Labbrapport i Statistik

Laboration 12

732G46

Mattias Hällgren, Michael Debebe

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2021-11-29

Innehåll

Introduktion

I denna laboration kommer två dataset analyseras.

I det första datasetet, har 15 försökerspersoner delats upp i 5 olika åldersgrupper, där de samtliga tre inom respektive åldersgrupp har fått gå på olika dieter inriktade på fet. Dieterna har varit extremt låg, ganska låg och moderat låg. Därefter har mätningar på skillnaderna skett i lipidnivåerna (gram per liter).

I det andra datasetet har det sammanställts data om produktivitetsförbättringar det senaste året för ett urval av företag, där de tillverkar elektronisk datautrustning. Resultat av bevisen mäts på en skala 0 till 100. Firmorna var klassifierade baserat på deras genomsnittliga utgifter för forskning och utveckling.

Målet med denna laboration är att göra analyser av blockförsök och bearbeta datan. Att skapa en kovariansmodell och Dietämföra skillnaden mellan en ANOVA och ANCOVA

Databehandling

```
prod_study <- read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/22_7")
colnames(prod_study) <- c("Curr_year","Utgifter", "Firma", "prior_year")
prod_study$Utgifter <- ifelse(prod_study$Utgifter == 1, "Låg",
ifelse(prod_study$Utgifter == 2, "Medel", "Hög"))
prod_study$Utgifter <- as.factor(prod_study$Utgifter)


Fat_diet <-read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/21_7")
colnames(Fat_diet) <- c("fat","Åldersgrupp","Diet")
Fat_diet$`Diet`<-as.factor(Fat_diet$`Diet`)
Fat_diet$`Åldersgrupp`<-as.factor(Fat_diet$`Åldersgrupp`)
```

Uppgifter

21.7 Fat in Diets

a) Why do you think that age of subject was used as a blocking variable?

Anledning varför ålder används som en block variabel är för att man vill veta även om ålder påverkar hur snabbt eller långsamt man tappar fetma i Dietämförelse med de olika åldersgrupperna. även se om en viss diet funkar bättre än en annan för just den åldersintervallen kan vara en intressant info att veta om.

b) Obtain the residuals, plot them against the fitted values. Also prepare a normal probability plot of the residuals. What are your findings?

```
model <- aov(formula = fat~Åldersgrupp+Diet, data = Fat_diet)
summary(model)
```

