Εθνικο Καποδιστριακο Πανεπιστημιο
Αθηνων

Τμημα Μαθηματικο

Μεταπτυχιακο Προγραμμα Σπουδων στη Στατιστικη
και Επιχειρησιακη Ερευνα

Μέθοδοι Markov Chain Monte Carlo για Μπεϋζιανή Συμπερασματολογία σε Γενικευμένα Γραμμικά Μοντέλα

Συγγραφέας:
Ειρήνη Μπόνη

Επιβλέπουσα:
Λουκία Μελιγκοτσίδου

Αθήνα
Οκτώβριος 2012

10 Οκτωβρίου 2012

'Σε εκείνους που οφείλω τα πάντα.
Στους γονείς μου και στο Γιώργο.'

Ευχαριστίες

Ευχαριστώ την επιβλέπουσα καθηγήτρια της διπλωματικής μου εργασίας κ. Λουκία Μελιγκοτσίδου για την πολύτιμη καθοδήγηση της.

Επίσης θα ήθελα να ευχαριστήσω το Βασίλη και τη Φιόρη για την ανιδιοτελή βοήθεια τους, το ενδιαφέρον τους, τον χρόνο που μου αφιέρωσαν και την συμπαράσταση τους σε όλη τη διάρκεια των μεταπτυχιακών μου σπουδών.

Ευχαριστώ επίσης ιδιαίτερα το Γιώργο και την οικογένειά μου για την στήριξη και κατανόησή τους ιδιαίτερα σε περιόδους πίεσης.

Τέλος θα ήθελα να αναφέρω ότι ο κώδικας που δημιουργήθηκε στη γλώσσα προγραμματισμού (R) για την πραγματοποίηση της παρούσας εργασίας είναι διαθέσιμος για κάθε ενδιαφερόμενο.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κεφάλαιο 1

Εισαγωγή

1.1 Αντικείμενο της Διπλωματικής Εργασίας

Στην παρούσα διπλωματική εργασία εξετάζονται *Markov Chain Monte Carlo* μέθοδοι σε γενικευμένα γραμμικά μοντέλα από τη σκοπιά της στατιστικής κατά *Bayes*. Συγκεκριμένα, εξετάζεται η γενική θεωρία των γενικευμένων γραμμικών μοντέλων με έμφαση στη λογιστική και τη *poisson* παλινδρόμηση, καθώς επίσης και ο τρόπος με τον οποίο προσεγγίζουμε τα παραπάνω προβλήματα από τη σκοπιά της Μπεϋζιανής θεωρίας. Η μπεϋζιανή συμπερασματολογία επιτρέπει την εξαγωγή πιθανοθεωρητικών συμπερασμάτων σχετικά με τις άγνωστες παραμέτρους του μοντέλου και την ενσωμάτωση σε αυτό εκ των προτέρων γνώσης με βάση την οποία οδηγούμαστε σε εκ των υστέρων (*posterior*) κατανομές στις οποίες εμπεριέχεται όλη η στατιστική συμπερασματολογία των αγνώστων αυτών παραμέτρων όπως αυτή έχει προκύψει από την Μπεϋζιανή ανάλυση. Χρησιμοποιώντας τους αλγόριθμους $MCMC$ μπορούμε να δημιουργήσουμε και να υπολογίσουμε σύνθετα μοντέλα που περιγράφουν σύνθετα προβλήματα τα ο-

ποία με τις παραδοσιακές μεθόδους δεν θα ήταν εύκολα να επιληθούν.

1.2 Δομή της Διπλωματικής Εργασίας

Η παρούσα διπλωματική εργασία διαρθρώνεται σε έξι κεφάλαια ως εξής: Στο δεύτερο κεφάλαιο γίνεται μια εισαγωγή στα γενικευμένα γραμμικά μοντέλα και αναφέρονται κάποιες βασικές έννοιες και ιδιότητες των μοντέλων που ανήκουν σε αυτήν την κατηγορία. Στο τρίτο κεφάλαιο περιγράφονται τα μοντέλα της λογιστικής και της *poisson* παλινδρόμησης. Στο τέταρτο κεφάλαιο αναπτύσσονται οι βασικές αρχές της Μπεϋζιανής Στατιστικής θεωρίας όπου μεταξύ άλλων δίνεται ο ορισμός της εκ των προτέρων κατανομής, της εκ των υστέρων κατανομής και του Θεωρήματος *Bayes*, καθώς επίσης δίνετε και μια εκτενής περιγραφή των αλγορίθμων *Markov Chain Monte Carlo*. Στο πέμπτο κεφάλαιο παρουσιάζονται εφαρμογές της λογιστικής και της *poisson* παλινδρόμησης σε πραγματικά δεδομένα. Τέλος, στο έκτο κεφάλαιο παρουσιάζεται και αναλύεται εκτενώς μία εφαρμογή της λογιστικής παλινδρόμησης με σκοπό την εκτίμηση της πιθανότητας χρεωκοπίας επιχειρήσεων.

Κεφάλαιο 2

Γενικευμένα Γραμμικά Μοντέλα

Στο κεφάλαιο αυτό παρουσιάζονται τα γενικευμένα γραμμικά μοντέλα ως μια επέκταση του απλού γραμμικού μοντέλου.

2.1 Ιστορική Αναδρομή

Τα γενικευμένα γραμμικά μοντέλα (*Generalised Linear Models*) προτάθηκαν από τους *John Nelder* και *Robert Weddeburn* το 1972 όπου με τίτλο *Generalized Linear Models* δημοσιεύτηκε στο περιοδικό: *Journal of the Royal Statistical Society* ως τρόπος ενοποίησης άλλων στατιστικών μοντέλων συμπεριλαμβανομένων της Λογιστικής παλινδρόμησης, της Γραμμικής παλινδρόμησης και της παλινδρόμησης *Poisson*.

Στη Στατιστική, αποτελούν μια μεγάλη κατηγορία μοντέλων που περιλαμβάνουν στοχαστικές αναπαραστάσεις οι οποίες χρησιμοποιούνται για την ανάλυση τόσο ποσοτικών (συνεχόμενων και διακριτών) όσο και ποιοτικών μεταβλητών. Για το λόγο αυτό τα τελευταία χρόνια μπορούν να θεωρηθούν πολύ σημαντικά

εργαλεία τόσο για τη μοντέρνα στατιστική θεωρία όσο και για τη στατιστική μοντελοποίηση. Αυτή η δημοτικότητά τους οφείλεται στην ευελιξία τους για την αντιμετώπιση ποικίλων στατιστικών προβλημάτων. Δεν είναι μόνο μια οικογένεια από μοντέλα που χρησιμοποιούνται ευρέως στην πράξη αλλά είναι ένας γενικός τρόπος σκέψης σχετικά με τη διαμόρφωση των στατιστικών μοντέλων.

Τα γενικευμένα γραμμικά μοντέλα (Γ.Γ.Μ.) μπορούν να θεωρηθούν ως μία φυσική επέκταση των κλασσικών γραμμικών μοντέλων που επιτρέπει στη μέση τιμή ενός πληθυσμού να εξαρτάται από μία γραμμική παράμετρο πρόβλεψης (*linear predictor*) μέσα από μία μη γραμμική συνδετική συνάρτηση (*link function*). Επίσης επιτρέπει στην κατανομή της εξαρτημένης μεταβλητής (*response variable*) να είναι οποιαδήποτε κατανομή από την εκθετική οικογένεια κατανομών, η οποία περιλαμβάνει τις πιο κοινές κατανομές όπως είναι η κανονική, η διωνυμική και η *Poisson*.

2.2 Απλό Γραμμικό Μοντέλο

Στη Στατιστική, η γραμμική παλινδρόμηση είναι μια προσέγγιση για τη μοντελοποίηση της σχέσης μιας μονοδιάστατης μεταβλητής Υ και μίας ή περισσοτέρων επεξηγηματικών μεταβλητών Χ. Στη γραμμική παλινδρόμηση οι άγνωστες παράμετροι εκτιμώνται από τα δεδομένα με την βοήθεια γραμμικών λειτουργιών. Τέτοιου είδους μοντέλα ονομάζονται γραμμικά μοντέλα.

Τα γραμμικά μοντέλα είναι μοντέλα τα οποία είναι γραμμικά ως προς τις παραμέτρους του μοντέλου και όχι αναγκαστικά ως προς την εξαρτημένη μεταβλητή. Ένα μοντέλο όπως το $y = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2$ ή το $y = \beta_0 + \beta_1 \cdot sinx$ είναι γραμμικά ως προς τις παραμέτρους του μοντέλου β ενώ το μοντέλο $y = \beta_0 + \beta_1 \cdot e^{-\beta_1 \cdot x}$

δεν είναι γραμμικό ως προς τις παραμέτρους του μοντέλου β.

Το γραμμικό μοντέλο είναι της μορφής

$$Y = X \cdot \beta + e$$

και με τη βοήθεια πινάκων περιγράφεται ως εξης:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & . & . & . & x_{1p} \\ x_{21} & . & . & . & x_{2p} \\ & & . & & \\ & & . & & \\ & & . & & \\ x_{n1} & . & . & . & x_{np} \end{bmatrix} * \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

όπου $Y = (y_1, ..., y_n)^T$ είναι η στήλη των παρατηρήσεων της εξαρτημένης μεταβλητής η οποία ονομάζεται και μεταβλητή απόκρισης (*response variable*) και ο πίνακας X, διάστασης $n \times p$, είναι ο πίνακας των τιμών των επεξηγηματικών ανεξάρτητων τυχαίων μεταβλητων (X_1,X_2,...,X_p). Κάθε γραμμή αναφέρεται σε μια διαφορετική παρατήρηση και κάθε στήλη αναφέρεται σε μια διαφορετική ανεξάρτητη μεταβλητή. Η στήλη των παραμέτρων $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ περιλαμβάνει τους συντελεστές των ανεξάρτητων μεταβλητών οι οποίοι θεωρούνται άγνωστοι και πρέπει να εκτιμηθούν. Η στήλη των υπολοίπων (*residuals*) $e = (e_1, e_2, ..., e_n)^T$ είναι η στήλη των τυχαίων σφαλμάτων (*random error terms*). Η υπόθεση που υιοθετούμε στο παραπάνω γραμμικό μοντέλο είναι ότι τα $e_1, e_2, ..., e_n$ είναι ανεξάρτητα και ομοιόμορφα κατανεμημένα και ακολουθούν την τη κανονική κατανομή $N(0, \sigma^2)$. Οπότε η μεταβλητή $Y | (X_1, X_2, ..., X_p) \sim N(X^T \cdot \beta, \sigma^2)$.

Σε αυτή τη περίπτωση

$$E(Y_i) = \mu_i = \sum_1^p x_{ij} \cdot \beta_j \quad i = 1, \ldots, n$$

και σε μορφή πινάκων γράφουμε :

$$\mu = X \cdot \beta$$

όπου μ είναι ένας πίνακας διάστασης $n \times 1$, X είναι ένας πίνακας διάστασης $n \times p$ και β είναι ένας πίνακας διάστασης $p \times 1$.

Παρατηρούμε ότι κυρίαρχο ρόλο έπαιξε η κανονική κατανομή την οποία υποθέσαμε ότι ακολουθούν τα τυχαία σφάλματα οπότε και η μεταβλητή απόκρισης (*response variable*). Πολλές φορές όμως αυτό δεν ισχύει, π.χ. η μεταβλητή απόκρισης μπορεί να ακολουθεί την διωνυμική κατανομή, δηλαδή τα αποτελέσματα να είναι της μορφής 0 ή αλλιώς αποτυχία και 1 ή αλλιώς επιτυχία.

Το μοντέλο αυτό μπορεί να γενικευθεί με πολλούς τρόπους. Εμείς θα ασχοληθούμε με την εξής γενίκευση:

- Οι ανεξάρτητες παρατηρήσεις ακολουθούν κατανομή διαφορετική της κανονικής κατανομής. Θα μπορούσε να είναι ακόμα και διακριτή. Αυτό βασίζεται στο γεγονός ότι πολλές από τις 'καλές' ιδιότητες της κανονικής κατανομής απαντώνται σε μια μεγαλύτερη κλάση κατανομών την *εκθετική οικογένεια κατανομών*.

- Η σχέση μεταξύ των εξαρτημένων και ανεξάρτητων μεταβλητών να μην είναι γραμμική. Η γραμμική έκφραση δεν εξαλείφεται πλήρως αλλά βρίσκεται μέσα σε κάποια άλλη συνάρτηση. Πιο συγκεκριμένα στο απλό γραμμικό μοντέλο εισάγουμε ένα γραμμικό εκτιμητή η (*linear predictor*) όπου μ=E(Y)=$X^T \cdot \beta$=η και σε αυτή τη περίπτωση βλέπουμε ότι τα μ και η είναι στην πραγματικότητα όμοια. Γενικεύοντας το παραπάνω μπορούμε να υποθέσουμε ότι η σχέση αυτή αντικαθίσταται από τη σχέση $\eta_i = g(\mu_i)$, όπου g οποιαδήποτε μονότονη συνάρτηση την οποία καλούμε συνάρτηση σύνδεσης (*link function*).

Στην περίπτωση αυτή έχουμε τα γενικευμένα γραμμικά μοντέλα των οποίων τα δεδομένα ακολουθούν κατανομές που ανήκουν στην εκθετική οικογένεια κατανομών και επιπρόσθετα η συνάρτηση σύνδεσης g μπορεί να είναι οποιαδήποτε μονότονη συνάρτηση.

2.3 Εκθετική Οικογένεια Κατανομών

Ορισμός

Έστω Υ μια τυχαία μεταβλητή, της οποίας η συνάρτηση πυκνότητας ή πιθανότητας $f(x;\underline{\theta})$ εξαρτάται από τη διανυσματική παράμετρο $\underline{\theta} = (\theta_1, \theta_2, ..., \theta_s)'$. Θα λέμε ότι η X, ή ισοδύμανα η κατανομή της X, ανήκει στην s - παραμετρική εκθετική οικογένεια κατανομών (s - *dimensional exponensial family*), αν και μόνο αν

ι. Το στήριγμα

$$S_f = [x \in \Re : f(x;\underline{\theta}) > 0]$$

της τυχαίας μεταβλητής X δεν εξαρτάται από την παράμετρο $\underline{\theta}$.

ιι. Η συνάρτηση πυκνότητας ή πιθανότητας $f(x;\underline{\theta})$ μπορεί να γραφεί σε μία από τις ισοδύναμες μορφές

1. $f(x;\underline{\theta}) = \exp[\sum_{k=1}^{s} \eta_k(\underline{\theta})T_k(x) - B(\underline{\theta}) + H(x)]$
2. $f(x;\underline{\theta}) = \exp[\sum_{k=1}^{s} \eta_k(\underline{\theta})T_k(x) - B(\underline{\theta})] \cdot h(x)$
3. $f(x;\underline{\theta}) = \beta(\underline{\theta}) \cdot [\exp[\sum_{k=1}^{s} \eta_k(\underline{\theta})T_k(x)] \cdot h(x)$

όπου οι $\eta_k(.)$, $\beta(.)$, $B(.)$, $h(.)$, $H(.)$, $T_k(.)$ είναι πραγματικές συναρτήσεις και ισχύει ότι $h(x) > 0$ για κάθε $x \in \Re$, $\beta(\underline{\theta}) > 0$ για κάθε $\underline{\theta} \in \Re^s$.

Στην εκθετική οικογένεια κατανομών συνηθίζεται να θεωρούμε σαν (φυσική) παράμετρο της κατανομής το διάνυσμα $\underline{\eta} = (\eta_1(\underline{\theta}), ..., \eta_s(\underline{\theta}))'$, οπότε με κατάλληλη αναπαραμετροποίηση η συνάρτηση πυκνότητας ή πιθανότητας παίρνει τη λεγόμενη κανονική μορφή $(canonical\ form)$:

$$f(x;\underline{\eta}) = [\exp[\textstyle\sum_{k=1}^{s} \eta_k T_k(x) - A(\underline{\eta})] \cdot h(x)$$

όπου A(.) πραγματική συνάρτηση.

Σαν φυσικός παραμετρικός χώρο της f ορίζουμε το σύνολο

$$Z = \left\{ \underline{\eta} \in \Re^s : \int f(x;\underline{\eta}) dx < \infty \right\}$$

εάν η τυχαία μεταβλητή X είναι συνεχείς,
ή το σύνολο

$$Z = \left\{ \underline{\eta} \in \Re^s : \sum f(x;\underline{\eta}) dx < \infty \right\}$$

εάν η τυχαία μεταβλητή X είναι διακριτή.

Η κανονική μορφή της εκθετικής οικογένειας κατανομών μπορεί να χρησιμοποιηθεί αποτελεσματικά για τον υπολογισμό των χαρακτηριστικών των τυχαίων μεταβλητών $T_k(x), k = 1, ..., s$, όπως δείχνει το επόμενο θεώρημα.

Θεώρημα 1 *Εάν X είναι μια τυχαία μεταβλητή που ανήκει στην εκθετική οικογένεια κατανομών ,τότε:*

- $$E[T_i(X)] = \frac{\partial A(\underline{\eta})}{\partial \eta_i}$$
- $$Cov(T_i(X), T_j(X)) = \frac{\partial^2 A(\underline{\eta})}{\partial \eta_i \cdot \partial \eta_j}$$

Πολλές γνωστές κατανομές ανήκουν στην εκθετική οικογένεια κατανομών. Για παράδειγμα η εκθετική κατανομή, η κατανομή *Poisson*, η κατανομή Γάμμα με μία παράμετρο, η διωνυμική κατανομή και η κανονική κατανομή με γνωστή διακύμανση ανήκουν στην E.O.K. και μπορούν να γραφούν στη κανονική της μορφή.

Παραδειγμα 1

Έστω $x \sim binomial(n, p)$ με συνάρτηση πυκνότητας πιθανότητας :

$f(x;p) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} \Rightarrow$

$f(x;p) = \exp[\ln \binom{n}{x} + x \ln p + (n-x)\ln(1-p)] \Rightarrow$

$f(x;p) = \exp[\ln \binom{n}{x} + x \ln p + n\ln(1-p) - x\ln(1-p)] \Rightarrow$

$f(x;p) = \exp[\ln \binom{n}{x} + x \ln \frac{p}{(1-p)} + n\ln(1-p)]$

Όπου $h(x) = \ln \binom{n}{x}$, $\mathrm{T}_k(x) = x$, $B(p) = n\ln(1-p)$, $\eta(p) = \log \frac{p}{(1-p)}$.

Άρα ανήκει στην εκθετική οικογένεια κατανομών. Παραμετροποιώντας την παραπάνω κατανομή σε σχέση με τη φυσική παράμετρο $\eta = \ln \frac{p}{(1-p)}$ έχουμε τη μορφή :

$f(x;\eta) = \exp[\ln \binom{n}{x} + x \cdot \eta - n\ln(1+e^{\eta})]$.

Παραδειγμα 2

Έστω $x \sim Poisson(\lambda)$ με συνάρτηση πυκνότητας πιθανότητας :

$$f(x;\lambda) = \frac{\lambda^x \cdot e^{-\lambda}}{x!} \Rightarrow$$

$$f(x;\lambda) = \exp[x\ln\lambda - \lambda - \ln x!]$$

Οπότε έχουμε: $h(x) = \ln x!$, $T_k(x) = x$, $B(\lambda) = -\lambda$, $\eta(\lambda) = \ln\lambda$.

Οπότε παρατηρούμε ότι ανήκει στην εκθετική οικογένεια κατανομών. Παραμετροποιώντας την παραπάνω κατανομή σε σχέση με τη φυσική παράμετρο $\eta = \ln\lambda$, η *Poisson* κατανομή μπορεί να γραφτεί στη μορφή :

$$f(x;\eta) = \exp[x\eta - \exp\eta - \ln x!]$$

Με παραμετρικό χώρο : $\Theta = \Re$.

Παραδειγμα 3

- Έστω $x \sim N(\mu, \sigma^2)$ με μ: άγνωστο και σ^2: γνωστό, με συνάρτηση πυκνότητας πιθανότητας :

$$f(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp[-\frac{(x-\mu)^2}{2\sigma^2}] \Rightarrow$$

$$f(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp[-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}].$$

Οπότε έχουμε:

$h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp[-\frac{x^2}{2\sigma^2}]$, $T_k(x) = x$, $B(\mu) = \frac{\mu^2}{2\sigma^2}$, $\eta(\mu) = \frac{\mu}{\sigma^2}$ και παρατηρούμε ότι ανήκει στην εκθετική οικογένεια κατανομών. Παραμετροποιώντας την παραπάνω κατανομή σε σχέση με τη φυσική παράμετρο $\eta(\mu) = \frac{\mu}{\sigma^2} = \eta \Rightarrow \mu = \eta \cdot \sigma^2$,

η κανονική κατανομή μπορεί να γραφτεί στη μορφή:

$$f(x;\eta) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp[-\frac{x^2}{2\sigma^2} + x \cdot \eta - A(\eta)]$$

όπου $B(\mu) = \frac{\mu^2}{2\sigma^2} = \frac{(\eta\cdot\sigma)^2}{2\sigma^2} = A(\eta)$.

Οπότε:

$$f(x;\eta) = h(x) \cdot exp[T(x) \cdot (\eta) - A(\eta).$$

- Έστω $x \sim N(\mu, \sigma^2)$ με μ: άγνωστο και σ^2: άγνωστο, με συνάρτηση πυκνότητας πιθανότητας :

$$f(x;\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp[-\frac{(x-\mu)^2}{2\sigma^2}] \Rightarrow$$

$$f(x;\mu) = \frac{1}{\sqrt{2\pi}} \cdot exp[-\frac{1 \cdot ln(\sigma^2)}{2} - \frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}]$$

Οπότε έχουμε ότι:

$h(x) = \frac{1}{\sqrt{2\pi}}$, $\eta_1(\mu,\sigma^2) = -\frac{1}{2\sigma^2}$, $T_1(x) = x^2$, $\eta_2(\mu,\sigma^2) = -\frac{\mu}{\sigma^2}$, $T_2(x) =$

x, $B(\mu,\sigma^2) = \frac{\mu^2}{2\sigma^2} + \frac{1 \cdot ln(\sigma^2)}{2}$ και παρατηρούμε ότι ανήκει στην εκθετική οικογένεια κατανομών. Παραμετροποιώντας την παραπάνω κατανομή σε σχέση με τη φυσική παράμετρο $\eta_1(\mu,\sigma^2) = \eta_1 = -\frac{1}{2\sigma^2} \Rightarrow \sigma^2 = -\frac{1}{2\eta_1}$ και $\eta_2(\mu,\sigma^2) = \eta_2 = \frac{\mu}{\sigma^2} \Rightarrow \mu = \sigma^2 \cdot \eta_2 = -\frac{\eta_2}{2\eta_1}$, η κανονική κατανομή μπορεί να γραφτεί στη μορφή:

$$f(x;\eta) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp[-\frac{x^2}{2\sigma^2} + x \cdot \eta - A(\eta)]$$

όπου $B(\mu) = \frac{\mu^2}{2\sigma^2} = \frac{(\eta\cdot\sigma)^2}{2\sigma^2} = A(\eta)$.

Οπότε:

$$f(x;\eta_1,\eta_2) = h(x) \cdot exp[\eta_1 T_1(x) + \eta_2 T_2(x) - A(\eta_1,\eta_2)].$$

2.4 Γενικευμένα Γραμμικά Μοντέλα

Η πρόοδος στη στατιστική θεωρία μαζί με την ανάπτυξη των υπολογιστών μας επέτρεψαν να δημιουργήσουμε μεθόδους ανάλογους με αυτές που έχουν αναπτυχθεί για τα γραμμικά μοντέλα σε περιπτώσεις που οι μεταβλητές απόκρισης ακολουθούν κατανομή διαφορετική από την κανονική, δεν είναι απαραίτητα συνεχείς (μπορεί να είναι κατηγορικές μεταβλητές) και δεν χρειάζεται να είναι στην απλή γραμμική μορφή : $Y = X \cdot \beta + e$. Μια σημαντική ανακάλυψη είναι ότι πολλές από τις χρήσιμες ιδιότητες της κανονικής κατανομής κατέχει η ομάδα κατανομών που ανήκει στην «εκθετική οικογένεια».

