ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ & ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

Φωτίου Διαμάντω

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Μπεϋζιανή επιλογή μεταβλητών με χρήση g-prior στα κανονικά γραμμικά μοντέλα»

Τριμελής Επιτροπή

Επιβλέπων: Φουσκάκης Δημήτριος, Αναπληρωτής Καθηγητής, ΕΜΠ

Λουλάκης Μιχαήλ, Αναπληρωτής Καθηγητής, ΕΜΠ

Ντζούφρας Ιωάννης, Καθηγητής, ΟΠΑ

Αθήνα, Οκτώβριος 2017

ΔΙΑΤΜΗΜΑΤΙΚΟ ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

ΚΑΤΕΥΘΥΝΣΗ

ΣΤΑΤΙΣΤΙΚΗ & ΠΙΘΑΝΟΤΗΤΕΣ

«Μπεϋζιανή επιλογή μεταβλητών με χρήση g-prior στα κανονικά γραμμικά μοντέλα»

Φωτίου Διαμάντω

Επιβλέπων: Δημήτριος Φουσκάκης

Διπλωματική εργασία που υποβλήθηκε στη σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών του Εθνικού Μετσόβιου Πολυτεχνείου, ως μέρος των απαιτήσεων για την απόκτηση του μεταπτυχιακού διπλώματος στις Εφαρμοσμένες Μαθηματικές Επιστήμες.

Αθήνα, Οκτώβριος 2017

Αφιερώνεται

στο Νικόλαο Πούλιο, τη μητέρα μου

και τον αδερφό μου

ΕΥΧΑΡΙΣΤΙΕΣ

Στα πλαίσια απόκτησης του μεταπτυχιακού διπλώματος στις Εφαρμοσμένες Μαθηματικές Επιστήμες που απονέμει η σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών του Εθνικού Μετσόβιου Πολυτεχνείου, θα ήθελα αρχικά να ευχαριστήσω τον επιβλέποντά μου κο Δημήτριο Φουσκάκη, Αναπληρωτή Καθηγητή, ΕΜΠ για την επιστημονική και πνευματική υποστήριξη που μου παρείχε καθ' όλη τη διάρκεια των προπτυχιακών και μεταπτυχιακών μου σπουδών καθώς επίσης και τα μέλη της τριμελούς μου επιτροπής κο Μιχαήλ Λουλάκη, Αναπληρωτή Καθηγητή, ΕΜΠ και κο Ιωάννη Ντζούφρα, Καθηγητή, ΟΠΑ.

Στη συνέχεια, θα ήθελα να ευχαριστήσω όσους από τους καθηγητές μου επηρέασαν θετικά τη συνολική μου πορεία με την αξιόλογη προσπάθεια και βοήθειά τους εκφράζοντας την επιθυμία μου να έχουν ένα υγιές, λαμπρό και δημιουργικό μέλλον.

Ιδιαιτέρως θα ήθελα να ευχαριστήσω το αγόρι μου Νικόλαο Πούλιο για την ψυχολογική υποστήριξη που μου παρείχε κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας αλλά και τη συνεχή στήριξή του στην παρούσα φάση τη ζωή μου, όπως επίσης τη μητέρα μου για τον αγώνα της να με μεγαλώσει, τον αδερφό μου για τα όμορφα χρόνια που περάσαμε μαζί, ευχόμενη να είναι πάντα καλά με την οικογένειά του καθώς επίσης και τις φίλες με τις οποίες περάσαμε μαζί όλες τις δυσκολίες των φοιτητικών μας χρόνων.

Τέλος θα ήθελα να εκφράσω την ευγνωμοσύνη μου στη γιαγιά μου για όλα όσα έκανε για μένα. Η δυναμικότητα του χαρακτήρα και η καλοσύνη της την έκαναν ένα ξεχωριστό άνθρωπο που θα ζει για πάντα στο μυαλό και στην καρδιά μου.

Αθήνα, Οκτώβριος 2017
Φωτίου Διαμάντω

Περιεχόμενα

Κεφάλαιο 3

Μπεϋζιανές μέθοδοι στην επιλογή μοντέλου και μεταβλητών

Κεφάλαιο 4

Μίξεις g-εκ των προτέρων κατανομών στην επιλογή μοντέλου και μεταβλητών

Κεφάλαιο 5

Εφαρμογές στην R με χρήση του πακέτου BAS για επιλογή μοντέλου και μεταβλητών με g-priors

Κεφάλαιο 6

ΣΥΝΟΨΗ ΠΕΡΙΕΧΟΜΕΝΩΝ

Το υλικό της παρούσης διπλωματικής εργασίας κατανέμεται σε 6 κεφάλαια εκ των οποίων τα 2 πρώτα περιλαμβάνουν καθαρά εισαγωγικές έννοιες επάνω στη θεωρία των γραμμικών μοντέλων και των θεμελιωδών αρχών της Μπεϋζιανής στατιστικής ούτως ώστε να είναι αποσαφηνισμένη κάθε βασική ορολογία που χρησιμοποιείται στα επόμενα κεφάλαια.

Συγκεκριμένα στο **Κεφάλαιο 1** γίνεται μια συνοπτική εισαγωγή στην απλή και την πολλαπλή γραμμική παλινδρόμηση ενώ στο **Κεφάλαιο 2** παρατίθεται το θεώρημα του Bayes και οι σημαντικότερες κατηγορίες εκ των προτέρων κατανομών που χρησιμοποιούνται στη Μπεϋζιανή στατιστική, με επισήμανση στη σπουδαιότητά τους.

Ακολουθούν το κύριο περιεχόμενο και οι κύριοι στόχοι της παρούσας εργασίας που εμπεριέχονται στα 3 τελευταία κεφάλαια όπου συγκεκριμένα:

Στο **Κεφαλαιο 3** ορίζεται το πρόβλημα της επιλογής μεταβλητών στα κανονικά γραμμικά μοντέλα σύμφωνα με τη Μπεϋζιανή προσέγγιση που θα υιοθετηθεί, ενώ επιπρόσθετα ορίζεται ο βασικός έλεγχος υποθέσεων που θα μελετηθεί για τη σύγκριση των 2^p δυνατών εμφωλευμένων μοντέλων, ενός κανονικού γραμμικού μοντέλου με p πεξηγηματικές μεταβλητές, καθώς και ένα μέτρο σύγκρισης αυτών που διατίθεται μέσω του παράγοντα Bayes. Επιπλέον ορίζονται δύο παράδοξα στα οποία οδηγούμαστε με χρήση διακεχυμένων (εννοώντας με μεγάλη διασπορά) εκ των προτέρων κατανομών:

- το παράδοξο των Lindley-Bartlett και
- το παράδοξο πληροφορίας,

διατυπώνοντας παράλληλα κάποια βασικά πλεονεκτήματα και μειονεκτήματα των Μπεϋζιανών μεθόδων στην επιλογή μεταβλητών.

Ακολούθως στο **Κεφάλαιο 4** μελετώνται διεξοδικά συγκεκριμένες εκ των προτέρων κατανομές των άγνωστων παραμέτρων που οδηγούν σε κλειστές μορφές του παράγοντα Bayes (Feng Liang et al. (2008)), όπως είναι η g εκ των προτέρων κατανομή του Zellner, η hyper-g εκ των προτέρων κατανομή,

η Zellner and Siow εκ των προτέρων κατανομή και η hyper-g/n.

Εν συνεχεία στο **Κεφάλαιο 5** παρουσιάζονται 3 εφαρμογές των μεθόδων που αναπτύσσονται στα κεφάλαια 3 και 4. Η μία υλοποιείται σε πραγματικά δεδομένα και οι άλλες δύο σε προσομοιωμένα δεδομένα με ανεξάρτητες και συσχετιζόμενες επεξηγηματικές μεταβλητές αντίστοιχα, ούτως ώστε να δούμε κατά πόσο ουσιαστική είναι η επίδραση των εκ των προτέρων κατανομών που θα χρησιμοποιηθούν στους συντελεστές ενός κανονικού γραμμικού μοντέλου στα εκ των υστέρων αποτελέσματα και κατά πόσο αυτή η επίδραση συμβάλλει στην ανίχνευση του πραγματικού μοντέλου μεταβλητών που είναι και το κύριο μέλημα της παρούσης διπλωματικής εργασίας.

Τέλος στο **Κεφάλαιο 6** παρουσιάζονται τα εξαγώμενα συμπεράσματα από την εφαρμογή της προτεινόμενης Μπεϋζιανής προσέγγισης στο πρόβλημα της επιλογής μεταβλητών στα κανονικά γραμμικά μοντέλα και στο παράρτημα ο κώδικας R που χρησιμοποιήθηκε για την υλοποίηση των παραδειγμάτων του Κεφαλαίου 5.

ΕΙΣΑΓΩΓΗ

Αρχικά θα εισαχθούμε στο πλαίσιο στο οποίο κινείται η Μπεϋζιανή θεωρία δίνοντας μια αναλυτική περιγραφή ευρέως διαδεδομένων μεθόδων που χρησιμοποιούνται σε ένα από τα κυριότερα πολυπαραμετρικά προβλήματα της στατιστικής, εκείνο που αφορά την επιλογή μεταβλητών στα κανονικά γραμμικά μοντέλα.

Ως απαραίτητα εργαλεία θα χρησιμοποιήσουμε έννοιες των πιθανοτήτων και της στατιστικής, καθώς τις περισσότερες φορές τα δεδομένα που συλλέγουμε από διάφορα φυσικά φαινόμενα περιλαμβάνουν αβεβαιότητα την οποία δεν μπορούμε να περιγράψουμε με απλά ντετερμινιστικά μοντέλα, με αποτέλεσμα να πρέπει να καταφύγουμε σε στατιστικές μεθόδους και μοντέλα με πιθανοκρατική περιγραφή προκειμένου να λύσουμε προβλήματα βέλτιστης επιλογής, εκτίμησης και πρόβλεψης.

Εξίσου απαραίτητο εργαλείο που θα χρησιμοποιήσουμε, είναι η ανάλυση των δεδομένων ως το αντικείμενο εκείνο που συνδυάζει έννοιες και ιδιότητες από τη θεωρία των πιθανοτήτων με τεχνικές της στατιστικής, με κύριο στόχο τη σύνοψη της πληροφορίας και την εξαγωγή συμπερασμάτων από ένα σύνολο δεδομένων το οποίο μπορεί να αποτελείται από πολλές παρατηρήσεις και να αφορά περισσότερα από ένα μεγέθη.

Είναι γεγονός ότι μια από τις κυριότερες μελέτες που συναντάμε συχνά σε πολλές στατιστικές εφαρμογές είναι η μελέτη της σχέσης δύο ή περισσοτέρων μεταβλητών. Παράδειγμα τέτοιας σχέσης έχουμε στη μελέτη του ύψους και του βάρους μιας ομάδας ανθρώπων, του εισοδήματος και της κατανάλωσης εργαζομένων σε μια εταιρεία κ.λ.π. Το πρώτο πρόβλημα που καλούμαστε να επιλύσουμε είναι να αποφασίσουμε αν υπάρχει σχέση σύνδεσης μεταξύ των εξεταζόμενων μεταβλητών και στη συνέχεια να προσδιορίσουμε τη σχέση αυτή (ισοδύναμα το μοντέλο των μεταβλητών) με βάση ορισμένες παρατηρήσεις.

Είναι προφανές ότι ένας από τους κύριους λόγους που καθιστούν τη μελέτη αυτή σημαντική, είναι ότι τα αποτελέσματά της χρησιμοποιούνται συχνά για προβλέψεις, όπως για παράδειγμα βλέπουμε συχνά ιδιωτικές εταιρείες ή κρατικές μονάδες να χρειάζεται να προβλέψουν μεταβλητές όπως η ζήτηση, τα επιτόκια, ο πληθωρισμός, οι τιμές πρώτων υλών, το εργατικό κόστος και πολλές άλλες.

Για το σκοπό αυτό στην επιστήμη της στατιστικής καταφεύγουμε συχνά στην κατασκευή ενός μοντέλου παλινδρόμησης που να περιγράφει ικανοποιητικά τη

σχέση μεταξύ μιας μεταβλητής απόκρισης Y και μιας ή περισσοτέρων επεξηγηματικών μεταβλητών X_1, X_2, ..., X_p. Προφανώς, η συμβολή όλο και περισσότερων επεξηγηματικών μεταβλητών στο μοντέλο, συνεισφέρει στην καλύτερη πρόβλεψη της μεταβλητής απόκρισης, αλλά για εξοικονόμηση χρόνου και οικονομία, δεν δύναται να συμπεριληφθούν όλες εκείνες οι μεταβλητές, αλλά ένα μέρος αυτών (μη εκ των προτέρων γνωστό), που να επηρεάζει ελαφρώς τις προβλέψεις μας. Άρα αφού αποφανθούμε για τη σχέση σύνδεσης της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών (στην περίπτωσή μας θα θεωρήσουμε αυτή τη σχέση να είναι γραμμική), το δεύτερο πρόβλημα που καλούμαστε να επιλύσουμε είναι αυτό που αφορά το ποιές μεταβλητές είναι απαραίτητες για να συμπεριληφθούν στο μοντέλο.

Ως αποτέλεσμα ενδελεχούς προσπάθειας, έχουν αναπτυχθεί διάφορες τεχνικές και μέθοδοι που αντιμετωπίζουν τα παραπάνω δύο προβλήματα είτε προσεγγίζοντάς τα με κλασική στατιστική είτε εφαρμόζοντας μια μπεϋζιανή προσέγγιση.

Στην παρούσα διπλωματική εργασία θα εξετάσουμε τη γενική θεωρία των γραμμικών μοντέλων με έμφαση στο πρόβλημα της βέλτιστης επιλογής μεταβλητών στα κανονικά γραμμικά μοντέλα, από τη σκοπιά της Μπεϋζιανής θεωρίας, καθώς αυτή μας επιτρέπει να ενσωματώσουμε τις εκ των προτέρων γνώσεις μας για τις άγνωστες παραμέτρους του μοντέλου, οι οποίες οδηγούν σε εκ των υστέρων κατανομές όπου εμπεριέχεται όλη η στατιστική πληροφορία που τις αφορά.

Κύριο μέλημά μας είναι στο σύνολο των 2^p συγκριθέντων εμφωλευμένων μοντέλων (όπου p το πλήθος των επεξηγηματικών μεταβλητών του κανονικού γραμμικού μοντέλου), να δοθεί έμφαση στον υπολογισμό της περιθώριας πιθανοφάνειας του κάθε εμφωλευμένου μοντέλου μιας και αυτή αποτελεί μία χρήσιμη ποσότητα που συμμετέχει στον υπολογισμό ενός μέτρου σύγκρισης των πιθανών μοντέλων, που είναι ο παράγοντας Bayes. Επιθυμούμε λοιπόν η πρότερη γνώση που θα χρησιμοποιήσουμε για τις άγνωστες παραμέτρους του κανονικού γραμμικού μοντέλου (πλήρους μοντέλου), να οδηγεί σε κλειστές μορφές της περιθώριας πιθανοφάνειας των εμφωλευμένων μοντέλων που έχουμε να συγκρίνουμε καθώς και του παράγοντα Bayes.

Δεδομένου λοιπόν ότι θέλουμε να ερμηνεύσουμε τη συμπεριφορά μιας μεταβλητής απόκρισης Y έχοντας στη διάθεσή μας ένα μεγάλο σύνολο επεξηγηματικών μεταβλητών X_1, X_2, ..., προσπαθούμε να δώσουμε απάντηση

στο εύλογο ερώτημα για το ποιό είναι το απαραίτητο υποσύνολο αυτών που πραγματικά επεξηγεί τη μεταβλητή απόκρισης ούτως ώστε να εξοικονομηθεί αφενός χρόνος και οικονομικό όφελος αλλά και αφετέρου η πρόβλεψη της μεταβλητής απόκρισης Y να γίνει με όσο το δυνατόν μεγαλύτερη ακρίβεια.

Επιθυμούμε δηλαδή, οι μεταβλητές που τελικώς θα επιλεχθούν να εξασφαλίζουν καλή προσαρμογή στα δεδομένα και οικονομία.

Κατά καιρούς και σύμφωνα με την κλασική στατιστική έχουν αναπτυχθεί διάφορες τεχνικές και κριτήρια για την αντιμετώπιση του προβλήματος της επιλογής μεταβλητών οι οποίες βασίζονται κατά κύριο λόγο σε στατιστικούς ελέγχους, στο συντελεστή προσδιορισμού R^2, στο κριτήριο Cp-Mallows και σε κριτήρια που αφορούν την πρόσθεση ή την αφαίρεση επεξηγηματικών μεταβλητών ή και τα δύο σε ένα μοντέλο, όπως για παράδειγμα η διαδικασία της διαδοχικής αφαίρεσης (Backward elimination), η διαδικασία της διαδοχικής πρόσθεσης ή προς τα εμπρός επιλογής (Forward selection) και η διαδικασία της κατά βήματα εμπρός-πίσω επιλογής (Stepwise selection). Ευρέως διαδεδομένοι μέθοδοι έχουν αναπτυχθεί επίσης και για την εκτίμηση των άγνωστων παραμέτρων του μοντέλου όπως για παράδειγμα η μέθοδος μεγίστης πιθανοφάνειας και η μέθοδος ελαχίστων τετραγώνων.

Η κύρια διαφορά όμως των μεθόδων που θα ασχοληθούμε στην παρούσα διπλωματική εργασία, έγκειται στο γεγονός ότι αυτές βασίζονται στην απλή ιδέα που έδωσε ο Thomas Bayes (1702 – 1761) ότι η μόνη ικανοποιητική περιγραφή της αβεβαιότητάς μας επιτυγχάνεται μέσω της πιθανότητας. Κατά συνέπεια όσα θα δούμε βασίζονται στο θεώρημα του Bayes όπου σύμφωνα με τη συγκεκριμένη προσέγγιση, οι άγνωστες παράμετροι του μοντέλου μας παύουν να θεωρούνται σταθερές όπως στην κλασική στατιστική αλλά θεωρούνται πλέον τυχαίες μεταβλητές οι οποίες περιγράφονται με πιθανότητες. Μπορούμε επομένως να δώσουμε σε αυτές εκ των προτέρων κατανομές τις οποίες μπορούμε να χρησιμοποιήσουμε και για να υπολογίσουμε την εκ των υστέρων κατανομή τους, οπότε να μπορέσουμε να λάβουμε κάποιες εκτιμήσεις γι αυτές αλλά και για να υπολογίσουμε τις εκ των υστέρων πιθανότητες των συγκριθέντων μοντέλων.

Ενώ λοιπόν η κλασική στατιστική, μας δίνει ένα σημείο εκτίμησης για τις άγνωστες παραμέτρους του μοντέλου μας, η Μπεϋζιανή στατιστική έρχεται να μας δώσει μια ολόκληρη εκ των υστέρων κατανομή η οποία περιγράφει ολοκληρωτικά την αβεβαιότητά μας γύρω από τις άγνωστες παραμέτρους του μοντέλου και μπορεί να χρησιμοποιηθεί για να τις εκτιμήσουμε όσο το

δυνατόν καλύτερα χρησιμοποιώντας είτε το μέσο της εν λόγω κατανομής, είτε τη διάμεσο ή κάποιο ποστημόριο αλλά και διαστήματα εμπιστοσύνης.

Επιπρόσθετα οδηγούμενοι και στον υπολογισμό των εκ των υστέρων κατανομών των μοντέλων (βάση των πρότερων κατανομών που έχουμε θέσει στους συντελεστές) μπορούμε να τα συγκρίνουμε και να καταλήξουμε σε μια βέλτιστη επιλογή μεταβλητών που είναι και το κύριο ζήτημα στο οποίο θα εφιστήσουμε την προσοχή μας.

Στην παρούσα διπλωματική εργασία και συγκεκριμένα στα Κεφάλαια 3 και 4 θα μελετήσουμε διεξοδικά για ένα κανονικό γραμμικό μοντέλο κάποια χρήσιμα εργαλεία σύγκρισης των εμφωλευμένων του μοντέλων που αφορούν:

- Την εκ των υστέρων πιθανότητα του κάθε μοντέλου.
- Το λόγο των εκ των υστέρων πιθανοτήτων των μοντέλων (posterior model odds).
- Το λόγο των περιθώριων πιθανοφανειών των μοντέλων (Bayes Factor).

ΚΕΦΑΛΑΙΟ 1

ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Στο παρόν κεφάλαιο περιλαμβάνεται μια εισαγωγή στη γενική θεωρία των γραμμικών μοντέλων και διατυπώνονται κάποιες βασικές έννοιες και ιδιότητες των μοντέλων της συγκεκριμένης κατηγορίας ούτως ώστε να γίνει εύληπτη η εφαρμογή των μεθόδων της Μπεϋζιανής θεωρίας που θα χρησιμοποιηθούν στα επόμενα κεφάλαια.

1.1 ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ ΚΑΙ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Σε πολλές στατιστικές εφαρμογές, όπως ήδη σημειώθηκε προγενέστερα στην εισαγωγή, είθισται να συναντάμε το πρόβλημα της μελέτης της σχέσης δύο ή περισσοτέρων μεταβλητών, που ανεξάρτητα από τους λόγους για τους οποίους η μελέτη της σχέσης αυτής είναι σημαντική, χρειάζεται να καταφύγουμε στην κατασκευή μιας μαθηματικής εξίσωσης (μοντέλου), που περιγράφει τη φύση της σχέσης που υφίσταται μεταξύ των υπό μελέτη μεταβλητών.

Η διαδικασία κατασκευής μιας μαθηματικής εξίσωσης για την περιγραφή ενός φαινομένου μπορεί να είναι ιδιαίτερα πολύπλοκη και απαιτεί μία εκ των προτέρων γνώση για τη σχέση μεταξύ της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών μιας και το πλήθος των διαφορετικών μαθηματικών μοντέλων που θα μπορούσαν να χρησιμοποιηθούν στα περισσότερα προβλήματα είναι σχεδόν άπειρο.

Σε πολλά ωστόσο πρακτικά προβλήματα, ενδιαφερόμαστε να μελετήσουμε πώς επηρεάζεται η μέση τιμή μιας μεταβλητής απόκρισης Y σε σχέση με την τιμή που έχει πάρει ένα σύνολο επεξηγηματικών μεταβλητών $X_1, X_2,..., X_p$, υποθέτοντας ότι η μέση τιμή της τυχαίας μεταβλητής Y είναι μια γραμμική συνάρτηση των παραμέτρων του μοντέλου.

Με την παραπάνω υπόθεση και για $p = 1$, καταλήγουμε στην απλή γραμμική παλινδρόμηση η οποία εκφράζεται από τη σχέση:

$$Y = \alpha X + \beta + \varepsilon, \qquad (1.1)$$

όπου σύμφωνα με την κλασσική στατιστική οι συντελεστές α, β θεωρούνται σταθερές με $\varepsilon \sim N(0,1)$ να είναι ο τυχαίος όρος που μετρά την απόκλιση της Y από το συστηματικό μέρος του μοντέλου, αφού στην πράξη είναι σχεδόν απίθανο οι τυχαίες μεταβλητές μας να συνδέονται με μία τέλεια γραμμική σχέση.

Γενικεύοντας την (1.1) για $p > 1$ πλήθος επεξηγηματικών μεταβλητών, έχουμε την περίπτωση της πολλαπλής γραμμικής παλινδρόμησης, η οποία εκφράζεται από την σχέση:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \boldsymbol{\varepsilon} \ (1.2),$$

για την οποία λαμβάνοντας τυχαίο δείγμα του πληθυσμού, υποθέτουμε ότι για δεδομένο $\mathbf{X} = \mathbf{x}$, η τυχαία μεταβλητή $Y \sim N(\mu, \sigma^2)$ όπου η μέση τιμή ισούται με:

$$\mu = E(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p,$$

με $x_1, x_2, \ldots, x_k$ τα παρατηρηθέντα δεδομένα και $\boldsymbol{\varepsilon}$ το διάνυσμα των ανεξάρτητων τυχαίων σφαλμάτων με $\varepsilon_i \sim N(0, \sigma^2),\ i = 1, 2, \ldots, n$ όπου n το μέγεθος του τυχαίου δείγματος.

Οι συντελεστές του μοντέλου $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ είναι άγνωστες σταθερές ή τυχαίες μεταβλητές, ανάλογα με την προσέγγιση που επιθυμούμε να κάνουμε και είναι αυτοί που θα μας απασχολήσουν στα επόμενα κεφάλαια αντιμετωπίζοντάς τους ως τυχαίες μεταβλητές κι αναζητώντας να τους δώσουμε κατάλληλες εκ των προτέρων κατανομές ούτως ώστε να πραγματοποιήσουμε τον βασικό έλεγχο υποθέσεων που θα περιγράψουμε αργότερα και που θα χρησιμοποιήσουμε για τη σύγκριση των υποψήφιων μοντέλων.

Για τις σχέσεις (1.1) και (1.2) που αναφέρονται όπως είδαμε στην απλή και την πολλαπλή γραμμική παλινδρόμηση, επισημάνθηκε η βασική υπόθεση της κανονικότητας για τη μεταβλητή απόκρισης δεδομένων των τιμών των επεξηγηματικών μεταβλητών και της κανονικότητας των τυχαίων σφαλμάτων.

Συνοψίζοντας την προηγηθείσα ανάλυση και τα στοιχεία που παρατέθηκαν και δεδομένου ότι η εκτίμηση των παραμέτρων των γραμμικών μοντέλων δεν μπορεί να γίνει αναλυτικά λόγω της πολυπλοκότητάς τους, το πρόβλημα που θα μας απασχολήσει στην παρούσα διπλωματική εργασία επικεντρώνεται στην

αναζήτηση μεθόδων για την πραγματοποίηση ενός ελέγχου υποθέσεων που θα χησιμοποιηθεί για τη σύγκριση των 2^p το πλήθος εμφωλευμένων μοντέλων, αφού για το κανονικό γραμμικό μοντέλο που παρουσιάζεται στη σχέση (1.2) υπάρχουν 2^p το πλήθος δυνατοί τρόποι να επιλέξουμε υποσύνολα των επεξηγηματικών μεταβλητών.

1.2 ΕΙΣΑΓΩΓΗ ΣΤΟ ΠΡΟΒΛΗΜΑ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ ΣΤΑ ΚΑΝΟΝΙΚΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Έχοντας θέσει τα διάφορα ερευνητικά μας ερωτήματα, αναζητούμε την απάντησή τους σε μία κατάλληλη μοντελοποίηση που βασίζεται στις ακόλουθες αρχές:

- Όλα τα μοντέλα είναι λάθος, πάραυτα υπάρχουν μερικά πιο χρήσιμα που περιγράφουν περιληπτικά την πραγματικότητα.
- Υπάρχουν πολλά διαθέσιμα μοντέλα που πρέπει να ελέγξουμε για να κάνουμε τη βέλτιστη επιλογή μεταβλητών (για ένα κανονικό γραμμικό μοντέλο p επεξηγηματικών μεταβλητών έχουμε 2^p το πλήθος εμφωλευμένα μοντέλα προς σύγκριση).
- Το μοντέλο μεταβλητών που θα επιλέξουμε πρέπει να προσαρμόζεται καλά στα δεδομένα μας.

Σκοπός:

Κατά την Μπεϋζιανή προσέγγιση που θα ακολουθήσουμε, πέρα από την κανονικότητα της μεταβλητής απόκρισης δοθέντος των τιμών των επεξηγηματικών μεταβλητών, θεωρούμε επίσης τους συντελεστές $\beta_0, \beta_1, \dots, \beta_p$ του κανονικού γραμμικού μοντέλου, τυχαίες μεταβλητές στις οποίες τοποθετούμε εκ των προτέρων κατανομές με σκοπό τον υπολογισμό των εκ των υστέρων πιθανοτήτων των συγκριθέντων μοντέλων.

Κύριο μέλημα τις παρούσης διπλωματικής εργασίας, είναι να παρουσιάσουμε κατάλληλες εκ των προτέρων κατανομές για τις άγνωστες παραμέτρους του κανονικού γραμμικού μοντέλου, ούτως ώστε να οδηγηθούμε σε εύκολους υπολογισμούς για τις εκ των υστέρων πιθανότητες των συγκριθέντων μοντέλων και σε κλειστές μορφές του παράγοντα Bayes, προκειμένου να αποφανθούμε για την επιλογή του βέλτιστου.

Για το λόγο αυτό στο Κεφάλαιο 2 θα γίνει μια εισαγωγή στο γενικό πλαίσιο στο οποίο κινείται η Μπεϋζιανή στατιστική συμπερασματολογία ξεκινώντας

από το βασικό εργαλείο υπολογισμού της εκ των υστέρων κατανομής που είναι το Θεώρημα του Bayes. Το κεφάλαιο αυτό αποσκοπεί μόνο στην κατανόηση των βασικών θεμελιωδών αρχών της Μπεϋζιανής προσέγγισης εισάγοντάς μας στη γενική φιλοσοφία κάτω από την οποία κινείται και κάνοντας μια συνοπτική παρουσίαση των σημαντικότερων κατηγοριών εκ των προτέρων κατανομών που χρησιμοποιούνται συχνά σε διάφορες πολυπαραμετρικές εφαρμογές.

Συγκεκριμένες επιλογές εκ των προτέρων κατανομών όπως η επιλογή της g εκ των προτέρων κατανομής του Zellner, της hyper-g εκ των προτέρων κατανομής και της hyper-g/n θα μελετηθούν διεξοδικά στο Κεφάλαιο 4.

ΚΕΦΑΛΑΙΟ 2

ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΜΠΕΥΖΙΑΝΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

2.1 ΕΙΣΑΓΩΓΗ

Το παρόν κεφάλαιο αποτελεί μια αφετηρία στην κατανόηση των βασικών αρχών της Μπεϋζιανής προσέγγισης, που καθορίζονται από τον κανόνα του Bayes και μέσω της θεωρίας των πιθανοτήτων (ως βασικό εργαλείο), παρέχουν τη δυνατότητα χειρισμού, ελέγχου και περιγραφής της αβεβαιότητάς μας.

Οι Μπεϋζιανές απόψεις και οι διάφορες μεθοδολογίες που έχουν αναπτυχθεί καθιστούν τη Μπεϋζιανή στατιστική χρήσιμη σε πρακτικό και θεωρητικό επίπεδο, με την αντιμετώπιση της άγνωστης παραμέτρου που αποτελεί αντικείμενο ερευνητικού σκοπού σε πολλά προβλήματα, ως τυχαίας ποσότητας, σε αντίθεση με την κλασσική στατιστική όπου η άγνωστη ποσότητα θεωρείται σταθερή. Η όλη στατιστική συμπερασματολογία για την άγνωστη παράμετρο $\boldsymbol{\theta}$ που προκύπτει από μία Μπεϋζιανή ανάλυση, ενσωματώνεται στην εκ των υστέρων κατανομή, η οποία εξαρτάται από τον καθορισμό της εκ των προτέρων κατανομής, της στηριζόμενης στις εκ των προτέρων γνώσεις και πεποιθήσεις του εκάστοτε στατιστικού και κατά συνέπεια το βασικό ερώτημα που απασχολεί τους περισσότερους ερευνητές είναι ποια ή ποιες εκ των προτέρων κατανομές είναι οι κατάλληλες για την ανάλυση.

Στο παρόν κεφάλαιο, διατυπώνεται το θεώρημα του Bayes στο οποίο ουσιαστικά παρουσιάζεται με συνεχή αναθεώρηση, ο τρόπος με τον οποίο συνδυάζονται τα δεδομένα με τις εκ των προτέρων πεποιθήσεις του στατιστικού προκειμένου να παραχθεί η εκ των υστέρων κατανομή στην οποία εμπεριέχεται όλη η πληροφορία για την άγνωστη παράμετρο. Επιπλέον, μιας και η επιλογή των εκ των προτέρων κατανομών καμιά φορά είναι ιδιαίτερα πολύπλοκη ή και αδύνατη (χαρακτηριστικό παράδειγμα η ύπαρξη, των πληροφοριακών, μη πληροφορικών και κατά Jeffrey's prior), παρουσιάζονται τρόποι να επιλέξουμε κατάλληλες εκ των προτέρων κατανομές (όπως για

παράδειγμα τις συζυγείς εκ των προτέρων κατανομές) ούτως ώστε να διευκολύνονται οι υπολογισμοί στο θεώρημα του Bayes.

Συνοψίζοντας τα παραπάνω, το παρόν Κεφάλαιο εστιάζεται στα ακόλουθα 4 βασικά σημεία:

- Στον καθορισμό της εκ των προτέρων κατανομής $\mathbf{f(\theta)}$.
- Στον καθορισμό του μοντέλου πιθανοφάνειας $\mathbf{f(y|\theta)}$.
- Στον καθορισμό της εκ των υστέρων κατανομής $\mathbf{f(\theta|y)}$.
- Στην διεξαγωγή συμπερασμάτων από την εκ των υστέρων κατανομή.

2.2 ΤΟ ΘΕΩΡΗΜΑ ΤΟΥ BAYES

Το θεώρημα του Bayes, στο αποτέλεσμα του οποίου θα στηρίξουμε την όλη μας συμπερασματολογία για την άγνωστη παράμετρο $\boldsymbol{\theta}$, διατυπώνεται σε όρους τυχαίων μεταβλητών ως εξής:

$$\mathbf{f(\theta|y)} = \frac{\mathbf{f(\theta)f(y|\theta)}}{\int \mathbf{f(\theta)f(y|\theta)}}$$

όπου:

- $\mathbf{y}^{\mathrm{T}} = (y_1, y_2, \dots, y_n)$ το τυχαίο δείγμα.
- $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_m) \in \boldsymbol{\Theta}$, το διάνυσμα των άγνωστων παραμέτρων.
- $\mathbf{f(y|\theta)} = \prod_{i=1}^{n} \mathbf{f(y_i|\theta)}$, η συνάρτηση πιθανοφάνειας, η οποία εκφράζει την πιθανότητα παρατήρησης διαφορετικών y_i κάτω από διαφορετικές τιμές της άγνωστης παραμέτρου $\boldsymbol{\theta}$, με $\mathbf{f(y_i|\theta)}$ να είναι η συνάρτηση πυκνότητας ή μάζας πιθανότητας που περιγράφει την τυχαία μεταβλητή Y_i.
- $\mathbf{f(\theta)}$ η εκ των προτέρων κατανομή για την άγνωστη παράμετρο, η οποία βασίζεται σε πληροφορίες που έχουμε από προηγούμενες έρευνες για την παράμετρο $\boldsymbol{\theta}$, η σε πεποιθήσεις που έχουμε γι αυτή, τη δεδομένη χρονική στιγμή που μελετάμε το πρόβλημα.
- $\mathbf{f(\theta|y)}$ η εκ των υστέρων κατανομή που περικλείει όλη τη στατιστική συμπερασματολογία για την αγνωστη παράμετρο $\boldsymbol{\theta}$ και προκύπτει ως το αποτέλεσμα αναπροσαρμογής των δεδομένων με την εκ των προτέρων γνώση.

Σημειώνουμε ότι στην περίπτωση που η άγνωστη παράμετρος είναι διακριτή, το ολοκλήρωμα στον παρονομαστή αντικαθίσταται από το άθροισμα:

$$\sum_i f(\theta_i)f(y|\theta_i).$$

2.2.1 ΕΝΑΛΛΑΚΤΙΚΗ ΜΟΡΦΗ ΤΟΥ ΘΕΩΡΗΜΑΤΟΣ BAYES

Παρατηρώντας ότι το ολοκλήρωμα του παρονομαστή (σταθερά κανονικοποίησης), στον τύπο του Bayes ειναι μία συνάρτηση μόνο του $\mathbf{y}$ αφού η ολοκλήρωση γίνεται ως προς $\boldsymbol{\theta}$, βλέπουμε ότι η εκ των υστέρων κατανομή είναι ανάλογη του αριθμητή και επομένως μια ισοδύναμη μορφή του θεωρήματος Bayes έχει ως εξής:

$$f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta}), \quad (2.2)$$

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

2.2.2 ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΕΥΚΟΛΟ ΥΠΟΛΟΓΙΣΜΟ ΤΗΣ ΕΚ ΤΩΝ ΥΣΤΕΡΩΝ ΚΑΤΑΝΟΜΗΣ

Σε πολλές περιπτώσεις, ο πλήρης υπολογισμός της εκ των υστέρων κατανομής είναι δύσκολος (λόγω του ολοκληρώματος) και ιδίως όταν η άγνωστη παράμετρος είναι πολυδιάστατη, το πρόβλημα υπολογισμού γίνεται ακόμη πιο περίπλοκο. Για το λόγο αυτό, έχουν προταθεί διάφορες λύσεις όπως η επιλογή των συζυγών εκ των προτέρων κατανομών (conjugate priors), οι ασυμπτωτικές προσεγγίσεις και η στοχαστική προσομοίωση από την εκ των υστέρων κατανομή με τους αλγορίθμους MCMC (Markov Chain Monte Carlo) όπως για παράδειγμα τον αλγόριθμο Metropolis Hastings και τον δειγματολήπτη Gibbs που βασίζονται στην ιδέα ότι οτιδήποτε θέλουμε να μάθουμε από μία κατανομή που στην περίπτωσή μας είναι η εκ των υστέρων κατανομή, μπορεί να επιτευχθεί απλά προσομοιώνοντας τυχαίες τιμές από αυτή.

2.3 ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΚΑΤΑΝΟΜΕΣ

Όπως έχουμε ήδη αναφέρει οι εκ των προτέρων (prior) κατανομές εκφράζουν τις πεποιθήσεις που έχουμε για την άγνωστη παράμετρο, πριν συλλέξουμε τα δεδομένα. Συνεπώς, μια στατιστική ανάλυση μπορεί να χαρακτηριστεί υποκειμενική εάν η επιλογή της εκ των προτέρων κατανομής δεν έχει γίνει με σωστό τρόπο. Μολονότι ωστόσο διαφορετικές εκ των προτέρων κατανομές οδηγούν σε διαφορετικά αποτέλεσματα, εάν έχουν επιλεγεί προσεκτικά, χάνουν την επίδρασή τους καθώς το μέγεθος του δείγματος αυξάνεται και συγκεντρώνονται περισσότερα δεδομένα.

Μια λογική επιλογή εκ των προτέρων κατανομών, γίνεται με τέτοιο τρόπο ώστε αυτές να πληρούν τις ακόλουθες δύο προϋποθέσεις:

- Να συσσωρεύουν τη γνώση και πληροφορία από προηγούμενες έρευνες και τη γνώμη ειδικών (πληροφοριακές prior).
- Να ανήκουν σε κάποια από τις γνωστές οικογένειες κατανομών και να διευκολύνουν τους υπολογισμούς στο θεώρημα του Bayes.

Εάν η πρώτη προϋπόθεση δεν δύναται να ικανοποιηθεί από άγνοια ή έλλειψη πληροφορίας για την άγνωστη παράμετρο, τότε φροντίζουμε η εκ των προτέρων κατανομή που θα χρησιμοποιήσουμε να είναι είναι μη πληροφοριακή, δηλαδή να υπόκειται στην κυριαρχία των δεδομένων για τον υπολογισμό της ύστερης κατανομής και προκειμένου να επιτευχθεί κάτι τέτοιο συνήθως επιλέγουμε διακεχυμένες πρότερες κατανομές (κατανομές με μεγάλη διασπορά).

Μερικές από τις πιο διαδεδομένες κατηγορίες εκ των προτέρων κατανομών είναι οι εξής:

- Οι συζυγείς εκ των προτέρων κατανομές (conjugate priors).
- Η μη πληροφοριακή εκ των προτέρων κατανομή του Jeffrey (Jeffrey's prior).
- Οι Ιεραρχικές εκ των προτέρων κατανομές (hyper priors).
- Οι εκ των προτέρων κατανομές που βασίζονται σε δυνάμεις της πιθανοφάνειας (power priors).

Για το σκοπό της παρούσης διπλωματικής εργασίας, δεν θα επεκταθούμε λεπτομερώς στις παραπάνω ευρέως διαδεδομένες κατηγορίες εκ των προτέρων κατανομών, αλλά πληροφοριακά θα δώσουμε μια απλή και συνοπτική περιγραφή των σημαντικότερων από αυτές τις κατηγορίες, στα πλαίσια μιας

μικρής εισαγωγής στις περιοχές της Μπεϋζιανής θεωρίας. Διεξοδικά θα αναφερθούμε στις ακόλουθες εκ των προτέρων κατανομές: τη g εκ των προτέρων κατανομή του Zellner για γραμμικά μοντέλα, τη hyper-g εκ των προτέρων κατανομή, τη Zellner and Siow εκ των προτέρων κατανομή και τη hyper-g/n.

2.3.1 ΣΥΖΥΓΕΙΣ ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΚΑΤΑΝΟΜΕΣ

Οι συζυγείς εκ των προτέρων κατανομές συνεισφέρουν στην αντιμετώπιση των υπολογιστικών δυσκολιών του ολοκληρώματος του κανόνα του Bayes, καθώς επιλέγονται με τέτοιο τρόπο ώστε δοθέντος ότι ανήκουν σε μια συγκεκριμένη οικογένεια κατανομών να καταλήγουν συνδυαζόμενες με τα δεδομένα σε εκ των υστέρων κατανομές που έχουν την ίδια συναρτησιακή μορφή, δηλαδή ανήκουν και αυτές στην ίδια οικογένεια και έχουν τις ίδιες ιδιότητες.

Οι συζυγείς εκ των προτέρων κατανομές είναι απλές και άμεσες στην κατασκευή υπό την προϋπόθεση ότι το δείγμα αποτελείται από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές και έχουν συχνά εκείνα τα επιθυμητά χαρακτηριστικά που είναι σημαντικά για την απλοποίηση των υπολογισμών, την ανάλυση και την ερμηνεία των δεδομένων. Δεν πρέπει να ξεχνάμε ωστόσο, ότι πρέπει να επιλέγονται εξίσου με τέτοιο τρόπο ώστε να είναι συμβατές με την εκ των προτέρων γνώση και εμπειρία των στατιστικών για το διάνυσμα των άγνωστων παραμέτρων.

2.3.2 ΜΗ ΠΛΗΡΟΦΟΡΙΑΚΕΣ ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΚΑΤΑΝΟΜΕΣ

Όπως έχουμε ήδη αναφέρει, σε περιπτώσεις που δεν έχουμε επαρκή πληροφορία για την άγνωστη παράμετρο $\boldsymbol{\theta}$ και κυρίως σε περιπτώσεις που έχουμε άγνοια γι αυτήν, χρησιμοποιούμε ως εκ των προτέρων κατανομές, κατανομές με μικρή ακρίβεια ούτως ώστε να εξασφαλίσουμε ότι ασυμπτωτικά η εκ των υστέρων κατανομή που θα προκύψει από τον κανόνα του Bayes θα καθορίζεται από τα δεδομένα του δείγματος και όχι από την εκ των προτέρων κατανομή. Δηλαδή επιθυμούμε να επιλέξουμε επίπεδες (flat) ή μη πληροφοριακές (non-informative) εκ των προτέρων κατανομές, εννοώντας ότι θέλουμε οριακά να έχουμε την εξής συμπεριφορά για την ακρίβεια:

$$\tau = \frac{1}{\sigma^2} \to 0$$

όπου σ^2 η διασπορά της εκ των προτέρων κατανομής.

Για να επιτευχθεί όμως η παραπάνω σύγκλιση θα πρέπει η κατανομή που θα επιλεγεί να έχει διασπορά που τείνει στο άπειρο με αποτέλεσμα να παύει να είναι μια καλά ορισμένη κατανομή αφού το ολοκλήρωμά της ως προς την αγνωστη παράμετρο παύει να είναι πεπερασμένο (μη γνήσια κατανομή). Παρ ολα αυτά μπορεί η χρήση μιας τέτοιας μη γνήσιας κατανομής να οδηγήσει σε εκ των υστέρων κατανομή με πεπερασμένο ολοκλήρωμα και γι αυτό δεν θεωρείται απαγορευτική στην Μπεϋζιανή στατιστική. Σε κάθε περίπτωση πρέπει να ειμαστε προσεκτικοί ως προς την ερμηνεία της εκ των υστέρων κατανομής και να ελέγχουμε ότι αυτή είναι καλά ορισμένη.

Γνωστά παραδείγματα μη πληροφοριακών εκ των προτέρων κατανομών αποτελεί η ομοιόμορφη κατανομή και η κατανομή του Jeffreys που ορίζεται ως:

$$f(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{\frac{1}{2}},$$

όπου $I(\boldsymbol{\theta})$ η πληροφορία του Fisher που ορίζεται από τον τύπο:

$$I(\boldsymbol{\theta}) = E_{\mathbf{y}|\boldsymbol{\theta}}\left[\left(\frac{d\log f(\mathbf{y}|\boldsymbol{\theta})}{d\boldsymbol{\theta}}\right)^2\right].$$

2.3.3 ΙΕΡΑΡΧΙΚΕΣ ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΚΑΤΑΝΟΜΕΣ

Υπάρχουν περιπτώσεις όπου η εκ των προτέρων κατανομή τυγχάνει να έχει συναρτησιακή μορφή που εξαρτάται από κάποια άγνωστη παράμετρο φ την οποία καλούμε υπερπαράμετρο. Δηλαδή μπορεί να είμαστε στην περίπτωση όπου το δείγμα μας προέρχεται από μια κατανομή $f(\mathbf{y}|\boldsymbol{\theta})$ και η εκ των προτέρων κατανομή είναι της μορφής $f(\boldsymbol{\theta}|\varphi)$ με φ άγνωστο οπότε απαιτεί τη δική του εκ των προτέρων κατανομή $f(\varphi)$.

Σε αυτή την περίπτωση απορρέουν δυο εκ των υστέρων κατανομές. Η πρώτη δίνεται από τον τύπο των Garlin & Louis 1996:

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y},\boldsymbol{\theta})}{f(\mathbf{y})} = \frac{\int f(\mathbf{y},\boldsymbol{\theta},\varphi)d\varphi}{\iint f(\mathbf{y},\boldsymbol{\theta},\varphi)d\varphi d\boldsymbol{\theta}} = \frac{\int f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}|\varphi)f(\varphi)d\varphi}{\iint f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}|\varphi)f(\varphi)d\varphi d\boldsymbol{\theta}},$$

ενώ η δεύτερη εκ των υστέρων κατανομή δίνεται από τον τύπο Gelman et al. 1995 & Bernardo et al. 1994:

$$f(\varphi|\mathbf{y}) = \frac{f(\mathbf{y}|\varphi)f(\varphi)}{f(\mathbf{y})} = \frac{[\int f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}|\varphi)d\boldsymbol{\theta}]f(\varphi)}{f(y)}$$

$$= \left[\frac{f(\mathbf{y}|\boldsymbol{\theta})f(\boldsymbol{\theta}|\varphi)}{f(\boldsymbol{\theta}|\mathbf{y})}\right]\frac{f(\varphi)}{f(\mathbf{y})} = \frac{f(\mathbf{y}|\varphi,\boldsymbol{\theta})f(\boldsymbol{\theta}|\varphi)f(\varphi)}{f(\boldsymbol{\theta}|\varphi,y)f(\mathbf{y})}$$

$$= \frac{f(\mathbf{y}|\varphi,\boldsymbol{\theta})f(\boldsymbol{\theta},\varphi)}{f(\boldsymbol{\theta}|\varphi,\mathbf{y})f(\mathbf{y})} = \frac{f(\varphi,\boldsymbol{\theta}|\mathbf{y})}{f(\boldsymbol{\theta}|\varphi,\mathbf{y})}. \quad (2.3)$$

Τέλος, οι δύο αυτές εκ των προτέρων κατανομές συνδέονται στη σχέση των Bernardo et al. 1994:

$$f(\boldsymbol{\theta}|\mathbf{y}) = \int f(\boldsymbol{\theta},\varphi|\mathbf{y})d\varphi = \int f(\boldsymbol{\theta}|\varphi,\mathbf{y})f(\mathbf{y}|\varphi)d\varphi. \quad (2.4)$$

Εναλλακτικά η παράμετρος $\boldsymbol{\varphi}$ θα μπορούσε να αντικατασταθεί από μια εκτιμήτριά της $\widehat{\boldsymbol{\varphi}}$ η οποία θα μπορούσε να προκύψει μεγιστοποιώντας ως προς $\boldsymbol{\varphi}$ την κατανομή $f(\mathbf{y}|\varphi)$. Στην περίπτωση αυτή η όλη στατιστική συμπερασματολογία θα βασίζονταν στην εκ των υστέρων κατανομή $f(\boldsymbol{\theta}|\mathbf{y},\widehat{\varphi})$ και μια τέτοια διαδικασία συχνά αναφέρεται ως εμπειρική μπεϋζιανή ανάλυση (empirical Bayes analysis) αφού χρησιμοποιούμε τα δεδομένα για να εκτιμήσουμε την εκ των προτέρων παράμετρο $\boldsymbol{\varphi}$. Η υπερπαράμετρος $\boldsymbol{\varphi}$ κατ' επέκταση θα μπορούσε να εξαρτάται από μια άλλη άγνωστη παράμετρο $\boldsymbol{\psi}$ κι ακολουθώντας την ίδια διαδικασία να καταλήγαμε σε μια ιεραρχική μοντελοποίηση.

