Devoir2-MAT3775

Jonathan Domingue 300246863

2023-10-16

```
## Question1
##Part a)

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.
```

```
library("plot3D")
```

```
## Warning: package 'plot3D' was built under R version 4.2.3
```

```
data1 = table.b1
X2= data1$x2
X7=data1$x7
X8= data1$x8
Y= data1$y


model= lm(y~x2+x7+x8, data=data1)
model
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8, data = data1)
##
## Coefficients:
## (Intercept)           x2           x7           x8
##   -1.808372     0.003598     0.193960    -0.004815
```

```
#Now we will show a summary of the model with the functionality summary in R
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0370 -0.7129 -0.2043  1.1101  3.7049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.808372   7.900859  -0.229 0.820899
## x2           0.003598   0.000695   5.177 2.66e-05 ***
## x7           0.193960   0.088233   2.198 0.037815 *
## x8          -0.004816   0.001277  -3.771 0.000938 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.706 on 24 degrees of freedom
## Multiple R-squared:  0.7863, Adjusted R-squared:  0.7596
## F-statistic: 29.44 on 3 and 24 DF,  p-value: 3.273e-08
```

Le model de regression linéaire est donc le suivant:

y= -1.8 +0.00360X2 + 0.194X7 -0.00482X8

Les coefficients associés a la variable x7 (percent rushing) est la plus élevée. Cependant nous notons que pour toutes les variables x2, x7 et x8, nous avons des plutot petites valeurs pour les coefficients, ce qui nous montre qu'ils n'ont pas une tres grande importance

Maintenant parlant de la signification de la Régression Le coefficient de la variable x2 $\hat{\beta}_1$ = -1.81

Le coefficient de la variable x8

$\hat{\beta}_3$ = -0.00482

sont statistiquement significatifs. i.e. ils ont une valeur de p <0.001

Cependant le coéfficient de $\hat{\beta}_0$ $\hat{\beta}_2$ ne sont pas statistiquement significatif car ils ont une valeur de p plus grande que 0.001

```r
##Part b)
##ANOVA Table

library(MPV)
library("plot3D")
data1 = table.b1
X2= data1$x2
X7=data1$x7
X8= data1$x8
Y= data1$y

model= lm(y~x2+x7+x8, data=data1)

anova(model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x2         1  76.193  76.193  26.172 3.100e-05 ***
## x7         1 139.501 139.501  47.918 3.698e-07 ***
## x8         1  41.400  41.400  14.221 0.0009378 ***
## Residuals 24  69.870   2.911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Avec un tableau ANOVA on peut tester la signification de la regression en utilisant la valeur F

Dans ce cas, le modele a une performance car le F-Statistic est de 29.44 dans son ensemble avec une valeur de p<0.001 En outre, le Adjusted R-squared value is R^2= 0.760 Ceci indique que le modèle de régression explique 76,0% de la variabilité de la mesure des résultats, ce qui est bon.

```r
## Part c)

library(MPV)
library("plot3D")
data1 = table.b1
X2= data1$x2
X7=data1$x7
X8= data1$x8
Y= data1$y

Full= lm(y~x2+x7+x8 , data=data1)

x2_model = lm(y~x2 , data=data1)

x7_model= lm( y~x7, data=data1 )

x8_model = lm(y~x8 , data=data1)

##Nous allons trouver les coefficients pour le model reduit avec x2
summary(x2_model)
```

```
##
```

```
## Call:
## lm(formula = y ~ x2, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7205 -2.0047  0.0448  2.2396  5.3491
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.195276   2.614037  -0.075  0.94102
## x2           0.003366   0.001197   2.811  0.00927 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.106 on 26 degrees of freedom
## Multiple R-squared:  0.233,  Adjusted R-squared:  0.2035
## F-statistic:   7.9 on 1 and 26 DF,  p-value: 0.009272
```

```
## Tableau anova pour modele reduit de RLS avec x2
anova(x2_model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## x2         1  76.193  76.193  7.8998 0.009272 **
## Residuals 26 250.771   9.645
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##Nous allons trouver les coefficients pour le model reduit avec x7
summary(x7_model)
```

```
##
## Call:
## lm(formula = y ~ x7, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3245 -1.6202  0.1368  2.0329  5.1474
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13.5135     6.1983  -2.180  0.03849 *
## x7            0.3521     0.1061   3.317  0.00269 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.972 on 26 degrees of freedom
## Multiple R-squared:  0.2974, Adjusted R-squared:  0.2704
## F-statistic: 11.01 on 1 and 26 DF,  p-value: 0.002689
```

```r
## Tableau ANOVA pour le model reduit de RLS avec x7
anova(x7_model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## x7         1  97.238  97.238  11.005 0.002689 **
## Residuals 26 229.726   8.836
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
##Nous allons trouver les coefficients pour le model reduit avec x8
summary(x8_model)
```

```
##
## Call:
## lm(formula = y ~ x8, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.804 -1.591 -0.647  2.032  4.580
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.788251   2.696233   8.081 1.46e-08 ***
## x8          -0.007025   0.001260  -5.577 7.38e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.393 on 26 degrees of freedom
## Multiple R-squared:  0.5447, Adjusted R-squared:  0.5272
## F-statistic:  31.1 on 1 and 26 DF,  p-value: 7.381e-06
```

```r
##Tableau ANOVA pour le model reduit de RLS avec x8
anova(x8_model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x8         1 178.09 178.092  31.103 7.381e-06 ***
## Residuals 26 148.87   5.726
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Les modeles de régression lineaire simples sont donc les suivants:

y= -0.195 + 0.003366x2

Le coeff de la variable x2 $\hat{\beta}_1 = 0.003366$ ne sont pas autant statistiquement significatifs car p >0.001 pour la statistique T Nous avons une valeur de p=0.00927

On va maintenant faire le calcul pour la variable x7:

y= -13.5 + 0.352x7
Le coefficient de la variable x7 n'est pas autant statistiquement significatif car p>0.001 pour la statistique T Nous avons une valeur p=0.00269 qui est legerement proche cependant de 0.001

