης πιθανοφάνειας. Ένας αρκετά γνωστός εκτιμητής μείωσης της μεροληψίας είναι ο εκτιμητής του Haldane (Haldane, 1955). Το σκεπτικό της μεθόδου αυτής είναι το εξής: Αν y είναι ο παρατηρούμενος αριθμός επιτυχιών σε μια διωνυμική δοκιμασία n επαναλήψεων με πιθανότητα επιτυχίας π . Ο προφανής εκτιμητής του log-odds ratio $\beta = \log\left(\frac{\pi}{1-\pi}\right)$ είναι ο $\hat{\beta} = \log\left(\frac{y}{n-y}\right)$, ο οποίος είναι επιπλέον και ο εκτιμητής μέγιστης πιθανοφάνειας.

Ωστόσο, ο $\hat{\beta}$ έχει το μειονέκτημα πως δεν παίρνει πεπερασμένη τιμή εάν $y = 0$ ή $y = n$ και εφόσον υπάρχει πεπερασμένη πιθανότητα το y να πάρει μια από τις δύο αυτές τιμές αυτόματα συμπεραίνουμε πως υπάρχει πεπερασμένη πιθανότητα ο εκτιμητής να μην ορίζεται. Η μέθοδος αυτή μας βοηθάει να ξεπεράσουμε αυτό

το πρόβλημα, όμως όπως θα δειχθεί αφαιρεί και ένα κομμάτι της μεροληψίας του εκτιμητή, πιο συγκεκριμένα ο εκτιμητής που προκύπτει έχει ασυμπτωτικό ανάπτυγμα μεροληψίας με κυρίαρχο όρο τάξης $\mathcal{O}(n^{-2})$.

Οι Haldane (1955) & Anscombe (1956) όρισαν τον νέο εκτιμητή:

$$\tilde{\beta}_\alpha = \log \left(\frac{y + a}{n - y + a} \right), \quad \alpha \in \mathbb{R} \quad (3.59)$$

και αναζήτησαν το α ώστε η αναμενόμενη τιμή του $\tilde{\beta}_\alpha$ να είναι κοντά στο β . Τότε, αφού $Y \sim B(n, \pi) \Rightarrow E(Y) = n\pi$ και $Var(Y) = n\pi(1 - \pi)$, ενώ επιπλέον $E\{(Y - n\pi)^3\} = n\pi(1 - \pi)(1 - 2\pi)$. Θέτουμε $U = \frac{Y - n\pi}{\sqrt{n}}$, τότε από το κεντρικό οριακό θεώρημα έχουμε ότι: $\frac{U}{\sqrt{\pi(1 - \pi)}} \xrightarrow{d} N(0, 1) \Rightarrow U = \mathcal{O}_p(1)$. Επιπλέον, λύνοντας ως προς Y έχουμε ότι: $Y = n\pi + U\sqrt{n}$. Αντικαθιστώντας, στη σχέση (3.59) και ύστερα από πράξεις έχουμε ότι:

$$\tilde{\beta}_\alpha - \beta = \log \left(1 + \left(\frac{U}{\sqrt{n}\pi} + \frac{\alpha}{n\pi} \right) \right) - \log \left(1 - \left(\frac{U}{\sqrt{n}q} - \frac{\alpha}{nq} \right) \right) \quad (3.60)$$

όπου $q = 1 - \pi$. Από το ανάπτυγμα Taylor του πρώτου και του δεύτερου όρου της σχέσης (3.60) έχουμε ότι:

$$\begin{aligned} \tilde{\beta}_\alpha - \beta &= \frac{U}{\sqrt{n}\pi} + \frac{\alpha}{n\pi} - \frac{U^2}{2n\pi^2} - \frac{\alpha U}{n^{3/2}\pi^2} + \frac{U^3}{3n^{3/2}\pi^3} + \\ &+ \frac{U}{\sqrt{n}q} - \frac{\alpha}{nq} + \frac{U^2}{2nq^2} - \frac{\alpha U}{n^{3/2}q^2} + \frac{U^3}{3n^{3/2}q^3} + \mathcal{O}_p(n^{-2}) \end{aligned} \quad (3.61)$$

Όμως, $E(U) = 0$, $E(U^2) = \pi(1 - \pi) = \pi q$ και $E(U^3) = n^{-1/2}\pi q(1 - 2\pi)$. Έτσι,

παίρνοντας την αναμενόμενη τιμή της διαφοράς $\tilde{\beta}_\alpha - \beta$ στην (3.61) βρίσκουμε ότι:

$$E(\tilde{\beta}_\alpha - \beta) = \frac{(1 - 2\pi)(2\alpha - 1)}{2n\pi q} + \mathcal{O}(n^{-2}) \quad (3.62)$$

είναι φανερό από την τελευταία σχέση πως η επιλογή του $\alpha = \frac{1}{2}$ είναι η επιθυμητή, γιατί τότε απαλείφεται ο $\mathcal{O}(n^{-1})$ όρος της μεροληψίας και τότε αυτή έχει ασυμπτωτική τάξη $\mathcal{O}(n^{-2})$, συνεπώς ο νέος εκτιμητής είναι βελτιωμένος ως προς τη μεροληψία, ενώ επιπλέον παίρνει πάντα πεπερασμένες τιμές. Έτσι, ο βελτιωμένος και πάντα πεπερασμένος εκτιμητής του log-odds ratio που κατέληξαμε είναι ο

$$\tilde{\beta}_{1/2} = \log \left(\frac{y + \frac{1}{2}}{n - y + \frac{1}{2}} \right) \quad (3.63)$$

3.5.4 Η μέθοδος μείωσης της μεροληψίας του Firth

Σε αυτήν την παράγραφο θα περιγραφεί η μέθοδος του Firth (Firth,1993) για τη μείωση της μεροληψίας του εκτιμητή μέγιστης πιθανοφάνειας στην περίπτωση της εκτίμησης των παραμέτρων της παλινδρόμησης στα γενικευμένα γραμμικά μοντέλα, με κανονική συνάρτηση σύνδεσης. Η ιδέα πίσω από τη μέθοδο του Firth είναι η ποινικοποίηση της συνάρτησης score (score function) με μία συνάρτηση $A_r(\beta)$ για την οποία οι υποθέσεις που γίνονται είναι πως:

1. Έχει ασυμπτωτική τάξη $\mathcal{O}_p(1)$.
2. Είναι τουλάχιστον τρεις φορές διαφορίσιμη.
3. Μπορεί να εξαρτάται από τα δεδομένα.

Η συγκεκριμένη μέθοδος είναι μια πολύ σημαντική μέθοδος. Ο Firth καταλήγει, όπως θα δείχθει στη συνέχεια, στο συμπέρασμα πως η κατάλληλη συνάρτηση ποινικοποίησης είναι η: $|I(\beta)|^{1/2}$ που είναι γνωστή στη βιβλιογραφία ως εκ των προτέρων Jeffrey (Jeffreys invariant prior). Επιπλέον, ο νέος εκτιμητής

διατηρεί τις ιδιότητες της ασυμπτωτικής συνέπειας και της κανονικότητας. Επιπρόσθετα, μια πολύ σπουδαία ιδιότητα που αποκτά ο βελτιωμένος εκτιμητής είναι πως είναι πάντοτε πεπερασμένος. Έτσι, η μέθοδος του Firth λύνει το πρόβλημα της ύπαρξης (έτσι όπως ορίζεται η ύπαρξη του από τους Albert & Anderson, 1984) του εκτιμητή μέγιστης πιθανοφάνειας σε περιπτώσεις όπου στα δεδομένα εμφανίζεται το φαινόμενο του πλήρη διαχωρισμού (complete separation) ή το πρόβλημα του σχεδόν πλήρη διαχωρισμού (quasi complete separation). Αναλυτική περιγραφή των δύο αυτών προβλημάτων και η επίλυση τους για την περίπτωση της λογιστικής παλινδρόμησης θα δοθεί σε επόμενο κεφάλαιο, προς το παρόν προσανατολιζόμαστε στην παρουσίαση της μεθόδου ποινικοποίησης του Firth. Για την παρουσίαση αυτών των αποτελεσμάτων έχουμε βασιστεί σε μεγάλο βαθμό στην διδακτορική διατριβή του (Kosmidis PhD thesis, 2007).

Έστω λοιπόν, η τροποποιημένη συνάρτηση score (score function) ώστε

$$U_r^*(\beta) = U_r(\beta) + A_r(\beta) \quad (3.64)$$

Έστω επίσης ότι $\hat{\beta}$ είναι ο αρχικός εκτιμητής μέγιστης πιθανοφάνειας, δηλαδή ότι $U_r(\hat{\beta}) = 0$ και β^* είναι ο βελτιωμένος εκτιμητής μέγιστης πιθανοφάνειας δηλαδή $U_r^*(\beta^*) = 0$. Θεωρούμε τη διαφορά $\gamma = \beta^* - \beta$. Από το ανάπτυγμα Taylor και με τη χρήση index notation έχουμε ότι:

$$\begin{aligned} U_r^*(\beta^*) &= U_r^*(\beta) + \gamma^s U_{rs}^*(\beta) + \frac{1}{2} \gamma^{st} U_{rst}^*(\beta) + \frac{1}{6} \gamma^{stu} U_{rstu}^*(\beta) + \mathcal{O}_p(n^{-1}) \\ \Rightarrow 0 &= U_r(\beta) + A_r(\beta) + \gamma^s U_{rs}(\beta) + \gamma^s A_{rs}(\beta) + \frac{1}{2} \gamma^{st} U_{rst}(\beta) + \frac{1}{6} \gamma^{stu} U_{rstu}(\beta) + \mathcal{O}_p(n^{-1}). \end{aligned}$$

Στην παραπάνω έκφραση ο όρος $\gamma^{st} A_{rst}(\beta)$ έχει τάξη $\mathcal{O}_p(n^{-1})$, ενώ ο όρος $\gamma^{stu} A_{rstu}(\beta)$ έχει τάξη $\mathcal{O}_p(n^{-3/2})$ και έτσι και οι δύο είναι ενσωματωμένοι στον $\mathcal{O}_p(n^{-1})$ όρο του αναπτύγματος. Από το μετασχηματισμό $U_{R_m} = v_{R_m} + H_{R_m}$, σχέση (3.35), και από τη σχέση $v_{rs} = -v_{r,s}$ (λόγω των σχέσεων του Bartlett)

έχουμε ότι:

$$\begin{aligned}
\Rightarrow 0 &= U_r(\beta) + A_r(\beta) + \gamma^s(-v_{r,s} + H_{rs}) + \gamma^s A_{rs}(\beta) + \frac{1}{2}\gamma^{st}(v_{rst} + H_{rst}) + \\
&+ \frac{1}{6}\gamma^{stu}(v_{rstu} + H_{rstu}) + \mathcal{O}_p(n^{-1}) \\
\Rightarrow \gamma^s v_{r,s} &= U_r(\beta) + A_r(\beta) + \gamma^s H_{rs} + \frac{1}{2}\gamma^{st}v_{rst} + \gamma^s A_{rs}(\beta) + \frac{1}{2}\gamma^{st}H_{rst} + \\
&+ \frac{1}{6}\gamma^{stu}v_{rstu} + \mathcal{O}_p(n^{-1}), \quad (\text{ο όρος } \gamma^{stu}H_{rstu} \text{ έχει ασυμπτωτική τάξη } \mathcal{O}_p(n^{-1}) \\
&\text{ και έτσι ενσωματώνεται στον γενικό όρο του ανάπτυγματος})
\end{aligned}$$

Ακολουθώντας τα ίδια ακριβώς βήματα όπως και στο ανάπτυγμα της διαφοράς $\tilde{\beta} - \beta$ έχουμε ότι:

$$\begin{aligned}
\gamma^{\tilde{r}} &= U^{\tilde{r}}(\beta) + A^{\tilde{r}}(\beta) + H_s^{\tilde{r}}\gamma^s + \frac{1}{2}v_{st}^{\tilde{r}}\gamma^{st} + A_s^{\tilde{r}}\gamma^s + \\
&+ \frac{1}{2}H_{st}^{\tilde{r}}\gamma^{st} + \frac{1}{6}v_{stu}^{\tilde{r}}\gamma^{stu} + \mathcal{O}_p(n^{-2})
\end{aligned}$$

