Labbrapport i Statistik

Laboration 10

732G46

Michael Debebe

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2022-05-06

Innehåll

Introduktion

I denna laboration kommer ett dataset analyseras.

I datasetet som kommer från en läskedryckstillverkare har företaget valt använda sig av fem agenter för att hantera den lyxigare distrubition av diverse produkter. 20 transaktioner för varje agent var slumpmässigt urvalda och tidsperiod som det tog i dagar tills varje transaktion hade gått igenom.

Målet med denna laboration är att skatta och testa faktornivåväntevärden och jämföra dessa. Göra inferens av linjära kontraster av faktornivåväntevärden, förstå vikten av simultan inferens och utföra nödvändig residualanalys.

```
sum(a)
```

```
## Error in eval(expr, envir, enclos): object 'a' not found
```

Databehandling

```
PremiumD <-read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/CH16PR12.txt") #Laddar
colnames(PremiumD) <-c("Days","Agent","Transactions")#Namnger kolumnerna
PremiumD$Agent <-as.factor(PremiumD$Agent)
```

Uppgifter

17.13 Refer to Premium distribution Problem 16.12

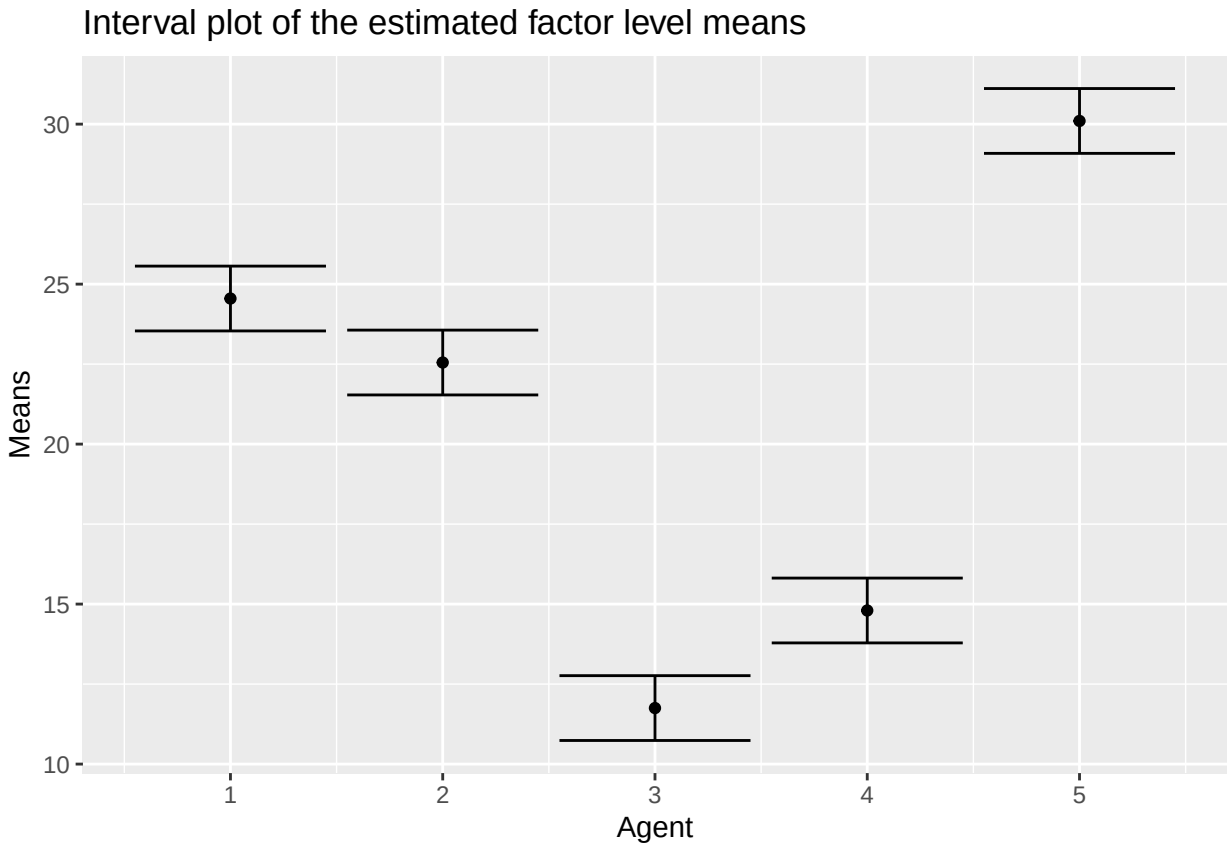
a) Prepare an interval plot of the estimated factor level means Y_bar, where intervals correspond to the confidence limit in 17.7 with a=.10

```
##             Df  Sum Sq  Mean Sq F value     Pr(>F)
## Agent        4 4430.10 1107.525 147.226 < 2.22e-16 ***
## Residuals   95  714.65    7.523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I utskriften ovan syns modellen som skapats och datan som ligger till grund för kommande tolkningar och analys. Det som syns är hur Agent som är variabeln är en starkt signifikant variabel, ett högt f-värde som tyder på att det råder variation inom grupperna.

a) Prepare an interval plot of the estimated factor level means Y_bar, where intervals correspondto the confidence limit in 17.7 with a=.10