On va maintenant faire le calcul pour la variable x8:

y= 21.79 + -0.007025x8 Le coefficient de la variable x8 est tres statistiquement significatif car p<0.001 pour la statistique T Nous avons une valeur p=7.381 e-6 qui est vraiment tres petite

x2 is not statistically significant (p > 0.001). x7 is marginally significant (p ~0.00269), which is close to 0.001. x8 is highly statistically significant (p < 0.001).

```
## Partie d)
##Des calculs fait avec R plus haut

multiple_r_squared <- 0.7863
adjusted_r_squared <- 0.7596

# Display the results
cat("Multiple R-squared:", multiple_r_squared, ", Adjusted R-squared:", adjusted_r_squared)
```

```
## Multiple R-squared: 0.7863 , Adjusted R-squared: 0.7596
```

```
## Partie e)
## We will simply create a full and a reduced Model
## Le premier model aura toutes les variables x2,x7 et x8
## Le model reduit aura les variables x2 et x8
##De cela nous allons pouvoire determiner la contribution de x7 dans le model

library(MPV)
library("plot3D")
data1 = table.b1
X2= data1$x2
X7=data1$x7
X8= data1$x8
Y= data1$y

Full= lm(y~x2+x7+x8 , data=data1)
Reduced= lm(y~x2+x8 , data= data1)

anova(Reduced, Full)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x2 + x8
## Model 2: y ~ x2 + x7 + x8
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     25 83.938
## 2     24 69.870  1    14.068 4.8324 0.03782 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

COMPARAISON DES DEUX MODEL REDUIT ET FULL

Model 1: SSE= 83.938 Model 2: SSE= 69.870

Dans ce cas, la statistique F est de 4.8324.

Pr(>F) (p-valeur) : Il s'agit de la p-valeur associée à la statistique F. Elle indique la probabilité d'obtenir une statistique F aussi extrême que celle observée si les deux modèles étaient équivalents (c'est-à-dire si l'ajout de x7 au Modèle 2 n'apportait pas d'amélioration significative par rapport au Modèle 1). La p-valeur est de 0.03782.

La p-valeur de 0.03782 est inférieure à un seuil de signification de 0.05 (indiqué par '*'), ce qui signifie que vous pouvez rejeter l'hypothèse nulle que les deux modèles sont équivalents. En d'autres termes, l'ajout de la variable x7 au Modèle 2 améliore significativement l'ajustement du modèle par rapport au Modèle 1. Par conséquent, x7 a une contribution significative à l'explication de la variance de la variable dépendante y.

Problem 3.3

```
## On va importer les donnes et ensuite faire le calcul
library(MPV)
data1 = table.b1
X2= data1$x2
X7=data1$x7
X8= data1$x8
Y= data1$y

Full_Model = lm(y~x2+x7+x8 , data=data1) ## Model complet
X7_Model = lm(y~x7 , data=data1 ) ## Model Lineaire Simple avec X7

## En utilisant le Model Complet avec une IC a 95%
confint(Full_Model, level=0.95)
```

a)Trouver un IC a 95% Beta7

```
##                     2.5 %       97.5 %
## (Intercept) -18.114944410 14.498200293
## x2            0.002163664  0.005032477
## x7            0.011855322  0.376065098
## x8           -0.007451027 -0.002179961
```

```
##

##Maintenant pour le model reduit simple avec uniquement la variable x7 IC de 95%
confint(X7_Model , level=0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) -26.254368 -0.7726759
## x7            0.133937  0.5702863
```

```r
## On va importer les donnes et ensuite faire le calcul
library(MPV)
data1 = table.b1

Full_Model = lm(y~x2+x7+x8 , data=data1) ## Model complet


new_data <- data.frame(x2 = 2300, x7 = 56.0, x8 = 2100)

predict(Full_Model , new_data , interval="confidence")
```

b)Trouvez un IC de 95 % pour le nombre moyen de matchs gagnés par une équipe lorsque x_2 = 2300, x_7=56,0 et x_8=2100

```
##        fit      lwr      upr
## 1 7.216424 6.436203 7.996645
```

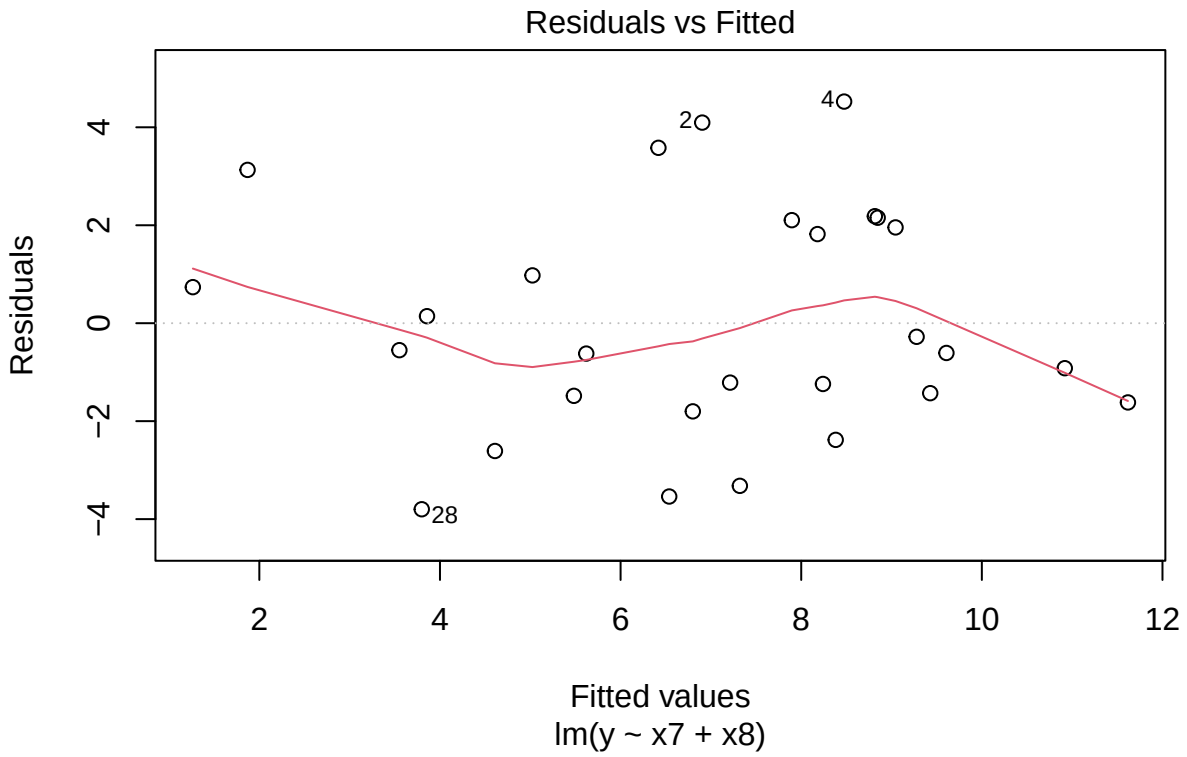
Problem 3.4: Suite

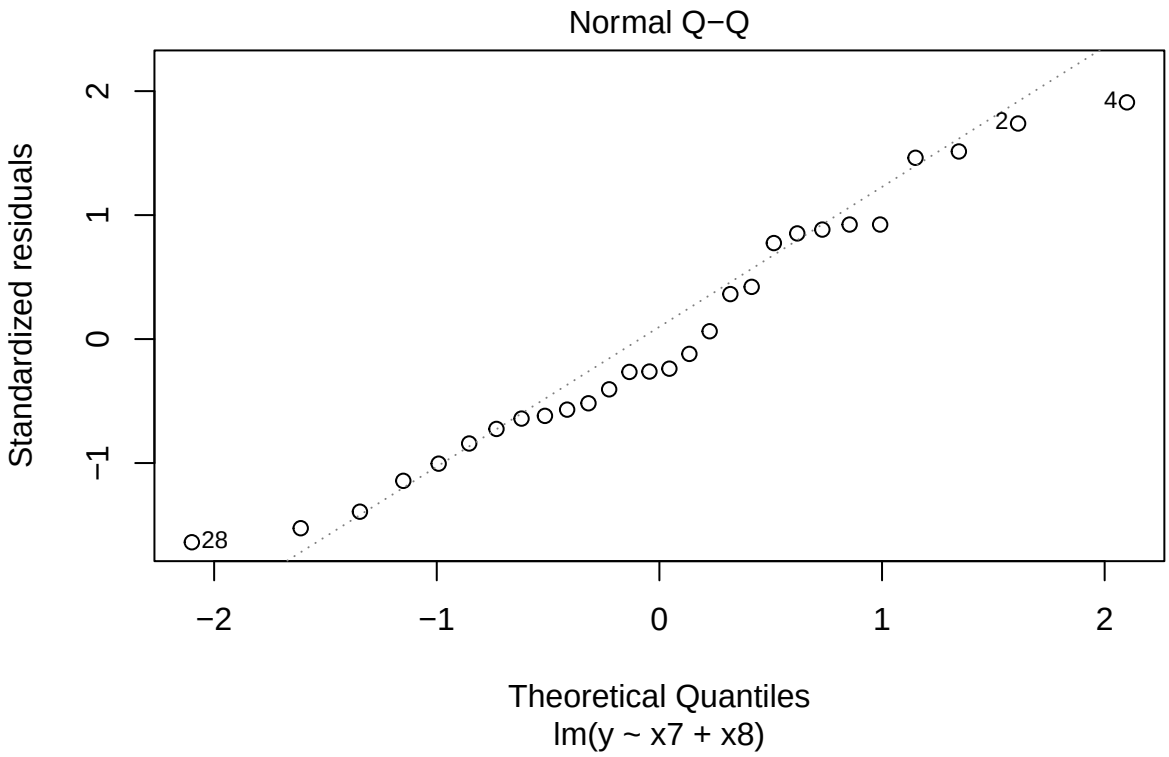
```r
library(MPV)
data1 = table.b1

