Διπλωματική Εργασία

Γκίνης Ιωάννης

Εθνικό Μετσόβιο Πολυτεχνείο

ΔΠΜΣ:
Μαθηματική Προτυποποίηση σε Σύγχρονες Τεχνολογίες και τη Χρηματοοικονομική

Τίτλος:

Μπεϋζιανή επιλογή μεταβλητών σε προβλήματα παλινδρόμησης με παρουσία ελλιπών τιμών

Τριμελής επιτροπή:

Επιβλέπων: Φουσκάκης Δημήτριος, Καθηγητής, ΕΜΠ
Ντζούφρας Ιωάννης, Καθηγητής, ΟΠΑ
Λουλάκης Μιχαήλ, Καθηγητής, ΕΜΠ

Αθήνα, Οκτώβριος 2023

Αφιερώνεται στην οικογένειά μου και ειδικά στη μητέρα μου που πάντα με στηρίζει.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέπων καθηγητή μου κο. Φουσκάκη Δημήτριο, καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου, για την καθοριστική βοήθεια και καθοδήγηση που μου παρείχε καθ'όλη τη διάρκεια εκπόνησης της παρούσας διπλωματικής εργασίας καθώς επίσης και τα μέλη της τριμελούς μου επιτροπής κο. Ντζούφρα Ιωάννη, καθηγητή του Οικονομικού Πανεπιστημίου Αθηνών και κο. Μιχαήλ Λουλάκη, καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου.

Είμαι ευγνώμων για την υπομονή, την ενθάρρυνση και την προθυμία τους να μοιραστούν τις γνώσεις τους, οι οποίες εμπλούτισαν την κατανόησή μου για την Μπεϋζιανή στατιστική. Αυτή η διπλωματική δεν θα ήταν δυνατή χωρίς την καθοδήγηση και την ενθάρρυνση τους και γι'αυτό τους είμαι πραγματικά ευγνώμων.

Τέλος, θα ήθελα να εκφράσω τις ευχαριστίες μου στην οικογένεια και τους φίλους μου για την αμέριστη υποστήριξη, κατανόηση και ενθάρρυνση σε όλη αυτή την προσπάθειά μου. Η πίστη που τόσο απλόχερα δείξανε σε μένα, υπήρξε κινητήριος δύναμη για την επίτευξη αυτού του έργου.

Περίληψη

Η παρούσα διατριβή εμβαθύνει στις θεμελιώδεις αρχές της Μπεϋζιανής στατιστικής και στην εφαρμογή τους στο πλαίσιο της επιλογής μεταβλητών. Ο πρωταρχικός στόχος είναι να διερευνηθεί διεξοδικά ο τρόπος με τον οποίο αποδίδουν οι Μπεϋζιανές μέθοδοι επιλογής μεταβλητών, όταν έρχονται αντιμέτωποι με δεδομένα που περιέχουν ελλιπείς τιμές. Η εμφάνιση ελλιπών τιμών σε δεδομένα αποτελεί ένα διάχυτο ζήτημα στις στατιστικές αναλύσεις, που συχνά προκύπτει από διάφορους μηχανισμούς, όπως η εντελώς τυχαία έλλειψη (MCAR), η τυχαία έλλειψη (MAR) και η μη τυχαία έλλειψη (NMAR).

Η μελέτη μας ξεκινά με την παροχή μιας εισαγωγής στην Μπεϋζιανή στατιστική, καλύπτοντας βασικές έννοιες, όπως το θεώρημα του Bayes, την επιλογή των εκ των προτέρων κατανομών και τους αλγόριθμους Markov Chain Monte Carlo (MCMC). Εν συνεχεία, αναλύουμε τις τρεις Μπεϋζιανές μεθόδους επιλογής μεταβλητών με τις οποίες θα ασχοληθούμε: Stochastic Search Variable Selection (SSVS), τον δειγματολήπτη Kuo & Mallick και τη μέθοδο Gibbs Variable Selection (GVS). Τέλος, εξετάζουμε σχολαστικά τις μεθόδους διαχείρισης ελλειπών τιμών, αναλύοντας μεθόδους απλής αντικατάστασης, πολλαπλής αντικατάστασης και μεθόδους αντικατάστασης μέσω της Μπεϋζιανής σκέψης και αποσαφηνίζουμε τις βασικές αρχές τους, συζητώντας τα δυνατά και τα αδύνατα σημεία τους στο χειρισμό των ελλιπών τιμών.

Όπως αναφέραμε και παραπάνω, η διπλωματική αυτή επικεντρώνεται σε τρεις Μπεϋζιανές μεθόδους επιλογής μεταβλητών: τη μέθοδο Stochastic Search Variable Selection (SSVS), τον δειγματολήπτη Kuo & Mallick και τη μέθοδο Gibbs Variable Selection (GVS). Αυτές οι μέθοδοι εφαρμόζονται και αξιολογούνται αυστηρά σε προσομοιωμένα δεδομένα που έχουν κατασκευαστεί εσκεμμένα ώστε να περιέχουν ελλιπείς τιμές αντιπροσωπεύοντας καθέναν από τους τρεις μηχανισμούς MCAR, MAR και NMAR.

Μέσω μίας συστηματικής ανάλυσης των αποτελεσμάτων, η παρούσα διατριβή εξάγει συμπεράσματα σχετικά με την προσαρμοστικότητα και αποτελεσματικότητα αυτών των μεθόδων επιλογής μεταβλητών (σε δεδομένα με ελλιπείς τιμές). Επιπλέον, παρέχει σημαντικές πληροφορίες σχετικά με την απόδοση τους, ρίχνοντας φως σε περιπτώσεις όπου ορισμένες μέθοδοι υπερέχουν ή αντιμετωπίζουν προβλήματα. Τα ευρήματα αυτής της μελέτης στοχεύουν να συμβάλουν στην κατανόηση των προβλημάτων που προκύπτουν στα αποτελέσματα των μεθόδων όταν εφαρμόζονται σε δεδομένα με ελλιπείς τιμές.

Περιεχόμενα

1 Εισαγωγή στην Μπεϋζιανή Στατιστική και MCMC **1**
1.1 Εισαγωγή . . . 1
1.2 Σύγκριση της Κλασικής με την Μπεϋζιανή Στατιστική . . . 1
1.3 Θεώρημα του Bayes . . . 3
1.4 Επιλογή της εκ των προτέρων κατανομής . . . 4
1.4.1 Συζυγείς εκ των προτέρων κατανομές . . . 5
1.4.2 Επίπεδες εκ των προτέρων κατανομές . . . 5
1.4.3 Εκ των προτέρων κατανομές αναφοράς . . . 6
1.4.4 Ασαφής εκ των προτέρων κατανομές . . . 7
1.4.5 Ιεραρχικές εκ των προτέρων κατανομές . . . 8
1.5 Περιθώρια πιθανοφάνεια . . . 9
1.6 Αλγόριθμοι Markov chain Monte Carlo . . . 10
1.6.1 Αλγόριθμος Metropolis-Hastings . . . 10
1.6.2 Δειγματολήπτης Gibbs . . . 12

2 Μπεϋζιανές μέθοδοι επιλογής μοντέλων **14**
2.1 Εισαγωγή . . . 14
2.1.1 Μειονεκτήματα του κλασικού ελέγχου υποθέσεων . . . 14
2.2 Παράγοντας του Bayes και Μπεϋζιανή στάθμιση μοντέλων . . . 15
2.3 Το παράδοξο των Lindley-Bartlett . . . 17
2.4 Τρόποι υπολογισμού της περιθώριας πιθανοφάνειας . . . 19
2.4.1 Μέθοδος του Laplace . . . 19
2.4.2 Κριτήριο BIC . . . 20
2.4.3 Αφελής Monte Carlo εκτιμητής και Δειγματολήπτης Σπουδαιότητας . . . 21
2.4.4 Εκτιμητής Αρμονικού Μέσου και Αντίστροφος Δειγματολήπτης Σπουδαιότητας . . . 22
2.5 Θετικά και αρνητικά των Μπεϋζιανών μεθόδων επιλογής μοντέλων . . . 23

3 Μπεϋζιανές μέθοδοι επιλογής μεταβλητών μοντέλου **25**
3.1 Εισαγωγή . . . 25
3.2 Επιλογή εκ των προτέρων κατανομών για συντελεστές πολλαπλού γραμμικού μοντέλου . . . 25
3.2.1 Χρήση της εκ των προτέρων κατανομής του Jeffreys . . . 26
3.2.2 Χρήση Κανονικής κατανομής ως εκ των προτέρων κατανομή . . . 28
3.2.3 Χρήση της εκ των προτέρων κατανομής του Zellner . . . 30
3.3 Stochastic Search Variable Selection (SSVS) . . . 33
3.4 Δειγματολήπτης Kuo &Mallick . . . 35
3.5 Gibbs Variable Selection (GVS) . . . 37

4 Μεθοδολογίες Διαχείρισης Ελλιπών τιμών **39**
4.1 Εισαγωγή . . . 39
4.2 Μηχανισμοί και δομές ελλιπών τιμών . . . 39
4.3 Απλές τεχνικές χειρισμού ελλιπών τιμών . . . 42
4.4 Πολλαπλή Αντικατάσταση . . . 44
4.5 Μπεϋζιανή προσέγγιση διαχείρισης ελλιπών τιμών . . . 47

5 Εφαρμογές σε προσομοιωμένα δεδομένα με ελλιπείς τιμές 50
5.1 Εισαγωγή 50
5.2 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε πλήρη δεδομένα 50
5.2.1 Εφαρμογή της μεθόδου SSVS 53
5.2.2 Εφαρμογή της μεθόδου Kuo &Mallick 58
5.2.3 Εφαρμογή της μεθόδου GVS 61
5.2.4 Πλήρης απαρίθμηση με τη χρήση του κριτηρίου BIC 65
5.3 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης 68
5.3.1 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές υπό το μηχανισμό MCAR 68
5.3.2 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές υπό το μηχανισμό MAR 78
5.3.3 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές υπό το μηχανισμό NMAR 86
5.4 Σύγκριση των Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης 96
5.5 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές 98
5.5.1 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές υπό το μηχανισμό MCAR 98
5.5.2 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές υπό το μηχανισμό MAR 110
5.5.3 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές υπό το μηχανισμό NMAR 120
5.6 Σύγκριση των Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές 131

6 Επίλογος και συμπεράσματα 133

Παράρτημα 136

Βιβλιογραφία 150

Κατάλογος Διαγραμμάτων και Σχημάτων

1.4.4.1 Απεικόνιση των Κανονικών κατανομών $N(0,1)$ και $N(0,100)$ . . 7
1.4.4.2 Απεικόνιση των εκ των υστέρων κατανομών με πρότερες κατανομές την $N(0,1)$ και την $N(0,100)$. 8
4.2.1 Θηκογράμματα του παρατηρήσιμου $\boldsymbol{U}_2$ και του μη παρατηρήσιμου $\boldsymbol{U}_2$ για κάθε μηχανισμό. 41
5.3.1.1 Θηκογράμματα των τιμών του Y_{miss} και του Y_{obs}. 69
5.3.2.1 Θηκογράμματα των τιμών του Y_{miss} και του Y_{obs}. 79
5.3.3.1 Θηκογράμματα των τιμών του Y_{miss} και του Y_{obs}. 87
5.4.1 Ραβδογράμματα των δεικτών $\boldsymbol{\gamma}$ στις μεθόδους SSVS, Kuo &Mallick και GVS σε δεδομένα με μηχανισμούς έλλειψης στη μεταβλητή απόκρισης Y. 97
5.5.1.1 Θηκογράμματα των τιμών των $X_{3\text{miss}}$, $X_{3\text{obs}}$ και $X_{7\text{miss}}$, $X_{7\text{obs}}$. . . 99
5.5.1.2 Θηκογράμματα των τιμών των $X_{10\text{miss}}$, $X_{10\text{obs}}$ και $X_{11\text{miss}}$, $X_{11\text{obs}}$. 101
5.5.2.1 Θηκογράμματα των τιμών των $X_{3\text{miss}}$, $X_{3\text{obs}}$ και $X_{7\text{miss}}$, $X_{7\text{obs}}$. . . 111
5.5.2.2 Θηκογράμματα των τιμών των $X_{10\text{miss}}$, $X_{10\text{obs}}$ και $X_{11\text{miss}}$, $X_{11\text{obs}}$. 113
5.5.3.1 Θηκογράμματα των τιμών των $X_{3\text{miss}}$, $X_{3\text{obs}}$ και $X_{7\text{miss}}$, $X_{7\text{obs}}$. . . 122
5.5.3.2 Θηκογράμματα των τιμών των $X_{10\text{miss}}$, $X_{10\text{obs}}$ και $X_{11\text{miss}}$, $X_{11\text{obs}}$. 123
5.6.1 Ραβδογράμματα των δεικτών $\boldsymbol{\gamma}$ στις μεθόδους SSVS, Kuo &Mallick και GVS σε δεδομένα με μηχανισμούς έλλειψης στις επεξηγηματικές μεταβλητές $\boldsymbol{X}$. 132

Κατάλογος Πινάκων

1.4.1.1 Χαρακτηριστικά παραδείγματα συζυγών εκ των προτέρων κατανομών. 5
2.2.1 Ερμηνεία του παράγοντα Bayes σύμφωνα με τους Kass and Raftery (1995). 16
4.2.1 Παράδειγμα μίας μονοτονικής έλλειψης και μίας μη-μονοτονικής έλλειψης στα δεδομένα. Κάθε στήλη αντιστοιχεί σε μία μεταβλητή και κάθε γραμμή αντιστοιχεί σε μία παρατήρηση. 42
5.2.1 Πίνακας συνδιακύμανσης της κανονικής κατανομής από όπου έχουμε προσομοιώσει τις τιμές των επεξηγηματικών μεταβλητών. 50
5.2.1.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ της μεθόδου SSVS στα πλήρη δεδομένα. 55
5.2.1.2 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου SSVS στα πλήρη δεδομένα. 57
5.2.2.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ του δειγματολήπτη Kuo &Mallick στα πλήρη δεδομένα. 60
5.2.2.2 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων του δειγματολήπτη Kuo &Mallick στα πλήρη δεδομένα. 60
5.2.3.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των συντελεστών του πλήρες μοντέλου. 62
5.2.3.2 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ της μεθόδου GVS στα πλήρη δεδομένα. 64
5.2.3.3 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου GVS στα πλήρη δεδομένα. 65
5.2.3.4 Τα 10 πρώτα μοντέλα με τις μικρότερες τιμές του κριτηρίου BIC στα πλήρη δεδομένα. 67
5.3.1.1.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MCAR). 72
5.3.1.1.2 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MCAR). 73
5.3.1.2.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ του δειγματολήπτη Kuo &Mallick σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MCAR). 74
5.3.1.2.2 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων του δειγματολήπτη Kuo &Mallick σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MCAR). 74
5.3.1.3.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των συντελεστών του πλήρες μοντέλου. 75

5.3.1.3.2 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MCAR). 76
5.3.1.3.3 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MCAR). 76
5.3.1.4.1 Τα 10 πρώτα μοντέλα με τις μικρότερες τιμές του κριτηρίου BIC σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MCAR). 77
5.3.2.1.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MAR). 81
5.3.2.1.2 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MAR). 81
5.3.2.2.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ του δειγματολήπτη Kuo &Mallick σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MAR). 82
5.3.2.2.2 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων του δειγματολήπτη Kuo &Mallick σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MAR). 83
5.3.2.3.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των συντελεστών του πλήρες μοντέλου. 83
5.3.2.3.2 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MAR). 84
5.3.2.3.3 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MAR). 85
5.3.2.4.1 Τα 10 πρώτα μοντέλα με τις μικρότερες τιμές του κριτηρίου BIC σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MAR). . 85
5.3.3.1.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (NMAR). 90
5.3.3.1.2 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (NMAR). 91
5.3.3.2.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ του δειγματολήπτη Kuo &Mallick σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (NMAR). 92
5.3.3.2.2 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων του δειγματολήπτη Kuo &Mallick σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (NMAR). 92

5.3.3.3.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των συντελεστών του πλήρες μοντέλου. 93
5.3.2.3.2 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (NMAR). 94
5.3.2.3.3 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (NMAR). 94
5.3.3.4.1 Τα 10 πρώτα μοντέλα με τις μικρότερες τιμές του κριτηρίου BIC σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (NMAR). 95
5.5.1.1.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MCAR). 104
5.5.1.1.2 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MCAR). 105
5.5.1.2.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ του δειγματολήπτη Kuo &Mallick σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MCAR). 106
5.5.1.2.2 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων του δειγματολήπτη Kuo &Mallick σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MCAR). . . . 106
5.5.1.3.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των συντελεστών του πλήρες μοντέλου. 107
5.5.1.3.2 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MCAR). 108
5.5.1.3.3 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MCAR). 108
5.5.1.4.1 Τα 10 πρώτα μοντέλα με τις μικρότερες τιμές του κριτηρίου BIC σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MCAR). 109
5.5.2.1.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MAR). 115
5.5.2.1.2 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MAR). 115
5.5.2.2.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ του δειγματολήπτη Kuo &Mallick σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MAR). 116

5.5.2.2.2 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων του δειγματολήπτη Kuo &Mallick σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MAR). 117
5.5.2.3.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των συντελεστών του πλήρες μοντέλου. 117
5.5.2.3.2 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MAR). 118
5.5.2.3.3 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MAR). 119
5.5.2.4.1 Τα 10 πρώτα μοντέλα με τις μικρότερες τιμές του κριτηρίου BIC σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MAR). 119
5.5.3.1.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (NMAR). 125
5.5.3.1.2 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (NMAR). 126
5.5.3.2.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ του δειγματολήπτη Kuo &Mallick σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (NMAR). 127
5.5.3.2.2 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων του δειγματολήπτη Kuo &Mallick σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (NMAR). . . . 127
5.5.3.3.1 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των συντελεστών του πλήρες μοντέλου. 128
5.5.3.3.2 Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος γ της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (NMAR). 129
5.5.3.3.3 Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (NMAR). 129
5.5.3.4.1 Τα 10 πρώτα μοντέλα με τις μικρότερες τιμές του κριτηρίου BIC σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (NMAR). 130
6.1 Ικανότητα ανίχνευσης του πραγματικού μοντέλου με χρήση των μεθόδων SSVS, Kuo &Mallick και GVS. 135

Συντομογραφίες & Ακρωνύμια

MCMC	Markov Chain Monte Carlo
KL	Kullback-Leibler
I.I.D	Independent and identically distributed
SCMH	Single component Metropolis-Hastings
MAP model	Maximum a posteriori probability model
PO	Posterior odds ratio
BF	Bayes factor
BMA	Bayesian model Averaging
BIC	Bayesian information criterion
MLE	Maximum likelihood estimation
RIS	Reverse importance sampling
SSVS	Stochastic Search Variable Selection
GVS	Gibbs Variable Selection
MCAR	Missing Completely at Random
MAR	Missing at Random
NMAR	Not Missing at Random
MICE	Multiple imputation through chained equations
PMM	Predictive mean matching

1 Εισαγωγή στην Μπεϋζιανή Στατιστική και MCMC

1.1 Εισαγωγή

Στην επιστήμη της Στατιστικής υπάρχουν δύο κύριες σχολές, η Κλασική (Frequentist Statistics) και η Μπεϋζιανή στατιστική (Bayesian Statistics) όπου με την ανάπτυξη της υπολογιστικής ικανότητας των υπολογιστών έχει γίνει όλο και πιο δημοφιλής. Η Κλασική στατιστική αντιμετωπίζει την πιθανότητα ως τη συχνότητα εμφάνισης ενός γεγονότος μετά από ένα μεγάλο αριθμό δοκιμών, ενώ υπό την Μπεϋζιανή προσέγγιση η πιθανότητα αντιμετωπίζεται ως ένα μέτρο αβεβαιότητας ή πίστης σε ένα γεγονός. Συνεπώς σε αυτό το κεφάλαιο θα διερευνήσουμε τις διαφορές μεταξύ αυτών των δύο προσεγγίσεων και θα μιλήσουμε για το θεώρημα του Bayes όπου είναι ο πυλώνας της Μπεϋζιανής σκέψης. Επιπλέον, θα εξετάσουμε τη σημασία της επιλογής κατάλληλων εκ των προτέρων κατανομών, καθώς αυτές οι κατανομές μπορούν να επηρεάσουν σε μεγάλο βαθμό τα αποτελέσματα της Μπεϋζιανής ανάλυσης. Τέλος, θα αναφέρουμε τους αλγορίθμους Markov Chain Monte Carlo (MCMC), οι οποίοι αποτελούν ισχυρά εργαλεία στην περίπτωση που είναι αδύνατη η εύρεση της εκ των υστέρων κατανομής σε κλειστή μορφή. Συνολικά λοιπόν, αυτό το κεφάλαιο θα παρέχει μια περιεκτική επισκόπηση των βασικών στοιχείων της Μπεϋζιανής στατιστικής.

1.2 Σύγκριση της Κλασικής με την Μπεϋζιανή Στατιστική

Η Μπεϋζιανή στατιστική και η κλασική στατιστική είναι δύο διαφορετικές προσεγγίσεις στατιστικής συμπερασματολογίας. Ενώ η κλασική στατιστική βασίζεται στην ιδέα της εμπειρικής πιθανότητας, η Μπεϋζιανή στατιστική βασίζεται στην έννοια της υποκειμενικής πιθανότητας.

- Εμπειρική πιθανότητα: Είναι η πιθανότητα που βασίζεται στην παρατηρούμενη συχνότητα γεγονότων σε μεγάλο αριθμό δοκιμών. Δηλαδή η εμπειρική πιθανότητα ή αλλιώς αντικειμενική πιθανότητα, μας λέει ότι η πιθανότητα ενός συμβάντος μπορεί να προσδιοριστεί με τη διεξαγωγή επαναλαμβανόμενων δοκιμών ενός πειράματος και την παρατήρηση του ποσοστού των φορών που συμβαίνει ή γίνεται το συμβάν. Για παράδειγμα αν στρίψουμε ένα δίκαιο νόμισμα πολλές φορές η πιθανότητα να πάρουμε κεφαλή θα είναι ίση με 0.5, καθώς αυτό το συμβάν θα πραγματοποιηθεί περίπου τις μισές φορές.

- Υποκειμενική πιθανότητα (Μπεϋζιανή πιθανότητα): Είναι η πιθανότητα που λαμβάνει υπόψιν τις πεποιθήσεις ενός ατόμου ή το βαθμό αβεβαιότητας για ένα γεγονός. Με λίγα λόγια μας αντικατοπτρίζει την υποκειμενική εκτίμηση ενός ατόμου (ή ατόμων) για την πιθανότητα του γεγονότος, έχοντας ως βάση τη δική του γνώση και εμπειρία.

Σε αντίθεση με την κλασική στατιστική που αντιμετωπίζει την παράμετρο $\boldsymbol{\theta}$ ως άγνωστη αλλά με σταθερή ποσότητα, στην Μπεϋζιανή στατιστική η παράμετρος $\boldsymbol{\theta}$ αντιμετωπίζεται ως τυχαία μεταβλητή. Η αντιμετώπιση της παραμέτρου $\boldsymbol{\theta}$ ως τυχαία μεταβλητή μας επιτρέπει να ενσωματώσουμε πρότερες πεποιθήσεις που έχουμε για αυτή

και μας δίνει τη δυνατότητα να ενημερώνουμε τις πεποιθήσεις αυτές καθώς γίνονται διαθέσιμα νέα δεδομένα, οδηγώντας μας έτσι σε πιο ακριβή και αξιόπιστα συμπεράσματα.

Η Μπεϋζιανή στατιστική και η κλασική στατιστική είναι δύο διαφορετικές προσεγγίσεις όπου έχουν τα θετικά και τα αρνητικά τους.

Θετικά στοιχεία της κλασικής στατιστικής:

1. Αποτελείται από απλούστερους υπολογισμούς σε σχέση με τη στατιστική κατά Bayes, κάνοντας έτσι την εφαρμογή της πιο εύκολη.

2. Μπορεί να μας φανεί χρήσιμη σε προβλήματα όπου έχουμε μεγάλο όγκο δεδομένων.

Αρνητικά στοιχεία της κλασικής στατιστικής:

1. Βασίζεται κατά μεγάλο βαθμό στις υποθέσεις που κάνουμε σχετικά με την κατανομή των δεδομένων, η οποία μπορεί να μην είναι πάντα η κατάλληλη.

2. Πολλές φορές χρειάζεται έναν μεγάλο όγκο δεδομένων για να μπορέσει να μας δώσει αξιόπιστα αποτελέσματα, πράγμα το οποίο στις περισσότερες περιπτώσεις δεν είναι εφικτό διότι αρκετές φορές ο αριθμός των παρατηρήσεων είναι περιορισμένος.

3. Δε χρησιμοποιεί κάποια εκ των προτέρων γνώση ή πεποίθηση για τις παραμέτρους στην ανάλυση, πράγμα το οποίο μας περιορίζει.

Θετικά στοιχεία της Μπεϋζιανή στατιστικής:

1. Μπορούμε να χρησιμοποιήσουμε εκ των προτέρων γνώσεις που έχουμε σχετικά με κάποια παράμετρο στην ανάλυσή μας.

2. Μπορεί να χρησιμοποιηθεί για να γίνουν πιο ακριβείς προβλέψεις ενσωματώνοντας στο μοντέλο μας νέα δεδομένα καθώς αυτά γίνονται διαθέσιμα (sequential updating).

3. Επιτρέπει τη σύγκριση πολλαπλών μοντέλων, καθιστώντας ευκολότερο τον εντοπισμό του καλύτερου μοντέλου για ένα σύνολο δεδομένων.

4. Παρέχει εκτιμήσεις της αβεβαιότητας με τη μορφή διαστημάτων αξιοπιστίας, τα οποία είναι χρήσιμα για τη λήψη αποφάσεων.

Αρνητικά στοιχεία της Μπεϋζιανή στατιστικής:

1. Έχει υπολογιστικά μεγάλο κόστος, ειδικά όταν έχουμε να κάνουμε με πολύπλοκα μοντέλα.

2. Η τελική συμπερασματολογία μας επηρεάζεται σε μεγάλο βαθμό από τις εκ των προτέρων γνώσεις ή πεποιθήσεις, προκαλώντας έτσι πολλές φορές μεροληψία στην ανάλυσή μας.

Συνεπώς συμπεραίνουμε ότι η Μπεϋζιανή και η κλασική στατιστική αποτελούν δύο διαφορετικές προσεγγίσεις, όπου η κάθε μία έχει τα πλεονεκτήματα της και τις αδυναμίες της. Ωστόσο, η επιλογή για το ποια προσέγγιση θα χρησιμοποιηθεί τελικά, εξαρτάται κατά κόρον από τις προτιμήσεις και υποθέσεις του ίδιου του ερευνητή. Είναι σημαντικό λοιπόν να κατανοηθούν οι θεμελιώδεις διαφορές μεταξύ αυτών των δύο προσεγγίσεων προκειμένου να ληφθεί μια τεκμηριωμένη απόφαση σχετικά με το ποια είναι καταλληλότερη προσέγγιση για εμάς. Ως εκ τούτου, και οι δύο μέθοδοι έχουν τη θέση τους στη σύγχρονη στατιστική και συνεχίζουν να ερευνώνται ενεργά και να εφαρμόζονται σε ποικίλα πεδία.

1.3 Θεώρημα του Bayes

Ο βασικός πυρήνας της Μπεϋζιανής στατιστικής είναι το θεώρημα του Bayes. Αποτελεί ένα απλό, αλλά θεμελιώδες αποτέλεσμα της θεωρίας πιθανοτήτων, το οποίο μας βοηθάει στο να βρίσκουμε δεσμευμένες πιθανότητες. Το θεώρημα αυτό προτάθηκε από τον αιδεσιμότατο Thomas Bayes, έναν μαθηματικό του 18ου αιώνα.

Το θεώρημα Bayes δίνεται από τον τύπο

$$\mathbb{P}(A|D) = \frac{\mathbb{P}(D|A) \cdot \mathbb{P}(A)}{\mathbb{P}(D)},$$

ενώ όταν μιλάμε για συναρτήσεις πυκνότητας πιθανότητας έχουμε

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{\int f(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}}, \tag{1.3.1}$$

όπου $p(\boldsymbol{\theta}|\boldsymbol{x})$ είναι η εκ των υστέρων συνάρτηση πυκνότητας πιθανότητας, $f(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta})$ είναι η πιθανοφάνεια (likelihood), $p(\boldsymbol{\theta})$ είναι η εκ των προτέρων συνάρτηση πυκνότητας πιθανότητας και $f(\boldsymbol{x}) = \int f(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ είναι η περιθώρια πιθανοφάνεια των δεδομένων (σταθερά κανονικοποίησης). Τέλος το $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^k$ αντιστοιχεί στις άγνωστες παραμέτρους, ενώ το $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ αντιστοιχεί στα δεδομένα που κατέχουμε.

Για να βρούμε την εκ των υστέρων συνάρτηση πυκνότητας πιθανότητας με τη βοήθεια του θεωρήματος Bayes, ακολουθούμε τα εξής βήματα:

1. Προσδιορίζουμε την εκ των προτέρων κατανομή του $\boldsymbol{\theta}$. Αυτή αντιπροσωπεύει τις πληροφορίες ή τις πεποιθήσεις που έχουμε για την παράμετρο $\boldsymbol{\theta}$ πριν ασχοληθούμε με τα δεδομένα.

2. Προσδιορίζουμε τη συνάρτηση πιθανοφάνειας, όπου περιγράφει την πυκνότητα παρατήρησης των δεδομένων δοθέντος την παράμετρο $\boldsymbol{\theta}$.

3. Υπολογίζουμε την περιθώρια πιθανοφάνεια ολοκληρώνοντας το γινόμενο της εκ των προτέρων κατανομής του $\boldsymbol{\theta}$ με την πιθανοφάνεια ως προς την παράμετρο $\boldsymbol{\theta}$.

4. Χρησιμοποιούμε το θεώρημα του Bayes και υπολογίζουμε την εκ των υστέρων συνάρτηση του $\boldsymbol{\theta}$.

5. Εξάγουμε συμπεράσματα για την παράμετρο $\boldsymbol{\theta}$ με τη βοήθεια της εκ των υστέρων κατανομής και κάνουμε εκτιμήσεις.

Από την (1.3.1) παρατηρούμε ότι η περιθώρια πιθανοφάνεια του $\boldsymbol{x}$ δεν είναι τίποτα άλλο παρά μία σταθερά στην $p(\boldsymbol{\theta}|\boldsymbol{x})$, καθώς η ολοκλήρωση γίνεται ως προς την παράμετρο $\boldsymbol{\theta}$. Επομένως μπορούμε να πούμε ότι η εκ των υστέρων κατανομή είναι ανάλογη του γινομένου της πιθανοφάνειας με την εκ των προτέρων κατανομή του $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto f(\boldsymbol{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}). \tag{1.3.2}$$

1.4 Επιλογή της εκ των προτέρων κατανομής

Στην Μπεϋζιανή στατιστική, η επιλογή της εκ των προτέρων κατανομής είναι πολύ σημαντική, καθώς παίζει καθοριστικό ρόλο στην κατασκευή της εκ των υστέρων κατανομής. Η επιλογή αυτή καθορίζεται από τυχόν πληροφορίες ή την έλλειψη αυτών που κατέχουμε για την άγνωστη παράμετρο $\boldsymbol{\theta}$ πριν ασχοληθούμε με τα δεδομένα.

Γενικά οι εκ των προτέρων κατανομές χωρίζονται σε δύο κατηγορίες οι οποίες είναι οι εξής:

- Πληροφοριακές εκ των προτέρων κατανομές
- Μη πληροφοριακές εκ των προτέρων κατανομές

όπου οι πληροφοριακές εκ των προτέρων κατανομές εκφράζουν τις πληροφορίες ή πεποιθήσεις που έχουμε σχετικά με την άγνωστη παράμετρο (παραδείγματα τέτοιων πληροφοριών περιλαμβάνουν γνώμες ειδικών, προηγούμενες μελέτες ή παλιά δεδομένα). Οι μη πληροφοριακές εκ των προτέρων κατανομές εκφράζουν την απώλεια των πρότερων γνώσεων που έχουμε για την άγνωστη παράμετρο.

Κάποιες υποκατηγορίες εκ των προτέρων κατανομών είναι:

- Συζυγείς εκ των προτέρων κατανομές (conjugate priors)
- Ιεραρχικές εκ των προτέρων κατανομές (Hyper priors).
- Επίπεδες εκ των προτέρων κατανομές (Flat priors)
- Ασαφής εκ των προτέρων κατανομές (vague priors)
- Η εκ των προτέρων κατανομή του Jeffreys (Jeffreys prior)
- Εκ των προτέρων κατανομές βασισμένες σε δυνάμεις της πιθανοφάνειας (Power priors)
- Εκ των προτέρων κατανομές αναφοράς (Reference priors)

Σε αυτήν τη διπλωματική εργασία θα δοθεί μια αρκετά συνοπτική επεξήγηση των παραπάνω εκ των προτέρων κατανομών, έτσι ώστε στα επόμενα κεφάλαια να έχουμε τις κατάλληλες γνώσεις όσο αναφορά τα βασικά στοιχεία της Μπεϋζιανής στατιστικής.

1.4.1 Συζυγείς εκ των προτέρων κατανομές

Μία εκ των προτέρων κατανομή ονομάζεται συζυγής εάν ανήκει στην ίδια οικογένεια κατανομών με την εκ των υστέρων κατανομή, καθιστώντας έτσι ευκολότερους τους υπολογισμούς εύρεσης της εκ των υστέρων κατανομής. Είναι απλές και άμεσες στην κατασκευή τους έχοντας όμως ως προϋπόθεση ότι το δείγμα μας αποτελείται από ανεξάρτητες και ισόνομες τυχαίες μεταβλητές, προερχόμενες από κατανομή που ανήκει στην εκθετική οικογένεια κατανομών.

Έστω $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ παρατηρήσεις προερχόμενες από ανεξάρτητες και ισόνομα κατανεμημένες τυχαίες μεταβλητές. Στον Πίνακα 1.4.1.1 δίνονται κάποια συνήθη παραδείγματα συζυγών εκ των προτέρων κατανομών.

$f(x_i\|\boldsymbol{\theta})$	Εκ των προτέρων $p(\boldsymbol{\theta})$	Εκ των υστέρων $p(\boldsymbol{\theta}\|\boldsymbol{x})$
$\text{Bin}(n, \boldsymbol{\theta})$	$\text{Beta}(\alpha, \beta)$	$\text{Beta}(\alpha + \boldsymbol{x}, n - \boldsymbol{x} + \beta)$
$\text{Poisson}(\boldsymbol{\theta})$	$\text{Gamma}(\alpha, \beta)$	$\text{Gamma}(\alpha + \sum_{i=1}^n x_i, \beta + n)$
$\text{Exp}(\boldsymbol{\theta})$	$\text{Gamma}(\alpha, \beta)$	$\text{Gamma}(\alpha + n, \beta + \sum_{i=1}^n x_i)$
$N(\boldsymbol{\theta}, \sigma^2)$	$N(\mu_0, \sigma_0^2)$	$N\left(s^2 \cdot \left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_i^n x_i}{\sigma^2}\right), s^2 = \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right)$

Πίνακας 1.4.1.1: Χαρακτηριστικά παραδείγματα συζυγών εκ των προτέρων κατανομών.

1.4.2 Επίπεδες εκ των προτέρων κατανομές

Οι επίπεδες εκ των προτέρων κατανομές αποτελούν μία υποκατηγορία των μη πληροφοριακών εκ των προτέρων κατανομών, καθώς δίνουν την ίδια βαρύτητα σε όλες τις τιμές τις παραμέτρου $\boldsymbol{\theta}$. Στην περίπτωση που ξέρουμε ότι το $\boldsymbol{\theta}$ ανήκει σε ένα κλειστό διάστημα (πχ. $\boldsymbol{\theta} \in [a, b]$), μπορούμε να χρησιμοποιήσουμε μια ομοιόμορφη κατανομή, κάτι το οποίο εκφράζει την πλήρη άγνοια που έχουμε για την παράμετρο που εκτιμάται. Στην περίπτωση όμως που η παράμετρος $\boldsymbol{\theta}$ δεν ανήκει σε ένα κλειστό διάστημα η χρήση της ομοιόμορφης κατανομής, ως πρότερη είναι λανθασμένη. Σε αυτή την περίπτωση η επίπεδη εκ των προτέρων ($p(\boldsymbol{\theta}) \propto c$) στο $\mathbb{R}$, δε θα αντιστοιχεί σε κατανομή. Τέτοιες κατανομές ονομάζονται μη γνήσιες (improper).

Μία μη γνήσια εκ των προτέρων είναι μία συνάρτηση που δε συμπεριφέρεται όπως μια παραδοσιακή κατανομή πιθανότητας, επειδή το ολοκλήρωμα της στο χώρο των παραμέτρων είναι ίσο με το άπειρο (δηλαδή $\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \infty$). Παρόλα αυτά, η εκ των υστέρων κατανομή μπορεί ακόμα να είναι καλά ορισμένη, αν και μόνο αν, η περιθώρια πιθανοφάνεια $f(\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} f(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ είναι και αυτή καλά ορισμένη για όλα τα $\boldsymbol{x} \in \boldsymbol{X}$. Από την (1.3.1) έχουμε

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{\int f(\boldsymbol{x}|\boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} = \frac{f(\boldsymbol{x}|\boldsymbol{\theta}) \cdot c}{\int f(\boldsymbol{x}|\boldsymbol{\theta}) \cdot c\, d\boldsymbol{\theta}} = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})}{\int f(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{\theta}}.$$

Οι μη γνήσιες κατανομές χρησιμοποιούνται μερικές φορές για λόγους ευκολίας, ειδικά όταν το σύνολο των δεδομένων είναι μεγάλο και η επιρροή που θα έχει η εκ των προτέρων κατανομή θα είναι πολύ μικρή. Ωστόσο θα πρέπει να είμαστε ιδιαίτερα προσεκτικοί στην ερμηνεία των εκ των υστέρων κατανομών που προέρχονται από μη γνήσιες εκ των προτέρων κατανομές, καθώς υπάρχει περίπτωση να οδηγηθούμε σε μη γνήσια εκ των υστέρων κατανομή (improper posterior) αν η πιθανοφάνεια είναι επίπεδη (πολύ σπάνιο ενδεχόμενο).

Όταν ασχολούμαστε με συνεχείς παραμέτρους, υπάρχει μία τάση να υποθέτουμε ομοιόμορφη εκ των προτέρων κατανομή σε ένα εύρος τιμών. Όμως, αυτή η προσέγγιση προκαλεί προβλήματα, διότι μία ομοιόμορφη κατανομή για την παράμετρο $\boldsymbol{\theta}$ δεν συνεπάγεται απαραίτητα ότι θα έχουμε ομοιόμορφη κατανομή σε κάποιο 1-1 μετασχηματισμό $\boldsymbol{\psi} = h(\boldsymbol{\theta})$. Για να ξεπεραστεί αυτό το ζήτημα, ο Jeffreys (1935) πρότεινε μία εκ των προτέρων κατανομή, η οποία είναι αμετάβλητη σε τέτοιους μετασχηματισμούς. Πρότεινε λοιπόν, μία μη πληροφοριακή πρότερη κατανομή η οποία είναι βασισμένη στον πίνακα πληροφορίας του Fisher $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$

$$p(\boldsymbol{\theta}) \propto \sqrt{\det \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})},$$

όπου το ij στοιχείο του πίνακα πληροφορίας του Fisher δίνεται από τον τύπο

$$I_{ij}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{x}|\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\theta_i\theta_j}\log(f(\boldsymbol{x}|\boldsymbol{\theta}))\right].$$

1.4.3 Εκ των προτέρων κατανομές αναφοράς

Ο Bernardo (1979) εισήγαγε την έννοια των πρότερων κατανομών αναφοράς, οι οποίες αναπτύχθηκαν περαιτέρω από τους Berger et al. (2009). Οι εκ των προτέρων κατανομές αναφοράς, επιλέγονται με τέτοιο τρόπο ώστε να μεγιστοποιείται η ασυμφωνία της $p(\boldsymbol{\theta})$ με τη $p(\boldsymbol{\theta}|\boldsymbol{x})$. Για να ποσοτικοποιήσουμε αυτή την ασυμφωνία θα χρησιμοποιήσουμε την απόκλιση Kullback-Leibler. Έχουμε

$$KL(p(\boldsymbol{\theta}|\boldsymbol{x}), p(\boldsymbol{\theta})) = \int p(\boldsymbol{\theta}|\boldsymbol{x})\log\left(\frac{p(\boldsymbol{\theta}|\boldsymbol{x})}{p(\boldsymbol{\theta})}\right)d\boldsymbol{\theta},$$

όπου αν $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{x})$ τότε $KL = 0$. Παρατηρούμε όμως ότι στο ολοκλήρωμα υπάρχει και η εκ των υστέρων κατανομή $p(\boldsymbol{\theta}|\boldsymbol{x})$, όπου για να βρεθεί χρειάζεται να χρησιμοποιήσουμε τα δεδομένα μας, πράγμα το οποίο έρχεται σε αντίθεση με τον τρόπο που βρίσκουμε τις εκ των προτέρων κατανομές. Οπότε θα πάρουμε τη μέση απόκλιση Kullback-Leibler:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[KL] &= \int f(\boldsymbol{x})KL(p(\boldsymbol{\theta}|\boldsymbol{x}), p(\boldsymbol{\theta}))\,d\boldsymbol{x} \\ &= \int\int p(\boldsymbol{\theta}|\boldsymbol{x})f(\boldsymbol{x})\log\left(\frac{p(\boldsymbol{\theta}|\boldsymbol{x})}{p(\boldsymbol{\theta})}\right)d\boldsymbol{\theta}\,d\boldsymbol{x}.\end{aligned}$$

Ξέρουμε όμως ότι $p(\boldsymbol{\theta}, \boldsymbol{x}) = p(\boldsymbol{\theta}|\boldsymbol{x})f(\boldsymbol{x})$, οπότε

$$\mathbb{E}_{\boldsymbol{x}}[KL] = \int\int p(\boldsymbol{\theta}, \boldsymbol{x})\log\left(\frac{p(\boldsymbol{\theta}, \boldsymbol{x})}{p(\boldsymbol{\theta})f(\boldsymbol{x})}\right)d\boldsymbol{\theta}\,d\boldsymbol{x}$$

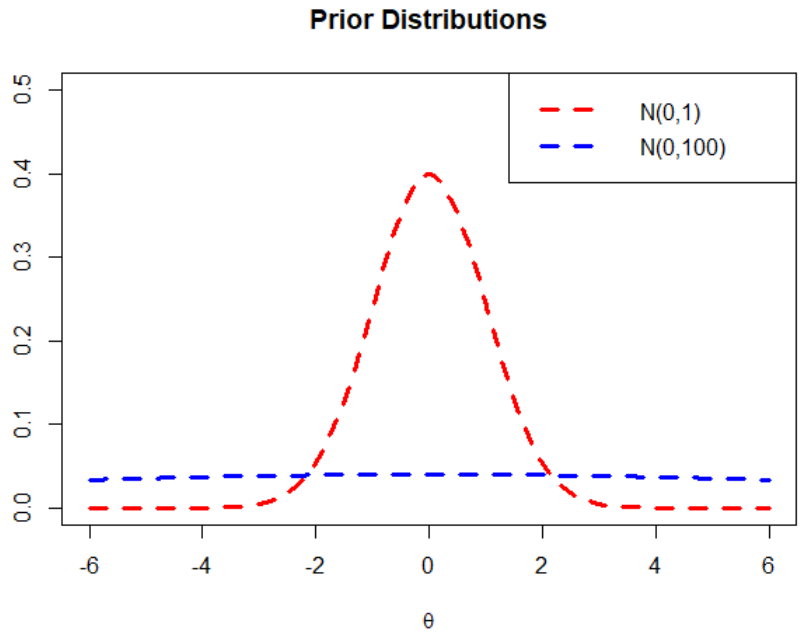
όπου το διπλό ολοκλήρωμα ονομάζεται αμοιβαία πληροφορία (mutual information).

Ψάχνουμε λοιπόν την εκ των προτέρων κατανομή $p(\boldsymbol{\theta})$ που θα μεγιστοποιεί την αμοιβαία πληροφορία μεταξύ του $\boldsymbol{\theta}$ και του $\boldsymbol{x}$, δίνοντας έτσι στα δεδομένα μας τη δυνατότητα να μπορούν να επηρεάσουν την εκ των υστέρων κατανομή.

1.4.4 Ασαφής εκ των προτέρων κατανομές

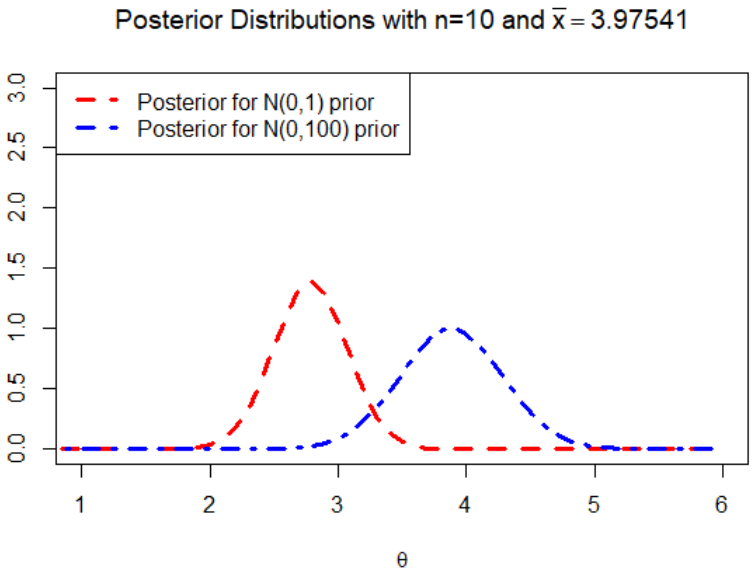
Όπως έχουμε ήδη αναφέρει και στην αρχή της παραγράφου (1.3), η χρήση των μη πληροφοριακών εκ των προτέρων κατανομών, γίνεται όταν υπάρχει έλλειψη γνώσης ή πληροφορίας σχετικά με την παράμετρο ενδιαφέροντος. Με αυτό τον τρόπο η εκ των υστέρων κατανομή που προέρχεται από το θεώρημα του Bayes θα καθορίζεται από τα δεδομένα μας. Ένας ακόμα τρόπος για να επιτευχθεί αυτό, είναι να χρησιμοποιήσουμε πρότερες κατανομές οι οποίες έχουν τεράστια διακύμανση (vague priors). Τέτοιες κατανομές μπορεί να είναι η Κανονική κατανομή με τεράστια διακύμανση (π.χ $N(\mu = 0, \sigma^2 = 10000)$) είτε μια κατανομή Γάμμα με μικρές τιμές στις παραμέτρους α και β (π.χ Gamma$(\alpha = 0.0001, \beta = 0.0001)$). Με τον τρόπο αυτό δίνουμε μία αρκετά μικρή αλλά όχι μηδενική πληροφορία σχετικά με την πραγματική τιμή της παραμέτρου $\boldsymbol{\theta}$. Έτσι η χρήση ασαφών πρότερων κατανομών μπορεί να θεωρηθεί ως ένας τρόπος να αφήσουμε τα δεδομένα να μιλήσουν από μόνα τους, καθώς η εκ των υστέρων κατανομή θα επηρεαστεί σε πολύ μεγάλο βαθμό από τη συνάρτηση πιθανοφάνειας.

Για παράδειγμα στο Σχήμα 1.4.4.1 απεικονίζεται μία κανονική Κατανομή με μικρή διασπορά (κόκκινη διακεκομμένη γραμμή) και μία Κανονική κατανομή με σχετικά μεγάλη διασπορά (μπλε διακεκομμένη γραμμή).



Σχήμα 1.4.4.1: Απεικόνιση των Κανονικών κατανομών $N(0, 1)$ και $N(0, 100)$

Στο Σχήμα 1.4.4.2 έχουμε υποθέσει ότι έχουμε μέγεθος δείγματος $n = 10$ και $\bar{x} = 3.97541$.



Σχήμα 1.4.4.2: Απεικόνιση των εκ των υστέρων κατανομών με πρότερες κατανομές την $N(0,1)$ και την $N(0,100)$.

Παρατηρούμε ότι η εκ των υστέρων κατανομή (όπου και αυτή είναι Κανονική) που αντιστοιχεί στην πρότερη κατανομή $N(0,100)$ επηρεάζεται σε μεγάλο βαθμό από το δείγμα, καθώς φαίνεται ότι η μέση τιμή μετακινήθηκε πολύ κοντά στην τιμή που ευνοεί η πιθανοφάνεια. Ενώ για την ύστερη κατανομή που αντιστοιχεί στην εκ των προτέρων κατανομή $N(0,1)$, βλέπουμε ότι χρειαζόμαστε παραπάνω δείγμα για να μπορέσουμε να εξαλείψουμε την εκ των προτέρων πληροφορία.

Οι όροι «ασαφή» και «μη πληροφοριακή» χρησιμοποιούνται συχνά εναλλακτικά, ωστόσο δεν έχουν την ίδια σημασία. Ενώ μια ασαφής εκ των προτέρων κατανομή είναι σχετικά μη πληροφοριακή και δεν ευνοεί έντονα ορισμένες τιμές της παραμέτρου $\boldsymbol{\theta}$, μπορεί όμως να μας δώσει κάποια ανεπιθύμητη πληροφορία για κάποιο μετασχηματισμό $h(\boldsymbol{\theta})$. Ένα συγκεκριμένο παράδειγμα που επεξηγεί την ανάγκη διάκρισης μεταξύ των δύο, είναι η περίπτωση που έχουμε αραιά δεδομένα (sparse data), όπου μια φαινομενικά «ασαφής» εκ των προτέρων κατανομή, μπορεί στην πραγματικότητα να επηρεάσει σημαντικά τυχόν συμπεράσματα της ανάλυσης (Lambert et al., 2005). Ως εκ τούτου, είναι απαραίτητο να γίνει διάκριση μεταξύ των όρων «ασαφή» και «μη πληροφοριακή», για να αποφευχθεί οποιαδήποτε ασάφεια στην Μπεϋζιανή συμπερασματολογία.

1.4.5 Ιεραρχικές εκ των προτέρων κατανομές

Μέχρι τώρα έχουμε υποθέσει ότι οι παράμετροι της εκ των προτέρων κατανομής της άγνωστης παραμέτρου $\boldsymbol{\theta}$ είναι γνωστές. Ωστόσο εάν δεν είναι γνωστές, τότε η μέχρι τώρα εμπειρία μας λέει ότι χρειάζεται να προσδιορίσουμε εκ των προτέρων κατανομή για τις παραμέτρους της πρότερης κατανομής του $\boldsymbol{\theta}$. Έστω ότι $\boldsymbol{\nu}$ είναι οι άγνωστοι παράμετροι της πρότερης κατανομής του $\boldsymbol{\theta}$, τότε $h(\boldsymbol{\nu})$ αντιστοιχεί στην εκ των προτέρων κατανομή του $\boldsymbol{\nu}$ η οποία συνήθως αναφέρεται ως υπέρ πρότερη (Hyperprior). Αυτή η κατανομή μας δίνει τη δυνατότητα να εξάγουμε συμπεράσματα σχετικά με το $\boldsymbol{\nu}$ με βάση

τα παρατηρούμενα δεδομένα. Η επιλογή της υπέρ πρότερης μπορεί να έχει σημαντικό αντίκτυπο στην εκ των υστέρων κατανομή και κατά συνέπεια στη συμπερασματολογία μας. Με τη βοήθεια του θεωρήματος Bayes (1.3.1) καταλήγουμε ότι η εκ των υστέρων κατανομή δίνεται από τον τύπο (Carlin and Louis, 1997):

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{x}) &= \frac{f(\boldsymbol{x},\boldsymbol{\theta})}{\int f(\boldsymbol{x},\boldsymbol{\theta})\,d\boldsymbol{\theta}} = \frac{\int f(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\nu})\,d\boldsymbol{\nu}}{\int\int f(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\nu})\,d\boldsymbol{\nu}\,d\boldsymbol{\theta}} \\ &= \frac{\int f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\nu})h(\boldsymbol{\nu})\,d\boldsymbol{\nu}}{\int\int f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\nu})h(\boldsymbol{\nu})\,d\boldsymbol{\nu}\,d\boldsymbol{\theta}}, \end{aligned}$$

ενώ η εκ των υστέρων κατανομή $p(\boldsymbol{\nu}|\boldsymbol{x})$ δίνεται από τον τύπο (Bernardo and Smith, 2009; Gelman et al., 1995)

$$\begin{aligned} p(\boldsymbol{\nu}|\boldsymbol{x}) &= \frac{f(\boldsymbol{x}|\boldsymbol{\nu})h(\boldsymbol{\nu})}{f(\boldsymbol{x})} = \frac{h(\boldsymbol{\nu})\left(\int f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\nu})\,d\boldsymbol{\theta}\right)}{f(\boldsymbol{x})} \\ &= \left(\frac{f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\nu})}{p(\boldsymbol{\theta}|\boldsymbol{x})}\right)\frac{h(\boldsymbol{\nu})}{f(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\nu})h(\boldsymbol{\nu})}{p(\boldsymbol{\theta}|\boldsymbol{\nu},\boldsymbol{x})f(\boldsymbol{x})} \\ &= \frac{f(\boldsymbol{x}|\boldsymbol{\nu},\boldsymbol{\theta})f(\boldsymbol{\theta},\boldsymbol{\nu})}{p(\boldsymbol{\theta}|\boldsymbol{\nu},\boldsymbol{x})f(\boldsymbol{x})} = \frac{p(\boldsymbol{\nu},\boldsymbol{\theta}|\boldsymbol{x})}{p(\boldsymbol{\theta}|\boldsymbol{\nu},\boldsymbol{x})}, \end{aligned}$$

όπου αυτές οι εκ των υστέρων κατανομές συνδέονται μεταξύ τους με τη σχέση (Bernardo and Smith, 2009)

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \int p(\boldsymbol{\theta},\boldsymbol{\nu}|\boldsymbol{x})\,d\boldsymbol{\nu} = \int p(\boldsymbol{\theta}|\boldsymbol{\nu},\boldsymbol{x})p(\boldsymbol{\nu}|\boldsymbol{x})\,d\boldsymbol{\nu}.$$

Εναλλακτικά μπορούμε να προχωρήσουμε μεγιστοποιώντας την πιθανοφάνεια $f(\boldsymbol{x}|\boldsymbol{\nu})$ ως προς $\boldsymbol{\nu}$ καταλήγοντας στην εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\boldsymbol{\nu}}$. Έτσι η συμπερασματολογία μας θα βασίζεται πλέον στην εκ των υστέρων κατανομή $p(\boldsymbol{\theta}|\boldsymbol{x},\hat{\boldsymbol{\nu}})$. Αυτή η προσέγγιση ονομάζεται Εμπειρική Μπεϋζιανή ανάλυση διότι χρησιμοποιούμε τα δεδομένα μας για την κατασκευή της εκτιμήτριας της παραμέτρου $\boldsymbol{\nu}$. Κατ'επέκταση και η παράμετρος $\boldsymbol{\nu}$ θα μπορούσε και αυτή με τη σειρά της να εξαρτάται από μια άλλη άγνωστη σε εμάς παράμετρο $\boldsymbol{\phi}$, όπου σε αυτή ορίζεται μια εκ των προτέρων κατανομή $s(\boldsymbol{\phi})$. Έτσι, με αυτό τον τρόπο καταλήγουμε σε κάτι που καλείται ιεραρχική μοντελοποίηση.

1.5 Περιθώρια πιθανοφάνεια

Η περιθώρια πιθανοφάνεια (marginal likelihood) έχει έναν κρίσιμο ρόλο στην Μπεϋζιανή συμπερασματολογία, καθώς αποτελεί τη σταθερά κανονικοποίησης της εκ των υστέρων κατανομής και όπως θα δούμε και στα επόμενα κεφάλαια, μας δίνει τη δυνατότητα να κάνουμε σύγκριση μοντέλων. Όπως αναφέραμε και παραπάνω, η περιθώρια πιθανοφάνεια δίνεται από τον εξής τύπο

$$f(\boldsymbol{x}) = \int_{\boldsymbol{\Theta}} f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,d\boldsymbol{\theta}.$$

Όπως θα δούμε και στα επόμενα κεφάλαια, ο υπολογισμός της περιθώριας πιθανοφάνειας μπορεί να φανεί υπολογιστικά δύσκολος, ειδικά όταν ασχολούμαστε με μεγάλο αριθμό παραμέτρων. Ωστόσο, υπάρχουν διάφορες υπολογιστικές τεχνικές,

όπως η δειγματολήπτης σπουδαιότητας, υπολογιστικές μέθοδοι Markov chain Monte Carlo (MCMC) και η μέθοδος Laplace, που μπορούν να χρησιμοποιηθούν έχοντας ως στόχο τη λύση αυτού του προβλήματος (Ενότητα 2.3).

1.6 Αλγόριθμοι Markov chain Monte Carlo

Πολλές φορές στην Μπεϋζιανή στατιστική είναι αδύνατος ο υπολογισμός της εκ των υστέρων κατανομής σε κλειστή μορφή. Σε αυτή την περίπτωση μπορεί να γίνει χρήση υπολογιστικών μεθόδων Markov chain Monte Carlo (MCMC), έχοντας ως τελικό στόχο να προσομοιώσουμε ψευδοτυχαίες τιμές από την εκ των υστέρων κατανομή.

Οι αλγόριθμοι MCMC μας δίνουν τη δυνατότητα να δημιουργήσουμε προσεγγιστικά διάφορες τιμές από την εκ των υστέρων κατανομή. Στην ουσία, αυτό που κάνουμε είναι ότι κατασκευάζουμε μία αλυσίδα Markov όπου ως στάσιμη κατανομή (equilibrium distribution) θα έχουμε την εκ των υστέρων κατανομή. Με αυτό τον τρόπο, η αλυσίδα μετά από κάποιες επαναλήψεις θα συγκλίνει στην εκ των υστέρων κατανομή, οπότε οι ψευδοτυχαίες τιμές που παράγονται από την αλυσίδα, μπορούν να χρησιμοποιηθούν στη συμπερασματολογία μας.

Οι αλγόριθμοι MCMC λειτουργούν ως εξής:

Αλγόριθμοι MCMC (MCMC scheme)

1. Ορίζουμε κάποιες αρχικές τιμές (προτεινόμενες τιμές) για τις παραμέτρους μας $\boldsymbol{\theta}^{(0)}$
2. Για $n = 1, \ldots, N$ γεννάμε ψευδοτυχαίες τιμές από τον αλγόριθμο μας.
3. Αν και εφόσον η αλυσίδα καταφέρει να συγκλίνει στη στάσιμη κατανομή (δηλαδή την εκ των υστέρων κατανομή), τότε μπορούμε να σταματήσουμε αφαιρώντας τις πρώτες m τιμές, μέχρι να επιτευχθεί σύγκλιση και να βασίσουμε τη συμπερασματολογία μας στις υπόλοιπες $N - m$ τιμές.

Εφόσον οι αλγόριθμοι MCMC είναι βασισμένοι σε αλυσίδες Markov, περιμένουμε το προσομοιωμένο δείγμα να μην είναι ανεξάρτητο και ισόνομα κατανεμημένο (I.I.D). Για την επίλυση αυτού του προβλήματος μπορούμε να εφαρμόσουμε λέπτυνση (Thinning) στο προσομοιωμένο δείγμα, κρατώντας μία ψευδοτυχαία τιμή ανά z επαναλήψεις του αλγόριθμου, όπου z ονομάζεται παράμετρος λέπτυνσης. Η επιλογή της παραμέτρου z γίνεται με τη βοήθεια διαγραμμάτων αυτοσυσχέτισης των ψευδοτυχαίων τιμών που παράγονται από την αλυσίδα.

1.6.1 Αλγόριθμος Metropolis-Hastings

Ο αλγόριθμος Metropolis-Hastings αναπτύχθηκε ανεξάρτητα από τον φυσικό Νικόλαο Μητρόπουλο (Metropolis et al., 1953) και τον μαθηματικό Hastings (1970). Ο Νικόλαος Μητρόπουλος ανέπτυξε αρχικά τον αλγόριθμο για την προσομοίωση της συμπεριφοράς ατόμων σε αέριο, ενώ ο Edward Hastings έδωσε μια γενικότερη έκδοση του αλγορίθμου έχοντας ως στόχο τη δημιουργία δειγμάτων από οποιαδήποτε συνάρτηση κατανομής πιθανότητας.

Για να χρησιμοποιήσουμε τον αλγόριθμο, ξεκινάμε θέτοντας κάποιες αρχικές τιμές στις παραμέτρους $\boldsymbol{\theta}^{(0)}$. Στη συνέχεια, στο $J-$βήμα με τη βοήθεια της κατανομής εισήγησης $g\left(\boldsymbol{\theta}^{(*)}|\boldsymbol{\theta}^{(J)}, \boldsymbol{x}\right)$ (proposal distribution) δημιουργούμε υποψήφιες τιμές (canditate values) των παραμέτρων $\boldsymbol{\theta}^{(*)}$ και αποφασίζουμε αν θα τις αποδεχτούμε με τη βοήθεια της πιθανότητας αποδοχής α (acceptance probability):

$$\alpha = \min\left(1, \frac{p\left(\boldsymbol{\theta}^{(*)}|\boldsymbol{x}\right) g\left(\boldsymbol{\theta}^{(J)}|\boldsymbol{\theta}^{(*)}, \boldsymbol{x}\right)}{p\left(\boldsymbol{\theta}^{(J)}|\boldsymbol{x}\right) g\left(\boldsymbol{\theta}^{(*)}|\boldsymbol{\theta}^{(J)}, \boldsymbol{x}\right)}\right). \tag{1.6.1.1}$$

Αλγόριθμος Metropolis-Hastings

1. Ορίζουμε αρχικές τιμές (προτεινόμενες τιμές) για τις παραμέτρους μας $\boldsymbol{\theta}^{(0)}$

   Για $J = 1, \ldots, N$ έχουμε:

2. Προσομοιώνουμε υποψήφιες τιμές $\boldsymbol{\theta}^{(*)}$ από την κατανομή εισήγησης $g\left(\boldsymbol{\theta}^{(*)}|\boldsymbol{\theta}^{(J)}, \boldsymbol{x}\right)$ όπου J είναι η τρέχουσα επανάληψη του αλγορίθμου.

3. Υπολογίζουμε την πιθανότητα αποδοχής

$$\alpha = \min\left(1, \frac{p\left(\boldsymbol{\theta}^{(*)}|\boldsymbol{x}\right) g\left(\boldsymbol{\theta}^{(J)}|\boldsymbol{\theta}^{(*)}, \boldsymbol{x}\right)}{p\left(\boldsymbol{\theta}^{(J)}|\boldsymbol{x}\right) g\left(\boldsymbol{\theta}^{(*)}|\boldsymbol{\theta}^{(J)}, \boldsymbol{x}\right)}\right)$$

4. Θέτουμε $\boldsymbol{\theta}^{(J+1)} = \boldsymbol{\theta}^{(*)}$ με πιθανότητα α και $\boldsymbol{\theta}^{(J+1)} = \boldsymbol{\theta}^{(J)}$ με πιθανότητα $(1-\alpha)$

Παρατηρούμε ότι, αν η κατανομή εισήγησης είναι συμμετρική, δηλαδή

$$g\left(\boldsymbol{\theta}^{(*)}|\boldsymbol{\theta}^{(J)}, \boldsymbol{x}\right) = g\left(\boldsymbol{\theta}^{(J)}|\boldsymbol{\theta}^{(*)}, \boldsymbol{x}\right),$$

τότε ο τύπος της πιθανότητας αποδοχής α (1.6.1.1) απλοποιείται σε:

$$\alpha = \min\left(1, \frac{p\left(\boldsymbol{\theta}^{(*)}|\boldsymbol{x}\right)}{p\left(\boldsymbol{\theta}^{(J)}|\boldsymbol{x}\right)}\right).$$

Επομένως, η πιθανότητα αποδοχής α θα εξαρτάται μόνο από την πυκνότητα της εκ των υστέρων κατανομής ($p(\boldsymbol{\theta}|\boldsymbol{x})$) στα σημεία $\boldsymbol{\theta}^{(*)}$ και $\boldsymbol{\theta}^{(J)}$. Από το πηλίκο $\frac{p\left(\boldsymbol{\theta}^{(*)}|\boldsymbol{x}\right)}{p\left(\boldsymbol{\theta}^{(J)}|\boldsymbol{x}\right)}$ συμπεραίνουμε πως ο αλγόριθμος θα τείνει να επισκέπτεται σημεία με μεγαλύτερη ύστερη πυκνότητα.

Μία άλλη εκδοχή του αλγορίθμου έχουμε όταν η κατανομή εισήγησης δεν εξαρτάται από την τρέχουσα τιμή

$$g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(J)}, \boldsymbol{x}) = g(\boldsymbol{\theta}|\boldsymbol{x}),$$

τότε ο τύπος της πιθανότητας αποδοχής είναι

$$\alpha = \min\left(1, \frac{p\left(\boldsymbol{\theta}^{(*)}|\boldsymbol{x}\right)}{p\left(\boldsymbol{\theta}^{(J)}|\boldsymbol{x}\right)} \frac{g\left(\boldsymbol{\theta}^{(J)}|\boldsymbol{x}\right)}{g\left(\boldsymbol{\theta}^{(*)}|\boldsymbol{x}\right)}\right).$$

Παρατηρούμε ότι στην περίπτωση που η $\boldsymbol{\theta}$ είναι πολυδιάστατη, ο αλγόριθμος Metropolis-Hastings ενημερώνει ταυτόχρονα ολόκληρο το διάνυσμα των παραμέτρων $\boldsymbol{\theta}$ με ένα μόνο βήμα. Κατά συνέπεια, η κατανομή εισήγησης χρειάζεται να έχει τις ίδιες διαστάσεις με το $\boldsymbol{\theta}$ δυσκολεύοντας έτσι τον αλγόριθμο. Αντ'αυτού, μπορούμε να δημιουργούμε μεμονωμένα την κάθε συνιστώσα έχοντας ως βάση διαφορετικές κατανομές εισήγησης για κάθε συνιστώσα του $\boldsymbol{\theta}$.

