PIE2 - Entregable 2

PauM PauF

2023-12-03

EXERCICI 1

Es vol predir la longitud del sèpal de plantes de IRIS en funció de l'amplada del sèpal i de l'espècie de IRIS. Llegiu les dades i contesteu les preguntes següents.

1) Especifiqueu quina és la variable resposta i quines son les explicatives així com el tipus de variable del que es tracta.

La variable resposta és la longitud del sèpal i les variables explicatives són l'amplada del sèpal (variable numèrica) i l'espècie d'IRIS (variable categòrica).

2) Quines són les preguntes que té sentit contestar?

Les preguntes que més ens interessaria contestar són:

- La longitud del sèpal canvia en funció de l'espècie?
- Quina relació hi ha entre la longitud del sèpal i la seva amplada? L'amplada influeix?
- Aquesta relació és igual per a totes les espècies?

3) Té sentit parlar d'interacció en aquest exercici?. Si és que sí, com l'interpretaríeu?

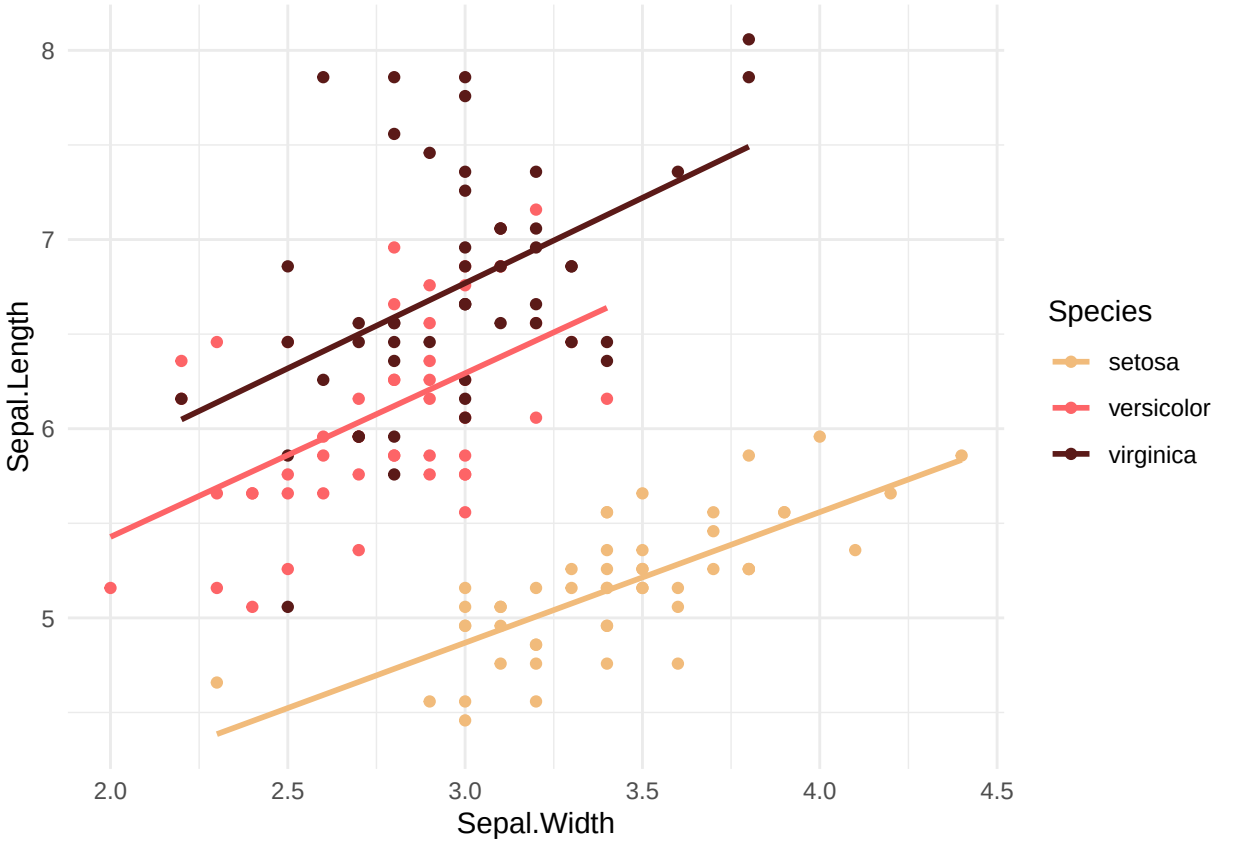
Sí, ja que podria ser que entre diferents espècies d'IRIS la relació entre l'amplada del sèpal i la longitud del sèpal fos diferent, i aleshores hauríem de tenir en compte l'interacció entre les dues variables explicatives per a que el nostre model expliqui millor les dades.

Si existís aquesta interacció, aleshores tindríem un coeficient d'interacció per a cada espècie (excepte per a l'espècie de referència), que s'interpretaria com a la diferència del que augmenta la longitud del sèpal per a cada unitat d'amplada entre una espècie amb l'espècie de referència.

4) Dibuixeu de forma exploratòria el vostre conjunt de dades i en base al gràfic o gràfics realitzats, traieu unes primeres conclusions, especifiqueu-ne tres (caldrà comprovar-les més endavant).

```
gplot <- ggplot(irisdat4) +
  aes(y = Sepal.Length, x = Sepal.Width, color = Species) +
  geom_point() +
  geom_smooth(se = FALSE, method = "lm") +
  theme_minimal() +
  scale_color_manual(values = wes_palette("GrandBudapest1", n = 3))

gplot
```



Clarament es veu que l'espècie és significativa. És a dir, que l'amplada del sèpal sí que depèen de l'espècie.

Si ens mirem cada espècie per separat, es veu una clara relació linial entre l'amplada i la llargada del sèpal, sobretot per a l'espècie setosa i versicolor. Per a l'espècie virginica també es veu linialitat però menys clara, ja que te més variabilitat. Si ens prenem totes les mostres com a una mateixa espècie aleshores es seguiria veient una mica de correlació positiva entre l'amplada i la llargada del sèpal però hi hauria molta variabilitat, i el model explicaria molt pitjor les dades. També s'observa que els tres pendents són molt similars, el que indica que la interacció entre les variables explicatives no és significativa.

I per últim, les dades de l'espècie setosa estan clarament separades de les altres dues però, en canvi, les espècies versicolor i virginica estan més juntes, i ja no queda tan clar si la diferència entre aquestes serà significativa o no.

5) Ajusteu un model de regressió lineal simple i contestes les següents preguntes:

```
model1 = lm(Sepal.Length ~ Sepal.Width, data=irisdat4)
summary(model1)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width, data = irisdat4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -1.5561 -0.6333 -0.1120  0.5579  2.2226
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.6844     0.4789   13.96   <2e-16 ***
## Sepal.Width  -0.2234     0.1551   -1.44    0.152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8251 on 148 degrees of freedom
## Multiple R-squared:  0.01382,    Adjusted R-squared:  0.007159
## F-statistic: 2.074 on 1 and 148 DF,  p-value: 0.1519
```

5.1) Quina és l'estimació de la variància que heu obtingut?
L'estimació de la variança és $0.8251^2 = 0.68078$.

5.2) Quina part de la variabilitat de la variable resposta és explicada pel model?
Com que $R^2 = 0.0138$, aleshores el model explica un 1.38% de la variabilitat de la variable resposta. És un percentatge molt petit, aquest model no ens serviria per res. Això ja quadra amb el que havíem dit a l'apartat 4).

5.3) Quin és el residu més gran que heu trobat?

```
max_res = irisdat4[ which.max(abs(residuals(model1))), ]
max_res = data.frame(max_res$X, max_res$Sepal.Length, max_res$Sepal.Width, max_res$Species)
names(max_res) = c("Mostra", "Sepal.Length", "Sepal.Width", "Species")
max_res
```

```
##   Mostra Sepal.Length Sepal.Width   Species
## 1    132     8.058206         3.8 virginica
```

El residu més gran és de 2.2226, que correspon a la mostra 132, amb valors Sepal length = 8.06 , Sepal width = 3.8 i espècie virginica.

5.4) Hi ha valors que tinguin un leverage superior al permès (feu servir la fita més acurada)? Quants?