όπου:

$$U^{\tilde{r}} = U_r v^{r,\tilde{r}} = \mathcal{O}_p(n^{1/2}) \cdot \mathcal{O}_p(n^{-1}) = \mathcal{O}_p(n^{-1/2}) \quad (3.65)$$

$$A^{\tilde{r}} = A_r v^{r,\tilde{r}} = \mathcal{O}_p(1) \cdot \mathcal{O}_p(n^{-1}) = \mathcal{O}_p(n^{-1}) \quad (3.66)$$

$$A_s^{\tilde{r}} = A_{rs} v^{r,\tilde{r}} = \mathcal{O}_p(1) \cdot \mathcal{O}_p(n^{-1}) = \mathcal{O}_p(n^{-1}) \quad (3.67)$$

$$H_s^{\tilde{r}} = H_{rs} v^{r,\tilde{r}} = \mathcal{O}_p(n^{1/2}) \cdot \mathcal{O}_p(n^{-1}) = \mathcal{O}_p(n^{-1/2}) \quad (3.68)$$

$$H_{st}^{\tilde{r}} = H_{rst} v^{r,\tilde{r}} = \mathcal{O}_p(n^{1/2}) \cdot \mathcal{O}_p(n^{-1}) = \mathcal{O}_p(n^{-1/2}) \quad (3.69)$$

$$v_{st}^{\tilde{r}} = v_{rst} v^{r,\tilde{r}} = \mathcal{O}_p(n) \cdot \mathcal{O}_p(n^{-1}) = \mathcal{O}_p(1) \quad (3.70)$$

$$v_{stu}^{\tilde{r}} = v_{rstu} v^{r,\tilde{r}} = \mathcal{O}_p(n) \cdot \mathcal{O}_p(n^{-1}) = \mathcal{O}_p(1) \quad (3.71)$$

ενώ και πάλι για ευκολία ως προς το συμβολισμό μπορούμε να επιλέξουμε ως τελικό άνω δείκτη (upper index) το r αντι για το $\tilde{r}$. Έτσι, το προηγούμενο ανάπτυγμα

γράφεται τώρα ως:

$$\begin{aligned}\gamma^r &= U^r(\beta) + A^r(\beta) + H_s^r \gamma^s + \frac{1}{2} v_{st}^r \gamma^{st} + A_s^r \gamma^s + \\ &+ \frac{1}{2} H_{st}^r \gamma^{st} + \frac{1}{6} v_{stu}^r \gamma^{stu} + \mathcal{O}_p(n^{-2}) \\ \Rightarrow \gamma^r &= U^r(\beta) + A^r(\beta) + H_s^r \gamma^s + \frac{1}{2} v_{st}^r \gamma^{st} + \mathcal{O}_p(n^{-3/2})\end{aligned}$$

Εφαρμόζοντας και πάλι την μέθοδο αντιστροφής (iterative substitution method)

η τελευταία σχέση γράφεται:

$$\gamma^r = U^r(\beta) + A^r(\beta) + H_s^r U^s + \frac{1}{2} v_{st}^r U^{st} + \mathcal{O}_p(n^{-3/2}) \quad (3.72)$$

Παίρνοντας τις αναμενόμενες τιμές στην (3.71) έχουμε ότι:

$$E(\gamma^r) = E(A^r(\beta)) + E(H_s^r U_s) + \frac{1}{2} v_{st}^r E(U^{st}) + \mathcal{O}(n^{-3/2})$$

Όμως έχουμε δείξει ότι:

$$E(H_s^r U_s(\beta)) = v^{r,s} v^{t,u} v_{stu}$$

$$E(U^{st}(\beta)) = v^{s,t}$$

$$v^{s,t} v_{st}^r = v^{r,s} v^{t,u} v_{stu}$$

Επομένως,

$$\begin{aligned}E(\gamma^r) &= E(A^r(\beta)) + \frac{1}{2} v^{r,s} v^{t,u} (v_{stu} + 2v_{st,u}) + \mathcal{O}(n^{-3/2}) \\ \Rightarrow E(\gamma^r) &= E(v^{r,s} A_s(\beta)) - \frac{1}{2} v^{r,s} v^{t,u} (v_{s,tu} + v_{s,t,u}) + \mathcal{O}(n^{-3/2}) \\ \Rightarrow E(\gamma^r) &= v^{r,s} E(A_s(\beta)) - \frac{1}{2} v^{r,s} v^{t,u} (v_{s,tu} + v_{s,t,u}) + \mathcal{O}(n^{-3/2})\end{aligned} \quad (3.73)$$

όπου από την πρώτη σχέση πήγαμε στη δεύτερη με εφαρμογή των τύπων του Bartlett. Τέλος, μπορούμε να φέρουμε τον προηγούμενο ανάπτυγμα σε όρους

null cumulants ανά παρατήρηση. Τότε έχουμε ότι:

$$E(\gamma^r) = n^{-1} k^{r,s} \left\{ -k^{t,u} \left(\frac{k_{s,t,u} + k_{s,tu}}{2} \right) + \alpha_s \right\} + \mathcal{O}(n^{-3/2}) \quad (3.74)$$

όπου $a_s = E(A_s(\beta))$

Παρατηρούμε ότι στην τελευταία σχέση εμφανίζεται ο $\mathcal{O}(n^{-1})$ όρος της μεροληψίας που δόθηκε από τους Cox & Snell, σχέση (3.51). Επομένως, για να έχει ο τροποποιημένος εκτιμητής μέγιστης πιθανοφάνειας μεροληψία με ασυμπτωτική τάξη μικρότερη από $\mathcal{O}(n^{-1})$ θα πρέπει να επιλέξουμε το α_s ώστε:

$$\begin{aligned} k^{r,s} a_s &= k^{r,s} k^{t,u} \left(\frac{k_{s,t,u} + k_{s,tu}}{2} \right) + \mathcal{O}(n^{-1/2}) \\ \Leftrightarrow k^{r,s} a_s &= -b_1^s + \mathcal{O}(n^{-1/2}) \\ \Leftrightarrow a_r &= -k_{r,s} b_1^s + \mathcal{O}(n^{-1/2}) \end{aligned}$$

Η τελευταία σχέση γράφεται σε μορφή πινάκων ως:

$$E(A) = -I(\beta) \frac{b_1(\beta)}{n} + \mathcal{O}(n^{-1/2})$$

Έτσι ο Firth (1993) κατέληξε στο συμπέρασμα πως οι δύο προφανείς μορφές του A_r που μπορούν να χρησιμοποιηθούν για την αφαίρεση του πρώτου όρου από το ασυμπτωτικό ανάπτυγμα της μεροληψίας του εκτιμητή μέγιστης πιθανοφάνειας είναι βασίζονται η μία στην αναμενόμενη πληροφορία και η άλλη στην παρατηρούμενη πληροφορία. Πιο συγκεκριμένα οι δύο μορφές του A_r είναι:

1. $A_r = A_r^{(E)} = -k_{r,s} b_1^s$, για την αναμενόμενη πληροφορία (expected information).
2. $A_r = A_r^{(O)} = n^{-1} U_{rs}(\beta) b_1^s$, για την παρατηρούμενη πληροφορία (observed information).

Στην περίπτωση της κανονικής συνάρτησης σύνδεσης (canonical link function) η παρατηρούμενη πληροφορία συμπίπτει με την αναμενόμενη. Έτσι μπορούμε να επεκτείνουμε και άλλο το προηγούμενο αποτέλεσμα ως εξής:

$$\alpha_r = -k_{r,s} b_1^s = -n k_{r,s} \frac{b_1^s}{n} = -k_{r,s} k^{s,t} k^{u,v} \left(\frac{k_{t,uv} + k_{t,u,v}}{2} \right)$$

$$\stackrel{k_{t,uv}=0}{=} -k_{r,s} k^{s,t} k^{u,v} \frac{k_{t,u,v}}{2} = -k^{u,v} \frac{k_{ruv}}{2}$$

Η προτελευταία ισότητα οφείλεται σε εφαρμογή της σύμβασης άθροισης του Einstein ενώ η τελευταία σε εφαρμογή της 3ης σχέσης του Bartlett. Έτσι τελικά, $\alpha_r = -k^{u,v} \frac{k_{ruv}}{2}$. Η σχέση αυτή γράφεται σε όρους πινάκων ως:

$$a_r = \frac{1}{2} \text{tr} \left\{ I^{-1} \left(\frac{\partial I}{\partial \beta_r} \right) \right\} = \frac{\partial}{\partial \beta_r} \left\{ \frac{1}{2} \log |I(\beta)| \right\} \quad (3.75)$$

Επομένως τελικά, η κατάλληλη τροποποίηση στη συνάρτηση score (score function) ώστε να προκύψει εκτιμητής μέγιστης πιθανοφάνειας απαλλαγμένος από τον $\mathcal{O}(n^{-1})$ όρο της μεροληψίας είναι:

$$U_r^*(\beta) = U_r(\beta) + \frac{\partial}{\partial \beta_r} \left\{ \frac{1}{2} \log |I(\beta)| \right\} \quad (3.76)$$

δηλαδή τροποποίηση στο λογάριθμο της συνάρτησης πιθανοφάνειας ως εξής:

$$l^*(\beta) = l(\beta) + \frac{1}{2} \log |I(\beta)| \quad (3.77)$$

ή απλά στη συνάρτηση πιθανοφάνειας ως:

$$L^*(\beta) = L(\beta) |I(\beta)|^{1/2} \quad (3.78)$$

Η συνάρτηση ποινικοποίηση $|I(\beta)|^{1/2}$ είναι γνωστή ως εκ των προτέρων Jeffrey (Jeffrey's invariant prior, Jeffrey 1946). Έτσι, ο Firth απέδειξε πως η ποινικοποίηση της συνάρτησης πιθανοφάνειας με την εκ των προτέρων Jeffrey οδηγεί σε έναν βελτιωμένο εκτιμητή ο οποίος είναι απαλλαγμένος από τον πρώτο όρο της

μεροληψίας, $\mathcal{O}(n^{-1})$.

Συνέπεια του β^*

Ορισμός 3.5.1. Η ακολουθία τυχαίων μεταβλητών $\{\beta_n\}$, λέμε ότι συγχλίνει κατά πιθανότητα προς την τυχαία μεταβλητή β αν:

$$\lim P(\|\beta_n - \beta\| \leq \epsilon) = 1$$

Ορισμός 3.5.2. Ένας εκτιμητής β^* μιας παραμέτρου β λέγεται συνεπής αν συγχλίνει κατά πιθανότητα στην παράμετρο β ή ισοδύναμα εάν $\beta^* - \beta = o_p(1)$

Έστω λοιπόν, β^* ο εκτιμητής μέγιστης πιθανοφάνειας που προκύπτει ως λύση της τροποποιημένης συνάρτησης score :

$$U_r^*(\beta) = U_r(\beta) + A_r(\beta).$$

Γνωρίζουμε πως ο εκτιμητής μέγιστης πιθανοφάνειας $\hat{\beta}$ που προκύπτει ως λύση της $U_r(\beta)$ είναι συνεπής εκτιμητής του β και επιπλέον η ποσότητα $\sqrt{n}(\hat{\beta} - \beta)$ ασυμπτωτικά ακολουθεί κανονική κατανομή με μέση τιμή 0 και πίνακα διακυμάνσεων συνδιακυμάνσεων τον αντίστροφο του πίνακα πληροφορίας του Fisher ανά παρατήρηση στο β . Έτσι,

$$(\hat{\beta} - \beta)^r = \mathcal{O}_p(n^{-1/2})$$

Έστω $\epsilon^r = (\beta^* - \hat{\beta})^r$. Η ποσότητα $A_r(\beta)$ έχει υποτεθεί ως μια $\mathcal{O}_p(1)$ ποσότητα και επομένως μπορούμε να γράψουμε ότι: $U_r^*(\beta) = U_r(\beta) + \mathcal{O}_p(1)$. Από το ανάπτυγμα Taylor της τροποποιημένης συνάρτησης score στο β^* γύρω από το $\hat{\beta}$ έχουμε ότι: $0 = U_r^*(\beta^*) = U_r^*(\hat{\beta}) + \epsilon^s U_{rs}^*(\hat{\beta})$. Αφού, το $\hat{\beta}$ αποτελεί λύση της

μη-τροποποιημένης συνάρτησης score μπορούμε να πάρουμε ότι:

$$\begin{aligned} U_r^*(\hat{\beta}) &= U_r(\hat{\beta}) + A_r(\hat{\beta}) \\ \Rightarrow U_r^*(\hat{\beta}) &= A_r(\hat{\beta}) \end{aligned}$$

Έτσι

$$\begin{aligned} 0 &= U_r^*(\hat{\beta}) + \epsilon^s U_{rs}^*(\hat{\beta}) \\ \Rightarrow \epsilon^r &= -A_s(\hat{\beta}) U^{*rs}(\hat{\beta}) \end{aligned}$$

Επιπλέον, γνωρίζουμε ότι το U_{rs}^* έχει ασυμπτωτική τάξη $\mathcal{O}_p(n)$ $\Rightarrow$ το U^{*rs} έχει ασυμπτωτική τάξη $\mathcal{O}_p(n^{-1})$ και αφού $A_r = \mathcal{O}_p(1)$ έχουμε ότι

$$\epsilon^r = \mathcal{O}_p(1) \cdot \mathcal{O}_p(n^{-1}) = \mathcal{O}_p(n^{-1})$$

Έτσι έχουμε ότι:

$$\begin{aligned} (\hat{\beta} - \beta)^r &= \hat{\beta}^r - \beta^r = \mathcal{O}_p(n^{-1/2}) \\ (\beta^* - \hat{\beta})^r &= \mathcal{O}_p(n^{-1}) \end{aligned}$$

Από την πρόσθεση αυτών των σχέσεων προκύπτει ότι:

$$(\beta^* - \beta)^r = \mathcal{O}_p \left(n^{\max\{-1, -\frac{1}{2}\}} \right) = \mathcal{O}_p(n^{-1/2})$$

Έτσι έχουμε ότι $(\beta^* - \beta)^r = \mathcal{O}_p(n^{-1/2})$, τότε από το Θεώρημα 3.2.5 που αποδείχθηκε στην αρχή του κεφαλαίου και συνδέει τα σύμβολα $\mathcal{O}_p$ και o_p για $\alpha = \frac{1}{2}$ και $t = \frac{1}{2}$ έχουμε ότι:

$$(\beta^* - \beta)^r = \text{o}_p(1)$$

και επομένως ο β^* είναι συνεπής εκτιμητής του β .

