Devoir3

Jonathan Domingue 300246863

2023-11-18

Exercise 5.5

```
## First we will load the data from the table into R
data <- data.frame(
  Weeks = c(4,5,6,7,8,9,10,11,12,13,14,15,16,17) ,
  DefectsPer10000= c(13.0,16.1,14.5,17.8,22.0,27.4,16.8,34.2,65.6,49.2,66.2,81.2,87.4,114.5)
)
```

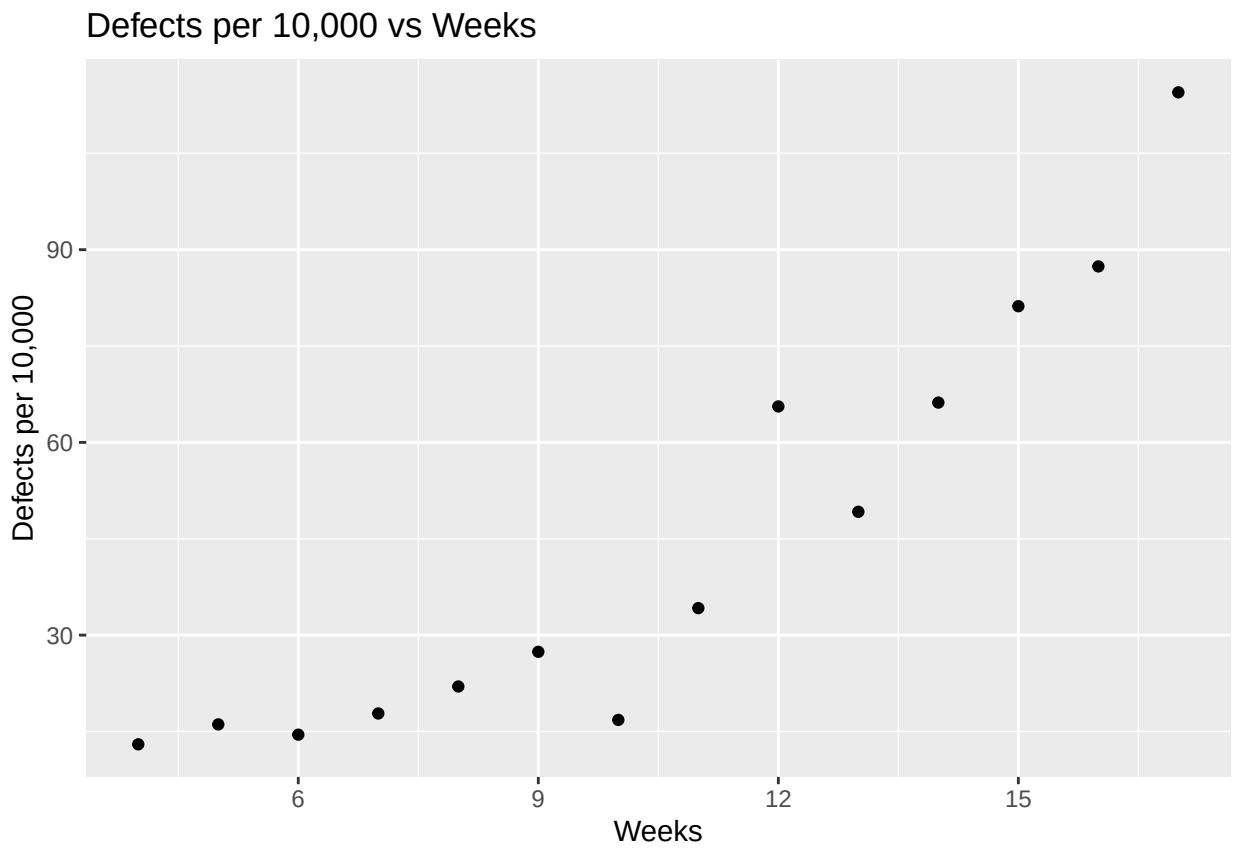
Visualisation de nos donnees

En premiere etape on va essayer de visualiser nos donnes pour voire en general ce qui se passe

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
ggplot(data=data, mapping= aes(x=Weeks , y= DefectsPer10000)) + geom_point() + labs(title = "Defects pe
```



Maintenant qu'on a une petite visualisation de nos donnes on peut commencer notre analyse

```
Model= lm ( DefectsPer10000 ~Weeks , data=data)

## We will then write the summary of the table which shows the T  stats and P-values
summary(Model)
```

```
##
## Call:
## lm(formula = DefectsPer10000 ~ Weeks, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.2688  -5.9229   0.5497   8.4203  22.4943
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -31.6982     9.7758  -3.243  0.00705 **
## Weeks         7.2767     0.8692   8.372 2.35e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.11 on 12 degrees of freedom
## Multiple R-squared:  0.8538, Adjusted R-squared:  0.8416
## F-statistic: 70.09 on 1 and 12 DF,  p-value: 2.354e-06
```

Analyse des Resultats Obtenus

Du summary obtenue, on peut voire que les p-value du coefficient Beta de weeks est inferieur a 0.05. Ceci nous fait savoir que la statistique est en effet significative

On peut egalement noter que le coefficient de R-squared et du Adjusted R-squared est plutot eleve et proche de 1, qui nous indique que les points sont plus ou moins sur la ligne de regression estimee.

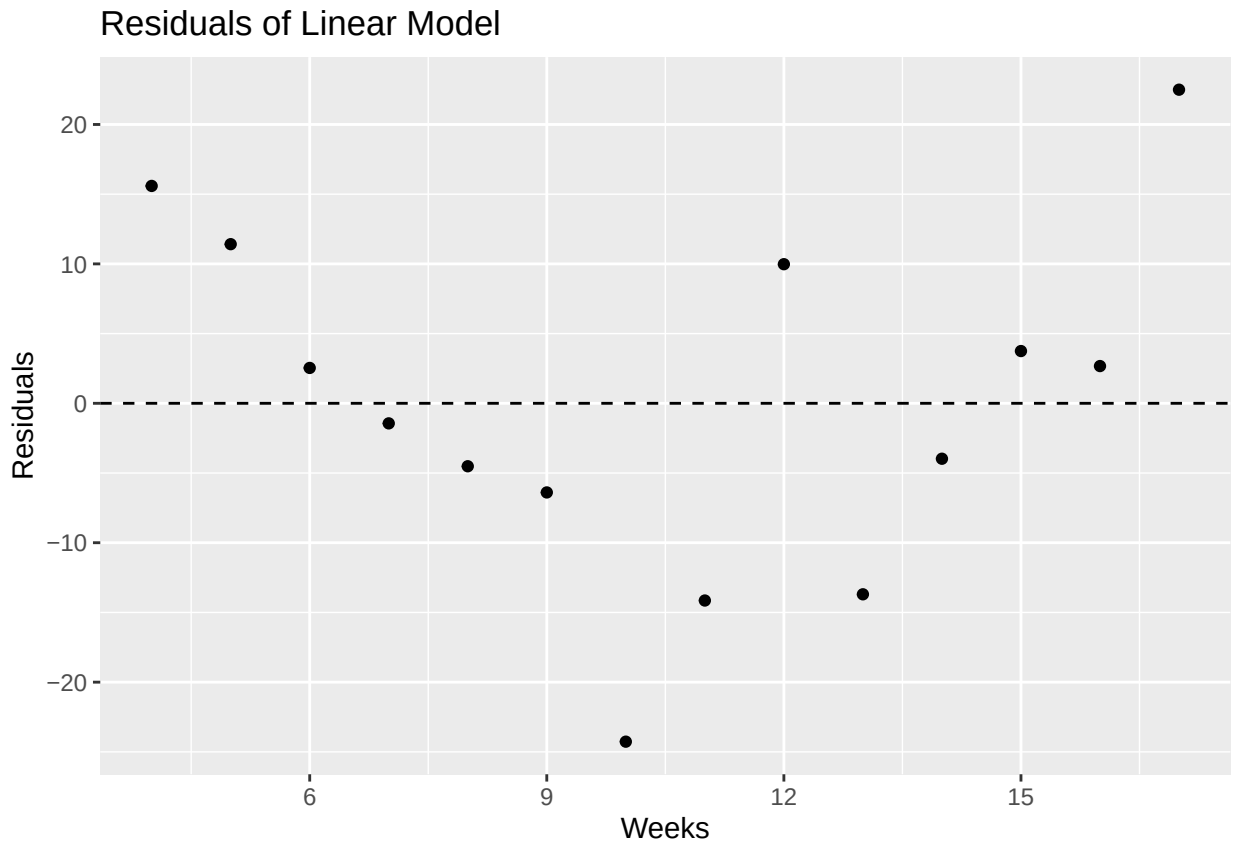
Le R-squared 0.85 signifie que la variable independante, qui dans ce cas est le nombres de semaines, explique 85% de la variation totale. Nous pouvons donc dire que le coefficient de Beta de Weeks est statistiquement significatif et que le modele semble bien fonctionner.

La prochaine etape sera maintenant de verifier si nous avons la Normalite pour les Residus

Test de la Normalite des Residus

Nous allons faire un graphique des residus pour potentiellement identifier les modeles non pris en compte par le modele