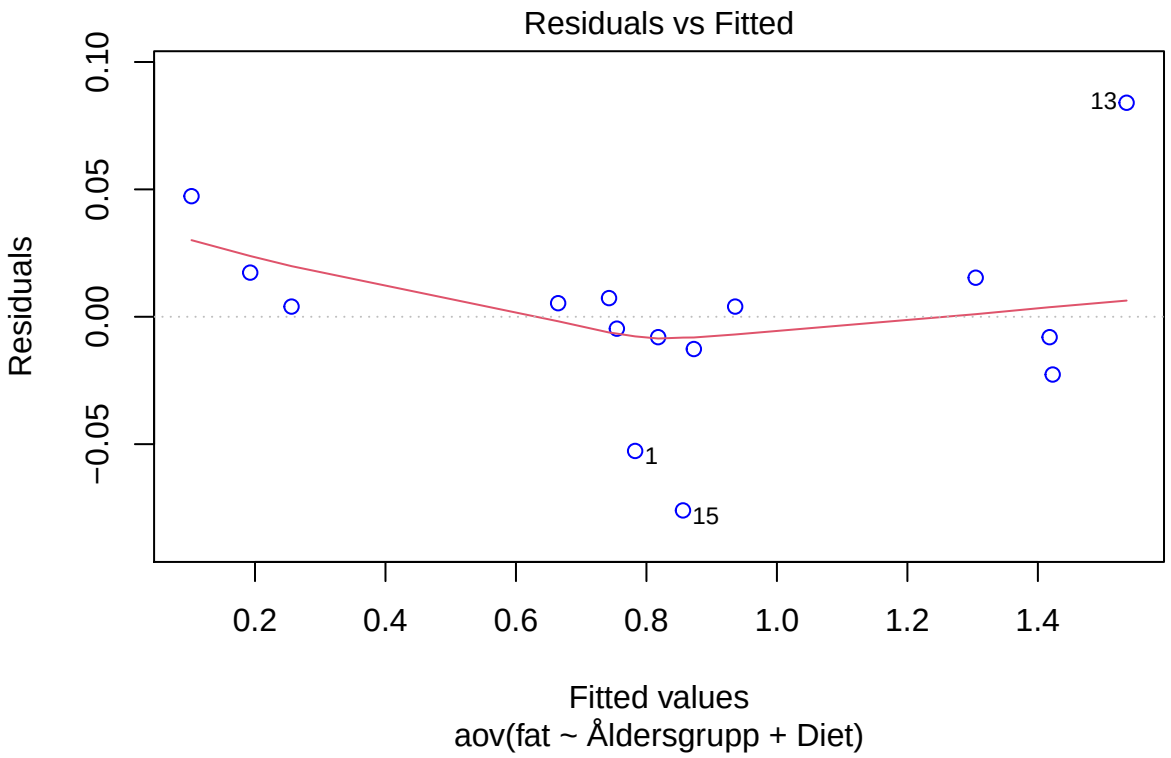
```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## Åldersgrupp  4 1.4190  0.3547   146.9 1.61e-07 ***
## Diet         2 1.3203  0.6601   273.4 4.33e-08 ***
## Residuals    8 0.0193  0.0024
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

För analysen av randomiserade Blockförsök kommer modellen som visas nedan användas.

$$Y_{iDiet} = \mu_{..} + \rho_i + \tau_{Diet} + \varepsilon_{iDiet}$$

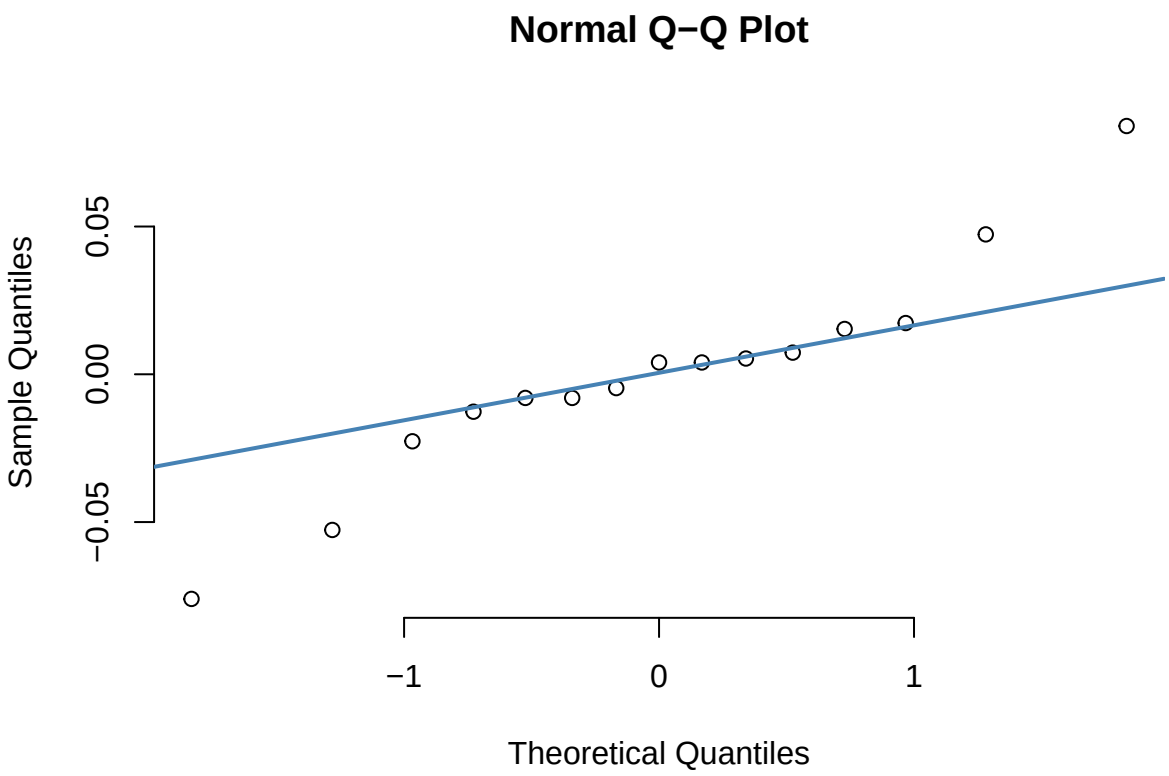
Där: $\mu \ldots$ är en konstant. ρ_i är konstanter för block (rad) effekt, subject to the restriction $\sum \rho_i = 0$ τ_{Diet} är konstant för behandlingseffekten, subject to the restriction $\sum \tau_{Diet} = 0$ ε_{iDiet} är oberoende. $N\left(0, \sigma^2\right)$ $i = 1, \ldots, n_b; Diet \doteq 1, \ldots, r$ Antaganden för modellen är att de är oberoende mätningar med normalfördelat medelvärde och Dietämn varians.