2.4 ΣΠΟΥΔΑΙΟΤΗΤΑ ΤΩΝ ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΚΑΤΑΝΟΜΩΝ

Στο παρόν Κεφάλαιο αναφέραμε συνοπτικά κάποιες από τις σημαντικότερες κατηγορίες εκ των προτέρων κατανομών, ειδικές περιπτώσεις των οποίων θα μελετήσουμε διεξοδικά στο Κεφάλαιο 4 καθώς η συμβολή των εκ των προτέρων κατανομών είναι καίριας σημασίας στην επιλογή μεταβλητών στα κανονικά γραμμικά μοντέλα.

Αυτό που πρέπει να επισημάνουμε προκειμένου να κατανοήσουμε το σκοπό της παρούσης διπλωματικής εργασίας είναι η σημαντικότητα των εκ των

προτέρων κατανομών και οι βασικές προϋποθέσεις που πρέπει να πληρούνται κατά τη διαδικασία της επιλογής τους, ως συμπλήρωμα των όσων έχουν ήδη αναφερθεί στην εισαγωγή του Κεφαλαίου. Έτσι λοιπόν συνοψίζουμε τις εξής παρατηρήσεις:

- Στην περίπτωση που η εκ των προτέρων κατανομή δεν είναι εντελώς παράλογη, η επιρροή της γίνεται ολοένα μικρότερη καθώς προστίθενται νέα δεδομένα.
- Από την στιγμή που η εκ των προτέρων κατανομή αντιπροσωπεύει τις πεποιθήσεις μας για την παράμετρο $\boldsymbol{\theta}$ προτού μελετηθούν τα δεδομένα μας, είναι φυσικό ότι η μεταγενέστερη ανάλυση είναι μοναδική για εμάς. Δηλαδή η εκ των προτέρων κατανομή που θέτει κάποιος άλλος, θα οδηγήσει σε διαφορετική μεταγενέστερη εκ των υστέρων συμπερασματολογία. Με αυτή την έννοια η ανάλυση είναι καθαρά υποκειμενική.
- Συχνά έχουμε μια γενική ιδέα για το ποια θα πρέπει να είναι η εκ των προτέρων κατανομή (πιθανότατα να μπορούμε να πούμε ποιος είναι ο μέσος και η διακύμανση της), χωρίς όμως να μπορούμε να είμαστε πιο συγκεκριμένοι για την μορφή της. Σε αυτές τις περιπτώσεις μπορούμε να χρησιμοποιήσουμε μια βολική μορφή της εκ των προτέρων κατανομής η οποία θα είναι το αποτέλεσμα των πεποιθήσεων μας και ταυτόχρονα θα απλοποιήσει τους μαθηματικούς υπολογισμούς.
- Πολλές φορές ο τελικός χρήστης ή ο ειδικός δεν εχει εκ των προτέρων πληροφορία ως προς την τιμή της παραμέτρου $\boldsymbol{\theta}$. Σε αυτές τις περιπτώσεις είναι αρκετά συνηθισμένο να χρησιμοποιούμε μια εκ των προτέρων κατανομή η οποία αντανακλά την άγνοια μας για την άγνωστη παράμετρο, επιλέγοντάς την με μεγάλη διασπορά.

ΚΕΦΑΛΑΙΟ 3

ΜΠΕΥΖΙΑΝΕΣ ΜΕΘΟΔΟΙ ΣΤΗΝ ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ ΚΑΙ ΜΕΤΑΒΛΗΤΩΝ

3.1 ΕΙΣΑΓΩΓΗ

Στο παρόν Κεφάλαιο θα εξετάσουμε το πρόβλημα που αφορά την επιλογή μοντέλου και μεταβλητών από τη σκοπιά της Μπεϋζιανής θεωρίας προκειμένου να παρέχουμε μια πιο ρεαλιστική προσέγγιση του συγκεκριμένου προβλήματος και να αντιμετωπίσουμε κάποια από τα μειονεκτήματα της κλασικής στατιστικής.

Διατυπώνοντας έναν έλεγχο υποθέσεων H_0 με εναλλακτική H_1, οι κλασικοί ερευνητές αποσκοπούν στη σύγκριση ενός μοντέλου και ενός υπό μοντέλου του, υπολογίζοντας κάτω από ένα συγκεκριμένο επίπεδο σημαντικότητας α την p τιμή του ελέγχου. Μικρές τιμές της p τιμής του ελέγχου οδηγούν σε απόρριψη της μηδενικής υπόθεσης ενώ μεγάλες τιμές οδηγούν τον κλασικό ερευνητή στο να μη μπορεί να αποφανθεί για το αν ισχύει ή δεν ισχύει η μηδενική υπόθεση. Ως προς την ερμηνεία της p τιμής, δοθέντος ότι η H_0 είναι αληθής, αν αυτό που παρατηρείται στο δείγμα είναι ακραίο, δηλαδή έχει πολύ μικρή πιθανότητα να συμβεί, τότε απορρίπτεται η μηδενική υπόθεση, ενώ στην αντίθετη περίπτωση, δηλαδή αν αυτό που παρατηρείται στο δείγμα δεν είναι ακραίο ή σπάνιο (όταν η H_0 είναι αληθής) τότε το δείγμα που έχει ληφθεί δεν δίνει αρκετές ενδείξεις για την απόρριψη της μηδενικής υπόθεσης και έχουμε «αποτυχία απόρριψης».

Με την παραπάνω στρατηγική ελλοχεύεται κάποιο «ρίσκο» γιατί και τα ακραία έστω και με πολύ μικρή πιθανότητα μπορεί να συμβούν. Έτσι εάν η μηδενική υπόθεση απορριφθεί μολονότι είναι αληθής, διότι συνέβη κάτι ακραίο (που μπορεί να οφείλεται στην τύχη), ο ερευνητής θα οδηγηθεί στο λεγόμενο σφάλμα τύπου 1 ενώ αν δεν την απορρίψει λανθασμένα, ενώ είναι αληθής η εναλλακτική υπόθεση, τότε θα οδηγηθεί στο λεγόμενο σφάλμα τύπου 2.

Η αποφυγή των παραπάνω μειονεκτημάτων, μπορεί να επιτευχθεί με την Μπεϋζιανή προσέγγιση του αντίστοιχου ελέγχου υποθέσεων τον οποίο θα τον διατυπώσουμε αναλυτικά στο παρόν Κεφάλαιο. Συγκεκριμένα η Μπεϋζιανή προσέγγιση του ελέγχου υποθέσεων είναι πιο απλή και όπως θα δούμε η σύγκριση των μοντέλων πραγματοποιείται με τον παράγοντα Bayes που είναι ο λόγος των περιθώριων πιθανοφανειών των μοντέλων. Αναλυτικότερα θα μελετήσουμε τον παράγοντα Bayes παρακάτω.

3.2 Ο ΒΑΣΙΚΟΣ ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ ΓΙΑ ΤΗ ΣΥΓΚΡΙΣΗ ΜΟΝΤΕΛΩΝ

Με ζητούμενο τη σύγκριση δύο μοντέλων $\mathbf{M_0}$ και $\mathbf{M_1}$ με άγνωστες παραμέτρους $\boldsymbol{\theta_0}$ και $\boldsymbol{\theta_1}$ αντίστοιχα, ο έλεγχος υποθέσεων που έχουμε να πραγματοποιήσουμε διατυπώνεται ως εξής:

$\mathrm{H_0}$: $\mathbf{Y}\sim\mathrm{M_0}$, με πιθανοφάνεια $\mathrm{f}(\mathbf{y}|\mathrm{M_0},\boldsymbol{\theta_0})$ και εκ των προτέρων κατανομή των παραμέτρων $\mathrm{f}(\boldsymbol{\theta_0}|\mathrm{M_0})$,

με εναλλακτική:

$\mathrm{H_1}$: $\mathbf{Y}\sim\mathrm{M_1}$, με πιθανοφάνεια $\mathrm{f}(\mathbf{y}|\mathrm{M_1},\boldsymbol{\theta_1})$ και εκ των ποροτέρων κατανομή των παραμέτρων $\mathrm{f}(\boldsymbol{\theta_1}|\mathrm{M_1})$.

Επιπλέον με τους ακόλουθους συμβολισμούς ορίζουμε:

- Το μοντέλο M.
- Την εκ των προτέρων κατανομή του μοντέλου $\mathrm{f}(\mathrm{M})$.
- Την εκ των υστέρων κατανομή του μοντέλου $\mathrm{f}(\mathrm{M}|\mathbf{y})$.
- Την περιθώρια πιθανοφάνεια του μοντέλου $\mathrm{f}(\mathbf{y}|\mathrm{M})$ (ή αλλιώς εκ των προτέρων προβλεπτική κατανομή του μοντέλου M), η οποία δίνεται από το ολοκλήρωμα:

$$\mathrm{f}(\mathbf{y}|\mathrm{M}) = \int \mathrm{f}(\mathbf{y}|\boldsymbol{\theta}_\mathrm{M},\mathrm{M})\mathrm{f}(\boldsymbol{\theta}_\mathrm{M}|\mathrm{M})\mathrm{d}\boldsymbol{\theta}_\mathrm{M} \qquad (3.2.1)$$

όπου $\boldsymbol{\theta}_\mathrm{M}$ το διάνυσμα των άγνωστων παραμέτρων του μοντέλου M, $\mathrm{f}(\mathbf{y}|\boldsymbol{\theta}_\mathrm{M},\mathrm{M})$ η πιθανοφάνεια του μοντέλου και $\mathrm{f}(\boldsymbol{\theta}_\mathrm{M}|\mathrm{M})$ η εκ των προτέρων κατανομή των άγνωστων παραμέτρων δεδομένου του μοντέλου M.

Το παραπάνω ολοκλήρωμα μπορεί να υπολογιστεί αναλυτικά όταν χρησιμοποιούνται συζυγείς εκ των προτέρων κατανομές αλλά σε μια πληθώρα περιπτώσεων ο υπολογισμός του καθίσταται δύσκολος.

Χρησιμοποιώντας τους παραπάνω συμβολισμούς, μπορούμε να ορίσουμε τον εκ των υστέρων λόγο των πιθανοτήτων των δύο μοντέλων καθώς και τον παράγοντα Bayes προκειμένου να τα συγκρίνουμε (ή ισοδύναμα να κάνουμε τον προαναφερθή έλεγχο υποθέσεων).

3.3 ΕΚ ΤΩΝ ΥΣΤΕΡΩΝ ΛΟΓΟΣ ΠΙΘΑΝΟΤΗΤΩΝ ΚΑΙ ΠΑΡΑΓΟΝΤΑΣ BAYES

Εφαρμόζοντας τον κανόνα του Bayes που περιγράψαμε στο Κεφάλαιο 2, μπορούμε να ορίσουμε τις εκ των υστέρων πιθανότητες των δύο μοντέλων $\mathrm{M_0}$ και $\mathrm{M_1}$ που επιθυμούμε να συγκρίνουμε.

Έτσι έχουμε:

$$\mathrm{f(M_0|y)} = \frac{\mathrm{f(y|M_0)f(M_0)}}{\sum_0^1 \mathrm{f(\,y|M_i)f(M_i)}}$$

και

$$\mathrm{f(M_1|y)} = \frac{\mathrm{f(y|M_1)f(M_1)}}{\sum_0^1 \mathrm{f(y|M_i)f(M_i)}}$$

Διαιρώντας κατά μέλη, προκύπτει **εκ των υστέρων λόγος πιθανοτήτων** των μοντέλων (**posterior model odds**) που δίνεται από τον τύπο:

$$\mathrm{PO_{01}} = \frac{\mathrm{f(M_0|y)}}{\mathrm{f(M_1|y)}} = \frac{\mathrm{f(y|M_0)}}{\mathrm{f(y|M_1)}}\frac{\mathrm{f(M_0)}}{\mathrm{f(M_1)}} \quad (3.3.1).$$

Στον δεύτερο μέλος του τύπου 3.3.1 παρατηρούμε να σχηματίζεται το **πηλίκο των εκ των προτέρων κατανομών** των δύο μοντέλων $\frac{\mathrm{f(M_0)}}{\mathrm{f(M_1)}}$ και το πηλίκο των περιθώριων πιθανοφανειών των δύο μοντέλων που υπολογίζονται σύμφωνα με το ολοκλήρωμα στον τύπο 3.2.1.

Το πηλίκο αυτό καθορίζει **τον παράγοντα Bayes (Bayes factor)**, δηλαδή έχουμε ότι:

$$\mathrm{BF_{01}} = \frac{\mathrm{f(y|M_0)}}{\mathrm{f(y|M_1)}} \quad (3.3.2).$$

Τιμές του παράγοντα Bayes μικρότερες της μονάδας σημαίνουν ότι το μοντέλο $\mathbf{M_1}$ προτιμάται έναντι του μοντέλου $\mathbf{M_0}$ και αντίστροφα όταν οι τιμές αυτές είναι μεγαλύτερες της μονάδος.

Συμβολίζοντας με $\mathbf{BF_{10}}$ τον παράγοντα Bayes του μοντέλου $\mathbf{M_1}$ έναντι του μοντέλου $\mathbf{M_0}$ μια ερμηνεία των τιμών του σύμφωνα με τους Kass & Raftery (1995) παρουσιάζεται στον πίνακα (3.3.1).

Log($\mathbf{BF_{10}}$)	**$\mathbf{BF_{10}}$**	**Ένδειξη εναντίον της $\mathbf{H_0}$**
0 - 1	1 - 3	αμελητέα
1 - 3	3 - 20	θετική
2 - 5	20 - 150	ισχυρή
> 5	> 150	πολύ ισχυρή

Πίνακας 3.3.1 **Ερμηνεία του παράγοντα Bayes του μοντέλου $\mathbf{M_1}$ έναντι του $\mathbf{M_0}$**

Η σύγκριση επομένως δύο μοντέλων κατά τη Μπεϋζιανή θεωρία βασίζεται κατά κύριο λόγο στην εκ των υστέρων πιθανότητά τους (επιλέγοντας ως καλύτερο εκείνο με την μεγαλύτερη εκ των υστέρων πιθανότητα) και στην ερμηνεία των τιμών του παράγοντα Bayes που απαιτεί τον υπολογισμό των περιθώριων πιθανοφανειών των μοντέλων.

Αμέσως στο σημείο αυτό αντιλαμβανόμαστε δύο προβλήματα κατά τον υπολογισμό του παράγοντα Bayes. Το ένα αφορά τον καθορισμό των εκ των προτέρων κατανομών των παραμέτρων και το άλλο τον υπολογισμό της περιθώριας κατανομής σύμφωνα με τον τύπο 3.2.1 που υπολογίζεται αναλυτικά όταν χρησιμοποιούνται συζυγείς εκ των προτέρων κατανομές αλλά σε πολλές περιπτώσεις όπως ήδη έχουμε αναφέρει, καθίσταται δύσκολος.

Όσον αφορά τον καθορισμό διακεχυμένων εκ των προτέρων κατανομών για τις παραμέτρους (δηλαδή εκ των προτέρων κατανομών με μεγάλη διασπορά προκειμένου να δηλώσουμε έλλειψη πληροφορίας), θα δούμε παρακάτω ότι μπορεί να οδηγηθούμε σε παράδοξες συμπεριφορές και πιο συγκεκριμένα θα αναφερθούμε στο παράδοξο των Lindley-Bartlett και το παράδοξο πληροφορίας.

Επιπρόσθετα, θα πρέπει να αποφεύγεται η χρήση μη γνήσιων εκ των προτέρων κατανομών για τις άγνωστες παραμέτρους, διότι όπως θα δούμε παρά το γεγονός ότι δεν παρουσιάζουν πρόβλημα κατά τον υπολογισμό της εκ των υστέρων κατανομής των άγνωστων παραμέτρων, αφού αυτή θα είναι μια γνήσια κατανομή (δηλαδή ολοκληρώσιμη στη μονάδα), οδηγούν σε ακαθόριστο παράγοντα Bayes κατά μια σταθερά κανονικοποίησης.

Λόγω του γεγονότος επομένως ότι πολλές φορές δεν είναι εύκολο ο παράγοντας Bayes να υπολογιστεί σε κλειστή μορφή, έχουν αναπτυχθεί

διάφορες μεθοδολογίες για την προσέγγισή του όπως είναι για παράδειγμα η δειγματοληψία σπουδαιότητας, οι MCMC μέθοδοι και οι Laplace προσεγγίσεις οι οποίες δεν θα αναλυθούν στην παρούσα διπλωματική εργασία. Στο Κεφάλαιο 4 θα μελετήσουμε κλειστές μορφές του παράγοντα Bayes κάτω από την επιλογή της g εκ των προτέρων κατανομής του Zellner και κάτω από τις μίξεις των g εκ των προτέρων κατανομών.

Η σχέση (3.3.2) για τον παράγοντα Bayes βοηθάει στη σύγκριση δύο μοντέλων. Η σύγκριση αυτή μπορεί να επεκταθεί και στην περίπτωση που έχουμε μεγαλύτερο πλήθος εξεταζόμενων μοντέλων. Συγκεκριμένα ας υποθέσουμε ότι έχουμε ως υποψήφια μοντέλα για σύγκριση τα $\mathbf{M_0, M_1, \dots, M_N}$. Μπορούμε να συγκρίνουμε κάθε μοντέλο $\mathbf{M_1, \dots, M_N}$ με το «μηδενικό» μοντέλο $\mathbf{M_0}$ (το οποίο θεωρούμε ως βάση για τη σύγκριση) και έτσι να πάρουμε ξεχωριστά τους παράγοντες Bayes $\mathbf{BF_{10}, BF_{20}, \dots, BF_{N0}}$. Στην περίπτωση αυτή η εκ των υστέρων πιθανότητα για κάθε μοντέλο $\mathbf{M_i, i = 0, 1, \dots, N}$ θα δίνεται ως εξής:

$$\mathbf{f(M_i|y) = \frac{w_i BF_{io}}{\sum_{j=1}^{N} w_j BF_{j0}}, i = 0, \dots, N} \quad \textbf{(3.3.3)},$$

όπου ο όρος $\mathbf{w_i}$ είναι ο εκ των προτέρων λόγος των συμπληρωματικών πιθανοτήτων για το μοντέλο $\mathbf{M_i}$ σε σχέση με το μοντέλο $\mathbf{M_0}$, με $\mathbf{w_0 = BF_{00} = 1}$.

Στο Κεφάλαιο 4 που θα ορίσουμε τη g εκ των προτέρων κατανομή θα δούμε πιο συγκεκριμένα πώς μπορούμε να χρησιμοποιήσουμε τον παράγοντα Bayes για να συγκρίνουμε δύο οποιαδήποτε μοντέλα.

3.4 ΜΠΕΥΖΙΑΝΗ ΣΤΑΘΜΙΣΗ ΜΟΝΤΕΛΩΝ

Οι εκ των υστέρων πιθανότητες των μοντέλων, θα μπορούσαν να χρησιμοποιηθούν ως βάρη προκειμένου να λάβουμε χρήσιμα συμπεράσματα για μια ποσότητα που μας ενδιαφέρει.

Έτσι αν υποθέσουμε ότι Δ είναι η ενδιαφερόμενη ποσότητα (π.χ μια μελλοντική πρόβλεψη) και συμβολίσουμε με $\mathbb{M}$ το χώρο με όλα τα πιθανά μοντέλα προς σύγκριση, τότε με τη Μπεϋζιανή στάθμιση μοντέλων (BMA-Bayesian Model Averaging) όπως θα διαπιστώσουμε λαμβάνουμε καλύτερες προβλέψεις από άλλες μεθόδους που βασίζονται σε μεμονωμένα μοντέλα.

Κάτω από την Μπεϋζιανή στάθμιση μοντέλων, χρησιμοποιώντας όλα τα μοντέλα $M \in \mathbb{M}$ μπορούμε να υπολογίσουμε την εκ των υστέρων κατανομή της ενδιαφερόμενης ποσότητας δοθέντος των παρατηρήσεων $\mathbf{y}$ ως:

$$f(\Delta|\mathbf{y}) = \sum_{M\in \mathbf{M}} f(M|\mathbf{y}) f(\Delta|M,\mathbf{y})$$

όπου $f(M|\mathbf{y})$ η εκ των υστέρων πιθανότητα του μοντέλου M και $f(\Delta|M,\mathbf{y})$ η εκ των υστέρων κατανομή της ενδιαφερόμενης ποσότητας Δ στο μοντέλο M.

Η προβλεπτική ικανότητα (logarithmic scoring) για κάθε μοντέλο ξεχωριστά μπορεί να μετρηθεί από την ποσότητα:

$$LS_M = -E\{\log[f(\Delta|M,\mathbf{y})]\},$$

η οποία είναι μεγαλύτερη ή ίση της αντίστοιχης ποσότητας:

$$LS = -E\{\log[f(\Delta|\,\mathbf{y})]\}$$

που μετρά την προβλεπτική ικανότητα του οποιοδήποτε μοντέλου κάτω από τη Μπεϋζιανή στάθμιση μοντέλων.

3.5 ΤΟ ΠΑΡΑΔΟΞΟ ΤΩΝ LINDLEY-BARTLETT

Έστω ότι ενδιαφερόμαστε να κάνουμε τον ακόλουθο έλεγχο υποθέσεων:

$$H_0 : Y_i \sim N(\theta_0, \sigma^2) \text{ για } \theta_0, \sigma^2 \text{ γνωστά}$$

με εναλλακτική

$$H_1 : Y_i \sim N(\theta, \sigma^2) \text{ για } \sigma^2 \text{ γνωστό}, i = 1, \dots, n$$

και θ άγνωστη παράμετρο που πρέπει να εκτιμηθεί.

Δηλαδή, έστω ότι υποθέτουμε ότι το μοντέλο M_0 κάτω από τη μηδενική υπόθεση δεν περιέχει άγνωστη παράμετρο, ενώ κάτω από την εναλλακτική περιέχει.

Υποθέτοντας επιπλέον ότι η εκ των προτέρων κατανομή για την άγνωστη παράμετρο δοθέντος του μοντέλου της εναλλακτικής υπόθεσης είναι $\theta|M_1 \sim N(\theta_0, \sigma_\theta^2)$ και δίνοντας στις υποθέσεις H_0 και H_1 τις εκ των προτέρων

πιθανότητες $f(H_0)$ και $f(H_1)$ αντίστοιχα, καταλήγουμε χρησιμοποιώντας τη σχέση 3.3.1, ότι ο εκ των υστέρων λόγος πιθανοτήτων (posterior model odds) θα είναι:

$$PO_{01} = \frac{f(H_0)}{f(H_1)}\sqrt{1+n\frac{\sigma_\theta^2}{\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\theta_0)^2-\sum_{i=1}^{n}(y-\bar{y})^2-\frac{n(\bar{y}-\theta_0)^2}{1+n\frac{\sigma_\theta^2}{\sigma^2}}\right]\right\} \quad (3.5.1).$$

Απόδειξη:

Επειδή κάτω από την H_0 δεν περιέχεται άγνωστη παράμετρος $\boldsymbol{\theta}$, η περιθώρια πιθανοφάνεια υπό του μοντέλου M_0 θα ταυτίζεται με την πιθανοφάνεια του μοντέλου για κανονικά δεδομένα και συνεπώς θα έχουμε ότι:

$$f(\mathbf{y}|M_0) = (2\pi\sigma^2)^{-n/2}e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta_0)^2},$$

επιπλέον αποδεικνύεται ότι η περιθώρια πιθανοφάνεια υπό του μοντέλου M_1 ισούται με:

$$f(\mathbf{y}|M_1) = \int (2\pi\sigma^2)^{-\frac{n}{2}}e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2}f(\theta)d\theta$$

$$= \int (2\pi\sigma^2)^{-\frac{n}{2}}e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2}\left(2\pi\sigma_\theta^2\right)^{-\frac{1}{2}}e^{-\frac{1}{2\sigma_\theta{}^2}(\theta-\theta_0)^2}d\theta$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}}\left(\frac{\sigma^2}{n\sigma_\theta{}^2+\sigma^2}\right)^{\frac{1}{2}}e^{\frac{1}{2\sigma_\theta{}^2}\left[\sum_{i=1}^{n}(y_i-\bar{y})^2+\frac{n(\bar{y}-\theta_0)^2}{1+\frac{n\sigma_\theta{}^2}{\sigma^2}}\right]}.$$

Αντικαθιστώντας όπως αναφέραμε τα παραπάνω στη σχέση 3.3.1 και δεδομένου ότι εκ των προτέρων κατανομή του μοντέλου M_0 κάτω από τη μηδενική υπόθεση είναι $f(H_0)$ και του μοντέλου M_1, $f(H_1)$ αντίστοιχα καταλήγουμε στη σχέση 3.5.1.

Δεδομένου επιπλέον ότι το τυχαίο δείγμα ακολουθεί κανονική κατανομή με μέση τιμή θ_0 και διασπορά σ^2, σε επίπεδο σημαντικότητας $\alpha = q$, ο

δειγματικός μέσος θα είναι $\bar{y} = \theta_0 \pm Z_{q/2}\, {}^{\sigma}/_{\sqrt{n}}$, οπότε αντικαθιστώντας στη σχέση 3.5.1 λαμβάνουμε ότι:

$$PO_{01} = \frac{f(H_0)}{f(H_1)}\sqrt{1 + n\frac{\sigma_\theta^2}{\sigma^2}}\exp\left\{-\frac{1}{2}\frac{n\sigma_\theta^2}{n\sigma_\theta^2 + \sigma^2}Z_{q/2}^2\right\} \quad (3.5.2).$$

Παρατηρήσεις:

Βασιζόμενος στον τύπο (3.5.2), ο Lindley (1957) παρατήρησε ότι ο λόγος των εκ των υστέρων συμπληρωματικών πιθανοτήτων εξαρτάται από το μέγεθος του δείγματος n και αυξάνεται όσο αυτό μεγαλώνει, δηλαδή για $n \to \infty \Rightarrow PO_{01} \to \infty$. Επιπλέον ο Bartlett (1957) παρατήρησε ότι όσο μεγαλύτερη είναι η διασπορά της εκ των προτέρων κατανομής τόσο μεγαλύτερος είναι ο παράγοντας Bayes. Στηριζόμενοι και στις δύο αυτές περιπτώσεις είναι φανερό ότι καταλήγουμε στο ίδιο παράδοξο, δηλαδή οδηγούμαστε υπέρ του μοντέλου της αρχικής υπόθεσης.

Υποδείξεις για την αποφυγή του παραδόξου Lindley –Bartlett:

- Η εξάρτηση του εκ των υστέρων λόγου πιθανοτήτων από το μέγεθος του δείγματος μπορεί να περιοριστεί επιλέγοντας για εκ των προτέρων διασπορά του θ, την $\frac{\sigma_\theta^2}{n}$ αντί για την σ_θ^2.
- Ο καθορισμός της διασποράς σ_θ^2 θα πρέπει να γίνεται με τέτοιο τρόπο ώστε στις περιπτώσεις έλλειψης επαρκούς πληροφορίας για την παράμετρο θ:
  - ✓ Να είναι αρκετά μεγάλη ώστε να αποφεύγεται η εκ των προτέρων μεροληψία σε κάθε μοντέλο.
  - ✓ Να είναι τόσο μεγάλη όσο δεν οδηγει στο παράδοξο των Lindley-Bartlett και στην πλήρη στήριξη της αρχικής υπόθεσης.

3.6 ΠΛΕΟΝΕΚΤΗΜΑΤΑ–ΜΕΙΟΝΕΚΤΗΜΑΤΑ ΤΩΝ ΜΠΕΥΖΙΑΝΩΝ ΜΕΘΟΔΩΝ ΣΤΗΝ ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ ΚΑΙ ΜΕΤΑΒΛΗΤΩΝ

Με τη χρήση Μπεϋζιανών μεθόδων για την επιλογή μοντέλου και μεταβλητών έχουμε τα ακόλουθα πλεονεκτήματα:

Α. Πλεονεκτήματα:

- ✓ Αποτελεσματική αναζήτηση του καταλληλότερου μοντέλου με τη χρήση MCMC μεθόδων (καταλληλότερο εκείνο με τη μεγαλύτερη εκ των υστέρων πιθανότητα).
- ✓ Αυτόματη επιλογή του καλύτερου μοντέλου (κατόπιν του καθορισμού του μοντέλου και των μεθόδων εκτίμησης).
- ✓ Καλή ερμηνεία και σύγκριση των εκ των υστέρων πιθανοτήτων των μοντέλων για την επιλογή εκείνου με τη μεγαλύτερη τιμή.
- ✓ Βοηθά στην επιλογή μιας κλάσης μοντέλων που προσαρμόζονται εξίσου καλά στα δεδομένα και έχουν εκ των υστέρων πιθανότητα που υπολογίζεται σε κλειστή μορφή.
- ✓ Με τη χρήση του παράγοντα Bayes στηριζόμαστε στην εκ των υστέρων πιθανότητα των συγκρινόμενων μοντέλων. Κατ' αυτόν τον τρόπο δεν έχουμε πάντα ότι το μοντέλο με τις περισσότερες επεξηγηματικές μεταβλητές θα είναι το καλύτερο (όπως συμβαίνει στην κλασική στατιστική με τον υπολογισμό του συντελεστή προσδιορισμού R^2). Υπολογίζοντας τον παράγοντα Bayes είναι δυνατό ένα μοντέλο με λιγότερες επεξηγηματικές μεταβλητές να προτιμηθεί.
- ✓ Ένα μοντέλο επιλέγεται ως πιο πιθανό. Αυτό είναι απλούστερα κατανοητό και δεν δημιουργεί ασαφή συμπεράσματα όπως στην περίπτωση του «δεν απορρίπτουμε την H_0» που συναντάμε στην κλασική στατιστική. Αμφίλογα συμπεράσματα όπως αυτό, δημιουργούν σύγχυση σε άτομα που δεν είναι εξοικειωμένα με στοιχειώδη γνώση στατιστικής και καλό είναι να αποφεύγονται.

Β. Μειονεκτήματα:

Σε πολλές περιπτώσεις το **κύριο μειονέκτημα** των Μπεϋζιανών μεθόδων στην επιλογή μοντέλου και μεταβλητών οφείλεται σε:

- ✓ Έλλειψη επαρκούς πληροφορίας για τις άγνωστες παραμέτρους με αποτέλεσμα τη χρήση μη πληροφοριακών εκ των προτέρων κατανομών που οδηγούν σε μη καλή ερμηνεία του παράγοντα Bayes (Lindley-Bartlett Paradox) και στην ανάγκη ανάπτυξης μεθοδολογιών για τον υπολογισμό του παράγοντα Bayes χωρίς τη χρήση εκ των προτέρων κατανομών. Τέτοιες μεθοδολογίες αποτελούν για παράδειγμα τα κριτήρια BIC (Bayesian Information Criterion ή Swartz Criterion) και AIC (Akaike Information Criterion).

Επιπλέον μειονεκτήματα μπορούν να θεωρηθούν:

- ✓ Η δυσκολία υπολογισμού των ολοκληρωμάτων. Σε πολλές περιπτώσεις δεν καθίσταται δυνατή η χρήση συζυγών εκ των προτέρων κατανομών ούτως ώστε να οδηγηθούμε σε μια κλειστή μορφή για τον υπολογισμό των εκ των υστέρων πιθανοτήτων του μοντέλου. Κατά συνέπεια δημιουργείται η ανάγκη για τον υπολογισμό αυτό να χρησιμοποιηθούν άλλες μέθοδοι όπως ασυμπτωτικές προσεγγίσεις, η μέθοδος Laplace και οι MCMC (Markov Chain Monte Carlo) μέθοδοι που επιτυγχάνουν προσομοιώνοντας τιμές, την κατασκευή μιας μαρκοβιανής αλυσίδας που συγκλίνει στην εκ των υστέρων κατανομή, όπως ο αλγόριθμος Metropolis-Hastings και ο Gibbs Sampler. Οι εν λόγω μέθοδοι μπορούν να χρησιμοποιηθούν και για τον υπολογισμό του παράγοντα Bayes (Kass and Raftery (1995), Ntzoufras (1999)).
- ✓ Η υπολογιστικά ασύμφορη αναζήτηση του βέλτιστου μοντέλου, ειδικά όταν το πλήθος των δυνατών υποψήφιων μοντέλων είναι πολύ μεγάλο.

Στα επόμενα δύο κεφάλαια (Κεφάλαιο 4 και Κεφάλαιο 5 αντίστοιχα) θα εξετάσουμε κάποιες συγκεκριμένες εκ των προτέρων κατανομές που χρησιμοποιούμε για την επιλογή μεταβλητών στα κανονικά γραμμικά μοντέλα και εν συντομία θα αναφερθούμε σε κάποιες μεθοδολογίες που χρησιμοποιούνται για τον υπολογισμό της περιθώριας πιθανοφάνειας, δηλαδή του ολοκληρώματος που παρουσιάσαμε στον τύπο 3.2.1.

ΚΕΦΑΛΑΙΟ 4

ΜΙΞΕΙΣ g - ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΚΑΤΑΝΟΜΩΝ ΣΤΗΝ ΜΠΕΥΖΙΑΝΗ ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ ΚΑΙ ΜΕΤΑΒΛΗΤΩΝ

4.1 ΕΙΣΑΓΩΓΗ

Τα γραμμικά μοντέλα, ο ορισμός των οποίων δόθηκε στο Κεφάλαιο 1, είναι ένα θεμελιώδες εργαλείο που χρησιμοποιούμε συχνά σε ένα μεγάλο εύρος προβλημάτων στην κλασική στατιστική. Ωστόσο η χρήση τους και στην Μπεϋζιανή στατιστική είναι εξίσου απαραίτητη καθώς παρέχουν μια εναλλακτική προσέγγιση αυτών των προβλημάτων ως προς τον ορισμό, την ερμηνεία και την ανάλυσή τους.

Τις περισσότερες φορές όπως αναφέραμε και στο Κεφάλαιο 3, το κύριο μέλημά μας στα γραμμικά μοντέλα είναι να αποφασίσουμε ποιές από τις επεξηγηματικές μεταβλητές θα χρησιμοποιήσουμε ούτως ώστε με το μικρότερο δυνατό κόστος να έχουμε καλή προσαρμογή στα δεδομένα. Για το λόγο αυτό στη Μπεϋζιανή στατιστική είναι σημαντικό να αποφανθούμε ποια εκ των προτέρων κατανομή θα χρησιμοποιήσουμε ώστε να αντιπροσωπεύει καλύτερα τις εκ των προτέρων πεποιθήσεις μας για το μοντέλο και τις άγνωστες παραμέτρους που εμπεριέχονται σε αυτό.

Στο Κεφάλαιο 2 ορίσαμε γενικότερα τις σημαντικότερες κατηγορίες εκ των προτέρων κατανομών οπως τις συζυγείς εκ των προτέρων κατανομές, τις μη πληροφοριακές και τις Ιεραρχικές εκ των προτέρων κατανομές.

Ο λόγος που αναφερθήκαμε στις παραπάνω κατηγορίες είναι διότι η επιλογή των εκ των προτέρων κατανομών για τις άγνωστες παραμέτρους, που στην ουσία βοηθάει στην ενσωμάτωση των εκ των προτέρων πεποιθήσεών μας για αυτές στο μοντέλο, κρύβει όλη την «ομορφιά» της μπεϋζιανής θεωρίας. Σε περίπτωση όμως που έχουμε ελλιπή πληροφορία για τις άγνωστες παραμέτρους του μοντέλου, θα πρέπει να αποφεύγεται η χρήση ακατάλληλων εκ των προτέρων κατανομών καθώς η χρήση μη γνησίων εκ των προτέρων κατανομών οδηγεί σε ακαθόριστο παράγοντα Bayes ενώ η χρήση διακεχυμένων εκ των προτέρων κατανομών μπορεί να οδηγήσει όπως επισημάναμε στην ενότητα 3.5 στο παράδοξο των Lindley-Bartlett.

Η ουσία στο να εφαρμόσει κανείς μια Μπεύζιανή προσέγγιση εντοπίζεται κυρίως στο να μπορέσει να αξιοποιήσει τις εκ των προτέρων πεποιθήσεις του, αλλά να είναι επίσης σε θέση να χειριστεί και την περίπτωση όπου δεν είναι διαθέσιμη καμία πληροφόρηση για τις άγνωστες παραμέτρους.

Στο παράδοξο των Lindley-Bartlett επισημάνθηκε η ευαισθησία των εκ των υστέρων πιθανοτήτων του μοντέλου και του παράγοντα Bayes αφενός στη χρήση εκ των προτέρων κατανομών με μεγάλη διασπορά για τους συντελεστές του μοντέλου και αφετέρου στη χρήση μεγάλου δείγματος. Και στις δύο αυτές περιπτώσεις είδαμε ότι εμφανίζεται η τάση να αναδεικνύεται ως πιο πιθανό από το χώρο των μοντέλων, εκείνο με τις λιγότερες μεταβλητές, γεγονός που οδηγεί σε αντιφάσεις μεταξύ των Μπεϋζιανών και κλασικών ελέγχων σημαντικότητας. Δεδομένου λοιπόν αυτού του παραδόξου, αλλά και δεδομένου ότι πολλές φορές έχουμε ελλιπή εκ των προτέρων γνώση για τους συντελεστές του μοντέλου, είναι φανερό ότι η επιλογή εκ των προτέρων κατανομής για αυτούς είναι μια δύσκολη υπόθεση και απαιτεί ιδιαίτερη προσοχή.

Συγκεκριμένα, στο παρόν Κεφάλαιο θα μελετήσουμε λεπτομερέστερα το θέμα της επιλογής των εκ των προτέρων κατανομών για τις άγνωστες παραμέτρους του μοντέλου, εστιάζοντας κυρίως στη Zellner's g prior, τη Zellner-Siow prior, τη hyper-g prior και τη hyper-$\mathbf{g/n}$ prior (με τις τελευταίες τρεις γνωστές ως μίξεις g εκ των προτέρων κατανομών). Σκοπός των εν λόγω επιλογών είναι η διαχείριση του προβλήματος που αφορά την επιλογή μεταβλητών που αναφέραμε στο Κεφάλαιο 3 και θα το επαναπροσδιορίσουμε συνοπτικά παρακάτω. Επίσης θα μελετήσουμε περιπτώσεις συγκεκριμένης επιλογής τιμών για την παράμετρο g.

4.2 ΤΟ ΠΡΟΒΛΗΜΑ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ ΣΤΑ ΚΑΝΟΝΙΚΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

Υποθέτουμε ότι δοθέντος των τιμών των επεξηγηματικών μεταβλητών, έχουμε ένα μοντέλο παλινδρόμησης με μεταβλητή απόκρισης $\boldsymbol{\Upsilon} = (\mathrm{Y}_1, \dots, \mathrm{Y}_\mathrm{n})^\mathrm{T}$ κανονικής κατανομής με μέση τιμή $\boldsymbol{\mu} = (\mu_1, \dots, \mu_\mathrm{n})^\mathrm{T}$ και πίνακα συνδιακύμανσης $\mathbf{I}_\mathrm{n}\sigma^2$ όπου $\mathbf{I}_\mathrm{n}$ ο μοναδιαίος πίνακας.

Δοθέντος του συνόλου των επεξηγηματικών μεταβλητών $\mathrm{X}_1, \dots, \mathrm{X}_\mathrm{p}$ όπου η μέση τιμή $\boldsymbol{\mu}$ εκφράζεται ως γραμμικός συνδυασμός τους, αναζητούμε το μικρότερο δυνατό υποσύνολο αυτών ούτως ώστε να έχουμε ταυτόχρονα και καλή προσαρμογή στα δεδομένα αλλά και οικονομικό όφελος.

Για την περιγραφή του μοντέλου χρησιμοποιούμε ένα διάνυσμα δεικτών $\boldsymbol{\gamma}$ με p διαστάσεις, όπου για τις συνιστώσες του θέτουμε $\gamma_j = 1$ αν η μεταβλητή X_j συμπεριληφθεί στο σύνολο των επεξηγηματικών μεταβλητών και $\gamma_j = 0$ αν δεν συμπεριληφθεί.

Κατ' αυτόν τον τρόπο η μέση τιμή $\boldsymbol{\mu_\gamma}$ υπό του μοντέλου $\mathbf{M_\gamma}$ θα μπορούσε να εκφραστεί υπό μορφή διανύσματος ως:

$$\mathrm{M}_{\boldsymbol{\gamma}}\colon\ \boldsymbol{\mu}_\gamma\ =\ \mathbf{I_n}\beta_0 + \mathbf{X_\gamma}\boldsymbol{\beta_\gamma}$$

όπου β_0 είναι ένας σταθερός όρος κοινός σε όλα τα μοντέλα, $\boldsymbol{\beta_\gamma}$ ένα $\mathrm{p}_{\boldsymbol{\gamma}}$ - διαστάσεων διάνυσμα από μη μηδενικούς συντελεστές παλινδρόμησης και $\mathbf{X_\gamma}$ ο $\mathrm{n\ x\ p}_{\boldsymbol{\gamma}}$ πίνακας σχεδιασμού για το μοντέλο $\mathrm{M}_{\boldsymbol{\gamma}}$.

Η Μπεϋζιανή προσέγγιση στην επιλογή μοντέλου και μεταβλητών βασίζεται στον καθορισμό των εκ των προτέρων κατανομών γα τις άγνωστες παραμέτρους $\boldsymbol{\theta_\gamma} = \left(\beta_0, \boldsymbol{\beta_\gamma}, \sigma^2\right) \in\ \boldsymbol{\Theta_\gamma}$ για κάθε μοντέλο $\mathrm{M}_{\boldsymbol{\gamma}}$, όπου δοθέντος μιας πρότερης κατανομής $\mathrm{f}(\mathrm{M}_{\boldsymbol{\gamma}})$, την οποία ανανεώνουμε, η εκ των υστέρων πιθανότητα για το καθένα δίνεται ως:

$$\mathrm{f}\left(\mathrm{M}_{\boldsymbol{\gamma}}\middle|\boldsymbol{y}\right) = \frac{\mathrm{f}\left(\mathrm{M}_{\boldsymbol{\gamma}}\right)\mathrm{f}\left(\mathbf{y}\middle|\mathrm{M}_{\boldsymbol{\gamma}}\right)}{\sum_\gamma\ \mathrm{f}\left(\mathrm{M}_{\boldsymbol{\gamma}}\right)\mathrm{f}\left(\mathbf{y}\middle|\mathrm{M}_{\boldsymbol{\gamma}}\right)} \quad (4.2.1).$$

Είναι προφανές ότι για να μπορέσει κανείς να κάνει τον παραπάνω υπολογισμό εφαρμόζοντας μια Μπεϋζιανή προσέγγιση, θα πρέπει πρώτα να υπολογίσει την περιθώρια πιθανοφάνεια του μοντέλου $\mathrm{M}_{\boldsymbol{\gamma}}$ που προκύπτει από την ακόλουθη ολοκλήρωση:

$$\mathrm{f}\left(\mathbf{y}\middle|\mathrm{M}_{\boldsymbol{\gamma}}\right) = \int \mathrm{f}\left(\mathbf{y}\middle|\boldsymbol{\theta_\gamma}, \mathrm{M}_{\boldsymbol{\gamma}}\right)\mathrm{f}\left(\boldsymbol{\theta_\gamma}\middle|\mathrm{M}_{\boldsymbol{\gamma}}\right)\mathrm{d}\boldsymbol{\theta_\gamma}, \qquad \boldsymbol{\theta_\gamma} \in\ \boldsymbol{\Theta_\gamma} \quad (4.2.2),$$

όπου $\mathrm{f}\left(\mathbf{y}\middle|\boldsymbol{\theta_\gamma}, \mathrm{M}_{\boldsymbol{\gamma}}\right)$ είναι η πιθανοφάνεια για το μοντέλο $\mathrm{M}_{\boldsymbol{\gamma}}$ δοθέντος του διανύσματος $\boldsymbol{\theta_\gamma}$, η οποία ορίζεται ως:

$$\mathrm{f}\left(\boldsymbol{y}\middle|\boldsymbol{\theta_\gamma}, \mathrm{M}_{\boldsymbol{\gamma}}\right) = \prod_{i=1}^{n} \mathrm{f}\left(\boldsymbol{y_i}\middle|\boldsymbol{\theta_\gamma}, \mathrm{M}_{\boldsymbol{\gamma}}\right),$$

με $\boldsymbol{y_i}$, $\mathrm{i} = 1, \ldots, \mathrm{n}$ τις παρατηρήσεις της μεταβλητής απόκρισης.

Για τους λόγους που έχουμε ήδη αναφέρει, προτείνεται να χρησιμοποιούνται γνήσιες εκ των προτέρων κατανομές στους συντελεστές κάθε μοντέλου και συγκεκριμένα για τα κανονικά γραμμικά μοντέλα προτείνονται γνήσιες εκ των προτέρων κατανομές που είναι βασισμένες στη συζυγή Κανονική-Γάμμα οικογένεια καθώς οδηγούν σε κλειστό υπολογισμό όλων των περιθώριων πιθανοφανειών. Όλα αυτά θα αναλυθούν ακολούθως εστιάζοντας στην $\mathbf{g}$ εκ των προτέρων κατανομή και στο πώς την επιλέγουμε, μιας και η συγκεκριμένη υιοθετήθηκε από πολλούς λόγω απλότητας, εύκολης ερμηνείας αλλά και απλών υπολογισμών.

4.3 ΕΠΙΛΟΓΗ ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΚΑΤΑΝΟΜΩΝ ΓΙΑ ΤΟΥΣ ΣΥΝΤΕΛΕΣΤΕΣ ΤΟΥ ΜΟΝΤΕΛΟΥ

Έστω ένα μοντέλο $\mathbf{M_\gamma}$ στο χώρο των εξεταζόμενων μοντέλων $\mathbf{M}$, με συντελεστές $\boldsymbol{\beta}_\gamma$. Προκειμένου να ενσωματώσουμε τις εκ των προτέρων πεποιθήσεις μας για τις άγνωστες παραμέτρους του μοντέλου, συνήθως επιλέγουμε ως πρότερη κατανομή την πολυμεταβλητή κανονική κατανομή. Δηλαδή θέτουμε:

$$f(\boldsymbol{\beta}_\gamma|M_\gamma)\sim N\left(\boldsymbol{\mu}_\gamma, \boldsymbol{\Sigma}_{(M_\gamma)}\right), \quad (4.3.1)$$

όπου $\boldsymbol{\mu}_\gamma$ και $\boldsymbol{\Sigma}_{(M_\gamma)}$ είναι αντίστοιχα ο εκ των προτέρων μέσος και ο εκ των προτέρων πίνακας συνδυακύμανσης των συντελεστών του μοντέλου $\mathbf{M_\gamma}$. Για να δηλώσουμε την άγνοιά μας για τους άγνωστες συντελεστές συνήθως επιλέγουμε $\boldsymbol{\mu}_\gamma = 0$ και $\boldsymbol{\Sigma}_{(M_\gamma)} = c^2\mathbf{V}_{(M_\gamma)}$ όπου $\mathbf{V}_{(M_\gamma)}$ είναι ένας πίνακας που καθορίζει τον εκ των προτέρων συσχετισμό των συντελεστών $\boldsymbol{\beta}_\gamma$ και c^2 μια σταθερά που αν επιλεγεί μικρή, η εκ των προτέρων κατανομή γίνεται πληροφοριακή ενώ αν επιλεγεί μεγάλη μπορεί να οδηγήσει στο παράδοξο των Lindley-Bartlett. Επομένως ο καθορισμός του $c^2\mathbf{V}_{(M_\gamma)}$ θα πρέπει να γίνεται με προσοχή.

