Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ιωάννης Ντάγκας

Μικτά Γραμμικά Μοντέλα

Τριμελής Επιτροπή:

Δημήτρης Φουσκάκης, Καθηγητής (Επιβλέπων)
Σχολή Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών

Μιχαήλ Λουλάκης, Καθηγητής
Σχολή Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών

Κωνσταντίνος Χρυσαφίνος, Καθηγητής
Σχολή Εφαρμοσμένων Μαθηματικών & Φυσικών Επιστημών

Αθήνα, Σεπτέμβριος 2023

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κ. Δημήτρη Φουσκάκη, για όλη τη βοήθεια και την καθοδήγησή του κατά τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας. Επίσης θα ήθελα να ευχαριστήσω τους γονείς μου, Χρήστο και Ρεβέκκα και τον αδερφό μου Παναγιώτη για την στήριξη και την υπομονή τους όλο αυτό το καιρό.

....

Ιωάννης Ντάγκας

Περίληψη

Σκοπός της παρούσας Διπλωματικής εργασίας, είναι η πλήρης επεξήγηση και ανάλυση ενός ιδιαίτερου στατιστικού μοντέλου, του μικτού γραμμικού μοντέλου. Τα θέματα συζήτησης και τα σημεία αναφοράς κατά τη μελέτη στατιστικών μοντέλων, είναι γνωστά και δεν επιδέχονται ιδιαίτερες αλλαγές από στατιστικό μοντέλο σε στατιστικό μοντέλο.

Επομένως θα ξεκινήσουμε την παρούσα εργασία, κάνοντας κατανοητό στους αναγνώστες, τι είναι ένα μικτό γραμμικό μοντέλο, ποιές είναι οι τρείς του βασικές κατηγορίες και πότε πρέπει να χρησιμοποιείται. Στην συνέχεια θα ορίσουμε μαθηματικά το μικτό γραμμικό μοντέλο και θα προσπαθήσουμε να εκτιμήσουμε τις παραμέτρους του με γνωστές μεθόδους στατιστικής ανάλυσης.

Καθώς τα στατιστικά πακέτα που έχουν αναπτυχθεί τα τελευταία χρόνια διευκολύνουν σε τεράστιο βαθμό τις αναλύσεις μας, θα δείξουμε μέσω παραδειγμάτων πως μπορούμε να προσαρμόσουμε ένα μικτό γραμμικό μοντέλο με την βοήθεια της R. Έπειτα θα είμαστε σε θέση να υπολογίσουμε και τις παραμέτρους του μοντέλου, ενώ ιδιαίτερη βαρύτητα θα δοθεί στην ερμηνεία των αποτελέσματων μας.

Τέλος καθώς το μοντέλο μας επιτρέπεται να έχει και μόνο κατηγορικές μεταβλητές, θα περιγράψουμε την μικτή ανάλυση διασποράς και θα γίνει αναφορά στις υποθέσεις που πρέπει να επαληθεύουν τα δεδομένα μας για την πραγματοποίηση της. Έπειτα θα μεταβούμε στην κύρια ανάλυση με την βοήθεια της R, ώστε να προσαρμόσουμε στατιστικά το μοντέλο μας και να καταλήξουμε σε χρήσιμα συμπεράσματα για το πρόβλημα μελέτης μας.

Abstract

The purpose of this dissertation is the full explanation and analysis of a particular statistical model, the mixed linear model. The topics of discussion and points of reference when studying statistical models are well known and are not subject to special changes from statistical model to statistical model.

We will therefore begin this dissertation by making readers understand what a mixed linear model is, what its three main categories are and when it should be used. We will then mathematically define the mixed linear model and attempt to estimate its parameters using well-known methods of statistical analysis.

As the statistical packages that have been developed in recent years greatly facilitate our analyses, we will show through examples how we can fit a mixed linear model with the help of R. We will then be able to estimate the parameters of the model, while particular attention will be paid to the interpretation of our results.

Finally, as our model is allowed to have only categorical variables, we will describe the mixed analysis of variance and discuss the assumptions that our data must verify in order to perform it. We will then move on to the main analysis using R to statistically fit our model and come to useful conclusions for our study problem.

Περιεχόμενα

Εισαγωγή

Σε μια εποχή που ορίζεται από την πληροφόρηση και τη λήψη αποφάσεων βάσει δεδομένων, ο τομέας της στατιστικής αναδεικνύεται σε πρωταρχικό παράγοντα εξαγωγής χρήσιμων πληροφοριών. Η στατιστική μας δίνει την δυνατότητα να λαμβάνουμε ακατέργαστα δεδομένα και να τα μετατρέπουμε σε ουσιαστικές γνώσεις, επιτρέποντας μας να κάνουμε προβλέψεις και ερμηνείες.

Ένα μεγάλο μέρος της επιστήμης της στατιστικής, ασχολείται με την κατασκευή και ανάλυση στατιστικών μοντέλων. Ένα στατιστικό μοντέλο, προσπαθεί να αποτυπώσει ένα πραγματικό πρόβλημα ή φαινόμενο με την χρήση μαθηματικών μοντέλων και στατιστικών υποθέσεων. Ένας ερευνητής από την μεριά του, προσπαθεί σε πρώτη φάση να εντοπίσει και ύστερα να ορίσει τις σχετικές μεταβλητές που πιστεύει ότι επηρεάζουν το πρόβλημα μελέτης του, και στην συνέχεια να επιλέξει το κατάλληλο στατιστικό μοντέλο κατά τον ίδιο για να περιγράψει την σχέση των μεταβλητών αυτών. Καθώς κανένα πραγματικό πρόβλημα δεν είναι εντελώς προβλέψιμο, τα στατιστικά μοντέλα ενσωματώνουν την αβεβαιότητα αυτή, με την ύπαρξη των τυχαίων σφαλμάτων.

Στην σύγχρονη εποχή, ερχόμαστε αντιμέτωποι με μια πληθώρα ερωτημάτων που απαιτούν απάντηση και με ένα τεράστιο όγκο διαθέσιμων δεδομένων. Οι ερευνητές ως απάντηση αυτού, έχουν αναπτύξει ένα ευρύ φάσμα στατιστικών μοντέλων για την αντιμετώπιση διαφορετικών τύπων δεδομένων, ερευνητικών ερωτημάτων και προβλημάτων. Ο τομέας της στατιστικής για τον παραπάνω λόγο, είναι ένας κλάδος που συνεχώς εξελίσσεται για να είναι σε θέση να αντιμετωπίσει με επιτυχία όλες τις νέες προκλήσεις.

Η ραγδαία εξέλιξη της τεχνολογίας έχει οδηγήσει σε τεράστια μεταβολή στην ποσότητα και διαθεσιμότητα των δεδομένων που έχουμε στην κατοχή μας. Από εκεί που πρίν λίγες δεκαετίες είχαμε πρόσβαση σε πολύ περιορισμένο αριθμό δεδομένων, πλέον έχουμε πρόσβαση σε τεράστιο όγκο δεδομένων. Μαζί με την εξέλιξη της τεχνολογίας, ήρθε και η εξέλιξη των υπολογιστών και ως αποτέλεσμα αυτού, η δημιουργία προγραμματιστικών προγραμμάτων σχεδιασμένα ειδικά για στατιστικούς υπολογισμούς και ανάλυση δεδομένων, κάνοντας πιο εύκολη την καθημερινότητα των στατιστικολόγων. Η ανάπτυξη των προγραμματιστικών προγραμμάτων ήταν αναγκαία για την αντιμετώπιση του τεράστιου όγκου δεδομένων και αποτελεί ένα ισχυρό εργαλείο και μια απαραίτητη γνώση για τους στατιστικολόγους.

Κεφάλαιο 1

Εισαγωγή στην θεωρία των Κανονικών Μικτών Γραμμικών Μοντέλων

1.1 Εισαγωγή

Το πρώτο κεφάλαιο θα αποτελέσει μία εισαγωγή στα κανονικά μικτά γραμμικά μοντέλα που θα μας απασχολήσουν καθ' όλη την διάρκεια της διπλωματικής εργασίας. Πιο συγκεκριμένα θα δοθεί έμφαση στην περιγραφή τους, ώστε με την ολοκλήρωση του κεφαλαίου να υπάρχει μια σαφή πρώτη εικόνα του αναγνώστη για τα μικτά γραμμικά μοντέλα, έτσι ώστε η μετάβαση στα επόμενα κεφάλαια να γίνει με ομαλό τρόπο.

Ωστόσο η αφετηρία του κεφαλαίου θα αφορά την περιγραφή ενός άλλου γνωστού μοντέλου, του μοντέλου σταθερών επιδράσεων, καθώς υπάρχει σύνδεση με τα μικτά γραμμικά μοντέλα. Στην συνέχεια θα γίνει η εισαγωγή στην πρώτη κατηγορία μικτών γραμμικών μοντέλων, το μοντέλο τυχαίων επιδράσεων και λόγω της σύνδεσης του με το μοντέλο σταθερών επιδράσεων, θα καταγραφούν οι διαφορές τους και θα αναλυθεί πλήρως πότε είναι κατάλληλη η χρησιμοποίηση του ενός μοντέλου έναντι του άλλου. Επίσης θα γίνει αναφορά στα τρία διαφορετικά είδη συγχωνεύσεων (pooling), τα οποία με την σειρά τους θα μας οδηγήσουν σε μία ακόμα σημαντική διαφορά μεταξύ ενός μοντέλου σταθερών επιδράσεων και ενός μικτού γραμμικού μοντέλου.

Τέλος θα ασχοληθούμε και θα περιγράψουμε πλήρως τις δυο εναπομείναντες κατηγορίες μικτών γραμμικών μοντέλων, το μοντέλο μοτίβου συνδιακυμάνσεων και το μοντέλο τυχαίων συντελεστών.

1.2 Μοντέλο Σταθερών Επιδράσεων

Στη συγκεκριμένη ενότητα θα περιγράψουμε το μοντέλο σταθερών επιδράσεων. Όπως σε κάθε στατιστικό μοντέλο, έτσι και στο μοντέλο σταθερών επιδράσεων χρησιμοποιούμε τις μεταβλητές που θεωρούμε στατιστικά σημαντικές για το πρόβλημα και τα δεδομένα που μελετάμε. Πιο συγκεκριμένα στο μοντέλο σταθερών επιδράσεων, θεωρούμε την επίδραση μιας μεταβλητής ως σταθερή επίδραση. Οι συντελεστές μιας σταθερής επίδρασης καλούνται παράμετροι (ή συντελεστές) σταθερών επιδράσεων, με την ιδιότητα που τους ξεχωρίζει να αποτελεί ότι λαμβάνουν σταθερές τιμές. Διαφορετικά μπορούμε να πούμε, ότι οι τιμές των συντελεστών είναι σταθερές για όλες τις παρατηρήσεις εντός μιας ομάδας ή εντός μιας κατηγορίας. Αυτό έχει ως αποτέλεσμα η σχέση μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών, να είναι η ίδια για όλα τα αντικείμενα μελέτης ή τα υποκείμενα του προβλήματος, εντός του ίδιου επιπέδου μιας κατηγορικής μεταβλητής. Απλοϊκά η μορφή του μοντέλου μπορεί να γραφεί ως εξής:

$$\text{Μοντέλο Σταθερών Επιδράσεων} = \text{Σταθερά} + \text{Σταθερές Επιδράσεις} + \text{Τυχαίο Σφάλμα}.$$

Αξίζει να σημειωθεί ότι δεν υπάρχει περιορισμός ως προς τον τύπο μιας μεταβλητής για να θεωρηθεί ότι η επίδραση της είναι σταθερή. Δηλαδή είτε η μεταβλητή είναι κατηγορική είτε ποσοτική, μπορεί εξίσου η επίδραση της να θεωρηθεί ως σταθερή επίδραση. Η διαφορά που θα προκύψει θα έχει να κάνει με τους συντελεστές τους, καθώς στην περίπτωση της ποσοτικής μεταβλητής έχουμε έναν συντελεστή, ενώ στην περίπτωση της κατηγορικής μεταβλητής ο αριθμός των συντελεστών ισούται με τον αριθμό των εικονικών μεταβλητών της αντίστοιχης κατηγορικής μεταβλητής. Στην συνέχεια θα εισάγουμε τον αναγνώστη σε ένα παράδειγμα, το οποίο θα χρησιμοποιηθεί και στις επόμενες ενότητες.

Παράδειγμα 1.2

Σκοπός της μελέτης είναι η σύγκριση δυο διαφορετικών μηχανημάτων Lazer που χρησιμοποιούνται στο χειρουργείο για την διόρθωση της μυωπίας. Έχουμε δεδομένα που αντιστοιχούν σε δέκα διαφορετικά ιατρικά κέντρα, τα οποία χρησιμοποιούν και τα δυο μηχανήματα. Επίσης για κάθε άτομο που υποβάλλεται στην χειρουργική επέμβαση, γνωρίζουμε προφανώς ποιο ιατρικό κέντρο επισκέφτηκε και ποιο μηχάνημα Lazer χρησιμοποίησε, την ηλικία του, τους βαθμούς μυωπίας πριν την χειρουργική επέμβαση (μέσος όρος μυωπίας των δυο ματιών) καθώς και έπειτα από τρείς μήνες. Δημιουργούμε το στατιστικό μοντέλο, λαμβάνοντας υπόψη τις εξής μεταβλητές:

- Ηλικία ανά γκρουπ (18-25,25-35,35-50,50+).
- Μυωπία πριν την χειρουργική επέμβαση.
- Μηχανήματα Lazer (A ή B).
- Ιατρικά Κέντρα (Ιατρικό Κέντρο 1, Ιατρικό Κέντρο 2, ...).

Συνοπτικά μπορούμε να αναφέρουμε ότι η μυωπία πριν την χειρουργική επέμβαση αποτελεί την μοναδική μας συνεχή μεταβλητή. Αντίθετα οι υπόλοιπες τρείς μεταβλητές είναι κατηγορικές, περιέχοντας τέσσερις, δυο και δέκα επιμέρους υποκατηγορίες αντίστοιχα. Αν επιθυμούμε να γίνουμε λίγο πιο συγκεκριμένοι για την επίδραση του ιατρικού κέντρου, θα πρέπει να προστεθεί στο μοντέλο μας καθώς ενδέχεται να υπάρχουν διαφορές μεταξύ των ιατρικών κέντρων που μπορεί να οφείλονται σε ιατρικούς λόγους ή σε λόγους που αγνοούμε. Η εξαρτημένη μεταβλητή στο πρόβλημα αυτό είναι η μυωπία μετά το πέρας των τριών μηνών από την χειρουργική επέμβαση, η οποία προφανώς και είναι συνεχής.

Αν οι επιδράσεις των παραπάνω μεταβλητών θεωρηθούν σταθερές επιδράσεις, τότε καταλήγουμε σε ένα μοντέλο σταθερών επιδράσεων. Συνηθίζεται όταν θεωρούμε την επίδραση μιας κατηγορικής μεταβλητής ως σταθερή, να αναφερόμαστε σε αυτήν ως κατηγορική σταθερή επίδραση, ενώ στην περίπτωση της ποσοτικής μεταβλητής, ως ποσοτική σταθερή επίδραση. Τέλος αντιστοιχεί ένας συντελεστής για την συνεχή σταθερή επίδραση της μυωπίας πριν την χειρουργική επέμβαση, ενώ αντιστοιχούν τρείς για την σταθερή επίδραση της ηλικίας.

1.3 Μοντέλο Τυχαίων Επιδράσεων

Στην προηγούμενη ενότητα αναφερθήκαμε στην περίπτωση που η επίδραση μιας μεταβλητής θεωρείται σταθερή επίδραση, με τους συντελεστές της να λαμβάνουν σταθερές τιμές. Ωστόσο υπάρχει η δυνατότητα κάτω από συγκεκριμένες προϋποθέσεις, να θεωρήσουμε την επίδραση μιας μεταβλητής ως τυχαία επίδραση. Οι συντελεστές μιας τυχαίας επίδραση καλούνται παράμετροι (ή συντελεστές) τυχαίων επιδράσεων και υποχρεωτικά ακολουθούν μια κατανομή. Στην περίπτωση μας καθώς μελετάμε κανονικά μικτά γραμμικά μοντέλα, η κατανομή που ακολουθούν είναι η κανονική.

Επομένως για να καταλήξουμε σε ένα μικτό γραμμικό μοντέλο, πρέπει να θεωρήσουμε στο μοντέλο μας τουλάχιστον μία παράμετρο ως σταθερή επίδραση και τουλάχιστον μία ως τυχαία. Αν δηλαδή χρησιμοποιήσουμε το Παράδειγμα 1.2 και θεωρήσουμε την επίδραση των μηχανημάτων Lazer ως σταθερή επίδραση και την επίδραση των ιατρικών κέντρων ως τυχαία, τότε θα καταλήξουμε σε ένα μικτό γραμμικό μοντέλο. Καθώς έχουμε και άλλες κατηγορίες μικτών γραμμικών μοντέλων που θα αναλυθούν στην συνέχεια, το συγκεκριμένο το οποίο είναι και το πιο διαδεδομένο ονομάζεται μοντέλο τυχαίων επιδράσεων. Πολύ απλά μπορεί να γραφεί ως εξής:

Μοντέλο Τυχαίων Επιδράσεων = Σταθερά + Σταθερές Επιδράσεις + Τυχαίες Επιδράσεις + Τυχαίο Σφάλμα.

Μία μεταβλητή πρέπει υποχρεωτικά να είναι κατηγορική και όχι ποσοτική, για να θεωρηθεί η επίδραση της τυχαία. Επομένως σε οποιοδήποτε μοντέλο τυχαίων επιδράσεων, θα έχουμε εκτός από την σταθερά και τουλάχιστον δυο συντελεστές τυχαίων επιδράσεων.

Έπειτα από τον ορισμό μιας τυχαίας επίδρασης, το ερώτημα που έχει προκύψει έχει να κάνει με το πότε οι επίδρασεις των μεταβλητών πρέπει να θεωρηθούν ως σταθερές επιδράσεις και πότε ως τυχαίες. Με άλλα λόγια πότε πρέπει να χρησιμοποιήσουμε το μοντέλο σταθερών επιδράσεων και πότε το μοντέλο τυχαίων επιδράσεων. Αναλυτική απάντηση θα δοθεί στην επόμενη ενότητα.

Υποθέτουμε ότι έχουμε μία τυχαία επίδραση, με έναν από τους συντελεστές της να συμβολίζεται ως β_i. Τότε ο συντελεστής β_i ορίζεται ως εξής:

$\beta_i \sim N(0, \sigma_\beta^2)$,

δηλαδή ακολουθεί την κανονική κατανομή με μέση τιμή μηδέν και διακύμανση σ_β^2, με την διακύμανση να αντιπροσωπεύει την μεταβλητότητα της τυχαίας επίδρασης. Αξίζει να σημειωθεί ότι όλοι οι συντελεστές μιας τυχαίας επίδρασης έχουν κοινή διακύμανση, επομένως ακολουθούν ακριβώς την ίδια κατανομή.

Αμέσως μετά τον ορισμό μιας παραμέτρου τυχαίων επιδράσεων, γίνεται κατανοητό ότι κατά τον υπολογισμό της διακύμανσης μιας μεμονωμένης παρατήρησης στο μοντέλο τυχαίων επιδράσεων, εκτός από την διακύμανση των υπολοίπων, έχουμε και την προσθήκη της διακύμανσης των τυχαίων επιδράσεων. Προφανώς οι σταθερές επιδράσεις δεν έχουν καμία συνεισφορά, καθώς είναι σταθεροί αριθμοί.

Στην συνέχεια θα καταλήξουμε σε ένα χρήσιμο συμπέρασμα για την συνδιακύμανση δυο διαφορετικών παρατηρήσεων. Για να γίνει πιο εύκολα κατανοητό θα χρησιμοποιηθεί το Παράδειγμα 1.2, για το οποίο θα υποθέσουμε την επίδραση του ιατρείου ως την μοναδική μας τυχαία επίδραση. Επομένως στην περίπτωση που έχουμε παρατηρητές που επισκέπτονται κοινό ιατρείο, χωρίς να μας ενδιαφέρουν οι υπόλοιπες μεταβλητές καθώς αποτελούν σταθερές επιδράσεις και επομένως οι συντελεστές τους είναι σταθεροί αριθμοί, η συνδιακύμανση οποιονδήποτε δυο ισούται με την διακύμανση του συντελεστή της τυχαίας επίδρασης του συγκεκριμένου ιατρείου, η οποία ωστόσο είναι κοινή για όλα τα ιατρεία. Σε αντίθετη περίπτωση, δηλαδή αν είχαμε ασθενείς από διαφορετικά ιατρεία η συνδιακύμανση θα ισούται με μηδέν, λόγω της ανεξαρτησίας των παρατηρήσεων από διαφορετικά ιατρεία. Γενικότερα θα λέγαμε ότι όταν έχουμε μία τυχαία επίδραση, ανεξάρτητα με τον αριθμό των σταθερών επιδράσεων, η συνδιακύμανση δυο παρατηρήσεων που ανήκουν στην ίδια υποομάδα της μεταβλητής που έχει τυχαία επίδραση, ισούται με την διακύμανση του συντελεστή της τυχαίας επίδρασης της αντίστοιχης υποομάδας, ενώ σε αντίθετη περίπτωση ισούται με μηδεν.

1.4 Σύγκριση Μοντέλου Σταθερών Επιδράσεων και Μοντέλου Τυχαίων Επιδράσεων

Οι δυο προηγούμενες ενότητες βοήθησαν τον αναγνώστη στην θεωρητική κατανόηση του μοντέλου σταθερών επιδράσεων και του μοντέλου τυχαίων επιδράσεων. Στην συγκεκριμένη ενότητα θα επικεντρωθούμε στην χρησιμότητα κάθε μοντέλου, απαντώντας ταυτόχρονα στο ερώτημα πότε πρέπει να χρησιμοποιείται το μοντέλο τυχαίων επιδράσεων και πότε το μοντέλο σταθερών επιδράσεων που θέσαμε στο Κεφάλαιο 1.3. Εναλλακτικά καλούμαστε να απαντήσουμε στο ερώτημα πότε θεωρούμε την επίδραση μιας μεταβλητής ως σταθερή και πότε ως τυχαία. Αξίζει να σημειωθεί ότι αποτελεί μία πολύ σημαντική ενότητα, αφού τα δυο αυτά μοντέλα είναι διαφορετικά και εσφαλμένη επιλογή θα επιφέρει λανθασμένα συμπεράσματα.

Μία σταθερή επίδραση έχει την δυνατότητα να είναι είτε ποσοτική είτε κατηγορική, ενώ μία τυχαία επίδραση μπορεί να είναι μόνο κατηγορική. Επομένως το ερώτημα για το πότε η επίδραση μίας μεταβλητής πρέπει να θεωρηθεί ως τυχαία και πότε ως σταθερή, περιορίζεται μόνο για τις κατηγορικές μεταβλητές.

Στην συνέχεια θα γίνει αναφορά στην καθοριστική σημασία του ερωτήματος του προβλήματος μας στην επιλογή μοντέλου. Για να προχωρήσουμε σε μια στατιστική μελέτη, προφανώς έχουμε ή μας έχουν ορίσει το πρόβλημα που εξετάζουμε, τα δεδομένα και προφανώς ποιο είναι το ζητούμενο. Στην περίπτωση μας παίζει σημαντικό ρόλο το ζητούμενο. Έτσι αν το συμπέρασμα μας θέλουμε να αφορά τα δεδομένα και μόνο που διαθέτουμε ή αλλιώς τις επιμέρους κατηγορίες που μας δίνονται, τότε η επιλογή είναι το μοντέλο σταθερών επιδράσεων. Αντίθετα αν το συμπέρασμα μας θέλουμε να αφορά έναν ευρύτερο πλυθυσμό που δεν έχουμε στα δεδομένα μας, τότε η κατάλληλη επιλογή είναι το μοντέλο τυχαίων επιδράσεων.

Για να γίνει απόλυτα κατανοητή η παραπάνω αναφορά θα χρησιμοποιήσουμε το Παράδειγμα 1.2. Υποθέτουμε ότι το ζητούμενο στο παράδειγμα μας είναι να καταλήξουμε στο πιο αποδοτικό μηχάνημα Lazer για την διόρθωση της μυωπίας. Επίσης υποθέτουμε ότι περισσότερα από δέκα ιατρικά κέντρα χρησιμοποιούν τα δυο μηχανήματα Lazer, ωστόσο έχουμε δεδομένα μόνο για δέκα από αυτά. Αν μας ζητείτε να καταλήξουμε στο παραπάνω συμπέρασμα βασισμένη στα δέκα ιατρικά κέντρα και μόνο που έχουμε στα δεδομένα μας, τότε θεωρούμε την επίδραση των ιατρικών κέντρων ως σταθερή και η επιλογή μας είναι το μοντέλο σταθερών επιδράσεων. Αντίθετα αν επιθυμούμε να καταλήξουμε σε ένα γενικότερο συμπέρασμα για την αποδοτικότητα των μηχανημάτων Lazer σε όλα τα ιατρικά κέντρα που χρησιμοποιούνται, τότε καλούμαστε να θεωρήσουμε την επίδραση των ιατρικών κέντρων ως τυχαία. Πραγματοποιώντας αυτήν την προσαρμογή, έχουμε επιλέξει να χρησιμοποιήσουμε το μοντέλο τυχαίων επιδράσεων. Τέλος με βάση τον ορισμό του προβλήματος, καθώς επιθυμούμε να καταλήξουμε σε συμπεράσματα για τα δυο συγκεκριμένα μηχανήματα Lazer, A και B, η επίδραση των μηχανημάτων Lazer είναι η σταθερή μας επίδραση.

Διαφορετικά θα λέγαμε ότι μία σταθερή επίδραση χρησιμοποιείται για να ελέγξουμε τις διαφορές των τιμών της μεταξύ συγκεκριμένων επιπέδων. Αντίθετα μία τυχαία επίδραση χρησιμοποιείται για να καταλήξουμε σε συμπεράσματα σχετικά με την κατανομή των τιμών της, δηλαδή τη διακύμανση μεταξύ των τιμών της απόκρισης στα διάφορα επίπεδα.

Στην περίπτωση που έχουμε μία κατηγορική μεταβλητή της οποίας το δείγμα εξαντλεί τον πληθυσμό της, τότε η μεταβλητή αυτή έχει σταθερή επίδραση, ενώ όταν επιλέγουμε το δείγμα της από έναν ευρύτερο πληθυσμό, τότε η επίδραση της αντίστοιχης μεταβλητής θεωρείται τυχαία. Δηλαδή οι επιδράσεις μεταβλητών όπως το φύλο, το μορφωτικό επίπεδο (πρωτοβάθμια-δευτεροβάθμια-τριτοβάθμια εκπαίδευση), τα ηλικιακά γκρουπ και άλλες αντίστοιχες μεταβλητές, θεωρούνται σταθερές επιδράσεις. Ο λόγος είναι ότι μελετάμε όλες τις πιθανές υποκατηγορίες των παραπάνω κατηγορικών μεταβλητών ή διαφορετικά ολόκληρο τον πληθυσμό τους. Αντίθετα αν για παράδειγμα επιλέγουμε ένα δείγμα ασθενών για την μελέτη μιας θεραπείες ή επιλέγουμε ένα δείγμα από διαφορετικά είδη πτηνών, τότε μιλάμε για τυχαίες επιδράσεις.

Σε ένα μοντέλο τυχαίων επιδράσεων, οι εκτιμήσεις των τυχαίων επιδράσεων "συρρικνώνονται" σε σύγκριση με τις αντίστοιχες των σταθερών επιδράσεων. Δηλαδή αν υποθέσουμε ότι έχουμε μια κατηγορική μεταβλητή και η επίδραση της θεωρηθεί τυχαία, οι μέσες τιμές των εκτιμήσεων των τυχαίων επιδράσεων θα είναι πιο κοντά στην συνολική μέση τιμή από ότι αν είχε θεωρηθεί η επίδραση της ίδιας μεταβλητής ως σταθερή. Η ιδιότητα αυτή των τυχαίων επιδράσεων, μπορεί να μας φανεί πάρα πολύ χρήσιμη στην περίπτωση που έχουμε μικρό αριθμό παρατηρήσεων ανά υποομάδα. Γνωρίζουμε ότι σε τέτοιες περιπτώσεις μπορούμε να καταλήξουμε σε ακραίες εκτιμήσεις λόγω του μικρού αριθμού παρατηρήσεων. Ωστόσο θεωρώντας την επίδραση αυτής της μεταβλητής ως τυχαία, λύνουμε το παραπάνω πρόβλημα καθώς ειδικότερα για τις υποομάδες που έχουν μικρό αριθμό παρατηρήσεων, οι μέσες τιμές τους συρρικνώνονται σε μεγάλο βαθμό προς την συνολική μέση τιμή.

Τέλος η διασταυρώμενη δοκιμή είναι ένας τύπος κλινικής δοκιμής που χρησιμοποιείται για την σύγκριση δυο ή περισσότερων θεραπειών ή γενικότερα διαδικασιών στην καταπολέμηση μιας ασθένειας ή μιας χρόνιας πάθησης. Κάθε παρατηρητής που συμμετέχει στην δοκιμή, λαμβάνει και τις δυο θεραπείες καταγραφόντας τα αποτελέσματα. Έτσι έχουμε στην διάθεση μας δεδομένα με την ανταπόκριση που είχαν οι δυο θεραπείες στον ίδιο οργανισμό. Ωστόσο ένα συχνό πρόβλημα που αντιμετωπίζουμε είναι ότι κάποιοι ασθενείς δεν ολοκληρώνουν την διαδικασία, επομένως καταλήγουμε να μελετάμε ελλειπή δεδομένα. Σε αυτήν την περίπτωση η επιλογή του μοντέλου τυχαίων επιδράσεων πλεονεκτεί έναντι του μοντέλου σταθερών επιδράσεων. Ο λόγος είναι ότι το μοντέλο τυχαίων επιδράσεων αντλεί πληροφορίες τόσο μεταξύ των ασθενών όσο και εντός των ασθενών, σε αντίθεση με το μοντέλο σταθερών επιδράσεων που αντλεί μόνο εντός των ασθενών, αγνοώντας ουσιαστικά τις παρατηρήσεις με ελλειπή δεδομένα. Το κέρδος μας με την επιλογή του μοντέλου τυχαίων επιδράσεων είναι η μείωση του τυπικού σφάλματος του μοντέλου, το οποίο μεγαλώνει όσο περισσότερες είναι οι ελλείψεις δεδομένων, κάνοντας τότε επιτακτική την επιλογή του μοντέλου τυχαίων επιδράσεων.

1.5 Διαφορετικά Είδη Συγχώνευσης

Στην ενότητα αυτή θα καταγράψουμε τα τρία διαφορετικά είδη συγχωνεύσεων με την βοήθεια ενός παραδείγματος, τα οποία με την σειρά τους θα μας βοηθήσουν να καταλήξουμε σε μία ακόμα, ιδιαίτερη διαφορά μεταξύ ενός μοντέλου σταθερών επιδράσεων και ενός μικτού γραμμικού μοντέλου.

Παράδειγμα 1.5

Υποθέτουμε ότι έχουμε δεδομένα που αφορούν τον καθαρό χρόνο τρεξίματος σε λεπτά, για 50 συμμετέχοντες στον ετήσιο αγώνα 10 χιλιομέτρων που πραγματοποιείται στην Καλαμάτα. Κάθε δρομέας είναι από 50 χρονών έως 60, έχοντας συνολικά 6 συμμετοχές στον ετήσιο αυτό αγώνα.

Στην πρώτη περίπτωση, για την αναλύση μας θα χρησιμοποιήσουμε την προσέγγιση πλήρης συγχώνευσης (complete-pooling), η οποία αποτελεί και το πρώτο είδος συγχώνευσης. Αποτελεί μια προσέγγιση στατιστικής μοντελοποίησης, όπου δεδομένα από διαφορετικές ομάδες ή κατηγορίες συνδυάζονται μάζι για να εκτιμηθούν οι παραμέτροι του μοντέλου. Πιο συγκεκριμένα η προσέγγιση πλήρης συγχώνευσης, χρησιμοποιεί όλα τα δεδομένα σαν να προέρχονται από έναν ενιαίο ομοιογενή πληθυσμό, αγνοώντας δηλαδή την ύπαρξη των υποομάδων και τις ιδιαιτερότητες που προκύπτουν ως αποτέλεσμα αυτού.

Εύκολα γίνεται κατανοητό ότι το μοντέλο πλήρης συγχώνευσης, δεν έχει αξιοποιήσει σωστά τα δεδομένα μας και έχουμε χάσει σημαντικές πληροφορίες, έχοντας συγκεντρώσει όλες τις παρατηρήσεις μας σε έναν πληθυσμό. Πιο συγκεκριμένα τα δυο βασικά μειονεκτήματα που προκύπτουν είναι τα εξής:

- Έχουμε κάνει την υπόθεση ότι κάθε παρατήρηση είναι ανεξάρτητη από τις άλλες.
- Συγκεκριμένα για το παράδειγμα μας, έχουμε δημιουργήσει μία γραμμική σχέση του χρόνου τρεξίματος και της ηλικίας, η οποία είναι η ίδια για κάθε δρομέα.

Οι παρατηρήσεις μεταξύ διαφορετικών δρομέων μπορεί να είναι ανεξάρτητες, ωστόσο είμαστε βέβαιοι ότι οι παρατηρήσεις εντός ενός δρομέα σχετίζονται. Επίσης το μοντέλο παρέχοντας μας την ίδια σχέση για τον χρόνο τρεξίματος και την ηλικία για κάθε δρομέα, αγνόησε το γεγονός πως ο κάθε άνθρωπος είναι ξεχωριστός και προφανώς ότι η σχέση δεν είναι ακριβής για τους δρομείς μεμονωμένα. Ενδεχομένως τα μειονεκτήματα που καταγράψαμε βασίστηκαν αρκετά στο Παράδειγμα 1.5, ωστόσο αξίζει να σημειωθεί ότι τα μειονεκτήματα αυτά ισχύουν γενικότερα για κάθε μοντέλο που χρησιμοποιεί την προσέγγιση πλήρης συγχώνευσης, έχοντας αυτή την δομή δεδομένων.

Στην συνέχεια για την ανάλυση μας θα χρησιμοποιήσουμε την προσέγγιση μη συγχώνευσης (no-pooling), η οποία αποτελεί και το δεύτερο είδος συγχώνευσης. Αντίθετα με την προηγούμενη περίπτωση, η προσέγγιση μη συγχώνευσης εξετάζει κάθε έναν από τους παρατηρητές ή τα αντικείμενα μελέτης ξεχωριστά. Δηλαδή στο παράδειγμα μας εξετάζει τους 50 δρομείς ξεχωριστά, δημιουργώντας ένα ξεχωριστό μοντέλο για κάθε δρομέα. Μέσω της προσέγγισης αυτής, επιτρέπουμε σε κάθε δρομέα να έχει μοναδική σταθερά και συντελεστή ηλικίας.

Εκ πρώτης όψεως, η προσέγγιση αυτή δεν εμφανίζει τα μειονεκτήματα της πλήρης συγχώνευσης. Ωστόσο και εδώ εμφανίζονται δυο σημαντικά μειονεκτήματα.

- Εξετάζοντας τον κάθε παρατηρητή ξεχωριστά, γίνεται η υπόθεση ότι η μία ομάδα δεν παρέχει σχετικές-χρήσιμες πληροφορίες για την αλλή, δηλαδή το μοντέλο αγνοεί τυχόν πληροφορίες μεταξύ των ομάδων. Το πρόβλημα αυτό διογκώνεται στην περίπτωση που έχουμε πολύ μικρό αριθμό παρατηρήσεων ανά ομάδα, καθώς αγνοώντας τις πληροφορίες μεταξύ των ομάδων εύκολα μπορούμε να καταλήξουμε σε ακραίες εκτιμήσεις.

- Το δεύτερο μειονέκτημα έχει να κάνει με τα συμπέρασματα τα οποία μπορούμε να καταλήξουμε. Έχουμε την δυνατότητα δηλαδή, να προβούμε σε συμπεράσματα μόνο για τις ομάδες που διαθέτουμε στα δεδομένα μας, αδυνατώντας δηλαδή να γενικεύσουμε τα συμπεράσματα μας.

Η προσέγγιση αυτή μοιάζει να είναι καλύτερη από την αρχική, ωστόσο έχει ακόμα δυο σημαντικές αδυναμίες. Στην συνέχεια θα παρουσιάσουμε το τελευταίο είδος συγχώνευσης, το οποίο έρχεται να λύσει τα παραπάνω προβλήματα.

Ωστόσο προτού παρουσιάσουμε το τελευταίο είδος συγχώνευσης, θα κάνουμε μία σημαντική σημείωση που έχει παραλειφθεί έως τώρα. Τα τρία διαφορετικά είδη συγχωνεύσεων, είναι τρείς διαφορετικοί τρόποι προσέγγισης των ιεραρχικών δεδομένων. Τα ιεραρχικά δεδομένα, αναφέρονται σε ένα τύπο δεδομένων, όπου οι μεμονωμένες παρατηρήσεις οργανώνονται και ομαδοποιούνται σε ομάδες ή συστάδες. Διαφορετικά θα λέγαμε, ότι κάθε ομάδα περιέχει πολλαπλές υποομάδες και αντίστοιχα αυτές οι υποομάδες μπορούν να περιέχουν υποομάδες μικρότερης τάξης. Για παράδειγμα μπορούμε να συλλέγουμε δεδομένα σχετικά με τις επιδόσεις των μαθητών, με τους μαθητές να ομαδοποιούνται με βάση τα σχολεία στα οποία φοιτούν. Αν επιστρέψουμε στο παράδειγμα μας, τότε το ανώτερο στρώμα της ιεραρχίας είναι ο πληθυσμός όλων των δρομέων, ενώ στο επόμενο στρώμα βρίσκονται οι 50 δρομείς που επιλέχθηκαν. Τέλος το τελευταίο στρώμα περιέχει τις πολλαπλές παρατηρήσεις ανά δρομέα.

Η τελευταία προσέγγιση που θα παρουσιάσουμε για την ανάλυση μας, είναι η μερική συγχώνευση (partial-pooling), η οποία ουσιαστικά είναι μία ενδιάμεση προσέγγιση μεταξύ της πλήρης συγχώνευσης και της μη συγχώνευσης στο πλαίσιο της μοντελοποίησης ιεραρχικών δεδομένων. Η προσέγγιση αυτή αναγνωρίζει ότι κάθε ομάδα μπορεί να έχει μοναδικά χαρακτηριστικά, αλλά αναγνωρίζει επίσης ότι όλες οι ομάδες συνδέονται μεταξύ τους και συνεπώς μπορεί να παρέχουν πολύτιμες πληροφορίες η μία για την άλλη. Επομένως η προσέγγιση μερικής συγχώνευσης, λαμβάνει πληροφορίες τόσο εντός των ομάδων όσο και μεταξύ των ομάδων, αξιοποιώντας όλες τις δυνατές πληροφορίες που μας δίνονται μέσω των δεδομένων μας.

Καταλήγουμε με βάση όσα έχουμε αναφέρει να συμπεράνουμε ότι η προσέγγιση μερικής συγχώνευσης, είναι η καλύτερη δυνατή όταν μελετάμε ιεραρχικά δεδομένα. Η προσέγγιση αυτή, εφαρμόζεται από ένα μικτό γραμμικό μοντέλο. Αντίθετα ένα μοντέλο σταθερών επιδράσεων δεν είναι σε θέση να εφαρμόσει την προσέγγιση μερικής συγχώνευσης. Εξαιτίας των πλεονεκτημάτων της μερικής συγχώνευσης, καταλήγουμε ότι ένα μικτό γραμμικό μοντέλο προτιμάται έναντι ενός μοντέλου σταθερών επιδράσεων κατά την μελέτη ιεραρχικών δεδομένων.

Τέλος είμαστε σε θέση να καταγράψουμε, ότι η προσέγγιση μερικής συγχώνευσης οδηγεί στο φαινόμενο της “συρρίκνωσης” των τυχαίων επιδράσεων που καταγράψαμε στο Κεφάλαιο 1.4.

1.6 Μοντέλο Μοτίβου Συνδιακυμάνσεων

Το μοντέλο τυχαίων επιδράσεων ήταν η πρώτη κατηγορία μικτού γραμμικού μοντέλου που περιγράψαμε, χωρίς ωστόσο να είναι η μοναδική. Στην συγκεκριμένη ενότητα θα γίνει αναφορά στην δεύτερη κατηγορία ενός μικτού γραμμικού μοντέλου, του μοντέλου μοτίβου συνδιακυμάνσεων. Το μοντέλο μοτίβου συνδιακύμανσεων χρησιμοποιείται μόνο στην περίπτωση που έχουμε δεδομένα επαναλαμβανόμενων μετρήσεων, δηλαδή έχουμε επαναλαμβανόμενες μετρήσεις για τον ίδιο παρατηρητή ή για το ίδιο αντικείμενο μελέτης.

Για να γίνει κατανοητό το μοντέλο, θα χρησιμοποιήσουμε το Παράδειγμα 1.2, πραγματοποιώντας ωστόσο μια μικρή τροποποίηση. Η τροποποίηση αυτή θα πραγματοποιηθεί για να καταλήξουμε σε δεδομένα επαναλαμβανόμενων μετρήσεων, δηλαδή θα υποθέσουμε ότι έχουμε συνολικά τρείς μετρήσεις για κάθε παρατηρητή μετά την χειρουργική επέμβαση, τον πρώτο, τον δεύτερο και τον τρίτο μήνα. Εύκολα αντιλαμβανόμαστε ότι οι επαναλαμβανόμενες μετρήσεις μετά την χειρουργική επέμβαση για τον ίδιο παρατηρητή, δεν είναι ανεξάρτητες μεταξύ τους. Ερχόμαστε δηλαδή αντιμέτωποι με μια νέα κατάσταση και ένα νέο πρόβλημα.

Το πρόβλημα αυτό έρχεται να το λύσει το μοντέλο μοτίβου συνδιακύμανσεων, καθώς είναι σε θέση να μοντελοποιεί την συσχέτιση μεταξύ των επαναλαμβανόμενων παρατηρήσεων για τον ίδιο παρατηρητή. Η συσχέτιση αυτή υποδηλώνεται μέσω των υπολοίπων, με τους τρείς πιθανούς τρόπους συσχέτισης να παρουσιάζονται στην συνέχεια:

- $corr(e_{ij}, e_{ij'}) = \rho$.
- $corr(e_{ij}, e_{ij'}) = \rho^{|j'-j|}$.
- $corr(e_{ij}, e_{ij'}) = \rho_{j,j'}$.
- $|\rho| < 1$.

Το corr συμβολίζει την συσχέτιση, ενώ το e_{ij} τα υπόλοιπα. Το i αντιπροσωπεύει τον παρατηρητή, το j την παρατήρηση που μελετάμε και το ρ είναι ενας σταθερός αριθμός με απόλυτη τιμή μικρότερη του 1. Στην συνέχεια θα αναλυθούν οι τρείς παραπάνω σχέσεις.

- Στην πρώτη περίπτωση έχουμε την πιο απλή προσέγγιση, καθώς θεωρούμε ότι η συσχέτιση για όλα τα ζεύγη μετρήσεων για τον ίδιο παρατηρητή, είναι σταθερή και ίση με ρ.
- Στην δεύτερη περίπτωση η συσχέτιση μεταξύ των ζεύγων μετρήσεων για τον ίδιο παρατηρητή, μειώνεται καθώς απομακρύνονται χρονικά.
- Η τρίτη περίπτωση είναι και η πιο σύνθετη, καθώς έχουμε διαφορετική συσχέτιση για κάθε ζεύγος μετρήσεων.

Εώς τώρα δεν έχουμε κάνει καμία αναφορά στις τυχαίες επιδράσεις, καθώς στο μοντέλο μοτίβου συνδιακύμανσεων δεν θεωρούμε την επίδραση καμίας μεταβλητής ως τυχαία. Το μοντέλο μοτίβου συνδιακυμάνσεων, είναι σε θέση να μοντελοποιεί τους τρείς προαναφερθείς τρόπους συσχέτισης μεταξύ των επαναλαμβανόμενων μετρήσεων για τον ίδιο παρατηρητή και θεωρεί τις επιδράσεις των επιθυμητών μεταβλητών ως σταθερές επιδράσεις. Επομένως καταλήγουμε σε ένα μοντέλο χωρίς καμία τυχαία επίδραση.

1.7 Μοντέλο Τυχαίων Συντελεστών

Η τρίτη και η τελευταία κατηγορία μικτών γραμμικών μοντέλων που θα παρουσιάσουμε, αντιστοιχεί και πάλι στην περίπτωση που έχουμε δεδομένα επαναλαμβανόμενων μετρήσεων, με το μοντέλο να καλείται ως μοντέλο τυχαίων συντελεστών. Αποτελεί μια εναλλακτική προσέγγιση και επιλογή από το μοντέλο μοτίβου συνδιακυμάνσεων, για την μοντελοποίηση τέτοιου είδους δεδομένων.

Γενικότερα, θα λέγαμε ότι ένα μοντέλο τυχαίων συντελεστών προσπαθεί να εξηγήσει αριθμητικά την σχέση μεταξύ της εξαρτημένης μεταβλητής και μίας συνδιακύμανσης. Μία συνδιακύμανση τώρα, γνωρίζουμε ότι είναι μία μεταβλητή η οποία εξηγεί μέρος της μεταβλητότητας της εξαρτημένης μεταβλητής. Στην περίπτωση μας για την επεξήγηση του μοντέλου θα χρησιμοποιήσουμε τον χρόνο ως την μεταβλητή συνδιακύμανσης.

Πιο συγκεκριμένα ένα μοντέλο τυχαίων συντελεστών, επιτρέπει μοναδικές σχέσεις μεταξύ της εξαρτημένης μεταβλητής και του χρόνου για κάθε υποκείμενο ή για κάθε αντικείμενο μελέτης. Για να γίνει αυτό, η επίδραση του χρόνου περιλαμβάνεται στο μοντέλο και μπορεί να μετρηθεί σε διαφορετικές μονάδες, όπως λεπτά, ώρες, ημέρες, εβδομάδες κ.λ.π. Εάν παραλείψουμε τυχόν άλλες επιδράσεις, τότε το μοντέλο μας σε αρχική μορφή μπορεί να γραφεί ως εξής:

- Εξαρτημ. Μεταβλητή = μ + β∗χρόνος + ...

Παρατηρούμε ότι έχουμε δημιουργήσει μια γραμμική σχέση μεταξύ της εξαρτημένης μεταβλητής και του χρόνου, με κλίση β. Στην συνέχεια θα γίνουμε πιο συγκεκριμένοι, χρησιμοποιώντας το Παράδειγμα 1.2 με τις τροποποιήσεις ωστόσο που έγιναν στην Ενότητα 1.6. Υποθέτουμε για την απλούστευση του μοντέλου μας, ότι μόνο η επίδραση των μηχανημάτων Lazer ήταν στατιστικά σημαντική. Επομένως προσθέτοντας και την επίδραση του χρόνου στο μοντέλο μας, τότε μπορεί να εκφραστεί ως εξής:

- $(B\alpha\theta\mu.M\upsilon\omega\pi\iota\alpha\varsigma)_{ij} = \mu + l_k + \beta * time_{ij} + e_{ij}$, όπου το l_k αντιπροσωπεύει την επίδραση των μηχανημάτων Lazer, A και B, το $time_{ij}$ αντιπροσωπεύει τον χρόνο παρατήρησης j για τον παρατηρητή i σε μήνες και τέλος το β είναι μία σταθερά που αντιπροσωπεύει την μέση μεταβολή των βαθμών μυωπίας ανά μήνα. Έχουμε δηλαδή, δημιουργήσει μία γραμμική σχέση μεταξύ των βαθμών μυωπίας και του χρόνου, με κλίση β.