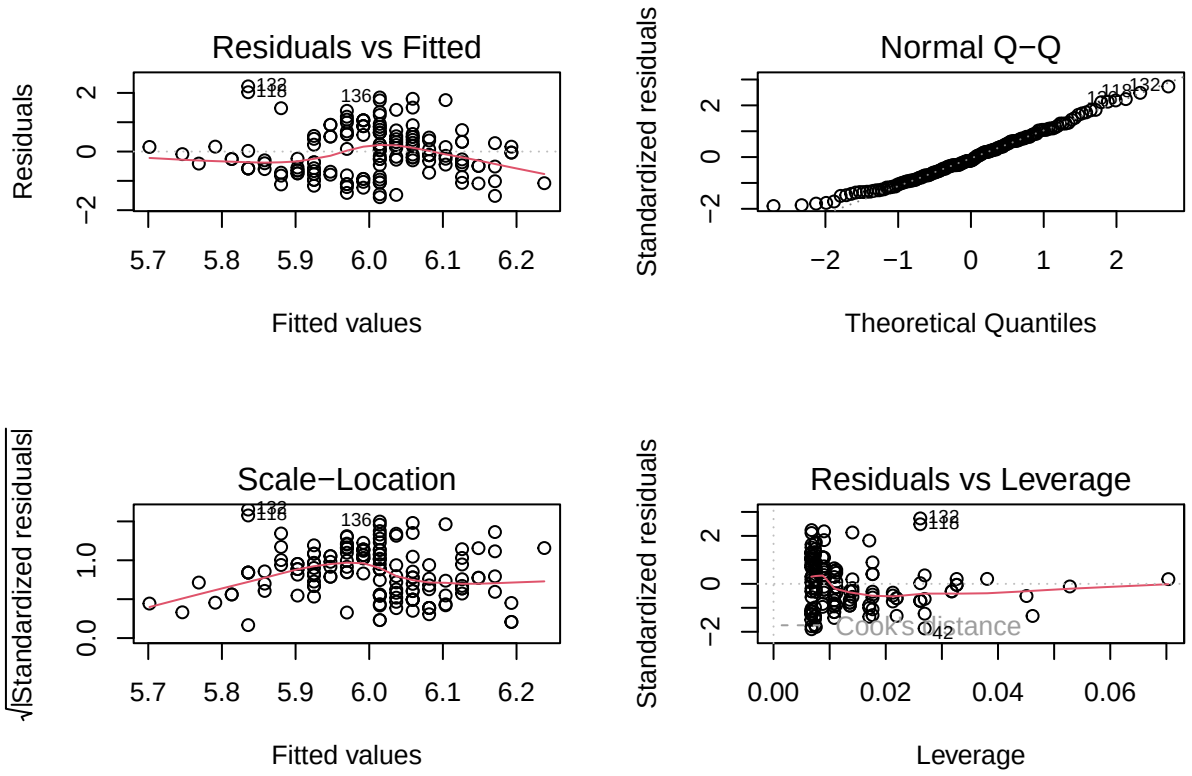
```
p = ncol(model.matrix(model1))
n = nrow(irisdat4)
cond_lev = 3*p/n
res = irisdat4[which(hatvalues(model1)>cond_lev),]
res = data.frame(res$X, res$Sepal.Length, res$Sepal.Width, res$Species)
names(res) = c("Mostra", "Sepal.Length", "Sepal.Width", "Species")
res
```

```
##   Mostra Sepal.Length Sepal.Width    Species
## 1     16     5.858206         4.4     setosa
## 2     33     5.358206         4.1     setosa
## 3     34     5.658206         4.2     setosa
## 4     61     5.158206         2.0 versicolor
```

Tenim 4 mostres que tenen un leverage superior al permès (superior a $3p/n$).

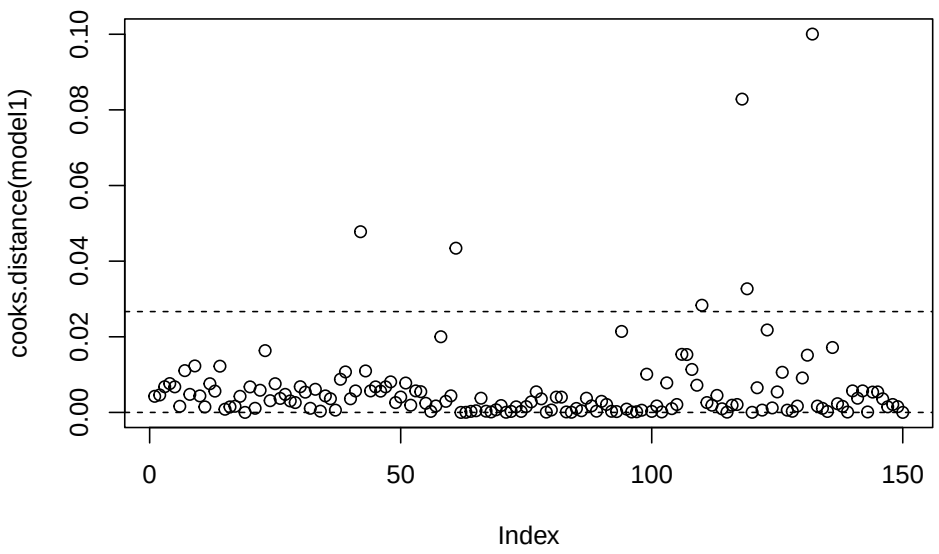
5.5) Feu el plot dels residus i comenteu tres coses de les que considereu importants.

```
par(mfrow = c(2,2))
plot(model1)
```



- Podem veure que la hipòtesis de normalitat de residus no s'acaba de complir, ja que a la gràfica de Residuals vs Fitted els errors no són del tot homogenis (son més grans al voltant del valor 6), i a l'extrem del qq-plot hi ha una mica de corbatura.
- La hipòtesis d'homoscedasticitat tampoc es compleix del tot, ja que al plot de Scale-Location veiem que la desviació augmenta fins al valor 6 i després torna a disminuïr una mica.
- El gràfic de Residuals vs Leverage està prou bé, però potser hi ha algunes mostres influents, com la 132 i 118. Per verure-ho millor les distàncies de Cook, podem fer un altre plot:

```
plot(cooks.distance(model1))
abline(h=c(0,4/n),lty=2)
```



Efectivament, tenim dades que sobrepassen de molt la distància de cook permesa. Ens les hauríem de mirar bé i si de cas, treure-les i tornar a ajustar el model.

6) Ajusteu un model que tingui com a explicatives les dues variables esmentades a l'enunciat i que no són la variable resposta (model 2).

```
model2 = lm(Sepal.Length ~ Sepal.Width + Species, data=irisdat4)
summary(model2)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width + Species, data = irisdat4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30711 -0.25713 -0.05325  0.19542  1.41253
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.4096     0.3698   6.517 1.09e-09 ***
## Sepal.Width         0.8036     0.1063   7.557 4.19e-12 ***
## Speciesversicolor   1.4587     0.1121  13.012  < 2e-16 ***
## Speciesvirginica    1.9468     0.1000  19.465  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.438 on 146 degrees of freedom
## Multiple R-squared:  0.7259, Adjusted R-squared:  0.7203
## F-statistic: 128.9 on 3 and 146 DF,  p-value: < 2.2e-16
```

6.1) Quants paràmetres té el vostre model? Tots els paràmetres son significatius?

El model 2 té 4 paràmetres i tots són significatius.

6.2) Com interpreteu els paràmetres obtinguts? (una frase per a cada paràmetre).

L'espècie de referència d'aquest model és la setosa. Aleshores, els paràmetres els interpretem com:

- Intercept: longitud del sèpal de l'espècie setosa (espècie de referència) quan l'amplada del sèpal és 0.
- Sepal.Width: la longitud que augmenta el sèpal quan augmentem en una unitat la seva amplada.
- Species versicolor: diferència de longitud del sèpal entre l'espècie versicolor i setosa (fixada una amplada de sèpal).
- Species virginica: diferència de longitud del sèpal entre l'espècie virginica i setosa

6.3) Quina part de la variabilitat de les dades explica el vostre model?

Ara el model explica molt millor les dades que el model d'abans; $R^2 = 0.7259$, i per tant el model explica el 72.6% de la variabilitat de les dades.

6.4) Quina és l'estimació de la variància residual que obteniu?

Obtenim una variança residual de $0.438^2 = 0.191844$.