```
ggplot(data = tt,aes(x=Agent,y=Means))+
  geom_point()+
  geom_errorbar(aes(ymin=lower,ymax=upper))+
  ggtitle(label = "Interval plot of the estimated factor level means")
```



I diagrammet ovanför syns konfidensintervallen av de olika agenternas medelvärden. Det som tydligt syns är att Agent 5 medelvärde är de högsta medan Agent 3 medelvärde är det lägsta. Det som blir intressant med denna typen av plot är att det till exempel syns hur Agent 1 nedersta konfidensgräns gränsar till Agent 2 övre konfidensgräns.

b) Test for all pairs of factor level means wheter or not they differ; use the Tukey procedure with a =.10.

$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5$

H_a : Minst två utav α skiljer sig från varandra

Formeln som används för att få ut konfidensintervallet för differensen mellan två faktor nivåer med Tukey procedure visas nedan, där $\hat{D}$ är differensen mellan två faktornivåer.

$\hat{D} \pm \frac{q_{k,n-k_i 1-\alpha}}{\sqrt{2}} \cdot \sqrt{s^2\left\{\hat{D}\right\}}$

$s^2\{\hat{D}\} = \frac{MSE}{n}$

$\hat{D}_{2-1} = 22.55 - 24.55 = -2.00$

$\hat{D}_{3-1} = 11.75 - 24.55 = -12.80$

$\hat{D}_{4-1} = 14.80 - 24.55 = -9.75$

$\hat{D}_{5-1} = 30.10 - 24.55 = 5.55$

$\hat{D}_{3-2} = 11.75 - 22.55 = -10.80$

$\hat{D}_{4-2} = 14.80 - 22.55 = -7.75$

$\hat{D}_{5-2} = 30.10 - 22.55 = 7.55$

$\hat{D}_{4-3} = 14.80 - 11.75 = 3.05$

$\hat{D}_{5-3} = 30.10 - 11.75 = 18.35$

$\hat{D}_{5-4} = 30.10 - 14.80 = 15.30$