Reduced_Model = lm(y~x7+x8 , data=data1) ## Model Réduit avec x7 et x8 comme regresseurs

summary(Reduced_Model)
```

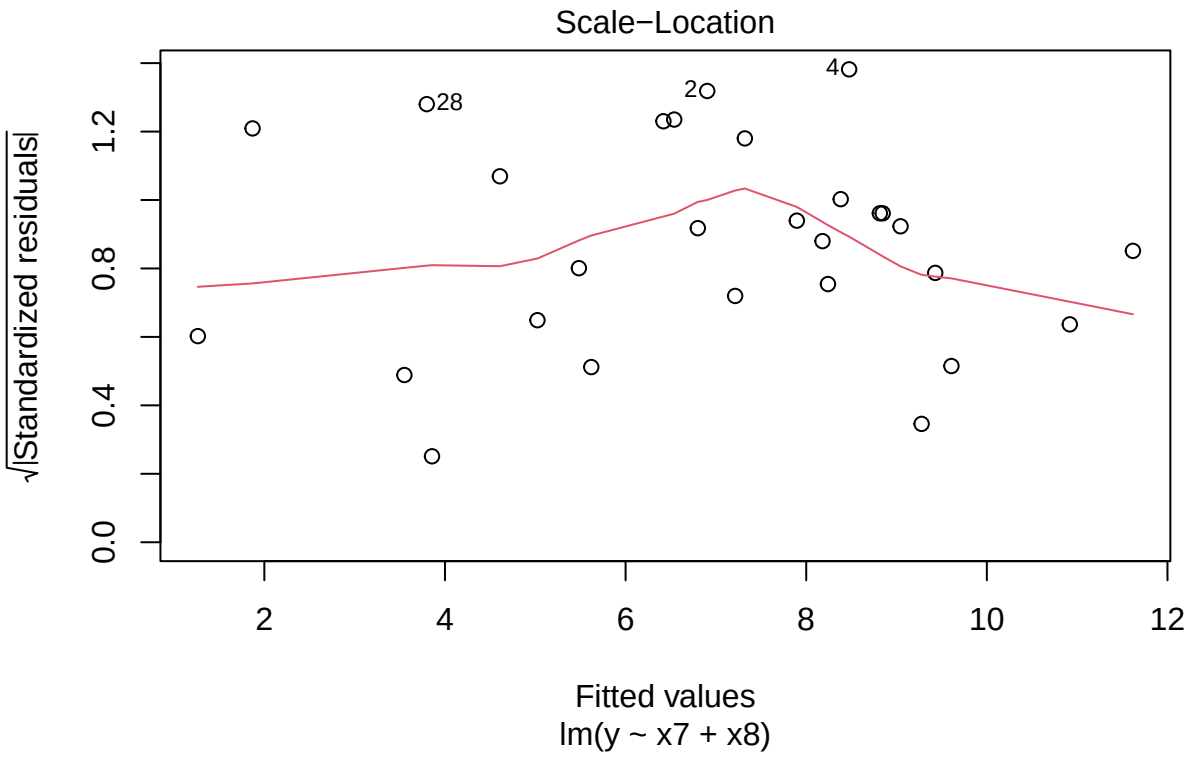
a)Ajuster Model Réduit avec x_7 et x_8 comme regresseurs

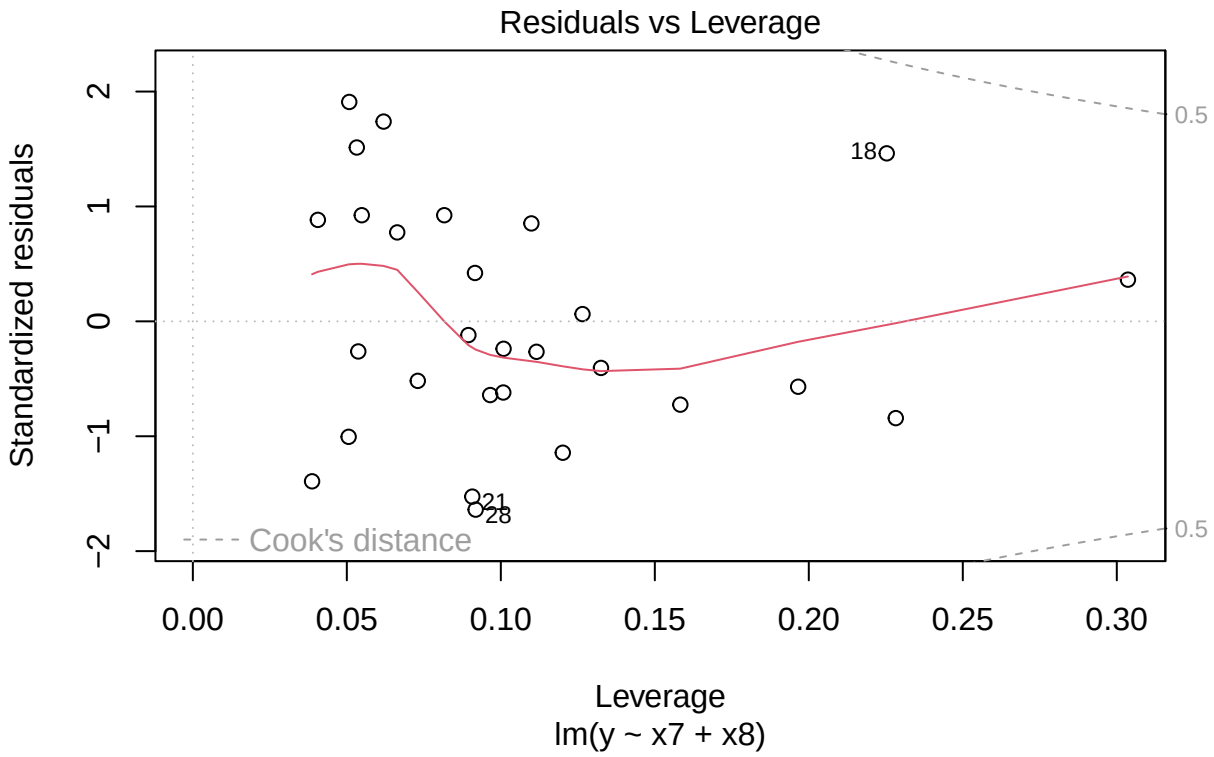
```
##
## Call:
## lm(formula = y ~ x7 + x8, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7985 -1.5166 -0.5792  1.9927  4.5248
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.944319   9.862484   1.819  0.08084 .
## x7           0.048371   0.119219   0.406  0.68839
## x8          -0.006537   0.001758  -3.719  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.432 on 25 degrees of freedom
## Multiple R-squared:  0.5477, Adjusted R-squared:  0.5115
## F-statistic: 15.13 on 2 and 25 DF,  p-value: 4.935e-05
```