6.5) Quina és la matriu del disseny (matriu X) associada al model que esteu ajustant?

La matriu del model la podem veure amb la següent comanda:

```
model.matrix(model2)
```

```
##     (Intercept) Sepal.Width Speciesversicolor Speciesvirginica
## 1             1         3.5                 0                0
## 2             1         3.0                 0                0
## 3             1         3.2                 0                0
## 4             1         3.1                 0                0
## 5             1         3.6                 0                0
## 6             1         3.9                 0                0
## 7             1         3.4                 0                0
## 8             1         3.4                 0                0
## 9             1         2.9                 0                0
## 10            1         3.1                 0                0
## 11            1         3.7                 0                0
## 12            1         3.4                 0                0
## 13            1         3.0                 0                0
## 14            1         3.0                 0                0
## 15            1         4.0                 0                0
## 16            1         4.4                 0                0
## 17            1         3.9                 0                0
## 18            1         3.5                 0                0
```

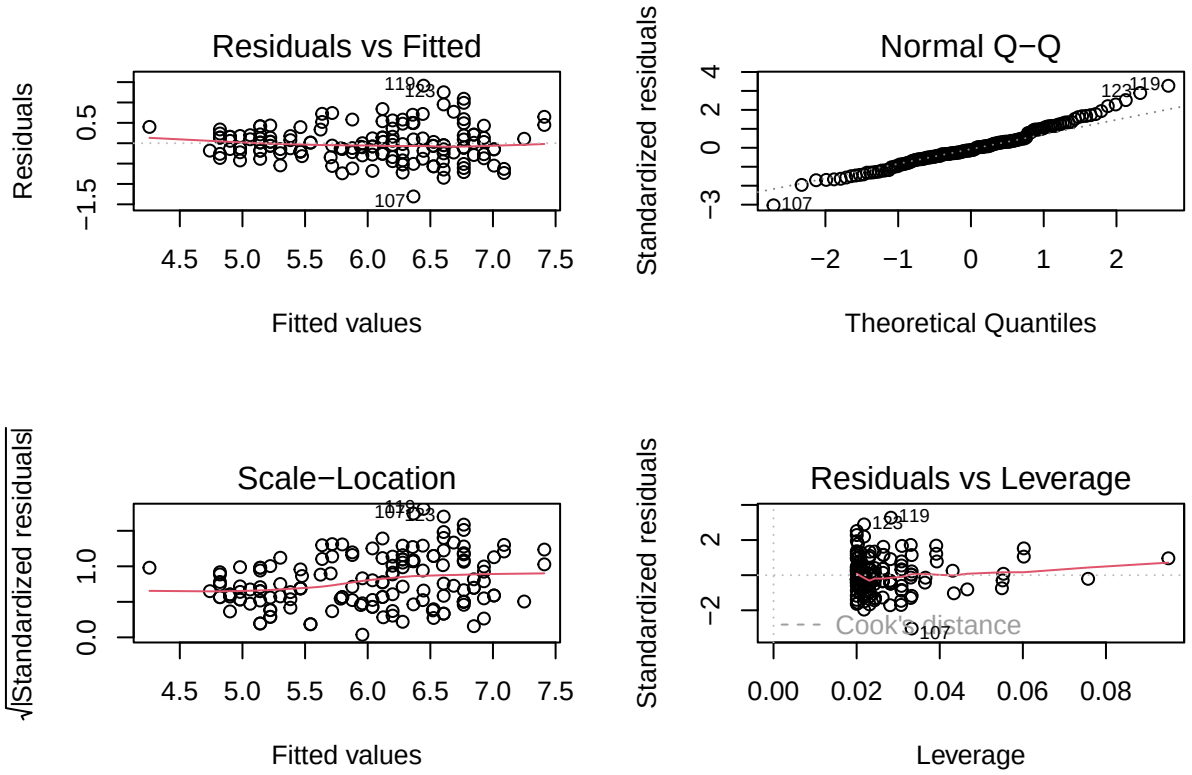
```
## 19          1         3.8                0              0
## 20          1         3.8                0              0
## 21          1         3.4                0              0
## 22          1         3.7                0              0
## 23          1         3.6                0              0
## 24          1         3.3                0              0
## 25          1         3.4                0              0
## 26          1         3.0                0              0
## 27          1         3.4                0              0
## 28          1         3.5                0              0
## 29          1         3.4                0              0
## 30          1         3.2                0              0
## 31          1         3.1                0              0
## 32          1         3.4                0              0
## 33          1         4.1                0              0
## 34          1         4.2                0              0
## 35          1         3.1                0              0
## 36          1         3.2                0              0
## 37          1         3.5                0              0
## 38          1         3.6                0              0
## 39          1         3.0                0              0
## 40          1         3.4                0              0
## 41          1         3.5                0              0
## 42          1         2.3                0              0
## 43          1         3.2                0              0
## 44          1         3.5                0              0
## 45          1         3.8                0              0
## 46          1         3.0                0              0
## 47          1         3.8                0              0
## 48          1         3.2                0              0
## 49          1         3.7                0              0
## 50          1         3.3                0              0
## 51          1         3.2                1              0
## 52          1         3.2                1              0
## 53          1         3.1                1              0
## 54          1         2.3                1              0
## 55          1         2.8                1              0
## 56          1         2.8                1              0
## 57          1         3.3                1              0
## 58          1         2.4                1              0
## 59          1         2.9                1              0
## 60          1         2.7                1              0
## 61          1         2.0                1              0
## 62          1         3.0                1              0
## 63          1         2.2                1              0
## 64          1         2.9                1              0
## 65          1         2.9                1              0
## 66          1         3.1                1              0
## 67          1         3.0                1              0
## 68          1         2.7                1              0
## 69          1         2.2                1              0
## 70          1         2.5                1              0
## 71          1         3.2                1              0
## 72          1         2.8                1              0
```

```
## 73          1         2.5               1               0
## 74          1         2.8               1               0
## 75          1         2.9               1               0
## 76          1         3.0               1               0
## 77          1         2.8               1               0
## 78          1         3.0               1               0
## 79          1         2.9               1               0
## 80          1         2.6               1               0
## 81          1         2.4               1               0
## 82          1         2.4               1               0
## 83          1         2.7               1               0
## 84          1         2.7               1               0
## 85          1         3.0               1               0
## 86          1         3.4               1               0
## 87          1         3.1               1               0
## 88          1         2.3               1               0
## 89          1         3.0               1               0
## 90          1         2.5               1               0
## 91          1         2.6               1               0
## 92          1         3.0               1               0
## 93          1         2.6               1               0
## 94          1         2.3               1               0
## 95          1         2.7               1               0
## 96          1         3.0               1               0
## 97          1         2.9               1               0
## 98          1         2.9               1               0
## 99          1         2.5               1               0
## 100         1         2.8               1               0
## 101         1         3.3               0               1
## 102         1         2.7               0               1
## 103         1         3.0               0               1
## 104         1         2.9               0               1
## 105         1         3.0               0               1
## 106         1         3.0               0               1
## 107         1         2.5               0               1
## 108         1         2.9               0               1
## 109         1         2.5               0               1
## 110         1         3.6               0               1
## 111         1         3.2               0               1
## 112         1         2.7               0               1
## 113         1         3.0               0               1
## 114         1         2.5               0               1
## 115         1         2.8               0               1
## 116         1         3.2               0               1
## 117         1         3.0               0               1
## 118         1         3.8               0               1
## 119         1         2.6               0               1
## 120         1         2.2               0               1
## 121         1         3.2               0               1
## 122         1         2.8               0               1
## 123         1         2.8               0               1
## 124         1         2.7               0               1
## 125         1         3.3               0               1
## 126         1         3.2               0               1
```

```
## 127           1         2.8                 0                 1
## 128           1         3.0                 0                 1
## 129           1         2.8                 0                 1
## 130           1         3.0                 0                 1
## 131           1         2.8                 0                 1
## 132           1         3.8                 0                 1
## 133           1         2.8                 0                 1
## 134           1         2.8                 0                 1
## 135           1         2.6                 0                 1
## 136           1         3.0                 0                 1
## 137           1         3.4                 0                 1
## 138           1         3.1                 0                 1
## 139           1         3.0                 0                 1
## 140           1         3.1                 0                 1
## 141           1         3.1                 0                 1
## 142           1         3.1                 0                 1
## 143           1         2.7                 0                 1
## 144           1         3.2                 0                 1
## 145           1         3.3                 0                 1
## 146           1         3.0                 0                 1
## 147           1         2.5                 0                 1
## 148           1         3.0                 0                 1
## 149           1         3.4                 0                 1
## 150           1         3.0                 0                 1
## attr(,"assign")
## [1] 0 1 2 2
## attr(,"contrasts")
## attr(,"contrasts")$Species
## [1] "contr.treatment"
```

