ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

Στατιστικά μοντέλα για δεδομένα διάρκειας ζωής

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΔΗΜΗΤΡΙΟΥ Ε. ΣΤΟΓΙΑΝΝΗ

Διπλωματούχου σχολής Ε.Μ.Φ.Ε. Ε.Μ.Π.

ΕΠΙΒΛΕΠΟΥΣΑ:

Χ. ΚΑΡΩΝΗ

Αναπλ. Καθηγήτρια Ε.Μ.Π.

ΑΘΗΝΑ, Οκτώβριος 2012

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών

Στατιστικά μοντέλα για δεδομένα διάρκειας ζωής

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

ΔΗΜΗΤΡΙΟΥ Ε. ΣΤΟΓΙΑΝΝΗ

Διπλωματούχου σχολής Ε.Μ.Φ.Ε. Ε.Μ.Π.

ΤΡΙΜΕΛΗΣ ΣΥΜΒΟΥΛΕΥΤΙΚΗ ΕΠΙΤΡΟΠΗ:

1. Χ. ΚΑΡΩΝΗ, Αναπλ. Καθ. Ε.Μ.Π. (Επιβλέπουσα)
2. Χ. ΚΟΥΚΟΥΒΙΝΟΣ, Καθ. Ε.Μ.Π.
3. Ν. ΛΗΜΝΙΟΣ, Καθ., University of Technology of Compiegne (UTC)

ΕΠΤΑΜΕΛΗΣ ΕΞΕΤΑΣΤΙΚΗ ΕΠΙΤΡΟΠΗ

1. Χ. ΚΑΡΩΝΗ, Αναπλ. Καθ. Ε.Μ.Π. (Επιβλέπουσα)
2. Χ. ΚΟΥΚΟΥΒΙΝΟΣ, Καθ. Ε.Μ.Π.
3. Ν. ΛΗΜΝΙΟΣ, Καθ., University of Technology of Compiegne (UTC),
4. Ι. ΒΟΝΤΑ, Επικ. Καθ. Ε.Μ.Π.
5. Β. ΠΑΠΑΝΙΚΟΛΑΟΥ, Καθ. Ε.Μ.Π.
6. Ι. ΣΠΗΛΙΩΤΗΣ, Αναπλ. Καθ. Ε.Μ.Π.
7. Π. ΟΙΚΟΝΟΜΟΥ, Λέκτορας Παν. Πατρών.

ΑΘΗΝΑ, Οκτώβριος 2012

Αφιερώνεται

στους γονείς μου, Δήμητρα και Βαγγέλη

στο θείο μου, Μανώλη

στην Κατερίνα

Η εκπόνηση της παρούσας διατριβής πραγματοποιήθηκε μέσω χρηματοδότησης (υποτροφία) από τα διαθέσιμα του Ειδικού Λογαριασμού Κονδυλίων Έρευνας (Ε.Λ.Κ.Ε.) του Εθνικού Μετσόβιου Πολυτεχνείου (Ε.Μ.Π.)

Συνοπτικά περιεχόμενα

Πίνακας Περιεχομένων

Περίληψη

Μια κατανομή που βρίσκει μεγάλη εφαρμογή σε στατιστικά μοντέλα για την ανάλυση δεδομένων διάρκειας ζωής είναι η αντίστροφη Γκαουσιανή κατανομή (Inverse Gaussian - IG) με σ.π.π.:

$$f\left(t;\mu,\lambda\right)=\sqrt{\frac{\lambda}{2\pi t^3}}\exp\left\{-\frac{\lambda\left(t-\mu\right)^2}{2\mu^2 t}\right\},\quad t>0,\quad \mu,\ \lambda>0.$$

Ένας σημαντικός λόγος για το εύρος των εφαρμογών της αποτελεί το γεγονός ότι προκύπτει ως κατανομή του χρόνου 1ης μετάβασης σε ανέλιξη Wiener στην παλινδρόμηση Κατωφλιού (Threshold regression ή First Hitting Time – FHT-regression). Βασικό στόχο της διδακτορικής διατριβής αποτελεί η συμβολή στην περαιτέρω ανάπτυξη του θεωρητικού υποβάθρου της παλινδρόμησης Κατωφλιού.

Αρχικά, γίνεται μία εκτενής μελέτη της αντίστροφης Γκαουσιανής κατανομής (IG distribution) και μελετώνται γενικές ιδιότητες της κατανομής που τη συμπεριλαμβάνουν ως γενικευμένο γραμμικό μοντέλο. Το γενικευμένο γραμμικό μοντέλο (GLM) παρουσιάστηκε από τους Nelder and Wedderburn (1972) και αποτελεί μία ενοποίηση γραμμικών και μη γραμμικών μοντέλων παλινδρόμησης, τα οποία επιτρέπουν στον πειραματιστή να διαλέξει για τη μεταβλητή απόκρισης μία κατανομή που είναι μέλος της εκθετικής οικογένειας κατανομών. Η Κανονική, η Διωνυμική, η Εκθετική και η αντίστροφη Γκαουσιανή κατανομή είναι μεταξύ άλλων, κάποιες από τις κατανομές αυτής της οικογένειας. Η αντίστροφη Γκαουσιανή κατανομή είναι το πιο σπάνια χρησιμοποιούμενο γενικευμένο γραμμικό μοντέλο.

Η IG ανήκει στην εκθετική οικογένεια κατανομών τάξης 2. Σε αυτή, συναντώνται οι εξής τέσσερις συναρτήσεις σύνδεσης: α) η κανονική, όπου $\eta=g\left(\mu\right)=\frac{1}{\mu^2}$, β) η αντίστροφη με $\eta=g\left(\mu\right)=\frac{1}{\mu}$, γ) η ταυτοτική με $\eta=g\left(\mu\right)=\mu$ και δ) η λογαριθμική, στην οποία $\eta=g\left(\mu\right)=\log\mu$. Σε κάποιες περιπτώσεις, η κανονική συνάρτηση μπορεί να επιλεχθεί εκ των προτέρων, συνήθως για ευκολία στην ερμηνεία. Η χρήση της κανονικής συνάρτησης σύνδεσης σε ένα GLM προσφέρει μερικά τεχνικά πλεονεκτήματα αλλά δεν είναι απαραίτητη, π.χ. η παλινδρόμηση Poisson δε γίνεται υποχρεωτικά με τη λογαριθμική συνάρτηση. Για κάποιες περιπτώσεις όμως, μπορεί να μην είναι προφανής η επιλογή της

κατάλληλης συνάρτησης σύνδεσης. Στην περίπτωση της συνηθισμένης $IG(\mu,\sigma^2)$ η κανονική συνάρτηση σύνδεσης είναι $\frac{1}{\mu^2}$ αλλά επίσης χρησιμοποιούνται η λογαριθμική, η ταυτοτική και η αντίστροφη.

Στηριζόμενοι στο έργο των Myers and Montgomery (1997), αλλά και στους Lewis et al. (2001a και 2001b), μελετούμε την επίδραση που έχει μία λανθασμένη επιλογής συνάρτησης σύνδεσης στις εκτιμήσεις της κάλυψης και της ακρίβειας (μήκος) ενός διαστήματος εμπιστοσύνης για την παράμετρο της μέσης τιμής της IG κατανομής. Τελικά, συμπεραίνουμε με τη βοήθεια προσομοιώσεων πως σε ένα IG GLM μεγάλο ρόλο κατέχει η επιλογή της κατάλληλης συνάρτησης σύνδεσης, ιδιαίτερα όταν αυτή δεν είναι η κανονική.

Στη συνέχεια μελετούμε τα διάφορα υπόλοιπα που μπορούν να χρησιμοποιηθούν για την IG κατανομή. Τα υπόλοιπα κατέχουν κεντρικό ρόλο στην προσαρμογή του γενικού γραμμικού μοντέλου και χρησιμοποιούνται ευρύτατα για την αξιολόγηση της καταλληλότητας όχι μόνο των γενικευμένων γραμμικών μοντέλων (McCullagh and Nelder, 1989), αλλά και των μοντέλων PH, AL και PO (Collett, 2003). Κατά τη διάρκεια της διατριβής μελετώνται τα διάφορα διαθέσιμα υπόλοιπα για την κατανομή IG και κατασκευάστηκαν αλγόριθμοι για την παραγωγή τους μέσω της R. Ειδικότερα, διερευνάται η σχέση μεταξύ των Pearson, Anscombe και Deviance υπολοίπων. Αποδεικνύεται πως σε ειδικές περιπτώσεις τα Anscombe και τα Deviance έχουν πολύ κοντινές τιμές στην IG. Τα διάφορα ευρήματα καταγράφονται τόσο θεωρητικά όσο και με τη βοήθεια προσομοιώσεων που έγιναν στην R.

Όπως και σε όλα τα στατιστικά μοντέλα, είναι σημαντικός ο εντοπισμός άτυπων σημείων (outliers), δεδομένων με ασυνήθιστα διαφορετικές τιμές και συμπεριφορά από τα υπόλοιπα, τα οποία ενδεχομένως έχουν μεγάλη επιρροή στην προσαρμογή του μοντέλου. Για το λόγο αυτό, έχουν αναπτυχθεί έλεγχοι εντοπισμού άτυπων σημείων για πολλές κατανομές. (Barnett and Lewis, 1994). Ωστόσο, δεν έχουν ακόμα αναπτυχθεί αντίστοιχες τεχνικές για την κατανομή IG. Βασισμένοι στο έργο των Chhikara και Folks (1989) και της Davis (1980), κατασκευάζουμε ελέγχους για τον εντοπισμό άτυπων τιμών των παραμέτρων (μ,λ) της IG σε δεδομένα ανεξάρτητων μονάδων με επαναλαμβανόμενα ανεξάρτητα γεγονότα ανά μονάδα. Οι διάφοροι έλεγχοι βασίζονται στη μεγιστοποίηση της τιμής του ελέγχου του λόγου των πιθανοφανειών για την ισότητα παραμέτρων, με διόρθωση Bonferroni για τις p-τιμές. Προσομοιώσεις γίνονται προκειμένου να επιβεβαιώσουμε την ακρίβεια των Bonferroni ελέγχων υπό τη μηδενική υπόθεση και να μελετήσουμε την ισχύ των ελέγχων υπό την εναλλακτική υπόθεση. Στη συνέχεια, χρησιμοποιείται μία εναλλακτική παραμέτρηση για τη συνάρτηση πυκνότητας πιθανότητας της IG, η οποία βρίσκεται σε αντιστοιχία με τις παραμέτρους (μ,λ) και δίνεται από τη σχέση:

$$f(t|m,\sigma^2,x_0)=\frac{x_0}{\sqrt{2\pi\sigma^2 t^3}}\exp\left[-\frac{(x_0+mt)^2}{2\sigma^2 t}\right],\ t>0,\ \ m<0,\sigma^2>0,x_0>0$$

Η παραπάνω σχέση είναι η παραμέτρηση της συνάρτησης πυκνότητας πιθανότητας του χρόνου πρώτης μετάβασης που ακολουθεί την IG και χρησιμοποιείται στην παλινδρόμηση Κατωφλιού.

Το ημι-παραμετρικό μοντέλο αναλογικής διακινδύνευσης (PH) του Cox είναι ευρέως διαδεδομένο σε εφαρμογές δεδομένων διάρκειας ζωής και αποτελεί συχνά την πρώτη επιλογή του πειραματιστή ως εργαλείο ανάλυσης της επιβίωσης. Ωστόσο, το συγκεκριμένο μοντέλο θέτει κάποιους περιορισμούς στην πιθανή μορφή της συνάρτησης διακινδύνευσης. Εναλλακτικά, όλο και περισσότερο έχουν αρχίσει να χρησιμοποιούνται μοντέλα βασισμένα στο χρόνο πρώτης διακοπής (FHT) μίας στοχαστικής ανέλιξης. Στη διατριβή συγκρίνουμε το μοντέλο του Cox και ένα μοντέλο FHT παλινδρόμησης βασισμένο σε ανέλιξη Wiener, το οποίο οδηγεί σε χρόνο πρώτης διακοπής που ακολουθεί την IG κατανομή.

Διαγνωστικές τεχνικές αναπτύσσονται για την καταλληλότητα του μοντέλου και γίνεται διερεύνηση πρακτικών θεμάτων στην προσαρμογή του μοντέλου παλινδρόμησης χρόνου πρώτης μετάβασης (IG FHTR). Στην παλινδρόμηση Κατωφλιού, μια συμμεταβλητή μπορεί να επηρεάζει τη διάρκεια ζωής με δύο τρόπους, και να υπάρχει ένας βαθμός μη αναγνωρισημότητας ή πολυσυγγραμικότητας στο μοντέλο. Η συχνή παρουσία αντιφατικών εκτιμήσεων των δύο επιδράσεων μιας συμμεταβλητής σε δημοσιευμένες εφαρμογές του FHT μοντέλου, μπορεί να αποτελεί επιβεβαίωση αυτής της δυσκολίας. Κατά τη διάρκεια της διατριβής γίνεται μία προσπάθεια να αποδοθούν εμπειρικές αποδείξεις σχετικά με τη δυνατότητα προσαρμογής του μοντέλου. Ειδικότερα, εξετάζεται εάν υπάρχει κάποια ένδειξη κατά τη διαδικασία προσαρμογής να τοποθετεί μία μεταβλητή σε λάθος παράμετρο. Επιπρόσθετα, ερευνούμε το φαινόμενο εμφάνισης αντίθετων προσήμων μίας μεταβλητής στις διάφορες παραμέτρους της κατανομής.

Στη συνέχεια προτείνουμε μία διαδικασία επιλογής μεταβλητών για την περίπτωση του IG FHT μοντέλου παλινδρόμησης. Η ύπαρξη μίας τέτοιας διαδικασίας θεωρείται αναγκαία για την Ανάλυση Επιβίωσης, ιδιαίτερα σε ιατρικές εφαρμογές στις οποίες συνήθως υπάρχει ένας μεγάλος αριθμός διαθέσιμων υποψήφιων μεταβλητών για τις επιμέρους αναλύσεις. Η προτεινόμενη διαδικασία αποτελείται από δύο διαδοχικές εφαρμογές της προσαρμοσμένης LASSO τεχνικής (adaptive LASSO) εκτελεσμένες από έναν αλγόριθμο ελαχίστων τετραγώνων. Η διαδικασία αποδεικνύεται αποτελεσματική για την ορθή αναγνώριση των μη–μηδενικών (στατιστικά σημαντικών) συντελεστών της παλινδρόμησης. Η μελέτη αυτή αποτελεί την πρώτη συνδρομή στη μεθοδολογία μοντελοποίησης, η οποία είναι απαραίτητο να αναπτυχθεί περαιτέρω για το παρόν μοντέλο παλινδρόμησης.

Μετά τη μοντελοποίηση και τον εντοπισμό άτυπων τιμών, πολύ σημαντική για την Ανάλυση Επιβίωσης θεωρείται η αναγνώριση σημείων επιρροής κατά την προσαρμογή του

μοντέλου. Με τον όρο επιρροή, εννοούμε την επίδραση της κάθε παρατήρησης στην προσαρμογή του μοντέλου. Στη βιβλιογραφία υπάρχει πληθώρα τεχνικών για την αναγνώριση σημείων επιρροής (Cook και Weisberg, 1982 και Therneau και Grambsch, 2000). Ωστόσο, δεν υπάρχει κάποια παρόμοια τεχνική για την παλινδρόμηση Κατωφλιού. Σκοπό αυτού του τμήματος της διατριβής αποτελεί η μελέτη και η ανάπτυξη μεθόδων για τον εντοπισμό σημείων επιρροής για την περίπτωση του IG FHT μοντέλου παλινδρόμησης. Αναπτύσσουμε μία μέθοδο βασισμένη στην τεχνική αφαίρεσης σημείου (Case Deletion Model - CDM), προκειμένου να μετρήσουμε την επιρροή της καθεμιάς παρατήρησης. Ακόμα, τα διάφορα στατιστικά μοντέλα συνήθως έχουν κάποιο βαθμό προσέγγισης, με αποτέλεσμα να είναι συνήθως λανθασμένα. Για το λόγο αυτό, η αξιολόγηση της επιρροής μικρών διαταραχών του μοντέλου είναι ιδιαίτερα σημαντική. Ο Cook (1986) ανέπτυξε μία τέτοια μέθοδο μέτρησης της τοπικής επιρροής, η οποία δεν περιορίζεται μόνο σε γραμμικά μοντέλα. Την τεχνική αυτή την επεκτείνουμε και την προσαρμόζουμε κατάλληλα για το FHT μοντέλο παλινδρόμησης. Η εγκυρότητα των διαφόρων θεωρητικών αποτελεσμάτων ελέγχεται με τη βοήθεια προσομοιώσεων.

Όλες οι μελέτες έγιναν με τη βοήθεια του στατιστικού πακέτου R, το οποίο έχει προγραμματιστικό περιβάλλον και είναι κατάλληλο για μελέτες με τη βοήθεια προσομοιώσεων. Επιπρόσθετα, χρησιμοποιήθηκε και πληθώρα στατιστικών πακέτων, όπως είναι τα SPSS, STATA, MINITAB, κυρίως για τις διάφορες εφαρμογές πραγματικών δεδομένων.

Δημήτρης Στογιάννης

Abstract

A distribution that is often applied in statistical models for lifetime data is the inverse Gaussian (IG) with p.d.f.:

$$f(t;\mu,\lambda)=\sqrt{\frac{\lambda}{2\pi t^3}}\exp\left\{-\frac{\lambda(t-\mu)^2}{2\mu^2 t}\right\},\quad t>0,\quad \mu,\ \lambda>0.$$

An important reason for the variety of its applications is that it arises as the distribution of the lifetime regarded as the first hitting time in a Wiener process in Threshold regression (First Hitting Time regression– FHT regression) (Lee and Whitmore, 2006). The main purpose of this thesis is to contribute to the development of the theoretical and mathematical framework of Threshold regression.

Firstly, we study in depth the inverse Gaussian distribution within the framework of generalized linear models (GLM). The GLM approach was first suggested by Nelder and Wedderburn (1972) and consists of a unification of linear and non-linear regression models, allowing the experimenter to select for the response variable a distribution which is a member of the exponential family. Normal, Binomial, Exponential and inverse Gaussian distributions are amongst the distributions of the exponential family. The inverse Gaussian distribution appears to be the most rarely used GLM. Two types of link functions exist in GLM: canonical and non-canonical. The canonical link is that function which equates the natural location parameter of the exponential family to the linear predictor. Four link functions can be used with the IG distribution: a) the canonical, where $\eta=g(\mu)=\frac{1}{\mu^2}$, b) the inverse, where $\eta=g(\mu)=\frac{1}{\mu}$, c) the identity, where $\eta=g(\mu)=\mu$ and d) the logarithmic, where $\eta=g(\mu)=\log\mu$. Sometimes, the canonical link can be pre-selected, but the appropriate choice of the link function is not always obvious. In the case of the usual $IG(\mu,\sigma^2)$ the canonical link function is $\frac{1}{\mu^2}$, but the logarithmic, the identical and the inverse links can also be used. Based on the work of Myers and Montgomery (1997) and Lewis et al. (2001a and 2001b), we investigate the impact of the choice of link function on the coverage and precision (length) of a confidence interval for the mean parameter μ of the IG distribution.

We find that in an IG GLM, the correct choice of the appropriate link function is important, especially when it is not the canonical.

We study several types of residuals that can be used with the IG distribution. Residuals can be used to explore the adequacy of fit of a model, in respect of choice of variance function, link function and terms in the linear predictor. They may also be used to evaluate the appropriateness of the generalized linear models (McCullagh and Nelder, 1989) and of the proportional hazards, accelerated life and proportional odds models as well (Collett, 2003). For the purpose of this thesis, we study the various residuals that can be used with the IG and we construct algorithms to obtain them in R. More specifically, we study the relationship between Pearson, Anscombe and deviance residuals. It is proved that in special cases, Anscombe and deviance residuals have similar values in the case of IG. The various findings are presented not only theoretically but also through simulation studies using R.

Because it is important in practical data analysis to identify observations that seem to be inconsistent with the rest of the data, outlier tests have been developed for many statistical distributions (Barnett and Lewis, 1994). Outlier tests for the IG are not available in the literature, even though this distribution is widely used in statistical modelling and more specifically in the analysis of lifetime data. Based on the work of Chhikara and Folks (1977) and Davis (1980), we construct tests for outlying values of the parameters (μ, λ) of this distribution when data are available from a sample of independent units and possibly with more than one event per unit. These outlier tests are constructed from likelihood ratio tests for equality of parameters. Simulation studies are used to confirm that Bonferroni tests have accurate size and to examine the powers of the tests. When the IG arises in threshold regression as the lifetime distribution, we use the alternative parameterization:

$$f(t|\, m, \sigma^2, x_0) = \frac{x_0}{\sqrt{2\pi\sigma^2 t^3}} \exp\left[-\frac{(x_0 + mt)^2}{2\sigma^2 t} \right],\ t > 0,\ \ m < 0, \sigma^2 > 0, x_0 > 0$$

of the p.d.f. The application of the outlier tests to (x_0,m) is shown.

Cox's widely used semi-parametric proportional hazards regression places restrictions on the possible shapes of the hazard function, even though it is not modeled directly. Various authors, but particularly Aalen and Gjessing (2001), have promoted an alternative way of approaching Survival Analysis, instead of through the hazard rate. Models based on the first hitting time of a stochastic process are among the alternatives and have the attractive feature of being based on a model of the underlying process. We review and compare the PH model and an FHT model based on a Wiener process which leads to an inverse Gaussian regression model. This particular model can also represent a "cured fraction" or long-term survivors. A case study of survival after coronary artery bypass grafting is used to examine the interpretation of the IG model, especially in relation to covariates that affect both of its parameters.

We develop some diagnostic techniques for the appropriateness of the model and we investigate some practical matters that arise when fitting an IG FTHR model. Various authors have commented that dependence of both parameters on the same covariate may imply multicollinearity. The frequent appearance of conflicting signs for the two coefficients of the same covariate may be related to this. We carry out simulation studies to examine the reality of this possible multicollinearity. Although there is some dependence between estimates, multicollinearity does not seem to be a major problem. Moreover, we examine whether the phenomenon of the conflicting signs of estimates may be due to model misspecification.

We propose a procedure for variable selection in the IG FHT regression model for lifetime data. This procedure meets an important need because in many studies, particularly in the field of medical applications, a large number of covariates are available and should be considered for inclusion in the final model. It consists of two applications of the adaptive LASSO implemented by a least squares approximation. The procedure is shown to be effective in identifying correctly the non-zero regression coefficients. This is the first contribution to the model-building methodology that needs to be developed for this model.

After modelling, in addition to the detection of outliers, the identification of influential observations is an issue of extreme importance in Survival Analysis. By influence we mean the impact of each point on the fit of a model. A number of different techniques for investigating influence diagnostics exist in the literature (Cook and Weisberg, 1982; Therneau and Grambsch, 2000). However, there is no influence diagnostics method for FHTR models. The purpose of this part of the thesis is to develop and propose influence diagnostics for the IG FHTR model. We construct a case-deletion diagnostic method (CDM) for the case of a FHTR model, where lifetimes follow the IG. Finally, we use the local influence approach to develop influence measures for identifying observations that have a disproportionate effect on the maximum likelihood estimate of parameters in models for lifetime data. Cook (1986) proposed a method based on differential geometry to assess the local influence of minor perturbations that can be applied to a wide variety of statistical models. We extend this technique for the case of the IG FHTR model.

All studies were conducted in R, a statistical package with programming interface, which is appropriate for simulation studies. Several other statistical software packages, including SPSS, STATA and MINITAB were used for a number of applications that were carried out for the purposes of this thesis.

Dimitris Stogiannis

Ευχαριστίες

Θα ήθελα ολόθερμα να ευχαριστήσω την Χρυσηΐδα Καρώνη, Αναπληρώτρια Καθηγήτρια Ε.Μ.Π. και επιβλέπουσα καθηγήτρια της διδακτορικής αυτής διατριβής, για την καθοδήγηση, τη μεθοδικότητα, τις συνεχείς συμβουλές και τη συνολική της υποστήριξη και συνεργασία, σε όλη τη διάρκεια εκπόνησης της διατριβής. Χωρίς τη συνδρομή της, το εγχείρημα αυτό δε θα ήταν δυνατό να ολοκληρωθεί.

Ιδιαίτερα, θα ήθελα να ευχαριστήσω τον Ιωάννη Κολέτσο, Επίκουρο Καθηγητή Ε.Μ.Π., ο οποίος με καθοδήγησε σε όλη τη διάρκεια των σπουδών μου, με συνεχή υποστήριξη, ενθάρρυνση και πολύτιμες συμβουλές. Επίσης, ευχαριστώ τον Δημήτρη Φουσκάκη, Επίκουρο Καθηγητή Ε.Μ.Π. για τις συμβουλές του κατά τη διάρκεια των μεταπτυχιακών μου σπουδών.

Ακόμα, θα ήθελα να ευχαριστήσω τους συναδέλφους μαθηματικούς Δήμο Γκουνταρούλη, Στρατούλα Χαριτίδου και Άννα Σκούντζου, για τις πολύτιμες συμβουλές και τη συμπαράσταση κατά τη διάρκεια εκπόνησης της διατριβής. Ευγνώμον είμαι στον Σπύρο Τζώρτσο, καθηγητή φιλόλογο, για τη φιλολογική επιμέλεια της διατριβής.

Επίσης, αυτή η προσπάθεια δε θα ήταν δυνατό να ολοκληρωθεί χωρίς την αμέριστη συμπαράσταση της οικογένειάς μου και ιδιαίτερα του θείου Μανώλη, καθηγητή Μαθηματικών και επιχειρηματία, ο οποίος έχει αναλάβει με σκληρό προσωπικό κόπο και συνεχή αγώνα, όντας πάντα χαμογελαστός, το “ευ ζειν” μου. Τέλος, ένα μεγάλο ευχαριστώ στην Κατερίνα, για την υπομονή της.

Δημήτρης Στογιάννης

Κεφάλαιο 1

Δεδομένα διάρκειας ζωής

1.1 Εισαγωγή στην ανάλυση επιβίωσης

Ανάλυση Επιβίωσης είναι η φράση που χρησιμοποιείται για να περιγράψει την ανάλυση δεδομένων διάρκειας ζωής (lifetime data) ή διάρκειας μέχρι να συμβεί κάποιο (συνήθως ανεπιθύμητο) γεγονός. Τα δεδομένα διάρκειας ζωής προκύπτουν σε πολλές επιστημονικές περιοχές αλλά ιδιαίτερα στην Ιατρική (Ανάλυση Επιβίωσης - Survival Analysis) και στις τεχνολογικές επιστήμες (Ανάλυση Αξιοπιστίας - Reliability). Κατεξοχήν παραδείγματα δεδομένων διάρκειας ζωής αποτελούν οι χρόνοι από την εκδήλωση συμπτωμάτων ασθένειας έως τη στιγμή του θανάτου από αυτήν την ασθένεια, χρόνος βλάβης κινητήρα πλοίου για πρώτη φορά, κ.τ.λ. Πέρα από την περίπτωση του χρόνου μέχρι την εκδήλωση ενός γεγονότος, περιλαμβάνει και άλλες περιπτώσεις όπου η απόκριση είναι μια μη αρνητική τυχαία μεταβλητή, όπως το φορτίο θραύσης ενός υλικού.

Είναι σημαντικό να διαθέτουμε στατιστικά μοντέλα που λαμβάνουν υπόψη την επίδραση συμμεταβλητών (όπως η ηλικία του ασθενή) στη διάρκεια ζωής. Κυριαρχούν δύο μεθοδολογικές προσεγγίσεις στην ανάλυση δεδομένων διάρκειας ζωής με συμμεταβλητές. Ενώ στις τεχνολογικές επιστήμες εμφανίζονται πιο συχνά παραμετρικά μοντέλα **επιταχυνόμενης ζωής** (Accelerated Life, AL), στις ιατρικές επιστήμες χρησιμοποιείται συνήθως το ημι-παραμετρικό μοντέλο **αναλογικής διακινδύνευσης του Cox** (Proportional Hazards, PH). Άλλες προσεγγίσεις, όπως το μοντέλο **αναλογικών συμπληρωματικών πιθανοτήτων** (Proportional Odds, PO), χρησιμοποιούνται σχετικά σπάνια, εν μέρει επειδή απουσιάζουν από τα γνωστά στατιστικά πακέτα.

Πρόσφατα έχει αναπτυχθεί ενδιαφέρον σε μοντέλα όπου η διάρκεια ζωής μοντελοποιείται ως ο χρόνος πρώτης μετάβασης (First Hitting Time - FHT) μίας (συνήθως λανθάνουσας) στοχαστικής ανέλιξης από μία αρχική κατάσταση έως ένα σύνορο. Όπως και

στα ευρέως διαδεδομένα μοντέλα της αναλογικής διακινδύνευσης και επιταχυνόμενης διακοπής, η επίδραση συμμεταβλητών εισάγεται στα FHT μοντέλα μέσω δομών παλινδρόμησης, η οποία είναι γνωστή στη βιβλιογραφία ως **παλινδρόμηση Κατωφλιού** (Threshold Regression - TR).

Η παρούσα διατριβή έχει δύο βασικούς σκοπούς. Ο πρώτος αφορά τη συμβολή στην ανάπτυξη του θεωρητικού μαθηματικού υποβάθρου της παλινδρόμησης Κατωφλιού, κυρίως με την ανάπτυξη διαγνωστικών τεχνικών για την εξέταση της καταλληλότητας του FHT μοντέλου, οι οποίες μέχρι στιγμής δεν υπάρχουν στη βιβλιογραφία. Ο δεύτερος αφορά την επέκταση των FHT μοντέλων για την περίπτωση επαναλαμβανόμενων γεγονότων στην ίδια μονάδα (recurrent events), η οποία απαιτείται για την εφαρμογή του μοντέλου στη μελέτη επισκευάσιμων συστημάτων και αλλού.

Στη συνέχεια του κεφαλαίου, θα δοθούν κάποιες εισαγωγικές έννοιες της Ανάλυσης Επιβίωσης. Ακολουθεί η περιγραφή μη-παραμετρικών μεθόδων εκτίμησης των συναρτήσεων επιβίωσης και διακινδύνευσης, και παρουσιάζεται ο μη-παραμετρικός έλεγχος υπόθεσης Log Rank. Παρουσιάζονται συνοπτικά τα γνωστά μοντέλα PH, AL και PO, τα οποία χρησιμοποιούνται στις διάφορες εφαρμογές, ενώ ακολουθεί αναλυτική περιγραφή των FHT μοντέλων. Τέλος, εισάγεται η έννοια της παλινδρόμησης Κατωφλιού, με την οποία θα ασχοληθούμε αναλυτικά στα υπόλοιπα κεφάλαια της διατριβής.

1.1.1 Ορισμοί

Η διάρκεια ζωής της μονάδας, *t*, μπορεί να θεωρηθεί ως η αριθμητική τιμή μίας τυχαίας μεταβλητής *T*, η οποία λαμβάνει **μόνο μη αρνητικές τιμές**. Οι διάφορες τιμές της ακολουθούν μία κατανομή πιθανότητας. Έστω λοιπόν ότι η *T*, ως τυχαία μεταβλητή έχει μία συνάρτηση πυκνότητας πιθανότητας $f(t)$. Η συνάρτηση κατανομής της *T* δίνεται από τον εξής τύπο:

$$F(t) = \mathrm{P}(T \le t) = \int_0^t f(u)du$$

και αναπαριστά την πιθανότητα ο χρόνος ζωής να έχει τιμή μικρότερη από την *t*. Με τη βοήθεια της συνάρτησης κατανομής μπορούμε να ορίσουμε τη **συνάρτηση επιβίωσης** $S(t)$, ως εξής:

$$S(t) = \mathrm{P}(T > t) = 1 - F(t) = 1 - \int_0^t f(u)du = \int_t^{+\infty} f(u)du, \qquad (1.1)$$

η οποία αναπαριστά την πιθανότητα η διάρκεια ζωής της μονάδας να υπερβεί τη χρονική στιγμή *t*.

Παρατήρηση: Από την (1.1), προκύπτει ότι $f(t)=\frac{d}{dt}F(t)=\frac{d}{dt}\left(1-S(t)\right)=-\frac{d}{dt}S(t)$.

Στη συνέχεια παρουσιάζουμε τη **συνάρτηση διακινδύνευσης**, $h(t)$, η οποία εκφράζει τον κίνδυνο θανάτου της μονάδας μέσα στο διάστημα $(t, t+\delta t)$, δεδομένου ότι έχει επιβιώσει μέχρι τη χρονική στιγμή *t* και ορίζεται ως εξής:

$$h(t)=\lim_{\delta t\to 0}\frac{\mathrm{P}\left(t\le T\le t+\delta t\middle| T\ge t\right)}{\delta t}=\frac{f(t)}{S(t)} \tag{1.2}$$

Η ποσότητα $h(t)\delta t$ είναι η πιθανότητα της επικείμενης διακοπής μίας μονάδας δεδομένου ότι έζησε μέχρι τη χρονική στιγμή *t*. Για παράδειγμα, αν ο χρόνος ζωής ατόμων σε πυρηνικό αντιδραστήρα μετριέται σε δευτερόλεπτα, τότε η ποσότητα $h(t)\delta t$ εκφράζει την πιθανότητα ενός ατόμου του οποίου ο πυρήνας δεν έχει διασπαστεί μέχρι το *t*-οστό δευτερόλεπτο, να υποστεί σχάση στο αμέσως επόμενο δευτερόλεπτο.

Μία άλλη χρήσιμη συνάρτηση είναι η **σωρευτική συνάρτηση διακινδύνευσης**, $H(t)$, η οποία ορίζεται ως εξής:

$$H(t)=\int_0^t h(u)\,du \tag{1.3}$$

Από τις (1.2), (1.3) έχουμε:

$$H(t)=-\log\left\{S(t)\right\} \tag{1.4}$$

Στην ανάλυση των δεδομένων διάρκειας ζωής, οι συναρτήσεις $S(t)$ και $H(t)$ εκτιμώνται από τους παρατηρούμενους χρόνους επιβίωσης. Ακολουθούν μη παραμετρικές μέθοδοι εκτίμησης, οι οποίες δεν απαιτούν την πρότερη (a-priori) γνώση της συνάρτησης πυκνότητας πιθανότητας της τυχαίας μεταβλητής T.

Παρατήρηση

Οι μέθοδοι ανάλυσης δεδομένων διάρκειας ζωής πρέπει να λάβουν υπόψη το φαινόμενο της **αποκοπής**. Όταν ένα πείραμα, στο οποίο καταγράφονται χρόνοι ζωής στατιστικών μονάδων, τερματίζεται, είναι πολύ πιθανό κάποιες από τις μονάδες αυτές να συνεχίζουν ακόμα να ζουν. Παρόλο που δε γνωρίζουμε ακριβώς τη διάρκεια ζωής μίας τέτοιας μονάδας, εντούτοις ξέρουμε ότι έχει ξεπεράσει το χρονικό διάστημα κατά το οποίο η μονάδα βρίσκεται υπό επιτήρηση μέσα στο πείραμα. Αυτή η πληροφορία πρέπει να περιληφθεί στην ανάλυση των δεδομένων.

1.1.3 Μη-παραμετρικές μέθοδοι εκτίμησης των συναρτήσεων επιβίωσης και διακινδύνευσης

Η βασική διαδικασία που ακολουθείται σε μία ανάλυση επιβίωσης, στην οποία το μόνο που γνωρίζουμε με βεβαιότητα είναι οι παρατηρήσεις μας (αποκομμένοι και μη χρόνοι διακοπής), είναι να προσπαθήσουμε να βρούμε τη μορφή των συναρτήσεων $S(t)$ και $H(t)$ (Collett, 2003). Στη συνέχεια, παρουσιάζονται τέτοιες μέθοδοι, οι οποίες καλούνται μη-παραμετρικές, μιας και δεν απαιτούν την ικανοποίηση συγκεκριμένων προϋποθέσεων σχετικά με την υποβόσκουσα κατανομή των χρόνων επιβίωσης.

Ένα γράφημα μίας εκτιμήτριας της συνάρτησης επιβίωσης, $\hat{S}(t)$, σε σχέση με το χρόνο είναι συνήθως ένας χρήσιμος τρόπος για την περιγραφή της επιβίωσης μίας ομάδας δεδομένων.

i. Στην περίπτωση που δεν υπάρχουν αποκοπές, μία μη-παραμετρική εκτίμηση της συνάρτησης επιβίωσης δίνεται από τον παρακάτω τύπο:

$$\hat{S}(t) = \frac{\text{αριθμός μονάδων με χρόνους επιβίωσης} > t}{n},$$

όπου *n* ο αριθμός των παρατηρήσεων.

ii. Στην περίπτωση, όμως, που έχουμε αποκομμένα δεδομένα, μία εκτίμηση της συνάρτησης επιβίωσης δίνεται από τη δειγματική εκτιμήτρια **Kaplan-Meier**. Αρχικά, πρέπει να διατάξουμε τις παρατηρήσεις μας σε αύξουσα σειρά έτσι ώστε $t_{(1)} \le t_{(2)} \le \ldots \le t_{(n)}$, όπου $t_{(j)}$ είναι **ο *j*-οστός χρόνος διακοπής**. Στη συνέχεια, η εκτιμήτρια Kaplan-Meier υπολογίζεται από τον παρακάτω τύπο:

$$\hat{S}(t) = \prod_{j:t_{(j)} \le t} \left(\frac{r_j - d_j}{r_j} \right) = \prod_{j:t_{(j)} \le t} \left(1 - \frac{d_j}{r_j} \right), \tag{1.5}$$

με r_j να συμβολίζει τον αριθμό των παρατηρήσεων που βρίσκονται σε κίνδυνο ακριβώς πριν από τη χρονική στιγμή $t_{(j)}$ (συμπεριλαμβανομένων και των αποκομμένων παρατηρήσεων της χρονικής στιγμής $t_{(j)}$) και d_j να είναι ο αριθμός των παρατηρήσεων που καταστρέφονται κατά τη χρονική στιγμή $t_{(j)}$. Μία εκτίμηση της διασποράς της Kaplan-Meier είναι η,

$$\text{Var}\left(\hat{S}(t)\right) = \left(\hat{S}(t)\right)^2 \sum_{j:t_{(j)} \le t} \frac{d_j}{r_j\left(r_j - d_j\right)},$$

από την οποία προκύπτει το τυπικό σφάλμα της Kaplan Meier, που δίνεται από τη σχέση:

$$se\left(\hat{S}(t)\right)=\left(\hat{S}(t)\right)\left(\sum_{j:\, t_{(j)}\le t}\frac{d_j}{r_j\left(r_j-d_j\right)}\right)^{\frac{1}{2}}, \tag{1.6}$$

η οποία είναι γνωστή ως τύπος του Greenwood (1926). Με τη βοήθεια της εκτιμήτριας της διασποράς της Kaplan Meier μπορούμε να κατασκευάσουμε διαστήματα εμπιστοσύνης για τη συνάρτηση επιβίωσης. Αν θεωρήσουμε ότι $\hat{S}(t)\sim \mathrm{N}\left(S(t),\sigma^2\right)$, τότε για την τυχαία μεταβλητή $Z=\frac{\hat{S}(t)-S(t)}{\sqrt{\sigma^2}}$ ισχύει ότι $Z\sim \mathrm{N}(0,1)$. Εάν αντικαταστήσουμε την τυπική απόκλιση με το τυπικό σφάλμα της σχέσης (1.6), προκύπτει τελικά το παρακάτω 95% διάστημα εμπιστοσύνης για τη συνάρτηση επιβίωσης:

$$\hat{S}(t)\pm 1.96\cdot \mathrm{se}\left(\hat{S}(t)\right)$$

Τέλος, αντί για την εκτιμήτρια της συνάρτησης διακινδύνευσης, εναλλακτικά μπορούμε να υπολογίσουμε μία εκτιμήτρια της σωρευτικής συνάρτησης διακινδύνευσης. Έτσι, από τη σχέση (1.4) και την εκτιμήτρια της Kaplan-Meier έχουμε $\hat{H}(t)=-\log\left\{\hat{S}(t)\right\}$.

Συχνά, αντί για την Kaplan-Meier προτιμούμε την εκτιμήτρια Nelson-Aalen της συνάρτησης επιβίωσης, η οποία δίνεται από τις ισοδύναμες σχέσεις:

$$\hat{H}(t)=\begin{cases}\displaystyle\sum_{j:t_{(j)}\le t}\frac{d_j}{r_j}, \text{ με } t\ge t_{(1)}\\ 0, \quad \text{αλλιώς}\end{cases}, \tag{1.7}$$

και

$$\hat{S}(t)=\begin{cases}\displaystyle\prod_{j:t_{(j)}\le t}\exp\left(-\frac{d_j}{r_j}\right), \text{ με } t\ge t_{(1)}\\ 0, \quad \text{αλλιώς}\end{cases} \tag{1.8}$$

Η εκτιμήτρια αυτή είναι γνωστή και ως εκτιμήτρια Altshuler (Collett, 2003) και η αντίστοιχη εκτιμήτρια της διασποράς της Nelson-Aalen είναι η $\hat{\mathrm{V}}\left(\hat{H}(t)\right)=\sum_{j:\, t_{(j)}\le t}\frac{d_j}{r_j^{\,2}}$ (Καρώνη, 2009).

1.1.4 Έλεγχος υπόθεσης: Μη παραμετρική μέθοδος Log Rank.

Πολύ συχνά θέλουμε να εξετάσουμε τους χρόνους επιβίωσης που προέρχονται από δύο διαφορετικές ομάδες δεδομένων. Στη βιβλιογραφία υπάρχει ένας μεγάλος αριθμός μεθόδων που μπορούν να αναδείξουν το μέγεθος των διαφοροποιήσεων μεταξύ μονάδων. Η πιο απλή λύση είναι να κάνουμε ένα γράφημα των Kaplan-Meier εκτιμητριών των δύο ομάδων.

Ωστόσο, **αφού διατάξουμε τους χρόνους διακοπής σε αύξουσα σειρά,** ένας εναλλακτικός τρόπος είναι να προχωρήσουμε σε έναν έλεγχο υπόθεσης:

H_0: Δε διαφοροποιείται η συνάρτηση επιβίωσης μεταξύ των δύο ομάδων δεδομένων

H_1: Διαφοροποιείται η συνάρτηση επιβίωσης μεταξύ των δύο ομάδων δεδομένων

Για τον έλεγχο της παραπάνω υπόθεσης χρησιμοποιούμε την ελεγχοσυνάρτηση

$$W_L = \frac{U_L^2}{V_L},$$

η οποία είναι γνωστή με το όνομα **log rank** και ακολουθεί την κατανομή χ_1^2, καθώς η συνάρτηση $\frac{U_L}{\sqrt{V_L}} \sim \mathrm{N}(0,1)$, όπου:

- $U_L = \sum_{j=1}^{r}\left(d_{1j} - \frac{n_{1j}d_j}{n_j}\right)$, όπου d_{1j} ο αριθμός των μονάδων της πρώτης ομάδας που παύουν να λειτουργούν τη χρονική στιγμή $t_{(j)}$ και $\frac{n_{1j}d_j}{n_j}$ η συχνότητα των διακοπών τη στιγμή $t_{(j)}$, δεδομένου ότι η μονάδα αυτή θα προέρχεται από την πρώτη ομάδα.
- $V_L = Var(U_L) = \sum_{j=1}^{r}\frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}$.

Ο έλεγχος log rank αποτελεί ένα μη-παραμετρικό έλεγχο, καθώς στην προηγούμενη διαδικασία δεν προσδιορίζονται οι συναρτήσεις επιβίωσης των δύο ομάδων. Τέλος, να αναφέρουμε πως ο έλεγχος log rank μπορεί να επεκταθεί και για την περίπτωση όπου έχουμε να συγκρίνουμε περισσότερες από δύο ομάδες δεδομένων.

1.2 Το μοντέλο αναλογικής διακινδύνευσης (Proportional hazards model - PH)

Οι μη-παραμετρικές μέθοδοι που παρουσιάστηκαν στην προηγούμενη παράγραφο μπορεί να είναι αρκετά χρήσιμες για την ανάλυση ενός μόνο δείγματος δεδομένων, ή στην περίπτωση όπου έχουμε να συγκρίνουμε δύο ή περισσότερες ομάδες δεδομένων. Ωστόσο, στις περισσότερες εφαρμογές, έχουμε για κάθε στατιστική μονάδα πολλές βοηθητικές πληροφορίες, οι οποίες ενδεχομένως να επηρεάζουν την επιβίωση της. Παραδείγματος χάρη, σε ένα πείραμα για ασθενείς που λαμβάνουν κάποια αγωγή, η αντίδραση της κάθε μονάδας σε αυτήν την αγωγή μπορεί να εξαρτάται από την ηλικία, το φύλο, το βάρος, την ψυχική διάθεση ή ακόμα και τον τρόπο ζωής της μονάδας. Ένας log rank έλεγχος δεν είναι ικανός να συμπεριλάβει τις παραπάνω **επεξηγηματικές μεταβλητές** και να ανακαλύψει ποια επιδρά περισσότερο στην επιβίωση.

Για το λόγο αυτό, αναπτύσσουμε ειδικά στατιστικά μοντέλα προκειμένου να εξερευνήσουμε τη σχέση μεταξύ της επιβίωσης του ασθενή και των διαφόρων επεξηγηματικών μεταβλητών. Συνήθως, το ενδιαφέρον κεντρίζεται στον κίνδυνο του θανάτου οποιαδήποτε στιγμή μετά την εκκίνηση παρακολούθησης της θεραπείας.

Συνήθης μεθοδολογία για τις διάφορες εφαρμογές αποτελεί η μοντελοποίηση της συνάρτησης διακινδύνευσης. Συγκεκριμένα, μπορεί να μελετηθεί η επίδραση που έχει η αγωγή στον κίνδυνο του θανάτου, όπως και το μέγεθος στο οποίο άλλες επεξηγηματικές μεταβλητές επηρεάζουν τη συνάρτηση διακινδύνευσης. Ένας ακόμα λόγος για τη μοντελοποίηση της συνάρτησης διακινδύνευσης είναι να αποκτήσουμε μία εκτιμήτρια της συνάρτησης ξεχωριστά για την κάθε μονάδα του δείγματος, αποτέλεσμα που μπορεί να οδηγήσει μέσω της σχέσης (1.4) σε εκτιμήτρια για τη συνάρτηση επιβίωσης.

1.2.1 Η έννοια της αναλογικής διακινδύνευσης

Μία βασική υπόθεση, η οποία υποβόσκει πίσω από μία πληθώρα μεθόδων της Ανάλυσης Επιβίωσης, είναι η υπόθεση της αναλογικής διακινδύνευσης. Σύμφωνα με την υπόθεση αυτή,

ο κίνδυνος του θανάτου οποιαδήποτε στιγμή για μία μονάδα σε μία ομάδα δεδομένων, είναι ανάλογος του κινδύνου του θανάτου για μια διαφορετική μονάδα μιας άλλης ομάδας δεδομένων την ίδια χρονική στιγμή.

Σε περίπτωση που οι συναρτήσεις διακινδύνευσης των δύο ομάδων είναι ανάλογες, τότε οι γραφικές απεικονίσεις των συναρτήσεων επιβίωσης των μονάδων των δύο ομάδων δε διασταυρώνονται μεταξύ τους. Ειδικότερα, ας υποθέσουμε πως $h_1(t)$ είναι ο κίνδυνος του θανάτου τη χρονική στιγμή t για μία μονάδα της ομάδας 1 και $h_2(t)$ είναι ο κίνδυνος θανάτου την ίδια χρονική στιγμή για μία μονάδα της ομάδας 2. Αν οι δύο κίνδυνοι είναι ανάλογοι, τότε μπορούμε να γράψουμε $h_1(t) = y \cdot h_2(t)$, όπου y σταθερά (ανεξάρτητη του t). Ολοκληρώνοντας, έχουμε:

$$e^{-\int h_1(t)dt} = e^{-\int y h_2(t)dt} \overset{(1.4)}{\Rightarrow}$$

$$S_1(t) = \{S_2(t)\}^y$$

Η παραπάνω σχέση δείχνει πως η $S_1(t)$ είναι συνεχώς μεγαλύτερη ή μικρότερη της $S_2(t)$ (ανάλογα με την τιμή του y), καθώς οι τιμές της συνάρτησης επιβίωσης είναι μεταξύ 0 και 1. Ένα τέτοιο αποτέλεσμα σημαίνει πως τα γραφήματα των συναρτήσεων επιβίωσης δε διασταυρώνονται.

Η πιο γνωστή μορφή μοντέλου αναλογικής διακινδύνευσης παρουσιάστηκε το 1972 από τον Cox και είναι γνωστό στη βιβλιογραφία ως **μοντέλο παλινδρόμησης του Cox** (Cox regression model) ή **μοντέλο αναλογικής διακινδύνευσης του Cox** (Cox proportional hazards model – PH model).

1.2.2 Παρουσίαση του μοντέλου αναλογικής διακινδύνευσης του Cox

Έστω ότι εκτελούμε ένα πείραμα στο οποίο παίρνουν μέρος n μονάδες και αντιμετωπίζουμε την περίπτωση όπου ο κίνδυνος θανάτου σε μία συγκεκριμένη χρονική στιγμή εξαρτάται από τις τιμές $x_1, x_2, ..., x_k$, των k επεξηγηματικών μεταβλητών $X_1, X_2, ..., X_k$. Στο γενικό μοντέλο αναλογικής διακινδύνευσης, ας συμβολίσουμε το σύνολο των τιμών των επεξηγηματικών μεταβλητών με το διάνυσμα $\boldsymbol{x}$. Έπίσης, έστω ότι $h_0(t)$ είναι η συνάρτηση διακινδύνευσης για μία μονάδα για την οποία **οι τιμές των συμμεταβλητών είναι μηδέν**. Η συνάρτηση αυτή καλείται **βασική συνάρτηση διακινδύνευσης**. Η συνάρτηση διακινδύνευσης για την *i*-οστή μονάδα μπορεί τότε να γραφτεί ως:

$$h_i(t) = h_0(t)\psi(\boldsymbol{x_i}) \tag{1.9}$$

όπου $\psi(\boldsymbol{x_i})$ είναι μία συνάρτηση των τιμών του διανύσματος των επεξηγηματικών μεταβλητών για την *i*-οστή μονάδα. Η $\psi(\cdot)$ μπορεί να ερμηνευτεί ως η διακινδύνευση τη χρονική στιγμή *t* για μία μονάδα της οποίας το διάνυσμα των επεξηγηματικών μεταβλητών είναι $\boldsymbol{x_i}$, σε σχέση με τη διακινδύνευση για μία μονάδα με $\boldsymbol{x} = \boldsymbol{0}$.

Στο μοντέλο αναλογικής διακινδύνευσης του Cox η σχετική διακινδύνευση γράφεται ως $\psi(\boldsymbol{x_i}) = \exp(\eta_i)$, όπου $\eta_i = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} = \boldsymbol{\beta' x}$, με $\boldsymbol{\beta}$ το διάνυσμα των συντελεστών των τιμών $x_1, x_2, ..., x_k$ των επεξηγηματικών μεταβλητών του μοντέλου. Η ποσότητα η_i αποτελεί το γραμμικό μέρος του μοντέλου και είναι γνωστή ως **προγνωστικός δείκτης** (prognostic index) ή **σκορ κινδύνου** (risk score) για την *i*-οστή μονάδα (Collett, 2003). Στην περίπτωση αυτή, η συνάρτηση αναλογικής διακινδύνευσης γίνεται:

$$h_i(t) = h_0(t)\exp(\beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki}) \tag{1.10}$$

Η παραπάνω σχέση μπορεί να γραφτεί και ως:

$$\log\left\{\frac{h_i(t)}{h_0(t)}\right\} = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki},$$

ώστε το μοντέλο αναλογικής διακινδύνευσης να θεωρηθεί σταθερό με το χρόνο για το λογάριθμο του λόγου της διακινδύνευσης.

Όπως είδαμε και στην προηγούμενη παράγραφο, βασική ιδιότητα του μοντέλου του Cox, καθώς και των άλλων μοντέλων αναλογικής διακινδύνευσης είναι ότι ο λόγος των συναρτήσεων διακινδύνευσης δύο μονάδων παραμένει σταθερός ως προς το χρόνο, δηλαδή ισχύει ότι:

$$\frac{h(t,\boldsymbol{x_1})}{h(t,\boldsymbol{x_2})}=\frac{h_0(t)e^{\beta' x_1}}{h_0(t)e^{\beta' x_2}}=\frac{e^{\beta' x_1}}{e^{\beta' x_2}}.$$

Να παρατηρήσουμε ότι το μοντέλο του Cox είναι ένα ημι-παραμετρικό μοντέλο, καθώς δε χρειάζεται να γνωρίζουμε τη μορφή της $h_0(t)$ για τις διάφορες εφαρμογές.

Η προσαρμογή του μοντέλου του Cox που παρουσιάστηκε στη σχέση (1.10) με τη βοήθεια ενός συνόλου δεδομένων, προϋποθέτει την εκτίμηση των άγνωστων συντελεστών των μεταβλητών του γραμμικού μέρους του μοντέλου. Επιπρόσθετα, ίσως χρειάζεται και η εκτίμηση της βασικής συνάρτησης διακινδύνευσης. Αποδεικνύεται πως τα δύο αυτά μέλη μπορούν να εκτιμηθούν ανεξάρτητα (Collett, 2003). Οι Kalbfleisch και Prentice (1973) παρουσίασαν μία προσεγγιστική μέθοδο εκτίμησης της βασικής συνάρτησης διακινδύνευσης. Οι εκτιμήτριες των συντελεστών των μεταβλητών του μοντέλου μπορούν να προκύψουν με τη μέθοδο μεγιστοποίησης της πιθανοφάνειας (Cox, 1973). Τέλος, να παρατηρήσουμε πως η πλειοψηφία των στατιστικών πακέτων περιέχει τις μεθόδους εκτίμησης του μοντέλου αναλογικής διακινδύνευσης, ένας από τους λόγους που το μοντέλο αυτό είναι ιδιαίτερα διαδεδομένο στις διάφορες εφαρμογές που συναντώνται στη βιβλιογραφία.

1.2.3 Έλεγχοι υπόθεσης για τις συμμεταβλητές στο προσαρμοσμένο μοντέλο

Θεωρώντας το μοντέλο της προηγούμενης παραγράφου, υποθέσεις όπως η $H_0:\boldsymbol{\beta}=\boldsymbol{\beta}^{(0)}$ (γενική-καθολική υπόθεση), μπορούν να ελεγχθούν με τις παρακάτω τρεις μεθόδους:

Μεγιστοποίηση του λογαρίθμου της πιθανοφάνειας (Maximum likelihood ratio test)

Κατά τη διαδικασία υπολογισμού των εκτιμητριών μέγιστης πιθανοφάνειας αγνοούμε τη βασική συνάρτηση διακινδύνευσης. Για το λόγο αυτό, η πιθανοφάνεια που προκύπτει καλείται "**μερική**". Ο συγκεκριμένος τρόπος υπολογισμού των εκτιμητριών του μοντέλου παρουσιάστηκε για πρώτη φορά από τον Cox (1972). Οι Delong et al. (1994) παρουσίασαν μία ισοδύναμη έκφραση για την ακριβή μερική πιθανοφάνεια, για την περίπτωση που υπάρχουν ισοπαλίες (ties), με υπολογιστικά πλεονεκτήματα. Αν, δηλαδή, συμβολίσουμε με $l_0\left(=l(\boldsymbol{\beta}^{(0)})\right)$ το λογάριθμο της μερικής πιθανοφάνειας υπολογισμένο στη θέση $\boldsymbol{\beta}^{(0)}$ και $l_1\left(=l(\hat{\boldsymbol{\beta}})\right)$ τη μεγιστοποιημένη τιμή του λογαρίθμου της μερικής πιθανοφάνειας για το

μοντέλο υπολογισμένο στην τελική εκτιμήτρια του $\boldsymbol{\beta}$, τότε η τιμή της ελεγχοσυνάρτησης $-2\left(l(\boldsymbol{\beta}^{(0)}) - l(\hat{\boldsymbol{\beta}})\right)$ συγκρίνεται με την χ^2 κατανομή προς αποδοχή ή απόρριψη της μηδενικής υπόθεσης.

Έλεγχος Wald (Wald test)

Ο έλεγχος του Wald απαιτεί την προσαρμογή ενός μόνο μοντέλου για τις ίδιες υποθέσεις και είναι χρήσιμος για μία πρώτη ένδειξη των σημαντικών μεταβλητών σε ένα μοντέλο με πολλές συμμεταβλητές. Η ελεγχοσυνάρτηση του Wald για τη μηδενική υπόθεση είναι η $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)})'\hat{\mathrm{I}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)})$, όπου $\hat{\mathrm{I}} = \mathrm{I}(\hat{\boldsymbol{\beta}})$ είναι ο εκτιμημένος πίνακας πληροφορίας. Για την περίπτωση που το μοντέλο μας περιέχει μία μόνο μεταβλητή, η παραπάνω σχέση μετατρέπεται στη συνηθισμένη ελεγχοσυνάρτηση $\frac{\hat{\beta} - \beta^{(0)}}{se\left(\hat{\beta}\right)}$, η οποία συγκρίνεται με την κατανομή $\mathrm{N}(0,1)$.

Σκορ έλεγχος (Score test)

Η τιμή της ελεγχοσυνάρτησης για ένα σκορ έλεγχο βασίζεται στην ποσότητα $\frac{\partial l}{\partial \beta}$ για β βαθμωτό και $U(\boldsymbol{\beta}) = \left(\frac{\partial l}{\partial \beta_1}, \frac{\partial l}{\partial \beta_2}, ..., \frac{\partial l}{\partial \beta_{\mathrm{k}}}\right)'$ για το διάνυσμα $\boldsymbol{\beta}$. Κάτω από γενικές συνθήκες κανονικότητας, σε περίπτωση που η μηδενική υπόθεση είναι $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}^{(0)}$ με εναλλακτική την $H_1 : \boldsymbol{\beta} \in \Omega - \left\{\boldsymbol{\beta}^{(0)}\right\}$, η ελεγχοσυνάρτηση γράφεται ως $U'\left(\boldsymbol{\beta}^{(0)}\right) \mathrm{I}_{\boldsymbol{\beta}^{(0)}}{}^{-1} U\left(\boldsymbol{\beta}^{(0)}\right)$.

Η κατανομή υπό τη μηδενική υπόθεση για κάθε έναν από τους τρεις αυτούς ελέγχους είναι ασυμπτωτικά μία χ^2 κατανομή με ***k* βαθμούς ελευθερίας**. Επίσης, οι τιμές που προκύπτουν από τους παραπάνω ελέγχους είναι ασυμπτωτικά ισοδύναμες, παρόλο που σε άπειρα δείγματα μπορεί να διαφοροποιούνται. Σε μία τέτοια περίπτωση ο έλεγχος του λόγου των πιθανοφανειών θεωρείται ως ο πιο αξιόπιστος έλεγχος και ο έλεγχος του Wald ο λιγότερο αξιόπιστος (Therneau και Grambsch, 2000). Να αναφέρουμε ακόμα πως όταν $k = 1$ και η μοναδική συμμεταβλητή του μοντέλου είναι κατηγορική, τότε η τιμή του σκορ ελέγχου είναι ίδια με την τιμή του ελέγχου του Wald. Τέλος, οι Kalbfleisch και Prentice (2002) περιγράφουν αναλυτικά ότι ο έλεγχος σκορ ισοδυναμεί με τον έλεγχο log rank για την περίπτωση του μοντέλου του Cox.

1.2.4 Γραφικός έλεγχος καταλληλότητας του μοντέλου αναλογικής διακινδύνευσης

Ένας γραφικός τρόπος ελέγχου της καταλληλότητας της υπόθεσης αναλογικής διακινδύνευσης προκύπτει από τη γραφική παράσταση της συνάρτησης $\ln\left\{-\ln S(t;\boldsymbol{x})\right\}$ με το χρόνο, αφού πρώτα προσδιορίσουμε εκτιμήσεις $S(t;\boldsymbol{x})$ για επιλεγμένες τιμές των $\boldsymbol{x}$. Να σημειώσουμε εδώ ότι η εκτίμηση της συνάρτησης επιβίωσης $S(t;\boldsymbol{x})=\left\{S_0(t)\right\}^{e^{\boldsymbol{\beta}'\boldsymbol{x}}}$, απαιτεί την εκτίμηση της άγνωστης βασικής συνάρτησης επιβίωσης $S_0(t)$. Το επιθυμητό γράφημα, λοιπόν, μπορεί να επιτευχθεί μόνο με τη χρησιμοποίηση μίας εκτιμήτριας για την άγνωστη συνάρτηση επιβίωσης, όπως είναι και η εκτιμήτρια του Breslow: $\hat{S}_0(t)=e^{-\hat{H}_0(t)}$, όπου

$$\hat{H}_0(t)=\sum_{t_{(j)}\le t}\left(\frac{d_j}{\sum_{i\in r_j}e^{\hat{\boldsymbol{\beta}}'\boldsymbol{x}}}\right), \tag{1.11}$$

με r_j να είναι το σύνολο των μονάδων σε κίνδυνο τη χρονική στιγμή $t_{(j)}$. Πλεονέκτημα της θεωρίας αυτής αποτελεί το γεγονός ότι ισχύει για οποιοδήποτε μοντέλο αναλογικής διακινδύνευσης και όχι μόνο για το μοντέλο του Cox. Για το λόγο αυτό θα τη χρησιμοποιήσουμε και στο μοντέλο της παλινδρόμησης Κατωφλιού, στο Κεφάλαιο 3.

1.2.5 Έλεγχος καταλληλότητας του προσαρμοσμένου μοντέλου μέσω υπολοίπων

Ένας άλλος βασικός τρόπος ελέγχου της καταλληλότητας ενός στατιστικού μοντέλου είναι η εξέταση των υπολοίπων μετά από την προσαρμογή του μοντέλου. Τα υπόλοιπα είναι ποσότητες που μπορούν να υπολογιστούν για κάθε μονάδα που παίρνει μέρος σε μία μελέτη και δείχνουν κατά πόσο τα δεδομένα συμφωνούν με τις προϋποθέσεις και τις προβλέψεις του μοντέλου, όχι μόνο συνολικά αλλά και μεμονωμένα. Στο μοντέλο του Cox, ο έλεγχος καταλληλότητας του μοντέλου μέσω υπολοίπων είναι μία αρκετά περίπλοκη διαδικασία λόγω του γεγονότος ότι υπόλοιπα που είναι εύκολο να υπολογιστούν, όπως αυτά ενός μοντέλου πολλαπλής γραμμικής παλινδρόμησης, δεν είναι διαθέσιμα. Ωστόσο, υπάρχουν πολλές εναλλακτικές λύσεις.

Για τα επόμενα, θεωρούμε ότι γνωρίζουμε τους χρόνους επιβίωσης *n* παρατηρήσεων, όπου *r* από αυτούς είναι χρόνοι διακοπής και οι υπόλοιποι *n-r* αντιστοιχούν σε χρόνους αποκοπής. Επίσης, έστω ότι προσαρμόζουμε στα δεδομένα αυτά το μοντέλο παλινδρόμησης του Cox της Παραγράφου 1.2.1 με τις *k* επεξηγηματικές μεταβλητές $X_1, X_2, ..., X_k$. Η εκτιμώμενη συνάρτηση διακινδύνευσης για την *i* μονάδα, $i=1,2,...,n$ δίνεται από τον τύπο:

$$\hat{h}_i(t) = \hat{h}_0(t) e^{\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i},$$

όπου $\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i = \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + ... + \hat{\beta}_k x_{ki}$ και $\hat{h}_0(t)$ μία εκτίμηση της βασικής συνάρτησης διακινδύνευσης.

Ένα είδος υπολοίπων που είναι ιδιαίτερα γνωστό στην Ανάλυση Επιβίωσης είναι τα υπόλοιπα **Cox-Snell**. Η ονομασία τους προήλθε από το γεγονός πως αποτελούν ένα συγκεκριμένο παράδειγμα του γενικού ορισμού των υπολοίπων, όπως αυτός δόθηκε από τους Cox και Snell (1968). Ο τύπος που δίνει το Cox-Snell υπόλοιπο για την *i*-οστή μονάδα είναι ο:

$$r_{Ci} = \exp\left(\hat{\boldsymbol{\beta}}' \boldsymbol{x}_i\right) \hat{H}_0(t_{(i)}), \tag{1.12}$$

όπου $\hat{H}_0(t_{(i)})$ μία εκτίμηση της βασικής σωρευτικής συνάρτησης διακινδύνευσης τη χρονική στιγμή $t_{(i)}$, που είναι ο παρατηρούμενος χρόνος επιβίωσης της μονάδας. Όπως είδαμε στην Παράγραφο 1.2.2, η μορφή της $h_0(t)$ δεν καθορίζεται συνήθως για τις διάφορες εφαρμογές στο μοντέλο του Cox, με αποτέλεσμα τα συγκεκριμένα υπόλοιπα να μη δίνουν πολύ καλά αποτελέσματα. Τα υπόλοιπα αυτά χρησιμοποιούνται κυρίως σε παραμετρικά μοντέλα επιβίωσης, στα οποία η συνάρτηση διακινδύνευσης καθορίζεται πλήρως.

Στη συνέχεια, με τη βοήθεια των Cox-Snell υπολοίπων μπορούμε να ορίσουμε τα **martingale υπόλοιπα** (martingale residuals) ως εξής:

$$r_{Mi} = \delta_i - r_{Ci} \tag{1.13}$$

Η τιμή των martingale υπολοίπων είναι μεταξύ $-\infty$ και μονάδας με την τιμή των υπολοίπων για τις αποκομμένες παρατηρήσεις ($\delta_i = 0$) να είναι εξ ορισμού αρνητική. Αντίστοιχα, η τιμή τους για τις μη-αποκομμένες παρατηρήσεις επιτυγχάνεται για $\delta_i = 1$. Αποδεικνύεται πως τα υπόλοιπα αυτά αθροίζουν στο μηδέν και ότι σε μεγάλα δείγματα τα martingale υπόλοιπα είναι ασυσχέτιστα μεταξύ τους και έχουν αναμενόμενη τιμή ίση με μηδέν. Ένα πλήθος συγγραφέων έχει ασχοληθεί με martingale μεθόδους από τις οποίες προκύπτουν εναλλακτικά τα παραπάνω υπόλοιπα. Ενδεικτικά αναφέρουμε τους Andersen et al. (1993) και Therneau και Grambsch (2000).

Ένα μειονέκτημα των martingale υπολοίπων είναι ότι δεν είναι συμμετρικά κατανεμημένα γύρω από το μηδέν, ακόμα και στην περίπτωση που το προσαρμοσμένο μοντέλο είναι σωστό. Ένα τέτοιο γεγονός μετατρέπει τα υπόλοιπα αυτά σε εργαλεία δύσκολα στο χειρισμό, καθώς δεν μπορούμε με ευκολία να ερμηνεύσουμε γραφήματα βασισμένα σε αυτά. Για το λόγο αυτό οι Therneau et al. (1990) εισήγαγαν τα **martingale υπόλοιπα τύπου deviance**, τα οποία είναι περισσότερο συμμετρικά κατανεμημένα γύρω από το μηδέν και ορίζονται ως εξής:

$$r_{Di} = \text{sgn}(r_{Mi})\sqrt{-2\{r_{Mi} + \delta_i \log(\delta_i - r_{Mi})\}} \quad (1.14)$$

όπου sgn(x) είναι η συνάρτηση πρόσημο και r_{Mi} το αντίστοιχο martingale υπόλοιπο για την i μονάδα. Λόγω της ύπαρξης της συνάρτησης sgn(r_{Mi}) εξασφαλίζουμε ότι το martingale υπόλοιπο τύπου deviance για την i-οστή μονάδα θα έχει το ίδιο πρόσημο με το αντίστοιχο martingale υπόλοιπο.

Μεγάλο μειονέκτημα των υπολοίπων που παρουσιάσαμε παραπάνω αποτελεί η ισχυρή τους εξάρτηση από τον παρατηρούμενο χρόνο επιβίωσης. Επιπρόσθετα, για τον υπολογισμό τους απαιτείται μία εκτίμηση της σωρευτικής συνάρτησης διακινδύνευσης $H_0(t)$. Ο Schoenfeld (1982) κατόρθωσε να ξεπεράσει τα μειονεκτήματα αυτά με την εισαγωγή των **Schoenfeld υπολοίπων**. Τα υπόλοιπα αυτά δε δίνουν μία τιμή για κάθε μονάδα, αλλά ένα σύνολο τιμών, μία για κάθε επεξηγηματική μεταβλητή του προσαρμοσμένου μοντέλου παλινδρόμησης του Cox και αφορούν μόνο τις μη-αποκομμένες παρατηρήσεις, δηλαδή τους χρόνους διακοπής. Το i-οστό Schoenfeld υπόλοιπο για τη μεταβλητή X_j δίνεται από τον παρακάτω τύπο:

$$r_{Pji} = \delta_i\{x_{ji} - \hat{\alpha}_{ji}\} \quad (1.15)$$

όπου x_{ji} είναι η τιμή της j-οστής επεξηγηματικής μεταβλητής για την i-οστή μονάδα και $\hat{\alpha}_{ji} = \dfrac{\sum_{k\in R(t_i)} x_{jk}\exp(\hat{\beta}' x_k)}{\sum_{k\in R(t_i)}\exp(\hat{\beta}' x_k)}$, με $R(t_i)$ να είναι το σύνολο των μονάδων σε κίνδυνο τη χρονική στιγμή t_i. Τα υπόλοιπα αυτά λαμβάνουν μη-μηδενικές τιμές για τις μη-αποκομμένες παρατηρήσεις (δηλαδή για τους χρόνους διακοπής).

Το i-οστό Schoenfeld υπόλοιπο για την επεξηγηματική μεταβλητή X_j αποτελεί μία εκτίμηση της i-οστής συντεταγμένης της συνάρτησης μερικής πιθανοφάνειας

$$\frac{\partial \log L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n}\delta_i\{x_{ji} - \alpha_{ji}\}, \text{ με } \alpha_{ji} = \frac{\sum_k x_{jk}\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x_k})}{\sum_k \exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x_k})}.$$

Ο i-οστός όρος του παραπάνω αθροίσματος υπολογισμένος στο $\hat{\boldsymbol{\beta}}$ είναι το Schoenfeld υπόλοιπο για τη μεταβλητή X_j. Να παρατηρήσουμε ότι τα υπόλοιπα αυτά αθροίζουν στο μηδέν.

Οι Grambsch και Therneau (1994) εναλλακτικά πρότειναν τα **scaled Schoenfeld** υπόλοιπα. Έστω ότι $\boldsymbol{r_{pi}} = (r_{P1i}, r_{P2i}, ..., r_{Pki})'$ είναι το διάνυσμα με τα Schoenfeld υπόλοιπα για την i-οστή μονάδα. Τότε, τα scaled Schoenfeld υπόλοιπα r^*_{Pji} είναι οι συντεταγμένες του

διανύσματος $r_{Pji}^{*} = r\,\mathrm{var}\left(\hat{\boldsymbol{\beta}}\right) r_{Pi}$, όπου r είναι ο αριθμός των διακοπών μεταξύ των n μονάδων και $\mathrm{var}\left(\hat{\boldsymbol{\beta}}\right)$ είναι ο πίνακας διασποράς-συνδιασποράς των εκτιμητριών των παραμέτρων στο προσαρμοσμένο μοντέλο του Cox.

Μια ιδέα που προκύπτει άμεσα από τα παραπάνω είναι να σχεδιάσουμε τα γραφήματα των scaled Schoenfeld υπολοίπων για τις επεξηγηματικές μεταβλητές του μοντέλου. Τα scaled Schoenfeld υπόλοιπα είναι ιδιαίτερα σημαντικά στο να ελέγχουν την υπόθεση της αναλογικής διακινδύνευσης, ύστερα από την προσαρμογή στα δεδομένα του μοντέλου παλινδρόμησης του Cox. Οι Grambsch και Therneau (1994) δείξανε ότι η αναμενόμενη τιμή του i-οστού scaled Schoenfeld υπολοίπου, r_{Pji}^{*}, για την επεξηγηματική μεταβλητή X_j, δίνεται από τον τύπο $E\left(r_{Pji}^{*}\right) \approx \beta_j(t_i) - \hat{\beta}_j$, όπου $\beta_j(t)$ ένας χρονικά εξαρτώμενος συντελεστής για τη μεταβλητή X_j, $\beta_j(t_i)$ είναι η τιμή του συντελεστή κατά τη στιγμή του i-οστού χρόνου διακοπής και $\hat{\beta}_j$ είναι η εκτιμήτρια του συντελεστή της μεταβλητής X_j στο προσαρμοσμένο μοντέλο του Cox. Συνεπώς, ένα γράφημα των ποσοτήτων $r_{Pji}^{*} + \hat{\beta}_j$, σε σχέση με τους χρόνους διακοπής, θα πρέπει να μας δίνει πληροφορίες σχετικά με τη μορφή του εξαρτημένου από το χρόνο συντελεστή $\beta_j(t)$ της μεταβλητής X_j. Συγκεκριμένα, μία οριζόντια γραμμή θα μας υποδεικνύει ότι ο συντελεστής της X_j είναι σταθερός και σαν αποτέλεσμα ικανοποιείται η υπόθεση αναλογικής διακινδύνευσης.

1.2.6 Έλεγχος καταλληλότητας της συγκεκριμένης συναρτησιακής μορφής του προσαρμοσμένου μοντέλου μέσω υπολοίπων

Επίσης, θα ήταν σημαντικό να ελέγξουμε αν έχουμε υιοθετήσει την πιο κατάλληλη συναρτησιακή μορφή για αυτές τις μεταβλητές. Ένα ενδεχομένως καλύτερα προσαρμοσμένο μοντέλο μπορεί να εξασφαλιστεί με τη χρησιμοποίηση κάποιου μετασχηματισμού των τιμών μιας μεταβλητής (π.χ. $\log(x_j)$ στη θέση της x_j, ή κάποια μη-γραμμική συνάρτηση των χρόνων διακοπής), στη θέση των αρχικών τιμών αυτής της μεταβλητής.

Ένας πολύ καλός τρόπος ελέγχου αυτής της πλευράς της καταλληλότητας του μοντέλου μας, βασίζεται στα martingale υπόλοιπα που προκύπτουν από την προσαρμογή ενός μοντέλου που δεν περιέχει καμία συμμεταβλητή. Στη συνέχεια, τα υπόλοιπα αυτά παρουσιάζονται σε γραφήματα συναρτήσει των τιμών κάθε μίας από τις συμμεταβλητές του μοντέλου. Όπως έδειξαν οι Therneau et al. (1990), ένα τέτοιο γράφημα θα πρέπει να μας υποδεικνύει την απαιτούμενη συναρτησιακή μορφή της κάθε συμμεταβλητής. Συγκεκριμένα, το γράφημα μίας ευθείας γραμμής μας υποδεικνύει την ανάγκη χρησιμοποίησης ενός γραμμικού όρου για τη συμμεταβλητή. Για να προεκτείνουμε την ιδέα αυτή, από τη στιγμή

που ήδη θεωρούμε γνωστή τη μορφή των συμμεταβλητών του μοντέλου μας και θέλουμε να ελέγξουμε τη μορφή κάποιας από αυτές, μπορούμε εύκολα να υπολογίσουμε τα martingale υπόλοιπα από το προσαρμοσμένο μοντέλο του Cox που περιέχει όλες τις συμμεταβλητές εκτός από εκείνη που ελέγχουμε. Στη συνέχεια, μπορούμε να δημιουργήσουμε το γράφημα των martingale υπολοίπων συναρτήσει των τιμών της συμμεταβλητής της οποίας τη μορφή θέλουμε να ελέγξουμε.

1.2.7 Παραμετρικά μοντέλα αναλογικής διακινδύνευσης (parametric PH models)

Όταν σε μία εφαρμογή δεδομένων διάρκειας ζωής χρησιμοποιείται το μοντέλο της αναλογικής διακινδύνευσης του Cox, δεν υπάρχει καμία ανάγκη να υποθέσουμε κάποια συγκεκριμένη μορφή για τη συνάρτηση πυκνότητας πιθανότητας των χρόνων διακοπής. Σαν αποτέλεσμα, η συνάρτηση διακινδύνευσης δεν περιορίζεται σε κάποια συγκεκριμένη συναρτησιακή μορφή, με αποτέλεσμα το μοντέλο να έχει ευελιξία και μεγάλο πεδίο εφαρμογών.

Αντιθέτως, όταν η υπόθεση μίας συγκεκριμένης κατανομής για τους χρόνους διακοπής ευσταθεί, τότε τα συμπεράσματα που θα προκύψουν βασισμένα στην υπόθεση αυτή θα είναι ακόμα πιο ακριβή. Τα μοντέλα εκείνα, στα οποία υποθέτουμε εκ των προτέρων μία γνωστή κατανομή για τους χρόνους διακοπής, είναι γνωστά στη βιβλιογραφία ως **παραμετρικά μοντέλα**. Κατανομές που κατέχουν κεντρικό ρόλο στην ανάλυση επιβίωσης και ειδικότερα στα παραμετρικά PH μοντέλα είναι η κατανομή Weibull, η οποία παρουσιάστηκε για πρώτη φορά το 1951 από τον W. Weibull στο πλαίσιο πειραμάτων βιομηχανικής αξιοπιστίας, η Εκθετική, η Gompertz, κ.α. (Collett, 2003). Τέτοιες κατανομές οδηγούν σε συναρτήσεις διακινδύνευσης που αυξάνονται ή μειώνονται μονότονα (Johnson και Kotz, 1970).

1.3 Το μοντέλο της επιταχυνόμενης διακοπής (Accelerated Failure Time model - AFT / Accelerated Life model - AL)

Παρόλο που το μοντέλο αναλογικής διακινδύνευσης βρίσκει ένα τεράστιο εύρος εφαρμογών στην Ανάλυση Επιβίωσης, υπάρχουν σχετικά λίγες διαθέσιμες συναρτήσεις πυκνότητας πιθανότητας για τις διάρκειες ζωής που μπορούν να χρησιμοποιηθούν με το μοντέλο αυτό. Ένα μοντέλο που περικλείει ένα ευρύτερο φάσμα δυνατών επιλογών για την κατανομή της διάρκειας ζωής, χωρίς περιορισμό ως προς τη μονοτονία της συνάρτησης διακινδύνευσης είναι το μοντέλο **Επιταχυνόμενης Διακοπής** (Accelerated Life model - AL).

Το AL είναι ένα γενικό μοντέλο διάρκειας ζωής, στο οποίο οι επεξηγηματικές μεταβλητές για την κάθε μονάδα θεωρείται πως δρουν πολλαπλασιαστικά στην κλίμακα του χρόνου. Επιπρόσθετα, αποτελεί ιδανικό εργαλείο για τις διάφορες εφαρμογές κατά τις οποίες η συνάρτηση διακινδύνευσης αλλάζει μονοτονία στο πέρασμα του χρόνου. Παραδείγματος

χάρη, οι πρώτες 10 περίπου ημέρες της μετεγχειρητικής πορείας της υγείας ασθενή που έχει υποστεί μεταμόσχευση καρδιάς θεωρούνται ως περίοδος αυξανόμενου κινδύνου θανάτου. Ωστόσο, όσο ο ασθενής απομακρύνεται χρονικά από τη στιγμή του χειρουργείου και το σώμα υιοθετεί το νέο όργανο, ο κίνδυνος του θανάτου συνεχώς απομακρύνεται. Εφαρμογές του μοντέλου επιταχυνόμενης διακοπής περιγράφονται στους Crowder et al. (1991).

1.3.1 Παρουσίαση του μοντέλου για την περίπτωση δύο ομάδων δεδομένων

Έστω πως ασθενείς είναι χωρισμένοι κατά τυχαίο τρόπο σε δύο ομάδες προκειμένου να τους χορηγηθεί μία αγωγή. Οι ασθενείς της πρώτης ομάδας (ομάδα ελέγχου) λαμβάνουν την υπάρχουσα αγωγή S, ενώ οι ασθενείς της δεύτερης ομάδας (ομάδα θεραπείας) λαμβάνουν μία νέα πειραματική θεραπεία N. Χρησιμοποιώντας το θεωρητικό υπόβαθρο των μοντέλων επιταχυνόμενης διακοπής, θεωρούμε πως η διάρκεια ζωής μίας μονάδας που έλαβε τη νέα θεραπεία είναι κάποιο πολλαπλάσιο της διάρκειας ζωής μίας μονάδας της ομάδας ελέγχου. Με αυτόν τον τρόπο, το αποτέλεσμα της νέας θεραπείας μπορεί να ερμηνευτεί ως **επιτάχυνση** ή **επιβράδυνση** του χρόνου.

Κάτω από αυτήν την υπόθεση, η πιθανότητα μία μονάδα που λαμβάνει τη νέα θεραπεία να ζήσει πέραν της χρονικής στιγμής *t*, ισούται με την πιθανότητα μία μονάδα που λαμβάνει την υπάρχουσα θεραπεία να ζήσει πέραν της χρονικής στιγμής t/φ, όπου φ άγνωστη σταθερά.

Έστω $S_S(t)$ και $S_N(t)$ οι συναρτήσεις επιβίωσης για τους ασθενείς των δύο ομάδων δεδομένων. Τότε το μοντέλο επιταχυνόμενης διακοπής επιβάλλει:

$$S_N(t) = S_S(t/\varphi), \text{ για οποιοδήποτε } t.$$

Μία ερμηνεία του παραπάνω μοντέλου είναι πως η διάρκεια ζωής της μονάδας που λαμβάνει τη νέα θεραπεία είναι φ φορές η διάρκεια ζωής που θα είχε η μονάδα εάν λάμβανε την υπάρχουσα θεραπεία. Η παράμετρος φ, λοιπόν, αναπαριστά την επίδραση της αγωγής στη βασική κλίμακα χρόνου. Όταν το σημείο του ενδιαφέροντος είναι ο θάνατος του ασθενή, τιμές του φ μικρότερες της μονάδας ισοδυναμούν με επιτάχυνση του χρόνου μέχρι το θάνατο της μονάδας που έλαβε τη νέα θεραπεία σε σχέση με μία μονάδα που έλαβε την κλασική θεραπεία. Όταν το σημείο του ενδιαφέροντος είναι η ανάρρωση ύστερα από κάποια περίοδο ασθένειας, τιμές του φ μικρότερες της μονάδας συναντώνται στην περίπτωση που η επίδραση της νέας θεραπείας ισοδυναμεί με επιτάχυνση του χρόνου ανάρρωσης. Κάτω από αυτές τις συνθήκες, η νέα θεραπεία θα είναι καλύτερη της κλασικής. Για το λόγο αυτό, ο παράγοντας φ^{-1} ορίζεται ως **παράγοντας επιτάχυνσης.**

Παρατήρηση

Αποδεικνύεται πως το *p*-οστό ποσοστιαίο σημείο της συνάρτησης επιβίωσης για έναν ασθενή που λαμβάνει τη νέα θεραπεία $t_N(p)$, είναι τέτοιο ώστε $t_N(p)=\varphi\cdot t_S(p)$, όπου $t_S(p)$ το αντίστοιχο ποσοστιαίο σημείο για έναν ασθενή που λαμβάνει την κλασική αγωγή.

Χρησιμοποιώντας τις εξισώσεις (1.1) και (1.2), προκύπτουν οι σχέσεις μεταξύ των συναρτήσεων πυκνότητας πιθανότητας και των συναρτήσεων διακινδύνευσης για τις δύο ομάδες δεδομένων:

$$f_N(t)=\varphi^{-1}f_S(t/\varphi)$$

$$h_N(t)=\varphi^{-1}h_S(t/\varphi)$$

Έστω ακόμα πως *X* είναι μία δείκτρια μεταβλητή με:

$$X=\begin{cases}0, & \text{για τους ασθενείς που λαμβάνουν την κλασική αγωγή}\\ 1, & \text{για τους ασθενείς που λαμβάνουν τη νέα αγωγή}\end{cases}$$

Η συνάρτηση διακινδύνευσης για τον *i*-οστό ασθενή γράφεται:

$$h_i(t)=\varphi^{-x_i}h_0\left(t/\varphi^{x_i}\right), \tag{1.16}$$

όπου x_i η αριθμητική τιμή της X για την i μονάδα. Θέτοντας $x_i=0$ στην (1.16), παρατηρούμε πως $h_0(t)$ είναι η συνάρτηση διακινδύνευσης για έναν ασθενή που λαμβάνει την κλασική θεραπεία. Όπως στο μοντέλο του Cox, έτσι και εδώ η $h_0(t)$ καλείται **βασική συνάρτηση διακινδύνευσης**. Σε αυτήν την περίπτωση, η συνάρτηση διακινδύνευσης για έναν ασθενή της ομάδας ελέγχου είναι:

$$h_S(t)=\varphi^{-1}h_0(t/\varphi)$$

Η παράμετρος φ πρέπει να είναι μη-αρνητική. Επομένως είναι βολικό να θέσουμε $\varphi=e^{\beta}$. Το μοντέλο επιταχυνόμενης διακοπής της σχέσης (1.16) γίνεται:

$$h_i(t)=e^{-\beta x_i}h_0\left(t/e^{\beta x_i}\right) \tag{1.17}$$

Η συνάρτηση διακινδύνευσης για τη μονάδα που λαμβάνει τη νέα θεραπεία είναι πλέον:

$$h_S(t)=e^{-\beta}h_0\left(t/e^{\beta}\right).$$

1.3.3 Το γενικό μοντέλο επιταχυνόμενης διακοπής

Το μοντέλο επιταχυνόμενης διακοπής της σχέσης (1.17) μπορεί να γενικευθεί για την περίπτωση που υπάρχουν *k* επεξηγηματικές μεταβλητές για την κάθε μονάδα. Τότε, η συνάρτηση διακινδύνευσης $h_i(t)$ για την *i*-οστή μονάδα τη χρονική στιγμή *t* είναι τέτοια ώστε:

$$h_i(t) = e^{-n_i} h_0\left(t/e^{n_i}\right), \tag{1.18}$$

όπου $n_i = \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} = \boldsymbol{\beta' x}$, με $\boldsymbol{\beta}$ το διάνυσμα των συντελεστών των τιμών $x_1, x_2, ..., x_k$ των επεξηγηματικών μεταβλητών του μοντέλου. Αντίστοιχα με το μοντέλο αναλογικής διακινδύνευσης, η βασική συνάρτηση διακινδύνευσης είναι ο κίνδυνος του θανάτου τη χρονική στιγμή *t* της μονάδας για την οποία οι τιμές των *k* επεξηγηματικών μεταβλητών είναι ίσες με το μηδέν. Η αντίστοιχη συνάρτηση επιβίωσης είναι:

$$S_i(t) = S_0\left(t/\exp(n_i)\right),$$

με $S_0(t)$ να είναι η βασική συνάρτηση επιβίωσης.

1.3.4 Η λογαριθμο-γραμμική μορφή του μοντέλου επιταχυνόμενης διακοπής

Ας θεωρήσουμε ένα λογαριθμο-γραμμικό μοντέλο για την τυχαία μεταβλητή T_i που σχετίζεται με τη ζωή της *i*-μονάδας, σύμφωνα με το οποίο:

$$\log T_i = \mu + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \sigma\varepsilon_i = \mu + \boldsymbol{\beta' x} + \sigma\varepsilon_i, \tag{1.19}$$

με $\boldsymbol{\beta}$ το διάνυσμα των συντελεστών των τιμών $x_1, x_2, ..., x_k$ των επεξηγηματικών μεταβλητών του μοντέλου και μ, σ δύο επιπλέον παράμετροι, γνωστές ως σταθερά και παράμετρος κλίμακας, αντίστοιχα. Η ποσότητα ε_i είναι μία τυχαία μεταβλητή που χρησιμοποιείται για να μοντελοποιήσει την απόκλιση των τιμών της $\log T_i$ από το γραμμικό μέρος του μοντέλου και θεωρείται πως έχει μία γνωστή κατανομή πιθανότητας. Στην παραπάνω μοντελοποίηση το διάνυσμα $\boldsymbol{\beta}$ αναπαριστά την επίδραση που έχει η κάθε επεξηγηματική μεταβλητή στις διάρκειες ζωής. Θετικές τιμές υποδεικνύουν πως οι διάρκειες ζωής αυξάνονται με την αύξηση των επεξηγηματικών μεταβλητών και αντίστροφα.

Για να δείξουμε τη σχέση μεταξύ αυτής της μοντελοποίησης και της σχέσης (1.18), ας θεωρήσουμε τη συνάρτηση επιβίωσης της T_i :

$$S_i(t) = P(T_i \geq t) = P\left(\exp\left(\mu + \boldsymbol{\beta' x} + \sigma\varepsilon_i\right) \geq t\right) = P\left(\exp\left(\mu + \sigma\varepsilon_i\right) \geq t/\exp\left(\boldsymbol{\beta' x}\right)\right)$$

Επειδή η $S_0(t)$ είναι η βασική συνάρτηση επιβίωσης μίας μονάδας με $\boldsymbol{x} = \mathbf{0}$, έπεται ότι:

$$S_0(t) = P\left(\exp(\mu + \sigma\varepsilon_i) \geq t\right),$$

οπότε καταλήγουμε στην

$$S_i(t) = S_0\left(t/\exp(\boldsymbol{\beta}'\boldsymbol{x})\right),$$

που είναι η γενική μορφή της συνάρτησης επιβίωσης για την *i*-μονάδα στο μοντέλο επιταχυνόμενης διακοπής, με **παράγοντα επιτάχυνσης** την ποσότητα $\exp(-\boldsymbol{\beta}'\boldsymbol{x})$. Λογαριθμίζοντας την παραπάνω σχέση και πολλαπλασιάζοντας την με (-1), με τη βοήθεια της σχέσης:

$$h(t) = -\frac{d}{dt}\log\left(S(t)\right)$$

προκύπτει η αντίστοιχη σχέση (1.18) ανάμεσα στις συναρτήσεις διακινδύνευσης με $\eta_i = \boldsymbol{\beta}'\boldsymbol{x}$.

1.3.5 Παραμετρικά AFT μοντέλα (parametric AFT models)

Αντίστοιχα με την περίπτωση των παραμετρικών μοντέλων αναλογικής διακινδύνευσης, στη βιβλιογραφία υπάρχει πληθώρα κατανομών για τους χρόνους διακοπής και για την περίπτωση ενός AFT μοντέλου. Κατανομές που κατέχουν κεντρικό ρόλο στην Ανάλυση Επιβίωσης και ειδικότερα στα παραμετρικά AFT μοντέλα, είναι η κατανομή Weibull, η log-logistic και η log-normal (Collett, 2003).

1.4 Το μοντέλο των αναλογικών λόγων συμπληρωματικών πιθανοτήτων (Proportional Odds model - PO)

Βασική ιδιότητα των μοντέλων αναλογικής διακινδύνευσης είναι ότι ο λόγος των συναρτήσεων διακινδύνευσης δύο μονάδων παραμένει σταθερός ως προς το χρόνο. Ωστόσο, όπως είδαμε και στην προηγούμενη παράγραφο, στις ιατρικές εφαρμογές είναι σύνηθες φαινόμενο οι συναρτήσεις διακινδύνευσης των ασθενών δύο ή περισσότερων ομάδων δεδομένων να συγκλίνουν με το χρόνο. Για παράδειγμα, σε μία μελέτη μακροχρόνιας παρακολούθησης της πορείας της υγείας ασθενών σε μία κλινική δοκιμή, η επίδραση της αγωγής στην επιβίωση ή στο αρχικό στάδιο της ασθένειας μπορεί να εξαφανιστεί. Παρόμοια, σε εφαρμογές όπου μία ομάδα ασθενών που πάσχουν από μία συγκεκριμένη ασθένεια συγκρίνονται με μία ομάδα υγιών ανθρώπων (ομάδα ελέγχου-control group), μία αποτελεσματική θεραπεία της ασθένειας θα οδηγούσε σε μία εξομοίωση της επιβίωσης των δύο ομάδων δεδομένων με το χρόνο.

