Linear Statistical Models

Week-3: Graded Assignment

Subjective Assignment: (Manual-grading) **Max. Marks: 35**

NOTE: All the plots should be properly labelled.

1. Consider the Linear Model:

$$y_{ij} = \mu + \mu_i + \epsilon_{ij} \quad ; \quad 1 \leq j \leq n_i \, , \, 1 \leq i \leq 3$$

Feel free to use the draft code available in the 'Supplementary Contents' → 'Week-3' on the portal.

(a) If $\epsilon_{ij} \sim N(0, 1)$, then find the distribution of y_{ij}. Elaborate on your answer. [3 Marks]

Solution:

Since, $\epsilon_{ij} \sim N(0, 1)$ and linear model is

$$y_{ij} = \mu + \mu_i + \epsilon_{ij} \quad ; \quad 1 \leq j \leq n_i \, , \, 1 \leq i \leq 3$$

Now,

$$E(y_{ij}) = E(\mu + \mu_i + \epsilon_{ij})$$
$$E(y_{ij}) = \mu + \mu_i + E(\epsilon_{ij})$$

Since, it is given that $E(\epsilon_{ij}) = 0$. Therefore, on putting the value of $E(\epsilon_{ij}) = 0$ in the above equation, we get

$$E(y_{ij}) = \mu + \mu_i \quad \ldots (*)$$

Now,

$$Var(y_{ij}) = Var(\mu + \mu_i + \epsilon_{ij})$$
$$Var(y_{ij}) = 0 + 0 + Var(\epsilon_{ij})$$

Since, it is given that $Var(\epsilon_{ij}) = 1$. Therefore, on putting the value of $Var(\epsilon_{ij}) = 1$ in the above equation, we get

$$Var(y_{ij}) = 0 + 0 + 1 = 1$$

Thus, the distribution of y_{ij} is $N(\mu + \mu_i, 1)$.

(b) Generate possible random values using R for μ, μ_i and n_i as generated in code draft provided. [1 Mark]

Note: *The value of n_i's should be greater than or equal to 100.*

Solution:

```
#Importing libraries
library(purrr)
install.packages("qpcR")
library("qpcR")
 k <- 3
 n<- rdunif(k, 120, 120) #Generating n_i
 mu<- rdunif(1, 10, -10) #Generating mu
 mu_i<-rdunif(k, 20, -20) #Generating mu_i
```

(c) Using R, generate the dataset, i.e. y_{ij}, and store it in a data frame. [3 Marks]

Solution:

```
for ( i in 1:k){
+ y_ij=rnorm(n[i],mean = mu+mu_i[i],sd = 1)
+ assign(paste0("y_",i),y_ij)
+ }
Y <- list(y_1,y_2,y_3)
```

(d) Using R, find the mean of y_i's denoted by y_{io} using the generated dataset.

Hint: $y_{io} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$ [1 Mark]

Solution:

```
y_i0=numeric(3)
for(i in 1:k){
+ y_i0[i]=mean(Y[[i]])
+ }
```

(e) Using R, compute the following: [3 Marks]

i. Sum of squares within group, i.e. SSW

Solution:

```
SSW=0
for(i in 1:k){
+ SSW = SSW +(sum((Y[[i]]-y_i0[i])^2))
+ }
print(SSW)
```

ii. Sum of squares between group, i.e. SSB

Solution:

```
y_bar=mean(c(y_1,y_2,y_3))
SSB=0
for(i in 1:k){
+ SSB = SSB + (n[i] * (y_i0[i]-y_bar)^2)
+ }
print(SSB)
```

iii. Total sum of squares, i.e. TSS

Solution:

```
TSS =0
for(i in 1:k){
+ TSS = TSS + (sum((Y[[i]]-y_bar)^2))
+ }
print(TSS)
```

(f) Using R, verify that $TSS = SSB + SSW$. [1 Mark]

Solution:
We can verify the $TSS = SSW + SSB$ by the following code:

```
all.equal(TSS,SSW+SSB)
```

(g) Iterate the steps in part $(c), (d)$ and (e) more than 100 times using R. [5 Marks]
***Hint**: Randomly generate a number greater than 100, and perform that many iterations.*

Solution:

```
rep=rdunif(1,102,100)
ssw=c()
ssb=c()
tss=c()
for(i in 1:rep){
+ for(i in 1:k){
+ y_ij=rnorm(n[i],mean=mu2+mu_i[i],sd=1)
+ assign(paste0("y_",i),y_ij)
+ }
+ Y = list(y_1,y_2,y_3)
+ SSW=0
+ for(i in 1:k){
+ SSW = SSW +(sum((Y[[i]]-y_i0[i])^2))
+ }
+ ssw=append(ssw, SSW)
+ y_bar=mean(c(y_1,y_2,y_3))
+ SSB=0
+ for(i in 1:k){
+ SSB = SSB + (n[i] * (y_i0[i]-y_bar)^2)
```

```
+ }
+ ssb=append(ssb,SSB)
+ TSS =0
+ for(i in 1:k){
+ TSS = TSS + (sum((Y[[i]]-y_bar)^2))
+ }
+ tss = append(tss, TSS)
+ }
```

(h) Using R, store the values of SSW, SSB and TSS computed in part (g) in a data frame. [2 Marks]

Solution:
We can store the values of SSW, SSB and TSS computed in part (g) in a data frame by the following code:

```
df=data.frame(ssw,ssb,tss)
```

(i) Using R, plot the histogram of values obtained for each of SSW, SSB and TSS. [3 Marks]

Solution:

```
library(ggplot2)
ggplot(data=df,aes(x=ssw))+geom_histogram()
ggplot(data=df,aes(x=ssb))+geom_histogram()
ggplot(data=df,aes(x=tss))+geom_histogram()
```

We can add title of the histogram by adding ggtitle() function.