```r
#plot residuals vs fitted
plot(model, which=1, col=c("blue"))
```



Plot för Dietämförelse mellan de anpassade värden och residualer här ser man att de håller sig ganska centralt. Plotten visar tre extremvärden(Observation 1,15 och 19).

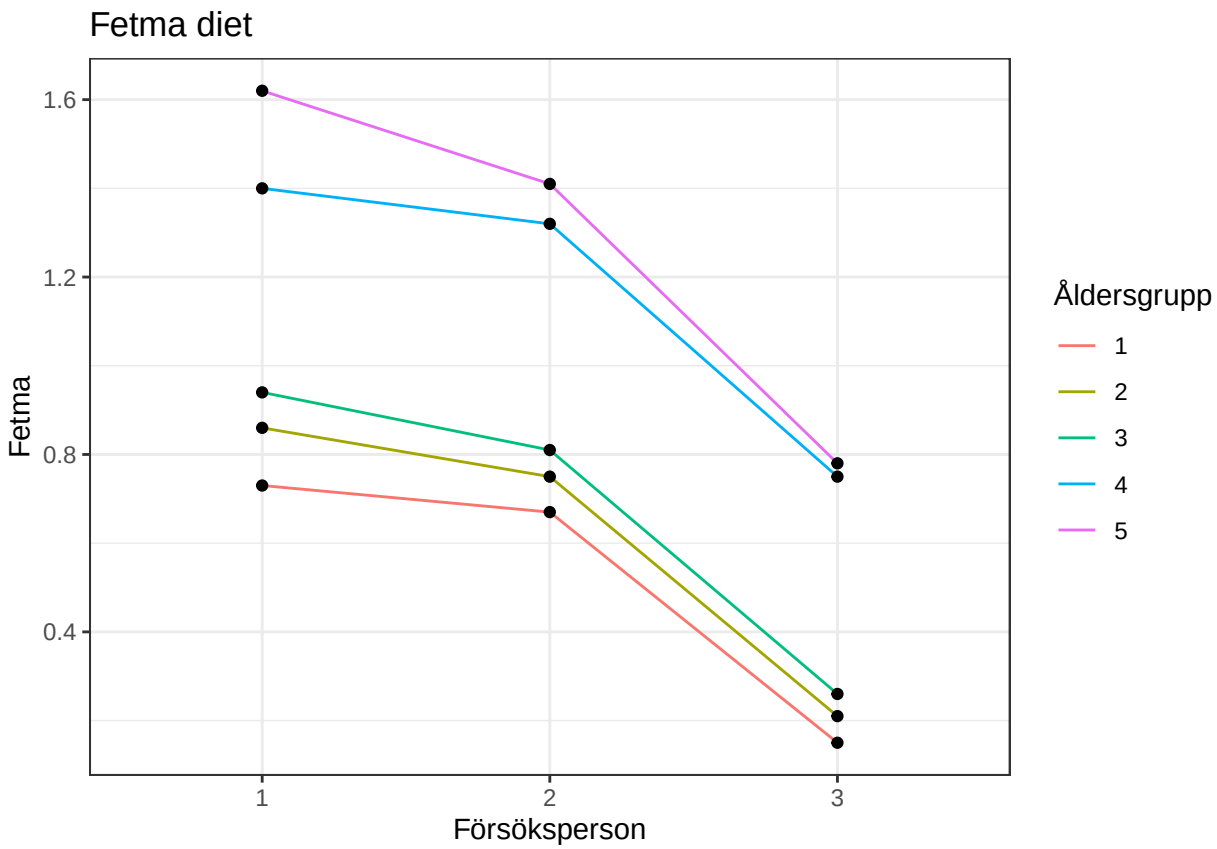
```r
#plot normal probability
qqnorm(y=model$residuals, pch = 1, frame = FALSE)
qqline(model$residuals, col = "steelblue", lwd = 2)
```



Ovanför är en Q-Q Plot som mäter normalfördelning här kan man se att värdena följer linjen men att den har några extremvärden. Detta tyder på att normalfördelningen inte är perfekt men ändå kan det tydas en normalfördelning.

c) Plot the responses Yij by blocks in format of figure 21.2 what does this plot suggest about the appropriateness of the no-interaction assumption here?

