**Εργασία Στατιστικής Υπολογιστικής**

**Ζωή Κελεπίρη 3290**

**Κατερίνα Ζαχαρία 3427**

***Προσπέλαση του αρχείου***

Αρχικά γίνεται import το dataset που μας δόθηκε με την χρήση της εντολής read\_sav το οποίο υπάρχει στο package heaven το οποίο κάνουμε install. Διαβάζουμε τα δεδομένα μας data με την χρήση της attach και στη συνέχεια γίνεται χρήση της περιγραφικής στατιστικής με την summary. Αναλυτικά, το κομμάτι κώδικα R.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

*Graphs*

**Γραφικές παραστάσεις όλων των μεταβλητών**

Αρχικά για κάθε μεταβλητή δημιουργήθηκε ιστόγραμμα και θηκόγραμμα, όπου οι μεταβλητές χρησιμοποιήθηκαν όπως είναι αλλά και με την ύπαρξη του λογαρίθμου. Παρακάτω απεικονίζονται οι εντολές που χρησιμοποιήθηκαν για αυτήν την υλοποίηση, χρησιμοποιήθηκε μια επαναληπτική διαδικασία for προκειμένου να γίνουν οι γραφικές παραστάσεις για κάθε μεταβλητή, για την εμφάνιση του αποτελέσματος χρησιμοποιήθηκε η par(mfrow)) όπου τα διαγράμματα εμφανίζονται ανά τρία σε 5 σειρές. Προκειμένου να μην υπάρχουν προβλήματα στην απεικόνιση των γραφημάτων έγινε χρήση της par(mar) όπου αλλάζει το μέγεθος της απεικόνισης σε κανονικό.

Graphical user interface, text, application

Description automatically generated

Diagram, engineering drawing

Description automatically generated

Diagram

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Στο παραπάνω boxplot παρατηρούμε το ενδοτεταρτημοριακό εύρος, την μέγιστη και ελάχιστη τιμή, την μέση τιμή καθώς και έκτοπα σημεία αν υπάρχουν. Οι μεταβλητές headcirumference, Gestation, mnocig, mheight, mppwt, fage,fheight, lowbwt, mage35 έχουν κάποια έκτοπα σημεία.

Graphical user interface

Description automatically generated

Αντίστοιχα και σε αυτό το σημείο παρατηρούνται έκτοπα σημεία σε μεταβλητές, όμως σε λιγότερες αυτή την φορά. Αυτές οι μεταβλητές είναι η mheight και fheight. Με την χρήση του log στις μεταβλητές παρατηρούμε ότι μειώθηκαν τα έκτοπα σημεία για αυτό και εντοπίζονται μόνο στις δυο αυτές μεταβλητές.

*Simple Graph For Variables*

Επιπλέον για κάθε μεταβλητή του dataset δημιουργήθηκε ένα ιστόγραμμα στο οποίο απεικονίζονται με κίτρινο χρώμα η συνάρτηση πυκνότητας πιθανότητας της μεταβλητής ενώ με κόκκινο χρώμα η μέση τιμή της. Οι εντολές που χρησιμοποιήθηκαν για την παραπάνω υλοποίηση είναι η density() για την πυκνότητα και η mean() για την μέση τιμή. Παρακάτω απεικονίζονται τα διαγράμματα για κάθε μεταβλητή.

A picture containing chart

Description automatically generated

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

*Other methods*

**ANOVA**

Στην συνέχεια εφαρμόστηκαν οι μέθοδοι της ανάλυσης διακύμανσης, επειδή οι μέθοδοι αυτοί μπορούν να εφαρμοστούν σε ερμηνευτικές μεταβλητές που είναι κατηγορικές, έγινε μετατροπή κάποιων μεταβλητών. Οι μεταβλητές οι οποίες έγιναν κατηγορικές είναι η smoker, , η lowbwt και η mage35, headcirumference και της fedyrs, με την χρήση του factor. Παρακάτω το αντίστοιχο στιγμιότυπο από την υλοποίηση που περιεγράφηκε.