Το μοντέλο που θα παρουσιαστεί σε αυτήν την παράγραφο εμφανίζει πολλές ομοιότητες (από άποψη δομής) με το μοντέλο αναλογικής διακινδύνευσης του Cox και

μπορεί να χρησιμοποιηθεί για την επίλυση παρόμοιων προβλημάτων. Σε αντίθεση με το μοντέλο του Cox, οι ρυθμοί διακινδύνευσης για διαφορετικές ομάδες ασθενών συγκλίνουν με το χρόνο. Αυτή η έννοια μπορεί να είναι αρκετά πιο χρήσιμη σε σχέση με ένα σταθερό λόγο συναρτήσεων διακινδύνευσης για την περίπτωση όπου αρχικά αποτελέσματα όπως διαφοροποιήσεις σε επίπεδο ασθένειας ή θεραπείας, εξαφανίζονται με την πάροδο του χρόνου. Τέλος, το θεωρητικό υπόβαθρο του μοντέλου υποδεικνύει πως οι διάφοροι προγνωστικοί παράγοντες έχουν πολλαπλασιαστική επίδραση στις πιθανότητες θανάτου κάτω από οποιαδήποτε χρονική περίοδο.

1.4.1 Εισαγωγή - ιστορική αναδρομή

Η έννοια των αναλογικών συμπληρωματικών πιθανοτήτων είναι γνωστή από επιδημιολογικές μελέτες ως ένα μέτρο του κατά προσέγγιση σχετικού κινδύνου ενός γεγονότος, όπως είναι η πραγματοποίηση ασθένειας ή θανάτου που συμβαίνει κατά την απουσία ή παρουσία ενός παράγοντα ενδιαφέροντος. Ωστόσο, είναι μία απλή ελεγχοσυνάρτηση που υπολογίζεται από ένα 2×2 πίνακα ενσωματώνοντας την κατάσταση των υποκειμένων της μελέτης στην κατάληξη της έρευνας. Σε μία μελέτη με μεγάλη περίοδο παρακολούθησης, είναι απαραίτητο ένα μέτρο του κινδύνου το οποίο θα συγκεντρώνει πληροφορία σχετικά με τις διάρκειες ζωής όλων των ασθενών από την είσοδο στη μελέτη και όχι απλά την πληροφορία για το αν ζούνε ή πεθάνανε.

Ακολουθώντας το έργο του Snell, ο Clayton (1974) ανέπτυξε μία εμπειρική λύση για το παραπάνω πρόβλημα. Σε μία επόμενη εργασία του (Clayton, 1976), παρουσίασε τη λύση για αποκομμένα δεδομένα, τα οποία εμφανίζονται συχνά στην Ανάλυση Επιβίωσης. Ο McCullagh (1980) γενίκευσε την ιδέα ενός σταθερού λόγου συμπληρωματικών πιθανοτήτων για την περίπτωση που υπάρχουν παραπάνω από δύο δείγματα, με τη βοήθεια ενός μοντέλου παλινδρόμησης, το οποίο ονόμασε μοντέλο **αναλογικών λόγων συμπληρωματικών πιθανοτήτων** (Proportional Odds model) και το οποίο προσάρμοσε με τη μέθοδο της μέγιστης πιθανοφάνειας.

Ο Bennett (1983a) προσάρμοσε το μοντέλο που πρότεινε ο McCullagh για την περίπτωση ενός ιατρικού προβλήματος και έδωσε εκτιμήτριες μέγιστης πιθανοφάνειας για τη συνεχή περίπτωση παρατηρήσεων με αποκομμένα δεδομένα. Παραμετρικές παραλλαγές του μοντέλου μπορούν να παραχθούν χρησιμοποιώντας τη log-logistic κατανομή για τα δεδομένα διάρκειας ζωής (Bennett, 1983b), γεγονός που μετατρέπει το μοντέλο σε ιδιαίτερα ελκυστικό για τις διάφορες εφαρμογές, καθώς η κατανομή αυτή είναι η μοναδική που μοιράζεται ταυτόχρονα τις ιδιότητες της επιταχυνόμενης διακοπής και των αναλογικών λόγων συμπληρωματικών πιθανοτήτων. Τέλος, περαιτέρω ανάπτυξη του μοντέλου έγινε από τους Bennett και Whitehead (1981), ενώ εφαρμογές περιγράφονται στους Crowder et al. (1991).

1.4.2 Παρουσίαση του μοντέλου

Ο λόγος των αναλογικών συμπληρωματικών πιθανοτήτων (proportional odds), θ, σε ένα ένα 2×2 πίνακα ορίζεται ως εξής:

$$\theta = \frac{p_1(1-p_2)}{p_2(1-p_1)}$$

όπου p_i, $i=1, 2$ είναι η αναλογία των παρατηρήσεων που αντιμετωπίζουν το γεγονός στην ομάδα i. Οι πιθανότητες πραγματοποίησης του γεγονότος για μία μονάδα στην ομάδα i είναι $\frac{p_i}{(1-p_i)}$. Γενικεύοντας το γεγονός σε "αποτυχία πραγματοποιείται μέχρι τη χρονική στιγμή t" για όλα τα $t>0$, ο ορισμός γίνεται:

$$\theta = \frac{F_1(t)(1-F_2(t))}{F_2(t)(1-F_1(t))}, \tag{1.20}$$

όπου $F_i(t)$ είναι η συνάρτηση κατανομής της πιθανότητας πραγματοποίησης των γεγονότων στην ομάδα i και $\frac{F_i(t)}{1-F_i(t)}$ είναι οι πιθανότητες αποτυχίας μέχρι τη χρονική στιγμή t. Στην περίπτωση που το συμβάν είναι αποτυχία, τότε $S_i(t)=1-F_i(t)$ είναι η συνάρτηση επιβίωσης για την ομάδα i. Η εξίσωση της (1.20) γίνεται:

$$\frac{F_1(t)}{(1-F_1(t))} = \frac{F_2(t)}{(1-F_2(t))}\theta$$

και αναπαριστά το λόγο των αναλογικών συμπληρωματικών πιθανοτήτων πραγματοποίησης του γεγονότος έως τη χρονική στιγμή t. Λογαριθμίζοντας την παραπάνω εξίσωση, έχουμε:

$$\log\left(\frac{F_1(t)}{(1-F_1(t))}\right) - \log\left(\frac{F_2(t)}{(1-F_2(t))}\right) = \log(\theta),$$

από την οποία προκύπτει πως η διαφορά στις συναρτήσεις των λογαρίθμων των συμπληρωματικών πιθανοτήτων μεταξύ των ομάδων 1 και 2 είναι σταθερή στο χρόνο. Ένα γράφημα μιας εμπειρικής εκτίμησης του λογαρίθμου των πιθανοτήτων με το χρόνο μπορεί να χρησιμοποιηθεί ως ένας πρωταρχικός έλεγχος της καταλληλότητας του μοντέλου των αναλογικών συμπληρωματικών πιθανοτήτων (Kaplan και Meier, 1958 και Altshuler, 1970).

Ο λόγος των συναρτήσεων διακινδύνευσης των ομάδων 1 και 2 συγκλίνει μονότονα από το θ στη μονάδα με το πέρασμα του χρόνου, ενώ ο λόγος των συναρτήσεων επιβίωσης των 2 ομάδων αποκλίνει από τη μονάδα στο θ με την πάροδο του χρόνου (Bennett, 1983a).

Όταν υπάρχουν παραπάνω από δύο ομάδες μονάδων, ή όταν πολλές διαφορετικές συμμεταβλητές μετρώνται για την κάθε μονάδα, μπορούμε να θεωρήσουμε ένα μοντέλο παλινδρόμησης συμπληρωματικών πιθανοτήτων θέτοντας απλά μία ξεχωριστή παράμετρο θ_i για τον *i*-οστό ασθενή και θέτοντας $\theta_i = \exp(\boldsymbol{\beta}'\boldsymbol{x_i})$, όπου $\boldsymbol{x_i} = (x_{i1}, ..., x_{ip})'$ το διάνυσμα των συμμεταβλητών για τον *i*-οστό ασθενή και $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ άγνωστες παράμετροι. Το μοντέλο γίνεται:

$$\frac{F(t;\theta_i)}{1-F(t;\theta_i)} = \frac{F_0(t)}{1-F_0(t)}\theta_i \tag{1.21}$$

όπου $F_0(t)$ είναι μία υποβόσκουσα άγνωστη συνάρτηση κατανομής πιθανότητας. Η πιθανοφάνεια ενός συνόλου U μίας ομάδας μη αποκομμένων δεδομένων, και ενός συνόλου C μίας ομάδας αποκομμένων παρατηρήσεων δίνεται από τη σχέση:

$$\prod_{i\in U} f(t_i;\theta_i)\prod_{i\in C}\left(1-F(t_i;\theta_i)\right)$$

όπου t_i είναι η στιγμή της αποτυχίας ή αποκοπής για την *i*-οστή μονάδα και *f* είναι η κατάλληλη συνάρτηση πυκνότητας πιθανότητας.

Για την προσαρμογή του μοντέλου των αναλογικών λόγων συμπληρωματικών πιθανοτήτων είναι απαραίτητη η εκτίμηση των αγνώστων παραμέτρων και της υποβόσκουσας συνάρτησης F_0. Η χρησιμοποιούμενη μέθοδος απαιτεί ένα μετασχηματισμό των χρόνων διακοπής στη log-logistic κατανομή, που περιγράφεται από τον Bennett (1983b). Τέλος, η εκτίμηση των παραμέτρων του μοντέλου πραγματοποιείται με τη μέθοδο μέγιστης πιθανοφάνειας χρησιμοποιώντας αριθμητικές μεθόδους τύπου Newton-Raphson.

Ωστόσο, υπάρχουν δύο κυρίως λόγοι για τους οποίους το παρόν μοντέλο δε χρησιμοποιείται συχνά στις διάφορες εφαρμογές. Ο πρώτος αφορά την απουσία του από την πλειοψηφία των ευρύτερα διαδεδομένων στατιστικών πακέτων που χρησιμοποιούνται στις εφαρμογές. Ένας δεύτερος λόγος είναι πως το μοντέλο αυτό δίνει συνήθως παρόμοια αποτελέσματα με το μοντέλο αναλογικής διακινδύνευσης του Cox, στο οποίο υπάρχει μία χρονικά μεταβαλλόμενη μεταβλητή που παράγει μη αναλογικούς κινδύνους (Collett, 2003).

1.6 Η παλινδρόμηση Κατωφλιού για την Ανάλυση Επιβίωσης: Μοντελοποίηση χρόνων διακοπής μίας στοχαστικής ανέλιξης που φτάνει σε κάποιο σύνορο

Στις προηγούμενες παραγράφους, παρουσιάσαμε τα κατεξοχήν μοντέλα που χρησιμοποιούνται για την ανάλυση δεδομένων διάρκειας ζωής και περιγράψαμε διαφορετικές μεθοδολογικές προσεγγίσεις στην Ανάλυση Επιβίωσης προβλημάτων με συμμεταβλητές. Μελετήσαμε το παραμετρικό μοντέλο επιταχυνόμενης ζωής (AL), το οποίο εμφανίζεται κυρίως στις τεχνολογικές επιστήμες και το μοντέλο αναλογικής διακινδύνευσης (PH), που χρησιμοποιείται συνήθως στις ιατρικές επιστήμες. Δυστυχώς, η πιο συνηθισμένη του μορφή που είναι το μοντέλο του Cox, εφαρμόζεται αυτομάτως στις περισσότερες περιπτώσεις χωρίς την απαιτούμενη λεπτομερή εξέταση της ισχύος των προϋποθέσεών του. Σαν αποτέλεσμα, πιθανώς να υπάρχουν πάρα πολλές εσφαλμένες εφαρμογές, στις οποίες δεν ισχύει η δομή της αναλογικής διακινδύνευσης. Τέλος, παρουσιάσαμε το μοντέλο αναλογικών λόγων συμπληρωματικών πιθανοτήτων (PO), το οποίο μαζί με το AL μοντέλο χρησιμοποιούνται σχετικά σπάνια στις διάφορες εφαρμογές, εν μέρει επειδή απουσιάζουν από τα γνωστά στατιστικά πακέτα.

Σε αυτά τα προβλήματα υπάρχει μεγάλη ανάγκη βελτίωσης της μεθοδολογίας, με την ανάπτυξη:

- ✓ μοντέλων με διαφορετικές δομές,
- ✓ τεχνικών εξέτασης της καταλληλότητας των μοντέλων αυτών
- ✓ εύχρηστων υπολογιστικών μεθόδων.

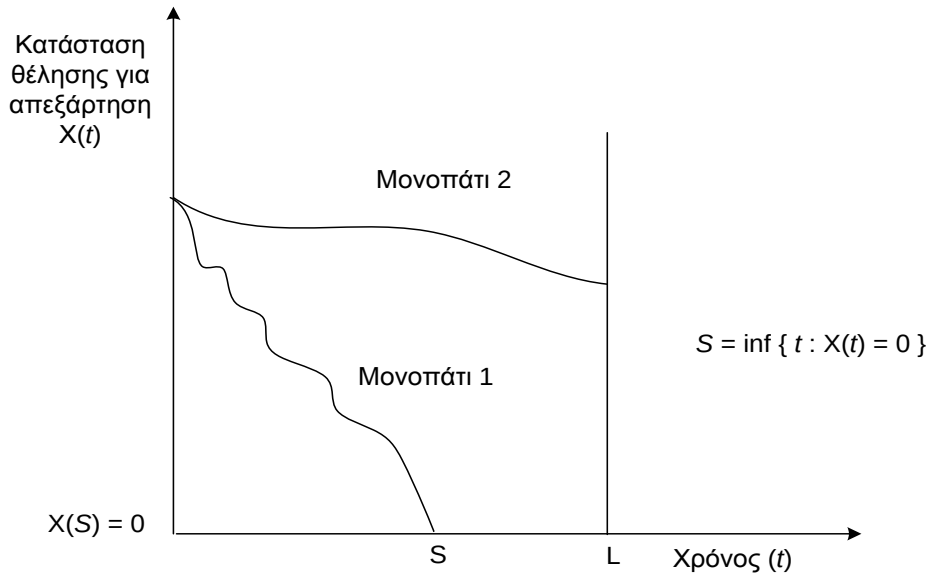
Επιπρόσθετα, η πλειοψηφία των μοντέλων που χρησιμοποιούνται για την περιγραφή δεδομένων διάρκειας ζωής έχει αναπτυχθεί κάτω από την υπόθεση πως το περιβάλλον δράσης των υπό μελέτη φαινομένων είναι **στατικό**. Ωστόσο, τα τελευταία χρόνια αποδεικνύεται πως τα μοντέλα αυτά αφορούν ειδικές περιπτώσεις γενικότερων μοντέλων. Τα νέα αυτά μοντέλα προσπαθούν να συμπεριλάβουν και να περιγράψουν τους διάφορους λανθάνοντες μηχανισμούς που ευθύνονται για την πορεία των υπό μελέτη φαινομένων και χρησιμοποιούν κατάλληλες στοχαστικές ανελίξεις ως εργαλεία για την περιγραφή των διαφόρων φαινομένων. Ο Singpurwalla (1995) παρουσιάζει μία αναλυτική καταγραφή - χαρτογράφηση όλων των διαφορετικών μοντέλων ως προς τις στρατηγικές που χρησιμοποιούνται για την ανάπτυξή τους.

1.6.1 Μοντέλα χρόνων πρώτης μετάβασης (First Hitting Time models - FHT models)

Εναλλακτικά με τον τρόπο μοντελοποίησης που παρουσιάσαμε στις προηγούμενες παραγράφους της διατριβής, πολλά είδη διάρκειας ζωής, ή διάρκειας μέχρι να συμβεί κάποιο γεγονός, μπορεί να ερμηνευτούν ως **χρόνοι πρώτης μετάβασης** (First Hitting Times) μιας

στοχαστικής ανέλιξης η οποία φτάνει για πρώτη φορά σε κάποιο σύνορο ή περνάει από κάποιο κατώφλι. Η στοχαστική αυτή ανέλιξη μπορεί να είναι είτε λανθάνουσα (π.χ. η μελέτη της πορείας της ψυχικής βούλησης ηρωινομανών προς απεξάρτηση, Caplehorn και Bell (1991), Σχήμα 1.1), είτε παρατηρήσιμη (π.χ. η καθημερινή ένδειξη του πυρετού ενός αρρώστου μέχρι να γιατρευτεί).

Τα **μοντέλα χρόνων πρώτης μετάβασης** (FHT models) έχουν μια μακρά ιστορία σε πολλά διαφορετικά πεδία εφαρμογών, τα οποία περιλαμβάνουν μεταξύ άλλων τη Φαρμακευτική, την Οικολογία, τη Μηχανική, τα Οικονομικά και την Κοινωνιολογία. Στη βιβλιογραφία, υπάρχει πληθώρα άρθρων που ασχολούνται με αυτά τα μοντέλα και τις εφαρμογές τους. Οι Eaton και Whitmore (1977) για παράδειγμα, εξετάζουν το FHT ως γενικό μοντέλο για την περιγραφή της παραμονής ασθενών σε νοσοκομείο. Οι Aalen και Gjessing (2001) παρέχουν μία επισκόπηση του θέματος. Παρόμοια, ο Lawless (2003) δίνει μία πλήρη περίληψη της θεωρίας, των μοντέλων και των μεθόδων. Ακόμα, οι Lee και Whitmore (2006) παρουσιάζουν, επίσης, μία επισκόπηση των FHT μοντέλων για δεδομένα διάρκειας ζωής.



Σχήμα 1.1: Πορεία της ποσότητας "θέληση για απεξάρτηση" δύο διαφορετικών ηρωινομανών της μελέτης των Caplehorn και Bell (1991)

Τα FHT μοντέλα είναι κατάλληλα για να περιγράψουν προβλήματα όπως ο χρόνος επιβίωσης ασθενή ύστερα από κάποια μεταμόσχευση, η διάρκεια μιας απεργίας (Lancaster, 1972), η επιθετική πορεία κάποιας μορφής καρκίνου που προκλήθηκε από έκθεση σε καρκινογόνα υλικά λόγω επαγγελματικής ιδιότητας (Hazelton et al., 2001, Lee et al., 2004), ο χρόνος επιβίωσης μιας επιχείρησης, η διάρκεια ενός γάμου, ακόμα και η μεταβατική περίοδος για την αλλαγή της τιμής ενός εμπορεύματος (Whitmore, 1979). Τα μοντέλα αυτά υιοθετούνται με αυξανόμενο ρυθμό από τους επιστήμονες για να περιγράψουν διάφορα

προβλήματα λόγω του ρεαλισμού και της ικανότητας εφαρμογής τους. Πρόσφατα, το γενικό ενδιαφέρον για τα μοντέλα αυτά έχει αυξηθεί και νέες περιοχές εφαρμογών έχουν εντοπιστεί.

1.6.2 Συστατικά ενός μοντέλου χρόνου πρώτης μετάβασης

Ένα μοντέλο χρόνου πρώτης μετάβασης αποτελείται από δύο βασικά συστατικά:

1. Μία **γονική στοχαστική ανέλιξη** (parent stochastic process) $\{X(t), t \in \mathcal{T}, x \in \mathcal{X}\}$ με αρχική τιμή $X(0) = x_0$, όπου $\mathcal{T}$ είναι ο παραμετρικός χώρος και $\mathcal{X}$ ο χώρος καταστάσεων της ανέλιξης.
2. Ένα **σύνορο *B***, όπου $B \subset \mathcal{X}$. Στα επόμενα, θα αναφερόμαστε στο σύνορο B ως σύνορο, φράγμα ή κατώφλι, ανάλογα με το γενικό πλαίσιο του κάθε προβλήματος.

Βασικό χαρακτηριστικό ενός FHT μοντέλου αποτελεί ο διαχωρισμός σε παρατηρήσιμο ή μη-παρατηρήσιμο μονοπάτι για τη στοχαστική ανέλιξη. Η πιο συνηθισμένη περίπτωση ανελίξεων που συναντάμε σε προβλήματα της καθημερινότητας, όπως π.χ. η πορεία της υγείας ενός ασθενή, είναι η κατηγορία των μη-παρατηρήσιμων ανελίξεων. Τέλος, ακόμα και το ίδιο το σύνορο μπορεί να έχει πολλά διαφορετικά χαρακτηριστικά. Παράδειγμα με παρατηρήσιμη ανέλιξη παρουσιάζεται στους Lee, DeGruttola και Schoenfeld (2000), οι οποίοι χρησιμοποιούν τη μέτρηση των κυττάρων CD4 ως ένδειξη για την πορεία της υγείας ασθενών με AIDS.

Αν συμβολίσουμε με $X(0) = x_0$ την αρχική τιμή της στοχαστικής ανέλιξης, και αν θεωρήσουμε ότι βρίσκεται έξω από το σύνορο B, τότε ο χρόνος πρώτης μετάβασης στο B είναι η τυχαία μεταβλητή S, η οποία ορίζεται ως εξής:

$$S = \inf\{t : X(t) \in B\} \tag{1.22}$$

Έτσι λοιπόν, ο χρόνος πρώτης μετάβασης είναι η χρονική στιγμή που η στοχαστική ανέλιξη εισέρχεται μέσα στο σύνορο B **για πρώτη φορά**. Στη βιβλιογραφία, η κατάσταση κατά την οποία η ανέλιξη εισέρχεται για πρώτη φορά μέσα στο σύνορο ($X(S) \in B$) αναφέρεται ως **κατάσταση κατωφλιού** (threshold state). Το σύνορο, λοιπόν, καθορίζει μία κατάσταση διακοπής για την ανέλιξη με τη συνηθισμένη έννοια του όρου που συναντάει κανείς στη θεωρία των στοχαστικών ανελίξεων. Στην περίπτωση που η γονική στοχαστική ανέλιξη είναι λανθάνουσα (π.χ. επίπεδα ψυχικής βούλησης αθλητών κατά τη διάρκεια του τελευταίου μήνα πριν από την Ολυμπιάδα), δεν υπάρχει κάποιος ευθύς τρόπος για να παρατηρήσουμε το χρόνο πρώτης μετάβασης της ανέλιξης, με αποτέλεσμα να χρειαζόμαστε παραπάνω εργαλεία, ώστε να μοντελοποιήσουμε ένα τέτοιο πρόβλημα.

Τα FHT μοντέλα εμφανίζονται με απόλυτα φυσικό τρόπο σε πολλούς τύπους στοχαστικών ανελίξεων, από την ανέλιξη Wiener έως τις Μαρκοβιανές αλυσίδες. Η

ουσιαστική δομή ενός FHT μοντέλου, για παράδειγμα, για μία στοχαστική ανέλιξη τύπου Wiener μπορεί να εξηγηθεί αναλυτικά και από το Σχήμα 1.1. Το συγκεκριμένο πρόβλημα αποτελεί ένα παράδειγμα δεδομένων διάρκειας ζωής στο οποίο ενδιαφερόμαστε για το χρόνο μέχρι να προκύψει το γεγονός "εγκατάλειψη της προσπάθειας του ηρωινομανή για απεξάρτηση". Υπάρχει μία ισχυρή πεποίθηση ότι πίσω από τη χρονική διάρκεια της πορείας της απεξάρτησης υποβόσκει η μη-μετρήσιμη ποσότητα "θέληση για απεξάρτηση". Έστω πως κατά τη χρονική στιγμή $t=0$ οι δύο ηρωινομανείς έχουν το ίδιο επίπεδο θέλησης για απεξάρτηση, $X(0)=x_0$. Παρατηρούμε πως η πορεία του επιπέδου της θέλησης για απεξάρτηση αλλάζει με την πάροδο του χρόνου με διαφορετικό τρόπο για τους δύο ηρωινομανείς. Ο ασθενής του πρώτου μονοπατιού φτάνει σε μηδενικό επίπεδο θέλησης για απεξάρτηση και εγκαταλείπει την προσπάθεια τη χρονική στιγμή *S*, κατά την οποία το πρώτο μονοπάτι της ανέλιξης "χτυπά" το σύνορο $\left(X(S)=0\right)$. Από την άλλη, ο ασθενής του δεύτερου μονοπατιού, ακόμα και μετά από χρόνο $L>S$ μέσα στην κλινική απεξάρτησης (τέλος περιόδου παρατήρησης) διατηρεί υψηλά επίπεδα θέλησης για απεξάρτηση. Η περίπτωση που μελετάμε εξετάζει την κατάσταση της υγείας ενός ασθενούς ως μία λανθάνουσα στοχαστική ανέλιξη $\{X(t)\}$.

Υπάρχουν περιπτώσεις μοντέλων χρόνων πρώτης μετάβασης, στις οποίες δεν υπάρχει καμία εγγύηση ότι η ανέλιξη $\{X(t)\}$ θα καταφέρει να φτάσει στο σύνορο *B*. Σε μία τέτοια περίπτωση ισχύει ότι $\mathrm{P}(S<\infty)<1$. Στα επόμενα, με $S=\infty$ θα εννοούμε την απουσία ενός πεπερασμένου χρόνου πρώτης μετάβασης με $\mathrm{P}(S=\infty)=1-\mathrm{P}(S<\infty)$. Τέλος, το βασικό μοντέλο χρόνου πρώτης μετάβασης της σχέσης (1.22) προϋποθέτει ένα σταθερό στο χρόνο σύνορο *B*. Ωστόσο, σε κάποιες εφαρμογές διαφοροποιείται με το χρόνο $\left(B(t)\right)$. Η διαφοροποίηση αυτή μπορεί να είναι είτε ντετερμινιστική είτε να ακολουθεί με τη σειρά της κάποια στοχαστική ανέλιξη.

1.6.3 Αντίστροφη Γκαουσιανή κατανομή (inverse Gaussian distribution) - Ιστορική αναδρομή

Στη διατριβή, θα μας απασχολήσουν δομές παλινδρόμησης για τα μοντέλα χρόνων πρώτης μετάβασης, στην οποία θα αναφερόμαστε και ως **παλινδρόμηση Κατωφλιού** (Threshold Regression), ή TR για συντομία. Η πιο γνωστή περίπτωση του TR μοντέλου αφορά δεδομένα που περιγράφονται από μία στοχαστική ανέλιξη τύπου Wiener. Στην περίπτωση αυτή, αποδεικνύεται (βλέπε Παράγραφο 2.1) πως ο χρόνος πρώτης μετάβασης ακολουθεί την **αντίστροφη Γκαουσιανή κατανομή** (inverse Gaussian distribution - IG). Σε αυτήν την παράγραφο παρουσιάζουμε κάποια ιστορικά στοιχεία για την κατανομή, ενώ αναλυτική παρουσίαση, καθώς και ιδιότητες δίνονται στο δεύτερο κεφάλαιο.

Το 1900 ο Bachelier κατέληξε στην Κανονική κατανομή προσπαθώντας να περιγράψει την κίνηση ενός κόκκου σιταριού, όταν πραγματοποιεί μονοδιάστατη κίνηση Brown. Η δουλεία του έχει πολλές ομοιότητες με τη σύγχρονη θεωρία της κίνησης Brown, ωστόσο έλειπε ένα μέτρο του μονοπατιού του χώρου των καταστάσεων, κάτι που προμήθευσε ο Wiener το 1923.

Η πρώτη ενασχόληση με την κατανομή του χρόνου πρώτης μετάβασης ανιχνεύεται στο έργο των Tweedie (1941) και Wald (1944). Ο Tweedie παρατήρησε την αντίστροφη σχέση που υπήρχε μεταξύ της γεννήτριας συνάρτησης αθροιστικών του χρόνου που χρειάζεται να καλυφθεί κάποιο διάστημα και της γεννήτριας συνάρτησης αθροιστικών του διαστήματος μετρημένης σε μονάδες χρόνου. Επίσης, ο Tweedie (1945) παρατήρησε αυτό το είδος σχέσης μεταξύ της Διωνυμικής και της αντίστροφης Διωνυμικής κατανομής, αλλά και μεταξύ της Poisson και της Εκθετικής κατανομής. Αργότερα (1957a), χρησιμοποίησε την ονομασία "αντίστροφη Γκαουσιανή" για την κατανομή του χρόνου πρώτης μετάβασης της κίνησης Brown. Εξέδωσε μία πολύ αναλυτική μελέτη πάνω στην κατανομή (1957a), στην οποία περιγράφονται πολλές από τις σημαντικές στατιστικές της ιδιότητες. Νωρίτερα, η αντίστροφη Γκαουσιανή κατανομή απασχόλησε φυσικούς όπως οι Einstein και Schrödinger λόγω της κίνησης Brown. Οι Chhikara και Folks (1977), πρότειναν να ονομαστεί "κατανομή του Tweedie" ως ένδειξη αναγνώρισης του έργου του. Πίνακες με ποσοστιαία σημεία της κατανομής δόθηκαν το 1988 στο εγχειρίδιο Χημείας και Φυσικής, CRC (Chemical Rubber Company). Ο Wald (1947), έδωσε μία ειδική περίπτωση της κατανομής. Τέλος, να αναφέρουμε πως η κατανομή συχνά συναντάται με το όνομα "κατανομή του Wald", ειδικά στην ρωσική βιβλιογραφία.

Ο Wasan (1969), έγραψε μία ανασκόπηση του πρότερου έργου πάνω στην κατανομή και ασχολήθηκε με συγκεκριμένες οριακές καταστάσεις και χαρακτηριστικά της κατανομής. Οι Johnson και Kotz (1970), έγραψαν μία περίληψη για τις μέχρι τότε γνωστές στατιστικές της ιδιότητες. Οι Chhikara και Folks (1977), μελέτησαν τις ομοιότητες των στατιστικών ιδιοτήτων της κατανομής με αυτές της Κανονικής κατανομής. Επιπρόσθετα ασχολήθηκαν με πολλές ενδιαφέρουσες στατιστικές μεθόδους βασισμένες στην IG κατανομή και τις παρουσίασαν μαζί με ιδιότητες και εφαρμογές της κατανομής σε μία εκτενή μελέτη (Chhikara και Folks, 1989). Παρουσίαση στατιστικών μεθόδων συμπερασματολογίας για παραμετρικά μοντέλα με την αντίστροφη Γκαουσιανή κατανομή δίνει και ο Lawless (2003). Τέλος, ο Seshadri (1993) έγραψε μία αναλυτική παρουσίαση της εκθετικής οικογένειας κατανομών κατά την οποία εστιάζει ειδικά στην αντίστροφη Γκαουσιανή κατανομή.

Πρόσφατα, η IG κατανομή έχει αποκτήσει έναν αρκετά ευέλικτο ρόλο σε μοντέλα στοχαστικών ανελίξεων συμπεριλαμβανομένης της θεωρίας των γενικευμένων γραμμικών μοντέλων (McCullagh και Nelder, 1983), της ανάλυσης αξιοπιστίας, της ανάλυσης δεδομένων διάρκειας ζωής (Padgett και Tsai, 1986), της θεωρίας μοντέλων επιταχυνόμενης

ζωής (Bhattacharyya και Fries, 1982), καθώς και σε περιπτώσεις κατανομών επιδιορθώσιμων συστημάτων (Chhikara και Folks, 1977). Τέλος, οι Hougaard (1984) και Feaganes and Suchindran (1991) ενθαρρύνουν τη χρήση της IG ως κατανομής της ευπάθειας (frailty). Τα τελευταία 20 χρόνια έχει γίνει ένας αρκετά μεγάλος αριθμός δημοσιεύσεων πάνω στην περιοχή.

1.7 Σκοπός της διατριβής

Βασικός στόχος της διδακτορικής διατριβής είναι η συμβολή στην περαιτέρω ανάπτυξη του θεωρητικού υποβάθρου της παλινδρόμησης Κατωφλιού. Ως σχετικά καινούργιο μοντέλο, απαιτεί πολύ ανάπτυξη ακόμα. Είναι σημαντικό να καταφέρουμε να ελέγξουμε την καταλληλότητα του προσαρμοσμένου μοντέλου, καθώς στόχος της μοντελοποίησης είναι η εύρεση εκείνου του μοντέλου, το οποίο επιτυγχάνει να εξηγήσει με τη μεγαλύτερη δυνατή ακρίβεια τα δεδομένα μας και να μας βοηθήσει να βγάλουμε χρήσιμα συμπεράσματα. Για την παλινδρόμηση Κατωφλιού όμως, δεν έχουν ακόμα αναπτυχθεί ούτε τεχνικές ελέγχου της καταλληλότητας του προτεινόμενου μοντέλου, ούτε διαγνωστικοί έλεγχοι. Επιθυμητή είναι και η επέκταση του μοντέλου για την περίπτωση επαναλαμβανόμενων γεγονότων στην ίδια μονάδα, η οποία απαιτείται για την εφαρμογή του μοντέλου στη μελέτη επισκευάσιμων συστημάτων και αλλού. Η παρούσα διδακτορική διατριβή συμβάλλει σε αυτή την επιστημονική εξέλιξη με την ανάπτυξη:

- τεχνικών για τον εντοπισμό άτυπων τιμών των παραμέτρων της κατανομής του χρόνου πρώτης μετάβασης στην παλινδρόμηση Κατωφλιού.
- του μοντέλου για την περίπτωση των επαναλαμβανόμενων δεδομένων.
- διαγνωστικών τεχνικών και ελέγχων για την καταλληλότητα του μοντέλου.
- μεθόδων επιλογής μεταβλητών.
- προγραμμάτων για την πρακτική εφαρμογή των παραπάνω τεχνικών και μεθόδων, ιδιαίτερα στη γλώσσα R.

1.8 Περιγραφή των κεφαλαίων της διδακτορικής διατριβής

Η διδακτορική διατριβή χωρίζεται σε πέντε ενότητες-κεφάλαια. Συγκεκριμένα, στο 1^{ο} κεφάλαιο δόθηκαν κάποιες εισαγωγικές έννοιες της Ανάλυσης Επιβίωσης και παρουσιάστηκαν συνοπτικά τα γνωστά μοντέλα PH, AL και PO, τα οποία χρησιμοποιούνται στις διάφορες εφαρμογές. Στη συνέχεια, έγινε μία περιγραφή των FHT μοντέλων και δόθηκε η έννοια της παλινδρόμησης Κατωφλιού. Στα κεφάλαια που ακολουθούν, δίνεται έμφαση στην παρουσίαση των ευρημάτων σε μορφή που διευκολύνει τους χρήστες πακέτων προγραμ-

ματιστικού ή μη, περιβάλλοντος, καθώς και στην περιγραφή όλων των επιμέρους διαδικασιών με τη βοήθεια ψευδοκώδικα.

Στο 2^{ο} κεφάλαιο παρουσιάζεται η αντίστροφη Γκαουσιανή κατανομή (IG distribution). Δίνονται γενικές ιδιότητες της κατανομής που συμπεριλαμβάνουν την IG ως γενικευμένο γραμμικό μοντέλο. Παρουσιάζονται υπόλοιπα για την κατανομή. Έλεγχοι υπόθεσης και έλεγχοι εντοπισμού άτυπων τιμών αναπτύσσονται για τις παραμέτρους της κατανομής. Η ιδέα της άτυπης τιμής επεκτείνεται στην περίπτωση της άτυπης μονάδας με επαναλαμβανόμενα γεγονότα (Karioti και Caroni, 2002). Τέλος, οι διάφοροι έλεγχοι παρουσιάζονται για μία εναλλακτική παραμέτρηση της κατανομής.

Στο Κεφάλαιο 3, αντιμετωπίζουμε την IG κατανομή ως μοντέλο της παλινδρόμησης Κατωφλιού. Το FHTR μοντέλο συγκρίνεται με το μοντέλο της αναλογικής διακινδύνευσης του Cox. Αναπτύσσονται διαγνωστικές τεχνικές για την καταλληλότητα του μοντέλου και γίνεται διερεύνηση πρακτικών θεμάτων στην προσαρμογή ενός IG FHTR μοντέλου. Παρουσιάζονται αποτελέσματα προσομοιώσεων. Τέλος, αναπτύσσεται μία τεχνική επιλογής μεταβλητών (variable selection).

Στο Κεφάλαιο 4, αναπτύσσεται μία τεχνική εντοπισμού σημείων επιρροής (influence) για την περίπτωση του IG FHTR μοντέλου, με τη βοήθεια της απόστασης των πιθανοφανειών. Εξετάζεται το θέμα της τοπικής επιρροής. Τα διάφορα θεωρητικά αποτελέσματα ελέγχονται με τη βοήθεια προσομοιώσεων. Τέλος, παρουσιάζεται και μία εφαρμογή των διαφόρων μέτρων επιρροής σε δεδομένα πραγματικών συνθηκών.

Το Κεφάλαιο 5 περιέχει συμπεράσματα και συνολικά σχόλια πάνω στη διατριβή και παρουσιάζει νέα πεδία έρευνας και ιδέες για μελλοντική δουλειά, τόσο πάνω στην IG κατανομή όσο και στην παλινδρόμηση Κατωφλιού. Τέλος, παρουσιάζονται επεκτάσεις για την περίπτωση των επαναλαμβανομένων γεγονότων για τη μονάδα.

Κεφάλαιο 2

Αντίστροφη Γκαουσιανή κατανομή

2.1 Εισαγωγή

Η επιλογή κατανομής για την περιγραφή δεδομένων συχνά γίνεται υπό το πρίσμα του πόσο καλά τα ίδια τα δεδομένα φαίνεται να προσαρμόζονται από τη συγκεκριμένη κατανομή. Αντιθέτως, στη Στατιστική μοντελοποίηση, η επιλογή της κατανομής γίνεται υπό το πρίσμα του πόσο καλά έχουμε καταφέρει να εξηγήσουμε το μηχανισμό της αποτυχίας ή επιτυχίας ενός φαινομένου. Παραδείγματος χάρη, είναι λογικό να θεωρήσουμε έναν αυξανόμενο ρυθμό αποτυχίας για την αναπαράσταση της διάρκειας ζωής σε μία κατάσταση που κυρίως σχετίζεται με τη γήρανση ή τη διαδικασία φθοράς. Ωστόσο, μία πρόωρη αποτυχία μπορεί να προκληθεί από πληθώρα άλλων λόγων, όπως τεχνολογικό ελάττωμα, λάθος χρήση, στιγμιαίος τραυματισμός κ.α.. Για το λόγο αυτό, είναι επιθυμητό να χρησιμοποιήσουμε τη φυσική περιγραφή ενός φαινομένου προκειμένου να επιλέξουμε την κατανομή που θα το περιγράψει.

Περιπτώσεις δεδομένων στα οποία συμβαίνουν συχνά πρόωρα γεγονότα όπως θάνατος ή επιδιόρθωση αποτελούν συνηθισμένο φαινόμενο σε δεδομένα διάρκειας ζωής. Μία κατανομή που είναι κατάλληλη για την περιγραφή δεδομένων με τέτοια συμπεριφορά είναι η αντίστροφη Γκαουσιανή κατανομή.

Στην παράγραφο αυτή, περιγράφεται η γέννηση της αντίστροφης Γκαουσιανής κατανομής ως κατανομής του χρόνου πρώτης μετάβασης σε ανέλιξη Wiener. Δίνονται αναλογίες της IG με την Κανονική κατανομή. Τέλος, γίνεται αναφορά σε εφαρμογές της κατανομής (βλέπε Παράγραφο 2.1.3).

2.1.1 Γέννηση της IG

Το βασικό χαρακτηριστικό των ανελίξεων με διακριτό χώρο καταστάσεων σε συνεχή χρόνο (ανέλιξη Poisson) είναι ότι σε ένα μικρό χρονικό διάστημα είτε παρατηρείται ριζική αλλαγή είτε δεν παρατηρείται καμία αλλαγή της κατάστασης της ανέλιξης. Πολύ συχνά στις εφαρμογές προκύπτει η ανάγκη να μελετήσουμε ανελίξεις με συνεχή χώρο καταστάσεων. Μέσα σε ένα μικρό χρονικό διάστημα μία τέτοια ανέλιξη μπορεί να υποστεί μικρές μόνο μεταβολές της κατάστασής της. Οι αναπαραστάσεις τέτοιων ανελίξεων αναμένεται να είναι συνεχείς συναρτήσεις. Παράδειγμα μίας τέτοιας ανέλιξης αποτελούν μόρια που αποβάλλονται μέσα σε κάποιο υγρό και κινούνται λόγω των γρήγορων, τυχαίων, βίαιων και διαδοχικών συγκρούσεων με τα γειτονικά μόρια. Αν προχωρούσαμε σε γράφημα της αλλαγής της πορείας ενός τέτοιου μορίου με το χρόνο θα περιμέναμε ένα συνεχές, ακανόνιστο και αλλοπρόσαλλο γράφημα, το οποίο θα αποτελούσε στην πραγματικότητα την αναπαράσταση μίας στοχαστικής ανέλιξης σε συνεχή χρόνο με συνεχή χώρο καταστάσεων. Το φυσικό αυτό φαινόμενο είναι γνωστό ως **κίνηση Brown** από τον βοτανολόγο Robert Brown, που το παρατήρησε για πρώτη φορά το 1827.

Στον απλό τυχαίο περίπατο, μεταβιβάσεις ενός βήματος επιτρέπεται να καταλήγουν μόνο στις κοντινότερες γειτονικές καταστάσεις. Τέτοιου είδους τοπικές αλλαγές μπορεί να θεωρηθούν ως το ανάλογο των διακριτών καταστάσεων του φαινομένου των συνεχών αλλαγών για την περίπτωση του συνεχούς χώρου καταστάσεων. Έτσι, αν θεωρήσουμε μικρά βήματα ενός μεγέθους Δ να συμβαίνουν σε μικρά χρονικά διαστήματα μεγέθους τ, τότε παίρνοντας τα όρια καθώς Δ και τ πηγαίνουν στο μηδέν, περιμένουμε να πάρουμε μία ανέλιξη της οποίας οι αναπαραστάσεις είναι συνεχείς συναρτήσεις του χρόνου, όπως περιγράφεται στους Cox και Miller (1965).

Έστω, λοιπόν, ένα μόριο ξεκινάει από την αρχή των αξόνων. Σε κάθε διάστημα τ κάνει ένα βήμα $Z(i)$ με πιθανότητες $\mathrm{P}(Z(i)=+\Delta)=p$ και $\mathrm{P}(Z(i)=-\Delta)=q=1-p$. Να παρατηρήσουμε ότι όλα τα βήματα είναι μεταξύ τους ανεξάρτητα. Η ροπογεννήτρια συνάρτηση της τυχαίας μεταβλητής Z για ένα βήμα είναι:

$$E\left(e^{-\theta z}\right)=pe^{-\theta\Delta}+qe^{\theta\Delta} \tag{2.1}$$

Στο χρόνο t πραγματοποιούνται $n=\frac{t}{\tau}$ βήματα και η τελική κατάσταση μετά από τα n αυτά βήματα, $X(t)$, είναι το άθροισμα n ανεξάρτητων τυχαίων μεταβλητών $Z(i)$, $i=1,2,...,n$, κάθε μία από τις οποίες έχει ροπογεννήτρια συνάρτηση που δίνεται από τον τύπο (2.1). Έτσι,

$X(t)=\sum_{i=1}^{n}Z(i)=X\left((n-1)\cdot\tau\right)+Z(n)$, με $X(0)=x_0$, $n=1,2,...$ και

$$E\left(e^{-\theta X(t)}\right)=E\left(e^{-\theta\sum_{i=1}^{n}Z_i}\right)=E\left(\prod_{i=1}^{n}e^{-\theta Z_i}\right)\overset{\text{ανεξαρτησία}}{=}\prod_{i=1}^{n}\left\{E\left(e^{-\theta Z_i}\right)\right\}=$$

$$=\prod_{i=1}^{n}\left\{E\left(e^{-\theta Z_i}\right)\right\}=\left\{E\left(e^{-\theta Z_i}\right)\right\}^n=\left(pe^{-\theta\Delta}+qe^{\theta\Delta}\right)^n=\left(pe^{-\theta\Delta}+qe^{\theta\Delta}\right)^{\frac{t}{\tau}}.$$

Επομένως,

$$E\left(e^{-\theta X(t)}\right)=\left(pe^{-\theta\Delta}+qe^{\theta\Delta}\right)^{\frac{t}{\tau}}$$

Τέλος, η μέση τιμή και η διασπορά της $X(t)$ είναι:

$$\mathrm{E}\left[X(t)\right]=(t/\tau)(p-q)\Delta \text{ και } V\left[X(t)\right]=(t/\tau)4pq\Delta^2.$$

Θέλουμε τώρα $\Delta\to 0$ και $\tau\to 0$ έτσι ώστε η μέση τιμή της ανέλιξης να είναι μ και η διασπορά της σ^2.

Αντικαθιστώντας, λοιπόν, στην παραπάνω σχέση τις ποσότητες $\Delta=\sigma\sqrt{\tau}$, $p=\frac{1}{2}\left(1+\frac{\mu\sqrt{\tau}}{\sigma}\right)$ και $q=\frac{1}{2}\left(1-\frac{\mu\sqrt{\tau}}{\sigma}\right)$ παίρνουμε ότι η ροπογεννήτρια συνάρτηση της $X(t)$ δίνεται από τον παρακάτω τύπο:

$$\mathrm{E}\left\{e^{-\theta X(t)}\right\}=\left[\frac{1}{2}\left(1+\frac{\mu\sqrt{\tau}}{\sigma}\right)e^{-\frac{\theta\sqrt{\tau}}{\sigma}}+\frac{1}{2}\left(1-\frac{\mu\sqrt{\tau}}{\sigma}\right)e^{\frac{\theta\sqrt{\tau}}{\sigma}}\right]^{\frac{t}{\tau}} \qquad (2.2)$$

Παίρνοντας τα όρια για $\tau\to 0$ και λογαριθμίζοντας τη σχέση (2.2), καταλήγουμε στην παρακάτω μορφή για τη **γεννήτρια συνάρτηση αθροιστικών** (γ.σ.α.-cumulant generating function):

$$\text{γ.σ.α.}=\pi(\theta;t)=\left[\mu\theta+\frac{\sigma^2\theta^2}{2}\right]t.$$

Αυτός είναι ο τύπος της γεννήτριας συνάρτησης αθροιστικών της Κανονικής κατανομής με μέση τιμή μt και διασπορά $\sigma^2 t$.

- Για $\mu=0$ η ανέλιξη $X(t)$ που προκύπτει, καλείται ανέλιξη Wiener ή ανέλιξη της κίνησης Brown. Στην περίπτωση αυτή, η συνάρτηση πυκνότητας πιθανότητας της $X(t)$ είναι συμμετρική γύρω από την αρχή των αξόνων για όλα τα *t*.
- Για $\mu\neq 0$, η παραπάνω ανέλιξη θα καλείται ανέλιξη Wiener με κατεύθυνση μ και παράμετρο διασποράς σ^2. Για μία τέτοια ανέλιξη, η προσαύξηση

$\Delta X(t) = X(t+\Delta t) - X(t)$ μέσα σε ένα μικρό διάστημα χρόνου Δt είναι ανεξάρτητη της $X(t)$ και έχει μέση τιμή και διασπορά ανάλογη του Δt.

Αποδεικνύεται (Cox και Miller, 1965) πως ο χρόνος πρώτης μετάβασης στο σύνορο $x = a$ έχει σ.π.π της αντίστροφης Γκαουσιανής κατανομής:

$$g(t) = \frac{a}{\sigma\sqrt{2\pi t^3}} \exp\left[-\frac{(a-\mu t)^2}{2\sigma^2 t} \right]$$

Με απλή αναπαραμέτρηση γράφεται στην πιο συνηθισμένη μορφή της σχέσης (2.4), η οποία παρουσιάζεται στη συνέχεια. Η ονομασία αντίστροφη Γκαουσιανή (IG) προκύπτει από τις σχέσεις των γ.σ.α. αυτής της κατανομής με την Κανονική.

Μία γεννήτρια συνάρτηση αθροιστικών συνήθως ορίζεται ως ο λογάριθμος μιας ροπογγενήτριας συνάρτησης, δηλαδή ως $\ln \mathrm{E}\left(\mathrm{e}^{\theta X}\right)$ ή ως $\ln \mathrm{E}\left(\mathrm{e}^{\theta X(t)}\right)$ αν υπάρχει εξάρτηση από το χρόνο *t*. Η ροπογεννήτρια συνάρτηση για την Κανονική κατανομή, $N\left(\mu, \sigma^2\right)$, είναι:

$$\text{ρ.σ.} = \exp\left\{ \mu\theta + \frac{\sigma^2\theta^2}{2} \right\} \tag{2.3}$$

ή $\exp\left\{ \left[\mu\theta + \frac{\sigma^2\theta^2}{2} \right] t \right\}$, αν μέση τιμή $= \mu t$ και διασπορά $= \sigma^2 t$.

Πιο συγκεκριμένα, για την IG με παραμέτρηση (μ, λ) (δηλαδή στη μορφή της σχέσης (2.4)), η ροπογεννήτρια συνάρτηση είναι:

$$\exp\left\{ \frac{\lambda}{\mu}\left(1 - \sqrt{1 - \frac{2\mu^2\theta}{\lambda}} \right) \right\}.$$

Λογαριθμίζοντας, η γ.σ.α. για την IG είναι:

$$\frac{\lambda}{\mu}\left(1 - \sqrt{1 - \frac{2\mu^2\theta}{\lambda}} \right)$$

Για την αντίστροφή της, θέτουμε $s = \frac{\lambda}{\mu}\left(1 - \sqrt{1 - \frac{2\mu^2\theta}{\lambda}} \right)$ και έχουμε:

$$\left(\frac{\mu s}{\lambda} - 1 \right)^2 = \left(1 - \frac{2\mu^2\theta}{\lambda} \right) \Rightarrow \frac{\mu^2 s^2}{\lambda^2} - \frac{2\mu s}{\lambda} + 1 = 1 - \frac{2\mu^2\theta}{\lambda} \Rightarrow \theta = \frac{\lambda}{2\mu^2}\left(\frac{2\mu s}{\lambda} - \frac{\mu^2 s^2}{\lambda^2} \right) \Rightarrow$$

$$\theta = \frac{s}{\mu} - \frac{s^2}{2\lambda},$$

η οποία με κατάλληλη αναπαραμέτρηση (π.χ. για $\theta^* = s, \mu^* = \frac{1}{\mu}, \sigma^2 = \frac{-1}{\lambda}$) καταλήγει στη σχέση (2.3). Δηλαδή, η αντίστροφη της γ.σ.α. της IG είναι η γ.σ.α. της Κανονικής κατανομής $N\left(\mu, \sigma^2\right)$.

Τέλος, μία γ.σ.α. ορίζεται και ως ο λογάριθμος μιας χαρακτηριστικής συνάρτησης, δηλαδή ως $\ln \mathrm{E}\left(\mathrm{e}^{i\theta X}\right)$ (αντίστοιχα $\ln \mathrm{E}\left(\mathrm{e}^{i\theta X(t)}\right)$). Σε αυτή την περίπτωση, η αντιστροφή της απαιτεί μιγαδική ανάλυση, σύμφωνα με τους Whitmore και Seshadri (1987).

2.1.2 Αναλογίες με την Κανονική κατανομή

Στη βιβλιογραφία, υπάρχει πληθος αποτελεσμάτων κατανομών δειγματοληψίας που υποδεικνύουν ομοιότητες της IG με την Κανονική κατανομή. Για παράδειγμα, αν θεωρήσουμε ένα δείγμα τυχαίων μεταβλητών, τις $X_1, X_2, ..., X_n \sim \mathrm{IG}$, αποδεικνύεται:

1. Ο δειγματικός μέσος, $\bar{X}$, ακολουθεί την IG.
2. Οι στατιστικές συναρτήσεις $\bar{X}$ και $\sum\left(1/X_i - 1/\bar{X}\right)$ είναι ανεξάρτητες.
3. Ο όρος στο εκθετικό της κατανομής είναι $\left(-\frac{1}{2}\right)$ φορές μία μεταβλητή που ακολουθεί την χ^2 κατανομή.
4. Ο ομοιόμορφα πιο ισχυρός αμερόληπτος (U.M.P) έλεγχος για τη μέση τιμή χρησιμοποιεί την *t* κατανομή του Student.

Στη βιβλιογραφία, υπάρχει πληθώρα παρόμοιων αποτελεσμάτων (π.χ. Chhikara και Folks, 1989). Ωστόσο, **δεν** υπάρχει κάποιος γενικός κανόνας δειγματοληψίας ή μετασχηματισμός, ο οποίος να μπορεί να παράγει την IG από την Κανονική κατανομή.

2.1.3 Εφαρμογές της IG κατανομής - Η παλινδρόμηση Κατωφλιού (Threshold Regression)

Στο πρώτο κεφάλαιο, παρουσιάσαμε τα μοντέλα χρόνων πρώτης μετάβασης (FHT models) μιας στοχαστικής ανέλιξης, η οποία φτάνει για πρώτη φορά σε κάποιο σύνορο. Για να μετατρέψουμε ένα FHT μοντέλο σε χρήσιμο εργαλείο για τις διάφορες εφαρμογές, πρέπει να το επεκτείνουμε με τέτοιο τρόπο, ώστε να είναι εφικτό να συμπεριλάβει και δομές παλινδρόμησης, όπως εξηγούν αναλυτικά οι Lee και Whitmore (2006). Τέτοιες δομές επιτρέπουν στις μεταβλητές να εξηγήσουν την έμφυτη διασπορά των δεδομένων λαμβάνοντας υπόψη τη μεταβλητότητα και μας οδηγούν σε ακριβείς εξαγωγές συμπερασμάτων. Στη συνέχεια, θα μας απασχολήσουν δομές παλινδρόμησης για τα μοντέλα χρόνων πρώτης μετάβασης, στην οποία θα αναφερόμαστε και ως **παλινδρόμηση**

Κατωφλιού (Threshold Regression), ή TR για συντομία. Η λέξη "κατώφλι" αναφέρεται στο γεγονός ότι ο χρόνος πρώτης μετάβασης "οπλίζεται" από την υποβόσκουσα στοχαστική ανέλιξη η οποία φτάνει σε κάποια κατάσταση κατωφλιού μέσα σε ένα σύνολο συνόρου, όπως ακριβώς περιγράφεται με περισσότερες λεπτομέρειες παρακάτω. Το μοντέλο παλινδρόμησης χρόνου πρώτης μετάβασης (first hitting time regression model) ή **παλινδρόμησης Κατωφλιού** (Whitmore, 1986) προσφέρει μια άλλη προσέγγιση με διαφορετική συμπεριφορά από τις άλλες, ιδιαίτερα ελκυστική, επειδή "βλέπει" τη διάρκεια ζωής ως την έκβαση μιας στοχαστικής ανέλιξης. Αναλυτική περιγραφή των TR μοντέλων θα δοθεί στο Κεφάλαιο 3.

Η έννοια της κίνησης Brown είναι κατάλληλη για την περιγραφή πολλών φαινομένων και ειδικά στις φυσικές επιστήμες. Καθώς ο χρόνος πρώτης μετάβασης σε μία ανέλιξη Wiener με κατεύθυνση μ ακολουθεί την IG κατανομή, είναι εύλογο να χρησιμοποιηθεί ως μοντέλο διάρκειας ζωής. Στην παρούσα διατριβή παρουσιάζεται πληθώρα παραδειγμάτων δεδομένων που ακολουθούν την IG, προσπαθώντας να μελετήσουμε την αξιοπιστία ή την επιβίωση ενός προϊόντος ή μίας συσκευής.

Τα τελευταία χρόνια η κατανομή IG έχει εμφανιστεί σε πολλές περιπτώσεις μοντέλων στοχαστικών ανελίξεων συμπεριλαμβανομένης της θεωρίας των Γενικευμένων Γραμμικών Μοντέλων (McCullagh and Nelder, 1989), της Ανάλυσης Αξιοπιστίας και Επιβίωσης (Padgett και Tsai, 1986), τα μοντέλα Επιταχυνόμενης Ζωής (Bhattacharyya και Fries, 1982) και κατανομές χρόνου επιδιόρθωσης (ειδικότερα σε περιπτώσεις πρόωρων αποτυχιών Chhikara και Folks, 1977). Ακόμα, στο βιομηχανικό Έλεγχο Ποιότητας, ο Edgeman (1989) δημιούργησε διαγράμματα ελέγχου βασισμένα στην IG. Τέλος, ο Hougaard (1984) και οι Feaganes και Suchidran (1991) συνηγορούν υπέρ της χρησιμοποίησης της IG για την κατανομή της ευπάθειας σε δεδομένα διάρκειας ζωής.

2.2 Παρουσίαση της IG

Στην παράγραφο αυτή παρουσιάζονται οι συναρτήσεις πυκνότητας πιθανότητας και κατανομής πιθανότητας της IG. Εισάγεται η μορφή που παρουσίασε ο Tweedie (1957a), καθώς οδηγεί στην ανάπτυξη αναλογίας των στατιστικών ιδιοτήτων της IG και της Κανονικής κατανομής. Τέλος, παρουσιάζεται η έννοια της χαρακτηριστικής συνάρτησης και της ροπογεννήτριας συνάρτησης για την κατανομή.

2.2.1 Συνάρτηση πυκνότητας πιθανότητας

Η σ.π.π. μίας τυχαίας μεταβλητής X που ακολουθεί την IG, όπως παρουσιάστηκε από τον Tweedie (1957a) δίνεται από τον τύπο:

$$f\left(x;\mu,\lambda\right)=\sqrt{\frac{\lambda}{2\pi x^3}}\exp\left\{-\frac{\lambda\left(x-\mu\right)^2}{2\mu^2 x}\right\},\quad x>0,\quad \mu,\ \lambda>0 \tag{2.4}$$

Η παράμετρος μ είναι η **μέση τιμή** της κατανομής και λ είναι η **παράμετρος κλίμακας**. Ο Tweedie παρουσίασε τρεις ακόμα ισοδύναμες μορφές της σχέσης (2.4), οι οποίες προκύπτουν ύστερα από αντικατάσταση των παραμέτρων (μ,λ) από τις (a,λ), (μ,φ) ή (φ,λ) σύμφωνα με τη σχέση $\mu=\frac{\lambda}{\varphi}=(2\alpha)^{-1/2}$:

$$f(x;\alpha,\lambda)=\sqrt{\frac{\lambda}{2\pi x^3}}\exp\left\{-\frac{\lambda}{2}\left(\alpha x-(2\alpha)^{1/2}+\frac{1}{2x}\right)\right\},\quad IG\left((2\alpha)^{-1/2},\lambda\right) \tag{2.5}$$

$$f(x;\mu,\varphi)=\sqrt{\frac{\mu\varphi}{2\pi x^3}}\,e^{\varphi}\exp\left\{-\frac{1}{2}\varphi\left(\frac{x}{\mu}+\frac{\mu}{x}\right)\right\},\quad IG(\mu,\varphi\mu) \tag{2.6}$$

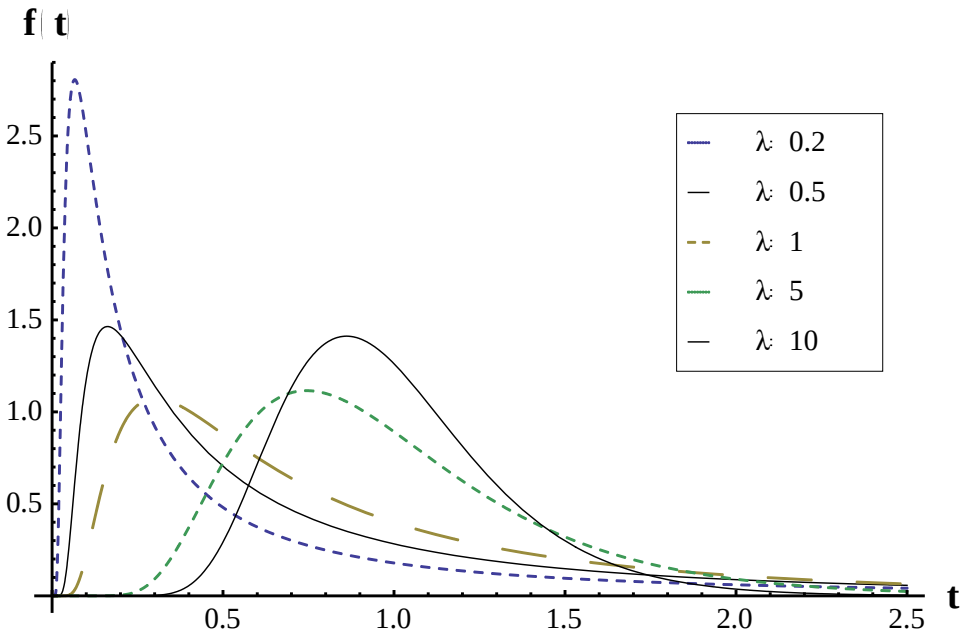
$$f(x;\varphi,\lambda)=\sqrt{\frac{\lambda}{2\pi x^3}}\,e^{\varphi}\exp\left\{-\frac{1}{2}\left(\frac{\varphi^2 x}{\lambda}+\frac{\lambda}{x}\right)\right\},\quad IG\left(\frac{\lambda}{\varphi},\lambda\right) \tag{2.7}$$

Οι μ, λ έχουν την ίδια φυσική διάσταση με την τυχαία μεταβλητή X. Ωστόσο, η παράμετρος $\varphi=\frac{\lambda}{\mu}$ παραμένει αναλλοίωτη κάτω από έναν μετασχηματισμό κλίμακας της X, όπως φαίνεται από τις παρακάτω σχέσεις:

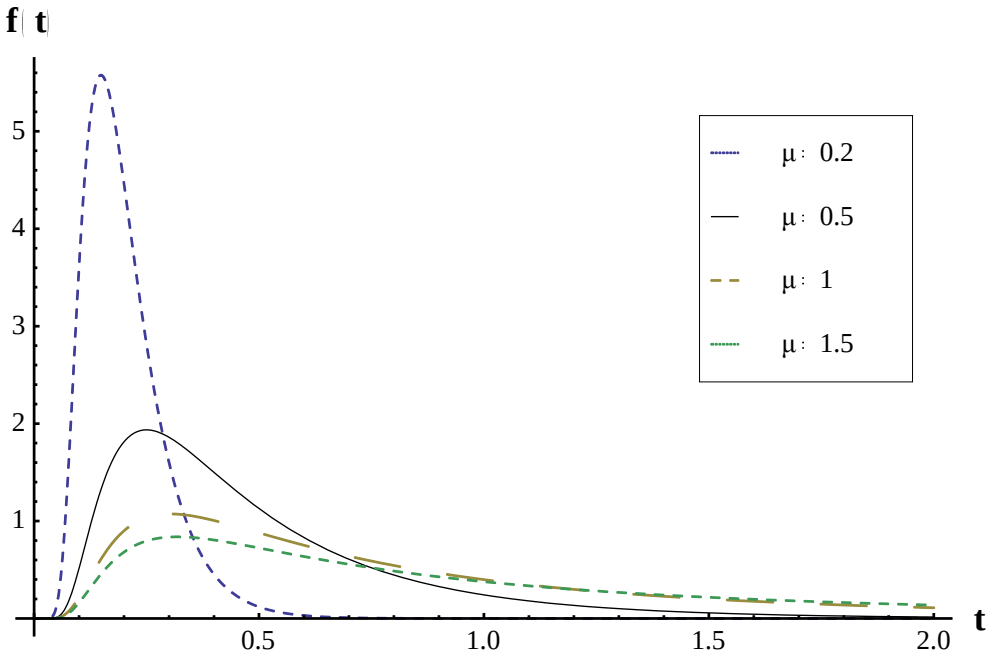
$$f(x;\mu,\lambda)=\mu^{-1}f\left(\frac{x}{\mu};1,\varphi\right)=\lambda^{-1}f\left(\frac{x}{\lambda};\varphi,1\right) \tag{2.8}$$

Η σ.π.π. μπορεί να υπολογιστεί αριθμητικά χρησιμοποιώντας οποιοδήποτε από τους τρείς ισοδύναμους τύπους της (2.8). Όπως θα δείξουμε στη συνέχεια (Παράγραφος 2.2.3), η συνάρτηση κατανομής πιθανότητας εξαρτάται από δύο μόνο ποσότητες, οι οποίες μπορεί να είναι οι $\frac{x}{\mu}$ και φ. Επομένως, η περίπτωση $\mu=1$ για την παραμετρική μορφή (μ,φ) της σχέσης (2.8) μπορεί να θεωρηθεί ως η κανονική μορφή.

Το σχήμα της κατανομής εξαρτάται μόνο από την παράμετρο φ. Επομένως, η φ θα αποτελεί την **παράμετρο σχήματος**. Η συνάρτηση πυκνότητας πιθανότητας της αντίστροφης Γκαουσιανής κατανομής αναπαριστά μία ευρεία κλάση κατανομών, οι οποίες κυμαίνονται από μία ιδιαίτερα λοξή έως μία συμμετρική, καθώς το φ μεταβάλλεται από 0 έως $+\infty$. Η ιδιότητα αυτή απεικονίζεται γραφικά στα Σχήματα 2.1 και 2.2. Οι καμπύλες των δύο σχημάτων προκύπτουν θέτοντας $\mu=1$ και μεταβάλλοντας το φ ή το λ για το Σχήμα 2.1, και θέτοντας $\lambda=1$ και ακολούθως μεταβάλλοντας το μ ή το φ για το Σχήμα 2.2.



Σχήμα 2.1: Γράφημα της σ.π.π για διάφορες τιμές του λ, με μ = 1



Σχήμα 2.2: Γράφημα της σ.π.π για διάφορες τιμές του μ, με λ = 1

Το γράφημα της σ.π.π. είναι μονοκόρυφο, με την κορυφή να επιτυγχάνεται στο σημείο:

$$\mu\left[\left(1+\frac{9}{4\varphi^2}\right)^{1/2}-\frac{3}{2\varphi}\right].$$

2.2.2 Χαρακτηριστική συνάρτηση και ροπογγενήτριες συναρτήσεις

Η χαρακτηριστική συνάρτηση μίας τυχαίας μεταβλητής $X \sim \mathrm{IG}(\mu, \lambda)$ δίνεται από τον τύπο:

$$\mathrm{E}\left[e^{itX}\right] = \int_0^{+\infty} e^{itX} f\left(x;\mu,\lambda\right)dx$$

Ύστερα από συνδυασμό των δύο όρων μέσα στο ολοκλήρωμα και απαλοιφή της σταθεράς στο δεύτερο μέλος της παρακάτω εξίσωσης,

$$\mathrm{E}\left[e^{itX}\right] = \exp\left\{\frac{\lambda}{\mu}\left[1-\left(1-\frac{2it\mu^2}{\lambda}\right)^{1/2}\right]\right\} \times \int_0^{+\infty} e^{itX} f\left(x;\mu\left(1-\frac{2it\mu^2}{\lambda}\right)^{-1/2},\lambda\right)dx,$$

καταλήγουμε στον εξής τύπο για τη χαρακτηριστική συνάρτηση, που συμβολίζεται με $C_X(t)$:

$$C_X(t) = \exp\left\{\frac{\lambda}{\mu}\left[1-\left(1-\frac{2it\mu^2}{\lambda}\right)^{1/2}\right]\right\} \tag{2.9}$$

Όλες οι θετικές και αρνητικές ροπές υπάρχουν. Οι θετικές ροπές προκύπτουν από παραγώγιση της χαρακτηριστικής συνάρτησης της σχέσης (2.9), ενώ οι αρνητικές με ολοκλήρωσή της (Cressie et al., 1981). Παίρνοντας την *r*-οστή παράγωγο της $C_X(t)$ και υπολογίζοντάς την στο σημείο $t = 0$, έχουμε:

$$\mathrm{E}\left[X^r\right] = \mu^r \sum_{k=0}^{r-1} \frac{(r-1+k)!}{(r-1-k)!}\left(2\frac{\lambda}{\mu}\right)^{-k} \tag{2.10}$$

Οι πρώτες τέσσερις ροπές γύρω από το μηδέν είναι:

$$\begin{aligned} &\mu \\ &\mu^2 + \frac{\mu^3}{\lambda} \\ &\mu^3 + 3\frac{\mu^4}{\lambda} + 3\frac{\mu^5}{\lambda^2} \\ &\mu^4 + 6\frac{\mu^5}{\lambda} + 15\frac{\mu^6}{\lambda^2} + 15\frac{\mu^7}{\lambda^3} \end{aligned} \tag{2.11}$$

Οι κεντρικές ροπές προκύπτουν είτε από τις (2.10) και (2.11), είτε από το ανάπτυγμα της δυναμοσειράς της γεννήτριας συνάρτησης αθροιστικών, το οποίο δίνεται από τη σχέση:

$$\ln \mathrm{E}\left(e^{tX}\right) = \frac{\lambda}{\mu}\left[1-\left(1-\frac{2it\mu^2}{\lambda}\right)^{1/2}\right], \; t < \frac{\lambda}{2\mu^2} \tag{2.12}$$

Για τις κεντρικές ροπές είναι βολικό να δουλεύουμε με τα αθροίσματα K_r, όπου K_r είναι ο συντελεστής της μεταβολής της ποσότητας $\left(\frac{t^r}{r!}\right)$ στο ανάπτυγμα της γεννήτριας συνάρτησης αθροιστικών, καθώς παραμένουν αναλλοίωτα (εκτός από τον όρο K_1) στους μετασχηματισμούς τυχαίων μεταβλητών. Εύλογα, προκύπτει από τη (2.12):

$$K_1 = \mu$$

και

$$K_r = 1\cdot 3\cdot 5\cdot \ldots \cdot (2r-3)\mu^{2r-1}\lambda^{1-r},\ r \geq 2 \tag{2.13}$$

Από την (2.11) ή την (2.13), η δεύτερη, τρίτη και τέταρτη κεντρική ροπή είναι:

$$\begin{aligned} \mu_2 &= \frac{\mu^3}{\lambda} \\ \mu_3 &= 3\frac{\mu^5}{\lambda^2} \\ \mu_4 &= 15\frac{\mu^7}{\lambda^3} + 3\frac{\mu^6}{\lambda^2} \end{aligned} \tag{2.14}$$

Επίσης, υπάρχει μία απλή σχέση ανάμεσα στις αρνητικές και τις θετικές ροπές, η οποία δίνεται από τη σχέση:

$$\mathrm{E}\left[X^{-r}\right] = \frac{\mathrm{E}\left[X^{r+1}\right]}{\mu^{2r+1}} \tag{2.15}$$

Ο συντελεστής της μεταβλητότητας της X είναι $\sqrt{\frac{\mu}{\lambda}}$. Τα μέτρα λοξότητας και ασυμμετρίας είναι αντίστοιχα:

$$\begin{aligned} \sqrt{\beta_1} &= 3\sqrt{\frac{\mu}{\lambda}} \\ \sqrt{\beta_2} &= 15\left(\frac{\mu}{\lambda}\right) + 3 \end{aligned} \tag{2.16}$$

2.2.3 Συνάρτηση κατανομής πιθανότητας

Υπολογίζοντας τη ροπογεννήτρια συνάρτηση για μία τυχαία μεταβλητή $X \sim \mathrm{IG}(\mu, \lambda)$ αποδεικνύεται, πως η τυχαία μεταβλητή $\lambda(X-\mu)^2/\mu^2 X$ ακολουθεί τη χ^2 κατανομή με ένα βαθμό ελευθερίας (Seshadri, 1993). Παρατηρώντας τη μορφή της σ.π.π. της σχέσης (2.4), προκύπτει πως η νέα μεταβλητή ισούται με (-2) φορές τον όρο που βρίσκεται στο εκθετικό.

Επομένως, παίρνουμε μία αναλογία με την Κανονική κατανομή, όπως παρουσιάσαμε και στην Παράγραφο 2.1.2. Βασισμένος σε αυτόν το μετασχηματισμό, ο Shuster (1968) παρουσίασε μία έκφραση της συνάρτησης κατανομής $F(x)$ της X σε όρους της τυποποιημένης Κανονικής κατανομής $\Phi(\cdot)$, η οποία δίνεται από τη σχέση:

$$F(x)=\Phi\left[\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu}-1\right)\right]+\exp^{2\lambda/\mu}\Phi\left[-\sqrt{\frac{\lambda}{x}}\left(1+\frac{x}{\mu}\right)\right] \tag{2.17}$$

Ας θεωρήσουμε την τυχαία μεταβλητή $\mathrm{Y}=\sqrt{\lambda}\dfrac{(X-\mu)}{\mu\sqrt{X}}$. Αποδεικνύεται πως για τη συνάρτηση κατανομής της Y, $G(y)$ ισχύει ότι $G(y)\to\Phi(y)$, καθώς $\varphi=\dfrac{\lambda}{\mu}\to+\infty$. Λόγω αυτού του αποτελέσματος και επειδή υπάρχει μία $1-1$ σχέση μεταξύ X και Y, αποδεικνύεται πως η κατανομή της X προσεγγίζει ασυμπτωτικά την Κανονική κατανομή, με μέση τιμή μ και διασπορά μ^3/λ (Wald, 1947). Να σημειώσουμε ότι μεγάλο πλεονέκτημα για την αντίστροφη Γκαουσιανή κατανομή αποτελεί το γεγονός πως έχει κλειστού τύπου συνάρτηση πυκνότητας πιθανότητας, καθώς και μία απλή στον υπολογισμό συνάρτηση κατανομής.

2.3 Χρήσιμοι μετασχηματισμοί και ιδιότητες της IG

Στην παράγραφο αυτή παρουσιάζονται μετασχηματισμοί κανονικοποίησης της IG και περιγράφεται μία μέθοδος παραγωγής ψευδοτυχαίων παρατηρήσεων από την αντίστροφη Γκαουσιανή κατανομή.

2.3.1 Κανονικοποίηση της IG

Δεν υπάρχει μετασχηματισμός που να οδηγεί σε κάποια τυπική μορφή για την κατανομή μίας IG μεταβλητής, παρόμοια με αυτήν της τυποποιημένης Κανονικής κατανομής. Ωστόσο, σε περίπτωση που συμβεί κάποια μετατροπή στην κλίμακα, είναι εφικτό να προκύψει μείωση του αριθμού των παραμέτρων της κατανομής και να πάρουμε μία αντίστροφη Γκαουσιανή κατανομή με **μία** μόνο παράμετρο. Μία κανονικοποίηση της αντίστροφης Γκαουσιανής κατανομής δόθηκε από τους Whitmore και Yalovsky (1978), χρησιμοποιώντας τον παρακάτω μετασχηματισμό:

$$W=\frac{1}{2\sqrt{\varphi}}+\sqrt{\varphi}\,\log\frac{X}{\mu} \tag{2.18}$$

Αποδεικνύεται πως όταν το φ είναι αρκετά μεγάλο, η τυχαία μεταβλητή W ακολουθεί προσεγγιστικά την τυποποιημένη Κανονική κατανομή. Καθώς ο πρώτος όρος της (2.18) μηδενίζεται για μεγάλες τιμές του φ, μπορεί να χρησιμοποιηθεί η τυχαία μεταβλητή

$$\sqrt{\varphi}\,\log\frac{X}{\mu} \tag{2.19}$$

για κανονικοποίηση της X. Οι Whitmore και Yalovsky (1978) παρατήρησαν ακόμα πως η τυχαία μεταβλητή

$$\sqrt{\varphi}\left(\frac{X}{\mu}-1\right) \tag{2.20}$$

είναι προσεγγιστικά Κανονικά κατανεμημένη, καθώς $\varphi\to+\infty$. Απέδειξαν επίσης πως η κατανομή της W που προκύπτει από τη σχέση (2.18) φτάνει πιο γρήγορα στην Κανονικότητα από τη μεταβλητή της σχέσης (2.20). Το ίδιο αποτέλεσμα ισχύει αντίστοιχα και για τη μεταβλητή που προκύπτει από τη σχέση (2.19), με την κατανομή της να φτάνει πιο γρήγορα στην Κανονικότητα από τη μεταβλητή της σχέσης (2.20).

2.3.2 Γέννηση ψευδοτυχαίων μεταβλητών από την IG

Μία συνηθισμένη τεχνική που συναντάται στη βιβλιογραφία για την παραγωγή τυχαίων παρατηρήσεων από μια κατανομή, είναι η αντιστροφή της συνάρτησης κατανομής πιθανότητας της μεταβλητής. Σε περίπτωση που η αντίστροφη της σ.κ. έχει κλειστή μορφή, απλή στον υπολογισμό, είναι εύκολο να παράγουμε τυχαίες τιμές για την κατανομή χρησιμοποιώντας μία γεννήτρια ομοιόμορφα κατανεμημένων τυχαίων παρατηρήσεων $\left(\mathrm{U}(0,1)\right)$. Η σ.κ. της IG δεν προσφέρεται για την παραπάνω μέθοδο, καθώς δεν αντιστρέφεται εύκολα.

Οι Michael et al. (1976), παρουσίασαν μία μέθοδο γέννησης ψευδοτυχαίων μεταβλητών χρησιμοποιώντας ένα μετασχηματισμό με πολλαπλές ρίζες. Η προσέγγισή τους αφορούσε την εύρεση μετασχηματισμού για την τυχαία μεταβλητή του ενδιαφέροντος και στη συνέχεια έπρεπε να χρησιμοποιήσουν τις πολλαπλές πιθανότητες που σχετίζονται με τις πολλαπλές ρίζες του μετασχηματισμού προκειμένου να διαλέξουν μία ρίζα για την τυχαία παρατήρηση.

Για την τυχαία μεταβλητή $X\sim\mathrm{IG}(\mu,\lambda)$, η μετασχηματισμένη τυχαία μεταβλητή:

$$Y^2=\frac{\lambda(X-\mu)^2}{\mu^2 X}$$

ακολουθεί τη χ^2 κατανομή με ένα βαθμό ελευθερίας και έχει δύο ρίζες, τις X_1, X_2 με

$$X_1=\frac{\mu}{2\lambda}\left[2\lambda+\mu Y^2-\sqrt{4\lambda\mu Y^2+\mu^2Y^4}\right] \tag{2.21}$$

και

$$X_2 = \frac{\mu^2}{X_1}. \tag{2.22}$$

Δεδομένης μίας τυχαίας τιμής για την Y^2 που γεννιέται από τη χ_1^2 κατανομή, οι Michael et al. (1976) έδωσαν την υπό συνθήκη πιθανότητα με την οποία επιλέγεται κάθε ρίζα. Η μικρότερη ρίζα X_1, επιλέγεται με πιθανότητα $\frac{\mu}{\mu + X_1}$ και η άλλη ρίζα X_2 με πιθανότητα $\frac{X_1}{\mu + X_1}$. Η συνολική διαδικασία συνοψίζεται στα εξής βήματα:

1. Γέννηση τυχαίων παρατηρήσεων από τη χ^2 κατανομή με ένα βαθμό ελευθερίας.
2. Για κάθε μία τιμή του βήματος 1, υπολογισμός της μικρότερης ρίζας X_1, όπως δίνεται παραπάνω.
3. Διεξαγωγή μίας δοκιμής Bernoulli με πιθανότητα επιτυχίας $p = \frac{\mu}{\mu + X_1}$.
4. Σε περίπτωση που η δοκιμή έχει επιτυχία, επιλέγεται η ρίζα X_1 για την τυχαία παρατήρηση από την IG με παραμέτρους μ και λ, αλλιώς επιλέγεται η μεγαλύτερη ρίζα.

2.4 Η IG ως γενικευμένο γραμμικό μοντέλο

Σε αυτήν την παράγραφο, γίνεται μία εισαγωγή στη θεωρία των γενικευμένων γραμμικών μοντέλων (Generalised Linear Models - GLMs). Παρουσιάζεται η IG ως γενικευμένο γραμμικό μοντέλο. Ακόμα, μελετάται η σημαντικότητα της επιλογής της σωστής συνάρτησης σύνδεσης στην προσαρμογή της IG σε ένα δείγμα δεδομένων. Επίσης, παρουσιάζεται ένας αλγόριθμος στην R για την παραπάνω μελέτη. Τέλος, δίνονται τα αποτελέσματα της μελέτης και ακολουθούν συμπεράσματα πάνω σε αυτά.

2.4.1 Γενικευμένα γραμμικά μοντέλα

Το γενικευμένο γραμμικό μοντέλο παρουσιάστηκε από τους Nelder και Wedderburn (1972) και αποτελεί μία ενοποίηση γραμμικών και μη γραμμικών μοντέλων παλινδρόμησης, τα οποία επιτρέπουν στον πειραματιστή να επιλέξει για τη μεταβλητή απόκρισης μία κατανομή που είναι μέλος μίας ιδιαίτερα πλούσιας και ευέλικτης οικογένειας, της εκθετικής οικογένειας κατανομών. Η Κανονική, η Διωνυμική, η Εκθετική, η Poisson, η Γάμμα, η αρνητική Διωνυμική και η αντίστροφη Γκαουσιανή κατανομή είναι μεταξύ άλλων, κάποιες από τις κατανομές αυτής της οικογένειας.

Η αντίστροφη Γκαουσιανή κατανομή είναι το πιο σπάνια χρησιμοποιούμενο γενικευμένο γραμμικό μοντέλο που υπάρχει στη βιβλιογραφία. Οι διάφορες πηγές πάντοτε το συμπεριλαμβάνουν σε πίνακες με οικογένειες γενικευμένων γραμμικών μοντέλων.

Ωστόσο, σπάνια γίνονται αναφορές σε αυτό. Ακόμα και οι McCullagh και Nelder (1989), απλά σημειώνουν την ύπαρξή του με ελάχιστες αναφορές σε αυτό. Στους Hardin και Hilbe (2001), δίνεται αναλυτική περιγραφή της IG ως γενικευμένο γραμμικό μοντέλο.

Ένα γενικευμένο γραμμικό μοντέλο αποτελείται από τρία μέρη: **1)** το τυχαίο μέρος (συμβολίζεται με Y, όπου $E(Y)=\mu$ και $V(Y)=\sigma^2$), το οποίο αναφέρεται στην κατανομή της απόκρισης (εξαρτημένης μεταβλητής) ή στη δομή του σφάλματος, **2)** το συστηματικό μέρος (συμβολίζεται με η), που αναφέρεται σε ένα παράγοντα πρόγνωσης που είναι γραμμικός συνδυασμός των μεταβλητών που συμμετέχουν στο πείραμα, ήτοι $\eta=\sum_{i=1}^{p}\beta_i x_i$ και **3)** μία συνάρτηση σύνδεσης ανάμεσα στο τυχαίο και το συστηματικό μέρος, η οποία συνδέει τη φυσική μέση τιμή για το συγκεκριμένο μέλος της εκθετικής οικογένειας κατανομών με το γραμμικό παράγοντα (ή γραμμική προβλέπουσα), δηλαδή $\eta=g(\mu)=\boldsymbol{x}'\boldsymbol{\beta}$.

Υπάρχουν δύο είδη συναρτήσεων σύνδεσης: οι κανονικές (canonical) και οι μη-κανονικές (non-canonical). Η κανονική είναι η συνάρτηση που εξισώνει τη φυσική παράμετρο θέσης της εκθετικής οικογένειας κατανομών με το γραμμικό προγνωστικό παράγοντα. Τέλος, η εκτίμηση των παραμέτρων στα γενικευμένα γραμμικά μοντέλα συνήθως γίνεται με τη μέθοδο μέγιστης πιθανοφάνειας. Υπάρχουν πολλά λογισμικά, τόσο ελεύθερα όσο και εμπορικά, τα οποία υποστηρίζουν το γενικευμένο γραμμικό μοντέλο, συμπεριλαμβανομένων της R, SAS, STATA, SPSS και S-PLUS.

Η κατανομή μιας τυχαίας μεταβλητής *Y* ανήκει στην εκθετική οικογένεια κατανομών, εάν η σ.π.π. της μεταβλητής μπορεί να πάρει τη μορφή:

$$f_Y(y;\theta,\varphi)=\exp\left\{\left(y\theta-b(\theta)\right)/\alpha(\varphi)+c(y,\varphi)\right\}, \tag{2.23}$$

για συγκεκριμένες συναρτήσεις $\alpha(\cdot)$, $b(\cdot)$ και $c(\cdot)$. Η παράμετρος φ (συχνά συμβολίζεται με σ^2) καλείται παράμετρος διασποράς και είναι σταθερή για τις διάφορες παρατηρήσεις. Εάν η φ είναι γνωστή, τότε το μοντέλο της σχέσης (2.23) είναι ένα μοντέλο της εκθετικής οικογένειας κατανομών με κανονική παράμετρο θ. Η κατανομή IG ανήκει στην εκθετική οικογένεια κατανομών τάξης 2. Θέτοντας $\sigma^2=1/\lambda$, η σ.π.π. της σχέσης (2.4) (συμβολίζεται με $\mathrm{IG}(\mu,\sigma^2)$) μπορεί να γραφτεί ως:

$$f(y;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi y^3\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2(\mu\sigma)^2 y}\right\}, \tag{2.24}$$

Σε εκθετική μορφή, η κατανομή της IG δίνεται από:

$$f(y;\mu,\sigma^2)=\exp\left\{-\frac{(y-\mu)^2}{2(\mu\sigma)^2 y}-\frac{1}{2}\ln\left(2\pi y^3\sigma^2\right)\right\}, \tag{2.25}$$

Ισοδύναμα,

$$f\left(y;\mu,\sigma^2\right)=\exp\left\{-\frac{y/\mu^2-2/\mu}{-2\sigma^2}+\frac{1/y}{-2\sigma^2}+\frac{\sigma^2}{-2\sigma^2}\ln\left(2\pi y^3\sigma^2\right)\right\} \tag{2.26}$$

Η θεωρία των γενικευμένων γραμμικών μοντέλων μας παρέχει τη συνάρτηση σύνδεσης και τις σωρευτικές συναρτήσεις στην **κανονική μορφή**:

$$g\left(\mu\right)=\theta=\frac{1}{2\mu^2}=\frac{1}{2}\mu^{-2} \tag{2.27}$$

$$\alpha\left(\varphi\right)=-\varphi=-\sigma^2 \tag{2.28}$$

$$b\left(\theta\right)=\frac{1}{\mu} \tag{2.29}$$

$$c\left(y;\varphi\right)=-\frac{1}{2}\left\{\log\left(2\pi\varphi x^3\right)+\frac{1}{\varphi y}\right\} \tag{2.30}$$

Το πρόσημο και η τιμή του συντελεστή συνήθως απλοποιούνται από τη συνάρτηση σύνδεσης. Στην IG, συναντώνται τέσσερις συναρτήσεις σύνδεσης:

α) η κανονική, όπου $\eta=g\left(\mu\right)=\frac{1}{\mu^2}$,

β) η αντίστροφη με $\eta=g\left(\mu\right)=\frac{1}{\mu}$,

γ) η ταυτοτική με $\eta=g\left(\mu\right)=\mu$ και

δ) η λογαριθμική, στην οποία $\eta=g\left(\mu\right)=\log\mu$.

2.4.2 Επιλογή της συνάρτησης σύνδεσης με τη βοήθεια διαστημάτων εμπιστοσύνης

Οι Myers και Montgomery (1997) παρουσίασαν ένα εγχειρίδιο για τα γενικευμένα γραμμικά μοντέλα. Σε αυτό, δίνουν πολλά παραδείγματα συγκρίνοντας μοντέλα κατασκευασμένα με τις μεθόδους ελαχίστων τετραγώνων και μετασχηματισμούς δεδομένων και μοντέλα κατασκευασμένα σύμφωνα με τη θεωρία των γενικευμένων γραμμικών μοντέλων. Σε κάθε περίπτωση, βρήκαν πως ένα καλύτερο μοντέλο είναι εφικτό με τη χρήση των GLM, όπου η έννοια "καλύτερο" μετριέται σε όρους απόδοσης στην εκτίμηση της απόκρισης και πρόγνωσης. Οι Myers και Montgomery χρησιμοποιούν το μήκος ενός διαστήματος εμπιστοσύνης για τη μέση απόκριση στη σύγκριση μοντέλων.

Όπως αναφέραμε προηγουμένως, ένα από τα τρία βασικά τμήματα ενός GLM είναι η συνάρτηση σύνδεσης. Σε κάποιες περιπτώσεις, η κανονική συνάρτηση μπορεί να επιλεχθεί εκ των προτέρων, συνήθως για ευκολία στην ερμηνεία. Η χρήση της κανονικής συνάρτησης σύνδεσης σε ένα GLM προσφέρει μερικά τεχνικά πλεονεκτήματα αλλά δεν είναι απαραίτητη, π.χ. η παλινδρόμηση Poisson δε γίνεται υποχρεωτικά με τη λογαριθμική συνάρτηση. Για κάποιες περιπτώσεις όμως, μπορεί να μην είναι προφανής η επιλογή της κατάλληλης συνάρτησης σύνδεσης. Στην περίπτωση της συνηθισμένης $IG\left(\mu,\sigma^2\right)$, η κανονική συνάρτηση σύνδεσης είναι $\frac{1}{\mu^2}$, αλλά επίσης χρησιμοποιούνται η λογαριθμική, η ταυτοτική και η αντίστροφη. Θέλουμε να εξετάσουμε πόσο σημαντική είναι η επιλογή της σωστής συνάρτησης σύνδεσης στην περίπτωση της IG.

Στηριζόμενοι στο έργο των Myers και Montgomery (1997), αλλά και στους Lewis et al. (2001a και 2001b), θα μελετήσουμε την επίδραση που θα έχει μία λανθασμένη επιλογή συνάρτησης σύνδεσης στις εκτιμήσεις της κάλυψης και της ακρίβειας (μήκος) ενός διαστήματος εμπιστοσύνης για την παράμετρο της μέσης τιμής της IG κατανομής.

2.4.3 Περιγραφή αλγορίθμου

Στη συνέχεια, παρουσιάζεται ο αλγόριθμος που χρησιμοποιήθηκε για την παραγωγή των πιθανοτήτων κάλυψης της παραμέτρου μ και στηρίζεται στην ιδέα των Lewis et al (2001b).

1. Επιλογή του μεγέθους *n* του δείγματος.
2. Δημιουργία δύο μεταβλητών, X_1 και X_2 με τιμές 0 και 1, ως παραγοντικό σχεδιασμό μεγέθους *n*.
3. Ορισμός του αληθινού μοντέλου, περιλαμβάνοντας και τη συνάρτηση σύνδεσης.
4. Δημιουργία δεδομένων.
5. Προσαρμογή του μοντέλου IG τέσσερις φορές, με τη σωστή συνάρτηση σύνδεσης (δηλαδή του αληθινού μοντέλου) και με τις άλλες τρείς γνωστές συναρτήσεις σύνδεσης της κατανομής.
6. Επανάληψη 500 φορών του βήματος 4 δίνοντας εμπειρικά τις πραγματικές πιθανότητες κάλυψης (coverage probabilities) του 95% διαστήματος εμπιστοσύνης της αληθινής παραμέτρου μ και της αντίστοιχης ακρίβειας του δ.ε., τόσο για κάθε μία παρατήρηση μ_i ξεχωριστά όσο και συνολικά για όλες τις παρατηρήσεις του σχεδιασμού.

Να παρατηρήσουμε πως στο βήμα 5 παίρνουμε αποτελέσματα για κάθε μ_i ξεχωριστά (αν υπάρχουν οκτώ σημεία σε κάθε σύνολο δεδομένων, εξετάζουμε εάν το καθένα καλύπτεται από το αντίστοιχο διάστημα εμπιστοσύνης). Σαν αποτέλεσμα, η τελική εμπειρική πιθανότητα κάλυψης βασίζεται σε $8\times 500 = 4,000$ σημεία.

