**Εθνικό Μετσόβιο Πολυτεχνείο**
Σχολή Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών
Τομέας Μαθηματικών

# ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Κωνσταντίνα Τιμολέων

## Μπεϋζιανά Κανονικά Γραμμικά Μοντέλα Παλινδρόμησης

**Τριμελής Επιτροπή:**
Δημήτρης Φουσκάκης, Καθηγητής (Επιβλέπων)
Σχολή Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών

Φίλια Βόντα, Καθηγήτρια
Σχολή Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών

Μιχάλης Λουλάκης, Αναπλ. Καθηγητής
Σχολή Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών

Αθήνα, Σεπτέμβριος 2021

**Ευχαριστίες**
Θα ήθελα να ευχαριστήσω τον επιβλέποντα και καθηγητή μου κ. Δημήτρη Φουσκάκη για την παραγωγική συνεργασία και την καθοριστική συμβολή του στην ολοκλήρωση της παρούσας διπλωματικής. Ακόμα, τα μέλη της εξεταστικής επιτροπής, κ. Φίλια Βόντα και κ. Μιχάλη Λουλάκη. Καθώς η εργασία αυτή σηματοδοτεί το τέλος των προπτυχιακών σπουδών μου στη Σχολή Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών, θα ήθελα επίσης να ευχαριστήσω την οικογένειά μου για τη διαρκή στήριξή τους και τέλος τους φίλους και συναδέλφους που πορευτήκαμε και εξελιχθήκαμε μαζί.

Κωνσταντίνα Τιμολέων

**Περίληψη**

Η Μπεϋζιανή παλινδρόμηση αποσκοπεί στην αποκρυπτογράφηση της γραμμικής σχέσης των παρατηρούμενων δεδομένων, μέσω του προσδιορισμού των λεγόμενων εκ των υστέρων κατανομών, ώστε να προβλέψει την τιμή της μεταβλητής απόκρισης νέων δεδομένων. Σε αντίθεση με την κλασική παλινδρόμηση των ελαχίστων τετραγώνων, δεν εστιάζει στην κατασκευή του μοντέλου που ταιριάζει καλύτερα στα δεδομένα· αντ' αυτού, επιχειρεί την ποσοτικοποίηση της αβεβαιότητας γύρω από το εκτιμώμενο μοντέλο. Ωστόσο, παρά την αποδεδειγμένη αποτελεσματικότητά της, σε ορισμένες περιπτώσεις εμφανίζει ζητήματα υπολογισιμότητας. Στην παρούσα εργασία στόχος μας είναι να αναδείξουμε τις αναλυτικές και υπολογιστικές μεθόδους που υπερπηδούν αυτά τα προβλήματα αλλά και να τις εφαρμόσουμε εμπράκτως. Οι στόχοι αυτοί εκπληρώνονται μέσα από την παρουσίαση των MCMC μεθόδων που συνοδεύουν τα Μπεϋζιανά μοντέλα, της συζυγούς ανάλυσης για την εξαγωγή της ύστερης πληροφορίας από την πρότερη, καθώς και των τρόπων με τους οποίους οι τεχνικές αυτές επιλύουν θέματα εφαρμογής της Μπεϋζιανής παλινδρόμησης. Το αποτέλεσμα, είναι ένα πλήρες εγχειρίδιο για την κατασκευή Μπεϋζιανών μοντέλων παλινδρόμησης το οποίο περιλαμβάνει όλους τους απαραίτητους τύπους για τη συμπερασματολογία και την πρόβλεψη στην ανάλυση παλινδρόμησης, εμπλουτίζοντας έτσι την περιορισμένη επί του θέματος ελληνική βιβλιογραφία. Η μελέτη ολοκληρώνεται με την παράθεση πρακτικών αποτελεσμάτων εφαρμογής των παραπάνω σε προσομοιωμένα δεδομένα αλλά και σε πραγματικά δεδομένα μοντέλου πρόβλεψης της ποιότητας του κρασιού vinho verde, φανερώνοντας την επιδεξιότητα της Μπεϋζιανής προσέγγισης στην γραμμική παλινδρόμηση.

**Λέξεις-κλειδιά**
Μπεϋζιανά μοντέλα παλινδρόμησης, Μπεϋζιανή στατιστική, Markov Chain Monte Carlo, ανάλυση παλινδρόμησης.

**Abstract**

Bayesian regression aims to decipher the linear relation in the observed data, by specifying the so-called posterior distributions, in order to predict the value of the dependent variable for new data. Unlike the ordinary least squares regression, it does not focus on constructing a best-fit model for the data; instead, Bayesian regression attempts to quantify the uncertainty surrounding the estimated model. However, inspite of its confirmed effectiveness, in some cases, it suffers from computability issues. In our present work, our goal is to introduce and apply the analytical and computational methods that surpass these issues. We fullfill these goals by presenting the MCMC algorithms that accompany Bayesian models, the conjugate analysis techniques for extracting posterior information from prior knowledge, as well as the manner in which these methods curate applicability concerns. The resulting outcome is a thorough manual for constructing Bayesian regression models which includes all the formulas for inference and prediction required in regression analysis, enriching the limited, on the matter, greek literature. Our study is concluded with the employment of the aforementioned in simulated and real data for the prediction of vinho verde wine quality, showcasing the dexterity of the Bayesian approach in linear regression.

**Keywords**
Bayesian regression models, Bayesian statistics, Markov Chain Monte Carlo, regression analysis.

# Περιεχόμενα

# Πίνακας ακρωνυμίων

| | |
|---|---|
| **μ.α.** | Μαρκοβιανή αλυσίδα |
| **σ.π.π.** | συνάρτηση πυκνότητας πιθανότητας |
| **τ.μ.** | τυχαία μεταβλητή |
| **IG** | αντίστροφη γάμμα (Inverse Gamma) κατανομή |
| **LOOCV** | Leave-One-Out Cross Validation |
| **M-H** | Metropolis-Hastings |
| **MCMC** | Markov Chain Monte Carlo |
| **MVSt** | πολυδιάστατη Student (MultiVariate Student) κατανομή |
| **NIG** | κανονική-αντίστροφη γάμμα (Normal-Inverse Gamma) κατανομή |
| **RMSE** | Root Mean Square Error |

# Εισαγωγή

Τα μοντέλα παλινδρόμησης συνιστούν ένα χρήσιμο στατιστικό εργαλείο, εφοδιασμένο με την ισχύ που του παρέχει η αυστηρά μαθηματική του θεμελίωση, με ποικίλες εφαρμογές στον κόσμο της τεχνολογίας, της οικονομίας και όχι μόνο. Η παρούσα διπλωματική εργασία μελετά τα κανονικά πολλαπλά μοντέλα παλινδρόμησης από τη σκοπιά της Μπεϋζιανής στατιστικής, εξερευνώντας την έννοια της πρότερης πληροφορίας -έννοια μείζονος σημασίας για τη σχολή αυτή της στατιστικής- καθώς και τρόπους ενσωμάτωσης αυτής στο μοντέλο. Η μελέτη περιλαμβάνει επίσης την ανάπτυξη και ανάλυση σχετικής μεθοδολογίας για την προσαρμογή Μπεϋζιανών μοντέλων παλινδρόμησης για τα διάφορα είδη πρότερης πληροφορίας καθώς και την έμπρακτη εφαρμογή των παραπάνω σε προσομοιωμένα και πραγματικά δεδομένα.

Η Μπεϋζιανή θεωρία, αν και άργησε χρονικά να γίνει αποδεκτή από την επιστημονική κοινότητα, τα τελευταία χρόνια έχει κερδίσει έδαφος λόγω της ανάμειξής της στις μεθόδους μηχανικής μάθησης και γενικότερα στο πλέον δημοφιλές πεδίο της τεχνητής νοημοσύνης. Επομένως, η μελέτη της κλασικής παλινδρόμησης μέσω αυτής της καίριας προσέγγισης παρουσιάζει ιδιαίτερο ενδιαφέρον. Παράλληλα, σχηματίζει ισχυρότατο υπολογιστικό δίδυμο με τις Markov Chain Monte Carlo μεθόδους με αποτέλεσμα το ξετύλιγμα της περίτεχνης λειτουργίας του ζεύγους να θεωρείται άξιο μελέτης.

Όσον αφορά τη διεθνή βιβλιογραφία, μια σύντομη έρευνα αποκαλύπτει πως η Μπεϋζιανή παλινδρόμηση έχει ισχυρή παρουσία, κυρίως σε πιο προχωρημένα κείμενα και βιβλία. Ωστόσο, ακόμα και τότε, είναι περιορισμένες οι περιπτώσεις όπου δίνεται αναλυτικά, με αποδείξεις, ο τρόπος εξαγωγής των μαθηματικών τύπων και λειτουργίας των μοντέλων.

Η κεντρική επιδίωξη της παρούσας εργασίας είναι μια ολοκληρωμένη και ενδελεχής παρουσίαση της λειτουργίας της Μπεϋζιανής παλινδρόμησης αλλά και των απαιτούμενων γνώσεων και τεχνικών που επιστρατεύει. Το ερευνητικό κομμάτι, που πραγματώνεται μέσα από την προσαρμογή διαφορετικών Μπεϋζιανών μοντέλων σε επιλεγμένα σύνολα δεδομένων, εστιάζει στη συγκριτική μελέτη της αποτελεσματικότητας της παλινδρόμησης για τους διάφορους τρόπους εισαγωγής πρότερης πληροφορίας. Για τον σκοπό αυτό υιοθετούμε μεθόδους συζυγούς ανάλυσης οι οποίες διευκολύνουν και απλουστεύουν τον σχηματισμό των μοντέλων.

Συμπληρωματικά, η έρευνά μας αποσκοπεί στη δημιουργία ενός χρηστικού και λεπτομερούς εγχειριδίου που αποσαφηνίζει και τεκμηριώνει με μαθηματικό τρόπο τη διαδικασία προσαρμογής Μπεϋζιανών μοντέλων παλινδρόμησης. Ελλείψει αντίστοιχου υλικού στην ελληνική γλώσσα, η απόπειρα αυτή θα είναι μια χρήσιμη προσθήκη στην ελληνική βιβλιογραφία.

Συνοπτικά, η εργασία δομείται ως εξής: αρχικά γίνεται αναφορά στα βασικότερα στοιχεία της Μπεϋζιανής στατιστικής, εστιάζοντας στον κανόνα του Bayes ο οποίος τη θεμελίωσε, και εισάγοντας τις απαραίτητες έννοιες της πρότερης και ύστερης πληροφορίας. Στη συνέχεια ασχολούμαστε εν συντομία με τις Markov Chain Monte Carlo μεθόδους και πώς αυτές υπεισέρχονται, με καταλυτικό τρόπο, στη Μπεϋζιανή συμπερασματολογία. Ακολούθως, αναλύεται η δομή των πολλαπλών κανονικών Μπεϋζιανών μοντέλων παλινδρόμησης και παρατίθενται τρεις διαφορετικές επιλογές για την εισαγωγή πρότερης πληροφορίας. Για κάθε επιλογή αναπτύσσονται αναλυτικοί τύποι για τον υπολογισμό διαφόρων ποσοτήτων ενδιαφέροντος στα πλαίσια της παλινδρόμησης. Τέλος, η θεωρία που αναπτύχθηκε τίθεται σε

εφαρμογή, πρώτα πάνω σε δεδομένα που έχουν προσομοιωθεί με τη βοήθεια υπολογιστή και κατόπιν σε πραγματικά δεδομένα που αφορούν την εκτίμηση ποιότητας κρασιού.

# 1 Εισαγωγή στη Μπεϋζιανή Στατιστική

Η Μπεϋζιανή Στατιστική (Bayesian statistics) παρέχει ένα άρτια θεμελιωμένο πλαίσιο εργασίας και μεθοδολογίας για την επαγωγική στατιστική συμπερασματολογία. Το χαρακτηριστικό της Μπεϋζιανής προσέγγισης (Bayesian approach) είναι η χρήση πιθανοτήτων και κατανομών για τη μοντελοποίηση της αβεβαιότητας και την ενσωμάτωση νέας πληροφορίας. Ο κλάδος αυτός της στατιστικής, που έχει διαδραματίσει καθοριστικό ρόλο στην εξέλιξη της στατιστικής συμπερασματολογίας (Kokolakis 2010), οφείλει το όνομά του στον Thomas Bayes.

## 1.1 Ιστορικά στοιχεία

Ο Thomas Bayes γεννήθηκε το 1701 (Bellhouse 2004) ή το 1702 (O'Connor & Robertson 2004) στο Λονδίνο και πέθανε στις 17 Απριλίου του 1761 στο Κεντ της Αγγλίας. Σπούδασε Λογική και Θεολογία στο Πανεπιστήμιο του Εδινβούργου και περίπου στα τριάντα του χρόνια χειροτονήθηκε ως Αιρετικός ιερέας (O'Connor & Robertson 2004), συνεχίζοντας την κληρονομιά του πατέρα του, Joshua Bayes.

Στη διάρκεια της ζωής του ασχολήθηκε εκτενώς με την επιστήμη, ή "φυσική φιλοσοφία" όπως ονομάζονταν τον 18ο αιώνα στο σύνολό τους οι φυσικές επιστήμες (Bellhouse 2004), αλλά και συγκεκριμένα με θέματα των μαθηματικών. Από τα ελάχιστα δείγματα της δουλειάς του που διασώθηκαν, φαίνεται ότι ο Bayes ασχολήθηκε αρχικά με τις άπειρες σειρές προτού αναπτύξει την πιθανοθεωρία του (Bellhouse 2004) η οποία έμελλε να επηρεάσει σημαντικά την επιστήμη της στατιστικής.

Όπως έχει συμβεί με πολλούς σπουδαίους επιστήμονες, η αναγνώριση του έργου του ήρθε μετά τον θάνατό του, όταν ο φίλος του, Richard Price, ύστερα από έκκληση συγγενών του θανόντος, εξέτασε τα γραπτά που ο Bayes είχε αφήσει πίσω του (Bellhouse 2004). Ετσι, ο Price προώθησε στην Βασιλική Εταιρεία (Royal Society) το άρθρο του Bayes με τίτλο "Essay towards solving a problem in the doctrine of chances" (Bayes & Price 1763) το οποίο και εκδόθηκε το 1764 στο περιοδικό Philosophical Transactions of the Royal Society of London (O'Connor & Robertson 2004).

## 1.2 Ο Κανόνας του Bayes και η Μπεϋζιανή Θεωρία

Παρόλο που η συνεισφορά του Bayes στην επιστήμη είναι ιδιαίτερα σημαντική στο σύνολό της, η υστεροφημία του οφείλεται κυρίως στον γνωστό "κανόνα" ή "θεώρημα του Bayes", που αποτελεί το απαύγασμα του "Essay towards solving a problem in the doctrine of chances". Η απλούστερη διατύπωση του κανόνα έχει ως εξής: αν $H$ μία υπόθεση και $D$ τα διαθέσιμα δεδομένα υπέρ ή κατά της υπόθεσης, ισχύει το εξής:

$$P(H|D) = \frac{P(D \mid H)P(H)}{P(D)}, \tag{1}$$

όπου $P(H)$ η πιθανότητα που αποδίδουμε στην υπόθεση $H$ πριν δούμε τα δεδομένα ενώ $P(H|D)$ η πιθανότητα που αποδίδουμε στην υπόθεση $H$ αφού λάβουμε τα δεδομένα $D$. Αντίστοιχα, $P(D)$ η πιθανότητα να παρατηρήσουμε τα συγκεκριμένα δεδομένα ενώ $P(D|H)$

η πιθανότητα να παρατηρήσουμε τα συγκεκριμένα δεδομένα δεδομένου ότι η υπόθεσή μας αληθεύει.

Λαμβάνοντας στοιχεία για τα $P(D), P(D|H)$, η αρχική μας πεποίθηση $P(H)$ ανανεώνεται στην $P(H|D)$ η οποία πλέον περιλαμβάνει την επικαιροποιημένη, μέσω των δεδομένων, γνώση για την εμφάνιση του $H$. Για τον λόγο αυτό, η $P(H)$ αναφέρεται ως πρότερη (prior) και η $P(H|D)$ ως ύστερη (posterior) πιθανότητα, όπου οι έννοιες prior και posterior ορίζονται δεδομένης μιας αρχικής κατάστασης πληροφορίας και σε σχέση με όποιες επιπλέον πληροφορίες (Bernando & Smith 2000). Γίνεται λοιπόν αντιληπτό ότι το θεώρημα δεν είναι απλώς ένας τρόπος υπολογισμού δεσμευμένων πιθανοτήτων, όπως συνήθως εκλαμβάνεται, αλλά ένας μηχανισμός που μας επιτρέπει να μαθαίνουμε από τα δεδομένα (Bernando & Smith 2000).

Με σκοπό να οριστεί αναλυτικά το πλαίσιο εφαρμογής της Μπεϋζιανής θεωρίας, υπογραμμίζεται πως για την υλοποίησή της υιοθετείται η υποκειμενική ερμηνεία της πιθανότητας (subjective interpretation) (Κοκολάκης & Φουσκάκης 2009). Αυτή η εναλλακτική προσέγγιση στην έννοια της πιθανότητας συνιστά και μία από τις διαφορές μεταξύ της Μπεϋζιανής και της κλασικής στατιστικής, αφού η δεύτερη στηρίζεται στην αντικειμενική και καθολική ερμηνεία των πιθανοτήτων. Υπό το πρίσμα της Μπεϋζιανής αντίληψης, η πιθανότητα αποτελεί προσωπική πεποίθηση, εκφράζει δηλαδή τον βαθμό αβεβαιότητας ή σιγουριάς ενός κατάλληλου συντελεστή (agent) για το προς μελέτη ενδεχόμενο (Hájek 2019). Η αβεβαιότητα, όπως αυτή μοντελοποιείται μέσω των προσωπικών πεποιθήσεων, είναι συμβατή με τον αυστηρό φορμαλισμό της πιθανότητας, επιτρέποντας τη χρήση των συνήθων πιθανοθεωρητικών μεθόδων. Παραπέμπουμε στους Bernando & Smith 2000 για μία αναλυτική τεκμηρίωση σχετικά με το επιχείρημα αυτό.

Μία ακόμη ειδοποιός διαφορά μεταξύ των δύο “σχολών" της στατιστικής, εντοπίζεται στην εκτιμητική. Ενώ στην μεθοδολογία της κλασικής στατιστικής η προς εκτίμηση παράμετρος, ή διάνυσμα παραμέτρων, έστω $\boldsymbol{\theta}$, θεωρείται άγνωστη αλλά σταθερή, στη Μπεϋζιανή προσέγγιση το $\boldsymbol{\theta}$ θεωρείται τυχαία μεταβλητή. Στη δεύτερη περίπτωση, σκοπός μας είναι η προσέγγιση της κατανομής του $\boldsymbol{\theta}$, μοντελοποιώντας με τη χρήση πιθανοτήτων την αβεβαιότητα γύρω από αυτή.

Επιστρέφουμε στο σημαντικότατο αποτέλεσμα που παρουσιάστηκε στη σχέση (1), ώστε να αναπτύξουμε με μεγαλύτερη σαφήνεια τη λειτουργία του. Ο κανόνας του Bayes μπορεί να διατυπωθεί για οποιαδήποτε ενδεχόμενα $A, B$, δίνοντας την πιθανότητα εμφάνισης του ενδεχομένου $A$ δοθείσης της εμφάνισης του ενδεχομένου $B$:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{2}$$

Εφαρμόζοντας τον τύπο ολικής πιθανότητας, για διαμέριση $A_1, A_2, ...A_n$ του δειγματικού χώρου, η σχέση (2) γίνεται:

$$P(A|B) = \frac{P(B|A)P(A)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}. \tag{3}$$

Η χρησιμότητα αυτού του αποτελέσματος θα δειχθεί με ένα απλό παράδειγμα: συγκεκριμένα, έστω ότι ενδιαφερόμαστε να μελετήσουμε την εμφάνιση ενός γονιδίου στο γονιδίωμα, δηλαδή κατά πόσο ένα χαρακτηριστικό φέρεται στο γενετικό υλικό. Έχουμε στη

διάθεσή μας κατάλληλο ιατρικό τεστ που ανιχνεύει αν ένα άτομο φέρει ή όχι το γονίδιο, με ένδειξη θετικό/$+$ (positive) ή αρνητικό/$-$ (negative) αντίστοιχα. Γνωρίζουμε ότι η ικανότητα του τεστ να εντοπίσει το γονίδιο σε άτομα που πράγματι το έχουν (true positive), ή αλλιώς *ευαισθησία* (*sensitivity*) του τεστ, είναι $90\%$. Ξέρουμε επίσης ότι το τεστ θα έχει όντως αρνητικό αποτέλεσμα στο $93\%$ των ατόμων που δεν φέρουν το γονίδιο (true negative), ποσοστό γνωστό και ως *ειδικότητα* (*specificity*) του τεστ.

Αν συμβολίσουμε με $G$ την εμφάνιση του γονιδίου και $G^C$ την απουσία αυτού, έχουμε, από προηγούμενες μελέτες ότι:

$$P(G) = 28\%$$

και άρα:

$$P(G^C) = 72\%.$$

Με βάση αυτόν τον συμβολισμό, το sensitivity και το specificity του τεστ δίνονται με τη μορφή πιθανοτήτων ως:

$$P(+|G) = 90\%$$

και:

$$P(-|G^C) = 93\%$$

αντίστοιχα.

Έχοντας υποβάλει ένα υποκείμενο στο εν λόγω τεστ και έχοντας λάβει θετικό αποτέλεσμα από αυτό, καλούμαστε να απαντήσουμε στο εξής ερώτημα: ποια είναι η πιθανότητα το υποκείμενο να είναι πράγματι φορέας του γονιδίου; Μία πρώτη, διαισθητική, απάντηση θα ήταν $90\% \cdot$ φυσικά αυτή είναι μία λανθασμένη απάντηση καθώς περιορίζει την πρόβλεψη στα στοιχεία που έχουμε για τη διαγνωστική ικανότητα του τεστ, χωρίς να συμπεριλαμβάνει την πιθανότητα εμφάνισης του γονιδίου στο άτομο, η οποία είναι ανεξάρτητη του αν έχει υποβληθεί σε διαγνωστικό έλεγχο ή όχι. Με ανάλογο τρόπο, εσφαλμένη θα ήταν και η απάντηση $28\%$ αφού βασίζεται αποκλειστικά στην πιθανότητα εμφάνισης του γονιδίου στον πληθυσμό αγνοώντας το γεγονός ότι έχουμε διενεργήσει κάποιον έλεγχο με θετικό αποτέλεσμα.

Για να απαντηθεί ικανοποιητικά το ερώτημα, αναζητούμε έναν τρόπο να συνδυαστούν οι παραπάνω πληροφορίες. Επιθυμούμε η αρχική γνώση να τροποποιηθεί κατάλληλα σε μία ανανεωμένη πεποίθηση σχετικά με την εμφάνιση του γονιδίου, ενόψει του αποτελέσματος του τεστ. Αυτή ακριβώς τη λειτουργία αναλαμβάνει να εκτελέσει ο κανόνας του Bayes: εφαρμόζοντας το θεώρημα μπορούμε να υπολογίσουμε την πιθανότητα $P(G|+)$ κάποιο άτομο να είναι φορέας του γονιδίου έχοντας λάβει θετική απάντηση στο διαγνωστικό τεστ ως εξής:

$$P(G|+) = \frac{P(+|G)P(G)}{P(+)}.$$

Με χρήση του θεωρήματος ολικής πιθανότητας για τον παρονομαστή, η παραπάνω εξίσωση μετασχηματίζεται ισοδύναμα:

$$\begin{aligned} P(G|+) &= \frac{P(+|G)P(G)}{P(+|G)P(G) + P(+|G^C)P(G^C)} \\ \Leftrightarrow P(G|+) &= \frac{P(+|G)P(G)}{P(+|G)P(G) + [1 - P(-|G^C)]\,P(G^C)}. \end{aligned}$$

Κατόπιν, με αντικατάσταση των διαθέσιμων δεδομένων προκύπτει:

$$P(G|+) = \frac{0,90 \times 0,28}{0,90 \times 0,28 + (1-0,93) \times 0,72}$$
$$\Leftrightarrow P(G|+) \approx 0,833.$$

Επομένως, ακόμα και αν ο έλεγχος διαγνώσει την ύπαρξη του γονιδίου στο γενετικό υλικό, η πιθανότητα το υποκείμενο να είναι πράγματι φορέας είναι λιγότερη από $85\%$. Δηλαδή, παρά τη σχετικά μεγάλη ευαισθησία του τεστ, υπάρχει αρκετό περιθώριο για false positives λόγω του μικρού ποσοστού εμφάνισης του γονιδίου.

Συνοψίζοντας, η διαδικασία που ακολουθήθηκε έχει ως εξής:

1. Εντοπίζουμε την αρχική/πρότερη (prior) πεποίθηση που έχουμε γύρω από την υπόθεση [εδώ $P(G)$].

2. Συλλέγουμε σχετικά δεδομένα ή αποδείξεις π.χ. μέσω μελέτης, διαγνωστικών τεστ κλπ. [εδώ $P(+|G)$].

3. Με χρήση του κανόνα του Bayes καταλήγουμε στην ύστερη πιθανότητα (posterior probability), την ενημερωμένη πεποίθηση σχετικά με την υπόθεση [εδώ $P(G|+)$].

Διασαφηνίζεται πως η αρίθμηση στα βήματα $1$ και $2$ δεν υπονοεί κάποια αυστηρή χρονική αλληλουχία μεταξύ τους· υπάρχει η δυνατότητα ανανέωσης της prior σε νέα posterior κάθε φορά που εμφανίζονται νέα δεδομένα ή εναλλακτικά η συλλογή όλων των δεδομένων και ύστερα ο υπολογισμός της posterior.

Μέσα από αυτή την πρώτη παρουσίαση του κανόνα του Bayes περιγράφεται το πώς μπορεί να χρησιμοποιηθεί σαν ένα σύστημα που μεταμορφώνει αρχικές και συλλεχθείσες πληροφορίες σε ύστερη γνώση (Zellner 1988), μια διαδικασία που καλείται και *Bayesian updating*. Η δράση του συστήματος αυτού δεν περιορίζεται στον υπολογισμό πιθανοτήτων αλλά επεκτείνεται και στην αναθεώρηση της πληροφορίας σχετικά με κατανομές και τις παραμέτρους αυτών, όπως θα δούμε στην γενίκευση του κανόνα του Bayes που ακολουθεί.

## 1.3 Γενίκευση του Θεωρήματος Bayes

Η παραπάνω διατύπωση και το αντίστοιχο παράδειγμα, αφορούν υπολογισμούς με χρήση πιθανοτήτων, γεγονός που ενέχει δυσκολίες στη διαδικασία της στατιστικής συμπερασματολογίας. Αν και δεν υπάρχει αμφιβολία ως προς την ορθότητα των υπολογισμών αυτών, στην πράξη οι περιπτώσεις που έχουμε απόλυτη γνώση της πιθανότητας εμφάνισης ενός ενδεχομένου είναι σπάνιες, αν όχι ανύπαρκτες. Επίσης, ακόμα και αν έχουμε επαρκή στοιχεία για τις τιμές των πιθανοτήτων, ενδέχεται διαφορετικοί ερευνητές να οδηγηθούν σε διαφορετικές εκτιμήσεις κατά τον υπολογισμό της πρότερης πιθανότητας η οποία δεν είναι κάποια σταθερή πληθυσμιακή ποσότητα αλλά βασίζεται στο διαθέσιμο δείγμα. Αυτές οι δυσκολίες προκύπτουν εξ' ορισμού των θέσεων της Μπεϋζιανής θεωρίας: η πιθανότητα αποτελεί μέτρο της αβεβαιότητας και είναι υποκειμενική.

Στην Μπεϋζιανή στατιστική λοιπόν προτιμάται η χρήση συναρτήσεων πιθανότητας αντί για τις ίδιες τις πιθανότητες αφού οι πρώτες μπορούν και συσσωρεύουν όλη την αβεβαιότητα γύρω από τις τιμές των δεύτερων. Εδώ η αβεβαιότητα σχετικά με την κατανομή οφείλεται στις

άγνωστες παραμέτρους αυτής, οι οποίες μάλιστα θεωρούνται τυχαίες μεταβλητές. Πλέον η ανάλυση συνίσταται στην αναπαράσταση της αρχικής αβεβαιότητας σχετικά με τις άγνωστες παραμέτρους και στη χρήση δεδομένων που θα προσφέρουν μια ανανεωμένη θεώρηση με μικρότερη αβεβαιότητα (Lynch 2007). Σε αυτό το πλαίσιο, το θεώρημα του Bayes γενικεύεται ώστε να ισχύει και για συναρτήσεις πιθανότητας τυχαίων μεταβλητών.

Η γενίκευση αυτή αφορά συναρτήσεις πυκνότητας πιθανότητας συνεχών τυχαίων μεταβλητών αλλά και συναρτήσεις μάζας πιθανότητας διακριτών μεταβλητών όπου τα ολοκληρώματα αντικαθίστανται από αθροίσματα. Ωστόσο, χάριν απλότητας, στην παρούσα διπλωματική θα χρησιμοποιούμε συμβολισμό που παραπέμπει στη μελέτη της συνεχούς περίπτωσης. Για τη διατύπωση λοιπόν του γενικευμένου θεωρήματος ορίζουμε πρώτα απ' όλα την πρότερη και την ύστερη συνάρτηση πυκνότητας πιθανότητας, όπως πριν ορίσαμε την πρότερη και την ύστερη πιθανότητα, ως:

$$p(\boldsymbol{\theta}) \text{ και } p(\boldsymbol{\theta}|\boldsymbol{x})$$

αντίστοιχα, όπου $\boldsymbol{\theta}$ το διάνυσμα των άγνωστων παραμέτρων κάποιας κατανομής και $x$ το σύνολο των διαθέσιμων δεδομένων. Στη θέση του $\boldsymbol{\theta}$ θα μπορούσε να βρίσκεται οποιαδήποτε τυχαία μεταβλητή, μονοδιάστατη ή πολυδιάστατη.

Σύμφωνα με τις παραπάνω παραδοχές, ο νόμος του Bayes διαμορφώνεται ως εξής:

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})}. \tag{4}$$

Η *πρότερη κατανομή* (*prior distribution*) $p(\boldsymbol{\theta})$ περιγράφει τη γνώση που έχουμε σχετικά με την παράμετρο χωρίς να λάβουμε υπόψη τα δεδομένα $\boldsymbol{x}$. Η *ύστερη κατανομή* (*posterior distribution*) $p(\boldsymbol{\theta}|\boldsymbol{x})$ εμπεριέχει την πληροφορία που μας παρέχουν τα δεδομένα και αποτελεί μια ανανεωμένη θεώρηση, όπως αυτή προέκυψε αφού "είδαμε" τα $\boldsymbol{x}$. Επισημαίνεται ότι η χρήση της λέξης "κατανομή" (distribution) δεν υποδηλώνει πως εργαζόμαστε με τη συνάρτηση κατανομής, αλλά αναφέρεται στις πληροφορίες που έχουμε για την κατανομή μέσω των παραμέτρων των συναρτήσεων πιθανότητας. Η ποσότητα $p(\boldsymbol{x})$ στον παρανομαστή ονομάζεται *περιθώρια πιθανοφάνεια* (*marginal likelihood*) του δείγματος ενώ ο όρος $p(\boldsymbol{x}|\boldsymbol{\theta})$ συνιστά την *πιθανοφάνεια* του δείγματος (*likelihood*), αφού αποτελεί συνάρτηση της παραμέτρου για τις παρατηρηθείσες τιμές του $\boldsymbol{x}$, και συμβολίζεται εναλλακτικά ως $L(\boldsymbol{\theta})$.

Αν αναλογιστούμε πως η περιθώρια πιθανοφάνεια είναι σταθερός αριθμός, για τα δεδομένα $\boldsymbol{x}$ που έχουν ήδη παρατηρηθεί, η Εξίσωση (4) μπορεί να γραφεί με τη μορφή αναλογίας:

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}) \tag{5}$$

απ' όπου φτάνουμε στην ακόλουθη διατύπωση η οποία ουσιαστικά ενσαρκώνει την κεντρική ιδέα της Μπεϋζιανής ανάλυσης:

$$\text{ύστερη κατανομή} \propto \text{πιθανοφάνεια} \times \text{πρότερη κατανομή}. \tag{6}$$

Το αποτέλεσμα της Σχέσης (6), όσο απλό και να φαίνεται εκ πρώτης όψεως, είναι ένα πολύτιμο εργαλείο: η αναλογία μεταξύ της ύστερης και του γινομένου της πρότερης με την πιθανοφάνεια φανερώνει το σχήμα της κατανομής της ύστερης (Bolstad 2010). Ωστόσο, για να έχουμε μια πλήρη περιγραφή της κατανομής οφείλουμε να προσδιορίσουμε τον παράγοντα κανονικοποίησης μέσω του παρονομαστή.

Για τον προσδιορισμό του παρονομαστή, η παρουσία του οποίου καθιστά το πηλίκο της Εξίσωσης (4) συνάρτηση πιθανότητας, θα χρησιμοποιήσουμε την ακόλουθη μορφή:

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

όπου $\Theta$ ο παραμετρικός χώρος. Στη γενική περίπτωση, η παραπάνω συνάρτηση-ολοκλήρωμα γράφεται ως:

$$p(\text{data}) = \int_{\Theta} p(\text{data}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{7}$$

και ονομάζεται *πρότερη προβλεπτική κατανομή* (*prior predictive distribution*). Ο υπολογισμός της πρότερης προβλεπτικής κατανομής πάνω σε ένα σετ δεδομένων ($data$) μας δίνει την ποσότητα στην οποία έχουμε ήδη αναφερθεί ως περιθώρια πιθανοφάνεια για τα συγκεκριμένα δεδομένα.

Η διαδικασία υπολογισμού του ολοκληρώματος της Εξίσωσης (7) που μας οδηγεί στην τιμή της περιθώριας πιθανοφάνειας δημιουργεί τις μεγαλύτερες δυσκολίες σε αυτό το κομμάτι της Μπεϋζιανής θεωρίας. Μάλιστα, αποτελεί τον κύριο λόγο για τον οποίο οι επιστήμονες για πολλά χρόνια απέφευγαν την πρακτική εφαρμογή της. Σε πολλές περιπτώσεις, απαιτείται να καταφύγουμε σε αριθμητικές μεθόδους οι οποίες στο παρελθόν ήταν εξαιρετικά χρονοβόρες, ελλείψει απαραίτητης υπολογιστικής δύναμης, ενώ δεν σπανίζουν οι περιπτώσεις που ούτε αυτές οι αριθμητικές μέθοδοι είναι εφαρμόσιμες. Μετά το 1990, η ραγδαία τεχνολογική ανάπτυξη, καθώς και η ανάπτυξη των *Markov Chain Monte Carlo (MCMC)* μεθόδων, συνέβαλλαν στην εδραίωση της Μπεϋζιανής στατιστικής (Ntzoufras 2009), αφήνοντας πίσω τα όποια προβλήματα υπολογισιμότητας. Ο τρόπος με τον οποίο οι MCMC μέθοδοι αναβίωσαν τη σημαντικότητα της Μπεϋζιανής θεωρίας θα μελετηθεί στην Ενότητα 2.

Στον αντίποδα, υπάρχουν προβλήματα ορισμένα με τέτοιο τρόπο ώστε η ανεύρεση της ύστερης κατανομής να γίνεται σχετικά εύκολα με αναλυτικό τρόπο. Η αντίστοιχη μεθοδολογία, όπου με τη χρήση ειδικών πρότερων κατανομών και της σχέσης αναλογίας (5) είμαστε σε θέση να εξάγουμε κλειστό τύπο για την εκ των υστέρων κατανομή, αποτελεί το αντικείμενο μελέτης της Υποενότητας που ακολουθεί.

## 1.4 Πρότερες κατανομές (Prior distributions)

Για την εφαρμογή του κανόνα του Bayes καλούμαστε να κάνουμε κάποιες αρχικές υποθέσεις σχετικά με την προς εξέταση κατανομή, που στη συνέχεια θα ανανεωθούν υπό το φως των διαθέσιμων δεδομένων. Η πρότερη κατανομή που θα υιοθετηθεί ενσαρκώνει τις υποθέσεις αυτές και παρέχει μια πρώτη, υποκειμενική εκτίμηση της κατανομής. Η υποκειμενικότητα της εκτίμησης έγκειται στο γεγονός ότι η επιλογή της πρότερης είναι στην ευχέρεια του ερευνητή επιτρέποντάς του να εκφράσει τις πληροφορίες που ήδη διαθέτει για το $\boldsymbol{\theta}$, να αποδώσει δηλαδή το τρέχον επίπεδο αβεβαιότητας. Αντιλαμβανόμαστε ότι η επιλογή αυτή πρέπει να γίνει προσεκτικά: θέλουμε η συναρτησιακή μορφή της πρότερης να αντικατοπτρίζει την αβεβαιότητα σχετικά με το $\boldsymbol{\theta}$ και οι επιλεχθείσες τιμές τυχόν υπερπαραμέτρων να αντικατοπτρίζουν τις πεποιθήσεις μας.

### 1.4.1 Συζυγείς πρότερες κατανομές (Conjugate prior distributions)

Όπως επισημάνθηκε παραπάνω, μια ισοδύναμη διατύπωση της Εξίσωσης (4) είναι η ακόλουθη:

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{8}$$

Είναι φανερό από την Εξίσωση (8) πως η εξαγωγή της ύστερης περιλαμβάνει τον υπολογισμό ολοκληρωμάτων που ωστόσο ενδέχεται να μην είναι εφικτός. Στην πραγματικότητα όμως υπάρχουν συγκεκριμένες επιλογές πρότερων που μπορούν να μας απαλλάξουν από τα αναλυτικά μη-διαχειρίσιμα ολοκληρώματα.

Ένα εργαλείο για την αποφυγή των ολοκληρωμάτων και τη διευκόλυνση της εκτίμησης της ύστερης κατανομής είναι οι *συζυγείς πρότερες κατανομές* (*conjugate prior distributions*): μία συζυγής πρότερη είναι μία πρότερη συνάρτηση πιθανότητας που ανήκει στην ίδια οικογένεια κατανομών με αυτή της ύστερης κατανομής. Επομένως, διαλέγοντας μια συζυγή πρότερη κατανομή που αντιπροσωπεύει ικανοποιητικά τις αρχικές μας πεποιθήσεις, είμαστε σε θέση να γνωρίζουμε πλήρως την ύστερη στην οποία θα οδηγηθούμε, όχι μόνο το σχήμα της, χωρίς τον υπολογισμό ολοκληρωμάτων. Η τεχνική που περιγράφηκε ονομάζεται *Conjugate Analysis* και είναι απόρροια της σχέσης αναλογίας μεταξύ ύστερης και πρότερης που είδαμε στην εναλλακτική διατύπωση του θεωρήματος Bayes στην Εξίσωση (5). Στο σημεία αυτό πρέπει να επισημανθεί ότι η αποτελεσματική λειτουργία αυτής της μεθοδολογίας προϋποθέτει και κατάλληλο, κατά περίπτωση, συνδυασμό της κατανομής των δεδομένων, και κατά συνέπεια της πιθανοφάνειας, με την κατανομή της πρότερης.

Αποδεικνύεται λοιπόν ότι υπάρχουν ζεύγη πρότερων κατανομών και συναρτήσεων πιθανοφάνειας για τα οποία γνωρίζουμε ότι η προκύπτουσα ύστερη κατανομή θα ανήκει στην ίδια οικογένεια με την πρότερη. Σε αυτή την περίπτωση θα λέμε ότι η πρότερη κατανομή είναι συζυγής της συνάρτησης πιθανοφάνειας. Στο παράδειγμα που ακολουθεί θα δείξουμε ότι η κανονική πρότερη κατανομή είναι συζυγής της κανονικής συνάρτησης πιθανοφάνειας, με γνωστή διασπορά, δηλαδή η κανονική κατανομή είναι συζυγής πρότερη του εαυτού της.

Υποθέτουμε ότι διαθέτουμε $n$ παρατηρήσεις από μία τυχαία μεταβλητή (τ.μ.) που παίρνει πραγματικές τιμές, επομένως μπορούμε να θεωρήσουμε ότι ακολουθεί την κανονική κατανομή. Συμβολίζοντας τις παρατηρήσεις με $x_i$, το διάνυσμα αυτών με $\boldsymbol{x}$, την άγνωστη μέση τιμή και τη γνωστή διασπορά τους με $\mu$ και $\sigma^2$ αντίστοιχα, η πιθανοφάνεια των δεδομένων θα είναι:

$$\begin{aligned} p(\boldsymbol{x}|\mu,\sigma^2) &= \prod_{i=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2+n\mu^2-2n\mu\bar{x}\right)\right\}. \end{aligned}$$

Αναζητούμε την κατανομή της μέσης τιμής $\mu$ δοθείσης της διασποράς $\sigma^2$ και επομένως καλούμαστε να διαλέξουμε μια πρότερη κατανομή $p(\mu|\sigma^2)$ την οποία θα πολλαπλασιάσουμε με την παραπάνω πιθανοφάνεια ώστε να καταλήξουμε στην ύστερη. Έστω ότι η πρότερη

γνώση μας για την κατανομή του $\mu|\sigma^2$ αποτυπώνεται στην κανονική κατανομή με μέση τιμή $\mu_0^2$ και διασπορά $\sigma_0^2$, δηλαδή:

$$\begin{aligned}\mu|\sigma^2 &\sim \mathcal{N}(\mu_0^2, \sigma_0^2) \Leftrightarrow \\ p(\mu|\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \\ &\propto \exp\left(-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right).\end{aligned}$$

Έχοντας προσδιορίσει τις δύο ποσότητες που, σύμφωνα με την Εξίσωση (5), απαιτούνται για τον υπολογισμό της ύστερης, μπορούμε να προχωρήσουμε στην αντικατάστασή τους στην παραπάνω έκφραση:

$$\begin{aligned}p(\mu|\boldsymbol{x}, \sigma^2) &\propto p(\boldsymbol{x}|\mu, \sigma^2) \times p(\mu|\sigma^2) \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 + n\mu^2 - 2n\mu\bar{x}\right)\right\} \qquad (9) \\ &\times \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\} \\ &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 + n\mu^2 - 2n\mu\bar{x}\right) - \frac{1}{2\sigma_0^2}\left(\mu^2 + \mu_0^2 - 2\mu_0\mu\right)\right\} \\ &\propto \exp\left\{-\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right) + 2\mu\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}\right) - \left(\frac{\mu_0^2}{2\sigma_0^2} + \frac{\sum_i x_i^2}{2\sigma^2}\right)\right\} \\ &\propto \exp\left\{-\left[\frac{\mu^2}{2}\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right) - 2\mu\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}\right) + \frac{1}{2}\left(\frac{\mu_0^2}{\sigma_0^2} + \frac{\sum_i x_i^2}{\sigma^2}\right)\right]\right\}. \qquad (10)\end{aligned}$$

Για την περαιτέρω επεξεργασία της ύστερης θα ορίσουμε τις ακόλουθες ποσότητες:

$$\begin{aligned}\tilde{\sigma}^2 &= \frac{\sigma^2\sigma_0^2}{n\sigma_0^2 + \sigma^2} = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} \\ \tilde{\mu} &= \tilde{\sigma}^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{x}}{\sigma^2}\right)\end{aligned}$$

οπότε η Εξίσωση (9) διαμορφώνεται ως εξής:

$$\begin{aligned}p(\mu|\boldsymbol{x}, \sigma^2) &\propto \exp\left[-\frac{1}{2\tilde{\sigma}^2}\left(\mu^2 - 2\mu\tilde{\mu} + \tilde{\mu}^2\right)\right] \\ &\propto \exp\left[-\frac{1}{2\tilde{\sigma}^2}\left(\mu - \tilde{\mu}\right)^2\right].\end{aligned}$$

Από τη μορφή της τελευταίας σχέσης προκύπτει ότι, όπως προβλέψαμε, η ύστερη συνάρτηση πυκνότητας πιθανότητας θα ανήκει στην κανονική κατανομή με μέση τιμή $\tilde{\mu}$ και διασπορά $\tilde{\sigma}^2$, δηλαδή:

$$p(\mu|\boldsymbol{x}, \sigma^2) = \frac{1}{\sqrt{2\pi\tilde{\sigma}^2}} \exp\left(-\frac{(\mu-\tilde{\mu})^2}{2\tilde{\sigma}^2}\right).$$

Ανακεφαλαιώνοντας, είδαμε ότι για κανονικά δεδομένα, άρα και κανονική πιθανοφάνεια, η χρήση κανονικής κατανομής ως πρότερη, μας οδήγησε σε κανονική ύστερη κατανομή. Πράγματι λοιπόν η κανονική πρότερη κατανομή είναι συζυγής της κανονικής πιθανοφάνειας.