```
ggplot(data=data, aes(x=Weeks, y= residuals(Model)) ) + geom_point() + geom_hline(yintercept=0, linetyp
  labs(title = "Residuals of Linear Model",
       x = "Weeks",
       y = "Residuals")
```



Analyse du Graphique Residuals vs Weeks

De ce graphique nous pouvons clairement voire que la variance n'est pas constante. Ca ressemble meme a une fonction quadratique.

Ceci met en evidence la non-linearite des données. Puisque nos variances ne sont pas constantes nous avons un cas de heterodescaticite qui peut conduire a des estimations inefficaces des coefficients et peut rendre les erreurs standard des coefficients biaisés.

Ce phenomene affecte a son tour les resultats des tests d'hypothese.

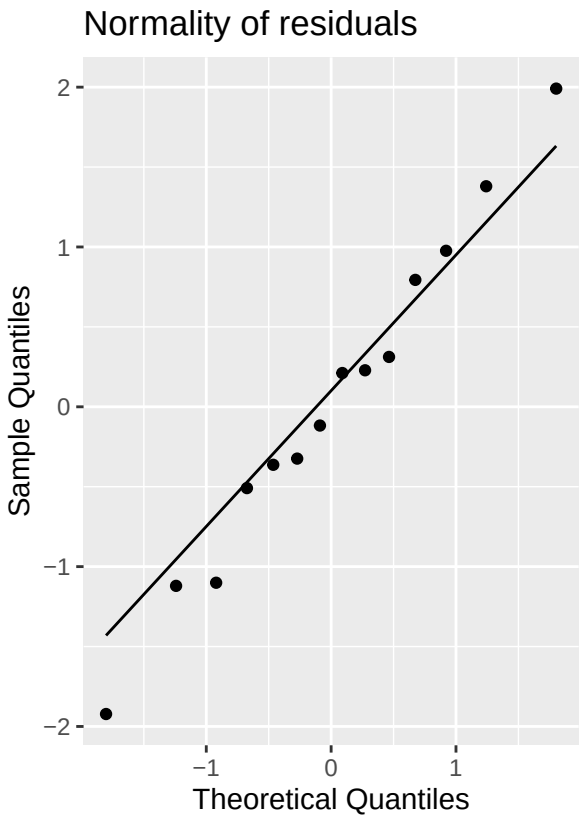
Conclusion: Nous allons devoire utiliser une transformation de Box Cox

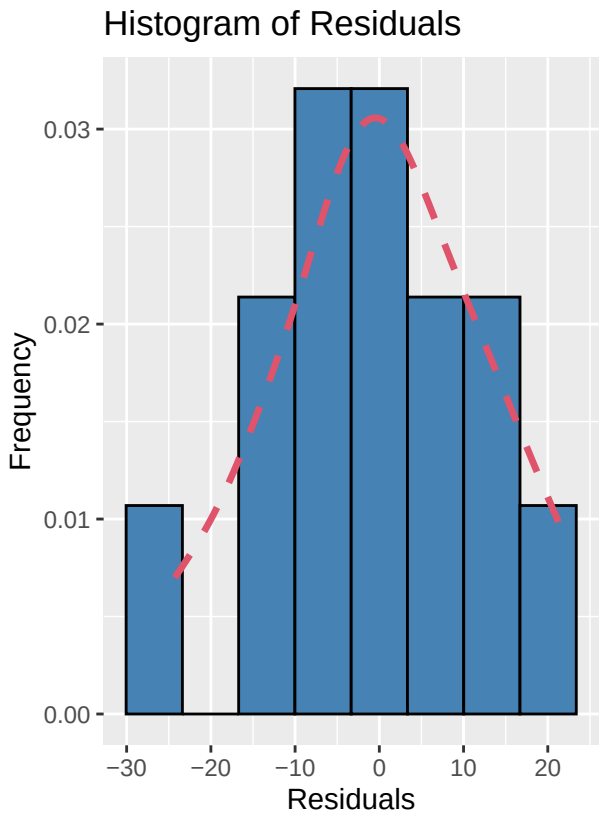
```
library(patchwork)
```

```
## Warning: package 'patchwork' was built under R version 4.2.3
```

```
p1<-ggplot( Model, aes(sample = rstandard(Model)) ) +
    geom_qq() +  stat_qq_line()+ labs(
  title    = "Normality of residuals",
  x        = "Theoretical Quantiles",
  y        = "Sample Quantiles")

p2<- ggplot(data, aes(x = Model$residuals)) +
    geom_histogram(aes(y = ..density..),bins = 8, fill = 'steelblue', color = 'black') +
    labs(title = 'Histogram of Residuals', x = 'Residuals', y = 'Frequency')+geom_density(lwd = 1.2, li
p1|p2
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```





La Transformation du Box Cox

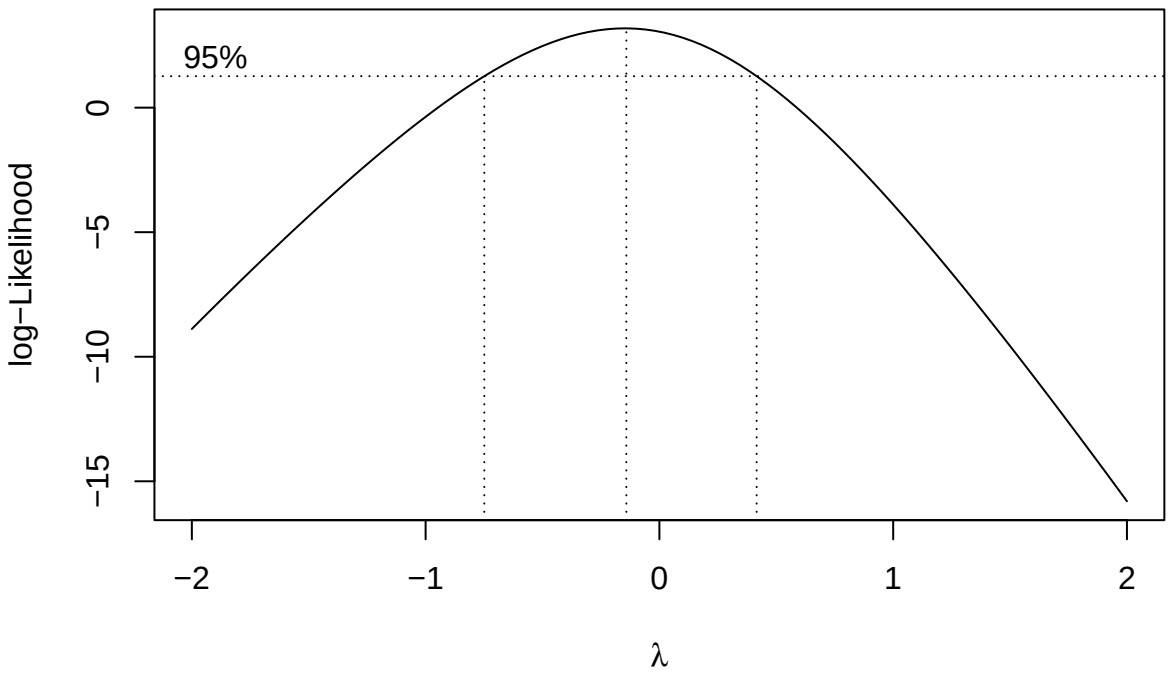
```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.2.3
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:patchwork':
##
##     area
```

```
bc <- boxcox(DefectsPer10000 ~Weeks , data= data)
```

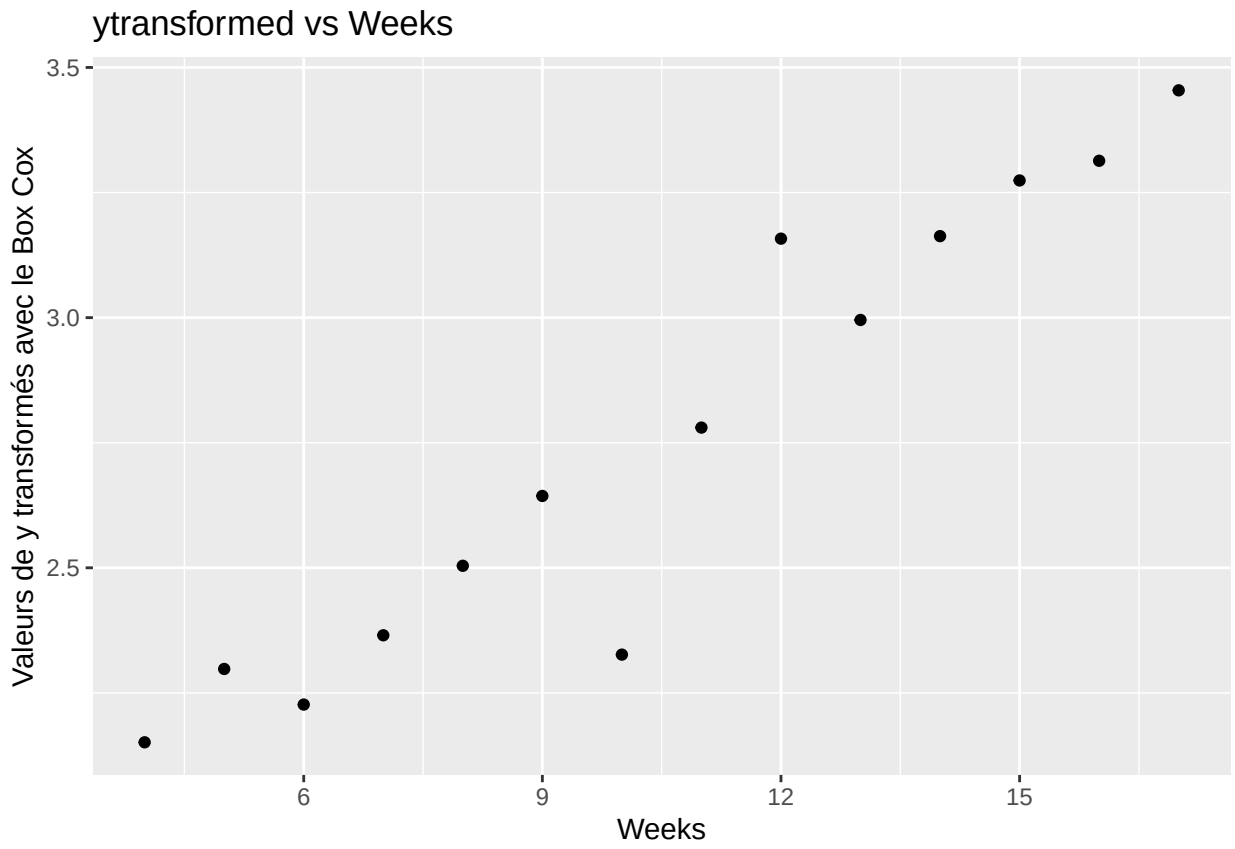



```
lambda_optimal <- bc$x[which.max(bc$y)]
lambda_optimal ##Printing the optimal value of lambda
```

```
## [1] -0.1414141
```