Ορίζοντας κατά αυτόν τον τρόπο το μοντέλο, έχουμε κάνει την υπόθεση ότι η σχέση των βαθμών μυωπίας και του χρόνου είναι η ίδια για παρατηρητές που έχουν χρησιμοποιήσει το ίδιο μηχάνημα Lazer. Ωστόσο στην αρχή επισημάνθηκε ότι δημιουργούμε μια ξεχωριστή γραμμική σχέση μεταξύ της εξαρτημένης μεταβλητής (βαθμοί μυωπίας) και του χρόνου για κάθε παρατηρητή. Για να επιτευχθεί αυτό, εισάγουμε δυο νέες μεταβλητές στο μοντέλο μας. Η πρώτη είναι η επίδραση των παρατηρητών, ενώ η δεύτερη είναι η επίδραση του χρόνου σε κάθε παρατηρητή. Το μοντέλο μας γράφεται ως εξής:

- $(B\alpha\theta\mu.M\upsilon\omega\pi\iota\alpha\varsigma)_{ij} = \mu + l_k + \beta * time_{ij} + p_i + (p\beta)_i * time_{ij} + e_{ij}$,

όπου το p_i είναι η επίδραση των παρατηρητών και αντιπροσωπεύει την διαφορά από την μέση σταθερά για τον i παρατηρητή, ενώ η $(p\beta)_i$ είναι η επίδραση του χρόνου σε κάθε παρατηρητή και αντιπροσωπεύει την διαφορά από την μέση κλίση για τον i παρατηρητή. Δηλαδή έχουμε καταφέρει να δημιουργήσουμε μια ξεχωριστή γραμμική σχέση μεταξύ των βαθμών μυωπίας και του χρόνου για κάθε παρατηρητή, το οποίο ήταν και το ζητούμενο μας.

Έχουμε ορίσει τις μεταβλητές που περιγράφουν το μοντέλο μας, ωστόσο δεν έχουμε κάνει ακόμα καμία αναφορά για το αν οι νέες μεταβλητές που έχουν προστεθεί, έχουν σταθερές ή τυχαίες επιδράσεις. Φαντάζει λογικό με βάση όσα έχουμε αναφέρει στην Ενότητα 1.4, να θεωρήσουμε τις επιδράσεις των δυο αυτών νέων μεταβλητών ως τυχαίες επιδράσεις.

Στην συνέχεια θα καταγράψουμε τις δυο σημαντικές διαφορές μεταξύ ενός μοντέλου τυχαίων συντελεστών και ενός μοντέλου τυχαίων επιδράσεων. Στο μοντέλο τυχαίων συντελεστών, έχουμε προσαρμόσει έναν όρο συνδιακυμάνσεων που είναι ο χρόνος, ως τυχαία επίδραση. Αντίθετα στο μοντέλο τυχαίων επιδράσεων δεν γίνεται καμία τέτοια προσαρμογή. Επίσης οι δυο νέες μεταβλητές που έχουν τυχαίες επιδράσεις, αποτελούν ουσιαστικά την προσωπική κλίση και σταθερά της γραμμής παλινδρόμησης κάθε παρατηρητή. Επομένως είναι λογικό να υποθέσουμε ότι υπάρχει μία συσχέτιση μεταξύ τους. Αντίθετα σε ένα μοντέλο τυχαίων επιδράσεων, οι τυχαίες επιδράσεις είναι ασυσχέτιστες μεταξύ τους.

Εξαιτίας τώρα της συσχέτισης των δυο τυχαίων επιδράσεων, για τον ορισμό τους θα χρησιμοποιήσουμε την διμεταβλητή κανονική κατανομή. Υποθέτουμε ότι $\sigma_p^2, \sigma_{p\beta}^2$ η διακύμανση της επίδρασης του παρατηρητή και της επίδρασης του χρόνου σε κάθε παρατηρητή αντίστοιχα και υποθέτουμε επίσης ότι η συνδιακύμανση ισούται με $\sigma_{p,p\beta}$, επομένως καταλήγουμε:

$$\begin{pmatrix} p_i \\ (p\beta)_i \end{pmatrix} \sim N(\mathbf{0}, \mathbf{G}),$$

όπου:

$$\mathbf{G} = \begin{pmatrix} \sigma_p^2 & \sigma_{p,p\beta} \\ \sigma_{p,p\beta} & \sigma_{p\beta}^2 \end{pmatrix}.$$

Αξίζει να σημειωθεί ότι στην περίπτωση ενός μοντέλου τυχαίων συντελεστών, συνηθίζεται όταν θεωρούμε την επίδραση μιας μεταβλητής ως τυχαία, να αναφερόμαστε σε αυτήν ως τυχαίο συντελεστή και όχι ως τυχαία επίδραση.

Συμπερασματικά θα λέγαμε ότι σε ένα μοντέλο τυχαίων συντελεστών, κάθε παρατηρητής ή κάθε αντικείμενο μελέτης επιτρέπεται να έχει μοναδική σταθερά και μοναδική κλίση, που περιγράφουν την σχέση μεταξύ της εξαρτημένης μεταβλητής και μίας μεταβλητής συνδιακύμανσης. Εναλλακτικά, το μοντέλο μπορεί να ερμηνευθεί ως ένα μοντέλο που επιτρέπει στη σταθερά και την κλίση της γραμμής παλινδρόμησης να διαφέρουν τυχαία μεταξύ των διαφόρων ομάδων ή των υποκειμένων της μελέτης.

1.8 Επίλογος

Το μοντέλο τυχαίων επιδράσεων, το μοντέλο μοτίβου συνδιακυμάνσεων και το μοντέλο τυχαίων συντελεστών ήταν οι τρείς κατηγορίες μικτών γραμμικών μοντέλων που παρουσιάστηκαν. Κάθε ένα έχει διαφορετική προσέγγιση και κάθε ένα έχει την δική του χρησιμότητα και το δικό του σημείο εφαρμογής κατά την μελέτη ενός μικτού γραμμικού μοντέλου.

Στο Κεφάλαιο 3, πριν χρησιμοποιήσουμε το στατιστικό πακέτο της R για την στατιστική επίλυση ενός μικτού γραμμικού μοντέλου, θα επανέλθουμε στην θεωρία τους καταγράφοντας κάποιες πιο ιδιαίτερες-συγκεκριμένες παρατηρήσεις, τις οποίες αμέσως μετά θα εφαρμόσουμε και θα παρατηρήσουμε με την βοήθεια του στατιστικού πακέτου της R.

Τέλος αξίζει να σημειωθεί ότι στην περίπτωση που έχουμε δεδομένα επαναλαμβανόμενων μετρήσεων, υπάρχει η δυνατότητα να χρησιμοποιηθεί και το μοντέλο τυχαίων επιδράσεων.

Κεφάλαιο 2

Κανονικά Μικτά Γραμμικά Μοντέλα

2.1 Εισαγωγή

Στο Κεφάλαιο 1 έγινε μια εισαγωγή σε θεωρητικό επίπεδο στα μικτά γραμμικά μοντέλα, με τον αναγνώστη να αποκτά μια πρώτη εικόνα γι'αυτά. Στο Κεφάλαιο 2 θα παρουσιάσουμε την μαθηματική τους πλευρά.

Καθώς το μοντέλο σταθερών επιδράσεων ήταν κύριο αντικείμενο μελέτης στο προηγούμενο κεφάλαιο, αναφορικά με τις διαφορές που εντοπίζονται με ένα μικτό γραμμικό μοντέλο, θα παρουσιάσουμε εν συντομία τον τύπο του και κάποια κύρια χαρακτηριστικά του. Έτσι εκτός από τις διαφορές σε θεωρητικό επίπεδο, θα είμαστε σε θέση να παρατηρήσουμε και αυτές που καταγράφονται καθαρά σε μαθηματικό.

Στην συνέχεια όπως είναι αναμενόμενο θα ασχοληθούμε με τα μικτά γραμμικά μοντέλα, με τα αντικείμενα μελέτης να είναι αντίστοιχα με οποιοδήποτε άλλο στατιστικό μοντέλο. Δηλαδή θα ξεκινήσουμε την αναφορά μας παραθέτοντας τον γενικό τύπο του μοντέλου και υπολογίζοντας τον πίνακα διακύμανσης του. Επίσης θα εκτιμήσουμε τους συντελεστές του, δηλαδή τις σταθερές και τυχαίες επιδράσεις και θα αναφερθούμε στην εκτίμηση των παραμέτρων διακύμανσης τους. Οι παράμετροι διακύμανσης σε ένα μικτό γραμμικό μοντέλο, περιλαμβάνουν την διακύμανση των τυχαίων επιδράσεων, την διακύμανση των υπολοίπων και την συνδιακύμανση των παραμέτρων των τυχαίων επιδράσεων εφόσον υπάρχει. Τέλος θα κατασκευάσουμε το διάστημα εμπιστοσύνης για τους συντελεστές των σταθερών επιδράσεων και θα παρουσιάσουμε τον έλεγχο του λόγου των πιθανοφανειών, καθώς και τα κριτήρια AIC, BIC.

2.2 Μοντέλο Σταθερών Επιδράσεων

Ο γενικός τύπος του μοντέλου σταθερών επιδράσεων για p συντελεστές και n παρατηρήσεις είναι ο εξής:

$$y_i = \mu + \alpha_1 x_{i1} + a_2 x_{i2} + ... + a_p x_{ip} + e_i, i = 1, ..., n. \tag{2.2}$$

- μ είναι η σταθερά του μοντέλου.
- $a_1, a_2, ..., a_p$ οι p συντελεστές (ή παράμετροι) των σταθερών επιδράσεων.
- $e_1, e_2, ..., e_n$ αντιπροσωπεύουν τα υπόλοιπα.

Η σχέση (2.2) γράφεται ως εξής για κάθε παρατήρηση:

$$\begin{aligned} y_1 &= \mu + \alpha_1 x_{11} + a_2 x_{12} + ... + a_p x_{1p} + e_1 \\ y_2 &= \mu + \alpha_1 x_{21} + a_2 x_{22} + ... + a_p x_{2p} + e_2 \end{aligned}$$

$$\dots$$
$$\dots$$
$$\dots$$
$$y_n = \mu + \alpha_1 x_{n1} + a_2 x_{n2} + \dots + a_p x_{np} + e_n.$$

Η σχέση (2.2) τώρα, με την χρήση πινάκων-διανυσμάτων γράφεται ως εξής:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{e}. \tag{2.3}$$

- $\mathbf{y} = (y_1, y_2, ..., y_n)'$ είναι το $(n\text{x}1)$ διάνυσμα τιμών της εξαρτημένης μεταβλητής.
- $\mathbf{X}$ ένας $(n\text{x}k)$ πίνακας για τις ανεξάρτητες μεταβλητές με $k = p + 1$, όπου p ο αριθμός των συντελεστών των σταθερών επιδράσεων.
- $\boldsymbol{\alpha} = (\mu, a_1, a_2, ..., a_p)'$ είναι το $\Big((p+1)\text{x}1\Big)$ διάνυσμα των συντελεστών των σταθερών επιδράσεων.
- $\mathbf{e} = (e_1, e_2, ..., e_n)'$ είναι το $(n\text{x}1)$ διάνυσμα των υπολοίπων.

Ο πίνακας $\mathbf{X}$ καλείται πίνακας σχεδιασμού και έχει την εξής μορφή:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

Η πρώτη στήλη του πίνακα σχεδιασμού συμπληρώνεται πάντα από τον αριθμό 1 λόγω της σταθεράς μ που υπάρχει σε κάθε μοντέλο. Οι επόμενες στήλες του περιέχουν τις τιμές των ανεξάρτητων μεταβλητών.

Οι δυο τρόποι γραφής ενός μοντέλου σταθερών επιδράσεων που παρουσιάστηκαν, είναι ισοδύναμοι. Εύκολα μπορούμε να καταλήξουμε από την διανυσματική μορφή στην αναλυτική, πραγματοποιώντας τις πράξεις μεταξύ των πινάκων-διανυσμάτων. Στην συνέχεια θα γίνουμε λίγο πιο συγκεκριμένοι για τους συντελεστές των σταθερών επιδράσεων και τις ανεξάρτητες μεταβλητές.

Οι συντελεστές των σταθερών επιδράσεων είναι οι άγνωστες παράμετροι του μοντέλου. Λαμβάνουν σταθερές τιμές εξού και το όνομα τους και καλούμαστε να τις υπολογίσουμε.

Αντίθετα οι τιμές $x_{i1}, x_{i2}, ..., x_{ip}$ είναι γνωστές με βάση τα δεδομένα που έχουμε. Καθώς μια σταθερή επίδραση μπορεί να είναι είτε κατηγορική είτε ποσοτική έχουμε δυο διαφορετικές περιπτώσεις. Στην γενική περίπτωση που εξετάζουμε μια κατηγορική σταθερή επίδραση με x επιμέρους υποκατηγορίες, έχουμε $x - 1$ εικονικές μεταβλητές στο μοντέλο μας. Οι ανεξάρτητες μεταβλητές στην περίπτωση αυτή, λαμβάνουν είτε την τιμή 1 είτε την τιμή 0. Η τιμή 1 εισάγεται στην περίπτωση που η παρατήρηση εντάσσεται στην επιμέρους υποκατηγορία, ενώ σε αντίθετη περίπτωση λαμβάνει την τιμή 0. Στην περίπτωση που εξετάζουμε μια ποσοτική σταθερή επίδραση, χρειαζόμαστε μόνο έναν συντελεστή στον γενικό τύπο του μοντέλου. Η τιμή της ανεξάρτητης μεταβλητής τότε, λαμβάνει αυτούσια την τιμή της ποσοτικής μεταβλητής που αντιστοιχεί στην παρατήρηση αυτή.

Παράδειγμα 2.2

Τα δεδομένα του προβλήματος έχουν προσομοιωθεί (επομένως είναι φανταστικά) και αντιστοιχούν σε ένα πείραμα στο οποίο μελετάται η αποτελεσματικότητα δυο διαφορετικών ειδών δίαιτας A, B σε διάστημα ενός μήνα. Έχουμε λοιπόν δέκα άντρες συμμετέχοντες, πέντε ακολουθούν πιστά την δίαιτα A και πέντε την δίαιτα B. Τα δεδομένα που έχουμε για τον κάθε συμμετέχοντα είναι τα κιλά πριν ακριβώς την έναρξη της δίαιτας και τα κιλά μετά το πέρας αυτού του μήνα. Προφανώς για να έχουμε πιο ασφαλή συμπεράσματα, θα μπορούσαμε να είχαμε δεδομένα όπως η ηλικία του κάθε συμμετέχοντος, εάν έχει κάποια ιατρικό θέμα που επηρεάζει την απώλεια κιλών κ.λ.π, αλλά δεν είναι αυτός ο σκοπός μας. Ο Πίνακας 2.1 που περιέχει τα δεδομένα μας είναι ο εξής:

Παρατηρητής	**Δίαιτα**	**Kg πριν την δίαιτα**	**Kg μετά την δίαιτα**
1	A	103.2	101.1
2	A	95.7	93.0
3	B	93.1	90.5
4	A	99.0	96.2
5	B	88.9	87.1
6	B	81.2	79.8
7	B	100.5	96.7
8	A	107.1	104.3
9	A	85.1	82.2
10	B	77.7	74.6

Πίνακας 2.1: Δεδομένα του Παραδείγματος 2.2.

Θεωρούμε την επίδραση της δίαιτας και την επίδραση των κιλών πριν την δίαιτα ως τις σταθερές επιδράσεις του μοντέλου μας. Επιθυμούμε η ομάδα αναφοράς στο παράδειγμα μας να είναι η δίαιτα A, επομένως εισάγουμε την εξής εικονική μεταβλητή:

$$x_{i1} = \begin{cases} 1, & \text{δίαιτα B} \\ 0, & \text{δίαιτα A} \end{cases}, \quad i = 1, 2, \ldots, 10.$$

Επομένως το μοντέλο μπορεί να εκφραστεί ως εξής:

$$y_i = \mu + \alpha_1 x_{i1} + \alpha_2 x_{i2} + e_i, i = 1, ..., 10. \tag{2.4}$$

Πιο συγκεκριμένα μπορούμε να πούμε ότι ο συντελεστής a_1 αντιστοιχεί στην δίαιτα B και ο a_2 αντιστοιχεί στα κιλά πριν την δίαιτα. Η σχέση (2.4) για τους παρατηρητές 1, 3 γράφεται ως εξής:

$$\begin{aligned} y_1 &= \mu + \alpha_1 x_{11} + \alpha_2 x_{12} + e_1 = \mu + 0\alpha_1 + 103.2\alpha_2 + e_1 \\ &= \mu + 103.2\alpha_2 + e_1. \end{aligned}$$

$$y_3 = \mu + \alpha_1 x_{31} + \alpha_2 x_{32} + e_3 = \mu + 1\alpha_1 + 93.1\alpha_2 + e_3$$
$$= \mu + \alpha_1 + 93.1\alpha_2 + e_3.$$

Στην συνέχεια θα γίνουμε αναλυτικότεροι για την κατανομή που ακολουθούν τα υπόλοιπα.

- $\mathbf{e} \sim N(0, \sigma^2 I)$ ή $e_i \sim N(0, \sigma^2)$.
- $E(\mathbf{e}) = \mathbf{0}$ ή $E(e_i) = 0$.
- $Var(\mathbf{e}) = \sigma^2 I$ ή $Var(e_i) = \sigma^2$.

Τα $\mathbf{X\alpha}$, $\mathbf{e}$ είναι ασυσχέτιστα, επομένως το μέσο διάνυσμα και η διακύμανση του $\mathbf{y}$ είναι τα εξής:

- $E(\mathbf{y}) = E(\mathbf{X\alpha} + \mathbf{e}) = E(\mathbf{X\alpha}) + E(\mathbf{e}) = \mathbf{X\alpha}$.
- $Var(\mathbf{y}) = Var(\mathbf{X\alpha} + \mathbf{e}) = Var(\mathbf{X\alpha}) + Var(\mathbf{e}) = \sigma^2 I$.

Συμπερασματικά μπορούμε να πούμε ότι το διάνυσμα των υπολοίπων $\mathbf{e}$, ακολουθεί την κανονική κατανομή με μέσο διάνυσμα $\mathbf{0}$ και πίνακα διακύμανσης $\sigma^2\mathbf{I}$. Το μέσο διάνυσμα της εξαρτημένης μεταβλητής $\mathbf{y}$ ισούται με $\mathbf{X\alpha}$ και ο πίνακας διακύμανσης του με $\sigma^2\mathbf{I}$.

2.3 Μικτά Γραμμικά Μοντέλα

Στην προηγούμενη ενότητα έγινε αναφορά στον μαθηματικό ορισμό του μοντέλου σταθερών επιδράσεων. Στην ενότητα αυτή θα επικεντρωθούμε στον μαθηματικό ορισμό ενός μικτού γραμμικού μοντέλου, με την κυριότερη διαφορά να αποτελεί η προσθήκη των τυχαίων επιδράσεων. Οι συντελεστές τους καθώς ασχολούμαστε με κανονικά μικτά γραμμικά μοντέλα, ακολουθούν την κανονική κατανομή. Ο γενικός τύπος του μοντέλου για p, q συντελεστές σταθερών και τυχαίων επιδράσεων αντίστοιχα και για n παρατηρήσεις είναι ο εξής:

$$y_i = \mu + \alpha_1 x_{i1} + a_2 x_{i2} + ... + a_p x_{ip} + \beta_1 z_{i1} + \beta_2 z_{i2} + ... + \beta_q z_{iq} + e_i, i = 1, ..., n. \tag{2.5}$$

- μ είναι η σταθερά του μοντέλου.
- $\alpha_1, \alpha_2, ..., \alpha_p$ οι p συντελεστές των σταθερών επιδράσεων.
- $\beta_1, \beta_2, ..., \beta_q$ οι q συντελεστές των τυχαίων επιδράσεων.
- $e_1, e_2, ..., e_n$ αντιπροσωπεύουν τα υπόλοιπα.

Ο γενικός τύπος που παραθέσαμε γράφεται ως εξής για κάθε παρατήρηση:

$$y_1 = \mu + \alpha_1 x_{11} + a_2 x_{12} + ... + a_p x_{1p} + \beta_1 z_{11} + \beta_2 z_{12} + ... + \beta_q z_{1q} + e_1$$
$$y_2 = \mu + \alpha_1 x_{21} + a_2 x_{22} + ... + a_p x_{2p} + \beta_1 z_{21} + \beta_2 z_{22} + ... + \beta_q z_{2q} + e_2$$
$$...$$
$$...$$
$$...$$
$$y_n = \mu + \alpha_1 x_{n1} + a_2 x_{n2} + ... + a_p x_{np} + \beta_1 z_{n1} + \beta_2 z_{n2} + ... + \beta_q z_{nq} + e_n.$$

Η σχέση (2.5) με την χρήση πινάκων-διανυσμάτων γράφεται ως εξής:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} + \mathbf{e}.$$

- $\mathbf{y} = (y_1, y_2, ..., y_n)'$ είναι το (nx1) διάνυσμα τιμών της εξαρτημένης μεταβλητής.
- $\mathbf{Z}$ ένας (nxq) πίνακας, όπου q ο αριθμός των συντελεστών των τυχαίων επιδράσεων.
- $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_q)'$ είναι το (qx1) διάνυσμα των συντελεστών (ή παραμέτρων) των τυχαίων επιδράσεων.

Για τα $\mathbf{X}, \boldsymbol{\alpha}$ δεν θα γίνει περαιτέρω αναφορά καθώς ισχύουν τα ίδια που καταγράφηκαν στην προηγούμενη ενότητα.

Οι συντελεστές των τυχαίων επιδράσεων είναι και αυτοί άγνωστοι και καλούμαστε να τους υπολογίσουμε. Για το διάνυσμα των συντελεστών των τυχαίων επιδράσεων ισχύει ότι:

- $\boldsymbol{\beta} \sim N(0, \mathbf{G})$, όπου $Var(\boldsymbol{\beta}) = \mathbf{G}$,

ενώ για ένα συντελεστή μιας τυχαίας επίδρασης έχουμε:

- $\beta_i \sim N(0, \sigma^2_\beta)$.

Επίσης για το διάνυσμα των υπολοίπων έχουμε ότι:

- $\mathbf{e} \sim N(0, \mathbf{R})$, όπου $Var(\mathbf{e}) = \mathbf{R}$.

Ο πίνακας $\mathbf{Z}$ ονομάζεται και αυτός πίνακας σχεδιασμού και έχει διάσταση (nxq). Καθορίζει τις τιμές των τυχαίων επιδράσεων και συμπληρώνεται με το ίδιο σκεπτικό με τον πίνακα σχεδιασμού $\mathbf{X}$ που εξετάσαμε στο Κεφάλαιο 2.2. Η μορφή του είναι η εξής:

$$\mathbf{Z} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1q} \\ z_{21} & z_{22} & \cdots & z_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nq} \end{pmatrix}.$$

Η διαφορά του στην μορφή με τον πίνακα σχεδιασμού $\mathbf{X}$, εντοπίζεται στην απουσία της πρώτης στήλης που υπήρχε στον πίνακα σχεδιασμού $\mathbf{X}$, λόγω της σταθεράς μ.

Γνωρίζοντας πλέον πως συμπληρώνονται οι δυο πίνακες σχεδιασμού $\mathbf{X}$, $\mathbf{Z}$ για να μπορούμε να γράψουμε ολοκληρωμένα τον τύπο του μοντέλου μένει να υπολογίσουμε τους συντελεστές των σταθερών-τυχαίων επιδράσεων, με τον τρόπο υπολογισμού τους να παρουσιάζεται στις επόμενες ενότητες.

Παράδειγμα 2.3

Τα δεδομένα του προβλήματος έχουν προσομοιωθεί (επομένως είναι φανταστικά) και αντιστοιχούν σε ένα πείραμα στο οποίο μελετάται η αποτελεσματικότητα δυο διαφορετικών μηχανών στην παραγωγική διαδικασία του τομέα της βιομηχανίας. Έχουμε λοιπόν τρείς διαφορετικούς εργάτες, με τον κάθε ένα να χρησιμοποιεί και τις δυο μηχανές καταγράφοντας στο τέλος την βαθμολογία. Τα δεδομένα βρίσκονται στον Πίνακα 2.2.

Παρατήρηση	**Εργάτης**	**Μηχανή**	**Βαθμολογία**
1	1	A	52.0
2	1	A	52.7
3	1	B	60.2
4	1	B	58.4
5	2	A	53.1
6	2	A	54.6
7	2	B	59.0
8	3	A	57.3
9	3	B	64.2

Πίνακας 2.2: Δεδομένα του Παραδείγματος 2.3.

Η επίδραση των μηχανών αποτελεί την σταθερή μας επίδραση, ενώ η επίδραση των εργατών αποτελεί την τυχαία. Επιθυμούμε η ομάδα αναφοράς στο παράδειγμα μας να είναι η μηχανή A, επομένως εισάγουμε την εξής εικονική μεταβλητή:

$$x_{i1} = \begin{cases} 1, & \text{μηχανή B} \\ 0, & \text{μηχανή A} \end{cases}, \quad i = 1, 2, \ldots, 9.$$

Επίσης στο μοντέλο μας θα έχουμε τρείς συντελεστές τυχαίων επιδράσεων β_1, β_2, β_3 που αντιπροσωπεύουν τους τρείς εργάτες κατά σειρά. Επομένως το μοντέλο μπορεί να εκφραστεί ως εξής:

$$y_i = \mu + \alpha_1 x_{i1} + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + e_i, i = 1, ..., 9, \tag{2.6}$$

ή

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} + \mathbf{e}. \tag{2.7}$$

Θα αγνοήσουμε τις σταθερές επιδράσεις του μοντέλου και θα επικεντρωθούμε στις τυχαίες. Ο πίνακας σχεδιασμού $\mathbf{Z}$ συμπληρώνεται ως εξής:

$$\mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Η πρώτη στήλη αντιστοιχεί στον εργάτη$_1$, η δεύτερη στον εργάτη$_2$ και η τελευταία στον εργάτη$_3$, ενώ κάθε γραμμή αντιστοιχεί σε μία παρατήρηση.

Αν προσθέσουμε στο μοντέλο μας την επίδραση της αλληλεπίδρασης εργάτης-μηχανή και την θεωρήσουμε τυχαία επίδραση, τότε θα έχουμε έξι νέες παραμέτρους, με τον πίνακα $\mathbf{Z}$ να γράφεται ως εξής:

$$\mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Η παράμετρος β_4 αντιστοιχεί στην αλληλεπίδραση εργάτης$_1$-μηχανή$_1$, η β_5 στην αλληλεπίδραση εργάτης$_1$-μηχανή$_2$, η β_6 στην αλληλεπίδραση εργάτης$_2$-μηχανή$_1$ κ.ο.κ.

Στα μικτά γραμμικά μοντέλα τα πράγματα δεν είναι τόσο απλά όπως στο μοντέλο σταθερών επιδράσεων για τον υπολογισμό της διακύμανσης του μοντέλου. Λαμβάνοντας υπόψη την γενική σχέση (2.7) του μοντέλου, θα καταλήξουμε σε πίνακα διακύμανσης. Γνωρίζοντας ότι οι σταθερές, τυχαίες επιδράσεις και τα υπόλοιπα είναι ασυσχέτιστα μεταξύ τους και αν θεωρήσουμε ότι $V = Var(\mathbf{y})$, τότε:

$$\mathbf{V} = Var(\mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\beta} + \mathbf{e}) = Var(\mathbf{X}\boldsymbol{\alpha}) + Var(\mathbf{Z}\boldsymbol{\beta}) + Var(\mathbf{e}).$$

Τα στοιχεία των $\boldsymbol{\alpha}$, $\mathbf{X}$ είναι σταθεροί αριθμοί, επομένως $Var(\mathbf{X}\boldsymbol{\alpha}) = 0$. Επίσης ο πίνακας $\mathbf{Z}$ αποτελείται από σταθερές οπότε καταλήγουμε:

$$\mathbf{V} = \mathbf{Z}Var(\boldsymbol{\beta})\mathbf{Z}' + Var(\mathbf{e}).$$

Γνωρίζουμε ότι $Var(\boldsymbol{\beta}) = \mathbf{G}$ και $Var(\mathbf{e}) = \mathbf{R}$. Επομένως καταλήγουμε στον εξής γενικό τύπο:

$$\mathbf{V} = Var(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}. \tag{2.9}$$

Έχοντας τον γενικό τύπο της διακύμανσης του διανύσματος $\mathbf{y}$, θα κάνουμε μια αναφορά στο μέσο διάνυσμα του. Καθώς το μέσο διάνυσμα των τυχαίων επιδράσεων ισούται με μηδέν, εύκολα καταλήγουμε ότι το μέσο διάνυσμα του $\mathbf{y}$ ισούται με $\mathbf{X}\boldsymbol{\alpha}$.

Στην συνέχεια θα παραθέσουμε την γενική μορφή των πινάκων $\mathbf{G}$, $\mathbf{R}$ ξεκινώντας με τον πίνακα $\mathbf{G}$, ο οποίος έχει διάσταση $(q\text{x}q)$, όπου q ο αριθμός των συντελεστών των τυχαίων επιδράσεων (ή τυχαίων συντελεστών).

$$\mathbf{G} = \begin{pmatrix} Var(\beta_1) & Cov(\beta_1, \beta_2) & \dots & Cov(\beta_1, \beta_q) \\ Cov(\beta_2, \beta_1) & Var(\beta_2) & \dots & Cov(\beta_2, \beta_q) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\beta_q, \beta_1) & Cov(\beta_q, \beta_2) & \dots & Var(\beta_q) \end{pmatrix}.$$

Ο πίνακας $\mathbf{R}$ έχει διάσταση $(n\text{x}n)$, όπου n ο αριθμός των παρατηρήσεων, με την γενική του μορφή να είναι η εξής:

$$\mathbf{R} = \begin{pmatrix} Var(e_1) & Cov(e_1, e_2) & \dots & Cov(e_1, e_q) \\ Cov(e_2, e_1) & Var(e_2) & \dots & Cov(e_2, e_q) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(e_q, e_1) & Cov(e_q, e_2) & \dots & Var(e_q) \end{pmatrix}.$$

Έχοντας τον γενικό τύπο των πινάκων $\mathbf{G}$, $\mathbf{R}$ θα τροποποιήσουμε την σχέση (2.9) για το μοντέλο τυχαίων επιδράσεων και το μοντέλο τυχαίων συντελεστών.

Υποθέτουμε ότι έχουμε δυο τυχαίες επιδράσεις και χωρίζουμε το διάνυσμα των τυχαίων επιδράσεων β σε δυο υποδιανύσματα, β_1 και β_2. Με αντίστοιχο τρόπο χωρίζουμε το $\mathbf{Z}$ σε $\mathbf{Z}_1$ και $\mathbf{Z}_2$. Έχουμε δηλαδή:

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

και

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 \end{bmatrix}.$$

Επομένως καταλήγουμε:

$$\mathbf{Z}\beta = \mathbf{Z}_1\beta_1 + \mathbf{Z}_2\beta_2.$$

Το διάνυσμα $\mathbf{G}$ γράφεται ως εξής:

$$\mathbf{G} = Var(\beta) = \begin{bmatrix} Var(\beta_1) & Cov(\beta_1, \beta_2') \\ Cov(\beta_2, \beta_1') & Var(\beta_2) \end{bmatrix} = \begin{bmatrix} \mathbf{G}_1 & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_2 \end{bmatrix}.$$

Επομένως η σχέση (2.9) γράφεται ως εξής:

$$\mathbf{V} = \mathbf{Z}_1\mathbf{G}_1\mathbf{Z}_1' + \mathbf{Z}_2\mathbf{G}_2\mathbf{Z}_2' + \mathbf{Z}_1\mathbf{G}_{12}\mathbf{Z}_2' + \mathbf{Z}_2\mathbf{G}_{21}\mathbf{Z}_1' + \mathbf{R}.$$

Ο παραπάνω τύπος αντιστοιχεί στην περίπτωση που έχουμε δυο τυχαίες επιδράσεις. Αν πάρουμε την γενική περίπτωση που έχουμε r τυχαίες επιδράσεις, τότε έχουμε:

$$\beta = \{\beta_i\}_{i=1}^{r} \qquad \text{και} \quad \mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_r] \;\; = \{\mathbf{Z}_i\}_{i=1}^{r}.$$

Για τον πίνακα $\mathbf{G}$ ισχύει ότι:

$$\mathbf{G} = \mathrm{Var}(\beta) = \{\mathbf{G}_{ii'}\}_{i,i'=1}^{r}.$$

Επομένως μπορούμε να καταλήξουμε στον γενικό τύπο της διακύμανσης για r τυχαίες επιδράσεις.

$$\mathbf{V} = \sum_{i=1}^{r} \mathbf{Z}_i\mathbf{G}_{ii}\mathbf{Z}_i' + \sum_{i=1}^{r} \sum_{\substack{i'=1 \\ i \neq i'}}^{r} \mathbf{Z}_i\mathbf{G}_{ii'}\mathbf{Z}_{i'}' + \mathbf{R}, \tag{2.10}$$

όπου:

$$\mathbf{G}_{ii} = \mathrm{Var}(\beta_i) \qquad \text{και} \qquad \mathbf{G}_{ii'} = \mathrm{Cov}(\beta_i, \beta'_{i'}).$$

Καθώς τα υπόλοιπα για τα δυο αυτά μοντέλα είναι ασυσχέτιστα μεταξύ τους, ο πίνακας $\mathbf{R}$ είναι διαγώνιος, επομένως αν υποθέσουμε ότι η διακύμανση των υπολοίπων ισούται με σ^2, η σχέση (2.10) γράφεται ως εξής:

$$\mathbf{V} = \sum_{i=1}^{r} \mathbf{Z}_i \mathbf{G}_{ii} \mathbf{Z}'_i + \sum_{i=1}^{r} \sum_{\substack{i'=1 \\ i \neq i'}}^{r} \mathbf{Z}_i \mathbf{G}_{ii'} \mathbf{Z}'_{i'} + \sigma^2 \mathbf{I}_n. \tag{2.11}$$

Για το μοντέλο τυχαίων συντελεστών ο τύπος (2.11) δεν μπορεί να απλοποιηθεί περισσότερο. Ωστόσο για το μοντέλο τυχαίων επιδράσεων, καθώς οι τυχαίες επιδράσεις είναι ασυσχέτιστες μεταξύ τους, καταλήγουμε ότι $\mathbf{G}_{ii'} = \mathbf{0}$. Δηλαδή $\mathbf{G} = \{\mathbf{G}_i\}_{i=1}^r$ με την σχέση (2.11) να γράφεται ως εξής:

$$\mathbf{V} = \sum_{i=1}^{r} \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}'_i + \sigma^2 \mathbf{I}_n.$$

Για κάθε τυχαία επίδραση γνωρίζουμε ότι οι παράμετροι της έχουν κοινή διακύμανση, επομένως ο πίνακας $\mathbf{G}$ για μία τυχαία επίδραση i που έχει q_i συντελεστές γράφεται ως εξής:

$$\mathbf{G} = \{\sigma_i^2 \mathbf{I}_{q_i}\}_{i=1}^r.$$

Επομένως ο τελικός τύπος για τον πίνακα διακύμανσης είναι ο εξής:

$$\mathbf{V} = \sum_{i=1}^{r} \mathbf{Z}_i \mathbf{Z}'_i \sigma_i^2 + \sigma^2 \mathbf{I}_n. \tag{2.12}$$

Καταλήγουμε ότι η σχέση (2.12) αντιστοιχεί σε ένα μοντέλο τυχαίων επιδράσεων, ενώ η σχέση (2.11) σε ένα μοντέλο τυχαίων συντελεστών.

Στην συνέχεια για κάθε μία από τις τρείς διαφορετικές κατηγορίες μικτών γραμμικών μοντέλων που έχουμε παρουσιάσει, θα υπολογίσουμε τον πίνακα διακύμανσης $\mathbf{V}$ με την βοήθεια ορισμένων παραδειγμάτων, χρησιμοποιώντας την γενική σχέση (2.9).

Μοντέλο Τυχαίων Επιδράσεων

Ο πίνακας $\mathbf{G}$ στο μοντέλο τυχαίων επιδράσεων είναι διαγώνιος, καθώς οι τυχαίες επιδράσεις είναι ασυσχέτιστες μεταξύ τους, και έχει διάσταση (qxq) όπου q ο αριθμός των συντελεστών των τυχαίων επιδράσεων. Η μορφή του είναι η εξής:

$$\mathbf{G} = \begin{pmatrix} Var(\beta_1) & 0 & \dots & 0 \\ 0 & Var(\beta_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Var(\beta_q) \end{pmatrix}.$$

Δηλαδή καταλήγουμε σε ένα διαγώνιο πίνακα που έχει κατά σειρά στην διαγώνιο την διακύμανση των συντελεστών των τυχαίων επιδράσεων που έχουμε ορίσει στο μοντέλο μας. Χρησιμοποιούμε το Παράδειγμα (2.3), κατά το οποίο οι επιδράσεις των εργατών θεωρούνται ως τυχαίες επιδράσεις με διακύμανση σ_e^2. Ο πίνακας $\mathbf{G}$ γράφεται ως εξής:

$$\mathbf{G} = \begin{pmatrix} \sigma_e^2 & 0 & 0 \\ 0 & \sigma_e^2 & 0 \\ 0 & 0 & \sigma_e^2 \end{pmatrix}.$$

Αν συμπεριλάβουμε στο μοντέλο μας την αλληλεπίδραση εργάτης-μηχανή, η οποία έχει οριστεί ως τυχαία επίδραση με διακύμανση που ισούται με σ_{em}^2, τότε έχουμε:

$$\mathbf{G} = \begin{pmatrix} \sigma_e^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_{em^2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_{em}^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_{em}^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma_{em}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_{em}^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_{em}^2 \end{pmatrix}.$$

Οι τρείς πρώτες τιμές της διαγωνίου αντιστοιχούν στους εργάτες, ενώ οι τελευταίες έξι αντιστοιχούν στην αλληλεπίδραση.

Ο πίνακας $\mathbf{R}$ τώρα, έχει διάσταση $(n\text{x}n)$ και αφού τα υπόλοιπα είναι ασυσχέτιστα μεταξύ τους καταλήγουμε σε διαγώνιο πίνακα με την εξής γενική μορφή:

$$\mathbf{R} = \begin{pmatrix} Var(e_1) & 0 & \dots & 0 \\ 0 & Var(e_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Var(e_n) \end{pmatrix}.$$

Αν υποθέσουμε ότι για το παράδειγμα μας η διακύμανση των υπολοίπων ισούται με σ^2, τότε ο πίνακας $\mathbf{R}$ γράφεται ως εξής:

$$\mathbf{R} = \begin{pmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{pmatrix}.$$

Έχοντας καταγράψει την γενική μορφή των πινάκων $\mathbf{R}$, $\mathbf{G}$ θα υπολογίσουμε την τελική μορφή του πίνακα $\mathbf{ZGZ}'$ με την βοήθεια του Παραδείγματος (2.3), αγνοώντας σε πρώτη φάση την αλληλεπίδραση.

Ο πίνακας $\mathbf{ZGZ}'$ μπορεί να προκύψει με βάση τους κανόνες πολλαπλασιασμού μεταξύ των πινάκων, ωστόσο στην περίπτωση που δεν έχει οριστεί κάποια αλληλεπίδραση ως τυχαία επίδραση και κάτω από συγκεκριμένες συνθήκες που θα αναφερθούν στην συνέχεια έχει σταθερή μορφή. Αυτό που επιδιώκουμε στα δεδομένα μας για κάθε τυχαία επίδραση ξεχωριστά, είναι να έχουμε συνεχόμενα τις παρατηρήσεις μας για κάθε παράμετρο τυχαίας επίδρασης. Έτσι καταλήγουμε σε μία μορφή του πίνακα $\mathbf{Z}$ όπως στο παράδειγμα μας, με συνεχόμενες τιμές 1 και συνεχόμενα 0 ανά στήλη. Τηρώντας αυτήν την προϋπόθεση ο πίνακας $\mathbf{ZGZ}'$ θα έχει συγκεκριμένη μορφή κάθε φορά. Η μορφή αυτή θα αποτελείται από τετραγωνικούς διαγώνιους υποπίνακες με κάθε έναν να αντιπροσωπεύει μια παράμετρο τυχαίας επίδρασης. Η διάσταση των υποπινάκων ισούται με τον αριθμό των παρατηρήσεων για την εκάστοτε παράμετρο. Τέλος ο αριθμός των διαγωνίων υποπινάκων ισούται με τον αριθμό των παραμέτρων των τυχαίων επιδράσεων που έχουμε ορίσει στο μοντέλο.

Ο πίνακας $\mathbf{ZGZ}'$ του Παραδείγματος (2.3) γράφεται ως εξής:

$$\mathbf{ZGZ}' = \begin{pmatrix} \sigma_e^2 & \sigma_e^2 & \sigma_e^2 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 \\ \sigma_e^2 & \sigma_e^2 & \sigma_e^2 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 \\ \sigma_e^2 & \sigma_e^2 & \sigma_e^2 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 \\ \sigma_e^2 & \sigma_e^2 & \sigma_e^2 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_e^2 & \sigma_e^2 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_e^2 & \sigma_e^2 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_e^2 & \sigma_e^2 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & \sigma_e^2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & \sigma_e^2 \end{pmatrix}.$$

Ο παραπάνω πίνακας αποτελείται από τρείς διαγώνιους υποπίνακες, όσες και οι παράμετροι τυχαίας επίδρασης. Ο κάθενας από αυτούς αντιστοιχεί σε μία παράμετρο, με τον πρώτο να έχει διάσταση (4x4) καθώς έχουμε 4 παρατηρήσεις για τον εργάτη$_1$, ο δεύτερος (3x3) κ.ο.κ.

Εύκολα τώρα καταλήγουμε στον ζητούμενο πίνακα $\mathbf{V}$, καθώς ο πίνακας $\mathbf{R}$ είναι διαγώνιος.

$$\mathbf{V} = \begin{pmatrix} \sigma_e^2+\sigma^2 & \sigma_e^2 & \sigma_e^2 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 \\ \sigma_e^2 & \sigma_e^2+\sigma^2 & \sigma_e^2 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 \\ \sigma_e^2 & \sigma_e^2 & \sigma_e^2+\sigma^2 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 \\ \sigma_e^2 & \sigma_e^2 & \sigma_e^2 & \sigma_e^2+\sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_e^2+\sigma^2 & \sigma_e^2 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_e^2 & \sigma_e^2+\sigma^2 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_e^2 & \sigma_e^2 & \sigma_e^2+\sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2+\sigma^2 & \sigma_e^2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & \sigma_e^2+\sigma^2 \end{pmatrix}.$$

Συμπερασματικά θα λέγαμε ότι ο πίνακας διακύμανσης $\mathbf{V}$, έχει αντίστοιχα διαγώνιους τετραγωνικούς υποπίνακες όπως ο πίνακας $\mathbf{ZGZ}'$. Επίσης η διαγώνιος του αποτελείται από το άθροισμα της διακύμανσης των εργατών και της διακύμανσης των υπολοίπων.

Στην περίπτωση που συμπεριλάβουμε την αλληλεπίδραση εργάτης-μηχανή στο μοντέλο μας τα πράγματα περιπλέκονται ακόμα περισσότερο. Προφανώς και πάλι μπορούμε να καταλήξουμε στον ζητούμενο πίνακα $\mathbf{V}$ πραγματοποιώντας τους πολλαπλασιασμούς μεταξύ των πινάκων.

Η επεξήγηση που θα δώσουμε για τον υπολογισμό του χωρίς την πραγματοποιήση των πολλαπλασιασμών, ισχύει για την γενική περίπτωση που έχουμε ορίσει την “A” μεταβλητή και την αλληλεπίδραση “A-B” ως τυχαίες επιδράσεις.

Με τον ίδιο τρόπο με το αρχικό μας παράδειγμα, δημιουργούμε τους τρείς διαγώνιους τετραγωνικούς υποπίνακες για τον κάθε εργάτη, ωστόσο αυτήν την φορά πρέπει να κατανοήσουμε που θα προστεθεί η διακύμανση της αλληλεπίδρασης. Ασχολούμαστε με την πρώτη παράμετρο τυχαίας επίδρασης, δηλαδή με τον εργάτη$_1$, γνωρίζοντας ότι ο πρώτος υποπίνακας έχει διάσταση (4x4) και δημιουργούμε τον εξής πίνακα:

$$
\begin{array}{c}
\\ E_1M_1 \\ E_1M_1 \\ E_1M_2 \\ E_1M_2
\end{array}
\begin{array}{c}
\begin{array}{cccc} E_1M_1 & E_1M_1 & E_1M_2 & E_1M_2 \end{array} \\
\begin{pmatrix}
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1
\end{pmatrix}
\end{array} .
$$

Το ερώτημα που τίθεται είναι πως δημιουργείται ένας τέτοιος πίνακας. Ασχολούμαστε μόνο με τις παρατηρήσεις που έχουμε για τον εργάτη$_1$ κατά σειρά, έτσι οι πρώτες δυο στήλες αντιπροσωπεύουν τις παρατηρήσεις που κατέγραψε για την μηχανή$_1$, ενώ οι δυο τελευταίες τις παρατηρήσεις για την μηχανή$_2$. Αντίστοιχα ισχύει το ίδιο και για τις γραμμές. Συμπληρώνουμε την θέση (1,1) του υποπίνακα μας με τον αριθμό 1, καθώς έχουμε ταύτιση της πρώτης γραμμής και της πρώτης στήλης αφού αποτυπώνουν παρατηρήσεις για την μηχανή$_1$. Αντίθετα στην θέση (1,3) τοποθετούμε το μηδέν, καθώς έχουμε παρατηρήσεις για διαφορετικές μηχανές. Με το ίδιο σκεπτικό συμπληρώνουμε και τον υπόλοιπο πίνακα, προσθέτοντας τελικά το σ^2_{em} στον πρώτο διαγώνιο υποπίνακα του $\mathbf{ZGZ'}$ στις αντίστοιχες θέσεις που έχουμε συμπληρώσει την μονάδα. Αντίστοιχη διαδικασία ακολουθείται και για τις υπόλοιπες παραμέτρους. Τελικά καταλήγουμε στον εξής πίνακα:

$$
\mathbf{ZGZ'} = \begin{pmatrix}
\sigma_e^2+\sigma_{em}^2 & \sigma_e^2+\sigma_{em}^2 & \sigma_e^2 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 \\
\sigma_e^2+\sigma_{em}^2 & \sigma_e^2+\sigma_{em}^2 & \sigma_e^2 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 \\
\sigma_e^2 & \sigma_e^2 & \sigma_e^2+\sigma_{em}^2 & \sigma_e^2+\sigma_{em}^2 & 0 & 0 & 0 & 0 & 0 \\
\sigma_e^2 & \sigma_e^2 & \sigma_e^2+\sigma_{em}^2 & \sigma_e^2+\sigma_{em}^2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_e^2+\sigma_{em}^2 & \sigma_e^2+\sigma_{em}^2 & \sigma_e^2 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_e^2+\sigma_{em}^2 & \sigma_e^2+\sigma_{em}^2 & \sigma_e^2 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_e^2 & \sigma_e^2 & \sigma_e^2+\sigma_{em}^2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2+\sigma_{em}^2 & \sigma_e^2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & \sigma_e^2+\sigma_{em}^2
\end{pmatrix} .
$$

Εύκολα τώρα καταλήγουμε στον ζητούμενο πίνακα $\mathbf{V}$, καθώς ο πίνακας $\mathbf{R}$ είναι διαγώνιος.

$$
\mathbf{V} = \begin{pmatrix}
\theta & \sigma_e^2+\sigma_{em}^2 & \sigma_e^2 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 \\
\sigma_e^2+\sigma_{em}^2 & \theta & \sigma_e^2 & \sigma_e^2 & 0 & 0 & 0 & 0 & 0 \\
\sigma_e^2 & \sigma_e^2 & \theta & \sigma_e^2+\sigma_{em}^2 & 0 & 0 & 0 & 0 & 0 \\
\sigma_e^2 & \sigma_e^2 & \sigma_e^2+\sigma_{em}^2 & \theta & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \theta & \sigma_e^2+\sigma_{em}^2 & \sigma_e^2 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_e^2+\sigma_{em}^2 & \theta & \sigma_e^2 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_e^2 & \sigma_e^2 & \theta & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \theta & \sigma_e^2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma_e^2 & \theta
\end{pmatrix} ,
$$

όπου $\theta = \sigma_e^2 + \sigma_{em}^2 + \sigma^2$.