```
plot(Reduced_Model)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x7 + x8)
28
2
4
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
28
2
4
2
4
6
8
10
12
Fitted values
lm(y ~ x7 + x8)



a) Tester la signification de la Régression

On peut utiliser les résultats obtenus plus haut pour tester la signification de la Régression

Commencons:

Les valeurs p associées à chaque coefficient pour stat T

Dans notre cas, “x8” a une valeur p de 0.00102, ce qui est généralement considéré comme significatif. Cela suggère que “x8” a un effet significatif sur la variable “y”.

Le F-Stat et sa valeur p-associée

La p-valeur de l’ANOVA globale est de 4.935e-05, ce qui est très faible, indiquant que le modèle global est significatif.

La valeur R-carré ajustée.

Une valeur de 0.5115 indique que le modèle explique une grande partie de la variance dans la variable dépendante.

En fonction de ces résultats, vous pouvez conclure que la régression est significative et que les prédicteurs “x7” et “x8” ont des effets significatifs sur la variable “y”

```r
##Des calculs fait avec R plus haut

multiple_r_squared <- 0.5477
adjusted_r_squared <- 0.5115

# Display the results
cat("Multiple R-squared:", multiple_r_squared, ", Adjusted R-squared:", adjusted_r_squared)
```

```
## Multiple R-squared: 0.5477 , Adjusted R-squared: 0.5115
```

Interpretation of Multiple R-Squared

In the new model (using only x7 and x8), R^2 is 0.5477, indicating that approximately 54.77% of the variance in the dependent variable is explained by x7 and x8. In the previous model (with x2, x7, and x8), R^2 is higher at 0.7863, showing that about 78.63% of the variance in the dependent variable is explained by x2, x7, and x8. The previous model has a higher R^2, which suggests a better overall fit.

Interpretation of Adjusted R-Squared

Adjusted R-squared takes into account the number of predictors in the model and penalizes the inclusion of unnecessary predictors. In the new model, the adjusted R2 is 0.5115, reflecting the adjusted fit of the model with x7 and x8. In the previous model, the adjusted R2 is higher at 0.7596, indicating that the inclusion of x2, x7, and x8 in the model provides a better adjusted fit.

CONCLUSION

The previous model (with x2, x7, and x8) has higher R-squared and adjusted R-squared values, indicating a better fit and explaining more of the variance in the dependent variable compared to the new model, which includes only x7 and x8. The inclusion of x2 in the previous model seems to improve its explanatory power significantly.

) Calculez un IC de 95 % pour beta7. Trouvez également un IC de 95 % pour le nombre myen de matchs gagnés par une équipe lorsque x_7 = 56,0 et x_8 = 2100. Comparez les longueurs de ces IC aux longueurs des IC correspondants du problème 3.3.

```r
## On va importer les donnes et ensuite faire le calcul
library(MPV)
data1 = table.b1

Full_Model = lm(y~x7+x8 , data=data1) ## Model complet


new_data <- data.frame(x7 = 56, x8 = 2100)

predict(Full_Model , new_data , interval="confidence")
```

```
##        fit      lwr      upr
## 1 6.926243 5.828643 8.023842
```

Interprétation

Apres en utilisant x7=56 et x8=2100

```
   fit      lwr      upr
```

1 6.926243 5.828643 8.023842

Longeur= 2.195199

Avant en utilisant x2 = 2300, x7 = 56.0, x8 = 2100

```
   fit      lwr      upr
```

1 7.216424 6.436203 7.996645

Longeur de IC= 1.560442

d) Que pouvons-nous conclure sur l'omission d'un regresseur important d'un Model?