```
Tukey<-TukeyHSD(AOV,ordered = FALSE,conf.level = 0.90)
Tukey
```

```
##   Tukey multiple comparisons of means
##     90% family-wise confidence level
##
## Fit: aov(formula = Days ~ Agent, data = PremiumD)
##
## $Agent
##       diff           lwr           upr       p adj
## 2-1  -2.00  -4.165429522   0.165429522 0.152049813
## 3-1 -12.80 -14.965429522 -10.634570478 0.000000000
## 4-1  -9.75 -11.915429522  -7.584570478 0.000000000
## 5-1   5.55   3.384570478   7.715429522 0.000000059
## 3-2 -10.80 -12.965429522  -8.634570478 0.000000000
## 4-2  -7.75  -9.915429522  -5.584570478 0.000000000
## 5-2   7.55   5.384570478   9.715429522 0.000000000
## 4-3   3.05   0.884570478   5.215429522 0.005924547
## 5-3  18.35  16.184570478  20.515429522 0.000000000
## 5-4  15.30  13.134570478  17.465429522 0.000000000
```

Genom att i detta fall enbart titta på p.adj syns det hur samtliga jämförelser har väldigt låga justerade p-värden förutom första genomförelsen, med tanke på att de är så extremt låga kan slutsatsen kring att det finns skillnader mellan samtliga grupper förutom emellan grupperna i första jämförelsen, detta är på 10 % signifikansnivå som statistiskt signifikanta differenser existerar. Alltså förkastas enbart H0 för första jämförelsen, på 10 % signifikansnivå, alltså finns det statistikt signifikanta skillnader i medelvärde mellan Agent 1 och Agent 2.

Vid efterfrågan av smalare intervall är Tukey att föredra gentemot Bonferroni, då det är generellt ”starkare”.

c) Construct a 90 percent confidence interval for mean time lapse for Agent 1

$$\mu_1 \pm t(0.90, 95) \cdot \sqrt{\frac{MSE}{n}} \Rightarrow \ 24.55 \pm 1.29 \cdot 0.613 = 23.76 \leq 24.55 \leq 25.34$$

I ett 90 procentigt konfidensintervall kommer medeltidsspannet för Agent 1 att ligga mellan 23.76 och 25.34, för att vara ett 90 % procentigt intervall är spannet relativt litet.

d) Obtain a 90 percent confidence interval for D = u2-u1. Interpret your interval estimate

$$\hat{D}_{2-1} = 22.55 - 24.55 = -2.00$$

$$t(.90, 95) = 1.29$$

$$S\left\{\hat{D}_{ij}\right\} - \sqrt{(MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \Rightarrow \sqrt{\left(7.52 \cdot \left(\frac{1}{20} + \frac{1}{20}\right)\right)} = 0.867$$

$$-2 \pm 1.29(0.867)$$

$$-3.12 : -0.882$$

I ett 90 procentigt konfidensintervall kommer skillnaderna agent 2 och 1 emellan att ligga mellan -3.12 och -0.882, intervallet täcker inte 0 vilket innebär att det är statistiskt signifikant, för att dessutom vara ett 90 % konfidensintervall är intervallet väldigt stort.

17.18 Refer to Premium Distrubition problem 16.12