Εικόνα που περιέχει κείμενο, εσωτερικό

Περιγραφή που δημιουργήθηκε αυτόματα

Η ANOVA αρχικά εφαρμόστηκε στην μεταβλητή απόκρισης Birthweight με μία από τις παρακάτω ερμηνευτικές μεταβλητές : headcirumference.categorical, lowbwt.categorical, fedyrs.categorical. Για τις μεταβλητές αυτές χρησιμοποιήθηκαν οι εντολές που σχετίζονται με τον πίνακα ανάλυσης της διακύμανσης και την αξιολόγηση του υποδείγματος.

* Μια κατηγορική μεταβλητή – headcirumference

Τα αποτελέσματα για την εύρεση σημαντικότητας της μεταβλητής σε σχέση με την Birthweight απεικονίζονται στις παρακάτω εικόνες. Παρακάτω το διάγραμμα για τις μεταβλητές.

Chart, box and whisker chart

Description automatically generated

Η πρώτη εντολή είναι ο πίνακας ανάλυσης διακύμανσης (analysis of variance) όπου ουσιαστικά αναλύεται η συνολική διασπορά σε αυτή που εξηγείται από το υπόδειγμα και σε αυτή που δεν εξηγείται. Μέρος της μεταβλητότητας που εξηγείται από την Birthweight είναι το SSA = 93 και αντίστοιχα το μέρος της μεταβλητότητας που δεν εξηγείται είναι το SSE = 98, με το συνολικό άθροισμα τετραγώνων να είναι το SSY = 191. To p-value<<0.05 και άρα η διαφορά ανάμεσα στις στάθμες του παράγοντα είναι στατιστικά σημαντική, ο παράγοντας headcirumference επιδρά στην μεταβλητή Birthweight. Οι δυο επόμενες εντολές αναφέρονται στην αξιολόγηση του υποδείγματος. Από τα αποτελέσματα συμπεραίνουμε ότι το επίπεδο αναφοράς είναι το 12 και στατιστικά σημαντικοί θεωρούνται το 13,14,15. Το 16 είναι οριακά στατιστικά σημαντικό. Επιπλέον η πιθανότητα είναι μικρότερη από το 0.05 γεγονός που αποδεικνύει ότι η ανεξάρτητη μεταβλητή επιδρά στην εξαρτημένη. Για την εύρεση των παραγόντων της ανεξάρτητης που είναι στατιστικά σημαντικοί γίνεται χρήση του Post Hoc Test. Από το παρακάτω στιγμιότυπο συμπεραίνουμε ότι τα επίπεδα 13-12, 14-12, 15-12 φαίνεται να έχουν στατιστικά σημαντική διαφορά (p< 0.05).

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματαΕικόνα που περιέχει κείμενο, πίνακας

Περιγραφή που δημιουργήθηκε αυτόματα

Το ίδιο μοτίβο εφαρμόστηκε και στις άλλες δυο μεταβλητές, για λόγους απλότητας παρατίθενται τα αποτελέσματα από την εφαρμογή τους και ένας σύντομος σχολιασμός.

* Μια κατηγορική μεταβλητή – fedyrs

Στην περίπτωση της fedyrs παρακάτω θα διαπιστώσουμε ότι δεν είναι στατιστικά σημαντική, για το λόγο ότι η πιθανότητα είναι μεγαλύτερη από 0.05. Για τον λόγο αυτό δεν προχωράμε στην χρήση του Post Hoc Test.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Στην συνέχεια χρησιμοποιήθηκε το μοντέλο ANOVA με περισσότερες από μία ερμηνευτικές μεταβλητές. Υλοποιήθηκαν τέσσερα τέτοια παραδείγματα τα οποία θα αναλυθούν συνοπτικά παρακάτω.