Η εκτίμηση των παραμέτρων του γραμμικού μοντέλου $Y = X \cdot \beta + e$ επεκτάθηκε στην εκτίμηση παραμέτρων συναρτήσεων της μορφής $Y = g(X \cdot \beta) + e$. Θεωρητικά οι διαδικασίες εκτίμησης είναι απλές. Στη πράξη απαιτούν ένα μεγάλο όγκο υπολογισμών οι οποίοι έγιναν εφικτοί μόνο μέσω υπολογιστών με τη βοήθεια αριθμητικών προσεγγίσεων μη γραμμικών συναρτήσεων. Θα δούμε

στη συνέχεια πως επεκτείνονται τα κλασσικά γραμμικά μοντέλα σε γενικευμένα γραμμικά μοντέλα.

Το γενικευμένο γραμμικό μοντέλο είναι ένα "εργαλείο" για την μοντελοποίηση μιας μεταβλητής απόκρισης Υ στη μορφή $g(X \cdot \beta) + e$, όπου η συνάρτηση g είναι μία αυθαίρετη συνδετική συνάρτηση. Τα μοντέλα αυτά επιτρέπουν κάποιο βαθμό μη γραμμικότητας στην σχέση μεταξύ των μεταβλητών X και Υ, διατηρώντας παράλληλα τον κεντρικό ρόλο για τον γραμμικό εκτιμητή $X \cdot \beta$ όπως και στο απλό γραμμικό μοντέλο.

Κάθε γενικευμένο γραμμικό μοντέλο αποτελείται από τρεις συνιστώσες:

1. Την κατανομή της μεταβητής απόκρισης
2. Μια γραμμική παράμετρο πρόβλεψης που περιέχει τις μεταβλητές παλλινδρόμησης x_i
3. Την συνάρτηση σύνδεσης η οποία ενώνει τη γραμμική παράμετρο πρόβλεψης με τη μέση τιμή της απόκρισης

Εξετάζοντας το μοντέλο μας πιο αναλυτικά παρατηρούμε ότι ορίζεται από ένα σύνολο τυχαίων μεταβλητών Υ=(Y_1,Y_2,...,Y_n), οι οποίες ονομάζονται μεταβλητές απόκρισης (*response variables*), κάθε μία από τις οποίες ακολουθεί μια κατανομή Y_i | (X_1,X_2,...,X_p) $\sim$ κατανομή(θ) από την εκθετική οικογένεια κατανομών με θ: διάνυσμα παραμέτρων και X_j: επεξηγηματικές μεταβλητές (*explanatory variables*). Υπάρχει συνάρτηση g, η οποία είναι μονότονη και διαφορήσιμη και η οποία επιλέγεται με τέτοιο τρόπο ώστε να ικανοποιούνται τυχόντες περιορισμοί για το ϑ. Η γραμμική παράμετρος πρόβλεψης του μοντέλου (*linear predictor*), μέσο της οποίας η κατανομή των Y_i εξατρτάται από τα x_i, είναι της μορφής :

$$\eta = \beta_0 + \beta_1 \cdot X_1 + ... + \beta_p \cdot X_p$$

ώστε ο τρόπος σύνδεσης των παραμέτρων ή κάποιας εκ των παραμέτρων της τυχαίας συνιστώσας με τη γραμμική παράμετρο πρόβλεψης του μοντέλου να

είναι μέσω μιας συνάρτησης σύνδεσης (*link function*) g έτσι ώστε :

$$g(\vartheta) = \eta = \beta_0 + \beta_1 \cdot X_1 + ... + \beta_p \cdot X_p.$$

Συνήθως έχουμε ένα τυχαίο δείγμα n παρατηρήσεων (Y_i,X_{1i},...,X_{pi}, i=1,...,p) με την βοήθεια του οποίου εκτιμούμε τα (β_0,β_1,.....,β_p).

Παραδειγμα 1

Έστω τυχαίες συνιστώσες (*random components*) $Y_1, \ldots, Y_n$ κανονικά κατανεμημένες. Η κανονική κατανομή ανήκει στην εκθετική οικογένεια κατανομών. Επιπλέον οι επεξηγηματικές μεταβλητές (*explanatory variables*) εισάγουν το γραμμικό μοντέλο διαμέσου ενός γραμμικού εκτιμητή (*linear predictor*):

$$\eta_i = x_i^T \cdot \beta$$

οπότε η σύνδεση μεταξύ των $E(Y) = \mu$ και του γραμμικού εκτιμητή η γίνεται διαμέσου της συνδετικής συνάρτησης:

$$\mu_i = \eta_i, \ \imath = 1, \ldots, n.$$

Παραδειγμα 2

Έστω τυχαία συνιστώσα (*random component*) $Y_i \sim binomial(n_i, p_i)$ όπου Y_i εκφράζει τον αριθμό επιτυχιών σε n_i επαναλήψεις και οι επεξηγηματικές μεταβλητές (*explanatory variables*) X_j μπορεί να είναι συνεχείς ή διακριτές, j = 1,...,p. Σε προηγούμενο παράδειγμα (παράγραφος 2.3 παράδειγμα 1) είδαμε ότι η διωνυμική κατανομή ανήκει στην εκθετική οικογένεια κατανομών με $A(\eta) = n \cdot log(1 + e^{\eta})$.

Από το θεώρημα 1 της παραγράφου 2.3 γνωρίζουμε ότι :

$$n \cdot p = E(Y) = A'(\eta) \Rightarrow$$

$$\eta = (A')^{-1}(n \cdot p) \Rightarrow$$

$$\eta = log(\frac{p}{1-p})$$

άρα η συνδετική συνάρτηση ($link\ function$) που προκύπτει είναι :

$$\log(p_i/(1-p_i)) = \eta_i = \beta_0 + \beta_1 \cdot X_{1i} + ... + \beta_p \cdot X_{pi}$$

οπότε :

$$g(p) = logit(p) \equiv log(p/(1-p)).$$

Παρατηρούμε ότι η συνάρτηση ορίζεται $g : \Re \to (0,1)$.

Παραδειγμα 3

Έστω τυχαία συνιστώσα ($random\ component$) $Y_i \sim Poisson(\lambda_i)$, όπου Y_i εκφράζει τον αριθμός εμφανίσεων γεγονότων σε ένα χρονικό/χωρικό διάστημα και οι επεξηγηματικές μεταβλητές ($explanatory\ variables$) X_j μπορεί να είναι συνεχείς ή διακριτές, $j = 1, ..., p$. Σε προηγούμενο παράδειγμα (παράγραφος 2.3 παράδειγμα 2) είδαμε ότι η κατανομή $Poisson$ ανήκει στην εκθετική οικογένεια κατανομών με $A(\lambda) = \lambda$.

Από το θεώρημα 1 της παραγράφου 2.3 γνωρίζουμε ότι :

$$\lambda = E(Y) = A'(\eta) \Rightarrow$$

$$\eta = (A')^{-1}(\lambda) \Rightarrow$$

$$\eta = log(\lambda)$$

άρα η συνδετική συνάρτηση ($link\ function$) η οποία προκύπτει είναι :

$$\log(\lambda_i) = \eta_i = \beta_0 + \beta_1 \cdot X_{1i} + ... + \beta_p \cdot X_{pi}.$$

Οπότε :

$$g(\lambda) = log(\lambda).$$

Στο $Poisson$ μοντέλο παρατηρούμε ότι $\mu = E(Y) > 0$ οπότε δεν μπορούμε να έχουμε $\mu = X^T \cdot \beta$ διότι για το δεξιό μέλος δεν ισχύει ο περιορισμος. Οπότε σε αυτή τη περίπτωση η συνάρτηση ορίζεται $g{:}\Re^+ \to \Re$.

2.4.1 Κλασική Εκτίμηση για Γενικευμένα Γραμμικά Μοντέλα

Το Γ.Γ.Μ. ορίζεται σε σχέση με ένα σύνολο από ανεξάρτητες τυχαίες αποκρίσεις $y_1, \cdots, y_n$ με μέσους $\mu_1, \cdots, \mu_n$, όπου κάθε μία από αυτές ακολουθούν κατανομές από την εκθετική οικογένεια κατανομών, που έχουν την ίδια συναρτησιακή μορφή. Δηλαδή

$$f(y_i; \eta_i) = \exp[\eta_i T_i(y_i) - A(\eta_i) + H(y_i)].$$

Το μοντέλο περιέχει μεταβλητές παλινδρόμησης τις $x_1, \cdots, x_p$, και κατασκευάζεται βάση της γραμμικής παραμέτρου πρόβλεψης:

$$\boldsymbol{\eta} = x^T \cdot \beta = \beta_0 + \sum_{i=1}^{p} \beta_i \cdot x_i.$$

Η σύνδεση ανάμεσα στην κατανομή των Y και στη γραμμική παράμετρο πρόβλεψης $\boldsymbol{\eta}$ γίνεται μέσω της συνάρτησης σύνδεσης:

$$\boldsymbol{\eta}_i = g(\mu_i), i = 1, \cdots, n$$

όπου $\mu_i \equiv E(Y_i), i = 1, \cdots, n$ οπότε

$$g(\mu_i) = g(E(Y_i)) = \boldsymbol{\eta}_i = x_i^T \cdot \beta, i = 1, \cdots, n.$$

Γνωρίζουμε ότι οι κατανομές των Y ανήκουν στην εκθετική οικογένεια κατανομών και από το θεώρημα 1 στη παράγραφο 2.3 έχουμε ότι η $E(Y) = A'(\eta)$ οπότε $\eta = (A')^{-1}(\mu)$ και δεδομένου του ότι $g(\mu) = \boldsymbol{\eta} = x_i^T\beta$ έχουμε $\eta = (A')^{-1}(g^{-1}(x_i^T\beta))$. Θεωρώντας ότι οι συναρτήσεις g και $(A')^{-1}$ είναι πανομοιότυπες έχουμε ότι $\eta = x_i^T\beta$. Ορίζουμε ως κανονική συνάρτηση σύνδεσης (*canonikal link function*) την $g(\mu) = (A')^{-1}(\mu)$, η οποία δηλώνει ότι υπάρχει σύνδεση μεταξύ του μέσου και της γραμμικής παραμέτρου πρόβλεψης με μέση τιμή απόκρισης:

$$E(y_i) = g^{-1}(\boldsymbol{\eta}_i) = g^{-1}(x_i^T \cdot \beta).$$

Όσον αφορά στις συναρτήσεις σύνδεσης υπάρχει μεγάλη ποικιλία. Ακολουθεί ένας πίνακας των κανονικών συνδετικών συναρτήσεων που χρησιμοποιούνται για διάφορες κατανομές από την εκθετική οικογένεια κατανομών στα Γ.Γ.Μ.
.

Πίνακας 2.1: Συνδετικές συναρτήσεις διαφόρων κατανομών

Κατανομή	Κανονική Συνδετική Συνάρτηση
Κανονική	$\boldsymbol{\eta}_i = \mu_i$
Διωνυμική	$\boldsymbol{\eta}_i = log(p/(1-p))$
Poisson	$\boldsymbol{\eta}_i = log(\mu_i)$
Εκθετική	$\boldsymbol{\eta}_i = \frac{1}{\mu_i}$
Γάμμα	$\boldsymbol{\eta}_i = \frac{1}{\mu_i}$
Αντίστροφη Κανονική	$\boldsymbol{\eta}_i = \frac{1}{\mu_i^2}$

Η εκτίμηση των παραμέτρων των Γ.Γ.Μ. δεν μπορεί να γίνει αναλυτικά, λόγω της πολυπλοκότητας τους. Έτσι πρέπει να καταφύγουμε σε κάποια μέθοδο αριθμητικής μεγιστοποίησης της πιθανοφάνειας. Μια τέτοια μέθοδος είναι ο αλγόριθμος *Newton-Raphson*.

Η μέθοδος *Newton-Raphson* έχει τα εξής βήματα:

Θέλουμε να μεγιστοποιήσουμε την συνάρτηση πιθανοφάνειας

$$logL(\vartheta) = l(\vartheta).$$

1. Δίνουμε αρχικές τιμές στο $\vartheta = (\vartheta_1, \dots \vartheta_p)$.

2. Παίρνουμε
$$\vartheta_{t+1} = \vartheta_t - \frac{l^{'}(\vartheta_t)}{l^{"}(\vartheta_t)}.$$

3. Επαναλαμβάνουμε το βήμα 2 έως ότου $\vartheta_{t+1} \approx \vartheta_t$, δηλαδή να επιτευχθεί σύγκλιση.

Πολλές φορές ανάλογα με τη μορφή της συνδετικής συνάρτησης είναι απαραίτητο να εφαρμόσουμε ένα διαφορετικό αλγόριθμο. Μία μορφή της μεθόδου *Newton-Raphson* όταν έχει χρησιμοποιηθεί η κανονική συνδετική συνάρτηση είναι ο αλγόριθμος *Fisher scoring* ο οποίος είναι μία ειδική περίπτωση του αλγόριθμου επαναλαμβανόμενων σταθμισμένων ελαχίστων τετραγώνων. Πρώτοι εφάρμοσαν τον αλγόριθμο αυτό για να εκτιμήσουν τις παραμέτρους των γενικευμένων γραμμικών μοντέλων οι *Nedler* και *Wedderburn* (1972).

Ο αλγόριθμος *Fisher scoring*, προσαρμοσμένος στις ανάγκες των γενικευμένων γραμμικών μοντέλων έχει τα εξής βήματα:

1. Δίνουμε αρχικές τιμές για τα ϑ.
2. Παίρνουμε
$$\vartheta_{t+1} = \vartheta_t + \frac{l^{'}(\vartheta_t)}{E(l^{"}(\vartheta_t))}$$
όπου:
$l^{'}(\vartheta_t) = X^T W z$
με
z είναι ένα διάνυσμα: $z_i = (Y_i - \mu_i) g^{'}(\mu_i)$
W είναι ένας διαγώνιος πίνακας: $W_{ii} = (g^{'}(\mu_i)^2 A^{"}(\eta_i))^{-1}, i = 1, \ldots, n$
και
$E(l^{"}(\vartheta_t)) = X^T W_t X$
οπότε συνδιάζοντας όλα τα παραπάνω έχουμε:
$\vartheta_{t+1} = (X^T W_t X)^{-1} X^T W_t (\eta_t + z_t)$
3. Επαναλαμβάνουμε το βήμα 2 μέχρι να επιτευχθεί σύγκλιση.

Η διαφορά των δύο παραπάνω αλγορίθμων είναι ότι ο αλγόριθμος *Newton-Raphson* χρησιμοποιεί την $l^{"}(\vartheta)$ δηλαδή την παρατηρούμενη πληροφορία του

Fisher, ενώ ο αλγόριθμος *Fisher scoring* χρησιμοποιεί την $E(l^{"}(\vartheta))$ δηλαδή την αναμενόμενη πληροφορία του *Fisher*.

Κεφάλαιο 3

Γενικευμένα Γραμμικά Μοντέλα για Διωνυμικά και *Poisson* Δεδομένα

Στα γενικευμένα γραμμικά μοντέλα τα δύο πιο σημαντικά μοντέλα που συναντάμε είναι αυτά της Λογιστικής και της *Poisson* παλινδρόμησης τα οποία έχουν πολλές εφαρμογές. Πρόκειται ουσιαστικά για Γ.Γ.Μ. για διωνυμικά και *Poisson* δεδομένα αντίστοιχα.

3.1 Γ.Γ.Μ. για Διωνυμικά Δεδομένα

Η λογιστική παλινδρόμηση είναι μια μέθοδος πολυπαραγοντικής στατιστικής ανάλυσης (*multivariate statistical analysis*) που χρησιμοποιεί ένα σύνολο ανεξαρτήτων μεταβλητών (*independent variables*) για τη διερεύνηση της κίνησης μιας κατηγορικής εξαρτημένης μεταβλητής (*dependent variable*).

Η λογιστική παλινδρόμηση (*Logistic Regression*) είναι χρήσιμη σε καταστάσεις στις οποίες επιθυμούμε την πρόβλεψη της ύπαρξης ή της απουσίας

ενός χαρακτηριστικού ή ενός συμβάντος. Η πρόβλεψη αυτή βασίζεται στην κατασκευή ενός γραμμικού μοντέλου και συγκεκριμένα στον προσδιορισμό των τιμών που παίρνουν οι συντελεστές ενός συνόλου (set) ανεξάρτητων μεταβλητών που χρησιμοποιούνται ως μεταβλητές πρόβλεψης ($predictor\ variables$).

Για την αντιμετώπιση τέτοιων προβλημάτων το υπόδειγμα της γραμμικής παλλινδρόμησης δεν είναι κατάλληλο για την εκτίμηση των τιμών της ανεξάρτητης μεταβλητής από τις μέσες τιμές των εξαρτημένων.

Σε μια τέτοια περίπτωση χρησιμοποιώντας την τιμή 1 για το ενδεχόμενο της "επιτυχίας" και τη τιμή 0 για το ενδεχόμενο της 'αποτυχίας', ο υπολογισμός της μέσης τιμής της εξαρτημένης δίτιμης μεταβλητής ορίζει την αναλογία p των επιτυχιών στο σύνολο των δυνατών τιμών της.

Η τεχνική με την οποία εκτιμάτε η πιθανότητα επιτυχίας p μιας δίτιμης μεταβλητής για ένα σύνολο τιμών μιας ή περισσότερων ανεξάρτητων μεταβλητών ονομάζετε λογιστική παλινδρόμηση.

Εκτός από την πρόβλεψη ένα μοντέλο λογιστικής παλινδρόμησης δίνει τη δυνατότητα να εκτιμήσουμε την επίδραση κάθε ανεξάρτητης μεταβλητής στη διαμόρφωση των τιμών της εξαρτημένης μεταβλητής. Στη λογιστική παλινδρόμηση, σε αντίθεση με την πολλαπλή παλινδρόμηση ($multiple\ regression$) είναι δυνατό να χρησιμοποιηθούν ως εξαρτημένες μεταβλητές εκτός από αναλογικές αριθμητικές μεταβλητές ($ratio\ scale$) και κατηγορικές μεταβλητές ($nominal\ scale$).

Η πιο διαδεδομένη βιβλιογραφικά έκφραση της λογιστικής παλινδρόμησης η οποία συνδέει την πιθανότητα επιτυχίας p με την ανεξάρτητη μεταβλητή X όταν:

$y_i \sim Binomial\,(n_i, p_i)$ είναι:

$$logit(p_i) = log\left(\frac{p_i}{1-p_i}\right) = \eta_i = x_i^T \cdot \beta$$

στην οποία έχουμε συνδετική συνάρτηση $logit$ όπως έχουμε δείξει στην παρά-

γραφο 2.4 στο παράδειγμα 2.

Το δεξί μέρος των εξισώσεων δημιουργείται από ένα γραμμικό συνδυασμό των ανεξάρτητων μεταβλητών που συμμετέχουν στο μοντέλο παλλινδρόμησης.

Το αριστερό μέρος περιέχει τις τιμές της εξαρτημένης μεταβλητής με τη μορφή του *logit* και ο όρος p εκφράζει την πιθανότητα του συμβάντος του γεγονότος που έχει ορισθεί ως επιτυχία του πειράματος.

Παρόμοιο με το λογιστικό υπόδειγμα είναι το κανονικό υπόδειγμα πιθανότητας. Η κύρια διαφορά τους είναι ότι σε αυτό για τον υπολογισμό της πιθανότητας p χρησιμοποιείται η αθροιστική συνάρτηση πιθανότητας της κανονικής κατανομής δηλαδή εάν:

$y_i \sim Bernoulli\left(\Phi\left(\eta_i\right)\right)$ τότε:

$$probit(p_i) = \Phi^{-1}\left(p_i\right)$$

οπότε εδώ έχουμε συνδετική συνάρτηση *probit*.

Η σχέση που υπάρχει ανάμεσα στις συνδετικές συναρτήσεις *probit* και *logit* για $Z = x_i^T \cdot \beta$ είναι:

$$Z^{logit} = \frac{\pi}{\sqrt{3}} Z^{probit}.$$

3.1.1 Το Μοντέλο της Λογιστικής Παλινδρόμησης

Αν προσπαθήσουμε να εκφράσουμε την πιθανότητα επιτυχίας p, μιας δίτιμης μεταβλητής Y, η οποία ακολουθεί Διωνυμικη κατανομή, με την βοήθεια ενός απλού γραμμικού μοντέλου

$$p = x_i^T \cdot \beta$$

όπου x οι τιμές μιας ανεξάρτητης μεταβλητής X, το κύριο πρόβλημα που θα αντιμετοπίσουμε είναι ότι αν και οι τιμές της p θεωρητικά δεν μπορούν να βρίσκονται εκτός του διαστήματος [0,1], οι τιμές της ποσότητας $x_i^T \cdot \beta$ κυμαίνονται σε όλο το εύρος των πραγματικών αριθμών. Για την αντιμετώπιση αυτού του

προβλήματος αναπαραμετροποιούμε το μοντέλο ως εξης. Θεωρούμε το φυσικό λογάριθμο του λόγου της πιθανότητας επιτυχίας προς την πιθανότητα αποτυχίας. Ονομάζουμε *odds* τον παραπάνω λόγο και έχουμε

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

καθώς και το λογάριθμο

$$log - odds = log\left(\frac{p}{1-p}\right).$$

Οι τιμές του μετασχηματισμένου λόγου κυμαίνονται στο διάστημα $(-\infty, +\infty)$.

Οπότε έχουμε ότι

$$log\left(\frac{p_i}{1-p_i}\right) = x_i^T \cdot \beta.$$

Θέτοντας $Z = x_i^T \cdot \beta$ και αντιλογαριθμίζοντας τα δύο μέλη της εξίσωσης έχουμε

$$p = \frac{e^Z}{1+e^Z}.$$

Η τελευταία σχέση αποτελεί την εκτίμηση της πιθανότητας επιτυχίας της δίτιμης μεταβλητής Y, για δεδομένες τιμές των επεξηγηματικών μεταβλητων.

Ερμηνεύοντας τους συντελεστές της Λογιστικής παλινδρόμησης παρατηρούμε ότι για αύξηση της επεξηγηματικής μεταβλητής κατά μία μονάδα τα *odds* αυξάνονται παλλαπλασταστικά κατα exp του αντίστοιχου β ή εναλλακτικά τα $log - odds$ αυξάνονται κατά αντίστοιχο β.

3.2 *Poisson* Παλινδρόμηση

Η κατανομή *Poisson* είναι η κατανομή των σπάνιων γεγονότων και χρησιμοποιείται όταν θέλουμε να μετρήσουμε τον αριθμό των ΄συμβάντων΄ στη μονάδα μέτρησης.Τα συμβάντα μπορεί να είναι για παράδειγμα το πλήθος των ανθρώπων που πάσχουν από μία ασθένια σε μία περιοχή. Η τυχαία μεταβλητή Y

που ακολουθεί την κατανομή $Poisson$ εκφράζει το πλήθος των συμβάντων στη μονάδα μέτρησης και η συνάρτηση πιθανότητας δίνεται από τον τύπο:

$$f(y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, i = 1, \ldots, n.$$

Στην $Poisson$ Παλινδρόμηση χρησιμοποιούμε τη λογαριθμική συνδετική συνάρτηση η οποία είναι της μορφής:

$$g(\mu_i) = log(\mu_i) = \eta_i = x_i^T\beta, i = 1, \ldots, n.$$

Αξίζει να επισημάνουμε ότι πολλές φορές στην $Poisson$ Παλινδρόμηση αντί να έχουμε αριθμό εμφανίσεων γεγονότων σε κάποια μονάδα (χρόνου, χώρου, κλπ.), έχουμε το ρυθμό εμφανίσεων των γεγονώτων. Για παράδειγμα ο αριθμός των ασθενών που αρρώστησαν διαφέρει ανάλογα με την περιοχή την οποία εξετάζουμε. Στην περίπτωση αυτή χρησιμοποιούμε κάποιους όρους σαν αντιστάθμισμα τους οποίους ανομάζουμε $offset$.