b) Estimate the following comparisions with 90 percent family confidence coefficent; use the Scheffé procedure

Formeln som användes för att kunna skapa familjekonfidensintervall med scheffes procedure visas nedan

$$\hat{L} \pm S\sqrt{MSE \cdot \sum_{i=1}^{k} \frac{c_i^2}{n_i}}$$

S ges på följande sätt som visas nedan, S används genomgående för samtliga intervall då det är familjekonfidens som beräknas.

$$S = \sqrt{(k-1) \cdot F_{k-1,n-k,1-\alpha}} \Rightarrow \sqrt{4 \cdot 2.005} = 2.832$$

$$F(.90; 4, 90) = 2.005$$

$D_1 = \mu_1 - \mu_2 \Rightarrow 24.55 - 22.55 = 2$

$S\{D_1\} = \sqrt{MSE \cdot \frac{cl^2}{n}} \Rightarrow \sqrt{7.52 \cdot \frac{(1)^2+(1)^2}{20}} = 0.8672$

$D_1 \pm 2.826 \cdot .8672 \Rightarrow \ -0.45065 \leq D_1 \leq 4.45065$

Differensen D1 kommer i ett 90 % familjekonfidensintervall att ligga mellan -0.45065 och 4.45065.

$D_2 = \mu_3 - \mu_4 \Rightarrow 11.75 - 14.80 = -3.05$

$S\{D_2\} = \sqrt{MSE \cdot \frac{c^2}{n}} \Rightarrow \sqrt{7.52 \cdot \frac{(1)^2+(1)^2}{20}} = 0.8672$

$D_2 \pm 2.826 \cdot .8672 \Rightarrow -5.50 \leq D_2 \leq -0.5993$ Differensen D2 kommer i ett 90 % familjekonfidensintervall att ligga mellan -5.50 och -0.5993.

$L_1 = \frac{\mu_1+\mu_2}{2} - \mu_3 \Rightarrow \frac{24.55+22.55}{2} - 30.10 = -6.55$

$S\{L_1\} = \sqrt{MSE \cdot \frac{c^2}{n}} \Rightarrow \sqrt{7.52 \cdot \frac{(2(1/2)^2)+(1)^2}{20}} = 0.751$

$L_1 \pm 2.826 \cdot .751 \Rightarrow \ -8.68 \leq L_1 \leq -4.423$

Differensen L1 kommer i ett 90 % familjekonfidensintervall att ligga mellan -8.68 och -4.423.

$L_2 = \frac{\mu_3+\mu_4}{2} - \mu_5 \Rightarrow \frac{11.75+14.80}{2} - 30.10 = -16.825$

$S\{L_2\} = \sqrt{MSE \cdot \frac{cl^2}{n}} \Rightarrow \sqrt{7.52 \cdot \frac{(2(1/2))^2+(1)^2}{20}} = 0.751$

$L_2 \pm 2.826 \cdot .751 \Rightarrow \ -18.95 \leq L_2 \leq -14.7$ Differensen i intervallet för L2 kommer i ett 90 % familjekonfidensintervall att ligga mellan -18.95 och -14.7

$L_3 = \frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2} \Rightarrow \frac{24.55+22.55}{2} - \frac{11.75+14.80}{2} = 10.275$

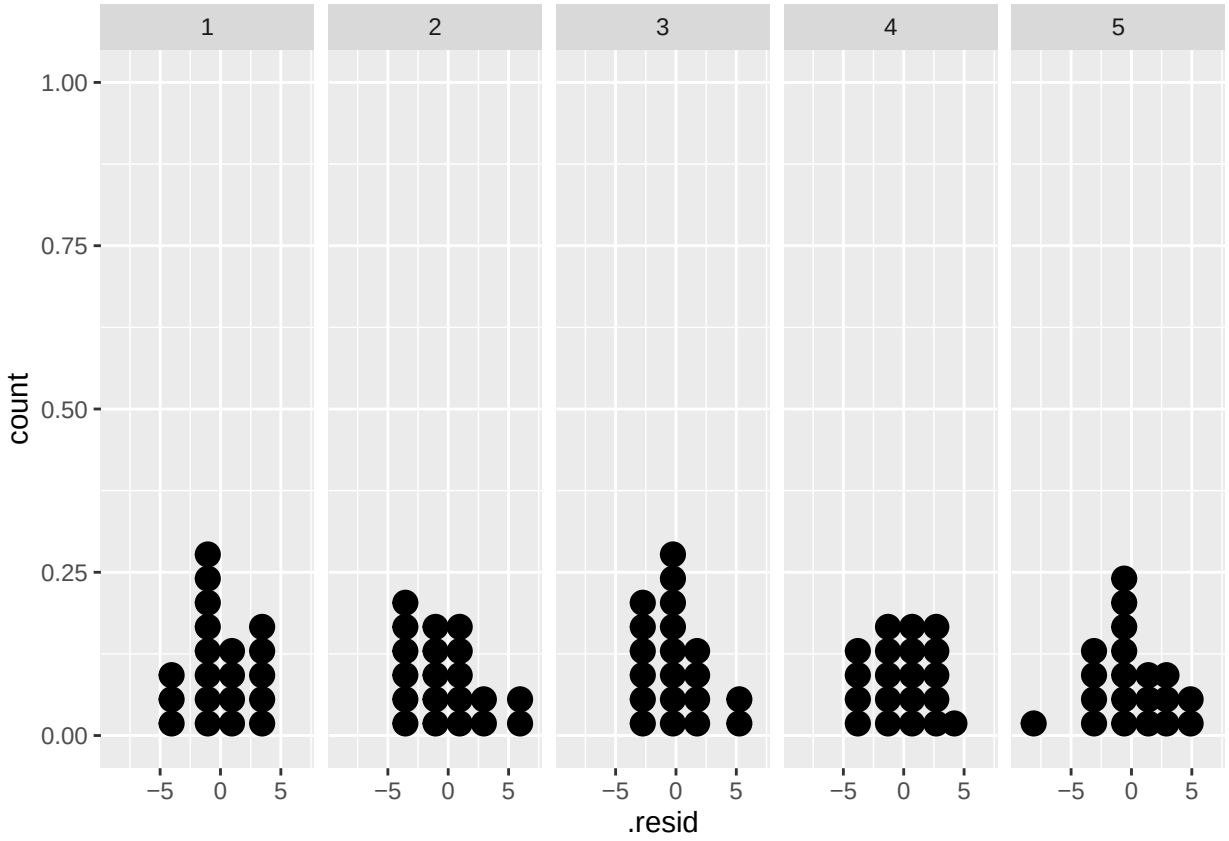
$S\{L_3\} = \sqrt{MSE \cdot \frac{cl^2}{n}} \Rightarrow \sqrt{7.52 \cdot \frac{(4(1/2))^2}{20}} = 0.6131$

$L_3 \pm 2.826 \cdot 0.6131 \Rightarrow \ 8.54 \leq L3 \leq 12.01$ Differensen i intervallet för L3 kommer i ett 90 % konfidensintervall att ligga mellan 8.54 och 12.01

18.9 Refer to Premium Distrubition problem 16.12

a) Obtain the residuals and prepare aligned residual dot plots by agent. What departues from ANOVA model (16.2) can be studied from these plots? What are your findings?