* Μια κατηγορική μεταβλητή – lowbwt

Από τα παραπάνω συμπεραίνουμε ότι ο παράγοντας lowbwt είναι στατιστικά σημαντικός αφού η πιθανότητα p-value είναι μικρότερη από 0.05 και το lowbwt με επίπεδο 1 είναι επίσης στατιστικά σημαντικό. Με την χρήση του Post Hoc Test καταλήξαμε ότι το επίπεδο 1-0 είναι στατιστικά σημαντικό.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

* Δυο κατηγορικές μεταβλητές – lowbwt, headcircumference

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Από την παραπάνω εικόνα, συμπεραίνουμε ότι το headcirumference και το lowbwt είναι στατιστικά σημαντικό, ενώ η αλληλεπίδραση τους δεν είναι στατιστικά σημαντική. Εάν αφαιρεθεί όπως παραπάνω θα είναι το ιδανικό. Το επίπεδο αναφοράς για την lowbwt είναι το 1 και για την headcirumference είναι το 12 κάποια από τα οποία είναι στατιστικά σημαντικά.

* Δυο κατηγορικές μεταβλητές – fedyrs, lowbwt

Μια πιο ολοκληρωμένη εφαρμογή της anova ακολουθήθηκε η ίδια διαδικασία με παραπάνω με την διαφορά ότι χρησιμοποιήθηκε η tapply για την εύρεση των μέσων τιμών καθώς και ένα διάγραμμα το οποίο οπτικοποιεί τα δεδομένα που περιεγράφηκαν.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Chart, bar chart

Description automatically generated

* Δυο κατηγορικές μεταβλητές – mage35, smoker

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Η παραπάνω υλοποίηση καταλήγει στο ότι η μεταβλητή smoker είναι στατιστικά σημαντική για την Birthweight ωστόσο η mage35 δεν είναι στατιστικά σημαντική όπως και η αλληλεπίδραση τους. Αφαιρώντας την αλληλεπίδραση τους παρατηρούμε ότι ακόμα η smoker είναι σημαντική και με την χρήση της anova καταλήγουμε στο ότι δεν υπάρχει στατιστικά σημαντική διαφορά ανάμεσα στα μοντέλα για το λόγο ότι 0.2691>0.05 και επομένως επιλέγουμε το δεύτερο μοντέλο υλοποίησης.

* Τρεις κατηγορικές μεταβλητές – smoker, lowbwt, mage35

Οι μέθοδοι που χρησιμοποιήθηκαν είναι παρόμοια με αυτές που αναφέρθηκαν παραπάνω. Το παρακάτω στιγμιότυπο απεικονίζει τα αποτελέσματα από την υλοποίηση αυτή.

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Από τα παραπάνω συμπεραίνουμε ότι στατιστικά σημαντική επίδραση έχουν οι παράγοντες smoker, lowbwt αφού η πιθανότητα τους είναι μικρότερη από 0.05. Αντίθετα η mage καθώς και όλες οι αλληλεπιδράσεις δεν είναι στατιστικά σημαντικές. Αναφορικά με την τελευταία εντολή που χρησιμοποιήθηκε, οι συντελεστές δείχνουν τη διαφορά της επίδρασης κάθε επιπέδου από ένα επίπεδο αναφοράς. Το επίπεδο αναφοράς και για τις τρεις μεταβλητές είναι το 0. Παρατηρούμε ότι οι επιδράσεις ανάμεσα στις μεταβλητές με επίπεδο 1 δεν είναι στατιστικά σημαντικές, ενώ για την smoker και την mage35 με επίπεδο το 1 είναι στατιστικά σημαντικές και άρα χρειαζόμαστε και τα δυο επίπεδα για την smoker και την mage35. Μπορούμε να αφαιρέσουμε την αλληλεπίδραση μεταξύ των μεταβλητών και τρέχουμε πάλι τις εντολές. Παρατηρούμε ότι στατιστικά σημαντικές είναι οι smoker και την mage35,αφού η πιθανότητα τους είναι μικρότερη από 0.05. Εφόσον παρατηρούμε σημαντικότητα σε κάποιες μεταβλητές εφαρμόζουμε το Post Hoc Test. Για την smoker, lowbwt και τα δυο επίπεδα είναι στατιστικά σημαντικά, ενώ για την mage35 δεν είναι κάποιο στατιστικά σημαντικό καθώς η πιθανότητα είναι μεγαλύτερη από 0.05. Τα αποτελέσματα απεικονίζονται παρακάτω.

