A4 - Anàlisi de la variància i repàs del curs

Proposta de solució

Semestre 2019.2

Índex

1 Introducció

Les dades que es faran servir per aquesta activitat correspon a les vendes de seients de cotxes infantils a 400 botigues diferents. Les variables són:

- Sales (Vendes unitàries, en milers, a cada ubicació)
- CompPrice (Preu cobrat pel competidor a cada ubicació)
- Income (Nivell d'ingressos comunitaris, en milers de dòlars)
- Advertising (Pressupost de publicitat local de l'empresa a cada ubicació, en milers de dòlars)
- Population (Mida de la població a la regió, en milers)
- Price (Preu per seients de cotxes a cada lloc)
- ShelveLoc (Un factor amb nivells Bad, Good i Medium que indica la qualitat de la ubicació dels prestatges dels seients del cotxe de cada lloc)
- Age (Edat mitjana de la població local)
- Education (Nivell educatiu a cada lloc)
- Urban (Un factor amb els nivells Yes i No per indicar si la botiga es troba en una ubicació urbana o rural)
- US (Un factor amb els nivells Yes i No per indicar si la botiga es troba als EUA o no)

Les dades de l'estudi estàn a l'arxiu `ChildCarSeats1.csv`

Nota: important a tenir en compte per a lliurar l'activitat:

Cal lliurar el fitxer Rmd i el fitxer de sortida (PDF o html). El fitxer de sortida ha d'incloure: el codi i el resultat de l'execució de la mateixa (pas a pas). S'ha de respectar la mateixa numeració dels apartats que l'enunciat.

2 Estadística descriptiva i visualització

a) Comproveu el tipus de variable que correspon a cada una de les variables. Quines són tipus numèric? Quines són tipus factor? Hi ha dades faltants?
b) Realitzeu una taula de les dades quantitatives on apareixi la mitja, la mitjana, la desviació standard i l'amplitud interquartílica (IQR, en anglès). Comenta els resultats.
c) Mostreu amb diversos diagrames de caixa la distribució de la variable `Sales` segons: `ShelveLoc`, `Urban` i `US`. Interpretar els gràfics breument.
d) Representeu gràficament les variables qualitatives.

```
mydata<-read.csv("ChildCarSeats1.csv", header=TRUE)
str(mydata)
```

```
## 'data.frame':    400 obs. of  11 variables:
##  $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
##  $ CompPrice  : int  138 111 113 117 141 124 115 136 132 132 ...
##  $ Income     : int  73 48 35 100 64 113 105 81 110 113 ...
##  $ Advertising: int  11 16 10 4 3 13 0 15 0 0 ...
##  $ Population : int  276 260 269 466 340 501 45 425 108 131 ...
##  $ Price      : int  120 83 80 97 128 72 108 120 124 124 ...
##  $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 3 2 3 3 ...
##  $ Age        : int  42 65 59 55 38 78 71 67 76 76 ...
##  $ Education  : int  17 10 12 14 13 16 15 10 10 17 ...
##  $ Urban      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
```

```
##  $ US          : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...
```

```
summary(mydata)
```

```
##      Sales          CompPrice       Income        Advertising
##  Min.   : 0.000   Min.   : 77   Min.   : 21.00   Min.   : 0.000
##  1st Qu.: 5.390   1st Qu.:115   1st Qu.: 42.75   1st Qu.: 0.000
##  Median : 7.435   Median :125   Median : 69.00   Median : 5.000
##  Mean   : 7.410   Mean   :125   Mean   : 68.66   Mean   : 6.635
##  3rd Qu.: 9.160   3rd Qu.:135   3rd Qu.: 91.00   3rd Qu.:12.000
##  Max.   :16.270   Max.   :175   Max.   :120.00   Max.   :29.000
##    Population        Price        ShelveLoc        Age          Education
##  Min.   : 10.0   Min.   : 24.0   Bad   : 96   Min.   :25.00   Min.   :10.0
##  1st Qu.:139.0   1st Qu.:100.0   Good  : 85   1st Qu.:39.75   1st Qu.:12.0
##  Median :272.0   Median :117.0   Medium:219   Median :54.50   Median :14.0
##  Mean   :264.8   Mean   :115.8                Mean   :53.32   Mean   :13.9
##  3rd Qu.:398.5   3rd Qu.:131.0                3rd Qu.:66.00   3rd Qu.:16.0
##  Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.0
##  Urban       US
##  No :118   No :142
##  Yes:282   Yes:258
##
##
##
##
```

a) Les variables `Sales`, `CompPrice`, `Income`, `Advertising`, `Population`, `Price`, `Age` i `Education` són variables quantitatives de tipus `int` o `num`, les altres són qualitatives de tipus `factor`. No hi ha dades faltants.

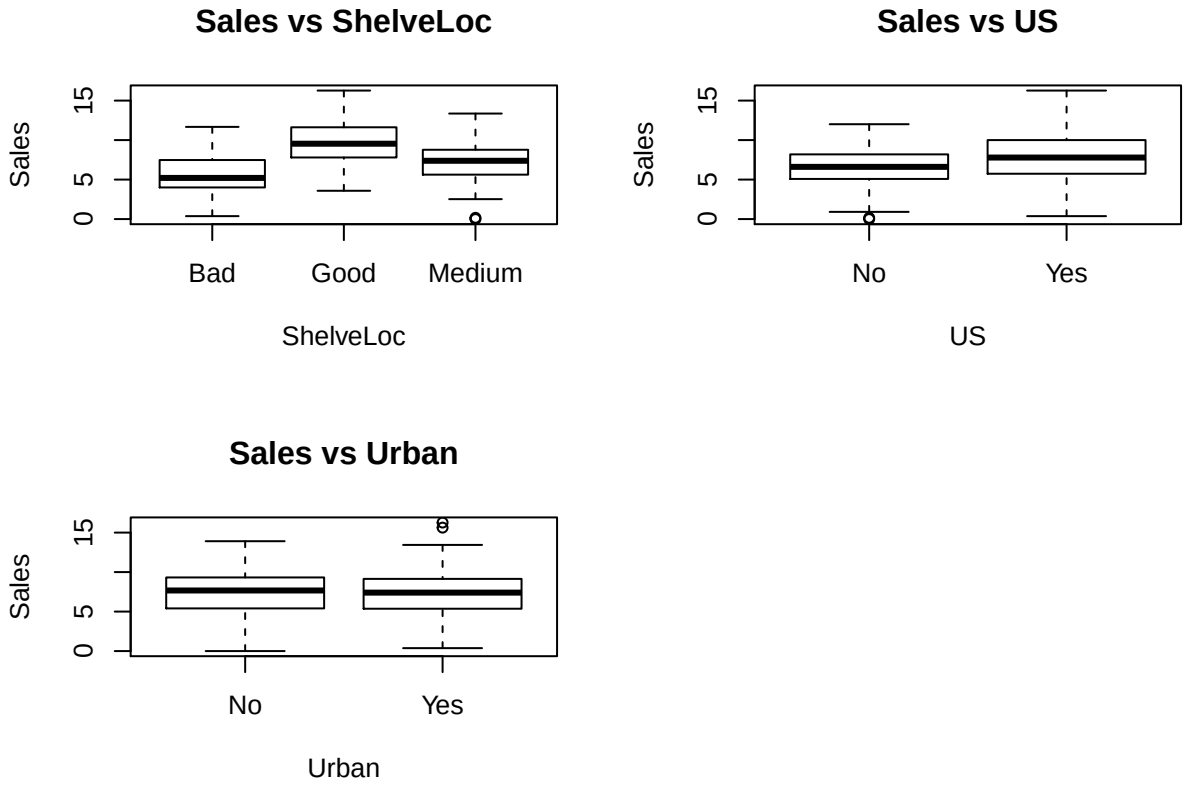
```
res <- c(1:6,8:9)
mean.n <- as.vector(sapply( mydata[,res ],mean,na.rm=TRUE ) )
std.n <- as.vector(sapply(mydata[,res ],sd, na.rm=TRUE))
median.n <- as.vector(sapply(mydata[,res],median, na.rm=TRUE))
IQR.n <- as.vector(sapply(mydata[,res],IQR, na.rm=TRUE))

kable(data.frame(variables= names(mydata)[res],
                 Mean = mean.n,
                 Median = median.n,
                 Standard.Dev = std.n,
                 IQR = IQR.n
                 ),
digits=2, caption="Estadístics de Tendència Central i Dispersió")
```