```
ggplot(AOV,aes(.resid))+
  geom_dotplot(binwidth = 8/4)+
  facet_grid(~Agent)
```



I dotploten ovanför syns fördelningarna för de olika Agenternas residualer, ingen av agenterna har vad som kan kallas en normalfördelning över sina residualer, däremot är detta förståeligt med tanke på det lilla urvalet av 20 observationer från varje agent, hade det däremot varit tvåhundra hade en mer precis analys kunnats genomföras. Med det datamaterialet som finns kan dock vissa slutsatser dras; såsom att residualerna för Agent 5 är något negativt skeva. Agent 4 är den agent som har "bäst spridning" bland sina residualer, däremot är dessa residualer långt ifrån normalfördelade dem med.

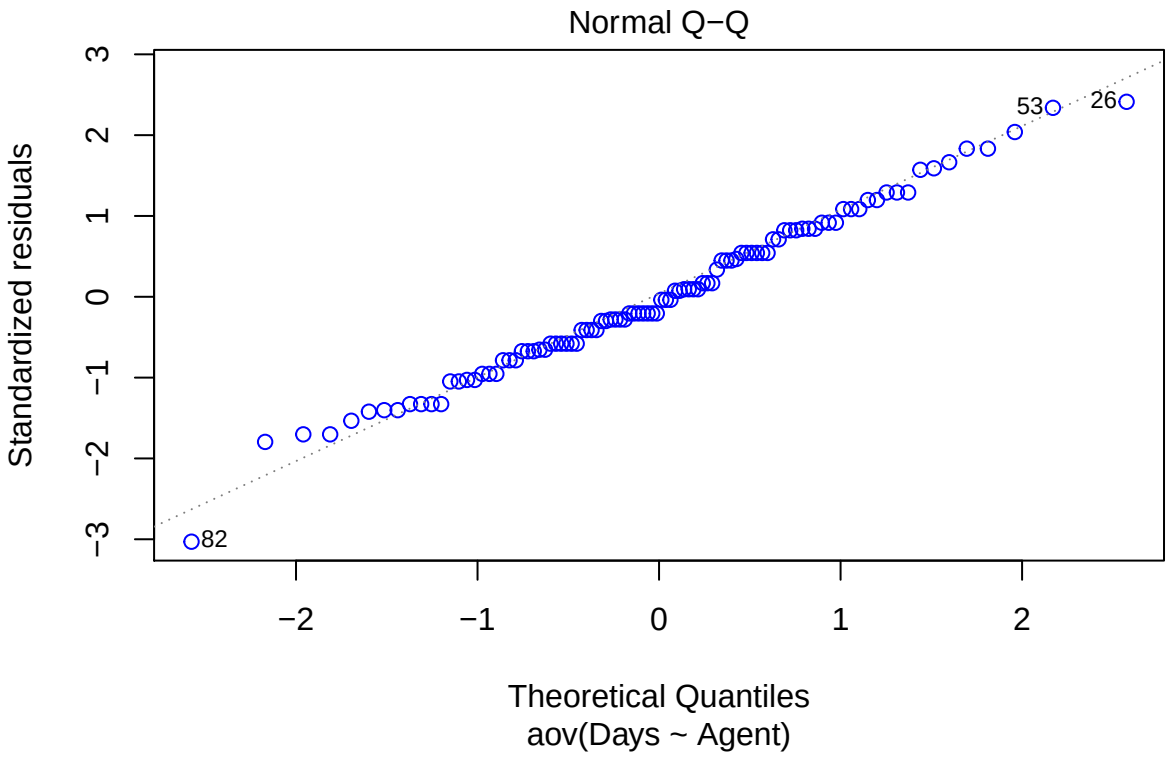
b) Prepare a normal probability plot of the residuals. Also obtain the coefficent of correlation between the ordered residuals and their expected values under normality. DOes the normality assumption appear to be reasonable here?