Έστω ότι $\boldsymbol{\theta}^{(J)} = \left(\theta_1^{(J)}, \theta_2^{(J)}, \ldots, \theta_k^{(J)}\right)$ είναι το διάνυσμα των παραμέτρων στην επανάληψη J. Ορίζουμε το διάνυσμα

$$\boldsymbol{\theta}_{-i}^{(J)} = \left(\theta_1^{(J+1)}, \theta_2^{(J+1)}, \ldots, \theta_{i-1}^{(J+1)}, \theta_{i+1}^{(J)}, \ldots, \theta_k^{(J)}\right).$$

Συνεπώς, η υποψήφια τιμή $\theta_i^{(*)}$ θα προσομοιώνεται με τη βοήθεια της κατανομής εισήγησης $g_i\left(\theta_i^{(*)}|\theta_i^{(J)}, \boldsymbol{\theta}_{-i}^{(J)}, \boldsymbol{x}\right)$. Ο αλγόριθμος αυτός αποτελεί μία ειδική περίπτωση του Metropolis-Hastings και ονομάζεται Single component Metropolis-Hastings (SCMH).

Αλγόριθμος Single component Metropolis-Hastings (SCMH)

Έστω ότι $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ είναι ένα διάνυσμα διάστασης k.

1. Ορίζουμε αρχικές τιμές (προτεινόμενες τιμές) για τις παραμέτρους μας $\boldsymbol{\theta}^{(0)}$

   Για $J = 1, \ldots, N$ έχουμε:

2. Προσομοιώνουμε υποψήφια τιμή για την παράμετρο $\theta_i^{(*)}$ του διανύσματος $\boldsymbol{\theta}^{(J)}$ από την κατανομή εισήγησης $g\left(\theta_i^{(*)}|\theta_i^{(J)}, \boldsymbol{\theta}_{-i}^{(J)}, \boldsymbol{x}\right)$ όπου J είναι η τρέχουσα επανάληψη του αλγορίθμου.

3. Υπολογίζουμε την πιθανότητα αποδοχής

$$\alpha = \min\left(1, \frac{p\left(\theta_i^{(*)}|\boldsymbol{\theta}_{-i}^{(J)}, \boldsymbol{x}\right) g\left(\theta_i^{(J)}|\theta_i^{(*)}, \boldsymbol{\theta}_{-i}^{(J)}, \boldsymbol{x}\right)}{p\left(\theta_i^{(J)}|\boldsymbol{\theta}_{-i}^{(J)}, \boldsymbol{x}\right) g\left(\theta_i^{(*)}|\theta_i^{(*)}, \boldsymbol{\theta}_{-i}^{(J)}, \boldsymbol{x}\right)}\right)$$

4. Θέτουμε $\theta_i^{(J+1)} = \theta_i^{(*)}$ με πιθανότητα α και $\theta_i^{(J+1)} = \theta_i^{(J)}$ με πιθανότητα $(1-\alpha)$

5. Επαναλαμβάνουμε τα βήματα 2,3 και 4 k φορές.

1.6.2 Δειγματολήπτης Gibbs

Ο δειγματολήπτης Gibbs είναι ένας αλγόριθμος MCMC που χρησιμοποιείται πολύ συχνά. Η θεμελίωση του έγινε πρώτη φορά από τους αδελφούς Geman and Geman (1984) και πήρε το όνομα του από τον Josiah Willard Gibbs, έναν Αμερικανό φυσικό και μαθηματικό του 19ου αιώνα που συνέβαλε σημαντικά στην ανάπτυξη της στατιστικής μηχανικής.

Ο δειγματολήπτης Gibbs αποτελεί μία ειδική περίπτωση του SCMH όπου:

$$g\left(\theta_i^{(*)}|\theta_i^{(J)}, \boldsymbol{\theta}_{-i}^{(J)}, \boldsymbol{x}\right) = p\left(\theta_i^{(*)}|\boldsymbol{\theta}_{-i}^{(J)}, \boldsymbol{x}\right)$$

και

$$g\left(\theta_i^{(J)}|\theta_i^{(*)}, \boldsymbol{\theta}_{-i}^{(J)}, \boldsymbol{x}\right) = p\left(\theta_i^{(J)}|\boldsymbol{\theta}_{-i}^{(*)}, \boldsymbol{x}\right).$$

Τότε παρατηρούμε ότι διαλέγοντας αυτή την κατανομή εισήγησης, στην ουσία δίνουμε στην πιθανότητα αποδοχής α την τιμή ένα και έτσι ο αλγόριθμος πάντα δέχεται την υποψήφια τιμή. Η $p\left(\theta_i^{(*)}|\boldsymbol{\theta}_{-i}^{(J)}, \boldsymbol{x}\right)$ καλείται πλήρους δέσμευσης εκ των υστέρων κατανομή (full conditional posterior distribution) για την παράμετρο θ_i.

Δειγματολήπτης Gibbs

Έστω ότι $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ είναι ένα διάνυσμα παραμέτρων διάστασης k.

1. Ορίζουμε αρχικές τιμές (προτεινόμενες τιμές) για τις παραμέτρους μας $\boldsymbol{\theta}^{(0)}$

   Για $J = 1, \ldots, N$ έχουμε:

2. Προσομοιώνουμε με σειρά ψευδοτυχαίες τιμές για τις παραμέτρους θ_i όπου $i = 1, \ldots, k$ από την κατανομή εισήγησης:

$$\begin{aligned}
\theta_1^{(J+1)} &\sim p\left(\theta_1^{(J)}|\theta_2^{(J)}, \theta_3^{(J)}, \ldots, \theta_k^{(J)}, \boldsymbol{x}\right) \\
\theta_2^{(J+1)} &\sim p\left(\theta_2^{(J)}|\theta_1^{(J+1)}, \theta_3^{(J)}, \ldots, \theta_k^{(J)}, \boldsymbol{x}\right) \\
\theta_3^{(J+1)} &\sim p\left(\theta_3^{(J)}|\theta_1^{(J+1)}, \theta_2^{(J+1)}, \theta_4^{(J)}, \ldots, \theta_k^{(J)}, \boldsymbol{x}\right) \\
&\vdots \\
\theta_k^{(J+1)} &\sim p\left(\theta_k^{(J)}|\theta_1^{(J+1)}, \theta_2^{(J+1)}, \theta_3^{(J+1)}, \ldots, \theta_{k-1}^{(J+1)}, \boldsymbol{x}\right)
\end{aligned}$$

3. Επαναλαμβάνουμε το βήμα 2, N φορές.

Ο δειγματολήπτης Gibbs προϋποθέτει πως μπορούμε και έχουμε τη δυνατότητα να προσομοιώσουμε τιμές από την κατανομή εισήγησης $p\left(\theta_i^{(*)}|\boldsymbol{\theta}_{-i}^{(J)}, \boldsymbol{x}\right)$ για όλα τα i. Ωστόσο, σε πραγματικά προβλήματα μπορεί για κάποιο ή κάποια θ_i να μας είναι αδύνατο να γεννήσουμε τιμές από την κατανομή εισήγησης χρησιμοποιώντας εύκολες μεθόδους (π.χ μέθοδο αντιστροφής ή μέθοδο απόρριψης). Σε αυτή την περίπτωση χρησιμοποιούμε τον αλγόριθμο Metropolis-Hastings με στόχο να προσομοιώσουμε τιμές από την $p\left(\theta_i^{(*)}|\boldsymbol{\theta}_{-i}^{(J)}, \boldsymbol{x}\right)$. Αυτή η μείξη των δύο αυτών αλγορίθμων στη βιβλιογραφία ονομάζεται Metropolis-within-Gibbs.

2 Μπεϋζιανές μέθοδοι επιλογής μοντέλων

2.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα αναπτύξουμε Μπεϋζιανές μεθόδους επιλογής μοντέλων. Αυτές οι προσεγγίσεις προσφέρουν πολλά πλεονεκτήματα σε σχέση με τις παραδοσιακές μεθόδους, όπως ο κλασικός έλεγχος υποθέσεων ή η διασταυρωμένη επικύρωση (cross-validation). Αντί να βασιζόμαστε σε αυθαίρετα κατώφλια ή υποθέσεις σχετικά με την κατανομή των δεδομένων, οι Μπεϋζιανές μέθοδοι ακολουθούν μια πιθανολογική (probabilistic) προσέγγιση, επιτρέποντάς μας να αναθέτουμε εκ των προτέρων πιθανότητες σε κάθε μοντέλο και με βάση τα παρατηρούμενα δεδομένα να τις ενημερώνουμε με τη βοήθεια του θεωρήματος Bayes. Αυτό μπορεί να οδηγήσει σε πιο ακριβή και αξιόπιστη επιλογή μοντέλων, ειδικά όταν πρόκειται για πολύπλοκα ή υψηλών διαστάσεων μοντέλα, όπου οι παραδοσιακές μέθοδοι μπορεί να αποτύχουν. Με αυτό τον τρόπο, οι μέθοδοι που θα αναπτύξουμε έχουν γίνει όλο και πιο δημοφιλή εργαλεία για την ανάλυση δεδομένων σε ένα ευρύ φάσμα πεδίων.

2.1.1 Μειονεκτήματα του κλασικού ελέγχου υποθέσεων

Στην κλασική στατιστική η θεμελίωση του ελέγχου υποθέσεων έγινε από τους Jerzy Neyman, Egon Pearson και τον Ronald Fisher ο οποίος ανέπτυξε τη θεωρία για την τιμή σημαντικότητας (p-value) (Biau et al., 2010). Τα βήματα του κλασικού ελέγχου υποθέσεως είναι τα εξής:

1. Ορίζουμε τη μηδενική υπόθεση H_0 και την εναλλακτική H_1.
2. Προσδιορίζουμε ένα επίπεδο σημαντικότητας α (συνήθως βάζουμε τις τιμές $0.01,\ 0.05$ ή 0.1) και επιλέγουμε ένα κατάλληλο στατιστικό (test statistic) T.
3. Προσδιορίζουμε την τιμή σημαντικότητας p, όπου είναι η πιθανότητα το στατιστικό T του δείγματος να έχει τιμή πιο ακραία σε σχέση με αυτή που παρατηρούμε από τα δεδομένα, βρισκόμενοι κάτω από τη μηδενική υπόθεση H_0.
4. Συγκρίνουμε την τιμή p με το επίπεδο σημαντικότητας α. Εάν η τιμή p είναι μικρότερη από α, απορρίπτουμε τη μηδενική υπόθεση. Ενώ, αν η τιμή p είναι μεγαλύτερη ή ίση με α, αποτυγχάνουμε να απορρίψουμε τη μηδενική υπόθεση.
5. Τέλος ερμηνεύουμε τα αποτελέσματα και εξάγουμε συμπεράσματα έχοντας ως βάση την απόφαση που πάρθηκε στο βήμα 4.

Ωστόσο, είναι σημαντικό να σημειωθεί ότι ο κλασικός έλεγχος υποθέσεων παρουσιάζει αρκετά μειονεκτήματα (Carlin and Louis, 1997). Αρχικά μας παρέχει μόνο στοιχεία απόρριψης ή αποτυχίας απόρριψης της μηδενικής υπόθεσης, αποτέλεσμα το οποίο δε μας λέει πόσο πιθανό είναι η μηδενική υπόθεση να είναι αληθής ή ψευδής. Επίσης το αποτέλεσμα του ελέγχου εξαρτάται σε μεγάλο βαθμό από το μέγεθος του δείγματος. Μια μικρή αλλαγή στο μέγεθος του δείγματος μπορεί να έχει σημαντικό αντίκτυπο στο αποτέλεσμα του ελέγχου, ακόμη και όταν η κατανομή των δεδομένων παραμένει η ίδια. Εφόσον η επιλογή του επιπέδου σημαντικότητας α αφήνεται στην υποκειμενική κρίση του ερευνητή, καταλαβαίνουμε ότι δεν υπάρχει σαφές κριτήριο για να αποφασίσουμε πότε μια τιμή p είναι αρκετά

μικρή ώστε να απορρίψουμε τη μηδενική υπόθεση, οδηγώντας έτσι σε πιθανή μεροληψία. Τέλος ο κλασικός έλεγχος υποθέσεων περιορίζεται σε απλές υποθέσεις και δεν είναι σε θέση να χειριστεί πολύπλοκες υποθέσεις όπως αυτές που συναντάμε στη σύγχρονη επιστήμη δεδομένων και τη μηχανική μάθηση.

2.2 Παράγοντας του Bayes και Μπεϋζιανή στάθμιση μοντέλων

Ο έλεγχος υποθέσεων κατά Bayes που αναπτύχθηκε με τη βοήθεια του Harold Jeffreys στις αρχές του 20ου αιώνα (Jeffreys, 1935), μας δίνει μια προσέγγιση η οποία λύνει τα παραπάνω προβλήματα που δημιουργεί ο κλασικός έλεγχος υποθέσεων. Μας δίνει την ευχέρεια να χρησιμοποιήσουμε πιο πολύπλοκες υποθέσεις στη θέση του H_0 και H_1.

Έστω, ότι θέλουμε να συγκρίνουμε δύο μοντέλα M_0 και M_1 με άγνωστες παραμέτρους $\boldsymbol{\theta}_0$ και $\boldsymbol{\theta}_1$ αντίστοιχα. Αρχικά ορίζουμε τις εκ των προτέρων πιθανότητες των δύο αυτών μοντέλων $p(M_0)$, $p(M_1)$ με $p(M_0) = 1 - p(M_1)$ και υπό την Μπεϋζιανή σκοπιά ο έλεγχος υπόθεσης στο συγκεκριμένο πρόβλημα είναι ο εξής:

- H_0: $\boldsymbol{X} \sim M_0$, όπου $f(\boldsymbol{x}|\boldsymbol{\theta}_0, M_0)$ είναι η πιθανοφάνεια και $p(\boldsymbol{\theta}_0|M_0)$ αποτελεί την εκ των προτέρων κατανομή των παραμέτρων του μοντέλου M_0.
- H_1: $\boldsymbol{X} \sim M_1$, όπου $f(\boldsymbol{x}|\boldsymbol{\theta}_1, M_1)$ είναι η πιθανοφάνεια και $p(\boldsymbol{\theta}_1|M_1)$ αποτελεί την εκ των προτέρων κατανομή των παραμέτρων του μοντέλου M_1.

Συνεχίζουμε υπολογίζοντας με τη βοήθεια του θεωρήματος του Bayes τις εκ των υστέρων πιθανότητες των δύο αυτών μοντέλων

$$p(M_0|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|M_0)p(M_0)}{f(\boldsymbol{x})}$$

και

$$p(M_1|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|M_1)p(M_1)}{f(\boldsymbol{x})},$$

όπου $f(\boldsymbol{x}|M_0)$ και $f(\boldsymbol{x}|M_1)$ είναι περιθώριες πιθανοφάνειες που δίνονται από το εξής ολοκλήρωμα:

$$f(\boldsymbol{x}|M_k) = \int f(\boldsymbol{x}|M_k, \boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|M_k)\, d\boldsymbol{\theta}_k, \;\; k = 0, 1, \ldots \tag{2.2.1}$$

Ο όρος που προκύπτει στον παρονομαστή $f(\boldsymbol{x})$ δίνεται από τον τύπο:

$$f(\boldsymbol{x}) = f(\boldsymbol{x}|M_0) \cdot p(M_0) + f(\boldsymbol{x}|M_1) \cdot p(M_1).$$

Ένας τρόπος για να αποφασίσουμε μεταξύ του μοντέλου M_0 και M_1 (δηλαδή μεταξύ των υποθέσεων H_0 και H_1) είναι να συγκρίνουμε τις εκ των υστέρων πιθανότητες $p(M_0|\boldsymbol{x})$, $p(M_1|\boldsymbol{x})$ και να δεχτούμε το μοντέλο με την υψηλότερη εκ των υστέρων πιθανότητα (maximum a posteriori probability model (MAP model)).

Παράλληλα μπορούμε να συγκρίνουμε τα μοντέλα M_0 και M_1, διαιρώντας τις εκ των υστέρων πιθανότητές τους και έτσι προκύπτει ο εκ των υστέρων λόγος πιθανότητας (posterior odds ratio) των δύο μοντέλων:

$$PO_{01} = \frac{p(M_0|\boldsymbol{x})}{p(M_1|\boldsymbol{x})} = \frac{f(\boldsymbol{x}|M_0)p(M_0)}{f(\boldsymbol{x}|M_1)p(M_1)}, \tag{2.2.2}$$

που περιέχει το πηλίκο των περιθώριων πιθανοφανειών και το πηλίκο των εκ των προτέρων πιθανοτήτων (prior odds ratio) των δύο αυτών μοντέλων. Το πηλίκο των περιθώριων πιθανοφανειών ονομάζεται ως παράγοντας του Bayes (Bayes factor) και έχει τον τύπο

$$BF_{01} = \frac{f(\boldsymbol{x}|M_0)}{f(\boldsymbol{x}|M_1)}, \tag{2.2.3}$$

οπότε από την (2.2.2) και (2.2.3) έχουμε

$$PO_{01} = \frac{p(M_0|\boldsymbol{x})}{p(M_1|\boldsymbol{x})} = BF_{01} \times \frac{p(M_0)}{p(M_1)}, \tag{2.2.4}$$

όπου παρατηρούμε ότι αν $p(M_0) = p(M_1)$ τότε $PO_{01} = BF_{01}$.

Η ερμηνεία των τιμών του παράγοντα Bayes BF_{10} (παράγοντας του Bayes του μοντέλου M_1 σε σχέση με το μοντέλο M_0) σύμφωνα με τους Kass and Raftery (1995) δίνεται από το Πίνακα 2.2.1.

$\log(BF_{10})$	BF_{10}	Ενδείξεις υπέρ της H_1
$0-1$	$1-3$	Αμελητέες
$1-3$	$3-20$	Θετικές
$3-5$	$20-150$	Ισχυρές
>5	>150	Πολύ ισχυρές

Πίνακας 2.2.1: Ερμηνεία του παράγοντα Bayes σύμφωνα με τους Kass and Raftery (1995).

Μέχρι τώρα έχουμε εξηγήσει πως γίνεται η σύγκριση δύο μόνο μοντέλων με τη βοήθεια του παράγοντα Bayes. Η σύγκριση αυτή μπορεί να επεκταθεί σε σύγκριση πολλών στο πλήθος μοντέλων. Έστω για παράδειγμα ότι θέλουμε να συγκρίνουμε τα μοντέλα $M_1, \ldots, M_c$, όπου στην H_0 υποθέτουμε ότι έχουμε το μοντέλο M_1. Η εκ των υστέρων πιθανότητα για ένα μοντέλο M_i όπου $i = 1, \ldots, c$ είναι

$$p(M_i|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|M_i)p(M_i)}{f(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|M_i)p(M_i)}{\sum_{j=1}^{c} f(\boldsymbol{x}|M_j)p(M_j)}, \tag{2.2.5}$$

όπου

- $p(M_i)$ είναι η εκ των προτέρων πιθανότητα του μοντέλου $i = 1, \ldots, c$ με $\sum_{i=1}^{c} p(M_i) = 1$.
- $f(\boldsymbol{x}|M_i)$ είναι η περιθώρια πιθανοφάνεια του μοντέλου M_i που δίνεται από τον τύπο (2.2.1).

Επίσης, αντί να διαλέξουμε ένα μόνο μοντέλο M_i με τη βοήθεια του παράγοντα Bayes και την εκ των υστέρων πιθανότητα $p(M_i|\boldsymbol{x})$ για να κάνουμε προβλέψεις, θα μπορούσαμε να χρησιμοποιήσουμε τις εκ των υστέρων πιθανότητες ενός συνόλου μοντέλων (ή όλων) ως βάρη προκειμένου να προβούμε σε πρόβλεψη κάποιας ποσότητας, έστω Δ. Η μέθοδος Μπεϋζιανής στάθμισης μοντέλων (Bayesian model Averaging (BMA)) χρησιμοποιείται όταν υπάρχει αβεβαιότητα επιλογής ενός μοντέλου και παρέχει τη δυνατότητα επιλογής ενός συνόλου μοντέλων $M_1, \ldots, M_c$. Έτσι μπορούμε να υπολογίσουμε την εκ των υστέρων κατανομή της ποσότητας Δ με τον εξής τύπο:

$$p(\Delta|\boldsymbol{x}) = \sum_{i=1}^{c} p(\Delta|M_i, \boldsymbol{x}) p(M_i|\boldsymbol{x}), \tag{2.2.6}$$

όπου $p(\Delta|M_i, \boldsymbol{x})$ είναι η εκ των υστέρων κατανομή της ποσότητας Δ στο μοντέλο M_i και $p(M_i|\boldsymbol{x})$ είναι η εκ των υστέρων πιθανότητα του μοντέλου M_i που δίνεται από τον τύπο (2.2.5).

Συνεπώς, έχοντας κάνει μία εισαγωγή στον παράγοντα Bayes και στην Μπεϋζιανή στάθμιση μοντέλων, καταλαβαίνουμε ότι οι επιλογές των εκ των προτέρων κατανομών των παραμέτρων παίζουν σημαντικό ρόλο στον υπολογισμό των εκ των υστέρων πιθανοτήτων των μοντέλων και ότι ο υπολογισμός του ολοκληρώματος (2.2.1), μπορεί να φανεί αρκετά δύσκολος σε περίπτωση που δεν έχουμε συζυγείς εκ των προτέρων κατανομές για τις παραμέτρους.

2.3 Το παράδοξο των Lindley-Bartlett

Το παράδοξο των Lindley-Bartlett είναι ένα στατιστικό φαινόμενο όπου η Μπεϋζιανή και η κλασική προσέγγιση στον έλεγχο υπόθεσης δίνουν διαφορετικά αποτελέσματα. Πήρε το όνομα του από τον Dennis Lindley, ο οποίος εισήγαγε για πρώτη φορά το παράδοξο το 1957 (Lindley, 1957).

Έστω, ότι έχουμε τον παρακάτω έλεγχο υπόθεσης και $\boldsymbol{x} = (x_1, \ldots, x_n)$ παρατηρήσεις:

- H_0: $X_i \sim N(\theta_0, \sigma^2)$, όπου θ_0, σ^2 είναι γνωστά
- H_1: $X_i \sim N(\theta \neq \theta_0, \sigma^2)$, όπου σ^2 είναι γνωστό, ενώ η παράμετρος θ είναι άγνωστη.

Το μοντέλο M_0 της μηδενικής υπόθεσης H_0 δεν περιέχει άγνωστη παράμετρο, ενώ το μοντέλο M_1 της εναλλακτικής υπόθεσης H_1 έχει την άγνωστη παράμετρο θ. Για την άγνωστη παράμετρο θ του μοντέλου M_1 ορίζουμε την εξής εκ των προτέρων κατανομή:

$$\theta|M_1 \sim N(\theta_0, \sigma_\theta^2).$$

Επίσης ορίζουμε τις εκ των προτέρων πιθανότητες των υποθέσεων $p(H_0)$, $p(H_1)$ και με τη βοήθεια της (2.2.1) βρίσκουμε τις περιθώριες πιθανοφάνειες για το μοντέλο M_0 και το M_1. Όπως αναφέραμε και παραπάνω, το μοντέλο M_0 δεν περιέχει άγνωστες παραμέτρους, οπότε η περιθώρια πιθανοφάνεια είναι ίση με:

$$f(\boldsymbol{x}|M_0) = (2\pi\sigma^2)^{-(n/2)} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \theta_0)^2\right).$$

Συνεχίζουμε υπολογίζοντας την περιθώρια πιθανοφάνεια του M_1:

$$\begin{aligned} f(\boldsymbol{x}|M_1) &= \int f(\boldsymbol{x}|M_1,\theta)\cdot p(\theta|M_1)\,d\theta \\ &= \int (2\pi\sigma^2)^{-(n/2)}\cdot \\ &\cdot \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right)\cdot(2\pi\sigma_\theta^2)^{-(1/2)}\exp\left(-\frac{1}{2\sigma_\theta^2}\sum_{i=1}^{n}(\theta-\theta_0)^2\right)d\theta \\ &= (2\pi\sigma^2)^{-(n/2)}\left(\frac{\sigma^2}{n\sigma_\theta^2+\sigma^2}\right)^2\exp\left(\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})+\frac{n\,(\bar{x}-\theta_0)^2}{1+n\left(\frac{\sigma_\theta}{\sigma}\right)^2}\right]\right), \end{aligned}$$

όπου αντικαθιστώντας τα $f(\boldsymbol{x}|M_0)$, $f(\boldsymbol{x}|M_1)$ στη σχέση (2.2.2) προκύπτει η γενική μορφή του PO_{01}

$$\begin{aligned} PO_{01} &= \sqrt{1+n\left(\frac{\sigma_\theta}{\sigma}\right)^2}\cdot \\ &\cdot \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(x_i-\theta_0)^2-\sum_{i=1}^{n}(x_i-\bar{x})-\frac{n\,(\bar{x}-\theta_0)^2}{1+n\left(\frac{\sigma_\theta}{\sigma}\right)^2}\right]\right)\frac{p(H_0)}{p(H_1)} \end{aligned} \tag{2.3.1}$$

όπου ξέρουμε ότι $\sum_{i=1}^{n}(x_i-\theta_0)^2=\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\theta_0)^2$, οπότε

$$\begin{aligned} PO_{01} &= \sqrt{1+n\left(\frac{\sigma_\theta}{\sigma}\right)^2}\exp\left(-\frac{1}{2\sigma^2}\left[n(\bar{x}-\theta_0)^2-\frac{n(\bar{x}-\theta_0)^2}{1+n\left(\frac{\sigma_\theta}{\sigma}\right)^2}\right]\right)\frac{p(H_0)}{p(H_1)} \\ &= \sqrt{1+n\left(\frac{\sigma_\theta}{\sigma}\right)^2}\exp\left(-\frac{1}{2\sigma^2}\left[\frac{\frac{n^2\sigma_\theta^2}{\sigma^4}(\bar{x}-\theta_0)^2}{1+n\left(\frac{\sigma_\theta^2}{\sigma^2}\right)}\right]\right)\frac{p(H_0)}{p(H_1)} \\ &= \sqrt{1+n\left(\frac{\sigma_\theta}{\sigma}\right)^2}\exp\left(-\frac{1}{2}\left[\frac{\frac{n^2\sigma_\theta^2}{\sigma^2}(\bar{x}-\theta_0)^2}{\sigma^2+n\sigma_\theta^2}\right]\right)\frac{p(H_0)}{p(H_1)}. \end{aligned}$$

Γνωρίζοντας ότι το τυχαίο δείγμα ακολουθεί την Κανονική κατανομή με μέση τιμή θ_0 και διακύμανση σ^2, διερευνούμε το σενάριο όπου τα δείγματα βρίσκονται στο όριο της περιοχής απόρριψης με επίπεδο σημαντικότητας $\alpha=q$ άρα $\bar{x}=\theta_0\pm z_{q/2}\frac{\sigma}{\sqrt{n}}$. Έτσι καταλήγουμε στην τελική μορφή του PO_{01}:

$$\begin{aligned} PO_{01} &= \sqrt{1+n\left(\frac{\sigma_\theta}{\sigma}\right)^2}\exp\left(-\frac{1}{2}\left[\frac{\frac{n^2\sigma_\theta^2}{\sigma^2}z_{q/2}^2\frac{\sigma^2}{n}}{\sigma^2+n\sigma_\theta^2}\right]\right)\frac{p(H_0)}{p(H_1)} \\ &= \sqrt{1+n\left(\frac{\sigma_\theta}{\sigma}\right)^2}\exp\left(-\frac{1}{2}\left[\frac{n\sigma_\theta^2}{\sigma^2+n\sigma_\theta^2}\right]z_{q/2}^2\right)\frac{p(H_0)}{p(H_1)}. \end{aligned} \tag{2.3.2}$$

Από τις σχέσεις (2.3.1) και (2.3.2) παρατηρούμε ότι ο λόγος των εκ των υστέρων πιθανοτήτων αυξάνεται όσο μεγαλώνει το μέγεθος του τυχαίου δείγματος. Δηλαδή όταν $n\to\infty$ τότε $PO_{01}\to\infty$, οπότε τείνουμε να δεχόμαστε τη μηδενική υπόθεση.

Πράγμα το οποίο έρχεται σε αντίθεση με την κλασική στατιστική όπου αν $n \to \infty$ τείνουμε να απορρίπτουμε την H_0 (Lindley, 1957). Επιπλέον παρατηρούμε ότι όσο αυξάνεται η τιμή της διασποράς της εκ των προτέρων κατανομής της παραμέτρου, αυξάνεται και ο λόγος των εκ των υστέρων πιθανοτήτων PO_{01} (δηλαδή $\sigma_\theta^2 \to \infty \Rightarrow PO_{01} \to \infty$) (Bartlett, 1957). Άρα και στις δύο περιπτώσεις η Μπεϋζιανή προσέγγιση δέχεται τη μηδενική υπόθεση.

Για να αφαιρέσουμε αυτή την εξάρτηση μεταξύ του λόγου των εκ των υστέρων πιθανοτήτων με το μέγεθος του δείγματος, μπορούμε να θέσουμε τη διασπορά της εκ των προτέρων κατανομής να εξαρτάται από το n. Δηλαδή να χρησιμοποιήσουμε $\frac{\sigma_\theta^2}{n}$ αντί για $\sigma_{\boldsymbol{\theta}}^2$.

Συνεπώς, από την παρατήρηση που έκανε ο Maurice Bartlett καταλαβαίνουμε ότι η επιλογή της διασποράς σ_θ^2 παρουσιάζει μια πρόκληση όταν δεν έχουμε πληροφορία για την παράμετρο θ, καθώς η επιλογή αυτή πρέπει να βρει μία ισορροπία μεταξύ του να είναι αρκετά μεγάλη ώστε να αποφευχθεί τυχόν μεροληψία από την εκ των προτέρων κατανομή σε κάθε μοντέλο και να μην είναι πολύ μεγάλη έτσι ώστε να ενεργοποιήσει το παράδοξο Lindley-Bartlett ευνοώντας το απλούστερο μοντέλο. Αυτή είναι μία λεπτή αντιστάθμιση που απαιτεί προσεκτική εξέταση.

2.4 Τρόποι υπολογισμού της περιθώριας πιθανοφάνειας

Στην Ενότητα 2.2 είδαμε ότι η περιθώρια πιθανοφάνεια χρησιμεύει στη σύγκριση μοντέλων βοηθώντας έτσι τους ερευνητές να προσδιορίσουν ποιο μοντέλο εξηγεί καλύτερα τα δεδομένα. Ωστόσο όπως αναφέραμε και στην Ενότητα 1.5, ο υπολογισμός της περιθώριας πιθανοφάνειας μπορεί να είναι δύσκολος, ειδικά όταν ασχολούμαστε με πολλές παραμέτρους. Συνεπώς η εκτίμηση της περιθώριας πιθανοφάνειας μπορεί να αποτελεί μία πρόκληση, ειδικά για πολύπλοκα μοντέλα (π.χ μη-γραμμικά μοντέλα), όμως διάφορες αριθμητικές και υπολογιστικές τεχνικές, όπως τεχνικές βασισμένες σε MCMC (Ενότητα 1.6), η μέθοδος Laplace και ο δειγματολήπτης σπουδαιότητας (importance sampling), έχουν αναπτυχθεί για την αντιμετώπιση αυτού του προβλήματος.

2.4.1 Μέθοδος του Laplace

Η μέθοδος του Laplace γενικά μας βοηθάει να προσεγγίσουμε το ολοκλήρωμα μίας συνάρτησης $\int h(\boldsymbol{\theta})\, d\boldsymbol{\theta}$, προσαρμόζοντας μία Κανονική κατανομή στο μέγιστο $\hat{\boldsymbol{\theta}}$ της $h(\boldsymbol{\theta})$ και υπολογίζοντας τον όγκο της. Ο πίνακας συνδιακύμανσης της Κανονικής κατανομής καθορίζεται από τον Εσσιανό πίνακα του $\log h(\boldsymbol{\theta})$ στο μέγιστο $\hat{\boldsymbol{\theta}}$ (MacKay, 1998). Στην Μπεϋζιανή στατιστική η μέθοδος του Laplace χρησιμοποιείται για την προσέγγιση της εκ των υστέρων κατανομής με χρήση μίας Κανονικής κατανομής που είναι κεντραρισμένη στη μέγιστη εκ των υστέρων εκτίμηση (maximum a posteriori estimate (MAP)). Άρα εφαρμόζουμε τη μέθοδο του Laplace με $h(\boldsymbol{\theta}) = f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Αυτό δικαιολογείται από το γεγονός ότι υπό ορισμένες συνθήκες, η εκ των υστέρων κατανομή προσεγγίζει την Κανονική όσο ο αριθμός του δείγματος μεγαλώνει (Gelman et al., 1995).

Ξεκινάμε ορίζοντας ότι $\hat{\boldsymbol{\theta}}_{MAP} \approx \boldsymbol{\theta}_{MAP} = \operatorname{argmax} p(\boldsymbol{\theta}|\boldsymbol{x})$, όπου το $\hat{\boldsymbol{\theta}}_{MAP}$ είναι μία

προσέγγιση της μέγιστης εκ των υστέρων εκτίμησης και υπολογίζεται με τη χρήση κάποιας υπολογιστικής μεθόδου. Επιπρόσθετα θεωρούμε την εξής προσέγγιση της εκ των υστέρων κατανομής $p(\boldsymbol{\theta}|\boldsymbol{x})$ στο $\hat{\boldsymbol{\theta}}_{MAP}$:

$$\hat{p}(\boldsymbol{\theta}|\boldsymbol{x}) = N\left(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}_{MAP}, \hat{\boldsymbol{\Sigma}}\right),$$

όπου $\hat{\boldsymbol{\Sigma}} \approx -\boldsymbol{H}^{-1}$ είναι μία προσέγγιση του Εσσιανού πίνακα του $\log(f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}))$ στο $\hat{\boldsymbol{\theta}}_{MAP}$. Έτσι, καταλήγουμε στην εξής προσέγγιση της περιθώριας πιθανοφάνειας:

$$\hat{f}(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{MAP})p(\hat{\boldsymbol{\theta}}_{MAP})}{N(\hat{\boldsymbol{\theta}}_{MAP}|\hat{\boldsymbol{\theta}}_{MAP}, \hat{\boldsymbol{\Sigma}})} = (2\pi)^{\frac{\dim(\boldsymbol{\theta})}{2}} \left|\hat{\boldsymbol{\Sigma}}\right|^{\frac{1}{2}} f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{MAP})p(\hat{\boldsymbol{\theta}}_{MAP}).$$

Ισοδύναμα, μπορούμε να καταλήξουμε στο ίδιο αποτέλεσμα αναπτύσσοντας το $\log\left(f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right)$ σε τετραγωνική μορφή γύρω από το $\hat{\boldsymbol{\theta}}_{MAP}$, δηλαδή έχουμε το εξής:

$$\log\left(f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right) \approx \log\left(f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{MAP})p(\hat{\boldsymbol{\theta}}_{MAP})\right) - \frac{1}{2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MAP}\right)^T \hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MAP}\right). \tag{2.4.1.1}$$

Για την περιθώρια πιθανοφάνεια έχουμε ότι

$$f(\boldsymbol{x}) = \int f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int \exp\{\log\left(f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right)\}\, d\boldsymbol{\theta},$$

χρησιμοποιώντας τη σχέση (2.4.1.1) έχουμε το εξής:

$$\begin{aligned} f(\boldsymbol{x}) &= \int \exp\{\log\left(f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\right)\}\, d\boldsymbol{\theta} \\ &\approx \int \exp\{\log\left(f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{MAP})p(\hat{\boldsymbol{\theta}}_{MAP})\right) - \frac{1}{2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MAP}\right)^T \hat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{MAP}\right)\}\, d\boldsymbol{\theta} \\ &= (2\pi)^{\frac{\dim(\boldsymbol{\theta})}{2}} \left|\hat{\boldsymbol{\Sigma}}\right|^{\frac{1}{2}} f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{MAP})p(\hat{\boldsymbol{\theta}}_{MAP}). \end{aligned} \tag{2.4.1.2}$$

Πολλές φορές οι αναλυτικοί υπολογισμοί των $\hat{\boldsymbol{\theta}}_{MAP}$ και $\hat{\boldsymbol{\Sigma}}$ είναι αρκετά δύσκολοι και χρονοβόροι. Επομένως για να αποφύγουμε αναλυτικούς υπολογισμούς, μπορούμε να χρησιμοποιήσουμε αλγορίθμους MCMC, έχοντας ως στόχο να εκτιμήσουμε τις ποσότητες $\hat{\boldsymbol{\theta}}_{MAP}$ και $\hat{\boldsymbol{\Sigma}}$.

Για την εκτίμηση των $\hat{\boldsymbol{\theta}}_{MAP}$ και $\hat{\boldsymbol{\Sigma}}$, οι Lewis and Raftery (1997) πρότειναν τη χρήση δειγμάτων από τον αλγόριθμο Metropolis-Hastings (Ενότητα 1.6.1). Αυτή η παραλλαγή ονομάζεται εκτιμητής Laplace-Metropolis.

2.4.2 Κριτήριο BIC

Το κριτήριο BIC (Bayesian information criterion) επιδιώκει να ελαχιστοποιήσει την επιρροή της εκ των προτέρων κατανομής όσο το δυνατόν περισσότερο. Επομένως, τη θέση της ποσότητας $\hat{\boldsymbol{\theta}}_{MAP}$ θα πάρει η ποσότητα που μεγιστοποιεί τη συνάρτηση πιθανοφάνειας, δηλαδή την ποσότητα $\hat{\boldsymbol{\theta}}_{MLE} \approx \boldsymbol{\theta}_{MLE} = \operatorname{argmax} f(\boldsymbol{x}|\boldsymbol{\theta})$. Από τον Schwarz (1978) έχουμε το εξής:

$$\text{BIC} = \dim(\boldsymbol{\theta})\log(n) - 2\log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{MLE}),$$

όπου η ποσότητα $\hat{\boldsymbol{\theta}}_{MLE}$ μπορεί να υπολογιστεί με τη βοήθεια κάποιας υπολογιστικής μεθόδου MCMC.

Για την εκτίμηση της περιθώριας πιθανοφάνειας μπορούμε να λογαριθμήσουμε τη σχέση (2.4.1.2) στο $\hat{\boldsymbol{\theta}}_{MLE}$ (Konishi and Kitagawa, 2008):

$$\log f(\boldsymbol{x}) \approx \log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{MLE}) + \underbrace{\log p(\hat{\boldsymbol{\theta}}_{MLE}) + \frac{\dim(\boldsymbol{\theta})}{2}\log(2\pi) - \frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}|}_{\text{όροι που ποινικοποίουν την πολυπλοκότητα του μοντέλου}}.$$

Υποθέτοντας ότι έχουμε ασαφή εκ των προτέρων Κανονική κατανομή για τις παραμέτρους και ότι ο Εσσιανός πίνακας είναι πλήρης τάξης (full rank), η λογαριθμοποιημένη μορφή της περιθώριας πιθανοφάνειας είναι η εξής:

$$\log f(\boldsymbol{x}) \approx \log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{MLE}) - \frac{\dim(\boldsymbol{\theta})}{2}\log(n),$$

οπότε η περιθώρια πιθανοφάνεια είναι η εξής:

$$f(\boldsymbol{x}) \approx \exp\{\log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_{MLE}) - \frac{\dim(\boldsymbol{\theta})}{2}\log(n)\} = \exp\{-\frac{1}{2}\text{BIC}\}, \quad n \to \infty,$$

και έτσι όσο ο αριθμός του δείγματος αυξάνεται καταλήγουμε ότι $\text{BIC} \approx -2\log f(\boldsymbol{x})$ ασυμπτωτικά.

Από τις παραπάνω σχέσεις καταλαβαίνουμε ότι μικρότερες τιμές του κριτηρίου BIC συνδέονται με καλύτερα μοντέλα. Επίσης παρατηρούμε ότι το κριτήριο BIC λαμβάνει υπόψη την πολυπλοκότητα του κάθε μοντέλου, αφού υψηλότερες τιμές στο BIC δίνονται σε μοντέλα με περισσότερο αριθμό παραμέτρων. Συγκεκριμένα η ποινή $\dim(\boldsymbol{\theta})\log(n)$ λαμβάνει υπόψιν την υπερπροσαρμογή (overfitting), καθώς αυξάνοντας τον αριθμό των παραμέτρων στο μοντέλο, αυξάνεται και η προσαρμογή.

2.4.3 Αφελής Monte Carlo εκτιμητής και Δειγματολήπτης Σπουδαιότητας

Ένας αρκετά αφελής τρόπος για να προσεγγίσουμε το ολοκλήρωμα είναι να χρησιμοποιήσουμε τη μέθοδο Monte carlo. Συγκεκριμένα το ολοκλήρωμα μπορεί να εκφραστεί ως $f(\boldsymbol{\theta}) = \mathbb{E}_{p(\boldsymbol{\theta})}[f(\boldsymbol{x}|\boldsymbol{\theta})]$, οπότε μπορούμε να δημιουργήσουμε ένα τυχαίο δείγμα μεγέθους J ($\{\boldsymbol{\theta}^{(j)}\}_{j=1}^{J}$) από την εκ των προτέρων κατανομή $p(\boldsymbol{\theta})$ και έπειτα να εκτιμήσουμε την περιθώρια πιθανοφάνεια με τον εξής τύπο:

$$\hat{f}(\boldsymbol{x}) = \frac{1}{J}\sum_{j=1}^{J} f(\boldsymbol{x}|\boldsymbol{\theta}^{(j)}), \quad \{\boldsymbol{\theta}^{(j)}\}_{j=1}^{J} \sim p(\boldsymbol{\theta}).$$

Σε αυτή την προσέγγιση παρατηρούμε ότι η εκτίμηση $\hat{f}(\boldsymbol{x})$ δεν θα είναι ιδανική (υψηλή διακύμανση) στην περίπτωση που η εκ των προτέρων κατανομή διαφέρει σημαντικά από την εκ των υστέρων κατανομή.

Σε αυτή την περίπτωση μπορούμε να προχωρήσουμε μέσω του δειγματολήπτη

σπουδαιότητας που στηρίζεται στην ακόλουθη ισότητα:

$$f(\boldsymbol{\theta}) = \int_{\boldsymbol{\Theta}} f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \mathbb{E}_{p^*}\left[\frac{f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p^*(\boldsymbol{\theta})}\right] = \int_{\boldsymbol{\Theta}} \frac{f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p^*(\boldsymbol{\theta})} p^*(\boldsymbol{\theta})\, d\boldsymbol{\theta},$$

όπου $p^*(\boldsymbol{\theta})$ είναι μία συνάρτηση πυκνότητας πιθανότητας στο $\mathbb{R}^{\dim(\boldsymbol{\theta})}$.

Προσομοιώνοντας J τιμές $\{\boldsymbol{\theta}^j\}_{j=1}^J$ από τη συνάρτηση πυκνότητας $p^*(\boldsymbol{\theta})$, η εκτίμηση της περιθώριας πιθανοφάνειας είναι η εξής:

$$\hat{f}(\boldsymbol{x}) = \frac{1}{J}\sum_{j=1}^{J} \frac{f\left(\boldsymbol{x}|\boldsymbol{\theta}^{(j)}\right) p\left(\boldsymbol{\theta}^{(j)}\right)}{p^*\left(\boldsymbol{\theta}^{(j)}\right)} = \frac{1}{J}\sum_{j=1}^{J} r_j f\left(\boldsymbol{x}|\boldsymbol{\theta}^{(j)}\right), \quad \{\boldsymbol{\theta}^{(j)}\}_{j=1}^J \sim p^*(\boldsymbol{\theta}),$$

όπου $r_j = \frac{p\left(\boldsymbol{\theta}^{(j)}\right)}{p^*\left(\boldsymbol{\theta}^{(j)}\right)}$. Από την παραπάνω σχέση παρατηρούμε ότι ο αφελής Monter Carlo εκτιμητής αντιστοιχεί στο δειγματολήπτη σπουδαιότητας όταν $p^*(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$.

2.4.4 Εκτιμητής Αρμονικού Μέσου και Αντίστροφος Δειγματολήπτης Σπουδαιότητας

Ο εκτιμητής αρμονικού μέσου (harmonic mean estimator) προέρχεται από την εξής μέση τιμή:

$$\begin{aligned}\mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{x})}\left[\frac{1}{f(\boldsymbol{x}|\boldsymbol{\theta})}\right] &= \int_{\boldsymbol{\Theta}} \frac{1}{f(\boldsymbol{x}|\boldsymbol{\theta})} p(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta} \\ &= \frac{1}{f(\boldsymbol{\theta})}\int_{\boldsymbol{\Theta}} \frac{1}{f(\boldsymbol{x}|\boldsymbol{\theta})} f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \frac{1}{f(\boldsymbol{\theta})}\int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \frac{1}{f(\boldsymbol{\theta})}.\end{aligned}$$

Εδώ η βασική ιδέα είναι να κάνουμε χρήση της εκ των υστέρων κατανομής $p(\boldsymbol{\theta}|\boldsymbol{x})$. Εφόσον χρειάζεται να πάρουμε δείγμα από την $p(\boldsymbol{\theta}|\boldsymbol{x})$, θα χρειαστεί να χρησιμοποιήσουμε κάποιον αλγόριθμο MCMC (Ενότητα 1.6). Ως εκ τούτου, ο εκτιμητής αρμονικού μέσου της περιθώριας πιθανοφάνειας είναι ο εξής:

$$\hat{f}(\boldsymbol{x}) = \left(\frac{1}{\frac{1}{J}\sum_{j=1}^{J}\frac{1}{f\left(\boldsymbol{x}|\boldsymbol{\theta}^{(j)}\right)}}\right), \quad \{\boldsymbol{\theta}^{(j)}\}_{j=1}^J \sim p(\boldsymbol{\theta}|\boldsymbol{x}). \tag{2.4.4.1}$$

Ο αρμονικός μέσος αποτελεί μία αρκετά απλή προσέγγιση εκτίμησης της περιθώριας πιθανοφάνειας, στην οποία έχει αποδειχθεί ότι είναι ευαίσθητη σε ακραίες τιμές, γεγονός που μπορεί να οδηγήσει σε αύξηση της διακύμανσης και έτσι σε αναξιόπιστες εκτιμήσεις. Λόγω των μειονεκτημάτων του, ο εκτιμητής αρμονικού μέσου δε συνιστάται ευρέως για πρακτική χρήση και πιο προηγμένες τεχνικές, όπως η μέθοδος του Laplace ή οι μέθοδοι που βασίζονται σε αλγόριθμους MCMC, προτιμώνται για ακριβή εκτίμηση της περιθώριας πιθανοφάνειας. Ωστόσο, ο εκτιμητής αυτός χρησιμεύει ως αφετηρία για την κατανόηση των προκλήσεων που εμπλέκονται στη σύγκριση μοντέλων στην Μπεϋζιανή στατιστική.

Ο αντίστροφος δειγματολήπτης σπουδαιότητας (reverse importance sampling (RIS)) ορίζεται από την εξής ισότητα (Gelfand and Dey, 1994):

$$\frac{1}{f(\boldsymbol{x})} = \mathbb{E}_{p(\boldsymbol{\theta}|\boldsymbol{x})}\left[\frac{h(\boldsymbol{\theta})}{f(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}\right],$$

όπου $h(\boldsymbol{\theta})$ είναι μία βοηθητική συνάρτηση με $\int_{\boldsymbol{\Theta}} h(\boldsymbol{\theta})\, d\boldsymbol{\theta} = 1$. Συνεπώς, η εκτίμηση της περιθώριας πιθανοφάνειας είναι η εξής:

$$\hat{f}(\boldsymbol{x}) = \left(\frac{1}{J}\sum_{j=1}^{J}\frac{h\left(\boldsymbol{\theta}^{(j)}\right)}{f\left(\boldsymbol{x}|\boldsymbol{\theta}^{(j)}\right)p\left(\boldsymbol{\theta}^{(j)}\right)}\right)^{-1}, \quad \{\boldsymbol{\theta}^{(j)}\}_{j=1}^{J} \sim p(\boldsymbol{\theta}|\boldsymbol{x}), \tag{2.4.4.2}$$

όπου για την προσομοίωση τιμών από την εκ των υστέρων κατανομή $p(\boldsymbol{\theta}|\boldsymbol{x})$ μπορούμε πάλι να χρησιμοποιήσουμε αλγόριθμους MCMC. Από τις σχέσεις (2.4.4.1) και (2.4.4.2) καταλαβαίνουμε ότι ο εκτιμητής αρμονικού μέσου αποτελεί μία ειδική περίπτωση του αντίστροφου δειγματολήπτη σπουδαιότητας, όπου $h(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$. Για την επιλογή της συνάρτησης $h(\boldsymbol{\theta})$, οι Robert and Wraith (2009) πρότειναν τη χρήση ομοιόμορφης σε περιοχές με υψηλή εκ των υστέρων πυκνότητα, ενώ οι Wang et al. (2018) για την $h(\boldsymbol{\theta})$ προτείνουν τη χρήση συνάρτησης που είναι τμηματικά σταθερή.

2.5 Θετικά και αρνητικά των Μπεϋζιανών μεθόδων επιλογής μοντέλων

Έχοντας μιλήσει για κάποιες βασικές Μπεϋζιανές μεθόδους επιλογής μοντέλων, είμαστε σε θέση να πούμε τα πλεονεκτήματα και τα μειονεκτήματα των μεθόδων αυτών. Κάποια βασικά πλεονεκτήματα είναι ότι:

- Μας δίνεται η δυνατότητα αυτόματης επιλογής του «καλύτερου» μοντέλου.
- Παρέχει ερμηνεύσιμα αποτελέσματα που μπορούν εύκολα να γίνουν κατανοητά και από το ευρύ κοινό.
- Παρέχουν τη δυνατότητα επιλογής ενός συνόλου εξίσου εύλογων μοντέλων σε περίπτωση που υπάρχει αβεβαιότητα επιλογής ενός μοντέλου.
- Είναι ευέλικτες και μπορούν να χειριστούν πολύπλοκα μοντέλα με πολλές παραμέτρους, όπως ιεραρχικά μοντέλα και μοντέλα με μη-γραμμικούς όρους.
- Υπάρχει η δυνατότητα επιλογής μοντέλων μέσω μεθόδων Markov chain Monte Carlo οι οποίοι και αυτοί με τη σειρά τους δίνουν αξιόλογα αποτελέσματα.

Κάποια βασικά μειονεκτήματα είναι:

- Η επιλογή μη-πληροφοριακών ή ασθενώς πληροφοριακών εκ των προτέρων κατανομών για τις παραμέτρους μπορεί να μας οδηγήσουν στο παράδοξο των Lindley-Bartlett. Κατ'επέκταση η χρήση των μη γνήσιων εκ των προτέρων (improper priors) θα δημιουργήσει έναν παράγοντα Bayes οποίος θα εξαρτάται από άγνωστες σταθερές κανονικοποίησης.

- Ο υπολογισμός του ολοκληρώματος που περιέχει η περιθώρια πιθανοφάνεια τις περισσότερες φορές είναι δύσκολος.
- Οι μέθοδοι αυτοί είναι απαιτητικές όταν το σύνολο των υποψήφιων μοντέλων είναι μεγάλο.

3 Μπεϋζιανές μέθοδοι επιλογής μεταβλητών μοντέλου

3.1 Εισαγωγή

Η επιλογή των κατάλληλων επεξηγηματικών μεταβλητών σε πολλαπλά γραμμικά μοντέλα αποτελεί ένα από τα πιο βασικά βήματα στην ανάλυση δεδομένων και γενικότερα στη Στατιστική μοντελοποίηση. Ωστόσο, είναι συχνά δύσκολο να προσδιοριστεί αυτό το υποσύνολο των επεξηγηματικών μεταβλητών που θα μας οδηγήσει σε ένα μοντέλο που είναι εύκολα εξηγήσιμο και οικονομικό στους υπολογισμούς.

Σε αυτό το κεφάλαιο θα ασχοληθούμε με το πολλαπλό γραμμικό μοντέλο, αλλά υπό την Μπεϋζιανή σκοπιά. Αυτό σημαίνει ότι οι παράμετροι του γραμμικού μας μοντέλου θα αποτελούν τυχαίες μεταβλητές, όπου θα χρειαστεί να προσδιορίσουμε εκ των προτέρων κατανομές που θα καθρεφτίζουν την πρότερη γνώση μας ή την έλλειψη αυτής για τις παραμέτρους.

Η επιλογή της εκ των προτέρων κατανομής όπως είδαμε στο Κεφάλαιο 2 και όπως θα δούμε στο παρόν κεφάλαιο, παίζει καθοριστικό ρόλο στην Μπεϋζιανή στατιστική διότι σε περίπτωση χρήσης ασαφών ή μη γνήσιων εκ των προτέρων κατανομών, μας οδηγούν σε πολλά προβλήματα (παράδοξο Lindley-Bartlett 2.2 και 2.3).

Συνοπτικά λοιπόν, σε αυτό το κεφάλαιο θα μιλήσουμε για υποψήφιες εκ των προτέρων κατανομές για τις παραμέτρους του μοντέλου και θα ορίσουμε τις μεθόδους Stochastic Search Variable Selection (SSVS), δειγματολήπτη Kuo & Mallick και Gibbs Variable Selection (GVS).

3.2 Επιλογή εκ των προτέρων κατανομών για συντελεστές πολλαπλού γραμμικού μοντέλου

Προτού ξεκινήσουμε να εμβαθύνουμε σε πρωτόγνωρες έννοιες για εμάς, καλό είναι να κάνουμε μια εισαγωγή στο γνωστό σε όλους μας πολλαπλό γραμμικό μοντέλο.

Έστω, ότι έχουμε τη μεταβλητή απόκρισης $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ και διαθέτουμε τις επεξηγηματικές μεταβλητές $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ με συντελεστές $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$. Στο κανονικό πολλαπλό γραμμικό μοντέλο έχουμε:

$$Y_i \sim N(\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_p X_{p,i}, \sigma^2),\ i = 1, \ldots, n, \tag{3.2.1}$$

και σε μορφή πινάκων έχουμε

$$\boldsymbol{Y} \sim N(\mathbf{1}_n \beta_0 + \boldsymbol{\mathcal{X}}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n), \quad \epsilon_i \sim N(0, \sigma^2) \tag{3.2.2}$$

όπου $\boldsymbol{\mathcal{X}}$ είναι ο πίνακας σχεδιασμού που έχει την εξής μορφή

$$\boldsymbol{\mathcal{X}} = \begin{bmatrix} X_{1,1} & X_{2,1} & \cdots & X_{p,1} \\ X_{1,2} & X_{2,2} & \cdots & X_{p,2} \\ \vdots & \vdots & \ddots & \vdots \\ X_{1,n} & X_{2,n} & \cdots & X_{p,n} \end{bmatrix} \in \mathbb{R}^{n \times p}$$

και $\boldsymbol{B} = (\beta_0, \boldsymbol{\beta})$.

Οι άγνωστες παράμετροι του παραπάνω μοντέλου είναι οι συντελεστές $\boldsymbol{B}$ και η διασπορά σ^2. Εφόσον έχουμε να κάνουμε με άγνωστες παραμέτρους, θα χρειαστεί να προσδιορίσουμε τις εκ των προτέρων κατανομές για το διάνυσμα των παραμέτρων $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}, \sigma^2)$, χρησιμοποιώντας εκ των προτέρων γνώσεις ή πληροφορίες (αλλιώς κάνουμε χρήση μη-πληροφοριακών εκ των προτέρων κατανομών). Στόχος μας είναι να βρούμε την εκ των υστέρων κατανομή για τις παραμέτρους του μοντέλου μας, αλλά και την εκ των υστέρων πιθανότητα του κάθε μοντέλου $M_k \in \mathcal{M}$ όπου $\mathcal{M}$ είναι το σύνολο όλων των πιθανών μοντέλων.

Αν $\boldsymbol{\beta} \in \mathbb{R}^p$, έστω $\boldsymbol{\gamma}$ είναι ένα διάνυσμα διάστασης p, όπου $\gamma_j = 1$ αν ο συντελεστής β_j του διανύσματος $\boldsymbol{\beta}$ συμπεριλαμβάνεται στο μοντέλο και $\gamma_j = 0$ αν δε συμπεριλαμβάνεται. Για παράδειγμα, αν

$$\boldsymbol{\gamma} = (1, 1, 0, 0, 1, 1, 0),$$

αυτό σημαίνει ότι το διάνυσμα $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ είναι διάστασης $p = 7$, όπου η πρώτη, δεύτερη, πέμπτη και έκτη συνιστώσα είναι διάφορες του μηδενός. Συνεπώς ο πίνακας σχεδιασμού $\boldsymbol{\mathcal{X}}_{\boldsymbol{\gamma}}$ θα περιέχει τις στήλες που αντιστοιχούν στις μη μηδενικές στήλες του $\boldsymbol{\gamma}$ και $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ θα είναι διάστασης $p_{\boldsymbol{\gamma}}$ και θα περιέχει τα στοιχεία του $\boldsymbol{\beta}$ όταν $\boldsymbol{\gamma}$ είναι 1 (η σταθερά β_0 βρίσκεται σε όλα τα μοντέλα). Ως εκ τούτου στο μοντέλο $M_{\boldsymbol{\gamma}}$, το $\boldsymbol{\mu}_{\boldsymbol{\gamma}}$ θα είναι το εξής:

$$\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \mathbf{1}_n \beta_0 + \boldsymbol{\mathcal{X}}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}.$$

Προτού ξεκινήσουμε να μιλάμε για διάφορες επιλογές εκ των προτέρων κατανομών, είναι λογικό να υποθέσουμε ότι $p+1 < n$, ο πίνακας $\boldsymbol{Q} = [\mathbf{1}_n \ \boldsymbol{\mathcal{X}}]$ είναι μέγιστης τάξης (δηλαδή οι γραμμές του είναι ανεξάρτητες) και ότι οι στήλες του πίνακας σχεδιασμού $\boldsymbol{\mathcal{X}}$ του πλήρους μοντέλου είναι κεντραρισμένες στον αντίστοιχο μέσο όρο τους σε κάθε στήλη.

3.2.1 Χρήση της εκ των προτέρων κατανομής του Jeffreys

Στην περίπτωση που δεν κατέχουμε καμία πληροφορία για τις παραμέτρους του μοντέλου, μπορούμε να χρησιμοποιήσουμε την εκ των προτέρων κατανομή του Jeffreys που είναι η εξής:

$$p(\beta_0, \boldsymbol{\beta}_{M_{\boldsymbol{\gamma}}}, \sigma^2) \propto \sigma^{-2}.$$

Γνωρίζουμε πως η πιθανοφάνεια του μοντέλου $M_{\boldsymbol{\gamma}}$ δίνεται από τον εξής τύπο:

$$f(\boldsymbol{Y}|\boldsymbol{\beta}_{M_{\boldsymbol{\gamma}}}, \boldsymbol{\gamma}, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \times \exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{Y} - \mathbf{1}_n\beta_0 - \boldsymbol{\mathcal{X}}_{M_{\boldsymbol{\gamma}}}\boldsymbol{\beta}_{M_{\boldsymbol{\gamma}}}\right)^T \left(\boldsymbol{Y} - \mathbf{1}_n\beta_0 - \boldsymbol{\mathcal{X}}_{M_{\boldsymbol{\gamma}}}\boldsymbol{\beta}_{M_{\boldsymbol{\gamma}}}\right)\right]. \tag{3.2.1.1}$$

Έστω $\bar{\boldsymbol{Y}} = \frac{1}{n}\sum_{i=1}^n y_i$ είναι ο μέσος της μεταβλητής απόκρισης και $\mathbf{1}_n^T \boldsymbol{\mathcal{X}} = \mathbf{0}_n^T$ καθώς οι στήλες του πίνακα σχεδιασμού είναι κεντραρισμένες στον αντίστοιχο μέσο όρο τους σε κάθε στήλη, ο όρος στο εκθετικό μπορεί να γραφτεί ως εξής:

$$\begin{aligned}
&\left(\boldsymbol{Y}-\mathbf{1}_n\beta_0-\boldsymbol{\mathcal{X}}_{M_\gamma}\boldsymbol{\beta}_{M_\gamma}\right)^T\left(\boldsymbol{Y}-\mathbf{1}_n\beta_0-\boldsymbol{\mathcal{X}}_{M_\gamma}\boldsymbol{\beta}_{M_\gamma}\right)=\\
&=\left(\boldsymbol{Y}-\bar{\boldsymbol{Y}}\mathbf{1}_n-\boldsymbol{\mathcal{X}}_{M_\gamma}\boldsymbol{\beta}_{M_\gamma}+\left(\bar{\boldsymbol{Y}}-\beta_0\right)\mathbf{1}_n\right)^T\left(\boldsymbol{Y}-\bar{\boldsymbol{Y}}\mathbf{1}_n-\boldsymbol{\mathcal{X}}_{M_\gamma}\boldsymbol{\beta}_{M_\gamma}+\left(\bar{\boldsymbol{Y}}-\beta_0\right)\mathbf{1}_n\right)\\
&=\left(\boldsymbol{Y}-\bar{\boldsymbol{Y}}\mathbf{1}_n-\boldsymbol{\mathcal{X}}_{M_\gamma}\boldsymbol{\beta}_{M_\gamma}\right)^T\left(\boldsymbol{Y}-\bar{\boldsymbol{Y}}\mathbf{1}_n-\boldsymbol{\mathcal{X}}_{M_\gamma}\boldsymbol{\beta}_{M_\gamma}\right)+\\
&+2\left(\bar{\boldsymbol{Y}}-\beta_0\right)\mathbf{1}_n^T\left(\boldsymbol{Y}-\bar{\boldsymbol{Y}}\mathbf{1}_n-\boldsymbol{\mathcal{X}}_{M_\gamma}\boldsymbol{\beta}_{M_\gamma}\right)+n\left(\bar{\boldsymbol{Y}}-\beta_0\right)^2\\
&=\left(\boldsymbol{Y}-\bar{\boldsymbol{Y}}\mathbf{1}_n-\boldsymbol{\mathcal{X}}_{M_\gamma}\boldsymbol{\beta}_{M_\gamma}\right)^T\left(\boldsymbol{Y}-\bar{\boldsymbol{Y}}\mathbf{1}_n-\boldsymbol{\mathcal{X}}_{M_\gamma}\boldsymbol{\beta}_{M_\gamma}\right)+n\left(\bar{\boldsymbol{Y}}-\beta_0\right)^2,
\end{aligned}$$

διότι $\mathbf{1}_n^T\left(\boldsymbol{Y}-\bar{\boldsymbol{Y}}\mathbf{1}_n-\boldsymbol{\mathcal{X}}_{M_\gamma}\boldsymbol{\beta}_{M_\gamma}\right)=\left(n\bar{\boldsymbol{Y}}-n\bar{\boldsymbol{Y}}\right)=0$. Άρα η πιθανοφάνεια δίνεται από τον εξής τύπο:

$$\begin{aligned}
f(\boldsymbol{Y}|\boldsymbol{\beta}_{M_\gamma},\boldsymbol{\gamma},\sigma^2)=&\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\cdot\exp\left[-\frac{n}{2\sigma^2}\left(\bar{\boldsymbol{Y}}-\beta_0\right)^2\right]\times\\
&\times\exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{Y}-\bar{\boldsymbol{Y}}\mathbf{1}_n-\boldsymbol{\mathcal{X}}_{M_\gamma}\boldsymbol{\beta}_{M_\gamma}\right)^T\left(\boldsymbol{Y}-\bar{\boldsymbol{Y}}\mathbf{1}_n-\boldsymbol{\mathcal{X}}_{M_\gamma}\boldsymbol{\beta}_{M_\gamma}\right)\right],
\end{aligned}\tag{3.2.1.2}$$

όπου βλέπουμε πως $\hat{\beta}_0=\bar{\boldsymbol{Y}}$ και $\hat{\boldsymbol{\beta}}_{M_\gamma}=\left(\boldsymbol{\mathcal{X}}_{M_\gamma}^T\boldsymbol{\mathcal{X}}_{M_\gamma}\right)^{-1}\boldsymbol{\mathcal{X}}_{M_\gamma}^T\boldsymbol{Y}$, ενώ ένας αμερόληπτος εκτιμητής του σ^2 είναι ο εξής

$$\hat{\sigma}^2=\frac{1}{n-p_\gamma-1}\left(\boldsymbol{Y}-\mathbf{1}_n\beta_0-\boldsymbol{\mathcal{X}}_{M_\gamma}\boldsymbol{\beta}_{M_\gamma}\right)^T\left(\boldsymbol{Y}-\mathbf{1}_n\beta_0-\boldsymbol{\mathcal{X}}_{M_\gamma}\boldsymbol{\beta}_{M_\gamma}\right)=\frac{s^2}{n-p_\gamma-1},$$

όπου η εκτίμηση μέγιστης πιθανοφάνειας είναι $\tilde{\sigma}^2=s^2/n$. Με χρήση αυτών των ισοτήτων έχουμε το εξής:

$$\begin{aligned}
f(\boldsymbol{Y}|\boldsymbol{\beta}_{M_\gamma},\boldsymbol{\gamma},\sigma^2)=&\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\times\\
&\times\exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{Y}-\hat{\beta}_0\mathbf{1}_n-\boldsymbol{\mathcal{X}}_{M_\gamma}\hat{\boldsymbol{\beta}}_{M_\gamma}\right)^T\left(\boldsymbol{Y}-\hat{\beta}_0\mathbf{1}_n-\boldsymbol{\mathcal{X}}_{M_\gamma}\hat{\boldsymbol{\beta}}_{M_\gamma}\right)\right]\times\\
&\times\exp\left[-\frac{n}{2\sigma^2}\left(\hat{\beta}_0-\beta_0\right)^2-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}_{M_\gamma}-\hat{\boldsymbol{\beta}}_{M_\gamma}\right)^T\boldsymbol{\mathcal{X}}_{M_\gamma}^T\boldsymbol{\mathcal{X}}_{M_\gamma}\left(\boldsymbol{\beta}_{M_\gamma}-\hat{\boldsymbol{\beta}}_{M_\gamma}\right)\right].
\end{aligned}\tag{3.2.1.3}$$

Ως εκ τούτου, οι από κοινού εκ των υστέρων κατανομή είναι η εξής:

$$\begin{aligned}
p(\beta_0,\boldsymbol{\beta}_{M_\gamma},\sigma^2|\boldsymbol{Y})\propto&\left(\sigma^{-2}\right)^{-n/2}\times\\
&\times\exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{Y}-\hat{\beta}_0\mathbf{1}_n-\boldsymbol{\mathcal{X}}_{M_\gamma}\hat{\boldsymbol{\beta}}_{M_\gamma}\right)^T\left(\boldsymbol{Y}-\hat{\beta}_0\mathbf{1}_n-\boldsymbol{\mathcal{X}}_{M_\gamma}\hat{\boldsymbol{\beta}}_{M_\gamma}\right)\right]\times\\
&\times\sigma^{-2}\exp\left[-\frac{n}{2\sigma^2}\left(\hat{\beta}_0-\beta_0\right)^2-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}_{M_\gamma}-\hat{\boldsymbol{\beta}}_{M_\gamma}\right)^T\boldsymbol{\mathcal{X}}_{M_\gamma}^T\boldsymbol{\mathcal{X}}_{M_\gamma}\left(\boldsymbol{\beta}_{M_\gamma}-\hat{\boldsymbol{\beta}}_{M_\gamma}\right)\right]\\
&\propto\left(\sigma^{-2}\right)^{-p_\gamma/2}\exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}_{M_\gamma}-\hat{\boldsymbol{\beta}}_{M_\gamma}\right)^T\boldsymbol{\mathcal{X}}_{M_\gamma}^T\boldsymbol{\mathcal{X}}_{M_\gamma}\left(\boldsymbol{\beta}_{M_\gamma}-\hat{\boldsymbol{\beta}}_{M_\gamma}\right)\right]\times\\
&\times\left(\sigma^{-2}\right)^{-1/2}\exp\left[-\frac{n}{2\sigma^2}\left(\hat{\beta}_0-\beta_0\right)^2\right]\times\\
&\times\left(\sigma^{-2}\right)^{-(n-p_\gamma-1)/2-1}\exp\left[-\frac{1}{2\sigma^2}s^2\right].
\end{aligned}$$

Συνεπώς, κοιτώντας το παραπάνω αποτέλεσμα, οι εκ των υστέρων κατανομές των παραμέτρων είναι οι εξής:

$$\begin{aligned}
&\beta_0|\sigma^2, \boldsymbol{Y} \sim N\left(\hat{\beta}_0, \sigma^2/n\right) \\
&\boldsymbol{\beta}_{M_\gamma}|\sigma^2, \boldsymbol{Y} \sim N\left(\hat{\boldsymbol{\beta}}_{M_\gamma}, \sigma^2\left(\boldsymbol{\mathcal{X}}_{M_\gamma}^T \boldsymbol{\mathcal{X}}_{M_\gamma}\right)^{-1}\right) \\
&\sigma^2|\boldsymbol{Y} \sim \mathrm{IG}\left((n - p_\gamma - 1)/2, s^2/2\right).
\end{aligned}$$

Αξίζει όμως να σημειωθεί ότι η χρήση μη γνήσιων (improper) εκ των προτέρων κατανομών δεν χρησιμοποιείται για την επιλογή μοντέλων, καθώς γίνεται προβληματικός ο υπολογισμός της περιθώριας πιθανοφάνειας.