Taula 1: Estadístics de Tendència Central i Dispersió

variables	Mean	Median	Standard.Dev	IQR
Sales	7.41	7.44	2.73	3.77
CompPrice	124.97	125.00	15.33	20.00
Income	68.66	69.00	27.99	48.25
Advertising	6.63	5.00	6.65	12.00
Population	264.84	272.00	147.38	259.50
Price	115.80	117.00	23.68	31.00
Age	53.32	54.50	16.20	26.25
Education	13.90	14.00	2.62	4.00

b) Els valor de mitja i mediana són bastant similars en totes les variables, on hi ha més diferència és en `Population` i `Advertising`. En canvi, l'estimació del valor de dispersió per desviació estandard és més petita que el IQR. També s'observa que les variables anteriors és on hi ha més diferències entre aquestes estimacions.

```
#c
par(mfrow=c(2,2))
boxplot(Sales ~ ShelveLoc , data = mydata, main = "Sales vs ShelveLoc")
boxplot(Sales ~ US , data = mydata, main = "Sales vs US")
boxplot(Sales ~ Urban , data = mydata, main = "Sales vs Urban")
par(mfrow=c(1,1))
```

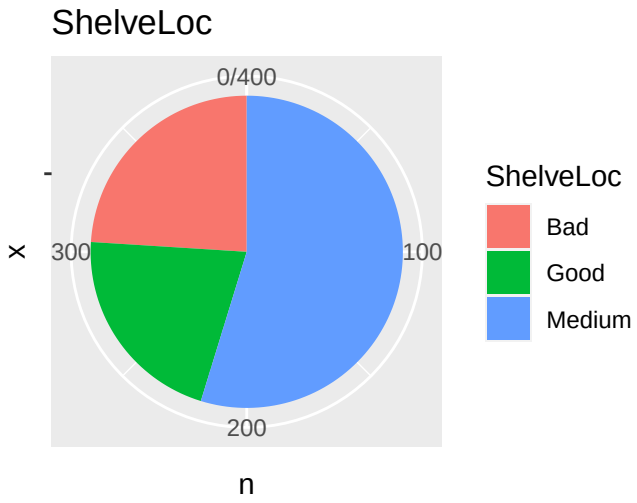


c) La localització del producte en l'expositor, `ShelveLoc`, mostra que influeix positivament a les vendes, és a dir, a millor localització més vendes. En següent gràfic, sembla que hi ha més vendes a les botigues d'EUA que fora d'EUA. Per últim, les vendes són molt similars entre botigues urbanes o no urbanes.

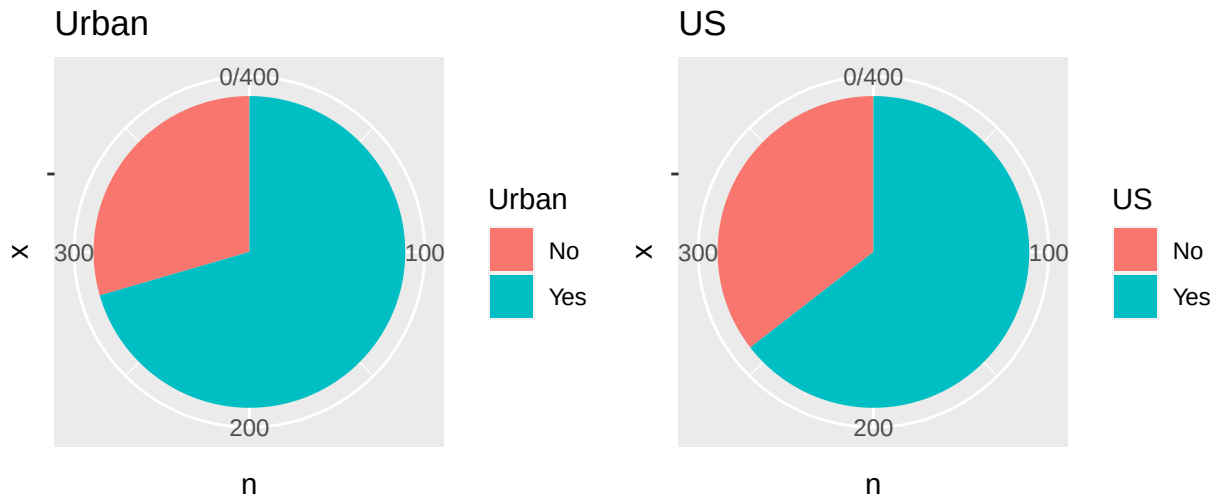
```
#d

kk <- summarize( group_by(mydata, ShelveLoc), n=length(ShelveLoc))
g1 <- ggplot( kk, aes(x="", y=n, fill=ShelveLoc)) +
geom_bar(width = 1, stat = "identity") +
coord_polar("y", start=0) + ggtitle("ShelveLoc")

kk <- summarize( group_by(mydata, Urban), n=length(Urban))
g2 <- ggplot( kk, aes(x="", y=n, fill=Urban)) +
geom_bar(width = 1, stat = "identity") +
coord_polar("y", start=0) + ggtitle("Urban")
```

```
kk <- summarize( group_by(mydata, US), n=length(US))
g3 <- ggplot( kk, aes(x="", y=n, fill=US)) +
geom_bar(width = 1, stat = "identity") +
coord_polar("y", start=0) + ggtitle("US")
```

```
g1
```




```
grid.arrange(g2,g3, ncol=2)
```



c) En les variables qualitatives s'observa que les freqüències de les classes són bastant desbalancejades.

3 Estadística inferencial

3.1 Interval de confiança de la variable `Price`

Calculeu l'interval de confiança al 95% de la variable `Price`. A partir del valor obtingut, expliqueu com s'interpreta el resultat de l'interval de confiança.

Nota: S'han de realitzar els càlculs manualment. No es poden fer servir funcions de `R` que calculin directament l'interval de confiança com `t.test` o similar. Si que podeu fer servir funcions com `qnorm`, `pnorm`, `qt` i `pt`

```
alpha<-0.05
n<-length(mydata$Price)
n
```

```
## [1] 400
```

```
#Aplicem el teorema del límit central ja que la mida mostral és superior a 30
mean <- mean(mydata$Price)
s <- sd(mydata$Price)
t <- qnorm(alpha/2, lower.tail= FALSE) # normal
#t<-qt(alpha/2,lower.tail=FALSE,df=n-1)  # t-student
li <- mean - t*s/ sqrt(n)
ui <- mean + t*s/ sqrt(n)
#Comprovació
cat( "(" , li , ",", ui , ")")
```

```
## ( 113.4747 , 118.1153 )
```

```
tt<-t.test( mydata$Price, conf.level=0.95 )
tt
```

```
##
##  One Sample t-test
##
## data:  mydata$Price
## t = 97.814, df = 399, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  113.4677 118.1223
## sample estimates:
## mean of x
##   115.795
```

```
tt$conf.int
```

```
## [1] 113.4677 118.1223
## attr(,"conf.level")
## [1] 0.95
```

Resulta l'interval de confiança (113.47 , 118.12). Sent la mitjana mostral 115.8 anys. La interpretació que en fem és que si repetissim un nombre elevat de mostres el mateix procediment resultaria que aproximadament el 95% dels intervals trobats contindrien el veritable valor de la mitjana del preu.

3.2 Test de comparació de dues mitjanes

Es pot acceptar que en les botigues d'EUA (variable `US`) la mitjana de vendes del seients de cotxes infantils (variable `Sales`) és superior a la mitjana de vendes en botigues fora d'EUA? Calculeu per a un nivell de confiança del 95%.

Nota: S'han de realitzar els càlculs manualment. No es poden fer servir funcions de `R` que calculin directament l'interval de confiança com `t.test` o similar. Si que podeu fer servir funcions com `qnorm`, `pnorm`, `qt` i `pt`

S'assumirà que la variable `Sales` té distribució normal.

Seguiu els passos que es detallen a continuació:

3.2.1 Escriure la hipòtesi nul · la i alternativa

Es tracta d'una comparació de mitjanes en poblacions normals independents. Formulem les hipòtesis: $H_0 : \mu_1 = \mu_2$ i $H_1 : \mu_1 > \mu_2$, on μ_1 denota la mitjana de les botigues d'EUA i μ_2 de les botigues fora d'EUA.

3.2.2 Justificar quin mètode aplicareu

Aplicarem un test de t de Student amb la variant de Welch. Com es pot veure en el test d'homogeneitat de variàncies, no assumirem igualtat de variàncies.

```
df_1<-mydata$Sales[mydata$US=="Yes"]
df_2<-mydata$Sales[mydata$US=="No"]
var.test(df_1,df_2)
```

```
##
##  F test to compare two variances
##
## data:  df_1 and df_2
## F = 1.667, num df = 257, denom df = 141, p-value = 0.0008679
```

```
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.237737 2.217092
## sample estimates:
## ratio of variances
##           1.666969
```

3.2.3 Realitzar els càlculs de l'estadístic de contrast, valor crític i p valor amb un nivell de confiança del 95%

```
n1<-length(df_1)
n2<-length(df_2)


m1<-mean(df_1)
m2<-mean(df_2)


s1<-sd(df_1)
s2<-sd(df_2)

dfree<-(s1^2/n1+s2^2/n2)^2/(s1^4/(n1^2*(n1-1))+s2^4/(n2^2*(n2-1)))


s<- sqrt( s1^2/n1 + s2^2/n2 )
t <- (m1-m2) / s
t
```

```
## [1] 4.970486
```

```
pvalue <- pt(t,df=dfree,lower.tail=FALSE)  #una cua
pvalue
```

```
## [1] 5.207902e-07
```

```
#con nivel de confianza del 95%
t.crit95 <- qt( 0.05, df=dfree, lower.tail=FALSE)
t.crit95
```

```
## [1] 1.649162
```

El p-valor del test és inferior al nivell de significació, per tant, rebutjarem la hipòtesis nul · la i acceptarem que la mitjana de vendes del seients de cotxes infantils és major a les botigues d'EUA.

