Labbrapport i Statistik

Laboration 9

732G46

Mattias Hällgren, Michael Debebe

Avdelningen för Statistik och maskininlärning
Institutionen för datavetenskap
Linköpings universitet

2021-11-07

Innehåll

Introduktion

I denna laboration kommer två dataset analyseras.

I det första datasetet kommer data från en studie där längden av sjukhusvistelser att analyseras i form av att respondenternas inkomstnivå kommer att tas i hänsyn. Detta genom att dela upp respodenterna i 4 olika grupper.

I det andra 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 kunnna formulera en envägs variansanalysmodell på två olika sätt och förklara hur modellparametrarna kan tolkas. Använda en datamängd och göra statistisk slutledning om modellens parameterar genom att använda R, samt förstå sambandet mellan en variansanalysmodell och en regressionsmodell.

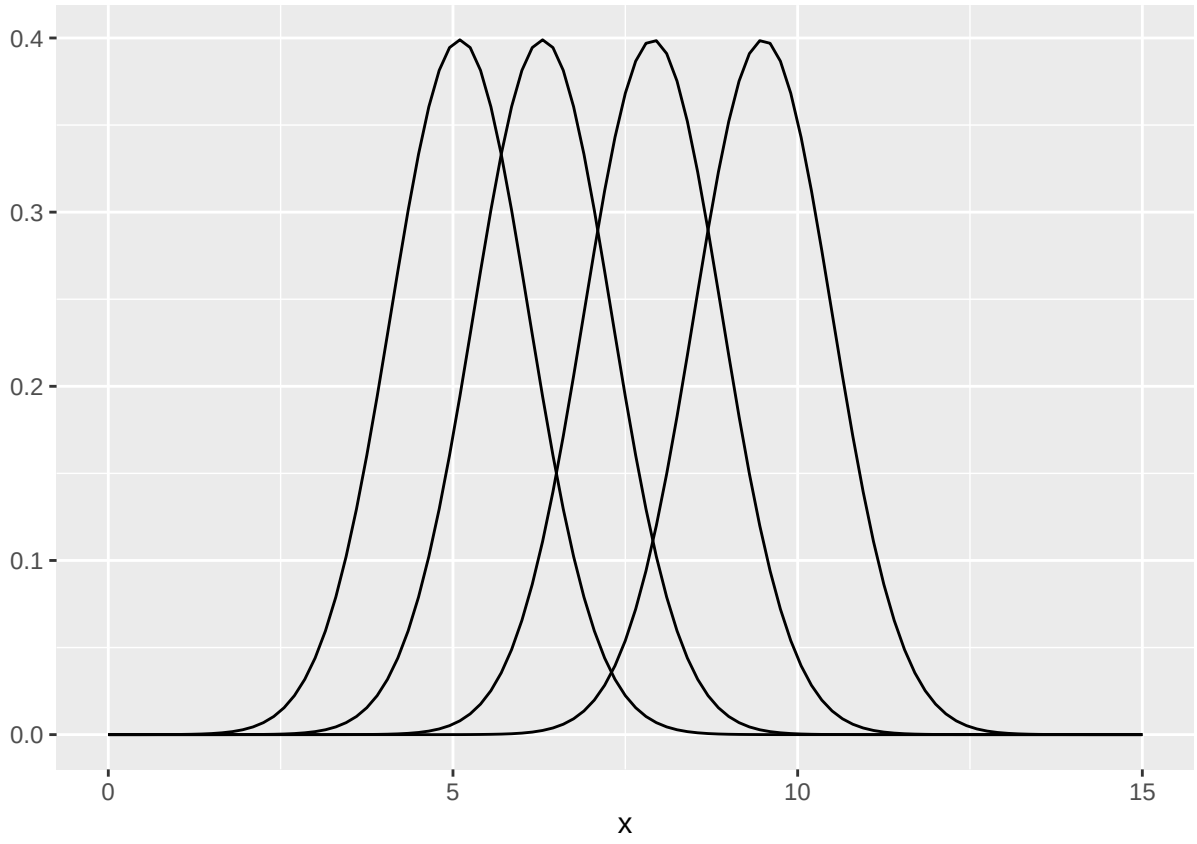
Databehandling

```
# datahantering
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)
sexsju <-read.table("https://raw.githubusercontent.com/MichaelDebebe/6555/main/16_7")
colnames(sexsju) <-c("low","mod","high")
```

16.5 In a study of length of hospital stay (in number of days) of persons in four income groups

a) Draw a representation of this model in the format of figure 16.2

```
ggplot(data = data.frame(x = c(0,15)), aes(x)) +
  stat_function(fun = dnorm, args = list(mean = 9.5),sd=2.8) + ylab("") +
   stat_function(fun = dnorm, args = list(mean = 5.1),sd=2.8) + ylab("")+
  stat_function(fun = dnorm, args = list(mean = 6.3),sd=2.8) + ylab("")+
  stat_function(fun = dnorm, args = list(mean = 7.9),sd=2.8) + ylab("")
```



Ovan syns normalfördelningstabellen för de 4 olika grupperna, samtliga grupper har samma populationsstandardavvikelse, då samtliga gruppers 95 % konfidens intervall berör varandra, skär samtliga linjer varandra, vilket även kan ses i grafen.