3.2.2 Χρήση Κανονικής κατανομής ως εκ των προτέρων κατανομή

Όταν κάνουμε επιλογή της εκ των προτέρων κατανομής για οποιαδήποτε παράμετρο αρχικά κοιτάμε τις τιμές που μπορεί να πάρει η παράμετρος (πεδίο τιμών της παραμέτρου). Δηλαδή αν μία άγνωστη σε εμάς παράμετρος $\boldsymbol{\theta}$ έχει πεδίο τιμών όλο το $\mathbb{R}$, εμείς θα διαλέξουμε μια κατανομή που να έχει στήριγμα το $\mathbb{R}$. Στη δικιά μας περίπτωση, οι συντελεστές ενός μοντέλου $M_\gamma \in \mathcal{M}$ έχουν και αυτές πεδίο τιμών όλο το $\mathbb{R}$. Συνεπώς, δρούμε με τον ίδιο τρόπο και συνήθως επιλέγουμε ως πρότερη κατανομή την πολυμεταβλητή Κανονική κατανομή:

$$p(\boldsymbol{B}_{M_\gamma}|M_\gamma) = N(\boldsymbol{\mu}_{M_\gamma}, \boldsymbol{\Sigma}_{M_\gamma}) \tag{3.2.2.1}$$

όπου $\boldsymbol{\mu}_{M_\gamma}$ είναι η εκ των προτέρων μέση τιμή και $\boldsymbol{\Sigma}_{M_\gamma}$ είναι ο πίνακας συνδιακύμανσης των παραμέτρων $\boldsymbol{B}_{M_\gamma} = \left(\beta_0, \boldsymbol{\beta}_{M_\gamma}\right)$. Στην περίπτωση που δεν έχουμε εκ των προτέρων πληροφορία για τους συντελεστές $\boldsymbol{B}_{M_\gamma}$, θα χρειαστεί να προσαρμόσουμε τις παραμέτρους της πρότερης κατανομής έτσι ώστε να δίνει την ελάχιστη πληροφορία στο συνολικό μοντέλο μας. Για την εκ των προτέρων μέση τιμή συνήθως διαλέγουμε $\boldsymbol{\mu}_{M_\gamma} = (0, \ldots, 0)$, ενώ ο πίνακας διακύμανσης συνδιακύμανσης μπορεί να γραφτεί ως $\boldsymbol{\Sigma}_{M_\gamma} = c^2 \boldsymbol{V}_{M_\gamma}$ όπου $\boldsymbol{V}_{M_\gamma}$ είναι ο εκ των προτέρων πίνακας συσχέτισης των συντελεστών $\boldsymbol{B}_{M_\gamma}$ και c^2 είναι μια σταθερά. Αν σε αυτή τη σταθερά ορίσουμε μία μεγάλη τιμή, τότε θα πάρουμε μια εκ των προτέρων κατανομή που δεν θα προσφέρει πληροφορία (vague prior) στο μοντέλο μας. Ας θυμηθούμε όμως ότι αρκετά μεγάλες τιμές στη διακύμανση μπορεί να οδηγήσουν στο παράδοξο των Lindley-Bartlett (2.2).

Στη γραμμική παλινδρόμηση χρησιμοποιούμε συνήθως συζυγείς εκ των προτέρων κατανομές για το διάνυσμα των παραμέτρων $\boldsymbol{\theta}$. Συγκεκριμένα μπορούμε να χρησιμοποιήσουμε Κανονική κατανομή για τους συντελεστές και αντίστροφη Γάμμα κατανομή για την άγνωστη παράμετρο σ^2 (ή Γάμμα κατανομή για την παράμετρο ακρίβειας $\tau = \frac{1}{\sigma^2}$)

$$\begin{aligned}
p(\boldsymbol{B}_{M_\gamma}, \sigma^2|M_\gamma) &= p(\boldsymbol{B}_{M_\gamma}|M_\gamma, \sigma^2) \cdot p(\sigma^2) = N(\boldsymbol{\mu}_{M_\gamma}, c^2 \boldsymbol{V}_{M_\gamma} \sigma^2) \cdot \mathrm{IG}(a, b) \\
&= \mathrm{NIG}(\boldsymbol{\mu}_{M_\gamma}, c^2 \boldsymbol{V}_{M_\gamma}, a, b),
\end{aligned}$$

οπότε οι εκ των προτέρων κατανομές που θα χρησιμοποιήσουμε είναι οι εξής:

$$p(\boldsymbol{B}_{M_k}|M_k, \sigma^2) = N(\boldsymbol{\mu}_{M_k}, c^2 \boldsymbol{V}_{M_k} \sigma^2),\ p(\sigma^2) = \mathrm{IG}(a, b), \tag{3.2.2.2}$$

όπου η διασπορά σ^2 θα είναι η ίδια για κάθε μοντέλο M_γ. Από τους Bernardo and Smith (2009) ξέρουμε ότι με τη χρήση αυτών των κατανομών, η από κοινού εκ των υστέρων κατανομή $p(\boldsymbol{B}_{M_\gamma}, \sigma^2 | M_\gamma, \boldsymbol{y})$ θα είναι επίσης μία Κανονική αντίστροφη Γάμμα (Normal-Inverse Gamma). Πράγματι από Banerjee (2008) έχουμε ότι η από κοινού εκ των προτέρων κατανομή των $(\boldsymbol{B}_{M_\gamma})$ είναι η εξής:

$$
\begin{aligned}
& p\left(\boldsymbol{B}_{M_\gamma}, \sigma^2 | M_\gamma\right) = p\left(\boldsymbol{B}_{M_\gamma} | \sigma^2, \boldsymbol{\gamma}\right) p\left(\sigma^2 | \boldsymbol{\gamma}\right) = N\left(\boldsymbol{\mu}_{M_\gamma}, \sigma^2 \boldsymbol{\Sigma}_{M_\gamma}\right) \times \mathrm{IG}(a, b) \\
& = \mathrm{NIG}\left(\boldsymbol{\mu}_{M_\gamma}, \boldsymbol{\Sigma}_{M_\gamma}, a, b\right) = \frac{b^a}{(2\pi)^{(p_\gamma+1)/2} \left|\boldsymbol{\Sigma}_{M_\gamma}\right|^{1/2} \Gamma(a)} \left(\frac{1}{\sigma^2}\right)^{a+p_\gamma/2+1} \times \\
& \times \exp\left[-\frac{1}{\sigma^2}\{b + \frac{1}{2}\left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}_{M_\gamma}\right)^T \boldsymbol{\Sigma}_{M_\gamma}^{-1} \left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}_{M_\gamma}\right)\}\right] \\
& \propto \left(\frac{1}{\sigma^2}\right)^{a+(p_\gamma+1)/2+1} \times \exp\left[-\frac{1}{\sigma^2}\{b + \frac{1}{2}\left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}_{M_\gamma}\right)^T \boldsymbol{\Sigma}_{M_\gamma}^{-1} \left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}_{M_\gamma}\right)\}\right].
\end{aligned}
$$

Ολοκληρώνοντας την παραπάνω από κοινού εκ των προτέρων κατανομή ως προς σ^2 έχουμε ότι (Banerjee, 2008):

$$
\begin{aligned}
& \int \mathrm{NIG}\left(\boldsymbol{\mu}_{M_\gamma}, \boldsymbol{\Sigma}_{M_\gamma}, a, b\right) d\sigma^2 = \frac{b^a}{(2\pi)^{(p_\gamma+1)/2} \left|\boldsymbol{\Sigma}_{M_\gamma}\right|^{1/2} \Gamma(a)} \times \\
& \times \int \left(\frac{1}{\sigma^2}\right)^{a+1} \exp\left[-\frac{1}{\sigma^2}\{b + \frac{1}{2}\left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}_{M_\gamma}\right)^T \boldsymbol{\Sigma}_{M_\gamma}^{-1} \left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}_{M_\gamma}\right)\}\right] d\sigma^2 \\
& = \frac{b^a}{(2\pi)^{(p_\gamma+1)/2} \left|\boldsymbol{\Sigma}_{M_\gamma}\right|^{1/2} \Gamma(a)} \times \\
& \times \int \exp\left[-\frac{1}{\sigma^2}\{b + \frac{1}{2}\left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}_{M_\gamma}\right)^T \boldsymbol{\Sigma}_{M_\gamma}^{-1} \left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}_{M_\gamma}\right)\}\right] d\sigma^2 \\
& = \frac{b^a \Gamma\left(a + \frac{(p_\gamma+1)}{2}\right)}{(2\pi)^{(p_\gamma+1)/2} \left|\boldsymbol{\Sigma}_{M_\gamma}\right|^{1/2} \Gamma(a)} \left[b + \frac{1}{2}\left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}_{M_\gamma}\right)^T \boldsymbol{\Sigma}_{M_\gamma}^{-1} \left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}_{M_\gamma}\right)\right]^{-\left(a+\frac{(p_\gamma+1)}{2}\right)} \\
& = \frac{\Gamma\left(a + \frac{(p_\gamma+1)}{2}\right)}{\pi^{(p_\gamma+1)/2} \left|(2a)\frac{b}{a}\boldsymbol{\Sigma}_{M_\gamma}\right|^{1/2} \Gamma(a)} \times \\
& \times \left[1 + \frac{\left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}_{M_\gamma}\right)^T \left[\frac{b}{a}\boldsymbol{\Sigma}_{M_\gamma}\right]^{-1} \left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}_{M_\gamma}\right)}{2a}\right]^{-\left(\frac{2a+(p_\gamma+1)}{2}\right)},
\end{aligned}
$$

όπου είναι μία πολυμεταβλητή κατανομή Student:

$$
\begin{aligned}
MVSt_\nu\left(\boldsymbol{\mu}_{M_\gamma}, \boldsymbol{G}\right) = & \frac{\Gamma\left(\frac{\nu+(p_\gamma+1)}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right) \pi^{(p_\gamma+1)/2} \left|\nu \boldsymbol{G}\right|^{1/2}} \times \\
& \times \left[1 + \frac{\left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}_{M_\gamma}\right)^T \boldsymbol{G}^{-1} \left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}_{M_\gamma}\right)}{2a}\right]^{-\frac{\nu+(p_\gamma+1)}{2}},
\end{aligned}
$$

με $\nu = 2a$ και $\boldsymbol{G} = \left(\frac{b}{a}\right)\boldsymbol{\Sigma}_{M_\gamma}$. Ξέρουμε πως η πιθανοφάνεια του μοντέλου μπορεί να δοθεί από τον εξής τύπο:

$$f(\boldsymbol{Y}|\boldsymbol{\beta}_{M_\gamma}, \boldsymbol{\gamma}, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{Y} - \boldsymbol{Q}_{M_\gamma}\boldsymbol{B}_{M_\gamma}\right)^T\left(\boldsymbol{Y} - \boldsymbol{Q}_{M_\gamma}\boldsymbol{B}_{M_\gamma}\right)\right],$$

όπου $\boldsymbol{Q}_{M_\gamma} = \left[\boldsymbol{1}_n\ \boldsymbol{\mathcal{X}}_{M_\gamma}\right]$. Με τη χρήση του θεωρήματος του Bayes καταλήγουμε ότι η απο κοινού εκ των υστέρων κατανομή είναι η εξής (Banerjee, 2008):

$$\begin{aligned} p\left(\boldsymbol{B}_{M_\gamma}, \sigma^2|\boldsymbol{y}\right) \propto \left(\frac{1}{\sigma^2}\right)^{a+\frac{n+(p_\gamma+1)}{2}+1} \times \\ \times \exp\left[-\frac{1}{\sigma^2}\left(b^* + \frac{1}{2}\left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}^*_{M_\gamma}\right)^T\boldsymbol{\Sigma}^{*-1}_{M_\gamma}\left(\boldsymbol{B}_{M_\gamma} - \boldsymbol{\mu}^*_{M_\gamma}\right)\right)\right], \end{aligned}$$

όπου είναι μία NIG $\left(\boldsymbol{\mu}^*_{M_\gamma}, \boldsymbol{\Sigma}^{*-1}_{M_\gamma}, a^*, b^*\right)$ και έχουμε ότι:

$$\begin{aligned} \boldsymbol{\mu}^*_{M_\gamma} &= \left(\boldsymbol{\Sigma}^{-1}_{M_\gamma} + \boldsymbol{Q}^T_{M_\gamma}\boldsymbol{Q}_{M_\gamma}\right)^{-1}\left(\boldsymbol{\Sigma}^{-1}_{M_\gamma}\boldsymbol{\mu}_{M_\gamma} + \boldsymbol{Q}^T_{M_\gamma}\boldsymbol{Y}\right), \\ \boldsymbol{\Sigma}^{*-1}_{M_\gamma} &= \left(\boldsymbol{\Sigma}^{-1}_{M_\gamma} + \boldsymbol{Q}^T_{M_\gamma}\boldsymbol{Q}_{M_\gamma}\right)^{-1}, \\ a^* &= a + n/2, \\ b^* &= b + \frac{1}{2}\left[\boldsymbol{\mu}^T_{M_\gamma}\boldsymbol{\Sigma}^{-1}_{M_\gamma}\boldsymbol{\mu}_{M_\gamma} + \boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{\mu}^{*T}_{M_\gamma}\boldsymbol{\Sigma}^{*-1}_{M_\gamma}\boldsymbol{\mu}^*_{M_\gamma}\right]. \end{aligned}$$

3.2.3 Χρήση της εκ των προτέρων κατανομής του Zellner

Ο Zellner (1986) πρότεινε μία μορφή εκ των προτέρων συζυγούς κατανομής ονόματι εκ των προτέρων κατανομή g ($g-$prior). Συγκεκριμένα έχουμε:

$$\boldsymbol{\beta}_{M_\gamma}|\beta_0, \sigma^2, M_\gamma \sim N\left(0, g\sigma^2\left(\boldsymbol{\mathcal{X}}^T_{M_\gamma}\boldsymbol{\mathcal{X}}_{M_\gamma}\right)^{-1}\right),\ p(\beta_0, \sigma^2) \propto \frac{1}{\sigma^2}, \tag{3.2.3.1}$$

όπου $\boldsymbol{\mathcal{X}}_{M_\gamma}$ είναι ο πίνακας σχεδιασμού του μοντέλου M_γ και $p(\beta_0, \sigma^2)$ είναι η εκ των προτέρων κατανομή του Jeffreys.

Η σύγκριση διαφορετικών μοντέλων όπως έχουμε αναφέρει στην Ενότητα 2.2, απαιτεί τον υπολογισμό των περιθώριων πιθανοφανειών, που στη συνέχεια μπορούν να χρησιμοποιηθούν για τον υπολογισμό του παράγοντα Bayes. Στην περίπτωση της εκ των προτέρων κατανομής g η περιθώρια πιθανοφάνεια ενός μοντέλου M_γ μπορεί να υπολογιστεί και έχει την εξής μορφή:

$$f(\boldsymbol{Y}|M_\gamma, g) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\pi^{\frac{n-1}{2}}n^{\frac{1}{2}}}||\boldsymbol{Y} - \bar{\boldsymbol{Y}}||^{-(n-1)}\frac{(1+g)^{(n-1-p_\gamma)/2}}{\left(1 + g\left(1 - R^2_\gamma\right)\right)^{(n-1)/2}}, \tag{3.2.3.2}$$

όπου R^2_γ είναι ο συντελεστής προσδιορισμού του μοντέλου M_γ όπου δίνεται από τον τύπο

$$R^2_\gamma = 1 - \frac{\left(\boldsymbol{Y} - \boldsymbol{1}_n\hat{\beta}_0 - \boldsymbol{\mathcal{X}}_{M_\gamma}\hat{\boldsymbol{\beta}}_{M_\gamma}\right)^T\left(\boldsymbol{Y} - \boldsymbol{1}_n\hat{\beta}_0 - \boldsymbol{\mathcal{X}}_{M_\gamma}\hat{\boldsymbol{\beta}}_{M_\gamma}\right)}{\left(\boldsymbol{Y} - \bar{\boldsymbol{Y}}\right)^T\left(\boldsymbol{Y} - \bar{\boldsymbol{Y}}\right)},$$

όπου $\hat{\boldsymbol{\beta}}_{M_\gamma} = \left(\boldsymbol{\mathcal{X}}_{M_\gamma}^T \boldsymbol{\mathcal{X}}_{M_\gamma}\right)^{-1} \boldsymbol{\mathcal{X}}_{M_\gamma}^T \boldsymbol{Y}$ είναι η εκτίμηση μέγιστης πιθανοφάνειας (Maximum likelihood estimation (MLE)) του $\boldsymbol{\beta}_{M_\gamma}$ και $\hat{\beta}_0 = \bar{\boldsymbol{Y}}$.

Για να βρούμε τον παράγοντα Bayes ενός μοντέλου M_γ σε σχέση με το μηδενικό μοντέλο (null model) M_0, όπου για το μηδενικό μοντέλο έχουμε $p_0 = 0$ και $R_0^2 = 0$, παίρνουμε τον λόγο των περιθώριων πιθανοφανειών τους. Άρα έχουμε

$$BF_{M_\gamma M_0} = \frac{f(\boldsymbol{Y}|M_\gamma, g)}{f(\boldsymbol{Y}|M_0, g)} = \frac{(1+g)^{(n-1-p_\gamma)/2}}{\left(1+g\left(1-R_\gamma^2\right)\right)^{(n-1)/2}}. \tag{3.2.3.3}$$

Για να συγκρίνουμε το μοντέλο M_γ με πίνακα σχεδιασμού $\boldsymbol{\mathcal{X}}_{M_\gamma}$ με το πλήρες μοντέλο M_F, μπορούμε να απαρτίσουμε τον πίνακα σχεδιασμού ως εξής $\boldsymbol{\mathcal{X}} = \left(\boldsymbol{\mathcal{X}}_{M_\gamma}, \boldsymbol{\mathcal{X}}_{M_{-\gamma}}\right)$ και έτσι θα έχουμε ότι

$$M_F: \quad \boldsymbol{Y} = \mathbf{1}_n\beta_0 + \boldsymbol{\mathcal{X}}_{M_\gamma}\boldsymbol{\beta}_{M_\gamma} + \boldsymbol{\mathcal{X}}_{M_{-\gamma}}\boldsymbol{\beta}_{M_{-\gamma}} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma^2\boldsymbol{I}_n\right),$$

με $\mathbf{1}_n = (1, 1, \ldots, 1)^T \in \mathbb{R}^n$ και $\boldsymbol{\mathcal{X}}_{M_{-\gamma}}$ είναι οι στήλες του πίνακα σχεδιασμού $\boldsymbol{\mathcal{X}}$ που δεν υπάρχουν στο μοντέλο M_γ. Ο έλεγχος υπόθεσης είναι ο εξής:

- H_0: $\boldsymbol{\beta}_{M_{-\gamma}} = \mathbf{0}$, όπου αντιστοιχεί στο μοντέλο M_γ.
- H_1: $\boldsymbol{\beta}_{M_{-\gamma}} \in \mathbb{R}^{p-p_\gamma}$, όπου αντιστοιχεί στο πλήρες μοντέλο M_F.

Εφόσον το $\boldsymbol{\beta}_{M_\gamma}$ υπάρχει και στο μοντέλο M_γ και στο πλήρες μοντέλο M_F, μπορούμε να το αντιμετωπίσουμε όπως το β_0 και έτσι θα έχουμε τις εξής εκ των προτέρων κατανομές:

$$\begin{aligned} M_\gamma: &\quad p(\beta_0, \boldsymbol{\beta}_{M_\gamma}, \sigma^2) \propto \frac{1}{\sigma^2} \\ M_F: &\quad p(\beta_0, \boldsymbol{\beta}_{M_\gamma}, \sigma^2) \propto \frac{1}{\sigma^2},\ \ \boldsymbol{\beta}_{M_{-\gamma}}|\beta_0, \boldsymbol{\beta}_{M_\gamma}, \sigma^2 \sim N\left(0, g\sigma^2\left(\boldsymbol{\mathcal{X}}_{M_{-\gamma}}^T\boldsymbol{\mathcal{X}}_{M_{-\gamma}}\right)^{-1}\right), \end{aligned} \tag{3.2.3.4}$$

με παράγοντα Bayes που έχει την εξής μορφή

$$BF_{M_\gamma M_F} = \frac{f(\boldsymbol{Y}|M_\gamma, g)}{f(\boldsymbol{Y}|M_F, g)} = (1+g)^{(n-1-p)/2}\left(1+g\frac{1-R_F^2}{1-R_\gamma^2}\right)^{(n-1-p_\gamma)/2}, \tag{3.2.3.5}$$

όπου R_F^2 είναι ο συντελεστής προσδιορισμού του πλήρες μοντέλου M_F. Συνεπώς, μπορούμε να συγκρίνουμε οποιοδήποτε μοντέλο με το μηδενικό ή το πλήρες μοντέλο και έτσι μπορούμε να συγκρίνουμε οποιοδήποτε ζεύγος από μοντέλα $(M_\gamma, M_{\gamma'})$ με τον εξής τύπο

$$BF_{M_\gamma M_{\gamma'}} = \frac{BF_{M_\gamma M_0}}{BF_{M_{\gamma'} M_0}}. \tag{3.2.3.6}$$

Η επιλογή της παραμέτρου g έχει αποτελέσει αντικείμενο εκτενούς έρευνας στη βιβλιογραφία, καθώς η εκ των υστέρων κατανομή των παραμέτρων και ο παράγοντας Bayes είναι εξαιρετικά ευαίσθητα σε αυτή την επιλογή. Συγκεκριμένα, όταν το $g \to \infty$ ενώ n και p_γ είναι σταθερά, τότε από τη σχέση (3.2.3.3) βλέπουμε ότι ο παράγοντας του

Bayes θα πάει στο 0, ευνοώντας το μηδενικό μοντέλο M_0 (το μικρότερο σε διάσταση μοντέλο) και καταλήγοντας στο παράδοξο των Lindlet-Bartlett (2.2). Αντ'αυτού, θα μπορούσαμε να υποθέσουμε ότι για το μοντέλο M_γ έχουμε ότι $R^2_\gamma \to 1$ (δηλαδή ότι ένα πολύ μεγάλο ποσοστό μεταβλητότητας των δεδομένων εξηγείται από το μοντέλο M_γ). Σε αυτή την περίπτωση θα περιμέναμε τον παράγοντα Bayes να πάει στο άπειρο καθώς θεωρούμε ότι το μοντέλο προσαρμόζεται τέλεια στα δεδομένα. Βλέπουμε όμως ότι

$$BF_{M_\gamma M_0} \to (1+g)^{n-1-p_\gamma},$$

που μας δείχνει ότι ο παράγοντας Bayes θα συγκλίνει σε μια σταθερά. Αυτό το φαινόμενο ονομάζεται παράδοξο πληροφορίας (Zellner, 1986).

Στη βιβλιογραφία κάποιες πολύ γνωστές επιλογές της παραμέτρου g είναι οι εξής:

- $g = n$, όπου σε αυτή την περίπτωση η πληροφορία που θα μας δίνει η εκ των προτέρων κατανομή των $\boldsymbol{\beta}_{M_\gamma}$ θα αντιστοιχεί στην πληροφορία που περιέχει μία παρατήρηση. Τέτοιες εκ των προτέρων κατανομές ονομάζονται εκ των προτέρων κατανομές μοναδιαίας πληροφορίας (Kass and Wasserman, 1995).
- $g = p^2$, που προτείνεται από το κριτήριο πληθωριστικού κινδύνου (Risk inflation criterion (RIC)) (Foster and George, 1994).
- $g = \max(n, p^2)$, όπου συνιστούν οι Fernandez et al. (2001).

Οι Zellner and Siow (1980) πρότειναν τη χρήση πολυμεταβλητής Cauchy εκ των προτέρων κατανομής στους συντελεστές παλινδρόμησης η οποία θεωρείται κατάλληλη όταν ασχολούμαστε με πολλά μοντέλα. Συγκεκριμένα έχουμε:

$$p(\boldsymbol{\beta}_{M_\gamma}|\sigma^2) \propto \frac{\Gamma(p_\gamma/2)}{\pi^{p_\gamma/2}} \left| \frac{\boldsymbol{\mathcal{X}}^T_{M_\gamma}\boldsymbol{\mathcal{X}}_{M_\gamma}}{n\sigma^2} \right|^{\frac{1}{2}} \left(1 + \boldsymbol{\beta}^T_{M_\gamma} \frac{\boldsymbol{\mathcal{X}}^T_{M_\gamma}\boldsymbol{\mathcal{X}}_{M_\gamma}}{n\sigma^2} \boldsymbol{\beta}_{M_\gamma}\right)^{-p_\gamma/2}, \; p(\beta_0, \sigma^2) \propto \frac{1}{\sigma^2}.$$

Σε σχέση με την εκ των προτέρων κατανομή g, η εκ των προτέρων Zellner & Siow δε χρησιμοποιείται τόσο πολύ στην επιλογή μοντέλων, καθώς είναι αδύνατος ο προσδιορισμός της περιθώριας πιθανοφάνειας ειδικά όταν ασχολούμαστε με μοντέλα μεγάλης διάστασης.

Για να ληφθεί υπόψη η αβεβαιότητα που περιβάλλει την επιλογή της παραμέτρου g, μπορεί να τοποθετηθεί μία υπέρ εκ των προτέρων κατανομή (hyperprior) για τη g. Αυτή η προσέγγιση θα δώσει τη δυνατότητα στα δεδομένα μας να επηρεάσουν την τιμή της g, η οποία πλέον αντιμετωπίζεται ως τυχαία μεταβλητή στην ανάλυση. Γνωστές επιλογές για την εκ των προτέρων κατανομές της g είναι:

- Η Zellner and Siow (1980) εκ των προτέρων κατανομή μπορεί να γραφτεί ως μία μείξη της μορφής:

$$p(\boldsymbol{\beta}_{M_\gamma}|\beta_0, \sigma^2, M_\gamma) \propto \int N\left(0, g\sigma^2 \left(\boldsymbol{\mathcal{X}}^T_{M_\gamma}\boldsymbol{\mathcal{X}}_{M_\gamma}\right)^{-1}\right) p(g)\, dg,$$
$$p(g) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} \exp\left(-\frac{n}{2g}\right).$$

- Μία άλλη επιλογή της εκ των προτέρων κατανομής g στην εκ των προτέρων κατανομή Zellner & Siow είναι η εξής (Liang et al., 2008):

$$\boldsymbol{\beta}_{M_\gamma}|\beta_0, \sigma^2, M_\gamma \sim N\left(0, g\sigma^2\left(\boldsymbol{\mathcal{X}}_{M_\gamma}^T \boldsymbol{\mathcal{X}}_{M_\gamma}\right)^{-1}\right)$$
$$p(g) = \frac{\alpha - 2}{2}(1+g)^{-\alpha/2},\ g > 0$$

όπου για $\alpha \leq 2$ είναι μη γνήσια ($p(g) \propto (1+g)^{-\alpha/2}$). Από Liang et al. (2008), κατάλληλη τιμή για το α είναι η τιμή 3.

- Οι Liang et al. (2008), επισημαίνουν ότι η έλλειψη συνέπειας υπό το μηδενικό μοντέλο μας παρακινεί σε μία τροποποίηση της hyper-g prior, η οποία ονομάζεται hyper-g/n prior όπου ενσωματώνει το μέγεθος του δείγματος n και δίνεται από τον εξής τύπο:

$$p(g) = \frac{\alpha - 2}{2n}\left(1 + \frac{g}{n}\right)^{-\alpha/2}.$$

3.3 Stochastic Search Variable Selection (SSVS)

Η στοχαστική επιλογή μεταβλητών (Stochastic Search Variable selection (SSVS)) των George and McCulloch (1993), είναι μία πολύ γνωστή Μπεϋζιανή μέθοδος επιλογής μεταβλητών σε στατιστικά μοντέλα παλινδρόμησης. Η χρήση του διανύσματος δείκτη $\boldsymbol{\gamma}$ που ορίσαμε στην Ενότητα 3.2, αποτελεί ένα από τα κύρια στοιχεία του αλγορίθμου, καθώς μας επιτρέπει να κάνουμε μεμονωμένες αλλαγές σε κάθε παράμετρο του μοντέλου.

Βασικό χαρακτηριστικό της SSVS είναι ότι όλα τα 2^p υποψήφια μοντέλα (όπου το β_0 βρίσκεται σε όλα τα μοντέλα) έχουν ίδια διάσταση p (και η πιθανοφάνεια είναι $f(\boldsymbol{Y}|\boldsymbol{\beta})$ για όλα τα μοντέλα), επιλύοντας έτσι σε μεγάλο βαθμό την υπολογιστική δυσκολία η οποία συχνά προκύπτει σε Μπεϋζιανά προβλήματα επιλογής μεταβλητών. Αυτό επιτυγχάνεται χρησιμοποιώντας ως εκ των προτέρων κατανομή για κάθε παράμετρο του διανύσματος $\boldsymbol{\beta}$ μία μείξη Κανονικών κατανομών. Συγκεκριμένα αν $\gamma_j = 0$ τότε ο συντελεστής β_j θα έχει ως εκ των προτέρων κατανομή την Κανονική κεντραρισμένη στο 0 και με μεγάλη τιμή ακρίβειας τ_j, έτσι ώστε το β_j να βρίσκεται πολύ κοντά στο 0. Αν $\gamma_j = 1$ τότε η εκ των προτέρων Κανονική κατανομή του συντελεστή β_j θα έχει μικρή τιμή ακρίβειας δίνοντας έτσι τη δυνατότητα στα δεδομένα να επηρεάσουν την τιμή του β_j (την εκ των υστέρων κατανομή). Αυτή η εκ των προτέρων κατανομή θα έχει την εξής μορφή:

$$\beta_j|\gamma_j \sim (1-\gamma_j)N\left(0, \tau_j^2\right) + \gamma_j N\left(0, c_j^2\tau_j^2\right) \tag{3.3.1}$$

όπου $\mathbb{P}(\gamma_j = 1) = 1 - \mathbb{P}(\gamma_j = 0)$ για κάθε $j = 0, 1, 2, \ldots, p$. Συνεπώς, θα διαλέξουμε το τ_j^2 να είναι σχετικά μεγάλο και συγχρόνως επιδιώκουμε ο όρος $c_j^2\tau_j^2$ να είναι μικρός. Με αυτή την επιλογή αν $\gamma_j = 0$ ο συντελεστής β_j θα έχει μία ισχυρά πληροφοριακή εκ των προτέρων Κανονική κατανομή με μέση τιμή τη μηδενική υπόθεση (συνήθως την τιμή 0). Αντ'αυτού, για $\gamma_j = 1$ ο συντελεστής β_j θα έχει μία ασαφή (vague) εκ των προτέρων Κανονική κατανομή. Οι George and McCulloch (1993) έχουν αναπτύξει μία ικανοποιητική διαδικασία για την επιλογή της παραμέτρου c_j^2, η οποία στηρίζεται

στο γεγονός ότι η παράμετρος c_j μπορεί να ερμηνευθεί ως τον εκ των προτέρων λόγο των συμπληρωματικών πιθανοτήτων όταν η ανεξάρτητη μεταβλητή X_j δεν υπάρχει στο μοντέλο και το β_j είναι πολύ κοντά στο 0. Ο τρόπος επιλογής των παραμέτρων c_j^2 τ_j^2 περιγράφεται με αναλυτικό τρόπο από τους George and McCulloch (1993).

Για το διάνυσμα $\boldsymbol{\beta}$ η εκ των προτέρων κατανομή θα είναι

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N_p(\mathbf{0}, \boldsymbol{D}_\gamma \boldsymbol{V} \boldsymbol{D}_\gamma) \tag{3.3.2}$$

με $\boldsymbol{D}_\gamma = \text{diag}\,(d_0\tau_0, d_1\tau_1, d_2\tau_2, \ldots, d_p\tau_p)$ όπου για $\gamma_j = 0$ έχουμε $d_j = 1$ και για $\gamma_j = 1$ τότε $d_j = c_j$ και $\boldsymbol{V}$ είναι ο εκ των προτέρων πίνακας συνδιακύμανσης. Στη θέση του πίνακα συνδιακύμανσης $\boldsymbol{V}$ μπορούμε να βάλουμε των μοναδιαίο πίνακα $\boldsymbol{I}_p$ πράγμα που σημαίνει ότι τα β_j θα είναι εκ των προτέρων ανεξάρτητα. Ωστόσο μία άλλη επιλογή είναι να θέσουμε $\boldsymbol{V} \propto (\boldsymbol{X}^T\boldsymbol{X})$ που αντιστοιχεί στην εκ των προτέρων κατανομή g, όπου ο πίνακας συνδιακύμανσης είναι ανάλογος του πίνακα σχεδιασμού (Ενότητα 3.2).

Η σταθερά β_0 μπορεί να θεωρηθεί ότι υπάρχει σε όλα τα 2^p υποψήφια μοντέλα, προσδιορίζοντας την εκ των προτέρων πιθανότητα ύπαρξης στο κάθε μοντέλο ίση με ένα. Έτσι η εκ των προτέρων κατανομή για το β_0 είναι η εξής:

$$\beta_0 \sim N(0, c_0^2\tau_0^2),$$

όπου $c_0^2\tau_0^2$ είναι μία αρκετά μικρή τιμή. Για τη διασπορά η εκ των προτέρων κατανομή θα είναι η αντίστροφη Γάμμα

$$\sigma^2|\boldsymbol{\gamma} \sim \text{IG}(\nu_\gamma/2, \nu_\gamma\lambda_\gamma/2)$$

όπου τα ν_γ και λ_γ μπορεί να εξαρτώνται από το $\boldsymbol{\gamma}$, δηλώνοντας έτσι εξάρτηση μεταξύ του $\boldsymbol{\beta}$ με το σ^2. Στην περίπτωση που η παράμετροι του σ^2 δεν εξαρτώνται από το $\boldsymbol{\gamma}$ τότε η εκ των προτέρων κατανομή του σ^2 έχει την εξής μορφή:

$$\sigma^2 \sim \text{IG}(a, b).$$

Οι δείκτες γ_j με τη σειρά τους θα έχουν προτέρων κατανομή Bernoulli με πιθανότητα επιτυχίας το 1/2

$$\gamma_j \sim \text{Bernoulli}(1/2), \quad j = 1, \ldots, p.$$

Με τη βοήθεια του δειγματολήπτη Gibbs (1.6.2) θα δημιουργήσουμε μία ακολουθία από $\boldsymbol{\gamma}^1, \ldots, \boldsymbol{\gamma}^k$ έχοντας ως στόχο τη σύγκλιση του αλγορίθμου στην εκ των υστέρων $p(\boldsymbol{\gamma}|\boldsymbol{Y})$.

Εφαρμογή δειγματολήπτη Gibbs στο SSVS

Έστω ότι $(\beta_0, \boldsymbol{\beta}, \boldsymbol{\gamma})$ είναι το ζεύγος διανυσμάτων που αντιπροσωπεύουν το μοντέλο (έστω $\boldsymbol{B} = (\beta_0, \boldsymbol{\beta})$).

1. Ορίζουμε αρχικές τιμές (προτεινόμενες τιμές) για τις παραμέτρους μας $\boldsymbol{B}_{(0)}$ και $\boldsymbol{\gamma}_{(0)}$. Για παράδειγμα $\boldsymbol{B}_{(0)} = \hat{\boldsymbol{B}}_{OLS}$ και $\boldsymbol{\gamma}_{(0)} = (1, 1, \ldots, 1)^T$.

2. Προσομοιώνουμε με σειρά ψευδοτυχαίες τιμές για τις παραμέτρους β_j και γ_j όπου $j = 0, \ldots, p$:

$$p(\beta_j|\boldsymbol{\beta}_{-j}, \boldsymbol{\gamma}, \boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma})p(\beta_j|\gamma_j)$$

και για $j = 1, \ldots, p$ έχουμε ότι $\gamma_j \sim \text{Bernoulli}\left(\frac{O_j}{1+O_j}\right)$ όπου

$$O_j = \frac{f(\boldsymbol{\beta}|\gamma_j = 1, \boldsymbol{\gamma}_{-j})}{f(\boldsymbol{\beta}|\gamma_j = 0, \boldsymbol{\gamma}_{-j})} \frac{f(\gamma_j = 1, \boldsymbol{\gamma}_{-j})}{f(\gamma_j = 0, \boldsymbol{\gamma}_{-j})}.$$

Το $\boldsymbol{\gamma}_{-j}$ είναι το διάνυσμα $\boldsymbol{\gamma}$ χωρίς το γ_j και γ_0 παραμένει πάντα 1.

Ο παραπάνω αλγόριθμος θα παράξει ένα δείγμα από $(\boldsymbol{B}^m, \boldsymbol{\gamma}^m)$ όπου $m = 1, \ldots, N$. Με τη βοήθεια αυτού του δείγματος έχουμε τη δυνατότητα να υπολογίσουμε την εκ των υστέρων πιθανότητα για κάθε ανεξάρτητη μεταβλητή. Η εκ των υστέρων πιθανότητα ύπαρξης της μεταβλητής X_i στο μοντέλο δίνεται από τον τύπο

$$\hat{\mathbb{P}}(\gamma_j = 1|\boldsymbol{Y}) = \frac{1}{N} \sum_{m=1}^{N} \gamma_j^m. \tag{3.3.3}$$

Η μέθοδος SSVS εισήγαγε δύο βασικές ιδέες οι οποίες βοήθησαν στην ανάπτυξη άλλων προσεγγίσεων που θα δούμε παρακάτω. Η πρώτη ιδέα ήταν η χρήση του διανύσματος δείκτη $\boldsymbol{\gamma}$, όπου έδωσε τη δυνατότητα να αναπτύξουμε αλγορίθμους βασισμένους στο δειγματολήπτη Gibbs. Η δεύτερη ιδέα πρότεινε τον περιορισμό των μη σημαντικών συντελεστών σε μια γειτονιά του μηδέν αντί να οριστούν ακριβώς ίσοι με το μηδέν. Αυτό διατήρησε τη διάσταση του προβλήματος σταθερή σε όλα τα μοντέλα, επιτρέποντας έτσι την υλοποίηση του δειγματολήπτη Gibbs, με τον οποίο μπορούμε άμεσα να εκτιμήσουμε τις εκ των υστέρων πιθανότητες ύπαρξης των ανεξάρτητων μεταβλητών στο μοντέλο, χωρίς να χρειάζεται να υπολογίσουμε την περιθώρια πιθανοφάνεια κάθε μοντέλου.

3.4 Δειγματολήπτης Kuo & Mallick

Στη μέθοδο SSVS όπως αναφέραμε και παραπάνω, οι μη σημαντικοί συντελεστές β_j δεν μπορούν να οριστούν ακριβώς στο μηδέν (αλλά σε μια γειτονιά του μηδέν), ενώ οι σημαντικοί συντελεστές θα πρέπει να οριστούν έτσι ώστε να διαφέρουν σημαντικά από το μηδέν. Επομένως η επιλογή των παραμέτρων της εκ των προτέρων κατανομής (3.3.1), πρέπει να πραγματοποιηθεί με μεγάλη προσοχή, έτσι ώστε οι μη σημαντικοί συντελεστές να μην επηρεάζουν το μοντέλο μας (να βρίσκονται πολύ κοντά στο μηδέν), ενώ πρέπει να δίνεται η δυνατότητα στα δεδομένα να επηρεάσουν τους σημαντικούς συντελεστές του μοντέλου μας. Ωστόσο οι Kuo and Mallick (1998) εισήγαγαν εκ των προτέρων κατανομές όπου οι μη σημαντικοί συντελεστές θα είναι ίσοι με το μηδέν. Σε αυτήν τη προσέγγιση, το διάνυσμα δείκτη $\boldsymbol{\gamma}$ και οι συντελεστές παλινδρόμησης $\boldsymbol{\beta}$ θεωρούνται εκ των προτέρων ανεξάρτητα. Συνεπώς θα έχουμε ανεξάρτητες εκ των προτέρων κατανομές έτσι ώστε $p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = p(\boldsymbol{\beta})p(\boldsymbol{\gamma})$ και η παλινδρόμηση θα έχει την εξής μορφή:

$$y_i = \beta_0 + \sum_{j=1}^{p} \gamma_j x_{ij} \beta_j + \epsilon_i, \ \epsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right), \tag{3.4.1}$$

οπότε η πιθανοφάνεια κάθε μοντέλου θα εξαρτάται από το διάνυσμα $\boldsymbol{\gamma}$.

Για το διάνυσμα $\boldsymbol{\gamma}$ μπορεί πάλι να θεωρηθεί ότι η εκ των προτέρων κατανομή

ακολουθεί μια κατανομή Bernoulli

$$\gamma_j \sim \text{Bernoulli}(\pi_j = 1/2), \quad j = 1, \ldots, p. \tag{3.4.2}$$

όπου στην περίπτωση που $\pi_j = 0.5$, σημαίνει ότι δίνουμε ίδιες πιθανότητες σε κάθε ανεξάρτητη μεταβλητή να υπάρχει στο μοντέλο. Αλλιώς μπορούμε να προσδιορίσουμε μία εκ των προτέρων κατανομή και για την παράμετρο $\pi_j \sim \text{Unif}(0,1)$ ή $\pi_j \sim \text{Beta}(1/2, 1/2)$. Εφόσον οι συντελεστές $\boldsymbol{\beta}$ είναι ανεξάρτητοι του διανύσματος $\boldsymbol{\gamma}$ τότε θα έχουμε

- Αν επιδιώκουμε εκ των προτέρων ανεξαρτησία τότε

$$\boldsymbol{\beta} \sim N_p(0, c^2 \boldsymbol{I}_p),$$

- Αν δεν θεωρήσουμε εκ των προτέρων ανεξαρτησία, μπορούμε να θεωρήσουμε ως πρότερη την g-prior

$$\boldsymbol{\beta}|\sigma^2 \sim N_p\left(0, g\sigma^2 \left(\boldsymbol{\mathcal{X}}^T \boldsymbol{\mathcal{X}}\right)^{-1}\right)$$

όπου προσδιορίζουμε μία τιμή για το g ή ορίζουμε επιπρόσθετα εκ των προτέρων κατανομή για την παράμετρο g, ενώ η εκ των προτέρων κατανομή της σταθεράς β_0 είναι η εξής:

$$\beta_0 \sim N\left(0, c_0^2\right).$$

Εφαρμογή δειγματολήπτη Gibbs στη μέθοδο Kuo & Mallick

Έστω ότι $(\beta_0, \boldsymbol{\beta}, \boldsymbol{\gamma})$ είναι το ζεύγος διανυσμάτων που αντιπροσωπεύουν το μοντέλο (έστω $\boldsymbol{B} = (\beta_0, \boldsymbol{\beta})$).

1. Ορίζουμε αρχικές τιμές (προτεινόμενες τιμές) για τις παραμέτρους μας $\boldsymbol{B}_{(0)}$ και $\boldsymbol{\gamma}_{(0)}$. Για παράδειγμα $\boldsymbol{B}_{(0)} = \hat{\boldsymbol{B}}_{OLS}$ και $\boldsymbol{\gamma}_{(0)} = (1, 1, \ldots, 1)^T$.

2. Προσομοιώνουμε με σειρά ψευδοτυχαίες τιμές για τις παραμέτρους β_j και γ_j όπου $j = 0, \ldots, p$:

$$p(\beta_j|\boldsymbol{\beta}_{-j}, \boldsymbol{\gamma}, \boldsymbol{Y}) = \begin{cases} f(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) p(\beta_j|\boldsymbol{\beta}_{-j}), & \gamma_j = 1 \\ p(\beta_j|\boldsymbol{\beta}_{-j}), & \gamma_j = 0 \end{cases}$$

και για $j = 1, \ldots, p$ έχουμε ότι $\gamma_j \sim \text{Bernoulli}\left(\frac{O_j}{1+O_j}\right)$ όπου

$$O_j = \frac{f(\boldsymbol{Y}|\boldsymbol{\beta}, \gamma_j = 1, \boldsymbol{\gamma}_{-j})}{f(\boldsymbol{Y}|\boldsymbol{\beta}, \gamma_j = 0, \boldsymbol{\gamma}_{-j})} \frac{f(\gamma_j = 1, \boldsymbol{\gamma}_{-j})}{f(\gamma_j = 0, \boldsymbol{\gamma}_{-j})}.$$

Το $\boldsymbol{\gamma}_{-j}$ είναι το διάνυσμα $\boldsymbol{\gamma}$ χωρίς το γ_j και γ_0 παραμένει πάντα 1.

3.5 Gibbs Variable Selection (GVS)

Η μέθοδος GVS που έχει εισαχθεί από Dellaportas et al. (2002), αποτελεί μια επέκταση των μεθόδων SSVS και Kuo & Mallick, καθώς παράλληλα χρησιμοποιεί το διάνυσμα $\boldsymbol{\gamma}$ για τον προσδιορισμό του κάθε μοντέλου και η παλινδρόμηση έχει την ίδια μορφή με τη μέθοδο Kuo & Mallick (3.4.1). Όμως η εκ των προτέρων κατανομή των συντελεστών $\boldsymbol{\beta}$ και του διανύσματος $\boldsymbol{\gamma}$ έχει τη μορφή

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = p(\boldsymbol{\beta}|\boldsymbol{\gamma})p(\boldsymbol{\gamma}), \tag{3.5.1}$$

όπου $p(\boldsymbol{\beta}|\boldsymbol{\gamma})$ είναι η εκ των προτέρων κατανομή των συντελεστών $\boldsymbol{\beta}$ με δεδομένο το διάνυσμα $\boldsymbol{\gamma}$ και $p(\boldsymbol{\gamma})$ είναι η εκ των προτέρων κατανομή του διανύσματος $\boldsymbol{\gamma}$. Η συντελεστές $\boldsymbol{\beta}$ μπορούν να χωριστούν στους συντελεστές $\boldsymbol{\beta}_\gamma$ που υπάρχουν στο μοντέλο με δομή $\boldsymbol{\gamma}$ και στους συντελεστές που λείπουν από αυτό $\boldsymbol{\beta}_{-\gamma}$. Άρα η (3.5.1) μπορεί να γραφτεί ως εξής:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}) = p(\boldsymbol{\beta}_\gamma, \boldsymbol{\beta}_{-\gamma}, \boldsymbol{\gamma}) = p(\boldsymbol{\gamma})p(\boldsymbol{\beta}_\gamma|\boldsymbol{\gamma})p(\boldsymbol{\beta}_{-\gamma}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}), \tag{3.5.2}$$

ενώ όπως διακρίνουμε η πιθανοφάνεια θα εξαρτάται μόνο από τους συντελεστές που υπάρχουν στο μοντέλο και από το διάνυσμα $\boldsymbol{\gamma}$

$$f(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2) = f(\boldsymbol{Y}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, \sigma^2). \tag{3.5.3}$$

Η εκ των υστέρων κατανομή του $\boldsymbol{\beta}$ στο μοντέλο $\boldsymbol{\gamma}$ θα είναι

$$p(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{\gamma}) = p(\boldsymbol{\beta}_\gamma, \boldsymbol{\beta}_{-\gamma}|\boldsymbol{Y}, \boldsymbol{\gamma}) = p(\boldsymbol{\beta}_\gamma|\boldsymbol{Y}, \boldsymbol{\gamma})p(\boldsymbol{\beta}_{-\gamma}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, \boldsymbol{Y})$$

όπου $p(\boldsymbol{\beta}_\gamma|\boldsymbol{Y}, \boldsymbol{\gamma})$ είναι η εκ των υστέρων κατανομή των παραμέτρων $\boldsymbol{\beta}_\gamma$ στο μοντέλο και $p(\boldsymbol{\beta}_{-\gamma}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}, \boldsymbol{Y})$ αποτελεί τη δεσμευμένη εκ των προτέρων κατανομή των παραμέτρων $\boldsymbol{\beta}_{-\gamma}$ στο μοντέλο $\boldsymbol{\gamma}$. Παρατηρούμε ότι το διάνυσμα $\boldsymbol{\beta}_{-\gamma}$ δεν επηρεάζεται από τα δεδομένα και δεν προσφέρει κάποια πληροφορία στην εκ των υστέρων κατανομή των παραμέτρων του μοντέλου $p(\boldsymbol{\beta}_\gamma|\boldsymbol{Y}, \boldsymbol{\gamma})$. Έτσι λοπόν η εκ των προτέρων κατανομή $p(\boldsymbol{\beta}_{-\gamma}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma})$ λέγεται ψευδο-πρότερη και ο παραπάνω τύπος θα έχει την εξής μορφή:

$$p(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{\gamma}) = p(\boldsymbol{\beta}_\gamma, \boldsymbol{\beta}_{-\gamma}|\boldsymbol{Y}, \boldsymbol{\gamma}) = p(\boldsymbol{\beta}_\gamma|\boldsymbol{Y}, \boldsymbol{\gamma})p(\boldsymbol{\beta}_{-\gamma}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma}). \tag{3.5.4}$$

Εφαρμογή δειγματολήπτη Gibbs στη μέθοδο GVS

Έστω ότι $(\beta_0, \boldsymbol{\beta}, \boldsymbol{\gamma})$ είναι το ζεύγος διανυσμάτων που αντιπροσωπεύουν το μοντέλο (έστω $\boldsymbol{B} = (\beta_0, \boldsymbol{\beta})$).

1. Ορίζουμε αρχικές τιμές (προτεινόμενες τιμές) για τις παραμέτρους μας $\boldsymbol{B}_{(0)}$ και $\boldsymbol{\gamma}_{(0)}$. Για παράδειγμα $\boldsymbol{B}_{(0)} = \hat{\boldsymbol{B}}_{OLS}$ και $\boldsymbol{\gamma}_{(0)} = (1, 1, \ldots, 1)^T$.

2. Προσομοιώνουμε με σειρά ψευδοτυχαίες τιμές για τις παραμέτρους $\boldsymbol{B}_\gamma = (\beta_0, \boldsymbol{\beta}_\gamma)$, $\boldsymbol{\beta}_{-\gamma}$ και $\boldsymbol{\gamma}$

$$p(\boldsymbol{B}_\gamma|\boldsymbol{\beta}_{-\gamma}, \boldsymbol{\gamma}, \boldsymbol{Y}) \propto f(\boldsymbol{Y}|\boldsymbol{B}, \boldsymbol{\gamma})p(\boldsymbol{B}_\gamma|\boldsymbol{\gamma})p(\boldsymbol{\beta}_{-\gamma}|\boldsymbol{\beta}_\gamma, \boldsymbol{\gamma})$$

3. Συνεχίζουμε προσομοιώνοντας ψευδοτυχαίες τιμές για το $\boldsymbol{\beta}_{-\gamma}$:

$$p(\boldsymbol{\beta}_{-\gamma}|\boldsymbol{\beta}_{\gamma}, \boldsymbol{\gamma}, \boldsymbol{Y}) \propto p(\boldsymbol{\beta}_{-\gamma}|\boldsymbol{\beta}_{\gamma}, \boldsymbol{\gamma})$$

και για $j = 1, \ldots, p$ έχουμε ότι $\gamma_j \sim \text{Bernoulli}\left(\frac{O_j}{1+O_j}\right)$ όπου

$$O_j = \frac{f(\boldsymbol{Y}|\boldsymbol{\beta}, \gamma_j = 1, \boldsymbol{\gamma}_{-j})}{f(\boldsymbol{Y}|\boldsymbol{\beta}, \gamma_j = 0, \boldsymbol{\gamma}_{-j})} \frac{f(\boldsymbol{\beta}|\gamma_j = 1, \boldsymbol{\gamma}_{-j})}{f(\boldsymbol{\beta}|\gamma_j = 0, \boldsymbol{\gamma}_{-j})} \frac{f(\gamma_j = 1, \boldsymbol{\gamma}_{-j})}{f(\gamma_i = 0, \boldsymbol{\gamma}_{-j})}.$$

Το $\boldsymbol{\gamma}_{-j}$ είναι το διάνυσμα $\boldsymbol{\gamma}$ χωρίς το γ_j και γ_0 παραμένει πάντα 1.

Βλέπουμε λοιπόν ότι η ψευδο-πρότερη κατανομή $p(\boldsymbol{\beta}_{-\gamma}|\boldsymbol{\beta}_{\gamma}, \boldsymbol{\gamma})$ δεν επηρεάζει τις εκ των υστέρων κατανομές παρά μόνο τη σύγκλιση του αλγορίθμου. Οι εκ των προτέρων κατανομές των παραμέτρων $\boldsymbol{\beta}$ και β_0 μπορούν να οριστούν ως εξής:

$$p(\beta_j|\gamma_j) = \gamma_j N(0, \Sigma_j) + (1 - \gamma_j) N(\bar{\mu}_j, S_j), \; p(\beta_0) = N(0, \Sigma_0), \tag{3.5.5}$$

όπου $\bar{\mu}_j$ και S_j αποτελούν παράμετροι της ψευδο-πρότερης κατανομής όπου πρέπει να επιλέξουμε έχοντας ως σκοπό τη γρηγορότερη σύγκλιση. Εφόσον διαλέξουμε την (3.5.5) ως την εκ των προτέρων κατανομή των συντελεστών (όπου είναι ανεξάρτητες για κάθε β_j) και γνωρίζοντας ότι η παλινδρόμηση έχει τη μορφή (3.4.1), τότε η από κοινού εκ των υστέρων κατανομή θα δίνεται από το τύπο:

$$p(\beta_j|\boldsymbol{\beta}_{-j}, \boldsymbol{\gamma}, \boldsymbol{Y}) \propto \begin{cases} f(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) N(0, \Sigma_j), & \gamma_j = 1 \\ N(\bar{\mu}_j, S_j), & \gamma_j = 0 \end{cases} \tag{3.5.6}$$

όπου καλές επιλογές για τις παραμέτρους $\bar{\mu}_j$ και S_j αποτελούν οι τιμές $\bar{\mu}_j = 0$ και $S_j = \Sigma_j/k^2$ με $k = 10$ (Ntzoufras, 1999). Βλέπουμε λοιπόν ότι και στη μέθοδο GVS χρησιμοποιούμε την εκ των προτέρων κατανομή που ανέπτυξαν οι George and McCulloch (1993). Δηλαδή για τις παραμέτρους που δεν υπάρχουν στο μοντέλο προσφέρουμε μια ψευδο-πρότερη με αρκετά μικρή διακύμανση, ενώ για τους συντελεστές του μοντέλου χρησιμοποιούμε εκ των προτέρων κατανομή με μεγάλη διακύμανση, αφήνοντας έτσι τους συντελεστές να μπορούν να επηρεαστούν από τα δεδομένα μας.

4 Μεθοδολογίες Διαχείρισης Ελλιπών τιμών

4.1 Εισαγωγή

Ένα από τα πιο κοινά προβλήματα που αντιμετωπίζουν οι ερευνητές είναι η έλλειψη δεδομένων στην έρευνά τους. Υπάρχουν διάφοροι λόγοι για την έλλειψη δεδομένων, που κυμαίνονται από μη ελεγχόμενους παράγοντες έως και προβλήματα του σχεδιασμού. Για να αντιμετωπίσουν την έλλειψη δεδομένων οι ερευνητές μπορούν να ακολουθήσουν μια πληθώρα μεθόδων όπου κύριος σκοπός τους είναι η επίλυση αυτού του προβλήματος.

Υπάρχουν διάφοροι τύποι ελλιπών τιμών, όπως η πλήρως τυχαία έλλειψη (Missing Completely at Random(MCAR)), η τυχαία έλλειψη (Missing At Random (MAR)) και η μη τυχαία έλλειψη (Not Missing At Random (NMAR)) (Rubin, 1976). Στην πρώτη περίπτωση η έλλειψη ονομάζεται πλήρως τυχαία (MCAR), όταν η πιθανότητα έλλειψης τιμών είναι ανεξάρτητη από τα παρατηρούμενα δεδομένα (και τα μη παρατηρούμενα δεδομένα) και παραμένει ίδια για όλες τις ομάδες που δημιουργούν τα δεδομένα μας. Στη δεύτερη περίπτωση η έλλειψη ονομάζεται τυχαία (MAR), όταν η πιθανότητα έλλειψης τιμών σχετίζεται με τα παρατηρούμενα δεδομένα αλλά είναι ανεξάρτητη από τα μη παρατηρούμενα δεδομένα. Τέλος στην τρίτη περίπτωση η έλλειψη ονομάζεται μη τυχαία (NMAR), όταν η πιθανότητα έλλειψης τιμών εξαρτάται και από τα μη παρατηρούμενα δεδομένα, δηλαδή κυμαίνεται σε διάφορες τιμές στο $[0, 1]$ για λόγους άγνωστους σε εμάς.

Σε αυτό το κεφάλαιο λοιπόν θα μιλήσουμε για τους παραπάνω μηχανισμούς ελλιπών παρατηρήσεων και θα δώσουμε κάποιες μεθόδους και Μπεϋζιανές προσεγγίσεις για τη διαχείριση των ελλιπών τιμών.

4.2 Μηχανισμοί και δομές ελλιπών τιμών

Έστω ότι έχουμε το πολλαπλό γραμμικό μοντέλο από την Ενότητα 3.2, με μεταβλητή απόκρισης $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ και p επεξηγηματικές $\boldsymbol{X}_j = (X_{j,1}, \ldots, X_{j,n})$, $j = 1, 2, \ldots, p$. Για να ορίσουμε ελλιπείς τιμές που μπορεί να υπάρχουν στη μεταβλητή απόκρισης $\boldsymbol{Y}$ ή σε κάποια επεξηγηματική μεταβλητή $\boldsymbol{X}_j$ μπορούμε να ακολουθήσουμε το ίδιο σκεπτικό με αυτό του κεφαλαίου 3 και να ορίσουμε ένα σύνολο διανυσμάτων $\{\boldsymbol{R} = \left(\boldsymbol{R}_{\boldsymbol{Y}}, \boldsymbol{R}_{\boldsymbol{X}_j}\right),\ j = 1, \ldots, p\}$, όπου

$$R_{Y_j} = \begin{cases} 0, & \text{αν η τιμή της } j \text{ συνιστώσας του } \boldsymbol{Y} \text{ λείπει} \\ 1, & \text{αν η τιμή της } j \text{ συνιστώσας του } \boldsymbol{Y} \text{ έχει παρατηρηθεί} \end{cases}$$

και

$$R_{X_{j,i}} = \begin{cases} 0, & \text{αν η τιμή της } i \text{ συνιστώσας του } \boldsymbol{X}_j \text{ λείπει} \\ 1, & \text{αν η τιμή της } i \text{ συνιστώσας του } \boldsymbol{X}_j \text{ έχει παρατηρηθεί} \end{cases}.$$

Η μεταβλητή $\boldsymbol{R}$ ονομάζεται δείκτης ελλιπών τιμών (Missing data indicator) και μας βοηθά να περιγράψουμε σύνολα δεδομένων με ελλιπείς παρατηρήσεις.

Για λόγους ευκολίας ορίζουμε ως $\boldsymbol{U} \in \mathbb{R}^{(n\times(p+1))}$ τον πίνακα των δεδομένων μας και $\boldsymbol{U} = \{\boldsymbol{U}_{obs}, \boldsymbol{U}_{miss}\}$ όπου $\boldsymbol{U}_{obs}$ είναι το παρατηρήσιμο κομμάτι και $\boldsymbol{U}_{miss}$ το μη παρατηρήσιμο κομμάτι του $\boldsymbol{U}$. Ο Rubin (1976) είπε ότι ο μηχανισμός με τον οποίο

καθορίζονται οι ελλιπείς τιμές χαρακτηρίζεται από τη δεσμευμένη κατανομή του $\boldsymbol{R}$ δοθέντος του $\boldsymbol{U}$, $p(\boldsymbol{R}|\boldsymbol{U}, \boldsymbol{\phi})$, όπου $\boldsymbol{\phi}$ αποτελεί ένα διάνυσμα με άγνωστες παραμέτρους. Πρότεινε λοιπόν τρεις διαφορετικούς μηχανισμούς: την πλήρως τυχαία έλλειψη (Missing Completely at Random (MCAR)), την τυχαία έλλειψη (Missing at Random) και τη μη τυχαία έλλειψη (Not Missing at Random (NMAR)):

- **MCAR**: Τα δεδομένα θα έχουν πλήρως τυχαία έλλειψη όταν η πιθανότητα έλλειψης της κάθε παρατήρησης δεν εξαρτάται από τις παρατηρούμενες και μη τιμές. Άρα έχουμε ότι

  $$p(\boldsymbol{R}|\boldsymbol{U}_{obs}, \boldsymbol{U}_{miss}, \boldsymbol{\phi}) = p(\boldsymbol{R}|\boldsymbol{\phi}).$$

  Για παράδειγμα, σε μία ιατρική μελέτη ένα τυχαίο υποσύνολο συμμετεχόντων μπορεί να μην εμφανιστεί στα ραντεβού παρακολούθησης ή να χάσει προγραμματισμένες εργαστηριακές εξετάσεις. Η απουσία τους αυτή είναι εντελώς τυχαία και δε σχετίζεται με μεταβλητές όπως η ηλικία, το φύλο ή η σοβαρότητα της νόσου.

- **MAR**: Εδώ η πιθανότητα έλλειψης μίας παρατήρησης θα εξαρτάται από το παρατηρήσιμο κομμάτι των δεδομένων ($\boldsymbol{U}_{obs}$). Δηλαδή έχουμε ότι

  $$p(\boldsymbol{R}|\boldsymbol{U}_{obs}, \boldsymbol{U}_{miss}, \boldsymbol{\phi}) = p(\boldsymbol{R}|\boldsymbol{U}_{obs}, \boldsymbol{\phi}).$$

  Για παράδειγμα, ας υποθέσουμε ότι διεξάγουμε μια έρευνα για να διερευνήσουμε τη σχέση μεταξύ εισοδήματος και κατάστασης υγείας. Σε αυτή την έρευνα συλλέγουμε δεδομένα σχετικά με την ηλικία, το φύλο, το εισόδημα και την κατάσταση της υγείας. Ωστόσο, διαπιστώνουμε ότι οι συμμετέχοντες με χαμηλό εισόδημα είναι πιο πιθανό να έχουν ελλιπή δεδομένα σχετικά με την κατάσταση της υγείας τους. Σε αυτή την περίπτωση, η έλλειψη δεν είναι εντελώς τυχαία διότι η πιθανότητα της έλλειψης εξαρτάται από το εισόδημα, αλλά λείπει τυχαία, επειδή μόλις ληφθεί υπόψη το εισόδημα, η πιθανότητα έλλειψης είναι η ίδια για όλες τις παρατηρήσεις με το ίδιο επίπεδο εισοδήματος.

- **NMAR**: Μη τυχαία έλλειψη έχουμε όταν η πιθανότητα έλλειψης μίας παρατήρησης εξαρτάται και από το μη παρατηρήσιμο κομμάτι των δεδομένων ($\boldsymbol{U}_{miss}$). Δηλαδή η κατανομή του δείκτη ελλιπών τιμών δεν απλοποιείται

  $$p(\boldsymbol{R}|\boldsymbol{U}_{obs}, \boldsymbol{U}_{miss}, \boldsymbol{\phi}) = p(\boldsymbol{R}|\boldsymbol{U}, \boldsymbol{\phi}),$$

  όπου σε αυτή την περίπτωση ο μηχανισμός αναφέρεται επίσης μερικές φορές ως μη αμελητέος (non-ignorable). Για παράδειγμα, ας υποθέσουμε ότι μια μελέτη μετρά το εισόδημα και τα επίπεδα εκπαίδευσης των συμμετεχόντων, αλλά ορισμένοι συμμετέχοντες αρνούνται να αναφέρουν το εισόδημά τους. Εάν εκείνοι που αρνούνται να δηλώσουν το εισόδημά τους το κάνουν επειδή έχουν υψηλότερο εισόδημα, τότε τα στοιχεία που λείπουν είναι NMAR. Ένα άλλο παράδειγμα είναι οι κλινικές δοκιμές όπου οι ασθενείς αποσύρονται επειδή παρουσιάζουν παρενέργειες από τη θεραπεία, γεγονός που οδηγεί σε έλλειψη δεδομένων σχετικά με την αποτελεσματικότητα της θεραπείας. Σε αυτή την περίπτωση τα δεδομένα που λείπουν είναι NMAR, επειδή οι τιμές που λείπουν σχετίζονται με την αποτελεσματικότητα της θεραπείας και όχι μόνο από τα παρατηρούμενα δεδομένα.