Per comprovar-ho podem usar la funció R.

```
#H1: mu_2 < mu_1  mu_1 es el valor No
t.test(Sales~US, data=mydata,m=0,var.equal=F,alt="less")
```

```
##
##  Welch Two Sample t-test
##
## data:  Sales by US
## t = -4.9705, df = 354.64, p-value = 5.208e-07
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -0.860059
## sample estimates:
```

```
##  mean in group No mean in group Yes
##          6.579789          7.866899
```

3.3 Contrast no paramètric

En l'apartat 2 hem assumit la normalitat de la variable vendes (`Sales`). Ara apliqueu un test no paramètric per respondre la pregunta anterior. Podeu usar una funció `R` per resoldre el contrast.

```
# u de mann-withney
wilcox.test(Sales~US,m=0,alt="less", data=mydata)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Sales by US
## W = 13552, p-value = 8.257e-06
## alternative hypothesis: true location shift is less than 0
```

3.3.1 Interpretar el resultat

Tenint en compte que el p-valor és inferior al 5%, acceptem la hipòtesis alternativa i podem afirmar que la diferència entre les vendes de seients de cotxes infantils entre botigues d'EUA i botigues fora d'EUA és superior.

4 Regressió

4.1 Model de regressió

a) Apliqueu un model de regressió lineal múltiple que tingui com a variables explicatives: `Price`, `Advertising`, `Age`, `Population`, `ShelveLoc`, `US`, i `Urban`, i com a variable dependent les vendes, variable `Sales`.

Especifiqueu el nivell base (usant la funció `relevel`): per a la variable `ShelveLoc`, la categoria "Bad", per a la variable `US`, la categoria "Yes", i per a la variable `Urban`, la categoria "Yes".

```
ShelveLocR <- relevel(mydata$ShelveLoc, ref='Bad' )
USR <- relevel(mydata$US, ref="Yes")
UrbanR <- relevel(mydata$Urban, ref="Yes")

mydata$ShelveLocR  <- ShelveLocR
mydata$USR <- USR
mydata$UrbanR <- UrbanR

mod <- lm(Sales~Price+ Advertising + Population +
            Age + ShelveLocR + UrbanR + USR, mydata )

summary(mod)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Advertising + Population + Age +
##     ShelveLocR + UrbanR + USR, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7157 -1.0064  0.0337  1.0892  4.5635
```

```
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      14.0896135  0.5732877  24.577  < 2e-16 ***
## Price            -0.0578467  0.0034628 -16.705  < 2e-16 ***
## Advertising       0.1138958  0.0176550   6.451 3.28e-10 ***
## Population       -0.0003501  0.0005830  -0.600    0.549
## Age              -0.0469978  0.0050552  -9.297  < 2e-16 ***
## ShelveLocRGood    4.3917379  0.2432240  18.056  < 2e-16 ***
## ShelveLocRMedium  1.9427228  0.2001982   9.704  < 2e-16 ***
## UrbanRNo         -0.1950090  0.1793677  -1.087    0.278
## USRNo            -0.2185299  0.2369814  -0.922    0.357
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.622 on 391 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6477
## F-statistic: 92.71 on 8 and 391 DF,  p-value: < 2.2e-16
```

b) Interpreteu el model ajustat. Expliqueu quina interpretació en feu de la contribució en el model de les variables regressores. Indiqueu com seria el model de regressió per una botiga fora d'EUA, no urbana i amb un `ShelveLoc` de tipus "Bad".

```
summary(mod)
```

```
##
## Call:
## lm(formula = Sales ~ Price + Advertising + Population + Age +
##     ShelveLocR + UrbanR + USR, data = mydata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7157 -1.0064  0.0337  1.0892  4.5635
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      14.0896135  0.5732877  24.577  < 2e-16 ***
## Price            -0.0578467  0.0034628 -16.705  < 2e-16 ***
## Advertising       0.1138958  0.0176550   6.451 3.28e-10 ***
## Population       -0.0003501  0.0005830  -0.600    0.549
## Age              -0.0469978  0.0050552  -9.297  < 2e-16 ***
## ShelveLocRGood    4.3917379  0.2432240  18.056  < 2e-16 ***
## ShelveLocRMedium  1.9427228  0.2001982   9.704  < 2e-16 ***
## UrbanRNo         -0.1950090  0.1793677  -1.087    0.278
## USRNo            -0.2185299  0.2369814  -0.922    0.357
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.622 on 391 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6477
## F-statistic: 92.71 on 8 and 391 DF,  p-value: < 2.2e-16
```

El model és significatiu, pvalor `< 2.2e-16` i el valor de R^2 ajustat és alt (0.65), la qual cosa indica capacitat predictiva del model per a estimar les vendes. S'observa que les variables `Price`, `Advertising`, `Age` i `ShelveLoc` són significatives.

Veiem que les variables `Price` i `Age` tenen un efecte de disminució de les vendes. Contràriament, la variable `Advertising` i `ShelveLoc` augmenta les vendes.

El model de regressió per una botiga fora d'EUA, no urbana i amb un `ShelveLoc` de tipus "Bad" és:

$$Sales = 14.0896 - 0.0578 * Price + 0.1139 * Advertising - 4 \times 10^{-4} * Population - 0.047 * Age - 0.195 - 0.2185$$

4.2 Predicció

Apliqueu el model de regressió per predir `Sales` d'una botiga fora d'EUA a una zona rural, amb `Price` de 131 dòlars, `Advertising` de 0 dòlars, `Population` de 139 milers de persones, `Age` de 40 anys i `ShelveLoc` de tipus ¨Bad¨.

Compareu el resultat amb el d'una botiga fora d'EUA a una zona rural, amb `Price` de 131 dòlars, `Advertising` de 9 mil de dòlars, `Population` de 139 milers de persones, `Age` de 40 anys i `ShelveLoc` de tipus ¨Good¨.

Expliqueu les diferències en funció dels coeficients del model de regressió.

```
new<-data.frame(Price=131,Advertising=0,Population=139,
                Age=40, ShelveLocR = 'Bad', USR='No', UrbanR='No')

predict(mod,new,type="response")
```

```
##        1
## 4.169583
```

```
new<-data.frame(Price=131,Advertising=9,Population=139,
                Age=40, ShelveLocR = 'Good', USR='No', UrbanR='No')

predict(mod,new,type="response")
```

```
##        1
## 9.586383
```

La diferència entre les variables explicatives és que a la segona botiga hi ha 9 mil dòlars de despeses d'`Advertesing` i s'exposa el producte a una localització "Good". Aquest canvi dona una predició de vendes alguna cosa més del doble que la botiga sense despeses d'`Advertising` i exposició del producte en una localització "Bad".

5 Anàlisi de la variància (ANOVA)

5.1 Anova d'un factor

5.1.1 Vendes i qualitat de la localització dintre de l'expositor

Realitzeu un ANOVA per contrastar la significació de la variable `ShelveLoc` en la variable `Sales`.

5.1.1.1 Hipòtesis nul · la i alternativa El factor `ShelveLoc` té 3 nivells, on "Bad" és el nivell més baix i "Good" el nivell més alt. Així les hipòtesis són:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad , \quad H_1 : \mu_i \neq \mu_j \quad \text{para algun } i, j$$