```
ggplot(data=Fat_diet, aes(x=Diet, y=fat,group=Åldersgrupp)) +
  geom_line(aes(color=Åldersgrupp))+
  geom_point() +
  labs(title = 'Fetma diet', x = 'Försöksperson', y = 'Fetma')+
  theme_bw()
```



Här kan man se att Åldersgrupp 4 och 5 ligger höst upp på tappad fett i gram per liter. Man kan även se att dieten som försöksperson 1 får fungerar bättre än diet 2 och 3 därav 3 är den sämre dieten av alla.

21.8 Fat in diets

a) Obtain the analysis of variance table

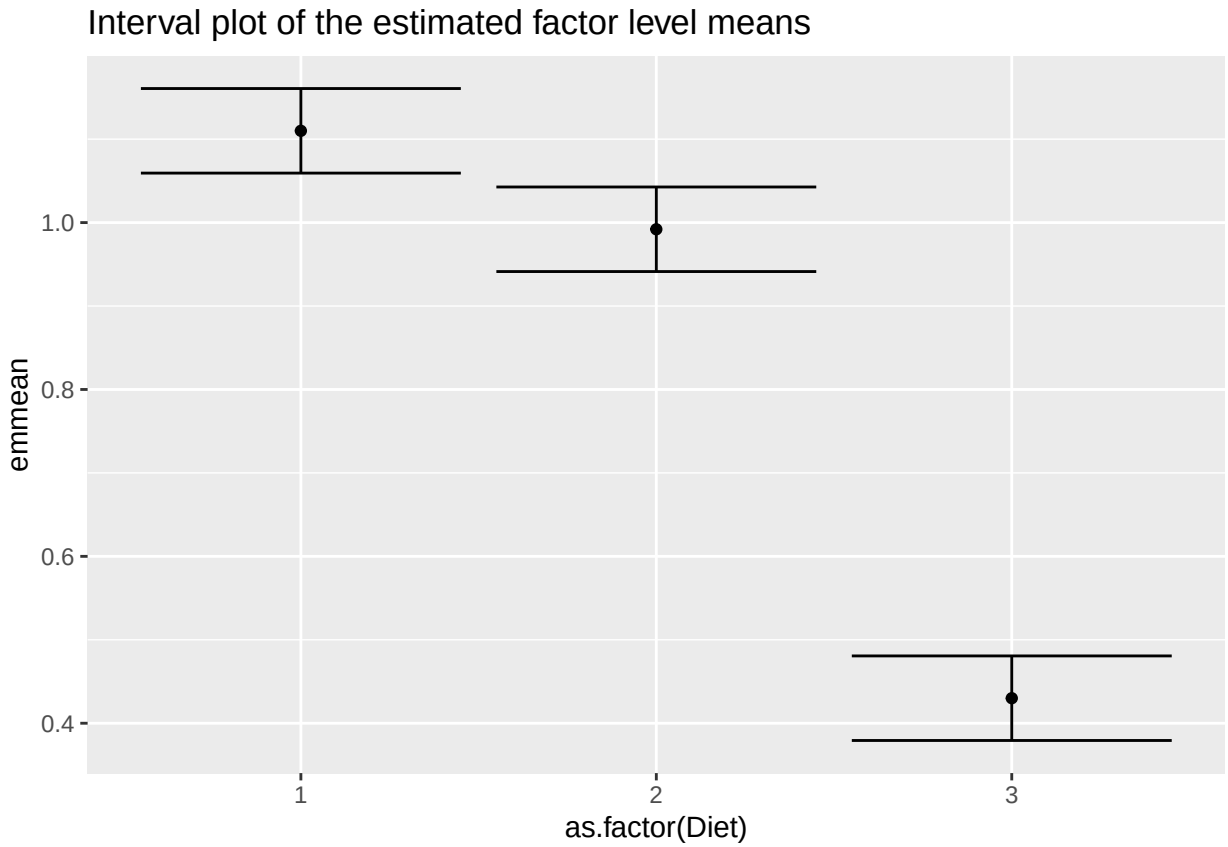
```
table <- aov(formula = fat~Åldersgrupp+Diet, data = Fat_diet)
anova(table)
```

```
## Analysis of Variance Table
##
## Response: fat
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Åldersgrupp  4 1.41896 0.35474  146.89 1.610e-07 ***
## Diet         2 1.32028 0.66014  273.35 4.326e-08 ***
## Residuals    8 0.01932 0.00242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ovan visas anova-tabellen för modellen, båda variablerna Block och Diet är starkt signifikanta och har otroligt låga p-värdet, i f-testen returnas höga teststatistikor. Residualerna även dom är väldigt låga.

b) Prepare a bar-interval graph of the estimated means, using 95 percent confidence intervals. Does it appear that the treatment means differ substantially here?

```
ggplot(data = Confint.model,aes(x=as.factor(Diet),y=emmean))+
geom_point()+
geom_errorbar(aes(ymin=lower.CL,ymax=upper.CL))+
ggtitle(label = "Interval plot of the estimated factor level means")
```



Ovan visas 95 % konfidensintervall för medelvärdena inom faktorgrupperna för faktor Diet, det syns tydligt hur grupp 3 är den grupp med längst medelvärde, grupp 2 är de näst största intervallet vars övre konfidensgräns gränsar till grupp 1's lägre gräns. Grupp 1 är den gruppen som har högst medelvärde. Utifrån denna graf ser det ut som att det finns signifikant skillnad mellan grupp 3 och grupp 1, däremot blir det svårt att säga om det ser ut och finnas signifikant skillnad i medelvärde mellan grupp 1 och grupp 2 baserat på enbart denna bild.