De notre transformation du Box Cox, nous avons obtenue que lambda est egale a -0.14 Puisque lambda n'est pas egal a zero nous allons faire une transformationd de puissance. Cela donnera le meilleur ajustement.

```
y_transformed <- (data$DefectsPer10000^lambda_optimal -1)/lambda_optimal

##Nouveau Model avec y qui a ete transforme
PowerModel <- lm(y_transformed ~ Weeks , data=data)

ggplot(data=data, mapping= aes(x=Weeks , y= y_transformed)) + geom_point() + labs(title = "ytransformed
```




```
summary(PowerModel)
```

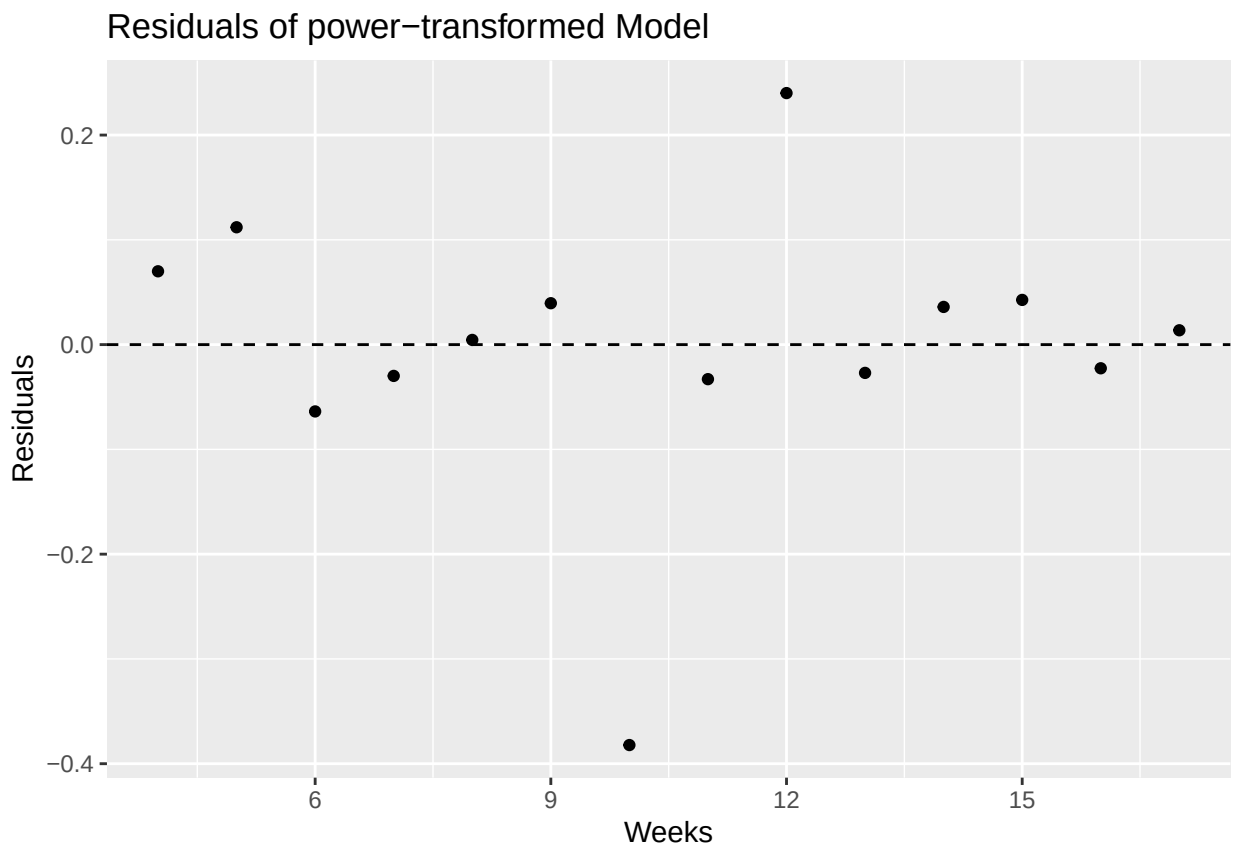
```
##
## Call:
## lm(formula = y_transformed ~ Weeks, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38219 -0.02914  0.00908  0.04189  0.24007
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.663022   0.103975   15.99 1.86e-09 ***
## Weeks       0.104565   0.009244   11.31 9.32e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1394 on 12 degrees of freedom
## Multiple R-squared:  0.9142, Adjusted R-squared:  0.9071
## F-statistic: 127.9 on 1 and 12 DF,  p-value: 9.318e-08
```

Analyse des Resultats Obtenus avec Le Box Cox

Comme cela etait prevu, on a un resultat ameliore de la partie precedente. On a une valeur de R-squared et Adjusted R-Squared plus eleve. Le nouveau Model avec le Box Cox explique 90.7% de la variation totale.

Résidus du Model Transforme en Puissance

```
# Un graphique des résidus du modèle transformé en puissance peut révéler si la transformation a été bé
ggplot(data=data, aes(x=Weeks, y=residuals(PowerModel))) +
  geom_point() +
  geom_hline(yintercept=0, linetype="dashed") +
  labs(title = "Residuals of power-transformed Model",
       x = "Weeks",
       y = "Residuals")
```



Comme nous pouvons le voire,le graphique des residus semble etre fermé. La variance des residus est plus ou moins constante pour les differentes valeurs de Weeks.

Nous avons obtenu l'homoscedacité.

Conclusion

Apres la transformation de Puissance avec le Box cox, on a obtenu un meilleur model, qui a une valeur plus eleve pour le R carre adjusted et pour lequelle on a obtenu l'homoscedacité

Exercise 5.23

Charger les donnees dans R

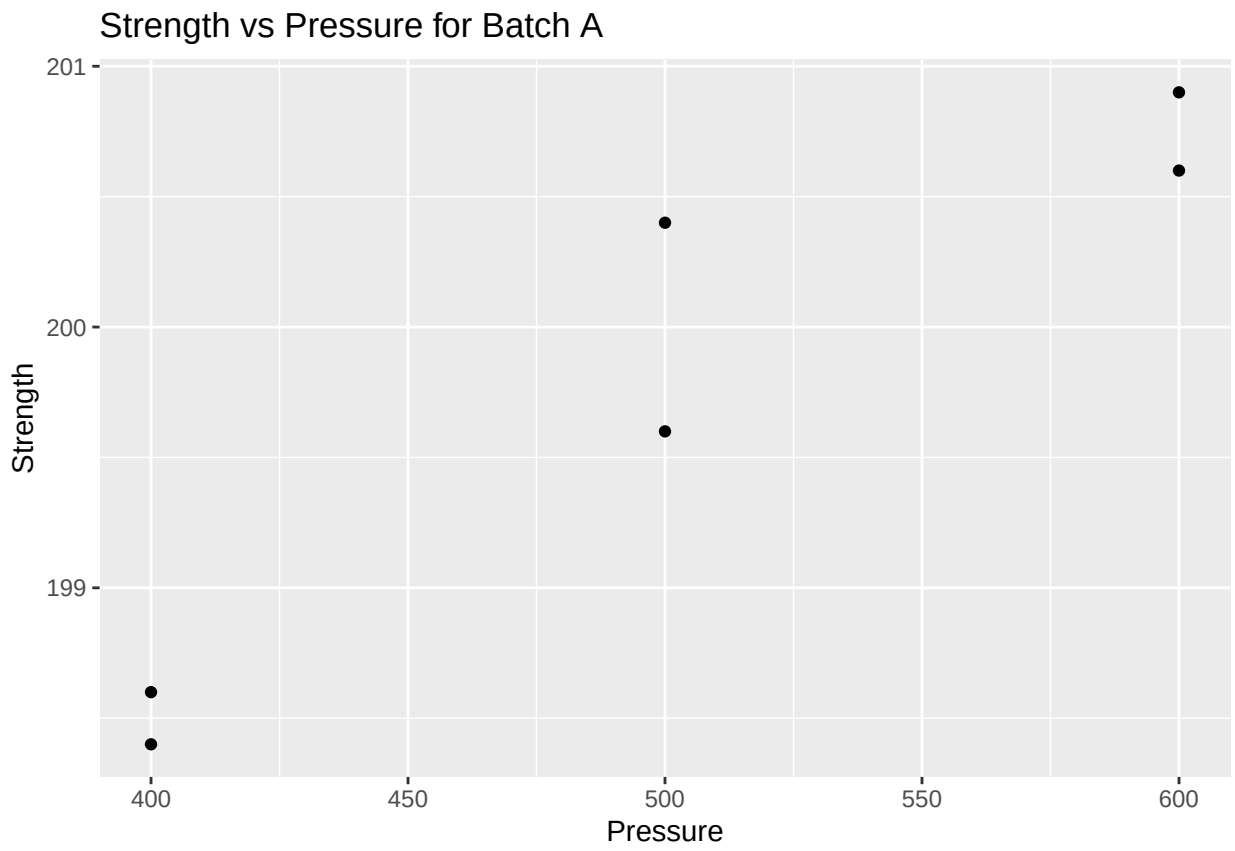
```
## First we will load the data from the table into R
data2 <- data.frame(

  StrengthBatchA= c(198.4 , 198.6 , 199.6 , 200.4 , 200.6, 200.9) ,
  StrengthBatchB = c(197.5 , 198.1 , 198.7 , 198.0 , 199.6 , 199.0) ,
  StrengthBatchC = c( 197.6 , 198.4 , 197.0 , 197.8 , 198.5 , 199.8) ,
  PressureBatchA= c(400,400,500,500,600,600) ,
  PressureBatchB = c(400,400,500,500,600,600) ,
  PressureBatchC = c(400,400,500,500,600,600)

)
```

```
## Nous allons egalement sauvegarder les donnees et faire un tableau avec tout
FullData <- data.frame(
  Batch = c("A", "A", "A", "A", "A", "A", "B", "B", "B", "B", "B", "B", "C", "C", "C", "C", "C", "C"),
  Pressure = c(400, 400, 500, 500, 600, 600, 400, 400, 500, 500, 600, 600, 400, 400, 500, 500, 600, 600
  Strength = c(198.4, 198.6, 199.6, 200.4, 200.6, 200.9, 197.5, 198.1, 198.7, 198.0, 199.6, 199.0, 197.
)
```