4.4 ΚΑΝΟΝΙΚΗ-ΑΝΤΙΣΤΡΟΦΗ ΓΑΜΜΑ ΣΥΖΥΓΗΣ ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΚΑΤΑΝΟΜΗ (NORMAL-INVERSE GAMMA (NIG) CONGUGATE PRIOR)

Η εκ των προτέρων κατανομή που δόθηκε στη σχέση (4.3.1) για τους συντελεστές του μοντέλου μπορεί να χρησιμοποιηθεί και στα κανονικά γραμμικά μοντέλα, δεδομένης της παραμέτρου σ^2. Συγκεκριμένα δηλαδή μπορούμε να θέσουμε:

$$f(\boldsymbol{\beta}_\gamma|\sigma^2,\ M_{\boldsymbol{\gamma}})\sim N\left(\boldsymbol{\mu}_\gamma, c^2\mathbf{V}_{(M_{\boldsymbol{\gamma}})}\sigma^2\right). \quad (4.4.1)$$

Ωστόσο πρέπει να ορίσουμε και μια εκ των προτέρων κατανομή για την διασπορά σ^2. Για το λόγο αυτό επιλέγουμε συνήθως Γάμμα(α, b) κατανομή για την ακρίβεια τ και κατά συνέπεια ως πρότερη κατανομή για τη διασπορά σ^2 μπορούμε να ορίσουμε μια Αντίστροφη Γάμμα(α, b) την οποία θεωρούμε κοινή σε όλα τα μοντέλα $M_{\boldsymbol{\gamma}}$. Κατ' αυτήν την επιλογή, η περιθώρια πιθανοφάνεια $f(\mathbf{y}|M_{\boldsymbol{\gamma}})$ του μοντέλου $M_{\boldsymbol{\gamma}}$ όπως αυτή δόθηκε από τη σχέση (3.2.1) γίνεται αναλυτικά υπολογίσιμη και το κύριο πρόβλημα που απαιτείται να αντιμετωπίσουμε είναι ο καθορισμός του $c^2\mathbf{V}_{(M_{\boldsymbol{\gamma}})}$.

Ως Κανονική-Αντίστροφη Γάμμα συζυγή (NIG) εκ των προτέρων κατανομή θεωρούμε την από κοινού εκ των προτέρων κατανομή των συντελεστών του μοντέλου και της διασποράς σ^2 που τη συμβολίζουμε $f(\boldsymbol{\beta}_\gamma, \sigma^2|M_{\boldsymbol{\gamma}})$ και απορρέει από το γινόμενο των πρότερων κατανομών που μόλις θέσαμε αντίστοιχα. Έχει αποδειχτεί ότι η εν λόγω εκ των προτέρων κατανομή είναι συζυγής καθώς η εκ των υστέρων $f(\boldsymbol{\beta}_\gamma, \sigma^2|M_{\boldsymbol{\gamma}}, \boldsymbol{y})$ είναι επίσης Κανονική-Αντίστροφη Γάμμα (NIG)(Bernardo and Smith (1994) ή O' Hagan (1994)).

4.5 Η g ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΚΑΤΑΝΟΜΗ ΤΟΥ ZELLNER

Η g εκ των προτέρων κατανομή του Zellner (1986) ορίζεται ως Κανονική-Αντίστροφη Γάμμα κατανομή (Normal Inverse Gamma (NIG)) με μέση τιμή κεντραρισμένη στο μηδέν, δηλαδή $\boldsymbol{\mu}_{\boldsymbol{\gamma}} = 0$ και πίνακα που καθορίζει τον εκ των προτέρων συσχετισμό των συντελεστών $\boldsymbol{\beta}_\gamma$ ίσο με $\mathbf{V}_{(M_{\boldsymbol{\gamma}})} = (\mathbf{X}_{\boldsymbol{\gamma}}{}^{\mathrm{T}}\mathbf{X}_{\boldsymbol{\gamma}})^{-1}$ όπου $\mathbf{X}_{\boldsymbol{\gamma}}$ είναι ο πίνακας σχεδιασμού του μοντέλου $M_{\boldsymbol{\gamma}}$, δηλαδή ο πίνακας που περιέχει τις επιδράσεις του διανύσματος των συντελεστών που είναι διαφορετικές από το μηδέν. Συγκεκριμένα δηλαδή για ένα μοντέλο $M_{\boldsymbol{\gamma}}$ με p_γ επεξηγηματικές μεταβλητές, ο Zellner θέτοντας $g = c^2$, εισήγαγε στην περίπτωση της κανονικής παλινδρόμησης την g εκ των προτέρων κατανομή για τους συντελεστές υπό την μορφή:

$$\boldsymbol{\beta}_\gamma|\sigma^2, \mathbf{X}_{\boldsymbol{\gamma}} \sim N_{p_\gamma}(\mathbf{0}, g\sigma^2(\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{X}_{\boldsymbol{\gamma}})^{-1})$$

με εκ των προτέρων κατανομή $f(\beta_o, \sigma^2) \propto \frac{1}{\sigma^2}$ κοινή για όλα τα μοντέλα $M_{\boldsymbol{\gamma}} \in \mathbf{M}$ (prior του Jeffreys).

4.5.1 ΚΛΕΙΣΤΕΣ ΜΟΡΦΕΣ ΤΟΥ ΠΑΡΑΓΟΝΤΑ BAYES ΜΕ ΕΠΙΛΟΓΗ ΤΗΣ g-PRIOR ΤΟΥ ZELLNER

Όπως είπαμε, το σημαντικότερο πλεονέκτημα της Zellner's g prior είναι ο αποτελεσματικός υπολογισμός του παράγοντα Bayes καθώς ο υπολογισμός της περιθώριας πιθανοφάνειας του μοντέλου, χρησιμοποιώντας τις εκ των προτέρων κατανομές για τους συντελεστές $\boldsymbol{\beta_\gamma}$, το σταθερό όρο β_o και τη διασπορά σ^2 που αναφέραμε στην ενότητα 4.5, δίνεται σε κλειστή μορφή ως:

$$f(\mathbf{Y}|M_{\boldsymbol{\gamma}}, g) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{\pi}^{(n-1)}\sqrt{n}} \left|\left|\mathbf{Y} - \bar{\mathbf{Y}}\right|\right|^{-(n-1)} \frac{(1+g)^{\frac{n-1-p_{\boldsymbol{\gamma}}}{2}}}{\left[1+g\left(1-{R_{\boldsymbol{\gamma}}}^2\right)\right]^{\frac{n-1}{2}}}$$

όπου ${R_{\boldsymbol{\gamma}}}^2$ είναι ο συντελεστής προσδιορισμού του μοντέλου παλινδρόμησης $M_{\boldsymbol{\gamma}}$ (Fully Bayes factors with a generalized g-prior , Yuzo Maruyama et. al).

Αμέσως τώρα θα δούμε δύο κλειστές μορφές του παράγοντα Bayes που αφορούν τη σύγκριση του μοντέλου $M_{\boldsymbol{\gamma}}$ ($\boldsymbol{\beta_\gamma} \in R^{p_\gamma}$) με το «μηδενικό» μοντέλο, δηλαδή το μοντέλο που περιέχει μόνο το σταθερό όρο (null model) και το «πλήρες» μοντέλο, δηλαδή το μοντέλο που περιέχει όλες τις p το πλήθος επεξηγηματικές μεταβλητές (full model) και στη συνέχεια θα δούμε πώς χρησιμοποιούμε αυτούς τους δύο για να παράγουμε τον παράγοντα Bayes με τον οποίο θα συγκρίνουμε δύο οποιαδήποτε μοντέλα $M_{\boldsymbol{\gamma}}$ και $M_{\boldsymbol{\gamma'}}$.

Α. ΠΑΡΑΓΟΝΤΑΣ BAYES ΓΙΑ ΣΥΓΚΡΙΣΗ ΜΕ ΤΟ «ΜΗΔΕΝΙΚΟ» ΜΟΝΤΕΛΟ

Από τον τύπο της περιθώριας πιθανοφάνειας που μόλις δόθηκε στην παρούσα ενότητα, μπορούμε να παρατηρήσουμε την ανεξαρτησία του «μηδενικού» μοντέλου M_N (null model) ($H_0: \boldsymbol{\beta_\gamma} = \mathbf{0}$) από την παράμετρο g καθώς έχουμε ότι ${R_{\boldsymbol{\gamma}}}^2 = 0$ και $p_{\boldsymbol{\gamma}} = 0$. Στηριζόμενοι σε αυτό και στον υπολογισμό της περιθώριας πιθανοφάνειας, ο παράγοντας Bayes για τη σύγκριση του μοντέλου $M_{\boldsymbol{\gamma}}$ με το μηδενικό μοντέλο M_N δίνεται ως:

$$BF_{M_\gamma M_N} = (1+g)^{(n-p_\gamma-1)/2}\left[1+g\left(1-{R_\gamma}^2\right)\right]^{-(n-1)/2}, \quad (4.5.1.1)$$

Απόδειξη:

$$BF_{M_\gamma M_N} = \frac{f(\mathbf{Y}|M_\gamma, g)}{f(\mathbf{Y}|M_\mathbf{N}, g)} = \frac{\frac{(1+g)^{\frac{n-1-p_\gamma}{2}}}{[1+g(1-R_\gamma{}^2)]^{\frac{n-1}{2}}}}{\frac{(1+g)^{\frac{n-1}{2}}}{(1+g)^{\frac{n-1}{2}}}}$$

$$= (1+g)^{(n-p_\gamma-1)/2}\,[1+g(1-R_\gamma{}^2)]^{-(n-1)/2}.$$

Β. ΠΑΡΑΓΟΝΤΑΣ BAYES ΓΙΑ ΣΥΓΚΡΙΣΗ ΜΕ ΤΟ ΠΛΗΡΕΣ ΜΟΝΤΕΛΟ

Για τη σύγκριση του μοντέλου M_γ με πίνακα σχεδιασμού $\mathbf{X_\gamma}$, με το πλήρες μοντέλο $M_\mathbf{F}$, μπορούμε να δούμε τον πίνακα σχεδιασμού του πλήρους μοντέλου χωρισμένο ως $\mathbf{X} = [\mathbf{I}, \mathbf{X_\gamma}, \mathbf{X_{-\gamma}}]$, έτσι ώστε το πλήρες μοντέλο $M_\mathbf{F}$ να μπορεί να γραφεί στη μορφή:

$$M_\mathbf{F} : \boldsymbol{\mu_F} = \mathbf{I}\beta_0 + \mathbf{X_\gamma}\boldsymbol{\beta_\gamma} + \mathbf{X_{-\gamma}}\boldsymbol{\beta_{-\gamma}},$$

όπου ο $\mathbf{X_{-\gamma}}$ αναφέρεται στις στήλες του πίνακα σχεδιασμού $\mathbf{X}$ που δεν περιλαμβάνονται στο μοντέλο M_γ. Το μοντέλο M_γ αντιστοιχεί στην υπόθεση $H_0: \boldsymbol{\beta_{-\gamma}} = \mathbf{0}$ ενώ το πλήρες μοντέλο $M_\mathbf{F}$ στην εναλλακτική $H_1: \boldsymbol{\beta_{-\gamma}} \in \mathbf{R^{p-p_\gamma}}$. Προκειμένου να συγκρίνουμε τα δύο αυτά μοντέλα ($\boldsymbol{\beta_\gamma}$ κοινό και στα δύο) με τη βοήθεια του παράγοντα Bayes χρησιμοποιούμε τις ακόλουθες εκ των προτέρων κατανομές:

$$M_\gamma : f(\beta_0, \sigma^2, \boldsymbol{\beta_\gamma}) \propto {}^1\!/_{\sigma^2}$$

και

$$M_\mathbf{F}: f(\beta_0, \sigma^2, \boldsymbol{\beta_\gamma}) \propto {}^1\!/_{\sigma^2}\ ,\ \boldsymbol{\beta_{-\gamma}}|\sigma^2 \sim N(\mathbf{0}, g\sigma^2(\mathbf{X_{-\gamma}}{}^T\mathbf{X_{-\gamma}})^{-1}),$$

με αποτέλεσμα ο παράγοντας Bayes να δίνεται στην ακόλουθη κλειστή μορφή:

$$BF_{M_\gamma M_F} = (1+g)^{-(n-p-1)/2}\left[1+g\frac{1-R_F{}^2}{1-R_\gamma{}^2}\right]^{(n-p_\gamma-1)/2} \quad (4.5.1.2),$$

όπου $R_F{}^2$ και $R_\gamma{}^2$ οι συντελεστές προσδιορισμού για το πλήρες μοντέλο $M_\mathbf{F}$ και το μοντέλο M_γ αντίστοιχα. Η απόδειξη παρατίθεται στο παράρτημα Α.

Σημειώνουμε ότι η σύγκριση ενός μοντέλου με το «μηδενικό» ή το πλήρες μοντέλο μπορεί να χρησιμοποιηθεί προκειμένου να συγκρίνουμε δύο οποιαδήποτε άλλα μοντέλα μεταξύ τους χρησιμοποιώντας το «μηδενικό» ή το πλήρες μοντέλο αντίστοιχα ως μοντέλο αναφοράς. Στην περίπτωση αυτή ο παράγοντας Bayes των δύο συγκρινόμενων μοντέλων θα δίνεται ως το πηλίκο των παραγόντων Bayes του κάθε μοντέλου ξεχωριστά σε σχέση με το μοντέλο αναφοράς.

Έτσι για παράδειγμα χρησιμοποιώντας ως μοντέλο αναφοράς το μηδενικό μοντέλο, ο παράγοντας Bayes για τα συγκρινόμενα μοντέλα $\mathbf{M_\gamma}$ και $\mathbf{M_{\gamma'}}$ θα δίνεται ως:

$$\mathbf{BF_{M_\gamma M_{\gamma'}}} = \frac{\mathbf{BF_{M_\gamma M_N}}}{\mathbf{BF_{M_{\gamma'} M_N}}}.$$

Σημειώνουμε ότι κατά ανάλογο τρόπο λαμβάνουμε τον παράγοντα Bayes εάν χρησιμοποιήσουμε ως μοντέλο αναφοράς το πλήρες μοντέλο, ή και οποιοδήποτε άλλο.

4.6 ΠΑΡΑΔΟΞΑ ΜΕ ΤΗΝ g ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΚΑΤΑΝΟΜΗ

Η επιλογή της $\mathbf{g}$ εκ των προτέρων κατανομής υπερισχύει έναντι άλλων λόγω του γεγονότος ότι υπάρχει μία μόνο υπερπαράμετρος που πρέπει να εκτιμηθεί. Ωστόσο, ο παράγοντας Bayes για την επιλογή μοντέλου με συγκεκριμένη επιλογή της υπερπαραμέτρου $\mathbf{g}$ μπορεί να οδηγήσει σε κάποιες ανεπιθύμητες καταστάσεις όπως θα συζητήσουμε παρακάτω.

4.6.1 ΤΟ ΠΑΡΑΔΟΞΟ ΤΩΝ LINDLEY-BARTLETT

Με μια Μπεϋζιανή προσέγγιση, η εκ των υστέρων κατανομή θα μπορούσε να είναι λογική ακόμα και αν η υπερπαράμετρος $\mathbf{g}$ έχει επιλεχθεί πολύ μεγάλη ούτως ώστε να είναι μη πληροφοριακή. Στην επιλογή μοντέλου ωστόσο, μια τέτοια επιλογή αποτελεί κακή ιδέα. Οι Lindley-Bartlett παρατήρησαν ότι στην ειδική περίπτωση που $\mathbf{n, p_\gamma}$ είναι σταθερά και η υπερπαράμετρος $\mathbf{g \to \infty}$, ο παράγοντας Bayes για σύγκριση του μοντέλου $\mathbf{M_\gamma}$ με το μηδενικό μοντέλο $\mathbf{M_N}$ που παρουσιάσαμε στη σχέση 4.5.1.1 τείνει στο 0 που σημαίνει ότι εξαναγκάζεται παρά την πληροφορία που παρέχεται από τα δεδομένα να υποστηρίζει ως πιο πιθανό το μικρότερο μοντέλο.

4.6.2 ΤΟ ΠΑΡΑΔΟΞΟ ΠΛΗΡΟΦΟΡΙΑΣ (INFORMATION PARADOX)

Υποθέτουμε ότι το μοντέλο M_γ παρουσιάζει τέλεια προσαρμογή στα δεδομένα. Δηλαδή υποθέτουμε ότι ο συντελεστής προσδιορισμού $R_\gamma^2 \to 1$, ή ισοδύναμα ο κλασικός στατιστικός ελεγχος $F_\gamma \to \infty$ με n, p_γ σταθερά. Σε αυτή την περίπτωση θα αναμέναμε πολύ υψηλή εκ των υστέρων πιθανότητα για το μοντέλο M_γ με απόρροια ο παράγοντας Bayes για τη σύγκρισή του με το μηδενικό μοντέλο M_N να τείνει στο ∞. Παρ' όλα αυτά, σε αυτή την περίπτωση ο παράγοντας Bayes που παρουσιάσαμε στη σχέση 4.5.1.1, με σταθερή τιμή για την υπερπαράμετρο g τείνει όπως βλέπουμε σε μια σταθερά $(1+g)^{(n-p_\gamma-1)/2}$ καθώς $R_\gamma^2 \to 1$, αφού εύκολα παρατηρούμε ότι ο δευτερος όρος του γινομένου στη σχέση 4.5.1.1. τείνει στη μονάδα. Η σύγκλιση αυτή μας οδηγεί σε αντίφαση με τις αναμενόμενες προσδοκίες.

4.7 ΤΡΟΠΟΙ ΕΠΙΛΟΓΗΣ ΤΗΣ g ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΚΑΤΑΝΟΜΗΣ

Στην Μπεϋζιανή επιλογή μοντέλου και μεταβλητών είναι απαραίτητο να επιλέξουμε κάποια τιμή για την υπερπαράμετρο g.

Κάτω από την επιλογή ομοιόμορφων εκ των προτέρων κατανομών για τα μοντέλα (George and Foster 2000), ο παράγοντας Bayes συμπεριφέρεται ως εξής:

- Μεγάλες τιμές για την υπερπαράμετρο g, με μέση τιμή $\mathbf{0}$ για τους συντελεστές $\boldsymbol{\beta_\gamma}$, οδηγούν σε επιλογή μοντέλων με λίγες παραμέτρους, εκείνες που αντιστοιχούν σε μεγάλους συντελεστές.
- Μικρές τιμές για την υπερπαράμετρο g, οδηγούν σε κορεσμένα μοντέλα με μικρούς συντελεστές.

Σύμφωνα λοιπόν με μελέτες που έχουν διεξαχθεί, συμπεριλαμβάνουμε τις ακόλουθες προτάσεις για την επιλογή της υπερπαραμέτρου g:

1. ***Εκ των προτέρων κατανομή μοναδιαίας πληροφορίας (Unit Information Prior)***: Από τη σκοπιά της κλασικής στατιστικής χρησιμοποιούνται κάποια κριτήρια ελέγχου «καλής προσαρμογής» τα οποία οδεύουν στην επιλογή μοντέλου. Ένα τέτοιο κριτήριο είναι ο συντελεστής προσδιορισμού R^2 και ακόμη ένα ενδιαφέρον κριτήριο είναι το κριτήριο BIC το οποίο λαμβάνει υπόψην του το γεγονός ότι ο συντελεστής προσδιορισμού R^2 αυξάνεται μονότονα καθώς αυξάνεται το πλήθος των παραμέτρων p. Οι Kass and Wasserman (1995) πρότειναν να χρησιμοποιούνται εκ των προτέρων κατανομές στις οποίες η παρεχόμενη πληροφορία για την παράμετρο είναι ίση με την

παρεχόμενη πληροφορία που εμπεριέχεται σε μία παρατήρηση. Έτσι στην περίπτωση της κανονικής γραμμικής παλινδρόμησης πρότειναν $\mathbf{g = n}$. Με αυτή την επιλογή οδηγούμαστε σε αποτελέσματα παρόμοια με εκείνα του κριτηρίου BIC.

2. ***Εκ των προτέρων κατανομή πληθωριστικού κινδύνου (Risk Inflation Criterion):*** Οι Foster and George (1994) μελέτησαν εκ των προτέρων κατανομές για επιλογή μοντέλου βασιζόμενες στο κριτήριο του πληθωριστικού κινδύνου RIC και πρότειναν για την υπερπαράμερο $\mathbf{g}$ την τιμή $\mathbf{g = p^2}$.

3. ***Benchmark εκ των προτέρων κατανομή(Benchmark Prior):*** Ο Fernandez et al. (2001) πραγματοποίησε μια μελέτη με διάφορες επιλογές για την υπερπαράμετρο $\mathbf{g}$ οι οποίες παρουσίαζαν εξάρτηση από το μέγεθος του δείγματος $\mathbf{n}$ και το πλήθος των παραμέτρων $\mathbf{p}$, καταλήγοντας στην επιλογή $\mathbf{g = \max(n, p^2)}$ η οποία αναφέρεται ως «Benchmark Prior» ή «BRIC» καθώς συνδυάζει τα κριτήρια BIC και RIC.

4. **Μπεϋζιανές εμπειρικές μέθοδοι:**

a) ***Local Empirical Bayes* g *prior*:** Η τοπική εμπειρική Μπεϋζιανή προσέγγιση βασίζεται στην εκτίμηση της υπερπαραμέτρου $\mathbf{g}$ ξεχωριστά για κάθε μοντέλο. Χρησιμοποιώντας την περιθώρια πιθανοφάνεια, ολοκληρώνοντας ως προς όλες τις παραμέτρους τη σχέση που δόθηκε στην ενότητα 4.5.1, μια Μπεϋζιανή εμπειρική εκτίμηση της υπερπαραμέτρου $\mathbf{g}$ δινεται εκτιμώντας τη μέγιστη τιμή της περιθώριας πιθανοφάνειας υπό τον περιορισμό να είναι μη αρνητική. Αυτό έχει ως αποτέλεσμα να λάβουμε την ακόλουθη εκτίμηση για το $\mathbf{g}$:

$$\hat{\mathbf{g}}^{\mathbf{EBL}} = \max\{\mathbf{F_\gamma - 1{,}0}\},$$

όπου

$$\mathbf{F_\gamma} = \frac{\mathbf{R_\gamma^2/p_\gamma}}{\mathbf{(1 - R_\gamma^2)/(n - 1 - p_\gamma)}}$$

είναι ο στατιστικός F έλεγχος για την υπόθεση $\boldsymbol{\beta_\gamma = 0}$.

b) ***Global Empirical Bayes* g *prior*:** Σύμφωνα με αυτή την επιλογή η εκτίμηση για την υπερπαράμετρο $\mathbf{g}$ είναι κοινή για όλα τα μοντέλα και

προκύπτει από την περιθώρια πιθανοφάνεια με Μπεϋζιανή στάθμιση μοντέλων ως:

$$\hat{g}^{EBG} = \text{argmax}_{g>0} \sum_{\gamma} f(M_\gamma) \frac{(1+g)^{(n-p_\gamma-1)/2}}{[1+g(1-R_\gamma^2)^{(n-1)/2}}.$$

Σημειώνουμε ότι η παραπάνω εκτίμηση δεν είναι υπολογίσιμη σε κλειστή μορφή, αλλά έχουν αναπτυχθεί ωστόσο αριθμητικές μέθοδοι που θα μπορούσαν να χρησιμοποιηθούν (George and Foster 2000).

Παρατηρήσεις:

Η εκ των προτέρων κατανομή μοναδιαίας πληροφορίας (Unit Information Prior), πληθωριστικού κινδύνου (Risk Inflation Cretirion) και η Benchmark εκ των προτέρων κατανομή δεν επιλύουν το πρόβλημα του παραδόξου πληροφορίας (Inofrmation Paradox), καθώς η επιλογή της υπερπαραμέτρου g δεν εξαρτάται από την πληροφορία των δεδομένων. Ωστόσο κάτω από τις δύο μπεϋζιανές εμπειρικές μεθόδους που μόλις αναφέραμε, έχουμε υπό μορφή θεωρήματος την ακόλουθη επιθυμητή συμπεριφορά.

Θεώρημα 4.7.1

Στην ενότητα του παραδόξου πληροφορίας με σταθερά $n, p < n$ και $R_\gamma^2 \to 1$, ο παράγοντας Bayes για τη σύγκριση των μοντέλων M_γ και M_N τείνει στο ∞ τόσο κάτω από την $\hat{g}^{EBL}$ (Local Empirical Bayes Estimation of g) όσο και κάτω από την $\hat{g}^{EBG}$ (Global Empirical Bayes Estimation of g)(Feng Liang et.al, 2007).

4.8 ΜΙΞΕΙΣ g ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΚΑΤΑΝΟΜΩΝ

Η ιδέα στις Μίξεις-g εκ των προτέρων κατανομών βασίζεται στο να χρησιμοποιήσουμε τη Zellner g prior θέτοντας εκ των προτέρων κατανομή στην υπερπαράμετρο g.

Υποθέτοντας λοιπόν ότι η υπερπαράμετρος $g > 0$ είναι τυχαία μεταβλητή και θέτοντάς της μία εκ των προτέρων κατανομή $f(g)$ (η οποία μπορεί να

εξαρτάται από το n), η περιθώρια πιθανοφάνεια του μοντέλου $\mathbf{f(Y|M_\gamma)}$ είναι ανάλογη του παράγοντα Bayes

$$\mathrm{BF_{M_\gamma M_N}} = \int_0^\infty (1+g)^{\frac{n-p_\gamma-1}{2}}\left[1+g\left(1-{R_\gamma}^2\right)\right]^{-\frac{n-1}{2}} f(g)dg \quad (4.8.1),$$

για σύγκριση με το «μηδενικό» μοντέλο.

Η εκ των υστέρων μέση τιμή του διανύσματος $\boldsymbol{\mu_\gamma}$, κάτω από την επιλογή του μοντέλου $\mathbf{M_\gamma \neq M_N}$ δίνεται από τη σχέση:

$$\mathrm{E}\left(\boldsymbol{\mu_\gamma}\,|\mathrm{M_\gamma}, \boldsymbol{\Upsilon}\right) = \mathrm{I_n}\widehat{\beta_0} + \mathrm{E}\left(\frac{g}{1+g}|\mathrm{M_\gamma}, \boldsymbol{\Upsilon}\right)\mathbf{X_\gamma}\widehat{\boldsymbol{\beta_\gamma}},$$

όπου $\widehat{\beta_0}$ και $\widehat{\boldsymbol{\beta_\gamma}}$ είναι οι εκτιμήτριες ελαχίστων τετραγώνων για το σταθερό όρο και τους συντελεστές του μοντέλου $\mathbf{M_\gamma}$ ενώ ο εκ των υστέρων μέσος του διανύσματος $\boldsymbol{\mu_\gamma}$ που απορρέει από ισοστάθμιση για όλα τα πιθανά μοντέλα $\mathbf{M_\gamma \neq M_N}$, δίνεται από τη σχέση:

$$\mathrm{E}\left(\boldsymbol{\mu_\gamma}|\boldsymbol{\Upsilon}\right) = \mathrm{I_n}\widehat{\beta_0} + \sum_{\gamma:\mathrm{M_\gamma \neq M_N}} f(\mathrm{M_\gamma}|\mathbf{Y})\mathrm{E}\left(\frac{g}{1+g}|\mathrm{M_\gamma}, \boldsymbol{\Upsilon}\right)\mathbf{X_\gamma}\widehat{\boldsymbol{\beta_\gamma}}.$$

Δεδομένου λοιπόν ότι η υπερπαράμτερος g δεν εμφανίζεται μόνο στους παράγοντες Bayes και στο πόσο πιθανό είναι ένα μοντέλο αλλά και στους εκ των υστέρων μέσους και τις προβλέψεις, ο καθορισμός της εκ των προτέρων κατανομής για την παράμετρο g είναι καθοριστικής σημασίας προκειμένου να αποφευχθούν δύσκολοι υπολογισμοί. Παρακάτω θα δούμε δύο περιπτώσεις επιλογής για την εκ των προτέρων κατανομή στην υπερπαράμετρο g.

4.8.1 HYPER-g ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΚΑΤΑΝΟΜΗ

Η hyper-g εκ των προτέρων κατανομή εισήχθη από τους Liang et.al (2008) με κύριο χαρακτηριστικό τη θεώρηση της υπερπαραμέτρου g ως τυχαία μεταβλητή προκειμένου με μία Μπεϋζιανή προσέγγιση η περιθώρια πιθανοφάνεια του μοντέλου να καταλήγει σε κλειστή μορφή ούτως ώστε να είναι εύκολα και αναλυτικά υπολογίσιμη. Αποτελεί μια προέκταση της κλασικής g εκ των προτέρων κατανομής, χρήσιμη για τα γραμμικά μοντέλα, σύμφωνα με την οποία η εκ των προτέρων κατανομή για την υπερπαράμετρο g μπορεί να είναι μια συνεχής κατανομή f(g).

Δεδομένου ότι καθώς το μέγεθος του δείγματος $n \to \infty$ η εκ των προτέρων κατανομή των συντελεστών του μοντέλου M_{γ}, συγκλίνει σε κανονική κατανομή ακόλουθης μορφής:

$$\beta_{\gamma} | \, \sigma^2, g, \gamma \sim N(0, gc\sigma^2 (X_{\gamma} W X_{\gamma})^{-1},$$

όπου c μια οποιαδήποτε σταθερά και W ο διαγώνιος πίνακας με τα βάρη, μια οικογένεια εκ των προτέρων κατανομών για την υπερπαράμετρο g δίνεται υπό τη μορφή:

$$f(g) = \frac{\alpha - 2}{2}(1+g)^{-\frac{\alpha}{2}}, \qquad g > 0 \ \& \ \alpha: \text{σταθερά} \quad (4.8.2)$$

η οποία όπως θα δούμε είναι γνήσια εκ των προτέρων κατανομή για $\alpha > 2$.

Αυτή η οικογένεια εκ των προτέρων κατανομών μελετήθηκε από τους Gui and George (2007) στο πρόβλημα της επιλογής μεταβλητών για γνωστή διασπορά και παρατηρήθηκε ότι για $\alpha \leq 2$ η hyper-g εκ των προτέρων κατανομή είναι μη ολοκληρώσιμη στη μονάδα ενώ για $1 < \alpha \leq 2$, η περιθώρια πιθανοφάνεια είναι υπολογίσιμη σε κλειστή μορφή λόγω του ότι είναι ευκολα υπολογίσιμη η εκ των υστέρων κατανομή της υπερπαραμέτρου g. Ωστόσο, αν και οι τιμές αυτές του α οδηγούν σε γνήσια εκ των υστέρων κατανομή, λόγω του γεγονότος ότι η υπερπαράμετρος g δεν περιλαμβάνεται στο «μηδενικό» μοντέλο ο παράγοντας Bayes παραμένει ακαθόριστος. Για το λόγο αυτό θα εφιστήσουμε την προσοχή μας στη μελέτη της συνεχούς συνάρτησης $f(g)$ για $\alpha > 2$. Αποδεικνύεται επίσης ότι για τον παράγοντα συρίκνωσης $g/(1+g)$, θα μπορούσε να χρησιμοποιηθεί μια εκ των προτέρων κατανομή:

$$\frac{g}{1+g} \sim \text{Beta}\left(1, \frac{\alpha}{2} - 1\right),$$

που είναι Beta κατανομή με μέση τιμή $2/\alpha$. Για $\alpha = 1$ η εκ των προτέρων κατανομή για τον παράγοντα συρίκνωσης είναι ομοιόμορφη. Για τιμές $\alpha > 4$ ο παράγοντας συρίκνωσης παίρνει τιμές κοντά στο 0 και έχουμε ανεπιθύμητη εκ των προτέρων συμπεριφορά ενώ για τιμές $\alpha = 3$ και $\alpha = 4$ λαμβάνει τιμές κοντά στη μονάδα. Επομένως τιμές $2 < \alpha \leq 4$ θεωρούνται λογικές για να δουλέψουμε και δίνουν το βασικό πλεονέκτημα στην hyper-g εκ των προτέρων κατανομή, να οδηγηθούμε σε κλειστή μορφή της εκ των υστέρων κατανομής για την υπερπαράμετρο g δοθέντος του μοντέλου M_{γ} που δίνεται ως:

$$f(g|\mathbf{Y}, M_{\boldsymbol{\gamma}}) = \frac{p_{\boldsymbol{\gamma}} + \alpha - 2}{2 {}_2F_1\left(\frac{n-2}{2}, 1; \frac{p_{\boldsymbol{\gamma}} + \alpha}{2}; R_{\boldsymbol{\gamma}}\right)^2} \frac{(1+g)^{(n-1-p_{\boldsymbol{\gamma}}-\alpha)/2}}{2} (1 + (1 - R_{\boldsymbol{\gamma}}{}^2)g)^{\frac{-(n-1)}{2}}$$

όπου ${}_2F_1(\alpha, b; c; z)$ είναι η Γκαουσιανή υπεργεωμετρική συνάρτηση Abramowitz και Milton (1970), δηλαδή:

$${}_2F_1(\alpha, b; c; z) = \frac{\Gamma(c)}{\Gamma(b)\Gamma(c-b)} \int_0^1 \frac{t^{b-1}(1-t)^{c-b-1}}{(1-tz)^{\alpha}} dt.$$

Το ολοκλήρωμα που παρουσιάζεται στην Γκαουσιανή υπεργεωμετρική συνάρτηση ${}_2F_1(\alpha, b; c; z)$ συγκλίνει για πραγματικό $|z| < 1$ με $c > b > 0$ και για $z = \pm 1$ αν και μόνο αν $c > a + b$ και $b > 0$. Σημειώνουμε επίσης ότι η σταθερά κανονικοποίησης στην εκ των προτέρων κατανομή της υπερπαραμέτρου g είναι ειδική περίπτωση της Γκαουσιανής υπεργεωμετρικής συνάρτησης ${}_2F_1(\alpha, b; c; z)$ για $z = 0$ και αναφέρουμε αυτή την οικογένεια εκ των προτέρων κατανομών (βλ. σχέση 4.8.1) ως hyper-g εκ των προτέρων κατανομές (hyper-g priors).

Μια άλλη ενδιαφέρουσα εμφάνιση της συνάρτησης ${}_2F_1(\alpha, b; c; z)$ είναι στον παράγοντα Bayes που χρησιμοποιούμε για τη σύγκριση του μοντέλου $M_{\boldsymbol{\gamma}}$ με το «μηδενικό μοντέλο». Συγκεκριμένα η σταθερά κανονικοποίησης στην εκ των υστέρων κατανομή για την υπερπαράμετρο g οδηγεί στο να έχουμε:

$$BF_{M_{\boldsymbol{\gamma}} M_N} = \frac{\alpha - 2}{2} \int_0^{\infty} (1+g)^{\frac{n-1-p_{\boldsymbol{\gamma}}-\alpha}{2}} \left(1 + (1 - R_{\boldsymbol{\gamma}}{}^2)g\right)^{\frac{-(n-1)}{2}} dg$$
$$= \frac{\alpha-2}{p_{\boldsymbol{\gamma}}+\alpha-2} \; {}_2F_1\left(\frac{n-1}{2}, 1; \frac{p_{\boldsymbol{\gamma}}+\alpha}{2}; R_{\boldsymbol{\gamma}}{}^2\right), \quad (4.8.3)$$

ο οποίος μπορεί εύκολα να υπολογιστεί.

Επιπρόσθετες ενδιαφέρουσες περιπτώσεις όπου εμφανίζεται η Γκαουσιανιανή υπεργεωμετρική συνάρτηση είναι στην εκ των υστέρων αναμενόμενη μέση τιμή της υπερπαραμέτρου g και του παράγοντα συρίκνωσης δοθέντος του μοντέλου.

Συγκεκριμένα έχουμε:

$$E(g|M_{\boldsymbol{\gamma}}, \mathbf{Y}) = \frac{2}{p_{\boldsymbol{\gamma}} + \alpha - 4} \frac{{}_2F_1\left(\frac{n-1}{2}, 2; \frac{p_{\boldsymbol{\gamma}} + \alpha}{2}; R_{\boldsymbol{\gamma}}{}^2\right)}{{}_2F_1\left(\frac{n-1}{2}, 1; \frac{p_{\boldsymbol{\gamma}} + \alpha}{2}; R_{\boldsymbol{\gamma}}{}^2\right)} \quad (4.8.4)$$

και

$$E\left(\frac{g}{1+g}\middle|M_{\gamma}, \mathbf{Y}\right) = \frac{\int g(1+g)^{\frac{n-1-p_{\gamma}-\alpha}{2}-1}\left(1+\left(1-{R_{\gamma}}^2\right)g\right)^{\frac{-(n-1)}{2}} dg}{\int (1+g)^{\frac{n-1-p_{\gamma}-\alpha}{2}}\left(1+\left(1-{R_{\gamma}}^2\right)g\right)^{\frac{-(n-1)}{2}} dg}$$

$$= \frac{2}{p_{\gamma}+\alpha}\frac{{}_2F_1\left(\frac{n-1}{2}, 2; \frac{p_{\gamma}+\alpha}{2}+1; {R_{\gamma}}^2\right)}{{}_2F_1\left(\frac{n-1}{2}, 1; \frac{p_{\gamma}+\alpha}{2}; {R_{\gamma}}^2\right)} \quad (4.8.5).$$

Για τον υπολογισμό της Γκαουσιανής υπεργεωμετρικής συνάρτησης έχουν αναπτυχθεί διάφορες υπορουτίνες οι οποίες ωστόσο παρουσιάζουν υπολογιστικές δυσκολίες για μεγάλο n και ${R_{\gamma}}^2$. Βελτιωμένες αριθμητικές μέθοδοι αποτελούν οι Laplace προσεγγίσεις (Tierney and Kadane 1986).

4.8.2 ZELLNER & SIOW ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΚΑΤΑΝΟΜΗ

Για τους ελέγχους υποθέσεων για τη μέση τιμή που παρουσιάσαμε στην ενότητα 4.5.1, ο Jeffreys (1961) απέρριψε την επιλογή κανονικών εκ των προτέρων κατανομών εξαιτίας του γεγονότος ότι αυτές οδηγούσαν στα παράδοξα που περιγράφηκαν στην ενότητα 4.6 και αντί αυτών πρότεινε τη χρήση της Caushy εκ των προτέρων κατανομής.

Επομένως η Zellner-Siow εκ των προτέρων κατανομή που θα δούμε αμέσως τώρα, χρησιμοποιείται ως μια απλή εκ των προτέρων κατανομή που ικανοποιεί βασικά κριτήρια συνέπειας στους προαναφερθείς στατιστικούς ελέγχους. Συγκεκριμένα οι Zellner-Siow (1980) πρότειναν πολυμεταβλητές Caushy εκ των προτέρων κατανομές για τους συντελεστές του μοντέλου παλινδρόμησης ως την καταλληλότερη επέκταση της δουλειάς του Jeffreys στο πρόβλημα της μέσης τιμής και η στρατηγική τους για τα συγκρινόμενα μοντέλα βασίζεται στον καθορισμό των δύο ακόλουθων εκ των προτέρων κατανομών για τους συντελεστές παλινδρόμησης του μοντέλου, το σταθερό όρο και τη διασπορά:

$$f\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\middle|\sigma^2\right) \propto \frac{\Gamma\left(\frac{p_{\gamma}}{2}\right)}{\pi^{\frac{p_{\gamma}}{2}}}\left|\frac{{X_{\gamma}}^T X_{\gamma}}{n\sigma^2}\right|^{\frac{1}{2}}\left(1+\frac{\sigma^2 {\boldsymbol{\beta}_{\boldsymbol{\gamma}}}^T {X_{\gamma}}^T X_{\gamma}\boldsymbol{\beta}_{\boldsymbol{\gamma}}}{n\sigma^2}\right)^{-\frac{p_{\gamma}}{2}}$$

και

$$f(\beta_0, \sigma^2) \propto \sigma^{-2}.$$

Η Zellner-Siow εκ των προτέρων κατανομή παρ όλα αυτά, δεν έγινε τόσο ευρέως γνωστή όπως η hyper-g εκ των προτέρων κατανομή και η εξήγηση εγκειται στο γεγονός ότι παρουσιάζει αρκετές δυσκολίες στον υπολογισμό της περιθώριας πιθανοφάνειας καθώς δεν υπάρχουν διαθέσιμες κλειστές μορφές υπολογισμού της. Αν και έγιναν αρκετές προσπάθειες προσέγγισης όπως οι Laplace προσεγγίσεις, παρατηρήθηκε ότι καθώς αυξάνετα η διάσταση του μοντέλου, μειώνεται η ακρίβεια των εν λόγω προσεγγίσεων.

Δεδομένου ότι η κατανομή Caushy μπορεί να αναπαρασταθεί ως μίξη κανονικών, μπορεί να χρησιμοποιηθεί μίξη των g εκ των προτέρων κατανομών με αντίστροφες γάμμα εκ των προτέρων κατανομές, δηλαδή να πάρουμε:

$$f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2) \propto \int N(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \left| 0, g\sigma^2 \left(X_{\boldsymbol{\gamma}}{}^T X_{\boldsymbol{\gamma}}\right)^{-1}\right) f(g)dg, \quad (4.8.6)$$

$$\text{όπου } f(g) = \frac{\left(\frac{n}{2}\right)^{1/2}}{\Gamma(\frac{1}{2})} g^{-3/2} e^{-n/(2g)} \quad (4.8.7).$$

Με την παραπάνω εκ των προτέρων κατανομή που δίνεται στη σχέση (4.8.6), αφού γίνει ολοκλήρωση ως προς $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ κατά τον υπολογισμό της περιθώριας πιθανοφάνειας του μοντέλου, λόγω της αναλογίας που περιγράψαμε στη σχέση (4.8.1), εύκολα μπορούμε να διαπιστώσουμε αντικαθιστώντας την εκ των προτέρων κατανομή που δίνεται από τη σχέση (4.8.7) ότι απομένει ο υπολογισμός ενός μονοδιάστατου ολοκληρώματος ως προς g το οποίο είναι ανεξάρτητο από τη διάσταση του μοντέλου. Αυτό το ολοκλήρωμα μπορεί να επιλυθεί με τυπικές τεχνικές ολοκλήρωσης ή χρησιμοποιώντας Laplace προσεγγίσεις.

Με τις Laplace προσεγγίσεις οδηγούμαστε σε έυκολους υπολογισμούς για την περιθώρια πιθανοφάνεια $f(Y|M_{\boldsymbol{\gamma}})$ αλλά και την εκ των υστέρων αναμενόμενη μέση τιμή του παράγοντα g/(1+g) που είναι αναγκαία για προβλέψεις, καθώς η εκ των υστέρων περιθώρια τιμή για την υπερπαράμετρο g αποτελεί λύση της κυβικής εξίσωσης (Feng Liang et.al, 2007).

4.9 ΣΥΝΕΠΕΙΑ

Στις προηγούμενες ενότητες αναφερθήκαμε σε συγκεκριμένες επιλογές g εκ των προτέρων κατανομών όπως την local empirical Bayes g prior, την global empirical Bayes g prior, την hyper-g prior και την Zellner and Siow.

Στην παράγραφο αυτή μας ενδιαφέρει να μελετήσουμε κάποιες από τις ασυμπτωτικές ιδιότητες των συγκεκριμένων επιλογών και συγκεκριμένα να εφιστήσουμε την προσοχή μας στη συμπεριφορά τους ως προς:

- το παράδοξο πληροφορίας που περιγράψαμε σε προηγούμενη ενότητα.
- την ασυμπτωτική συνέπεια της εκ των υστέρων πιθανότητας του μοντέλου καθώς $n \to \infty$.
- την ασυμπτωτική συνέπεια των προβλέψεων.

Συγκεκριμένα, στο παράδοξο πληροφορίας οι ασυμπτωτικές ιδιότητες των local empirical Bayes και global empirical Bayes g prior περιγράφτηκαν μέσω του θεωρήματος 4.7.1 ενώ για να επιλύσουμε το εν λόγω παράδοξο κάτω από την επιλογή της hyper-g prior και της Zellner and Siow χρειάζεται να ικανοποιούνται κάποιες επιπλέον απαιτήσεις.

Θεώρημα 4.9.1:

Για την επίλυση του παραδόξου πληροφορίας κάτω από την επιλογή των g εκ των προτέρων κατανομών που περιγράψαμε, θα πρέπει για όλα τα n και $\mathrm{p} < \mathrm{n}$ να ικανοποιείται η ακόλουθη συνθήκη:

$$\int_0^{\infty} (1+\mathrm{g})^{\frac{\mathrm{n}-1-\mathrm{p}_\gamma}{2}} \mathrm{f}(\mathrm{g})\mathrm{dg} = \infty \quad \forall\, \mathrm{p}_\gamma \le \mathrm{p} \quad \textbf{(4.9.1 (α))}$$

Απόδειξη:

Στην εκφυλισμένη προσέγγιση, δηλαδή κάτω από τη μηδενική υπόθεση, έχοντας $\mathrm{f}(\mathrm{g})$ να είναι η εκ των προτέρων κατανομή που έχουμε θέσει στην υπερπαράμετρο g, η περιθώρια πιθανοφάνεια του μοντέλου είναι ανάλογη του παράγοντα Bayes:

$$\mathrm{BF}_{\mathrm{M}_\gamma \mathrm{M}_\mathrm{N}} = \int_0^{\infty} (1+\mathrm{g})^{\frac{\mathrm{n}-\mathrm{p}_\gamma-1}{2}} \left[1+\mathrm{g}\left(1-\mathrm{R}_\gamma{}^2\right)\right]^{-\frac{\mathrm{n}-1}{2}} \mathrm{f}(\mathrm{g})\mathrm{dg}\,.$$

Η σχέση αυτή παρατηρούμε ότι είναι μονοτονικά αύξουσα ως προς $\mathrm{R}_\gamma{}^2$, επομένως όταν $\mathrm{R}_\gamma{}^2 \to 1$ τότε $\mathrm{BF}_{\mathrm{M}_\gamma \mathrm{M}_\mathrm{N}} \to \int_0^{\infty} (1+\mathrm{g})^{\frac{\mathrm{n}-\mathrm{p}_\gamma-1}{2}} \mathrm{f}(\mathrm{g})\mathrm{dg}$ λόγω του θεωρήματος της μονότονης σύγκλισης. Συνεπώς για την επίλυση του παραδόξου πληροφορίας αναγκαία και ικανή συνθήκη είναι η μη ολοκληρωσιμότητα της συνάρτησης $\int_0^{\infty} (1+\mathrm{g})^{\frac{\mathrm{n}-\mathrm{p}_\gamma-1}{2}} \mathrm{f}(\mathrm{g})$.