Έτσι αντί να έχουμε:

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

έχουμε:

$$log\mu_i = loge_i + \beta_0' + \beta_1' x_i$$

Όπου το $loge_i$ παίζει το ρόλο του $offset$ για την i παρατήρηση. Ερμηνεύοντας τους συντελεστές της $Poisson$ παλινδρόμησης βλέπουμε ότι για αύξηση της μεταβλητής x κατά μία μονάδα, παρατηρείται αύξηση του λογαρίθμου της αναμενόμενης τιμής της μεταβλητής y, $log\mu_i = log(E(y_i))$, κατά αντίστοιχο β.

Κεφάλαιο 4

Μπεϋζιανή Στατιστική

Το κεφάλαιο αυτό αφιερώνεται στη παρουσίαση της στατιστικής θεωρείας του *Bayes* καθώς επίσης και στην αναλυτική περιγραφή των *Markov Chain Monte Carlo* μεθόδων που χρησιμοποιούμε στην παρούσα εργασία.

4.1 Ιστορική Αναδρομή

Στις αρχές του 21ου αιώνα η Μπεϋζιανή στατιστική βρίσκεται να παίζει ένα πολύ σημαντικό ρόλο στη στατιστική συμπερασματολογία. Έως τα τέλη της δεκαετίας του 1980, θεωρούσαμε τη Μπεϋζιανή στατιστική ως μια ενδιαφέρουσα εναλλακτική θεωρία της Κλασικής στατιστικής. Η βασική διαφορά ανάμεσα στη Κλασική στατιστική θεωρία και στη Μπεϋζιανή προσέγγιση είναι ότι η δεύτερη θεωρεί τις παραμέτρους ως τυχαίες μεταβλητές οι οποίες χαρακτηρίζονται από μια εκ των προτέρων κατανομή. Η εκ των προτέρων κατανομή συνδιάζεται με την συνάρτηση πιθανοφάνειας και μας δίνουν την εκ των υστέρων κατανομή των παραμέτρων στις οποίες είναι βασισμένη η στατιστική συμπερασματολογία.

Παρόλο που το κύριο εργαλείο της Μπεϋζιανής θεωρίας είναι η θεωρία πιθανοτήτων, για πολλά χρόνια οι Μπεϋζιανοί θεωρούνταν ως μειονότητα για

αρκετούς λόγους. Η βασική αντίρρηση από τους κλασικούς στατιστικούς για την Μπεϋζιανή θεωρία, εντοπίζεται στο γεγονός ότι τα συμπεράσματα εξαρτώνται από την επιλογή της εκ των προτέρων κατανομής. Όπως η ιστορία μας έδειξε, ο κύριως λόγος για τον οποίο η Μπεϋζιανή θεωρία καθυστέρησε να εδρεωθεί ως μία αποδεκτή προσέγγιση για την ανάλυση στατιστικών δεδομένων ήταν το ότι για τον υπολογισμό της εκ των υστέρων κατανομής εμπλέκονταν οι εκ των προτέρων πεποιθήσεις μας.

Ασυμπτωτικές μέθοδοι έχουν βρεθεί για να λύνουμε συγκεκριμένα προβλήματα, αλλα δεν πρέπει να το γενικεύσουμε. Έως τις αρχές του 1990 δύο ομάδες στατιστικών εισήγαγαν την *Markov chain Monte Carlo* ($MCMC$) μέθοδο (*Gelfand and Smith*, 1990 και *Gelfand et al.*,1990) στην Μπεϋζιανή συμπερασματολογία. Στο τομέα της φυσικής χρησιμοπούν την $MCMC$ μέθοδο από το 1950. Ο *Nick Metropolis* και οι συνεργάτες του ανέπτυξαν έναν από τους πρώτους πιο εξελιγμένους ηλεκτρονικούς υπολογιστές εκείνης της εποχής και δοκίμασαν τις θεωρίες της φυσικής που είχαν ανακαλύψει χρησιμοποιώντας $MCMC$ τεχνικές. Η εφαρμογή της $MCMC$ μεθόδου σε συνδιασμό με την ραγδαία ανάπτυξη των υπολογιστών έκαναν αυτό το πολύ χρήσιμο υπολογιστικό εργαλείο πολύ δημοφιλές μέσα σε λίγα χρόνια. Η Μπεϋζιανή στατιστική ξαφνικά έγινε πολύ γνωστή και άνοιξε νέους ορίζοντες για την στατιστική έρευνα. Χρησιμοποιώντας τους αλγόριθμους $MCMC$, μπορούμε να δημιουργήσουμε και να εκτιμήσουμε σύνθετα μοντέλα που περιγράφουν πολύπλοκα προβλήματα τα οποία με τις παραδοσιακές μεθόδους δεν θα ήταν δυνατό να επιλυθούν. Από το 1990, όταν η $MCMC$ μέθοδο πρωτοεμφανίστηκε στην επιστήμη της στατιστικής, πολλές σημαντικές εργασίες γράφτηκαν. Κατά τη διάρκεια 1990-1995, η έρευνα πάνω στους αλγόριθμους $MCMC$ επικεντρώθηκε στο να εφαρμοστεί αυτή η νέα για εκείνη την εποχή μεθοδο σε διάφορα πολύ δημοφιλή μοντέλα [*Gelman and Rubin* (1992), *Gelfand, Smith and Lee* (1992), *Gilks*

and Wild(1992), *Delaportas and Smith* (1993)]. Η ανάπτυξη των $MCMC$ μεθόδων προόθησε επίσης στην ανάπτυξη των τυχαίων επιδράσεων και των ιεραρχικών μοντέλων. Η ανάπτυξη σε συνδιασμό με την επέκταση των $MCMC$ μέθοδων χρησιμοποιείται στη στατιστική έρευνα από τα μέσα της δεκαετίας του 1990.

Στο παρόν κεφάλαιο θα δούμε αρχικά μία μικρή εισαγωγή της Μπεϋζιανής θεωρίας και στη συνέχεια θα επικεντρώσουμε το ενδιαφέρον μας στην ανάλυση των πιο διαδεδομένων $MCMC$ μέθοδων οι οποίες χρησιμοποιούνται ευρέως στη Μπεϋζιανή στατιστική.

4.2 Βασικές Αρχές της Μπεϋζιανής Θεωρίας

Το πλαίσιο στο οποίο κινείται η συμπερασματολογία κατά *Bayes* είναι παρόμοιο με αυτό της κλασικής στατιστικής, δηλαδή υπάρχει η άγνωστη παράμετρος ϑ του πληθυσμού την οποία θέλουμε να εκτιμήσουμε, καθώς και η πιθανότητα $f(x|\vartheta)$ η οποία καθορίζει την πιθανότητα παρατήρησης διαφορετικών x, κάτω από διαφορετικές τιμές της παραμέτρου ϑ. Όμως η θεμελιώδης διαφορά είναι ότι το ϑ χρησιμοποιείται σαν τυχαία ποσότητα. Αν και η διαφορά αυτή μπορεί να φανεί όχι και τόσο ουσιαστική, οδηγεί σε μία τελείως διαφορετική προσέγγιση, ως προς την ερμηνεία, από αυτήν της κλασικής στατιστικής.

Στην ουσία, η συμπερασματολογία μας θα βασιστεί στην $f(\vartheta|x)$ και όχι στην $f(x|\vartheta)$, δηλαδή στην πιθανότητα της κατανομής της παραμέτρου δεδομένης της x (δεδομένα) και όχι της x δεδομένης της παραμέτρου. Σε πολλές περιπτώσεις αυτό οδηγεί σε περισσότερο φυσικά συμπεράσματα σε σχέση με την κλασική στατιστική.

Για να μπορέσει όμως να επιτευχθεί αυτό θα πρέπει να καθορίσουμε την εκ

των προτέρων κατανομή $f(\vartheta)$ (*prior probability distribution*), η οποία αντιπροσωπεύει τις «πεποιθήσεις» μας για την κατανομή της παραμέτρου ϑ προτού αποκτήσουμε οποιαδήποτε πληροφορία για τα δεδομένα μας.

Η ιδέα της εκ των προτέρων κατανομής της παραμέτρου αποτελεί και την «καρδιά» της θεωρίας κατά *Bayes*, και βασιζόμενοι στο αν μιλάμε σε έναν υπερασπιστή ή σε έναν αντιμαχόμενο της συγκεκριμένης μεθοδολογίας, η εκ των προτέρων κατανομή μπορεί να αποτελέσει το μεγαλύτερο πλεονέκτημα ή το σοβαρότερο μειονέκτημα έναντι της κλασικής στατιστικής.

4.3 Θεώρημα του *Bayes*

Η Μπεϋζιανή στατιστική διαφέρει από τη κλασική στατιστική θεωρία στο ότι όλες οι άγνωστες παράμετροι θεωρούνται ως τυχαίες μεταβλητές. Για το λόγο αυτό, η εκ των προτέρων κατανομή πρέπει να καθορίζεται από την αρχή. Η εκ των προτέρων κατανομή εκφράζει την εκ των προτέρων γνώση και πεποίθηση μας πρωτού λάβουμε υπόψιν μας τα δεδομένα. Έτσι η συμπερασματολογία δεν προκύπτει μόνο από τη μελέτη των δεδομένων, δηλαδή την πιθανοφάνεια $f(\vartheta/y)$, αλλά τελικά από τη συνάρτηση πιθανότητας της κατανομής της παραμέτρου δοθέντος των δεδομένων $f(y/\vartheta)$, η οποία καλείται εκ των υστέρων κατανομή (*posterior distribution*).

Ο υπολογισμός της δίνεται από το θεώρημα του *Bayes*:

$$f(\vartheta/y) = \frac{f(y/\vartheta) \cdot f(\vartheta)}{f(y)},$$

όπου η συνάρτηση πιθανοφάνειας δίνετε από το τύπο:

$$f(y/\vartheta) = \prod_{i=1}^{n} f(y_i/\vartheta)$$

Εάν ο παραμετρικός χώρος ϑ είναι διακριτός τότε η σταθερά κανονικοποίησης

δίνεται από το άθροισμα:

$$f(y) = \sum_{\vartheta_i \in \Theta} f(\vartheta_i) \cdot f(y/\vartheta_i)$$

Εάν ο παραμετρικός χώρος ϑ είναι συνεχής τότε η σταθερά κανονικοποίησης δίνεται από το ολοκλήρωμα:

$$f(y) = \int_{\vartheta_i \in \Theta} f(\vartheta_i) \cdot f(y/\vartheta_i) d\vartheta$$

Δεδομένου του ότι η σταθερά κανονικοποίησης είναι συνάρτηση μόνο των δεδομένων y και δεν εξαρτάται καθόλου από την παράμετρο ϑ συχνά στη βιβλιογραφία παρουσιάζεται το θεώρημα $Bayes$ ως:

$$f(\vartheta/y) \propto f(y/\vartheta) \cdot f(\vartheta)$$

δηλαδή η εκ των υστέρων κατανομή είναι ανάλογη του γινομένου της εκ των προτέρων κατανομής πολλαπλασιασμένη με τη συνάρτηση πιθανοφάνειας.

Παραδειγμα 1

Έστω $y_1, y_2, \cdots, y_n$ ένα τυχαίο δείγμα που ακολουθεί την $Poisson$ κατανομή με συνάρτηση πιθανοφάνειας:

$$f(y/\vartheta) = \frac{e^{-n\vartheta} \cdot \vartheta^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y!} \propto e^{-n\vartheta} \cdot \vartheta^{\sum_{i=1}^n y_i}, i = 1, \cdots, n$$

και με εκ των προτέρων κατανομή την

$$f(\vartheta) = e^{-\vartheta}, \vartheta > 0.$$

Εφαρμόζοντας το θεώρημα $Bayes$ έχουμε :

$$\begin{aligned} f(\vartheta/y) &= f(\vartheta) \cdot f(y/\vartheta) \\ &\propto e^{-\vartheta} \cdot \frac{e^{-n\vartheta} \cdot \vartheta^{\sum_{i=1}^n y_i}}{\prod_{i=1}^n y!} \\ &\propto e^{-\vartheta} \cdot e^{-n\vartheta} \cdot \vartheta^{\sum_{i=1}^n y_i} \end{aligned}$$

$$\propto e^{-(n+1)\vartheta} \cdot \vartheta^{\sum_{i=1}^{n} y_i + 1 - 1}$$

Άρα

$$f(\vartheta/y) \sim Gamma(\sum_{i=1}^{n} y_i + 1, n + 1)$$

4.4 Εκ των Προτέρων Κατανομές

Έχουμε ήδη δει ότι η βασική διαφορά μεταξύ την Μπεϋζιανής θεωρίας και της κλασικής στατιστικής είναι ότι σύμφωνα με την Μπεϋζιανή στατιστική οι άγνωστες παράμετροι χρησιμοποιούνται σαν τυχαίες μεταβλητές και για αυτόν τον λόγο η χρησιμοποίηση του θεωρήματος του *Bayes* απαιτεί τον καθορισμό εκ των προτέρων κατανομών για τις μεταβλητές αυτές. Ο καθορισμός της εκ των προτέρων κατανομής είναι ένας πολύ σημαντικός παράγοντας για την Μπεϋζιανή συμπερασματολογία, καθώς επιρρεάζει την εκ των υστέρων κατανομή. Διαφορετική εκ των προτέρων κατανομή οδηγεί σε διαφορετικά συμπεράσματα. Η προσωπική επιλογή της εκ των προτέρων κατανομής καθιστά την ανάλυση υποκειμενική. Ωστόσο μια «λογική» επιλογή της εκ των προτέρων κατανομής χάνει την επίδρασή της καθώς συγκεντρώνονται όλο και περισσότερα δεδομένα. Συνήθως, ιδιαίτερη έμφαση δίνεται στον προσδιορισμό του μέσου και της διασποράς της εκ των προτέρων κατανομής. Ο μέσος μας παρέχει ένα εκ των προτέρων σημειακό εκτιμητή για την παράμετρο που μελετάμε, ενώ η διασπορά εκφράζει την αβεβαιώτητα σχετικά με αυτή την εκτίμηση. Όταν έχουμε μια εκ των προτέρων ισχυρή πεποίθηση ότι η εκτίμησή μας είναι ακριβής, τότε η διασπορά πρέπει να είναι μικρή, ενώ η μεγάλη αβεβαιότητα σχετικά με την εκτίμηση μας οδηγεί σε αυξημένη τιμή της διασποράς. Εάν είναι διαθέσιμη εκ των προτέρων πληροφορία τότε αυτή μπορούμε να τη συνοψίσουμε σε μία «κατάλληλη» συναρτησιακή μορφή της κατανομής που να μας διευκολίνει υπολογιστικά. Η διαδικασία αυτή ονομάζεται «*elicitation*» της εκ των προτέρων πληροφορίας.

Σε πολλές περιπτώσεις όμως η διαθέσιμη εκ των προτέρων πληροφορία είναι περιορισμένη. Στην περίπτωση αυτή επιθυμούμε η πληροφορία που προκύπτει από τα δεδομένα να κυριαρχίσει στον υπολογισμό της εκ των υστέρων κατανομής. Σε αυτή τη περίπτωση χρειαζόμαστε να προσδιορίσουμε μια εκ των προτέρων κατανομή η οποία να μην επιρρεάζει την εκ των υστέρων κατανομή. Μία τέτοια κατανομή συχνά καλείται μη πληρφοριακή ($noninformative$) εκ των προτέρων κατανομή. Πέρα από τη διαφοροποίηση των εκ των προτέρων κατανομών σε πληροφοριακέν και μη πληροφοριακέν υπάρχουν διάφορες κατηγορίες εκ των προτέρων κατανομών.

Μερικές από τις κατηγορίες των εκ των προτέρων κατανομών είναι οι ακόλουθες:

- Συζυγείς εκ των προτέρων κατανομές ($Conjugate\ priors$)
- Ακατάλληλες εκ των προτέρων κατανομές ($Improper\ priors$)
- Μη πληροφοριακή εκ των προτέρων κατανομή του $Jeffrey$ ($Jeffrey's\ Prior$)

4.4.1 Συζυγείς εκ των προτέρων κατανομές

Η χρησιμοποίηση του θεωρήματος του $Bayes$ συνεπάγεται αρκετές υπολογιστικές δυσκολίες, που αφορούν στον υπολογισμός της σταθεράς κανονικοποίησης. Για το λόγο αυτό για αρκετό διάστημα η Μπεϋζιανή θεωρία περιορίστηκε σε κατανομές που διευκόλυναν τον υπολογισμό των εκ των υστέρων κατανομών. Τέτοιες κατανομές είναι οι συζυγείς εκ των προτέρων κατανομές. Ως συζυγείς ορίζονται οι εκ των προτέρων κατανομές οι οποίες εφαρμόζοντας το θεώρημα του $Bayes$ καταλήγουν σε εκ των υστέρων κατανομές που ανήκουν στην ίδια οικογένεια κατανομών με την εκ των προτέρων κατανομή. Είναι σημαντικό, να σημειωθεί ότι η χρήση των συζυγών εκ των προτέρων κατανομών δε γίνεται μό-

νο για λόγους ευκολίας. Θα ήταν λάθος να εννοηθεί ότι ο μόνος λόγος χρήσης τους είναι η απλοποίηση των υπολογισμών που προσφέρουν. Χρησιμοποιούμε τις κατανομές αυτές όταν είναι συμβατές με τις πεποιθήσεις μας, όταν περιγράφουν την προηγούμενη γνώση που έχουμε για την παράμετρο. Δημιουργείται όμως το ερώτημα, σε ποιες περιπτώσεις μπορούμε να χρησιμοποιήσουμε ή να εντοπίσουμε μία οικογένεια συζυγών κατανομών; Η μόνη περίπτωση στην οποία οι συζυγείς κατανομές προκύπτουν εύκολα, είναι για τα υποδείγματα που ανήκουν στην εκθετική οικογένεια κατανομών.

Έστω δεδομένα y που ανήκουν στην οικογένεια εκθετικών κατανομών.
Η συνάρτηση πιθανοφάνειας θα έχει τη μορφή :

$$f(y/\vartheta) = \prod_{i=1}^{n}[h(y_i)] \cdot \beta(\vartheta)^n \cdot exp[\sum_{i=1}^{n} T(y_i) \cdot \eta(\vartheta)]$$

$$\propto \beta(\vartheta)^n \cdot exp[\sum_{i=1}^{n} T(y_i) \cdot \eta(\vartheta)]$$

Η εκ των προτέρων κατανομή θα έχει τη μορφή :

$$f(\vartheta) \propto \beta(\vartheta)^d \cdot exp[b \cdot \eta(\vartheta)]$$

Και εφαρμόζοντας το θεώρημα του $Bayes$ έχουμε :

$$f(\vartheta/y) \propto f(\vartheta) \cdot f(y/\vartheta)$$

$$\propto \beta(\vartheta)^d \cdot exp[b \cdot \eta(\vartheta)] \cdot \beta(\vartheta)^n \cdot exp[\sum_{i=1}^{n} T(y_i) \cdot \eta(\vartheta)]$$

$$= \beta(\vartheta)^{n+d} \cdot exp[[b + \sum_{i=1}^{n} T(y_i)] \cdot \eta(\vartheta)]$$

$$= \beta(\vartheta)^{D} \cdot exp[B \cdot \beta(\vartheta)]$$

όπου $D = n + d$ $B = \beta + \sum_{i=1}^{n} T(y_i)$.
Οπότε προκύπτει μία εκ των υστέρων κατανομή η οποία ανήκει στην ίδια οικογένεια κατανομών με την εκ των προτέρων κατανομή, αλλά με προσαρμοσμένες παραμέτρους.

Παράδειγμα

Έστω $y_1, y_2, \cdots, y_n$ ένα τυχαίο δείγμα που ακολουθεί την $Poisson$ κατανομή, με εκ των προτέρων κατανομή μια $Gamma(p, q)$:
$f(\vartheta) = \frac{p^q}{\Gamma(p)} \cdot \vartheta^{p-1} \cdot exp[-q\vartheta]$, $\mu\epsilon$ $p > 0$,$q > 0$,$\vartheta > 0$.
Εφαρμόζουμε το θεώρημα του $Bayes$:

$$f(\vartheta/y) = f(\vartheta) \cdot f(y/\vartheta)$$

$$\propto \vartheta^{p-1} \cdot exp[-q\vartheta] \cdot e^{-n\vartheta} \cdot \vartheta^{\sum_{i=1}^{n} y_i}$$

$$\propto \vartheta^{p-1+\sum_{i=1}^{n} y_i} \cdot exp[-(q+n)\vartheta]$$

$$\equiv Gamma(p + \sum_{i=1}^{n} y_i, q + n)$$

Συνεπώς, η εκ των υστέρων κατανομή του ϑ είναι η Γάμμα με παραμέτρους $p + \sum_{i=1}^{n} y_i$ και $q + n$ που εξαρτώνται από τα δεδομένα. Δηλαδή ανήκει στην ίδια οικογένεια κατανομών με την εκ των προτέρων κατανομή. Η Γάμμα κατανομή είναι η συζυγής εκ των προτέρων κατανομή για το μοντέλο της *Poisson* κατανομής διότι η κατανομή *Poisson* όπως αποδείξαμε και σε προηγούμενο παράδειγμα (παράγραφος 2.3 παράδειγμα 2) ανήκει στην εκθετική οικογένεια κατανομών και η μορφή στην οποία γράφεται είναι:

$$f(x; \vartheta) = \exp[x \ln \vartheta - \vartheta - \ln x!]$$

όπου :

$h(x) = \ln x!$, $T_k(x) = x$, $B(\vartheta) = -\vartheta$, $\eta(\vartheta) = \ln \vartheta$.

οπότε η συζυγής εκ των προτέρων κατανομή της είναι της μορφής

$$f(\vartheta) \propto \beta(\vartheta)^d \cdot exp[b \cdot \eta(\vartheta)]$$

$$\propto exp(-\vartheta d) \cdot exp[b \cdot \ln \vartheta]$$

$$\propto exp(-\vartheta d) \cdot \vartheta^d]$$

Άρα

$$\vartheta \sim Gamma(b + 1, d).$$

4.4.2 Ακατάλληλες εκ των προτέρων κατανομές

Κατά την εφαρμογή των μεθόδων της Μπεϋζιανής συμπερασματολογίας, για την υπό εκτίμηση παράμετρο χρησιμοποιείται η προηγούμενη πληροφορία που έχουμε για την τιμή της, δηλαδή αυτή που έχουμε πριν γίνει η εφαρμογή στα

δεδομένα του παρόντος προβλήματος. Σε αυτό ακριβώς το σημείο έγκειται η βασική διαφορά της Μπεϋζιανής από την κλασική στατιστική, στη χρήση της εκ των προτέρων κατανομής, αφού η παράμετρος θεωρείται τυχαία μεταβλητή. Κάθε πρόβλημα είναι ξεχωριστό και έχει το δικό του περιεχόμενο, από όπου πηγάζουν οι εκ των προτέρων πληροφορίες. Υπάρχουν περιπτώσεις όπου είναι πιθανό να μην έχουμε επαρκή εκ των προτέρων πληροφορία για την παράμετρο, δηλαδή η εκ των προτέρων πεποιθήσεις μας να μην είναι αντικειμενικές ή αξιόπιστες απέναντι στα δεδομένα. Τότε, θέλουμε να επιλέξουμε εκ των προτέρων κατανομές που να επηρεάζουν όσο το δυνατό λιγότερο τις εκ των υστέρων κατανομές. Έτσι δημιουργούμε εκ των προτέρων κατανομές με μικρή ακρίβεια ώστε να διαφυλάξουμε ότι η εκ των προτέρων κατανομή θα έχει πολύ μικρή επίδραση στον σχηματισμό της εκ των υστέρων κατανομής. Αυτές οι κατανομές καλούνται επίπεδες ($flat$) ή μη-πληροφοριακές ($non-inforative$). Αν θεωρήσουμε την ομοιόμορφη κατανομή σε κάποιο διάστημα των πραγματικών αριθμών, τότε αυτή είναι μία κατάλληλη ($proper$) εκ των προτέρων κατανομή. Αν όμως θεωρήσουμε ότι $f(\vartheta) \propto 1, \vartheta \in \Re$, τότε η ομοιόμορφη στο σύνολο των πραγματικών αριθμών δεν είναι κατάλληλη κατανομή ($improper\ prior$).