Från laborationen går det att i figur 3 läsa om hur grafens utseende innebär att samplingfördelningen för de fyra grupperna skiljer sig åt.

b) Suppose that 100 persons from each income group are randomly selected for the study. Find E{MSTR} and E{MSE}.

```
b_upp <-c(5.1,6.3,7.9,9.5)
zigma <-2.8
k <-4
medel <-sum((b_upp-mean(b_upp))^2)
E_MSTR <-zigma^2+(100*medel)/(k-1)
E_MSTR
```

```
## [1] 374.51
```

```
E_MSE <-zigma^2
E_MSE
```

```
## [1] 7.84
```

Man kan tydligt se att MSTR har ett högre värde än MSE. Detta betyder att det är en betydlig större skillnad när man jämför grupperna med varandra än att jämföra grupperna var för sig.

c) If u2 = 5.6 and u3 = 9.0, what would E{MSTR} be? Why is E{MSTR} substaintially larger than E{MSE} here? What is the implication of this?

```
c_upp <-c(5.1,5.6,9.0,9.5)
c_medel <-sum((c_upp-mean(c_upp))^2)
E_MSTR_c <-zigma^2+(100*c_medel)/(k-1)
E_MSTR_c
```

```
## [1] 523.17
```

Detta eftersom att avstånden grupperna emellan ökar då ett ganska centralt medelvärde sjunker och det andra centrala värdet ökar, alltså ökas avståenden bland medelvärdena “dubbelt”.

16.16 Refer to the Problem 16.5. What are the values of ti if the ANOVA model is expressed in the factor effects formulation (16.62) and u is defined by (16.63)?

```
Medel_varde <-mean(b_upp)
Medel_varde
```

```
## [1] 7.2
```

```
ti_value <-b_upp-Medel_varde
rbind(b_upp,ti_value)
```

```
##          [,1] [,2] [,3] [,4]
## b_upp     5.1  6.3  7.9  9.5
## ti_value -2.1 -0.9  0.7  2.3
```

```
b_upp
```

```
## [1] 5.1 6.3 7.9 9.5
```

$$Medelvärde = (5.1 + 6.3 + 7.9 + 9.5)/4 = 7.2$$

Medelvärdet räknas ut genom att ta parametrarna delat med antalet så får vi ut värdet 7.2

$$ti1 = 5.1 - 7.2 = -2.1$$
$$ti2 = 6.3 - 7.2 = -0.9$$
$$ti3 = 7.9 - 7.2 = 0.7$$
$$ti4 = 7.9 - 7.2 = 2.3$$

Ti-värdena ges genom att subrathera medelvärdet från vektorn med siffror, den vektorn som ges visas på tredje raden bland siffrorna ovan.

16.12 Premium distribution

b) Obtain the fitted values

```
AOV <-aov(Days~Agent, data=PremiumD)
Coef <-coef(AOV)
Coef
```

```
## (Intercept)      Agent2      Agent3      Agent4      Agent5
##       24.55       -2.00      -12.80       -9.75        5.55
```

En modell anpassas och det intressanta blir att de olika gruppernas medelvärden fås ut på sådant sätt att den första gruppen blir interceptet, därefter adderas de olika summorna till interceptet för att få ut de enskilda gruppernas medelvärden, som visas nedan.

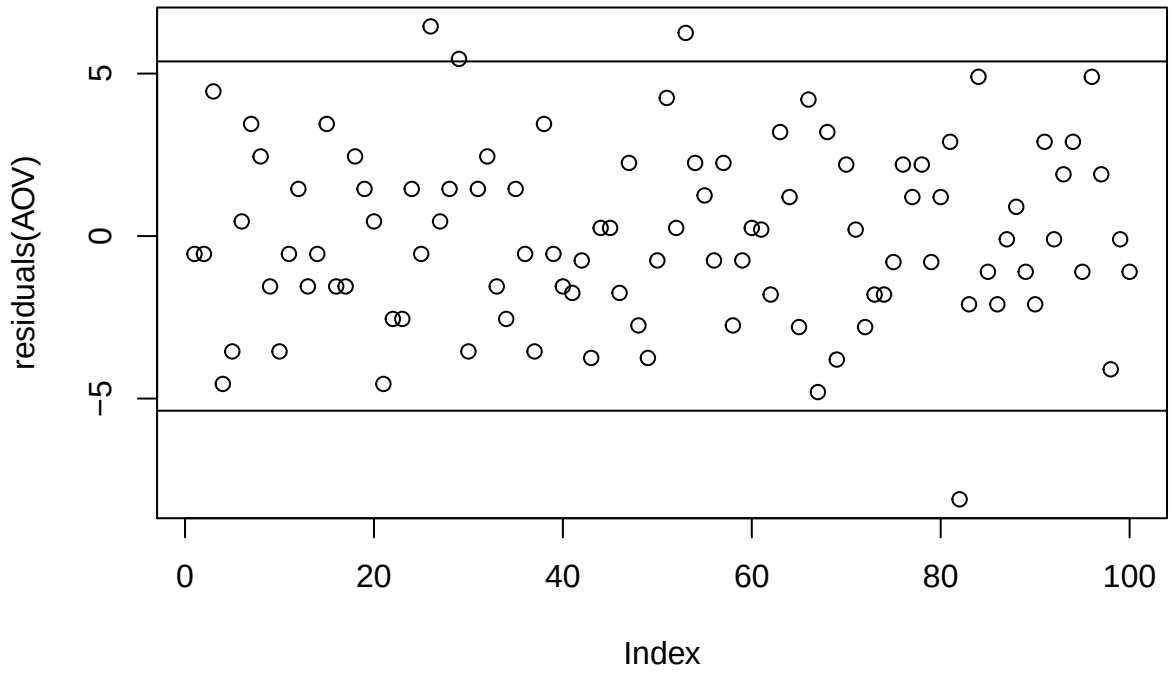
```
InterCP<-c(0,24.55,24.55,24.55,24.55)
Fitted_values <-Coef+InterCP
Fitted_values
```

```
## (Intercept)      Agent2      Agent3      Agent4      Agent5
##       24.55       22.55       11.75       14.80       30.10
```