Στην ειδική περίπτωση μάλιστα που επιλεγεί η μικρότερη δυνατή τιμή για το μέγεθος του δείγματος, δηλαδή στην περίπτωση που $n = p + 2$ θα πρέπει να ικανοποιείται ότι:

$$\int_0^{\infty}(1+g)^{\frac{1}{2}} f(g)dg = \infty \quad \textbf{(4.9.1 (β))}$$

Σημείωση:

Η συνθήκη που μόλις περιγράψαμε στο θεώρημα 4.9.1 ικανοποιείται από την εκ των προτέρων Zellner & Siow κατανομή, ενώ για την hyper-g εκ των προτέρων κατανομή απαιτείται να ικανοποιείται επιπρόσθετα ο περιορισμός:

$$\alpha \leq n - p_\gamma + 1,$$

όπου στην περίπτωση του ελάχιστου δείγματος, προτείνεται $2 < \alpha \leq 3$.

4.10 ΣΥΝΕΠΕΙΑ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΟΥ

Ο ορισμός για την ασυμπτωτική συνέπεια των εκ των υστέρων πιθανοτήτων του μοντέλου καθώς $n \to \infty$, προτάθηκε από τον Fernades et al. (2001), σύμφωνα με τον οποίο:

$$\text{plim}_n\, f\left(M_\gamma \middle| \boldsymbol{\Upsilon}\right) = 1,$$

όπου M_γ είναι το πραγματικό μοντέλο και plim η σύγκλιση κατά πιθανότητα με μέτρο πιθανότητας τη δειγματική κατανομή κάτω από το μοντέλο M_γ που έχουμε υποθέσει ως πραγματικό.

Εναλλακτικά, μια ισοδύναμη διατύπωση του παραπάνω ορισμού, με χρήση του παράγοντα Bayes δίνεται ως ακολούθως:

$$\text{plim}_n\, BF_{M_{\gamma'}M_\gamma} = 0 \quad \forall\, M_{\gamma'} \neq M_\gamma \quad (4.10.1)$$

Για κάθε μοντέλο $M_{\gamma'}$, που δεν περιλαμβάνει το πραγματικό μοντέλο M_γ, υπό την υπόθεση ότι:

$$\lim_{n\to\infty} \frac{\boldsymbol{\beta}_{\gamma'}'\mathbf{X}_{\gamma'}'(1 - P_{\gamma'})\mathbf{X}_\gamma\boldsymbol{\beta}_\gamma}{n} = b_{\gamma'} \epsilon\, (0, \infty), \quad (1)$$

όπου $P_{\gamma'}$ είναι ο πίνακας προβολής του πίνακα σχεδιασμού $\mathbf{X_{\gamma}}$, ο Fernades et al. (2001) απέδειξε ότι η συνέπεια διατηρείται μόνο για τα κριτήρια BRIC και BIC.

Θεώρημα 4.10.1

Υπό την υπόθεση της οριακής συμπεριφοράς που περιγράφτηκε από τη σχέση (1), όταν για το πραγματικό μοντέλο M_{γ}, ισχύει $M_{\gamma} \neq M_N$, οι εκ των υστέρων πιθανότητες $f(M_{\gamma}|\mathbf{Y})$, κάτω από την επιλογή των empirical Bayes, Zellner & Siow και hyper-g εκ των προτέρων κατανομών, είναι συνεπείς στην Μπεϋζιανή επιλογή μοντέλου, ενώ στην περίπτωση που ισχύει $M_{\gamma} = M_N$, η συνέπεια διατηρείται μόνο κάτω από την επιλογή της Zellner & Siow εκ των προτέρων κατανομής (Feng Liang et al. (2008)).

4.11 HYPER-g/n ΕΚ ΤΩΝ ΠΡΟΤΕΡΩΝ ΚΑΤΑΝΟΜΗ

Μια σημαντική παρατήρηση σε ο,τι αφορά τη συνέπεια του «μηδενικού» μοντέλου κάτω από την επιλογή της Zellner & Siow εκ των προτέρων κατανομής είναι ότι η παράμετρος g εξαρτάται από το μέγεθος του δείγματος n, κάτι το οποίο δεν ισχύει στην περίπτωση των empirical Bayes και hyper-g εκ των προτέρων κατανομών.

Ωστόσο, αν και δεν ισχύει η συνέπεια για το «μηδενικό» μοντέλο σύμφωνα με τον ορισμό που δόθηκε για τη συγκεκριμένη επιλογή εκ των προτέρων κατανομών (emperical Bayes & hyper-g), το «μηδενικό» μοντέλο παραμένει ως το μοντέλο με την υψηλότερη εκ των υστέρων πιθανότητα η οποία φράσσεται από τη μονάδα.

Βάση της παρατήρησης αυτής, η έλλειψη της συνέπειας στην περίπτωση του «μηδενικού» μοντέλου, έδωσε την ώθηση να τροποποιηθεί η hyper-g εκ των προτέρων κατανομή σε hyper-g/n, η οποία δίνεται από την ακόλουθη σχέση:

$$f(g) = \frac{\alpha - 2}{2n}\left(1 + \frac{g}{n}\right)^{-\alpha/2}, \quad (4.11.1)$$

με τη σταθερά κανονικοποίησης για την εκ των προτέρων κατανομή, να αποτελεί μια ειδική περίπτωση της Γκαουσιανής υπεργεωμετρικής.

4.12 ΣΥΝΕΠΕΙΑ ΠΡΟΒΛΕΨΗΣ

Το ενδιαφέρον μας στη Μπεϋζιανή επιλογή μοντέλου εστιάζεται στο να μπορέσουμε να κάνουμε συνεπείς προβλέψεις. Συγκεκριμένα, δοθέντος των παρατηρηθέντων δεδομένων $(\mathbf{\Upsilon}, \mathbf{X_1}, \dots, \mathbf{X_P})$ και νέων τιμών $\mathbf{x}^* \in \mathrm{R}^{\mathrm{p}}$ για τις επεξηγηματικές μεταβλητές, μας ενδιαφέρει να προβλέψουμε την τιμή $\mathbf{\Upsilon}^*$ της μεταβλητής απόκρισης.

Στη Μπεϋζιανή επιλογή μοντέλου, η βέλτιστη εκτίμηση για την τιμή της μεταβλητής απόκρισης υπό την απαίτηση της ελαχιστοποίησης του τετραγωνικού σφάλματος, δίνεται μέσω του μηχανισμού της Μπεϋζιανής στάθμισης μοντέλου BMA (Bayesian Model Averaging), σύμφωνα με τον οποίο η πρόβλεψη δίνεται ως:

$$\widehat{\Upsilon}_{\mathrm{n}}^{*} = \widehat{\beta_0} + \sum_{\gamma} \mathbf{x}_{\gamma}^{*\prime}\, \widehat{\boldsymbol{\beta}}_{\gamma}\, \mathrm{f}\left(\mathrm{M}_{\gamma} \middle| \mathbf{\Upsilon}\right) \int_0^{\infty} \frac{\mathrm{g}}{1+\mathrm{g}}\, \mathrm{f}(\mathrm{g}|\, \mathrm{M}_{\gamma}, \mathbf{Y}) \mathrm{dg}, \quad (4.12.1)$$

όπου η $\mathrm{f}(\mathrm{g}|\mathrm{M}_{\gamma}, \mathbf{Y})$ μπορεί να αντικατασταθεί εκτιμώντας την παράμετρο g είτε χρησιμοποιώντας την τοπική εμπειρική Bayes εκτίμηση $\hat{\mathrm{g}}^{\mathrm{EBL}}$, ή την ολική εμπειρική εκτίμηση $\hat{\mathrm{g}}^{\mathrm{EBG}}$ μέσω κάποιου αλγορίθμου, όπως είναι ο EM αλγόριθμος.

Όταν η πραγματική δειγματική κατανομή είναι γνωστή που σημαίνει ότι $(\mathrm{M}_{\gamma}, \alpha, \boldsymbol{\beta}_{\gamma}, \sigma^2)$ είναι γνωστά, τότε η βέλτιστη πρόβλεψη για την τιμή της μεταβλητής απόκρισης Υ^* δίνεται από τον μέσο της. Έτσι λέμε ότι ο εκτιμητής $\widehat{\Upsilon}_{\mathrm{n}}^{*}$ που δόθηκε από τη σχέση 4.12.1 είναι συνεπής εκτιμητής αν και μόνο αν:

$$\mathrm{plim}_{\mathrm{n}} \widehat{\Upsilon}_{\mathrm{n}}^{*} = \mathrm{E}(\Upsilon^*) = \mathrm{a} + \mathbf{x}_{\gamma}^{*\prime} \boldsymbol{\beta}_{\gamma}, \quad (4.12.2)$$

όπου plim η σύγκλιση κατά πιθανότητα με μέτρο πιθανότητας τη δειγματική κατανομή υπό του μοντέλου M_{γ}.

Θεώρημα 4.12.1

Ο BMA εκτιμητής $\widehat{\Upsilon}_{\mathrm{n}}^{*}$ που δίνεται από τη σχέση 4.12.1, είναι συνεπής εκτιμητής σύμφωνα με τον ορισμό που δίνεται από τη σχέση 4.12.2, τόσο κάτω από την επιλογή των empirical Bayes εκ των προτέρων κατανομών όσο και κάτω από τη hyper-g, τη hyper-$\mathrm{g/n}$ και τη Zellner & Siow εκ των προτέρων κατανομή.

Απόδειξη:

Για $M_\gamma = M_N$, λόγω της συνέπειας των εκτιμητών των ελαχίστων τετραγώνων έχουμε ότι $\|\hat{\beta}_\gamma\| \to 0$ και από αυτό έπεται η συνέπεια του BMA εκτιμητή. Για $M_\gamma \neq M_N$, έχουμε ότι: $\text{plim}_n\, f(M_\gamma|\boldsymbol{\Upsilon}) = 1,$ επομένως κάνοντας χρήση της συνέπειας των εκτιμητών των ελαχίστων τετραγώνων αρκεί να δείξουμε ότι:

$$\text{plim}_n \int_0^\infty \frac{g}{1+g}\, f(g|\, M_\gamma, \mathbf{Y}) dg = 1. \quad \text{(a)}$$

Το ολοκλήρωμα στην παραπάνω σχέση μπορεί να γραφεί ως:

$$\frac{\int_0^\infty \frac{g}{1+g} L(g) f(g) dg}{\int_0^\infty L(g) f(g) dg}, \quad \text{(b)}$$

όπου $L(g) = (1+g)^{-p_\gamma/2}[1 - R_\gamma^2 \frac{g}{1+g}]^{-(n-1)/2}$.

Εφαρμόζοντας προσέγγιση Laplace στο κλάσμα που δίνεται στη σχέση (b), λαμβάνουμε ότι:

$$\int_0^\infty \frac{g}{1+g}\, f(g|\, M_\gamma, \mathbf{Y}) dg = \frac{\hat{g}_\gamma^{EBL}}{1+\hat{g}_\gamma^{EBL}}\left(1 + O\left(\frac{1}{n}\right)\right),$$

όμως $\text{plim}_n \hat{g}_\gamma^{EBL} = \infty$ υπό του μοντέλου M_γ και βάση αυτού καταλήγουμε στο ζητούμενο της σχέσης (a).

Συνοψίζοντας, σύμφωνα με τον πρώτο ορισμό της συνέπειας που μελετήσαμε δηλαδή απαιτώντας $\text{plim}_n\, f(M_\gamma|\boldsymbol{\Upsilon}) = 1$ σύμφωνα με το θεώρημα 4.10.1, είδαμε ότι για $M_\gamma \neq M_N$ με όλες τις επιλογές των g εκ των προτέρων κατανομών που μελετήσαμε έχουμε συνέπεια του μοντέλου, ενώ για $M_\gamma = M_N$, η συνέπεια διατηρείται μόνο κάτω από την επιλογογή των Zellner & Siow και hyper-g/n εκ των προτέρων κατανομών. Σύμφωνα όμως με το δεύτερο ορισμό για τη συνέπεια του εκτιμητή πρόβλεψης και σύμφωνα με το θεώρημα 4.12.1 είδαμε ότι όλες οι επιλογές των g εκ των προτέρων κατανομών που μελετήσαμε διατηρούν τη συνέπεια ακόμα και κάτω από το «μηδενικό» μοντέλο.

ΚΕΦΑΛΑΙΟ 5

ΕΦΑΡΜΟΓΕΣ ΣΤΗΝ R ΜΕ ΧΡΗΣΗ ΤΟΥ ΠΑΚΕΤΟΥ BAS ΓΙΑ ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ ΚΑΙ ΜΕΤΑΒΛΗΤΩΝ ΜΕ ΧΡΗΣΗ g-PRIOR

ΕΙΣΑΓΩΓΗ

Στο παρόν Κεφάλαιο θα εφαρμόσουμε αρχικά σε δείγμα πραγματικών δεδομένων τη Μπεϋζιανή προσέγγιση με τη χρήση των g-prior που περιγράψαμε, ούτως ώστε από το δυναμοσύνολο των πιθανών υποψήφιων μοντέλων να διακρίνουμε τα πρώτα καλύτερα με βάση την εκ των υστέρων τους πιθανότητα. Η υλοποίηση θα γίνει με χρήση της R και του στατιστικού πακέτου BAS (Bayesian Adaptive Sampling) στο δείγμα των πραγματικών δεδομένων (UScrime) που μπορεί κανείς να το βρει απευθείας διαθέσιμο στο πακέτο MASS της R ή μέσω της ιστοσελίδας:

http://lib.stat.cmu.edu/datasets/1993.expo/.

Μετέπειτα από την ανάλυση στα πραγματικά δεδομένα, θα ακολουθήσουν δύο διαδοχικά παραδείγματα με ανάλυση σε προσομοιωμένα δεδομένα, όπου στο πρώτο η προσομοίωση θα γίνει υπό την προϋπόθεση οι επεξηγηματικές μεταβλητές να είναι γραμμικώς ανεξάρτητες, ενώ στο δεύτερο υπό την προϋπόθεση να υπάρχει συσχέτιση.

5.1 ΣΥΝΤΟΜΗ ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΠΑΚΕΤΟΥ BAS

Το πακέτο BAS χρησιμοποιείται στα γραμμικά μοντέλα για Μπεϋζιανή στάθμιση μοντέλων με στοχαστική ή ντετερμινιστική δειγματοληψία χωρίς επανάθεση από την εκ των υστέρων κατανομή των μοντέλων. Η μορφή των εκ των προτέρων κατανομών που χρησιμοποιείται στους συντελεστές βασίζεται στη Zellner's g prior και σε μίξεις g-prior περιλαμβάνοντας ως επιλογές την Zellner and Siow prior, τη hyper-g prior, τη hyper-g/n prior, τη local and grobal empirical g-prior και κριτήρια όπως τα AIC και BIC.

Η χρήση του εν λόγω πακέτου στη γραμμική παλινδρόμηση αποσκοπεί στην εύρεση του καταλληλότερου μοντέλου (δηλαδή του μοντέλου με την υψηλότερη εκ των υστέρων πιθανότητα) και επομένως διαμέσου των εφαρμογών που θα πραγματοποιήσουμε σε πραγματικά και σε προσομοιωμένα

δεδομένα, θα δούμε αν οι αλγόριθμοι που περιλαμβάνει είναι αποδοτικοί και αξιόπιστοι καθώς θα εξετάσουμε αν είναι αποτελεσματική η επίδραση των εκ των προτέρων κατανομών που θα χρησιμοποιηθούν στους συντελεστές, στα αποτελέσματα των εκ των υστέρων κατανομών. Για p μικρότερο από 20-25 εξετάζει όλα τα μοντέλα όπως θα συμβεί και στα τρέχοντα παραδείγματα, ειδάλλως πραγματοποιεί τυχαία ή ντετερμινιστική δειγματοληψία χωρίς επανάθεση στο χώρο των μοντέλων.

Η βασική συνάρτηση που θα χρησιμοποιήσουμε από το εν λόγω πακέτο είναι η:

bas.lm(formula, data, n.models=NULL, prior="ZS-null", alpha=NULL, modelprior=uniform(),initprobs="Uniform",method="BAS", update=NULL, bestmodel=NULL, bestmarg=NULL, prob.local=0.0, prob.rw=0.5, Burnin.iterations=NULL, MCMC.iterations=NULL, lambda=NULL, delta=0.025**)**

της οποίας η αναλυτική ερμηνεία των παραμέτρων παρατίθεται στο πακέτο BAS της R.

Ακολούθως παρουσιάζουμε συνοπτικά κάποιες από τις σημαντικότερες παραμέτρους της εν λόγω συνάρτησης, που θα χρησιμοποιήσουμε μετέπειτα με προκαθορισμένες ή συγκεκριμένες τιμές στα επόμενα 3 παραδείγματα στις ενότητες 5.2, 5.3 και 5.4 αντίστοιχα.

formula	Η γραμμική φόρμουλα του μοντέλου, Y~.
data	Το σύνολο των δεδομένων.
n.models	Ο αριθμός των μοντέλων για τη δειγματοληψία. Εαν είναι null, το BAS εξετάζει όλα τα μοντέλα εκτός της περίπτωσης p>25.
prior	Η εκ των προτέρων κατανομή που τίθεται στους συντελεστές παλινδρόμησης. Περιλαμβάνονται οι επιλογές AIC, BIC, g-prior, ZS-null, ZS-full, hyper-g, hyper-g-n, EB-local και EB-global.
alpha	Προαιρετική υπερπαράμετρος. Για την Zellner-g prior προτείνεται η τιμή g ενώ για τη hyper-g η τιμή 3 ή τιμές που ανήκουν στο διάστημα (2,4).
modelprior	Οικογένεια εκ των προτέρων κατανομών για τα μοντέλα. Περιλαμβάνονται οι επιλογές Uniform, Bernoulli και beta.binomial. Συνήθης επιλογή είναι η Uniform που χρησιμοποιείται για να δώσει ίση εκ των προτέρων πιθανότητα σε κάθε μοντέλο.

initprops	Οι αρχικές πιθανότητες που χρησιμοποιούνται για τη δειγματοληψία χωρίς επανάθεση. Για την επιλογή "uniform" κάθε μεταβλητή πρόβλεψης έχει την ίδια πιθανότητα να ληφθεί στο δείγμα. Υπάρχουν και άλλες επιλογές διαθέσιμες (βλ. πακέτο BAS).
method	Παράμετρος που καθορίζει τη μέθοδο δειγματοληψίας που θα χρησιμοποιηθεί. Περιλαμβάνει τις μεθόδους "BAS"(που είναι και η default επιλογή), "MCMC" και "MCMC+BAS", με τις τελευταίες να χρησιμοποιούνται για μεγάλης διαστάσεως πρόβλημα.
Burnin.iterations	Ο αριθμός των επαναλήψεων που δεν θα ληφθεί υπόψην κατά την επιλογή της μεθόδου MCMC.
MCMC.iterations	Ο Αριθμός των επαναλήψεων που θα χρησιμοποιηθεί στον αλγόριθμο MCMC.

Το πακέτο BAS περιλαμβάνει ακόμη τις ακόλουθες σημαντικές συναρτήσεις:

- **update( ):** Ανανεώνει τα αποτελέσματα χρησιμοποιώντας διαφορετική εκ των προτέρων κατανομή στους συντελεστές του μοντέλου χωρίς να χρειαστεί να ξανα τρέξει ο αλγόριθμος.
- **summary( ):** Επιστρέφει προκαθορισμένα τα 5 πρώτα μοντέλα με τη μεγαλύτερη εκ των υστέρων πιθανότητα και κατά αντιστοιχία τους παράγοντες Bayes, τις εκ των υστέρων πιθανότητες, τις τιμές του συντελεστή προσδιορισμού, τη διάσταση και το λογάριθμο της περιθώριας πιθανοφάνειας.
- **coef( )**: Επιστρέφει τους περιθώριους εκ των υστέρων μέσους, τις τυπικές αποκλίσεις και τις περιθώριες εκ των υστέρων πιθανότητες των συντελεστών.
- **fitted( )**: Επιστρέφει τις προσαρμοσμένες τιμές βάση του μοντέλου με τη μεγαλύτερη πιθανότητα.
- **predict( )**: Επιστρέφει τις προβλεπόμενες τιμές σύμφωνα με τη Μπεϋζιανή στάθμιση μοντέλων.
- **plot( )**: Επιστρέφει 4 διαγράμματα, όπου το πρώτο αφορά τα υπόλοιπα με τις προσαρμοσμένες τιμές, το δεύτερο τις πιθανότητες του κάθε μοντέλου κατά την αναζήτηση, το τρίτο την λογαριθμοποιημένη περιθώρια πιθανοφάνεια ανάλογα με τη διάσταση του μοντέλου και το τέταρτο τις επεξηγηματικές μεταβλητές με τις συμπεριλαμβανόμενες περιθώριες πιθανότητες.
- **image( )**: Δημιουργεί ένα χάρτη με τα μοντέλα που συμμετάσχουν στη δειγματοληψία.

5.2 ΠΑΡΑΔΕΙΓΜΑ 1:

Μπεϋζιανή επιλογή μοντέλου και μεταβλητών σε πραγματικά δεδομένα.

Α) Σύντομη περιγραφή των δεδομένων UScrime:

Οι εγκληματολόγοι ενδιαφέρονται να μελετήσουν την επίδραση του καθεστώτος τιμωρίας στο ποσοστό εγκληματικότητας σε n=47 κράτη μέλη της Αμερικής για το έτος 1960, ως γραμμικής συνάρτησης 15 επεξηγηματικών μεταβλητών. Οι εν λόγω επεξηγηματικές μεταβλητές καθώς και η μεταβλητή απόκρισης που θα χρησιμοποιηθούν κατά την Μπεϋζιανή επιλογή μοντέλου και μεταβλητών στην παρούσα εφαρμογή, περιγράφονται ακολούθως:

Επεξηγηματικές μεταβλητές:

M: Το ποσοστό ανδρών ηλικίας 14-24.
S_0: Μία δείκτρια μεταβλητή για τις νότιες πολιτείες.
E_d: Η μέση εκπαίδευση.
P_{01}: Οι αστυνομικές δαπάνες το 1960.
P_{02}: Οι αστυνομικές δαπάνες το 1959.
LF: Το ποσοστό συμμετοχής στο εργατικό δυναμικό.
$M.F$: Ο αριθμός ανδρών ανά 1000 άτομα.
Pop: Ο πληθυσμός του κράτους.
NW: Το πλήθος των λευκών ανά 1000 ανθρώπους.
U_1: Το αστικό ποσοστό ανεργίας ανδρών 14-24 ετών.
U_2: Το αστικό ποσοστό ανεργίας ανδρών 35-39 ετών.
GDP: Το ακαθάριστο εγχώριο προϊόν ανά κεφαλή.
$Ineq$: Η εισοδηματική ανισότητα.
$Prob$: Η πιθανότητα φυλάκισης.
$Time$: Ο μέσος χρόνος υπηρέτησης στις κρατικές φυλακές.

Μεταβλητή απόκρισης:

Y: Το ποσοστό εγκληματικότητας σε συγκεκριμένη πληθυσμιακή κατηγορία.

Β) Εξαγωγή και ανάλυση των αποτελεσμάτων με χρήση του πακέτου BAS:

Χρησιμοποιώντας τα πραγματικά δεδομένα που μόλις περιγράψαμε αλλά και τις προαναφερθείς συναρτήσεις του πακέτου BAS, αναζητούμε το μοντέλο με

τη μεγαλύτερη εκ των υστέρων πιθανότητα που τηρεί την υπόθεση της καλής προσαρμογής.

Αρχικά εγκαθιστούμε στην R το πακέτο BAS και φορτώνουμε αυτό με τα δεδομένα τα οποία λαμβάνουμε για λόγους ευκολίας σε λογαριθμική μορφή (εκτός της $2^{ης}$ στήλης) με τις εντολές:

- install.packages('BAS')
- library(MASS)
- data(Uscrime)
- Uscrime[,-2]<-log(Uscrime[,-2])

Εν συνεχεία εφαρμόζουμε τη συνάρτηση bas.lm( ) για να εντοπίσουμε την κλάση με τα υψηλότερης εκ των υστέρων πιθανότητας μοντέλα, ανανεώνοντας την τιμή της παραμέτρου prior εφαρμόζοντας όλα τα κριτήρια που μας ενδιαφέρουν, ούτως ώστε λαμβάνοντας δείγματα χωρίς επανάθεση από την εκ των υστέρων κατανομή των μοντέλων να εξετάσουμε στην περίπτωσή μας όλα τα 2^{15} δυνατά μοντέλα (αφού $p = 15 < 25$), καταλήγοντας στη βέλτιστη επιλογή και σύγκριση των μεθόδων. Η εντολή που χρησιμοποιούμε για το σκοπό αυτό στην R και για το κριτήριο AIC είναι η:

- **crime.aic<-**bas.lm(y~., data = Uscrime, prior = "AIC", modelprior = uniform())

και με αντίστοιχο τρόπο ή με την εντολή update ανανεώνουμε την prior και εκχωρούμε τα αντίστοιχα αποτελέσματα στις μεταβλητές:

crime.bic, crime.g, crime.hg, crime.hgn, crime.EBL, crime.EBG , crime.ZSN & crime.ZSF.

Ακολούθως εξάγουμε και παρουσιάζουμε σε πίνακες τα εκ των υστέρων αποτελέσματα για τα 4 πρώτα μοντέλα με τη μεγαλύτερη εκ των υστέρων πιθανότητα, για κάθε μία από τις παραπάνω περιπτώσεις, με χρήση της εντολής:

- **summary(crime.aic,4)**

και κατά αντιστοιχία για τις υπόλοιπες prior.

Τα εξαγώμενα αποτελέσματα παρουσιάζονται ξεχωριστά στους επόμενους πίνακες και η σύγκριση πραγματοποιείται μεταγενέστερα από αυτούς.

	P(B != 0 | Y)	model 1	model 2	model 3	model 4
Intercept	1.0000000	1.00000	1.0000000	1.0000000	1.0000000
M	0.9771969	1.00000	1.0000000	1.0000000	1.0000000
So	0.3617534	0.00000	0.0000000	0.0000000	0.0000000
Ed	0.9985813	1.00000	1.0000000	1.0000000	1.0000000
Po1	0.7356143	1.00000	1.0000000	1.0000000	1.0000000
Po2	0.4668874	0.00000	0.0000000	0.0000000	0.0000000
LF	0.3380030	0.00000	0.0000000	0.0000000	1.0000000
M.F	0.3917988	1.00000	0.0000000	0.0000000	1.0000000
Pop	0.5715658	1.00000	0.0000000	0.0000000	1.0000000
NW	0.9181190	1.00000	1.0000000	1.0000000	1.0000000
U1	0.4111466	0.00000	0.0000000	1.0000000	0.0000000
U2	0.8636117	1.00000	1.0000000	1.0000000	1.0000000
GDP	0.6375222	1.00000	1.0000000	1.0000000	1.0000000
Ineq	0.9998382	1.00000	1.0000000	1.0000000	1.0000000
Prob	0.9884011	1.00000	1.0000000	1.0000000	1.0000000
Time	0.6452531	1.00000	1.0000000	1.0000000	1.0000000
BF	NA	1.00000	0.8978253	0.7595417	0.7178133
PostProbs	NA	0.01580	0.0142000	0.0120000	0.0113000
R2	NA	0.86340	0.8506000	0.8558000	0.8672000
dim	NA	12.00000	10.0000000	11.0000000	13.0000000
logmarg	NA	-13.41039	-13.5181654	-13.6854257	-13.7419314

Πίνακας 5.2.1. Μοντέλα υψηλότερης εκ των υστέρων πιθανότητας (AIC)

	P(B != 0 | Y)	model 1	model 2	model 3	model 4
Intercept	1.0000000	1.00000	1.0000000	1.0000000	1.0000000
M	0.9093806	1.00000	1.0000000	1.0000000	1.0000000
So	0.2286218	0.00000	0.0000000	0.0000000	0.0000000
Ed	0.9919748	1.00000	1.0000000	1.0000000	1.0000000
Po1	0.6872631	1.00000	1.0000000	1.0000000	1.0000000
Po2	0.4037022	0.00000	0.0000000	0.0000000	0.0000000
LF	0.1607246	0.00000	0.0000000	0.0000000	0.0000000
M.F	0.1677401	0.00000	0.0000000	0.0000000	0.0000000
Pop	0.3591253	0.00000	0.0000000	0.0000000	1.0000000
NW	0.7757744	1.00000	1.0000000	1.0000000	1.0000000
U1	0.2263200	0.00000	0.0000000	0.0000000	0.0000000
U2	0.6959277	1.00000	1.0000000	1.0000000	1.0000000
GDP	0.3634938	0.00000	0.0000000	1.0000000	0.0000000
Ineq	0.9992075	1.00000	1.0000000	1.0000000	1.0000000
Prob	0.9462122	1.00000	1.0000000	1.0000000	1.0000000
Time	0.4085486	1.00000	0.0000000	1.0000000	0.0000000
BF	NA	1.00000	0.7609295	0.5431578	0.5203179
PostProbs	NA	0.03470	0.0264000	0.0189000	0.0181000
R2	NA	0.84200	0.8265000	0.8506000	0.8375000
dim	NA	9.00000	8.0000000	10.0000000	9.0000000
logmarg	NA	-22.15855	-22.4317627	-22.7689035	-22.8118635

Πίνακας 5.2.2 Μοντέλα υψηλότερης εκ των υστέρων πιθανότητας (BIC)

P(B != 0 | Y)	model 1	model 2	model 3	model 4
Intercept	1.0000000	1.00000	1.0000000	1.0000000	1.0000000
M	0.7086633	1.00000	1.0000000	1.0000000	1.0000000
So	0.4117573	0.00000	0.0000000	0.0000000	0.0000000
Ed	0.8336964	1.00000	1.0000000	1.0000000	1.0000000
Po1	0.6434759	1.00000	1.0000000	0.0000000	1.0000000
Po2	0.5650151	0.00000	0.0000000	1.0000000	0.0000000
LF	0.3796794	0.00000	0.0000000	0.0000000	0.0000000
M.F	0.3719708	0.00000	0.0000000	0.0000000	0.0000000
Pop	0.4518090	0.00000	0.0000000	0.0000000	1.0000000
NW	0.5897842	1.00000	1.0000000	1.0000000	1.0000000
U1	0.3807743	0.00000	0.0000000	0.0000000	0.0000000
U2	0.5168192	1.00000	1.0000000	1.0000000	1.0000000
GDP	0.4528283	0.00000	0.0000000	0.0000000	0.0000000
Ineq	0.9456756	1.00000	1.0000000	1.0000000	1.0000000
Prob	0.7462329	1.00000	1.0000000	1.0000000	1.0000000
Time	0.4225926	1.00000	0.0000000	0.0000000	0.0000000
BF	NA	1.00000	0.9792021	0.8331256	0.8125369
PostProbs	NA	0.00180	0.0017000	0.0015000	0.0014000
R2	NA	0.84200	0.8265000	0.8229000	0.8375000
dim	NA	9.00000	8.0000000	8.0000000	9.0000000
logmarg	NA	17.41451	17.3934954	17.2319417	17.2069187

Πίνακας 5.2.3. Μοντέλα υψηλότερης εκ των υστέρων πιθανότητας (g)

P(B != 0 | Y)	model 1	model 2	model 3	model 4
Intercept	1.0000000	1.00000	1.0000000	1.0000000	1.0000000
M	0.8540880	1.00000	1.0000000	1.0000000	1.0000000
So	0.2909157	0.00000	0.0000000	0.0000000	0.0000000
Ed	0.9725279	1.00000	1.0000000	1.0000000	1.0000000
Po1	0.6655088	1.00000	1.0000000	0.0000000	1.0000000
Po2	0.4600328	0.00000	0.0000000	1.0000000	0.0000000
LF	0.2211288	0.00000	0.0000000	0.0000000	0.0000000
M.F	0.2233114	0.00000	0.0000000	0.0000000	0.0000000
Pop	0.3850323	0.00000	0.0000000	0.0000000	1.0000000
NW	0.6998971	1.00000	1.0000000	1.0000000	1.0000000
U1	0.2703076	0.00000	0.0000000	0.0000000	0.0000000
U2	0.6209142	1.00000	1.0000000	1.0000000	1.0000000
GDP	0.3784518	0.00000	0.0000000	0.0000000	0.0000000
Ineq	0.9957802	1.00000	1.0000000	1.0000000	1.0000000
Prob	0.8993760	1.00000	1.0000000	1.0000000	1.0000000
Time	0.3870606	1.00000	0.0000000	0.0000000	0.0000000
BF	NA	1.00000	0.9008736	0.6146928	0.6023051
PostProbs	NA	0.01640	0.0148000	0.0101000	0.0099000
R2	NA	0.84200	0.8265000	0.8229000	0.8375000
dim	NA	9.00000	8.0000000	8.0000000	9.0000000
logmarg	NA	25.11529	25.0108948	24.6286524	24.6082939

Πίνακας 5.2.4. Μοντέλα υψηλότερης εκ των υστέρων πιθανότητας (EBL)

P(B != 0 | Y)	model 1	model 2	model 3	model 4
Intercept	1.0000000	1.0000	1.0000000	1.0000000	1.0000000
M	0.8557940	1.0000	1.0000000	1.0000000	1.0000000
So	0.2892361	0.0000	0.0000000	0.0000000	0.0000000
Ed	0.9744588	1.0000	1.0000000	1.0000000	1.0000000
Po1	0.6645813	1.0000	1.0000000	1.0000000	0.0000000
Po2	0.4588179	0.0000	0.0000000	0.0000000	1.0000000
LF	0.2179496	0.0000	0.0000000	0.0000000	0.0000000
M.F	0.2205374	0.0000	0.0000000	0.0000000	0.0000000
Pop	0.3843497	0.0000	0.0000000	1.0000000	0.0000000
NW	0.7012110	1.0000	1.0000000	1.0000000	1.0000000
U1	0.2686459	0.0000	0.0000000	0.0000000	0.0000000
U2	0.6212074	1.0000	1.0000000	1.0000000	1.0000000
GDP	0.3782397	0.0000	0.0000000	0.0000000	0.0000000
Ineq	0.9964608	1.0000	1.0000000	1.0000000	1.0000000
Prob	0.9015429	1.0000	1.0000000	1.0000000	1.0000000
Time	0.3862507	1.0000	0.0000000	0.0000000	0.0000000
BF	NA	1.0000	0.8761666	0.6158209	0.6094915
PostProbs	NA	0.0158	0.0139000	0.0097000	0.0096000
R2	NA	0.8420	0.8265000	0.8375000	0.8229000
dim	NA	9.0000	8.0000000	9.0000000	8.0000000
logmarg	NA	25.0410	24.9087983	24.5561981	24.5458670

Πίνακας 5.2.5. Μοντέλα υψηλότερης εκ των υστέρων πιθανότητας (EBG)

P(B != 0 | Y)	model 1	model 2	model 3	model 4
Intercept	1.0000000	1.00000	1.0000000	1.0000000	1.000000
M	0.8429514	1.00000	1.0000000	1.0000000	1.000000
So	0.2952809	0.00000	0.0000000	0.0000000	0.000000
Ed	0.9669550	1.00000	1.0000000	1.0000000	1.000000
Po1	0.6624773	1.00000	1.0000000	0.0000000	1.000000
Po2	0.4654536	0.00000	0.0000000	1.0000000	0.000000
LF	0.2260716	0.00000	0.0000000	0.0000000	0.000000
M.F	0.2278912	0.00000	0.0000000	0.0000000	0.000000
Pop	0.3848058	0.00000	0.0000000	0.0000000	1.000000
NW	0.6861940	1.00000	1.0000000	1.0000000	1.000000
U1	0.2724634	0.00000	0.0000000	0.0000000	0.000000
U2	0.6075464	1.00000	1.0000000	1.0000000	1.000000
GDP	0.3770189	0.00000	0.0000000	0.0000000	0.000000
Ineq	0.9946277	1.00000	1.0000000	1.0000000	1.000000
Prob	0.8888800	1.00000	1.0000000	1.0000000	1.000000
Time	0.3815292	1.00000	0.0000000	0.0000000	0.000000
BF	NA	1.00000	0.9264345	0.6399928	0.612358
PostProbs	NA	0.01490	0.0138000	0.0095000	0.009100
R2	NA	0.84200	0.8265000	0.8229000	0.837500
dim	NA	9.00000	8.0000000	8.0000000	9.000000
logmarg	NA	23.13839	23.0619774	22.6920911	22.647951

Πίνακας 5.2.6. Μοντέλα υψηλότερης εκ των υστέρων πιθανότητας (Hyper-g)

P(B != 0 | Y)	model 1	model 2	model 3	model 4
Intercept	1.0000000	1.00000	1.0000000	1.0000000	1.0000000
M	0.8478593	1.00000	1.0000000	1.0000000	1.0000000
So	0.2720167	0.00000	0.0000000	0.0000000	0.0000000
Ed	0.9723166	1.00000	1.0000000	1.0000000	1.0000000
Po1	0.6639407	1.00000	1.0000000	0.0000000	1.0000000
Po2	0.4491879	0.00000	0.0000000	1.0000000	0.0000000
LF	0.2007827	0.00000	0.0000000	0.0000000	0.0000000
M.F	0.2035064	0.00000	0.0000000	0.0000000	0.0000000
Pop	0.3660300	0.00000	0.0000000	0.0000000	1.0000000
NW	0.6860167	1.00000	1.0000000	1.0000000	1.0000000
U1	0.2497849	0.00000	0.0000000	0.0000000	0.0000000
U2	0.6069948	1.00000	1.0000000	1.0000000	1.0000000
GDP	0.3551769	0.00000	0.0000000	0.0000000	0.0000000
Ineq	0.9961247	1.00000	1.0000000	1.0000000	1.0000000
Prob	0.8935637	1.00000	1.0000000	1.0000000	1.0000000
Time	0.3655911	1.00000	0.0000000	0.0000000	0.0000000
BF	NA	1.00000	0.9711425	0.6567878	0.5946605
PostProbs	NA	0.01790	0.0174000	0.0118000	0.0107000
R2	NA	0.84200	0.8265000	0.8229000	0.8375000
dim	NA	9.00000	8.0000000	8.0000000	9.0000000
logmarg	NA	23.51641	23.4871244	23.0960121	22.9966418

Πίνακας 5.2.7. Μοντέλα υψηλότερης εκ των υστέρων πιθανότητας (hyper-g/n)

P(B != 0 | Y)	model 1	model 2	model 3	model 4
Intercept	1.0000000	1.00000	1.0000000	1.0000000	1.000000
M	0.8535720	1.00000	1.0000000	1.0000000	1.000000
So	0.2737083	0.00000	0.0000000	0.0000000	0.000000
Ed	0.9746605	1.00000	1.0000000	1.0000000	1.000000
Po1	0.6651553	1.00000	1.0000000	0.0000000	1.000000
Po2	0.4490097	0.00000	0.0000000	1.0000000	0.000000
LF	0.2022374	0.00000	0.0000000	0.0000000	0.000000
M.F	0.2049659	0.00000	0.0000000	0.0000000	0.000000
Pop	0.3696150	0.00000	0.0000000	0.0000000	1.000000
NW	0.6944069	1.00000	1.0000000	1.0000000	1.000000
U1	0.2525834	0.00000	0.0000000	0.0000000	0.000000
U2	0.6149388	1.00000	1.0000000	1.0000000	1.000000
GDP	0.3601179	0.00000	0.0000000	0.0000000	0.000000
Ineq	0.9965359	1.00000	1.0000000	1.0000000	1.000000
Prob	0.8991841	1.00000	1.0000000	1.0000000	1.000000
Time	0.3717976	1.00000	0.0000000	0.0000000	0.000000
BF	NA	1.00000	0.9416178	0.6369712	0.594453
PostProbs	NA	0.01820	0.0172000	0.0116000	0.010800
R2	NA	0.84200	0.8265000	0.8229000	0.837500
dim	NA	9.00000	8.0000000	8.0000000	9.000000
logmarg	NA	23.65111	23.5909572	23.2000822	23.130999

Πίνακας 5.2.8. Μοντέλα υψηλότερης εκ των υστέρων πιθανότητας (ZS-null)

P(B != 0 | Y)	model 1	model 2	model 3	model 4
Intercept	1.0000000	1.000000	1.0000000	1.0000000	1.000000
M	0.8800119	1.000000	1.0000000	1.0000000	1.000000
So	0.3580105	0.000000	0.0000000	0.0000000	0.000000
Ed	0.9735415	1.000000	1.0000000	1.0000000	1.000000
Po1	0.6759217	1.000000	1.0000000	1.0000000	1.000000
Po2	0.4971666	0.000000	0.0000000	0.0000000	0.000000
LF	0.2974159	0.000000	0.0000000	0.0000000	0.000000
M.F	0.3008883	0.000000	0.0000000	0.0000000	0.000000
Pop	0.4578802	0.000000	0.0000000	1.0000000	0.000000
NW	0.7529388	1.000000	1.0000000	1.0000000	1.000000
U1	0.3475336	0.000000	0.0000000	0.0000000	1.000000
U2	0.6781381	1.000000	1.0000000	1.0000000	1.000000
GDP	0.4695913	0.000000	1.0000000	0.0000000	0.000000
Ineq	0.9948620	1.000000	1.0000000	1.0000000	1.000000
Prob	0.9187742	1.000000	1.0000000	1.0000000	1.000000
Time	0.4711992	1.000000	1.0000000	0.0000000	1.000000
BF	NA	1.000000	0.8733009	0.6342861	0.626701
PostProbs	NA	0.010000	0.0087000	0.0063000	0.006200
R2	NA	0.842000	0.8506000	0.8375000	0.847500
dim	NA	9.000000	10.0000000	9.0000000	10.000000
logmarg	NA	7.425266	7.2897905	6.9700105	6.9579800

Πίνακας 5.2.9. Μοντέλα υψηλότερης εκ των υστέρων πιθανότητας (ZS-full)

Πίνακες 5.2.1 – 5.2.9 Μοντέλα υψηλότερης εκ των υστέρων πιθανότητας από την εφαρμογή της μπεϋζιανή επιλογής μοντέλου και μεταβλητών με χρήση των g-prior και του πακέτου BAS

Πίνακες 5.2.1 - 5.29 - Παρατηρήσεις:

Οι πίνακες 5.2.1-5.2.9 περιλαμβάνουν 1 στήλη όπου δηλώνονται όλες οι επεξηγηματικές μεταβλητές, συμπεριλαμβανομένου και του σταθερού όρου καθώς επίσης ο παράγοντας Bayes (BF) για το κάθε μοντέλο, η εκ των υστέρων πιθανότητα αυτού (Postprobs), ο συντελεστής προσδιορισμού (R^2), η διάσταση του κάθε μοντέλου (dim) ανάλογα με το πόσες επεξηγηματικές μεταβλητές θα συμπεριληφθούν τελικώς στα μοντέλα με τις υψηλότερες εκ των υστέρων πιθανότητες, καθώς επίσης και η τιμή της λογαριθμοποιημένης περιθώριας πιθανοφάνειας κάτω από την εκάστοτε επιλεγμένη εκ των προτέρων κατανομή.

Οι υπόλοιπες στήλες αντιστοιχούν στα 4 πρώτα μοντέλα με τη μεγαλύτερη εκ των υστέρων πιθανότητα, όπου σε κάθε γραμμή έχουμε την τιμή 1 εφ όσον η αντίστοιχη επεξηγηματική μεταβλητή της $1^{ης}$ στήλης συμπεριληφθεί στο μοντέλο και την τιμή 0 εάν δεν συμπεριληφθεί. Οι υπόλοιπες γραμμές των 4

στηλών συμπληρώνονται με την τιμή του παράγοντα Bayes, της εκ των υστέρων πιθανότητας, του συντελεστή προσδιορισμού R^2, της διάστασης και της λογαριθμοποιημένης περιθώριας πιθανοφάνειας του εκάστοτε μοντέλου.

Παρατηρώντας τα εκ των υστέρων αποτελέσματα για τα 4 πρώτα μοντέλα, διαπιστώνουμε ότι τόσο κάτω από την επιλογή του κριτηρίου AIC όσο και κάτω από την επιλογή της prior ZS-full υπάρχει η τάση να επιλέγονται μοντέλα υψηλότερης διάστασης καθώς επίσης και η τάση (συγκριτικά με τις υπόλοιπες prior) να αποδίδεται μεγαλύτερη εκ των υστέρων πιθανότητα συμπερίληψης σε μεταβλητές που τελικώς δεν συμπεριλαμβάνονται στο εκάστοτε μοντέλο. Το τελευταίο ισχύει και κάτω από την επιλογή της g prior. Σημειώνουμε επίσης ότι κάτω από αυτές τις τρεις επιλογές, εχουμε τις μικρότερες εκ των υστέρων πιθανότητες (PostProbs) για τα 4 πρώτα μοντέλα ενώ επιπλέον τόσο κάτω από την επιλογή των κριτηρίων AIC και BIC όσο και κάτω από της επιλογή της ZS-full prior, παρατηρούμε τις μικρότερες τιμές για τη λογαριθμοποιημένη περιθώρια πιθανοφάνεια.

Εστιάζοντας την προσοχή μας στα εκ των υστέρων αποτελέσματα για το πρώτο προτεινόμενο μοντέλο (μοντέλο υψηλότερης εκ των υστέρων πιθανότητας), παρατηρούμε ότι αυτά είναι παραπλήσια κάτω από τις επιλογές, BIC, EBL, EBG, hyper-g, hyper-g/n και ZS-null, σύμφωνα με τις οποίες ενσωματώνονται στο μοντέλο 9 επεξηγηματικές μεταβλητές, σημειώνοντας ότι ο σταθερός όρος ενσωματώνεται πάντα στη λίστα των μοντέλων υψηλότερης εκ των υστέρων πιθανότητας κάτω από οποιαδήποτε επιλογή.

Η διερεύνηση των εκ των υστέρων αποτελεσμάτων για τις δυνατές επιλογές εκ των προτέρων κατανομών που μόλις περιγράφηκαν και που είδαμε ότι οι hyper-g, hyper-g/n, EBL, EBG και ZS-null είναι οι πιο αποδοτικές, θα συνεχιστεί παρακάτω με τη σύγκριση των εκ των υστέρων μέσων και τυπικών σφαλμάτων των συντελεστών καθώς και επιπρόσθετων διαγνωστικών ελέγχων.

Για την εξαγωγή των περιθώριων εκ των υστέρων μέσων, τυπικών αποκλίσεων και εκ των υστέρων πιθανοτήτων για τη μη μηδενικότητα των συντελεστών για prior = AIC, χρησιμοποιούμε την εντολή:

- coef(crime.aic)

και κατα αντιστοιχία την ίδια εντολή για τις υπόλοιπες περιπτώσεις των εκ των προτέρων κατανομών που έχουμε αναφέρει. Τα εκ των υστέρων αποτελέσματα για τους συντελεστές παρουσιάζονται συγκριτικά για όλες τις περιπτώσεις αυτές στους πίνακες της ακόλουθης παραγράφου.

Συγκριτικοί πίνακες εκ των υστέρων αποτελεσμάτων για τους συντελεστές:

Περιθώριοι κ των υστέρων μέσοι για τους συντελεστές (2^{15} υποψήφια μοντέλα)

Πραγματικά δεδομένα	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
Intercept	6.72	6.72	6.73	6.72	6.72	6.72	6.72	6.72	6.72
M	1.41	1.28	0.74	1.11	1.13	1.13	1.14	1.14	1.23
S_0	0.04	0.03	0.04	0.04	0.04	0.04	0.04	0.04	0.05
Ed	2.10	2.03	1.17	1.80	1.85	1.84	1.85	1.86	1.96
Po1	0.62	0.63	0.46	0.59	0.60	0.59	0.59	0.60	0.61
Po2	0.22	0.30	0.28	0.32	0.32	0.32	0.32	0.32	0.31
LF	0.13	0.05	0.13	0.07	0.06	0.07	0.06	0.06	0.10
M.F	-0.52	-0.07	0.08	-0.03	-0.02	-0.04	-0.04	-0.03	-0.13
Pop	-0.04	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.03
NW	0.09	0.08	0.04	0.06	0.07	0.07	0.07	0.07	0.08
U1	-0.08	-0.03	-0.01	-0.02	-0.02	-0.03	-0.03	-0.02	-0.04
U2	0.32	0.24	0.13	0.20	0.20	0.21	0.21	0.21	0.25
GDP	0.39	0.22	0.21	0.21	0.21	0.21	0.22	0.21	0.28
Ineq	1.47	1.43	1.06	1.37	1.39	1.38	1.38	1.39	1.46
Prob	-0.27	-0.24	-0.14	-0.21	-0.21	-0.22	-0.22	-0.22	-0.24
Time	-0.17	-0.11	-0.05	-0.08	-0.08	-0.09	-0.09	-0.09	-0.11

Πίνακας 5.2.10.