Κανονικότητα του β^*

Θεώρημα 3.5.3 (Το λήμμα του Slutsky). Έστω X_n, Y_n δύο ακολουθίες τυχαίων στοιχείων (*random elements*) αν η X_n συγκλίνει κατά κατανομή σε ένα τυχαίο στοιχείο X ($X_n \xrightarrow{d} X$) και η Y συγκλίνει κατά πιθανότητα σε μία σταθερά c ($Y_n \xrightarrow{p} c$) τότε:

1. $X_n + Y_n \xrightarrow{d} X + c$
2. $X_n Y_n \xrightarrow{d} cX$
3. $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$, υπό την προϋπόθεση ότι ορίζονται οι ποσότητες στους παρονομαστές.

Για την ασυμπτωτική κανονικότητα του β^* , αφού

$$\sqrt{n} \gamma^r = \sqrt{n} v^{r,s} U_s + \mathcal{O}_p(n^{-1/2})$$

και πάλι χρησιμοποιώντας το θεώρημα 3.2.5 που συνδέει $\mathcal{O}_p$ και $\mathcal{o}_p$ για $\alpha = \frac{1}{2}$ και για $t = \frac{1}{2}$ έχουμε ότι: $\sqrt{n} \gamma^r = \sqrt{n} v^{r,s} U_s + \mathcal{o}_p(1)$. Έτσι, από εφαρμογή του κεντρικού οριακού θεωρήματος για το τυποποιημένο διάνυσμα score (score vector) και από το λήμμα του Slutsky έχουμε ότι η ποσότητα:

$$\sqrt{n}(\beta^{*r} - \beta^r)$$

περιγράφεται ασυμπτωτικά από την κανονική κατανομή με μέση τιμή 0 και πίνακα διακυμάνσεων-συνδιακυμάνσεων τον αντρίστροφο του πίνακα πληροφορίας του Fisher ανά παρατήρηση $k^{r,s} = n v^{r,s}$.

3.6 Το πρόβλημα των διαχωρισμένων δεδομένων

Σε αυτήν την ενότητα θα γίνει μια σύντομη αναφορά σε μια σπουδαία ιδιότητα της μεθόδου του Firth, αφού προκύπτει πως η ποινικοποίηση της συνάρτησης πιθανοφάνειας με την εκ των προτέρων κατανομή Jeffreys οδηγεί πάντοτε σε πεπερασμένους εκτιμητές. Το πρόβλημα αυτό εμφανίζεται συχνά στην περίπτωση της λογιστικής παλινδρόμησης, αλλά και στο μοντέλο του Cox ιδιαίτερα όταν έχουμε στη διάθεση μας ένα μικρό σύνολο δεδομένων. Το πρόβλημα των μη-πεπερασμένων εκτιμητών μέγιστης πιθανοφάνειας είναι γνωστό ως πρόβλημα του διαχωρισμού (separation problem) ενώ συχνά συναντάται και ως το πρόβλημα της μονότονης πιθανοφάνειας (monotone likelihood).

Το πρόβλημα τέθηκε από τους Albert and Anderson (1984). Στην περίπτωση της λογιστικής παλινδρόμησης λέμε ότι υπάρχει διαχωρισμός στα δεδομένα όταν υπάρχει ένα υπερεπίπεδο T ώστε τα δειγματικά σημεία για τα οποία $y_i = 0$ να βρίσκονται στη μια πλευρά του T , ενώ για $y_i = 1$ να βρίσκονται στην άλλη πλευρά του υπερεπιπέδου. Γενικότερα, το πρόβλημα αυτό δημιουργείται όταν μεταξύ των ανεξάρτητων μεταβλητών υπάρχει μια μεταβλητή ή ένας γραμμικός συνδυασμός μεταβλητών ώστε να προβλέπουν με τέλεια τις τιμές της μεταβλητής απόκρισης. Οι Albert και Anderson στην εργασία τους σημειώνουν πως τα n δειγματικά σημεία ενός συνόλου δεδομένων μπορούν να εμφανίσουν ένα από τα παρακάτω φαινόμενα:

- Πλήρως διαχωρισμένα δεδομένα (completely separated data)
- Σχεδόν πλήρως διαχωρισμένα δεδομένα (quasi-complete separation data)
- Αλληλοεπικαλυπτόμενα δεδομένα (overlaped data).

Στη συνέχεια δίνονται οι ορισμοί και τα θεωρήματα που δόθηκαν από τους Albert and Anderson (1984) όπως αυτά αναδιατυπώνονται στην διδακτορική διατριβή του Konis (Konis, PhD thesis, 2007).

Ορισμός 3.6.1. (Albert-Anderson, 1984 - Konis PhD thesis, 2007)

Λέμε ότι υπάρχει πλήρης διαχωρισμός (complete separation) ανάμεσα στα n δειγματικά σημεία, αν υπάρχει διάνυσμα $\beta \in \mathbb{R}^p$ ώστε

$$\begin{cases} \beta^T x_i < 0, & \text{όταν } y_i = 0 \\ \beta^T x_i > 0, & \text{όταν } y_i = 1 \end{cases}$$

για $i = 1, 2, \dots, n$

Το επόμενο θεώρημα είναι το πρώτο από τα 2 βασικά αποτελέσματα των Albert & Anderson (1984). Για τις αποδείξεις αυτών των θεωρημάτων παραπέμπουμε στην συγκεκριμένη εργασία αλλά και στην εργασία των Lesaffe & Albert (1989). Επιπλέον, διορθώσεις πάνω στις αρχικές αποδείξεις των θεωρημάτων των Albert & Anderson μπορούν να βρεθούν στην εργασία των Santner & Duffy (1986).

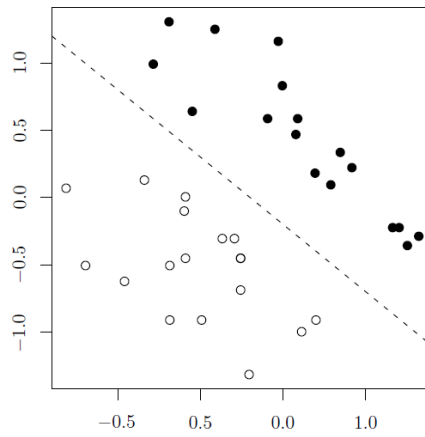
Θεώρημα 3.6.2. (Albert-Anderson, 1984 - Konis PhD thesis, 2007) Αν υπάρχει πλήρης διαχωρισμός ανάμεσα στα δειγματικά σημεία, τότε ο εκτιμητής μέγιστης πιθανοφάνειας $\hat{\beta}$ δεν υπάρχει και μάλιστα:

$$\sup_{\beta \in \mathbb{R}^p} L(\beta; y, x) = 1$$

Εδώ σημειώνουμε πως σε αυτήν την παράγραφο η αναφορά στην λέξη "ύπαρξη" του εκτιμητή μέγιστης πιθανοφάνειας έχει την έννοια της ύπαρξης που δόθηκε από τους Albert & Anderson στην εργασία τους, όπου η ύπαρξη του εκτιμητή ταυτίζεται με την έννοια του πεπερασμένου εκτιμητή. Εάν υπάρχει πλήρης διαχωρισμός στα δεδομένα τότε υπάρχει υπερεπίπεδο διάστασης $p - 1$ που χωρίζει

τον $\mathbb{R}^p$ σε δύο κομμάτια τέτοια ώστε οι ποσότητες x_i να βρίσκονται γνήσια στο ένα κομμάτι όταν $i \in E_1$ και γνήσια στο άλλο κομμάτι όταν $i \in E_2$. Στο επόμενο γράφημα φαίνεται μια πιθανή μορφή που μπορεί να έχουν τα πλήρως διαχωρισμένα δεδομένα, όταν $p = 3$.

Σχήμα 3.1: Η μορφή των πλήρως διαχωρισμένων δεδομένων, για $p = 3$.



Αν δεν υπάρχει διάνυσμα να διαχωρίζει τέλεια τα n δειγματικά σημεία, τότε μπορεί να υφίσταται ένα άλλο είδος διαχωρισμού ανάμεσα στα σημεία αυτά.

Ορισμός 3.6.3. (Albert-Anderson, 1984 - Konis PhD thesis, 2007) Από ένα διάνυσμα $\beta \neq 0, \beta \in \mathbb{R}^p$ λέμε ότι προκύπτει σχεδόν πλήρης διαχωρισμός (quasi-complete separation) εάν

$$\begin{cases} \beta^T x_i \leq 0, & \text{όταν } y_i = 0 \\ \beta^T x_i \geq 0, & \text{όταν } y_i = 1 \end{cases}$$

με την ισότητα να ισχύει τουλάχιστον για ένα $i, i = 1, 2, \dots, n$

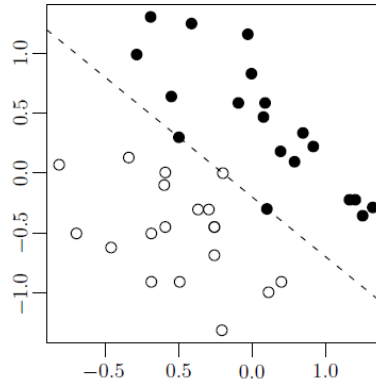
Θεώρημα 3.6.4. (Albert-Anderson, 1984 - Konis PhD thesis, 2007) Αν υπάρχει σχεδόν πλήρης διαχωρισμός (quasi-complete separation) ανάμεσα στα

δειγματικά σημεία τότε ο εκτιμητής μέγιστης πιθανοφάνειας $\hat{\beta}$ δεν υπάρχει και μάλιστα

$$\sup_{\beta \in \mathbb{R}^p} L(\beta; y, x) < 1$$

Στο επόμενο γράφημα δίνεται η μορφή που μπορούν να έχουν τα σχεδόν-πλήρη διαχωρισμένα δεδομένα.

Σχήμα 3.2: Η μορφή των σχεδόν-πλήρως διαχωρισμένων δεδομένων για $p = 3$.



Τέλος αν στα δεδομένα δεν παρουσιάζεται κανένα από τα δύο προηγούμενα φαινόμενα τότε αυτά αλληλοεπικαλύπτονται.

Ορισμός 3.6.5. (Albert-Anderson, 1984 - Konis PhD thesis, 2007) Λέμε ότι τα δεδομένα παρουσιάζουν το φαινόμενο της αλληλοεπικάλυψης (overlap) όταν για κάθε διάνυσμα β είτε $\beta^T x_i \geq 0$, για τουλάχιστον ένα $i \in E_1$ είτε $\beta^T x_i \geq 0$, για τουλάχιστον ένα $i \in E_2$.

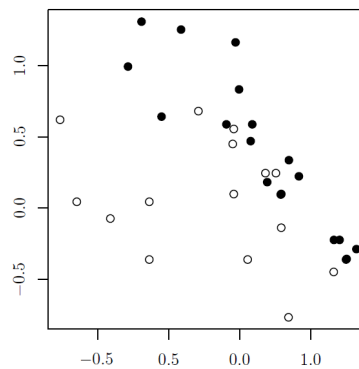
Η περίπτωση των αλληλοεπικαλυπτόμενων δεδομένων είναι μια επιθυμητή ιδιότητα κάτι που φαίνεται και από το επόμενο θεώρημα που αναφέρεται στην

εργασία των Albert & Anderson (1984) όμως αποδείχθηκε για πρώτη φορά από τον Silvapulle (1981).

Θεώρημα 3.6.6. (Silvapulle, 1981) *Αν τα δειγματικά σημεία αλληλοεπικαλύπτονται (overlapping data) τότε ο εκτιμητής μέγιστης πιθανοφάνειας $\hat{\beta}$ υπάρχει και είναι μοναδικός.*

Στο επόμενο γράφημα δίνεται μια πιθανή μορφή που μπορεί να έχουν τα αλληλοεπικαλυπτόμενα δεδομένα (overlapping data).