5.1.1.2 Model Calculeu l'anàlisi de variància, utilitzant la funció `aov` o `lm`. Interpreteu el resultat de l'anàlisi, tenint en compte els valors Sum Sq, Mean SQ, F i Pr (> F).

```r
#AOV
myaov <- aov(Sales ~ ShelveLoc, mydata)
kk <- summary( myaov )
kk
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## ShelveLoc     2  832.8   416.4   77.02 <2e-16 ***
## Residuals   397 2146.5     5.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El factor estudiat és significatiu

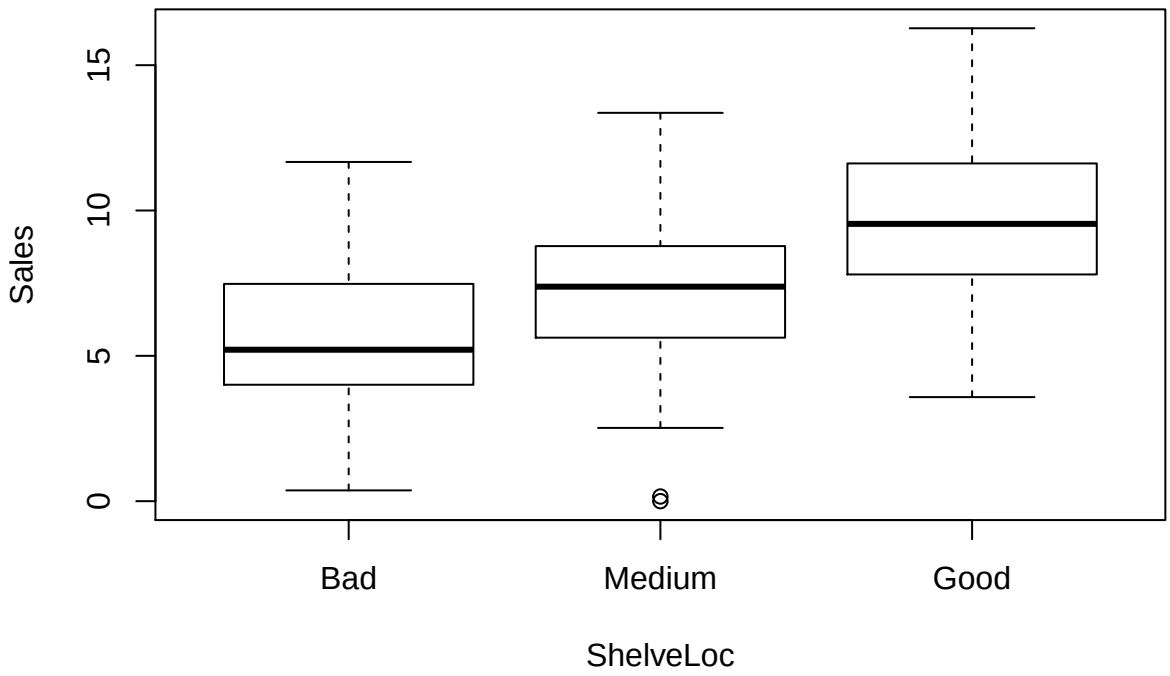
5.1.1.3 Càlculs

a) Mostreu gràficament la distribució de vendes, `Sales`, segons el factor `ShelveLoc` ordenat segons la qualitat: "Bad", "Medium" i "Good". Pots fer servir la funció `reorder`.

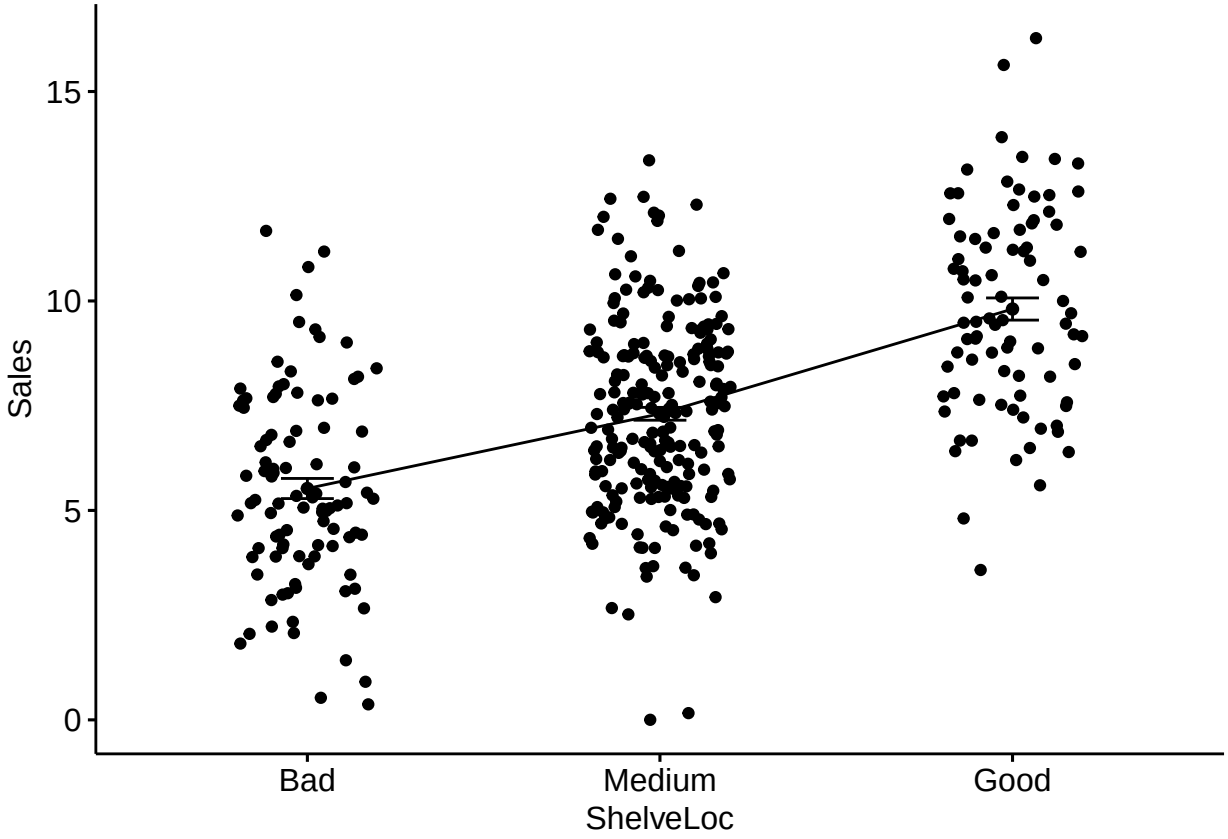
b) Per tal d'aprofundir en la comprensió del model ANOVA, calculeu manualment la suma de quadrats intra i la suma de quadrats entre grups. Els resultats han de coincidir amb el resultat del model ANOVA. Com a referència, pots obtenir les fórmules de López-Roldán i Fachelli (2015), pàgines 29-33.

c) També proporcioneu l'estimació dels efectes dels nivells del factor `ShelveLoc`. I l'estimació de la variància de l'error.

```r
# a) Plotting
mydata$ShelveLoc <- reorder(mydata$ShelveLoc, mydata$Sales, median)
boxplot( Sales ~ ShelveLoc, mydata)
```



```
ggline(mydata , x = "ShelveLoc", y = "Sales",
add = c("mean_se", "jitter"),
order = c("Bad","Medium","Good"),
ylab = "Sales", xlab = "ShelveLoc", font.label = list(size=6,face="plain"))
```




```
# b) Càlcul manual
E1<-mydata$Sales[mydata$ShelveLoc=="Bad"]
E2<-mydata$Sales[mydata$ShelveLoc=="Medium"]
E3<-mydata$Sales[mydata$ShelveLoc=="Good"]

n1<-length(E1)
n2<-length(E2)
n3<-length(E3)

mean.E1 <- mean(E1)
mean.E2 <- mean(E2)
mean.E3 <- mean(E3)

global.mean <- mean(mydata$Sales)

k<-3
SumOfSq <- function( x, mean ){
  sum <- sum( (x-mean)^2 )
  return (sum)
}
```

```r
within <- SumOfSq( E1, mean.E1 ) + SumOfSq( E2, mean.E2) + SumOfSq( E3, mean.E3 )


within
```

```
## [1] 2146.483
```

```r
between<- SumOfSq( mean.E1, global.mean)*n1 +
SumOfSq( mean.E2, global.mean)*n2 +
SumOfSq( mean.E3, global.mean)*n3

between
```

```
## [1] 832.8471
```

Comprovació amb la funció `lm()` o `aov()`.

```r
# Comprovació
mod.aov<-aov(Sales~ShelveLoc,data=mydata)
anova(mod.aov)
```

```
## Analysis of Variance Table
##
## Response: Sales
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## ShelveLoc   2  832.85  416.42  77.019 < 2.2e-16 ***
## Residuals 397 2146.48    5.41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interpretació dels resultats:

Mean Sq relativo a `ShelveLoc` és la variabilitat entre grups (la part explicada pel model) i Mean Sq `Residuals` és la variabilitat dintre de grups (la part no explicada pel model). El p-valor es inferior al 5% cosa que implica que rebutjem la hipòtesis nul · la d'igualtat de mitjanes. Direm que el factor `ShelveLoc` és significatiu.

Estimació dels efectes

```r
# c) Estimació del efectes
model.tables(mod.aov,type="mean")
```

```
## Tables of means
## Grand mean
##
## 7.409975
##
##  ShelveLoc
##        Bad  Medium   Good
##      5.523   7.307  9.808
## rep 96.000 219.000 85.000
```

```r
model.tables(mod.aov,type="effects")
```

```
## Tables of effects
##
##  ShelveLoc
##        Bad   Medium   Good
##     -1.887  -0.1034  2.398
## rep 96.000 219.0000 85.000
```

L'estimació de la variància de l'error és 5.41.