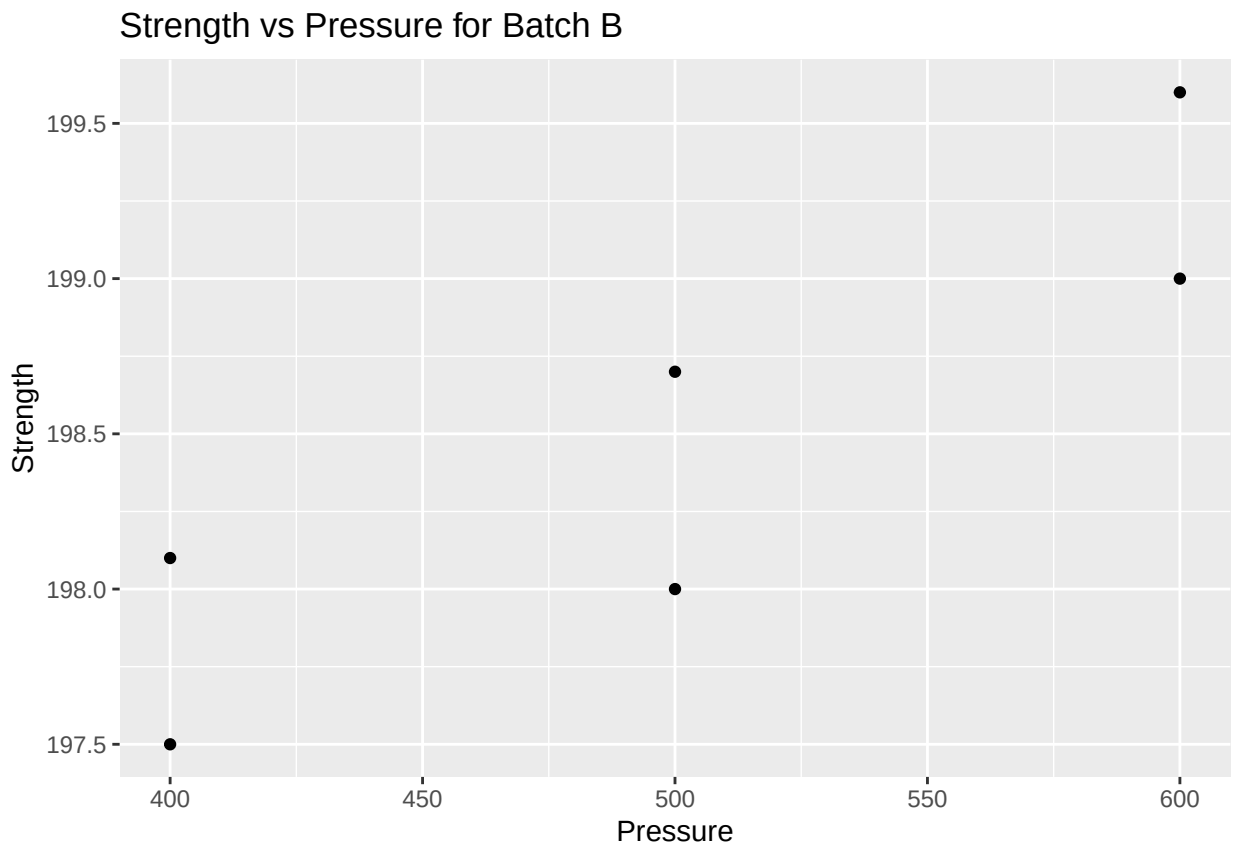
Visualisation des donnés

En premier lieu on va essayer de visualiser le jeu de données pour voire un peu ce qui se passe

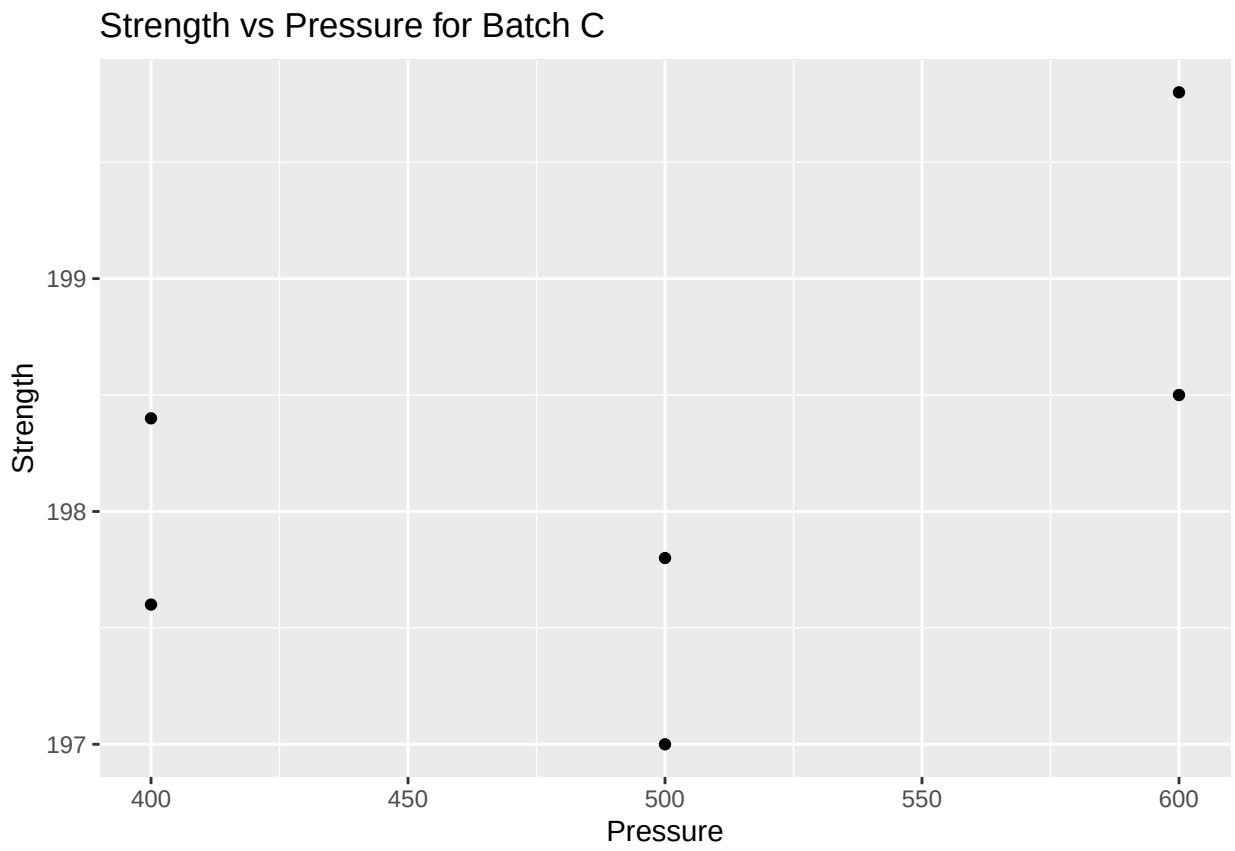
```
library(ggplot2)

ggplot(data=data2, mapping= aes(x=PressureBatchA , y= StrengthBatchA)) + geom_point() + labs(title = "S
```




```
ggplot(data=data2, mapping= aes(x=PressureBatchB , y= StrengthBatchB)) + geom_point() + labs(title = "S
```




```
ggplot(data=data2, mapping= aes(x=PressureBatchC , y= StrengthBatchC)) + geom_point() + labs(title = "S
```



Statistiqies Descriptives des données

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
summary_stats <- FullData %>% group_by(Batch, Pressure) %>% summarise(Mean_Strength = mean(Strength), SD
```

```
## `summarise()` has grouped output by 'Batch'. You can override using the
## `.groups` argument.
```

```r
print(summary_stats)
```

```
## # A tibble: 9 x 5
## # Groups:   Batch [3]
##   Batch Pressure Mean_Strength SD_Strength Count
##   <chr>    <dbl>         <dbl>       <dbl> <int>
## 1 A          400          198.       0.141     2
## 2 A          500          200        0.566     2
## 3 A          600          201.       0.212     2
## 4 B          400          198.       0.424     2
## 5 B          500          198.       0.495     2
## 6 B          600          199.       0.424     2
## 7 C          400          198        0.566     2
## 8 C          500          197.       0.566     2
## 9 C          600          199.       0.919     2
```

Analyse des Differents Batchs