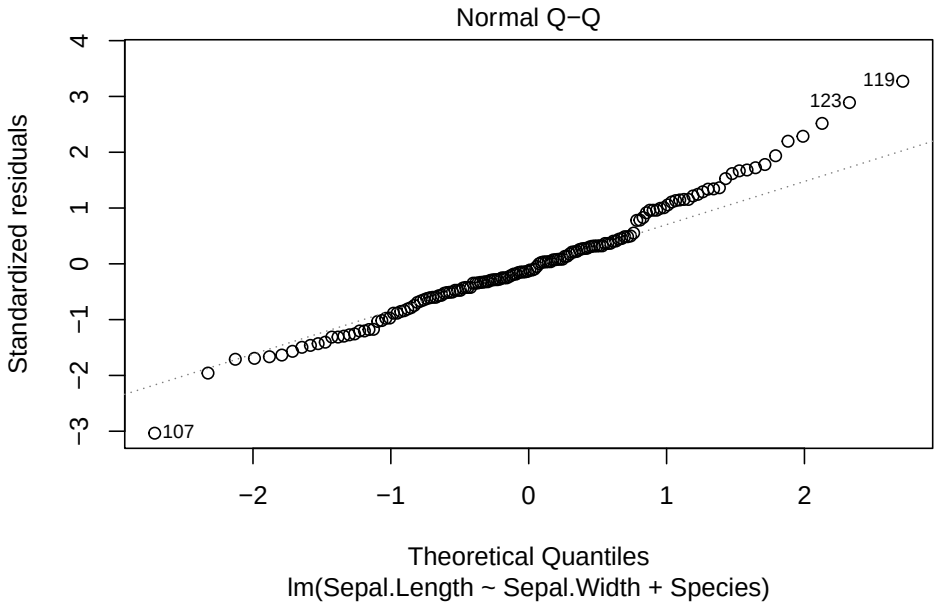
6.6) Feu els plots dels residus que creieu necessaris i treieu-ne tres conclusions que considereu importants.

```
par(mfrow = c(2,2))
plot(model2)
```



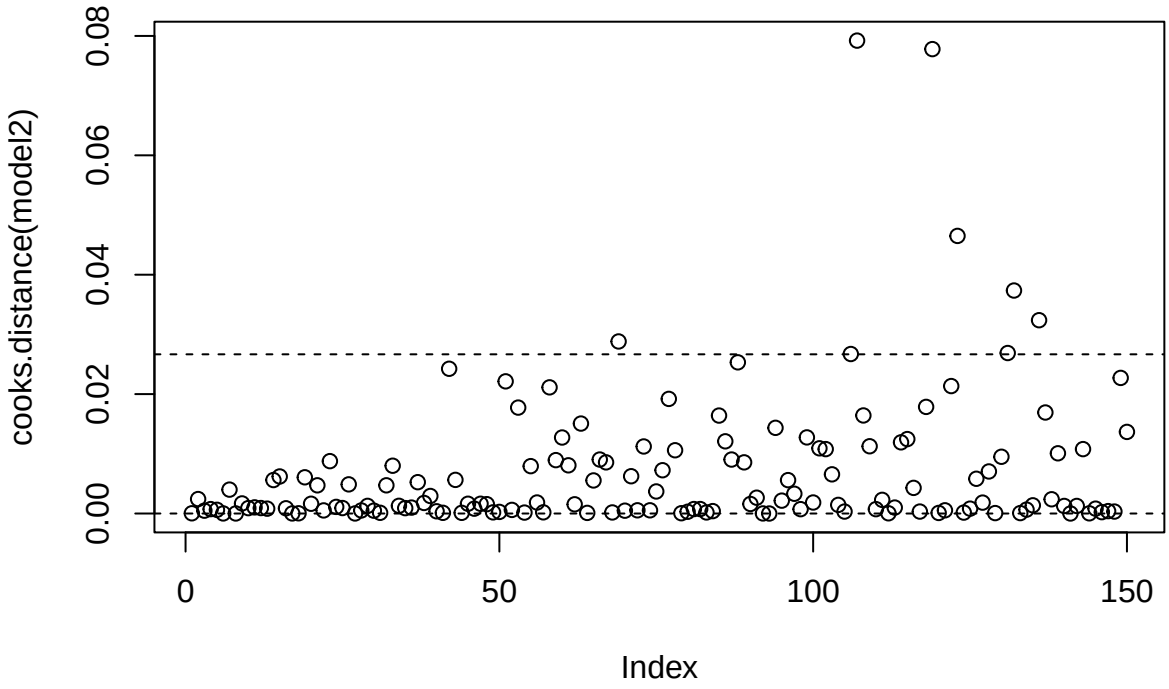
(mirem-nos el qq-plot més gran:)

```
plot(model2, which=2)
```



- L'hipòtesi d'homoscedasticitat ara es compleix més que amb el model anterior, ja que al plot de Residuals vs Fitted la variància dels residus és més homogènia o, equivalentment, al plot de Scale-Location la desviació augmenta molt poc, a diferència d'abans.
- L'hipòtesi de normalitat ha empitjorat bastant, ja que al qqplot tenim molta més desviació respecte la distribució normal, sobretot a l'extrem superior. Segurament tenim dades influents que ens desvien el model lineal.
- Del plot de Residuals vs Leverage es veu que hi ha observacions que podrien ser influents. Podem fer el mateix que abans per veure les distàncies de Cook:

```
plot(cooks.distance(model2))
abline(h=c(0,4/n),lty=2)
```



Seguim tenint dades molt influents i que tenen una distància de Cook per sobre del doble de la permesa, però respecte el model 1, ara les distàncies almenys ens han disminuït una mica.

6.7) Creieu que és necessari treure alguna observació del conjunt de dades?. Justifiqueu la vostra resposta. En cas que sigui necessàri feu-ho i torneu a ajustar el model.

Si, convindria treure les observacions influents que acabem de veure. Per tant, eliminem aquelles mostres que tinguin una Cook's distance superior a $4/n$, i tornem a entrenar el model linial amb el nou dataset. Les mostres influents són:

```
irisdat4$cook = cooks.distance(model2)
cond_cook = irisdat4$cook > 4/n
```

```r
indexs_true <- which(cond_cook)
indexs_true
```

```
## [1]  69 106 107 119 123 131 132 136
```

Sempre va bé situar les mostres al conjunt de dades:

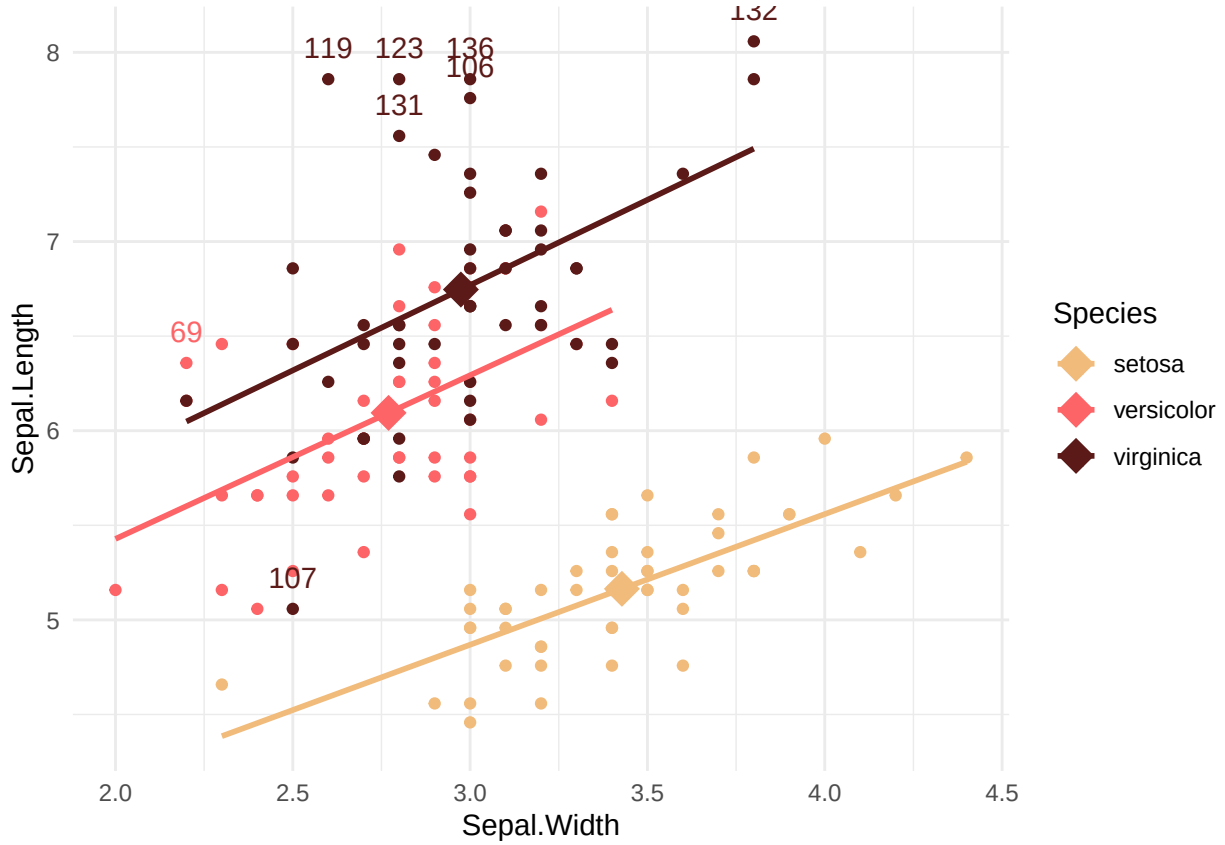
```r
gplot <- ggplot(irisdat4) +
  aes(y = Sepal.Length, x = Sepal.Width, color = Species) +
  geom_point() +
  geom_smooth(se = FALSE, method = "lm") +
  theme_minimal()


# Mostres de cook's distance gran
indexs = c(69, 106, 107, 119, 123, 131, 132, 136)
mostres_cook <- irisdat4[irisdat4$X %in% indexs, ]

# Afegir etiquetes a les mostres destacades
gplot <- gplot + geom_text(data = mostres_cook, aes(label = X), vjust = -1)


# Càlcul dels centroids per cada espècie
centroids <- irisdat4 %>%
  group_by(Species) %>%
  summarize(mean_Sepal.Length = mean(Sepal.Length),
            mean_Sepal.Width = mean(Sepal.Width))

gplot <- gplot +
  geom_point(data = centroids, aes(x = mean_Sepal.Width, y = mean_Sepal.Length),
             shape = 18, size = 6)
gplot <- gplot + scale_color_manual(values = wes_palette("GrandBudapest1", n = 3))
gplot
```



Els rombes representen els centroides de cada espècie. Es pot veure que la major part de les mostres influents pertanyen a l'espècie virginica.

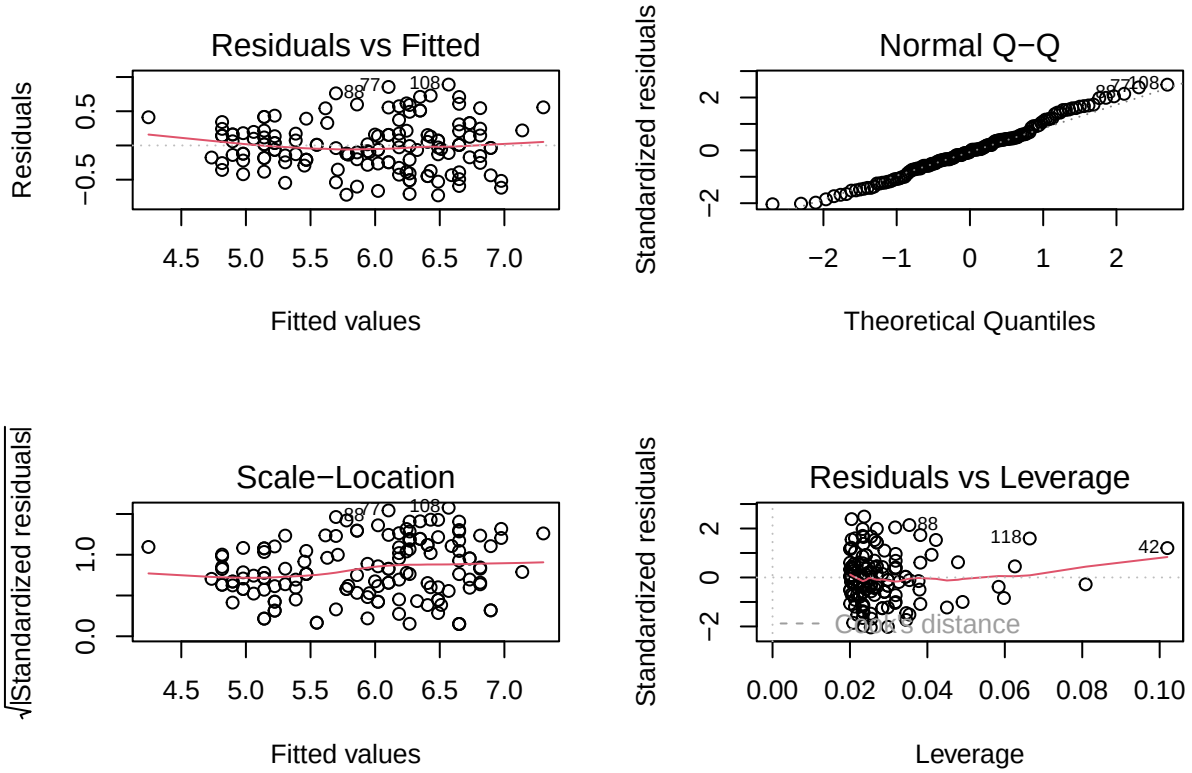
Ara, ajustem de nou el model 2 (li direm model2.1) amb el nou dataset:

```
irisdat4.1 <- irisdat4[- c(69, 106, 107, 119, 123, 131, 132, 136), ]
model2.1 = lm(Sepal.Length ~ Sepal.Width + Species, data=irisdat4.1)
summary(model2.1)
```

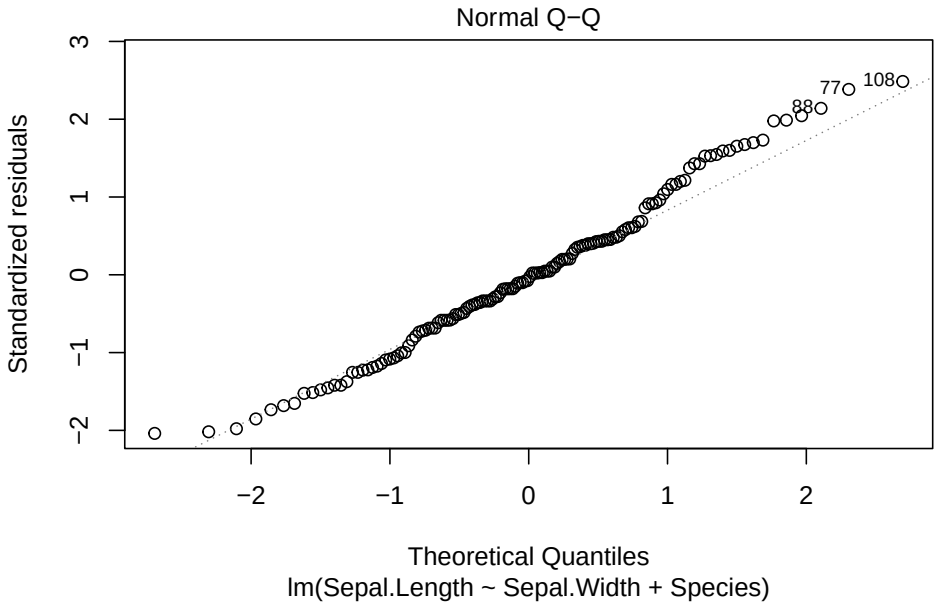
```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width + Species, data = irisdat4.1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72911 -0.23928 -0.01764  0.19352  0.88950
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.37431    0.31936   7.435 1.00e-11 ***
## Sepal.Width        0.81386    0.09196   8.851 3.76e-15 ***
## Speciesversicolor  1.45066    0.09400  15.433  < 2e-16 ***
## Speciesvirginica   1.83422    0.08581  21.376  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.3623 on 138 degrees of freedom
## Multiple R-squared:  0.7745, Adjusted R-squared:  0.7696
## F-statistic:   158 on 3 and 138 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(model2.1)
```




```
plot(model2.1, which=2)
```



Amb el model 2 tenim un R^2 de 0.7259, i amb el model 2.1 és de 0.7745. Ens ha agumentat, per tant el model 2.1 ara expica millor les dades. Pel que fa als residus, ara estan millor que amb el model 2. Els plots de Residual vs Fitted i Scale-Location estan quasi impecables, i el qq-plot a millorat un pèl, però segueix havent-hi molta corbatura respecte la distribució normal cap a l'extrem superior.