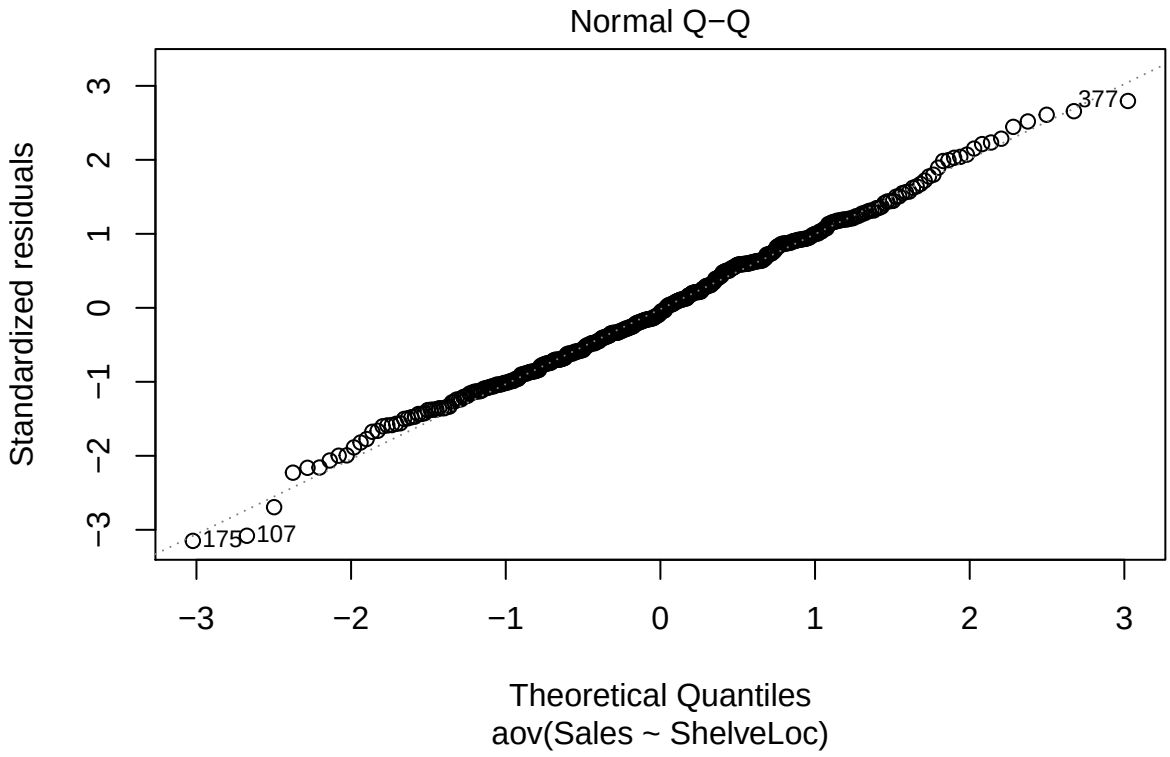
5.2 Adequació del model

Mostra visualment l'adequació del model ANOVA. Podeu fer servir plot sobre el model ANOVA resultant. En els apartats següents es demana la interpretació d'aquests gràfics.

5.2.1 Normalitat dels residus

Interpreteu la normalitat dels residus a partir de l'gràfic Normal Q-Q que es mostra en l'apartat anterior.

```
plot(mod.aov,which=2)
```

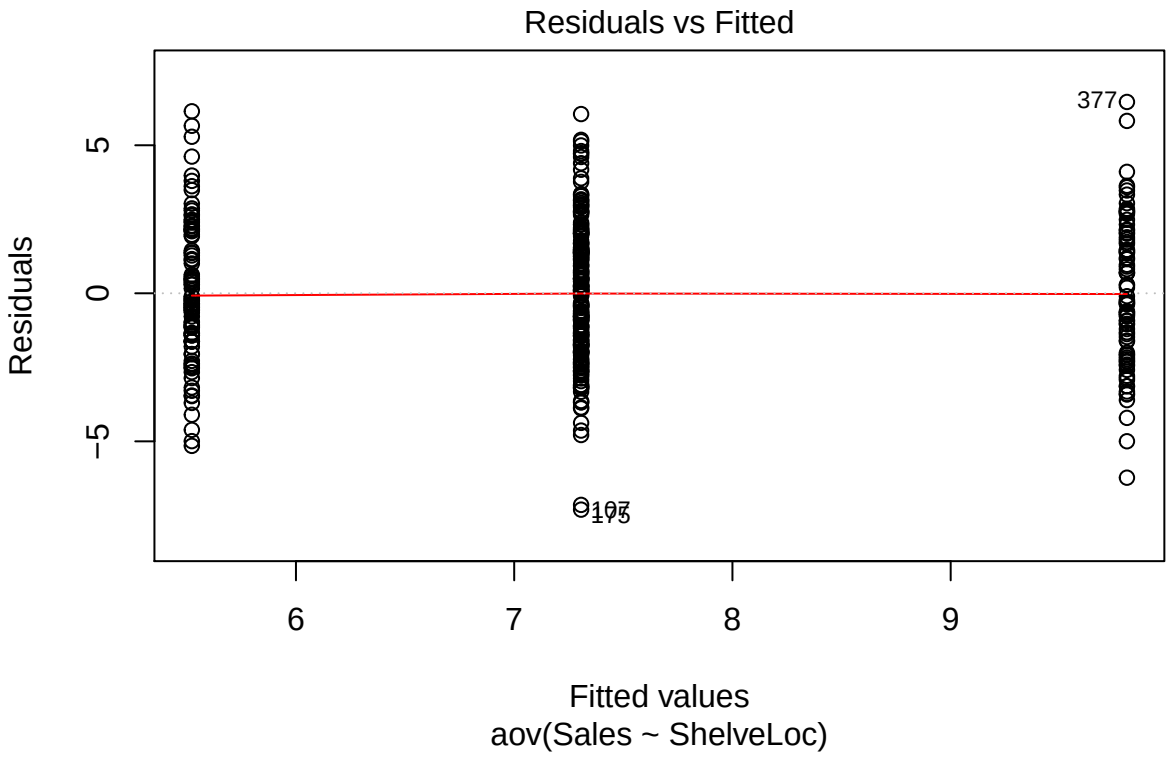


La majoria de residus es disposen d'acord als quantils teorics. Com és habitual hi ha un cert desajust als extrems de la distribució.

5.2.2 Homocedasticitat dels residus

Els gràfics "Residuals vs Fitted", "Scale-Location" i "Residuals vs Factor levels" donen informació sobre la homocedasticitat dels residus. Interpreteu aquests gràfics.

```
#par(mfrow=c(2,2))
plot(mod.aov,which=1)
```




```
#par(mfrow=c(1,1))
```

A nivell visual, sembla que la suposició d'homocedasticitat es manté. Per a totes 3 condicions experimentals la dispersió dels residus és bastant similar.

5.3 ANOVA no paramètric

Si la validació de les premisses de normalitat i homocedasticitat no es verifiquen es pot aplicar un test no paramètric, per exemple el test de Kruskal-Wallis.