c) Test whether or not the mean reductions in lipid level differ for the three diets use alpha = .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

$$H0 : j_1 = j_2 = j_3 = 0$$

$$H1 : j_1 \neq \ j_2 \neq \ j_3 = 0$$

```
anova(table)
```

```
## Analysis of Variance Table
##
## Response: fat
##              Df  Sum Sq Mean Sq F value    Pr(>F)
## Åldersgrupp   4 1.41896 0.35474  146.89 1.610e-07 ***
## Diet          2 1.32028 0.66014  273.35 4.326e-08 ***
## Residuals     8 0.01932 0.00242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F = \frac{0.660140}{0.002415} = 273.3499$$
$$F(.95, 2, 8) = 4.46$$

I fall av att F > 4.46 förkastas $H0$, om inte antas $H0$.

I detta fall där teststatistiskan överstiger det kritiska värdet ($273.3499 > 4.46$) kan vi förkasta nollhypotesen om att det inte finns statistiskt signifikant skillnad i medeljusteringar av lipidnivå mellan de tre olika dieterna. P-värdet från testet summerar till 0.0000000432 vilket är starkt signifikant.

d) Estimate L1 = U1-U2 and L2 = U2-U3. Using the bonferroni procedure with a 95 percent family confidence coefficient. State your findings.

$D1 = 1.11 - 0.992 = 0.118$

$F = (1 - (0.05/4), 8) = 2.751524$

$s\left\{\bar{D}_{iDiet\cdot}\right\} = \sqrt{\frac{2*MSE}{n}} \Rightarrow \sqrt{\frac{2*0.0024}{5}} = 0.0301$

$D \pm B \cdot s\left\{D\right\} = 0.118 \pm 2.752 \cdot 0.0301 \ = 0.0327 : 0.203$

För differensen mellan faktor 1 och 2 kommer differensen i ett 95 procentigt konfidensintervall att ligga mellan 0.0327 och 0.203. Då intervallet inte täcker 0 är det fortfarande statistiskt signifikant. Det är dessutom ett relativt stort intervall vilket visar en osäkerhet.

$D2 = 0.992 - 0.43 = 0.562$

$F = (1 - (0.05/4), 8) = 2.751524$