```r
## Batch A
ModelStrengthA = lm(StrengthBatchA ~ PressureBatchA , data=data2)
summary(ModelStrengthA)
```

```
##
## Call:
## lm(formula = StrengthBatchA ~ PressureBatchA, data = data2)
##
## Residuals:
##      1      2      3      4      5      6
## -0.225 -0.025 -0.150  0.650 -0.275  0.025
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.941e+02  9.583e-01 202.567 3.56e-09 ***
## PressureBatchA 1.125e-02  1.892e-03   5.947  0.00401 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3783 on 4 degrees of freedom
## Multiple R-squared:  0.8984, Adjusted R-squared:  0.873
## F-statistic: 35.37 on 1 and 4 DF,  p-value: 0.00401
```

```r
## Batch B
ModelStrengthB = lm(StrengthBatchB ~ PressureBatchB , data=data2)
summary(ModelStrengthB)
```

```
##
## Call:
## lm(formula = StrengthBatchB ~ PressureBatchB, data = data2)
##
## Residuals:
##       1       2       3       4       5       6
## -0.2333  0.3667  0.2167 -0.4833  0.3667 -0.2333
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.947e+02  1.028e+00 189.493 4.65e-09 ***
## PressureBatchB 7.500e-03  2.028e-03   3.697   0.0209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4057 on 4 degrees of freedom
## Multiple R-squared:  0.7736, Adjusted R-squared:  0.717
## F-statistic: 13.67 on 1 and 4 DF,  p-value: 0.02088
```

```
## Batch B
ModelStrengthC = lm(StrengthBatchC ~ PressureBatchC , data=data2)
summary(ModelStrengthC)
```

```
##
## Call:
## lm(formula = StrengthBatchC ~ PressureBatchC, data = data2)
##
## Residuals:
##         1         2         3         4         5         6
## -0.008333  0.791667 -1.183333 -0.383333 -0.258333  1.041667
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.953e+02  2.310e+00  84.556 1.17e-07 ***
## PressureBatchC 5.750e-03  4.559e-03   1.261    0.276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9118 on 4 degrees of freedom
## Multiple R-squared:  0.2845, Adjusted R-squared:  0.1056
## F-statistic: 1.591 on 1 and 4 DF,  p-value: 0.2758
```

Des analyses faites plus haut, nous pouvons venir a la conclusion que la statistique est significative pour le Batch A et B. On observe des p-value inferieur a 0.05 qui demontre que la statistque est significative.

De plus nous voyons des valeurs de R-carre plutot elevees pour le Batch A et le Batch B.(0.873 et 0.717 respectivement). Cela nous permet de savoir que la variation du Model pour A est explique a 87.3% pour A et 71.7% pour B.

Cependant pour C on voit que les resultats des p-values sont plutot grandes. La valeur de R-carre et R-carre ajuste est egalement plutot petites et pres de zero. La variation du Model est explique a 10.56% par C.

Analyse des Batchs Ensemble

```
ModelTout= lm (StrengthBatchA+StrengthBatchB+StrengthBatchC ~PressureBatchA+PressureBatchB+PressureBatch
summary(ModelTout)
```

```
##
## Call:
## lm(formula = StrengthBatchA + StrengthBatchB + StrengthBatchC ~
##     PressureBatchA + PressureBatchB + PressureBatchC, data = data2)
##
## Residuals:
##       1       2       3       4       5       6
## -0.4667  1.1333 -1.1167 -0.2167 -0.1667  0.8333
##
## Coefficients: (2 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.842e+02  2.376e+00 245.892 1.64e-09 ***
## PressureBatchA 2.450e-02  4.689e-03   5.225  0.00641 **
## PressureBatchB        NA         NA      NA       NA
## PressureBatchC        NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9379 on 4 degrees of freedom
## Multiple R-squared:  0.8722, Adjusted R-squared:  0.8402
## F-statistic:  27.3 on 1 and 4 DF,  p-value: 0.006407
```

```
# Subset the data for Batch A and Batch B
data_reduit <- FullData[FullData$Batch %in% c("A", "B"), ]

# Display the new dataframe
print(data_reduit)
```

```
##    Batch Pressure Strength
## 1      A      400    198.4
## 2      A      400    198.6
## 3      A      500    199.6
## 4      A      500    200.4
## 5      A      600    200.6
## 6      A      600    200.9
## 7      B      400    197.5
## 8      B      400    198.1
## 9      B      500    198.7
## 10     B      500    198.0
## 11     B      600    199.6
## 12     B      600    199.0
```

```
# Fit the model for batches A and B (excluding C)
ModelReduit <- lm(Strength ~ Pressure, data = data_reduit)

# Display the summary of the model
summary(ModelReduit)
```

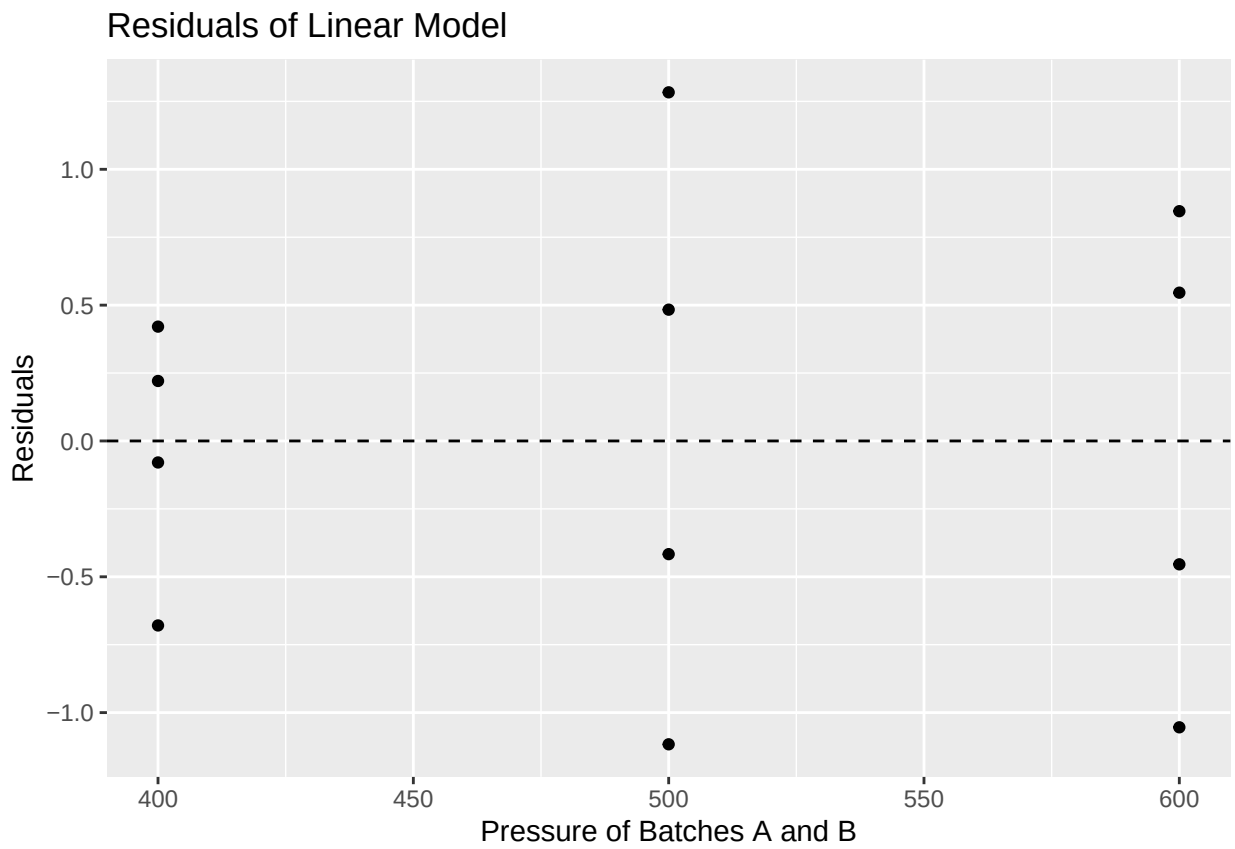
```
##
## Call:
## lm(formula = Strength ~ Pressure, data = data_reduit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11667 -0.51042  0.07083  0.49896  1.28333
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.944e+02  1.425e+00 136.483  < 2e-16 ***
## Pressure    9.375e-03  2.812e-03   3.334  0.00757 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7953 on 10 degrees of freedom
## Multiple R-squared:  0.5264, Adjusted R-squared:  0.4791
## F-statistic: 11.12 on 1 and 10 DF,  p-value: 0.007566
```

Nous pouvons observer que le Model Reduit sans C a clairement ameliore la valeur de R carre Ajusté d'approximativement 0.33 a 0.48. Le model reduit explique 48% de la variation totale.Ca nous fait savoir qu'il semble que le model reduit pourrait mieux expliquer la variation totale du model

Nous notons aussi que pour le Model complet et Reduit nous avions tous deux des p-value inferieur a 0.05 pour le T statistique et la F-Statistique. Cela nous demontrait que la statistique etait significative

Tester la Normalite des Residus