We can observe that when we omit an important regressor in the model, the range of the confidence interval becomes more narrow

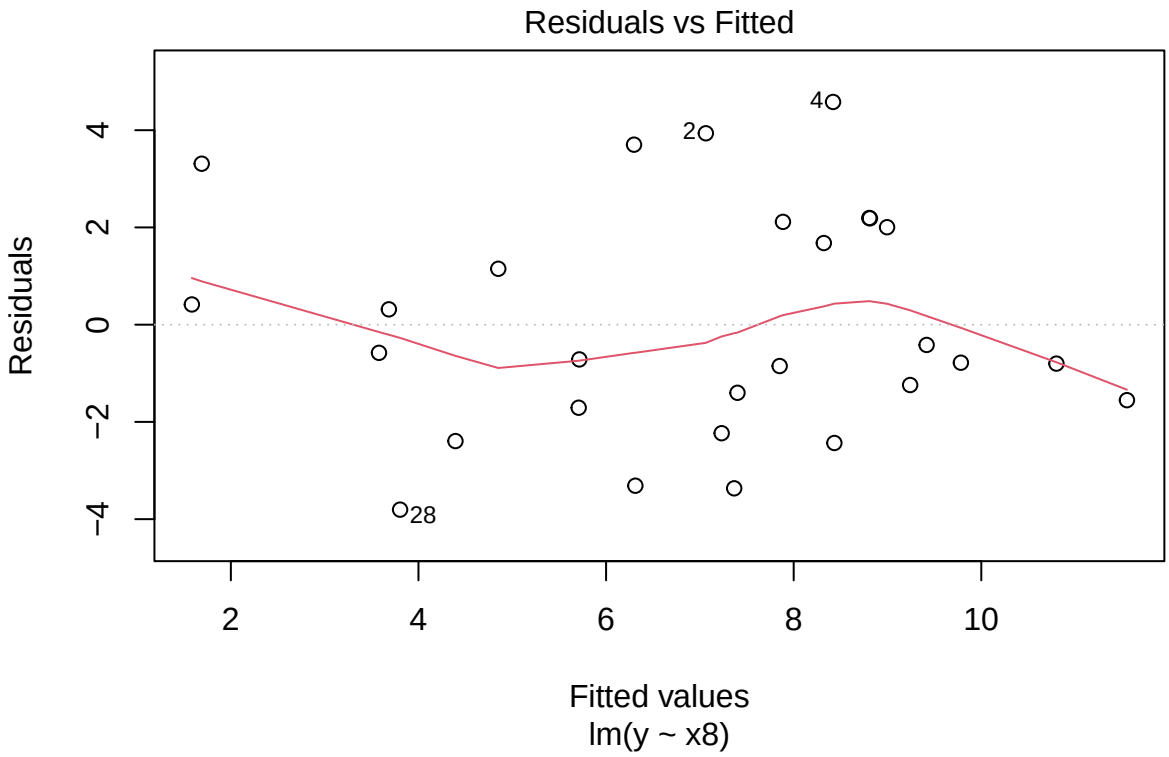
Cela suggère que l'estimation "Avant" est plus précise (car l'IC est plus étroit) et que le nombre moyen de matchs gagnés est légèrement plus élevé dans la situation "Avant".

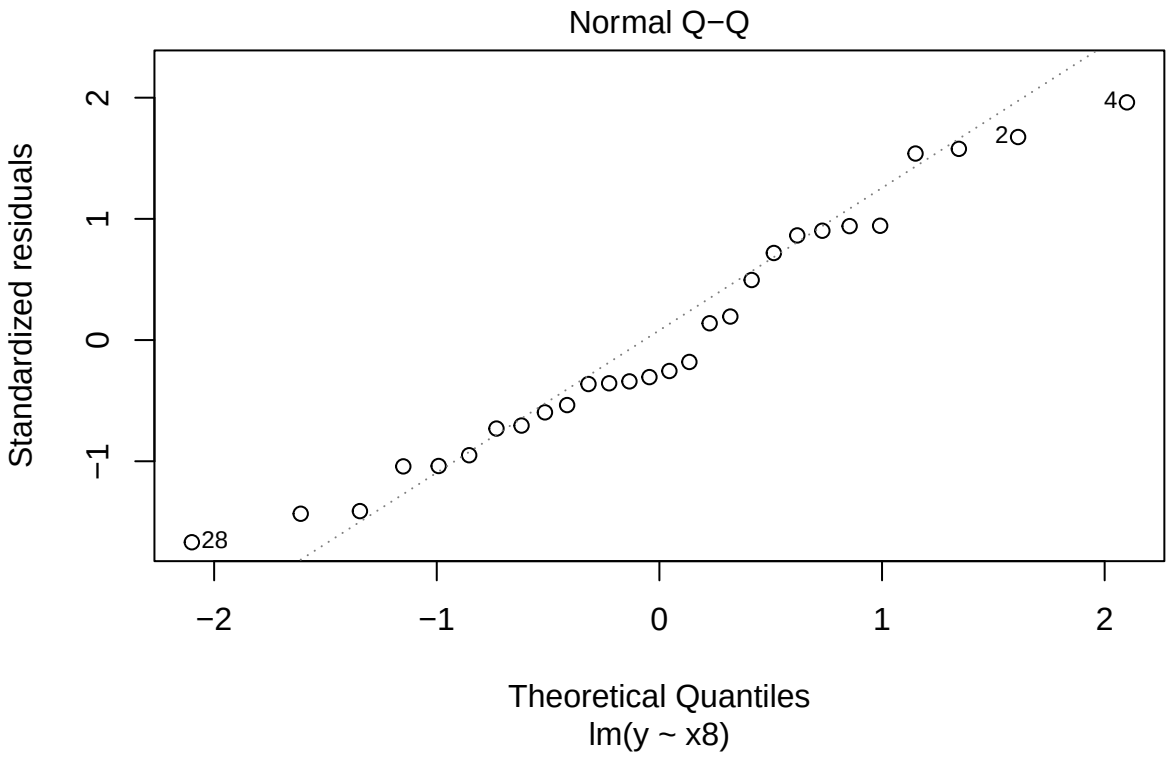
Probleme 4.1: Ligue National de Football du Devoire 1