Σχήμα 3.3: Η μορφή των αλληλοεπικαλυπτόμενων δεδομένων για $p = 3$.



Το πρόβλημα του διαχωρισμού αποτελεί ένα πολύ σημαντικό πρόβλημα το οποίο μπορεί να προκύψει κατά την ανάλυση δεδομένων ιδιαίτερα σε περιπτώσεις μικρών συνόλων δεδομένων, αφού τότε βλάπτεται η αξιοπιστία του εκτιμητή μέγιστης πιθανοφάνειας. Λύση στο συγκεκριμένο πρόβλημα δίνεται από την τροποποίηση της συνάρτηση πιθανοφάνειας με τη μέθοδο που προτάθηκε από τον Firth (1993) και αναλύθηκε σε προηγούμενη ενότητα, έτσι ένας λόγος για τον οποίο υπερέχει η συγκεκριμένη μέθοδος σε σχέση με αυτή των McCullagh & Cordeiro (1991) είναι πως η μέθοδος του Firth δίνει πάντοτε πεπερασμένους εκτιμητές. Η μέθοδος των

McCullagh & Cordeiro δε θα μπορούσε να εφαρμοστεί σε τέτοιες περιπτώσεις διότι εξαρτάται άμεσα από την αρχική τιμή του εκτιμητή μέγιστης πιθανοφάνειας του οποίου τη μεροληψία επιχειρεί να διορθώσει. Σχετικά, τώρα με τη λύση του προβλήματος αυτού οι Heinze & Schemper σε δύο εργασίες τους (2001,2002) προσάρμοσαν σε σύνολα δεδομένων ένα μοντέλο λογιστικής παλινδρόμησης και ένα μοντέλο ανάλογων κινδύνων του Cox. Σε αυτά τα σύνολα δεδομένων η χρήση των κλασικών μεθόδων δε μπορούσε να εφαρμοστεί αφού είχε διαπιστωθεί η ύπαρξη διαχωρισμού στα δεδομένα. Από την τροποποίηση των συναρτήσεων πιθανοφάνειας με τη μέθοδο του Firth προέκυψαν πεπερασμένοι εκτιμητές. Τα δύο αυτά σύνολα δεδομένων είναι αυτά που θα χρησιμοποιηθούν στη συνέχεια για την εφαρμογή της μεθόδου του Firth.

Σημειώνουμε πως οι Heinze & Schemper στις εργασίες τους δεν έδωσαν θεωρήματα και αποδείξεις για τη συγκεκριμένη ιδιότητα της μεθόδου του Firth. Τα συγκεκριμένα θεωρήματα δόθηκαν σε πρόσφατη εργασία των Kosmidis and Firth, 2018. Στη συγκεκριμένη εργασία δίνεται αναλυτικά μια σειρά θεωρημάτων και αποδείξεων για την περίπτωση της λογιστικής παλινδρόμησης και έτσι "κλείνει" το θεωρητικό κενό που άφησε η εργασία των Heinze & Schemper για την λογιστική παλινδρόμηση. Στη συνέχεια παραθέτουμε τα θεωρήματα και κάποιες από τις αποδείξεις που υπάρχουν στη συγκεκριμένη πρόσφατη εργασία. Τέλος, σημειώνουμε ότι στη βιβλιογραφία δεν έχει βρεθεί κάποια αναφορά για την κατασκευή και την απόδειξη παρόμοιων θεωρημάτων για την περίπτωση του μοντέλου των αναλογικών κινδύνων του Cox.

Θεωρούμε την εκτίμηση του β που προκύπτει από τη μεγιστοποίηση του ποινικοποιημένου λογαρίθμου της συνάρτησης πιθανοφάνειας που έχει τη μορφή:

$$l^{(\alpha)}(\beta) = l(\beta) + \alpha \log \det I(\beta) \quad (3.79)$$

όπου α μια θετική σταθερά και $I(\beta)$ ο πίνακας πληροφορίας του Fisher στο β .

Θεώρημα 3.6.7. (*Kosmidis and Firth, 2018*) Αν κάποια συνιστώσα της γραμμικής προβλέπουσας $\eta = X\beta$ απειρίζεται, τότε η ποινικοποιημένη λογαριθμική πιθανοφάνεια $l^{(\alpha)}(\beta)$ θα απειρίζεται αρνητικά.

Θεώρημα 3.6.8. (*Kosmidis and Firth, 2018*) Το διάνυσμα $\hat{\beta}^{(\alpha)}$ που μεγιστοποιεί την $l^{(\alpha)}(\beta)$ έχει όλες του τις συνιστώσες πεπερασμένες, για κάθε θετικό αριθμό α .

Απόδειξη. Αν κάποια συνιστώσα του β απειρίζεται τότε τουλάχιστον μια συνιστώσα της γραμμικής προβλέπουσας $\eta(\beta)$ θα απειρίζεται. Τότε από το προηγούμενο θεώρημα η $l^{(\alpha)}(\beta)$ θα απειρίζεται αρνητικά και έτσι η $l^{(\alpha)}(\beta)$ δεν μεγιστοποιείται. Καθώς από την υπόθεση το $\beta^{(\alpha)}$ μεγιστοποιεί την $l^{(\alpha)}(\beta)$, τότε αναγκαστικά θα έχει όλες του τις συνιστώσες πεπερασμένες. $\square$

Έτσι με βάση τα προηγούμενα δύο θεωρήματα εξάγουμε το συμπέρασμα πως από την οικογένεια συναρτήσεων που ορίζονται από την (3.79) προκύπτουν πάντοτε πεπερασμένοι εκτιμητές μέγιστης πιθανοφάνειας. Αν επιλέξουμε $\alpha = 1/2$ τότε αναγόμεστε στην περίπτωση της ποινικοποίησης που προτάθηκε από τον Firth (1993). Εξίσου σημαντικά είναι και τα επόμενα δύο θεωρήματα, από τα οποία συμπεραίνουμε πως ύστερα από την ποινικοποίηση της συνάρτησης πιθανοφάνειας με την εκ των προτέρων κατανομή Jeffreys προκύπτουν εκτιμητές μέγιστης πιθανοφάνειας οι οποίοι έχουν υποστεί μια συρρίκνωση (shrinkage) ως προς την αρχή των αξόνων σε σχέση με τους αντίστοιχους που προκύπτουν από την κλασική πιθανοφάνεια. Για να αποδείξουμε τα συγκεκριμένα θεωρήματα θα χρειαστούμε κάποια αλγεβρικά λήμματα τα οποία θα διατυπώσουμε όμως θα τα χρησιμοποιήσουμε χωρίς την αντιστοιχη απόδειξη, για την οποία παραπέμπουμε στο βιβλίο

των Magnus & Neudecker (1999, Κεφ. 11).

Λήμμα 3.6.9. Έστω A ένας θετικά ορισμένος (*positive definite*) $n \times n$ πίνακας με ιδιοτιμές (*eigenvalues*) $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$. Τότε

$$\min_{Q^T Q = I_p} \det(Q^T A Q) = \prod_{t=1}^p \lambda_t$$

$$\max_{Q^T Q = I_p} \det(Q^T A Q) = \prod_{t=1}^p \lambda_{n-p+t}$$

Λήμμα 3.6.10. Αν A και B είναι δύο διαγώνιοι (*diagonal*) $n \times n$ πίνακες με μη-αρνητικά διαγώνια στοιχεία a_r και b_r αντίστοιχα, ώστε $a_r \geq b_r$ για κάθε $r = 1, 2, \dots, n$ τότε αν ο X είναι ένας $n \times p$ πίνακας, θα ισχύει ότι

$$\det(X^T A X) > \det(X^T B X)$$

Λήμμα 3.6.11. Αν A και B δύο θετικά ημι-ορισμένοι (*positive semidefinite*) πίνακες που έχουν την ίδια τάξη (*order*) και θ ένας αριθμός ώστε $0 < \theta < 1$ τότε

$$\det A^\theta \det B^{(1-\theta)} \leq \det\{\theta A + (1-\theta)B\}$$

Θεώρημα 3.6.12. (*Kosmidis and Firth, 2018*) Έστω β η p -διάστατη παράμετρος της παλινδρόμησης του μοντέλου της λογιστικής παλινδρόμησης. Αν επιπλέον $I(\beta)$ είναι ο πίνακας πληροφορίας του Fisher στο β και X ο πίνακας της παλινδρόμησης που υποτίθεται ως πίνακας πλήρους βαθμίδας (*full rank*), τότε η συνάρτηση $\det I(\beta)$ μεγιστοποιείται ολικά για $\beta = 0$.

Απόδειξη. Αφού ο πίνακας X είναι πίνακας πλήρους βαθμίδας, τότε μπορούμε να τον ορθογωνοποιήσουμε με τη μέθοδο Gram-Schmidt. Τότε $X = QR$ όπου

Q ένας $n \times p$ πίνακας τέτοιος ώστε $Q^T Q = I_p$ και ο R είναι ένας μη-ιδιάζων (non-singular) $p \times p$ πίνακας. Τότε μπορεί να δειχθεί ότι:

$$\det\{X^T W(\beta)X\} = \frac{\det\{Q^T W(\beta)Q\}}{(\det R)^2}$$

Παρατηρούμε ότι η ποσότητα στον παρονομαστή έχει θετικό πρόσημο ενώ επιπλέον είναι ανεξάρτητη του β . Άρα οι ποσότητες $\det\{X^T W(\beta)X\}$ και $\det\{Q^T W(\beta)Q\}$ έχουν τα ίδια στάσιμα σημεία (stationary points) για τις ίδιες τιμές του β . Έτσι αν εντοπίσουμε το β για το οποίο μεγιστοποιείται η $\det\{Q^T W(\beta)Q\}$ τότε αυτό θα είναι και το β για το οποίο μεγιστοποιείται η $\det\{X^T W(\beta)X\}$. Τότε τα διαγώνια στοιχεία του πίνακα W είναι οι ιδιοτιμές του, $w_r, r = 1, 2, \dots, n$ τις οποίες και διατάσσουμε ώστε: $w_{(1)} \leq w_{(2)} \leq \dots \leq w_{(n)}$. Τότε από το Λήμμα 3.7.9 θα ισχύει ότι:

$$\prod_{t=1}^p w_{(t)}(\beta) \leq \det\{Q^T W(\beta)Q\} \leq \prod_{t=1}^p w_{(n-p+t)}(\beta). \quad (3.80)$$

Τότε από το πρώτο κεφάλαιο και πιο συγκεκριμένα από τη σχέση (1.32) γνωρίζουμε πως για την περίπτωση της κανονικής συνάρτησης σύνδεσης (canonical link function) στο μοντέλο της λογιστικής παλινδρόμησης, όπου $\phi = 1$, το r -διαγώνιο στοιχείο του πίνακα W δίνεται από τη σχέση:

$$w_r(\pi_r) = \text{Var}(Y_r) = \pi_r(1 - \pi_r) \Rightarrow w(\eta) = \frac{e^\eta}{(1 + e^\eta)^2} \quad (3.81)$$

όπου η είναι η γραμμική προβλέπουσα. Εύκολα μπορούμε να μεγιστοποιήσουμε την συνάρτηση $w(\eta)$ και τότε βρίσκουμε ότι

$$0 < w(\eta) \leq \frac{1}{4}$$

με το μέγιστο να επιτυγχάνεται όταν: $\eta = 0$. Τότε για $\beta = 0$ στην σχέση (3.80) παίρνουμε ότι:

$$\frac{1}{4^p} \leq \det\{Q^T W(0)Q\} \leq \frac{1}{4^p}$$

Άρα για $\beta = 0$ η συνάρτηση $\det\{Q^T W(\beta)Q\}$ μεγιστοποιείται με τιμή μεγίστου $\frac{1}{4^p}$. Επομένως, για $\beta = 0$ μεγιστοποιείται και η συνάρτηση $\det\{Q^T W(\beta)Q\} = \det I(\beta)$ $\square$

Θεώρημα 3.6.13. (*Kosmidis, PhD thesis, 2007*) Ας είναι β η p -διάστατη παράμετρος της παλινδρόμησης του λογιστικού μοντέλου και π το n -διάστατο διάνυσμα των πιθανοτήτων επιτυχίας. Αν $I(\beta)$ είναι ο πίνακας πληροφορίας του Fisher υπολογισμένος στο β συμβολίζουμε με $\tilde{I}$ τον πίνακα πληροφορίας του Fisher υπολογισμένο στο π . Τότε, η συνάρτηση $f(\pi) = \det \tilde{I}(\pi)$ είναι $\log$ -κοίλη ($\log$ -concave) ως προς π .