Γενικά η χρήση ακατάλληλων εκ των προτέρων κατανομών είναι αποδεκτή, εφόσον ελεγχθεί ότι η εκ των υστέρων κατανομή που προκύπτει είναι κατάλληλη.

4.4.3 Μη πληροφοριακή εκ των προτέρων κατανομή του $Jeffrey$

Μία προσέγγιση που χρησιμοποιείται ευρέως στον καθορισμό της εκ των προτέρων άγνοιάς μας και είναι συνεπής σε 1-1 μετασχηματισμούς των παραμέτρων βασίζεται στην πληροφορία του $Fisher$ και εισηγήθηκε από τον $Jeffrey$ ως λύση στο πρόβλημα ότι η ομοιόμορφη δεν αποδίδει μια σταθερή ανάλυση

όταν οι παράμετροι μετασχηματίζονται.

Η εκ των προτέρων κατανομή του $Jeffrey$,$(Jeffrey's\ prior)$ ορίζεται ως:

$$J_{\Theta}(\vartheta) \propto |I(\vartheta)|^{\frac{1}{2}}$$

όπου $I(\vartheta)$ η πληροφορία του $Fisher$ που δίνεται από τον τύπο:

$$I(\vartheta) = -E\left[\frac{d^2 log f(y/\vartheta)}{d\vartheta^2}\right] = E\left[\left(\frac{d log f(y/\vartheta)}{d\vartheta}\right)^2\right].$$

Παραδειγμα 1

Έστω $y_1, \ldots, y_n$ ένα τυχαίο δείγμα από την $Binomial(N_i, \vartheta)$.

Τότε :

$$f(y/\vartheta) = \prod_{i=1}^{n} \left(\frac{N_i}{y_i}\right) \cdot \vartheta^{\sum y_i} \cdot (1-\vartheta)^{N-\sum y_i}$$

όπου :$N = \sum_{i=1}^{n} N_i$

$$log f(y/\vartheta) = \sum y_i log(\vartheta) + (N - \sum y_i) log(1-\vartheta) + C$$

$$\frac{d log f(y/\vartheta)}{d\vartheta} = \frac{\sum y_i}{\vartheta} - \frac{N - \sum y_i}{1-\vartheta}$$

$$\frac{d^2 log f(y/\vartheta)}{d\vartheta^2} = -\frac{\sum y_i}{\vartheta^2} - \frac{N - \sum y_i}{(1-\vartheta)^2}$$

οπότε η πληροφορία του $Fisher$ δίνεται από τον τύπο :

$$I(\vartheta) = -E\left[\frac{d^2 log f(y/\vartheta)}{d\vartheta^2}\right]$$

$$= \frac{E(\sum y_i)}{\vartheta^2} + \frac{N - E(\sum y_i)}{(1-\vartheta)^2}$$

$$= \frac{N\vartheta}{\vartheta^2} + \frac{N - N\vartheta}{(1-\vartheta)^2}$$

$$= N \cdot \left(\frac{1}{\vartheta} + \frac{1}{1-\theta}\right)$$

$$= \frac{N}{\vartheta \cdot (1-\vartheta)}$$

Το οποίο οδηγεί στην εκ των προτέρων κατανομή του $Jeffrey$:

$$J(\vartheta) \propto \vartheta^{-\frac{1}{2}} \cdot (1-\vartheta)^{-\frac{1}{2}}$$

Η οποία στην περίπτωσή μας είναι μια κατάλληλη Βητα(1/2,1/2) κατανομή.

4.5 Πολυπαραμετρικά Προβλήματα

Όλα τα παραδείγματα τα οποία εξετάσαμε μέχρι αυτό το σημείο, είχανε μία μονάχα παράμετρο, τον μέσο ή την διακύμανση του πληθυσμού. Τα περισσότερα στατιστικά προβλήματα περιλαμβάνουν στατιστικά μοντέλα τα οποία περιέχουν περισσότερες από μία άγνωστες παραμέτρους. Μπορεί να υπάρξει η περίπτωση όπου μονάχα μία από τις παραμέτρους να έχει ενδιαφέρον, αλλά συνήθως θα υπάρχουν και άλλες παράμετροι των οποίων οι τιμές θα είναι άγνωστες. Η μέθοδος ανάλυσης των πολυπαραμετρικών προβλημάτων στην Μπεϋζιανή στατιστική είναι πολύ πιο άμεση (τουλάχιστον ως προς τις αρχές της), σε σχέση με αυτήν που χρησιμοποιείται στην κλασική στατιστική. Στην περίπτωση των πολυπαραμετρικών προβλημάτων, έχουμε ένα διάνυσμα $\vartheta = (\vartheta_1, \ldots, \vartheta_d)$από παραμέτρους για το οποίο θέλουμε να εξάγουμε κάποια συμπεράσματα. Καθορίζουμε μία εκ των προτέρων (πολυμεταβλητή) κατανομή $f(\vartheta)$ για το διάνυσμα ϑ, και συνδυαζόμενο με το μοντέλο πιθανοφάνειας $f(y|\vartheta)$ μέσω του θεωρήματος του $Bayes$, υπολογίζεται η εκ των υστέρων κατανομή του ϑ όπως και προηγουμένως :

$$f(\vartheta/y) = \frac{f(\vartheta) \cdot f(y/\vartheta)}{\int f(\vartheta) \cdot f(y/\vartheta) d\vartheta}.$$

Φυσικά, εκ των υστέρων κατανομή θα είναι και αυτή τώρα μία πολυμεταβλητή κατανομή. Ωστόσο, σύμφωνα με τη Μπευζιανή προσέγγιση η συμπερασματολογία για οποιαδήποτε υποομάδα παραμέτρων του διανύσματος ϑ μπορεί να υπολογιστεί άμεσα χρησιμοποιώντας την απο κοινού κατανομή. Για παράδειγμα, η δεσμευμένη εκ των υστέρων κατανομή για το ϑ_i δεδομένων των τιμών όλων των άλλων παραμέτρων, ϑ_{-i}, δίνεται από τον τύπο :

$$f_i(\vartheta_i/y, \vartheta_{-i}) \propto f(\vartheta/y)$$

όπου οι τιμές των ϑ_{-i} θεωρούνται γνωστές.

Παρόλα αυτά, ακριβής μπευζιανή συμπερασματολογία για την παράμετρο ϑ_i μπορεί να γίνει μόνο ολοκληρώνοντας την εκ των υστέρων κατανομή ως προς όλες τις υπόλοιπες παραμέτρους του διανύσματος ϑ_{-i}

$$f(\vartheta_i/y) = \int f(\vartheta/y)d\vartheta_{-i}.$$

Αυτό μας δίνει την περιθώρια εκ των υστέρων κατανομή της παραμέτρου ϑ_i μετά την εξάλειψη των υπόλοιπων παραμέτρων. Αυτή μπορεί να χρησιμοποιηθεί για να εξάγουμε συμπεράσματα για την παράμετρο ϑ_i.

Παρόλο που δεν χρειάζεται καινούργια θεωρία για να γενικευτεί το πρόβλημα στις d-διαστάσεις, μία σειρά από πρακτικά προβλήματα δημιουργούνται:

1. Καθορισμός των εκ των προτέρων κατανομών (prior Specification)
Οι εκ των προτέρων κατανομές τώρα είναι πολυδιάστατες κατανομές. Αυτό σημαίνει ότι ο καθορισμός των εκ των προτέρων κατανομών πρέπει να αντιπροσωπεύει τις εκ των προτέρων πεποιθήσεις μας όχι απλά για κάθε μία παράμετρο ξεχωριστά, αλλά θα πρέπει να αντιπροσωπεύει και τις πεποιθήσεις μας σχετικά με την ανεξαρτησία ανάμεσα σε διαφορετικούς συνδυασμούς παραμέτρων (εάν μία παράμετρος θεωρηθεί μεγάλη, είναι δυνατόν κάποια άλλη παράμετρος να θεωρηθεί μικρή;). Η επιλογή κατάλληλων οικογενειών από εκ των προτέρων κατανομές και ο συνοψισμός των εκ των προτέρων ειδικών πληροφοριών με αυτόν τον τρόπο, είναι αισθητά πιο πολύπλοκο πρόβλημα.

2. Υπολογισμός (computation)

Ακόμα και στα προβλήματα της μίας διάστασης, είδαμε την χρησιμότητα των συζυγών οικογενειών κατανομών στην απλούστευση της εκ των υστέρων ανάλυσης μέσα από το θεώρημα του *Bayes*. Με τα πολυδιάστατα προβλήματα, τα ολοκληρώματα γίνονται ακόμα δυσκολότερα για υπολογισμό. Αυτό κάνει την χρήση των συζυγών εκ των προτέρων κατανομών ακόμα πιο απαραίτητη, και φανερώνει την ανάγκη για υπολογιστικές τεχνικές, με σκοπό να βγάλουμε συμπεράσματα για το πότε η χρήση των συζυγών κατανομών είναι κατάλληλη και πότε είναι ακατάλληλη.

3. Ερμηνεία (interpritation)

Ολόκληρη η εκ των υστέρων συμπερασματολογία, περιλαμβάνεται στην εκ των υστέρων κατανομή, η οποία έχει τόσες διαστάσεις όσες και η παράμετρος ϑ. Η δομή της εκ των υστέρων κατανομής μπορεί να είναι ιδιαίτερα πολύπλοκη, και μπορεί να απαιτεί συγκεκριμένη υποδομή (για παράδειγμα έναν υπολογιστή με δυνατότητες παροχής καλών γραφικών) για να μπορέσει να δοθεί έμφαση στις πιο σημαντικές σχέσεις τις οποίες περιλαμβάνει.

Παρόλα τα πρακτικά προβλήματα, είναι πολύ σημαντικό να τονίσουμε το γεγονός για μία ακόμα φορά, ότι για τα πολυπαραμετρικά προβλήματα χρησιμοποιείται η ίδια θεωρία όπως και για τα προβλήματα μίας διάστασης. Το πλαίσιο της Μπεϋζιανής θεωρίας τονίζει ότι όλα τα συμπεράσματα πηγάζουν από τους βασικούς κανόνες των πιθανοτήτων.

Λόγω των υπολογιστικών δυσκολιών εξαιτίας των ολοκληρωμάτων που παρουσιάζονται στην ανάλυση των πολυπαραμετρικών προβλημάτων στην Μπε-

ϋζιανή στατιστικη, έχουν αναπτυχθεί αλγόριθμοι οι οποίοι μας επιτρέπουν να αντιμετοπίζουμε πολύπλοκα προβλήματα τα οποία ήταν αδύνατον να χειριστούμε. Αυτές οι μέθοδοι είναι κατάλληλες για τον υπολογισμό της εκ των υστέρων κατανομής (μέσω προσομοίωσης) στη Μπεϋζιανή συμπερασματολογία.

4.6 Αλγόριθμοι *Marcov Chain Monte Carlo* (*MCMC*)

4.6.1 Εισαγωγή

Δύο από τα σημαντικότερα προβλήματα της υπολογιστικής στατιστικής είναι η προσομοίωση παρατηρήσεων από κάποια κατανομή f (κατανομή στόχος) και τα ολοκληρώματα που πρέπει να επιλυθούν για τον υπολογισμό της εκ των υστέρων κατανομής τα οποία πολλές φορές παρουσιάζουν μεγάλη δυσκολία. Πιθανόν να μην είναι δυνατή η χρήση συζυγών εκ των προτέρων κατανομών για τον υπολογισμό της από κοινού εκ των υστέρων κατανομής σε κλειστή μορφή, οπότε στις περιπτώσεις αυτές θέλοντας να πάρουμε δείγμα από την εκ των υστέρων κατανομή χρησιμοποιούνται ασυμπτωτικές προσεγγίσεις για τον υπολογισμό των εκ των υστέρων πιθανοτήτων.

Όσο περισσότερο αυξάνεται η πολυπλοκότητα των παραπάνω προβλημάτων ή/και η διάσταση της παραμέτρου, τόσο η αντιμετώπιση τους με «άμεσες» τεχνικές γίνεται όλο και πιο δύσκολη, αν όχι αδύνατη. Για αυτό το λόγω μία πληθώρα τεχνικών *Monte Carlo* (*MC*) και *Markov Chain Monte Carlo* (*MCMC*) έχουν προταθεί στη βιβλιογραφία. Οι τεχνικές *MC* παράγουν ανεξάρτητες παρατηρήσεις είτε απευθείας από την εκ των υστέρων κατανομή, είτε από κάποια

διαφορετική κατανομή (κατανομή πρότασης). Μία από τις σημαντικότερες μεθόδους MC είναι η μέθοδος της Αποδοχής Απόρριψης. Οι τεχνικές ($MCMC$) προσομοιώνουν τιμές τυχαίων μεταβλητών με χρήση ηλεκτρονικού υπολογιστή από την εκ των υστέρων κατανομή. Για κάθε παρατήρηση κατασκευάζεται μια αλυσίδα $Markov$. Οι ιδιότητες των τιμών της μαρκοβιανής αλυσίδας δίνει τη δυνατότητα στην επόμενη τιμή κάθε παρατήρησης να εξαρτάται από την παρούσα τιμή, όχι όμως από την προηγούμενη. Το πλεονέκτημα αυτής της μεθόδου, είναι ότι όταν ο αλγόριθμος της προσομοίωσης επαναλαμβάνεται πολλές φορές, η προσέγγιση της εκ των υστέρων κατανομής βελτιώνεται σε κάθε βήμα. Έτσι, δίνεται η ικανότητα στους Μπεϋζιανούς να εκτιμούν με ακρίβεια τις εκ των υστέρων κατανομές. Οι πιο δημοφιλείς μέθοδοι $MCMC$, είναι ο αλγόριθμος $Metropolis-Hastings$, ο αλγόριθμος $random-walk\ Metropolis-Hastings$, ο αλγόριθμος *Gibbs sampler* και ο αλγόριθμος προσαύξησης δεδομένων (*Data Augmentation*).

4.6.2 Γενικός αλγόριθμος $MCMC$

Πριν αναλύσουμε τους πιο δημοφιλείς από τους $MCMC$ αλγόριθμους θα μιλήσουμε για το γενικό πλαίσιο στο οποίο κινούνται.
Τα βήματα ενός γενικού αγλόρυθμου $MCMC$ είναι :

1. Χωρίζουμε τις παραμέτρους σε d ομάδες $\vartheta_1, \ldots, \vartheta_d$ καθένα από τα οποία είναι διάστασης ≥ 1.
2. Δίνουμε αρχικές τιμές στο διάνυσμα των παραμέτρων $\vartheta^{(0)} = (\vartheta_1^0, \ldots, \vartheta_d^0)$.
3. Ανανεώνουμε την τιμή ϑ_1^0 σε ϑ_1^1 σύμφωνα με την δεσμευμένη εκ των υ-

στέρων κατανομή της παραμέτρου ϑ_1 :

$$f(\vartheta_1/y, \vartheta_2^0, \ldots, \vartheta_d^0).$$

$\vdots$

4. Ανανεώνουμε την τιμή ϑ_i^0 σε ϑ_i^1 σύμφωνα με την δεσμευμένη εκ των υστέρων κατανομή της παραμέτρου ϑ_i :

$$f(\vartheta_i/y, \vartheta_1^1, \ldots, \vartheta_{i-1}^1, \vartheta_{i+1}^0, \ldots, \vartheta_{d-1}^0).$$

$\vdots$

5. Ανανεώνουμε την τιμή ϑ_d^0 σε ϑ_d^1 σύμφωνα με την δεσμευμένη εκ των υστέρων κατανομή της παραμέτρου ϑ_d :

$$f(\vartheta_d/y, \vartheta_1^1, \ldots, \vartheta_{d-1}^1).$$

6. Επαναλαμβάνουμε τη διαδικασία εως ότου να επιτευχθεί σύγκλιση.

4.6.3 Αλγόριθμος *Metropolis − Hastings*

Η ονομασία αυτού του αλγορίθμου προέρχεται από τη δημοσίευση των *Metropolis et al.* (1953) και *Hastings* (1970). Αυτές θεωρούνται ως οι βασικές αναφορές για την ανάπτυξη αυτού του αλγορύθμου, συμπεριλαμβανομένων εκείνων των *Metropolis−Ulam* όπου το 1949 εισήγαγαν την φιλοσοφία *Monte Carlo*. Επιπρόσθετα σημαντική ήταν η συμβολή των *Barker* (1965) που εισήγαγε μια εναλλακτική μέθοδο αποδοχής απόρριψης και *Peskun* (1973) που εισήγαγε μια *partial ordering* έτσι ώστε να γίνει πιο κατανοητή η σημασία της πρότασης του *Barker* [*Monte Carlo Strategies in Scientific Computing*].

Η πρωτότυπη δημοσίευση από τους *Metropolis et al.* (*Equation of State Calculations by Fast Computing Machines*) (1953) διατύπωσε αρχικά τον

Metropolis algorithm ο οποίος ασχολείται με τον υπολογισμό ατομικών και μοριακών συστημάτων και δημοσιεύτηκε στο επιστημονικό περιοδικό *Journal of Chemical Physics*. Έπειτα, ο *Hastings*(1970) γενίκευσε τον προυπάρχον αλγόριθμο και δημιούργησε τον $Metropolis - Hastings\ algorithm$ ο οποίος θεωρείται και ως η γενίκευση όλων των $(MCMC)$ μεθόδων. Η τελική γενίκευση του $Metropolis - Hastings\ algorithm$ διατυπώθηκε από τον *Green* και δημοσιεύτηκε με τίτλο *Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination* στο περιοδικό *Biometrika* (1995). Η τελευταία αυτή γενίκευση επιτρέπει να πέρνουμε δείγματα από παραμετρικούς χώρους με διαφορετικές διαστάσεις.

Ας υποθέσουμε ότι έχουμε την εκ των υστέρων κατανομή $f(\vartheta/y)$ από την οποία επιθυμούμε να να παράγουμε ένα δείγμα μεγέθους N. Έστω $\vartheta^{(j)}$ το διάνυσμα των παραγόμενων παραμέτρων της j επανάληψης του αλγορίθμου.
Ο αλγόριθμος $Metropolis-Hastings$ μπορεί να περιγραφεί με τα εξής βήματα :

1. Δίνουμε αρχικές τιμές στο διάνυσμα των παραμέτρων $\vartheta^{(0)}$.
2. Έστω ότι βρισκόμαστε στη $j - οστη$ επανάληψη $\vartheta_1^{(j)}, \ldots, \vartheta_d^{(j)}$, για $j = 1, \ldots, d$. Στόχος μας είναι να ανανεώσουμε το $\vartheta_1^{(j)}$ σε $\vartheta_1^{(j+1)}$. Για το σκοπό αυτό προτείνουμε μία υποψήφια τιμή $\vartheta_1^{(can)}$ από μία τυχαία επιλεγμένη κατανομή με συνάρτηση πυκνότητας πιθανότητας $q(\vartheta_1^{(can)}/\vartheta_1^{(j)}, \cdots, \vartheta_d^{(j)})$.
3. Αποδεχόμαστε ως νέα τιμή για την παράμετρο ϑ_1 την :

$$\vartheta_1^{(j+1)} = \begin{cases} \vartheta_1^{(can)}, με\ πιθανοτητα\ p \\ \vartheta_1^{(j)}, με\ πιθανοτητα\ 1-p \end{cases}.$$

Όπου :

$$p = min\left\{1, \frac{f(\vartheta_1^{(can)}/y, \vartheta_2^{(j)}, \ldots, \vartheta_d^{(j)}}{f(\vartheta_1^{(j)}/y, \vartheta_2^{(j)}, \ldots, \vartheta_d^{(j)}} \cdot \frac{q(\vartheta_1^{(j)}/y, \vartheta_1^{(can)}, \ldots, \vartheta_d^{(j)}}{q(\vartheta_1^{(can)}/\vartheta_1^{(j)}, \vartheta_2^{(j)}, \ldots, \vartheta_d^{(j)}}\right\}$$

- Στην πραγματικότητα ο τρόπος με τον οποίο διεξάγεται το βήμα της αποδοχής της τιμής $\vartheta_1^{(can)}$ είναι :
  Προσομοιώνουμε μία τιμή $u \sim U(0,1)$.
  Εάν $u \prec p$ τότε θέτουμε

$$\vartheta_1^{(j+1)} = \vartheta_1^{(can)}$$

  Αλλιώς θέτουμε

$$\vartheta_1^{(j+1)} = \vartheta_1^{(j)}$$

  .
- Επαναλαμβάνουμε εως ότου να επιτευχθεί σύγκλιση.

4.6.4 Αλγόριθμος *random−walk Metropolis−Hastings* με Κανονικές Προσαυξήσεις

Μία επιλογή για τη γεννήτρια τυχαίων υποψήφιων τιμών $\vartheta_1^{(can)}$ είναι να επιλέξουμε η κατανομή με συνάρτηση πυκνότητας πιθανότητας $q(\vartheta_1^{(can)}/\vartheta_1^{(j)}, \vartheta_2^{(j)}, \ldots, \vartheta_d^{(j)})$ να είναι η Κανονική κατανομή για το $\vartheta_1^{(can)}$ με μέση τιμή $\vartheta_1^{(j)}$ και μια κατάλληλα επιλεγμένη διασπορά u. Τη τιμή της διασποράς u την επιλέγουμε έτσι ώστε ο ρυθμός αποδοχής της $\vartheta_1^{(j)}$ να είναι περίπου στο διάστημα (0.2 ,0.5). Εάν πάρουμε u πολύ μεγάλο τότε ο αλγόριθμος με μεγάλη πιθανότητα θα προτείνει

τιμές μακριά από το $\vartheta_1^{(j)}$ οπότε θα απορρίπτεται συχνά διότι θα έχω μικρή πιθανότητα αποδοχής. Εάν πάρουμε μία τιμή για το u πολύ μικρή, τότε με μεγάλη πιθανότητα θα προτείνονται τιμές πολύ κοντά στη τρέχουσα $\vartheta_1^{(j)}$. Άρα σε αυτή τη περίπτωση ο αλγόρυθμος θα δέχεται πολύ συχνά.

Διαλέγουμε λοιπόν κατανομή q έτσι ώστε :

$$q(\vartheta_1^{(can)}/\vartheta_1^{(j)}, \vartheta_2^{(j)}, \ldots, \vartheta_d^{(j)}) = q(\vartheta_1^{(can)}/\vartheta_1^{(j)}) \equiv N(\vartheta_1^{(j)}, u)$$

Σε αυτή τη περίπτωση η πιθανότητα αποδοχής ισούτε με το λόγο των δεσμευμένων εκ των υστέρων κατανομών :

$$p = min\left\{1, \frac{f(\vartheta_1^{(can)}/y, \vartheta_2^{(j)}, \ldots, \vartheta_d^{(j)}}{f(\vartheta_1^{(j)}/y, \vartheta_2^{(j)}, \ldots, \vartheta_d^{(j)}} \cdot \frac{q(\vartheta_1^{(j)}/\vartheta_1^{(can)}}{q(\vartheta_1^{(can)}/\vartheta_1^{(j)}}\right\} = min\left\{1, \frac{f(\vartheta_1^{(can)}/y, \vartheta_2^{(j)}, \ldots, \vartheta_d^{(j)}}{f(\vartheta_1^{(j)}/y, \vartheta_2^{(j)}, \ldots, \vartheta_d^{(j)}}\right\}$$

Διότι :

$$q(\vartheta_1^{(j)}/\vartheta_1^{(can)}) = \frac{1}{\sqrt{2}} \cdot exp\left\{-\frac{1(\vartheta_1^{(can)} - \vartheta_1^{(j)})^2}{2u}\right\}$$

και

$$q(\vartheta_1^{(can)}/\vartheta_1^{(j)}) = \frac{1}{\sqrt{2}} \cdot exp\left\{-\frac{1(\vartheta_1^{(j)} - \vartheta_1^{(can)})^2}{2u}\right\}$$

οπότε παρατηρούμε ότι η q είναι συμμετρική κατανομή και ισχύει :

$$q(\vartheta_1^{(j)}/\vartheta_1^{(can)}) = q(\vartheta_1^{(can)}/\vartheta_1^{(j)}).$$

Οπότε ο αλγόριθμος $random-walk\ Metropolis-Hastings$ με κανονικές προσαυξήσεις μπορεί να περιγραφεί με τα εξής βήματα :

1. Δίνουμε αρχικές τιμές στο διάνυσμα των παραμέτρων $\vartheta^{(0)}$.