### 1.4.2 Μη-πληροφοριακές πρότερες και η πρότερη του Jeffreys

Η χρήση συζυγών πρότερων επιλύει πολλά προβλήματα μη-υπολογισιμότητας. Ταυτόχρονα όμως επηρεάζει έντονα την προκύπτουσα εκ των υστέρων κατανομή αφού στην πραγματικότητα η επιλογή μιας πρότερης συνάρτησης πιθανότητας καθορίζει την οικογένεια κατανομών στην οποία θα ανήκει η ύστερη. Ωστόσο, υπάρχουν περιπτώσεις στις οποίες είτε δεν έχουμε επαρκή πρότερη γνώση για το $\boldsymbol{\theta}$ είτε προτιμάμε να υιοθετήσουμε μια πιο αντικειμενική προσέγγιση. Για τον σκοπό αυτό, καταφεύγουμε στη χρήση *μη-πληροφοριακών πρότερων* (*non-informative* ή *uninformative priors*) δηλαδή κατανομών που δεν κατευθύνουν το αποτέλεσμα του Bayesian updating αλλά αφήνουν τα δεδομένα να μας οδηγήσουν στο κατάλληλο συμπέρασμα.

Μια συνήθης επιλογή μη-πληροφοριακής πρότερης είναι η *επίπεδη πρότερη* (*flat prior*), η οποία θεωρεί όλες τις τιμές του παραμετρικού χώρου ισοπίθανες, δηλαδή:

$$p(\boldsymbol{\theta}) \propto c, c = \text{σταθερό}.$$

Από τη σχέση (5) αντιλαμβανόμαστε πως η ύστερη κατανομή θα είναι ανάλογη της πιθανοφάνειας και μόνο, με αποτέλεσμα τα δεδομένα να έχουν τον πρώτο λόγο στη διαμόρφωση της ύστερης. Η υιοθέτηση τέτοιων κατανομών, όπως για παράδειγμα η ομοιόμορφη, φαίνεται να είναι η προφανής λύση όταν επιδιώκεται να διατηρήσουμε συγκρατημένη στάση απέναντι στις εκ τω προτέρων πληροφορίες σχετικά με το $\boldsymbol{\theta}$. Παρ' όλα αυτά, ανακύπτουν δύο σημαντικά προβλήματα κατά την εφαρμογή τους.

Το πρώτο ζήτημα αφορά στο γεγονός ότι σε μη-συμπαγείς παραμετρικούς χώρους, όπως στην περίπτωση $\boldsymbol{\theta} \in (-\infty, +\infty)$, οι επίπεδες πρότερες δεν συνιστούν έγκυρες συναρτήσεις πυκνότητας πιθανότητας αφού το $\int_{-\infty}^{+\infty} p(\boldsymbol{\theta})$ απειρίζεται. Αυτού του είδους οι πρότερες ονομάζονται *improper* και μερικές φορές αναφέρονται και ως "ακατάλληλες" στην ελληνική βιβλιογραφία. Ο χαρακτηρισμός αυτών των πρότερων ως improper δεν αποτελεί εμπόδιο στη χρήση τους, αρκεί η προκύπτουσα ύστερη να είναι μία καλά ορισμένη κατανομή.

Το δεύτερο και σημαντικότερο ζήτημα είναι πως οι επίπεδες πρότερες δεν παραμένουν αναλλοίωτες σε μετασχηματισμούς των παραμέτρων, με αποτέλεσμα σε αυτό το σενάριο να χάνουν τον μη-πληροφοριακό χαρακτήρα τους. Για παράδειγμα, έστω μία τυχαία μεταβλητή που ακολουθεί την κατανομή Bernoulli με παράμετρο $\theta \in (0, 1)$ για την οποία παράμετρο χρησιμοποιούμε την επίπεδη πρότερη $p(\theta) = 1$. Η επιλογή της πρότερης αντιπροσωπεύει την έλλειψη πληροφοριών σχετικά με το $\theta$ πριν τη συλλογή δεδομένων από τη διεξαγωγή τυχόν πειράματος. Ένας έγκυρος μετασχηματισμός του $\theta$ προκύπτει από τη σχέση:

$$\psi = \log(\theta/(1-\theta))$$

και εύκολα υπολογίζεται ότι η κατανομή του $\psi$ θα είναι η παρακάτω:

$$p(\psi) = \frac{e^{\psi}}{(1+e^{\psi})^2}$$

η οποία προφανώς δεν είναι επίπεδη. Συμβαίνει λοιπόν το εξής παράδοξο: ενώ ισχυριζόμαστε ότι δεν έχουμε πρότερη πληροφορία για το $\theta$, δεν φαίνεται να συμβαίνει το ίδιο και για το $\log(\theta/(1-\theta))$ (Liu & Wasserman 2014).

Τη λύση στο πρόβλημα του μη-αναλλοίωτου των επίπεδων πρότερων σε μετασχηματισμούς των παραμέτρων έρχεται να δώσει ο Sir Harold Jeffreys (Jeffreys 1961). Ο Jeffreys ισχυρίστηκε πως η κατασκευή μη-πληροφοριακής πρότερης, που είναι ταυτόχρονα αναλλοίωτη από μετασχηματισμούς, μπορεί να επιτευχθεί μέσω του υπολογισμού της πληροφορίας κατά Fisher (Fisher information). Η πληροφορία κατά Fisher συμβολίζεται ως $I(\boldsymbol{\theta})$ και είναι ένας δείκτης που συνοψίζει την ποσότητα της πληροφορίας σχετικά με το άγνωστο διάνυσμα παραμέτρων $\boldsymbol{\theta}$ που εμπεριέχεται σε μια παρατηρήσιμη τυχαία μεταβλητή.

Εστιάζοντας στην περίπτωση μονοδιάστατης παραμέτρου $\theta$, η πληροφορία κατά Fisher δίνεται από τον ακόλουθο τύπο:

$$I(\theta) = E_{\boldsymbol{X}|\theta}\left[\left(\frac{\partial}{\partial\theta}\log p(\boldsymbol{X} \mid \theta)\right)^2\right] = -E_{\boldsymbol{X}|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log p(\boldsymbol{X} \mid \theta)\right]. \tag{11}$$

Αντίστοιχα στην πολυδιάστατη περίπτωση $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_d)$ ορίζεται ο $d \times d$ πίνακας πληροφορίας κατά Fisher του οποίου τα στοιχεία δίνονται ως εξής:

$$\begin{aligned}\left[I\left(\boldsymbol{\theta}\right)\right]_{ij} &= E_{\boldsymbol{X}|\boldsymbol{\theta}}\left[\left(\frac{\partial}{\partial\theta_i}\log p\left(\boldsymbol{X} \mid \boldsymbol{\theta}\right)\right)\left(\frac{\partial}{\partial\theta_j}\log p\left(\boldsymbol{X} \mid \boldsymbol{\theta}\right)\right)\right] \\ &= -E_{\boldsymbol{X}|\boldsymbol{\theta}}\left[\left(\frac{\partial}{\partial^2\theta_i\partial\theta_j}\log p\left(\boldsymbol{X} \mid \boldsymbol{\theta}\right)\right)\right], \text{ όπου } i, j = 1, ...d.\end{aligned} \tag{12}$$

Συγκεκριμένα, ο Jeffreys προτείνει την ακόλουθη μορφή πρότερης για την μονοδιάστατη περίπτωση:

$$p(\theta) \propto [I(\theta)]^{\frac{1}{2}} \tag{13}$$

και κατά συνέπεια την εξής πρότερη στην περίπτωση πολυδιάστατης παραμέτρου $\boldsymbol{\theta}$:

$$p(\boldsymbol{\theta}) \propto \left[\det I(\boldsymbol{\theta})\right]^{\frac{1}{2}}. \tag{14}$$

Επιστρατεύοντας τις παραπάνω προτάσεις, καταφέρνουμε να κατασκευάσουμε μη-πληροφοριακές πρότερες που δεν επηρεάζονται από αλλαγές στη μορφή των παραμέτρων της κατανομής.

Επιστρέφοντας στο παράδειγμα της κατανομής Bernoulli($\theta$), αποδεικνύεται (Zhu & Lu 2004) σύμφωνα με τα παραπάνω, ότι μια κατάλληλη πρότερη θα είναι η:

$$p(\theta) \propto \frac{1}{\sqrt{p(1-p)}},$$

καθώς ικανοποιεί την απαίτηση περί αναλλοίωτου.

## 1.5 Προβλεπτικές κατανομές

Ήδη, από την Εξίσωση (7), είδαμε μία μορφή προβλεπτικής κατανομής, την πρότερη προβλεπτική κατανομή. Συνοπτικά, πρόκειται για τη συνάρτηση πυκνότητας πιθανότητας των πιθανών συνόλων δεδομένων που μπορεί να προκύψουν, δοθέντων των αρχικών

υποθέσεων σχετικά με την κατανομή των παραμέτρων, όπως αυτές εκφράζονται μέσω της πρότερης κατανομής, αλλά και με βάση την πιθανοφάνεια που έχουμε ορίσει. Η κατανομή ορίζεται πριν δούμε τα δεδομένα και, όπως ήδη αναφέραμε, η απόκτηση των δεδομένων $\boldsymbol{x}$ μας οδηγεί σε μία τιμή που αποτελεί την περιθώρια πιθανοφάνεια:

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

Εν ολίγοις, η περιθώρια πιθανοφάνεια είναι η πιθανοφάνεια του μοντέλου που έχει ως κατανομή δείγματος την πρότερη προβλεπτική κατανομή.

Έχοντας συλλέξει τις παρατηρήσεις μας, προβαίνουμε στην εφαρμογή του κανόνα του Bayes ο οποίος κατά τα γνωστά θα μας οδηγήσει στην ύστερη κατανομή. Η ύστερη κατανομή, μας επιτρέπει να βγάλουμε συμπεράσματα για μελλοντικές παρατηρήσεις, έστω $x^*$, δεδομένων των αρχικών παρατηρήσεων και υποθέσεων οι οποίες εισήχθησαν στο θεώρημα Bayes. Ακολουθώντας παρόμοια συλλογιστική πορεία με αυτή για την κατασκευή της πρότερης προβλεπτικής κατανομής, μπορούμε να κατασκευάσουμε την *ύστερη προβλεπτική κατανομή* (*posterior predictive distribution*), χρησιμοποιώντας τώρα, όπως είναι φυσικό, την ύστερη κατανομή $p(\boldsymbol{\theta}|\boldsymbol{x})$. Επομένως καταλήγουμε στην ακόλουθη κατανομή για τις μελλοντικές παρατηρήσεις:

$$p(x^*|\boldsymbol{x}) = \int_{\Theta} p(x^*|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta}. \tag{15}$$

Η Εξίσωση (15) περιγράφει πώς περιμένουμε να συμπεριφέρονται τα καινούρια δεδομένα, βάσει των πρόσφατων υποθέσεων και ευρημάτων μας. Μάλιστα, μπορεί να χρησιμοποιηθεί τόσο για σημειακές εκτιμήσεις όσο και για εκτίμηση διαστημάτων εμπιστοσύνης αλλά και ελέγχους υποθέσεων.

# 2 Εισαγωγή στα Markov Chain Monte Carlo

Η συλλογή αλγορίθμων που αποκαλούμε Markov Chain Monte Carlo (MCMC) διακρίνεται από ένα τεράστιο εύρος επιστημονικών -και μη- πεδίων εφαρμογής, ίσως το μεγαλύτερο μεταξύ παρόμοιων αλγορίθμων. Ως πολύτιμο και βασικό υπολογιστικό εργαλείο της Μπεϋζιανής Στατιστικής, αποτελεί αναπόσπαστο κομμάτι της.

Οι απαρχές των MCMC μεθόδων εντοπίζονται στα μέσα του 20ου αιώνα, και συγκεκριμένα το 1949 στο Los Alamos των Η.Π.Α., όταν οι Metropolis και Ulam, στα πλαίσια της πυρετώδους έρευνας για την κατασκευή πυρηνικών όπλων (Richey 2010), οδηγήθηκαν στην έκδοση του άρθρου στο οποίο εισήχθη η *Μέθοδος Monte Carlo* (Metropolis & Ulam 1949). Ενώ αρχικά τα MCMC χρησιμοποιήθηκαν στα πλαίσια επίλυσης προβλημάτων της φυσικής, πολλές αναφορές και μελέτες αυτών εκδόθηκαν τα ακόλουθα χρόνια και στη στατιστική βιβλιογραφία, χωρίς όμως να αντιμετωπίζονται θερμά από τη στατιστική κοινότητα. Ήταν η ανάπτυξη του λογισμικού BUGS (Bayesian inference Using Gibbs Sampling) το 1991, που συνέβαλε σημαντικά, μεταξύ άλλων, στην μαζική υιοθέτηση των MCMC αλγορίθμων (Robert & Casella 2011), αφού προσέφερε τα κατάλληλα υπολογιστικά μέσα για την υλοποίηση των μεθόδων.

Η περαιτέρω εξέλιξη της τεχνολογίας γενικότερα επισφράγισε την επιτυχία των MCMC και εξασφάλισε την εδραίωσή τους. Χρησιμοποιώντας την τυχαιότητα και την προσομοίωση, τα MCMC καταφέρνουν να υπερπηδήσουν τα υπολογιστικά εμπόδια που αντιμετωπίζουν αναλυτικές και αριθμητικές μέθοδοι σε αντίστοιχες περιπτώσεις.

## 2.1 Μαρκοβιανές αλυσίδες και Μέθοδος Monte Carlo: Τα δομικά στοιχεία των MCMC

Η ονομασία των MCMC περιέχει τις δύο ιδέες που έχουν συνδυαστεί για τη δημιουργία τους: Μαρκοβιανές Αλυσίδες (Markov Chains) και Μέθοδος Monte Carlo. Κάθε μία παρουσιάζεται συνοπτικά στη συνέχεια.

Η βασική σύλληψη πάνω στην οποία στηρίχθηκε η Monte Carlo μεθοδολογία είναι πως μπορούμε να λάβουμε οποιαδήποτε πληροφορία για μια κατανομή (π.χ. μέση τιμή, τυπική απόκλιση, τεταρτημόρια) προσομοιώνοντας τιμές από αυτή. Όπως έχουμε ήδη διαπιστώσει, δεν σπανίζουν οι περιπτώσεις όπου υπολογισμοί με χρήση αναλυτικού τύπου δεν είναι εφικτοί ή δημιουργούν δυσκολίες. Ωστόσο, σύμφωνα με τους Metropolis και Ulam (Metropolis & Ulam 1949), μπορούμε να παρακάμψουμε τέτοιους χρονοβόρους, ή/και αδύνατους, υπολογισμούς γεννώντας απλώς τυχαίο δείγμα από την προς μελέτη κατανομή.

Πράγματι, ο Νόμος των Μεγάλων Αριθμών (Law of Large Numbers) μας εξασφαλίζει τη σύγκλιση του μέσου όρου ανεξάρτητων δειγμάτων μιας κατανομής στη μέση τιμή της κατανομής αυτής, όταν διαθέτουμε μεγάλο μέγεθος δείγματος. Συγκεκριμένα, αν $X_1, X_2, ...$ ανεξάρτητες και ισόνομες διακριτές τυχαίες μεταβλητές από την κατανομή $\pi$, ο Ισχυρός Νόμος των Μεγάλων Αριθμών έχει ως εξής:

$$\mathbb{P}\left[\lim_{n\to\infty} \| \frac{f(X_1) + f(X_2) + ... + f(X_n)}{n} - \mathbb{E}_\pi[f] \| = 0\right] = 1. \tag{16}$$

Η παραπάνω Εξίσωση δηλώνει τη σχεδόν σίγουρη σύγκλιση (almost sure convergence) του

αριθμητικού μέσου στη μέση τιμή της κατανομής, πράγμα που σημαίνει ότι όταν το $n$ είναι μεγάλο οι δύο ποσότητες είναι κοντά με πιθανότητα 1.

Με παρόμοιο τρόπο διατυπώνεται και ο Ασθενής Νόμος των Μεγάλων Αριθμών που εξασφαλίζει την κατά πιθανότητα σύγκλιση:

$$\lim_{n\to\infty} \mathbb{P}\left[ || \frac{f(X_1) + f(X_2) + ... + f(X_n)}{n} - \mathbb{E}_\pi[f] \,||> \epsilon \right] = 0. \tag{17}$$

Σε αυτή την περίπτωση, το συμπέρασμα είναι πως η πιθανότητα ο αριθμητικός μέσος να απέχει περισσότερο από $\epsilon$ από τη μέση τιμή της κατανομής τείνει στο μηδέν.

Παραπάνω αναφερθήκαμε, χάριν ευκολίας, σε διακριτές τυχαίες μεταβλητές. Φυσικά και τα δύο Θεωρήματα έχουν ισχύ και στην περίπτωση συνεχών τυχαίων μεταβλητών, με τις απαραίτητες αλλαγές στη διατύπωση.

Επομένως, αρκεί να έχουμε πρόσβαση σε μεγάλο δείγμα από την κατανομή που μας ενδιαφέρει και ο νόμος των μεγάλων αριθμών μας εγγυάται την ικανοποιητική προσέγγιση της μέσης τιμής. Το ανεξάρτητο δείγμα συνήθως λαμβάνεται μέσω κάποιου αλγορίθμου προσομοίωσης, όπως η μέθοδος απόρριψης ή αλλιώς rejection sampling (Casella, Robert, & Wells 2004).

Φυσικά, δεν είναι πάντα εύκολο να πάρουμε δείγματα από την κατανομή $\pi$ και μάλιστα υπάρχουν πολλές ενδιαφέρουσες εφαρμογές όπου η διαδικασία είναι εξαιρετικά δύσκολη όπως όταν εργαζόμαστε με πολυδιάστατες κατανομές. Ακόμα, μπορεί να μην έχουμε πλήρη εικόνα σχετικά με τον τύπο της κατανομής $\pi$ αλλά να τη γνωρίζουμε μόνο μέσω κάποιας σχέσης αναλογίας. Το τελευταίο αυτό ζήτημα δεν μας είναι άγνωστο καθώς αποτελεί σύνηθες πρόβλημα κατά την εκτίμηση της ύστερης κατανομής στα πλαίσια της Μπεϋζιανής συμπερασματολογίας.

Όταν ανακύπτουν τέτοιου είδους προβλήματα, δεν είναι υπολογιστικά βιώσιμη η παραγωγή ανεξάρτητου δείγματος με κανέναν σχετικό αλγόριθμο. Ωστόσο, οι Metropolis και Ulam ουδέποτε εξέφρασαν κάποια απαίτηση ανεξαρτησίας των τυχαίων μεταβλητών στο άρθρο τους, οπότε θεωρητικά θα μπορούσαμε να υλοποιήσουμε την ιδέα τους ακόμα και στην περίπτωση που διαθέτουμε εξαρτημένο δείγμα. Προκύπτει λοιπόν ένα νέο ερώτημα: πώς μπορούμε να δημιουργήσουμε εξαρτημένο δείγμα από μία κατανομή που δεν γνωρίζουμε απαραίτητα;

Σε αυτό το σημείο υπεισέρχεται η έννοια των Μαρκοβιανών Αλυσίδων: μια *Μαρκοβιανή Αλυσίδα* (*μ.α.*) είναι μια στοχαστική διαδικασία με την ιδιότητα η πιθανότητα μετάβασης στην επόμενη κατάσταση να εξαρτάται αποκλειστικά από την τρέχουσα κατάσταση του συστήματος (Λουλάκης 2015). Η ιδιότητα αυτή, που ονομάζεται Μαρκοβιανή, πρακτικά σημαίνει πως η μόνη κατάσταση που επηρεάζει το μέλλον της αλυσίδας είναι η τρέχουσα, καθώς το παρελθόν δεν μας προσφέρει περαιτέρω πληροφορίες για την εξέλιξή της.

Σε μια τέτοια ανέλιξη η τυχαία μεταβλητή του κάθε βήματος θα έχει δεσμευμένη εξάρτηση από την τυχαία μεταβλητή του αμέσως προηγούμενου βήματος. Συμπεραίνουμε λοιπόν πως οι τυχαίες μεταβλητές που περιγράφουν τις καταστάσεις του συστήματος εμφανίζουν ανά δύο δεσμευμένη εξάρτηση και σαν συλλογή θα αποτελούν ένα εξαρτημένο σύνολο δεδομένων. Συνεπώς, οι τυχαίες μεταβλητές που απαρτίζουν την μ.α. μπορούν να χρησιμοποιηθούν σαν δείγμα και η ίδια η μ.α. σαν γεννήτρια τυχαίου, αλλά όχι ανεξάρτητου, δείγματος.

Βέβαια, για να μας είναι χρήσιμο το εν λόγω δείγμα, θα πρέπει η κατανομή των τυχαίων μεταβλητών να είναι η ζητούμενη κατανομή $\pi$. Για τον σκοπό αυτό, αρκεί η αλυσίδα να έχει ως *στάσιμη ή αναλλοίωτη κατανομή* (*stationary* ή *invariant distribution*) την $\pi$. Η ύπαρξη και η μοναδικότητα της στάσιμης κατανομής δεν είναι εν γένει δεδομένες αλλά στην περίπτωση που εξασφαλίζονται γνωρίζουμε ότι μόλις η αλυσίδα φτάσει να χαρακτηρίζεται από τη στάσιμη κατανομή της σε κάποιο βήμα, θα τη διατηρήσει σε όλα τα ακόλουθα βήματα (Gamerman D. 2006). Αυτή η συμπεριφορά είναι πολύτιμη όταν επιχειρούμε την παραγωγή δείγματος από συγκεκριμένη κατανομή: αν καταφέρουμε να κατασκευάσουμε μ.α. η οποία εγγυημένα θα έχει στάσιμη κατανομή μένει μονάχα να την τροφοδοτήσουμε με την κατάλληλη αρχική κατανομή ώστε μετά από μία σειρά βημάτων να καταλήξει στην $\pi$. Σε κάθε επόμενο βήμα, οι τυχαίες μεταβλητές θα κατανέμονται σύμφωνα με την $\pi$ και θα σχηματίζουν σιγά-σιγά το τυχαίο δείγμα που θα αξιοποιηθεί στη Monte Carlo εκτίμηση.

Για να διατυπώσουμε τις συνθήκες που εξασφαλίζουν την ύπαρξη μοναδικής αναλλοίωτης κατανομής θα χρειαστούμε τους παρακάτω χαρακτηρισμούς των Μαρκοβιανών αλυσίδων:

**Ορισμός 1** *Μία μ.α. ονομάζεται μη υποβιβάσιμη (irreducible) αν από την τρέχουσα κατάσταση, όποια και να είναι αυτή, μπορούμε να φτάσουμε σε οποιαδήποτε άλλη κατάσταση σε πεπερασμένο χρόνο.*

**Ορισμός 2** *Λέμε ότι μία κατάσταση είναι γνησίως επαναληπτική (positive recurrent) όταν η αναμενόμενη τιμή του χρόνου πρώτης επιστροφής στην κατάσταση είναι πεπερασμένη. Εν' ολίγοις, για μια γνωσίως επαναληπτική κατάσταση γνωρίζουμε με πιθανότητα 1 ότι αν ξεκινήσουμε από αυτή θα την επισκεφτούμε ξανά σε πεπερασμένο χρόνο. Αντίστοιχα, μία αλυσίδα θα είναι γνησίως επαναληπτική αν κάθε κατάστασή της είναι γνησίως επαναληπτική.*

**Ορισμός 3** *Αν ο μέγιστος κοινός διαιρέτης των δυνατών χρόνων επιστροφής σε μία κατάσταση είναι το 1, η κατάσταση ονομάζεται απεριοδική (aperiodic). Για να χαρακτηρίσουμε μια αλυσίδα ως απεριοδική, πρέπει όλες οι καταστάσεις αυτής να είναι απεριοδικές.*

Αποδεικνύεται ότι κάθε μη υποβιβάσιμη και γνησίως επαναληπτική Μαρκοβιανή αλυσίδα θα έχει μοναδική αναλλοίωτη κατανομή (Λουλάκης 2015). Αν επιπλέον η αλυσίδα είναι και απεριοδική, θα έχει ως *ασυμπτωτική κατανομή* (*limiting distribution*) την αναλλοίωτη κατανομή της. Αυτό σημαίνει ότι σε βάθος χρόνου και όσο αφήνουμε την αλυσίδα να τρέχει, αυτή θα υιοθετήσει, και θα διατηρήσει, τη στάσιμη κατανομή της. Η επιπλέον απαίτηση της απεριοδικότητας, παρόλο που δεν επηρεάζει την ύπαρξη αναλλοίωτης κατανομής, είναι καθοριστική όσον αφορά τη σύγκλιση των πιθανοτήτων μετάβασης (Gamerman D. 2006).

Το αποτέλεσμα που έρχεται να συμπληρώσει την απόπειρά μας για την κατασκευή Monte Carlo εκτιμητών χωρίς την απαίτηση της ανεξάρτησίας του δείγματος είναι το *Εργοδικό Θεώρημα* (*Ergodic Theorem*). Το εν λόγω θεώρημα αφορά την ασυμπτωτική συμπεριφορά χρονικών μέσων όρων διαφόρων συναρτησιακών της αλυσίδας (Λουλάκης 2015), όπως ακριβώς και ο Νόμος των Μεγάλων Αριθμών στην περίπτωση ανεξάρτητων τ.μ. Αναλυτικότερα, το Εργοδικό Θεώρημα για γνησίως επαναληπτικές αλυσίδες σε διακριτό χώρο, το οποίο ορίζεται με ανάλογο τρόπο και για μ.α. σε συνεχείς χώρους, διατυπώνεται ως εξής:

**Θεώρημα 1** *Έστω $\{X_n\}_{n\geq 0}$ μη υποβιβάσιμη, γνησίως επαναληπτική αλυσίδα στον χώρο καταστάσεων $\mathbb{X}$ και $f : \mathbb{X} \to \mathbb{R}$ φραγμένη συνάρτηση με $\boldsymbol{E}_\pi[f] < \infty$. Ισχύει:*

$$\boldsymbol{P}\left[\frac{1}{n}\sum_{k=1}^{n} f(X_k) \to \boldsymbol{E}_\pi[f]\right] = 1,$$

*όπου $\boldsymbol{E}_\pi[f] = \sum_{x\in\boldsymbol{X}} f(x)\pi(x)$ η μέση τιμή της $f(X)$ όταν η $X$ ακολουθεί την αναλλοίωτη κατανομή $\pi$ της αλυσίδας.*

Επί της ουσίας, μέσω του Εργοδικού Θεωρήματος έχουμε εγγυημένη προσέγγιση της αναμενόμενης τιμής της συνάρτησης από τον αριθμητικό της μέσο, για κατάλληλα επιλεγμένη αλυσίδα. Μόλις επιτευχθεί η στασιμότητα, λαμβάνουμε εκτιμήσεις για συναρτήσεις της αλυσίδας, δηλαδή του δείγματος, όπως για παράδειγμα η μέση τιμή και η τυπική απόκλιση.

Μετά από αυτή την παρατήρηση, θα πρέπει πλέον να είναι αντιληπτός ο καταλυτικός ρόλος των MCMC στη Μπεϋζιανή μεθοδολογία. Όταν θέλουμε να πάρουμε πληροφορίες σχετικά με την εκ των υστέρων κατανομή $p(\boldsymbol{\theta}|x)$ μπορούμε να υπερπηδήσουμε τις δυσκολίες που αναφέραμε στη λήψη δείγματος από αυτή και αντ' αυτού να κατασκευάσουμε μία μη υποβιβάσιμη, γνησίως επαναληπτική και απεριοδική αλυσίδα με στάσιμη κατανομή την $p(\boldsymbol{\theta}|x)$. Μόλις η αλυσίδα φτάσει στην κατανομή ισορροπίας της, είμαστε σε θέση να συλλέξουμε το δείγμα και να υπολογίσουμε όποια ποσότητα μας ενδιαφέρει σχετικά με αυτό.

Οι MCMC μέθοδοι συνέβαλαν στην άρση υπολογιστικών περιορισμών σε πολλούς επιστημονικούς τομείς, όπως η μηχανική στατιστική και η σωματιδιακή φυσική, και έστρωσαν το έδαφος για την ανάπτυξη της Μπεϋζιανής στατιστικής. Στη συνέχεια θα μελετήσουμε τον αλγόριθμο Metropolis-Hastings και τη μέθοδο δειγματοληψίας του Gibbs (Gibbs sampling) που αποτελούν τα κύρια εργαλεία στοχαστικής προσομοιώσης για την Μπεϋζιανή συμπερασματολογία.

## 2.2 Ο Αλγόριθμος Metropolis-Hastings

Ο αλγόριθμος Metropolis-Hastings (M-H) φέρει τα ονόματα των συντακτών των δύο βασικών έργων που συνετέλεσαν στην ανάπτυξη και τον εμπλουτισμό της μεθόδου (Metropolis, A. Rosenbluth, M. Rosenbluth, H. Teller, & E. Teller 1953 και Hastings 1970). Πρόκειται για μια σειρά βημάτων και οδηγιών που πραγματώνει όσα περιγραφήκαν θεωρητικά στην προηγούμενη ενότητα: μας επιτρέπει να πάρουμε δείγμα από μία συνάρτηση πιθανότητας, της οποίας γνωρίζουμε τη γενική μορφή αλλά όχι τον παράγοντα κανονικοποίησης, κατασκευάζοντας μια Μαρκοβιανή αλυσίδα που έχει σαν στάσιμη κατανομή τη ζητούμενη.

Στο εξής θα θεωρούμε ως κατανομή-στόχο, την οποία μέχρι τώρα γράφαμε ως $\pi$, την εκ των υστέρων κατανομή $p(\boldsymbol{\theta}|x)$ και θα τη συμβολίζουμε αναλόγως, ώστε να γίνει εμφανής η σύνδεση και η συνδρομή των MCMC με το κομμάτι αυτό της Μπεϋζιανής στατιστικής. Επομένως, η μ.α. του αλγορίθμου θα αποτελείται από τιμές του διανύσματος των παραμέτρων, $\boldsymbol{\theta}$, και θέλουμε να έχει ως στάσιμη κατανομή την ύστερη. Για απλότητα στο συμβολισμό και χωρίς βλάβη της γενικότητας, στην παρούσα ενότητα θα θεωρούμε επίσης μονοδιάστατη την παράμετρο $\theta$.

Διαισθητικά, μπορούμε να πούμε ότι σχηματίζουμε την μ.α. εξερευνώντας τον παραμετρικό χώρο, λαμβάνοντας προτάσεις για τον προορισμό κάθε βήματος από μία κατανομή,

έστω $q$. Η κατανομή αυτή ονομάζεται *κατανομή εισήγησης* (*proposal distribution*) και παράγει τις υποψήφιες καταστάσεις για τη μετάβαση της αλυσίδας, λαμβάνοντας υπόψιν μόνο την τρέχουσα κατάστασή της, γεγονός που δικαιολογεί τον χαρακτηρισμό της ως Μαρκοβιανή. Τα βήματα γίνονται αποδεκτά ή απορρίπτονται σύμφωνα με την *πιθανότητα αποδοχής* (*acceptance probability*) που συμβολίζεται ως $\alpha$. Οι δύο αυτές ποσότητες ορίζουν έναν *πυρήνα μετάβασης* (*transition kernel*), που συμβολίζουμε με $P$, που κυβερνά τις κινήσεις της αλυσίδας, καθορίζει δηλαδή το πού και αν θα μετακινηθεί.

Η αλυσίδα συνεχίζει το ταξίδι της στον παραμετρικό χώρο μέχρις ότου προταθεί, και γίνει αποδεκτή, τιμή για το $\theta$ που έχει παραχθεί από τη στάσιμη κατανομή, δηλαδή την ύστερη. Η περίοδος που ο αλγόριθμος τρέχει χώρις να έχει επιτύχει στασιμότητα, ονομάζεται *περίοδος burn-in* (*burn-in period*). Κατόπιν, σε κάθε βήμα θα παράγεται δείγμα από την ύστερη και ο αλγόριθμος θα έχει εκπληρώσει το σκοπό του.

Μπορούμε να συνοψίσουμε απλά και με πρακτικό τρόπο τον αλγόριθμο Metropolis-Hastings ως εξής:

1. Θέσε αριθμό επανάληψης $t = 0$ και επίλεξε μια αρχική τιμή $\theta^{(0)}$.

2. Προσομοίωσε $\theta^*$ από την κατανομή $q(\theta|\theta^{(t)})$.

3. Αποδέξου το προτεινόμενο $\theta^*$ με πιθανότητα $\alpha(\theta^*|\theta^{(t)})$. Σε περίπτωση αποδοχής θέσε $\theta^{(t+1)} = \theta^*$ ενώ σε περίπτωση απόρριψης $\theta^{(t+1)} = \theta^{(t)}$ (δηλαδή η αλυσίδα δεν κινείται).

4. Αύξησε τον αριθμό επανάληψης από $t$ σε $t + 1$ και επίστρεψε στο βήμα 2 μέχρι να επιτευχθεί σύγκλιση.

Επισημαίνεται πως για την εκτέλεση του βήματος 3, απαιτείται αρχικά η δημιουργία μίας ανεξάρτητης ομοιόμορφης στο $[0, 1]$ ποσότητας $u$. Αν $u \leq \alpha$ η μετάβαση γίνεται αποδεκτή ενώ αν $u > \alpha$ η κίνηση δεν επιτρέπεται.

Με σκοπό να ξετυλίξουμε τη λειτουργία του αλγορίθμου, θα μελετήσουμε πιο διεξοδικά την έννοια και τα συστατικά του πυρήνα μετάβασης. Ειδικότερα, ο πυρήνας μετάβασης ορίζει μια μεικτή κατανομή για τη νέα θέση $\theta^{(t+1)}$ της αλυσίδας (Gamerman D. 2006):

$$P(\theta^{(t+1)}|\theta^{(t)}) = q(\theta^{(t+1)}|\theta^{(t)})\alpha(\theta^*|\theta^{(t)}) + \boldsymbol{I}_{\theta^{(t+1)}=\theta^{(t)}}[1 - \int q(\theta^*|\theta^{(t)})\alpha(\theta^{(t+1)}|\theta^{(t)})d\theta^*]. \quad (18)$$

Η γενική αυτή έκφραση, όπως δόθηκε στην Εξίσωση (18), είναι στην πραγματικότητα το άθροισμα της πυκνότητας πιθανότητας που προκύπτει από την αποδοχή του υποψήφιου $\theta^*$ :

$$P(\theta^{(t+1)}|\theta^{(t)}) = q(\theta^{(t+1)}|\theta^{(t)})\alpha(\theta^{(t+1)}|\theta^{(t)}), \text{ για } \theta^{(t+1)} \neq \theta^{(t)} \quad (19)$$

και της πιθανότητας η αλυσίδα να παραμείνει στο ίδιο σημείο, να έχουμε δηλαδή απορρίψει κάθε πιθανή τιμή υποψήφιου $\theta^*$ :

$$P(\theta^{(t+1)}|\theta^{(t)}) = 1 - \int q(\theta^*|\theta^{(t)})\alpha(\theta^*|\theta^{(t)})d\theta^*, \text{ για } \theta^{(t+1)} = \theta^{(t)}. \quad (20)$$

Περνώντας στην πιθανότητα αποδοχής, στην οποία οφείλεται και η δεξιοτεχνία της μεθόδου, αποδείχθηκε από τον Hastings ότι η σύγκλιση εξασφαλίζεται όταν η πρώτη είναι

ίση με:

$$\alpha(\theta^*|\theta^{(t)}) = \min\left\{1, \frac{\frac{p(\theta^*)}{q(\theta^*|\theta^{(t)})}}{\frac{p(\theta^{(t)})}{q(\theta^{(t)}|\theta^*)}}\right\}, \tag{21}$$

ανεξάρτητα από την επιλογή κατανομής εισήγησης. Αυτό σημαίνει πως μπορούμε να ξεκινήσουμε με μια αυθαίρετη επιλογή για τη μορφή της $q$ και να εμπιστευτούμε τον αλγόριθμο να βρει τον δρόμο του για τη στάσιμη κατανομή μέσα στον παραμετρικό χώρο. Φυσικά, η ταχύτητα σύγκλισης του αλγορίθμου καθορίζεται σε πολύ μεγάλο βαθμό από τη συναρτησιακή μορφή της κατανομής με την οποία τον τροφοδοτούμε, επομένως αν μας ενδιαφέρει να αυξήσουμε την αποδοτικότητα χρειάζεται να μιμηθούμε τη συναρτησιακή μορφή της $p$, τις κορυφές, τις ουρές της κ.ο.κ. Σημειώνεται πως το ζήτημα επιλογής αρχικής τιμής $\theta^{(0)}$ για την παράμετρο ποσώς επηρεάζει την λειτουργία και την ταχύτητα του Metropolis-Hastings.

Το *ποσοστό αποδοχής* (*acceptance rate*) των προτεινόμενων για την παράμετρο τιμών, όπως αυτές γεννιούνται από την $q$, αποτελεί μια σημαντική τιμή για την ταχύτητα σύγκλισης του αλγορίθμου. Ένα μεγάλο ποσοστό σημαίνει ότι η αλυσίδα κάνει μικρά βήματα και έτσι αυξάνεται σημαντικά ο απαιτούμενος χρόνος για την επαρκή εξερεύνηση του χώρου. Αντίθετα, μικρό ποσοστό αποδοχής συνεπάγεται μεγάλα βήματα οπότε η αλυσίδα δεν καλύπτει όλο τον χώρο αφού πολλές από τις υποψήφιες τιμές θα έχουν μικρή πιθανότητα να προταθούν.

Επίσης, άμεση είναι και η σύνδεση μεταξύ του ποσοστού αποδοχής του αλγορίθμου και της διασποράς της κατανομής εισήγησης: πολύ μικρή διασπορά έχει ως αποτέλεσμα την αργή ανάμειξη της αλυσίδας που οδηγεί σε μεγάλα ποσοστά αποδοχής ενώ από την άλλη, μεγάλη διασπορά οδηγεί σε χαμηλό ποσοστό αποδοχής. Η εξάρτηση αυτή του ποσοστού από τη διασπορά της κατανομής εισήγησης αξιοποιείται για την προσαρμογή της (tuning) κατά το διάστημα του burn-in, τροποποιώντας τη διασπορά ώστε να καθοδηγήσει τον αλγόριθμο στο επιθυμητό ποσοστό αποδοχής.

### 2.2.1 Συμμετρικές κατανομές εισήγησης

Ήδη αναφερθήκαμε στην ελευθερία που παρέχεται σχετικά με την επιλογή της $q$. Μία πιθανή κατηγορία κατανομών εισήγησης είναι οι συμμετρικές κατανομές που ορίζουν και συμμετρικές αλυσίδες για τον αλγόριθμο.

**Ορισμός 4** *Μία αλυσίδα είναι συμμετρική (symmetric) αν ο πυρήνας μετάβασής της είναι συμμετρικός για κάθε ζευγάρι καταστάσεων $\theta, \phi$, δηλαδή $P(\theta, \phi) = P(\phi, \theta)$.*

Λόγω της δομής του πυρήνα μετάβασης στον αλγόριθμο M-H, η συμμετρικότητα του πυρήνα ισοδυναμεί με συμμετρικότητα της κατανομής εισήγησης. Θα ισχύει λοιπόν η σχέση $q(\theta|\phi) = q(\phi|\theta)$ για κάθε ζεύγος καταστάσεων $\theta, \phi$. Η πιθανότητα αποδοχής δεν μένει ανεπηρέαστη από τη συμμετρικότητα της κατανομής εισήγησης και έτσι αποκτά μία απλούστερη μορφή, σύμφωνα με την παρακάτω Εξίσωση:

$$\alpha(\theta^*|\theta^{(t)}) = \min\left\{1, \frac{p(\theta^*)}{p(\theta^{(t)})}\right\}. \tag{22}$$

Παρατηρώντας τον λόγο που εμφανίζεται στη συνάρτηση ελαχίστου της Εξίσωσης (21), ο οποίος καλείται και *Hastings' ratio*, διαπιστώνουμε ότι ο αλγόριθμος δέχεται πάντα κινήσεις

που μας οδηγούν σε θέσεις με μεγαλύτερη πυκνότητα πιθανότητας. Αυτό συμβαίνει επειδή σε μια τέτοια περίπτωση το πηλικό θα είναι μεγαλύτερο του $1$ και, λόγω της συνάρτησης ελαχίστου, η πιθανότητα αποδοχής θα γίνεται ακριβώς $1$. Κινήσεις που μας απομακρύνουν από τιμές υψηλής συχνότητας γίνονται επίσης αποδεκτές με πιθανότητα ίση με το πηλίκο της Εξίσωσης (21) αλλά είναι η απόρριψη τέτοιων κινήσεων που κρατά τον αλγόριθμο, στα περισσότερα βήματα, σε περιοχές μεγάλης πυκνότητας της ύστερης (Sherlock, Fearnhead, & G. Roberts 2010). Έτσι, η αλυσίδα σταδιακά μεταβαίνει σε περιοχές με υψηλότερη πυκνότητα, χωρίς αυτό να σημαίνει ότι δεν θα υπάρχουν βήματα στα οποία θα πέφτουμε σε θέσεις μικρότερης πυκνότητας.

Καθ' αυτόν τον τρόπο, είναι πιο πιθανό να επισκεφθούμε τα σημεία που βρίσκονται πιο κοντά στο $\theta^{(t)}$, καθιστώντας το προκύπτον δείγμα αποτέλεσμα ενός τυχαίου περιπάτου (random walk) στον παραμετρικό χώρο. Η ιδιότητα που μόλις περιγράφηκε οφείλεται για την ονομασία αυτής της παραλλαγής του αλγορίθμου Metropolis-Hastings ως *Random Walk Metropolis-Hastings* και αποτελεί μάλιστα την αρχική μορφή του αλγορίθμου Metropolis (βλ. Metropolis, A. Rosenbluth, M. Rosenbluth, H. Teller, & E. Teller 1953). Αναφέρεται ενδεικτικά ότι μία συνήθης επιλογή για την μορφή της $q$ είναι αυτή της κανονικής κατανομής.

Όσον αφορά το ποσοστό αποδοχής, έχει αποδειχθεί πως η βέλτιστη τιμή για την μονοδιάστατη περίπτωση Random Walk M-H είναι περίπου $44\%$ ενώ για την πολυδιάστατη είναι κοντά στο $23\%$ (Gelman, G.O. Roberts, & Gilks 1996).

## 2.3 Η Μέθοδος δειγματοληψίας Gibbs

Η δειγματοληψία Gibbs (Gibbs sampling) αποτελεί έναν ακόμα αλγόριθμο της οικογένειας των MCMC που μπορεί να δώσει λύση στο πρόβλημα υπολογισμού της εκ των υστέρων κατανομής. Αναπτύχθηκε από τους αδελφούς Geman (S. Geman & D. Geman 1984) ως εργαλείο για την επεξεργασία εικόνων (image processing) όπου η κατανομή από την οποία θέλουμε να λάβουμε δείγμα είναι η κατανομή Gibbs. Λίγα χρόνια αργότερα, προτάθηκε η χρήση του αλγορίθμου και για άλλες μορφές ύστερων κατανομών, πέρα από την Gibbs η οποία είχε ήδη χαρίσει το όνομά της στη μέθοδο.