$s\left\{\bar{D}_{iDiet\cdot}\right\} = \sqrt{\frac{2*MSE}{n}} \Rightarrow \sqrt{\frac{2*0.0024}{5}} = 0.0301$

$D \pm B \cdot s\left\{D\right\} = 0.562 \pm 2.752 \cdot 0.0301 \ = 0.4767 : 0.6473$

För differensen mellan faktor 2 och 3 kommer differensen i ett 95 procentigt konfidensintervall att ligga mellan 0.4767 och 0.6473. Intervallet täcker inte noll vilket gör det signifikant, intervallet är väldigt likt de övre intervallet i differsen mellan grupp 1 och 2 inom faktorn.

21.8

e) Test whether or not blocking effects are present use alpha = .05.State the alternatives, decision rule and conclusion. What is the P-Value of the test?

$$H0 : i_1 = i_2 = i_3 = 0$$
$$H1 : i_1 \neq \ i_2 \neq \ i_3 = 0$$

$$F = \frac{0.354740}{0.002415} = 146.8903$$

$ Ifall av att värdet från F-testet överstiger det kritiska vÃrdet kan vi med 95 % konfidens förkasta nollhypotesen om att det inte finns några skillnader i medelvärden mellan blocken.$

$$F = (0.95, 4, 8) = 3.84$$

I detta fall där teststatistiskan överstiger det kritiska värdet ($146.9 > 3.84$) kan vi förkasta nollhypotesen om att det inte finns statistiskt signifikant skillnad i medeljusteringar av lipidnivå mellan de fem olika Åldersgruppen. P-värdet från testet summerar till 0.000000161 vilket är starkt signifikant.

Uppgift 2

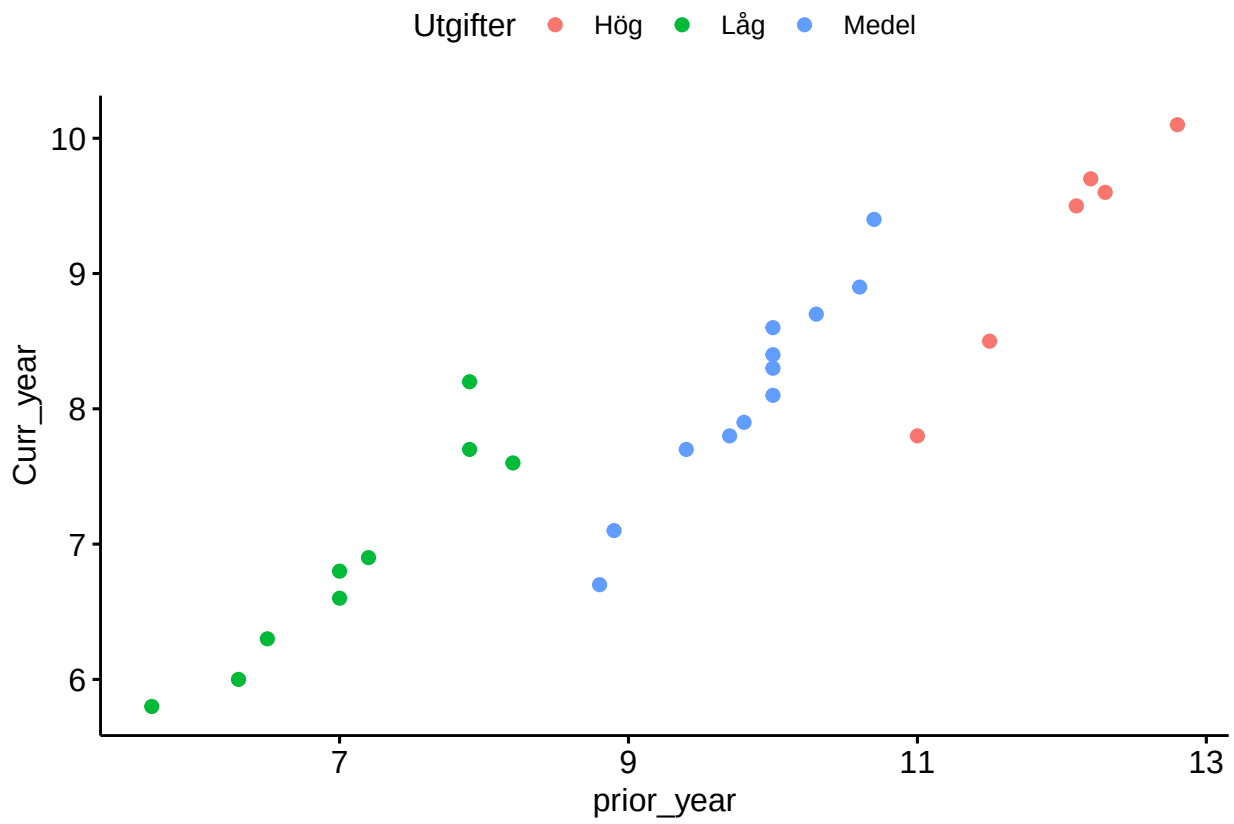
a) Formulera en korrekt kovariansmodell. Beskriv modellens delar, vad det betyder för datasetet och modellens antagande.

$$Y - X\beta + E$$

där **X** är designmatrisen bestående av: - en kolumn 1:or för μ - $r - 1$ kolumner med effektkodning för Faktor A - $c - 1$ kolumner med effektkodning för Faktor **B**

b) Skapa ett spridningsdiagram för data med olika färger för faktornivåerna. Finns det någon relation mellan responsvariabeln och kovariaten? syns det några faktoreffekter?