Παράδειγμα (Van Buuren, 2018)

Έστω ότι τα δεδομένα μας $\boldsymbol{U} = (\boldsymbol{U}_1, \boldsymbol{U}_2)$ έχουν παραχθεί από μία δισδιάστατη Κανονική κατανομή όπου τα $\boldsymbol{U}_1$ και $\boldsymbol{U}_2$ έχουν θετική συσχέτιση και ίση με 0.6. Έστω ότι οι ελλιπείς τιμές υπάρχουν μόνο στο $\boldsymbol{U}_2$ και δημιουργούνται με τη βοήθεια του εξής μοντέλου:

$$\mathbb{P}(R_{\boldsymbol{U}_2} = 0) = \phi_0 + \frac{e^{\boldsymbol{U}_1}}{1 + e^{\boldsymbol{U}_1}}\phi_1 + \frac{e^{\boldsymbol{U}_2}}{1 + e^{\boldsymbol{U}_2}}\phi_2$$

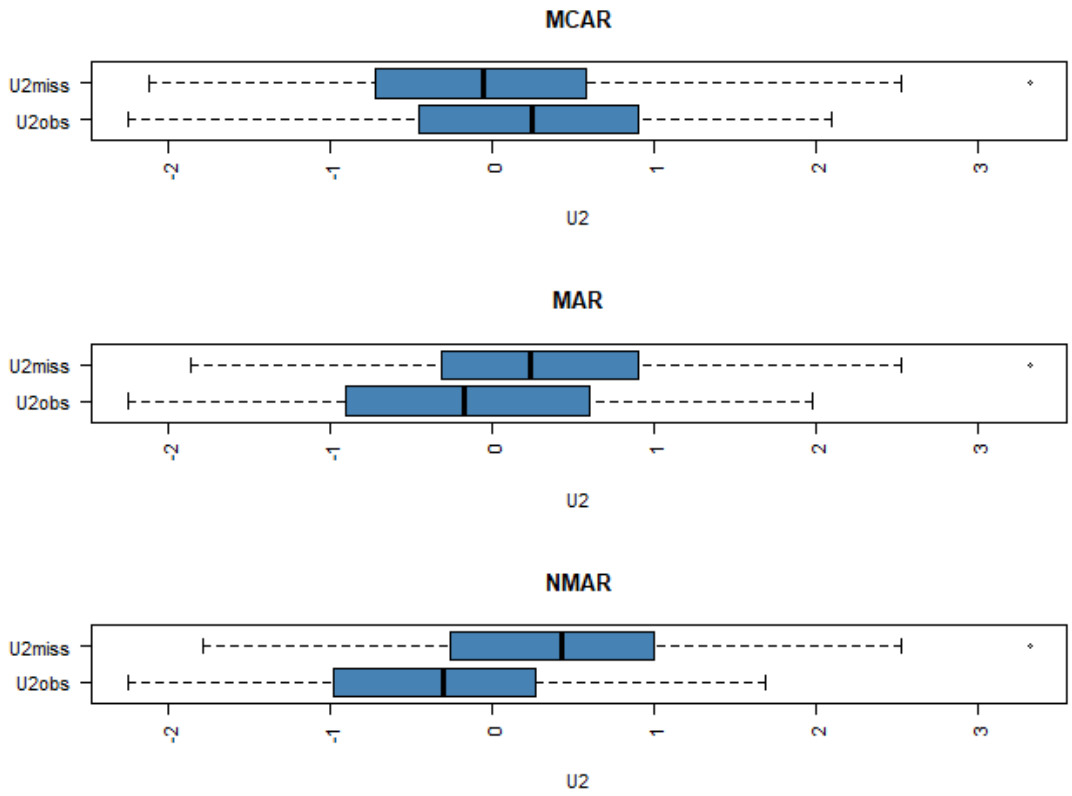
όπου $\boldsymbol{\phi} = (\phi_0, \phi_1, \phi_2)$. Για να καταλήξουμε στον κάθε μηχανισμό, προσδιορίζουμε τις εξής τιμές για το διάνυσμα των παραμέτρων $\boldsymbol{\phi}$

$$\begin{aligned}
&\textbf{MCAR}: \quad \boldsymbol{\phi}_{\textbf{MCAR}} = (0.5, 0, 0) \\
&\textbf{MAR}: \quad \boldsymbol{\phi}_{\textbf{MAR}} = (0, 1, 0) \\
&\textbf{NMAR}: \quad \boldsymbol{\phi}_{\textbf{NMAR}} = (0, 0, 1).
\end{aligned}$$

Έτσι τα μοντέλα που προκύπτουν είναι τα εξής:

$$\begin{aligned}
&\textbf{MCAR}: \quad \mathbb{P}(R_{\boldsymbol{U}_2} = 0) = 0.5 \\
&\textbf{MAR}: \quad \text{logit}\left(\mathbb{P}(R_{\boldsymbol{U}_2} = 0)\right) = \boldsymbol{U}_1 \\
&\textbf{NMAR}: \quad \text{logit}\left(\mathbb{P}(R_{\boldsymbol{U}_2} = 0)\right) = \boldsymbol{U}_2,
\end{aligned}$$

όπου $\text{logit}(p) = \left(p/\left(1-p\right)\right)$ για κάθε $0 < p < 1$.



Διάγραμμα 4.2.1: Θηκογράμματα του παρατηρήσιμου $\boldsymbol{U}_2$ και του μη παρατηρήσιμου $\boldsymbol{U}_2$ για κάθε μηχανισμό.

Στο Διάγραμμα 4.2.1 παρατηρούμε ότι υπό το μηχανισμό MCAR, το παρατηρήσιμο και μη παρατηρήσιμο κομμάτι του $\boldsymbol{U}_2$ δεν διαφέρουν πολύ. Όμως όταν

πάμε στο μηχανισμό NMAR βλέπουμε ότι η γραμμή της διαμέσου του μη παρατηρήσιμου $\boldsymbol{U}_2$ ξεπερνά το τρίτο τεταρτημόριο του παρατηρήσιμου $\boldsymbol{U}_2$, οπότε υπάρχει σημαντική διαφορά μεταξύ τους. Πράγμα το οποίο ήταν αναμενόμενο διότι έχουμε μη τυχαία έλλειψη.

Ένα σημαντικό πρόβλημα στη μελέτη τέτοιου είδους δεδομένων, προκύπτει όταν σε πολλές από τις μεταβλητές υπάρχουν ελλιπείς τιμές. Στην περίπτωση που οι μεταβλητές μπορούν να ταξινομηθούν με τέτοιο τρόπο ώστε όταν η τιμή της μεταβλητή $\boldsymbol{U}_j$ λείπει, τότε οι τιμές των μεταβλητών $\boldsymbol{U}_k$ όπου $k > j$ θα είναι και αυτές ελλιπείς. Αυτή η έλλειψη στα δεδομένα ονομάζεται μονότονη (monotone). Αν μία τέτοια ταξινόμηση των δεδομένων δεν μπορεί να δημιουργηθεί, τότε η έλλειψη στα δεδομένα καλείται μη-μονότονη (non-monotone). Ένα παράδειγμα μονοτονικής και μη-μονοτονικής ταξινόμησης δίνεται από τον Πίνακα 4.2.1.

Monotone pattern

Observations					
	Obs	Missing	Missing	Missing	Missing
	Obs	Obs	Missing	Missing	Missing
	Obs	Obs	Obs	Missing	Missing
	Obs	Obs	Obs	Obs	Missing
	Obs	Obs	Obs	Obs	Obs

Non-Monotone pattern

Observations					
	Obs	Missing	Obs	Missing	Missing
	Obs	Obs	Missing	Obs	Missing
	Obs	Obs	Missing	Obs	Missing
	Obs	Missing	Obs	Obs	Missing
	Obs	Obs	Missing	Obs	Obs

Πίνακας 4.2.1: Παράδειγμα μίας μονοτονικής έλλειψης και μίας μη-μονοτονικής έλλειψης στα δεδομένα. Κάθε στήλη αντιστοιχεί σε μία μεταβλητή και κάθε γραμμή αντιστοιχεί σε μία παρατήρηση.

Συνήθως η μονοτονική έλλειψη εμφανίζεται συχνά σε διαχρονικές μελέτες (Longitudinal studies). Για παράδειγμα, σε μία διαχρονική μελέτη όπου τα δεδομένα συλλέγονται σε πολλά χρονικά σημεία, εάν ένας συμμετέχων εγκαταλείψει τη μελέτη μετά από ένα συγκεκριμένο χρονικό διάστημα, τότε όλα τα επόμενα δεδομένα για αυτόν το συμμετέχοντα θα λείπουν, δημιουργώντας έτσι μία μονότονη έλλειψη. Ενώ η μη-μονοτονική έλλειψη αποτελεί ένα πιο συχνό φαινόμενο σε μελέτες, διότι είναι πιο πιθανό οι συμμετέχοντες να έχασαν μία επίσκεψη σε μία διαχρονική μελέτη ή να μην απάντησαν σε κάποιες ερωτήσεις κάποιου ερωτηματολογίου, διότι δεν κατάλαβαν τις ερωτήσεις ή είχαν πάρει την απόφαση να μην απαντήσουν.

4.3 Απλές τεχνικές χειρισμού ελλιπών τιμών

Η πιο εύκολη μέθοδος διαχείρισης ελλιπών τιμών είναι η διαγραφή (Listwise deletion) των παρατηρήσεων που περιέχουν τουλάχιστον μία ελλιπή τιμή. Δηλαδή

συνεχίζουμε την ανάλυση μας μόνο με τις πλήρης παρατηρήσεις αγνοώντας έτσι την πληροφορία που περιέχεται σε παρατηρήσεις με ελλιπείς τιμές. Αυτή η μέθοδος μπορεί να είναι απλή στην εφαρμογή, αλλά μπορεί να οδηγήσει σε μεροληπτικές εκτιμήσεις εάν τα δεδομένα που λείπουν δεν είναι εντελώς τυχαία (MCAR). Μπορεί επίσης να οδηγήσει σε μεγάλη απώλεια πληροφορίας καθώς μειώνει το μέγεθος του δείγματος της ανάλυσης.

Συνεπώς, αν και η διαγραφή των παρατηρήσεων που περιέχουν ελλιπείς τιμές, αποτελεί μία αρκετά εύκολη μέθοδος που χρησιμοποιείται ως προεπιλογή από πολλά στατιστικά πακέτα (όπως Stata και R), δεν παύει να δημιουργεί πολλά προβλήματα στην ανάλυση μας. Τα προβλήματα αυτά είναι τα εξής:

1. Αφαιρώντας παρατηρήσεις που περιέχουν τουλάχιστον μία ελλιπή τιμή, μειώνεται σημαντικά η διαθέσιμη πληροφορία που έχουμε από τα δεδομένα. Για παράδειγμα η χρήση αυτής της μεθόδου σε μελέτες που έχουν πολλές μεταβλητές μπορεί να έχει ως αποτέλεσμα την παράλειψη πολλών παρατηρήσεων, διότι η πιθανότητα εμφάνισης ελλιπών τιμών είναι μεγάλη.

2. Σε πολλές μελέτες όπου το μέγεθος του δείγματος είναι αρκετά μικρό, η χρήση αυτής της μεθόδου θα προκαλέσει σοβαρή μείωση στο μέγεθος του δείγματος.

3. Εάν η έλλειψη τιμών δεν είναι MCAR, τότε η διαγραφή παρατηρήσεων μπορεί να οδηγήσει σε μεροληπτικές εκτιμήσεις.

Επομένως η διαγραφή παρατηρήσεων αποτελεί έναν πολύ απλοϊκό τρόπο, ο οποίος πολλές φορές προκαλεί πολλά προβλήματα στην ανάλυσή μας, ειδικά όταν ο μηχανισμός έλλειψης δεν είναι πλήρως τυχαίος (όπου σε πραγματικά δεδομένα ισχύει σχεδόν πάντα).

Μία άλλη σχετικά απλή μέθοδος, είναι η μέθοδος απλής αντικατάστασης (single imputation). Η μέθοδος αυτή χρησιμοποιεί κάποιο μέτρο θέσης για να αντικαταστήσει τις ελλιπείς τιμές. Κάποια από αυτά τα μέτρα είναι τα εξής

- **Δειγματικός μέσος** (mean)
- **Διάμεσος** (median)
- **Επικρατούσα τιμή** (mode)
- **Κοντινότερος γείτονας** (Nearest Neighbor)

Εφαρμόζοντας την παραπάνω απλή αντικατάσταση, υποθέτουμε ότι κάθε ελλιπής τιμή θα ισούται με το μέτρο θέσης που έχουμε επιλέξει. Αυτό μπορεί να οδηγήσει σε πιθανή υποεκτίμηση της διασποράς, ειδικά όταν οι ελλιπείς τιμές είναι πολλές σε κάθε μεταβλητή. Επιπλέον αυτή η αντικατάσταση δε λαμβάνει υπόψιν τυχόν συσχετίσεις που μπορεί να υπάρχουν μεταξύ μεταβλητών. Άρα καταλαβαίνουμε ότι η παραπάνω αντικατάσταση προκαλεί πληθώρα προβλημάτων και γενικά αποφεύγεται ως μέθοδος διαχείρισης ελλιπών τιμών.

Μία εξίσου γνωστή μέθοδος είναι η αντικατάσταση με τη βοήθεια μοντέλου παλινδρόμησης (regression imputation). Σε αυτή τη μέθοδο, χρησιμοποιείται ένα μοντέλο παλινδρόμησης για την εκτίμηση των ελλιπών τιμών έχοντας ως βάση τα

διαθέσιμα δεδομένα. Συγκεκριμένα, η μεταβλητή που περιέχει ελλιπείς τιμές αντιστοιχεί στη μεταβλητή απόκρισης, ενώ οι άλλες μεταβλητές στο σύνολο δεδομένων χρησιμοποιούνται ως επεξηγηματικές μεταβλητές. Στη συνέχεια, οι προβλεπόμενες τιμές από αυτή την παλινδρόμηση αντικαθιστούν τις ελλιπείς τιμές. Δηλαδή αν η μεταβλητή $\boldsymbol{Y}$ περιέχει ελλιπείς τιμές, τότε η i ελλιπής τιμή της μεταβλητής $\boldsymbol{Y}$ θα αντικατασταθεί από το $\hat{Y}_i$

$$\hat{Y}_i = \hat{\alpha}_0 + \hat{\alpha}_1 X_{1i} + \hat{\alpha}_2 X_{2i} + \cdots + \hat{\alpha}_p X_{pi},$$

όπου οι συντελεστές $\hat{\alpha}_j$ έχουν εκτιμηθεί με τη βοήθεια της μεθόδου ελαχίστων τετραγώνων.

Ένα πλεονέκτημα αυτής της μεθόδου είναι ότι μπορεί να παράγει καλύτερα αποτελέσματα σε σύγκριση με απλούστερες μεθόδους που είδαμε παραπάνω. Ωστόσο υπάρχει και εδώ πιθανή μεροληψία εάν ο μηχανισμός έλλειψης των δεδομένων δεν είναι MCAR. Κάποια σημαντικά μειονεκτήματα αυτής της μεθόδου είναι τα εξής:

1. Πολλές φορές υποθέτει γραμμική σχέση μεταξύ των μεταβλητών, η οποία μπορεί να μην ισχύει στην πραγματικότητα.

2. Αγνοεί την αβεβαιότητα που σχετίζεται με τις τιμές $\hat{Y}_i$, κάτι το οποίο μπορεί να προκαλέσει υποεκτίμηση των τυπικών σφαλμάτων και υπερεκτίμηση της στατιστικής σημαντικότητας στους συντελεστές β_j στο τελικό μοντέλο ενδιαφέροντος.

3. Μπορεί να εισάγει μεροληψία στην ανάλυση μας, εάν το μοντέλο παλινδρόμησης για την αντικατάσταση των ελλιπών τιμών είναι εσφαλμένα καθορισμένο ή η σχέση μεταξύ των μεταβλητών είναι μη γραμμική.

Επομένως οι παραπάνω απλές τεχνικές που αναφέραμε βλέπουμε ότι είναι αρκετά απλές και εύκολες στην εφαρμογή τους, όμως μπορούν να εισάγουν μεροληψία στην ανάλυση εάν ο μηχανισμός έλλειψης δεν είναι MCAR. Ως εκ τούτου, θα πρέπει να εξετάσουμε προσεκτικά την καταλληλότητα και τα μειονεκτήματα των μεθόδων απλής αντικατάστασης πριν επιλέξουμε να τις εφαρμόσουμε στην ανάλυση μας.

4.4 Πολλαπλή Αντικατάσταση

Στη μέθοδο πολλαπλής αντικατάστασης (multiple imputation) (Little and Rubin, 1987) σε αντίθεση με την απλή αντικατάσταση, οι ελλιπείς τιμές αντικαθίστανται με ένα σύνολο τιμών m, που συνήθως δημιουργείται με τη βοήθεια κάποιου μοντέλου (συνήθως με τη χρήση μοντέλων παλινδρόμησης). Έτσι σαν αποτέλεσμα θα έχουμε m διαφορετικά δείγματα ίδιου μεγέθους. Σε αυτά τα δείγματα οι πλήρης παρατηρήσεις παραμένουν οι ίδιες, ενώ οι ελλιπείς τιμές έχουν αντικατασταθεί από τις m τιμές που έχουν εκτιμηθεί.

Τα βήματα της μεθόδου είναι τα εξής:

1. Για κάθε ελλιπή τιμή, υπολογίζουμε m ξεχωριστές τιμές με χρήση κάποιου μοντέλου έχοντας ως βάση τις παρατηρούμενες μεταβλητές:

$$Y_{miss}^{(1)}, Y_{miss}^{(2)}, \ldots, Y_{miss}^{(m)} \sim f(\boldsymbol{Y}_{miss} | \boldsymbol{Y}_{obs})$$

όπου $Y_{miss}^{(j)}$ για $j = 1, \ldots, m$ είναι η j τιμή για μία ελλιπή τιμή της μεταβλητής $\boldsymbol{Y}$, $\boldsymbol{Y}_{obs}$ αντιστοιχεί στις παρατηρούμενες τιμές του δείγματος και $f(\boldsymbol{Y}_{miss}|\boldsymbol{Y}_{obs})$ είναι το αντίστοιχο μοντέλο.

2. Πραγματοποιούμε την ανάλυσή μας για κάθε δείγμα ξεχωριστά, λαμβάνοντας έτσι m σύνολα εκτιμήσεων για κάθε παράμετρο που μας ενδιαφέρει. Για παράδειγμα, εάν επιθυμούμε να εφαρμόσουμε γραμμική παλινδρόμηση με $\boldsymbol{Y}$ ως μεταβλητή απόκρισης και $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ ως επεξηγηματικές μεταβλητές, τότε το μοντέλο παλινδρόμησης θα εφαρμοστεί ξεχωριστά σε κάθε δείγμα:

$$\boldsymbol{Y}^{(j)} = \beta_0^{(j)} + \beta_1^{(j)}\boldsymbol{X}_1 + \cdots + \beta_p^{(j)}\boldsymbol{X}_p + \boldsymbol{\epsilon}^{(j)},$$

όπου $\boldsymbol{Y}^{(j)}$ είναι η μεταβλητή απόκρισης για το δείγμα j με $j = 1, \ldots, m$ και $\beta_0^{(j)}, \ldots, \beta_p^{(j)}$ είναι οι συντελεστές του μοντέλου στο δείγμα j.

3. Συνδυάζουμε τις m εκτιμήσεις που έχουμε για τις παραμέτρους $\boldsymbol{\beta}$ παίρνοντας τη μέση τιμή για κάθε συντελεστή β_k όπου $k = 0, \ldots, p$:

$$\bar{\beta}_k = \frac{1}{m}\sum_{j=1}^{m}\hat{\beta}_k^{(j)}, \ k = 0, \ldots, p,$$

όπου $\bar{\boldsymbol{\beta}} = \left(\bar{\beta}_0, \bar{\beta}_1, \ldots, \bar{\beta}_p\right)$ και με ολική διακύμανση που δίνεται από τον τύπο

$$\mathcal{T} = \bar{\Sigma} + \left(1 + \frac{1}{m}\right)B.$$

Το $\bar{\Sigma}$ έχει τη μορφή

$$\bar{\Sigma} = \frac{1}{m}\sum_{j=1}^{m}\Sigma_j,$$

όπου Σ_j είναι ο πίνακας διακύμανσης-συνδιακύμανσης του $\hat{\boldsymbol{\beta}}^{(j)}$ από το δείγμα j και το B δίνεται από τον τύπο

$$B = \frac{1}{m-1}\sum_{j=1}^{m}\left(\hat{\boldsymbol{\beta}}^{(j)} - \bar{\boldsymbol{\beta}}\right)\left(\hat{\boldsymbol{\beta}}^{(j)} - \bar{\boldsymbol{\beta}}\right)^T.$$

Συνεπώς η ολική διακύμανση $\mathcal{T}$ προέρχεται από (Little and Rubin, 1987):

- $\bar{\Sigma}$: Τη διακύμανση εντός ομάδων.
- B: Τη διακύμανση λόγω των ελλιπών τιμών (Διακύμανση μεταξύ των ομάδων).
- $\frac{B}{m}$: Τη μεταβλητότητα που πηγάζει από το γεγονός ότι δημιουργούμε πεπερασμένα στο πλήθος δείγματα m για την εκτίμηση του $\bar{\boldsymbol{\beta}}$.

4. Διεξάγουμε συμπεράσματα με βάση τις συνδυασμένες εκτιμήσεις και την ολική διακύμανση (πχ. υπολογίζουμε τυπικά σφάλματα και διαστήματα εμπιστοσύνης). Για την τιμή του m επιλέγεται μία σχετικά μικρή τιμή, καθώς μεγάλες τιμές του m θα κάνουν τη μέθοδο υπολογιστικά δύσκολη. Στη βιβλιογραφία το m πολλές φορές ισούται με $m = 3$, $m = 5$ ή $m = 10$.

Η μέθοδος πολλαπλής αντικατάστασης έχει και αυτή με τη σειρά της κάποια μειονεκτήματα τα οποία μπορούν να επηρεάσουν σημαντικά την ανάλυσή μας. Αποτελεί μία αρκετά δύσκολη και υπολογιστικά χρονοβόρα διαδικασία, ειδικά όταν ασχολούμαστε με πολύπλοκα μοντέλα και δεδομένα που έχουν ελλιπείς τιμές σε πολλές μεταβλητές. Επιπλέον, η λάθος επιλογή του μοντέλου $f(\boldsymbol{Y}_{miss}|\boldsymbol{Y}_{obs})$ μπορεί να οδηγήσει σε διαφορετικά αποτελέσματα, οπότε η επιλογή του μοντέλου αποτελεί ένα πολύ σημαντικό βήμα της μεθόδου. Παρόλα αυτά, η μέθοδος πολλαπλής αντικατάστασης χρησιμοποιείται ευρέως και συνιστάται στη βιβλιογραφία για το χειρισμό των ελλιπών τιμών.

Μία δημοφιλής προσέγγιση της μεθόδου πολλαπλής αντικατάστασης είναι η μέθοδος πολλαπλής αντικατάστασης με τη χρήση αλυσιδωτών εξισώσεων (multiple imputation through chained equations (MICE)). Η μέθοδος αυτή αναπτύχθηκε από τους Van Buuren and Oudshoorn (1999) ως μία εναλλακτική προσέγγιση της κλασικής μεθόδου πολλαπλής αντικατάστασης. Είναι μία ιδιαίτερα χρήσιμη μέθοδος όταν αντιμετωπίζουμε δεδομένα με μη-μονότονη έλλειψη και μπορεί να χειριστεί τόσο συνεχείς όσο και κατηγορηματικές μεταβλητές.

Τα βήματα της μεθόδου MICE είναι τα εξής (Van Buuren and Oudshoorn, 1999) :

Μέθοδος MICE

1. Αφαιρούμε όλες τις παρατηρήσεις που δεν περιέχουν παρατηρήσιμες τιμές σε κάθε μεταβλητή.
2. Συμπληρώνουμε τις ελλιπείς τιμές κάθε μεταβλητής με εύλογες αρχικές τιμές (π.χ μέσες τιμές, διάμεσοι κ.λπ).
3. Πραγματοποιούμε απλή αντικατάσταση με χρήση κάποιου μοντέλου σε κάθε μεταβλητή.
4. Αντικαθιστούμε τις αρχικές τιμές με τις τιμές που βρήκαμε από το βήμα 3 και επαναλαμβάνουμε το βήμα 3 k φορές.
5. Επαναλαμβάνουμε τα βήματα 1,2,3 και 4 m φορές. Έτσι σαν αποτέλεσμα θα έχουμε m διαφορετικά δείγματα.

Η αντικατάσταση των ελλιπών τιμών στη μέθοδο MICE (βήματα 3 και 4), μπορεί να πραγματοποιηθεί με τη χρήση κάποιου μοντέλου όπως είδαμε στην Ενότητα 4.3 ή με τη μέθοδο Predictive Mean Matching (PMM) όπου είναι και η προεπιλογή για συνεχείς μεταβλητές στη γλώσσα προγραμματισμού R. Στη μέθοδο PMM ξεκινάμε υπολογίζοντας την προβλεπόμενη τιμή για κάθε ελλιπή τιμή χρησιμοποιώντας κάποιο μοντέλο. Έπειτα για κάθε ελλιπή τιμή, δημιουργούμε ένα σύνολο από υποψήφιους δωρητές (συνήθως 3, 5 ή 10) από τις πλήρεις περιπτώσεις που κατέχουν προβλεπόμενες τιμές οι οποίες βρίσκονται κοντά στη προβλεπόμενη τιμή της ελλιπής τιμής. Από τους υποψήφιους δωρητές, διαλέγουμε τυχαία έναν και η παρατηρούμενη τιμή αυτού αντικαθιστά την ελλιπή τιμή. Έτσι με αυτό τον τρόπο, η αντικατάσταση των ελλιπών τιμών γίνεται με τη βοήθεια των πλήρων τιμών της κάθε μεταβλητής.

4.5 Μπεϋζιανή προσέγγιση διαχείρισης ελλιπών τιμών

Έχοντας μέχρι τώρα ορίσει κάποιες κλασικές μεθόδους για τη διαχείριση ελλιπών τιμών, είμαστε στη θέση να μιλήσουμε για το πως η Μπεϋζιανή στατιστική επιλύει το πρόβλημα των ελλιπών τιμών στα δεδομένα. Όπως έχουμε αναφέρει και στο Κεφάλαιο 1, στην Μπεϋζιανή συμπερασματολογία οι παράμετροι αντιμετωπίζονται ως τυχαίες μεταβλητές. Συνεπώς, θα ακολουθήσουμε το ίδιο σκεπτικό για την κατασκευή της Μπεϋζιανής μεθόδου διαχείρισης ελλιπών τιμών Data Augmentation (Little and Rubin, 1987).

Έστω ότι οι ελλιπείς τιμές βρίσκονται στη μεταβλητή απόκρισης $\boldsymbol{Y}$, η κατανομή της οποίας εξαρτάται από την παράμετρο $\boldsymbol{\theta}$ και ο μηχανισμός έλλειψης είναι MAR. Δηλαδή έχουμε ότι $\boldsymbol{Y} = \{\boldsymbol{Y}_{miss}, \boldsymbol{Y}_{obs}\}$, όπου $\boldsymbol{Y}_{obs}$ αποτελεί το πλήρως παρατηρήσιμο κομμάτι της μεταβλητής $\boldsymbol{Y}$, ενώ $\boldsymbol{Y}_{miss}$ το μη παρατηρήσιμο κομμάτι. Σε αυτή την περίπτωση ορίζουμε το $\boldsymbol{Y}_{miss}$ ως μία άγνωστη παράμετρο η οποία θα έχει και αυτή ύστερη κατανομή η οποία θα ισούται με την εκ των υστέρων προβλεπτική κατανομή όπου δίνεται από τον εξής τύπο

$$\begin{aligned} p(\boldsymbol{Y}_{miss}|\boldsymbol{Y}_{obs}) &= \int p(\boldsymbol{Y}_{miss}, \boldsymbol{\theta}|\boldsymbol{Y}_{obs})\, d\boldsymbol{\theta} \\ &= \int p(\boldsymbol{Y}_{miss}|\boldsymbol{Y}_{obs}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})\, d\boldsymbol{\theta} \\ &= \int p(\boldsymbol{Y}_{miss}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})\, d\boldsymbol{\theta}, \end{aligned} \tag{4.5.1}$$

όπου $p(\boldsymbol{Y}_{miss}|\boldsymbol{\theta})$ είναι η δειγματοληπτική κατανομή των $\boldsymbol{Y}_{miss}$ και $p(\boldsymbol{\theta}|\boldsymbol{Y}_{obs})$ είναι η εκ των υστέρων κατανομή των παραμέτρων $\boldsymbol{\theta}$. Η ισότητα $p(\boldsymbol{Y}_{miss}|\boldsymbol{Y}_{obs}, \boldsymbol{\theta}) = p(\boldsymbol{Y}_{miss}|\boldsymbol{\theta})$ ισχύει διότι η πληροφορία από το παρατηρούμενο κομμάτι του $\boldsymbol{Y}$ είναι ενσωματωμένη στην παράμετρο $\boldsymbol{\theta}$. Ως εκ τούτου, στην Μπεϋζιανή προσέγγιση παρατηρούμε ότι η πολλαπλή αντικατάσταση λαμβάνει υπόψιν την αβεβαιότητα που έχουμε για την $\boldsymbol{Y}_{miss}$ δοθέντος του $\boldsymbol{\theta}$, αλλά και την αβεβαιότητα που έχουμε για την εκτίμηση του $\boldsymbol{\theta}$.

Η εκ των υστέρων κατανομή του $\boldsymbol{\theta}$ για τις πλήρες παρατηρήσεις ξέρουμε ότι θα είναι ανάλογη με την πιθανοφάνεια και την εκ των προτέρων κατανομή του $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}) \propto f(\boldsymbol{Y}_{obs}|\boldsymbol{\theta}) p(\boldsymbol{\theta}). \tag{4.5.2}$$

Αντιμετωπίζοντας το $\boldsymbol{Y}_{miss}$ ως άγνωστη παράμετρο, η από κοινού εκ των υστέρων κατανομή των $(\boldsymbol{Y}_{miss}, \boldsymbol{\theta})$ είναι

$$\begin{aligned} p\left(\boldsymbol{Y}_{miss}, \boldsymbol{\theta}|\boldsymbol{Y}_{obs}\right) &= p(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{miss}) p(\boldsymbol{Y}_{miss}|\boldsymbol{Y}_{obs}) \\ &= p(\boldsymbol{Y}_{miss}|\boldsymbol{Y}_{obs}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}) \\ &= p(\boldsymbol{Y}_{miss}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}) \end{aligned}$$

όπου με τη βοήθεια της 4.5.2 έχουμε ότι

$$p\left(\boldsymbol{Y}_{miss}, \boldsymbol{\theta}|\boldsymbol{Y}_{obs}\right) \propto p(\boldsymbol{Y}_{miss}|\boldsymbol{\theta}) f(\boldsymbol{Y}_{obs}|\boldsymbol{\theta}) p(\boldsymbol{\theta}). \tag{4.5.3}$$

Επομένως η πολλαπλή αντικατάσταση υπό την Μπεϋζιανή σκέψη ξεκινάει παίρνοντας τυχαίο δείγμα από την ύστερη κατανομή της $\boldsymbol{\theta}$ $(p(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}))$ και έπειτα παίρνοντας τυχαίο δείγμα από την εκ των υστέρων δεσμευμένη προβλεπτική κατανομή της παραμέτρου

$\boldsymbol{Y}_{miss}$ ($p(\boldsymbol{Y}_{miss}|\boldsymbol{\theta})$). Κάνοντας αυτά τα δύο βήματα N φορές θα καταλήξουμε να έχουμε γεννήσει ένα δείγμα μεγέθους N (Little and Rubin, 1987).

Οι Tanner and Wong (1987), παρατήρησαν ότι μπορούμε να γεννήσουμε ψευδοτυχαίες τιμές από την από κοινού εκ των υστέρων κατανομή της $\boldsymbol{Y}_{miss}$ και $\boldsymbol{\theta}$ με τη βοήθεια του δειγματολήπτη Gibbs (Ενότητα 1.6.2). Συγκεκριμένα οι Tanner and Wong (1987) ξεκινούν από την εκ των υστέρων δεσμευμένη προβλεπτική κατανομή της παραμέτρου $\boldsymbol{Y}_{miss}$

$$\boldsymbol{Y}_{miss}^{(j+1)} \sim p\left(\boldsymbol{Y}_{miss}|\boldsymbol{\theta}^{(j)}, \boldsymbol{Y}_{obs}\right) = p\left(\boldsymbol{Y}_{miss}|\boldsymbol{\theta}^{(j)}\right)$$

και συνεχίζουν με την εκ των υστέρων κατανομή της $\boldsymbol{\theta}$

$$\boldsymbol{\theta}^{(j+1)} \sim p\left(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}, \boldsymbol{Y}_{miss}^{(j+1)}\right).$$

Μέχρι τώρα βασική προϋπόθεση ήταν ότι ο μηχανισμός έλλειψης είναι τυχαίος (MAR). Στην περίπτωση όμως που ο μηχανισμός έλλειψης τιμών είναι μη αμελητέος (non-ignorable) και έχουμε μηχανισμό έλλειψης NMAR, η εκ των υστέρων προβλεπτική κατανομή θα δεσμεύεται και από το δείκτη ελλιπών τιμών $\boldsymbol{R}$, οπότε θα έχουμε

$$\begin{aligned} p(\boldsymbol{Y}_{miss}|\boldsymbol{Y}_{obs}, \boldsymbol{R}) &= \int\int p(\boldsymbol{Y}_{miss}, \boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{Y}_{obs}, \boldsymbol{R})\, d\boldsymbol{\theta}\, d\boldsymbol{\phi} \\ &= \int\int p(\boldsymbol{Y}_{miss}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{Y}_{obs}, \boldsymbol{R}) p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{Y}_{obs}, \boldsymbol{R})\, d\boldsymbol{\theta}\, d\boldsymbol{\phi}. \end{aligned} \tag{4.5.4}$$

Στην περίπτωση που έχουμε να αντιμετωπίσουμε ελλιπείς τιμές στις επεξηγηματικές μεταβλητές $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, θα χρειαστεί να ορίσουμε μοντέλο με παραμέτρους $\boldsymbol{\theta}_{\boldsymbol{X}}$ για τις ελλιπείς τιμές και ας βρισκόμαστε σε μηχανισμό έλλειψης τιμών MAR ή MCAR. Επομένως η ύστερη προβλεπτική κατανομή των q ελλιπών τιμών της i γραμμής $\boldsymbol{X}_{i,miss} = (X_{i,miss,1}, \ldots, X_{i,miss,q})^T$ είναι η εξής:

$$p(\boldsymbol{X}_{i,miss}|\boldsymbol{Y}, \boldsymbol{X}_{i,obs}) = \int\int p(\boldsymbol{X}_{i,miss}, \boldsymbol{\theta}, \boldsymbol{\theta}_{\boldsymbol{X}}|\boldsymbol{Y}, \boldsymbol{X}_{i,obs})\, d\boldsymbol{\theta}\, d\boldsymbol{\theta}_{\boldsymbol{X}} \tag{4.5.5}$$

και η από κοινού εκ των υστέρων κατανομή είναι

$$\begin{aligned} p(\boldsymbol{X}_{i,miss}, \boldsymbol{\theta}, \boldsymbol{\theta}_{\boldsymbol{X}}|\boldsymbol{Y}, \boldsymbol{X}_{i,obs}) &= p(\boldsymbol{X}_{i,miss}, \boldsymbol{\theta}, \boldsymbol{\theta}_{\boldsymbol{X}}|Y_i, \boldsymbol{Y}_{-i}, \boldsymbol{X}_{i,obs}) \\ &\propto p(Y_i|\boldsymbol{X}_{i,miss}, \boldsymbol{X}_{i,obs}, \boldsymbol{\theta}) p(\boldsymbol{X}_{i,miss}|\boldsymbol{X}_{i,obs}, \boldsymbol{\theta}_{\boldsymbol{X}}) p(\boldsymbol{\theta}|\boldsymbol{Y}_{-i}, \boldsymbol{X}_{i,obs}) p(\boldsymbol{\theta}_{\boldsymbol{X}}|\boldsymbol{X}_{i,obs}), \end{aligned} \tag{4.5.6}$$

όπου

$$\boldsymbol{X}_{i,obs} = (X_{i,obs,1}, \ldots, X_{i,obs,u})^T \quad \text{με } u + q = p$$

και

$$\boldsymbol{Y}_{-i} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)^T,\ \boldsymbol{\theta}_{\boldsymbol{X}} = (\boldsymbol{\theta}_{\boldsymbol{X}_1}, \ldots, \boldsymbol{\theta}_{\boldsymbol{X}_u})^T.$$

Για να μπορέσουμε να χρησιμοποιήσουμε των δειγματολήπτη Gibbs, θα χρειαστεί να γράψουμε την πιθανοφάνεια $p(\boldsymbol{X}_{i,miss}|\boldsymbol{X}_{i,obs}, \boldsymbol{\theta}_{\boldsymbol{X}})$ ως γινόμενο μονοδιάστατων κατανομών (Ibrahim et al., 2002; Erler, 2019), οπότε έχουμε

$$\begin{aligned} &p(\boldsymbol{X}_{i,miss}|\boldsymbol{X}_{i,obs}, \boldsymbol{\theta}_{\boldsymbol{X}}) = \\ &\quad p(X_{i,miss,1}|\boldsymbol{X}_{i,obs}, \boldsymbol{\theta}_{\boldsymbol{X}_1}) \prod_{u=2}^{p} p(X_{i,miss,u}|\boldsymbol{X}_{i,obs}, \boldsymbol{\theta}_{\boldsymbol{X}_u}, X_{i,miss,1}, \ldots, X_{i,miss,u-1}). \end{aligned} \tag{4.5.7}$$

Συνεπώς, έχοντας γράψει την πιθανοφάνεια των $\boldsymbol{X}_{i,miss}$ με αυτήν τη μορφή και έχοντας προσδιορίσει τις εκ των υστέρων κατανομές $p(\boldsymbol{\theta}|\boldsymbol{Y}_{-i}, \boldsymbol{X}_{i,obs})$ και $p(\boldsymbol{\theta}_{\boldsymbol{X}}|\boldsymbol{X}_{i,obs})$, μας δίνεται η δυνατότητα να χρησιμοποιήσουμε το δειγματολήπτη Gibbs για να προσομοιώσουμε τιμές από την από κοινού εκ των υστέρων κατανομή των ελλιπών τιμών και των παραμέτρων.

5 Εφαρμογές σε προσομοιωμένα δεδομένα με ελλιπείς τιμές

5.1 Εισαγωγή

Στο παρόν κεφάλαιο θα ξεκινήσουμε εφαρμόζοντας τις μεθόδους SSVS, Kuo & Mallick και GVS σε προσομοιωμένα δεδομένα για τα οποία γνωρίζουμε το πραγματικό μοντέλο. Έπειτα θα δημιουργήσουμε ελλιπείς τιμές στα δεδομένα αυτά (amputation) όπου θα προέρχονται από τους μηχανισμούς MCAR, MAR και MCAR. Στη συνέχεια μέσω της Μπεϋζιανής προσέγγισης θα εκτιμήσουμε αυτές τις τιμές και εφαρμόζοντας πάλι τις μεθόδους επιλογής μεταβλητών SSVS, Kuo & Mallick, GVS θα συγκρίνουμε τα αποτελέσματα ως προς την επιλογή των μεταβλητών του μοντέλου και θα διακρίνουμε τυχόν αλλαγές στα αποτελέσματα καθώς αλλάζουμε το μηχανισμό έλλειψης τιμών. Στο παρόν Κεφάλαιο αξίζει να σημειωθεί πως στην επιλογή των εκ των προτέρων κατανομών δεν λάβαμε υπόψιν το παράδοξο Lindley-Bartlett.

5.2 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε πλήρη δεδομένα

Υποθέτουμε ότι τα δεδομένα έχουν μέγεθος $n = 80$ και περιέχουν $p = 12$ επεξηγηματικές μεταβλητές, οι οποίες προέρχονται από μία πολυμεταβλητή Κανονική κατανομή και η μεταβλητή απόκρισης $\boldsymbol{Y}$ σχετίζεται γραμμικά με κάποιες από τις p επεξηγηματικές μεταβλητές.

Η πολυμεταβλητή Κανονική κατανομή που θα χρησιμοποιήσουμε για να προσομοιώσουμε τις τιμές των 12 επεξηγηματικών μεταβλητών έχει μέση τιμή $\boldsymbol{\mu} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ και πίνακα συνδιακύμανσης που δίνεται από τον Πίνακα 5.2.1:

	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_4$	$\boldsymbol{X}_5$	$\boldsymbol{X}_6$	$\boldsymbol{X}_7$	$\boldsymbol{X}_8$	$\boldsymbol{X}_9$	$\boldsymbol{X}_{10}$	$\boldsymbol{X}_{11}$	$\boldsymbol{X}_{12}$
$\boldsymbol{X}_1$	1.00	0.24	0.15	0.19	0.35	0.16	0.37	0.07	0.27	0.11	0.23	0.25
$\boldsymbol{X}_2$	0.24	1.00	0.12	0.24	0.18	0.27	0.11	0.02	0.02	0.26	0.24	0.28
$\boldsymbol{X}_3$	0.15	0.12	1.00	0.39	0.15	0.32	0.17	0.00	0.20	0.16	0.07	0.41
$\boldsymbol{X}_4$	0.19	0.24	0.39	1.00	0.43	0.21	0.28	0.20	0.10	0.16	0.30	0.26
$\boldsymbol{X}_5$	0.35	0.18	0.15	0.43	1.00	0.05	0.18	0.26	0.12	0.20	0.30	0.28
$\boldsymbol{X}_6$	0.16	0.27	0.32	0.21	0.05	1.00	0.42	0.14	0.00	0.01	0.17	0.17
$\boldsymbol{X}_7$	0.37	0.11	0.17	0.28	0.18	0.42	1.00	0.09	0.41	0.41	0.08	0.21
$\boldsymbol{X}_8$	0.07	0.02	0.00	0.20	0.26	0.14	0.09	1.00	0.00	0.17	0.14	0.36
$\boldsymbol{X}_9$	0.27	0.02	0.20	0.10	0.12	0.00	0.41	0.00	1.00	0.24	0.35	0.05
$\boldsymbol{X}_{10}$	0.11	0.26	0.16	0.16	0.20	0.01	0.41	0.17	0.24	1.00	0.02	0.01
$\boldsymbol{X}_{11}$	0.23	0.24	0.07	0.30	0.30	0.17	0.08	0.14	0.35	0.02	1.00	0.24
$\boldsymbol{X}_{12}$	0.25	0.28	0.41	0.26	0.28	0.17	0.21	0.36	0.05	0.01	0.24	1.00

Πίνακας 5.2.1: Πίνακας συνδιακύμανσης της κανονικής κατανομής από όπου έχουμε προσομοιώσει τις τιμές των επεξηγηματικών μεταβλητών.

Τις τιμές της μεταβλητής απόκρισης τις προσομοιώνουμε από το παρακάτω μοντέλο:

$$\boldsymbol{Y} = 2 + 2.5 \cdot \boldsymbol{X}_3 - 2 \cdot \boldsymbol{X}_5 + 1 \cdot \boldsymbol{X}_9 + 0.5 \cdot \boldsymbol{X}_{11} + \boldsymbol{\epsilon},$$

όπου το ϵ ακολουθεί την Κανονική κατανομή $N(0,1)$. Γνωρίζοντας το πραγματικό μοντέλο είμαστε σε θέση να παρατηρήσουμε πόσο καλά οι μέθοδοι επιλογής μεταβλητών μπορούν να το εντοπίσουν. Συνοπτικά λοιπόν, θα ξεκινήσουμε προσομοιώνοντας τα δεδομένα μας με τη βοήθεια της R και χρησιμοποιώντας το πρόγραμμα WinBUGS θα εφαρμόσουμε τις Μπεϋζιανές μεθόδους επιλογής μεταβλητών SSVS, Kuo & Mallick, GVS που αναφέραμε στο Κεφάλαιο 3 και την πλήρη εξερεύνηση (full enumaration) με χρήση του κριτηρίου BIC (Βλ. 2.4.2).

Κώδικας προσομοίωσης δεδομένων στην R

```
#We load the MASS library into R allowing us to simulate the
#covariates using a multivariate normal distribution.
library(MASS)
#We set a unique seed so that we can replicate the same simulation
#and get the same results
set.seed(100001)
#We choose the sample size to be equal to 80.
samplesize <- 80
#This represent the mean vector of the
#multivariate normal distribution
meanvector <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
#and the variance covariance matrix which is equal to the identity
    matrix.
Variance_covariance <-  Πίνακας 5.2.1
# We sample an nxp matrix, where each column represents a covariate.
X <- mvrnorm(n = samplesize,
             mu = meanvector,
             Sigma = Variance_covariance)
X <- as.data.frame(X)
#We name each column of the data frame to the
#corresponding variable Xj, j=1,...,p
colnames(X) <- paste0('X', 1:(ncol(X)))
#We simulate n points from the univariate normal distribution with
#mu=0 and sd=1. This is our error.
Error <- rnorm(samplesize,mean=0, sd=1)
#We pick our beta's for the real model
b0 <- 2
b <- rep(0,ncol(X))
b[3] <- 2.5
b[5] <- -2
b[9] <- 1
b[11] <- 0.5
#and we simulate Y using the following linear model
#(which will be our real model).
Y <- b0 + b[3]*X$X3 + b[5]*X$X5 + b[9]*X$X9 + b[11]*X$X11 + Error
#Lastly, we merge the vector Y with the dataframe X to get
#our complete simulated dataset.
FullDataset <- cbind(Y,X)
```

Έχοντας εφαρμόσει τον παραπάνω κώδικα, είμαστε σε θέση με τη βοήθεια της R και

του WinBUGS να εφαρμόσουμε τις μεθόδους SSVS, Kuo & Mallick και GVS. Για να γίνει αυτή η εφαρμογή πιο εύκολη, θα χρησιμοποιήσουμε τη βιβλιοθήκη της R ονόματι R2WinBUGS. Με αυτήν τη βιβλιοθήκη μπορούμε μέσω της R να ενεργοποιήσουμε το πρόγραμμα WinBUGS, να εφαρμόσουμε τις μεθόδους και έπειτα να αποθηκεύσουμε τα αποτελέσματα στην R. Η μόνη εντολή που θα μας είναι χρήσιμη από τη βιβλιοθήκη R2WinBUGS είναι η εντολή bugs που είναι η εξής

R2WinBUGS

```
bugs(data, inits, parameters.to.save, model.file="model.bug",
    n.chains=..., n.iter=..., n.burnin=...,
    n.thin=..., n.sims = ..., debug=FALSE,
    bugs.directory="c:/Program Files/WinBUGS14/",
    program=c("WinBUGS", "OpenBUGS", "winbugs", "openbugs"))
```

όπου συγκεκριμένα οι παράμετροι ενδιαφέροντος αυτής της εντολής είναι οι εξής

data: Μία λίστα (List) που περιέχει τα δεδομένα μας.

inits: Μία λίστα που θα περιέχει τις αρχικές τιμές για το μοντέλο που εφαρμόζουμε στο WinBUGS. Αν inits= NULL, τότε οι αρχικές τιμές προσδιορίζονται από το WinBUGS.

parameters.to.save: Ένα διάνυσμα που περιέχει τις παραμέτρους (σε μορφή χαρακτήρων) που θέλουμε να αποθηκεύσουμε τα αποτελέσματα.

model.file: Περιέχει το αρχείο το οποίο είναι γραμμένο σε κώδικα WinBUGS. Το αρχείο μπορεί να είναι είτε .bugs ή αρχείο .txt. Για να μπορέσει να ανιχνεύσει το αρχείο η εντολή bugs, χρειάζεται να δώσουμε την ακριβή διαδρομή (path) του αρχείου.

n.chains: Ο αριθμός των Μαρκοβιανών αλυσίδων. Προκαθορισμένα η εντολή έχει n.chains=3.

n.iter: Αριθμός των συνολικών επαναλήψεων ανά αλυσίδα (μαζί με το burn-in period). Προκαθορισμένα η εντολή έχει n.iter=2000.

n.burnin: Ο αριθμός επαναλήψεων που αφαιρούμε στην αρχή του αλγορίθμου. Προκαθορισμένη τιμή στην εντολή είναι n.burnin = n.iter/2, όπου αυτό σημαίνει ότι αφαιρούμε τις μισές επαναλήψεις από την προσομοίωσή μας.

n.thin: Παράμετρος λέπτυνσης (thinning) του αλγορίθμου. Προκαθορισμένη τιμή στην εντολή είναι n.thin=max(1, floor(n.chains * (n.iter-n.burnin) / 1000)), που σημαίνει ότι θα έχουμε n.thin > 1 όταν έχουμε τουλάχιστον 2000 προσομοιώσεις.

n.sims: Ο αριθμός των επαναλήψεων που κρατάμε μετά από τη λέπτυνση και burn-in period.

debug: Αν debug=FALSE τότε το πρόγραμμα WinBUGS κλείνει αυτόματα με το που τελειώσει να τρέχει ο κώδικας, ενώ αν debug=TRUE το παράθυρο του WinBUGS παραμένει ανοιχτό.

bugs.directory: Περιέχει το path του εκτελέσιμου αρχείου (executable file) του WinBUGS.

program: Η έκδοση του προγράμματος BUGS που χρησιμοποιούμε.

5.2.1 Εφαρμογή της μεθόδου SSVS

Ξεκινάμε γράφοντας τον κώδικα του WinBUGS σε ένα αρχείο .txt (SSVS.txt).

Μέθοδος SSVS στο WinBUGS. Ntzoufras (2002)

```
#Linear regression variable selection using SSVS
Model SSVS; {
#We standardise the X's and the coefficients
for(j in 1:p){
b[j] <- beta[j]/sd(x[,j])
for(i in 1:n){
z[i,j] <- (x[i,j]-mean(x[,j]))/sd(x[,j])
     }
standardized_means[j] <- b[j]*mean(x[,j])
  }
b0 <- beta0 - sum(standardized_means[])
#Linear regression model
for(i in 1:n){
#Normal distribution for the Y's
y[i] ~ dnorm (mu[i],tau)
mu[i] <- beta0 + beta[1]* z[i,1] +beta[2]* z[i,2]+ beta[3]* z[i,3]
    + beta[4]* z[i,4]+ beta[5]* z[i,5] + beta[6]* z[i,6]+ beta[7]*
    z[i,7] + beta[8]* z[i,8] + beta[9]* z[i,9] + beta[10]* z[i,10]+
    beta[11]* z[i,11] + beta[12]* z[i,12]
  }
#Normal diffuse prior for the constant of the model
beta0 ~ dnorm(0, 0.0001)
#Mixture of normals as a prior
for(j in 1:p){
beta[j] ~ dnorm(muprior[j],tauprior[j])
muprior[j] <- 0
tauprior[j] <- gamma[j]*0.001 + (1-gamma[j])*10
#Priors for variable indicator
gamma[j] ~ dbern(0.5)
}
#Model code
for(j in 1:p){
modelindicators[j] <- gamma[j]*pow(2,j-1)
}
modelind <- 1 + sum(modelindicators[])
# Vector with the model indicators
for(j in 1: models){
permodelselect[j] <- equals(modelind,j)
}
```

```
#Diffuse gamma prior for tau
tau ~ dgamma(1.0E-3,1.0E-3)
# normal errors
sigma <- sqrt(1/tau)
}
```

Οι εκ των προτέρων κατανομές που έχουμε προσδιορίσει για τη σταθερά του μοντέλου β_0 και την ακρίβεια τ είναι οι εξής

$$\beta_0 \sim N(0, 0.0001), \quad \tau \sim \text{Gamma}(0.001, 0.001),$$

όπου βλέπουμε ότι αποτελούν ασαφής εκ των προτέρων κατανομές. Για τους συντελεστές $\boldsymbol{\beta}$ του μοντέλου η εκ των προτέρων κατανομή θα έχει τη μορφή της (3.3.1)

$$\beta_j|\gamma_j \sim (1-\gamma_j)N(0, \tau_0 = 10) + \gamma_j N(0, \tau_1 = 0.001), \ j = 1, \ldots, 12,$$

οπότε η σταθερά του γραμμικού μοντέλου θα υπάρχει σε όλα τα μοντέλα (δηλαδή $\gamma_0 = 1$). Η επιλογή των τ_0 και τ_1 έγινε λαμβάνοντας υπόψη τους George and McCulloch (1997) όπου πρότειναν ότι $\tau_0/\tau_1 \leq 10000$ έτσι ώστε να μην προκύψουν υπολογιστικά προβλήματα. Τέλος για κάθε γ_j επιλέγουμε ως εκ των προτέρων κατανομή μία κατανομή Bernoulli με πιθανότητα επιτυχίας ίση με 0.5.

Θα χρησιμοποιήσουμε την εντολή bugs για να εκτελέσουμε τον παραπάνω κώδικα του WinBUGS και θα αποθηκεύσουμε τις προσομοιωμένες τιμές των ύστερων κατανομών των παραμέτρων beta0, beta, gamma και permodelselect όπου

beta0, beta: Κανονικοποιημένοι συντελεστές και σταθερά του μοντέλου.

gamma: Δείκτες ύπαρξης των επεξηγηματικών μεταβλητών στο μοντέλο.

sigma: Τυπική απόκλιση των σφαλμάτων, $\sigma = \tau^{-1}$.

permodelselect: Πιθανότητες του κάθε μοντέλου.

Κώδικας εκτέλεσης της μεθόδου SSVS στο WinBUGS μέσω της εντολής bugs στην R

```
#We list our initial values and the dataset
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
   tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=FullDataset$Y, x
   =as.matrix(FullDataset[,2:13]),p=12, n=80, models=4096)
#We use the bugs function to execute our bugs code
FUllmodelssvs <- bugs(inits=inits,data=data,model.file =
   "ssvs.txt",parameters = c("beta0","beta", "gamma",
   "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
   = 12000,n.thin = 5,n.sims = 10000,
bugs.directory = "c:/Program Files/WinBUGS14/",debug=FALSE)
```

```
FUllmodelssvs
#Here we have the summary of the model parameters
FUllmodelparametersssvs <- FUllmodelssvs$summary[1:26,]
FUllmodelparametersssvs
FUllmodelssssvs <- FUllmodelssvs$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
head(FUllmodelssssvs[order(FUllmodelssssvs[,2], decreasing=TRUE),
   ],n=10)
```

Αρχικές τιμές για τους συντελεστές $\boldsymbol{\beta}$ και το σταθερό όρο $\boldsymbol{\beta}_0$ του μοντέλου έχουμε την τιμή μηδέν, ενώ για το διάνυσμα δείκτη $\boldsymbol{\gamma}$ την τιμή ένα για κάθε $j = 1, \ldots, 12$. Στη λίστα που περιέχει τα δεδομένα έχουμε τη μεταβλητή απόκρισης Y, τον πίνακα που περιέχει τις επεξηγηματικές μεταβλητές, το μέγεθος του δείγματος και τον αριθμό όλων των δυνατών μοντέλων ($2^{12} = 4096$). Τρέχουμε τη μέθοδο SSVS με μία αλυσίδα και συνολικά 12000 επαναλήψεις όπου αφαιρούμε τις πρώτες 2000 επαναλήψεις και έχουμε λέπτυνση n.thin=5.

Οι περιγραφικοί δείκτες των προσομοιωμένων τιμών των εκ των υστέρων κατανομών των παραμέτρων του μοντέλου και του διανύσματος δείκτη $\boldsymbol{\gamma}$ της μεθόδου SSVS που μας δίνει το WinBUGS δίνονται στον Πίνακα 5.2.1.1.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.409	0.1217	0.003237	2.166	2.409	2.647
beta[1]	−0.2409	0.1281	0.002895	−0.4861	−0.24	0.01287
beta[2]	−0.05399	0.1364	0.002842	−0.3216	−0.05412	0.2111
beta[3]	2.732	0.1618	0.003706	2.4	2.727	3.051
beta[4]	−0.06293	0.1319	0.002742	−0.3194	−0.06268	0.2119
beta[5]	−1.775	0.1393	0.002912	−2.049	−1.775	−1.499
beta[6]	0.02894	0.152	0.003752	−0.266	0.02678	0.3319
beta[7]	−0.03451	0.1693	0.004135	−0.3754	−0.03905	0.307
beta[8]	−0.1042	0.1283	0.00249	−0.3474	−0.1063	0.1489
beta[9]	1.035	0.1849	0.003616	0.6716	1.042	1.379
beta[10]	−0.04897	0.1532	0.003213	−0.3613	−0.04583	0.241
beta[11]	0.5275	0.1404	0.002844	0.2504	0.5281	0.805
beta[12]	0.05626	0.1507	0.003262	−0.2404	0.05388	0.3439
sigma	1.059	0.08926	0.002024	0.907	1.054	1.247
Δείκτες γ						
gamma[1]	0.015	0.1216	0.002459	0.0	0.0	0.0
gamma[2]	0.012	0.1089	0.002198	0.0	0.0	0.0
gamma[3]	1.0	0.0	$2.236E-12$	1.0	1.0	1.0
gamma[4]	0.0115	0.1066	0.002887	0.0	0.0	0.0
gamma[5]	1.0	0.0	$2.236E-12$	1.0	1.0	1.0
gamma[6]	0.0145	0.1195	0.002772	0.0	0.0	0.0
gamma[7]	0.009	0.09444	0.002072	0.0	0.0	0.0
gamma[8]	0.014	0.1175	0.002946	0.0	0.0	0.0
gamma[9]	0.63	0.4828	0.009579	0.0	1.0	1.0
gamma[10]	0.0115	0.1066	0.002404	0.0	0.0	0.0
gamma[11]	0.0535	0.225	0.005244	0.0	0.0	1.0
gamma[12]	0.014	0.1175	0.00269	0.0	0.0	0.0

Πίνακας 5.2.1.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ της μεθόδου SSVS στα πλήρη δεδομένα.

Ο Πίνακας 5.2.1.1 περιλαμβάνει επτά στήλες όπου η πρώτη στήλη αντιστοιχεί στις παραμέτρους του μοντέλου και οι υπόλοιπες στήλες περιέχουν τη μέση τιμή, την τυπική απόκλιση, το Monte Carlo σφάλμα, τη διάμεσο και το διάστημα αξιοπιστίας για κάθε

παράμετρο του μοντέλου παλινδρόμησης. Επιπροσθέτως ο Πίνακας 5.2.1.1 μας δίνει περιγραφικούς δείκτες για τους δείκτες $\boldsymbol{\gamma}$.

Από τον Πίνακα 5.2.1.1 παρατηρούμε ότι ο σταθερός όρος β_0 και οι συντελεστές $\beta_3, \beta_5, \beta_9$ και β_{11} είναι στατιστικά σημαντικοί διότι τα διαστήματα αξιοπιστίας δεν περιέχουν την τιμή μηδέν. Για τους υπόλοιπους συντελεστές βλέπουμε ότι η μέση τιμή βρίσκεται κοντά στο μηδέν, η τυπική τους απόκλιση συγκριτικά με τη μέση τιμή είναι αρκετά υψηλή και αυτό έχει ως αποτέλεσμα τα διαστήματα αξιοπιστίας να συμπεριλαμβάνουν και την τιμή μηδέν. Από το γεγονός αυτό απορρέει ότι για τα πλήρη δεδομένα η μέθοδος SSVS δεν κατάφερε να ανιχνεύσει ότι το πραγματικό μοντέλο που έχουμε εμείς προσδιορίσει στην προσομοίωση των δεδομένων περιέχει και τον συντελεστή β_{11}. Ωστόσο είναι σωστό να υπενθυμίσουμε ότι στη μέθοδο SSVS οι μη στατιστικά σημαντικοί συντελεστές υπάρχουν στο μοντέλο αλλά έχουν τιμές πολύ κοντά στο μηδέν και επηρεάζουν έτσι αμυδρά το μοντέλο. Έτσι η διάσταση του μοντέλου της μεθόδου SSVS δεν θα είναι ίση με τη διάσταση του πραγματικού μοντέλου.

Στους δείκτες $\boldsymbol{\gamma}$ που αντιστοιχούν στις επεξηγηματικές μεταβλητές $\boldsymbol{X}_3$ και $\boldsymbol{X}_5$ διαπιστώνουμε ότι έχουν την τιμή ένα και για την $\boldsymbol{X}_9$ έχουμε ότι $\gamma_9 = 0.63$, ενώ ο δείκτης γ_{11} και οι δείκτες των επεξηγηματικών μεταβλητών που δε συμπεριλαμβάνονται στο πραγματικό μοντέλο έχουν τιμές πάρα πολύ κοντά στο μηδέν και κατέχουν μεγάλες τυπικές αποκλίσεις.

Συνεχίζουμε βρίσκοντας τις εκ των υστέρων πιθανότητες όλων των μοντέλων που προκύπτουν από τα δεδομένα μας. Στον κώδικα της μεθόδου SSVS στο WinBUGS το κάθε μοντέλο διακρίνεται με τη βοήθεια ενός δείκτη. Έτσι το αποτέλεσμα που αποθηκεύουμε στην R θα περιέχει τους δείκτες και τις εκ των υστέρων πιθανότητες που αντιστοιχούν στο κάθε μοντέλο. Για να μπορέσουμε να καταλάβουμε ποιο από τα 2^p μοντέλα αντιστοιχεί στον κάθε δείκτη θα χρησιμοποιήσουμε τις παρακάτω δύο συναρτήσεις στην R. Οι δύο συναρτήσεις αυτές είναι οι εξής

Συνάρτηση εύρεσης μοντέλου στην R

```
ind<-function(p)
{
#if we have 0,1 or 2 covariates
if(p == 0) {     return(t <- 0)  }
  else if(p == 1) {return(t <- rbind(0, 1)) }
  else if(p == 2) {
    return(t <- rbind(c(0, 0), c(1, 0), c(0, 1), c(1, 1)))
  }
  #for p>2
  else {
    t <- rbind(cbind(ind(p - 1), rep(0, 2^(p - 1))), cbind(
      ind(p - 1), rep(1, 2^(p - 1))))
    return(t)
  }}
```

Συνάρτηση εύρεσης μοντέλου με χρήση του δείκτη στην R

```
#This function uses the ind() function from above to extract
#the model structure from each model indicator
print.ind<-function(p)
{    t <- ind(p)
ee <- NULL
for(i in 2:nrow(t)) {
  e <- NULL
  L <- T
  for(j in 1:ncol(t)) {
    if(t[i, j] == 1 & L == T) {
      e <- paste(e, "x", j, sep = "")
      L <- F
    }
    else if(t[i, j] == 1 & L == F) {
      e <- paste(e, "+ x", j, sep = "")
    }}
  ee <- c(ee,e)}
return(c("intercept",ee))
}
modelssss <- print.ind(12)
#The indicator 1301 corresponds to the real model
modelssss[1301]
#[1] "x3+ x5+ x9+ x11"
```

Έχοντας τα αποτελέσματα του WinBUGS και με τη βοήθεια των παραπάνω συναρτήσεων, είμαστε σε θέση να δημιουργήσουμε έναν πίνακα (Πίνακας 5.2.1.2) ο οποίος θα περιέχει τα δέκα μοντέλα, όπου η αλυσίδα μας έχει επισκεφθεί τις περισσότερες φορές. Δηλαδή τα δέκα μοντέλα με τις μεγαλύτερες εκ των υστέρων πιθανότητες.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	277	0.5400	0.4985
$\boldsymbol{X}_3 + \boldsymbol{X}_5$	21	0.3190	0.4662
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.0320	0.1760
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_{11}$	1045	0.0150	0.1215
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9$	309	0.0080	0.0891
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10}$	789	0.0075	0.0862
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{12}$	2325	0.0075	0.0862
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	278	0.0065	0.0803
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	285	0.0060	0.0772
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9$	405	0.0060	0.0772

Πίνακας 5.2.1.2: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου SSVS στα πλήρη δεδομένα.

Από τον Πίνακα 5.2.1.2 βλέπουμε ότι το μοντέλο με τις επεξηγηματικές μεταβλητές $\boldsymbol{X}_3$, $\boldsymbol{X}_5$ και $\boldsymbol{X}_9$ κατέχει τη μεγαλύτερη εκ των υστέρων πιθανότητα σε σχέση με τα υπόλοιπα εννέα μοντέλα, ενώ το πραγματικό μοντέλο κατέχει μία αρκετά μικρή πιθανότητα. Αυτό συμβαίνει διότι η ύστερη μέση τιμή του συντελεστή β_{11} είναι αρκετά μικρή και είναι κοντά στο μηδέν.

5.2.2 Εφαρμογή της μεθόδου Kuo & Mallick

Όπως και στη μέθοδο SSVS έτσι και εδώ ξεκινάμε γράφοντας τον κώδικα του δειγματολήπτη Kuo & Mallick στο WinBUGS σε ένα αρχείο .txt (KuoMallick.txt).

Μέθοδος Kuo & Mallick στο WinBUGS. Ntzoufras (2002)

```
#Linear regression variable selection using Kuo Mallick
Model KuoMallick; {
#We standardise the Xs and the coefficients
for(j in 1:p){
b[j] <- beta[j]/sd(x[,j])
for(i in 1:n){
z[i,j] <- (x[i,j]-mean(x[,j]))/sd(x[,j])
     }
standardized_means[j] <- b[j]*mean(x[,j])
  }
b0 <- beta0 - sum(standardized_means[])
#Linear regression model
for(i in 1:n){
y[i] ~ dnorm (mu[i],tau)
mu[i] <- beta0 + gamma[1]*beta[1]* z[i,1] +gamma[2]*beta[2]*
    z[i,2]+ gamma[3]*beta[3]* z[i,3] + gamma[4]*beta[4]* z[i,4]+
    gamma[5]*beta[5]* z[i,5] + gamma[6]*beta[6]* z[i,6]+
    gamma[7]*beta[7]* z[i,7] + gamma[8]*beta[8]* z[i,8] +
    gamma[9]*beta[9]* z[i,9] + gamma[10]*beta[10]* z[i,10]+
    gamma[11]*beta[11]* z[i,11] + gamma[12]*beta[12]* z[i,12]
  }
#Normal diffuse prior for the constant and coeficients of the model
beta0 ~ dnorm(0, 0.0001)
for(j in 1:p){
beta[j] ~ dnorm(muprior[j],tauprior[j])
muprior[j] <- 0
tauprior[j] <- 0.0001
#Priors for variable indicator
gamma[j] ~ dbern(0.5)
}
#Model code and vector with the model indicators
for(j in 1:p){
modelindicators[j] <- gamma[j]*pow(2,j-1)
}
modelind <- 1 + sum(modelindicators[])
for(j in 1: models){
permodelselect[j] <- equals(modelind,j)
}
#Diffuse gamma prior for tau
tau ~ dgamma(1.0E-3,1.0E-3)
sigma <- sqrt(1/tau)
}
```

Στο δειγματολήπτη Kuo & Mallick (Ενότητα 3.4) ξέρουμε ότι η εκ των προτέρων κατανομή των συντελεστών $\boldsymbol{\beta}$ είναι ανεξάρτητη της πρότερης κατανομής του διανύσματος δείκτη $\boldsymbol{\gamma}$. Κατά συνέπεια στο δειγματολήπτη Kuo & Mallick η εκ των προτέρων κατανομή των συντελεστών του μοντέλου θα είναι η εξής

$$\beta_j \sim N(0, \tau_j = 0.0001),$$

όπου αυτό σημαίνει ότι κάθε συντελεστής του μοντέλου θα έχει μία ασαφή εκ των προτέρων κανονική κατανομή με μέση τιμή το μηδέν. Χρειάζεται επίσης να σημειωθεί ότι η χρήση αυτής της εκ των προτέρων κατανομής υπονοεί ότι οι συντελεστές του μοντέλου είναι ανεξάρτητοι. Για τη σταθερά του μοντέλου β_0 και την ακρίβεια τ οι εκ των προτέρων κατανομές είναι οι εξής:

$$\beta_0 \sim N(0, \tau_0 = 0.0001), \quad \tau \sim \text{Gamma}(0.001, 0.001).$$

Σε αντίθεση με τη μέθοδο SSVS, ο δειγματολήπτης Kuo & Mallick ενσωματώνει τους δείκτες γ_j στο μοντέλο. Έτσι με τη βοήθεια της σχέσης 3.4.1 έχουμε

$$Y_i = \beta_0 + \gamma_1\beta_1 X_{i1} + \gamma_2\beta_2 X_{i1} + \cdots + \gamma_p\beta_p X_{ip} + \epsilon_i, \quad \text{με } \epsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right),$$

όπου και πάλι για κάθε γ_j, $j = 1, \ldots, p$ επιλέγουμε ως εκ των προτέρων κατανομή μία κατανομή Bernoulli με πιθανότητα επιτυχίας ίση με 0.5.