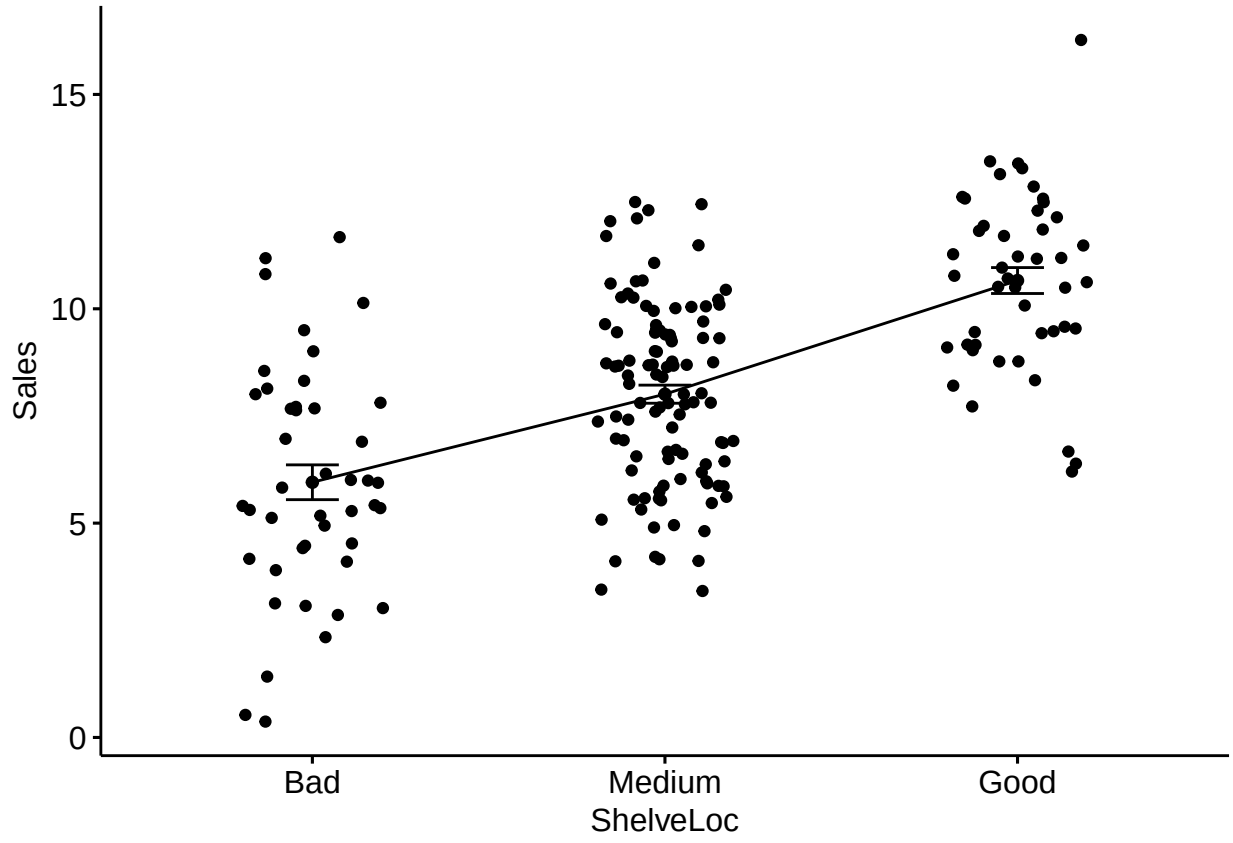
Apliqueu un test de Kruskal-Wallis per contrastar si hi ha diferències entre les botigues segon on s'exposa (`ShelveLoc`) pel que fa a les vendes (`Sales`) només a les observacions amb `Advertising` major que el valor de la mediana.

```
subsetmydata <- mydata[mydata$Advertising > median(mydata$Advertising),]
summary(subsetmydata)
```

```
##      Sales          CompPrice       Income        Advertising
##  Min.   : 0.370   Min.   : 85.0   Min.   : 21.00   Min.   : 6.00
##  1st Qu.: 6.015   1st Qu.:114.2   1st Qu.: 45.00   1st Qu.:10.00
##  Median : 8.425   Median :124.5   Median : 69.00   Median :12.00
##  Mean   : 8.170   Mean   :124.2   Mean   : 69.59   Mean   :12.59
##  3rd Qu.:10.095   3rd Qu.:134.0   3rd Qu.: 90.00   3rd Qu.:15.00
##  Max.   :16.270   Max.   :175.0   Max.   :120.00   Max.   :29.00
##    Population        Price        ShelveLoc        Age          Education
##  Min.   : 16.0   Min.   : 55.0   Bad   : 44   Min.   :25.00   Min.   :10.00
##  1st Qu.:183.2   1st Qu.:101.2   Medium:104   1st Qu.:41.00   1st Qu.:11.00
##  Median :312.5   Median :116.5   Good  : 46   Median :53.00   Median :14.00
```

```
##  Mean   :294.2   Mean   :116.4                Mean   :53.27   Mean   :13.79
##  3rd Qu.:415.5   3rd Qu.:131.0                3rd Qu.:66.75   3rd Qu.:16.00
##  Max.   :509.0   Max.   :191.0                Max.   :80.00   Max.   :18.00
##  Urban       US       ShelveLocR   USR      UrbanR
##  No : 57   No :  4   Bad   : 44   Yes:190   Yes:137
##  Yes:137   Yes:190   Good  : 46   No :  4   No : 57
##                      Medium:104
##
##
##
```

```
ggline(subsetmydata , x = "ShelveLoc", y = "Sales",
add = c("mean_se", "jitter"),
order = c("Bad","Medium","Good" ),
ylab = "Sales", xlab = "ShelveLoc", font.label = list(size=6,face="plain"))
```




```
krus<-kruskal.test(Sales~ShelveLoc,data=subsetmydata)
krus
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Sales by ShelveLoc
## Kruskal-Wallis chi-squared = 64.571, df = 2, p-value = 9.52e-15
```

5.3.1 Interpretació dels resultats:

Tenint en compte el p-valor, descartem la igualtat de mitjanes de `Sales` segons la qualitat del lloc d'exposició al subgrup d'observacions amb `Advertising` major que el valor de la mediana.

6 ANOVA multifactorial

A continuació, es vol avaluar l'efecte de més d'un factor sobre la variable `Sales` on el primer factor sempre serà `ShelveLoc`. Primer es realitzarà l'anàlisi on el segon factor és `US` i després, l'anàlisi on el segon factor és `Urban`.

6.1 Factors: `ShelveLoc` i `US`

6.1.1 Anàlisis visual dels efectes principals i possibles interaccions

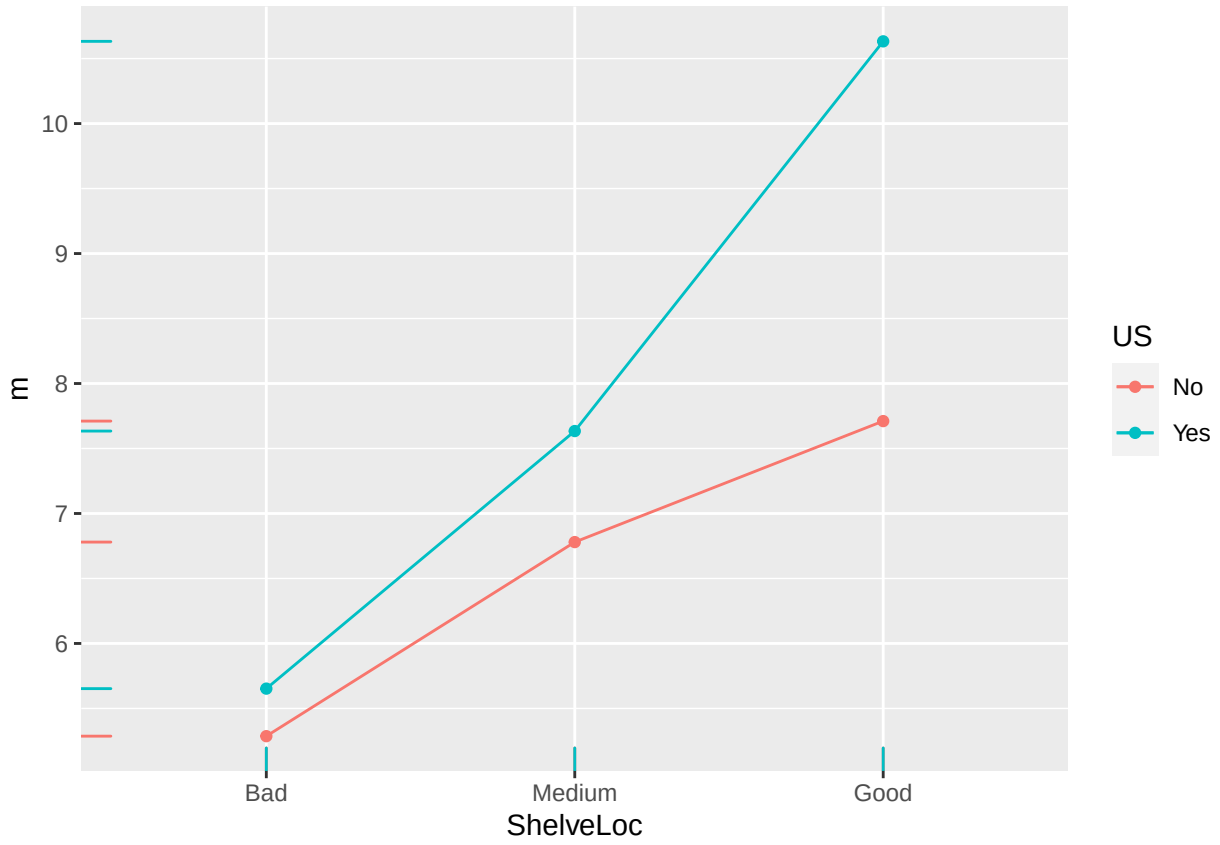
Dibuixeu en un gràfic la variable `Sales` en funció de `ShelveLoc` i en funció de `US`. El gràfic ha de permetre avaluar si hi ha interacció entre els dos factors. Per això, es recomana seguir els passos:

1. Agrupeu el conjunt de dades per `ShelveLoc` i per `US`. Calcular la mitjana de vendes per a cada grup. Per realitzar aquest procés, es pot fer amb les funcions `group_by` i `summarise` de la llibreria `dplyr`.
2. Mostreu el conjunt de dades en forma de taula, on es mostri la mitjana de cada grup segons `ShelveLoc` i `US`.
3. Mostreu en un gràfic el valor mitjà de la variable `Sales` per a cada factor. Pots inspirar-te en els gràfics de López-Roldán i Fachelli (2015), p.38. Pots realitzar aquest tipus de gràfic usant la funció `ggplot` de la llibreria `ggplot2`.
4. Interpreteu el resultat sobre si només hi ha efectes principals o hi ha interacció entre els factors. Si hi ha interacció, explicar com s'observa aquesta interacció en el gràfic.

```
mydata %>% group_by(ShelveLoc, US) -> DS2
DS3 <- summarise( DS2, m=mean(Sales), sd=sd(Sales), n=length(Sales))
DS3
```

```
## # A tibble: 6 x 5
## # Groups:   ShelveLoc [3]
##   ShelveLoc US        m    sd     n
##   <fct>     <fct> <dbl> <dbl> <int>
## 1 Bad       No     5.29  2.02    34
## 2 Bad       Yes    5.65  2.53    62
## 3 Medium    No     6.78  2.23    84
## 4 Medium    Yes    7.63  2.24   135
## 5 Good      No     7.71  1.65    24
## 6 Good      Yes   10.6   2.20    61
```

```
ggplot(DS3, aes(x=ShelveLoc, y=m, group=US, color=US)) +
geom_point() + geom_line() + geom_rug()
```



Sembla que hi ha interacció entre els factors `ShelveLoc` i `US` pel que fa a la variable `Sales`. Sembla que no hi ha paral · lelisme entre els segments del nivells Medium i Good per botigues d'EUA i fora d'EUA.