(j) In each of the above plotted histogram, fit the density curve of 'Chi-square Distribution' to verify the following: [8 Marks]

i. $SSW \sim \chi^2$ with degree of freedom $= 2$.

Solution:

```
library(ggplot2)
ssw_chisq=ggplot(data=df,aes(x=ssw))+geom_histogram()
+stat_function(fun=dchisq,args=list(df=2))
ssb_chisq=ggplot(data=df,aes(x=ssb))+geom_histogram()
+stat_function(fun=dchisq,args=list(df=sum(n)-3))
tss_chisq=ggplot(data=df,aes(x=tss))+geom_histogram()
+stat_function(fun=dchisq,args=list(df=sum(n)-1))
```

ii. $SSB \sim \chi^2$ with degree of freedom $= n - 3$.
where, $n = n_1 + n_2 + n_3$

Solution:

```
ssb_chisq=ggplot(data=df,aes(x=ssb))+geom_histogram()
+stat_function(fun=dchisq,args=list(df=sum(n)-3))
```

iii. $TSS \sim \chi^2$ with degree of freedom $= n-1$.

Solution:

```
tss_chisq=ggplot(data=df,aes(x=tss))+geom_histogram()
+stat_function(fun=dchisq,args=list(df=sum(n)-1))
```

Note: Based on the obtained plot comment on each of the obtained histograms.

2. Suppose $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$ be a linear model, where

$$\underset{\sim}{Y} = (y_1, y_2, y_3, \ldots, y_n)^T,$$

$$\underset{\sim}{\beta} = (\beta_1, \beta_2, \beta_3, \ldots, \beta_m)^T,$$

$$X = ((x_{ij})) \ ; 1 \leq i \leq n, 1 \leq j \leq m$$

and

$$\underset{\sim}{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3, \ldots, \epsilon_n)^T$$

where, $\underset{\sim}{\epsilon}$ has mean $\underset{\sim}{0}$ and variance covariance matrix $\sigma^2 I_{n\times n}$.

Let $\underset{\sim}{p}^T\underset{\sim}{\beta} = p_1\beta_1 + p_2\beta_2 + \ldots + p_m\beta_m \ \forall \ \underset{\sim}{p} \in \mathbb{R}^m$, then for a $\underset{\sim}{l} \in \mathbb{R}^n$ show that:

$$E[\underset{\sim}{l}^T\underset{\sim}{Y}] = p^T\underset{\sim}{\beta}, \ \ \forall \ \underset{\sim}{\beta} \in \mathbb{R}^m$$

. [5 Marks]

Solution:

We have to show that

$$E[\underset{\sim}{l}^T\underset{\sim}{Y}] = p^T\underset{\sim}{\beta}, \ \ \forall \ \underset{\sim}{\beta} \in \mathbb{R}^m$$

Now,

$$\underset{\sim}{l}^T\underset{\sim}{Y} = (l_1, l_2, \ldots, l_n)\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$= l_1y_1 + l_2y_2 + \ldots + l_ny_n$$

Since, we know that $E(\underset{\sim}{Y}) = X\underset{\sim}{\beta}$. Thus,

$$E[\underset{\sim}{l}^T\underset{\sim}{Y}] = E[l_1y_1 + l_2y_2 + \ldots + l_ny_n]$$

$$E[\underset{\sim}{l}^T \underset{\sim}{Y}] = l_1 E[y_1] + l_2 E[y_2] + \ldots + l_n E[y_n]$$

$$= l_1(x_{11}\beta_1 + x_{12}\beta_2 + \ldots + x_{1m}\beta_m) + \ldots + l_n(x_{n1}\beta_1 + x_{n2}\beta_2 + \ldots + x_{nm}\beta_m)$$

$$= (l_1 x_{11} + l_2 x_{21} + \ldots + l_n x_{n1})\beta_1 + \ldots + (l_1 x_{1m} + l_2 x_{12} + \ldots + l_n x_{nm})\beta_m$$

$$= p_1\beta_1 + p_2\beta_2 + \ldots + p_m\beta_m$$

$$= \underset{\sim}{p}^T \underset{\sim}{\beta} \quad \text{where } \underset{\sim}{p} = (p_1, p_2, p_3, \ldots, p_m)^T \in \mathbb{R}^m$$

Thus,

$$E[\underset{\sim}{l}^T \underset{\sim}{Y}] = \underset{\sim}{p}^T \underset{\sim}{\beta}, \ \ \forall \ \underset{\sim}{\beta} \in \mathbb{R}^m$$