Ο πίνακας διακύμανσης $\mathbf{V}$ έχει στην διαγώνιο του το άθροισμα της διακύμανσης των εργατών, της αλληλεπίδρασης και των υπολοίπων.

Μοντέλο Τυχαίων Συντελεστών

Σε αντίθεση με το μοντέλο τυχαίων επιδράσεων, εδώ γνωρίζουμε από το Κεφάλαιο 1.6 ότι υπάρχει συσχέτιση μεταξύ των τυχαίων συντελεστών του μοντέλου. Επομένως δεν καταλήγουμε σε διαγώνιο πίνακα, με την μορφή του πίνακα $\mathbf{G}$ να είναι αυτή που παρουσιάστηκε αρχικά.

Υποθέτουμε ότι έχουμε τρία άτομα που εφαρμόζουν τις δίαιτες A, B καταγράφοντας το σωματικό τους βάρος σε τυχαίες χρονικές στιγμές. Η δομή των δεδομένων παρουσιάζεται στον Πίνακα 2.3.

Παρατηρητής	**Δίαιτα**	**Χρόνος (σε ημέρες)**
1	A	t_{11}
1	A	t_{12}
1	A	t_{13}
1	A	t_{14}
2	B	t_{21}
2	B	t_{22}
2	B	t_{23}
3	A	t_{31}
3	A	t_{32}

Πίνακας 2.3: Δομή των δεδομένων.

Θεωρούμε την επίδραση των παρατηρητών (β_1, β_2, β_3) και την επίδραση του χρόνου σε κάθε παρατηρητή ξεχωριστά ($\beta_{1,t}$, $\beta_{2,t}$, $\beta_{3,t}$) ως τυχαίους συντελεστές. Ο πίνακας $\mathbf{Z}$ είναι ο εξής:

$$\mathbf{Z} = \begin{pmatrix} 1 & t_{11} & 0 & 0 & 0 & 0 \\ 1 & t_{12} & 0 & 0 & 0 & 0 \\ 1 & t_{13} & 0 & 0 & 0 & 0 \\ 1 & t_{14} & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & t_{21} & 0 & 0 \\ 0 & 0 & 1 & t_{22} & 0 & 0 \\ 0 & 0 & 1 & t_{23} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & t_{31} \\ 0 & 0 & 0 & 0 & 1 & t_{32} \end{pmatrix} .$$

Υποθέτουμε ότι η διακύμανση των παρατηρητών ισούται με σ_p^2, η διακύμανση της επίδρασης του χρόνου ισούται με σ_{pt}^2 και η από κοινού συνδιακύμανση ισούται με $\sigma_{p,t}$. Τότε ο πίνακας $\mathbf{G}$ γράφεται:

$$\mathbf{G} = \begin{pmatrix} \sigma_p^2 & \sigma_{p,t} & 0 & 0 & 0 & 0 \\ \sigma_{p,t} & \sigma_{pt}^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_p^2 & \sigma_{p,t} & 0 & 0 \\ 0 & 0 & \sigma_{p,t} & \sigma_{pt}^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_p^2 & \sigma_{p,t} \\ 0 & 0 & 0 & 0 & \sigma_{p,t} & \sigma_{pt}^2 \end{pmatrix}.$$

Όπως βλέπουμε και με την βοήθεια του πίνακα $\mathbf{G}$, η συσχέτιση εμφανίζεται μόνο μεταξύ των συντελεστών για τον ίδιο παρατηρητή, για παράδειγμα β_1 με $\beta_{1,t}$.

Ο πίνακας $\mathbf{R}$ είναι διαγώνιος, επομένως ισχύει:

$$\mathbf{R} = \begin{pmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{pmatrix}.$$

Ο πίνακας $\mathbf{ZGZ}'$ που προκύπτει είναι ο εξής:

$$\mathbf{ZGZ}' = \begin{pmatrix} \nu_{1,11} & \nu_{1,12} & \nu_{1,13} & \nu_{1,14} & 0 & 0 & 0 & 0 & 0 \\ \nu_{1,12} & \nu_{1,22} & \nu_{1,23} & \nu_{1,24} & 0 & 0 & 0 & 0 & 0 \\ \nu_{1,13} & \nu_{1,23} & \nu_{1,33} & \nu_{1,34} & 0 & 0 & 0 & 0 & 0 \\ \nu_{1,14} & \nu_{1,24} & \nu_{1,34} & \nu_{1,44} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \nu_{2,11} & \nu_{2,12} & \nu_{2,13} & 0 & 0 \\ 0 & 0 & 0 & 0 & \nu_{2,12} & \nu_{2,22} & \nu_{2,23} & 0 & 0 \\ 0 & 0 & 0 & 0 & \nu_{2,13} & \nu_{2,23} & \nu_{2,33} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \nu_{3,11} & \nu_{3,12} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \nu_{3,12} & \nu_{3,22} \end{pmatrix}.$$

όπου $\nu_{i,jk} = \sigma_p^2 + (t_{ij} + t_{ik})\sigma_{p,t} + t_{ij}t_{ik}\sigma_{pt}^2$.

Ο πίνακας $\mathbf{ZGZ}'$ έχει και αυτός διαγώνιους τετραγωνικούς υποπίνακες με διάσταση ίση με τις μετρήσεις μας για τον κάθε παρατηρητή. Εύκολα προκύπτει ο πίνακας διακύμανσης $\mathbf{V}$, αφού ο πίνακας $\mathbf{R}$ είναι διαγώνιος.

$$\mathbf{V} = \begin{pmatrix} \nu_{1,11}+\sigma^2 & \nu_{1,12} & \nu_{1,13} & \nu_{1,14} & 0 & 0 & 0 & 0 & 0 \\ \nu_{1,12} & \nu_{1,22}+\sigma^2 & \nu_{1,23} & \nu_{1,24} & 0 & 0 & 0 & 0 & 0 \\ \nu_{1,13} & \nu_{1,23} & \nu_{1,33}+\sigma^2 & \nu_{1,34} & 0 & 0 & 0 & 0 & 0 \\ \nu_{1,14} & \nu_{1,24} & \nu_{1,34} & \nu_{1,44}+\sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \nu_{2,11}+\sigma^2 & \nu_{2,12} & \nu_{2,13} & 0 & 0 \\ 0 & 0 & 0 & 0 & \nu_{2,12} & \nu_{2,22}+\sigma^2 & \nu_{2,23} & 0 & 0 \\ 0 & 0 & 0 & 0 & \nu_{2,13} & \nu_{2,32} & \nu_{2,33}+\sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \nu_{3,11}+\sigma^2 & \nu_{3,12} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \nu_{3,12} & \nu_{3,22}+\sigma^2 \end{pmatrix}.$$

Μοντέλο Μοτίβου Συνδιακυμάνσεων

Στο μοντέλο μοτίβου συνδιακυμάνσεων όπως έχουμε αναφέρει, μοντελοποιούμε την συσχέτιση μεταξύ των επαναλαμβανόμενων μετρήσεων για τον ίδιο παρατηρητή και δεν εισάγουμε κάποια τυχαία επίδραση ή τυχαίο συντελεστή. Επομένως ο πίνακας $\mathbf{G}$ είναι μηδενικός και ο υπολογισμός του πίνακα διακύμανσης ανάγεται στον υπολογισμό του πίνακα $\mathbf{R}$. Η γενική του μορφή είναι η εξής:

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{R}_2 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{R}_3 & \mathbf{0} & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{R}_4 & \mathbf{0} & \dots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{R}_5 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Δηλαδή έχουμε διαγώνιους τετραγωνικούς υποπίνακες $\mathbf{R}_i$, με κάθε ένα να αντιπροσωπεύει ένα παρατηρητή ή ένα αντικείμενο μελέτης, με διάσταση ίση με τις παρατηρήσεις που διαθέτουμε για τον εκάστοτε παρατηρητή.

Αν τώρα θεωρήσουμε ότι $Cov(e_i, e_j) = \theta_{ij}$, τότε μπορούμε να ορίσουμε την γενική μορφή του υποπίνακα $\mathbf{R}_i$ ως εξής:

$$\mathbf{R}_i = \begin{pmatrix} \sigma_1^2 & \theta_{12} & \theta_{13} & \theta_{14} & \dots \\ \theta_{21} & \sigma_2^2 & \theta_{23} & \theta_{24} & \dots \\ \theta_{31} & \theta_{32} & \sigma_3^2 & \theta_{34} & \dots \\ \theta_{41} & \theta_{42} & \theta_{43} & \sigma_4^2 & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Υποθέτουμε ότι έχουμε τρία άτομα που εφαρμόζουν τις δίαιτες Α, Β καταγράφοντας το σωματικό τους βάρος σε διαφορετικές επισκέψεις. Η δομή των δεδομένων παρουσιάζεται στον Πίνακα 2.4.

Παρατηρητής	**Δίαιτα**	**Επισκέψεις**
1	A	1
1	A	2
1	A	3
1	A	4
2	B	1
2	B	2
2	B	3
3	A	1
3	A	2

Πίνακας 2.4: Δομή των δεδομένων.

Θεωρούμε την πιο απλή συσχέτιση που μπορούμε να έχουμε στο μοντέλο, δηλαδή ότι είναι σταθερή και ίση με ρ. Τότε ο πίνακας $\mathbf{R}$ γράφεται ως εξής:

$$\mathbf{R} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & 0 & 0 & 0 & 0 & 0 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & 0 & 0 & 0 & 0 & 0 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & 0 & 0 & 0 & 0 & 0 \\ \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sigma^2 & \rho\sigma^2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \rho\sigma^2 & \sigma^2 \end{pmatrix}.$$

Με τον υπολογισμό του πίνακα $\mathbf{R}$ καταλήγουμε στο ζητούμενο μας που είναι ο υπολογισμός του πίνακα διακύμανσης $\mathbf{V}$, καθώς υπάρχει ταύτιση μεταξύ τους.

2.4 Εκτίμηση των Σταθερών Επιδράσεων

Στην προηγούμενη ενότητα έγινε αναφορά στον μαθηματικό ορισμό των μικτών γραμμικών μοντέλων. Καταγράψαμε ότι έχουμε την δυνατότητα να υπολογίσουμε τους συντελεστές των σταθερών και των τυχαίων επιδράσεων, ώστε να είμαστε σε θέση να γράψουμε αναλυτικά τον τύπο του μοντέλου. Στην ενότητα αυτή θα επικεντρωθούμε στην εκτίμηση των σταθερών επιδράσεων, δηλαδή στο διάνυσμα $\boldsymbol{\alpha}$. Προφανώς η εκτίμηση τους με τους αριθμητικούς τρόπους που θα περιγράψουμε στην συνέχεια δεν είναι πρακτική, καθώς μπορεί να γίνει πολύ πιο εύκολα με την βοήθεια οποιουδήποτε στατιστικού πακέτου, ωστόσο οφείλουμε να μελετήσουμε και να κατανοήσουμε το θεωρητικό και υπολογιστικό κομμάτι των διαφόρων μεθόδων εκτίμησης των σταθερών επιδράσεων.

2.4.1 Συνάρτηση Πιθανοφάνειας

Η εκτίμηση της παραμέτρου $\boldsymbol{\alpha}$ θα γίνει με την βοήθεια της συνάρτησης πιθανοφάνειας. Αποτελεί μια γνωστή συνάρτηση, καθώς εμφανίζεται για πρώτη φορά κατά την μελέτη του πιο απλού στατιστικού μοντέλου, του απλού γραμμικού μοντέλου και μετέπειτα χρησιμοποιείται και εξετάζεται κατά την μελέτη και πιο σύνθετων στατιστικών μοντέλων. Η συνάρτηση πιθανοφάνειας στην περίπτωση μελέτης ανεξάρτητων παρατηρήσεων, ορίζεται ως το γινόμενο της συνάρτησης πυκνότητας για κάθε παρατήρηση. Καθώς η μελέτη μας αφορά τα μικτά γραμμικά μοντέλα οι παρατηρήσεις δεν είναι ανεξάρτητες μεταξύ τους, επομένως η συνάρτηση πιθανοφάνειας βασίζεται στην πολυμεταβλητή κανονική κατανομή, της οποίας ο ορισμός παρουσιάζεται αμέσως μετά.

Ορισμός: Έστω ένα τυχαίο διάνυσμα $\mathbf{X}$ με διάσταση p, τότε λέγεται ότι ακολουθεί την πολυμεταβλητή κανονική κατανομή εάν η συνάρτηση πυκνότητας πιθανότητας δίνεται από τον εξής τύπο:

$$f(\mathbf{X}) = f(X_1, X_2, ..., X_p) = \left(\tfrac{1}{2\pi}\right)^{p/2} |\mathbf{\Sigma}|^{-1/2} exp\Big(-\tfrac{1}{2}(\mathbf{X}-\mathbf{m})'\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{m})\Big), \tag{2.13}$$

όπου $\mathbf{m}$ είναι το μέσο διανύσμα και $\mathbf{\Sigma}$ ο συμμετρικός-θετικά ορισμένος πίνακας διακύμανσης. Τότε γράφουμε ότι $\mathbf{X} \sim N_p(\mathbf{m}, \mathbf{\Sigma})$.

Γνωρίζουμε ότι το μέσο διάνυσμα του $\mathbf{y}$ ισούται με $\mathbf{X\alpha}$ και ο πίνακας διακύμανσης του με $\mathbf{V}$, επομένως αν χρησιμοποιήσουμε την σχέση (2.13) και εκμεταλλευτούμε το γεγονός πως η συνάρτηση πιθανοφάνειας ισούται με την συνάρτηση της πυκνότητας πιθανότητας για μια πολυμεταβλητή κανονική κατανομή, καταλήγουμε στον εξής τύπο για την συνάρτηση πιθανοφάνειας:

$$L = L(\boldsymbol{\gamma}, \boldsymbol{\alpha}; \mathbf{y}) = \frac{exp\Big(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha})\Big)}{(2\pi)^{(1/2)n}|\mathbf{V}|^{1/2}},$$

όπου $\boldsymbol{\gamma}$ το διάνυσμα των παραμέτρων διακύμανσης του μοντέλου. Δηλαδή η συνάρτηση πιθανοφάνειας ορίζεται ως προς τις παραμέτρους $\boldsymbol{\alpha}$, $\boldsymbol{\gamma}$.

Ωστόσο για τους υπολογισμούς χρησιμοποιείται η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας, η οποία δίνεται από την εξής σχέση:

$$l = -\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\alpha}) - \frac{1}{2}log|\mathbf{V}| - \frac{1}{2}nlog(2\pi). \tag{2.14}$$

Ο στόχος μας είναι η εύρεση της παραμέτρου $\boldsymbol{\alpha}$ που μεγιστοποιεί την παραπάνω συνάρτηση. Κάτι τέτοιο μπορεί να συμβεί με δυο διαφορετικούς τρόπους, οι οποίοι θα παρουσιαστούν στις επόμενες υποενότητες.

2.4.2 Εκτίμηση με την μέθοδο Μέγιστης Πιθανοφάνειας (ML)

Στην υποενότητα αυτή θα αναφερθούμε στην πρώτη μέθοδο, την μέθοδο μέγιστης πιθανοφάνειας (συμβολίζεται ως ML) θεωρώντας ότι ο πίνακας διακύμανσης $\mathbf{V}$ είναι γνωστός. Η διαδικασία για τον υπολογισμό των σταθερών επιδράσεων, μπορεί να είναι ανεξάρτητη του πίνακα $\mathbf{V}$, ωστόσο καθώς έχει συμμετοχή στον τελικό τύπο είναι υποχρεωτική η γνώση του. Διαφορετικά είμαστε αναγκασμένοι να τον αντικαταστήσουμε με μία εκτιμήτρια $\hat{\mathbf{V}}$.

Για την εύρεση της εκτιμήτριας μέγιστης πιθανοφάνειας, δηλαδή της παραμέτρου $\hat{\boldsymbol{\alpha}}$, θα εργαστούμε με αντίστοιχο τρόπο με αυτόν για την μεγιστοποίηση μίας συνάρτησης. Δηλαδή θα παραγωγίσουμε την λογαριθμοποιημένη συνάρτηση πιθανοφάνειας ως πρός το διάνυσμα της παραμέτρου $\boldsymbol{\alpha}$ και στην συνέχεια θα θέσουμε την σχέση ίση με το μηδέν. Έτσι θα είμαστε σε θέση να υπολογίσουμε την εκτιμήτρια $\hat{\boldsymbol{\alpha}}$ που μεγιστοποιεί την συνάρτηση αυτήν. Ωστόσο καθώς η συνάρτηση πιθανοφάνειας δεν είναι ορισμένη μόνο ως προς το $\boldsymbol{\alpha}$, θα έχουμε μερική παραγώγιση.

Για τον υπολογισμό της μερικής παραγώγου της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας ως προς το $\boldsymbol{\alpha}$, θα εργαστούμε αρχικά με την γενική μορφή, κατά την οποία έχουμε ότι το μέσο διάνυσμα ισούται με $\boldsymbol{\mu}$ και ο πίνακας διακύμανσης με $\mathbf{V}$. Τότε η σχέση (2.14) γράφεται ως εξής:

$$l = -\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu}) - \frac{1}{2}log|\mathbf{V}| - \frac{1}{2}nlog(2\pi).$$

Υποθέτουμε ότι κάθε στοιχείο του $\boldsymbol{\mu}$ μπορεί να γραφεί ως συνάρτηση του διανύσματος $\boldsymbol{\theta}$, δηλαδή ισχύει ότι: $\boldsymbol{\mu}$=$\boldsymbol{\mu}(\boldsymbol{\theta})$.

Τότε η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας γράφεται ως εξής:

$$l = -\frac{1}{2}\Big(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta})\Big)'\mathbf{V}^{-1}\Big(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta})\Big) - \frac{1}{2}log|\mathbf{V}| - \frac{1}{2}nlog(2\pi).$$

Η πρώτη παράγωγος της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας ως προς $\boldsymbol{\theta}$ γράφεται ως εξής:

$$\frac{\partial \mathbf{l}}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\theta}}\mathbf{V}^{-1}\Big(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\theta})\Big).$$

Στην περίπτωση μας έχουμε:

- $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\alpha}$.
- $\boldsymbol{\theta} = \boldsymbol{\alpha}$.
- $\dfrac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\theta}} = \mathbf{X}'$.

Επομένως η πρώτη παράγωγος της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας ως προς $\boldsymbol{\alpha}$, γράφεται ως εξής:

$$\frac{\partial \mathbf{l}}{\partial \boldsymbol{\alpha}} = \mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}),$$

από την οποία καταλήγουμε:

$$\mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}) = \mathbf{0}.$$

Επομένως η εκτιμήτρια $\hat{\boldsymbol{\alpha}}$ ισούται με:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

Η διακύμανση του $\hat{\boldsymbol{\alpha}}$ είναι η εξής:

$$\begin{aligned}
Var(\hat{\boldsymbol{\alpha}}) &= Var((\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}) \\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}Var(\mathbf{y})\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}.
\end{aligned}$$

Επομένως με την βοήθεια της μεθόδου μέγιστης πιθανοφάνειας υπολογίσαμε την εκτιμήτρια της παραμέτρου $\boldsymbol{\alpha}$ καθώς και την διακύμανση της.

2.4.3 Εκτίμηση με την Υπολειπόμενη μέθοδο Μέγιστης Πιθανοφάνειας (REML)

Η υπολειπόμενη μέθοδος μέγιστης πιθανοφάνειας (συμβολίζεται ως REML) αποτελεί την δεύτερη και τελευταία μέθοδο στην οποία θα αναφερθούμε για την εκτίμηση των σταθερών επιδράσεων. Αρχικά θα γίνει μια γενική περιγραφή της μεθόδου, προτού προχωρήσουμε στην περίπτωση του μικτού γραμμικού μοντέλου.

Όπως μας προϊδεάζει και το όνομα της μεθόδου εργαζόμαστε με τα υπόλοιπα. Καλούμαστε δηλαδή να ορίσουμε την συνάρτηση πιθανοφάνειας συναρτήσει των υπολοίπων και όχι βάση του διανύσματος $\mathbf{y}$. Πραγματοποιώντας την διαφοροποίηση αυτή, καταφέρνουμε να εξαλείψουμε τις σταθερές επιδράσεις από την συνάρτηση πιθανοφάνειας, δηλαδή τον όρο $\boldsymbol{\alpha}$. Στην συνέχεια καλούμαστε να υπολογίσουμε την συνάρτηση πιθανοφάνειας συναρτήσει των υπολοίπων. Ένας τρόπος για να συμβεί αυτό είναι η εύρεση ενός στατιστικού μοντέλου που περιγράφει τα υπόλοιπα, όμως στην περίπτωση μας θα αποφευχθεί καθώς είναι αρκετά πολύπλοκη διαδικασία και δεν είναι αυτός ο σκοπός μας.

Στην περίπτωση του μικτού γραμμικού μοντέλου, ορισμένοι συγγραφείς ορίζουν τα υπόλοιπα ως $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}} - \mathbf{Z}\hat{\boldsymbol{\beta}}$, ενώ κάποιοι άλλοι ως $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}}$. Δηλαδή είτε αποφασίζουν να αφαιρέσουν τον όρο $\mathbf{Z}\hat{\boldsymbol{\beta}}$ μιας και τον θεωρούν όρο σφάλματος είτε όχι. Στην εργασία αυτή, θα ορίσουμε τα πλήρη υπόλοιπα ως $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}}$ για να τα διαφοροποιήσουμε από τα συνήθη υπόλοιπα, $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}} - \mathbf{Z}\hat{\boldsymbol{\beta}}$. Στην συνέχεια θα ορίσουμε την συνάρτηση πιθανοφάνειας χρησιμοποιώντας τα πλήρη υπόλοιπα. Για τον ορίσμο της συνάρτησης πιθανοφάνειας θα χρησιμοποιήσουμε χωρίς απόδειξη ότι τα πλήρη υπόλοιπα είναι ένας γραμμικός συνδυασμός του $\mathbf{y}$ και επίσης ότι τα πλήρη υπόλοιπα και η εκτιμήτρια $\hat{\boldsymbol{\alpha}}$ είναι ανεξάρτητα. Έτσι η συνάρτηση πιθανοφάνειας για το διάνυσμα των σταθερών επιδράσεων $\boldsymbol{\alpha}$ και το διάνυσμα των παραμέτρων διακύμανσης $\boldsymbol{\gamma}$, μπορεί να γραφεί ως το γινόμενο της πιθανοφάνειας βασισμένη στα πλήρη υπόλοιπα και το $\hat{\boldsymbol{\alpha}}$ ως εξής:

$$L(\boldsymbol{\gamma}, \boldsymbol{\alpha}; \mathbf{y}) = L(\boldsymbol{\gamma}; \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}}) L(\boldsymbol{\alpha}; \hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma}).$$

Η συνάρτηση πιθανοφάνειας για το διάνυσμα των σταθερών επιδράσεων $\boldsymbol{\alpha}$ και το διάνυσμα των παραμέτρων διακύμανσης $\boldsymbol{\gamma}$ γράφεται ως εξής:

$$L(\boldsymbol{\gamma}, \boldsymbol{\alpha}; \mathbf{y}) \propto |\mathbf{V}|^{-1/2} exp\Big(-\tfrac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})\Big).$$

Χρησιμοποιώντας την πολυμεταβλητή κανονική κατανομή για την εκτιμήτρια $\hat{\boldsymbol{\alpha}}$, η οποία έχει γνωστή διακύμανση που υπολογίστηκε στην προηγούμενη ενότητα, καταλήγουμε στον εξής τύπο:

$$L(\boldsymbol{\alpha}; \hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma}) \propto |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|^{-1/2} exp\Big(\tfrac{1}{2}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha})\Big).$$

Με την βοήθεια της αρχικής σχέσης η συνάρτηση πιθανοφάνειας βασισμένη στα πλήρη υπόλοιπα γράφεται ως εξής:

$$L(\boldsymbol{\gamma}; \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}}) = \frac{L(\boldsymbol{\gamma}, \boldsymbol{\alpha}; \mathbf{y})}{L(\boldsymbol{\alpha}; \hat{\boldsymbol{\alpha}}, \boldsymbol{\gamma})}.$$

Επομένως καταλήγουμε:

$$L(\boldsymbol{\gamma}; \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}}) \propto |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|^{-1/2}|\mathbf{V}|^{-1/2} exp\Big(-\tfrac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}})\Big).$$

Η παραπάνω συνάρτηση αποτελεί την συνάρτηση πιθανοφάνειας υπολογισμένη με την μέθοδο REML, με την λογαριθμοποιημένη συνάρτηση πιθανοφάνειας να δίνεται από τον παρακάτω τύπο:

$$l = log(L(\boldsymbol{\gamma}; \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}})) = -\frac{1}{2}\Big(log|\mathbf{V}| - log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|^{-1} + (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\alpha}})\Big) - \frac{1}{2}nlog(2\pi).$$

Καταλήξαμε στην λογαριθμοποιημένη συνάρτηση πιθανοφάνειας στην οποία όπως έγινε αναφορά και στην αρχή της ενότητας απουσιάζει η σταθερή επίδραση $\boldsymbol{\alpha}$. Είναι εμφανές ότι υπάρχει η εκτιμήτρια $\hat{\boldsymbol{\alpha}}$, ωστόσο δεν είναι γραμμένη συναρτήσει του $\boldsymbol{\alpha}$.

Έχοντας την λογαριθμοποιημένη συνάρτηση για την μέθοδο REML είμαστε πλέον σε θέση να υπολογίσουμε την εκτιμήτρια $\hat{\boldsymbol{\alpha}}$. Ο τρόπος για να συμβεί αυτό είναι αντίστοχος με αυτόν της προηγούμενης ενότητας, δηλαδή παραγωγίζοντας την λογαριθμοποιημένη συνάρτηση πιθανοφάνειας ως προς την εκτιμήτρια $\hat{\boldsymbol{\alpha}}$. Καθώς οι δυο λογαριθμοποιημένες συναρτήσεις πιθανοφάνειας που υπολογίστηκαν με τις μεθόδους ML, REML διαφέρουν μόνο ως προς τον όρο $|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|^{-1}$ θα προκύψουν τα ίδια αποτελέσματα, δηλαδή:

- $\hat{\boldsymbol{\alpha}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$.
- $Var(\hat{\boldsymbol{\alpha}}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$.

2.4.4 Σύγκριση Μεθόδων ML - REML

Για την εκτίμηση των σταθερών επιδράσεων υπάρχουν και άλλες μέθοδοι, όπως των επαναληπτικών γενικευμένων ελαχίστων τετραγώνων, ωστόσο θεωρήθηκε σκόπιμο να αναφερθούμε σε αυτές τις δυο, καθώς είναι αυτές που χρησιμοποιούνται σε ένα μικτό γραμμικό μοντέλο. Επίσης στο επόμενο κεφάλαιο θα δούμε ότι κατά την προσαρμογή ενός μικτού γραμμικού μοντέλου με την βοήθεια της R, χρησιμοποιείται είτε η μέθοδος ML είτε η REML.

Κατά την περιγραφή τους οι δυο αυτές μέθοδοι είχαν ως κοινή αφετηρία την εύρεση της παραμέτρου, ουσιαστικά του διανύσματος $\boldsymbol{\alpha}$, που μεγιστοποιεί την λογαριθμοποιημένη συνάρτηση πιθανοφάνειας. Κάθε μία το έκανε με διαφορετικό τρόπο, ωστόσο η κατάληξη μετά από τους υπολογισμούς ήταν η ίδια. Σε αυτήν την ενότητα θα γίνει μια σύγκριση μεταξύ των δυο αυτών μεθόδων, αναφέροντας για αυτόν τον σκοπό κάποιες επιπλέον λεπτομέρειες που δεν ειπώθηκαν στις δυο προηγούμενες ενότητες.

Η διαφορά που εντοπίζεται μεταξύ των δυο αυτών μεθόδων, είναι ότι η διακύμανση της εκτιμήτριας που υπολογίστηκε με την βοήθεια της μεθόδου ML, μπορεί να είναι μεροληπτική κάτω από συγκεκριμένες προϋποθέσεις. Αν προσπαθήσουμε να ερμηνεύσουμε την έννοια μεροληψία, τότε από μαθηματικής πλευράς δηλώνει ότι η μέση τιμή της εκτιμήτριας της διακύμανσης διαφέρει από την πραγματική τιμή της, ενώ πρακτικά δηλώνει ότι η εκτιμήτρια υπερεκτιμά ή υποτιμά την πραγματική τιμή της διακύμανσης. Αξίζει να σημειωθεί παρόλο που δεν έχει γίνει ακόμα αναφορά, ότι και για την εκτίμηση των τυχαίων επιδράσεων χρησιμοποιείται η μέθοδος ML ή η μέθοδος REML. Το πρόβλημα της μεροληψίας ενδέχεται να εμφανιστεί και για την διακύμανση της εκτιμήτριας των τυχαίων επιδράσεων, στην περίπτωση που ο υπολογισμός τους γίνει με την μέθοδο ML. Οι προϋποθέσεις για την εμφάνιση της μεροληψίας στις δυο παραπάνω περιπτώσεις είναι κοινές και θα καταγραφούν στην συνέχεια.

Αμφότερες οι προϋποθέσεις έχουν να κάνουν με τα δεδομένα μας. Το πρόβλημα της μεροληψίας ενδέχεται να εμφανιστεί στην περίπτωση που έχουμε μη ισορροπημένα δεδομένα, δηλαδή έχουμε διαφορετικό αριθμό παρατηρήσεων ανά ομάδα. Η δεύτερη και η τελευταία προϋπόθεση έχει να κάνει με το μέγεθος του δείγματος, με το πρόβλημα της μεροληψίας να διογκώνεται με την μείωση του μεγέθους του δείγματος. Η εμφάνιση της μεροληψίας οδήγησε στην μελέτη και την προτίμηση της μεθόδου REML κατά την μελέτη ενός μικτού γραμμικού μοντέλου, καθώς με τον τρόπο που ορίστηκε και την προσέγγιση του $\boldsymbol{\alpha}$ που θεωρήθηκε παράμετρος και όχι σταθερά, δεν αντιμετωπίζει το ίδιο πρόβλημα.

Είναι σημαντικό να γίνει κατανοητό ότι οι τιμές των συντελεστών των σταθερών επιδράσεων και με τις δυο μεθόδους δεν θα διαφέρουν σημαντικά μεταξύ τους. Η διαφορά που θα προκύψει θα είναι σημαντική για τα τυπικά τους σφάλματα στην περίπτωση που εμφανίζεται το φαινόμενο της μεροληψίας, καθώς θα είναι μεροληπτικά προς τα κατώ.

Τέλος η σύγκριση δυο διαφορετικών μοντέλων αποτελεί μια συχνή μελέτη στην στατιστική, καθώς μας βοηθάει να επιλέξουμε το καταλληλότερο μοντέλο με βάση τα δεδομένα και το πρόβλημα που μελετάμε. Στην συγκεκριμένη περίπτωση ενδιαφερόμαστε για την σύγκριση μικτών γραμμικών μοντέλων που διαφέρουν ως προς τις παραμέτρους που έχουν χρησιμοποιηθεί. Μία γνωστή μέθοδος για την σύγκριση δυο διαφορετικών μοντέλων, αποτελεί ο έλεγχος του λόγου των πιθανοφανειών. Ο έλεγχος αυτός απαιτεί την τιμή της υπολογισμένης συνάρτησης πιθανοφάνειας και για τα δυο μοντέλα. Στην περίπτωση που τα δυο μοντέλα διαφέρουν μόνο ως προς τις σταθερές τους επιδράσεις, τότε μπορεί να γίνει χρήση μόνο της συνάρτησης πιθανοφάνειας που υπολογίστηκε με την μέθοδο ML, καθώς η μέθοδος REML απαιτεί τις ίδιες ακριβώς σταθερές επιδράσεις.

2.5 Εκτίμηση (ή Πρόβλεψη) των Τυχαίων Επιδράσεων

Έπειτα από την εκτίμηση των σταθερών επιδράσεων ακολουθεί η εκτίμηση των τυχαίων. Οι σταθερές επιδράσεις όπως έχουμε αναφέρει είναι σταθεροί αριθμοί, επομένως υπάρχει η δυνατότητα εκτίμησης τους. Αντίθετα για τις τυχαίες επιδράσεις ταιριάζει περισσότερο ο όρος πρόβλεψη, μιας και είναι τυχαίες μεταβλητές. Ωστόσο επιλέγουμε να χρησιμοποιήσουμε από εδώ και πέρα την πιο διαδεδομένη ορολογία, που είναι η εκτίμηση των τυχαίων επιδράσεων.

Κάθε παράμετρος τυχαίας επίδρασης με βάση τον ορισμό της, ακολουθεί την κανονική κατανομή με αναμενόμενη τιμή μηδέν και διακύμανση σ_β^2. Αν χρησιμοποιήσουμε το Παράδειγμα 2.3, κατά το οποίο έχουμε προσαρμόσει την επίδραση των εργατών ως τυχαία, τότε γίνεται κατανοητό ότι οι εκτιμήσεις των τιμών των τριών εργατών θα έχουν μεγάλη ομοιότητα αφού προέρχονται από την ίδια κατανομή. Ωστόσο αυτή η ομοιότητα δεν φαντάζει και τόσο λογική και για αυτόν τον λόγο η προσέγγιση μας θα είναι λίγο διαφορετική για τις παραμέτρους κάθε τυχαίας επίδρασης ξεχωριστά. Δηλαδή εφόσον έχουμε φτάσει στο σημείο να υπολογίσουμε την τιμή της διακύμανσης, τότε έχουμε την κατανομή σε πλήρη μορφή για κάθε παράμετρο τυχαίας επίδρασης. Έπειτα αυτό που επιδιώκουμε είναι να προβλέψουμε την θέση κάθε παραμέτρου τυχαίων επιδράσεων, στην προκειμένη περίπτωση κάθε εργάτη, στην συγκεκριμένη κανονική κατανομή.

Στην συνέχεια θα γίνει η εκτίμηση της τυχαίας επίδρασης $\boldsymbol{\beta}$, με την βοήθεια της μεθόδου ML. Ορίζουμε την συνάρτηση πιθανοφάνειας ως προς τα $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, δηλαδή τις σταθερές-τυχαίες επιδράσεις και το διάνυσμα των παραμέτρων διακύμανσης $\boldsymbol{\gamma}$. Η συνάρτηση πιθανοφάνειας μπορεί να γραφεί ως το γινόμενο της πιθανοφάνειας για το $\mathbf{y}$ δεδομένου όμως του διανύσματος των τυχαίων επιδράσεων β και του β ως εξής:

$$L(\alpha, \beta, \gamma; \mathbf{y}) = L(\alpha, \gamma_R; \gamma|\beta)L(\gamma_G; \beta),$$

όπου:

- γ_R =οι παράμετροι διακύμανσης του πίνακα $\mathbf{R}$.
- γ_G =οι παράμετροι διακύμανσης του πίνακα $\mathbf{G}$.

Χρησιμοποιούμε την πολυμεταβλητή κανονική κατανομή για τις μεταβλητές $\mathbf{y}|\beta$, β, γνωρίζοντας για την $\mathbf{y}|\beta$, ότι το μέσο διάνυσμα της ισούται με $\mathbf{X}\alpha + \mathbf{Z}\beta$ και έχει διακύμανση $\mathbf{R}$. Επομένως καταλήγουμε στην παρακάτω σχέση:

$$L(\alpha, \beta, \gamma; \mathbf{y}) \propto |\mathbf{R}|^{-1/2} exp\Big(-\tfrac{1}{2}(\mathbf{y} - \mathbf{X}\alpha - \mathbf{Z}\beta)'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\alpha - \mathbf{Z}\beta)\Big)$$
$$*|\mathbf{G}|^{-1/2} exp(-\tfrac{1}{2}\beta'\mathbf{G}^{-1}\beta).$$

Υπολογίζουμε την λογαριθμοποιήμενη συνάρτηση πιθανοφάνειας, η οποία ισούται με:

$$l = -\frac{1}{2}\Big(log|\mathbf{R}| + (\mathbf{y} - \mathbf{X}\alpha - \mathbf{Z}\beta)'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\alpha - \mathbf{Z}\beta) + log|\mathbf{G}| + \beta'\mathbf{G}^{-1}\beta\Big) - \frac{1}{2}nlog(2\pi).$$

Η εκτιμήτρια των τυχαίων επιδράσεων με την μέθοδο ML, θα προκύψει παραγωγίζοντας την παραπάνω σχέση ως προς β.

$$\begin{aligned} \frac{\partial \mathbf{l}}{\partial \beta} &= \mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\alpha - \mathbf{Z}\beta) - \mathbf{G}^{-1}\beta \\ &= -(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\beta + \mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\alpha). \end{aligned}$$

Θέτοντας την παραπάνω σχέση ίση με το μηδέν και λύνοντας την εξίσωση που θα προκύψει ως προς β καταλήγουμε στον εξής τύπο:

$$\hat{\beta} = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\alpha).$$

Λαμβάνοντας υπόψη ότι $\mathbf{R} = \sigma^2\mathbf{I}$ για το μοντέλο τυχαίων επιδράσεων και το μοντέλο τυχαίων συντελεστών, η σχέση γράφεται ως εξής:

$$\hat{\beta} = (\mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1}/\sigma^2)^{-1}\mathbf{Z}'(\mathbf{y} - \mathbf{X}\alpha).$$

Η σχέση απλοποιείται ακόμα περισσότερο αν χρησιμοποιήσουμε την γενική σχέση για την διακύμανση του μοντέλου, δηλαδή $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$.

$$\hat{\beta} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\alpha).$$

Τέλος η διακύμανση της εκτιμήτριας $\hat{\beta}$ ισούται με:

$$Var(\hat{\beta}) = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G} - \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}.$$

Για τον υπολογισμό των τυχαίων επιδράσεων με την βοήθεια της μεθόδου REML, ακολουθούμε αντίστοιχη διαδικασία χρησιμοποιώντας ωστόσο την συνάρτηση πιθανοφάνειας που υπολογίστηκε με την μέθοδο REML στην Ενότητα 2.4.3. Καθώς οι δυο λογαριθμοποιημένες συναρτήσεις πιθανοφάνειας υπολογισμένες με την μέθοδο ML και REML διαφέρουν μόνο ως προς τον όρο $|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|^{-1}$, θα καταλήξουμε στα ίδια αποτελέσματα. Ωστόσο όπως αναφέραμε και στην προηγούμενη ενότητα, η διακύμανση της εκτιμήτριας των τυχαίων επιδράσεων υπολογισμένη με την μέθοδο ML μπορεί να είναι μεροληπτική κάτω από τις συνθήκες που καταγράψαμε.

2.6 Εκτίμηση των Παραμέτρων Διακύμανσης

Στην συγκεκριμένη ενότητα θα ασχοληθούμε με την εκτίμηση των παραμέτρων διακύμανσης ενός μικτού γραμμικού μοντέλου. Η διακύμανση των υπολοίπων είναι γνωστό από την μελέτη πιο απλών στατιστικών μοντέλων, ότι είναι δυνατόν να εκτιμηθεί με την βοήθεια της μεθόδου ML ή χρησιμοποιώντας στην επίλυση του προβλήματος την ανάλυση διασποράς, γνωστή και ως ANOVA. Σε ένα μικτό γραμμικό μοντέλο για την εκτίμηση των παραμέτρων διακύμανσης μπορούν να χρησιμοποιηθούν οι μέθοδοι ML, REML και η ANOVA. Η ANOVA όπως γνωρίζουμε χρησιμοποιείται κάτω από συγκεκριμένες συνθήκες, απαιτώντας οι ανεξάρτητες μεταβλητές να είναι κατηγορικές και τα δεδομένα μας ισορροπημένα. Καθώς η μέθοδος αυτή δεν είναι διαδεδομένη και δεν προτιμάται για ένα μικτό γραμμικό μοντέλο και λόγω του ότι η R χρησιμοποιεί μόνο τις μεθόδους ML, REML για την εκτίμηση τους, θα γίνει αναφορά μόνο σε αυτές.

2.6.1 Μέθοδοι ML - REML

Το σκεπτικό και για τις δυο μεθόδους είναι το ίδιο, καθώς καλούμαστε να υπολογίσουμε τις παραμέτρους διακύμανσης που μεγιστοποιούν την λογαριθμοποιημένη συνάρτηση, η οποία έχει υπολογιστεί και για τις δυο μεθόδους στο Κεφάλαιο 2.3. Σε αντίθεση με τα πιο απλά στατιστικά μοντέλα, στα οποία η εκτίμηση της διακύμανσης των υπολοίπων με την μέθοδο ML γίνεται με σχετική ευκολία, στα μικτά γραμμικά μοντέλα για την εκτίμηση συνολικά των παραμέτρων διακύμανσης, κάτι τέτοιο δεν ισχύει και για τις δυο μεθόδους. Ο λόγος είναι ότι η σχέση που προκύπτει έπειτα από την παραγώγιση της λογαριθμοποιημένης συνάρτησης ως προς τις παραμέτρους διακύμανσεις δεν είναι γραμμική. Επίσης οι σχέσεις που προκύπτουν είναι τουλάχιστον δυο.

Καλούμαστε δηλαδή να λύσουμε ένα σύστημα μη γραμμικών εξισώσεων. Για να επιτευχθεί αυτό είμαστε αναγκασμένοι να χρησιμοποιήσουμε γνωστές μεθόδους λύσης μη γραμμικών εξισώσεων, όπως η Newton-Raphson. Καθώς η εκτίμηση τους με την βοήθεια ενός στατιστικού πακέτου είναι εφικτή και αρκετά πιο απλή, θεωρείται σκόπιμο να παραλειφθεί η απόδειξη μιας και είναι αρκετά πολύπλοκη.

2.7 Εξέταση της σημαντικότητας των Σταθερών Επιδράσεων

Στην Ενότητα αυτή θα αναφερθούμε σε δυο γνωστές μεθόδους για τον έλεγχο της στατιστικής σημαντικότητας των σταθερών επιδράσεων. Αξίζει να σημειωθεί πως οι μέθοδοι που θα παρουσιαστούν, είναι αυτές που θα χρησιμοποιηθούν και στο στατιστικό πακέτο της R για τον έλεγχο της στατιστικής σημαντικότητας των σταθερών επιδράσεων.

2.7.1 Διάστημα Εμπιστοσύνης

Το διάστημα εμπιστοσύνης αποτελεί μια γνωστή διαδικασία υπολογισμού του εύρους των πιθανών τιμών για τις παραμέτρους ενός στατιστικού μοντέλου. Στο μικτό γραμμικό μοντέλο το διάστημα εμπιστοσύνης κατασκευάζεται κατά κύριο λόγο για τις σταθερές επιδράσεις. Ένα διάστημα εμπιστοσύνης 95% για έναν συντελεστή μιας σταθερής επίδρασης με εκτιμώμενη τιμή x και τυπικό σφάλμα SE δίνεται από τον παρακάτω τύπο:

$$(x - 1.96SE, x + 1.96SE).$$

Εκτός του εύρους των πιθανών τιμών για τις σταθερές επιδράσεις λαμβάνουμε μία επιπλέον πληροφορία. Αν το μηδέν εμπεριέχεται στο διάστημα εμπιστοσύνης, τότε είμαστε βέβαιοι κατά 95% ότι η σταθερή επίδραση που αντιστοιχεί σε αυτό το διάστημα εμπιστοσύνης δεν είναι στατιστικά σημαντική, επομένως μπορεί να αφαιρεθεί από το μοντέλο μας.

2.7.2 Υπολογισμός t-τιμών για τους συντελεστές των Σταθερών Επιδράσεων

Μια διαφορετική προσέγγιση για την εξέταση της σημαντικότητας των σταθερών επιδράσεων, είναι ο υπολογισμός των t-τιμών τους. Καθώς οι σταθερές επιδράσεις είναι σταθεροί αριθμοί, ο υπολογισμός των t-τιμών τους για έναν τυχαίο συντελεστή α_1 μπορεί να συμβεί ως εξής:

$$t - value = \frac{\hat{\alpha_1}}{se(\hat{\alpha_1})},$$

όπου $\hat{\alpha_1}$ η εκτίμηση του συντελεστή και $se(\hat{\alpha_1})$ το τυπικό σφάλμα της εκτιμήτριας $\hat{\alpha_1}$. Η μηδενική υπόθεση (H_0) ορίζει ότι ο συντελεστής ισούται με μηδέν, ενώ η εναλλακτική υπόθεση (H_1) ορίζει ότι ο συντελεστής είναι διάφορος του μηδενός.

Η εύρεση της t-τιμής δεν αρκεί για να αποφανθούμε αν η σταθερή μας επίδραση είναι στατιστικά σημαντική. Γνωρίζουμε ότι χρειαζόμαστε την p-τιμή της και για αυτόν τον σκοπό χρειαζόμαστε τους βαθμούς ελευθερίας του ελέγχου.

Ωστόσο για τα μικτά γραμμικά μοντέλα ο προσδιορισμός των βαθμών ελευθερίας είναι αρκετά πολύπλοκη διαδικασία, λόγω της παρουσίας των τυχαίων επιδράσεων και της συσχέτισης που ενδεχομένως υπάρχει εντός των δεδομένων. Υπάρχουν αρκετές προσεγγίσεις για τον υπολογισμό τους, ωστόσο δεν θα επεκταθούμε. Στο Κεφάλαιο 3 με την βοήθεια του στατιστικού πακέτου της R, θα χρησιμοποιήσουμε την προσέγγιση των Kenward και Roger (Small sample inference for fixed effects from REML, Biometrics, 1997) για τον υπολογισμό τους.

2.8 Κριτήρια - Έλεγχοι για την Σύγκριση Μικτών Γραμμικών Μοντέλων

Στην ενότητα αυτή θα εισάγουμε τον αναγνώστη στους ελέγχους-κριτήρια που μπορούμε να χρησιμοποιήσουμε για την σύγκριση διαφορετικών μικτών γραμμικών μοντέλων. Πιο συγκεκριμένα θα παρουσιάσουμε τον έλεγχο του λόγου των πιθανοφανειών και τα κριτήρια AIC-BIC. Οι μέθοδοι αυτοί είναι γνωστοί καθώς χρησιμοποιούνται κατά κόρον για την σύγκριση στατιστικών μοντέλων που διαφέρουν ως προς τις παραμέτρους τους. Στις επόμενες δυο υποενότητες θα καταγράψουμε την μαθηματική τους πλευρά, με την εφαρμογή τους σε παραδείγματα να παρουσιάζεται στο επόμενο κεφάλαιο.

2.8.1 Ο Έλεγχος του λόγου των Πιθανοφανειών

Ο έλεγχος του λόγου των πιθανοφανειών για να πραγματοποιηθεί απαιτεί την τιμή της συνάρτησης πιθανοφάνειας. Ήδη γνωρίζουμε ότι η τιμή της μπορεί να υπολογιστεί είτε με την μέθοδο ML είτε με την μέθοδο REML. Επομένως ο έλεγχος του λόγου των πιθανοφανειών μπορεί να βασίζεται είτε στην μέθοδο ML είτε στην μέθοδο REML. Ήδη καταγράψαμε ότι στην περίπτωση που τα δυο μοντέλα διαφέρουν μόνο ως προς τις σταθερές τους επιδράσεις, τότε η επιλογή της μεθόδου ML είναι μονόδρομος.

Υποθέτουμε ότι έχουμε δυο μοντέλα, το A με συνάρτηση πιθανοφάνειας L_1 και το B με συνάρτηση πιθανοφάνειας L_2, με το μοντέλο A να έχει επιπλέον σταθερές ή τυχαίες επιδράσεις ή και τις δυο ταυτόχρονα σε σχέση με το B. Η συνάρτηση για να πραγματοποιηθεί ο έλεγχος του λόγου των πιθανοφανειών μεταξύ των δυο μοντέλων A, B είναι η εξής:

$$2(log(L_1) - log(L_2)) \sim \chi^2_n,$$

η οποία ακουλουθεί την χ^2_n κατανομή με n βαθμούς ελευθερίας, όπου το n ισούται με τον αριθμό των παραμέτρων που διαφέρουν τα δυο μοντέλα.