Κώδικας εκτέλεσης της μεθόδου Kuo & Mallick στο WinBUGS μέσω της εντολής bugs στην R

```
#We list our initial values and the dataset
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
   tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=FullDataset$Y, x
   =as.matrix(FullDataset[,2:13]),p=12, n=80, models=4096)
#We use the bugs function to execute our bugs code
FUllmodelKuoMallick <- bugs(inits=inits,data=data,model.file =
   "KuoMallick.txt",parameters = c("beta0","beta", "gamma",
   "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
   = 5000,n.thin = 1,n.sims = 3000,bugs.directory = "c:/Program
   Files/WinBUGS14/",debug=FALSE)
#Here we have the summary of the model parameters
FUllmodelparametersKuoMallick <- FUllmodelKuoMallick$summary[1:26,]
FUllmodelparametersKuoMallick
FUllmodelsKuoMallick <- FUllmodelKuoMallick$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
head(FUllmodelsKuoMallick[order(FUllmodelsKuoMallick[, 2],
   decreasing=TRUE), ],n=10)
```

Οι αρχικές τιμές θα παραμείνουν ίδιες, έτσι ώστε οι αλυσίδες και των δύο μεθόδων να ξεκινούν έχοντας τους συντελεστές του μοντέλου ίσους με το μηδέν και το διάνυσμα $\boldsymbol{\gamma}$ να έχει την τιμή ένα για κάθε $j = 1, \ldots, 12$. Στην εντολή bugs στην R έχουμε κρατήσει όλες τις παραμέτρους σταθερές και απλά αλλάξαμε το αρχείο .txt στο αρχείο KuoMallick.txt, όπου το περιεχόμενό του αναγράφεται παραπάνω. Οι

περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου, του διανύσματος $\boldsymbol{\gamma}$ και των πιθανοτήτων των καλύτερων μοντέλων του δειγματολήπτη Kuo & Mallick που μας δίνει το WinBUGS, δίνονται από τους Πίνακες 5.2.2.1 και 5.2.2.2.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.409	0.1203	0.002057	2.178	2.408	2.643
beta[1]	−0.3556	100.2	1.691	−197.6	−0.2012	192.5
beta[2]	−2.384	99.62	1.573	−203.3	−1.294	191.8
beta[3]	2.672	0.136	0.005604	2.397	2.673	2.939
beta[4]	0.4182	99.1	1.968	−194.6	0.6471	191.3
beta[5]	−1.827	0.1586	0.01115	−2.123	−1.829	−1.496
beta[6]	−2.401	102.4	2.062	−213.1	−0.9348	192.7
beta[7]	0.5717	100.6	2.049	−198.2	−0.4081	200.7
beta[8]	$3.535E-4$	98.37	1.922	−185.9	−0.8793	203.3
beta[9]	1.076	0.1412	0.007464	0.8074	1.071	1.358
beta[10]	1.36	99.65	1.726	−196.1	1.754	191.5
beta[11]	−1.975	61.99	1.038	−152.1	0.5371	148.7
beta[12]	0.155	100.5	1.759	−199.5	0.436	197.2
sigma	1.108	0.1051	0.00652	0.9269	1.103	1.34
Δείκτες γ						
gamma[1]	0.014	0.1175	0.005931	0.0	0.0	0.0
gamma[2]	0.001667	0.04079	0.001095	0.0	0.0	0.0
gamma[3]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[4]	0.001667	0.04079	$9.785E-4$	0.0	0.0	0.0
gamma[5]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[6]	0.002	0.04468	$9.017E-4$	0.0	0.0	0.0
gamma[7]	0.006333	0.07933	0.003377	0.0	0.0	0.0
gamma[8]	0.004667	0.06815	0.003674	0.0	0.0	0.0
gamma[9]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[10]	0.002667	0.05157	0.001928	0.0	0.0	0.0
gamma[11]	0.6153	0.4865	0.06134	0.0	1.0	1.0
gamma[12]	0.003333	0.05764	0.001887	0.0	0.0	0.0

Πίνακας 5.2.2.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ του δειγματολήπτη Kuo & Mallick στα πλήρη δεδομένα.

Από τον Πίνακα 5.2.2.1 παρατηρούμε ότι η σταθερά β_0 και οι συντελεστές β_3, β_5 και β_9 είναι σημαντικοί διότι τα διαστήματα αξιοπιστίας δεν περιέχουν την τιμή μηδέν, ενώ ο συντελεστής β_{11} και οι υπόλοιποι συντελεστές που δεν υπάρχουν στο πραγματικό μοντέλο κατέχουν μεγάλες τιμές τυπικών αποκλίσεων και έτσι τα διαστήματα αξιοπιστίας περιέχουν την τιμή μηδέν. Στους δείκτες $\boldsymbol{\gamma}$ διαπιστώνουμε ότι τα $\gamma_3, \gamma_5, \gamma_9$ έχουν την τιμή ένα και το $\gamma_{11} = 0.6153$, όπου έχει σχετικά μεγάλη τυπική απόκλιση.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.5896	0.4919
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	277	0.37465	0.4841
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1302	0.0123	0.1103
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	405	0.0040	0.0631
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9$	341	0.0036	0.0604
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1365	0.0026	0.0515
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10} + \boldsymbol{X}_{11}$	1813	0.0026	0.0515
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3349	0.0026	0.0515
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1303	0.0016	0.0407
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1461	0.0013	0.0364

Πίνακας 5.2.2.2: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων του δειγματολήπτη Kuo & Mallick στα πλήρη δεδομένα.

Από τον Πίνακα 5.2.2.2 διαπιστώνουμε ότι και ο δειγματολήπτης Kuo & Mallick κατάφερε να ανιχνεύσει το πραγματικό μοντέλο, όμως βλέπουμε πως η αλυσίδα επέλεξε αρκετές φορές και το μοντέλο που δεν περιέχει το συντελεστή β_{11}.

5.2.3 Εφαρμογή της μεθόδου GVS

Για τη μέθοδο GVS θα χρειαστεί να προσαρμόσουμε αρχικά το μοντέλο που περιέχει όλες τις επεξηγηματικές μεταβλητές, έχοντας ως στόχο να πάρουμε τις ύστερες εκτιμήσεις των παραμέτρων της ψευδο-πρότερης εκ των προτέρων κατανομής όπου αναφέραμε στην Ενότητα 3.5.

Κώδικας εκτέλεσης του Full μοντέλου στο WinBUGS

```
#Full model run to extract the mu and sigma for the pseudoprior
Model pseudoprior;
{
for(j in 1:p){
b[j] <- beta[j]/sd(x[,j])
for(i in 1:n){
z[i,j] <- (x[i,j]-mean(x[,j]))/sd(x[,j])
     }
standardized_means[j] <- b[j]*mean(x[,j])
  }
b0 <- beta0 - sum(standardized_means[])
for(i in 1:n){
y[i] ~ dnorm (mu[i],tau)
mu[i] <- beta0 + beta[1]* z[i,1] + beta[2]* z[i,2]+ beta[3]* z[i,3]
    + beta[4]* z[i,4]+ beta[5]* z[i,5] + beta[6]* z[i,6]+ beta[7]*
    z[i,7] + beta[8]* z[i,8] + beta[9]* z[i,9] + beta[10]* z[i,10]+
    beta[11]* z[i,11] + beta[12]* z[i,12]
  }
beta0 ~ dnorm(0, 1.0E-5)
for(j in 1:p){
beta[j] ~ dnorm(0,1.0E-5)
}
tau ~ dgamma(1.0E-3,1.0E-3)
sigma <- sqrt(1/tau)
}
```

Κώδικας εκτέλεσης του Full μοντέλου μέσω της εντολής bugs

```
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2)}
data <- list(y=FullDataset$Y, x
    =as.matrix(FullDataset[,2:13]),p=12, n=80)
FUllmodelpilotrun <- bugs(inits=inits, data=data, model.file =
    "Pilotrun.txt",parameters = c("beta0","beta", "sigma"),n.chains
    = 1, n.burnin = 2000, n.iter = 7000,n.thin = 1,n.sims =
```

```
5000,bugs.directory = "c:/Program Files/WinBUGS14/",debug=FALSE)
```

Συγκεκριμένα για τις παραμέτρους στο πλήρες μοντέλο έχουμε ορίσει τις εξής εκ των προτέρων κατανομές:

$$\beta_0 \sim N\left(0, \tau_0 = 10^{-5}\right), \quad \beta_j \sim N\left(0, \tau_j = 10^{-5}\right), \quad \tau \sim \text{Gamma}(0.001, 0, 001),$$

όπου $j = 1, \ldots, p$.

Τρέχοντας τον παραπάνω κώδικα στην R και προσαρμόζοντας το πλήρες μοντέλο στο πρόγραμμα WinBUGS (αρχείο Pilotrun.txt) με μία αλυσίδα και συνολικά 7000 επαναλήψεις, όπου οι 2000 είναι burned-in, καταλήγουμε στον Πίνακα 5.2.3.1 όπου μας δίνονται οι περιγραφικοί δείκτες των προσομοιωμένων τιμών των ύστερων κατανομών των παραμέτρων της ψευδο-πρότερης κατανομής όπου ορίσαμε στην Ενότητα 3.5.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.412	0.1206	0.001681	2.176	2.412	2.648
beta[1]	−0.3085	0.1455	0.001895	−0.5893	−0.3083	−0.02529
beta[2]	−0.05654	0.1616	0.001968	−0.3776	−0.05543	0.2582
beta[3]	2.741	0.1794	0.002639	2.391	2.739	3.095
beta[4]	−0.08608	0.1477	0.001739	−0.3822	−0.08342	0.2101
beta[5]	−1.791	0.1471	0.001984	−2.073	−1.792	−1.501
beta[6]	0.0436	0.1799	0.002773	−0.301	0.04189	0.4069
beta[7]	−0.007025	0.2172	0.002803	−0.44	−0.006926	0.4116
beta[8]	−0.1149	0.1469	0.001647	−0.4069	−0.1132	0.1745
beta[9]	1.085	0.169	0.002226	0.7554	1.085	1.418
beta[10]	−0.1012	0.1902	0.002705	−0.4824	−0.1036	0.2744
beta[11]	0.6239	0.1606	0.002026	0.3036	0.6251	0.9357
beta[12]	0.06587	0.1884	0.002617	−0.3068	0.0641	0.4375

Πίνακας 5.2.3.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των συντελεστών του πλήρες μοντέλου.

Στη μέθοδο GVS (Ενότητα 3.5) με τη χρήση της σχέσης (3.5.5), η εκ των προτέρων κατανομή των παραμέτρων β_j θα είναι μία μείξη Κανονικών κατανομών που έχει την εξής μορφή:

$$\beta_j|\gamma_j \sim \gamma_j N(0, \tau_1 = 0.0001) + (1 - \gamma_j) N\left(\bar{\mu}_j, \frac{1}{\bar{s}_j^2}\right), \quad j = 1, \ldots, p,$$

όπου οι παράμετροι $\bar{\mu}_j$ και $\bar{s}_j$ της ψευδο-πρότερης κατανομής προσδιορίζονται με τη βοήθεια του πίνακα 5.2.3.1. Ωστόσο καλές επιλογές για τις παραμέτρους $\bar{\mu}_j$ και $\bar{s}_j$ είναι οι τιμές που προτάθηκαν από Ntzoufras (1999).

Για τη σταθερά του μοντέλου β_0 και την ακρίβεια τ, οι εκ των προτέρων κατανομές είναι οι εξής:

$$\beta_0 \sim N(0, \tau_0 = 0.0001), \quad \tau \sim \text{Gamma}(0.001, 0.001).$$

Όπως και στο δειγματολήπτη Kuo & Mallick, έτσι και εδώ το διάνυσμα $\boldsymbol{\gamma}$ ενσωματώνεται στην παλινδρόμηση, οπότε έχουμε ότι:

$$Y_i = \beta_0 + \gamma_1\beta_1 X_{i1} + \gamma_2\beta_2 X_{i2} + \cdots + \gamma_p\beta_p X_{ip} + \epsilon_i, \quad \text{με } \epsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right),$$

όπου και εδώ, για κάθε δείκτη γ_j προσδιορίζουμε ως εκ των προτέρων κατανομή μία κατανομή Bernoulli με πιθανότητα επιτυχίας ίση με 0.5.

Γράφουμε τον κώδικα της μεθόδου GVS στο WinBUGS σε ένα αρχείο .txt (GVS.txt).

Μέθοδος GVS στο WinBUGS. Ntzoufras (2002)

```
#Linear regression variable selection using GVS
Model GVS; {
#We standardise the X's and the coefficients
for(j in 1:p){
b[j] <- beta[j]/sd(x[,j])
for(i in 1:n){
z[i,j] <- (x[i,j]-mean(x[,j]))/sd(x[,j])
     }
standardized_means[j] <- b[j]*mean(x[,j])
  }
b0 <- beta0 - sum(standardized_means[])
#Linear regression model
for(i in 1:n){
y[i] ~ dnorm (mu[i],tau)
mu[i] <- beta0 + gamma[1]*beta[1]* z[i,1] +gamma[2]*beta[2]*
    z[i,2]+ gamma[3]*beta[3]* z[i,3] + gamma[4]*beta[4]* z[i,4]+
    gamma[5]*beta[5]* z[i,5] + gamma[6]*beta[6]* z[i,6]+
    gamma[7]*beta[7]* z[i,7] + gamma[8]*beta[8]* z[i,8] +
    gamma[9]*beta[9]* z[i,9] + gamma[10]*beta[10]* z[i,10]+
    gamma[11]*beta[11]* z[i,11] + gamma[12]*beta[12]* z[i,12]
  }
#Normal diffuse prior for the constant of the model
beta0 ~ dnorm(0, 0.0001)
#Mixture of normals as a prior
for(j in 1:p){
beta[j] ~ dnorm(muprior[j],tauprior[j])
muprior[j] <- (1-gamma[j])*mean[j]
tauprior[j] <- gamma[j]*0.0001 + (1-gamma[j])/(se[j]*se[j])
#Priors for variable indicator
gamma[j] ~ dbern(0.5)
}
#Model code and vector with the model indicators
for(j in 1:p){
modelindicators[j] <- gamma[j]*pow(2,j-1)
}
modelind <- 1 + sum(modelindicators[])
for(j in 1: models){
permodelselect[j] <- equals(modelind,j)
}#Diffuse gamma prior for tau
tau ~ dgamma(1.0E-3,1.0E-3)
sigma <- sqrt(1/tau)
}
```

Χρησιμοποιούμε λοιπόν τις τιμές από τον Πίνακα 5.2.3.1, τρέχουμε τον παρακάτω κώδικα στην R και έτσι με τη βοήθεια του WinBUGS (εφαρμόζουμε τη μέθοδο GVS) καταλήγουμε στους Πίνακες 5.2.3.2 και 5.2.3.3.

Εφαρμογή της μεθόδου GVS

```
#We list our initial values and the dataset
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=FullDataset$Y, x
=as.matrix(FullDataset[,2:13]),p=12, n=80, models=4096, mean=
as.vector(colMeans(FUllmodelpilotrun$sims.array[1:5000,1,]))[2:13],
se=as.vector(FUllmodelpilotrun$sd$beta))
FUllmodelgvs <- bugs(inits=inits,data=data,model.file =
"GVS.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 5000,n.thin = 1,n.sims = 3000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=TRUE)
#Here we have the summary of the model parameters
FUllmodelparametersgvs <- FUllmodelgvs$summary[1:26,]
FUllmodelsgvs <- FUllmodelgvs$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
head(FUllmodelsgvs[order(FUllmodelsgvs[, 2], decreasing=TRUE),
    ],n=10)
```

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.411	0.1235	0.002183	2.168	2.412	2.654
beta[1]	−0.309	0.1468	0.002468	−0.5977	−0.3079	−0.02685
beta[2]	−0.05882	0.1563	0.002764	−0.3495	−0.05989	0.246
beta[3]	2.678	0.1306	0.002655	2.419	2.68	2.93
beta[4]	−0.08696	0.1505	0.00284	−0.3843	−0.08765	0.2158
beta[5]	−1.849	0.1544	0.004954	−2.13	−1.853	−1.544
beta[6]	0.04455	0.1804	0.003675	−0.2977	0.03967	0.4071
beta[7]	−0.0136	0.2225	0.004618	−0.47	−0.01075	0.4116
beta[8]	−0.1144	0.1483	0.003031	−0.4071	−0.1167	0.1801
beta[9]	1.063	0.1363	0.003634	0.8104	1.055	1.342
beta[10]	−0.102	0.1884	0.003251	−0.4635	−0.1016	0.2669
beta[11]	0.568	0.1479	0.003251	0.2823	0.5679	0.8617
beta[12]	0.06052	0.1881	0.00367	−0.3197	0.05895	0.4248
sigma	1.097	0.1025	0.002719	0.9224	1.087	1.317
Δείκτες γ						
gamma[1]	0.008667	0.09269	0.001772	0.0	0.0	0.0
gamma[2]	0.001333	0.03649	$6.453E-4$	0.0	0.0	0.0
gamma[3]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[4]	0.001	0.03161	$5.66E-4$	0.0	0.0	0.0
gamma[5]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[6]	0.002	0.04468	$7.723E-4$	0.0	0.0	0.0
gamma[7]	0.005667	0.07506	0.001557	0.0	0.0	0.0
gamma[8]	0.001	0.03161	$5.591E-4$	0.0	0.0	0.0
gamma[9]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[10]	0.001667	0.04079	$7.13E-4$	0.0	0.0	0.0
gamma[11]	0.727	0.4455	0.01708	0.0	1.0	1.0
gamma[12]	0.001	0.03161	$5.694E-4$	0.0	0.0	0.0

Πίνακας 5.2.3.2: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ της μεθόδου GVS στα πλήρη δεδομένα.

Στον Πίνακα 5.2.3.2 εύκολα παρατηρούμε ότι οι συντελεστές του πραγματικού μοντέλου είναι σημαντικοί, διότι τα διαστήματα αξιοπιστίας τους δεν περιέχουν την τιμή μηδέν. Στους δείκτες $\boldsymbol{\gamma}$ βλέπουμε ότι οι δείκτες γ_3, γ_5 και γ_9 έχουν την τιμή ένα, ενώ όπως είδαμε και στον Πίνακα 5.2.2.1, ο δείκτης γ_{11} είναι μικρότερος του ένα και κατέχει σχετικά μεγάλη τυπική απόκλιση.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.7093	0.4541
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	277	0.2683	0.4431
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1302	0.0086	0.0927
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9$	341	0.0040	0.0631
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1333	0.0016	0.0407
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1365	0.0016	0.0407
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10} + \boldsymbol{X}_{11}$	1813	0.0016	0.0407
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1303	0.0013	0.0364
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1309	0.0010	0.0316
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3349	0.0010	0.0316

Πίνακας 5.2.3.3: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου GVS στα πλήρη δεδομένα.

Από τον Πίνακα 5.2.3.3 παρατηρούμε ότι το μοντέλο με τη μεγαλύτερη εκ των υστέρων πιθανότητα αντιστοιχεί στο πραγματικό μοντέλο, ενώ έπειτα ακολουθεί το μοντέλο χωρίς το συντελεστή β_{11}. Αυτό είναι λογικό διότι ο συντελεστής β_{11} είναι απόλυτα μικρότερος από τους άλλους συντελεστές που υπάρχουν στο πραγματικό μοντέλο.

5.2.4 Πλήρης απαρίθμηση με τη χρήση του κριτηρίου BIC

Η πλήρης απαρίθμηση (full enumeration) είναι μία μέθοδος επιλογής μοντέλου που χρησιμοποιείται στη στατιστική και τη μηχανική μάθηση. Σε αυτή τη μέθοδο προσαρμόζουμε κάθε πιθανό μοντέλο στα δεδομένα έχοντας ως σκοπό τον υπολογισμό του κριτηρίου BIC (Ενότητα 2.4.2). Το μοντέλο με τη χαμηλότερη τιμή στο κριτήριο BIC επιλέγεται στη συνέχεια ως το προτιμώμενο μοντέλο, υποδεικνύοντας έτσι την καλύτερη αντιστάθμιση μεταξύ προσαρμογής και πολυπλοκότητας μοντέλου. Αυτή η εξαντλητική διαδικασία διασφαλίζει ότι κανένα μοντέλο δεν αγνοείται.

Αν και η πλήρης απαρίθμηση σε συνδυασμό με το κριτήριο BIC εγγυάται μια εξαντλητική εξερεύνηση όλων των δυνατών μοντέλων και ταυτόχρονα μας διασφαλίζει τη βέλτιστη επιλογή μεταξύ των υποψηφίων, μπορεί εξίσου να γίνει και υπολογιστικά δύσκολη, κυρίως για εκτεταμένα σύνολα δεδομένων ή περίπλοκα μοντέλα με πολλές παραμέτρους.

Ωστόσο η μέθοδος αυτή είναι ιδιαίτερα χρήσιμη σε καταστάσεις όπου ο χώρος των μοντέλων είναι σχετικά μικρός και διαχειρίσιμος, επιτρέποντας μια εξαντλητική αξιολόγηση όλων των πιθανών υποψήφιων μοντέλων. Δηλαδή έχοντας ένα μικρό αριθμό επεξηγηματικών μεταβλητών (συνήθως $p < 15$), είμαστε σε θέση να χρησιμοποιήσουμε πλήρη απαρίθμηση χωρίς κάποιο ιδιαίτερο υπολογιστικό κόστος.

Εμείς στο προσομοιωμένο δείγμα μας έχουμε συνολικά $p = 12$ επεξηγηματικές μεταβλητές. Συνεπώς μας είναι εφικτό να χρησιμοποιήσουμε πλήρη απαρίθμηση χωρίς να υποστούμε μεγάλο υπολογιστικό κόστος.

Για να εφαρμόσουμε τη μέθοδο αυτή στα δεδομένα μας, δημιουργούμε μία συνάρτηση στην R με όνομα BICminimum, η οποία θα προσαρμόζει όλα τα πιθανά μοντέλα στα δεδομένα μας, θα υπολογίζει το κριτήριο BIC για κάθε μοντέλο και θα μας δίνει τα πρώτα 10 μοντέλα με τη μικρότερη τιμή του BIC.

Πλήρης απαρίθμηση με χρήση του κριτηρίου BIC

```
#This function implements the full enumeration method
#using the BIC criterion
BICminimum<-function(p,data){
#We create all possible models
  lista<-list()
  for(i in 1:p){
    lista[[i]]<-c(0,i)
  }
  gridd<-expand.grid(lista)
  colnames(gridd)<-colnames(data[,2:ncol(data)])
  allpossiblemodels<-apply(gridd,1,
  function(x)paste(colnames(data)[1],paste("~"),
  paste(colnames(data[,2:ncol(data)])[x],collapse="+")))
  allpossiblemodels[1]<-paste(colnames(data)[1],paste("~1"))
  datasetframe<-as.data.frame(data)
  #Here we apply the lm() function for every combination
  #of covariates
  modelsS<-lapply(allpossiblemodels,lm,data=datasetframe)
  # All possible model coefficients
  allpossiblecoefficients<-lapply(modelsS,coef)
  #BIC values for every model.
  allBIC<-lapply(modelsS,BIC)
  # Top 10 models with the lowest BIC value
  top10models <-
     allpossiblecoefficients[head(order(unlist(allBIC)), 10)]
  # The BIC values of those 10 models
  top10BIC <- allBIC[head(order(unlist(allBIC)), 10)]
  #We return the 10 models, their corresponding BIC's and
  #the model codes
  return(list(top10models,top10BIC,head(order(unlist(allBIC)), 10)))
}
```

Η συνάρτηση BICminimum που δημιουργήσαμε περιέχει δύο παραμέτρους, την παράμετρο p που αντιστοιχεί στον αριθμό των επεξηγηματικών μεταβλητών (εμείς στο δείγμα μας έχουμε ότι $p = 12$) και την παράμετρο data όπου αντιστοιχεί στο δείγμα μας. Τρέχουμε τον παρακάτω κώδικα στην R και εφαρμόζουμε τη συνάρτηση BICminimun στο προσομοιωμένο δείγμα και έτσι καταλήγουμε στον Πίνακα 5.2.3.4.

Εφαρμογή της συνάρτησης BICminimum στην R

```
#We implement the BICminimum function we created above
```

```
Lowest10BICmodelsFULL <-BICminimum(p=12,FullDataset)
#Coefficients of the models
Lowest10BICmodelsFULL[[1]][1:10]
#and their corresponding BIC values.
Lowest10BICmodelsFULL[[2]][1:10]
#Model code
Lowest10BICmodelsFULL[[3]]
```

Παράμετροι μοντέλου	Αριθμός μοντέλου	Κριτήριο BIC
$\boldsymbol{X}_1+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1302	255.1372
$\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1301	256.6603
$\boldsymbol{X}_1+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{10}+\boldsymbol{X}_{11}$	1814	257.9006
$\boldsymbol{X}_1+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_8+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1430	258.0106
$\boldsymbol{X}_1+\boldsymbol{X}_3+\boldsymbol{X}_4+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1310	258.5427
$\boldsymbol{X}_1+\boldsymbol{X}_2+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1304	259.0012
$\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_7+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1365	259.0287
$\boldsymbol{X}_1+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_7+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1366	259.2298
$\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{10}+\boldsymbol{X}_{11}$	1813	259.2606
$\boldsymbol{X}_3+\boldsymbol{X}_4+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1309	259.4572

Πίνακας 5.2.3.4: Τα 10 πρώτα μοντέλα με τις μικρότερες τιμές του κριτηρίου BIC στα πλήρη δεδομένα.

Από τον Πίνακα 5.2.3.4 βλέπουμε ότι το μοντέλο με τη μικρότερη τιμή του κριτηρίου BIC, αντιστοιχεί στο μοντέλο που περιέχει τους συντελεστές του πραγματικού μοντέλου μαζί με το συντελεστή β_1, ενώ το μοντέλο με τη δεύτερη μικρότερη τιμή BIC αντιστοιχεί στο πραγματικό μοντέλο.

Μοντέλο με τη μικρότερη τιμή του κριτηρίου BIC

```
summary(lm(Y~ X1+X3+X5+X9+X11,data=FullDataset))
Coefficients: #Partial console output
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.0934     0.1196  17.508  < 2e-16 ***
X1          -0.2812     0.1181  -2.381   0.0198 *
X3           2.6066     0.1131  23.045  < 2e-16 ***
X5          -1.8202     0.1271 -14.319  < 2e-16 ***
X9           1.0571     0.1223   8.640 7.96e-13 ***
X11          0.5439     0.1191   4.565 1.94e-05 ***
---
#We load the car library in R, in order to use the function vif
library(car)
vif(lm(Y~ X1+X3+X5+X9+X11,data=FullDataset))
#Console output
      X1       X3       X5       X9      X11
1.146096 1.063569 1.232490 1.093022 1.282533
```

Από τα παραπάνω αποτελέσματα στην R βλέπουμε ότι ο συντελεστής β_1 είναι στατιστικά σημαντικός για επίπεδο σημαντικότητας $\alpha = 0.05$. Επίσης διαπιστώνουμε

ότι στο μοντέλο αυτό ο παράγοντας VIF (variance inflation factor) είναι αρκετά μικρός για κάθε επεξηγηματική μεταβλητή, οπότε το μοντέλο μας δεν πάσχει από πολυσυγγραμμικότητα.

5.3 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης

Όταν αντιμετωπίζουμε δεδομένα που περιέχουν ελλιπείς τιμές στη μεταβλητή απόκρισης στο πλαίσιο της Μπεϋζιανής επιλογής μεταβλητών, η επιλογή των μεθόδων αντικατάστασης καθίσταται κρίσιμη για ακριβή και αξιόπιστα αποτελέσματα. Σε περιπτώσεις όπου ο μηχανισμός έλλειψης ακολουθεί MCAR, MAR ή NMAR (Ενότητα 4.2), η επιλογή μίας κατάλληλης μεθόδου αντικατάστασης μπορεί να μας επηρεάσει σε μεγάλο βαθμό τα τελικά αποτελέσματα. Σε αυτή την ενότητα θα εμβαθύνουμε στη σύγκριση των τριών Μπεϋζιανών μεθόδων επιλογής μεταβλητών stochastic search variable selection (SSVS), Kuo & Mallick και Gibbs variable selection, έχοντας λάβει υπόψιν κάθε φορά διαφορετικούς μηχανισμούς έλλειψης τιμών. Θα διερευνήσουμε πώς αποδίδει κάθε μέθοδος επιλογής μεταβλητών και θα αξιολογήσουμε την αποτελεσματικότητά τους ως προς την εύρεση του πραγματικού μοντέλου. Μέσω μιας τέτοιας ολοκληρωμένης αξιολόγησης, στόχος μας είναι να ρίξουμε φως στα πλεονεκτήματα και τους περιορισμούς αυτών των μεθόδων όταν ερχόμαστε αντιμέτωποι με δείγματα που περιέχουν ελλιπείς τιμές στη μεταβλητή απόκρισης.

5.3.1 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές υπό το μηχανισμό MCAR

Η κατασκευή του μηχανισμού MCAR είναι αρκετά εύκολη και απλή, εφόσον η έλλειψη στη μεταβλητή απόκρισης θα είναι πλήρως τυχαία και έτσι η πιθανότητα έλλειψης κάθε παρατήρησης δε θα εξαρτάται ούτε από τις παρατηρούμενες, αλλά και ούτε και από τις μη παρατηρούμενες τιμές. Κατασκευάζουμε λοιπόν στην R ένα μηχανισμό έλλειψης MCAR για τη μεταβλητή απόκρισης.

Κατασκευή μηχανισμού MCAR στη μεταβλητή απόκρισης

```
#We create a Bernoulli trial for every observation
#with probability p=0.65 of being observed.
miss.Y <- rbinom(samplesize, 1, prob = 0.65)
# # of observations that remain observed after the MCAR mechanism
sum(miss.Y==1)
MCARDataset <- FullDataset
# if miss.y=1 then the observation remains observed
#and if miss.y=0 then it becomes a missing value.
MCARDataset$Y <- ifelse(miss.Y==1,MCARDataset$Y,NA)
```

Στον παραπάνω κώδικα δημιουργήσαμε ένα μηχανισμό έλλειψης MCAR θεωρώντας μία κατανομή Bernoulli με πιθανότητα επιτυχίας $p = 0.65$ (πιθανότητα η

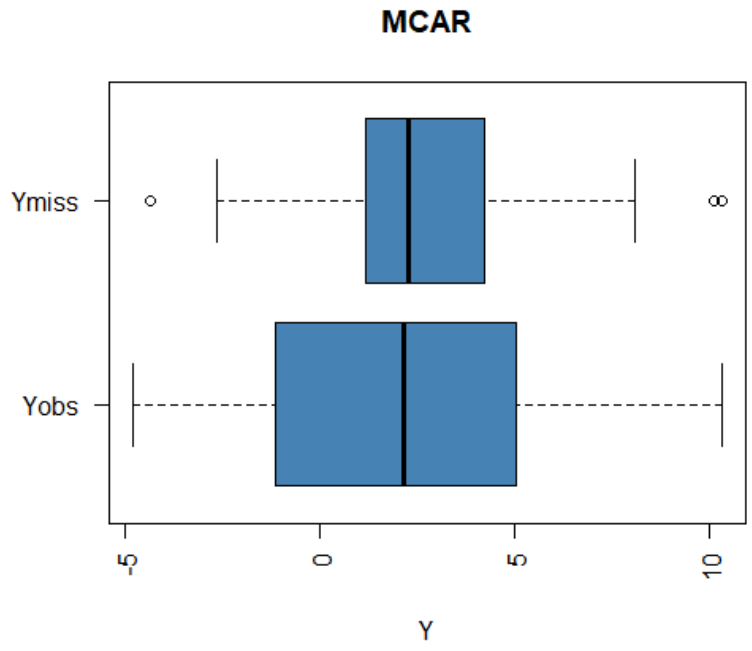
κάθε παρατήρηση να παραμείνει παρατηρήσιμη). Έτσι με αυτό το τρόπο κάθε παρατήρηση γίνεται ελλιπής με πιθανότητα $1 - p = 0.35$.

Κοιτώντας το Διάγραμμα 4.2.1 στο παράδειγμα του Van Buuren (2018), περιμένουμε ότι τα θηκογράμματα των Y_{miss} και Y_{obs} δεν έχουν σημαντικές διαφορές μεταξύ τους.

Κατασκευή θηκογραμμάτων

```
#Number of observations that remain observed after the
#MCAR mechanism
sum(miss.Y==1)
#console output is 53.
#Number of missing values in y.
sum(is.na(MCARDataset$Y))
#console output is 27.
#These are the observations that we took as missing
YmissMCAR <- FullDataset$Y[is.na(MCARDataset$Y)]
#Boxplot within the missing and the remaining observations.
boxplot(MCARDataset$Y, YmissMCAR, horizontal=TRUE, col='steelblue',
las=2,main = "MCAR", xlab = "Y", names = c("Yobs","Ymiss"))
```

Με τη χρήση του παραπάνω κώδικα καταλήγουμε στο Διάγραμμα 5.3.1.1, όπου διαπιστώνουμε ότι δεν υπάρχουν σημαντικές διαφορές μεταξύ των τιμών των Y_{miss} και Y_{obs}.



Διάγραμμα 5.3.1.1: Θηκογράμματα των τιμών του Y_{miss} και του Y_{obs}.

Για να αντικαταστήσουμε αυτές τις παρατηρήσεις θα χρησιμοποιήσουμε την ίδια

λογική που αναπτύξαμε στην Ενότητα 4.5, δηλαδή θα κάνουμε χρήση της εκ των υστέρων προβλεπτικής κατανομής η οποία είναι η εξής:

$$p\left(\boldsymbol{Y}_{miss}|\boldsymbol{Y}_{obs}\right) = \int p\left(\boldsymbol{Y}_{miss}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}\right)\, d\boldsymbol{\theta},$$

όπου $p\left(\boldsymbol{Y}_{miss}|\boldsymbol{\theta}\right)$ είναι η δειγματοληπτική κατανομή των $\boldsymbol{Y}_{miss}$ και $p\left(\boldsymbol{\theta}|\boldsymbol{Y}_{obs}\right)$ είναι η ύστερη κατανομή των παραμέτρων του μοντέλου. Γράφουμε λοιπόν τον κώδικα του μοντέλου αντικατάστασης στο WinBUGS σε ένα αρχείο .txt (imputationresponse.txt). Αυτό το μοντέλο θα χρησιμοποιηθεί για να αντικαταστήσουμε τις ελλιπείς τιμές της μεταβλητής απόκρισης ανεξάρτητα απ'το μηχανισμό έλλειψης που διαθέτουμε στο προσομοιωμένο δείγμα μας.

Μοντέλο αντικατάστασης ελλιπών τιμών στη μεταβλητή απόκρισης (imputationresponse.txt)

```
Model imputationresponse;
{
for(j in 1:p){
b[j] <- beta[j]/sd(x[,j])
for(i in 1:n){
z[i,j] <- (x[i,j]-mean(x[,j]))/sd(x[,j])
     }
standardized_means[j] <- b[j]*mean(x[,j])
  }
b0 <- beta0 - sum(standardized_means[])
#Linear regression model
for(i in 1:n){
y[i] ~ dnorm (mu[i],tau)
mu[i] <- beta0 + beta[1]* z[i,1] + beta[2]* z[i,2]+ beta[3]* z[i,3]
    + beta[4]* z[i,4]+ beta[5]* z[i,5] + beta[6]* z[i,6]+ beta[7]*
    z[i,7] + beta[8]* z[i,8] + beta[9]* z[i,9] + beta[10]* z[i,10]+
    beta[11]* z[i,11] + beta[12]* z[i,12]
  }
#Normal diffuse prior for the constant of the model
beta0 ~ dnorm(0, 0.0001)
#Normal diffuse priors for all the coefficients of the model
for(j in 1:p){
beta[j] ~ dnorm(0,1.0E-5)
}
#Diffuse gamma prior for tau
tau ~ dgamma(1.0E-3,1.0E-3)
sigma <- sqrt(1/tau)
}
```

Τρέχουμε το παραπάνω αρχείο .txt, που περιέχει το μοντέλο αντικατάστασης των ελλιπών τιμών σε κώδικα του WinBUGS με τη βοήθεια της εντολής bugs στην R. Βασικός μας στόχος είναι να αποκτήσουμε τιμές της μεταβλητής απόκρισης Y, τις οποίες μπορούμε να τις αντικαταστήσουμε στη θέση των ελλιπών τιμών που δημιουργήθηκαν μέσω ενός MCAR μηχανισμού. Για να το καταφέρουμε αυτό, αρκεί να προσδιορίσουμε το $\boldsymbol{Y}$ ως παράμετρο στην εντολή bugs.

Εφαρμογή Μοντέλου αντικατάστασης ελλιπών τιμών στη μεταβλητή απόκρισης

```
#MCAR for Responce missing values
#Initial values and data values
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2)}
data <- list(y=MCARDataset$Y, x
    =as.matrix(MCARDataset[,2:13]),p=12, n=80)
#We run the imputationresponse.txt
MCARImputationResponse <- bugs(inits=inits,data=data,model.file =
"imputationresponse.txt",parameters = c("beta0","beta", "gamma",
"sigma","y"),n.chains = 1, n.burnin = 2000, n.iter
= 7000,n.thin = 1,n.sims = 5000,bugs.directory = "c:/Program
Files/WinBUGS14/",debug=FALSE)
MCARImputationResponse
colMeans(MCARImputationResponse$sims.array[1:5000,1,])
#The imputed values we estimated
colMeans(MCARImputationResponse$sims.array[1:5000,1,15:
(sum(is.na(MCARDataset$Y))+14)])
```

Έχοντας αποκτήσει τις τιμές της μεταβλητής Y, μπορούμε να τις αντικαταστήσουμε στις ελλιπείς τιμές κάνοντας χρήση του παρακάτω κώδικα στην R,

Αντικατάσταση των ελλιπών τιμών στη μεταβλητή απόκρισης

```
MCARIMPUTEDDataset <- MCARDataset
#We impute the values of Y
MCARIMPUTEDDataset$Y <-
    replace(MCARIMPUTEDDataset$Y,is.na(MCARIMPUTEDDataset$Y),
as.vector(colMeans(MCARImputationResponse$sims.array[1:5000,
1,15:(sum(is.na(MCARDataset$Y))+14)])))
```

όπου στο MCARIMPUTEDDataset έχουμε αντικαταστήσει τις ελλιπείς τιμές με τις τιμές που βρήκαμε μέσω της εκ των υστέρων προβλεπτικής κατανομής. Συνεπώς στις παρακάτω υποενότητες θα εφαρμόσουμε και θα συγκρίνουμε τα αποτελέσματα των μεθόδων SSVS, Kuo & Mallick, GVS και full enumeration στο δείγμα MCARIMPUTEDDataset.

5.3.1.1 Εφαρμογή της μεθόδου SSVS

Για να εφαρμόσουμε τη μέθοδο SSVS, θα χρησιμοποιήσουμε το αρχείο SSVS.txt που διατυπώσαμε στην Ενότητα 5.2.1. Με τη χρήση του παρακάτω κώδικα, υλοποιούμε τη μέθοδο SSVS και έτσι καταλήγουμε στους Πίνακες 5.3.1.1.1 και 5.3.1.1.2. Αυτοί οι πίνακες περιέχουν τους περιγραφικούς δείκτες των προσομοιωμένων τιμών των εκ των υστέρων κατανομών των παραμέτρων του μοντέλου, του διανύσματος $\boldsymbol{\gamma}$ και τις πιθανότητες των καλύτερων μοντέλων της μεθόδου SSVS.

Εφαρμογή της μεθόδου SSVS

```
#Initial values and dataset
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=MCARIMPUTEDDataset$Y, x
    =as.matrix(MCARIMPUTEDDataset[,2:13]),p=12, n=80, models=4096)
#We implement the ssvs method
MCARmodelssvs <- bugs(inits=inits,data=data,model.file =
    "SSVS.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 12000,n.thin = 5,n.sims = 10000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=FALSE)
MCARmodelparametersssvs <- MCARmodelssvs$summary[1:26,]
MCARmodelparametersssvs
MCARmodelsssvs <- MCARmodelssvs$summary[27:4122,]
MCARmodelsssvs
#We print the first 10 models with the highest posterior probability
head(MCARmodelsssvs[order(MCARmodelsssvs[, 2], decreasing=TRUE),
    ],n=10)
```

Από τον παραπάνω κώδικα παρατηρούμε ότι το μόνο πράγμα που αλλάζει με τον κώδικα της Ενότητας 5.2.1 είναι είναι το δείγμα (αντί για το πλήρες δείγμα, έχουμε το MCARIMPUTEDDataset).

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.455	0.09313	0.002478	2.27	2.455	2.637
beta[1]	−0.2901	0.1028	0.00234	−0.4858	−0.2884	−0.08864
beta[2]	−0.07899	0.1107	0.002303	−0.3025	−0.0794	0.1305
beta[3]	2.984	0.1279	0.003088	2.728	2.981	3.233
beta[4]	−0.09416	0.106	0.002202	−0.2997	−0.09359	0.1255
beta[5]	−1.664	0.1076	0.00226	−1.876	−1.664	−1.449
beta[6]	−0.1761	0.1243	0.00316	−0.4173	−0.178	0.07107
beta[7]	0.2057	0.1419	0.003623	−0.07791	0.2035	0.4843
beta[8]	−0.05688	0.1027	0.002027	−0.2524	−0.05801	0.1482
beta[9]	0.8544	0.1313	0.002865	0.606	0.8533	1.123
beta[10]	−0.0399	0.1257	0.002786	−0.2907	−0.03828	0.1997
beta[11]	0.7345	0.1178	0.00243	0.5049	0.7322	0.9763
beta[12]	0.09612	0.1238	0.002791	−0.1454	0.0956	0.3338
sigma	0.81	0.06916	0.001556	0.6917	0.8068	0.9568
Δείκτες γ						
gamma[1]	0.0175	0.1311	0.002644	0.0	0.0	0.0
gamma[2]	0.0115	0.1066	0.0022	0.0	0.0	0.0
gamma[3]	1.0	0.0	$2.236E-12$	1.0	1.0	1.0
gamma[4]	0.012	0.1089	0.002972	0.0	0.0	0.0
gamma[5]	1.0	0.0	$2.236E-12$	1.0	1.0	1.0
gamma[6]	0.0175	0.1311	0.003174	0.0	0.0	0.0
gamma[7]	0.01	0.0995	0.002212	0.0	0.0	0.0
gamma[8]	0.012	0.1089	0.002792	0.0	0.0	0.0
gamma[9]	0.3235	0.4678	0.01026	0.0	0.0	1.0
gamma[10]	0.01	0.0995	0.002181	0.0	0.0	0.0
gamma[11]	0.167	0.373	0.00684	0.0	0.0	1.0
gamma[12]	0.0135	0.1154	0.002511	0.0	0.0	0.0

Πίνακας 5.3.1.1.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MCAR).

Από τον Πίνακα 5.3.1.1.1 βλέπουμε ότι ο συντελεστής β_1 και οι συντελεστές του

πραγματικού μοντέλου είναι σημαντικοί (τα διαστήματα αξιοπιστίας δεν περιέχουν την τιμή μηδέν). Στους δείκτες $\boldsymbol{\gamma}$ έχουμε ότι $\gamma_3, \gamma_5 = 1$, ενώ $\gamma_9 = 0.3235$ και $\gamma_{11} = 0.167$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5$	21	0.5060	0.5000
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	277	0.2500	0.4331
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_{11}$	1045	0.1015	0.3020
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.0465	0.2106
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6$	53	0.0085	0.0918
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5$	22	0.0080	0.0891
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_{12}$	2069	0.0075	0.0862
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5$	23	0.0070	0.0833
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7$	85	0.0060	0.0772
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5$	29	0.0050	0.0705

Πίνακας 5.3.1.1.2: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MCAR).

Από τον Πίνακα 5.3.1.1.2 διαπιστώνουμε ότι το μοντέλο με τη μεγαλύτερη εκ των υστέρων πιθανότητα αντιστοιχεί στο μοντέλο με τους συντελεστές β_3 και β_5, ενώ στο πραγματικό μοντέλο βλέπουμε ότι έχουμε εκ των υστέρων πιθανότητα ίση με 0.0465.

5.3.1.2 Εφαρμογή της μεθόδου Kuo & Mallick

Για να εφαρμόσουμε τη μέθοδο Kuo & Mallick θα χρησιμοποιήσουμε το αρχείο KuoMallick.txt που δώσαμε στην Ενότητα 5.2.2.

Εφαρμογή της μεθόδου Kuo & Mallick

```
#Initial values and data values
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
   tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=MCARIMPUTEDDataset$Y, x
   =as.matrix(MCARIMPUTEDDataset[,2:13]),p=12, n=80, models=4096)
#We implement the Kuo & Mallick sampler
MCARmodelKuoMallick <- bugs(inits=inits,data=data,model.file =
   "KuoMallick.txt",parameters = c("beta0","beta", "gamma",
   "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
   = 5000,n.thin = 1,n.sims = 3000,bugs.directory = "c:/Program
   Files/WinBUGS14/",debug=FALSE)
MCARmodelKuoMallick
MCARmodelparametersKuoMallick <- MCARmodelKuoMallick$summary[1:26,]
MCARmodelparametersKuoMallick
MCARmodelsKuoMallick <- MCARmodelKuoMallick$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
head(MCARmodelsKuoMallick[order(MCARmodelsKuoMallick[, 2],
   decreasing=TRUE), ],n=10)
```

Εφαρμόζουμε τη μέθοδο Kuo & Mallick με χρήση του παραπάνω κώδικα και έτσι καταλήγουμε στους Πίνακες 5.3.1.2.1 και 5.3.1.2.2.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.455	0.09271	0.001605	2.279	2.454	2.637
beta[1]	−0.01434	97.17	1.699	−194.9	−0.2665	192.4
beta[2]	−2.281	99.66	1.554	−203.3	−1.101	191.8
beta[3]	2.896	0.1006	0.002049	2.701	2.894	3.097
beta[4]	0.4464	99.25	1.979	−194.6	0.7797	191.5
beta[5]	−1.742	0.1058	0.002466	−1.947	−1.742	−1.534
beta[6]	−2.316	102.2	2.049	−213.1	−0.7495	192.2
beta[7]	0.4027	100.9	2.04	−198.9	−0.7733	200.7
beta[8]	−0.1083	98.55	1.926	−186.6	−1.573	203.3
beta[9]	0.9867	0.1001	0.001766	0.7918	0.9875	1.184
beta[10]	1.369	99.66	1.727	−196.1	1.966	191.5
beta[11]	0.652	0.1085	0.002419	0.4394	0.6521	0.8674
beta[12]	0.07807	100.6	1.761	−199.5	0.6774	197.2
sigma	0.8544	0.07175	0.001678	0.73	0.8508	1.004
Δείκτες γ						
gamma[1]	0.068	0.2517	0.02465	0.0	0.0	1.0
gamma[2]	0.001667	0.04079	0.00135	0.0	0.0	0.0
gamma[3]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[4]	0.0	0.0	$1.826E-12$	0.0	0.0	0.0
gamma[5]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[6]	0.00333	0.05764	0.001421	0.0	0.0	0.0
gamma[7]	0.001333	0.03649	$7.932E-4$	0.0	0.0	0.0
gamma[8]	$6.667E-4$	0.02581	$6.578E-4$	0.0	0.0	0.0
gamma[9]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[10]	0.002	0.04468	0.001395	0.0	0.0	0.0
gamma[11]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[12]	$6.667E-4$	0.02581	$6.7E-4$	0.0	0.0	0.0

Πίνακας 5.3.1.2.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ του δειγματολήπτη Kuo & Mallick σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MCAR).

Στον Πίνακα 5.3.1.2.1 παρατηρούμε ότι οι συντελεστές που αντιστοιχούν στους συντελεστές του πραγματικού μοντέλου είναι σημαντικοί, ενώ στους δείκτες $\boldsymbol{\gamma}$ έχουμε ότι $\gamma_3, \gamma_5, \gamma_9, \gamma_{11} = 1$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.9236	0.2655
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1302	0.0666	0.2494
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1333	0.0020	0.0446
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10} + \boldsymbol{X}_{11}$	1813	0.0020	0.0446
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1303	0.0016	0.0407
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1334	0.0013	0.0364
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1365	0.0013	0.0364
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1429	0.0006	0.0258
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3349	0.0006	0.0258
Intercept	1	0.0000	0.0000

Πίνακας 5.3.1.2.2: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων του δειγματολήπτη Kuo & Mallick σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MCAR).

Στον Πίνακα 5.3.1.2.2 έχουμε το μοντέλο με τη μεγαλύτερη εκ των υστέρων πιθανότητα (0.9236) αντιστοιχεί στο πραγματικό μοντέλο. Ως εκ τούτου, η μέθοδος Kuo & Mallick κατάφερε να ανιχνεύσει το πραγματικό μοντέλο.

5.3.1.3 Εφαρμογή της μεθόδου GVS

Προτού εφαρμόσουμε τη μέθοδο GVS, θα χρειαστεί να υλοποιήσουμε αρχικά το

μοντέλο που περιέχει όλες τις επεξηγηματικές μεταβλητές (όπως και στην Ενότητα 5.2.3), με στόχο να πάρουμε τις εκτιμήσεις των παραμέτρων της ψευδο-πρότερης εκ των προτέρων κατανομής (Πίνακας 5.3.1.3.1).

Κώδικας εκτέλεσης του Full μοντέλου

```
#We list our initial values and the dataset
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
   tau=0.2)}
data <- list(y=MCARIMPUTEDDataset$Y, x
   =as.matrix(MCARIMPUTEDDataset[,2:13]),p=12, n=80)
MCARmodelpilotrun <- bugs(inits=inits, data=data, model.file =
   "Pilotrun.txt",parameters = c("beta0","beta", "sigma"),n.chains
   = 1, n.burnin = 2000, n.iter = 7000,n.thin = 1,n.sims =
   5000,bugs.directory = "c:/ProgramFiles/WinBUGS14/",debug=FALSE)
```

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.458	0.0917	0.001278	2.278	2.458	2.637
beta[1]	−0.349	0.1106	0.001441	−0.5625	−0.3488	−0.1336
beta[2]	−0.05577	0.1229	0.001497	−0.2999	−0.05493	0.1835
beta[3]	3.025	0.1364	0.002007	2.758	3.023	3.294
beta[4]	−0.131	0.1123	0.001323	−0.3561	−0.1289	0.09427
beta[5]	−1.677	0.1118	0.001509	−1.892	−1.678	−1.457
beta[6]	−0.2273	0.1368	0.002109	−0.4893	−0.2286	0.04904
beta[7]	0.3052	0.1652	0.002131	−0.02397	0.3053	0.6236
beta[8]	−0.04147	0.1117	0.001252	−0.2635	−0.04017	0.1785
beta[9]	0.8819	0.1285	0.001693	0.6313	0.8822	1.136
beta[10]	−0.1237	0.1446	0.002057	−0.4136	−0.1256	0.1619
beta[11]	0.826	0.1221	0.001541	0.5825	0.827	1.063
beta[12]	0.07494	0.1433	0.00199	−0.2085	0.0736	0.3576

Πίνακας 5.3.1.3.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των συντελεστών του πλήρες μοντέλου.

Εφαρμογή της μεθόδου GVS

```
#We list our initial values and the dataset
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
   tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=MCARIMPUTEDDataset$Y, x
   =as.matrix(MCARIMPUTEDDataset[,2:13]),p=12,n=80,models=4096,mean=
as.vector(colMeans(MCARmodelpilotrun$sims.array[1:5000,1,]))[2:13],
se=as.vector(MCARmodelpilotrun$sd$beta))
MCARmodelgvs <- bugs(inits=inits,data=data,model.file =
"GVS.txt",parameters = c("beta0","beta", "gamma",
   "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
   = 5000,n.thin = 1,n.sims = 3000,bugs.directory = "c:/Program
   Files/WinBUGS14/",debug=TRUE)
MCARmodelparametersgvs <- MCARmodelgvs$summary[1:26,]
MCARmodelsgvs <- MCARmodelgvs$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
```

```
head(MCARmodelsgvs[order(MCARmodelsgvs[, 2],decreasing=TRUE),],n=10)
```

Εφαρμόζοντας τον παραπάνω κώδικα στην R καταλήγουμε στους Πίνακες 5.3.1.3.2 και 5.3.1.3.3.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.457	0.09616	0.001686	2.267	2.457	2.647
beta[1]	−0.3461	0.1118	0.001921	−0.5676	−0.345	−0.1298
beta[2]	−0.05757	0.1189	0.002101	−0.2796	−0.05844	0.1743
beta[3]	2.895	0.09885	0.001558	2.698	2.894	3.085
beta[4]	−0.1315	0.1144	0.002158	−0.3577	−0.132	0.09857
beta[5]	−1.744	0.1058	0.002512	−1.947	−1.743	−1.534
beta[6]	−0.2261	0.1371	0.002769	−0.4863	−0.23	0.04916
beta[7]	0.3007	0.1699	0.003536	−0.04694	0.303	0.6236
beta[8]	−0.04111	0.1128	0.002303	−0.2637	−0.04288	0.1828
beta[9]	0.9859	0.09989	0.002187	0.7957	0.9843	1.197
beta[10]	−0.1242	0.1433	0.00247	−0.3992	−0.124	0.1562
beta[11]	0.6552	0.1092	0.002288	0.4427	0.6564	0.8634
beta[12]	0.07094	0.1429	0.002789	−0.2183	0.06968	0.3479
sigma	0.8554	0.07221	0.001386	0.7299	0.8512	1.012
Δείκτες γ						
gamma[1]	0.04833	0.2145	0.004261	0.0	0.0	1.0
gamma[2]	0.001667	0.04079	$7.156E-4$	0.0	0.0	0.0
gamma[3]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[4]	0.001667	0.04079	$7.156E-4$	0.0	0.0	0.0
gamma[5]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[6]	0.005	0.07053	0.00138	0.0	0.0	0.0
gamma[7]	0.001667	0.04079	$8.628E-4$	0.0	0.0	0.0
gamma[8]	$3.333E-4$	0.01825	$3.289E-4$	0.0	0.0	0.0
gamma[9]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[10]	0.001	0.03161	$5.66E-4$	0.0	0.0	0.0
gamma[11]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[12]	$6.667E-4$	0.02581	$4.694E-4$	0.0	0.0	0.0

Πίνακας 5.3.1.3.2: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MCAR).

Στον Πίνακα 5.3.1.3.2 βλέπουμε ότι οι συντελεστές που αντιστοιχούν στους συντελεστές του πραγματικού μοντέλου είναι σημαντικοί και οι δείκτες $\gamma_3, \gamma_5, \gamma_9, \gamma_{11} = 1$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.9403	0.2369
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1302	0.0480	0.2138
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1333	0.0046	0.0681
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1303	0.0016	0.0407
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1365	0.0016	0.0407
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1309	0.0010	0.0316
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10} + \boldsymbol{X}_{11}$	1813	0.0010	0.0316
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3349	0.0006	0.0258
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1310	0.0003	0.0182
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1341	0.0003	0.0182

Πίνακας 5.3.1.3.3: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MCAR).

Από τον Πίνακα 5.3.1.3.3 καταλαβαίνουμε ότι η μέθοδος GVS κατάφερε να ανιχνεύσει το πραγματικό μοντέλο.

5.3.1.4 Πλήρης απαρίθμηση με τη χρήση του κριτηρίου BIC

Για να εφαρμόσουμε πλήρη απαρίθμηση, θα χρησιμοποιήσουμε τη συνάρτηση BICminimum που κατασκευάσαμε στην R.

Εφαρμογή της συνάρτησης BICminimum στην R

```
Lowest10BICmodelsMCAR <-BICminimum(p=12,MCARIMPUTEDDataset)
#Coefficients of the models
Lowest10BICmodelsMCAR[[1]][1:10]
#and their corresponding BIC values.
Lowest10BICmodelsMCAR[[2]][1:10]
#MOdel code
Lowest10BICmodelsMCAR[[3]]
```

Εφαρμόζοντας τη συνάρτηση BICminimum καταλήγουμε στον Πίνακα 5.3.1.4.1.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Κριτήριο BIC
$X_1 + X_3 + X_5 + X_9 + X_{11}$	1302	217.3453
$X_1 + X_3 + X_5 + X_6 + X_9 + X_{11}$	1334	220.1944
$X_1 + X_2 + X_3 + X_5 + X_9 + X_{11}$	1304	220.6633
$X_1 + X_3 + X_5 + X_9 + X_{11} + X_{12}$	3350	220.8007
$X_1 + X_3 + X_4 + X_5 + X_9 + X_{11}$	1310	220.8404
$X_1 + X_3 + X_5 + X_8 + X_9 + X_{11}$	1430	221.0908
$X_1 + X_3 + X_5 + X_7 + X_9 + X_{11}$	1366	221.1472
$X_1 + X_3 + X_5 + X_9 + X_{10} + X_{11}$	1814	221.1606
$X_3 + X_5 + X_9 + X_{11}$	1301	221.4986
$X_1 + X_3 + X_5 + X_6 + X_7 + X_9 + X_{11}$	1398	221.7615

Πίνακας 5.3.1.4.1: Τα 10 πρώτα μοντέλα με τις μικρότερες τιμές του κριτηρίου BIC σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MCAR).

Από το Πίνακα 5.3.1.4.1 (όπως και στον Πίνακα 5.2.3.4) βλέπουμε ότι το μοντέλο με τη μικρότερη τιμή του κριτηρίου BIC, αντιστοιχεί στο μοντέλο που περιέχει τους συντελεστές του πραγματικού μοντέλου μαζί με το συντελεστή β_1, ενώ το ένατο μοντέλο με τη μικρότερη τιμή BIC αντιστοιχεί στο πραγματικό μοντέλο.

Μοντέλο με τη μικρότερη τιμή του κριτηρίου BIC

```
#We take the summary of the model with lowest BIC value
summary(lm(Y~ X1+X3+X5+X9+X11,data=MCARIMPUTEDDataset))
#Partial console output
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.14100    0.09442 22.676 < 2e-16 ***
X1         -0.26919    0.09326 -2.886  0.0051 **
X3          2.78919    0.08932 31.228 < 2e-16 ***
X5         -1.67184    0.10038 -16.655 < 2e-16 ***
X9          1.01093    0.09660 10.465 3.02e-16 ***
X11         0.63960    0.09407  6.799 2.32e-09 ***
---
```

5.3.2 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές υπό το μηχανισμό MAR

Στην κατασκευή του μηχανισμού έλλειψης MAR θα πρέπει η πιθανότητα έλλειψης μιας παρατήρησης στη μεταβλητή απόκρισης να εξαρτάται από το παρατηρήσιμο κομμάτι του δείγματος. Με αυτό το τρόπο η δημιουργία του μηχανισμού MAR που θα φτιάξουμε στην R πρέπει να φτιαχτεί με τέτοιο τρόπο ώστε να αντικατοπτρίζει τη δέσμευση που ορίσαμε (για το μηχανισμό MAR) στην Ενότητα 4.2. Έτσι ξεκινάμε δημιουργώντας στην R ένα μηχανισμό έλλειψης MAR για τη μεταβλητή απόκρισης.

Κατασκευή μηχανισμού MAR στη μεταβλητή απόκρισης

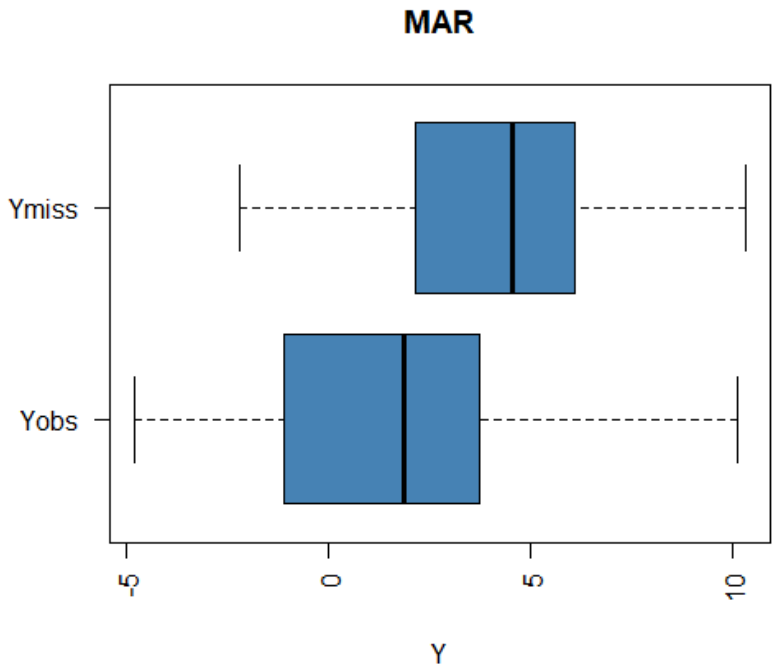
```
MARDataset <- FullDataset
#We create an MAR missing mechanism which depends on X10 and X5.
sort.X10X5 <- sort(FullDataset$X10-FullDataset$X5, decreasing=TRUE)
mar <- quantile(sort.X10X5, probs =0.65)
# if X10-X5 > mar then the observation has a missing probability of
   0.85.
y.mar <- ifelse(FullDataset$X10-FullDataset$X5>=mar, rbinom(80, 1,
   0.15), 1)
y.mar
# # of observations that remain observed after the MAR mechanism
#console output is 55
sum(y.mar==1)
# if y.mar=1 then the observation remains observed
#and if y.mar=0 then it becomes a missing value.
MARDataset$Y <- ifelse(y.mar==1,MARDataset$Y,NA)
#Number of missing values in y.
#console output is 25.
sum(is.na(MARDataset$Y))
```

Ο παραπάνω μηχανισμός που φτιάξαμε στην R βλέπουμε ότι εξαρτάται από τη διαφορά των επεξηγηματικών μεταβλητών X_{10} και X_5. Συγκεκριμένα όταν η διαφορά αυτών των δύο μεταβλητών ξεπερνά ένα οριοθετημένο κατώφλι, τότε η πιθανότητα έλλειψης της τιμής της μεταβλητής απόκρισης είναι 0.85, ενώ αν η διαφορά αυτή είναι μικρότερη από το κατώφλι τότε η τιμή της Y παραμένει παρατηρήσιμη με πιθανότητα 1.

Κατασκευή θηκογραμμάτων

```
#These are the observations that we took as missing
YmissMAR <- FullDataset$Y[is.na(MARDataset$Y)]
#Boxplot within the missing and the remaining observations.
boxplot(MARDataset$Y, YmissMAR, horizontal=TRUE, col='steelblue',
las=2,main = "MAR", xlab = "Y", names = c("Yobs","Ymiss"))
```

Υλοποιούμε τον παραπάνω κώδικα στην R και καταλήγουμε στο Διάγραμμα 5.3.2.1 όπου απεικονίζει τα Θηκογράμματα των τιμών των Y_{miss} και Y_{obs}.



Διάγραμμα 5.3.2.1: Θηκογράμματα των τιμών του Y_{miss} και του Y_{obs}.

Από το Διάγραμμα 5.3.2.1 παρατηρούμε ότι η διάμεσος των τιμών του Y_{miss} ξεπερνά το 3ο τεταρτημόριο των τιμών του Y_{obs}, οπότε είναι πιθανό να υπάρχει στατιστικά σημαντική διαφορά μεταξύ των Y_{miss} και Y_{obs}. Για να αντικαταστήσουμε αυτές τις τιμές θα χρησιμοποιήσουμε το μοντέλο που έχουμε διατυπώσει στην Ενότητα 5.3.1 (imputationresponse.txt).

Εφαρμογή Μοντέλου αντικατάστασης ελλιπών τιμών στη μεταβλητή απόκρισης

```
#MAR for Responce missing values
#Initial values and data inputs
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2)}
data <- list(y=MARDataset$Y, x =as.matrix(MARDataset[,2:13]),p=12,
    n=80)
#We run the imputationresponse.txt
MARImputationResponse <- bugs(inits=inits,data=data,model.file =
"imputationresponse.txt",parameters = c("beta0","beta", "gamma",
"sigma","y"),n.chains = 1, n.burnin = 2000, n.iter= 7000,n.thin = 1,
n.sims = 5000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=FALSE)
colMeans(MARImputationResponse$sims.array[1:5000,1,])
#The imputed values we estimated
colMeans(MARImputationResponse$sims.array[1:5000,1,15:
(sum(is.na(MARDataset$Y))+14)])
```

Συνεχίζουμε αντικαθιστώντας τις ελλιπείς τιμές κάνοντας χρήση του παρακάτω κώδικα στην R.

Αντικατάσταση των ελλιπών τιμών στη μεταβλητή απόκρισης

```
MARIMPUTEDDataset <- MARDataset
#We impute the values of Y
MARIMPUTEDDataset$Y <-
    replace(MARIMPUTEDDataset$Y,is.na(MARIMPUTEDDataset$Y),
as.vector(colMeans(MARImputationResponse$sims.array[1:5000,
1,15:(sum(is.na(MARDataset$Y))+14)])))
```

Ως εκ τούτου στις παρακάτω υποενότητες θα εφαρμόσουμε τις μεθόδους SSVS, Kuo & Mallick, GVS full enumeration και θα συγκρίνουμε τα αποτελέσματά τους.