Πίνακας 5.2.10 - Παρατηρήσεις:

Στον πίνακα 5.2.10 παρατηρούμε ότι οι περιθώριες εκ των υστέρων εκτιμήσεις για τους μέσους των αντίστοιχων συντελεστών υπό των εκ των προτέρων κατανομών hyper-g, hyper-g/n, EBL, EBG και ZS-null είναι με ελάχιστες (σχεδόν αμελητέες) διαφορές ίδιες.

Οι εκ των προτέρων κατανομές AIC και ZS-null που είδαμε προηγουμένως να παρουσιάζουν μια αδυναμία ως προς την ανίχνευση του πραγματικού μοντέλου, εξακολουθούν φανερά να διαφωνούν με τις υπόλοιπες ακόμη και ως προς την εκτίμηση των περιθώριων εκ των υστέρων μέσων των συντελεστών, γεγονός που σε συνδυασμό με τα προαναφερθή σχόλιά μας, μας κάνει να αμφιβάλλουμε ως προς την αποδοτικότητά τους συγκριτικά με τις υπόλοιπες.

Επιπλέον με τη χρήση των εκ των προτέρων κατανομών BIC και g που πρωτίστως είδαμε να οδηγούν και αυτές στην ανίχνευση του πραγματικού μοντέλου, έχουμε να επισημάνουμε ότι όσον αφορά τις εκτιμήσεις των εκ των υστέρων μέσων των συντελεστών παρατηρούμε μια διαφοροποίηση συγκριτικά με τις τιμές αυτών κάτω από τις hyper-g, hyper-g/n, EBL, EBG και ZS-null prior καθώς υπό της prior BIC θα λέγαμε ότι φαινεται οι συντελεστές να δίνουν

μεγαλύτερη βαρύτητα στις αντίστοιχες επεξηγηματικές μεταβλητές, ενώ υπό της g-prior μικρότερη.

Η ανάλυση των αποτελεσμάτων συνεχίζεται με την παρουσίαση των περιθώριων εκ των υστέρων τυπικών σφαλμάτων των συντελεστών στον πίνακα 5.2.11.

Περιθώριες εκ των υστέρων τυπικές αποκλίσεις για τους συντελεστές (2^{15} υποψήφια μοντέλα)

Πραγματικά δεδομένα	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
Intercept	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
M	0.54	0.63	0.62	0.70	0.67	0.70	0.70	0.70	0.68
S_0	0.09	0.08	0.10	0.09	0.09	0.09	0.09	0.09	0.10
Ed	0.56	0.59	0.71	0.65	0.64	0.64	0.63	0.63	0.66
Po1	0.54	0.52	0.52	0.54	0.54	0.54	0.53	0.53	0.57
Po2	0.52	0.50	0.52	0.52	0.52	0.52	0.52	0.52	0.56
LF	0.42	0.28	0.43	0.33	0.32	0.33	0.33	0.32	0.41
M.F	1.30	0.75	1.04	0.85	0.81	0.85	0.84	0.81	1.05
Pop	0.05	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.05
NW	0.23	0.06	0.05	0.06	0.06	0.06	0.06	0.06	0.06
U1	0.23	0.17	0.18	0.18	0.18	0.18	0.18	0.18	0.22
U2	0.43	0.22	0.19	0.22	0.22	0.22	0.22	0.22	0.24
GDP	0.43	0.38	0.36	0.37	0.37	0.37	0.37	0.37	0.42
Ineq	0.36	0.36	0.42	0.37	0.37	0.37	0.37	0.37	0.39
Prob	0.10	0.11	0.11	0.12	0.12	0.12	0.12	0.12	0.12
Time	0.19	0.17	0.14	0.16	0.16	0.16	0.16	0.16	0.18

Πίνακας 5.2.11.

Πίνακας 5.2.11 - Παρατηρήσεις:

Στον πίνακα 5.2.11 βλέπουμε και πάλι ως προς τις περιθώριες εκ των υστέρων τυπικές αποκλίσεις των συντελεστών να υπάρχει ουσιαστική διαφοροποίηση των αποτελεσμάτων με χρήση των εκ των προτέρων κατανομών AIC και ZS-full, με μεγαλύτερα τυπικά σφάλματα και στις δύο αυτές περιπτώσεις συγκριτικά με τις άλλες μεθόδους, γεγονός που μας κάνει να τις εμπιστευτούμε λιγότερο.

Σχετικά με την επιλογή των hyper-g, hyper-g/n, EBL, EBG και ZS-null παρατηρούμε εντυπωσιακή συμφωνία στα εκ των υστέρων τυπικά σφάλματα των συντελεστών με την hyper-g/n να βελτιώνει κάπως την hyper-g και την EBG αντίστοιχα να βελτιώνει την EBL. Τα τυπικά σφάλματα σύμφωνα με τις hyper-g, EBL και EBG είναι σχεδόν ίδια ενώ με τις hyper-g/n και ZS-null λίγο μικρότερα. Ωστόσο η διαφοροποίηση των περιθώριων εκ των υστέρων τυπικών σφαλμάτων με τις εν λόγω prior επί της ουσίας φαίνεται να είναι

αμελητέα και μπορούμε να εμπιστευτούμε οποιαδήποτε από αυτές, αρκεί βέβαια να δούμε ότι συμφέρουν και ως προς το χρόνο εκτέλεσης αλλά και ως προς την αξιοποίηση υπολογιστικών πόρων και δυνατοτήτων.

Επιπλέον με την επιλογή της g-prior μολονότι καταλήγουμε στην ανίχνευση του πραγματικού μοντέλου είδαμε ότι η εκ των υστέρων πιθανότητα που δίνει σε αυτό είναι αρκετά μικρή ενώ τώρα παρατηρούμε και τα περιθώρια εκ των υστέρων τυπικά σφάλματα των συντελεστών να μην έρχονται σε πλήρη συμφωνία με την επιλογή των αποδοτικών prior που αναφέραμε. Μολονότι με την επιλογή της g-prior φαίνεται για κάποιους συντελεστές τα τυπικά σφάλματα να μειώνονται (χωρίς σημαντική μείωση), παρατηρούμε ταυτόχρονα ότι για κάποιους άλλους συντελεστές τα τυπικά σφάλματα είναι μεγαλύτερα με σημαντική θα λέγαμε διαφορά. Αυτό μπορεί να οφείλεται στην εξάρτηση της g-prior από το μέγεθος του δείγματος και το πλήθος των παραμέτρων που έχουν ληφθεί στην παρούσα εφαρμογή.

Τέλος με την επιλογή της BIC prior τα τυπικά σφάλματα φαίνονται να είναι λίγο μικρότερα, ωστόσο ως γνωστό από προηγούμενες έρευνες που έχουν διεξαχθεί το κριτήριο BIC αποτελεί μεν μια βελτιωμένη εκδοχή του κριτηρίου AIC ως προς τη συνέπεια υπό την έννοια ότι ανιχνεύει το πιο φειδωλό μοντέλο με τη μικρότερη απόκλιση από το πραγματικό, αλλά ως προς την αποδοτικότητα (υπό την έννοια ανίχνευσης του μοντέλου που ελαχιστοποιεί το τετραγωνικό σφάλμα) καθώς $n \to \infty$, υστερεί του κριτηρίου AIC κάτι το οποίο δεν φαίνεται στην παρούσα εφαρμογή λόγω επιλογής μικρού μεγέθους δειγματος. Βάση επομένως αυτής της επισήμανσης και του γεγονότος ότι το κριτήριο BIC λαμβάνει υπόψην του το γεγονός ότι ο συντελεστής προσδιορισμού αυξάνεται μονοτονικά καθώς αυξάνεται το πλήθος των παραμέτρων αλλά και του γεγονότος ότι από την έως τώρα ανάλυση οι επιλογές των hyper-g, hyper-g/n, EBL, EBG και ZS-null φαίνεται να υπερτερούν ως προς την αξιοπιστία έναντι των prior AIC και ZS-full, θα μπορούσαμε μιας και η διαφορά των τυπικών σφαλμάτων υπό την prior BIC δεν είναι σημαντική συγκριτικά με τις προαναφερθείς, να τις θεωρήσουμε ως πιο αποδοτικές.

Μετέπειτα που θα πραγματοποιηθεί η παρούσα ανάλυση σε προσομοιωμένα δεδομένα μεγαλύτερου μεγέθους, η σύγκριση αυτή θα είναι ευκολότερη.

Στον αμέσως επόμενο πίνακα παρουσιάζουμε την περιθώρια εκ των υστέρων πιθανότητα οι συντελεστές να είναι μη μηδενικοί για κάθε μία από τις διαφορετικές επιλογές εκ των προτέρων κατανομών.

Περιθώρια εκ των υστέρων πιθανότητα p(B != 0) (2^{15} υποψήφια μοντέλα)

Πραγματικά δεδομένα	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
Intercept	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
M	0.98	0.91	0.71	0.84	0.85	0.85	0.86	0.85	0.88
S_0	0.36	0.23	0.41	0.30	0.27	0.30	0.29	0.27	0.36
Ed	1.00	0.99	0.83	0.97	0.97	0.97	0.97	0.97	0.97
Po1	0.74	0.69	0.64	0.66	0.66	0.67	0.66	0.66	0.68
Po2	0.47	0.40	0.57	0.47	0.45	0.46	0.46	0.45	0.50
LF	0.34	0.16	0.38	0.23	0.20	0.22	0.22	0.20	0.30
M.F	0.39	0.17	0.38	0.23	0.20	0.22	0.22	0.20	0.30
Pop	0.57	0.36	0.45	0.38	0.37	0.39	0.38	0.37	0.46
NW	0.92	0.78	0.59	0.69	0.69	0.70	0.70	0.70	0.75
U1	0.41	0.23	0.38	0.27	0.25	0.27	0.27	0.25	0.35
U2	0.86	0.70	0.52	0.61	0.61	0.62	0.62	0.61	0.68
GDP	0.64	0.36	0.45	0.38	0.36	0.38	0.38	0.36	0.47
Ineq	1.00	1.00	0.95	1.00	1.00	1.00	1.00	1.00	0.99
Prob	0.99	0.95	0.74	0.89	0.90	0.90	0.90	0.90	0.92
Time	0.65	0.41	0.42	0.38	0.37	0.39	0.39	0.37	0.47

Πίνακας 5.2.12.

Πίνακας 5.2.12 - Παρατηρήσεις:

Στον πίνακας 5.2.12 παρατηρούμε τις περιθώριες εκ των υστέρων πιθανότητες οι συντελεστές του μοντέλου των 15 επεξηγηματικών μεταβλητών να είναι μη μηδενικοί.

Παρατηρούμε ότι οι κοινές μεταβλητές που με υψηλή πιθανότητα θα συμπεριληφθούν στο μοντέλο (ανεξαρτήτως της prior που θα χρησιμοποιηθεί) είναι ο σταθερός όρος που θα συμπεριληφθεί με πιθανότητα 1 σε όλες τις περιπτώσεις και οι μεταβλητές M, Ed, Po1, M.F, Pop, NW, U2, GDP, Ineq, Prob και Time που συμπεριλαμβάνονται με εξίσου υψηλή πιθανότητα.

Όσον αφορά τις συγκεκριμένες μεταβλητές παρατηρούμε ότι κάτω από την επιλογή των hyper-g, hyper-g/n, EBL, EBG και ZS-null, συμπεριλαμβάνονται με παραπλήσια πιθανότητα οι αντίστοιχοι συντελεστές να είναι μη μηδενικοί, σε αντιδιαστολή με την επιλογή των prior AIC, BIC, g και ZS-full όπου οι πιθανότητες μη μηδενικότητας των συντελεστών βάση των οποίων γίνεται η ενσωμάτωση των εν λόγω επεξηγηματικών μεταβλητών είναι μεγαλύτερες.

Επιπλέον το σημαντικό που παρατηρούμε κι εδώ και που έχουμε ήδη επισημάνει σε προηγούμενη παρατήρηση είναι ότι βάση των prior AIC και ZS-full θα λέγαμε ότι επικρατεί η τάση να δίνεται υψηλότερη πιθανότητα για μη μηδενικότητα των συντελεστών που αντιστοιχούν στις μεταβλητές που δεν συμπεριλαμβάνονται στο μοντέλο, σε αντίθεση με τις υπόλοιπες prior και όσον αφορά τις επιλογές των prior όπου έχουμε τις λιγότερες διαφοροποιήσεις στις εκ των υστέρων πιθανότητες, δηλαδή στις περιπτώσεις των hyper-g, hyper-g/n, EBL, EBG και Zellner-Siow-null, έχουμε να παρατηρήσουμε ότι με τις hyper-g/n και Zellner-Siow-null οι εκ των υστέρων πιθανότητες για τις μεταβλητές που δεν συμπεριλαμβάνονται στο μοντέλο είναι πιο μικρές από τις αντίστοιχες που λαμβάνουμε με τις υπόλοιπες prior, γεγονός που μας κάνει να τις θεωρήσουμε πιο αποδοτικές για την ανίχνευση του πραγματικού μοντέλου.

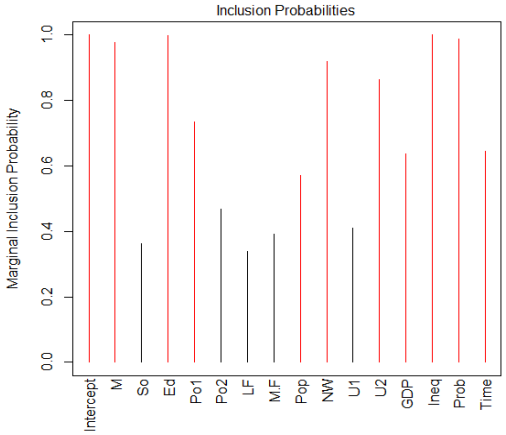
Επιλογή μεταβλητών με p(B != 0)>0.5 (Median Probability Model):

Στον πίνακα 5.2.13 έχουμε συγκεντρωτικά με * τις συμπεριλαμβανόμενες επεξηγηματικές μεταβλητές απαιτώντας **p(B != 0)>0.5 (Median Probability Model)**, για κάθε prior στο μοντέλο υψηλότερης εκ των υστέρων πιθανότητας.

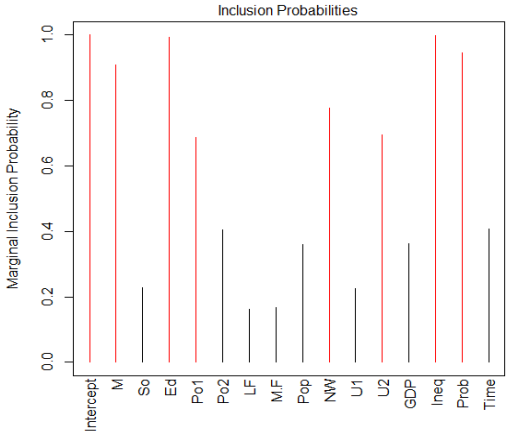
Πραγματικά δεδομένα	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
Intercept	*	*	*	*	*	*	*	*	*
M	*	*	*	*	*	*	*	*	*
S_0									
Ed	*	*	*	*	*	*	*	*	*
Po1	*	*	*	*	*	*	*	*	*
Po2			*						
LF									
M.F									
Pop	*								
NW	*	*	*	*	*	*	*	*	*
U1									
U2	*	*	*	*	*	*	*	*	*
GDP	*								
Ineq	*	*	*	*	*	*	*	*	*
Prob	*	*	*	*	*	*	*	*	*
Time	*								

Πίνακας 5.2.13. *** Στατιστικά σημαντικές επεξηγηματικές μεταβλητές για p(B != 0)>0.5**

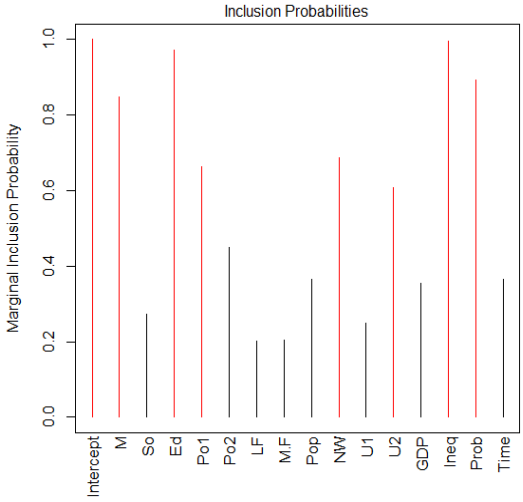
Η έως τώρα ανάλυση των εκ των υστέρων αποτελεσμάτων εκτείνεται με την παρουσίαση πρόσθετων διαγνωστικών ελέγχων.



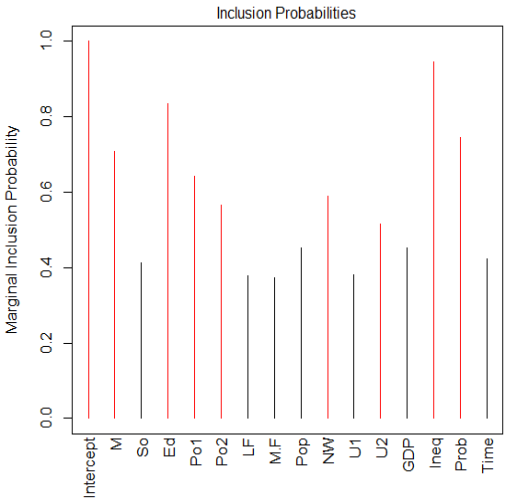
Σχήμα 1: Prior = AIC



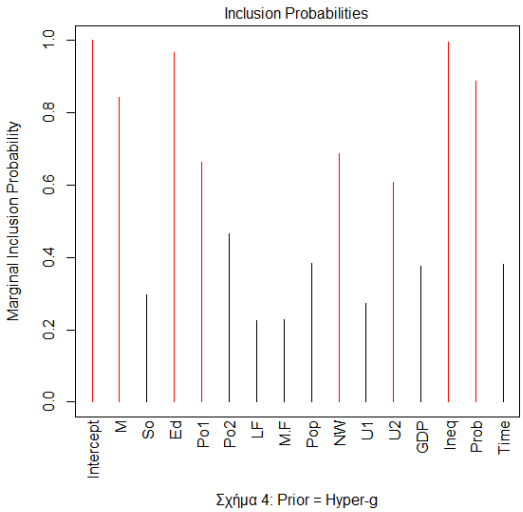
Σχήμα 2: Prior = BIC



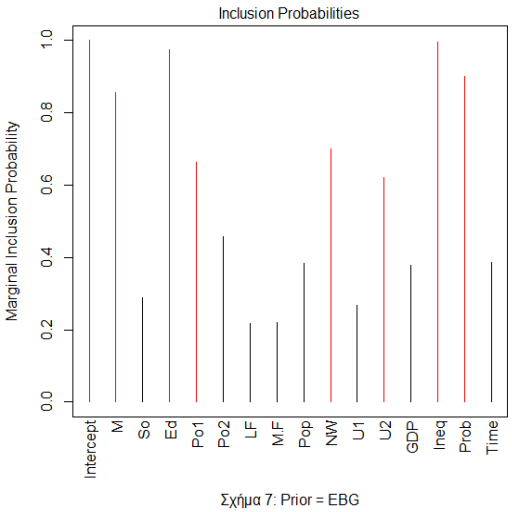
Σχήμα 5: Prior = Hyper-g/n



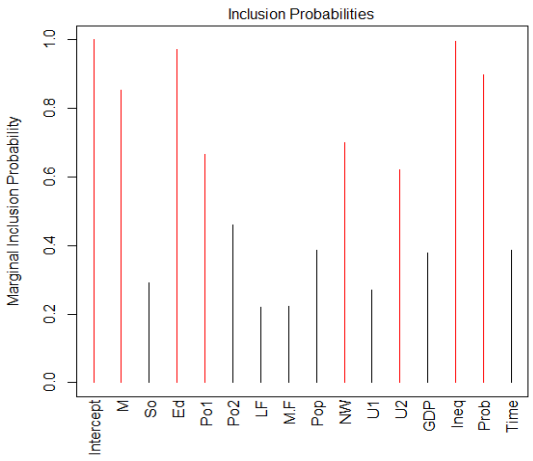
Σχήμα 3: Prior = g



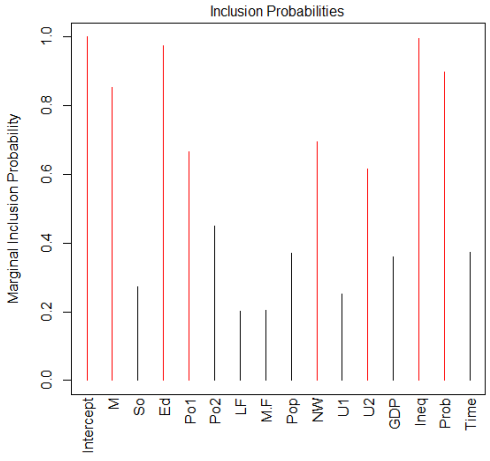
Σχήμα 4: Prior = Hyper-g



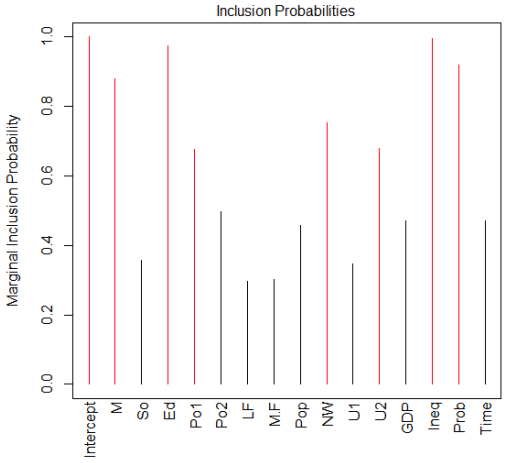
Σχήμα 7: Prior = EBG



Σχήμα 6: Prior = EBL



Σχήμα 8: Prior = ZS-null



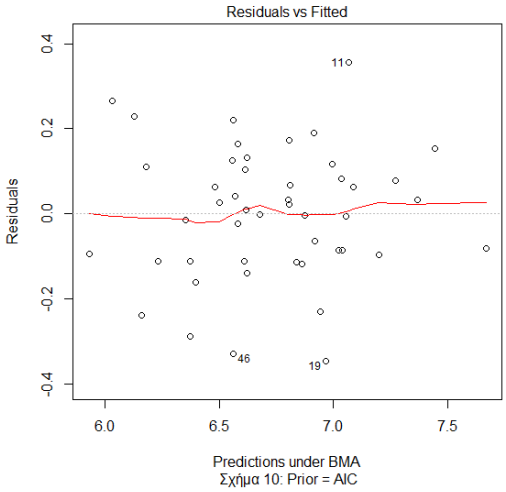
Σχήμα 9: Prior = ZS-full

Σχήματα 1-9: Εκ των υστέρων πιθανότητες ενσωμάτωσης των επεξ. μεταβλητών.

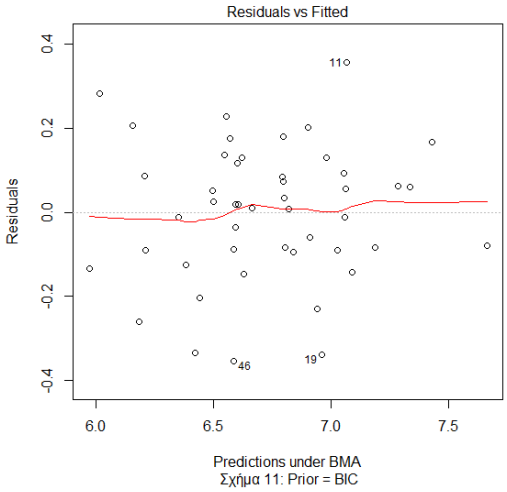
Σχήματα 1-9 - Παρατηρήσεις:

Στα σχήματα 1-9 επιβεβαιώνουμε γραφικώς τις παρατηρήσεις που διατυπώσαμε κατά την ανάλυση του πίνακα 5.2.12, εφιστώντας την προσοχή μας στην τάση που παρουσιάζεται σύμφωνα με τις prior AIC και ZS-full να απονέμουν (συγκριτικά με τις υπόλοιπες prior), μεγαλύτερη εκ των υστέρων πιθανότητα στη μη μηδενικότητα των συντελεστών που αντιστοιχούν στις επεξηγηματικές μεταβλητές που δεν είναι στατιστικά σημαντικές.

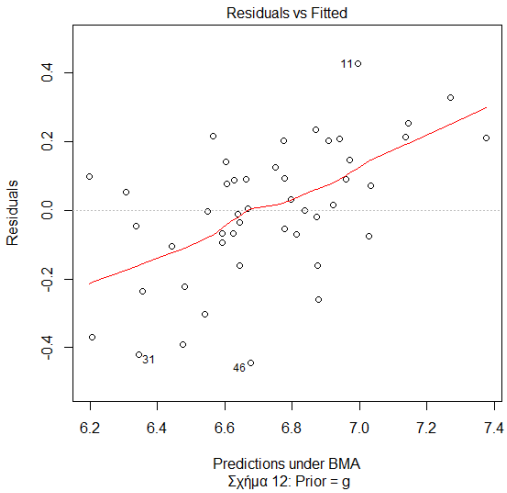
Εν συνεχεία παρουσιάζουμε τα υπόλοιπα με τις προσαρμοσμένες τιμές σε κάθε ξεχωριστή περίπτωση, απ τα οποία θα προσπαθήσουμε να βγάλουμε κάποια χρήσιμα συμπεράσματα για τη γραμμικότητα του μοντέλου, τη διασπορά των σφαλμάτων και τις ακραίες τιμές.



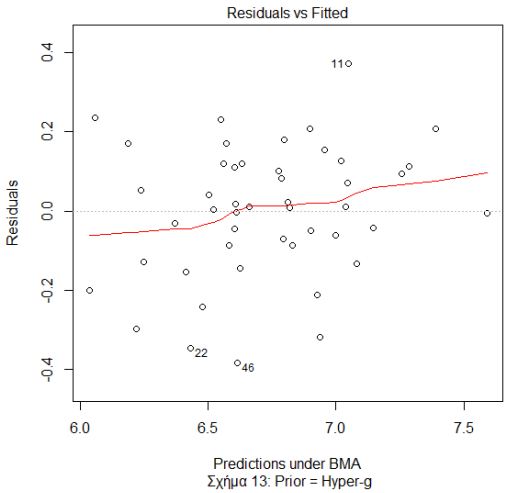
Σχήμα 10: Prior = AIC



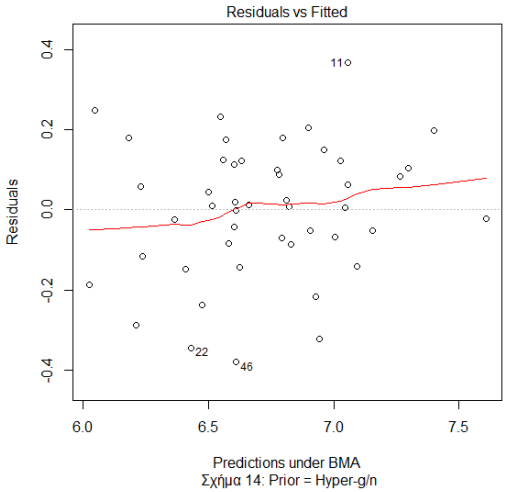
Σχήμα 11: Prior = BIC



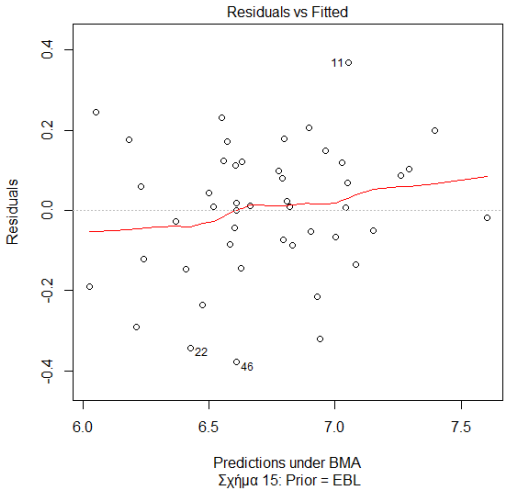
Σχήμα 12: Prior = g



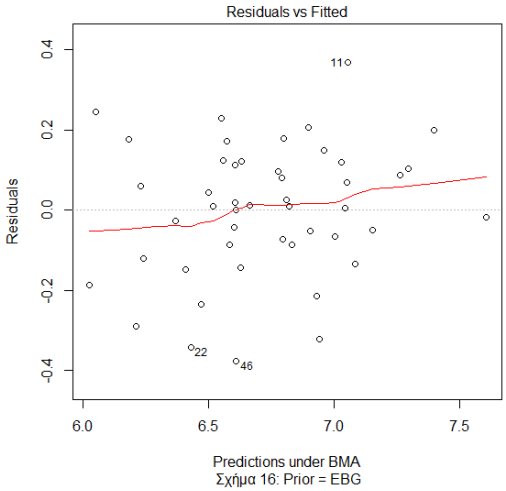
Σχήμα 13: Prior = Hyper-g

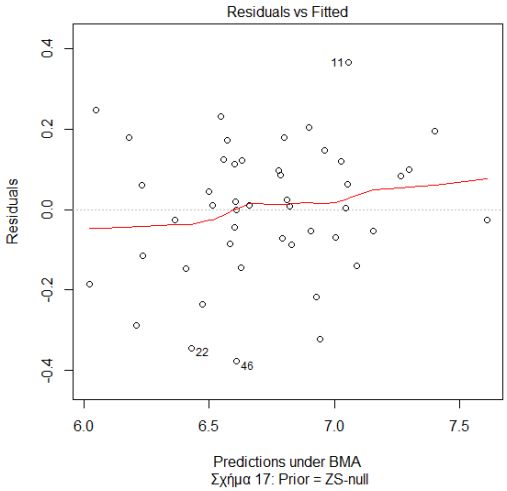


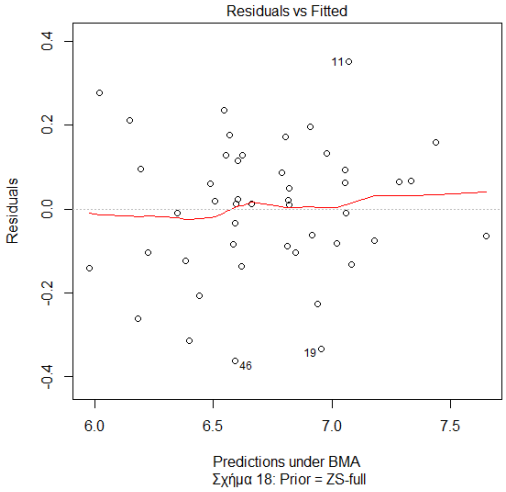
Σχήμα 14: Prior = Hyper-g/n



Σχήμα 15: Prior = EBL







Σχήματα 10 – 18: Υπόλοιπα με προσαρμοσμένες τιμές.

Παρατηρήσεις:

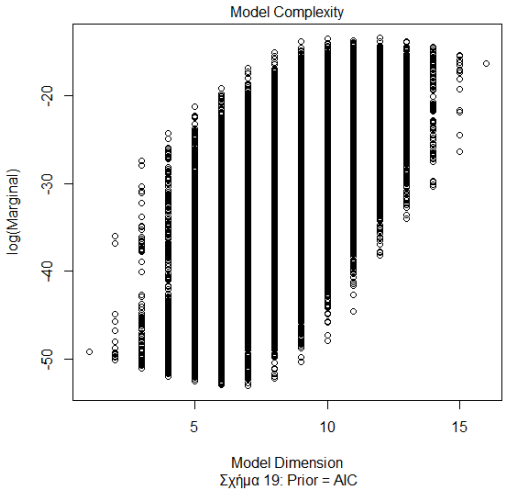
Στα σχήματα 10-18 βλέπουμε τα υπόλοιπα σε σχέση με τις προσαρμοσμένες τιμές.

Σε όλες τις περιπτώσεις (εκτός στην περίπτωση της prior g (σχήμα 12) όπου παρατηρούμε μια αύξουσα συμπεριφορά των υπολοίπων), μπορούμε να πούμε ότι η υπόθεση της γραμμικότητας φαίνεται λογική, καθώς τα υπόλοιπα φαίνονται τυχαία κατανεμημένα γυρω από το 0, ενώ επίσης φαίνεται να ικανοποιείται και η υπόθεση της ομοσκεδαστικότητας.

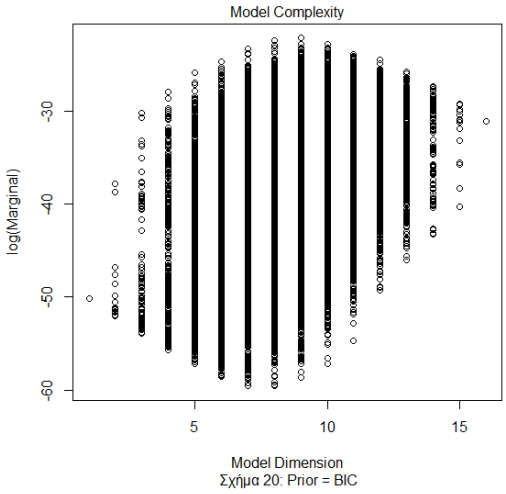
Επίσης ως προς τη συμπεριφορά των υπολοίπων σε σχέση με τις προσαρμοσμένες τιμές μπορούμε να διαχωρίσουμε τις επιλεγμένες prior σε δύο κατηγορίες. Στη μία μπορούμε να ενσωματώσουμε τις prior AIC, BIC και

ZS-full (σχήματα 10-11 και σχήμα 18), αφού και στις 3 αυτές περιπτώσεις ανιχνεύονται οι ίδιες ακραίες παρατηρήσεις ($11^η$, $19^η$ και $46^η$) και ταυτόχρονα παρατηρείται η ίδια συναρτησιακή μορφή στην καμπύλη, που τείνει να γίνει ευθεία γύρω από το μηδέν, ενώ στη άλλη κατηγορία μπορούμε να ενσωματώσουμε τις prior hyper-g, hyper-g/n, EBL, EBG και ZS-null αφού κάτω από τις συγκεκριμένες επιλογές έχουμε τις ίδιες ακραίες παρατηρήσεις ($11^η$, $22^η$ και $46^η$) και η αντίστοιχη καμπύλη στα σχήματα 13-17 να παρουσιάζει παραπλήσια μορφή.

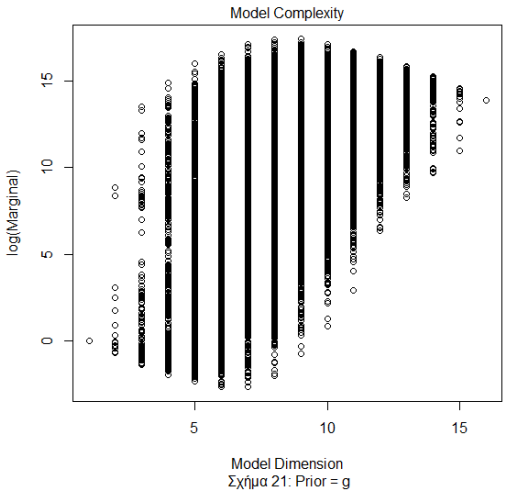
Συνοψίζοντας, φαίνεται και από τα εν λόγω σχήματα να επιβεβαιώνονται τα όσα ήδη έχουν ειπωθεί ως προς την προσαρμογή του ανιχνευθέντος μοντέλου και τη σύγκριση των επιλεγμένων prior.



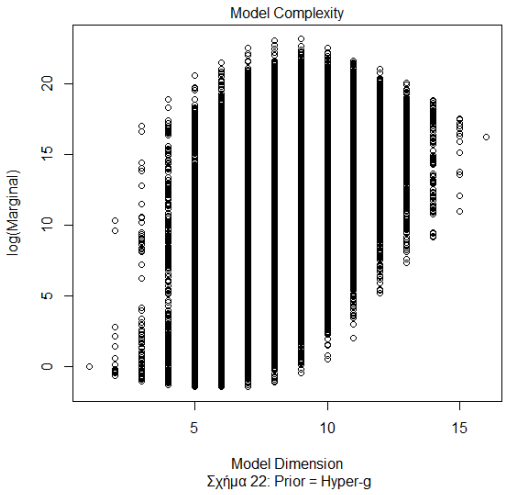
Σχήμα 19: Prior = AIC



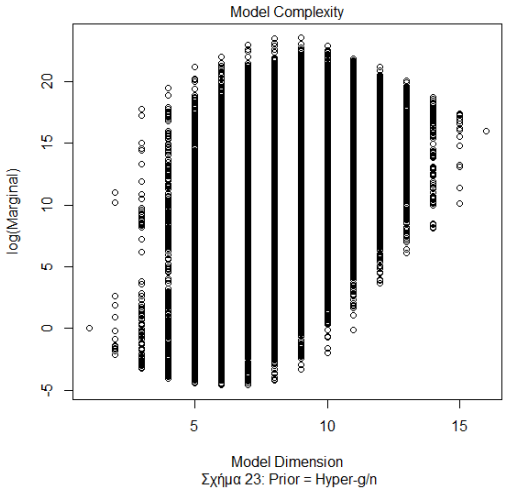
Σχήμα 20: Prior = BIC



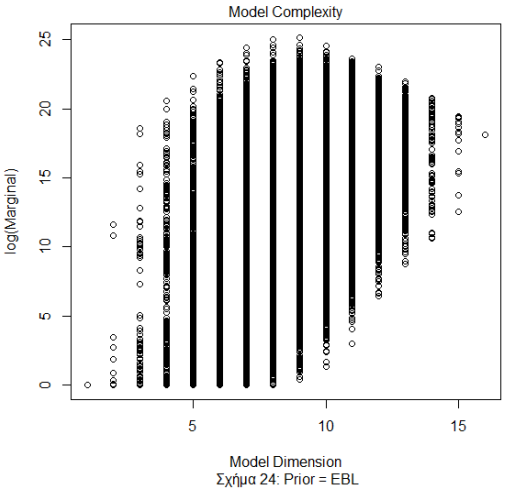
Σχήμα 21: Prior = g



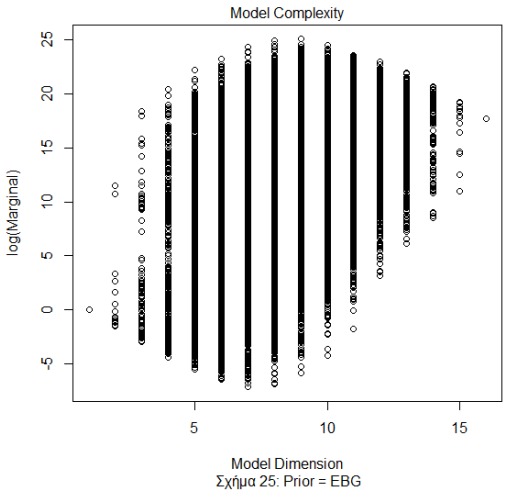
Σχήμα 22: Prior = Hyper-g



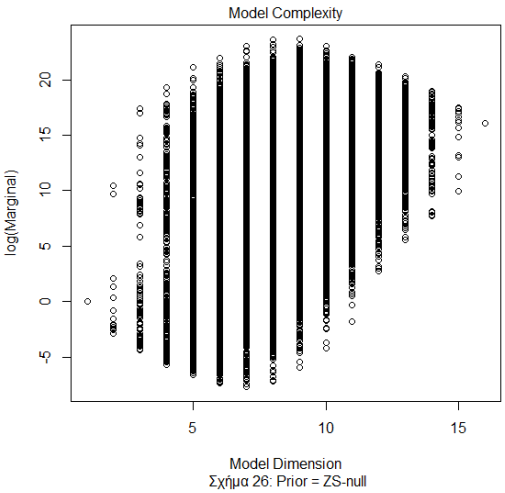
Σχήμα 23: Prior = Hyper-g/n



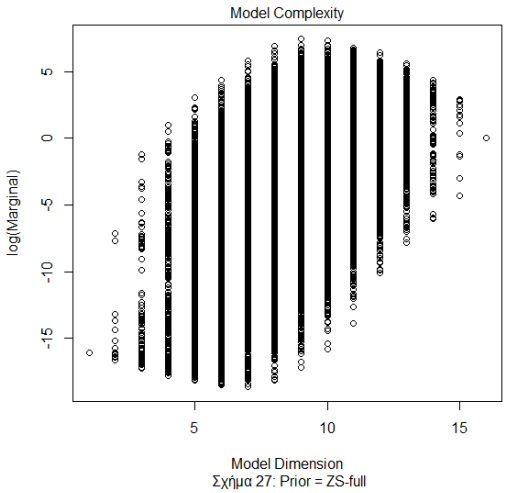
Σχήμα 24: Prior = EBL



Σχήμα 25: Prior = EBG



Σχήμα 26: Prior = ZS-null

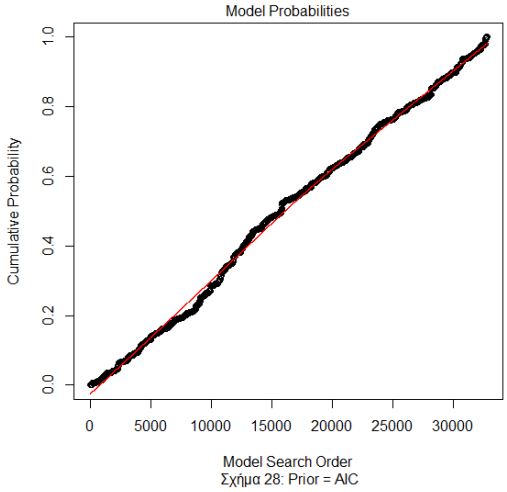


Σχήμα 27: Prior = ZS-full

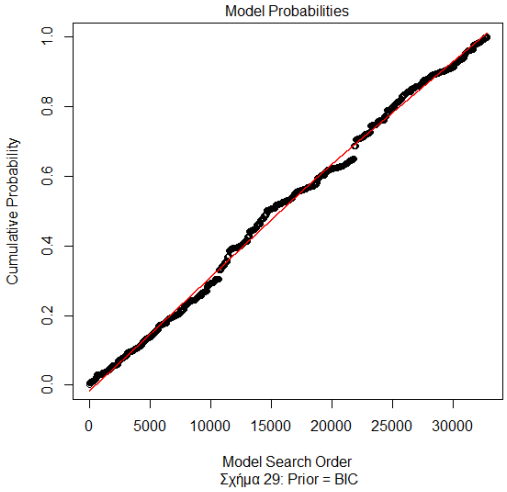
Σχήματα 19 -27: Πολυπλοκότητα μοντέλου.

Παρατηρήσεις

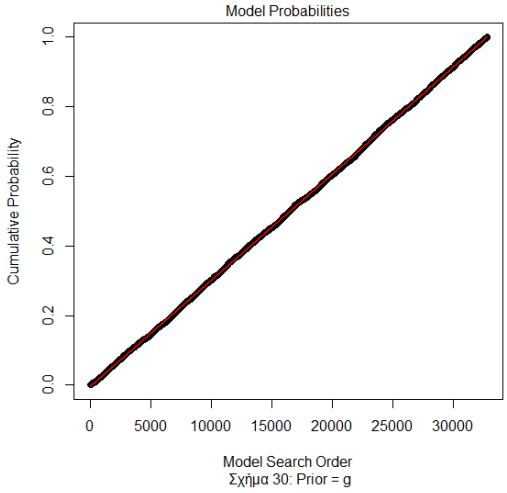
Στα σχήματα 19-27 παρατηρούμε πώς κυμαίνεται η τιμή της λογαριθμοποιημένης περιθώριας πιθανοφάνειας ανάλογα με την prior που έχουμε επιλέξει και ανάλογα με τη διάσταση του μοντέλου. Σύμφωνα με το κριτήριο AIC η τιμή αυτή μεγιστοποιείται όταν η διάσταση του μοντέλου είναι 12, ενώ σύμφωνα με όλες τις άλλες επιλογές εκ των προτέρων κατανομών η τιμή της λογαριθμοποιημένης πιθανοφάνειας μεγιστοποιείται όταν στο μοντέλο συμπεριλαμβάνονται 9 επεξηγηματικές μεταβλητές. Είναι φανερό ότι για τις επιλογές των prior BIC και AIC παρουσιάζονται πολύ μικρές τιμές για την λογαριθμοποιημένη πιθανοφάνεια ενώ ακολουθούν οι prior ZS-full και g με κάπως υψηλότερες τιμές και στη συνέχεια οι prior hyper-g, hyper-g/n, ZS-null, EBG και EBL με τις υψηλότερες τιμές λογαριθμοποιημένης πιθανοφάνειας για το μοντέλο που συμπεριλαμβάνει 9 επεξηγηματικές μεταβλητές και με πολύ μικρές διαφοροποιήσεις μεταξύ τους.