2.4.4 Αποτελέσματα μελέτης

Στους πίνακες που ακολουθούν παρουσιάζουμε τις πιθανότητες κάλυψης, καθώς και την ακρίβεια των διαστημάτων εμπιστοσύνης για τέσσερις διαφορετικές περιπτώσεις, ανάλογα με το μέγεθος του δείγματος και τη συνάρτηση σύνδεσης του αληθινού μοντέλου. Για το αληθινό μοντέλο χρησιμοποιήσαμε την κανονική και τη λογαριθμική συνάρτηση σύνδεσης. Σύμφωνα με τον αλγόριθμο που παρουσιάσαμε στην προηγούμενη παράγραφο, κάθε φορά προσαρμόζαμε το γενικό γραμμικό μοντέλο με δύο μεταβλητές, τις X_1 και X_2 (παραγοντικός σχεδιασμός μεγέθους *n*), σε IG δεδομένα. Έγιναν δύο διαφορετικές επιλογές μεγέθους δείγματος $(n=8$ και $n=16)$. Σε όλες τις προσομοιώσεις, θεωρήσαμε την παράμετρο λ του αληθινού μοντέλου ίση με τη μονάδα. 500 προσομοιώσεις έγιναν για την κάθε περίπτωση.

Πρώτη περίπτωση

Αληθινό μοντέλο: $$g(\mu)=\frac{1}{\mu^2}=100+50x_1+30x_2,$$

με $x_1=(1,1,0,0,1,1,0,0)$ και $x_2=(1,0,1,0,1,0,1,0)$, $n=8$, $\lambda=1$, 500 προσομοιώσεις.

Οι πιθανότητες κάλυψης για την κάθε μ_i δίνονται στον Πίνακα 2.1:

Αληθινό μοντέλο	**Συνάρτηση σύνδεσης που χρησιμοποιήθηκε στο προσαρμοσμένο μοντέλο**			
Κανονική	**Κανονική**	**Λογαριθμική**	**Αντίστροφη**	**Ταυτοτική**
Αριθμός παρατήρησης	**Πιθανότητες κάλυψης**			
1	90.4	88.0	89.6	85.6
2	84.0	84.4	84.4	83.8
3	88.0	88.2	88.2	88.0
4	87.0	85.8	86.4	83.4
5	90.4	88.0	89.6	85.6
6	84.0	84.4	84.4	83.8
7	88.0	88.2	88.2	88.0
8	87.0	85.8	86.4	83.4
Μέση κάλυψη (%)	**87.4**	**86.6**	**87.2**	**85.2**

Πίνακας 2.1: Πιθανότητες κάλυψης της παραμέτρου *μ* για την κανονική συνάρτηση σύνδεσης στο αληθινό μοντέλο και μέγεθος δείγματος $n=8$.

Οι αντίστοιχες τιμές για την ακρίβεια (μήκος) του διαστήματος εμπιστοσύνης για την κάθε μ_i δίνονται στον Πίνακα 2.2:

Αληθινό μοντέλο	Συνάρτηση σύνδεσης που χρησιμοποιήθηκε στο προσαρμοσμένο μοντέλο			
Κανονική	**Κανονική**	**Λογαριθμική**	**Αντίστροφη**	**Ταυτοτική**
Αριθμός παρατήρησης	**Ακρίβεια (μήκος) διαστήματος εμπιστοσύνης**			
1	0.043	0.045	0.044	0.043
2	0.052	0.052	0.052	0.052
3	0.060	0.058	0.059	0.060
4	0.077	0.068	0.073	0.077
5	0.043	0.045	0.044	0.043
6	0.052	0.052	0.052	0.052
7	0.060	0.058	0.059	0.060
8	0.077	0.0685	0.073	0.077

Πίνακας 2.2: Ακρίβεια του διαστήματος εμπιστοσύνης της παραμέτρου μ για την κανονική συνάρτηση σύνδεσης στο αληθινό μοντέλο και μέγεθος δείγματος $n=8$.

Δεύτερη περίπτωση

Αληθινό μοντέλο: $g(\mu)=\log\mu=-5+2x_1+x_2$,

με $x_1=(1,1,0,0,1,1,0,0)$ και $x_2=(1,0,1,0,1,0,1,0)$, $n=8$, $\lambda=1$, 500 προσομοιώσεις.

Οι πιθανότητες κάλυψης για την κάθε μ_i δίνονται στον Πίνακα 2.3:

Αληθινό μοντέλο	Συνάρτηση σύνδεσης που χρησιμοποιήθηκε στο προσαρμοσμένο μοντέλο			
Λογαριθμική	**Λογαριθμική**	**Κανονική**	**Αντίστροφη**	**Ταυτοτική**
Αριθμός παρατήρησης	**Πιθανότητες κάλυψης**			
1	87.0	100	100	37.6
2	86.2	99.4	90.0	85.4
3	85.6	47.4	67.6	98.4
4	87.8	4.4	38.8	98.8
5	87.0	100	100	37.6
6	86.2	99.4	90.0	85.4
7	85.6	47.4	67.6	98.4
8	87.8	4.4	38.8	98.8
Μέση κάλυψη (%)	**86.7**	**62.8**	**74.1**	**80.1**

Πίνακας 2.3: Πιθανότητες κάλυψης της παραμέτρου μ για τη λογαριθμική συνάρτηση σύνδεσης στο αληθινό μοντέλο και μέγεθος δείγματος $n=8$.

Οι αντίστοιχες τιμές για την ακρίβεια (μήκος) του διαστήματος εμπιστοσύνης για την κάθε μ_i δίνονται στον Πίνακα 2.4:

Αληθινό μοντέλο	**Συνάρτηση σύνδεσης που χρησιμοποιήθηκε στο προσαρμοσμένο μοντέλο**			
Λογαριθμική	**Λογαριθμική**	**Κανονική**	**Αντίστροφη**	**Ταυτοτική**
Αριθμός παρατήρησης	**Ακρίβεια (μήκος) διαστήματος εμπιστοσύνης**			
1	0.075	0.551	0.888	0.075
2	0.024	0.095	0.061	0.025
3	0.006	0.011	0.012	0.006
4	0.001	0.009	0.007	0.001
5	0.075	0.551	0.888	0.075
6	0.024	0.095	0.061	0.025
7	0.006	0.011	0.012	0.006
8	0.001	0.009	0.007	0.001

Πίνακας 2.4: Ακρίβεια του διαστήματος εμπιστοσύνης της παραμέτρου μ για την κανονική συνάρτηση σύνδεσης στο αληθινό μοντέλο και μέγεθος δείγματος $n = 8$.

Τρίτη περίπτωση

Αληθινό μοντέλο:

$$g(\mu) = \frac{1}{\mu^2} = 100 + 50x_1 + 30x_2,$$

με $x_1 = (1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0)$ και $x_2 = (1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0)$,

$n = 16$, $\lambda = 1$, 500 προσομοιώσεις.

όμοια με τις δύο προηγούμενες περιπτώσεις, οι πιθανότητες κάλυψης για την κάθε μ_i δίνονται στον Πίνακα 2.5:

Αληθινό μοντέλο	**Συνάρτηση σύνδεσης που χρησιμοποιήθηκε στο προσαρμοσμένο μοντέλο**			
Κανονική	**Κανονική**	**Λογαριθμική**	**Αντίστροφη**	**Ταυτοτική**
Αριθμός παρατήρησης	**Πιθανότητες κάλυψης**			
1	90.2	91.4	90.6	92.2
2	91.0	92.4	90.8	91.4
3	89.4	90.6	90.0	90.0
4	90.4	90.6	90.8	91.0
5	90.2	91.4	90.6	92.2
6	91.0	92.4	90.8	91.4
7	89.4	90.6	90.0	90.0
8	90.4	90.6	90.8	91.0
9	90.2	91.4	90.6	92.2
10	91.0	92.4	90.8	91.4
11	89.4	90.6	90.0	90.0
12	90.4	90.6	90.8	91.0
13	90.2	91.4	90.6	92.2
14	91.0	92.4	90.8	91.4
15	89.4	90.6	90.0	90.0
16	90.4	90.6	90.8	91.0
Μέση κάλυψη (%)	**90.3**	**91.3**	**90.6**	**91.2**

Πίνακας 2.5: Πιθανότητες κάλυψης της παραμέτρου μ για την κανονική συνάρτηση σύνδεσης στο αληθινό μοντέλο και μέγεθος δείγματος $n = 16$.

Με τις αντίστοιχες τιμές για την ακρίβεια (μήκος) του διαστήματος εμπιστοσύνης για την κάθε μ_i να δίνονται στον Πίνακα 2.6:

Αληθινό μοντέλο	**Συνάρτηση σύνδεσης που χρησιμοποιήθηκε στο προσαρμοσμένο μοντέλο**			
Κανονική	**Κανονική**	**Λογαριθμική**	**Αντίστροφη**	**Ταυτοτική**
Αριθμός παρατήρησης	**Ακρίβεια (μήκος) διαστήματος εμπιστοσύνης**			
1	0.031	0.033	0.032	0.031
2	0.038	0.039	0.038	0.038
3	0.043	0.043	0.043	0.043
4	0.055	0.049	0.052	0.055
5	0.031	0.033	0.032	0.031
6	0.038	0.039	0.038	0.038
7	0.043	0.043	0.043	0.043
8	0.055	0.049	0.052	0.055
9	0.031	0.033	0.032	0.031
10	0.038	0.039	0.038	0.038
11	0.043	0.043	0.043	0.043
12	0.055	0.049	0.052	0.055
13	0.031	0.033	0.032	0.031
14	0.038	0.039	0.038	0.038
15	0.043	0.043	0.043	0.043
16	0.055	0.049	0.052	0.055

Πίνακας 2.6: Ακρίβεια του διαστήματος εμπιστοσύνης της παραμέτρου μ για την κανονική συνάρτηση σύνδεσης στο αληθινό μοντέλο και μέγεθος δείγματος $n=16$.

Τέταρτη περίπτωση

Αληθινό μοντέλο:

$$g(\mu)=\log\mu=-5+2x_1+x_2,$$

με $x_1=(1,1,0,0,1,1,0,0,1,1,0,0,1,1,0,0)$ και $x_2=(1,0,1,0,1,0,1,0,1,0,1,0,1,0,1,0)$,

$n=16$, $\lambda=1$, 500 προσομοιώσεις.

Οι πιθανότητες κάλυψης για την κάθε μ_i δίνονται στον Πίνακα 2.7:

Αληθινό μοντέλο	**Συνάρτηση σύνδεσης που χρησιμοποιήθηκε στο προσαρμοσμένο μοντέλο**			
Λογαριθμική	**Λογαριθμική**	**Κανονική**	**Αντίστροφη**	**Ταυτοτική**
Αριθμός παρατήρησης	**Πιθανότητες κάλυψης**			
1	91.4	100	100	19.4
2	91.8	99.8	82.6	33.8
3	92.0	4.2	17.2	99.0
4	93.4	0.00	0.00	99.2
5	91.4	100	100	19.4
6	91.8	99.8	82.6	33.8
7	92.0	4.2	17.2	99.0
8	93.4	0.00	0.00	99.2
9	91.4	100	100	19.4
10	91.8	99.8	82.6	33.8
11	92.0	4.2	17.2	99.0
12	93.4	0.00	0.00	99.2
13	91.4	100	100	19.4
14	91.8	99.8	82.6	33.8
15	92.0	4.2	17.2	99.0
16	93.4	0.00	0.00	99.2
Μέση κάλυψη (%)	**92.2**	**51.0**	**50.0**	**62.9**

Πίνακας 2.7: Πιθανότητες κάλυψης της παραμέτρου μ για τη λογαριθμική συνάρτηση σύνδεσης στο αληθινό μοντέλο και μέγεθος δείγματος $n = 16$.

Τέλος, οι αντίστοιχες τιμές για την ακρίβεια (μήκος) του διαστήματος εμπιστοσύνης για την κάθε μ_i δίνονται στον Πίνακα 2.8:

Αληθινό μοντέλο	**Συνάρτηση σύνδεσης που χρησιμοποιήθηκε στο προσαρμοσμένο μοντέλο**			
Λογαριθμική	**Λογαριθμική**	**Κανονική**	**Αντίστροφη**	**Ταυτοτική**
Αριθμός παρατήρησης	**Ακρίβεια (μήκος) διαστήματος εμπιστοσύνης**			
1	0.058	0.353	0.573	0.058
2	0.019	0.062	0.039	0.019
3	0.005	0.007	0.007	0.005
4	0.001	0.006	0.005	0.001
5	0.058	0.353	0.573	0.058
6	0.019	0.062	0.039	0.019
7	0.005	0.007	0.007	0.005
8	0.001	0.006	0.005	0.001
9	0.058	0.353	0.573	0.058
10	0.019	0.062	0.039	0.019
11	0.005	0.007	0.007	0.005
12	0.001	0.006	0.005	0.001
13	0.058	0.353	0.573	0.058
14	0.019	0.062	0.039	0.019
15	0.005	0.007	0.007	0.005
16	0.001	0.006	0.005	0.001

Πίνακας 2.8: Ακρίβεια του διαστήματος εμπιστοσύνης της παραμέτρου μ για τη λογαριθμική συνάρτηση σύνδεσης στο αληθινό μοντέλο και μέγεθος δείγματος $n = 16$.

Συμπεράσματα

Παρατηρούμε πως όταν η συνάρτηση σύνδεσης του αληθινού μοντέλου είναι η κανονική (πρώτη περίπτωση), ενδεχόμενη λανθασμένη επιλογή για τη συνάρτηση σύνδεσης κατά την προσαρμογή του IG GLM δεν επηρεάζει τα αποτελέσματα της προσαρμογής. Επίσης, τα διαστήματα εμπιστοσύνης για την παράμετρο μ είναι ιδιαίτερα μικρά σε μήκος και για τις τέσσερις συναρτήσεις σύνδεσης, δηλαδή η προσαρμογή είναι κάθε φορά αρκετά ακριβής. Τα αποτελέσματα ισχυροποιούνται με την αύξηση του μεγέθους του δείγματος (τρίτη περίπτωση).

Αντιθέτως, όταν η σωστή συνάρτηση σύνδεσης είναι η λογαριθμική (δεύτερη περίπτωση), οι πιθανότητες κάλυψης της παραμέτρου μ μειώνονται αισθητά για τα μοντέλα

με τις άλλες τρείς συναρτήσεις σύνδεσης τόσο για μεμονωμένες παρατηρήσεις όσο και συνολικά. Παρόμοια αποτελέσματα είχαμε όταν η συνάρτησης σύνδεσης του αληθινού μοντέλου ήταν η αντίστροφη. Επίσης, τα μήκη των διαστημάτων εμπιστοσύνης είναι μεγαλύτερα, με αποτέλεσμα να μειώνεται η ακρίβεια. Τα αποτελέσματα γίνονται ακόμα πιο αισθητά με την αύξηση του μεγέθους του δείγματος (τέταρτη περίπτωση).

Σύμφωνα με τα αποτελέσματα των παραπάνω προσομοιώσεων, φαίνεται πως σε ένα IG GLM παίζει μεγάλο ρόλο η επιλογή της κατάλληλης συνάρτησης σύνδεσης, ιδιαίτερα όταν αυτή δεν είναι η κανονική. Οι Hardin και Hilbe (σελ. 110, 2001) επιβεβαιώνουν πως "η λογαριθμική συνάρτηση σύνδεσης είναι η καταλληλότερη εναλλακτική της κανονικής"

2.5 Υπόλοιπα για την αντίστροφη Γκαουσιανή κατανομή

Τα υπόλοιπα ή κατάλοιπα (residuals) κατέχουν κεντρικό ρόλο στην προσαρμογή ενός γενικού γραμμικού μοντέλου και χρησιμοποιούνται ευρύτατα για την αξιολόγηση της καταλληλότητας όχι μόνο των γενικευμένων γραμμικών μοντέλων (McCullagh και Nelder, 1989), αλλά και των μοντέλων PH, AL και PO που παρουσιάστηκαν στο πρώτο κεφάλαιο της διατριβής (Collett, 2003). Ο Anscombe (1961) παρουσίασε μία εκτενή περιγραφή ελέγχων καταλληλότητας βασισμένων στα υπόλοιπα. Οι Cox και Snell (1968) έδωσαν ένα γενικό ορισμό των υπολοίπων. Πλέον η βιβλιογραφία έχει επεκταθεί πολύ και εφαρμογές των υπολοίπων συναντώνται σε ολόκληρο το φάσμα της στατιστικής μοντελοποίησης γενικότερα.

Υπό το πλαίσιο της θεωρίας των γραμμικών μοντέλων, το $n \times 1$ διάνυσμα τυχαίων μεταβλητών $\boldsymbol{Y}$ έχει τη μορφή:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \tag{2.31}$$

όπου $\boldsymbol{X}$ γνωστός πίνακας, $\boldsymbol{\beta}$ διάνυσμα άγνωστων παραμέτρων και $\boldsymbol{\varepsilon}$ ένα $n \times 1$ διάνυσμα μη-παρατηρήσιμων, ανεξάρτητων, Κανονικά κατανεμημένων τυχαίων μεταβλητών με μέση τιμή 0 και σταθερή διασπορά. Εάν $\hat{\boldsymbol{\beta}}$ είναι η εκτιμήτρια ελαχίστων τετραγώνων για το $\boldsymbol{\beta}$, τα υπόλοιπα, $\hat{\boldsymbol{\varepsilon}}$, δίνονται από τη σχέση:

$$\boldsymbol{Y} = \boldsymbol{X\hat{\beta}} + \hat{\boldsymbol{\varepsilon}}, \tag{2.32}$$

Δεδομένου πως ο αριθμός των παραμέτρων είναι σχετικά μικρός σε σύγκριση με το *n*, οι περισσότερες από τις ιδιότητες των $\hat{\boldsymbol{\varepsilon}}$ είναι κοντά σε εκείνες των τυχαίων σφαλμάτων $\boldsymbol{\varepsilon}$, δηλαδή θα πρέπει να έχουν περίπου τις ιδιότητες ενός τυχαίου δείγματος από την Κανονική κατανομή.

Για τα γενικευμένα γραμμικά μοντέλα απαιτείται μία επέκταση του ορισμού των υπολοίπων, εφαρμοσμένη σε κάθε κατανομή που μπορεί να αντικαταστήσει την Κανονική (McCullagh και Nelder, 1989).

Στην παράγραφο αυτή, παρουσιάζονται οι διάφορες ειδικές μορφές των γενικών υπολοίπων που είναι διαθέσιμα για την κατανομή IG και περιγράφονται αλγόριθμοι για την παραγωγή τους μέσω της R. Παρουσιάζουμε μία σχέση μεταξύ των Pearson, Anscombe και Deviance υπολοίπων. Ακόμα, αποδεικνύεται πως σε ειδικές περιπτώσεις, τα Anscombe και τα Deviance έχουν πολύ κοντινές τιμές στην IG. Τα διάφορα ευρήματα παρουσιάζονται τόσο θεωρητικά όσο και με τη βοήθεια της R.

2.5.1 Pearson, Anscombe και Deviance υπόλοιπα στο IG GLM μοντέλο

Στην κατανομή IG, τα υπόλοιπα Pearson έχουν την εξής μορφή:

$$r_P = \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i^{3/2}} \tag{2.33}$$

Ένα μειονέκτημα των Pearson υπολοίπων είναι πως η κατανομή των r_P για μη-Κανονικές κατανομές είναι συνήθως ιδιαίτερα λοξή, με αποτέλεσμα να μην έχουν ιδιότητες παρόμοιες με αυτές των Κανονικά κατανεμημένων υπολοίπων.

Τα υπόλοιπα που πρότεινε ο Anscombe (1961) διαμορφώνονται ως εξής για το IG μοντέλο:

$$r_A = \frac{\ln y_i - \ln \hat{\mu}_i}{\hat{\mu}_i^{1/2}} \tag{2.34}$$

Τέλος, εάν θέσουμε $d_i = \frac{\left(y_i - \hat{\mu}_i\right)^2}{\hat{\mu}_i^2 y_i}$, ορίζονται τα Deviance υπόλοιπα ως εξής:

$$r_d = \operatorname{sgn}\left(y_i - \hat{\mu}_i\right)\sqrt{d_i} = \operatorname{sgn}\left(y_i - \hat{\mu}_i\right)\frac{\left(y_i - \hat{\mu}_i\right)}{\hat{\mu}_i\sqrt{y_i}}, \tag{2.35}$$

Παρόλο που τα Anscombe και τα Deviance υπόλοιπα φαίνεται να έχουν πολύ διαφορετικές μορφές για την IG κατανομή, οι τιμές που παίρνουν για συγκεκριμένα y_i και $\hat{\mu}_i$ είναι συχνά πολύ κοντινές.

Θέτοντας $y_i = c\hat{\mu}_i$, οι σχέσεις (2.33), (2.34) και (2.35) δίνουν αντίστοιχα:

$$r_A = \frac{\ln c\hat{\mu}_i - \ln \hat{\mu}_i}{\hat{\mu}_i^{1/2}} = \frac{\ln c}{\hat{\mu}_i^{1/2}} \tag{2.36}$$

Επίσης:

$$r_d = \operatorname{sgn}\left(c\hat{\mu}_i - \hat{\mu}_i\right)\sqrt{\frac{\left(c\hat{\mu}_i - \hat{\mu}_i\right)^2}{\hat{\mu}_i^2 c\hat{\mu}_i}} = \operatorname{sgn}\left(\hat{\mu}_i\left(c-1\right)\right)\sqrt{\frac{\hat{\mu}_i^2\left(c-1\right)^2}{\hat{\mu}_i^2 c\hat{\mu}_i}} \quad \Leftrightarrow$$

$$r_d = \operatorname{sgn}\left(\hat{\mu}_i (c-1)\right)\sqrt{\frac{(c-1)^2}{c\hat{\mu}_i}} \overset{\mu>0}{=} \operatorname{sgn}(c-1)\frac{|c-1|}{\sqrt{c\hat{\mu}_i}}. \qquad (2.37)$$

Τέλος,

$$r_P = \frac{c\hat{\mu}_i - \hat{\mu}_i}{\hat{\mu}_i^{3/2}} = \frac{\hat{\mu}_i (c-1)}{\hat{\mu}_i \hat{\mu}_i^{1/2}} = \frac{c-1}{\hat{\mu}_i^{1/2}}. \qquad (2.38)$$

Στον Πίνακα 2.1 δίνεται μία σύγκριση των τιμών των τριών IG υπολοίπων για επιλεγμένες τιμές της σταθεράς *c*,

	r_A	r_d	r_P
c	$\ln c$	$\operatorname{sgn}(c-1)\frac{\lvert c-1\rvert}{\sqrt{c}}$	$c-1$
0.3	-1.20397	-1.27802	-0.7
0.5	-0.69315	-0.70711	-0.5
0.7	-0.35667	-0.35857	-0.3
0.9	-0.10536	-0.10541	-0.1
1.1	0.09531	0.09535	0.1
1.5	0.40547	0.40825	0.5
2	0.69315	0.70711	1.0
2.5	0.91629	0.94868	1.5
3	1.09861	1.15470	2.0
3.5	1.25276	1.33631	2.5

Πίνακας 2.9: Σύγκριση των διαφόρων IG υπολοίπων

Διαπιστώνουμε λοιπόν πως για μικρές τιμές της σταθεράς *c*, τα υπόλοιπα Anscombe και Deviance λαμβάνουν πολύ κοντινές τιμές.

Από τις σχέσεις (2.34) ή (2.36), έχουμε:

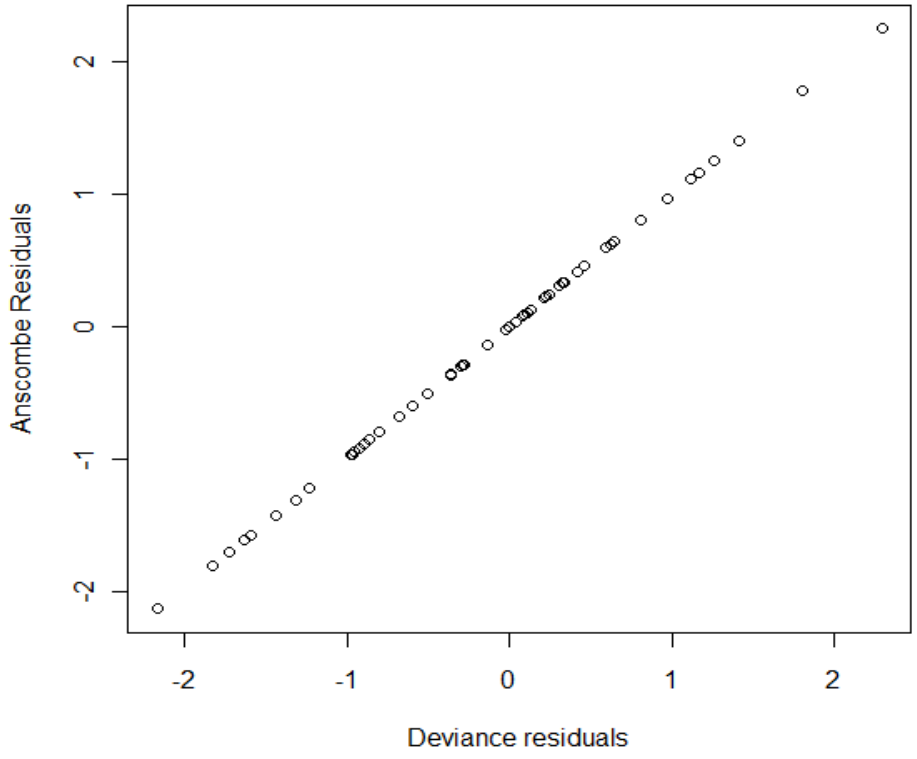
$$r_A = \frac{\ln y_i - \ln \hat{\mu}_i}{\hat{\mu}_i^{1/2}} = \frac{\ln \frac{y_i}{\hat{\mu}_i}}{\hat{\mu}_i^{1/2}}.$$

Επίσης, οι σχέσεις (2.35) ή (2.37) δίνουν:

$$r_d = \operatorname{sgn}\left(y_i - \hat{\mu}_i\right)\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i y_i^{1/2}} = \operatorname{sgn}\left(y_i - \hat{\mu}_i\right)\frac{\frac{y_i}{\hat{\mu}_i} - 1}{y_i^{1/2}}.$$

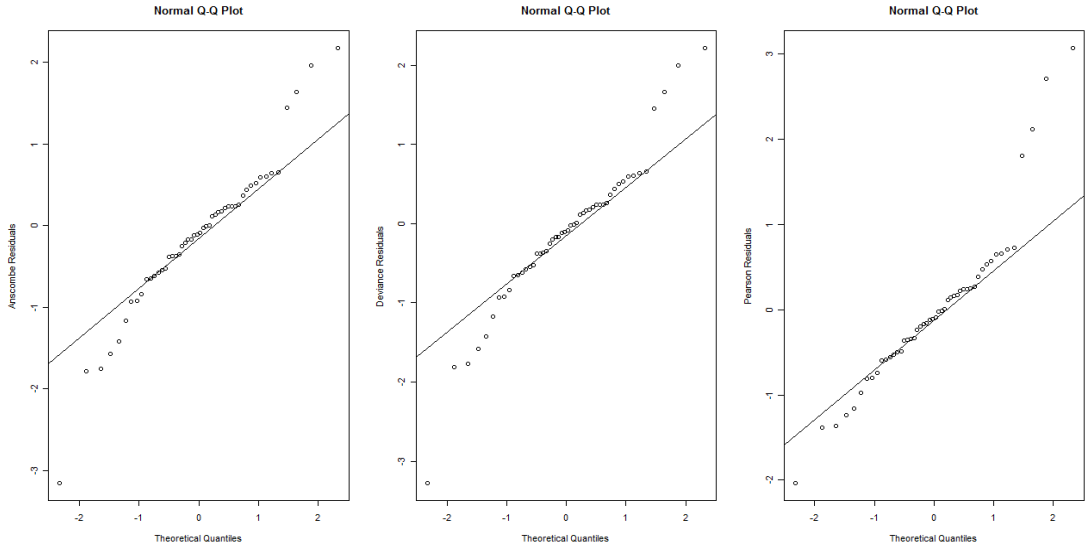
Καθώς $\ln\frac{y_i}{\hat{\mu}_i}=\ln\left(\frac{y_i}{\hat{\mu}_i}-1+1\right)\approx\frac{y_i}{\hat{\mu}_i}-1$ για μικρές τιμές της ποσότητας $\frac{y_i}{\hat{\mu}_i}-1$, οι οποίες επιτυγχάνονται για $\hat{\mu}_i$ κοντά στο y_i, παρατηρούμε πως για το IG μοντέλο τα υπόλοιπα Anscombe και Deviance είναι παρόμοια.

Το Σχήμα 2.3 παρουσιάζει ένα γράφημα των Deviance και των Anscombe υπολοίπων πάνω σε ένα άλλο σύνολο 50 παρατηρήσεων από την $IG(\mu=0.1,\lambda=1)$, στα οποία έχει προσαρμοστεί το γενικό γραμμικό μοντέλο με δύο μεταβλητές, τη X_1 από τη Διωνυμική κατανομή, τη $X_2\sim U(0,1)$ και την κανονική συνάρτηση σύνδεσης. Στο συγκεκριμένο παράδειγμα, η μέση τιμή των ποσοτήτων $\frac{y_i}{\hat{\mu}_i}-1=0.0001177$, για $i=1,...,50$. Αντίστοιχα, η μέση τιμή των Deviance υπολοίπων είναι -0.1302, ενώ των Anscombe -0.1281. Από το Σχήμα 2.3 επιβεβαιώνεται και γραφικά πως τα δύο υπόλοιπα έχουν σχεδόν ίδιες τιμές.



Σχήμα 2.3: Γράφημα των Deviance υπολοίπων ως προς τα Anscombe υπόλοιπα

Το Σχήμα 2.4 παρουσιάζει τα αντίστοιχα Q-Q γραφήματα για τα παραπάνω υπόλοιπα. Τα τρία γραφήματα του Σχήματος 2.4 επιβεβαιώνουν την πεποίθηση πως τα Anscombe και Deviance υπόλοιπα έχουν παρόμοιες τιμές. Επίσης, φαίνεται πως τα Pearson ίσως αποτελούν καλύτερη επιλογή για την περίπτωση που μελετάμε, παρόλο που και τα άλλα δύο υπόλοιπα δε φαίνεται να είναι ακατάλληλα ως επιλογή για το συγκεκριμένο μοντέλο.



Σχήμα 2.4: QQ-plots για τα υπόλοιπα Deviance, Anscombe και Pearson

2.5.2 Cox - Snell και Martingale υπόλοιπα στο IG GLM μοντέλο

Όπως είδαμε και στην Παράγραφο 1.2.5, ένα είδος υπολοίπων που χρησιμοποιείται ευρύτατα στην ανάλυση δεδομένων διάρκειας ζωής και κυρίως σε παραμετρικά μοντέλα είναι το υπόλοιπο Cox-Snell. Με τη βοήθεια της εκτιμήτριας Nelson-Aalen της σχέσης (1.8), η σχέση (1.12) μας δίνει το Cox-Snell υπόλοιπο για την *i* - μονάδα:

$$r_{Ci} = -\ln \hat{S}(y_i) \tag{2.39}$$

Τα Cox-Snell υπόλοιπα έχουν σχετικά ανόμοιες ιδιότητες με εκείνες των υπολοίπων που χρησιμοποιούνται στη γραμμική παλινδρόμηση. Συγκεκριμένα, δεν είναι κατανεμημένα γύρω από το 0 και δεν επιτρέπεται να είναι αρνητικά. Τέλος, καθώς θεωρείται πως έχουν εκθετική κατανομή όταν ένα κατάλληλο μοντέλο προσαρμόζεται στα δεδομένα, παρουσιάζουν μία αρκετά λοξή κατανομή με μέση τιμή και διασπορά για την *i*-μονάδα, ίση με 1 (Collett, 2003).

Στη συνέχεια, με τη βοήθεια των Cox-Snell υπολοίπων μπορούμε να ορίσουμε τα **martingale υπόλοιπα** (martingale residuals), ως εξής:

$$r_{M_i} = \delta_i - r_{Ci} = \delta_i + \ln \hat{S}\left(y_i\right) \tag{2.40}$$

Η τιμή των martingale υπολοίπων είναι μεταξύ -∞ και μονάδας με την τιμή των υπολοίπων για τα αποκομμένα δεδομένα ($\delta_i = 0$) να είναι εξ' ορισμού αρνητική. Αποδεικνύεται, πως τα υπόλοιπα αυτά αθροίζουν στο μηδέν και ότι σε μεγάλα δείγματα τα martingale υπόλοιπα είναι ασυσχέτιστα μεταξύ τους και έχουν αναμενόμενη τιμή ίση με μηδέν. Ένα πλήθος συγγραφέων έχει ασχοληθεί με martingale μεθόδους από τις οποίες προκύπτουν εναλλακτικά

τα παραπάνω υπόλοιπα. Ενδεικτικά, αναφέρουμε τους Andersen et al. (1993) και Therneau και Grambsch (2000).

Οι Therneau, Grambsch και Fleming (1990) εισήγαγαν τα Deviance υπόλοιπα στην περιοχή των στοχαστικών ανελίξεων με τη χρησιμοποίηση των martingale υπολοίπων. Για παράδειγμα, το martingale υπόλοιπο τύπου deviance για το μοντέλο του Cox, στο οποίο δεν υπάρχουν χρονικά μεταβαλλόμενες μεταβλητές, δίνεται από τη σχέση:

$$r_{D_i} = \operatorname{sgn}\left(r_{M_i}\right)\left\{-2\left[r_{M_i} + \delta_i \ln\left(\delta_i - r_{M_i}\right)\right]\right\}^{1/2}, \tag{2.41}$$

Για τη συγκεκριμένη παραμετρική περίπτωση που μελετάμε, το υπόλοιπο martingale, r_{M_i}, δίνεται από τη σχέση (2.40).

2.5.3 Δύο αλγόριθμοι για τη μελέτη των διαφόρων μορφών υπολοίπων

Στη συνέχεια παρουσιάζουμε δύο αλγόριθμους που κατασκευάστηκαν στην R προκειμένου να μελετήσουμε τις διάφορες μορφές υπολοίπων που αναπτύξαμε παραπάνω για την IG κατανομή.

Σκιαγράφηση 1ου αλγορίθμου

1. Επιλογή των τιμών των παραμέτρων μ και λ της κατανομής IG, καθώς και του μεγέθους του δείγματος.
2. Ορισμός μεταβλητών: Κατασκευή δύο μεταβλητών, $X_1 \sim U(0,1)$ και X_2 από τη Διωνυμική κατανομή.
3. Παραγωγή των παρατηρήσεων y_i από την κατανομή $IG(\mu, \lambda)$ για τις επιλεγμένες τιμές των παραμέτρων μ και λ.
4. Προσαρμογή του γενικευμένου γραμμικού μοντέλου με κατάλληλη συνάρτηση σύνδεσης.
5. Ορισμός των διαφόρων τύπων υπολοίπων που παρουσιάστηκαν στην προηγούμενη παράγραφο, χρησιμοποιώντας τις σχέσεις (2.33), (2.34), (2.35), (2.39), (2.40) και (2.41). Διάταξη των έξι διανυσμάτων με τις τιμές των υπολοίπων σε αύξουσα σειρά.
6. Q-Q γραφήματα για κάθε έναν τύπο υπολοίπου.

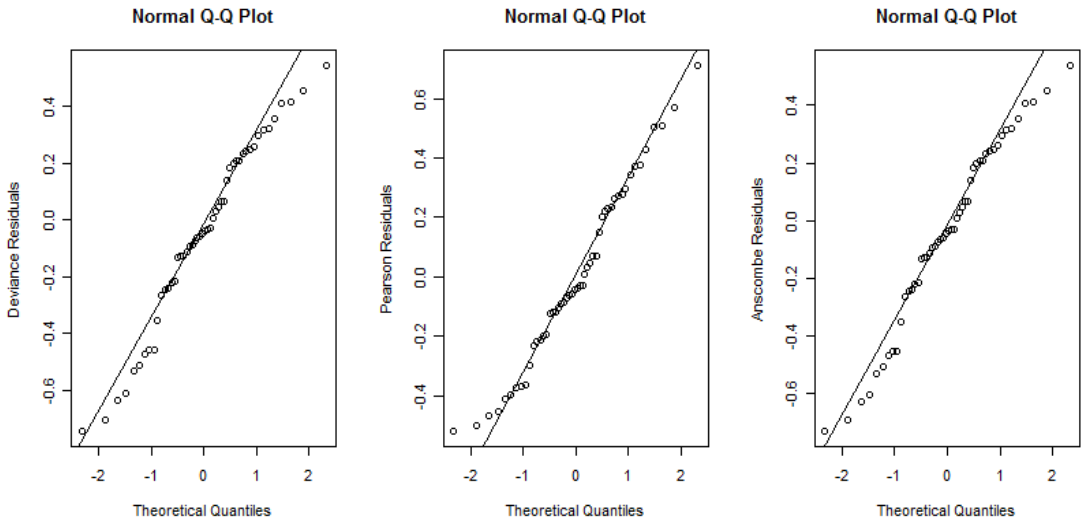
Σκιαγράφηση 2ου αλγορίθμου

1. Βήματα 1 – 5 ίδια με πριν.
2. Αποθήκευση της τιμής του υπολοίπου κάποιας παρατήρησης (τυχαία επιλογή παρατήρησης, π.χ. της 25ης) σε κάθε επανάληψη, για κάθε έναν από τους έξι τύπους υπολοίπων.
3. Επανάληψη των βημάτων 3, 4 και 5 1,000 φορές.
4. Q-Q γραφήματα για κάθε έναν τύπο υπολοίπου.

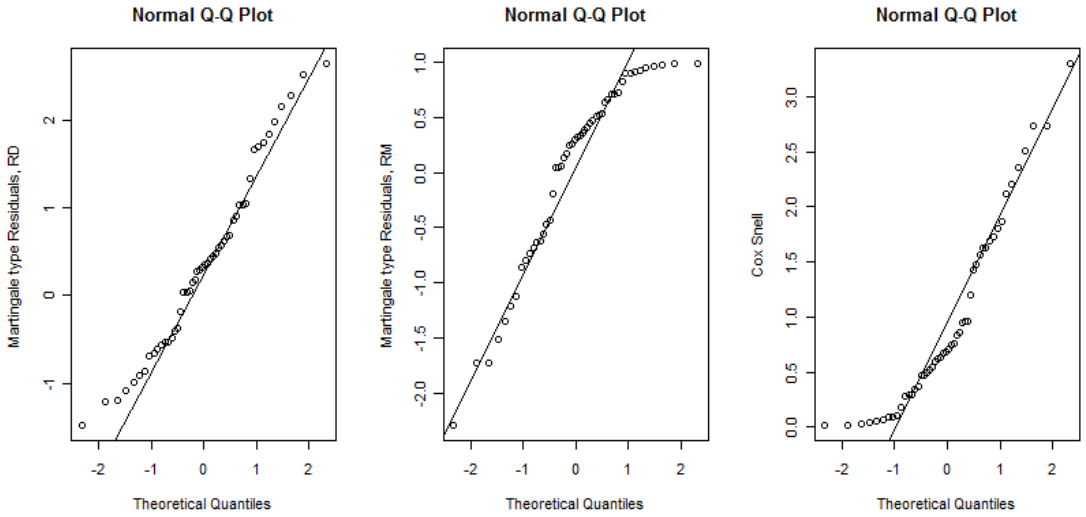
2.5.4 Αποτελέσματα προσομοιώσεων

Σύμφωνα με τον 1ο αλγόριθμο που παρουσιάσαμε στην προηγούμενη παράγραφο, για τα Σχήματα 2.5 έως 2.12 που ακολουθούν, έχουμε προσαρμόσει σε IG δεδομένα το γενικό γραμμικό μοντέλο με δύο μεταβλητές, τη X_1 από τη Διωνυμική κατανομή, τη $X_2 \sim U(0,1)$ και έχουμε χρησιμοποιήσει την κανονική συνάρτηση σύνδεσης. Έγιναν δύο διαφορετικές επιλογές για το μέγεθος του δείγματος $(n=50 \text{ και } n=150)$ και δύο επιλογές για τα ζευγάρια των παραμέτρων μ και λ της IG κατανομής $(1,10)$ και $(1,7)$.

Τα Σχήματα 2.5 και 2.6 παρουσιάζουν γραφήματα των έξι διαφορετικών τύπων υπολοίπων που ορίσαμε στις Παραγράφους 2.5.1 και 2.5.2. σε ένα σύνολο 50 παρατηρήσεων από την $IG(\mu=1, \lambda=10)$.



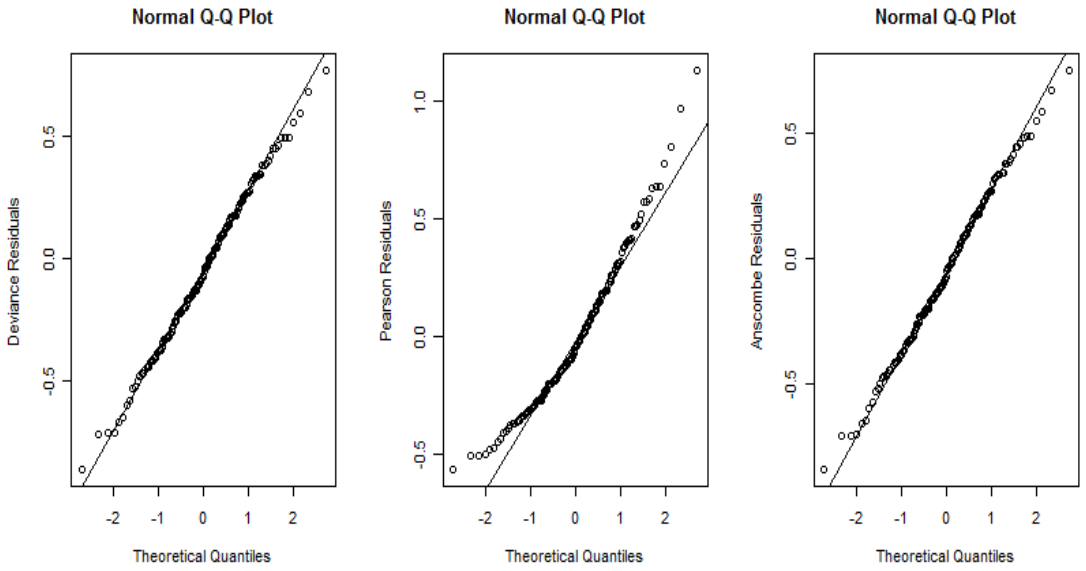
Σχήμα 2.5: Γραφήματα των Deviance, Pearson και Anscombe υπολοίπων για *n* = 50 και $(\mu, \lambda)=(1,10)$



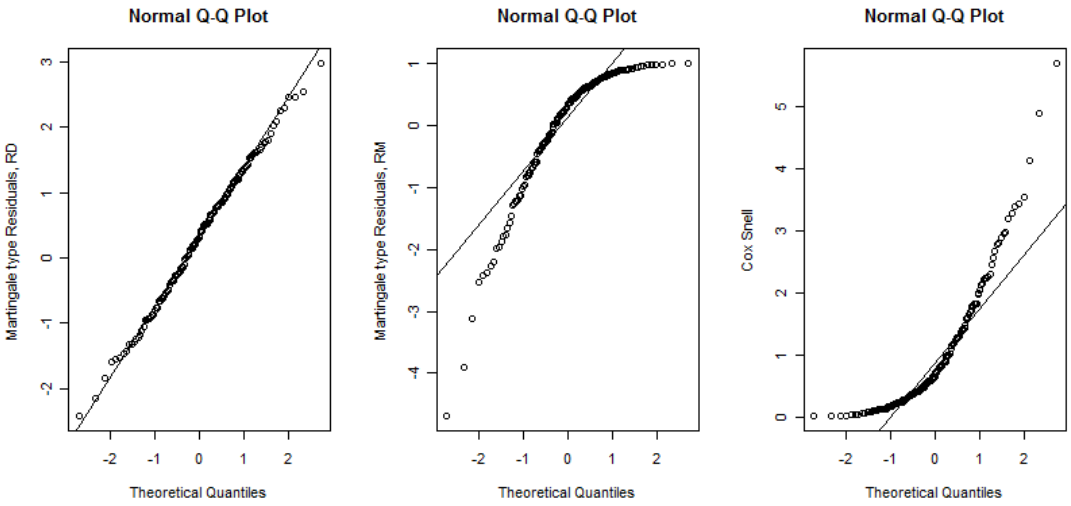
Σχήμα 2.6: Γραφήματα των r_{D_i}, r_{M_i} και Cox-Snell **υπολοίπων για** $n = 50$ **και** $(\mu, \lambda) = (1, 10)$

Για τις συγκεκριμένες επιλογές παραμέτρων της κατανομής, φαίνεται πως μόνο τέσσερις από τους έξι τύπους υπολοίπων είναι κατάλληλοι για το μοντέλο IG. Τα Q-Q γραφήματα για τα martingale τύπου υπόλοιπα r_{M_i} και για τα Cox-Snell υπόλοιπα υποδεικνύουν πως δεν ικανοποιούνται οι προϋποθέσεις για την IG κατανομή.

Τα Σχήματα 2.7 και 2.8 παρουσιάζουν γραφήματα των έξι διαφορετικών τύπων υπολοίπων που ορίσαμε στις Παραγράφους 2.5.1 και 2.5.2 σε ένα σύνολο 150 παρατηρήσεων από την $IG(\mu = 1, \lambda = 10)$.



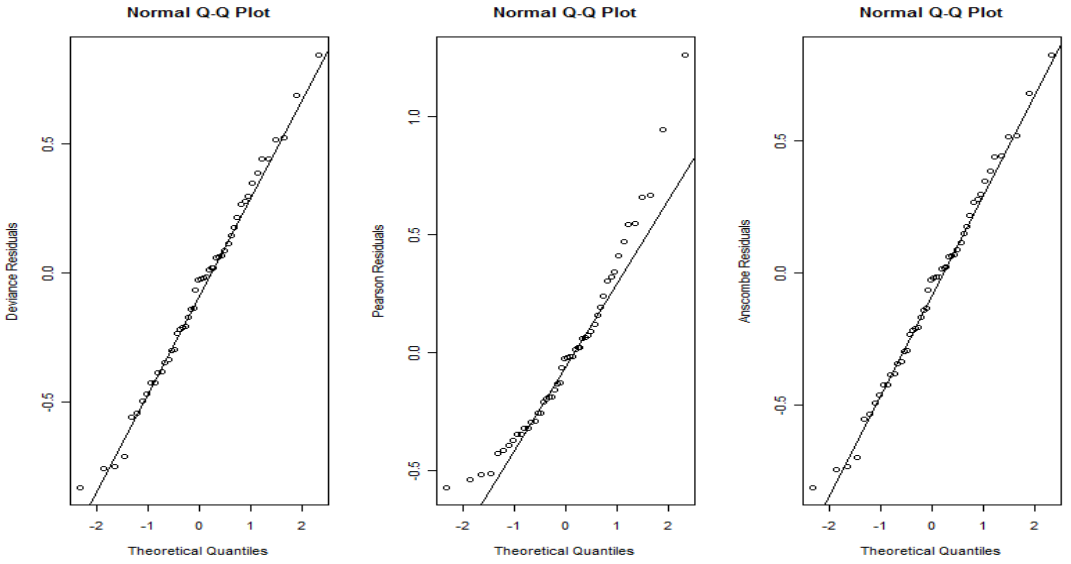
Σχήμα 2.7: Γραφήματα των Deviance, Pearson και Anscombe υπολοίπων για $n = 150$ **και** $(\mu, \lambda) = (1, 10)$



Σχήμα 2.8: Γραφήματα των r_{D_i}, r_{M_i} και Cox-Snell υπολοίπων για $n = 150$ και $(\mu, \lambda) = (1,10)$

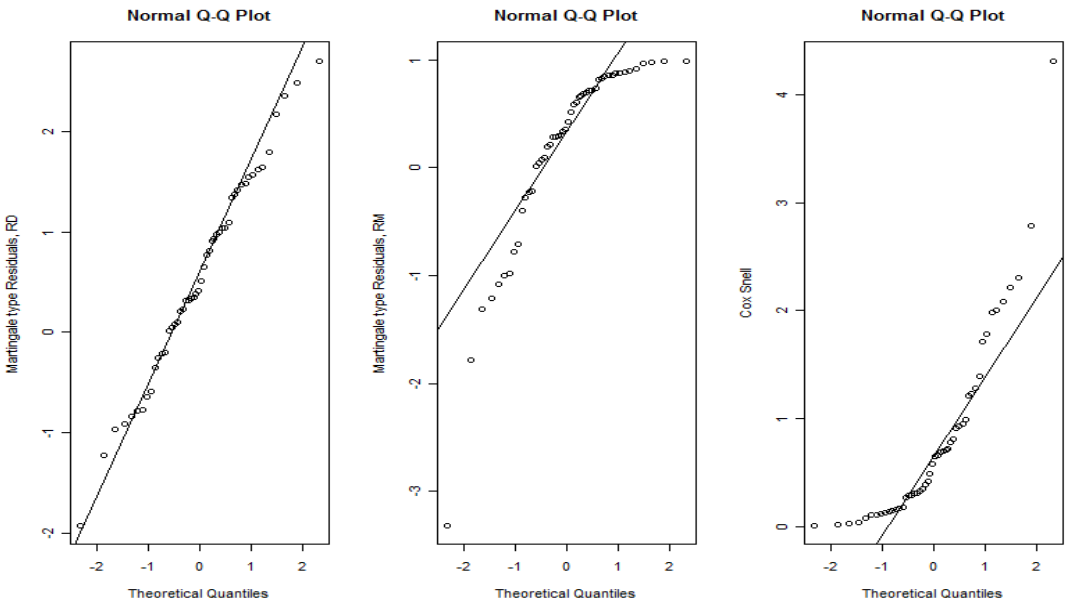
Τα συμπεράσματα που βγάλαμε από τα Σχήματα 2.5 και 2.6 φαίνεται να ισχυροποιούνται με την αύξηση του μεγέθους του δείγματος.

Τα Σχήματα 2.9 και 2.10 παρουσιάζουν γραφήματα των 6 διαφορετικών τύπων υπολοίπων που ορίσαμε στις Παραγράφους 2.5.1 και 2.5.2 σε ένα σύνολο 50 παρατηρήσεων από την $IG(\mu = 1, \lambda = 7)$.



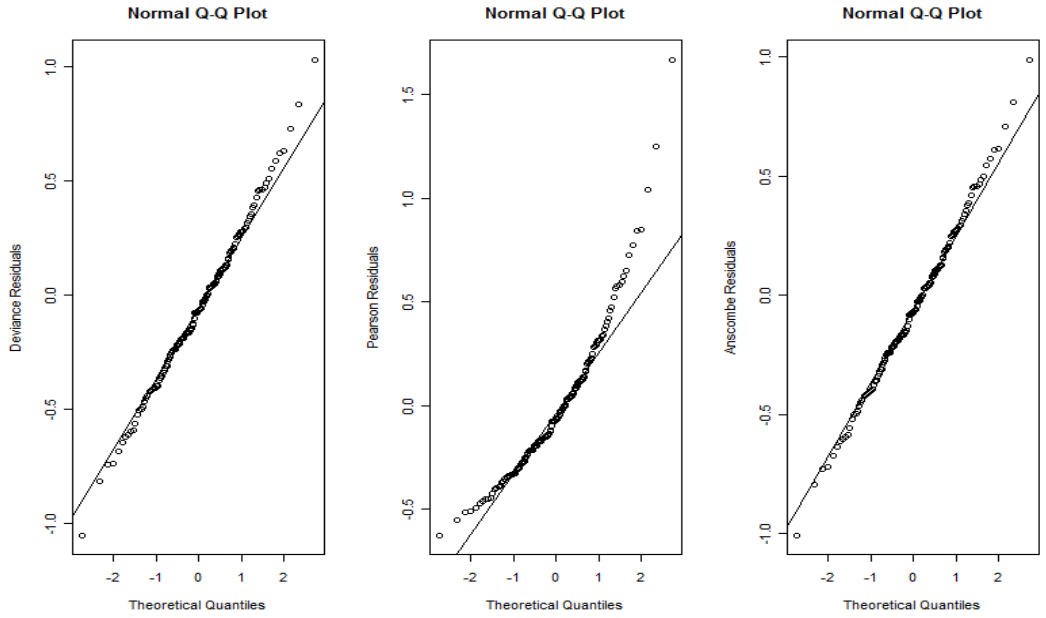
Σχήμα 2.9: Γραφήματα των Deviance, Pearson και Anscombe υπολοίπων για $n = 50$ και $(\mu, \lambda) = (1,7)$

Για τις συγκεκριμένες επιλογές παραμέτρων της κατανομής, φαίνεται πως μόνο τρείς από τους έξι τύπους υπολοίπων είναι κατάλληλοι για το μοντέλο IG. Συγκεκριμένα, τα Q-Q γραφήματα για τα martingale τύπου υπόλοιπα r_{M_i}, για τα Cox-Snell υπόλοιπα, αλλά ίσως και για τα Pearson υπόλοιπα, υποδεικνύουν πως δεν ικανοποιούνται οι προϋποθέσεις για την IG κατανομή.

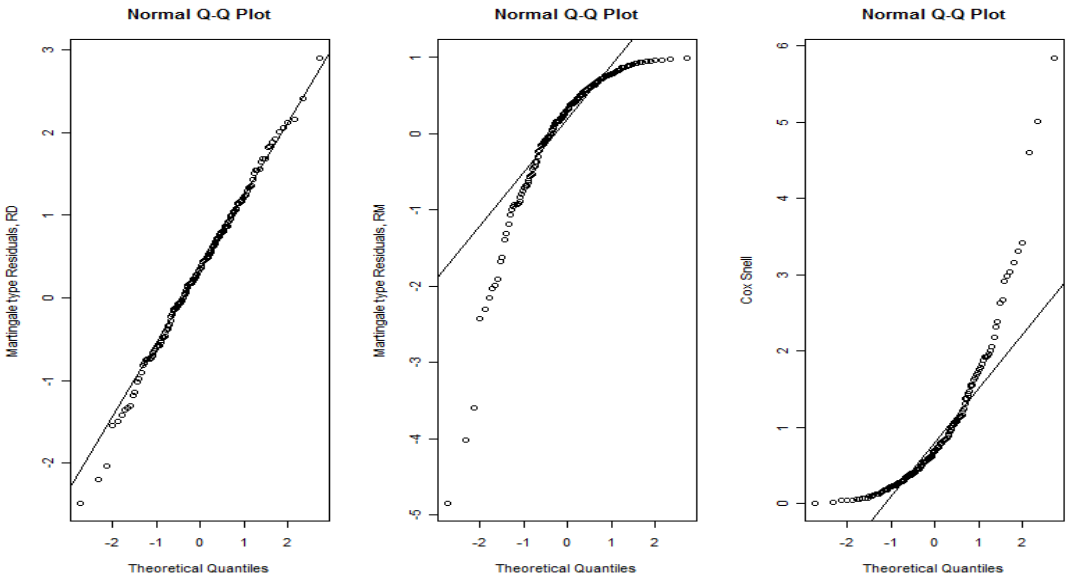


Σχήμα 2.10: Γραφήματα των r_{D_i}, r_{M_i} και Cox-Snell **υπολοίπων για** $n = 50$ **και** $(\mu, \lambda) = (1,7)$

Τα Σχήματα 2.11 και 2.12 παρουσιάζουν γραφήματα των έξι διαφορετικών τύπων υπολοίπων που ορίσαμε στις Παραγράφους 2.5.1 και 2.5.2 σε ένα σύνολο 150 παρατηρήσεων από την $IG(\mu = 1, \lambda = 7)$.



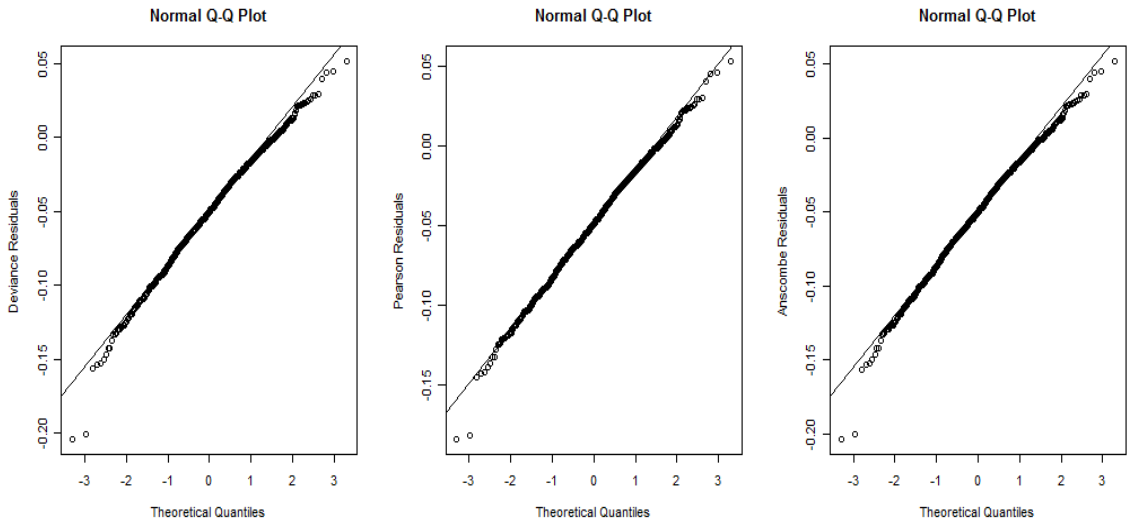
Σχήμα 2.11: Γραφήματα των Deviance, Pearson και Anscombe υπολοίπων για n **=150 και** $(\mu, \lambda) = (1,7)$



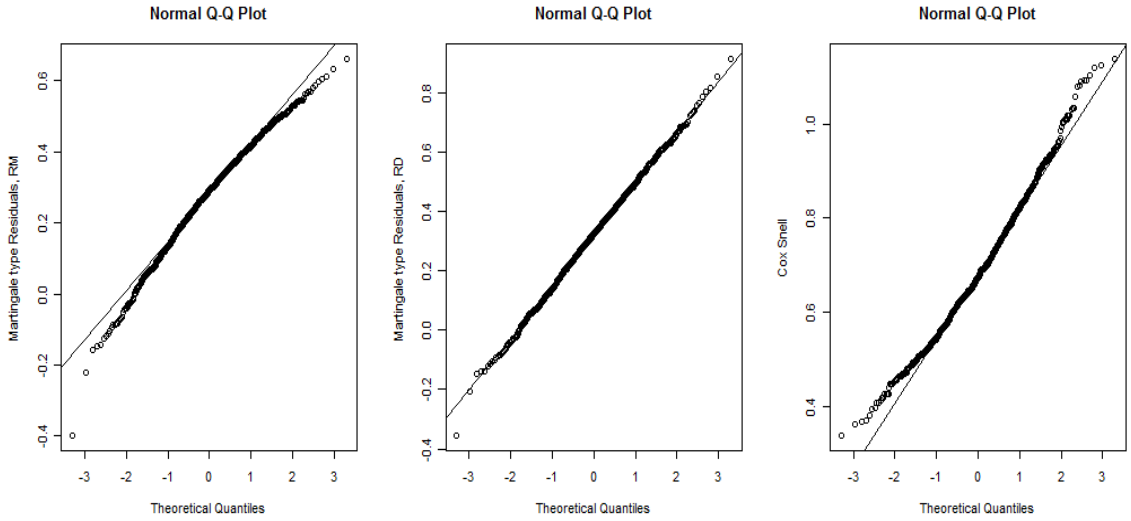
Σχήμα 2.12: Γραφήματα των r_{D_i}, r_{M_i} και Cox-Snell **υπολοίπων για** $n = 150$ **και** $(\mu, \lambda) = (1,7)$

Τα συμπεράσματα που βγάλαμε από τα Σχήματα 2.11 και 2.12 φαίνεται να ισχυροποιούνται με την αύξηση του μεγέθους του δείγματος. Βέβαια, να παρατηρήσουμε πως για τις συγκεκριμένες επιλογές των παραμέτρων, η κατανομή IG παρουσιάζει μεγάλη κύρτωση με πολύ παχιές ουρές.

Σύμφωνα με τον 2ο αλγόριθμο που παρουσιάσαμε στην προηγούμενη παράγραφο, για τα Σχήματα 2.13 και 2.14 που ακολουθούν, έχουμε προσαρμόσει σε 50 IG δεδομένα το γενικό γραμμικό μοντέλο με δύο μεταβλητές, τη X_1 από τη Διωνυμική κατανομή και τη $X_2 \sim U(0,1)$. Χρησιμοποιήσαμε την κανονική συνάρτηση σύνδεσης και οι προσομοιώσεις έγιναν για επιλογή των παραμέτρων μ και λ της IG κατανομής ίση με (1,10). Πραγματοποιήθηκαν 1,000 επαναλήψεις του αλγορίθμου και σε κάθε μια επανάληψη αποθηκεύσαμε τα διάφορα υπόλοιπα μίας μόνο παρατήρησης $(t = 25)$. Στα Σχήματα 2.13 και 2.14 παρουσιάζονται τα Q-Q γραφήματα των 6 διαφορετικών τύπων υπολοίπων, τα οποία ορίσαμε στις Παραγράφους 2.5.1 και 2.5.2.



Σχήμα 2.13: Γραφήματα των Deviance, Pearson και Anscombe υπολοίπων για *n* =50 και $(\mu,\lambda)=(1,10)$



Σχήμα 2.14: Γραφήματα των r_{D_i}, r_{M_i} και Cox-Snell υπολοίπων για *n* = 50 και $(\mu,\lambda)=(1,10)$

Ενδείξεις για αναχώρηση από την κανονικότητα, έχουμε για μια ακόμη φορά για τα τύπου martingale υπόλοιπα r_{M_i}, ίσως και για τα Cox-Snell. Παρόμοια αποτελέσματα έχουμε για την περίπτωση όπου τα δεδομένα είναι από την $IG(\mu=1,\lambda=7)$.

Συμπεραίνουμε, λοιπόν, πως οι παραπάνω τύποι υπολοίπων που εξετάσαμε (Deviance, Pearson, Anscombe και τύπου martingale υπόλοιπα r_{D_i}) και χρησιμοποιούνται ευρύτατα τόσο σε γραμμικά όσο και σε μη γραμμικά μοντέλα, μπορούν να χρησιμοποιηθούν και στην περίπτωση ενός γενικευμένου γραμμικού μοντέλου για την κατανομή IG. Από την άλλη, τα υπόλοιπα τύπου Cox-Snell και martingale r_{M_i}, φαίνεται να είναι μη κατάλληλα.

2.6 Εκτίμηση παραμέτρων για την IG

Στην παρούσα παράγραφο, παρουσιάζονται οι εκτιμήτριες μέγιστης πιθανοφάνειας για τις παραμέτρους μ και λ της IG κατανομής.

2.6.1 Εκτιμήτριες μέγιστης πιθανοφάνειας

Έστω ένα τυχαίο δείγμα $X_1, X_2, ..., X_n \sim \text{IG}(\mu, \lambda)$. Η συνάρτηση πιθανοφάνειας είναι:

$$L = \left(\frac{\lambda}{2\pi}\right)^{n/2} \left(\prod_{i=1}^{n} x_i^{-3/2}\right) \exp\left[-\lambda \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\mu^2 x_i}\right], \ \mu > 0, \ \lambda > 0 \tag{2.42}$$

και οι αντίστοιχες εκτιμήτριες μέγιστης πιθανοφάνειας (Ε.Μ.Π.) για τις παραμέτρους μ και λ είναι:

$$\begin{aligned} \hat{\mu} &= \bar{X} \\ \frac{1}{\hat{\lambda}} &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{X_1} - \frac{1}{\bar{X}}\right) \end{aligned}, \tag{2.43}$$

όπου $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

Επίσης, $\hat{\mu} \sim IG(\mu, n\lambda)$ και η ελεγχοσυνάρτηση $\frac{n\lambda}{\hat{\lambda}}$ ακολουθεί την κατανομή χ^2_{n-1} (Tweedie, 1957a).

2.7 Έλεγχοι υπόθεσης για την IG κατανομή

Η Davis (1980) χρησιμοποίησε τον έλεγχο του λόγου των πιθανοφανειών προκειμένου να ελέγξει υποθέσεις που αφορούν τις παραμέτρους μ και λ της IG κατανομής για τις περιπτώσεις ενός και δύο δειγμάτων δεδομένων.

Στη συνέχεια, παρουσιάζονται δύο έλεγχοι υπόθεσης για τις παραμέτρους της IG κατανομής για την περίπτωση των δύο δειγμάτων δεδομένων, με τη βοήθεια του λόγου των πιθανοφανειών. Εισάγεται η έννοια της άτυπης τιμής (outlier). Οι παραπάνω έλεγχοι προσαρμόζονται για την περίπτωση εντοπισμού άτυπων τιμών των παραμέτρων της IG σε δεδομένα *k* ανεξάρτητων μονάδων με επαναλαμβανόμενα ανεξάρτητα γεγονότα ανά μονάδα. Οι προτεινόμενοι έλεγχοι μελετώνται με τη βοήθεια της R. Τέλος, κατασκευάζονται αντίστοιχοι έλεγχοι για μία εναλλακτική παραμέτρηση της IG και παρουσιάζεται μία εφαρμογή σε ένα πρόβλημα πραγματικών συνθηκών.

2.7.1 Έλεγχος υπόθεσης για την παράμετρο μ της IG

Έστω ένα δείγμα n παρατηρήσεων $X_i \sim IG(\mu,\lambda)$ και ένα δείγμα m παρατηρήσεων $Y_i \sim IG(\nu,\lambda)$ με λ άγνωστο, αλλά κοινό μεταξύ των δύο δειγμάτων. Θέλουμε να προχωρήσουμε στον έλεγχο της υπόθεσης:

$H_0: \mu = \nu$

$H_1: \mu \neq \nu$,

Η από κοινού σ.π.π ή συνάρτηση πιθανοφάνειας δίνεται από τον τύπο:

$$L(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}) = L(\mu,\nu,\lambda|\, x_1,\ ...,x_n, y_1,...,y_n) = f_{\boldsymbol{x},\boldsymbol{y}} = \prod_{i=1}^{n} f_{x_i}(x_i,\boldsymbol{\theta}) \cdot \prod_{j=1}^{m} f_{y_j}(y_j,\boldsymbol{\theta}) =$$

$$= \prod_{i=1}^{n}\left(\frac{\lambda}{2\pi x_i^3}\right)^{1/2} \exp\left(-\frac{\lambda(x_t-\mu)^2}{2\mu^2 x_i}\right) \prod_{j=1}^{m}\left(\frac{\lambda}{2\pi y_j^3}\right)^{1/2} \exp\left(-\frac{\lambda(y_j-\nu)^2}{2\nu^2 y_j}\right) =$$

$$L(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}) = \left(\frac{\lambda}{2\pi}\right)^{\frac{n+m}{2}} \prod_{i=1}^{n}\frac{1}{x_i^{3/2}} \prod_{j=1}^{m}\frac{1}{y_j^{3/2}} \exp\left(-\frac{\lambda}{2}\left[\sum_{i=1}^{n}\frac{(x_t-\mu)^2}{\mu^2 x_i} + \sum_{j=1}^{m}\frac{(y_j-\nu)^2}{\nu^2 y_j}\right]\right) \qquad (2.44)$$

Υπό τη μηδενική υπόθεση

Η συνάρτηση πιθανοφάνειας της σχέσης (2.44) γίνεται:

$$L(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}) = \left(\frac{\lambda}{2\pi}\right)^{\frac{n+m}{2}} \prod_{i=1}^{n}\frac{1}{x_i^{3/2}} \prod_{j=1}^{m}\frac{1}{y_j^{3/2}} \exp\left(-\frac{\lambda}{2}\left[\sum_{i=1}^{n}\frac{(x_t-\mu)^2}{\mu^2 x_i} + \sum_{j=1}^{m}\frac{(y_j-\mu)^2}{\mu^2 y_j}\right]\right)$$

Λογαριθμίζοντας την παραπάνω σχέση έχουμε,

$$\ln L(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}) = \ln L = \frac{n+m}{2}\ln\frac{\lambda}{2\pi} - \frac{3}{2}\left[\sum_i \ln x_i + \sum_j \ln y_j\right] - \frac{\lambda}{2}\left[\sum_{i=1}^{n}\frac{(x_t-\mu)^2}{\mu^2 x_i} + \sum_{j=1}^{m}\frac{(y_j-\mu)^2}{\mu^2 y_j}\right]$$

Τέλος, με μερική παραγώγιση προκύπτουν οι εκτιμήτριες μέγιστης πιθανοφάνειας:

$$\frac{\partial \ln L(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y})}{\partial \mu} = 0 \quad \Leftrightarrow$$

$$\frac{\partial}{\partial \mu}\left[\frac{n+m}{2}\ln\frac{\lambda}{2\pi} - \frac{3}{2}\left[\sum_i \ln x_t + \sum_j \ln y_j\right] - \frac{\lambda}{2}\left[\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\mu^2 x_i} + \sum_{j=1}^{m}\frac{(y_j-\mu)^2}{\mu^2 y_j}\right]\right] = 0 \quad \Leftrightarrow$$

$$\frac{\partial}{\partial\mu}\left[\frac{\sum_{i=1}^{n}x_i}{\mu^2}-\frac{2n}{\mu}+\sum_{i=1}^{n}\frac{1}{x_i}+\frac{\sum_{j=1}^{m}y_j}{\mu^2}-\frac{2m}{\mu}+\sum_{j=1}^{m}\frac{1}{y_j}\right]=0\Leftrightarrow$$

$$-2\mu^{-3}\left(\sum_{i=1}^{n}x_i+\sum_{j=1}^{m}y_j\right)+2\mu^{-2}\left(n+m\right)=0$$

$$\tilde{\mu}=\frac{n\bar{x}+m\bar{y}}{n+m} \tag{2.45}$$

$$\frac{\partial\ln L\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)}{\partial\lambda}=0\quad\Leftrightarrow$$

$$\frac{\partial}{\partial\lambda}\left[\frac{n+m}{2}\ln\frac{\lambda}{2\pi}-\frac{3}{2}\left[\sum_i\ln x_i+\sum_j\ln y_j\right]-\frac{\lambda}{2}\left[\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{\mu^2x_i}+\sum_{j=1}^{m}\frac{\left(y_j-\mu\right)^2}{\mu^2y_j}\right]\right]=0\quad\Leftrightarrow$$

$$\frac{n+m}{\lambda}-\left[\sum_{i=1}^{n}\frac{\left(x_t-\mu\right)^2}{\mu^2x_t}+\sum_{j=1}^{m}\frac{\left(y_j-\mu\right)^2}{\mu^2y_j}\right]=0\quad\Leftrightarrow$$

$$\tilde{\lambda}=\frac{n+m}{\sum_{i=1}^{n}\frac{\left(x_i-\tilde{\mu}\right)^2}{\tilde{\mu}^2x_i}+\sum_{j=1}^{m}\frac{\left(y_j-\tilde{\mu}\right)^2}{\tilde{\mu}^2y_j}} \tag{2.46}$$

Υπό την εναλλακτική υπόθεση

Η συνάρτηση πιθανοφάνειας της σχέσης (2.44) γίνεται:

$$L\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)=\left(\frac{\lambda}{2\pi}\right)^{\frac{n+m}{2}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left(-\frac{\lambda}{2}\left[\sum_{i=1}^{n}\frac{\left(x_t-\mu\right)^2}{\mu^2x_i}+\sum_{j=1}^{m}\frac{\left(y_j-\mathrm{v}\right)^2}{\mathrm{v}^2y_j}\right]\right)$$

Λογαριθμίζοντας την παραπάνω σχέση, έχουμε:

$$\ln L\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)=\ln L=\frac{n+m}{2}\ln\frac{\lambda}{2\pi}-\frac{3}{2}\left[\sum_i\ln x_i+\sum_j\ln y_j\right]-\frac{\lambda}{2}\left[\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{\mu^2x_i}+\sum_{j=1}^{m}\frac{\left(y_j-\mathrm{v}\right)^2}{\mathrm{v}^2y_j}\right]$$

Τέλος, με μερική παραγώγιση προκύπτουν οι εκτιμήτριες μέγιστης πιθανοφάνειας:

$$\frac{\partial\ln L\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)}{\partial\mu}=0\quad\Leftrightarrow$$

$$\frac{\partial}{\partial\mu}\left[\frac{n+m}{2}\ln\frac{\lambda}{2\pi}-\frac{3}{2}\left[\sum_i \ln x_i+\sum_j \ln y_j\right]-\frac{\lambda}{2}\left[\sum_{i=1}^{n}\frac{(x_t-\mu)^2}{\mu^2 x_i}+\sum_{j=1}^{m}\frac{(y_j-\nu)^2}{\nu^2 y_j}\right]\right]=0 \quad \Leftrightarrow$$

$$\frac{\partial}{\partial\mu}\left[\frac{\sum_{i=1}^{n}x_i}{\mu^2}-\frac{2n}{\mu}+\sum_{i=1}^{n}\frac{1}{x_i}\right]=0 \quad \Leftrightarrow -2\mu^{-3}\left(\sum_{i=1}^{n}x_i\right)+2\mu^{-2}n=0 \quad \Leftrightarrow$$

$$\hat{\mu}=\frac{\sum_{i=1}^{n}x_i}{n}=\bar{x} \tag{2.47}$$

Όμοια,

$$\frac{\partial \ln L\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)}{\partial \nu}=0 \quad \Leftrightarrow$$

$$\hat{\nu}=\frac{\sum_{j=1}^{m}y_j}{m}=\bar{y} \tag{2.48}$$

Τέλος, για την Ε.Μ.Π. της παραμέτρου λ έχουμε:

$$\frac{\partial \ln L\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)}{\partial\lambda}=0 \quad \Leftrightarrow$$

$$\frac{\partial}{\partial\lambda}\left[\frac{n+m}{2}\ln\frac{\lambda}{2\pi}-\frac{3}{2}\left[\sum_i \ln x_i+\sum_j \ln y_j\right]-\frac{\lambda}{2}\left[\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\mu^2 x_i}+\sum_{j=1}^{m}\frac{(y_j-\nu)^2}{\nu^2 y_j}\right]\right]=0 \quad \Leftrightarrow$$

$$\frac{n+m}{\lambda}-\left[\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{\mu^2 x_i}+\sum_{j=1}^{m}\frac{(y_j-\nu)^2}{\nu^2 y_j}\right]=0$$

Αντικαθιστώντας στην παραπάνω σχέση τις ποσότητες $\hat{\mu}$ και $\hat{\nu}$ των σχέσεων (2.47) και (2.48), καταλήγουμε στην Ε.Μ.Π. για την παράμετρο λ:

$$\hat{\lambda}=\frac{n+m}{\sum_{i=1}^{n}\frac{(x_i-\hat{\mu})^2}{\hat{\mu}^2 x_i}+\sum_{j=1}^{m}\frac{(y_j-\hat{\nu})^2}{\hat{\nu}^2 y_j}}=\frac{n+m}{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)} \tag{2.49}$$

Προκειμένου να χρησιμοποιήσουμε τον έλεγχο του λόγου των πιθανοφανειών, έχουμε τα εξής:

$$\sup_{\boldsymbol{\theta}|H_0} p\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)=\sup_{\boldsymbol{\theta}|H_0} p\left(\mu,\nu,\lambda\middle|\boldsymbol{x},\boldsymbol{y}\right) \overset{\text{ανεξαρτησία}}{=} \sup_{\boldsymbol{\theta}|H_0} p\left(\mu,\lambda\middle|x_1,\ ...,\ x_n\right)\cdot\sup_{\boldsymbol{\theta}|H_0} p\left(\mu,\lambda\middle|y_1,...,y_n\right)=$$

$$= \sup_{\boldsymbol{\theta}|H_0}\left[\prod_{i=1}^{n}\left(\frac{\lambda}{2\pi x_i^3}\right)^{1/2}\exp\left(-\frac{\lambda\left(x_i-\mu\right)^2}{2\mu^2 x_i}\right)\right]\cdot\sup_{\boldsymbol{\theta}|H_0}\left[\prod_{j=1}^{m}\left(\frac{\lambda}{2\pi y_j^3}\right)^{1/2}\exp\left(-\frac{\lambda\left(y_j-\mu\right)^2}{2\mu^2 y_j}\right)\right]=$$

$$= \sup_{\boldsymbol{\theta}|H_0}\left[\left(\frac{\lambda}{2\pi}\right)^{\frac{n+m}{2}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left(-\frac{\lambda}{2}\left[\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{\mu^2 x_i}+\sum_{j=1}^{m}\frac{\left(y_j-\mu\right)^2}{\mu^2 y_j}\right]\right)\right]=$$

$$= \left(\frac{\tilde{\lambda}}{2\pi}\right)^{\frac{n+m}{2}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left(-\frac{\tilde{\lambda}}{2}\left[\sum_{i=1}^{n}\frac{\left(x_i-\tilde{\mu}\right)^2}{\tilde{\mu}^2 x_i}+\sum_{j=1}^{m}\frac{\left(y_j-\tilde{\mu}\right)^2}{\tilde{\mu}^2 y_j}\right]\right)=$$

$$\sup_{\boldsymbol{\theta}|H_0} p\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)=\left(\frac{\tilde{\lambda}}{2\pi}\right)^{\frac{n+m}{2}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left(-\frac{n+m}{2}\right) \qquad (2.50)$$

Με παρόμοια συλλογιστική, παίρνουμε:

$$\sup_{\boldsymbol{\theta}|H_1} p\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)=\sup_{\boldsymbol{\theta}|H_1} p\left(\mu,\mathrm{v},\lambda\middle|\boldsymbol{x},\boldsymbol{y}\right)\overset{\text{ανεξαρτησία}}{=}\sup_{\boldsymbol{\theta}|H_1} p\left(\mu,\lambda\middle|x_1,\ ...,\ x_n\right)\cdot\sup_{\boldsymbol{\theta}|H_1} p\left(\mathrm{v},\lambda\middle|y_1,...,y_n\right)=$$

$$= \sup_{\boldsymbol{\theta}|H_1}\left[\prod_{i=1}^{n}\left(\frac{\lambda}{2\pi x_i^3}\right)^{1/2}\exp\left(-\frac{\lambda\left(x_i-\mu\right)^2}{2\mu^2 x_i}\right)\right]\cdot\sup_{\boldsymbol{\theta}|H_1}\left[\prod_{j=1}^{m}\left(\frac{\lambda}{2\pi y_j^3}\right)^{1/2}\exp\left(-\frac{\lambda\left(y_j-\mathrm{v}\right)^2}{2\mathrm{v}^2 y_j}\right)\right]=$$

$$= \sup_{\boldsymbol{\theta}|H_1}\left[\left(\frac{\lambda}{2\pi}\right)^{\frac{n+m}{2}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left(-\frac{\lambda}{2}\left[\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{\mu^2 x_i}+\sum_{j=1}^{m}\frac{\left(y_j-\mathrm{v}\right)^2}{\mathrm{v}^2 y_j}\right]\right)\right]=$$

$$= \left(\frac{\hat{\lambda}}{2\pi}\right)^{\frac{n+m}{2}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left(-\frac{\hat{\lambda}}{2}\left[\sum_{i=1}^{n}\frac{\left(x_i-\hat{\mu}\right)^2}{\hat{\mu}^2 x_i}+\sum_{j=1}^{m}\frac{\left(y_j-\hat{\mathrm{v}}\right)^2}{\hat{\mathrm{v}}^2 y_j}\right]\right)$$

Αντικαθιστώντας στην παραπάνω σχέση τις $\hat{\mu}$, $\hat{\mathrm{v}}$ και $\hat{\lambda}$ των σχέσεων (2.47), (2.48) και (2.49) καταλήγουμε στη σχέση:

$$\sup_{\boldsymbol{\theta}|H_1} p\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)=\sup_{\boldsymbol{\theta}|H_1} p\left(\mu,\mathrm{v},\lambda\middle|\boldsymbol{x},\boldsymbol{y}\right)=\left(\frac{\hat{\lambda}}{2\pi}\right)^{\frac{n+m}{2}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\cdot\exp\left(-\frac{n+m}{2}\right) \qquad (2.51)$$

Λόγω των σχέσεων (2.50) και (2.51), ο λόγος των πιθανοφανειών γίνεται:

$$\lambda^*=\frac{\sup_{\boldsymbol{\theta}|H_0} p\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)}{\sup_{\boldsymbol{\theta}|H_1} p\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)}=\frac{\left(\frac{\tilde{\lambda}}{2\pi}\right)^{\frac{n+m}{2}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\cdot\exp\left(-\frac{n+m}{2}\right)}{\left(\frac{\hat{\lambda}}{2\pi}\right)^{\frac{n+m}{2}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\cdot\exp\left(-\frac{n+m}{2}\right)}\quad\Leftrightarrow$$

$$\lambda^* = \left(\frac{\tilde{\lambda}}{\hat{\lambda}}\right)^{\frac{n+m}{2}} = \left(\frac{\dfrac{n+m}{\sum_{i=1}^{n}\dfrac{\left(x_i-\tilde{\mu}\right)^2}{\tilde{\mu}^2 x_i}+\sum_{j=1}^{m}\dfrac{\left(y_j-\tilde{\mu}\right)^2}{\tilde{\mu}^2 y_j}}}{\dfrac{n+m}{\sum_{i=1}^{n}\left(\dfrac{1}{x_i}-\dfrac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\dfrac{1}{y_j}-\dfrac{1}{\bar{y}}\right)}} \right)^{\frac{n+m}{2}} \quad \Leftrightarrow$$

$$\lambda^* = \left(\frac{\sum_{i=1}^{n}\left(\dfrac{1}{x_i}-\dfrac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\dfrac{1}{y_j}-\dfrac{1}{\bar{y}}\right)}{\sum_{i=1}^{n}\dfrac{\left(x_i-\tilde{\mu}\right)^2}{\tilde{\mu}^2 x_i}+\sum_{j=1}^{m}\dfrac{\left(y_j-\tilde{\mu}\right)^2}{\tilde{\mu}^2 y_j}} \right)^{\frac{n+m}{2}} \tag{2.52}$$

Κάνοντας πράξεις στον παρανομαστή, έχουμε:

$$\sum_{i=1}^{n}\frac{\left(x_i-\tilde{\mu}\right)^2}{\tilde{\mu}^2 x_i}+\sum_{j=1}^{m}\frac{\left(y_j-\tilde{\mu}\right)^2}{\tilde{\mu}^2 y_j} = \frac{\sum_{i=1}^{n}x_i}{\tilde{\mu}^2}-\frac{2n}{\tilde{\mu}}+\sum_{i=1}^{n}\frac{1}{x_i}+\frac{\sum_{j=1}^{m}y_j}{\tilde{\mu}^2}-\frac{2m}{\tilde{\mu}}+\sum_{j=1}^{m}\frac{1}{y_j} =$$

$$= \frac{n\bar{x}\left(n+m\right)^2}{\left(n\bar{x}+m\bar{y}\right)^2}+\frac{m\bar{y}\left(n+m\right)^2}{\left(n\bar{x}+m\bar{y}\right)^2}+\sum_{i=1}^{n}\frac{1}{x_i}+\sum_{j=1}^{m}\frac{1}{y_j}-\frac{2n\left(n+m\right)}{n\bar{x}+m\bar{y}}-\frac{2m\left(n+m\right)}{n\bar{x}+m\bar{y}} =$$

$$= \sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)+\frac{n\bar{y}+m\bar{x}}{\bar{x}\bar{y}}-\frac{\left(n+m\right)^2}{n\bar{x}+m\bar{y}} =$$

$$= \sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)+\frac{nm\left(\bar{x}-\bar{y}\right)^2}{\bar{x}\bar{y}\left(n\bar{x}+m\bar{y}\right)}$$

Τελικά, η σχέση (2.52) γίνεται:

$$\lambda^* = \left(\frac{\sum_{i=1}^{n}\left(\dfrac{1}{x_i}-\dfrac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\dfrac{1}{y_j}-\dfrac{1}{\bar{y}}\right)}{\sum_{i=1}^{n}\left(\dfrac{1}{x_i}-\dfrac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\dfrac{1}{y_j}-\dfrac{1}{\bar{y}}\right)+\dfrac{nm\left(\bar{x}-\bar{y}\right)^2}{\bar{x}\bar{y}\left(n\bar{x}+m\bar{y}\right)}} \right)^{\frac{n+m}{2}} =$$

$$= \left(1 \Big/ \left[1+\frac{nm\left(\bar{x}-\bar{y}\right)^2}{\bar{x}\bar{y}\left(n\bar{x}+m\bar{y}\right)\left\{\sum_{i=1}^{n}\left(\dfrac{1}{x_i}-\dfrac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\dfrac{1}{y_j}-\dfrac{1}{\bar{y}}\right)\right\}} \right] \right)^{\frac{n+m}{2}}$$

Θέτοντας

$$Q=\frac{nm\left(\overline{x}-\overline{y}\right)^2}{\overline{xy}\left(n\overline{x}+m\overline{y}\right)\left\{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\overline{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\overline{y}}\right)\right\}}, \tag{2.53}$$

έχουμε ότι:

$$\lambda^*=\left(\frac{1}{1+Q}\right)^{\frac{n+m}{2}} \tag{2.54}$$

Η απόρριψη της μηδενικής υπόθεσης για μικρές τιμές του λ^* ισοδυναμεί με απόρριψη της μηδενικής υπόθεσης για μεγάλες τιμές του $\left|\sqrt{Q}\right|$, το οποίο είναι ισοδύναμο με τον ομοιόμορφα πιο ισχυρό αμερόληπτο έλεγχο που παρουσίασε ο Chhikara (1975). Η ελεγχοσυνάρτηση $\left[\left(n+m-2\right)Q\right]^{1/2}$ ακολουθεί την *t* κατανομή με $\left(n+m-2\right)$ βαθμούς ελευθερίας.