5.3.2.1 Εφαρμογή της μεθόδου SSVS

Ξεκινάμε εφαρμόζοντας τη μέθοδο SSVS χρησιμοποιώντας το αρχείο SSVS.txt που θεωρήσαμε στην Ενότητα 5.2.1.

Εφαρμογή της μεθόδου SSVS

```
#Initial values
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
# dataset
data <- list(y=MARIMPUTEDDataset$Y, x
    =as.matrix(MARIMPUTEDDataset[,2:13]),p=12, n=80, models=4096)
#We implement the ssvs method
MARmodelssvs <- bugs(inits=inits,data=data,model.file =
    "SSVS.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 12000,n.thin = 5,n.sims = 10000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=FALSE)
MARmodelssvs
MARmodelparametersssvs <- MARmodelssvs$summary[1:26,]
MARmodelparametersssvs
MARmodelsssvs <- MARmodelssvs$summary[27:4122,]
MARmodelsssvs
#We print the first 10 models with the highest posterior probability
head(MARmodelsssvs[order(MARmodelsssvs[, 2], decreasing=TRUE),
    ],n=10)
```

Με χρήση του παραπάνω κώδικα στην R (χρήση της εντολής bugs της βιβλιοθήκης R2WinBUGS), υλοποιούμε στο WinBUGS τη μέθοδο SSVS και έτσι καταλήγουμε στους Πίνακες 5.3.2.1.1 και 5.3.2.1.2, που περιέχουν τους περιγραφικούς δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των συντελεστών του μοντέλου, τους περιγραφικούς δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των $\boldsymbol{\gamma}$ και

τα μοντέλα με τις μεγαλύτερες εκ των υστέρων πιθανότητες αντίστοιχα.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.25	0.0988	0.00263	2.052	2.25	2.442
beta[1]	−0.3684	0.1083	0.002468	−0.5734	−0.3671	−0.1547
beta[2]	0.09646	0.1162	0.002401	−0.1389	0.09599	0.3175
beta[3]	2.863	0.1343	0.003192	2.591	2.859	3.129
beta[4]	0.01574	0.1111	0.00232	−0.2023	0.01686	0.2477
beta[5]	−1.637	0.1137	0.002395	−1.861	−1.636	−1.411
beta[6]	−0.01803	0.1298	0.003255	−0.269	−0.0202	0.2383
beta[7]	0.06207	0.1466	0.003669	−0.2316	0.05934	0.3512
beta[8]	−0.2326	0.1083	0.002113	−0.439	−0.2339	−0.01386
beta[9]	0.8505	0.1388	0.00294	0.5881	0.8482	1.132
beta[10]	−0.1656	0.1315	0.002836	−0.4288	−0.1631	0.08466
beta[11]	0.57	0.1169	0.002365	0.343	0.5714	0.8024
beta[12]	0.049	0.1293	0.002901	−0.2077	0.04727	0.2999
sigma	0.8596	0.07308	0.00166	0.7346	0.8555	1.016
Δείκτες γ						
gamma[1]	0.0225	0.1483	0.0031	0.0	0.0	0.0
gamma[2]	0.0125	0.1111	0.002308	0.0	0.0	0.0
gamma[3]	1.0	0.0	$2.236E-12$	1.0	1.0	1.0
gamma[4]	0.0105	0.1019	0.002616	0.0	0.0	0.0
gamma[5]	0.9995	0.02236	$4.914E-4$	1.0	1.0	1.0
gamma[6]	0.0145	0.1195	0.002772	0.0	0.0	0.0
gamma[7]	0.009	0.09444	0.002073	0.0	0.0	0.0
gamma[8]	0.0165	0.1274	0.003117	0.0	0.0	0.0
gamma[9]	0.3255	0.4686	0.009766	0.0	0.0	1.0
gamma[10]	0.012	0.1089	0.002402	0.0	0.0	0.0
gamma[11]	0.0605	0.2384	0.005075	0.0	0.0	1.0
gamma[12]	0.0135	0.1154	0.002602	0.0	0.0	0.0

Πίνακας 5.3.2.1.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MAR).

Στον παραπάνω πίνακα παρατηρούμε ότι οι συντελεστές β_1, β_8 και οι συντελεστές του πραγματικού μοντέλο είναι σημαντικοί, καθώς στα διαστήματα αξιοπιστίας δεν υπάρχει η τιμή μηδέν. Ωστόσο στους δείκτες $\boldsymbol{\gamma}$ βλέπουμε ότι τα γ_3, γ_5 είναι περίπου ένα, ενώ $\gamma_9 = 0.3255$ με αρκετά μεγάλη τυπική απόκλιση και $\gamma_{11} = 0.0605$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5$	21	0.5670	0.4956
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	277	0.2780	0.4481
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_{11}$	1045	0.0345	0.1825
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.0185	0.1347
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5$	22	0.0125	0.1111
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5$	23	0.0085	0.0918
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6$	53	0.0085	0.0918
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8$	149	0.0085	0.0918
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_{12}$	2069	0.0080	0.0891
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_{10}$	533	0.0065	0.0803

Πίνακας 5.3.2.1.2: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MAR).

Στον Πίνακα 5.3.2.1.2 βλέπουμε ότι πάλι το μοντέλο που έχει τη μεγαλύτερη εκ των υστέρων πιθανότητα στη μέθοδο SSVS είναι το μοντέλο που περιέχει τους συντελεστές β_3 και β_5, ενώ το πραγματικό μοντέλο έχει πιθανότητα ίση με 0.0185. Συνεπώς, η μέθοδος SSVS δεν κατάφερε να ανιχνεύσει το πραγματικό μοντέλο.

5.3.2.2 Εφαρμογή της μεθόδου Kuo & Mallick

Εφαρμόζουμε το δειγματολήπτη Kuo & Mallick κάνοντας χρήση του αρχείου KuoMallick.txt που διατυπώσαμε στην Ενότητα 5.2.2.

Εφαρμογή της μεθόδου Kuo & Mallick

```
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=MARIMPUTEDDataset$Y, x
    =as.matrix(MARIMPUTEDDataset[,2:13]),p=12, n=80, models=4096)
MARmodelKuoMallick <- bugs(inits=inits,data=data,model.file =
    "KuoMallick.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 5000,n.thin = 1,n.sims = 3000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=FALSE)
MARmodelparametersKuoMallick <- MARmodelKuoMallick$summary[1:26,]
MARmodelsKuoMallick <- MARmodelKuoMallick$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
head(MARmodelsKuoMallick[order(MARmodelsKuoMallick[, 2],
    decreasing=TRUE), ],n=10)
```

Εφαρμόζουμε τη μέθοδο Kuo & Mallick με χρήση του παραπάνω κώδικα και έτσι καταλήγουμε στους Πίνακες 5.3.2.2.1 και 5.3.2.2.2.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.25	0.1015	0.001761	2.057	2.249	2.449
beta[1]	0.08535	95.95	1.69	−193.5	−0.3258	191.5
beta[2]	−2.272	99.68	1.553	−203.3	−1.294	191.8
beta[3]	2.838	0.1104	0.002365	2.624	2.836	3.058
beta[4]	0.3693	99.12	1.966	−194.6	0.6471	191.3
beta[5]	−1.793	0.117	0.00295	−2.021	−1.794	−1.564
beta[6]	−2.362	102.4	2.054	−213.1	−0.9348	192.7
beta[7]	0.4584	100.8	2.039	−198.9	−0.6131	200.7
beta[8]	−0.2164	97.98	1.94	−185.9	−0.3627	202.2
beta[9]	0.8295	0.1103	0.00199	0.6158	0.8299	1.049
beta[10]	1.592	99.38	1.711	−196.1	1.151	191.5
beta[11]	0.5672	0.1193	0.002956	0.3334	0.5671	0.8036
beta[12]	0.07804	100.6	1.761	−199.5	0.6774	197.2
sigma	0.9358	0.07937	0.002139	0.796	0.931	1.1
Δείκτες γ						
gamma[1]	0.095	0.2932	0.02974	0.0	0.0	1.0
gamma[2]	$3.333E-4$	0.01825	$3.289E-4$	0.0	0.0	0.0
gamma[3]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[4]	0.001	0.03161	$7.409E-4$	0.0	0.0	0.0
gamma[5]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[6]	0.001667	0.04079	$7.13E-4$	0.0	0.0	0.0
gamma[7]	0.003	0.05469	0.001395	0.0	0.0	0.0
gamma[8]	0.01133	0.1059	0.007326	0.0	0.0	0.0
gamma[9]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[10]	0.011	0.1043	0.006285	0.0	0.0	0.0
gamma[11]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[12]	$6.667E-4$	0.02581	$6.7E-4$	0.0	0.0	0.0

Πίνακας 5.3.2.2.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ του δειγματολήπτη Kuo & Mallick σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MAR).

Στον Πίνακα 5.3.2.2.1 παρατηρούμε ότι οι συντελεστές που αντιστοιχούν στους συντελεστές του πραγματικού μοντέλου είναι σημαντικοί, ενώ στους δείκτες $\boldsymbol{\gamma}$ έχουμε ότι $\gamma_3, \gamma_5, \gamma_9, \gamma_{11} = 1$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.8770	0.3284
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1302	0.0940	0.2918
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1429	0.0113	0.1058
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10} + \boldsymbol{X}_{11}$	1813	0.0110	0.1043
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1365	0.0030	0.0546
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1309	0.0010	0.0316
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1334	0.0010	0.0316
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1333	0.0006	0.0258
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3349	0.0006	0.0258
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1303	0.0003	0.0182

Πίνακας 5.3.2.2.2: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων του δειγματολήπτη Kuo & Mallick σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MAR).

Από τον Πίνακα 5.3.2.2.2 βλέπουμε ότι ο δειγματολήπτης Kuo & Mallick κατάφερε να ανιχνεύσει το πραγματικό μοντέλο, καθώς το μοντέλο με τη μεγαλύτερη εκ των υστέρων πιθανότητα αντιστοιχεί στο πραγματικό μοντέλο.

5.3.2.3 Εφαρμογή της μεθόδου GVS

Ξεκινάμε υλοποιώντας το πλήρες μοντέλο με στόχο να πάρουμε τις εκτιμήσεις των παραμέτρων της ψευδο-πρότερης εκ των προτέρων κατανομής (Πίνακας 5.3.2.3.1).

Κώδικας εκτέλεσης του Full μοντέλου

```
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0), tau=0.2)}
data <-list(y=MARIMPUTEDDataset$Y, x
    =as.matrix(MARIMPUTEDDataset[,2:13]),p=12, n=80)
MARmodelpilotrun <- bugs(inits=inits, data=data, model.file =
    "Pilotrun.txt",parameters = c("beta0","beta", "sigma"),n.chains = 1,
    n.burnin = 2000, n.iter = 7000,n.thin = 1,n.sims = 5000,bugs.directory
    = "c:/Program Files/WinBUGS14/",debug=FALSE)
```

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.253	0.09742	0.001358	2.061	2.252	2.443
beta[1]	0.1403	0.1305	0.00159	−0.119	0.1412	0.3946
beta[2]	0.1403	0.1305	0.00159	−0.119	0.1412	0.3946
beta[3]	2.884	0.1449	0.002132	2.601	2.883	3.17
beta[4]	0.001124	0.1193	0.001405	−0.2381	0.003277	0.2404
beta[5]	−1.64	0.1188	0.001603	−1.867	−1.64	−1.405
beta[6]	−0.03308	0.1453	0.00224	−0.3114	−0.03446	0.2604
beta[7]	0.1247	0.1755	0.002264	−0.225	0.1248	0.4629
beta[8]	−0.2419	0.1187	0.00133	−0.4778	−0.2405	−0.008154
beta[9]	0.9141	0.1365	0.001798	0.6479	0.9144	1.183
beta[10]	−0.2604	0.1536	0.002185	−0.5684	−0.2624	0.04296
beta[11]	0.643	0.1297	0.001637	0.3843	0.644	0.8949
beta[12]	0.02913	0.1522	0.002114	−0.2719	0.0277	0.3294

Πίνακας 5.3.2.3.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των συντελεστών του πλήρες μοντέλου.

Εφαρμογή της μεθόδου GVS

```
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=MARIMPUTEDDataset$Y, x
    =as.matrix(MARIMPUTEDDataset[,2:13]),p=12,n=80,models=4096,mean=
as.vector(colMeans(MARmodelpilotrun$sims.array[1:5000,1,]))[2:13],
se=as.vector(MARmodelpilotrun$sd$beta))
MARmodelgvs <- bugs(inits=inits,data=data,model.file =
"GVS.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 5000,n.thin = 1,n.sims = 3000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=TRUE)
#Here we have the summary of the model parameters
MARmodelparametersgvs <- MARmodelgvs$summary[1:26,]
MARmodelsgvs <- MARmodelgvs$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
head(MARmodelsgvs[order(MARmodelsgvs[, 2], decreasing=TRUE), ],n=10)
```

Εφαρμόζοντας τον παραπάνω κώδικα στην R καταλήγουμε στους Πίνακες 5.3.2.3.2 και 5.3.1.3.3.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.251	0.1056	0.00186	2.042	2.252	2.461
beta[1]	−0.4326	0.1208	0.002256	−0.6699	−0.4309	−0.1973
beta[2]	0.1383	0.1265	0.00223	−0.09742	0.1375	0.3847
beta[3]	2.835	0.1103	0.001853	2.61	2.834	3.046
beta[4]	$4.108E-4$	0.1216	0.002294	−0.2397	$-1.387E-4$	0.245
beta[5]	−1.788	0.1234	0.003243	−2.019	−1.788	−1.532
beta[6]	−0.0322	0.1456	0.00296	−0.3088	−0.03614	0.2606
beta[7]	0.1196	0.1805	0.003692	−0.25	0.1222	0.4629
beta[8]	−0.2416	0.1199	0.002464	−0.4779	−0.2434	−0.003629
beta[9]	0.8322	0.1131	0.002777	0.6236	0.8285	1.069
beta[10]	−0.2606	0.152	0.002644	−0.5531	−0.2605	0.03688
beta[11]	0.5746	0.1213	0.002519	0.3381	0.5768	0.8051
beta[12]	0.02488	0.1519	0.002962	−0.2824	0.02354	0.319
sigma	0.9401	0.08387	0.001987	0.7974	0.9352	1.121
Δείκτες γ						
gamma[1]	0.09767	0.2969	0.007811	0.0	0.0	1.0
gamma[2]	0.001667	0.04079	$7.156E-4$	0.0	0.0	0.0
gamma[3]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[4]	$6.667E-4$	0.02581	$4.694E-4$	0.0	0.0	0.0
gamma[5]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[6]	0.002667	0.05157	$8.74E-4$	0.0	0.0	0.0
gamma[7]	0.003333	0.05764	0.001169	0.0	0.0	0.0
gamma[8]	0.01033	0.1011	0.001991	0.0	0.0	0.0
gamma[9]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[10]	0.007667	0.08722	0.001395	0.0	0.0	0.0
gamma[11]	0.9683	0.1751	0.006728	0.0	1.0	1.0
gamma[12]	$6.667E-4$	0.02581	$4.694E-4$	0.0	0.0	0.0

Πίνακας 5.3.2.3.2: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MAR).

Στον Πίνακα 5.3.2.3.2 βλέπουμε ότι οι συντελεστές $\beta_3, \beta_5, \beta_9$ και β_{11} είναι σημαντικοί και $\gamma_3, \gamma_5, \gamma_9 = 1$ και $\gamma_{11} = 0.9683$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1301	0.8526	0.3544
$\boldsymbol{X}_1+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1302	0.0896	0.2857
$\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_9$	277	0.0303	0.1715
$\boldsymbol{X}_1+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_8+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1430	0.0066	0.0813
$\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{10}+\boldsymbol{X}_{11}$	1813	0.0066	0.0813
$\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_8+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1429	0.0036	0.0604
$\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_6+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1333	0.0026	0.0515
$\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_7+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1365	0.0023	0.0482
$\boldsymbol{X}_2+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1303	0.0016	0.0407
$\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_7+\boldsymbol{X}_9$	341	0.0010	0.0316

Πίνακας 5.3.2.3.3: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MAR).

Από τον Πίνακα 5.3.2.3.3 μας είναι φανερό ότι η μέθοδος GVS κατάφερε να ανιχνεύσει το πραγματικό μοντέλο (το μοντέλο με τους συντελεστές $\beta_3, \beta_5, \beta_9$ και β_{11}), καθώς η εκ των υστέρων πιθανότητα είναι ίση με 0.8526.

5.3.2.4 Πλήρης απαρίθμηση με τη χρήση του κριτηρίου BIC

Εφαρμόζουμε πλήρη απαρίθμηση χρησιμοποιώντας τη συνάρτηση BICminimum που έχουμε κατασκευάσει στην R.

Εφαρμογή της συνάρτησης BICminimum στην R

```
Lowest10BICmodelsMAR <-BICminimum(p=12,MARIMPUTEDDataset)
#Coefficients of the models
Lowest10BICmodelsMAR[[1]][1:10]
#and their corresponding BIC values.
Lowest10BICmodelsMAR[[2]][1:10]
#Model code
Lowest10BICmodelsMAR[[3]]
```

Εφαρμόζοντας τη συνάρτηση BICminimum καταλήγουμε στον Πίνακα 5.3.2.4.1.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Κριτήριο BIC
$\boldsymbol{X}_1+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_8+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1430	227.9548
$\boldsymbol{X}_1+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_8+\boldsymbol{X}_9+\boldsymbol{X}_{10}+\boldsymbol{X}_{11}$	1942	229.942
$\boldsymbol{X}_1+\boldsymbol{X}_2+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_8+\boldsymbol{X}_9+\boldsymbol{X}_{11}+\boldsymbol{X}_{12}$	3478	230.3918
$\boldsymbol{X}_1+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{10}+\boldsymbol{X}_{11}$	1814	230.4899
$\boldsymbol{X}_1+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1302	230.7303
$\boldsymbol{X}_1+\boldsymbol{X}_2+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_8+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1432	231.3143
$\boldsymbol{X}_1+\boldsymbol{X}_2+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_8+\boldsymbol{X}_9+\boldsymbol{X}_{10}+\boldsymbol{X}_{11}$	1944	231.9664
$\boldsymbol{X}_1+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_6+\boldsymbol{X}_8+\boldsymbol{X}_9+\boldsymbol{X}_{10}+\boldsymbol{X}_{11}$	1462	232.1592
$\boldsymbol{X}_1+\boldsymbol{X}_3+\boldsymbol{X}_4+\boldsymbol{X}_5+\boldsymbol{X}_8+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1438	232.1716
$\boldsymbol{X}_1+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_7+\boldsymbol{X}_8+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1494	232.2082

Πίνακας 5.3.2.4.1: Τα 10 πρώτα μοντέλα με τις μικρότερες τιμές του κριτηρίου BIC σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (MAR).

Από το Πίνακα 5.3.2.4.1 βλέπουμε ότι το μοντέλο με τη μικρότερη τιμή του κριτηρίου BIC, αντιστοιχεί στο μοντέλο που περιέχει τους συντελεστές του

πραγματικού μοντέλου μαζί με τους συντελεστές β_1 και β_8, ενώ το πραγματικό μοντέλο δε βρίσκεται στον παραπάνω πίνακα.

Μοντέλο με τη μικρότερη τιμή του κριτηρίου BIC

```
#We take the summary of the model with lowest BIC value
summary(lm(Y~ X1+X3+X5+X8+X9+X11,data=MARIMPUTEDDataset))
#Partial console output
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.92339    0.09902 19.425  < 2e-16 ***
X1         -0.37084    0.09954 -3.726 0.000381 ***
X3          2.75980    0.09385 29.407  < 2e-16 ***
X5         -1.61791    0.11139 -14.525 < 2e-16 ***
X8         -0.28800    0.11018 -2.614 0.010867 *
X9          0.81771    0.10192  8.023 1.27e-11 ***
X11         0.59102    0.09884  5.979 7.59e-08 ***
---
```

5.3.3 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές υπό το μηχανισμό NMAR

Στο μηχανισμό έλλειψης NMAR ξέρουμε ότι θα χρειαστεί η πιθανότητα έλλειψης μιας παρατήρησης στη μεταβλητή απόκρισης να εξαρτάται και από το μη παρατηρήσιμο κομμάτι του δείγματος. Δηλαδή ο μηχανισμός έλλειψης NMAR στην R θα πρέπει να φτιαχτεί με τέτοιο τρόπο ώστε η πιθανότητα έλλειψης να εξαρτάται και από το Y_{miss}. Συνεπώς ξεκινάμε δημιουργώντας στην R ένα μηχανισμό έλλειψης NMAR για τη μεταβλητή απόκρισης.

Για τη κατασκευή του μηχανισμού έλλειψης NMAR στη μεταβλητή απόκρισης θα χρησιμοποιήσουμε το λογιστικό μοντέλο. Συγκεκριμένα το μοντέλο για την πιθανότητα παρατήρησης της κάθε τιμής της μεταβλητής απόκρισης $\boldsymbol{Y}$ θα είναι το εξής:

$$\text{logit}\left(\mathbb{P}\left(R_{Y_i}=1\right)\right) = 1 \cdot Y_i - 0.2 \cdot X_{5,i}, \quad i = 1, \ldots, 80,$$

όπου παρατηρούμε ότι η πιθανότητα παρατήρησης κάθε τιμής της μεταβλητής απόκρισης εξαρτάται και από το ίδιο το $\boldsymbol{Y}$ και από την επεξηγηματική μεταβλητή $\boldsymbol{X}_5$ (ενώ R αποτελεί το δείκτη ελλιπών τιμών όπου είδαμε και στην Ενότητα 4.2).

Κατασκευή μηχανισμού NMAR στη μεταβλητή απόκρισης

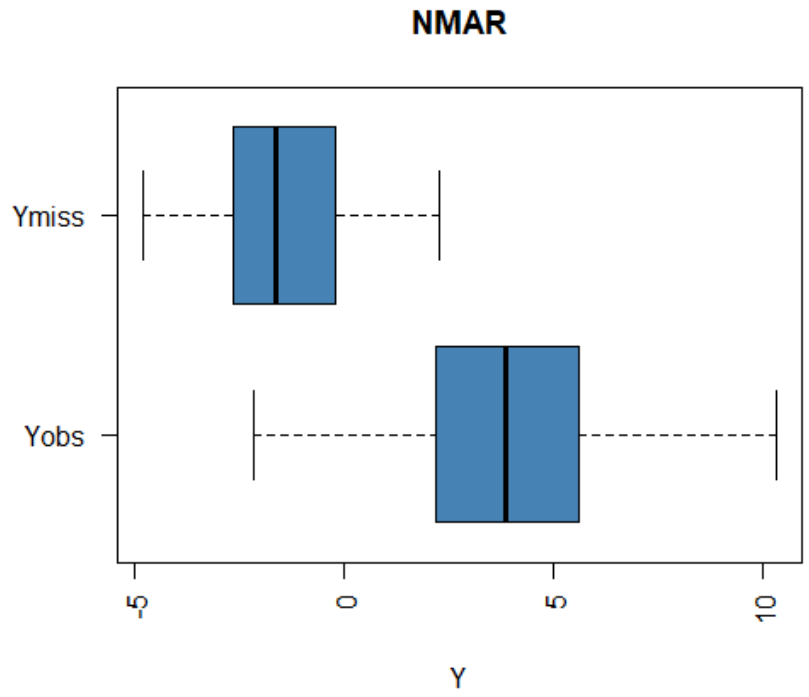
```
NMARDataset <- FullDataset
# We create an NMAR missing mechanism using a logistic model.
logistic <- function(x) exp(x) / (1 + exp(x))
modNMAR <- 1 * FullDataset$Y -0.2 * FullDataset$X5
r2.nmar <-  rbinom(length(FullDataset$Y), 1, logistic(modNMAR))
```

```
# # of observations that remain observed after the NMAR mechanism
sum(r2.nmar==1)
#console output is
#if r2.nmar=0 then the observation becomes a missing value.
NMARDataset$Y[r2.nmar==0] <- NA

# Number of missing values in y.
sum(is.na(NMARDataset$Y))
#console output is 25.
```

Κατασκευή θηκογραμμάτων

```
#These are the observations that we took as missing
YmissNMAR <- FullDataset$Y[is.na(NMARDataset$Y)]
#Boxplot within the missing and the remaining observations.
boxplot(NMARDataset$Y, YmissNMAR, horizontal=TRUE, col='steelblue',
las=2,main = "NMAR",xlab = "Y", names = c("Yobs","Ymiss"))
```

Με χρήση του παραπάνω κώδικα καταλήγουμε στο Διάγραμμα 5.3.3.1 όπου απεικονίζει τα Θηκογράμματα των τιμών των Y_{miss} και Y_{obs}.



Διάγραμμα 5.3.3.1: Θηκογράμματα των τιμών του Y_{miss} και του Y_{obs}.

Από το Διάγραμμα 5.3.3.1 εύκολα παρατηρούμε ότι υπάρχει διαφορά στα θηκογράμματα μεταξύ του παρατηρήσιμου και του μη παρατηρήσιμου μέρους των τιμών του Y. Συγκεκριμένα βλέπουμε ότι το 4ο τεταρτημόριο των τιμών του Y_{miss}

βρίσκεται περίπου στο 1ο τεταρτημόριο των τιμών του Y_{obs}, οπότε είναι αρκετά πιθανό να υπάρχει στατιστικά σημαντική διαφορά μεταξύ των τιμών των Y_{miss} και Y_{obs}.

Εφόσον ο μηχανισμός έλλειψης τιμών εξαρτάται από την ίδια τη μεταβλητή $\boldsymbol{Y}$, θα χρειαστεί να ενσωματώσουμε ένα δείκτη miss ο οποίος μας λέει αν λείπει μία παρατήρηση ή όχι και προσδιορίζουμε ένα λογιστικό μοντέλο για την πιθανότητα έλλειψης με coef= $\log(0.6)$ (δεν ενσωματώνουμε το μοντέλο που μας έχει δημιουργήσει τις ελλιπείς τιμές, καθώς στην πραγματικότητα δεν το γνωρίζουμε). Αυτό σημαίνει ότι πως η αύξηση της μεταβλητής απόκρισης κατά μία μονάδα προκαλεί μείωση της πιθανότητας έλλειψης (ο λόγος ακέραιων τιμών (odds) μειώνεται), ενώ για την παράμετρο a βλέπουμε πως έχουμε προσδιορίσει μία λογιστική εκ των προτέρων κατανομή με $\mu = 0$ και $s = 1$ όπου αντιστοιχεί στη πιθανότητα έλλειψης όταν το Y_i έχει την τιμή μηδέν.

Μοντέλο αντικατάστασης ελλιπών τιμών στη μεταβλητή απόκρισης υπό το μηχανισμό έλλειψης NMAR (imputationresponsenon.txt)

```
#Imputation model for NMAR mechanism missingness
Model imputationresponse;
{
   for(j in 1:p){
      b[j] <- beta[j]/sd(x[,j])
      for(i in 1:n){
         z[i,j] <- (x[i,j]-mean(x[,j]))/sd(x[,j])
      }
      standardized_means[j] <- b[j]*mean(x[,j])
   }
   b0 <- beta0 - sum(standardized_means[])
   for(i in 1:n){
      y[i] ~ dnorm (mu[i],tau)
      mu[i] <- beta0 + beta[1]* z[i,1] + beta[2]* z[i,2]+ beta[3]*
          z[i,3] + beta[4]* z[i,4]+ beta[5]* z[i,5] + beta[6]*
          z[i,6]+ beta[7]* z[i,7] + beta[8]* z[i,8] + beta[9]*
          z[i,9] + beta[10]* z[i,10]+ beta[11]* z[i,11] + beta[12]*
          z[i,12]
      #selection model for missing data mechanism
      miss[i] ~ dbern(prob[i])
      logit(prob[i]) <- a + coef*y[i]
   }
   a ~ dlogis(0, 1)
   coef <- log(0.6)
   beta0 ~ dnorm(0, 0.0001)
   for(j in 1:p){
      beta[j] ~ dnorm(0,1.0E-5)
   }
   tau ~ dgamma(1.0E-3,1.0E-3)
   sigma <- sqrt(1/tau)
}
```

Εφαρμογή Μοντέλου αντικατάστασης ελλιπών τιμών στη μεταβλητή απόκρισης

```
#NMAR for Responce missing values
#Initial values and data inputs
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2)}
#dataset
data <- list(y=NMARDataset$Y, x
    =as.matrix(NMARDataset[,2:13]),p=12, n=80,
    miss=rep(1,times=80)-r2.nmar)
#We run the imputationresponse.txt
NMARImputationResponse <- bugs(inits=inits,data=data,model.file =
"imputationresponsenon.txt",parameters = c("beta0","beta", "gamma",
"sigma","y"),n.chains = 1, n.burnin = 2000, n.iter= 7000,n.thin = 1,
n.sims = 5000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=FALSE)
colMeans(NMARImputationResponse$sims.array[1:5000,1,])
#The imputed values we estimated
colMeans(NMARImputationResponse$sims.array[1:5000,1,15:
(sum(is.na(NMARDataset$Y))+14)])
```

Συνεχίζουμε αντικαθιστώντας τις ελλιπείς τιμές στη μεταβλητή απόκρισης κάνοντας χρήση του παρακάτω κώδικα.

Αντικατάσταση των ελλιπών τιμών στη μεταβλητή απόκρισης

```
NMARIMPUTEDDataset <- NMARDataset
#We impute the values of Y
NMARIMPUTEDDataset$Y <-
    replace(NMARIMPUTEDDataset$Y,is.na(NMARIMPUTEDDataset$Y),
as.vector(colMeans(NMARImputationResponse$sims.array[1:5000,
1,15:(sum(is.na(NMARDataset$Y))+14)])))
```

Επομένως στις παρακάτω υποενότητες θα εφαρμόσουμε τις μεθόδους SSVS, Kuo & Mallick, GVS , full enumeration και θα συγκρίνουμε τα αποτελέσματά τους.

5.3.3.1 Εφαρμογή της μεθόδου SSVS

Για να εφαρμόσουμε τη μέθοδο SSVS θα χρειαστεί να χρησιμοποιήσουμε το αρχείο SSVS.txt που έχουμε διατυπώσει στην Ενότητα 5.2.1 και με τη βοήθεια της εντολής bugs υλοποιήσουμε το αρχείο στο WinBUGS.

Με τη χρήση του παρακάτω κώδικα (στην R) υλοποιούμε τη μέθοδο SSVS στο WinBUGS και έτσι καταλήγουμε στους Πίνακες 5.3.3.1.1 και 5.3.3.1.2.

Εφαρμογή της μεθόδου SSVS

```
#Initial values and data values
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0), tau=0.2,
    gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=NMARIMPUTEDDataset$Y, x
    =as.matrix(NMARIMPUTEDDataset[,2:13]),p=12, n=80, models=4096)
#We implement the ssvs method
NMARmodelssvs <- bugs(inits=inits,data=data,model.file =
    "SSVS.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter =
    12000,n.thin = 5,n.sims = 10000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=FALSE)
NMARmodelparametersssvs <- NMARmodelssvs$summary[1:26,]
NMARmodelparametersssvs
NMARmodelsssvs <- NMARmodelssvs$summary[27:4122,]
NMARmodelsssvs
#We print the first 10 models with the highest posterior probability
head(NMARmodelsssvs[order(NMARmodelsssvs[, 2], decreasing=TRUE), ],n=10)
```

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.406	0.09822	0.002616	2.212	2.406	2.597
beta[1]	0.00688	0.1066	0.00242	−0.1946	0.008794	0.218
beta[2]	−0.2961	0.1158	0.002468	−0.5294	−0.2958	−0.07142
beta[3]	2.704	0.1339	0.003203	2.431	2.699	2.968
beta[4]	0.05762	0.1104	0.00232	−0.1604	0.05796	0.2849
beta[5]	−1.881	0.1129	0.002357	−2.101	−1.881	−1.656
beta[6]	−0.002236	0.1293	0.00321	−0.2548	−0.001904	0.2532
beta[7]	−0.1541	0.1459	0.003623	−0.4467	−0.1583	0.1392
beta[8]	−0.1041	0.1074	0.002132	−0.3098	−0.1044	0.1112
beta[9]	1.152	0.1424	0.002806	0.8558	1.157	1.417
beta[10]	0.02331	0.1299	0.002889	−0.2425	0.02689	0.271
beta[11]	0.4658	0.1154	0.002309	0.2396	0.4668	0.6965
beta[12]	0.1126	0.129	0.002852	−0.1431	0.1117	0.3567
sigma	0.8537	0.07214	0.001642	0.7309	0.8505	1.006
Δείκτες γ						
gamma[1]	0.012	0.1089	0.002415	0.0	0.0	0.0
gamma[2]	0.018	0.133	0.00299	0.0	0.0	0.0
gamma[3]	1.0	0.0	$2.236E-12$	1.0	1.0	1.0
gamma[4]	0.0115	0.1066	0.002805	0.0	0.0	0.0
gamma[5]	1.0	0.0	$2.236E-12$	1.0	1.0	1.0
gamma[6]	0.0145	0.1195	0.002772	0.0	0.0	0.0
gamma[7]	0.0095	0.097	0.002427	0.0	0.0	0.0
gamma[8]	0.0135	0.1154	0.002956	0.0	0.0	0.0
gamma[9]	0.818	0.3858	0.007351	0.0	1.0	1.0
gamma[10]	0.0105	0.1019	0.002302	0.0	0.0	0.0
gamma[11]	0.0325	0.1773	0.003866	0.0	0.0	1.0
gamma[12]	0.014	0.1175	0.002499	0.0	0.0	0.0

Πίνακας 5.3.3.1.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (NMAR).

Στον παραπάνω πίνακα παρατηρούμε ότι ο συντελεστής β_2 και οι συντελεστές του πραγματικού μοντέλου είναι σημαντικοί, καθώς στα διαστήματα αξιοπιστίας δεν υπάρχει η τιμή μηδέν. Ωστόσο στους δείκτες $\boldsymbol{\gamma}$ βλέπουμε ότι τα γ_3, γ_5 είναι 1, ενώ $\gamma_9 = 0.818$ με αρκετά μεγάλη τυπική απόκλιση και $\gamma_{11} = 0.0325$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	277	0.7150	0.4515
$\boldsymbol{X}_3 + \boldsymbol{X}_5$	21	0.1615	0.3680
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.0240	0.1530
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	279	0.0125	0.1111
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{12}$	2325	0.0100	0.0995
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9$	309	0.0095	0.0970
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	278	0.0085	0.0918
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9$	405	0.0080	0.0891
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9$	341	0.0075	0.0862
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10}$	789	0.0070	0.0833

Πίνακας 5.3.3.1.2: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (NMAR).

Στο Πίνακα 5.3.3.1.2 έχουμε τα πρώτα δέκα μοντέλα με τις μεγαλύτερες εκ των υστέρων πιθανότητες. Συγκεκριμένα βλέπουμε ότι το μοντέλο που επισκέφτηκε η αλυσίδα μας περισσότερο είναι αυτό με τους συντελεστές β_3, β_5 και β_9, ενώ το πραγματικό μοντέλο βρίσκεται στην 3η θέση με πιθανότητα 0.0240.

5.3.3.2 Εφαρμογή της μεθόδου Kuo & Mallick

Εφαρμόζουμε το δειγματολήπτη Kuo & Mallick κάνοντας χρήση του αρχείου KuoMallick.txt που ορίσαμε στην Ενότητα 5.2.2.

Εφαρμογή της μεθόδου Kuo & Mallick

```
#Initial values
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
#dataset
data <- list(y=NMARIMPUTEDDataset$Y, x
    =as.matrix(NMARIMPUTEDDataset[,2:13]),p=12, n=80, models=4096)
#We implement the Kuo & Mallick sampler
NMARmodelKuoMallick <- bugs(inits=inits,data=data,model.file =
    "KuoMallick.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 5000,n.thin = 1,n.sims = 3000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=FALSE)
NMARmodelparametersKuoMallick <- NMARmodelKuoMallick$summary[1:26,]
NMARmodelparametersKuoMallick
NMARmodelsKuoMallick <- NMARmodelKuoMallick$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
head(NMARmodelsKuoMallick[order(NMARmodelsKuoMallick[, 2],
    decreasing=TRUE), ],n=10)
```

Εφαρμόζουμε τη μέθοδο Kuo & Mallick με χρήση του παραπάνω κώδικα και έτσι καταλήγουμε στους Πίνακες 5.3.3.2.1 και 5.3.3.2.2.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.406	0.0991	0.001692	2.219	2.405	2.599
beta[1]	−0.2039	100.9	1.694	−198.8	0.4422	194.2
beta[2]	−0.966	91.8	1.532	−190.5	−0.3325	186.9
beta[3]	2.706	0.1138	0.005277	2.476	2.708	2.925
beta[4]	0.3694	99.12	1.966	−194.6	0.6471	191.3
beta[5]	−1.869	0.1313	0.00969	−2.114	−1.877	−1.598
beta[6]	−2.316	102.2	2.042	−213.1	−0.7495	192.2
beta[7]	0.6559	100.6	2.047	−198.2	−0.3326	200.7
beta[8]	−0.08208	98.54	1.926	−186.6	−1.571	203.3
beta[9]	1.176	0.1218	0.007529	0.9474	1.172	1.429
beta[10]	1.369	99.66	1.727	−196.1	1.966	191.5
beta[11]	−1.783	52.99	0.94	−143.4	0.495	130.1
beta[12]	0.155	100.5	1.759	−199.5	0.436	197.2
sigma	0.9128	0.09201	0.006794	0.7534	0.9059	1.113
Δείκτες γ						
gamma[1]	0.001667	0.04079	0.001176	0.0	0.0	0.0
gamma[2]	0.1533	0.3603	0.0407	0.0	0.0	1.0
gamma[3]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[4]	0.001	0.03161	$7.409E-4$	0.0	0.0	0.0
gamma[5]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[6]	0.003333	0.05764	0.001421	0.0	0.0	0.0
gamma[7]	0.006333	0.07933	0.003439	0.0	0.0	0.0
gamma[8]	0.001333	0.03649	$9.303E-4$	0.0	0.0	0.0
gamma[9]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[10]	0.002	0.04468	0.001395	0.0	0.0	0.0
gamma[11]	0.719	0.4495	0.05863	0.0	1.0	1.0
gamma[12]	0.003333	0.05764	0.001887	0.0	0.0	0.0

Πίνακας 5.3.3.2.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ του δειγματολήπτη Kuo & Mallick σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (NMAR).

Στον Πίνακα 5.3.3.2.1 παρατηρούμε ότι οι συντελεστές που αντιστοιχούν στους συντελεστές του πραγματικού μοντέλου είναι σημαντικοί και παράλληλα στους δείκτες $\boldsymbol{\gamma}$ έχουμε ότι $\gamma_3, \gamma_5, \gamma_9$ και $\gamma_{11} = 0.719$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.5540	0.4971
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	1303	0.2753	0.4467
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1303	0.1516	0.3587
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1365	0.0033	0.0576
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9$	341	0.0030	0.0546
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3349	0.0026	0.0515
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1333	0.0023	0.0482
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9$	309	0.0010	0.0316
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1302	0.0010	0.0316
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10} + \boldsymbol{X}_{11}$	1813	0.0010	0.0316

Πίνακας 5.3.3.2.2: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων του δειγματολήπτη Kuo & Mallick σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (NMAR).

Από τον Πίνακα 5.3.3.2.2 διαπιστώνουμε ότι το πραγματικό μοντέλο κατέχει μία αρκετά υψηλή εκ των υστέρων πιθανότητα. Ως εκ τούτου μπορούμε να πούμε πως ο δειγματολήπτης Kuo & Mallick κατάφερε να ανιχνεύσει το πραγματικό μοντέλο.

5.3.3.3 Εφαρμογή της μεθόδου GVS

Εφαρμόζουμε το πλήρες μοντέλο με στόχο να πάρουμε τις εκτιμήσεις των παραμέτρων της ψευδο-πρότερης εκ των προτέρων κατανομής (Πίνακας 5.3.3.3.1).

Κώδικας εκτέλεσης του Full μοντέλου

```
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2)}
data <- list(y=NMARIMPUTEDDataset$Y, x
    =as.matrix(NMARIMPUTEDDataset[,2:13]),p=12, n=80)
NMARmodelpilotrun <- bugs(inits=inits, data=data, model.file =
    "Pilotrun.txt",parameters = c("beta0","beta", "sigma"),n.chains
    = 1, n.burnin = 2000, n.iter = 7000,n.thin = 1,n.sims =
    5000,bugs.directory = "c:/Program Files/WinBUGS14/",debug=FALSE)
```

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.409	0.09722	0.001355	2.218	2.408	2.599
beta[1]	0.005863	0.1173	0.001528	−0.2205	0.006001	0.2342
beta[2]	−0.3635	0.1303	0.001587	−0.6223	−0.3626	−0.1098
beta[3]	2.67	0.1447	0.002128	2.388	2.669	2.956
beta[4]	0.07638	0.1191	0.001402	−0.1623	0.07853	0.3151
beta[5]	−1.905	0.1185	0.001599	−2.132	−1.905	−1.671
beta[6]	0.04231	0.145	0.002236	−0.2355	0.04093	0.3352
beta[7]	−0.2132	0.1751	0.002259	−0.5622	−0.2131	0.1242
beta[8]	−0.1256	0.1184	0.001328	−0.361	−0.1242	0.1077
beta[9]	1.169	0.1362	0.001795	0.9037	1.17	1.438
beta[10]	0.06123	0.1533	0.002181	−0.2461	0.05928	0.364
beta[11]	0.5018	0.1294	0.001633	0.2437	0.5028	0.7532
beta[12]	0.169	0.1519	0.00211	−0.1314	0.1676	0.4687

Πίνακας 5.3.3.3.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των συντελεστών του πλήρες μοντέλου.

Εφαρμογή της μεθόδου GVS

```
#We list our initial values and the dataset
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=NMARIMPUTEDDataset$Y, x
    =as.matrix(NMARIMPUTEDDataset[,2:13]),p=12,n=80,models=4096,mean=
as.vector(colMeans(NMARmodelpilotrun$sims.array[1:5000,1,]))[2:13],
se=as.vector(NMARmodelpilotrun$sd$beta))
NMARmodelgvs <- bugs(inits=inits,data=data,model.file =
"GVS.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 5000,n.thin = 1,n.sims = 3000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=TRUE)
#Here we have the summary of the model parameters
NMARmodelparametersgvs <- NMARmodelgvs$summary[1:26,]
NMARmodelsgvs <- NMARmodelgvs$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
head(NMARmodelsgvs[order(NMARmodelsgvs[, 2], decreasing=TRUE),
    ],n=10)
```

Εφαρμόζοντας τον παραπάνω κώδικα στην R καταλήγουμε στους Πίνακες 5.3.2.3.2

και 5.3.1.3.3.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.656	0.0939	0.001654	2.47	2.657	2.841
beta[1]	0.005325	0.1184	0.001993	−0.2273	0.006375	0.2329
beta[2]	−0.3601	0.1228	0.002169	−0.5936	−0.3612	−0.1227
beta[3]	2.714	0.1072	0.002074	2.496	2.715	2.92
beta[4]	0.07564	0.1214	0.002291	−0.1642	0.07512	0.3197
beta[5]	−1.893	0.1228	0.003687	−2.115	−1.896	−1.642
beta[6]	0.04224	0.1461	0.002988	−0.2357	0.03871	0.3343
beta[7]	−0.2178	0.179	0.003659	−0.5864	−0.2157	0.1224
beta[8]	−0.1252	0.1195	0.002442	−0.3612	−0.1271	0.1122
beta[9]	1.162	0.1145	0.003538	0.9528	1.157	1.398
beta[10]	0.05995	0.1527	0.002642	−0.2386	0.06061	0.3579
beta[11]	0.4977	0.1204	0.002811	0.2616	0.4994	0.7325
beta[12]	0.1647	0.1517	0.002962	−0.1427	0.1635	0.4584
sigma	0.9	0.08599	0.002424	0.7514	0.8934	1.084
Δείκτες γ						
gamma[1]	$3.333E-4$	0.01825	$3.35E-4$	0.0	0.0	0.0
gamma[2]	0.1397	0.3466	0.009961	0.0	0.0	1.0
gamma[3]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[4]	$6.667E-4$	0.02581	$4.694E-4$	0.0	0.0	0.0
gamma[5]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[6]	0.004333	0.06569	0.001161	0.0	0.0	0.0
gamma[7]	0.006333	0.07933	0.001432	0.0	0.0	0.0
gamma[8]	0.001	0.03161	$5.591E-4$	0.0	0.0	0.0
gamma[9]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[10]	0.003333	0.05764	0.001268	0.0	0.0	0.0
gamma[11]	0.857	0.3501	0.01341	0.0	1.0	1.0
gamma[12]	$3.333E-4$	0.01825	$3.35E-4$	0.0	0.0	0.0

Πίνακας 5.3.2.3.2: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (NMAR).

Στον Πίνακα 5.3.2.3.2 βλέπουμε ότι οι συντελεστές $\beta_3, \beta_5, \beta_9$ και β_{11} είναι σημαντικοί και $\gamma_3, \gamma_5, \gamma_9 = 1$, $\gamma_{11} = 0.857$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.7056	0.4558
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	277	0.1390	0.3460
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1303	0.1383	0.3453
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1333	0.0040	0.0631
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9$	341	0.0033	0.0576
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1813	0.0033	0.0576
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1365	0.0030	0.0546
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1429	0.0010	0.0316
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	279	0.0006	0.025
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1302	0.0003	0.0182

Πίνακας 5.3.2.3.3: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (NMAR).

Από τον Πίνακα 5.3.2.3.3 μας είναι φανερό ότι η μέθοδος GVS κατάφερε να ανιχνεύσει το πραγματικό μοντέλο, καθώς η εκ των υστέρων πιθανότητα είναι ίση με 0.7056.

5.3.3.4 Πλήρης απαρίθμηση με τη χρήση του κριτηρίου BIC

Εφαρμόζουμε πλήρη απαρίθμηση χρησιμοποιώντας τη συνάρτηση BICminimum που έχουμε κατασκευάσει στην R.

Εφαρμογή της συνάρτησης BICminimum στην R

```
Lowest10BICmodelsNMAR <-BICminimum(p=12,NMARIMPUTEDDataset)
#Coefficients of the models
Lowest10BICmodelsNMAR[[1]][1:10]
#and their corresponding BIC values.
Lowest10BICmodelsNMAR[[2]][1:10]
#Model code
Lowest10BICmodelsNMAR[[3]]
```

Εφαρμόζοντας τη συνάρτηση BICminimum καταλήγουμε στον Πίνακα 5.3.3.4.1.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Κριτήριο BIC
$\boldsymbol{X}_2+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1303	221.3815
$\boldsymbol{X}_2+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_7+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1367	224.3996
$\boldsymbol{X}_2+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_8+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1431	224.9061
$\boldsymbol{X}_2+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{10}+\boldsymbol{X}_{11}$	1815	225.1329
$\boldsymbol{X}_2+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_6+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1335	225.2073
$\boldsymbol{X}_2+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{11}+\boldsymbol{X}_{12}$	3351	225.2399
$\boldsymbol{X}_2+\boldsymbol{X}_3+\boldsymbol{X}_4+\boldsymbol{X}_5+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1311	225.7559
$\boldsymbol{X}_1+\boldsymbol{X}_2+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_{99}+\boldsymbol{X}_{11}$	1304	225.7567
$\boldsymbol{X}_2+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_7+\boldsymbol{X}_9+\boldsymbol{X}_{11}+\boldsymbol{X}_{12}$	3415	227.8781
$\boldsymbol{X}_2+\boldsymbol{X}_3+\boldsymbol{X}_5+\boldsymbol{X}_7+\boldsymbol{X}_8+\boldsymbol{X}_9+\boldsymbol{X}_{11}$	1495	227.9917

Πίνακας 5.3.3.4.1: Τα 10 πρώτα μοντέλα με τις μικρότερες τιμές του κριτηρίου BIC σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης (NMAR).

Από το Πίνακα 5.3.3.4.1 βλέπουμε ότι το μοντέλο με τη μικρότερη τιμή του κριτηρίου BIC, αντιστοιχεί στο μοντέλο που περιέχει τους συντελεστές του πραγματικού μοντέλου μαζί με το συντελεστή β_2.

Μοντέλο με τη μικρότερη τιμή του κριτηρίου BIC

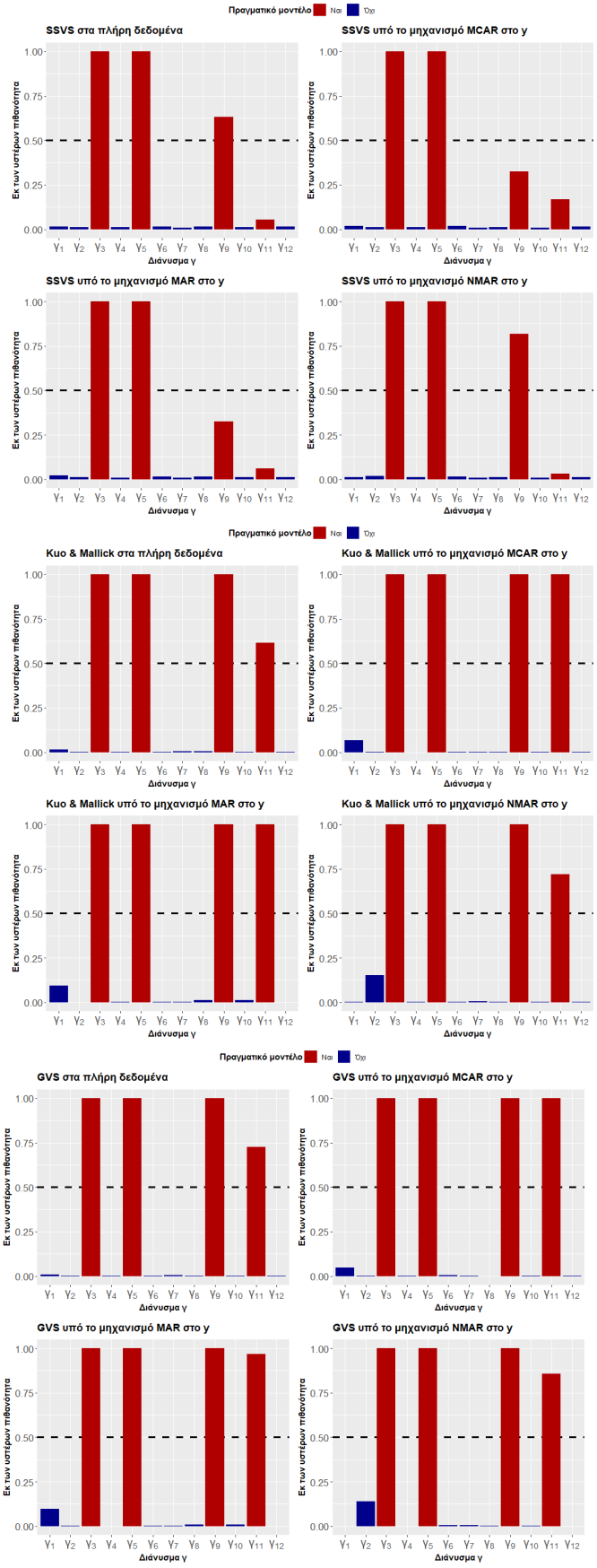
```
#We take the summary of the model with lowest BIC value
summary(lm(Y~ X1+X3+X5+X8+X9+X11,data=MARIMPUTEDDataset))
#Partial console output
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.25310   0.09140 24.652 < 2e-16 ***
X2         -0.24567   0.09530 -2.578  0.0119 *
X3          2.41348   0.08774 27.508 < 2e-16 ***
X5         -1.75327   0.09614 -18.237 < 2e-16 ***
X9          1.09039   0.09693 11.250 < 2e-16 ***
X11         0.54301   0.09500  5.716 2.16e-07 ***
---
```

5.4 Σύγκριση των Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης

Στην προηγούμενη ενότητα είδαμε ότι η εμφάνιση ελλιπών τιμών στη μεταβλητή απόκρισης, δεν επηρεάζει σε μεγάλο βαθμό την αποτελεσματικότητα των Μπεϋζιανών μεθόδων επιλογής μεταβλητών, ανεξάρτητα από το μηχανισμό έλλειψης που κατέχουμε. Σε αυτή την ενότητα πρωταρχικός μας στόχος είναι να συγκρίνουμε τα αποτελέσματα των μεθόδων SSVS, Kuo & Mallick και GVS και να προβούμε σε συμπεράσματα ως προς την προσαρμοστικότητα και αποτελεσματικότητα των μεθόδων όταν τα δεδομένα κατέχουν ελλιπείς τιμές στη μεταβλητή απόκρισης.

Το Διάγραμμα 5.4.1 περιέχει τα ραβδογράμματα των δεικτών $\boldsymbol{\gamma}$ από τις μεθόδους SSVS, Kuo & Mallick και GVS στα πλήρη δεδομένα, αλλά και σε δεδομένα με ελλιπείς τιμές στη μεταβλητή απόκρισης Y (από μηχανισμούς έλλειψης MCAR, MAR και NMAR).

Από τα ραβδογράμματα του Διαγράμματος 5.4.1 βλέπουμε ότι η μέθοδος SSVS δίνει αρκετά χαμηλές τιμές στο δείκτη γ_{11}, στα πλήρη δεδομένα άλλα και σε δεδομένα που περιέχουν ελλιπείς τιμές στο Y. Συνεπώς καταλαβαίνουμε ότι η μέθοδος SSVS δεν είναι κατάλληλη για να ανιχνεύσει οριακά σημαντικές μεταβλητές στο μοντέλο. Αντιθέτως για το δειγματολήπτη Kuo & Mallick και τη μέθοδο GVS παρατηρούμε ότι οι δείκτες γ_3, γ_5, γ_9 και γ_{11} κατέχουν τιμές αρκετά υψηλές και πολύ κοντά στο ένα (εκτός του δείκτη γ_{11} στην περίπτωση που έχουμε πλήρη δεδομένα).



Διάγραμμα 5.4.1: Ραβδογράμματα των δεικτών $\boldsymbol{\gamma}$ στις μεθόδους SSVS, Kuo & Mallick και GVS σε δεδομένα με μηχανισμούς έλλειψης στη μεταβλητή απόκρισης Y.

5.5 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές

Αυτή η ενότητα εμβαθύνει σε μια ενδελεχή σύγκριση των τριών Μπεϋζιανών μεθόδων επιλογής μεταβλητών: Stochastic search variable selection (SSVS), το δειγματολήπτη Kuo & Mallick και τη μέθοδο Gibbs variable selection (GVS). Ο πρωταρχικός στόχος αυτής της σύγκρισης είναι η αξιολόγηση αυτών των μεθόδων υπό τη παρουσία ελλιπών τιμών στις επεξηγηματικές μεταβλητές του προσομοιωμένου δείγματος. Ωστόσο για τη δημιουργία των ελλιπών τιμών στις ανεξάρτητες μεταβλητές θα χρησιμοποιήσουμε πάλι τους μηχανισμούς έλλειψης τιμών που έχουμε αναπτύξει στο Κεφάλαιο 4 (MCAR, MAR και NMAR).

Αυτή η συγκριτική εξερεύνηση θα προσφέρει πολύτιμες γνώσεις σχετικά με την προσαρμοστικότητα και την αποτελεσματικότητα αυτών των μεθόδων, εξοπλίζοντας έτσι τους ερευνητές με μια ολοκληρωμένη κατανόηση της προσαρμοστικότητας και της αποτελεσματικότητας αυτών των μεθόδων.

5.5.1 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές υπό το μηχανισμό MCAR

Για να κατασκευάσουμε μηχανισμούς MCAR για κάποιες από τις επεξηγηματικές μεταβλητές μπορούμε να χρησιμοποιήσουμε τον ίδιο κώδικα που αναπτύξαμε στην Ενότητα 5.3.1. Συγκεκριμένα, θα δημιουργήσουμε ελλιπείς τιμές στις μεταβλητές X_3, X_7, X_{10} και X_{11} (όπου οι επεξηγηματικές μεταβλητές X_3 και X_{11} βρίσκονται στο πραγματικό μοντέλο) με χρήση του μηχανισμού έλλειψης MCAR. Για να δημιουργήσουμε το μηχανισμό MCAR για κάθε μεταβλητή, θα κάνουμε χρήση μίας κατανομής Bernoulli με πιθανότητα επιτυχίας (πιθανότητα έλλειψης $1-p$) p για κάθε παρατήρηση.

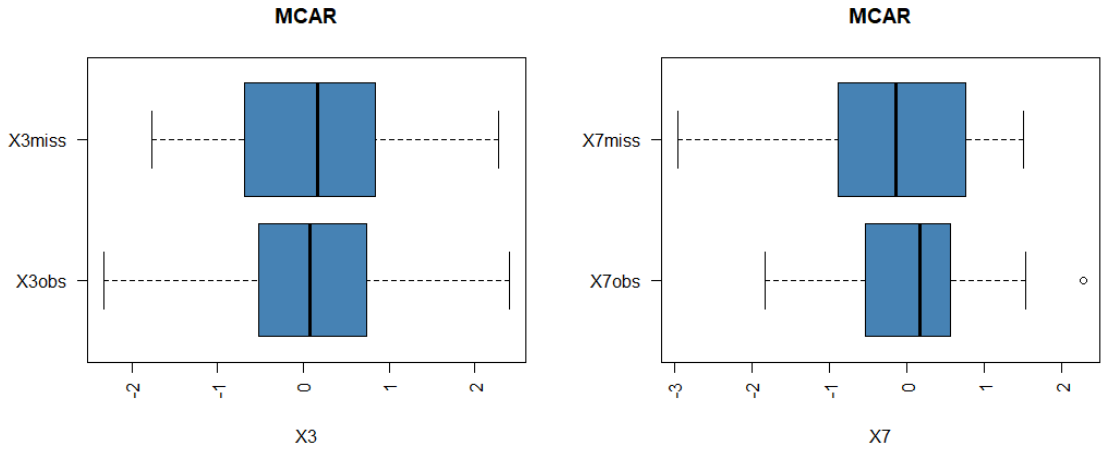
Κατασκευή μηχανισμού MCAR στις επεξηγηματικές μεταβλητές X_3 και X_7

```
#We create a Bernoulli trial for every observation
#with probability p=0.65 of being observed.
miss.X3 <- rbinom(samplesize, 1, prob = 0.65)
# # of observations that remain observed after the MCAR mechanism
sum(miss.X3==1)
MCARDatasetX <- FullDataset
# if miss.x3=1 then the observation remains observed
#and if miss.x3=0 then it becomes a missing value.
MCARDatasetX$X3 <- ifelse(miss.X3==1,MCARDatasetX$X3,NA)
#We create a Bernoulli trial for every observation
#with probability p=0.6 of being observed.
miss.X7 <- rbinom(samplesize, 1, prob = 0.6)
# # of observations that remain observed after the MCAR mechanism
sum(miss.X7==1)
# if miss.x7=1 then the observation remains observed
#and if miss.x7=0 then it becomes a missing value.
MCARDatasetX$X7 <- ifelse(miss.X7==1,MCARDatasetX$X7,NA)
```

Κατασκευή θηκογραμμάτων

```
#Number of observations that remain observed after the
#MCAR mechanism
sum(miss.X3==1)
#console output is 49.
#Number of missing values in X3.
sum(is.na(MCARDatasetX$X3))
#console output is 31.
#These are the observations that we took as missing
X3missMCAR <- FullDataset$X3[is.na(MCARDatasetX$X3)]
#Boxplot within the missing and the remaining observations.
boxplot(MCARDatasetX$X3, X3missMCAR, horizontal=TRUE,
    col='steelblue', las=2,main = "MCAR",
xlab = "X3", names = c("X3obs","X3miss"))
#Number of observations that remain observed after the
#MCAR mechanism
sum(miss.X7==1)
#console output is 46.
#Number of missing values in X7.
sum(is.na(MCARDatasetX$X7))
#console output is 34.
#These are the observations that we took as missing
X7missMCAR <- FullDataset$X7[is.na(MCARDatasetX$X7)]
#Boxplot within the missing and the remaining observations.
boxplot(MCARDatasetX$X7, X7missMCAR, horizontal=TRUE,
    col='steelblue', las=2,main = "MCAR",
xlab = "X7", names = c("X7obs","X7miss"))
```

Κάνοντας χρήση του παραπάνω κώδικα καταλήγουμε στο Διάγραμμα 5.5.1.1, όπου βλέπουμε ότι δεν υπάρχουν σημαντικές διαφορές μεταξύ των θηκογραμμάτων των τιμών των $X_{3\text{miss}}$, $X_{3\text{obs}}$ και $X_{7\text{miss}}$, $X_{7\text{obs}}$.



Διάγραμμα 5.5.1.1: Θηκογράμματα των τιμών των $X_{3\text{miss}}$, $X_{3\text{obs}}$ και $X_{7\text{miss}}$, $X_{7\text{obs}}$.

Κατασκευή μηχανισμού MCAR στις επεξηγηματικές μεταβλητές X_{10} και X_{11}

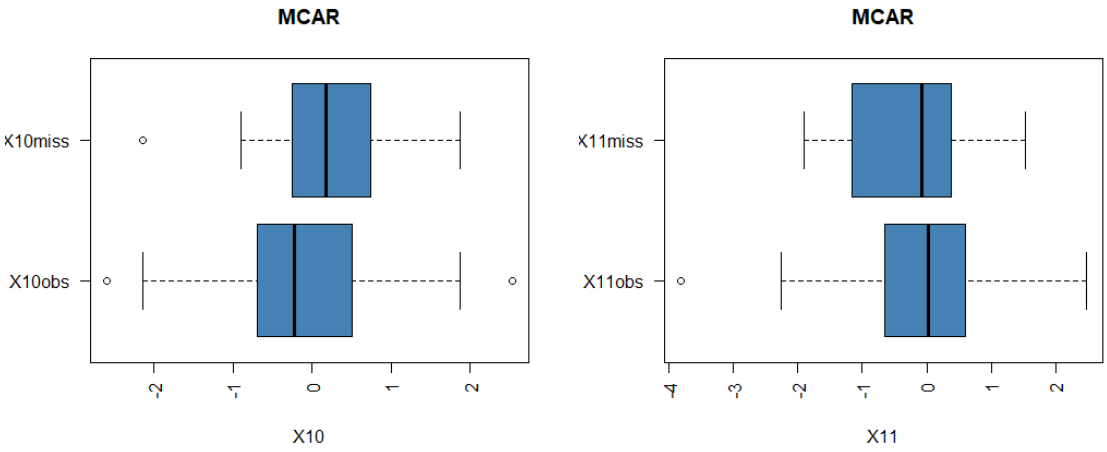
```
#We create a Bernoulli trial for every observation
#with probability p=0.7 of being observed.
miss.X10 <- rbinom(samplesize, 1, prob = 0.7)
# if miss.x10=1 then the observation remains observed
#and if miss.x10=0 then it becomes a missing value.
MCARDatasetX$X10 <- ifelse(miss.X10==1,MCARDatasetX$X10,NA)
#We create a Bernoulli trial for every observation
#with probability p=0.6 of being observed.
miss.X11 <- rbinom(samplesize, 1, prob = 0.6)
# if miss.x11=1 then the observation remains observed
#and if miss.x11=0 then it becomes a missing value.
MCARDatasetX$X11 <- ifelse(miss.X11==1,MCARDatasetX$X11,NA)
```

Κατασκευή θηκογραμμάτων

```
#Number of observations that remain observed after the
#MCAR mechanism
sum(miss.X10==1)
#console output is 50.
#Number of missing values in X10.
sum(is.na(MCARDatasetX$X10))
#console output is 30.
#These are the observations that we took as missing
X10missMCAR <- FullDataset$X10[is.na(MCARDatasetX$X10)]
#Boxplot within the missing and the remaining observations.
boxplot(MCARDatasetX$X10, X10missMCAR, horizontal=TRUE,
    col='steelblue', las=2,main = "MCAR",
xlab = "X10", names = c("X10obs","X10miss"))
#Number of observations that remain observed after the
#MCAR mechanism
sum(miss.X11==1)
#console output is 54.
#Number of missing values in X11.
sum(is.na(MCARDatasetX$X11))
#console output is 26.
#These are the observations that we took as missing
X11missMCAR <- FullDataset$X11[is.na(MCARDatasetX$X11)]
#Boxplot within the missing and the remaining observations.
boxplot(MCARDatasetX$X11, X11missMCAR, horizontal=TRUE,
    col='steelblue', las=2,main = "MCAR",
xlab = "X11", names = c("X11obs","X11miss"))
```

Κάνοντας χρήση του παραπάνω κώδικα, καταλήγουμε στο Διάγραμμα 5.5.1.2, όπου και εδώ διαπιστώνουμε ότι δεν υπάρχουν σημαντικές διαφορές μεταξύ των

θηκογραμμάτων των τιμών των $X_{10\text{miss}}$, $X_{10\text{obs}}$ και $X_{11\text{miss}}$, $X_{11\text{obs}}$.



Διάγραμμα 5.5.1.2: Θηκογράμματα των τιμών των $X_{10\text{miss}}$, $X_{10\text{obs}}$ και $X_{11\text{miss}}$, $X_{11\text{obs}}$.

Για να αντικαταστήσουμε αυτές τις ελλιπείς τιμές που δημιουργήσαμε, θα χρησιμοποιήσουμε την ίδιο σκεπτικό που αναπτύξαμε στην Ενότητα 4.5. Γράφουμε λοιπόν τον κώδικα του μοντέλου αντικατάστασης στο WinBUGS σε ένα αρχείο .txt (imputationcovariates.txt). Αυτό το μοντέλο θα χρησιμοποιηθεί για να αντικαταστήσουμε τις ελλιπείς τιμές στις επεξηγηματικές μεταβλητές ανεξάρτητα απ'το μηχανισμό έλλειψης που προσφέρουμε στο προσομοιωμένο δείγμα μας.