```r
# Un graphique des résidus permet d'identifier les modèles non pris en compte par le modèle.
ggplot(data=data_reduit, aes(x=Pressure, y=residuals(ModelReduit) ) ) +
  geom_point() +
  geom_hline(yintercept=0, linetype="dashed") +
  labs(title = "Residuals of Linear Model",
       x = "Pressure of Batches A and B",
       y = "Residuals")
```



Comme nous pouvons voire, nous avons une variance plus ou moins constante des residus. Nous avons la normalite des residus. On peut dire qu'on a ce qu'on appele homoscedascite.

Nous n'avaons donc pas besoin de faire de Transformation de Box Cox

Approche: Variable Categorielles

Dans notre cas ici nous savons aussi que nous avons des variables categorielles (categorical variables) La variable categorielle est le Batch qui sont des categories A, B et C pour la pression. Nous allons donc aussi essayer d'explorer un peu dans cette direction.

A noter que nous voyons dans l'énoncé qu'il ya une interaction entre le Batch et la Pression

```
Model2 <- lm(Strength ~ Pressure* Batch , data = FullData)
summary(Model2)
```

```
##
## Call:
## lm(formula = Strength ~ Pressure * Batch, data = FullData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1833 -0.2521 -0.0875  0.3292  1.0417
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     194.125000   1.560946 124.364  < 2e-16 ***
```

```
## Pressure          0.011250   0.003081   3.651  0.00332 **
## BatchB            0.608333   2.207510   0.276  0.78756
## BatchC            1.183333   2.207510   0.536  0.60172
## Pressure:BatchB  -0.003750   0.004357  -0.861  0.40631
## Pressure:BatchC  -0.005500   0.004357  -1.262  0.23084
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6162 on 12 degrees of freedom
## Multiple R-squared:  0.788,  Adjusted R-squared:  0.6996
## F-statistic: 8.919 on 5 and 12 DF,  p-value: 0.0009869
```

Tester la Significativité: Variables Categorielles

```
Model_MR = lm(Strength ~ Pressure, data=FullData)
Model_MC = lm (Strength ~Pressure*Batch , data=FullData)

anova(Model_MR , Model_MC) ##Analyse du Tableaau ANOVA
```

```
## Analysis of Variance Table
##
## Model 1: Strength ~ Pressure
## Model 2: Strength ~ Pressure * Batch
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     16 13.4861
## 2     12  4.5567  4    8.9294 5.8789 0.007399 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic is 5.8789, and the p-value (Pr(>F)) is 0.007399. The small p-value suggests that there is significant evidence to reject the null hypothesis that Model 1 is as good as Model 2.

Therefore, including the interaction term "Pressure * Batch" significantly improves the fit compared to the model with only "Pressure."

The p-value is less than 0.01, indicating high significance.

Comparaison avec si on excluait C pour Approche Variable Categorielle

```
ModelReduit_Cat = lm(Strength ~Pressure* Batch, data= data_reduit)
summary(ModelReduit_Cat)
```

```
##
## Call:
## lm(formula = Strength ~ Pressure * Batch, data = data_reduit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4833 -0.2333 -0.0875  0.2542  0.6500
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     194.125000   0.993595 195.376 5.27e-16 ***
## Pressure          0.011250   0.001961   5.736 0.000436 ***
## BatchB            0.608333   1.405155   0.433 0.676501
## Pressure:BatchB  -0.003750   0.002774  -1.352 0.213339
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3922 on 8 degrees of freedom
## Multiple R-squared:  0.9078, Adjusted R-squared:  0.8733
## F-statistic: 26.27 on 3 and 8 DF,  p-value: 0.0001708
```

```
ModelReduit_NoCat = lm(Strength ~Pressure, data = data_reduit)
summary(ModelReduit_NoCat)
```

```
##
## Call:
## lm(formula = Strength ~ Pressure, data = data_reduit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.11667 -0.51042  0.07083  0.49896  1.28333
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.944e+02  1.425e+00 136.483  < 2e-16 ***
## Pressure    9.375e-03  2.812e-03   3.334  0.00757 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7953 on 10 degrees of freedom
## Multiple R-squared:  0.5264, Adjusted R-squared:  0.4791
## F-statistic: 11.12 on 1 and 10 DF,  p-value: 0.007566
```

Nous voyons clairement que le meilleur sera lorsqu'on prend l'association avec le Batch A et B.

On observe un R-carre ajuste a 0.87, en comparaison avec 0.48 pour l'autre.

Ceci veut dire que pour ModelReduit_Cat la variation totale est expliquee a 87% par le Model. A noter que pour tous les models encore une fois nous notons que les p-value sont inferieur a 0.05 pour la T-Statistique et F-Statistique qui nous montre que la Statistique est significative.

Conclusion

Apres avoir fais plusieurs analyses, nous sommes arrivés a la conclusion que le meilleur Model est ModelReduit_Cat .

Exercise 6.12: Pinot Noir

Regression Linéaire

```
library(MPV)
```

```
## Warning: package 'MPV' was built under R version 4.2.3
```

```
## Loading required package: lattice
```

```
## Loading required package: KernSmooth
```

```
## KernSmooth 2.23 loaded
## Copyright M. P. Wand 1997-2009
```

```
## Loading required package: randomForest
```

```
## Warning: package 'randomForest' was built under R version 4.2.3
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
##
## Attaching package: 'MPV'
```

```
## The following object is masked from 'package:MASS':
##
##     cement
```

```
data= table.b11
## We will fit  a Model using the Clarity, Aroma, Body, Flavor, Oakiness
library(MPV)
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.3
```

```
## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'tidyr' was built under R version 4.2.3

## Warning: package 'readr' was built under R version 4.2.3

## Warning: package 'purrr' was built under R version 4.2.3

## Warning: package 'stringr' was built under R version 4.2.3

## Warning: package 'forcats' was built under R version 4.2.3

## Warning: package 'lubridate' was built under R version 4.2.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## v readr     2.1.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x randomForest::combine() masks dplyr::combine()
## x dplyr::filter()         masks stats::filter()
## x dplyr::lag()            masks stats::lag()
## x randomForest::margin()  masks ggplot2::margin()
## x dplyr::select()         masks MASS::select()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.2.3

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.2.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.2.3

##
## Attaching package: 'car'
##
## The following object is masked from 'package:purrr':
##
##     some
##
## The following object is masked from 'package:dplyr':
##
##     recode
```

```r
library(MASS)
library(olsrr)
```

```
## Warning: package 'olsrr' was built under R version 4.2.3
```

```
##
## Attaching package: 'olsrr'
##
## The following object is masked from 'package:MPV':
##
##     cement
##
## The following object is masked from 'package:MASS':
##
##     cement
##
## The following object is masked from 'package:datasets':
##
##     rivers
```

```r
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.2.3
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:randomForest':
##
##     combine
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(ggcorrplot)
```

```
## Warning: package 'ggcorrplot' was built under R version 4.2.3
```

```r
# Influence analysis
library(broom)
```

```
## Warning: package 'broom' was built under R version 4.2.3
```

```r
library(dplyr)
library(ggplot2)
model1 <- lm(Quality ~ .-Region, data = data)
summary(model1)
```

```
##
## Call:
## lm(formula = Quality ~ . - Region, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.85552 -0.57448 -0.07092  0.67275  1.68093
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.9969     2.2318   1.791 0.082775 .
## Clarity       2.3395     1.7348   1.349 0.186958
## Aroma         0.4826     0.2724   1.771 0.086058 .
## Body          0.2732     0.3326   0.821 0.417503
## Flavor        1.1683     0.3045   3.837 0.000552 ***
## Oakiness     -0.6840     0.2712  -2.522 0.016833 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.163 on 32 degrees of freedom
## Multiple R-squared:  0.7206, Adjusted R-squared:  0.6769
## F-statistic: 16.51 on 5 and 32 DF,  p-value: 4.703e-08
```

Interprétation: La F-Statistique a une p-value inferieur a 0.05. Ceci nous demontre que la statistique est significative. Dans son ensemble, le model est significatif. On peut rejeter l'hypothese null que tous les Beta sont egale a zero

Pour la T-statistique, le coefficient de Beta pour Flavor et Oakiness sont inferieur a 0.05, ce qui nous montre que la statistique est significative. On peut rejeter l'hypothese null pour les Beta de ces deux predicteurs.

Pour la T-Statistique, le coefficient de Beta pour les predicteurs Clarity, Aroma et Body sont supérieur a 0.05. La statistique n'est donc pas significative. On ne peut pas rejeter l'hypothese nulle que le coefficient du predicteur est egale a zero.