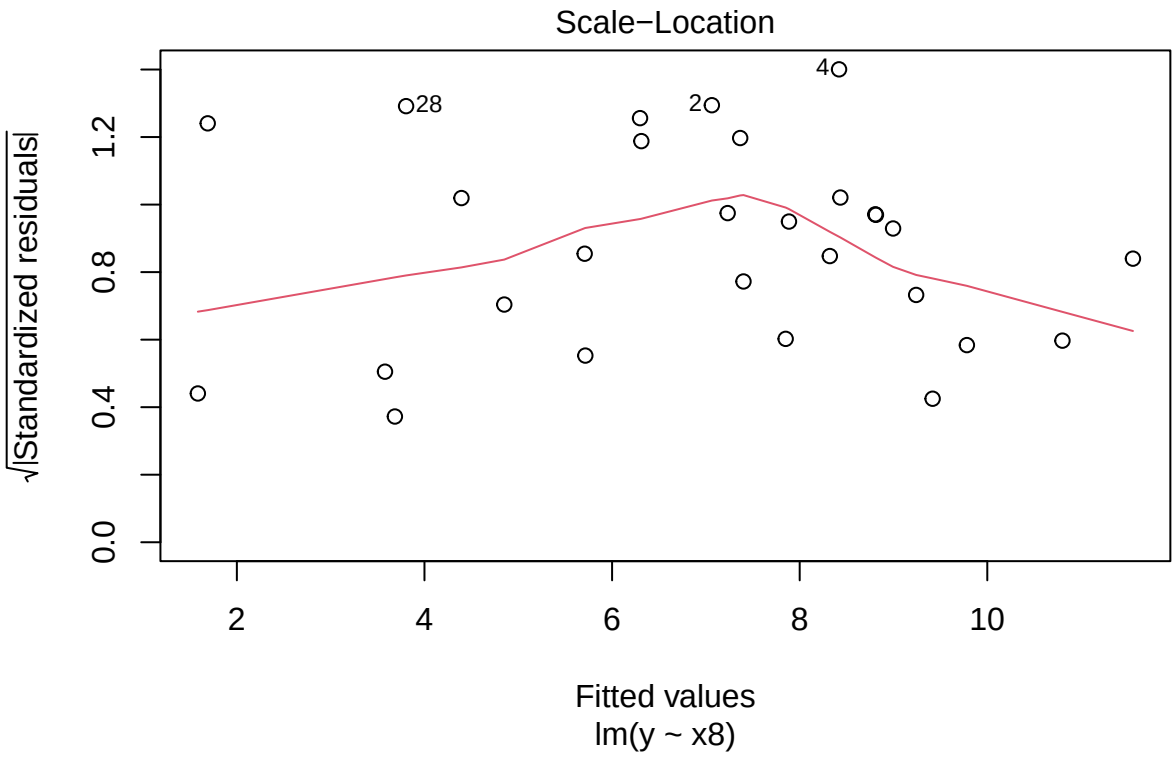
(a) Construisez un diagramme de probabilité normale des résidus. L'hypothèse de normalité semble-t-elle poser un problème ?