Συγκεκριμένα, η εκ των υστέρων κατανομή στον αλγόριθμο MCMC θα είναι η από κοινού εκ των υστέρων κατανομή 4.5.6 και η πλήρους δέσμευσης εκ των υστέρων κατανομή θα είναι η εξής:

$$p(\boldsymbol{X}_{i,miss}|\boldsymbol{\theta},\boldsymbol{\theta}_{\boldsymbol{X}},\boldsymbol{Y},\boldsymbol{X}_{i,obs}) \propto p(Y_i|\boldsymbol{X}_{i,obs},\boldsymbol{\theta})p(\boldsymbol{X}_{i,miss}|\boldsymbol{\theta}_{\boldsymbol{X}}). \tag{5.4.1.1}$$

Όμως στο πρόγραμμα WinBUGS αρκεί να ορίσουμε την κατανομή του $\boldsymbol{X}_{i,miss}$. Άρα για τις επεξηγηματικές μεταβλητές X_3, X_7, X_{10} και X_{11} έχουμε τα εξής μοντέλα:

$$\boldsymbol{X}_{i,3,miss} \sim N(\mu_3,\tau_3), \quad \boldsymbol{X}_{i,7,miss} \sim N(\mu_7,\tau_7),$$

$$\boldsymbol{X}_{i,10,miss} \sim N(\mu_{10},\tau_{10}), \quad \boldsymbol{X}_{i,11,miss} \sim N(\mu_{11},\tau_{11}),$$

όπου οι εκ των προτέρων κατανομές των παραμέτρων είναι οι εξής:

$$\mu_3 \sim N(0,\tau=10^{-4}), \quad \tau_3 \sim \text{Gamma}(10^{-4},10^{-4}),$$

$$\mu_7 \sim N(0,\tau=10^{-4}), \quad \tau_7 \sim \text{Gamma}(10^{-4},10^{-4}),$$
$$\mu_{10} \sim N(0,\tau=10^{-4}), \quad \tau_{10} \sim \text{Gamma}(10^{-4},10^{-4}),$$
$$\mu_{11} \sim N(0,\tau=10^{-4}), \quad \tau_{11} \sim \text{Gamma}(10^{-4},10^{-4}).$$

Μοντέλο αντικατάστασης ελλιπών τιμών στις επεξηγηματικές μεταβλητές (imputationcovariates.txt)

```
Model imputationcovariates;
{
for(j in 1:p){
b[j] <- beta[j]/sd(x[,j])
for(i in 1:n){
z[i,j] <- (x[i,j]-mean(x[,j]))
}
standardized_means[j] <- b[j]*mean(x[,j])
}
b0 <- beta0 - sum(standardized_means[])

for(i in 1:n){
y[i] ~ dnorm (mu[i],tau)
mu[i] <- beta0 + beta[1]* z[i,1] + beta[2]* z[i,2]+ beta[3]* z[i,3]
   + beta[4]* z[i,4]+ beta[5]* z[i,5] + beta[6]* z[i,6]+ beta[7]*
   z[i,7] + beta[8]* z[i,8] + beta[9]* z[i,9] + beta[10]* z[i,10]+
   beta[11]* z[i,11] + beta[12]* z[i,12]

x[i,3] ~ dnorm(mu.x3, tau.x3)
x[i,7] ~ dnorm(mu.x7, tau.x7)
x[i,10] ~ dnorm(mu.x10, tau.x10)
x[i,11] ~ dnorm(mu.x11, tau.x11)
}

beta0 ~ dnorm(0, 0.0001)
for(j in 1:p){
beta[j] ~ dnorm(0,1.0E-5)
}
tau ~ dgamma(1.0E-4,1.0E-4)
mu.x3 ~ dnorm(0,1.0E-4)
tau.x3 ~ dgamma(1.0E-4,1.0E-4)
mu.x7 ~ dnorm(0,1.0E-4)
tau.x7 ~ dgamma(1.0E-4,1.0E-4)
mu.x10 ~ dnorm(0,1.0E-4)
tau.x10 ~ dgamma(1.0E-4,1.0E-4)
mu.x11 ~ dnorm(0,1.0E-4)
tau.x11 ~ dgamma(1.0E-4,1.0E-4)
sigma <- sqrt(1/tau)
}
```

Με τη βοήθεια της εντολής bugs στην R, τρέχουμε το παραπάνω αρχείο .txt το οποίο περιέχει το μοντέλο αντικατάστασης των ελλιπών τιμών στις επεξηγηματικές μεταβλητές σε κώδικα του WinBUGS. Βασικός μας στόχος είναι να αποκτήσουμε τιμές για τις επεξηγηματικές μεταβλητές X_3, X_7, X_{10} και X_{11}, τις οποίες μπορούμε να τις αντικαταστήσουμε στη θέση των ελλιπών τιμών που δημιουργήθηκαν μέσω του MCAR μηχανισμού.

Εφαρμογή Μοντέλου αντικατάστασης ελλιπών τιμών στις επεξηγηματικές μεταβλητές

```
#MCAR for missing values in the covariates.
#Initial values
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, mu.x3=0, tau.x3=1,mu.x7=0, tau.x7=1, mu.x10=0,
    tau.x10=1, mu.x11=0, tau.x11=1)}
#Data inputs
data <- list(y=MCARDatasetX$Y, x
    =as.matrix(MCARDatasetX[,2:13]),p=12, n=80)
#We run the imputationcovariates.txt
MCARImputationCovariates <- bugs(inits=inits,data=data,model.file =
"imputationcovariates.txt",parameters = c("beta0","beta",
    "sigma","x[,3]","x[,7]","x[,10]","x[,11]"),n.chains = 1,
    n.burnin = 2000, n.iter = 7000,n.thin = 1,n.sims =
    5000,bugs.directory = "c:/Program
Files/WinBUGS14/",debug=FALSE)
MCARImputationCovariates
#The imputed values we estimated
colMeans(MCARImputationCovariates$sims.array[1:5000,1,15:(dim(
MCARImputationCovariates$sims.array)[3]-1)])
```

Έχοντας αποκτήσει τις τιμές των επεξηγηματικών μεταβλητών, μπορούμε να τις αντικαταστήσουμε στις ελλιπείς τιμές κάνοντας χρήση του παρακάτω κώδικα στην R.

Αντικατάσταση των ελλιπών τιμών στις επεξηγηματικές μεταβλητές

```
MCARIMPUTEDDatasetX <- MCARDatasetX
imputedvalues <-
    as.vector(colMeans(MCARImputationCovariates$sims.array[1:5000,
1,15:(dim(MCARImputationCovariates$sims.array)[3]-1)]))
counter <- 1
for (j in 1:nrow(MCARIMPUTEDDatasetX)) {
   for (i in 1:ncol(MCARIMPUTEDDatasetX)) {
      #If the xji point is missing then we impute it
      #with the corresponding value taken from the imputation model
      if(is.na(MCARIMPUTEDDatasetX[j,i])==TRUE){
         MCARIMPUTEDDatasetX[j,i] <- imputedvalues[counter]
         counter <- counter + 1
      }
   }
}
```

Επομένως στις παρακάτω υποενότητες θα εφαρμόσουμε και θα συγκρίνουμε τα αποτελέσματα των μεθόδων SSVS, Kuo & Mallick, GVS και full enumeration στο δείγμα MCARIMPUTEDDatasetX.

5.5.1.1 Εφαρμογή της μεθόδου SSVS

Για να εφαρμόσουμε τη μέθοδο SSVS (όπως και στην Ενότητα 5.3), θα χρησιμοποιήσουμε το αρχείο SSVS.txt που διατυπώσαμε στην Ενότητα 5.2.1. Με τη χρήση του παρακάτω κώδικα, υλοποιούμε τη μέθοδο SSVS και έτσι καταλήγουμε στους Πίνακες 5.5.1.1.1 και 5.5.1.1.2.

Εφαρμογή της μεθόδου SSVS

```
#Initial values and dataset and we implement the ssvs method
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=MCARIMPUTEDDatasetX$Y, x
    =as.matrix(MCARIMPUTEDDatasetX[,2:13]),p=12, n=80, models=4096)
MCARmodelXssvs <- bugs(inits=inits,data=data,model.file =
    "SSVS.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 12000,n.thin = 5,n.sims = 10000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=FALSE)
MCARmodelXparametersssvs <- MCARmodelXssvs$summary[1:26,]
MCARmodelsXssvs <- MCARmodelXssvs$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
head(MCARmodelsXssvs[order(MCARmodelsXssvs[, 2], decreasing=TRUE),
    ],n=10)
```

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.41	0.09388	0.002497	2.221	2.409	2.593
beta[1]	−0.4292	0.1047	0.002402	−0.6292	−0.4297	−0.221
beta[2]	0.04051	0.1082	0.002163	−0.1796	0.04157	0.2501
beta[3]	2.81	0.1227	0.002609	2.56	2.807	3.058
beta[4]	−0.1188	0.1046	0.002324	−0.3246	−0.1199	0.09976
beta[5]	−1.486	0.1075	0.002343	−1.696	−1.486	−1.274
beta[6]	0.001307	0.1103	0.00259	−0.2088	0.001886	0.2233
beta[7]	0.07721	0.1036	0.002379	−0.1253	0.07827	0.2705
beta[8]	−0.3046	0.1001	0.00188	−0.4977	−0.3046	−0.104
beta[9]	1.018	0.1216	0.002207	0.7812	1.019	1.251
beta[10]	−0.2721	0.1023	0.001989	−0.4846	−0.2733	−0.07195
beta[11]	0.6678	0.1066	0.00215	0.4621	0.6688	0.8812
beta[12]	0.03029	0.1166	0.002512	−0.2003	0.02893	0.2537
sigma	0.817	0.06988	0.001576	0.7002	0.8124	0.9662
Δείκτες γ						
gamma[1]	0.0295	0.1692	0.003395	0.0	0.0	1.0
gamma[2]	0.0115	0.1066	0.0022	0.0	0.0	0.0
gamma[3]	1.0	0.0	$2.236E-12$	1.0	1.0	1.0
gamma[4]	0.012	0.1089	0.002972	0.0	0.0	0.0
gamma[5]	0.9945	0.07396	0.001621	1.0	1.0	1.0
gamma[6]	0.0145	0.1195	0.002772	0.0	0.0	0.0
gamma[7]	0.009	0.09444	0.002073	0.0	0.0	0.0
gamma[8]	0.0185	0.1348	0.003269	0.0	0.0	0.0
gamma[9]	0.6185	0.4858	0.008926	0.0	1.0	1.0
gamma[10]	0.015	0.1216	0.002738	0.0	0.0	0.0
gamma[11]	0.106	0.3078	0.006604	0.0	0.0	1.0
gamma[12]	0.0135	0.1154	0.002602	0.0	0.0	0.0

Πίνακας 5.5.1.1.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MCAR).

Από τον Πίνακα 5.5.1.1.1 βλέπουμε ότι οι συντελεστές β_1, β_8, β_{10} και οι συντελεστές του πραγματικού μοντέλου είναι σημαντικοί. Στους δείκτες $\boldsymbol{\gamma}$ έχουμε ότι $\gamma_3 = 1$, $\gamma_5 = 0.9945$ και $\gamma_9 = 0.6125$, ενώ βλέπουμε ότι ο δείκτης γ_{11} ισούται με 0.106 όπου είναι μία αρκετά μικρή τιμή.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	277	0.4895	0.5000
$\boldsymbol{X}_3 + \boldsymbol{X}_5$	21	0.3025	0.4594
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.0565	0.2309
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_{11}$	1045	0.0340	0.1812
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	278	0.0135	0.1154
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5$	22	0.0095	0.0970
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9$	309	0.0070	0.0833
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10}$	789	0.0070	0.0833
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9$	85	0.0065	0.0803
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{12}$	29	0.0065	0.0803

Πίνακας 5.5.1.1.2: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MCAR).

Από τον Πίνακα 5.5.1.1.2 διαπιστώνουμε ότι το μοντέλο με τη μεγαλύτερη εκ των υστέρων πιθανότητα αντιστοιχεί στο μοντέλο που περιέχει τους συντελεστές β_3, β_5 και β_9, ενώ στο πραγματικό μοντέλο βλέπουμε ότι έχουμε εκ των υστέρων πιθανότητα ίση με 0.0565.

5.5.1.2 Εφαρμογή της μεθόδου Kuo & Mallick

Εφαρμόζουμε το δειγματολήπτη Kuo & Mallick κάνοντας χρήση του αρχείου KuoMallick.txt που ορίσαμε στην Ενότητα 5.2.2.

Εφαρμογή της μεθόδου Kuo & Mallick

```
#Initial values and data values
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=MCARIMPUTEDDatasetX$Y, x
    =as.matrix(MCARIMPUTEDDatasetX[,2:13]),p=12, n=80, models=4096)
#We implement the Kuo & Mallick sampler
MCARmodelXKuoMallick <- bugs(inits=inits,data=data,model.file =
    "KuoMallick.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 5000,n.thin = 1,n.sims = 3000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=FALSE)
MCARmodelXparametersKuoMallick <-
    MCARmodelXKuoMallick$summary[1:26,]
MCARmodelsXKuoMallick <- MCARmodelXKuoMallick$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
head(MCARmodelsXKuoMallick[order(MCARmodelsXKuoMallick[, 2],
    decreasing=TRUE), ],n=10)
```

Εφαρμόζουμε τη μέθοδο Kuo & Mallick με χρήση του παραπάνω κώδικα και έτσι καταλήγουμε στους Πίνακες 5.5.1.2.1 και 5.5.1.2.2.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.409	0.105	0.001763	2.207	2.408	2.615
beta[1]	0.3062	80.05	1.398	−182.0	−0.4353	176.4
beta[2]	−2.277	99.68	1.554	−203.3	−1.294	191.8
beta[3]	2.738	0.1256	0.00717	2.481	2.739	2.982
beta[4]	0.5143	98.96	1.975	−194.6	0.5256	191.3
beta[5]	−1.657	0.1426	0.01131	−1.921	−1.663	−1.356
beta[6]	−2.338	102.4	2.056	−213.1	−0.8398	192.7
beta[7]	0.5506	100.8	2.043	−198.2	−0.5584	200.7
beta[8]	0.9239	89.96	1.627	−177.1	−0.3716	197.2
beta[9]	1.017	0.1298	0.007899	0.7784	1.014	1.296
beta[10]	1.482	96.33	1.709	−194.7	−0.2863	189.2
beta[11]	−0.5774	34.76	0.7065	−95.83	0.6031	74.39
beta[12]	0.1551	100.5	1.759	−199.5	0.436	197.2
sigma	0.9657	0.1121	0.01041	0.7657	0.9604	1.202
Δείκτες γ						
gamma[1]	0.374	0.4839	0.06273	0.0	0.0	1.0
gamma[2]	$6.667E-4$	0.02581	$4.608E-4$	0.0	0.0	0.0
gamma[3]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[4]	0.004333	0.06569	0.003955	0.0	0.0	0.0
gamma[5]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[6]	0.002333	0.04825	$9.491E-4$	0.0	0.0	0.0
gamma[7]	0.003	0.05469	0.001615	0.0	0.0	0.0
gamma[8]	0.1703	0.3759	0.04648	0.0	0.0	1.0
gamma[9]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[10]	0.06067	0.2387	0.02692	0.0	0.0	1.0
gamma[11]	0.8737	0.3322	0.04295	0.0	1.0	1.0
gamma[12]	0.003333	0.05764	0.001887	0.0	0.0	0.0

Πίνακας 5.5.1.2.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ του δειγματολήπτη Kuo & Mallick σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MCAR).

Στον Πίνακα 5.5.1.2.1 παρατηρούμε ότι οι συντελεστές του πραγματικού μοντέλου είναι σημαντικοί, ενώ στους δείκτες $\boldsymbol{\gamma}$ έχουμε ότι $\gamma_3, \gamma_5, \gamma_9 = 1$ και $\gamma_{11} = 0.8737$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.4546	0.4980
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1302	0.1906	0.3928
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1430	0.1336	0.3403
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	277	0.1180	0.3226
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10} + \boldsymbol{X}_{11}$	1814	0.0380	0.1912
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1429	0.0226	0.1488
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10} + \boldsymbol{X}_{11}$	1813	0.0160	0.1254
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9$	405	0.0070	0.0833
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9 + \boldsymbol{X}_{10} + \boldsymbol{X}_{11}$	1942	0.0053	0.0728
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1310	0.0036	0.0604

Πίνακας 5.5.1.2.2: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων του δειγματολήπτη Kuo & Mallick σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MCAR).

Από τον Πίνακα 5.5.1.2.2 διαπιστώνουμε ότι το πραγματικό μοντέλο κατέχει εκ των υστέρων πιθανότητα που είναι ίση με 0.4546. Ως εκ τούτου κοιτώντας τον πίνακα παρατηρούμε πως ο δειγματολήπτης Kuo & Mallick δίνει τη μεγαλύτερη εκ των υστέρων πιθανότητα στο πραγματικό μοντέλο.

5.5.1.3 Εφαρμογή της μεθόδου GVS

Εφαρμόζουμε το πλήρες μοντέλο με στόχο να πάρουμε τις εκτιμήσεις των παραμέτρων της ψευδο-πρότερης εκ των προτέρων κατανομής (Πίνακας 5.5.1.3.1).

Κώδικας εκτέλεσης του Full μοντέλου

```
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0), tau=0.2)}
data <- list(y=MCARIMPUTEDDatasetX$Y, x
    =as.matrix(MCARIMPUTEDDatasetX[,2:13]),p=12, n=80)
MCARmodelXpilotrun <- bugs(inits=inits, data=data, model.file =
    "Pilotrun.txt",parameters = c("beta0","beta", "sigma"),n.chains = 1,
    n.burnin = 2000, n.iter = 7000,n.thin = 1,n.sims = 5000,bugs.directory
    = "c:/Program Files/WinBUGS14/",debug=FALSE)
```

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.412	0.09237	0.001288	2.231	2.412	2.593
beta[1]	−0.5068	0.1118	0.001455	−0.7218	−0.5051	−0.2864
beta[2]	0.05333	0.1162	0.001423	−0.1759	0.05363	0.2802
beta[3]	2.818	0.1248	0.001863	2.576	2.82	3.064
beta[4]	−0.1389	0.1084	0.001315	−0.3546	−0.139	0.07405
beta[5]	−1.466	0.1116	0.001502	−1.682	−1.467	−1.247
beta[6]	0.01157	0.1143	0.001817	−0.2099	0.009922	0.238
beta[7]	0.09791	0.1077	0.001517	−0.1145	0.09797	0.3072
beta[8]	−0.34	0.1057	0.001193	−0.5482	−0.3387	−0.1304
beta[9]	1.046	0.1118	0.00147	0.8304	1.047	1.269
beta[10]	−0.314	0.1088	0.00156	−0.537	−0.3152	−0.1037
beta[11]	0.7383	0.1096	0.001381	0.5202	0.7388	0.9516
beta[12]	0.03829	0.1276	0.001772	−0.214	0.0371	0.2899

Πίνακας 5.5.1.3.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των συντελεστών του πλήρες μοντέλου.

Εφαρμογή της μεθόδου GVS

```
#We list our initial values and the dataset
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=MCARIMPUTEDDatasetX$Y, x
    =as.matrix(MCARIMPUTEDDatasetX[,2:13]),p=12, n=80, models=4096,
    mean=as.vector(colMeans(MCARmodelXpilotrun$sims.array[1:5000,1,])
)[2:13], se=as.vector(MCARmodelXpilotrun$sd$beta))
MCARmodelXgvs <- bugs(inits=inits,data=data,model.file =
"GVS.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 5000,n.thin = 1,n.sims = 3000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=TRUE)
MCARmodelXparametersgvs <- MCARmodelXgvs$summary[1:26,]
MCARmodelsXgvs <- MCARmodelXgvs$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
head(MCARmodelsXgvs[order(MCARmodelsXgvs[, 2], decreasing=TRUE),
    ],n=10)
```

Εφαρμόζοντας τον παραπάνω κώδικα στην R καταλήγουμε στους Πίνακες 5.5.1.3.2 και 5.5.1.3.3.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.411	0.1065	0.001853	2.2	2.412	2.621
beta[1]	−0.4785	0.1192	0.002627	−0.716	−0.4765	−0.2399
beta[2]	0.05152	0.1127	0.001981	−0.1584	0.0508	0.2709
beta[3]	2.744	0.1187	0.003134	2.508	2.744	2.974
beta[4]	−0.1398	0.1106	0.002105	−0.3582	−0.1413	0.08254
beta[5]	−1.657	0.1444	0.005381	−1.918	−1.665	−1.351
beta[6]	0.01211	0.1147	0.002334	−0.2054	0.009075	0.2425
beta[7]	0.09492	0.1108	0.002302	−0.1317	0.09645	0.3056
beta[8]	−0.345	0.1088	0.002485	−0.5587	−0.3467	−0.1283
beta[9]	1.011	0.1243	0.003502	0.7797	1.005	1.275
beta[10]	−0.3195	0.1092	0.001984	−0.5345	−0.3182	−0.1039
beta[11]	0.6151	0.1359	0.004018	0.3434	0.6179	0.8756
beta[12]	0.03465	0.1274	0.002486	−0.2235	0.03358	0.2813
sigma	0.9453	0.09992	0.003618	0.7637	0.9402	1.16
Δείκτες γ						
gamma[1]	0.426	0.4945	0.02272	0.0	0.0	1.0
gamma[2]	0.001667	0.04079	$7.156E-4$	0.0	0.0	0.0
gamma[3]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[4]	0.005	0.07053	0.001377	0.0	0.0	0.0
gamma[5]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[6]	0.002333	0.04825	$8.27E-4$	0.0	0.0	0.0
gamma[7]	0.002333	0.04825	$9.591E-4$	0.0	0.0	0.0
gamma[8]	0.1683	0.3742	0.01537	0.0	0.0	1.0
gamma[9]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[10]	0.139	0.3459	0.008538	0.0	0.0	1.0
gamma[11]	0.969	0.1733	0.006182	0.0	1.0	1.0
gamma[12]	0.001	0.03161	$5.694E-4$	0.0	0.0	0.0

Πίνακας 5.5.1.3.2: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MCAR).

Στον Πίνακα 5.5.1.3.2 βλέπουμε ότι οι συντελεστές $\beta_1, \beta_3, \beta_5, \beta_8, \beta_9, \beta_{10}$ και β_{11} είναι σημαντικοί και $\gamma_3, \gamma_5, \gamma_9 = 1$ και $\gamma_{11} = 0.969$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.4483	0.4974
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1302	0.2173	0.4125
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1430	0.1443	0.3514
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10} + \boldsymbol{X}_{11}$	1813	0.0760	0.2650
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10} + \boldsymbol{X}_{11}$	1814	0.0490	0.2159
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	277	0.0303	0.1715
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9 + \boldsymbol{X}_{10} + \boldsymbol{X}_{11}$	1942	0.0116	0.1073
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1429	0.0100	0.0995
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1309	0.0023	0.0482
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1333	0.0016	0.0407

Πίνακας 5.5.1.3.3: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MCAR).

Από τον Πίνακα 5.5.1.3.3 βλέπουμε πως η μέθοδος GVS δίνει τη μεγαλύτερη εκ των υστέρων πιθανότητα στο πραγματικό μοντέλο (ίση με 0.4483), ενώ έπειτα ακολουθούν τα μοντέλα που περιέχουν και τους συντελεστές β_1 και β_8.

5.5.1.4 Πλήρης απαρίθμηση με τη χρήση του κριτηρίου BIC

Εφαρμόζουμε πλήρη απαρίθμηση χρησιμοποιώντας τη συνάρτηση BICminimum που έχουμε κατασκευάσει στην R.

Εφαρμογή της συνάρτησης BICminimum στην R

```
Lowest10BICmodelsMCARX <-BICminimum(p=12,MCARIMPUTEDDatasetX)
#Coefficients of the models
Lowest10BICmodelsMCARX[[1]][1:10]
#their corresponding BIC values and model codes.
Lowest10BICmodelsMCARX[[2]][1:10]
Lowest10BICmodelsMCARX[[3]]
```

Εφαρμόζοντας τη συνάρτηση BICminimum καταλήγουμε στον Πίνακα 5.5.1.4.1.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Κριτήριο BIC
$X_1 + X_3 + X_5 + X_8 + X_9 + X_{10} + X_{11}$	1942	221.5333
$X_1 + X_3 + X_4 + X_5 + X_8 + X_9 + X_{10} + X_{11}$	1950	224.3223
$X_1 + X_3 + X_5 + X_7 + X_8 + X_9 + X_{10} + X_{11}$	2006	224.6631
$X_1 + X_2 + X_3 + X_5 + X_8 + X_9 + X_{10} + X_{11}$	1944	225.5346
$X_1 + X_3 + X_5 + X_8 + X_9 + X_{10} + X_{11} + X_{12}$	3990	225.5618
$X_1 + X_3 + X_5 + X_6 + X_8 + X_9 + X_{10} + X_{11}$	1974	225.7191
$X_1 + X_3 + X_5 + X_8 + X_9 + X_{11}$	1430	226.588
$X_1 + X_3 + X_4 + X_5 + X_7 + X_8 + X_9 + X_{10} + X_{11}$	2014	227.3532
$X_1 + X_2 + X_3 + X_4 + X_5 + X_8 + X_9 + X_{10} + X_{11}$	1952	228.0221
$X_1 + X_3 + X_4 + X_5 + X_8 + X_9 + X_{10} + X_{11} + X_{12}$	3998	228.2468

Πίνακας 5.5.1.4.1: Τα 10 πρώτα μοντέλα με τις μικρότερες τιμές του κριτηρίου BIC σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MCAR).

Από τον Πίνακα 5.5.1.4.1 βλέπουμε ότι το μοντέλο με τη μικρότερη τιμή του κριτηρίου BIC, αντιστοιχεί στο μοντέλο που περιέχει τους συντελεστές του πραγματικού μοντέλου μαζί με τους συντελεστές β_1, β_8 και β_{10}, ενώ το πραγματικό μοντέλο δε βρίσκεται στον παραπάνω πίνακα.

Μοντέλο με τη μικρότερη τιμή του κριτηρίου BIC

```
#We take the summary of the model with lowest BIC value
summary(lm(Y~ X1+X3+X5+X8+X9+X10+X11,data=MCARIMPUTEDDatasetX))
#Partial console output
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.98973   0.09404 21.158 < 2e-16 ***
X1         -0.45135   0.09758 -4.626 1.61e-05 ***
X3          2.81354   0.09463 29.732 < 2e-16 ***
X5         -1.47285   0.10217 -14.416 < 2e-16 ***
X8         -0.36096   0.10606 -3.403 0.00109 **
X9          1.06520   0.10336 10.305 8.16e-16 ***
X10        -0.34530   0.11501 -3.002 0.00368 **
X11         0.74632   0.10316  7.234 4.08e-10 ***
```

5.5.2 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές υπό το μηχανισμό MAR

Στη κατασκευή των μηχανισμών έλλειψης MAR θα χρησιμοποιήσουμε το λογιστικό μοντέλο όπου αναφέραμε στην Ενότητα 5.3.3. Ειδικότερα, τα μοντέλα για την πιθανότητα παρατήρησης της κάθε τιμής των επεξηγηματικών μεταβλητών $\boldsymbol{X}_3$, $\boldsymbol{X}_7$, $\boldsymbol{X}_{10}$ και $\boldsymbol{X}_{11}$ θα είναι τα εξής:

$$\begin{aligned}
\text{logit}\left(\mathbb{P}\left(R_{X_{3,i}}=1\right)\right) &= X_{12,i}+C,\\
\text{logit}\left(\mathbb{P}\left(R_{X_{7,i}}=1\right)\right) &= X_{6,i}+0.2\cdot X_{4,i}+C,\\
\text{logit}\left(\mathbb{P}\left(R_{X_{10,i}}=1\right)\right) &= 2\cdot X_{9,i}+X_{2,i}+C,\\
\text{logit}\left(\mathbb{P}\left(R_{X_{11,i}}=1\right)\right) &= 2\cdot X_{5,i}+X_{1,i}+C,
\end{aligned}$$

όπου R αποτελεί το δείκτη ελλιπών τιμών που έχουμε αναφέρει στην Ενότητα 4.2, το C είναι μία παράμετρος που συντονίζει (tuning parameter) το ποσοστό έλλειψης τιμών (εμείς παρακάτω θα έχουμε ότι $C=0.8$) και $i=1,\ldots,80$.

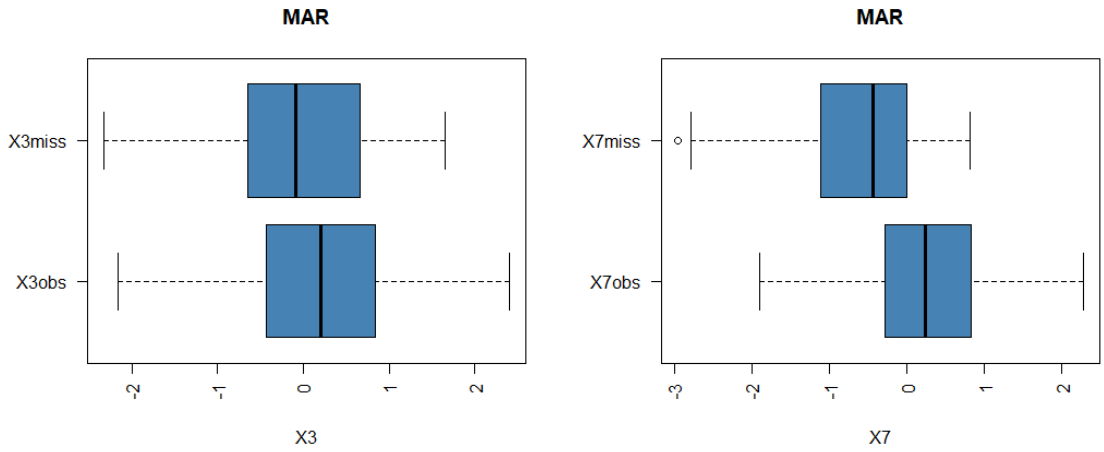
Κατασκευή μηχανισμού MAR στις επεξηγηματικές μεταβλητές X_3 και X_7

```
MARDatasetX <- FullDataset
#Tuning parameter
C <- 0.8
#Logistic model.
logistic <- function(x) exp(x) / (1 + exp(x))
modMARX3 <- 1*FullDataset$X12 + 0.4*FullDataset$X6 + C
r2.marx3 <-  rbinom(length(FullDataset$X3), 1, logistic(modMARX3))
# # of observations that remain observed after the MAR mechanism
sum(r2.marx3==1)
#console output is 54
#if r2.marx3=0 then the observation becomes a missing value
MARDatasetX$X3[r2.marx3==0] <- NA
#Number of missing values in x3.
sum(is.na(MARDatasetX$X3))
#console output is 26.

#Logistic model.
modMARX7 <- 1*FullDataset$X12 + 0.4*FullDataset$X6 + C
r2.marx7 <-  rbinom(length(FullDataset$X7), 1, logistic(modMARX7))
# # of observations that remain observed after the MAR mechanism
sum(r2.marx7==1)
#console output is 55
#if r2.marx7=0 then the observation becomes a missing value
MARDatasetX$X7[r2.marx7==0] <- NA
#Number of missing values in x7.
sum(is.na(MARDatasetX$X7))
#console output is 25.
```

Κατασκευή θηκογραμμάτων

```
#Number of observations that remain observed
sum(r2.marx3==1)
#console output is 54.
#Number of missing values in X3.
sum(is.na(MARDatasetX$X3))
#console output is 26.
#These are the observations that we took as missing
X3missMAR <- FullDataset$X3[is.na(MARDatasetX$X3)]
#Boxplot within the missing and the remaining observations.
boxplot(MARDatasetX$X3, X3missMAR, horizontal=TRUE,
    col='steelblue', las=2,main = "MAR",
xlab = "X3", names = c("X3obs","X3miss"))
#Number of observations that remain observed
sum(r2.marx7==1)
#console output is 55.
#Number of missing values in X7.
sum(is.na(MARDatasetX$X7))
#console output is 25.
#These are the observations that we took as missing
X7missMAR <- FullDataset$X7[is.na(MARDatasetX$X7)]
#Boxplot within the missing and the remaining observations.
boxplot(MARDatasetX$X7, X7missMAR, horizontal=TRUE,
    col='steelblue', las=2,main = "MAR",
xlab = "X7", names = c("X7obs","X7miss"))
```

Χρησιμοποιώντας τον παραπάνω κώδικα καταλήγουμε στο Διάγραμμα 5.5.2.1, όπου βλέπουμε ότι δεν υπάρχουν σημαντικές διαφορές μεταξύ των θηκογραμμάτων των τιμών των $X_{3\text{miss}}$, $X_{3\text{obs}}$. Αντιθέτως όμως βλέπουμε ότι η διάμεσος των τιμών του $X_{7\text{obs}}$ ξεπερνά το 3ο τεταρτημόριο του θηκογράμματος των τιμών του $X_{7\text{miss}}$.



Διάγραμμα 5.5.2.1: Θηκογράμματα των τιμών των $X_{3\text{miss}}$, $X_{3\text{obs}}$ και $X_{7\text{miss}}$, $X_{7\text{obs}}$.

Κατασκευή μηχανισμού MAR στις επεξηγηματικές μεταβλητές X_{10} και X_{11}

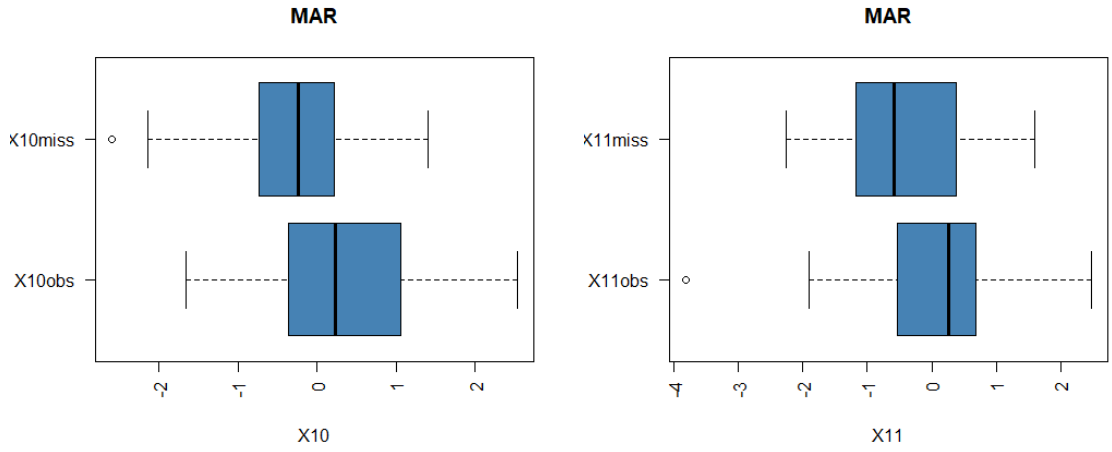
```
#Tuning parameter
C <- 0.8
#Logistic model.
logistic <- function(x) exp(x) / (1 + exp(x))
modMARX10 <- 2*FullDataset$X9 +1*FullDataset$X2 + C
r2.marx10 <-  rbinom(length(FullDataset$X10), 1, logistic(modMARX10))
# # of observations that remain observed after the MAR mechanism
sum(r2.marx10==1)
#console output is 54
#if r2.marx10=0 then the observation becomes a missing value
MARDatasetX$X10[r2.marx10==0] <- NA
#Number of missing values in x10.
sum(is.na(MARDatasetX$X10))
#console output is 26.

#Logistic model.
modMARX11 <- 2*FullDataset$X5+FullDataset$X1 + C
r2.marx11 <-  rbinom(length(FullDataset$X11), 1, logistic(modMARX11))
# # of observations that remain observed after the MAR mechanism
sum(r2.marx11==1)
#console output is 55
#if r2.marx11=0 then the observation becomes a missing value
MARDatasetX$X11[r2.marx11==0] <- NA
#Number of missing values in x11.
sum(is.na(MARDatasetX$X11))
#console output is 25.
```

Κατασκευή θηκογραμμάτων

```
#These are the observations that we took as missing
X10missMAR <- FullDataset$X10[is.na(MARDatasetX$X10)]
#Boxplot within the missing and the remaining observations.
boxplot(MARDatasetX$X10, X10missMAR, horizontal=TRUE,
    col='steelblue', las=2,main = "MAR",
xlab = "X10", names = c("X10obs","X10miss"))

#These are the observations that we took as missing
X11missMAR <- FullDataset$X11[is.na(MARDatasetX$X11)]
#Boxplot within the missing and the remaining observations.
boxplot(MARDatasetX$X1, X11missMAR, horizontal=TRUE,
    col='steelblue', las=2,main = "MAR",
xlab = "X11", names = c("X11obs","X11miss"))
```

Υλοποιώντας τον παραπάνω κώδικα στην R προκύπτει το Διάγραμμα 5.5.2.2, όπου παρατηρούμε πως δεν υπάρχουν σημαντικές διαφορές μεταξύ των

θηκογραμμάτων των τιμών των $X_{10\text{miss}}$, $X_{10\text{obs}}$, ενώ βλέπουμε ότι και εδώ το 1ο τεταρτημόριο του θηκογράμματος των τιμών του $X_{11\text{obs}}$ (οριακά) ξεπερνά τη διάμεσο των τιμών του $X_{11\text{miss}}$.



Διάγραμμα 5.5.2.2: Θηκογράμματα των τιμών των $X_{10\text{miss}}$, $X_{10\text{obs}}$ και $X_{11\text{miss}}$, $X_{11\text{obs}}$.

Για να αντικαταστήσουμε τις τιμές θα χρησιμοποιήσουμε το μοντέλο που δημιουργήσαμε στην Ενότητα 5.5.1 (imputationcovariates.txt).

Εφαρμογή Μοντέλου αντικατάστασης ελλιπών τιμών στις επεξηγηματικές μεταβλητές

```
#MAR for missing values in the covariates.
#Initial values
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, mu.x3=0, tau.x3=1,mu.x7=0, tau.x7=1, mu.x10=0,
    tau.x10=1, mu.x11=0, tau.x11=1)}
#Data inputs
data <- list(y=MARDatasetX$Y, x
    =as.matrix(MARDatasetX[,2:13]),p=12, n=80)
#We run the imputationcovariates.txt
MARImputationCovariates <- bugs(inits=inits,data=data,model.file =
"imputationcovariates.txt",parameters = c("beta0","beta",
    "sigma","x[,3]","x[,7]","x[,10]","x[,11]"),n.chains = 1,
    n.burnin = 2000, n.iter = 7000,n.thin = 1,n.sims =
    5000,bugs.directory = "c:/Program
Files/WinBUGS14/",debug=FALSE)
MARImputationCovariates
#The imputed values we estimated
colMeans(MARImputationCovariates$sims.array[1:5000,1,15:(dim(
MARImputationCovariates$sims.array)[3]-1)])
```

Έχοντας αποκτήσει τις τιμές των επεξηγηματικών μεταβλητών, μπορούμε να τις αντικαταστήσουμε στις ελλιπείς τιμές κάνοντας χρήση του παρακάτω κώδικα στην R.

Αντικατάσταση των ελλιπών τιμών στις επεξηγηματικές μεταβλητές

```
MARIMPUTEDDatasetX <- MARDatasetX
imputedvalues <-
    as.vector(colMeans(MARImputationCovariates$sims.array[1:5000,
1,15:(dim(MARImputationCovariates$sims.array)[3]-1)]))
counter <- 1
for (j in 1:nrow(MARIMPUTEDDatasetX)) {
   for (i in 1:ncol(MARIMPUTEDDatasetX)) {
      #If the xji point is missing then we impute it
      #with the corresponding value taken from the imputation model
      if(is.na(MARIMPUTEDDatasetX[j,i])==TRUE){
         MARIMPUTEDDatasetX[j,i] <- imputedvalues[counter]
         counter <- counter + 1
      }
   }
}
```

Ως εκ τούτου, στις παρακάτω υποενότητες θα εφαρμόσουμε τις μεθόδους SSVS, Kuo & Mallick, GVS και full enumeration στο δείγμα MARIMPUTEDDatasetX.

5.5.2.1 Εφαρμογή της μεθόδου SSVS

Για να εφαρμόσουμε τη μέθοδο SSVS (όπως και στην Ενότητα 5.3), θα χρησιμοποιήσουμε το αρχείο SSVS.txt που διατυπώσαμε στην Ενότητα 5.2.1. Με τη χρήση του παρακάτω κώδικα, υλοποιούμε τη μέθοδο SSVS και έτσι καταλήγουμε στους Πίνακες 5.5.2.1.1 και 5.5.2.1.2.

Εφαρμογή της μεθόδου SSVS

```
#Initial values and dataset and we implement the ssvs method
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=MARIMPUTEDDatasetX$Y, x
    =as.matrix(MARIMPUTEDDatasetX[,2:13]),p=12, n=80, models=4096)
MARmodelXssvs <- bugs(inits=inits,data=data,model.file =
    "SSVS.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 12000,n.thin = 5,n.sims = 10000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=FALSE)
MARmodelXparametersssvs <- MARmodelXssvs$summary[1:26,]
MARmodelsXssvs <- MARmodelXssvs$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
head(MARmodelsXssvs[order(MARmodelsXssvs[, 2], decreasing=TRUE),
    ],n=10)
```

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.41	0.09115	0.002419	2.228	2.409	2.587
beta[1]	−0.0667	0.1003	0.002247	−0.2604	−0.06543	0.1317
beta[2]	−0.03134	0.1054	0.002142	−0.2444	−0.03098	0.1779
beta[3]	2.751	0.1164	0.002443	2.518	2.75	2.986
beta[4]	0.09852	0.09857	0.002188	−0.094	−0.0969	0.2986
beta[5]	−1.933	0.1023	0.002098	−2.133	−1.935	−1.728
beta[6]	−0.09165	0.1144	0.002684	−0.309	−0.09456	0.1361
beta[7]	−0.1378	0.1062	0.002333	−0.3499	−0.1381	0.07279
beta[8]	$1.165E-4$	0.1026	0.001934	−0.1926	$-9.612E-5$	0.203
beta[9]	0.9846	0.1124	0.00224	0.7635	0.9847	1.208
beta[10]	0.09879	0.1015	0.002177	−0.1108	0.09988	0.2983
beta[11]	0.5349	0.09128	0.001791	0.353	0.5366	0.7081
beta[12]	0.2597	0.1173	0.002557	0.03036	0.2564	0.4849
sigma	0.7929	0.06988	0.001529	0.6808	0.7887	0.9362
Δείκτες γ						
gamma[1]	0.0125	0.1111	0.00241	0.0	0.0	1.0
gamma[2]	0.011	0.1043	0.002201	0.0	0.0	0.0
gamma[3]	1.0	0.0	$2.236E-12$	1.0	1.0	1.0
gamma[4]	0.012	0.1089	0.002802	0.0	0.0	0.0
gamma[5]	1.0	0.0	$2.236E-12$	1.0	1.0	1.0
gamma[6]	0.0155	0.1235	0.002911	0.0	0.0	0.0
gamma[7]	0.008	0.08908	0.002042	0.0	0.0	0.0
gamma[8]	0.0115	0.1066	0.002794	0.0	0.0	0.0
gamma[9]	0.556	0.4969	0.01058	0.0	1.0	1.0
gamma[10]	0.01	0.0995	0.001951	0.0	0.0	0.0
gamma[11]	0.0435	0.204	0.004573	0.0	0.0	1.0
gamma[12]	0.019	0.1365	0.002667	0.0	0.0	0.0

Πίνακας 5.5.2.1.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MAR).

Από τον Πίνακα 5.5.2.1.1 βλέπουμε ότι ο συντελεστής β_1 και οι συντελεστές του πραγματικού μοντέλου είναι σημαντικοί (τα διαστήματα αξιοπιστίας δεν περιέχουν την τιμή μηδέν). Στους δείκτες $\boldsymbol{\gamma}$ έχουμε ότι $\gamma_3, \gamma_5 = 1$, ενώ $\gamma_9 = 0.556$ και $\gamma_{11} = 0.0435$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	277	0.4790	0.4996
$\boldsymbol{X}_3 + \boldsymbol{X}_5$	21	0.3895	0.4877
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.0270	0.1621
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_{11}$	1045	0.0110	0.1043
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_{12}$	2069	0.0085	0.0918
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{12}$	2325	0.0075	0.0862
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9$	309	0.0070	0.0833
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6$	53	0.0065	0.0803
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	285	0.0055	0.0739
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5$	22	0.0050	0.0705

Πίνακας 5.5.2.1.2: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MAR).

Από τον Πίνακα 5.5.2.1.2 διαπιστώνουμε ότι το μοντέλο με τη μεγαλύτερη εκ των υστέρων πιθανότητα αντιστοιχεί στο μοντέλο με τους συντελεστές β_3, β_5 και β_9, ενώ στο πραγματικό μοντέλο βλέπουμε ότι έχουμε εκ των υστέρων πιθανότητα ίση με 0.0270.

5.5.2.2 Εφαρμογή της μεθόδου Kuo & Mallick

Εφαρμόζουμε το δειγματολήπτη Kuo & Mallick κάνοντας χρήση του αρχείου KuoMallick.txt που ορίσαμε στην Ενότητα 5.2.2.

Εφαρμογή της μεθόδου Kuo & Mallick

```
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=MARIMPUTEDDatasetX$Y, x
    =as.matrix(MARIMPUTEDDatasetX[,2:13]),p=12, n=80, models=4096)
MARmodelXKuoMallick <- bugs(inits=inits,data=data,model.file =
    "KuoMallick.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 5000,n.thin = 1,n.sims = 3000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=FALSE)
MARmodelXparametersKuoMallick <- MARmodelXKuoMallick$summary[1:26,]
MARmodelsXKuoMallick <- MARmodelXKuoMallick$summary[27:4122,]
head(MARmodelsXKuoMallick[order(MARmodelsXKuoMallick[, 2],
    decreasing=TRUE), ],n=10)
```

Εφαρμόζουμε τη μέθοδο Kuo & Mallick με χρήση του παραπάνω κώδικα και έτσι καταλήγουμε στους Πίνακες 5.5.2.2.1 και 5.5.2.2.2.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.41	0.08863	0.001534	2.238	2.409	2.583
beta[1]	−0.218	100.9	1.695	−198.8	0.4364	194.2
beta[2]	−2.272	99.68	1.553	−203.3	−1.294	191.8
beta[3]	2.792	0.09672	0.001836	2.604	2.79	2.986
beta[4]	0.3695	99.12	1.966	−194.6	0.6471	191.3
beta[5]	−1.894	0.09499	0.001836	−2.076	−1.896	−1.704
beta[6]	−2.387	102.3	2.064	−213.1	−0.8398	192.2
beta[7]	0.4584	100.8	2.039	−198.9	−0.6131	200.7
beta[8]	−0.1083	98.55	1.926	−186.6	−1.573	203.3
beta[9]	0.9828	0.09403	0.001651	0.7989	0.9832	1.168
beta[10]	1.33	99.77	1.736	−196.6	2.268	191.5
beta[11]	0.6096	0.09529	0.001973	0.4226	0.609	0.7986
beta[12]	0.1035	99.1	1.749	−197.6	0.2957	197.1
sigma	0.817	0.06809	0.0014	0.6979	0.813	0.9617
Δείκτες γ						
gamma[1]	0.002	0.04468	0.001383	0.0	0.0	0.0
gamma[2]	$3.333E-4$	0.01825	$3.289E-4$	0.0	0.0	0.0
gamma[3]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[4]	0.001	0.03161	$7.409E-4$	0.0	0.0	0.0
gamma[5]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[6]	0.002333	0.04825	0.001063	0.0	0.0	0.0
gamma[7]	0.003	0.05469	0.001395	0.0	0.0	0.0
gamma[8]	$6.667E-4$	0.02581	$6.578E-4$	0.0	0.0	0.0
gamma[9]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[10]	$6.667E-4$	0.02581	$4.651E-4$	0.0	0.0	0.0
gamma[11]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[12]	0.01167	0.1074	0.006992	0.0	0.0	0.0

Πίνακας 5.5.2.2.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ του δειγματολήπτη Kuo & Mallick σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MAR).

Στον Πίνακα 5.5.2.2.1 παρατηρούμε ότι οι συντελεστές που αντιστοιχούν στους

συντελεστές του πραγματικού μοντέλου είναι σημαντικοί, ενώ στους δείκτες $\boldsymbol{\gamma}$ έχουμε ότι $\gamma_3, \gamma_5, \gamma_9, \gamma_{11} = 1$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.9790	0.1434
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3349	0.0110	0.1043
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1365	0.0030	0.0546
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1333	0.0023	0.0482
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1302	0.0020	0.0446
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1309	0.0010	0.0316
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10} + \boldsymbol{X}_{11}$	1813	0.0006	0.0258
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3477	0.0003	0.0182
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1303	0.0000	0.0000
Intercept	1	0.0000	0.0000

Πίνακας 5.5.2.2.2: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων του δειγματολήπτη Kuo & Mallick σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MAR).

Από τον Πίνακα 5.5.2.2.2 διαπιστώνουμε ότι το πραγματικό μοντέλο κατέχει μία αρκετά υψηλή εκ των υστέρων πιθανότητα. Ως εκ τούτου μπορούμε να πούμε πως ο δειγματολήπτης Kuo & Mallick κατάφερε να ανιχνεύσει το πραγματικό μοντέλο.

5.5.2.3 Εφαρμογή της μεθόδου GVS

Εφαρμόζουμε το πλήρες μοντέλο με στόχο να πάρουμε τις εκτιμήσεις των παραμέτρων της ψευδο-πρότερης εκ των προτέρων κατανομής (Πίνακας 5.5.2.3.1).

Κώδικας εκτέλεσης του Full μοντέλου

```
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0), tau=0.2)}
data <- list(y=MARIMPUTEDDatasetX$Y, x
    =as.matrix(MARIMPUTEDDatasetX[,2:13]),p=12, n=80)
MARmodelXpilotrun <- bugs(inits=inits, data=data, model.file =
    "Pilotrun.txt",parameters = c("beta0","beta", "sigma"),n.chains = 1,
    n.burnin = 2000, n.iter = 7000,n.thin = 1,n.sims = 5000,bugs.directory
    = "c:/Program Files/WinBUGS14/",debug=FALSE)
```

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.412	0.09011	0.001256	2.235	2.412	2.588
beta[1]	−0.0878	0.108	0.001412	−0.2997	−0.08824	0.1248
beta[2]	0.05415	0.1136	0.001463	−0.2807	−0.05397	0.1672
beta[3]	2.733	0.1181	0.00177	2.498	2.733	2.965
beta[4]	0.1055	0.1023	0.001258	−0.09985	0.1055	0.3093
beta[5]	−1.946	0.1073	0.001467	−2.152	−1.947	−1.732
beta[6]	−0.07799	0.1211	0.001936	−0.3129	−0.07753	0.1624
beta[7]	−0.1689	0.1137	0.001579	−0.388	−0.1681	0.0508
beta[8]	−0.01648	0.1119	0.001267	−0.2389	−0.01548	0.205
beta[9]	1.026	0.1051	0.00135	0.8213	1.027	1.231
beta[10]	0.1202	0.1114	0.001559	−0.1074	0.1186	0.3369
beta[11]	0.5701	0.09685	0.00122	0.3774	0.5701	0.7588
beta[12]	0.3116	0.1305	0.001813	0.05339	0.3104	0.5691

Πίνακας 5.5.2.3.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των συντελεστών του πλήρες μοντέλου.

Εφαρμογή της μεθόδου GVS

```
#We list our initial values and the dataset
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=MARIMPUTEDDatasetX$Y, x
    =as.matrix(MARIMPUTEDDatasetX[,2:13]),p=12, n=80, models=4096,
    mean=as.vector(colMeans(MARmodelXpilotrun$sims.array[1:5000,1,])
)[2:13], se=as.vector(MARmodelXpilotrun$sd$beta))
MARmodelXgvs <- bugs(inits=inits,data=data,model.file =
"GVS.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 5000,n.thin = 1,n.sims = 3000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=TRUE)
#Here we have the summary of the model parameters
MARmodelXparametersgvs <- MARmodelXgvs$summary[1:26,]
MARmodelsXgvs <- MARmodelXgvs$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
head(MARmodelsXgvs[order(MARmodelsXgvs[, 2], decreasing=TRUE),
    ],n=10)
```

Εφαρμόζοντας τον παραπάνω κώδικα στην R καταλήγουμε στους Πίνακες 5.5.2.3.2 και 5.5.2.3.3.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.411	0.09188	0.001613	2.229	2.412	2.593
beta[1]	−0.08827	0.109	0.001835	−0.3024	−0.08732	0.1212
beta[2]	−0.05559	0.1097	0.001949	−0.2593	−0.05646	0.1584
beta[3]	2.791	0.09502	0.001533	2.6	2.79	2.973
beta[4]	0.1049	0.1043	0.001967	−0.101	0.1044	0.3146
beta[5]	−1.895	0.09446	0.002041	−2.076	−1.894	−1.711
beta[6]	−0.07733	0.1214	0.002472	−0.3077	−0.08053	0.1667
beta[7]	−0.1715	0.1166	0.002395	−0.4112	−0.1701	0.05352
beta[8]	−0.01609	0.113	0.002307	−0.2391	−0.01788	0.2082
beta[9]	0.9829	0.09374	0.00197	0.805	0.9815	1.176
beta[10]	0.1197	0.1105	0.001915	−0.09239	−0.1199	0.3358
beta[11]	0.6131	0.09476	0.001684	0.4263	0.616	0.7953
beta[12]	0.3072	0.1304	0.002545	0.04445	0.3056	0.5602
sigma	0.8171	0.0688	0.00128	0.6986	0.8132	0.9657
Δείκτες γ						
gamma[1]	0.0	0.0	$1.826E-12$	0.0	0.0	0.0
gamma[2]	0.001333	0.03649	$6.423E-4$	0.0	0.0	0.0
gamma[3]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[4]	0.001	0.03161	$5.66E-4$	0.0	0.0	0.0
gamma[5]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[6]	0.003333	0.05764	0.001069	0.0	0.0	0.0
gamma[7]	0.001333	0.03649	$6.453E-4$	0.0	0.0	0.0
gamma[8]	$3.333E-4$	0.01825	$3.289E-4$	0.0	0.0	0.0
gamma[9]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[10]	0.001	0.03161	$5.66E-4$	0.0	0.0	0.0
gamma[11]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[12]	0.009333	0.09616	0.0019	0.0	0.0	0.0

Πίνακας 5.5.2.3.2: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MAR).

Στον Πίνακα 5.5.2.3.2 βλέπουμε ότι οι συντελεστές $\beta_3, \beta_5, \beta_9, \beta_{11}$ και β_{12} είναι

σημαντικοί και $\gamma_3, \gamma_5, \gamma_9, \gamma_{11} = 1$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.9833	0.1280
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3349	0.0086	0.0927
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1333	0.0026	0.0515
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1365	0.0013	0.0364
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1303	0.0010	0.0316
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10} + \boldsymbol{X}_{11}$	1813	0.0010	0.0316
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1309	0.0006	0.0258
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1341	0.0003	0.0182
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1429	0.0003	0.0182
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3351	0.0003	0.0182

Πίνακας 5.5.2.3.3: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MAR).

Από τον Πίνακα 5.5.2.3.3 μας είναι φανερό ότι η μέθοδος GVS κατάφερε να ανιχνεύσει το πραγματικό μοντέλο, καθώς η εκ των υστέρων πιθανότητα είναι ίση με 0.9833.

5.5.2.4 Πλήρης απαρίθμηση με τη χρήση του κριτηρίου BIC

Εφαρμόζουμε πλήρη απαρίθμηση χρησιμοποιώντας τη συνάρτηση BICminimum που έχουμε κατασκευάσει στην R.

Εφαρμογή της συνάρτησης BICminimum στην R

```
Lowest10BICmodelsMARX <-BICminimum(p=12,MARIMPUTEDDatasetX)
#Coefficients of the models
Lowest10BICmodelsMARX[[1]][1:10]
#and their corresponding BIC values.
Lowest10BICmodelsMARX[[2]][1:10]
#Model code
Lowest10BICmodelsMARX[[3]]
```

Εφαρμόζοντας τη συνάρτηση BICminimum καταλήγουμε στο Πίνακα 5.5.2.4.1.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Κριτήριο BIC
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	1942	212.847
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	213.8796
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3413	213.9204
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3381	214.8865
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3350	215.7223
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1365	216.3527
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3357	216.7076
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1333	216.7646
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9 + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3477	216.8472
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10} + \boldsymbol{X}_{11} + \boldsymbol{X}_{12}$	3861	216.9115

Πίνακας 5.5.2.4.1: Τα 10 πρώτα μοντέλα με τις μικρότερες τιμές του κριτηρίου BIC σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (MAR).

Από το Πίνακα 5.5.2.4.1 βλέπουμε ότι το μοντέλο με τη μικρότερη τιμή του κριτηρίου BIC, αντιστοιχεί στο μοντέλο που περιέχει τους συντελεστές του πραγματικού μοντέλου μαζί με το συντελεστή β_{12}, ενώ το πραγματικό μοντέλο κατέχει τη 2η χαμηλότερη τιμή του BIC.

Μοντέλο με τη μικρότερη τιμή του κριτηρίου BIC

```
#We take the summary of the model with lowest BIC value
summary(lm(Y~ X3+X5+X9+X11+X12,data=MARIMPUTEDDatasetX))
#Partial console output
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.73318   0.08995 19.267 < 2e-16 ***
X3          2.72982   0.09783 27.904 < 2e-16 ***
X5         -1.94983   0.09467 -20.595 < 2e-16 ***
X9          1.04543   0.09448 11.065 < 2e-16 ***
X11         0.65275   0.10303  6.336 1.66e-08 ***
X12         0.22327   0.09808  2.276  0.0257 *
---
```

5.5.3 Εφαρμογή Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές υπό το μηχανισμό NMAR

Για τη κατασκευή των μηχανισμών έλλειψης NMAR στις επεξηγηματικές μεταβλητές θα χρησιμοποιήσουμε πάλι το λογιστικό μοντέλο όπως και στην Ενότητα 5.5.2. Έτσι για τις επεξηγηματικές μεταβλητές $\boldsymbol{X}_3$, $\boldsymbol{X}_7$, $\boldsymbol{X}_{10}$ και $\boldsymbol{X}_{11}$ τα μοντέλα έλλειψης του μηχανισμού NMAR είναι τα εξής:

$$\begin{aligned}
\text{logit}\left(\mathbb{P}\left(R_{X_{3,i}}=1\right)\right) &= X_{3,i}+C,\\
\text{logit}\left(\mathbb{P}\left(R_{X_{7,i}}=1\right)\right) &= X_{7,i}+C,\\
\text{logit}\left(\mathbb{P}\left(R_{X_{10,i}}=1\right)\right) &= X_{10,i}+C,\\
\text{logit}\left(\mathbb{P}\left(R_{X_{11,i}}=1\right)\right) &= X_{11,i}+C,
\end{aligned}$$

όπου εδώ η παράμετρος C είναι ίση με 0.5 και $i=1,\ldots,80$.

Κατασκευή μηχανισμού NMAR στις επεξηγηματικές μεταβλητές X_3 και X_7

```
NMARDatasetX <- FullDataset
#Tuning parameter
C <- 0.5
#Logistic model.
logistic <- function(x) exp(x) / (1 + exp(x))
modNMARX3 <- FullDataset$X3 + C
r2.nmarx3 <-  rbinom(length(FullDataset$X3), 1, logistic(modMARX3))
# # of observations that remain observed after the NMAR mechanism
```

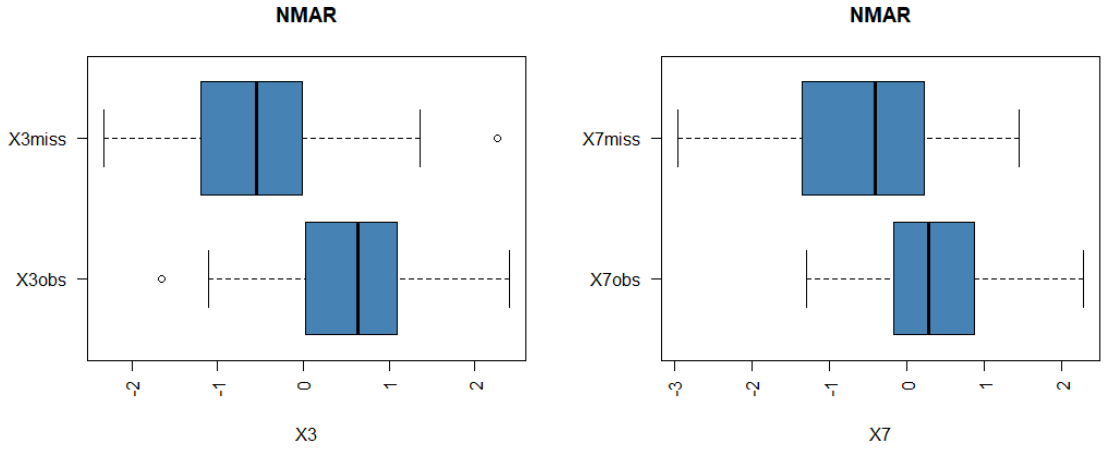
```
sum(r2.nmarx3==1)
#console output is 48
#if r2.nmarx3=0 then the observation becomes a missing value
NMARDatasetX$X3[r2.nmarx3==0] <- NA
#Number of missing values in x3.
sum(is.na(NMARDatasetX$X3))
#console output is 32.

#Logistic model.
modNMARX7 <- FullDataset$X7 + C
r2.nmarx7 <-  rbinom(length(FullDataset$X7), 1, logistic(modNMARX7))
# # of observations that remain observed after the NMAR mechanism
sum(r2.nmarx7==1)
#console output is 42
#if r2.nmarx7=0 then the observation becomes a missing value
NMARDatasetX$X7[r2.nmarx7==0] <- NA
#Number of missing values in x7.
sum(is.na(NMARDatasetX$X7))
#console output is 38.
```

Κατασκευή θηκογραμμάτων

```
#Number of observations that remain observed
sum(r2.nmarx3==1)
#console output is 48.
#Number of missing values in X3.
sum(is.na(NMARDatasetX$X3))
#console output is 32.
#These are the observations that we took as missing
X3missNMAR <- FullDataset$X3[is.na(NMARDatasetX$X3)]
#Boxplot within the missing and the remaining observations.
boxplot(NMARDatasetX$X3, X3missNMAR, horizontal=TRUE,
    col='steelblue', las=2,main = "NMAR",
xlab = "X3", names = c("X3obs","X3miss"))
#Number of observations that remain observed
sum(r2.nmarx7==1)
#console output is 42.
#Number of missing values in X7.
sum(is.na(NMARDatasetX$X7))
#console output is 38.
#These are the observations that we took as missing
X7missNMAR <- FullDataset$X7[is.na(NMARDatasetX$X7)]
#Boxplot within the missing and the remaining observations.
boxplot(NMARDatasetX$X7, X7missNMAR, horizontal=TRUE,
    col='steelblue', las=2,main = "NMAR",
xlab = "X7", names = c("X7obs","X7miss"))
```

Εφαρμόζοντας τον παραπάνω κώδικα στην R καταλήγουμε στο Διάγραμμα 5.5.3.1,

όπου όπως περιμέναμε υπάρχουν σημαντικές διαφορές μεταξύ του θηκογράμματος των τιμών του $X_{3\text{miss}}$ με αυτό των τιμών του $X_{3\text{obs}}$ και του θηκογράμματος των τιμών του $X_{7\text{miss}}$ με αυτό των τιμών του $X_{7\text{obs}}$.



Διάγραμμα 5.5.3.1: Θηκογράμματα των τιμών των $X_{3\text{miss}}$, $X_{3\text{obs}}$ και $X_{7\text{miss}}$, $X_{7\text{obs}}$.

Κατασκευή μηχανισμού NMAR στις επεξηγηματικές μεταβλητές X_{10} και X_{11}

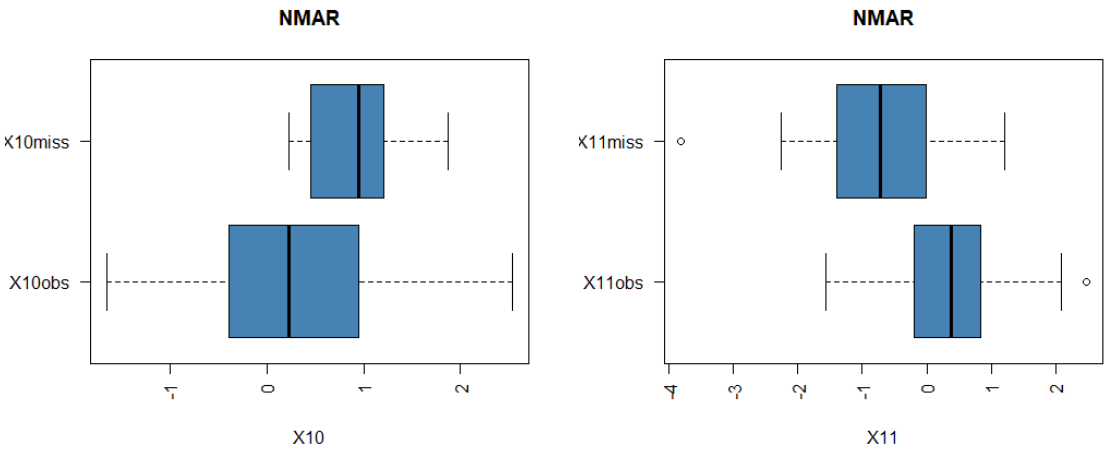
```
#Tuning parameter
C <- 0.5
#Logistic model.
logistic <- function(x) exp(x) / (1 + exp(x))
modNMARX10 <- FullDataset$X10 + C
r2.nmarx10 <- rbinom(length(FullDataset$X10), 1,
    logistic(modNMARX10))
# # of observations that remain observed after the NMAR mechanism
sum(r2.nmarx10==1)
#console output is 49
#if r2.nmarx10=0 then the observation becomes a missing value
NMARDatasetX$X10[r2.nmarx10==0] <- NA
#Number of missing values in x10.
sum(is.na(NMARDatasetX$X10))
#console output is 31.

#Logistic model.
modNMARX11 <- FullDataset$X11 + C
r2.nmarx11 <- rbinom(length(FullDataset$X11), 1,
    logistic(modNMARX11))
# # of observations that remain observed after the NMAR mechanism
sum(r2.nmarx11==1)
#console output is 49
#if r2.nmarx11=0 then the observation becomes a missing value
NMARDatasetX$X11[r2.nmarx11==0] <- NA
```

```
#Number of missing values in x11.
sum(is.na(NMARDatasetX$X11))
#console output is 31.
```

Κατασκευή θηκογραμμάτων

```
#These are the observations that we took as missing
X10missNMAR <- FullDataset$X10[is.na(NMARDatasetX$X10)]
#Boxplot within the missing and the remaining observations.
boxplot(NMARDatasetX$X10, X10missNMAR, horizontal=TRUE,
    col='steelblue', las=2,main = "NMAR",
xlab = "X10", names = c("X10obs","X10miss"))

#These are the observations that we took as missing
X11missNMAR <- FullDataset$X11[is.na(NMARDatasetX$X11)]
#Boxplot within the missing and the remaining observations.
boxplot(NMARDatasetX$X1, X11missNMAR, horizontal=TRUE,
    col='steelblue', las=2,main = "NMAR",
xlab = "X11", names = c("X11obs","X11miss"))
```

Υλοποιώντας τον παραπάνω κώδικα στην R προκύπτει το Διάγραμμα 5.5.3.2, όπου παρατηρούμε ότι και εδώ υπάρχουν σημαντικές διαφορές μεταξύ του θηκογράμματος των τιμών του $X_{11\text{miss}}$ με το θηκόγραμμα των τιμών του $X_{11\text{obs}}$, ενώ βλέπουμε ότι η διάμεσος του θηκογράμματος των τιμών του $X_{10\text{obs}}$ (οριακά) δεν ξεπερνά το 3ο τεταρτημόριο των τιμών του $X_{10\text{miss}}$.