Αναφορικά με τη λειτουργία της μεθόδου, δεν διαφέρει πολύ από όσα περιγράφηκαν παραπάνω για τον αλγόριθμο Metropolis-Hastings και μάλιστα θεωρείται υποπερίπτωση αυτού. Η βασική ιδέα παραμένει ίδια: σχηματίζουμε μια Μαρκοβιανή αλυσίδα με στάσιμη κατανομή την ζητούμενη, αφήνουμε την αλυσίδα να τρέξει μέχρι να επέλθει στασιμότητα και από αυτό το σημείο και μετά μπορούμε να αρχίσουμε να συλλέγουμε τις μεταβλητές των καταστάσεων της μ.α. ως δείγμα.

Η ειδοποιός διαφορά μεταξύ των υλοποιήσεων των δύο αλγορίθμων αφορά την κατανομή εισήγησης καθώς στην περίπτωση του Gibbs επιλέγουμε να λαμβάνουμε τις υποψήφιες νέες θέσεις της αλυσίδας από την *πλήρους δέσμευσης εκ των υστέρων κατανομή* (*full conditional probability distribution*), την οποία θα αποσαφηνίσουμε ακολούθως. Τείνουμε λοιπόν να επιστρατεύουμε τη δειγματοληψία Gibbs όταν η πλήρους δέσμευσης ύστερη είναι ευκολότερα υπολογίσιμη, όπως συμβαίνει σε αρκετές περιστάσεις. Ακόμα, σε αυτόν τον αλγόριθμο η πιθανότητα αποδοχής είναι ίση με $1$ που σημαίνει ότι αποδεχόμαστε όποια κίνηση μας "κληρώσει" η κατανομή εισήγησης.

Γενικά, για ένα διάνυσμα $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_d)'$ ορίζουμε ως κατανομή πλήρους δέσμευσης την δεσμευμένη κατανομή του $\theta_i$ δεδομένων όλων των υπόλοιπων τιμών $\theta_j, j \neq i$ και

γράφουμε:

$$g_i(\theta_i) = g(\theta_i|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_d) = g(\theta_i|\boldsymbol{\theta}_{-i}). \tag{23}$$

Επομένως, η χρήση της πλήρους δέσμευσης εκ των υστέρων κατανομής ως κατανομή εισήγησης $q$, συνεπάγεται την ακόλουθη Εξίσωση:

$$q_i(\theta_i^*|\theta_i^{(t)}, \boldsymbol{\theta}_{-i}^{(t)}) = p(\theta_i^*|\boldsymbol{\theta}_{-i}^{(t)}) \tag{24}$$

όπου συμβολίζουμε με $\theta_i^{(t)}$ την τιμή της i-οστής συνιστώσας του διανύσματος παραμέτρων τη χρονική στιγμή $t$ και με $\boldsymbol{\theta}_{-i}^{(t)}$ τις τιμές του διανύσματος των παραμέτρων χωρίς το $\theta_i$ τη χρονική στιγμή $t$ ενώ $p$ όπως πάντα η κατανομή-στόχος και $\theta_i^*$ το υποψήφιο σημείο.

Όταν μπορούμε να απομονώσουμε τις κατανομές πλήρους δέσμευσης, υπό την έννοια ότι είναι απολύτως γνωστές και έχουμε τη δυνατότητα να λάβουμε δείγμα από αυτές, η δειγματοληψία Gibbs προσφέρει μία υπολογιστικά οικονομική εναλλακτική για την παραγωγή των τιμών της παραμέτρου. Ειδικότερα, το κάθε βήμα της μεθόδου περιλαμβάνει την ανανέωση της τιμής κάθε συνιστώσας του $\boldsymbol{\theta}$ χρησιμοποιώντας ως γεννήτορα την πλήρους δέσμευσης εκ των υστέρων κατανομή για τη διαδοχική δημιουργία τιμών. Επειδή η πιθανότητα μετάβασης είναι ίση με 1, ο πυρήνας μετάβασης είναι πιο απλοϊκός σε σχέση με αυτόν του M-H και αποτελείται αποκλειστικά από την κατανομή εισήγησης $q$. Η δέσμευση της $q_i$ γίνεται ως προς τις τιμές $\theta_i$ που έχουν προλάβει να ανανεωθούν στο τρέχον βήμα και ως προς εκείνες τις συνιστώσες που θα ανανεωθούν ακολούθως. Βάσει αυτής της τακτικής, η τελευταία συνιστώσα που ανανεώνεται στο εκάστοτε βήμα θα λάβει την τιμή της από την πλήρους δέσμευσης ύστερη, με δεδομένη την προηγούμενη τιμή της ίδιας της συνιστώσας και τις ανανεωμένες τιμές όλων των υπολοίπων.

Συνοπτικά, και κατ' αναλογία με τον Metropolis-Hastings, ο αλγόριθμος της δειγματοληψίας Gibbs έχει ως εξής:

1. Θέσε αριθμό επανάληψης $t = 0$ και επίλεξε μια αρχική τιμή $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, ..., \theta_d^{(0)})$.

2. Χτίσε το νέο διάνυσμα $\boldsymbol{\theta}^{(t+1)} = (\theta_1^{(t+1)}, ..., \theta_d^{(t+1)})$ λαμβάνοντας διαδοχικά τις τιμές

$$\begin{aligned}
\theta_1^{(t+1)} &\sim p(\theta_1|\theta_2^{(t)}, ..., \theta_d^{(t)}) \\
\theta_2^{(t+1)} &\sim p(\theta_2|\theta_1^{(t+1)}, \theta_3^{(t)}, ..., \theta_d^{(t)}) \\
&\vdots \\
\theta_d^{(t+1)} &\sim p(\theta_d|\theta_1^{(t+1)}, ..., \theta_{d-1}^{(t+1)})
\end{aligned}$$

3. Αύξησε τον αριθμό επανάληψης από $t$ σε $t + 1$ και επίστρεψε στο βήμα 2 μέχρι να επιτευχθεί σύγκλιση.

# 3 Πολλαπλά Κανονικά Μοντέλα Γραμμικής Παλινδρόμησης στη Μπεϋζιανή στατιστική

Όταν υπάρχουν ενδείξεις ή υποψίες ότι η τιμή μίας μεταβλητής επηρεάζεται γραμμικά από τις τιμές άλλων μεταβλητών, καταφεύγουμε στην προσαρμογή πολλαπλών μοντέλων γραμμικής παλινδρόμησης (Καρώνη & Οικονόμου 2017). Η προς μελέτη μεταβλητή καλείται μεταβλητή απόκρισης ενώ οι μεταβλητές που υποψιαζόμαστε ότι την επηρεάζουν καλούνται επεξηγηματικές. Η γενική μορφή ενός τέτοιου μοντέλου έχει ως εξής:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{25}$$

Στην παραπάνω σχέση, ο $X$ ονομάζεται πίνακας σχεδιασμού και είναι ένας $n \times p$ πίνακας του οποίου η πρώτη στήλη αποτελείται από μονάδες ενώ οι υπόλοιπες $k = p - 1$ στήλες του αντιστοιχούν στις $n$ παρατηρήσεις κάθε επεξηγηματικής μεταβλητής. Με $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)'$ συμβολίζουμε το $p \times 1$ διάνυσμα των παραμέτρων και με $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)'$ το $n \times 1$ διάνυσμα των τυχαίων σφαλμάτων.

Ο χαρακτηρισμός ενός πολλαπλού γραμμικού μοντέλου παλινδρόμησης ως κανονικό απαιτεί την υπόθεση κανονικότητας του διανύσματος $\boldsymbol{\epsilon}$ των τυχαίων σφαλμάτων. Για την ακρίβεια, υποθέτουμε ότι η κατανομή του $\boldsymbol{\epsilon}$ είναι η $n-$διάστατη κανονική με μέση τιμή μηδέν, δηλαδή:

$$\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n). \tag{26}$$

Στα πλαίσια της ανάλυσης παλινδρόμησης, ο προσδιορισμός του διανύσματος $\boldsymbol{\beta}$ θεωρείται μείζονος σημασίας καθώς αποκαλύπτει την ακριβή γραμμική σχέση μεταξύ του $\boldsymbol{y}$ και των $x_i$, $i = 1, ..., k$, δηλαδή της μεταβλητής απόκρισης και των τιμών των επεξηγηματικών μεταβλητών. Στην κλασική στατιστική, η εκτίμηση των παραμέτρων γίνεται συνήθως με την μέθοδο ελαχίστων τετραγώνων, η οποία παρέχει τύπο για τις εκτιμήτριες σε κλειστή μορφή.

Εφόσον το διάνυσμα $\boldsymbol{\beta}$ συνίσταται από τις παραμέτρους του μοντέλου, μπορεί κάλλιστα να εκτιμηθεί με τη χρήση Μπεϋζιανής μεθοδολογίας, αν θεωρήσουμε ότι οι συνιστώσες $\beta_i$, καθώς και η διασπορά $\sigma^2$, είναι τυχαίες μεταβλητές (Fahrmeir, Kneib, Lang, & Marx 2013). Όπως ακριβώς γίνεται και στην περίπτωση εκτίμησης παραμέτρων κατανομών, πρώτα υιοθετούμε μία, από κοινού, κατανομή για τα $\boldsymbol{\beta}$, $\sigma^2$, αντιπροσωπευτική της πρότερης γνώσης που διαθέτουμε. Κατόπιν, σύμφωνα με τον γενικευμένο νόμο του Bayes, πολλαπλασιάζοντας με την πιθανοφάνεια των δεδομένων, εν προκειμένω των τιμών της μεταβλητής απόκρισης, καταλήγουμε στη μορφή της ύστερης κατανομής των παραμέτρων. Τέλος, διαιρώντας με την περιθώρια κατανομή των δεδομένων, εφόσον αυτό είναι εφικτό, αποκτούμε τον αναλυτικό τύπο της από κοινού εκ των υστέρων κατανομής των παραμέτρων.

Ο γενικευμένος νόμος του Bayes για τη διαδικασία ανανέωσης της πρότερης γνώσης στα πολλαπλά κανονικά γραμμικά μοντέλα διαμορφώνεται ως εξής:

$$p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) = \frac{p(\boldsymbol{\beta}, \sigma^2)p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)}{p(\boldsymbol{y})}. \tag{27}$$

Λαμβάνοντας υπόψη τη μορφή του μοντέλου στην Εξίσωση (25), την υπόθεση της κανονικότητας των σφαλμάτων όπως ορίζεται στην Εξίσωση (26) καταλήγουμε στην ακόλουθη κατανομή για τις τιμές $\boldsymbol{y}$ της μεταβλητής απόκρισης:

$$\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2 \sim N_n(X\boldsymbol{\beta}, \sigma^2 I_n). \tag{28}$$

Θεωρούμε λοιπόν $n-$διάστατη κανονική κατανομή για την απόκριση, δεσμευμένη ως προς τις παραμέτρους $\boldsymbol{\beta}$ και $\sigma^2$, δηλαδή κανονική πιθανοφάνεια για τα δεδομένα. Το αποτέλεσμα αυτό είναι ιδιαίτερα χρήσιμο καθώς η κανονική πιθανοφάνεια υποβοηθά τη δημιουργία συζυγών ζευγών πρότερων και ύστερων κατανομών. Όπως είδαμε και στο Κεφάλαιο 1, οι συζυγείς πρότερες διευκολύνουν την εξαγωγή της ύστερης κατανομής, αποφεύγοντας πολύπλοκους, και ενίοτε αδύνατους, υπολογισμούς.

Στη συνέχεια θα δούμε αναλυτικά διάφορες δημοφιλείς επιλογές πρότερων για τα $\boldsymbol{\beta}$ και $\sigma^2$ στα πολλαπλά κανονικά γραμμικά μοντέλα παλινδρόμησης.

## 3.1 Κανονική-Αντίστροφη Γάμμα από κοινού πρότερη

Η συνήθης επιλογή για την πρότερη κατανομή των συντελεστών $\boldsymbol{\beta}$ της παλινδρόμησης, δοθέντος του $\sigma^2$, είναι αυτή της πολυμεταβλητής κανονικής με μέση τιμή $\boldsymbol{m}$ και πίνακα συνδιασποράς $\sigma^2 V$, δηλαδή:

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2 &\sim N_p(\boldsymbol{m}, \sigma^2 V) \Leftrightarrow \\ p\left(\boldsymbol{\beta} \mid \sigma^2\right) &= \frac{1}{(2\pi)^{\frac{p}{2}} \left(\sigma^2\right)^{\frac{p}{2}} |V|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{m})' V^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right\}. \end{aligned} \tag{29}$$

Μια τέτοια επιλογή μπορεί να χαρακτηριστεί φυσική καθώς στο κλασικά μοντέλα παλινδρόμησης η κατανομή των συντελεστών θεωρείται προσεγγιστικά πολυδιάστατη κανονική (Fahrmeir, Kneib, Lang, & Marx 2013).

Όσον αφορά την κατανομή του $\sigma^2$, επιλέγουμε την αντίστροφη γάμμα κατανομή [inverse gamma (IG) distribution] με παράμετρο σχήματος (shape parameter) $a$ και παράμετρο κλίμακας (scale parameter) $b$ της οποίας η συνάρτηση πυκνότητας πιθανότητας δίνεται ακολούθως:

$$\begin{aligned} \sigma^2 &\sim IG(a, b) \Leftrightarrow \\ p\left(\sigma^2\right) &= \frac{b^a}{\Gamma(a)} \frac{1}{\left(\sigma^2\right)^{a+1}} \exp\left\{-\frac{b}{\sigma^2}\right\}. \end{aligned} \tag{30}$$

Τα $a$, $b$ ονομάζονται υπερπαράμετροι (hyperparameters) καθώς αποτελούν παραμέτρους της πρότερης κατανομής του $\sigma^2$ και όχι του ίδιου του μοντέλου.

Δοθέντων των παραπάνω κατανομών, εξάγουμε την από κοινού κατανομή:

$$\begin{aligned} p\left(\boldsymbol{\beta}, \sigma^2\right) &= p\left(\boldsymbol{\beta} \mid \sigma^2\right) \times p\left(\sigma^2\right) \\ &= \frac{1}{(2\pi)^{\frac{p}{2}} \left(\sigma^2\right)^{\frac{p}{2}} |V|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{m})' V^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right\} \\ &\times \frac{b^a}{\Gamma(a)} \frac{1}{\left(\sigma^2\right)^{a+1}} \exp\left\{-\frac{b}{\sigma^2}\right\} \\ &= \frac{b^a}{\Gamma(a)} \frac{1}{(2\pi)^{\frac{p}{2}} \left(\sigma^2\right)^{a+1+\frac{p}{2}} |V|^{\frac{1}{2}}} \exp\left\{-\frac{2b + (\boldsymbol{\beta}-\boldsymbol{m})' V^{-1}(\boldsymbol{\beta}-\boldsymbol{m})}{2\sigma^2}\right\}. \end{aligned}$$

Από την τελευταία ισότητα καταλήγουμε ότι η από κοινού κατανομή είναι η κανονική-αντίστροφη γάμμα [normal-inverse gamma (NIG)] με παραμέτρους αυτές των περιθώριων κατανομών, δηλαδή:

$$\boldsymbol{\beta}, \sigma^2 \sim NIG(\boldsymbol{m}, V, a, b). \tag{31}$$

**Εκ των υστέρων κατανομή**

Η Εξίσωση (31) διευκολύνει την ανάλυση καθώς η κανονική-αντίστροφη γάμμα κατανομή είναι συζυγής της κανονικής πιθανοφάνειας, από την οποία χαρακτηρίζεται το μοντέλο. Λόγω του συζυγούς χαρακτήρα της πρότερης, απευθείας συμπεραίνουμε ότι η ύστερη κατανομή θα είναι επίσης κανονική-αντίστροφη γάμμα. Με χρήση του θεωρήματος του Bayes, υπολογίζεται η μορφή της ύστερης, ενόψει των δεδομένων $\boldsymbol{y}$, ως ποσότητα ανάλογη του γινομένου της πιθανοφάνειας με την από κοινού κατανομή των $\boldsymbol{\beta}, \sigma^2$:

$$\begin{aligned} p\left(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}\right) \propto & L\left(\boldsymbol{\beta}, \sigma^2\right) \times p\left(\boldsymbol{\beta} \mid \sigma^2\right) \times p\left(\sigma^2\right) \\ \propto & \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-X\boldsymbol{\beta})'(\boldsymbol{y}-X\boldsymbol{\beta})\right\} \\ & \times \frac{1}{\left(\sigma^2\right)^{\frac{p}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{m})'V^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right\} \\ & \times \frac{1}{\left(\sigma^2\right)^{a+1}} \exp\left\{-\frac{b}{\sigma^2}\right\}. \end{aligned}$$

Με σκοπό να μετασχηματίσουμε την έκφραση για την ύστερη κατανομή σε μια πιο εύχρηστη μορφή, συγκεντρώνουμε τους εκθετικούς όρους σε έναν όπως φαίνεται ακολούθως:

$$\exp\left\{-\frac{1}{2\sigma^2}\left[2b + (\boldsymbol{y}-X\boldsymbol{\beta})'(\boldsymbol{y}-X\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{m})'V^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right]\right\}.$$

Κατόπιν, ορίζοντας τις ποσότητες:

$$\tilde{V} = \left(X'X + V^{-1}\right)^{-1} \tag{32}$$

$$\tilde{\boldsymbol{m}} = \tilde{V}\left(V^{-1}\boldsymbol{m} + X'\boldsymbol{y}\right) \tag{33}$$

επεξεργαζόμαστε το εσωτερικό του εκθετικού ως εξής:

$$\begin{aligned} & (\boldsymbol{y}-X\boldsymbol{\beta})'(\boldsymbol{y}-X\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{m})'V^{-1}(\boldsymbol{\beta}-\boldsymbol{m}) \\ =& \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'X'\boldsymbol{y} + \boldsymbol{\beta}'X'X\boldsymbol{\beta} + \boldsymbol{\beta}'V^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'V^{-1}\boldsymbol{m} + \boldsymbol{m}'V^{-1}\boldsymbol{m} \\ =& \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\beta}'\left(X'X + V^{-1}\right)\boldsymbol{\beta} - 2\boldsymbol{\beta}'\left(X'\boldsymbol{y} + V^{-1}\boldsymbol{m}\right) + \boldsymbol{m}'V^{-1}\boldsymbol{m} \\ =& \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\beta}'\tilde{V}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\tilde{V}^{-1}\tilde{V}\left(X'\boldsymbol{y} + V^{-1}\boldsymbol{m}\right) + \boldsymbol{m}'V^{-1}\boldsymbol{m} \\ =& \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{\beta}'\tilde{V}^{-1}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\tilde{V}^{-1}\tilde{\boldsymbol{m}} + \boldsymbol{m}'V^{-1}\boldsymbol{m} \\ =& \boldsymbol{y}'\boldsymbol{y} + (\boldsymbol{\beta}-\tilde{\boldsymbol{m}})'\tilde{V}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}}) - \tilde{\boldsymbol{m}}'\tilde{V}^{-1}\tilde{\boldsymbol{m}} + \boldsymbol{m}'V^{-1}\boldsymbol{m}. \end{aligned}$$

Προκύπτει συνεπώς η παρακάτω μορφή της ύστερης:

$$\begin{aligned} p\left(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}\right) \propto & \frac{1}{\left(\sigma^2\right)^{\frac{n+p}{2}+a+1}} \\ & \times \exp\left\{-\frac{1}{2\sigma^2}\left[2b + \boldsymbol{y}'\boldsymbol{y} - \tilde{\boldsymbol{m}}'\tilde{V}^{-1}\tilde{\boldsymbol{m}} + \boldsymbol{m}'V^{-1}\boldsymbol{m} + (\boldsymbol{\beta}-\tilde{\boldsymbol{m}})'\tilde{V}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})\right]\right\}. \end{aligned}$$

Μένει τώρα να θέσουμε τα παρακάτω:

$$\tilde{a} = a + \frac{n}{2} \tag{34}$$

$$\tilde{b} = b + \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} - \tilde{\boldsymbol{m}}'\tilde{V}^{-1}\tilde{\boldsymbol{m}} + \boldsymbol{m}'V^{-1}\boldsymbol{m}\right) \tag{35}$$

και λαμβάνουμε τελικά την μορφή της ύστερης, η οποία αντιστοιχεί σε κανονική-αντίστροφη γάμμα κατανομή με παραμέτρους $\tilde{\boldsymbol{m}}, \tilde{V}, \tilde{a}, \tilde{b}$, όπως αυτές ορίστηκαν στις σχέσεις (32)-(35):

$$p\left(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}\right) \propto \frac{1}{(\sigma^2)^{\tilde{a}+1+\frac{p}{2}}} \exp\left\{-\frac{1}{\sigma^2}\left[\tilde{b} + \frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})'\tilde{V}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})\right]\right\} \Leftrightarrow$$

$$\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y} \sim NIG\left(\tilde{\boldsymbol{m}}, \tilde{V}, \tilde{a}, \tilde{b}\right). \tag{36}$$

Θα κατασκευάσουμε τώρα μία ισοδύναμη έκφραση για το $\tilde{b}$ της οποίας η χρησιμότητα θα φανεί στους υπολογισμούς που θα ακολουθήσουν. Ξεκινάμε λοιπόν από την Εξίσωση (35), αντικαθιστούμε την έκφραση για το $\tilde{\boldsymbol{m}}$ από την Εξίσωση (33) και εκτελούμε απλές αλγεβρικές πράξεις:

$$\begin{aligned}
\tilde{b} &= b + \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} - \tilde{\boldsymbol{m}}'\tilde{V}^{-1}\tilde{\boldsymbol{m}} + \boldsymbol{m}'V^{-1}\boldsymbol{m}\right) \\
&= b + \frac{1}{2}\left[\boldsymbol{y}'\boldsymbol{y} - \left(V^{-1}\boldsymbol{m} + X'\boldsymbol{y}\right)'\tilde{V}\tilde{V}^{-1}\tilde{V}\left(V^{-1}\boldsymbol{m} + X'\boldsymbol{y}\right) + \boldsymbol{m}'V^{-1}\boldsymbol{m}\right)] \\
&= b + \frac{1}{2}\left[\boldsymbol{y}'\boldsymbol{y} - \left(V^{-1}\boldsymbol{m} + X'\boldsymbol{y}\right)'\tilde{V}\left(V^{-1}\boldsymbol{m} + X'\boldsymbol{y}\right) + \boldsymbol{m}'V^{-1}\boldsymbol{m}\right] \\
&= b + \frac{1}{2}\left\{\boldsymbol{y}'\boldsymbol{y} - \left[\left(V^{-1}\boldsymbol{m}\right)'\tilde{V}V^{-1}\boldsymbol{m} + 2\left(X'\boldsymbol{y}\right)'\tilde{V}V^{-1}\boldsymbol{m} + \left(X'\boldsymbol{y}\right)'\tilde{V}\left(X'\boldsymbol{y}\right)\right] + \boldsymbol{m}'V^{-1}\boldsymbol{m}\right\} \\
&= b + \frac{1}{2}\left[\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{m}'V^{-1}\tilde{V}V^{-1}\boldsymbol{m} - 2\boldsymbol{y}'X\tilde{V}V^{-1}\boldsymbol{m} - \boldsymbol{y}'X\tilde{V}X'\boldsymbol{y} + \boldsymbol{m}'V^{-1}\boldsymbol{m}\right] \\
&= b + \frac{1}{2}\left[\boldsymbol{y}'\left(I_n - X\tilde{V}X'\right)\boldsymbol{y} - 2\boldsymbol{y}'X\tilde{V}V^{-1}\boldsymbol{m} + \boldsymbol{m}'\left(V^{-1} - V^{-1}\tilde{V}V^{-1}\right)\boldsymbol{m}\right]
\end{aligned} \tag{37}$$

Επεξεργαζόμαστε ξεχωριστά τους δύο τελευταίους όρους της αγκύλης ώστε να εμφανίσουμε και σε αυτούς την ποσότητα $\left(I_n - X\tilde{V}X'\right)$. Συγκεκριμένα, για τον δεύτερο όρο, παρατηρούμε το εξής:

$$\begin{aligned}
&\tilde{V}\tilde{V}^{-1} = I_p \\
\Leftrightarrow &\tilde{V}\left(V^{-1} + X'X\right) = I_p \\
\Leftrightarrow\ &\tilde{V}V^{-1} + \tilde{V}X'X = I_p \\
\Leftrightarrow\ &\tilde{V}V^{-1} = I_p - \tilde{V}X'X \\
\Leftrightarrow\ &X\tilde{V}V^{-1} = X - X\tilde{V}X'X \\
\Leftrightarrow\ &X\tilde{V}V^{-1} = \left(I_n - X\tilde{V}X'\right)X.
\end{aligned} \tag{38}$$

Για τον μετασχηματισμό του τρίτου όρου θα χρειαστούμε την ταυτότητα πινάκων Sherman-Woodbury-Morrison για τετραγωνικούς αντιστρέψιμους πίνακες $A, D$ και ορθογώνιους πίνακες $B, C$:

$$(A + BDC)^{-1} = A^{-1} - A^{-1}B\left(D^{-1} + CA^{-1}B\right)^{-1}CA^{-1}.$$

Αφού αντικαταστήσουμε την αναλυτική έκφραση για το $\tilde{V}$ στον δεύτερο όρο, εφαρμόζουμε την ταυτότητα για $A = V, D = (X'X)^{-1}$ και $B = C = I_p$ και προκύπτει:

$$\begin{aligned}\boldsymbol{m}' \left(V^{-1} - V^{-1}\tilde{V}V^{-1}\right) \boldsymbol{m} &= \boldsymbol{m}' \left(V^{-1} - V^{-1}\left(V^{-1} + X'X\right)V^{-1}\right) \boldsymbol{m} \\ &= \boldsymbol{m}' \left[V + (X'X)^{-1}\right]^{-1} \boldsymbol{m}.\end{aligned}$$

Εφαρμόζουμε για δεύτερη φορά την ταυτότητα, επιλέγοντας τώρα $A = (X'X)^{-1}, D = V$ και $B, C$ όπως πριν, οπότε έχουμε:

$$\begin{aligned}\boldsymbol{m}' \left[V + (X'X)^{-1}\right]^{-1} \boldsymbol{m} &= \boldsymbol{m}' \left[X'X - X'X\left(X'X + V^{-1}\right)^{-1} X'X\right] \boldsymbol{m} \\ &= \boldsymbol{m}' \left[X'\left(I_n - X\tilde{V}X'\right)X\right] \boldsymbol{m}. \qquad (39)\end{aligned}$$

Χρησιμοποιούμε άλλη μία φορά την ταυτότητα πινάκων που παρουσιάσαμε παραπάνω με $A = I_n, B = C = I_p$ και $D = XVX'$ και έχουμε:

$$(I_n + XVX')^{-1} = I_n - X\left(V^{-1} + X'X\right)^{-1} X' = I_n - X\tilde{V}X' \qquad (40)$$

Αντικαθιστώντας τις Εξισώσεις (38), (39) και (40) στην (37), λαμβάνουμε:

$$\tilde{b} = b + \frac{1}{2}\left[\boldsymbol{y}' \left(I_n - XVX'\right)^{-1} \boldsymbol{y} - 2\boldsymbol{y}' \left(I_n - XVX'\right)^{-1} X\boldsymbol{m} + \boldsymbol{m}'X' \left(I_n - XVX'\right)^{-1} X\boldsymbol{m}\right].$$

Απομονώνοντας τον κοινό παράγοντα καταλήγουμε στην ακόλουθη μορφή η οποία όπως προαναφέρθηκε θα χρησιμοποιηθεί σε υπολογισμούς μετέπειτα:

$$\tilde{b} = b + \frac{1}{2}\left[(\boldsymbol{y} - X\boldsymbol{m})' \left(I_n - XVX'\right)^{-1} (\boldsymbol{y} - X\boldsymbol{m})\right]. \qquad (41)$$

**Περιθώριες ύστερες κατανομές**

Είναι σύνηθες σε πολλά προβλήματα να εμπλέκονται περισσότερες από μία παράμετροι οι οποίες είτε είναι άγνωστες, είτε δεν μπορούν να παρατηρηθούν, είτε δεν αποτελούν σημείο ενδιαφέροντος στην εκάστοτε ανάλυση. Όταν συμβαίνει αυτό, είναι χρήσιμο να μπορούν να απομονωθούν οι ποσότητες που πραγματικά μας απασχολούν και να εξατομικευτεί αναλόγως η μελέτη.

Στη Μπεϋζιανή στατιστική η δυνατότητα αυτή δίνεται μέσω των περιθώριων κατανομών που προκύπτουν από απλή, συνήθως, ολοκλήρωση της από κοινού κατανομής. Για να εξάγουμε τις περιθώριες ύστερες κατανομές των παραμέτρων $\boldsymbol{\beta}$ και $\sigma^2$, αρκεί να ολοκληρώσουμε την Εξίσωση (36)) ως προς την άλλη παράμετρο.

Αναλυτικότερα, η περιθώρια ύστερη κατανομή του διανύσματος $\boldsymbol{\beta}$ των συντελεστών μπορεί να υπολογιστεί από το εξής ολοκλήρωμα:

$$p\left(\boldsymbol{\beta} \mid \boldsymbol{y}\right) = \int p\left(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}\right) d\sigma^2$$

$$
\begin{aligned}
&\propto \int \frac{1}{(\sigma^2)^{\tilde{a}+1+\frac{p}{2}}} \exp\left\{-\frac{1}{\sigma^2}\left[\tilde{b}+\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})'\tilde{V}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})\right]\right\} d\sigma^2 \\
&\propto \int IG\left(\tilde{a}+\frac{p}{2}, \tilde{b}+\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})'\tilde{V}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})\right) d\sigma^2 \\
&\propto \left[\tilde{b}+\frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})'\tilde{V}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})\right]^{-(\tilde{a}+p/2)} \\
&\propto \left[1+\frac{(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})'\tilde{V}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})}{2\tilde{b}}\right]^{-(\tilde{a}+p/2)} \\
&\propto \left[1+\frac{\tilde{a}}{\tilde{a}}\frac{(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})'\tilde{V}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})}{2\tilde{b}}\right]^{-(2\tilde{a}+p)/2} \\
&\propto \left[1+\frac{(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})'\tilde{a}\tilde{V}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})}{2\tilde{a}\tilde{b}}\right]^{-(2\tilde{a}+p)/2}.
\end{aligned}
$$

Ορίζοντας τώρα τις ποσότητες:

$$
\begin{aligned}
\Sigma &= \frac{\tilde{b}}{\tilde{a}}\tilde{V} \\
\nu &= 2\tilde{a},
\end{aligned}
$$

η περιθώρια ύστερη σ.π.π. γίνεται:

$$
p\left(\boldsymbol{\beta} \mid \boldsymbol{y}\right) \propto \left[1+\frac{(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})'\Sigma^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})}{\nu}\right]^{-(\nu+p)/2}.
$$

Η τελευταία έκφραση μας παραπέμπει στην συνάρτηση πυκνότητας πιθανότητας της πολυδιάστατης κατανομής Student, με παράμετρο θέσης $\tilde{\boldsymbol{m}}$, παράμετρο κλίμακας $\Sigma = \frac{\tilde{b}}{\tilde{a}}\tilde{V}$ και $\nu = 2\tilde{a}$ βαθμούς ελευθερίας. Κατά συνέπεια, η πλήρης έκφραση της σ.π.π. της ύστερης κατανομής του διανύσματος των συντελεστών θα είναι η ακόλουθη:

$$
\begin{aligned}
&\boldsymbol{\beta} \mid \boldsymbol{y} \sim MVSt_{\nu}\left(\tilde{m}, \Sigma\right) \Leftrightarrow \\
&p\left(\boldsymbol{\beta} \mid \boldsymbol{y}\right) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\pi^{\frac{p}{2}}|\nu\Sigma|^{\frac{1}{2}}}\left(1+\frac{(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})'\Sigma^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})}{\nu}\right)^{-\frac{\nu+p}{2}}. \qquad (42)
\end{aligned}
$$

Αντίστοιχα, για την περιθώρια ύστερη κατανομή του $\sigma^2$, ολοκληρώνουμε την από κοινού ύστερη ως προς $\boldsymbol{\beta}$:

$$
\begin{aligned}
p\left(\sigma^2 \mid \boldsymbol{y}\right) &= \int p\left(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}\right) d\boldsymbol{\beta} \\
&= \int \frac{1}{(\sigma^2)^{\tilde{a}+1+\frac{p}{2}}} \exp\left\{-\frac{1}{\sigma^2}\left[\tilde{b}+(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})'\tilde{V}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})\right]\right\} d\boldsymbol{\beta} \\
&\propto \frac{1}{(\sigma^2)^{\tilde{a}+1+\frac{p}{2}}} \exp\left(-\frac{\tilde{b}}{\sigma^2}\right).
\end{aligned}
$$

Όπως έχουμε ήδη δει, η παραπάνω σ.π.π. αντιστοιχεί στην αντίστροφη γάμμα κατανομή με παραμέτρους $\tilde{a}, \tilde{b}$. Άρα, για την σ.π.π. της εκ των υστέρων κατανομής της διασποράς $\sigma^2$, θα ισχύει:

$$\begin{aligned} \sigma^2 \mid \boldsymbol{y} &\sim IG\left(\tilde{a}, \tilde{b}\right) \Leftrightarrow \\ p\left(\sigma^2 \mid \boldsymbol{y}\right) &= \frac{\tilde{a}^{\tilde{b}}}{\Gamma(\tilde{a})} \frac{1}{(\sigma^2)^{\tilde{a}+1+\frac{p}{2}}} \exp\left(-\frac{\tilde{b}}{\sigma^2}\right). \end{aligned} \tag{43}$$

**Περιθώρια πιθανοφάνεια**

Μία αναδρομή στα εισαγωγικά στοιχεία της Μπεϋζιανής στατιστικής, όπως αυτά περιγράφηκαν στην παρούσα εργασία, φανερώνει την εκτενή αναφορά στα προβλήματα μη υπολογισιμότητας που ουκ ολίγες φορές δημιουργεί η παρουσία του παρονομαστή $p(\boldsymbol{y})$. Η ποσότητα αυτή ονομάσαμε περιθώρια πιθανοφάνεια και, στα πλαίσια εκφράσεων όπως η (27), αποτελεί τον παράγοντα κανονικοποίησης που ολοκληρώνει την εικόνα μας για την εκ των υστέρων κατανομή.

Παρόλα αυτά, χάρη στην ανάλυση που πραγματοποιήθηκε, δίνεται πλέον η δυνατότητα υπολογισμού και της περιθώριας πιθανοφάνειας, τον οποίο είχαμε αρχικά παρακάμψει. Πιο αναλυτικά, παρατηρούμε ότι:

$$\begin{aligned} p(\boldsymbol{y}) &= \int\int p\left(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2\right) p\left(\boldsymbol{\beta}, \sigma^2\right) d\boldsymbol{\beta} d\sigma^2 \\ &= \int p\left(\boldsymbol{y} \mid \sigma^2\right) p\left(\sigma^2\right) d\sigma^2. \end{aligned} \tag{44}$$

Επομένως αρκεί να προσδιορίσουμε με κάποιο τρόπο την περιθώρια κατανομή $p\left(\boldsymbol{y} \mid \sigma^2\right)$ ώστε να υπολογιστεί η περιθώρια πιθανοφάνεια. Αυτό μπορεί να επιτευχθεί ολοκληρώνοντας το γινόμενο της πιθανοφάνειας με την περιθώρια κατανομή του $\boldsymbol{\beta}|\sigma^2$ ως προς $\boldsymbol{\beta}$, όπως και γίνεται ακολούθως:

$$\begin{aligned} p\left(\boldsymbol{y} \mid \sigma^2\right) &= \int L(\boldsymbol{\beta}, \sigma^2) p\left(\boldsymbol{\beta} \mid \sigma^2\right) d\boldsymbol{\beta} \\ &= \int p\left(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2\right) p\left(\boldsymbol{\beta} \mid \sigma^2\right) d\boldsymbol{\beta} \\ &= \int N_n\left(X\boldsymbol{\beta}, \sigma^2 I_n\right) \times N\left(\boldsymbol{m}, \sigma^2 V\right) d\boldsymbol{\beta} \\ &= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n+p}{2}} |V|^{1/2}} \\ &\times \int \exp\left[-\frac{1}{2\sigma^2}\left\{(\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{m})' V^{-1} (\boldsymbol{\beta} - \boldsymbol{m})\right\}\right] d\boldsymbol{\beta}. \end{aligned}$$

Για να προχωρήσουμε, χρειάζεται να μετασχηματίσουμε το εσωτερικό του εκθετικού όπως κάναμε και κατά τον σχηματισμό της από κοινού ύστερης κατανομής. Αντικαθιστώντας ταυτόχρονα την ισοδύναμη έκφραση για το $\tilde{b}$ από την Εξίσωση (41), χωρίς τον όρο $b$,

λαμβάνουμε:

$$
\begin{aligned}
p\left(\boldsymbol{y} \mid \sigma^2\right) &= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n+p}{2}}\left|V\right|^{1/2}} \\
&\times \int \exp\left[-\frac{1}{2\sigma^2}\left\{\left(\boldsymbol{y}-X\boldsymbol{m}\right)'\left(I_n+XVX'\right)^{-1}\left(\boldsymbol{y}-X\boldsymbol{m}\right)\right\}\right] d\boldsymbol{\beta} \\
&\times \int \exp\left[-\frac{1}{2\sigma^2}+\left\{\left(\boldsymbol{\beta}-\tilde{\boldsymbol{m}}\right)'\tilde{V}^{-1}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{m}}\right)\right\}\right] d\boldsymbol{\beta} \\
&= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n+p}{2}}\left|V\right|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-X\boldsymbol{m}\right)'\left(I_n+XVX'\right)^{-1}\left(\boldsymbol{y}-X\boldsymbol{m}\right)\right\} \\
&\times \int \exp\left[-\frac{1}{2\sigma^2}\left\{\left(\boldsymbol{\beta}-\tilde{\boldsymbol{m}}\right)'\tilde{V}^{-1}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{m}}\right)\right\}\right] d\boldsymbol{\beta} \\
&= \frac{\left(2\pi\sigma^2\right)^{\frac{p}{2}}\left|\tilde{V}\right|^{1/2}}{\left(2\pi\sigma^2\right)^{\frac{n+p}{2}}\left|V\right|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-X\boldsymbol{m}\right)'\left(I_n+XVX'\right)^{-1}\left(\boldsymbol{y}-X\boldsymbol{m}\right)\right\} \\
&= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}}\left(\frac{\left|\tilde{V}\right|}{\left|V\right|}\right)^{1/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-X\boldsymbol{m}\right)'\left(I_n+XVX'\right)^{-1}\left(\boldsymbol{y}-X\boldsymbol{m}\right)\right\} \\
&= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}\left|I_n+XVX'\right|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-X\boldsymbol{m}\right)'\left(I_n+XVX'\right)^{-1}\left(\boldsymbol{y}-X\boldsymbol{m}\right)\right\}.
\end{aligned}
\tag{45}
$$

Η τελευταία ισότητα προκύπτει από την ταυτότητα πινάκων:

$$\left|A+BDC\right| = \left|A\right|\left|D\right|\left|D^{-1}+CA^{-1}B\right|,$$

για $A = I_n, B = X, D = V, C = X'$, δηλαδή, πιο συγκεκριμένα:

$$\left|I_n+XVX'\right| = \left|V\right|\left|V^{-1}+X'X\right| = \left(\frac{\left|V\right|}{\left|\tilde{V}\right|}\right).$$

Παρατηρώντας τη μορφή της (45), διαπιστώνουμε ότι πρόκειται για τη συνάρτηση πυκνότητας πιθανότητας $n-$διάστατης κανονικής κατανομής, ως εξής:

$$p\left(\boldsymbol{y} \mid \sigma^2\right) = N_n\left(X\boldsymbol{m}, \sigma^2\left(I_n+XVX'\right)\right). \tag{46}$$

Αφού καταλήξαμε στο παραπάνω αποτέλεσμα, απομένει να ολοκληρωθεί ως προς τη διασπορά το γινόμενο της νεοευρεθείσας δεσμευμένης κατανομής με την πρότερη κατανομή της διασποράς. Εφόσον η πρώτη είναι κανονική κατανομή και η δεύτερη αντίστροφη γάμμα, προφανώς το γινόμενό τους θα αντιστοιχεί σε κανονική-αντίστροφη γάμμα κατανομή. Συγκεκριμένα:

$$p(\boldsymbol{y}) = \int p\left(\boldsymbol{y} \mid \sigma^2\right) p\left(\sigma^2\right) d\sigma^2$$

$$
\begin{aligned}
&= \int N_n\left(X\boldsymbol{m}, \sigma^2\left(I_n + XVX'\right)\right) IG(a,b) d\sigma^2 \\
&= \int NIG\left(X\boldsymbol{m}, \left(I_n + XVX'\right), a, b\right) d\sigma^2 \\
&\propto \int \frac{1}{(\sigma^2)^{a+1+\frac{n}{2}}} \exp\left\{-\frac{1}{\sigma^2}\left[b + \frac{1}{2}\left(\boldsymbol{y} - X\boldsymbol{m}\right)'\left(I_n + XVX'\right)^{-1}\left(\boldsymbol{y} - X\boldsymbol{m}\right)\right]\right\} d\sigma^2 \\
&\propto \int IG\left(a + \frac{n}{2}, b + \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{m})'\left(I_n + XVX'\right)^{-1}\left(\boldsymbol{y} - X\boldsymbol{m}\right)\right) d\sigma^2 \\
&\propto \left[b + \frac{1}{2}(\boldsymbol{y} - X\boldsymbol{m})'\left(I_n + XVX'\right)^{-1}\left(\boldsymbol{y} - X\boldsymbol{m}\right)\right]^{-(a+n/2)} \\
&\propto \left[1 + \frac{(\boldsymbol{y} - X\boldsymbol{m})'\left(I_n + XVX'\right)^{-1}\left(\boldsymbol{y} - X\boldsymbol{m}\right)}{2b}\right]^{-(a+n/2)} \\
&\propto \left[1 + \frac{a}{a}\frac{(\boldsymbol{y} - X\boldsymbol{m})'\left(I_n + XVX'\right)^{-1}\left(\boldsymbol{y} - X\boldsymbol{m}\right)}{2b}\right]^{-(2a+n)/2} \\
&\propto \left[1 + \frac{(\boldsymbol{y} - X\boldsymbol{m})' a \left(I_n + XVX'\right)^{-1}\left(\boldsymbol{y} - X\boldsymbol{m}\right)}{2ab}\right]^{-(2a+n)/2}.
\end{aligned}
$$

Ορίζοντας τώρα τις ποσότητες:

$$
\begin{aligned}
\Sigma &= \frac{b}{a}\left(I_n + XVX'\right) \\
\nu &= 2a,
\end{aligned}
$$

η περιθώρια πιθανοφάνεια γίνεται:

$$
p\left(\boldsymbol{y}\right) \propto \left[1 + \frac{(\boldsymbol{y} - X\boldsymbol{m})'\Sigma^{-1}(\boldsymbol{y} - X\boldsymbol{m})}{\nu}\right]^{-(\nu+n)/2}. \tag{47}
$$

Μελετώντας τη μορφή της Εξίσωσης (47) και αναλογιζόμενοι το γεγονός ότι η διαδικασία που προηγήθηκε είναι ανάλογη με αυτή που ακολουθήθηκε για την εξαγωγή της ύστερης περιθώριας κατανομής $p(\boldsymbol{\beta}|\boldsymbol{y})$, συμπεραίνουμε πως το αποτέλεσμα είναι και πάλι η πολυδιάστατη Student κατανομή. Αυτή τη φορά, έχουμε παράμετρο θέσης $X\boldsymbol{m}$, παράμετρο κλίμακας $\frac{b}{a}(I + XVX')$ και $\nu = 2a$ βαθμούς ελευθερίας, όπου $a$ η παράμετρος σχήματος της πρότερης αντίστροφης γάμμα κατανομής του $\sigma^2$, δηλαδή:

$$
p(\boldsymbol{y}) = MVSt_{2a}\left(X\boldsymbol{m}, \frac{b}{a}\left(I_n + XVX'\right)\right),
$$

και πιο αναλυτικά:

$$
p(\boldsymbol{y}) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\pi^{\frac{p}{2}}|\nu(I + XVX')|^{\frac{1}{2}}}\left(1 + \frac{(\boldsymbol{y} - X\boldsymbol{\beta})'(I + XVX')^{-1}(\boldsymbol{y} - X\boldsymbol{\beta})}{\nu}\right)^{-\frac{\nu+n}{2}}. \tag{48}
$$

Ολοκληρώνοντας τη μελέτη για τον προσδιορισμό της περιθώριας πιθανοφάνειας για κανονική-αντίστροφη γάμμα πρότερη, θα δείξουμε μια χρήσιμη ιδιότητα της $NIG$ πρότερης. Συγκεκριμένα, γράφουμε την περιθώρια πιθανοφάνεια σύμφωνα με την αρχική της μορφή:

$$p(\boldsymbol{y}) = \int\int p\left(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2\right) p\left(\boldsymbol{\beta}, \sigma^2\right) d\boldsymbol{\beta} d\sigma^2 \Leftrightarrow$$
$$p(\boldsymbol{y}) = \int\int N_n\left(X\boldsymbol{\beta}, \sigma^2 I_n\right) \times NIG\left(\boldsymbol{m}, V, a, b\right) d\boldsymbol{\beta} d\sigma^2 = MVSt_{2a}\left(X\boldsymbol{m}, \frac{b}{a}\left(I_n + XVX'\right)\right). \tag{49}$$

Η τελευταία ισότητα προκύπτει επειδή ήδη δείξαμε ότι το αριστερό μέλος αντιστοιχεί στην πολυδιάστατη Student κατανομή που ορίστηκε παραπάνω. Πρακτικά αυτό σημαίνει πως ολοκληρώνοντας το γινόμενο κανονικής-αντίστροφης γάμμα κατανομής με μία κανονική κατανομή, προκύπτει πολυδιάστατη Student κατανομή, αποτέλεσμα που θα αξιοποιήσουμε στη συνέχεια.