ANOVA

```
modele_anova <- anova(model1)
modele_anova
```

```
## Analysis of Variance Table
##
## Response: Quality
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Clarity    1  0.125   0.125  0.0926 0.7628120
## Aroma      1 77.353  77.353 57.2351 1.286e-08 ***
## Body       1  6.414   6.414  4.7461 0.0368417 *
## Flavor     1 19.050  19.050 14.0953 0.0006946 ***
## Oakiness   1  8.598   8.598  6.3616 0.0168327 *
## Residuals 32 43.248   1.352
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interprétation : Le test ANOVA montre s'il existe une différence statistiquement significative dans le rendement moyen du blé associé aux différents niveaux de nos prédicteurs.

L'analyse de variance (ANOVA) a été réalisée pour évaluer si la qualité du vin Pinot Noir est significativement influencée par différents predicteurs, notamments Clarity, Aroma, Body, Flavor et Oakiness. Les résultats de l'ANOVA sont présentés dans le tableau ci-dessus.

Clarity : La valeur p (Pr(>F)) est de 0,7628120, ce qui est supérieur au seuil de signification de 0,05. Cela suggère qu'il n'y a pas de différence statistiquement significative dans la qualité(variable reponse) moyenne en fonction des différents niveaux de Clarity.

Aroma : La valeur p est très petite (1,286e-08), indiquant une différence hautement significative. Ainsi, il existe des preuves suggérant que la qualité moyenne varie de manière significative selon les différents niveaux de Aroma.

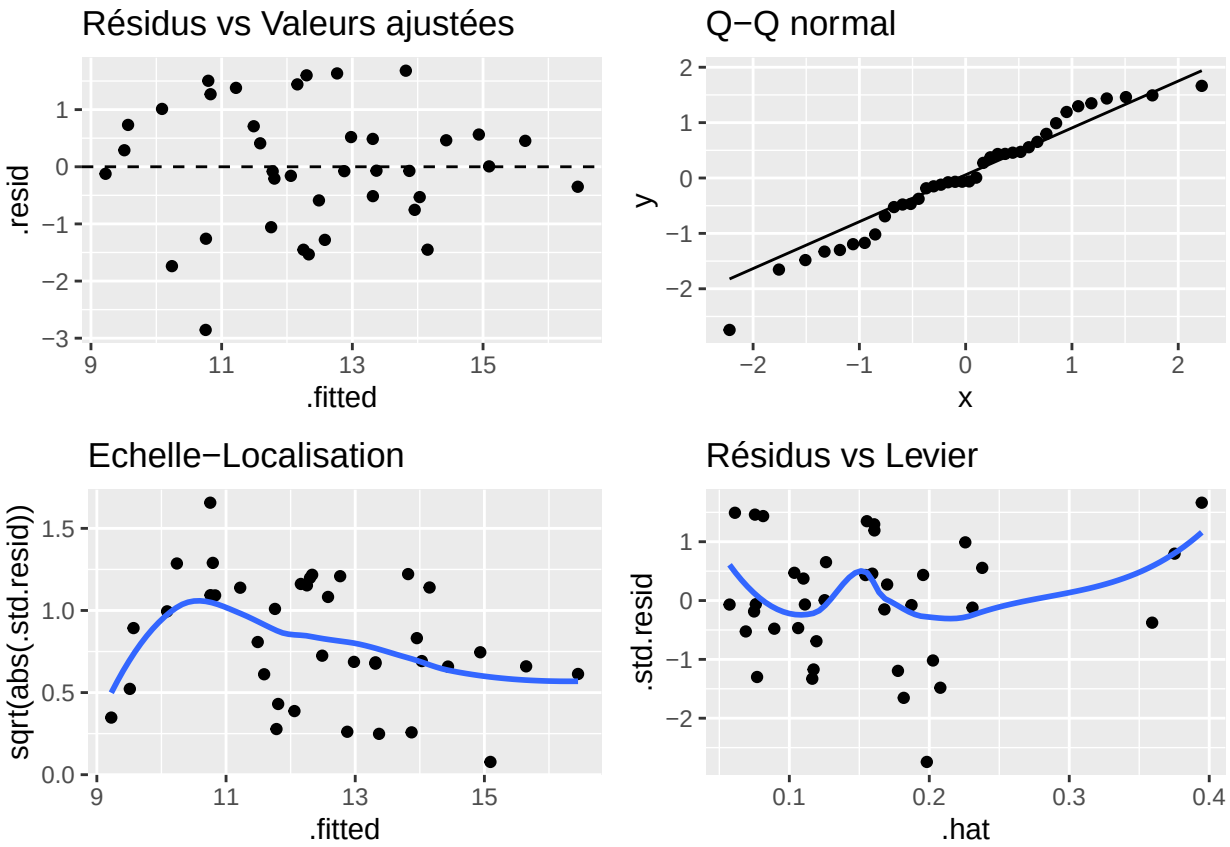
Body : La valeur p est de 0,0368417, ce qui est inférieur à 0,05. Cela suggère une différence statistiquement significative dans la qualité moyenne en fonction des différents niveaux de Body.

Flavor : La valeur p est très petite (0,0006946), indiquant une différence hautement significative. De manière similaire à Aroma, il existe des preuves suggérant que la qualité moyenne varie de manière significative selon les différents niveaux de Flavor.

Oakiness : La valeur p est de 0,0168327, ce qui est inférieur à 0,05. Cela suggère une différence statistiquement significative dans la qualité moyenne en fonction des différents niveaux de Oakiness.

Diagnostics du modèle

```
donnees_diag <- broom::augment(model1)

p1 <- ggplot(donnees_diag, aes(.fitted, .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "dashed") +
  labs(title = "Résidus vs Valeurs ajustées")

p2 <- ggplot(donnees_diag, aes(sample = .std.resid)) +
  stat_qq() +
  stat_qq_line() +
  labs(title = "Q-Q normal")

p3 <- ggplot(donnees_diag, aes(.fitted, sqrt(abs(.std.resid)))) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(title = "Echelle-Localisation")

p4 <- ggplot(donnees_diag, aes(.hat, .std.resid)) +
  geom_point() +
  geom_smooth(se = FALSE) +
  labs(title = "Résidus vs Levier")

grid.arrange(p1, p2, p3, p4, ncol = 2)
```



Interprétation: Puisque nous voyons plus ou moins une distribution uniforme en haut et au bas, on peut voire que la variance des residus est plus ou moins constante.

Puisque pour le graphe Q-Q Normal on a plutot une ligne droite en rejoignant les points, on peut donc dire que les residus suivent une distribution Normale.

Le graphique Echelle-Localisation montre si les résidus sont répartis également sur les gammes de prédicteurs (homoscédasticité).

Le graphique Résidus vs Levier identifie les cas influents qui pourraient affecter notre régression de manière disproportionnée.

Sélection du modèle

```
modele_etape <- stepAIC(model1, direction = "both")
```

```
## Start:  AIC=16.92
## Quality ~ (Clarity + Aroma + Body + Flavor + Oakiness + Region) -
##     Region
##
##             Df Sum of Sq    RSS    AIC
## - Body       1    0.9118 44.160 15.709
## <none>                   43.248 16.916
## - Clarity    1    2.4577 45.706 17.016
## - Aroma      1    4.2397 47.488 18.470
## - Oakiness   1    8.5978 51.846 21.806
```

```
## - Flavor    1   19.8986 63.147 29.299
##
## Step:  AIC=15.71
## Quality ~ Clarity + Aroma + Flavor + Oakiness
##
##            Df Sum of Sq    RSS    AIC
## - Clarity   1    1.6936 45.853 15.139
## <none>                  44.160 15.709
## + Body      1    0.9118 43.248 16.916
## - Aroma     1    5.3545 49.514 18.058
## - Oakiness  1    8.0807 52.241 20.094
## - Flavor    1   27.3280 71.488 32.014
##
## Step:  AIC=15.14
## Quality ~ Aroma + Flavor + Oakiness
##
##            Df Sum of Sq    RSS    AIC
## <none>                  45.853 15.139
## + Clarity   1    1.6936 44.160 15.709
## + Body      1    0.1477 45.706 17.016
## - Aroma     1    6.6026 52.456 18.251
## - Oakiness  1    6.9989 52.852 18.537
## - Flavor    1   25.6888 71.542 30.043
```

```
summary(modele_etape)
```

```
##
## Call:
## lm(formula = Quality ~ Aroma + Flavor + Oakiness, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5707 -0.6256  0.1521  0.6467  1.7741
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.4672     1.3328   4.852 2.67e-05 ***
## Aroma         0.5801     0.2622   2.213 0.033740 *
## Flavor        1.1997     0.2749   4.364 0.000113 ***
## Oakiness     -0.6023     0.2644  -2.278 0.029127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.161 on 34 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.6776
## F-statistic: 26.92 on 3 and 34 DF,  p-value: 4.203e-09
```

Nous pouvons conclure que l'utilisation de `Aroma, Flavor, Oakiness` produit le $R^2 = 68\%$ ajusté que l'utilisation de toutes les colonnes.

Utilisation du modèle réduit (Aroma + Flavor + Oakiness)