Διάγραμμα 5.5.3.2: Θηκογράμματα των τιμών των $X_{10\text{miss}}$, $X_{10\text{obs}}$ και $X_{11\text{miss}}$, $X_{11\text{obs}}$.

Για να μπορέσουμε να αντικαταστήσουμε τις τιμές θα χρησιμοποιήσουμε το μοντέλο που δημιουργήσαμε στην Ενότητα 5.5.1 (imputationcovariates.txt).

Εφαρμογή Μοντέλου αντικατάστασης ελλιπών τιμών στις επεξηγηματικές μεταβλητές

```
#NMAR for missing values in the covariates.
#Initial values
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, mu.x3=0, tau.x3=1,mu.x7=0, tau.x7=1, mu.x10=0,
    tau.x10=1, mu.x11=0, tau.x11=1)}
#Data inputs
data <- list(y=NMARDatasetX$Y, x
    =as.matrix(NMARDatasetX[,2:13]),p=12, n=80)
#We run the imputationcovariates.txt
NMARImputationCovariates <- bugs(inits=inits,data=data,model.file =
"imputationcovariates.txt",parameters = c("beta0","beta",
    "sigma","x[,3]","x[,7]","x[,10]","x[,11]"),n.chains = 1,
    n.burnin = 2000, n.iter = 7000,n.thin = 1,n.sims =
    5000,bugs.directory = "c:/Program
Files/WinBUGS14/",debug=FALSE)
NMARImputationCovariates
#The imputed values we estimated
colMeans(NMARImputationCovariates$sims.array[1:5000,1,15:(dim(
NMARImputationCovariates$sims.array)[3]-1)])
```

Έχοντας αποκτήσει τις τιμές των επεξηγηματικών μεταβλητών, μπορούμε να τις αντικαταστήσουμε στις ελλιπείς τιμές κάνοντας χρήση του παρακάτω κώδικα στην R.

Αντικατάσταση των ελλιπών τιμών στις επεξηγηματικές μεταβλητές

```
NMARIMPUTEDDatasetX <- NMARDatasetX
imputedvalues <-
    as.vector(colMeans(NMARImputationCovariates$sims.array[1:5000,
1,15:(dim(NMARImputationCovariates$sims.array)[3]-1)]))
counter <- 1
for (j in 1:nrow(NMARIMPUTEDDatasetX)) {
    for (i in 1:ncol(NMARIMPUTEDDatasetX)) {
    #If the xji point is missing then we impute it
   #with the corresponding value taken from the imputation model
   if(is.na(NMARIMPUTEDDatasetX[j,i])==TRUE){
   NMARIMPUTEDDatasetX[j,i] <- imputedvalues[counter]
   counter <- counter + 1
      }
   }
}
```

Έτσι, στις παρακάτω υποενότητες θα εφαρμόσουμε τις μεθόδους SSVS, Kuo & Mallick, GVS και full enumeration στο δείγμα NMARIMPUTEDDatasetX.

5.5.3.1 Εφαρμογή της μεθόδου SSVS

Για να εφαρμόσουμε τη μέθοδο SSVS , θα χρησιμοποιήσουμε το αρχείο SSVS.txt που διατυπώσαμε στην Ενότητα 5.2.1. Με τη χρήση του παρακάτω κώδικα, υλοποιούμε τη μέθοδο SSVS και έτσι καταλήγουμε στους Πίνακες 5.5.3.1.1 και 5.5.3.1.2.

Εφαρμογή της μεθόδου SSVS

```
#Initial values and dataset and we implement the ssvs method
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=NMARIMPUTEDDatasetX$Y, x
    =as.matrix(NMARIMPUTEDDatasetX[,2:13]),p=12, n=80, models=4096)
NMARmodelXssvs <- bugs(inits=inits,data=data,model.file =
    "SSVS.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 12000,n.thin = 5,n.sims = 10000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=FALSE)
NMARmodelXparametersssvs <- NMARmodelXssvs$summary[1:26,]
NMARmodelsXssvs <- NMARmodelXssvs$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
head(NMARmodelsXssvs[order(NMARmodelsXssvs[, 2], decreasing=TRUE),
    ],n=10)
```

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.41	0.09562	0.002541	2.219	2.409	2.596
beta[1]	0.02967	0.1028	0.002311	−0.1683	0.03145	0.2335
beta[2]	−0.3068	0.1107	0.002345	−0.5202	−0.3048	−0.08851
beta[3]	2.604	0.1167	0.002489	2.368	2.604	2.835
beta[4]	0.1032	0.1022	0.002323	−0.09163	0.1023	0.3098
beta[5]	−1.704	0.1072	0.002261	−1.912	−1.707	−1.488
beta[6]	0.2416	0.1105	0.002636	0.0276	0.2412	0.4611
beta[7]	−0.2524	0.09898	0.002059	−0.4483	−0.2525	−0.05991
beta[8]	0.02444	0.1044	0.001959	−0.1749	0.02346	0.232
beta[9]	1.247	0.1124	0.002186	1.02	1.245	1.465
beta[10]	−0.1332	0.1052	0.002104	−0.355	−0.1312	0.06996
beta[11]	−0.04886	0.09793	0.002037	−0.2394	−0.04907	0.1366
beta[12]	0.06646	0.1235	0.002663	−0.1763	0.06416	0.3073
sigma	0.8316	0.07021	0.001599	0.7146	0.8272	0.9806
Δείκτες γ						
gamma[1]	0.012	0.1089	0.002415	0.0	0.0	0.0
gamma[2]	0.0185	0.1348	0.00305	0.0	0.0	0.0
gamma[3]	1.0	0.0	$2.236E-12$	1.0	1.0	1.0
gamma[4]	0.012	0.1089	0.002802	0.0	0.0	0.0
gamma[5]	1.0	0.0	$2.236E-12$	1.0	1.0	1.0
gamma[6]	0.018	0.133	0.003159	0.0	0.0	0.0
gamma[7]	0.011	0.1043	0.002326	0.0	0.0	0.0
gamma[8]	0.0115	0.1066	0.002794	0.0	0.0	0.0
gamma[9]	0.9365	0.2439	0.005405	0.0	1.0	1.0
gamma[10]	0.0115	0.1066	0.00241	0.0	0.0	0.0
gamma[11]	0.011	0.1043	0.002424	0.0	0.0	1.0
gamma[12]	0.014	0.1175	0.00269	0.0	0.0	0.0

Πίνακας 5.5.3.1.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (NMAR).

Από τον Πίνακα 5.5.3.1.1 βλέπουμε ότι οι συντελεστές $\beta_2, \beta_6, \beta_7$ και οι συντελεστές $\beta_3, \beta_5, \beta_9$ του πραγματικού μοντέλου είναι σημαντικοί (τα διαστήματα αξιοπιστίας δεν περιέχουν την τιμή μηδέν), ενώ παρατηρούμε ότι το διάστημα αξιοπιστίας του συντελεστή β_{11} περιέχει την τιμή μηδέν. Στους δείκτες $\boldsymbol{\gamma}$ έχουμε ότι $\gamma_3, \gamma_5 = 1$, ενώ $\gamma_9 = 0.9365$ και $\gamma_{11} = 0.011$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	277	0.8335	0.3726
$\boldsymbol{X}_3 + \boldsymbol{X}_5$	21	0.0560	0.2299
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	279	0.0150	0.1215
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9$	309	0.0145	0.1195
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{12}$	2325	0.0120	0.1089
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.0100	0.0995
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	278	0.0100	0.0995
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	285	0.0085	0.0918
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9$	285	0.0085	0.0918
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10}$	789	0.0085	0.0918

Πίνακας 5.5.3.1.2: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου SSVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (NMAR).

Από τον Πίνακα 5.5.3.1.2 διαπιστώνουμε ότι το μοντέλο με τη μεγαλύτερη εκ των υστέρων πιθανότητα αντιστοιχεί στο μοντέλο με τους συντελεστές β_3, β_5 και β_9, ενώ στο πραγματικό μοντέλο βλέπουμε ότι έχουμε εκ των υστέρων πιθανότητα ίση με 0.0100.

5.5.3.2 Εφαρμογή της μεθόδου Kuo & Mallick

Εφαρμόζουμε το δειγματολήπτη Kuo & Mallick κάνοντας χρήση του αρχείου KuoMallick.txt που ορίσαμε στην Ενότητα 5.2.2.

Εφαρμογή της μεθόδου Kuo & Mallick

```
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=NMARIMPUTEDDatasetX$Y, x
    =as.matrix(NMARIMPUTEDDatasetX[,2:13]),p=12, n=80, models=4096)
NMARmodelXKuoMallick <- bugs(inits=inits,data=data,model.file =
    "KuoMallick.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 5000,n.thin = 1,n.sims = 3000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=FALSE)
NMARmodelXparametersKuoMallick <-
    NMARmodelXKuoMallick$summary[1:26,]
NMARmodelsXKuoMallick <- NMARmodelXKuoMallick$summary[27:4122,]
head(NMARmodelsXKuoMallick[order(NMARmodelsXKuoMallick[, 2],
    decreasing=TRUE), ],n=10)
```

Εφαρμόζουμε τη μέθοδο Kuo & Mallick με χρήση του παραπάνω κώδικα και έτσι καταλήγουμε στους Πίνακες 5.5.3.2.1 και 5.5.3.2.2.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.409	0.09968	0.001726	2.216	2.408	2.604
beta[1]	−0.2536	100.9	1.694	−198.8	0.4422	194.2
beta[2]	−2.024	98.76	1.555	−202.5	−0.3574	191.8
beta[3]	2.717	0.1066	0.001977	2.51	2.715	2.93
beta[4]	0.3695	99.12	1.966	−194.6	0.6471	191.3
beta[5]	−1.691	0.1051	0.00215	−1.887	−1.693	−1.479
beta[6]	−2.404	102.5	2.063	−213.1	−1.008	192.7
beta[7]	0.4706	100.4	2.021	−198.2	−0.2565	200.7
beta[8]	−0.1083	98.55	1.926	−186.6	−1.573	203.3
beta[9]	1.166	0.1052	0.001894	0.961	1.165	1.375
beta[10]	1.58	99.37	1.714	−196.1	0.6445	191.5
beta[11]	−2.692	99.63	2.016	−204.6	−3.118	190.4
beta[12]	0.155	100.5	1.759	−199.5	0.436	197.2
sigma	0.9188	0.07627	0.001487	0.7843	0.9136	1.079
Δείκτες γ						
gamma[1]	0.001	0.03161	$9.868E-4$	0.0	0.0	0.0
gamma[2]	0.017	0.1293	0.006304	0.0	0.0	0.0
gamma[3]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[4]	0.001	0.03161	$7.409E-4$	0.0	0.0	0.0
gamma[5]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[6]	0.001	0.03161	$5.625E-4$	0.0	0.0	0.0
gamma[7]	0.009333	0.09616	0.004718	0.0	0.0	0.0
gamma[8]	$6.667E-4$	0.02581	$6.578E-4$	0.0	0.0	0.0
gamma[9]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[10]	0.01233	0.1104	0.006384	0.0	0.0	0.0
gamma[11]	0.002	0.04468	0.001701	0.0	0.0	0.0
gamma[12]	0.003333	0.05764	0.001887	0.0	0.0	0.0

Πίνακας 5.5.3.2.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ του δειγματολήπτη Kuo & Mallick σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (NMAR).

Στον Πίνακα 5.5.3.2.1 παρατηρούμε ότι οι συντελεστές β_3, β_5 και β_9 είναι σημαντικοί, ενώ βλέπουμε ότι το διάστημα αξιοπιστίας του συντελεστή β_{11} περιέχει την τιμή μηδέν. Παράλληλα στους δείκτες $\boldsymbol{\gamma}$ έχουμε ότι $\gamma_3, \gamma_5, \gamma_9 = 1$, ενώ για το συντελεστή β_{11} έχουμε ότι $\gamma_{11} = 0.002$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	277	0.9543	0.2087
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	279	0.0150	0.1215
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10}$	789	0.0123	0.1103
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9$	341	0.0073	0.0853
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{12}$	2325	0.0033	0.0576
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9$	343	0.0020	0.0446
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.0020	0.0446
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	278	0.0010	0.0316
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	285	0.0010	0.0316
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9$	309	0.0010	0.0316

Πίνακας 5.5.3.2.2: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων του δειγματολήπτη Kuo & Mallick σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (NMAR).

Από τον Πίνακα 5.5.3.2.2 διαπιστώνουμε ότι το μοντέλο που περιέχει τους συντελεστές β_3, β_5 και β_9 κατέχει την υψηλότερη εκ των υστέρων πιθανότητα. Επομένως αυτό σημαίνει ότι η μέθοδος Kuo & Mallick δεν κατάφερε να βρει το πραγματικό μοντέλο, καθώς η εκ των υστέρων πιθανότητα του είναι ίση με 0.0020.

5.5.3.3 Εφαρμογή της μεθόδου GVS

Εφαρμόζουμε το πλήρες μοντέλο με στόχο να πάρουμε τις εκτιμήσεις των παραμέτρων της ψευδο-πρότερης εκ των προτέρων κατανομής (Πίνακας 5.5.3.3.1).

Κώδικας εκτέλεσης του Full μοντέλου

```
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2)}
data <- list(y=NMARIMPUTEDDatasetX$Y, x
    =as.matrix(NMARIMPUTEDDatasetX[,2:13]),p=12, n=80)
NMARmodelXpilotrun <- bugs(inits=inits, data=data, model.file =
    "Pilotrun.txt",parameters = c("beta0","beta", "sigma"),n.chains
    = 1, n.burnin = 2000, n.iter = 7000,n.thin = 1,n.sims =
    5000,bugs.directory = "c:/Program Files/WinBUGS14/",debug=FALSE)
```

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.412	0.09472	0.00132	2.226	2.412	2.597
beta[1]	0.03062	0.1112	0.001407	−0.1857	−0.02983	0.2501
beta[2]	−0.3589	0.1211	0.001518	−0.5986	−0.3578	−0.1199
beta[3]	2.576	0.1195	0.001784	2.341	2.575	2.808
beta[4]	0.1136	0.1066	0.001296	−0.09954	0.1134	0.328
beta[5]	−1.715	0.1125	0.001539	−1.931	−1.715	−1.492
beta[6]	0.2913	0.1163	0.001848	0.06608	0.2894	0.5234
beta[7]	−0.2952	0.1043	0.001489	−0.4993	−0.2948	−0.09067
beta[8]	0.02497	0.1139	0.00128	−0.202	0.02594	0.2496
beta[9]	1.263	0.1096	0.001465	1.047	1.265	1.477
beta[10]	−0.1308	0.1151	0.001651	−0.3618	−0.1319	0.09451
beta[11]	0.04819	0.1059	0.001344	−0.2599	−0.04809	0.1583
beta[12]	0.09858	0.1386	0.001925	−0.1756	0.09728	0.372

Πίνακας 5.5.3.3.1: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των συντελεστών του πλήρες μοντέλου.

Εφαρμογή της μεθόδου GVS

```
inits <- function(){list(beta0=0,beta=c(0,0,0,0,0,0,0,0,0,0,0,0),
    tau=0.2, gamma=c(1,1,1,1,1,1,1,1,1,1,1,1))}
data <- list(y=NMARIMPUTEDDatasetX$Y, x
    =as.matrix(NMARIMPUTEDDatasetX[,2:13]),p=12, n=80, models=4096,
    mean=as.vector(colMeans(NMARmodelXpilotrun$sims.array[1:5000,1,])
)[2:13], se=as.vector(NMARmodelXpilotrun$sd$beta))
NMARmodelXgvs <- bugs(inits=inits,data=data,model.file =
"GVS.txt",parameters = c("beta0","beta", "gamma",
    "sigma","permodelselect"),n.chains = 1, n.burnin = 2000, n.iter
    = 5000,n.thin = 1,n.sims = 3000,bugs.directory = "c:/Program
    Files/WinBUGS14/",debug=TRUE)
NMARmodelXparametersgvs <- NMARmodelXgvs$summary[1:26,]
NMARmodelsXgvs <- NMARmodelXgvs$summary[27:4122,]
#We print the first 10 models with the highest posterior probability
head(NMARmodelsXgvs[order(NMARmodelsXgvs[, 2], decreasing=TRUE),
    ],n=10)
```

Εφαρμόζοντας τον παραπάνω κώδικα στην R καταλήγουμε στους Πίνακες 5.5.3.3.2 και 5.5.3.3.3.

Παράμετροι μοντέλου	Μέση τιμή	Τυπική απόκλιση	Σφάλμα MC	2.5%	Διάμεσος	97.5%
beta0	2.411	0.1033	0.001816	2.207	2.412	2.615
beta[1]	0.03011	0.1122	0.001889	−0.1903	0.03111	0.2458
beta[2]	−0.358	0.1174	0.002076	−0.5777	−0.3591	−0.1315
beta[3]	2.715	0.1052	0.001676	2.508	2.716	2.916
beta[4]	0.113	0.1086	0.002049	−0.1016	0.1125	0.3314
beta[5]	−1.691	0.1047	0.002175	−1.895	−1.691	−1.486
beta[6]	0.2916	0.1171	0.002404	0.06836	0.2888	0.5263
beta[7]	−0.297	0.1074	0.002238	−0.5176	−0.2958	−0.09074
beta[8]	0.02533	0.1149	0.002347	−0.2015	0.02354	0.2535
beta[9]	1.166	0.105	0.002171	0.9645	0.165	1.381
beta[10]	−0.1322	0.1147	0.002002	−0.3568	−0.1315	0.092
beta[11]	−0.04697	0.1059	0.00178	−0.2564	−0.04606	0.1547
beta[12]	0.09467	0.1384	0.002701	−0.1851	0.09349	0.3626
sigma	0.8171	0.0688	0.00128	0.6986	0.8132	0.9657
Δείκτες γ						
gamma[1]	$3.333E-4$	0.01825	$3.35E-4$	0.0	0.0	0.0
gamma[2]	0.02333	0.151	0.002853	0.0	0.0	0.0
gamma[3]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[4]	$6.667E-4$	0.02581	$4.651E-4$	0.0	0.0	0.0
gamma[5]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[6]	0.002333	0.04825	$8.27E-4$	0.0	0.0	0.0
gamma[7]	0.007333	0.08532	0.001682	0.0	0.0	0.0
gamma[8]	$6.667E-4$	0.02581	$4.608E-4$	0.0	0.0	0.0
gamma[9]	1.0	0.0	$1.826E-12$	1.0	1.0	1.0
gamma[10]	0.009333	0.09616	0.001491	0.0	0.0	0.0
gamma[11]	0.002	0.04468	$7.746E-4$	0.0	0.0	0.0
gamma[12]	$3.333E-4$	0.01825	$3.35E-4$	0.0	0.0	0.0

Πίνακας 5.5.3.3.2: Περιγραφικοί δείκτες των προσομοιωμένων τιμών της ύστερης κατανομής των παραμέτρων του μοντέλου και του διανύσματος $\boldsymbol{\gamma}$ της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (NMAR).

Στον Πίνακα 5.5.3.3.2 βλέπουμε ότι οι συντελεστές $\beta_2, \beta_3, \beta_5, \beta_7$ και β_9 είναι σημαντικοί και $\gamma_3, \gamma_5, \gamma_9 = 1$ και $\gamma_{11} = 0.002$.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Μέση τιμή	Τυπική απόκλιση
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	277	0.9543	0.2087
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	279	0.0230	0.1499
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{10}$	789	0.0093	0.0961
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_7 + \boldsymbol{X}_9$	341	0.0070	0.0833
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_6 + \boldsymbol{X}_9$	309	0.0023	0.0482
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1301	0.0013	0.0364
$\boldsymbol{X}_3 + \boldsymbol{X}_4 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	285	0.0006	0.0258
$\boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_8 + \boldsymbol{X}_9$	405	0.0006	0.0258
$\boldsymbol{X}_1 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9$	278	0.0003	0.0182
$\boldsymbol{X}_2 + \boldsymbol{X}_3 + \boldsymbol{X}_5 + \boldsymbol{X}_9 + \boldsymbol{X}_{11}$	1303	0.0003	0.0182

Πίνακας 5.5.3.3.3: Περιγραφικοί δείκτες των εκ των υστέρων πιθανοτήτων των 10 πρώτων μοντέλων της μεθόδου GVS σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (NMAR).

Από τον Πίνακα 5.5.3.3.3 φαίνεται πως και η μέθοδος GVS ευνοεί το μοντέλο με του συντελεστές β_3, β_5 και β_9 διότι η εκ των υστέρων πιθανότητα είναι ίση με 0.9543. Αντιθέτως βλέπουμε ότι στο πραγματικό μοντέλο έχουμε μία αρκετά μικρή εκ των υστέρων πιθανότητα (ίση με 0.0013).

5.5.3.4 Πλήρης απαρίθμηση με τη χρήση του κριτηρίου BIC

Εφαρμόζουμε πλήρη απαρίθμηση χρησιμοποιώντας τη συνάρτηση BICminimum που έχουμε κατασκευάσει στην R.

Εφαρμογή της συνάρτησης BICminimum στην R

```
Lowest10BICmodelsNMARX <-BICminimum(p=12,NMARIMPUTEDDatasetX)
#Coefficients of the models
Lowest10BICmodelsNMARX[[1]][1:10]
#and their corresponding BIC values.
Lowest10BICmodelsNMARX[[2]][1:10]
#Model code
Lowest10BICmodelsNMARX[[3]]
```

Εφαρμόζοντας τη συνάρτηση BICminimum καταλήγουμε στον Πίνακα 5.5.3.4.1.

Παράμετροι μοντέλου	Αριθμός μοντέλου	Κριτήριο BIC
$X_2+X_3+X_5+X_6+X_7+X_9$	375	222.9534
$X_2+X_3+X_5+X_6+X_7+X_9+X_{10}$	887	224.987
$X_2+X_3+X_5+X_6+X_7+X_9+X_{12}$	2423	225.1052
$X_2+X_3+X_4+X_5+X_6+X_7+X_9$	383	225.8568
$X_2+X_3+X_5+X_6+X_7+X_9+X_{11}$	1399	226.9812
$X_1+X_2+X_3+X_5+X_6+X_7+X_7+X_9$	376	227.0011
$X_2+X_3+X_5+X_7+X_9$	343	227.2031
$X_2+X_3+X_5+X_6+X_9$	311	227.26
$X_2+X_3+X_5+X_6+X_7+X_8+X_9$	503	227.3001
$X_2+X_3+X_5+X_9$	279	227.3904

Πίνακας 5.5.3.4.1: Τα 10 πρώτα μοντέλα με τις μικρότερες τιμές του κριτηρίου BIC σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (NMAR).

Από το Πίνακα 5.5.3.4.1 βλέπουμε ότι το μοντέλο με τη μικρότερη τιμή του κριτηρίου BIC, αντιστοιχεί στο μοντέλο που περιέχει τους συντελεστές του πραγματικού μοντέλου μαζί με το συντελεστή β_{12}, ενώ το πραγματικό μοντέλο βλέπουμε πως δε βρίσκεται στον παραπάνω πίνακα.

Μοντέλο με τη μικρότερη τιμή του κριτηρίου BIC

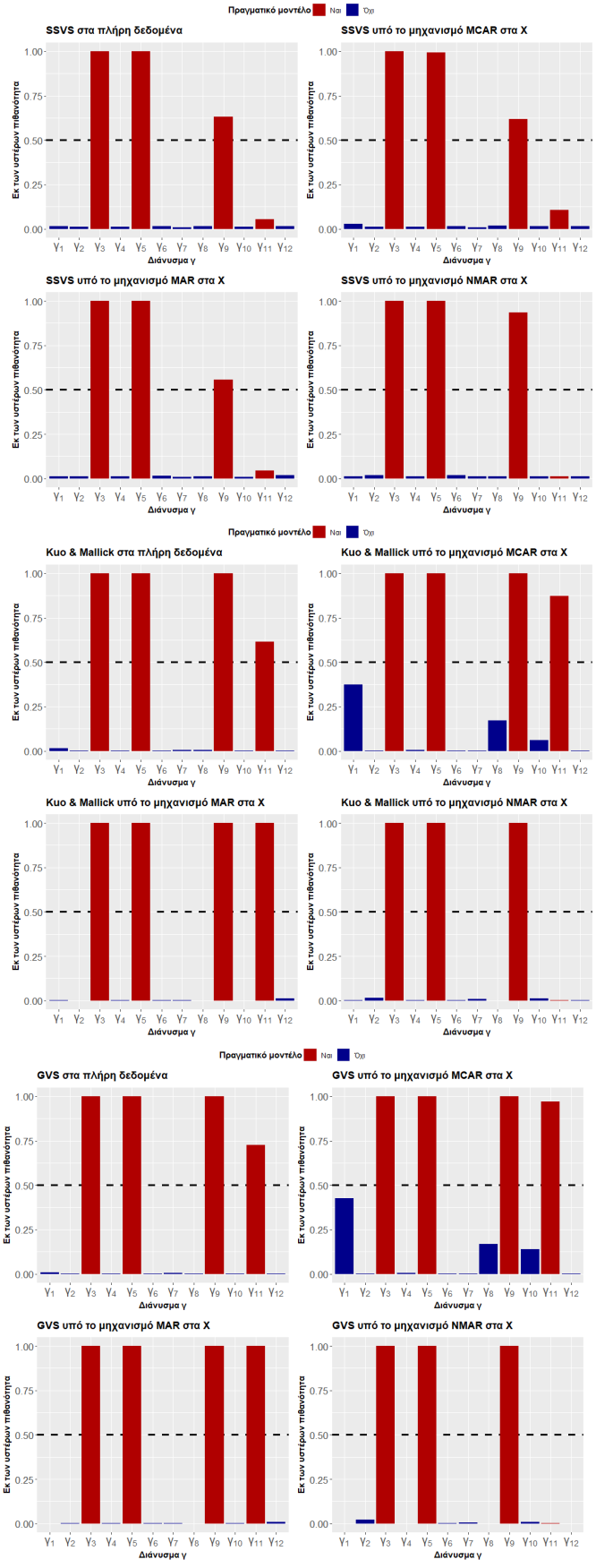
```
#We take the summary of the model with lowest BIC value
summary(lm(Y~ X2+X3+X5+X6+X7+X9,data=NMARIMPUTEDDatasetX))
#Partial console output
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.74374   0.11514 15.145 < 2e-16 ***
X2         -0.34398   0.10032 -3.429 0.00100 ***
X3          2.88332   0.11042 26.112 < 2e-16 ***
X5         -1.64599   0.09393 -17.523 < 2e-16 ***
X6          0.34312   0.11898  2.884 0.00516 **
X7         -0.50656   0.17504 -2.894 0.00501 **
X9          1.21526   0.09826 12.368 < 2e-16 ***
```

5.6 Σύγκριση των Μπεϋζιανών μεθόδων επιλογής μεταβλητών σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές

Όπως είδαμε στην προηγούμενη ενότητα, η επιλογή των μεθόδων επιλογής μεταβλητών μπορεί να επηρεάσει σημαντικά την ποιότητα και την αξιοπιστία των αποτελεσμάτων, ειδικά όταν πρόκειται για σύνολα δεδομένων που περιέχουν ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (μέσω κάποιου MCAR, MAR ή NMAR μηχανισμού). Υλοποιώντας στην Ενότητα 5.5 μία εκτενή εφαρμογή των μεθόδων SSVS, Kuo & Mallick και GVS, είδαμε ότι η απόδοσή τους μπορεί να ποικίλλει υπό διαφορετικούς μηχανισμούς έλλειψης δεδομένων. Σε αυτή την ενότητα χρησιμοποιώντας τα αποτελέσματα της Ενότητας 5.5, θα συγκρίνουμε τις τρεις αυτές μεθόδους με στόχο να δώσουμε πολύτιμα συμπεράσματα σχετικά με την προσαρμοστικότητα και την αποτελεσματικότητα των μεθόδων αυτών.

Το Διάγραμμα 5.6.1 περιέχει τα ραβδογράμματα των δεικτών $\boldsymbol{\gamma}$ από τις μεθόδους SSVS, GVS και το δειγματολήπτη Kuo & Mallick στα πλήρη δεδομένα, αλλά και σε δεδομένα με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές (με μηχανισμούς έλλειψης MCAR, MAR και NMAR).

Από τα ραβδογράμματα της μεθόδου SSVS στο Διάγραμμα 5.6.1 καταλαβαίνουμε ότι η SSVS δεν κατάφερε να ανιχνεύσει την επεξηγηματική μεταβλητή $\boldsymbol{X}_{11}$ που βρίσκεται στο πραγματικό μοντέλο, καθώς ο δείκτης γ_{11} κατέχει τιμές κοντά στο μηδέν σε όλες τις περιπτώσεις. Ωστόσο από τα ραβδογράμματα του δειγματολήπτη Kuo & Mallick και της μεθόδου GVS στους μηχανισμούς MCAR και MAR παρατηρούμε ότι κατάφεραν να ανιχνεύσουν το πραγματικό μοντέλο, καθώς οι τιμές των δεικτών $\boldsymbol{\gamma}$ που αντιστοιχούν στις επεξηγηματικές μεταβλητές του πραγματικού μοντέλου κατέχουν τιμές πολύ κοντά στο ένα. Όμως στην περίπτωση που έχουμε το μηχανισμό NMAR βλέπουμε πως ο δείκτης γ_{11} κατέχει μία αρκετά μικρή τιμή. Κατά συνέπεια λοιπόν καταλαβαίνουμε πως όταν η έλλειψη παρατηρήσεων στις επεξηγηματικές μεταβλητές είναι NMAR, ο δειγματολήπτης Kuo & Mallick και η μέθοδος GVS δεν καταφέρνουν να ανιχνεύσουν την επεξηγηματική μεταβλητή $\boldsymbol{X}_{11}$, που είναι οριακά σημαντική στο πραγματικό μοντέλο. Λαμβάνοντας έτσι υπόψιν και τα συμπεράσματα που λάβαμε από την Ενότητα 5.4, είναι εύκολα αντιληπτό ότι η μέθοδος GVS και ο δειγματολήπτης Kuo & Mallick υπερτερούν της μεθόδου SSVS στη συγκεκριμένη εφαρμογή.



Διάγραμμα 5.6.1: Ραβδογράμματα των δεικτών $\boldsymbol{\gamma}$ στις μεθόδους SSVS, Kuo & Mallick και GVS σε δεδομένα με μηχανισμούς έλλειψης στις επεξηγηματικές μεταβλητές $\boldsymbol{X}$.

6 Επίλογος και συμπεράσματα

Στην αρχή αυτής της διπλωματικής εργασίας έγινε μία ενδελεχής ανασκόπηση της κλασικής θεωρίας της Μπεϋζιανής στατιστικής, έχοντας ως απώτερο σκοπό την κατανόησή της. Συγκεκριμένα στο Κεφάλαιο 1 πραγματοποιήθηκε μία σύγκριση της κλασικής στατιστική με την Μπεϋζιανή. Στη σύγκριση αυτή είδαμε πως αν και αυτές οι προσεγγίσεις διαφέρουν στις φιλοσοφίες και τις μεθοδολογίες τους, δεν παύουν να εμπλουτίζουν τον κόσμο της ανάλυσης δεδομένων. Η κλασική στατιστική, βασισμένη στην ιδέα της εμπειρικής πιθανότητας, μας προσφέρει εύκολα κατανοήσημες τεχνικές εκτίμησης παραμέτρων, ενώ η Μπεϋζιανή στατιστική, καθοδηγούμενη από την κομψότητα της θεωρίας πιθανοτήτων, επεκτείνει τα σύνορα της στατιστικής συμπερασματολογίας με την ικανότητά της να ενσωματώνει εκ των προτέρων γνώσεις και να ποσοτικοποιεί την αβεβαιότητα των αποτελεσμάτων. Μέσω αυτής της σύγκρισης, είδαμε ότι κάθε προσέγγιση έχει τα δικά της πλεονεκτήματα και μειονεκτήματα και έτσι η επιλογή για το ποια προσέγγιση θα χρησιμοποιηθεί, εξαρτάται από τις προτιμήσεις του ίδιου του ερευνητή. Εν συνεχεία, στο Κεφάλαιο 1, ορίσαμε το θεώρημα του Bayes, μιλήσαμε για τις διάφορες μορφές εκ των προτέρων κατανομών και τέλος δώσαμε μία ολοκληρωμένη εισαγωγή σε κάποιους από τους πιο γνωστούς αλγορίθμους Markov Chain Monte Carlo (MCMC).

Στο Κεφάλαιο 2 αναπτύξαμε τη θεωρία των Μπεϋζιανών μεθόδων επιλογής μοντέλων και του παράγοντα του Bayes. Η επιλογή μοντέλων μέσω του παράγοντα Bayes μας προσφέρει μία πιθανολογική προσέγγιση στη σύγκριση μοντέλων και μας δίνει τη δυνατότητα να επιλέξουμε μοντέλα που επιτυγχάνουν μια ισορροπία μεταξύ της προσαρμοστικότητας και της πολυπλοκότητας. Ωστόσο, σε αυτή την ανάπτυξη θεωρίας και ιδεών, ήρθαμε αντιμέτωποι με το παράδοξο των Lindley-Bartlett, όπου μας προσγειώνει στο γεγονός ότι η επιλογή των παραμέτρων της εκ των προτέρων κατανομής παίζει σημαντικό ρόλο σε Μπεϋζιανά προβλήματα επιλογής μοντέλων με χρήση του παράγοντα του Bayes, διότι μπορεί να οδηγήσει σε τυχόν μεροληψίες ή στην προτίμηση του απλούστερου μοντέλου. Παράλληλα διακρίναμε ότι η περιθώρια πιθανοφάνεια όπου αποτελεί ένα θεμελιώδες συστατικό του παράγοντα του Bayes, απαιτεί συχνά δύσκολους σε εμάς υπολογισμούς, ειδικά όταν ασχολούμαστε με πολλές παραμέτρους. Συνεπώς οι υπολογιστικές τεχνικές, όπως η μέθοδος Laplace, ο δειγματολήπτης σπουδαιότητας και μέθοδοι βασισμένοι σε MCMC ήταν φυσικό επακόλουθο να αναπτυχθούν για την εκτίμηση της περιθώριας πιθανοφάνειας. Στο τέλος του Κεφαλαίου 2 μιλήσαμε για τις θετικές και αρνητικές πτυχές των Μπεϋζιανών μεθόδων επιλογής μοντέλων. Κατά την αξιολόγηση των θετικών στοιχείων, επισημάνθηκε ότι οι Μπεϋζιανές μέθοδοι επιλογής μοντέλων, παρέχουν ερμηνεύσιμα αποτελέσματα και τη δυνατότητα επιλογής ενός συνόλου εύλογων μοντέλων, σε περίπτωση που υπάρχει αβεβαιότητα επιλογής ενός μόνο μοντέλου. Παρ'όλα αυτά, η ευαισθησία των αποτελεσμάτων ως προς την επιλογή της εκ των προτέρων κατανομής, καθιστά ένα πρόβλημα το οποίο πρέπει να εξεταστεί σε μεγάλο βαθμό από τον ερευνητή.

Στο Κεφάλαιο 3 ασχοληθήκαμε με το πολλαπλό γραμμικό μοντέλο υπό την Μπεϋζιανή προσέγγιση και επισημάναμε τη χρήση συζυγών εκ των προτέρων κατανομών για του συντελεστές του μοντέλου, καθώς καθιστούν τους υπολογισμούς των εκ των υστέρων κατανομών πιο εύκολους. Συγκεκριμένα, έγινε χρήση της

ασαφής Κανονικής κατανομής για τους συντελεστές του μοντέλου (και χρήση της αντίστροφης Γάμμα κατανομής για την άγνωστη παράμετρο σ^2). Επίσης προσδιορίσαμε διάφορες παραλλαγές της εκ των προτέρων κατανομής g του Zellner (1986) και μιλήσαμε για τη σημαντικότητα που έχει η επιλογή της παραμέτρου g. Στη συνεχεία παρουσιάσαμε τη μέθοδο Stochastic Search Variable Selection (SSVS), η οποία χρησιμοποιεί το διάνυσμα δείκτη $\boldsymbol{\gamma}$ για την επιλογή των συντελεστών και μία μείξη Κανονικών κατανομών ως εκ των προτέρων κατανομή των συντελεστών β_i. Επιπλέον μελετήσαμε το δειγματολήπτη Kuo & Mallick, όπου εδώ το διάνυσμα $\boldsymbol{\gamma}$ συμπεριλαμβάνεται στο μοντέλο και είναι ανεξάρτητο της εκ των προτέρων κατανομής των συντελεστών του μοντέλου. Τέλος στο κεφάλαιο αυτό, παρουσιάσαμε τη μέθοδο Gibbs Variable Selection (GVS), η οποία ενώνει τα σκεπτικά των μεθόδων SSVS και Kuo & Mallick.

Στο Κεφάλαιο 4, εξερευνήσαμε τους διάφορους μηχανισμούς έλλειψης τιμών: Εντελώς τυχαία έλλειψη (MCAR), τυχαία έλλειψη (MAR) και μη τυχαία έλλειψη (NMAR). Συζητήσαμε μια σειρά από τεχνικές για την αντιμετώπιση αυτού του διάχυτου προβλήματος, ξεκινώντας από μεθόδους απλής αντικατάστασης (single imputation), όπου οι ελλιπείς τιμές αντικαθίστανται με μεμονωμένες εκτιμήσεις, ενώ στη συνέχεια εμβαθύναμε σε μεθόδους πολλαπλής αντικατάστασης (multiple imputation). Στην αναζήτησή μας αυτή, εξερευνήσαμε επίσης μεθόδους διαχείρισης ελλιπών τιμών υπό την Μπεϋζιανή προσέγγιση, οι οποίες αξιοποιούν τη δύναμη της Μπεϋζιανής στατιστικής για το χειρισμό των ελλιπών τιμών. Έτσι, εξοπλισμένοι πλέον με μία ολοκληρωμένη κατανόηση αυτών των μεθοδολογίων, είμαστε έτοιμοι να ξεκινήσουμε μία βαθύτερη εξερεύνηση των Μπεϋζιανών μεθόδων επιλογής μεταβλητών, κάνοντας διάφορες εφαρμογές σε προσομοιωμένα δεδομένα που περιέχουν ελλιπείς τιμές.

Στο Κεφάλαιο 5 ξεκινήσαμε μία ολοκληρωμένη εξέταση των Μπεϋζιανών μεθόδων επιλογής μεταβλητών, εφαρμόζοντας τις μεθόδους SSVS, Kuo & Mallick και GVS. Αυτό που έκανε αυτή την εφαρμογή ιδιαίτερα ενδιαφέρουσα ήταν η σκόπιμη εισαγωγή ελλιπών τιμών είτε στις τιμές της μεταβλητής απόκρισης, είτε σε τιμές των επεξηγηματικών μεταβλητών και η αντικατάσταση αυτών με χρήση Μπεϋζιανών μεθόδων στα προσομοιωμένα δεδομένα. Αυτές οι ελλιπείς τιμές δημιουργήθηκαν στα δεδομένα μας ακολουθώντας τους μηχανισμούς έλλειψης MCAR, MAR και NMAR. Αυτή η εφαρμογή μας επέτρεψε να εξετάσουμε εξονυχιστικά την απόδοση αυτών των μεθόδων υπό διαφορετικούς μηχανισμούς έλλειψης τιμών, παρέχοντάς μας πολύτιμες πληροφορίες σχετικά με την προσαρμοστικότητα και την αποτελεσματικότητα τους σε διάφορα σενάρια ελλιπών τιμών.

Μέσω αυτής της εκτεταμένης εφαρμογής των Μπεϋζιανών μεθόδων επιλογής μεταβλητών, δηλαδή των μεθόδων SSVS, Kuo & Mallick και GVS σε προσομοιωμένα δεδομένα που χαρακτηρίζονται από ελλιπείς τιμές, είτε στη μεταβλητή απόκρισης είτε στις ανεξάρτητες μεταβλητές με τη χρήση διαφορετικών μηχανισμών έλλειψης τιμών (MCAR, MAR και NMAR), προέκυψαν αρκετά αξιοσημείωτα συμπεράσματα. Σε γενικές γραμμές, τόσο η GVS όσο και ο Kuo & Mallick απέδωσαν καλύτερα από τη μέθοδο SSVS στα περισσότερα σενάρια. Η καλύτερη απόδοσή τους ήταν ιδιαίτερα εμφανής σε περιπτώσεις που τα δεδομένα χαρακτηρίζονταν από ελλιπείς τιμές. Ωστόσο, είναι σημαντικό να επισημανθεί ένα σημαντικό εύρημα, όταν ασχολούμαστε

με ελλιπείς τιμές στις επεξηγηματικές μεταβλητές οι οποίες έχουν δημιουργηθεί με τη χρήση του μηχανισμού NMAR. Οι μέθοδοι Kuo & Mallick και GVS δείχνουν την τάση να παραβλέπουν μία επεξηγηματική μεταβλητή, η οποία είναι οριακά σημαντική, κάτι το οποίο φαίνεται από τον Πίνακα 6.1.

Μέθοδοι	Ανίχνευση του πραγματικού μοντέλου						
	Πλήρες δεδομένα	Ελλιπείς τιμές στη μεταβλητή απόκρισης			Ελλιπείς τιμές στις επεξηγηματικές μεταβλητές		
		MCAR	MAR	NMAR	MCAR	MAR	NMAR
SSVS	✗	✗	✗	✗	✗	✗	✗
Kuo & Mallick	✓	✓	✓	✓	✓	✓	✗
GVS	✓	✓	✓	✓	✓	✓	✗

Πίνακας 6.1: Ικανότητα ανίχνευσης του πραγματικού μοντέλου με χρήση των μεθόδων SSVS, Kuo & Mallick και GVS.

Αυτή η παρατήρηση υπογραμμίζει τη σημασία της εξέτασης του είδους της έλλειψης και της πιθανής επίδρασής της στα αποτελέσματα των μεθόδων επιλογής μεταβλητών.

Αυτά τα ευρήματα τονίζουν την κρίσιμη σημασία που έχει ο προσεκτικός χειρισμός των ελλιπών τιμών, ειδικά όταν προέρχονται από το μηχανισμό NMAR. Για να διασφαλιστεί η ακεραιότητα των Μπεϋζιανών μεθόδων επιλογής μεταβλητών, πρέπει να ελαχιστοποιηθούν ή ιδανικά να αποτραπούν οι ελλιπείς τιμές στις επεξηγηματικές μεταβλητές οι οποίες δημιουργήθηκαν με NMAR μηχανισμούς. Για την επίτευξη αυτού του στόχου μπορούν να χρησιμοποιηθούν διάφορες στρατηγικές. Αρχικά, η χρήση καλά σχεδιασμένων ερωτηματολογίων και δομημένων συνεντεύξεων μπορεί να συμβάλει στη μείωση των ελλιπών τιμών. Επιπλέον, η εφαρμογή μέτρων ποιοτικού ελέγχου κατά τη συλλογή δεδομένων για τον εντοπισμό και την αντιμετώπιση πιθανών πηγών μεροληψίας ή ελλιπών τιμών. Αντιμετωπίζοντας προληπτικά το ζήτημα αυτό, οι ερευνητές μπορούν να ενισχύσουν την αξιοπιστία και την ακρίβεια των μεθόδων επιλογής μεταβλητών.

Παράρτημα

Φόρτωμα της βιβλιοθήκης ggplot

```
#Loading ggplot library
library(ggplot2)

gammas <- c("\u03b3_1","\u03b3_2","\u03b3_3","\u03b3_4","\u03b3_5",
"\u03b3_6","\u03b3_7","\u03b3_8","\u03b3_9","\u03b3_10",
"\u03b3_11","\u03b3_12")
gammas <- factor(gammas, levels = gammas)
# Labels used for all the barplots below
fillers <- c("Yes","No","Yes","No",
"Yes","No","No","No",
"Yes","No","Yes","No")
labelss <- c(expression("\u03b3"[1]),expression("\u03b3"[2]),
expression("\u03b3"[3]),expression("\u03b3"[4]),
expression("\u03b3"[5]),expression("\u03b3"[6]),
expression("\u03b3"[7]),expression("\u03b3"[8]),
expression("\u03b3"[9]),expression("\u03b3"[10]),
expression("\u03b3"[11]),expression("\u03b3"[12]))
```

Ραβδόγραμμα των δεικτών γ της μεθόδου SSVS στα πλήρη δεδομένα

```
#SSVS
Fulldatassvsgamma <- FUllmodelparametersssvs[14:25,]
Fulldatassvsgamma <- as.data.frame(Fulldatassvsgamma)
Barplotssvsfulldata <- ggplot( Fulldatassvsgamma ,aes(x = gammas ,
    y = Fulldatassvsgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)

#Printing the barplot of the index vector gamma
Barplotssvsfulldata + theme(axis.text.x = element_text(size = 15))
    + theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ του δειγματολήπτη Kuo & Mallick στα πλήρη δεδομένα

```
#Kuo & Mallick
FulldataKuoMallickgamma <- FUllmodelparametersKuoMallick[14:25,]
FulldataKuoMallickgamma <- as.data.frame(FulldataKuoMallickgamma)
BarplotKuoMallickfulldata <- ggplot(FulldataKuoMallickgamma ,aes(x
    = gammas , y = FulldataKuoMallickgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)

#Printing the barplot of the index vector gamma
BarplotKuoMallickfulldata + theme(axis.text.x = element_text(size =
    15)) + theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ της μεθόδου GVS στα πλήρη δεδομένα

```
#GVS
Fulldatagvsgamma <- FUllmodelparametersgvs[14:25,]
Fulldatagvsgamma <- as.data.frame(Fulldatagvsgamma)
Barplotgvsfulldata <- ggplot( Fulldatagvsgamma ,aes(x = gammas , y
    = Fulldatagvsgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)

#Printing the barplot of the index vector gamma
Barplotgvsfulldata + theme(axis.text.x = element_text(size = 15)) +
    theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ της μεθόδου SSVS υπό το μηχανισμό έλλειψης MCAR στη μεταβλητή απόκρισης

```
#SSVS while having missing values in the response variable
#using an MCAR mechanism.
MCARdatassvsgamma <- MCARmodelparametersssvs[14:25,]
MCARdatassvsgamma <- as.data.frame(MCARdatassvsgamma)
BarplotssvsMCARdata <- ggplot( MCARdatassvsgamma ,aes(x = gammas ,
    y = MCARdatassvsgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotssvsMCARdata + theme(axis.text.x = element_text(size = 15))
    + theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ του δειγματολήπτη Kuo & Mallick υπό το μηχανισμό έλλειψης MCAR στη μεταβλητή απόκρισης

```
#Kuo & Mallick while having missing values in the response variable
#using an MCAR mechanism.
MCARdataKuoMallickgamma <- MCARmodelparametersKuoMallick[14:25,]
MCARdataKuoMallickgamma <- as.data.frame(MCARdataKuoMallickgamma)
BarplotKuoMallickMCARdata <- ggplot( MCARdataKuoMallickgamma ,aes(x
    = gammas , y = MCARdataKuoMallickgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotKuoMallickMCARdata + theme(axis.text.x = element_text(size =
    15)) + theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ της μεθόδου GVS υπό το μηχανισμό έλλειψης MCAR στη μεταβλητή απόκρισης

```
#GVS while having missing values in the response variable
#using an MCAR mechanism.
MCARdatagvsgamma <- MCARmodelparametersgvs[14:25,]
MCARdatagvsgamma <- as.data.frame(MCARdatagvsgamma)
BarplotgvsMCARdata <- ggplot( MCARdatagvsgamma ,aes(x = gammas , y
    = MCARdatagvsgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotgvsMCARdata + theme(axis.text.x = element_text(size = 15)) +
    theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ της μεθόδου SSVS υπό το μηχανισμό έλλειψης MAR στη μεταβλητή απόκρισης

```
#SSVS while having missing values in the response variable
#using an MAR mechanism.
MARdatassvsgamma <- MARmodelparametersssvs[14:25,]
MARdatassvsgamma <- as.data.frame(MARdatassvsgamma)
BarplotssvsMARdata <- ggplot( MARdatassvsgamma ,aes(x = gammas , y
    = MARdatassvsgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotssvsMARdata + theme(axis.text.x = element_text(size = 15)) +
    theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ του δειγματολήπτη Kuo & Mallick υπό το μηχανισμό έλλειψης MAR στη μεταβλητή απόκρισης

```
#Kuo & Mallick while having missing values in the response variable
#using an MAR mechanism.
MARdataKuoMallickgamma <- MARmodelparametersKuoMallick[14:25,]
MARdataKuoMallickgamma <- as.data.frame(MARdataKuoMallickgamma)
BarplotKuoMallickMARdata <- ggplot( MARdataKuoMallickgamma ,aes(x =
    gammas , y = MARdataKuoMallickgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotKuoMallickMARdata + theme(axis.text.x = element_text(size =
    15)) + theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ της μεθόδου GVS υπό το μηχανισμό έλλειψης MAR στη μεταβλητή απόκρισης

```
#GVS while having missing values in the response variable
#using an MAR mechanism.
MARdatagvsgamma <- MARmodelparametersgvs[14:25,]
MARdatagvsgamma <- as.data.frame(MARdatagvsgamma)
BarplotgvsMARdata <- ggplot( MARdatagvsgamma ,aes(x = gammas , y =
    MARdatagvsgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotgvsMARdata + theme(axis.text.x = element_text(size = 15)) +
    theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ της μεθόδου SSVS υπό το μηχανισμό έλλειψης NMAR στη μεταβλητή απόκρισης

```
#SSVS while having missing values in the response variable
#using an NMAR mechanism.
NMARdatassvsgamma <- NMARmodelparametersssvs[14:25,]
NMARdatassvsgamma <- as.data.frame(NMARdatassvsgamma)
BarplotssvsNMARdata <- ggplot(NMARdatassvsgamma ,aes(x = gammas , y
    = NMARdatassvsgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotssvsNMARdata + theme(axis.text.x = element_text(size = 15))
    + theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ του δειγματολήπτη Kuo & Mallick υπό το μηχανισμό έλλειψης NMAR στη μεταβλητή απόκρισης

```
#Kuo & Mallick while having missing values in the response variable
#using an NMAR mechanism.
NMARdataKuoMallickgamma <- NMARmodelparametersKuoMallick[14:25,]
NMARdataKuoMallickgamma <- as.data.frame(NMARdataKuoMallickgamma)
BarplotKuoMallickNMARdata <- ggplot( NMARdataKuoMallickgamma ,aes(x
    = gammas , y = NMARdataKuoMallickgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotKuoMallickNMARdata + theme(axis.text.x = element_text(size =
    15)) + theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ της μεθόδου GVS υπό το μηχανισμό έλλειψης NMAR στη μεταβλητή απόκρισης

```
#GVS while having missing values in the response variable
#using an NMAR mechanism.
NMARdatagvsgamma <- NMARmodelparametersgvs[14:25,]
NMARdatagvsgamma <- as.data.frame(NMARdatagvsgamma)
BarplotgvsNMARdata <- ggplot( NMARdatagvsgamma ,aes(x = gammas , y
    = NMARdatagvsgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotgvsNMARdata + theme(axis.text.x = element_text(size = 15)) +
    theme(axis.text.y = element_text(size = 13))
```

Φόρτωμα της βιβλιοθήκης ggpubr και απεικόνιση των ραβδογραμμάτων για κάθε μέθοδο

```
#Loading ggpubr library
library(ggpubr)

#SSVS FOR Y
ggarrange(Barplotssvsfulldata + theme(axis.text.x =
    element_text(size = 15)) + theme(axis.text.y = element_text(size
    = 13)),
BarplotssvsMCARdata + theme(axis.text.x = element_text(size = 15))
    + theme(axis.text.y = element_text(size = 13)),
BarplotssvsMARdata + theme(axis.text.x = element_text(size = 15)) +
    theme(axis.text.y = element_text(size = 13)),
BarplotssvsNMARdata + theme(axis.text.x = element_text(size = 15))
    + theme(axis.text.y = element_text(size = 13)),
ncol = 2, nrow = 2,common.legend = TRUE)

#Kuo & Mallick FOR Y
ggarrange(BarplotKuoMallickfulldata + theme(axis.text.x =
    element_text(size = 15)) + theme(axis.text.y = element_text(size
    = 13)),
BarplotKuoMallickMCARdata + theme(axis.text.x = element_text(size =
```

```
    15)) + theme(axis.text.y = element_text(size = 13)),
BarplotKuoMallickMARdata + theme(axis.text.x = element_text(size =
    15)) + theme(axis.text.y = element_text(size = 13)),
BarplotKuoMallickNMARdata + theme(axis.text.x = element_text(size =
    15)) + theme(axis.text.y = element_text(size = 13)),
ncol = 2, nrow = 2,common.legend = TRUE)

#GVS FOR Y
ggarrange(Barplotgvsfulldata + theme(axis.text.x =
    element_text(size = 15)) + theme(axis.text.y = element_text(size
    = 13)),
BarplotgvsMCARdata + theme(axis.text.x = element_text(size = 15)) +
    theme(axis.text.y = element_text(size = 13)),
BarplotgvsMARdata + theme(axis.text.x = element_text(size = 15)) +
    theme(axis.text.y = element_text(size = 13)),
BarplotgvsNMARdata + theme(axis.text.x = element_text(size = 15)) +
    theme(axis.text.y = element_text(size = 13)),
ncol = 2, nrow = 2,common.legend = TRUE)
```

Ραβδόγραμμα των δεικτών γ της μεθόδου SSVS υπό το μηχανισμό έλλειψης MCAR στις επεξηγηματικές μεταβλητές

```
#SSVS while having missing values in the explanatory variables
#using an MCAR mechanism.
MCARdataXssvsgamma <- MCARmodelXparametersssvs[14:25,]
MCARdataXssvsgamma <- as.data.frame(MCARdataXssvsgamma)
BarplotXssvsMCARdata <- ggplot( MCARdataXssvsgamma ,aes(x = gammas
    , y = MCARdataXssvsgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotXssvsMCARdata + theme(axis.text.x = element_text(size = 15))
    + theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ του δειγματολήπτη Kuo & Mallick υπό το μηχανισμό έλλειψης MCAR στις επεξηγηματικές μεταβλητές

```
#Kuo & Mallick while having missing values in the explanatory
#variables using an MCAR mechanism.
MCARdataXKuoMallickgamma <- MCARmodelXparametersKuoMallick[14:25,]
MCARdataXKuoMallickgamma <- as.data.frame(MCARdataXKuoMallickgamma)
BarplotXKuoMallickMCARdata <- ggplot( MCARdataXKuoMallickgamma
    ,aes(x = gammas , y = MCARdataXKuoMallickgamma$mean,
    fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotXKuoMallickMCARdata + theme(axis.text.x = element_text(size
    = 15)) + theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ της μεθόδου GVS υπό το μηχανισμό έλλειψης MCAR στις επεξηγηματικές μεταβλητές

```
#GVS while having missing values in the explanatory variables
#using an MCAR mechanism.
MCARdataXgvsgamma <- MCARmodelXparametersgvs[14:25,]
MCARdataXgvsgamma <- as.data.frame(MCARdataXgvsgamma)
BarplotXgvsMCARdata <- ggplot( MCARdataXgvsgamma ,aes(x = gammas ,
    y = MCARdataXgvsgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
```

```
BarplotXgvsMCARdata + theme(axis.text.x = element_text(size = 15))
    + theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ της μεθόδου SSVS υπό το μηχανισμό έλλειψης MAR στις επεξηγηματικές μεταβλητές

```
#SSVS while having missing values in the explanatory variables
#using an MAR mechanism.
MARdataXssvsgamma <- MARmodelXparametersssvs[14:25,]
MARdataXssvsgamma <- as.data.frame(MARdataXssvsgamma)
BarplotXssvsMARdata <- ggplot( MARdataXssvsgamma ,aes(x = gammas ,
    y = MARdataXssvsgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotXssvsMARdata + theme(axis.text.x = element_text(size = 15))
    + theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ του δειγματολήπτη Kuo & Mallick υπό το μηχανισμό έλλειψης MAR στις επεξηγηματικές μεταβλητές

```
#Kuo & Mallick while having missing values in the explanatory
#variables using an MAR mechanism.
MARdataXKuoMallickgamma <- MARmodelXparametersKuoMallick[14:25,]
MARdataXKuoMallickgamma <- as.data.frame(MARdataXKuoMallickgamma)
BarplotXKuoMallickMARdata <- ggplot(MARdataXKuoMallickgamma ,aes(x
    = gammas , y = MARdataXKuoMallickgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
```

```
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotXKuoMallickMARdata + theme(axis.text.x = element_text(size =
    15)) + theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ της μεθόδου GVS υπό το μηχανισμό έλλειψης MAR στις επεξηγηματικές μεταβλητές

```
#GVS while having missing values in the explanatory variables
#using an MAR mechanism.
MARdataXgvsgamma <- MARmodelXparametersgvs[14:25,]
MARdataXgvsgamma <- as.data.frame(MARdataXgvsgamma)
BarplotXgvsMARdata <- ggplot( MARdataXgvsgamma ,aes(x = gammas , y
    = MARdataXgvsgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotXgvsMARdata + theme(axis.text.x = element_text(size = 15)) +
    theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ της μεθόδου SSVS υπό το μηχανισμό έλλειψης NMAR στις επεξηγηματικές μεταβλητές

```
#SSVS while having missing values in the explanatory variables
#using an NMAR mechanism.
NMARdataXssvsgamma <- NMARmodelXparametersssvs[14:25,]
NMARdataXssvsgamma <- as.data.frame(NMARdataXssvsgamma)
BarplotXssvsNMARdata <- ggplot( NMARdataXssvsgamma ,aes(x = gammas
    , y = NMARdataXssvsgamma$mean, fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
```

```
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotXssvsNMARdata + theme(axis.text.x = element_text(size = 15))
    + theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ του δειγματολήπτη Kuo & Mallick υπό το μηχανισμό έλλειψης NMAR στις επεξηγηματικές μεταβλητές

```
#Kuo & Mallick while having missing values in the explanatory
#variables using an NMAR mechanism.
NMARdataXKuoMallickgamma <- NMARmodelXparametersKuoMallick[14:25,]
NMARdataXKuoMallickgamma <- as.data.frame(NMARdataXKuoMallickgamma)
BarplotXKuoMallickNMARdata <- ggplot(NMARdataXKuoMallickgamma
    ,aes(x = gammas , y = NMARdataXKuoMallickgamma$mean,
    fill=fillers)) +
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotXKuoMallickNMARdata + theme(axis.text.x = element_text(size
    = 15)) + theme(axis.text.y = element_text(size = 13))
```

Ραβδόγραμμα των δεικτών γ της μεθόδου GVS υπό το μηχανισμό έλλειψης NMAR στις επεξηγηματικές μεταβλητές

```
#GVS while having missing values in the explanatory variables
#using an NMAR mechanism.
NMARdataXgvsgamma <- NMARmodelXparametersgvs[14:25,]
NMARdataXgvsgamma <- as.data.frame(NMARdataXgvsgamma)
BarplotXgvsNMARdata <- ggplot( NMARdataXgvsgamma ,aes(x = gammas ,
    y = NMARdataXgvsgamma$mean, fill=fillers)) +
```

```
geom_bar(stat = "identity")+
scale_fill_manual(values = c("#B00000","darkblue")) +
geom_hline(yintercept=0.50, linetype='dashed', col =
    'black',lwd=1.3)+
labs(title =
    substitute(paste(bold('Title'))),x=substitute(paste(bold('Title
    for x axis'))),fill = substitute(paste(bold('Title for
    legend'))),
y=substitute(paste(bold('Title for y axis'))))+
geom_col() +
scale_x_discrete(labels=labelss)
#Printing the barplot of the index vector gamma
BarplotXgvsNMARdata + theme(axis.text.x = element_text(size = 15))
    + theme(axis.text.y = element_text(size = 13))
```

Φόρτωμα της βιβλιοθήκης ggpubr και απεικόνιση των ραβδογραμμάτων για κάθε μέθοδο

```
#Loading ggpubr library
library(ggpubr)

#SSVS FOR X
ggarrange(Barplotssvsfulldata + theme(axis.text.x =
    element_text(size = 15)) + theme(axis.text.y = element_text(size
    = 13)),
BarplotXssvsMCARdata + theme(axis.text.x = element_text(size = 15))
    + theme(axis.text.y = element_text(size = 13)),
BarplotXssvsMARdata + theme(axis.text.x = element_text(size = 15))
    + theme(axis.text.y = element_text(size = 13)),
BarplotXssvsNMARdata + theme(axis.text.x = element_text(size = 15))
    + theme(axis.text.y = element_text(size = 13)),
ncol = 2, nrow = 2,common.legend = TRUE)

#Kuo & Mallick FOR X
ggarrange(BarplotKuoMallickfulldata + theme(axis.text.x =
    element_text(size = 15)) + theme(axis.text.y = element_text(size
    = 13)),
BarplotXKuoMallickMCARdata + theme(axis.text.x = element_text(size
    = 15)) + theme(axis.text.y = element_text(size = 13)),
BarplotXKuoMallickMARdata + theme(axis.text.x = element_text(size =
    15)) + theme(axis.text.y = element_text(size = 13)),
BarplotXKuoMallickNMARdata + theme(axis.text.x = element_text(size
    = 15)) + theme(axis.text.y = element_text(size = 13)),
ncol = 2, nrow = 2,common.legend = TRUE)

#GVS FOR X
ggarrange(Barplotgvsfulldata + theme(axis.text.x =
```

```
    element_text(size = 15)) + theme(axis.text.y = element_text(size
    = 13)),
BarplotXgvsMCARdata + theme(axis.text.x = element_text(size = 15))
    + theme(axis.text.y = element_text(size = 13)),
BarplotXgvsMARdata + theme(axis.text.x = element_text(size = 15)) +
    theme(axis.text.y = element_text(size = 13)),
BarplotXgvsNMARdata + theme(axis.text.x = element_text(size = 15))
    + theme(axis.text.y = element_text(size = 13)),
ncol = 2, nrow = 2,common.legend = TRUE)
```

Βιβλιογραφία

S. Banerjee. Bayesian linear model: Gory details. 2008.

M.S. Bartlett. Comment on D.V. Lindleys statistical paradox. *Biometrika,*, 44: 533–534, 1957.

J.O. Berger, J.M. Bernardo, and D. Sun. The formal definition of reference priors. *The Annals of Statistics,*, 37:905–938, 2009.

J.M. Bernardo. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(2):113–128, 1979.

J.M. Bernardo and A.F.M. Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

D.J. Biau, B.M. Jolles, and R. Porcher. P value and the theory of hypothesis testing: an explanation for new researchers. *Clinical Orthopaedics and Related Research®*, 468:885–892, 2010.

B.P. Carlin and T.A. Louis. Bayes and empirical Bayes methods for data analysis, 1997.

P. Dellaportas, J.J Forster, and I. Ntzoufras. On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12(1):27–36, 2002.

N. Erler. Bayesian imputation of missing covariates. 2019.

C. Fernandez, E. Ley, and M.F.J. Steel. Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100(2):381–427, 2001.

D.P. Foster and E.I. George. The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975, 1994.

A.E. Gelfand and D.K. Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514, 1994.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 1995.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

E.I. George and R.E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

E.I. George and R.E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, pages 339–373, 1997.

W.K. Hastings. Monte carlo sampling methods using Markov chains and their applications. 1970.

J.G. Ibrahim, M.H. Chen, and S.R. Lipsitz. Bayesian methods for generalized linear models with covariates missing at random. *Canadian Journal of Statistics*, 30(1): 55–78, 2002.

H. Jeffreys. Some tests of significance, treated by the theory of probability. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 31, pages 203–222. Cambridge University Press, 1935.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

R.E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934, 1995.

S. Konishi and G. Kitagawa. Information criteria and statistical modeling. 2008.

L. Kuo and B. Mallick. Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81, 1998.

P.C. Lambert, A.J. Sutton, P.R. Burton, K.R. Abrams, and D.R. Jones. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24(15):2401–2428, 2005.

S.M. Lewis and A.E. Raftery. Estimating Bayes factors via posterior simulation with the Laplace—Metropolis estimator. *Journal of the American Statistical Association*, 92(438):648–655, 1997.

F. Liang, R. Paulo, G. Molina, M.A. Clyde, and J.O. Berger. Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.

D.V. Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.

R.J.A. Little and D.B. Rubin. Multiple imputation for nonresponse in surveys. *John Wiley & Sons, Inc.. doi*, 10:9780470316696, 1987.

D.J.C. MacKay. Choice of basis for Laplace approximation. *Machine Learning*, 33: 77–86, 1998.

N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

I. Ntzoufras. Aspects of Bayesian model and variable selection using MCMC. *Unpublished Ph. D. Thesis, Department of Statistics, Athens University of Economics and Business, Athens, Greece*, 1999.

I. Ntzoufras. Gibbs variable selection using BUGS. *Journal of Statistical Software*, 7:1–19, 2002.

I. Ntzoufras. Bayesian variable selection – an introductory tutorial. ISA, 2011.

K. Perrakis and I. Ntzoufras. Stochastic search variable selection (SSVS). *Wiley StatsRef: Statistics Reference Online*, pages 1–6, 2015.

C.P. Robert and D. Wraith. Computational methods for Bayesian model choice. In *Aip Conference Proceedings*, volume 1193, pages 251–262. American Institute of Physics, 2009.

D.B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464, 1978.

M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

S. Van Buuren. *Flexible imputation of missing data*. CRC Press, 2018.

S. Van Buuren and K. Oudshoorn. *Flexible multivariate imputation by MICE*. Leiden: TNO, 1999.

Y.B. Wang, M.H. Chen, L. Kuo, and P.O. Lewis. A new Monte Carlo method for estimating marginal likelihoods. *Bayesian Analysis*, 13(2):311, 2018.

A. Zellner. On assessing prior distributions and Bayesian regression analysis using g-prior distributions. In P. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. North-Holland, Amsterdam, 1986.

A. Zellner. *An introduction to Bayesian inference in econometrics*. John Wiley & Sons, 1996.

A. Zellner and A. Siow. Posterior odds ratios for selected regression hypothesis (with discussion). In J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds., *Bayesian Statistics*, Vol. 1, pages 585–606 & 618–647 (discussion). Oxford University Press, 1980.