2.7.2 Έλεγχος υπόθεσης για την παράμετρο λ της IG

Έστω ένα δείγμα *n* παρατηρήσεων $X_i \sim IG\left(\mu,\lambda\right)$ και ένα δείγμα *m* παρατηρήσεων $Y_i \sim IG\left(\kappa,\nu\right)$. Θέλουμε να προχωρήσουμε στον έλεγχο της υπόθεσης:

$\mathrm{H}_0: \lambda=\nu$

$\mathrm{H}_1: \lambda\neq\nu$,

Η από κοινού σ.π.π ή συνάρτηση πιθανοφάνειας δίνεται από τον τύπο:

$$L\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right)=L\left(\mu,\kappa,\lambda,\nu|\,x_1,\ ...,x_n,y_1,...,y_n\right)=f_{\boldsymbol{x},\boldsymbol{y}}=\prod_{i=1}^{n}f_{x_i}\left(x_i,\boldsymbol{\theta}_1\right)\cdot\prod_{j=1}^{m}f_{y_j}\left(y_j,\boldsymbol{\theta}_2\right)=$$

$$=\frac{\lambda^{\frac{n}{2}}\nu^{\frac{m}{2}}}{\left(2\pi\right)^{\frac{n+m}{2}}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left(-\frac{\lambda}{2}\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{\mu^2x_i}\right)\exp\left(-\frac{\nu}{2}\sum_{j=1}^{m}\frac{\left(y_j-\kappa\right)^2}{\kappa^2y_j}\right)\quad\Leftrightarrow$$

$$L\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right)=\frac{\lambda^{\frac{n}{2}}\nu^{\frac{m}{2}}}{\left(2\pi\right)^{\frac{n+m}{2}}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left(-\frac{\lambda}{2}\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{\mu^2x_i}\right)\exp\left(-\frac{\nu}{2}\sum_{j=1}^{m}\frac{\left(y_j-\kappa\right)^2}{\kappa^2y_j}\right) \tag{2.55}$$

Υπό τη μηδενική υπόθεση

Η συνάρτηση πιθανοφάνειας της σχέσης (2.55) γίνεται:

$$L\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)=\left(\frac{\lambda}{2\pi}\right)^{\frac{n+m}{2}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left(-\frac{\lambda}{2}\left[\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{\mu^2 x_t}+\sum_{j=1}^{m}\frac{\left(y_j-\kappa\right)^2}{\kappa^2 y_j}\right]\right)$$

Λογαριθμίζοντας την παραπάνω σχέση έχουμε:

$$\ln L\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)=\ln L=\frac{n+m}{2}\ln\frac{\lambda}{2\pi}-\frac{3}{2}\left[\sum_i \ln x_t+\sum_j \ln y_j\right]-\frac{\lambda}{2}\left[\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{\mu^2 x_i}+\sum_{j=1}^{m}\frac{\left(y_j-\kappa\right)^2}{\kappa^2 y_j}\right]$$

Τέλος, με μερική παραγώγιση και παρόμοια συλλογιστική με την περίπτωση του ελέγχου για την ισότητα των μέσων τιμών, προκύπτουν οι εκτιμήτριες μέγιστης πιθανοφάνειας για τις παραμέτρους μ, n και λ:

$$\tilde{\mu}=\bar{x} \tag{2.56}$$

$$\tilde{\kappa}=\bar{y} \tag{2.57}$$

και

$$\tilde{\lambda}=\frac{n+m}{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)} \tag{2.58}$$

Υπό την εναλλακτική υπόθεση

Η συνάρτηση πιθανοφάνειας της σχέσης (2.55) γίνεται:

$$L\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)=\frac{\lambda^{\frac{n}{2}}\nu^{\frac{m}{2}}}{\left(2\pi\right)^{\frac{n+m}{2}}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left(-\frac{\lambda}{2}\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{\mu^2 x_i}\right)\exp\left(-\frac{\nu}{2}\sum_{j=1}^{m}\frac{\left(y_j-\kappa\right)^2}{\kappa^2 y_j}\right)$$

Λογαριθμίζοντας την παραπάνω σχέση, έχουμε:

$$\ln L\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)=\frac{n}{2}\ln\lambda+\frac{m}{2}\ln\nu-\frac{n+m}{2}\ln 2\pi-\frac{3}{2}\left[\sum_i \ln x_i+\sum_j \ln y_j\right]-$$

$$-\frac{\lambda}{2}\left[\sum_{i=1}^{n}\frac{\left(x_i-\mu\right)^2}{\mu^2 x_i}+\sum_{j=1}^{m}\frac{\left(y_j-\kappa\right)^2}{\kappa^2 y_j}\right]$$

Όμοια, με μερική παραγώγιση της παραπάνω σχέσης, προκύπτουν οι εκτιμήτριες μέγιστης πιθανοφάνειας των παραμέτρων μ, n, λ και ν:

$$\hat{\mu}=\frac{\sum_{i=1}^{n}x_i}{n}=\bar{x} \tag{2.59}$$

$$\hat{\kappa} = \frac{\sum_{j=1}^{m} y_j}{m} = \bar{y} \tag{2.60}$$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i} - \frac{1}{\bar{x}}\right)} \tag{2.61}$$

και

$$\hat{\nu} = \frac{m}{\sum_{j=1}^{m}\left(\frac{1}{y_j} - \frac{1}{\bar{y}}\right)} \tag{2.62}$$

Προκειμένου να χρησιμοποιήσουμε τον έλεγχο του λόγου των πιθανοφανειών, έχουμε τα εξής:

$$\sup_{\boldsymbol{\theta}|H_0} p\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right) = \sup_{\boldsymbol{\theta}|H_0} p\left(\mu,\kappa,\lambda\middle|\boldsymbol{x},\boldsymbol{y}\right) \overset{\text{ανεξαρτησία}}{=} \sup_{\boldsymbol{\theta}|H_0} p\left(\mu,\lambda\middle|x_1,\ ...,\ x_n\right)\cdot\sup_{\boldsymbol{\theta}|H_0} p\left(\kappa,\lambda\middle|y_1,...,y_n\right) =$$

$$= \sup_{\boldsymbol{\theta}|H_0}\left[\prod_{i=1}^{n}\left(\frac{\lambda}{2\pi x_i^3}\right)^{1/2}\exp\left(-\frac{\lambda\left(x_i-\mu\right)^2}{2\mu^2 x_i}\right)\right]\cdot\sup_{\boldsymbol{\theta}|H_0}\left[\prod_{j=1}^{m}\left(\frac{\lambda}{2\pi y_j^3}\right)^{1/2}\exp\left(-\frac{\lambda\left(y_j-\kappa\right)^2}{2\kappa^2 y_j}\right)\right] =$$

$$= \left(\frac{\tilde{\lambda}}{2\pi}\right)^{\frac{n+m}{2}}\prod_{i=1}^{n}x_i^{-3/2}\prod_{j=1}^{m}y_j^{-3/2}\exp\left(-\frac{\tilde{\lambda}}{2}\left[\sum_{i=1}^{n}\frac{\left(x_i-\tilde{\mu}\right)^2}{\tilde{\mu}^2 x_i}+\sum_{j=1}^{m}\frac{\left(y_j-\tilde{\kappa}\right)^2}{\tilde{\kappa}^2 y_j}\right]\right) \quad \Leftrightarrow$$

$$\sup_{\boldsymbol{\theta}|H_0} p\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right) = \left(\frac{\tilde{\lambda}}{2\pi}\right)^{\frac{n+m}{2}}\prod_{i=1}^{n}x_i^{-3/2}\prod_{j=1}^{m}y_j^{-3/2}\exp\left(-\frac{n+m}{2}\right) \tag{2.63}$$

Με παρόμοια συλλογιστική για την πιθανοφάνεια του δείγματος υπό την εναλλακτική υπόθεση, έχουμε:

$$\sup_{\boldsymbol{\theta}|H_1} p\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right) = \sup_{\boldsymbol{\theta}|H_1} p\left(\mu,\kappa,\lambda,\nu\middle|\boldsymbol{x},\boldsymbol{y}\right) \overset{\text{ανεξαρτησία}}{=} \sup_{\boldsymbol{\theta}|H_1} p\left(\mu,\lambda\middle|x_1,\ ...,\ x_n\right)\cdot\sup_{\boldsymbol{\theta}|H_1} p\left(\kappa,\nu\middle|y_1,...,y_n\right) =$$

$$= \sup_{\boldsymbol{\theta}|H_1}\left[\prod_{i=1}^{n}\left(\frac{\lambda}{2\pi x_i^3}\right)^{1/2}\exp\left(-\frac{\lambda\left(x_i-\mu\right)^2}{2\mu^2 x_i}\right)\prod_{j=1}^{m}\left(\frac{\nu}{2\pi y_j^3}\right)^{1/2}\exp\left(-\frac{\nu\left(y_j-\kappa\right)^2}{2\kappa^2 y_j}\right)\right] =$$

$$= \frac{\hat{\lambda}^{\frac{n}{2}}\hat{\nu}^{\frac{m}{2}}}{\left(2\pi\right)^{\frac{n+m}{2}}}\prod_{i=1}^{n}x_i^{-3/2}\prod_{j=1}^{m}y_j^{-3/2}\exp\left(-\frac{\hat{\lambda}}{2}\sum_{i=1}^{n}\frac{\left(x_i-\hat{\mu}\right)^2}{\hat{\mu}^2 x_i}\right)\exp\left(-\frac{\hat{\nu}}{2}\sum_{j=1}^{m}\frac{\left(y_j-\hat{\kappa}\right)^2}{\hat{\kappa}^2 y_j}\right) =$$

$$= \frac{\hat{\lambda}^{\frac{n}{2}}\hat{\nu}^{\frac{m}{2}}}{\left(2\pi\right)^{\frac{n+m}{2}}}\prod_{i=1}^{n}x_i^{-3/2}\prod_{j=1}^{m}y_j^{-3/2}\exp\left(-\frac{n+m}{2}\right)$$

$$\sup_{\boldsymbol{\theta}|H_1} p\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right)=\frac{\hat{\lambda}^{\frac{n}{2}}\hat{\nu}^{\frac{m}{2}}}{\left(2\pi\right)^{\frac{n+m}{2}}}\prod_{i=1}^{n}x_i^{-3/2}\prod_{j=1}^{m}y_j^{-3/2}\exp\left(-\frac{n+m}{2}\right) \tag{2.64}$$

Λόγω των σχέσεων (2.63) και (2.64), ο λόγος των πιθανοφανειών γίνεται:

$$\lambda^*=\frac{\sup_{\boldsymbol{\theta}|H_0} p\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right)}{\sup_{\boldsymbol{\theta}|H_1} p\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right)}=\frac{\left(\frac{\tilde{\lambda}}{2\pi}\right)^{\frac{n+m}{2}}\prod_{i=1}^{n}x_i^{-3/2}\prod_{j=1}^{m}y_j^{-3/2}\exp\left(-\frac{n+m}{2}\right)}{\frac{\hat{\lambda}^{\frac{n}{2}}\hat{\nu}^{\frac{m}{2}}}{\left(2\pi\right)^{\frac{n+m}{2}}}\prod_{i=1}^{n}x_i^{-3/2}\prod_{j=1}^{m}y_j^{-3/2}\exp\left(-\frac{n+m}{2}\right)} \quad \Leftrightarrow$$

$$\lambda^*=\frac{\left(\tilde{\lambda}\right)^{\frac{n+m}{2}}}{\hat{\lambda}^{\frac{n}{2}}\hat{\nu}^{\frac{m}{2}}}=\frac{\left(\frac{n+m}{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)}\right)^{\frac{n+m}{2}}}{\left(\frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)}\right)^{\frac{n}{2}}\left(\frac{m}{\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)}\right)^{\frac{m}{2}}} \quad \Leftrightarrow$$

$$\lambda^*=\frac{\left(n+m\right)^{\frac{n+m}{2}}}{\left(n\right)^{\frac{n}{2}}\left(m\right)^{\frac{m}{2}}}\frac{\left(\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)\right)^{\frac{-m}{2}}\left(\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)\right)^{\frac{m}{2}}}{\left(1+\frac{\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)}{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)}\right)^{\frac{n+m}{2}}} \quad \Leftrightarrow$$

Θέτοντας

$$Q_1=\frac{\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)}{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)}, \tag{2.65}$$

έχουμε ότι:

$$\lambda^*=\frac{\left(n+m\right)^{\frac{n+m}{2}}}{\left(n\right)^{\frac{n}{2}}\left(m\right)^{\frac{m}{2}}}\left(\frac{1}{1+Q_1}\right)^{\frac{n+m}{2}}Q_1^{\frac{m}{2}} \tag{2.66}$$

Παρατηρούμε ότι:

$$\ln\lambda^* = \ln\left[\frac{(n+m)^{\frac{n+m}{2}}}{(n)^{\frac{n}{2}}(m)^{\frac{m}{2}}}\right] + \frac{n+m}{2}\ln\left(\frac{1}{1+Q_1}\right) + \frac{m}{2}\ln Q_1 \quad \Leftrightarrow$$

$$\frac{d\ln\lambda^*}{dQ_1} = -\frac{n+m}{2}\frac{1}{1+Q_1} + \frac{m}{2}\frac{1}{Q_1} = \frac{-(m+n)Q_1 + m(1+Q_1)}{2Q_1(1+Q_1)} \quad \Leftrightarrow$$

$$\frac{d\ln\lambda^*}{dQ_1} = \frac{m-nQ_1}{2Q_1(1+Q_1)}$$

Η λ^* είναι μία αύξουσα συνάρτηση για

$$\frac{d\ln\lambda^*}{dQ_1} > 0 \quad \Leftrightarrow \quad m - nQ_1 > 0 \quad \Leftrightarrow \quad Q_1 < \frac{m}{n}$$

και φθίνουσα για

$$\frac{d\ln\lambda^*}{dQ_1} < 0 \quad \Leftrightarrow \quad m - nQ_1 < 0 \quad \Leftrightarrow \quad Q_1 > \frac{m}{n}.$$

Η απόρριψη της μηδενικής υπόθεσης για μικρές τιμές του λ^*, ισοδυναμεί με απόρριψη της μηδενικής υπόθεσης για πολύ μεγάλες ή πολύ μικρές τιμές του Q_1. Το Q_1 είναι ο λόγος δύο ανεξάρτητων χ^2 κατανομών. Επομένως, $Q_1\frac{(n-1)}{(m-1)} \sim F_{m-1,n-1}$, οπότε για ακόμα μία φορά ο λόγος των πιθανοφανειών μας οδηγεί σε μία ελεγχοσυνάρτηση που ακολουθεί μία γνωστή κατανομή.

2.7.3 Η έννοια της άτυπης τιμής (outliers)

Όπως και σε όλα τα στατιστικά μοντέλα, είναι σημαντικός ο εντοπισμός παρατηρήσεων οι οποίες δεν είναι συνεπείς με τα υπόλοιπα δεδομένα. Στη βιβλιογραφία, τέτοιες παρατηρήσεις αναφέρονται ως **άτυπα σημεία** ή άτυπες τιμές (outliers) και ενδεχομένως έχουν μεγάλη επιρροή στην προσαρμογή του μοντέλου. Για το λόγο αυτό, έχουν αναπτυχθεί έλεγχοι εντοπισμού άτυπων σημείων για πολλές κατανομές. Ωστόσο, δεν έχουν ακόμα αναπτυχθεί σχετικές τεχνικές για την κατανομή IG. Συχνά, το άτυπο σημείο μοντελοποιείται μέσω της μεταβολής μιας παραμέτρου της κατανομής, π.χ. $\mathrm{N}(\mu,\sigma^2) \to \mathrm{N}(\mu+\alpha,\sigma^2)$.

Στη συνέχεια, παρουσιάζονται έλεγχοι για τον εντοπισμό άτυπων τιμών των παραμέτρων (μ,λ) της IG σε δεδομένα ανεξάρτητων μονάδων με επαναλαμβανόμενα ανεξάρτητα γεγονότα ανά μονάδα. Οι έλεγχοι αυτοί αποτελούν επεκτάσεις των δύο ελέγχων

που παρουσιάσαμε στις Παραγράφους 2.7.1 και 2.7.2. Η έννοια της άτυπης τιμής θα αναφέρεται σε κάποια μονάδα με διαφορετικές τιμές των παραμέτρων από τις υπόλοιπες μονάδες ενός δείγματος. Η περίπτωση των επαναλαμβανόμενων παρατηρήσεων για την κάθε μονάδα γίνεται, προκειμένου να διευκολύνουμε την εφαρμογή των επαναλαμβανόμενων δεδομένων σε δεδομένα διάρκειας ζωής.

Οι διάφοροι έλεγχοι βασίζονται στη μεγιστοποίηση της τιμής του ελέγχου του λόγου των πιθανοφανειών για την ισότητα παραμέτρων, με διόρθωση Bonferroni για τις *p*-τιμές. Επειδή η ελεγχοσυνάρτηση στην περίπτωση της ισότητας της παραμέτρου λ μεταξύ μονάδων ακολουθεί την κατανομή *F* με διαφορετικούς βαθμούς ελευθερίας ανά μονάδα, εφαρμόζεται ένας μετασχηματισμός από την *F* στην τυποποιημένη Κανονική κατανομή. Στη συνέχεια, θα γίνουν προσομοιώσεις προκειμένου να επιβεβαιώσουμε την ακρίβεια των Bonferroni ελέγχων υπό τη μηδενική υπόθεση και να μελετήσουμε την ισχύ των ελέγχων υπό την εναλλακτική υπόθεση. Τα διάφορα αποτελέσματα θα παρουσιαστούν για τη συνηθισμένη αναπαραμέτρηση της σχέσης (2.4) που συναντάται στη βιβλιογραφία, όπου τα δεδομένα είναι της αντίστροφης Γκαουσιανής κατανομής.

2.7.4 Έλεγχος εντοπισμού άτυπων τιμών για την παράμετρο μ της IG

Ας υποθέσουμε πως έχουμε ένα δείγμα από *k* ανεξάρτητες μονάδες και πως για την *i* μονάδα, υπάρχουν n_i ανεξάρτητες εμφανίσεις μίας τυχαίας μεταβλητής *T*, η οποία ακολουθεί την αντίστροφη Γκαουσιανή κατανομή με σ.π.π.:

$$f\left(t;\mu_i,\lambda_i\right)=\sqrt{\frac{\lambda_i}{2\pi t^3}}\exp\left\{-\frac{\lambda_i\left(t-\mu_i\right)^2}{2\mu_i^2 t}\right\},\quad t>0$$

και παραμέτρους $\mu_i>0$ και $\lambda_i>0$. Ένας έλεγχος της μηδενικής υπόθεσης ότι όλες οι *k* μονάδες έχουν την ίδια τιμή για την παράμετρο μ, έναντι της εναλλακτικής υπόθεσης πως η μονάδα *i*, $i=1,2,...,k$ έχει διαφορετική τιμή από τις υπόλοιπες, μπορεί να εκφραστεί ως εξής:

$$\mathrm{H}_0:\mu_1=\mu_2=\mu_3=...=\mu$$

$$\mathrm{H}_1:\ \mu_i\neq\mu$$

Ο παραπάνω έλεγχος για τη μονάδα *i*, $i=1,2,...,k$, μας οδηγεί στην ελεγχοσυνάρτηση Q_i της σχέσης (2.53) για τον έλεγχο του λόγου των πιθανοφανειών. Στη συνέχεια, θα αποδείξουμε πως μεγάλες τιμές της ελεγχοσυνάρτησης οδηγούν σε απόρριψη της H_0. Στην περίπτωση που η πιθανή άτυπη τιμή δεν έχει εντοπιστεί εκ των προτέρων, η κατάλληλη ελεγχοσυνάρτηση είναι η $\max_i Q_i$ και το σύστημα που τελικά εντοπίζεται σαν άτυπη τιμή είναι η μονάδα εκείνη, στην οποία επιτυγχάνεται το μέγιστο.

Ο έλεγχος της υπόθεσης H_i έναντι της H_0 είναι ένας έλεγχος της ισότητας της παραμέτρου μ μεταξύ δύο ομάδων, όπου η μία αντιστοιχεί αποκλειστικά στη μονάδα *i* και η άλλη αποτελείται από τα δεδομένα όλων των υπόλοιπων μονάδων εκτός από την *i*. Στον έλεγχο του λόγου των πιθανοφανειών που δόθηκε από την Davis (1980) και παρουσιάστηκε στην Παράγραφο 2.7.1, έχουμε προχωρήσει στη επιπρόσθετη υπόθεση ότι οι παράμετροι λ_i είναι ίσοι. Υιοθετώντας τα αποτελέσματα της στη δική μας περίπτωση η ελεγχοσυνάρτηση που προκύπτει είναι:

$$Q_i = \frac{N_{(i)} n_i \left(\bar{t}_{(i)} - \bar{t}_i\right)^2}{\bar{t}_{(i)}\bar{t}_i N \bar{t}\left[\sum_{j\neq i}\sum_{l=1}^{n_j}\left(\frac{1}{t_{jl}} - \frac{1}{\bar{t}_{(i)}}\right) + \sum_{l=1}^{n_i}\left(\frac{1}{t_{il}} - \frac{1}{\bar{t}_i}\right)\right]},$$

όπου:

- ✓ $\bar{t}_i = \sum_{l=1}^{n_i} \frac{t_{il}}{n_i}$ είναι η μέση τιμή των παρατηρήσεων της μονάδας *i*,
- ✓ $\bar{t}_{(i)} = \sum_{j\neq i}\sum_{l=1}^{n_j} \frac{t_{jl}}{N_{(i)}}$ είναι η μέση τιμή όλων των παρατηρήσεων εκτός εκείνων της μονάδας *i*,
- ✓ $\bar{t}$ είναι η μέση τιμή όλων των παρατηρήσεων,
- ✓ $N_{(i)} = \sum_{j\neq i} n_j$ είναι ο συνολικός αριθμός των παρατηρήσεων εκτός από εκείνες της μονάδας *i* και
- ✓ *N* είναι ο συνολικός αριθμός των παρατηρήσεων.

Είδαμε πως η στατιστική συνάρτηση $\left\{(N-2)Q_i\right\}^{1/2}$ ακολουθεί την *t* κατανομή με $N_{(i)} + n_i - 2 = N - 2$ βαθμούς ελευθερίας. Ένας έλεγχος Bonferroni σε επίπεδο a % για τον εντοπισμό μίας άτυπης τιμής πραγματοποιείται συγκρίνοντας τις ακραίες τιμές των Q_i με τα άνω και κάτω $a/2k$ ποσοστιαία σημεία της *t* κατανομής. Η διαδικασία αυτή είναι συντηρητική, αλλά είναι γνωστό πως τα πραγματικά και θεωρητικά επίπεδα σημαντικότητας σε ελέγχους για τον εντοπισμό μιας άτυπης τιμής που κατασκευάζονται με τον παραπάνω τρόπο έχουν την τάση να είναι πολύ κοντά μεταξύ τους.

Περιγραφή αλγορίθμων

Στη συνέχεια, παραθέτουμε τον ψευδοκώδικα του αλγορίθμου που χρησιμοποιήσαμε στην R, προκειμένου να κάνουμε προσομοιώσεις για τον παραπάνω έλεγχο υπόθεσης.

1. Παραγωγή ενός δείγματος *k* ανεξάρτητων μονάδων από την IG με ανεξάρτητο αριθμό επαναλαμβανομένων γεγονότων ανά μονάδα.
2. Υπολογίζουμε τις ποσότητες Q_i, $\left|\sqrt{(\mathrm{N}-2)Q_i}\right|$ και $\max\left|\sqrt{(\mathrm{N}-2)Q_i}\right|$ για τις οποίες γνωρίζουμε πως η ελεγχοσυνάρτηση $\sqrt{(\mathrm{N}-2)Q_i} \sim t_{(\mathrm{N}-2)}$, όπου $\mathrm{N} = \sum_{i=1}^{k} n_i$.
3. Χρησιμοποιούμε τις ανισότητες Bonferroni, ώστε να υπολογίσουμε το ποσοστιαίο σημείο t_a:

$$\mathrm{P}\left(\max\left|\sqrt{(\mathrm{N}-2)Q_i}\right| > t_a\right) = a \Rightarrow \cdots \Rightarrow \mathrm{P}\left(\sqrt{(\mathrm{N}-2)Q_i} > t_a\right) = \frac{a}{2k} \quad \Rightarrow$$

$$\mathrm{P}\left(\sqrt{(\mathrm{N}-2)Q_i} \le t_a\right) = 1 - \frac{a}{2k} \Rightarrow F\left(t_a\right) = 1 - \frac{a}{2k} \Rightarrow$$

$$t_a = F^{-1}\left(1 - \frac{a}{2k}\right)$$

4. Επαναλαμβάνουμε τα βήματα 1 και 2 για 10,000 φορές και μετράμε πόσες φορές η ελεγχοσυνάρτηση:

$$\max\left|\sqrt{(\mathrm{N}-2)Q_i}\right| > t_a$$

5. Επαναλαμβάνουμε την παραπάνω διαδικασία υπό την εναλλακτική υπόθεση, προκειμένου να αποκτήσουμε την ισχύ του ελέγχου.

2.7.5 Έλεγχος εντοπισμού άτυπων τιμών για την παράμετρο *λ* της IG

Για τον έλεγχο της υπόθεσης:

$\mathrm{H}_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_k = \lambda$

$\mathrm{H}_1 : \lambda_i \neq \lambda$,

εργαζόμαστε με παρόμοια συλλογιστική με αυτήν της προηγούμενης παραγράφου. Η κατάλληλη ελεγχοσυνάρτηση για τη μεγιστοποίηση του λόγου των πιθανοφανειών, αυτή τη φορά χωρίς περιορισμό για το μ_i, δόθηκε και πάλι από την Davis (1980) και παρουσιάστηκε στην Παράγραφο 2.7.2. Τα αποτελέσματά της υιοθετημένα στη δική μας περίπτωση δημιουργούν την εξής ελεγχοσυνάρτηση:

$$F_i = \frac{\left(N_{(i)} - 1\right) Q_{1i}}{n_i - 1}, \tag{2.67}$$

όπου

$$Q_{1i} = \frac{\sum_{l=1}^{n_i} \left(\frac{1}{t_{il}} - \frac{1}{\overline{t}_i} \right)}{\sum_{j \neq i} \sum_{l=1}^{n_j} \left(\frac{1}{t_{jl}} - \frac{1}{\overline{t}_{(i)}} \right)}$$

με κατανομή υπό τη μηδενική υπόθεση, $F_i \sim F_{n_i - 1,\, N_{(i)} - 1}$.

Να σημειώσουμε πως ο έλεγχος αυτός δεν μπορεί να χρησιμοποιηθεί για την περίπτωση μονάδων με μία μόνο παρατήρηση $\left(n_i = 1\right)$. Ωστόσο, τέτοιες μονάδες μπορούν να συμπεριληφθούν στους υπολογισμούς άλλων μονάδων. Ο αριθμός των *k* μονάδων για τον υπολογισμό των κρίσιμων τιμών των ελέγχων Bonferroni θα ήταν σε αυτήν την περίπτωση ο αριθμός των μονάδων με επαναλαμβανόμενες παρατηρήσεις, δηλαδή $n_i \geq 2$.

Οι ελεγχοσυναρτήσεις F_i έχουν διαφορετικούς βαθμούς ελευθερίας για τα διάφορα *i*, εκτός εάν ο αριθμός των παρατηρήσεων n_i είναι ίδιος για κάθε μονάδα, έστω *n*. Σε αυτήν την περίπτωση κάθε στατιστικό F_i ακολουθεί την *F* κατανομή με $\left(n-1\right)$ και $n\left(k-1\right)$ βαθμούς ελευθερίας. Ένας έλεγχος Bonferroni σε επίπεδο $a\%$ για τον εντοπισμό μίας άτυπης τιμής πραγματοποιείται συγκρίνοντας τις ακραίες τιμές των F_i με τα άνω και κάτω $a/2k$ ποσοστιαία σημεία της *F* κατανομής.

Στη γενική περίπτωση με άνισους βαθμούς ελευθερίας, προχωράμε χρησιμοποιώντας έναν προσεγγιστικό μετασχηματισμό κανονικοποίησης, ο οποίος δίνεται στον Viveros (1990), έτσι ώστε όλες οι ελεγχοσυναρτήσεις να ακολουθούν την ίδια κατανομή. Ο μετασχηματισμός αυτός έχει χρησιμοποιηθεί επιτυχώς από την Caroni (2010) για ένα παρόμοιο σκοπό. Εάν η τυχαία μεταβλητή F_i ακολουθεί την κατανομή *F* με $2m_i$ και $2M_{(i)}$ βαθμούς ελευθερίας, αποδεικνύεται πως η τυχαία μεταβλητή Z_i ακολουθεί προσεγγιστικά την τυποποιημένη Κανονική κατανομή με

$$Z_i = \frac{F_i^{c_i} - \mu_i}{\sigma_i},$$

όπου

$$c_i = \frac{M_{(i)} - m_i}{3M - 2}$$

$$\mu_i = \frac{\Gamma(m_i + c_i)\Gamma(m_i - c_i)\left(M_{(i)}/m_i\right)^{c_i}}{\Gamma\left(M_{(i)}\right)\Gamma(m_i)}$$

$$\sigma_i^2 + \mu_i^2 = \frac{\Gamma(m_i + 2c_i)\Gamma\left(M_{(i)} - 2c_i\right)\left(M_{(i)}/m_i\right)^{2c_i}}{\Gamma\left(M_{(i)}\right)\Gamma(m_i)},$$

με $M = \sum_{i=1}^{k} m_i$. Έλεγχοι εντοπισμού άτυπων τιμών για κανονικά δεδομένα με γνωστή μέση τιμή και διασπορά μπορούν να εφαρμοστούν στο σύνολο των τιμών z_i . Ο προφανής έλεγχος για μία μοναδική άτυπη τιμή βασίζεται στη μεγαλύτερη τιμή από τις $|z_i|$. Η παρουσία άτυπης τιμής δηλώνεται με έναν έλεγχο Bonferroni σε επίπεδο $a\%$ στην περίπτωση που η παραπάνω τιμή ξεπεράσει το άνω $a/2k$ ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής.

Περιγραφή αλγορίθμου

Στη συνέχεια παραθέτουμε τον ψευδοκώδικα του αλγορίθμου που χρησιμοποιήσαμε στην R, προκειμένου να κάνουμε προσομοιώσεις για τον παραπάνω έλεγχο υπόθεσης.

1. Παραγωγή ενός δείγματος *k* ανεξάρτητων μονάδων από την IG με ανεξάρτητο αριθμό επαναλαμβανομένων γεγονότων ανά μονάδα.
2. Υπολογίζουμε τις ποσότητες c_i , μ_i , σ_i , F_i και Z_i , για τις οποίες γνωρίζουμε πως η ελεγχοσυνάρτηση $Z_i \sim N(0,1)$.
3. Όταν η ελεγχοσυνάρτηση F_i έχει μία πολύ μεγάλη ή μία πολύ μικρή τιμή (ωστόσο πάντοτε θετική), η αντίστοιχη Z_i αποκτάει μία πολύ μεγάλη τιμή (θετική ή αρνητική) κατ' ακολουθία. Οπότε, θέτουμε $Z = \max|Z_i|$ και προχωράμε σε ανισότητες Bonferroni, υπολογίζοντας το ποσοστιαίο σημείο της τυποποιημένης Κανονικής κατανομής, z_a :

$$\mathrm{P}(Z > z_a) = a \Rightarrow \mathrm{P}\left(\max|Z_i| > z_a\right) = a \Rightarrow \cdots \Rightarrow \mathrm{P}\left(|Z_i| > z_a\right) = \frac{a}{k} \Rightarrow \mathrm{P}\left(Z_i > z_a\right) = \frac{a}{2k} \Rightarrow$$

$$\mathrm{P}\left(Z_i \le z_a\right) = 1 - \frac{a}{2k} \Rightarrow \Phi\left(z_a\right) = 1 - \frac{a}{2k} \Rightarrow$$

$$z_a = \Phi^{-1}\left(1 - \frac{a}{2k}\right)$$

4. Επαναλαμβάνουμε τα βήματα 1 και 2 10,000 φορές και μετράμε πόσες φορές η ελεγχοσυνάρτηση:

$$\max|Z_i| > z_a .$$

5. Επαναλαμβάνουμε την παραπάνω διαδικασία υπό την εναλλακτική υπόθεση, προκειμένου να αποκτήσουμε την ισχύ του ελέγχου.

2.7.6 Αποτελέσματα της μελέτης

Οι επιδόσεις των ελεγχοσυναρτήσεων που παρουσιάστηκαν προηγουμένως εξετάστηκαν με τη βοήθεια προσομοιώσεων, αρχικά κάτω από τη μηδενική υπόθεση, ώστε να ελεγχθεί το μέγεθος των ελέγχων και στη συνέχεια κάτω από την εναλλακτική υπόθεση, ώστε να αποκτήσουμε κάποια ένδειξη της ισχύς τους. Για κάθε επιλεγμένη τιμή του αριθμού k των μονάδων, n_i των παρατηρήσεων και των παραμέτρων (μ,λ), δημιουργούμε 10,000 σύνολα δεδομένων. Με το συγκεκριμένο μέγεθος δείγματος, τα Διωνυμικά τυπικά σφάλματα των εκτιμώμενων ποσοστών κοντά στο 5% και 1% είναι 0.22 και 0.10, αντίστοιχα.
Η μέθοδος των Michael et al (Παράγραφος 2.3.2) χρησιμοποιήθηκε για να δημιουργήσουμε ψευδο-τυχαίες μεταβλητές από την αντίστροφη Γκαουσιανή κατανομή. Όλοι οι υπολογισμοί πραγματοποιήθηκαν στην R.

Όπως έχουμε δει και στα προηγούμενα, δεν υπάρχει κάποια συγκεκριμένη προτιμώμενη μορφή της αντίστροφης Γκαουσιανής κατανομής στη βιβλιογραφία, από την οποία ένας απλός μετασχηματισμός να μας μεταφέρει σε οποιαδήποτε άλλη IG κατανομή. Ωστόσο, στην Παράγραφο 2.2 αποδείχτηκε πως εάν η $X \sim IG(\mu,\lambda)$ και $Y = \theta X$, τότε η $Y \sim IG(\theta\mu,\theta\lambda)$. Σαν επακόλουθο, οποιαδήποτε $IG(\mu,\lambda)$ μπορεί να παραχθεί από την $\mathrm{IG}(1,\lambda/\mu)$ ή την $\mathrm{IG}(\mu/\lambda,1)$.

Όπως είδαμε στην Παράγραφο 2.2.1, ο λόγος λ/μ διέπει τον τρόπο συμπεριφοράς της IG κατανομής. Δημιουργήσαμε δεδομένα για τιμές της λ/μ ίσες με 1, 5 και 0.2. Ο Πίνακας 2.10 δείχνει τη συμπεριφορά του ελέγχου για τον εντοπισμό μίας άτυπης τιμής για την παράμετρο μ, κάτω από τη μηδενική υπόθεση.

Όπως ήταν αναμενόμενο, τα πραγματικά μεγέθη των ελέγχων είναι πολύ κοντά στις θεωρητικές τιμές, με μία μικρή μόνο τάση προς συντηρητική συμπεριφορά. Ο έλεγχος για μία άτυπη τιμή για την παράμετρο λ συμπεριφέρεται με τον ίδιο ακριβώς τρόπο κάτω από τη μηδενική υπόθεση στην περίπτωση που όλα τα n_i είναι ίσα (Πίνακας 2.11).

Μονάδες x Παρατηρήσεις	Συνολικός αριθμός γεγονότων	$\lambda/\mu=1$		$\lambda/\mu=5$		$\lambda/\mu=0.2$	
		1 %	5 %	1 %	5 %	1 %	5%
5×1	5	0.93	4.79	0.83	4.90	1.09	4.91
5×2	10	1.00	5.14	0.94	4.99	1.01	5.08
10×1	10	1.11	4.63	0.88	5.05	1.07	5.25
25×1	25	0.95	5.12	1.15	5.23	1.01	4.94
25×2	50	1.03	4.71	0.93	4.95	1.05	4.91
6,8,10,12,14	50	0.96	4.55	0.87	4.88	0.88	4.83
50×2	100	0.91	4.72	0.86	4.85	0.85	5.27
50×6	300	0.88	4.62	1.00	4.74	0.80	4.62

Πίνακας 2.10: Προσομοιωμένες πιθανότητες υπέρβασης (σε 10,000 επαναλήψεις) στο θεωρητικό 5% επίπεδο σημαντικότητας των ελέγχων Bonferroni για διαφορετική τιμή του μ σε μία μονάδα

Μονάδες x Παρατηρήσεις	Συνολικός αριθμός γεγονότων	$\lambda/\mu=1$		$\lambda/\mu=5$		$\lambda/\mu=0.2$	
		1 %	5 %	1 %	5 %	1 %	5%
10×2	20	1.10	5.04	1.09	5.15	0.99	4.74
5×4	20	0.94	4.83	1.07	4.93	1.07	4.69
25×2	50	1.03	5.18	0.90	5.21	0.98	4.69
10×5	50	1.06	4.69	1.03	4.97	0.98	4.82
5×10	50	0.90	4.71	0.91	4.58	1.03	5.02
20×5	100	0.99	5.22	1.14	4.88	0.92	5.00
10×10	100	0.91	4.75	0.94	4.80	1.00	4.98
5×20	100	0.97	4.78	1.00	4.74	1.09	4.64
25×8	200	0.88	4.90	1.03	4.68	0.85	4.51
10×20	200	1.18	5.12	0.81	4.48	1.07	5.01

Πίνακας 2.11: Προσομοιωμένες πιθανότητες υπέρβασης (σε 10,000 επαναλήψεις) στο θεωρητικό 5% επίπεδο σημαντικότητας των ελέγχων Bonferroni για διαφορετική τιμή του λ σε μία μονάδα: ίσοι αριθμοί επαναλαμβανόμενων γεγονότων ανά μονάδα.

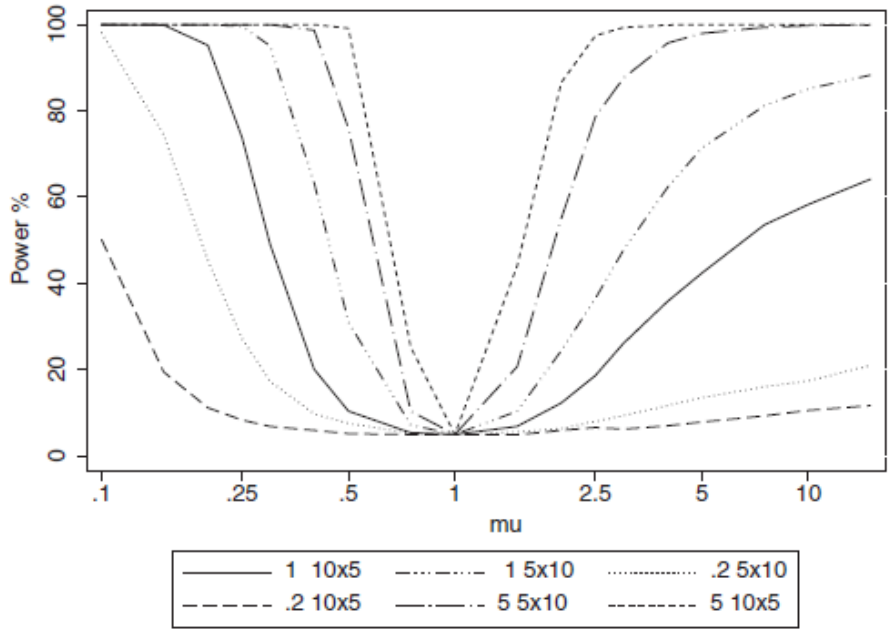
Στην περίπτωση αυτή, κάθε ελεγχοσυνάρτηση Q_{1i} ακολουθεί την ίδια κατανομή και ο μετασχηματισμός που περιγράφηκε στην προηγούμενη παράγραφο δεν είναι απαραίτητος. Ο Πίνακας 2.12 παρουσιάζει αποτελέσματα για τον έλεγχο αυτό, όταν τα n_i δεν είναι όλα ίσα μεταξύ τους. Στην περίπτωση αυτή, εκτός από το δειγματοληπτικό σφάλμα και την εκτίμηση που εισήχθη από την προσέγγιση Bonferroni, περαιτέρω ανακρίβεια εισάγεται λόγω της χρησιμοποίησης του προσεγγιστικού μετασχηματισμού από την F στην τυποποιημένη Κανονική κατανομή.

		$\lambda/\mu=1$		$\lambda/\mu=5$		$\lambda/\mu=0.2$	
Μονάδες x Παρατηρήσεις	**Συνολικός αριθμός γεγονότων**	1 %	5 %	1 %	5 %	1 %	5%
2, 3, 4, 5, 6	20	1.40	5.23	1.29	5.04	1.52	5.57
6, 8, 10, 12, 14	50	1.01	4.78	1.11	4.79	0.98	4.63
5, 5, 10, 15, 15	50	0.97	4.89	0.97	4.71	1.07	4.85
12, 16, 20, 24, 28	100	0.99	4.45	0.95	4.90	1.12	4.95
5 x 6, 5 x 14	100	0.74	4.25	0.95	4.84	0.74	4.73
5 x (6, 8, 12, 14)	200	0.55	4.11	0.72	4.32	0.83	4.63

Πίνακας 2.12: Προσομοιωμένες πιθανότητες υπέρβασης (σε 10,000 επαναλήψεις) στο θεωρητικό 5% επίπεδο σημαντικότητας των ελέγχων Bonferroni για διαφορετική τιμή του λ σε μία μονάδα: άνισοι αριθμοί επαναλαμβανόμενων γεγονότων ανά μονάδα, με την προσέγγιση της τυποποιημένης Κανονικής κατανομής από την F.

Τα αποτελέσματα υποδεικνύουν πως ο έλεγχος είναι ακόμα πιο συντηρητικός κάτω από αυτές τις περιστάσεις, αλλά και πάλι όχι υπερβολικά πάνω από τα επιτρεπτά όρια.

Καμπύλες που δείχνουν την ισχύ του ελέγχου για μια άτυπη τιμή για την παράμετρο μ δίνονται στο Σχήμα 2.15. Σε αυτές τις προσομοιώσεις, η τιμή $\mu=1$ χρησιμοποιήθηκε για $k-1$ μονάδες, ενώ η τιμή της διέφερε για την εναπομείνασα μονάδα. Συγκρίνοντας τις καμπύλες για 5 και 10 μονάδες αντίστοιχα, είναι προφανές από αυτά τα αποτελέσματα πως η ισχύς, για ένα δεδομένο αριθμό παρατηρήσεων είναι ουσιωδώς μεγαλύτερη σε περίπτωση που η άτυπη τιμή βρίσκεται σε μικρότερο αριθμό μονάδων αλλά με περισσότερες παρατηρήσεις. Επίσης, για συγκεκριμένες επιλογές μονάδων και παρατηρήσεων, η ισχύς αυξάνει όσο ο λόγος λ/μ αυξάνει. Μικρές τιμές της παραπάνω αναλογίας υποδεικνύουν έντονη ασυμμετρία της IG κατανομής, περίπτωση στην οποία ενδέχεται να είναι ιδιαίτερα δύσκολος ο εντοπισμός άτυπων τιμών.

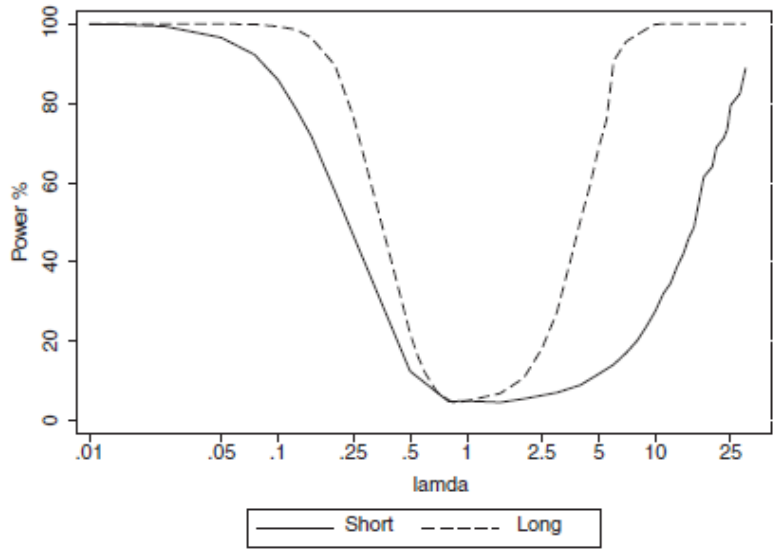


Σχήμα 2.15: Ισχύς σε ποσοστό (%) του ελέγχου βασισμένου στο $\max Q_i$, σε ονομαστικό 5% επίπεδο, ως συνάρτηση της άτυπης τιμής μ για έξι διαφορετικές επιλογές δεδομένων. Στη λεζάντα παρουσιάζονται οι λόγοι λ/μ και οι αριθμοί μονάδων επί τον αριθμό των παρατηρήσεων.

Καμπύλες που δείχνουν την ισχύ του ελέγχου μίας άτυπης τιμής για την παράμετρο λ δίνονται στο Σχήμα 2.16. Να υπενθυμίσουμε πως η ελεγχοσυνάρτηση $\hat{\lambda}_i = n_i \Big/ \sum_{l=1}^{n_i} \left(\frac{1}{t_{il}} - \frac{1}{\bar{t}_i} \right)$ που προκύπτει από τη σχέση (2.43) είναι ο εκτιμητής μέγιστης πιθανοφάνειας των λ_i και η κατανομή του δίνεται από $n_i \lambda_i \sim \chi^2_{n_i - 1}$. Εάν η μονάδα i είναι όντως άτυπη και εάν οι υπόλοιπες μονάδες έχουν την ίδια εκτιμώμενη τιμή $\hat{\lambda}$ για την παράμετρο λ, τότε:

$$\frac{\lambda_i \left(N_{(i)} - 1 \right) Q_{1i}}{\lambda \left(n_i - 1 \right)} \sim F_{n_i - 1,\, N_{(i)} - 1} \tag{2.68}$$

Ως εκ τούτου, κάτω από την εναλλακτική υπόθεση, η ελεγχοσυνάρτηση της σχέσης (2.68) είναι απλά λ_i / λ φορές πολλαπλάσια της ελεγχοσυνάρτησης F_i υπό τη μηδενική υπόθεση (σχέση (2.67)). Συνεπώς, η ισχύς του ελέγχου των δειγμάτων εξαρτάται μόνο από το λόγο λ / λ_i και το ίδιο πρέπει να ισχύει και για τον έλεγχο εντοπισμού άτυπων τιμών. Καμπύλες δίνονται για μία επιλογή με άνισα μεγέθη παρατηρήσεων ανά μονάδα. Η ισχύς είναι πολύ μεγαλύτερη, όταν η άτυπη τιμή εμφανίζεται σε μία μεγαλύτερη σειρά με πολλές παρατηρήσεις, παρά σε μία μικρότερη σειρά με λίγες παρατηρήσεις.



Σχήμα 2.16: Ισχύς σε ποσοστό (%) του ελέγχου βασισμένου στο $\max Q_i$, σε ονομαστικό 5% επίπεδο, ως συνάρτηση της άτυπης τιμής λ για μία επιλογή δεδομένων, η οποία αποτελείται από πέντε μονάδες με έξι παρατηρήσεις για την καθεμιά και πέντε μονάδες με δεκατέσσερις. Η άτυπη τιμή τοποθετείται είτε στη μονάδα με τις λίγες παρατηρήσεις (6), είτε στη μονάδα με τις πολλές (14).

2.7.7 Μία εναλλακτική παραμέτρηση για την IG

Στη συνέχεια, θα χρησιμοποιήσουμε μία εναλλακτική παραμέτρηση για τη συνάρτηση πυκνότητας πιθανότητας της IG, η οποία βρίσκεται σε αντιστοιχία με τις παραμέτρους (μ, λ) και δίνεται από τον τύπο:

$$f(t|m,\sigma^2,x_0)=\frac{x_0}{\sqrt{2\pi\sigma^2 t^3}}\exp\left[-\frac{(x_0+mt)^2}{2\sigma^2 t}\right],\ t>0, \qquad (2.69)$$

$$m<0,\ \sigma^2>0,\ x_0>0$$

Η σχέση (2.69) είναι η παραμέτρηση της συνάρτησης πυκνότητας πιθανότητας του χρόνου πρώτης μετάβασης που ακολουθεί την IG και παρουσιάστηκε στην Παράγραφο 2.2.1. Η αντιστοιχία μεταξύ των παραμέτρων των εξισώσεων (2.4) και (2.69) δίνεται από τους εξής τύπους:

$$\lambda=x_0^2,\ \mu=x_0/|m| \qquad (2.70)$$

Για τις περιπτώσεις που μας ενδιαφέρουν, θέτουμε αυθαίρετα $\sigma^2=1$, χωρίς βλάβη της γενικότητας.

Συνεπώς, ο έλεγχος που παρουσιάσαμε στην Παράγραφο 2.7.1 για την ισότητα των παραμέτρων $\mu_1=\mu_2$ δύο διαφορετικών δειγμάτων από την IG, μετατρέπεται σε έναν έλεγχο για την ισότητα $x_{01}/|m_1|=x_{02}/|m_2|$. Παρόλα αυτά, ο έλεγχος κατασκευάστηκε με την

προϋπόθεση ίσων τιμών για την παράμετρο λ και άρα για την x_0. Σαν αποτέλεσμα, ο έλεγχος για τα δύο δείγματα για τις τιμές της παραμέτρου μ είναι ισοδύναμος με έναν έλεγχο για την ισότητα των τιμών της m μεταξύ των μονάδων και αντίστοιχα για τον έλεγχο εντοπισμού των άτυπων τιμών.

Έλεγχος υπόθεσης για την παράμετρο *m* της IG

Έστω ένα δείγμα n παρατηρήσεων $X_i \sim IG(m_1, x_{01})$ και ένα δείγμα m παρατηρήσεων $Y_i \sim IG(m_2, x_{02})$, με $x_{01} = x_{02} = x_0$. Θέλουμε να προχωρήσουμε στον έλεγχο της υπόθεσης:

$\mathrm{H}_0 : m_1 = m_2$

$\mathrm{H}_1 : m_1 \neq m_2$,

Η από κοινού σ.π.π ή συνάρτηση πιθανοφάνειας δίνεται από τον τύπο:

$$L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) = L(m_1, m_2, x_0 | x_1, \ ..., x_n, y_1, ..., y_n) = f_{x,y} = \prod_{i=1}^{n} f_{x_i}(x_i, \boldsymbol{\theta}) \cdot \prod_{j=1}^{m} f_{y_j}(y_j, \boldsymbol{\theta}) =$$

$$= \prod_{i=1}^{n} \frac{x_{01}}{\sqrt{2\pi x_i^3}} \exp\left(-\frac{(x_{01} - m_1 x_i)^2}{2x_i} \right) \prod_{j=1}^{m} \frac{x_{02}}{\sqrt{2\pi y_j^3}} \exp\left(-\frac{(x_{02} - m_2 y_j)^2}{2y_j} \right) \overset{x_{01}=x_{02}=x_0}{=}$$

$$L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) = \left(\frac{x_0}{\sqrt{2\pi}} \right)^{n+m} \prod_{i=1}^{n} \frac{1}{x_i^{3/2}} \prod_{j=1}^{m} \frac{1}{y_j^{3/2}} \exp\left(-\sum_{i=1}^{n} \frac{(x_0 - m_1 x_i)^2}{2x_i} \right) \exp\left(-\sum_{j=1}^{m} \frac{(x_0 - m_2 y_j)^2}{2y_j} \right) \qquad (2.71)$$

Υπό τη μηδενική υπόθεση

Η συνάρτηση πιθανοφάνειας της σχέσης (2.44) γίνεται:

$$L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y}) = \left(\frac{x_0}{\sqrt{2\pi}} \right)^{n+m} \prod_{i=1}^{n} \frac{1}{x_i^{3/2}} \prod_{j=1}^{m} \frac{1}{y_j^{3/2}} \exp\left[-\frac{1}{2} \left(\sum_{i=1}^{n} \frac{(x_0 - m_1 x_i)^2}{2x_i} + \sum_{j=1}^{m} \frac{(x_0 - m_1 y_j)^2}{2y_j} \right) \right]$$

Λογαριθμίζοντας την παραπάνω σχέση και μετά από μερική παραγώγιση, προκύπτουν οι εκτιμήτριες μέγιστης πιθανοφάνειας:

$$\frac{\partial \ln L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})}{\partial m_1} = 0 \quad \Leftrightarrow$$

$$\tilde{m}_1 = \frac{x_0 (m+n)}{n x_i + m y_j} \qquad (2.72)$$

$$\frac{\partial \ln L\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right)}{\partial x_0}=0 \quad \Leftrightarrow$$

$$\tilde{x}_0^2=\frac{n+m}{\left(\sum_{i=1}^{n}\frac{1}{x_i}+\sum_{j=1}^{m}\frac{1}{y_j}\right)-\frac{\left(m+n\right)^2}{nx_i+my_j}} \tag{2.73}$$

Υπό την εναλλακτική υπόθεση

Η συνάρτηση πιθανοφάνειας της σχέσης (2.71) γίνεται:

$$L\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right)=\left(\frac{x_0}{\sqrt{2\pi}}\right)^{n+m}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\exp\left[-\frac{1}{2}\left(\sum_{i=1}^{n}\frac{\left(x_0-m_1x_i\right)^2}{2x_i}\right)\right]\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\cdot\exp\left(-\frac{1}{2}\sum_{j=1}^{m}\frac{\left(x_0-m_2y_j\right)^2}{2y_j}\right)$$

Λογαριθμίζοντας την παραπάνω σχέση και μετά από μερική παραγώγιση, προκύπτουν οι εκτιμήτριες μέγιστης πιθανοφάνειας:

$$\frac{\partial \ln L\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right)}{\partial m_1}=0 \quad \Leftrightarrow \quad \frac{\partial}{\partial m_1}\left[-\frac{1}{2}\left(\sum_{i=1}^{n}\frac{\left(x_0-m_1x_i\right)^2}{2x_i}\right)\right]=0 \quad \Leftrightarrow$$

$$\hat{m}_1=\frac{\hat{x}_0}{\bar{x}} \tag{2.74}$$

Όμοια,

$$\frac{\partial \ln L\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right)}{\partial m_2}=0 \quad \Leftrightarrow \quad \frac{\partial}{\partial m_2}\left[-\frac{1}{2}\left(\sum_{j=1}^{m}\frac{\left(x_0-m_1y_j\right)^2}{2y_j}\right)\right]=0 \quad \Leftrightarrow$$

$$\hat{m}_2=\frac{\hat{x}_0}{\bar{y}} \tag{2.75}$$

$$\frac{\partial \ln L\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right)}{\partial x_0}=0 \quad \Leftrightarrow$$

$$\frac{\left(n+m\right)}{x_0}-x_0\left(\sum_{i=1}^{n}\frac{1}{x_i}+\sum_{j=1}^{m}\frac{1}{y_j}\right)+m_1n+m_2m=0$$

Τέλος, αντικαθιστώντας στην παραπάνω σχέση τις ποσότητες $\hat{m}_1$ και $\hat{m}_2$ των σχέσεων (2.74) και (2.75), καταλήγουμε στην Ε.Μ.Π. για την παράμετρο x_0:

$$\frac{1}{\hat{x}_0}\left[\left(n+m\right)-\hat{x}_0^2\left(\sum_{i=1}^{n}\frac{1}{x_i}+\sum_{j=1}^{m}\frac{1}{y_j}\right)+\hat{x}_0\cdot\frac{\hat{x}_0}{\bar{x}}n+\hat{x}_0\cdot\frac{\hat{x}_0}{\bar{y}}m\right]=0 \quad \Leftrightarrow$$

$$\hat{x}_0^2 = \frac{(n+m)}{\left[\left(\sum_{i=1}^{n}\frac{1}{x_i}+\sum_{j=1}^{m}\frac{1}{y_j}\right)-\left(\frac{n}{\bar{x}}+\frac{m}{\bar{y}}\right)\right]} \quad \Leftrightarrow \tag{2.76}$$

Προκειμένου να χρησιμοποιήσουμε τον έλεγχο του λόγου των πιθανοφανειών, έχουμε τα εξής:

$$\sup_{\boldsymbol{\theta}|H_0} p\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right) = \sup_{\boldsymbol{\theta}|H_0} p\left(m_1, m_2, x_0|\boldsymbol{x},\boldsymbol{y}\right) \overset{\text{ανεξαρτησία}}{=} \sup_{\boldsymbol{\theta}|H_0} p\left(m_1, x_0|\, x_1,\ ..., x_n\right)\cdot\sup_{\boldsymbol{\theta}|H_0} p\left(m_2, x_0|y_1, ..., y_n\right) =$$

$$= \sup_{\boldsymbol{\theta}|H_0}\left[\prod_{i=1}^{n}\frac{x_0}{\sqrt{2\pi}x_i^{3/2}}\exp\left(-\sum_{i=1}^{n}\frac{\left(x_0-m_1x_i\right)^2}{2x_i}\right)\right]\cdot\sup_{\boldsymbol{\theta}|H_0}\left[\prod_{j=1}^{m}\frac{x_0}{\sqrt{2\pi}y_j^{3/2}}\exp\left(-\sum_{j=1}^{m}\frac{\left(x_0-m_2y_j\right)^2}{2y_j}\right)\right] =$$

$$= \left(\frac{\tilde{x}_0}{\sqrt{2\pi}}\right)^{n+m}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left[-\left(\sum_{i=1}^{n}\frac{\left(\tilde{x}_0-m_1x_i\right)^2}{2x_i}+\sum_{j=1}^{m}\frac{\left(\tilde{x}_0-m_1y_j\right)^2}{2y_j}\right)\right] \Rightarrow$$

$$\sup_{\boldsymbol{\theta}|H_0} p\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right) = \left(\frac{\tilde{x}_0}{\sqrt{2\pi}}\right)^{n+m}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\cdot\exp\left(-\frac{n+m}{2}\right) \tag{2.77}$$

Με παρόμοια συλλογιστική:

$$\sup_{\boldsymbol{\theta}|H_1} p\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right) = \sup_{\boldsymbol{\theta}|H_1} p\left(m_1, m_2, x_0|\boldsymbol{x},\boldsymbol{y}\right) \overset{\text{ανεξαρτησία}}{=} \sup_{\boldsymbol{\theta}|H_1} p\left(m_1, x_0|\, x_1,\ ..., x_n\right)\cdot\sup_{\boldsymbol{\theta}|H_1} p\left(m_2, x_0|y_1, ..., y_n\right) =$$

$$= \left(\frac{\hat{x}_0}{\sqrt{2\pi}}\right)^{n+m}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left[-\frac{1}{2}\left(\sum_{i=1}^{n}\frac{\left(\hat{x}_0-m_1x_i\right)^2}{x_i}+\sum_{j=1}^{m}\frac{\left(\hat{x}_0-m_1y_j\right)^2}{y_j}\right)\right] =$$

Ύστερα από πράξεις, καταλήγουμε στα εξής:

$$\sup_{\boldsymbol{\theta}|H_1} p\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right) = \left(\frac{\hat{x}_0}{\sqrt{2\pi}}\right)^{n+m}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left(-\frac{n+m}{2}\right) = \tag{2.78}$$

Λόγω των (2.77) και (2.78), ο λόγος των πιθανοφανειών γίνεται:

$$\lambda^* = \frac{\sup_{\boldsymbol{\theta}|H_0} p\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right)}{\sup_{\boldsymbol{\theta}|H_1} p\left(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}\right)} = \frac{\left(\frac{\tilde{x}_0}{\sqrt{2\pi}}\right)^{n+m}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left(-\frac{n+m}{2}\right)}{\left(\frac{\hat{x}_0}{\sqrt{2\pi}}\right)^{n+m}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left(-\frac{n+m}{2}\right)} \quad \Leftrightarrow$$

$$\lambda^* = \frac{\left(\frac{\tilde{x}_0}{\sqrt{2\pi}}\right)^{n+m}}{\left(\frac{\hat{x}_0}{\sqrt{2\pi}}\right)^{n+m}} \quad \Leftrightarrow \lambda^* = \left(\frac{\frac{n+m}{\left(\sum_{i=1}^{n}\frac{1}{x_i}+\sum_{j=1}^{m}\frac{1}{y_j}\right)-\frac{(n+m)^2}{n\bar{x}+m\bar{y}}}}{\frac{n+m}{\left(\sum_{i=1}^{n}\frac{1}{x_i}+\sum_{j=1}^{m}\frac{1}{y_j}\right)-\left(\frac{n+m}{\bar{x}+\bar{y}}\right)}} \right)^{\frac{n+m}{2}}$$

$$\lambda^* = \left(\frac{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)}{\left(\sum_{i=1}^{n}\frac{1}{x_i}+\sum_{j=1}^{m}\frac{1}{y_j}\right)-\frac{(n+m)^2}{n\bar{x}+m\bar{y}}} \right)^{\frac{n+m}{2}} \tag{2.79}$$

Κάνοντας πράξεις στον παρανομαστή, έχουμε:

$$\left(\sum_{i=1}^{n}\frac{1}{x_i}+\sum_{j=1}^{m}\frac{1}{y_j}\right)-\frac{(n+m)^2}{n\bar{x}+m\bar{y}} = \sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)+\frac{n}{\bar{x}}+\frac{m}{\bar{y}}-\frac{(n+m)^2}{n\bar{x}+m\bar{y}} =$$

$$\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)+\frac{(n\bar{y}+m\bar{x})(n\bar{x}+m\bar{y})-(n^2+m^2+2nm)\bar{x}\bar{y}}{\bar{x}\bar{y}(n\bar{x}+m\bar{y})} =$$

$$\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)+\frac{nm(\bar{x}-\bar{y})^2}{\bar{x}\bar{y}(n\bar{x}+m\bar{y})}$$

Τελικά, η σχέση (2.79) γίνεται:

$$\lambda^* = \left(\frac{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)}{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)+\frac{nm(\bar{x}-\bar{y})^2}{\bar{x}\bar{y}(n\bar{x}+m\bar{y})}} \right)^{\frac{n+m}{2}} =$$

$$= \left(1 \Big/ \left[1+\frac{nm(\bar{x}-\bar{y})^2}{\bar{x}\bar{y}(n\bar{x}+m\bar{y})\left\{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)\right\}} \right] \right)^{\frac{n+m}{2}}$$

Θέτοντας

$$Q = \frac{nm(\bar{x}-\bar{y})^2}{\bar{x}\bar{y}(n\bar{x}+m\bar{y})\left\{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)\right\}}, \tag{2.80}$$

έχουμε ότι:

$$\lambda^* = \left(\frac{1}{1+Q}\right)^{\frac{n+m}{2}} \tag{2.81}$$

Η απόρριψη της μηδενικής υπόθεσης για μικρές τιμές του λ^* ισοδυναμεί με απόρριψη της μηδενικής υπόθεσης για μεγάλες τιμές του $\left|\sqrt{Q}\right|$, το οποίο είναι ισοδύναμο με τον ομοιόμορφα πιο ισχυρό αμερόληπτο έλεγχο που παρουσίασε ο Chhikara (1975). Η ελεγχοσυνάρτηση $\left[(n+m-2)Q\right]^{1/2}$ ακολουθεί την t κατανομή με $(n+m-2)$ βαθμούς ελευθερίας.

Με παρόμοιο τρόπο, παρουσιάζουμε στη συνέχεια τον έλεγχο της Παραγράφου 2.7.2 για την ισότητα των τιμών της παραμέτρου λ, ως έλεγχο για την ισότητα των τιμών της παραμέτρου x_0. Αφού ο έλεγχος δεν επιβάλλει ισότητα για τις τιμές της μ, οι οποίες όπως είδαμε αντιστοιχούν στην ποσότητα $x_0/|m|$, τότε δεν υπάρχει κανένας περιορισμός.

Έλεγχος υπόθεσης για την παράμετρο x_0 της IG

Έστω ένα δείγμα n παρατηρήσεων $X_i \sim IG(m_1, x_{01})$ και ένα δείγμα m παρατηρήσεων $Y_i \sim IG(m_2, x_{02})$. Θέλουμε να προχωρήσουμε στον έλεγχο της υπόθεσης:

$\mathrm{H}_0: x_{01} = x_{02}$

$\mathrm{H}_1: x_{01} \neq x_{02}$,

Η από κοινού σ.π.π. ή συνάρτηση πιθανοφάνειας δίνεται από τον τύπο:

$$L(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}) = L(m_1, m_2, x_{01}, x_{02}|\, x_1, \ldots, x_n, y_1, \ldots, y_n) = f_{\boldsymbol{x},\boldsymbol{y}} = \prod_{i=1}^{n} f_{x_i}(x_i, \boldsymbol{\theta}) \cdot \prod_{j=1}^{m} f_{y_j}(y_j, \boldsymbol{\theta}) =$$

$$= \prod_{i=1}^{n} \frac{x_{01}}{\sqrt{2\pi x_i^3}} \exp\left(-\frac{(x_{01} - m_1 x_i)^2}{2x_i}\right) \prod_{j=1}^{m} \frac{x_{02}}{\sqrt{2\pi y_j^3}} \exp\left(-\frac{(x_{02} - m_2 y_j)^2}{2y_j}\right) \overset{x_{01}=x_{02}}{=}$$

Υπό τη μηδενική υπόθεση

Η συνάρτηση πιθανοφάνειας της σχέσης (2.71) γίνεται:

$$L(\boldsymbol{\theta}|\boldsymbol{x},\boldsymbol{y}) = \left(\frac{x_{01}}{\sqrt{2\pi}}\right)^{\frac{n+m}{2}} \prod_{i=1}^{n} \frac{1}{x_i^{3/2}} \prod_{j=1}^{m} \frac{1}{y_j^{3/2}} \exp\left[-\left(\sum_{i=1}^{n} \frac{(x_{01} - m_1 x_i)^2}{2x_i} + \sum_{j=1}^{m} \frac{(x_{01} - m_2 y_j)^2}{2y_j}\right)\right]$$

Λογαριθμίζοντας την παραπάνω σχέση και μετά από μερική παραγώγιση, προκύπτουν οι εκτιμήτριες μέγιστης πιθανοφάνειας για τις παραμέτρους $\tilde{m}_1$, $\tilde{m}_2$ και $\tilde{x}_{01}$:

$$\tilde{m}_1 = \frac{\tilde{x}_{01}}{\bar{x}} \tag{2.82}$$

$$\tilde{m}_2 = \frac{\tilde{x}_{01}}{\bar{y}} \tag{2.83}$$

και

$$\tilde{x}_{01}^2 = \frac{n+m}{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)} \tag{2.84}$$

Υπό την εναλλακτική υπόθεση

Η συνάρτηση πιθανοφάνειας της σχέσης (2.71) γίνεται:

$$L\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right) = \frac{\left(x_{01}^2\right)^{\frac{n}{2}}\left(x_{02}^2\right)^{\frac{m}{2}}}{\left(2\pi\right)^{\frac{n+m}{2}}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left[-\left(\sum_{i=1}^{n}\frac{\left(x_{01}-m_1x_i\right)^2}{2x_i}\right)\right]\exp\left[-\left(\sum_{j=1}^{m}\frac{\left(x_{02}-m_2y_j\right)^2}{2y_j}\right)\right]$$

Λογαριθμίζοντας την παραπάνω σχέση και μετά από μερική παραγώγιση, προκύπτουν οι εκτιμήτριες μέγιστης πιθανοφάνειας των παραμέτρων $\hat{m}_1$, $\hat{m}_2$, $\hat{x}_{01}$ και $\hat{x}_{02}$:

$$\hat{m}_1 = \frac{\hat{x}_{01}}{\bar{x}} \tag{2.85}$$

$$\hat{m}_2 = \frac{\hat{x}_{02}}{\bar{y}} \tag{2.86}$$

$$\hat{x}_{01}^2 = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{n}{\bar{x}}\right)} \tag{2.87}$$

και

$$\hat{x}_{01}^2 = \frac{m}{\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{m}{\bar{y}}\right)} \tag{2.88}$$

Προκειμένου να χρησιμοποιήσουμε τον έλεγχο του λόγου των πιθανοφανειών, έχουμε τα εξής:

$$\sup_{\boldsymbol{\theta}|H_0} p\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)=\sup_{\boldsymbol{\theta}|H_0} p\left(m_1,m_2,x_{01},x_{02}\middle|\boldsymbol{x},\boldsymbol{y}\right)\overset{\text{ανεξαρτησία}}{=}\sup_{\boldsymbol{\theta}|H_0} p\left(m_1,x_{01}\middle|x_1,\ ...,x_n\right)\cdot\sup_{\boldsymbol{\theta}|H_0} p\left(m_2,x_{01}\middle|y_1,...,y_n\right)=$$

$$=\frac{\left(\tilde{x}_{01}^2\right)^{\frac{n+m}{2}}}{\left(2\pi\right)^{\frac{n+m}{2}}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\exp\left[-\frac{1}{2}\left(\sum_{i=1}^{n}\frac{\left(\tilde{x}_{01}-m_1x_i\right)^2}{x_i}\right)\right]\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left[-\frac{1}{2}\left(\sum_{j=1}^{m}\frac{\left(\tilde{x}_{02}-m_2y_j\right)^2}{y_j}\right)\right]\quad\Leftrightarrow$$

$$\sup_{\boldsymbol{\theta}|H_0} p\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)=\left(\frac{\tilde{x}_{01}^2}{2\pi}\right)^{\frac{n+m}{2}}\prod_{i=1}^{n}x_i^{-3/2}\prod_{j=1}^{m}y_j^{-3/2}\exp\left(-\frac{n+m}{2}\right)\tag{2.89}$$

Για την πιθανοφάνεια του δείγματος υπό την εναλλακτική υπόθεση, έχουμε:

$$\sup_{\boldsymbol{\theta}|H_1} p\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)=\sup_{\boldsymbol{\theta}|H_1} p\left(m_1,x_{01}\middle|x_1,\ ...,x_n\right)\cdot\sup_{\boldsymbol{\theta}|H_1} p\left(m_2,x_{02}\middle|y_1,...,y_n\right)$$

Και μετά από πράξεις, καταλήγουμε:

$$\sup_{\boldsymbol{\theta}|H_1} p\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)=\frac{\left(\hat{x}_{01}^2\right)^{\frac{n}{2}}\left(\hat{x}_{02}^2\right)^{\frac{m}{2}}}{\left(2\pi\right)^{\frac{n+m}{2}}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left[-\frac{n+m}{2}\right]\tag{2.90}$$

Λόγω των σχέσεων (2.89) και (2.90), ο λόγος των πιθανοφανειών γίνεται:

$$\lambda^*=\frac{\sup_{\boldsymbol{\theta}|H_0} p\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)}{\sup_{\boldsymbol{\theta}|H_1} p\left(\boldsymbol{\theta}\middle|\boldsymbol{x},\boldsymbol{y}\right)}=\frac{\left(\frac{\tilde{x}_{01}^2}{2\pi}\right)^{\frac{n+m}{2}}\prod_{i=1}^{n}x_i^{-3/2}\prod_{j=1}^{m}y_j^{-3/2}\exp\left(-\frac{n+m}{2}\right)}{\frac{\left(\hat{x}_{01}^2\right)^{\frac{n}{2}}\left(\hat{x}_{02}^2\right)^{\frac{m}{2}}}{\left(2\pi\right)^{\frac{n+m}{2}}}\prod_{i=1}^{n}\frac{1}{x_i^{3/2}}\prod_{j=1}^{m}\frac{1}{y_j^{3/2}}\exp\left[-\frac{n+m}{2}\right]}\quad\Leftrightarrow$$

$$\lambda^*=\frac{\left(\tilde{\lambda}\right)^{\frac{n+m}{2}}}{\hat{\lambda}^{\frac{n}{2}}\hat{\nu}^{\frac{m}{2}}}=\frac{\left(\frac{n+m}{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)}\right)^{\frac{n+m}{2}}}{\left(\frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)}\right)^{\frac{n}{2}}\left(\frac{m}{\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)}\right)^{\frac{m}{2}}}\quad\Leftrightarrow$$

$$\lambda^{*}=\frac{\left(n+m\right)^{\frac{n+m}{2}}}{\left(n\right)^{\frac{n}{2}}\left(m\right)^{\frac{m}{2}}}\frac{\left(\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)\right)^{\frac{n}{2}}\left(\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)\right)^{\frac{m}{2}}}{\left(\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)+\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)\right)^{\frac{n+m}{2}}} \quad \Leftrightarrow$$

Θέτοντας

$$Q_1=\frac{\sum_{j=1}^{m}\left(\frac{1}{y_j}-\frac{1}{\bar{y}}\right)}{\sum_{i=1}^{n}\left(\frac{1}{x_i}-\frac{1}{\bar{x}}\right)}, \tag{2.91}$$

έχουμε ότι:

$$\lambda^{*}=\frac{\left(n+m\right)^{\frac{n+m}{2}}}{\left(n\right)^{\frac{n}{2}}\left(m\right)^{\frac{m}{2}}}\left(\frac{1}{1+Q_1}\right)^{\frac{n+m}{2}}Q_1^{\frac{m}{2}} \tag{2.92}$$

Παρατηρούμε ότι

$$\ln\lambda^{*}=\ln\left[\frac{\left(n+m\right)^{\frac{n+m}{2}}}{\left(n\right)^{\frac{n}{2}}\left(m\right)^{\frac{m}{2}}}\right]+\frac{n+m}{2}\ln\left(\frac{1}{1+Q_1}\right)+\frac{m}{2}\ln Q_1 \quad \Leftrightarrow$$

$$\frac{d\ln\lambda^{*}}{dQ_1}=-\frac{n+m}{2}\frac{1}{1+Q_1}+\frac{m}{2}\frac{1}{Q_1}=\frac{-\left(m+n\right)Q_1+m\left(1+Q_1\right)}{2Q_1\left(1+Q_1\right)} \quad \Leftrightarrow$$

$$\frac{d\ln\lambda^{*}}{dQ_1}=\frac{m-nQ_1}{2Q_1\left(1+Q_1\right)}$$

Η λ^{*} είναι μία αύξουσα συνάρτηση για $\frac{d\ln\lambda^{*}}{dQ_1}>0 \Leftrightarrow m-nQ_1>0 \Leftrightarrow Q_1<\frac{m}{n}$ και φθίνουσα για $\frac{d\ln\lambda^{*}}{dQ_1}<0 \Leftrightarrow m-nQ_1<0 \Leftrightarrow Q_1>\frac{m}{n}$.

Η απόρριψη της μηδενικής υπόθεσης για μικρές τιμές του λ^{*} ισοδυναμεί με απόρριψη της μηδενικής υπόθεσης για πολύ μεγάλες ή πολύ μικρές τιμές του Q_1. Το Q_1 είναι ο λόγος δύο ανεξάρτητων χ^2 κατανομών. Επομένως, $Q_1\frac{\left(n-1\right)}{\left(m-1\right)}\sim F_{m-1,n-1}$, οπότε για ακόμα μία φορά ο λόγος των πιθανοφανειών μας οδηγεί σε μία ελεγχοσυνάρτηση που ακολουθεί μία γνωστή κατανομή.

2.7.8 Εφαρμογή σε πρόβλημα ρεαλιστικών συνθηκών

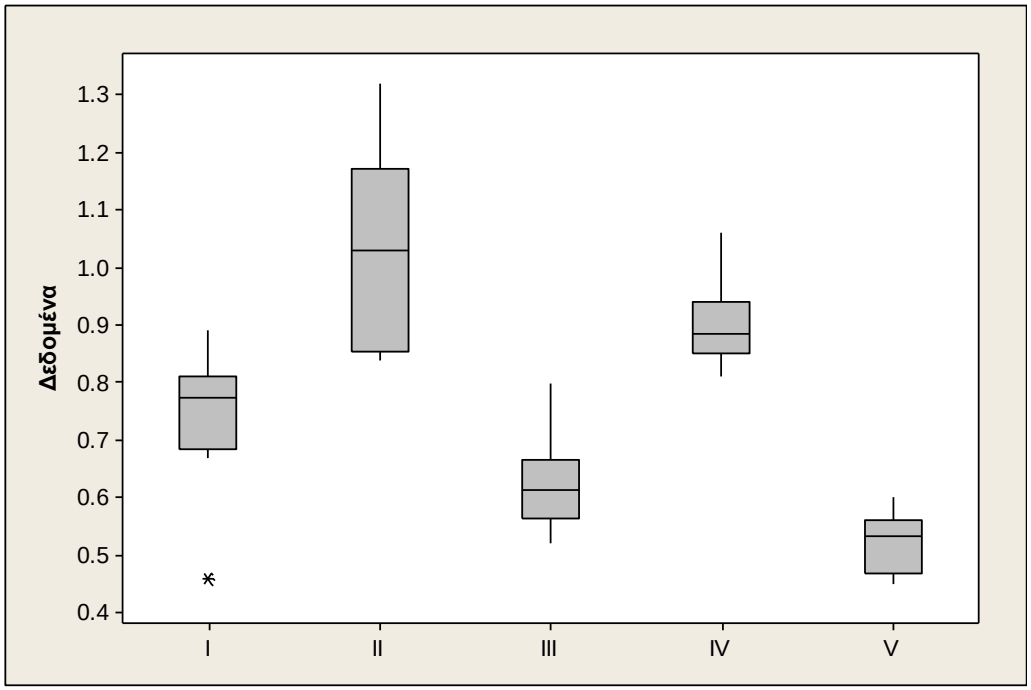
Για μία ακόμα πιο αναλυτική εικόνα των ελέγχων που παρουσιάστηκαν στις προηγούμενες παραγράφους, χρησιμοποιούμε τα δεδομένα του Πίνακα 2.13 που αποτελούνται από παρατηρήσεις που αφορούν την αντοχή των επιπτώσεων δέκα δειγμάτων αποτελούμενων από πέντε παρτίδες μονωτικών υλικών το καθένα (Ostle, 1963). Πολλοί συγγραφείς έχουν αναλύσει τα δεδομένα υποθέτοντας μία αντίστροφη Γκαουσιανή κατανομή, συμπεριλαμβανομένων των Fries και Bhattacharyya (1983), οι οποίοι αισθάνθηκαν πως η φύση των υλικών δικαιολογεί τη χρησιμοποίηση της IG με σταθερές παραμέτρους μέσα στις παρτίδες, αλλά με την παράμετρο μ να διαφέρει μεταξύ των παρτίδων.

Μέρη μονωτικού υλικού				
I	**II**	**III**	**IV**	**V**
0.89	0.86	0.52	0.86	0.52
0.69	1.17	0.52	1.06	0.53
0.46	1.18	0.80	0.81	0.47
0.85	1.32	0.64	0.97	0.47
0.73	1.03	0.63	0.90	0.57
0.67	0.84	0.58	0.93	0.54
0.78	0.89	0.65	0.87	0.56
0.77	0.84	0.60	0.88	0.55
0.80	1.03	0.71	0.89	0.45
0.79	1.06	0.59	0.82	0.60

Πίνακας 2.13: Δεδομένα ορίων αντοχής 10 δειγμάτων από 5 μέρη μονωτικού υλικού. Σταθερές τιμές των παραμέτρων μέσα σε κάθε μέρος, αλλά η παράμετρος μ πιθανότατα διαφέρει μεταξύ των 5 μερών.

Ωστόσο, γραφήματα τύπου box-plots υποδεικνύουν την ύπαρξη άτυπων τιμών μέσα στις παρτίδες. Ειδικότερα, στο Σχήμα 2.17 φαίνεται πως μία παρατήρηση στην πρώτη παρτίδα εμφανίζει αισθητά χαμηλότερη τιμή για την παράμετρο μ από τις άλλες παρατηρήσεις της ίδιας παρτίδας. Επιπρόσθετα, ολόκληρη η τέταρτη παρτίδα φαίνεται πως εμφανίζει μικρότερη διαδπορά σε σχέση με τις υπόλοιπες παρτίδες. Έτσι λοιπόν, στην παρτίδα I, ο έλεγχος για την ύπαρξη άτυπης τιμής για την παράμετρο μ (με $k=10, n_i=1$ και $N(i)=9$), δίνει μία ακραία τιμή ίση με 5.55 για την ελεγχοσυνάρτηση, που αντιστοιχεί στην ιδιαίτερα χαμηλή τρίτη μέτρηση. Σε συνδυασμό με την κατανομή t_9, η παρατήρηση αυτή μπορεί να θεωρηθεί άτυπη τιμή σε 1% επίπεδο σημαντικότητας. Ελέγχοντας στη συνέχεια για διαφορές μεταξύ των παρτίδων, ο έλεγχος για μία άτυπη τιμή για την παράμετρο λ χωρίς να υποθέτουμε ισότητα για την παράμετρο μ μεταξύ των παρτίδων (με $k=5, n_i=10$ και $N(i)=40$), δίνει μία ακραία τιμή ίση με 0.062 στο χαμηλό άκρο της κατανομής $F_{9,39}$. Η

τιμή αυτή, που αντιστοιχεί στην τέταρτη παρτίδα, είναι επίσης σημαντική σε 1% επίπεδο σημαντικότητας με αποτέλεσμα η παρτίδα αυτή να θεωρείται άτυπη.



Σχήμα 2.17: Box-plot γραφήματα για τις παρτίδες I, II, III, IV και V των μονωτικών υλικών.

Κεφάλαιο 3

Παλινδρόμηση μοντέλων χρόνου πρώτης μετάβασης

Στο παρόν κεφάλαιο, περιγράφονται γενικές ιδιότητες και χαρακτηριστικά του μοντέλου της παλινδρόμησης Κατωφλιού. Παρουσιάζεται η περίπτωση ενός FHTR μοντέλου όπου η στοχαστική ανέλιξη είναι τύπου Wiener. Το μοντέλο του Cox συγκρίνεται με το μοντέλο της παλινδρόμησης Κατωφλιού. Αναπτύσσονται διαγνωστικές τεχνικές για την καταλληλότητα του μοντέλου και γίνεται διερεύνηση πρακτικών θεμάτων που προκύπτουν κατά την προσαρμογή ενός IG FHTR μοντέλου. Τέλος, αναπτύσσεται μία τεχνική επιλογής μεταβλητών (variable selection).

3.1 IG ως μοντέλο παλινδρόμησης Κατωφλιού (Threshold Regression)

Όπως είδαμε στα δύο προηγούμενα κεφάλαια και ιδιαίτερα στις Παραγράφους 1.2 -1.4 , υπάρχει μεγάλη ποικιλία μοντέλων παλινδρόμησης τα οποία είναι κατάλληλα για την ανάλυση δεδομένων διάρκειας ζωής σε συνδυασμό με μετρήσιμες μεταβλητές. Τα πιο συνηθισμένα μοντέλα στη Μηχανική και τις Θετικές Επιστήμες είναι παραμετρικά, στα οποία οι παράμετροι της κατανομής της διάρκειας ζωής μιας μονάδας (για παράδειγμα η κατανομή Weibull) εξαρτώνται από συμμεταβλητές.

Στο πρώτο κεφάλαιο της διατριβής, δόθηκε αναλυτική περιγραφή των FHT μοντέλων, τα οποία χρησιμοποιούνται τα τελευταία χρόνια στην Ανάλυση Επιβίωσης (Whitmore, 1986;

Lee και Whitmore, 2006). Στη συνέχεια της παραγράφου, θα εισάγουμε δομές παλινδρόμησης για τα FHT μοντέλα και θα ασχοληθούμε με ένα συγκεκριμένο μοντέλο χρόνου πρώτης μετάβασης, στο οποίο η υποβόσκουσα στοχαστική ανέλιξη είναι τύπου Wiener.

3.1.1 Μοντέλα παλινδρόμησης χρόνου πρώτης μετάβασης (FHTR models)

Στα FHT μοντέλα της Παραγράφου 1.5 , γίνεται η υπόθεση πως η διάρκεια ζωής της μονάδας καθορίζεται από μία υποβόσκουσα στοχαστική ανέλιξη, η οποία αναπαριστά την κατάσταση της πορείας της υγείας της μονάδας. Η μονάδα φτάνει σε ένα σημείο κλινικού τερματισμού, όταν η στοχαστική ανέλιξη φτάσει σε μια κρίσιμη κατάσταση για πρώτη φορά. Η κλίμακα του χρόνου μπορεί να είναι ημερολογιακή ή και κάποιο άλλο επιχειρησιακό μέτρο προόδου/εξέλιξης κάποιας διαδικασίας. Ένα απλό παράδειγμα, είναι ένα μοντέλο βασισμένο σε ανέλιξη Wiener (Cox και Miller, 1965), στο οποίο η υποβόσκουσα κατάσταση $\mathrm{X}(t)$ ακολουθεί ένα τυχαίο περίπατο στο χρόνο t με ανεξάρτητες και ισόνομες προσαυξήσεις $N\left(\mu\delta t,\sigma^2\delta t\right)$ σε μη-επικαλυπτόμενα διαστήματα δt . Η διάρκεια ζωής είναι ο χρόνος μέχρι η ανέλιξη να φτάσει σε μία κατάσταση κατωφλιού, όπως $\mathrm{X}(t)=0$. Ωστόσο, στα περισσότερα παραδείγματα η γονική στοχαστική ανέλιξη είναι λανθάνουσα, όπως όταν αναπαριστά την πορεία της υγείας ενός ασθενή.

Η **παλινδρόμηση Κατωφλιού** (Threshold Regression) αναφέρεται σε μοντέλα χρόνων πρώτης μετάβασης **με δομές παλινδρόμησης**, οι οποίες καθιστούν εύκολη τη σύνδεση της διάρκειας ζωής με δεδομένα διαφόρων μεταβλητών. Οι παράμετροι της ανέλιξης, η κατάσταση κατωφλιού και η κλίμακα του χρόνου μπορεί να εξαρτώνται από μεταβλητές.

3.1.2 Παρουσίαση του μοντέλου IG FHTR

Για τα επόμενα, θεωρούμε την πιο συνηθισμένη περίπτωση που συναντάται στη βιβλιογραφία, όπου η λανθάνουσα στοχαστική ανέλιξη για την κατάσταση της υγείας μίας μονάδας σε μια κλίμακα χρόνου t είναι τύπου Wiener, $\{\mathrm{X}(t), t\geq 0\}$ με παραμέτρους m για τη μέση τιμή, σ^2 για τη διασπορά της ανέλιξης και αρχική τιμή $\mathrm{X}(0)=x_0>0$. Στο δεύτερο κεφάλαιο, είδαμε πως ο χρόνος πρώτης μετάβασης T της ανέλιξης σε κάποιο σύνολο συνόρου ακολουθεί την αντίστροφη Γκαουσιανή κατανομή με συνάρτηση πυκνότητας πιθανότητας:

$$f(t|m,\sigma^2,x_0)=\frac{x_0}{\sqrt{2\pi\sigma^2 t^3}}\exp\left[-\frac{(x_0+mt)^2}{2\sigma^2 t}\right],\ t>0, \tag{3.1}$$

$$-\infty<m<+\infty,\ \sigma^2>0,\ x_0>0$$

Η αντίστοιχη συνάρτηση κατανομής δίνεται από τον τύπο:

$$F(t|\, m, \sigma^2, x_0) = \Phi\left[-\frac{(x_0 + mt)}{\sqrt{\sigma^2 t}}\right] + \exp(-2x_0 m / \sigma^2)\Phi\left[\frac{mt - x_0}{\sqrt{\sigma^2 t}}\right], \tag{3.2}$$

όπου με $\Phi(\cdot)$ η συνάρτηση κατανομής της τυποποιημένης Κανονικής κατανομής (Seshadri, 1993). Στην πραγματικότητα, η τιμή σ^2 μπορεί να θεωρηθεί ίση με τη μονάδα, καθώς η κλίμακα της λανθάνουσας ανέλιξης είναι αυθαίρετη.

Στην παλινδρόμηση Κατωφλιού, τόσο το αρχικό σημείο εκκίνησης της ανέλιξης x_0 όσο και η παράμετρος της κλίσης *m* μπορούν να συνδεθούν με τις *k* μεταβλητές του εκάστοτε προβλήματος που αναπαριστώνται με το διάνυσμα $\boldsymbol{z} = (1, z_1, ..., z_k)$, με τη βοήθεια κατάλληλων γραμμικών συναρτήσεων σύνδεσης, ως εξής:

$$\begin{aligned} \ln x_0 &= \gamma_0 + \gamma_1 z_1 + ... + \gamma_k z_k = \boldsymbol{\gamma}'\boldsymbol{z} \\ m &= \beta_0 + \beta_1 z_1 + ... + \beta_k z_k = \boldsymbol{\beta}'\boldsymbol{z} \end{aligned} \tag{3.3}$$

(Lee και Whitmore, 2006). Η μονάδα του διανύσματος $\boldsymbol{z}$ επιτρέπει την ύπαρξη σταθερού όρου στη δομή παλινδρόμησης που παρουσιάσαμε. Η πρώτη σχέση του τύπου (3.3) χρησιμοποιείται για να συνδέσει το αρχικό σημείο εκκίνησης της στοχαστικής ανέλιξης με τις διάφορες μεταβλητές του προβλήματος μέσω μιας λογαριθμικής συνάρτησης, ενώ η δεύτερη χρησιμοποιείται για να συνδέσει την παράμετρο κλίσης της ανέλιξης με τις μεταβλητές του προβλήματος. Τα διανύσματα $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, .., \gamma_k)'$ και $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)'$ είναι τα διανύσματα με τους συντελεστές της παλινδρόμησης, όπου γ_0 και β_0 σταθεροί όροι. Να παρατηρήσουμε πως θέτοντας συγκεκριμένα στοιχεία των διανυσμάτων $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ ίσα με το μηδέν, οι μεταβλητές δρουν μόνο σε μία ή και στις δύο παραμέτρους.

Η αναμενόμενη διάρκεια ζωής δίνεται από τη σχέση:

$$E(T) = \frac{x_0}{|m|} \tag{3.4}$$

Σε περίπτωση που η μέση τιμή *m* της ανέλιξης είναι θετική, η ανέλιξη φτάνει στο σύνορο με πιθανότητα:

$$p_0 = \exp\left(-2x_0 m / \sigma^2\right) < 1,$$

ενώ υπάρχει και θετική πιθανότητα να μη φτάσει ποτέ στο σύνορο:

$$p_1 = 1 - p_0 = 1 - \exp\left(-2x_0 m / \sigma^2\right) \tag{3.5}$$

Σε μία τέτοια περίπτωση, μπορεί να υπάρχουν κάποιες μονάδες με άπειρη "διάρκεια ζωής", με αποτέλεσμα να μη συμβεί ποτέ για αυτές τις μονάδες η πρώτη μετάβαση σε πεπερασμένο

χρόνο. Ένα τέτοιο αποτέλεσμα μπορεί να μας προμηθεύσει με ένα μοντέλο για παρατηρήσεις που θεραπεύονται με το πέρασμα του χρόνου, ή για την περίπτωση που έχουμε μακροχρόνιους επιζώντες (Whitmore, 1979; Balka et al., 2009).

Το γεγονός ότι $P(T=\infty)=1-\exp\left(-2x_0 m/\sigma^2\right)>0$, σχετίζεται σε κάποια FHT μοντέλα με το φαινόμενο των ανταγωνιζόμενων κινδύνων. Γενικά, εάν το μοντέλο λάβει υπόψη όλους τους ανταγωνιζόμενους κίνδυνους που απειλούν την επιβίωση της μονάδας, τότε κάποια στιγμή η ανέλιξη θα φτάσει οπωσδήποτε στο κατώφλι. Στην περίπτωση όμως που το μοντέλο λάβει υπόψη μόνο έναν ή κάποιους από αυτούς, υπάρχει θετική πιθανότητα ο χρόνος πρώτης μετάβασης να είναι άπειρος, ώστε να συμπεριλάβει και όλες εκείνες τις μονάδες που δεν υπόκεινται τελικά σε κάποιον από τους περιορισμένους κινδύνους που έχει λάβει υπόψη το μοντέλο. Επιτρέποντας $m>0$ (αντίστοιχα $\mu<0$ για την περίπτωση που έχουμε την παραμέτρηση (μ,λ) - βλέπε εξίσωση (2.4)), δημιουργείται μία "ελαττωματική" κατανομή, όπως πρώτος σχολίασε ο Whitmore (1979). Όταν $m>0$, η σχέση (3.4) ισχύει υπό συνθήκη, δηλαδή εάν με D_s συμβολίσουμε την κατάσταση κατωφλιού:

$$E(T|D_s)=\frac{x_0}{|m|} \tag{3.6}$$

Δηλαδή, $E(T|D_s)$ είναι η αναμενόμενη διάρκεια ζωής για τις μονάδες των οποίων η ανέλιξη φτάνει στο σύνορο για τη συγκεκριμένη υπό μελέτη αιτία αποτυχίας.

Σε περίπτωση που η μέση τιμή *m* είναι αρνητική, η ανέλιξη κατευθύνεται προς το μηδέν με πιθανότητα τη μονάδα. Η ανέλιξη αυτή φαίνεται να αποτελεί ένα κατάλληλο μοντέλο για την περιγραφή πολλών φυσικών διαδικασιών που παρουσιάζουν τυχαίες αποκλίσεις με την πάροδο του χρόνου. Είναι επίσης ένα αρκετά ρεαλιστικό μοντέλο για να περιγράψει έννοιες όπως η κατάσταση υγείας, καθώς, παρόλο που η υγεία μίας μονάδας τείνει να μειώνεται όσο οι άνθρωποι γερνάνε, ωστόσο, σε μία μικρότερη κλίμακα χρόνου η κατάσταση της υγείας είναι επιρρεπής σε ταλαντώσεις οι οποίες μπορούν να περιγραφούν ιδανικά από τις κινήσεις μίας ανέλιξης Wiener. Για την παλινδρόμηση Κατωφλιού, οι κυριότερες αναφορές σε εφαρμογές έχουν γίνει από τους Whitmore (1979, 1983, 1986, 1995), Doksum και Hoyland (1992), Doksum και Normand (1995), Lu (1995), Whitmore και Schenkelberg (1997), Lee et al (2004) κ.α..

Στην πιο συνηθισμένη παραμέτρηση της IG στη βιβλιογραφία, η συνάρτηση πυκνότητας πιθανότητας μιας τυχαίας μεταβλητής *T*, δίνεται από τη σχέση (2.4) του δεύτερου κεφαλαίου:

$$f(t;\mu,\lambda)=\sqrt{\frac{\lambda}{2\pi t^3}}\exp\left\{-\frac{\lambda(t-\mu)^2}{2\mu^2 t}\right\},\quad t>0,\quad \mu,\ \lambda>0$$

Οι παράμετροι των δύο εκδοχών συμπίπτουν με τη βοήθεια της σχέσης (2.70):

$$\lambda = x_0^2,\ \mu = \frac{x_0}{|m|}$$

Τέλος, να παρατηρήσουμε πως ο περιορισμός $\mu > 0$ της (2.4) αντιστοιχεί στον περιορισμό $m < 0$ της (3.1).

3.1.3 Μελέτη της προσαρμογής ενός FHTR μοντέλου

Στη συνέχεια του κεφαλαίου, θα παρουσιαστούν διάφορα θεωρητικά αποτελέσματα σχετικά με το μοντέλο παλινδρόμησης IG FHTR. Προς υποστήριξη της εγκυρότητας και ορθότητας των διαφόρων θεωρητικών αποτελεσμάτων, έχουν γίνει πολλές προσομοιώσεις, τα αποτελέσματα των οποίων παρουσιάζονται κατά μήκος του κεφαλαίου μέσα από πληθώρα πινάκων, διαγραμμάτων και γραφημάτων.

Στην αρχή κάθε μελέτης, δημιουργούνται ψευδοτυχαίες τιμές από διάφορες γνωστές κατανομές (Κανονική, Weibull, Ομοιόμορφη, Εκθετική, πολυμεταβλητή Κανονική, κ.τ.λ.). Στη συνέχεια, κατασκευάζονται οι παράμετροι του αληθινού FHT μοντέλου παλινδρόμησης στα ίδια δεδομένα, με τη βοήθεια της σχέσης (3.3) και τη χρήση των προσομοιωμένων τιμών των διαφόρων μεταβλητών που συμμετέχουν στην εκάστοτε μελέτη. Το πλήθος των μεταβλητών και οι τιμές των συντελεστών στις γραμμικές εκτιμήτριες ποικίλουν ανάλογα το είδος της μελέτης. Οι μέσες τιμές επιλέγονται με τέτοιο τρόπο, ώστε $m < 0$ για την πλειοψηφία των μονάδων. Για την περίπτωση που κάποια μονάδα έχει $m > 0$, με αποτέλεσμα η απορρόφηση από το κατώφλι να μην είναι εξασφαλισμένη, μία ψευδοτυχαία τιμή δημιουργείται από την Ομοιόμορφη κατανομή στο διάστημα $[0,1]$. Στην περίπτωση που η τιμή αυτή είναι μικρότερη από την τιμή $\exp(-2x_0 m)$, δημιουργούμε μία νέα διάρκεια ζωής για τη μονάδα. Στην αντίθετη περίπτωση, θεωρούμε τη διάρκεια ζωής άπειρη (στην πραγματικότητα της δίνουμε μία μεγάλη δεξιά αποκομμένη τιμή). Στη συνέχεια, οι τιμές των μεταβλητών που δημιουργήθηκαν για την κάθε μονάδα, χρησιμοποιούνται για να κατασκευάσουν τις αντίστοιχες τιμές των IG παραμέτρων, με τη χρήση των εξισώσεων της σχέσης (2.70).