**Προβλεπτικές κατανομές**

Εκτιμώντας Μπεϋζιανά τους συντελεστές της παλινδρόμησης, επιχειρούμε να αποκρυπτογραφήσουμε τη γραμμική σχέση μεταξύ των επεξηγηματικών μεταβλητών και της μεταβλητής απόκρισης. Κατόπιν, μπορούμε να χρησιμοποιήσουμε την νεοαποκτηθείσα γνώση για την πρόβλεψη της απόκρισης, δοθέντος ενός νέου $m \times p$ πίνακα παρατηρήσεων από τις $k$ επεξηγηματικές μεταβλητές, επαυξημένο όπως πάντα με μία στήλη από μονάδες ώστε να ενσωματώσουμε στο διάνυσμα $\boldsymbol{\beta}$ και τον σταθερό όρο.

Αν συμβολίσουμε με $X^*$ τον νέο $m \times p$ πίνακα σχεδιασμού, η μεταβλητή απόκρισης $\boldsymbol{y}^*$ θα κατανέμεται κανονικά, και συγκεκριμένα θα είναι:

$$\boldsymbol{y}^* | \boldsymbol{\beta}, \sigma^2 \sim N_m(X^*\boldsymbol{\beta}, \sigma^2 I_m)$$

όπου τα $\boldsymbol{\beta}, \sigma^2$ εξακολουθούν να είναι άγνωστα· ό,τι πληροφορία έχουμε για αυτά εμπεριέχεται στις ύστερες κατανομές. Επομένως, η ύστερη προβλεπτική κατανομή, με βάση την ύστερη κατανομή των παραμέτρων όπως αυτή προέκυψε από τα δεδομένα $\boldsymbol{y}$, εκτιμάται μέσω του παρακάτω ολοκληρώματος:

$$\begin{aligned} p(\boldsymbol{y}^*|\boldsymbol{y}) &= \int\int p(\boldsymbol{y}^*|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}) d\boldsymbol{\beta} d\sigma^2 \\ &= \int\int N_m(X^*\boldsymbol{\beta}, \sigma^2 I_m) NIG(\tilde{\boldsymbol{m}}, \tilde{V}, \tilde{a}, \tilde{b}) d\boldsymbol{\beta} d\sigma^2. \end{aligned}$$

Το ολοκλήρωμα αυτού του γινομένου έχουμε παρατηρήσει ήδη από την Εξίσωση (49) ότι οδηγεί σε πολυδιάστατη Student κατανομή κι έτσι προκύπτει:

$$p(\boldsymbol{y}^*|\boldsymbol{y}) = MVSt_{2\tilde{a}}\left(X^*\tilde{\boldsymbol{m}}, \frac{\tilde{b}}{\tilde{a}}\left(I_m + X^*\tilde{V}X^{*\prime}\right)\right). \tag{50}$$

Δηλαδή η κατανομή κάθε νέας πρόβλεψης, δοθέντος του δείγματος $\boldsymbol{y}$ θα είναι πολυδιάστατη Student με $2\tilde{a}$ βαθμούς ελευθερίας, παράμετρο θέσης $X^*\tilde{\boldsymbol{m}}$ και παράμετρο κλίμακας

$$\frac{\tilde{b}}{\tilde{a}}\left(I_m + X^*\tilde{V}X^{*\prime}\right).$$

Σαν τελικό σχόλιο, αναφέρουμε πως υπάρχουν δύο παράγοντες που προκαλούν αβεβαιότητα στην πρόβλεψη του $\tilde{\boldsymbol{y}}$ και η ύπαρξή τους αποτυπώνεται στην παράμετρο κλίμακας της προκύπτουσας κατανομής. Η πρώτη και κύρια πηγή μεταβλητότητας είναι αυτή που περιέχεται στο μοντέλο λόγω του $\sigma^2$. Η δεύτερη πηγή εντοπίζεται στην εκ των υστέρων αβεβαιότητα για τις παραμέτρους $\boldsymbol{\beta}, \sigma^2$, αφού έχουν εκτιμηθεί μέσω πεπερασμένου δείγματος (Banerjee 2008).

### 3.1.1 Η πρότερη του Zellner

Η παραπάνω ανάλυση κρύβει ένα πολύ λεπτό σημείο: αυτό του προσδιορισμού του πίνακα συνδιασποράς $V$ που εμφανίζεται στην κατανομή του $\boldsymbol{\beta}$. Ακόμα και αν διαθέτουμε επαρκή πρότερη πληροφορία για τις παραμέτρους του μοντέλου, η κατάλληλη επιλογή τιμών για τις υπερπαραμέτρους εξακολουθεί να είναι σύνθετη ενώ παράλληλα επηρεάζει σε πολύ μεγάλο βαθμό τη διασπορά στο ύστερο μοντέλο (Marin & Robert 2014).

Για τον προσδιορισμό λοιπόν των υπερπαραμέτρων προτάθηκε μία πιο αντικειμενική τακτική που κατά μία έννοια παρακάμπτει το πρόβλημα: πρόκειται για την πρότερη του Zellner, ή αλλιώς *Zellner's g-prior* (Zellner 1986). Τα κύρια χαρακτηριστικά αυτής της προσέγγισης είναι πως μας επιτρέπει να εισάγουμε πρότερη πληροφορία για το διάνυσμα $\boldsymbol{\beta}$ ενώ ταυτόχρονα μας απαλλάσσει από την ανάγκη να καθορίσουμε τη συνδιακύμανση των συμμεταβλητών, θέτοντας τον πίνακα συνδιασποράς σταθερά ίσο με $g(X'X)^{-1}$.

Όσον αφορά τους συντελεστές της παλινδρόμησης, έχουμε την ελευθερία να ενσωματώσουμε έστω και ασθενή εκ των προτέρων πληροφορία σχετικά με την κατανομή τους. Εντούτοις, η μελέτη απλουστεύεται ακόμα περισσότερο αν θεωρήσουμε την επίδραση των συμμεταβλητών στο μοντέλο κεντραρισμένη στο μηδέν. Ειδικότερα, στην περίπτωση αυτή, παίρνουμε:

$$\boldsymbol{\beta}|\sigma^2 \sim N_n(\mathbf{0}, g\sigma^2(X'X)^{-1}). \tag{51}$$

Η υπερπαράμετρος $g$ είναι μια χαρακτηριστική θετική ποσότητα της μεθόδου, εξ' ου και το όνομα g-prior, και για την επιλογή της τιμής της υπάρχουν ποικίλες προτάσεις· παραπέμπουμε στο Fernandez, Ley, & Steel 2001 σχετικά. Στην παρούσα εργασία θα εργαστούμε με το $g = n$ που ερμηνεύεται ως η προσθήκη πρότερης πληροφορίας η οποία ισοδυναμεί με μία και μόνο παρατήρηση (Ntzoufras 2009) και προτιμάται ελλείψει πρότερης πληροφορίας.

Όλες οι κατανομές που προκύπτουν από την πρότερη του Zellner μπορούν εύκολα να ευρεθούν από τις αντίστοιχες Εξισώσεις της Υποενότητας 3, αντικαθιστώντας τις τιμές των παραμέτρων που επιβάλλει η επιλογή της πρότερης αυτής. Ακολουθούν οι αναλυτικές εκφράσεις αυτών:

#### Από κοινού πρότερη κατανομή

Χρησιμοποιούμε την Εξίσωση (31), αντικαθιστώντας τις εξής τιμές για τις παραμέτρους:

$$\begin{aligned} \boldsymbol{m} &= \mathbf{0} \\ V &= g\left(X'X\right)^{-1}, \end{aligned}$$

και προκύπτει η από κοινού πρότερη κατανομή των $\boldsymbol{\beta}, \sigma^2$ η οποία θα είναι και πάλι κανονική-

αντίστροφη γάμμα:

$$p\left(\boldsymbol{\beta},\sigma^2\right) \propto \frac{1}{\left(g\sigma^2\right)^{a+1+\frac{p}{2}}} \exp\left\{-\frac{1}{g\sigma^2}\left[\frac{2b+\boldsymbol{\beta}'(X'X)\boldsymbol{\beta}}{2}\right]\right\} \Leftrightarrow$$
$$\boldsymbol{\beta},\sigma^2 \sim NIG(\mathbf{0}, g\sigma^2(X'X)^{-1}, a, b). \tag{52}$$

**Εκ των υστέρων κατανομή**

Γνωρίζουμε ήδη από την ανάλυση που πρηγήθηκε πως η κατανομή των $\boldsymbol{\beta},\sigma^2$, αφού έχουμε δει τα δεδομένα $\boldsymbol{y}$, θα είναι επίσης κανονική-αντίστροφη γάμμα, με ανανεωμένες τις τιμές των παραμέτρων. Πριν δώσουμε την έκφραση για την ύστερη, υπολογίζουμε τις παραμέτρους της, σύμφωνα με τις Εξισώσεις (32) - (35):

$$\begin{aligned}\tilde{V} &= \left(X'X + V^{-1}\right)^{-1} = \left(X'X + \frac{1}{g}X'X\right)^{-1} = \left(\frac{g}{g}X'X + \frac{1}{g}X'X\right)^{-1} \\ &= \left(\frac{g+1}{g}X'X\right)^{-1} = \frac{g}{g+1}\left(X'X\right)^{-1} \end{aligned} \tag{53}$$

$$\tilde{\boldsymbol{m}} = \tilde{V}\left(V^{-1}\boldsymbol{m} + X'\boldsymbol{y}\right) = \tilde{V}\left(X'\boldsymbol{y}\right) = \frac{g}{g+1}\left(X'X\right)^{-1}X'\boldsymbol{y} \tag{54}$$

$$\tilde{a} = a + \frac{n}{2} \tag{55}$$

$$\begin{aligned}\tilde{b} &= b + \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} - \tilde{\boldsymbol{m}}'\tilde{V}^{-1}\tilde{\boldsymbol{m}} + \boldsymbol{m}'V^{-1}\boldsymbol{m}\right) \\ &= b + \frac{1}{2}\left[\boldsymbol{y}'\boldsymbol{y} - \frac{g}{g+1}\boldsymbol{y}'X\left(X'X\right)^{-1}\frac{g+1}{g}\left(X'X\right)\frac{g}{g+1}\left(X'X\right)^{-1}X'\boldsymbol{y}\right] \\ &= b + \frac{1}{2}\left[\boldsymbol{y}'\boldsymbol{y} - \frac{g}{g+1}\boldsymbol{y}'X\left(X'X\right)^{-1}X'\boldsymbol{y}\right] \\ &= b + \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} - \tilde{m}'X'\boldsymbol{y}\right). \end{aligned} \tag{56}$$

Καταλήγουμε συνεπώς στην ακόλουθη κατανομή:

$$p\left(\boldsymbol{\beta},\sigma^2 \mid \boldsymbol{y}\right) \propto \frac{1}{\left(\sigma^2\right)^{\tilde{a}+1+\frac{p}{2}}} \exp\left\{-\frac{1}{g\sigma^2}\left[\tilde{b} + \frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})'\frac{g+1}{g}\left(X'X\right)(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})\right]\right\} \Leftrightarrow$$
$$\boldsymbol{\beta},\sigma^2 \mid \boldsymbol{y} \sim NIG(\tilde{\boldsymbol{m}}, \frac{g}{g+1}\left(X'X\right)^{-1}, \tilde{a}, \tilde{b}). \tag{57}$$

**Περιθώριες ύστερες κατανομές**

Ακολουθούν οι περιθώριες εκ των υστέρων κατανομές των $\boldsymbol{\beta}$ και $\sigma^2$, όπως δόθηκαν και στις Εξισώσεις (42) (43), με κατάλληλη προσαρμογή των παραμέτρων σύμφωνα με τις Εξισώσεις (53)-(56).

Ξεκινάμε από την εκ των υστέρων κατανομή του $\boldsymbol{\beta}$ που θα είναι πολυδιάστατη Student:

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto \left[1 + \frac{(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})'\Sigma^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})}{\nu}\right]^{-(\nu+p)/2} \Leftrightarrow$$
$$\boldsymbol{\beta} \mid \boldsymbol{y} \sim MVSt_{2\tilde{a}}\left(\tilde{\boldsymbol{m}}, \Sigma\right),$$

όπου απαιτείται ο εκ νέου προσδιορισμός της παραμέτρου κλίμακας $\Sigma$ ως εξής:

$$\Sigma = \frac{\tilde{b}}{\tilde{a}}\tilde{V} = \frac{\tilde{b}}{\tilde{a}}\frac{g}{g+1}\left(X'X\right)^{-1}. \tag{58}$$

Ακολουθεί η ύστερη κατανομή του $\sigma^2$, που όπως αναμένεται είναι η αντίστροφη γάμμα:

$$p\left(\sigma^2 \mid \boldsymbol{y}\right) \propto \frac{1}{\left(\sigma^2\right)^{\tilde{a}+1+\frac{p}{2}}}\exp\left(-\frac{\tilde{b}}{\sigma^2}\right) \Leftrightarrow$$
$$\sigma^2 \mid \boldsymbol{y} \sim IG\left(\tilde{a}, \tilde{b}\right).$$

**Περιθώρια πιθανοφάνεια**

Η κατανομή της περιθώριας πιθανοφάνειας με χρήση της πρότερης του Zellner, δίνεται απευθείας από την Εξίσωση (48) για τις νέες τιμές των $\boldsymbol{m}$ και $V$:

$$p(\boldsymbol{y}) = MVSt_{2a}\left(\boldsymbol{0}, \frac{b}{a}\left(I_n + XVX'\right)\right).$$

**Προβλεπτικες κατανομές**

Τέλος, για την προβλεπτική κατανομή της απόκρισης $\boldsymbol{y}^*$ που αντιστοιχεί σε νέα δεδομένα με πίνακα σχεδιασμού $X^*$ διαστάσεων $m \times p$, ανατρέχουμε στην Εξίσωση (15):

$$p\left(\boldsymbol{y}^* \mid \boldsymbol{y}\right) = MVSt_{2a^*}\left(X\tilde{\boldsymbol{m}}, \frac{\tilde{b}}{\tilde{a}}\left(I + X^*\tilde{V}X^{*\prime}\right)\right).$$

## 3.2 Improper πρότερη για τη διασπορά

Ο προσδιορισμός της εκ των προτέρων κατανομής της διασποράς είναι μια διαδικασία που κρύβει δυσκολίες καθώς συχνά περιλαμβάνει και τον προσδιορισμό των τιμών των υπερπαραμέτρων, καθιστώντας τη μελέτη ακόμα πιο σύνθετη. Ειδικότερα, όταν δεν υπάρχουν προηγούμενες μελέτες ή δεδομένα στα οποία μπορούμε να βασιστούμε, είναι συνετό να μην τροφοδοτείται το μοντέλο με πρότερες κατανομές παραπλανητικές ως προς το πραγματικό επίπεδο γνώσης που έχουμε για τις παραμέτρους.

Αποσκοπώντας στην απλοποίηση αυτής της διαδικασίας, επιλέγουμε μία πρότερη κατανομή που δεν περιέχει παρά ελάχιστη πληροφορία, όπως η ακόλουθη, improper, κατανομή:

$$p(\sigma^2) \propto \frac{1}{\sigma^2}. \tag{59}$$

Η παραπάνω έκφραση μας εξυπηρετεί αρκετά καθώς μπορεί να θεωρηθεί ως οριακή περίπτωση σ.π.π. της αντίστροφης γάμμα κατανομής $IG(a, b)$ με $a \to 0, b \to 0$. Μάλιστα, με έναν απλό μετασχηματισμό, διαπιστώνουμε ότι η υιοθέτηση της πρότερης $p(\sigma^2) \propto 1/\sigma^2$ είναι ισοδύναμη με την ομοιόμορφη πρότερη για το $\ln(\sigma^2)$. Η παρατήρηση αυτή, σε συνδυασμό με την υιοθέτηση κανονικής πρότερης για το $\boldsymbol{\beta}$, επιτρέπει τη χρήση της μεθοδολογίας της προηγούμενης ενότητας, αξιοποιώντας για άλλη μια φορά τις ιδιότητες των συζυγών πρότερων.

Διατηρώντας την πρότερη της προηγούμενης Υποενότητας για την κατανομή του $\boldsymbol{\beta}|\sigma^2$, όπως δόθηκε από την Εξίσωση (29), και συνδυάζοντάς τη με την πρότερη της Εξίσωσης (59), εξάγουμε την από κοινού πρότερη κατανομή των παραμέτρων της παλινδρόμησης και την παραθέτουμε ακολούθως:

$$\begin{aligned} p\left(\boldsymbol{\beta}, \sigma^2\right) &\propto p\left(\boldsymbol{\beta} \mid \sigma^2\right) \times p\left(\sigma^2\right) \\ &\propto \frac{1}{\left(\sigma^2\right)^{\frac{p}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\boldsymbol{m})' V^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right\} \times \frac{1}{\sigma^2} \\ &\propto \frac{1}{\left(\sigma^2\right)^{1+\frac{p}{2}}} \exp\left\{-\frac{(\boldsymbol{\beta}-\boldsymbol{m})' V^{-1}(\boldsymbol{\beta}-\boldsymbol{m})}{2\sigma^2}\right\} \Leftrightarrow . \end{aligned}$$

Το αποτέλεσμα, όπως αναμενόταν, είναι η κανονική-αντίστροφη γάμμα πρότερη,

$$\boldsymbol{\beta}, \sigma^2 \sim NIG(\boldsymbol{m}, V, a, b) \tag{60}$$

όπου $a \to 0, b \to 0$.

**Εκ των υστέρων κατανομή**

Λόγω του συζυγούς χαρακτήρα της κανονικής-αντίστροφης γάμμα πρότερης ενόψει κανονικής πιθανοφάνειας, γνωρίζουμε πως και η ύστερη θα περιγράφεται από την ίδια κατανομή, με ανανεωμένες τιμές παραμέτρων:

$$\begin{aligned} p\left(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}\right) \propto & L\left(\boldsymbol{\beta}, \sigma^2\right) \times p\left(\boldsymbol{\beta}, \sigma^2\right) \\ \propto & \frac{1}{\left(\sigma^2\right)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-X\boldsymbol{\beta})'(\boldsymbol{y}-X\boldsymbol{\beta})\right\} \\ & \times \frac{1}{\left(\sigma^2\right)^{1+\frac{p}{2}}} \exp\left\{-\frac{(\boldsymbol{\beta}-\boldsymbol{m})' V^{-1}(\boldsymbol{\beta}-\boldsymbol{m})}{2\sigma^2}\right\}. \end{aligned}$$

Τα βήματα που ακολουθούνται για τον αναλυτικό προσδιορισμό των νέων παραμέτρων δεν διαφέρουν από αυτά που ακολουθήθηκαν στην Υποενότητα 3.1 που προηγήθηκε. Εν προκειμένω, πρέπει και πάλι να συγκεντρώσουμε τους εκθετικούς όρους σε έναν ως εξής:

$$\exp\left\{-\frac{1}{2\sigma^2}\left[(\boldsymbol{y}-X\boldsymbol{\beta})'(\boldsymbol{y}-X\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{m})' V^{-1}(\boldsymbol{\beta}-\boldsymbol{m})\right]\right\}.$$

Κατόπιν, επεξεργαζόμαστε και μετασχηματίζουμε το εσωτερικό του εκθετικού με τρόπο εντελώς ανάλογο με την προηγούμενη Υποενότητα, με τη βοήθεια των νέων παραμέτρων $\tilde{V}$ και $\tilde{\boldsymbol{m}}$, ακριβώς όπως αυτές ορίστηκαν στις Εξισώσεις (32) και (33), δηλαδή:

$$\begin{aligned} \tilde{V} &= \left(X'X + V^{-1}\right)^{-1} \\ \tilde{\boldsymbol{m}} &= \tilde{V}\left(V^{-1}\boldsymbol{m} + X'\boldsymbol{y}\right). \end{aligned}$$

Απομένει λοιπόν να δοθούν οι νέες τιμές για τις παραμέτρους σχήματος και κλίμακας, $\tilde{a}$ και $\tilde{b}$, που αντιστοιχούν στο inverse gamma κομμάτι της ύστερης, όπου και έχουμε

διαφοροποίηση από τις αντίστοιχες Εξισώσεις (34) και (35) που προέκυψαν από την κανονική-αντίστροφη γάμμα πρότερης:

$$\tilde{a} = \frac{n}{2} \tag{61}$$

$$\tilde{b} = \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} - \tilde{\boldsymbol{m}}'\tilde{V}^{-1}\tilde{\boldsymbol{m}} + \boldsymbol{m}'V^{-1}\boldsymbol{m}\right). \tag{62}$$

Τελικά, λαμβάνουμε την ακόλουθη κανονική-αντίστροφη γάμμα εκ των υστέρων κατανομή για τις παραμέτρους της παλινδρόμησης:

$$p\left(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}\right) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}+1+\frac{p}{2}}} \exp\left\{-\frac{1}{\sigma^2}\left[\tilde{b} + \frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})'\tilde{V}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})\right]\right\} \Leftrightarrow$$

$$\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y} \sim NIG\left(\tilde{m}, \tilde{V}, \tilde{a}, \tilde{b}\right). \tag{63}$$

**Περιθώριες ύστερες κατανομές**

Για την εύρεση των εκ των υστέρων κατανομών των $\boldsymbol{\beta}$ και $\sigma^2$, δεν έχουμε παρά να αντικαταστήσουμε τις Εξισώσεις (61) και (62), που μας δίνουν τις διαφοροποιημένες από την προηγούμενη Υποενότητα ύστερες τιμές των παραμέτρων, στις Εξισώσεις (42) και (43). Όσον αφορά τις παραμέτρους $\tilde{\boldsymbol{m}}, \tilde{V}$ παραμένουν όπως έχουν οριστεί στην Υποενότητα 3.1.

Αμέσως φανερώνεται πως για άλλη μια φορά η περιθώρια ύστερη κατανομή του $\boldsymbol{\beta}$ θα είναι πολυδιάστατη Student ενώ του $\sigma^2$ αντίστροφη-γάμμα. Επίσης, ειδικά για την κατανομή του $\boldsymbol{\beta}$, θα χρειαστεί να ορίσουμε την ακόλουθη ποσότητα:

$$\Sigma = \frac{2\tilde{b}}{\tilde{a}}\tilde{V},$$

η οποία αποτελεί την παράμετρο κλίμακας της Student κατανομής. Συνεπώς, οι εν λόγω κατανομές διαμορφώνονται ως εξής:

$$p\left(\boldsymbol{\beta} \mid \boldsymbol{y}\right) = \frac{\Gamma\left(n + \frac{p}{2}\right)}{\Gamma\left(n\right)\pi^{\frac{p}{2}}|n\Sigma|^{\frac{1}{2}}}\left(1 + \frac{(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})'\Sigma^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})}{2n}\right)^{-\left(n+\frac{p}{2}\right)} \Leftrightarrow$$

$$\boldsymbol{\beta} \mid \boldsymbol{y} \sim MVSt_n\left(\tilde{\boldsymbol{m}}, \Sigma\right), \tag{64}$$

$$p\left(\sigma^2 \mid \boldsymbol{y}\right) = \frac{\tilde{a}^{\tilde{b}}}{\Gamma(\tilde{a})}\frac{1}{(\sigma^2)^{\frac{n}{2}+1+\frac{p}{2}}}\exp\left(-\frac{\tilde{b}}{\sigma^2}\right) \Leftrightarrow$$

$$\sigma^2 \mid \boldsymbol{y} \sim IG\left(\tilde{a}, \tilde{b}\right). \tag{65}$$

**Περιθώρια πιθανοφάνεια**

Σύμφωνα με την Εξίσωση (44), μπορούμε να προσδιορίσουμε την περιθώρια πιθανοφάνεια $p\left(\boldsymbol{y}\right)$

$$p\left(\boldsymbol{y} \mid \sigma^2\right) = N_n\left(X\boldsymbol{m}, \sigma^2\left(I_n + XVX'\right)\right). \tag{66}$$

$$
\begin{aligned}
p(\boldsymbol{y}) &= \int p\left(\boldsymbol{y} \mid \sigma^2\right) p\left(\sigma^2\right) d\sigma^2 \\
&= \int N_n\left(X\boldsymbol{m}, \sigma^2\left(I_n + XVX'\right)\right) IG(a,b) d\sigma^2 \\
&= \int NIG\left(X\boldsymbol{m}, \left(I_n + XVX'\right), a, b\right) d\sigma^2 \\
&\propto \int \frac{1}{(2\pi)^{\frac{n}{2}}\left(\sigma^2\right)^{1+\frac{n}{2}} |I_n + XVX'|^{\frac{1}{2}}} \\
&\times \exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}\left(\boldsymbol{y} - X\boldsymbol{m}\right)'\left(I_n + XVX'\right)^{-1}\left(\boldsymbol{y} - X\boldsymbol{m}\right)\right]\right\} d\sigma^2 \\
&\propto \frac{1}{(2\pi)^{\frac{n}{2}} |I_n + XVX'|^{\frac{1}{2}}} \\
&\times \int \frac{1}{\left(\sigma^2\right)^{1+\frac{n}{2}}} \exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}\left(\boldsymbol{y} - X\boldsymbol{m}\right)'\left(I_n + XVX'\right)^{-1}\left(\boldsymbol{y} - X\boldsymbol{m}\right)\right]\right\} d\sigma^2
\end{aligned}
$$

Παρατηρούμε ότι το ολοκλήρωμα της τελευταίας σχέσης είναι ανάλογο της συνάρτησης πυκνότητας πιθανότητας της αντίστροφης γάμμα κατανομής με παραμέτρους:

$$
\begin{aligned}
a &= \frac{n}{2} \\
b &= \frac{1}{2}\left(\boldsymbol{y} - X\boldsymbol{m}\right)'\left(I_n + XVX'\right)^{-1}\left(\boldsymbol{y} - X\boldsymbol{m}\right).
\end{aligned}
$$

Επομένως, πολλαπλασιάζοντας και διαιρώντας με τις σταθερές της συγκεκριμένης αντίστροφης γάμμα σ.π.π., το ολοκλήρωμα θα γίνει ίσο με τη μονάδα και η έκφραση για την περιθώρια πιθανοφάνεια διαμορφώνεται ως εξής:

$$
\begin{aligned}
p(\boldsymbol{y}) &\propto \frac{\Gamma\left(\frac{n}{2}\right)}{(2\pi)^{\frac{n}{2}} |I_n + XVX'|^{\frac{1}{2}}} \left[\frac{1}{\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{m})'\left(I_n + XVX'\right)^{-1}\left(\boldsymbol{y} - X\boldsymbol{m}\right)}\right]^{n/2} \Leftrightarrow \\
&\propto \frac{\Gamma\left(\frac{n}{2}\right)}{(\pi)^{\frac{n}{2}} |I_n + XVX'|^{\frac{1}{2}}} \left[(\boldsymbol{y} - X\boldsymbol{m})'\left(I_n + XVX'\right)^{-1}\left(\boldsymbol{y} - X\boldsymbol{m}\right)\right]^{-\frac{n}{2}}. \qquad (67)
\end{aligned}
$$

Η Εξίσωση (67) δεν αντιστοιχεί σε συνάρτηση πυκνότητας πιθανότητας κάποιας γνωστής κατανομής, ωστόσο παρατηρούμε ότι περιέχει τον πυρήνα $p-$διάστατης Student.

**Προβλεπτικές κατανομές**

Η πρόβλεψη της κατανομής του διανύσματος $\boldsymbol{y}^*$ που αφορά νέα δεδομένα με πίνακα σχεδιασμού $X^*$, διαστάσεων $m \times p$, γίνεται άμεσα με την αντικατάσταση των ανανεωμένων τιμών (61) και (62) στην Εξίσωση (50). Οι παράμετροι $\tilde{\boldsymbol{m}}, \tilde{V}$ παραμένουν και πάλι όπως έχουν οριστεί στην Υποενότητα 3.1. Έτσι προκύπτει η πολυδιάστατη Student κατανομή με $n$ βαθμούς ελευθερίας:

$$
p(\boldsymbol{y}^*|\boldsymbol{y}) = MVSt_n\left(X^*\tilde{\boldsymbol{m}}, \frac{2\tilde{b}}{\tilde{n}}\left(I_m + X^*\tilde{V}X^{*\prime}\right)\right). \qquad (68)
$$

## 3.3 Μη-πληροφοριακή από κοινού πρότερη

Στην ενότητα που προηγήθηκε, παρουσιάστηκε ένας ευρέως διαδεδομένος τρόπος για να εμπλουτίσουμε μόνο μερικώς το μοντέλο με πρότερη πληροφορία. Στην παρούσα ενότητα, το ενδιαφέρον στρέφεται στις περιπτώσεις όπου επιλέγουμε να μην χρησιμοποιήσουμε καθόλου πρότερη γνώση, είτε επειδή απλώς δεν υπάρχει, είτε επειδή δεν θέλουμε να επηρεάσουμε την ύστερη κατανομή, αφήνοντας τα δεδομένα να "μιλήσουν" από μόνα τους. Αυτή η προσέγγιση πραγματώνεται μέσω της κατασκευής μη-πληροφοριακών πρότερων ή vague ή noniformative ή uninformative priors όπως αναφέρονται στην αγγλική βιβλιογραφία.

Η δημοφιλέστερη επιλογή μη-πληροφοριακής από κοινού πρότερης είναι η εξής:

$$p\left(\boldsymbol{\beta}, \sigma^2\right) \propto \frac{1}{\sigma^2}. \tag{69}$$

Η παραπάνω μορφή μπορεί να προκύψει αν θεωρήσουμε ξανά την improper αντίστροφη γάμμα πρότερη της Εξίσωσης (59) για τη διασπορά $\sigma^2$, αυτή τη φορά με παραμέτρους:

$$a = -p/2, b \to 0,$$

σε συνδυασμό με την κανονική εκ των προτέρων κατανομή της Εξίσωσης (29) για το διάνυσμα $\boldsymbol{\beta}$ των παραμέτρων της παλινδρόμησης, θέτοντας όμως $V^{-1} \to 0$ για τον πίνακα συνδιακύμανσης (O'Hagan & Forster 2004). Η Εξίσωση (69) μπορεί να θεωρηθεί οριακή περίπτωση κανονικής-αντίστροφης γάμμα πρότερης, οπότε και αξιοποιούμε τις συζυγείς της ιδιότητες για την εξαγωγή συμπερασμάτων σχετικά με την ύστερη κατανομή. Από την μορφή της (69), είναι φανερό πως η συγκεκριμένη πρότερη είναι και improper.

Ένα ακόμη σημαντικό χαρακτηριστικό αυτής της πρότερης, όπως θα δούμε στη συνέχεια, είναι ότι οδηγεί σε εκτιμήτρια $\tilde{\boldsymbol{m}}$ για τους συντελεστές του μοντέλου που συμπίπτει με τη γνωστή μορφή της εκτιμήτριας ελαχίστων τετραγώνων. Συνεπώς, στο πλαίσιο απλής σύγκρισης των εκτιμητριών της Μπεϋζιανής και της κλασικής παλινδρόμησης μπορούμε να σχολιάσουμε ότι η πρώτη, στην απλούστερη μορφή της, δηλαδή χωρίς την ενσωμάτωση πρότερης πληροφορίας, εκτιμά το $\boldsymbol{\beta}$ τόσο καλά όσο και η δεύτερη.

Σημειώνεται πως η περιθώρια πιθανοφάνεια $p(\boldsymbol{y})$ δεν μπορεί να οριστεί στην παρούσα περίπτωση.

#### Εκ των υστέρων κατανομή

Εφόσον η από κοινού πρότερη υπάγεται στην γενικότερη περίπτωση κανονικής-αντίστροφης γάμμα πρότερης της Υποενότητας 3.1, ακολουθούμε την ίδια διαδικασία για την εξαγωγή της εκ των υστέρων από κοινού κατανομής, η οποία θα είναι επίσης κανονική-αντίστροφη γάμμα, όπως συνεπάγεται από τη σχέση συζυγίας.

Κατά τα γνωστά, επιστρατεύουμε τον κανόνα του Bayes για να λάβουμε τη μορφή της ύστερης:

$$\begin{aligned} p\left(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}\right) &\propto L(\boldsymbol{\beta}, \sigma^2) \times p\left(\boldsymbol{\beta}, \sigma^2\right) \\ &\propto \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta})\right\} \times \frac{1}{\sigma^2} \\ &= \frac{1}{(\sigma^2)^{\frac{n}{2}+1}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta})\right\}. \end{aligned}$$

Κατόπιν, θέτουμε τις παρακάτω ποσότητες:

$$\tilde{V} = (X'X)^{-1} \tag{70}$$

$$\tilde{\boldsymbol{m}} = \tilde{V}X'\boldsymbol{y} \tag{71}$$

και με τη βοήθεια αυτών επεξεργαζόμαστε το εσωτερικό του εκθετικού:

$$\begin{aligned}
&(\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta}) \\
=&\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'X'\boldsymbol{y} + \boldsymbol{\beta}'X'X\boldsymbol{\beta} \\
=&\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}(X'X)^{-1}(X'X)X'\boldsymbol{y} + \boldsymbol{\beta}'X'X\boldsymbol{\beta} \\
=&\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}\tilde{V}^{-1}\tilde{\boldsymbol{m}} + \boldsymbol{\beta}'\tilde{V}^{-1}\boldsymbol{\beta} \\
=&\boldsymbol{y}'\boldsymbol{y} + (\boldsymbol{\beta} - \tilde{\boldsymbol{m}})'\tilde{V}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{m}}) - \tilde{\boldsymbol{m}}'\tilde{V}^{-1}\tilde{\boldsymbol{m}}.
\end{aligned}$$

Ενσωματώνοντας την νέα έκφραση για τον εκθετικό όρο, προκύπτει η παρακάτω μορφή της ύστερης:

$$p\left(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}\right) \propto \frac{1}{(\sigma^2)^{\frac{n}{2}-\frac{p}{2}+\frac{p}{2}+1}} \exp\left\{-\frac{1}{2\sigma^2}\left[\boldsymbol{y}'\boldsymbol{y} + (\boldsymbol{\beta} - \tilde{\boldsymbol{m}})'\tilde{V}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{m}}) - \tilde{\boldsymbol{m}}'\tilde{V}^{-1}\tilde{\boldsymbol{m}}\right]\right\}.$$

Τέλος, θέτοντας τις ακόλουθες ποσότητες:

$$\tilde{a} = \frac{n-p}{2} = a + \frac{n}{2} \tag{72}$$

$$\tilde{b} = \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} - \tilde{\boldsymbol{m}}'\tilde{V}^{-1}\tilde{\boldsymbol{m}}\right) \tag{73}$$

λαμβάνουμε την τελική μορφή της ύστερης, η οποία αντιστοιχεί σε κανονική-αντίστροφη γάμμα κατανομή με παραμέτρους $\tilde{\boldsymbol{m}}, \tilde{V}, \tilde{a}, \tilde{b}$, όπως αυτές ορίστηκαν στις σχέσεις (70)-(73):

$$\begin{aligned}
p\left(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}\right) \propto \frac{1}{(\sigma^2)^{\tilde{a}+1+\frac{p}{2}}} \exp\left\{-\frac{1}{\sigma^2}\left[\tilde{b} + \frac{1}{2}(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})'\tilde{V}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})\right]\right\} \Leftrightarrow \\
\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y} \sim NIG\left(\tilde{\boldsymbol{m}}, \tilde{V}, \tilde{a}, \tilde{b}\right).
\end{aligned} \tag{74}$$

Αν παρατηρήσουμε την έκφραση (71) που, εκτός από την παράμετρο θέσης της ύστερης κατανομής, αποτελεί την Μπεϋζιανή εκτιμήτρια του διανύσματος $\boldsymbol{\beta}$ των συντελεστών του μοντέλου, διαπιστώνουμε ότι αυτή ταυτίζεται με την γνωστή εκτιμήτρια ελαχίστων τετραγώνων (ordinary least squares) $\hat{\boldsymbol{\beta}}_{OLS}$:

$$\tilde{\boldsymbol{m}} = \tilde{V}X'\boldsymbol{y} = \left(X'X\right)^{-1}X'\boldsymbol{y} = \hat{\boldsymbol{\beta}}_{OLS}.$$

Επομένως, μπορούμε κατά μία έννοια να θεωρήσουμε την Μπεϋζιανή παλινδρόμηση με υιοθέτηση μη-πληροφοριακής πρότερης ισοδύναμη με την κλασική παλινδρόμηση των εκτιμητριών ελαχίστων τετραγώνων.

Τέλος, παρουσιάζουμε μια ακόμα χρήσιμη έκφραση για το $\tilde{b}$, συναρτήσει της ποσότητας:

$$s^2 = \frac{1}{n-p}(\boldsymbol{y} - X\boldsymbol{\beta})'(\boldsymbol{y} - X\boldsymbol{\beta}).$$

Για τον σκοπό αυτό, μετασχηματίζουμε αναλόγως την Εξίσωση (73) παρακάτω:

$$\begin{aligned}
\tilde{b} &= \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} - \tilde{\boldsymbol{m}}'\tilde{V}^{-1}\tilde{\boldsymbol{m}}\right) \\
&= \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'X\tilde{V}'\tilde{V}^{-1}\tilde{V}X'\boldsymbol{y}\right) \\
&= \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'X\tilde{V}'X'\boldsymbol{y}\right) \\
&= \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} - \tilde{\boldsymbol{m}}'X'\boldsymbol{y}\right) \\
&= \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} - 2\tilde{\boldsymbol{m}}'X'\boldsymbol{y} + \tilde{\boldsymbol{m}}'X'\boldsymbol{y}\right) \\
&= \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} - 2\tilde{\boldsymbol{m}}'X'\boldsymbol{y} + \tilde{\boldsymbol{m}}'X'\tilde{V}^{-1}\tilde{V}X'\boldsymbol{y}\right) \\
&= \frac{1}{2}\left(\boldsymbol{y}'\boldsymbol{y} - 2\tilde{\boldsymbol{m}}'X'\boldsymbol{y} + \tilde{\boldsymbol{m}}'X'X\tilde{\boldsymbol{m}}\right) \\
&= \frac{1}{2}\left(\boldsymbol{y} - X\tilde{\boldsymbol{m}}\right)'\left(\boldsymbol{y} - X\tilde{\boldsymbol{m}}\right) \\
\Leftrightarrow \tilde{b} &= \frac{n-p}{2}s^2. \qquad (75)
\end{aligned}$$

**Περιθώριες ύστερες κατανομές**

Η συνήθης διαδικασία ολοκλήρωσης για την εξαγωγή των περιθώριων ύστερων κατανομών που ακολουθήθηκε στις ενότητες που προηγήθηκαν, δύναται να εφαρμοστεί απαράλλακτη και σε αυτό το εδάφιο.

Συνεπώς, άμεσα συμπεραίνουμε πως η εκ των υστέρων περιθώρια κατανομή $p\left(\boldsymbol{\beta} \mid \boldsymbol{y}\right)$ του $\boldsymbol{\beta}$ θα είναι και πάλι πολυδιάστατη Student. Ειδικότερα:

$$\begin{aligned}
p\left(\boldsymbol{\beta} \mid \boldsymbol{y}\right) &= \int p\left(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}\right) d\sigma^2 \\
&\propto \left(1 + \frac{(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})'\tilde{V}^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})}{2\tilde{b}}\right)^{-(\tilde{a}+\frac{p}{2})}.
\end{aligned}$$

Αντικαθιστώντας τις αναλυτικές εκφράσεις των παραμέτρων $\tilde{V}, \tilde{a}$ και χρησιμοποιώντας την εναλλακτική έκφραση για το $\tilde{b}$ της Εξίσωσης (75), έχουμε:

$$\begin{aligned}
p\left(\boldsymbol{\beta} \mid \boldsymbol{y}\right) &\propto \left(1 + \frac{(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})'\left(X'X\right)(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})}{(n-p)s^2}\right)^{-(\frac{n-p}{2}+\frac{p}{2})} \\
&= \left(1 + \frac{(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})'\Sigma^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})}{(n-p)}\right)^{-\left[\frac{(n-p)+p}{2}\right]} \\
&= \left(1 + \frac{(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})'\Sigma^{-1}(\boldsymbol{\beta} - \tilde{\boldsymbol{m}})}{\nu}\right)^{-\left(\frac{\nu+p}{2}\right)}.
\end{aligned}$$

Καταλήγουμε στην ακόλουθη κατανομή η οποία εκ πρώτης όψεως μας θυμίζει αυτή της σχέσης (42):

$$p\left(\boldsymbol{\beta} \mid \boldsymbol{y}\right) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\pi^{\frac{p}{2}}|\nu\Sigma|^{\frac{1}{2}}}\left(1+\frac{(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})'\Sigma^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{m}})}{\nu}\right)^{-\frac{\nu+p}{2}} \Leftrightarrow$$
$$\boldsymbol{\beta} \mid \boldsymbol{y} \sim MVSt_{n-p}\left(\tilde{\boldsymbol{m}}, \Sigma\right). \tag{76}$$

Στην πραγματικότητα διαφέρουν ως προς τις τιμές των παραμέτρων αφού στην προκειμένη έχουμε $\nu = n-p$ βαθμούς ελευθερίας, η παράμετρος θέσης $\tilde{\boldsymbol{m}}$ είναι ίση με την εκτιμήτρια ελαχίστων τετραγώνων και η παράμετρος κλίμακας δίνεται ως $\Sigma = \frac{\tilde{b}}{\tilde{a}}\tilde{V} = s^2(X'X)^{-1}$.