```
ggscatter(
  prod_study, x = "prior_year", y = "Curr_year",
  color = "Utgifter")
```



Ovanför är ett spridningsdiagram där kovarianten är pÃ x-axeln, antal pÃ y-axeln och Level som visar grupperna.

c) Anpassa en kovariansmodell. Pröva om det finns nÃgon kostnadsnivåeffekt med signifikansnivå på 5% visa era beräkningar för testet.

```
#Ancova model
contrasts(prod_study$Utgifter)<-contr.sum(n=3)
Ancova <-aov(Curr_year~Utgifter+prior_year,data = prod_study)
Anova(Ancova,type="III")
```

```
## Anova Table (Type III tests)
##
## Response: Curr_year
##              Sum Sq Df F value    Pr(>F)
## (Intercept)  0.8622  1  15.052 0.0007582 ***
## Utgifter     4.1958  2  36.623 7.095e-08 ***
## prior_year  14.0447  1 245.176 9.274e-14 ***
## Residuals    1.3175 23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

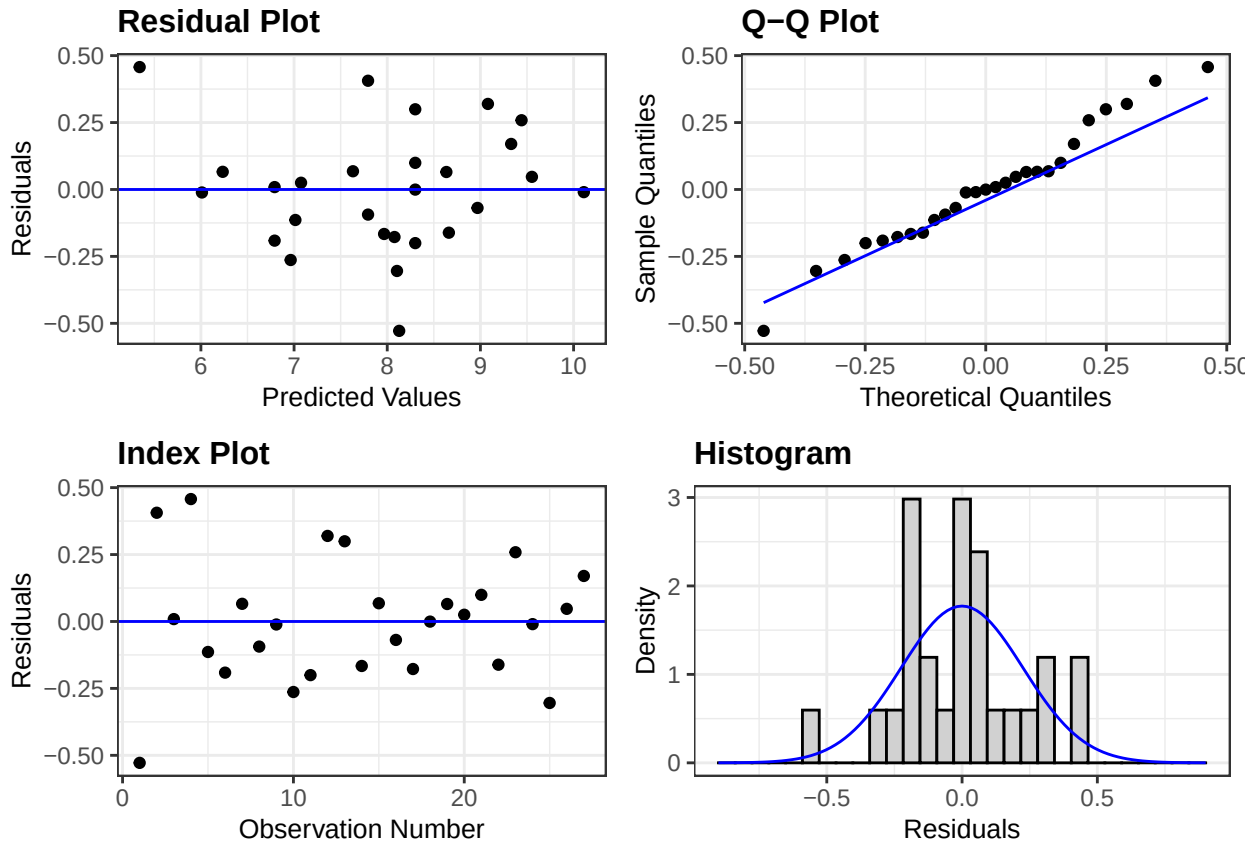
$$H0 : i_1 = i_2 = i_3 = 0$$
$$H1 : i_1 \neq \ i_2 \neq \ i_3 = 0$$

$$F = \frac{4.195826/2}{1.3175/23} = 36.623$$
$$F = (0.95, 2, 23) = 3.42$$

Ifall av att värdet från F-testet $F > 3.42$ överstiger det kritiska värdet kan vi med 95 % konfidens förkasta nollhypotesen om att det inte finns några skillnader i medelvärden mellan Åldersgruppen.

I detta fall där teststatistiskan överstiger det kritiska värdet ($36.623 > 3.42$) kan vi förkasta nollhypotesen om att det inte finns statistiskt signifikant skillnad i medeljusteringar av lipidnivå mellan de fem olika Åldersgruppen. P-värdet från testet summerar till $7.0954e - 08$ vilket är starkt signifikant.

d) Ta fram residualerna och kolla modellantaganden på ett relevant sätt.



Det syns tydligt hur spridningen i det första diagrammet (Upp vänster) inte är jämn med flera utomliggande observationer. Sett till normalfördelningstabellen (Upp höger) är det ett stort antal observationer som ligger utanför linjen. Index plotten är i detta fall den enda plotten som uppvisar en någorlunda jämn varians, histogrammet kan inte ses som normalfördelat och slutsatsen blir att vi ej kan se resiudalerna som normalfördelade.

Antaganden för att kunna använda en kovariansmodell är att residualerna är normalfördelade med jämn varians samt oberoende, då variansen inte är jämn trots att residualerna är oberoende uppfyller inte denna modell antagandena.