Text

Description automatically generated

**ANCOVA**

Τέλος έγινε χρήση της συνδυακύμανσης ερμηνευτικών μεταβλητών που είναι είτε συνεχείς είτε κατηογορικές. Έγινε εφαρμογή πέντε τέτοιων παραδειγμάτων τα οποία θα αναλυθούν παρακάτω. Για κάθε περίπτωση εξετάστηκε η συνδυακύμανση δυο μεταβλητών καθώς και η ύπαρξη η όχι της αλληλεπίδρασης των ανεξάρτητων μεταβλητών όπως επίσης και η σύγκριση των δυο αυτών μεθόδων.

* Μια μεταβλητή συνεχής και μία κατηγορική – mheight, lowbwt

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Με βάση τα παραπάνω και οι δυο μεταβλητές στατιστικά σημαντικές αφού οι πιθανότητες μικρότερες από το 0.05.

* Μια μεταβλητή συνεχής και μία κατηγορική – mnocig, smoker

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Χωρίς την αλληλεπίδραση τους προκύπτει ότι και οι δυο μεταβλητές είναι στατιστικά σημαντικές.

* Μια μεταβλητή συνεχής και μία κατηγορική – headcirumference, mage35

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Το headcirumference είναι στατιστικά σημαντικό σε αντίθεση με το mage35.

* Μια μεταβλητή συνεχής και μία κατηγορική – Gestation, lowbwt

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Οι δυο μεταβλητές είναι στατιστικά σημαντικές.

* Μια μεταβλητή συνεχής και μία κατηγορική – fage, fedyrs

Εικόνα που περιέχει κείμενο

Περιγραφή που δημιουργήθηκε αυτόματα

Καμία από τις μεταβλητές δεν είναι στατιστικά σημαντική

*Birthweight Correlations*

Πριν αρχίσουμε να εξηγούμε τις συσχετίσεις που ερευνήσαμε, παρουσιάζουμε τις γραφικές που χρησιμοποιήσαμε για να δούμε την κατανομή της Birthweight, την οποία αρχικά θεωρούμε ως την εξαρτημένη μεταβλητή μας.

Chart

Description automatically generated

Στα διαγράμματα βλέπουμε ότι η μεταβλητή ακολουθεί κανονική κατανομή (Normal Q-Q Plot), ότι το εύρος τιμών των δεδομένων είναι από 4+ μέχρι 10, και ότι η διάμεσος βρίσκεται στο 7. Από το ιστόγραμμα και την εντολή που φαίνεται στην πιο κάτω εικόνα, που επιστρέφει αρνητικό αριθμό, παρατηρούμε ότι υπάρχει αριστερή ουρά.



**Σύγκριση δύο δειγμάτων**

* Διακύμανση μεταξύ των μεταβλητών mnocig και fnocig.

1. qf – Ποσοστιαία σημεία

Graphical user interface, text

Description automatically generated Chart, histogram

Description automatically generated

Το F.ratio είναι μεγαλύτερο από το ποσοστιαίο σημείο του 97.5%, οπότε κατατάσσεται στο 2,5% του πληθυσμού των δειγμάτων και άρα η μηδενική υπόθεση (ισότητας των διακυμάνσεων) απορρίπτεται.

1. pf – Πιθανότητα/Συχνότητα

Ένας άλλος τρόπος, πιο αποδοτικός, για να το εξετάσουμε είναι το pf, μέσω της πράξης *«2\*(1-pf(F.ratio,83,83))»*, όπου το 83 είναι οι βαθμοί ελευθερίας. Το αποτέλεσμα που επιστρέφει είναι *0.005133703,* το οποίο είναι κάτω από 0.05 και άρα είναι στατιστικά σημαντική διαφορά.