2. Έστω ότι βρισκόμαστε στη $j-οστη$ επανάληψη του αλγορίθμου όπου $\vartheta_1^{(j)}, \ldots, \vartheta_d^{(j)}$, για $j = 1, \ldots, d$. Στόχος μας είναι να ανανεώσουμε το $\vartheta_1^{(j)}$ σε $\vartheta_1^{(j+1)}$. Για το σκοπό αυτό προτείνουμε μία υποψήφια τιμή $\vartheta_1^{(can)}$ από μία επιλεγμένη κατανομή με συνάρτηση πυκνότητας πιθανότητας $N(\vartheta_1^{(j)}, u)$.

3. Αποδεχόμαστε ως νέα τιμή για την παράμετρο ϑ_1 την :

$$\vartheta_1^{(j+1)} = \begin{cases} \vartheta_1^{(can)}, με\ πιθανότητα\ p \\ \vartheta_1^{(j)}, με\ πιθανότητα\ 1-p \end{cases}$$

Όπου :

$$p = min\left\{1, \frac{f(\vartheta_1^{(can)}/y, \vartheta_2^{(j)}, \ldots, \vartheta_d^{(j)}}{f(\vartheta_1^{(j)}/y, \vartheta_2^{(j)}, \ldots, \vartheta_d^{(j)}} \cdot \frac{q(\vartheta_1^{(j)}/\vartheta_1^{(can)}}{q(\vartheta_1^{(can)}/\vartheta_1^{(j)}}\right\} = min\left\{1, \frac{f(\vartheta_1^{(can)}/y, \vartheta_2^{(j)}, \ldots, \vartheta_d^{(j)}}{f(\vartheta_1^{(j)}/y, \vartheta_2^{(j)}, \ldots, \vartheta_d^{(j)}}\right\}$$

- Στην πραγματικότητα ο τρόπος με τον οποίο διεξάγεται το βήμα της αποδοχής της τιμής $\vartheta_1^{(can)}$ είναι :
  Προσομοιώνουμε μία τιμή $u \sim U(0,1)$.
  Εάν $u \prec p$ τότε θέτουμε

$$\vartheta_1^{(j+1)} = \vartheta_1^{(can)}.$$

  Αλλιώς θέτουμε

$$\vartheta_1^{(j+1)} = \vartheta_1^{(j)}.$$

- Επαναλαμβάνουμε εως ότου να επιτευχθεί σύγκλιση.

4.6.5 Ο Δειγματολήπτης *Gibbs*

Ο δειγματολήπτης *Gibbs* είναι μία ειδική περίπτωση του *Markov Chain Monte Carlo algorithm*. Πήρε το όνομά του από τον *J.W.Gibbs* ο οποίος βρήκε την αναλογία μεταξύ ενός δειγματοληπτικού αλγορίθμου και της στατιστικής φυσικής. Ο αλγόριθμος περιγράφτηκε λίγο αργότερα και από τους αδερφούς *Stuart* και *Donald Geman* (1984). Στη γενική του μορφή ο δειγματολήπτης

Gibbs είναι ειδική περίπτωση του αλγόριθμου $Metropolis-Hastings$. Ο δειγματολήπτης *Gibbs* εφαρμόζεται όταν η από κοινού κατανομή είναι άγνωστη ή όταν είναι ανέφικτο να πάρουμε δείγμα από αυτή κατευθείαν, αλλά οι δεσμευμένες εκ των υστέρων κατανομές είναι γνωστές για κάθε μεταβλητή κι άρα είναι εύκολο να πάρουμε δείγμα από αυτές. Ο δειγματολήτης *Gibbs* καταφέρνει να προσομοιώσει δείγμα απο αυτή την εκ των υστέρων κατανομή προσομοιώνοντας ακολουθιακά και επαναληπτικά από τις δεσμευμένες κατανομές των επί μέρους παραμέτρων.

Έστω ότι θέλουμε να λάβουμε δείγμα από την πολυπαραμετρική εκ των υστέρων κατανομή $f(\vartheta_1, \ldots, \vartheta_d/y)$ (κατανομή στόχος).

Τα βήματα του αλγορίθμου είναι τα ακόλουθα:

1. Δίνουμε αρχικές τιμές στο διάνυσμα της παραμέτρου $\vartheta^0 = (\vartheta_1^0, \ldots, \vartheta_d^0)$.
2. Έστω ότι βρισκόμαστε στη $j-οστη$ επανάληψη του αλγορίθμου όπου $\vartheta^{j-1} = (\vartheta_1^{(j-1)}, \ldots, \vartheta_d^{(j-1)})$, για $j = 1, \ldots, d$.
   - Προσομοιώσουμε τη νέα τιμή $\vartheta_1^{(j)}$, για την πρώτη παράμετρο από τη δεσμευμένη εκ των υστέρων κατανομή της παραμέτρου ϑ_1 από την :
   $$f(\vartheta_1/y, \vartheta_2^{(j-1)}, \ldots, \vartheta_d^{(j-1)}).$$
   - Προσομοιώσουμε τη νέα τιμή $\vartheta_2^{(j)}$, για την δεύτερη παράμετρο από τη δεσμευμένη εκ των υστέρων κατανομή της παραμέτρου ϑ_2 από την :
   $$f(\vartheta_2/y, \vartheta_1^{(j)}, \vartheta_3^{(j-1)}, \ldots, \vartheta_d^{(j-1)}).$$
   - Προσομοιώσουμε τη νέα τιμή $\vartheta_d^{(j)}$, για την d-οστή παράμετρο από τη δεσμευμένη εκ των υστέρων κατανομή της παραμέτρου ϑ_d από την :
   $$f(\vartheta_d/y, \vartheta_1^{(j)}, \vartheta_2^{(j)}, \ldots, \vartheta_{d-1}^{(j)}).$$
3. Επαναλαμβάνουμε τη διαδικασία μέχρι να επιτευχθεί σύγκλιση.

Κάτω από συνθήκες ομαλότητας αποδεικνύεται ότι η Μαρκοβιανή αλυσίδα συγκλίνει στη στάσιμη κατανομή $f(y/\vartheta)$. Έτσι μετά από μία περίοδος $burn-in$(η οποία αποτελείται από έναν αριθμό επαναλήψεων όπου οι τιμές που λαμβάνουμε για τις παραμέτρους απορρίπτονται), τα δείγματα που παράγονται από τον αλγόριθμο *Gibbs* μπορούν να θεωρηθούν ως δείγματα από την από κοινού εκ των υστέρων κατανομή των παραμέτρων $f(y/\vartheta)$.

Παρατηρούμε ότι ο αλγόριθμος *Gibbs* μπορεί να χρησιμοποιηθεί μόνο σε περιπτώσεις όπου οι δεσμευμένες εκ των υστέρων κατανομές είναι γνωστές ή ευκολα προσομοιώσιμες. Εκεί έγκυται και η διαφορά του από τον αλγόριθμο $Metropolis-Hastings$ ο οποίος μας επιτρέπει να λαμβάνουμε δείγμα από εκ των υστέρων κατανομές οι οποίες δεν είναι γνωστες. Υπάρχει όμως και η περίπτωση κάποιες από τις δεσμευμένες κατανομές να είναι γνωστές και κάποιες όχι. Σε αυτή τη περίπτωση χρησιμοποιούμε ένα μικτό αλγόριθμο ο οποίος προσομοιώνει δείγμα από τις μεν γνωστές δεσμευμένες κατανομές με βάσει τον δειγματολήπτη *Gibbs* και από τις δε εκ των υστέρων κατανομές που δεν είναι σε κλειστή μορφή με βάση τον αλγόριθμο $Metropolis-Hastings$.

4.6.6 Αλγόριθμος Αύξησης Δεδομένων

Ο αλγόριθμος αύξησης δεδομένων (*Data Augmentation*) πήρε το όνομά του από τους *Taner* και *Wong* (1987) οι οποίοι τον χρησιμοποίησαν για να περιγράψουν έναν επαναληπτικό αλγόριθμο για την προσέγγιση της εκ των υστέρων κατανομής.

Η χρήση της μεθόδου Αύξηση δεδομένων (*Data Augmentation*) είναι μία τεχνική η οποία επιτρέπει τη χρήση του αλγορίθμου *Gibbs* σε κάποιες κατηγορίες προβλημάτων που εξάρχής δεν είναι εφικτό. Αυτό γίνεται με την εισαγωγή κάποιων νέων τυχαίων μεταβητών (Z) στο πρόβλημα. Δεσμεύοντας ως προς αυτές τις νέες μεταβλητές γίνεται εύκολη στο χειρισμό η συνάρτηση πιθανοφά-

νειας οπότε καταλήγουμε σε μία εκ των υστέρων κατανομή σε κλειστή μορφή.

Έστω ότι θέλουμε να πάρουμε δείγμα από την εκ των υστέρων κατανομή $f(\vartheta|Y)$. Για να μπορέσουμε να πάρουμε δείγμα από τη κατανομή αυτή, σε περιπτώσεις όπου η συνάρτηση $f(Y|\vartheta)$ είναι δύσκολη στο χειρισμό εισάγουμε νέες τυχαίες μεταβλητές Z, έτσι ώστε να δημιουργήσουμε την $f(Y, Z|\vartheta)$ την οποία είναι εύκολο να χειριστούμε. Έτσι προκύπτει μία εκ των υστέρων κατανομή της μορφής $f(\vartheta, Z|Y) \propto f(Y, Z|\vartheta) \cdot f(\vartheta)$ από την οποία καταλήγουμε στη ζητούμενη δεσμευμένη εκ των υστέρων κατανομή $f(\vartheta|Y, Z)$.

Γενικά αυτός ο αλγόρυθμος χρησιμοποιείται :

- Σε προβλήματα έλλειψης δεδομένων (*missing data problems*).

  Σε αυτή τη περίπτωση κατανομή των παρατηρούμενων δεδομένων y πέρνει τη μορφή :

  $$f(y/\vartheta) = \int f(y, z/\vartheta)dz$$

  όπου z τα δεδομένα που λείπουν. και $f(y, z/\vartheta)$ είναι η από κοινού κατανομή των y, z.

- Σε περιπτώσεις όπου η συνάρτηση πιθανοφάνειας είναι δύσκολη στο χειρισμό αλλά με την εισαγωγή νέων μεταβλητών το αρχικό πρόβλημα λύνεται.

Κεφάλαιο 5

Μπεϋζιανή Εκτίμηση για Γενικευμένα Γραμμικά Μοντέλα

Στο κεφάλαιο αυτό παρουσιάζεται η Μπεϋζιανή μέθοδος εκτίμησης των γενικευμένων γραμμικών μοντέλων χρησιμοποιώντας αλγόριθμους $MCMC$. Για την καλύτερη παρουσίαση των μεθόδωνχρησιμοποιούνται παραδείγματα με πραγματικά σετ δεδομένων.

5.1 Μπεϋζιανή Εκτίμηση για Μοντέλα Λογιστικής Παλινδρόμησης

Θα περιγράψουμε την διαδικασία κατασκευής ενός μοντέλου λογιστικής παλινδρόμησης χρησιμοποιώντας δεδομένα σχετικά με τις λοιμόξεις που προκύπτουν από την γέννηση με καισσαρική τομή, *Caesarian data* (*Fahrmeir and Tutz*,2001).

Τρείς παράγοντες επιρροής έχουν μελετηθεί:

- ένας δείκτης που μας πληροφορεί για το εάν η καισσαρική τομή ήταν προγραμματισμένη ή όχι ($x_1 = 1$: όχι, $x_1 = 0$: ναι), το οποίο στο πίνακα αποτελεσμάτων αναφέρεται ως *noplan* (x_1),
- ένας δείκτης του κατά πόσον πρόσθετοι παράγοντες κινδύνου ήταν παρόντες κατά τη στιγμή της γέννησης, π.χ.υπέρβαρη ή διαβητική μητέρα ($x_2 = 1$: ναι, $x_2 = 0$: ναι), το οποίο στο πίνακα αποτελεσμάτων αναφέρεται ως $factor$ (x_2),
- μία ένδειξη για το αν δόθηκαν αντιβιωτικά για προληπτικούς λόγους ($x_3 = 1$: ναι, $x_3 = 0$: όχι), το οποίο στο πίνακα αποτελεσμάτων αναφέρεται ως *antib* (x_3).

Η μεταβλητή αποκρίσεως y_i είναι ο αριθμός των μολύνσεων που παρατηρήθηκαν μεταξύ των n_i ασθενών που έχουν τους ίδιους παράγοντες κινδύνου και p_i είναι η πιθανότητα λοίμωξης για κάθε έγκυο με αυτούς τους παράγοντες κινδύνου.

Τα δεδομένα δίνονται στον πίνακα 5.1 και μπορούν να μοντελοποιηθούν υποθέτοντας ότι:

$$y_i \sim Binomial\,(n_i, p_i)$$

5.1.1 Ανάλυση με Βάση τη Κλασική Στατιστική

Τα αποτελέσματα ακολουθόντας την κλασική στατιστική συνοψίζονται στον πίνακα 5.2. Παρατηρώντας τα αποτελέσματα του πίνακα με βάση τη κλασική στατιστική (πίνακας 5.2) παρατηρούμε ότι όλες οι μεταβλητές είναι στατιστικά σημαντικές. Επίσης παρατηρούμε ότι ενώ οι παράγοντες *noplan* και $factor$ επιδρούν θετικά και άρα υπάρχει θετική συσχέτιση με την εμφάνιση λοίμωξης, ο παράγοντας *antib* επιδρά αρνητικά και άρα υπάρχει αρνητική συσχέτιση με

Πίνακας 5.1: *Caesarian data*

i	$y_i(yes)$	no	x_1	x_2	x_3
1	8	32	0	0	0
2	0	2	0	0	1
3	28	30	0	1	0
4	1	17	0	1	1
5	0	9	1	0	0
6	0	0	1	0	1
7	23	3	1	1	0
8	11	87	1	1	1

την εμφάνιση λοίμωξης κάτι το οποίο ήταν αναμενόμενο δεδομενου ότι στην έγκυο έχει χορηγηθεί προληπτικά αντιβίωση. Προσπαθώντας να ερμηνεύσουμε για παράδειγμα το κατά πόσο ο παράγοντας κινδύνου *noplan*, δηλαδή η ύπαρξη προγραμματισμένης καισσαρικής ή όχι, επιδρά στην εμφάνιση λοίμωξης από τη γέννηση με καισσαρική τομή και κρατώντας σταθερούς τους δύο παράγοντες κινδύνου $factor$ και $antib$, παρατηρούμε ότι τα $odds$ εμφάνισης λοίμωξης αυξάνουν για την περίπτωση προγραμματισμένης καισσαρικής κατα $exp(1.0720)$ ή τα $log(odds)$ αυξάνουν για την ίδια περίπτωση κατά 1.0720. Αντίστοιχα ερμηνεύονται και οι άλλοι δύο συντελεστές.

5.1.2 Ο Αλγόριθμος Επαναλαμβανόμενων Σταθμισμένων Ελαχίστων Τετραγώνων του *Gamerman*

Ο αλγόριθμος *Gamerman*, ο οποίος πήρε το όνομα του από τον εφευρέτη του *Dani Gamerman*, είναι μια ειδική περίπτωση των αλγορίθμων *Metropolis* και $Metropolis - Hastings$ στην οποία η γεννήτρια τυχαίων υποψήφιων τιμών

Πίνακας 5.2: Αποτελέσματα εφαρμόζοντας τη κλασική στατιστική

$coefficients$	$estimate\ std$	$error$	$z\ value$	$Pr(>\|z\|)$
$(intercept)$	-1.8926	0.4124	-4.589	4.45e-06
$noplan$	1.0720	0.4254	2.520	0.0117
$factor$	2.0299	0.4553	4.459	8.25e-06
$antib$	-3.2544	0.4813	-6.761	1.37e-11

προέρχεται από μία επαναληπτική μέθοδο σταθμισμένων εκτιμητών ελαχίστων τετραγώνων (*iterated weighted least squares* - $IWLS$). Για κάθε επανάληψη χρησιμοποιείται για τις παρατηρήσεις ένα σύνολο από βάρη. Τα βάρη κατασκευάζονται με την εφαρμογή μιας συνάρτησης βάρους στα τρέχοντα κατάλοιπα. Η γεννήτρια τυχαίων υποψήφιων τιμών χρησιμοποιεί τις τιμές των παραμέτρων της τρέχουσας επανάληψης, που αποτελούν τη κατανομή από την οποία θα δημιουργήσει μια νέα τυχαία προτεινόμενη τιμή.

Η μέθοδος αυτή μπορεί να χρησιμοποιηθεί για την απόκτηση τυχαίων δειγμάτων από τον εκ των υστέρων κατανομή των παραμέτρων παλινδρόμησης σε ένα γενικευμένο γραμμικό μοντέλο (GLM).

Για τα διωνυμικά δεδομένα y_i έχουμε:

$$f\left(y_i/\beta\right) = \begin{pmatrix} n_i \\ y_i \end{pmatrix} p_i^{y_i} \left(1-p_i\right)^{(n_i - y_i)}$$

$$f\left(y/\beta\right) \propto \prod_{i=1}^{n} \frac{\left(exp\left(x_i^T\beta\right)\right)^{y_i}}{\left[1+exp\left(x_i^T\beta\right)\right]^{y_i}} \cdot \frac{1^{n_i-y_i}}{\left[1+exp\left(x_i^T\beta\right)\right]^{n_i-y_i}} = \prod_{i=1}^{n} \frac{exp\left(x_i^T\beta y_i\right)}{\left[1+exp\left(x_i^T\beta\right)\right]^{n_i}}$$

Χρησιμοποιώντας το μοντέλο της λογιστικής παλινδρόμησης για τα παραπάνω δεδομένα:

$$logit\left(p_i\right) = x_i^T\beta$$

$$\Rightarrow log\left(\frac{p_i}{1-p_i}\right) = x_i^T\beta$$

$$\Rightarrow \frac{p_i}{1-pi} = exp\left(x_i^T\beta\right)$$

$$\Rightarrow p_i = exp\left(x_i^T\beta\right) - p_i \cdot exp\left(x_i^T\beta\right)$$

$$\Rightarrow p_i \cdot \left(1 + exp\left(x_i^T\beta\right)\right) = exp\left(x_i^T\beta\right)$$

$$\Rightarrow p_i = \frac{exp\left(x_i^T\beta\right)}{1 + exp\left(x_i^T\beta\right)}$$

Οπότε η εκ των υστέρων κατανομή για τα β είναι της μορφής:

$$f\left(\beta/y\right) \propto f\left(y/\beta\right) \cdot f\left(\beta\right) \propto \left\{\prod_{i=1}^{n} \frac{exp\left(x_i^T\beta y_i\right)}{\left[1 + exp\left(x_i^T\beta\right)\right]^{n_i}}\right\} \cdot \exp\left(-\frac{\left(\beta - \mu_0\right)^T \cdot C_0^{-1} \cdot \left(\beta - \mu_0\right)}{2}\right)$$

Για τη λογιστική παλινδρόμηση με διωνυμικά δεδομένα, η κανονική συνδετική συνάρτηση είναι της μορφής:

$$g\left(\mu_i\right) = log\left(\frac{p_i}{1-p_i}\right) = log\left(\frac{n_i \cdot p_i}{n_i - n_i \cdot p_i}\right) = log\left(\frac{\mu_i}{n_i - \mu_i}\right)$$

Και χρησιμοποιώντας τον εξής μετασχηματισμό των δεδομένων:

$$\widetilde{y}_i\left(\beta\right) = h_i + \left(y_i - \mu_i\right) g'\left(\mu_i\right)$$

Με βάρη:

$$W_i(\beta) = 1/g'\left(\mu_i\right)$$

Δεδομένου του ότι το γενικευμένο γραμμικό μοντέλο με βάρη είναι της μορφής:

$$\widetilde{y}\left(\beta^{(t-1)}\right) \sim N\left(X\beta, W^{-1}\left(\beta^{(t-1)}\right)\right)$$

και ο γενικός τύπος του GLS εκτιμητή του διανύσματος β γράφεται ως:

$$\beta^{(t)} = \left(X^T W\left(\beta^{(t-1)}\right) X\right)^{-1} X^T W\left(\beta^{(t-1)}\right) \widetilde{y}\left(\beta^{(t-1)}\right)$$

χρησιμοποιώντας τη στατιστική κατά $Bayes$ και παίρνοντας μια εκ των προτέρων κανονική κατανομή για τις παραμέτρους του μοντέλου μας έχουμε:

$$\widetilde{y}\left(\beta^{(t-1)}\right) \sim N\left(X\beta, W^{-1}\left(\beta^{(t-1)}\right)\right)$$

$$\beta \sim N\left(\mu_0, C_0\right)$$

οπότε πιο αναλυτικά οι παραπάνω κατανομές γράφονται στη μορφή:

$$f\left(\beta\right) = \frac{1}{\sqrt{2\pi}} \cdot \left|C_0\right|^{-1/2} \cdot exp\left\{-\frac{1}{2} \cdot \left(\beta - \mu_0\right)^T \cdot \left(C_0\right)^{-1} \cdot \left(\beta - \mu_0\right)\right\}$$

$$f\left(\widetilde{y}\left(\beta^{(t-1)}\right)\right) = \frac{W\left(\beta^{(t-1)}\right)^{n/2}}{\left(\sqrt{2\pi}\right)^n} \cdot exp\left\{-\frac{W\left(\beta^{(t-1)}\right)}{2} \cdot \sum\left(\widetilde{y}_i\left(\beta^{(t-1)}\right) - x_i^T\beta\right)^2\right\}$$

όπου επειδή έχουμε πίνακες:

$$\sum\left(\widetilde{y}_i\left(\beta^{(t-1)}\right) - x_i^T\beta\right)^2 = \left(\widetilde{y}\left(\beta^{(t-1)}\right) - X\beta\right)^T \cdot \left(\widetilde{y}\left(\beta^{(t-1)}\right) - X\beta\right)$$

οπότε έχουμε:

$$f\left(\widetilde{y}\left(\beta^{(t-1)}\right)\right) = \frac{W\left(\beta^{(t-1)}\right)^{n/2}}{\left(\sqrt{2\pi}\right)^n} \cdot exp\left\{-\frac{1}{2} \cdot \left(\widetilde{y}\left(\beta^{(t-1)}\right) - X\beta\right)^T \cdot W\left(\beta^{(t-1)}\right) \cdot \left(\widetilde{y}\left(\beta^{(t-1)}\right) - X\beta\right)\right\}$$

άρα

$$f\left(\beta/\widetilde{y}\left(\beta^{(t-1)}\right)\right) \propto \left(\widetilde{y}\left(\beta^{(t-1)}\right)/\beta\right) \cdot f\left(\beta\right) \propto$$

$$exp\left\{-\frac{1}{2} \cdot \left(\widetilde{y}\left(\beta^{(t-1)}\right) - X\beta\right)^T \cdot W\left(\beta^{(t-1)}\right) \cdot \left(\widetilde{y} - X\beta\right)\right\} \cdot exp\left\{-\frac{1}{2} \cdot \left(\beta - \mu_0\right)^T \cdot \left(C_0\right)^{-1} \cdot \left(\beta - \mu_0\right)\right\}$$

$$= exp\left\{-\frac{1}{2} \cdot \left[\beta^T \cdot \left(\left(C_0\right)^{-1} + X^T \cdot W\left(\beta^{(t-1)}\right) \cdot X\right) \cdot \beta - \right.\right.$$

$$- \beta^T \cdot \left(\left(C_0\right)^{-1} \cdot \mu_0 + X^T \cdot W\left(\beta^{(t-1)}\right) \cdot \widetilde{y}\left(\beta^{(t-1)}\right)\right) -$$

$$\left.\left. - \left(\mu_0^T \cdot \left(C_0\right)^{-1} + \widetilde{y}\left(\beta^{(t-1)}\right)^T \cdot W\left(\beta^{(t-1)}\right) \cdot X\right) \cdot \beta\right.\right.$$