Όσον αφορά την εκ των υστέρων κατανομή του $\sigma^2$, αξιοποιώντας τη σχέση (43), διαπιστώνουμε ότι θα ακολουθεί αντίστροφη γάμμα κατανομή με παραμέτρους $\tilde{a} = \frac{(n-p)}{2}$ και $\tilde{b} = \frac{(n-p)}{2}s^2$:

$$p\left(\sigma^2 \mid \boldsymbol{y}\right) = \frac{\tilde{a}^{\tilde{b}}}{\Gamma(\tilde{a})}\frac{1}{(\sigma^2)^{\tilde{a}+1+\frac{p}{2}}}\exp\left(-\frac{\tilde{b}}{\sigma^2}\right) \Leftrightarrow$$
$$\sigma^2 \mid \boldsymbol{y} \sim IG\left(\tilde{a}, \tilde{b}\right). \tag{77}$$

**Προβλεπτικές κατανομές**

Κατά τα γνωστά, έστω $X^*$ ένας $m \times p$ πίνακας με νέες παρατηρήσεις. Μας ενδιαφέρει η κατανομή της απόκρισης $\boldsymbol{y}^*$ των καινούριων παρατηρήσεων, δοθέντος του διανύσματος $\boldsymbol{y}$, η οποία γράφεται ως εξής:

$$p(\boldsymbol{y}^*|\boldsymbol{y}) = \int\int p(\boldsymbol{y}^*|\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta},\sigma^2|\boldsymbol{y})d\boldsymbol{\beta}d\sigma^2.$$
$$= \int\int N_m(X^*\boldsymbol{\beta},\sigma^2 I_m)NIG(\hat{\boldsymbol{\beta}}_{OLS}, s^2(X'X), \frac{n-p}{2}, \frac{n-p}{2}s^2)d\boldsymbol{\beta}d\sigma^2.$$

Στο σημείο αυτό παρατηρούμε ότι μπορούμε και πάλι να αξιοποιήσουμε την παρατήρηση που διατυπώθηκε μέσω της Εξίσωσης (49) και άμεσα να εξάγουμε τη ζητούμενη κατανομή, που όπως όλες οι προβλεπτικές κατανομές που έχουμε καταλήξει μέχρι τώρα είναι πολυδιάστατη Student. Πιο συγκεκριμένα, έχουμε:

$$p(\boldsymbol{y}^*|\boldsymbol{y}) = MVSt_{n-p}\left(X^*\hat{\boldsymbol{\beta}}_{OLS}, s^2\left(I_m + X^*(X'X)^{-1}X^{*\prime}\right)\right). \tag{78}$$

Επομένως η προβλεπτική κατανομή είναι πολυδιάστατη Student με βαθμούς ελευθερίας:

$$\nu = 2a = 2\frac{n-p}{2} = n-p,$$

παράμετρο θέσης:

$$X^*\hat{\boldsymbol{\beta}}_{OLS},$$

και παράμετρο κλίμακας:

$$\frac{\tilde{b}}{\tilde{a}}\left(I_m + X^*(X'X)^{-1}X^{*\prime}\right) = \frac{\frac{n-p}{2}s^2}{\frac{n-p}{2}} = s^2\left(I_m + X^*(X'X)^{-1}X^{*\prime}\right).$$

# 4 Εφαρμογές σε προσομοιωμένα και πραγματικά δεδομένα

Το τελευταίο κομμάτι της εργασίας αφορά την πρακτική εφαρμογή των τεχνικών που περιγράφηκαν θεωρητικά ως τώρα. Θα χρησιμοποιήσουμε πρώτα προσομοιωμένα, με τη βοήθεια γεννητριών τυχαίων αριθμών, δεδομένα και στη συνέχεια πραγματικά δεδομένα, διαθέσιμα στο UCI Machine Learning Repository.

Σε αμφότερα τα προσομοιωμένα και πραγματικά δεδομένα, θα επικεντρωθούμε στη Μπεϋζιανή παλινδρόμηση με χρήση των πρότερων που έχουν παρουσιαστεί στην Ενότητα 3. Το προγραμματιστικό εργαλείο που θα επιστρατευτεί εδώ είναι η στατιστική γλώσσα προγραμματισμού R®, στο περιβάλλον ανάπτυξης του RStudio®.

## 4.1 Εφαρμογή σε προσομοιωμένα δεδομένα

Για την εξοικείωσή μας με την εφαρμογή της Μπεϋζιανής παλινδρόμησης, προσομοιώνουμε αρχικά τέσσερα σετ δεδομένων, ώστε να γνωρίζουμε εξ αρχής τη δομή τους. Σε αυτά, δοκιμάζουμε τα τρία συζυγών πρότερων που μελετήθηκαν και συγκρίνουμε τις προκύπτουσες ύστερες με τις πραγματικές τιμές.

### 4.1.1 Παραγωγή και παρουσίαση δεδομένων

**Πρώτο σύνολο δεδομένων**

Θα δημιουργήσουμε $50$ παρατηρήσεις από $15$ μεταβλητές της τυποποιημένης κανονικής κατανομής:

$$x_i \sim N(0,1), i = 1, ..., 15,$$

οι οποίες θα έχουν τον ρόλο των επεξηγηματικών μεταβλητών για τη μεταβλητή απόκρισης $\boldsymbol{y}$. Συγκεκριμένα, το μοντέλο θα δίνεται ως εξής:

$$\begin{aligned} \boldsymbol{y} &\sim N(X\boldsymbol{\beta}, \sigma^2), \\ X\boldsymbol{\beta} &= 6 + 8x_1 + 3x_4 + 10x_7 - 12x_{12} + 4x_{15}, \\ \sigma^2 &= 1.5^2, \end{aligned}$$

όπου $X$ ο πίνακας σχεδιασμού, διαστάσεων $n \times p$ με $n = 50$ (ο αριθμός των παρατηρήσεων) και $p = 16$ ($15$ συμμεταβλητές και ο σταθερός όρος). Από το γινόμενο $X\boldsymbol{\beta}$ διαπιστώνουμε πως το διάνυσμα $\boldsymbol{\beta}$ των παραμέτρων της παλινδρόμησης, το οποίο θα επιχειρήσουμε να εκτιμήσουμε Μπεϋζιανά, θα είναι:

$$\boldsymbol{\beta} = (6, 8, 0, 0, 3, 0, 0, 10, 0, 0, 0, 0, -12, 0, 0, 4). \tag{79}$$

Η υλοποίηση των παραπάνω θα γίνει με τη χρήση της εντολής `rnorm(k,a,b)` της R, η οποία παράγει `k` το πλήθος τυχαίους αριθμούς από την κανονική κατανομή με μέση τιμή `a` και τυπική απόκλιση `b`. Για τον πλήρη κώδικα που αναπτύχθηκε παραπέμπουμε στο Α.1.

**Δεύτερο σύνολο δεδομένων**

Για το δεύτερο σετ δεδομένων, θα χρησιμοποιήσουμε την ίδια μέθοδο με αυτή που ακολουθήθηκε για το πρώτο, με τη διαφορά ότι τώρα η διασπορά θα είναι $\sigma^2 = 2.5^2$. Επομένως το μοντέλο διαμορφώνεται ως εξής:

$$\begin{aligned} x_i &\sim N(0,1), i = 1, ..., 15, \\ \boldsymbol{y} &\sim N(X\boldsymbol{\beta}, \sigma^2), \\ X\boldsymbol{\beta} &= 6 + 8x_1 + 3x_4 + 10x_7 - 12x_{12} + 4x_{15}, \\ \sigma^2 &= 2.5^2. \end{aligned}$$

**Τρίτο σύνολο δεδομένων**

Στο τρίτο σετ που προσομοιώνουμε, η μεθοδολογία παραμένει ίδια για τις $10$ πρώτες συμμεταβλητές, η οποίες εξακολουθούν να κατανέμονται σύμφωνα με την τυποποιημένη κανονική κατανομή:

$$x_i \sim N(0,1), i = 1, ..., 10.$$

Ωστόσο, έχουμε διαφοροποίηση στην παραγωγή των τελευταίων $5$ συμμεταβλητών οι οποίες αντί να προέρχονται από την τυποποιημένη κανονική κατανομή, κατανέμονται κανονικά με διασπορά ίση με $1$ και μέση τιμή που προκύπτει ως γραμμικός συνδυασμός των επεξηγηματικών μεταβλητών $x_1$ έως $x_5$, δηλαδή:

$$\begin{aligned} x_{i'} &\sim N(\boldsymbol{\mu}', 1), i' = 11, ..., 15, \\ \boldsymbol{\mu}' &= 0.3x_1 + 0.5x_2 + 0.7x_3 + 0.9x_4 + 1.1x_5. \end{aligned}$$

Η δομή αυτή, δημιουργεί γραμμική εξάρτηση μεταξύ των πρώτων και των τελευταίων $5$ επεξηγηματικών μεταβλητών και μας προϊδεάζει για τις πιθανές επιδράσεις της στην προσαρμογή και ερμηνεία του μοντέλου παλινδρόμησης. Ακολουθεί ο Πίνακας 1 με τις τιμές του συντελεστή συσχέτισης για τα ζεύγη μεταβλητών $(x_k, x_l), k = 1, 2, 3, 4, l = 11, 12, 13, 14, 15$:

**Τρίτο σύνολο προσομοιωμένων δεδομένων**

| | $\boldsymbol{x_{11}}$ | $\boldsymbol{x_{12}}$ | $\boldsymbol{x_{13}}$ | $\boldsymbol{x_{14}}$ | $\boldsymbol{x_{15}}$ |
|---|---|---|---|---|---|
| $\boldsymbol{x_1}$ | 0.44 | 0.44 | 0.36 | 0.37 | 0.33 |
| $\boldsymbol{x_2}$ | 0.13 | 0.28 | 0.27 | 0.33 | 0.18 |
| $\boldsymbol{x_3}$ | 0.27 | 0.22 | 0.22 | 0.28 | 0.13 |
| $\boldsymbol{x_4}$ | 0.61 | 0.59 | 0.62 | 0.57 | 0.59 |
| $\boldsymbol{x_5}$ | 0.66 | 0.58 | 0.58 | 0.53 | 0.69 |

Πίνακας 1: Συντελεστές συσχέτισης για τα ζεύγη $(x_k, x_l), k = 1, ..., 4, l = 11, ..., 15$ στο τρίτο προσομοιωμένο σύνολο δεδομένων.

Όπως αναμενόταν, οι τιμές του συντελεστή φανερώνουν θετική γραμμική συσχέτιση στα παραπάνω ζεύγη με την εντονότερη να παρατηρείται στα ζεύγη $(x_4, x_l)$ και $(x_5, x_l)$. Αυτή

η παρατήρηση εξηγείται από το γεγονός ότι οι μεταβλητές $x_4$ και $x_5$ έχουν τη μεγαλύτερη συνεισφορά στο γραμμικό συνδυασμό που παράγει τα $x_l, l = 11, 12, 13, 14, 15$ αφού πολλαπλασιάζονται με συντελεστή $0.9$ και $1.1$ αντίστοιχα.

Όσον αφορά τη μεταβλητή απόκρισης, γεννάται με πανομοιότυπο τρόπο με αυτόν που χρησιμοποιήθηκε στο πρώτο σύνολο δεδομένων:

$$\begin{aligned} \boldsymbol{y} &\sim N(X\boldsymbol{\beta}, \sigma^2), \\ X\boldsymbol{\beta} &= 6 + 8x_1 + 3x_4 + 10x_7 - 12x_{12} + 4x_{15}, \\ \sigma^2 &= 1.5^2. \end{aligned}$$

**Τέταρτο σύνολο δεδομένων**

Όπως το δεύτερο σετ προέκυψε από μια τροποποίηση του πρώτου, έτσι και το τέταρτο σύνολο θα ταυτίζεται σχεδόν με το τρίτο, με μόνη διαφορά πως επιλέγουμε τυπική απόκλιση ίση με $2.5$ αντί του $1.5$. Έτσι έχουμε τη μορφή του τελευταίου μοντέλου προσομοιωμένων δεδομένων:

$$\begin{aligned} x_i &\sim N(0, 1), i = 1, ..., 10, \\ x_{i'} &\sim N(\boldsymbol{\mu}', 1), i' = 11, ..., 15, \\ \boldsymbol{\mu}' &= 0.3x_1 + 0.5x_2 + 0.7x_3 + 0.9x_4 + 1.1x_5, \\ \boldsymbol{y} &\sim N(X\boldsymbol{\beta}, \sigma^2), \\ X\boldsymbol{\beta} &= 6 + 8x_1 + 3x_4 + 10x_7 - 12x_{12} + 4x_{15}, \\ \sigma^2 &= 2.5^2. \end{aligned}$$

Φυσικά, και σε αυτή την περίπτωση, είμαστε προετοιμασμένοι για το φαινόμενο της πολυσυγγραμικότητας που θα προκαλέσει η ύπαρξη των εξαρτημένων από τα $x_i, i = 1, .., 5$ μεταβλητών κατά την προσαρμογή του μοντέλου. Παρακάτω, ο Πίνακας 2 δίνει τις τιμές του συντελεστή συσχέτισης για τα ζεύγη των γραμμικά εξαρτημένων μεταβλητών. Τα αποτελέσματα προσομοιάζουν αυτά που σχολιάστηκαν και παραπάνω, στον Πίνακα 1 για το τρίτο δείγμα, με τις υψηλότερες τιμές του συντελεστή να εμφανίζονται στα ζεύγη που αφορούν τις μεταβλητές $x_4$ και $x_5$.

**Τέταρτο σύνολο προσομοιωμένων δεδομένων**

| | $\boldsymbol{x_{11}}$ | $\boldsymbol{x_{12}}$ | $\boldsymbol{x_{13}}$ | $\boldsymbol{x_{14}}$ | $\boldsymbol{x_{15}}$ |
|---|---|---|---|---|---|
| $\boldsymbol{x_1}$ | 0.08 | -0.06 | 0.12 | 0.11 | 0.22 |
| $\boldsymbol{x_2}$ | 0.06 | 0.16 | -0.05 | 0.08 | 0.02 |
| $\boldsymbol{x_3}$ | 0.23 | 0.36 | 0.36 | 0.42 | 0.31 |
| $\boldsymbol{x_4}$ | 0.48 | 0.46 | 0.47 | 0.56 | 0.52 |
| $\boldsymbol{x_5}$ | 0.70 | 0.60 | 0.64 | 0.53 | 0.64 |

Πίνακας 2: Συντελεστές συσχέτισης για τα ζεύγη $(x_k, x_l), k = 1, ..., 4, l = 11, ..., 15$ στο τέταρτο προσομοιωμένο σύνολο δεδομένων.

### 4.1.2 Προσαρμογή Μπεϋζιανών μοντέλων παλινδρόμησης

Τα τέσσερα σύνολα δεδομένων που προσομοιώθηκαν θα μελετηθούν στα πλαίσια της Μπεϋζιανής παλινδρόμησης. Με τη βοήθεια της Μπεϋζιανής μεθοδολογίας θα επιχειρήσουμε να εκτιμήσουμε το γραμμικό μοντέλο που δημιούργησε τα δεδομένα, δηλαδή το διάνυσμα $\boldsymbol{\beta}$. Εφόσον γνωρίζουμε την πραγματική τιμή του $\boldsymbol{\beta}$, μπορούμε πολύ εύκολα να αξιολογήσουμε την απόδοση και να υπολογίσουμε την απόκλιση της εκτίμησης.

Στο κάθε σετ δεδομένων θα εφαρμοστούν τρία είδη πρότερων κατανομών: κανονική-αντίστροφη γάμμα από κοινού πρότερη, improper πρότερη για τη διασπορά και μη-πληροφοριακή πρότερη, σύμφωνα με τα ευρήματα της Ενότητας 3. Κάθε επιλογή πρότερης συνοδεύεται από πίνακες με τα αποτελέσματα των εκτιμητών σε κάθε σύνολο. Η ανάλυση συμπληρώνεται με πίνακες που αναγράφουν το σχετικό επί τοις εκατό σφάλμα εκτίμησης των (μη-μηδενικών) συντελεστών για κάθε σύνολο και κάθε επιλογή πρότερης.

Αφού το κεντρικό σημείο ενδιαφέροντος της μελέτης είναι ο προσδιορισμός της μέσης τιμής του $\boldsymbol{\beta}$ ύστερα από το Bayesian updating, αρκεί να υπολογίσουμε σε κάθε περίπτωση το διάνυσμα $\tilde{\boldsymbol{m}}$, όπως προκύπτει από τον συμβολισμό της προαναφερθείσας Ενότητας. Η διαδικασία αυτή απλουστεύεται μέσω των αναλυτικών τύπων που έχουν δοθεί για τον άμεσο υπολογισμό του $\tilde{\boldsymbol{m}}$.

Υπενθυμίζουμε πως, παρά τις διαφορές τους, και τα τέσσερα σύνολα χαρακτηρίζονται από το ίδιο διάνυσμα παραμέτρων $\boldsymbol{\beta}$. Επομένως, ο στόχος κάθε εκτίμησης είναι η καλύτερη δυνατή προσέγγιση των τιμών:

$$\boldsymbol{\beta} = (6, 8, 0, 0, 3, 0, 0, 10, 0, 0, 0, 0, -12, 0, 0, 4).$$

**Κανονική-αντίστροφη γάμμα από κοινού πρότερη**

Αρχικά, χρησιμοποιούμε τη NIG πρότερη που παρουσιάστηκε στην Υποενότητα 3.1 και αξιοποιούμε τις προτάσεις του Zellner για την επιλογή των $\boldsymbol{m}$ και $V$ σύμφωνα με την Εξίσωση (51). Θέτοντας σε εφαρμογή όσα έχουν ήδη δειχθεί για αυτή την περίπτωση πρότερης κατανομής, λαμβάνουμε άμεσα από την Εξίσωση (33) την εκ των υστέρων τιμή του διανύσματος της παραμέτρου θέσης, $\tilde{\boldsymbol{m}}$. Για τις παραμέτρους $a, b$ του IG μέρους της πρότερης επιλέγουμε μικρές τιμές, και συγκεκριμένα $a = 0.001 = b$, αφού δεν γνωρίζουμε εκ των προτέρων την κατανομή του $\sigma^2$ και η επιλογή τέτοιων τιμών συμβολίζει αυτή μας την άγνοια.

Στον Πίνακα 3 που ακολουθεί φαίνονται οι τιμές των συνιστωσών του $\tilde{\boldsymbol{m}}$, δηλαδή των $\tilde{m}_i$, όπως προέκυψαν από την προσαρμογή μοντέλου παλινδρόμησης με την συγκεκριμένη πρότερη, στο εκάστοτε δείγμα. Η τελευταία στήλη του πίνακα αποτελείται από τις πραγματικές τιμές των παραμέτρων της παλινδρόμησης που παρήγαγαν τα δεδομένα.

Εκ πρώτης όψεως, η NIG πρότερη φαίνεται να δίνει ικανοποιητικά αποτελέσματα. Οι μη-μηδενικές συνιστώσες του $\boldsymbol{\beta}$ προσεγγίζονται με ακρίβεια σε μεγάλο βαθμό. Ακόμα, παρόλο που οι εκτιμήσεις των μηδενικών συνιστωσών δεν είναι ακριβώς ίσες με μηδέν, λαμβάνουν γενικά μικρές τιμές, μακριά από τη μονάδα. Εξαίρεση αποτελούν οι εκτιμήσεις $\tilde{m}_5$ στο τρίτο και τέταρτο σετ δεδομένων με τιμές $0.97$ και $0.88$ αντίστοιχα, όμως ακόμα και σε αυτή την περίπτωση τα νούμερα δεν είναι ανησυχητικά.

Συμβουλευόμενοι τους Πίνακες 6 και 7 παρακάτω, διαπιστώνουμε πως συνολικά, για τη NIG πρότερη, οι μεγαλύτερες αποκλίσεις από τις πραγματικές τιμές των συντελεστών παλινδρόμησης παρατηρούνται στο δείγμα $4$ ενώ αντίστοιχα, η μικρότερη συνολική απόκλιση

| | NIG prior | | | | |
|---|---|---|---|---|---|
| $\tilde{m}_i$ | **1ο Σετ** $(\sigma^2 = 1.5^2)$ | **2ο Σετ** $(\sigma^2 = 2.5^2)$ | **3ο Σετ** $(\sigma^2 = 1.5^2)$ | **4ο Σετ** $(\sigma^2 = 2.5^2)$ | $\beta_i$ |
| $\tilde{m}_0$ | 5.71 | 6.19 | 6.02 | 5.22 | 6 |
| $\tilde{m}_1$ | 8.19 | 8.33 | 8.12 | 7.45 | 8 |
| $\tilde{m}_2$ | $-0.22$ | $-0.57$ | 0.29 | $-0.63$ | 0 |
| $\tilde{m}_3$ | 0.29 | 0 | 0.63 | $-0.58$ | 0 |
| $\tilde{m}_4$ | 2.86 | 2.78 | 3.81 | 2.55 | 3 |
| $\tilde{m}_5$ | 0.05 | $-0.06$ | 0.97 | $-0.88$ | 0 |
| $\tilde{m}_6$ | $-0.11$ | 0.26 | 0.20 | $-0.32$ | 0 |
| $\tilde{m}_7$ | 9.80 | 9.64 | 9.93 | 10.58 | 10 |
| $\tilde{m}_8$ | $-0.53$ | 0.41 | $-0.15$ | $-0.39$ | 0 |
| $\tilde{m}_9$ | $-0.28$ | 0.12 | 0.12 | 0.10 | 0 |
| $\tilde{m}_{10}$ | 0.29 | 0.28 | $-0.12$ | $-0.33$ | 0 |
| $\tilde{m}_{11}$ | $-0.50$ | 0.11 | $-0.42$ | 0.19 | 0 |
| $\tilde{m}_{12}$ | $-12.21$ | $-11.87$ | $-12.40$ | $-11.12$ | $-12$ |
| $\tilde{m}_{13}$ | $-0.15$ | $-0.25$ | $-0.06$ | $-0.46$ | 0 |
| $\tilde{m}_{14}$ | $-0.07$ | 0.70 | 0.12 | $-0.21$ | 0 |
| $\tilde{m}_{15}$ | 3.50 | 3.51 | 3.83 | 4.50 | 4 |

Πίνακας 3: Οι εκτιμώμενες τιμές των παραμέτρων της παλινδρόμησης με τη χρήση της κανονικής-αντίστροφης γάμμα πρότερης στα προσομοιωμένα δεδομένα.

παρατηρείται για τα δεδομένα του δείγματος 2.

**Improper πρότερη για τη διασπορά**

Συνεχίζουμε τη μελέτη των προσομοιωμένων δεδομένων εφαρμόζοντας τη μεθοδολογία της Υποενότητας 3.2, δηλαδή εισάγοντας improper πρότερη για τη διασπορά $\sigma^2$ και συγκεκριμένα:

$$p(\sigma^2) \propto \frac{1}{\sigma^2}$$

όπως δείξαμε και στην Εξίσωση (59). Όσον αφορά την πρότερη επιλογή της παραμέτρου θέσης $\boldsymbol{m}$ και του πίνακα συνδιασποράς $V$ της από κοινού κατανομής $p(\boldsymbol{\beta}, \sigma^2)$ θα χρησιμοποιήσουμε και πάλι την προσέγγιση του Zellner.

Η improper πρότερη οδηγεί συνολικά σε διαφορετική ύστερη από αυτή που οδηγούμαστε μέσω της NIG πρότερης. Εντούτοις, ο τύπος υπολογισμού του $\tilde{\boldsymbol{m}}$ είναι πανομοιότυπος και στις δύο περιπτώσεις πρότερων και εξαρτάται από την επιλογή των $\boldsymbol{m}, V$, σύμφωνα με την εξίσωση (33):

$$\tilde{\boldsymbol{m}} = \tilde{V}\left(V^{-1}\boldsymbol{m} + X'\boldsymbol{y}\right).$$

Συνεπώς, αφού έχουμε επιλέξει ίδιες τιμές για τα $\boldsymbol{m}, V$ τόσο στην NIG πρότερη όσο και στην improper, αναμένεται να οδηγηθούμε στην ίδια ύστερη τιμή της παραμέτρου θέσης, δηλαδή στις ίδιες εκτιμώμενες τιμές των παραμέτρων της παλινδρόμησης.

Ο Πίνακας 4 που ακολουθεί επιβεβαιώνει αυτή την πρόβλεψη καθώς ταυτίζεται απολύτως με τα αποτελέσματα του Πίνακα 3 σε όλα τα σετ δεδομένων. Αντίστοιχα, ο σχολιασμός των προκυπτουσών τιμών $\tilde{m}_i$ δεν διαφέρει από όσα ήδη επισημάνθηκαν για τις εκτιμήσεις που έδωσε η προηγούμενη πρότερη.

| | **Improper prior for** $\sigma^2$ | | | | |
|---|---|---|---|---|---|
| $\tilde{m}_i$ | **1ο Σετ** $(\sigma^2 = 1.5^2)$ | **2ο Σετ** $(\sigma^2 = 2.5^2)$ | **3ο Σετ** $(\sigma^2 = 1.5^2)$ | **4ο Σετ** $(\sigma^2 = 2.5^2)$ | $\beta_i$ |
| $\tilde{m}_0$ | 5.71 | 6.19 | 6.02 | 5.22 | 6 |
| $\tilde{m}_1$ | 8.19 | 8.33 | 8.12 | 7.45 | 8 |
| $\tilde{m}_2$ | −0.22 | −0.57 | 0.29 | −0.63 | 0 |
| $\tilde{m}_3$ | 0.29 | 0 | 0.63 | −0.58 | 0 |
| $\tilde{m}_4$ | 2.86 | 2.78 | 3.81 | 2.55 | 3 |
| $\tilde{m}_5$ | −0.05 | −0.06 | 0.97 | −0.88 | 0 |
| $\tilde{m}_6$ | −0.11 | 0.26 | 0.20 | −0.32 | 0 |
| $\tilde{m}_7$ | 9.80 | 9.64 | 9.93 | 10.58 | 10 |
| $\tilde{m}_8$ | −0.53 | 0.41 | −0.15 | −0.39 | 0 |
| $\tilde{m}_9$ | −0.28 | 0.12 | 0.12 | 0.10 | 0 |
| $\tilde{m}_{10}$ | 0.29 | 0.28 | −0.12 | −0.33 | 0 |
| $\tilde{m}_{11}$ | −0.50 | 0.12 | −0.42 | 0.19 | 0 |
| $\tilde{m}_{12}$ | −12.21 | −11.87 | −12.40 | −11.12 | −12 |
| $\tilde{m}_{13}$ | −0.15 | −0.25 | −0.06 | −0.46 | 0 |
| $\tilde{m}_{14}$ | −0.07 | 0.70 | 0.12 | −0.21 | 0 |
| $\tilde{m}_{15}$ | 3.50 | 3.51 | 3.83 | 4.50 | 4 |

Πίνακας 4: Οι εκτιμώμενες τιμές των παραμέτρων της παλινδρόμησης με τη χρήση improper πρότερης για τη διασπορά στα προσομοιωμένα δεδομένα.

**Μη πληροφοριακή πρότερη**

Οι δοκιμές με τα προσομοιωμένα δεδομένα κλείνουν με την προσαρμογή μοντέλου παλινδρόμησης όπου χρησιμοποιείται μη-πληροφοριακή από κοινού πρότερη για τις παραμέτρους $(\boldsymbol{\beta}, \sigma^2)$. Αυτή η επιλογή πρότερης αποτέλεσε το αντικείμενο μελέτης της Ενότητας 3.3 και συνεπώς έχουμε και πάλι στη διάθεσή μας αναλυτικούς τύπους για τον υπολογισμό των εκ των υστέρων τιμών των παραμέτρων.

Από τις Εξισώσεις (70)-(73) προκύπτουν απευθείας οι τιμές των παραμέτρων και των υπερπαραμέτρων ύστερα από το Bayesian updating. Ειδικότερα, η Εξίσωση (71) δίνει την εκ των υστέρων εκτίμηση των συντελεστών της παλινδρόμησης $\tilde{\boldsymbol{m}}$ που σε αυτή την περίπτωση ταυτίζεται με την εκτιμήτρια των ελαχίστων τετραγώνων. Συνεπώς, τα αποτελέσματα του μοντέλου σε κάθε δείγμα θα ταυτίζονται με αυτά της γνωστής παλινδρόμησης ελαχίστων τετραγώνων της κλασικής στατιστικής.

Όπως έγινε και για τις δύο πρότερες που προηγήθηκαν, ακολουθεί ο Πίνακας 5 με τις εκτιμήσεις των $\beta_i$ σε κάθε σύνολο που εφαρμόστηκε το εν λόγω μοντέλο.

Όσον αφορά τις μηδενικές παραμέτρους, και σύμφωνα με τον Πίνακα 5, οι τιμές των

| | **Non-informative prior** | | | | |
|---|---|---|---|---|---|
| $\tilde{m}_i$ | **1ο Σετ** $(\sigma^2 = 1.5^2)$ | **2ο Σετ** $(\sigma^2 = 2.5^2)$ | **3ο Σετ** $(\sigma^2 = 1.5^2)$ | **4ο Σετ** $(\sigma^2 = 2.5^2)$ | $\beta_i$ |
| $\tilde{m}_0$ | 5.83 | 6.31 | 6.14 | 5.32 | 6 |
| $\tilde{m}_1$ | 8.35 | 8.50 | 8.28 | 7.60 | 8 |
| $\tilde{m}_2$ | $-0.22$ | $-0.58$ | 0.29 | $-0.64$ | 0 |
| $\tilde{m}_3$ | 0.29 | 0 | 0.64 | $-0.59$ | 0 |
| $\tilde{m}_4$ | 2.92 | 2.83 | 3.88 | 2.60 | 3 |
| $\tilde{m}_5$ | $-0.05$ | $-0.06$ | 0.99 | $-0.90$ | 0 |
| $\tilde{m}_6$ | $-0.11$ | 0.26 | 0.21 | $-0.33$ | 0 |
| $\tilde{m}_7$ | 10.00 | 9.84 | 10.13 | 10.79 | 10 |
| $\tilde{m}_8$ | $-0.54$ | 0.42 | $-0.15$ | $-0.39$ | 0 |
| $\tilde{m}_9$ | $-0.29$ | 0.12 | 0.13 | 0.10 | 0 |
| $\tilde{m}_{10}$ | 0.29 | 0.29 | $-0.13$ | $-0.33$ | 0 |
| $\tilde{m}_{11}$ | $-0.51$ | 0.12 | $-0.43$ | 0.20 | 0 |
| $\tilde{m}_{12}$ | $-12.46$ | $-12.10$ | $-12.65$ | $-11.34$ | $-12$ |
| $\tilde{m}_{13}$ | $-0.16$ | $-0.25$ | $-0.06$ | $-0.47$ | 0 |
| $\tilde{m}_{14}$ | $-0.07$ | 0.71 | 0.12 | $-0.22$ | 0 |
| $\tilde{m}_{15}$ | 3.57 | 3.58 | 3.90 | 4.59 | 4 |

Πίνακας 5: Οι εκτιμώμενες τιμές των παραμέτρων της παλινδρόμησης με τη χρήση μη-πληροφοριακής πρότερης στα προσομοιωμένα δεδομένα.

εκτιμητών τους είναι αρκετά κοντά στο μηδέν χωρίς να πλησιάζουν σημαντικά τη μονάδα. Για άλλη μια φορά, εξαίρεση αποτελεί το $\tilde{m}_5$ στο τρίτο και τέταρτο σύνολο δεδομένων που λαμβάνει τιμές $0,99$ και $-0,90$ αντίστοιχα.

Ανατρέχοντας στους Πίνακες 6 και 7, διαπιστώνουμε πως οι εκτιμητές των μη-μηδενικών παραμέτρων είναι ικανοποιητικοί με τη μεγαλύτερη απόκλιση να παρατηρείται στην εκτίμηση του $\beta_4$ στο τρίτο σετ, όπου το $\tilde{m}_4$ λαμβάνει τιμή κατά $29,60\%$ μεγαλύτερη. Η χαμηλότερη απόκλιση εντοπίζεται στην εκτίμηση του $\beta_7$ του πρώτου σετ όπου το $\tilde{m}_7$ απέχει μόλις $0,02\%$ από την επιθυμητή τιμή.

Ακόμα, η μη-πληροφοριακή πρότερη φαίνεται να σημειώνει μεγαλύτερες αποκλίσεις στα δύο τελευταία δείγματα, στα όποια έχουμε επισημάνει την ύπαρξη πολυσυγγραμμικότητας. Ωστόσο, οι εκτιμήσεις που προκύπτουν επιλέγοντας να μην εισάγουμε πρότερη πληροφορία στο μοντέλο ώστε να μην το επηρεάσουμε, κρίνονται στο σύνολό τους αποτελεσματικές ενώ δεν καταγράφονται αποκλίσεις που θα μπορούσαν να θεωρηθούν επιβλαβείς για τη δομή του μοντέλου.

### 4.1.3 Σύγκριση μεταξύ εκ των προτέρων κατανομών και συμπεράσματα

Οι Πίνακες 6 και 7 περιέχουν τα επί τοις εκατό σχετικά σφάλματα εκτίμησης των μη-μηδενικών συντελεστών για κάθε επιλογή πρότερης στο εκάστοτε προσομοιωμένο σύνολο δεδομένων, διευκολύνοντας έτσι τη συγκριτική αξιολόγηση των πρότερων. Επειδή, όπως προαναφέρθηκε, τα αποτελέσματα μέσω της NIG πρότερης ταυτίζονται με όσα προκύπτουν

από την improper πρότερη για τη διασπορά, στο εξής όποιο σχόλιο γίνεται σχετικά με την NIG πρότερη θα αναφέρεται επίσης και στην improper.

Ξεκινώντας από τα αποτελέσματα που αφορούν το πρώτο σετ προσομοιωμένων δεδομένων όπως φαίνονται στον Πίνακα 6, φαίνεται πως οι εκτιμήσεις από τη NIG πρότερη εμφανίζουν συνολικά ελαφρώς μεγαλύτερες αποκλίσεις από τις εκτιμήσεις της μη-πληροφοριακής πρότερης. Γενικότερα όμως οι τιμές των αποκλίσεων είναι μικρές και μόνο η εκτίμηση του $\beta_{15}$, η οποία παρουσιάζει και τη μικρότερη ακρίβεια στο συγκεκριμένο σετ και για τις δύο πρότερες, εμφανίζει σφάλμα μεγαλύτερο του $10\%$. Η ακριβέστερη εκτίμηση για την NIG πρότερη συναντάται στον συντελεστή της επεξηγηματικής μεταβλητής $x_{12}$ ενώ για τη μη-πληροφοριακή πρότερη στον συντελεστή του $x_7$.

| | **NIG prior** | | **Improper variance** | | **Non-informative** | |
|---|---|---|---|---|---|---|
| $\boldsymbol{\beta_i}$ | **1ο Σετ** | **2ο Σετ** | **1ο Σετ** | **2ο Σετ** | **1ο Σετ** | **2ο Σετ** |
| $\beta_0$ | 4.76 % | −3.19 % | 4.76 % | −3.19 % | 2.86 % | −5.26 % |
| $\beta_1$ | −2.36 % | −4.11 % | −2.36 % | −4.11 % | −4.41 % | −6.19 % |
| $\beta_4$ | 4.61 % | 7.43 % | 4.61 % | 7.43 % | 2.70 % | 5.58 % |
| $\beta_7$ | 1.98 % | 3.58 % | 1.98 % | 3.58 % | 0.02 % | 1.65 % |
| $\beta_{12}$ | −1.78 % | 1.12 % | −1.78 % | 1.12 % | −3.81 % | −0.86 % |
| $\beta_{15}$ | 12.56 % | −2.15 % | 12.56 % | −2.15 % | 10.81 % | 10.39 % |

Πίνακας 6: Επί τοις εκατό σχετικό σφάλμα εκτίμησης των μη-μηδενικών συντελεστών για κάθε μία από τις τρεις επιλογές πρότερων, στο πρώτο και δεύτερο προσομοιωμένο σύνολο δεδομένων.

Παρόμοια συμπεριφορά απαντάται και στο δεύτερο σετ προσομοιωμένων δεδομένων, όπου οι συνολικές αποκλίσεις των εκτιμήσεων είναι αυξημένες σε σχέση με το πρώτο, γεγονός που πιθανότατα οφείλεται στην αυξημένη διασπορά των τυχαίων μεταβλητών του δεύτερου σετ. Απόκλιση άνω του $10\%$ βλέπουμε και πάλι στην εκτίμηση του $\beta_{15}$, αυτή τη φορά μόνο από τη μη-πληροφοριακή πρότερη η οποία παρ' όλα αυτά έχει συνολικά ελαφρώς μικρότερες κατ' απόλυτη τιμή σφάλματα. Η εκτίμηση με την μικρότερη ακρίβεια είναι αυτή του $\beta_4$ για τη NIG πρότερη και του $\beta_{15}$ για τη μη-πληροφοριακή ενώ και οι δύο πρότερες εμφανίζουν τη μεγαλύτερη ακρίβεια εκτίμησης στον συντελεστή του $x_{12}$.

Το τρίτο σύνολο προσομοιωμένων δεδομένων είναι το μοναδικό από τα τέσσερα σύνολα στο οποίο η μη-πληροφοριακή πρότερη έχει χειρότερη απόδοση από την κανονική-αντίστροφη γάμμα ενώ μάλιστα φαίνεται να "φουσκώνει" τις τιμές των εκτιμητών. Σε αυτό το σύνολο παρατηρείται επίσης το μεγαλύτερο σφάλμα εκτίμησης, και για τις δύο επιλογές πρότερων. Συγκεκριμένα, η εκτίμηση του $\beta_4$ είναι κατά $27\%$ μεγαλύτερη από την πραγματική τιμή για τη NIG πρότερη και σχεδόν $30\%$ για τη μη-πληροφοριακή, καθιστώντας το σφάλμα αξιοσημείωτο. Στον αντίποδα, η κανονική-αντίστροφη γάμμα πρότερη εμφανίζει μεγάλη ακρίβεια στην προσέγγιση του σταθερού όρου ενώ η μη-πληροφοριακή επιτυγχάνει την καλύτερη προσέγγιση στον συντελεστή της επεξηγηματικής μεταβλητής $x_7$.

Στο τέταρτο σύνολο συναντάμε αυξημένες συνολικές αποκλίσεις σε σχέση με το τρίτο, όπως αντίστοιχα είδαμε αύξηση των σφαλμάτων από το πρώτο στο δεύτερο σύνολο. Στην πραγματικότητα, πρόκειται για το σετ με τις μεγαλύτερες παρατηρούμενες συνολικές αποκλίσεις,

με μικρό προβάδισμα της μη-πληροφοριακής πρότερης σε σχέση με την NIG ως προς την ακριβεία. Τα σφάλματα στην εκτίμηση των συντελεστών των μεταβλητών $x_4, x_{15}$ και του σταθερού όρου ξεπερνούν το $10\%$ για όλες τις πρότερες ενώ η ακριβέστερη εκτίμηση είναι αυτή του συντελεστή του $x_7$ και του $x_1$ για τη NIG και τη μη-πληροφοριακή πρότερη αντίστοιχα.

| | **NIG prior** | | **Improper variance** | | **Non-informative** | |
|---|---|---|---|---|---|---|
| $\boldsymbol{\beta_i}$ | **3ο Σετ** | **4ο Σετ** | **3ο Σετ** | **4ο Σετ** | **3ο Σετ** | **4ο Σετ** |
| $\beta_0$ | $-0.37$ % | 13.01 % | $-0.37$ % | 13.01 % | $-2.38$ % | 11.28 % |
| $\beta_1$ | $-1.48$ % | 6.85 % | $-1.48$ % | 6.85 % | $-3.51$ % | 4.99 % |
| $\beta_4$ | $-27.06$ % | 15.00 % | $-27.06$ % | 15.00 % | $-29.60$ % | 13.30 % |
| $\beta_7$ | 0.66 % | $-5.77$ % | 0.66 % | $-5.77$ % | $-1.33$ % | $-7.89$ % |
| $\beta_{12}$ | $-3.37$ % | 7.36 % | $-3.37$ % | 7.36 % | $-5.44$ % | 5.51 % |
| $\beta_{15}$ | 4.35 % | $-12.44$ % | 4.35 % | $-12.44$ % | 2.43 % | $-14.69$ % |

Πίνακας 7: Επί τοις εκατό σχετικό σφάλμα εκτίμησης των μη-μηδενικών συντελεστών για κάθε μία από τις τρεις επιλογές πρότερων, στο τρίτο και τέταρτο προσομοιωμένο σύνολο δεδομένων.

Επιστρέφοντας στους Πίνακες 1 και 2, είναι εμφανές πως η επεξηγηματική μεταβλητή $x_4$, που συμμετέχει ενεργά στο μοντέλο καθώς έχει μη-μηδενικό συντελεστή παλινδρόμησης, εμφανίζει υψηλή συσχέτιση με τις συμμεταβλητές $x_{11}$ έως $x_{15}$ στο τρίτο και τέταρτο σύνολο δεδομένων. Σε συνδυασμό με τα σχόλια που προηγήθηκαν αναφορικά με τις αποκλίσεις στις εκτιμήσεις των αντίστοιχων συντελεστών στα σύνολα αυτά, έχουμε βάσιμες υποψίες πως η γραμμική εξάρτηση των συμμεταβλητών επηρεάζει την ποιότητα της εκτίμησης.

Ακόμα, παρατηρήσαμε το μοτίβο της αύξησης των σφαλμάτων εκτίμησης από το πρώτο στο δεύτερο και από το τρίτο στο τέταρτο σύνολο δεδομένων, τα οποία έχουν κατασκευαστεί με τρόπο ώστε το δεύτερο μέλος κάθε ζεύγους να παρουσιάζει μεγαλύτερη διασπορά από το πρώτο. Θα μπορούσαμε λοιπόν να υποθέσουμε πως η αυξημένη διασπορά των τυχαίων μεταβλητών που απαρτίζουν το σύνολο δεδομένων είναι ένας ακόμη παράγοντας που υποβαθμίζει την ποιότητα της εκτίμησης των συντελεστών παλινδρόμησης.

Πέρα από τις επί μέρους παρατηρήσεις, η συνολική απόδοση των πρότερων στη δημιουργία εκτιμητών κρίνεται ικανοποιητική. Σε κάθε περίπτωση, η μεθοδολογία κατάφερε να εντοπίσει την υποβόσκουσα δομή των προσομοιωμένων δεδομένων μέσα από έναν μικρό αριθμό παρατηρήσεων. Συνεπώς, ολοκληρώνοντας την επεξεργασία και την ανάλυση του πιλοτικού αυτού παραδείγματος εφαρμογής της Μπεϋζιανής παλινδρόμησης, τα συμπεράσματα είναι θετικά. Οι όποιες επιπλοκές και αποκλίσεις μπορούν να θεωρηθούν δικαιολογημένες και αναμενόμενες όταν εργαζόμαστε με ποσότητες που χαρακτηρίζονται από στοχαστική συμπεριφορά. Εν ολίγοις, διαπιστώσαμε ιδίοις όμμασι πως η Μπεϋζιανή στατιστική παρέχει έναν αποτελεσματικό και αξιόπιστο μηχανισμό για την ανάλυση παλινδρόμησης, ακόμα και για δεδομένα με ιδιάζουσα συμπεριφορά.

## 4.2 Εφαρμογή σε πραγματικά δεδομένα: Εκτίμηση ποιότητας κρασιού

Η εφαρμογή της ανάλυσης στα προσομοιωμένα δεδομένα ήταν ένα σκαλοπάτι για τη μετάβαση από τη θεωρητική περιγραφή στην πρακτική εφαρμογή των μοντέλων. Οι μέθοδοι τώρα θα εφαρμοστούν σε πιο ρεαλιστικό πλαίσιο δηλαδή σε πραγματικά δεδομένα που έχουν συλλεχθεί χωρίς να είναι εξ αρχής γνωστή η αιτιακή σχέση μεταξύ συμμεταβλητών και απόκρισης.