Η μηδενική υπόθεση (H_0) ορίζει ότι οι επιπλέον επιδράσεις είναι ίσες με το μηδέν, ενώ η εναλλακτική υπόθεση (H_1) ορίζει ότι οι επιπλέον επιδράσεις είναι διάφορες του μηδενός. Διαφορετικά θα λέγαμε ότι αν επαληθευτεί η μηδενική υπόθεση επιλέγουμε το μοντέλο B, αλλιώς επιλέγουμε το μοντέλο A.

2.8.2 Κριτήρια AIC-BIC

Τα AIC, BIC αποτελούν κριτήρια επιλογής του βέλτιστου μοντέλου. Στην γενική περίπτωση το κριτήριο AIC όριζεται από την σχέση:

$$AIC = 2df - 2ln\hat{L},$$

όπου df οι ενεργοί βαθμοί ελευθερίας και $\hat{L}$ η μεγιστοποιημένη συνάρτηση πιθανοφάνειας υπολογισμένη με την μέθοδο ML ή με την μέθοδο REML.

Στην γενική περίπτωση το κριτήριο BIC ορίζεται από την σχέση:

$$BIC = dfln(n) - 2ln\hat{L},$$

όπου df οι ενεργοί βαθμοί ελευθερίας, n ο αριθμός των παρατηρήσεων και $\hat{L}$ η μεγιστοποιημένη συνάρτηση πιθανοφάνειας υπολογισμένη με την μέθοδο ML ή με την μέθοδο REML.

Παρατηρούμε ότι στους τύπους των δυο κριτηρίων αναφερθήκαμε στους ενεργούς βαθμούς ελευθερίας, οι οποίοι χρησιμοποιούνται για την διόρθωση της μεροληψίας. Για πολύπλοκα μοντέλα, όπως τα μικτά γραμμικά μοντέλα λόγω της παρουσίας των τυχαίων επιδράσεων, αναφερόμαστε σε ενεργούς βαθμούς ελευθερίας, καθώς δεν είναι απλό να υπολογιστεί ο αριθμός των παραμέτρων ως ο αριθμός των βαθμών ελευθερίας, μιας και οι βαθμοί ελευθερίας που χρησιμοποιούνται από το μοντέλο είναι λιγότεροι από τον αριθμό των παραμέτρων.

Ο υπολογισμός των ενεργών βαθμών ελευθερίας σε ένα μικτό γραμμικό μοντέλο δεν είναι απλή διαδικασία. Υπάρχουν αρκετές διαφορετικές προσεγγίσεις ωστόσο δεν θα επεκταθούμε. Στην συνέχεια και πιο συγκεκριμένα στο Κεφάλαιο 3, θα χρησιμοποιηθούν και τα δυο κριτήρια με τη βοήθεια της R για την σύγκριση διαφορετικών μικτών γραμμικών μοντέλων. Καθώς ο τρόπος υπολογισμού των ενεργών βαθμών ελευθερίας μέσω της R και κατ'επέκταση οι τιμές των δυο κριτηρίων AIC, BIC προκύπτουν με την ίδια προσέγγιση δεν αντιμετωπίζουμε κάποιο πρόβλημα ορθότητας.

Συμπερισματικά θα λέγαμε ότι οι τιμές των κριτηρίων AIC, BIC μπορεί να είναι βασισμένες είτε στην μέθοδο ML είτε στην μέθοδο REML. Το προτιμητέο μοντέλο με βάση τα κριτήρια AIC, BIC είναι εκείνο με την μικρότερη τιμή.

Κεφάλαιο 3

Ανάλυση Μικτών Γραμμικών Μοντέλων με τη βοήθεια της R

3.1 Εισαγωγή

Στο προηγούμενο κεφάλαιο παρουσιάσαμε την μαθηματική πλεύρα ενός μικτού γραμμικού μοντέλου, εστιάζοντας στον γενικό του τύπο, στην διακύμανση του μοντέλου, στον υπολογισμό των σταθερών επιδράσεων καθώς και στην εκτίμηση των τυχαίων επιδράσεων και της διακύμανσης τους. Οι υπολογισμοί-εκτιμήσεις των παραπάνω εννοιών έγιναν με την χρήση γνωστών μεθόδων, οι οποίες χρησιμοποιούνται κατά κόρον και σε άλλα στατιστικά μοντέλα. Ωστόσο εύκολα γίνεται αντιληπτό, ότι κατά την μελέτη ενός πραγματικού προβλήματος με ένα μεγάλο αριθμό παρατηρήσεων κάτι τέτοιο δεν είναι ούτε πρακτικό ούτε εφικτό.

Την λύση στο παραπάνω πρόβλημα έρχεται να μας την δώσει το στατιστικό πακέτο της R. Με την βοήθεια του, στις επόμενες ενότητες θα παρουσιάσουμε πως μπορούμε να προσαρμόσουμε ένα μικτό γραμμικό μοντέλο και να υπολογίσουμε όσα αναφέρθηκαν σε θεωρητικό επίπεδο στο Κεφάλαιο 2. Για τον σκοπό αυτό θα βασιστούμε σε συγκεκριμένα παραδείγματα, που με την βοήθεια τους θα δείξουμε πως προσεγγίζεται ένα μικτό γραμμικό μοντέλο με την χρήση της R, για όσο το δυνατόν περισσότερες διαφορετικές περιπτώσεις. Εκτός από τις εντολές που απαιτούνται στην R για τον υπολογισμό και την εξαγωγή των τιμών αυτών, θα δοθεί έμφαση στην ερμηνεία των αποτελεσμάτων.

Ένα μικτό γραμμικό μοντέλο μπορεί να ανταποκριθεί σε ένα μεγάλο εύρος διαφορετικών δεδομένων, όπως επίσης και σε ένα μεγάλο αριθμό διαφορετικών περιπτώσεων. Υπάρχει η δυνατότητα τα δεδομένα να έχουν μία ή περισσότερες πηγές ομαδοποίησης, η οποία μπορεί να είναι ισορροπημένη ή όχι. Η ομαδοποίηση ενδέχεται επίσης να είναι και ιεραρχική. Ωστόσο αξίζει να σημειωθεί ότι η προσέγγιση των μικτών γραμμικών μοντέλων με την βοήθεια της R, για οποιαδήποτε δομή δεδομένων είναι κατά μεγάλο ποσοστό η ίδια. Εφόσον γίνουν αντιληπτά όσα ειπωθούν στις επόμενες ενότητες, ο χειρισμός των διαφόρων δεδομένων θα γίνεται με σχετική ευκολία.

Στα δυο πρώτα κεφάλαια, έχουμε αναφέρει ότι θεωρούμε την επίδραση μιας μεταβλητής ως τυχαία επίδραση ή ως τυχαίο συντελεστή. Ωστόσο κατά την μελέτη των μικτών γραμμικών μοντέλων με την βοήθεια της R, δεν συνηθίζεται να χρησιμοποιούνται οι όροι αυτοί. Οι δυο νέοι όροι που χρησιμοποιούνται, είναι η τυχαία σταθερά και η τυχαία κλίση. Στην συνέχεια θα κάνουμε μία σύντομη πρώτη αναφορά για τους δυο αυτούς όρους, ωστόσο τόσο στην Ενότητα 3.2, όσο και στις επόμενες μέσω και των παραδειγμάτων θα γίνουν όλα πιο σαφή. Οι δυο νέοι αυτοί όροι, είναι ουσιαστικά συγκεκριμένοι τύποι των τυχαίων επιδράσεων και των τυχαίων συντελεστών και αναφέρονται σε συγκεκριμένα πράγματα.

Μία τυχαία σταθερά, αντιπροσωπεύει την μεταβλητότητα μεταξύ των διαφορετικών ομάδων ή συστάδων μιας επίδρασης. Αν θεωρήσουμε την επίδραση μιας κατηγορικής μεταβλητής ως τυχαία σταθερά, τότε υποθέτουμε ότι οι τιμές της μεταβάλλονται τυχαία εντός των επιπέδων της κατηγορικής μεταβλητής.

Μία τυχαία κλίση τώρα, χρησιμοποιείται για να ληφθεί υπόψη η μεταβλητότητα της επίδρασης μιας συνδιακύμανσης στις διαφορετικές ομάδες μιας κατηγορικής μεταβλητής. Πιο συγκεκριμένα μία τυχαία κλίση, επιτρέπει στην επίδραση μιας συνδιακύμανσης να μεταβάλλεται τυχαία στα διάφορα επίπεδα μιας κατηγορικής μεταβλητής.

Διαφορετικά θα λέγαμε, ότι μία τυχαία σταθερά επιτρέπει σε κάθε ομάδα να έχει την δική της μοναδική σταθερά, ενώ μία τυχαία κλίση επιτρέπει σε κάθε ομάδα να έχει την δική της μοναδική κλίση.

Με βάση όσα καταγράψαμε νωρίτερα, αν θυμηθούμε τον ορισμό του Μοντέλου Τυχαίων Συντελεστών στο Κεφάλαιο 1.7, εύκολα συμπεραίνουμε ότι η επίδραση των παρατηρητών είναι μία τυχαία σταθερά, ενώ η επίδραση του χρόνου σε κάθε παρατηρητή αποτελεί την τυχαία μας κλίση.

3.2 Περιπτώσεις με Τυχαία Σταθερά και Τυχαία Κλίση

Στην ενότητα αυτή θα μελετήσουμε δυο βασικές περιπτώσεις, εξηγώντας πλήρως πως μπορούν να αναλυθούν με την βοήθεια της R. Καθώς και οι δυο περιπτώσεις έχουν κάποια βασικά χαρακτηριστικά, στην συνέχεια θα αναφερθούμε σε αυτά προτού μεταβούμε στο παράδειγμα που θα αναλύσουμε με την βοήθεια της R. Η επεξήγηση των δυο αυτών περιπτώσεων αλλά και το παράδειγμα το οποίο θα αναλύσουμε στην συνέχεια, θα εξετάσει την περίπτωση που έχουμε μία μόνο σταθερή επίδραση, η οποία είναι αυστηρά ποσοτική. Σε αντίστοιχα συμπεράσματα μπορούμε να καταλήξουμε και για την περίπτωση που εκτός απο την ποσοτική σταθερή επίδραση, έχουμε επιπλέον σταθερές κατηγορικές επιδράσεις.

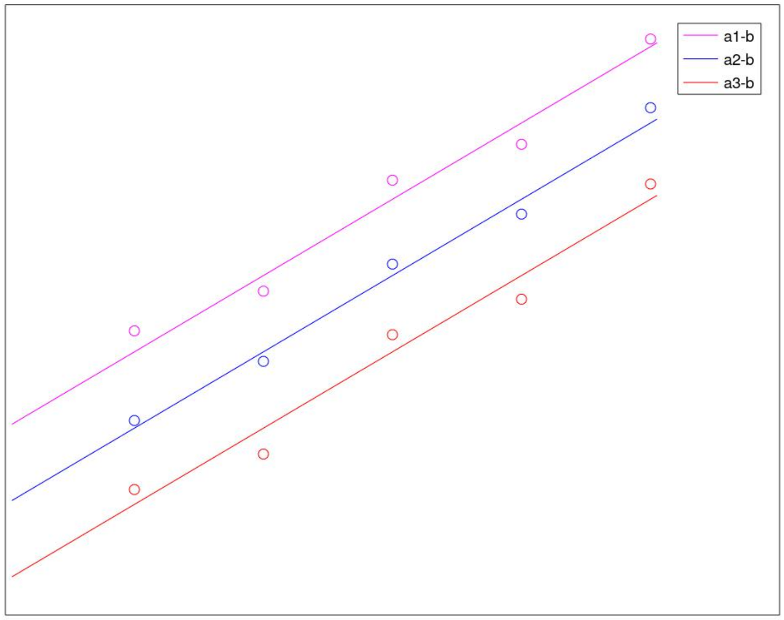
Για την πρώτη περίπτωση, ως πρώτο βήμα θα παρουσιάσουμε το διάγραμμα της γραμμικής παλινδρόμησης του μοντέλου, δηλαδή την σχέση της ποσοτικής σταθερής επίδρασης με την εξαρτημένη μεταβλητή. Το διάγραμμα θα σχεδιαστεί για τρία τυχαία στοιχεία της μεταβλητής ως προς την οποία υπάρχει ομαδοποίηση.

Από το Διάγραμμα 3.2, διαπιστώνουμε ότι έχουμε σταθερή κλίση και τυχαία σταθερά. Δηλαδή το μοντέλο επιτρέπει στις σταθερές να διαφέρουν τυχαία μεταξύ των ομάδων, καθορίζοντας έτσι την τιμή της εξαρτημένης μεταβλητής για κάθε μεμονωμένη παρατήρηση, αφού η κλίση είναι κοινή. Το μοντέλο μπορεί να εκφραστεί ως εξής:

$$y_{ij} = \mu + bx_{ij} + a_j + e_{ij}.$$

Οι συντελεστές μ, b αποτελούν την μέση σταθερά και κλίση αντίστοιχα, ενώ ο a_j αποτελεί την τυχαία σταθερά, η οποία είναι και ο λόγος που έχουμε απόκλιση από την μέση σταθερά. Από εδώ και στο εξής θα καλούμε ένα τέτοιο μοντέλο ως μοντέλο τυχαίων σταθερών, το οποίο ουσιαστικά είναι το μοντέλο τυχαίων επιδράσεων που παρουσιάσουμε στην Ενότητα 1.3, αλλά με διαφορετική ονομασία.

Για την δεύτερη περίπτωση, θα κινηθούμε ανάλογα παρουσιάζοντας πρώτα το διάγραμμα της γραμμικής παλινδρόμησης του μοντέλου για τρία τυχαία στοιχεία της μεταβλητής ως προς την



Διάγραμμα 3.2: Διάγραμμα Γραμμικής Παλινδρόμησης για το Μοντέλο Τυχαίων Σταθερών.

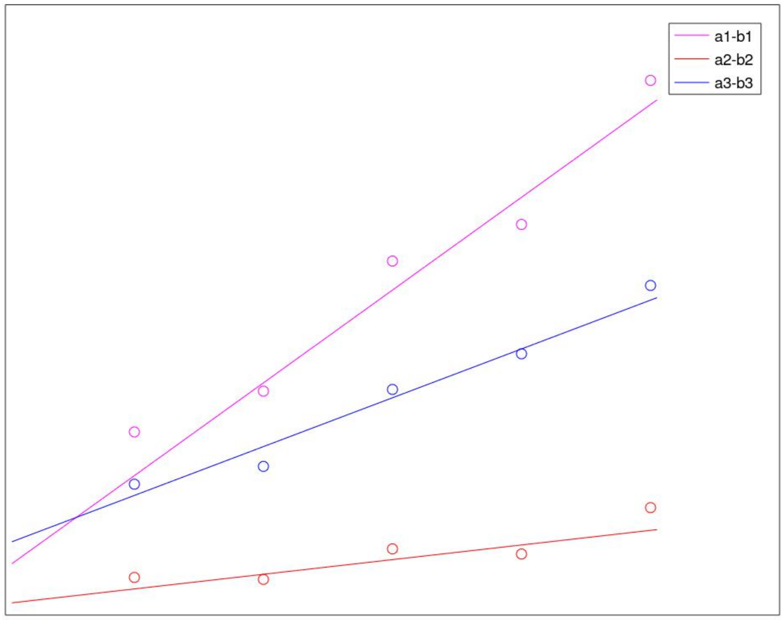
οποία υπάρχει ομαδοποίηση. Από το Διάγραμμα 3.3, διαπιστώνουμε ότι εκτός από τυχαία σταθερά, έχουμε και τυχαία κλίση. Δηλαδή το μοντέλο επιτρέπει και στις σταθερές και στις κλίσεις να διαφέρουν τυχαία μεταξύ των ομάδων, καθορίζοντας έτσι την τιμή της εξαρτημένης μεταβλητής για κάθε μεμονωμένη παρατήρηση. Το μοντέλο σε αυτήν την περίπτωση μπορεί να εκφραστεί ως εξής:

$$y_{ij} = \mu + a_j + (b + c_j)x_{ij} + e_{ij}.$$

Οι συντελεστές μ, b αποτελούν την μέση σταθερά και κλίση αντίστοιχα, ενώ οι a_j, c_j αποτελούν την τυχαία σταθερά και κλίση και αποτελούν τον λόγο που έχουμε απόκλιση από την μέση σταθερά και την μέση κλίση. Ουσιαστικά έχουμε μία σταθερή κλίση κοινή για όλα τα υποκείμενα του προβλήματος και μία τυχαία κλίση που μεταβάλλεται εντός των επιπέδων της μεταβλητής ως προς την οποία υπάρχει ομαδοποίηση. Από εδώ και στο εξής θα καλούμε ένα τέτοιο μοντέλο ως μοντέλο τυχαίων κλίσεων, το οποίο ουσιαστικά είναι το μοντέλο τυχαίων συντελεστών που παρουσιάσουμε στην Ενότητα 1.7, αλλά με διαφορετική ονομασία.

Έχοντας ολοκληρώσει την θεωρητική μας αναφορά στα δυο αυτά μοντέλα, αλλά και στις τυχαίες σταθερές και κλίσεις, γίνεται κατανοητό ότι η προσθήκη των δυο αυτών ειδικών όρων ήταν αναγκαίοι, καθώς κατά την μελέτη ενός μικτού γραμμικού μοντέλου με την βοήθεια της R, καλούμαστε να γίνουμε πιο συγκεκριμένοι.

Ανακεφαλαιώνοντας πρέπει να γίνει κατανοητό ότι ένα μοντέλο τυχαίων σταθερών, απαιτεί την ύπαρξη μόνο τυχαίων σταθερών στο μοντέλο. Αντίθετα ένα μοντέλο τυχαίων κλίσεων, απαιτεί την ύπαρξη τουλάχιστον μίας τυχαίας κλίσης και τουλάχιστον μίας τυχαίας σταθεράς.



Διάγραμμα 3.3: Διάγραμμα Γραμμικής Παλινδρόμησης για το Μοντέλο Τυχαίων Κλίσεων.

Δηλαδή κανένα από τα δυο μοντέλα δεν ενδιαφέρεται για τις σταθερές επιδράσεις που έχουν χρησιμοποιηθεί στο μοντέλο, είτε είναι μόνο ποσοτικές είτε είναι μόνο κατηγορικές είτε υπάρχει συνδυασμός τους. Επιλέχθηκε η επεξήγηση να γίνει για μία ποσοτική σταθερή επίδραση, ώστε να μπορεί να κατασκευαστεί και το διάγραμμα γραμμικής παλινδρόμησης του μοντέλου, με σκοπό να υπάρχει και οπτικοποίηση του προβλήματος για την πλήρη κατανόηση του. Στο Κεφάλαιο 3.3 θα ασχοληθούμε με ένα παράδειγμα το οποίο θα έχει μόνο κατηγορικές σταθερές επιδράσεις.

3.2.1 Μοντέλο Τυχαίων Σταθερών

Το πρώτο μοντέλο που θα εξετάσουμε με την βοήθεια της R, θα είναι το μοντέλο τυχαίων σταθερών. Στην συνέχεια με την βοήθεια ενός απλού παραδείγματος θα έχουμε την δυνατότητα να προσαρμόσουμε το μοντέλο, όπως επίσης να καταγράψουμε και να ερμηνεύσουμε τις τιμές που εξάγονται μέσω της R με την χρήση απλών συναρτήσεων.

Για την ανάλυση μας θα χρησιμοποιήσουμε τα δεδομένα sleepstudy της βιβλιοθήκης lme4 της R. Έχουμε λοιπόν 18 συμμετέχοντες με συνολικά δέκα παρατηρήσεις ο καθένας. Το πείραμα εξετάζει τον μέσο χρόνο αντίδρασης (σε milli-second) ανά ημέρα για άτομα που έχουν περιορισμένο ύπνο. Πιο συγκεκριμένα την ημέρα μηδέν τα υποκείμενα έχουν κοιμηθεί κανονικά καταγράφοντας τον χρόνο αντίδρασης, ενώ για τις επόμενες εννιά ημέρες και επομένως εννιά παρατηρήσεις έχουν περιορίσει τον ύπνο τους σε τρείς μόνο ώρες.

Έπειτα από την ανάγνωση του προβλήματος γίνεται κατανοητό ότι η μία επίδραση που θα μας απασχολήσει, είναι των ημερών και ότι η εξαρτημένη μεταβλητή του προβλήματος είναι ο χρόνος αντίδρασης. Επίσης φαντάζει λογικό να προσθέσουμε και την επίδραση των παρατηρητών στο μοντέλο μας, καθώς κάθε υποκείμενο είναι διαφορετικό. Οι παρατηρητές αποτελούν ένα δείγμα από ένα ευρύτερο πληθυσμό και επίσης δεν ενδιαφερόμαστε να καταλήξουμε σε συμπεράσματα για τους δεκαοκτώ συγκεκριμένους συμμετέχοντες, επομένως η επίδραση των υποκειμένων θεωρείται τυχαία.

Στην ανάλυση μας για τις διάφορες συναρτήσεις που θα χρησιμοποιήσουμε, θα χρειαστούμε την βιβλιοθήκη lme4 της R. Η συνάρτηση που χρησιμοποιείται κατά κύριο λόγο για την προσαρμογή ενός μικτού γραμμικού μοντέλου είναι η lmer. Η συνάρτηση lmer μοιάζει αρκετά με την συνάρτηση lm, που συνήθως χρησιμοποιείται κατά την προσαρμογή μοντέλων γραμμικής παλινδρόμησης. Ωστόσο στην περίπτωση μας έχουμε ένα νέο δεδομένο, καθώς πρέπει να προσθέσουμε και τις τυχαίες επιδράσεις. Στο παράδειγμα μας για να προσθέσουμε την τυχαία σταθερά του υποκειμένου, καλούμαστε να εισάγουμε το εξής όρισμα: (1 | Subject). Το παραπάνω όρισμα δηλώνει ότι επιτρέπουμε στην σταθερά, που αντιπροσωπεύεται από το 1, να μεταβάλλεται ανάλογα με το υποκείμενο ή γενικότερα να μεταβάλλεται εντός των επιπέδων της επίδρασης του δεξιού μέλους. Η μεταβλητή Reaction αποτελεί την εξαρτημένη μεταβλητή, ενώ η επίδραση των ημερών (Days) αποτελεί την σταθερή επίδραση. Τελικά το μοντέλο προσαρμόζεται ως εξής:

```
> random_intercept<-lmer(Reaction~Days+(1|Subject),
+ data=sleepstudy)
> random_intercept

Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (1 | Subject)
   Data: sleepstudy
REML criterion at convergence: 1786.465
Random effects:
 Groups   Name        Std.Dev.
 Subject  (Intercept) 37.12
 Residual             30.99
Number of obs: 180, groups:  Subject, 18
Fixed Effects:
(Intercept)         Days
     251.41        10.47
```

Έχουμε καταφέρει λοιπόν, με την βοήθεια της συνάρτησης lmer να προσαρμόσουμε το μοντέλο μας και να παρουσιάσουμε κάποιες βασικές πληροφορίες για αυτό. Το μοντέλο μπορεί να εκφραστεί ως εξής:

$$y_{ij} = \alpha_0 + \alpha_1 x_j + s_{0i} + e_{ij},$$

όπου το y_{ij} δηλώνει την τιμή του χρόνου αντίδρασης του i υποκειμένου την x_j ημέρα, τα α_0 και α_1 αποτελούν την σταθερή σταθερά και κλίση αντίστοιχα, ενώ το s_{0i} αποτελεί την τυχαία σταθερά. Προς το παρών δεν θα επεκταθούμε στα αποτελέσματα της R, καθώς με την χρήση της επόμενης συνάρτησης (summary) θα λάβουμε εκτός από τις παραπάνω πληροφορίες ακόμα περισσότερες.

```
> summary(random_intercept)

Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (1 | Subject)
   Data: sleepstudy

REML criterion at convergence: 1786.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2257 -0.5529  0.0109  0.5188  4.2506

Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 1378.2   37.12
 Residual              960.5   30.99
Number of obs: 180, groups:  Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept) 251.4051     9.7467   25.79
Days         10.4673     0.8042   13.02

Correlation of Fixed Effects:
     (Intr)
Days -0.371
```

Από τα παραπάνω αποτελέσματα γίνεται αντιληπτό ότι έχουμε υπολογίσει και παρουσιάσει αρκετές χρήσιμες πληροφορίες για το μοντέλο μας. Πιο συγκεκριμένα οι πληροφορίες που μας δίνονται κατά σειρά είναι οι εξής:

- Έχουμε προσαρμόσει ένα μικτό γραμμικό μοντέλο χρησιμοποιώντας την μέθοδο REML.
- Τον τύπο του μοντέλου και το όνομα του πλαισίου δεδομένων που χρησιμοποιήθηκε.
- Την τιμή της μεγιστοποιημένης συνάρτησης πιθανοφάνειας με την χρήση της μεθόδου REML.
- Για τα τα υπόλοιπα δίνονται: η ελάχιστη/μέγιστη τιμής τους, η διάμεσος και οι τιμές του πρώτου και τρίτου τεταρτημορίου.

- Για τις τυχαίες επιδράσεις ανά στήλη δίνονται: ο παράγοντας ομαδοποίησης των τυχαίων επιδράσεων, το όνομα της τυχαίας επίδρασης και τέλος οι εκτιμήσεις των τιμών της διακύμανσης και της τυπικής τους απόκλισης.

- Ο αριθμός των παρατηρήσεων, το όνομα της μεταβλητής ως προς την οποία υπάρχει ομαδοποίηση καθώς και ο συνολικός αριθμός των ομάδων.

- Για τις σταθερές επιδράσεις ανά στήλη δίνονται: τα ονόματα τους, οι εκτιμώμενες τιμές των συντελεστών τους, το τυπικό τους σφάλμα και η t-τιμή του στατιστικού ελέγχου σημαντικότητας των συντελεστών. Τέλος υπολογίζεται και η τιμή της συσχέτισης των δυο αυτών σταθερών επιδράσεων.

Οι συναρτήσεις lmer και summary, συνέβαλαν στο να υπολογίσουμε όσα καταγράφηκαν προηγουμένως. Ωστόσο εκτός από τις τιμές που παρουσιάστηκαν ήδη με την βοήθεια τους μπορούμε να εξάγουμε κάποιες επιπλέον χρήσιμες πληροφορίες, στις οποίες θα αναφερθούμε στην συνέχεια. Αρχικά θα επικεντρωθούμε στα αποτελέσματα των σταθερών επιδράσεων.

Γενικότερα σε περιπτώσεις με μία μόνο ποσοτική σταθερή επίδραση, η τιμή του συντελεστή της σταθεράς (intercept), δηλώνει την αναμενόμενη τιμή της εξαρτημένης μεταβλητής όταν η ανεξάρτητη μεταβλητή λάβει την τιμή μηδέν. Στην περίπτωση μας, η αναμενόμενη τιμή του χρόνου αντίδρασης την ημέρα 0 είναι 251.405ms.

Επίσης σε περιπτώσεις με μία μόνο ποσοτική σταθερή επίδραση, η τιμή του συντελεστή της κλίσης δηλώνει την μέση μεταβολή της εξαρτημένης μεταβλητής όταν αυξηθεί κατά μία μονάδα η ανεξάρτητη μεταβλητή. Στην περίπτωση μας, η τιμή του μέσου χρόνου αντίδρασης καθώς μεταβαίνουμε στην επόμενη ημέρα, αυξάνεται κατά 10.467ms. Συνδυάζοντας τα δυο παραπάνω συμπεράσματα, η μέση τιμή του χρόνου αντίδρασης την ημέρα 1 ισούται με: 261.872ms (251.405+10.467). Δηλαδή έχουμε την δυνατότητα να υπολογίσουμε την μέση τιμή του χρόνου αντίδρασης για οποιαδήποτε ημέρα.

Τέλος η συσχέτιση των σταθερών επιδράσεων, αφορά τους συντελεστές και όχι τις μεταβλητές. Αποτελεί ένα μέτρο πρόβλεψης που μας βοηθάει να κατανοήσουμε πως θα κινηθεί η τιμή του ενός συντελεστή, στην περίπτωση που επαναληφθεί το πείραμα και γνωρίζουμε ότι η τιμή του άλλου συντελεστή αυξήθηκε ή μειώθηκε. Στην περίπτωση μας εάν υποθέσουμε ότι αυξηθεί η εκτιμώμενη τιμή του συντελεστή των ημερών, τότε το πιο πιθανό είναι να μειωθεί η τιμή του σταθερού όρου, λόγω της αρνητικής τιμής της συσχέτισης.

Για τις παρατηρήσεις που θα καταγράψουμε για την τυχαία σταθερά, θα χρειαστούμε την αναμενόμενη τιμή του χρόνου αντίδρασης την ημέρα 0, η οποία είναι 251.405ms και την εκτίμηση της τυπικής απόκλισης της επίδρασης του υποκειμένου, η οποία ισούται με 37.12ms. Επίσης γνωρίζοντας πως οι τυχαίες επιδράσεις ακολουθούν την κανονική κατανομή, θα εκμεταλλευτούμε το γεγονός πως οι τιμές τους κυμαίνονται στο 68% των περιπτώσεων ± 1 τυπική απόκλιση από την μέση τιμή, ενώ στο 95% των περιπτώσεων κυμαίνονται ± 2 τυπικές αποκλίσεις από την μέση τιμή.

Αν επιστρέψουμε στο παράδειγμα μας, τότε μπορούμε να καταλήξουμε ότι τα υποκείμενα την ημέρα 0 θα έχουν προσωπικές σταθερές που θα κυμαίνονται από $(251.405 - 37.12, 251.405 + 37.12) = (214.285, 288.525)$ms στο 68% των περιπτώσεων, ενώ στο 95% των περιπτώσεων τα υποκείμενα θα έχουν προσωπικές σταθερές που θα κυμαίνονται από $(251.405-2*37.12, 251.405+2*37.12) = (177.165, 325.645)$ms. Σε αντίστοιχα συμπεράσματα μπορούμε να καταλήξουμε και για τις υπόλοιπες ημέρες, εφόσον πρώτα υπολογίσουμε τις αναμενόμενες τιμές τους.

Τέλος για τα υπόλοιπα θα χρησιμοποιήσουμε την εκτίμηση της τυπικής απόκλισης, η οποία ισούται με 30.99. Η τυπική απόκλιση αντιπροσωπεύει την μεταβλητότητα του χρόνου αντίδρασης τις δέκα διαφορετικές ημέρες γύρω από τις γραμμές παλινδρόμησης. Πιο συγκεκριμένα αφότου σχεδιαστεί μια προσωπική γραμμή παλινδρόμησης για κάθε υποκείμενο, τότε οι πραγματικές μετρήσεις τους θα κυμαίνονται με πιθανότητα 95%, $61.98 (= 2 * 30.99)$ μονάδες κατακόρυφα πάνω ή κάτω από τις γραμμές παλινδρόμησης.

Αξίζει να σημειωθεί ότι για τις σταθερές επιδράσεις μας δίνονται μόνο οι t-τιμές του στατιστικού ελέγχου και όχι οι p-τιμές. Ο λόγος είναι ότι δεν υπάρχει ξεκάθαρη επιλογή ως προς τους βαθμούς ελευθερίας που πρέπει να επιλεχθούν για τον υπολογισμό τους και για αυτό η συνάρτηση lmer αποφεύγει να τις υπολογίσει. Προφανώς υπάρχουν αρκετές προσεγγίσεις, χωρίς όμως να υιοθετούνται από την συνάρτηση lmer υπό τον φόβο λανθασμένων-προσεγγιστικών υπολογισμών. Ωστόσο τόσο για τις σταθερές όσο και για τις τυχαίες επιδράσεις, θα αναφέρουμε αργότερα διαφορετικές μεθόδους για τον έλεγχο της σημαντικότητας τους.

Η εκτίμηση των συντελεστών των τυχαίων επιδράσεων δεν εξάγεται στα αποτελέσματα της συνάρτησης summary. Για την εκτίμηση τους θα χρησιμοποιήσουμε την συνάρτηση ranef. Ενδεικτικά θα παρουσιάσουμε τις τιμές για τα πρώτα πέντε υποκείμενα.

```
> ranef(random_intercept)$Subject %>% head(5)

    (Intercept)
308   40.783710
309  -77.849554
310  -63.108567
330    4.406442
331   10.216189
```

Παρατηρούμε ότι για κάθε υποκείμενο μας δίνεται η τιμή της τυχαίας σταθεράς του. Διαφορετικά χρησιμοποιείται η συνάρτηση coef, η οποία εκτιμά το άθροισμα της σταθεράς της σταθερής και της τυχαίας επίδρασης. Ενδεικτικά και πάλι θα παρουσιάσουμε τις πέντε πρώτες τιμές.

```
> coef(random_intercept)$Subject %>% head(5)

    (Intercept)
308    292.1888
309    173.5556
310    188.2965
```

```
330    255.8115
331    261.6213
```

Οι δυο παραπάνω συναρτήσεις είναι αυτές που χρησιμοποιούνται για την εκτίμηση των τυχαίων επιδράσεων, με την κάθε μία να έχει την δικιά της χρησιμότητα ανάλογα με το τι αποτελέσματα επιθυμούμε να παρουσιάσουμε.

Επανερχόμαστε και πάλι στο ζήτημα της σημαντικότητας των σταθερών και των τυχαίων επιδράσεων. Για τον έλεγχο της σημαντικότητας των σταθερών επιδράσεων, η πιο συνηθισμένη και εύκολη επιλογή είναι ο υπολογισμός του διαστήματος εμπιστοσύνης. Έτσι ένα διάστημα εμπιστοσύνης 95% κατασκευάζεται στην R ως εξής:

```
> confint(random_intercept,oldNames=FALSE)

Computing profile confidence intervals ...
                            2.5 %    97.5 %
sd_(Intercept)|Subject  26.007120  52.93598
sigma                   27.813847  34.59105
(Intercept)            231.992326 270.81788
Days                     8.886551  12.04802
```

Για την σταθερή επίδραση των ημερών, παρατηρούμε ότι το μηδέν δεν εμπεριέχεται στο διάστημα εμπιστοσύνης, επομένως καταλήγουμε ότι αποτελεί στατιστικά σημαντική επίδραση για το μοντέλο μας. Τα διαστήματα εμπιστοσύνης των δυο πρώτων γραμμών, αντιστοιχούν στην επίδραση των υποκειμένων και στα υπόλοιπα αντίστοιχα και υπολογίζουν ένα διάστημα εμπιστοσύνης 95% για τις τυπικές τους αποκλίσεις.

Ωστόσο τα διαστήματα εμπιστοσύνης για τις σταθερές επιδράσεις, έχουν και μια διαφορετική ερμηνεία. Δηλαδή μπορούμε να καταγράψουμε ότι η τιμή του χρόνου αντίδρασης την ημέρα 0, θα κυμαίνεται από 232ms έως 270.818ms το 95% των περιπτώσεων. Επίσης καθώς μεταβαίνουμε στην επόμενη ημέρα, η αύξηση της τιμής του χρόνου αντίδρασης κυμαίνεται από 8.887ms έως 12.048ms το 95% των περιπτώσεων.

Για την εξέταση της σημαντικότητας των τυχαίων επιδράσεων καταλήγουμε στα κριτήρια AIC, BIC για τα δυο "ανταγωνιστικά" μοντέλα, με το ένα να περιέχει την τυχαία επίδραση που επιθυμούμε να εξετάσουμε ενώ το άλλο όχι. Αναλυτικότερη αναφορά και εφαρμογή των κριτηρίων θα γίνει στην επόμενη υποενότητα.

Τέλος θα καταγράψουμε τρείς επιπλέον βοηθητικές συναρτήσεις χωρίς όμως την παρουσίαση των αποτελεσμάτων τους. Για την χρησιμοποίηση των δυο πρώτων είναι απαραίτητη η βιβλιοθήκη merTools. Οι συναρτήσεις είναι οι εξής:

```
> predict(random_intercept)
> predictInterval(random_intercept)
> VarCorr(random_intercept)
```

Τα αποτελέσματα που μας δίνουν οι τρείς αυτές συναρτήσεις κατά σειρά είναι τα εξής:

- Εξάγει τις προβλεπόμενες τιμές της εξαρτημένης μεταβλητής, δηλαδή τις τιμές που αναμένεται να λάβει με βάση την εκτίμηση των συντελεστών του μοντέλου. Διαφορετικά θα λέγαμε ότι αναμένουμε τις τιμές αυτές για την εξαρτημένη μεταβλητή εφόσον χρησιμοποιηθούν νέα δεδομένα.
- Μαζί με τις προβλεπόμενες τιμές της εξαρτημένης μεταβλητής, δίνεται ένα διάστημα πρόβλεψης των τιμών αυτών το οποίο μπορεί να τροποποιηθεί.
- Δίνεται ξεχωριστά η τυπική απόκλιση των τυχαίων επιδράσεων καθώς και η τιμή της συσχέτισης εφόσον υπάρχει.

Τέλος αν εξαιρέσουμε τις τρείς παραπάνω συναρτήσεις, όλες οι υπόλοιπες είναι αυτές που χρησιμοποιούνται κατά κύριο λόγο σε μια ανάλυση ενός μικτού γραμμικού μοντέλου, ανεξαρτήτου δομής δεδομένων, τυχαίων επιδράσεων ή οτιδήποτε άλλο.

3.2.2 Μοντέλο Τυχαίων Κλίσεων

Για την εξέταση του μοντέλου τυχαίων κλίσεων θα χρησιμοποιήσουμε και πάλι τα δεδομένα sleepstudy που μας παρέχει η βιβλιοθήκη lme4 της R, έχοντας ωστόσο μια διαφοροποίηση σε σχέση με την προσέγγιση της προηγούμενης ενότητας. Έτσι λοιπόν, θα γίνει η υπόθεση ότι το κάθε υποκείμενο επηρεάζεται διαφορετικά με το πέρας των ημερών λόγω της έλλειψης ύπνου. Επομένως η σταθερή επίδραση των ημερών δεν είναι αρκετή, καθώς είναι κοινή για όλα τα υποκείμενα. Γι'αυτόν τον λόγο θα εισάγουμε μια νέα τυχαία επίδραση, με την βοήθεια της οποίας θα επιτρέπουμε στην επίδραση των ημερών να διαφέρει ανά υποκείμενο. Το όρισμα για να πραγματοποιηθεί κάτι τέτοιο στην R, λαμβάνοντας υπόψη και την τυχαία σταθερά των υποκειμένων, είναι το εξής: (1 + Days | Subject). Ο συμβολισμός αυτός εκτός από ότι επιτρέπει στην σταθερά που αντιπροσωπεύεται από το 1 να μεταβάλλεται εντός των υποκειμένων, επιτρέπει επίσης στην επίδραση του αριστερού μέλους (των ημερών) να διαφέρει ανά επίπεδο της επίδρασης του δεξιού μέλους (του υποκειμένου). Δηλαδή με την χρήση του ορίσματος αυτού, εκτός από την τυχαία σταθερά των υποκειμένων, έχουμε και μία τυχαία κλίση που αντιπροσωπεύει την επίδραση των ημερών σε κάθε υποκείμενο ξεχωριστά. Το μοντέλο προσαρμόζεται ως εξής:

```
> random_slope<-lmer(Reaction~Days+(1+Days|Subject),
+ data=sleepstudy)
> random_slope
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (1 + Days | Subject)
   Data: sleepstudy
REML criterion at convergence: 1743.628
Random effects:
 Groups   Name        Std.Dev. Corr
 Subject  (Intercept) 24.741
          Days         5.922   0.07
 Residual             25.592
Number of obs: 180, groups:  Subject, 18
Fixed Effects:
(Intercept)         Days
     251.41        10.47
```

Το μοντέλο το οποίο προσαρμόσαμε με την βοήθεια της συνάρτησης lmer μπορεί να εκφραστεί ως εξής:

$$y_{ij} = \alpha_0 + \alpha_1 x_j + s_{0i} + s_{1i} x_j + e_{ij},$$

όπου το y_{ij} δηλώνει την τιμή του χρόνου αντίδρασης του i υποκειμένου την x_j ημέρα, τα α_0 και α_1 αποτελούν την σταθερή σταθερά και κλίση αντίστοιχα, ενώ τα s_{0i}, s_{1i} αποτελούν την τυχαία σταθερά και κλίση. Αξίζει να σημειωθεί ότι τα βήματα που ακολουθούμε προς το παρών είναι αντίστοιχα με αυτά της προηγούμενης ενότητας. Για τον λόγο αυτό, θα χρησιμοποιήσουμε την συνάρτηση summary με σκοπό να λάβουμε πλήρεις πληροφορίες για το μοντέλο μας.

```
> summary(random_slope)

Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (1 + Days | Subject)
   Data: sleepstudy

REML criterion at convergence: 1743.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.9536 -0.4634  0.0231  0.4634  5.1793

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 Subject  (Intercept) 612.10   24.741
          Days         35.07    5.922   0.07
 Residual             654.94   25.592
Number of obs: 180, groups:  Subject, 18

Fixed effects:
```

```
              Estimate Std. Error t value
(Intercept)  251.405      6.825  36.838
Days          10.467      1.546   6.771

Correlation of Fixed Effects:
     (Intr)
Days -0.138
```

Οι πληροφορίες που λαμβάνουμε είναι οι ίδιες με την προηγούμενη περίπτωση, αν εξαιρέσουμε την προσθήκη της νέας τυχαίας επίδρασης και επομένως της συσχέτισης (στήλη Corr) που δημιουργείται λόγω της ύπαρξης δυο τυχαίων επιδράσεων. Οι τιμές και η ερμηνεία των σταθερών επιδράσεων παραμένουν οι ίδιες, καθώς δεν έχει αλλάξει κάτι για αυτές. Αντίθετα η διακύμανση της τυχαίας σταθεράς έχει μεταβληθεί λόγω της προσθήκης μιας νέας τυχαίας επίδρασης, με την ερμηνεία της να γίνεται με αντίστοιχο τρόπο. Στην συνέχεια όταν γίνει ο υπολογισμός των τυχαίων επιδράσεων θα παρατηρήσουμε ότι και οι εκτιμήσεις για την τυχαία σταθερά του υποκειμένου είναι διαφορετικές, κάτι που ήταν αναμενόμενο.

Για τις παρατηρήσεις που θα καταγράψουμε για την τυχαία κλίση, θα χρειαστούμε την εκτίμηση της τυπικής τους απόκλισης και την εκτίμηση του συντελεστή της σταθερής επίδρασης Days. Τότε καταλήγουμε ότι οι προσωπικές κλίσεις των υποκειμένων, θα κυμαίνονται από $(10.467 - 5.922, 10.467 + 5.922) = (4.545, 16.389)$ms στο 68% των περιπτώσεων, ενώ στο 95% των περιπτώσεων τα υποκείμενα θα έχουν προσωπικές κλίσεις που θα κυμαίνονται από $(10.467 - 2 * 5.922, 10.467 + 2 * 5.922) = (-1.377, 22.311)$ms.

Η συσχέτιση των τυχαίων επιδράσεων, αποτελεί επίσης ένα μέτρο πρόβλεψης της συμπεριφοράς των συντελεστών. Δηλαδή μια θετική συσχέτιση προβλέπει ότι η αύξηση της τιμής του ενός συντελεστή θα οδηγήσει κατά πάσα πιθανότητα και στην αύξηση του άλλου. Ωστόσο μας δίνει μια επιπλέον χρήσιμη πληροφορία όσον αφορά την από κοινού συνδιακύμανση των τυχαίων επιδράσεων. Γνωρίζουμε από το Κεφάλαιο 1, ότι $(s_{0i}, s_{1i}) \sim N(\mathbf{0}, \mathbf{G})$, όπου:

$$\mathbf{G} = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_2^2 \end{pmatrix},$$

με το $\sigma_{1,2}$ ή $\sigma_{2,1}$ να υπολογίζεται ως εξής:

$$\sigma_{1,2} = \sigma_{2,1} = 0.07 * \sqrt{\sigma_1^2 \sigma_2^2} = 10.255.$$

Δηλαδή έχοντας την διακύμανση και την συσχέτιση των δυο επιδράσεων από τα αποτελέσματα της R, με την βοήθεια της παραπάνω σχέσης μπορούμε να υπολογίσουμε την από κοινού συνδιακύμανση των δυο τυχαίων επιδράσεων.

Για την εκτίμηση των τυχαίων επιδράσεων αυτήν την φορά, εκτός από την εκτίμηση της επίδρασης του υποκειμένου, καλούμαστε να εκτιμήσουμε και τις τιμές της τυχαίας κλίσης, το οποίο γίνεται με την βοήθεια της συνάρτησης ranef. Ενδεικτικά θα παρουσιάσουμε τις πέντε πρώτες τιμές.

```
> ranef(random_slope)

$Subject
    (Intercept)        Days
308   2.2585509   9.1989758
309 -40.3987381  -8.6196806
310 -38.9604090  -5.4488565
330  23.6906196  -4.8143503
331  22.2603126  -3.0699116
```

Αντίστοιχα με την προηγούμενη ενότητα, με την χρήση της συνάρτησης coef μπορούμε να έχουμε το άθροισμα των σταθερών και των τυχαίων επιδράσεων. Στην περίπτωση μας στην πρώτη στήλη θα έχουμε το άθροισμα των σταθερών τους, ενώ στην δεύτερη το άθροισμα των κλίσεων τους.

```
> coef(random_slope)

$Subject
    (Intercept)       Days
308    253.6637 19.6662617
309    211.0064  1.8476053
310    212.4447  5.0184295
330    275.0957  5.6529356
331    273.6654  7.3973743
332    260.4447 10.1951090
```

Στο μοντέλο αυτό με την χρήση της συνάρτησης confint για την κατασκευή των διαστημάτων εμπιστοσύνης, εκτός της κατασκευής τους για τους συντελεστές των σταθερών επιδράσεων και τις τυπικές αποκλίσεις των τυχαίων επιδράσεων, κατασκευάζεται επίσης ένα διάστημα εμπιστοσύνης για την συσχέτιση της τυχαίας σταθεράς με την τυχαία κλίση. Προτού παρουσιάσουμε τα αποτελέσματα, οφείλουμε να αναφέρουμε ότι η συσχέτιση λαμβάνει τιμές στο διάστημα [-1,1].

```
> confint(random_slope,oldNames=FALSE)

Computing profile confidence intervals ...
                                 2.5 %     97.5 %
sd_(Intercept)|Subject       14.3814182 37.7159953
cor_Days.(Intercept)|Subject -0.4815008  0.6849863
sd_Days|Subject               3.8011641  8.7533808
sigma                        22.8982669 28.8579965
```

```
(Intercept)                     237.6806955 265.1295147
Days                              7.3586533  13.5759188
```

Παρατηρούμε ότι σε σύγκριση με τα διαστήματα εμπιστοσύνης που παρουσιάστηκαν στην προηγούμενη ενότητα, έχουμε επιπλέον το διάστημα εμπιστοσύνης για την τυπική απόκλιση της τυχαίας κλίσης και στην δεύτερη γραμμή έχουμε το διάστημα εμπιστοσύνης για την συσχέτιση των δυο τυχαίων επιδράσεων.

Επανερχόμαστε στο ζήτημα της σημαντικότητας των τυχαίων επιδράσεων. Όπως αναφέρθηκε στην Ενότητα 3.2.1, τα κριτήρια AIC, BIC αποτελούν μια λύση για την σύγκριση των μοντέλων. Έτσι υπολογίζονται οι τιμές των κριτηρίων αυτών για τα "ανταγωνιστικά" μοντέλα, με το μοντέλο που επιλέγεται τελικά να είναι αυτό που λαμβάνει την μικρότερη τιμή για τα κριτήρια αυτά. Στην συνέχεια θα παρουσιάσουμε τις τιμές των κριτηρίων αυτών για τα δυο μοντέλα που έχουμε προσαρμόσει έως τώρα:

```
> AIC(random_intercept)

[1] 1794.465

> BIC(random_intercept)

[1] 1807.237

> AIC(random_slope)

[1] 1755.628

> BIC(random_slope)

[1] 1774.786
```

Από τα παραπάνω αποτελέσματα καταλήγουμε ότι το δεύτερο μοντέλο, δηλαδή το μοντέλο με την τυχαία σταθερά και κλίση, είναι αυτό που προσαρμόζεται καλύτερα στα δεδομένα και στο πρόβλημα μας. Για το συμπέρασμα αυτό μας είχαν ήδη προϊδεάσει οι τιμές της τυπικής απόκλισης των υποκειμένων και των υπολοίπων, οι οποίες μειώθηκαν αισθητά με την προσθήκη της τυχαίας κλίσης.