(I)

οπότε εάν

$$\beta/\widetilde{y}\left(\beta^{(t-1)}\right) \sim N\left(\mu_1, C_1\right)$$

η εκ των υστέρων κατανομή του $\beta/\widetilde{y}$ είναι:

$$f\left(\beta/\widetilde{y}\left(\beta^{(t-1)}\right)\right) \propto exp\left\{-\frac{1}{2} \cdot \left(\beta - \mu_1\right)^T \cdot \left(C_1\right)^{-1} \cdot \left(\beta - \mu_1\right)\right\}$$

$$\propto exp\left\{-\frac{1}{2}\cdot\left(\beta^T-\mu_1^T\right)\cdot(C_1)^{-1}\cdot(\beta-\mu_1)\right\}$$

$$\propto exp\left\{-\frac{1}{2}\cdot\left(\beta^T\cdot(C_1)^{-1}\cdot\beta-\beta^T\cdot(C_1)^{-1}\cdot\mu_1-\mu_1^T(C_1)^{-1}\cdot\beta+\mu_1^T(C_1)^{-1}\cdot\mu_1\right]\right\}$$

$$\propto exp\left\{-\frac{1}{2}\cdot\left(\beta^T\cdot(C_1)^{-1}\cdot\beta-\beta^T\cdot(C_1)^{-1}\cdot\mu_1-\mu_1^T(C_1)^{-1}\cdot\beta\right)\right\}$$

(II)

άρα συνδιάζοντας τις σχέσεις (I) και (II) έχουμε ότι:

$$(C_1)^{-1}=\left((C_0)^{-1}+X^T\cdot W\left(\beta^{(t-1)}\right)\cdot X\right)$$

και

$$\mu_1=(C_1)\cdot\left((C_0)^{-1}\cdot\mu_0+X^T\cdot W\left(\beta^{(t-1)}\right)\widetilde{y}\left(\beta^{(t-1)}\right)\right)$$

Τελικά η ζητούμενη εκ των υστέρων κατανομή είναι:

$$f\left(\beta/\widetilde{y}\left(\beta^{(t-1)}\right)\right)\sim N\left(\beta^{(t)},\left((C_0)^{-1}+X^T\cdot W\left(\beta^{(t-1)}\right)\cdot X\right)^{-1}\right)$$

όπου

$$\beta^{(t)}=C_1\cdot\left((C_0)^{-1}\cdot\mu_0+X^T\cdot W\left(\beta^{(t-1)}\right)\widetilde{y}\left(\beta^{(t-1)}\right)\right)$$

Εάν το $n\to\infty$ η *posterior* κατανομή παίρνει τη μορφή:

$$\beta/y\sim N\left(\widehat{\beta},\left((C_0)^{-1}+X^T\cdot W\left(\beta^{(t-1)}\right)\cdot X\right)^{-1}\right)$$

Για να υλοποιήσουμε τον αλγόριθμο *gamerman's metropolis − hastings IWLS algorithm* χρειαζόμαστε την πιθανότητα αποδοχής - απόρριψης που

δίνεται από τον τύπο

$$p = min\left\{1, \frac{f\left(\beta^{can}/y\right)}{f\left(\beta/y\right)} \cdot \frac{q\left(\beta/\beta^{can}, y\right)}{q\left(\beta^{can}/\beta, y\right)}\right\}$$

Όπου

$$\frac{f(\beta^{can}/y)}{f(\beta/y)} \cdot \frac{q(\beta/\beta^{can},y)}{q(\beta^{can}/\beta,y)} = \frac{\left\{\prod_{i=1}^{n} \frac{exp\left(x_i^T \beta^{can} y_i\right)}{\left[1+exp\left(x_i^T can\right)\right]^{n_i}}\right\} \cdot \exp\left(-\frac{(\beta^{can}-\mu_0)^T \cdot C_0^{-1} \cdot (\beta^{can}-\mu_0)}{2}\right)}{\left\{\prod_{i=1}^{n} \frac{exp\left(x_i^T \beta y_i\right)}{\left[1+exp\left(x_i^T \beta\right)\right]^{n_i}}\right\} \cdot \exp\left(-\frac{(\beta-\mu_0)^T \cdot C_0^{-1} \cdot (\beta-\mu_0)}{2}\right)} \cdot \frac{|C_{1can}|^{-1/2} \cdot (\beta-\mu_{1can})^T \cdot (C_{1can})^{-1} \cdot (\beta-\mu_{1can})}{|C_1|^{-1/2} \cdot (\beta^{can}-\mu_1)^T \cdot (C_1)^{-1} \cdot (\beta^{can}-\mu_1)}$$

Και δεχόμαστε ως γεννήτρια τυχαίων τιμών την κατανομή:

$$q\left(\beta_{can}/\beta, y\right) \; N\left(f\left(\beta\right), \left(C_0^{-1} + X^T \cdot W\left(\beta\right) \cdot X\right)^{-1}\right)$$

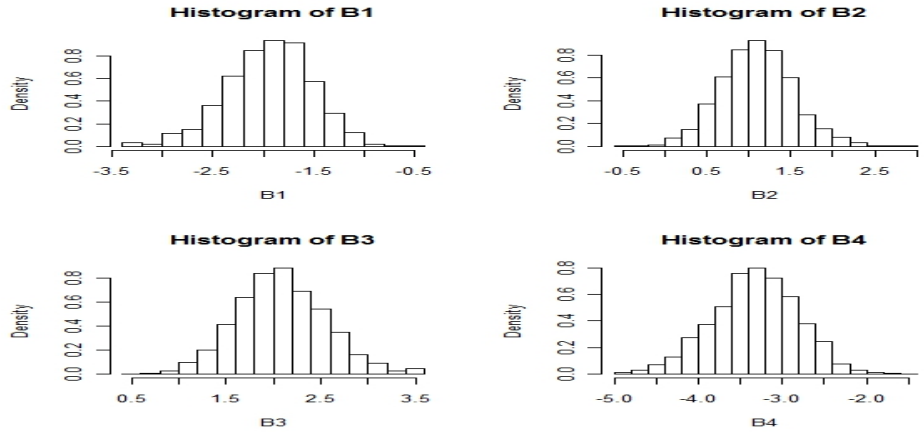
Με

$$f\left(\beta\right) = \left(C_0^{-1} + X^T \cdot W\left(\beta\right) \cdot X\right)^{-1} \cdot \left(C_0^{-1} \cdot \mu_0 + X^T \cdot W\left(\beta\right) \cdot \widetilde{y}\left(\beta\right)\right)$$

Τα αποτελέσματα ακολουθόντας τον αλγόριθμο επαναλαμβανόμενων σταθμισμένων ελαχίστων τετραγώνων ($IWLS$) του $Gamerman$ και χρησιμοποιώντας 5.000 επαναλήψεις με περίοδο ζεστάματος ($burn-in$) τις 500 πρώτες επαναλήψεις συνοψίζονται στον πίνακα 5.3 και το ποσοστό αποδοχής είναι 75.26%.

Πίνακας 5.3: Αποτελέσματα $gamerman's\ metropolis-hastings\ IWLS$ αλγορίθμου

$coefficients$	$posterior\ mean$	$posterior\ std$
$(intercept)$	-1.949641	0.4138379
$noplan$	1.109156	0.4342377
$factor$	2.078270	0.4450228
$antib$	-3.305292	0.4802814

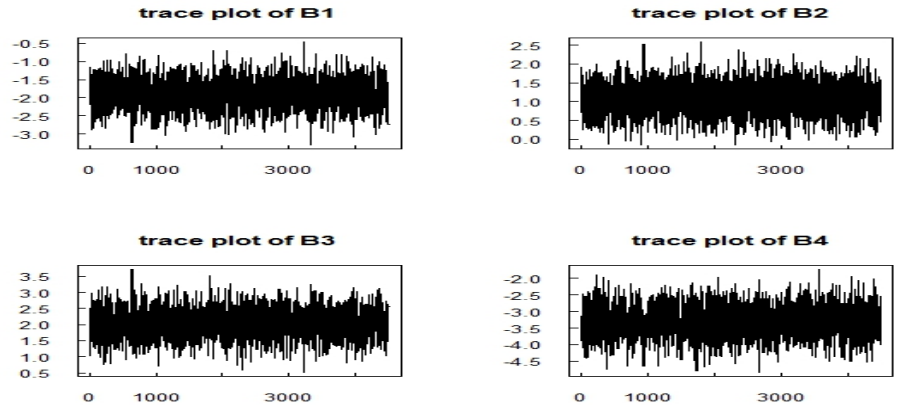


Σχήμα 5.1: *Ιστογράμματα για τα β βασισμένα στον gamerman's metropolis − hastings IWLS αλγόριθμο*

Επιπρόσθετα τα αποτελέσματα του παραπάνω αλγορίθμου φαίνονται στα γραφήματα 5.1 και 5.2.

Παρατηρούμε ότι τα ιστογράμματα στο γράφημα 5.1, τα οποία βασίζονται σε 4.500 επαναλήψεις από την εκ των υστέρων κατανομή, παρουσιάζουν συμμετρικές κατανομές. Επίσης γίνεται φανερό ότι το μηδέν δεν περιέχεται κοντά στο κέντρο των εκ των υστέρων κατανομών όποτε μπορούμε με ασφάλεια να ισχυριστούμε ότι οι παράμετροί μας είναι χρήσιμοι για το μοντέλο μας και δε μπορούμε να τις παραλείψουμε.

Το γράφημα *traceplot* (γράφημα 5.2) μας δείχνει ότι μετά τη περίοδο ζεστάματος ο αλγόριθμος έχει συγκλίνει.



Σχήμα 5.2: *Γράφημα για τα* β *του gamerman's metropolis − hastings IWLS αλγορίθμου*

5.1.3 Αλγόριθμος Αύξησης Δεδομένων

Θα περιγράψουμε την διαδικασία κατασκευής ενός μοντέλου *probit* Παλινδρόμησης για τα δεδομένα σχετικά με τις λοιμόξεις που προκύπτουν από την γέννηση με καισσαρική τομή (*Fahrmeir and Tutz*,2001). Θα μελετήσουμε τους ίδιους παράγοντες επιρροής κάνοντας χρήση όμως αυτή τη φορά της μεθόδου αύξησης δεδομένων :

Θεωρούμε λοιπόν ότι:

$$y_i \sim Bernoulli\left(\Phi\left(\eta_i\right)\right)$$

με $\Phi\left(\cdot\right)$ την αθροιστική συνάρτηση της κανονικής κατανομής, $\eta_i = x_i^T \cdot \beta$ ο γραμμικός προσδιορισμός του μοντέλου και $\beta \sim N\left(\mu_0, C_0\right)$. Σε ένα *probit* μοντέλο δεδομένου του ότι $\mu_i = \Phi\left(\eta_i\right)$ έχουμε ότι η συνάρτηση σύνδεσης είναι της μορφής $g\left(\mu_i\right) = \Phi^{-1}\left(\mu_i\right)$.

Η $y_i \in \{0,1\}, i = 1, \ldots, n$ λοιπόν είναι μία δίτιμη απαντητική μεταβλητή η οποία πέρνει τη τιμή 1 εάν παρατηρήθηκε λοίμοξη στη συγκεκριμένη ασθενή

και τη τιμή 0 διαφορετικά οπότε χρησιμοποιώντας τις βοηθήτικες μεταβλητές z_i μετασχηματίζουμε τα y_i έτσι ώστε :

$$y_i = \begin{cases} 1 & \text{εάν} \quad z_i > 0 \\ 0 & \text{διαφορετικά} \end{cases} \tag{5.1}$$

με $z_i \sim N(x_i^T \cdot \beta, 1)$

$\beta \sim N(\mu_0, C_0)$

Η από κοινού εκ των υστέρων κατανομή είναι της μορφής :

$$f(z, \beta/y) = f(y, z/\beta) \cdot f(\beta) \propto \left\{ \prod_{i=1}^{n} f\left(y_i/z_i, \beta\right) \cdot P\left(z_i\right) \right\} \cdot f\left(\beta\right) \propto$$

$$\propto \left\{ \prod_{y_i=1} exp\left[-\frac{1}{2}(z_i - x_i^T\beta)^2 \right] \cdot I(z_i > 0) \right\} \cdot \left\{ \prod_{y_i=0} exp\left[-\frac{1}{2}(z_i - x_i^T\beta)^2 \right] \cdot I(z_i < 0) \right\}$$

$$\cdot exp\left(-\frac{(\beta-\mu_0)^T \cdot C_0^{-1} \cdot (\beta-\mu_0)}{2} \right).$$

Οπότε η δεσμευμένη εκ των υστέρων κατανομή του β είναι της μορφής :

$$\beta/z, y \sim N\left(\left(C_0^{-1} + X^T X\right)^{-1} \left(C_0^{-1}\mu_0 + X^T z\right), \left(C_0^{-1} + X^T X\right)^{-1} \right)$$

Όπου η δεσμευμένη εκ των υστέρων κατανομή των z_i δίνετε από την κανονική *truncated* κατανομή η οποία έχει τη μορφή :

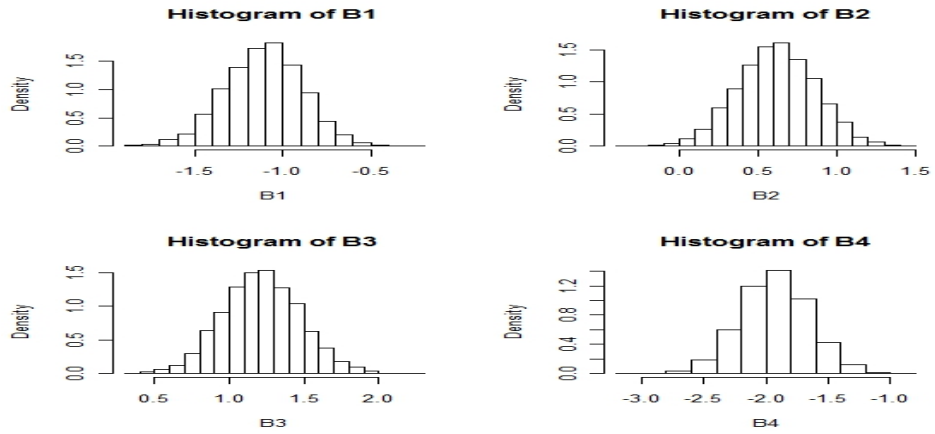
$$z_i/\beta, y = \begin{cases} N\left(x_i^T\beta, 1\right) \cdot I\left[z_i > 0\right] & \text{εάν} \quad y_i = 1 \\ N\left(x_i^T\beta, 1\right) \cdot I\left[z_i \leq 0\right] & \text{διαφορετικά} \end{cases} \tag{5.2}$$

Τα αποτελέσματα ακολουθόντας την μέθοδο της αύξησης δεδομένων και χρησιμοποιώντας 5.000 επαναλήψεις με περίοδο συγκλισης τις 500 πρώτες επαναλήψεις συνοψίζονται στον πίνακα 5.4:

Πίνακας 5.4: Αποτελέσματα αλγορίθμου αύξησης δεδομένων

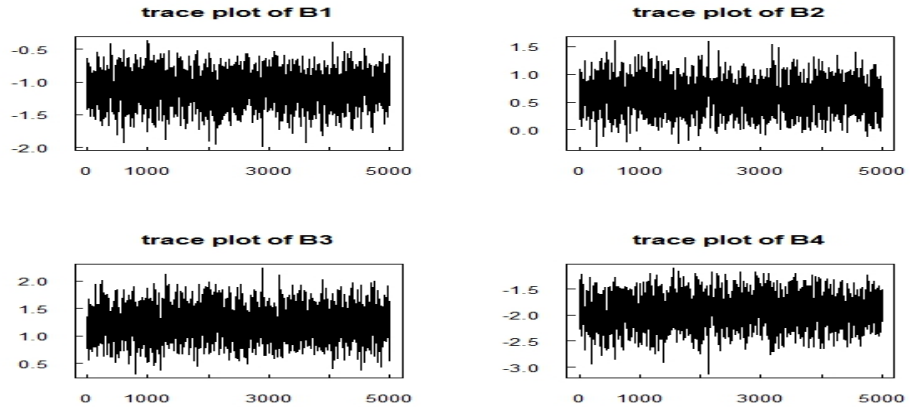
$coefficients$	$posterior\ mean$	$posterior\ std$
$(intercept)$	-1.0994610	0.2186791
$noplan$	0.5971601	0.2493319
$factor$	1.2090687	0.2520427
$antib$	-1.9090932	0.2738041

Επιπρόσθετα τα αποτελέσματα του παραπάνω αλγορίθμου για τα β φαίνονται στα γραφήματα 5.3 και 5.4.



Σχήμα 5.3: *Ιστογράμματα της probit Παλινδρόμησης με data augmentation αλγόριθμο για τα β*

Από τα ιστογράμματα (γράφημα 5.3) γίνεται φανερό ότι έχουμε συμμετρικές κατανομές όπου το μηδέν δεν εμφανίζεται στο κέντρο τους. Αυτό μας δίνει τη δυνατότητα να κατανοήσουμε την ακρίβεια της εκτίμησης μας και με βεβαιώτητα να ισχυριστούμε ότι χρειαζόμαστε όλες τις παραμέτρους.



Σχήμα 5.4: *Γράφημα της probit Παλινδρόμησης με data augmentation αλγόριθμο*

Στο γράφημα *traceplot* (γράφημα 5.4) μπορούμε να παρατηρήσουμε ότι η σύγκλιση φαίνεται να έχει επιτυχθεί μετά από περίοδο ζεστάματος ($burn - in$) τις 500 πρώτες παρατηρήσεις.

5.1.4 Μοντέλο *logit* με Τυχαίες Επιδράσεις

Γενικεύοντας το μοντέλο της λογιστικής παλινδρόμησης, χρησιμοποιώντας για κάθε παρατήρηση μια επιπρόσθετη τυχαία επίδραση ($random\ effect$) η οποία υποθέτουμε ότι ακολουθεί την κανονική κατανομή έτσι ώστε $b_i \sim N(0, \omega^{-1})$, έχουμε:

$$logit\left(p_i\right) = x_i^T\beta + b_i$$

$$\Rightarrow log\left(\frac{p_i}{1-p_i}\right) = x_i^T\beta + b_i$$

$$\Rightarrow \frac{p_i}{1-pi} = exp\left(x_i^T\beta + b_i\right)$$

$$\Rightarrow p_i = exp\left(x_i^T\beta + b_i\right) - p_i \cdot exp\left(x_i^T\beta b_i right\right)$$

$$\Rightarrow p_i \cdot \left(1 + exp\left(x_i^T\beta + b_i\right)\right) = exp\left(x_i^T\beta + b_i\right)$$

$$\Rightarrow p_i = \frac{exp\left(x_i^T\beta + b_i\right)}{1 + exp\left(x_i^T\beta + b_i\right)}$$

Δεδομένου του ότι οι εκ των προτέρων κατανομές είναι:

$$\beta \sim N(\mu_0, C_0)$$

$$b_i \sim N(0, \omega^{-1})$$

$$\omega \sim Gamma(c, d)$$

Η εκ των υστέρων κατανομή για τα β είναι της μορφής:

$$f\left(\beta/y\right) \propto f\left(y/\beta, b\right) \cdot f\left(\beta\right) \cdot f\left(b/\omega\right) \cdot f\left(\omega\right) \propto$$

$$\left\{\prod_{i=1}^{n} \frac{exp\left(x_i^T\beta + b_i\right)^{y_i}}{\left[1 + exp\left(x_i^T\beta + bi\right)\right]^{n_i}}\right\} \cdot \exp\left(-\frac{\left(\beta - \mu_0\right)^T \cdot C_0^{-1} \cdot \left(\beta - \mu_0\right)}{2}\right) \cdot \prod_{i=1}^{n} exp\left\{\frac{\omega b_i^2}{2} \cdot \omega^{(c-1)exp(-d\omega)}\right\}$$

Για τη λογιστική παλινδρόμηση με διωνυμικά δεδομένα χρησιμοποιούμε τη κανονική συνδετική συνάρτηση:

$$g(\mu_i) = log\left(\frac{p_i}{1-p_i}\right) = log\left(\frac{n_i \cdot p_i}{n_i - n_i \cdot p_i}\right) = log\left(\frac{\mu_i}{n_i - \mu_i}\right)$$

και παράγουμε τα β, δεδομένων των b τα οποία θεωρούμε και ως $offset$, χρησιμοποιώντας τον εξής μετασχηματισμό των δεδομένων:

$$\widetilde{y}_i(\beta) = \eta_i + (y_i - \mu_i)\, g'(\mu_i) - bi = x_i^T \beta + (y_i - \mu_i)\, g'(\mu_i)$$

Με αντίστοιχα βάρη: $W_i(\beta) = 1/g'(\mu_i)$
Όμοια για την παραγωγή των b_i, δεδομένου του ότι θεωρούμε τον όρο $x_i^T\beta$ σαν $offset$, χρησιμοποιούμε τον εξής μετασχηματισμό των δεδομένων:

$$\widetilde{y}_i(b_i) = \eta_i + (y_i - \mu_i)\, g'(\mu_i) - bi = b_i + (y_i - \mu_i)\, g'(\mu_i)$$

με αντίτοιχα βάρη: $W_i(b_i) = 1/g'(\mu_i)$
Για να υλοποιήσουμε τον αλγόριθμο επαναλαμβανόμενων σταθμισμένων ελαχίστων τετραγώνων ($IWLS$) του $Gamerman$ με τυχαίες επιδράσεις χρειαζόμαστε την πιθανότητα αποδοχής - απόρριψης που δίνεται από τον τύπο

$$p = min\left\{1, \frac{f(\beta^{can}/y)}{f(\beta/y)} \cdot \frac{q(\beta/\beta^{can}, y)}{q(\beta^{can}/\beta, y)}\right\}$$

Και δεχόμαστε ως γεννήτρια τυχαίων τιμών την κατανομή:

$$b_i^{can} \sim N\left(\frac{W_i(b_i)\,\widetilde{y}_i(b_i)}{\omega + W_i(b_i)}, \frac{1}{\omega + W_i(b_i)}\right)$$

Τα αποτελέσματα ακολουθόντας τον αλγόριθμο επαναλαμβανόμενων σταθμισμένων ελαχίστων τετραγώνων ($IWLS$) του *Gamerman* με τυχαίες επιδράσεις και χρησιμοποιώντας 5.000 επαναλήψεις με περίοδο ζεστάματος ($burn\ in$) τις 500 πρώτες επαναλήψεις για το β και για το b συνοψίζονται στους πίνακες 5.5, 5.6 αντίστοιχα. Επίσης, το ποσοστό αποδοχής για το β είναι 70.82%, καθώς και τα αποτελεματα για το ποσοστό αποδοχής για το b παρουσιάζονται στο πίνακα 5.7.

Πίνακας 5.5: Αποτελέσματα $gamerman's\ metropolis - hastings\ IWLS$ αλγορίθμου

$coefficients$	$posterior\ mean$	$posterior\ std$
$(intercept)$	-2.2843408	0.9778951
$noplan$	0.6124758	1.3276814
$factor$	2.9187720	1.5547793
$antib$	3.5411593	1.1074935

Επιπρόσθετα τα αποτελέσματα του παραπάνω αλγορίθμου για τα B φαίνονται γραφήματα 5.5 και 5.6.