Απόδειξη. Αν $W(\beta)$ είναι ο πίνακας W υπολογισμένος στο β συμβολίζουμε τον πίνακα W υπολογισμένο στο π ως $\tilde{W}$ ενώ γνωρίζουμε ότι το r -διαγώνιο στοιχείο του δίνεται από τη σχέση: $w_r(\pi_r) = \pi_r(1 - \pi_r)$. Έστω $\theta \in (0, 1)$, τότε για οποιοδήποτε ζεύγος των n διανυσμάτων με πιθανότητες π , ϕ και αφού η συνάρτηση $w(z) = z(1 - z)$ είναι κοίλη θα ισχύει ότι:

$$w_r(\theta\pi_r + (1 - \theta)\phi_r) \geq \theta w_r(\pi_r) + (1 - \theta)w_r(\phi_r). \quad (3.82)$$

Τότε με βάση το Λήμμα 3.7.10 η σχέση (3.82) γράφεται ως:

$$\det(X^T \tilde{W}(\theta\pi + (1 - \theta)\phi)X) \geq \det(\theta X^T \tilde{W}(\pi)X + (1 - \theta)X^T \tilde{W}(\phi)X)$$

και καθώς ο λογάριθμος είναι μια γνήσιως αύξουσα συνάρτηση η τελευταία σχέση γράφεται ως:

$$\begin{aligned} \log \left(\det(X^T \tilde{W}(\theta\pi + (1 - \theta)\phi)X) \right) &\geq \log \left(\det(\theta X^T \tilde{W}(\pi)X + (1 - \theta)X^T \tilde{W}(\phi)X) \right) \\ (\text{από το λήμμα 3.7.11}) &\geq \log \left(\det(X^T \tilde{W}(\pi)X)^\theta \det(X^T \tilde{W}(\phi)X)^{1-\theta} \right) \\ &= \theta \log \left(\det(X^T \tilde{W}(\pi)X) \right) + (1 - \theta) \log \left(\det(X^T \tilde{W}(\phi)X) \right) \end{aligned}$$

Έτσι, έχουμε ότι ,

$$\begin{aligned} \log \left(\det(X^T \tilde{W} (\theta\pi + (1-\theta)\phi)X) \right) &\geq \theta \log \left(X^T \tilde{W} (\pi)X \right) + \\ &\quad + (1-\theta) \log \left(\det(X^T \tilde{W} (\phi)X) \right) \\ \Rightarrow \log \left(\det \tilde{I} (\theta\pi + (1-\theta)\phi) \right) &\geq \theta \log(\det \tilde{I} (\pi)) + (1-\theta) \log \det \tilde{I} (\phi) \end{aligned}$$

□

Επομένως, με βάση τα Θεωρήματα 3.7.12 και 3.7.13 οι Kosmidis και Firth (2018) χρησιμοποιώντας μια μετρική που βασίζεται στην ορίζουσα του πίνακα πληροφορίας του Fisher καταλήγουν στο συμπέρασμα πως η ποινικοποίηση της συνάρτησης πιθανοφάνειας, εκτός από το γεγονός ότι παράγει πάντα πεπερασμένους εκτιμητές, δίνει μια συρρίκνωση (shrinkage) των αρχικών εκτιμητών, όταν αυτοί υπάρχουν, γύρω από την αρχή $\beta = 0$.

3.7 Εφαρμογή της μεθόδου του Firth

3.7.1 Εφαρμογή στη λογιστική παλινδρόμηση

Σε αυτήν την παράγραφο θα γίνει αναφορά στη μορφή που έχει η συνάρτηση score (score function), για το μοντέλο της λογιστικής παλινδρόμησης, όταν αυτή ποινικοποιηθεί με την εκ των προτέρων κατανομή Jeffreys. Στη συνέχεια θα εφαρμοστεί η μέθοδος του Firth σε ένα πραγματικό σύνολο δεδομένων. Πριν από αυτά όμως κρίνεται σκόπιμο να γίνει αναφορά στην εργασία του Firth (1992a) στην οποία δίνεται η κλειστή μορφή που έχει η συνάρτηση score (score function) στα γενικευμένα γραμμικά μοντέλα με κανονική συνάρτηση σύνδεσης, όταν εφαρμόζεται η μέθοδος του Firth.

Θεωρούμε ένα γενικευμένο γραμμικό μοντέλο για τις ανεξάρτητες παρατηρήσεις $y_1, y_2, \dots, y_n$. Αφού υποθέτουμε πως υπάρχει κανονική συνάρτηση σύνδεσης θα ισχύει ότι:

$$\theta_i = \sum_{j=1}^p \beta_j X_{ij}.$$

Σε αυτήν την περίπτωση, όπως έχουμε αποδείξει σε προηγούμενο κεφάλαιο, (βλέπε σχέση (1.30)» ο πίνακας πληροφορίας του Fisher έχει την ακόλουθη μορφή

$$I(\beta) = \frac{1}{\phi} X^T W X$$

όπου $W = \frac{1}{\phi} \text{diag}\{k_{2i}\}$ με k_{2i} να είναι ο δεύτερης τάξης cumulant στην μεταβλητής απόκρισης, δηλαδή η διακύμανση της $y_i, i = 1, 2, \dots, n$. Έτσι από την αντικατάσταση αυτής της μορφής του πίνακα πληροφορίας του Fisher στη σχέση του Firth και τη σχέση

$$\frac{\partial W}{\partial \beta_j} = W_j = \frac{1}{\phi} \text{diag}\left\{\frac{\partial k_{2i}}{\partial \beta_j}\right\} = \frac{1}{\phi^2} \text{diag}\{k_{3i} X_{ij}\}$$

ύστερα από αλγεβρικές πράξεις καταλήγουμε στο συμπέρασμα:

$$0 = U_r(\beta) + \frac{1}{2\phi} \sum_{i=1}^n \left(\frac{k_{3i}}{k_{2i}} \right), \quad j = 1, 2, \dots, p \quad (3.83)$$

όπου h_i είναι τα διαγώνια στοιχεία του πίνακα

$$H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2}$$

Από τη λύση του συστήματος εξισώσεων που ορίζεται από την (3.83) προκύπτει ο βελτιωμένος εκτιμητής μέγιστης πιθανοφάνειας. Έτσι, με βάση τα προηγούμενα για να εξάγουμε τη μορφή της συνάρτησης score για την περίπτωση της λογιστικής παλινδρόμησης αρκεί να υπολογίσουμε τον λόγο $\frac{k_{3i}}{k_{2i}}$. Οι ποσότητες αυτές είναι γνωστές, ενώ στην λογιστική παλινδρόμηση όπως και σε άλλα γενικευμένα γραμμικά μοντέλα ισχύει ότι $\phi = 1$. Τότε

$$\frac{k_{3i}}{k_{2i}} = \frac{2\pi_i^3 - 3\pi_i^2 + \pi_i}{\pi_i(1 - \pi_i)} = -2\pi_i + 1$$

ενώ από προηγούμενο κεφάλαιο γνωρίζουμε ότι $U_r(\beta) = \sum_{i=1}^n (y_i - \mu_i) X_{ij}$, $j = 0, 1, \dots, p-1$. Έτσι για την περίπτωση της λογιστικής παλινδρόμησης, η μορφή που έχει η τροποποιημένη συνάρτηση score (modified score function) είναι η εξής:

$$U_r^*(\beta) = \sum_{i=1}^n \left(y_i + \frac{h_i}{2} - h_i \pi_i - \pi_i \right) X_{ij}, \quad j = 1, 2, \dots, p \quad (3.84)$$

Στη συνέχεια θα παρουσιαστούν τα αποτελέσματα που προέκυψαν από την ανάλυση ενός πραγματικού συνόλου δεδομένων. Η πηγή του συνόλου αυτού είναι η εργασία των Heinze & Schemper (2002). Το δείγμα αποτελείται από 79 άτομα τα οποία διαγνώστηκαν με καρκίνο του ενδομητρίου (endometrial cancer). Ο σκοπός αυτής της έρευνας ήταν να κατασκευαστεί ένα μοντέλο πρόβλεψης που θα μπορεί να εξηγήει την ιστολογία του ενδομητρίου (histology of endometrium) με τη χρήση κάποιων επεξηγηματικών μεταβλητών. Στη συγκεκριμένη εργασία, σκοπός μας είναι να δείξουμε τον δραστικό τρόπο με τον οποίο επιδρά η μέθοδος του Firth και όχι να προτείνουμε κάποιο τελικό μοντέλο πρόβλεψης. Στο σύνολο δεδομένων υπάρχουν οι εξής μεταβλητές:

- **HG**: Είναι η εξαρτημένη μεταβλητή που περιγράφει την ιστολογία του ενδομητρίου και κωδικοποιείται με 0 (κατηγορίες I- II) και 1 (III-IV) , στις κατηγορίες αυτές ανήκουν 30 και 49 ασθενείς αντίστοιχα.
- **NV**: Κατηγορική μεταβλητή που υποδεικνύει την ύπαρξη ή όχι νεοαγγείωσης (neovascularization). Με 1 κωδικοποιείται η ύπαρξη νεοαγγείωσης, ενώ με 0 η απουσία. Στην πρώτη κατηγορία ανήκουν συνολικά 13 ασθενείς ενώ στη δεύτερη ανήκουν 66 ασθενείς.
- **PI**: Συνεχής μεταβλητή που καταγράφει το δείκτη παλμικότητας της μητριάς αρτηρίας (pulsality index of arteria uterina).

- **EH:** Συνεχής μεταβλητή που καταγράφει το ύψος του ενδομητρίου (height of endometrium).

Από την προσαρμογή του μοντέλου με όλες τις προηγούμενες ανεξάρτητες μεταβλητές προέκυψαν τα αποτελέσματα που φαίνονται στον πίνακα 3.1: Αυτό που είναι φανερό από την προηγούμενη προσαρμογή είναι πως η εκτίμηση που έχουμε για την μεταβλητή NV είναι μη ερμηνεύσιμη αφού το τυπικό σφάλμα του εκτιμητή είναι υπερβολικά μεγάλο. Επιπρόσθετα, το δείγμα της έρευνας είναι σχετικά μικρό αφού αποτελείται μόλις από 79 άτομα κάτι που μας δείχνει πως η μεροληψία έχει επηρεάσει σίγουρα τα αποτελέσματά μας. Επιπλέον, είναι απαραίτητο να διερευνηθεί η μεταβλητή NV ως προς το διαχωρισμό. Αυτό μπορεί να γίνει με την κατασκευή ενός 2×2 πίνακα της NV με την εξαρτημένη μεταβλητή HG. Αυτός ο πίνακας δίνεται παρακάτω:

	NV(0)	NV(1)
HG(0)	49	0
HG(1)	17	13

Από τον προηγούμενο πίνακα βλέπουμε πως πράγματι τα δεδομένα είναι σχεδόν πλήρως διαχωρισμένα, αφού δεν υπάρχει παρατήρηση στον συνδυασμό $HG=0$ και $NV=1$. Έτσι η μεταβλητή NV προβλέπει τέλεια την μεταβλητή HG. Τότε σύμφωνα με το Θεώρημα 3.7.4 συμπεραίνουμε πως ο εκτιμητής μέγιστης πιθανοφάνειας για την μεταβλητή NV δεν είναι πεπερασμένος. Προκειμένου να εξακριβώσουμε τον προηγούμενο ισχυρισμό μπορούμε να σχεδιάσουμε το γράφημα της συνάρτησης πιθανοφάνειας ξεχωριστά για την καθεμιά μεταβλητή θεωρώντας τις υπόλοιπες μεταβλητές ως "ενοχλητικές" παραμέτρους (nuisance parameters). Η μέθοδος αυτή είναι γνωστή ως μέθοδος της profile πιθανοφάνειας (profile

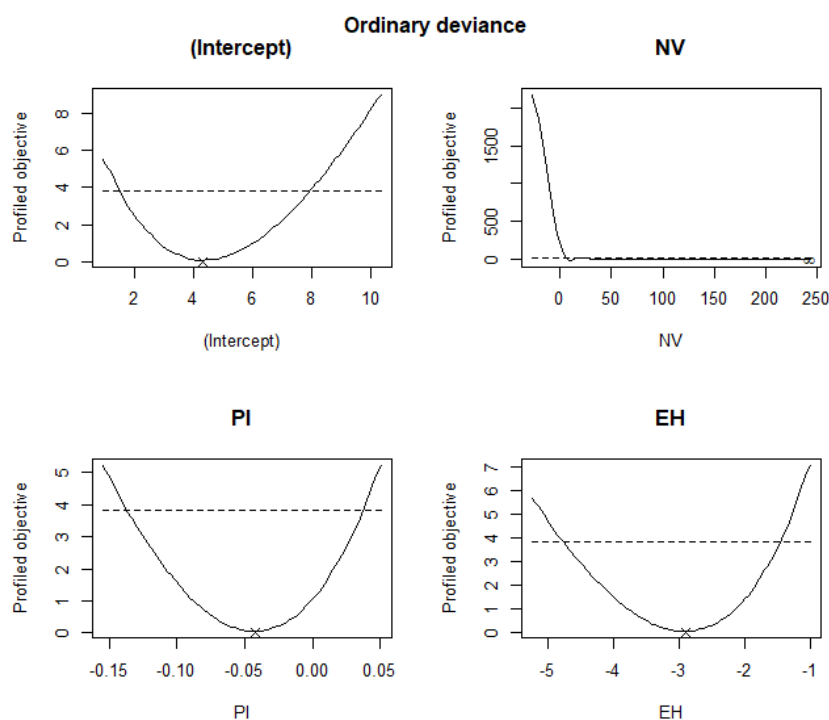
Πίνακας 3.1

<i>Μοντέλο λογιστικής παλινδρόμησης:</i>	
	HG
NV (1)	18.186
(se)	(1,715.751)
EH	−2.903***
(se)	(0.846)
PI	−0.042
(se)	(0.044)
Constant	4.305***
(se)	(1.637)
Observations	79
Log Likelihood	−27.697
Akaike Inf. Crit.	63.393