Τέλος αξίζει να σημειωθεί ότι ο έλεγχος του λόγου των πιθανοφανειών, δεν προτιμάται στην περίπτωση που τα δυο μοντέλα διαφέρουν ως προς τις τυχαίες τους επιδράσεις. Ο λόγος είναι ότι ο έλεγχος της μηδενικής υπόθεσης είναι αρκετά συντηρικός, δίνοντας μεγαλύτερη εκτίμηση για την p-τιμή. Αυτό οφείλεται στο γεγονός πως ο έλεγχος πραγματοποιείται σε μία οριακή συνθήκη, $\sigma^2 \geq 0$. Παρόλο που μπορούμε να καταλήξουμε σε σωστά συμπεράσματα στην περίπτωση που η p-τιμή είναι αρκετά μικρή, ο ελέγχος αυτός αποφεύγεται.

3.3 Ένα Διαφορετικό Παράδειγμα

Το παράδειγμα της προηγούμενης ενότητας ήταν αρκετά συγκεκριμένο, περιέχοντας δυο επιδράσεις με την επίδραση των ημερών να αποτελεί την διακριτή σταθερή επίδραση. Το παράδειγμα το οποίο θα παρουσιάσουμε στην συνέχεια, εκτός από το γεγονός πως θα έχει τέσσερις διαφορετικές επιδράσεις, επομένως πολλές διαφορετικές παραλλαγές, θα περιέχει επίσης και κατηγορικές μεταβλητές προσαρμοσμένες ως σταθερές επιδράσεις. Αυτό σε πρώτη ανάγνωση σηματοδοτεί αλλαγές στην ερμηνεία των τιμών των συντελεστών των σταθερών επιδράσεων.

Γενικότερα μια κατηγορική μεταβλητή με p υποομάδες, της οποίας η επίδρασης της έχει θεωρηθεί ως σταθερή επίδραση, έχει $p-1$ συντελεστές με προεπιλεγμένη ψευδοκωδικοποίηση. Οι τιμές τους δηλώνουν την διαφορά από την ομάδα αναφοράς, η οποία έχει καθοριστεί από την R και η μέση τιμή της υποδηλώνεται μέσω της σταθεράς (Intercept).

3.3.1 Ανάλυση και Σύγκριση των διαφόρων Μοντέλων

Το παράδειγμα με το οποίο θα ασχοληθούμε στην Ενότητα 3.3 είναι το εξής:

Παράδειγμα 3.3

Τα δεδομένα του προβλήματος αντιστοιχούν σε ένα πείραμα στο οποίο μελετάται η συχνότητα ομιλίας (σε Hz) έξι παρατηρητών, τριών γυναικών και τριών ανδρών. Εξετάζεται λοιπόν η σχέση της συχνότητας ομιλίας για δυο διαφορετικές συνθήκες, την ευγενική και την ανεπίσημη, λαμβάνοντας επίσης υπόψη και την επίδραση του φύλου. Για να συμβεί αυτό καταγράφεται η τιμή της συχνότητας ομιλίας σε επτά διαφορετικά σενάρια, με το κάθε ένα να περιέχει την ευγενική και ανεπίσημη συνθήκη. Για παράδειγμα σε ένα σενάριο για την ευγενική συνθήκη οι μαθητές ζητούσαν από έναν καθηγητή τους μια χάρη, ενώ για την ανεπίσημη συνθήκη ζητούσαν μια χάρη από έναν συμμαθητή τους. Τέλος το σενάριο είχε την δυνατότητα να έχει κοινή εκδοχή για την ευγενική και την ανεπίσημη συνθήκη, δηλαδή ένα σενάριο αφορούσε την περίπτωση "συγγνώμη που ήρθα πολύ αργά".

Σειρά έχει η εισαγωγή των δεδομένων στην R και στην συνέχεια η μετατροπή των μεταβλητών, εκτός της συχνότητας, σε παράγοντες. Επίσης με την βοήθεια της συνάρτησης str θα παρουσιάσουμε μία λίστα, η οποία για κάθε μεταβλητή (ή αλλιως στήλη των δεδομένων) περιέχει τον τύπο της (αριθμητικός, παράγοντας κ.λ.π.). Αν είναι παράγοντας τότε εμφανίζει τον αριθμό των επιπέδων και τα ονόματα τους, ενώ αν είναι αριθμητική εμφανίζει τις πέντε πρώτες τιμές της.

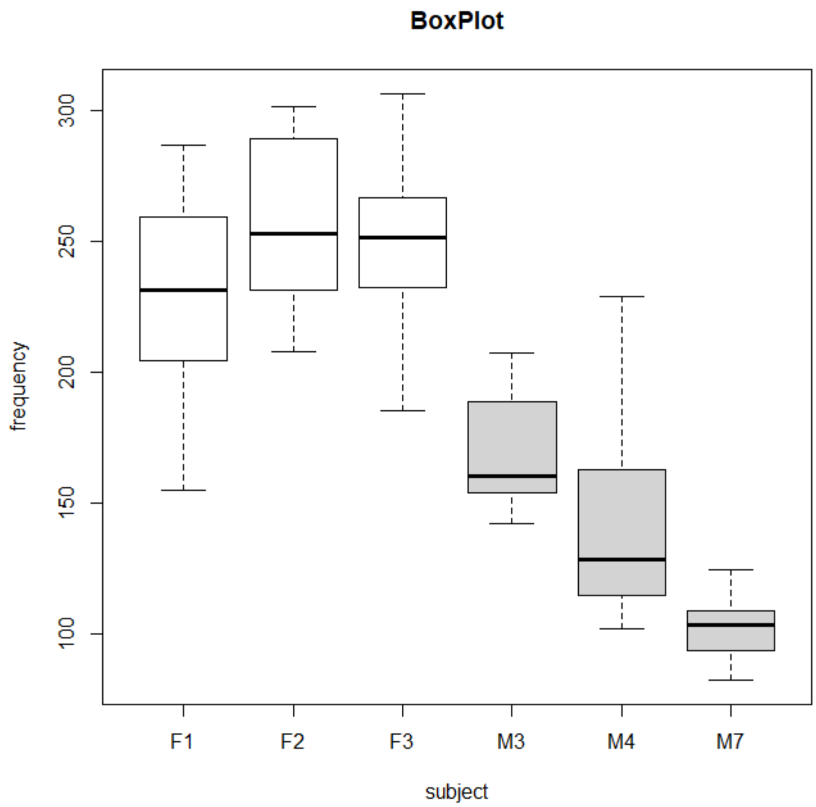
```
> politeness=
+ read.csv("http://www.bodowinter.com/tutorial/politeness
+ _data.csv")
> politeness[,"gender"]=factor(politeness[,"gender"])
> politeness[,"subject"]=factor(politeness[,"subject"])
> politeness[,"attitude"]=factor(politeness[,"attitude"])
```

```
> politeness[,"scenario"]=factor(politeness[,"scenario"])
> str(politeness)

data.frame':   84 obs. of  5 variables:
 $ subject  : Factor w/ 6 levels "F1","F2","F3",..
 $ gender   : Factor w/ 2 levels "F","M"
 $ scenario : Factor w/ 7 levels "1","2","3","4",..
 $ attitude : Factor w/ 2 levels "inf","pol"
 $ frequency: num  213 204 285 260 204 ...
```

Αν γίνουμε πιο συγκεκριμένοι, τότε ο συμβολισμός των τριών γυναικών είναι: F1, F2, F3, ενώ των τριών ανδρών είναι: M3, M4, M7, με τα επτά σεναρία να συμβολίζονται από 1 έως 7. Τέλος η ευγενική συνθήκη συμβολίζεται ως pol, ενώ η ανεπίσημη ως inf με το όνομα της μεταβλητής να είναι attitude.

Οι μεταβλητές που έχουμε ως επιλογή για το μοντέλο μας είναι αρκετές. Για να αποκτήσουμε μια πρώτη εικόνα για αυτές και τις διαφορές που ενδέχεται να έχουν μεταξύ των επιπέδων τους, μπορούμε να κατασκευάσουμε ένα θηκοδιάγραμμα. Ενδεικτικά θα παρουσιάσουμε το θηκοδιάγραμμα για την συχνότητα συναρτήσει των υποκειμένων.

```
> boxplot(frequency~subject,col=c("white","white","white",
+ "lightgray","lightgray","lightgray"),
+ politeness,main="BoxPlot")
```



Διάγραμμα 3.4: Θηκοδιάγραμμα Συχνότητας-Υποκειμένων για τα δεδομένα του Παραδείγματος 3.3.

Με την βοήθεια του Διαγράμματος 3.4, διαπιστώνουμε ότι υπάρχουν διαφοροποιήσεις στην τιμή της συχνότητας για τους διάφορους παρατηρητές, με την διακύμανση των τιμών εντός των υποκειμένων να φαίνεται αρκετά υψηλή. Επιπλέον τα τρία πρώτα θηκοδιαγράμματα αντιστοιχούν στις γυναίκες, ενώ τα επόμενα τρία στους άνδρες. Εύκολα διακρίνουμε ότι οι γυναίκες τείνουν να έχουν υψηλότερη συχνότητα ομιλίας σε σύγκριση με τους άνδρες. Αντίστοιχα διαγράμματα για να αποκτήσουμε μια πρώτη εικόνα, μπορούμε να κατασκευάσουμε και για τις μεταβλητές scenario και attitude ή για οποιοδήποτε συνδυασμό εξ'αυτών.

Οι επιδράσεις που έχουμε την δυνατότητα να συμπεριλάβουμε στο μοντέλο μας με οποιοδήποτε τρόπο είναι το υποκείμενο, το σενάριο, το φύλο και η συμπεριφορά. Οι δυο τελευταίες με βάση όσα έχουν ειπωθεί αποτελούν τις σταθερές επιδράσεις. Για την επίδραση της συμπεριφοράς αν θέλουμε να γίνουμε πιο συγκεκριμένοι, τότε μπορούμε να καταγράψουμε ότι μελετάμε δυο συνθήκες και ενδιαφερόμαστε να καταλήξουμε σε συμπεράσματα που αφορούν τις δυο αυτές συνθήκες και μόνο. Αντίθετα το υποκείμενο και το σενάριο αποτελούν τις τυχαίες επιδράσεις. Τα συμπεράσματα στα οποία θέλουμε να καταλήξουμε δεν επιθυμούμε να είναι βασισμένα ούτε στα επτά αυτά σενάρια ούτε στους έξι παρατηρητές, ο οποίος αποτελεί και τον κυριότερο λόγο που λαμβάνουμε αυτήν την απόφαση.

Το πρώτο μοντέλο που θα προσαρμόσουμε θα περιέχει την επίδραση της συμπεριφοράς και των υποκειμένων, καταλήγοντας έτσι στο εξής μοντέλο:

```
> model1<-lmer(frequency~attitude+(1|subject),data=politeness)
> summary(model1)

Linear mixed model fit by REML ['lmerMod']
Formula: frequency ~ attitude + (1 | subject)
   Data: politeness

REML criterion at convergence: 804.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2953 -0.6018 -0.2005  0.4774  3.1772

Random effects:
 Groups   Name        Variance Std.Dev.
 subject  (Intercept) 3982     63.10
 Residual              851     29.17
Number of obs: 83, groups:  subject, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  202.588     26.151   7.747
attitudepol  -19.376      6.407  -3.024

Correlation of Fixed Effects:
            (Intr)
attitudepol -0.121
```

Το μοντέλο που έχουμε προσαρμόσει με τη βοήθεια της συνάρτησης lmer είναι το εξής:

$$y_i = \alpha_0 + \alpha_1 I(attitude = pol) + \beta_{1,j[i]} + e_i, i = 1, 2, ..., 84,$$

όπου το y_i δηλώνει την τιμή της συχνότητας ομιλίας για την i παρατήρηση, τα α_0 και α_1 αποτελούν τις σταθερές σταθερές, ενώ το $\beta_{1,j[i]}$ αποτελεί την τυχαία σταθερά. Το $I(attitude = pol)$ αποτελεί μία συνάρτηση δείκτη και λαμβάνει την τιμή 1 στην περίπτωση που ισχύει η συνθήκη της παρένθεσης, ενώ σε αντίθετη περίπτωση λαμβάνει την τιμή 0. Τέλος το $j[i]$ δηλώνει το επίπεδο της ομάδας $(F_1, F_2, F_3, M_1, M_2, M_3)$ για το υποκείμενο i.

Το νέο δεδομένο που εντοπίζεται στην παραπάνω σύνοψη, αφορά τις σταθερές επιδράσεις λόγω της ύπαρξης της κατηγορικής επίδρασης attitude. Η ομάδα αναφοράς που έχει ορίσει η R είναι η ανεπίσημη συμπεριφορά. Με την βοήθεια των παραπάνω αποτελεσμάτων, καταλήγουμε ότι η μέση συχνότητα ομιλίας για την ανεπίσημη συνθήκη είναι αυξημένη κατά 19.376Hz σε σύγκριση με την ευγενική. Τέλος η μέση συχνότητα ομιλίας για την ανεπίσημη συνθήκη είναι

202.588Hz, ενώ για την ευγενική συνθήκη είναι 183.212Hz.

Η τυπική απόκλιση των υποκειμένων ισούται με 63.10Hz, με την ανεπίσημη συνθήκη να έχει μέσο όρο σταθεράς 202.588Hz. Επομένως μπορούμε να καταλήξουμε ότι τα υποκείμενα για την ανεπίσημη συνθήκη, θα έχουν προσωπικές σταθερές που θα είναι έως 63.1Hz χαμηλότερες ή υψηλότερες από τον μέσο όρο της ομάδας αυτής στο 68% των περιπτώσεων και μέχρι 126.2Hz χαμηλότερες ή υψηλότερες στο 95% των περιπτώσεων. Αντίστοιχα με τον ίδιο τρόπο καταλήγουμε σε ανάλογα συμπεράσματα και για την επίσημη συνθήκη.

Στην συνέχεια θα διαφοροποιήσουμε το μοντέλο μας, προσθέτοντας την επίδραση του φύλου αντί της συμπεριφοράς και έχοντας ως τυχαίες σταθερές τις επιδράσεις των υποκειμένων και των σεναρίων. Το μοντέλο στο οποίο καταλήγουμε είναι το εξής:

```
> model2<-lmer(frequency~gender+(1|subject)+(1|scenario)
+ ,data=politeness)
> summary(model2)

Linear mixed model fit by REML ['lmerMod']
Formula: frequency ~ gender + (1 | subject) + (1 | scenario)
   Data: politeness

REML criterion at convergence: 792.4

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.47496 -0.57398 -0.07435  0.61985  2.85496

Random effects:
 Groups   Name        Variance Std.Dev.
 scenario (Intercept) 206.0    14.35
 subject  (Intercept) 610.8    24.71
 Residual             751.0    27.40
Number of obs: 83, groups:  scenario, 7; subject, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)   246.99      15.84   15.59
genderM      -108.24      21.06   -5.14

Correlation of Fixed Effects:
        (Intr)
genderM -0.664
```

Το μοντέλο που έχουμε προσαρμόσει με τη βοήθεια της συνάρτησης lmer είναι το εξής:

$$y_i = \alpha_0 + \alpha_1 I(gender = male) + \beta_{1,j[i]} + \beta_{2,k[i]} + e_i, i = 1, 2, ..., 84,$$

όπου το y_i δηλώνει την τιμή της συχνότητας ομιλίας για την i παρατήρηση, τα α_0, α_1 αποτελούν τις σταθερές σταθερές, ενώ τα $\beta_{1,j[i]}$, $\beta_{2,k[i]}$ αποτελούν τις τυχαίες σταθερές. Τέλος το $j[i]$ δηλώνει το επίπεδο της ομάδας $(F_1, F_2, F_3, M_1, M_2, M_3)$ για το υποκείμενο i και το $k[i]$ δηλώνει το επίπεδο της ομάδας του σεναρίου (1,2,...,7) για το υποκείμενο i.

Αντίστοιχα με την προηγούμενη περίπτωση θα ερμηνεύσουμε το σταθερό μέρος του μοντέλου, με την ομάδα αναφοράς για την κατηγορική επίδραση του φύλου να είναι οι γυναίκες. Καταλήγουμε ότι η μέση συχνότητα ομιλίας για τις γυναίκες είναι αυξημένη κατά 108.24Hz σε σύγκριση με τους άνδρες. Η μέση συχνότητα ομιλίας για τις γυναίκες είναι 246.99Hz, ενώ για τους άνδρες είναι 138.75Hz. Το αποτέλεσμα αυτό ήταν αναμένομενο με βάση το Διάγραμμα 3.3.

Αν στο προηγούμενο μοντέλο συμπεριλάβουμε και την επίδραση της συμπεριφοράς τότε το μοντέλο γράφεται ως εξής:

```
> model3<-lmer(frequency~attitude+gender+(1|subject)+
+ (1|scenario),data=politeness)
> summary(model3)

Linear mixed model fit by REML ['lmerMod']
Formula: frequency ~ attitude + gender + (1 | subject) +
+ (1 | scenario)

   Data: politeness

REML criterion at convergence: 775.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2591 -0.6236 -0.0772  0.5388  3.4795

Random effects:
 Groups   Name        Variance Std.Dev.
 scenario (Intercept) 219.5    14.81
 subject  (Intercept) 615.6    24.81
 Residual             645.9    25.41
Number of obs: 83, groups:  scenario, 7; subject, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  256.846     16.116  15.938
attitudepol  -19.721      5.584  -3.532
genderM     -108.516     21.013  -5.164

Correlation of Fixed Effects:
            (Intr) atttdp
```

```
attitudepol -0.173
genderM     -0.652  0.004
```

Το μοντέλο που έχουμε προσαρμόσει με τη βοήθεια της συνάρτησης lmer είναι το εξής:

$$y_i = \alpha_0 + \alpha_1 I(attitude = pol) + \alpha_2 I(gender = male) + \beta_{1,j[i]} + \beta_{2,k[i]} + e_i, i = 1, 2, ..., 84,$$

όπου το y_i δηλώνει την τιμή της συχνότητας ομιλίας για την i παρατήρηση, τα α_0, α_1, α_2 αποτελούν τις σταθερές σταθερές, ενώ τα $\beta_{1,j[i]}$, $\beta_{2,k[i]}$ αποτελούν τις τυχαίες σταθερές. Τέλος το $j[i]$ δηλώνει το επίπεδο της ομάδας $(F_1, F_2, F_3, M_1, M_2, M_3)$ για το υποκείμενο i και το $k[i]$ δηλώνει το επίπεδο της ομάδας του σεναρίου (1,2,...,7) για το υποκείμενο i.

Αντίστοιχα με τις προηγούμενες περιπτώσεις, οι ομάδες αναφοράς είναι η γυναίκα και η ανεπίσημη συνθήκη. Λόγω της ύπαρξης και των δυο σταθερών επιδράσεων μπορούμε να καταλήξουμε στα εξής συμπεράσματα:

- Η μέση συχνότητα ομιλίας των γυναικών για την ανεπίσημη συνθήκη είναι 256.846Hz.
- Η μέση συχνότητα ομιλίας των γυναικών για την επίσημη συνθήκη είναι 237.125Hz (256.846−19.721).
- Η μέση συχνότητα ομιλίας των ανδρών για την ανεπίσημη συνθήκη είναι 148.33Hz (256.846−108.516).
- Η μέση συχνότητα ομιλίας των ανδρών για την επίσημη συνθήκη είναι 128.609Hz (256.846 −19.721−108.516).

Έχουμε δηλαδή την μέση τιμή της συχνότητας ομιλίας για οποιοδήποτε συνδυασμό των επιπέδων των σταθερών επιδράσεων, φύλο και συμπεριφορά. Επομένως έχουμε την δυνατότητα να εντοπίσουμε και να καταγράψουμε τυχόν διαφορές τους.

Για την σημαντικότητα των δυο σταθερών επιδράσεων, αντί της χρήσης του διαστήματος εμπιστοσύνης θα εισάγουμε μια νέα συνάρτηση. Η συνάρτηση αυτή ανήκει στο πακέτο afex και έχει όνομα mixed. Αναλύσαμε στην Ενότητα 3.2 τους λόγους για τους οποίους η συνάρτηση lmer δεν υπολογίζει τις p-τιμές. Για τον σκοπό αυτό δημιουργήθηκε η συνάρτηση mixed, η οποία υπολογίζει τις p-τιμές για τα μικτά γραμμικά μοντέλα που έχουν προσαρμοστεί με την βοήθεια της συνάρτησης lmer, χρησιμοποιώντας την προσέγγιση των Kenward-Rogers για τους βαθμούς ελευθερίας του ελέγχου. Η συνάρτηση mixed λαμβάνει ως ορίσματα το όνομα του μοντέλου (ή αλλιώς αναλυτικά τον τύπο του) και τα δεδομένα μας. Τα αποτελέσματα που μας δίνει η συνάρτηση είναι τα εξής:

```
> mixed(model3,data=politeness)

Contrasts set to contr.sum for the following variables:
attitude, gender, subject
```

```
Formula (the first argument) converted to formula.
Fitting one lmer() model. [DONE]
Calculating p-values. [DONE]
Mixed Model Anova Table (Type 3 tests, S-method)

Model: frequency ~ attitude + gender +
+ (1 | subject) + (1 | scenario)
Data: politeness
    Effect       df          F p.value
1 attitude 1, 70.05 12.47 ***    <.001
2   gender  1, 4.01  26.67 **     .007
---
Signif. codes:  0 ą***  0.001 ą**  0.01 ą*  0.05 ą+  0.1 ą   1
Warning message:
Due to missing values, reduced number of observations to 83
```

Παρατηρούμε ότι οι p-τιμές και για τις δυο σταθερές επιδράσεις, συμπεριφορά και φύλο, είναι μικρότερες του 0.05. Επομένως απορρίπτουμε την μηδενική υπόθεση, καταλήγοντας ότι και οι δυο είναι στατιστικά σημαντικές μεταβλητές για το μοντέλο μας.

Η ύπαρξη δυο σταθερών επιδράσεων, γεννά το ερώτημα αν η αλληλεπίδραση τους πρέπει να προστεθεί στο μοντέλο, δηλαδή αν είναι στατιστικά σημαντική επίδραση. Την απάντηση αυτή αφού προσαρμόσουμε ένα μοντέλο που περιέχει την αλληλεπίδραση, θα μας την δώσει ο έλεγχος του λόγου των πιθανοφανειών. Για να γίνει ένας τέτοιος έλεγχος χρησιμοποιείται η συνάρτηση anova, η οποία έχει ως προεπιλογή την μέθοδο ML. Με βάση όσα έχουμε καταγράψει στο Κεφάλαιο 2, εφόσον τα δυο μοντέλα διαφέρουν μόνο ως προς τις σταθερές τους επιδράσεις, η μέθοδος ML είναι η επιλογή μας. Επομένως ο ελέγχος είναι ο εξής:

```
> model_with_interaction<-lmer(frequency~attitude+gender+
+ attitude*gender+(1|subject)+(1|scenario),data=politeness)
> anova(model3,model_with_interaction)

refitting model(s) with ML (instead of REML)
Data: politeness
Models:
model3
model_with_interaction:
                       npar    AIC    BIC  logLik deviance
model3                    6 807.10 821.61 -397.55   795.10
model_with_interaction    7 807.11 824.04 -396.55   793.11

 Chisq   Df   Pr(>Chisq)

1.9963    1    0.1577 **
```

```
---
Signif. codes:  0 ą***  0.001 ą**  0.01 ą*  0.05 ą.  0.1 ą   1
```

Τα αποτελέσματα που λαμβάνουμε με την βοήθεια της R είναι τα εξής:

Οι βαθμοί ελευθερίας των μοντέλων, οι τιμές των κριτηρίων AIC, BIC, οι τιμές της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας υπολογισμένες με την μέθοδο ML και τέλος για τον στατιστικό έλεγχο του λόγου των πιθανοφανειών, η τιμή του, οι βαθμοί ελευθερίας και η p-τιμή του.

Η p-τιμή του ελέγχου του λόγου των πιθανοφανειών είναι μεγαλύτερη του 0.05, επομένως καταλήγουμε ότι δεν έχουμε λόγο να απορρίψουμε την μηδενική υπόθεση, δηλαδή ότι η αλληλεπίδραση δεν αποτελεί στατιστικά σημαντική επίδραση. Έχουμε δηλαδή καταλήξει στις σταθερές επιδράσεις που θα περιέχει το μοντέλο μας, χωρίς να έχει γίνει κάτι ανάλογο για τις τυχαίες επιδράσεις.

Μία επιπλέον παράμετρο που δεν έχουμε συνυπολογίσει, είναι το ενδεχόμενο η επίδραση της συμπεριφοράς να διαφέρει για τα διάφορα σενάρια όπως επίσης και για τους διάφορους παρατηρητές. Διαφορετικά θα λέγαμε, ότι επιτρέπουμε στα δυο επίπεδα της επίδρασης της συμπεριφοράς να μεταβάλλονται εντός των υποκειμένων και των σεναρίων. Δηλαδή για κάθε υποκείμενο και σενάριο θα έχουμε δυο ξεχωριστές τυχαίες επιδράσεις, μία τυχαία σταθερά και μία τυχαία κλίση, με την τυχαία κλίση να αντιπροσωπεύει την ευγενική συνθήκη. Τα ορίσματα για να καταλήξουμε σε αυτές τις τυχαίες επιδράσεις με βάση όσα αναφέρθηκαν στην Ενότητα 3.2.2 είναι τα εξής: (1+attitude | scenario), (1+attitude | subject).

Όπως γίνεται κατανοητό οι επιλογές για το τελικό μοντέλο είναι αρκετές. Οι δυο σταθερές επιδράσεις δεν επηρεάζονται από τις διαφορετικές τυχαίες επιδράσεις που θα έχει κάθε μοντέλο, ούτε ως προς την σημαντικότητα τους αλλά ούτε και οι τιμές τους. Επομένως έχοντας ως αφετηρία τις δυο αυτές επιδράσεις, οι διαφοροποιήσεις των μοντέλων θα αφορούν τις τυχαίες επιδράσεις. Ενδεικτικά στον παρακάτω πίνακα θα παρουσιάσουμε τις τιμές τους κριτηρίου BIC για συγκεκριμένους συνδυασμούς των τυχαίων επιδράσεων. Οι στήλες 1, 2 έχουν ισχύ όταν έχουμε τα εξής ορίσματα: (1 | subject), (1 | scenario), ενώ οι στήλες 1, 3 και 2, 4 έχουν ισχύ με τα εξής ορίσματα (1+attitude | subject), (1+attitude | scenario).

Τυχαία Σταθερά 1	Τυχαία Σταθερά 2	Τυχαία Κλίση 1	Τυχαία Κλίση 2	BIC
✓	✗	✗	✗	808.8381
✗	✓	✗	✗	827.0947
✓	✓	✗	✗	801.9678
✓	✗	✓	✗	817.6517
✓	✓	✓	✗	810.7806
✓	✓	✗	✓	810.4531
✓	✓	✓	✓	819.2661

Από τα παραπάνω αποτελέσματα καταλήγουμε ότι το τρίτο μοντέλο κατά σειρά είναι αυτό που προσαρμόζεται καλύτερα στα δεδομένα μας. Το μοντέλο αυτό περιέχει τις δυο σταθερές επιδράσεις, καθώς και τις δυο τυχαίες σταθερές και είναι ουσιαστικά το model3 που αναλύσαμε νωρίτερα.

Παραθέτουμε τα κυριότερα αποτελέσματα για το μοντέλο αυτό:

```
Formula: frequency ~ attitude + gender + (1 | subject) +
+ (1 | scenario)

   Data: politeness

Random effects:
 Groups   Name        Variance Std.Dev.
 scenario (Intercept) 219.5    14.81
 subject  (Intercept) 615.6    24.81
 Residual             645.9    25.41
Number of obs: 83, groups:  scenario, 7; subject, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  256.846     16.116  15.938
attitudepol  -19.721      5.584  -3.532
genderM     -108.516     21.013  -5.164

Computing profile confidence intervals ...
                  2.5 %     97.5 %
(Intercept)  226.084997 287.60823
attitudepol  -30.739625  -8.69669
genderM     -149.340247 -67.68308
```

Στην αρχική μας αναφορά για το μοντέλο, έχουμε καταγράψει την μέση τιμή της συχνότητας ομιλίας για οποιοδήποτε συνδυασμό των δυο επιπέδων των σταθερών επιδράσεων. Χρησιμοποιώντας τα παραπάνω αποτελέσματα, θα καταγράψουμε κάποια επιπλέον συμπεράσματα τα οποία δεν παρουσιάστηκαν νωρίτερα:

- Η μέση συχνότητα ομιλίας των γυναικών είναι αυξημένη κατά 108.516Hz σε σύγκριση με τους άνδρες για την ίδια συνθήκη.
- Η μέση συχνότητα ομιλίας για την ανεπίσημη συνθήκη είναι αυξημένη κατά 19.721Hz σε σύγκριση με την ευγενική συνθήκη για το ίδιο φύλο.

Διαφορετικά χρησιμοποιώντας τα διαστήματα εμπιστοσύνης καταλήγουμε στα εξής συμπεράσματα:

- Η συχνότητα ομιλίας για το γυναικείο φύλο στην ανεπίσημη συνθήκη κυμαίνεται από 226.085Hz έως 287.608Hz το 95% των περιπτώσεων.

- Η συχνότητα ομιλίας για την ευγενική συνθήκη είναι μειώμενη μεταξύ 30.741Hz και 8.701Hz σε σύγκριση με την ανεπίσημη συνθήκη για το ίδιο φύλο το 95% των περιπτώσεων.

- Η συχνότητα ομιλίας για τους άνδρες είναι μειώμενη μεταξύ 149.34Hz και 67.683Hz σε σύγκριση με τις γυναίκες για την ίδια συνθήκη το 95% των περιπτώσεων.

Η διακύμανση των τιμών των υποκειμένων ισούται με 615.6 και είναι κάτι παραπάνω από διπλάσια από αυτήν για τις τιμές των σεναρίων που ισούται με 219.5. Η συνολική διακύμανση είναι 1481, με αρκετά μεγάλη συμβολή να έχει η μεταβλητότητα μεταξύ των διαφόρων υποκειμένων, η οποία είναι περίπου 41.5% (615.6/1481).

Τέλος η τυπική απόκλιση των σεναρίων και των υποκειμένων ισούται με 14.81Hz, 24.81Hz αντίστοιχα, με τις δυο τυχαίες επιδράσεις να είναι ασυσχέτιστες μεταξύ τους. Επομένως η τυπική απόκλιση του αθροισμάτος τους ισούται με: 28.894Hz ($\sqrt{14.81^2 + 24.81^2}$), με την ανεπίσημη συνθήκη για το γυναικείο φύλο να έχει μέσο όρο σταθεράς 256.846Hz. Μπορούμε να καταλήξουμε δηλαδή, ότι οι γυναίκες στην ανεπίσημη συνθήκη, θα έχουν προσωπικές σταθερές που θα είναι έως 28.894Hz χαμηλότερες ή υψηλότερες από τον μέσο όρο αυτό στο 68% των περιπτώσεων και μέχρι 57.788Hz χαμηλότερες ή υψηλότερες στο 95% των περιπτώσεων. Αντίστοιχα με τον ίδιο τρόπο καταλήγουμε σε ανάλογα συμπεράσματα για οποιοδήποτε συνδυασμό των επιπέδων των δυο αυτών σταθερών επιδράσεων.

3.3.2 Όλοι οι πιθανοί τρόποι προσαρμογής Τυχαίων Επιδράσεων στην R

Στις Ενότητες 3.2, 3.3 παρουσιάστηκαν δυο διαφορετικοί τρόποι προσαρμογής τυχαίων επιδράσεων με την βοήθεια της R. Η πρώτη περίπτωση η οποία είναι και η απλούστερη, αφορά μία επίδραση, έστω A, η οποία θεωρείται τυχαία σταθερά με το όρισμα να είναι το εξής: (1 | A). Τότε για κάθε επίπεδο της επίδρασης λαμβάνουμε μια ξεχωριστή εκτίμηση.

Η δεύτερη περίπτωση είναι πιο σύνθετη καθώς εμπλέκει δυο διαφορετικές επιδράσεις. Χρησιμοποιείται στην περίπτωση που θεωρούμε πως κάθε επίπεδο της επίδρασης A, επηρεάζεται διαφορετικά από την επίδραση B, με το όρισμα να είναι το εξής: (1 + B | A). Στην περίπτωση που η μεταβλητή B είναι ποσοτική, τότε έχουμε μία τυχαία σταθερά και μία τυχαία κλίση, με την τυχαία κλίση να αντιπροσωπεύει την ποσοτική επίδραση B. Για κάθε επίπεδο της επίδρασης A, λαμβάνουμε εκτιμήσεις τόσο για την τυχαία σταθερά όσο και για την τυχαία κλίση. Η περίπτωση αυτή παρουσιάστηκε στην Ενότητα 3.2.2. Αντίθετα αν η μεταβλητή B είναι κατηγορική, τότε έχουμε μία τυχαία σταθερά και $p-1$ τυχαίες κλίσεις, όπου p ο αριθμός των επιπέδων της επίδρασης B. Αντίστοιχα για κάθε επίπεδο της επίδρασης A, έχουμε εκτίμηση της τυχαίας σταθεράς και των $p-1$ τυχαίων κλίσεων, με κάθε τυχαία κλίση να ταυτίζεται με μία από τις υποομάδες της επίδρασης B, εξαιρώντας την ομάδα αναφοράς. Τέλος αξίζει να σημειωθεί ότι η τυχαία σταθερά και στις δυο περιπτώσεις, αντιπροσωπεύει την επίδραση του δεξιού μέλους, δηλαδή την επίδραση A.

Χρησιμοποιώντας το Παράδειγμα 3.3, θα προσαρμόσουμε ένα μοντέλο το οποίο θα περιέχει τους δυο αυτούς τρόπους προσαρμογής των τυχαίων επιδράσεων, παρουσιάζοντας συνοπτικά τα κυριότερα σημεία που επιθυμούμε να τονίσουμε. Αξίζει να σημειωθεί ότι στην περίπτωση μας η επίδραση B είναι η συμπεριφορά, η οποία είναι μια κατηγορική επίδραση με δυο υποομάδες.

```
> model4<-lmer(frequency~attitude+gender+(1|subject)+
+ (1+attitude|scenario),data=politeness)

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 scenario (Intercept) 203.49   14.265
          attitudepol  71.02    8.427   0.00
 subject  (Intercept) 616.51   24.830
 Residual             627.81   25.056

$scenario
  (Intercept) attitudepol
1  -13.657911   0.7432393
2    3.810621   6.8764573
3   12.780617  -3.6926454

$subject
   (Intercept)
F1  -13.936316
F2   10.440585
F3    3.495731
```

Το μοντέλο που έχουμε προσαρμόσει με τη βοήθεια της συνάρτησης lmer είναι το εξής:

$$y_i = \alpha_0 + \alpha_1 I(attitude = pol) + \alpha_2 I(gender = male) + \beta_{1,j[i]} + \beta_{2,k[i]} \\ + \beta_{3,k[i]} I(attitude = pol) + e_i, i = 1, 2, ..., 84,$$

όπου το y_i δηλώνει την τιμή της συχνότητας ομιλίας για την i παρατήρηση, τα α_0, α_1, α_2 αποτελούν τις σταθερές σταθερές, ενώ τα $\beta_{1,j[i]}$, $\beta_{2,k[i]}$ αποτελούν τις τυχαίες σταθερές και το $\beta_{3,k[i]}$ αποτελεί την τυχαία κλίση. Το $j[i]$ δηλώνει το επίπεδο της ομάδας $(F_1, F_2, F_3, M_1, M_2, M_3)$ για το υποκείμενο i και το $k[i]$ δηλώνει το επίπεδο της ομάδας του σεναρίου (1,2,...,7) για το υποκείμενο i.

Συνοπτικά θα λέγαμε ότι λαμβάνουμε μία τυχαία σταθερά για την επίδραση των υποκειμένων και για κάθε επίπεδο της μία εκτίμηση. Καθώς η επίδραση της συμπεριφοράς είναι μια κατηγορική επίδραση με δυο υποομάδες, έχουμε μία τυχαία σταθερά και μία τυχαία κλίση. Επομένως για κάθε σενάριο έχουμε εκτιμήσεις για την τυχαία σταθερά και την τυχαία κλίση, με την τυχαία σταθερά να αντιπροσωπεύει την επίδραση των σεναρίων, ενώ την τυχαία κλίση να αντιπροσωπεύει την επίσημη συμπεριφορά. Στις περιπτώσεις αυτές χρήζει προσοχής η ερμηνεία της τιμής της διακύμανσης της τυχαίας κλίσης. Η τιμή της διακύμανσης της τυχαίας κλίσης, δηλώνει την διακύμανση των παρατηρήσεων της επίσημης συμπεριφοράς εντός των σενάριων και ισούται με 71.02.

Η τελευταία περίπτωση δεν είναι η μοναδική όταν επιθυμούμε να θεωρήσουμε τις επιδράσεις δυο μεταβλητών ως τυχαίες, οι οποίες συνδέονται μεταξύ τους. Σε αντίθεση με την προηγούμενη περίπτωση, στην συγκεκριμένη έχουμε μόνο τυχαίες κλίσεις. Δηλαδή η επίδραση του δεξιού μέλους, η οποία ήταν η τυχαία σταθερά απουσιάζει. Για να πραγματοποιηθεί κάτι τέτοιο, το όρισμα που πρέπει να εισάγουμε είναι το εξής: (0 + B | A). Γενικότερα στην περίπτωση που η μεταβλητή B είναι διακριτή, τότε έχουμε μία τυχαία κλίση που αντιπροσωπεύει την επίδραση B και μία εκτίμηση της για κάθε επίπεδο της επίδρασης A. Αντίθετα αν η μεταβλητή B είναι κατηγορική, τότε έχουμε p τυχαίες κλίσεις, όπου p ο αριθμός των υποομάδων της επίδρασης B, με κάθε τυχαία κλίση να ταυτίζεται με μία από τις υποομάδες της επίδρασης B. Δηλαδή για κάθε επίπεδο της επίδρασης A, έχουμε p ξεχωριστές εκτιμήσεις. Η χρήση του παραπάνω ορίσματος στο Παράδειγμα 3.3 εξάγει τα εξής αποτελέσματα:

```
> model5<-lmer(frequency~attitude+gender+(1|subject)+
+ (0+attitude|scenario),data=politeness)

Random effects:
 Groups   Name         Variance Std.Dev. Corr
 scenario attitudeinf 203.5    14.27
          attitudepol 275.4    16.59    0.86
 subject  (Intercept) 616.5    24.83
 Residual             627.8    25.06

$scenario
  attitudeinf attitudepol
1  -13.658245  -12.914386
2    3.810137   10.687341
3   12.781172    9.087555
```

Το μοντέλο που έχουμε προσαρμόσει με τη βοήθεια της συνάρτησης lmer είναι το εξής:

$$y_i = \alpha_0 + \alpha_1 I(attitude = pol) + \alpha_2 I(gender = male) + \beta_{1,j[i]} + \beta_{2,k[i]} I(attitude = inf)$$
$$+ \beta_{3,k[i]} I(attitude = pol) + e_i, i = 1, 2, ..., 84,$$

όπου το y_i δηλώνει την τιμή της συχνότητας ομιλίας για την i παρατήρηση, τα α_0, α_1, α_2 αποτελούν τις σταθερές σταθερές, ενώ το $\beta_{1,j[i]}$ αποτελεί την τυχαία σταθερά και τα $\beta_{2,k[i]}$, $\beta_{3,k[i]}$ αποτελούν τις τυχαίες κλίσεις. Τέλος το $j[i]$ δηλώνει το επίπεδο της ομάδας $(F_1, F_2, F_3, M_1, M_2, M_3)$ για το υποκείμενο i και το $k[i]$ δηλώνει το επίπεδο της ομάδας του σεναρίου (1,2,...,7) για το υποκείμενο i.

Συνοπτικά θα λέγαμε ότι έχουμε δυο τυχαίες κλίσεις, όσες και ο αριθμός των υποομάδων της συμπεριφοράς, με την πρώτη να αντιπροσωπεύει την ανεπίσημη ενώ την δεύτερη την ευγενική. Για κάθε σενάριο λαμβάνουμε εκτιμήσεις για τις δυο αυτές τυχαίες κλίσεις. Αντίστοιχα με την προηγούμενη περίπτωση οι τιμές της διακύμανσης των τυχαίων κλίσεων, δηλώνουν την διακύμανση των παρατηρήσεων της ανεπίσημης και επίσημης συμπεριφοράς εντός των σεναρίων.

Όπως υπάρχει η δυνατότητα να θεωρήσουμε μία αλληλεπίδραση ως σταθερή επίδραση, αντίστοιχα μπορούμε να θεωρήσουμε και μία αλληλεπίδραση ως τυχαία. Το όρισμα που πρέπει να εισάγουμε για να πραγματοποιηθεί αυτό είναι το εξής: (1 | A:B). Χρησιμοποιώντας το όρισμα αυτό, λαμβάνουμε μία τυχαία σταθερά και έχουμε ξεχωριστή εκτίμηση για κάθε συνδυασμό των επιπέδων των δυο αυτών επιδράσεων. Η χρήση του παραπάνω ορίσματος στο Παράδειγμα 3.3 εξάγει τα εξής αποτελέσματα:

```
> model6<-lmer(frequency~attitude+gender+(1|subject)+
+ (1|scenario:attitude),data=politeness)

Random effects:
 Groups            Name        Variance Std.Dev.
 scenario:attitude (Intercept) 238.5    15.44
 subject           (Intercept) 616.9    24.84
 Residual                      628.2    25.06

$`scenario:attitude`
      (Intercept)
1:inf  -15.164085
1:pol   -8.214203
2:inf   -3.037781
2:pol   14.428515
3:inf   17.265382
3:pol    2.475939
```

Το μοντέλο που έχουμε προσαρμόσει με τη βοήθεια της συνάρτησης lmer είναι το εξής:

$$y_i = \alpha_0 + \alpha_1 I(attitude = pol) + \alpha_2 I(gender = male) + \beta_{1,j[i]} + (\beta_2\beta_3)_{k[i]l[i]} + e_i, i = 1, ..., 84,$$

όπου το y_i δηλώνει την τιμή της συχνότητας ομιλίας για την i παρατήρηση, τα α_0, α_1, α_2 αποτελούν τις σταθερές σταθερές, ενώ τα $\beta_{1,j[i]}$, $(\beta_2\beta_3)_{k[i]l[i]}$ αποτελούν τις τυχαίες σταθερές. Τέλος το $j[i]$ δηλώνει το επίπεδο της ομάδας $(F_1, F_2, F_3, M_1, M_2, M_3)$ για το υποκείμενο i, το $k[i]$ δηλώνει το επίπεδο της ομάδας του σεναρίου (1,2,...,7) για το υποκείμενο i και το $l[i]$ δηλώνει το επίπεδο της ομάδας της συμπεριφοράς (inf,pol) για το υποκείμενο i.

Συνοπτικά μπορούμε να πούμε ότι λαμβάνουμε μία τυχαία σταθερά για την αλληλεπίδραση συμπεριφορά-σενάριο και για κάθε συνδυασμό των επιπέδων τους, γίνεται εκτίμηση των τιμών τους. Η τιμή της διακύμανσης της αλληλεπίδρασης, δηλώνει την διακύμανση των παρατηρήσεων για όλους τους πιθανούς συνδυασμούς των επιπέδων των μεταβλητών συμπεριφορά και σενάριο.

Οι παραπάνω είναι όλοι οι πιθανοί τρόποι προσαρμογής τυχαίων επιδράσεων με την βοήθεια της R.

3.4 Χειρισμός των Φωλιασμένων Δεδομένων

Στην αρχή του Κεφαλαίου 3, επισημάνθηκε ότι η προσέγγιση για ένα μικτό γραμμικό μοντέλο με την βοήθεια της R για οποιαδήποτε δομή δεδομένων είναι κατά μεγάλο ποσοστό η ίδια. Μια μικρή διαφοροποίηση στην προσέγγιση και πιο συγκεκριμένα στα ορίσματα της συνάρτησης lmer, ενδέχεται να προκύψουν στην περίπτωση που μελετάμε φωλιασμένα δεδομένα. Προτού γίνουμε πιο συγκεκριμένοι, θα καταγράψουμε εν συντομία ποιά ακριβώς είναι η δομή των φωλιασμένων δεδομένων με την χρήση ενός απλού παραδείγματος.

Το παράδειγμα μας αφορά σχολικές τάξεις σε σχολεία. Υποθέτουμε ότι έχουμε τρία σχολεία και εννιά σχολικές τάξεις. Τότε αν οι τρείς πρώτες τάξεις ανήκουν στο σχολείο ένα, οι επόμενες τρείς στο σχολείο δυο και οι τελευταίες τρείς στο σχολείο τρία, έχουμε δομή φωλιασμένων δεδομένων. Αντίστοιχα μπορούμε να έχουμε ασθενείς που παρακολουθούνται από γιατρούς ή ασθενείς που νοσηλεύονται σε νοσοκομεία. Δηλαδή θέλουμε ένα συγκεκριμένο αριθμό ασθενών να παρακολουθούνται για παράδειγμα μόνο από τον γιατρό ένα, χωρίς αυτοί οι ασθενείς να έχουν ανάμειξη με οποιοδήποτε άλλον γιατρό, με το ίδιο να αναμένεται και για τους υπόλοιπους ασθενείς.

Ενδιαφέρον επίσης παρουσιάζει ο συμβολισμός τους, καθώς ενδέχεται να είναι η αιτία της διαφοροποίησης των ορισμάτων. Ανατρέχουμε και πάλι στο παράδειγμα με τα σχολεία και τις σχολικές τάξεις, για το οποίο έχουμε δυο διαφορετικές επιλογές. Ξεκινώντας με τα σχολεία, η επιλογή του συμβολισμού είναι κοινή και είναι η εξής: Σ1, Σ2, Σ3. Η διαφοροποίηση έχει να κάνει με τον συμβολισμό των τάξεων. Στην πρώτη περίπτωση ο συμβολισμός τους είναι: Τ1, Τ2, ..., Τ8, Τ9. Αντίθετα στην δεύτερη περίπτωση ο συμβολισμός τους είναι: Τ1, Τ2, Τ3, Τ1, Τ2, Τ3, Τ1, Τ2, Τ3, με την πρώτη τριάδα να ανήκει στο Σ1, την δεύτερη στο Σ2 κ.ο.κ.

Εύκολα κάποιος μπορεί να ισχυριστεί ότι ο συμβολισμός των δεδομένων αποτελεί μια ιδιότητα τους και δεν σχετίζεται με την προσαρμογή του μικτού γραμμικού μοντέλου. Ωστόσο στην συνέχεια θα αποδείξουμε ότι ο ισχυρισμός αυτός είναι λανθασμένος.

Υποθέτουμε ότι έχουμε ένα παράδειγμα στο οποίο θέλουμε η επίδραση των τάξεων να αποτελεί την τυχαία μας σταθερά, με το όρισμα που πρέπει να εισάγουμε να είναι γνωστό από τις προηγούμενες ενότητες και να είναι το εξής: (1 | Class). Για την πρώτη περίπτωση συμβολισμού των τάξεων δεν αντιμετωπίζουμε κανένα πρόβλημα, καθώς έχουμε διαφορετικό συμβολισμό για κάθε μία από τις εννιά τάξεις, επομένως θα έχουμε εννιά διαφορετικές εκτιμήσεις.