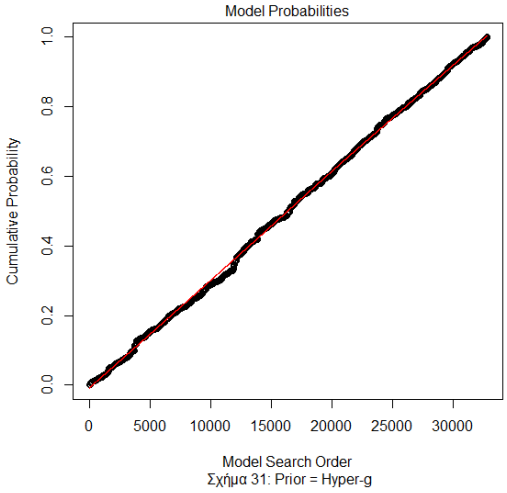
Σχήμα 28: Prior = AIC



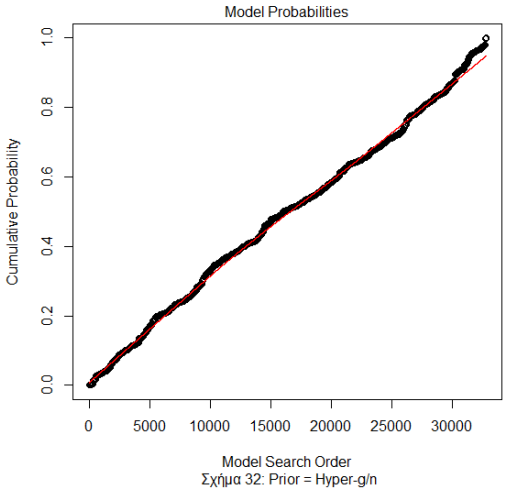
Σχήμα 29: Prior = BIC

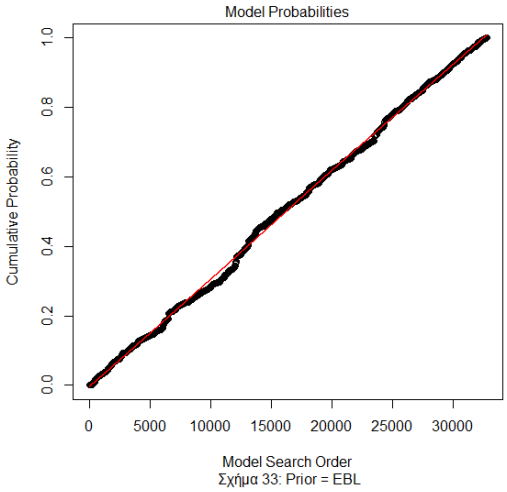


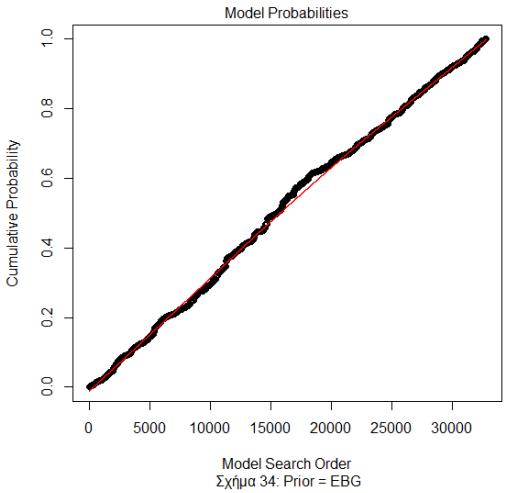
Σχήμα 30: Prior = g

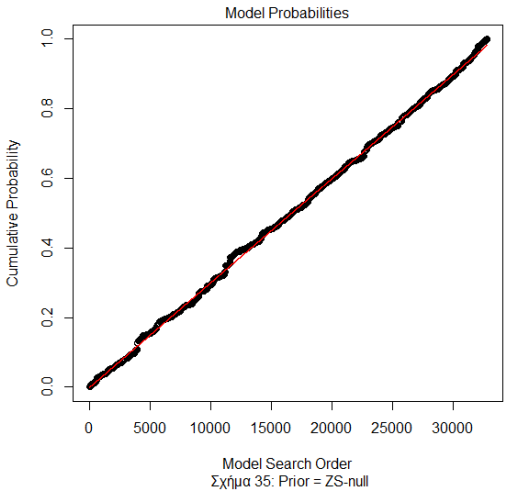


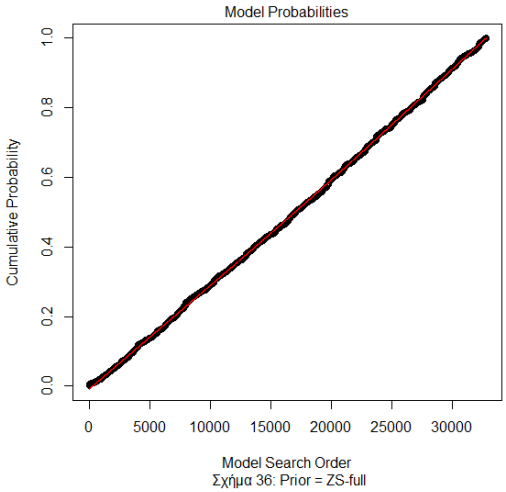
Σχήμα 31: Prior = Hyper-g







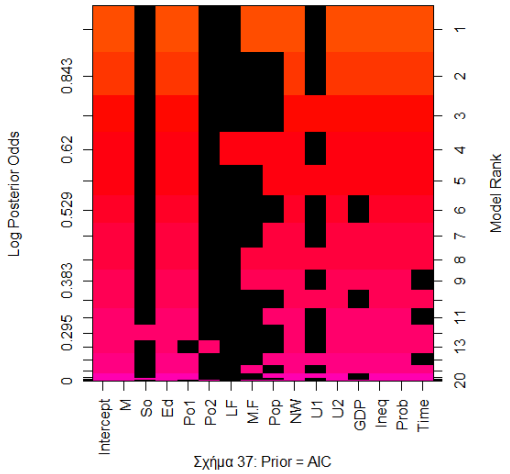




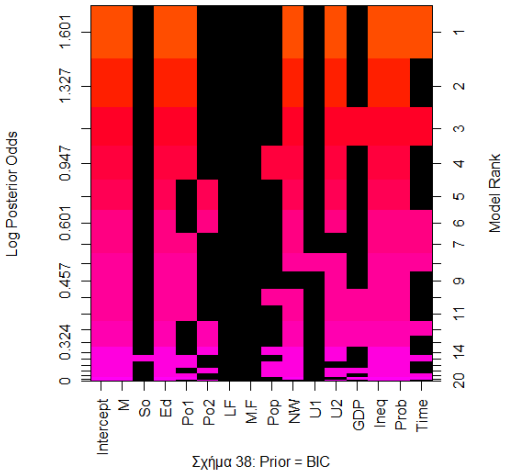
Σχήματα 28 -36: Πιθανότητα μοντέλου.

Παρατηρήσεις:

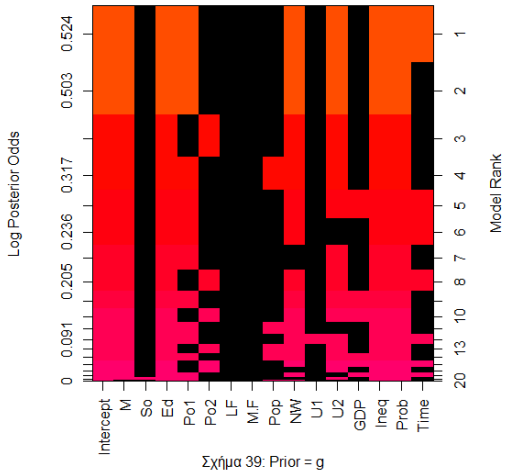
Στα σχήματα 28-36 παρατηρούμε ότι ανεξάρτητα από την prior που χρησιμοποιούμε η πιθανότητα του μοντέλου αυξάνεται συσσωρευτικά καθώς προστίθονται νέα μοντέλα κατά τη δειγματοληψία.



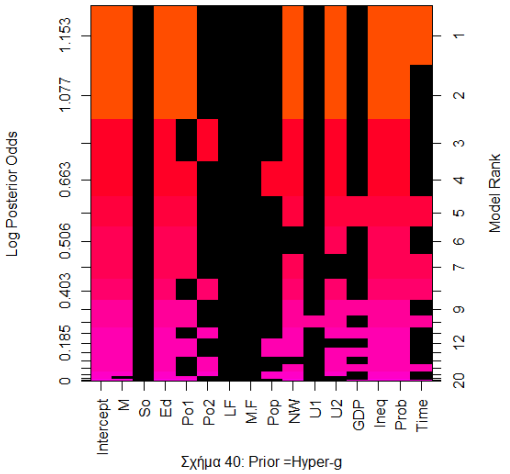
Σχήμα 37: Prior = AIC



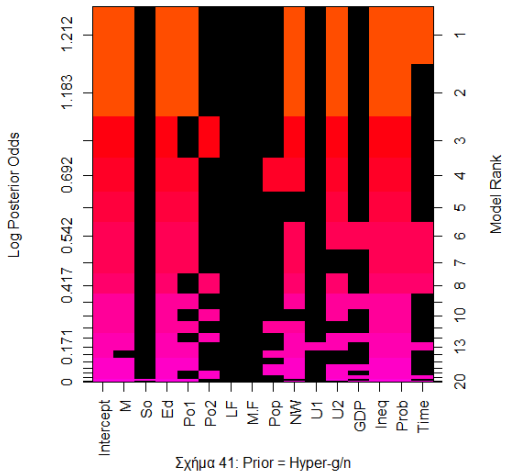
Σχήμα 38: Prior = BIC



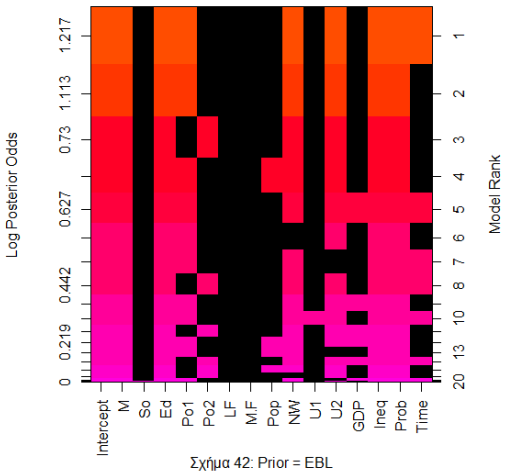
Σχήμα 39: Prior = g



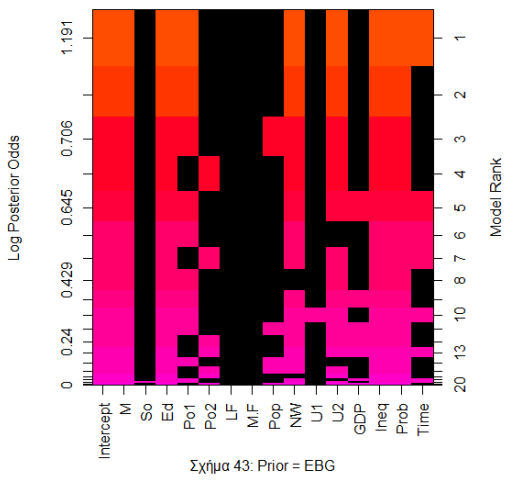
Σχήμα 40: Prior =Hyper-g



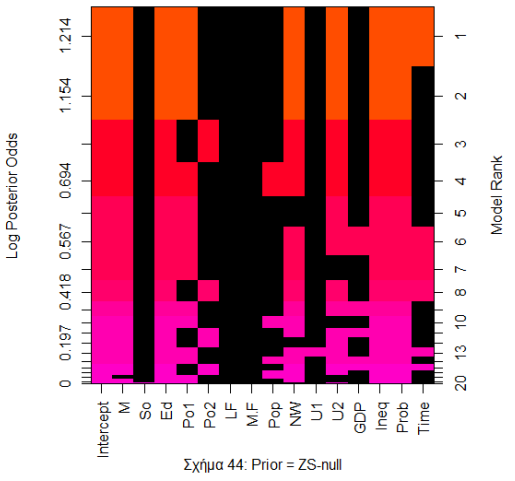
Σχήμα 41: Prior = Hyper-g/n



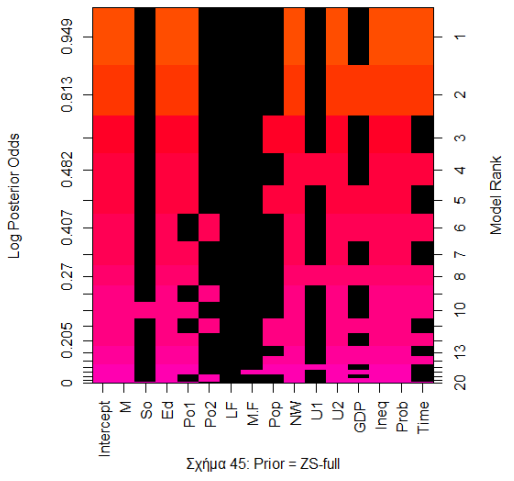
Σχήμα 42: Prior = EBL



Σχήμα 43: Prior = EBG



Σχήμα 44: Prior = ZS-null



Σχήμα 45: Prior = ZS-full

Σχήματα 37-45: Απεικόνιση στο χώρο των μοντέλων.

Παρατηρήσεις

Στα σχήματα 37-45, παρουσιάζουμε μια απεικόνιση στο χώρο των μοντέλων για κάθε prior ξεχωριστά, η οποία περιλαμβάνει τα 20 πρώτα μοντέλα(άξονας y δεξιά), με τις συμπεριλαμβανόμενες και τις μη συμπεριλαμβανόμενες επεξηγηματικές μεταβλητές στο καθένα (άξονας x) καθώς και την τιμή του λογαρίθμου του εκ των υστέρων λόγου πιθανοτήτων για το κάθε μοντέλο (άξονας y αριστερά). Με μαύρο απεικονίζονται οι εξαιρούμενες από το μοντέλο επεξηγηματικές μεταβλητές ενώ με οποιοδήποτε άλλο χρώμα οι επεξηγηματικές μεταβλητές που συμπεριλαμβάνονται σε αυτό. Όσες από τις επεξηγηματικές μεταβλητές ανήκουν στο ίδιο χρώμα σημαίνει ότι συμπεριλαμβάνονται στο ίδιο μοντέλο, με την αντίστοιχη εκ των υστέρων πιθανότητα.

Με τις παραπάνω απεικονίσεις στο χώρο των μοντέλων ολοκληρώνουμε τους γραφικούς διαγνωστικούς ελέγχους καλής προσαρμογής, αναδεικνύοντας και πάλι τις hyper-g, hyper-g/n, EBL, EBG και ZS-null ως πιο αποδοτικές έναντι των AIC, BIC και ZS-full prior. Συγκεκριμένα βλέπουμε και εδώ ότι για τα 20 πρώτα μοντέλα ο εκ των υστέρων λόγος πιθανοτήτων λαμβάνει μικρότερες τιμές βάση των prior AIC, ZS-full και g έναντι των υπολοίπων prior. Οι τιμές αυτές παρουσιάζονται να είναι μεγαλύτερες με βάση την prior BIC αλλά ωστόσο έχουν αναφερθεί κάποια μειονεκτήματα της εν λόγω prior στις παρατηρήσεις του πίνακα 5.2.11 και επομένως απομένει να παρατηρήσουμε ότι οι τιμές του εκ των υστέρων λόγου των πιθανοτήτων εξετάζοντας τις αποδοτικότερες prior, κατατάσσονται σε αύξουσα σειρά ως εξής:

- $\log \text{Posterior Odds}_{\text{hyper}-\text{g}}$
- $\log \text{Posterior Odds}_{\text{EBG}}$
- $\log \text{Posterior Odds}_{\text{hyper}-\frac{\text{g}}{\text{n}}}$
- $\log \text{Posterior Odds}_{\text{ZS}-\text{null}}$
- $\log \text{Posterior Odds}_{\text{EBL}}$.

Εώς τώρα παρουσιάσαμε την υπό μελέτη Μπεϋζιανή επιλογή μοντέλου και μεταβλητών σε πραγματικά δεδομένα (μικρού δείγματος) και προβήκαμε σε μία πρώτη απόπειρα να διερευνήσουμε την επιρροή των εκ των προτέρων κατανομών που χρησιμοποιήσαμε για τους συντελεστές του μοντέλου, στα εκ των υστέρων αποτελέσματα. Η ανάλυση θα επαναληφθεί σε προσομοιωμένα δεδομένα, παρεμβάλλοντας 2 ακόμα παραδείγματα πριν την εξαγωγή των τελικών συμπερασμάτων, στα οποία θα χρησιμοποιηθεί μεγαλύτερο μέγεθος δείγματος και ευρύτερο πλήθος παραμέτρων.

Στην αμέσως επόμενη ενότητα ακολουθεί πρώτα μία σύντομη περιγραφή για τα δεδομένα του παραδείγματος 2, για το οποίο ο σχετικός κώδικας R εντάσσεται στο παράρτημα Β και κατόπιν πραγματοποιείται η αντίστοιχη σύγκριση των εκ των προτέρων κατανομών που μελετήσαμε στα πραγματικά δεδομένα.

5.3 <u>ΠΑΡΑΔΕΙΓΜΑ 2</u>:

Μπεϋζιανή επιλογή μοντέλου και μεταβλητών σε προσομοιωμένα δεδομένα με ανεξάρτητες επεξηγηματικές μεταβλητές

Στο παρόν παράδειγμα η επιλογή μοντέλου που θα πραγματοποιήσουμε θα στηριχτεί σε δείγμα μεγέθους $\mathrm{n} = 100$ προσομοιωμένων τιμών.

Υποθέτουμε ότι έχουμε $\mathrm{p} = 20$ επεξηγηματικές μεταβλητές οι οποίες είναι μεταξύ τους γραμμικώς ανεξάρτητες και προσομοιώνουμε για αυτές τιμές από την τυποποιημένη κανονική κατανομή $\mathrm{N}(0,1)$, ενώ επίσης υποθέτουμε ότι η μεταβλητή απόκρισης Y σχετίζεται γραμμικά με τις παραπάνω επεξηγηματικές μεταβλητές, δηλαδή ότι έχουμε ένα πολλαπλό γραμμικό μοντέλο της μορφής

$$\mathrm{Y} = \beta_0 + \beta_1 \mathrm{X}_1 + \cdots + \beta_\mathrm{p} \mathrm{X}_\mathrm{p} + \boldsymbol{\varepsilon}.$$

Θέτουμε $\beta_0 = 3$ & $\boldsymbol{\beta} = (1, 0, 1.8, 0, 0, 0, 2, 0, 0, 0, 0, 0, 1.5, 0, 0, 0, 0, 0, 0,0)$ και προσομοιώνουμε τιμές για τη μεταβλητή απόκρισης Y ούτως ώστε το πραγματικό μοντέλο να είναι της μορφής:

$$\mathrm{Y_i} \sim \mathrm{N}\left(\beta_0 + \mathrm{X_{i1}} + 1.8\,\mathrm{X_{i3}} + 2\mathrm{X_{i7}} + 1.5\mathrm{X_{i13}}, 2.5\right), \quad \mathrm{i} = 1, \ldots, 100.$$

Εν συνεχεία εκτελούμε στην R τις εντολές που χρησιμοποιήσαμε και στο παράδειγμα 1 με τα πραγματικά δεδομένα, θέτοντας ομοιόμορφη εκ των προτέρων κατανομή στο χώρο των μοντέλων ώστε αυτά να θεωρηθούν ισοπίθανα και ώστε επιπρόσθετα με τα σχόλια που αναφέραμε στο προηγούμενο παράδειγμα, να υπεισέλθουν στην ανάλυσή μας περισσότερα στοιχεία που αφορούν τη συνέπεια των εκ των προτέρων κατανομών που μελετάμε, ως λογικά συμπεράσματα των πολλαπλών επαναλήψεων της Μπεϋζιανής επιλογής μοντέλου (50 επαναλήψεις) που θα εφαρμόσουμε στα δεδομένα που προέρχονται εκάστως από την προσομοίωση που περιγράψαμε στην αρχή της παραγράφου.

Ακολουθεί μια σειρά από συγριτικούς πίνακες για την εκάστοτε prior.

Ποσοστό ενσωμάτωσης της εκάστοτε επεξηγηματικής μεταβλητής στο HPM (Highest Probability Model)

ποσοστό %	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
Intercept	100	100	100	100	100	100	100	100	100
X1	98	98	88	96	98	98	98	100	100
X2	24	12	8	8	4	4	4	8	6
X3	100	100	100	100	100	100	100	100	100
X4	18	6	4	6	2	2	6	4	8
X5	14	2	4	2	4	8	4	4	12
X6	16	2	2	10	4	6	8	8	4
X7	100	100	100	100	100	100	100	100	100
X8	22	4	2	12	0	8	4	4	14
X9	12	6	2	6	2	6	6	2	8
X10	10	4	2	6	0	8	2	2	10
X11	20	6	0	12	0	10	14	8	8
X12	26	6	0	2	6	6	16	10	18
X13	100	100	100	100	100	100	100	98	100
X14	26	2	4	10	10	8	4	4	2
X15	12	6	2	4	0	10	12	8	16
X16	24	8	2	0	0	8	8	2	8
X17	20	2	2	6	4	4	4	10	16
X18	22	0	4	8	4	6	8	2	14
X19	6	0	2	6	8	4	2	0	12
X20	26	8	4	10	10	2	12	4	10

Μέση διάσταση μοντέλου (για όλες τις επαναλήψεις)

	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
dim	8	6	7	6	6	6	6	6	7

Πίνακας 5.3.1 50 επαναλήψεις του παραδείγματος 2 σε προσομοιωμένα δεδομένα (με ανεξαρτησία)

– Πραγματικό μοντέλο : $Y_i \sim N(3 + X_1 + 1.8\,X_3 + 2X_7 + 1.5X_{13}, 2.5), \quad i = 1, \dots, 100.$

Πίνακας 5.3.1 – Παρατηρήσεις:

Σε κάθε επανάληψη του αλγορίθμου (κώδικας παραρτήματος Β) για την εύρεση του προτεινόμενου πρώτου μοντέλου στην κλάση των υψηλότερης εκ των υστέρων πιθανότητας μοντέλων που παράγουμε με την εκάστοτε prior, κάθε επεξηγηματική μεταβλητή είτε ενσωματώνεται στο μοντέλο είτε όχι.

Στον πίνακα 5.3.1 παρουσιάζουμε την συχνότητα με την οποία κάθε επεξηγηματική μεταβλητή συμπεριλαμβάνεται στο πρώτο προτεινόμενο μοντέλο για όλες τις επαναλήψεις καθώς και τη μέση διάσταση του ενδεδειγμένου μοντέλου, όπου με χρώμα τονίζουμε εκείνες τις επεξηγηματικές μεταβλητές που συνθέτουν το πραγματικό μοντέλο και επιθυμούμε στα εκ των

υστέρων αποτελέσματα που θα δούμε παρακάτω να διακριθούν ως οι πιο στατιστικά σημαντικές μεταβλητές ή ισοδύναμα ως οι μεταβλητές με τη μεγαλύτερη εκ των υστέρων πιθανότητα συμπερίληψης.

Στον εν λόγω διαθέσιμο πίνακα παρατηρούμε ότι με όλες τις επιλογές των εκ των προτέρων κατανομών που μελετάμε οδηγούμαστε σε σχεδόν 100% συμπερίληψη των επεξηγηματικών μεταβλητών που συνθέτουν το πραγματικό μοντέλο, αφού και στις 50 επαναλήψεις του αλγορίθμου οι συγκεκριμένες μεταβλητές ενσωματώνονται 49 και 50 φορές.

Επιπρόσθετα παρατηρούμε ότι οι μεταβλητές που δεν θα έπρεπε να συμπεριλαμβάνονται στην εκάστοτε επανάληψη στο προτεινόμενο μοντέλο, (αφού είναι οι μεταβλητές εκείνες που δεν συμμετάσχουν στο πραγματικό), υπό του κριτηρίου AIC και της prior ZS-full έχουν μεγαλύτερη συχνότητα εμφάνισης συγκριτικά με τις υπόλοιπες prior, γεγονός που μπορεί να θεωρηθεί ως μειονέκτημα των εν λόγω εκ των προτέρων κατανομών έναντι των υπολοίπων. Ως προς την ίδια παρατήρηση ακολουθούν οι εμπερικές Bayes prior, η hyper-g, η BIC και η g με μικρές μεταξύ τους διαφοροποιήσεις ως προς τη συχνότητα εμφάνισης των μη στατιστικά σημαντικών μεταβλητών, ενώ είναι οφθαλμοφανές ότι η εν λόγω συχνότητα είναι αρκετά μικρή στις περιπτώσεις επιλογής των hyper-g/n και ZS-null κάτι το οποίο σε συνδυασμό με τις θεωρητικές ιδιότητες που περιγράψαμε στο Κεφάλαιο 4 μας οδηγεί στο να διεξάγουμε και εμπειρικά το συμπέρασμα ότι οι συγκεκριμένες επιλογές εκ των προτέρων κατανομών πλεονεκτούν των υπολοίπων.

Τέλος, ως προς τη διάσταση του ενδεδειγμένου μοντέλου υψηλότερης εκ των υστέρων πιθανότητας με την εκάστοτε prior, παρατηρούμε ότι με τις επιλογές των AIC, g και ZS-full επικρατεί η τάση να επιλέγονται μοντέλα με περισσότερες επεξηγηματικές μεταβλητές από αυτές που πραγματικά χρειάζονται για την ερμηνεία της μεταβλητής απόκρισης.

Ακολούθως θα μελετήσουμε εκτενέστερα τις υπό συζήτηση εκ των προτέρων κατανομές για τους συντελεστές, εξετάζοντας τη μέση τιμή των εκτιμήσεων των ενδιαφερομένων παραμέτρων που παρουσιάζουμε στον πίνακα 5.3.2 κατά την ανίχνευση του πραγματικού μοντέλου στις 50 επαναλήψεις.

Ανίχνευση του πραγματικού μοντέλου στις 50 επαναλήψεις

50 επαναλήψεις	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
Αριθμός εμφάνισης του πραγματικού μοντέλου	2	26	37	16	27	24	15	23	7

Μέση τιμή των εκ των υστέρων μέσων των ενδιαφερόμενων παραμέτρων κατά την ανίχνευση του πραγματικού μοντέλου στις 50 επαναλήψεις

	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
Prob(true model)	0.002	0.07	0.07	0.01	0.04	0.01	0.02	0.03	0.00
Log(marginal likelihood)	-324.3	-331.0	39.91	40.40	1132.6	38.67	39.84	39.61	19.25
Intercept	3.42	2.98	3.09	3.22	3.09	3.07	2.96	2.97	3.37
X1	0.89	0.99	0.85	1.01	0.98	0.88	0.95	0.99	0.88
X2	-0.10	-0.03	0.00	0.01	0.01	0.04	-0.03	0.01	0.02
X3	1.72	1.71	1.74	1.67	1.82	1.64	1.62	1.82	1.66
X4	0.05	0.01	-0.02	0.00	0.01	0.00	0.00	0.02	-0.01
X5	0.02	-0.02	0.00	0.02	-0.02	-0.01	0.01	0.03	-0.03
X6	-0.02	-0.01	-0.01	0.04	-0.01	-0.01	0.02	0.00	-0.01
X7	2.14	1.94	1.95	1.88	1.96	1.86	2.00	1.86	2.00
X8	-0.01	0.00	0.01	-0.01	0.03	-0.01	0.00	-0.01	-0.01
X9	-0.05	-0.01	0.00	-0.01	0.01	0.01	0.02	-0.01	0.02
X10	-0.04	0.00	-0.01	0.02	0.01	0.02	0.00	0.01	-0.04
X11	0.01	0.02	0.00	0.01	0.02	0.00	-0.01	-0.01	0.03
X12	0.04	0.01	0.00	-0.02	-0.02	-0.01	0.01	0.00	-0.01
X13	1.40	1.41	1.44	1.46	1.46	1.44	1.43	1.45	1.24
X14	0.01	0.01	-0.01	0.01	0.02	0.03	0.00	0.02	-0.01
X15	0.00	0.00	0.01	0.00	-0.02	0.00	0.02	0.02	-0.03
X16	-0.01	0.01	-0.01	-0.06	0.00	0.01	-0.02	0.02	0.04
X17	-0.04	-0.02	0.00	0.01	0.01	-0.01	0.04	0.01	0.04
X18	0.02	-0.02	-0.02	-0.03	-0.01	0.02	0.00	-0.01	0.02
X19	0.04	-0.02	0.02	0.01	-0.01	-0.03	-0.02	0.01	0.00
X20	0.01	0.01	0.01	0.00	0.00	-0.03	0.01	-0.01	-0.04

Πίνακας 5.3.2 **50 επαναλήψεις του παραδείγματος 2 σε προσομοιωμένα δεδομένα (με ανεξαρτησία)**

Πίνακας 5.3.2 – Παρατηρήσεις:

Στον πίνακα 5.3.2 παρατηρούμε ότι με βάση το κριτήριο AIC και την prior ZS-full ο αριθμός ανίχνευσης του πραγματικού μοντέλου στις 50 επαναλήψεις που αναφέραμε για το παράδειγμα 2, είναι αρκετά μικρότερος έναντι του αριθμού ανίχνευσης του πραγματικού μοντέλου που λαμβάνουμε με τις υπόλοιπες prior.

Επιπρόσθετα οι εκτιμήσεις που λαμβάνουμε υπό του κριτηρίου AIC και της ZS-full prior για τους συντελεστές του πραγματικού μοντέλου κατά την ανίχνευσή του στις 50 επαναλήψεις, δεν είναι τόσο κοντά στις πραγματικές τιμές όπως συμβαίνει με τις υπόλοιπες prior.

Υπό του κριτηρίου BIC και των prior g, EBL, EBG, hyper-g, hyper-g/n και ZS-null είναι φανερό (με αμελητέες διαφοροποιήσεις και καλύτερα εκ των υστέρων αποτελέσματα βάση των hyper-g/n και ZS-null) ότι το πραγματικό μοντέλο ανιχνεύεται περισσότερες φορές και η εκ των υστέρων πιθανότητα που του αποδίδεται είναι μεγαλύτερη από την αντίστοιχη που παρατηρούμε βάση του κριτηρίου AIC και της prior ZS-full.

Με την εως τώρα ανάλυση φαίνεται τα συμπεράσματα που λαμβάνουμε εμπειρικά να συνάδουν με όσα αναφέρθηκαν στα θεωρήματα του Κεφαλαίου 4 κατά τη μελέτη των ασυμπτωτικών ιδιοτήτων και της συνέπειας του μοντέλου.

Συνεχίζουμε εμπειρικά τις παρατηρήσεις μας και τα εξαγώμενα συμπεράσματα θα αναπτυχθούν αναλυτικότερα στο κεφάλαιο 6.

Μέση τιμή των εκ των υστέρων τυπικών αποκλίσεων των ενδιαφερόμενων παραμέτρων κατά την ανίχνευση του πραγματικού μοντέλου στις 50 επαναλήψεις

50 επαναλήψεις	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
Prob(true model)	0.00	0.02	0.03	0.01	0.02	0.01	0.01	0.02	0.00
Log(marginal likelihood)	9.15	8.46	7.47	7.46	6.35	5.00	6.53	5.00	1.21
Intercept	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.24	0.25
X1	0.31	0.29	0.29	0.27	0.27	0.29	0.28	0.27	0.31
X2	0.21	0.14	0.13	0.15	0.14	0.17	0.16	0.14	0.20
X3	0.26	0.27	0.26	0.25	0.26	0.26	0.27	0.26	0.27
X4	0.17	0.13	0.14	0.15	0.13	0.15	0.16	0.14	0.15
X5	0.16	0.13	0.14	0.16	0.15	0.15	0.16	0.18	0.18
X6	0.16	0.12	0.12	0.17	0.15	0.19	0.16	0.14	0.13
X7	0.27	0.27	0.27	0.26	0.26	0.27	0.26	0.26	0.27
X8	0.14	0.12	0.14	0.18	0.16	0.15	0.13	0.15	0.19
X9	0.19	0.15	0.13	0.14	0.16	0.16	0.17	0.15	0.16
X10	0.16	0.14	0.14	0.15	0.14	0.17	0.17	0.14	0.17
X11	0.21	0.16	0.15	0.16	0.15	0.16	0.16	0.16	0.16
X12	0.15	0.12	0.12	0.17	0.15	0.17	0.15	0.17	0.18
X13	0.29	0.27	0.27	0.25	0.25	0.26	0.26	0.26	0.28
X14	0.16	0.13	0.12	0.18	0.15	0.16	0.15	0.14	0.14
X15	0.16	0.12	0.13	0.15	0.14	0.17	0.14	0.16	0.18
X16	0.14	0.13	0.13	0.18	0.14	0.15	0.14	0.16	0.16
X17	0.19	0.11	0.13	0.17	0.15	0.15	0.16	0.16	0.18
X18	0.15	0.14	0.15	0.15	0.14	0.17	0.14	0.14	0.18
X19	0.17	0.14	0.13	0.18	0.13	0.16	0.14	0.15	0.16
X20	0.16	0.13	0.12	0.18	0.15	0.16	0.14	0.14	0.17

Πίνακας 5.3.3 **50 επαναλήψεις του παραδείγματος 2 σε προσομοιωμένα δεδομένα (με ανεξαρτησία)**

Πίνακας 5.3.3 – Παρατηρήσεις:

Στον πίνακα 5.3.3 παρατηρούμε σε τι τιμές κυμαίνονται κατά μέσο όρο τα τυπικά σφάλματα των υπό εξέταση μεταβλητών κατά την ανίχνευση του πραγματικού μοντέλου με την εκάστοτε prior, όπου αντιλαμβανόμαστε κυρίως για τις επεξηγηματικές μεταβλητές που το αφορούν (δηλαδή τις στατιστικά σημαντικές) ότι αυτά είναι μεγαλύτερα για τις περιπτώσεις των AIC, ZS-full prior και BIC. Υπό των prior hyper-g/n και ZS-null είναι ορατό ότι έχουμε τα μικρότερα τυπικά σφάλματα ενώ οι τιμές αυτών κυμαίνονται κάπως υψηλότερα στις περιπτώσεις των g, hyper-g και EB-prior, χωρίς όμως ουσιαστικές διαφοροποιήσεις.

Μέση τιμή των εκ των υστέρων πιθανοτήτων p(B != 0) κατά την ανίνευση του πραγματικού μοντέλου στις 50 επαναλήψεις

50 επαναλήψεις	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
Intercept	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
X1	0.94	0.93	0.86	0.99	0.97	0.92	0.98	0.95	0.91
X2	0.38	0.16	0.16	0.24	0.19	0.27	0.25	0.20	0.33
X3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
X4	0.31	0.15	0.16	0.24	0.18	0.23	0.24	0.20	0.27
X5	0.31	0.15	0.15	0.24	0.21	0.23	0.24	0.27	0.32
X6	0.30	0.14	0.13	0.30	0.21	0.31	0.25	0.21	0.25
X7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
X8	0.28	0.13	0.16	0.29	0.24	0.24	0.20	0.22	0.30
X9	0.35	0.19	0.15	0.21	0.23	0.26	0.25	0.21	0.27
X10	0.30	0.17	0.17	0.23	0.21	0.27	0.27	0.20	0.30
X11	0.35	0.19	0.17	0.25	0.23	0.25	0.23	0.23	0.27
X12	0.30	0.13	0.14	0.27	0.22	0.27	0.23	0.24	0.30
X13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
X14	0.28	0.13	0.13	0.29	0.21	0.24	0.22	0.21	0.25
X15	0.29	0.14	0.14	0.24	0.20	0.17	0.21	0.23	0.31
X16	0.27	0.15	0.15	0.29	0.20	0.24	0.21	0.22	0.28
X17	0.33	0.12	0.15	0.28	0.23	0.23	0.24	0.22	0.30
X18	0.28	0.16	0.17	0.25	0.20	0.28	0.22	0.21	0.30
X19	0.32	0.15	0.16	0.28	0.19	0.25	0.22	0.22	0.28
X20	0.30	0.14	0.14	0.29	0.21	0.25	0.22	0.29	0.30
Πίνακας 5.3.4	**50 επαναλήψεις του παραδείγματος 2 σε προσομοιωμένα δεδομένα (με ανεξαρτησία)**								

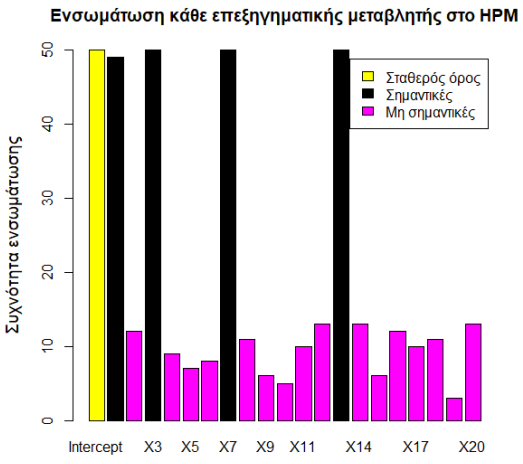
Πίνακας 5.3.4 – Παρατηρήσεις:

Στον πίνακα 5.3.4 βλέπουμε πώς κατανέμονται κατά μέσο όρο οι εκ των υστέρων πιθανότητες συμπερίληψης της κάθε επεξηγηματικής μεταβλητής κατά την ανίχνευση του πραγματικού μοντέλου για κάθε prior ξεχωριστά.

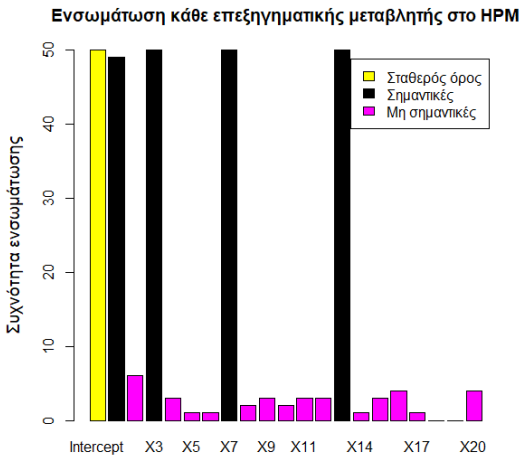
Είναι εμφανές ότι σε όλες τις περιπτώσεις ο σταθερός όρος ενσωματώνεται με πιθανότητα 1 στο πραγματικό μοντέλο, ενώ κάτι αντίστοιχο συμβαίνει και για τις επεξηγηματικές μεταβλητές που πραγματικά επεξηγούν τη μεταβλητή απόκρισης. Συγκεκριμένα βλέπουμε ότι οι επεξηγηματικές μεταβλητές X3, X7 και X13 συμπεριλαμβάνονται με εκ των υστέρων πιθανότητα 1 ενώ η επεξηγηματική μεταβλητή X1 με εκ των υστέρων πιθανότητα που τείνει και αυτή στη μονάδα (με κάπως μικρότερες τιμές στις περιπτώσεις των AIC, g και ZS-full prior).

Από την άλλη, βλέπουμε έκδηλα ότι κάτω από τις επιλογές των AIC και ZS-full prior για τις υπόλοιπες επεξηγηματικές μεταβλητές που δεν επιδρούν στο πραγματικό μοντέλο, δίνεται μεγαλύτερη εκ των υστέρων πιθανότητα οι συντελεστές που τους αντιστοιχούν να είναι μη μηδενικοί σε παραλληλισμό με τις υπόλοιπες prior.

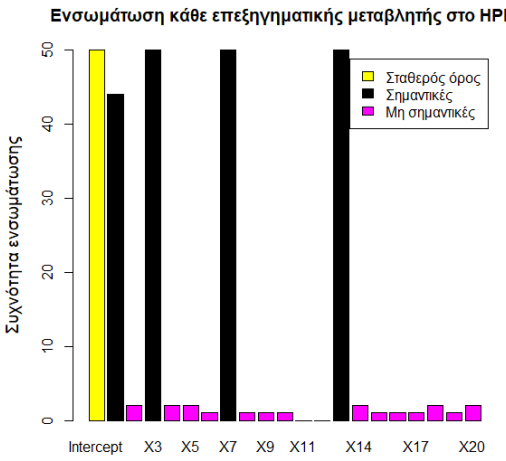
Παρακάτω απεικονίζουμε γραφικά τη συχνότητα εμφάνισης της κάθε επεξηγηματικής μεταβλητής στα μοντέλα υψηλότερης εκ των υστέρων πιθανότητας HPM (Hightest Propability Model) για τις 50 επαναλήψεις του παραδείγματος 2 σε προσομοιωμένα δεδομένα καθώς και τους (κατά μέσο όρο) εκ των υστέρων μέσους των συντελεστών και τις (κατά μέσο όρο) εκ των υστέρων πιθανότητες συμπερίληψης PIP (Posterior Inclusion Propabilities) των επεξηγηματικών μεταβλητών κατά την ανίχνευση του πραγματικού μοντέλου στις 50 επαναλήψεις για κάθε επιλεγμένη prior ξεχωριστά ως παιρετέρω ανάλυση των όσων έχουν διατυπωθεί στις παρατηρήσεις των πινάκων 5.3.1 – 5.3.4.



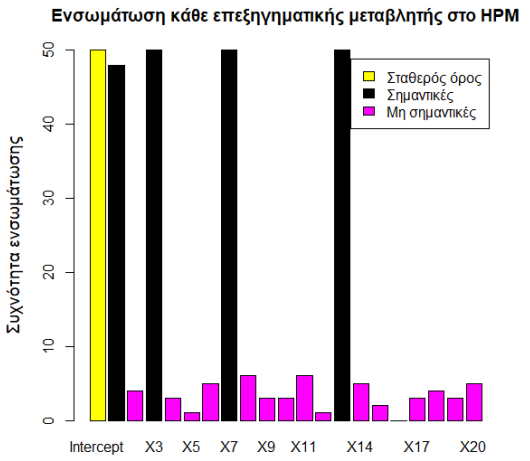
Ραβδόγραμμα 1: AIC prior (50 επαναλήψεις - παράδειγμα 2)



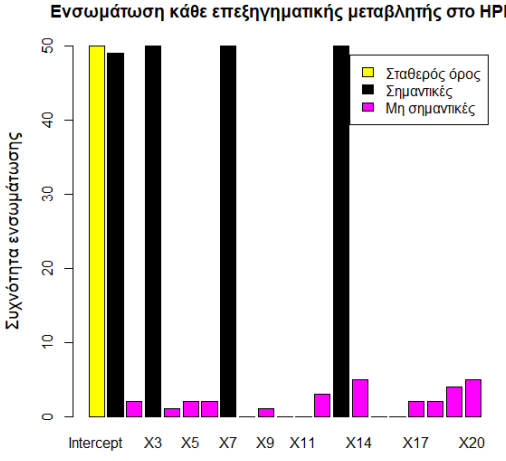
Ραβδόγραμμα 2: BIC prior (50 επαναλήψεις - παράδειγμα 2)



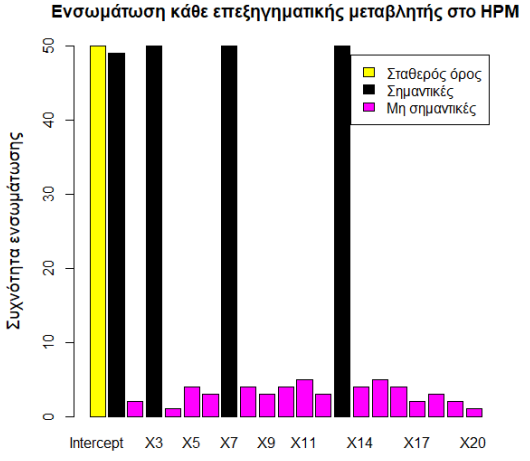
Ραβδόγραμμα 3: g prior (50 επαναλήψεις - παράδειγμα 2)



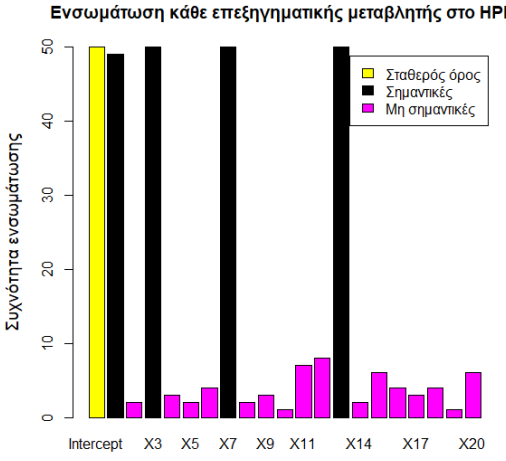
Ραβδόγραμμα 4: hyper-g prior (50 επαναλήψεις - παράδειγμα 2)



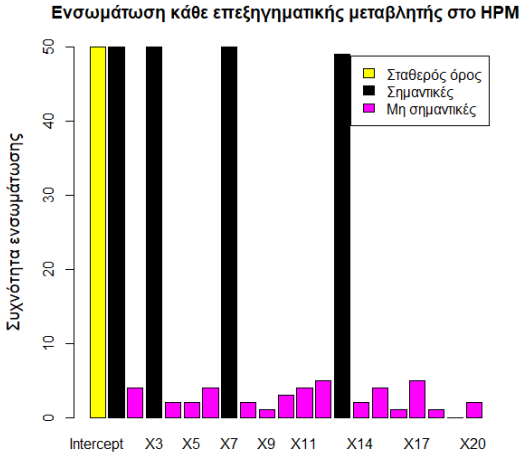
Ραβδόγραμμα 5: hyper-g/n prior (50 επαναλήψεις - παράδειγμα 2)



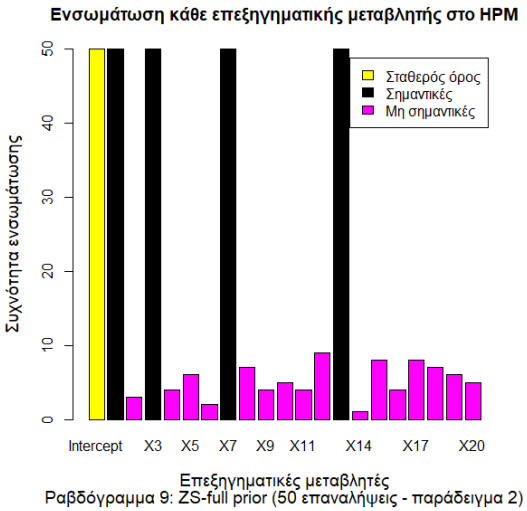
Ραβδόγραμμα 6: EBL prior (50 επαναλήψεις - παράδειγμα 2)



Ραβδόγραμμα 7: EBG prior (50 επαναλήψεις - παράδειγμα 2)



Ραβδόγραμμα 8: ZS-null prior (50 επαναλήψεις - παράδειγμα 2)



Ραβδογράμματα 1-9 (α): Συχνότητα ενσωμάτωσης κάθε επεξηγηματικής μεταβλητής στο εκάστοτε μοντέλο υψηλότερης εκ των υστέρων πιθανότητας HPM για 50 επαναλήψεις:

Ραβδογράμματα 1-9 (α) - Παρατηρήσεις:

Συνοδευτικά με την παρεχόμενη πληροφορία του πίνακα 5.3.1, εξακριβώνουμε και γραφικά με τα παραπάνω ραβδογράμματα τα όσα παραθέσαμε στις παρατηρήσεις του εν λόγω πίνακα.

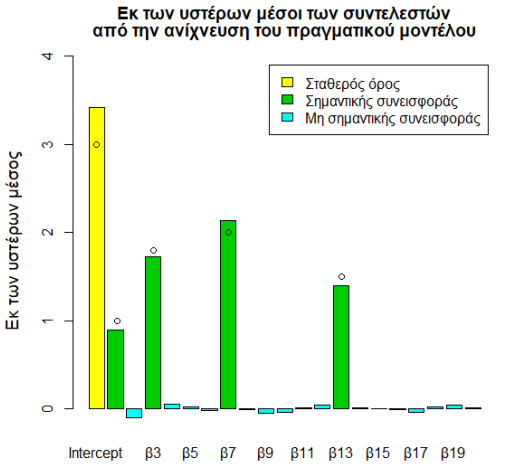
Συγκεκριμένα διαπιστώνουμε ότι με τις επιλογές των prior AIC και ZS-full, επικρατεί μεγαλύτερη τάση συγκριτικά με τις υπόλοιπες να ενσωματώνονται στατιστικά μη σημαντικές μεταβλητές στο υψηλότερης εκ των υστέρων πιθανότητας μοντέλο. Εν αντιθέσει, με τις επιλογές των hyper-g/n και ZS-null παρατηρούμε σχεδόν 100% συχνότητα εμφάνισης των στατιστικά σημαντικών επεξηγηματικών μεταβλητών και πολύ μικρότερη συχνότητα ενσωμάτωσης των στατιστικά μη σημαντικών, όταν οι επιλογές αυτές υποβάλλονται σε σύγκριση με τις υπόλοιπες.

Αξίζει να σημειωθεί επίσης ότι με την prior BIC τα αποτελέσματα ως προς τη συχνότητα ενσωμάτωσης των στατιστικά σημαντικών και μη σημαντικών επεξηγηματικών μεταβλητών φαίνονται πολύ ικανοποιητικά, κάτι το οποίο είναι λογικό μιας και το μέγεθος του δείγματος n=100 είναι αρκετά μεγαλύτερο του πλήθους των επεξηγηματικών μεταβλητών p=20. Ωστόσο δεν πρέπει να ξεχνάμε ότι με την prior BIC παρατηρήσαμε μεγαλύτερα τυπικά σφάλματα και ότι σε γενικές γραμμές επικρατεί η τάση να επιλέγεται το πιο φειδωλό μοντέλο.