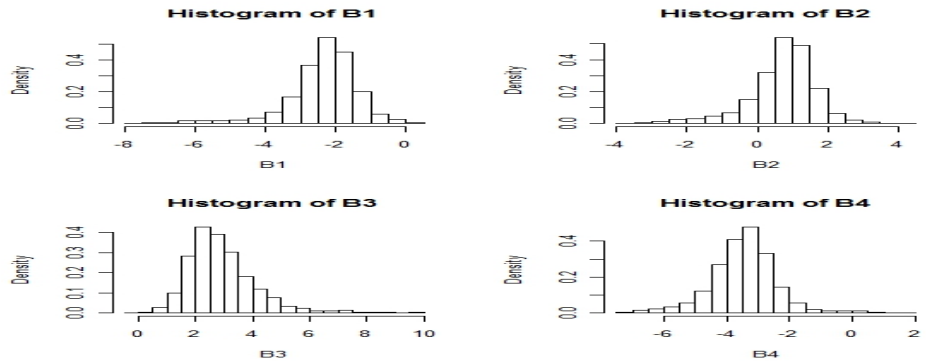
Πίνακας 5.6: Αποτελέσματα $gamerman's\ metropolis-hastings\ IWLS$ αλγορίθμου

$coefficients$	$posterior\ mean$	$posterior\ std$	$coefficients$	$posterior\ mean$	$posteriorstd$
b_1	0.57860874	0.9747781	b_5	-0.47640504	1.1032402
b_2	-0.05216705	1.0824121	b_6	0.42778052	0.8889192
b_3	-0.59581430	1.1188614	b_7	0.17421156	0.9308143
b_4	-0.15600399	1.0114384			

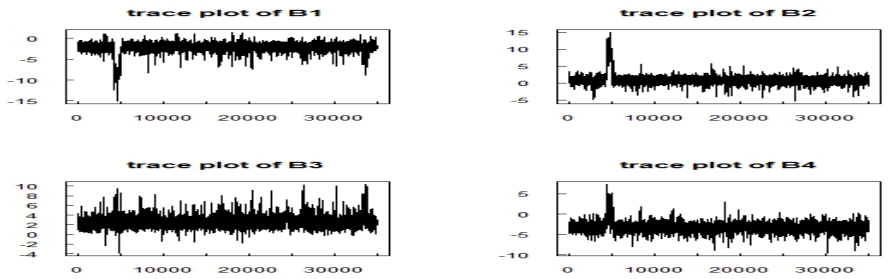
Πίνακας 5.7: Ποσοστό αποδοχής του $gamerman's\ metropolis-hastings\ IWLS$ αλγορίθμου

$coefficients$	ποσοστό	$coefficients$	ποσοστό
b_1	¨93.98'	b_5	¨94.36'
b_2	¨99.02'	b_6	¨92.1'
b_3	¨99.6'	b_7	¨92.14'
b_4	¨92.22'		

Στα ιστογράμματα (γράφημα 5.5) παρατηρούμε ότι οι εκ των υστέρων κατανομές παρουσιάζουν ασυμετρίες και ειδικά σε αυτή που αναφέρεται στην παράμετρο B_2, το μηδέν βρίσκεται κοντά στο κέντρο της. Από αυτό μπορούμε να ισχυριστούμε ότι η παράμετρος αυτή μπορεί να παραληφθεί.
Το γράφημα $traceplot$ (γράφημα 5.6) μας πληροφορεί ότι ο αλγόριθμος δεν έχει συγκλίνει ακόμα και μετά από περίοδο ζεστάματος ($burn-in$) τις 5000 πρώτες παρατηρήσεις.



Σχήμα 5.5: *Ιστογράμματα του gamerman's metropolis – hastings IWLS αλγορίθμου για τα B*



Σχήμα 5.6: *Γράφημα του gamerman's metropolis – hastings IWLS αλγορίθμου για τα B*

5.1.5 Αναπαραμετροποιώντας το Αρχικό Μοντέλο

Για να απλοποιηθεί ο αλγόριθμος αναπαραμετροποιούμε το μοντέλο μας και αντί για τα β θεωρούμε το $\eta_i/\beta \sim N(x_i^T\beta, \omega^{-1})$, και με βάση αυτό το μετασχηματισμό το μοντέλο μας χρησιμοποιώντας το μοντέλο της λογιστικής παλινδρόμησης για τα παραπάνω δεδομένα γράφετε ως εξής:

$$logit\,(p_i) = \eta_i$$

$$\Rightarrow log\left(\frac{p_i}{1-p_i}\right) = \eta_i$$

$$\Rightarrow \frac{p_i}{1-pi} = exp(\eta_i)$$

$$\Rightarrow p_i = \frac{exp(\eta_i)}{1+exp(\eta_i)}$$

Δεδομένου του ότι οι εκ των προτέρων κατανομές ειναι:

$$\eta_i \sim N\left(x_i^T\beta, \omega^{-1}\right)$$

$$\beta \sim N\left(\mu_0, C_0\right)$$

$$\omega \sim Gamma\left(c, d\right)$$

Η εκ των υστέρων κατανομή για τα β παίρνει τη μορφή:

$$f\left(\beta/\eta\right) \sim N\left(\left(C_0^{-1}+\omega X^T X\right)^{-1}\left(C_0^{-1}\mu_0+\omega X^T\eta\right), \left(C_0^{-1}+\omega X^T X\right)^{-1}\right)$$

Και δεχόμαστε ως γεννήτρια τυχαίων τιμών την κατανομή:

$$\eta_i^{can} \sim N\left(\frac{\omega x_i^T\beta + W_i\left(\eta_i\right)\widetilde{y}_i\left(\eta_i\right)}{\omega + W_i\left(\eta_i\right)}, \frac{1}{\omega + W_i\left(\eta_i\right)}\right)$$

Όπου

$$\widetilde{y}_i\left(\eta_i\right) = \eta_i + \left(y_i - \mu_i\right)g'(\mu_i)$$

$$W_i\left(\eta_i\right) = 1/g'\left(\mu_i\right)$$

Τα αποτελέσματα ακολουθόντας τον $gamerman's\ metropolis-hastings\ IWLS\ algorithm$ και χρησιμοποιώντας 5.000 επαναλήψεις με περίοδο συγκλισης τις 500 πρώτες επαναλήψεις, για τα β, b καθώς και τα ποσοστά αποδοχής για τα η συνοψίζονται στους πίνακες 5.8, 5.9 και 5.10 αντίστοιχα.

Επιπρόσθετα τα αποτελέσματα του παραπάνω αλγορίθμου για τα B φαίνονται στα γραφήματα 5.7 και 5.8 αντίστοιχα.

Πίνακας 5.8: Αποτελέσματα $gamerman's\ metropolis-hastings\ IWLS$ αλγορίθμου με αναπαραμέτρηση για τα β

$coefficients$	$posterior\ mean$	$posterior\ std$
$(intercept)$	-2.1021314	0.7324329
$noplan$	0.8159975	0.8402021
$factor$	2.5033345	0.9333387
$antib$	-3.3825748	0.9157245

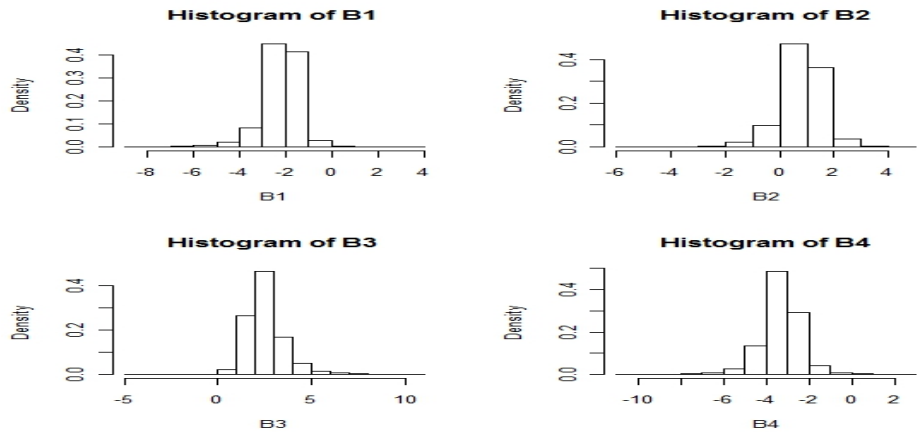
Πίνακας 5.9: Αποτελέσματα $gamerman's\ metropolis-hastings\ IWLS$ αλγορίθμου με αναπαραμέτρηση για τα b

$coefficients$	$posterior\ mean$	$posterior\ std$	$coefficients$	$posterior\ mean$	$posteriorstd$
b_1	0.403049808	0.6950633	b_5	-0.428859364	0.7492074
b_2	-0.005628309	0.7314508	b_6	0.365990699	0.6918776
b_3	-0.361692321	0.6699003	b_7	0.067641571	0.6526819
b_4	-0.048124123	0.6064760			

Στα ιστογράμματα (γράφημα 5.7) παρατηρούμε ότι οι εκ των υστέρων κατανομές παρουσιάζουν ασυμετρίες και ειδικά σε αυτή που αναφέρεται στην παράμετρο B_2, το μηδέν βρίσκεται κοντά στο κέντρο της. Από αυτό μπορούμε να ισχυριστούμε ότι η παράμετρος αυτή μπορεί να παραληφθεί.
Το γράφημα $traceplot$ (γράφημα 5.8) μας πληροφορεί ότι ο αλγόριθμος δεν έχει συγκλίνει ακόμα και μετά από περίοδο ζεστάματος $(burn-in)$ τις 5000 πρώτες παρατηρήσεις.

Πίνακας 5.10: Ποσοστό αποδοχής του $gamerman's\ metropolis - hastings$ $IWLS$ αλγορίθμου με αναπαραμέτρηση για τα η

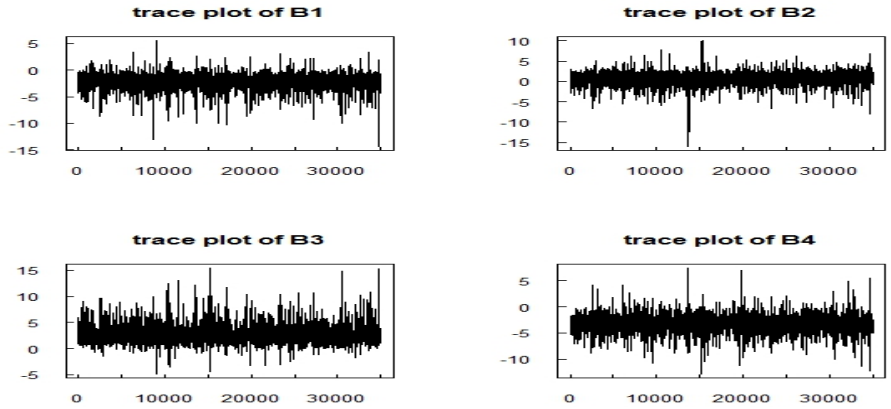
$coefficients$	ποσοστό	$coefficients$	ποσοστό	$coefficients$	ποσοστό
η_1	¨94.46'	η_3	¨99.48'	η_5	¨96.44'
η_2	¨99.48'	η_4	¨94.7'	η_6	¨95.28'
η_7	¨93.86'				



Σχήμα 5.7: *Ιστογράμματα του gamerman's metropolis − hastings IWLS αλγορίθμου με αναπαραμέτρηση για το B*

5.1.6 Συμπεράσματα

Συγκρίνοντας τους παραπάνω αλγορίθμους που αφορούν στις γεννήσεις με καισαρική τομή *Caesarian data* (*Fahrmeir and Tutz*, 2001), βλέπουμε ότι ο αλγόριθμος σταθμισμένων ελαχίστων τετραγώνων του *Gamerman* όπως και ο αλγόριθμος αύξησης δεδομένων έχουν καλύτερη σύγκλιση σε σχέση με τους άλλους δύο αλγορίθμους (τον αλγόριθμο του *Gamerman* με τυχαίες επιδράσεις και τον αλγόριθμο με αναπαραμέτρηση του αρχικού μας μοντέλου) και αυτό



Σχήμα 5.8: *Γράφημα gamerman's metropolis − hastings IWLS αλγορίθμου με αναπαραμέτρηση για τα B*

οφείλετε κατά κύριο λόγο στη πολυπλοκότητα των δύο τελευταίων. Ανάμεσα στους δύο αρχικούς αλγορίθμους (τον αλγόριθμο σταθμισμένων ελαχίστων τετραγώνων του *Gamerman* και τον αλγόριθμο αύξησης δεδομένων) παρατηρούμε ότι ενώ ο $Metropolis-Hastings$ αλγόριθμος στη γενική του περίπτωση συνήθως συγκλίνει πιο δύσκολα σε σχέση με τον αλγόριθμο του *Gibbs*, στη συγκεκριμένη περίπτωση όπου έχουμε την γεννήτρια προτεινόμενων τιμών του *Gamerman* ο αλγόριθμος γίνεται καλύτερος από τον *Gibbs* ως προς τη σύγκλιση.

5.2 Μπεϋζιανή Εκτίμηση Μοντέλων Παλινδρόμησης *Poisson*

Αυτό το σύνολο δεδομένων περιέχει πληροφορίες σχετικά με την εμφάνιση του καρκίνου του χείλους σε 56 περιοχές της Σκωτίας κατά τη διάρκεια των ετών 1975-1980, *Incidence of lip cancer in* 56 *areas in Scotland* (*Breslow, N.E. and Clayton, D.G.*, 1993). Για κάθε περιφέρεια, ο παρατηρούμενος και ο αναμενόμενος αριθμός των περιπτώσεων καρκίνου του χείλους δίνονται, κάτι το οποίο καθιστά δυνατό τον υπολογισμό της $SMR's$ (*standardized morbidity ratio* - τυποποιημένο ποσοστό νοσηρότητας) ως μία εκτίμηση για το σχετικό κίνδυνο. Επίσης, δεδομένο είναι το ποσοστό των ατόμων που εργάζονται στη δασοκομία, την αλιεία ή τη γεωργία.

Τα δεδομένα δίνονται στον πίνακα 5.13.

Παρατηρούμε λοιπόν ότι:

- ο δείκτης i μας πληροφορεί για την εκάστοτε περιοχή,
- το y είναι ο παρατηρούμενος αριθμός ασθενών σε κάθε περιοχή,
- το e είναι ο αναμενόμενος αριθμός ασθενών σε κάθε περιοχή,
- το x είναι το ποσοστό των ατόμων που εργάζονται σε εξωτερικό χώρο (*outdoors*).

Τα δεδομένα αυτά μπορούν να μοντελοποιηθούν υποθέτοντας ότι :

$$y_i \sim Poisson\,(\mu_i)$$

5.2.1 Ανάλυση με Βάση τη Κλασική Στατιστική

Τα αποτελέσματα ακολουθόντας την κλασική στατιστική συνοψίζονται στον πίνακα 5.11.

Πίνακας 5.11: Αποτελέσματα εφαρμόζοντας τη κλασική στατιστική

$coefficients$	$estimate\ std$	$error$	$z\ value$	$Pr(>\|z\|)$
$(intercept)$	-0.202973	0.066189	-3.067	0.00217
$outdoors$	0.026246	0.005995	4.378	1.20e-05

Παρατηρώντας τα αποτελέσματα του πίνακα 5.11 διαπιστώνουμε ότι οι μεταβλητές είναι στατιστικά σημαντικές. Από τον παραπάνω πίνακα γίνεται επίσης φανερό ότι υπαρχει θετική συσχέτιση ανάμεσα στην εμφάνιση του καρκίνου του χείλους σε 56 περιοχές της Σκωτίας και στο ποσοστό των ατόμων που δούλευαν σε εξωτερικό χώρο ($outdoors$). Πιο συγκεκριμένα παρατηρούμε ότι τα $odds$ εμφάνισης του καρκίνου του χείλους αυξάνονται για την περίπτωση ανθρώπων που εργάζονταν σε εξωτερικό χωρο κατά $exp(0.026246)$.

5.2.2 Ο Αλγόριθμος Επαναλαμβανόμενων Σταθμισμένων Ελαχίστων Τετραγώνων του *Gamerman*

Θεωρώντας ότι η εκ των προτέρων κατανομή για τα β είναι:

$$\beta \sim N\left(\mu_0, C_0\right)$$

Εάν υποθέσουμε ότι $\mu_i = e_i\lambda_i$ και $\lambda_i = exp\left(x_i^T\beta\right)$

Τότε χρησιμοποιώντας τη λογαριθμική συνάρτηση σύνδεσης για $poisson$ δεδομένα: $log\left(\mu_i\right) = \eta_i = log\left(e_i\right) + x_i^T\beta$

Χρησιμοποιώντας τον εξής μετασχηματισμό των δεδομένων:

$$\widetilde{y}_i\left(\beta\right) = h_i + \left(y_i - \mu_i\right)g'\left(\mu_i\right) - log\left(e_i\right)$$

Με βάρη:

$$W_i = 1/g'\left(\mu_i\right) = \mu_i$$

Για να υλοποιήσουμε τον αλγόριθμο επαναλαμβανόμενων σταθμισμένων ελαχίστων τετραγώνων ($IWLS$) του $Gamerman$ χρειαζόμαστε την πιθανότητα αποδοχής - απόρριψης που δίνεται από τον τύπο

$$p = min\left\{1, \frac{f\left(\beta^{can}/y\right)}{f\left(\beta/y\right)} \cdot \frac{q\left(\beta/\beta^{can}, y\right)}{q\left(\beta^{can}/\beta, y\right)}\right\}$$

Και δεχόμαστε ως γεννήτρια τυχαίων τιμών την κατανομή:

$$q\left(\beta_{can}/\beta, y\right)\ N\left(f\left(\beta\right), \left(C_0^{-1} + X^T \cdot W\left(\beta\right) \cdot X\right)^{-1}\right)$$

Με

$$f\left(\beta\right) = \left(C_0^{-1} + X^T \cdot W\left(\beta\right) \cdot X\right)^{-1} \cdot \left(C_0^{-1} \cdot \mu_0 + X^T \cdot W\left(\beta\right) \cdot \widetilde{y}\left(\beta\right)\right)$$

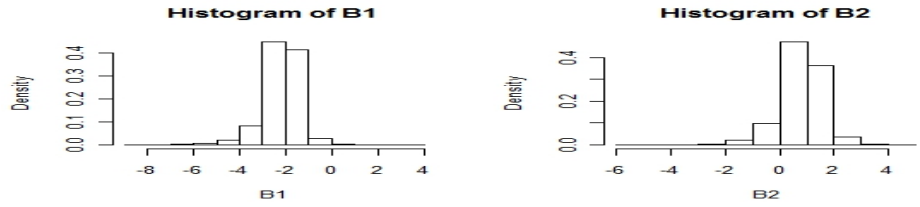
Τα αποτελέσματα ακολουθόντας τον τον αλγόριθμο επαναλαμβανόμενων σταθμισμένων ελαχίστων τετραγώνων ($IWLS$) του $Gamerman$ και χρησιμοποιώντας 5.000 επαναλήψεις με περίοδο συγκλισης τις 500 πρώτες επαναλήψεις συνοψίζονται στον πίνακα 5.12 και το ποσοστό αποδοχής είναι 86.4%.

Πίνακας 5.12: Αποτελέσματα της $Poisson$ Παλινδρόμησης με $gamerman's\ metropolis - hastings\ IWLS\ algorithm$

$coefficients$	$posterior\ mean$	$posterior\ std$
$(intercept)$	-0.20542695	0.067159607
$outdoors$	0.02619939	0.006102588

Επιπρόσθετα τα αποτελέσματα του παραπάνω αλγορίθμου για το β φαίνονται στα γραφήματα 5.9 και 5.10.

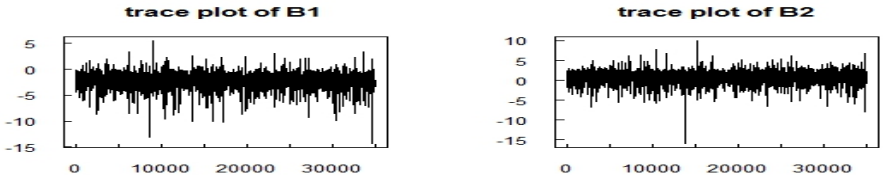
Παρατηρώντας τα γραφήματα που παρουσιάζονται τα ιστογράμματα (γράφημα 5.9) είναι φανερό ότι έχουμε εκ των υστέρων συμμετρικές κατανομές. Δεδομένου του ότι το μηδέν βρίσκεται κοντά στο κέντρο της εκ των υστέρων



Σχήμα 5.9: *Ιστόγραμμα για τα β βασισμένα στον αλγόριθμο Poisson Παλινδρόμησης με gamerman's metropolis − hastings IWLS*

κτανομής της παραμέτρου B_2 μπορούμε να τη θεωρήσουμε ως ασήμαντη και να τη παραλήψουμε με ασφάλεια.

Στο *traceplot* γράφημα (γράφημα 5.10) παρατηρούμε ότι η δεν έχει επιτευχθεί σύγκλιση ακόμα και μετά από περίοδο ζεστάματος ($Burn-in$) τις 5000 πρώτες παρατηρήσεις.



Σχήμα 5.10: *Γράφημα της Poisson Παλινδρόμησης με gamerman's metropolis − hastings IWLS algorithm*

Πίνακας 5.13: *Incidence of lip cancer in* 56 *areas in Scotland*

i	y	e	x	i	y	e	x
1	9	1.4	10	29	16	14.4	7
2	39	8.7	16	30	11	10.2	16
3	11	3.0	10	31	5	4.8	10
4	9	2.5	16	32	3	2.9	16
5	15	4.3	7	33	7	7.0	0
6	8	2.4	7	34	8	8.5	0
7	26	8.1	10	35	11	12.3	0
8	7	2.3	10	36	9	10.1	1
9	6	2.0	10	37	11	12.7	1
10	20	6.6	16	38	8	9.4	1
11	13	4.4	16	39	6	7.2	0
12	5	1.8	16	40	4	5.3	1
13	3	1.1	16	41	10	18.8	1
14	8	3.3	7	42	8	15.8	1
15	17	7.8	1	43	2	4.3	7
16	9	4.6	10	44	6	14.6	1
17	2	1.1	1	45	19	50.7	0
18	7	4.2	7	46	3	8.2	10
19	9	5.5	1	47	2	5.6	7
20	7	4.4	10	48	3	9.3	7
21	16	10.5	1	49	28	88.7	7
22	31	22.7	10	50	6	19.6	24
23	11	8.8	7	51	1	3.4	24
24	7	5.6	10	52	1	3.6	7
25	19	15.5	7	53	1	5.7	16
26	15	12.5	10	54	1	7.0	24
27	7	6.0	24	55	0	4.2	7
28	10	9.0	16	56	0	1.8	7

Κεφάλαιο 6

Εφαρμογή της Λογιστικής Παλινδρόμησης στην Εκτίμηση της Πιθανότητας Χρεωκοπίας Επιχειρήσεων

Στο παρόν κεφάλαιο θα ασχοληθούμε με τη μελέτη της εκτίμησης της πιθανότητας χρεωκοπίας επιχειρήσεων λαμβάνοντας υπόψιν ότι τα συμπτώματα εμφανίζονται στον ισολογισμό της επιχείρησης για ένα ή δύο χρόνια πριν η εταιρεία πτωχεύσει πραγματικά. Η ακριβής πρόβλεψη της οικονομικής αποτυχίας μπορεί να δόσει χρόνο στα εταιρικά στελέχη να αναλάβουν δράση και να σώσουν την επιχείρηση από οικονομική αφερεγγυότητα. Ως εκ τούτου, η ακριβής πρόβλεψη της πτώχευσης είναι ένα σημαντικό και ευρέως μελετημένο θέμα στο λογιστικό τομέα. Στατιστικές τεχνικές εφαρμόζονται όλο και περισσότερο στην πρόβλεψη τέτοιων οικονομικών αποτυχιών.

6.1 Δεδομένα

Τα δεδομένα που χρησιμοποιήθηκαν προέρχονται από τα ¨1968, 1969, 1970, 1971, 1972 *Moody's Industrial Manuals*¨, (*Bankruptcy Data*, πίνακας 6.3). Στα δεδομένα αυτά αναφέρονται τέσσερις τύποι ετήσιων οικονομικών δεικτών για 21 επιχειρήσεις που πτώχευσαν σε περίπου 2 χρόνια μετά την πραγματοποίηση της έρευνας και για 25 επιχειρήσεις οι οποίες παρέμειναν υγιείς περίπου την ίδια χρονική περίοδο.