Αντίθετα γίνεται εύκολα αντιληπτό ότι στην δεύτερη περίπτωση το όρισμα αυτό θα δημιουργήσει προβλήματα, καθώς έχουμε ίδιους συμβολισμούς για τις τρείς τάξεις κάθε σχολείου. Πιο συγκεκριμένα αντί για εννιά διαφορετικές εκτιμήσεις για τις σχολικές τάξεις, λόγω του συμβολισμού αυτού θα λάβουμε μόνο τρείς (Τ1, Τ2, Τ3), το οποίο προφανώς και δεν είναι σωστό. Για να αντιμετωπιστεί αυτό το πρόβλημα και να έχουμε ξεχωριστή εκτίμηση για κάθε σχολική ταξή, είμαστε αναγκασμένοι να χρησιμοποιήσουμε το όρισμα της αλληλεπίδρασης το οποίο είναι το εξής: (1 | school:class). Κατά αυτόν τον τρόπο θα λάβουμε εκτίμηση για κάθε συνδυασμό σχολείου-σχολική τάξη που διαθέτουμε στα δεδομένα μας.

Παραμένουμε στην ίδια περίπτωση συμβολισμού, υποθέτοντας ότι θέλουμε να προσθέσουμε και την επίδραση των σχολείων ως μία τυχαία σταθερά. Τότε με βάση όσα έχουμε ήδη αναφέρει, τα όρισματα που πρέπει να εισάγουμε είναι τα εξής: (1 | school:class), (1 | school). Αντί του αθροίσματος των παραπάνω ορισμάτων, υπάρχει ένα καινούργιο, ισοδύναμο όρισμα το οποίο χρησιμοποιείται στις περιπτώσεις αυτές και είναι αυστηρά για φωλιασμένα δεδομένα. Το νέο όρισμα είναι το εξής: (1 | school/class), με τα αποτελέσματα που λαμβάνουμε να είναι ακριβώς τα ίδια και για τις δυο περιπτώσεις. Γενικότερα το όρισμα αυτό, δηλώνει ότι η μεταβλητή του δεξιού μέλους της διαγωνίου είναι φωλιασμένη στην μεταβλητή του αριστερού μέλους, δηλαδή στην περίπτωση μας ότι οι σχολικές τάξεις είναι φωλιασμένες στα σχολεία. Τέλος κατά την χρήση του λαμβάνουμε μια τυχαία σταθερά για την επίδραση του αριστερού μέλους, δηλαδή του σχολείου, όπως επίσης και μια τυχαία σταθερά για την επίδραση του δεξιού μέλους, δηλαδή των σχολικών τάξεων, με την ιδιαιτερότητα της δεύτερης τυχαίας σταθεράς να αποτελεί ότι η μεταβλητή αυτή είναι φωλιασμένη σε αυτήν του αριστερού μέλους, καταλήγοντας τελικά σε μια τυχαία αλληλεπίδραση.

Συμπερασματικά θα λέγαμε ότι η αντιμετωπίση των φωλιασμένων δεδομένων εξαρτάται από την ονομασία των στοιχείων των μεταβλητών. Στην περίπτωση που έχουμε κοινές ονομασίες, τότε είμαστε αναγκασμένοι να προσφύγουμε στις λύσεις που υποδείχθηκαν νωρίτερα. Σε αντίθετη περίπτωση εφαρμόζουμε όσα παρουσιάστηκαν στις προηγούμενες ενότητες. Ωστόσο πάντα υπάρχει η λύση της μετανομασίας των στοιχείων των μεταβλητών, ώστε κάθε ένα να έχει μοναδική ονομασία αποφεύγοντας τελικά τέτοιου είδους προβλήματα.

3.5 Σύγκριση Διαφορετικών Μοντέλων

Στην ενότητα αυτή θα συγκεντρώσουμε τις διαφορετικές περιπτώσεις που ενδέχεται να συναντήσουμε όταν επιθυμούμε να συγκρίνουμε δυο ή περισσότερα μικτά γραμμικά μοντέλα. Ο χειρισμός τους απαιτεί μεγάλη προσοχή, καθώς εύκολα μπορούμε να καταλήξουμε σε εσφαλμένα αποτελέσματα. Υπενθυμίζουμε ότι κατά την προσαρμογή των μοντέλων μας με την βοήθεια της συνάρτησης lmer, χρησιμοποιήθηκε η μέθοδος REML για τους λόγους που έχουμε καταγράψει στο Κεφάλαιο 2.

Η πρώτη περίπτωση αφορά την σύγκριση μοντέλων που διαφέρουν μόνο ως προς τις τυχαίες τους επιδράσεις, δηλαδή η διαφορά των δυο μοντέλων είναι ότι το ένα περιέχει μία ή περισσότερες τυχαίες επιδράσεις σε σύγκριση με το άλλο. Η επιλογή που προτιμάται σε αυτήν την περίπτωση είναι η χρήση των κριτηρίων AIC, BIC για τα δυο "ανταγωνιστικά" μοντέλα. Το μοντέλο με την μικρότερη τιμή είναι και αυτό που προτιμάται. Αξίζει να σημειωθεί ότι κατά την χρήση των συναρτήσεων AIC, BIC η μέθοδος που χρησιμοποιείται για τον υπολογισμό τους, ταυτίζεται με αυτήν που έχουμε χρησιμοποιήσει στην προσαρμογή του μοντέλου, δηλαδή την μέθοδο REML.

Η δεύτερη περίπτωση αφορά την σύγκριση μοντέλων που διαφέρουν αυτήν την φορά μόνο ως προς τις σταθερές τους επιδράσεις, δηλαδή η διαφορά των δυο μοντέλων είναι ότι το ένα περιέχει μία ή περισσότερες σταθερές επιδράσεις σε σύγκριση με το άλλο. Η πιο συνήθης επιλογή είναι η χρήση της συνάρτησης anova για τα δυο "ανταγωνιστικά" μοντέλα, με την οποία λαμβάνουμε αποτελέσματα για τον έλεγχο του λόγου των πιθανοφανειών και τα κριτήρια AIC, BIC. Ωστόσο στο Κεφάλαιο 2 επισημάναμε ότι ο έλεγχος του λόγου των πιθανοφανείων για να πραγματοποιηθεί με την βοήθεια της μεθόδου REML, απαιτεί τις ίδιες ακριβώς σταθερές επιδράσεις. Η συνάρτηση anova τώρα, έχει ως προεπιλογή την μέθοδο ML, οπότε δεν χρειάζεται να προσθέσουμε κάποιο επιπλέον όρισμα. Τέλος οι τιμές των κριτηρίων AIC, BIC προφανώς υπολογίζονται και αυτές με βάση την μέθοδο ML.

Οι δυο παραπάνω περιπτώσεις είναι ειδικές, προϋποθέτοντας η αποκλειστική διαφορά των μοντέλων να είναι μία ή περισσότερες σταθερές ή τυχαίες επιδράσεις σε σύγκριση με το άλλο. Σε μία πιο γενική περίπτωση που δεν είμαστε βέβαιοι για τον τρόπο προσαρμογής των τυχαίων επιδράσεων, τότε χρησιμοποιούνται και πάλι τα κριτήρια AIC, BIC για τα επιλεγμένα μοντέλα. Καθώς υπάρχει περίπτωση να καταλήξουμε σε ισοδύναμες τιμές, η επιλογή του μοντέλου εξαρτάται από την πολυπλοκότητα του. Δηλαδή εφόσον οι τιμές των κριτηρίων είναι ίσες δεν έχει νόημα να επιλέξουμε το πιο σύνθετο, καταλήγοντας στο πιο απλό μοντέλο. Αντίστοιχα στην περίπτωση που επιθυμούμε άμεση σύγκριση για δυο μοντέλα που διαφέρουν και ως προς τις σταθερές και ως προς τις τυχαίες τους επιδράσεις, τότε η επιλογή είναι και πάλι τα κριτήρια AIC, BIC.

3.6 Συνάρτηση lme

Η συνάρτηση lmer είναι η πιο βασική και διαδεδομένη συνάρτηση που χρησιμοποιείται κατά την προσαρμογή ενός μικτού γραμμικού μοντέλου, χωρίς ωστόσο να είναι η μοναδική. Η εναλλακτική επιλογή είναι η χρήση της συνάρτησης lme, του πακέτου nlme της R. Μικρές διαφορές μεταξύ των δυο συναρτήσεων εντοπίζονται στα ορίσματα τους και στα αποτελέσματα που δίνει η σύνοψη των μοντέλων.

Για την παρουσίαση της νέας αυτής συνάρτησης, θα χρησιμοποιηθούν τα δεδομένα sleepstudy του Κεφαλαίου 3.2. Στην συνέχεια θα καταγράψουμε πως προσαρμόζεται ένα μοντέλο τυχαίας σταθεράς, παρουσιάζοντας επίσης τα αποτελέσματα που μας δίνει η συνάρτηση summary.

```
> model7<-lme(Reaction~Days,random=~1|Subject,data=sleepstudy)
> summary(model7)

Linear mixed-effects model fit by REML
  Data: sleepstudy
       AIC      BIC    logLik
  1794.465 1807.192 -893.2325

Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:    37.12383 30.99123

Fixed effects:  Reaction ~ Days
                Value Std.Error  DF  t-value p-value
(Intercept) 251.40510  9.746716 161 25.79383       0
Days         10.46729  0.804221 161 13.01543       0
```

```
 Correlation:
     (Intr)
Days -0.371

Standardized Within-Group Residuals:
       Min          Q1         Med          Q3         Max
-3.2256707 -0.5528788  0.0108521  0.5187971  4.2506162

Number of Observations: 180
Number of Groups: 18
```

Έχοντας ως αφετηρία τον ορισμό του μοντέλου, παρατηρούμε ότι δεν έχει αλλάξει κάτι για την προσαρμογή των σταθερών επιδράσεων, με την διαφοροποίηση να αφορά μόνο τις τυχαίες επιδράσεις. Η συνάρτηση lme απαιτεί ένα καινούργιο όρισμα, το random, στο οποίο δηλώνονται ξεχωριστά οι τυχαίες επιδράσεις του μοντέλου. Στο παράδειγμα μας υποδείξαμε τον τρόπο προσθήκης μιας τυχαίας σταθεράς στο μοντέλο μας, η οποία στην προκειμένη περίπτωση είναι η Subject. Γενικότερα για να προσθέσουμε μία τυχαία σταθερά (έστω A) χρησιμοποιώντας την συνάρτηση lme, πρέπει το όρισμα random να ισούται με: $\sim 1 \mid A$.

Για την σύνοψη του μοντέλου θα επισημάνουμε μόνο τις διαφορές που εντοπίζονται μεταξύ των δυο συναρτήσεων. Παρατηρούμε ότι η συνάρτηση lme εξάγει τις τιμές των κριτηρίων AIC, BIC σε αντίθεση με την συνάρτηση lmer. Ωστόσο η σημαντικότερη διαφορά έχει να κάνει με τις p-τιμές, τις οποίες για λόγους που αναλύσαμε νωρίτερα η συνάρτηση lmer δεν υπολογίζει. Αντίθετα η συνάρτηση lme εκτός από τις τιμές των στατιστικών ελέγχων, παραθέτει τους βαθμούς ελευθερίας και τις p-τιμές του, δίνοντας την ευκαιρία να προβούμε σε συμπεράσματα για την σημαντικότητα τους χωρίς την χρήση άλλης συνάρτησης.

Αξίζει να σημειωθεί ότι δεν εντοπίζονται διαφορές στις τιμές και στα τυπικά σφάλματα των σταθερών επιδράσεων ή αντίστοιχα στις τιμές και στην τυπική απόκλιση των τυχαίων επιδράσεων. Αντίστοιχα το ίδιο ισχύει και για τα κριτήρια AIC, BIC καθώς και την τιμή της συνάρτησης πιθανοφάνειας.

Η προεπιλογή της R και για αυτήν την συνάρτηση είναι η μέθοδος REML. Οι συναρτήσεις για την εκτίμηση των τυχαίων επιδράσεων, ranef και coef, μπορούν να χρησιμοποιηθούν και πάλι δίνοντας αντίστοιχα αποτελέσματα. Η διαφοροποίηση για τις συμπληρωματικές συναρτήσεις συναντάται κατά την κατασκευή ενός διαστήματος εμπιστοσύνης, για το οποίο χρησιμοποιείται η συνάρτηση intervals αντί της συνάρτησης confint, έχοντας ως όρισμα το όνομα του μοντέλου.

Τα αντίστοιχα ορίσματα $(1 + A \mid B)$, $(0 + A \mid B)$ για την συνάρτηση lme είναι τα εξής:

- random=$\sim 1 + A \mid B$.
- random=$\sim 0 + A \mid B$.

Το πλεονέκτημα της συνάρτησης lme σε σύγκριση με την συνάρτηση lmer, είναι ο υπολογισμός των p-τιμών για τον στατιστικό έλεγχο σημαντικότητας των σταθερών επιδράσεων. Αντίθετα ένα από τα μειονεκτήματα της συνάρτησης lme, είναι ότι δεν μπορεί να προσαρμόσει ένα μοντέλο με ελλειπή δεδομένα. Επίσης μπορεί τα ορίσματα για την εισαγωγή των τυχαίων επιδράσεων που καταγράψαμε νωρίτερα να είναι απλά, ωστόσο ο συνδυασμός τους σε ένα μοντέλο, έστω και δυο τυχαίων σταθερών, προϋποθέτει την συγγραφή ενός σύνθετου κώδικα, το οποίο προφανώς και δεν είναι το ζητούμενο.

Καταλήγοντας η επιλογή και η χρήση της συνάρτησης lme μπορεί να προτιμηθεί έναντι της lmer, στην περίπτωση που έχουμε πλήρη δεδομένα και ένα απλό μοντέλο που χρειάζεται μόνο ένα από τα παραπάνω ορίσματα για να προσαρμοστούν οι τυχαίες του επιδράσεις.

3.7 Βιβλιοθήκες και Συναρτήσεις της R

Στην υποενότητα αυτή θα συγκεντρώσουμε τις βιβλιοθήκες, καθώς και τις συναρτήσεις που χρησιμοποιήθηκαν καθόλη την διάρκεια της ανάλυσης των μικτών γραμμικών μοντέλων με την βοήθεια της R. Για τις συναρτήσεις θα αναφερθούν εν συντομία τα ορίσματα τους καθώς και η χρησιμότητα τους. Οι βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι εξής:

- **lme4**: Για την χρήση των συναρτήσεων lmer, confint, ranef, coef.
- **afex**: Για την χρήση της συνάρτησης mixed.
- **nlme**: Για την χρήση των συναρτήσεων lme, intervals.
- **merTools**: Για την χρήση των βοηθητικών συναρτήσεων predict, preditInterval.

Οι συναρτήσεις που χρησιμοποιήθηκαν είναι οι εξής:

- **lmer** (formula,data,REML)

Η formula που εισάγεται διαχωρίζεται με το σύμβολο $\sim$, έχοντας στο αριστερό της μέλος την εξαρτημένη μεταβλητή και στο δεξί της μέλος τις σταθερές και τυχαίες επιδράσεις που επιθυμούμε να προσθέσουμε στο μοντέλο μας. Επίσης εισάγεται το πλαίσιο δεδομένων και εξαιτίας του γεγονότος πως η προεπιλογή της συνάρτησης είναι η μέθοδος REML, αν επιθυμούμε να χρησιμοποιήσουμε την μέθοδο ML, τότε το όρισμα REML λαμβάνει την λογική τιμή FALSE. Τέλος η συνάρτηση χρησιμοποιείται για την προσαρμογή ενός μικτού γραμμικού μοντέλου.

- **summary** (όνομα του μοντέλου)

Το όρισμα που δέχεται η συνάρτηση summary είναι το όνομα του μοντέλου. Χρησιμοποιώντας την λαμβάνουμε πληροφορίες για τα υπόλοιπα, την συνάρτηση πιθανοφάνειας και τις σταθερές-τυχαίες επιδράσεις.

- **ranef** (όνομα του μοντέλου)

Το όρισμα που λαμβάνει η συνάρτηση ranef είναι το όνομα του μοντέλου και εξάγει τις τιμές των τυχαίων επιδράσεων.

- **coef** (όνομα του μοντέλου)

Το όρισμα που λαμβάνει η συνάρτηση coef είναι το όνομα του μοντέλου και εξάγει το άθροισμα των τιμών των σταθερών και τυχαίων επιδράσεων (ξεχωριστά σταθερών και κλίσεων).

- **confint** (όνομα του μοντέλου)

Το όρισμα που λαμβάνει η συνάρτηση confint είναι το όνομα του μοντέλου και κατασκευάζει ένα διάστημα εμπιστοσύνης 95% για τους συντελεστές των σταθερών επιδράσεων και για τις τυπικές αποκλίσεις των τυχαίων επιδράσεων.

- **AIC, BIC** (όνομα του μοντέλου)

Το όρισμα που λαμβάνουν οι συναρτήσεις AIC, BIC είναι το όνομα του μοντέλου και εξάγουν τις τιμές των κριτηρίων αυτών.

- **anova** (model1,model2,refit)

Τα ορίσματα που λαμβάνει η συνάρτηση anova είναι τα ονόματα των δυο μοντέλων ως προς τα οποία θέλουμε να πραγματοποιηθεί ο έλεγχος του λόγου των πιθανοφανειών με την χρήση της μεθόδου ML. Καθώς η προεπιλογή της συνάρτησης είναι η μέθοδος ML, αν επιθυμούμε να χρησιμοποιήσουμε την μέθοδο REML, τότε το όρισμα refit λαμβάνει την λογική τιμή FALSE. Εξάγει την τιμή του ελέγχου του λόγου των πιθανοφανειών, τους βαθμούς ελευθερίας του και την p-τιμή του. Τέλος υπολογίζει και τις τιμές των κριτηρίων AIC, BIC.

- **mixed** (όνομα του μοντέλου,data)

Τα ορίσματα που εισάγονται στην συνάρτηση mixed είναι το όνομα του μοντέλου (ή αλλιώς ο τύπος του) και το πλαίσιο δεδομένων. Λαμβάνουμε αποτελέσματα για τις p-τιμές του στατιστικού ελέγχου σημαντικότητας των σταθερών επιδράσεων, για ένα μοντέλο που έχει προσαρμοστεί με την βοήθεια της συνάρτησης lmer, χρησιμοποιώντας την προσέγγιση των Kenward-Rogers για τους βαθμούς ελευθερίας.

- **lme** (formula,random,data)

Η formula που εισάγεται διαχωρίζεται με το σύμβολο $\sim$, έχοντας στο αριστερό της μέλος την εξαρτημένη μεταβλητή και στο δεξί της μέλος τις σταθερές επιδράσεις. Οι τυχαίες επιδράσεις εισάγονται με την βοήθεια του ορίσματος random και επίσης καλούμαστε να εισάγουμε και το πλαίσιο δεδομένων μας. Τέλος η συνάρτηση χρησιμοποιείται για την προσαρμογή ενός μικτού γραμμικού μοντέλου χωρίς ελλειπή δεδομένα.

- **intervals** (όνομα του μοντέλου)

Το όρισμα που λαμβάνει η συνάρτηση intervals είναι το όνομα του μοντέλου και χρησιμοποιείται για την κατασκευή ενός διαστήματος εμπιστοσύνης 95% για τους συντελεστές των σταθερών επιδράσεων και για τις τυπικές αποκλίσεις των τυχαίων επιδράσεων, στην περίπτωση που το μοντέλο μας έχει προσαρμοστεί με την βοήθεια της συνάρτησης lme.

- **boxplot** (formula,col,data,main)

Η formula που εισάγεται διαχωρίζεται με το σύμβολο $\sim$, έχοντας στο αριστέρο της μέλος το όνομα της εξαρτημένης μεταβλητής και στο δεξί της μέλος την ανεξάρτητη μεταβλητή ως προς την οποία θα κατασκευαστεί το θηκοδιάγραμμα. Με τα ορίσματα col, main ορίζουμε το χρώμα και τον τίτλο του θηκοδιαγράμματος αντίστοιχα. Τέλος καλούμαστε να εισάγουμε και το πλαίσιο δεδομένων μας.

- **str** (όνομα του πλαισίου δεδομένων)

Το όρισμα που λαμβάνει η συνάρτηση str είναι το πλαίσιο δεδομένων και εξάγει μία λίστα, η οποία για κάθε μεταβλητή (ή αλλιώς στήλη των δεδομένων) περιέχει τον τύπο της, τα επίπεδα της και τις τιμές της.

Τέλος οι βοηθητικές συναρτήσεις που επισημάνθηκαν στην Ενότητα 3.2.1 θεωρήθηκε σκόπιμο να παραλειφθούν καθώς έχει ήδη γίνει αναλυτική αναφορά.

3.8 Επίλογος

Από την ανάγνωση του Κεφαλαίου 3, γίνεται κατανοητό ότι η σωστή χρήση των συναρτήσεων lmer ή lme, δηλαδή η προσαρμογή ενός μικτού γραμμικού μοντέλου με την βοήθεια της R, δεν εμφανίζει ιδιαίτερες δυσκολίες. Οι δυσκολίες που ερχόμαστε αντιμέτωποι σε πρώτη φάση, έχουν να κάνουν περισσότερο με τις σταθερές και τις τυχαίες επιδράσεις που θα συμπεριλάβουμε στο μοντέλο μας και τον τρόπο που θα μοντελοποιηθούν οι τυχαίες επιδράσεις. Αφού καταλήξουμε στο τελικό μοντέλο, οφείλουμε να αναφέρουμε ότι τα αποτελέσμα της σύνοψης του μοντέλου από μόνα τους, δεν μας λένε απολύτως τίποτα. Η μεγαλύτερη δυσκολία εμφανίζεται τώρα, καθώς θα πρέπει να είμαστε σε θέση να ερμηνεύσουμε τα αποτελέσματα αυτά για οποιοδήποτε μικτό γραμμικό μοντέλο. Η πλειοψηφία των συμπερασμάτων και ειδικότερα αυτά που προκύπτουν μέσω των σταθερών επιδράσεων, πρέπει να παρουσιαστούν με τρόπο που να είναι κατανοητά για οποιοδήποτε γνωρίζει το αρχικό μας πρόβλημα, χωρίς απαραίτητα να γνωρίζει τι είναι ένα μικτό γραμμικό μοντέλο.

Τα μοντέλα που παρουσιάστηκαν στο Κεφάλαιο 3 προσαρμόστηκαν χωρίς κανένα πρόβλημα. Ωστόσο η συνθήκη αυτή δεν συναντάται πάντα, καθώς ένα μικτό γραμμικό μοντέλο κατά την προσαρμογή του μπορεί να εμφανίσει προβλήματα σύγκλισης. Υπάρχουν αρκετές διαφορετικές παράμετροι που μπορούν να οδηγήσουν στο πρόβλημα αυτό. Σίγουρα μία από αυτές έχει να κάνει με τα δεδομένα, με την πιθανότητα εμφάνισης του προβλήματος σύγκλισης να αυξάνεται στην περίπτωση που έχουμε μικρό αριθμό παρατηρήσεων ή μη ισορροπημένα δεδομένα. Επίσης η αύξηση του αριθμού των τυχαίων επιδράσεων κάνει το μοντέλο μας πιο πολύπλοκο, επομένως αυξάνει τις πιθανότητες εμφάνισης του προβλήματος σύγκλισης. Οι παραπάνω είναι δυο από τους πολλούς λόγους εμφάνισης του προβλήματος σύγκλισης.

Το πρόβλημα σύγκλισης χρησιμοποιώντας την συνάρτηση lmer γίνεται αντιληπτό με την έξοδο του παρακάτω σφάλματος: "boundary (singular) fit: see help('isSingular')". Αξίζει να τονιστεί ότι παρόλο την έξοδο του σφάλματος αυτού, η R προσαρμόζει κανονικά το μοντέλο μας. Λόγω όμως των προβλημάτων σύγκλισης, αναμένουμε μη αξιόπιστες ή μη ακριβείς εκτιμήσεις των παραμέτρων του μοντέλου. Στην περίπτωση αυτή, προτείνεται η διαφοροποίηση του μοντέλου ώστε να ξεπεράστει το πρόβλημα της σύγκλισης.

Τέλος ένα άλλο ζήτημα που μας απασχόλεισαι στο Κεφάλαιο 3, ήταν οι p-τιμές των σταθερών επιδράσεων, με τις τιμές τους να εξάγονται για τις συναρτήσεις lme και mixed αλλά όχι για την συνάρτηση lmer. Η συνάρτηση lmer ακολουθεί μία πιο συντηρητική προσέγγιση και για αυτόν τον λόγο αποφεύγει να τις υπολογίσει. Αυτό ωστόσο δεν σημαίνει ότι οι p-τιμές των δυο άλλων συναρτήσεων είναι προσεγγιστικές και ελλοχεύουν τον κίνδυνο εξαγωγής λανθασμένων τιμών. Πρέπει να γίνει κατανοητό ότι οι τιμές τους είναι ακριβείς και μπορούν να ληφθούν άφοβα υπόψη για την εξέταση της σημαντικότητας των σταθερών επιδράσεων.

Κεφάλαιο 4

Μικτή Ανάλυση Διασποράς

4.1 Εισαγωγή

Η ανάλυσης διασποράς (ANOVA) αποτελεί ένα από τα πιο γνωστά εργαλεία ανάλυσης στην στατιστική. Έχει ως σκοπό να ελέγξει την επίδραση ενός ή περισσότερων παραγόντων συγκρίνοντας τους μέσους όρους διαφορετικών ομάδων. Η ανάλυση διασποράς με ένα παράγοντα, καθώς και με δυο παράγοντες με ή χωρίς αλληλεπίδραση αποτελούν τις πιο βασικές και γνωστές μεθόδους του συγκεκριμένου εργαλείου ανάλυσης. Η γνώση τους εισάγει τον αναγνώστη στο σκεπτικό της συγκεκριμένης μεθόδου καθώς και στις βασικές του έννοιες.

Στην συγκεκριμένη ενότητα θα γίνει προσέγγιση του μικτού γραμμικού μοντέλου με την βοήθεια της ANOVA. Η πρώτη επισήμανση μας πρέπει να αφορά το γεγονός πως καμία από τις βασικές μεθόδους που αναφέρθηκαν δεν χρησιμοποιείται αυτούσια στην προσέγγιση του μοντέλου. Γι'αυτόν τον σκοπό θα εισάγουμε μια νέα μέθοδο, η οποία καλείται μικτή ανάλυση διασποράς ή μικτή ANOVA, την οποία και θα αποκαλούμε από εδω και πέρα ως μικτή ANOVA για συντομία και θα παρουσιαστεί στην συνέχεια του κεφαλαίου.

Αρχικά θα γίνει μια θεωρητική αναφορά της μεθόδου, περιγράφοντας την πλήρως, καθώς και θα αποτυπωθούν χαρακτηριστικές περιπτώσεις-παραδείγματα στα οποία χρησιμοποιείται. Έτσι θα κατανοήσουμε ουσιαστικά σε τι είδους προβλήματα αναφέρεται και ταιριάζει η μικτή ANOVA. Έπειτα με την βοήθεια συγκεκριμένων παραδειγμάτων και της R θα παρουσιάσουμε πως μπορούμε να αναλύσουμε πλήρως μια μικτή ANOVA δυο ή τριών παραγόντων με σκοπό να καταλήξουμε στην εξαγωγή χρήσιμων συμπερασμάτων.

4.2 Περιγραφή της Μικτής Ανάλυσης Διασποράς

Προτού ξεκινήσουμε με την περιγραφή της μικτής ANOVA θα γίνει μια σύντομη αναφορά σε βασικές έννοιες που θα χρησιμοποιηθούν σε όλη την διάρκεια του κεφαλαίου. Η αφετηρία μας θα αφορά τις ανεξάρτητες μεταβλητές, οι οποίες θα αναφέρονται ως παράγοντες (ή factors), ενώ οι ομάδες τους ως στάθμες (ή levels). Καθώς στην μελέτη μας θα έχουμε υποχρεωτικά δυο ή περισσότερους παράγοντες, η εξαρτημένη μεταβλητή θα λαμβάνει τιμές για συγκεκριμένους συνδυασμούς από στάθμες των παραγόντων. Κάθε τέτοιος συνδυασμός ονομάζεται αγωγή (ή treatment).

Μια μικτή ANOVA λοιπόν, είναι ένας συνδυασμός μιας ANOVA μεταξύ των υποκειμένων και μιας ANOVA εντός των υποκειμένων. Πιο συγκεκριμένα απαιτούνται τουλάχιστον δυο παραγόντες, με τουλάχιστον έναν εκ των οποίων να μεταβάλλεται εντός των υποκειμένων και τουλάχιστον έναν μεταξύ των υποκειμένων. Η εξαρτημένη μεταβλητή στην περίπτωση μας είναι συνεχής, ενώ οι παράγοντες αποτελούν αυστηρά κατηγορικές μεταβλητές. Η κύρια διαφορά της με μια ANOVA δυο παραγόντων, που απαιτεί δυο παράγοντες μεταξύ των υποκειμένων ή με μια επαναλαμβανόμενη ANOVA δυο κατευθύνσεων, που απαιτεί δυο παράγοντες εντός των υποκειμένων, είναι η υποχρεωτική παρουσία παραγόντων τόσο μεταξύ όσο και εντός των υποκείμενων, κάτι που δεν συμβαίνει στις δυο προαναφερθείς ANOVA.

Έπειτα από την ανάγνωση του ορισμού της μικτής ANOVA, είναι κατανοητό ότι θα μας απασχολήσουν οι έννοιες εντός και μεταξύ των υποκειμένων. Είναι πολύ σημαντικό ο αναγνώστης να κατανοήσει πληρώς τις δυο αυτές έννοιες, ώστε να είναι σε θέση να αντιληφθεί ποιοί παράγοντες είναι εντός και ποιοί μεταξύ των υποκειμένων. Εκτός από το γεγονός πως η ύπαρξη τους είναι προαπαιτούμενο για την μικτή ANOVA, θα χρειαστούν και στην συνέχεια στην ανάλυση με την βοήθεια της R. Γι'αυτούς τους λόγους στην συνέχεια θα γίνει μια πιο λεπτομέρης αναφορά.

Η πρώτη έννοια στην οποία θα αναφερθούμε είναι αυτή μεταξύ των υποκειμένων. Για να ορίσουμε έναν τέτοιο παράγοντα πρέπει τα υποκείμενα μας να έχουν χωριστεί σε δυο ή περισσότερες ανεξάρτητες στάθμες της κατηγορικής μεταβλητής. Κάποια χαρακτηριστικά παραδείγματα αποτελούν το φύλο, η εθνικότητα, το ηλιακακό γκρουπ κλπ. Αντίθετα η ερμηνεία του παράγοντα εντός των υποκειμένων είναι πιο σύνθετη. Αυτήν την φορά το υποκείμενο καλείται να έχει "συμμετοχή" σε όλες τις στάθμες του παράγοντα. Πιο συγκεκριμένα για το ίδιο υποκείμενο πρέπει να έχει καταγραφεί η τιμή της εξαρτημένης μεταβλητής σε όλες τις στάθμες του συγκεκριμένου παράγοντα, με τον χρόνο να αποτελεί το πιο χαρακτηριστικό παράδειγμα. Στην επόμενη ενότητα θα εντοπίσουμε τους παράγοντες αυτούς, με την βοήθεια κάποιων χαρακτηριστικών περιπτώσεων και παραδειγμάτων που συναντάμε στην μικτή ANOVA.

Η αλληλέπιδραση που εμφανίζεται λόγω της υπάρξης περισσότερων από δυο παραγόντων στο μοντέλο παίζει σημαντικό ρόλο στην ανάλυση του προβλήματος. Είναι σημαντικό να συμπεράνουμε γρήγορα κατά την ανάλυση μας εάν είναι στατιστικά σημαντική, καθώς η προσέγγιση και η ανάλυση του προβλήματος αλλάζει με την απουσία της. Πρακτικά όμως τι σημαίνει η ύπαρξη αλληλεπίδρασης όταν διαθέτουμε δυο παράγοντες Α, Β. Ότι η ταυτόχρονη δράση δυο συγκεκριμένων στάθμεων (αγωγή) των παραγόντων Α, Β δεν ισούται με το άθροισμα των ξεχωριστών επιδράσεων των δυο αυτών στάθμεων. Διαφορετικά η επιρροή που θα έχει η στάθμη ενός παράγοντα στην εξαρτημένη μεταβλητή, εξαρτάται από την επιλογή της στάθμης του άλλου παράγοντα.

Ένα από τα ζητήματα που δεν επισημάνθηκαν έως τώρα, είναι σε τι είδους συμπεράσματα καταλήγουμε εφαρμόζοντας μια μικτή ANOVA. Καθώς βρισκόμαστε σε πρώιμο στάδιο, αυτό που μπορούμε να αναφέρουμε είναι ότι η χρήση μιας μικτής ANOVA έχει ως σκοπό την σύγκριση των μέσων όρων ομάδων που διασταυρώνονται με δυο ή περισσότερους παράγοντες, οι οποίοι έχουν τις ιδιότητες που καταγράφαμε προηγουμένως. Αναλυτικότερη αναφορά στα συμπεράσματα τα οποία μπορούν να προκύψουν, καθώς και στον τρόπο επίτευξης τους και στο γενικότερο σκεπτικό θα δωθούν κατά την χρησιμοποίηση του στατιστικού πακέτου της R.

Αξίζει να σημειωθεί ότι όπως και σε οποιαδήποτε ANOVA, έτσι και στην μικτή ANOVA υπάρχει η δυνατότητα να έχουμε περισσότερους από τρείς παράγοντες, με την προϋπόθεση ότι τουλάχιστον ένας μεταβάλλεται εντός και τουλαχίστον ένας μεταξύ των υποκειμένων. Προφανώς όσους περισσότερους παράγοντες διαθέτουμε τόσο πιο πολύπλοκη, σύνθετη και χρονοβόρα γίνεται η ανάλυση. Εμείς θα περιοριστούμε στην μελέτη των περιπτώσεων των δυο και των τριών παραγόντων. Προβλήματα που περιέχουν περισσότερους από τρείς παράγοντες αντιμετωπίζονται με το ίδιο σκεπτικό.

4.3 Περιπτώσεις - Παραδείγματα

Στην υποενότητα αυτή θα αναφερθούμε σε τρείς χαρακτηριστικές περιπτώσεις που μπορούμε να συναντήσουμε κατά την μελέτη μιας μικτής ANOVA. Εκτός από την αναφορά των περιπτώσεων, θα γίνουμε ακόμα πιο συγκεκριμένοι με την καταγραφή παραδειγμάτων τα οποία θα βοηθήσουν στην πλήρη κατανόηση, τόσο της μικτής ANOVA αλλά και των εννοιών εντός και μεταξύ των υποκειμένων.

Στην πρώτη περίπτωση που θα εξετάσουμε θα έχουμε ως παράγοντα εντός των υποκειμένων τον χρόνο, ενώ ως παράγοντα μεταξύ των υποκειμένων μία συνθήκη. Για να γίνουμε ακόμα πιο συγκεκριμένοι θα δώσουμε ένα παράδειγμα. Στόχος της ανάλυσης είναι η σύγκριση δυο θεραπειών A, B ως προς την αποτελεσματικότητα με το πέρας του χρόνου στην μείωση της αρτηριακής πίεσης των ασθενών. Οι ασθενείς είτε λαμβάνουν την θεραπεία A είτε την B και έστω ότι έχουμε συνολικά τέσσερις μετρήσεις ανά δέκα ημέρες. Εύκολα διαπιστώνουμε ότι το υποκείμενο μας είναι ο ασθενής ενώ η εξαρτημένη μεταβλητή είναι η αρτηριακή πίεση. Προφανώς ο χρόνος αποτελεί τον παράγοντα εντός των υποκειμένων, ενώ οι θεραπείες τον παράγοντα μεταξύ των υποκειμένων. Με την πραγματοποίηση της μικτής ANOVA θα καταλήξουμε εάν η μέση αρτηριακή πίεση των ασθενών διαφέρει σημαντικά μεταξύ των δυο θεραπειών σε οποιαδήποτε από τις τέσσερις χρονικές στιγμές.

Στην δεύτερη περίπτωση και πάλι ο παράγοντας εντός των υποκειμένων είναι ο χρόνος, ωστόσο αυτήν την φορά ο παράγοντας μεταξύ των υποκειμένων είναι ένα χαρακτηριστικό του δείγματος. Προχωράμε στο παράδειγμα μας. Σκοπός της συγκεκριμένης έρευνας είναι να συγκρίνει την περιεκτικότητα του αλκοόλ στο αίμα μεταξύ ανδρών και γυναικών έπειτα από την κατανάλωση συγκεκριμένης ποσότητας αλκοόλ. Οι μετρήσεις λαμβάνονται σε τρείς συγκεκριμένες χρονικές στιγμές. Το υποκείμενο και πάλι είναι ο ασθενής, η εξαρτημένη μεταβλητή είναι η περιεκτικότητα του αλκοόλ στο αίμα, ενώ οι παράγοντες εντός και μεταξύ των υποκειμένων είναι ο χρόνος και το φύλο αντίστοιχα. Με την πραγματοποίηση της μικτής ANOVA αυτήν την φορά, μπορούμε να πούμε ότι θα καταλήξουμε εάν η μέση περιεκτικότητα του αλκοόλ στο αίμα διαφέρει σημαντικά μεταξύ των δυο φύλων σε οποιαδήποτε από τις τρείς χρονικές στιγμές.

Στην τελευταία περίπτωση δεν εμφανίζεται ο χρόνος, αλλά αυτήν την φορά έχουμε ως παράγοντα εντός των υποκειμένων μια συνθήκη, ενώ ως παράγοντα μεταξύ των υποκειμένων ένα χαρακτηριστικό του δείγματος. Έστω ότι αυτήν την φορά θέλουμε να εξετάσουμε την επίδραση της έντασης της άσκησης στην κατάθλιψη, λαμβάνοντας υπόψη τις διαφορές που μπορεί να προκύψουν μεταξύ των δυο φύλλων. Πιο συγκεκριμένα όσον αφορά την ένταση της άσκησης, έχουμε τρία είδη: χαμηλή, μεσαία και υψηλή. Η διαφορά με τα δυο προηγούμενα παραδείγματα έγκειται στο γεγονός πως ο κάθε συμμετέχων καλείται να περάσει και από τα τρία επίπεδα έντασης της άσκησης και να καταγράψει τα αποτελέσματα. Επομένως η εξαρτημένη μεταβλητή είναι η κατάθλιψη, η οποία μετράται με την χρήση ενός ειδικού δείκτη, ενώ οι παράγοντες εντός και μεταξύ των υποκειμένων είναι η ένταση της άσκησης και το φύλο αντίστοιχα. Τέλος κατά την πραγματοποίηση της μικτής ANOVA θα καταλήξουμε εάν η μέση βαθμολογία της κατάθλιψης διαφέρει σημαντικά μεταξύ δυο συγκεκριμένων στάθμεων του παράγοντα ένταση της άσκησης για οποιοδήποτε φύλο.

Οι παραπάνω αποτελούν τρείς από τις πολλές περιπτώσεις που μπορούμε να συναντήσουμε κατά την μελέτη μιας μικτής ANOVA δυο παραγόντων. Αξίζει να σημειωθεί ότι εύκολα μπορούν να γίνουν μικτές ANOVA τριών παραγόντων, εφόσον προστεθεί ένας ακόμα κατάλληλος παράγοντας. Όσον αφορά τα παραδείγματα, εκτός από την κατανόηση του μοντέλου και των εννοιών εντός και μεταξύ των παραγόντων, μας βοήθησαν να έχουμε μια πρώτη εικόνα για τα συμπεράσματα στα οποία ενδεχομένως μας ζητείτε να καταλήξουμε. Έτσι έχουμε σχεδόν ολοκληρώσει την θεωρητική μας αναφορά στο μοντέλο και είμαστε σε θέση να περάσουμε στο πρακτικό κομμάτι με την βοήθεια της R.

4.4 Μικτή ANOVA δυο Παραγόντων με τη βοήθεια της R

Στην συγκεκριμένη ενότητα θα ασχοληθούμε με την ανάλυση της μικτής ANOVA δυο παραγόντων με την βοήθεια της R. Εκτός από τις βασικές υποθέσεις που αποτυπώθηκαν στην Ενότητα 4.2, υπάρχουν και άλλες τέσσερις που παραλείφθηκαν, για τον απλούστατο λόγο ότι συνδέονται με την R για την επαλήθευση τους. Επομένως έπειτα από την αναφορά των υποθέσεων αυτών και την εισαγωγή του αναγνώστη στο παράδειγμα το οποίο θα μελετήσουμε, θα ακολουθήσει με την βοήθεια της R η επαλήθευση αυτών των υποθέσεων καθώς και η εφαρμογή της μικτής ANOVA, με σκοπό να καταλήξουμε στα ζητούμενα συμπεράσματα.

4.4.1 Υποθέσεις για τα Δεδομένα

Εφόσον καταλήξουμε σε θεωρητικό επίπεδο ότι στο πρόβλημα μας ταιριάζει η μικτή ANOVA, πριν την κύρια ανάλυση οφείλουμε να επαληθεύσουμε τις τέσσερις υποθέσεις που θα αναφερθούν στην συνέχεια. Με την επιβεβαίωση τους θα είμαστε σίγουροι ότι εκτός από το πρόβλημα και στα δεδομένα ταιριάζει μία τέτοια ανάλυση. Ωστόσο οφείλουμε να επισημάνουμε ότι στην περίπτωση που μελετάμε πραγματικά δεδομένα και όχι εικονικά είναι συχνό και λογικό να μην επαληθεύονται μία ή περισσότερες εξ αυτών. Οι τέσσερις υποθέσεις είναι οι εξής:

- **Υπόθεση 1**: Δεν θα πρέπει να υπάρχουν σημαντικά ακραίες τιμές σε καμία αγωγή του προβλήματος.

- **Υπόθεση 2**: Η εξαρτημένη μεταβλητή θα πρέπει να ακολουθεί κατά προσέγγιση την κανονική κατανομή για κάθε αγωγή του προβλήματος.

- **Υπόθεση 3**: Η διακύμανση της εξαρτημένης μεταβλητής πρέπει να είναι ίση μεταξύ των ομάδων των παραγόντων μεταξύ των υποκειμένων. Η υπόθεση αυτή είναι γνωστή και ως ομοιογένεια των διαφορών.

- **Υπόθεση 4**: Η διακύμανση των διαφορών των παρατηρούμενων τιμών μεταξύ των ομάδων του παράγοντα εντός των υποκειμένων πρέπει να είναι ίση για όλες τις ομάδες του παράγοντα μεταξύ των υποκειμένων. Η υποθέση αυτή είναι γνωστή και ως σφαιρικότητα.

Με την αναφορά των παραπάνω υποθέσεων έχουμε ουσιαστικά ολοκληρώσει τη θεωρητική αναφορά για την μικτή ANOVA. Έτσι είμαστε πλέον έτοιμοι να περάσουμε στο παράδειγμα το οποίο και θα χρησιμοποιήσουμε για την ανάλυση με την χρήση της R.

4.4.2 Εισαγωγή στο Παράδειγμα

Στο παράδειγμα το οποίο θα μελετήσουμε θα χρησιμοποιήσουμε τα δεδομένα depression από το πακέτο datarium της R. Τα δεδομένα αντιστοιχούν σε ένα πείραμα στο οποίο μελετάται μια θεραπεία για την κατάθλιψη. Έχουμε λοιπόν δυο ομάδες ασθενών (1: ctr - 2: treatment) με την πρώτη ομάδα να μην λαμβάνει την θεραπεία ενώ την δεύτερη να είναι αυτή που την λαμβάνει. Οι ασθενείς αυτοί παρακολουθήθηκαν σε τέσσερις διαφορετικές χρονικές στιγμές (t_0: προ-θεραπείας , t_1: 1 μήνα μετά την έναρξη του πειράματος , t_2: 3 μήνες μετά την έναρξη του πειράματος , t_3: 6 μήνες μετά την έναρξη του πειράματος) και σε κάθε χρονική στιγμή καταγράφηκε η βαθμολογία της κατάθλιψης. Για την καταγραφή της βαθμολογίας της κατάθλιψης χρησιμοποιήθηκε ένα τυποποιημένο ερωτηματολόγιο, με τιμές που κυμαίνονται από 1 έως 500. Οι υψηλότερες τιμές στην κλίμακα αυτή υποδηλώνουν πιο σοβαρά συμπτώματα κατάθλιψης. Τέλος καλούμαστε να καταλήξουμε εάν οποιαδήποτε αλλαγή στην βαθμολογία της κατάθλιψης είναι αποτέλεσμα της αλληλεπίδρασης των παραγόντων θεραπείας και χρόνου, όπως επίσης να εξετάσουμε και ενδεχόμενη διαφορετική επιρροή συγκεκριμένων αγωγών στην μέση βαθμολογία της κατάθλιψης.

Οι δυο παράγοντες που θα εξετάσουμε στο πρόβλημα μας είναι η θεραπεία (treatment) και ο χρόνος (time) με τον πρώτο να αποτελεί τον παράγοντα μεταξύ των υποκειμένων, ενώ τον δεύτερο τον παράγοντα εντός των υποκειμένων. Προφανώς η εξαρτημένη μεταβλητή αποτελείται από την βαθμολογία της κατάθλιψης (score). Συνολικά έχουμε 96 τιμές δεδομένων και 24 συμμετέχοντες.

Έχοντας συμπεριλάβει τους δυο αυτούς παράγοντες και προφανώς την αλληλεπίδραση τους, το μοντέλο μας μπορεί να εκφραστεί ως εξής:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, i = 1, ..., 4, j = 1, 2, k = 1, 2, ..., 96,$$

όπου μ είναι η σταθερά, α_i είναι η κύρια επίδραση της i στάθμης του παράγοντα χρόνου, β_j είναι η κύρια επίδραση της j στάθμης του παράγοντα θεραπεία, $(\alpha\beta)_{ij}$ είναι η αλληλεπίδραση που οφείλεται στον συνδυασμό των δυο παραγόντων χρόνου και θεραπείας στην αγωγή (ij) και ε_{ijk} είναι το τυχαίο σφάλμα της k παρατήρησης της εξαρτημένης μεταβλητής y στην αγωγή (ij).

4.4.3 Εισαγωγή στην R και Επαλήθευση των Υποθέσεων

Προτού ξεκινήσουμε με την επαλήθευση των τεσσάρων υποθέσεων, πρέπει να εισάγουμε τα δεδομένα στην R, καθώς και τις απαραίτητες "βιβλιοθήκες" που θα χρειαστούμε καθόλη την διάρκεια της ανάλυσης, με τις οποίες και θα ξεκινήσουμε.

```
> library(tidyverse)
> library(ggpubr)
> library(rstatix)
```

Τα δεδομένα εισάγονται ως εξής:

```
> data("depression",package="datarium")
```

Καλούμαστε τώρα να δημιουργήσουμε μια ενιαία στήλη για τις τέσσερις χρονικές στιγμές που διαθέτουμε την οποία θα ονομάσουμε time, όπως επίσης και μια στήλη score με την βαθμολογία της κατάθλιψης σε κάθε χρονική στιγμή. Τέλος θα μετατρέψουμε τις στήλες id και time σε παράγοντες.

```
> depression<-depression %>%
+ gather(key="time",value="score",t0,t1,t2,t3) %>%
+ convert_as_factor(id,time)
```

Πλέον έχουμε φέρει τα δεδομένα στην κατάλληλη μορφή ώστε να ξεκινήσουμε την επαλήθευση των τεσσάρων υποθέσεων που επισημάνθηκαν στην Υποενότητα 4.4.1, καθώς έχουμε δημιουργήσει τις στήλες που χρειαζόμασταν (id, time, treatment, score).