Note: *p<0.1; **p<0.05; ***p<0.01

likelihood). Εφαρμόζοντας αυτή τη μέθοδο τα γραφήματα των συναρτήσεων πιθανοφάνειας είναι τα παρακάτω:

Σχήμα 3.4: Διαγράμματα για την διευρέυση της ύπαρξης των εκτιμητών



Στον κάθετο άξονα των προηγούμενων γραφημάτων βρίσκονται οι τιμές της συνάρτησης απόκλισης (deviance function). Γνωρίζουμε στο σημείο ελαχιστοποίησης της βρίσκεται ο εκτιμητής μέγιστης πιθανοφάνειας. Παρατηρούμε ότι οι συναρτήσεις απόκλισης όλων των μεταβλητών είναι αυστηρά κυρτές (strictly convex) και συνεπώς ελαχιστοποιούνται μοναδικά με εξαίρεση την μεταβλητή NV η οποία απειρίζεται. Από την προσαρμογή του μοντέλου με την μέθοδο του Firth πήραμε τα εξής αποτελέσματα:

Προσαρμογή ποινικοποιημένου μοντέλου:

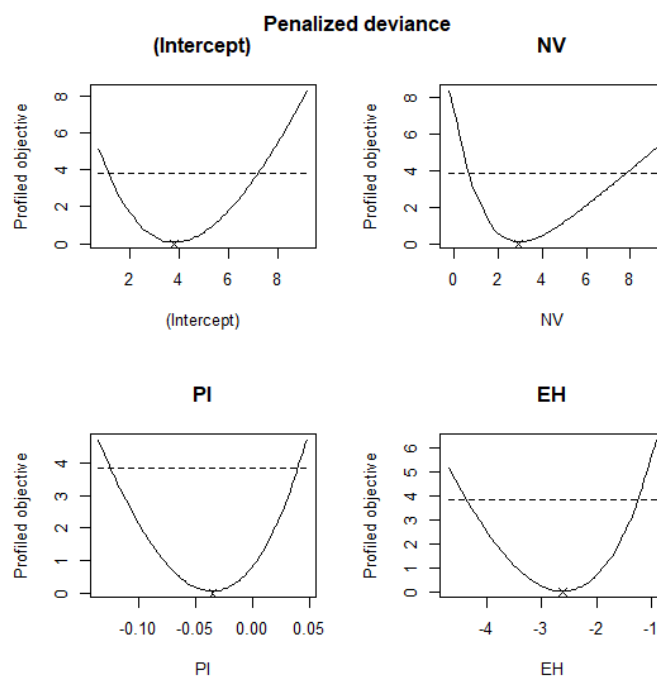
	HG
NV (1)	2.929*
(se)	(1.551)
EH	−2.604***
(se)	(0.776)
PI	−0.035
(se)	(0.040)
Constant	3.775**
(se)	(1.489)

Note:

*p<0.1; **p<0.05; *** p<0.01

Παρατηρούμε ότι η επίδραση της μεθόδου του Firth μειώνει δραστικά το τυπικό σφάλμα της εκτίμησης. Εδώ είναι επιτακτική η εφαρμογή της μεθόδου αφού από τον 2×2 πίνακα και το γράφημα που προηγήθηκε συμπεραίνουμε πως ο εκτιμητής μέγιστης πιθανοφάνειας για την μεταβλητή NV δεν υπάρχει. Τέλος, τα αντίστοιχα γραφήματα που προκύπτουν με βάση την profile πιθανοφάνεια, μετά την ποινικοποίηση, δίνονται παρακάτω. Από τα γραφήματα αυτά βλέπουμε πως πλέον οι εκτιμητές μέγιστης πιθανοφάνειας είναι πεπερασμένοι και μοναδικοί αφού οι συναρτήσεις απόκλισης είναι αυστηρά κυρτές (strictly convex). Έτσι η μέθοδος του Firth διόρθωσε το πρόβλημα των απειριζόμενων εκτιμητών ενώ με βάση τη θεωρητική της κατασκευή έχει μειώσει και την μεροληψία του εκτιμητή.

Σχήμα 3.5: Διαγράμματα των τροποποιημένων εκτιμητών μέγιστης πιθανοφάνειας



Έτσι το μοντέλο που καταλήγουμε είναι το:

$$\log \left(\frac{\hat{\pi}}{1 - \hat{\pi}} \right) = 3.775 + 2.929 \text{NV}(1) - 2.604 \text{EH} - 0.035 \text{PI}$$

3.7.2 Εφαρμογή στο μοντέλο του Cox

Σε αυτήν την παράγραφο θα δωθούν οι εκφράσεις που παίρνει η συνάρτηση της μερικής πιθανοφάνειας όταν σε αυτή εφαρμοστεί η ποινικοποίηση με τη μέθοδο του Firth. Για τα αποτελέσματα αυτής της παραγράφου θα στηριχθούμε στην εργασία των Heinze and Schemper (2001) όπου οι αντίστοιχες ποσότητες υπολογίζονται. Στο τέλος της ενότητας θα εφαρμόσουμε και πάλι τη μέθοδο του Firth για ένα πραγματικό σύνολο δεδομένων. Πριν όμως εφαρμόσουμε τη μέθοδο του Firth στο μοντέλο του Cox κρίνεται σκόπιμο να κάνουμε την εξής παρατήρηση: Η εργασία των Heinze and Schemper (2001) μπορεί να εφαρμόζει την μέθοδο αυτή για το μοντέλο του Cox όμως σε αυτή υπάρχει ένα μεγάλο θεωρητικό κενό. Η απόδειξη της μεθόδου του Firth, σύμφωνα με την οποία η ποινικοποίηση της συνάρτησης πιθανοφάνειας με την εκ των προτέρων κατανομή Jeffreys οδηγεί σε εκτιμητή μέγιστης πιθανοφάνειας απαλλαγμένο από τον $\mathcal{O}(n^{-1})$ όρο της μεροληψίας, έγινε υπό με την προϋπόθεση ότι ισχύουν οι σχέσεις του Bartlett:

$$k_r = 0$$

$$k_{rs} + k_{r,s} = 0$$

$$k_{rst} + k_{rs,t}[3] + k_{r,s,t} = 0$$

ενώ υπό την υπόθεση της κανονικής συνάρτησης σύνδεσης προέκυψε ότι $k_{r,st} = 0$. Ο συνδυασμός των προηγούμενων σχέσεων οδήγησε τον Firth στο τελικό του

συμπέρασμα. Οι αντίστοιχες σχέσεις δεν αποδείχθηκαν από τους Heinze and Schemper (2001) και εκεί έγκειται και η παρατήρηση σχετικά με το θεωρητικό κενό που άφησε η εργασία τους. Το συγκεκριμένο κενό έκλεισε με την πρόσφατη εργασία των Nagashima and Sato (2017) οι οποίοι απέδειξαν τις προηγούμενες σχέσεις για το μοντέλο του ανάλογων κινδύνων του Cox. Έστω από αυτήν την παρατήρηση η εφαρμογή της μεθόδου είναι άμεση. Έστω ένα τυχαίο δείγμα n ατόμων. Έστω ότι παρατηρούμε m διαφορετικούς μεταξύ τους χρόνους επιβίωσης, τους οποίους διατάσσουμε και τους συμβολίζουμε με $t_{(j)}, 1 \leq j \leq m$. Συμβολίζουμε με d_j τον αριθμό των ατόμων που εμφανίζουν το υπό μελέτη γεγονός την $t_{(j)}$ χρονική στιγμή, ενώ s_j είναι το διάνυσμα που περιέχει ως συνιστώσες τα αθροίσματα των ανεξάρτητων μεταβλητών των d_j ατόμων. Τέλος, με $R_{t_{(j)}}$ συμβολίζουμε το σύνολο ρίσκου (risk set) την $t_{(j)}$ χρονική στιγμή, δηλαδή το σύνολο όλων των ατόμων που έχουν επιβιώσει ακριβώς πριν την χρονική στιγμή $t_{(j)}$. Τότε όπως είδαμε αναλυτικά σε προηγούμενο κεφάλαιο η συνάρτηση πιθανοφάνειας έχει την ακόλουθη μορφή:

$$\log L(\beta) = \sum_{j=1}^m \left[\beta s_j - d_j \log \left\{ \sum_{h \in R_{t_{(j)}}} \exp(x_h \beta) \right\} \right].$$

Γνωρίζουμε ότι η τροποποιημένη με τη μέθοδο του Firth συνάρτηση score έχει τη μορφή:

$$U_r^*(\beta) = U_r(\beta) + a_r$$

όπου $a_r = \text{tr}\{I(\beta)^{-1} \left[\frac{\partial I(\beta)}{\partial \beta_r} \right]\}$. Επομένως, το μόνο που χρειαζόμαστε για την εφαρμογή αυτής της μεθόδου είναι να μπορούμε να έχουμε τις εκφράσεις των παραγώγων μέχρι και τρίτης τάξης του λογαρίθμου της συνάρτησης πιθανοφάνειας.

Οι παράγωγοι αυτοί είναι οι εξής:

$$\begin{aligned}\frac{\partial \log L(\beta)}{\partial \beta_r} &= \sum_{j=1}^m s_{jr} - \frac{d_j \sum_{h \in R_{t(j)}} x_{hr} \exp(x_h \beta)}{\sum_{h \in R_{t(j)}} \exp(x_h \beta)} \cdot \\ \frac{\partial^2 \log L(\beta)}{\partial \beta_r \partial \beta_s} &= - \sum_{j=1}^m d_j \left[\frac{\sum_{h \in R_{t(j)}} x_{hr} x_{hs} \exp(x_h \beta)}{\sum_{h \in R_{t(j)}} \exp(x_h \beta)} \right. \\ &\quad \left. - \frac{\sum_{h \in R_{t(j)}} x_{hr} \exp(x_h \beta) \sum_{h \in R_{t(j)}} x_{hs} \exp(x_h \beta)}{\left\{ \sum_{h \in R_{t(j)}} \exp(x_h \beta) \right\}^2} \right] \cdot \\ \frac{\partial^3 \log L(\beta)}{\partial \beta_r \partial \beta_s \partial \beta_t} &= - \sum_{j=1}^m d_j \left[\left(\frac{S_{j,rst}}{S_{j,0}} - \frac{S_{j,rs} S_{j,t}}{S_{j,0}^2} \right) - \frac{S_{j,s}}{S_{j,0}} \left(\frac{S_{j,rt}}{S_{j,0}} - \frac{S_{j,r} S_{j,t}}{S_{j,0}^2} \right) \right. \\ &\quad \left. - \frac{S_{j,r}}{S_{j,0}} \left(\frac{S_{j,st}}{S_{j,0}} - \frac{S_{j,s} S_{j,t}}{S_{j,0}^2} \right) \right], \quad \mu \varepsilon \quad 1 \leq r, s, t \leq k\end{aligned}$$

όπου

$$\begin{aligned}S_{j,0} &= \sum_{h \in R_{t(j)}} \exp(x_h \beta) \\ S_{j,r} &= \sum_{h \in R_{t(j)}} x_{hr} \exp(x_h \beta) \\ S_{j,s} &= \sum_{h \in R_{t(j)}} x_{hs} \exp(x_h \beta) \\ S_{j,t} &= \sum_{h \in R_{t(j)}} x_{ht} \exp(x_h \beta) \\ S_{j,rs} &= \sum_{h \in R_{t(j)}} x_{hr} x_{hs} \exp(x_h \beta)\end{aligned}$$

$$S_{j,rt} = \sum_{h \in R_{t(j)}} x_{hr} x_{ht} \exp(x_h \beta)$$

$$S_{j,st} = \sum_{h \in R_{t(j)}} x_{hs} x_{ht} \exp(x_h \beta)$$

$$S_{j,rst} = \sum_{h \in R_{t(j)}} x_{hr} x_{hs} x_{ht} \exp(x_h \beta)$$

Η συγκεκριμένη παράγραφος θα κλείσει με την παρουσίαση των αποτελεσμάτων που προέκυψαν από την ανάλυση ενός πραγματικού συνόλου δεδομένων το οποίο εμφανίζεται στην εργασία των Heinze and Schemper (2001). Στο συγκεκριμένο σύνολο δεδομένων καταγράφεται ο χρόνος επιβίωσης 100 ατόμων τα οποία έπασχαν από καρκίνο του μαστού (breast cancer), 74 εκ των οποίων είχαν λογοκριμένο χρόνο επιβίωσης (censored survival time). Πιο συγκεκριμένα, στο συγκεκριμένο σύνολο δεδομένων καταγράφονται οι παρακάτω μεταβλητές, οι 4 πρώτες είναι οι ανεξάρτητες μεταβλητές (explanatory variables) οι μεταβλητές αυτές έχουν διχοτομηθεί και κωδικοποιηθεί με 0 και 1, όπου ως 1 υποδεικνύεται μια αρνητική κλινικά κατάσταση.