Caldrà corroborar si la interacció observada gràficament es confirma.

```
mod2<-lm(Sales~ShelveLoc*US,data=mydata)
anova(mod2)
```

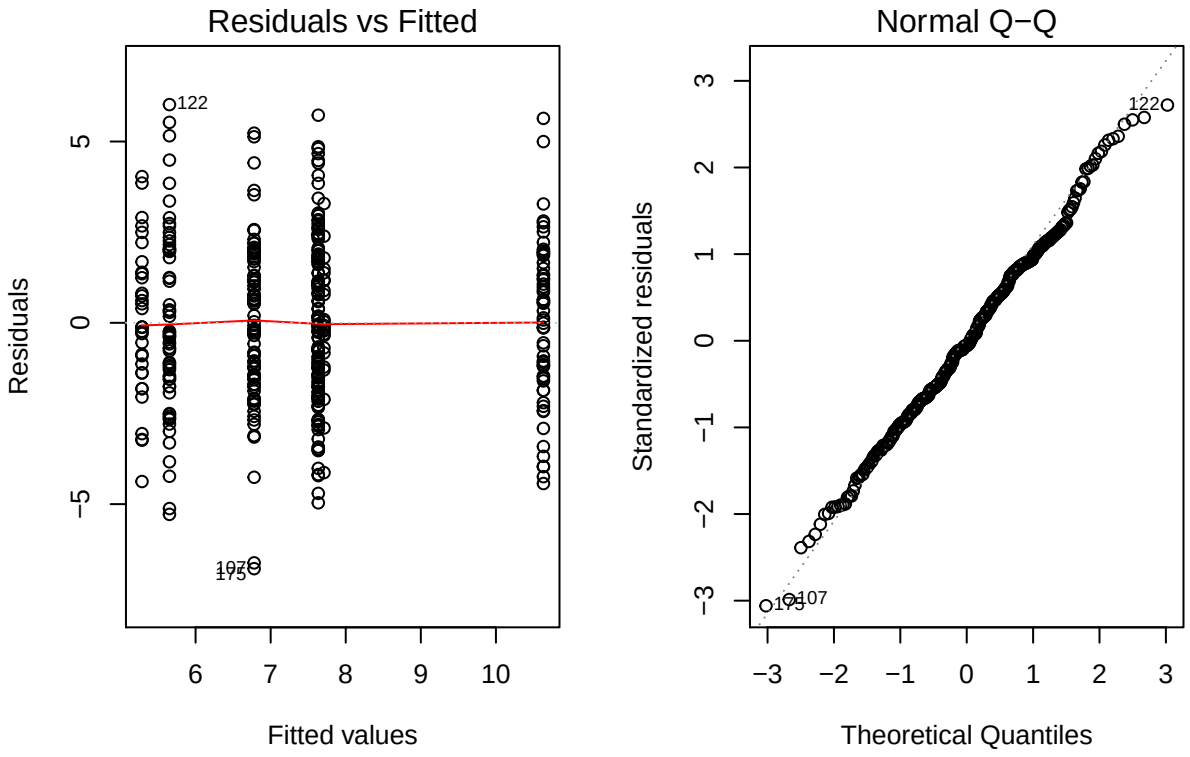
```
## Analysis of Variance Table
##
## Response: Sales
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## ShelveLoc      2  832.85  416.42 83.7640 < 2.2e-16 ***
## US             1  115.69  115.69 23.2716 2.014e-06 ***
## ShelveLoc:US   2   72.06   36.03  7.2477 0.0008107 ***
## Residuals    394 1958.73    4.97
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tant els factors principals com la interacció entre els factors són significatius. Per tant, les vendes en funció de ShelveLoc, són diferents segons la botiga estigui ubicada als EUA o no.

6.1.2 Adequació del model

Interpreteu l'adequació del model ANOVA obtingut usant els gràfics de residus.

```
par(mfrow=c(1,2),cex=0.8)
plot(mod2,which=1)
plot(mod2,which=2)
```



Visualment podem admetre que hi ha tendència normal en la distribució dels residus i també sembla observar-se homocedasticitat.

6.2 Factors: `ShelveLoc` i `Urban`

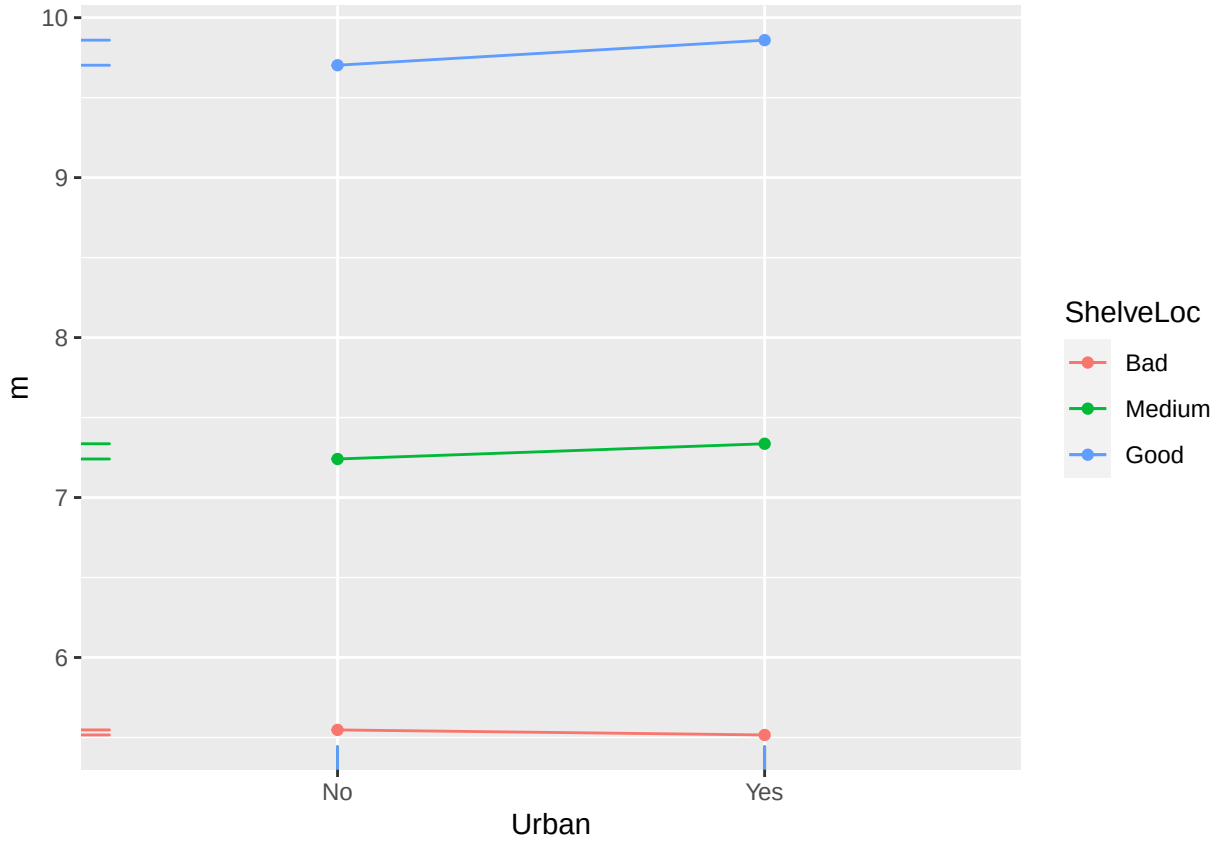
6.2.1 Anàlisis visual dels efectes principals i possibles interaccions

Dibuixeu en un gràfic la variable `Sales` en funció de `ShelveLoc` i en funció de `Urban`. El gràfic ha de permetre avaluar si hi ha interacció entre els dos factors. Per això, es recomana seguir els passos:

1. Agrupeu el conjunt de dades per `ShelveLoc` i per `Urban`. Calcular la mitjana de vendes per a cada grup. Per realitzar aquest procés, es pot fer amb les funcions `group_by` i `summarise` de la llibreria `dplyr`.
2. Mostreu el conjunt de dades en forma de taula, on es mostri la mitjana de cada grup segons `ShelveLoc` i `Urban`.
3. Mostreu en un gràfic el valor mitjà de la variable `Sales` per a cada factor. Pots inspirar-te en els gràfics de López-Roldán i Fachelli (2015), p.38. Pots realitzar aquest tipus de gràfic usant la funció `ggplot` de la llibreria `ggplot2`.
4. Interpreteu el resultat sobre si només hi ha efectes principals o hi ha interacció entre els factors. Si hi ha interacció, explicar com s'observa aquesta interacció en el gràfic.