7) Amb el conjunt de dades que us hagiu quedat al final de l'apartat 6), ajusteu un model que assumeixi que les dues variables explicatives del model 2 interaccionen (model3).

```
model3 = lm(Sepal.Length ~ Sepal.Width * Species, data=irisdat4.1)
summary(model3)
```

```
##
## Call:
## lm(formula = Sepal.Length ~ Sepal.Width * Species, data = irisdat4.1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.74092 -0.24130  0.00719  0.18941  0.89355
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    2.7972     0.4714   5.934 2.33e-08 ***
## Sepal.Width                    0.6905     0.1367   5.051 1.39e-06 ***
## Speciesversicolor              0.6127     0.6719   0.912   0.3634
## Speciesvirginica               1.2631     0.7202   1.754   0.0817 .
## Sepal.Width:Speciesversicolor  0.2726     0.2190   1.245   0.2154
## Sepal.Width:Speciesvirginica   0.1731     0.2274   0.761   0.4479
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.3627 on 136 degrees of freedom
## Multiple R-squared:  0.7772, Adjusted R-squared:  0.769
## F-statistic:  94.9 on 5 and 136 DF,  p-value: < 2.2e-16
```

7.1) Us surt significativa la interacció?

No, no ens surt significativa.

7.2) Quina és l'estimació de la variància residual?

Obtenim una estimació de la variància de 0.1315627, que és quasi la mateixa que teníem amb el model 2.

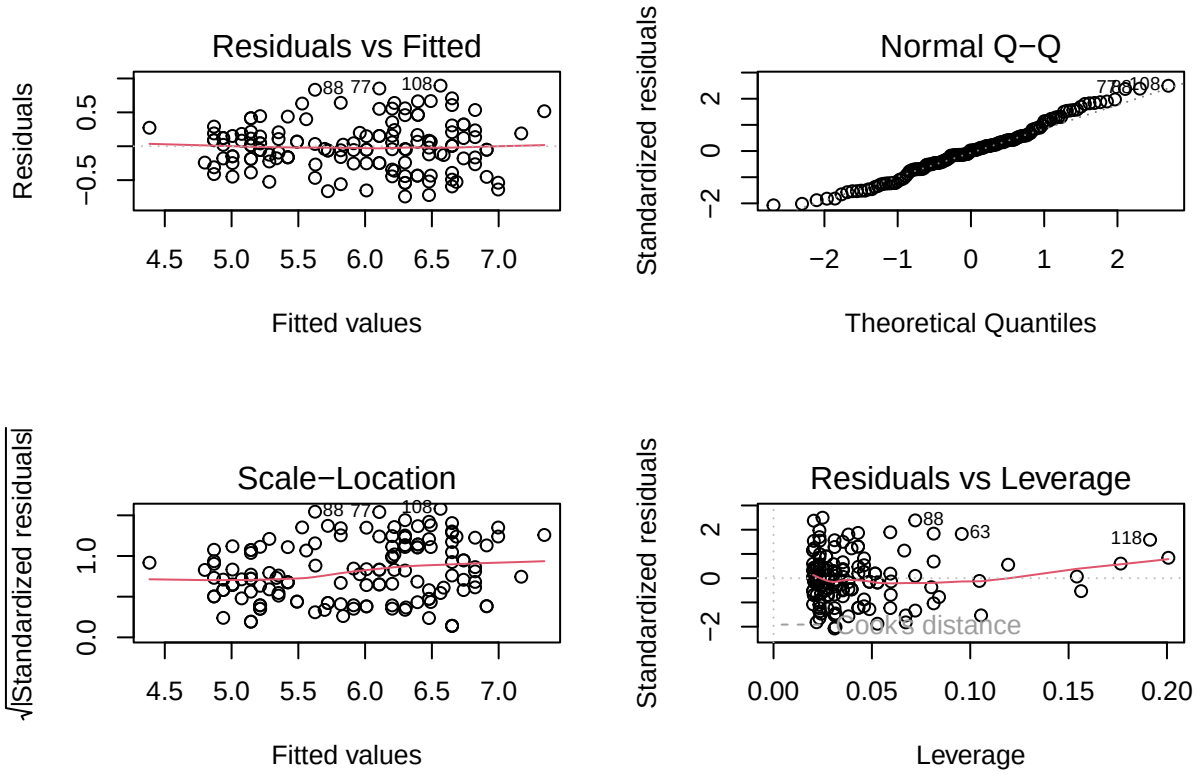
7.3) Interpreteu els paràmetres que us surtin significatius (una frase per a cada paràmetre).

Els paràmetres significatius que hem obtingut són:

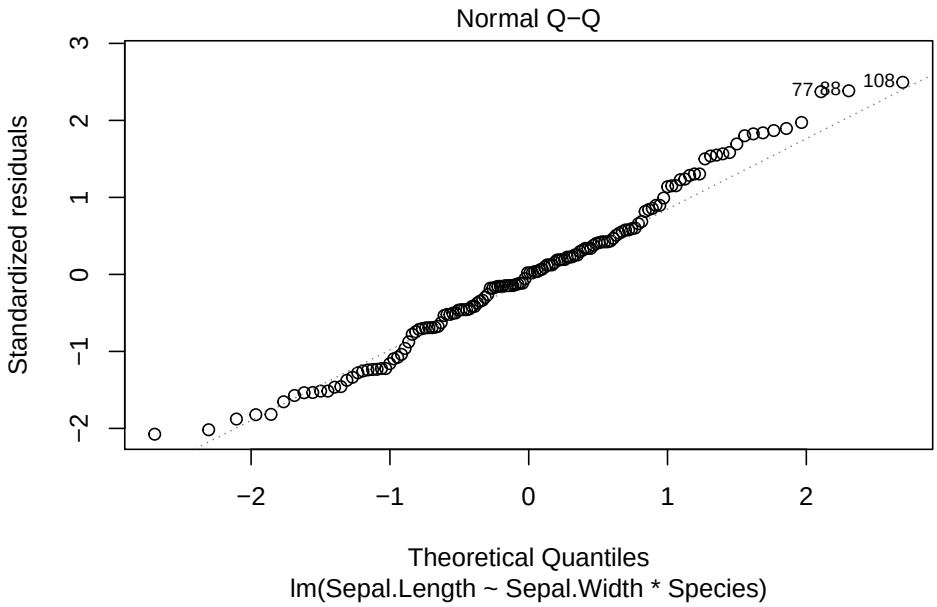
- Intercept: longitud del sèpal de l'espècie setosa (espècie de referència) quan l'amplada del sèpal és 0.
- Sepal.Width: la longitud que augmenta el sèpal quan augmentem en una unitat la seva amplada.
- Speciesvirginica: diferència de longitud del sèpal entre l'espècie virginica i setosa (fixada una amplada de sèpal).

7.4) Han millorat els gràfics de residus respecte als trobats amb el model 2? Justifiqueu la vostra resposta.

```
par(mfrow = c(2,2))
plot(model3)
```




```
plot(model3, which=2)
```



Com es pot veure, els residus no han millorat gaire. Potser una mica a la gràfica de Residual vs Fitted, però

el qq-plot continua mostrant bastanta desviació a l'extrem superior.

9) Feu una taula comparativa dels tres models ajustats i que us permeti decidir quin dels tres és el millor. Justifiqueu en paraules la vostra elecció.