Ακραίες Τιμές

Ο πρώτος έλεγχος αφορά την ύπαρξη ακραίων τιμών στα δεδομένα και πολύ εύκολα ο έλεγχος γίνεται ως εξής:

```
> group_by(depression,time,treatment) %>%
+ identify_outliers(score)

A tibble: 4 x 6
  treatment time  id    score is.outlier is.extreme
  <fct>     <fct> <fct> <dbl> <lgl>      <lgl>
1 ctr       t0    5       150 TRUE       FALSE
2 ctr       t0    8       447 TRUE       FALSE
3 treated   t0    18      138 TRUE       FALSE
4 treated   t0    22      150 TRUE       FALSE
```

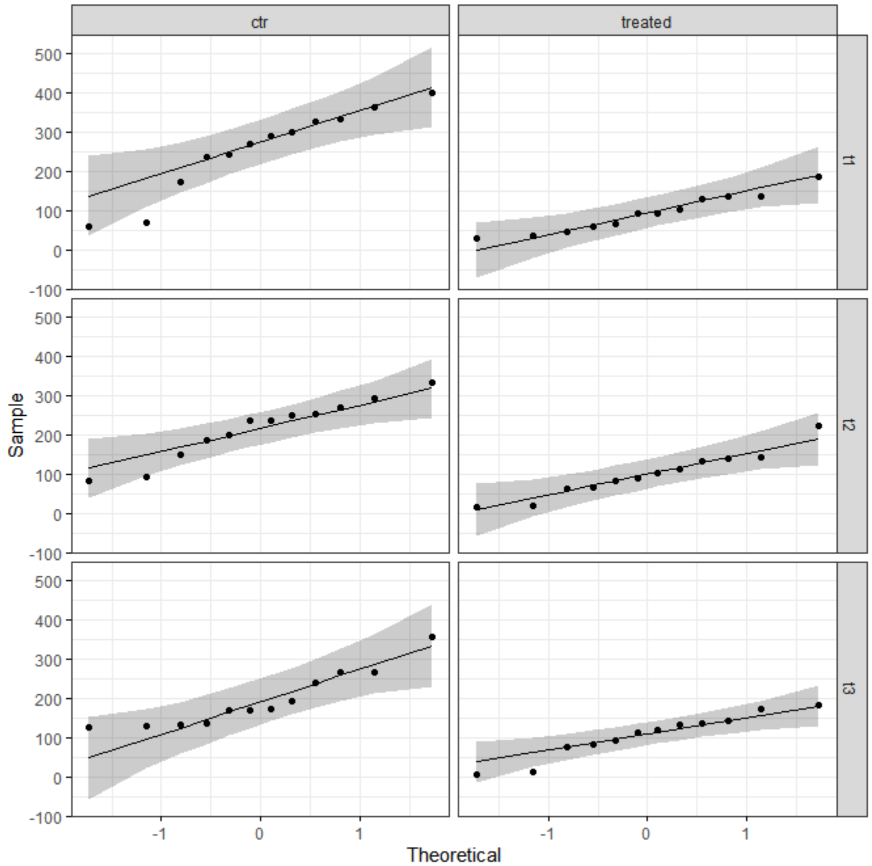
Πριν σχολιάσουμε τα αποτελέσματα, θα γίνουμε λίγο πιο αναλυτικοί για τις συναρτήσεις group_by, %>%, καθώς χρησιμοποιούνται κατά κόρον. Η συνάρτηση (ή διαφορετικά τελεστής) %>% συνδέει την προηγούμενη με την επόμενη γραμμή του κώδικα. Η συνάρτηση group_by τώρα, ομαδοποιεί τις γραμμές των δεδομένων μας ως προς τους παράγοντες που έχει στα ορίσματα της. Ο συνδυασμός των δυο αυτών συναρτήσεων στην προκειμένη περίπτωση, δηλώνει στην R ότι αναζητούμε τις ακραίες τιμές της εξαρτημένης μεταβλητής στις αγωγές του προβλήματος.

Από τα παραπάνω αποτελέσματα βλέπουμε ότι η R έχει εντοπίσει τέσσερις παρατηρήσεις, τις οποίες έχει προσδιορίσει πλήρως δίνοντας το id και την στάθμη των δυο παραγόντων στην οποία ανήκει η κάθε μια. Η στήλη is.outlier σε συνδυασμό με την λογική τιμή TRUE για κάθε παρατήρηση μας βοηθάει να καταλάβουμε ότι οι τέσσερις αυτές παρατηρήσεις είναι "ασυνήθιστες", δηλαδή είναι στα όρια της ακραίας τιμής. Ωστόσο η στήλη που μας ενδιαφέρει για να κατανοήσουμε εάν τελικά υπάρχουν ακραίες τιμές, είναι η is.extreme στην οποία σε συνδυασμό με την λογική τιμή FALSE που υπάρχει σε κάθε παρατήρηση, μας βοηθάει να καταλήξουμε ότι καμία από αυτές τις παρατηρήσεις δεν αποτελεί εν τέλει ακραία.

Υπόθεση Κανονικότητας

Για την επιβεβαίωση της υπόθεσης της κανονικότητας, έχουμε δυο επιλογές. Η πρώτη είναι με την κατασκευή ενός διαγράμματος Q-Q. Δημιουργούμε λοιπόν ένα ξεχωριστό διάγραμμα Q-Q για κάθε αγωγή του προβλήματος, με την κατακόρυφη στήλη να αποτελείται από τα δειγματικά ποσοστημόρια, ενώ την οριζόντια από τα θεωρητικά. Το διάγραμμα είναι το εξής:

```
> ggqqplot(depression,"score",ggtheme=theme_bw())+
+ facet_grid(time~treatment))
```



Διάγραμμα 4.4: Διάγραμμα Q-Q για τα δεδομένα του Παραδείγματος της Υποενότητας 4.4.2.

Διαφορετικά μπορούμε να χρησιμοποιήσουμε τον έλεγχο Shapiro-Wilk για την υπόθεση κανονικότητας. Ο έλεγχος παρουσιάζεται αμέσως μετά.

```
> group_by(depression,time,treatment) %>%
+ shapiro_test(score)

# A tibble: 8 x 5
  treatment time  variable statistic      p
  <fct>     <fct> <chr>        <dbl>  <dbl>
1 ctr       t0    score        0.945 0.572
2 treated   t0    score        0.836 0.0248
3 ctr       t1    score        0.918 0.268
4 treated   t1    score        0.953 0.680
5 ctr       t2    score        0.951 0.652
6 treated   t2    score        0.957 0.734
7 ctr       t3    score        0.873 0.0706
8 treated   t3    score        0.932 0.398
```

Με την βοήθεια του Διαγράμματος 4.4 και τον έλεγχο Shapiro-Wilk, συμπεραίνουμε ότι εκτός της αγωγής “treated-t_0”, τα δεδομένα μας είναι κανονικά κατανεμημένα για κάθε αγωγή του προβλήματος.

Ομοιογένεια των Διαφορών

Η εξέταση της ομοιογένειας των διαφορών θα γίνει με την βοήθεια του Levene’s ελέγχου. Ο έλεγχος θα πραγματοποιηθεί σε κάθε στάθμη του παράγοντα time, ο οποίος αποτελεί και τον παράγοντα εντός των υποκειμένων. Αν δεν απορριφθεί η μηδενική υπόθεση του ελέγχου, τότε οι στάθμες του παράγοντα treatment έχουν ίσες διακυμάνσεις, ενώ σε αντίθετη περίπτωση διαφέρει τουλάχιστον μία.

```
> group_by(depression,time) %>%
+ levene_test(score~treatment)

# A tibble: 4 x 5
  time    df1   df2 statistic      p
  <fct> <int> <int>     <dbl>  <dbl>
1 t0        1    22    0.0753 0.786
2 t1        1    22    3.62   0.0701
3 t2        1    22    0.573  0.457
4 t3        1    22    0.339  0.567
```

Για τον παραπάνω στατιστικό έλεγχο δίνονται οι βαθμοί ελευθερίας, καθώς και οι F, p τιμές του. Από τα παραπάνω αποτελέσματα παρατηρούμε ότι και στις τέσσερις στάθμες του παράγοντα time, $p > 0.05$ επομένως δεν έχουμε λόγο να απορρίψουμε την μηδενική υπόθεση, δηλαδή να καταλήξουμε ότι οι στάθμες του παράγοντα treatment έχουν ίσες διακυμάνσεις σε κάθε στάθμη του παράγοντα time.

Υπόθεση της Σφαιρικότητας

Η υπόθεση της σφαιρικότητας δεν χρειάζεται να ελεγχθεί ξεχωριστά, καθώς γίνεται αυτόματος έλεγχος κατά την εφαρμογή του ANOVA test και πιο συγκεκριμένα με την χρήση της συνάρτησης anova_test. Εάν υπάρχει κάποιος παράγοντας που παραβιάζει αυτήν την υπόθεση, τότε διορθώνεται αυτόματα με την βοήθεια της συνάρτησης get_anova_table.

Ο έλεγχος της υπόθεσης της σφαιρικότητας γίνεται με το τεστ σφαιρικότητας του Mauchly. Στην περίπτωση που παραβιάζεται η υπόθεση της σφαιρικότητας, χρειάζεται να διορθώσουμε την p-τιμή των κύριων επιδράσεων της αγωγής, χρησιμοποιώντας είτε τη διόρθωση Huynh-Feldt είτε τη διόρθωση Greenhouse-Geisser. Ωστόσο δεν θα γίνει αναλυτικότερη αναφορά ούτε στο τεστ σφαιρικότητας του Mauchly, ούτε στις δυο διορθώσεις και θα βασιστούμε στα αποτελέσματα που λαμβάνουμε με τη βοήθεια της R.

4.4.4 Κύρια Ανάλυση και Εξαγωγή Συμπερασμάτων

Έπειτα από την επαλήθευση των υποθέσεων, είμαστε σε θέση να χρησιμοποιήσουμε την μικτή ANOVA για την ανάλυση των δεδομένων μας. Οφείλουμε να επισημάνουμε ότι όταν θέλουμε να εφαρμόσουμε μια μικτή ANOVA δυο παραγόντων, τα βήματα που πρέπει να ακολουθήσουμε είναι συγκεκριμένα. Γι'αυτόν τον σκοπό στην συνέχεια θα αναφέρουμε εν συντομία τα συγκεκριμένα βήματα.

- Βήμα 1: Ο πρώτος έλεγχος που θα πραγματοποιηθεί θα αφορά την σημαντικότητα του όρου της αλληλεπίδρασης των δυο παραγόντων και είναι το πιο σημαντικό βήμα καθώς η μετέπειτα ανάλυση εξαρτάται από αυτόν.

Ο όρος της αλληλεπίδρασης είναι στατιστικά σημαντικός.

- Βήμα $2i$): Στην περίπτωση που ο όρος της αλληλεπίδρασης είναι στατιστικά σημαντικός, ο επόμενος έλεγχος που πραγματοποιείται είναι μια απλή ανάλυση κύριων επιδράσεων, κατά την οποία εξετάζουμε την επίδραση του παράγοντα Α σε όλες τις στάθμες του παράγοντα Β. Επομένως καταλήγουμε στις στάθμες του παράγοντα Β, στις οποίες είναι στατιστικά σημαντική η απλή κύρια επίδραση Α.

- Βήμα $3i$): Στο βήμα αυτό θα χρησιμοποιήσουμε μόνο τις στάθμες του παράγοντα Β, που με βάση τα αποτελέσματα του ελέγχου του βήματος $2i$) θεωρήθηκαν ως στατιστικά σημαντικές. Στις συγκεκριμένες στάθμες, θα πραγματοποιήσουμε πολλαπλές συγκρίσεις κατά ζεύγη μεταξύ των υποομάδων του παράγοντα Α, ώστε να καταλήξουμε εν τέλει μεταξύ ποιών υποομάδων του παράγοντα Α εντοπίζεται διαφορά.

Ο όρος της αλληλεπίδρασης δεν είναι στατιστικά σημαντικός.

- Βήμα $2ii$): Στην περίπτωση που ο όρος της αλληλεπίδρασης δεν είναι στατιστικά σημαντικός, καλούμαστε να ερμηνεύσουμε με την βοήθεια του ελέγχου του βήματος 1, την στατιστική σημαντικότητα των δυο κύριων επιδράσεων.

- Βήμα $3ii$): Τέλος πραγματοποιούμε πολλαπλές συγκρίσεις κατά ζεύγη μεταξύ των στάθμεων του ή των παραγόντων που είναι στατιστικά σημαντικοί σύμφωνα με το βήμα $2ii$). Επομένως θα καταλήξουμε σε τυχόν διαφορές μεταξύ των ομάδων εντός των παραγόντων.

Καλούμαστε πλέον να περάσουμε στο πρακτικό κομμάτι, ακολουθώντας τα παραπάνω βήματα και εφαρμόζοντας τους κατάλληλους ελέγχους με την βοήθεια της R. Ο πρώτος έλεγχος, που είναι ένας έλεγχος μικτής ANOVA δυο κατευθύνσεων, για να πραγματοποιηθεί πρέπει μαζί με την εντολή να εισάγουμε τα δεδομένα, την στήλη αρίθμησης των υποκειμένων του προβλήματος, την εξαρτημένη μεταβλητή καθώς και τους παράγοντες εντός και μεταξύ των υποκειμένων. Επομένως ο έλεγχος είναι ο εξής:

```
> mixed.anova<-anova_test(data=depression,dv=score,
+ wid=id,between=treatment,within=time)
> get_anova_table(mixed.anova)
```

```
ANOVA Table (type II tests)

           Effect  DFn   DFd      F        p p<.05   ges
1       treatment 1.00 22.00 17.886 3.45e-04     * 0.330
2            time 2.19 48.19 37.615 5.05e-11     * 0.402
3  treatment:time 2.19 48.19  7.164 1.00e-03     * 0.113
```

Από τα παραπάνω αποτελέσματα και σε ότι αφορά την αλληλεπίδραση, βλέπουμε ότι μπορούμε να απορρίψουμε την μηδενική υπόθεση, αφού p$= 0.001 < 0.05$. Επομένως καταλήγουμε ότι η αλληλεπίδραση είναι στατιστικά σημαντική και επηρεάζει την τιμή της βαθμολογίας της κατάθλιψης. Είναι λοιπόν περιττό να ελέγξουμε τις κύριες επιδράσεις.

Εφόσον η αλληλεπίδραση είναι στατιστικά σημαντική, ακολουθούμε το βήμα $2i$) και πιο συγκεκριμένα επιλέγουμε να εξετάσουμε την επίδραση του παράγοντα time, σε όλες τις στάθμες του παράγοντα treatment. Αντίστοιχα μπορούσαμε να εξετάσουμε την επίδραση του παράγοντα treatment σε όλες τις στάθμες του παράγοντα time.

```
> one.way<-group_by(depression,treatment) %>%
+ anova_test(dv=score,wid=id,within=time) %>%
+ get_anova_table() %>%
+ adjust_pvalue(method="bonferroni")
> one.way

# A tibble: 2 x 6
  treatment Effect   DFn   DFd     F    p.adj
  <fct>     <chr>  <dbl> <dbl> <dbl>    <dbl>
1 ctr       time       3    33  7.78 9.2 e- 4
2 treated   time       3    33 47.0  1.04e-11
```

Από τα παραπάνω αποτελέσματα και λαμβάνοντας υπόψη τις p.adj τιμές, συμπεραίνουμε ότι μπορούμε να απορρίψουμε την μηδενική υπόθεση και στις δυο περιπτώσεις, αφού p< 0.05. Επομένως η απλή κύρια επίδραση του χρόνου, είναι στατιστικά σημαντική και στις δυο στάθμες του παράγοντα treatment. Αναμένεται δηλαδή να υπάρχουν διαφορές στην μέση βαθμολογία της κατάθλιψης μεταξύ διαφορετικών χρονικών στιγμών. Επομένως για τον τελευταίο έλεγχο, θα πραγματοποιήσουμε πολλαπλές συγκρίσεις κατά ζεύγη μεταξύ των υποομάδων του παράγοντα time σε κάθε στάθμη του παράγοντα treatment, με σκοπό την εύρεση αυτών των χρονικών στιγμών.

Προτού πραγματοποιήσουμε τον έλεγχο αυτό, αξιζεί να σημειωθεί ότι και στον προηγούμενο και στον επόμενο έλεγχο στα αποτελέσματα παραλείπονται οι ανούσιες στήλες που μας δίνει η R. Επίσης για την εξαγωγή τον p τιμών χρησιμοποιείται η διόρθωση κατά Bonferroni για μεγαλύτερη ακρίβεια στα αποτελέσματα.

Κατά την πραγματοποίηση πολλαπλών συγκρίσεων, αυξάνεται η πιθανότητα σφάλματος τύπου I, δηλαδή η απόρριψη της μηδενικής υπόθεσης, ενώ είναι πραγματικά αληθής. Ένας τρόπος αντιμετώπισης του προβλήματος αυτού, είναι η χρήση της μεθόδου Bonferroni με βάση την οποία όταν πραγματοποιούμε N συνολικά τέτοιες συγκρίσεις αναπροσαρμόζουμε την p-τιμή κάθε ελέγχου. Αν υποθέσουμε ότι το αρχικό επίπεδο σημαντικότητας συμβολίζεται με p, τότε το νέο επίπεδο σημαντικότητας p' προσαρμοσμένο κατά Bonferroni προκύπτει ως εξής: $p' = p/N$.

```
# χρησιμοποιείται paired έλεγχος γιατί έχουμε επαναλαμβανόμενες παρατηρήσεις ως
προς τον χρόνο
> test<-group_by(depression,treatment) %>%
+ pairwise_t_test(score~time,paired=TRUE,
+ p.adjust.method="bonferroni")
> test

# A tibble: 12 x 6
   treatment group1 group2    n1    n2       p.adj
   <fct>     <chr>  <chr>  <int> <int>       <dbl>
 1 ctr       t0     t1        12    12 0.362
 2 ctr       t0     t2        12    12 0.008
 3 ctr       t0     t3        12    12 0.029
 4 ctr       t1     t2        12    12 0.08
 5 ctr       t1     t3        12    12 0.385
 6 ctr       t2     t3        12    12 1
 7 treated   t0     t1        12    12 0.00000762
 8 treated   t0     t2        12    12 0.00000738
 9 treated   t0     t3        12    12 0.000261
10 treated   t1     t2        12    12 1
11 treated   t1     t3        12    12 1
12 treated   t2     t3        12    12 1
```

Με την βοήθεια του παραπάνω ελέγχου και λαμβάνοντας υπόψη τις p.adj τιμές της διόρθωσης κατά Bonferroni καταλήγουμε στα εξής συμπεράσματα:

- Η μέση βαθμολογία της κατάθλιψης όταν δεν πραγματοποιείται η θεραπεία (δηλαδή στην στάθμη ctr) είναι στατιστικά σημαντικά διαφορετική μεταξύ των χρονικών στιγμών $\mathbf{t}_0$ vs $\mathbf{t}_2$, $\mathbf{t}_0$ vs $\mathbf{t}_3$.

- Αντίθετα η μέση βαθμολογία της κατάθλιψης όταν δεν πραγματοποιείται η θεραπεία (δηλαδή στην στάθμη ctr) **δεν** διαφέρει στατιστικά σημαντικά μεταξύ των χρονικών στιγμών $\mathbf{t}_0$ vs $\mathbf{t}_1$, $\mathbf{t}_1$ vs $\mathbf{t}_2$, $\mathbf{t}_1$ vs $\mathbf{t}_3$, $\mathbf{t}_2$ vs $\mathbf{t}_3$.

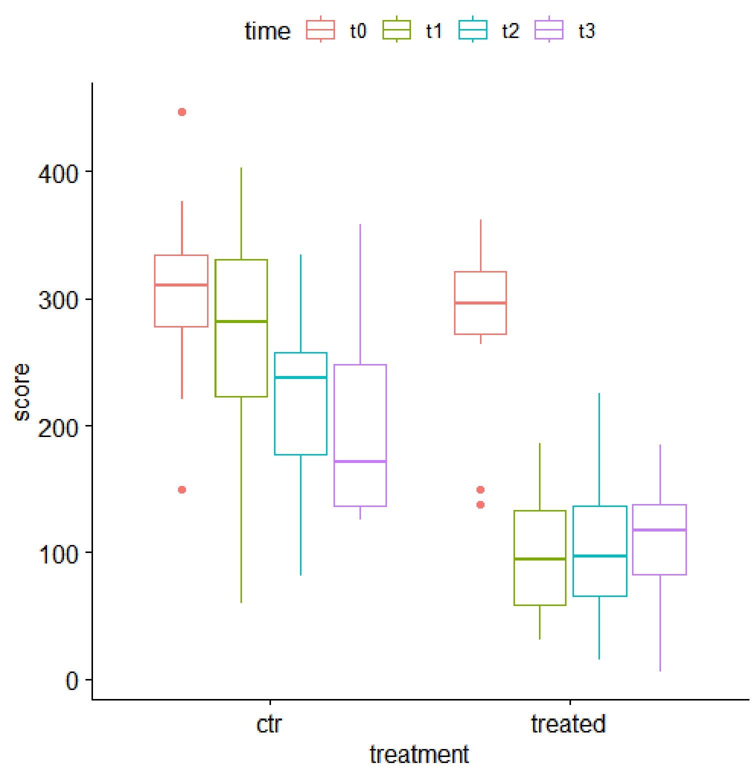
- Η μέση βαθμολογία της κατάθλιψης όταν πραγματοποιείται η θεραπεία (δηλαδή στην στάθμη treated) είναι στατιστικά σημαντικά διαφορετική μεταξύ των χρονικών στιγμών

$\mathbf{t}_0$ vs $\mathbf{t}_1$, $\mathbf{t}_0$ vs $\mathbf{t}_2$, $\mathbf{t}_0$ vs $\mathbf{t}_3$.

- Αντίθετα η μέση βαθμολογία της κατάθλιψης όταν πραγματοποιείται η θεραπεία (δηλαδή στην στάθμη treated) **δεν** διαφέρει στατιστικά σημαντικά μεταξύ των χρονικών στιγμών $\mathbf{t}_1$ vs $\mathbf{t}_2$, $\mathbf{t}_1$ vs $\mathbf{t}_3$, $\mathbf{t}_2$ vs $\mathbf{t}_3$.

Με την ερμηνεία των αποτελεσμάτων του τελευταίου ελέγχου έχουμε ουσιαστικά ολοκληρώσει την ανάλυση μας. Υπάρχει επίσης η δυνατότητα κατασκευής ενός θηκοδιαγράμματος, ώστε να παρατηρήσουμε όσα αποδείχθηκαν προηγουμένως, το οποίο και θα παρουσιαστεί στην συνέχεια.

```
> boxplot<-ggboxplot(depression,x="treatment",y="score",
+ color="time")
```



Διάγραμμα 4.5: Θηκοδιάγραμμα για τα δεδομένα του παραδείγματος της Υποενότητας 4.4.2

Με την βοήθεια του Διαγράμματος 4.5, παρατηρούμε ότι στην περίπτωση που οι ασθενείς λαμβάνουν την θεραπεία, η βαθμολογία της κατάθλιψης μειώνεται σημαντικά από την πρώτη χρονική στιγμή (t_1) και διατηρείται περίπου στο ίδιο επίπεδο μέχρι το τέλος του πειράματος. Αντίθετα όταν δεν χορηγείται θεραπεία, η πρώτη σημαντική μείωση που είναι ωστόσο μικρή, πραγματοποιείται την χρονική στιγμή t_2.

Αξίζει να σημειωθεί ότι η επιλογή της κύριας επίδρασης στον δεύτερο έλεγχο (one.way), λόγω της διατύπωσης του προβλήματος και πιο συγκεκριμένα του ερωτήματος του Παραδείγματος 4.4.2 είναι ελεύθερη. Στην περίπτωση όμως που στο πρόβλημα αναφερόταν, ότι καλούμαστε να εξετάσουμε εάν η μέση βαθμολογία της κατάθλιψης διαφέρει σημαντικά μεταξύ των θεραπειών σε οποιαδήποτε από τις τέσσερις χρονικές στιγμές, τότε υποχρεωτικά θα έπρεπε να επιλέξουμε να εξετάσουμε την επίδραση του παράγοντα treatment σε όλες τις στάθμες του παράγοντα time, ώστε με την βοήθεια και του τελευταίου ελέγχου να απαντούσαμε στο ερώτημα αυτό. Επομένως η διατύπωση του προβλήματος και το ερώτημα που τίθεται ενδεχομένως να παίξει σημαντικό ρόλο στην επιλογή που θα κάνουμε.

Όπως αναφέρθηκε και στην αρχή του κεφαλαίου, η σημαντικότητα του όρου της αλληλεπίδρασης παίζει σπουδαίο ρόλο, καθώς η μετέπειτα ανάλυση εξαρτάται από αυτόν. Γι'αυτόν τον λόγο θα γίνει η υπόθεση ότι ο όρος αυτός δεν είναι στατιστικά σημαντικός, έτσι ώστε να παρουσιάσουμε συνοπτικά και την άλλη περίπτωση.

Επομένως με βάση τον πρώτο έλεγχο mixed.anova, αποδεικνύεται ότι και οι δυο κύριες επιδράσεις είναι στατιστικά σημαντικές, αφού $p < 0.05$. Δηλαδή για κάθε παράγοντα ξεχωριστά, απομένει να γίνουν πολλαπλές συγκρίσεις κατά ζεύγη χρησιμοποιώντας την διόρθωση κατά Bonferroni.

```
> depression %>%
+ pairwise_t_test(score~time paired=TRUE,
+ p.adjust.method="bonferroni")

> depression %>%
+ pairwise_t_test(score~treatment,p.adjust.method="bonferroni")

# A tibble: 6 x 6
  .y.   group1 group2    n1    n2       p.adj
  <chr> <chr>  <chr>  <int> <int>       <dbl>
1 score t0     t1        24    24 0.0000537
2 score t0     t2        24    24 0.000000243
3 score t0     t3        24    24 0.00000478
4 score t1     t2        24    24 0.518
5 score t1     t3        24    24 1
6 score t2     t3        24    24 1
```

Ενδεικτικά παρουσιάσουμε τα αποτελέσματα μόνο του πρώτου ελέγχου, τα οποία και θα ερμηνεύσουμε λαμβάνοντας υπόψη τις p.adj τιμές της διόρθωσης κατά Bonferroni.

- Η μέση βαθμολογία της κατάθλιψης είναι στατιστικά σημαντικά διαφορετική μεταξύ των χρονικών στιγμών $\mathbf{t}_0$ vs $\mathbf{t}_1$, $\mathbf{t}_0$ vs $\mathbf{t}_2$, $\mathbf{t}_0$ vs $\mathbf{t}_3$.

- Αντίθετα η μέση βαθμολογία της κατάθλιψης δεν διαφέρει στατιστικά σημαντικά μεταξύ των χρονικών στιγμών $\mathbf{t}_1$ vs $\mathbf{t}_2$, $\mathbf{t}_1$ vs $\mathbf{t}_3$, $\mathbf{t}_2$ vs $\mathbf{t}_3$.

4.5 Μικτή ANOVA τριών Παραγόντων με τη βοήθεια της R

Στην συγκεκριμένη ενότητα θα ασχοληθούμε με την ανάλυση της μικτής ANOVA τριών παραγόντων με την βοήθεια της R. Όπως και μια μικτή αναλύση διασποράς δυο παραγόντων, έτσι και η συγκεκριμένη έχει κάποια βασικά βήματα που πρέπει να ακολουθήσουμε, ώστε να πραγματοποιήσουμε την στατιστική μας ανάλυση και να καταλήξουμε στα ζητούμενα συμπεράσματα, τα οποία και θα αναφέρουμε στην συνέχεια. Καθώς μια μικτή ANOVA τριών παραγόντων έχει την δυνατότητα είτε να έχει δυο παράγοντες εντός των υποκειμένων και ένα μεταξύ, είτε το αντίθετο, θα μελετήσουμε και τις δυο περιπτώσεις με την δεύτερη να γίνεται πιο επιγραμματικά.

4.5.1 Μεθοδολογία

Αντίστοιχα με την περίπτωση της μικτής ANOVA δυο παραγόντων, τα βήματα είναι συγκεκριμένα και αποτελούν οδηγό για κάθε μικτή ANOVA τριών παραγόντων. Ωστόσο προτού χρησιμοποιηθεί ο παρακάτω οδηγός, πρέπει να έχουν επαληθευτεί οι υποθέσεις που απαιτούνται για να πραγματοποιηθεί μία τέτοια ανάλυση. Τα βήματα είναι τα εξής:

- Βήμα 1: Ο πρώτος έλεγχος αφορά τον όρο της αλληλεπίδρασης των τριών παραγόντων και είναι το πιο σημαντικό βήμα καθώς η μετέπειτα ανάλυση εξαρτάται από αυτόν.

Ο όρος της αλληλεπίδρασης των τρίων παραγόντων είναι στατιστικά σημαντικός.

- Βήμα 2*i*: Στην περίπτωση που ο όρος της αλληλεπίδρασης είναι στατιστικά σημαντικός, για τον επόμενο έλεγχο επιλέγουμε μια αμφίδρομη αλληλεπίδραση (έστω ΑΒ) και εξετάζουμε την επίδραση της σε κάθε στάθμη του παράγοντα Γ. Επομένως καταλήγουμε στις στάθμες του παράγοντα Γ, στις οποίες είναι στατιστικά σημαντική η απλή αμφίδρομη αλληλεπίδραση ΑΒ.

- Βήμα 3*i*: Έπειτα θα πραγματοποιήσουμε μία απλή ανάλυση κύριων επιδράσεων. Πιο συγκεκριμένα θα χρησιμοποιήσουμε τις στάθμες του παράγοντα Γ, που με βάση τα αποτελέσματα του βήματος 2*i* θεωρήθηκαν ως στατιστικά σημαντικές. Στις συγκεκριμένες στάθμες θα εξέτασουμε την επίδραση του παράγοντα Α σε όλες τις στάθμες του παράγοντα Β, καταλήγοντας με αυτόν τον τρόπο στις στάθμες του παράγοντα Β, στις οποίες είναι στατιστικά σημαντική η απλή κύρια επίδραση Α.

- Βήμα 4*i*: Στο τελευταίο βήμα θα χρησιμοποιήσουμε μόνο τις στάθμες των παραγόντων Γ, Β που με βάση τα βήματα 2*i*, 3*i* αντίστοιχα, θεωρήθηκαν ως στατιστικά σημαντικές. Στις συγκεκριμένες αγωγές των δυο παραγόντων, θα πραγματοποιήσουμε πολλαπλές συγκρίσεις κατά ζεύγη μεταξύ των υποομάδων του παράγοντα Α, ώστε να καταλήξουμε εν τέλει σε ποιές αγωγές και για ποιές υποομάδες του παράγοντα Α εντοπίζεται διαφορά.

Ο όρος της αλληλεπίδρασης των τριών παραγόντων δεν είναι στατιστικά σημαντικός.

- Βήμα $2ii$: Στην περίπτωση που ο όρος της αλληλεπίδρασης δεν είναι στατιστικά σημαντικός, τότε το πρόβλημα μετατρέπεται σε μικτή ANOVA δυο παραγόντων και είμαστε σε θέση να ακολουθήσουμε όσα αναφέραμε στην Ενότητα 4.4.

Για τα βήματα $2i$, $3i$ τόσο η επιλογή της αμφίδρομης αλληλεπίδρασης όσο και της κύριας επίδρασης είναι ελεύθερη, εκτός και αν μέσω του ερωτήματος του προβλήματος ζητείτε κάτι συγκεκριμένο, οπότε καλούμαστε να επιλέξουμε συγκεκριμένη αλληλεπίδραση και κύρια επίδραση ώστε να απαντήσουμε στο ερώτημα αυτό.

4.5.2 Ανάλυση Μικτής ANOVA τριών Παραγόντων με δυο Παράγοντες εντός των Υποκειμένων

Η πρώτη περίπτωση που θα ασχοληθούμε θα αφορά την μικτή ANOVA τριών παραγόντων με δυο παράγοντες εντός των υποκειμένων και ένα μεταξύ. Το παράδειγμα με το οποίο θα ασχοληθούμε είναι το εξής:

Παράδειγμα 4.5.2

Τα δεδομένα του παραδείγματος έχουν προσομοιωθεί (επομένως είναι φανταστικά) με σκοπό να μπορούν να ανταποκριθούν σε μια μικτή ανάλυση διασποράς τριών παραγόντων με δυο παράγοντες εντός των υποκειμένων. Το πρόβλημα αντιστοιχεί σε ένα πείραμα στο οποίο μελετάται η επίδραση δυο διαφορετικών τύπων ασκήσεων στο άγχος, λαμβάνοντας υπόψη τις διαφορές μεταξύ των δυο φύλων. Έχουμε λοιπόν δυο ειδών ασκήσεις, A και B, με κάθε συμμετέχων να πραγματοποιεί και τις δυο ασκήσεις διαδοχικά, πρώτα την A και έπειτα την B, καταγράφοντας τις τιμές σε διάφορες χρονικές στιγμές. Πιο συγκεκριμένα οι συμμετέχοντες παρακολουθήθηκαν σε τρείς διαφορετικές χρονικές στιγμές (t_1: 10 ημέρες μετά την έναρξη του πειράματος, t_2: 20 ημέρες μετά την έναρξη του πειράματος, t_3: 30 ημέρες μετά την έναρξη του πειράματος) και σε κάθε χρονική στιγμή καταγράφηκε η βαθμολογία του άγχους. Για την καταγραφή της βαθμολογίας του άγχους χρησιμοποιήθηκε ένα τυποποιημένο ερωτηματολόγιο, με τιμές που κυμαίνονται από 10 έως 100. Οι υψηλότερες τιμές στην κλίμακα αυτή υποδηλώνουν πιο σοβαρά συμπτώματα άγχους. Τέλος καλούμαστε να καταλήξουμε εάν οποιαδήποτε αλλαγή της βαθμολογίας του άγχους είναι αποτέλεσμα της αλληλεπίδρασης των τριών παραγόντων, όπως επίσης να εξετάσουμε και τις διαφορές που ενδέχεται να προκύψουν για την μέση βαθμολογία του άγχους, μεταξύ των διαφορετικών χρονικών στιγμών ανάλογα με το φύλο και το είδος της άσκησης.

Οι τρείς παράγοντες που θα εξετάσουμε στο πρόβλημα μας, είναι το φύλο (gender), ο χρόνος (time) και το είδος της άσκησης (exercise), με τον πρώτο παράγοντα να αποτελεί τον παράγοντα μεταξύ των υποκειμένων, ενώ τους δυο τελευταίους τους παράγοντες εντός των υποκειμένων. Προφανώς η εξαρτημένη μεταβλητή αποτελείται από την βαθμολογία του άγχους (score), ενώ ονομάζουμε τα δεδομένα ως stress.

Έχοντας συμπεριλάβει τους τρείς αυτούς παράγοντες και όλες τις πιθανές αλληλεπιδράσεις, το μοντέλο μας μπορεί να εκφραστεί ως εξής:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl},$$

$$i = 1, 2, j = 1, 2, 3, k = 1, 2, l = 1, 2, ..., 120,$$

όπου μ είναι η σταθερά, α_i είναι η κύρια επίδραση της i στάθμης του παράγοντα φύλου, β_j είναι η κύρια επίδραση της j στάθμης του παράγοντα χρόνου, γ_k είναι η κύρια επίδραση της k στάθμης του παράγοντα άσκηση, $(\alpha\beta)_{ij}$ είναι η αλληλεπίδραση που οφείλεται στον συνδυασμό των δυο παραγόντων φύλου και χρόνου στην αγωγή (ij), $(\alpha\gamma)_{ik}$ είναι η αλληλεπίδραση που οφείλεται στον συνδυασμό των δυο παραγόντων φύλου και άσκηση στην αγωγή (ik), $(\beta\gamma)_{jk}$ είναι η αλληλεπίδραση που οφείλεται στον συνδυασμό των δυο παραγόντων χρόνου και άσκησης στην αγωγή (jk), $(\alpha\beta\gamma)_{ijk}$ είναι η αλληλεπίδραση που οφείλεται στον συνδυασμό των τριών παραγόντων φύλου, χρόνου και άσκησης στην αγωγή (ijk) και ε_{ijkl} είναι το τυχαίο σφάλμα της l παρατήρησης της εξαρτημένης μεταβλητής y στην αγωγή (ijk).

Στο πείραμα μας συμμετέχουν συνολικά 20 άτομα, 10 άνδρες και 10 γυναίκες, οι οποίοι πρώτα ακολουθούν την άσκηση A και καταγράφουν την βαθμολογία στις τρείς διαφορετικές χρονικές στιγμές και έπειτα ακολουθούν την άσκηση B εκτελώντας ακριβώς την ίδια διαδικάσια. Ενδεικτικά θα παρουσιάσουμε τις 20 πρώτες από τις συνολικά 120 τιμές δεδομένων.

```
    id gender time exercise    score
1    1      M   t1        A 41.50072
2    2      M   t1        A 39.30003
3    3      M   t1        A 42.93013
4    4      M   t1        A 42.87831
5    5      M   t1        A 42.70582
6    6      M   t1        A 42.30035
7    7      M   t1        A 41.88945
8    8      M   t1        A 40.01117
9    9      M   t1        A 39.26681
10  10      M   t1        A 39.03956
11  11      F   t1        A 43.07726
12  12      F   t1        A 41.51939
13  13      F   t1        A 40.43556
14  14      F   t1        A 37.40233
15  15      F   t1        A 44.36228
16  16      F   t1        A 37.49909
17  17      F   t1        A 47.25567
18  18      F   t1        A 44.96414
19  19      F   t1        A 38.77505
20  20      F   t1        A 36.00753
```

Ο κώδικας για την προσομοίωση των δεδομένων μας, όπως επίσης και το σύνολο των 120 γραμμών των δεδομένων μας, παρουσιάζεται στο τέλος της εργασίας στην ειδική ενότητα Παράρτημα. Στην συνέχεια θα παρουσιάσουμε ένα πίνακα με την μέση τιμή και την τυπική απόκλιση κάθε αγωγής.

```
> group_by(stress,time,exercise,gender) %>%
+ get_summary_stats(score,type="mean_sd")

# A tibble: 12 x 7
   gender time  exercise variable     n  mean    sd
   <fct>  <fct> <fct>    <chr>    <dbl> <dbl> <dbl>
 1 F      t1    A        score       10  41.1  3.74
 2 M      t1    A        score       10  41.2  1.61
 3 F      t1    B        score       10  35.4  2.61
 4 M      t1    B        score       10  32.9  2.48
 5 F      t2    A        score       10  39.7  3.03
 6 M      t2    A        score       10  34.2  3.02
 7 F      t2    B        score       10  33.0  3.26
 8 M      t2    B        score       10  33.2  3.96
 9 F      t3    A        score       10  43.5  1.37
10 M      t3    A        score       10  51.4  3.93
11 F      t3    B        score       10  38.5  2.33
12 M      t3    B        score       10  33.1  3.63
```

Τέλος προτού προχωρήσουμε στην επαλήθευση των υποθέσεων θα χρειαστούμε τις αντίστοιχες βιβλιοθήκες που χρησιμοποιήθηκαν για την μικτή ANOVA δυο παραγόντων, όπως επίσης θα πρέπει και να μετατρέψουμε τις τέσσερις πρώτες στήλες (id, gender, time , exercise) σε παράγοντες.

Επαλήθευση των Υποθέσεων

Ακραίες Τιμές

Ο έλεγχος των ακραίων τιμών γίνεται ως εξής:

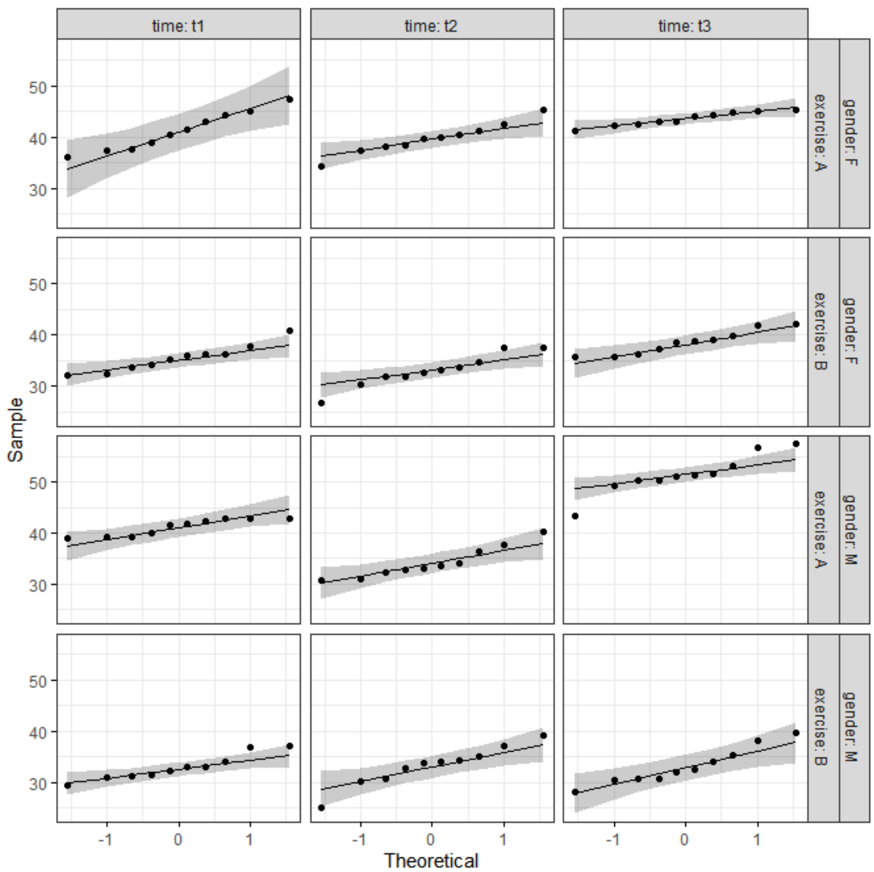
```
> group_by(stress,gender,exercise,time) %>%
+ identify_outliers(score)

# A tibble: 7 x 7
  gender time  exercise    id score is.outlier is.extreme
  <fct>  <fct> <fct>    <fct> <dbl> <lgl>      <lgl>
1 F      t2    A           20  45.3 TRUE       FALSE
2 F      t1    B           20  40.8 TRUE       FALSE
3 F      t2    B           12  26.6 TRUE       FALSE
4 M      t3    A            4  56.7 TRUE       FALSE
5 M      t3    A            6  57.4 TRUE       FALSE
6 M      t3    A            7  43.3 TRUE       FALSE
7 M      t2    B            8  24.9 TRUE       FALSE
```

Από τα παραπάνω αποτελέσματα συμπεραίνουμε ότι δεν υπάρχουν ακραίες τιμές στα δεδομένα μας.

Υπόθεση Κανονικότητας

Για την υπόθεση της κανονικότητας θα μας βοηθήσει το διάγραμμα Q-Q.

```
> ggqqplot(stress,"score",ggtheme=theme_bw())+
+ facet_grid(gender+exercise~time,labeller="label_both")
```



Διάγραμμα 4.6: Διάγραμμα Q-Q για τα δεδομένα του Παραδείγματος της Υποενότητας 4.5.3.

Με την βοήθεια του Διαγράμματος 4.6, κατά προσέγγιση συμπεραίνουμε ότι τα δεδομένα μας είναι κανονικά κατανεμημένα για κάθε αγωγή του προβλήματος, αφού η πλειοψηφία των σημείων βρίσκεται κοντά στην ιδανική ευθεία.

Ομοιογένεια των Διαφορών

Η εξέταση της ομοιογένειας των διαφορών θα γίνει με την βοήθεια του Levene's ελέγχου. Ο έλεγχος θα πραγματοποιηθεί σε κάθε αγωγή των παραγόντων exercise, time, οι οποίοι αποτελούν και τους παράγοντες εντός των υποκειμένων.

```
> group_by(stress,exercise,time) %>%
+ levene_test(score~gender)

# A tibble: 6 x 6
  time  exercise   df1   df2 statistic       p
  <fct> <fct>    <int> <int>     <dbl>   <dbl>
1 t1    A            1    18  7.39      0.0141
2 t2    A            1    18  0.000600  0.981
3 t3    A            1    18  2.46      0.134
4 t1    B            1    18  0.0163    0.900
5 t2    B            1    18  0.106     0.748
6 t3    B            1    18  1.11      0.305
```

Με την βοήθεια του παραπάνω ελέγχου και λαμβάνοντας υπόψη τις p-τιμές, παρατηρούμε ότι οι στάθμες του παράγοντα gender έχουν ίσες διακυμάνσεις σε κάθε αγωγή των δυο άλλων παραγόντων, εκτός της αγωγής "t_1-A", δηλαδή την χρονική στιγμή t_1 για την άσκηση A. Επομένως αφού οι υποθέσεις μας επαληθεύτηκαν είμαστε σε θέση να περάσουμε στην κύρια ανάλυση.

Κύρια Ανάλυση

Ο πρώτος έλεγχος, που είναι ένας έλεγχος μικτής ANOVA τριών κατευθύνσεων, γίνεται με σκοπό να εξετάσουμε την στατιστική σημαντικότητα του όρου αλληλεπίδρασης των τριών παραγόντων. Ο έλεγχος που πραγματοποιείται είναι ο εξής:

```
> mixed.anova<-anova_test(data=stress,dv=score,
+ wid=id,between=gender,within=c(time,exercise))
> get_anova_table(mixed.anova)

ANOVA Table (type II tests)

                Effect DFn DFd      F        p p<.05   ges
1               gender   1  18  4.003 6.10e-02       0.023
2                 time   2  36 51.377 2.84e-11     * 0.473
```

```
3              exercise   1  18  168.645  1.40e-10     * 0.629
4           gender:time   2  36    4.592  1.70e-02     * 0.074
5       gender:exercise   1  18    8.723  9.00e-03     * 0.081
6         time:exercise   2  36   14.002  3.17e-05     * 0.236
7  gender:time:exercise   2  36   20.772  1.00e-06     * 0.315
```

Από τα παραπάνω αποτελέσματα και σε ότι αφορά την αλληλεπίδραση των τριών παραγόντων, παρατηρούμε ότι μπορούμε να απορρίψουμε την μηδενική υπόθεση, αφού p$<$ 0.05. Επομένως καταλήγουμε ότι η αλληλεπίδραση των τριών παραγόντων είναι στατιστικά σημαντική και επηρεάζει την τιμή της βαθμολογίας του άγχους. Είναι λοιπόν περιττό να ελέγξουμε τις αμφίδρομες αλληλεπιδράσεις.

Για τον επόμενο έλεγχο επιλέγουμε να εξετάσουμε την επίδραση της απλής αμφίδρομης αλληλεπίδρασης gender-time στις στάθμες του παράγοντα exercise. Ο έλεγχος που πραγματοποιείται είναι ο εξής:

```
> two.way<-group_by(stress,exercise) %>%
+ anova_test(dv=score,wid=id,within=time,between=gender)
> get_anova_table(two.way)

# A tibble: 6 x 7
  exercise Effect        DFn   DFd     F        p    ges
  <fct>    <chr>       <dbl> <dbl> <dbl>    <dbl>  <dbl>
1 A        gender       1      18    1.40 2.53e- 1  0.021
2 A        time         2      36   59.2  4.18e-12  0.705
3 A        gender:time  2      36   24.1  2.27e- 7  0.493
4 B        gender       1      18   11.6  3    e- 3  0.162
5 B        time         1.45  26.0   3.75 5    e- 2  0.127
6 B        gender:time  1.45  26.0   3.83 4.7  e- 2   0.13
```

Αξίζει να σημειωθεί ότι η στατιστική σημαντικότητα της απλής αμφίδρομης αλληλεπίδρασης, γίνεται αποδεκτή σε επίπεδο σημαντικότητας 0.025, προσαρμοσμένο κατά Bonferroni. Στην περίπτωση μας, διεξάγουμε δυο ελέγχους στις στάθμες του παράγοντα exercise για την απλή αμφίδρομη αλληλεπίδραση, επομένως διαιρούμε το αρχικό επίπεδο σημαντικότητας 0.05 με τον αριθμό των ελέγχων, δηλαδή το δυο. Επομένως με αυτόν τον τρόπο καταλήγουμε στο νέο επίπεδο σημαντικότητας 0.025.