- **T**: Στάδιο όγκου (tumor stage).
- **N**: Κατάσταση κόμβων (nodal status).
- **G**: Ιστολογική ταξινόμηση (histological grading).
- **CD**: Ανοσοαντιδραστικότητα καθεψίνης D (cathepsin D immunoreactivity).
- **CENS**: Δίτιμη μεταβλητή που περιγράφει τους χρόνους επιβίωσης. (0=λογοκριμένος χρόνος, 1=πλήρως καταγεγραμμένος).
- **TIME**: Συνεχής μεταβλητή που καταγράφει τον χρόνο επιβίωσης.

Τα αποτελέσματα που πήραμε προσαρμόζοντας το μοντέλο ανάλογων κινδύνων του Cox με όλες τις προηγούμενες ανεξάρτητες μεταβλητές δίνονται στη συνέχεια.

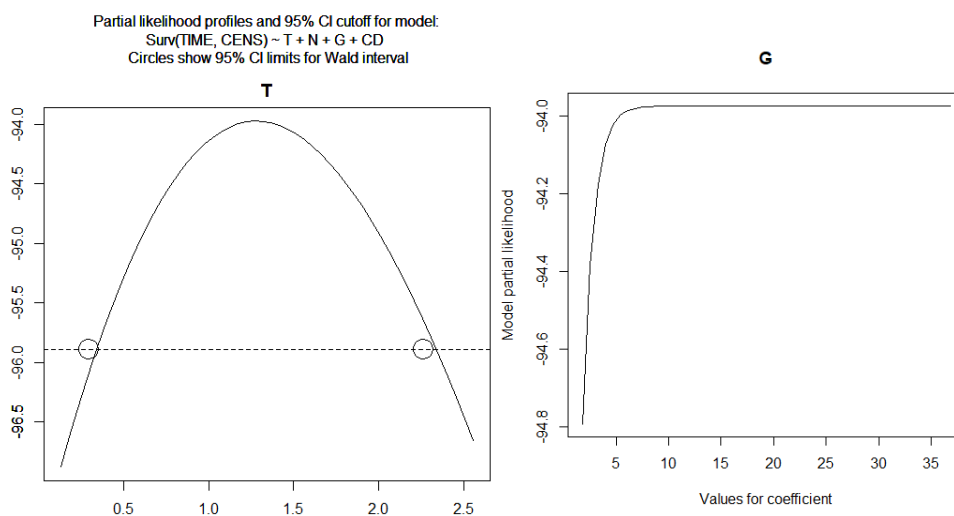
Πίνακας 3.2

<i>Μοντέλο ανάλογων κινδύνων του Cox:</i>	
	TIME
T (1)	1.279**
(se)	(0.502)
N (1)	0.946**
(se)	(0.425)
G (1)	18.421
(se)	(4,316.755)
CD (1)	0.400
(se)	(0.443)
Observations	100
Wald Test	18.890*** (df = 4)
LR Test	39.356*** (df = 4)

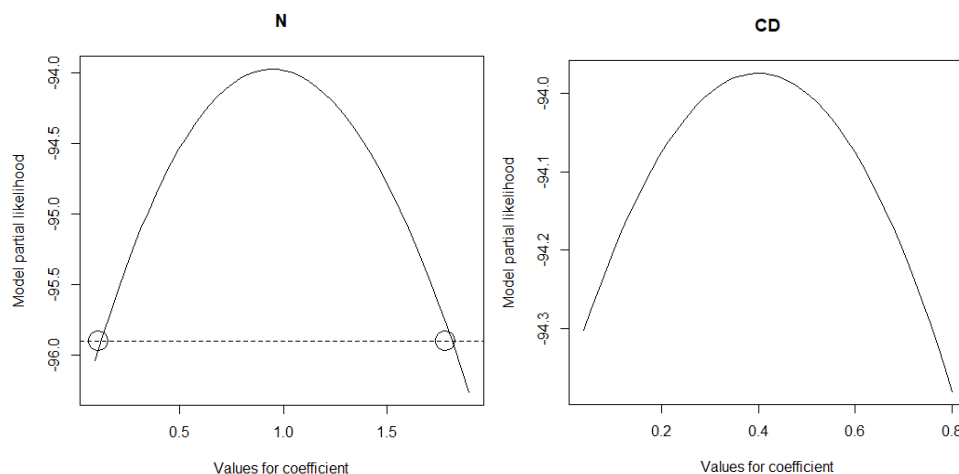
Note: *p<0.1; **p<0.05; ***p<0.01

Απο τον προηγούμενο πίνακα παρατηρούμε ότι για την μεταβλητή G έχουμε μια ασυνήθιστα μεγάλη τιμή του εκτιμητή, η οποία μεταφράζεται σε σχετικό κίνδυνο (relative risk) 100031931. Επιπλέον, το τυπικό σφάλμα αυτού του εκτιμητή είναι πολύ μεγάλο. Η αδυναμία μας να ερμηνεύσουμε το αποτέλεσμα αλλά και το μη αποδεκτό τυπικό σφάλμα του εκτιμητή μας υποδεικνύουν πως η εκτίμηση έχει επηρεαστεί από τη μεροληψία, ενώ κρίνεται απαραίτητο να κατασκευαστούν και σε αυτήν την περίπτωση τα διαγράμματα των profile πιθανοφανειών ώστε να διαπιστώσουμε εάν οι εκτιμητές των παραμέτρων είναι όντως πεπερασμένοι. Τα διαγράμματα αυτά δίνονται στη συνέχεια.

Σχήμα 3.6: Διαγράμματα profile πιθανοφανειών



Σχήμα 3.7: Διαγράμματα profile πιθανοφανειών



Στα προηγούμενα διαγράμματα παίρνουμε τις profile πιθανοφάνειες για κάθε μια από τις ανεξάρτητες μεταβλητές. Παρατηρούμε ότι για τις μεταβλητές N, CD, T οι συναρτήσεις είναι αυστηρά κοίλες (strictly concave) και έτσι αυτές μεγιστοποιούνται μοναδικά. Αντίθετα στην περίπτωση της μεταβλητής G βλέπουμε ότι εμφανίζεται το φαινόμενο της μονότονης πιθανοφάνειας (monotone likelihood) με αποτέλεσμα να μην υπάρχει ο εκτιμητής μέγιστης πιθανοφάνειας. Επομένως, και σε αυτό το μοντέλο το αποτέλεσμα που μας έδωσε το στατιστικό πακέτο της **R** οφείλεται σε σφάλμα. Με βάση όλα τα προηγούμενα συμπεραίνουμε πως και σε αυτήν την περίπτωση είναι αναγκαία η ποινικοποίηση της πιθανοφάνειας με τη μέθοδο του Firth αφού αυτή η μέθοδος διαθέτει όλο το θεωρητικό υπόβαθρο για να δώσει λύση στα προβλήματα που προέκυψαν. Τα αποτελέσματα που προέκυψαν ύστερα από την τροποποίηση της συνάρτησης πιθανοφάνειας δίνονται στον επόμενο πίνακα.

Πίνακας 3.3

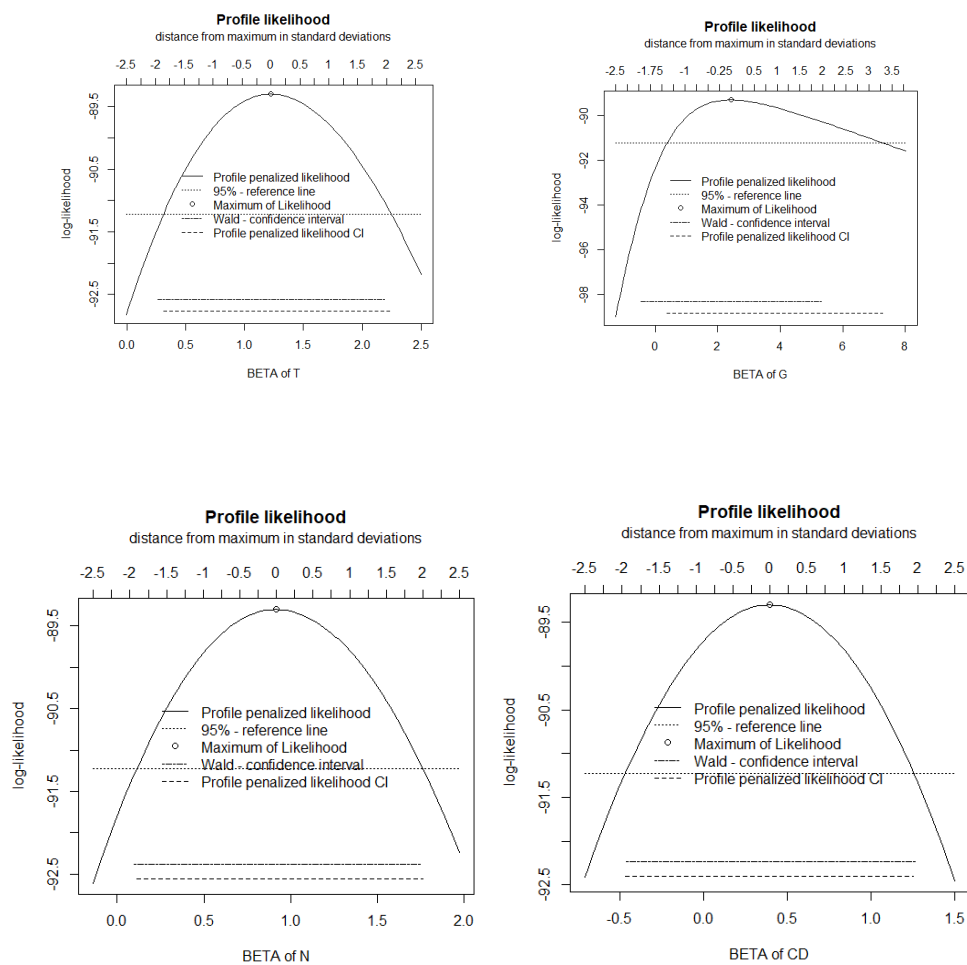
<i>Μοντέλο του Cox, ύστερα από την ποινικοποίηση:</i>	
	TIME
T (1)	1.224**
(se)	(0.491)
N (1)	0.918**
(se)	(0.422)
G (1)	2.424**
(se)	(1.473)
CD (1)	0.397
(se)	(0.441)
Observations	100
Wald Test	23.200*** (df = 4)
LR Test	35.95142*** (df = 4)

Note:

*p<0.1; **p<0.05; ***p<0.01

Προκειμένου να διαπιστώσουμε ότι οι εκτιμήσεις δεν οφείλονται και πάλι σε σφάλμα του αλγορίθμου κατασκευάζουμε τα νέα διαγράμματα της profile ποινικοποιημένης πιθανοφάνειας. Τα διαγράμματα αυτά δίνονται στη συνέχεια.