Per poder fer una taula comparativa dels tres models alhora amb la funció *anova( )*, tots els models que li passem a la funció han d'estar entrenats amb el mateix dataset. Per tant, podem tornar a ajustar el model1. Realment no caldira perquè el model1 ja es veu que no serà el millor dels tres, ja que explica molt malament les dades al no tenir en compte les diferents espècies, i té un R-squared de 0.0138227, que és molt més baix que el dels models 2 i 3.

```
model1.new_data = lm(Sepal.Length ~ Sepal.Width, data=irisdat4.1)
anova(model1.new_data, model2.1, model3)
```

```
## Analysis of Variance Table
##
## Model 1: Sepal.Length ~ Sepal.Width
## Model 2: Sepal.Length ~ Sepal.Width + Species
## Model 3: Sepal.Length ~ Sepal.Width * Species
##   Res.Df    RSS Df Sum of Sq        F Pr(>F)
## 1    140 78.946
## 2    138 18.110  2    60.836 231.2060 <2e-16 ***
## 3    136 17.893  2     0.217   0.8249 0.4405
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Obtenim que el millor model és el 2.1. Això vol dir que aquest, amb menys paràmetres que el model3, explica les dades quasi igual de bé. També es pot veure comparant els adjusted R-squared dels models: el model 2.1 té un adj. R-squared de 0.7696185, i el model 3 té un adj. R-squared de 0.7690325, que és lleugerament inferior al model 2.1 i per tant "no val la pena" afegir els paràmetres de l'interacció.

10) Pel model escollit, prediu quina és la longitud esperada d'un sèpal que tingui una amplada de 3.8 per a cadascúna de les espècies que tenim? Expliciteu tant l'estimació puntual com l'intèrval de confiança al 95%.

```
newdata = data.frame(Sepal.Width = c(3.8, 3.8, 3.8),
                     Species = c("setosa", "versicolor", "virginica"))
predict(model2.1, newdata = newdata, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 5.466960 4.740389 6.193531
## 2 6.917622 6.170745 7.664499
## 3 7.301176 6.561477 8.040875
```

Per tant, la predicció de la llargada del sèpal d'una flor que te una amplada de sèpal de 3.8 és:

- 5.466960 si és de l'espècie setosa
- 6.917622 si és de l'espècie versicolor
- 7.301176 si és de l'espècie virginica

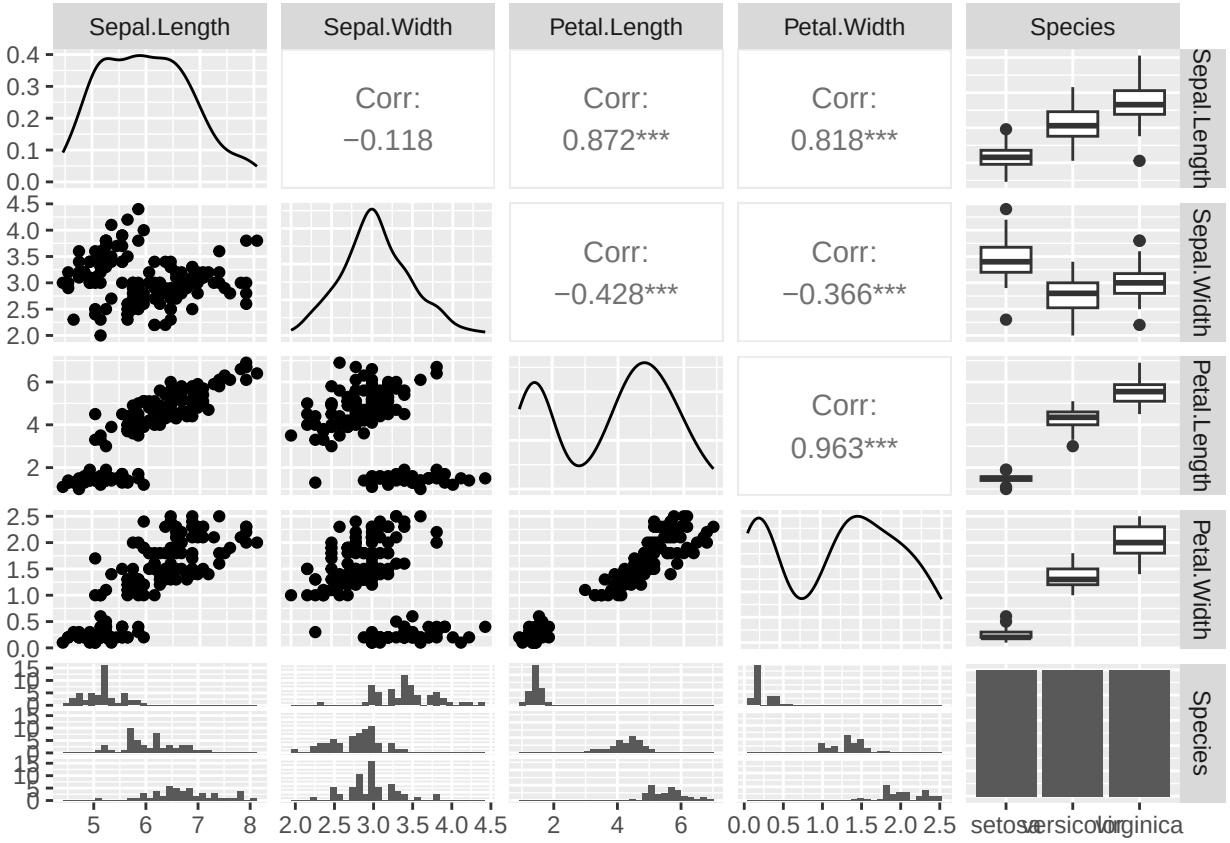
Dels intèrvals de confiança s'interpreta que el 95% de les flors que tinguin una amplada de sèpal de 3.8 i que siguin de l'espècie setosa, tindràn una llargada del sèpal dins de l'intèrval (4.740389, 6.193531). (ídem amb els intèrvals de les altres espècies).

Exercici 2

Treballarem el mateix conjunt de dades que a l'Exercici 1. Ara però volem predir la Longitud d'un pètal.Podeu fer servir la resta de variables que tenim com a explicatives.

1) Porteu a terme l'scatterplotmatrix de les dades. Què observeu?

```
irisdat4_pairs = irisdat4[,!(names(irisdat4) %in% c("X","X.1", "cook"))]
ggpairs(irisdat4_pairs)
```



Tres coses que veiem són:

- La varaible Petal.Length té molta correlació amb les variabes Petal.Width i Sepal.Length (cosa bona, ja que Petal.Length és la variable resposta)
- Les variables Petal.Width i Sepal.Length tenen bastanta correlació, cosa que voldríem evitar entre dues variables explicatives. Potser ens convindria agafar només una d'aquestes dues.
- Sembla ser que hi ha diferències significatives entre la longitud del pètal de les diferents espècies.

2) Ajusteu un model de regresió linel múltiple amb les explicatives que teniu (model 4). Per a aquest model contesteu les preguntes següents:

```
model4 = lm(Petal.Length ~ Petal.Width+Sepal.Length+Sepal.Width,data = irisdat4)
summary(model4)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width + Sepal.Length + Sepal.Width,
##     data = irisdat4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99333 -0.17656 -0.01004  0.18558  1.06909
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.37806    0.30374  -1.245    0.215
## Petal.Width   1.44679    0.06761  21.399   <2e-16 ***
## Sepal.Length  0.72914    0.05832  12.502   <2e-16 ***
## Sepal.Width  -0.64601    0.06850  -9.431   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.319 on 146 degrees of freedom
## Multiple R-squared:  0.968,  Adjusted R-squared:  0.9674
## F-statistic:  1473 on 3 and 146 DF,  p-value: < 2.2e-16
```

2.1) Quin és el valor de R2 i R2 adj?

Multiple R-squared: 0.968, Adjusted R-squared: 0.9674

2.2) Quins son els residus mínim i màxim?

El mínim és de -0.99333 i el màxim és de 1.06909

2.3) Hi ha alguna observació amb leverage?

```
p = ncol(model.matrix(model4))
n = nrow(irisdat4)
cond_lev = 3*p/n
res = irisdat4[which(hatvalues(model4)>cond_lev),]
res = data.frame(res$X, res$Petal.Length, res$Species)
res
```

```
##   res.X res.Petal.Length res.Species
## 1   132              6.4   virginica
```

Només tenim una observació amb leverage; la 132.

2.4) Hi ha alguna observació influent?