1. var.test – Πλήρης έλεγχος

Εφαρμόζοντας το var.test παρατηρούμε ακριβώς τα ίδια αποτελέσματα με το Β, οπότε συμπεραίνουμε πως, αν και μικρή η διαφορά, είναι στατιστικά σημαντική.

Text, letter

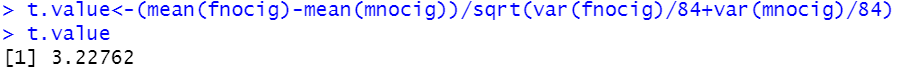
Description automatically generated

* Μέση τιμή μεταξύ των μεταβλητών mnocig και fnocig.

Από την στιγμή που οι δύο πιο πάνω μεταβλητές έχουν στατιστική σημαντική διαφορά στις διακυμάνσεις, μπορούμε να ελέγξουμε και την διαφορά στις μέσες τιμές τους.

1. qt με υπολογισμό t.value





Συγκρίνοντας τα δύο παρατηρούμε ότι αν πάρουμε απόλυτη τιμή του t.value, θα είναι μεγαλύτερη από το qt και άρα πέφτει στην απορριπτική περιοχή. 🡪 Μεγάλη στατιστικά σημαντική διαφορά.

1. t.test – Πλήρης έλεγχος

Το αποτέλεσμα είναι ακριβώς το ίδιο με το A. (Μεγάλη στατιστικά σημαντική διαφορά)

Text, letter

Description automatically generated

* Έλεγχος ανεξαρτησίας

1. fisher.test μεταξύ smoker και mage35 (και οι 2 δυαδικές)

Μετράμε το σύνολο των εγγραφών που έχουν 0 ή 1 σε κάθε μια μεταβλητή και τους συνδυασμούς των 2. Δημιουργούμε το πινακάκι με την εντολή matrix και βλέπουμε αντίστοιχα για 00, 01, 10, 11 ότι υπάρχουν 38, 38, 2 και 6 εγγραφές.   
Με το fisher.test παρατηρούμε ότι η συσχέτιση τους δεν είναι στατιστικά σημαντική   
(p-value = 0.2694 > 0.05)

Graphical user interface, text

Description automatically generated

1. chisq.test μεταξύ smoker και lowbwt (και οι 2 δυαδικές)

Κάνουμε την ίδια διαδικασία με πάνω και έχουμε 38, 34, 2 και 10 εγγραφές αντίστοιχα για τους συνδυασμούς 00, 01, 10, 11.   
Με το chisq.test παρατηρούμε ότι η πιθανότητα εξάρτησης τους είναι στατιστικά σημαντική 🡪 Εξαρτημένα. (p-value = 0.04478 < 0.05)

Text

Description automatically generated

* Έλεγχος συσχέτισης

Για λόγους χώρου, εμφανίζονται μόνο οι συνδυασμοί των πρώτων 4 μεταβλητών.

1. cor με μέθοδο Pearson

Text

Description automatically generated

Η συσχέτιση του length με το Birthweight (π.χ.) φτάνει το 49.84%

1. cor με μέθοδο SpearmanText

   Description automatically generated

Η συσχέτιση του length με το Birthweight με αυτή την μέθοδο φτάνει το 49.36%

Δηλαδή οι διαφορές των δύο μεθόδων δεν είναι πολύ μεγάλες.

1. Συσχέτιση μεταξύ Birthweight και mheight με cor.test

Η συσχέτιση φτάνει το 45% και είναι πολύ σημαντική επειδή το p-value<<0.05.

Text

Description automatically generated

**Παλινδρόμηση με μεταβλητή απόκρισης την Birthweight και μια αριθμητική ερμηνευτική μεταβλητή**

Σε αυτό το κομμάτι θα εφαρμόσουμε διάφορα μοντέλα (lm, polynomial, log, exp) για να βρούμε την παλινδρόμηση.