Σχήμα 3.8: Διαγράμματα profile πιθανοφανειών



Παρατηρούμε ότι η ποινικοποίηση με τη μέθοδο του Firth λύνει το πρόβλημα, αφού αφενός προκύπτει ένας ερμηνεύσιμος εκτιμητής μέγιστης πιθανοφάνειας. Ο εκτιμητής αυτός υπάρχει, είναι πεπερασμένος και μοναδικός αφού οι profile πιθανοφάνειες για όλες τις ανεξάρτητες μεταβλητές είναι αυστηρά κοίλες (strictly concave) συναρτήσεις και επομένως μεγιστοποιούνται μοναδικά. Τέλος, το μοντέλο που καταλήγουμε είναι το

$$\hat{h}(t) = \hat{h}_0(t) \exp\{1.224T(1) + 0.918N(1) + 2.424G(1) + 0.397CD(1)\}$$

ΕΠΙΛΟΓΟΣ

Η μεροληψία του εκτιμητή μέγιστης πιθανοφάνειας αποτελεί ένα από τα πιο σημαντικά προβλήματα που μπορεί να αντιμετωπίσει ένας ερευνητής στην ανάλυση δεδομένων σε περιπτώσεις μικρών δειγμάτων (small sample sizes). Επιπρόσθετα, σημαντικό πρόβλημα μπορεί να δημιουργηθεί και σε περιπτώσεις μεγάλων δειγμάτων όπου το υπο μελέτη γεγονός δεν παρατηρείται συχνά (rare event). Στην μεταπτυχιακή διατριβή αυτή παρουσιάστηκαν παραδείγματα στα οποία η μεροληψία του εκτιμητή αλλοίωσε τα αποτελέσματα και οδηγούσε σε παραπλανητικά συμπεράσματα. Ιδιαίτερα, μια τέτοια κατάσταση μπορεί να προκαλέσει εσφαλμένες εκτιμήσεις για σοβαρά κλινικά ερωτήματα τα οποία μοντελοποιούνται πολλές φορές μέσω κάποιου γενικευμένου γραμμικού μοντέλου ή με τη χρήση κάποιου μοντέλου ανάλυσης επιβίωσης. Σκοπός της διατριβής ήταν αρχικά να παρουσιάσει αυτά τα μοντέλα (Κεφάλαια 1 και 2) και στη συνέχεια να δώσει μια ανασκόπηση κάποιων μεθόδων αντιμετώπισης του προβλήματος της μεροληψίας (Κεφάλαιο 3). Η βασική μέθοδος στην οποία εστίασε η εργασία είναι η μέθοδος ποινικοποιημένης πιθανοφάνειας του Firth (Firth, 1993) η σημασία της οποίας γίνεται σαφέστερη μέσω των εφαρμογών της σε πραγματικά σύνολα δεδομένων. Τέλος, αξίζει να

αναφερθεί πως πολλοί ερευνητές αγνοούν την ύπαρξη ενός τέτοιου προβλήματος, ιδιαίτερα σε περιπτώσεις μεγάλων δειγμάτων όπου η εμφάνιση του γεγονότος είναι σπάνια, [Altman et al. (2016)]. Έτσι, πολλές από τις μελέτες παρέχουν εκτιμήσεις στις οποίες η ύπαρξη της μεροληψίας είναι αρκετά έντονη με αποτέλεσμα να δημιουργεί προβλήματα στην προσπάθεια των ερευνητών να συνθέσουν, μέσω της μετά-ανάλυσης (meta-analysis), την πληροφορία που είναι διαθέσιμη αφού τότε τα μεροληπτικά αποτελέσματα των μεμονομένων ερευνών (individual studies) συγκεντρώνονται στους συγκεντρωτικούς εκτιμητές (summary estimates). Λαμβάνοντας, υπόψιν όλα αυτά οι προτάσεις για μελλοντική έρευνα μπορούν να συνοψιστούν στις εξής 3 κατευθύνσεις:

1. Επέκταση και εφαρμογή των μεθόδων που παρουσιάστηκαν και ιδιαίτερα της μεθόδου του Firth και σε άλλες περιπτώσεις μοντέλων (για παράδειγμα, μη γραμμικά μοντέλα (non-linear models)).
2. Εφαρμογή των τροποποιημένων πιθανοφανειών σε μοντέλα μετά-ανάλυσης (meta-analysis) και μετά-ανάλυσης δικτύου (network meta-analysis) με σκοπό την βελτίωση της ακρίβειας των εκτιμήσεων.
3. Ανάπτυξη στατιστικού λογισμικού για αυτές τις προτάσεις προκειμένου να γίνει εύκολη η εφαρμογή τους σε αναλύσεις δεδομένων.

ΠΕΡΙΛΗΨΗ

Η μέθοδος της μέγιστης πιθανοφάνειας είναι η πιο διαδεδομένη μέθοδος εκτίμησης παραμέτρων σε μοντέλα παλινδρόμησης. Η συγκεκριμένη μέθοδος οφείλει τη δημοφιλία της στην ευνοϊκή συμπεριφορά των εκτιμήσεων σε περιπτώσεις μεγάλων συνόλων δεδομένων. Αντίθετα, σε περιπτώσεις μικρών συνόλων δεδομένων οι εκτιμητές που προκύπτουν με βάση αυτή τη μέθοδο είναι συχνά μεροληπτικοί με τη μεροληψία να επιδρά σημαντικά στα αποτελέσματα. Στόχος της μεταπτυχιακής διατριβής είναι: α) η ανασκόπηση μοντέλων παλινδρόμησης που χρησιμοποιούνται συχνά σε πραγματικά προβλήματα με κατηγορικά δεδομένα όπως η λογιστική παλινδρόμηση και το μοντέλο του Cox για την ανάλυση επιβίωσης, β) η ανασκόπηση μεθόδων που στοχεύουν στη μείωση της μεροληψίας των εκτιμήσεων των παραμέτρων σε μοντέλα παλινδρόμησης. Η διατριβή εστιάζει σε μεθόδους που βασίζονται στο ανάπτυγμα του Taylor καθώς πολλές αναφορές παρουσιάζουν αυτές τις μεθόδους ως ανώτερες σε σχέση με άλλες μεθόδους που υπάρχουν στη βιβλιογραφία. Ιδιαίτερη βαρύτητα δίνεται στη μέθοδο μείωσης της μεροληψίας που προτάθηκε από τον Firth (1993) διότι ο τροποποιημένος εκτιμητής μέγιστης πιθανοφάνειας διαθέτει κάποιες επιπρόσθετες θεωρητικές αλλά και πρακτικές ιδι-

ότητες. Στην μεταπτυχιακή διατριβή γίνεται εφαρμογή της μεθόδου του Firth σε δύο πραγματικά σύνολα δεδομένων με τη βοήθεια του στατιστικού πακέτου της R. Τέλος, προτείνονται θέματα για περαιτέρω έρευνα.

ABSTRACT

The maximum likelihood method is the most widely used method for the parameter estimation in regression models. This method owes its popularity to the desirable properties of the corresponding estimators in the presence of a large sample. However, when a large sample is not available the corresponding estimators are usually biased and the bias affects significantly the final results. The objective of this thesis is: a) to review some regression models for categorical data such as logistic regression and Cox model for survival analysis, b) to review methods whose aim is bias reduction in parameter estimation for regression models. This thesis focuses on methods based on the Taylor expansion, for there is a plethora of references in the literature which are referred to the superiority of those methods. The main method which is presented in that thesis is the method proposed by Firth (1993). The modified bias-reduced estimator has some superior properties over the traditional maximum likelihood estimator. After describing the theoretical background of different methods for the bias reduction of the maximum likelihood estimator we applied Firth's method in two real datasets. In our analyses we used the statistical package

of R. Finally, topics for further research are discussed.

Βιβλιογραφία

- Agresti, A. (2003). *Categorical data analysis*. John Wiley & Sons.
- Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.
- Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **71**(1):1–10.
- Anderson, J. and Richardson, S. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics*, **21**(1):71–78.
- Bartlett, M. S. (1953). Approximate confidence intervals. ii. more than one unknown parameter. *Biometrika*, **40**(3/4):306–317.
- Bewick, V., Cheek, L., and Ball, J. (2004). Statistics review 12: survival analysis. *Critical Care*, **8**(5):389–394.
- Bishop, Y. M., Fienberg, S. E., Holland, P. W., Light, R. J., and Mosteller, F. (1977). Discrete multivariate analysis: Theory and practice.
- Collett, D. (2015). *Modelling survival data in medical research*. Chapman and Hall/CRC.
- Cordeiro, G. M. and McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **53**(3):629–643.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**(2):187–202.
- Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**(2):269–276.
- Cox, D. R. and Snell, E. J. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, **30**(2):248–275.

- Dardis, C. (2018). *survMisc: Miscellaneous Functions for Survival Data*. R package version 0.5.5.
- Efron, B. et al. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, **3**(6):1189–1242.
- Firth, D. (1992a). Bias reduction, the jeffreys prior and glim. In *Advances in GLIM and Statistical Modelling*, pages 91–100. Springer.
- Firth, D. (1992b). Generalized linear models and jeffreys priors: an iterative weighted least-squares approach. In *Computational Statistics*, pages 553–557. Springer.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**(1):27–38.
- Greenland, S., Mansournia, M. A., and Altman, D. G. (2016). Sparse data bias: a problem hiding in plain sight. *BMJ*, **352**:i1981.
- Greenwood, M. et al. (1926). A report on the natural duration of cancer. *A Report on the Natural Duration of Cancer.*, (33).
- Haldane, B. J. (1956). The estimation and significance of the logarithm of a ratio of frequencies. *Annals of human genetics*, **20**(4):309–311.
- Heinze, G. and Ploner, M. (2018a). *coxphf: Cox Regression with Firth’s Penalized Likelihood*. R package version 1.13.
- Heinze, G. and Ploner, M. (2018b). *logistf: Firth’s Bias-Reduced Logistic Regression*. R package version 1.23.
- Heinze, G. and Schemper, M. (2001). A solution to the problem of monotone likelihood in cox regression. *Biometrics*, **57**(1):114–119.
- Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, **21**(16):2409–2419.

-
- Hinkley, D. V. and Cox, D. (1979). *Theoretical statistics*. Chapman and Hall/CRC.
- Hosmer Jr, D. W. and Lemeshow, S. (1999). Applied survival analysis: regression modelling of time to event data (1999). *Eur Orthodontic Soc*, pages 561–2.
- Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
- Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A*, **186**(1007):453–461.
- Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on cox’s regression and life model. *Biometrika*, **60**(2):267–278.
- Kalbfleisch, J. D. and Prentice, R. L. (2011). *The statistical analysis of failure time data*. John Wiley & Sons.
- Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, **53**(282):457–481.
- Kassambara, A. and Kosinski, M. (2018). *survminer: Drawing Survival Curves using 'ggplot2'*. R package version 0.4.3.
- Kauermann, G. and Norrie, J. (2014). *Generalized Linear Models: Overview*. American Cancer Society.
- Kelley, T. L. (1928). Crossroads in the mind of man: A study of differentiable mental abilities.
- King, G. and Zeng, L. (2001). Explaining rare events in international relations. *International Organization*, **55**(3):693–715.
- Kleinbaum, D. G. and Klein, M. (2010). *Survival analysis*. Springer.
- Knight, K. (2000). *Mathematical Statistics*. Chapman& Hall/CRC.

- Konis, K. (2007). *Linear programming algorithms for detecting separated data in binary logistic regression models*. PhD thesis, University of Oxford.
- Kosmidis, I. (2007). *Bias reduction in exponential family nonlinear models*. PhD thesis, University of Warwick.
- Kosmidis, I. (2017). *brglm: Bias Reduction in Binary-Response Generalized Linear Models*. R package version 0.6.1.
- Kosmidis, I. and Firth, D. (2018). Jeffreys' prior, finiteness and shrinkage in binomial-response generalized linear models. *arXiv preprint arXiv:1812.01938*.
- Lesaffre, E. and Albert, A. Partial separation in logistic discrimination. *Journal of the Royal Statistical Society. Series B, Methodological*, **51**(1).
- Magnus, J. and Neudecker, H. (1999). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley & Sons.
- Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, **22**(4):719–748.
- McCullagh, P. (2018). *Tensor methods in statistics*. Courier Dover Publications.
- McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall.
- Nagashima, K. and Sato, Y. (2017). Information criteria for firth's penalized partial likelihood approach in cox regression models. *Statistics in Medicine*, **36**(21):3422–3436.
- Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, **135**(3):370–384.

-
- Ntzoufras, I. (2011). *Bayesian modeling using WinBUGS*. John Wiley & Sons.
- Pace, L. and Salvan, A. (1997). *Principles of statistical inference from a neo-Fisherian perspective*. World Scientific Publishing Company.
- Pavlou, M., Ambler, G., Seaman, S., De Iorio, M., and Omar, R. Z. (2016). Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in medicine*, **35**(7):1159–1177.
- R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- R Core Team (2018). *foreign: Read Data Stored by 'Minitab', 'S', 'SAS', 'SPSS', 'Stata', 'Systat', 'Weka', 'dBase', ...* R package version 0.8-71.
- Santner, T. J. and Duffy, D. E. (1986). A note on a. albert and ja anderson's conditions for the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, **73**(3):755–758.
- Schaefer, R. L. (1983). Bias correction in maximum likelihood logistic regression. *Statistics in Medicine*, **2**(1):71–78.
- Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69**(1):239–241.
- Silvapulle, M. J. (1981). On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, **43**(3):310–313.
- Skovgaard, I. (1986). A note on the differentiation of cumulants of log likelihood derivatives. *International Statistical Review/Revue Internationale de Statistique*, **54**(1):29–32.
- Therneau, T. M. (2015). *A Package for Survival Analysis in S*. version 2.38.
- Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge university

press.

van Smeden, M., Moons, K. G., de Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J., and Reitsma, J. B. (2018). Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*, **0**(0):1–20.