```
irisdat4$cook = cooks.distance(model4)
cond_cook = irisdat4$cook > 4/n

indexs_true <- which(cond_cook)
indexs_true
```

```
##  [1]  15  33  42 108 115 118 132 135 136 142 146
```

Podem veure que tenim bastantes mostres influents en aquest model.

2.5) Quina variable té el VIF més gran? És acceptable aquest valor?

```
vif(model4)
```

```
##  Petal.Width Sepal.Length  Sepal.Width
##     3.889961     3.415733     1.305515
```

El VIF més gran és de 3.8, de la variable Petal.Width, que vol dir que hi ha com a mínim dues variables que estan moderadament correlades, però és un valor acceptable (inferior a 5). Com que l'altra variable que té un VIF gran és la de Sepal.Length, implica que l'amplada del pètal i la longitud del sèpal estan correlades, tal com ja podiem veure al ggpairs de l'apartat 1).

2.6) Interpreteu els paràmetres associats a totes les variables que creieu us surtin significatives (una frase per a cada paràmetre).

- Petal.Width: el que augmenta la llargada del pètal per cada unitat que augmenti la seva amplada
- Sepal.Length: el que augmenta la llargada del pètal per cada unitat que augmenti la longitut del sèpal
- Sepal.Width: el que augmenta la llargada del pètal per cada unitat que augmenti l'amplada del sèpal

2.7) Podem afirmar que el coeficient de la variable Amplada del pètal és estadísticament igual a 1.5?

Podem veure els intèrvals de confiança dels coeficients del nostre model de la següent manera:

```
confint(model4)
```

```
##                   2.5 %     97.5 %
## (Intercept)  -0.9783543  0.2222243
## Petal.Width   1.3131702  1.5804166
## Sepal.Length  0.6138790  0.8443979
## Sepal.Width  -0.7813871 -0.5106378
```

Per tant, sí que podem afirmar que, amb un nivell de confiança del 95%, el coeficient de la variable amplada del pètal és estadísticament igual a 1.5

3) Amplieu el model anterior afegint la variable espècie (model5), i justifiqueu quin dels dos models (m4 i m5) és millor (feu una taula comparativa).

```
model5 = lm(Petal.Length ~ Petal.Width+Sepal.Length+Sepal.Width+Species, data = irisdat4)
summary(model5)
```

```
##
## Call:
## lm(formula = Petal.Length ~ Petal.Width + Sepal.Length + Sepal.Width +
##     Species, data = irisdat4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.78396 -0.15708  0.00193  0.14730  0.65418
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1.20718    0.27369  -4.411 2.00e-05 ***
## Petal.Width        0.60222    0.12144   4.959 1.97e-06 ***
## Sepal.Length       0.60801    0.05024  12.101  < 2e-16 ***
## Sepal.Width       -0.18052    0.08036  -2.246   0.0262 *
## Speciesversicolor  1.46337    0.17345   8.437 3.14e-14 ***
## Speciesvirginica   1.97422    0.24480   8.065 2.60e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2627 on 144 degrees of freedom
## Multiple R-squared:  0.9786, Adjusted R-squared:  0.9778
## F-statistic:  1317 on 5 and 144 DF,  p-value: < 2.2e-16
```

```
anova(model4, model5)
```

```
## Analysis of Variance Table
##
## Model 1: Petal.Length ~ Petal.Width + Sepal.Length + Sepal.Width
## Model 2: Petal.Length ~ Petal.Width + Sepal.Length + Sepal.Width + Species
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1    146 14.8529
## 2    144  9.9397  2    4.9133 35.59 2.756e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

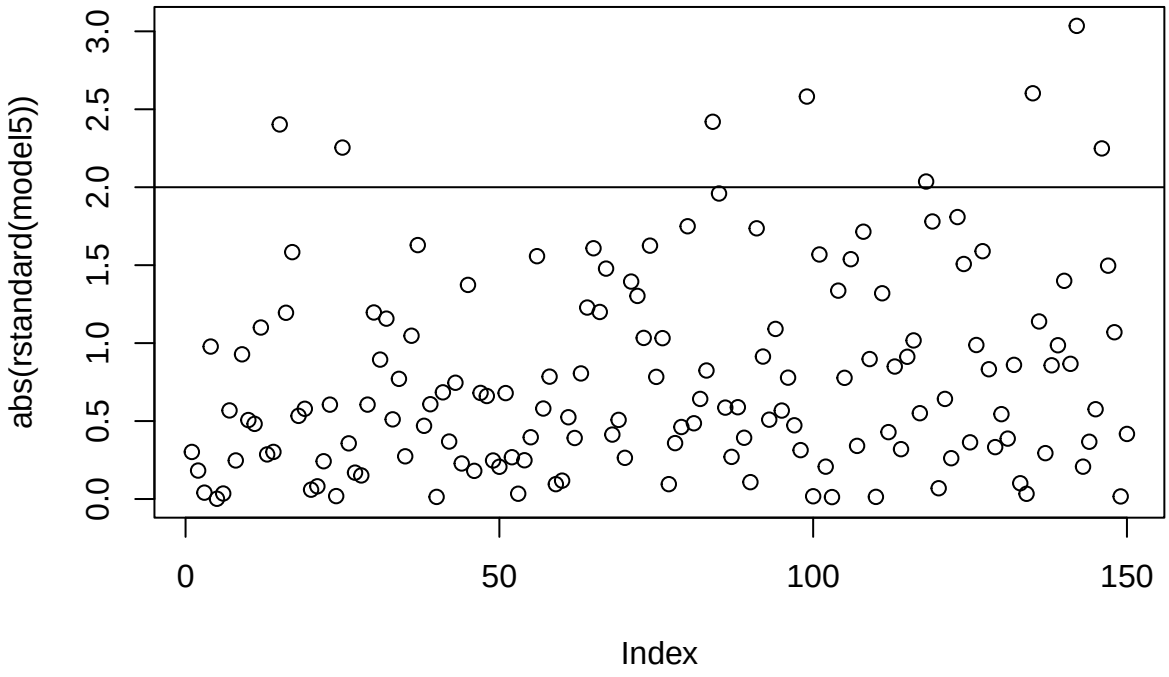
```
# H0: model4 millor
# H1: model5 millor
```

Obtenim que el millor model és el model5, ja que, tot i utilizar més paràmetres que el model4, val la pena ja que aconsegueix explicar millor les dades. També es pot veure comparant els adjusted R-squared, i es veu que el del model5 és lleugerament superior.

4) Pel model que hagiu escollit,

4.1) Quants residus estandaritzats cauen fora de l'inferval (-2; 2)? Creieu que és un percentatge escaient?

```
plot(abs(rstandard(model5)))
abline(a=2, b=0)
```



Tenim exactament 8 residus fora de l'interval (-2, 2), que representen un 5.33% dels residus. Aquest és un percentatge adequat, ja que indica que els residus segueixen una distribució normal.

4.2) Quina és la longitud del pètal esperada, per a una flor que tingui una amplada del pètal igual a 2.1, una longitud del sèpal igual a 5 i una amplada del sèpal igual a 3.5? Trobeu l'estimació puntual i per interval de confiança per a cadascúna de les espècies.

```
newdata = data.frame(Petal.Width = c(2.1, 2.1, 2.1),
                     Sepal.Width = c(3.5, 3.5, 3.5),
                     Sepal.Length = c(5, 5, 5),
                     Species = c("setosa", "versicolor", "virginica"))

predict(model5, newdata = newdata, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 2.465683 1.778414 3.152952
```

```
## 2 3.929054 3.348609 4.509499
## 3 4.439906 3.869825 5.009986
```

Per tant, la predicció de la llargada del pètal d'una flor d'aquestes característiques és:

- 2.465683 si és de l'espècie setosa
- 3.929054 si és de l'espècie versicolor
- 4.439906 si és de l'espècie virginica

Dels intèrvals de confiança s'interpreta que 95% de les flors que tinguin aquestes característiques, tindràn una longitud del pètal dins de l'intèrval:

- (1.778414, 3.152952) si és de l'espècie setosa
- (3.348609, 4.509499) si és de l'espècie versicolor
- (3.869825, 5.009986) si és de l'espècie virginica