Οι προγνωστικοί παράγοντες είναι:

$x_1 = \frac{CF}{TD}$ = ταμειακές ροές προς το συνολικό χρέος

$x_2 = \frac{NI}{TA}$ = καθαρό εισοδήμα προς το σύνολο του ενεργητικού

$x_3 = \frac{CA}{CL}$ = κυκλοφορούντα στοιχεία ενεργητικού προς βραχυπρόθεσμες υποχρεώσεις

$x_4 = \frac{CA}{NS}$ = κυκλοφορούντα στοιχεία ενεργητικού προς καθαρές πωλήσεις

και η απαντητική μεταβλητή είναι :

$$y_i = \begin{cases} 1 & \text{εάν η επιχείρηση είναι οικονομικά υγιής} \\ 0 & \text{διαφορετικά} \end{cases} \tag{6.1}$$

6.2 Ανάλυση με Βάση τη Κλασική Στατιστική

Προσαρμόζουμε το παρακάτω πλήρες μοντέλο *probit* Παλινδρόμησης στα δεδομένα μας

$$y_i \sim Bernoulli(p_i)$$

$$\vartheta_i = \Phi^{-1}(p_i) = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_4, i = 1, \ldots, 46$$

Τα αποτελέσματα ακολουθόντας την κλασική στατιστική συνοψίζονται στον πίνακα 6.1.

Πίνακας 6.1: Αποτελέσματα εφαρμόζοντας τη κλασική στατιστική

$coefficients$	$estimate\ std$	$error$	$z\ value$	$Pr(>\|z\|)$
$\beta_1(intercept)$	-2.754	1.227	-2.244	0.02484
$\beta_2(x_1 = \frac{CF}{TD})$	4.167	3.169	1.315	0.18860
$\beta_3(x_2 = \frac{NI}{TA})$	-2.527	6.855	-0.369	0.71243
$\beta_4(x_3 = \frac{CA}{CL})$	1.821	0.603	3.019	0.00253
$\beta_5(x_4 = \frac{CA}{NS})$	-1.779	1.604	-1.109	0.26733

Παρατηρώντας τα αποτελέσματα του πίνακα 6.1 γίνεται φανερό ότι οι μεταβλητές $x_1 = \frac{CF}{TD}$, $x_2 = \frac{NI}{TA}$ και $x_4 = \frac{CA}{NS}$ δεν είναι στατιστικά σημαντικές. Η μεταβλητή $x_1 = \frac{CF}{TD}$ μας παρέχει ένδειξη για την ικανότητα της επιχείρησης να καλύπτει το συνολο του χρέους της με ετήσιες ταμειακές ροές από λειτουργικές δραστηριότητες. Η μεταβλητή $x_2 = \frac{NI}{TA}$ μας παρέχει ένα πρότυπο για το πόσο αποτελεσματική είναι η οικονομική διαχείρηση της επιχείρησης. Η μεταβλητή $x_3 = \frac{CA}{CL}$ μας δείχνει την ικανότητα της εκάστοτε επιχείρησης να ανταποκρίνεται στις τρέχουσες υποχρεώσεις της και επιπρόσθετα η μεταβλητή αυτή είναι ικανή να διαφοροποιήσει και να προσδιορίσει προβληματικές επιχειρήσεις. Τέλος η μεταβλητή $x_4 = \frac{CA}{NS}$ μετρά την ικανότητα της επιχήρησης να αξιοποιήσει περιουσιακά της στοιχεία για να βελτιώσει τις πωλήσεις. Επίσης από τον ίδιο πίνακα γίνεται φανερό ότι υπάρχει θετική συσχέτιση ανάμεσα στην πιθανότητα μη εμφάνισης χρεωκοπίας της επιχείρησης και στη μεταβλητή x_3. Αυτό σημαίνει ότι η αύξηση της μεταβλητής x_3 κατά μία μονάδα αντιστοιχεί σε αύξηση του $log - odds$ της πιθανότητας να είναι υγιής η επιχείρηση κατά 1.821.

Επειδή οι μεταβλητές $x_1 = \frac{CF}{TD}$, $x_2 = \frac{NI}{TA}$ και $x_4 = \frac{CA}{NS}$ στο πλήρες μοντέλο δεν είναι στατιστικά σημαντικές πρέπει να ακολουθήσουμε κάποια διαδικασία επιλογής μοντέλου για το πρόβλημα. Ακολουθώντας την $Backward$ διαδικασία επιλογής επεξηγηματικών μεταβλητών καταλήγουμε στο μοντέλο που περιέχει

τις μεταβλητές $x_1 = \frac{CF}{TD}$ και $x_3 = \frac{CA}{CL}$, δηλαδή στο $\vartheta_i = \Phi^{-1}(p_i) = \beta_1 + \beta_2 x_1 + \beta_4 x_3, i = 1, \ldots, 46$. Τα αποτελέσματα που προκύπτουν συνοψίζονται στον πίνακα 6.2.

Πίνακας 6.2: Αποτελέσματα εφαρμόζοντας τη κλασική στατιστική

$coefficients$	$estimate\ std$	$error$	$z\ value$	$Pr(>\|z\|)$
$\beta_1(intercept)$	-3.0363	0.9785	-3.103	0.00192
$\beta_2(x_1 = \frac{CF}{TD})$	3.6684	1.5459	2.373	0.01764
$\beta_4(x_3 = \frac{CA}{CL})$	1.5501	0.5048	3.071	0.00213

6.3 Μπεϋζιανή Ανάλυση

Χρησιμοποιώντας τη διαδικασία κατασκευής ενός μοντέλου *probit* παλινδρόμησης που περιγράψαμε στη παράγραφο 5.1.3 κάνοντας χρήση της μεθόδου αύξησης δεδομένων και χρησιμοποιώντας 5.000 επαναλήψεις με περίοδο συγκλισης τις 500 πρώτες επαναλήψεις, τα αποτελέσματα συνοψίζονται στον πίνακα 6.3.

Πίνακας 6.3: Αποτελέσματα αλγορίθμου αύξησης δεδομένων

$coefficients$	$posterior\ mean$	$posterior\ std$
$\beta_1(intercept)$	-3.025318	1.2415073
$\beta_2(x_1 = \frac{CF}{TD})$	4.503453	3.2984774
$\beta_3(x_2 = \frac{NI}{TA})$	-1.804783	7.4718844
$\beta_4(x_3 = \frac{CA}{CL})$	2.006851	0.6190486
$\beta_5(x_4 = \frac{CA}{NS})$	-2.057761	1.7823257

Επιπρόσθετα τα αποτελέσματα του παραπάνω αλγορίθμου για τα β φαίνονται στα γραφήματα 6.1 και 6.2. Παρατηρώντας τα ιστογράμματα (γράφημα 6.1) των εκ των υστέρων κατανομών γίνεται φανερό ότι το μηδέν βρίσκεται κοντά στο κέντρο της κατανομής των παραμέτρων $b2$, $b3$ και $b5$. Στο γράφημα *traceplot* (γράφημα 6.2) μπορούμε να παρατηρήσουμε ότι η σύγκλιση φαίνεται να έχει επιτευχθεί μετά από περίοδιο ζεστάματος ($Burn - in$) τις 500 πρώτες παρατηρήσεις.

Ακολουθώντας την μέθοδο επιλογής μοντέλου της παραγράφου 6.1.1 επιλέγουμε τελικά το μοντέλο που περιέχει τις μεταβλητές $x_1 = \frac{CF}{TD}$ και $x_3 = \frac{CA}{CL}$.

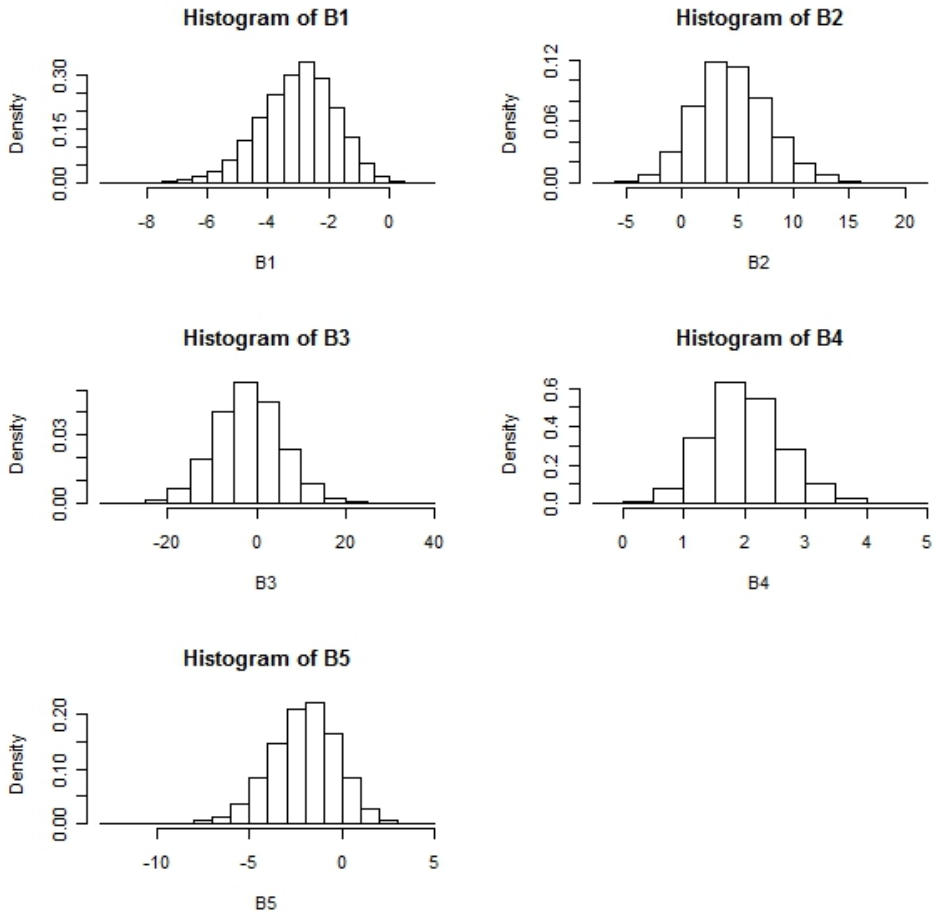
Χρησιμοποιώντας την ίδια διαδικασία κατασκευής ενός μοντέλου *probit* παλινδρόμησης που περιγράψαμε στη παράγραφο 5.1.3 κάνοντας χρήση της μεθόδου αύξησης δεδομένων και χρησιμοποιώντας 5.000 επαναλήψεις με περίοδο συγκλισης τις 500 πρώτες επαναλήψεις, αλλά κρατώντας μόνο τις μεταβλητές x_1 και x_3 τα αποτελέσματα συνοψίζονται στον πίνακα 6.4:

Πίνακας 6.4: Αποτελέσματα αλγορίθμου αύξησης δεδομένων

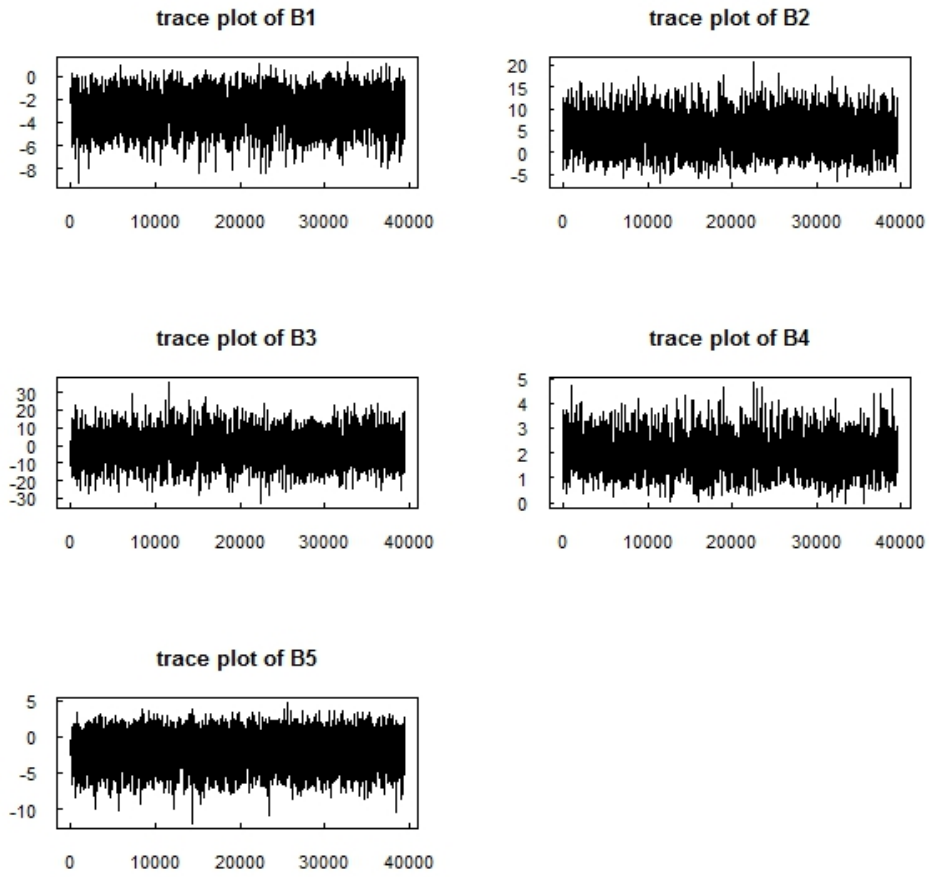
$coefficients$	$posterior\ mean$	$posterior\ std$
$\beta_1(intercept)$	-3.223574	0.8845349
$\beta_2(x_1 = \frac{CF}{TD})$	3.979249	1.5395127
$\beta_4(x_3 = \frac{CA}{CL})$	1.643386	0.4599705

Επιπρόσθετα τα αποτελέσματα του παραπάνω αλγορίθμου για τα β φαίνονται στα γραφήματα 6.3 και 6.4. Παρατηρώντας τα ιστογράμματα (γράφημα 6.3) των εκ των υστέρων κατανομών γίνεται φανερό ότι το μηδέν δε βρίσκεται κοντά στο κέντρο της κατανομής των παραμέτρων. Από αυτό μπορούμε να ισχυριστούμε με ασφάλεια ότι τις παραμέτρους αυτές δεν μπορούμε να τις παραλήψουμε. Στο γράφημα *traceplot* (γράφημα 6.4) μπορούμε να παρατηρήσουμε ότι η σύγκλιση φαίνεται να έχει επιτευχθεί μετά από περίοδιο ζεστάματος ($Burn - in$) τις 500 πρώτες παρατηρήσεις.

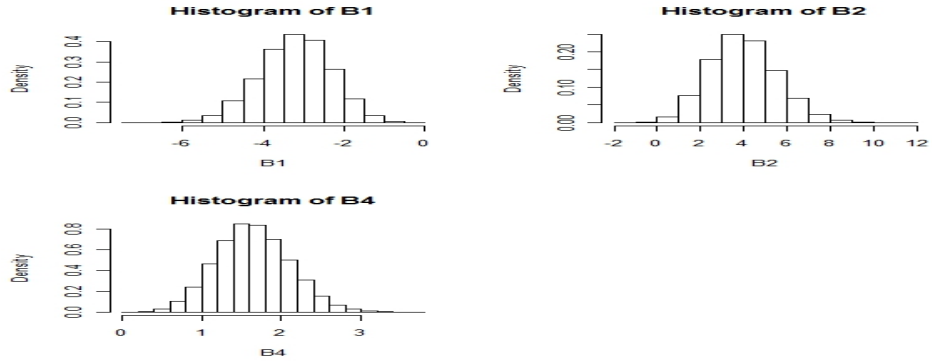
Ερμηνεύοντας τα αποτελέσματα του πίνακα 6.4 γίνεται φανερό ότι υπάρχει θετική συσχέτιση ανάμεσα στην πιθανότητα μη εμφάνισης χρεωκοπίας της επιχείρησης και στη μεταβλητή x_1 όπως και στη μεταβλητή x_3. Αυτό σημαίνει ότι η αύξηση της μεταβλητής x_3, κατά μία μονάδα αντιστοιχεί σε αύξηση του $log - odds$ της πιθανότητας να είναι υγιής η επιχείρηση κατά 1.643386. Αντίστοιχα η αύξηση της μεταβλητής x_1, κατά μία μονάδα αντιστοιχεί σε αύξηση του $log - odds$ της πιθανότητας να είναι υγιής η επιχείρηση κατά 3.979249.



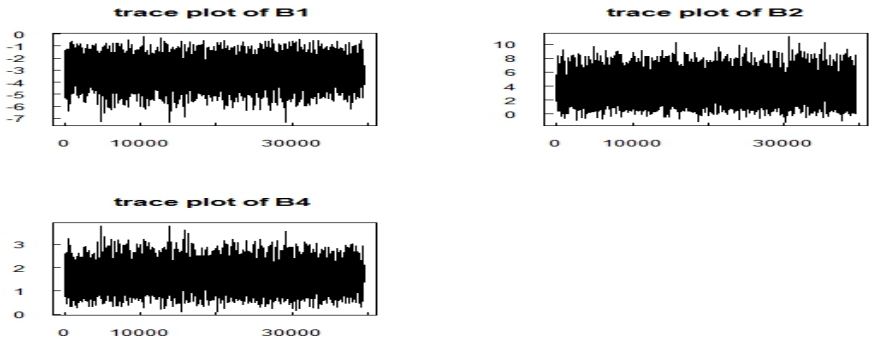
Σχήμα 6.1: *Ιστογράμματα της probit Παλινδρόμησης με data augmentation αλγόριθμο για τα* β



Σχήμα 6.2: *Γράφημα της probit Παλινδρόμησης με data augmentation αλγόριθμο*



Σχήμα 6.3: *Ιστογράμματα της probit Παλινδρόμησης με data augmentation αλγόριθμο για τα β*



Σχήμα 6.4: *Γράφημα της probit Παλινδρόμησης με data augmentation αλγόριθμο*

Πίνακας 6.5: *Bankruptcy Data*

Row	$firm$ $i=0$	x_1	x_2	x_3	x_4	*Row*	$firm$ $i=1$	x_1	x_2	x_3	x_4
1	0	-0.45	-0.41	1.09	0.45	1	1	0.51	0.10	2.49	0.54
2	0	-0.56	-0.31	1.51	0.16	2	1	0.08	0.02	2.01	0.53
3	0	0.06	0.02	1.01	0.40	3	1	0.38	0.11	3.27	0.35
4	0	-0.07	-0.09	1.45	0.26	4	1	0.19	0.05	2.25	0.33
5	0	-0.10	-0.09	1.56	0.67	5	1	0.32	0.07	4.24	0.63
6	0	-0.14	-0.07	0.71	0.28	6	1	0.31	0.05	4.45	0.69
7	0	0.04	0.01	1.50	0.71	7	1	0.12	0.05	2.52	0.69
8	0	-0.07	-0.06	1.37	0.40	8	1	-0.02	0.02	2.05	0.35
9	0	0.07	-0.01	1.37	0.34	9	1	0.22	0.08	2.35	0.40
10	0	-0.14	-0.14	1.42	0.43	10	1	0.17	0.07	1.80	0.52
11	0	-0.23	-0.30	0.33	0.18	11	1	0.15	0.05	2.17	0.55
12	0	0.07	0.02	1.31	0.25	12	1	-0.10	-0.01	2.50	0.58
13	0	0.01	0.00	2.15	0.70	13	1	0.14	-0.03	0.46	0.26
14	0	-0.28	-0.23	1.19	0.66	14	1	0.14	0.07	2.61	0.52
15	0	0.15	0.05	1.88	0.27	15	1	0.15	0.06	2.23	0.56
16	0	0.37	0.11	1.99	0.38	16	1	0.16	0.05	2.31	0.20
17	0	-0.08	-0.08	1.51	0.42	17	1	0.29	0.06	1.84	0.38
18	0	0.05	0.03	1.68	0.95	18	1	0.54	0.11	2.33	0.48
19	0	0.01	0.00	1.26	0.60	19	1	-0.33	-0.09	3.01	0.47
20	0	0.12	0.11	1.14	0.17	20	1	0.48	0.09	1.24	0.18
21	0	-0.28	-0.27	1.27	0.51	21	1	0.56	0.11	4.29	0.44
						22	1	0.20	0.08	1.99	0.30
						23	1	0.47	0.14	2.92	0.45
						24	1	0.17	0.04	2.45	0.14
					87	25	1	0.58	0.04	5.06	0.13

Βιβλιογραφία

[1] Nicholas Metropolis, S. Ulam (1949). The Monte Carlo Method. Journal of the American Statistical Association. (Vol. 44, No. 247., pp. 335-341).

[2] Metropolis, N. and etal (1953). Equations of state calculations by fast computational machine. Journal of Chemical Physics. (21:1087-1091).

[3] Barker A.A. (1965). Monte Carlo calculations of the radial distribution functions for a proton-electron plasma. Aust. J. Phys. (18,119-133).

[4] W. K. Hastings (1970). Monte Carlo Sampling Methods Using Markov Chains and Their Applications. Biometrika. (Vol. 57, No. 1., pp. 97-109).

[5] John Nelder Robert Weddeburn (1972). Generalized Linear Models. Journal of the Royal Statistical Society.(Series A (General) (Blackwell Publishing) 135 (3): 370•384).

[6] P. H. Peskun (1973). Optimum Monte-Carlo Sampling Using Markov Chains. Biometrika. (Vol. 60, No. 3, pp. 607-612)

[7] Stuart Geman, and Donald Geman (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. IEEE Transactions on Pattern Analysis and Machine Intelligence. (6(6)721-741).

[8] Martin A. Tanner, Wing Hung Wong (1987). The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association. (Vol. 82, No. 398., pp. 528-540).

[9] McCullagh, P. and Nelder, J.A. (1989). Generalised Linear Models, 2nd edition. Chapman Hall, London.

[10] Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. Journal of the American Statistical Association. (85, 972-85).

[11] Alan E. Gelfand, Adrian F. M. Smith (1990).Sampling-Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association. (Vol. 85, No. 410, pp. 398-409).

[12] Gelfand, A. E., Smith, A. F. M., and Lee, T.-M. (1992). Bayesian analysis of constrained parameter and trucated data problems using Gibbs sampling. Journal of the American Statistical Association.(87, 523-32).

[13] Andrew Gelman and Donald B. Rubin (1992).Inference from Iterative Simulation Using Multiple Sequences. Statist. Sci. (Volume 7, Number 4, 457-472).

[14] W. R. Gilks and P. Wild (1992). Adaptive Rejection Sampling for Gibbs Sampling. Journal of the Royal Statistical Society. (Series C (Applied Statistics), Vol. 41, No. 2, pp. 337-348)

[15] Delaportas, P., and Smith, A.F.M. (1993). Bayesian inference for generalized linear and proportional hazards models via Gibbs sampling. Appl. Statist. (42: 443-459).

[16] Albert, J. Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. Journal of the American Statistical Association (88, 669-679).

[17] Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. Journal of the American Statistical Association. (88(421): 9-25).

[18] Peter J. Green (1995). Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. Biometrika. (Vol. 82, No. 4 , pp. 711-732).

[19] Fahrmeir L., Tutz G. (2001). Multivariate Statistical Modelling Based on Generalised Linear Models. Springer, Berlin.

[20] Richard A. Johnson and Dean W. Wichern, Prentice Hall, (2002). Applied Mulivariate Statistical Analysis, 5th Edition.

[21] Meligkotsidou, L. (2006). Computationally Intensive Methods II.

[22] Meligkotsidou, L. (2008). Markov Chain Monde Carlo Methods in Bayesian Inference.

[23] Ntzoufras, I. (2008). Bayesian Modeling Using WinBUGS. Wiley, New Jersey.

[24] Chu-Yu Chung, Yi Su, Xiangmin Zhang. Bayesian Generalized Regression for Bankruptcy Prediction.

[25] Ανδρομάχη Σκουφά (2008). Λογιστική Παλινδρόμηση.