Αντικείμενο μελέτης σε αυτή την ενότητα θα αποτελέσει η εκτίμηση της ποιότητας του κρασιού, σε κλίμακα από $0$ έως $10$, ως συνάρτηση $11$ φυσικοχημικών ιδιοτήτων τους. Πρόκειται για κρασιά Πορτογαλικής προέλευσης που προέρχονται από την τοπική ποικιλία vinho verde η οποία μάλιστα αποτελεί Προστατευόμενη Ονομασία Προέλευσης (Π.Ο.Π.) της επαρχίας Mihno στη βορειοδυτική Πορτογαλία. Τα δεδομένα συλλέχθηκαν χρησιμοποιώντας

| | | **Ελάχιστη τιμή** | | **Μέση τιμή** | | **Μέγιστη τιμή** | |
|---|---|---|---|---|---|---|---|
| **ΑΑ** | **Χαρακτηριστικό [μονάδα]** | **Λ** | **Κ** | **Λ** | **Κ** | **Λ** | **Κ** |
| 1 | Μη-πτητική οξύτητα (fixed acidity) [$g/dm^3$] | 3.8 | 4.6 | 6.855 | 8.32 | 14.2 | 15.9 |
| 2 | Πτητική οξύτητα (volatile acidity) [$g/dm^3$] | 0.08 | 0.12 | 0.2782 | 0.5278 | 1.1 | 1.58 |
| 3 | Κιτρικό οξύ (citric acid) [$g/dm^3$] | 0 | 0 | 0.3342 | 0.271 | 1.66 | 1 |
| 4 | Υπολείμματα ζάχαρης (residual sugar) [$g/dm^3$] | 0.6 | 0.9 | 6.391 | 2.539 | 65.8 | 15.5 |
| 5 | Χλωριούχα (chlorides) [$g/dm^3$] | 0.009 | 0.012 | 0.04577 | 0.08747 | 0.346 | 0.611 |
| 6 | Ελεύθερο διοξείδιο του θείου (free sulfur dioxide) [$mg/dm^3$] | 2 | 1 | 35.31 | 15.87 | 289 | 72 |
| 7 | Συνολικό διοξείδιο του θείου (total sulfur dioxide)[$mg/dm^3$] | 9 | 6 | 138.4 | 46.47 | 440 | 289 |
| 8 | Πυκνότητα (density) [$g/cm^3$] | 0.9871 | 0.9901 | 0.994 | 0.9967 | 1.039 | 1.0037 |
| 9 | pH | 2.72 | 2.74 | 3.188 | 3.311 | 3.82 | 4.01 |
| 10 | Θειϊκά άλατα (sulphates) [$g/dm^3$] | 0.22 | 0.33 | 0.4898 | 0.6581 | 1.08 | 2 |
| 11 | Αλκοόλ (alcohol) [vol.%] | 8 | 8.4 | 10.51 | 10.42 | 14.2 | 14.9 |
| - | Ποιότητα (quality) | 3 | 3 | 5.878 | 5.636 | 9 | 8 |

Πίνακας 8: Τα χαρακτηριστικά του συνόλου δεδομένων, για τα λευκά (Λ) και κόκκινα (Κ) κρασιά, συνοδευόμενα από περιγραφικά μέτρα για τις παρατηρήσεις.

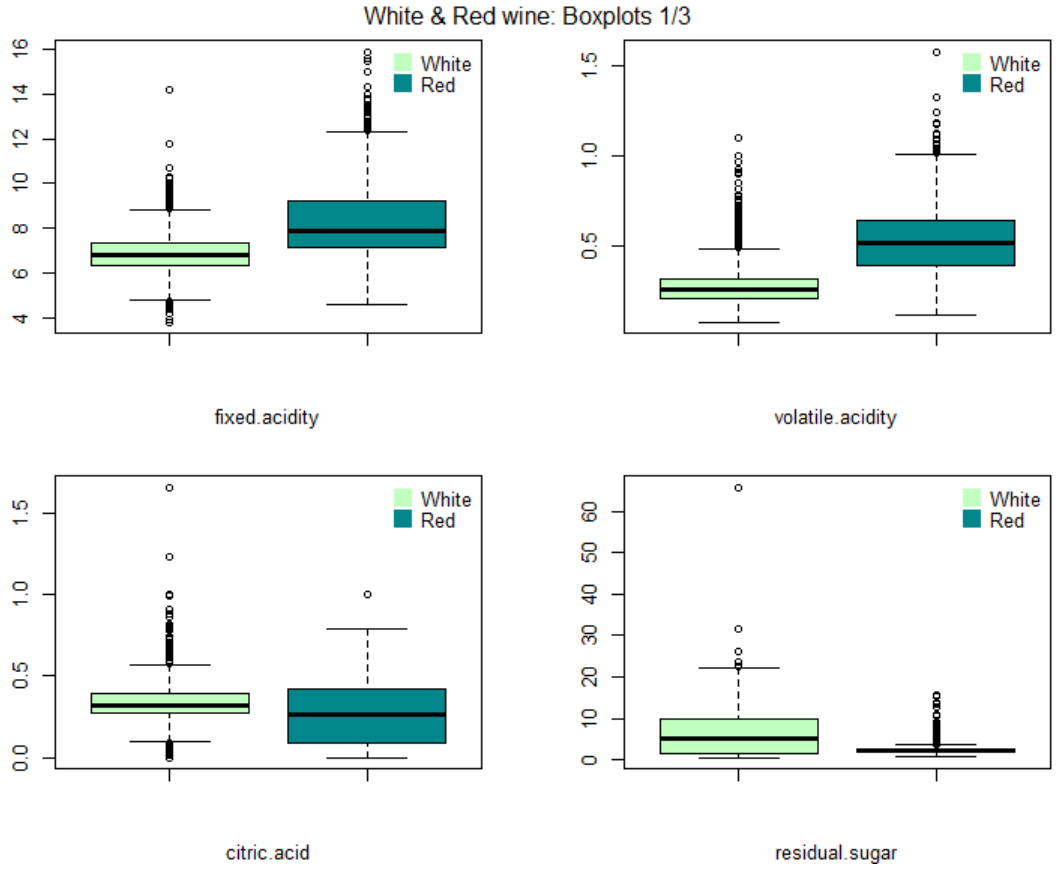
ειδικό αυτοματοποιημένο σύστημα από τον Comissão de Viticultura da Região dos Vinhos Verdes (CVRVV), οργανισμός αρμόδιος για τη διασφάλιση ποιότητας του vinho verde, και

ανακτήθηκαν για τους σκοπούς της παρούσας εργασίας μέσω του UCI Machine Learning Repository (Cortez, Cerdeira, Almeida, Matos, & Reis 2009), όπου και είναι δημοσίως διαθέσιμα.

### 4.2.1 Παρουσίαση δεδομένων

Το σύνολο δεδομένων περιέχει $4.898$ καταγραφές βαθμολογίας λευκού και $1.599$ καταγραφές βαθμολογίας κόκκινου κρασιού σε κλίμακα $0 - 10$, κάθε μία συνοδευόμενη από $11$ μετρήσεις φυσικοχημικών ιδιοτήτων του εκάστοτε κρασιού. Σημειώνεται πως θα προσαρμόσουμε ξεχωριστά μοντέλα για τα λευκά και κόκκινα κρασιά της ποικιλίας vinho verde.

Στον Πίνακα 8 παρουσιάζονται τα χαρακτηριστικά που περιέχονται στο σύνολο δεδομένων μαζί με περιγραφικά μέτρα για τις διαθέσιμες παρατηρήσεις αυτών, για λευκά και κόκκινα κρασιά ξεχωριστά. Στον πίνακα έχουμε συμπεριλάβει και την μεταβλητή απόκρισης που αφορά στην ποιότητα (quality) του κρασιού. Η αρίθμηση του κάθε χαρακτηριστικού θα διατηρηθεί και ως δείκτης της εκάστοτε επεξηγηματικής μεταβλητής στην οποία θα μετατραπεί. Στη συνέχεια δίνεται ένας σύντομος ορισμός κάθε χαρακτηριστικού του συνόλου δεδομένων, ενώ παρεμβάλλονται και θηκογράμματα για κάθε μεταβλητή, (βλ. Διαγράμματα 1-4).

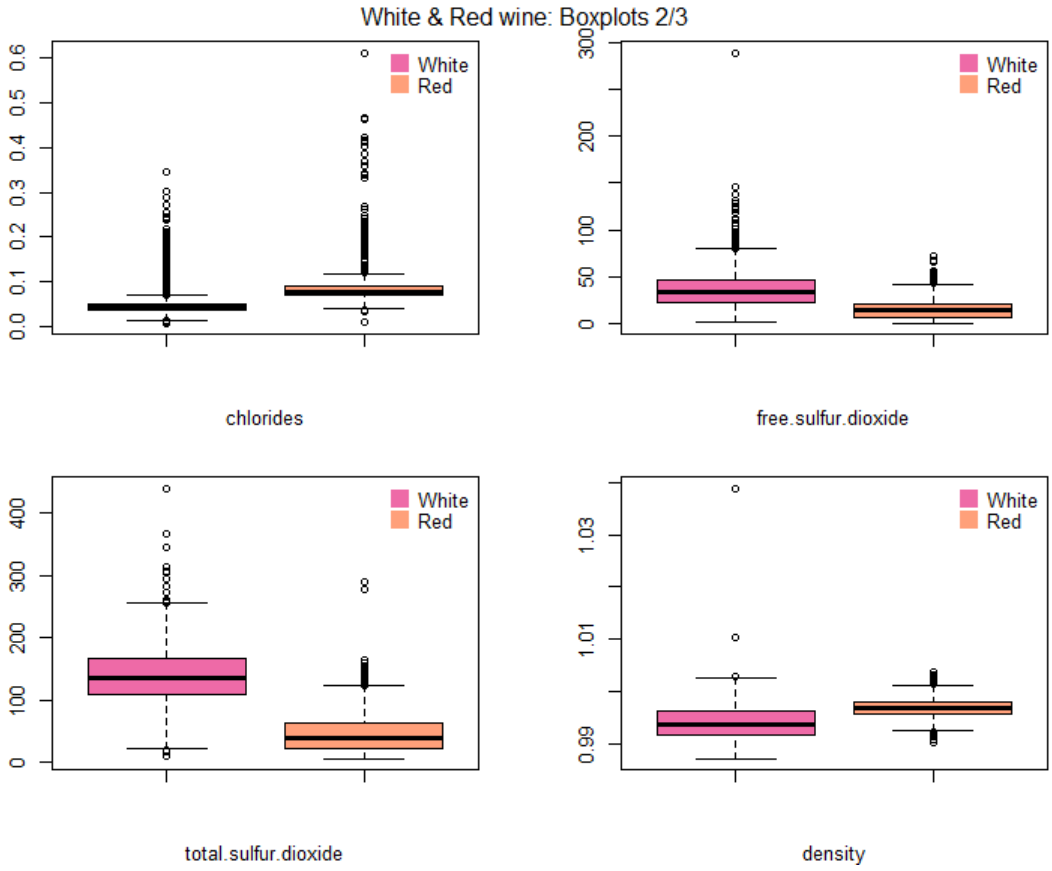



Διάγραμμα 1: Θηκόγραμμα για τις τέσσερις πρώτες επεξηγηματικές μεταβλητές για λευκά (White) και κόκκινα (Red) κρασιά.

Ο σχολιασμός αυτός αποσκοπεί στην ευκολότερη, και με βάση τη λογική, αξιολόγηση των αποτελεσμάτων αφού θα έχουμε πάρει μια πρώτη γεύση από τις αιτιακές σχέσεις που κρύβουν τα δεδομένα. Ειδικότερα:

1. Μη-πτητική οξύτητα (fixed acidity)
   Η μη-πτητική οξύτητα οφείλεται στην παρουσία οξέων που δεν εξατμίζονται εύκολα. Γενικότερα θέλουμε να υπάρχουν σε μεγάλη ποσότητα στο κρασί, ειδάλλως θα είναι άγευστο. Ωστόσο, υπερβολικά μεγάλες τιμές προκαλούν ξινή γεύση. Στα δεδομένα που μελετάμε, η μέτρηση αφορά συγκεκριμένα το τρυγικό οξύ (tartaric acid), ένα από τα είδη οξέων που συμβάλλουν στην αυξημένη τιμή μη-πτητικής οξύτητας.

2. Πτητική οξύτητα (volatile acidity)
   Η πτητική οξύτητα αφορά τα οξέα του κρασιού που βρίσκονται σε αέρια, όχι υγρή κατάσταση, με αποτέλεσμα να είναι ευκολότερα ανιχνεύσιμα στη μυρωδιά παρά στη γεύση. Ο κύριος παράγοντας της πτητικής οξύτητας είναι το οξικό οξύ (acetic acid), το γνωστό ξύδι και επομένως δεν είναι επιθυμητή η παρουσία του σε μεγάλη ποσότητα. Μάλιστα, η υψηλή πτητική οξύτητα αποτελεί ένδειξη πως το κρασί είναι χαλασμένο.




Διάγραμμα 2: Θηκόγραμμα για τις επεξηγηματικές μεταβλητές 5 έως 8, για λευκά (White) και κόκκινα (Red) κρασιά.

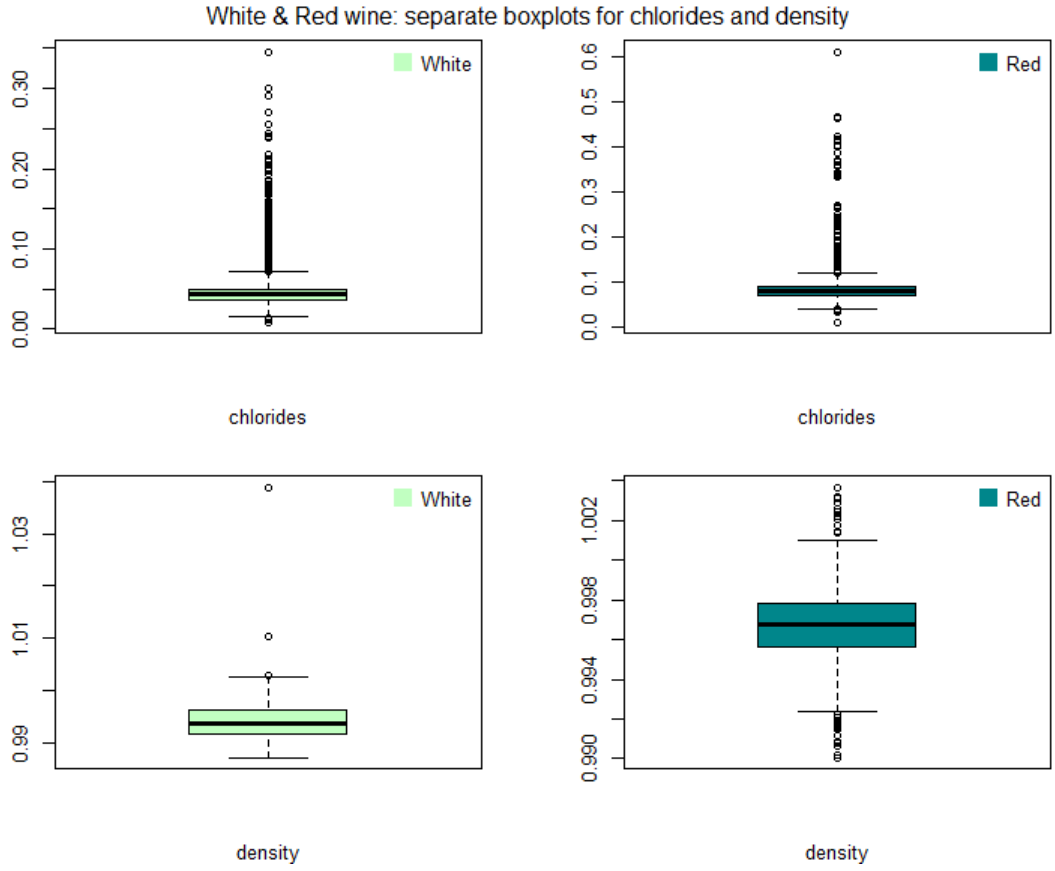
3. Κιτρικό οξύ (citric acid)
   Το κιτρικό οξύ χαρίζει γεύση και φρεσκάδα στο κρασί.

4. Υπολείμματα ζάχαρης (residual sugar)
   Πρόκειται για τα φυσικά σάκχαρα του σταφυλιού που παραμένουν αυτούσια στο κρασί όταν ολοκληρώνεται η ζύμωση του αλκοόλ.

5. Χλωριούχα (chlorides)
   Η ποσότητα αλάτων στο κρασί.

6. Ελεύθερο διοξείδιο του θείου (free sulfur dioxide)
   Το ελεύθερο διοξείδιο του θείου συμβάλλει στην πρόληψη της οξείδωσης και της ανάπτυξης μικροβίων. Είναι καλό να υπάρχει, αλλά όχι σε υπερβολική ποσότητα.

7. Συνολικό διοξείδιο του θείου (total sulfur dioxide)
   Πρόκειται για το σύνολο των διοξειδίων του θείου, σε ελεύθερη αλλά και δεσμευμένη μορφή. Η παρουσία του σε ποσότητες άνω των $50\ ppm$ γίνεται αισθητή, και όχι ευχάριστα, στη γεύση.

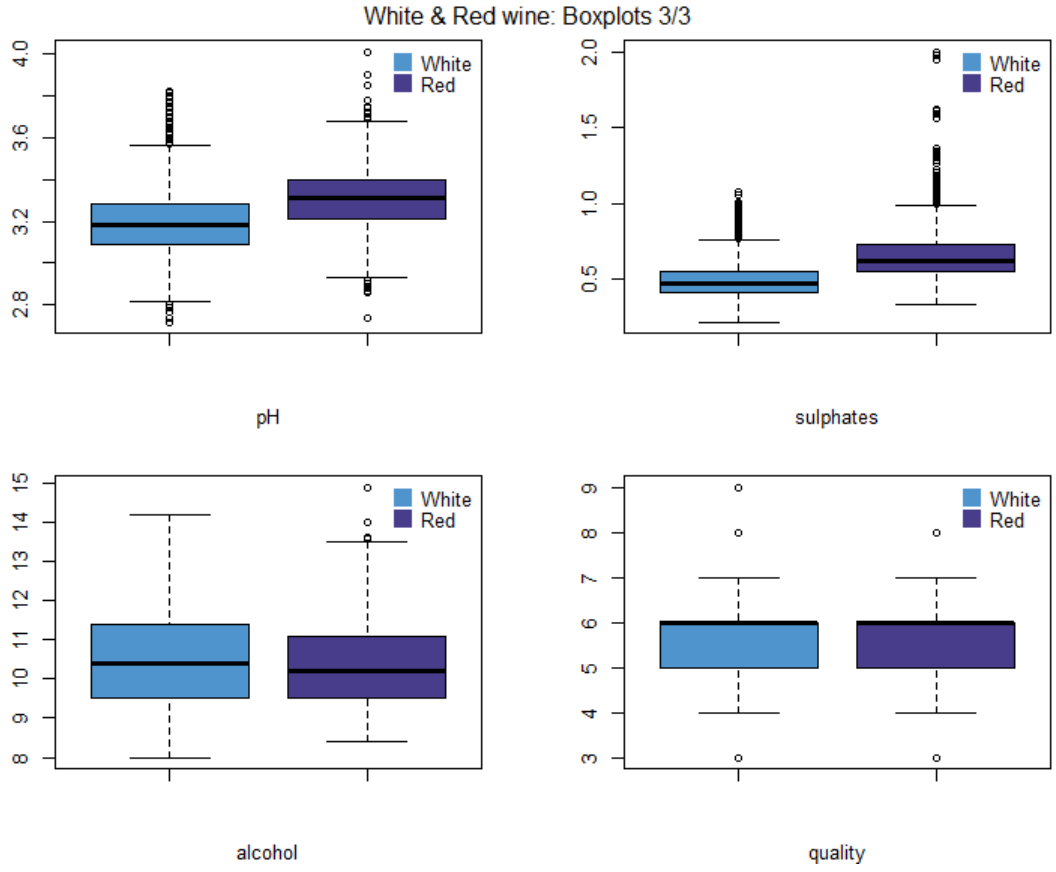



Διάγραμμα 3: Θηκόγραμμα για τις επεξηγηματικές μεταβλητές chlorides και density, ξεχωριστά, για λευκά (White) και κόκκινα (Red) κρασιά.

8. Πυκνότητα (density)
   Η πυκνότητα του κρασιού είναι σχετικά κοντά σε αυτή του νερού, δηλαδή πέριξ του $1$ $g/cm^3$, αλλά εξαρτάται από τα επίπεδα ζάχαρης και αλκοόλ.

9. Θειϊκά άλατα (sulphates)
   Τα θειϊκά άλατα αποτελούν πρόσθετα του κρασιού με αντιμικροβιακή και αντιοξειδωτική λειτουργία.

10. Αλκοόλ (alcohol)
    Τα συνηθέστερα επίπεδα αλκοόλ στο κρασί είναι μεταξύ $11$ και $13\%$, ενώ το φάσμα του εκτείνεται από $5.5$ μέχρι και $20\%$.



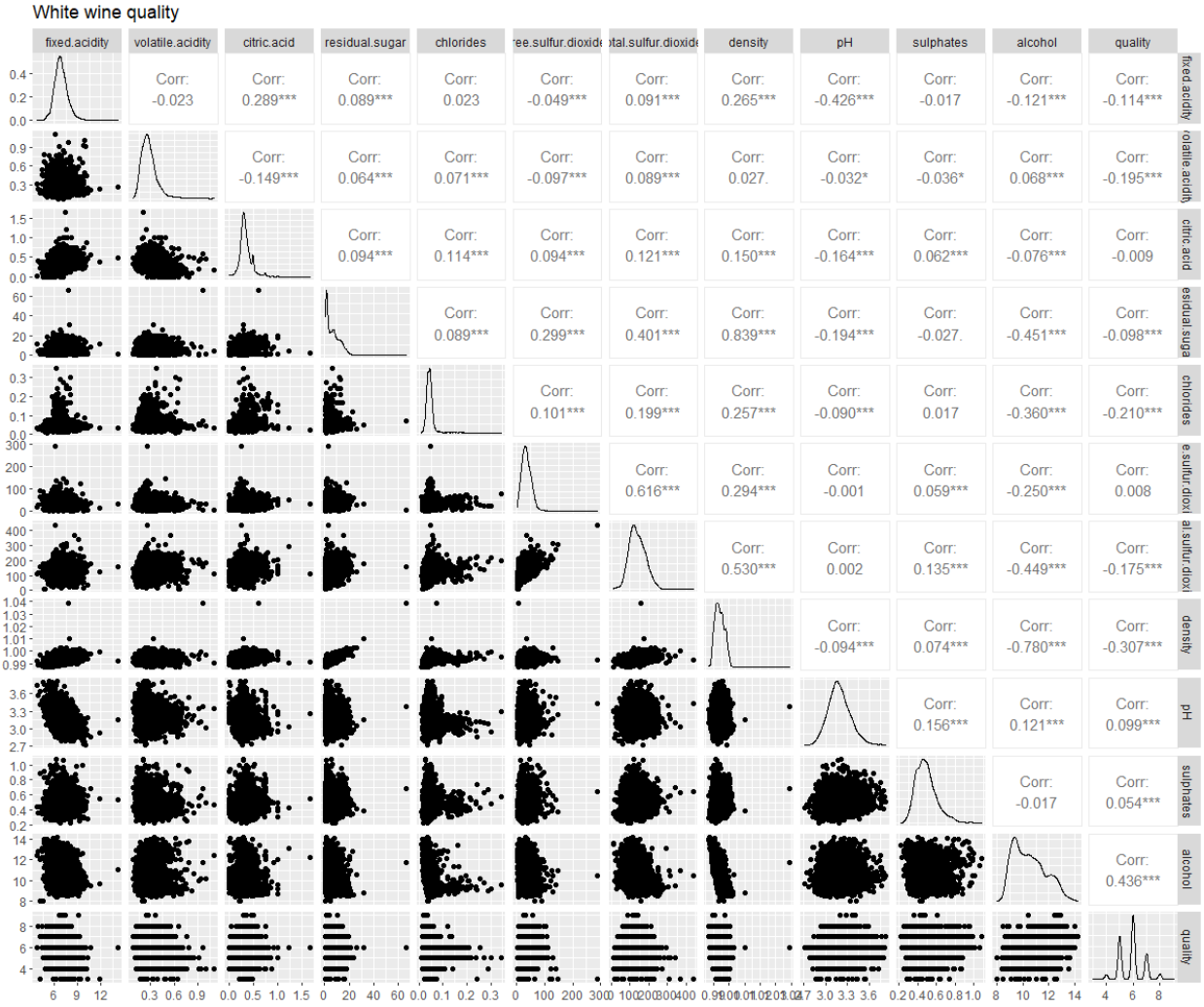

Διάγραμμα 4: Θηκόγραμμα για τις επεξηγηματικές μεταβλητές 9 έως 11 για λευκά (White) και κόκκινα (Red) κρασιά.

11. pH
    Το κρασί είναι γενικά όξινο διάλυμα οπότε το pH του παίρνει τιμές κάτω από το $7$, συνήθως μεταξύ $3$ και $4$.

12. Ποιότητα (quality)
    Πρόκειται για τη μεταβλητή της οποίας τη συμπεριφορά θέλουμε να εξερευνήσουμε, δοθέντων των τιμών των παραπάνω χαρακτηριστικών. Το κρασί βαθμολογείται σε κλίμακα

$0 - 10$ με το 10 να θεωρείται το καλύτερο. Επισημαίνεται ότι στο σύνολο δεδομένων παρατηρούνται μόνο ακέραιες τιμές εντός του διαστήματος, χωρίς αυτό να σημαίνει ότι ένα κρασί δεν μπορεί να λάβει και δεκαδικό αριθμό ως βαθμολογία.

Η παρουσίαση των δεδομένων ολοκληρώνεται με τα γραφήματα που παράγονται μέσω της εντολής `ggpairs()` της βιβλιοθήκης `ggplot2` και απεικονίζουν τη συσχέτιση κάθε ζεύγους μεταβλητών του συνόλου δεδομένων, τόσο στα λευκά όσο και στα κόκκινα κρασιά. Στα δεδομένα του διαγράμματος φροντίσαμε να συμπεριλάβουμε και τη μεταβλητή απόκρισης ώστε να έχουμε μια εικόνα για τη συσχέτισή της με το κάθε $x_i$ ξεχωριστά. Τα προκύπτοντα γραφήματα φαίνονται στα Διαγράμματα 5 και 6 παρακάτω.

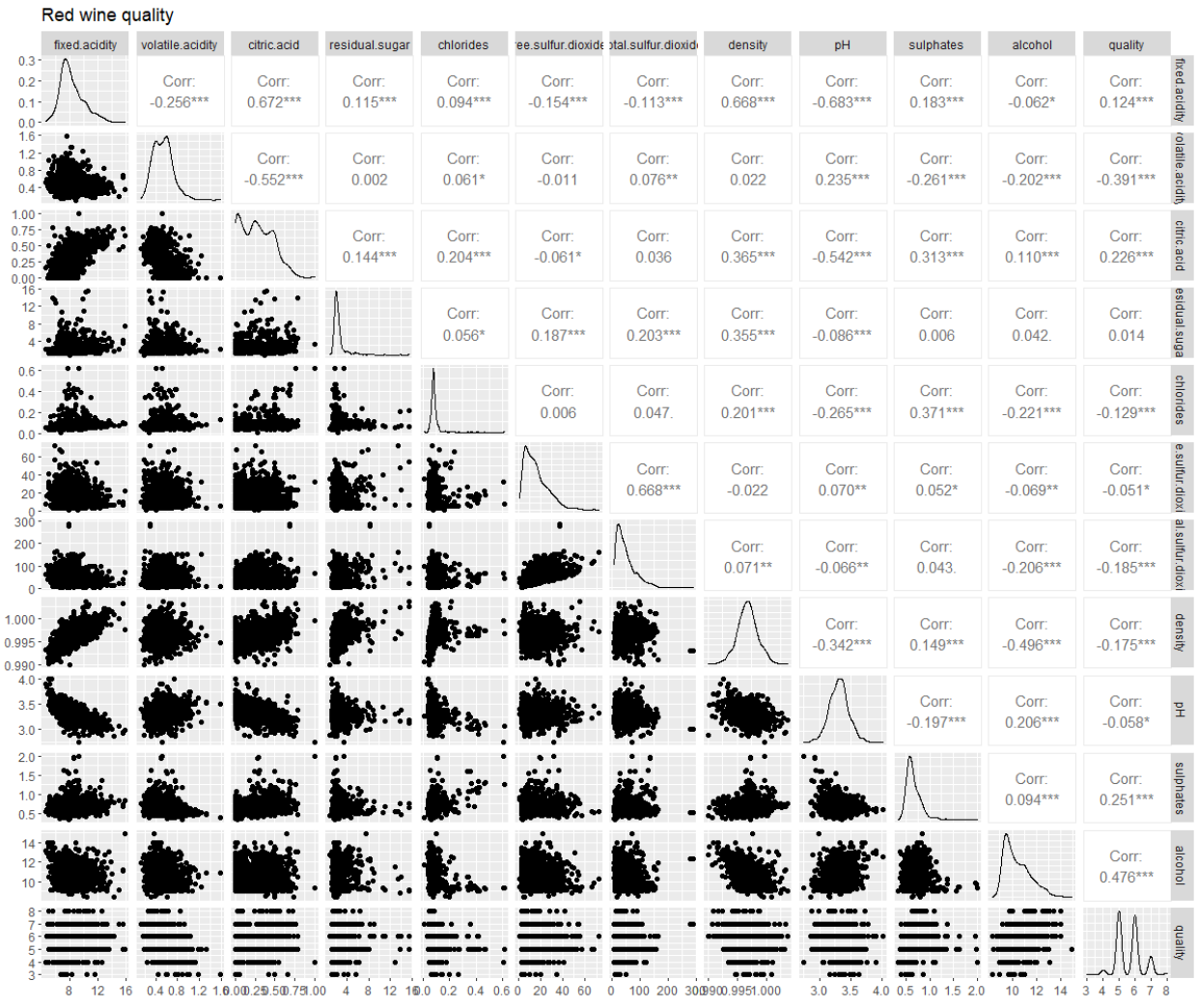



Διάγραμμα 5: Διάγραμμα που απεικονίζει την κατανομή των επεξηγηματικών μεταβλητών και τις μεταξύ τους ανά δύο συσχετίσεις για τα λευκά κρασιά.

Ξεκινάμε από το Διάγραμμα 5 που αφορά τα λευκά κρασιά. Στη διαγώνιό του φαίνεται η κατανομή της εκάστοτε επεξηγηματικής μεταβλητής όπου κατά κύριο λόγο βλέπουμε καμπανοειδείς ασύμμετρες κατανομές. Ιδιαίτερο ενδιαφέρον παρουσιάζει η μεταβλητή

`residual.sugar` που θυμίζει περισσότερο εκθετική κατανομή αλλά και το `pH` του οποίου η κατανομή μπορεί να θεωρηθεί κανονική. Ακόμα, η μεταβλητή `alcohol` που μοιάζει να έχει τρεις κορυφές.

Στρέφουμε τώρα την προσοχή μας στην τελευταία δεξιά στήλη του Διαγράμματος 5, όπου αναγράφεται η τιμή του συντελεστή συσχέτισης μεταξύ της ποιότητας του κρασιού και όλων των καταγεγραμμένων ιδιοτήτων του. Η μεγαλύτερη θετική συσχέτιση εμφανίζεται μεταξύ της ποιότητας και της περιεκτικότητας σε αλκοόλ, με τιμή $0.436$ ενώ η μεγαλύτερη αρνητική συσχέτιση εντοπίζεται μεταξύ ποιότητας και πυκνότητας, με τιμή $-0.307$. Παρατηρούμε ακόμα πως η ποιότητα του κρασιού δεν παρουσιάζει αξιοσημείωτη συσχέτιση με την ποσότητα κιτρικού οξέως, ούτε με αυτή των ελεύθερων διοξειδίων του θείου.




Διάγραμμα 6: Διάγραμμα που απεικονίζει την κατανομή των επεξηγηματικών μεταβλητών και τις μεταξύ τους ανά δύο συσχετίσεις για τα κόκκινα κρασιά.

Συνεχίζουμε με τον σχολιασμό του Διαγράμματος 6 που αφορά τα δεδομένα για το κόκκινο κρασί και δεν διαφοροποιείται σημαντικά σε σχέση με το Διάγραμμα των δεδομένων

του λευκού κρασιού. Στη διαγώνιο κυριαρχούν οι ασύμμετρες καμπανοειδείς κατανομές. Η κατανομή του `volatile.acidity` φαίνεται να έχει δύο κορυφές ενώ η κατανομή του `citric.acid` τρεις.

Όσον αφορά τους συντελεστές συσχέτισης, έχουμε ενδείξεις ότι η γραμμική συσχέτιση μεταξύ της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών `residual.sugar`, `free.sulfur.dioxide` και `pH` είναι αδύναμη. Η ισχυρότερη θετική συσχέτιση εμφανίζεται μεταξύ `quality` και `alcohol`, όπως και στα δεδομένα λευκού κρασιού, με τιμή συντελεστή συσχέτισης $0.476$. Η ισχυρότερη αρνητική συσχέτιση της ποιότητας εντοπίζεται με την μεταβλητή `volatile.acidity`, με τιμή συντελεστή συσχέτισης $-0.391$.

Κρίνεται σκόπιμο να υπογραμμιστεί σε αυτό το σημείο πως οι τιμές του συντελεστή συσχέτισης αποκαλύπτουν μόνο κατά πόσο μία μεταβλητή μπορεί να εξηγηθεί γραμμικά από κάποια άλλη. Οι τιμές του συντελεστή δεν προβλέπουν ούτε αποκλείουν την ύπαρξη σχέσης αιτίου-αιτιατού μεταξύ των μεταβλητών· θα αφήσουμε τους Μπεϋζιανούς εκτιμητές να αποκρυπτογραφήσουν τις αιτιακές σχέσεις του μοντέλου στη συνέχεια της μελέτης.

### 4.2.2 Έλεγχος προϋποθέσεων πολλαπλού γραμμικού μοντέλου

Στην κλασική στατιστική, η εκτίμηση των παραμέτρων της παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων απαιτεί την ικανοποίηση κάποιων βασικών προϋποθέσεων όσον αφορά τα δεδομένα (Φουσκάκης 2013). Αυτές είναι:

1. Γραμμική σχέση μεταξύ της δεσμευμένης μέσης τιμής της μεταβλητής απόκρισης και των επεξηγηματικών μεταβλητών

2. Κανονική κατανομή των σφαλμάτων

3. Ομοσκεδαστικότητα

4. Ανεξαρτησία των σφαλμάτων

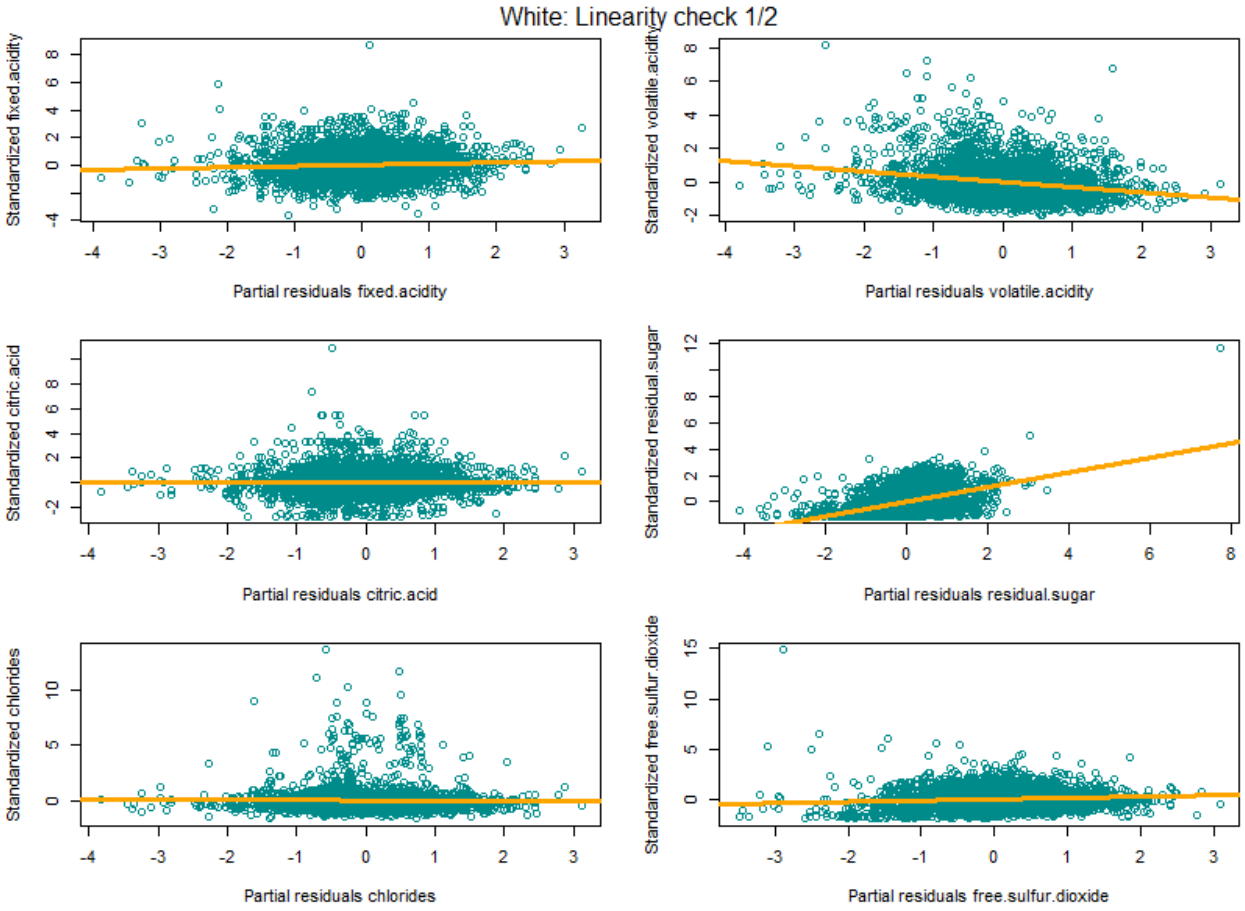
Η απαίτηση αυτή δεν παύει να ισχύει στην περίπτωση της Μπεϋζιανής παλινδρόμησης αφού η δομή του μοντέλου παραμένει η ίδια, με μόνη διαφορά πως πλέον οι παράμετροι θεωρούνται τυχαίες μεταβλητές και όχι σταθερές ποσότητες, προσέγγιση που συνιστά και τη θεμελιώδη διαφοροποίηση των δύο σχολών της στατιστικής.

Ο έλεγχος των τεσσάρων αυτών προϋποθέσεων πραγματοποιείται ακολούθως, ξεχωριστά για τα δεδομένα λευκού και κόκκινου κρασιού, με τη βοήθεια της εκτιμήτριας ελαχίστων τετραγώνων, η οποία έχουμε αποδείξει ότι ταυτίζεται με την περίπτωση χρήσης μη-πληροφοριακής πρότερης. Επιλέγουμε η παρουσίαση των ελέγχων να προηγηθεί της εφαρμογής της Μπεϋζιανής παλινδρόμησης ώστε να γίνει φανερό στον αναγνώστη πως τα δεδομένα είναι κατάλληλα για μια τέτοια ανάλυση.

#### Γραμμικότητα

Στη σχέση (27) θεωρήσαμε πως η δεσμευμένη μέση τιμή της μεταβλητής απόκρισης συνδέεται γραμμικά με τις μεταβλητές που συνιστούν τον πίνακα σχεδιασμού. Στην πολλαπλή παλινδρόμηση, η διάσταση του μοντέλου δεν μας επιτρέπει να ελέγξουμε τον ισχυρισμό μας με διάγραμμα διασποράς όπως θα κάναμε στην απλή παλινδρόμηση.

Η διαδικασία που ακολουθείται εδώ, γίνεται για κάθε μία μεταβλητή ξεχωριστά και, υποθέτοντας ότι οι υπόλοιπες επεξηγηματικές μεταβλητές συνδέονται γραμμικά με τη δεσμευμένη μέση τιμή της ανεξάρτητης μεταβλητής, ελέγχεται αν και η τιμή της προς εξέταση μεταβλητής παρουσιάζει επίσης γραμμική σύνδεση με τη δεσμευμένη μέση τιμή της $\boldsymbol{y}$.




Διάγραμμα 7: Έλεγχος γραμμικότητας με χρήση των μερικών υπολοίπων για τις επεξηγηματικές μεταβλητές 1 έως 6 στα δεδομένα λευκού κρασιού.

Για παράδειγμα, έστω ότι εξετάζουμε τη γραμμική σχέση ανάμεσα στην μεταβλητή απόκρισης και την επεξηγηματική μεταβλητή $X_j$, θεωρώντας τη δεδομένη για όλες τις υπόλοιπες επεξηγηματικές μεταβλητές. Τότε για την $i-$οστή παρατήρηση $y_i$ θα ισχύει:

$$y_i \approx \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_{j-1} x_{i(j-1)} + p_j(x_{ij}) + \hat{\beta}_{j+1} x_{i(j+1)} + ... \hat{\beta}_p x_{ip}, i = 1, ..., n, \tag{80}$$

όπου $\hat{\beta}_i$ κάποιος εκτιμητής των συντελεστών της παλινδρόμησης (για παράδειγμα εκτιμήτριες ελαχίστων τετραγώνων). Από την παραπάνω έκφραση, αρκεί να δείξουμε ότι η συνάρτηση $p_j(x_{ij})$ είναι γραμμική. Αν αναλογιστούμε και την ακόλουθη εξίσωση που μας δίνει το $i-$οστό υπόλοιπο $\hat{\epsilon}_i$ :

$$\hat{\epsilon}_i = y_i - \left( \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_p x_{ip} \right), i = 1, ..., n, \tag{81}$$

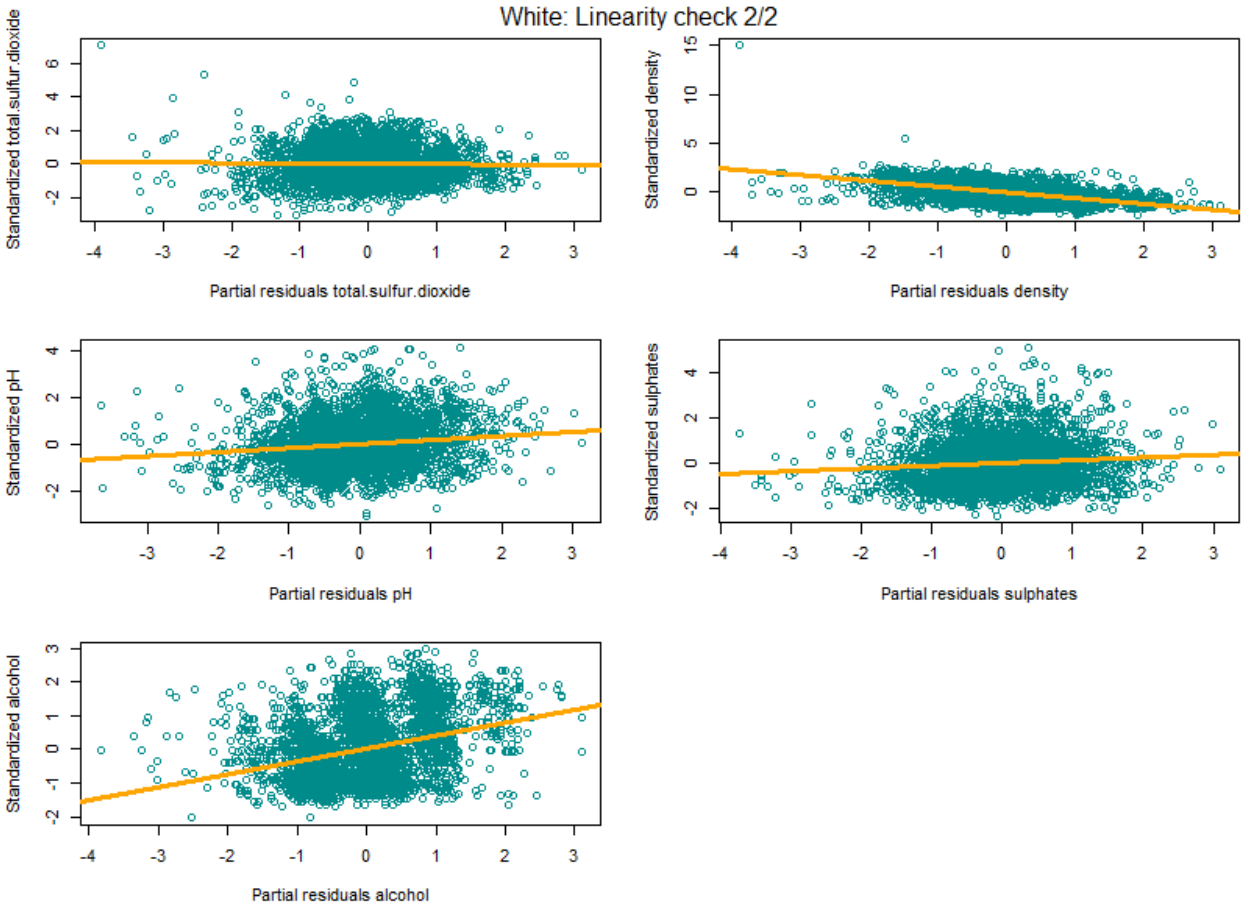
τότε από τις (80), (81) καταλήγουμε στην Εξίσωση:

$$p_j(x_{ij}) \approx \hat{\beta}_j x_{ij} + \hat{\epsilon}_i, i = 1, ..., n. \tag{82}$$

Το δεξί μέλος της Εξίσωσης (81) αποτελεί τον ορισμό των μερικών υπολοίπων (partial residuals) $P_{ij}$ και συνεπώς η γραμμικότητα της συνάρτησης $p_j(x_{ij})$, και κατ' επέκταση της σχέσης μεταξύ της μεταβλητής $X_j$ και της απόκρισης, ελέγχεται μέσω διαγράμματος διασποράς των σημείων $(x_{ij}, P_{ij})$.