Από τα παραπάνω αποτελέσματα και λαμβάνοντας υπόψη τις p-τιμές και το νέο επίπεδο σημαντικότητας, συμπεραίνουμε ότι η απλή αμφίδρομη αλληλεπίδραση gender-time είναι στατιστικά σημαντική για την άσκηση A του παράγοντα exercise, ενώ δεν είναι στατιστικά σημαντική για την άσκηση B. Στην συνέχεια θα εξετάσουμε την επίδραση του παράγοντα time στις στάθμες του παράγοντα gender, παρουσιάζοντας όμως μόνο τα αποτελέσματα για την άσκηση A αφού μόνο αυτά μας αφορούν.

```
> time.effect<-group_by(stress,exercise,gender) %>%
+ anova_test(dv=score,wid=id,within=time) %>%
+ get_anova_table()
> time.effect

# A tibble: 2 x 8
  gender exercise Effect   DFn   DFd      F         p    ges
  <fct>  <fct>    <chr>  <dbl> <dbl>  <dbl>     <dbl>  <dbl>
1 F      A        time    1.16  10.5   3.38  9.1 e- 2  0.246
2 M      A        time    2     18    92.9   3.26e-10  0.86
```

Αντίστοιχα με τον προηγούμενο έλεγχο και στον συγκεκριμένο η στατιστική σημαντικότητα της απλής κύριας επίδρασης γίνεται αποδεκτή σε επίπεδο σημαντικότητας 0.025, προσαρμοσμένο κατά Bonferroni. Στην περίπτωση μας, διεξάγουμε δυο ελέγχους στις στάθμες του παράγοντα gender για την απλή κύρια επίδραση time, επομένως διαιρούμε το αρχικό επίπεδο σημαντικότητας 0.05 με τον αριθμό των ελέγχων, δηλαδή το δυο. Επομένως με αυτόν τον τρόπο καταλήγουμε στο νέο επίπεδο σημαντικότητας 0.025.

Από τα παραπάνω αποτελέσματα και λαμβάνοντας υπόψη τις p-τιμές και το νέο επίπεδο σημαντικότητας, διαπιστώνουμε ότι για την άσκηση A, η απλή κύρια επίδραση του παράγοντα time είναι στατιστικά σημαντική για το αντρικό φύλο, ενώ δεν είναι στατιστικά σημαντική για το γυναικείο. Ο τελευταίος έλεγχος που θα πραγματοποιηθεί θα έχει ως σκοπό την εύρεση των χρονικών στιγμών που εντοπίζεται η διαφορά στην μέση βαθμολογία του άγχους. Σύμφωνα με τους προηγούμενους ελέγχους, τα αποτελέσματα που θα παρουσιάσουμε θα είναι μόνο για το αντρικό φύλο όταν πραγματοποιεί την άσκηση A.

```
> test<-group_by(stress,gender,exercise) %>%
+ pairwise_t_test(score~time,paired=TRUE,
+ p.adjust.method="bonferroni") %>%
+ select(-statistic,-df ,-p,-p.adj.signif)
> test

# A tibble: 3 x 8
   gender exercise .y.   group1 group2    n1    n2      p.adj
 * <fct>  <fct>    <chr> <chr>  <chr>  <int> <int>      <dbl>
 1 M      A        score t1     t2        10    10 0.00027
 2 M      A        score t1     t3        10    10 0.0000729
 3 M      A        score t2     t3        10    10 0.00000238
```

Με την βοήθεια του παραπάνω ελέγχου και λαμβάνοντας υπόψη τις p.adj τιμές της διόρθωσης κατά Bonferroni καταλήγουμε στα εξής συμπεράσματα:

- Η μέση βαθμολογία του άγχους για τους **άνδρες** όταν πραγματοποιούν την άσκηση **A**, είναι στατιστικά σημαντικά διαφορετική μεταξύ των χρονικών στιγμών $\mathbf{t}_1$ vs $\mathbf{t}_2$, $\mathbf{t}_1$ vs $\mathbf{t}_3$, $\mathbf{t}_2$ vs $\mathbf{t}_3$.

Με την ανάλυση των αποτελεσμάτων του παραπάνω ελέγχου έχουμε ουσιαστικά ολοκληρώσει την ανάλυση μας. Αντίστοιχα με την μικτή ANOVA δυο παραγόντων, μπορούμε να κατασκευάσουμε ένα θηκοδιάγραμμα για να παρατηρήσουμε όσα αποδείξαμε προηγουμένως.

Αν για τις αγωγές που διαφέρουν μεταξύ τους, χρειάζεται να αποφανθούμε για ποιά από τις τρείς η μέση βαθμολογία του άγχους είναι μεγαλύτερη ή μικρότερη, μπορεί εύκολα να γίνει με την χρήση ενός θηκοδιαγράμματος ή διαφορετικά με την βοήθεια του πίνακα που κατασκευάστηκε νωρίτερα και περιέχει την μέση τιμή και την τυπική απόκλιση των παρατηρήσεων για κάθε αγωγή του προβλήματος.

Δηλαδή λαμβάνοντας υπόψη τις τιμές του πίνακα, η μέση βαθμολογία του άγχους για τους άνδρες όταν πραγματοποιούν την άσκηση A, λαμβάνει την μεγαλύτερη τιμή την χρονική στιγμή t_3, έπειτα την χρονική στιγμή t_2, ενω ακολουθεί τελευταία η χρονική στιγμή t_1.

4.5.3 Ανάλυση Μικτής ANOVA τριών Παραγόντων με δυο Παράγοντες μεταξύ των Υποκειμένων

Η δεύτερη και ταυτόχρονα τελευταία περίπτωση που θα ασχοληθούμε, θα αφορά την μικτή ANOVA τριών παραγόντων με δυο παράγοντες μεταξύ των υποκειμένων και έναν εντός. Το παράδειγμα με το οποίο θα ασχοληθούμε είναι το εξής:

Παράδειγμα 4.5.3

Τα δεδομένα του παραδείγματος έχουν προσομοιωθεί (επομένως είναι φανταστικά) με σκοπό να είναι σε θέση να ανταποκριθούν σε μια μικτή ανάλυση διασποράς τριών παραγόντων με δυο παράγοντες μεταξύ των υποκειμένων. Το πρόβλημα αντιστοιχεί σε ένα πείραμα στο οποίο μελετάται η επίδραση δυο διαφορετικών προγραμμάτων εκγύμνασης σε αθλητές της καλαθοσφαίρισης, για την βελτίωση του ύψους του άλματος. Έχουμε λοιπόν δυο διαφορετικά ειδικά προγράμματα εκγύμνασης (για ευκολία και συντομία θα αναφέρονται ως προγράμματα A, B) και κάθε συμμετέχων πραγματοποιεί ένα από τα δυο προγράμματα. Οι συμμετέχοντες παρακολουθήθηκαν σε δυο διαφορετικές χρονικές στιγμές (t_1: 5 μέρες μετά την έναρξη του προγράμματος εκγύμνασης, t_2: 30 μέρες μετά την έναρξη του προγράμματος εκγύμνασης) και σε κάθε χρονική στιγμή καταγράφηκε το ύψος του άλματος τους σε εκατοστά. Επίσης λαμβάνεται υπόψη και η ηλικία των συμμετεχόντων στο πείραμα, χωρίζοντας τους σε τρία ηλικιακά γκρούπ τα οποία είναι τα εξής: 18-23, 24-29, 30-35. Τέλος καλούμαστε να καταλήξουμε εάν οποιαδήποτε αλλαγή στο ύψος του άλματος είναι αποτέλεσμα της αλληλεπίδρασης των τριών παραγόντων, όπως επίσης να εξετάσουμε και τις διαφορές που ενδέχεται να προκύψουν για το μέσο ύψος του άλματος, μεταξύ των τριών ηλικιακών γκρουπ ανάλογα με την χρονική στιγμή και το πρόγραμμα εκγύμνασης που ακολουθούν.

Οι τρείς παράγοντες που θα χρησιμοποιήσουμε στο πρόβλημα μας είναι το ηλικιακό γκρουπ (agegroup), το πρόγραμμα εκγύμνασης (program) και ο χρόνος (time) με τους δυο πρώτους παράγοντες να αποτελούν τους παράγοντες μεταξύ των υποκειμένων, ενώ τον τελευταίο να αποτελεί τον παράγοντα εντός των υποκειμένων. Προφανώς η εξαρτημένη μεταβλητή αποτελείται από το ύψος του άλματος (height), ενώ ονομάζουμε τα δεδομένα ως df.

Έχοντας συμπεριλάβει τους τρείς αυτούς παράγοντες και όλες τις πιθανές αλληλεπιδράσεις, το μοντέλο μας μπορεί να εκφραστεί ως εξής:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl},$$
$$i = 1, ..., 3, j = 1, 2, k = 1, 2, l = 1, 2, ..., 120,$$

όπου μ είναι η σταθερά, α_i είναι η κύρια επίδραση της i στάθμης του παράγοντα ηλικιακό γκρουπ, β_j είναι η κύρια επίδραση της j στάθμης του παράγοντα πρόγραμμα εκγύμνασης, γ_k είναι η κύρια επίδραση της k στάθμης του παράγοντα χρόνου, $(\alpha\beta)_{ij}$ είναι η αλληλεπίδραση που οφείλεται στον συνδυασμό των δυο παραγόντων ηλικιακό γκρουπ και πρόγραμμα εκγύμνασης στην αγωγή (ij), $(\alpha\gamma)_{ik}$ είναι η αλληλεπίδραση που οφείλεται στον συνδυασμό των δυο παραγόντων ηλικιακό γκρουπ και χρόνου στην αγωγή (ik), $(\beta\gamma)_{jk}$ είναι η αλληλεπίδραση που οφείλεται στον συνδυασμό των δυο παραγόντων πρόγραμμα εκγύμνασης και χρόνου στην αγωγή (jk), $(\alpha\beta\gamma)_{ijk}$ είναι η αλληλεπίδραση που οφείλεται στον συνδυασμό των τριών παραγόντων ηλικιακό γκρουπ, πρόγραμμα εκγύμνασης και χρόνου στην αγωγή (ijk) και ε_{ijkl} είναι το τυχαίο σφάλμα της l παρατήρησης της εξαρτημένης μεταβλητής y στην αγωγή (ijk).

Στο πείραμα συμμετέχουν συνολικά 60 αθλητές με τους μισούς να ακολουθούν το πρόγραμμα A και τους άλλους μισούς το πρόγραμμα B, έχοντας επίσης 20 αθλητές σε κάθε ηλικιακό γκρουπ. Ενδεικτικά θα παρουσιάσουμε τις 20 πρώτες από τις συνολικά 120 τιμές δεδομένων, οι οποίες είναι οι εξής:

```
   id program agegroup time   height
1   1       A    18-23   t1 81.99862
2   2       A    18-23   t1 86.91579
3   3       A    18-23   t1 83.50700
4   4       A    18-23   t1 82.41670
5   5       A    18-23   t1 83.74661
6   6       A    18-23   t1 83.82926
7   7       A    18-23   t1 86.59186
8   8       A    18-23   t1 84.42510
9   9       A    18-23   t1 85.83972
10 10       A    18-23   t1 83.19073
11 11       A    24-29   t1 85.77187
12 12       A    24-29   t1 84.53678
13 13       A    24-29   t1 85.49728
14 14       A    24-29   t1 83.22941
15 15       A    24-29   t1 82.27768
16 16       A    24-29   t1 89.85602
17 17       A    24-29   t1 86.65676
18 18       A    24-29   t1 82.41406
19 19       A    24-29   t1 83.88048
20 20       A    24-29   t1 82.56478
```

Ο κώδικας για την προσομοίωση των δεδομένων μας, όπως επίσης και το σύνολο των 120 γραμμών των δεδομένων μας, παρουσιάζεται στο τέλος της εργασίας στην ειδική ενότητα Παράρτημα. Στην συνέχεια θα παρουσιάσουμε ένα πίνακα με την μέση τιμή και την τυπική απόκλιση κάθε αγωγής.

```
> group_by(df,time,agegroup,program) %>%
+ get_summary_stats(height,type="mean_sd")

# A tibble: 12 x 7
   program agegroup time  variable      n  mean    sd
   <fct>   <fct>    <fct> <chr>     <dbl> <dbl> <dbl>
 1 A       18-23    t1    height       10  84.2  1.69
 2 B       18-23    t1    height       10  78.7  3.10
 3 A       24-29    t1    height       10  84.7  2.37
 4 B       24-29    t1    height       10  85.0  2.46
 5 A       30-35    t1    height       10  75.0  2.28
 6 B       30-35    t1    height       10  83.1  2.89
 7 A       18-23    t2    height       10  84.2  2.28
 8 B       18-23    t2    height       10  83.3  1.29
 9 A       24-29    t2    height       10  84.6  3.85
10 B       24-29    t2    height       10  87.2  3.78
11 A       30-35    t2    height       10  84.3  3.05
12 B       30-35    t2    height       10  81.5  1.85
```

Η επιβεβαίωση των υποθέσεων έχει γίνει ήδη μια φορά για την μικτή ANOVA τριών παραγόντων με δυο παράγοντες εντός των υποκειμένων. Η διαδικασία και στην περίπτωση που έχουμε δυο παράγοντες μεταξύ των υποκειμένων είναι αντίστοιχη και λόγω του γεγονότος πως δεν έχει να μας προσφέρει κάτι καινούργιο, θεωρείται σκόπιμο να παραλειφθεί. Η μόνη αναφορά που θα γίνει θα είναι για τον Levene's έλεγχο, ο οποίος πρέπει να πραγματοποιηθεί και για τους δυο παράγοντες μεταξύ των υποκειμένων, δηλαδή για το όρισμα agegroup*program. Τέλος οφείλουμε να αναφέρουμε ότι τα δεδομένα μας επαληθεύουν τις υποθέσεις.

Κύρια Ανάλυση

Ο πρώτος έλεγχος, που είναι ένας έλεγχος μικτής ANOVA τριών κατευθύνσεων, γίνεται με σκοπό να εξετάσουμε την στατιστική σημαντικότητα του όρου αλληλεπίδρασης των τριών παραγόντων. Ο έλεγχος που πραγματοποιείται είναι ο εξής:

```
> mixed.anova<-anova_test(data=df,dv=height,
+ wid=id,within=time,between=c(agegroup,program))
> get_anova_table(mixed.anova)
```

```
ANOVA Table (type II tests)

                 Effect DFn DFd      F        p p<.05   ges
1              agegroup   2  54 22.969 6.05e-08     * 0.337
2               program   1  54  0.276 6.01e-01       0.003
3                  time   1  54 30.658 9.31e-07     * 0.186
4      agegroup:program   2  54 11.333 7.77e-05     * 0.201
5         agegroup:time   2  54  3.371 4.20e-02     * 0.048
6          program:time   1  54  2.203 1.44e-01       0.016
7 agegroup:program:time   2  54 30.067 1.68e-09     * 0.309
```

Από τα παραπάνω αποτελέσματα και σε ότι αφορά την αλληλεπίδραση των τριών παραγόντων, παρατηρούμε ότι μπορούμε να απορρίψουμε την μηδενική υπόθεση, αφού $p < 0.05$. Επομένως καταλήγουμε ότι η αλληλεπίδραση των τριών παραγόντων είναι στατιστικά σημαντική και επηρεάζει την τιμή του ύψους του άλματος. Είναι λοιπόν περιττό να ελέγξουμε τις αμφίδρομες αλληλεπιδράσεις.

Για τον επόμενο έλεγχο επιλέγουμε να εξετάσουμε την επίδραση της απλής αμφίδρομης αλληλεπίδρασης agegroup-program στις στάθμες του παράγοντα time.

```
> two.way<-group_by(df,time) %>%
+ anova_test(dv=height,wid=id,between=c(agegroup,program))
> get_anova_table(two.way)

# A tibble: 6 x 6
  time  Effect              DFn   DFd     F        p
* <fct> <chr>             <dbl> <dbl> <dbl>    <dbl>
1 t1    agegroup              2    54 27.0   7.59e- 9
2 t1    program               1    54  2.08  1.55e- 1
3 t1    agegroup:program      2    54 37.5   6.19e-11
4 t2    agegroup              2    54  5.92  5   e- 3
5 t2    program               1    54  0.253 6.17e- 1
6 t2    agegroup:program      2    54  4.41  1.7 e- 2
```

Αντίστοιχα με την Ενότητα 4.5.2, το επίπεδο σημαντικότητας και εδώ είναι 0.025. Επομένως από τα παραπάνω αποτελέσματα και λαμβάνοντας υπόψη τις p-τιμές και το νέο επίπεδο σημαντικότητας, συμπεραίνουμε ότι η απλή αμφίδρομη αλληλεπίδραση agegroup-program είναι στατιστικά σημαντική και στις δυο στάθμες του παράγοντα time, t_1 και t_2. Στην συνέχεια θα εξετάσουμε την επίδραση του παράγοντα agegroup στις στάθμες του παράγοντα program ως εξής:

```
> agegroup.effect<-group_by(df,time,program) %>%
+ anova_test(dv=height,wid=id,between=agegroup) %>%
+ get_anova_table()
> agegroup.effect

# A tibble: 4 x 7
  program time  Effect     DFn   DFd      F        p
* <fct>   <fct> <chr>    <dbl> <dbl>  <dbl>    <dbl>
1 A       t1    agegroup     2    27 65.9   4.12e-11
2 B       t1    agegroup     2    27 13.1   1.08e- 4
3 A       t2    agegroup     2    27  0.053 9.48e- 1
4 B       t2    agegroup     2    27 12.9   1.18e- 4
```

Αντίστοιχα με την Ενότητα 4.5.2, το επίπεδο σημαντικότητας και εδώ είναι 0.025. Επομένως με την βοήθεια του ελέγχου και λαμβάνοντας υπόψη τις p-τιμές και το νέο επίπεδο σημαντικότητας, διαπιστώνουμε ότι για την χρονική στιγμή t_1, η απλή κύρια επίδραση του παράγοντα agegroup είναι στατιστικά σημαντική και στις δυο στάθμες του παράγοντα program. Αντίθετα την χρονική στιγμή t_2, η επίδραση του παράγοντα agegroup είναι στατιστικά σημαντική μόνο στην στάθμη B του παράγοντα program. Ο τελευταίος έλεγχος που θα πραγματοποιήσουμε θα έχει ως σκοπό την εύρεση των ηλικιακών γκρουπ που εντοπίζεται η διαφορά στην μέση τιμή του ύψους του άλματος.

```
> test<-group_by(df,time,program) %>%
+ pairwise_t_test(height~agegroup,p.adjust.method=
+ "bonferroni") %>%
+ select(-p,-p.signif,-p,-p.adj.signif)
> test

# A tibble: 9 x 8
   program time  .y.    group1 group2    n1    n2    p.adj
   <fct>   <fct> <chr>  <chr>  <chr>  <int> <int>    <dbl>
 1 A       t1    height 18-23  24-29     10    10 1    e+ 0
 2 A       t1    height 18-23  30-35     10    10 7.97e-10
 3 A       t1    height 24-29  30-35     10    10 3.07e-10
 4 B       t1    height 18-23  24-29     10    10 9.65e- 5
 5 B       t1    height 18-23  30-35     10    10 5.13e- 3
 6 B       t1    height 24-29  30-35     10    10 4.38e- 1
 7 B       t2    height 18-23  24-29     10    10 6.42e- 3
 8 B       t2    height 18-23  30-35     10    10 3.83e- 1
 9 B       t2    height 24-29  30-35     10    10 1    e- 4
```

Με την βοήθεια του παραπάνω ελέγχου και λαμβάνοντας υπόψη τις p.adj τιμές της διόρθωσης κατά Bonferroni καταλήγουμε στα εξής συμπεράσματα:

- Το μέσο ύψος του άλματος για το πρόγραμμα εκγύμνασης **A** την χρονική στιγμή $\mathbf{t}_1$, είναι στατιστικά σημαντικά διαφορετικό μεταξύ των ηλικιακών γκρουπ 18-23 vs 30-35 και 24-29 vs 30-35.

- Αντίθετα το μέσο ύψος του άλματος για το πρόγραμμα εκγύμνασης **A** την χρονική στιγμή $\mathbf{t}_1$, δεν διαφέρει στατιστικά σημαντικά μεταξύ των ηλικιακών γκρουπ 18-23 vs 24-29.

- Το μέσο ύψος του άλματος για το πρόγραμμα εκγύμνασης **B** την χρονική στιγμή $\mathbf{t}_1$, είναι στατιστικά σημαντικά διαφορετικό μεταξύ των ηλικιακών γκρουπ 18-23 vs 24-29 και 18-23 vs 30-35.

- Αντίθετα το μέσο ύψος του άλματος για το πρόγραμμα εκγύμνασης **B** την χρονική στιγμή $\mathbf{t}_1$, δεν διαφέρει στατιστικά σημαντικά μεταξύ των ηλικιακών γκρουπ 24-29 vs 30-35.

- Το μέσο ύψος του άλματος για το πρόγραμμα εκγύμνασης **B** την χρονική στιγμή $\mathbf{t}_2$, είναι στατιστικά σημαντικά διαφορετικό μεταξύ των ηλικιακών γκρουπ 18-23 vs 24-29 και 24-29 vs 30-35.

- Αντίθετα το μέσο ύψος του άλματος για το πρόγραμμα εκγύμνασης **B** την χρονική στιγμή $\mathbf{t}_2$, δεν διαφέρει στατιστικά σημαντικά μεταξύ των ηλικιακών γκρουπ 18-23 vs 30-35.

Με την ανάλυση των αποτελεσμάτων του τελευταίου ελέγχου, έχουμε ουσιαστικά ολοκληρώσει την ανάλυση μας. Έτσι λοιπόν έχουμε πλέον καλύψει κάθε περίπτωση για μία μικτή ANOVA τριών παραγόντων, δείχνοντας πως πρέπει να αντιμετωπίζονται τέτοιου είδους προβλήματα.

4.6 Βιβλιοθήκες και Συναρτήσεις της R

Στην υποενότητα αυτή θα συγκεντρώσουμε τις βιβλιοθήκες, καθώς και τις συναρτήσεις που χρησιμοποιήθηκαν καθόλη την διάρκεια της ανάλυσης της μικτής ANOVA με την βοήθεια της R. Για τις συναρτήσεις θα αναφερθούν εν συντομία τα ορίσματα τους καθώς και η χρησιμότητα τους. Οι βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι εξής:

- **tidyverse**: Για την τροποποίηση των δεδομένων.

- **ggpubr**: Για την δημιουργία των διαγραμμάτων.

- **rstatix**: Περιέχει όλες τις συναρτήσεις που χρησιμοποιήθηκαν στην εκτέλεση των στατιστικών ελέγχων.

Οι συναρτήσεις που χρησιμοποιήθηκαν είναι οι εξής:

- **anova_test** (data,dv,wid,between,within)

Τα ορίσματα που λαμβάνει η συνάρτηση anova_test είναι το πλαίσιο δεδομένων (data), το όνομα του διανύσματος της εξαρτημένης μεταβλητής (dv), το όνομα της στήλης που περιέχει την αρίθμηση των υποκειμένων του προβλήματος (wid), τον (ή τους) παράγοντα μεταξύ των υποκειμένων (between) και τον (ή τους) παράγοντα εντός των υποκειμένων (within). Χρησιμοποιείται για τον υπολογισμό της μικτής ANOVA.

- **get_anova_table()**

Το όρισμα που λαμβάνει η συνάρτηση get_anova_table είναι το όνομα που έχουμε ορίσει στην συνάρτηση anova_test. Με την χρήση της, εξάγεται ο πίνακας ANOVA που υπολογίστηκε με την βοήθεια της συνάρτησης anova_test και εφόσον παραβιάζεται η υπόθεση της σφαιρικότητας τότε γίνεται αυτόματα διόρθωση.

- **pairwise_t_test** (formula,paired,p.adjust.method)

Η formula που εισάγεται διαχωρίζεται με το σύμβολο ∼, έχοντας στο αριστερό της μέλος την εξαρτημένη μεταβλητή και στο δεξί της μέλος τον παράγοντα ως προς τον οποίο θέλουμε να πραγματοποιηθούν οι πολλαπλές συγκρίσεις κατά ζεύγη. Εφόσον έχουμε επαναλαμβανόμενες παρατηρήσεις για τον παράγοντα αυτό, το όρισμα paired λαμβάνει την λογική τιμή TRUE. Τέλος με την βοήθεια του ορίσματος p.adjust.method μπορεί να γίνει χρήση οποιασδήποτε προσέγγισης επιθυμούμε για τις p-τιμές, με την προσωπική επιλογή μας να είναι η μέθοδος Bonferroni.

- **identify_outliers** (data,variable)

Τα ορίσματα που λαμβάνει η συνάρτηση identify_outliers είναι το πλαίσιο δεδομένων και το όνομα της εξαρτημένης μεταβλητής για την οποία θα πραγματοποιηθεί ο έλεγχος για τον εντοπισμό των ακραίων τιμών.

- **levene_test** (data,formula)

Τα ορίσματα που λαμβάνει η συνάρτηση levene_test είναι το πλαίσιο δεδομένων και μία formula στην οποία εισάγονται οι μεταβλητές για τις οποίες θα πραγματοποιηθεί ο έλεγχος. Το αριστέρο μέλος αποτελείται από την εξαρτημένη μεταβλητή, ενώ το δεξί μέλος από τον (ή τους) παράγοντα μεταξύ των υποκειμένων. Τέλος διαχωρίζονται με το σύμβολο ∼.

- **ggboxplot** (data,x,y,color)

Τα ορίσματα που λαμβάνει η συνάρτηση ggboxplot είναι το πλαίσιο δεδομένων (data), τα ονόματα των μεταβλητών που θα χρησιμοποιηθούν στον οριζόντιο άξονα (x) και στον κατακόρυφο (y) και τέλος το όνομα της μεταβλητής που θα οριστεί με βάση τα διαφορετικά χρώματα των θηκοδιαγραμμάτων (color).

- **ggqqplot** (data,x,ggtheme,facet_grid)

Τα ορίσματα που λαμβάνει το διάγραμμα Q-Q για την κατασκευή του, είναι το πλαίσιο δεδομένων (data), το όνομα της μεταβλητής ως προς την οποία θα σχεδιαστεί (x) και τέλος τους παράγοντες ως προς τους οποίους θα ομαδοποιηθεί και θα διαιρεθεί σε πολλαπλά πλαίσια (facet_grid), οι οποίοι χωρίζονται με το σύμβολο ∼.

- **convert_as_factor** (data,name)

Τα ορίσματα που λαμβάνει η συνάρτηση convert_as_factor είναι το πλαίσιο δεδομένων και το όνομα της μεταβλητής (ουσιαστικά την στήλη των δεδομένων) που επιθυμούμε να μετατραπεί σε παράγοντα.

- **gather** (data,key,value,...)

Τα ορίσματα που λαμβάνει η συνάρτηση gather είναι το πλαίσιο δεδομένων (data), το όνομα της στήλης προς δημιουργία (key), το όνομα της στήλης των τιμών προς δημιουργία (value) και τέλος τα ονομάτα των στηλών που θα συγκεντρωθούν στις δυο προαναφερθείς στήλες (...). Η στήλη key περιέχει τα ονόματα των στηλών που θα συγκεντρωθούν, ενώ η στήλη value τις τιμές τους. Επομένως η παραπάνω συνάρτηση χρησιμοποιείται για την συλλογή πολλαπλών στηλών και την μετατροπή τους σε ένα ζεύγος νέων στηλών key-value.

- **group_by**(data,group)

Τα ορίσματα που λαμβάνει η συνάρτηση group_by είναι το πλαίσιο δεδομένων και το όνομα (ή ονόματα) της στήλης των δεδομένων ως προς την οποία θέλουμε να ομαδοποιήσουμε τις γραμμές των δεδομένων μας.

- **get_summary_stats** (data,type)

Τα ορίσματα που λαμβάνει η συνάρτηση get_summary_stats είναι το πλαίσιο δεδομένων και τι τύπο συνοπτικών στατιστικών επιθυμούμε να εξάγουμε. Στην περίπτωση μας επιλέχθηκε ο τύπος mean_sd, για την παρουσίαση της μέσης τιμής και της τυπικής απόκλισης κάθε αγωγής.

- %>%

Η παραπάνω συνάρτηση ουσιαστικά αποτελεί έναν τελεστή, ο οποίος εισάγεται στο τέλος της εντολής ώστε να συνδέσει την προηγούμενη με την επόμενη εντολή του κώδικα. Πιο συγκεκριμένα είτε θα προωθήσει μια τιμή είτε ένα αποτελέσμα στην επόμενη γραμμή του κώδικα.

Οι παραπάνω συναρτήσεις έχουν πολλά περισσότερα ορίσματα από αυτά που καταγράφηκαν, ωστόσο θεωρήθηκε σκόπιμο να γίνει αναφορά μόνο σε αυτά που χρησιμοποιήθηκαν κατά την διάρκεια των αναλύσεων μας. Τέλος η αξιοποίηση του τελεστή %>% στον κώδικα μας, βοήθησε στην σύνδεση ορισμένων από τις παραπάνω συναρτήσεις.

4.7 Επίλογος

Στο Κεφάλαιο 4 μελετήσαμε αναλυτικά τις περιπτώσεις που έχουμε μέχρι και τρείς παράγοντες σε ένα πρόβλημα μικτής ANOVA. Καθώς υπάρχει και το ενδεχόμενο στο πρόβλημα που θα συναντήσουμε να εμφανίζονται περισσότεροι, η αντιμετώπιση θα πρέπει να είναι ανάλογη. Εάν υποθέσουμε ότι έχουμε τέσσερις παράγοντες που επαληθεύουν τις προϋποθέσεις, συνοπτικά θα λέγαμε ότι ο πρώτος έλεγχος πρέπει να αφορά την αλληλεπίδραση των τεσσάρων παραγόντων. Στην περίπτωση που δεν είναι στατιστικά σημαντική, τότε θα καταλήγαμε να ακολουθήσουμε τα βήματα μίας μικτής ANOVA τριών παραγόντων, ενώ στην αντίθετη περίπτωση καλούμαστε να εξετάσουμε την επίδραση της αλληλεπίδρασης των τριών παραγόντων σε κάθε στάθμη του τέταρτου παράγοντα. Η συνέχεια της διαδικασίας για την ολοκλήρωση της ανάλυσης είναι ανάλογη και με βάση τα όσα έχουμε αναφέρει στις Ενότητες 4.4, 4.5.

Γενικότερα η μικτή ANOVA είναι μια σχετικά απλή μέθοδος, όσον αφορά την διαδικασία, τα βήματα και τον κώδικα που πρέπει γράψουμε στο στατιστικό πακέτο της R για την μελέτη και τον υπολογισμό της. Ωστόσο η δυσκολία της εμφανίζεται κατά την ερμηνεία των αποτελεσμάτων, καθώς θα πρέπει να είμαστε αναλυτικοί και ακριβείς στα λεγόμενα μας, λαμβάνοντας πάντα υπόψη ποιο είναι το ερώτημα του προβλήματος, ώστε εν τέλει να το απαντήσουμε.

Τέλος είναι πολύ σημαντικό να γίνεται εύκολα αντιληπτή κατά την ανάγνωση ενός προβλήματος μια ενδεχόμενη χρήση της μικτής ANOVA για την λύση του προβλήματος. Για να επιτευχθεί αυτό είναι κομβικό να έχει γίνει πλήρως κατανοητή η θεωρία του Κεφαλαίου 4.2. Τα κυριότερα σημεία αφορούν την αναγκαιότητα δυο παραγόντων, τουλάχιστον ενός μεταξύ και τουλάχιστον ενός εντός των υποκειμένων, όπως επίσης να έχουν γίνει σαφείς οι έννοιες εντός και μεταξύ.

Παράρτημα

Η προσομοίωση των δεδομένων της Ενότητας 4.5.2 γίνεται ως εξής στην R:

```
> id <- rep(1:20, 6)
> gender <- rep(c(rep("M", 10), rep("F", 10)), 6)
> time <- rep(c(rep("t1", 20), rep("t2", 20), rep("t3", 20)), 2)
> exercise <- rep(c(rep("A", 60), rep("B", 60)))
> score <- c(
+   rnorm(10, mean = 41.2, sd = 1.61),
+   rnorm(10, mean = 41.1, sd = 1.61),
+   rnorm(10, mean = 34.2, sd = 3.02),
+   rnorm(10, mean = 39.7, sd = 3.03),
+   rnorm(10, mean = 51.4, sd = 3.93),
+   rnorm(10, mean = 43.5, sd = 1.37),
+   rnorm(10, mean = 32.9, sd = 2.48),
+   rnorm(10, mean = 35.4, sd = 2.61),
+   rnorm(10, mean = 33.2, sd = 3.96),
+   rnorm(10, mean = 33.0, sd = 3.26),
+   rnorm(10, mean = 33.1, sd = 3.63),
+   rnorm(10, mean = 38.5, sd = 2.33)
+ )
> stress <- data.frame(id = id, gender = gender, time = time,
+ exercise = exercise, score = score)
> stress <- stress %>%
+ convert_as_factor(id, time, exercise, gender)
```

Οι 120 τιμές των δεδομένων που χρησιμοποιήθηκαν για την ανάλυση μας στο Κεφάλαιο 4.5.2 είναι οι εξής:

```
    id gender time exercise    score
1    1      M   t1        A 42.60260
2    2      M   t1        A 42.54558
3    3      M   t1        A 41.21322
4    4      M   t1        A 37.12445
5    5      M   t1        A 39.83567
6    6      M   t1        A 39.34479
7    7      M   t1        A 41.91712
8    8      M   t1        A 39.06406
9    9      M   t1        A 43.17977
10  10      M   t1        A 39.05753
11  11      F   t1        A 40.01736
```

```
12  12       F    t1         A 40.45291
13  13       F    t1         A 43.91367
14  14       F    t1         A 37.79378
15  15       F    t1         A 36.12622
16  16       F    t1         A 45.46494
17  17       F    t1         A 38.05593
18  18       F    t1         A 37.06927
19  19       F    t1         A 35.27304
20  20       F    t1         A 35.10349
21   1       M    t2         A 37.13706
22   2       M    t2         A 31.27280
23   3       M    t2         A 30.68417
24   4       M    t2         A 43.24250
25   5       M    t2         A 33.03697
26   6       M    t2         A 35.03205
27   7       M    t2         A 35.97603
28   8       M    t2         A 32.85025
29   9       M    t2         A 35.64205
30  10       M    t2         A 35.22941
31  11       F    t2         A 38.93765
32  12       F    t2         A 44.05116
33  13       F    t2         A 46.16125
34  14       F    t2         A 44.43555
35  15       F    t2         A 39.83991
36  16       F    t2         A 41.65633
37  17       F    t2         A 38.37728
38  18       F    t2         A 38.32313
39  19       F    t2         A 45.19515
40  20       F    t2         A 37.04311
41   1       M    t3         A 49.84216
42   2       M    t3         A 50.56911
43   3       M    t3         A 50.31177
44   4       M    t3         A 55.91269
45   5       M    t3         A 48.47994
46   6       M    t3         A 47.56137
47   7       M    t3         A 50.49787
48   8       M    t3         A 51.20415
49   9       M    t3         A 51.53060
50  10       M    t3         A 49.21283
51  11       F    t3         A 47.76618
52  12       F    t3         A 43.93065
53  13       F    t3         A 44.50226
54  14       F    t3         A 42.71526
55  15       F    t3         A 43.02302
56  16       F    t3         A 41.63059
57  17       F    t3         A 43.75701
```

```
58   18       F    t3         A  44.88559
59   19       F    t3         A  44.70807
60   20       F    t3         A  42.63462
61    1       M    t1         B  33.00635
62    2       M    t1         B  30.23278
63    3       M    t1         B  31.42259
64    4       M    t1         B  32.03314
65    5       M    t1         B  37.50142
66    6       M    t1         B  31.00232
67    7       M    t1         B  33.31295
68    8       M    t1         B  32.90301
69    9       M    t1         B  30.19533
70   10       M    t1         B  32.51431
71   11       F    t1         B  32.13332
72   12       F    t1         B  38.62429
73   13       F    t1         B  34.12448
74   14       F    t1         B  32.68522
75   15       F    t1         B  34.44078
76   16       F    t1         B  34.54988
77   17       F    t1         B  38.19668
78   18       F    t1         B  35.33642
79   19       F    t1         B  37.20381
80   20       F    t1         B  33.70698
81    1       M    t2         B  37.90374
82    2       M    t2         B  34.12023
83    3       M    t2         B  32.55710
84    4       M    t2         B  30.79029
85    5       M    t2         B  24.57713
86    6       M    t2         B  36.71039
87    7       M    t2         B  26.83496
88    8       M    t2         B  35.21920
89    9       M    t2         B  39.67368
90   10       M    t2         B  26.89877
91   11       F    t2         B  34.90124
92   12       F    t2         B  33.13828
93   13       F    t2         B  34.50929
94   14       F    t2         B  31.27216
95   15       F    t2         B  29.91368
96   16       F    t2         B  40.73089
97   17       F    t2         B  36.16432
98   18       F    t2         B  30.10834
99   19       F    t2         B  32.20149
100  20       F    t2         B  30.32348
101   1       M    t3         B  37.53696
102   2       M    t3         B  33.77743
103   3       M    t3         B  28.66864
```

```
104  4       M    t3         B  28.89280
105  5       M    t3         B  28.55401
106  6       M    t3         B  32.72946
107  7       M    t3         B  29.09915
108  8       M    t3         B  37.48288
109  9       M    t3         B  42.99042
110 10       M    t3         B  29.78058
111 11       F    t3         B  46.56693
112 12       F    t3         B  42.79088
113 13       F    t3         B  36.07048
114 14       F    t3         B  39.51401
115 15       F    t3         B  35.43491
116 16       F    t3         B  35.17119
117 17       F    t3         B  33.84055
118 18       F    t3         B  34.66693
119 19       F    t3         B  44.23287
120 20       F    t3         B  36.96984
```

Η προσομοίωση των δεδομένων της Ενότητας 4.5.3 γίνεται ως εξής στην R:

```
> id <- rep(1:60, 2)
> program <- rep(c(rep("A", 30), rep("B", 30)), 2)
> agegroup <- rep(c(rep("18-23", 10),
+ rep("24-29", 10), rep("30-35", 10)), 4)
> time <- rep(c("t1", "t2"), each = 60)
> height <- c(
+   rnorm(10, 84.2, 1.69),
+   rnorm(10, 84.7, 2.37),
+   rnorm(10, 75.0, 2.28),
+   rnorm(10, 78.7, 3.10),
+   rnorm(10, 85.0, 2.46),
+   rnorm(10, 83.1, 2.89),
+   rnorm(10, 84.2, 2.28),
+   rnorm(10, 84.6, 3.85),
+   rnorm(10, 84.3, 3.05),
+   rnorm(10, 83.3, 1.29),
+   rnorm(10, 87.2, 3.78),
+   rnorm(10, 81.5, 1.85)
+ )
> df <- data.frame(id = id, program = program,
+ agegroup = agegroup, time = time, height = height)
> df <- df %>%
+ convert_as_factor(id, program, agegroup, time)
```

Οι 120 τιμές των δεδομένων που χρησιμοποιήθηκαν για την ανάλυση μας στο Κεφάλαιο 4.5.3 είναι οι εξής:

```
    id program agegroup time   height
1    1       A    18-23   t1 81.99862
2    2       A    18-23   t1 86.91579
3    3       A    18-23   t1 83.50700
4    4       A    18-23   t1 82.41670
5    5       A    18-23   t1 83.74661
6    6       A    18-23   t1 83.82926
7    7       A    18-23   t1 86.59186
8    8       A    18-23   t1 84.42510
9    9       A    18-23   t1 85.83972
10  10       A    18-23   t1 83.19073
11  11       A    24-29   t1 85.77187
12  12       A    24-29   t1 84.53678
13  13       A    24-29   t1 85.49728
14  14       A    24-29   t1 83.22941
15  15       A    24-29   t1 82.27768
16  16       A    24-29   t1 89.85602
17  17       A    24-29   t1 86.65676
18  18       A    24-29   t1 82.41406
19  19       A    24-29   t1 83.88048
20  20       A    24-29   t1 82.56478
21  21       A    30-35   t1 77.07646
22  22       A    30-35   t1 75.35249
23  23       A    30-35   t1 73.00981
24  24       A    30-35   t1 73.11260
25  25       A    30-35   t1 72.95724
26  26       A    30-35   t1 74.87194
27  27       A    30-35   t1 73.20723
28  28       A    30-35   t1 77.05166
29  29       A    30-35   t1 79.57719
30  30       A    30-35   t1 73.51970
31  31       B    18-23   t1 85.62982
32  32       B    18-23   t1 78.61550
33  33       B    18-23   t1 79.66085
34  34       B    18-23   t1 76.39282
35  35       B    18-23   t1 76.95564
36  36       B    18-23   t1 74.40920
37  37       B    18-23   t1 78.29796
38  38       B    18-23   t1 80.36187
39  39       B    18-23   t1 80.03723
40  40       B    18-23   t1 76.24534
```

```
41  41        B     24-29   t1  85.71920
42  42        B     24-29   t1  86.02268
43  43        B     24-29   t1  88.43404
44  44        B     24-29   t1  84.16988
45  45        B     24-29   t1  83.00797
46  46        B     24-29   t1  89.51492
47  47        B     24-29   t1  84.35254
48  48        B     24-29   t1  83.66506
49  49        B     24-29   t1  82.41350
50  50        B     24-29   t1  82.29536
51  51        B     30-35   t1  80.92580
52  52        B     30-35   t1  85.36533
53  53        B     30-35   t1  87.19730
54  54        B     30-35   t1  85.69906
55  55        B     30-35   t1  81.70914
56  56        B     30-35   t1  83.28614
57  57        B     30-35   t1  80.43929
58  58        B     30-35   t1  80.39228
59  59        B     30-35   t1  86.35853
60  60        B     30-35   t1  79.28096
61   1        A     18-23   t2  84.87198
62   2        A     18-23   t2  81.88090
63   3        A     18-23   t2  83.16402
64   4        A     18-23   t2  83.82244
65   5        A     18-23   t2  89.71953
66   6        A     18-23   t2  82.71079
67   7        A     18-23   t2  85.20261
68   8        A     18-23   t2  84.76053
69   9        A     18-23   t2  81.84051
70  10        A     18-23   t2  84.34135
71  11        A     24-29   t2  88.66224
72  12        A     24-29   t2  85.82953
73  13        A     24-29   t2  84.65922
74  14        A     24-29   t2  83.33641
75  15        A     24-29   t2  78.68463
76  16        A     24-29   t2  87.76878
77  17        A     24-29   t2  80.37507
78  18        A     24-29   t2  86.65233
79  19        A     24-29   t2  89.98739
80  20        A     24-29   t2  80.42284
81  21        A     30-35   t2  89.69639
82  22        A     30-35   t2  87.14171
83  23        A     30-35   t2  82.59503
84  24        A     30-35   t2  84.92474
85  25        A     30-35   t2  82.16503
86  26        A     30-35   t2  81.98661
```

```
87  27        A     30-35   t2 81.08637
88  28        A     30-35   t2 81.64545
89  29        A     30-35   t2 88.11728
90  30        A     30-35   t2 83.20349
91  31        B     18-23   t2 84.62373
92  32        B     18-23   t2 84.58670
93  33        B     18-23   t2 83.72153
94  34        B     18-23   t2 81.06649
95  35        B     18-23   t2 82.82702
96  36        B     18-23   t2 82.50827
97  37        B     18-23   t2 84.17861
98  38        B     18-23   t2 82.32598
99  39        B     18-23   t2 84.99852
100 40        B     18-23   t2 82.32174
101 41        B     24-29   t2 89.41924
102 42        B     24-29   t2 83.22159
103 43        B     24-29   t2 82.59949
104 44        B     24-29   t2 95.87176
105 45        B     24-29   t2 85.08605
106 46        B     24-29   t2 87.19456
107 47        B     24-29   t2 88.19220
108 48        B     24-29   t2 84.88872
109 49        B     24-29   t2 87.83922
110 50        B     24-29   t2 87.40313
111 51        B     30-35   t2 80.97940
112 52        B     30-35   t2 81.57676
113 53        B     30-35   t2 81.36529
114 54        B     30-35   t2 85.96783
115 55        B     30-35   t2 79.86000
116 56        B     30-35   t2 79.10517
117 57        B     30-35   t2 81.51823
118 58        B     30-35   t2 82.09860
119 59        B     30-35   t2 82.36686
120 60        B     30-35   t2 80.46225
```

Βιβλιογραφία

Διεθνής Βιβλιογραφία

Brown, H. & Prescott, R. (2006). Applied Mixed Models in Medicine. John Wiley & Sons, Ltd, Chichester, p.1-105.

Gelman, A. & Hill, J. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, Cambridge, p.251-298.

McCulloch, C. E. & Searle, S. R. (2001). Generalized, Linear, and Mixed Models. John Wiley & Sons, Inc., New York, p.156-218, 302-310.

Seltman, H. J. (2018). Experimental Design and Analysis. Carnegie Mellon University, Pennsylvania, p.357-378.

Meier, L. (2022). ANOVA and Mixed Models: A Short Introduction Using R (online version). Chapman & Hall, United Kingdom, Chapter 6.

Johnson, A. A. & Ott, M. Q. & Dogucu, M. (2021). Bayes Rules! An Introduction to Applied Bayesian Modeling (online version). Chapman & Hall, United Kingdom, Chapter 15.

Ιστότοποι

https://bookdown.org/steve_midway/DAR/random-effects.html

https://www.ssc.wisc.edu/sscc/pubs/MM/MM_TestEffects.html

https://yury-zablotski.netlify.app/post/mixed-effects-models-1/

https://yury-zablotski.netlify.app/post/mixed-models/

https://joshuawiley.com/MonashHonoursStatistics/LMM_Comparison.html

https://towardsdatascience.com/maximum-likelihood-ml-vs-reml-78cf79bef2cf

http://users.stat.umn.edu/~gary/classes/5303/handouts/REML.pdf

https://www.datanovia.com/en/lessons/mixed-anova-in-r/

https://bookdown.org/mike/data_analysis/linear-mixed-models.html

https://m-clark.github.io/posts/2020-03-01-random-categorical/

https://m-clark.github.io/mixed-models-with-R/random_intercepts.html

https://m-clark.github.io/mixed-models-with-R/random_slopes.html

https://stats.stackexchange.com/questions/48671/what-is-restricted-maximum

-likelihood-and-when-should-it-be-used

https://stats.stackexchange.com/questions/4700/what-is-the-difference-between-fixed-effect-random-effect-and-mixed-effect-mode

https://stats.stackexchange.com/questions/24452/how-to-interpret-variance-and-correlation-of-random-effects-in-a-mixed-effects-m

https://stats.stackexchange.com/questions/13166/rs-lmer-cheat-sheet

https://stats.stackexchange.com/questions/228800/crossed-vs-nested-random-effects-how-do-they-differ-and-how-are-they-specified?noredirect=1&lq=1

https://stats.stackexchange.com/questions/475549/demonstrating-complete-pooling-no-pooling-and-partial-pooling-regression-in-r

https://stats.stackexchange.com/questions/57240/how-do-i-interpret-the-correlations-of-fixed-effects-in-my-glmer-output

https://stats.stackexchange.com/questions/23778/how-should-mixed-effects-models-be-compared-and-or-validated

https://stats.stackexchange.com/questions/214119/model-selection-testing-the-need-for-random-effects-terms-in-longitudinal-data?rq=1

https://www.rdocumentation.org/

https://statistics.laerd.com/spss-tutorials/mixed-anova-using-spss-statistics.php

https://www.amstatisticalconsulting.com/three-way-repeated-measures-anova/

https://www.itl.nist.gov/div898/handbook/pmc/section5/pmc542.htm

https://www.ssc.wisc.edu/sscc/pubs/MM/MM_TestEffects.html#testing-mixed-models-parameters

https://www.bristol.ac.uk/cmm/learning/videos/random-intercepts.html

https://www.bristol.ac.uk/cmm/learning/videos/random-slopes.html

https://bodowinter.com/tutorial/bw_LME_tutorial2.pdf

https://stat.ethz.ch/~meier/teaching/anova/random-and-mixed-effects-models.html

https://alexander-pastukhov.github.io/notes-on-statistics/effective-degrees-of-freedom-number-of-parameters.html

https://www.sciencedirect.com/science/article/pii/S0047259X14001869