Το πρώτο παράδειγμα ερευνά την συσχέτιση μεταξύ της Birthweight με την Gestation, εξετάζοντας 3 διαφορετικά μοντέλα. Πρώτα, εξάγουμε μέσω του summary την ευθεία που περιγράφει καλύτερα την μεταξύ τους σχέση, η οποία είναι Birthweight = -7.1991 + 0.3693 \* Gestation.

Text, letter

Description automatically generated

Επίσης, μπορούμε να βγάλουμε το συμπέρασμα ότι είναι στατιστικά σημαντική η διαφορά τους αφού, όπως φαίνεται και στο κόκκινο κουτί, p-value<<0.05. Εκτελώντας και summary(aov(μοντέλου)), επιβεβαιώνουμε και τα αποτελέσματα μέσω της ANOVA. Ένα άλλο χρήσιμο συμπέρασμα που μπορούμε να εξάγουμε είναι ότι το r2 = 52% περίπου, και άρα η Birthweight περιγράφει περίπου 52% την μεταβλητότητα της Gestation.

Για τις δύο αυτές μεταβλητές, εξετάζουμε ακόμα δύο μοντέλα. Το επόμενο που θα αναλύσουμε είναι το πολυωνυμικό μοντέλο. Δοκιμάζουμε να βάλουμε το Gestation στην Δευτέρα και άρα να έχουμε το μοντέλο *lm(Birthweight~Gestation+Gestation2)*. Η παρατήρηση που κάνουμε είναι ότι το τετράγωνο του Gestation δεν είναι στατιστικά σημαντικό, οπότε το μοντέλο αυτό δεν περιγράφει ούτε καλύτερα ούτε χειρότερα την συσχέτιση από ότι το γραμμικό.

Graphical user interface, text, application, email

Description automatically generated

Συγκεκριμένα, με την εντολή anova(model, linear) εξετάζουμε την διαφορά και βλέπουμε ότι δεν έχουν στατιστικά σημαντική και άρα λόγω απλοποίησης προτιμούμε το γραμμικό.

Graphical user interface, text, application, email

Description automatically generated

Το τελευταίο μοντέλο είναι το *lm(log(Birthweight)~Gestation)* το οποίο είναι επίσης στατιστικά σημαντικό.

A screenshot of a computer

Description automatically generated with low confidence

Αντίστοιχη έρευνα κάνουμε και σε σχέση με το mmpwt – mother’s pre-pregnancy weight, και το length. Ο λόγος που επιλέξαμε αυτά τα δύο είναι για να δούμε πρώτα αν το βάρος της μητέρας με το βάρος του μωρού έχουν σχέση, και αν το μήκος του μωρού παίζει ρόλο στο βάρος του.

Όλες οι συσχετίσεις με το Birthweight και με ποιες συγκεκριμένες τιμές μεταβλητών επηρεάζεται, φαίνονται στο πιο κάτω δέντρο. Για μικρό βάρος μωρού (lowbwt=1) η συσχέτιση είναι 5.208, ενώ για μεγάλο βάρος ελέγχει και τις άλλες μεταβλητές, π.χ. Gestation < 39.5.

Diagram, schematic

Description automatically generated

**Πολλαπλή παλινδρόμηση με μεταβλητή απόκρισης την Birthweight και πολλές αριθμητικές ερμηνευτικές μεταβλητές**

Για αυτή την κατηγορία προβλημάτων έχουμε κάνει 4 διαφορετικά μοντέλα αλλά θα αναφέρουμε μόνο το ένα. Σκοπός μας εδώ είναι να δημιουργήσουμε κατάλληλα μοντέλα που να δείχνουν την επιρροή που έχουν πάνω από 2 αριθμητικές μεταβλητές, και οι αλληλεπιδράσεις τους, πάνω στην Birthweight. Άρα, κάνουμε ανανέωση κάθε φορά το μοντέλο αφαιρώντας datapoints, αλληλεπιδράσεις, κ.ά.