e) Anpasssa en envägs ANOVA utan kovariat och Dietämför resultatet med ovan.

```
ANCOVA2 <- aov(data = prod_study, Curr_year ~ Utgifter)
Anova(ANCOVA2,TYPE="III")
```

```
## Anova Table (Type II tests)
##
## Response: Curr_year
##           Sum Sq Df F value    Pr(>F)
## Utgifter  20.125  2   15.72 4.331e-05 ***
## Residuals 15.362 24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

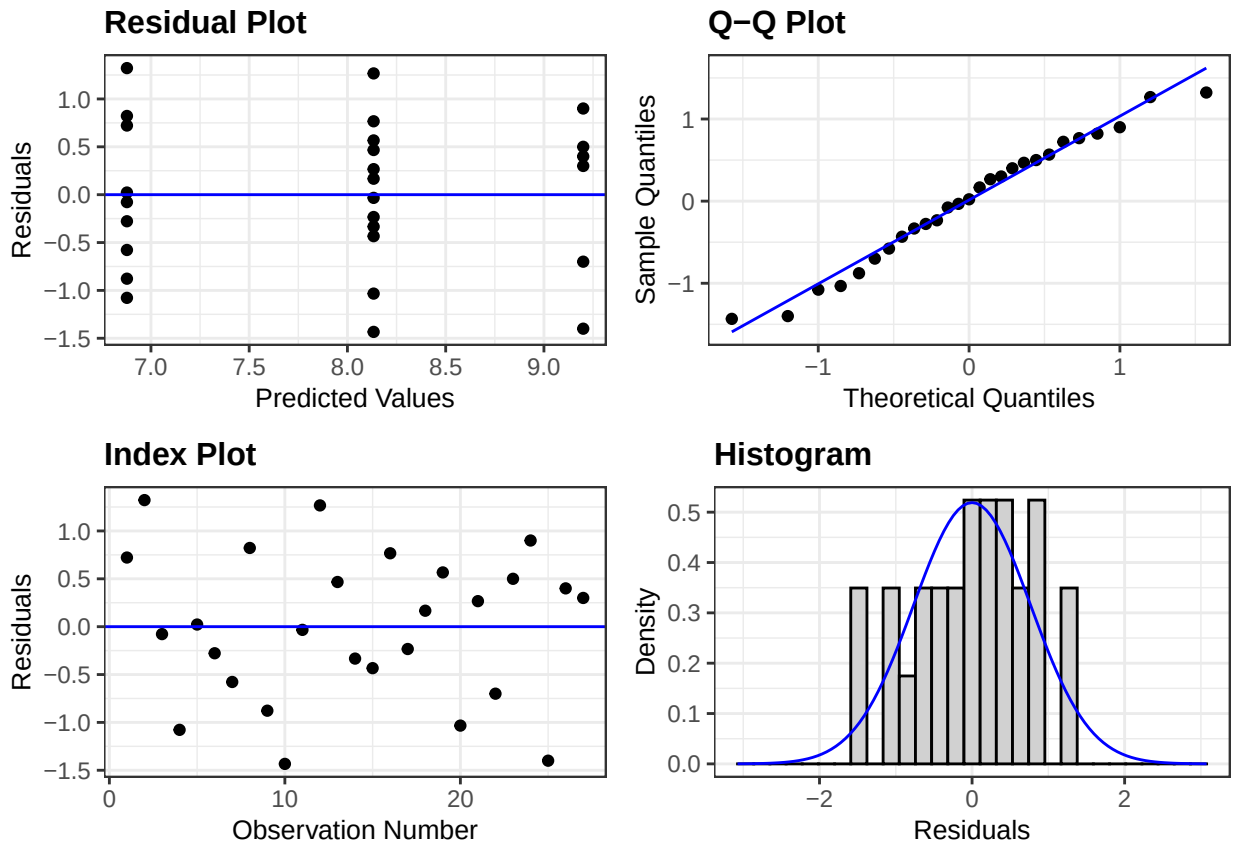
$$H0 : i_1 = i_2 = i_3 = 0$$
$$H1 : i_1 \neq \ i_2 \neq \ i_3 = 0$$

$$F = \frac{20.12518/2}{15.36222/24} = 15.72053$$

$ Ifall av att värdet från F-testet överstiger det kritiska vÃrdet kan vi med 95 % konfidens förkasta nollhypotesen om att det inte finns några skillnader i medelvärden mellan Åldersgruppen.$

$$F = (0.95, 2, 24) = 3.403$$

I detta fall där teststatistiskan överstiger det kritiska värdet ($15.72 > 3.403$) kan vi förkasta nollhypotesen om att det inte finns statistiskt signifikant skillnad i medeljusteringar av lipidnivå mellan de fem olika Åldersgruppen. P-värdet från testet summerar till $4.331e - 05$ vilket är starkt signifikant.



Vad gäller residualerna för modellen utan kovarit kan vi se att residualerna i samtliga av ovanstående plottar är normalfördelade och oberoende, vilket medför att modellen utfyller samtliga antaganden.

Vad gäller jämförelse mellan modellerna är SS och MSE i modellen utan kovariat (15.362 respektive 0.64) medan modellen med kovariat har SS och MSE som summerar till (1.3175 respektive 0.057), det är förståeligt att modellen utan kovariat har högre standardfel då kovariaten kan förklara en stor del av variationen i responsvariabeln.

Vad gäller variabeln utgifter är medelfelet för den däremot större i modellen utan kovariat då den som variabel får "förklara medelfelet ensamt", i modellen utan kovariat summerar SS för utgifter till 20.125 medan de i modellen med kovariat summerar till 4.196.