Όλοι οι υπολογισμοί γίνονται με τη βοήθεια του υπολογιστικού στατιστικού πακέτου R. Σε κάθε προσομοιωμένο δείγμα, οι διάρκειες ζωής των μονάδων δημιουργούνται ως ψευδοτυχαίες τιμές από την κατανομή IG με τις τιμές των παραμέτρων μ και λ που προκύπτουν από το αληθινό μοντέλο χρησιμοποιώντας τη ρουτίνα *rinvgauss* της βιβλιοθήκης *statmod* (Smyth, 2012). Η εν λόγω ρουτίνα χρησιμοποιεί την τεχνική που πρότειναν οι Michael et al. (1976) και παρουσιάσαμε στο δεύτερο κεφάλαιο της διατριβής για την παραγωγή τυχαίων παρατηρήσεων από την IG χρησιμοποιώντας έναν

μετασχηματισμό με πολλαπλές ρίζες. Τέλος, το FHT μοντέλο παλινδρόμησης προσαρμόζεται με ελαχιστοποίηση της συνάρτησης του αρνητικού λογαρίθμου της πιθανοφάνειας χρησιμοποιώντας τη ρουτίνα *optim*.

3.2 Σύγκριση των μοντέλων FHT και Cox

Σε αυτήν την παράγραφο, συγκρίνουμε το μοντέλο του Cox και ένα μοντέλο FHT παλινδρόμησης βασισμένο σε ανέλιξη Wiener, το οποίο οδηγεί σε χρόνο πρώτης μετάβασης που ακολουθεί την IG κατανομή. Παρουσιάζεται μία εφαρμογή σε ασθενείς που έχουν υποστεί επέμβαση ανοιχτής καρδιάς της στεφανιαίας αρτηρίας, προκειμένου να εξεταστεί η ερμηνεία του μοντέλου, ειδικά στην περίπτωση που κάποια μεταβλητή επιδρά και στις δύο παραμέτρους.

3.2.1 Εισαγωγή

Η συντριπτική πλειοψηφία των εφαρμογών στη Βιοστατιστική χρησιμοποιεί το μοντέλο του Cox. Επεκτάσεις του μοντέλου του Cox, ειδικά η στρωματοποιημένη έκδοση καθώς και η εισαγωγή χρονικά εξαρτώμενων μεταβλητών, αυξάνουν την ευελιξία σε σημαντικό βαθμό, ωστόσο το βασικό σημείο παραμένει η υπόθεση της αναλογικής διακινδύνευσης (Therneau και Grambsch, 2000; Collett, 2003). Όπως είδαμε και στο πρώτο κεφάλαιο, το μοντέλο του Cox καθορίζει μόνο τον τρόπο με τον οποίο το διάνυσμα των συμμεταβλητών $\boldsymbol{x}$ δρα σε μία βασική συνάρτηση διακινδύνευσης $h_0(t)$ μέσω της σχέσης:

$$h_1(t) = h_0(t)\mathrm{e}^{\boldsymbol{\beta}'\boldsymbol{x}} \tag{3.7}$$

με $\boldsymbol{\beta}$ το διάνυσμα των συντελεστών των υπό εκτίμηση μεταβλητών του μοντέλου. Η βασική συνάρτηση διακινδύνευσης $h_0(t)$ δεν καθορίζεται, χωρίς ωστόσο να είναι απόλυτα ελεύθερη να πάρει οποιαδήποτε μορφή. Η υπόθεση της αναλογικής διακινδύνευσης (PH assumption) συνεπάγεται ένα περιορισμό ως προς τις επιτρεπόμενες μορφές της συνάρτησης διακινδύνευσης, παρόλο που δε μοντελοποιείται απευθείας. Στην πραγματικότητα, η βασική συνάρτηση διακινδύνευσης πρέπει να είναι μία μονότονη συνάρτηση: αύξουσα (που αποτελεί και την πιο συνηθισμένη περίπτωση που συναντάται στη βιβλιογραφία), φθίνουσα ή σταθερή. Μία μη-μονότονη συνάρτηση διακινδύνευσης δεν ικανοποιεί την υπόθεση PH. Οι μόνες γνωστές κατανομές διάρκειας ζωής που ικανοποιούν την παραπάνω απαίτηση είναι οι Weibull (συμπεριλαμβανομένης και της Εκθετικής σαν ειδική περίπτωση), Gompertz και η Gamma, με αποτέλεσμα η παραμετρική PH παλινδρόμηση να είναι περιορισμένη στη χρησιμοποίηση τους. Συγκεκριμένα, η ανάγκη για αυστηρή μονοτονία αποκλείει τη δυνατότητα εφαρμογής του παραμετρικού ή μη-παραμετρικού PH μοντέλου, όταν η εμπειρική συνάρτηση διακινδύνευσης αυξάνει μέχρι ένα μέγιστο σημείο και μετά φθίνει

(γνωστό στη βιβλιογραφία της Αξιοπιστίας ως το "αυξανόμενο-και-μετά-μειούμενο ρυθμό αποτυχίας" - IDFR), ή όταν έχει το σχήμα "μπανιέρας" (bath tub shape) κατά το οποίο μία αρχική μείωση ακολουθείται από μία περίοδο σταθερής διακινδύνευσης και στη συνέχεια από μία αύξηση. Τα παραπάνω σχήματα είναι αρκετά συνηθισμένα σε κάποιες περιοχές εφαρμογών. Για παράδειγμα, οι Gore et al. (1984), περιγράφουν περιπτώσεις όπου οι ρυθμοί των θανάτων αυξάνουν τα πρώτα χρόνια της περιόδου παρακολούθησης, αλλά στη συνέχεια μειώνονται και συγκλίνουν μεταξύ των ομάδων. Το AL μοντέλο, αντιθέτως, δεν έχει περιορισμούς ως προς τη μονοτονία της συνάρτησης διακινδύνευσης. Για παράδειγμα, στα AL μοντέλα χρησιμοποιείται συνήθως η log-normal κατανομή των χρόνων ζωής και η συνάρτηση διακινδύνευσης έχει πάντοτε μέγιστη τιμή. Επίσης, το AL μοντέλο επιτρέπει στα γραφήματα των συναρτήσεων διακινδύνευσης να διασταυρωθούν, κάτι που δεν είναι δυνατό κάτω από την PH υπόθεση (Zhang και Peng, 2009).

Ακόμα και στις περιπτώσεις εκείνες που η διατύπωση της PH υπόθεσης είναι σωστή για το πραγματικό, άγνωστο μοντέλο του πληθυσμού, δεν ισχύει απαραίτητα για το προσαρμοσμένο μοντέλο ή ακόμα και για τα διαθέσιμα δεδομένα. Η PH ιδιότητα παύει να ισχύει στην περίπτωση που το μοντέλο είναι λανθασμένα προσδιορισμένο λόγω παράλειψης κάποιας μεταβλητής που θα έπρεπε να έχει συμπεριληφθεί. Οι Hutton και Monaghan (2002) αποδεικνύουν πως το AL είναι πιο εύρωστο/ανθεκτικό από το PH σε ενδεχόμενο λανθασμένο προσδιορισμό. Επιπρόσθετα, η PH ιδιότητα χάνεται, εάν τα δεδομένα έχουν αποκτηθεί μέσω μεροληπτικής δειγματοληψίας (Economou και Caroni, 2009).

Δεδομένων των παραπάνω περιορισμών, είναι πιθανό οι αμέτρητες εφαρμογές του μοντέλου του Cox να περιέχουν αρκετές για τις οποίες δεν αποτελούσε την κατάλληλη επιλογή. Ο ίδιος ο Cox ισχυρίζεται πως η εξίσωση (3.7) "προορίζεται για την αναπαράσταση της συμπεριφοράς των χρόνων βλάβης ως βολικό, ευέλικτο και εν τούτοις ολοκληρωτικά εμπειρικό" (Cox, 1972, p.200) και ολοκλήρωσε την εργασία του με τη δήλωση πως το μοντέλο "ως βάση για περαιτέρω εμπειρική μείωση των δεδομένων… δείχνει ευέλικτη και ικανοποιητική" (Cox, 1972, p.201). Ο Freedman (2008, p.117) εύλογα αντιτίθεται σε αυτόν τον ισχυρισμό ρωτώντας: "αν το μοντέλο είναι λάθος, γιατί οι εκτιμήτριες των παραμέτρων να αποτελούν μία καλή περίληψη των δεδομένων?"

3.2.2 Σύγκριση FHT μοντέλων με τα μοντέλα αναλογικής διακινδύνευσης

Σε αντίθεση με τη μοντελοποίηση της συνάρτησης διακινδύνευσης, η ιδέα της χρησιμοποίησης ενός FHT μοντέλου της λανθάνουσας ανέλιξης, που περιγράφει το χρόνο μέχρι το συμβάν, μπορεί να οδηγήσει σε διάφορες κατανομές για τη διάρκεια ζωής, ανάλογα με τη φύση της υποβόσκουσας ανέλιξης. Ωστόσο, η πιο συνηθισμένη περίπτωση που συναντάται στη βιβλιογραφία αφορά ένα μοντέλο στο οποίο η ανέλιξη είναι τύπου Wiener,

στο οποίο ο χρόνος πρώτης μετάβασης ακολουθεί την IG. Οι Fries και Bhattacharyya (1983) δικαιολογούν το ενδιαφέρον της IG ως κατανομής της διάρκειας ζωής λόγω:

α) της ευελιξίας του και επειδή παίρνει ποικιλία μορφών.

β) των βολικών ιδιοτήτων για τις κατανομές δειγματοληψίας της, οι οποίες συχνά είναι ανάλογες με την Κανονική κατανομή (Tweedie, 1957a).

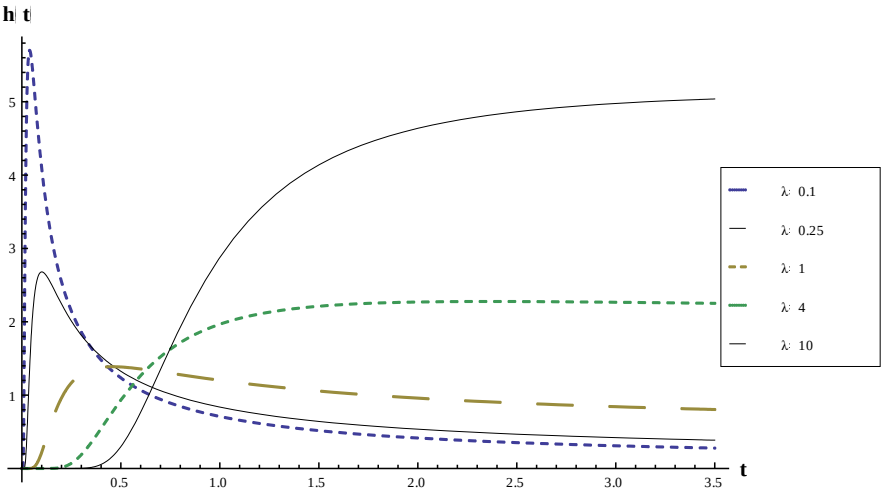
γ) του γεγονότος πως "η παραγωγή του από μια εύλογη στοχαστική μοντελοποίηση της ανέλιξης της βλάβης συχνά προσφέρει μία φυσική υποστήριξη της εμπειρικής του προσαρμογής".

Ωστόσο, το πρώτο από τα παραπάνω σχόλια απαιτεί κάποιες περαιτέρω διευκρινήσεις. Αναλυτική περιγραφή της συνάρτησης πυκνότητας πιθανότητας για την IG κατανομή δόθηκε στην Παράγραφο 2.2 . Όπως είδαμε, η πιο χρησιμοποιούμενη μορφή στη βιβλιογραφία, δίνεται από τη σχέση (2.4). Η αντίστοιχη συνάρτηση επιβίωσης $S(t;\mu,\lambda)=P(T>t)$, είναι:

$$S(t;\mu,\lambda)=\Phi\left[\sqrt{\frac{\lambda}{t}}\left(1-\frac{t}{\mu}\right)\right]-e^{2\lambda/\mu}\,\Phi\left[-\sqrt{\frac{\lambda}{t}}\left(1+\frac{t}{\mu}\right)\right],\ t>0,\ \mu,\ \lambda>0 \qquad (3.8)$$

Η μορφή της συνάρτησης διακινδύνευσης $h(t)=f(t)/S(t)$ της IG παρουσιάστηκε λεπτομερώς από τους Chhikara και Folks (1977), οι οποίοι έδειξαν ότι έχει πάντοτε τη μορφή IDFR, για κάθε λ και κάθε μ. Το γράφημα της $h(t)$ για διάφορες τιμές των μ και λ δίνεται στο Σχήμα 3.1. Αποδεικνύεται πως η κορυφή για την $h(t)$ επιτυγχάνεται στο σημείο t^* επιλύοντας την εξίσωση:

$$h(t)=\frac{3}{2t}+\frac{\lambda}{2t^2}-\frac{\lambda}{2\mu^2}$$



Σχήμα 3.1: Συνάρτηση διακινδύνευσης για την κατανομή IG για διάφορες τιμές του λ, με μ = 1.

Ο Πίνακας 3.1 δίνει τις διάφορες τιμές του t^* για διάφορες επιλογές της παραμέτρου λ (για $\mu = 1$). Επίσης, δίνει τις αντίστοιχες τιμές της $S(t^*)$. Επειδή η τιμή της παραμέτρου μ είναι ίση με 1, η αναμενόμενη διάρκεια ζωής είναι και αυτή ίση με τη μονάδα. Φαίνεται εύκολα πως για μικρές τιμές της ποσότητας λ/μ, το μέγιστο της συνάρτησης διακινδύνευσης επιτυγχάνεται γρήγορα με αποτέλεσμα ένα μεγάλο ποσοστό μονάδων (σχεδόν 0.9) να επιζεί μετά τη χρονική στιγμή t^*. Συνεπώς, κάποιες πρόωρες αποτυχίες υπάρχουν, αλλά μετά από αυτήν τη σχετικά μικρή χρονική περίοδο η συνάρτηση διακινδύνευσης φθίνει. Αντιθέτως, για μεγάλες τιμές της ποσότητας λ/μ, η πλειοψηφία των μονάδων έχει ήδη φτάσει στην αποτυχία μέχρι η συνάρτηση διακινδύνευσης να φτάσει στο μέγιστό της. Συνεπώς, η συγκεκριμένη κατάσταση περιγράφει έναν αυξανόμενο ρυθμό διακινδύνευσης αρκετά αποτελεσματικά. Έτσι λοιπόν, η συνάρτηση διακινδύνευσης της IG, παρόλο που τεχνικά έχει πάντοτε μία μέγιστη τιμή και είναι μονοκόρυφη, μπορεί να συμπεριφερθεί είτε σαν IFR ή DFR μοντέλο. Για το λόγο αυτό θεωρείται ιδιαίτερα ευέλικτη.

λ	t^*	$S(t^*)$	$h(t^*)$	$h(\infty)$	$h(t^*)/h(\infty)$
0.10	0.039	0.879	5.70	0.05	114
0.25	0.101	0.854	2.68	0.125	21.4
1	0.457	0.674	1.39	0.5	2.78
2	1.035	0.353	1.52	1	1.52
4	2.363	0.020	2.28	2	1.14
10	3.605	3×10^{-6}	5.05	5	1.01

Πίνακας 3.1: Χρονική στιγμή t^* κατά την οποία επιτυγχάνεται η κορυφή της συνάρτησης διακινδύνευσης της IG, τιμές των συναρτήσεων διακινδύνευσης και επιβίωσης κατά τη χρονική αυτή στιγμή και ασυμπτωτικές τιμές της συνάρτησης διακινδύνευσης

Η συνάρτηση διακινδύνευσης της IG οριακά προσεγγίζει την τιμή $\lambda/2\mu^2$, καθώς $t \to +\infty$ (Chhikara and Folks, 1977). Ας θεωρήσουμε για τη συνάρτηση πυκνότητας πιθανότητας του χρόνου πρώτης μετάβασης την εναλλακτική παραμέτρηση της παλινδρόμησης Κατωφλιού της σχέσης (3.1):

$$f(t|m,\sigma^2,x_0) = \frac{x_0}{\sqrt{2\pi\sigma^2 t^3}} \exp\left[-\frac{(x_0+mt)^2}{2\sigma^2 t}\right],\ t>0, \qquad (3.9)$$

$$-\infty < m < +\infty,\ \sigma^2 > 0,\ x_0 > 0$$

Σε αντιστοιχία με τα προηγούμενα, το όριο της συνάρτησης διακινδύνευσης είναι $m^2/2$.

Έτσι λοιπόν, προκύπτουν τα εξής:

1. Το όριο της συνάρτησης διακινδύνευσης $h(t)$ είναι ανεξάρτητο του σημείου εκκίνησης x_0 της ανέλιξης.
2. Για μεγάλες τιμές του *t*, ο λόγος $h(t:\boldsymbol{x_1})/h(t:\boldsymbol{x_2})$ δύο μονάδων με διανύσματα συμμεταβλητών $\boldsymbol{x_1}$ και $\boldsymbol{x_2}$ είναι ανεξάρτητος του *t*.

Η πρώτη παρατήρηση μπορεί να ερμηνευτεί λέγοντας πως οι αρχικές διαφορές μεταξύ των μονάδων τείνουν να χάσουν τη σημαντικότητά τους με την πάροδο του χρόνου, σε αντίθεση με την περίπτωση PH, όπου οι αρχικές διαφορές μεταξύ των μονάδων διατηρούνται για πάντα. Αυτό αποτελεί ένα σημαντικό πλεονέκτημα των μοντέλων FHT σε σχέση με τα PH μοντέλα, που μπορεί να τα μετατρέψει σε ένα πιο ρεαλιστικό εργαλείο σε πολλές εφαρμογές. Ωστόσο, η δεύτερη παρατήρηση υποδηλώνει πως για μεγάλες τιμές του *t*, η υπόθεση αναλογικής διακινδύνευσης ισχύει (με τη σταθερά της αναλογικότητας να εξαρτάται αποκλειστικά από την κλίση της ανέλιξης και όχι από τις αρχικές συνθήκες). Συνεπώς, οι κίνδυνοι για μία συμμεταβλητή που επηρεάζει μόνο το αρχικό σημείο εκκίνησης της ανέλιξης συγκλίνουν με την πάροδο του χρόνου, αλλά οι κίνδυνοι για μία συμμεταβλητή που επηρεάζει την κλίση είναι τελικά ανάλογοι. Επίσης, οι κίνδυνοι στα μοντέλα αναλογικών συμπληρωματικών πιθανοτήτων συγκλίνουν με την πάροδο του χρόνου (Bennett, 1983a).

Οι Lee και Whitmore (2010) μελετούν λεπτομερώς τη σύνδεση μεταξύ των μοντέλων FHT και PH. Αποδεικνύουν πως τα μοντέλα αναλογικής διακινδύνευσης προκύπτουν από τα FHT μοντέλα με την αλλαγή της κλίμακας του χρόνου ή με αλλαγή του συνόρου. Υπό αυτήν την έννοια, θεωρούν τα μοντέλα PH ως μια ειδική περίπτωση των FHT μοντέλων.

3.2.3 Η περίπτωση των θεραπευμένων μονάδων (Cured Fraction)

Στην Παράγραφο 3.1.2, παρουσιάσαμε την περίπτωση των "θεραπευμένων μονάδων" "μακροπρόθεσμα επιζώντων", σύμφωνα με την οποία ένα ποσοστό $1-p_0$ του πληθυσμού δε θα αντιμετωπίσει ποτέ την αποτυχία εάν $m>0$. Ο Whitmore (1979), ήταν ο πρώτος που χρησιμοποίησε την κατανομή IG για την αναπαράσταση αυτού του χαρακτηριστικού των παρατηρήσεων. Μία πιο συνηθισμένη προσέγγιση είναι η χρησιμοποίηση μοντέλων μίξης κατανομών (mixture models) της μορφής:

$$S(t)=(1-\pi)+\pi S_0(t), \quad 0<\pi<1 \tag{3.10}$$

για την περιγραφή πληθυσμών που αποτελούνται από ένα ποσοστό θεραπευμένων μονάδων 1-π και ένα ποσοστό μονάδων π που αντιμετωπίζουν την αποτυχία σύμφωνα με μία συνάρτηση επιβίωσης $S_0(t)$. Τα μοντέλα αυτά εισήχθησαν από τον Boag (1949). Όταν η κατανομή IG προκύπτει από μία υποβόσκουσα ανέλιξη Wiener, αναπαριστά την περίπτωση

των μακροπρόθεσμα επιζώντων λίγο διαφορετικά από το μοντέλο της σχέσης (3.10). Οι μακροπρόθεσμα επιζώντες της σχέσης (3.10) ανήκουν σε αυτήν την ομάδα από την αρχή. Για παράδειγμα, έχουν θεραπευτεί από μία ασθένεια με τη βοήθεια θεραπείας, οπότε δεν είναι πλέον υποψήφιοι να πεθάνουν από την ασθένεια αυτή. Αυτός ο χωρισμός των δύο ομάδων μπορεί να είναι ρεαλιστικός ή όχι, ανάλογα με την εφαρμογή. Αντιθέτως, στο FHT μοντέλο η μακροπρόθεσμη επιβίωση προκύπτει σαν αποτέλεσμα της τυχαίας πορείας της ανέλιξης. Η δυναμική αυτή αναπαράσταση της ανέλιξης από το FHT μοντέλο είναι ιδιαίτερα ελκυστική (Singpurwalla, 1995). Για παράδειγμα, είναι πιο εύλογο να θεωρήσουμε πως η ενδεχόμενη επιστροφή στη φυλακή ενός παραβάτη ο οποίος έχει εκτίσει την ποινή του (όπως στην ανάλυση των Maller και Zhou (1994)) είναι αποτέλεσμα της περιόδου μετά την αποφυλάκισή του, παρά ότι αποτελεί ένα σταθερό και προαποφασισμένο ενδεχόμενο τη στιγμή της αποφυλάκισης. Μοντέλα θεραπείας βασισμένα στην ανέλιξη Wiener παρουσιάζονται λεπτομερώς στους Balka et al. (2009), με πολλές επεκτάσεις συμπεριλαμβανομένων και των μοντέλων μίξης.

3.2.4 Εφαρμογή του FHT μοντέλου σε δεδομένα PH

Στη συνέχεια, εξετάζουμε τι πρόκειται να συμβεί στην περίπτωση που το IG μοντέλο προσαρμοστεί σε δεδομένα τα οποία ικανοποιούν την υπόθεση PH της αναλογικής διακινδύνευσης και αντιστρόφως. Τα παρακάτω αποτελέσματα προέκυψαν με τη δημιουργία δεδομένων υπό το ένα μοντέλο και εν συνεχεία με την προσαρμογή τους στο άλλο. Σε αυτήν την πρώτη περίπτωση, PH δεδομένα ($n=100$) δημιουργήθηκαν από την κατανομή Weibull με παράμετρο σχήματος ίση με ένα και παράμετρο κλίμακας $\exp(\boldsymbol{\beta}'\boldsymbol{z})$, όπου οι ανεξάρτητες μεταβλητές στο διάνυσμα $\boldsymbol{z}$ δημιουργήθηκαν ως εξής:

1. Η z_1 αποτελείται από 50 μηδενικά και 50 μονάδες.
2. Η z_2 ακολουθεί την κατανομή Bernoulli με παράμετρο 0.5.
3. Η z_3 ακολουθεί την Ομοιόμορφη κατανομή στο διάστημα $U[0,1]$.
4. Τέλος, η $z_4 \sim N(2,1)$.

Το διάνυσμα των συντελεστών είναι $\boldsymbol{\beta}=(2,-1,2,-1)$. Το φαινόμενο της αποκοπής ελήφθη υπ' όψιν στη μελέτη με τη δημιουργία μίας ανεξάρτητης στιγμής αποκοπής για κάθε μία μονάδα από την Ομοιόμορφη κατανομή. Με τον τρόπο αυτό, επετεύχθη ένα ποσοστό 33% αποκομμένων παρατηρήσεων. Κατά την προσαρμογή του IG FHT μοντέλου, όλες οι μεταβλητές επιτρεπόταν να εισαχθούν στις γραμμικές εκτιμήτριες και των δύο παραμέτρων x_0 και m. Στατιστικά σημαντικοί συντελεστές σε ε.σ. 1% προέκυψαν για τις μεταβλητές z_1, z_3 και z_4 στην παράμετρο x_0 και για τις μεταβλητές z_2 και z_4 για την παράμετρο m. Συνεπώς, με αυτό το παράδειγμα, η ανάλυση αναγνώρισε μεταβλητές οι οποίες ήταν

σημαντικές μόνο για την παράμετρο x_0 (x_1 και z_3), μόνο για την παράμετρο m (z_2) και μεταβλητές σημαντικές και για τις δύο παραμέτρους ταυτόχρονα (z_4, με συντελεστές που βρίσκονται σε συμφωνία ανάμεσα στις δύο παραμέτρους). Το αποτέλεσμα αυτό συνιστά ένα μεγάλο εύρος πιθανών αποτελεσμάτων.

Τη δεύτερη φορά, δεδομένα από την κατανομή IG δημιουργήθηκαν για τον πίνακα των συμμεταβλητών και για το επίπεδο της αποκοπής όπως και πριν. Η γραμμική εκτιμήτρια για την παράμετρο x_0 ήταν $1-0.5z_1-0.5z_3$, ενώ για την παράμετρο m ήταν $1-z_2-z_4$. Όλοι οι παράγοντες έχουν αρνητική επίδραση στη διάρκεια ζωής με αποτέλεσμα να είναι αναμενόμενη η εμφάνιση θετικών συντελεστών σε ενδεχόμενη παλινδρόμηση με το μοντέλο του Cox, καθώς αναμένεται να αυξήσουν τη διακινδύνευση. Ωστόσο, τα αποτελέσματα της προσαρμογής με το μοντέλο του Cox σε αυτά τα δεδομένα παρουσίασαν στατιστικά σημαντικές μόνο τις μεταβλητές z_2 και z_4 σε 5% ε.σ. (παρόλο που το μέγεθος του δείγματος είναι μεγάλο). Επίσης, η z_2 εμφάνισε αρνητικό συντελεστή.

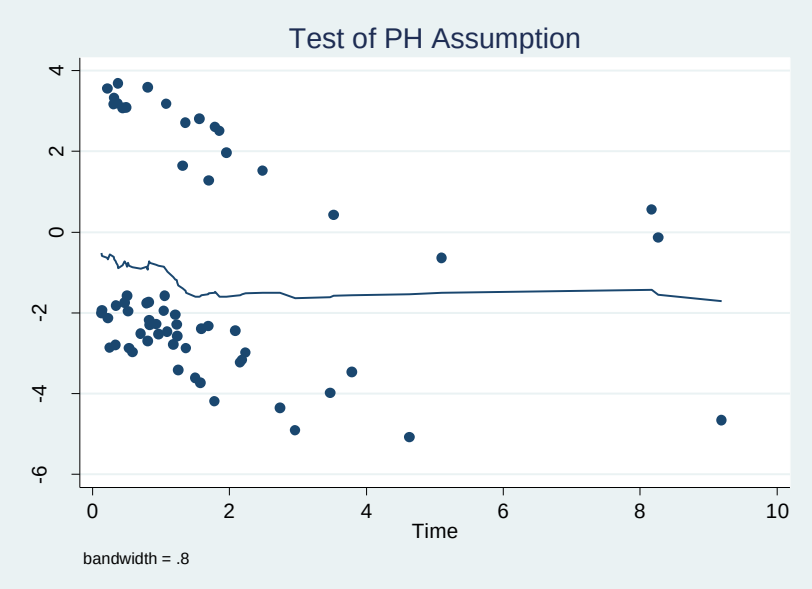
Για το μοντέλο του Cox, έχει αναπτυχθεί ένας μεγάλος αριθμός διαγνωστικών τεχνικών, όπως παρουσιάσαμε στις Παραγράφους 1.2.3 έως 1.2.6 της διατριβής. Αυτές δεν αποδείχθηκαν ιδιαίτερα ωφέλιμες για την απόδειξη πως το μοντέλο δεν ήταν όντως PH.

	ρ	χ^2	Βαθμοί ελευθερίας	P-τιμή
z_1	-0.19440	2.33	1	0.1271
z_2	-0.17988	2.56	1	0.1094
z_3	-0.13132	1.13	1	0.2881
z_4	0.08566	0.37	1	0.5419
Καθολικός έλεγχος		5.73	4	0.2198

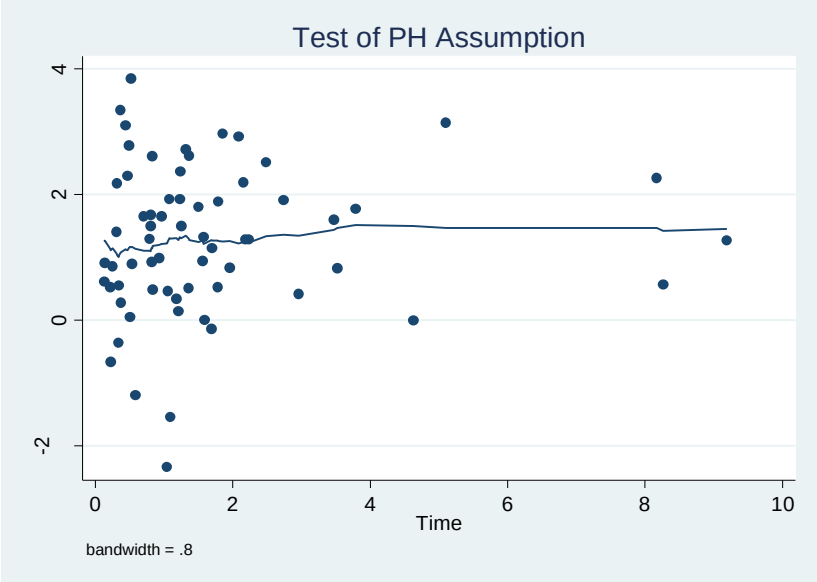
Πίνακας 3.2: Έλεγχος της καταλληλότητας του προσαρμοσμένου μοντέλου μελετώντας την πορεία των εκτιμητριών του μοντέλου με το πέρασμα του χρόνου.

Η στήλη ρ του Πίνακα 3.2 δίνει το γνωστό συντελεστή συσχέτισης του Pearson, ο οποίος ελέγχει εάν ο συντελεστής της κάθε μεταβλητής ξεχωριστά μεταβάλλεται με το πέρασμα του χρόνου.

Οι ελεγχοσυναρτήσεις για τον έλεγχο της PH υπόθεσης (Therneau και Grambsch, 2000), για κάθε μία μεταβλητή ξεχωριστά, αλλά και για το σύνολο των μεταβλητών ήταν ξεκάθαρα στατιστικά μη-σημαντικές, όπως δείχνει και ο Πίνακας 3.2 (p-value > 0.10). Τέλος, τα γραφήματα των Schoenfeld υπολοίπων, τα οποία παρουσιάστηκαν στην Παράγραφο 1.2.5 (Therneau και Grambsch, 2000), μπορεί να ερμηνευτούν ως ενδείξεις αναχώρησης από την PH για μικρές διάρκειες ζωής μόνο για τη z_2, αλλά όχι και για τη z_4, όπως δείχνουν τα Σχήματα 3.2 και 3.3.



Σχήμα 3.2: Γράφημα των Scaled – Schoenfeld υπολοίπων με το χρόνο για τη μεταβλητή z_2 .

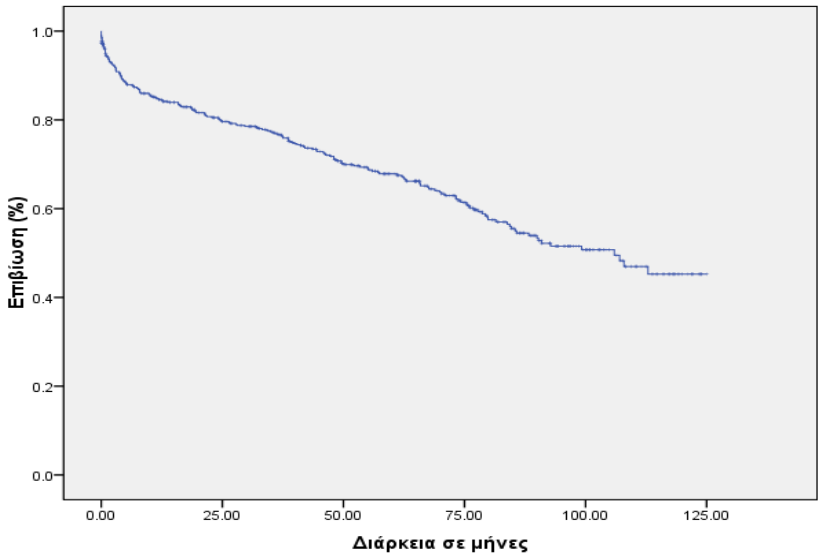


Σχήμα 3.3: Γράφημα των Scaled – Schoenfeld υπολοίπων με το χρόνο για τη μεταβλητή z_4 .

Τα δύο παραπάνω παραδείγματα αποτελούν απλές αναπαραστάσεις ικανές να δείξουν πως τα δύο μοντέλα FHT και PH μπορούν να καταλήξουν σε εντελώς διαφορετικά αποτελέσματα. Μεγάλη ανάγκη υπάρχει για την ανάπτυξη διαγνωστικών τεχνικών για τα FHT μοντέλα.

3.2.6 Εφαρμογή σε πρόβλημα πραγματικών συνθηκών

Μία μελέτη που δημοσιεύτηκε από τους DeRose et al. (2005) περιγράφει την εφαρμογή του μοντέλου παλινδρόμησης του Cox σε δεδομένα που αφορούν τη διάρκεια ζωής 544 ασθενών οι οποίοι είχαν υποστεί επέμβαση ανοιχτής καρδιάς στη στεφανιαία αρτηρία με τη μέθοδο εισαγωγής μοσχεύματος σε κάποιο νοσοκομείο μεταξύ των ετών 1992 και 2002. Κατά τη διάρκεια της παρακολούθησης, 192 ασθενείς (35.3%) πέθαναν. Το κριτήριο για τη συμμετοχή των ασθενών στη μελέτη ήταν η χαμηλή ένδειξη του δείκτη κλάσματος εξώθησης της αριστερής κοιλιακής περιοχής της καρδιάς (το κλάσμα του αίματος που αντλείται από την αριστερή κοιλιακή χώρα σε κάθε καρδιακό παλμό). Για να θεωρηθεί μία τιμή χαμηλή, θα πρέπει να είναι χαμηλότερη από 25% σε σχέση με την τιμή 50% ή παραπάνω των υγιών ανθρώπων. Το Σχήμα 3.4 παρουσιάζει την εκτιμήτρια Kaplan Meier της επιβίωσης για όλο το δείγμα. Σκοπός της μελέτης ήταν η χρησιμοποίηση προ-εγχειρητικών παραγόντων κινδύνου για την κατασκευή ενός προγνωστικού σκορ για την πορεία της επιβίωσης.



Σχήμα 3.4: Γράφημα της εκτιμήτριας Kaplan – Meier της επιβίωσης (σε μήνες) των 544 ασθενών που υπέστησαν επέμβαση επέμβαση ανοιχτής καρδιάς.

Η στρατηγική που ακολουθήθηκε υιοθετείται συχνά σε ιατρικές εφαρμογές επιβίωσης στις οποίες υπάρχει ένα μεγάλο δείγμα διαθέσιμων μεταβλητών. Αρχικά, ο κάθε υποψήφιος παράγοντας κινδύνου ελέγχθηκε ξεχωριστά σε μία μονομεταβλητή ανάλυση με το μοντέλο του Cox. Εκείνοι οι παράγοντες που αποδείχθηκαν στατιστικά σημαντικοί $(p-\text{value} < 0.05)$ παρέμειναν στη μελέτη, όπου εισήχθησαν σε μία πολυμεταβλητή ανάλυση με το μοντέλο του Cox. Αυτές ήταν οι εξής έντεκα μεταβλητές: *ηλικία, δείκτης μάζας σώματος* (BMI), *επέμβαση έκτακτης ανάγκης, επείγουσα επέμβαση, τωρινή καρδιακή ανεπάρκεια* (CHF), *προηγούμενη καρδιακή ανεπάρκεια* (PHF), *περιφερειακή αγγειακή νόσος* (PVD), *χρόνια αποφρακτική πνευμονική ασθένεια* (COPD), *αρτηριοσκλήρωση, ανοσοποιητική δυσλειτουργία και*

χειρουργείο με πάλλουσα καρδιά (OPCAB). Οι μεταβλητές *ηλικία* και *δείκτης μάζας σώματος* ήταν συνεχείς, ενώ οι υπόλοιπες είναι ψευδομεταβλητές με τιμή 1 για "ναι" και 0 για "όχι". Επιπρόσθετα, η τεχνική της Διαδοχικής Αφαίρεσης (Backward Elimination) χρησιμοποιήθηκε στο μοντέλο του Cox, προκειμένου να μειώσει ακόμα περισσότερο τον αριθμό των έντεκα μεταβλητών. Οι προ-εγχειρητικοί παράγοντες κινδύνου που παρέμειναν με αυτή τη διαδικασία στο τελικό μοντέλο παρουσιάζονται στον Πίνακα 3.3.

Στη συνέχεια, χρησιμοποιήσαμε το IG FHT μοντέλο παλινδρόμησης στα ίδια δεδομένα. Στη συνάρτηση πυκνότητας πιθανότητας της σχέσης (3.1), οι παράμετροι x_0 και m μοντελοποιήθηκαν ως εξής:

$$\ln x_0 = \boldsymbol{\beta}'\boldsymbol{x}\ , \quad m = \boldsymbol{\gamma}'\boldsymbol{z}\ ,$$

όπου $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$ είναι τα διανύσματα των υπό εκτίμηση συντελεστών της παλινδρόμησης. Όπως έχουμε αναφέρει, τα δύο σύνολα μεταβλητών $\boldsymbol{x}$ και $\boldsymbol{z}$ μπορεί να είναι εξ' ολοκλήρου ή μερικώς διαφορετικά. Ωστόσο, στη δική μας εφαρμογή όλες οι έντεκα μεταβλητές που ελέγχθηκαν με το μοντέλο του Cox χρησιμοποιήθηκαν σαν πιθανοί προγνωστικοί παράγοντες και για τις δύο παραμέτρους x_0 και m.

			FHT παλινδρόμηση			
	Παλινδρόμηση Cox		**$\ln x_0$**		***m***	
	Συντελ.	**Τυπ. απόκ.**	**Συντελ.**	**Τυπ. απόκ.**	**Συντελ.**	**Τυπ. απόκ.**
Ηλικία	0.049	(0.008)	-0.029	(0.004)	-0.003	(0.001)
Έκτακτη επέμβαση	0.580	(0.178)	-0.569	(0.126)		
PHF	0.430	(0.158)				
PVD	0.507	(0.153)	-0.376	(0.094)		
COPD	0.366	(0.158)	0.952	(0.101)	-0.125	(0.022)
BMI			0.019	(0.010)		
Επείγουσα επέμβαση			0.773	(0.123)		
CHF			-1.018	(0.087)	0.083	(0.022)
Αρτηριο-σκλήρωση			-0.508	(0.108)		
Δυσλειτουργία Ανοσοποιητικού			1.334	(0.417)		
OPCAB					0.089	(0.036)
Σταθερά			1.739	(0.362)	0.351	(0.071)

Πίνακας 3.3: εκτιμήτριες των παραμέτρων (στατιστικά σημαντικές σε ε.σ. 5%) από την παλινδρόμηση Cox και την FHT παλινδρόμηση

Το μοντέλο προσαρμόστηκε με το στατιστικό πακέτο STATA 8 χρησιμοποιώντας τις εντολές *ml model* και *ml maximize* για τον καθορισμό και την προσαρμογή του μοντέλου με τη μέθοδο μεγιστοποίησης της πιθανοφάνειας. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 3.3. Η πλειοψηφία των μεταβλητών, εκτός από τις OPCAB και PHF, είναι στατιστικά σημαντικές για το αρχικό σημείο εκκίνησης της ανέλιξης. Ένα τέτοιο αποτέλεσμα είναι αναμενόμενο. Το γεγονός πως συμπεριλήφθηκαν εκ των προτέρων στη μελέτη σαν πιθανοί προγνωστικοί παράγοντες της επιβίωσης μετά από επέμβαση ανοιχτής καρδιάς στη στεφανιαία αρτηρία υποδεικνύει πως η ιατρική κοινότητα τις θεωρεί ως σημαντικούς παράγοντες για την περιγραφή της πορείας της ασθένειας.

Αντιθέτως, μόνο τέσσερις μεταβλητές φαίνεται να είναι στατιστικά σημαντικές στην παλινδρόμηση για την παράμετρο της κλίσης *m*, η οποία περιγράφει τον τρόπο με τον οποίο εξελίσσεται η ασθένεια. Αυτές είναι οι *ηλικία,* OPCAB, CHF και COPD. Το αποτέλεσμα αυτό είναι πιο κοντινό με εκείνο της ανάλυσης που πραγματοποιήθηκε με το μοντέλο του Cox, η οποία εξετάζει την πορεία της συνάρτησης διακινδύνευσης με την πάροδο του χρόνου, τη στιγμή που έχει επιλεγεί ένας μικρός αριθμός από παράγοντες κινδύνου. Ωστόσο, οι πέντε σημαντικές μεταβλητές από το μοντέλο του Cox και οι τέσσερις από το FHT μοντέλο έχουν κοινές μόνο τις μεταβλητές *ηλικία* και COPD. Χρησιμοποιώντας τους τέσσερις FHT προγνωστικούς παράγοντες στο μοντέλο του Cox, η προσαρμογή φαίνεται να είναι λιγότερο ισχυρή (-2 φορές η πιθανοφάνεια δίνει τιμή 2141.5 σε σύγκριση με την 2119.22: το μοντέλο χωρίς κανένα παράγοντα έχει τιμή 2197.52).

Όλοι οι παράγοντες που παρέμειναν στην ανάλυση για την παράμετρο *m* στο μοντέλο FHT έχουν μικρούς συντελεστές σχετικά με την τιμή του σταθερού όρου, με αποτέλεσμα οι εκτιμώμενες τιμές για την παράμετρο *m* να είναι θετικές για όλους εκτός από οκτώ (1.5%) από τους ασθενείς της μελέτης. Το γεγονός πως γενικά $\hat{m} > 0$ συμφωνεί με την εντύπωση που δίνεται στο Σχήμα 3.4, πως η καμπύλη της επιβίωσης δε συγκλίνει προς το μηδέν: υπάρχει μεγάλο ποσοστό μακροπρόθεσμων επιζώντων.

Η παλινδρόμηση Κατωφλιού χωρίζει τις μεταβλητές σε δύο σύνολα: εκείνες που σχετίζονται με την επιβίωση λόγω σχέσης με την αρχική κατάσταση της υγείας του ασθενή και εκείνες που σχετίζονται με την επιβίωση, επειδή συνδέονται με την αλλαγή της κατάστασης της υγείας μετά την επέμβαση. Αυτά τα δύο σύνολα δεν είναι αμοιβαία αποκλειόμενα. Στην πραγματικότητα, οι μεταβλητές *ηλικία*, CHF και COPD εμφανίζονται και στις δύο εκτιμήτριες των συντελεστών της παλινδρόμησης $\hat{x}_0$ και $\hat{m}$. Ο συντελεστής της μεταβλητής *ηλικία* έχει το ίδιο (αρνητικό) πρόσημο στις γραμμικές εκτιμήτριες τόσο του x_0 όσο και του *m*. Συνεπώς, μεγαλύτερη ηλικία έχει την τάση να συνδέεται με χαμηλότερο αρχικό σημείο εκκίνησης για την κατάσταση της υγείας του ασθενή (μικρότερο x_0) και πιο απότομη κλίση προς το κατώφλι (αρνητικό *m*). Να υπενθυμίσουμε πως μικρότερο *m* σημαίνει και μεγαλύτερη τιμή για την πιθανότητα να πραγματοποιηθεί μετάβαση στο

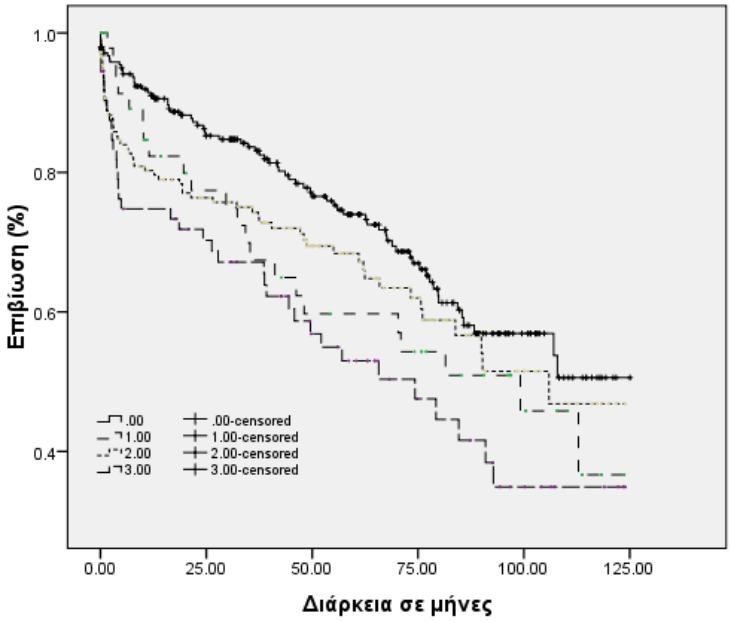
κατώφλι, $p_0 = \exp(-2x_0 m)$, με αποτέλεσμα ο ασθενής να πεθάνει. Οι δύο αυτές επιδράσεις είναι προς την ίδια κατεύθυνση: ο πιο μεγάλος σε ηλικία ασθενής έχει χειρότερη πορεία από ένα νεότερο ασθενή.

Από την άλλη, για τις μεταβλητές CHF και COPD οι συντελεστές της παλινδρόμησης των x_0 και m έχουν αντίθετα πρόσημα. Τωρινή καρδιακή ανεπάρκεια δείχνει να συνδέεται με ένα χαμηλότερο αρχικό σημείο εκκίνησης της κατάστασης της υγείας, αλλά και με μία πιο αργή κλίση προς το σύνορο. Αντιθέτως, ενδεχόμενη χρόνια αποφρακτική πνευμονική ασθένεια φαίνεται να συνδέεται με καλύτερο αρχικό σημείο εκκίνησης της κατάστασης της υγείας, αλλά με πιο γρήγορη πορεία προς το κατώφλι. Το φαινόμενο αυτό, μιας μεταβλητής η οποία να έχει αντίθετες επιδράσεις στις δύο παραμέτρους, απασχόλησε επίσης και τους Lee et al. (2009) σε μία ανάλυση θνησιμότητας λόγω καρκίνου στον πνεύμονα ανάμεσα σε εργάτες των σιδηροδρόμων. Στην πραγματικότητα, στην ανάλυση αυτή, και οι τρεις μεταβλητές που χρησιμοποιήθηκαν στη μελέτη εμφάνισαν το φαινόμενο των αντίθετων προσήμων. Ένας τρόπος που βρήκαν προκειμένου να απομονώσουν την καθαρή επίδραση των μεταβλητών αυτών, ήταν να ασχοληθούν με την αναμενόμενη διάρκεια ζωής μέχρι το θάνατο για κάθε δυνατό συνδυασμό των τιμών των μεταβλητών. Η δική μας εφαρμογή είναι πιο σύνθετη από τη δική τους, λόγω της ύπαρξης και μακροπρόθεσμων επιζώντων, γεγονός που υποδεικνύει πως ασχολούμαστε με υπο συνθήκη κατανομή των χρόνων επιβίωσης. Επιπρόσθετα, καθώς η IG κατανομή μπορεί να γίνει ιδιαιτέρως λοξή, η χρησιμοποίηση των μέσων τιμών μπορεί να είναι παραπλανητική. Αντιθέτως, στον Πίνακα 3.4 παρατηρούμε τις πιθανότητες μακροχρόνιας επιβίωσης για κάθε ένα συνδυασμό των μεταβλητών CHF και COPD. Οι τιμές αυτές προέκυψαν με τον υπολογισμό της εκτιμώμενης τιμής της ποσότητας $1-\exp(2x_0 m)$ για τον κάθε ασθενή και στη συνέχεια λαμβάνοντας τη μέση τιμή τους.

COPD	**Πρόσφατη CHF**	**Πιθανότητα Επιβίωσης**
Όχι	Όχι	0.441
Ναι	Όχι	0.297
Όχι	Ναι	0.285
Ναι	Ναι	0.280

Πίνακας 3.4: Εκτιμώμενη πιθανότητα μακροχρόνιων επιζώντων με την παρουσία / απουσία των COPD και CHF

Τα αποτελέσματα αυτά μπορεί να συγκριθούν με τις Kaplan Meier καμπύλες για κάθε συνδυασμό των μεταβλητών CHF και COPD, όπως φαίνεται στο Σχήμα 3.5. Παρατηρούμε πως παρόλο που υπάρχουν διαφορετικά μοτίβα εμφάνισης προσήμων που συνδέονται με τις μεταβλητές COPD και CHF, η παρουσία οποιασδήποτε εκ των δύο οδηγεί σε πιθανότερο ενδεχόμενο θάνατο.

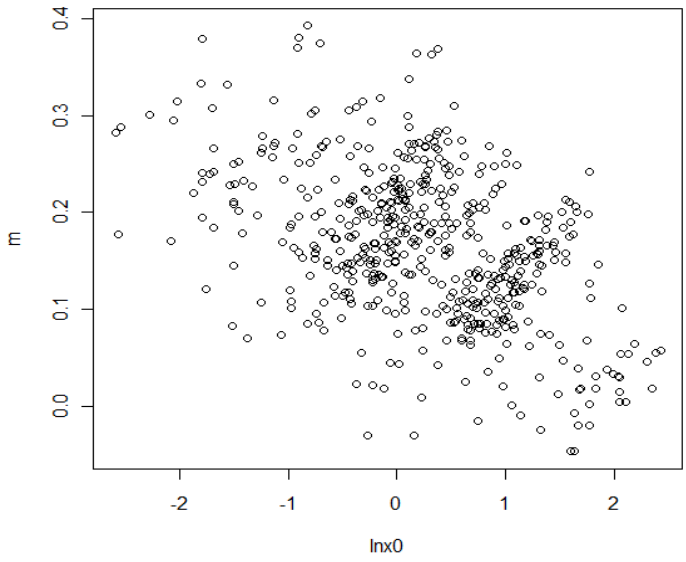


Σχήμα 3.5: Εκτιμήτριες Kaplan – Meier της επιβίωσης των 4 συνδυασμών των μεταβλητών COPD και CHF (0 κανένα, 1 μόνο COPD, 2 μόνο CHF, 3 και τα 2).

Παρόλο που τα μοντέλα χρόνων πρώτης μετάβασης υπάρχουν στη βιβλιογραφία για κάποια μεγάλη χρονική περίοδο, μόλις τον τελευταίο καιρό έχουν αρχίσει να γίνονται ιδιαίτερα δημοφιλή. Κάποιες από τις ελκυστικές τους ιδιότητες συζητήθηκαν και στην παρούσα παράγραφο. Σε αντίθεση με το μοντέλο αναλογικής διακινδύνευσης του Cox και άλλες συνηθισμένες προσεγγίσεις, προσφέρουν την ευκαιρία μίας πιο αποκαλυπτικής μοντελοποίησης η οποία προχωράει πέρα από μία απλή περιγραφή του προβλήματος (Lee et al., 2010). Πρόσφατες επεκτάσεις αυξάνουν την ευελιξία τους ακόμα παραπέρα, για παράδειγμα με την εισαγωγή μοντέλων που επιτρέπουν την ύπαρξη χρονικά μεταβαλλόμενων μεταβλητών (Lee et al., 2010) και την ημι-παραμετρική μοντελοποίηση των μεταβλητών απόκρισης (Yu et al., 2009). Δημοσιευμένες εφαρμογές παρουσιάζουν την ικανότητά τους να αναπαριστούν περιπτώσεις που τα PH μοντέλα δεν μπορούν, όπως είναι η περίπτωση κατά την οποία οι συναρτήσεις διακινδύνευσης διασταυρώνονται (Zhang και Peng, 2009).

Ωστόσο, συγκεκριμένα προβλήματα παραμένουν. Ένα από αυτά τα προβλήματα, το οποίο έχουν επίσης παρατηρήσει και άλλοι ερευνητές και εμφανίστηκε και στη δική μας εφαρμογή, αφορά την ύπαρξη αντίθετων προσήμων της ίδιας συμμεταβλητής στις παραμέτρους x_0 και *m*. Οι Lee και Whitmore (2006) υποδεικνύουν πως οι εκτιμήτριες των παραμέτρων στην FHT παλινδρόμηση ενδέχεται να παρουσιάζουν σημαντική πολυσυγγραμικότητα, καθώς η μέση διάρκεια ζωής είναι $x_0/|m|$. Το συμπέρασμα αυτό παρουσιάζεται και γραφικά στο Σχήμα 3.6, όπου φαίνεται μία ιδιαίτερα υψηλή αρνητική συσχέτιση (ο συντελεστής συσχέτισης του Pearson είναι -0.379) ανάμεσα στις τιμές του $\hat{x}_0$ και του $\hat{m}$ για τις μονάδες στα δεδομένα μας. Επιπρόσθετα, παρατηρούμε πως στα ίδια

συμπεράσματα μπορούμε να καταλήξουμε από το συνδυασμό ενός υψηλού αρχικού σημείου εκκίνησης x_0 και απότομης κλίσης m, ή από ένα μικρότερο αρχικό σημείο εκκίνησης συνοδευόμενο από πιο αργή κλίση. Υπάρχει, λοιπόν, ένας έμφυτος βαθμός μη αναγνωρισημότητας (έλλειψης διακριτικής ικανότητας) των παραμέτρων στο μοντέλο. Το γεγονός αυτό δεν έχει οδηγήσει ποτέ σε τεχνικά προβλήματα κατά τη διάρκεια μεγιστοποίησης των πιθανοφανειών σε κανένα από τα σύνολα δεδομένων που εξετάσαμε, αλλά μπορεί να είναι υπεύθυνο για την έλλειψη σαφήνειας σχετικά με την ακρίβεια του ρόλου μιας μεταβλητής που αφορά ταυτόχρονα και τις δύο παραμέτρους. Το συγκεκριμένο πρόβλημα δε θα πρέπει να παρουσιάζεται σε εφαρμογές όπου οι παράμετροι εξαρτώνται από διαφορετικές μεταβλητές.



Σχήμα 3.6: Γράφημα των εκτιμητριών της παραμέτρου κλίσης *m*, έναντι των εκτιμητριών του αρχικού σημείου εκκίνησης της ανέλιξης

Στη βιβλιογραφία, υπάρχει εκτενής περιγραφή στις εφαρμογές του μοντέλου του Cox. Η τεράστια συσσωρευμένη εμπειρία έχει αναπτύξει και έναν άτυπο, εμπειρικό κανόνα, ο οποίος επιτρέπει τη χρήση του μοντέλου του Cox, μόνο στην περίπτωση που υπάρχουν τουλάχιστον δέκα παρατηρήσεις για κάθε μία μεταβλητή του προβλήματος (Concato et al., 1995; Peduzzi et al., 1995; Vittinghoff και McCulloch, 2007). Σαν αποτέλεσμα, η ανάλυση προστατεύεται από προβλήματα εκτίμησης, όπως μεροληψία και ανακριβή ποσοστά διαστημάτων εμπιστοσύνης. Προς το παρόν, δεν υπάρχουν αντίστοιχες μελέτες για την περίπτωση του FHT μοντέλου, προκειμένου να μελετήσουμε πότε το μέγεθος ενός δείγματος δεδομένων είναι κατάλληλο για την εφαρμογή ενός FHT μοντέλου σε αυτό.

Για την περίπτωση του IG μοντέλου που θεωρήσαμε σε αυτήν την παράγραφο, ένα αντίστοιχο ζήτημα που μας απασχολεί, αποτελεί το εάν μία μεταβλητή που εισάγεται στο μοντέλο και για τις δύο παραμέτρους θα έπρεπε να θεωρηθεί σαν δύο διαφορετικές

μεταβλητές για το συγκεκριμένο σκοπό. Επιπρόσθετα, μπορεί να σημαίνει πως για την άρση της μερικής έλλειψης διακριτικής ικανότητας που αναφέραμε προηγουμένως απαιτείται ένα ακόμα μεγαλύτερο μέγεθος δείγματος. Στην επόμενη παράγραφο, θα μελετήσουμε τα συγκεκριμένα ζητήματα που προκύπτουν κατά την προσαρμογή του μοντέλου FHT.

3.3 Διερεύνηση πρακτικών θεμάτων που προκύπτουν κατά την προσαρμογή του FHT μοντέλου παλινδρόμησης για δεδομένα διάρκειας ζωής.

3.3.1 Εισαγωγή

Πολλοί συγγραφείς υποδεικνύουν το γεγονός πως η ύπαρξη της ίδιας μεταβλητής και στις δύο παραμέτρους στο μοντέλο IG FHT μπορεί να επιφέρει ένα βαθμό έλλειψης διακριτικής ικανότητας ή πολυσυγγραμικότητας στο μοντέλο. Η συχνή εμφάνιση του φαινομένου των αντίθετων προσήμων για τους δύο συντελεστές της συμμεταβλητής στις δύο παραμέτρους μπορεί να σχετίζεται με αυτό το φαινόμενο.

Στην παράγραφο που ακολουθεί, γίνεται μία προσπάθεια να αποδώσουμε εμπειρικές αποδείξεις σχετικά με τη δυνατότητα προσαρμογής του μοντέλου της παλινδρόμησης Κατωφλιού. Ειδικότερα, εξετάζουμε εάν υπάρχει κάποια τάση κατά τη διαδικασία προσαρμογής να τοποθετείται μία μεταβλητή σε λάθος παράμετρο. Επιπρόσθετα, μελετάμε το φαινόμενο εμφάνισης αντίθετων προσήμων μίας μεταβλητής στις διάφορες παραμέτρους της κατανομής.

3.3.2 Επίδραση των συμμεταβλητών

Όπως είδαμε και στην προηγούμενη παράγραφο, είναι προφανές από τη σχέση (3.4) πως ενδεχόμενη μεταβολή σε οποιαδήποτε από τις δύο παραμέτρους x_0 και m μπορεί να επιφέρει παρόμοιες επιδράσεις στην αναμενόμενη διάρκεια ζωής. Για παράδειγμα, μικρότερες διάρκειες ζωής μπορεί να προκύπτουν λόγω χαμηλότερων τιμών για το αρχικό σημείο εκκίνησης, ή λόγω πιο απότομης κλίσης της πορείας της ανέλιξης μέχρι την πρώτη διακοπή, ή και ως συνδυασμός των παραπάνω δύο κινήσεων. Έτσι λοιπόν, μία μεταβλητή z_i μπορεί να επηρεάσει τις διάρκειες ζωής με δύο διαφορετικούς τρόπους, μέσω των συντελεστών β_i και γ_i. Εάν οι εκτιμήτριες των συντελεστών της παλινδρόμησης έχουν το ίδιο πρόσημο, τότε οι επιδράσεις αυτές βρίσκονται σε συμφωνία. Για παράδειγμα, όταν β_i και γ_i είναι και οι δύο θετικοί, τότε μεγαλύτερες τιμές για τη z_i σχετίζονται με υψηλότερο σημείο εκκίνησης και λιγότερο απότομη κλίση της ανέλιξης, δύο δράσεις που υποδεικνύουν πως μία μονάδα με μεγαλύτερη τιμή για τη z_i αναμένεται να ζήσει περισσότερο από μία μονάδα με χαμηλότερη τιμή για τη z_i. Από την άλλη, αντίθετα πρόσημα για τους συντελεστές β_i και γ_i βρίσκονται

σε ασυμφωνία με την έννοια πως ο ένας υποδεικνύει καλύτερο αποτέλεσμα, ενώ ο άλλος χειρότερο.

Σε δημοσιευμένες εργασίες (για παράδειγμα Eberly et al., 2001; Lee et al., 2009), το φαινόμενο των αντίθετων προσήμων φαίνεται να είναι πολύ συχνό, γεγονός που οδηγεί σε δυσκολίες στην ερμηνεία των τελικών αποτελεσμάτων. Όπως είδαμε και στην προηγούμενη παράγραφο, στην εργασία των Lee et al. (2009) σε μία ανάλυση θνησιμότητας λόγω καρκίνου στον πνεύμονα ανάμεσα σε εργάτες των σιδηροδρόμων και οι τρεις μεταβλητές που χρησιμοποιήθηκαν στη μελέτη εμφάνισαν το φαινόμενο των αντίθετων προσήμων. Για παράδειγμα, για τη θνησιμότητα λόγω καρκίνου στον πνεύμονα, κάθε εκτιμήτρια των συντελεστών της παλινδρόμησης εμφάνισε ένα θετικό πρόσημο για το αρχικό σημείο εκκίνησης, αλλά ένα αρνητικό για την κλίση της ανέλιξης, όπως φαίνεται στον Πίνακα 3.5.

	Variable	**Estimate**	**P-value**
$\ln x_0$			
	Μηχανικός	1.15681	< 0.001
	Καπνιστής	0.08389	< 0.001
	Εκτεθειμένος σε αμίαντο	0.08202	< 0.001
	Σταθερά	2.40497	< 0.001
m			
	Μηχανικός	-0.89114	< 0.001
	Καπνιστής	-0.16693	< 0.001
	Εκτεθειμένος σε αμίαντο	-0.07824	< 0.001
	Σταθερά	-0.27750	< 0.001

Πίνακας 3.5:

Στην περίπτωση της μεταβλητής *μηχανικός* οι συγγραφείς πρότειναν πως εργάτες που ακολουθούσαν το επάγγελμα του μηχανοδηγού μπορεί να ήταν σε καλύτερη πρότερη φυσική κατάσταση από άλλους εργάτες, αλλά λόγω συνεχούς έκθεσης σε αναθυμιάσεις καυσαερίων ντίζελ, σταδιακά έχασαν πιο απότομα την καλή αυτή κατάσταση της υγείας τους σε σχέση με άλλους εργάτες. Ωστόσο, είναι δύσκολο να συμπεράνει κανείς πώς εργάτες που κάπνιζαν ή που ήταν εκτεθειμένοι σε αμίαντο θα έπρεπε να έχουν εκ των προτέρων καλύτερη κατάσταση υγείας.

Κάποιες φορές μπορεί να υπάρχουν λόγοι, ώστε να επιτρέψουμε σε κάποια μεταβλητή να συνδεθεί μόνο με μία από τις δύο παραμέτρους του μοντέλου. Είναι ευκολότερο να δικαιολογήσουμε την ανάγκη απαγόρευσης μίας μεταβλητής να συνδεθεί με την παράμετρο

του αρχικού σημείου της ανέλιξης, παρά με την παράμετρο της κλίσης. Για παράδειγμα, η αγωγή που πρόκειται να ληφθεί σε μία τυχαιοποιημένη κλινική δοκιμή λογικά δε θα πρέπει να συνδέεται με το x_0 (Pennell et al., 2010). Οι Aalen και Gjessing (2001) πρότειναν οι μεταβλητές να χωριστούν σε δύο είδη, εκείνες που μετράνε πόσο πολύ έχει προχωρήσει η πορεία της ανέλιξης και εκείνες που αναπαριστούν αιτιακές επιρροές στην ανάπτυξη της ανέλιξης. Οι πρώτες προτείνεται να μοντελοποιηθούν ασκώντας επιρροή στην παράμετρο x_0, ενώ οι τελευταίες στην παράμετρο *m*. Ωστόσο, ορισμένες μεταβλητές που εμφανίζονται συχνά στις διάφορες εφαρμογές της Ανάλυσης Επιβίωσης, όπως η ηλικία ενός ασθενή επιβάλλεται να συνδεθούν και με τις δύο παραμέτρους ταυτόχρονα. Έτσι λοιπόν, κάποιες φορές μπορεί να είναι απαραίτητο κατά την προσαρμογή του μοντέλου να επιτρέψουμε σε κάποια μεταβλητή να επιδράσει και στις δύο παραμέτρους.

Η πιθανότητα μία μεταβλητή να έχει παρόμοια αποτελέσματα στις διάρκειες ζωής μέσω της επίδρασής της στο x_0 ή το *m*, εγείρει την απορία για το εάν αυτά τα αποτελέσματα μπορεί στην πραγματικότητα να διαχωριστούν κατά την προσαρμογή του μοντέλου, όταν δεν είναι εφικτό να αποκλειστούν εκ των προτέρων από κάποια εκ των δύο παραμέτρων. Έτσι, έχει παρατηρηθεί πως "δεν είναι ακόμα ξεκάθαρο εάν προκύπτουν θέματα διάκρισης σε περίπτωση που μεταβλητές περιλαμβάνονται και στις δύο παραμέτρους" (Eberly et al., 2001) και "θα είναι δύσκολο να αποδώσουμε την επίδραση σε κάποιο συγκεκριμένο συστατικό του μοντέλου… οι εκτιμήτριες των επιδράσεων των συμμεταβλητών του αρχικού σημείου και της κλίσης μπορεί να είναι συγγραμικές" (Lee et al., 2006).

Ο κύριος σκοπός της παραγράφου αυτής είναι να παρουσιάσουμε εμπειρικές αποδείξεις σχετικά με ενδεχόμενη δυνατότητα διάκρισης των επιδράσεων μιας μεταβλητής στις διάρκειες ζωής μέσω των επιδράσεων στις παραμέτρους x_0 και *m*. Επιπρόσθετα, θα μελετήσουμε εάν ενδεχόμενη πιθανή εγγενής έλλειψη διακριτικής ικανότητας ή πολυσυγγραμικότητα επιφέρει πρακτικές επιπλοκές. Θα προσανατολιστούμε ιδιαίτερα στο να ανακαλύψουμε εάν υπάρχει κάποια τάση κατά τη διαδικασία προσαρμογής του μοντέλου να τοποθετεί μία μεταβλητή σε "λάθος" παράμετρο. Τέλος, θα μελετήσουμε εάν είναι πιθανό η παρατηρούμενη εμφάνιση αντίθετων προσήμων στους συντελεστές της παλινδρόμησης να αυξάνεται σε περίπτωση που η μεταβλητή έπρεπε εκ των προτέρων να συνδέεται μόνο με μία και όχι με δύο παραμέτρους.

3.3.3 Σχεδιασμός της μελέτης

Δεδομένα δημιουργήθηκαν από ένα IG FHT μοντέλο παλινδρόμησης με μία ή δύο μεταβλητές, για συγκεκριμένες τιμές των συντελεστών $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$ της παλινδρόμησης, κάποιες από τις οποίες μπορεί να είναι ίσες με το μηδέν. Η στρατηγική για την επιλογή των μη-μηδενικών συντελεστών περιγράφεται στη συνέχεια. Στην παρούσα μελέτη, δεν υπήρξε

κανένας περιορισμός στην προσαρμογή του IG FHTR μοντέλου, δηλαδή κάθε μεταβλητή είχε το δικαίωμα να εισέλθει και στις δύο παραμέτρους κατά τη διαδικασία προσαρμογής.

Η μεταβλητή z_1 είναι Ομοιόμορφα κατανεμημένη στο διάστημα $[0,1]$. Στις περιπτώσεις που χρησιμοποιήθηκε και μία δεύτερη μεταβλητή z_2, αυτή δημιουργήθηκε από την Ομοιόμορφη κατανομή στο διάστημα $[1,2]$, ανεξάρτητα από την πρώτη.

Επειδή οι αριθμητικές τιμές των συντελεστών $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$ δεν έχουν κάποια ευθεία και ακριβή ερμηνεία (Eberly et al., 2001), οι τιμές τους επιλέχθηκαν σύμφωνα με το μέγεθος της επίδρασης τους στη μέση (αναμενόμενη) διάρκεια ζωής της σχέσης (3.4). Η ιδέα αυτή μας βοήθησε να επιλέξουμε τις τιμές για τους συντελεστές $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$ με τέτοιο τρόπο, ώστε να είναι κατά μία έννοια συγκρίσιμοι. Πιο συγκεκριμένα, ας θεωρήσουμε την περίπτωση "μία μεταβλητή επιδρά και στις δύο παραμέτρους". Σύμφωνα με τις σχέσεις (3.3) και (3.4), η επίδραση στη μέση διάρκεια ζωής μιας μοναδιαίας αύξησης της μεταβλητής z_1 δίνεται από τον τύπο:

$$E\left(\mathrm{T}_{z_1+1}\right)=\frac{x_0}{|m|}=\frac{\exp\left(\gamma_0+\gamma_1 z_1+\gamma_1\right)}{\left|\beta_0+\beta_1 z_1+\beta_1\right|} \tag{3.11}$$

Εάν η μεταβλητή z_1 δεν επιδρά στην παράμετρο m $\left(\beta_1=0\right)$, αλλά επιδρά στην παράμετρο x_0 $\left(\gamma_1\neq 0\right)$, η σχέση (3.11) γίνεται:

$$E\left(\mathrm{T}_{z_1+1}\right)=\frac{\exp\left(\gamma_0+\gamma_1 z_1+\gamma_1\right)}{\left|\beta_0\right|}=\frac{\exp\left(\gamma_0+\gamma_1 z_1\right)\exp\left(\gamma_1\right)}{\left|\beta_0\right|}=\frac{\exp\left(\gamma_0+\gamma_1 z_1\right)}{\left|\beta_0\right|}\exp\left(\gamma_1\right) \quad \Rightarrow$$

$$E\left(\mathrm{T}_{z_1+1}\right)=\frac{\exp\left(\gamma_0+\gamma_1 z_1\right)}{\left|\beta_0\right|}\exp\left(\gamma_1\right)=\mathrm{E}\left(\mathrm{T}_{z_1}\right)e^{\gamma_1} \quad \Rightarrow$$

$$\mathrm{E}\left(\mathrm{T}_{z_1+1}\right)=\mathrm{E}\left(\mathrm{T}_{z_1}\right)e^{\gamma_1}, \tag{3.12}$$

για κάθε z_1. Σαν αποτέλεσμα, μπορούμε να επιλέξουμε τιμές για το συντελεστή γ_1, οι οποίες θα επιφέρουν συγκεκριμένες ποσοστιαίες μεταβολές στην $\mathrm{E}(\mathrm{T})$. Για παράδειγμα, μία μείωση 10% μπορεί να δημιουργηθεί για $\gamma_1=-0.1054$.

Για την περίπτωση όπου η μεταβλητή επιδρά μόνο στην παράμετρο m $\left(\beta_1\neq 0\right)$, αλλά όχι στη x_0 $\left(\gamma_1=0\right)$, η σχέση (3.11) γίνεται:

$$E\left(\mathrm{T}_{z_1+1}\right)=\frac{\exp\left(\gamma_0\right)}{\left|\beta_0+\beta_1 z_1+\beta_1\right|}=\frac{\exp\left(\gamma_0\right)}{\left|\beta_0+\beta_1 z_1+\beta_1\right|}=\frac{\exp\left(\gamma_0\right)/\left|\beta_0+\beta_1 z_1\right|}{\left|\beta_0+\beta_1 z_1+\beta_1\right|/\left|\beta_0+\beta_1 z_1\right|} \quad \Rightarrow$$

$$E\left(\mathrm{T}_{z_1+1}\right)=\frac{\exp(\gamma_0)}{|\beta_0+\beta_1 z_1|}\cdot\left(\frac{|\beta_0+\beta_1 z_1+\beta_1|}{|\beta_0+\beta_1 z_1|}\right)^{-1}=E\left(\mathrm{T}_{z_1}\right)\cdot\frac{|\beta_0+\beta_1 z_1|}{|\beta_0+\beta_1 z_1+\beta_1|}\quad\Rightarrow$$

$$E\left(\mathrm{T}_{z_1+1}\right)=E\left(\mathrm{T}_{z_1}\right)\cdot\frac{|\beta_0+\beta_1 z_1|}{|\beta_0+\beta_1 z_1+\beta_1|}\tag{3.13}$$

Η παραπάνω σχέση εξαρτάται από τις διάφορες τιμές της μεταβλητής z_1. Επιτρέποντας στη z_1 να πάρει οποιαδήποτε τιμή μέσα στο διάστημα $[0,1]$ και στη συνέχεια θέτοντας στην παραπάνω σχέση $z_1=0$, τότε η ποσότητα $\frac{|\beta_0|}{|\beta_0+\beta_1|}$ αναπαριστά την αλλαγή από το ένα ακραίο σημείο του διαστήματος $[0,1]$ $(z_1=0)$, στο άλλο $(z_1=1)$. Σε αυτήν την ειδική περίπτωση, εάν θέσουμε $\beta_0=-1$, τότε μία μείωση 10% της αναμενόμενης διάρκειας ζωής λόγω μιας μοναδιαίας αύξησης στη μεταβλητή z_1, επιτυγχάνεται λύνοντας την εξίσωση:

$$\frac{1}{|\beta_1-1|}=0.9,\tag{3.14}$$

δίνοντας σαν αποτέλεσμα την τιμή $\beta_1=-0.1111$, ή την τιμή $\beta_1=2.1111$. Επιλέγουμε την πρώτη λύση, καθώς θέλουμε να επικεντρωθούμε στην περίπτωση $m<0$.

Για τη διερεύνηση της περίπτωσης $m>0$, θέτουμε $\beta_0=0.5$ και η εξίσωση που αντιστοιχεί στη σχέση (3.14) γίνεται:

$$\frac{1}{|1+2\beta_1|}=0.9,$$

από την οποία επιλέγουμε τη μεγαλύτερη από τις δύο λύσεις της β_1, προκειμένου να εξασφαλίσουμε πως $m>0$. Οι τιμές του γ_0 επιλέχθηκαν με τέτοιο τρόπο, ώστε η πιθανότητα (της σχέσης (3.5)) η ανέλιξη να μη φτάσει ποτέ στο σύνορο, υπολογισμένη στη μέση τιμή της μεταβλητής z_1 να είναι περίπου ίση με 0.5:

$$p_1=1-p_0=1-\exp\left(-2x_0m\right)<0.5\quad\Leftrightarrow$$

$$1-e^{(-2x_0m)}<\frac{1}{2}\quad\Leftrightarrow\quad e^{(-2x_0m)}>\frac{1}{2}\quad\Leftrightarrow\quad -2x_0m>\ln\frac{1}{2}\quad\Leftrightarrow$$

$$e^{\gamma_0+\frac{\gamma_1}{2}}\left(\beta_0+\frac{\beta_1}{2}\right)<0.35$$

Οι τιμές των συντελεστών που επιλέχτηκαν σύμφωνα με αυτή τη μέθοδο χρησιμοποιήθηκαν και για την περίπτωση των δύο μεταβλητών.

Για κάθε μία μονάδα στο προσομοιωμένο δείγμα, οι τιμές των IG παραμέτρων κατασκευάστηκαν από τη σχέση (3.3), ύστερα από τη δημιουργία της μεταβλητής (ή των μεταβλητών) για την κάθε μονάδα. Για την περίπτωση που κάποια μονάδα είχε $m > 0$, εργαστήκαμε με τον τρόπο που περιγράψαμε στην Παράγραφο 3.1.3.

Σχετικά με τα μεγέθη των προσομοιωμένων δειγμάτων, περιμέναμε πως όσο μικρότερο είναι το δείγμα, τόσο πιο εμφανή θα είναι τα διάφορα προβλήματα κατά την προσαρμογή του μοντέλου. Έτσι λοιπόν, επιλέξαμε να μη χρησιμοποιήσουμε ιδιαίτερα μεγάλο μέγεθος δείγματος. Επιπρόσθετα, η ύπαρξη σχετικά μικρού δείγματος, όπως για παράδειγμα είκοσι ή τριάντα παρατηρήσεων δεν αποτελεί ρεαλιστική περίπτωση, καθώς θα είναι δύσκολη η διάκριση ανάμεσα στο FHT και σε άλλα μοντέλα διάρκειας ζωής. Για τους παραπάνω λόγους, ασχοληθήκαμε με τις περιπτώσεις $n = 50$ και $n = 100$, για τα μεγέθη των προσομοιωμένων δειγμάτων.

Όλοι οι υπολογισμοί έγιναν στην R. Ψευδοτυχαίες τιμές από την IG μεταβλητή δημιουργήθηκαν χρησιμοποιώντας τη ρουτίνα *rinvgauss* της βιβλιοθήκης *statmod* (Smyth, 2012). Το FHT μοντέλο παλινδρόμησης προσαρμόστηκε με ελαχιστοποίηση της συνάρτησης του αρνητικού λογαρίθμου της πιθανοφάνειας χρησιμοποιώντας τη ρουτίνα *optim*. Υποθέσεις της μορφής $\beta_i = 0$ και $\gamma_j = 0$ ελέγχθηκαν με τον έλεγχο του λόγου των πιθανοφανειών (συγκρίνοντας τα μοντέλα πριν και μετά την προσθήκη του υπό εξέταση όρου) και με τον έλεγχο Wald (το τετράγωνο του λόγου της εκτίμησης του συντελεστή προς το ασυμπτωτικά τυπικό σφάλμα της εκτίμησης αυτής), χρησιμοποιώντας τις ασυμπτωτικές χ^2 κατανομές για αυτές τις ελεγχοσυναρτήσεις.

3.3.4 Αποτελέσματα για την περίπτωση της μιας συμμεταβλητής

Δεδομένα δημιουργήθηκαν κάτω από τρία βασικά σενάρια για την περίπτωση της μίας μεταβλητής:

1. Η μεταβλητή επιδρά μόνο στο αρχικό σημείο εκκίνησης x_0 της ανέλιξης ($\gamma_1 \neq 0$, $\beta_1 = 0$).
2. Η μεταβλητή επιδρά μόνο στην παράμετρο κλίσης *m* της ανέλιξης ($\gamma_1 = 0$, $\beta_1 \neq 0$).
3. Η μεταβλητή επιδρά και στις δύο παραμέτρους της IG κατανομής ($\gamma_1 \neq 0$, $\beta_1 \neq 0$).

Για το τρίτο σενάριο πρέπει να γίνει μια επιπλέον διερεύνηση για τις περιπτώσεις που οι συντελεστές β_1 και γ_1 έχουν το ίδιο ή διαφορετικό πρόσημο, δηλαδή έχουν επίδραση προς την ίδια ή αντίθετη κατεύθυνση. Τέλος, για όλα τα σενάρια οι περιπτώσεις $m < 0$ και $m > 0$ πρέπει επίσης να διαχωριστούν.

Αρχικά θεωρούμε τα δύο σενάρια στα οποία η μεταβλητή επιδρά μόνο σε μία παράμετρο, με $m < 0$. Ο Πίνακας 3.6 παρουσιάζει τα προσομοιωμένα μεγέθη των ελέγχων για την υπόθεση $\gamma_1 = 0$, όταν $\beta_1 \neq 0$ και την υπόθεση $\beta_1 = 0$, όταν $\gamma_1 \neq 0$. Τα

αποτελέσματα παρουσιάζονται για τους ελέγχους με το λόγο των πιθανοφανειών (LR) και του Wald, σε ονομαστικό επίπεδο 5%, χρησιμοποιώντας την χ^2 κατανομή με ένα βαθμό ελευθερίας. Η τιμή κάθε κελιού του πίνακα είναι βασισμένη σε πεντακόσιες προσομοιώσεις, εκτός από τις μέσες τιμές της τελευταίας γραμμής. Αυτή η διαδικασία δίνει ένα διωνυμικό τυπικό σφάλμα, με τιμή κάτω από μία ποσοστιαία μονάδα σε ένα εκτιμημένο ποσοστό περίπου ίσο με 5%. Η μέση τιμή της κάθε στήλης, που παρουσιάζεται στην τελευταία γραμμή του πίνακα, βασίζεται σε 4,500 επαναλήψεις, με αποτέλεσμα το διωνυμικό τυπικό σφάλμα της να είναι 0.3%.

	Μέγεθος (%) του ελέγχου για την υπόθεση $\gamma_1 = 0$				Μέγεθος (%) του ελέγχου για την υπόθεση $\beta_1 = 0$			
	Αληθινό μοντέλο: $\gamma_1 = 0$ και $\beta_1 \neq 0$				Αληθινό μοντέλο: $\beta_1 = 0$ και $\gamma_1 \neq 0$			
Μέγεθος επίδρασης της β_1 ή της γ_1	$n = 50$		$n = 100$		$n = 50$		$n = 100$	
	Wald	LR	Wald	LR	Wald	LR	Wald	LR
+75%	5.6	5.2	5.6	5.4	4.4	4.6	5.8	5.8
+50%	5.6	5.6	4.6	4.2	5.0	5.0	4.4	4.6
+25%	4.2	3.8	5.6	5.2	6.6	7.0	6.2	6.2
+10%	6.4	5.8	5.0	4.8	4.2	4.4	7.4	7.6
0	5.8	5.6	3.4	3.4	4.6	4.6	4.4	4.4
-10%	7.8	7.0	5.8	5.6	3.4	3.4	7.2	7.2
-25%	8.0	7.2	5.6	5.2	5.6	5.6	6.8	7.4
-50%	7.2	6.2	5.8	5.6	4.6	4.8	6.4	6.6
-75%	6.2	6.4	5.6	5.8	4.8	5.2	5.6	5.8
Μέση τιμή	6.3	5.9	5.2	5.0	4.8	5.0	6.0	6.2

Πίνακας 3.6: Προσομοιώσεις για την περίπτωση όπου μία μεταβλητή επιδρά είτε στο *m* είτε στο $\mathbf{x_0}$, είτε και στις δύο παραμέτρους, όταν $m < 0$. Μέγεθος (%) των ελέγχων Wald και LR για τις υποθέσεις $\gamma_1 = 0$ όταν $\beta_1 \neq 0$ και $\beta_1 = 0$ όταν $\gamma_1 \neq 0$, σε θεωρητικό επίπεδο 5%, με $\beta_0 = -1$ και $\gamma_0 = 1$.

Ο Πίνακας 3.7 παρουσιάζει τα αντίστοιχα αποτελέσματα για την περίπτωση που $m > 0$. Γενικά, σε όλη τη διάρκεια της μελέτης, δεν καταλήξαμε σε διαφορετικά αποτελέσματα ανάλογα με το πρόσημο του *m*. Για το λόγο αυτό, στη συνέχεια θα παρουσιάσουμε αποτελέσματα μόνο για την περίπτωση όπου $m < 0$.

Τα αποτελέσματα στους Πίνακες 3.6 και 3.7 υποδεικνύουν πως το μέγεθος των ελέγχων Wald και LR παραμένουν κοντά στο ονομαστικό επίπεδο 5%, ακόμα και όταν το μέγεθος του δείγματος είναι μόλις $n = 50$. Τα μεγέθη των ελέγχων δε φαίνεται να επηρεάζονται από

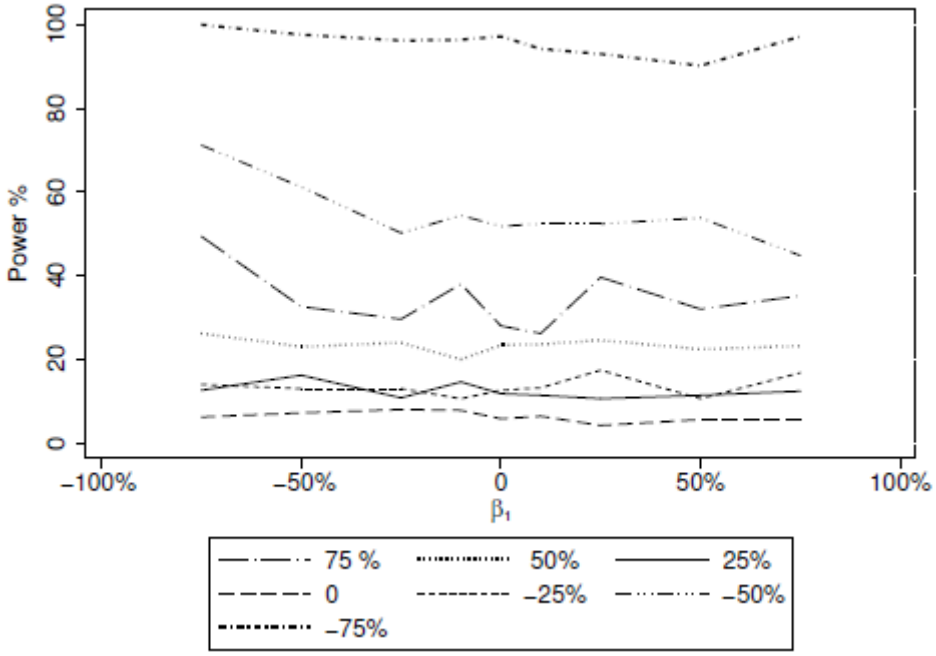
την τιμή της άλλης παραμέτρου. Τα αποτελέσματα των δύο ελέγχων φαίνεται να είναι πολύ κοντά γενικά. Επιπρόσθετα, οι έλεγχοι Wald και LR διαφωνούν σε ένα πολύ μικρό αριθμό μεμονωμένων επαναλήψεων (μικρότερο από 1% - τα δεδομένα δεν παρουσιάζονται στους παραπάνω πίνακες). Η γενική αυτή ισοδυναμία παρατηρήθηκε κατά τη διάρκεια ολόκληρης της μελέτης. Συνεπώς, για τα επόμενα, θα παρουσιάζουμε μόνο τα αποτελέσματα του ελέγχου LR.

	Μέγεθος (%) του ελέγχου για την υπόθεση $\gamma_1 = 0$				Μέγεθος (%) του ελέγχου για την υπόθεση $\beta_1 = 0$			
	Αληθινό μοντέλο: $\gamma_1 = 0$ και $\beta_1 \neq 0$				Αληθινό μοντέλο: $\beta_1 = 0$ και $\gamma_1 \neq 0$			
Μέγεθος επίδρασης της β_1 ή γ_1	$n = 50$		$n = 100$		$n = 50$		$n = 100$	
	Wald	LR	Wald	LR	Wald	LR	Wald	LR
+75%	6.0	5.4	6.0	6.2	3.2	4.4	4.6	5.4
+50%	3.8	3.8	4.6	4.6	3.6	5.2	4.6	5.0
+25%	5.8	5.8	3.6	3.6	3.6	3.8	4.2	4.8
+10%	5.4	5.6	6.0	5.8	4.6	6.0	4.4	5.2
0	5.2	5.4	4.6	5.0	5.2	5.4	4.0	4.2
-10%	5.6	5.4	4.2	4.2	3.8	4.2	3.8	4.4
-25%	6.0	6.0	6.0	5.8	3.6	4.2	5.8	6.6
-50%	5.0	4.0	5.8	5.8	4.6	5.6	4.8	5.4
-75%	6.4	6.4	6.0	6.0	3.6	4.8	5.0	5.2
Μέση τιμή	5.5	5.3	5.2	5.2	4.0	4.8	4.6	5.1

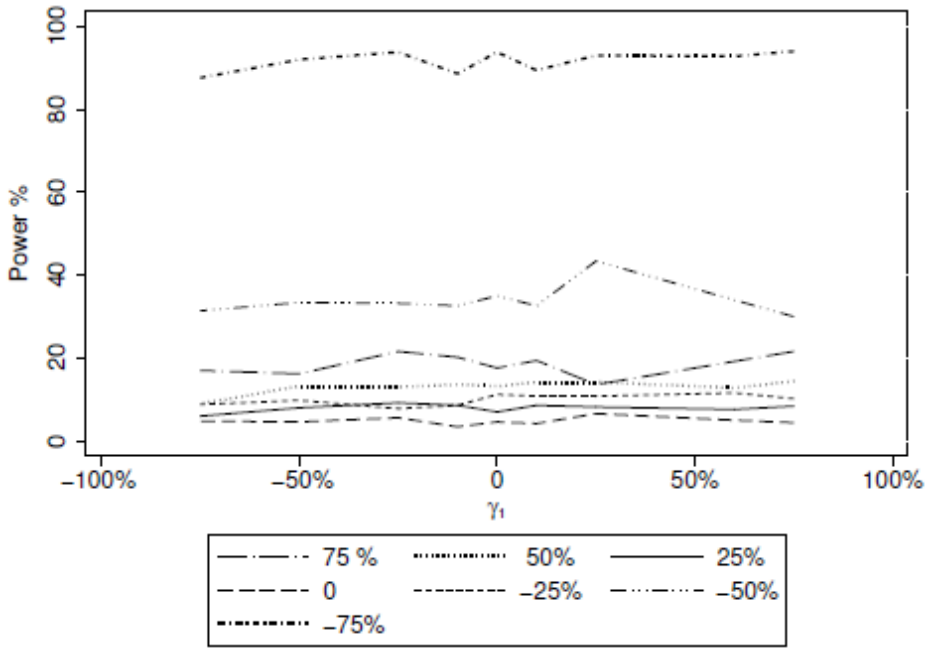
Πίνακας 3.7: Προσομοιώσεις για την περίπτωση όπου μία μεταβλητή επιδρά είτε στο *m* είτε στο $\mathbf{x_0}$, είτε και στις δύο παραμέτρους, όταν $m > 0$. Μέγεθος (%) των ελέγχων Wald και LR για τις υποθέσεις $\gamma_1 = 0$ όταν $\beta_1 \neq 0$ και $\beta_1 = 0$ όταν $\gamma_1 \neq 0$, σε θεωρητικό επίπεδο 5%, με $\beta_0 = 0.5$.

Τα αποτελέσματα για το τρίτο σενάριο, με β_1 και γ_1 ταυτόχρονα μη-μηδενικά παρουσιάζονται στα Σχήματα 3.7 και 3.8. Τα γραφήματα αυτά παρουσιάζουν την ισχύ των ελέγχων για τις υποθέσεις $\gamma_1 = 0$ (Σχήμα 3.7) και $\beta_1 = 0$ (Σχήμα 3.8), χρησιμοποιώντας τον έλεγχο του Wald σε ένα ονομαστικό 5% επίπεδο σημαντικότητας, για δείγματα μεγέθους $n = 50$. Για παράδειγμα, όταν οι τιμές των συντελεστών β_1 και γ_1 αναπαριστούν μία αύξηση κατά 75% στην αναμενόμενη διάρκεια ζωής, οι προσομοιωμένες τιμές της ισχύος του ελέγχου $\gamma_1 = 0$ είναι 35.2%, σύμφωνα με το Σχήμα 3.7. Όμοια, όταν οι τιμές των β_1 και γ_1 αναπαριστούν μία αύξηση 75% στην αναμενόμενη διάρκεια ζωής, οι προσομοιωμένες τιμές της ισχύος που προκύπτουν από το Σχήμα 3.8 είναι 21.6% για την απόρριψη της $\beta_1 = 0$. Και στα δύο σχήματα, παρατηρούμε ότι όσο μεγαλώνει η αρνητική αύξηση στη

μεταβολή της αναμενόμενης διάρκειας ζωής που επιφέρει η επιλογή του εκάστοτε συντελεστή, μεγαλώνει αντίστοιχα και η ισχύς του αντίστοιχου ελέγχου για το συντελεστή.