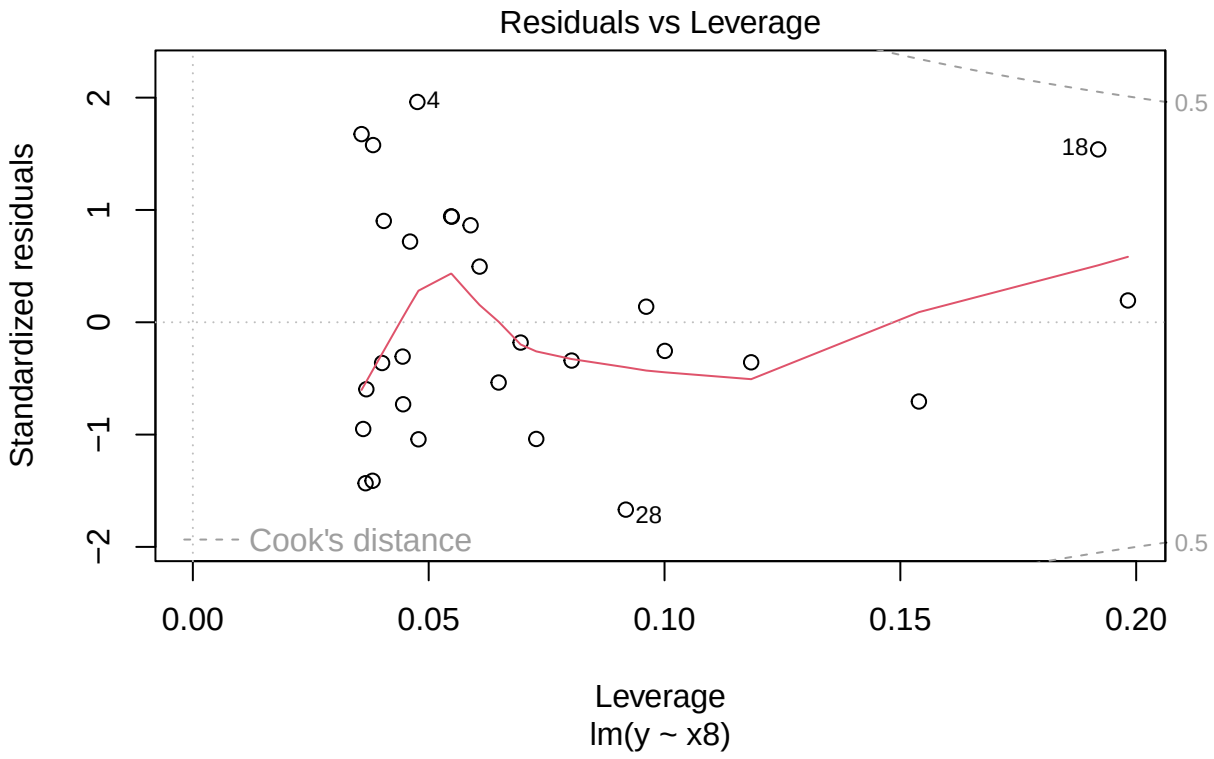
(b) Construire et interpréter un graphique des résidus par rapport à la réponse prévue

```r
## On va importer les donnes et ensuite faire le calcul
library(MPV)
data1 = table.b1

Model= lm(y~x8 , data=data1 )
plot(Model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x8)
4
2
28
−4
−2
0
2
4
2
4
6
8
10


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x8)
28
2
4
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
28
2
4
2
4
6
8
10
Fitted values
lm(y ~ x8)



Interprétation

a)Oui puisque les points suivent plutot la ligne droite pour le graphique de standardized residuals vs Theoretical Quantiles, on peut conclure que oui en effet on peut supposer la normalite pour les résidus

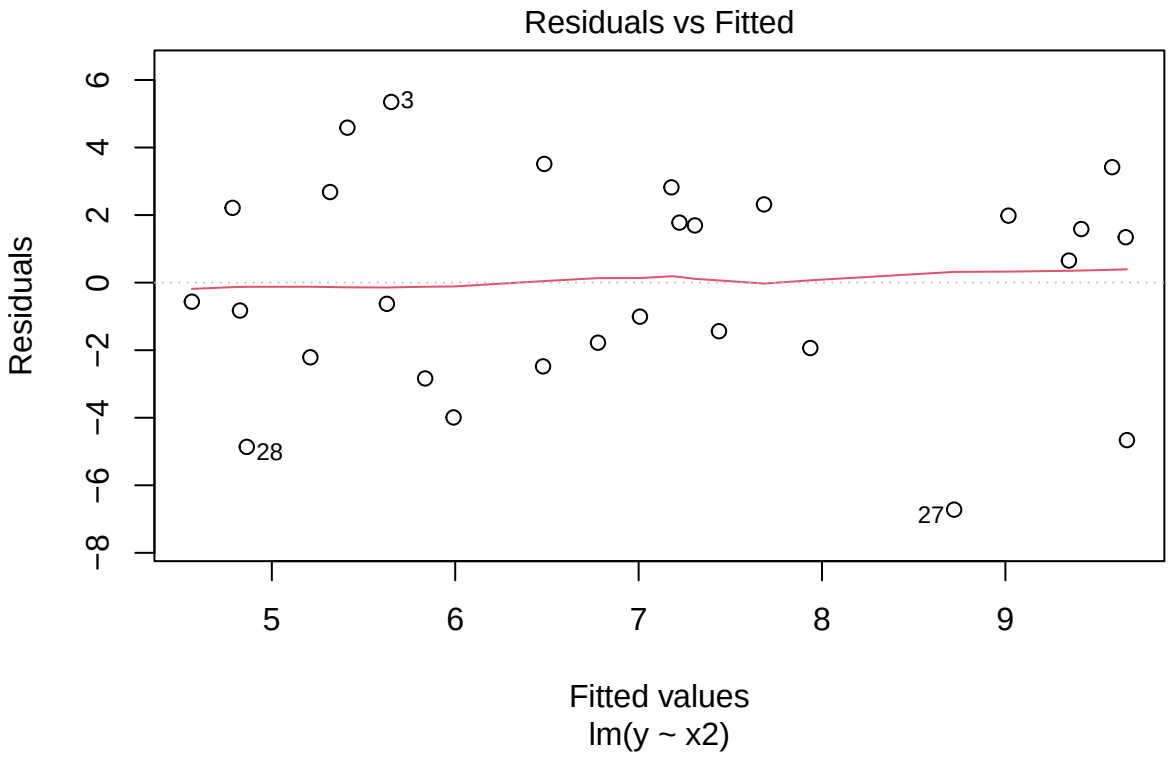
b) Pour interpréter on va examiner le graphique de Residuals vs Fitted Values

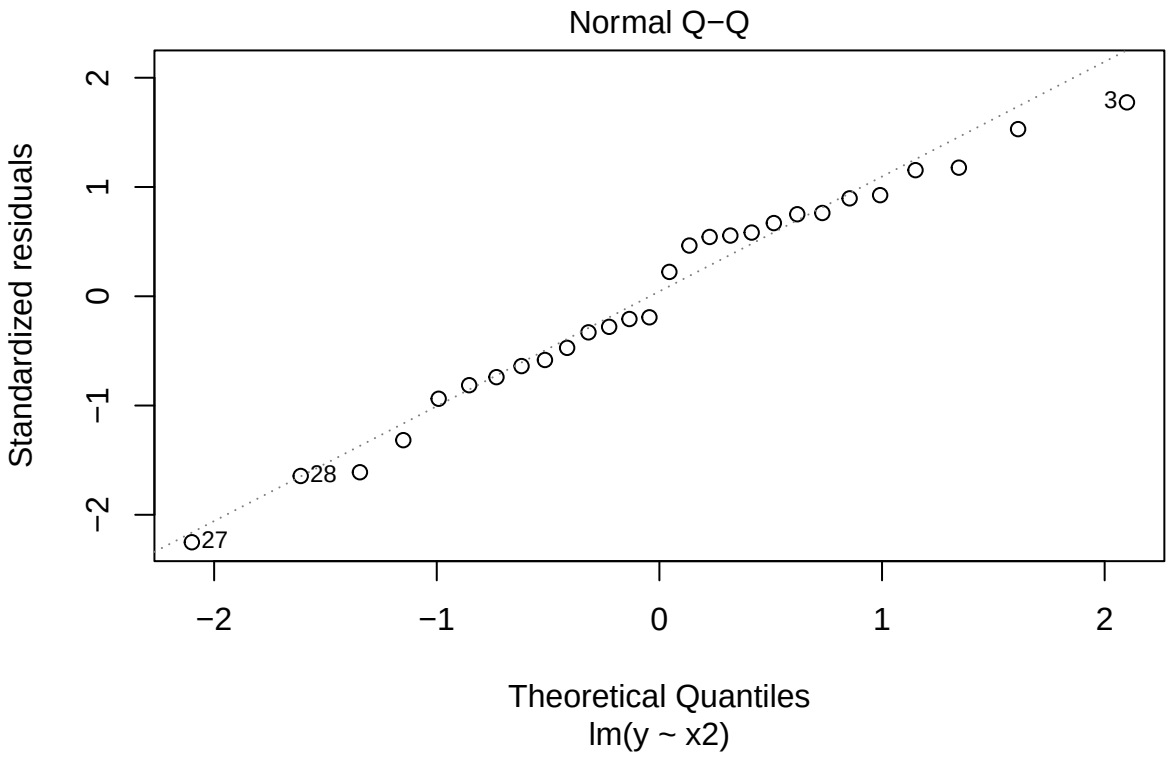
L'objectif de ce graphique est de vérifier si les résidus ont une variance constante le long de la plage des valeurs prédites. Dans notre cas, nous pouvons observer que les points ne sont pas vraiment plus ou moins repartis autour de zero et semblent legerement changer pour les differentes fitted values