Χρησιμοποιώντας τις τυποποιημένες τιμές των παρατηρήσεων κάθε μεταβλητής, σχηματίζουμε αυτά ακριβώς τα γραφήματα, τα οποία φαίνονται στα Διαγράμματα 7, 8, 9 και 10.

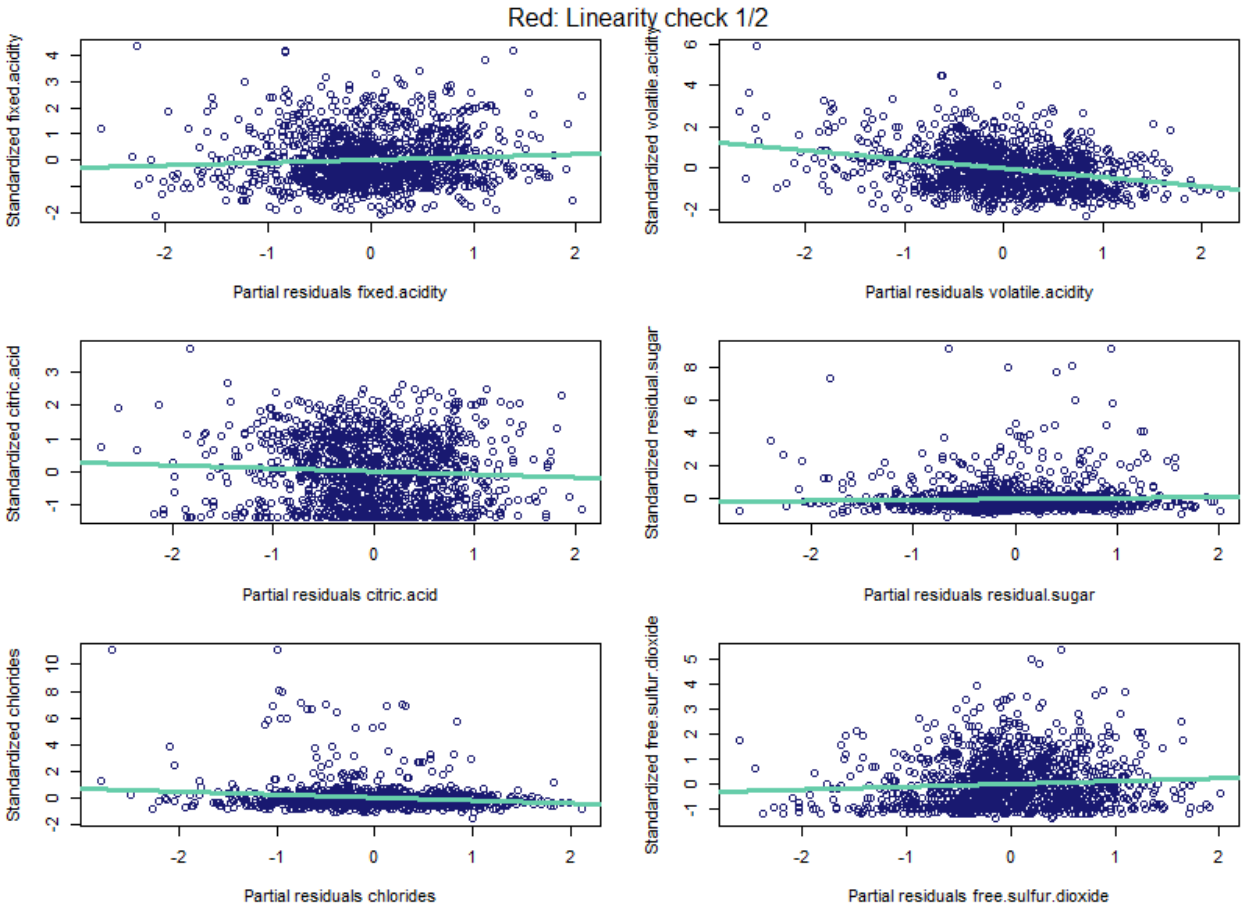
Εξετάζοντας το Διάγραμμα 7, διαπιστώνουμε ότι οι επεξηγηματικές μεταβλητές $1$ έως $6$ δεν εμφανίζουν τέλεια γραμμική σχέση με την μεταβλητή απόκρισης. Ωστόσο, μπορούμε να δεχθούμε πως η υπόθεση της γραμμικότητας ευσταθεί μερικώς, με εντονότερες αποκλίσεις από τη γραμμικότητα στο `volatile.acidity` ($x_2$) και το `chlorides` ($x_5$).




Διάγραμμα 8: Έλεγχος γραμμικότητας με χρήση των μερικών υπολοίπων για τις επεξηγηματικές μεταβλητές 7 έως 11 στα δεδομένα λευκού κρασιού.

Από το Διάγραμμα 8, φαίνεται να υπάρχουν κάποια ζητήματα γραμμικότητας για

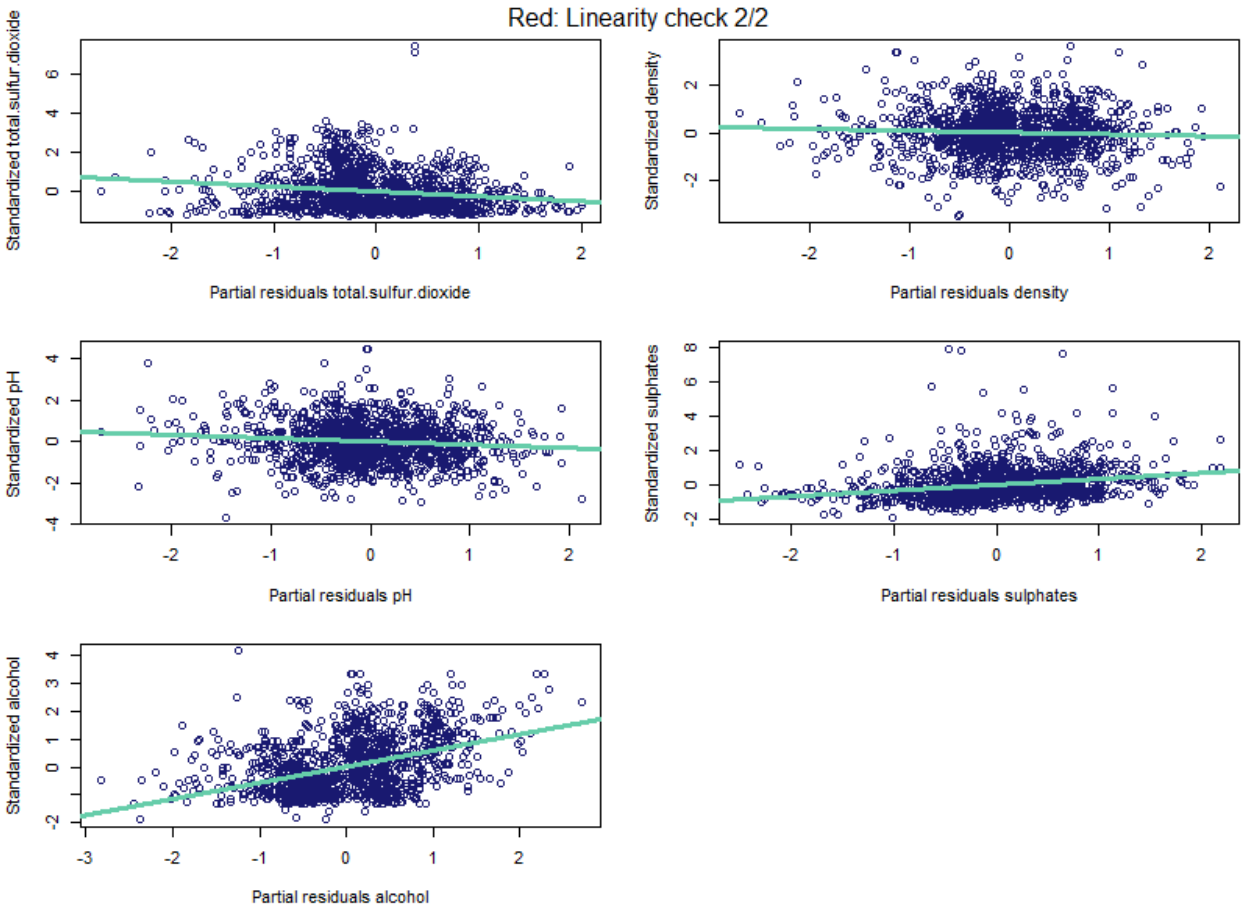
τις μεταβλητές `pH`, `sulphates`, `alcohol` ($x_9, x_{10}, x_{11}$ αντίστοιχα), ενώ για τις υπόλοιπες μεταβλητές θεωρούμε ότι δεν παραβιάζεται έντονα η υπόθεση της γραμμικότητας.




Διάγραμμα 9: Έλεγχος γραμμικότητας με χρήση των μερικών υπολοίπων για τις επεξηγηματικές μεταβλητές 1 έως 6 στα δεδομένα κόκκινου κρασιού.

Μελετώντας το Διάγραμμα 9, εντοπίζουμε ζητήματα γραμμικότητας όσον αφορά τις μεταβλητές `fixed acidity` ($x_1$), `citric acid` ($x_3$) και `free.sulfur.dioxide` ($x_6$). Αντιθέτως, η σχέση των μεταβλητών `residual.sugar` ($x_4$) και `chlorides` ($x_5$) με την μεταβλητή απόκρισης κρίνεται ικανοποιητικά γραμμική, γεγονός που μάλλον οφείλεται στο ότι αμφότερες οι επεξηγηματικές μεταβλητές λαμβάνουν μικρές τιμές. Από την άλλη βλέπουμε ότι υπάρχει ένας πεπερασμένος αριθμός παρατηρήσεων με αρκετά μεγαλύτερες τιμές που δεν μπορεί να εξηγηθεί από τη γραμμικότητα. Τέλος, προσεγγιστικά γραμμική φαίνεται και η σχέση με τη μεταβλητή `volatile.acidity` ($x_2$).

Στο Διάγραμμα 10 δεν εμφανίζεται ανησυχητική παραβίαση της γραμμικότητας από κάποια μεταβλητή ενώ μικρότερες αποκλίσεις από αυτή παρουσιάζουν οι μεταβλητές `total.sulfur.dioxide` ($x_7$) και `sulphates` ($x_{10}$). Όπως παρατηρήσαμε και στο Διάγραμμα 9 που προηγήθηκε, υπάρχει ένα σύνολο τιμών των μεταβλητών $x_7$ και $x_{10}$ του οποίου η ύπαρξη δεν δικαιολογείται από τη γραμμική σχέση επεξηγηματικής μεταβλητής και απόκρισης.



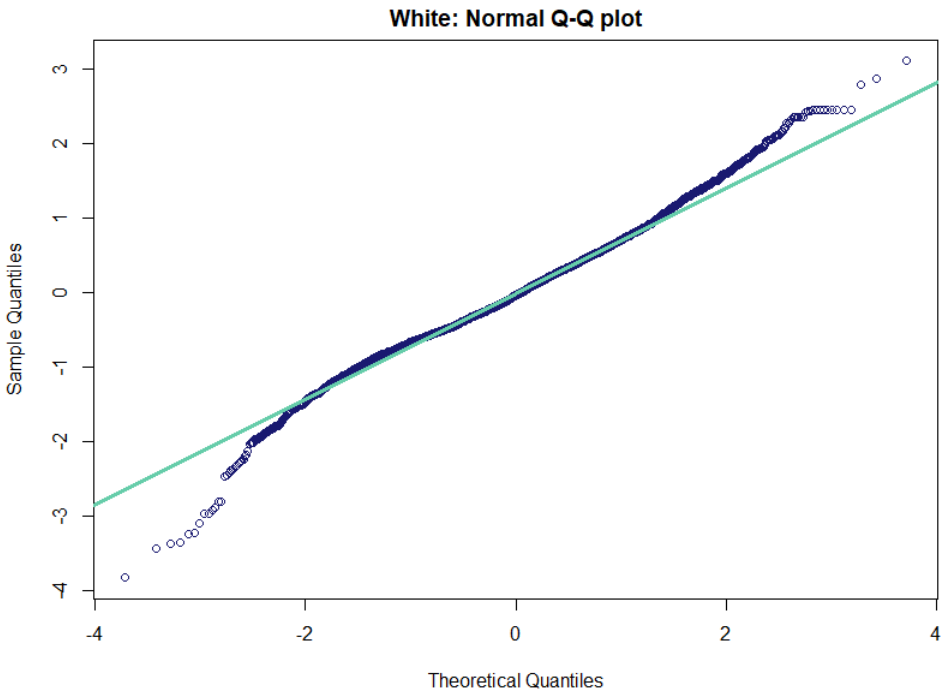

Διάγραμμα 10: Έλεγχος γραμμικότητας με χρήση των μερικών υπολοίπων για τις επεξηγηματικές μεταβλητές 7 έως 11 στα δεδομένα κόκκινου κρασιού.

**Κανονικότητα σφαλμάτων**

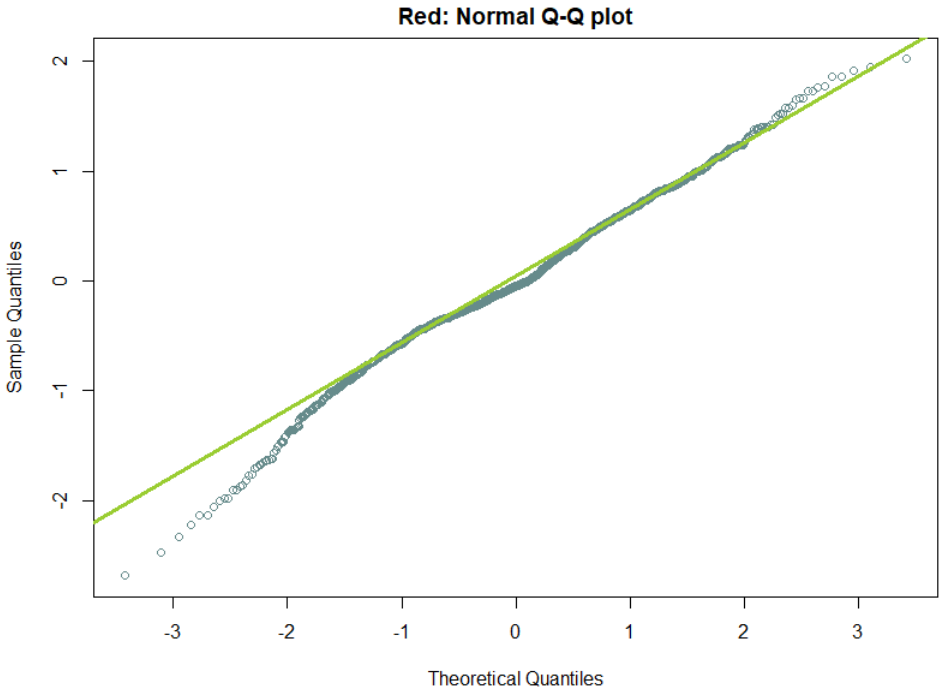
Η βασική υπόθεση που έγινε για την κατασκευή του μοντέλου, ήταν πως τα τυχαία σφάλματα $\epsilon_i, i = 1, ..., n$ ακολουθούν την κανονική κατανομή. Εφόσον τα τυχαία σφάλματα είναι άγνωστες ποσότητες, θα θεωρήσουμε τα υπόλοιπα $\epsilon_i, i = 1, ..., n$, που αποτελούν εκτιμήσεις των τυχαίων σφαλμάτων και θα ελέγξουμε την κανονικότητα αυτών, όπως συνηθίζεται.

Χρησιμοποιούμε το κανονικό διάγραμμα Q-Q (normal Q-Q plot), που συγκρίνει τα ποσοστιαία σημεία της εμπειρικής κατανομής των δεδομένων με τα ποσοστιαία σημεία της τυποποιημένης κανονικής κατανομής, και προσθέτουμε σε αυτό την ευθεία που διέρχεται από το πρώτο και το τρίτο τεταρτημόριο. Για να δεχθούμε την υπόθεση της κανονικότητας, πρέπει τα σημεία του γραφήματος να σχηματίζουν περίπου μια ευθεία, χωρίς να παρουσιάζεται καμπυλότητα ή μεγάλες αποκλίσεις στις ουρές της κατανομής.

Ακολουθώντας τη διαδικασία που περιγράφηκε για τα δεδομένα λευκού κρασιού, προκύπτει το Διάγραμμα 11 και τα αποτελέσματα είναι σχετικά ικανοποιητικά. Στην πλειοψηφία τους τα δεδομένα είναι σε ευθεία με αποκλίσεις στις άκρες που δεν ξεπερνούν τα αποδεκτά




Διάγραμμα 11: Έλεγχος της κανονικότητας των υπολοίπων για τα δεδομένα λευκού κρασιού.



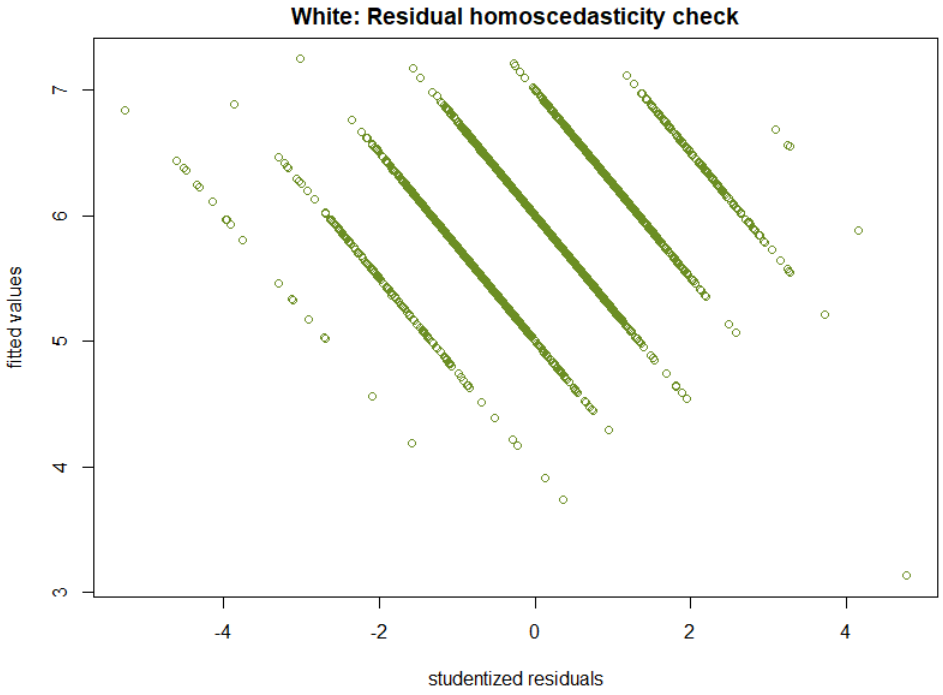

Διάγραμμα 12: Έλεγχος της κανονικότητας των υπολοίπων για τα δεδομένα κόκκινου κρασιού.

όρια ώστε να παραπέμπουν σε κάποια κατανομή με πιο παχιές ουρές. Συμπερασματικά, η προϋπόθεση της κανονικότητας των σφαλμάτων ικανοποιείται για τα δεδομένα.

Εκτελούμε τον αντίστοιχο έλεγχο στα δεδομένα κόκκινου κρασιού και λαμβάνουμε το Διάγραμμα 12. Παρατηρώντας τα σημεία, συμπεραίνουμε πως συμμορφώνονται σχετικά ικανοποιητικά στην υπόθεση της κανονικότητας.

**Ομοσκεδαστικότητα**

Η τρίτη προϋπόθεση της οποίας πρέπει να ελέγξουμε την ισχύ είναι αυτή της ομοσκεδαστικότητας των δεδομένων. Τα δεδομένα χαρακτηρίζονται από ομοσκεδαστικότητα όταν η διασπορά της δεσμευμένης κατανομής της μεταβλητής απόκρισης δοθέντων των τιμών $X$ των επεξηγηματικών μεταβλητών παραμένει σταθερή για τις διάφορες τιμές του $X$ ή, ισοδύναμα, όταν η διασπορά των τυχαίων σφαλμάτων $\epsilon_i$ παραμένει σταθερή για τις διάφορες τιμές του $X$. Η παραβίαση της υπόθεσης της ομοσκεδαστικότητας ονομάζεται ετεροσκεδαστικότητα και απαιτεί μετασχηματισμό του $\boldsymbol{y}$ για την αντιμετώπισή της.



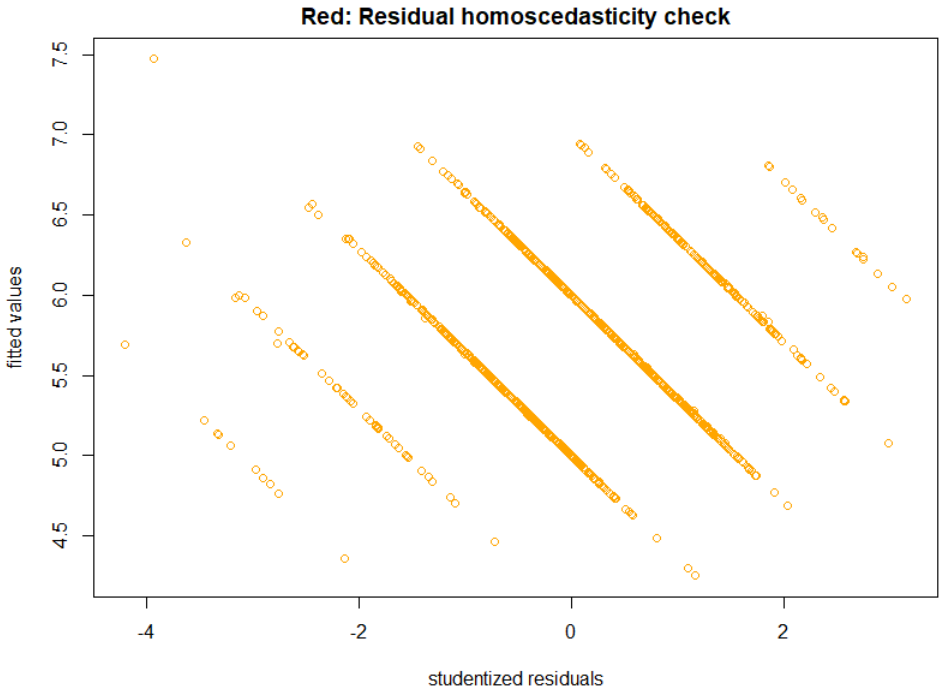

Διάγραμμα 13: Έλεγχος της υπόθεσης της ομοσκεδαστικότητας για τα δεδομένα λευκού κρασιού.

Στη θέση των άγνωστων τυχαίων σφαλμάτων θα χρησιμοποιήσουμε και πάλι τα υπόλοιπα και συγκεκριμένα τα τυποποιημένα κατά Student υπόλοιπα (studentized residuals) που ορίζονται ως εξής:

$$r_i = \frac{\hat{\epsilon}_i}{s_{y|x}\sqrt{1-h_{ii}}}, i = 1, ..., n \tag{83}$$

όπου $h_{ii}$ τα διαγώνια στοιχεία του πίνακα προβολής (ή hat matrix) $H = \tilde{X}\left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'$. Όταν ισχύει η υπόθεση της ομοσκεδαστικότητας, τα τυποποιημένα κατά Student υπόλοιπα έχουν διασπορά σταθερά ίση με 1 και μας επιτρέπουν την ευκολότερη ανίχνευση ασυνήθιστης συμπεριφοράς στο διάγραμμα διασποράς τους με τις προβλεπόμενες παρατηρούμενες τιμές $\hat{y}_i$. Ως ασυνήθιστη συμπεριφορά θεωρούμε την εμφάνιση κωνικού σχήματος (ή σχήματος βεντάλιας) στο γράφημα $(r_i, \hat{y}_i)$ ή γενικότερα μοτίβων που φανερώνουν αύξηση της διασποράς των σφαλμάτων καθώς αυξάνονται οι τιμές των $\hat{y}_i$ και κατά συνέπεια αποτελούν ένδειξη ετεροσκεδαστικότητας. Στο Διάγραμμα 13 που ακολουθεί, εκτελούμε αυτόν ακριβώς τον έλεγχο που περιγράφηκε.

Προχωρώντας στον σχολιασμό του Διαγράμματος 13 που αφορά τα δεδομένα λευκού κρασιού, διαπιστώνουμε πως δεν έχουμε ενδείξεις ότι τα δεδομένα χαρακτηρίζονται από ετεροσκεδαστικότητα. Παρατηρούμε ότι, καθώς οι τιμές των $\hat{y}_i$ αυξάνονται, οι πλάγιες ισαπέχουσες γραμμές διατηρούν σχεδόν ίδια μήκη. Επομένως, η διασπορά των τυποποιημένων κατά Student υπολοίπων, και κατ' επέκταση των τυχαίων σφαλμάτων, κινείται σε παρόμοια επίπεδα για τις διάφορες τιμές του $X$. Τελικά, αφού δεν φαίνεται τα υπόλοιπα να διασπείρονται με άτακτο τρόπο, μπορούμε να δεχθούμε την εγκυρότητα της υπόθεσης της ομοσκεδαστικότητας.



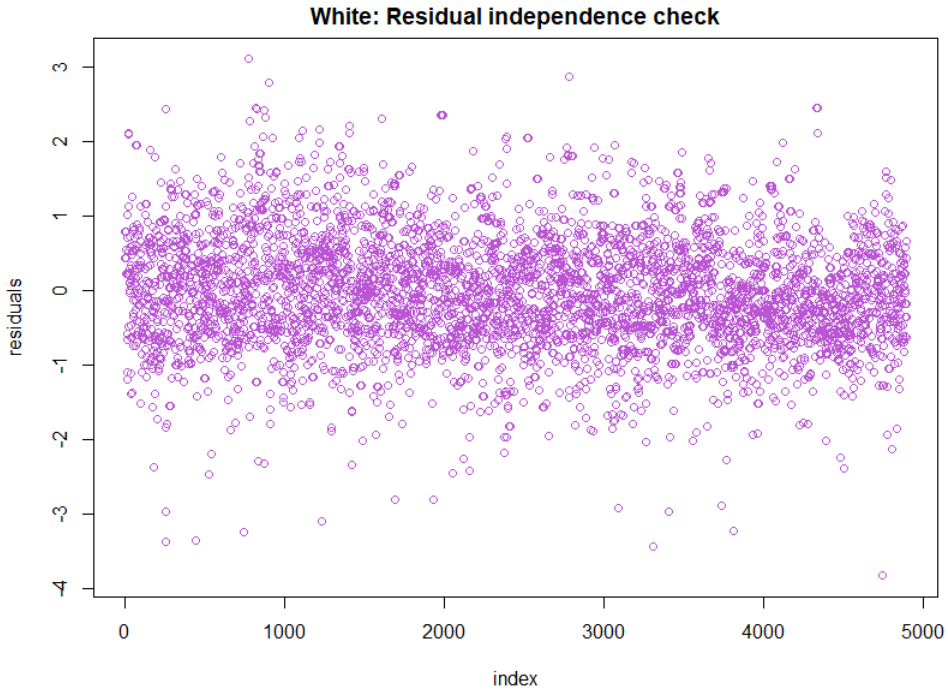

Διάγραμμα 14: Έλεγχος της υπόθεσης της ομοσκεδαστικότητας για τα δεδομένα κόκκινου κρασιού.

Τα αποτελέσματα του ελέγχου της ομοσκεδαστικότητας των υπολοίπων στις παρατηρήσεις για το κόκκινο κρασί είναι αντίστοιχα με αυτά που παρουσιάστηκαν για το λευκό (βλ. Διάγραμμα 13). Πιο συγκεκριμένα, δεν έχουμε ισχυρές ενδείξεις για την ύπαρξη

ετεροσκεδαστικότητας αφού οι παράλληλες γραμμές δεν διαφέρουν ιδιαίτερα σε μήκος.

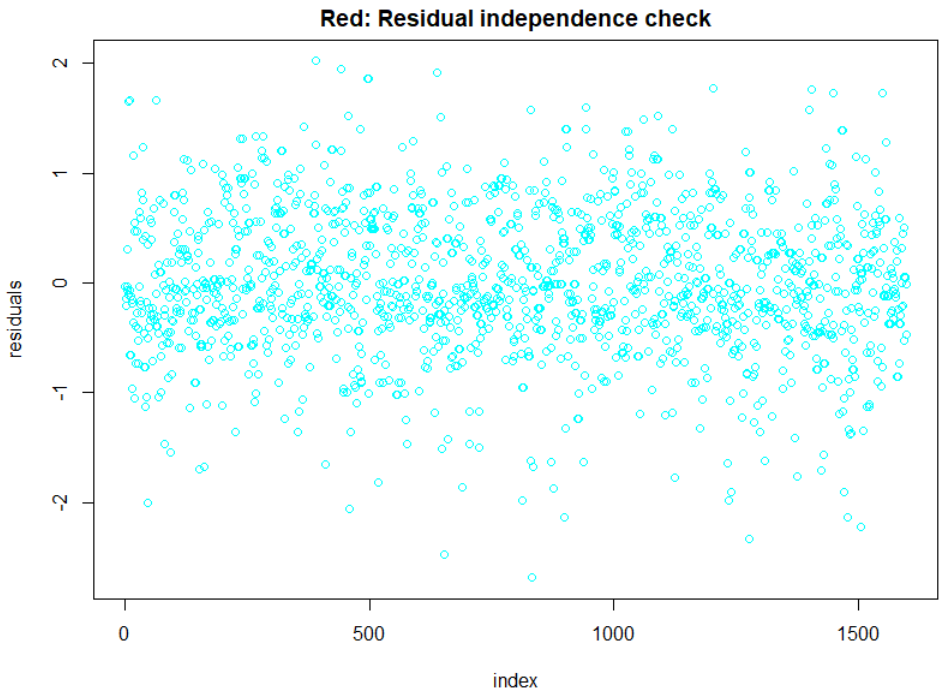
**Ανεξαρτησία των σφαλμάτων**

Τέλος, θα ασχοληθούμε με μία ακόμα βασική υπόθεση που έγινε κατά την κατασκευή του μοντέλου: την υπόθεση ότι τα σφάλματα είναι ανεξάρτητες τυχαίες μεταβλητές. Κατά τα γνωστά, τα υπόλοιπα θα μας βοηθήσουν να βεβαιώσουμε ότι η υπόθεση δεν παραβιάζεται και έτσι θα δημιουργήσουμε το διάγραμμα διασποράς τους σε σχέση με τη σειρά των δεδομένων. Αν η υπόθεση της ανεξαρτησίας ευσταθεί, το διάγραμμα δεν πρέπει να εμφανίζει κάποια συστηματική συμπεριφορά των υπολοίπων, αντιθέτως πρέπει να συμπεριφέρονται τυχαία, ειδάλλως αντιμετωπίζουμε πρόβλημα αυτοσυσχέτισης (autocorrelation) και οφείλουμε να προσαρμόσουμε το μοντέλο.




Διάγραμμα 15: Έλεγχος της υπόθεσης της ανεξαρτησίας των σφαλμάτων για τα δεδομένα λευκού κρασιού.

Ο εν λόγω έλεγχος είναι χρήσιμος όταν γνωρίζουμε τη σειρά με την οποία έχουν συλλεχθεί τα δεδομένα, όπου και έχει νόημα να παραστήσουμε ακολουθιακά τα υπόλοιπα. Ωστόσο, για λόγους πληρότητας, συμπεριλαμβάνουμε το γράφημα $(i, \hat{\epsilon}_i)$ στην ανάλυσή μας και, όπως φαίνεται στα Διαγράμματα 15 και 16, δεν προκύπτει πρόβλημα αυτοσυσχέτισης σε κανένα από τα δύο δείγματα κρασιού.




Διάγραμμα 16: Έλεγχος της υπόθεσης της ανεξαρτησίας των σφαλμάτων για τα δεδομένα κόκκινου κρασιού.

### 4.2.3 Προσαρμογή Μπεϋζιανών μοντέλων παλινδρόμησης

Σε αντιστοιχία με τη μελέτη που διεξήχθη για τα προσομοιωμένα δεδομένα, θα εφαρμόσουμε τις τρεις επιλογές πρότερων σε μοντέλα που αφορούν τα πραγματικά δεδομένα και θα τις συγκρίνουμε ως προς την ακρίβειά τους. Έτσι, στην παρούσα ενότητα, κατασκευάζουμε τρία πολλαπλά κανονικά Μπεϋζιανά μοντέλα παλινδρόμησης για τα δεδομένα λευκού κρασιού και άλλα τρία για τα δεδομένα κόκκινου κρασιού.

Επισημαίνεται ξανά πως, χάρη στην ανάλυση της Ενότητας 3, διαθέτουμε έτοιμους αναλυτικούς τύπους για τον υπολογισμό των ύστερων τιμών των παραμέτρων και για τις τρεις επιλογές πρότερων κατανομών. Το ενδιαφέρον μας εστιάζεται στην εκ των υστέρων μέση τιμή της παραμέτρου θέσης του $\boldsymbol{\beta}$, δηλαδή στο $\tilde{\boldsymbol{m}}$, που αποτελεί εκτίμηση του διανύσματος των παραμέτρων της παλινδρόμησης.

#### Κανονική-αντίστροφη γάμμα πρότερη

Αρχικά θα εφαρμοστεί στα δεδομένα η κανονική-αντίστροφη γάμμα πρότερη σε συνδυασμό με την πρότερη του Zellner. Χρησιμοποιώντας τη γνωστή πλέον μεθοδολογία και ορίζοντας, κατά τα γνωστά, τις ακόλουθες τιμές για τις παραμέτρους των κατανομών:

$$\begin{aligned} \boldsymbol{m} &= \mathbf{0}, \\ V &= g(X'X)^{-1}, \end{aligned}$$

$$g = n,$$
$$a = 0.001,$$
$$b = 0.001,$$

για τους λόγους που εξηγήσαμε στην Ενότητα 3, προκύπτει ο Πίνακας 9 παρακάτω. Ο εν λόγω Πίνακας περιλαμβάνει τις εκτιμήσεις των συνιστωσών του διανύσματος $\tilde{\boldsymbol{m}}$, ξεχωριστά για τα δεδομένα λευκού και κόκκινου κρασιού, βασισμένες στο μοντέλο παλινδρόμησης που ορίσαμε.

Το σύνολο των επιλογών για τις παραμέτρους, και τις υπερπαραμέτρους, αποσκοπεί στην κατασκευή μιας πιο αντικειμενικής από κοινού πρότερης. Ο λόγος που χτίζουμε την πρότερη κατανομή με τέτοιο τρόπο είναι ότι, παρόλο που ενδέχεται να υποψιαζόμαστε τη δομή του μοντέλου, δεν διαθέτουμε εμπεριστατωμένη πρότερη πληροφορία σχετικά με την κατανομή των παραμέτρων της παλινδρόμησης. Τελικά, χρησιμοποιούμε την μέση τιμή της εκ των υστέρων κατανομής των παραμέτρων ως σημειακή εκτίμηση αυτών.

| | | *NIG prior* | |
|---|---|---|---|
| **i** | **Επεξηγηματική μεταβλητή** | **Λευκό κρασί** | **Κόκκινο κρασί** |
| | | $\tilde{m}_i$ | $\tilde{m}_i$ |
| 0 | (intercept) | 150.16 | 21.95 |
| 1 | fixed.acidity | 0.07 | 0.02 |
| 2 | volatile.acidity | -1.86 | -1.08 |
| 3 | citric.acid | 0.02 | -0.18 |
| 4 | residual.sugar | 0.08 | 0.02 |
| 5 | chlorides | -0.25 | -1.87 |
| 6 | free.sulfur.dioxide | $4\times10^{-3}$ | $4\times10^{-3}$ |
| 7 | total.sulfur.dioxide | $3\times10^{-4}$ | $3\times10^{-4}$ |
| 8 | density | -150.25 | -17.87 |
| 9 | pH | 0.07 | -0.41 |
| 10 | sulphates | 0.06 | 0.92 |
| 11 | alcohol | 0.19 | 0.28 |

Πίνακας 9: Μέση τιμή της εκ των υστέρων κατανομής των παραμέτρων της παλινδρόμησης με χρήση της πρότερης του Zellner στα δεδομένα για λευκό και κόκκινο κρασί.

Τα ευρήματα που προκύπτουν από την προσαρμογή του μοντέλου είναι ποιοτικά παρόμοια για τα δύο είδη κρασιού. Πιο αναλυτικά, οι επεξηγηματικές μεταβλητές με τους μεγαλύτερους εκτιμώμενους συντελεστές στα μοντέλα είναι οι:

- `volatile.acidity`
- `density`
- `chlorides` (μόνο στα κόκκινα κρασιά).

Αντίθετα, οι επεξηγηματικές μεταβλητές στις οποίες αντιστοιχούν πολύ μικρές τιμές συντελεστών, κάτω από $10^{-1}$, για τα δεδομένα κόκκινου κρασιού, είναι οι:

- `fixed.acidity`
- `residual.sugar`
- `free.sulfur.dioxide`
- `total.sulfur.dioxide`,

ενώ στην περίπτωση του λευκού κρασιού η λίστα συμπληρώνεται με τα:

- `citric.acid`
- `chlorides`
- `pH`
- `sulphates`.

Οφείλει να επισημανθεί το γεγονός ότι μικρές ή μεγάλες τιμές των συντελεστών δεν συνεπάγονται αυτόματα μικρή ή μεγάλη επίδραση της αντίστοιχης επεξηγηματικής μεταβλητής στο συγκεκριμένο μοντέλο. Για να αποφανθούμε σχετικά, πρέπει πρώτα να λάβουμε υπόψη το εύρος των τιμών της μεταβλητής. Παραδείγματος χάριν, ο συντελεστής του $x_6$ (`free.sulfur.dioxide`) με τιμή $4\times10^{-3}$ στις παρατηρήσεις λευκών κρασιών είναι ένας φαινομενικά ασήμαντος αριθμός. Ωστόσο, το γινόμενό του με τιμές του $x_6$ άνω του $50$, αριθμός που εμφανίζεται με σημαντική συχνότητα, προσαυξάνει κατά $0,2$ στην ποιότητα του κρασιού, συνεισφορά καθόλου αμελητέα στην κλίμακα $0-10$.

Επιπλέον, είναι αξιοσημείωτο το γεγονός ότι τα πρόσημα των συντελεστών των επεξηγηματικών μεταβλητών του μοντέλου στην πλειοψηφία τους δεν διαφοροποιούνται μεταξύ των δύο ειδών κρασιού, που σημαίνει ότι κατά κύριο λόγο οι προς μελέτη ιδιότητες επηρεάζουν με τον ίδιο τρόπο την ποιότητα και στα δύο είδη. Μάλιστα, το μοντέλο επιβεβαιώνει κάποιες αρχικές υποψίες που είχαμε για τη σχέση εξαρτημένων και ανεξάρτητων μεταβλητών. Ο λόγος για το αρνητικό πρόσημο στον συντελεστή του $x_2$ (`volatile.acidity`) αφού πράγματι αναμέναμε η υψηλή πτητική οξύτητα, που προκαλεί γεύση ξυδιού, να μειώνει την ποιότητα του κρασιού.

Εξαίρεση για την παρατήρηση σχετικά με τα πρόσημα αποτελούν οι ιδιότητες που περιγράφονται από τα $x_3$ (`citric.acid`) και $x_9$ (`pH`). Συγκεκριμένα, φαίνεται πως η ποσότητα κιτρικού οξέως επιδρά θετικά στην ποιότητα του λευκού κρασιού και αρνητικά σε αυτή του κόκκινου κρασιού. Βέβαια, το αποτέλεσμα αυτό για τα δεδομένα του κόκκινου κρασιού, έρχεται σε αντίθεση με τη διαίσθησή μας καθώς η ύπαρξη κιτρικού οξέως ενισχύει τη φρεσκάδα και τη γεύση του κρασιού, συμβάλλοντας στην βελτίωση της ποιότητάς του. Από την άλλη, μεγαλύτερο pH, δηλαδή μικρότερη οξύτητα, φαίνεται να αυξάνει την ποιότητα του λευκού κρασιού ενώ μικρότερο pH, δηλαδή μεγαλύτερη οξύτητα, οδηγεί σε αυξημένη ποιότητα στο κόκκινο κρασί.

Συγκεντρωτικά, οι επεξηγηματικές μεταβλητές με θετικό συντελεστή για το μοντέλο λευκού κρασιού είναι:

1. `fixed.acidity`
2. `citric.acid`
3. `residual.sugar`
4. `free.sulfur.dioxide`
5. `total.sulfur.dioxide`
6. `pH`
7. `sulphates`
8. `alcohol`.

Ακολουθεί η αντίστοιχη λίστα για το μοντέλο κόκκινου κρασιού, όπου θετικό συντελεστή βρίσκουμε στις μεταβλητές:

1. `fixed.acidity`
2. `residual.sugar`
3. `free.sulfur.dioxide`
4. `total.sulfur.dioxide`
5. `sulphates`
6. `alcohol`.

Οι μεγάλες θετικές τιμές των $\tilde{m}_0$, σε συνδυασμό με το αρνητικό πρόσημο των μεγαλύτερων κατ' απόλυτη τιμή συντελεστών των επεξηγηματικών μεταβλητών, αποκαλύπτουν τον μηχανισμό που διαμορφώνει τη μεταβλητή απόκρισης: κάθε παρατήρηση κρασιού ξεκινά με μία υψηλή τιμή για την ποιότητα ($150.16$ στα λευκά και $21.95$ στα κόκκινα κρασιά) η οποία μειώνεται σημαντικά με την επίδραση των περισσότερων συμμεταβλητών και αυξάνει μερικώς με την επίδραση των υπολοίπων συμμεταβλητών. Με αυτή τη δομή, η κλίμακα της βαθμολογίας της ποιότητας διαμορφώνεται στο $0 - 10$.