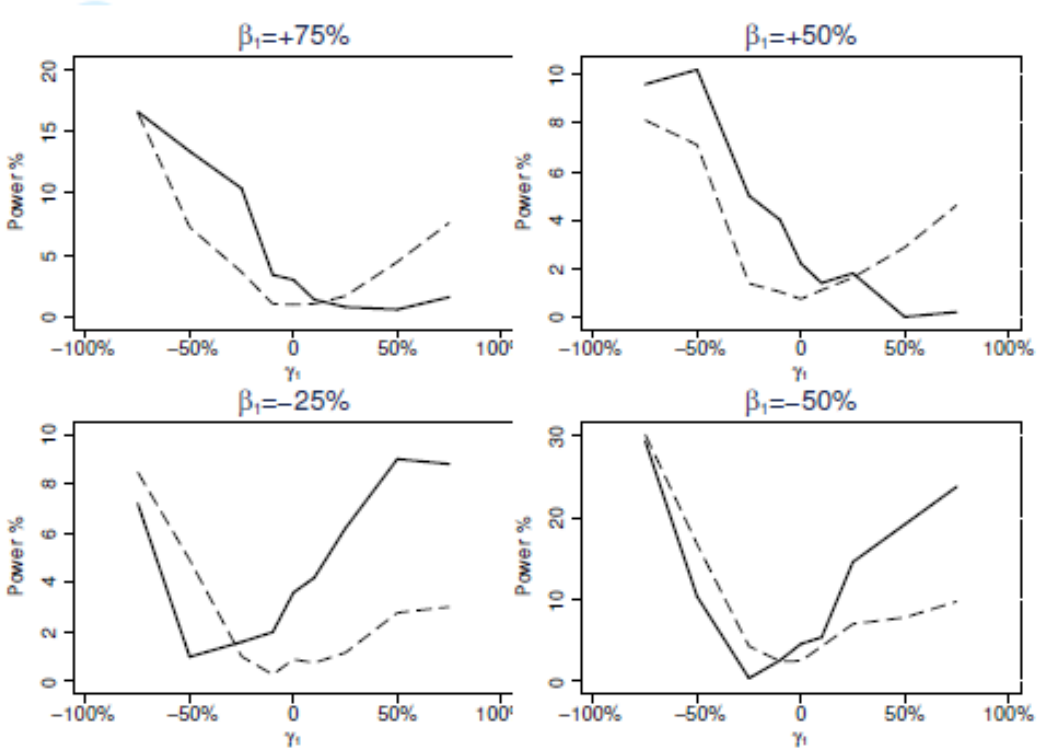


Σχήμα 3.7: Ισχύς του LR ελέγχου της υπόθεσης $\gamma_1 = 0$, για διαφορετικές τιμές της παραμέτρου γ_1 στο αληθινό μοντέλο, σε σχέση με την αληθινή τιμή του β_1. Οι τιμές των συντελεστών όπως στον Πίνακα 3.6, με *n* = 50.



Σχήμα 3.8: Ισχύς του LR ελέγχου της υπόθεσης $\beta_1 = 0$, για διαφορετικές τιμές της παραμέτρου β_1 στο αληθινό μοντέλο, σε σχέση με την αληθινή τιμή του γ_1. Οι τιμές των συντελεστών όπως στον Πίνακα 3.6, με *n* = 50.

Στις ίδιες προσομοιώσεις από τις οποίες προέκυψαν τα αποτελέσματα των Σχημάτων 3.7 και 3.8 μετρήθηκε επίσης και το ποσοστό των φορών στις οποίες και οι δύο μηδενικές υποθέσεις $\beta_1 = 0$ και $\gamma_1 = 0$ απορρίφθηκαν ταυτόχρονα. Τα δεδομένα αυτά παρουσιάζονται στο Σχήμα 3.9 για επιλεγμένους συνδυασμούς των συντελεστών β_1 και γ_1, μαζί με το αναμενόμενο ποσοστό των φορών που θα απορρίπταμε ταυτόχρονα και τις δύο υποθέσεις, εάν οι έλεγχοι των δύο υποθέσεων $\beta_1 = 0$ και $\gamma_1 = 0$ δρούσαν ανεξάρτητα.



Σχήμα 3.9: Ισχύς των LR ελέγχου της ταυτόχρονης απόρριψης των υποθέσεων $\beta_1 = 0$ και $\gamma_1 = 0$ (συνεχόμενη γραμμή), σε σύγκριση με την αναμενόμενη ισχύ εάν οι δύο έλεγχοι γινόντουσαν ξεχωριστά (διακεκομμένη γραμμή).

Για παράδειγμα, όταν οι τιμές των β_1 και γ_1 μαζί αναπαριστούν μία αύξηση 75%, οι προσομοιωμένες τιμές της ισχύος που προκύπτουν από τα Σχήματα 3.7 και 3.8 είναι 35.2% για την απόρριψη της $\gamma_1 = 0$ και 21.6% για την απόρριψη της $\beta_1 = 0$. Το αναμενόμενο ποσοστό των ταυτόχρονων απορρίψεων κάτω από καθεστώς ανεξαρτησίας είναι $35.2 \times 21.6/100 = 7.6\%$ (διακεκομμένη γραμμή), αλλά το παρατηρημένο ποσοστό είναι μόνο 1.6% (συνεχόμενη γραμμή).

Το Σχήμα 3.9 συγκρίνει τις προσομοιωμένες τιμές της ταυτόχρονης απόρριψης των υποθέσεων $\beta_1 = 0$ και $\gamma_1 = 0$, σε αντίθεση με την αναμενόμενη ισχύ των δύο ελέγχων εάν $\beta_1 = 0$ και $\gamma_1 = 0$, εάν ήταν ανεξάρτητοι. Όταν η μεταβλητή επιδρά και στις δύο παραμέτρους x_0 και m προς την ίδια κατεύθυνση (δηλαδή όταν οι συντελεστές β_1 και γ_1 έχουν το ίδιο πρόσημο), η διακεκομμένη γραμμή (αναμενόμενη ισχύς) βρίσκεται πάνω από

τη συνεχόμενη γραμμή (πραγματική ισχύς). Το γεγονός αυτό σημαίνει πως οι δύο υποθέσεις απορρίπτονται ταυτόχρονα λιγότερο συχνά από όταν οι έλεγχοι ήταν ανεξάρτητοι. Επομένως, πρέπει να υπάρχει μία τάση για το προσαρμοσμένο μοντέλο να συγκεντρώνει τα δύο παρόμοια αποτελέσματα σε μία μόνο παράμετρο. Αντιθέτως, όταν β_1 και γ_1 έχουν αντίθετα πρόσημα, οπότε η επίδραση της μεταβλητής γίνεται στις δύο παραμέτρους προς αντίθετη κατεύθυνση, οι δύο υποθέσεις απορρίπτονται ταυτόχρονα πιο συχνά από όταν οι έλεγχοι ήταν ανεξάρτητοι.

Οι επιδράσεις αυτές μπορεί να είναι ιδιαίτερα μεγάλες. Για παράδειγμα, όταν οι τιμές των β_1 και γ_1 αντιστοιχούν μαζί σε μία μείωση 50% στην $E(S)$, η συχνότητα των ταυτόχρονων απορρίψεων των δύο υποθέσεων είναι 10.4% σε σύγκριση με την αναμενόμενη τιμή 20.1%, εάν οι δύο έλεγχοι δρούσαν ανεξάρτητα.

Οι Πίνακες 3.8 και 3.9 δείχνουν πως τα β_1 και γ_1 αντίστοιχα εκτιμώνται χωρίς μεροληψία. Οι ρίζες των μέσων τετραγωνικών σφαλμάτων (RMSE) φαίνεται να είναι μεγάλες, αλλά το μέγεθος του δείγματος είναι μόλις $n = 50$.

	Μέγεθος επίδρασης και τιμή για την παράμετρο β_1				
	+75%	**+25%**	**0**	**-25%**	**-75%**
γ_1	0.43	0.20	0	-0.33	-3.0
+75%	0.4543	0.1833	0.0442	-0.3083	-3.1131
	(0.3909)	(0.4470)	(0.5165)	(0.6140)	(0.9014)
+25%	0.4065	0.2192	-0.0754	-0.3087	-2.9670
	(0.4435)	(0.5259)	(0.5618)	(0.5805)	(0.8168)
0	0.4432	0.1818	-0.0229	-0.3323	-2.9712
	(0.4072)	(0.4906)	(0.4844)	(0.6113)	(0.8751)
-25%	0.4565	0.2743	0.0159	-0.3031	-3.0967
	(0.4205)	(0.4822)	(0.0159)	(0.6979)	(0.9657)
-75%	0.4570	0.1054	0.0657	-0.4043	-3.0493
	(0.4818)	(0.6062)	(0.5915)	(0.6961)	(1.0058)
Μέση τιμή	0.4435	0.1928	0.0055	-0.3313	-3.0395
	(0.4289)	(0.5132)	(0.5483)	(0.6411)	(0.9154)

Πίνακας 3.8: Μέσες τιμές και RMSE των εκτιμήσεων της β_1 για τις διάφορες τιμές της παραμέτρου γ_1. Οι τιμές των συντελεστών όπως στον Πίνακα 3.6, με *n* = 50.

	Μέγεθος επίδρασης και τιμή για την παράμετρο γ_1				
	+75%	**+25%**	**0**	**-25%**	**-75%**
$\boldsymbol{\beta_1}$	0.5596	0.2231	0	-0.2877	-1.3893
+75%	0.5680	0.2745	-0.0146	-0.3013	-1.3743
	(0.3671)	(0.4167)	(0.3738)	(0.3537)	(0.3452)
+25%	0.5907	0.2129	0.0252	-0.3259	-1.3421
	(0.3683)	(0.4467)	(0.3746)	(0.3561)	(0.3911)
0	0.5315	0.2692	0.0004	-0.2864	-1.4238
	(0.4036)	(0.4166)	(0.3664)	(0.4095)	(0.3792)
-25%	0.5347	0.1954	0.0174	-0.2999	-1.3711
	(0.4179)	(0.3819)	(0.3998)	(0.4387)	(0.3861)
-75%	0.5502	0.1699	0.0522	-0.2899	-1.4294
	(0.2824)	(0.3082)	(0.3025)	(0.2816)	(0.3202)
Μέση τιμή	0.5550	0.2244	0.0161	-0.3007	-1.3881
	(0.3078)	(0.3669)	(0.3648)	(0.3719)	(0.3654)

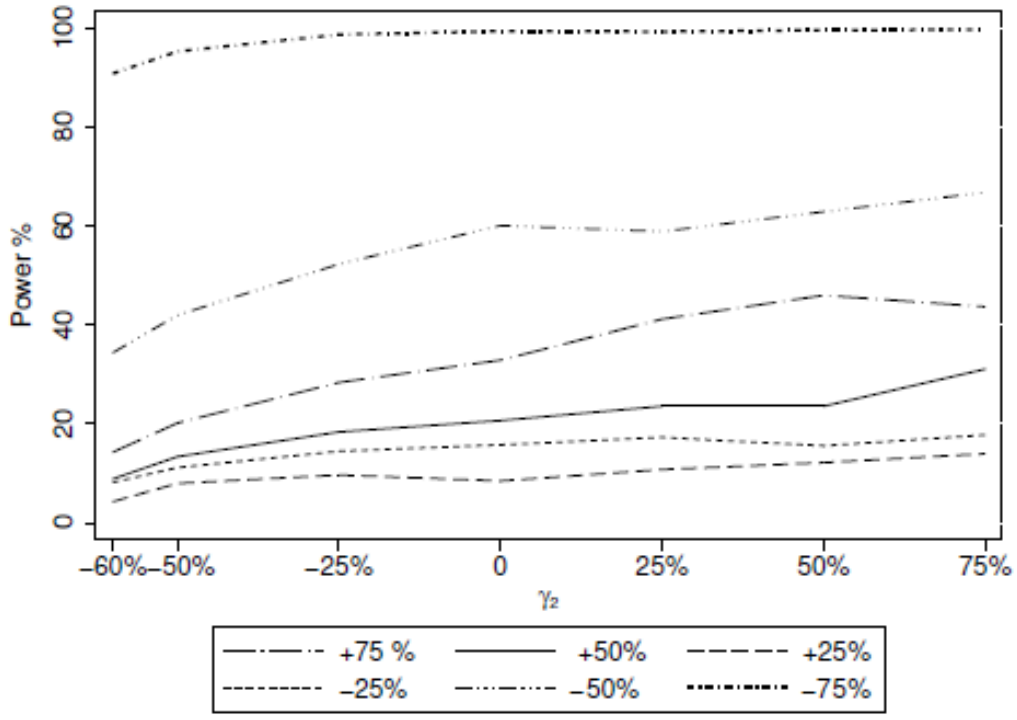
Πίνακας 3.9: Μέσες τιμές και RMSE των εκτιμήσεων της γ_1 για τις διάφορες τιμές της παραμέτρου β_1 . Οι τιμές των συντελεστών όπως στον Πίνακα 3.6, με *n* = 50.

3.3.5 Αποτελέσματα για την περίπτωση των δύο συμμεταβλητών

Η μελέτη για την περίπτωση της μιας μεταβλητής μας παρέχει ήδη κάποιες πληροφορίες σχετικά με την εκτίμηση των συντελεστών, όταν μία μεταβλητή εμφανίζεται στους γραμμικούς εκτιμητές και των δύο παραμέτρων. Για το λόγο αυτό, στην περίπτωση των δύο μεταβλητών ασχοληθήκαμε με το σενάριο στο οποίο οι μεταβλητές του αληθινού μοντέλου επιδρούν σε διαφορετικές παραμέτρους. Συγκεκριμένα, η πρώτη μεταβλητή επιδρά στην παράμετρο *m*, ενώ η δεύτερη μεταβλητή επιδρά στο x_0. Συνεπώς, κατά τη δημιουργία των δεδομένων, ο συντελεστής β_1 μπορούσε να λάβει πληθώρα τιμών, αλλά ο συντελεστής β_2 ήταν πάντα ίσος με το μηδέν. Όμοια, ο συντελεστής γ_2 μπορούσε να πάρει πολλές διαφορετικές τιμές, αλλά ο συντελεστής γ_1 ήταν πάντοτε μηδέν. Ωστόσο, κανένας περιορισμός δεν επιβλήθηκε στις παραμέτρους, κατά την προσαρμογή του IG FHTR μοντέλου. Δηλαδή, και οι δύο μεταβλητές είχαν το δικαίωμα να εισέλθουν και στις δύο παραμέτρους. Οι τιμές των συντελεστών της παλινδρόμησης υπολογίστηκαν με παρόμοιο τρόπο με τη μελέτη προσομοιώσεων για την περίπτωση της μίας μεταβλητής με $m < 0$.

Έλεγχοι LR του λόγου των πιθανοφανειών πραγματοποιήθηκαν για την εκτίμηση της κάθε παραμέτρου ξεχωριστά, χρησιμοποιώντας τα κρίσιμα σημεία της ασυμπτωτικά χ^2 κατανομής. Δοκιμάστηκαν τέσσερα διαφορετικά σχήματα επιλογών των τιμών των συντελεστών β_1 και γ_2 στο αληθινό μοντέλο ως προς τη συνεισφορά τους στη μεταβολή της αναμενόμενης διάρκειας ζωής: (θετική μεταβολή και για τους δύο, μόνο για το β_1, μόνο για το γ_2, αρνητική μεταβολή και για τους δύο). Τα αποτελέσματα παρουσιάζονται στο ονομαστικό επίπεδο σημαντικότητας 5%. Τα μεγέθη των ελέγχων για τις υποθέσεις $\beta_2 = 0$ και $\gamma_1 = 0$ ήταν πολύ κοντά σε αυτό το επίπεδο. Σε τριάντα έξι σύνολα των χιλίων προσομοιωμένων δειγμάτων για το κάθε σύνολο, για διάφορους συνδυασμούς τιμών των συντελεστών β_1 και γ_2 με $n = 50$, ο μέσος όρος απόρριψης των υποθέσεων $\beta_2 = 0$ και $\gamma_1 = 0$ ήταν 5.54% (τυπική απόκλιση 0.81) και 6.11% (0.92), αντίστοιχα. Για $n = 100$, οι αντίστοιχες τιμές ήταν 5.15% (0.64) και 5.63% (0.83). Συμπεραίνουμε, λοιπόν, πως ο έλεγχος για την υπόθεση $\beta_2 = 0$ φαίνεται να είναι πιο ακριβής από τον έλεγχο της υπόθεσης $\gamma_1 = 0$, αλλά μόλις κατά μισή μόνο ποσοστιαία μονάδα.

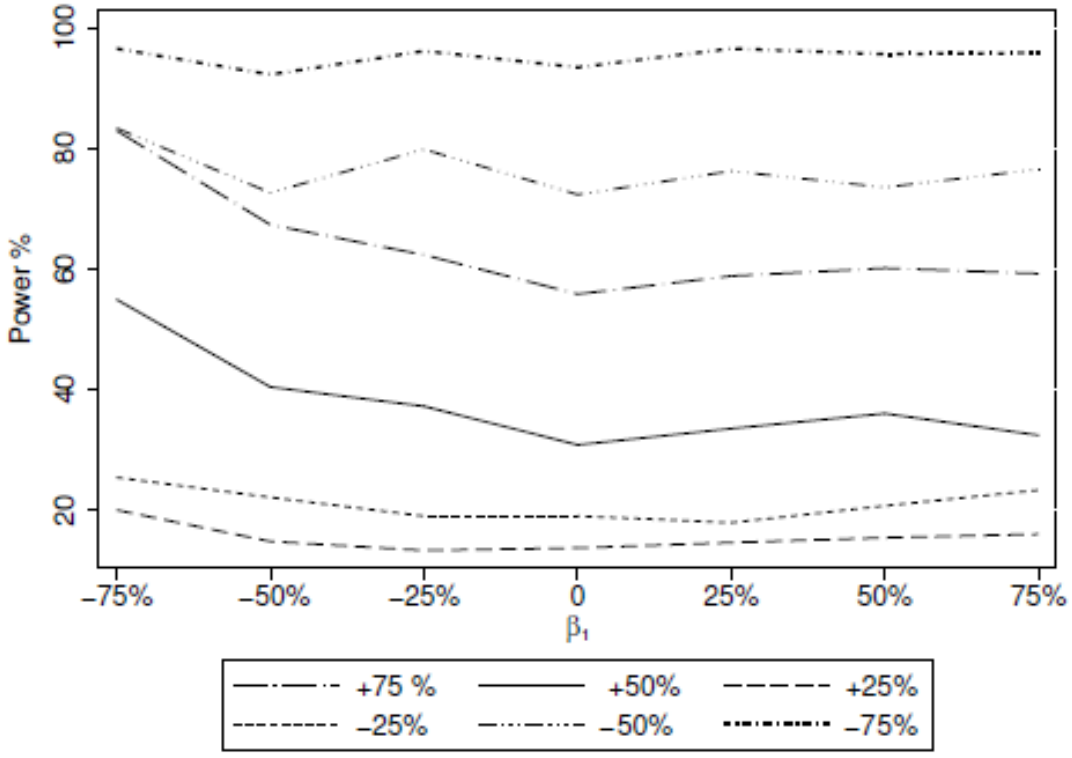
Το Σχήμα 3.10 δείχνει το ποσοστό των προσομοιωμένων επαναλήψεων στις οποίες η μηδενική υπόθεση $\beta_1 = 0$ απορρίπτεται στο ονομαστικό επίπεδο σημαντικότητας 5%, για διάφορους συνδυασμούς τιμών των β_1 και γ_2. Φαίνεται πως η ισχύς του LR ελέγχου για την απόρριψη της $\beta_1 = 0$ αυξάνεται με την αύξηση του γ_2 για σταθερό β_1.



Σχήμα 3.10: Ισχύς του LR ελέγχου της υπόθεσης $\beta_1 = 0$, για διαφορετικές τιμές της παραμέτρου β_1 στο αληθινό μοντέλο, σε σχέση με την αληθινή τιμή του γ_2.

Το Σχήμα 3.11 δείχνει τα αντίστοιχα ποσοστά για την απόρριψη της υπόθεσης $\gamma_2 = 0$ σε ένα 5% ονομαστικό επίπεδο. Στην περίπτωση αυτή, υπάρχει μία τάση για την ισχύ του ελέγχου να μειώνεται, καθώς αυξάνεται η τιμή του β_1 για σταθερό γ_2.

Για την εξήγηση αυτών των αποτελεσμάτων, έχουμε τα εξής: από τη σχέση $E(T) = \frac{x_0}{|m|}$ για την αναμενόμενη διάρκεια ζωής της μονάδας φαίνεται πως μεγαλύτερες τιμές του γ_2 υπονοούν και μεγαλύτερες διάρκειες ζωής. Το αποτέλεσμα αυτό διευκολύνει την ανακάλυψη μεταβολών στις διάρκειες ζωής λόγω αλλαγών στην παράμετρο *m* στον παρανομαστή. Αντίστροφα, μεγαλύτερες τιμές του β_1 υπονοούν μικρότερες αναμενόμενες διάρκειες ζωής, με αποτέλεσμα να είναι σχετικά πιο δύσκολο να αναγνωριστούν μεταβολές λόγω αλλαγών στο x_0.



Σχήμα 3.11: Ισχύς του LR ελέγχου της υπόθεσης $\gamma_2 = 0$, για διαφορετικές τιμές της παραμέτρου γ_2 στο αληθινό μοντέλο, σε σχέση με την αληθινή τιμή του β_1.

Οι στήλες που βρίσκονται στην αριστερή πλευρά του Πίνακα 3.10, παρουσιάζουν τα ποσοστά των φορών που οι εκτιμήσεις των β_1 και γ_1 έχουν το ίδιο πρόσημο ως συνάρτηση της αληθινής τιμής του $\beta_1 \neq 0$ (με $\gamma_1 = 0$). Οι τιμές που εμφανίζονται στα κελιά του πίνακα αποτελούν μέσες τιμές ποσοστών σε ένα εύρος τιμών για την παράμετρο γ_2 ανάμεσα σε $\pm 75\%$, καθώς δεν υπήρχε εμφανής εξάρτηση στο $\gamma_2 = 0$. Τα αποτελέσματα δείχνουν μία απότομα αυξανόμενη τάση για το $\hat{\gamma}_1$ να έχει το ίδιο πρόσημο με το $\hat{\beta}_1$ όσο η αληθινή

επίδραση του β_1 αυξάνει. Όμοια, οι στήλες που βρίσκονται στη δεξιά πλευρά του Πίνακα 3.10 παρουσιάζουν τα ποσοστά των φορών που οι εκτιμήσεις των β_2 και γ_2 έχουν το ίδιο πρόσημο ως συνάρτηση της αληθινής τιμής του $\gamma_2 \neq 0$ (με $\beta_1 = 0$).

Μέγεθος της επίδρασης του β_1	Ίδιο πρόσημο για $\hat{\beta}_1$, $\hat{\gamma}_1$ (%)		Μέγεθος της επίδρασης του γ_2	Ίδιο πρόσημο για $\hat{\beta}_2$, $\hat{\gamma}_2$ (%)	
	$n=50$	$n=100$		$n=50$	$n=100$
+75%	38.7	43.3	+75%	44.4	47.7
+50%	32.8	37.7	+50%	38.3	45.3
+25%	28.1	30.4	+25%	26.5	32.9
-25%	28.4	33.0	-25%	32.5	37.8
-50%	42.6	48.7	-50%	48.9	49.8
-75%	52.7	52.0	-75%	49.3	49.7

Πίνακας 3.10: Ποσοστά 1,000 προσομοιώσεων στις οποίες οι εκτιμήτριες των συντελεστών β_1, γ_1 ή β_2, γ_2 έχουν το ίδιο πρόσημο, σε σχέση με τις πραγματικές τιμές του β_1 ή γ_1.

Όπως και πριν, τα ποσοστά αυτά αποτελούν μέσο όρο ενός εύρους τιμών του β_2. Και εδώ είναι εμφανής η αυξανόμενη τάση του $\hat{\beta}_2$ να έχει το ίδιο πρόσημο με το $\hat{\gamma}_2$, καθώς η αληθινή επίδραση του γ_2 αυξάνει. Με διαφορετική διατύπωση:

"όταν μία μεταβλητή έχει ισχυρή επίδραση σε μία παράμετρο, υπάρχει μία τάση να εμφανίζει μία επίδραση προς την ίδια κατεύθυνση (του ίδιου προσήμου) και στην άλλη παράμετρο."

Το φαινόμενο αυτό ενισχύεται με την αύξηση του μεγέθους του δείγματος.

3.3.6 Συμπεράσματα της μελέτης

Πολλοί συγγραφείς έχουν παρατηρήσει τις πιθανές δυσκολίες που μπορεί να προκύψουν κατά την προσαρμογή ενός FHT μοντέλου, λόγω του γεγονότος πως η ίδια μεταβλητή μπορεί να εισέλθει στην παλινδρόμηση σε δύο μέρη, μέσω της επίδρασής της στους γραμμικούς εκτιμητές των παραμέτρων, υποδεικνύοντας πως το μοντέλο έχει έμφυτα κάποιο βαθμό έλλειψης διακριτικής ικανότητας ή πολυσυγγραμικότητας. Τα αποτελέσματα στις διάφορες πρακτικές εφαρμογές που παραθέσαμε παραπάνω, φαίνεται να επιβεβαιώνουν αυτήν τη δυσκολία, και το Σχήμα 3.6 την επιδεικνύει με την ύπαρξη αρνητικής συσχέτισης

μεταξύ των εκτιμητριών των παραμέτρων x_0 και m για τις μονάδες. Η παρούσα μελέτη σκόπευε να αποκτήσει εμπειρικές αποδείξεις, ώστε να ξεκαθαρίσει αυτά τα θέματα.

Τα αποτελέσματα των προσομοιώσεων προτείνουν πως οι εκτιμήσεις των συντελεστών της παλινδρόμησης συμπεριφέρονται αρκετά καλά κάτω από συγκεκριμένες προϋποθέσεις. Τα μεγέθη των ελέγχων Wald και του λόγου των πιθανοφανειών LR, είναι σωστά, ακόμα και όταν το μέγεθος του δείγματος είναι σχετικά μικρό, όπως στις περιπτώσεις που θεωρήσαμε ότι δεν υπάρχει φαινομενική μεροληψία. Ωστόσο, τα αποτελέσματα επιβεβαιώνουν την ύπαρξη κάποιας εξάρτησης ανάμεσα στις εκτιμήτριες των διάφορων συντελεστών. Συγκεκριμένα,

- ✓ Όταν μία μεταβλητή επιδρά ταυτόχρονα και στις δύο παραμέτρους προς την ίδια κατεύθυνση, υπάρχει μία τάση μόνο για τον ένα από τους δύο συντελεστές της να αναγνωριστεί ως στατιστικά σημαντικός.
- ✓ Αντιθέτως, όταν η μεταβλητή επιδρά ταυτόχρονα και στις δύο παραμέτρους προς διαφορετικές κατευθύνσεις, υπάρχει μία τάση και για τους δύο συντελεστές της να αναγνωριστούν ως στατιστικά σημαντικοί.

3.3.7 Η προσαρμογή ενός FHT μοντέλου σε Weibull δεδομένα

Τα προηγούμενα συμπεράσματα είναι βασισμένα στην προσαρμογή ενός FHT μοντέλου σε δεδομένα παραγόμενα από ένα FHT μοντέλο. Επιτρέπουν τη λάθος εκτίμηση των προδιαγραφών μόνο στην περίπτωση που η διαδικασία προσαρμογής επιτρέπει εσφαλμένα σε μία μεταβλητή να εισέλθει και στις δύο παραμέτρους, ενώ θα έπρεπε να εισέλθει μόνο σε μία.

Για το λόγο αυτό, παραθέτουμε ακόμα μία σύντομη έρευνα σχετικά με την επίδραση ενός λανθασμένα υιοθετημένου FHT μοντέλου, όταν θα έπρεπε να έχει επιλεχτεί για την προσαρμογή των δεδομένων κάποιο διαφορετικό μοντέλο. Ένα αντίστοιχο παράδειγμα παρουσιάσαμε στην προηγούμενη παράγραφο της διατριβής όπου προσαρμόσαμε το IG FHTR μοντέλο σε δεδομένα κατασκευασμένα από μία Weibull παλινδρόμηση με τέσσερις μεταβλητές, εκ των οποίων η μία επιδρούσε μόνο στο m, δύο μόνο στο x_0 και η τέταρτη και στις δύο παραμέτρους (με τους συντελεστές της μεταβλητής να έχουν ίδια πρόσημα και στις δύο παραμέτρους). Φαίνεται πιθανό και λογικό η λάθος αυτή εκτίμηση των προδιαγραφών να είναι υπεύθυνη για τη δημιουργία ενός μεγάλου εύρους επιδράσεων.

Προκειμένου να μελετήσουμε περαιτέρω το συγκεκριμένο πρόβλημα, προχωρήσαμε σε προσομοιώσεις και προσαρμόσαμε το IG FHT μοντέλο σε δεδομένα κατασκευασμένα από μία Weibull παλινδρόμηση με δύο μεταβλητές. Η πρώτη μεταβλητή ακολουθούσε την Ομοιόμορφη κατανομή $U[0,1]$, ενώ η δεύτερη την $N(2,1)$. Όπως και προηγουμένως, δημιουργήσαμε και αποκομμένα δεδομένα με τον τρόπο που περιγράψαμε στην Παράγραφο

3.1.3, τα οποία καταλάμβαναν το 30% του συνολικού δείγματος. Δύο τιμές από το συντελεστή κάθε Weibull παλινδρόμησης χρησιμοποιήθηκαν στις προσομοιώσεις. Οι τιμές αυτές επιλέχθηκαν με τέτοιο τρόπο, ώστε η ισχύς του ελέγχου για την απόρριψη της μηδενικής υπόθεσης των μηδενικών συντελεστών, όταν το Weibull μοντέλο προσαρμόζεται στα δεδομένα, να είναι περίπου 75% και 95% για τις δύο μεταβλητές αντίστοιχα. Αποτελέσματα της προσαρμογής του IG FHT μοντέλου σε αυτά τα δεδομένα παρουσιάζονται στον Πίνακα 3.11 για πεντακόσιες προσομοιώσεις κάθε συνδυασμού αληθινών τιμών της Weibull παλινδρόμησης και μεγέθους δείγματος ίσο με εκατό.

	Αληθινές τιμές των συντελεστών της Weibull παλινδρόμησης			
	(1.1, 0.3)	**(1.1, 0.5)**	**(1.5, 0.3)**	**(1.5, 0.5)**
Απόρριψη της $\beta_1 = 0$ (%)	22.9	24.9	29.2	29.3
Απόρριψη της $\gamma_1 = 0$ (%)	59.9	53.6	58.1	62.9
Απόρριψη και των 2 (%)	11.2	10.9	10.3	11.3
Με ίδιο πρόσημο (%)	0.6	0.3	1.6	3.0
Απόρριψη της $\beta_2 = 0$ (%)	23.0	29.1	27.1	34.2
Απόρριψη της $\gamma_2 = 0$ (%)	52.4	59.6	54.4	61.5
Απόρριψη και των 2 (%)	8.8	9.7	11.4	11.7
Με ίδιο πρόσημο (%)	0.4	4.4	1.3	5.3

Πίνακας 3.11: Αποτελέσματα των LR ελέγχων για τους συντελεστές του προσαρμοσμένου FHT μοντέλου σε δεδομένα δημιουργημένα από ένα μοντέλο Weibull παλινδρόμησης με δύο μεταβλητές, για *n* = 100.

Παρατηρούμε πως:

α) Η FHT προσαρμογή είναι δύο φορές πιο πιθανό να τοποθετήσει την επίδραση μιας μεταβλητής στο x_0 απ' ότι στο *m*,

β) Και οι δύο συντελεστές για την ίδια υπόθεση απορρίπτονται ταυτόχρονα λιγότερο συχνά απ' όταν οι δύο συντελεστές είναι ανεξάρτητοι (όταν οι έλεγχοι γίνονται ξεχωριστά),

γ) Όταν απορρίπτονται και οι δύο συντελεστές για την ίδια υπόθεση, υπάρχει μία τάση να εμφανίζουν αντίθετα πρόσημα.

Εάν θεωρήσουμε το τελευταίο από τα παραπάνω συμπεράσματα μαζί με τα προηγούμενα αποτελέσματα, τα οποία δεν υποδεικνύουν μεγάλα θέματα πολυσυγγραμικότητας στην FHT παλινδρόμηση, φαίνεται πως το εύρημα αντίθετων κατευθύνσεων για τις δύο επιδράσεις της ίδιας μεταβλητής (που φαίνεται να είναι συχνό φαινόμενο στις δημοσιευμένες εργασίες)

ίσως εμφανίζεται κυρίως όταν γίνεται εσφαλμένη προσαρμογή του FHT μοντέλου. Αυτό το συμπέρασμα με τη σειρά του υποδεικνύει την αναγκαιότητα κατασκευής διαγνωστικών ελέγχων και ελέγχων καλής προσαρμογής για το FHT μοντέλο.

3.4 Επιλογή μεταβλητών

3.4.1 Εισαγωγή

Σε αυτήν την παράγραφο, προτείνουμε μία διαδικασία επιλογής μεταβλητών για την περίπτωση του IG FHT μοντέλου παλινδρόμησης. Η ύπαρξη μίας τέτοιας διαδικασίας θεωρείται αναγκαία για την Ανάλυση Επιβίωσης, ιδιαίτερα σε ιατρικές εφαρμογές, στις οποίες συνήθως υπάρχει ένας μεγάλος αριθμός διαθέσιμων υποψήφιων μεταβλητών για τις επιμέρους αναλύσεις. Η προτεινόμενη διαδικασία αποτελείται από δύο διαδοχικές εφαρμογές της προσαρμοσμένης LASSO τεχνικής (adaptive LASSO), μία για κάθε παράμετρο του μοντέλου, εκτελεσμένες από έναν αλγόριθμο ελαχίστων τετραγώνων. Η διαδικασία αποδεικνύεται αποτελεσματική για την ορθή αναγνώριση των μη-μηδενικών (στατιστικά σημαντικών) συντελεστών της παλινδρόμησης. Η παρούσα μελέτη αποτελεί την **πρώτη** συνδρομή στη μεθοδολογία μοντελοποίησης, η οποία είναι απαραίτητη να αναπτυχθεί περαιτέρω για το παρόν μοντέλο παλινδρόμησης.

3.4.2 Διαδικασία επιλογής μεταβλητών

Πολλές διαδικασίες έχουν προταθεί στη βιβλιογραφία για την επιλογή του καλύτερου δυνατού υποσυνόλου από ένα αρχικό σύνολο υποψηφίων μεταβλητών. Όσον αφορά τη δομή παλινδρόμησης της σχέσης (3.3), σκοπός είναι να εντοπίσουμε ποιούς συντελεστές μπορούμε να θέσουμε ίσους με το μηδέν, επιπρόσθετα με όσους είναι ήδη μηδέν εκ των προτέρων. Σύμφωνα με τις παραδοσιακές μεθόδους, η διαδικασία που ακολουθείται συνήθως είναι η προσαρμογή διαφόρων μοντέλων (τα οποία επιλέγονται χειροκίνητα ή μέσω κάποιας αυτοματοποιημένης μεθόδου όπως είναι οι ευρέως διαδεδομένες "κατά βήματα" τεχνικές, οι οποίες χρησιμοποιούνται σε πολλά μοντέλα και κυρίως στο γενικό γραμμικό μοντέλο) και η επιλογή του καλύτερου ανάλογα με την τιμή που δίνει η προσαρμογή του κάθε μοντέλου σε κάποιο συγκεκριμένο κριτήριο. Δύο κριτήρια, τα οποία χρησιμοποιούνται συνήθως, είναι το AIC, με τιμή:

$$\mathrm{AIC} = \frac{-2L\left(M_k\right) + 2p}{n},$$

όπου με $L\left(M_k\right)$ συμβολίζεται ο λογάριθμος της πιθανοφάνειας του μοντέλου και το BIC, με τιμή:

$$\mathrm{BIC} = D\left(M_k\right) - \left(df\right)\ln\left(n\right),$$

όπου με $D(M_k)$ συμβολίζεται η ποσότητα Deviance του μοντέλου. Τα δύο αυτά κριτήρια χρησιμοποιούν το λογάριθμο της πιθανοφάνειας μαζί με κάποιο όρο ποινής, ο οποίος σχετίζεται με τον αριθμό των παραμέτρων στο μοντέλο. Σε αντίθεση με το AIC, στο κριτήριο BIC ο όρος ποινής αυξάνεται απότομα με την αύξηση του μεγέθους του δείγματος (Hardin και Hilbe, 2001).

Περισσότερο σύγχρονες προσεγγίσεις χρησιμοποιούν τεχνικές συρρίκνωσης, οι οποίες υπολογίζουν εκτιμήτριες για όλους τους συντελεστές ταυτόχρονα και επιλέγουν μεταβλητές θέτοντας κάποιους συντελεστές ίσους με το μηδέν. Ένα παράδειγμα μιας τέτοιας τεχνικής είναι η LASSO (Tibshirani, 1996). Οι Wang και Leng (2007) έδειξαν πως η προσέγγιση της ευθείας των ελαχίστων τετραγώνων (Least Square Approximation-LSA) προσφέρει μία βολική υλοποίηση της προσαρμοσμένης τεχνικής LASSO (aLASSO) (Zou, 2006). Η τεχνική που προτείνουμε σε αυτήν την παράγραφο για την επιλογή του καλύτερου συνόλου μεταβλητών στην παλινδρόμηση κατωφλιού είναι βασισμένη στην τεχνική aLASSO.

Ο αλγόριθμος LSA μετασχηματίζει πολλούς διαφορετικούς τύπους αντικειμενικών συναρτήσεων του αλγόριθμου LASSO σε ασυμπτωτικά ισοδύναμα προβλήματα ελαχίστων τετραγώνων. Έστω η αντικειμενική συνάρτηση της προσαρμοσμένης τεχνικής LASSO για την εκτίμηση της παραμέτρου του ενδιαφέροντος $\boldsymbol{\theta}$, δίνεται από τον τύπο:

$$n^{-1}l_n(\boldsymbol{\theta})+\sum_{j=1}^{d}\lambda_j\left|\theta_j\right|, \tag{3.15}$$

όπου l_n είναι η συνάρτηση του αρνητικού λογαρίθμου της πιθανοφάνειας, $\tilde{\boldsymbol{\theta}}$ η εκτιμήτρια μέγιστης πιθανοφάνειας (ε.μ.π.) της $\boldsymbol{\theta}$ και λ_j βελτιωτικές παράμετροι. Αποδεικνύεται πως εάν η $\tilde{\boldsymbol{\theta}}$ είναι $\sqrt{n}$ -συνεπής εκτιμήτρια και ακολουθεί ασυμπτωτικά την Κανονική κατανομή, τότε $\sqrt{n}\left(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)\dot{\sim}N(0,\Sigma)$, όπου Σ είναι ο ασυμπτωτικός πίνακας διασποράς-συνδιασποράς του $\tilde{\boldsymbol{\theta}}$.

Το ανάπτυγμα του πρώτου όρου της σχέσης (3.15) σε σειρά Taylor γύρω από την εκτιμήτρια μέγιστης πιθανοφάνειας $\tilde{\boldsymbol{\theta}}$ μας δίνει:

$$n^{-1}l_n(\boldsymbol{\theta})\approx n^{-1}l_n\left(\tilde{\boldsymbol{\theta}}\right)+n^{-1}\dot{l}_n\left(\tilde{\boldsymbol{\theta}}\right)^{\mathrm{T}}\left(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\right)+\frac{1}{2}\left(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\right)^{\mathrm{T}}\left\{\frac{1}{n}\ddot{l}_n\left(\tilde{\boldsymbol{\theta}}\right)\right\}\left(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\right), \tag{3.16}$$

όπου $\dot{l}_n(\cdot)$ και $\ddot{l}_n(\cdot)$ είναι οι πρώτη και δεύτερη μερική παράγωγος της συνάρτησης $l_n(\cdot)$, αντίστοιχα. Καθώς $\tilde{\boldsymbol{\theta}}$ είναι η ε.μ.π. της $l_n(\cdot)$, ισχύει ότι $\dot{l}_n\left(\tilde{\boldsymbol{\theta}}\right)=0$. Επομένως, η προσέγγιση της σχέσης (3.16) γίνεται:

$$n^{-1}l_n(\boldsymbol{\theta})\approx n^{-1}l_n\left(\tilde{\boldsymbol{\theta}}\right)+\frac{1}{2}\left(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\right)^{\mathrm{T}}\left\{\frac{1}{n}\ddot{l}_n\left(\tilde{\boldsymbol{\theta}}\right)\right\}\left(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\right), \tag{3.17}$$

Αγνοώντας τη σταθερά $l_n\left(\tilde{\boldsymbol{\theta}}\right)$ και το συντελεστή $\frac{1}{2}$, η αντικειμενική συνάρτηση της σχέσης (3.17) μπορεί να απλοποιηθεί ακόμα περισσότερο:

$$n^{-1}l_n(\boldsymbol{\theta}) \approx \left(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\right)^{\mathrm{T}} \left\{\frac{1}{n}\ddot{l}_n\left(\tilde{\boldsymbol{\theta}}\right)\right\}\left(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\right), \tag{3.18}$$

Ακόμα, είναι εύλογο πως $E\left(n^{-1}\ddot{l}_n\left(\tilde{\boldsymbol{\theta}}\right)\right) \approx \Sigma^{-1}$, με αποτέλεσμα η ποσότητα $\hat{\Sigma}^{-1} = n^{-1}\ddot{l}_n\left(\tilde{\boldsymbol{\theta}}\right)$ να είναι μία πολύ καλή εκτιμήτρια για την Σ^{-1}. Σαν αποτέλεσμα, η συνάρτηση ελαχίστων τετραγώνων:

$$\left(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\right)' \hat{\Sigma}^{-1}\left(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\right) \tag{3.19}$$

είναι μία απλή προσέγγιση της αρχικής συνάρτησης $n^{-1}l_n\left(\boldsymbol{\theta}\right)$. Θα αναφερόμαστε στη σχέση (3.19) ως την προσέγγιση ελαχίστων τετραγώνων (LSA) για το πρόβλημα. Με τον LSA ορισμένο από τη σχέση (3.19), το αρχικό LASSO πρόβλημα της σχέσης (3.15) μετατρέπεται στο ασυμπτωτικά ισοδύναμο πρόβλημα ελαχίστων τετραγώνων:

$$Q(\boldsymbol{\theta}) = \left(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\right)' \hat{\Sigma}^{-1}\left(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}}\right) + \sum_{j=1}^{d}\lambda_j\left|\theta_j\right|, \tag{3.20}$$

με καθολική εκτιμήτρια την ποσότητα $\hat{\boldsymbol{\theta}}$, η οποία ελαχιστοποιεί το παραπάνω L_1 πρόβλημα ελαχίστων τετραγώνων της σχέσης (3.19). Η τελική μορφή απαιτεί μόνο την ύπαρξη της εκτιμήτριας $\hat{\Sigma}$ ενός συνεπή πίνακα διασποράς-συνδιασποράς, απαίτηση που ικανοποιείται για τα περισσότερα μοντέλα παλινδρόμησης.

Τα γενικευμένα γραμμικά μοντέλα είναι μία από τις ειδικές περιπτώσεις, οι οποίες μπορούν να εισαχθούν σε αυτό το θεωρητικό πλαίσιο κατευθείαν. Συνεπώς, ο αλγόριθμος LSA είναι άμεσα εφαρμόσιμος και στην περίπτωση του FHTR μοντέλου με σταθερό x_0, η οποία περιγράφτηκε στην Παράγραφο 2.4 . Ο LSA αλγόριθμος έχει ως μοναδικό προαπαιτούμενο τη γνώση του διανύσματος της εκτιμήτριας μέγιστης πιθανοφάνειας και της τιμής του αντίστοιχου πίνακα διασποράς-συνδιασποράς για την εκτιμήτρια αυτή. Τα δύο αυτά προαπαιτούμενα μπορούν να προκύψουν εύκολα με τη χρησιμοποίηση έτοιμων αλγορίθμων σχετικά με την προσαρμογή ενός GLM και υπάρχουν στην πλειοψηφία των στατιστικών πακέτων.

Για τη γενική περίπτωση ενός FHTR μοντέλου η δυσκολία προκύπτει, επειδή σε αντίθεση με το γενικό γραμμικό μοντέλο, το FHTR μοντέλο έχει δύο παραμέτρους και αντίστοιχα δύο γραμμικές συναρτήσεις σύνδεσης, μία για την κάθε παράμετρο. Αποφασίσαμε να εφαρμόσουμε τον αλγόριθμο LSA δύο φορές, ως εξής: αρχικά αποκτούμε τις εκτιμήτριες μέγιστης πιθανοφάνειας $\hat{\boldsymbol{\beta}}$ και $\hat{\boldsymbol{\gamma}}$, όπως και τον αντίστοιχο πίνακα

διασποράς-συνδιασποράς που προκύπτει από τον Εσσιανό πίνακα του προβλήματος. Στη συνέχεια, εφαρμόζουμε τον LSA αλγόριθμο μόνο για το διάνυσμα $\boldsymbol{\beta}$. Οι συντελεστές του $\boldsymbol{\gamma}$ δε λαμβάνονται υπόψη, με αποτέλεσμα η τεχνική συρρίκνωσης να μην εφαρμόζεται σε αυτούς. Στη συνέχεια, εκτελούμε ακριβώς την ίδια διαδικασία για το διάνυσμα $\boldsymbol{\gamma}$ των συντελεστών, ώστε η συρρίκνωση να εφαρμοστεί μόνο σε αυτούς.

Η προτεινόμενη διαδικασία αντιμετωπίζει τα δύο μέρη της δομής παλινδρόμησης της σχέσης (3.3) ξεχωριστά. Αυτό μπορεί να θεωρηθεί υπερ-απλούστευση της διαδικασίας, ειδικά κάτω από το πρίσμα των προτάσεων πως η εμφάνιση της ίδιας μεταβλητής και στα δύο μέρη της δομής υπονοεί κάποιο βαθμό έλλειψης διακριτικής ικανότητας του FHTR μοντέλου (Eberly et al., 2001; Lee και Whitmore, 2006). Ως εκ τούτου, μπορεί να μην είναι δυνατό να ξεχωρίσουμε τον τρόπο με τον οποίο η μεταβλητή επιδρά στις διάρκειες ζωής των μονάδων (μέσω των συντελεστών $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$). Ωστόσο, η μελέτη της προηγούμενης παραγράφου σχετικά με τις επιδόσεις της εκτιμήτριας μέγιστης πιθανοφάνειας σε δεδομένα δημιουργημένα από ένα IG FHTR μοντέλο έδειξε περιορισμένη σχέση ανάμεσα στις εκτιμήσεις των συντελεστών $\hat{\beta}_i$ και $\hat{\gamma}_i$ που αντιστοιχούν στην ίδια μεταβλητή z_i. Στην υπόλοιπη παράγραφο, εφαρμόζουμε τη διαδικασία LSA αρχικά σε προσομοιωμένα δεδομένα και στη συνέχεια σε ένα δείγμα από αληθινά δεδομένα.

3.4.3 Σχεδιασμός της μελέτης

Τα δεδομένα για τη μελέτη των προσομοιώσεων δημιουργήθηκαν από μοντέλα IG FHTR με τρείς, έξι, ή είκοσι μεταβλητές για συγκεκριμένες τιμές των συντελεστών στα διανύσματα $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$. Το μηδέν ήταν επιτρεπόμενη επιλογή για κάποιο συντελεστή. Το μέγεθος κάθε προσομοιωμένου δείγματος κυμάνθηκε από 100 έως 400, ανάλογα με το πλήθος των μεταβλητών που χρησιμοποιήθηκαν. Ψευδοτυχαίες τιμές των μεταβλητών δημιουργήθηκαν από την πολυμεταβλητή Κανονική κατανομή χρησιμοποιώντας τη ρουτίνα *rmvnorm* της βιβλιοθήκης *mvtnorm* του στατιστικού πακέτου R (Genz et al., 2011). Κάθε μεταβλητή ήταν Κανονικά κατανεμημένη με μοναδιαία διασπορά. Οι συσχετίσεις μεταξύ των μεταβλητών z_i και z_j ορίστηκαν ίσες με $0.5^{|i-j|}$. Οι μέσες τιμές επιλέχτηκαν με τέτοιο τρόπο ώστε $m < 0$ για την πλειοψηφία των μονάδων. Ο χειρισμός των περιπτώσεων όπου $m > 0$ και ο τρόπος παραγωγής των τιμών των IG παραμέτρων στο προσομοιωμένο δείγμα έγινε όπως περιγράψαμε στην παράγραφο 3.1.3.

Δεν εισαγάγαμε στη μελέτη αποκομμένες παρατηρήσεις. Όλοι οι υπολογισμοί έγιναν με τη βοήθεια του στατιστικού πακέτου R. Οι εκτιμήτριες μέγιστης πιθανοφάνειας υπολογίστηκαν με ελαχιστοποίηση της συνάρτησης του αρνητικού λογαρίθμου της πιθανοφάνειας, χρησιμοποιώντας είτε μία τεχνική τύπου *Newton-Raphson*, που βρίσκεται στη ρουτίνα *nlm*, είτε την τεχνική *simulated annealing* της μεθόδου SANN, που βρίσκεται

στη ρουτίνα *optim*. Ο αλγόριθμος LSA εφαρμόστηκε με τη βοήθεια του πακέτου lars (Hastie και Efron, 2011), όπως προτείνουν οι Wang και Leng (2007).

Πεντακόσιες επαναλήψεις προσομοιώσεων έγιναν για κάθε επιλογή μεγέθους δείγματος, αριθμού μεταβλητών και διανυσμάτων των συντελεστών $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$. Τα αποτελέσματα για κάθε επιλογή παρουσιάζονται περιληπτικά χρησιμοποιώντας τρία μέτρα: το μέσο μέγεθος του μοντέλου (AMS), το ποσοστό των φορών που αναγνωρίζεται το σωστό μοντέλο (CM), καθώς και η διάμεσος της ποσότητας "σχετικό σφάλμα του μοντέλου" (RME). Η ποσότητα MS είναι ο αριθμός των μη-μηδενικών συντελεστών που παραμένουν στο μοντέλο μετά από την εφαρμογή του αλγορίθμου LSA, χωρίς να μετράμε τις σταθερές οι οποίες δεν υπόκεινται σε συρρίκνωση. Το μοντέλο που αναγνωρίζεται από τον αλγόριθμο LSA θεωρείται σωστό, όταν συμπίπτουν όλα τα μηδενικά στις τελικές εκτιμήσεις του προσαρμοσμένου μοντέλου και στα διανύσματα των πραγματικών συντελεστών των παραμέτρων. Η ποσότητα RME (Fan και Li, 2001) συγκρίνει το σφάλμα στις τελικές εκτιμήτριες με το σφάλμα των εκτιμητριών μέγιστης πιθανοφάνειας. Δηλαδή, συγκρίνει τα σφάλματα πριν και μετά την εφαρμογή του αλγορίθμου LSA. Οι ποσότητες αυτές υπολογίστηκαν ξεχωριστά για τους συντελεστές των διανυσμάτων $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$, αλλά και για όλους τους συντελεστές μαζί. Οι σταθεροί όροι β_0 και γ_0 δε συμπεριλαμβάνονται σε αυτούς τους υπολογισμούς.

3.4.4 Αποτελέσματα για το μοντέλο IG GLM

Αρχικά, ασχοληθήκαμε με την ειδική περίπτωση που προκύπτει όταν η παράμετρος x_0 δεν εξαρτάται από μεταβλητές, με αποτέλεσμα το FHTR μοντέλο να μπορεί να εκφραστεί ως ένα GLM. Από τη σχέση (3.3):

$$\mu \propto |m|^{-1} = |\mathbf{z}'\beta|^{-1}$$

Ένα τέτοιο αποτέλεσμα συνεπάγεται ότι το μοντέλο FHTR είναι στην πραγματικότητα ισοδύναμο με ένα μοντέλο IG GLM, με την αντίστροφη συνάρτηση σύνδεσης.

Εξετάσαμε την εφαρμογή του αλγόριθμού LSA σε δεδομένα δημιουργημένα από έξι μεταβλητές, με διάνυσμα συντελεστών $\boldsymbol{\beta} = (1.5, -0.75, 0, 1.5, -0.75, 0)$. Για το υπόλοιπο της παραγράφου, δε θα παραθέτουμε τις τιμές των σταθερών όρων κατά την παρουσίαση των διανυσμάτων των συντελεστών $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$. Αποτελέσματα παρουσιάζονται στον Πίνακα 3.12 για ένα εύρος των τιμών της παραμέτρου $x_0 = \exp(\gamma_0)$. Εκτός από τους δείκτες AMS, CM και RME παρουσιάζουμε ξεχωριστά για κάθε μία μεταβλητή και το ποσοστό των φορών που η εκτίμηση του συντελεστή είναι μη-μηδενικός αριθμός μετά την εφαρμογή του αλγορίθμου LSA.

	Μη μηδενικοί συντελεστές μετά την εφαρμογή του αλγορίθμου LSA								
	β_1	β_2	β_3	β_4	β_5	β_6	**AMS**	**CM**	**RME**
Αληθινή τιμή	**1.5**	**-0.75**	**0**	**1.5**	**-0.75**	**0**	**4**		
n **= 100**	**%**	**%**	**%**	**%**	**%**	**%**		**%**	
$\gamma_0 = 3$	100	100	4.6	100	100	6.2	4.11	90.0	1.04
2	100	100	25.8	100	100	5.6	4.31	70.2	0.99
1	100	99.8	11.2	100	98.8	7.4	4.17	81.4	0.97
0	100	96.0	11.6	100	88.8	12.4	4.09	66.6	0.97
-1	98.4	46.6	10.6	94.8	58.4	12.6	3.21	25.8	1.18
-2	50.8	12.8	7.8	43.6	14.4	7.8	1.37	1.6	1.23
-3	17.2	3.8	2.4	14.6	4.2	3.2	0.45	0.0	0.74
n **= 200**									
$\gamma_0 = 3$	100	100	4.8	100	100	9.6	4.14	87.2	1.01
2	100	100	2.0	100	100	3.6	4.06	94.8	0.84
1	100	100	4.2	100	100	6.2	4.10	90.2	0.82
0	100	98.6	6.4	100	98.8	8.0	4.12	84.4	0.91
-1	100	87.4	7.8	100	88.2	7.4	3.91	66.6	0.93
-2	78.0	29.4	5.0	79.4	30.6	6.2	2.29	11.2	1.58
-3	41.2	5.4	3.8	28.4	9.2	3.2	0.91	0.2	1.17

Πίνακας 3.12: Αποτελέσματα για την εκτίμηση ενός IG GLM με έξι μεταβλητές, $m = z'\beta$ με σταθερό x_0, από πεντακόσιες επαναλήψεις για κάθε αληθινή τιμή του σταθερού όρου γ_0.

Η αποτελεσματικότητα του LSA στην ορθή αναγνώριση των μηδενικών συντελεστών είναι εμφανής, όπως ακριβώς προκύπτει και από τα αποτελέσματα των Wang και Leng (2007). Επίσης, είναι αξιοπρόσεκτο το γεγονός πως η πιθανότητα αναγνώρισης μη-μηδενικών επιδράσεων εξαρτάται από την τιμή του σταθερού όρου γ_0. Η πιθανότητα αυτή μειώνεται, καθώς η τιμή του γ_0 μειώνεται. Το αποτέλεσμα αυτό υποδεικνύει πως όσο το x_0 ελαττώνεται, η αντίστοιχη ισχύς της αναγνώρισης των μη-μηδενικών συντελεστών επίσης ελαττώνεται. Σαν αποτέλεσμα, οι επιδράσεις θα πρέπει να είναι ισχυρότερες, ώστε να μπορούν να εντοπιστούν από ένα μικρό x_0. Τέλος, όπως είναι αναμενόμενο τα παραπάνω αποτελέσματα είναι περισσότερο εμφανή για μικρότερο μέγεθος δείγματος.

3.4.5 Αποτελέσματα για σταθερό *m*

Στη συνέχεια, παρουσιάζουμε την ειδική περίπτωση όπου οι διάφορες μονάδες έχουν την ίδια κλίση προς το κατώφλι, αλλά με ένα διαφορετικό αρχικό επίπεδο εκκίνησης της ανέλιξης. Το μοντέλο αυτό δεν είναι πλέον GLM. Αποτελέσματα των προσομοιώσεων από

την εφαρμογή του αλγόριθμου LSA σε δεδομένα δημιουργημένα υπό το μοντέλο αυτό παρουσιάζονται στον Πίνακα 3.13. Ο τρόπος της εξάρτησης των παραμέτρων του μοντέλου από μεταβλητές δίνεται από το διάνυσμα $\gamma = (0.6, -0.3, 0, 0.6, -0.3, 0)$. Όπως και πριν, χρησιμοποιούνται διάφορες τιμές της ποσότητας $m = \beta_0$.

	Μη μηδενικοί συντελεστές μετά την εφαρμογή του αλγορίθμου LSA								
	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	**AMS**	**CM**	**RME**
Αληθινή τιμή	**0.6**	**-0.3**	**0**	**0.6**	**-0.3**	**0**	**4**		
***n* = 100**	**%**	**%**	**%**	**%**	**%**	**%**		**%**	
β_0 **= –0.1**	100	96.0	11.6	100	95.6	13.4	4.17	73.4	0.85
-0.25	100	90.6	13.4	100	91.2	10.2	4.05	65.8	0.91
-0.5	100	98.4	11.0	100	96.2	12.6	4.18	75.0	0.91
-1	100	96.4	13.6	100	98.8	14.8	4.24	71.2	0.91
-1.5	100	99.6	4.0	100	99.8	12.4	4.21	79.8	0.89
-2	100	99.6	8.6	100	100	8.2	4.16	83.8	0.86
-2.5	100	100	6.8	100	100	9.2	4.16	84.8	0.88
-5	100	100	5.6	100	100	7.4	4.13	87.8	0.83
-10	100	100	4.6	100	100	5.2	4.10	90.2	0.75
***n* = 200**									
β_0 **= –0.1**	100	99.8	6.4	100	99.0	8.2	4.13	85.0	0.86
-0.25	100	100	5.2	100	99.6	8.6	4.13	86.4	0.84
-0.5	100	99.6	5.0	100	100	5.0	4.10	89.6	0.77
-1	100	100	3.2	100	100	3.8	4.07	93.0	0.78
-1.5	100	100	6.2	100	100	5.6	4.12	89.2	0.86
-2	100	100	3.4	100	100	6.0	4.09	90.8	0.80
-2.5	100	100	3.2	100	100	4.0	4.07	93.0	0.78
-5	100	100	3.8	100	100	2.4	4.06	94.2	0.77
-10	100	100	2.0	100	100	3.6	4.06	94.8	0.74

Πίνακας 3.13: Αποτελέσματα για την εκτίμηση ενός μοντέλου με έξι μεταβλητές, $\ln(x_0) = z'\gamma$ με σταθερό *m*, από πεντακόσιες επαναλήψεις για κάθε αληθινή τιμή του σταθερού όρου β_0.

Για ακόμα μία φορά είναι ξεκάθαρη η αποτελεσματικότητα του LSA στην ορθή αναγνώριση των μη-μηδενικών συντελεστών. Οι τιμές της ποσότητας RME κυμαίνονται ανάμεσα σε 0.75 και 0.9. Υπάρχει μία μικρή τάση να υπερεκτιμηθεί το μέγεθος του δείγματος, όπως φαίνεται και στις πρώτες γραμμές του Πίνακα 3.12. Για μέγεθος δείγματος παρατηρήσεων $n = 200$, σε σχέση με μέγεθος $n = 100$, η ακρίβεια της ποσότητας AMS βελτιώνεται ελάχιστα. Ωστόσο, το ποσοστό των ορθών μοντέλων αυξάνεται απότομα.

Σύμφωνα με τις παραπάνω δύο περιπτώσεις που εξετάσαμε, μεγαλύτερες τιμές της παραμέτρου $|m|$ οδηγούν σε καλύτερα αποτελέσματα. Αρχικά, φαίνεται παράξενο ότι μεγαλύτερες τιμές για το $|m|$ (Πίνακας 3.12) και μικρότερες τιμές για το x_0 (Πίνακας 3.13) έχουν αντίθετες επιδράσεις στην ισχύ της διαδικασίας, καθώς η μέση τιμή της IG κατανομής είναι $E(T)=\frac{x_0}{|m|}$. Επομένως, αυξάνοντας το $|m|$ και μειώνοντας το x_0, είναι δράσεις οι οποίες αναμένεται να επηρεάζουν τη μέση τιμή με τον ίδιο τρόπο. Ωστόσο, η διασπορά δίνεται από τον τύπο $V(T)=\frac{x_0}{|m|^3}$, και συμπεριφέρεται διαφορετικά από τη μέση τιμή στις δύο περιπτώσεις. Όσο η τιμή του x_0 μειώνεται, καθώς το $|m|$ παραμένει σταθερό, η διασπορά μειώνεται με ρυθμό ίδιο με αυτόν της μέσης τιμής και ο συντελεστής μεταβλητότητας της κατανομής αυξάνεται, με αποτέλεσμα να είναι δυσκολότερο να εντοπιστούν συστηματικές επιδράσεις των μεταβλητών ενάντια σε αυτήν την αυξημένη μεταβλητότητα. Αντιθέτως, καθώς η τιμή του $|m|$ αυξάνεται κρατώντας ταυτόχρονα σταθερό το x_0, η $V(T)$ μειώνεται πιο απότομα από την $E(T)$ και ο συντελεστής μεταβλητότητας της κατανομής μειώνεται, με αποτέλεσμα να είναι ευκολότερο να εντοπιστούν εξαρτήσεις των παραμέτρων από τις μεταβλητές.

3.4.6 Αποτελέσματα για το γενικό μοντέλο IG FHTR

Για τη γενική περίπτωση ενός FHTR μοντέλου όπου υπάρχουν δύο παράμετροι και δύο γραμμικές συναρτήσεις σύνδεσης, εφαρμόσαμε τον αλγόριθμο LSA δύο φορές. Αρχικά, θεωρούμε ένα IG FHTR μοντέλο με τρεις μεταβλητές. Το διάνυσμα $\boldsymbol{\beta}$ με τους συντελεστές των μεταβλητών της παραμέτρου *m* στο σωστό μοντέλο σε όλες τις περιπτώσεις ήταν ίσο με $(2,-0.8,0)$. Το διάνυσμα $\boldsymbol{\gamma}$ αποτελείται από ένα θετικό, έναν αρνητικό και ένα μηδενικό συντελεστή σε όλες τις περιπτώσεις. Τα διανύσματα με τους συντελεστές των μεταβλητών που χρησιμοποιήσαμε για την παράμετρο x_0 είναι Α $(0.35,-0.25,0)$, Β$(-0.25,0.35,0)$, Γ $(0,-0.25,0.35)$ και Δ $(0,0.35,-0.25)$.

Αποτελέσματα των προσομοιώσεων δίνονται στον Πίνακα 3.14 για τέσσερις διαφορετικές επιλογές του διανύσματος $\boldsymbol{\gamma}$. Στις περιπτώσεις Α και Β οι θέσεις των μηδενικών στα διανύσματα $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$ συμπίπτουν, ενώ στις περιπτώσεις Γ και Δ βρίσκονται σε διαφορετικές θέσεις. Τα γραμμοσκιασμένα κελιά αντιστοιχούν στις περιπτώσεις με αληθινούς μηδενικούς συντελεστές.

Τα αποτελέσματα μεταξύ των τεσσάρων περιπτώσεων δεν είναι ευθέως συγκρίσιμα, καθώς δεν υπάρχει κάποια ξεκάθαρη ερμηνεία για τις τιμές των συντελεστών (Eberly et al,

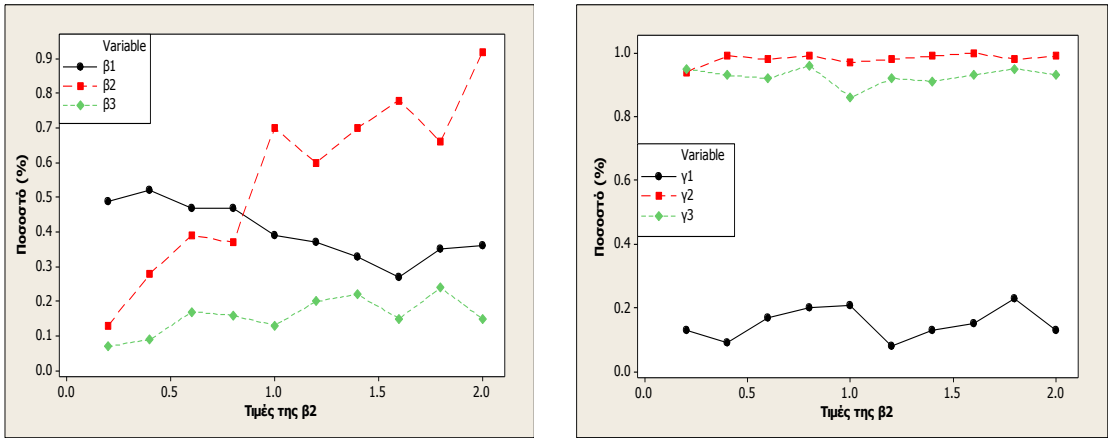
2001). Ωστόσο, τα ποσοστά των φορών που ένας αρχικά μηδενικός συντελεστής του αληθινού μοντέλου αναγνωρίζεται ορθά από τον αλγόριθμο LSA είναι πολύ κοντά μεταξύ των $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$. Επίσης, είναι πολύ κοντά μεταξύ των τεσσάρων διαφορετικών αληθινών μοντέλων. Τα μέσα μεγέθη των μοντέλων είναι ελαφρώς υπερεκτιμημένα και για τα δύο διανύσματα $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ ανεξαρτήτως μεγέθους δείγματος. Ωστόσο, με την αύξηση του μεγέθους του δείγματος από $n=100$ σε $n=200$ αυξάνεται το ποσοστό των μοντέλων που αναγνωρίζονται ορθά από τον LSA. Τα σχετικά σφάλματα των μοντέλων (δεν παρουσιάζονται στον πίνακα) κυμαίνονται από 0.9 έως 1.0, με τα περισσότερα να βρίσκονται κοντά στο 0.95.

	$n=100$				$n=200$			
	Α	**Β**	**Γ**	**Δ**	**Α**	**Β**	**Γ**	**Δ**
β_1 **(%)**	100	99.8	100	100	100	100	100	100
β_2 **(%)**	86.8	86.0	83.6	81.6	100	97.8	97.8	92.8
β_3 **(%)**	21.2	27.4	18.0	27.2	14.8	17.8	12.8	14.4
γ_1 **(%)**	100	94.4	15.6	25.8	100	100	13.2	10.2
γ_2 **(%)**	91.0	99.2	93.8	99.4	99.6	100	100	100
γ_3 **(%)**	20.4	21.8	100	95.6	10.4	17.6	100	100
AMS (%)	4.19	4.29	4.11	4.30	4.25	4.33	4.24	4.17
AMS - β **(%)**	2.08	2.13	2.02	2.09	2.15	2.16	2.11	2.07
AMS - γ **(%)**	2.11	2.15	2.09	2.21	2.10	2.18	2.13	2.10
CM (%)	54.6	50.8	56.0	46.0	79.6	72.4	74.6	73.0
CM - β **(%)**	70.2	62.2	69.0	61.8	85.2	81.0	85.2	78.8
CM - γ **(%)**	74.6	73.0	79.4	71.2	89.6	82.4	86.8	89.8

Πίνακας 3.14: Αποτελέσματα για την εκτίμηση ενός FHTR μοντέλου με τρεις μεταβλητές, με διάνυσμα $\boldsymbol{\beta}=(2,-0.8,0)$. Οι συντελεστές των μεταβλητών στο $\boldsymbol{\gamma}$ είναι Α $(0.35,-0.25,0)$, Β $(-0.25,0.35,0)$, Γ $(0,-0.25,0.35)$ και Δ $(0,0.35,-0.25)$. Τα γραμμοσκιασμένα κελιά αντιστοιχούν στις περιπτώσεις με αληθινούς μηδενικούς συντελεστές.

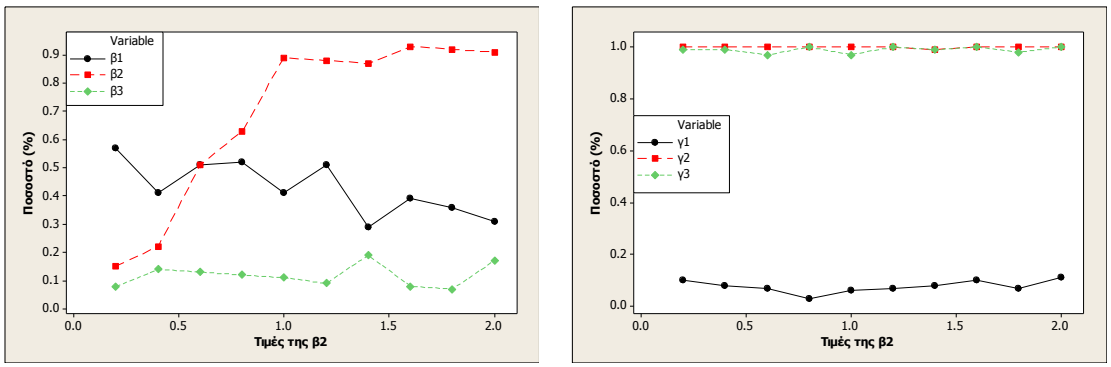
Στη συνέχεια, εξετάσαμε τον τρόπο με τον οποίο τα διάφορα αποτελέσματα εξαρτώνται από το μέγεθος ενός συντελεστή. Θεωρήσαμε μία επιλογή διανυσμάτων $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$, στην οποία οι συντελεστές $\beta_1, \beta_2, \gamma_2$ και γ_3 ήταν μη-μηδενικοί, οι συντελεστές β_3 και γ_1 ήταν ίσοι με το μηδέν και μεταβάλλαμε τις τιμές των συντελεστών β_2 ή γ_2 του αληθινού μοντέλου. Διαγράμματα με τις μεταβολές του ποσοστού των φορών που η εκτίμηση κάθε συντελεστή είναι μη-μηδενική παρουσιάζονται στο Σχήμα 3.12 για τις μεταβολές του β_2 και

στο Σχήμα 3.14 για τις μεταβολές του γ_2. Τα αποτελέσματα αυτά αφορούν μέγεθος δείγματος $n=100$.



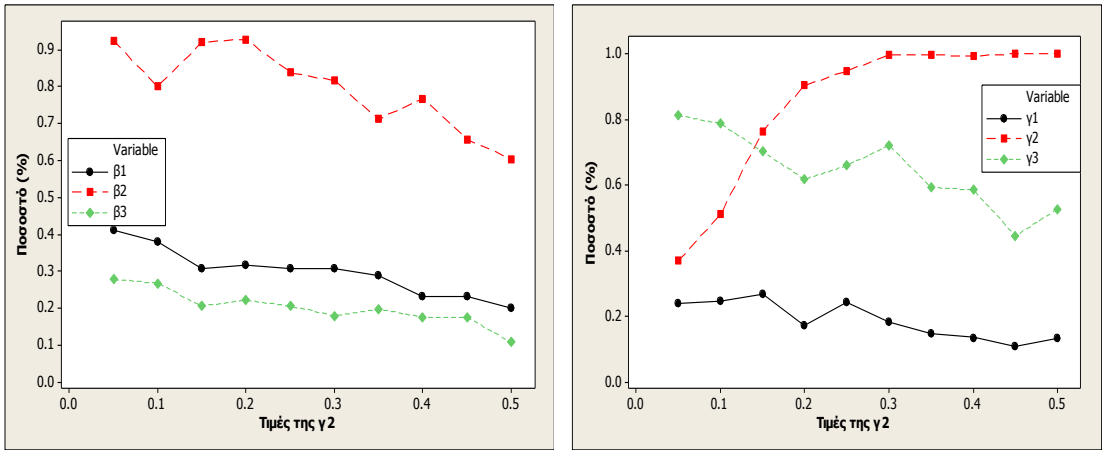
Σχήμα 3.12: Ποσοστό μη-μηδενικών εκτιμητών των συντελεστών των μεταβλητών σε σχέση με τις αληθινές τιμές του συντελεστή β2. Οι συντελεστές της μέσης τιμής στα αριστερά, οι συντελεστές του σημείου εκκίνησης της ανέλιξης στα δεξιά.

Από το Σχήμα 3.12 φαίνεται πως τα αποτελέσματα για το διάνυσμα $\boldsymbol{\gamma}$ δεν επηρεάζονται από τις τιμές του συντελεστή β_2. Ωστόσο, υπάρχει μία ξεκάθαρη τάση του συντελεστή β_1 να είναι μη - μηδενικός όλο και πιο συχνά, καθώς μειώνεται η τιμή του β_2. Δηλαδή, φαίνεται πως η διαδικασία της εκτίμησης εμφανίζει μία τάση να σχετίζει την επίδραση με τη "λάθος" μεταβλητή. Βέβαια, το αποτέλεσμα αυτό συνδέεται και με την ιδιαίτερα μεγάλη συσχέτιση (0.5) των μεταβλητών z_1 και z_2. Οι ποσότητες AMS και τα ποσοστά αναγνώρισης του σωστού μοντέλου για το $\boldsymbol{\gamma}$ δε διέφεραν σημαντικά για τις διάφορες τιμές του συντελεστή β_2 (τα αποτελέσματα δεν παρουσιάζονται). Ωστόσο, όπως ήταν αναμενόμενο, οι αντίστοιχες ενδείξεις για το διάνυσμα $\boldsymbol{\beta}$ μειώθηκαν σημαντικά με την εξασθένηση του β_2. Αντίστοιχα αποτελέσματα για τις μεταβολές του β_2 και για μέγεθος δείγματος $n=200$ παρουσιάζονται γραφικά στο Σχήμα 3.13.



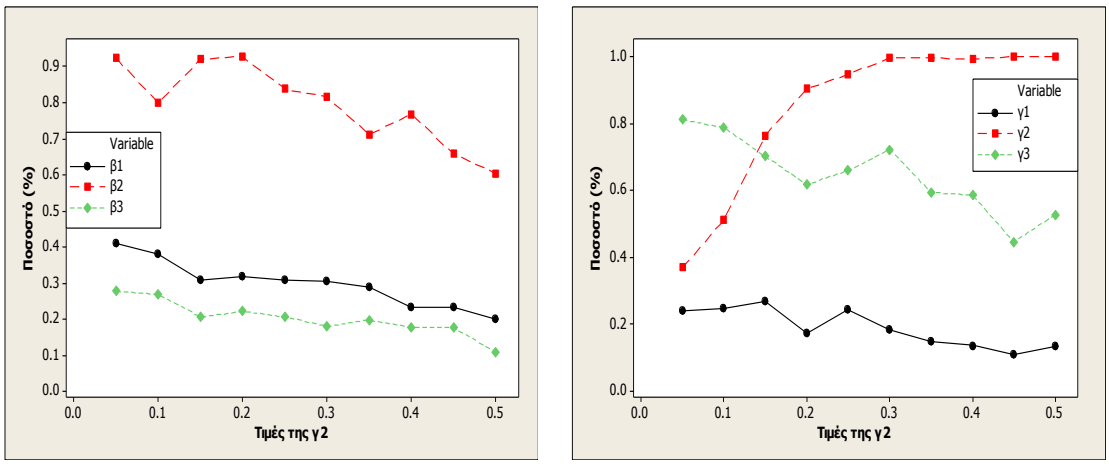
Σχήμα 3.13: Ποσοστό μη-μηδενικών εκτιμητών των συντελεστών των μεταβλητών σε σχέση με τις αληθινές τιμές του συντελεστή γ2. Οι συντελεστές του σημείου εκκίνησης της ανέλιξης στα αριστερά, οι συντελεστές της μέσης τιμής στα δεξιά.

Το Σχήμα 3.14 δείχνει αντίστοιχα πως η εκτίμηση του συντελεστή γ_3 έχει μία τάση να είναι μη-μηδενική πιο συχνά, καθώς μειώνεται η τιμή του γ_2. Ωστόσο, στην περίπτωση αυτή υπάρχει επίσης και κάποια επίδραση στις εκτιμήσεις των συντελεστών του διανύσματος $\boldsymbol{\beta}$.



Σχήμα 3.14: Ποσοστό μη-μηδενικών εκτιμητών των συντελεστών των μεταβλητών σε σχέση με τις αληθινές τιμές του συντελεστή γ2. Οι συντελεστές του σημείου εκκίνησης της ανέλιξης στα αριστερά, οι συντελεστές της μέσης τιμής στα δεξιά.

Οι συντελεστές αυτοί είναι μη-μηδενικοί πιο συχνά, καθώς μειώνεται το γ_2. Η ποσότητα AMS για το διάνυσμα $\boldsymbol{\beta}$ αυξήθηκε για μικρότερες τιμές του γ_2. Ωστόσο, δε συνέβη το ίδιο για το ποσοστό των ορθά αναγνωρισμένων μοντέλων. Αντίστοιχα αποτελέσματα για τις μεταβολές του γ_2 και για μέγεθος δείγματος $n=200$ δίνονται στο Σχήμα 3.15.



Σχήμα 3.15: Ποσοστό μη-μηδενικών εκτιμητών των συντελεστών των μεταβλητών σε σχέση με τις αληθινές τιμές του συντελεστή γ2. Οι συντελεστές του σημείου εκκίνησης της ανέλιξης στα αριστερά, οι συντελεστές της μέσης τιμής στα δεξιά.

Για την περίπτωση των έξι μεταβλητών, το αληθινό διάνυσμα των συντελεστών των μεταβλητών της παραμέτρου *m* ήταν σε όλες τις προσομοιώσεις ίσο με $(2, 1.5, -0.5, -0.7, 0, 0)$. Το διάνυσμα $\boldsymbol{\gamma}$ έχει δύο θετικούς, δύο αρνητικούς και δύο μηδενικούς συντελεστές.

Τα αποτελέσματα παρουσιάζονται στον Πίνακα 3.15 για τέσσερα διαφορετικά διανύσματα γ. Τα μηδενικά στα διανύσματα β και γ συμπίπτουν στην περίπτωση Ε με διάνυσμα γ το $(2,1.5,-0.5,-0.7,0,0)$ και στην περίπτωση Z με διάνυσμα $(-0.3,-0.2,0.35,0.25,0,0)$. Μηδενικά προκύπτουν για διαφορετικές μεταβλητές στις περιπτώσεις ΣΤ $(0,0,0.3,0,-0.2,0.35,-0.25)$ και Η $(0.3,-0.2,0,0,0.35,-0.25)$.

	$n=100$				$n=200$			
	E	**Z**	**ΣΤ**	**H**	**E**	**Z**	**ΣΤ**	**H**
β_1 **(%)**	100	100	100	100	100	100	100	100
β_2 **(%)**	99.8	97.2	93.8	97.8	99.8	99.4	99.0	100
β_3 **(%)**	55.6	59.8	43.2	25.6	60.0	82.4	72.8	44.6
β_4 **(%)**	93.0	74.0	93.6	42.4	60.6	64.4	89.2	85.0
β_5 **(%)**	34.8	27.4	30.8	47.0	23.8	37.2	21.4	42.8
β_6 **(%)**	49.6	35.0	28.8	36.8	27.8	23.6	26.0	28.8
γ_1 **(%)**	99.4	99.6	21.8	99.6	100	100	18.4	100
γ_2 **(%)**	98.8	78.2	22.4	81.6	98.4	97.8	23.0	97.3
γ_3 **(%)**	99.2	99.8	96.4	28.2	99.6	100	100	21.0
γ_4 **(%)**	95.6	99.0	80.4	22.0	99.8	100	99.0	20.0
γ_5 **(%)**	26.0	20.6	99.8	99.6	11.4	19.4	100	100
γ_6 **(%)**	23.6	30.0	97.2	97.0	10.4	17.8	100	100
AMS (%)	8.74	8.21	8.08	7.78	7.92	8.42	8.49	8.39
AMS - β (%)	4.32	3.93	3.90	3.50	3.72	40.7	4.08	4.01
AMS - γ (%)	4.43	4.27	4.18	4.28	4.20	4.35	4.40	4.38
CM (%)	8.8	12.4	9.2	0.8	13.4	23.6	24.8	10.4
CM - β (%)	14.2	17.6	16.6	3.4	14.4	27.2	36.4	13.8
CM - γ (%)	55.0	43.8	48.0	47.6	79.4	66.0	62.8	63.6

Πίνακας 3.15: Αποτελέσματα για την εκτίμηση ενός FHTR μοντέλου με έξι μεταβλητές, με διάνυσμα $\beta=(2,1.5,-0.5,-0.7,0,0)$. Τα γραμμοσκιασμένα κελιά αντιστοιχούν στις περιπτώσεις με αληθινούς μηδενικούς συντελεστές.

Τα αποτελέσματα δείχνουν πως η ποσότητα AMS δεν είναι πάντοτε υπερεκτιμημένη. Τα ποσοστά των ορθά αναγνωρισμένων μοντέλων είναι πολύ πιο χαμηλά από την περίπτωση

του Πίνακα 3.14. Ένα τέτοιο αποτέλεσμα όμως είναι αναμενόμενο, καθώς υπάρχουν περισσότερες μεταβλητές στη μελέτη. Οι ποσότητες RME ήταν πολύ κοντά με αυτές για την περίπτωση των τριών μεταβλητών.

	Εκτίμηση του β			**Εκτίμηση του γ**		
		Μη μηδενικές εκτιμήσεις (%)			**Μη μηδενικές εκτιμήσεις (%)**	
Μεταβλητή	**Αληθινό β**	**$n=300$**	**$n=400$**	**Αληθινό γ**	**$n=300$**	**$n=400$**
1	0	8.0	5.8	0.5	100	100
2	0	8.8	15.4	0.2	100	99.8
3	0	34.8	40.4	-0.35	100	100
4	2	95.6	92.6	-0.2	98.2	99.8
5	1.5	62.2	30.8	0	11.0	14.6
6	-1.5	9.8	52.0	0	28.8	18.4
7	0	10.8	2.4	0	16.2	8.8
8	0	1.0	29.4	0	14.4	14.8
9	-1	19.6	10.0	0.5	100	100
10	0	21.8	13.2	0.2	99.4	99.8
11	0	12.0	2.8	-0.35	100	100
12	0	5.8	22.8	-0.2	99.4	100
13	2	85.0	79.8	0	29.6	30.4
14	1.5	44.0	70.2	0	15.4	18.8
15	-1.5	37.2	27.4	0	36.6	28.2
16	-1	17.2	46.6	0	24.0	13.4
17	0	4.8	6.0	0.5	100	100
18	2	94.0	93.2	0.2	100	100
19	1.5	46.6	93.6	0	13.0	9.6
20	0	27.2	39.0	0	19.4	26.0
	AMS-β	6.26	7.99	**AMS-γ**	12.1	11.8
	RME-β	0.729	0.761	**RME-γ**	0.717	0.734

Πίνακας 3.16: Αποτελέσματα για την εκτίμηση ενός μοντέλου με είκοσι μεταβλητές σε πεντακόσιες επαναλήψεις.

Τέλος, ο Πίνακας 3.16 παρουσιάζει τα αποτελέσματα για την περίπτωση ενός μοντέλου με ένα πολύ μεγάλο αριθμό μεταβλητών, $p=20$ για μεγέθη δειγμάτων $n=300$ και $n=400$. Οι επιλογές των συντελεστών στα διανύσματα $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$ περιλαμβάνουν δέκα

μηδενικούς συντελεστές σε κάθε διάνυσμα. Η ποσότητα AMS για το διάνυσμα $\boldsymbol{\beta}$ δεν αναγνωρίζει σωστά αυτόν τον αριθμό, ωστόσο έρχεται πιο κοντά στην αληθινή τιμή, καθώς αυξάνεται το μέγεθος του δείγματος. Για τη συγκεκριμένη επιλογή δεδομένων τα αποτελέσματα για το διάνυσμα $\boldsymbol{\beta}$ είναι λιγότερο ξεκάθαρα από τα αντίστοιχα αποτελέσματα για το διάνυσμα $\boldsymbol{\gamma}$, ωστόσο το ποσοστό των ορθών μοντέλων είναι μηδενικό και για τα δύο διανύσματα.

Τα αποτελέσματα αυτά πρέπει να ερμηνευτούν υπό το πρίσμα των συσχετίσεων που υπάρχουν ανάμεσα στις μεταβλητές. Για παράδειγμα, το ιδιαίτερα υψηλό ποσοστό των μη - μηδενικών εκτιμήσεων του συντελεστή β_{20} μπορεί να οφείλεται στη μεγάλη συσχέτιση (0.5) της μεταβλητής z_{20} με τη μεταβλητή z_{19}, η οποία έχει μη μηδενικό συντελεστή στο αληθινό μοντέλο $\beta_{19} = 1.5$.

3.4.7 Εφαρμογή σε πραγματικά δεδομένα

Ως συνέχεια των προσομοιώσεων, παραθέτουμε και την εφαρμογή της προτεινόμενης μεθόδου επιλογής μεταβλητών σε ένα σύνολο δεδομένων που αφορούν μία κλινική δοκιμή ενός ναρκωτικού για την αντιμετώπιση και θεραπεία μίας ασθένειας, η οποία αφορά πρώιμη κύρωση της χοληδόχου κύστης (PBC). Τα δεδομένα υπάρχουν στο πακέτο *survival* του στατιστικού υπολογιστικού πακέτου R (Therneau, 2011). 17 μεταβλητές είναι διαθέσιμες για τη μελέτη, συμπεριλαμβανομένης και μίας ψευδομεταβλητής που υποδεικνύει την χορηγούμενη αγωγή που λαμβάνει κάθε ασθενής. Επειδή το πείραμα ήταν τυχαιοποιημένο, το αρχικό σημείο εκκίνησης της ανέλιξης, x_0, δε μπορεί να συνδεθεί με την αγωγή (Pennell et al., 2010), με αποτέλεσμα ο συντελεστής του διανύσματος $\boldsymbol{\gamma}$ που αντιστοιχεί στην ακολουθούμενη αγωγή, να τεθεί εκ των προτέρων ίσος με το μηδέν. Όλες οι υπόλοιπες μεταβλητές επιτρέπεται να εισέλθουν και στις δύο γραμμικές εκτιμήτριες των παραμέτρων. Προηγούμενες αναλύσεις πάνω στο συγκεκριμένο σύνολο δεδομένων που υπάρχουν στη βιβλιογραφία έδειξαν πως αρκετές μεταβλητές παραμένουν στο μοντέλο μετά την εφαρμογή διαφόρων μεθόδων επιλογής μεταβλητών (Hoeting et al., 1999; Tibshirani, 1997; Wang και Leng, 2007; Zhu και Fan, 2011).

Ο Πίνακας 3.17 παρουσιάζει τις εκτιμήτριες μέγιστης πιθανοφάνειας των συντελεστών των μεταβλητών και τις εκτιμήτριες, όπως αυτές προκύπτουν ύστερα από την εφαρμογή του LSA αλγορίθμου. Δώδεκα από τους δεκαέξι συντελεστές των μεταβλητών στο διάνυσμα $\boldsymbol{\gamma}$ παραμένουν μη-μηδενικοί και μετά την εφαρμογή του αλγορίθμου, υποδεικνύοντας τη σχέση τους με το αρχικό σημείο εκκίνησης της ανέλιξης, δηλαδή με την κατάσταση του ασθενή στην αρχή του πειράματος. Το αποτέλεσμα αυτό φαίνεται αρκετά λογικό, καθώς αρχικά οι μεταβλητές αυτές συμπεριλήφθηκαν στη μελέτη λόγω της πεποίθησης που υπήρχε πως σχετίζονται με την ασθένεια. Επιπρόσθετα, μόνο τέσσερις μεταβλητές έχουν μη-μηδενικούς

συντελεστές στο διάνυσμα $\boldsymbol{\beta}$, υποδεικνύοντας πως σχετίζονται με την εξέλιξη της κατάστασης της υγείας των ασθενών μετά από την είσοδό τους στο πείραμα.

	Εκτιμήτριες μέγιστης πιθανοφάνειας		**Εκτιμήτριες μετά από την εφαρμογή του LSA αλγορίθμου**	
Μεταβλητή	$\ln x_0$	m	$\ln x_0$	m
Αγωγή		-0.046 (0.076)		0
Ηλικία	-0.002 (0.007)	-0.013 (0.007)	0	-0.005
Φύλο	-0.198 (0.218)	- 0.259 (0.194)	-0.131	0
Ασκίτης	0.004 (0.290)	-0.192 (0.284)	0	0
Ηπατομεγαλία	0.197 (0.164)	-0.207 (0.164)	0.175	0
Spiders	-0.271 (0.143)	0.095 (0.140)	-0.226	0
Όιδημα	-0.878 (0.303)	0.203 (0.248)	-0.845	0
Χολερυθρίνη	0.015 (0.105)	-0.192 (0.101)	0	-0.088
Χοληστερόλη	0.545 (0.164)	-0.580 (0.213)	0.503	-0.536
Λευκωματίνη	-1.114 (0.546)	1.434 (0.599)	-1.048	0.950
Ουροδόχος κύστη	-0.248 (0.115)	0.078 (0.109)	-0.231	0
Αλκαλική φωσφατάση	0.105 (0.110)	-0.095 (0.093)	0.054	0
Ast (SGOT)	-0.245 (0.189)	0.043 (0.165)	-0.160	0
Τριγλυκερίδια	-0.318 (0.181)	0.249 (0.163)	-0.223	0
Αιμοπετάλια	-0.135 (0.191)	-0.005 (0.172)	0	0
Protime	-1.539 (0.896)	0.217 (0.817)	-1.227	0
Στάδιο	-0.309 (0.109)	0.151 (0.109)	-0.308	0

Πίνακας 3.17: Αποτελέσματα της εφαρμογής του αλγορίθμου LSA σε δεδομένα πρώιμης κύρωσης της χοληδόχου κύστης. n = 276. Η εξαρτημένη μεταβλητή είναι η επιβίωση σε χρόνια. Οι τιμές των μεταβλητών Bilirubin έως και Protime έχουν μετατραπεί σε λογαρίθμους.

3.4.9 Συμπεράσματα

Η διπλή εφαρμογή του αλγορίθμου LSA για την εκτέλεση της προσαρμοσμένης LASSO τεχνικής φαίνεται να δουλεύει αποτελεσματικά στην επιλογή των κατάλληλων μεταβλητών για την περίληψή τους στους δύο γραμμικούς εκτιμητές του FHTR μοντέλου. Το αποτέλεσμα αυτό συναντά μία ανάγκη, καθώς στη μεθοδολογία της FHTR μοντελοποίησης δεν υπάρχουν προς το παρόν τεχνικές ανάπτυξης μοντέλου. Άλλες διαδικασίες, όπως για παράδειγμα διαγνωστικοί έλεγχοι καταλληλότητας του FHTR μοντέλου, θα έπρεπε επίσης να αναπτυχθούν, όπως γίνεται και με άλλα ευρέως διαδεδομένα μοντέλα. Παραδείγματα αποτελούν τα μοντέλα αναλογικής διακινδύνευσης και επιταχυνόμενης ζωής για την ανάλυση δεδομένων διάρκειας ζωής, τα οποία έχουν δοκιμαστεί και χρησιμοποιηθεί ευρύτατα.