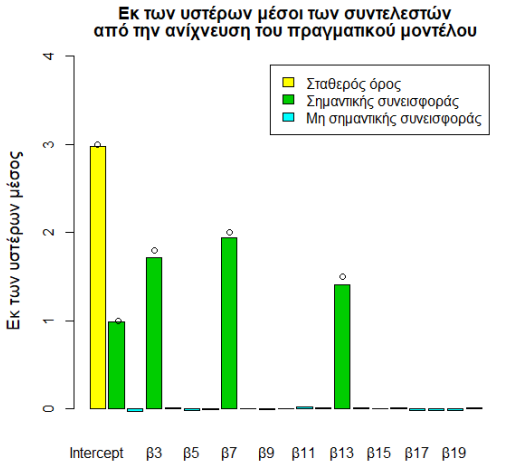
Όσον αφορά την εξέταση του ραβδογράμματος υπό την επιλογή της g prior, παρατηρούμε μία πολύ ικανοποιητική εικόνα αλλά ωστόσο και σε αυτή την

περίπτωση δεν πρέπει να ξεχνάμε ότι μολονότι η g prior αποτελεί μια δημοφιλή επιλογή λόγω του ότι οδηγεί σε κλειστές μορφές της περιθώριας πιθανοφάνειας και του παράγοντα Bayes, εν τούτοις παρουσιάζει δυσκολίες ως προς τον καθορισμό της τιμής που θα δοθεί στην παράμετρο g, όπου εν προκειμένου εξετάσαμε την περίπτωση της μοναδιαίας πληροφορίας (για n αρκετά μεγάλο).

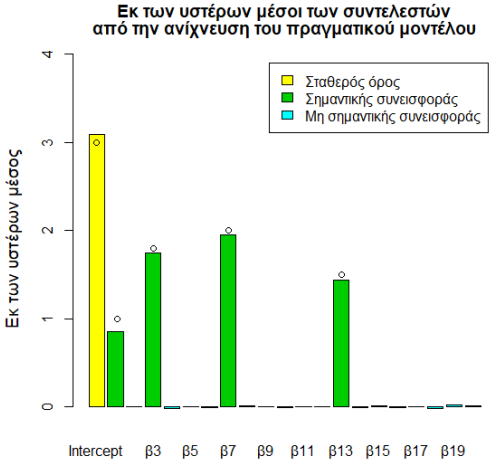
Τέλος υπό των prior hyper-g, EBL και EBG τα αποτελέσματα φαίνονται αρκετά ικανοποιητικά και με πολύ μικρές διαφοροποιήσεις. Ωστόσο υπενθυμίζουμε ότι η hyper-g μολονότι επιλύει το παράδοξο πληροφορίας χρειάζεται να ικανοποιεί μια επιπλέον ακόμη συνθήκη συγκριτικά με την prior ZS και επιπλέον μολονότι αυτή και οι EB priors είναι συνεπείς για μεγάλο n, χάνουν τη συνέπειά τους όταν το πραγματικό μοντέλο είναι το μηδενικό μοντέλο (εν αντιθέσει της ZS). Οι δε EB priors έχουν ακόμη το μειονέκτημα να παρουσιάζουν πρόβλημα όταν το πλήθος των παραμέτρων εξαρτάται από το μέγεθος του δείγματος.



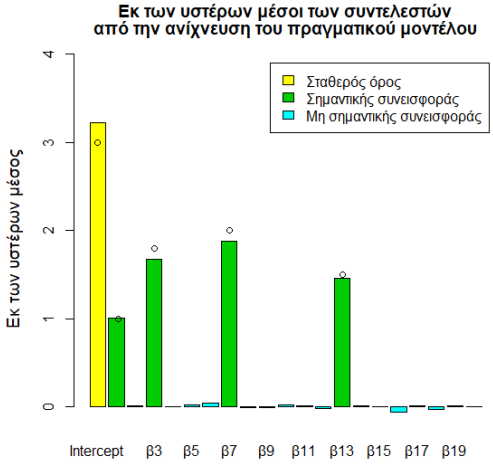
Ραβδόγραμμα 1: AIC prior (50 επαναλήψεις - παράδειγμα 2)



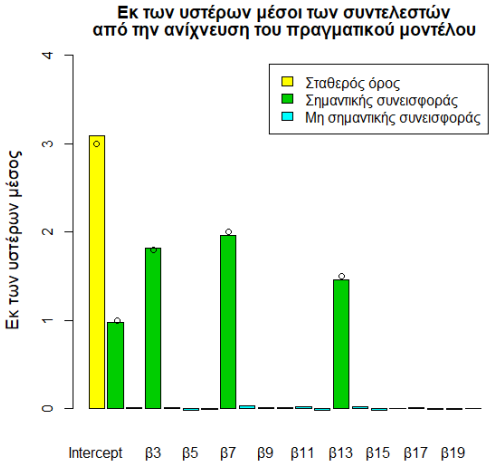
Ραβδόγραμμα 2: BIC prior (50 επαναλήψεις - παράδειγμα 2)



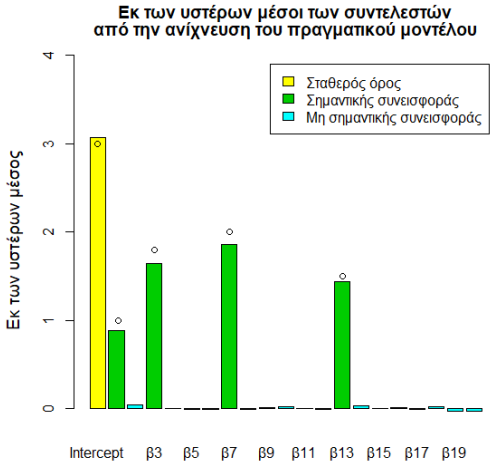
Συντελεστές επεξηγηματικών μεταβλητών
Ραβδόγραμμα 3: g prior (50 επαναλήψεις - παράδειγμα 2)



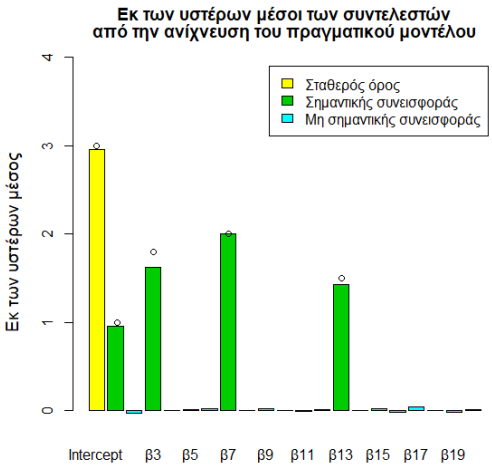
Συντελεστές επεξηγηματικών μεταβλητών
Ραβδόγραμμα 4: hyper-g prior (50 επαναλήψεις - παράδειγμα 2)



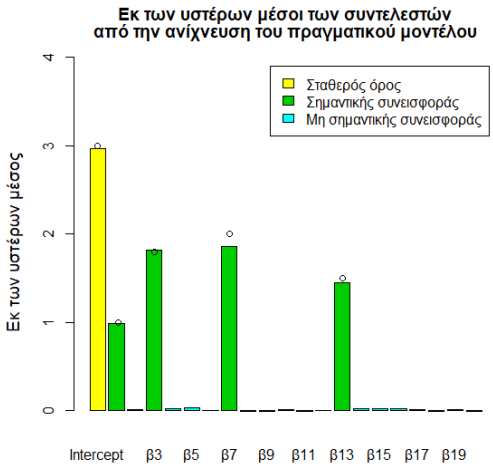
Συντελεστές επεξηγηματικών μεταβλητών
Ραβδόγραμμα 5: hyper-g/n prior (50 επαναλήψεις - παράδειγμα 2)



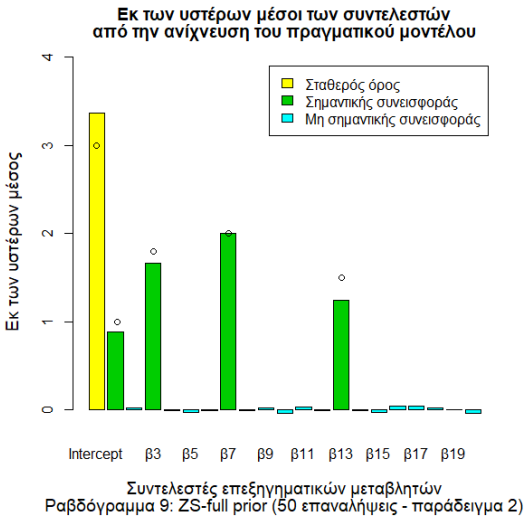
Συντελεστές επεξηγηματικών μεταβλητών
Ραβδόγραμμα 6: EBL prior (50 επαναλήψεις - παράδειγμα 2)



Συντελεστές επεξηγηματικών μεταβλητών
Ραβδόγραμμα 7: EBG prior (50 επαναλήψεις - παράδειγμα 2)



Συντελεστές επεξηγηματικών μεταβλητών
Ραβδόγραμμα 8: ZS-null prior (50 επαναλήψεις - παράδειγμα 2)

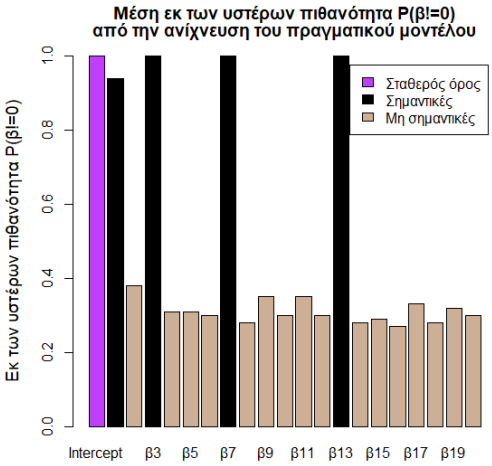


Ραβδογράμματα 1-9 (β): Εκ των υστέρων μέσοι των συντελεστών των επεξ. μεταβλητών κατά μέσο όρο από την ανίχνευση του πραγματικού μοντέλου στις 50 επαναλήψεις:

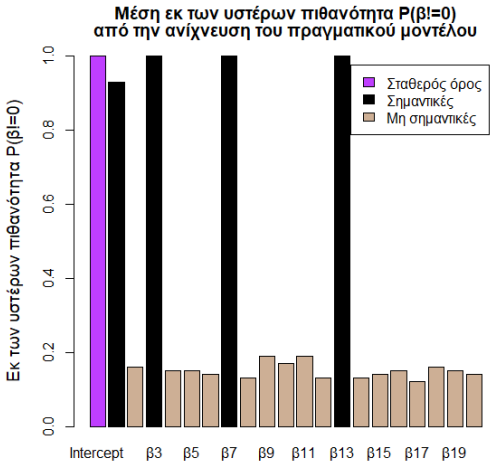
Ραβδογράμματα 1-9 (β) – Παρατηρήσεις:

Τα παραπάνω ραβδογράμματα αναπαριστούν την εκτιμώμενη μέση τιμή των συντελεστών των επεξηγηματικών μεταβλητών κατά την ανίχνευση του πραγματικού μοντέλου στις 50 επαναλήψεις, όπου το κίτρινο χρώμα αντιστοιχεί στο σταθερό όρο, το πράσινο στις στατιστικά σημαντικές μεταβλητές και η κουκίδα στην τιμή των συντελεστών του πραγματικού μοντέλου.

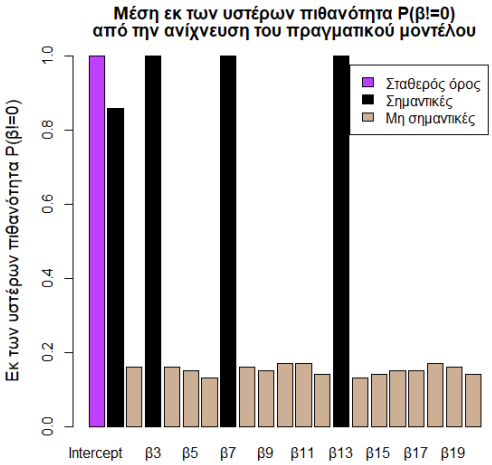
Είναι αντιληπτό για άλλη μια φορά ότι με τις prior AIC και ZS-full οδηγούμαστε σε μεγαλύτερες αποκλίσεις από την πραγματική τιμή των συντελεστών των στατιστικά σημαντικών επεξηγηματικών μεταβλητών έναντι των υπολοίπων prior. Με τις prior BIC, hyper-g και EBL, λαμβάνουμε καλύτερες εκτιμήσεις αλλά έχουμε και πάλι αποκλίσεις από τις πραγματικές τιμές οι οποίες δείχνουν να είναι κάπως μικρότερες υποκείπτοντας στην g-prior και την EBG ενώ τις καλύτερες εκτιμήσεις (σχεδόν ταυτιζόμενες με τις πραγματικές τιμές) τις λαμβάνουμε κάτω από την hyper-g/n και ZS-null prior.



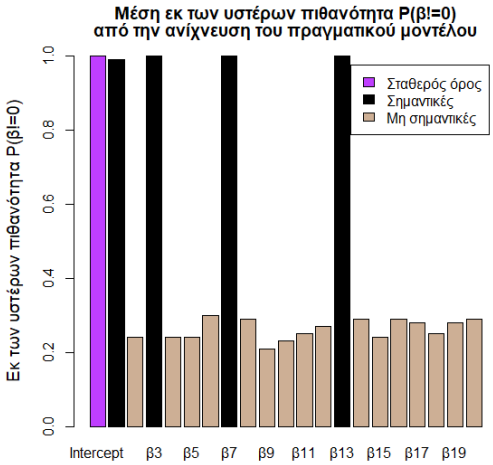
Ραβδόγραμμα 1: AIC prior (50 επαναλήψεις - παράδειγμα 2)



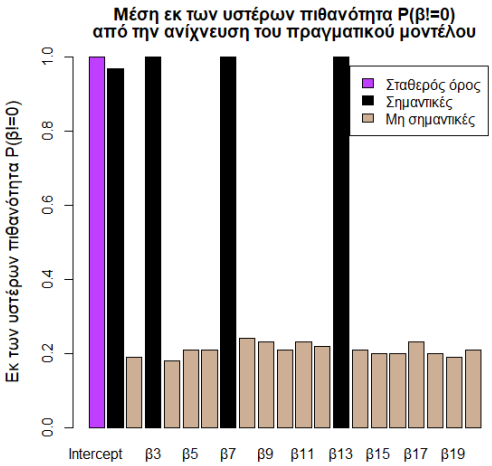
Ραβδόγραμμα 2: BIC prior (50 επαναλήψεις - παράδειγμα 2)



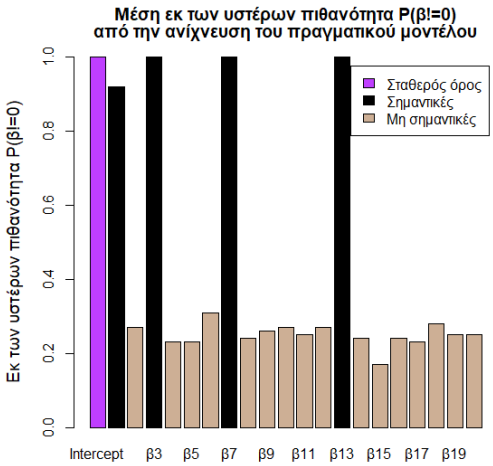
Ραβδόγραμμα 3: g prior (50 επαναλήψεις - παράδειγμα 2)



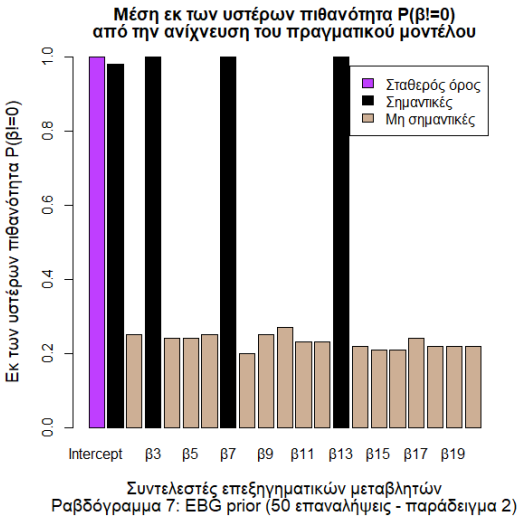
Ραβδόγραμμα 4: hyper-g prior (50 επαναλήψεις - παράδειγμα 2)



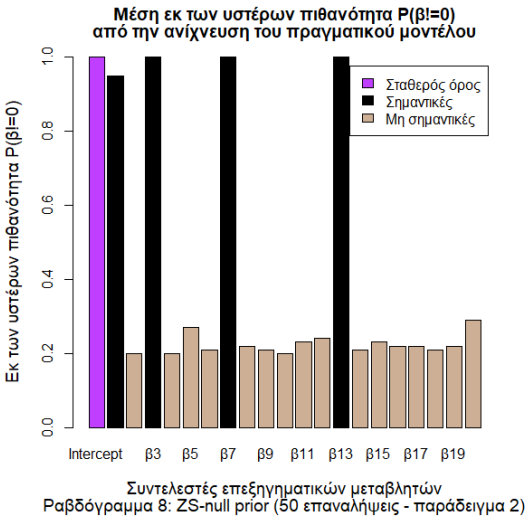
Ραβδόγραμμα 5: hyper-g/n prior (50 επαναλήψεις - παράδειγμα 2)



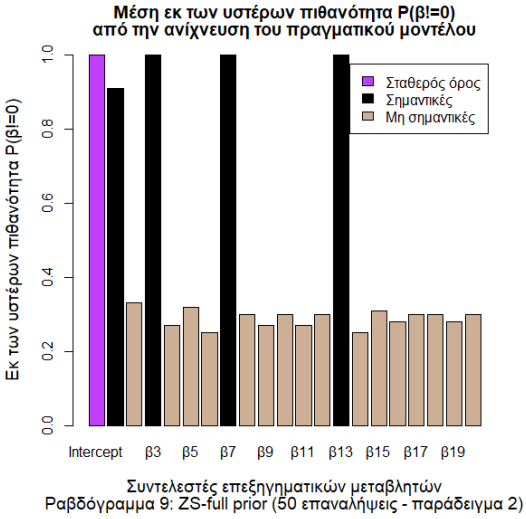
Ραβδόγραμμα 6: EBL prior (50 επαναλήψεις - παράδειγμα 2)



Ραβδόγραμμα 7: EBG prior (50 επαναλήψεις - παράδειγμα 2)



Ραβδόγραμμα 8: ZS-null prior (50 επαναλήψεις - παράδειγμα 2)



Ραβδόγραμμα 9: ZS-full prior (50 επαναλήψεις - παράδειγμα 2)

Ραβδογράμματα 1-9 Μέση εκ των υστέρων πιθανότητα p(β!=0) από την ανίχνευση του πραγματικού μοντέλου στις 50 επαναλήψεις:

Ραβδογράμματα 1-9 – Παρατηρήσεις:

Η συγκριτική μας άποψη για τις 9 υπό μελέτη priors ολοκληρώνεται με την παρατήρηση των ραβδογραμμάτων 1-9 όπου συμπηρωματικά με την ανάλυση του πίνακα 5.3.4 και όλα τα προαναφερθή σχόλια παρατηρούμε ότι με όλες τις δυνατές α-priori επιλογές, οι εκ των υστέρων πιθανότητες συμπερίληψης των στατιστικά σημαντικών επεξηγηματικών μεταβλητών τείνουν στη μονάδα.

Η AIC και η ZS-full prior αντιπαραβαλλόμενες με τις υπόλοιπες φαίνεται να δίνουν μικρότερη εκ των υστέρων πιθανότητα συμπερίληψης για τη μεταβλητή X1 (πράγμα κακό διότι η X1 συμμετέχει στο πραγματικό μοντέλο) και αφετέρου μεγαλύτερες εκ των υστέρων πιθανότητες συμπερίληψης σε στατιστικά μη σημαντικές μεταβλητές όπως π.χ στη X2 (γεγονός εξίσου κακό).

Στην αναμέτρηση αυτή η hyper-g/n και η ZS-null priors φαίνεται να αντικατοπτρίζουν καλύτερα την πραγματικότητα μιας και παρουσιάζουν σχεδόν μοναδιαία εκ των υστέρων πιθανότητα συμπερίληψης για το σταθερό όρο και όλες τις στατιστικά σημαντικές επεξηγηματικές μεταβλητές, δηλαδή τις X1, X3, X7 και X13 (μεταβλητές του πραγματικού μοντέλου), ενω ταυτόχρονα δίνουν και πολύ μικρή εκ των υστέρων πιθανότητα συμπερίληψης στις στατιστικά μη σημαντικές επεξηγηματικές μεταβλητές.

Οι υπόλοιπες prior συμπεριλαμβάνονται στην κλάση με τα ενδιάμεσα αποτελέσματα. Συγκεκριμένα η BIC prior μολονότι φαίνεται να παρέχει μικρές εκ των υστέρων πιθανότητες συμπερίληψης στις στατιστικά μη σημαντικές μεταβλητές, υστερεί περισσότερο από τις hyper-g/n και ZS-null ως προς την εκ των υστέρων πιθανότητα συμπερίληψης που δίνει στην επεξηγηματική μεταβλητή X1 κάτι το οποίο ισχύει και με την g prior. Όσο για τις hyper-g, EBL και EBG φαίνεται να σημειώνονται σχεδόν αμυδρές μεταξύ τους διαφοροποιήσεις ως προς τις εκ των υστέρων πιθανότητες συμπερίληψης της κάθε επεξηγηματικής μεταβλητής και του σταθερού όρου, οι οποίες εμφανίζονται να είναι αρκετά ικανοποιητικές.

Για την κάλυψη των προθέσεων της παρούσας διπλωματικής εργασίας, συνεχίζουμε στην επόμενη ενότητα με ένα ακόμα παράδειγμα για να μελετήσουμε την επίδραση των εξεταζόμενων εκ των προτέρων κατανομών των συντελεστών στα εκ των υστέρων αποτέλεσματα, στην περίπτωση όπου έχουμε συσχετιζόμενα δεδομένα.

5.4 ΠΑΡΑΔΕΙΓΜΑ 3:

Μπεϋζιανή επιλογή μοντέλου και μεταβλητών σε προσομοιωμένα δεδομένα με συσχετιζόμενες επεξηγηματικές μεταβλητές

Στο παρόν παράδειγμα θα εργαστούμε ομοιοτρόπως με το παράδειγμα 2. Δηλαδή θα προσομοιώσουμε δεδομένα για τις επεξηγηματικές μεταβλητές και τη μεταβλητή απόκρισης θεωρώντας ότι ισχύουν οι ίδιες υποθέσεις που διατυπώσαμε εκεί, με μόνη διαφορά ότι θα θεωρήσουμε πως κάποιες από τις επεξηγηματικές μας μεταβλητές συσχετίζονται μεταξύ τους. Σημειώνουμε επιπλέον ότι η διαδικασία της προσομοίωσης θα επαναληφθεί για 100 φορές.

Οι υποθέσεις που κάνουμε (πέραν της γραμμικότητας του μοντέλου) έχουν ως εξής

1. $n = 100$, το μέγεθος του δείγματος.

2. $p = 20$, το πλήθος των διαθέσιμων επεξηγηματικών μεταβλητών.
3. $\beta_0 = 3$ & $\boldsymbol{\beta} = (1, 0, 1.8, 0, 0, 0, 2, 0, 0, 0, 0, 0, 1.5, 0, 0, 0, 0, 0, 0,0)$, οι τιμές για το σταθερό όρο και τους συντελεστές των επεξηγηματικών μεταβλητών.
4. $Y_i \sim N(\alpha + X_{i1} + 1.8\, X_{i3} + 2X_{i7} + 1.5X_{i13}, 2.5),\ i = 1, \dots, 100,$ το πραγματικό μοντέλο.
5. $X_{ij} \sim N(0,1),\ i = 1, \dots, 100$ και $j \in \{1, \dots, 20\} \backslash \{9,11\}$.
6. $X9 = 0.8X3 + \mathbf{e}$ με $e_i \sim N(0,0.2)$ για $i = 1, \dots, 100$.
7. $X11 = 0.9X1 + \mathbf{e}$ με $e_i \sim N(0,0.2)$ για $i = 1, \dots, 100$.
8. $p(M_\gamma) = \frac{1}{2^p}$, η εκ των προτέρων κατανομή για το μοντέλο M_γ.

Σημειώνουμε ότι ο αλγόριθμος που υλοποιεί την παραπάνω διαδικασία είναι παραπλήσιος με εκείνον που χρησιμοποιήσαμε στο παράδειγμα 2, συμπεριλαμβάνοντας απλά και τις συσχετίσεις που δηλώθηκαν στις υποθέσεις 6-7.

Παρακάτω παρουσιάζουμε κατά αντιστοιχία με το προηγηθέν παράδειγμα, τους πίνακες των εκ των υστέρων αποτελεσμάτων και τα αντίστοιχα ραβδογράμματα.

Συχνότητα ενσωμάτωσης της εκάστοτε επεξηγηματικής μεταβλητής στα top πρώτα μοντέλα

100 επαναλήψεις	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
Intercept	100	100	100	100	100	100	100	100	100
X1	70	67	62	65	64	64	68	70	56
X2	22	7	9	9	9	9	6	2	13
X3	84	81	79	79	85	81	81	82	81
X4	15	0	13	6	3	6	11	3	9
X5	18	7	8	9	7	5	7	2	12
X6	20	10	15	5	9	6	5	10	15
X7	100	100	100	100	100	100	100	100	100
X8	23	2	15	5	6	10	6	7	5
X9	26	20	28	25	19	22	26	21	25
X10	20	4	16	0	2	3	12	5	12
X11	49	30	41	39	35	40	32	30	49
X12	16	6	10	5	4	8	7	6	13
X13	100	100	100	100	100	100	100	100	100
X14	19	8	12	2	7	12	9	5	17
X15	15	6	10	10	3	6	7	2	11
X16	12	6	6	6	7	6	6	10	16
X17	18	4	11	8	6	9	8	5	11
X18	25	9	5	3	3	6	7	6	15
X19	17	6	12	5	4	7	4	6	9
X20	22	6	10	5	7	8	9	6	7

Μέση διάσταση μοντέλου (για όλες τις επαναλήψεις)

	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
dim	8	6	7	6	6	6	6	6	7

Πίνακας 5.4.1 **100 επαναλήψεις του παραδείγματος 3 σε προσομοιωμένα δεδομένα (με συσχέτιση)**

Πίνακας 5.4.1-Παρατηρήσεις:

Κατ' αντιστοιχία με τα αποτελέσματα που λάβαμε κατά την υλοποίηση του παραδείγματος 2, επιβεβαιώνουμε και εδώ μέσα από τη συχνότητα εμφάνισης της εκάστοτε επεξηγηματικής μεταβλητής στον πίνακα 5.4.1, ότι όσον αφορά την αποδοτικότητα των υπό μελέτη prior, οι AIC και ZS-full φαίνεται να υστερούν για το λόγο ότι έχουν την τάση στην εκάστοτε επανάληψη του αλγορίθμου να ενσωματώνουν στατιστικά μη σημαντικές μεταβλητές και την τάση να επιλέγουν μοντέλα υψηλότερης διάστασης συγκριτικά με τις υπόλοιπες prior. Οι prior BIC, hyper-g/n και ZS-null φαίνονται πιο αποδοτικές από την άποψη ότι ενσωματώνουν λιγότερες φορές στατιστικά μη σημαντικές μεταβλητές και επιλέγουν μοντέλα μικρότερης διάστασης από τις AIC και ZS-full ενώ οι prior g, hyper-g, EBL και EBG έχουν παραπλήσια αποτελέσματα και είναι σε μία ενδιάμεση κατάσταση με τη hyper-g να φαίνεται να βελτιώνει τη g prior και την EBG να βελτιώνει την EBL. Αξίζει ωστόσο να σημειωθεί ότι

η υψηλή αποδοτικότητα της BIC prior οφείλεται και στο γεγονός ότι το μέγεθος του δείγματος είναι αρκετά μεγαλύτερο από το πλήθος των επεξηγηματικών μεταβλητών.

Επιπρόσθετα διαπιστώνουμε ότι με οποιαδήποτε prior, για το σταθερό όρο και τις στατιστικά σημαντικές επεξηγηματικές μεταβλητές (β_0, X1, X3, X7, X13), παρατηρούμε πολύ υψηλή συχνότητα εμφάνισης με το σταθερό όρο την X7 και τη X13 να επιλέγονται σε όλες τις επαναλήψεις του αλγορίθμου ενώ τη X1 και τη X3 να επιλέγονται λιγότερες φορές λόγω της συσχέτισής τους με τις X11 και X9 αντίστοιχα. Αυτός είναι ο λόγος επίσης που που οι στατιστικά μη σημαντικές επεξηγηματικές μεταβλητές X9 και X11 επιλέγονται περισσότερες φορές απ ότι τις είδαμε να εμφανίζονται κατά την υλοποίηση του παραδείγματος 2, σε σημείο που τείνουν να επιλεχθούν ως στατιστικά σημαντικές.

Η διαφορά της συχνότητας εμφάνισης των εν λόγω επεξηγηματικών μεταβλητών συγκριτικά με τη συχνότητα εμφάνισης των υπολοίπων στατιστικά μη σημαντικών επεξηγηματικών μεταβλητών είναι περισσότερο ευδιάκριτη υπό των επιλογών g, hyper-g, hyper-g/n, EBL, EBG και ZS-null μιας και υπό των prior AIC και ZS-full η τάση να ενσωματώνονται αρκετές φορές οι στατιστικά μη σημαντικές επεξηγηματικές μεταβλητές επικρατεί για σχεδόν όλες.

Ισοσταθμικά εκ των υστέρων αποτελέσματα από την ανίχνευση του πραγματικού μοντέλου

100 επαναλήψεις	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
Αριθμός εμφάνισης του πραγματικού μοντέλου	2	22	10	28	31	17	23	26	5

Μέση τιμή των εκ των υστέρων μέσων των ενδιαφερόμενων παραμέτρων κατά την ανίχνευση του πραγματικού μοντέλου στις 100 επαναλήψεις

	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
Prob(true model)	0.00	0.05	0.00	0.01	0.02	0.01	0.01	0.02	0.00
Log(marginal likelihood)	-321	-332	28.10	41.03	40.70	44.13	41.53	38.64	20.06
Intercept	2.45	3.04	3.17	2.99	3.07	2.96	3.02	3.02	3.31
X1	1.05	0.83	0.47	0.86	0.94	0.79	0.77	0.76	0.96
X2	-0.05	-0.01	0.00	-0.03	0.02	-0.01	-0.01	0.00	-0.04
X3	1.67	1.65	0.91	1.39	1.60	1.61	1.39	1.45	1.53
X4	-0.04	-0.01	-0.02	0.02	-0.01	0.01	-0.02	0.01	0.02
X5	-0.02	0.02	-0.01	0.00	-0.01	-0.01	-0.02	0.04	-0.06
X6	-0.11	0.01	0.03	0.00	0.01	0.01	0.02	0.00	0.02
X7	1.56	1.95	1.57	1.95	1.96	2.04	1.85	1.91	2.10
X8	-0.02	-0.02	0.00	0.02	0.00	0.01	0.02	0.01	0.01
X9	0.28	0.17	0.39	0.39	0.21	0.04	0.42	0.26	0.41
X10	-0.01	0.01	0.02	0.00	0.02	0.02	-0.02	0.00	-0.01
X11	0.08	0.14	0.10	0.14	0.08	0.21	0.17	0.17	0.07
X12	0.02	-0.02	-0.01	-0.01	0.00	0.02	-0.01	0.01	0.05
X13	1.78	1.46	1.04	1.47	1.32	1.42	1.42	1.49	1.48
X14	-0.02	-0.01	-0.03	0.00	0.01	-0.01	-0.02	-0.01	0.00
X15	0.10	0.00	0.02	0.00	0.01	0.01	0.02	0.02	0.04
X16	-0.06	-0.01	-0.03	0.02	-0.03	0.01	0.01	0.02	-0.01
X17	-0.01	0.01	0.00	0.02	0.00	-0.02	-0.02	0.00	0.08
X18	-0.05	0.00	-0.01	-0.02	0.00	0.01	-0.01	-0.01	0.03
X19	0.02	0.00	0.00	0.01	-0.03	-0.01	-0.01	-0.02	0.01
X20	0.01	-0.01	-0.03	-0.01	0.01	-0.02	0.00	-0.01	0.02

Πίνακας 5.4.2 **100 επαναλήψεις του παραδείγματος 3 σε προσομοιωμένα δεδομένα (με συσχέτιση)**

Πίνακας 5.4.2-Παρατηρήσεις:

Στον πίνακα 5.4.2, τα αποτελέσματα του οποίου προέκυψαν από τις 100 επαναλήψεις του αλγορίθμου που υλοποιήθηκε για το παράδειγμα 3, παρατηρούμε αρχικά ότι όσον αφορά τη σύγκριση των υπό μελέτη εκ των προτέρων κατανομών ως προς τον αριθμό ανίχνευσης του πραγματικού μοντέλου και τις εκ των υστέρων εκτιμήσεις των συντελεστών του, ισχύουν τα ίδια που έχουμε αναφέρει κατά το σχολιασμό του πίνακα 5.3.2.

Αξίζει ωστόσο να σημειωθεί ότι οι συσχετίσεις που έχουμε συμπεριλάβει σύμφωνα με τις υποθέσεις (6) και (7) επενεργούν στο να μειωθεί ο αριθμός ανίχνευσης του πραγματικού μοντέλου, μολονότι έχουμε διπλασιάσει τον αριθμό επαναλήψεων του αλγορίθμου, ενώ επιπρόσθετα οι συσχετίσεις αυτές φαίνεται να επενεργούν και στους εκτιμώμενους μέσους των στατιστικά σημαντικών επεξηγηματικών μεταβλητών όπου παρατηρούμε ότι κυρίως για τις επεξηγηματικές μεταβλητές X1 και X3 (που αφορούν το πραγματικό μοντέλο και οι οποίες συσχετίζονται με τις X11 και X9) οι εκτιμώμενοι μέσοι των συντελεστών δεν προέκυψαν τόσο κοντά στις πραγματικές τιμές όπως προέκυψαν στην περίπτωση του παραδείγματος 2 όπου υποθέσαμε ανεξαρτησία.

Παράλληλα βλέπουμε οι στατιστικά μη σημαντικές επεξηγηματικές μεταβλητές X11 και X9 να αποκτούν μια όχι και τόσο αμελητέα επίδραση και άρα είναι φανερό ότι με τις εκ των προτέρων κατανομές που μελετάμε υπάρχει η τάση υπό την υπόθεση ότι οι επεξηγηματικές μεταβλητές συσχετίζονται μεταξύ τους, οι στατιστικά μη σημαντικές επεξηγηματικές μεταβλητές να τείνουν να θεωρηθούν στατιστικά σημαντικές.

Μέση τιμή των εκ των υστέρων τυπικών αποκλίσεων των ενδιαφερόμενων παραμέτρων κατά την ανίχνευση του πραγματικού μοντέλου στις 100 επαναλήψεις

100 επαναλήψεις	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
Prob(true model)	0.00	0.02	0.00	0.01	0.01	0.01	0.01	0.01	0.00
Log(marginal likelihood)	2.59	5.84	4.18	6.90	8.09	7.14	6.74	9.59	1.22
Intercept	0.25	0.25	0.25	0.25	0.25	0.24	0.25	0.25	0.28
X1	0.86	0.74	0.72	0.78	0.79	0.74	0.76	0.73	0.97
X2	0.17	0.11	0.17	0.16	0.13	0.14	0.14	0.15	0.18
X3	0.84	0.80	0.77	0.85	0.79	0.85	0.82	0.81	1.05
X4	0.16	0.14	0.16	0.15	0.15	0.15	0.14	0.16	0.16
X5	0.16	0.13	0.17	0.15	0.17	0.14	0.15	0.15	0.19
X6	0.19	0.16	0.17	0.15	0.15	0.17	0.17	0.16	0.18
X7	0.29	0.27	0.24	0.26	0.26	0.26	0.25	0.26	0.30
X8	0.14	0.13	0.15	0.16	0.16	0.14	0.16	0.17	0.19
X9	0.99	0.92	0.91	1.01	0.92	1.00	0.96	0.96	1.24
X10	0.16	0.11	0.17	0.16	0.16	0.15	0.15	0.15	0.18
X11	0.90	0.77	0.77	0.82	0.83	0.78	0.79	0.77	1.02
X12	0.23	0.13	0.15	0.16	0.14	0.15	0.15	0.14	0.18
X13	0.27	0.27	0.23	0.27	0.27	0.26	0.26	0.26	0.30
X14	0.15	0.12	0.17	0.13	0.15	0.15	0.16	0.13	0.16
X15	0.20	0.13	0.16	0.16	0.16	0.15	0.14	0.16	0.17
X16	0.19	0.12	0.18	0.16	0.16	0.14	0.15	0.14	0.16
X17	0.15	0.15	0.14	0.15	0.15	0.15	0.15	0.14	0.21
X18	0.20	0.12	0.16	0.16	0.14	0.15	0.16	0.14	0.18
X19	0.22	0.11	0.17	0.16	0.16	0.15	0.15	0.14	0.15
X20	0.14	0.11	0.15	0.15	0.15	0.15	0.13	0.14	0.18

Πίνακας 5.4.3 **100 επαναλήψεις του παραδείγματος 3 σε προσομοιωμένα δεδομένα (με συσχέτιση)**

Πίνακας 5.4.3-Παρατηρήσεις:

Συμπληρωματικά με όσα αναφέραμε προγενέστερα, στον πίνακα 5.4.3 παρατηρούμε πώς κατανέμονται τα τυπικά σφάλματα για την εκ των υστέρων πιθανότητα του μοντέλου, την περιθώρια πιθανοφάνεια και τους συντελεστές των επεξηγηματικών μεταβλητών κατά την ανίχνευση του πραγματικού μοντέλου στις 100 επαναλήψεις και με τις συσχετίσεις που περιγράψαμε.

Για συνεχή φορά παρατηρούμε ότι τα τυπικά σφάλματα των συντελεστών είναι μεγαλύτερα στις περιπτώσεις των AIC και ZS-full prior ενώ το τυπικό σφάλμα για την εκ των υστέρων πιθανότητα του μοντέλου κυμαίνεται μεταξύ 0.00 και 0.01 για όλες τις prior με λίγο υψηλότερη τιμή και ίση με 0.02 για την prior BIC.

Τέλος, συγκριτικά με τα αποτελέσματα για τα τυπικά σφάλματα που είχαμε στο παράδειγμα 2 η ουσιαστική διαφορά που μπορούμε να προσδιορίσουμε αφορά τις τιμές των τυπικών σφαλμάτων των συντελεστών των επεξηγηματικών μεταβλητών στις οποίες συμπεριλάβαμε τη συσχέτιση, όπου αναμφισβήτητα παρουσιάζονται να είναι κατά πολύ μεγαλύτερες κάτω από οποιοιαδήποτε a-priori επιλογή.

Μέση τιμή των εκ των υστέρων πιθανοτήτων p(B != 0) κατά την ανίνευση του πραγματικού μοντέλου στις 100 επαναλήψεις

100 επαναλήψεις	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
Intercept	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
X1	0.80	0.68	0.61	0.70	0.71	0.67	0.68	0.67	0.96
X2	0.31	0.13	0.38	0.24	0.19	0.22	0.22	0.22	0.27
X3	0.87	0.85	0.73	0.78	0.86	0.86	0.80	0.82	0.77
X4	0.31	0.16	0.37	0.24	0.22	0.23	0.23	0.22	0.25
X5	0.31	0.14	0.38	0.24	0.25	0.21	0.24	0.23	0.29
X6	0.42	0.19	0.37	0.22	0.21	0.26	0.26	0.23	0.27
X7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
X8	0.28	0.15	0.36	0.25	0.22	0.22	0.24	0.24	0.30
X9	0.37	0.27	0.52	0.39	0.30	0.35	0.35	0.34	0.45
X10	0.30	0.12	0.38	0.24	0.23	0.24	0.24	0.21	0.29
X11	0.45	0.37	0.53	0.43	0.42	0.45	0.44	0.43	0.49
X12	0.45	0.14	0.35	0.24	0.19	0.24	0.23	0.20	0.30
X13	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
X14	0.28	0.14	0.39	0.21	0.21	0.23	0.24	0.18	0.25
X15	0.40	0.15	0.37	0.26	0.23	0.23	0.22	0.23	0.26
X16	0.36	0.14	0.39	0.24	0.23	0.22	0.24	0.19	0.26
X17	0.33	0.17	0.36	0.25	0.22	0.23	0.23	0.20	0.35
X18	0.38	0.13	0.37	0.24	0.20	0.23	0.24	0.20	0.30
X19	0.42	0.12	0.37	0.24	0.23	0.23	0.23	0.20	0.25
X20	0.28	0.13	0.37	0.23	0.21	0.25	0.21	0.19	0.27

Πίνακας 5.4.4 **100 επαναλήψεις του παραδείγματος 3 σε προσομοιωμένα δεδομένα (με συσχέτιση)**

Πίνακας 5.4.4-Παρατηρήσεις:

Ενώ η εκ των υστέρων πιθανότητα συμπερίληψης των στατιστικά σημαντικών επεξηγηματικών μεταβλητών και του σταθερού όρου ήταν σχεδόν παντού ίση με 1 κατά την υλοποίηση του παραδείγματος 2 όπου υποθέσαμε ανεξαρτησία μεταξύ των επεξηγηματικών μεταβλητών, εδώ διαπιστώνουμε ότι η ένταξη των συσχετίσεων (σύμφωνα με τις υποθέσεις (6) και (7) που αναφέραμε στην αρχή του παραδείγματος 3) δείχνει να μειώνει την εκ των υστέρων πιθανότητα συμπερίληψης των στατιστικά σημαντικών επεξηγηματικών μεταβλητών X1 και X3 και να αυξάνει την εκ των υστέρων πιθανότητα συμπερίληψης των

στατιστικά μη σημαντικών επεξηγηματικών μεταβλητών X9 και X11. Οι υψηλές συσχετίσεις επομένως μεταξύ των επεξηγηματικών μεταβλητών επηρεάζουν κατά τέτοιο τρόπο τις εκ των υστέρων πιθανότητες συμπερίληψης ώστε να οδηγούν στο να θεωρήσουμε ακόμα και τις στατιστικά μη σημαντικές επεξηγηματικές μεταβλητές, ως στατιστικά σημαντικές.

Η AIC και η ZS-full όπως έχει ήδη αναφερθεί φαίνεται να δίνουν μεγαλύτερη εκ των υστέρων πιθανότητα συμπερίληψης σε στατιστικά μη σημαντικές επεξηγηματικές μεταβλητές συγκριτικά με τις υπόλοιπες prior. Αυτές οι πιθανότητες συμπερίληψης κατόπιν της ένταξης των συσχετίσεων φαίνεται συγκριτικά και με τα αποτελέσματα του παραδείγματος 2 να αυξάνονται ακόμα και στις περιπτώσεις των αποδοτικών prior. Ειδικά για τις επεξηγηματικές μεταβλητές X9 και X11 που συσχετίζονται με τις X3 και X1 του πραγματικού μοντέλου βλέπουμε να λαμβάνουν εκ των υστέρων πιθανότητες συμπερίληψης με τιμές μεταξύ 0.30 και 0.53 που είναι αρκετά υψηλές.

Εκ των υστέρων πιθανότητα συμπερίληψης των X9, X11 (υπό της υπόθεσης της ανεξαρτησίας)

	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
X9	0.35	0.19	0.15	0.21	0.23	0.26	0.25	0.21	0.27
X11	0.35	0.19	0.17	0.25	0.23	0.25	0.23	0.23	0.27

Εκ των υστέρων πιθανότητα συμπερίληψης των X9, X11 (με την εισαγωγή συσχετίσεων)

	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
X9	0.37	0.27	0.52	0.39	0.30	0.35	0.35	0.34	0.45
X11	0.45	0.37	0.53	0.43	0.42	0.45	0.44	0.43	0.49

Σημείωση: X9, X11 στατιστικά μη σημαντικές αφού X1, X3 επαρκείς για το πραγματικό μοντέλο.

Πίνακας 5.4.5 Αποδοτικότητα των prior στις περιπτώσεις ανεξαρτησίας και συσχέτισης των επεξ.μεταβλ.

Πίνακας 5.4.5-Παρατηρήσεις:

Στον πίνακα 5.4.5 παραθέτουμε συγκριτικά τις εκ των υστέρων πιθανότητες συμπερίληψης για τις μη στατιστικά σημαντικές επεξηγηματικές μεταβλητές X9 και X11 πριν και μετά την ενσωμάτωση της συσχέτισης.

Παρατηρούμε ότι στην περίπτωση του παραδείγματος 2 που υποθέσαμε ανεξαρτησία μεταξύ των επεξηγηματικών μεταβλητών οι τιμές των εκ των υστέρων πιθανοτήτων των εν λόγω επεξηγηματικών μεταβλητών προέκυψαν αρκετά μικρές για όλες τις prior εκτός των AIC και ZS-full οι οποίες εξ' αρχής

παρουσίαζαν την τάση να δίνουν υψηλές εκ των υστέρων πιθανότητες συμπερίληψης και στις στατιστικά μη σημαντικές επεξηγηματικές μεταβλητές. Κατόπιν όμως της ενσωμάτωσης των συσχετίσεων μεταξύ των επεξηγηματικών μεταβλητών X9-X3 και X11-X1 κατά την υλοποίηση του παραδείγματος 3, οι τιμές των εκ των υστέρων πιθανοτήτων συμπερίληψης των στατιστικά μη σημαντικών επεξηγηματικών μεταβλητών X9 και X11 αυξήθηκαν αισθητά με αποτέλεσμα να τείνουν να θεωρηθούν στατιστικά σημαντικές όπως έχουμε ήδη προαναφέρει.

Συχνότητα εμφάνισης του πραγματικού μοντέλου για ανεξάρτητες και συσχετισμένες επεξ. μεταβλητές

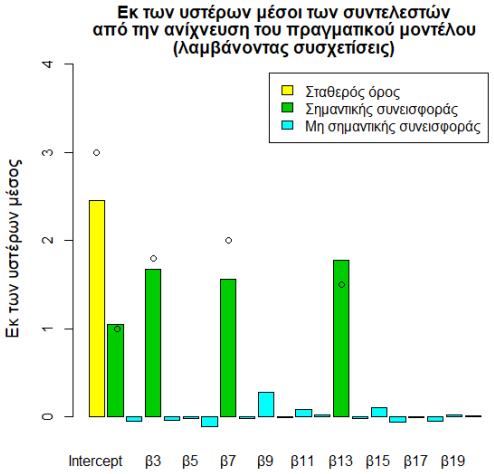
	AIC	BIC	g	Hyper g	Hyper g/n	EBL	EBG	ZS null	ZS full
50 επαναλ. (ανεξαρτησία)	**2**	**26**	**37**	**16**	**27**	**24**	**15**	**23**	**7**
100 επαναλ. (συσχέτιση)	**2**	**22**	**10**	**28**	**31**	**17**	**23**	**26**	**5**

Πίνακας 5.4.6 **Αριθμός ανίχνευσης του πραγματικού μοντέλου**

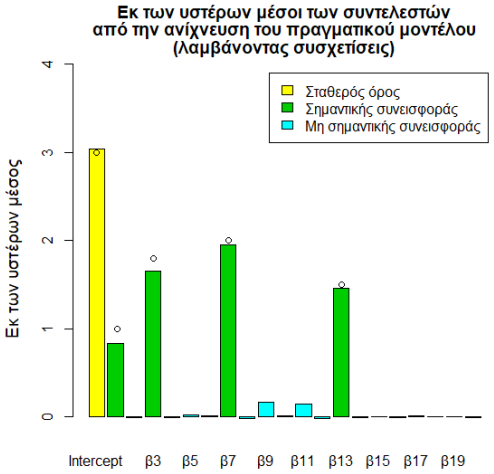
Πίνακας 5.4.6-Παρατηρήσεις:

Στον πίνακα 5.4.6 παρουσιάζουμε συγκριτικά τον αριθμό ανίχνευσης του πραγματικού μοντέλου σύμφωνα με τα παραδείγματα 2 και 3 διαπιστώνοντας ότι στη 2η περίπτωση, οι συμπεριλαμβανόμενες συσχετίσεις συνέλαβαν στο να αυξηθεί η πολυπλοκότητα του προβλήματος της επιλογής μοντέλου, μιας και μολονότι διπλασιάστηκε ο αριθμός των επαναλήψεων του αλγορίθμου, το πραγματικό μοντέλο ανιχνεύθηκε περίπου τόσες φορές όσο και στην πρώτη περίπτωση. Αυτό αναμένονταν λογικό αφού οι στατιστικά μη σημαντικές επεξηγηματικές μεταβλητές X9 και X11 ξεκίνησαν να θεωρούνται στατιστικά σημαντικές. Ωστόσο, παρά την ένταξη των συσχετίσεων οι εκ των προτέρων κατανομές hyper-g/n και ZS-null εξακολουθούν να φαίνονται και πάλι πιο αποδοτικές αφού ο αριθμός ανίχνευσης του πραγματικού μοντέλου στις 100 επαναλήψεις είναι μεγαλύτερος από κάθε άλλη επιλογή. Παρ' όλα αυτά η αποδοτικότητα των παραπάνω επιλογών μειώνεται καθώς ενσωματώνονται υψηλές συσχετίσεις μεταξύ των επεξηγηματικών μεταβλητών.