```
qqp <-qqnorm(rstudent(AOV),plot.it = FALSE)
cor(qqp$x,qqp$y)
```

```
## [1] 0.994406746
```

En väldigt hög korrelation mellan residualerna och deras förväntande värden returneras, det går nästan att säga att dom direkt beror på varandra (Korrelerar), däremot kan vi inte säga att det finns kausalitet mellan dom på just enbart detta värde.

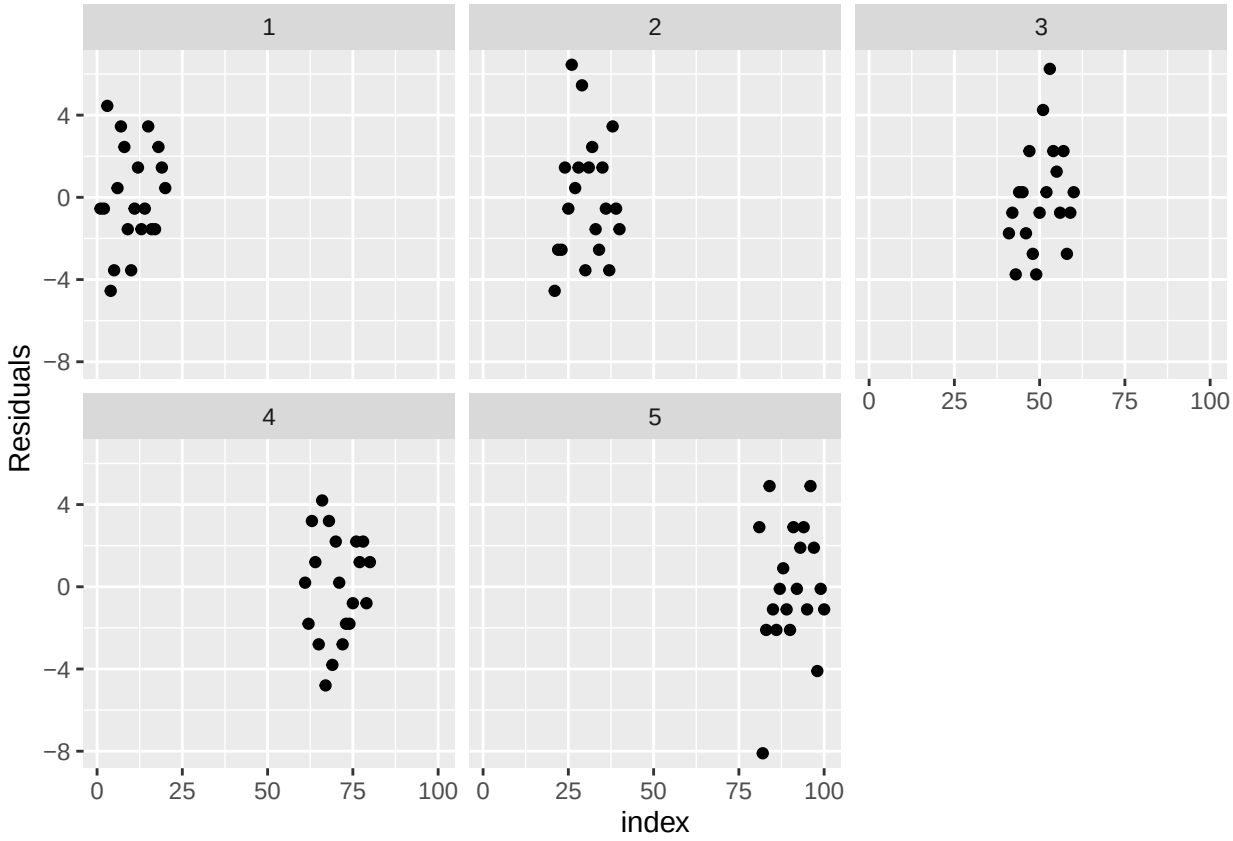
```
plot(AOV,which = 2,col="blue")
```



Ovan visas en normalfördelningstabell utav residualerna för modellen, det som blir intressant är hur korrelationen mellan variablerna nästan var 1 men ändå returneras diagrammet med flera avvikande värden och en moderat linje.

c) The observations within each factor level are in time order. Prepare residual sequences plot and interpret them. What are your findings?

```
index <-1:length(AOV$residuals)
KK <-as.data.frame(cbind(AOV$residuals,index,PremiumD$Agent))
ggplot(KK,aes(x=index,y=`V1`),color=Agent)+geom_point()+facet_wrap(~PremiumD$Agent)+ylab("Residuals")
```



Gemensamt ser variansen jämn ut för residualerna, detta trots att Agent 2,3 och 5 har avvikande observationer. Bortsett från detta syns en jämn och fin varians i tabellen vilket medför att de kan ses som normalfördelade. Totalt rör det sig om ca 3 avvikande observationer vilket för ett dataset på 100 observationer inte ses som överhängande fara.

Lärdomar

Målet med denna laboration är att skatta och testa faktornivåväntevärden och jämföra dessa. Göra inferens av linjära kontraster av faktornivåväntevärden, förstå vikten av simultan inferens och utföra nödvändig residualanalys.