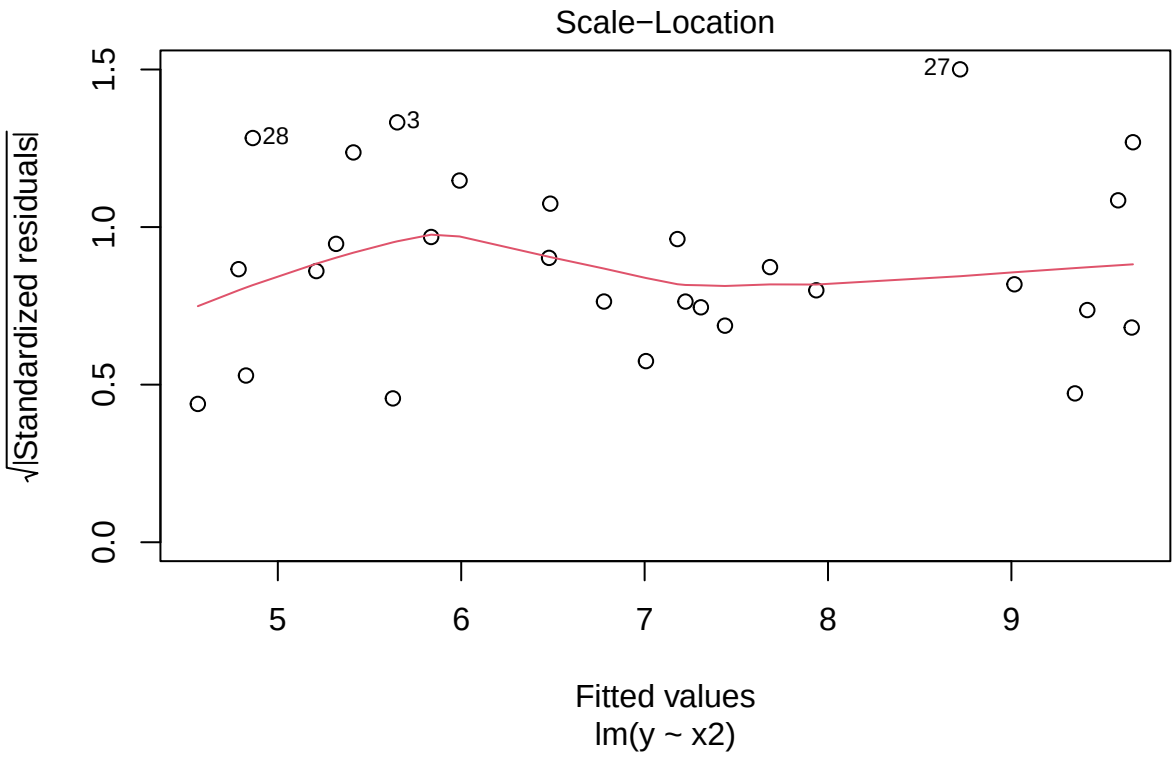
Cela nous fait savoir que la variance des residus change lorsque les fitted values changent. Il ya un espece genre de graphe qui baisse et qui augmente qu'on peut observer

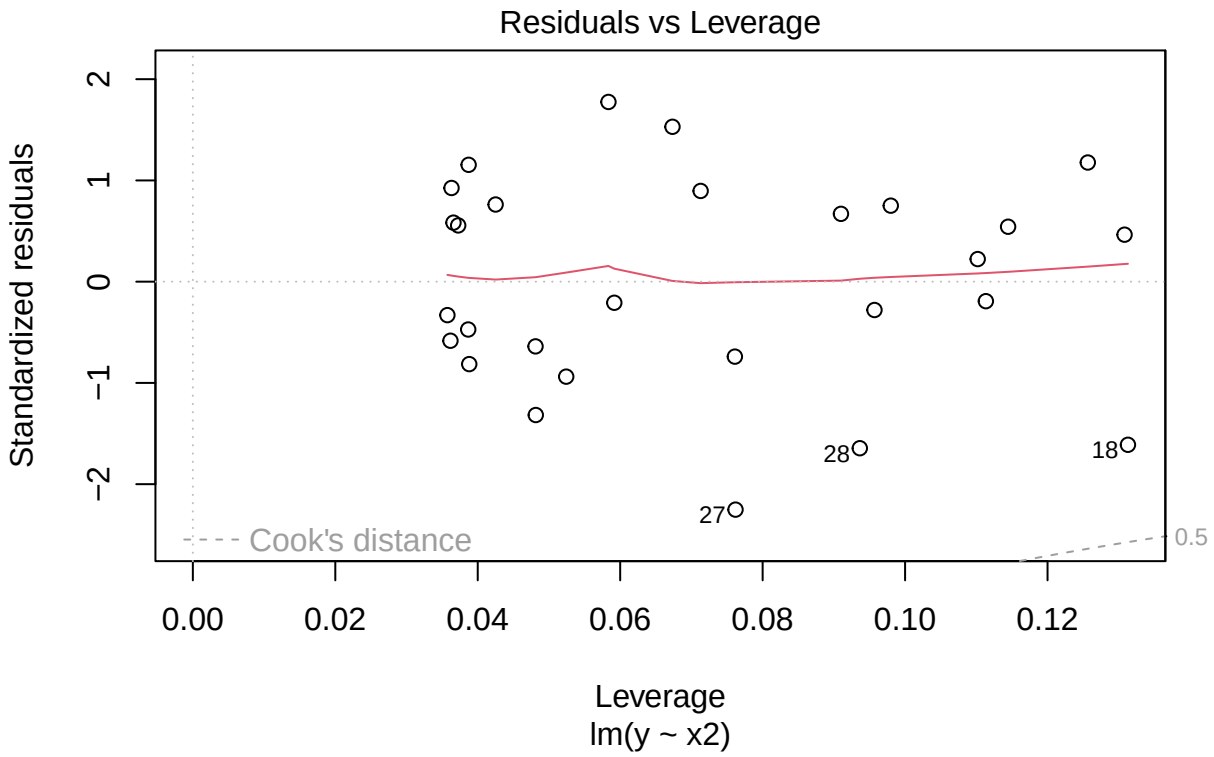
c) Maintenant on va tracer les Residus par rapport a x2

```
## On va importer les donnes et ensuite faire le calcul
library(MPV)
data1 = table.b1

Model= lm(y~x2, data=data1)
plot(Model)
```


Residuals vs Fitted
Residuals
Fitted values
lm(y ~ x2)
3
28
27
6
4
2
0
−2
−4
−6
−8
5
6
7
8
9


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(y ~ x2)
28
27
3
−2
−1
0
1
2


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
28
3
27
5
6
7
8
9
Fitted values
lm(y ~ x2)



Interprétation

Lorsqu'on observe le graphique de Residuals vs Fitted Values avec la variable x2 on peut clairement observer que nous avons un bien meilleur graphique que avec x8.

La variance des résidus pour les differentes Fitted Values sont autour de zero et sont presque toujours constantes qui est excellent!

Oui ce graphique indique en effet que le model sera amélioré avec l'ajout de x8