Κεφάλαιο 4

Διαγνωστικοί έλεγχοι για το FHTR μοντέλο

4.1 Εισαγωγή

Όπως είδαμε και στα προηγούμενα κεφάλαια της διατριβής, τα μοντέλα χρόνου πρώτης μετάβασης γενικότερα και ειδικά το μοντέλο εκείνο, στο οποίο ο χρόνος πρώτης μετάβασης ακολουθεί την αντίστροφη Γκαουσιανή κατανομή, μας προσφέρουν συχνά μία ελκυστική αναπαράσταση δεδομένων διάρκειας ζωής. Μετά τη μοντελοποίηση του προβλήματος, μεγάλο ρόλο για την Ανάλυση Επιβίωσης κατέχει η αναγνώριση των άτυπων τιμών, με τις οποίες ασχοληθήκαμε στο δεύτερο κεφάλαιο της διατριβής, καθώς και ο εντοπισμός σημείων επιρροής (influential observations). Με την έννοια επιρροή, εννοούμε την επίδραση του κάθε σημείου του δείγματος στην προσαρμογή του μοντέλου.

Στη βιβλιογραφία, υπάρχει ένας μεγάλος αριθμός διαφορετικών τεχνικών για τον εντοπισμό σημείων επιρροής (Cook και Weisberg, 1982; Therneau και Grambsch, 2000). Ωστόσο, δεν υπάρχει κάποια μέθοδος για την περίπτωση ενός IG FHTR μοντέλου. Σκοπός του κεφαλαίου, είναι να αναπτύξουμε και να παρουσιάσουμε διαγνωστικούς ελέγχους για το συγκεκριμένο μοντέλο.

Στη συνέχεια της παραγράφου, θα παρουσιάσουμε την τεχνική αφαίρεσης σημείου CDM (Case Deletion Method), για την περίπτωση ενός FHTR μοντέλου στο οποίο οι διάρκειες ζωής ακολουθούν την IG κατανομή, προκειμένου να μετρηθεί η επιρροή της καθεμιάς παρατήρησης. Η μέθοδος CDM είναι μία τεχνική με μεγάλη ευκολία εφαρμογής. Ένας τρόπος να μετρήσουμε την επιρροή της i-οστής παρατήρησης στις εκτιμήσεις των συντελεστών της παλινδρόμησης είναι να κατασκευάσουμε διαγνωστικούς ελέγχους για την

περίπτωση του μοντέλου από το οποίο απουσιάζει η *i*-οστή παρατήρηση. Επιπρόσθετα, χρησιμοποιούμε την ιδέα αυτή προκειμένου να μετατρέψουμε την απόσταση των πιθανοφανειών (likelihood distance) του μοντέλου με όλες τις παρατηρήσεις και εκείνου χωρίς την *i*-οστή παρατήρηση, σε χρήσιμο εργαλείο εντοπισμού των περιπτώσεων όπου μόνο μία από τις δύο παραμέτρους της κατανομής είναι ειδικού ενδιαφέροντος.

Ακόμα, η αξιολόγηση της επιρροής μικρών διαταραχών (local influence) ενός μοντέλου αποτελεί ένα ιδιαίτερα χρήσιμο εργαλείο. Η τεχνική του Cook (1986) για τη μέτρηση της τοπικής επιρροής επεκτείνεται και προσαρμόζεται κατάλληλα για το μοντέλο παλινδρόμησης IG FHTR.

Βασική ιδιότητα του μοντέλου της παλινδρόμησης Κατωφλιού αποτελεί το γεγονός πως και οι δύο παράμετροι του μοντέλου εξαρτώνται από συμμεταβλητές, σε αντίθεση με άλλα μοντέλα παλινδρόμησης ή και με τα γενικευμένα γραμμικά μοντέλα, στα οποία συνήθως μόνο μία παράμετρος εξαρτάται από μεταβλητές. Για το FHTR μοντέλο, δηλαδή, έχουμε δύο γραμμικές προβλέπουσες, μία για το x_0 και μία για το *m*. Ωστόσο, μπορεί να μην είναι δυνατό να διακρίνουμε επιρροές στα $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$. Η βασική ερώτηση που μας απασχολεί και ερευνήσαμε και στο προηγούμενο κεφάλαιο, είναι κατά πόσο η επίδραση της ίδιας μεταβλητής και στις δύο παραμέτρους μπορεί να διαχωριστεί στην πραγματικότητα, ή αν υπάρχει ενδεχόμενος βαθμός έλλειψης διακριτικής ικανότητας στο μοντέλο (Eberly et al., 2001; Lee & Whitmore, 2006). Η μελέτη του προηγούμενου κεφαλαίου υποδεικνύει πως αυτό δεν αποτελεί ιδιαίτερα μεγάλο πρόβλημα όταν το FHT είναι το κατάλληλο μοντέλο για την προσαρμογή. Έτσι λοιπόν, έχει ενδιαφέρον να μελετήσουμε μέτρα επιρροής, προκειμένου να εξετάσουμε ξεχωριστά την επιρροή των εκτιμητριών των παραμέτρων x_0 και *m*.

Στα προηγούμενα κεφάλαια της διατριβής, έχουμε παρουσιάσει αναλυτικά το FHTR μοντέλο για την περίπτωση όπου οι διάρκειες ζωής ακολουθούν την IG. Η πιο συχνά χρησιμοποιούμενη μορφή της σ.π.π. μιας τυχαίας μεταβλητής που ακολουθεί την IG με παραμέτρους μ και λ δίνεται από τη σχέση:

$$f\left(t;\mu,\lambda\right)=\frac{\sqrt{\lambda}}{\sqrt{2\pi t^3}}\exp\left(-\frac{\lambda\left(t-\mu\right)^2}{2\mu^2 t}\right),\quad t>0,\,\mu,\lambda>0 \tag{4.1}$$

Όμοια, η συνηθισμένη παραμέτρηση της σ.π.π. της IG, όταν προκύπτει από ένα FHTR μοντέλο βασισμένο σε μια ανέλιξη Wiener δίνεται από τη σχέση:

$$f\left(t;m,\sigma^2,x_0\right)=\frac{x_0}{\sqrt{2\pi\sigma^2 t^3}}\exp\left(-\frac{\left(x_0+mt\right)^2}{2\sigma^2 t}\right),\quad t>0,-\infty<m<+\infty,\sigma^2>0,x_0>0\,, \tag{4.2}$$

όπου οι παράμετροι *m* και x_0 αντιστοιχούν σε αυτές της σχέσης (4.1) με τη βοήθεια της σχέσης:

$$\lambda = x_0^2 \ , \ \mu = \frac{x_0}{|m|} \tag{4.3}$$

Στο υπόλοιπο κεφάλαιο, παρουσιάζουμε τη μέθοδο μέγιστης πιθανοφάνειας για το IG FHTR μοντέλο, χωρίς αποκομμένα δεδομένα. Ακολουθεί η ανάπτυξη διαγνωστικών ελέγχων βασισμένων στην τεχνική CDM και δίνονται οι αντίστοιχες προσεγγίσεις βήματος των εκτιμητών των συντελεστών του μοντέλου. Μελετάται η τοπική επιρροή. Τα διάφορα αποτελέσματα δίνονται μέσα από προσομοιώσεις, αλλά και από τη μελέτη ενός πραγματικού συνόλου δεδομένων. Τέλος, ακολουθούν συμπεράσματα και σχόλια.

4.2 Η απόσταση των πιθανοφανειών (Likelihood Distance) και η γενικευμένη απόσταση του Cook (Generalized Cook's Distance)

Σε αυτήν την παράγραφο, παρουσιάζουμε τη μέθοδο της απόστασης των πιθανοφανειών και τη γενικευμένη απόσταση του Cook, δύο πολύ γνωστές μεθόδους για τον εντοπισμό σημείων επιρροής. Στη συνέχεια, τις προσαρμόζουμε για την περίπτωση ενός IG FHTR μοντέλου.

4.2.1 Εισαγωγή στην τεχνική αφαίρεσης σημείου (CDM)

Προσπαθώντας να δώσουμε έναν ορισμό για την έννοια της επιρροής της *i*-οστής παρατήρησης στην εκτίμηση μέγιστης πιθανοφάνειας, $\hat{\boldsymbol{\theta}}$, μπορούμε να χρησιμοποιήσουμε την ποσότητα $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)}$, όπου $\hat{\boldsymbol{\theta}}_{(i)}$ είναι το αποτέλεσμα της προσαρμογής του μοντέλου που περιλαμβάνει όλα τα σημεία του δείγματος εκτός από την παρατήρηση *i*. Εάν η *i*-οστή παρατήρηση δεν έχει μεγάλη επιρροή στο προσαρμοσμένο δείγμα, τότε η εκτίμηση μέγιστης πιθανοφάνειας όλου του δείγματος δε θα πρέπει να έχει πολύ διαφορετική τιμή από την εκτίμηση μέγιστης πιθανοφάνειας του δείγματος χωρίς την *i*-οστή παρατήρηση, δηλαδή θα πρέπει να ισχύει $\hat{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}}_{(i)}$ (ισοδύναμα $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)} \approx \mathbf{0}$).

Προσεγγίζοντας την έννοια της επιρροής, οι Cook και Weisberg (1982) εξηγούν πως μέτρα της επιρροής της *i*-οστής παρατήρησης μπορεί να υπολογιστούν εύκολα και άμεσα βασισμένα στην ποσότητα $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{(i)}$. Ωστόσο, μία τέτοια διαδικασία ενδεχομένως να κοστίζει σε υπολογιστικό χρόνο, με αποτέλεσμα να είναι δύσκολη στην εφαρμογή της, καθώς απαιτείται ο υπολογισμός $(n+1)$ εκτιμητριών μέγιστης πιθανοφάνειας, καθεμιά από τις οποίες μπορεί να απαιτεί επαναλήψεις κατά τη διαδικασία του αλγορίθμου βελτιστοποίησης. Να παρατηρήσουμε πως τα τελευταία χρόνια η επιστήμη των υπολογιστών έχει προοδεύσει σημαντικά στην κατεύθυνση της βελτιστοποίησης των υπολογιστικών διαδικασιών, με αποτέλεσμα ο υπολογιστικός χρόνος να μην αποτελεί πλέον σημαντική παράμετρο σε μία

έρευνα για την εύρεση σημείων επιρροής σε ένα δείγμα δεδομένων. Ωστόσο, η συγκεκριμένη παράμετρος δεν πρέπει να αγνοηθεί σε περιπτώσεις μεγάλων δειγμάτων.

Η εκτιμήτρια μέγιστης πιθανοφάνειας μπορεί να συμβολιστεί με $\hat{\boldsymbol{\theta}}$ και προκύπτει ως λύση της εξίσωσης:

$$u(\boldsymbol{\theta}) = \dot{l}(\boldsymbol{\theta}) = \frac{d\, l(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = \boldsymbol{\theta}, \tag{4.4}$$

όπου $l(\cdot)$ είναι η συνάρτηση του λογαρίθμου της πιθανοφάνειας, για την οποία ισχύει ότι $l(\boldsymbol{\theta}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}) = l_{(i)}(\boldsymbol{\theta}) + l_i(\boldsymbol{\theta})$, με $l_{(i)}(\boldsymbol{\theta}) = \sum_{j \neq i} l_j(\boldsymbol{\theta})$.

Για το υπόλοιπο της παραγράφου, οι ποσότητες $l(\boldsymbol{\theta}), l_i(\boldsymbol{\theta}), l_{(i)}(\boldsymbol{\theta})$ συμβολίζουν το λογάριθμο της πιθανοφάνειας ολόκληρου του δείγματος, του μοντέλου που έχει μόνο την *i*-οστή παρατήρηση και του μοντέλου χωρίς αυτή, αντίστοιχα.

Στόχος μας είναι να βρούμε την ποσότητα $\hat{\boldsymbol{\theta}}_{(i)}$, η οποία μεγιστοποιεί το λογάριθμο της πιθανοφάνειας του δείγματος χωρίς την *i*-οστή παρατήρηση, λύνοντας την εξίσωση $u_{(i)}\left(\boldsymbol{\theta}_{(i)}\right) = \boldsymbol{\theta}$, όπου:

$$u_{(i)}(\boldsymbol{\theta}) = \dot{l}_{(i)}(\boldsymbol{\theta}) = \frac{d\, l_{(i)}(\boldsymbol{\theta})}{d\boldsymbol{\theta}}.$$

Παρατήρηση: $u(\boldsymbol{\theta})$ είναι η συνάρτηση σκορ του δείγματος και $H(\boldsymbol{\theta})$ ο αντίστοιχος Εσσιανός πίνακας.

Από τη συνηθισμένη ανάπτυξη σε σειρά Taylor γύρω από το σημείο $\hat{\boldsymbol{\theta}}$, προκύπτει:

$$u_{(i)}(\boldsymbol{\theta}) = u_{(i)}\left(\hat{\boldsymbol{\theta}}\right) + H_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right) + \ldots, \tag{4.5}$$

όπου $H_{(i),jk}(\boldsymbol{\theta}) = \dfrac{\partial^2 l_{(i)}(\boldsymbol{\theta})}{\partial\theta_j \partial\theta_k}$.

Όταν αντιμετώπισαν το πρόβλημα του υπολογιστικού χρόνου, οι Cook και Weisberg πρότειναν την εναλλακτική θεώρηση μιας τετραγωνικής προσέγγισης του λογαρίθμου της πιθανοφάνειας του μοντέλου που προκύπτει ύστερα από τη διαγραφή της *i*-οστής παρατήρησης:

$$l_{(i)}(\boldsymbol{\theta}) \approx l_{(i)}\left(\hat{\boldsymbol{\theta}}\right) + \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^T \dot{l}_{(i)}\left(\hat{\boldsymbol{\theta}}\right) + \frac{1}{2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^T \ddot{l}_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right), \tag{4.6}$$

όπου $\dot{l}_{(i)}\left(\hat{\boldsymbol{\theta}}\right)$ είναι το διάνυσμα της κλίσης με το *j*-οστό στοιχείο $\frac{\partial l_{(i)}(\boldsymbol{\theta})}{\partial \theta_j}$ υπολογισμένο στο σημείο $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}$. Επίσης, η ποσότητα $\ddot{l}_{(i)}\left(\hat{\boldsymbol{\theta}}\right)$ έχει ως (j,k)-οστό στοιχείο το $\frac{\partial^2 l_{(i)}(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}$, υπολογισμένο στο σημείο $\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}$.

Αποδεικνύεται (Kennedy και Gentle, 1980, Κεφάλαιο 10) πως εάν η ποσότητα $-\ddot{l}_{(i)}\left(\hat{\boldsymbol{\theta}}\right)$ είναι θετικά ορισμένη, τότε η τετραγωνική προσέγγιση μεγιστοποιείται στη θέση:

$$\hat{\boldsymbol{\theta}}_{(i)}^{1}=\hat{\boldsymbol{\theta}}-\left(\ddot{l}_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)^{-1}\dot{l}_{(i)}\left(\hat{\boldsymbol{\theta}}\right), \tag{4.7}$$

όπου $\hat{\boldsymbol{\theta}}_{(i)}^{1}$ είναι η προσέγγιση βήματος της ποσότητας $\boldsymbol{\theta}_{(i)}$, καθώς είναι η ίδια με αυτή στην οποία θα καταλήγαμε ύστερα από μία μοναδική επανάληψη της μεθόδου Newton-Raphson χρησιμοποιώντας το διάνυσμα $\hat{\boldsymbol{\theta}}$ ως σημείο εκκίνησης για τη μεγιστοποίηση. Όταν η *i*-οστή παρατήρηση αφορά σημείο επιρροής, η διαφορά $\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_{(i)}^{1}$ αναμένεται να είναι αρκετά μεγαλύτερη από τις αντίστοιχες διαφορές των υπόλοιπων σημείων του δείγματος.

Στην περίπτωση του μοντέλου FHTR, μπορούμε να χρησιμοποιήσουμε απευθείας την ποσότητα $\hat{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_{(i)}$, ως εργαλείο για τη μέτρηση της επιρροής της *i*-οστής παρατήρησης στην εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\boldsymbol{\theta}}=\left(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\gamma}}\right)$.

Εάν αντικαταστήσουμε το $\boldsymbol{\theta}$ με $\hat{\boldsymbol{\theta}}_{(i)}$ στη σχέση (4.5), καταλήγουμε στην αντίστοιχη διαδικασία τύπου Newton Raphson:

$$\boldsymbol{0}\approx u_{(i)}\left(\hat{\boldsymbol{\theta}}\right)+H_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\left(\hat{\boldsymbol{\theta}}_{(i)}-\hat{\boldsymbol{\theta}}\right) \tag{4.8}$$

Η παραπάνω εξίσωση γράφεται και στην εξής μορφή:

$$\hat{\boldsymbol{\theta}}_{(i)}-\hat{\boldsymbol{\theta}}\approx-\left(H_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)^{-1}u_{(i)}\left(\hat{\boldsymbol{\theta}}\right) \tag{4.9}$$

Τελικά, έχουμε:

$$\hat{\boldsymbol{\theta}}_{(i)}\approx\hat{\boldsymbol{\theta}}-\left(H_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)^{-1}u_{(i)}\left(\hat{\boldsymbol{\theta}}\right), \tag{4.10}$$

όπου

$$u_{(i)}\left(\hat{\boldsymbol{\theta}}\right)=\left.\frac{\partial l_{(i)}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\hat{\theta}}=\left.\frac{\partial\left(l(\boldsymbol{\theta})-l_i(\boldsymbol{\theta})\right)}{\partial\boldsymbol{\theta}}\right|_{\hat{\theta}}=-\left.\frac{\partial l_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right|_{\hat{\theta}} \tag{4.11}$$

και

$$H_{(i)jk}\left(\hat{\boldsymbol{\theta}}\right)=\left.\frac{\partial^2 l_{(i)}(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\right|_{\hat{\theta}}=\left.\frac{\partial^2\left(l(\boldsymbol{\theta})-l_i(\boldsymbol{\theta})\right)}{\partial\theta_j\partial\theta_k}\right|_{\hat{\theta}}=\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\right|_{\hat{\theta}}-\left.\frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\right|_{\hat{\theta}}=H_{jk}\left(\hat{\boldsymbol{\theta}}\right)-H_{i,jk}\left(\hat{\boldsymbol{\theta}}\right) \tag{4.12}$$

4.2.2 Υπολογισμός της απόστασης των πιθανοφανειών (Likelihood Distance)

Μπορούμε να ορίσουμε την απόσταση των πιθανοφανειών LD_i, ως εξής:

$$LD_i=2\left[l\left(\hat{\boldsymbol{\theta}}\right)-l\left(\hat{\boldsymbol{\theta}}_{(i)}\right)\right] \tag{4.13}$$

Ή αντίστοιχα, χρησιμοποιώντας την εκτιμήτρια που προκύπτει από τη διαδικασία ενός βήματος, που συναντάται στους Cook και Weisberg,

$$LD_i^1=2\left[l\left(\hat{\boldsymbol{\theta}}\right)-l\left(\hat{\boldsymbol{\theta}}_{(i)}^1\right)\right] \tag{4.14}$$

Εναλλακτικά, τα παραπάνω μέτρα μπορεί να ερμηνευτούν σε όρους μιας ασυμπτωτικής περιοχής εμπιστοσύνης (Cox και Hinkley, 1974)

$$\left\{\boldsymbol{\theta}:2\left[l\left(\hat{\boldsymbol{\theta}}\right)-l(\boldsymbol{\theta})\right]\le\chi^2(\alpha,q)\right\}, \tag{4.15}$$

όπου $\chi^2(\alpha,q)$ είναι το άνω α ποσοστιαίο σημείο της χ^2 κατανομής με q βαθμούς ελευθερίας, όπου q είναι η διάσταση του διανύσματος $\boldsymbol{\theta}$. Σαν αποτέλεσμα, μπορούμε να μετρήσουμε την ποσότητα LD_i συγκρίνοντάς την με την κατανομή $\chi^2(q)$.

Η ποσότητα LD_i μπορεί να **προσεγγιστεί** από την ανάπτυξη σε σειρά Taylor της ποσότητας $l\left(\hat{\boldsymbol{\theta}}_{(i)}\right)$ γύρω από το $\hat{\boldsymbol{\theta}}$:

$$l\left(\hat{\boldsymbol{\theta}}_{(i)}\right)\approx l\left(\hat{\boldsymbol{\theta}}\right)+\left(\boldsymbol{\theta}_{(i)}-\hat{\boldsymbol{\theta}}\right)^T\dot{l}\left(\hat{\boldsymbol{\theta}}\right)+\frac{1}{2}\left(\boldsymbol{\theta}_{(i)}-\hat{\boldsymbol{\theta}}\right)^T\ddot{l}\left(\hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}_{(i)}-\hat{\boldsymbol{\theta}}\right)+\ldots$$

Επειδή $\dot{l}_{(i)}\left(\hat{\boldsymbol{\theta}}\right)=\boldsymbol{0}$, ύστερα από κάποιες απλοποιήσεις σταθερών ποσοτήτων, παίρνουμε από την παραπάνω εξίσωση:

$$LD_i\approx\left(\hat{\boldsymbol{\theta}}_{(i)}-\hat{\boldsymbol{\theta}}\right)^T\left(-H\left(\hat{\boldsymbol{\theta}}\right)\right)\left(\hat{\boldsymbol{\theta}}_{(i)}-\hat{\boldsymbol{\theta}}\right) \tag{4.16}$$

Στη συνέχεια, θα ασχοληθούμε με το γεγονός πως για το FHTR μοντέλο, έχουμε δύο γραμμικές προβλέπουσες, μία για το x_0 και μία για το m. Θα προσπαθήσουμε να διακρίνουμε επιρροές στα $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$. Για το λόγο αυτό, θα μελετήσουμε μέτρα επιρροής προκειμένου να εξετάσουμε ξεχωριστά την επιρροή των εκτιμήσεων των παραμέτρων x_0 και m.

Η απόσταση των πιθανοφανειών μπορεί εύκολα να μετατραπεί, ώστε να συμπεριλάβει περιπτώσεις στις οποίες απαιτείται να μελετηθεί μόνο ένα υποσύνολο $\boldsymbol{\theta}_1$ του διανύσματος $\boldsymbol{\theta}$, των παραμέτρων του προβλήματος. Ας θεωρήσουμε $\boldsymbol{\theta}^T = \left(\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T\right)$ και $\hat{\boldsymbol{\theta}}_{(i)}^T = \left(\hat{\boldsymbol{\theta}}_{1(i)}^T, \hat{\boldsymbol{\theta}}_{2(i)}^T\right)$. Μία ασυμπτωτική περιοχή για το $\boldsymbol{\theta}_1$ δίνεται από:

$$\left\{\boldsymbol{\theta}_1 : 2\left[l\left(\hat{\boldsymbol{\theta}}\right) - l\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\left(\boldsymbol{\theta}_1\right)\right)\right] \le \chi^2\left(\alpha, q_1\right)\right\},$$

όπου q_1 είναι η διάσταση του $\boldsymbol{\theta}_1$ και η ποσότητα:

$$l\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\left(\boldsymbol{\theta}_1\right)\right) = \max_{\boldsymbol{\theta}_2}\left[l\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right)\right]$$

είναι ο λογάριθμος της πιθανοφάνειας μεγιστοποιημένος στο χώρο των τιμών της παραμέτρου $\boldsymbol{\theta}_2$ όταν $\boldsymbol{\theta}_1$ είναι σταθερό (Cox και Hinkley, 1974). Αντίστοιχα, η απόσταση των πιθανοφανειών της εκτιμήτριας $\hat{\boldsymbol{\theta}}_1$ όταν η *i*-οστή παρατήρηση έχει διαγραφεί, δηλαδή η $\hat{\boldsymbol{\theta}}_{1(i)}$, γίνεται:

$$LD_i\left(\boldsymbol{\theta}_1 \mid \boldsymbol{\theta}_2\right) = 2\left[l\left(\hat{\boldsymbol{\theta}}\right) - l\left(\hat{\boldsymbol{\theta}}_{1(i)}, \boldsymbol{\theta}_2\left(\hat{\boldsymbol{\theta}}_{1(i)}\right)\right)\right] = 2\left[l\left(\hat{\boldsymbol{\theta}}\right) - \max_{\boldsymbol{\theta}_2}\left[l\left(\hat{\boldsymbol{\theta}}_{1(i)}, \boldsymbol{\theta}_2\right)\right]\right], \qquad (4.17)$$

με ένα παρόμοιο μέτρο να αποκτάται στην περίπτωση που αντικαταστήσουμε τον εκτιμητή του ενός βήματος με εκείνον που προκύπτει από ολόκληρη την επαναληπτική διαδικασία βελτιστοποίησης. Στη συνέχεια, το μέτρο αυτό συγκρίνεται με την $\chi^2\left(q_1\right)$ κατανομή για λόγους βαθμονόμησης.

Ας θεωρήσουμε την περίπτωση ενός FHTR μοντέλου, όπου οι παράμετροι του ενδιαφέροντος είναι οι x_0 και *m*, καθεμία από τις οποίες προκύπτει από διαφορετική συνάρτηση σύνδεσης, όπως δείξαμε στο προηγούμενο κεφάλαιο. Θα μελετήσουμε την περίπτωση ενός δείγματος *n* μη-αποκομμένων παρατηρήσεων από την IG, δηλαδή: $t_1, t_2, ..., t_n \sim IG\left(x_0, m, \sigma^2 = 1\right)$.

Το διάνυσμα των παραμέτρων είναι:

$$\boldsymbol{\theta}^T = \left(\theta_1, \theta_2\right) = \left(x_0, m\right)$$

Και το αντίστοιχο διάνυσμα των εκτιμητριών των παραμέτρων για το μοντέλο χωρίς την *i*-οστή παρατήρηση:

$$\hat{\boldsymbol{\theta}}_{(i)}^T = \left(\hat{\theta}_{1(i)}, \hat{\theta}_{2(i)}\right) = \left(\hat{x}_{0(i)}, \hat{m}_{(i)}\right)$$

Υπολογίζοντας την απόσταση των πιθανοφανειών, όταν η παράμετρος του ενδιαφέροντος είναι η x_0

Το μέτρο που μας παρέχει η ασυμπτωτική περιοχή για την απόσταση των πιθανοφανειών για την παράμετρο x_0 γίνεται:

$$\left\{x_0 : 2\left[l\left(\hat{x}_0, \hat{m}\right) - l\left(x_0, m\left(x_0\right)\right)\right] \le \chi^2\left(\alpha, q_1\right)\right\} \quad \Leftrightarrow$$

$$LD_i\left(x_0 \mid m\right) = 2\left[l\left(\hat{x}_0, \hat{m}\right) - l\left(\hat{x}_0, m\left(\hat{x}_0\right)\right)\right] = 2\left[l\left(\hat{x}_0, \hat{m}\right) - \max_m\left[l\left(\hat{x}_0, m\right)\right]\right] \qquad (4.18)$$

Η αντίστοιχη ασυμπτωτική περιοχή εμπιστοσύνης για την απόσταση των πιθανοφανειών για την ποσότητα $\hat{x}_{0(i)}$ (δηλαδή για την περίπτωση του μοντέλου χωρίς την *i*-οστή παρατήρηση) γίνεται:

$$\left\{\hat{x}_{0(i)} : 2\left[l\left(\hat{x}_0, \hat{m}\right) - l\left(\hat{x}_{0(i)}, m\left(\hat{x}_{0(i)}\right)\right)\right] \le \chi^2\left(\alpha, q_1\right)\right\},$$

ή σε μία διαφορετική μορφή,

$$LD_i\left(x_{0(i)} \mid m\right) = 2\left[l\left(\hat{x}_0, \hat{m}\right) - l\left(\hat{x}_{0(i)}, m\left(\hat{x}_{0(i)}\right)\right)\right] = 2\left[l\left(\hat{x}_0, \hat{m}\right) - \max_m\left[l\left(\hat{x}_{0(i)}, m\right)\right]\right] \qquad (4.19)$$

Συνεπώς, χρειάζεται να αποκτήσουμε τις εκτιμήτριες μέγιστης πιθανοφάνειας $\hat{x}_0, \hat{m}$, τις αντίστοιχες για το μοντέλο χωρίς την *i*-οστή παρατήρηση $\hat{x}_{0(i)}, \hat{m}_{(i)}$ και τέλος να βελτιστοποιήσουμε την πιθανοφάνεια όλου του δείγματος, υπολογισμένη στα σημεία $\left(\hat{x}_0, \hat{m}\right)$ και $\left(\hat{x}_{0(i)}, m\left(\hat{x}_{0(i)}\right)\right)$. Έτσι λοιπόν, έχουμε:

$$t_1, t_2, ..., t_n \sim IG\left(x_0, m, \sigma^2 = 1\right)$$

Από την εξίσωση (4.2), παίρνουμε:

$$f\left(t; m, x_0\right) = \frac{x_0}{\sqrt{2\pi t^3}} \exp\left(-\frac{\left(x_0 + mt\right)^2}{2t}\right), \quad t > 0, -\infty < m < +\infty, x_0 > 0$$

Βήμα 1: Υπολογισμός του λογαρίθμου της πιθανοφάνειας

$$L\left(x_0, m\right) = \prod_{i=1}^{n} f_i\left(t; m, x_0\right) \quad \Rightarrow$$

$$l\left(x_0, m\right) = \log L\left(x_0, m\right) = \log\left[\prod_{i=1}^{n}\left[\frac{x_0}{\sqrt{2\pi t_i^3}} \exp\left(-\frac{\left(x_0 + mt_i\right)^2}{2t_i}\right)\right]\right] =$$

$$= \sum_{i=1}^{n}\left\{\log\left[\frac{x_0}{\sqrt{2\pi t_i^3}}\exp\left(-\frac{\left(x_0+mt_i\right)^2}{2t_i}\right)\right]\right\} = \sum_{i=1}^{n}\left\{\log\left(x_0\right)-\log\left(2\pi t_i^3\right)^{\frac{1}{2}}-\frac{\left(x_0+mt_i\right)^2}{2t_i}\right\} =$$

$$= n\log\left(x_0\right)-\frac{n}{2}\log\left(2\pi\right)-\frac{3}{2}\sum_{i=1}^{n}\log\left(t_i\right)-\sum_{i=1}^{n}\frac{\left(x_0^2+2x_0mt_i+m^2t_i^2\right)}{2t_i} =$$

$$= n\log\left(x_0\right)-\frac{n}{2}\log\left(2\pi\right)-\frac{3}{2}\sum_{i=1}^{n}\log\left(t_i\right)-\frac{x_0^2}{2}\sum_{i=1}^{n}\frac{1}{t_i}-n\,x_0\,m-\frac{m^2}{2}\sum_{i=1}^{n}t_i \quad \Rightarrow$$

$$l\left(x_0,m\right)=n\log\left(x_0\right)-\frac{n}{2}\log\left(2\pi\right)-\frac{3}{2}\sum_{i=1}^{n}\log\left(t_i\right)-\frac{x_0^2}{2}\sum_{i=1}^{n}\frac{1}{t_i}-n\,x_0\,m-\frac{m^2}{2}\sum_{i=1}^{n}t_i \tag{4.20}$$

Βήμα 2: Υπολογισμός των ε.μ.π. ολόκληρου του δείγματος

ε.μ.π. για την παράμετρο *m*

$$\frac{\partial\, l\left(x_0,m\right)}{\partial m}=0 \quad\Leftrightarrow\quad -nx_0-\frac{2}{2}mn\bar{t}=0 \quad\Leftrightarrow$$

$$\hat{m}=-\frac{x_0}{\bar{t}}=-\frac{\hat{x}_0}{\bar{t}} \tag{4.21}$$

Ε.μ.π. για την παράμετρο x_0

Για λόγους ευκολίας θέτουμε $\log x_0 := x_0' \quad\Leftrightarrow\quad x_0=\exp\left(x_0'\right)$. Τότε,

$$l\left(x_0,m\right)=\log L\left(x_0,m\right)=nx_0'-\frac{n}{2}\log\left(2\pi\right)-\frac{3}{2}\sum_{i=1}^{n}\log\left(t_i\right)-\sum_{i=1}^{n}\frac{\left(\left(\mathrm{e}^{x_0'}\right)^2+2\mathrm{e}^{x_0'}mt_i+m^2t_i^2\right)}{2t_i}.$$

$$\frac{\partial\, l\left(x_0,m\right)}{\partial x_0}=0 \quad\Leftrightarrow\quad n-\frac{1}{2}\left[2\mathrm{e}^{2x_0'}\sum_{i=1}^{n}\frac{1}{t_i}+2\mathrm{e}^{x_0'}mn\right]=0 \quad\Leftrightarrow$$

$$n-\frac{1}{2}\left[\sum_{i=1}^{n}\frac{1}{t_i}-\mathrm{e}^{x_0'}mn\right]=0 \quad\overset{x_0\,=\,\exp\left(x_0'\right)}{\Leftrightarrow}$$

$$n-x_0^2\sum_{i=1}^{n}\frac{1}{t_i}-n\,x_0\,m=0 \quad\overset{\text{αντικαθιστούμε το } m \text{ με } \hat{m}}{\Leftrightarrow}\quad n-x_0^2\sum_{i=1}^{n}\frac{1}{t_i}+n\,x_0\frac{x_0}{\bar{t}}=0 \quad\Leftrightarrow$$

$$\hat{x}_0^2=\frac{n}{\left(\sum_{i=1}^{n}\frac{1}{t_i}-\frac{n}{\bar{t}}\right)} \tag{4.22}$$

Παρατήρηση: Επειδή έχουμε $x_0 > 0$ στο μοντέλο FHTR, πάντοτε επιλέγουμε τη θετική εκτιμήτρια της εξίσωσης (4.22), $\hat{x}_0 = +\sqrt{\dfrac{n}{\left(\sum_{i=1}^{n}\dfrac{1}{t_i} - \dfrac{n}{\bar{t}}\right)}}$.

Η εξίσωση (4.19) μας παρέχει το επιθυμητό μέτρο της επιρροής. Τα τελικά βήματα είναι να υπολογίσουμε τις ποσότητες $l(\hat{x}_0, \hat{m})$ και $l\left(\hat{x}_{0(i)}, m\left(\hat{x}_{0(i)}\right)\right)$.

Βήμα 3: Υπολογισμός της ποσότητας $l(\hat{x}_0, \hat{m})$

Εάν θέσουμε $\hat{m}(x_0) = \hat{m}$ και αντικαταστήσουμε το m με $\hat{m}$ και το x_0 με $\hat{x}_0$, παίρνουμε από την εξίσωση (4.20):

$$L(x_0, m) = \prod_{i=1}^{n} f_i(t; m, x_0) \quad \Rightarrow$$

$$l(x_0, m) = n\log(x_0) - \frac{n}{2}\log(2\pi) - \frac{3}{2}\sum_{i=1}^{n}\log(t_i) - \frac{x_0^2}{2}\sum_{i=1}^{n}\frac{1}{t_i} - n x_0 m - \frac{m^2}{2}\sum_{i=1}^{n} t_i \quad \Rightarrow$$

$$l(\hat{x}_0, \hat{m}) = n\log(\hat{x}_0) - \frac{n}{2}\log(2\pi) - \frac{3}{2}\sum_{i=1}^{n}\log(t_i) - \frac{\hat{x}_0^2}{2}\sum_{i=1}^{n}\frac{1}{t_i} - n\hat{x}_0\hat{m} - \frac{\hat{m}^2}{2}\sum_{i=1}^{n} t_i \quad \Rightarrow$$

$$l(\hat{x}_0, \hat{m}) = n\log(\hat{x}_0) - \frac{n}{2}\log(2\pi) - \frac{3}{2}\sum_{i=1}^{n}\log(t_i) - \frac{\hat{x}_0^2}{2}\sum_{i=1}^{n}\frac{1}{t_i} + \frac{n\hat{x}_0^2}{\bar{t}} - \frac{\hat{x}_0^2}{2\bar{t}^2}\sum_{i=1}^{n} t_i \quad \Rightarrow$$

$$l(\hat{x}_0, \hat{m}) = n\log\left(\sqrt{\frac{n}{\left(\sum_{i=1}^{n}\frac{1}{t_i} - \frac{n}{\bar{t}}\right)}}\right) - \frac{n}{2}\log(2\pi) - \frac{3}{2}\sum_{i=1}^{n}\log(t_i) - \frac{n}{2\left(\sum_{i=1}^{n}\frac{1}{t_i} - \frac{n}{\bar{t}}\right)}\sum_{i=1}^{n}\frac{1}{t_i} +$$

$$+\frac{n^2}{\bar{t}\left(\sum_{i=1}^{n}\frac{1}{t_i} - \frac{n}{\bar{t}}\right)} - \frac{n}{2\bar{t}^2\left(\sum_{i=1}^{n}\frac{1}{t_i} - \frac{n}{\bar{t}}\right)} n\bar{t} \quad \Rightarrow$$

Τελικά,

$$l(\hat{x}_0, \hat{m}) = \log L\left(\sqrt{\frac{n}{\left(\sum_{i=1}^{n}\frac{1}{t_i} - \frac{n}{\bar{t}}\right)}}, -\frac{\sqrt{\dfrac{n}{\left(\sum_{i=1}^{n}\dfrac{1}{t_i} - \dfrac{n}{\bar{t}}\right)}}}{\bar{t}}\right) =$$

$$n\log\left(\sqrt{\frac{n}{\left(\sum_{i=1}^{n}\frac{1}{t_i}-\frac{n}{\overline{t}}\right)}}\right)-\frac{n}{2}\log(2\pi)-\frac{3}{2}\sum_{i=1}^{n}\log(t_i)-\frac{1}{\left(\sum_{i=1}^{n}\frac{1}{t_i}-\frac{n}{\overline{t}}\right)}\left\{\frac{n}{2}\sum_{i=1}^{n}\frac{1}{t_i}-\frac{n^2}{\overline{t}}+\frac{n^2}{2\overline{t}}\right\}\Rightarrow$$

$$l(\hat{x}_0,\hat{m})=\frac{n}{2}\log n-\frac{n}{2}\log\left(\left(\sum_{i=1}^{n}\frac{1}{t_i}-\frac{n}{\overline{t}}\right)\right)-\frac{n}{2}\log(2\pi)-\frac{3}{2}\sum_{i=1}^{n}\log(t_i)-\frac{n}{2} \qquad (4.23)$$

Βήμα 4: Υπολογισμός των ε.μ.π. για τις ποσότητες $\hat{x}_{0(i)}$, $\hat{m}_{(i)}$ και $m\left(\hat{x}_{0(i)}\right)$

ε.μ.π. για τις παραμέτρους $x_{0(i)}$ και $m_{(i)}$.

Προκειμένου να υπολογίσουμε την ποσότητα $l\left(\hat{x}_{0(i)},m\left(\hat{x}_{0(i)}\right)\right)$, πρώτα πρέπει να εκτιμήσουμε τις ποσότητες $\hat{x}_{0(i)},m\left(\hat{x}_{0(i)}\right)$ με τον ίδιο ακριβώς τρόπο που εργαστήκαμε για τις ποσότητες $\hat{x}_0,\hat{m}$. Εκτός από το γεγονός ότι στο δείγμα μας υπάρχουν $(n-1)$ παρατηρήσεις αντί για *n*, τίποτα δεν αλλάζει από τη διαδικασία που ακολουθήσαμε προηγουμένως για την εκτίμηση των ποσοτήτων $x_{0(i)},m_{(i)}$. Αρχικά, υπολογίζουμε τις ε.μ.π. για τα $x_{0(i)}$ και $m_{(i)}$ για την περίπτωση του μοντέλου χωρίς την *i*-οστή παρατήρηση:

$$\hat{x}_{0(i)}^2=\frac{n-1}{\left(\sum_{i=1}^{n-1}\frac{1}{t_i}-\frac{n-1}{\overline{t}_{(i)}}\right)} \quad\Rightarrow\quad \hat{x}_{0(i)}=+\sqrt{\frac{n-1}{\left(\sum_{i=1}^{n-1}\frac{1}{t_i}-\frac{n-1}{\overline{t}_{(i)}}\right)}}, \qquad (4.24)$$

όπου $\overline{t}_{(i)}=-\frac{\sum_{i=1}^{n-1}t_i}{n-1}$.

και

$$\hat{m}_{(i)}=-\frac{\hat{x}_{0(i)}}{\overline{t}_{(i)}}=-\frac{\sqrt{\frac{n-1}{\left(\sum_{i=1}^{n-1}\frac{1}{t_i}-\frac{n-1}{\overline{t}_{(i)}}\right)}}}{\overline{t}_{(i)}}=-\frac{\sqrt{n-1}}{\overline{t}_{(i)}\sqrt{\left(\sum_{i=1}^{n-1}\frac{1}{t_i}-\frac{n-1}{\overline{t}_{(i)}}\right)}} \qquad (4.25)$$

(Υπενθύμιση: για τον υπολογισμό της ποσότητας $\hat{x}_{0(i)}$ επιλέγουμε μόνο τη θετική ρίζα).

ε.μ.π. για την ποσότητα $\hat{m}\left(\hat{x}_{0(i)}\right)$

Σημείωση: υποθέτουμε ότι $\hat{x}_{0(i)}$ είναι σταθερή. Σκοπός είναι να υπολογίσουμε την εκτιμήτρια της ποσότητας $m\left(\hat{x}_{0(i)}\right)$ από την πιθανοφάνεια με βάση ολόκληρο το δείγμα **και όχι** από την πιθανοφάνεια του δείγματος με τις $(n-1)$ παρατηρήσεις.

$$l(x_0, m) = \log L(x_0, m) = \log\left[\prod_{i=1}^{n}\left[\frac{x_0}{\sqrt{2\pi t_i^3}}\exp\left(-\frac{(x_0 + mt_i)^2}{2t_i}\right)\right]\right] =$$

$$= n\log(x_0) - \frac{n}{2}\log(2\pi) - \frac{3}{2}\sum_{i=1}^{n}\log(t_i) - \frac{x_0^2}{2}\sum_{i=1}^{n}\frac{1}{t_i} - n x_0\, m - \frac{m^2}{2}\sum_{i=1}^{n} t_i \quad \Rightarrow$$

$$l\left(\hat{x}_{0(i)}, m\left(\hat{x}_{0(i)}\right)\right) \overset{m\left(\hat{x}_{0(i)}\right)=m}{=} l\left(\hat{x}_{0(i)}, m\right) = n\log\left(\hat{x}_{0(i)}\right) - \frac{n}{2}\log(2\pi) - \frac{3}{2}\sum_{i=1}^{n}\log(t_i) -$$

$$-\frac{\hat{x}_{0(i)}^2}{2}\sum_{i=1}^{n}\frac{1}{t_i} - n\hat{x}_{0(i)}\, m - \frac{m^2}{2}\sum_{i=1}^{n} t_i \quad \Rightarrow$$

$$\frac{\partial\, l\left(\hat{x}_{0(i)}, m\right)}{\partial m} = 0 \quad \Leftrightarrow \quad -n\hat{x}_{0(i)} - \frac{2}{2} m\sum_{i=1}^{n} t_i = 0 \quad \Leftrightarrow$$

$$\hat{m}\left(\hat{x}_{0(i)}\right) = \max_m l\left(\hat{x}_{0(i)}, m\left(\hat{x}_{0(i)}\right)\right) = -\frac{n\hat{x}_{0(i)}}{\sum_{i=1}^{n} t_i} = -\frac{\hat{x}_{0(i)}}{\overline{t}} \qquad (4.26)$$

Τελικά, από τις σχέσεις (4.21), (4.22), (4.23), (4.24) και (4.26) μπορούμε να υπολογίσουμε την απόσταση των δύο πιθανοφανειών για το x_0 από τη σχέση (4.19):

Βήμα 5: Υπολογισμός της απόστασης των πιθανοφανειών για το x_0 .

$$LD_i\left(x_{0(i)} \mid m\right) = 2\left[l\left(\hat{x}_0, \hat{m}\right) - l\left(\hat{x}_{0(i)}, m\left(\hat{x}_{0(i)}\right)\right)\right] = 2\left[l\left(\hat{x}_0, \hat{m}\right) - \max_m\left[l\left(\hat{x}_{0(i)}, m\right)\right]\right] =$$

$$= 2\left[\log L\left(\sqrt{\frac{n}{\left(\sum_{i=1}^{n}\frac{1}{t_i} - \frac{n}{\overline{t}}\right)}}, -\frac{\sqrt{\left(\sum_{i=1}^{n}\frac{1}{t_i} - \frac{n}{\overline{t}}\right)}}{\overline{t}}\right) - \log L\left(\sqrt{\frac{n-1}{\left(\sum_{i=1}^{n-1}\frac{1}{t_i} - \frac{n-1}{\overline{t}_{(i)}}\right)}}, -\frac{\sqrt{\left(\sum_{i=1}^{n-1}\frac{1}{t_i} - \frac{n-1}{\overline{t}_{(i)}}\right)}}{\overline{t}}\right)\right]$$

Υπολογισμός της απόστασης των πιθανοφανειών για την παράμετρο *m*

Το αντίστοιχο μέτρο της απόστασης των πιθανοφανειών για τη $\hat{m}_{(i)}$ μέσω ασυμπτωτικής περιοχής (όταν η *i*-οστή παρατήρηση έχει διαγραφεί) γίνεται:

$$\left\{\hat{m}_{(i)} : 2\left[l\left(\hat{m}, \hat{x}_0\right) - l\left(\hat{m}_{(i)}, x_0\left(\hat{m}_{(i)}\right)\right)\right] \le \chi^2\left(\alpha, q_1\right)\right\},$$

ή σε μια άλλη μορφή,

$$LD_i\left(m_{(i)} \mid x_0\right) = 2\left[l\left(\hat{m}, \hat{x}_0\right) - l\left(\hat{m}_{(i)}, x_0\left(\hat{m}_{(i)}\right)\right)\right] \tag{4.27}$$

Εργαζόμενοι όπως και πριν, χρειαζόμαστε να αποκτήσουμε τις ε.μ.π. $\hat{x}_0$, $\hat{m}$, από ολόκληρο το δείγμα, τις αντίστοιχες για το δείγμα χωρίς την *i*-οστή παρατήρηση, $\hat{x}_{0(i)}, \hat{m}_{(i)}$. Τέλος, βελτιστοποιούμε την πιθανοφάνεια ολόκληρου του δείγματος, υπολογισμένη στα σημεία $\left(\hat{x}_0, \hat{m}\right)$ και $\left(\hat{m}_{(i)}, x_0\left(\hat{m}_{(i)}\right)\right)$.

Παρατήρηση: Οι υπολογισμοί των ποσοτήτων $\hat{x}_0$, $\hat{m}$ και $l\left(\hat{x}_0, \hat{m}\right)$ είναι ακριβώς οι ίδιοι όπως και προηγουμένως. Για το λόγο αυτό, θα παραθέσουμε μόνο τα τελικά αποτελέσματα.

Βήμα 1: Υπολογισμός της πιθανοφάνειας

$$l\left(m, x_0\right) = \log L\left(m, x_0\right) = n\log\left(x_0\right) - \frac{n}{2}\log\left(2\pi\right) - \frac{3}{2}\sum_{i=1}^{n}\log\left(t_i\right) - \frac{x_0^2}{2}\sum_{i=1}^{n}\frac{1}{t_i} - n\,x_0\,m - \frac{m^2}{2}\sum_{i=1}^{n}t_i$$

Βήμα 2: Υπολογισμός των ε.μ.π. ολόκληρου του δείγματος

ε.μ.π. για την ποσότητα x_0

$$\hat{x}_0^2 = \frac{n}{\left(\sum_{i=1}^{n}\frac{1}{t_i} - \frac{n}{\overline{t}}\right)} \quad \Leftrightarrow \quad \hat{x}_0 = +\sqrt{\frac{n}{\left(\sum_{i=1}^{n}\frac{1}{t_i} - \frac{n}{\overline{t}}\right)}}$$

ε.μ.π. για την ποσότητα *m*

$$\hat{m} = m\left(\hat{x}_0\right) = -\frac{\hat{x}_0}{\overline{t}} = -\frac{\hat{x}_0}{\overline{t}} = -\frac{\sqrt{n}}{\overline{t}\sqrt{\left(\sum_{i=1}^{n}\frac{1}{t_i} - \frac{n}{\overline{t}}\right)}}$$

Βήμα 3: Υπολογισμός της $l(\hat{x}_0, \hat{m})$

$$l(\hat{m}, \hat{x}_0) = \log L\left(-\frac{\sqrt{n}}{\bar{t}\sqrt{\left(\sum_{i=1}^{n}\frac{1}{t_i} - \frac{n}{\bar{t}}\right)}}, \quad \sqrt{\frac{n}{\left(\sum_{i=1}^{n}\frac{1}{t_i} - \frac{n}{\bar{t}}\right)}} \right) =$$

$$= \frac{n}{2}\log n - \frac{n}{2}\log\left(\left(\sum_{i=1}^{n}\frac{1}{t_i} - \frac{n}{\bar{t}}\right)\right) - \frac{n}{2}\log(2\pi) - \frac{3}{2}\sum_{i=1}^{n}\log(t_i) - \frac{n}{2}$$

Βήμα 4: Υπολογισμός των ε.μ.π. $\hat{m}_{(i)}$, $\hat{x}_{0(i)}$ και $\hat{x}_0\left(\hat{m}_{(i)}\right)$

ε.μ.π. για τις ποσότητες $m_{(i)}$ και $x_{0(i)}$

Προκειμένου να υπολογίσουμε την ποσότητα $l\left(\hat{m}_{(i)}, \hat{x}_0\left(\hat{m}_{(i)}\right)\right)$, πρέπει πρώτα να βρούμε τις εκτιμήτριες τις $\hat{m}_{(i)}$, $\hat{x}_0\left(\hat{m}_{(i)}\right)$ με τον ίδιο τρόπο που εργαστήκαμε και για τις ε.μ.π., $\hat{x}_0$, $\hat{m}$. Εκτός από το γεγονός πως υπάρχουν $(n-1)$ παρατηρήσεις στο δείγμα αντί για *n*, δεν αλλάζει τίποτα στη διαδικασία εκτίμησης των $m_{(i)}$ and $x_{0(i)}$. Αρχικά βρίσκουμε τις ε.μ.π. για τα $x_{0(i)}$ και $m_{(i)}$ για το δείγμα χωρίς την *i*-οστή παρατήρηση:

$$\hat{x}_{0(i)}^2 = \frac{n-1}{\left(\sum_{i=1}^{n-1}\frac{1}{t_i} - \frac{n-1}{\bar{t}_{(i)}}\right)} \quad \Rightarrow \quad \hat{x}_{0(i)} = +\sqrt{\frac{n-1}{\left(\sum_{i=1}^{n-1}\frac{1}{t_i} - \frac{n-1}{\bar{t}_{(i)}}\right)}},$$

όπου $\bar{t}_{(i)} = -\dfrac{\sum_{i=1}^{n-1} t_i}{n-1}$ και

$$\hat{m}_{(i)} = -\frac{\hat{x}_{0(i)}}{\bar{t}_{(i)}}$$

(Υπενθύμιση: για τον υπολογισμό της ποσότητας $\hat{x}_{0(i)}$ επιλέγουμε τη θετική ρίζα).

ε.μ.π. για την ποσότητα $x_0\left(\hat{m}_{(i)}\right)$

Σημείωση: Υποθέτουμε την ποσότητα $\hat{m}_{(i)}$ σταθερή. Στόχος είναι ο υπολογισμός της ποσότητας $\hat{x}_0\left(\hat{m}_{(i)}\right)$ από το λογάριθμο της πιθανοφάνειας ολόκληρου του δείγματος **και όχι** από το δείγμα χωρίς με τις $(n-1)$ παρατηρήσεις!

$$l(m, x_0) = \log L(m, x_0) = \log\left[\prod_{i=1}^{n}\left[\frac{x_0}{\sqrt{2\pi t_i^3}}\exp\left(-\frac{(x_0+mt_i)^2}{2t_i}\right)\right]\right] =$$

$$= n\log(x_0) - \frac{n}{2}\log(2\pi) - \frac{3}{2}\sum_{i=1}^{n}\log(t_i) - \frac{x_0^2}{2}\sum_{i=1}^{n}\frac{1}{t_i} - n\,x_0\,m - \frac{m^2}{2}\sum_{i=1}^{n}t_i \quad \Rightarrow$$

$$l\left(\hat{m}_{(i)}, x_0\left(\hat{m}_{(i)}\right)\right) \underset{\hat{m}_{(i)}\ \text{σταθερή}}{\overset{x_0\left(\hat{m}_{(i)}\right)=x_0}{=}} l\left(\hat{m}_{(i)}, x_0\right) = n\log(x_0) - \frac{n}{2}\log(2\pi) - \frac{3}{2}\sum_{i=1}^{n}\log(t_i) -$$

$$-\frac{x_0^2}{2}\sum_{i=1}^{n}\frac{1}{t_i} - n\,x_0\,\hat{m}_{(i)} - \frac{\hat{m}_{(i)}^{\,2}}{2}\sum_{i=1}^{n}t_i \quad \Rightarrow$$

$$\frac{\partial\, l\left(\hat{m}_{(i)}, x_0\left(\hat{m}_{(i)}\right)\right)}{\partial x_0} = 0 \quad \Leftrightarrow \quad \frac{n}{x_0} - \frac{2x_0}{2}\sum_{i=1}^{n}\frac{1}{t_i} - n\hat{m}_{(i)} = 0 \quad \Leftrightarrow \quad x_0^2\sum_{i=1}^{n}\frac{1}{t_i} + n\hat{m}_{(i)}x_0 - n = 0 \quad \Leftrightarrow$$

Τελικά, η διακρίνουσα του τριωνύμου δίνεται από τον τύπο:

$$D = \left(n\hat{m}_{(i)}\right)^2 - 4(-n)\sum_{i=1}^{n}\frac{1}{t_i} = \left(n\hat{m}_{(i)}\right)^2 + 4n\sum_{i=1}^{n}\frac{1}{t_i}$$

και οι ρίζες της εξίσωσης:

$$\hat{x}_{01,2} = \frac{-n\hat{m}_{(i)} \pm \sqrt{D}}{2\sum_{i=1}^{n}\frac{1}{t_i}} = 0 \quad \Leftrightarrow \quad \hat{x}_0 = \frac{-n\hat{m}_{(i)} + \sqrt{\left(n\hat{m}_{(i)}\right)^2 + 4n\sum_{i=1}^{n}\frac{1}{t_i}}}{2\sum_{i=1}^{n}\frac{1}{t_i}}$$

$$\hat{x}_0\left(\hat{m}_{(i)}\right) = \max_{x_0} L\left(\hat{m}_{(i)}, x_0\left(\hat{m}_{(i)}\right)\right) \frac{-n\hat{m}_{(i)} + \sqrt{\left(n\hat{m}_{(i)}\right)^2 + 4n\sum_{i=1}^{n}\frac{1}{t_i}}}{2\sum_{i=1}^{n}\frac{1}{t_i}} \qquad (4.28)$$

Τελικά, από τις σχέσεις (4.21), (4.22), (4.23), (4.24) και (4.28) μπορούμε να υπολογίσουμε την απόσταση των πιθανοφανειών για την παράμετρο *m* από τη σχέση (4.27):

Βήμα 5: Υπολογισμός της απόστασης των πιθανοφανειών για την παράμετρο *m*.

$$LD_i\left(m_{(i)} \mid x_0\right) = 2\left[l\left(\hat{m}, \hat{x}_0\right) - l\left(\hat{m}_{(i)}, x_0\left(\hat{m}_{(i)}\right)\right)\right] = \left[l\left(\hat{m}, \hat{x}_0\right) - \max_{x_0}\left[l\left(\hat{m}_{(i)}, x_0\right)\right]\right] =$$

$$= 2\left[\log L\left(-\frac{\sqrt{n}}{\bar{t}\sqrt{\left(\sum_{i=1}^{n}\frac{1}{t_i} - \frac{n}{\bar{t}}\right)}}, \sqrt{\frac{n}{\left(\sum_{i=1}^{n}\frac{1}{t_i} - \frac{n}{\bar{t}}\right)}},\right) - \log L\left(-\frac{\sqrt{n-1}}{\bar{t}_{(i)}\sqrt{\left(\sum_{i=1}^{n-1}\frac{1}{t_i} - \frac{n-1}{\bar{t}_{(i)}}\right)}}, \frac{-n\hat{m}_{(i)} + \sqrt{\left(n\hat{m}_{(i)}\right)^2 + 4n\sum_{i=1}^{n}\frac{1}{t_i}}}{2\sum_{i=1}^{n}\frac{1}{t_i}}\right)\right]$$

4.2.3 Η γενικευμένη απόσταση του Cook

Οι Qu και Xie (2011), ανάπτυξαν διαγνωστικούς ελέγχους για τα μοντέλα παλινδρόμησης log-Birnbaum-Saunders. Στο άρθρο τους, χρησιμοποιούν την έννοια της γενικευμένης απόστασης του Cook, η οποία εισήχθηκε αρχικά από τους Cook και Weisberg (1982), προκειμένου να μετρήσουν την αλλαγή ανάμεσα στις ποσότητες $\hat{\theta}$ και $\hat{\theta}_{(i)}$. Η απόσταση αυτή εκφράζεται ως εξής:

$$GD_i = \left(\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}\right)^T M \left(\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}\right), \tag{4.29}$$

όπου *M* είναι ένας θετικά ορισμένος πίνακας που μετράει τους σταθμισμένους συνδυασμούς των στοιχείων για την απόσταση $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$. Οι Cook και Weisberg (1982) πρότειναν αρκετές επιλογές για τον πίνακα *M*.

Στο άρθρο τους, οι Qu και Xie πρότειναν ως ελκυστική επιλογή για τον πίνακα M, τη χρησιμοποίηση του πίνακα πληροφορίας, $M = -\ddot{l}\left(\hat{\theta}\right) = -H\left(\hat{\theta}\right)$. Μία αντίστοιχη πρόταση, είναι να χρησιμοποιήσουμε τις σχέσεις (4.9), (4.29) και τον πίνακα $M = \ddot{l}_{(i)}\left(\hat{\theta}\right) = H_{(i)}\left(\hat{\theta}\right)$:

$$GD_i = \left(\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}\right)^T \left(H_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)\left(\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}\right) =$$

$$= \left(-\left(H_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)^{-1} \cdot u_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)^T \left(H_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)\left(-\left(H_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)^{-1} \cdot u_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right) =$$

$$= \left(u_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)^T \cdot \left(\left(H_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)^{-1}\right)^T \left(H_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right) \cdot \left(H_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)^{-1} u_{(i)}\left(\hat{\boldsymbol{\theta}}\right) =$$

$$= \left(u_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)^T \left(\left(H_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)^{-1}\right)^T u_{(i)}\left(\hat{\boldsymbol{\theta}}\right) \Rightarrow$$

$$GD_i = \left(u_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)^T \left(\left(H_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)^{-1}\right)^T u_{(i)}\left(\hat{\boldsymbol{\theta}}\right) \tag{4.30}$$

Εναλλακτικά, αντί να αντικαταστήσουμε στη σχέση (4.29) τη διαφορά $\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}$ με τον προσεγγιστικό ισοδύναμο όρο $-\left(\ddot{l}_{(i)}\left(\hat{\boldsymbol{\theta}}\right)\right)^{-1} \dot{l}_{(i)}\left(\hat{\boldsymbol{\theta}}\right)$ της σχέσης (4.9), προτείνουμε την απευθείας χρήση της σχέσης (4.29) επιλέγοντας ως πίνακα M τον $\left(-H\left(\hat{\boldsymbol{\theta}}\right)\right)^{-1}$.

Τέλος, μία ακόμα πρόταση, η οποία και θα χρησιμοποιηθεί στη συνέχεια, είναι να χρησιμοποιήσουμε για τη διαφορά $\left(\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}\right)$ τον προσεγγιστικό τύπο της σχέσης (4.9) και για τον πίνακα M να επιλέξουμε τον Εσσιανό πίνακα του προβλήματος, $H\left(\hat{\boldsymbol{\theta}}\right)$. Έτσι, η σχέση (4.29) γίνεται:

$$GD_i = \left(\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}\right)^T H\left(\hat{\theta}\right)\left(\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}\right), \tag{4.31}$$

Στη συνέχεια και δεδομένου πως έχουμε δύο γραμμικές προβλέπουσες, θα εργαστούμε με τον ίδιο τρόπο που εργαστήκαμε για το LD_i, προκειμένου να εξετάσουμε ξεχωριστά την επιρροή των εκτιμητριών των παραμέτρων x_0 και *m*, χρησιμοποιώντας το μέτρο GD_i.

Έτσι λοιπόν, το μέτρο που προκύπτει από την ποσότητα GD_i για την παράμετρο x_0 γίνεται:

$$GD_{x_0} = \left(\hat{\boldsymbol{\theta}}_{(i)} - \tilde{\boldsymbol{\theta}}_{x_0}\right)^T H\left(\hat{\theta}\right)\left(\hat{\boldsymbol{\theta}}_{(i)} - \tilde{\boldsymbol{\theta}}_{x_0}\right), \tag{4.32}$$

όπου $\tilde{\boldsymbol{\theta}}_{x_0} = \left(\hat{x}_0, \hat{m}\left(\hat{x}_0\right)\right)$. Τέλος, το αντίστοιχο μέτρο για την παράμετρο *m*:

$$GD_m = \left(\hat{\boldsymbol{\theta}}_{(i)} - \tilde{\boldsymbol{\theta}}_m\right)^T H\left(\hat{\theta}\right)\left(\hat{\boldsymbol{\theta}}_{(i)} - \tilde{\boldsymbol{\theta}}_m\right), \tag{4.33}$$

όπου $\tilde{\boldsymbol{\theta}}_m = \left(\hat{m}, \hat{x}_0\left(\hat{m}\right)\right)$.

4.3 Μέτρηση της τοπικής επιρροής

Τα διάφορα στατιστικά μοντέλα συνήθως χρησιμοποιούν κάποιο βαθμό προσέγγισης, με αποτέλεσμα να είναι σχεδόν πάντα λανθασμένα. Λόγω αυτής της ανακρίβειας, είναι σημαντική η αξιολόγηση της επιρροής μικρών διαταραχών του μοντέλου.

4.3.1 Η περίπτωση του IG FHTR μοντέλου

Ο Cook (1986), πρότεινε τη χρησιμοποίηση της Διαφορικής Γεωμετρίας προκειμένου να εκτιμήσει την τοπική επιρροή τέτοιων μικρών διαταραχών, οι οποίες έχουν ως αποτέλεσμα την αναχώρηση από τις υποθέσεις του στατιστικού μοντέλου. Χρησιμοποιώντας διαφορετικές περιπτώσεις διαταραχών, η προτεινόμενη προσέγγιση έχει εφαρμοστεί με επιτυχία σε πληθώρα αναλύσεων. Η προτεινόμενη μέθοδος είναι απλή στην εφαρμογή. Οι Poon και Tang (2010), χρησιμοποίησαν την ιδέα της προσέγγισης της τοπικής επιρροής για να κατασκευάσουν μέτρα επιρροής, προκειμένου να ανιχνεύουν παρατηρήσεις με δυσανάλογη επίδραση στην εκτίμηση μέγιστης πιθανοφάνειας των παραμέτρων σε μοντέλα για δεδομένα διάρκειας ζωής. Η μέθοδος τους μπορεί να εφαρμοστεί σε μεγάλη ποικιλία μοντέλων. Ωστόσο, δεν έχει χρησιμοποιηθεί για το IG FHTR μοντέλο. Στη συνέχεια, αναπτύσσουμε τη μέθοδο για την περίπτωση αυτή.

Θα χρησιμοποιήσουμε τη συνηθισμένη παραμέτρηση της σ.π.π. της IG, όταν προκύπτει από ένα FHTR μοντέλο βασισμένο σε μια ανέλιξη Wiener και δίνεται από τη σχέση (4.2):

$$f\left(t; m, x_0\right) = \frac{x_0}{\sqrt{2\pi t^3}} \exp\left(-\frac{\left(x_0 + mt\right)^2}{2t} \right), \quad t > 0, -\infty < m < +\infty, x_0 > 0 .$$

Έστω πως έχουμε *n* παρατηρήσεις, δ_i είναι δείκτης αποκοπής, $f\left(t\right)$ η παραπάνω σ.π.π., $S\left(t\right)$ η αντίστοιχη συνάρτηση επιβίωσης και $\boldsymbol{\theta}$ είναι το διάνυσμα των προς εκτίμηση παραμέτρων του μοντέλου. Η συνάρτηση πιθανοφάνειας δίνεται από τη σχέση:

$$L\left(\boldsymbol{\theta}\right) = \prod_{i=1}^{n} \left[f\left(t_i, \boldsymbol{\theta}\right) \right]^{\delta_i} \left[S\left(t_i, \boldsymbol{\theta}\right) \right]^{\left(1-\delta_i\right)} \tag{4.34}$$

ή ισοδύναμα:

$$l\left(\boldsymbol{\theta}\right) = \log L\left(\boldsymbol{\theta}\right) = \sum_{i=1}^{n} \left[\delta_i \log f\left(t_i, \boldsymbol{\theta}\right) + \left(1-\delta_i\right) \log S\left(t_i, \boldsymbol{\theta}\right) \right] \tag{4.35}$$

Η εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\boldsymbol{\theta}}$ του $\boldsymbol{\theta}$, προκύπτει από τη βελτιστοποίηση της σχέσης (4.35). Προκειμένου να εκτιμήσουμε τον τρόπο με τον οποίο μεμονωμένες παρατηρήσεις επηρεάζουν την ε.μ.π., εισάγουμε βάρη $\boldsymbol{\omega} = \left(\omega_1, ..., \omega_n\right)'$, τα οποία υποδεικνύουν κάποια επιπλέον διαταραχή στο μοντέλο. Η αντίστοιχη "διαταραγμένη" συνάρτηση του λογαρίθμου της πιθανοφάνειας γίνεται:

$$l\left(\boldsymbol{\theta} | \boldsymbol{\omega}\right) = \sum_{i=1}^{n} \omega_i \left[\delta_i \log f\left(t_i, \boldsymbol{\theta}\right) + \left(1-\delta_i\right) \log S\left(t_i, \boldsymbol{\theta}\right) \right] \tag{4.36}$$

Από τη σχέση (4.36) παρατηρούμε πως εάν $\boldsymbol{\omega}=\boldsymbol{\omega}_0=(1,1,...,1)'$, τότε $l(\boldsymbol{\theta}|\boldsymbol{\omega}_0)=l(\boldsymbol{\theta})$. Εάν πραγματοποιηθεί μία μικρή διαταραχή του ω_i από το $\omega_i=1$ και οδηγήσει σε πολύ διαφορετική εκτίμηση της παραμέτρου $\boldsymbol{\theta}$, τότε η *i*-παρατήρηση θεωρείται σημείο επιρροής.

Ο Cook (1986) παρουσίασε μία σειρά εύκολων και υπολογιστικά γρήγορων πράξεων, οι οποίες μπορούν να εντοπίσουν τα σημεία επιρροής του δείγματος. Θεωρούμε τον πίνακα:

$$\Pi=-\Delta'\left(\ddot{l}\left(\hat{\boldsymbol{\theta}}\right)\right)^{-1}\Delta_{|\boldsymbol{\theta}=\hat{\boldsymbol{\theta}},\,\boldsymbol{\omega}=\boldsymbol{\omega}_0}, \tag{4.37}$$

όπου

$$\ddot{l}\left(\hat{\boldsymbol{\theta}}\right)=\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{4.38}$$

είναι ένας πίνακας διάστασης $p\times p$ και

$$\Delta=\left.\frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\omega}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}},\,\boldsymbol{\omega}=\boldsymbol{\omega}_0} \tag{4.39}$$

είναι ένας πίνακας διάστασης $p\times n$. Από τις σχέσεις (4.34) και (4.36), έχουμε:

$$\ddot{l}\left(\hat{\boldsymbol{\theta}}\right)=\sum_{i=1}^{n}\left[\delta_i\left.\frac{\partial^2\log f(t_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}+(1-\delta_i)\left.\frac{\partial^2\log S(t_i,\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right] \tag{4.40}$$

Ακόμα, το στοιχείο $(i,\,j)$ του πίνακα Δ δίνεται από τη σχέση:

$$\frac{\partial^2 l(\boldsymbol{\theta}|\boldsymbol{\omega})}{\partial\omega_i\,\partial\theta_j}=\delta_i\left.\frac{\partial\log f(t_i,\boldsymbol{\theta})}{\partial\theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}+(1-\delta_i)\left.\frac{\partial\log S(t_i,\boldsymbol{\theta})}{\theta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{4.41}$$

Ένας τρόπος εντοπισμού σημείων επιρροής είναι με τη βοήθεια του πίνακα Π. Εάν το *i*-οστό στοιχείο του ιδιοδιανύσματος που αντιστοιχεί στη μεγαλύτερη ιδιοτιμή του Π είναι μεγάλο, τότε η *i*-οστή παρατήρηση είναι σημείο επιρροής. Οι ιδιοτιμές και τα ιδιοδιανύσματα του Π υπολογίζονται εύκολα και γρήγορα με τη βοήθεια υπολογιστικών πακέτων.

Εναλλακτικά, μπορούμε να υπολογίσουμε την ποσότητα:

$$B_i=\frac{\Pi_{ii}}{\sqrt{tr\left(\Pi^2\right)}},\ i=1,2,...,n \tag{4.42}$$

Αποδεικνύεται (Poon και Poon, 1999), πως η ποσότητα B_i λαμβάνει τιμές μέσα στο μοναδιαίο διάστημα και πως παρατηρήσεις με μεγάλη τιμή για το B_i θεωρούνται σημεία επιρροής. Από τις σχέσεις (4.37), (4.40), (4.41) και (4.42) φαίνεται πως τα μέτρα επιρροής μπορεί να υπολογιστούν εύκολα όταν είναι διαθέσιμες οι ποσότητες $f(t_i,\boldsymbol{\theta})$ και $S(t_i,\boldsymbol{\theta})$.

Διαγράμματα θέσης (index plots) χρησιμοποιούνται για τον εντοπισμό παρατηρήσεων με μεγάλη τιμή στο ιδιοδιάνυσμα ή την ποσότητα B_i. Επιπρόσθετα, έχει προταθεί από τους Poon και Poon (1999) ένα πλαφόν, προκειμένου να αυτοματοποιηθεί η διαδικασία εντοπισμού των σημείων επιρροής και δίνεται από τη σχέση:

$$c = \frac{2tr(\Pi)}{n\sqrt{tr\Pi^2}}, i = 1, 2, ..., n \tag{4.43}$$

Κάθε παρατήρηση με τιμή για το B_i που ξεπερνάει το πλαφόν αυτό, θεωρείται πιθανό σημείο επιρροής. Να παρατηρήσουμε πως η ποσότητα c είναι το διπλάσιο της μέσης τιμής των $B_i, i = 1, 2, ..., n$.

4.3.2 Η περίπτωση του IG FHT μοντέλου, χωρίς μεταβλητές και σημεία αποκοπής

Έστω $X_1, ..., X_n \sim IG(m, x_0)$, με $x_0 = \exp(\gamma_0)$ και $m = \beta_0$, όπου γ_0, β_0 σταθερές.

Σ.π.π: $$f(t; m, x_0) = \frac{x_0}{\sqrt{2\pi t^3}} \exp\left(-\frac{(x_0 + mt)^2}{2t} \right), \quad t > 0, -\infty < m < +\infty, x_0 > 0$$

$$\log f(t; m, x_0) = \log x_0 - \frac{1}{2}\log(2\pi) - \frac{3}{2}\log t - \frac{(x_0 + mt)^2}{2t}$$

$$l(\boldsymbol{\theta}|\boldsymbol{\omega}) = \sum_{j=1}^{n} \omega_j \log f(t_j, \boldsymbol{\theta}) = \left(\log x_0 - \frac{1}{2}\log(2\pi) \right) \sum_{j=1}^{n} \omega_j - \frac{3}{2}\sum_{j=1}^{n} \omega_j \log t_j - \frac{1}{2}\sum_{j=1}^{n} \frac{\omega_j (x_0 + mt_j)^2}{t_j} \Rightarrow$$

$$l_j = \left(\log x_0 - \frac{1}{2}\log(2\pi) \right) \omega_j - \frac{3}{2}\omega_j \log t_j - \frac{1}{2}\frac{\omega_j (x_0 + mt_j)^2}{t_j}$$

Επομένως, σύμφωνα με τη σχέση (4.39) έχουμε:

$$\Delta_{ij} = \frac{\partial l_j}{\partial \theta_i}, i = 1, \ldots, p \text{ (εδώ } p = 2) \tag{4.44}$$

Τελικά,

$$\Delta_{1j} = \frac{\partial l_j}{\partial m} = -\frac{1}{2}\frac{\omega_j 2t_j (x_0 + mt_j)}{t_j} = -\omega_j (x_0 + mt_j)$$

Επίσης,

$$\Delta_{2j}=\frac{\partial l_j}{\partial x_0}=\frac{1}{x_0}\omega_j-\frac{1}{2}\frac{\omega_j 2\left(x_0+mt_j\right)}{t_j}=\frac{\omega_j}{x_0}-\frac{\omega_j\left(x_0+mt_j\right)}{t_j}$$

Εάν οι διάφοροι υπολογισμοί γίνουν για το ω_0, έχουμε:

$$\Delta_{1j}=-\left(\hat{x}_0+\hat{m}t_j\right) \tag{4.45}$$

και

$$\Delta_{2j}=\frac{1}{\hat{x}_0}-\frac{\left(\hat{x}_0+\hat{m}t_j\right)}{t_j} \tag{4.46}$$

Απομένει ο υπολογισμός των ε.μ.π. των παραμέτρων x_0 και *m*.

Για την παράμετρο *m* έχουμε:

$$\frac{\partial l_j}{\partial\theta_1}=\frac{\partial l_j}{\partial m}=-\frac{1}{2}\sum_{j=1}^{n}\frac{\omega_j 2t_j\left(x_0+mt_j\right)}{t_j}=-\sum_{j=1}^{n}\omega_j\left(x_0+mt_j\right)$$

Για την παράμετρο x_0 έχουμε:

$$\frac{\partial l_j}{\partial\theta_2}=\frac{\partial l_j}{\partial x_0}=\frac{1}{x_0}\sum_{j=1}^{n}\omega_j-\frac{1}{2}\sum_{j=1}^{n}\frac{\omega_j 2\left(x_0+mt_j\right)}{t_j}$$

Εάν οι διάφοροι υπολογισμοί γίνουν για το ω_0, έχουμε:

$$\frac{\partial l_j}{\partial\theta_1}=\frac{\partial l_j}{\partial m}=-\sum_{j=1}^{n}\left(x_0+mt_j\right)=0\ \Rightarrow n\hat{x}_0+\hat{m}n\bar{t}=0\ \Rightarrow$$

$$\hat{m}=-\frac{\hat{x}_0}{\bar{t}} \tag{4.47}$$

$$\frac{\partial l_j}{\partial\theta_2}=\frac{\partial l_j}{\partial x_0}=\frac{1}{x_0}-x_0\sum_{j=1}^{n}\frac{1}{t_j}+nm=0\ \ \Rightarrow\ \frac{1}{\hat{x}_0}-\hat{x}_0\sum_{j=1}^{n}\frac{1}{t_j}+n\hat{m}=0\ \ \Rightarrow$$

$$\frac{1}{-\hat{m}\bar{t}}=-\hat{m}\bar{t}\sum_{j=1}^{n}\frac{1}{t_j}+n\hat{m}=-\hat{m}\bar{t}\left(\sum_{j=1}^{n}\frac{1}{t_j}-\frac{1}{\bar{t}}\right)\ \ \Rightarrow$$

$$\hat{m}^2=\frac{n}{\bar{t}^2\left(\sum_{j=1}^{n}\frac{1}{t_j}-\frac{1}{\bar{t}}\right)} \tag{4.48}$$

και ύστερα από πράξεις, η σχέση (4.47) με τη βοήθεια της σχέσης (4.48) δίνει:

$$\hat{x}_0^{\,2} = \hat{m}^2 \bar{t}^{\,2} = \frac{n}{\left(\sum_{j=1}^{n} \frac{1}{t_j} - \frac{1}{\bar{t}} \right)} \tag{4.49}$$

Επίσης,

$$\frac{\partial^2 l_j}{\partial \theta_1^2} = \frac{\partial^2 l_j}{\partial m^2} = -\sum_{j=1}^{n} \omega_j t_j = -n\bar{t}^{\,2} \quad \text{στο } \boldsymbol{\omega} = \boldsymbol{\omega}_0 .$$

Όμοια,

$$\frac{\partial^2 l_j}{\partial \theta_2^2} = \frac{\partial^2 l_j}{\partial x_0^2} = -\frac{1}{x_0^2} \sum_{j=1}^{n} \omega_j - x_0 \sum_{j=1}^{n} \frac{\omega_j}{t_j} = -\frac{n}{x_0^2} - \sum_{j=1}^{n} \frac{1}{t_j} \quad \text{στο } \boldsymbol{\omega} = \boldsymbol{\omega}_0 .$$

Ακόμα,

$$\frac{\partial^2 l_j}{\partial \theta_1 \partial \theta_2} = \frac{\partial^2 l_j}{\partial m \partial x_0} = \frac{\partial^2 l_j}{\partial x_0 \partial m} = -\sum_{j=1}^{n} \omega_j = -n \quad \text{στο } \boldsymbol{\omega} = \boldsymbol{\omega}_0 .$$

Τελικά, έχουμε για τις ποσότητες Δ_{1j} και Δ_{2j}:

$$\Delta_{1j} = -\left(\hat{x}_0 + \hat{m} t_j \right) = -\hat{m}\left(t_j - \bar{t} \right) \tag{4.50}$$

και

$$\Delta_{2j} = \frac{1}{\hat{x}_0} - \frac{\left(\hat{x}_0 + \hat{m} t_j \right)}{t_j} = \frac{1}{\hat{x}_0} - \hat{x}_0 \left(\frac{1}{t_j} - \frac{1}{\bar{t}} \right) \tag{4.51}$$

Επομένως, $\Delta = \left(\Delta_{1j}, \Delta_{2j} \right)$ και ο πίνακας Π υπολογίζεται από τη σχέση (4.37).

4.3.3 Η περίπτωση του IG FHT μοντέλου, χωρίς σημεία αποκοπής

Έστω $X_1, ..., X_n \sim IG\left(m, x_0 \right)$, με $x_0 = \exp\left(\boldsymbol{\beta}' \boldsymbol{u} \right)$ και $m = \boldsymbol{\gamma}' \boldsymbol{v}$,

όπου $\boldsymbol{\theta} = \left(\beta_1, ..., \beta_p, \gamma_1, ..., \gamma_q \right)$ και $u_1 = v_1 = 1$.

Σ.π.π.: $$f\left(t; m, x_0 \right) = \frac{x_0}{\sqrt{2\pi t^3}} \exp\left(-\frac{\left(x_0 + mt \right)^2}{2t} \right), \quad t > 0, -\infty < m < +\infty, x_0 > 0$$

$$l\left(\boldsymbol{\theta} | \boldsymbol{\omega} \right) = \sum_{j=1}^{n} \omega_j \log f\left(t_j, \boldsymbol{\theta} \right) = \left(\log x_0 - \frac{1}{2} \log\left(2\pi \right) \right) \sum_{j=1}^{n} \omega_j - \frac{3}{2} \sum_{j=1}^{n} \omega_j \log t_j - \frac{1}{2} \sum_{j=1}^{n} \frac{\omega_j \left(x_0 + m t_j \right)^2}{t_j}$$

Για την παράμετρο x_0, έχουμε:

$$\frac{\partial l}{\partial \theta_i}=\frac{\partial l}{\partial \beta_i}=\sum_{j=1}^{n}\omega_j\left[u_{ij}-\frac{2\left(x_0+mt_j\right)}{2t_j}\frac{\partial x_0}{\partial \beta_i}\right]=\sum_{j=1}^{n}\omega_j\left[u_{ij}-\frac{x_0\left(x_0+mt_j\right)u_{ij}}{t_j}\right]\ \Rightarrow$$

$$\frac{\partial l}{\partial \theta_i}=\sum_{j=1}^{n}\omega_j u_{ij}\left[1-\frac{x_0\left(x_0+mt_j\right)}{t_j}\right],\ i=1,...,p$$

Για την παράμετρο *m*, έχουμε:

$$\frac{\partial l}{\partial \theta_{p+i}}=\frac{\partial l}{\partial \gamma_i}=\sum_{j=1}^{n}\omega_j\left[-\frac{2\left(x_0+mt_j\right)t_j}{2t_j}v_{ij}\right]=\sum_{j=1}^{n}\omega_j v_{ij}\left(-x_0+mt_j\right),\ i=1,...,q$$

Επομένως, σύμφωνα με τη σχέση (4.39) έχουμε:

$$\Delta_{ij}=\frac{\partial l_j}{\partial \theta_i}=\frac{\partial l}{\partial \beta_i}=u_{ij}\left(1-\frac{x_0\left(x_0+mt_j\right)}{t_j}\right),\ i=1,2,...,p\ ,\ j=1,...,n \qquad (4.52)$$

Επίσης,

$$\Delta_{p+i,j}=\frac{\partial l_j}{\partial \theta_{p+i}}=\frac{\partial l}{\partial \gamma_i}--=v_{ij}x_0\left(x_0+mt_j\right),\ i=1,2,...,q\ ,\ j=1,...,n \qquad (4.53)$$

4.4 Έλεγχος των θεωρητικών αποτελεσμάτων

Στη συνέχεια, παρουσιάζονται τα αριθμητικά αποτελέσματα από τη μελέτη στην οποία προχωρήσαμε, προκειμένου να επιβεβαιώσουμε ή να απορρίψουμε την ικανότητα των διαφόρων προτεινόμενων μέτρων για την επιρροή και να επιλέξουμε το καλύτερο για την περίπτωση του FHTR μοντέλου.

4.4.1 Σχεδιασμός της μελέτης

Δεδομένα δημιουργήθηκαν από το μοντέλο παλινδρόμησης IG FHTR για επιλεγμένες τιμές των συντελεστών $\boldsymbol{\beta}$ και $\boldsymbol{\gamma}$. Εξετάστηκαν οι περιπτώσεις του μοντέλου χωρίς μεταβλητές (IG GLM) και του μοντέλου IG FHTR με τρεις μεταβλητές. Για το αληθινό μοντέλο, χρησιμοποιήθηκαν οι μεταβλητές, ως εξής:

α) Δημιουργία μιας μεταβλητής από την Κανονική κατανομή με μέση τιμή -3 και μοναδιαία διασπορά.

β) Δημιουργία μίας μεταβλητής από την Ομοιόμορφη κατανομή στο διάστημα (0,1).

γ) Δημιουργία μιας μεταβλητής από τη Διωνυμική κατανομή με πιθανότητα επιτυχίας 0.5.

Κάθε παρατήρηση στο προσομοιωμένο δείγμα δημιουργήθηκε από την αντίστροφη Γκαουσιανή κατανομή με σ.π.π. της σχέσης (4.1), με τη χρήση της ρουτίνας *rinvgauss* του

πακέτου *statmod*, όπως είδαμε και στο προηγούμενο κεφάλαιο. Για κάθε παρατήρηση στο προσομοιωμένο δείγμα, οι τιμές των παραμέτρων της IG κατανομής κατασκευάστηκαν από τη σχέση:

$$\ln x_0 = \gamma_0 + \gamma_1 z_1 + ... + \gamma_k z_k = \boldsymbol{\gamma' z}$$
$$m = \beta_0 + \beta_1 z_1 + ... + \beta_k z_k = \boldsymbol{\beta' z} \quad ,$$

ύστερα από τη δημιουργία των τιμών των μεταβλητών για τη συγκεκριμένη παρατήρηση.

Ασχοληθήκαμε μόνο με την περίπτωση $m < 0$ για όλες τις παρατηρήσεις, προκειμένου η απορρόφηση της ανέλιξης από το σύνορο να είναι εξασφαλισμένη. Χρησιμοποιήθηκαν προσομοιωμένα δείγματα μεγέθους 25, 50 και 100 παρατηρήσεων για την περίπτωση όπου δεν υπήρχαν μεταβλητές στο μοντέλο και δείγματα μεγέθους 50, 75 και 100 παρατηρήσεων για την περίπτωση όπου στο αληθινό μοντέλο υπήρχαν τρεις μεταβλητές. Άτυπες τιμές κατασκευάστηκαν για τις παραμέτρους *m* και x_0 στο αληθινό μοντέλο, ως εξής:

Έστω ότι επιθυμούμε να κατασκευάσουμε άτυπη τιμή για την παράμετρο *m*. Τότε, για κάποιο *i*, $i = 1, 2, ..., n$, θέτουμε:

$m_i = m_i + c$, όπου c σταθερά, με $c > 0$

Με όμοιο τρόπο, εάν επιθυμούμε να κατασκευάσουμε άτυπη τιμή για την παράμετρο x_0, τότε για κάποιο *i*, $i = 1, 2, ..., n$, θέτουμε:

$\ln x_{0i} = \ln x_{0i} + c$, όπου c σταθερά, με $c > 0$

Για κάθε ξεχωριστή περίπτωση που μελετήθηκε, είχαμε διαφορετική τιμή της σταθεράς *c* και διαφορετικά διανύσματα συντελεστών για τις παραμέτρους x_0 και *m* στο αληθινό μοντέλο, με αποτέλεσμα οι διάφορες περιπτώσεις που παρουσιάζονται στη συνέχεια να μην είναι μεταξύ τους συγκρίσιμες. Όλοι οι υπολογισμοί έγιναν στο στατιστικό πακέτο R. 500 επαναλήψεις έγιναν για κάθε ξεχωριστή περίπτωση που μελετήθηκε. Η προσαρμογή του FHT μοντέλου παλινδρόμησης έγινε με ελαχιστοποίηση της συνάρτησης του αρνητικού λογαρίθμου της πιθανοφάνειας χρησιμοποιώντας τη ρουτίνα *optim* και τη μέθοδο βελτιστοποίησης BFGS.

4.4.2 Αποτελέσματα για το μοντέλο IG GLM

Αρχικά ασχοληθήκαμε με την ειδική περίπτωση όπου η παράμετρος x_0 δεν εξαρτάται από μεταβλητές, με αποτέλεσμα το FHTR μοντέλο να μπορεί να εκφραστεί ως ένα GLM, όπως εργαστήκαμε και στην περίπτωση της επιλογής μεταβλητών, στην Παράγραφο 3.4.4 του Κεφαλαίου 3. Στην περίπτωση που μελετάμε, το μοντέλο FHTR είναι στην πραγματικότητα ισοδύναμο με ένα μοντέλο IG GLM, με την αντίστροφη συνάρτηση σύνδεσης.

Άτυπη τιμή κατασκευάστηκε για την παράμετρο x_0, για κάποιο i, $i = 1, 2, ..., n$ με τον τρόπο που περιγράψαμε στην Παράγραφο 4.4.1. Εξετάσαμε την εφαρμογή των μέτρων LD_i, LDx_0, LD_m, GD_i, GDx_0, GD_m και B_i που παρουσιάσαμε στην Παράγραφο 4.2, σε δεδομένα δημιουργημένα από ένα σταθερό αληθινό FHT μοντέλο χωρίς μεταβλητές ούτε για το x_0, αλλά ούτε και για το m.

Σε κάθε επανάληψη του αλγορίθμου, ελέγχουμε εάν το κάθε προτεινόμενο μέτρο είναι ικανό να αναγνωρίσει απευθείας την άτυπη τιμή. Επίσης, για το κάθε μέτρο που μελετάμε, κατασκευάζουμε ένα πλαφόν, το οποίο ισούται με το διπλάσιο του μέσου όρου των τιμών που λαμβάνει το κάθε μέτρο σε κάθε επανάληψη. Για το B_i, το πλαφόν αυτό υπολογίζεται μέσω της σχέσης (4.43). Στη συνέχεια, ελέγχουμε εάν η άτυπη παρατήρηση έχει τιμή για το κάθε ξεχωριστό μέτρο που να ξεπερνάει αυτό το πλαφόν.

Αποτελέσματα παρουσιάζονται στους πίνακες που ακολουθούν για τιμές των συντελεστών β_0 και γ_0 στο αληθινό μοντέλο ίσες με -2 και 0.5 αντίστοιχα. Εκτός από τα ποσοστά εντοπισμού της άτυπης τιμής από τους διάφορους δείκτες μέτρησης της επιρροής, υπολογίζουμε και τα ποσοστά στα οποία η άτυπη τιμή είναι μεγαλύτερη και από το αντίστοιχο πλαφόν.

	Ποσοστό των επαναλήψεων (%), στις οποίες η τιμή του μέτρου στη θέση της άτυπης x_0 παρατήρησης είναι:					
	Η μεγαλύτερη τιμή του μέτρου			**Μεγαλύτερη από ένα πλαφόν**		
	***n*=25**	***n*=50**	***n*=100**	***n*=25**	***n*=50**	***n*=100**
LD_i	69.8	56.2	47.4	93.0	92.4	92.6
GD_i	64.4	54.4	45.6	92.4	92.2	92.6
B_i	79.8	67.6	57.0	96.8	96.8	97.0

Πίνακας 4.1: Στοιχεία σχετικά με την αποτελεσματικότητα των μέτρων LD_i, GD_i και B_i

Η αποτελεσματικότητα των μέτρων LD_i, GD_i και B_i στην ορθή αναγνώριση του σημείου επιρροής είναι εμφανής, με το μέτρο B_i να κρίνεται ως το πιο αξιόπιστο, παρουσιάζοντας τα μεγαλύτερα ποσοστά εντοπισμού της άτυπης τιμής. Τα παραπάνω αποτελέσματα είναι περισσότερο εμφανή για μικρότερο μέγεθος δείγματος.

Αξιοπρόσεκτο είναι το γεγονός πως, παρόλο που η άτυπη τιμή βρίσκεται σε μία παρατήρηση για το x_0, εντούτοις τα μέτρα LD_m και GD_m φαίνεται να είναι πιο αποτελεσματικά στην αναγνώριση της άτυπης τιμής από τα μέτρα LDx_0 και GDx_0. Επιπρόσθετα, τα ποσοστά των φορών που η άτυπη τιμή ξεπερνάει το πλαφόν είναι αρκετά

υψηλά κάθε φορά, ανεξάρτητα από το μέτρο του ενδιαφέροντος και τον αριθμό των παρατηρήσεων του δείγματος, γεγονός που μετατρέπει την τιμή του πλαφόν σε ένα αξιόπιστο κάτω φράγμα για την άτυπη τιμή.

	Ποσοστό των επαναλήψεων (%), στις οποίες η τιμή του μέτρου στη θέση της άτυπης x_0 παρατήρησης είναι:					
	Η μεγαλύτερη τιμή του μέτρου			**Μεγαλύτερη από ένα πλαφόν**		
	***n*=25**	***n*=50**	***n*=100**	***n*=25**	***n*=50**	***n*=100**
LD_{x_0}	49.8	38.8	29.4	78.8	78.2	81.0
GD_{x_0}	49.8	38.8	29.4	78.8	78.4	81.0
LD_m	79.2	64.8	55.0	94.6	94.6	94.6
GD_m	79.2	64.8	55.0	94.6	94.6	94.6

Πίνακας 4.2: Στοιχεία σχετικά με την αποτελεσματικότητα των μέτρων LD_{x_0}, LD_m, GD_{x_0} και GD_m

Να παρατηρήσουμε πως οι δείκτες LD_m και GD_m, καθώς και οι δείκτες LDx_0 και GDx_0 δίνουν ακριβώς την ίδια τιμή, παρόλο που υπολογίζονται με δύο διαφορετικούς προσεγγιστικούς τρόπους, γεγονός που επιβεβαιώνει την ορθότητα των διάφορων υπολογισμών. Τέλος, οι δείκτες LDx_0 και GDx_0 εμφανίζουν περιορισμένη ικανότητα εντοπισμού της άτυπης τιμής, ειδικά σε σχέση με τους δείκτες LD_m και GD_m, τη στιγμή που η άτυπη τιμή βρίσκεται στην παράμετρο x_0.