```r
# Adjust model
model <- lm(Quality ~ Aroma + Flavor + Oakiness, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = Quality ~ Aroma + Flavor + Oakiness, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5707 -0.6256  0.1521  0.6467  1.7741
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.4672     1.3328   4.852 2.67e-05 ***
## Aroma         0.5801     0.2622   2.213 0.033740 *
## Flavor        1.1997     0.2749   4.364 0.000113 ***
## Oakiness     -0.6023     0.2644  -2.278 0.029127 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.161 on 34 degrees of freedom
## Multiple R-squared:  0.7038, Adjusted R-squared:  0.6776
## F-statistic: 26.92 on 3 and 34 DF,  p-value: 4.203e-09
```

L'analyse d'Influence

Calcul des mesures d'influence

Nous calculerons la **distance de Cook** pour identifier les points influents. En effet, avec le Distance de Cook un point est qualifié d'influent si: `Distance de Cook` : Supérieure à $4/n$, où n est le nombre d'observations.

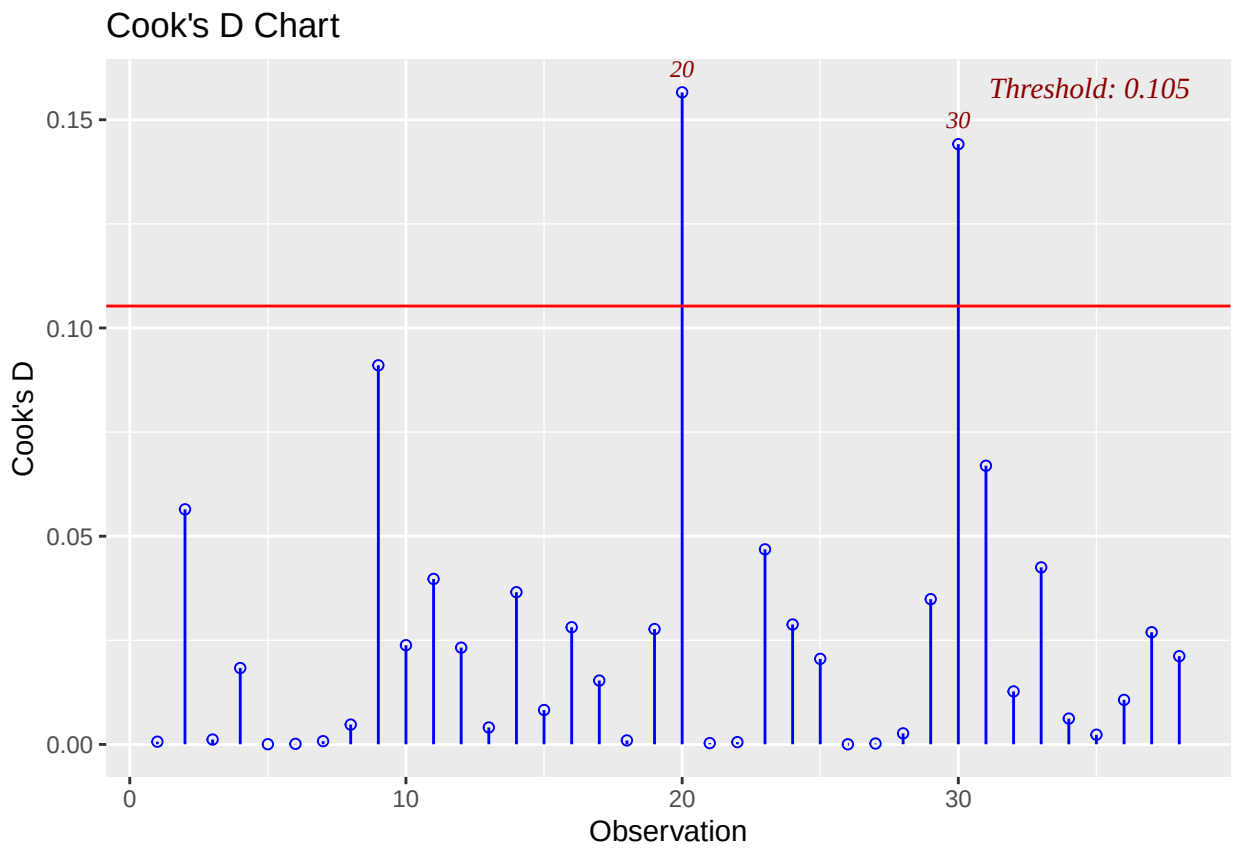
La distance de cook

Étapes du calcul de la distance de Cook :

1. supprimer les observations une à une.
2. refaire le modèle de régression sur les (p-1) observations restantes
3. examiner dans quelle mesure toutes les valeurs ajustées changent lorsque la ième observation est supprimée.

Une observation est considérée comme influente si la valeur absolue de sa valeur DFFITS est supérieure à $4/n$, où n est le nombre d'observations

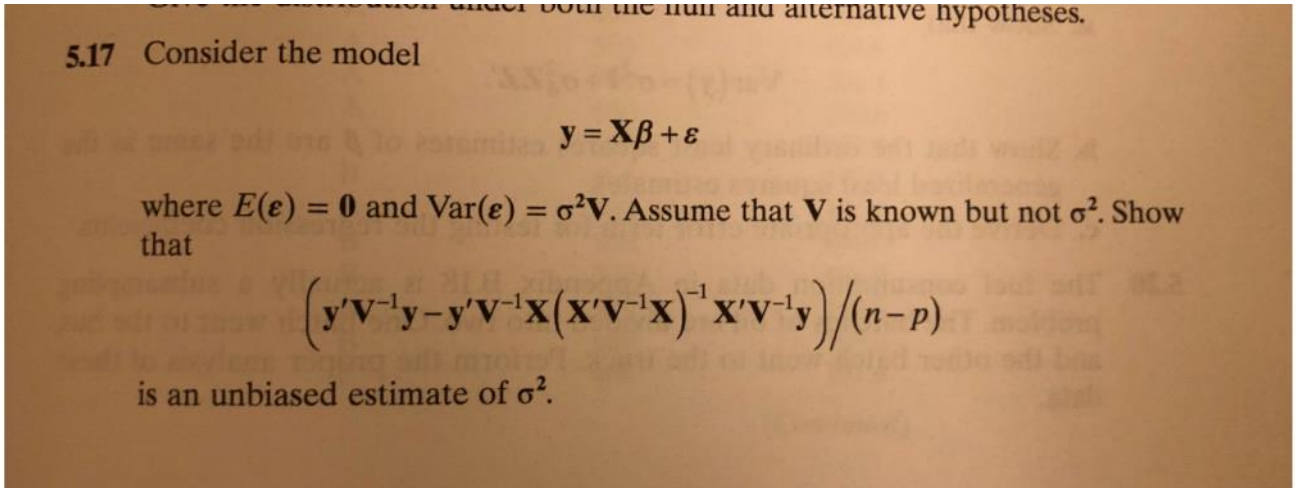
```r
ols_plot_cooksd_chart(model)
```



De ce graphe. On peut voir que pour lè observation de 20 et 30 sont les points influents car ils ont une valeur DFFITS est supérieure à $4/n$

Exercise 5.27

Saturday, November 18, 2023 7:32 PM



...the null and alternative hypotheses.

5.17 Consider the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$. Assume that $\mathbf{V}$ is known but not σ^2. Show that

$$\left(\mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - \mathbf{y}'\mathbf{V}^{-1}\mathbf{X}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\right)\Big/(n-p)$$

is an unbiased estimate of σ^2.

To show that it is an unbiased estimate of σ^2

We need to show

$$E\left[\frac{y'\bar{v}^{-1}y - y'\bar{v}^{-1}x(x'\bar{v}^{-1}x)^{-1}x'\bar{v}^{-1}y}{(n-p)}\right] = y'\bar{v}^{-1}y - y'\bar{v}^{-1}x(x'\bar{v}^{-1}x)^{-1}x'\bar{v}^{-1}y$$

$\frac{1}{(n-p)}$ is a constant

Therefore using the properties of the Expected value

$$E\left[\frac{y'\bar{v}^{-1}y - y'\bar{v}^{-1}x(x'\bar{v}^{-1}x)^{-1}x'\bar{v}^{-1}y}{(n-p)}\right]$$

$$= \frac{1}{n-p}E\left[y'\bar{v}^{-1}y - y'\bar{v}^{-1}x(x'\bar{v}^{-1}x)^{-1}x'\bar{v}^{-1}y\right]$$

Again using properties of Expected value

$E[X+Y] = E[X] + E[Y]$

$$= \frac{1}{n-p}\left(E[y'V^{-1}y] - E[y'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}y]\right)$$

i) We know that $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ where y_i, $i = 1, 2 \dots n$ are the observed response variables

y is a $n \times 1$ Matrix

ii) $y' = [y_1 \; y_2 \dots y_n]$

y' is a $1 \times n$ Matrix (y' is the transpose of y)

iii) From knowing y' is a $1 \times n$ Matrix, we can deduce that V is a $n \times m$ matrix

but here $m = n$, as we can have the inverse of V (Linear Algebra)

$\therefore$ V is an $n \times n$ Matrix

Now lets continue the calculations above using this information

$E[X+Y] = E[X] + E[Y]$

$$= \frac{1}{n-p}\left(E[y'V^{-1}y] - E[y'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}y]\right) \quad \text{—(1)}$$

$y'V^{-1}y$: $[y_1 \; y_2 \dots y_n]$ ($1 \times n$) $\begin{bmatrix} V_{1,1} & V_{1,2} \dots & V_{1,n} \\ \vdots & & \vdots \\ V_{n,1} & V_{n,2} \dots & V_{n,n} \end{bmatrix}$ ($n \times n$) $\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ ($n \times 1$) $= [\mu]$ (one)

$1 \times n \;\; n \times n \;\; n \times 1 = 1 \times 1$ Matrix

$y'V^{-1}y$ evaluates to a 1×1 matrix

1×1 Matrix are considered as scalars

$\therefore E[y'V^{-1}y] = y'V^{-1}y$ (using properties of Expected Value)

$\therefore$ (2) : $= \frac{1}{n-p}\left(y'V^{-1}y - E[y'V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}y]\right)$

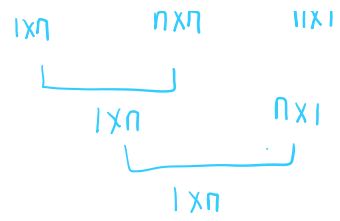
On regarde

$y'V^{-1}X \;\; (X'V^{-1}X)^{-1} \;\; X'V^{-1} \;\; y$

$1 \times n \;\; n \times n \;\; n \times n \;\; n \times n \;\; n \times n \;\; n \times 1$

$1 \times n \;\; n \times n \;\; n \times 1$

$n \times 1$



We can see that: $y'\bar{v}^{-1}X(X'\bar{v}^{-1}X)^{-1}X'\bar{v}^{-1}y$ evaluates to a 1x1 Matrix

$$\frac{1}{n-p}\left(y'\bar{v}^{-1}y - E[y'\bar{v}^{-1}X(X'\bar{v}^{-1}X)^{-1}X'\bar{v}^{-1}y]\right)$$

Again using the properties of the Expected

$E[A]=A$, where A is a 1x1 Matrix considered as a scalar

$$= \frac{1}{n-p}\left(y'\bar{v}^{-1}y - y'\bar{v}^{-1}X(X'\bar{v}^{-1}X)^{-1}X'\bar{v}^{-1}y\right)$$

$$= \frac{y'\bar{v}^{-1}y - y'\bar{v}^{-1}X(X'\bar{v}^{-1}X)^{-1}X'\bar{v}^{-1}y}{n-p}$$

Puisque $E[\hat{\sigma}^2]=\sigma^2$, nous pouvons dire que nous avons un estimateur non-biaisé pour σ^2

Q.E.D