Το αρχικό μοντέλο που σκεφτήκαμε να ελέγξουμε είναι το *lm(Birthweight~mppwt\*fheight\*length+I(mppwt^2)+I(fheight^2)+I(length^2)).*

Το μοντέλο μας είναι γραμμικό μέχρι τρίτης τάξης (όλες οι μεταβλητές, τα τετράγωνά τους, και αλληλεπιδράσεις 2ης τάξης και 3ης τάξης).

Text

Description automatically generated

Παρατηρούμε ότι κανένα δεν είναι στατιστικά σημαντικό, οπότε αρχίζουμε να αφαιρούμε παράγοντες. Πρώτα αφαιρούμε την αλληλεπίδραση της 3ης τάξης: mppwt:fheight:length. Προκύπτουν τα εξής:

Text

Description automatically generated

Συνεχίζοντας να αφαιρούμε μη σημαντικές αλληλεπιδράσεις, φτάνουμε στο 4ο μοντέλο, το οποίο επιστρέφει τα πιο κάτω. Σειρά έχει να αφαιρέσουμε το τετράγωνο του length.

Text

Description automatically generated

Μετά από αρκετές επαναλήψεις, καταλήγουμε στο μοντέλο *lm(Birthweight~mppwt+fheight+length+mppwt:fheight).* Εμφανίζουμε κάποιες γραφικές για να δούμε αν υπάρχει διαφορά στην ετεροσκεδαστικότητα.

Chart, scatter chart

Description automatically generated

Όπως φαίνεται και στο γράφημα, υπάρχει διαφορά στην διασπορά και άρα βάζουμε log στο μοντέλο και συνεχίζουμε να το αναπτύσσουμε βάζοντας αρχικά log(Birthweight). Το μοντέλο, όπως ήδη παρατηρούμε από το Normal Q-Q Plot, έχει ακραίες τιμές. Οπότε, η επόμενη αλλαγή είναι να αφαιρέσουμε το δείγμα 10. Η σημαντικότητα των παραγόντων εξακολουθεί να είναι στατιστικά σημαντική και άρα αφαιρούμε ακόμα ένα υπόδειγμα (22), όπου και φτάνουμε στο τελικό μας μοντέλο:

Table

Description automatically generated

*Other Correlations*

**Παλινδρόμηση με μεταβλητή απόκρισης την length και μια αριθμητική ερμηνευτική μεταβλητή.**

Για την length εξετάζουμε τις εξής συσχετίσεις: mheight, fheight και mppwt. Δοκιμάζουμε γραμμικά, πολυωνυμικά και λογαριθμικά μοντέλα.

Το γραμμικό μοντέλο της length σε σχέση με το mheight είναι αρκετά πιο αποδοτικό από το πολυωνυμικό τους μοντέλο, χωρίς το δεύτερο να μην δείχνει την στατιστικά σημαντική διαφορά τους.

**Text

Description automatically generated** Text

Description automatically generated

**Chart, scatter chart

Description automatically generated**

Συγκρίνοντας τα δύο μοντέλα μέσω της anova, επιβεβαιώνουμε και το πιο πάνω.

Text

Description automatically generated

Αντίστοιχα ασχολούμαστε και με τις υπόλοιπες μεταβλητές.

**Παλινδρόμηση με μεταβλητή απόκρισης την Gestation και μια αριθμητική ερμηνευτική μεταβλητή.**

Στον πρώτο μας έλεγχο με το mppwt, βλέπουμε ότι είναι στατιστικά σημαντική η διαφορά.

Text, letter

Description automatically generated

Δοκιμάζουμε επίσης την Gestation με το fheight και mheight. Τα αποτελέσματα παρουσιάζονται πιο κάτω. Η mheight επηρεάζει περισσότερο την μεταβλητότητα της Gestation από ότι το fheight.

Text, letter

Description automatically generated A screenshot of a computer

Description automatically generated with medium confidence