	Ποσοστό των επαναλήψεων (%), στις οποίες η τιμή του δείκτη στη θέση της άτυπης *m* παρατήρησης είναι:					
	Η μεγαλύτερη τιμή του δείκτη			**Μεγαλύτερη από ένα πλαφόν**		
	***n*=25**	***n*=50**	***n*=100**	***n*=25**	***n*=50**	***n*=100**
LD_i	67.0	62.6	59.6	77.2	76.8	77.4
GD_i	65.2	61.4	59.4	76.2	76.4	77.4
B_i	70.8	66.8	62.2	79.8	79.4	81.0

Πίνακας 4.3: Στοιχεία σχετικά με την αποτελεσματικότητα των μέτρων LD_i, GD_i και B_i

Στη συνέχεια ασχοληθήκαμε με την περίπτωση όπου η ατυπη τιμή κατασκευάστηκε για την παράμετρο *m*, για κάποιο *i*, $i = 1, 2, ..., n$. Τα αποτελέσματα παρουσιάζονται στους Πίνακες 4.3 και 4.4.

	Ποσοστό των επαναλήψεων (%), στις οποίες η τιμή του δείκτη στη θέση της άτυπης *m* παρατήρησης είναι:					
	Η μεγαλύτερη τιμή του δείκτη			**Μεγαλύτερη από ένα πλαφόν**		
	***n*=25**	***n*=50**	***n*=100**	***n*=25**	***n*=50**	***n*=100**
LD_{x_0}	61.2	56.8	54.8	72.4	72.6	71.8
GD_{x_0}	61.2	56.8	54.8	72.4	72.8	71.8
LD_m	70.4	65.2	61.2	78.2	78.0	78.0
GD_m	70.4	65.2	61.2	78.2	78.0	78.0

Πίνακας 4.4: Στοιχεία σχετικά με την αποτελεσματικότητα των μέτρων LD_{x_0}, LD_m, GD_{x_0} και GD_m

Η αποτελεσματικότητα των μέτρων LD_i, GD_i και B_i στην ορθή αναγνώριση του σημείου επιρροής είναι και σε αυτήν την περίπτωση εμφανής, με το μέτρο B_i να κρίνεται ως το πιο αξιόπιστο. Και σε αυτήν την περίπτωση, τα μέτρα LD_m και GD_m είναι πιο αποτελεσματικά στην αναγνώριση της άτυπης τιμής από τα LDx_0 και GD_{x_0}. Το συγκεκριμένο αποτέλεσμα είναι αναμενόμενο, καθώς η άτυπη τιμή βρίσκεται στην παράμετρο *m*. Τα παραπάνω αποτελέσματα είναι και εδώ περισσότερο εμφανή για μικρότερο μέγεθος δείγματος. Τέλος, τα ποσοστά των φορών που η άτυπη τιμή ξεπερνάει το πλαφόν είναι αρκετά υψηλά κάθε φορά, ανεξάρτητα από το μέτρο του ενδιαφέροντος και τον αριθμό των παρατηρήσεων του δείγματος.

4.4.3 Αποτελέσματα προσομοιώσεων για το γενικό FHTR μοντέλο

Για τη γενική περίπτωση ενός FHTR μοντέλου, όπου υπάρχουν δύο παράμετροι και δύο γραμμικές συναρτήσεις σύνδεσης, εξετάσαμε την εφαρμογή των μέτρων LD_i, LDx_0, LD_m, GD_i, GDx_0, GD_m και B_i, σε δεδομένα δημιουργημένα από ένα αληθινό FHT μοντέλο με 3 μεταβλητές, οι οποίες επιδρούν και στις δύο παραμέτρους, x_0 και *m* του μοντέλου. Ασχοληθήκαμε με τρεις διαφορετικές περιπτώσεις:

α) Δεν υπάρχει άτυπη τιμή σε καμία παράμετρο του μοντέλου.

β) Η άτυπη τιμή βρίσκεται σε κάποια μονάδα για την παράμετρο x_0.

γ) Η άτυπη τιμή βρίσκεται σε κάποια μονάδα για την παράμετρο *m*.

Για την περίπτωση όπου δεν υπάρχει άτυπη τιμή σε καμία παράμετρο του αληθινού μοντέλου, ελέγχουμε εάν το μέτρο του ενδιαφέροντος είναι ικανό να αναγνωρίσει την παρατήρηση με τη μεγαλύτερη τιμή για την κάθε παράμετρο. Για τις άλλες δύο περιπτώσεις, σε κάθε επανάληψη του αλγορίθμου ελέγχουμε εάν το αντίστοιχο μέτρο του ενδιαφέροντος είναι ικανό να αναγνωρίσει απευθείας την άτυπη τιμή. Επίσης, για το κάθε μέτρο που μελετάμε, εισάγουμε ένα πλαφόν και ελέγχουμε εάν η άτυπη παρατήρηση έχει τιμή για το κάθε ξεχωριστό μέτρο που να ξεπερνάει αυτό το πλαφόν.

Η περίπτωση όπου δεν υπάρχει outlier σε κάποια παράμετρο

Για τα επόμενα, χρησιμοποιήθηκαν τρεις μεταβλητές ως εξής:

α) μία μεταβλητή από την Κανονική κατανομή με μέση τιμή -3 και διασπορά 1.

β) μία μεταβλητή από την Ομοιόμορφη κατανομή στο διάστημα (0,1) και

γ) μία μεταβλητή από τη Διωνυμική κατανομή με πιθανότητα επιτυχίας 0.5.

Αληθινό μοντέλο:

$$\ln x_0 = 0.25 - 0.5z_1 - 0.25z_2 + 0.5z_3$$
$$m = 1 + 2z_1 + 3.5z_2 - 1.5z_2$$

Α) Αποτελέσματα για την παράμετρο x_0

Αποτελέσματα παρουσιάζονται στους Πίνακες 4.5 και 4.6, όπου σε κάθε επανάληψη υπολογίζουμε το μέσο αριθμό των παρατηρήσεων με τιμή για το μέτρο του ενδιαφέροντος μεγαλύτερη από ένα πλαφόν.

	Μέσος όρος των παρατηρήσεων που ξεπερνούν το πλαφόν του δείγματος σε κάθε επανάληψη		
	***n* = 50**	***n*=75**	***n*=100**
LD_i	5.54	8.67	10.63
GD_i	5.49	8.32	10.54
B_i	6.87	10.0	13.20

Πίνακας 4.5: Στοιχεία σχετικά με την αποτελεσματικότητα των μέτρων LD_i, GD_i και B_i

Ο δείκτης B_i επιτρέπει σε μεγαλύτερο αριθμό παρατηρήσεων να υπερβούν το πλαφόν, ανεξαρτήτως μεγέθους δείγματος.

	Μέσος όρος των παρατηρήσεων που ξεπερνούν το πλαφόν του δείγματος σε κάθε επανάληψη		
	***n* = 50**	***n*=75**	***n*=100**
LD_{x_0}	5.45	8.04	9.82
GD_{x_0}	5.49	8.12	9.85
LD_m	4.70	7.06	8.43
GD_m	4.66	7.00	8.39

Πίνακας 4.6: Στοιχεία σχετικά με την αποτελεσματικότητα των μέτρων LD_{x_0}, LD_m, GD_{x_0} και GD_m

Για ακόμα μία φορά, οι δείκτες LD_m και GD_m, καθώς και οι δείκτες LD_{x_0} και GD_{x_0} δίνουν σχεδόν ίδια αποτελέσματα, με τους δύο πρώτους να έχουν τη μεγαλύτερη διακριτική ικανότητα από όλους τους δείκτες, επιτρέποντας σε λίγες κάθε φορά παρατηρήσεις να υπερβαίνουν το πλαφόν, για οποιοδήποτε μέγεθος δείγματος .

Β) Αποτελέσματα για την παράμετρο *m*

Τα διάφορα αποτελέσματα παρουσιάζονται στους Πίνακες 4.7 και 4.8 που ακολουθούν. Παρόμοια εικόνα έχουμε με την προηγούμενη περίπτωση για το μέσο όρο των παρατηρήσεων που ξεπερνούν το πλαφόν του δείγματος σε κάθε επανάληψη, με το δείκτη B_i να είναι ο λιγότερο "αυστηρός" ως προς τον αριθμό των παρατηρήσεων που υπερβαίνουν κάθε φορά το πλαφόν. Τα παραπάνω αποτελέσματα, ίσως συνδέονται με την επιλογή του αρχικού διανύσματος των συντελεστών των δύο παραμέτρων του αληθινού μοντέλου.

	Μέσος όρος των παρατηρήσεων που ξεπερνούν το πλαφόν του δείγματος σε κάθε επανάληψη		
	***n* = 50**	***n*=75**	***n*=100**
LD_i	5.91	8.53	9.55
GD_i	5.32	8.29	9.25
B_i	6.74	9.83	13.05

Πίνακας 4.7: Στοιχεία σχετικά με την αποτελεσματικότητα των μέτρων LD_i, GD_i και B_i

	Μέσος όρος των παρατηρήσεων που ξεπερνούν το πλαφόν του δείγματος σε κάθε επανάληψη		
	***n* = 50**	***n*=75**	***n*=100**
LD_{x_0}	5.45	7.78	8.99
GD_{x_0}	5.48	7.84	9.09
LD_m	5.06	7.25	7.38
GD_m	4.93	7.12	7.38

Πίνακας 4.8: Στοιχεία σχετικά με την αποτελεσματικότητα των μέτρων LD_{x_0}, LD_m, GD_{x_0} και GD_m

Η περίπτωση όπου το outlier βρίσκεται στην παράμετρο x_0

Για τα επόμενα, χρησιμοποιήθηκαν τρεις μεταβλητές, όπως και στην περίπτωση που δεν υπάρχει άτυπη τιμή σε κάποια παράμετρο του αληθινού μοντέλου.

Αληθινό μοντέλο:

$$\ln x_0 = 0.25 - 0.5z_1 - 0.25z_2 + 0.5z_3$$
$$m = 1 + 2z_1 + 3.5z_2 - 1.5z_2$$

Η αποτελεσματικότητα των διαφόρων μέτρων στην ορθή αναγνώριση του σημείου επιρροής είναι εμφανής, με το δείκτη LD_i να εμφανίζει τα υψηλότερα ποσοστά εντοπισμού της άτυπης x_0-τιμής.

	Ποσοστό των επαναλήψεων (%), στις οποίες η τιμή του μέτρου στη θέση της άτυπης x_0 παρατήρησης είναι:					
	Η μεγαλύτερη τιμή του μέτρου			Μεγαλύτερη από ένα πλαφόν		
	***n* = 50**	***n*=75**	***n*=100**	***n* = 50**	***n*=75**	***n*=100**
LD_i	71.4	80.8	97.4	99.6	100	100
GD_i	55.8	72.8	98.2	96.4	100	100
B_i	50.8	79.6	90.8	95.6	99.6	99.6

Πίνακας 4.9: Στοιχεία σχετικά με την αποτελεσματικότητα των μέτρων LD_i, GD_i και B_i

Για όλα τα μέτρα, η τιμή αυτή ξεπερνάει το πλαφόν του κάθε μέτρου σχεδόν πάντοτε. Ακόμα, τα ποσοστά όλων των μέτρων αυξάνονται ραγδαία με την αύξηση του μεγέθους του

δείγματος και οι όποιες αρχικές διαφορές υπάρχουν μεταξύ των τριών μέτρων του Πίνακα 4.9 έχουν την τάση να εξαλείφονται.

	Ποσοστό των επαναλήψεων (%), στις οποίες η τιμή του μέτρου στη θέση της άτυπης x_0 παρατήρησης είναι:					
	Η μεγαλύτερη τιμή του μέτρου			**Μεγαλύτερη από ένα πλαφόν**		
	***n* = 50**	***n*=75**	***n*=100**	***n* = 50**	***n*=75**	***n*=100**
LD_{x_0}	74.8	83.4	98.0	99.4	100	100
GD_{x_0}	74.0	82.6	97.4	99.2	100	100
LD_m	73.8	82.2	97.4	99.8	100	100
GD_m	74.0	82.2	97.2	99.8	100	100

Πίνακας 4.10: Στοιχεία σχετικά με την αποτελεσματικότητα των μέτρων LD_{x_0}, LD_m, GD_{x_0} και GD_m

Τα ποσοστά ορθής αναγνώρισης της άτυπης τιμής για τους δείκτες LD_m και GD_m είναι αναμενόμενα ελαφρώς χαμηλότερα από τα αντίστοιχα για τους δείκτες LD_{x_0} και GDx_0, καθώς η άτυπη τιμή βρίσκεται σε μία παρατήρηση για το x_0. Επίσης, τα ποσοστά αναγνώρισης της άτυπης τιμής είναι παρόμοια για τα ζευγάρια δεικτών LD_m - GD_m και LD_{x_0} - GDx_0, αποτέλεσμα αναμενόμενο, που επιβεβαιώνει και τα θεωρητικά αποτελέσματα των προηγούμενων παραγράφων. Επιπρόσθετα, οι δείκτες LDx_0 και GDx_0 εμφανίζουν ελαφρώς μεγαλύτερη ικανότητα εντοπισμού της άτυπης τιμής από τους δείκτες LD_m και GD_m. Και αυτό το αποτέλεσμα θεωρείται αναμενόμενο, τη στιγμή που η άτυπη τιμή βρίσκεται στην παράμετρο x_0. Βέβαια και για τα δύο ζευγάρια δεικτών, τα ποσοστά αναγνώρισης της άτυπης τιμής είναι ιδιαίτερα υψηλά. Τέλος, το μέτρο B_i φαίνεται να είναι αρκετά αξιόπιστο στον εντοπισμό των σημείων επιρροής, με ποσοστά αναγνώρισης της άτυπης τιμής υψηλότερα από τα υπόλοιπα μέτρα. Συνολικά, τα πιο ικανοποιητικά αποτελέσματα φαίνεται να παρουσιάζουν τα μέτρα LD_i, LD_{x_0} και LD_m.

Η περίπτωση όπου το outlier βρίσκεται στην παράμετρο *m*

Για τα επόμενα, χρησιμοποιήθηκαν τρεις μεταβλητές, όπως στην προηγούμενη περίπτωση.

Αληθινό μοντέλο:

$$\ln x_0 = 0.25 - 0.5z_1 - 0.25z_2 + 0.5z_3$$
$$m = 1 - 2z_1 + 3.5z_2 - 1.5z_3$$

Η αποτελεσματικότητα των διαφόρων μέτρων στην ορθή αναγνώριση του σημείου επιρροής σε αυτήν την περίπτωση δεν είναι ιδιαίτερα εμφανής και φθίνει με την αύξηση του μεγέθους του δείγματος.

	Ποσοστό των επαναλήψεων (%), στις οποίες η τιμή του μέτρου στη θέση της άτυπης *m* παρατήρησης είναι:					
	Η μεγαλύτερη τιμή του μέτρου			**Μεγαλύτερη από ένα πλαφόν**		
	***n* = 50**	***n*=75**	***n*=100**	***n* = 50**	***n*=75**	***n*=100**
LD_i	35.8	8.6	5.2	96.6	45.6	62.2
GD_i	39.4	6.2	5.0	98.0	45.0	61.4
B_i	13.6	20.6	15.0	33.8	67.0	55.4

Πίνακας 4.11: Στοιχεία σχετικά με την αποτελεσματικότητα των μέτρων LD_i, GD_i και B_i

Ο δείκτης B_i φαίνεται να είναι ο πιο σταθερός, παρουσιάζοντας τις μικρότερες αποκλίσεις για τα διάφορα μεγέθη δειγμάτων και για τις δύο διαφορετικές ποσότητες που μελετάμε. Βέβαια, για όλα τα μέτρα, η άτυπη *m*-τιμή αναγνωρίζεται σε ικανοποιητικό βαθμό από το πλαφόν του κάθε μέτρου, ιδιαίτερα για δείγματα μεγέθους $n = 75$ και $n = 100$.

	Ποσοστό των επαναλήψεων (%), στις οποίες η τιμή του μέτρου στη θέση της άτυπης *m* παρατήρησης είναι:					
	Η μεγαλύτερη τιμή του μέτρου			**Μεγαλύτερη από ένα πλαφόν**		
	***n* = 50**	***n*=75**	***n*=100**	***n* = 50**	***n*=75**	***n*=100**
LD_{x_0}	38.4	7.0	3.4	98.2	49.0	65.0
GD_{x_0}	36.4	7.8	3.6	96.6	48.8	65.0
LD_m	28.6	12.4	12.6	94.0	63.6	75.6
GD_m	28.0	12.0	10.8	92.8	62.6	75.0

Πίνακας 4.12: Στοιχεία σχετικά με την αποτελεσματικότητα των μέτρων LD_{x_0}, LD_m, GD_{x_0} και GD_m

Τα ποσοστά ορθής αναγνώρισης της άτυπης τιμής για τους δείκτες LD_m και GD_m είναι αναμενόμενα ελαφρώς υψηλότερα από τα αντίστοιχα για τους δείκτες LD_{x_0} και GDx_0, καθώς η άτυπη τιμή βρίσκεται σε μία παρατήρηση για το *m*. Επίσης, τα ποσοστά αναγνώρισης της άτυπης τιμής είναι παρόμοια για τα ζευγάρια δεικτών LD_i - GD_i, LD_m -

GD_m και LD_{x_0} - GDx_0, αποτέλεσμα που επιβεβαιώνει και την ορθότητα των θεωρητικών αποτελεσμάτων που παρουσιάστηκαν στις προηγούμενες παραγράφους. Επιρπόσθετα οι δείκτες LD_{x_0} και GDx_0 εμφανίζουν ελαφρώς περιορισμένη ικανότητα εντοπισμού της άτυπης τιμής σε σχέση με τους δείκτες LD_m και GD_m, αποτέλεσμα αναμενόμενο, τη στιγμή που η άτυπη τιμή βρίσκεται στην παράμετρο *m*.

Συμπεράσματα

Τα τρία προτεινόμενα μέτρα LD_i, GD_i και B_i που παρουσιάσαμε, κρίνονται ικανοποιητικά για τον εντοπισμό σημείων επιρροής, με το τελευταίο να είναι το πιο αξιόπιστο για την περίπτωση που δεν υπάρχουν μεταβλητές στο μοντέλο, καθώς εμφανίζει τα υψηλότερα ποσοστά ορθής αναγνώρισης της άτυπης τιμής. Για τις άλλες δύο περιπτώσεις που μελετήθηκαν, φαίνεται πως το πιο αξιόπιστο μέτρο είναι ο δείκτης LD_i, ο οποίος είναι συνεπής στην ορθότητα αναγνώρισης της άτυπης τιμής, σε όποια παράμετρο και αν βρίσκεται αυτή. Τα ποσοστά ορθής αναγνώρισης της άτυπης τιμής από τα διάφορα μέτρα αυξάνονται με την αύξηση του μεγέθους του δείγματος. Ακόμα, τα διάφορα μέτρα φαίνεται να είναι συνολικά πιο αποτελεσματικά στην περίπτωση που η άτυπη τιμή βρίσκεται στην παράμετρο x_0. Ωστόσο αυτό το αποτέλεσμα ίσως να ωφείλεται και στο μέγεθος της άτυπης τιμής, αλλά και από την επιλογή του διανύσματος των συντελεστών στο αληθινό μοντέλο, όπως δείχνει η πρώτη περίπτωση που μελετήσαμε, του μοντέλου IG GLM.

Οι δείκτες LD_m και GD_m, καθώς και οι δείκτες LD_{x_0} και GD_{x_0}, έχουν πάντοτε σχεδόν ίδια τιμή. Το γεγονός αυτό επιβεβαιώνει την ορθότητα των διαφόρων θεωρητικών υπολογισμών, καθώς για την κατασκευή τους έχει ακολουθηθεί διαφορετική προσεγγιστική διαδικασία.

4.4.4 Εφαρμογή σε πραγματικά δεδομένα

Ως συνέχεια της μελέτης, παραθέτουμε την εφαρμογή των διαφόρων μέτρων της επιρροής σε ένα μικρό σύνολο δεδομένων, αποτελούμενο από δέκα μη-αποκομμένες παρατηρήσεις. Οι παρατηρήσεις αυτές αφορούν ένα πείραμα που έγινε στην Αμερικανική αεροπορική βάση Wright-Patterson (Brown και Potts, 1977). Σκοπός της μελέτης ήταν να αξιολογήσει τις διάρκειες ζωής κάποιων ειδικών ρουλεμαν που χρησιμοποιούνται σε υψηλής ταχύτητας τουρμπίνες αεροσκαφών, ως προς την αντοχή τους σε κυκλικές περιστροφές μέσω συγκεκριμένων μεθόδων.

Οι παρατηρήσεις του Πίνακα 4.13 αφορούν διάρκειες ζωής των ρουλεμάν (μετρημένες σε εκατομμύρια κυκλικές περιστροφές) σε μία από αυτές τις μεθόδους.

t_i	5.88	6.74	6.90	6.98	7.21	8.14	8.59	9.80	12.28	**25.46**

Πίνακας 4.13: Χρόνοι για το πρόβλημα

Αρχικά, παραθέτουμε τον Πίνακα 4.14, στον οποίο δίνονται οι ποσότητες $\hat{x}_{0(i)}$ και $\hat{m}_{(i)}$, από τις οποίες φαίνεται πως η τελευταία παρατήρηση διαφέρει αισθητά από τις υπόλοιπες.

A/A	$\hat{x}_{0(i)}$	$\hat{m}_{(i)}$	**A/A**	$\hat{x}_{0(i)}$	$\hat{m}_{(i)}$
1	2.02	-0.74	**6**	1.95	-0.70
2	1.98	-0.72	**7**	1.94	-0.70
3	1.97	-0.71	**8**	1.94	-0.71
4	1.97	-0.71	**9**	1.95	-0.74
5	1.97	-0.71	**10**	**2.59**	**-1.65**

Πίνακας 4.14: Τιμές των δεικτών, $\hat{x}_{0(i)}$ και $\hat{m}_{(i)}$

Επίσης, είναι εμφανές πως και οι τρεις δείκτες του Πίνακα 4.15 αναγνωρίζουν την τελευταία παρατήρηση ως πιθανό σημείο επιρροή, με το δείκτη LD_i να εμφανίζει τη μεγαλύτερη διακριτική ικανότητα από τους άλλους δύο. Εντυπωσιακό το γεγονός πως η τιμή του συγκεκριμένου δείκτη στη θέση που βρίσκεται η άτυπη τιμή είναι έως και 300 μεγαλύτερη από την τιμή του δείκτη στις υπόλοιπες παρατηρήσεις του δείγματος. Ακόμα, να παρατηρήσουμε πως τα μέτρα LD_i και GD_i έχουν σχεδόν ίδια τιμή σε όλες τις παρατηρήσεις, εκτός από την τελευταία, στην οποία βρίσκεται η άτυπη τιμή. Στη θέση αυτή η τιμή του LD_i είναι περισσότερο από δύο φορές μεγαλύτερη από την αντίστοιχη τιμή του GD_i, γεγονός που υποδεικνύει την καλύτερη διακριτική ικανότητα του πρώτου δείκτη. Να παρατηρήσουμε ωστόσο πως και η τιμή που λαμβάνει ο δείκτης GD_i (7.9429) είναι έως και 100 φορές μεγαλύτερη από τις υπόλοιπες τιμές του ίδιου δείκτη στο δείγμα.

A/A	LD_i	GD_i	B_i	**A/A**	LD_i	GD_i	B_i
1	0.1256	0.1339	0.0443	**6**	0.0519	0.0532	0.0099
2	0.0619	0.0622	0.0289	**7**	0.0527	0.0529	0.0058
3	0.0580	0.0586	0.0264	**8**	0.0536	0.0499	0.0002
4	0.0565	0.0573	0.02515	**9**	0.0685	0.0643	0.0166
5	0.0536	0.0549	0.0217	**10**	**19.347**	**7.9429**	**0.8213**

Πίνακας 4.15: Τιμές των δεικτών LD_i, GD_i και B_i

Ομοίως, και οι τρεις δείκτες του Πίνακα 4.16 αναγνωρίζουν την τελευταία παρατήρηση ως πιθανό σημείο επιρροής. Ακόμα, να παρατηρήσουμε πως οι τιμές των ποσοτήτων LD_i και GD_i βρίσκονται αναμενόμενα πολύ κοντά μεταξύ τους. Παρόμοια συμπεριφορά έχουμε και για τα ζευγάρια δεικτών LD_m - GD_m και LD_{x_0} - GD_{x_0}.

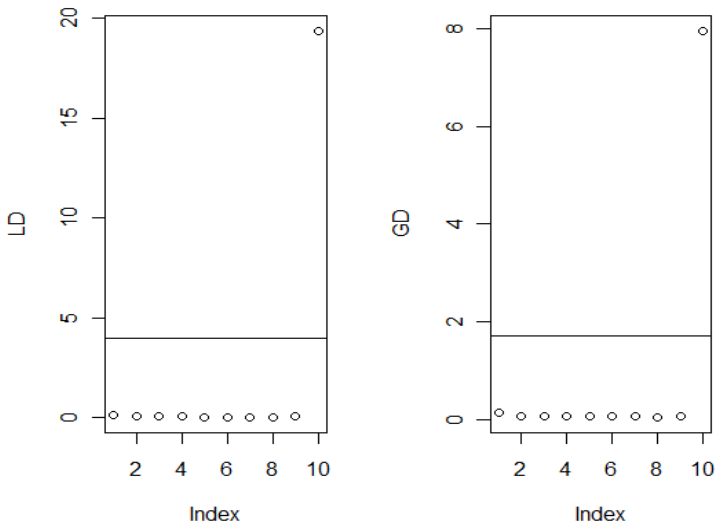
A/A	LD_{x_0}	GD_{x_0}	LD_m	GD_m
1	0.0203	0.0199	0.0021	0.0021
2	0.0017	0.0017	0.0027	0.0264
3	0.0043	0.0043	0.0312	0.0309
4	0.0059	0.0059	0.0334	0.0332
5	0.0112	0.0114	0.0391	0.0388
6	0.0348	0.0358	0.0517	0.0512
7	0.04366	0.0451	0.0519	0.0514
8	0.0536	0.0556	0.0388	0.0385
9	0.0259	0.02655	0.0009	0.0008
10	**10.9677**	**9.6761**	**1.8204**	**2.4773**

Πίνακας 4.16: Τιμές των δεικτών LD_{x_0}, LD_m, GD_{x_0} και GD_m

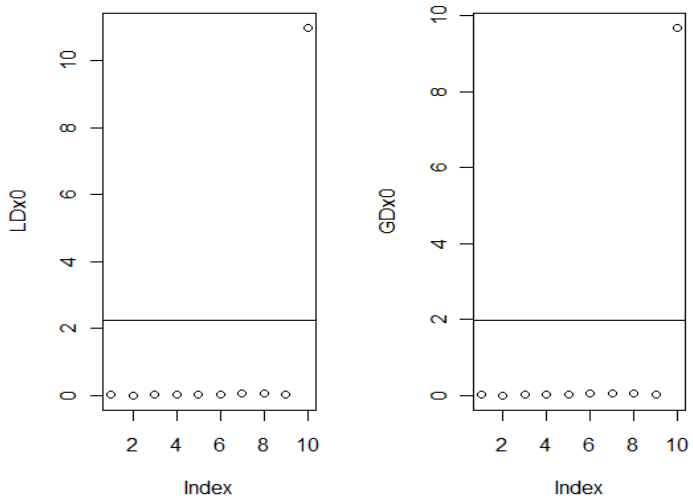
Στα Σχήματα 4.1 έως 4.4, παρουσιάζουμε και τα αντίστοιχα γραφήματα δεικτών (index plots) για τα παραπάνω μέτρα επιρροής. Σε αυτά, δίνεται και η αντίστοιχη τιμή πλαφόν για τον κάθε δείκτη, η οποία παρουσιάζεται στον Πίνακα 4.17.

LD_i	3.9858	GD_i	1.7061
LD_{x_0}	2.2338	GD_{x_0}	1.9765
LD_m	3.6959	GD_m	5.0093
B_i	0.2001		

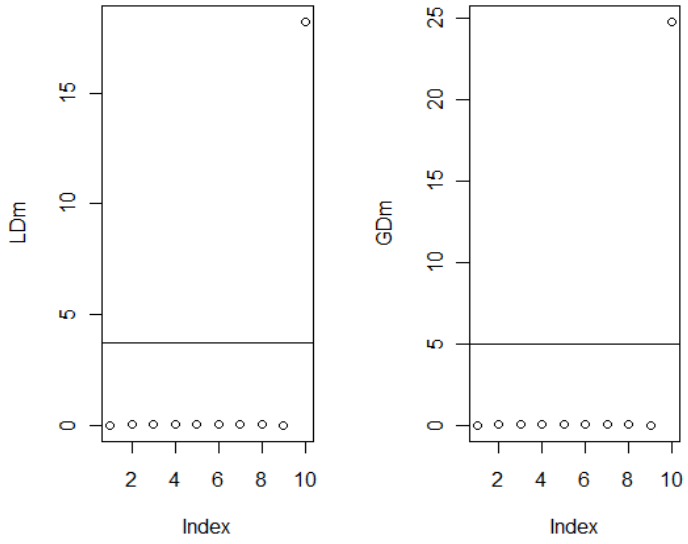
Πίνακας 4.17: Πίνακας με τις τιμές των πλαφόν για το εκάστοτε μέτρο επιρροής



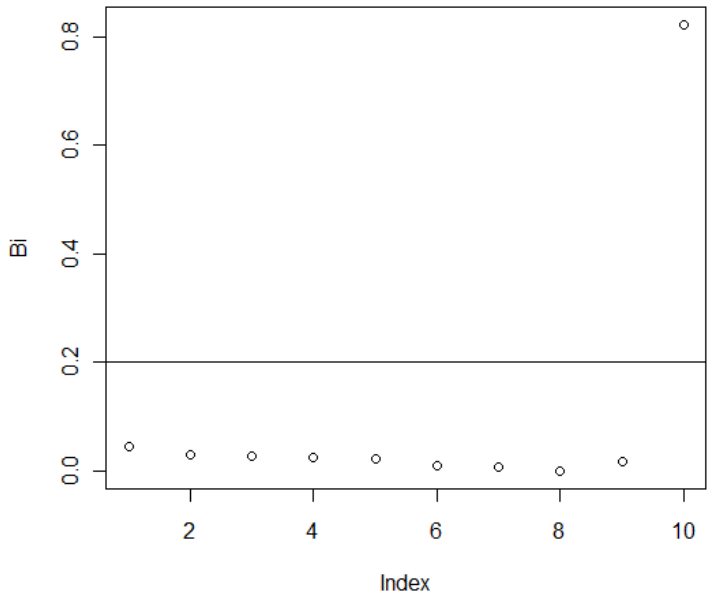
Σχήμα 4.1: Γράφημα με τις τιμές των δεικτών LD_i και GD_i



Σχήμα 4.2: Γράφημα με τις τιμές των δεικτών LD_{x_0} και GD_{x_0}



Σχήμα 4.3: Γράφημα με τις τιμές των δεικτών LD_m και GD_m



Σχήμα 4.4 Γράφημα με τις τιμές του δείκτη B_i

Οι διάφοροι θεωρητικοί αριθμητικοί υπολογισμοί επιβεβαιώνονται και από τα αντίστοιχα γραφήματα δεικτών για τα διάφορα μέτρα επιρροής: Η δέκατη παρατήρηση αναγνωρίζεται από τα Σχήματα 4.1 έως 4.4 ως πιθανό σημείο επιρροής.

Κεφάλαιο 5

Συμπεράσματα και μελλοντική δουλειά στο FHTR μοντέλο

Όπως είδαμε και στα προηγούμενα κεφάλαια της διατριβής, τα μοντέλα FHTR δεν αποτελούν απλώς παραλλαγή προσεγγίσεων που υπάρχουν ήδη στη βιβλιογραφία, με θεωρητικό μόνο ενδιαφέρον. Είναι μία ριζικά διαφορετική προσέγγιση, καθώς τα μοντέλα αυτά έχουν διαφορετική συμπεριφορά από τις άλλες προσεγγίσεις που χρησιμοποιούνται συνήθως (Lee et al., 2010). Επιπρόσθετα, μπορούν να χρησιμοποιηθούν και σε περιπτώσεις όπου άλλα μοντέλα είναι ακατάλληλα, όπως για παράδειγμα τα μοντέλα αναλογικής διακινδύνευσης είναι κατάλληλα μόνο για τις περιπτώσεις εκείνες στις οποίες η συνάρτηση διακινδύνευσης είναι μονοτονική.

Από τα παραπάνω, αποδεικνύεται ότι η σπουδαιότητα των FHT μοντέλων ξεπερνάει το ακαδημαϊκό ενδιαφέρον που συγκεντρώνουν, με την προσθήκη σημαντικών εργαλείων για την ανάλυση δεδομένων σε πολλές επιστημονικές περιοχές. Στα προηγούμενα κεφάλαια της διατριβής έγινε αναλυτική παρουσίαση του FHTR μοντέλου, για την περίπτωση όπου η γονική στοχαστική ανέλιξη είναι τύπου Wiener και κατά συνέπεια ο χρόνος πρώτης μετάβασης της ανέλιξης ακολουθεί την IG.

Στην υπόλοιπη παράγραφο παρουσιάζουμε κάποια συνοπτικά συμπεράσματα που προέκυψαν από τη διατριβή. Επιπρόσθετα, ακολουθούν δύο θεματικές ενότητες, οι οποίες θα αποτελέσουν σε ερευνητικό επίπεδο τη φυσική συνέχεια της παρούσας διατριβής. Θα γίνει μία περιγραφή κάποιων στοχαστικών ανελίξεων που μπορεί να χρησιμοποιηθούν εναλλακτικά με την ανέλιξη Wiener στο μοντέλο FHTR και θα παρουσιαστούν οι αντίστοιχες κατανομές των χρόνων πρώτης διακοπής για την κάθε μία ανέλιξη. Τέλος, εισάγεται αναλυτικά η έννοια των επαναλαμβανομένων δεδομένων για τη μονάδα και

παρουσιάζονται δομές για δεδομένα διάρκειας ζωής με επαναλαμβανόμενα γεγονότα για το FHTR μοντέλο.

5.1 Συμπεράσματα

Η αντίστροφη Γκαουσιανή κατανομή είναι το πιο σπάνια χρησιμοποιούμενο γενικευμένο γραμμικό μοντέλο που υπάρχει στη βιβλιογραφία. Αρχικά, έγινε μία εκτενής μελέτη της κατανομής IG και μελετήθηκαν γενικές ιδιότητες που συμπεριλαμβάνουν την IG ως γενικευμένο γραμμικό μοντέλο.

Στηριζόμενοι στο έργο των Myers και Montgomery (1997) και των Lewis et al. (2001a και b) μελετήσαμε την επίδραση που προκαλεί μία λανθασμένη επιλογή συνάρτησης σύνδεσης στις εκτιμήσεις της κάλυψης και της ακρίβειας (μήκος) ενός διαστήματος εμπιστοσύνης για την παράμετρο της μέσης τιμής της IG κατανομής. Είδαμε πως όταν η συνάρτηση σύνδεσης του αληθινού μοντέλου είναι η κανονική, ενδεχόμενη λανθασμένη επιλογή για τη συνάρτηση σύνδεσης κατά την προσαρμογή του IG GLM δεν επηρεάζει τα αποτελέσματα της προσαρμογής. Τα διαστήματα εμπιστοσύνης για την παράμετρο μ ήταν ιδιαίτερα μικρά σε μήκος για τις διαφορετικές συναρτήσεις σύνδεσης που ελέγχθηκαν. Με διαφορετική διατύπωση, η προσαρμογή φαίνεται να είναι κάθε φορά αρκετά ακριβής. Τα αποτελέσματα ισχυροποιούνται με την αύξηση του μεγέθους του δείγματος.

Παρατηρήσαμε επίσης πως σε ένα IG GLM είναι πολύ σημαντική η επιλογή της κατάλληλης συνάρτησης σύνδεσης, ιδιαίτερα όταν αυτή δεν είναι η κανονική. Όταν η συνάρτηση σύνδεσης του αληθινού μοντέλου είναι η λογαριθμική, οι πιθανότητες κάλυψης της παραμέτρου μ μειώνονται αισθητά για τα μοντέλα με τις άλλες τρείς συναρτήσεις σύνδεσης, τόσο για μεμονωμένες παρατηρήσεις όσο και συνολικά. Στην περίπτωση που η συνάρτηση σύνδεσης είναι η λογαριθμική, τα μήκη των διαστημάτων εμπιστοσύνης για την παράμετρο μ είναι μεγαλύτερα από την περίπτωση της κανονικής συνάρτησης σύνδεσης, με αποτέλεσμα να μειώνεται η ακρίβεια της προσαρμογής. Τα αποτελέσματα φαίνεται να ισχυροποιούνται με την αύξηση του μεγέθους του δείγματος.

Στη συνέχεια, μελετήθηκαν τα διάφορα υπόλοιπα που μπορούν να χρησιμοποιηθούν για την IG κατανομή και κατασκευάστηκαν αλγόριθμοι για την παραγωγή τους μέσω του στατιστικού πακέτου R. Ειδικότερα, ερευνήθηκε η σχέση μεταξύ των Pearson, Anscombe και Deviance υπολοίπων, τα οποία χρησιμοποιούνται ευρύτατα τόσο σε γραμμικά όσο και σε μη γραμμικά μοντέλα. Αποδείχθηκε πως σε ειδικές περιπτώσεις, τα υπόλοιπα Anscombe έχουν τιμές πολύ κοντινές με τις αντίστοιχες των Deviance υπολοίπων στην IG. Επίσης, φαίνεται πως τα Pearson ίσως αποτελούν την καλύτερη επιλογή από τα τρία διαθέσιμα υπόλοιπα για το συγκεκριμένο μοντέλο. Ακόμα δεκτή επιλογή για την περίπτωση ενός γενικευμένου γραμμικού μοντέλου για την κατανομή IG αποτελούν και τα martingale

υπόλοιπα τύπου deviance, r_{D_i}. Αντιθέτως, τα υπόλοιπα τύπου martingale, r_{M_i}, και τα υπόλοιπα Cox-Snell φαίνεται να μην είναι κατάλληλα για την κατανομή IG.

Βασισμένοι στο έργο του Chhikara (1975) και της Davis (1980), κατασκευάσαμε ελέγχους για τον εντοπισμό άτυπων τιμών των παραμέτρων (μ, λ) της IG σε δεδομένα ανεξάρτητων μονάδων με επαναλαμβανόμενα ανεξάρτητα γεγονότα ανά μονάδα. Οι διάφοροι έλεγχοι βασίστηκαν στη μεγιστοποίηση της τιμής του ελέγχου του λόγου των πιθανοφανειών για την ισότητα των παραμέτρων, με διόρθωση Bonferroni για τις *p*-τιμές. Επιβεβαιώθηκε η ακρίβεια των Bonferroni ελέγχων υπό τη μηδενική υπόθεση και μελετήθηκε η ισχύς των ελέγχων υπό την εναλλακτική υπόθεση. Τα πραγματικά μεγέθη των ελέγχων είναι πολύ κοντά στις θεωρητικές τιμές, με μία μικρή μόνο τάση προς συντηρητική συμπεριφορά και για τις δύο περιπτώσεις των παραμέτρων μ και λ. Ο έλεγχος για μία άτυπη τιμή για την παράμετρο λ συμπεριφέρεται με τον ίδιο ακριβώς τρόπο όπως ο έλεγχος για την παράμετρο μ κάτω από τη μηδενική υπόθεση, στην περίπτωση που όλες οι μονάδες έχουν τον ίδιο αριθμό επαναλαμβανόμενων γεγονότων. Στην περίπτωση που οι μονάδες δεν έχουν τον ίδιο αριθμό επαναλαμβανόμενων γεγονότων, εκτός από το δειγματοληπτικό σφάλμα και την εκτίμηση που εισήχθη από την προσέγγιση Bonferroni, περαιτέρω ανακρίβεια εισάγεται λόγω της χρησιμοποίησης ενός προσεγγιστικού μετασχηματισμού από την *F* στην τυποποιημένη Κανονική κατανομή. Τα διάφορα αποτελέσματα υποδεικνύουν πως ο έλεγχος είναι ακόμα πιο συντηρητικός κάτω από αυτές τις περιστάσεις, αλλά και πάλι όχι υπερβολικά πάνω από τα επιτρεπτά όρια.

Ακόμα, παρουσιάστηκαν διάφορες καμπύλες που δείχνουν την ισχύ του ελέγχου για μια άτυπη τιμή για τις παραμέτρους μ και λ υπό την εναλλακτική υπόθεση και παρουσιάστηκαν οι περιπτώσεις εκείνες, στις οποίες η ισχύς του ελέγχου αυξάνεται. Αποδείχτηκε πως για συγκεκριμένες επιλογές μονάδων και παρατηρήσεων, η ισχύς αυξάνει όσο ο λόγος λ/μ αυξάνει. Μικρές τιμές της παραπάνω αναλογίας υποδεικνύουν έντονη ασυμμετρία της IG κατανομής, λόγος για τον οποίο ενδέχεται να είναι ιδιαίτερα δύσκολος ο εντοπισμός άτυπων τιμών. Στη συνέχεια, χρησιμοποιήσαμε μία εναλλακτική παραμέτρηση για τη συνάρτηση πυκνότητας πιθανότητας της IG, η οποία βρίσκεται σε αντιστοιχία με τις παραμέτρους (μ, λ) και χρησιμοποιείται στην παλινδρόμηση Κατωφλιού.

Στη συνέχεια, παρουσιάστηκε αναλυτικά το μοντέλο της παλινδρόμησης Κατωφλιού (FHTR) και συγκρίναμε το μοντέλο του Cox με ένα μοντέλο FHT παλινδρόμησης βασισμένο σε ανέλιξη Wiener, το οποίο οδηγεί σε χρόνο πρώτης μετάβασης που ακολουθεί την IG κατανομή. Δείξαμε πως τα μοντέλα FHT εμφανίζουν σημαντικά πλεονεκτήματα σε σχέση με τα PH μοντέλα, γεγονός που τα μετατρέπει σε πιο ρεαλιστικό εργαλείο σε πολλές εφαρμογές. Είδαμε πως τα FHT μοντέλα προσφέρουν την ευκαιρία μίας πιο αποκαλυπτικής μοντελοποίησης, η οποία προχωράει πέρα από μία απλή περιγραφή του προβλήματος. Ακόμα, μελετήσαμε την περίπτωση που ένα IG μοντέλο προσαρμόστηκε σε δεδομένα τα

οποία ικανοποιούν την υπόθεση PH της αναλογικής διακινδύνευσης και αντιστρόφως και προέκυψε πως τα δύο μοντέλα αναγνώρισαν ως σημαντικές διαφορετικές μεταβλητές. Επιπρόσθετα, οι μεταβλητές που αναγνωρίστηκαν ως σημαντικές και στα δύο μοντέλα, δεν είχαν απαραίτητα τα ίδια πρόσημα στη διακινδύνευση, με αποτέλεσμα να έχουν συνεισφορά προς διαφορετικές κατευθύνσεις.

Τα προβλήματα που ανέδειξε η σύγκριση των δύο μοντέλων, καθώς και οι διάφορες αναφορές σε δημοσιευμένες εργασίες για την ανάγκη ανάπτυξης διαγνωστικών τεχνικών για την καταλληλότητα του μοντέλου, οδήγησαν στην περαιτέρω διερεύνηση κάποιων πρακτικών θεμάτων κατά την προσαρμογή του μοντέλου παλινδρόμησης χρόνου πρώτης μετάβασης (IG FHTR). Εμπειρικές αποδείξεις δόθηκαν σχετικά με τη δυνατότητα προσαρμογής του μοντέλου. Ειδικότερα, εξετάστηκε εάν υπάρχει κάποια ένδειξη κατά τη διαδικασία προσαρμογής να τοποθετείται μία μεταβλητή σε λάθος παράμετρο.

Τα αποτελέσματα της μελέτης έδειξαν πως οι εκτιμήσεις των συντελεστών της παλινδρόμησης συμπεριφέρονται αρκετά καλά κάτω από συγκεκριμένες προϋποθέσεις. Τα μεγέθη των ελέγχων Wald και του λόγου των πιθανοφανειών LR, είναι σωστά ακόμα και για σχετικά μικρά μεγέθη δείγματος και δε φαίνεται να υπάρχει σημαντική μεροληψία. Ωστόσο, τα διάφορα αποτελέσματα επιβεβαιώνουν την ύπαρξη κάποιας εξάρτησης ανάμεσα στις εκτιμήσεις των διάφορων συντελεστών. Συγκεκριμένα, όταν μία μεταβλητή επιδρά ταυτόχρονα και στις δύο παραμέτρους προς την ίδια κατεύθυνση, υπάρχει μία τάση μόνο για τον ένα από τους δύο συντελεστές της να αναγνωριστεί ως στατιστικά σημαντικός. Αντιθέτως, όταν η μεταβλητή επιδρά ταυτόχρονα και στις δύο παραμέτρους προς διαφορετικές κατευθύνσεις, υπάρχει μία τάση και για τους δύο συντελεστές της να αναγνωριστούν ως στατιστικά σημαντικοί.

Επιπρόσθετα, μελετήσαμε το φαινόμενο εμφάνισης αντίθετων προσήμων μίας μεταβλητής στις διάφορες παραμέτρους της κατανομής και δείξαμε πως όταν μία μεταβλητή έχει ισχυρή επίδραση σε μία παράμετρο, υπάρχει μία τάση να εμφανίζει μία επίδραση προς την ίδια κατεύθυνση (με το ίδιο πρόσημο) και στην άλλη παράμετρο. Το φαινόμενο αυτό ενισχύεται με την αύξηση του μεγέθους του δείγματος.

Τέλος, μελετήσαμε την επίδραση ενός λανθασμένα υιοθετημένου FHT μοντέλου, όταν θα έπρεπε να έχει επιλεχτεί για την προσαρμογή των δεδομένων κάποιο διαφορετικό μοντέλο. Φάνηκε πως το εύρημα αντίθετων κατευθύνσεων για τις δύο επιδράσεις της ίδιας μεταβλητής ίσως εμφανίζεται κυρίως όταν γίνεται εσφαλμένη προσαρμογή του FHT μοντέλο. Αυτό το συμπέρασμα με τη σειρά του ισχυροποίησε την αναγκαιότητα κατασκευής διαγνωστικών ελέγχων και ελέγχων καλής προσαρμογής για το FHT μοντέλο.

Στη συνέχεια, προτείναμε μία διαδικασία επιλογής μεταβλητών για την περίπτωση του IG FHT μοντέλου παλινδρόμησης, η οποία αποτελείται από δύο διαδοχικές εφαρμογές της προσαρμοσμένης LASSO τεχνικής (adaptive LASSO) εκτελεσμένες από έναν αλγόριθμο

ελαχίστων τετραγώνων. Η διαδικασία αποδείχθηκε αποτελεσματική για την ορθή αναγνώριση των μη μηδενικών (στατιστικά σημαντικών) συντελεστών της παλινδρόμησης. Η μελέτη αυτή αποτελεί την πρώτη συνδρομή στη μεθοδολογία μοντελοποίησης, η οποία είναι απαραίτητη να αναπτυχθεί περαιτέρω για το παρόν μοντέλο παλινδρόμησης.

Επίσης, ασχοληθήκαμε με τον εντοπισμό σημείων επιρροής για την περίπτωση του IG FHT μοντέλου παλινδρόμησης, όπου μέχρι τώρα δεν υπάρχει κάποια παρόμοια τεχνική για την παλινδρόμηση Κατωφλιού. Κατά τη διάρκεια της διατριβής αναπτύξαμε μία μέθοδο βασισμένη στην τεχνική αφαίρεσης σημείου (Case Deletion Method - CDM), προκειμένου να μετρήσουμε την επιρροή της καθεμιάς παρατήρησης ξεχωριστά. Τέλος, προσαρμόσαμε για το FHT μοντέλο παλινδρόμησης μία τεχνική που είχε προτείνει ο Cook για τη μέτρηση της τοπικής επιρροής στο μοντέλου. Τα διάφορα θεωρητικά αποτελέσματα επιβεβαιώθηκαν με επιτυχία μέσα από παραδείγματα αριθμητικών δεδομένων με τα τρία βασικά προτεινόμενα μέτρα που προέκυψαν να κρίνονται ως ικανοποιητικά για τον εντοπισμό σημείων επιρροής, Για την περίπτωση του IG FHTR μοντέλου με συμμεταβλητές το πιο αξιόπιστο μέτρο είναι ο δείκτης LD_i, ο οποίος είναι συνεπής στην ορθότητα αναγνώρισης της άτυπης τιμής, σε όποια παράμετρο και αν βρίσκεται. Ακόμα, φάνηκε πως τα ποσοστά ορθής αναγνώρισης της άτυπης τιμής από τα διάφορα μέτρα αυξάνονται με την αύξηση του μεγέθους του δείγματος και πως τα διάφορα μέτρα φαίνεται να είναι συνολικά πιο αποτελεσματικά στην περίπτωση που η άτυπη τιμή βρίσκεται στην παράμετρο x_0.

5.2 Κατανομές για το FHTR μοντέλο

Στη βιβλιογραφία υπάρχει ποικιλία ανελίξεων, οι οποίες μπορεί να χρησιμοποιηθούν ως γονικές για το FHTR μοντέλο παλινδρόμησης. Στα προηγούμενα κεφάλαια της διατριβής ασχοληθήκαμε αποκλειστικά με την περίπτωση όπου η ανέλιξη-γονέας είναι τύπου Wiener και ο χρόνος πρώτης μετάβασης ακολουθεί την κατανομή IG.

Ωστόσο, εκτός από το μοντέλο που βασίζεται στην ανέλιξη Wiener, ένα ακόμα μοντέλο που χρησιμοποιείται στην πράξη φαίνεται να είναι εκείνο που προκύπτει από την ανέλιξη Γάμμα, η οποία διαφέρει σημαντικά από τη Wiener λόγω των μη αρνητικών προσαυξήσεων (Dufresne et al., 1991). Το συγκεκριμένο μοντέλο, προσφέρει ένα ελκυστικό εργαλείο για την περιγραφή της αύξησης της σοβαρότητας ή του μεγέθους ενός ελαττώματος, όπως είναι για παράδειγμα η εξέλιξη μίας ρωγμής σε ένα δοκάρι μέχρι κατάρρευσης, ή γενικώς μέχρι τη διακοπή της λειτουργίας. Σχετικές δημοσιεύσεις περιλαμβάνουν οι Lawless και Crowder (2004).

Εκτός από την ανέλιξη Γάμμα, υπάρχει ποικιλία FHT μοντέλων ανάλογα με τη μορφή της υποβόσκουσας στοχαστικής ανέλιξης, όπως η Erlang, η Poisson και η Ricciardo-Sato (Lee και Whitmore, 2006). Μελλοντικός στόχος είναι να εργαστούμε για την περίπτωση των

ανελίξεων αυτών. Στη συνέχεια, παρουσιάζουμε αυτές τις περιπτώσεις μέσα από απλά, αντιπροσωπευτικά παραδείγματα.

5.2.1 Ανέλιξη Γάμμα και χρόνος πρώτης μετάβασης που ακολουθεί την αντίστροφη Γάμμα

Ας θεωρήσουμε μία γονική στοχαστική ανέλιξη $\{X(t), t \geq 0\}$, με σημείο εκκίνησης $X(0) = x_0 > 0$. Έστω $X(t) = x_0 - Z(t)$, όπου $\{Z(t), t \geq 0\}$ είναι μία στοχαστική ανέλιξη τύπου Γάμμα, με παράμετρο κλίμακας ξ, παράμετρο σχήματος $n(t)$ και $Z(0) = 0$.

Ο χρόνος πρώτης μετάβασης, S, της γονικής ανέλιξης στο σημείο μηδέν $(X(t) = 0)$, ακολουθεί την αντίστροφη Γάμμα κατανομή, η οποία ορίζεται από την ισότητα:

$$P(S > t) = P(Z(t) < x_0)$$

Η ισότητα προκύπτει από το γεγονός πως η ανέλιξη Γάμμα έχει μονότονα (μη - φθίνοντα) μονοπάτια. Πράγματι, έστω $\{Z(t), t \geq 0\} \sim Ga(\xi, n(t))$, με $n(t) = nt$ και $Z(0) = 0$. Τότε:

$$P(S > t) = P(X(t) > 0) = P(x_0 - Z(t) > 0) = P(x_0 > Z(t)) = P(Z(t) < x_0)$$

ή

$$P(S \leq t) = P(X(t) \leq 0) = P(x_0 - Z(t) \leq 0) = P(Z(t) \geq x_0) = P\left(\frac{1}{Z(t)} \leq \frac{1}{x_0}\right) \Rightarrow$$

$$P(S \leq t) = P\left(Z^*(t) \leq \frac{1}{x_0}\right), \text{ με } Z^*(t) = \frac{1}{Z(t)}.$$

Δηλαδή, $Z^*(t) \sim$ Inv. Gamma. Αλλά, η συνάρτηση επιβίωσης είναι:

$$S(t) = P(S > t) = P(Z(t) < x_0) = P(Ga(\xi, nt) < x_0)$$

Στα διάφορα στατιστικά πακέτα, υπάρχουν έτοιμες υπολογιστικές ρουτίνες για τη σ.π.π. της Γάμμα κατανομής, γεγονός που καθιστά εύκολη στον υπολογισμό τη σ.π.π. της S.

Ο Singpurwalla (1995) και οι Lawless και Crowder (2004) θεωρούν την ανέλιξη Γάμμα ως μοντέλο για την περιγραφή της υποβάθμισης ενός συστήματος. Οι Park και Padgett (2005) θεωρούν μαζί τη γεωμετρική κίνηση Brown και την ανέλιξη Γάμμα σε ένα μοντέλο επιταχυνόμενης χειροτέρευσης.

5.2.2 Ανέλιξη Poisson και χρόνος πρώτης μετάβασης που ακολουθεί την κατανομή Erlang

Ο χρόνος *S* έως την πραγματοποίηση του *m* γεγονότος σε μία ανέλιξη Poisson $\{N(t), t \geq 0\}$ με παράμετρο λ, αποδεικνύεται ότι ακολουθεί την κατανομή Erlang, με παραμέτρους *m* και *λ*. Για να προχωρήσουμε στη συνηθισμένη αναπαραμέτρηση, θεωρούμε μία γονική ανέλιξη $\{X(t), t \geq 0\}$, με αρχική τιμή $X(0) = x_0 = m$. Έστω επίσης ότι $X(t) = x_0 - N(t)$, όπου $\{N(t), t \geq 0\}$ είναι η ανέλιξη Poisson που ορίσαμε προηγουμένως. Ο χρόνος πρώτης μετάβασης είναι η χρονική στιγμή $t = S$, όπου $X(t) = 0$. Το συγκεκριμένο FHT μοντέλο μπορεί να χρησιμοποιηθεί για να περιγράψει το χρόνο μέχρι την αποτυχία ενός μηχανικού συστήματος που αποτελείται από *m* εξαρτήματα τοποθετημένα παράλληλα, με ανεξάρτητες και ισόνομες διάρκειες ζωής που ακολουθούν την Εκθετική κατανομή και τοποθετούνται διαδοχικά σε λειτουργία, όταν παρουσιάζονται αποτυχίες.

5.2.3 Ανέλιξη Ornstein–Uhlenbeck και χρόνος πρώτης μετάβασης τύπου Ricciardi–Sato

Η ανέλιξη τύπου Ornstein–Uhlenbeck (OU) αποτελεί μία παραλλαγή της ανέλιξης Wiener, με τάση να επιστρέφει προς ένα σταθερό σημείο ισορροπίας, με αποτέλεσμα να έχει την ιδιότητα της ομοιόστασης. Οι Aalen και Gjessing (2001) πρότειναν την κατανομή του χρόνου πρώτης μετάβασης για μία τέτοια ανέλιξη ως ένα μοντέλο διάρκειας ζωής και παρήγαγαν συναφή αποτελέσματα. Επιπρόσθετα, σημειώνουν πως η μορφή της κατανομής του χρόνου πρώτης μετάβασης συναντάται στους Ricciardi και Sato (1988), οι οποίοι την έχουν ερευνήσει σε βάθος. Τελευταία, οι Erich και Pennell (2012) χρησιμοποιούν μοντέλα παλινδρόμησης Κατωφλιού στα οποία η ανέλιξη της διάρκειας ζωής είναι τύπου Ornstein-Uhlenbeck.

5.3 Επαναλαμβανόμενα δεδομένα στην παλινδρόμηση Κατωφλιού

Στην επιστήμη και την τεχνολογία, υπάρχει ενδιαφέρον για την κατανόηση των μηχανισμών, οι οποίοι είναι υπεύθυνοι για τη δημιουργία παρατηρήσεων που εμφανίζονται με κάποιο μοτίβο επανάληψης στη διάρκεια του χρόνου. Στους μηχανισμούς αυτούς, αναφερόμαστε ως ανελίξεις επαναλαμβανόμενων δεδομένων και τα δεδομένα που απορρέουν από αυτές, καλούνται επαναλαμβανόμενα γεγονότα.

Δομές επαναλαμβανόμενων δεδομένων εμφανίζονται συχνά σε πεδία όπως η Φαρμακευτική και η Δημόσια Υγεία, στις Επιχειρήσεις και τη Βιομηχανία, την Αξιοπιστία, τις Κοινωνικές Επιστήμες και στον κλάδο των Ασφαλίσεων. Η βιβλιογραφία της στατιστικής ανάλυσης των επαναλαμβανόμενων παρατηρήσεων έχει αυξηθεί ραγδαία τα τελευταία

χρόνια και μία μεγάλη ποικιλία μοντέλων και μεθόδων έχει αναπτυχθεί. Παραδείγματα, περιλαμβάνουν επιληπτικές κρίσεις στη Νευρολογία, κατάγματα στην Ορθοπεδική, δημιουργία όγκων στην Ογκολογία. Στο χώρο των επιχειρήσεων ένα παράδειγμα αφορά τη συμπλήρωση αιτήσεων για αξιώσεις εγγύησης σε τροχαία ατυχήματα, ενώ στο χώρο της Βιομηχανίας παράδειγμα αποτελούν οι διαδοχικές διακοπές λειτουργίας ενός συστήματος.

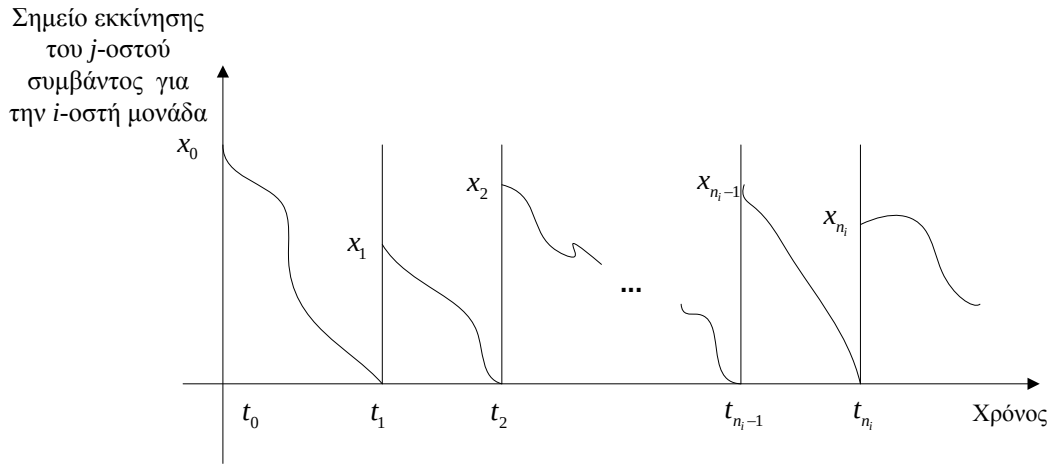
Οι γενικές μέθοδοι ανάλυσης που έχουν αναπτυχθεί, περιγράφονται από τους Cook & Lawless (2007). Στις τεχνολογικές επιστήμες, η ανάλυση επισκευάσιμων συστημάτων έχει μακρά ιστορία αλλά όχι σε σχέση με συμμεταβλητές πέρα από την απλή μορφή της σύγκρισης δυο ομάδων μηχανημάτων. Ωστόσο, είναι σημαντικό να υπάρχουν μοντέλα για επαναλαμβανόμενες παρατηρήσεις στις ίδιες μονάδες (recurrent events), όπως στην περίπτωση διαδοχικών διακοπών της λειτουργίας επισκευάσιμου (repairable) μηχανήματος.

Όπως έχει ήδη παρατηρηθεί, δεν υπάρχει έως τώρα ανάπτυξη των FHT μοντέλων για την περίπτωση επαναλαμβανόμενων γεγονότων στις ίδιες στατιστικές μονάδες (recurrent events). Μόλις πρόσφατα, οι Whitmore et al. (2012) ασχολήθηκαν με τη μοντελοποίηση μιας ανέλιξης επαναλαμβανόμενων δεδομένων, στην οποία οι διάρκειες ζωής ακολουθούν την ανέλιξη Wiener και οι χρόνοι πρώτης μετάβασης είναι ανεξάρτητοι και ισόνομοι, με ίσα διαδοχικά διαστήματα μεταξύ των βημάτων της ανέλιξης.

Στο δεύτερο κεφάλαιο της διατριβής παρουσιάστηκαν έλεγχοι για την εξέταση της ισότητας ή μη των παραμέτρων δύο δειγμάτων δεδομένων που ακολουθούν την κατανομή IG και στη συνέχεια κατασκευάστηκαν έλεγχοι για τον εντοπισμό άτυπων τιμών των παραμέτρων (μ, λ) της IG σε δεδομένα ανεξάρτητων μονάδων με επαναλαμβανόμενα ανεξάρτητα γεγονότα ανά μονάδα. Σε αυτήν την παράγραφο θα περιγράψουμε δομές επαναλαμβανόμενων παρατηρήσεων για την περίπτωση ενός IG FHTR μοντέλου παλινδρόμησης.

5.3.1 Δομές δεδομένων για επαναλαμβανόμενες παρατηρήσεις

Υπάρχει μεγάλη ποικιλία ως προς τις δομές δεδομένων που είναι διαθέσιμες για τις διάφορες μελέτες που έχουν πραγματοποιηθεί με τα μοντέλα της παλινδρόμησης Κατωφλιού. Πιο συγκεκριμένα, θα ασχοληθούμε με την περίπτωση όπου έχουμε επαναλαμβανόμενα δεδομένα διαθέσιμα για την κάθε μονάδα, όπως φαίνεται στο Σχήμα 5.1.



Σχήμα 5.1: Η περίπτωση των επαναλαμβανόμενων γεγονότων για μια μονάδα

Στην περίπτωση αυτή, η δομή για μία μονάδα *i*, $i=1,\ldots,m$ μπορεί να συνοψιστεί ως εξής:

Χρονικές στιγμές:

$t_1 \le \ldots \le t_{n_i}$

Δείκτες αποκοπής:

$f_1 = 0, \ldots, f_{n_i-1} = 0, f_{n_i} = 0 \text{ ή } 1,$

Ενδείξεις καταστάσεων της γονικής ανέλιξης:

$x_0, x_1, \ldots, x_{n_i}$

Διανύσματα συμμεταβλητών:

$\mathbf{z}_1, \ldots, \mathbf{z}_{n_i-1}$

Επομένως, για κάθε μία μονάδα *i*, $i=1,\ldots,m$ μπορούμε να δημιουργήσουμε ένα διάνυσμα της μορφής $\left(t_{i,j}, f_{i,j}, x_{i,j}, \mathbf{z}_{i,j}\right), j=1,\ldots,n_i$, όπου $0 \le t_{i,0} \le t_{i,1} \le \ldots \le t_{i,n_i}$, με το οποίο να την περιγράφουμε. Εδώ, $t_{i,j}$ είναι η χρονική στιγμή εμφάνισης του *j*-οστού συμβάντος, $f_{i,j}$ είναι δείκτης αποκοπής σχετικά με το εάν η στιγμή $t_{i,j}$ είναι στιγμή μετάβασης στο κατώφλι, $x_{i,j}$ είναι η κατάσταση της ανέλιξης τη χρονική στιγμή $t_{i,j}$ και $\mathbf{z}_{i,j}$ είναι το διάνυσμα των συμμεταβλητών της *j*-οστής παρατήρησης για την *i*-οστή μονάδα. Να παρατηρήσουμε πως οι χρονικές στιγμές $t_{i,j}$ υποδηλώνουν και τις στιγμές εμφάνισης των καινούργιων αρχικών σημείων επανεκκίνησης της διαδικασίας, θεωρούμε δηλαδή μηδενικό χρόνο από τη στιγμή που θα πραγματοποιηθεί ένα συμβάν, έως τη στιγμή όπου το σύστημα επανέρχεται σε κάποιο αρχικό σημείο εκκίνησης της διαδικασίας. Ακόμα, οι δομές δεδομένων μπορεί να έχουν μία σειρά από χαρακτηριστικά, όπως φαίνεται στη συνέχεια:

1. Τα σύνολα των δεδομένων συνήθως περιλαμβάνουν ένα δείγμα από μονάδες *i*, $i=1,\ldots,n$, με γονικές στοχαστικές διαδικασίες $\{X_i(t)\}$ για την κάθε μονάδα και σύνολα συνόρων B_i. Οι ανελίξεις των μονάδων συνήθως θεωρούνται αμοιβαία ανεξάρτητες.
2. Όταν υπάρχουν ανταγωνιζόμενοι κίνδυνοι αποτυχίας, τότε η αιτία της αποτυχίας *d*, θα καταγράφεται για την κάθε μονάδα.
3. Τα δεδομένα είναι επαναλαμβανόμενα, εάν υπάρχει παραπάνω από μία καταγραφή για τη μονάδα, δηλαδή εάν $n_i > 1$ για κάποιες μονάδες.
4. Σε περίπτωση που η γονική ανέλιξη είναι μη-παρατηρήσιμη, τότε το δείγμα δε θα έχει καμία παρατήρηση $x_{i,j}$, παρόλο που μπορεί να υπάρχουν παρατηρήσεις στο διάνυσμα συμμεταβλητών $\mathbf{z}_{i,j}$.

Στις περισσότερες εφαρμογές ως κλίμακα μέτρησης του χρόνου χρησιμοποιείται η ημερολογιακή κλίμακα. Υπάρχουν βέβαια και περιπτώσεις όπου και άλλες κλίμακες μπορούν να χρησιμοποιηθούν εναλλακτικά ή/και επιπρόσθετα. Για παράδειγμα, η φθορά που πραγματοποιείται στα λάστιχα μιας μηχανής δεν αυξάνεται μόνο με το πέρασμα του χρόνου, αλλά και με την καθημερινή χρήση της μηχανής σε χιλιόμετρα.

5.3.2 Η περίπτωση των χρονικά εξαρτημένων μεταβλητών

Οι δομές που περιγράψαμε προηγουμένως, μπορεί να συμπεριλάβουν και την περίπτωση των χρονικά εξαρτημένων μεταβλητών, ένα φαινόμενο που προκύπτει αρκετά συχνά στην περιοχή της Ανάλυσης Επιβίωσης. Κάποια συμμεταβλητή για την *i*-οστή μονάδα ενδέχεται να είναι χρονικά εξαρτημένη, με αποτέλεσμα να έχει διαφορετικές τιμές για τα διάφορα j, $j=1, 2, ..., n_i$, επομένως οφείλουμε να συμπεριλάβουμε και αυτήν την περίπτωση στη μοντελοποίηση.

Σε μοντέλα σαν και αυτό, συνήθως υποθέτουμε πως οι χρόνοι $t_{i,j}$, $t_{i',j'}$ είναι ανεξάρτητοι για διαφορετικές μονάδες, $i \neq i'$, και πως είναι υπό συνθήκη ανεξάρτητοι για την ίδια μονάδα, $i = i'$, δεδομένου των συμμεταβλητών. Επομένως, για την απλή αυτή περίπτωση, η έννοια της ανεξαρτησίας σημαίνει πως τα n_i επαναλαμβανόμενα δεδομένα για την *i* μονάδα μπορεί να συμπεριφερθούν ως μεμονωμένα γεγονότα n_i ανεξάρτητων μονάδων.

Στη συνέχεια, μπορούμε να εισάγουμε κάποια επιπλέον εξάρτηση ανάμεσα στις χρονικές στιγμές $t_{i,j}$ που ανήκουν στην ίδια μονάδα, συμπεριλαμβάνοντας κάποιες επιπλέον μεταβλητές:

1. Αριθμός προηγούμενων γεγονότων.
2. Συνολικός χρόνος μέχρι τώρα.

3. Αμέσως προηγούμενη χρονική στιγμή, $t_{i,j}$.

Τέλος, για κάποια εφαρμογή που περιλαμβάνει την ηλικία της μονάδας, τότε με το να υπολογίσουμε το χρόνο που έχει περάσει έως τώρα, είναι σαν να μετατρέπουμε την ηλικία σε χρονικά εξαρτώμενη μεταβλητή.
Η πιθανοφάνεια γράφεται ως:

$$L = \prod_{i=1}^{m} \prod_{j=1}^{n_i} L\left(t_{i,j} \middle| x_{0i,j}, m_{i,j}\right), \tag{5.1}$$

όπου οι χρονικές στιγμές $t_{i,j}$ αποτελούν παρατηρούμενους χρόνους για $j < n_i$, ενώ η τιμή μπορεί να είναι και αποκομμένη.

Συγκεκριμένα, για την περίπτωση του FHTR μοντέλου, οι ποσότητες $x_{0i,j}$ και $m_{i,j}$ συνδέονται με τις διάφορες μεταβλητές του προβλήματος, με τη βοήθεια γραμμικών συνδετικών συναρτήσεων και μπορεί να έχουν διαφορετικές τιμές για κάθε *j* (όπως και για κάθε *i*) και επιπλέον περιλαμβάνουν:

1. Αριθμός των προηγούμενων $j-1$ γεγονότων, ως μεταβλητή για την *j* παρατήρηση.
2. Συνολικός χρόνος μέχρι τώρα, $\sum_{l<j} t_{i,l}$.
3. Αμέσως προηγούμενη χρονική στιγμή, $t_{i,j-1}$.

Βιβλιογραφία

1. Aalen, O.O., Gjessing, H.K. (2001). Understanding the shape of the hazard rate: A process point of view. *Statistical Science*, **16**, 1-22.
2. Ackerberg, D.A., Machado, M.P., Riordan, M.H. (2006). Benchmarking for productivity improvement: A health-care application, *International Economic Review*, **47**, 161-201.
3. Altshuler, B. (1970). Theory for the measurement of competing risks in animal experiments, *Mathematical Biosciences*, **6**, 1-11.
4. Andersen, P.K., Borgan, O, Gill, R.D., Keiding, N. (1993). *Statistical Methods Based on Counting Processes*, Springer, New York.
5. Anscombe, F.J. (1961). *Examination of residuals*, Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability (University of California Press), 1-36.
6. Anscombe, F.J., Tukey, J.W. (1963). The examination and analysis of residuals, *Technometrics*, **5**, 141-160.
7. Bachelier, L. (1900). Theorie de la speculation, *Annales scientifiques de l'École Normale Supérieure*, Paris, **17**, 3, 21-86.
8. Balka, J., Desmond, A.F., McNicholas, P.D. (2009). Review and implementation of cure models based on first hitting times for Wiener processes, *Lifetime Data Analysis*, **15**, 147-176.
9. Barnett, V., Lewis, T. (1994). *Outliers in Statistical Data*, 3rd edition, John Wiley, Chichester.
10. Bhattacharyya G.K., Fries, A. (1982). *Inverse Gaussian regression and accelerated life tests in Survival Analysis*, ed. J. Crowley and R.A. Johnson, IMS Lecture Notes, Monograph Series, **2**, 101-118.
11. Bennett, S., Whitehead, J. (1981). Fitting logistic and log-logistic regression models to censored data using GLIM, *GLIM Newsletter*, **4**, 12–19 with correction **5**, 3.
12. Bennett, S. (1983a). Analysis of survival data by the proportional odds model, *Statistics in Medicine*, **2**, 273-277.

13. Bennett, S. (1983b). Log-logistic regression models for survival data, *Applied Statistics*, **32**, 165 - 171.
14. Boag, J.W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy, *Journal of the Royal Statistical Society*, Series B, **11**, 15–53.
15. Brown, P.F., Potts, J.R. (1977). Evaluation of powder processed turbine engine ball bearings, Air Force Aero-Propulsion Laboratory, Technical Report No. AFAPL-TR, 77-26.
16. Caroni, C. (2010). Testing for outliers in the power-law growth parameters of sets of recurrent events data, *Statisical Methodology*,**7**, 632–637.
17. Clayton, D.G. (1974). Some odds ratio statistics for the analysis of ordered categorical data, *Biometrika*, **61**, 525-531.
18. Clayton, D.G. (1976). An odds ratio comparison for ordered categorical data with censored observations, *Biometrika*, **63**, 405-408.
19. Caplehorn, J.R.M., Bell, J. (1991). Methadone dosage and retention of patients in maintenance treatment, *The Medical Journal of Australia*, **154**, 195-199.
20. Chhikara, R.S. (1975). Optimum tests for comparison of two Inverse Gaussian distribution means, *Australian Journal of Statistics*, **17**, 77-83.
21. Chhikara, R.S., Folks, J.L. (1977). The inverse Gaussian distribution as a lifetime model, *Technometrics*, **4**, 461-468.
22. Chhikara, R.S., Folks, J.L. (1989). *The Inverse Gaussian Distribution*, Marcel Dekker, Inc, New York.
23. Collett, D. (2003). *Modelling Survival Data in Medical Research*, Chapman and Hall, London.
24. Concato, J., Peduzzi, P., Holford, T.R., Feinstein, A.R. (1995). Importance of events per independent variable in proportional hazards analysis: I. Background, goals, and general strategy, *Journal of Clinical Epidemiology*, **48**, 1495-1501.
25. Cook, R.D., Lawless, J.F. (2007). *The Statistical Analysis of Recurrent Events*, Springer, New York.
26. Cook, R.D., Weisberg, S. (1982). *Residuals and Influence in Regression*, Chapman and Hall.
27. Cook, R.D. (1986). Assessment of local influence, *Journal of the Royal Statistical Society*, Series B, **48**, 133-169.
28. Cox, D.R., Miller, H.D. (1965). *The Theory of Stochastic Processes*. Chapman and Hall, London.
29. Cox, D.R., Snell, E.J. (1968). A general definition of residuals (with discussion). *Journal of the Royal Statistical Society*, Series A, **30**, 248-275.

30. Cox, D.R. (1972). Regression models and life-tables**,** *Journal of the Royal Statistical Society*, Series B, **34**, 187-220.
31. Cox, D. R., Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman and Hall, New York.
32. Cressie, N., Davis, A.S., Folks, J.L., Policello II, G.E. (1981), The moment generating function and negative integer moments, *American Statistician*, **35**, 148-150.
33. Crowder M.J., Kimber, A.C., Smith, R.L. and Sweeting, T.J. (1991). *Statistical Analysis of Reliability Data*, Chapman and Hall/CRC, London.
34. Davis, A.S. (1980). Use of the likelihood ratio test on the inverse Gaussian distribution, *American Statistician*, **34,** 108–110.
35. Delong D.M., Guirguis, G.H., So, Y.C. (1994). Efficient computation of subset selection probabilities with application to Cox regression. *Biometrika*, **81**, 607-611.
36. DeRose, J.J., Toumpoulis, I.K., Balaram, S.K., Ioannidis, J.P., Belsey, S., Ashton, R.C., Swistel, D.G., Anagnostopoulos, C.E. (2005). Preoperative prediction of long-term survival after coronary artery bypass grafting in patients with low left ventricular ejection fraction, *Journal of Thoracic and Cardiovascular Surgery*, **129**, 314–321.
37. Doksum, K.A., Hoyland, A. (1992). Models for variable-stress accelerated life testing experiments based on Wiener processes and the inverse Gaussian distribution, *Technometrics*, **34**, 74–82.
38. Doksum, K.A., Normand, S.L. (1995). Gaussian models for degradation processes: I. Methods for the analysis of biomarker data, *Lifetime Data Analysis*, **1**, 131-144.
39. Dufresne, F., Gerber, H.U., Shiu, E.S.W. (1991). Risk theory with the Gamma process, *Astin Bulletin*, **21**, 177-192.
40. Eaton, W.W., Whitmore, G.A. (1977). Length of stay as a stochastic process: A general approach and application to hospitalization for schizophrenia, *Journal of Mathematical Sociology*, **5**, 273-292.
41. Eberly LE, Grambsch P, Connett JE (2001). Comment on paper by Aalen & Gjessing, *Statistical Science*, **16**, 16-19.
42. Economou, P., Caroni, C. (2009). Fitting parametric frailty and mixture models under biased sampling, *Journal of Applied Statistics*, **36**, 53–66.
43. Edgeman, R.L. (1989). Assessing the inverse Gaussian distribution assumption, *IEEE Transactions on Reliability,* **39**, 352-355.
44. Embury, S.H., Elias, L., Heller, P.H, Hood, C.E., Greenberg, P.L., Schier, S.L. (1977). Remission maintenance therapy in acute myelogenous leukemia, *Western Journal of Medicine*, **126**, 267-272.

45. Erich, R., Pennell, M. (2012). *Ornstein-Uhlenbeck Threshold Regression Models for Time to Event Data*, ICSA, the 21st Applied Statistics Symposium, June 23-26, Boston, USA.

46. Everrit, B.S., Hothorn, T. (2006). *A Handbook of Statistical Analysis Using R*, Chapman and Hall, London.

47. Fan, J., Li, J.R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96**, 1348-1360.

48. Feaganes J.R., Suchindran C.M. (1991). *Weibull regression with unobservable heterogeneity, an application*, American Statistical Association, Proceedings of the Social Statistics Section, 160-165.

49. Freedman, D.A. (2008). Survival analysis: A primer, *American Statistician*, **62**, 110-119.

50. Fries, A., Bhattacharyya, G.K. (1983). Analysis of two-factor experiments under an inverse Gaussian model, *Journal of American Statistical Association*, **78**, 820-826.

51. Garshick, E., Laden,F., Hart, J., Rosner, B., Smith, T.J., Dockery, D.W., Speizer, F.E. (2004). Lung cancer in railroad workers exposed to diesel exhaust, *Enviromental Health Perspectives*, **112**, 1539-1543.

52. Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Bornkamp, B., Hothorn, T. (2011). Package "mvtnorm", available at http://cran.r- project/web/packages/ mvtnorm/mvtnorm.pdf.

53. Gore, S.M., Pocock, S.J., Kerr, G.R. (1984). Regression models and non-proportional hazards in the analysis of breast cancer survival, *Applied Statistics*, **33**,176-195.

54. Grambsch, P.M., Therneau, T.M. (1994). Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika*, **81**, 515-526.

55. Greenwood, M., (1926). The errors of sampling of the survivorship tables, *Reports on Public Health and Statistical Subjects*, 33, Appendix 1, HMSO, London.

56. Hansen, B.E. (2000). Sample splitting and threshold estimation, *Econometrica*, **68**, 575-603.

57. Hardin, J.W., Hilbe, J.M.. (2007). *Generalised Linear Models and Extensions*, 2nd edition, Stata Press, College Station, Texas.

58. Hastie T., Efron, B. (2011). Package `lars'; available at http://cran.r-project/web /packages/lars/lars.pdf.

59. Hazelton, W. D., Luebeck, E. G., Heidenreich, W. F., Moolgavkar, S. H. (2001). Analysis of a historical cohort of Chinese tin miners with arsenic, radon, cigarette smoke, and pipe smoke exposures using the biologically based two-stage clonal expansion model, *Radiation Research*, **156**, 78-94.

60. Hoeting, J.A., Madigan, D., Raftery, A., Volinsky, C. (1999). Bayesian model averaging: a tutorial (with discussions), *Statistical Science*, **14**, 382-417.

61. Hougaard P. (1984). Life table methods for heterogeneous populations: Distributions describing the heterogeneity, *Biometrika*, **71**, 75-83.

62. Hutton, J.L., Monaghan, P.F. (2002). Choice of parametric accelerated life and proportional hazards models for survival data: Asymptotic results, *Lifetime Data Analysis*, **8**, 375-393.

63. Johnson, N.L., Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions*, 1, Houghton Mifflin, Boston.

64. Kalbfleisch, J.D., Prentice R. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. Wiley, N.Y.

65. Kaplan, E.L., Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**, 457-481.

66. Karioti, V., Caroni, C. (2002). Detecting outlying series in sets of short time series, *Computational Statistics and Data Analysis*, **39,** 351-364.

67. Karioti, V., Caroni, C. (2006). Properties Of The GAR(1) Model For Time Series Of Counts, *Journal of Modern Applied Statistical Methods*, **5**, 140-151.

68. Kennedy, Jr., W.J., Gentle, J.E. (1980). *Statistical Computing*, Marcel Dekker Inc, New York.

69. Lancaster, T. (1972). A stochastic model for the duration of a strike, *Journal of the Royal Statistical Society*, Series A, **135**, 257-271.

70. Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd ed. Wiley, New York.

71. Lawless, J., Crowder, M. (2004). Covariates and random effects in a Gamma process model with application to degradation and failure, *Lifetime Data Analysis*, **10**, 213-227.

72. Lee, M.L.T., De Gruttola, V., Schoenfeld, D. (2000). A model for markers and latent health status, *Journal of the Royal Statistical Society*, Series B, **62**, 747-762.

73. Lee, M.L.T., Garshick, E., Whitmore, G.A., Laden, F., Hart, J. (2004). Assesing lung cancer risk to rail workers using a first hitting time regression model, *Envirometrics*, **15**, 1-12.

74. Lee, M.L.T., Whitmore, G.A. (2006). Threshold regression for survival analysis: Modelling event times by a stochastic process reaching a boundary, *Statistical Science*, **21**, 501-503.

75. Lee M-L.T., Whitmore GA, Laden F, Hart JE & Garshick E. (2009). A case-control study relating railroad worker mortality to diesel exhaust exposure using a threshold regression model, *Journal of Statistical Planning and Inference*, **139**, 1633-1642.

76. Lee, M-L.T., Whitmore, G.A. (2010). Proportional hazards and threshold regression: Their theoretical and practical connections, *Lifetime Data Analysis*, **16**, 196-214.

77. Lee, M-L.T., Whitmore, G.A., Rosner, B.A. (2010). Threshold regression for survival data with time-varying covariates, *Statistics in Medicine*, **29**, 896-905.

78. Lewis, S.L., Montgomery, D.C., Myers, R.H. (2001a). Examples of designed experiments with nonnormal responses, *Journal of Quality Technology*, **33**, 265-278.

79. Lewis, S.L., Montgomery, D.C., Myers, R.H. (2001b). Confidence interval coverage for designed experiments analyzed with GLMs, *Journal of Quality Technology*, **33**, 279-292.

80. Lin J.G., Wei B.C., Zhang N.S. (2004). Varying dispersion diagnostics for inverse Gaussian regression models, *Journal of Applied Statistics*, **31**, 1157-1170.

81. Maller, P. A., Zhou, S. (1994). Testing for sufficient follow-up and outliers in survival data, *Journal of American Statistical Association*, **89**, 1499-1506.

82. McCullagh, P. (1980). Regression models for ordinal data (with discussion), *Journal of the Royal Statistical Society*, Series B, **42**, 109-142.

83. McCullagh, P., Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edition, Chapman and Hall, New York, London.

84. Michael, J.R., Schucany, W., Haas, R., (1976). Generating random variables using transformation with multiple roots, *American Statistician*, **30**, 88-90.

85. Myers, R.H., Montgomery, D.C. (1997). A tutorial on generalized linear models, *Journal of Quality Technology*, **29**, 274-291.

86. Nelder, J., Wedderburn, R. (1972). Generalized linear models, *Journal of the Royal Statistical Society*, Series A, **135**, 3, 370-384.

87. Ostle, B. (1963). *Statistics in Research*, 2nd ed., Iowa State University Press, Ames, IA.

88. Padgett W.J., Tsai, S.K. (1986). Prediction intervals for future observations from the inverse Gaussian distribution, *IEEE Transactions on Reliability*, R**35**, 406-408.

89. Park, C., Padgett, D.W. (2005). Accelerated Degradation Models for Failure Based on Geometric Brownian Motion and Gamma Processes, Lifetime Data Analysis, **11**, 511-527.

90. Peduzzi, P., Concato, J., Feinstein, A.R., Holford, T.R. (1995). Importance of events per independent variable in proportional hazards regression analysis: II. Accuracy and precision of regression estimates, *Journal of Clinical Epidemiology*, **48**, 1503-1510.

91. Pennell, M.L., Whitmore, G.A., Lee, M.L.T. (2010). Bayesian random-effects threshold regression with application to survival data with non-proportional hazards, *Biostatistics*, **11**, 111-126.

92. Poon, W-Y., Poon, Y-S. (1999). Conformal normal curvature and assessment of local influence, *Journal of the Royal Statistical Society*, Series B, **61**, 51-61.

93. Poon, W-Y., Tang, M-L. (2010). Influence measure in maximum likelihood estimate for models of lifetime data, *Journal of Applied Statistics*, **28**, 737-742.

94. Qu, H, Xie, F.C. (2011). Diagnostics analysis for log-Birnbaum–Saunders regression models with censored data, *Statistica Neerlandica*, **65**, 1-21

95. Rabe-Hesketh, S., Everrit, B.S. (2000). *A Handbook of Statistical Analysis Using Stata*, Chapman and Hall, London.

96. Ricciardi, L.M., Sato, S. (1988). First passage-time density and moments of the Ornstein-Uhlenbeck process, *Journal of Applied Probability*, **25**, 43-57.

97. Ross, S.M., (1996). *Stochastic Processes*, 2nd edition, Wiley, New York.

98. Rykov, V.V., Balakrisnan, N., Nikulin, M.S., (2010). *Mathematical and Statistical Models and Methods in Reliability. Applications to Medicine, Finance, and Quality Control*, Birkhauser, Boston.

99. Schoenfeld, D.A. (1982). Partial residuals for the proportional hazards regression model, *Biometrika*, **69**, 239-241.

100. Seshadri, V. (1993). *The Inverse Gaussian Distribution*, Clarendon Press, Oxford.

101. Shuster, J.J. (1968). On the inverse Gaussian distribution, *Journal of the American Statistical Association*, **63**, 1514-1516.

102. Singpurwalla, N. (1995). Survival in dynamic environments, *Statistical Science*, **10**, 86-103.

103. Smyth, G., Hu, Y., Dunn, P., Phipson, B. (2011). Package "statmod", available at http://cran.r-project/web/packages/statmod/statmod.pdf.

104. Snell, E.J. (1964). A scaling procedure for ordered categorical data, *Biometrics*, **20**, 592- 607.

105. Tableman, M., Kim, J.M. (2004). *Survival Analysis Using S*. Chapman and Hall, London.

106. Therneau, T.M., Grambsch, P.M., Fleming, T.R. (1990). Martingale-based residuals for survival models, *Biometrika*, **77**, 147-60.

107. Therneau, T.M., Grambsch, P.M. (2000). *Modelling Survival Data: Extending the Cox Model*. Springer, New York.

108. Therneau, T.M. (2011). Package "survival", available at http://cran.r- project/web/ packages/survival/survival.pdf.

109. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society*, **58**, 267-288.

110. Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model, *Statistics in Medicine*, **16**, 385-395.
111. Tweedie, M.C.K. (1941). A mathematical investigation of some electrophoretic measurements on colloids, M.Sc. Thesis, University of Reading, England.
112. Tweedie, M.C.K. (1945). Inverse statistical variates, *Nature*, 155-453.
113. Tweedie, M.C.K. (1957a). Statistical properties of the inverse Gaussian distributions I, *Annals of Mathematical Statistics*, **28**, 362-377.
114. Tweedie, M.C.K. (1957b). Statistical properties of the inverse Gaussian distributions II, *Annals of Mathematical Statistics*, **28**, 696-705.
115. Van Noortwijk, J.M. (2009). A survey of the application of gamma processes in maintenance, Reliability Engineering & System Safety, **94**, 2-21.
116. Vittinghoff, E., McCulloch, C.E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression, *American Journal of Epidemiology*, **165**, 710–718.
117. Viveros, R. (1990). An approximate normalizing power transformation for the F distribution, *Communication in Statistics-Simulation and Computation*, **19**, 57–69.
118. Wald, A. (1944). On cumulative sums of random variables, *Annals of Mathematical Statistics*, **15**, 283-296.
119. Wald, A. (1947). *Sequential Analysis*,Wiley, New York.
120. Wang, H., Leng, C. (2007). Unified LASSO estimation by least squares approximation, *Journal of the American Statistical Association*, **102**, 1039-1048.
121. Wasan, M.T. (1969). On an inverse Gaussian process, *Scandinavian Actuarial Journal*, **60**, 69-96.
122. Wiener N. (1923). Differential space, *Journal of Mathematical Physics*, **2**, 131-174.
123. Whitmore, G.A. (1975). The inverse Gaussian distribution as a model of hospital stay, *Health Services Research*, **10**, 297-302.
124. Whitmore, G.A, Yalovsky, M. (1978). A normalizing logarithmic transformation for inverse Gaussian random variables, *Technometrics*, **20**, 207-208.
125. Whitmore, G.A. (1979). An inverse Gaussian model for labour turnover, *Journal of the Royal Statistical Society*, Series A, **142**, 468-478.
126. Whitmore, G.A. (1983). A regression method for censored inverse-Gaussian data, *Canadian Journal of Statist*ics, **11**, 305-315.
127. Whitmore, G.A. (1986). First passage time models for duration data regression structures and competing risks, *The Statistician*, **35**, 207-219.
128. Whitmore, G.A., Seshadri, V. (1987). A heuristic derivation of the inverse Gaussian distribution, *American Statistician*, **41**, 4, 280-281.

129. Whitmore, G.A. (1995). Estimating degradation by a Wiener diffusion process subject to measurement error, *Lifetime Data Analysis*, **1**, 307-319.

130. Whitmore, G.A., Schenkelberg, F. (1997). Modelling accelerated degradation data using Wiener diffusion with a time scale transformation. *Lifetime Data Analysis*, **3**, 27-45.

131. Whitmore, G.A., Crowder, M.J., Lawless, M.J. (1998). Failure inference from a marker process based on a bivariate Wiener model, *Lifetime Data Analysis*, **4**, 229-251.

132. Whitmore, G.A., Ramsay, T., Aaron, S.D. (2012). Recurrent first hitting times in Wiener diffusion under several observation schemes, *Lifetime Data Analysis*, **18**, 157-176.

133. Yu, Z., Tu, W., Lee, M.L.T. (2009). A semi-parametric threshold regression analysis of sexually transmitted infections in adolescent women, *Statistics in Medicine*, **28**, 3029-3042.

134. Zhang, J., Peng, Y. (2009). Crossing hazard functions in common survival models, *Statistics & Probability Letters*, **79**, 2124-2130.

135. Zhu, M., Fan, G. (2011). Variable selection by ensembles for the Cox model, *Journal of Statistical Computation and Simulation*, **81**,1983-1992.

136. Zou, H. (2006). The adaptive Lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418-1429.

137. Καρώνη, Χ. (2009). *Μοντέλα Αξιοπιστίας και Επιβίωσης*, Εκδόσεις Συμεών, Αθήνα.

138. Κοκολάκης, Γ.Ε. (2006). *Σημειώσεις Στοχαστικών Ανελίξεων*, Εκδόσεις Ε.Μ.Π, Αθήνα.

139. Κοκολάκης, Γ.Ε., Φουσκάκης, Δ. (2009). *Στατιστική: Θεωρία και Εφαρμογές*, Εκδόσεις Συμεών, Αθήνα.