Ερμηνεύουμε πιο αναλυτικά τις προκύπτουσες τιμές των εκτιμητών, αρχικά για το δείγμα λευκού κρασιού. Επιστρατεύοντας την NIG πρότερη με τις προτάσεις του Zellner, η εκτιμώμενη ευθεία παλινδρόμησης για τη μέση ποιότητα $\boldsymbol{y}_\lambda$ του λευκού κρασιού vinho verde είναι η εξής:

$$\begin{aligned}\hat{\boldsymbol{y}}_\lambda = {} & 150.16 - 0.07x_1 - 1.86x_2 + 0.02x_3 + 0.08x_4 \\ & - 0.25x_5 + 4 \times 10^{-3}x_6 - 3 \times 10^{-4}x_7 - 150.25x_8 \\ & + 0.07x_9 + 0.06x_{10} + 0.19x_{11}. \end{aligned} \tag{84}$$

Γενικότερα, στην ανάλυση παλινδρόμησης, οι συντελεστές $\beta_i$ εκφράζουν την αναμενόμενη μεταβολή στη μεταβλητή απόκρισης, όταν η τιμής της επεξηγηματικής μεταβλητής $x_i$ αυξηθεί κατά μία μονάδα, δεδομένου ότι οι υπόλοιπες επεξηγηματικές μεταβλητές του

μοντέλου μένουν σταθερές (Φουσκάκης 2013). Συνεπώς, ενδεικτικά αναφέρουμε πως η εκτίμηση της τιμής του συντελεστή του $x_2$ υποδηλώνει πως αν η πτητική οξύτητα αυξηθεί κατά ένα $g/dm^3$, ενώ οι υπόλοιπες μεταβλητές είναι σταθερές, η ποιότητα θα μειωθεί κατά $1.86$. Αντίστοιχα, από την τιμή του συντελεστή του $x_8$, βλέπουμε πως αν η πυκνότητα αυξηθεί κατά ένα $g/cm^3$, για σταθερή τιμή των υπόλοιπων επεξηγηματικών μεταβλητών, η ποιότητα μειώνεται κατά $150.25$.

Ανάλογα συμπεράσματα προκύπτουν και στο δείγμα κόκκινου κρασιού όπου το μοντέλο εκτιμάται ως εξής:

$$\begin{aligned}\hat{\boldsymbol{y}}_\kappa = &21.95 - 0.02x_1 - 1.08x_2 - 0.18x_3 + 0.02x_4 \\ &- 1.87x_5 + 4 \times 10^{-3}x_6 - 3 \times 10^{-4}x_7 - 17.87x_8 \\ &- 0.41x_9 + 0.92x_{10} + 0.28x_{11}. \end{aligned} \tag{85}$$

Κατά τη συνήθη ερμηνεία των συντελεστών της παλινδρόμησης, σχολιάζουμε ενδεικτικά την τιμή του $\tilde{m}_5$ η οποία προκαλεί μείωση της μεταβλητής κατά $1,87$ όταν η μεταβλητή $x_7$ (`chlorides`) αυξηθεί κατά ένα $g/dm^3$ και θεωρώντας πως οι υπόλοιπες μεταβλητές μένουν σταθερές.

**Improper πρότερη για τη διασπορά**

Τα αποτελέσματα του Πίνακα 10 αφορούν την εφαρμογή του μοντέλου της improper πρότερης για τη διασπορά, που αναπτύχθηκε στην Υποενότητα 3.2, στα δεδομένα του κρασιού vinho verde. Συνδυάζουμε τη συγκεκριμένη πρότερη για το $\sigma^2$ με την πρότερη του Zellner για τις τιμές των $\boldsymbol{m}$ και $V$ και έτσι η ύστερη μεταβλητή ενδιαφέροντος $\tilde{\boldsymbol{m}}$ παίρνει τις ίδιες τιμές με την περίπτωση NIG πρότερης που προηγήθηκε.

| **i** | **Επεξηγηματική μεταβλητή** | *Improper prior for variance* | |
|---|---|---|---|
| | | **Λευκό κρασί** | **Κόκκινο κρασί** |
| | | $\tilde{m}_i$ | $\tilde{m}_i$ |
| 0 | (intercept) | 150.16 | 21.95 |
| 1 | fixed.acidity | 0.07 | 0.02 |
| 2 | volatile.acidity | -1.86 | -1.08 |
| 3 | citric.acid | 0.02 | -0.18 |
| 4 | residual.sugar | 0.08 | 0.02 |
| 5 | chlorides | -0.25 | -1.87 |
| 6 | free.sulfur.dioxide | $4\times10^{-3}$ | $4\times10^{-3}$ |
| 7 | total.sulfur.dioxide | $3\times10^{-4}$ | $3\times10^{-4}$ |
| 8 | density | -150.25 | -17.87 |
| 9 | pH | 0.07 | -0.41 |
| 10 | sulphates | 0.06 | 0.92 |
| 11 | alcohol | 0.19 | 0.28 |

Πίνακας 10: Μέση τιμή της εκ των υστέρων κατανομής των παραμέτρων της παλινδρόμησης με χρήση improper πρότερης για τη διασπορά στα δεδομένα για λευκό και κόκκινο κρασί.

Εφόσον οι προκύπτουσες εκτιμήσεις των συντελεστών της παλινδρόμησης δεν διαφοροποιούνται, τις παραθέτουμε απλά ακολούθως ενώ οποιαδήποτε μελλοντική αναφορά στα αποτελέσματα της κανονικής-αντίστροφης γάμμα πρότερης θα συμπεριλαμβάνει και την παρούσα εφαρμογή. Υπενθυμίζουμε πως παρόμοιο σχόλιο είχε γίνει και στην αντίστοιχη εφαρμογή σε προσομοιωμένα δεδομένα.

**Μη-πληροφοριακή από κοινού πρότερη**

Η τελευταία εκ των προτέρων κατανομή που θα χρησιμοποιηθεί για τη Μπεϋζιανή παλινδρόμηση με τα δεδομένα κρασιού, είναι η μη-πληροφοριακή πρότερη. Με αυτή την επιλογή, κατασκευάζουμε ένα εντελώς αντικειμενικό εκτιμητή για το $\boldsymbol{\beta}$ ο οποίος μάλιστα έχουμε αναφέρει ότι θα συμπίπτει κατά μέση τιμή με την εκτιμήτρια ελαχίστων τετραγώνων. Παρακάτω, στον Πίνακα 11, φαίνονται οι τιμές των εκτιμητών:

| | | *Non-informative prior* | |
|---|---|---|---|
| **i** | **Επεξηγηματική μεταβλητή** | **Λευκό κρασί** $\tilde{m}_i$ | **Κόκκινο κρασί** $\tilde{m}_i$ |
| 0 | (intercept) | 150.19 | 21.97 |
| 1 | fixed.acidity | 0.07 | 0.02 |
| 2 | volatile.acidity | -1.86 | -1.08 |
| 3 | citric.acid | 0.02 | -0.18 |
| 4 | residual.sugar | 0.08 | 0.02 |
| 5 | chlorides | -0.24 | -1.87 |
| 6 | free.sulfur.dioxide | $4\times10^{-3}$ | $4\times10^{-3}$ |
| 7 | total.sulfur.dioxide | $3\times10^{-4}$ | $3\times10^{-4}$ |
| 8 | density | -150.28 | -17.88 |
| 9 | pH | 0.07 | -0.41 |
| 10 | sulphates | 0.06 | 0.92 |
| 11 | alcohol | 0.19 | 0.28 |

Πίνακας 11: Μέση τιμή της εκ των υστέρων κατανομής των παραμέτρων της παλινδρόμησης με χρήση μη-πληροφοριακής πρότερης για τη διασπορά στα δεδομένα για λευκό και κόκκινο κρασί.

Συνεπώς προκύπτουν οι εξής εκτιμώμενες ευθείες παλινδρόμησης για τη μέση ποιότητα $\boldsymbol{y}_\lambda$ των λευκών κρασιών και τη μέση ποιότητα $\boldsymbol{y}_\kappa$ των κόκκινων κρασιών:

$$\begin{aligned}\hat{\boldsymbol{y}}_\lambda &= 150.19 - 0.07x_1 - 1.86x_2 + 0.02x_3 + 0.08x_4 \\ &\quad - 0.24x_5 + 4\times10^{-3}x_6 - 3\times10^{-4}x_7 - 150.28x_8 \\ &\quad + 0.07x_9 + 0.06x_{10} + 0.19x_{11}. \end{aligned} \tag{86}$$

$$\begin{aligned}\hat{\boldsymbol{y}}_{\kappa} = 21.97 &- 0.02x_1 - 1.08x_2 - 0.18x_3 + 0.02x_4 \\ &- 1.87x_5 + 4 \times 10^{-3}x_6 - 3 \times 10^{-4}x_7 - 17.88x_8 \\ &- 0.41x_9 + 0.92x_{10} + 0.28x_{11}. \end{aligned} \tag{87}$$

Συγκρίνοντας τις Εξισώσεις (86) και (87) με το ζεύγος (84), (85) που εμφανίστηκε στην ανάλυση που έγινε για τα μοντέλα με χρήση της NIG πρότερης, διαπιστώνουμε ότι οι διαφορές μεταξύ των μοντέλων που παρήγαγαν οι πρότερες είναι πολύ μικρές και ασήμαντες. Συμβουλευόμενοι και τους Πίνακες 12, 13 εντοπίζουμε επακριβώς τις αποκλίσεις μεταξύ των εκτιμήσεων:

- Λευκό κρασί:

    1. Ο σταθερός όρος εμφανίζεται αυξημένος κατά $3 \times 10^{-2}$ στο μοντέλο της μη-πληροφοριακής πρότερης.
    2. Η εκτίμηση του συντελεστή της μεταβλητής `chlorides` παρουσιάζει αύξηση κατά $10^{-2}$ στο μοντέλο της μη-πληροφοριακής πρότερης.
    3. Η εκτίμηση του συντελεστή της μεταβλητής `density` εμφανίζει αύξηση κατά $3 \times 10^{-2}$ στο μοντέλο της μη-πληροφοριακής πρότερης.

| | | *Λευκό κρασί* | | |
|---|---|---|---|---|
| **i** | **Επεξηγηματική μεταβλητή** | **NIG** | **Improper variance** | **Non-informative** |
| | | $\tilde{m}_i$ | $\tilde{m}_i$ | $\tilde{m}_i$ |
| 0 | (intercept) | 150,16 | 150.16 | 150.19 |
| 1 | fixed.acidity | 0.07 | 0.07 | 0.07 |
| 2 | volatile.acidity | -1.86 | -1.86 | -1.86 |
| 3 | citric.acid | 0.02 | 0.02 | 0.02 |
| 4 | residual.sugar | 0.08 | 0.08 | 0.08 |
| 5 | chlorides | -0.25 | -0.25 | -0.24 |
| 6 | free.sulfur.dioxide | $4\times10^{-3}$ | $4\times10^{-3}$ | $4\times10^{-3}$ |
| 7 | total.sulfur.dioxide | $3\times10^{-4}$ | $3\times10^{-4}$ | $3\times10^{-4}$ |
| 8 | density | -150.25 | -150.25 | -150.28 |
| 9 | pH | 0.07 | 0.07 | 0.07 |
| 10 | sulphates | 0.06 | 0.06 | 0.06 |
| 11 | alcohol | 0.19 | 0.19 | 0.19 |

Πίνακας 12: Μέση τιμή της εκ των υστέρων κατανομής των παραμέτρων της παλινδρόμησης για τα τρία είδη πρότερων στα δεδομένα για λευκό κρασί.

- Κόκκινο κρασί:

    1. Ο σταθερός όρος εμφανίζεται αυξημένος κατά $2 \times 10^{-2}$ στο μοντέλο της μη-πληροφοριακής πρότερης.
    2. Η εκτίμηση του συντελεστή της μεταβλητής `density` εμφανίζει αύξηση κατά $10^{-2}$ στο μοντέλο της μη-πληροφοριακής πρότερης.

Όπως ήδη επισημάνθηκε, οι διαφοροποιήσεις αυτές πολύ μικρές και συνεπώς τα συμπεράσματα που προκύπτουν από την εφαρμογή της μη-πληροφοριακής πρότερης είναι πανομοιότυπα με αυτά που διατυπώθηκαν για το μοντέλο της κανονικής-αντίστροφης γάμμα πρότερης.

| | | *Κόκκινο κρασί* | | |
|---|---|---|---|---|
| **i** | **Επεξηγηματική μεταβλητή** | **NIG** | **Improper variance** | **Non-informative** |
| | | $\tilde{m}_i$ | $\tilde{m}_i$ | $\tilde{m}_i$ |
| 0 | (intercept) | 21.95 | 21.95 | 21.97 |
| 1 | fixed.acidity | 0.02 | 0.02 | 0.02 |
| 2 | volatile.acidity | -1.08 | -1.08 | -1.08 |
| 3 | citric.acid | -0.18 | -0.18 | -0.18 |
| 4 | residual.sugar | 0.02 | 0.02 | 0.02 |
| 5 | chlorides | -1.87 | -1.87 | -1.87 |
| 6 | free.sulfur.dioxide | $4\times10^{-3}$ | $4\times10^{-3}$ | $4\times10^{-3}$ |
| 7 | total.sulfur.dioxide | $3\times10^{-4}$ | $3\times10^{-4}$ | $3\times10^{-4}$ |
| 8 | density | -17.87 | -17.87 | -17.88 |
| 9 | pH | -0.41 | -0.41 | -0.41 |
| 10 | sulphates | 0.92 | 0.92 | 0.92 |
| 11 | alcohol | 0.28 | 0.28 | 0.28 |

Πίνακας 13: Μέση τιμή της εκ των υστέρων κατανομής των παραμέτρων της παλινδρόμησης για τα τρία είδη πρότερων στα δεδομένα για κόκκινο κρασί.

### 4.2.4 Προβλέψεις

Η διαδικασία πρόβλεψης της τιμής της μεταβλητής απόκρισης σε νέα δεδομένα διαφοροποιείται από αυτή που ακολουθείται στην κλασική παλινδρόμηση. Στα πλαίσια της Μπεϋζιανής στατιστική βασιζόμαστε στην προβλεπτική κατανομή η οποία λαμβάνει υπόψη και την αβεβαιότητα που υπάρχει για τις άγνωστες παραμέτρους του μοντέλου.

Στο κεφάλαιο που προηγήθηκε είδαμε ότι και οι τρεις πρότερες που επιλέξαμε να παρουσιάσουμε οδηγούν σε προβλεπτικές κατανομές της οικογένειας της πολυδιάστατης Student κατανομής. Παρά τις επί μέρους διαφορές όσον αφορά τους βαθμούς ελευθερίας και την παράμετρο κλίμακας, σε όλες τις περιπτώσεις η παράμετρος θέσης δίνεται ως:

$$X^* \tilde{\boldsymbol{m}},$$

όπου $\tilde{\boldsymbol{m}}$ η εκ των υστέρων μέση τιμή του διανύσματος $\boldsymbol{\beta}$ των συντελεστών της παλινδρόμησης, όπως αυτή ορίζεται κατά περίπτωση για την εκάστοτε επιλογή πρότερης, και $X^*$ ο πίνακας

σχεδιασμού των νέων παρατηρήσεων. Καθώς η παράμετρος θέσης της MVSt προβλεπτικής κατανομής συμπίπτει με τη μέση τιμή της, μπορούμε να τη χρησιμοποιήσουμε ως σημειακή εκτίμηση της τιμής της μεταβλητής απόκρισης. Έτσι, η τιμή $\boldsymbol{y}^*$ μπορεί να εκτιμηθεί ως εξής:

$$\hat{\boldsymbol{y}^*} = X^* \tilde{\boldsymbol{m}}.$$

Εφόσον δεν έχουμε πρόσβαση σε εξ' ολοκλήρου νέες παρατηρήσεις, θα χρησιμοποιήσουμε τη μέθοδο *Leave-One-Out Cross Validation* (*LOOCV*) για να εφαρμόσουμε και να αξιολογήσουμε τη διαδικασία πρόβλεψης. Σύμφωνα με τη μέθοδο αυτή, εκπαιδεύουμε το μοντέλο παλινδρόμησης χρησιμοποιώντας όλα τα διανύσματα παρατηρήσεων, πλην ενός. Στη συνέχεια, χρησιμοποιούμε το μοντέλο που προέκυψε ώστε να προβλέψουμε την τιμή της μεταβλητής απόκρισης για την παρατήρηση που παραλείψαμε και τη συγκρίνουμε με την πραγματική καταγεγραμμένη τιμή της. Η διαδικασία επαναλαμβάνεται παραλείποντας κάθε φορά κι από μία παρατήρηση μέχρις ότου έχουν όλες μείνει εκτός του συνόλου εκπαίδευσης ακριβώς μία φορά.

Η αξιολόγηση του μοντέλου γίνεται με τη βοήθεια της ρίζας του μέσου τετραγωνικού σφάλματος ή αλλιώς του *Root Mean Square Error* (*RMSE*) και, όπως υποδεικνύει και η ονομασία του, προκύπτει από τον υπολογισμό της ρίζας του αριθμητικού μέσου των τετραγώνων των αποκλίσεων της προβλεπόμενης τιμής από την πραγματική. Ο αριθμητικός μέσος ορίζεται μεταξύ των $n$ διαφορετικών μοντέλων που προσαρμόζονται στα πλαίσια του LOOCV ενώ η απόκλιση αφορά την εκτιμώμενη και την πραγματική τιμή της μεταβλητής απόκρισης για την παρατήρηση που έχει παραλειφθεί στο $i-$οστό μοντέλο, $i = 1, ..., n$. Πιο αναλυτικά, στην περίπτωσή μας θα δίνεται ως εξής:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{\left(\hat{y}_i - y_i\right)^2}{n}}. \tag{88}$$

Επιθυμούμε οι τιμές του RMSE να είναι μικρές ώστε το σφάλμα πρόβλεψης να είναι μικρό και το μοντέλο να μπορεί να χαρακτηριστεί χρήσιμο και ακριβές. Το ποιες ακριβώς τιμές του RMSE θεωρούνται μικρές εξαρτάται από το είδος και τη μονάδα μέτρησης της προς μελέτη μεταβλητής.

Στον Πίνακα 14 φαίνονται συγκεντρωτικά οι τιμές του RMSE για τις τρεις επιλογές πρότερης αφού εφαρμόσουμε Leave-One-Out Cross Validation ξεχωριστά για τα δεδομένα κόκκινου και λευκού κρασιού.

| | *RMSE* | |
|---|---|---|
| **Είδος πρότερης** | **Λευκό κρασί** | **Κόκκινο κρασί** |
| NIG | 0.75 | 0.65 |
| Improper variance | 0.75 | 0.65 |
| Non-informative | 0.75 | 0.65 |

Πίνακας 14: Υπολογισμός του Root Mean Square Error για τις προβλεπόμενες τιμές της μεταβλητής απόκρισης μέσω Leave-One-Out Cross Validation, για τις τρεις επιλογές πρότερης κατανομής στους δύο τύπους κρασιού.

Παρατηρώντας τον Πίνακα 14, φαίνεται πως τόσο στα δεδομένα κόκκινου όσο και στα δεδομένα λευκού κρασιού, οι τιμές του RMSE ταυτίζονται μεταξύ των διαφόρων ειδών πρότερων. Το γεγονός αυτό δεν μας προκαλεί έκπληξη καθώς τα μοντέλα παλινδρόμησης για τις διάφορες πρότερες που προέκυψαν παραπάνω δεν διαφέρουν παρά ελάχιστα σε κάθε τύπο κρασιού. Επομένως, ανεξαρτήτως πρότερης, η τιμή του RMSE για το μοντέλο λευκού κρασιού διαμορφώνεται στις $0.75$ μονάδες ενώ για το μοντέλο κόκκινου κρασιού στις $0.65$. Καμία από τις δύο εκτιμήσεις δεν μπορεί να θεωρηθεί αμελητέα καθώς και οι δύο ξεπερνούν τη μισή μονάδα, τιμή που μπορεί να κάνει τη διαφορά μεταξύ ενός χαλασμένου και ενός μέτριου κρασιού, ενός μέτριου και ενός πολύ καλού. Παρ' όλα αυτά, συνολικά δεν θα θεωρούσαμε ανησυχητικές τις τιμές του RMSE στα δεδομένα κόκκινου κρασιού ενώ κρατάμε τις επιφυλάξεις μας για το προκύπτον μοντέλο από τα δεδομένα λευκού κρασιού.

### 4.2.5 Συμπεράσματα και σχόλια

Έχοντας ολοκληρώσει την ανάλυση των δεδομένων λευκού και κόκκινου κρασιού vinho verde, μπορούμε πλέον να σχολιάσουμε τα αποτελέσματα και να διατυπώσουμε τα ερευνητικά μας συμπεράσματα.

Όσον αφορά τα προκύπτοντα μοντέλα, συμβουλευόμαστε ξανά τους Πίνακες 12 και 13 ώστε να καταλήξουμε στους σημαντικότερους παράγοντες που επηρεάζουν την ποιότητα του κρασιού. Καταρχήν, δεν πρέπει να μας παραπλανήσει η μεγάλη τιμή του συντελεστή της πυκνότητας (`density`), καθώς, όπως είπαμε, η ίδια η μεταβλητή παίρνει τιμές πάρα πολύ κοντά στο $1$. Έτσι, τελικά, και για τα δύο είδη κρασιού, ο συντελεστής της πυκνότητας λειτουργεί σαν παράγοντας που εξισορροπεί την παρουσία της εξίσου αυξημένης τιμής του σταθερού όρου.

Παραβλέποντας λοιπόν αυτόν τον όρο, μπορούμε να ξεχωρίσουμε ως σημαντικές συμμεταβλητές στα δεδομένα λευκού κρασιού τις ακόλουθες:

- πτητική οξύτητα (`volatile.acidity`), με αρνητική επίδραση στην ποιότητα,
- χλωριούχα (`chlorides`), με αρνητική επίδραση στην ποιότητα,
- αλκοόλ (`alcohol`), με θετική επίδραση στην ποιότητα.

Για τα δεδομένα κόκκινου κρασιού, οι συμμεταβλητές που επηρεάζουν την ποιότητα είναι οι εξής:

- πτητική οξύτητα (`volatile.acidity`), με αρνητική επίδραση στην ποιότητα,
- κιτρικό οξύ (`citric.acid`), με αρνητική επίδραση στην ποιότητα,
- χλωριούχα (`chlorides`), με αρνητική επίδραση στην ποιότητα,
- pH, με θετική επίδραση στην ποιότητα,
- θειϊκά άλατα `sulphates`, με θετική επίδραση στην ποιότητα,
- αλκοόλ (`alcohol`), με θετική επίδραση στην ποιότητα.

Σχολιάζοντας τώρα τα διαθέσιμα δεδομένα και παρόλο που τα μοντέλα που προέκυψαν από αυτά είχαν καλή απόδοση, αναφέρουμε ότι θα μπορούσαν να έχουν συμπεριληφθεί και οι μετρήσεις μεταβλητών πέρα από αυτές που αφορούν φυσικοχημικές ιδιότητες, όπως η χρονιά παραγωγής, ο χρόνος ζύμωσης κ.α. Επίσης, διαπιστώνουμε από τη μορφή της μεταβλητής απόκρισης ότι δεν είναι στην πραγματικότητα συνεχής, αφού οι τιμές που μπορούν να αποδοθούν στην ποιότητα κρασιού είναι τελικά πεπερασμένες. Μάλιστα, στα διαθέσιμα δεδομένα δεν υπήρχαν καταγραφές ποιότητας με δεκαδικά ψηφία, μόνο ακέραιες τιμές. Το γεγονός αυτό επηρεάζει σαφώς το μοντέλο αφού εξ' αρχής υποθέτουμε κανονική κατανομή της εξαρτημένης μεταβλητής, ωστόσο πρόκειται για μια απόλυτα ρεαλιστική εφαρμογή της παλινδρόμησης.

Ειδικότερα, η χρησιμότητα της ανάπτυξης μοντέλου πρόβλεψης της ποιότητας κρασιού δεν περιορίζεται στις μελέτες του οργανισμού που σύλλεξε τα δεδομένα. Αντιθέτως, η ενσωμάτωση αυτού του μηχανισμού στη διαδικασία ελέγχου ποιότητας μπορεί να βοηθήσει σημαντικά τη διαδικασία παραγωγής κρασιού, από τη μεγαλύτερη βιομηχανία μέχρι τον μικρότερο παραγωγό. Ένα απλό δείγμα από το παραγόμενο κρασί μπορεί να εισαχθεί στο κατάλληλο μοντέλο το οποίο θα προβλέψει την ποιότητά του διευκολύνοντας την ανίχνευση χαλασμένων δειγμάτων και την παραγωγή προϊόντων που θα έχουν μεγαλύτερη απήχηση στο καταναλωτικό κοινό.

Κλείνοντας με αυτόν τον τρόπο την παρούσα εργασία, θέλουμε να υπογραμμίσουμε το συναρπαστικό γεγονός ότι τα μαθηματικά μπορούν να τυποποιήσουν και να αυτοματοποιήσουν διαδικασίες από όλες τις εκφάνσεις της ζωής, από τις πιο απλές στις πιο πολύπλοκες, διαδικασίες που παλαιότερα βασίζονταν στην ανθρώπινη διαίσθηση.

# A Κώδικας στην R

## A.1 Προσομοίωση δεδομένων

**Πρώτο σύνολο δεδομένων**

```
n<-50 #number of observations for each covariate
p<-16 #an intercept and 15 covariates

lab<-list(seq(1:n),c("intercept","x1","x2","x3","x4","x5","x6","x7",
                     "x8","x9","x10","x11","x12","x13","x14","x15"))

X<-matrix(nrow=n,ncol=p,dimnames=lab) #creating an empty design matrix
X[,1]<-(rep(1,times=n)) #adding a column of 1s for the intercept

for (i in 2:p) {
  X[,i]<-rnorm(n,0,1) #filling each column with 50 values from N(0,1)
}

#real coefficient values aka beta
coeffs<-as.matrix(c(6,8,0,0,3,0,0,10,0,0,0,0,-12,0,0,4))

#first sample, sigma=1.5
sigma<-1.5
mu<-X%*%coeffs #the mean value
y<-rnorm(n,mu,sigma)
```

**Δεύτερο σύνολο δεδομένων**

```
n<-50 #number of observations for each covariate
p<-16 #an intercept and 15 covariates

lab<-list(seq(1:n),c("intercept","x1","x2","x3","x4","x5","x6","x7",
                     "x8","x9","x10","x11","x12","x13","x14","x15"))

X<-matrix(nrow=n,ncol=p,dimnames=lab) #creating an empty design matrix
X[,1]<-(rep(1,times=n)) #adding a column of 1s for the intercept

for (i in 2:p) {
  X[,i]<-rnorm(n,0,1) #filling each column with 50 values from N(0,1)
}

#real coefficient values aka beta
coeffs<-as.matrix(c(6,8,0,0,3,0,0,10,0,0,0,0,-12,0,0,4))

#second sample, sigma=2.5
sigma<-2.5
mu<-X%*%coeffs
y<-rnorm(n,mu,sigma)
```

**Τρίτο σύνολο δεδομένων**

```
n<-50 #number of observations for each covariate
p<-16 #an intercept and 15 covariates

lab<-list(seq(1:n),c("intercept","x1","x2","x3","x4","x5","x6","x7",
                     "x8","x9","x10","x11","x12","x13","x14","x15"))

X<-matrix(nrow=n,ncol=p,dimnames=lab) #creating an empty design matrix
X[,1]<-(rep(1,times=n)) #adding a column of 1s for the intercept

for (i in 2:11) {
  X[,i]<-rnorm(n,0,1) #filling columns 2-11 with 50 values from N(0,1)
}

mu1<-X[,2:6]%*%c(0.3,0.5,0.7,0.9,1.1)

for (i in 12:p) {
  X[,i]<-rnorm(n,mu1,1) #filling columns 12-16 with 50 values from N(mu1,1)
}

#real coefficient values
coeffs<-as.matrix(c(6,8,0,0,3,0,0,10,0,0,0,0,0,-12,0,0,4))

#third sample, sigma=1.5
sigma<-1.5
mu<-X%*%coeffs
y<-rnorm(n,mu,sigma)
```

**Τέταρτο σύνολο δεδομένων**

```
n<-50 #number of observations for each covariate
p<-16 #an intercept and 15 covariates

lab<-list(seq(1:n),c("intercept","x1","x2","x3","x4","x5","x6","x7",
                     "x8","x9","x10","x11","x12","x13","x14","x15"))

X<-matrix(nrow=n,ncol=p,dimnames=lab) #creating an empty design matrix
X[,1]<-(rep(1,times=n)) #adding a column of 1s for the intercept

for (i in 2:11) {
  X[,i]<-rnorm(n,0,1) #filling columns 2-11 with 50 values from N(0,1)
}

mu1<-X[,2:6]%*%c(0.3,0.5,0.7,0.9,1.1)

for (i in 12:p) {
  X[,i]<-rnorm(n,mu1,1) #filling columns 12-16 with 50 values from N(mu1,1)
}

#real coefficient values
coeffs<-as.matrix(c(6,8,0,0,3,0,0,10,0,0,0,0,0,-12,0,0,4))
```

```
#fourth sample, sigma=2.5
sigma<-2.5
mu<-X%*%coeffs
y<-rnorm(n,mu,sigma)
```

## A.2 Έλεγχος προϋποθέσεων γραμμικού μοντέλου

```
#loading the data
winequality.red<- read.csv("winequality-red.csv",header = TRUE, sep = ";")

#fitting a linear model
LModel<-lm(quality~., data=winequality.red)

##CHECKING LINEAR MODEL ASSUMPTIONS

##1. LINEARITY CHECK
#Prepping the data for linearity check
res_labels<-list(labels(residuals(LModel,"partial"))[[1]],
        paste("Partial residuals",labels(residuals(LModel,"partial"))[[2]],sep=" "))
std_labels<-list(labels(residuals(LModel,"partial"))[[1]],
        paste("Standardized",labels(residuals(LModel,"partial"))[[2]],sep=" "))
partial_residuals<-matrix(data=as.matrix(residuals(LModel,"partial")),
        nrow = dim(residuals(LModel,"partial"))[1],
        ncol=(dim(residuals(LModel,"partial"))[2]),
        dimnames = res_labels)
standardized_data<-matrix(nrow = (dim(residuals(LModel,"partial"))[1]),
        ncol=(dim(residuals(LModel,"partial"))[2]),dimnames = std_labels)

#Plots for linearity check
#Variables 1 through 6
par(mar=c(4,4,2,2))
par(mfrow=c(3,2))

for(i in 1:6){
  standardized_data[,i]<-(winequality.red[,i]-mean(winequality.red[,i]))
        /sd(winequality.red[,i])
  plot(partial_residuals[,i],standardized_data[,i], xlab=res_labels[[2]][i],
        ylab=std_labels[[2]][i],col="midnightblue")
  abline(lm(standardized_data[,i]~partial_residuals[,i]),col="mediumaquamarine",
        lwd = 3)
}

mtext("Red: Linearity check 1/2", side = 3, line = -1.5, outer = TRUE)

#Variables 7 through 11
par(mar=c(4,4,2,2))
par(mfrow=c(3,2))

for(i in 7:11){
  standardized_data[,i]<-(winequality.red[,i]-mean(winequality.red[,i]))
        /sd(winequality.red[,i])
```

```
  plot(partial_residuals[,i],standardized_data[,i], xlab=res_labels[[2]][i],
       ylab=std_labels[[2]][i],col="midnightblue")
  abline(lm(standardized_data[,i]~partial_residuals[,i]),col="mediumaquamarine",
       lwd = 3)
}

mtext("Red: Linearity check 2/2", side = 3, line = -1.5, outer = TRUE)

#2. NORMALITY OF RESIDUALS CHECK
#Checking normality assumptions
par(mfrow=c(1,1))
qqnorm(LModel$residuals,col="paleturquoise4", main="Red: Normal Q-Q plot")
qqline(LModel$residuals,col="yellowgreen",lwd=3)

#3. HOMOSCEDASTICITY CHECK
#Checking homoscedasticity
par(mfrow=c(1,1))
plot(rstudent(LModel),fitted(LModel),xlab = "studentized residuals",
       ylab="fitted values",main="Red: Residual homoscedasticity check",
       col="orange")

#4. RESIDUAL INDEPENDENCE HYPOTHESIS
#Checking residual independence
par(mfrow=c(1,1))
plot(1:dim(winequality.red)[[1]],LModel$residuals,xlab="index",
       ylab="residuals",main="Red: Residual independence check",
       col="cyan")
```

## A.3 Υλοποίηση Μπεϋζιανής παλινδρόμησης

### A.3.1 Κανονική-αντίστροφη γάμμα πρότερη

```
##NIG PRIOR
library(matrix)

winequality.white<- read.csv("winequality-white.csv",header = TRUE, sep = ";")

n<-dim(winequality.white)[[1]]
p<-dim(winequality.white)[[2]]

X<-as.matrix(cbind(rep(1,times=n),winequality.white[,1:11]))
y<-as.matrix(winequality.white$quality)

#Zellner's prior hyperparameter
g<-n

#Normal prior for coefficients
mu_beta<-as.matrix(rep(0,times=p))
V_beta<-as.matrix(g*solve(t(X)%*%X))

#Inverse gamma prior for variance
a<-0.001 #shape
```

```
b<-0.001 #scale

#posterior parameters
V_new<-solve(t(as.matrix(X))%*%X+solve(V_beta))
mu_new<-V_new%*%(solve(V_beta)%*%mu_beta+t(X)%*%y)
a_new<-a+n/2
b_new<-b+(t(y)%*%y+t(mu_beta)%*%solve(V_beta)%*%mu_beta
        -t(mu_new)%*%solve(V_new)%*%mu_new)/2
```

### A.3.2 Improper πρότερη για τη διασπορά

```
##IMPROPER PRIOR FOR VARIANCE
library(matlib)

winequality.white<- read.csv("winequality-white.csv",header = TRUE, sep = ";")

n<-dim(winequality.white)[[1]]
p<-dim(winequality.white)[[2]]

X<-as.matrix(cbind(rep(1,times=n),winequality.white[,1:11]))
y<-as.matrix(winequality.white$quality)

##IMPROPER PRIOR FOR VARIANCE
#Zellner's prior hyperparameter
g=n

#Normal prior for coefficients
mu_beta<-rep(0,times=p)
V_beta<-g*inv(t(X)%*%X)

#Inverse gamma prior for variance
a<-0.001 #shape
b<-0.001 #scale

#posterior parameters
V_new<-solve(t(as.matrix(X))%*%as.matrix(X)+solve(as.matrix(V_beta)))
mu_new<-V_new%*%(solve(as.matrix(V_beta))%*%mu_beta+t(as.matrix(X))%*%y)
a_new<-a
b_new<-b+(t(y)%*%y+t(mu_beta)%*%solve(as.matrix(V_beta))%*%mu_beta
        -t(mu_new)%*%solve(V_new)%*%mu_new)/2
```

### A.3.3 Μη πληροφοριακή πρότερη

```
library(matrix)

winequality.white<- read.csv("winequality-white.csv",header = TRUE, sep = ";")

n<-dim(winequality.white)[[1]]
p<-dim(winequality.white)[[2]]

X<-as.matrix(cbind(rep(1,times=n),winequality.white[,1:11]))
```

```
y<-as.matrix(winequality.white$quality)

##NON INFORMATIVE PRIOR
#Non-informative prior for coefficients
mu_beta<-rep(0,times=p)

#Non-informative prior for variance
a<-p/2 #shape
b<-0.001 #scale

#posterior parameters
V_new<-solve(t(as.matrix(X))%*%as.matrix(X))
mu_new<-V_new%*%t(X)%*%y
a_new<-(n-p)/2
b_new<-(t(y)%*%y-t(mu_new)%*%solve(V_new)%*%mu_new)/2
```

## A.4 Προβλέψεις

```
#loading the library for RMSE calculation
library(Metrics)

#loading the library for matrix calculations
library(matlib)

#loading data
winequality.white<- read.csv("winequality-white.csv",header = TRUE, sep = ";")

n<-dim(winequality.white)[[1]]
p<-dim(winequality.white)[[2]]

X<-as.matrix(cbind(rep(1,times=n),winequality.white[,1:11]))
actual<-as.matrix(winequality.white$quality)

#initializing the array which will contain the fitted values of the dependent variable
predicted <- NULL

#initializing the RMSE array
rmse<- NULL
```

### A.4.1 Κανονική-αντίστροφη γάμμα πρότερη

```
#leave-one-out Cross validation and model fitting using NIG prior
for(i in 1:n){
  #leaving one out
  validation<-X[i,]
  training<-X[-i,]
  y_training<-actual[-i,]

  #NIG prior
  mu_beta<-as.matrix(rep(0,times=p))
  V_beta<-as.matrix(g*solve(t(training)%*%training))
```

```
  #posterior parameter estimation using NIG prior
  V_new<-solve(t(as.matrix(training))%*%training+solve(V_beta))
  mu_new<-V_new%*%(solve(V_beta)%*%mu_beta+t(training)%*%y_training)

  #predicting dependent variable values
  predicted[i]<-validation%*%mu_new
}

#calculating Root Mean Square Error
white_NIG_rmse<-rmse(actual,predicted)
```

### A.4.2 Improper πρότερη για τη διασπορά

```
#leave-one-out Cross validation and model fitting using Improper prior
for(i in 1:n){
  #leaving one out
  validation<-X[i,]
  training<-X[-i,]
  y_training<-actual[-i,]

  #Improper prior
  mu_beta<-as.matrix(rep(0,times=p))
  V_beta<-as.matrix(g*solve(t(training)%*%training))

  #posterior parameter estimation using Improper prior
  V_new<-solve(t(as.matrix(training))%*%training+solve(V_beta))
  mu_new<-V_new%*%(solve(V_beta)%*%mu_beta+t(training)%*%y_training)

  #predicting dependent variable values
  predicted[i]<-validation%*%mu_new
}

#calculating Root Mean Square Error
white_improper_rmse<-rmse(actual,predicted)
```

### A.4.3 Μη πληροφοριακή πρότερη

```
#leave-one-out Cross validation and model fitting using Non-informative prior
for(i in 1:n){
  #leaving one out
  validation<-X[i,]
  training<-X[-i,]
  y_training<-actual[-i,]

  #posterior parameter estimation using Non-informative prior
  V_new<-solve(t(as.matrix(training))%*%as.matrix(training))
  mu_new<-V_new%*%t(training)%*%y_training

  #predicting dependent variable values
  predicted[i]<-validation%*%mu_new
```

```
}

#calculating Root Mean Square Error
white_non_rmse<-rmse(actual,predicted)
```

# Βιβλιογραφικές Αναφορές στα Ελληνικά


Καρώνη, Χ. & Οικονόμου, Π. (2017). *Στατιστικά Μοντέλα Παλινδρόμησης*. Εκδόσεις Συμεών.
Κοκολάκης, Γ. & Φουσκάκης, Δ. (2009). *Στατιστική Θεωρία και Εφαρμογές*. Εκδόσεις Συμεών.
Λουλάκης, Μ. (2015). *Στοχαστικές Διαδικασίες*. Εκδόσεις Κάλλιπος.
Φουσκάκης, Δ. (2013). *Ανάλυση Δεδομένων με Χρήση της R*. Εκδόσεις Τσότρας.


# Διεθνείς Βιβλιογραφικές Αναφορές


Banerjee, S. (2008). *Bayesian Linear Model : Gory Details. Pubh7440 Notes*. URL: `https://bit.ly/3ANtxTE`.

Bayes, T. & Price, R. (1763). "An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S." In: *Philosophical Transactions (1683-1775)* 53, pp. 370–418.

Bellhouse, D.R. (2004). "The Reverend Thomas Bayes, FRS: A Biography to Celebrate the Tercentenary of His Birth". In: *Statistical Science* 19.1, pp. 3–32.

Bernando, J. & Smith, A. (2000). *Bayesian Theory*. John Wiley & Sons.

Bolstad, W. (2010). *Understanding Computational Bayesian Statistics*. John Wiley & Sons.

Casella, G., Robert, C., & Wells, M. (2004). "Generalized Accept-Reject sampling schemes". In: *Institute of Mathematical Statistics Lecture Notes - Monograph Series* 45, pp. 342–347.

Cortez, P., Cerdeira, A., Almeida, A., Matos, T., & Reis, J. (2009). "Modeling wine preferences by data mining from physicochemical properties". In: *Decision Support Systems* 47.4, pp. 547–553.

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression*. Berlin, Heidelberg: Springer.

Fernandez, C., Ley, E., & Steel, M. (2001). "Benchmark priors for Bayesian model averaging". In: *Journal of Econometrics* 100.2, pp. 381–427.

Gamerman D., Lopes H. (2006). *Markov Chain Monte Carlo Stochastic Simulation for Bayesian Inference*. Second Edition. Chapman and Hall/CRC.

Gelman, A., Roberts, G.O., & Gilks, W.R. (1996). "Efficient Metropolis jumping rules". In: *Bayesian Statistics* 5, pp. 599–608.

Geman, S. & Geman, D. (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". In: *IEEE Transactions on pattern analysis and machine intelligence* 6, pp. 721–741.

Hájek (2019). "Interpretations of Probability". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2019. Metaphysics Research Lab, Stanford University.

Hastings, W.K. (1970). "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1, pp. 97–109.

Jeffreys, H. (1961). *Theory of probability*. International series of monographs on physics (Oxford, England). Oxford : Clarendon Press.

Kokolakis, G. (2010). "Bayesian Statistical Analysis". In: *International Encyclopedia of Education*. Third Edition. Oxford: Elsevier, pp. 37–45.

Liu, H. & Wasserman, L. (2014). “Bayesian Inference”. In: Statistical Machine Learning (to be published). Chap. 12.

Lynch, S. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. Springer, New York, NY.

Marin, J.M. & Robert, C. (2014). *Bayesian Essentials with R*. Second Edition. Springer Texts in Statistics. New York: Springer-Verlag.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, H., & Teller, E. (1953). “Equation of State Calculations by Fast Computing Machines”. In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092.

Metropolis, N. & Ulam, S. (1949). “The Monte Carlo Method”. In: *Journal of the American Statistical Association* 44.247, pp. 335–341.

Ntzoufras, I. (2009). *Bayesian Modeling Using WinBUGS*. New York: John Wiley & Sons.

O’Connor, J. & Robertson, E. (2004). *Thomas Bayes Biography*. URL: `https://bit.ly/3o9JmR8`.

O’Hagan, A. & Forster, J. (2004). *Kendall’s Advanced Theory of Statistics, volume 2B: Bayesian Inference*. Second Edition. London, UK: Arnold.

Richey, M. (2010). “The Evolution of Markov Chain Monte Carlo Methods”. In: *The American Mathematical Monthly* 117.5, pp. 383–413.

Robert, C. & Casella, G. (2011). “A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data”. In: *Statistical Science* 26.1.

Sherlock, C., Fearnhead, P., & Roberts, G. (2010). “The Random Walk Metropolis: Linking Theory and Practice Through a Case Study”. In: *Statistical Science* 25.2.

Zellner, A. (1986). “On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions”. In: *Studies in Bayesian Econometrics and Statistics* 6: *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pp. 233–243.

Zellner, A. (1988). “Optimal Information Processing and Bayes’s Theorem”. In: *The American Statistician* 42.4, pp. 278–280.

Zhu, M. & Lu, A. (2004). “The Counter-intuitive Non-informative Prior for the Bernoulli Family”. In: *Journal of Statistics Education* 12.