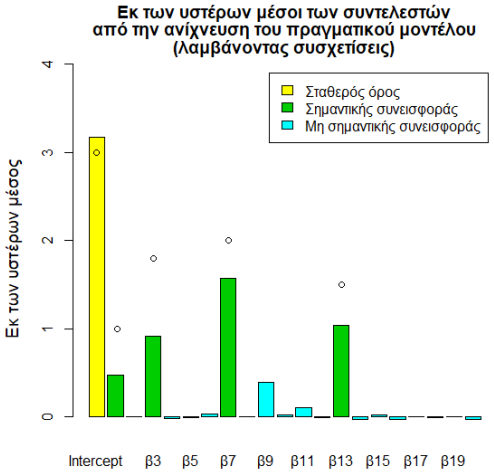
Οι πληροφορίες που αντλήσαμε από τα εκ των υστέρων αποτελέσματα του αλγορίθμου κατά την υλοποίηση του παραδείγματος 3 (για προσομοιωμένα δεδομένα με συσχέτιση), παρατέθηκαν κατά το σχολιασμό των πινάκων 5.4.1-5.4.6 και συνοψίζονται εικονικά στα επόμενα ραβδογράμματα.



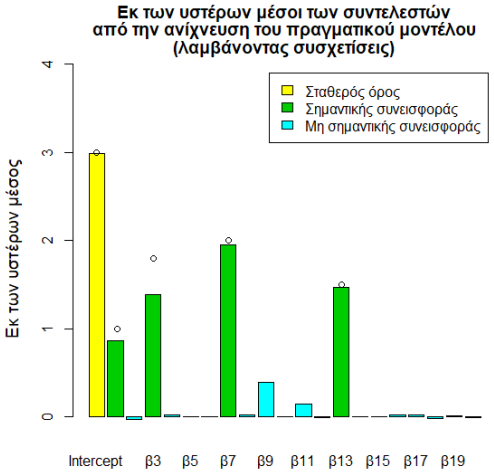
Συντελεστές επεξηγηματικών μεταβλητών
Ραβδόγραμμα 1: AIC prior (100 επαναλήψεις - παράδειγμα 3)



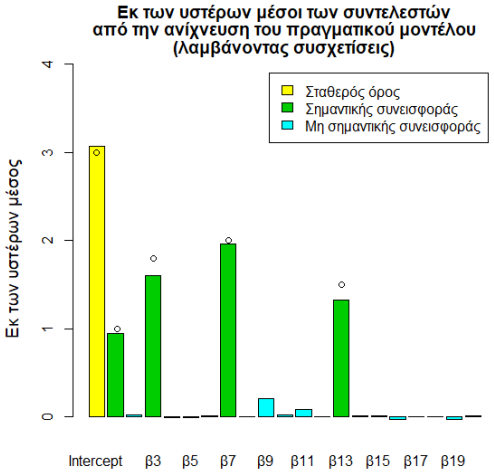
Συντελεστές επεξηγηματικών μεταβλητών
Ραβδόγραμμα 2: BIC prior (100 επαναλήψεις - παράδειγμα 3)



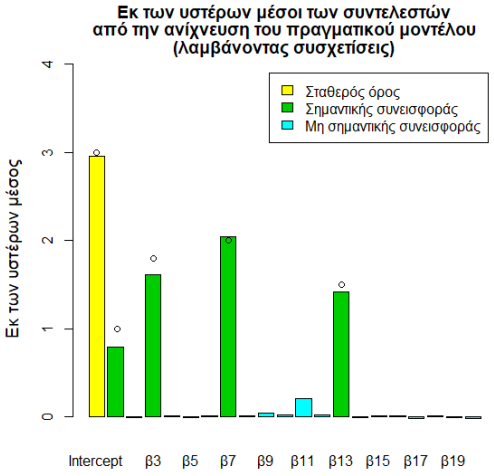
Συντελεστές επεξηγηματικών μεταβλητών
Ραβδόγραμμα 3: g prior (100 επαναλήψεις - παράδειγμα 3)



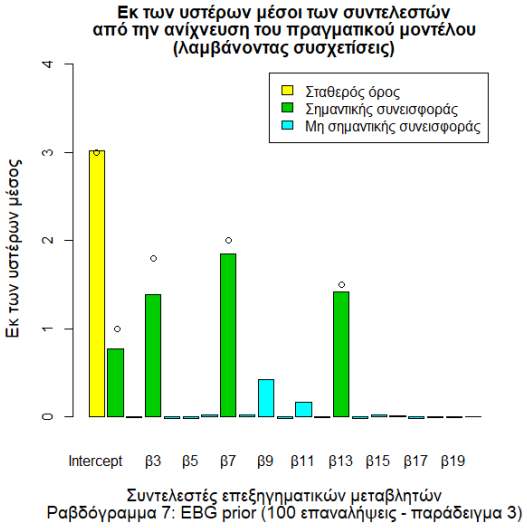
Συντελεστές επεξηγηματικών μεταβλητών
Ραβδόγραμμα 4: hyper-g prior (100 επαναλήψεις - παράδειγμα 3)

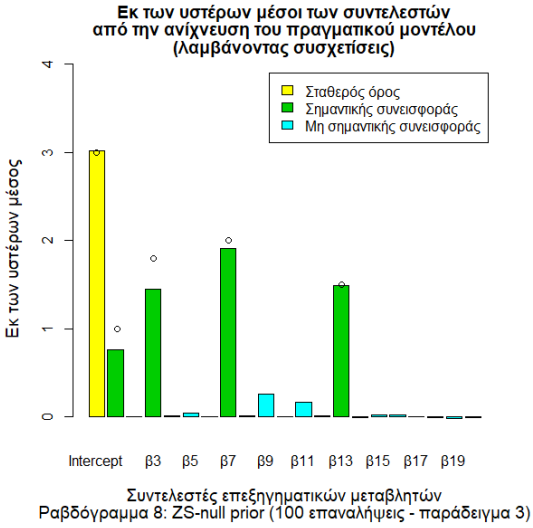


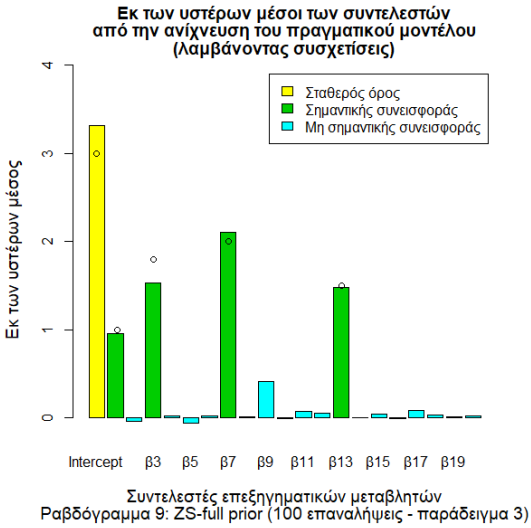
Συντελεστές επεξηγηματικών μεταβλητών
Ραβδόγραμμα 5: hyper-g/n prior (100 επαναλήψεις - παράδειγμα 3



Συντελεστές επεξηγηματικών μεταβλητών
Ραβδόγραμμα 6: EBL prior (100 επαναλήψεις - παράδειγμα 3)



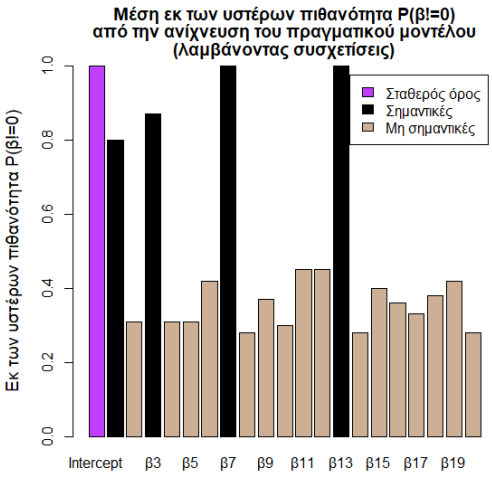




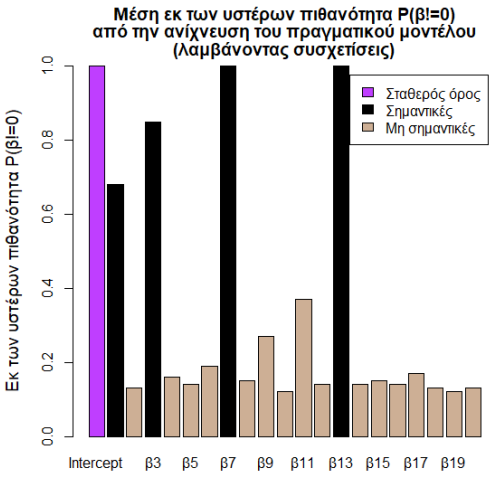
Ραβδογράμματα 1-9 (α): Εκ των υστέρων μέσοι των συντελεστών των επεξ. μεταβλητών κατά μέσο όρο από την ανίχνευση του πραγματικού μοντέλου στις 100 επαναλήψεις (λαμβάνοντας συσχετίσεις).

Ραβδογράμματα 1-9 (α) – Παρατηρήσεις:

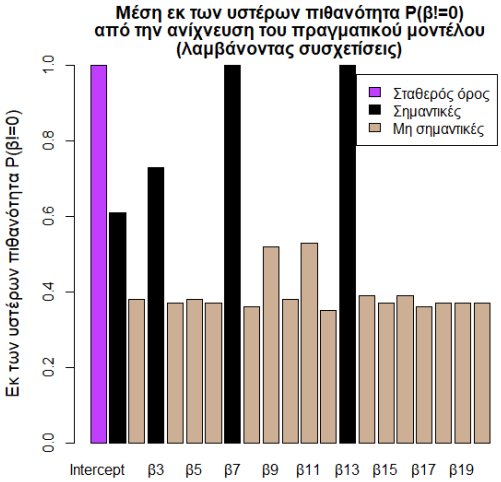
Συμπληρωματικά με όσα αναφέραμε στα αντίστοιχα διαγράμματα του παραδείγματος 2, αυτό που έχουμε να παρατηρήσουμε επιπλέον εδώ είναι ότι η μέση εκ των υστέρων τιμή για τους συντελεστές των επεξηγηματικών μεταβλητών X9 και X11 που στο προηγούμενο παράδειγμα ήταν σχεδόν μηδενική, πλέον με την ενσωμάτωση των συσχετίσεων έχει αυξηθεί αισθητά.



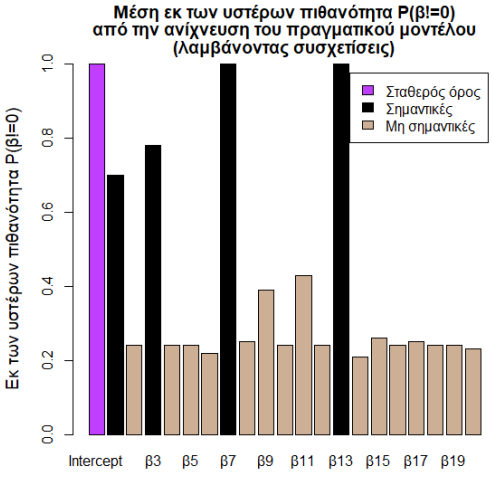
Ραβδόγραμμα 1: AIC prior (100 επαναλήψεις - παράδειγμα 3)



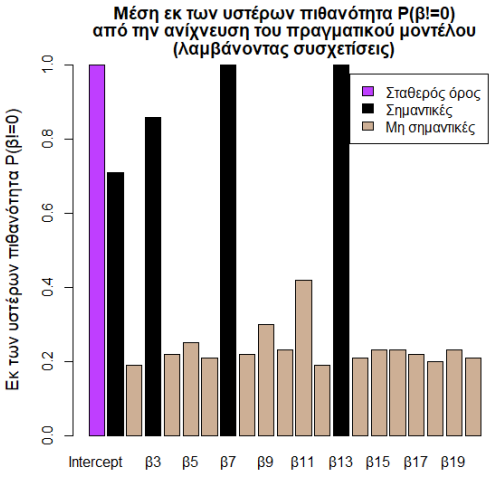
Ραβδόγραμμα 2: BIC prior (100 επαναλήψεις - παράδειγμα 3)



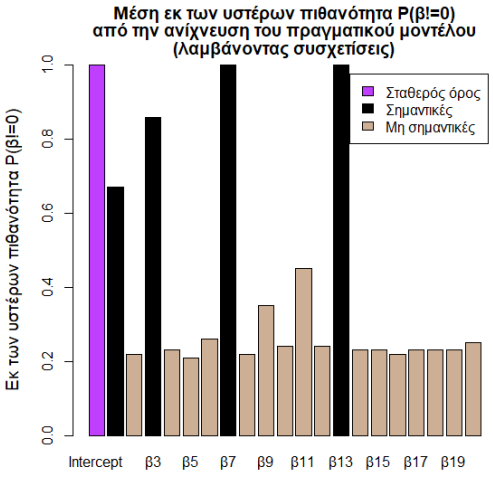
Ραβδόγραμμα 3: g prior (100 επαναλήψεις - παράδειγμα 3)



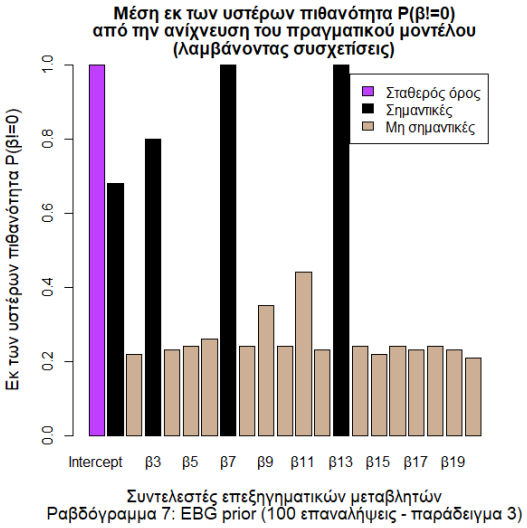
Ραβδόγραμμα 4: hyper-g prior (100 επαναλήψεις - παράδειγμα 3)



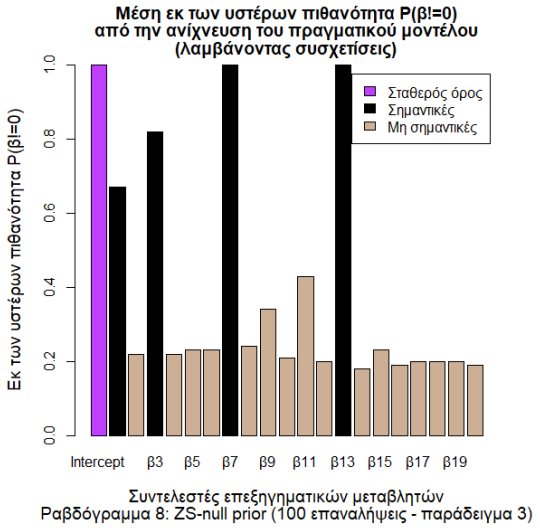
Ραβδόγραμμα 5: hyper-g/n prior (100 επαναλήψεις - παράδειγμα 3



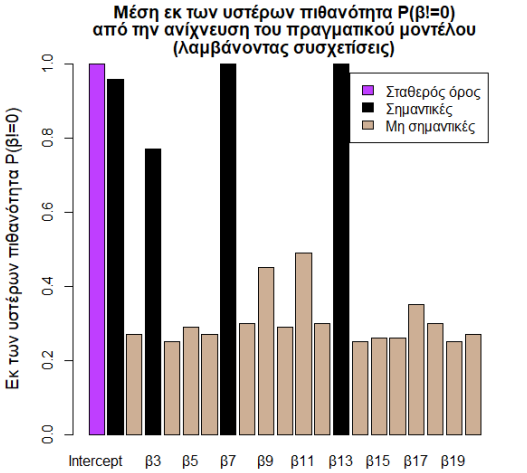
Ραβδόγραμμα 6: EBL prior (100 επαναλήψεις - παράδειγμα 3)



Ραβδόγραμμα 7: EBG prior (100 επαναλήψεις - παράδειγμα 3)



Ραβδόγραμμα 8: ZS-null prior (100 επαναλήψεις - παράδειγμα 3)



Ραβδόγραμμα 9: ZS-full prior (100 επαναλήψεις - παράδειγμα 3)

Ραβδογράμματα 1-9 (β): Μέση εκ των υστέρων πιθανότητα p(β!=0) από την ανίχνευση του πραγματικού μοντέλου στις 100 επαναλήψεις:

Ραβδογράμματα 1-9 (β) – Παρατηρήσεις:

Ομοίως και εδώ, συνοδευτικά με όσα αναφέρθηκαν στα αντίστοιχα ραβδογράμματα του παραδείγματος 2, είναι ορατό ότι η ενσωμάτωση των συσχετίσεων προκάλεσε με όλες τις εκ των προτέρων κατανομές, αισθητή αύξηση της εκ των υστέρων πιθανότητας συμπερίληψης των επεξηγηματικών μεταβλητών X9 και X11, εκτός της περίπτωσης του κριτηρίου AIC που δεν είναι τόσο διακριτή η αλλαγή αυτή δεδομένου του γεγονότος ότι όπως έχει ήδη αναφερθεί το συγκεκριμένο κριτήριο έδινε εξαρχής υψηλές εκ των υστέρων

πιθανότητες συμπερίληψης σε στατιστικά μη σημαντικές επεξηγηματικές μεταβλητές.

ΚΕΦΑΛΑΙΟ 6

ΑΝΑΚΕΦΑΛΑΙΩΣΗ-ΣΥΜΠΕΡΑΣΜΑΤΑ

Μελετώντας θεωρητικά και εμπειρικά από τη σκοπιά της Μπεϋζιανής στατιστικής το πρόβλημα της επιλογής μοντέλου και μεταβλητών που πραγματευτήκαμε στην παρούσα διπλωματική εργασία, ανακεφαλαιώνουμε τα κύρια σημεία του ενδιαφέροντός μας, εξάγωντας κάποια χρήσιμα συμπεράσματα για την ανοιχτή περιοχή της Μπεϋζιανής στατιστικής που μελετά την επίδραση των διαφόρων εκ των προτέρων κατανομών των συντελεστών του μοντέλου, στα εκ των υστέρων αποτελέσματα.

Περιορίζοντας την ανάλυσή μας στην περίπτωση του κανονικού γραμμικού μοντέλου, το κύριο ζήτημα που διερευνήσαμε, αφορούσε το πώς από ένα σύνολο 2^p το πλήθος υποψήφιων συγκριθέντων μοντέλων θα μπορούσαμε με μια Μπεϋζιανή προσέγγιση να επιλέξουμε εκείνο με την καλύτερη προσαρμογή στα διαθέσιμα δεδομένα, ούτως ώστε να εξασφαλίσουμε ταυτόχρονα οικονομία και εύκολη ερμηνεία αλλά και να αντιμετωπίσουμε κάποια από τα μειονεκτήματα της κλασικής προσέγγισης που αναφέρθηκαν στο κύριο περιεχόμενο της παρούσας εργασίας.

Η προσέγγιση που ακολουθήσαμε, αποσκοπούσε στην επιλογή του μοντέλου υψηλότερης εκ των υστέρων πιθανότητας HPM (Highest Probability Model), για το οποίο ήταν απαραίτητο να θέσουμε μία εκ των προτέρων κατανομή $\mathrm{f}(\mathrm{M}_\gamma)$ για κάθε μοντέλο M_γ και μία εκ των προτέρων κατανομή για τις άγνωστες παραμέτρους του μοντέλου (θεωρώντας το σταθερό όρο κοινό για όλα τα μοντέλα).

Στα πρώτα 2 κεφάλαια παρακάμψαμε το κυρίως περιεχόμενο της παρούσας διπλωματικής εργασίας ούτως ώστε να ορίσουμε το πρόβλημα της επιλογής μεταβλητών στα κανονικά γραμμικά μοντέλα και να αναφερθούμε σε κάποιες βασικές έννοιες της Μπεϋζιανής στατιστικής απαραίτητες για να κατανοήσουμε τη σπουδαιότητα και το ρόλο των εκ των προτέρων κατανομών.

Στο Κεφάλαιο 3 μεταβήκαμε στο κύριο μέρος της διπλωματικής εργασίας για να ορίσουμε κάποιες ευρέως διαδομένες εκ των προτέρων κατανομές που χρησιμοποιούνται για τους συντελεστές του μοντέλου ούτως ώστε να διευκολύνονται οι υπολογισμοί των εκ των υστέρων κατανομών και να λαμβάνουμε κλειστές μορφές και σαφή ερμηνεία για κάποια μέτρα σύγκρισης των υποψήφιων μοντέλων όπως είναι η εκ των υστέρων περιθώρια πιθανοφάνεια και ο παράγοντας Bayes.

Ξεκινήσαμε λοιπόν παρουσιάζοντας τη g prior του Zellner, παραθέτοντας κάποια γνωστά κριτήρια όπως το κριτήριο μοναδιαίας πληροφορίας, τα κριτήρια AIC και BIC και τις EB (local and global) διδικασίες που συμβάλλουν στο να εκτιμήσουμε την υπερπαράμετρο g και συνεχίσαμε με την παρουσίαση της hyper-g prior, της Zellner and Siow prior και της hyper-g/n (γνωστές ως μίξεις g prior).

Δεδομένου ότι η επιλογή της υπερπαραμέτρου g ήταν καθοριστικής σημασίας, καθώς είδαμε ότι ακατάλληλες τιμές μπορεί να οδηγήσουν στο παράδοξο των Lindley-Bartlett και το παράδοξο πληροφορίας, μεταβήκαμε στο Κεφάλαιο 4 όπου μελετήσαμε κάποιες από τις ασυμπτωτικές ιδιότητες των παραπάνω εκ των προτέρων κατανομών και θέματα συνέπειας που αφορούσαν το κατα πόσο οι παραπάνω επιλογές επιλύουν το παράδοξο πληροφορίας και οδηγούν στην επιλογή ενός συνεπούς μοντέλου για να γίνονται συνεπείς προβλέψεις.

Η μελέτη της συνέπειας ολοκληρώθηκε εμπειρικά με 3 παραδείγματα στο Κεφάλαιο 5, εφαρμόζοντας την Μπεϋζιανή προσέγγιση για επιλογή μοντέλου και μεταβλητών πρώτα 1 φορά σε πραγματικά δεδομένα και κατόπιν 50 φορές σε προσομοιωμένα δεδομένα με την προϋπόθεση ότι οι επεξηγηματικές μεταβλητές ήταν ανεξάρτητες και 100 φορές σε προσομοιωμένα δεδομένα όπου θεωρήσαμε ότι μεταξύ κάποιων επεξηγηματικών μεταβλητών υπήρχε συσχέτιση. Για την ανάλυση χρησιμοποιήσαμε το πακέτο BAS της R, δοκιμάζοντας όλες τις προτεινόμενες τιμές για την παράμετρο prior και θέτοντας ομοιόμορφη εκ των προτέρων κατανομή για κάθε μοντέλο (παράμετρος modelprior) ούτως ώστε όλα τα μοντέλα να θεωρηθούν ισοπίθανα.

Στον παρόν Κεφάλαιο συνδυάζοντας θεωρία και πράξη από τα προηγούμενα κεφάλαια, παρουσιάζουμε τα εξαγώμενα συμπεράσματα, σύμφωνα με τα οποία:

Μολονότι η g prior του Zellner αποτελεί μία δημοφιλή επιλογή λόγω του γεγονότος ότι οδηγεί σε απλούς υπολογισμούς, εύκολη ερμηνεία και κλειστές μορφές για την περιθώρια πιθανοφάνεια του μοντέλου και τον παράγοντα Bayes, είδαμε ότι αν η τιμή της υπερπαραμέτρου g δεν επιλεγεί σωστά μπορεί να οδηγήσει σε ανεπιθύμητες καταστάσεις οι οποίες περιγράφτηκαν ορίζοντας το παράδοξο των Lindley-Bartlett και το παράδοξο πληροφορίας, καταλήγοντας στο συμπέρασμα ότι θα πρέπει να αποφεύγονται διακεχυμένες αλλά και μη γνήσιες εκ των προτέρων κατανομές, μιας και αυτές με τη σειρά τους οδηγούν σε ασάφειες ως προς την ερμηνεία του παράγοντα Bayes ή ακόμα και σε ακαθόριστες μορφές.

Προκειμένου να αντιμετωπιστούν οι δυσκολίες που αφορούσαν την επιλογή τιμής για την υπερπαράμετρο g, αναπτύχθηκαν οι EB (local and global) διδιαδικασίες οι οποίες συνδυάζουν τα κριτήρια AIC και BIC διατηρώντας την καλή ιδιότητα της g να οδηγούν και αυτές σε κλειστές μορφές των περιθώριων πιθανοφανειών και του παράγοντα Bayes. Επιπρόσθετα όμως οι EB διαδικασίες είδαμε ότι έχουν το πλεονέκτημα για σταθερά $n, p < n$ και $R_Y^2 \to 0$ να επιλύουν και το παράδοξο πληροφορίας, ενώ εμπειρικά διαπιστώσαμε ότι ήταν ικανές να ανιχνεύσουν αρκετές φορές το πραγματικό μοντέλο, με υψηλή εκ των υστέρων πιθανότητα.

Όσον αφορά την επιλογή της prior AIC όπως και την επιλογή της prior ZS-full διαπιστώσαμε ότι δεν ήταν τόσο αποδοτικές καθώς υστερούσαν και ως προς τον αριθμό ανίχνευσης του πραγματικού μοντέλου, και ως προς την τιμή της εκ των υστέρων πιθανότητας του μοντέλου αλλά και ως προς το γεγονός ότι έδιναν υψηλές εκ των υστέρων πιθανότητες συμπερίληψης σε στατιστικά μη σημαντικές επεξηγηματικές μεταβλητές, ενώ είχαν ταυτόχρονα την τάση να επιλέγουν μοντέλα μεγαλύτερης διάστασης από την πραγματική.

Το κριτήριο BIC από την άλλη, μολονότι εμπειρικά φάνηκε να παρουσιάζει πολύ ικανοποιητικά εκ των υστέρων αποτελέσματα (δεδομένου ότι επιλέχθηκε μεγάλο μέγεθος δείγματος n συγκριτικά με το πλήθος των άγνωστων παραμέτρων), ως γνωστόν υστερεί να χειριστεί προβλήματα μεγάλης διαστάσεως ενώ παρουσιάζει την τάση να επιλέγει τα πιο φειδωλά μοντέλα και είδαμε παράλληλα κάπως μεγαλύτερα τυπικά σφάλματα για τους εκτιμώμενους μέσους των συντελεστών.

Κατά συνέπεια οι EB priors από τη στιγμή που συνδυάζουν τα κριτήρια AIC και BIC αναμένεται αφενός να επιλέγουν και αυτές μεγαλύτερης

διάστασης μοντέλα και αφετέρου να μην είναι τόσο αποδοτικές για μεγάλης διαστάσεως προβλήματα. Το τελευταίο διαπιστώθηκε και εμπειρικά αφού αυξάνοντας στα προσομοιωμένα δεδομένα τη διάσταση από 15 επεξηγηματικές μεταβλητές σε 20 παρουσιάστηκαν μεγάλες απαιτήσεις μνήμης και χρονοκαθυστέρηση στην εκτέλεση ενώ δεν θα ήταν απίθανο να παρουσιαστεί απροσδόκητη διακοπή της εκτέλεσης επιλέγοντας ακόμα μεγαλύτερη διάσταση. Στην περίπτωση μάλιστα που το πραγματικό μοντέλο θα επιλέγονταν να είναι το «μηδενικό», οι EB διαδικασίες δε θα το ανίχνευαν ποτέ.

Για λόγους συνέπειας προτάθηκαν οι μίξεις g εκ των προτέρων κατανομών και συγκεκριμένα η hyper-g prior και η Zellner and Siow prior οι οποίες επιλύουν το παράδοξο πληροφορίας κάτω από συγκεκριμένους περιορισμούς που αναφέρθηκαν στο Κεφάλαιο 4, ενώ επίσης (όπως και οι EB prior) παρουσιάζουν συνεπείς εκ των υστέρων πιθανότητες μοντέλου κάτω από οποιοδήποτε μοντέλο πέραν του εκφυλισμένου. Στην τελευταία περίπτωση η συνέπεια διατηρείται μόνο υπό της Zellner and Siow prior λόγος που οδήγησε στο να αναπτυχθεί και η hyper-g/n prior καθώς παρατηρήθηκε ότι υπό τους Zellner and Siow, η prior που προτάθηκε για την υπερπαράμτερο g είχε εξάρτηση από το μέγεθος του δείγματος n.

Πράγματι όπως διαπιστώθηκε και εμπειρικά η hyper-g/n και η ZS-null φάνηκε να έχουν την καλύτερη αποδοτικότητα καθώς ακόμα και στο παράδειγμα όπου συμπεριλάβαμε συσχετίσεις ανίχνευσαν περισσότερες φορές από οποιαδήποτε άλλη το πραγματικό μοντέλο. Ωστόσο η ένταξη συσχετίσεων είδαμε ότι μειώνει την αποδοτικότητά τους καθώς ξεκίνησαν και οι μη στατιστικά σημαντικές μεταβλητές να γίνονται σημαντικές αυξάνοντας την εκ των υστέρων τους πιθανότητα συμπερίληψης.

Πάντως όπως διαπιστώσαμε εμπειρικά, για τη συγκεκριμένη επιλογή μεγέθους δείγματος και πλήθους επεξηγηματικών μεταβλητών τα εκ των υστέρων αποτελέσματα κάτω από τις EB priors, τη hyper-g, τη hyper-g/n και τη ZS-null διαφοροποιούνται ελάχιστα και όπως έχουμε αναφέρει στο Κεφάλαιο 4 ο BMA (Bayesian Model Average) εκτιμητής πρόβλεψης είναι συνεπής εκτιμητής για όλες αυτές τις περιπτώσεις.

ΠΑΡΑΡΤΗΜΑ Α

Κεφάλαιο - 4.5.1: Κλειστές μορφές του παράγοντα Bayes

Απόδειξη σχέσης 4.5.1.1:

Θέλουμε νδο: $\mathrm{BF}_{\mathrm{M_\gamma M_F}} = (1+\mathrm{g})^{(\mathrm{n-p-1})/2}\left[1+\mathrm{g}\frac{1-\mathrm{R_F^2}}{1-\mathrm{R_\gamma^2}}\right]^{(\mathrm{n-p_\gamma-1})/2}$.

Από τον ορισμό του παράγοντα Bayes έχουμε:

$$\mathrm{BF}_{\mathrm{M_\gamma M_F}} = \frac{\mathrm{f(Y|M_\gamma, g)}}{\mathrm{(Y|M_F, g)}} = \frac{\frac{(1+\mathrm{g})^{\frac{\mathrm{n-1-p_\gamma}}{2}}}{\left[1+\mathrm{g}\left(1-\mathrm{R_\gamma^2}\right)\right]^{\frac{\mathrm{n-1}}{2}}}}{\frac{(1+\mathrm{g})^{\frac{\mathrm{n-1-p}}{2}}}{[1+\mathrm{g}(1-\mathrm{R_F^2})]^{\frac{\mathrm{n-1}}{2}}}}$$

$$= (1+\mathrm{g})^{-(\mathrm{n-1-p})/2}(1+\mathrm{g})^{\frac{\mathrm{n-1-p_\gamma}}{2}}\left[\frac{1+\mathrm{g}(1-\mathrm{R_F^2})}{1+\mathrm{g}\left(1-\mathrm{R_\gamma^2}\right)}\right]^{(\mathrm{n-1})/2}.$$

Αρκεί λοιπόν νδο:

$$\left[1+\mathrm{g}\frac{1-\mathrm{R_F^2}}{1-\mathrm{R_\gamma^2}}\right]^{(\mathrm{n-p_\gamma-1})/2} = (1+\mathrm{g})^{\frac{\mathrm{n-1-p_\gamma}}{2}}\left[\frac{1+\mathrm{g}(1-\mathrm{R_F^2})}{1+\mathrm{g}\left(1-\mathrm{R_\gamma^2}\right)}\right]^{(\mathrm{n-1})/2}.$$

Έχουμε:

$$\left[1+\mathrm{g}\frac{1-\mathrm{R_F^2}}{1-\mathrm{R_\gamma^2}}\right]^{(\mathrm{n-p_\gamma-1})/2} = \left[\frac{1-\mathrm{R_\gamma^2}+\mathrm{g}(1-\mathrm{R_F^2})}{1-\mathrm{R_\gamma^2}}\right]^{(\mathrm{n-p_\gamma-1})/2}$$

$$= \left[\frac{1-\mathrm{R_\gamma^2}+\mathrm{g}(1-\mathrm{R_F^2})}{1-\mathrm{R_\gamma^2}}\frac{(1+\mathrm{g})}{(1+\mathrm{g})}\right]^{(\mathrm{n-p_\gamma-1})/2}$$

$$= (1+\mathrm{g})^{\frac{\mathrm{n-1-p_\gamma}}{2}}\left[\frac{1-\mathrm{R_\gamma^2}+\mathrm{g}(1-\mathrm{R_F^2})}{1-\mathrm{R_\gamma^2}+\mathrm{g}(1-\mathrm{R_\gamma^2})}\right]^{(\mathrm{n-p_\gamma-1})/2}.$$

Απομένει λοιπόν νδο:

$$\left[\frac{1-\mathrm{R_\gamma^2}+\mathrm{g}(1-\mathrm{R_F^2})}{1-\mathrm{R_\gamma^2}+\mathrm{g}(1-\mathrm{R_\gamma^2})}\right]^{(\mathrm{n-p_\gamma-1})/2} = \left[\frac{1+\mathrm{g}(1-\mathrm{R_F^2})}{1+\mathrm{g}\left(1-\mathrm{R_\gamma^2}\right)}\right]^{(\mathrm{n-1})/2}.$$

Ξεκινώντας από το β μέλος έχουμε:

$$\left[\frac{1+\mathrm{g}(1-\mathrm{R_F^2})}{1+\mathrm{g}\left(1-\mathrm{R_\gamma^2}\right)}\right]^{\frac{\mathrm{n-1}}{2}} =$$

$$= \left[\left(\frac{1+g(1-R_F^2)-R_\gamma^2}{1+g\left(1-R_\gamma^2\right)-R_\gamma^2}\right)^{\frac{p_\gamma}{n-1}}\left(\frac{1+g(1-R_F^2)-R_\gamma^2}{1+g\left(1-R_\gamma^2\right)-R_\gamma^2}\right)^{-\frac{p_\gamma}{n-1}}\frac{1+g(1-R_F^2)}{1+g\left(1-R_\gamma^2\right)}\right]^{\frac{n-1}{2}}$$

$$\left[\left(\frac{1+g(1-R_F^2)-R_\gamma^2}{1+g\left(1-R_\gamma^2\right)-R_\gamma^2}\right)\left(\frac{1+g(1-R_F^2)-R_\gamma^2}{1+g\left(1-R_\gamma^2\right)-R_\gamma^2}\right)^{\frac{p_\gamma}{n-1}-1}\left(\frac{1+g(1-R_F^2)-R_\gamma^2}{1+g\left(1-R_\gamma^2\right)-R_\gamma^2}\right)^{-\frac{p_\gamma}{n-1}}\frac{1+g(1-R_F^2)}{1+g\left(1-R_\gamma^2\right)}\right]^{\frac{n-1}{2}}$$

$$= \left(\frac{1+g(1-R_F^2)-R_\gamma^2}{1+g\left(1-R_\gamma^2\right)-R_\gamma^2}\right)^{(n-1)/2}\left(\left(\frac{1+g(1-R_F^2)-R_\gamma^2}{1+g\left(1-R_\gamma^2\right)-R_\gamma^2}\right)^{-\frac{p_\gamma}{n-1}}\right)^{(n-1)/2}\left(\left(\frac{1+g(1-R_F^2)-R_\gamma^2}{1+g\left(1-R_\gamma^2\right)-R_\gamma^2}\right)^{\frac{p_\gamma}{n-1}-1}\right)^{(n-1)/2}$$

$$\left(\frac{1+g(1-R_F^2)}{1+g\left(1-R_\gamma^2\right)}\right)^{(n-1)/2}$$

$$= \left(\frac{1+g(1-R_F^2)-R_\gamma^2}{1+g\left(1-R_\gamma^2\right)-R_\gamma^2}\right)^{(n-1)/2}\left(\frac{1+g(1-R_F^2)-R_\gamma^2}{1+g\left(1-R_\gamma^2\right)-R_\gamma^2}\right)^{-p_\gamma/2}\left(\frac{1+g(1-R_F^2)-R_\gamma^2}{1+g\left(1-R_\gamma^2\right)-R_\gamma^2}\right)^{(p_\gamma-n+1)/2}\left(\frac{1+g(1-R_F^2)}{1+g\left(1-R_\gamma^2\right)}\right)^{(n-1)/2}.$$

Αρκεί λοιπόν ο όρος:

$$A = \left(\frac{1+g(1-R_F^2)-R_\gamma^2}{1+g\left(1-R_\gamma^2\right)-R_\gamma^2}\right)^{(p_\gamma-n+1)/2}\left(\frac{1+g(1-R_F^2)}{1+g\left(1-R_\gamma^2\right)}\right)^{(n-1)/2}$$

να ισούται με τη μονάδα.

Πράγματι παρατηρούμε ότι για $R_\gamma^2 \to R_F^2$, έχουμε $A \to 1$.

Επίσης για $R_\gamma^2 \to 0 \leftrightarrow p_\gamma \to 0$, έχουμε:

$$A = \left(\frac{1+g(1-R_F^2)-R_\gamma^2}{1+g\left(1-R_\gamma^2\right)-R_\gamma^2}\right)^{\frac{-n+1}{2}+\frac{n-1}{2}} = 1$$

και επομένως αποδείχθη το ζητούμενο.

ΠΑΡΑΡΤΗΜΑ Β

Διευκρίνηση:

Ο παρακάτω κώδικας R, αφορά το παράδειγμα 2 με τα προσομοιωμένα δεδομένα και τις ανεξάρτητες επεξηγηματικές μεταβλητές για 50 επαναλήψεις και επιλεγμένη prior τη hyper-g/n. Ο ίδιος κώδικας εφαρμόζεται για όλες τις τιμές της παραμέτρου prior. Ο Κώδικας για το παράδειγμα 1, με τα πραγματικά δεδομένα παρουσιάστηκε στο κυρίως μέρος της διπλωματικής εργασίας ενώ για το παράδειγμα 3 είναι ίδιος με του παραδείγματος 2 με τη μόνη διαφορά ότι αλλάζουμε το πληθος των επαναλήψεων σε 100 και συμπεριλαμβάνουμε σε κάθε επανάληψη στα κύρια βήματα της εκτέλεσης τις συσχετίσεις που περιγράφηκαν στις υποθέσεις (6) και (7) του παράδειγματος 3 του κεφαλαίου 5.

Παράδειγμα 2:

```
# φόρτωση βιβλιοθηκών

library(MASS)

library(BAS)


# Αρχικοποίηση των n, p, b

n<-100

p<-20

b<-matrix(0, nrow=21, ncol=1)

b[1]<-3 # ο σταθερός όρος

b[2]<-1

b[4]<-1.8

b[8]<-2

b[14]<-1.5


Y<-matrix(0, nrow=n, ncol=1) # μεταβλητή απόκρισης

colnames(Y)<-"Y"


# συχνότητα εμφάνισης των επεξηγηματικών μεταβλητών στις 50 επαναλήψεις

TimesXInvolved<-matrix(0, nrow=p+1, ncol=1)
```

```
# αριθμός ανίχνευσης του πραγματικού μοντέλου

TimesRealModelFound <- 0


# συνολικά εκ των υστέρων αποτελέσματα από την ανίχνευση του πραγματικού μοντέλου

TotalPostMeanX<-matrix(0, nrow=p+1, ncol=1)

TotalPostSdX<-matrix(0, nrow=p+1, ncol=1)

TotalPostProbCoef <-matrix(0, nrow=p+1, ncol=1)

TotalPostLogMarginLkl<-0

TotalProbRealModel<-0

TotalDim<-0

AllProbRealModel<-NULL

AllPostLogMarginLkl<-NULL


# κύρια βήματα εκτέλεσης του αλγορίθμου για την ανίχνευση του πραγματικού μοντέλου στις 50 επαναλήψεις

 for (i in 1:50)

{

  TrueModelHasFound <- as.logical("True")

  ExtraVariableExists<-as.logical("False")

  X<-matrix(rnorm(n*p), nrow=n, ncol=p) # επεξηγηματικές μεταβλητές

  colnames(X)<-paste("X", seq(1:p), sep="")

  for (j in 1:100)

  {

   Y[j]<- rnorm(1, b[1]+b[2]*X[j,1]+b[4]*X[j,3]+b[8]*X[j,7]+b[14]*X[j,13], 2.5)

  }

  SimulatedData<-data.frame(cbind(X,Y)) # το dataset των προσομοιωμένων τιμών

  Results<-bas.lm(Y ~ ., SimulatedData, prior="hyper-g-n",modelprior=uniform(),alpha=3)

  HighestProbModel<-summary(Results)[,2]  # μοντέλο υψηλότερης εκ των υστέρων πιθανότητας

  TotalDim<-TotalDim+HighestProbModel[25]

  for (k in 1:(p+1)){
```

```
  TimesXInvolved[k]<-TimesXInvolved[k]+HighestProbModel[k]
  }
 #θέσεις που αντιστοιχούν στις συμπεριλαμβανόμενες επεξηγηματικές μεταβλητές
  IndexesOfXEqual1<-which(HighestProbModel %in% 1)
  for (m in 1:length(IndexesOfXEqual1))
  { ExtraVariableExists<-!is.element(IndexesOfXEqual1[m], c(1,2,4,8,14,22 ))  # η θέση 22 αφορά τον BF
# τσεκάρουμε αν έχει συμπεριληφθεί και μεταβλητή που δεν αφορά το πραγματικό μοντέλο
    if (ExtraVariableExists) {
      TrueModelHasFound <-as.logical("False")}
  }
  if (TrueModelHasFound)
  {
   TimesRealModelFound<-TimesRealModelFound+1
   TotalProbRealModel<-TotalProbRealModel + HighestProbModel[23]
   TotalPostLogMarginLkl<-TotalPostLogMarginLkl + HighestProbModel[26]
   AllProbRealModel<-cbind(AllProbRealModel,HighestProbModel[23])
   AllPostLogMarginLkl<-cbind(AllPostLogMarginLkl, HighestProbModel[26])
   TotalPostMeanX<-TotalPostMeanX+coef(Results)[[1]]
   TotalPostSdX<-TotalPostSdX+coef(Results)[[2]]
   TotalPostProbCoef<-TotalPostProbCoef+coef(Results)[[3]]
  }
}
#Εξαγώμενα αποτελέσματα
TimesXInvolved
TimesRealModelFound # (<>0)
TotalDim/50 # κατά μέσο όρο επιλεγόμενη διάσταση για το HPM
Round(TotalPorobRealModel/TimesRealModelFound,2)
Round(TotalPostLogMarginLkl/TimesRealModelFound,2)
```

```
sd(AllProbRealModel)

sd(AllPostLogMarginLkl)

round(TotalPostMeanX/TimesRealModelFound, 2) # εκ των υστέρων μέσοι των συντελεστών

round(TotalPostSdX/TimesRealModelFound, 2) # εκ των υστέρων τυπικές αποκλίσεις

round(TotalPostProbCoef/TimesRealModelFound, 2) # εκ των υστέρων πιθανότητες συμπερίληψης
```

Βιβλιογραφία

[1] Anabel Forte, Gonzalo Garcia-Donato and Mark F.J. Steel. Methods and Tools for Bayesian Variable Selection and Model Averaging in Normal Linear Regression.

[2] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. Bayesian Data Analysis, Second Edition. Chapman & Hall/CRC, July 2003.

[3] Carmen Fernandez, Eduardo Ley, Mark F.J. Steel (2000). In Benchmark priors for Bayesian model averaging. Journal of Econometrics 100 (2001) 381-427

[4] Clyde, M. (1999) Bayesian model averaging and model search strategies (with discussion). In Bayesian Statistics 6 (eds J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith), pp. 157–185. John Wiley.

[5] Dellaportas, P., Forster, J.J. and Ntzoufras, I. (1998) On Bayesian model and variable selection using MCMC. Working Paper. Athens University of Economics and Business.

[6] Feng Liang, Rui Paulo, German Molina, , M. ((2008)). Mixtures of g Priors for Bayesian Variable Selection. American Statistical Association Journal of the American Statistical Association March 2008.

[7] Fouskakis, D. (n.d.). Bayesian Variable Selection (An Introduction Tutorial*)*. Retrieved from www.math.ntua.gr./~fouskakis/bvs.pdf.

[8] George Cassela, E.M. (2007). In Objective Bayesian Model Selection : Some Methods, Some Theory. International Workshop on Statistical Modelling 2007.

[9] Hoeting, J. A. (n.d.). Methodology for Bayesian Model Averaging: An Update. Colorado State University.

[10] Jennifer A. Hoeting, D. M. (1999). Bayesian Model Averaging: A Tutorial. Statistical Science 1999, Vol. 14, No. 4, 382–417.

[11] Kass, R.A. and Raftery, A.E. (1995) Bayes' factor. Journal of the American Statistical Association, 90,773–795

[12] Lopes, H. F. (October, 30th 2002). Bayesian Model Selection. Universidade Federal do Rio de Janeiro, BRAZIL.

[13] Martin Feldkircher and Stefan Zeugner (2009). Benchmark Priors Revisited: On Adaptive Shrinkage and the Supermodel Effect in Bayesian Model Averaging. International Monetary Fund, 2009.

[14] Merlise Clyde , Joyee Ghosh and Michael Littman. Bayesian Adaptive Sampling for Variable Selection and Model Averaging.

[15] Steve Geinitz (2009). Prior Covariance Choices and the g Prior. Seminar on Bayesian Linear Model Institute Mathematics University Zurich.

[16] Yuzo Maruyama and Edward I. George (2011). In Fully Bayes factors with a generalized g prior. The Annals of Statistics 2011, Vol. 39, No. 5, 2740–2765, Institute of Mathematical Statistics, 2011.