```
## # A tibble: 6 x 5
## # Groups:   ShelveLoc [3]
##   ShelveLoc Urban     m    sd     n
##   <fct>     <fct> <dbl> <dbl> <int>
## 1 Bad       No     5.55  2.36    22
## 2 Bad       Yes    5.52  2.37    74
```

```
## 3 Medium    No     7.24  2.50    68
## 4 Medium    Yes    7.34  2.16   151
## 5 Good      No     9.70  2.01    28
## 6 Good      Yes    9.86  2.64    57
```



S'aprecia un molt petit efecte d'interacció pel nivell Good del factor `ShelveLoc`. De totes maneres cal fer la comprovació de la significació.

```
## Analysis of Variance Table
##
## Response: Sales
##                  Df  Sum Sq Mean Sq F value Pr(>F)
## ShelveLoc         2  832.85  416.42 76.4693 <2e-16 ***
## Urban             1    0.57    0.57  0.1051  0.746
## ShelveLoc:Urban   2    0.33    0.17  0.0305  0.970
## Residuals       394 2145.58    5.45
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

L'unic factor significatiu és `ShelveLoc`. El factor `Urban` i la interacció no són significatius.

7 Comparacions múltiples

Prenent com a referència el model ANOVA multifactorial, amb els factors `ShelveLoc` i `US`, aplicar el test de comparació múltiple Scheffé. Interpreteu el resultat del test i indicar quins grups són diferents significativament entre si.

```r
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 3.6.3
```

```
##
## Attaching package: 'DescTools'
```

```
## The following objects are masked from 'package:psych':
##
##     AUC, ICC, SD
```

```
## The following object is masked from 'package:data.table':
##
##     %like%
```

```r
mod2<-aov(Sales~ShelveLoc*US,data=mydata) # cal usar aov() enlloc de lm()
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: Sales
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## ShelveLoc      2  832.85  416.42 83.7640 < 2.2e-16 ***
## US             1  115.69  115.69 23.2716 2.014e-06 ***
## ShelveLoc:US   2   72.06   36.03  7.2477 0.0008107 ***
## Residuals    394 1958.73    4.97
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
mod2_HocTest <- PostHocTest( mod2, method="scheffe")
mod2_HocTest
```

```
##
##   Posthoc multiple comparisons of means : Scheffe Test
##     95% family-wise confidence level
##
## $ShelveLoc
##                 diff    lwr.ci   upr.ci    pval
## Medium-Bad  1.783659 0.8709244 2.696393 1.1e-07 ***
## Good-Bad    4.284730 3.1741731 5.395288 < 2e-16 ***
## Good-Medium 2.501072 1.5481614 3.453982 7.6e-14 ***
##
## $US
##            diff    lwr.ci   upr.ci    pval
## Yes-No 1.120046 0.3408931 1.899199 0.00041 ***
##
## $`ShelveLoc:US`
##                             diff      lwr.ci    upr.ci    pval
## Medium:No-Bad:No      1.49323529 -0.02244876  3.008919  0.0566 .
## Good:No-Bad:No        2.42406863  0.43607176  4.412065  0.0059 **
## Bad:Yes-Bad:No        0.36565465 -1.22562714  1.956936  0.9884
## Medium:Yes-Bad:No     2.34745752  0.91664152  3.778274 2.2e-05 ***
## Good:Yes-Bad:No       5.34585824  3.74996364  6.941753 < 2e-16 ***
## Good:No-Medium:No     0.93083333 -0.79505778  2.656724  0.6611
## Bad:Yes-Medium:No    -1.12758065 -2.37607827  0.120917  0.1070
## Medium:Yes-Medium:No  0.85422222 -0.18202245  1.890467  0.1824
## Good:Yes-Medium:No    3.85262295  2.59825133  5.106995 < 2e-16 ***
```

```
## Bad:Yes-Good:No      -2.05841398 -3.85105999 -0.265768  0.0126 *
## Medium:Yes-Good:No   -0.07661111 -1.72846959  1.575247  1.0000
## Good:Yes-Good:No      2.92178962  1.12504769  4.718532 2.7e-05 ***
## Medium:Yes-Bad:Yes    1.98180287  0.83782621  3.125780 5.1e-06 ***
## Good:Yes-Bad:Yes      4.98020360  3.63546283  6.324944 < 2e-16 ***
## Good:Yes-Medium:Yes   2.99840073  1.84801625  4.148785 1.2e-13 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
str(mod2_HocTest$`ShelveLoc:US`)
```

```
##  num [1:15, 1:4] 1.493 2.424 0.366 2.347 5.346 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:15] "Medium:No-Bad:No" "Good:No-Bad:No" "Bad:Yes-Bad:No" "Medium:Yes-Bad:No" ...
##   ..$ : chr [1:4] "diff" "lwr.ci" "upr.ci" "pval"
```

Val a dir que el test de comparacions múltiples s'aplica quan s'observen diferències significatives entre els grups, a fi de determinar entre quins grups es troben aquestes diferències. Pel que fa al factor `ShelveLoc` es detecten diferències 2 a 2. Pel que fa al factor `US` si que hi ha diferències però de fet al tractar-se d'un factor de dos nivells no caldria fer la correcció. Pel que fa a les interaccions, veiem que hi ha algunes de significatives:

```
pval.sel <- mod2_HocTest$`ShelveLoc:US`[,4] < 0.05

kable(mod2_HocTest$`ShelveLoc:US`[pval.sel,],
digits=c(3,3,3,5), caption="Interaccions significatives")
```

Taula 2: Interaccions significatives

	diff	lwr.ci	upr.ci	pval
Good:No-Bad:No	2.424	0.436	4.412	0.00591
Medium:Yes-Bad:No	2.347	0.917	3.778	0.00002
Good:Yes-Bad:No	5.346	3.750	6.942	0.00000
Good:Yes-Medium:No	3.853	2.598	5.107	0.00000
Bad:Yes-Good:No	-2.058	-3.851	-0.266	0.01256
Good:Yes-Good:No	2.922	1.125	4.719	0.00003
Medium:Yes-Bad:Yes	1.982	0.838	3.126	0.00001
Good:Yes-Bad:Yes	4.980	3.635	6.325	0.00000
Good:Yes-Medium:Yes	2.998	1.848	4.149	0.00000

8 Conclusions

Podem concloure (amb una confiança del 95%, amb un nivell de significació del 5%) que:

- El conjunt de dades té 11 variables. Les variables `Sales` , `CompPrice` , `Income` , `Advertising` , `Population` , `Price` , `Age` i `Education` són variables quantitatives de tipus `int` o `num` , les altres són qualitatives de tipus factor. No hi ha dades faltants.
- L'interval de confiança de la variable `Price` està entre (113.4747 , 118.1153).
- La mitjana de vendes (`Sales`) a les botigues d'EUA és significativament major que a les botigues fora d'EUA.
- Les variables `Price`, `Advertising`, `Age` i `ShelveLoc` són variables significatives en el model de regressió lineal múltiple per explicar la variable `Sales`

- Els factors `ShelveLoc` i `US` i la interacció són significatius en `Sales`.

9 Comentaris importants sobre l'activitat

1. **No es pot inspeccionar ni corregir de manera manual** el fitxer de dades. Per exemple, **no** es poden realitzar instruccions d'aquest tipus:

   ```
   data[1,5] <- 32.5
   ```

   Aquest tipus de transformacions s'han de fer amb funcionalitats de cerca (buscar els registres que tenen errors o inconsistències) i després fer les correccions oportunes amb funcionalitats de R. Així el procediment de neteja és útil, independentment del fitxer de dades i de la posició i valors concrets de l'arxiu.

2. **No es poden fer llistats complets de les dades del fitxer a pantalla**, perquè generen arxius de sortida excessivament grans. Si es desitja validar el resultat d'una instrucció sobre les dades, es pot usar la funció **head** que mostra les primeres files de la taula de dades o **tail** que mostra les últimes.

Puntuació de l'activitat

- Apartat 2 (10%)
- Apartat 3 (10%)
- Apartat 4 (10%)
- Apartat 5 (20%)
- Apartat 6 (20%)
- Apartat 7 (10%)
- Apartat 8 (10%)
- Qualitat del informe dinàmic (10%)