Det som syns utifrån denna vektor av siffror är hur agent 5 har ett väldigt högt medelvärde medan agent 3 har det lägsta medelvärdet följt av agent 4 och därefter 2.

c) Obtain the residuals, do they sum to zero?

```
SSD <-sd(residuals(AOV))
plot(residuals(AOV))
abline(h=2*SSD)
abline(h=-2*SSD)
```



I grafen som visas ovan, syns residualerna för modellen, med två linjer som ska symbolisera två standardavvikelser under medelvärdet och två standardavvikelser över medelvärdet, det som syns är att majoriteten av observationeran ligger innanför linjerna. Däremot är det tre observationer som ligger utanför, vilket inte alls är särskilt farligt, hade det däremot varit 10 + hade datasetet kunna ifrågasättas. Däremot är den tredje observationen som också är den nedre en väldigt kraftigt avvikande observation.

c) Obtain the residuals, do they sum to zero

```
sum(residuals(AOV))
```

```
## [1] 3.8025e-15
```

Residualerna summerar till 0.00000116319 vilket tolkas som 0.

d) Obtain the analysis of the variance table

```
anova(aov(Days~Agent, data=PremiumD))
```

```
## Analysis of Variance Table
##
## Response: Days
##           Df Sum Sq Mean Sq F value Pr(>F)
## Agent      4   4430    1108     147 <2e-16 ***
## Residuals 95    715       8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I tabellen som returneras får returneras faktumet kring att Agent är en signifikant variabel, ett väldigt lågt P-värdet returernas också.

e) Test wheter or not the mean time lapse differs for five agent: use a = .10

State the alternatives, decision rule and conclusion.

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5 = 0$$

$$H_1 : \tau_1 \neq \tau_2 \neq \tau_3 \neq \tau_4 \neq \tau_5 = 0$$

```
anova(aov(Days~Agent, data=PremiumD))
```

```
## Analysis of Variance Table
##
## Response: Days
##           Df Sum Sq Mean Sq F value Pr(>F)
## Agent      4   4430    1108     147 <2e-16 ***
## Residuals 95    715       8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$F = \frac{MSR}{MSE} = \frac{1107.5}{7.5} = 147.7$$
$$F(.90, 4, 95) = 2.005$$

I detta fall överstiger teststatistiskan det kritiska värdet $147 > 2.005$ vilket innebär att vi kan förkasta H_0 om att alla agenter har likadana tidsspan och dra slutsatsen om att det skiljer sig agenterna emellan.

f) What is the P-value of the test in part (e)? Explain how the same conclusion

as in part (e) can be reached by knowing the P-value.

```
anova(aov(Days~Agent, data=PremiumD))
```

```
## Analysis of Variance Table
##
## Response: Days
##           Df Sum Sq Mean Sq F value Pr(>F)
## Agent      4   4430    1108     147 <2e-16 ***
## Residuals 95    715       8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Genom att anovatabellerna innehåller teststatisikor från de testet som de genomför, i detta fall är det f-test som genomförs medan det i en lm-modell genomförs t-test. Genom att få ut ett så lågt p-värde 2^-16 samtidigt som att tre stjärnor visas till höger om variabeln Agent visas det hur variabeln är en signifikant variabel, samt på vilken signifikansnivå variabeln är signifikant.

Lärdomar

Målet med denna laboration var att kunnna formulera en envägs variansanalysmodell på två olika sätt och förklara hur modellparametrarna kan tolkas. I efterhand känner vi att vi har skapat oss en bra uppfattning i hur envägs- variansanalyser genomförs, detta var många steg vi körde fast på i denna labb men det löste sig!