Impact du CO2 sur la température de la planète

Pour de l'aide sur R Markdown on peut aller sur

http://rmarkdown.rstudio.com

https://lms.fun-mooc.fr/c4x/UPSUD/42001S02/asset/RMarkdown.html

Introduction

On s'intéresse à la dépendance entre la température de surface de la planète et les émissions de C02 et leurs évolutions dans le temps. L'analyse de la température de surface GISS ver. 4 (GISTEMP v4) est une estimation du changement global de la température de surface. Les données proviennent de NOAA GHCN v4 (stations météorologiques) et ERSST v5 (zones océaniques). Plus de détails sont accessibles sur ce lien https://data.giss.nasa.gov/gistemp/. Concernant le CO2, il est mesuré sur le Mauna Loa (sommet de l'archipel d'Hawai) depuis la fin des annnées 50. Détails sur https://gml.noaa.gov/ccgg/trends/

Le fichier "Data/climat_CO2.txt" contient 3 colonnes :

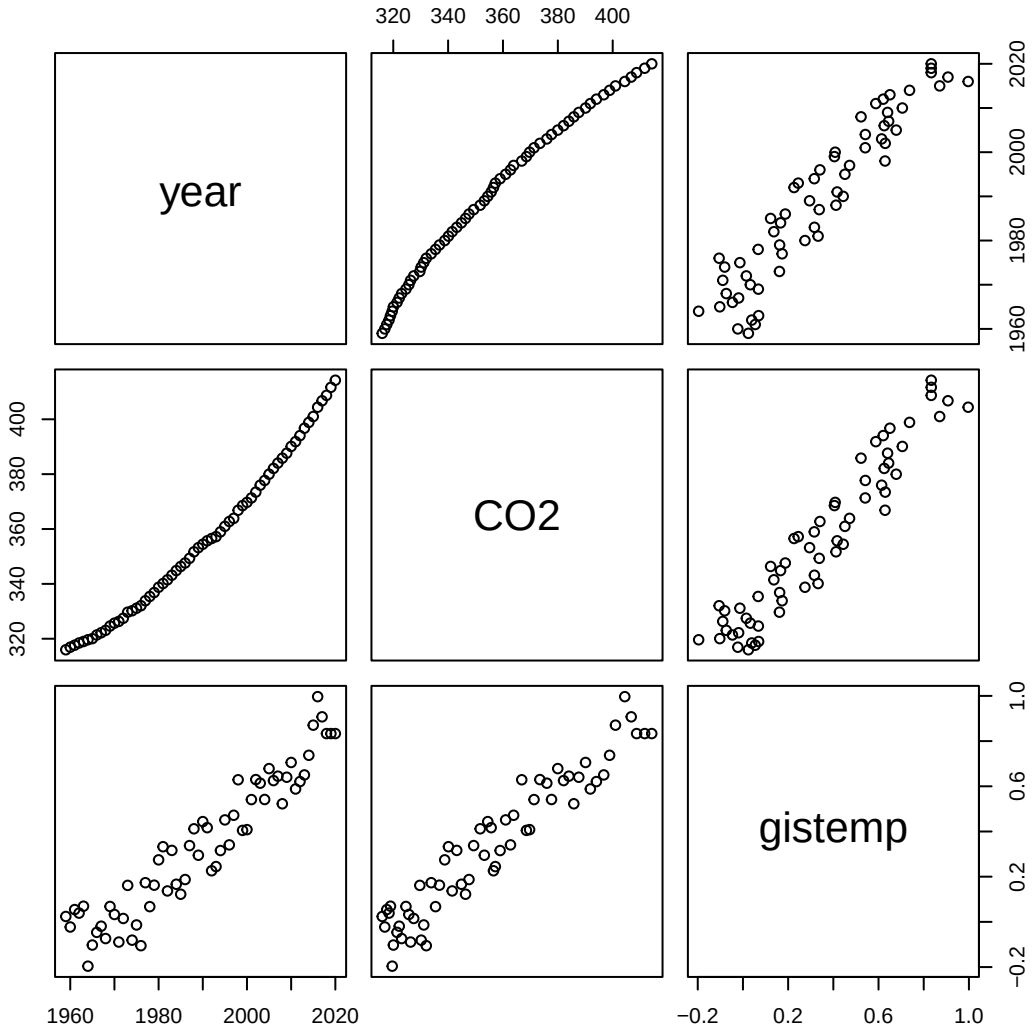
- **year** : années entre 1959 et 2020
- **CO2** : CO2 exprimé en fraction molaire dans l'air sec, micromol/mol, abrégé en ppm. Voir www.esrl.noaa.gov/gmd/ccgg/trends/ pour plus de détails.
- **gistemp** : anomalies de température en dégrés Celsius. Il s'agit des écarts entre la température annuelle et la température moyenne de la période 1951-1980.

1 - On commence par lire et visualiser les données du fichier

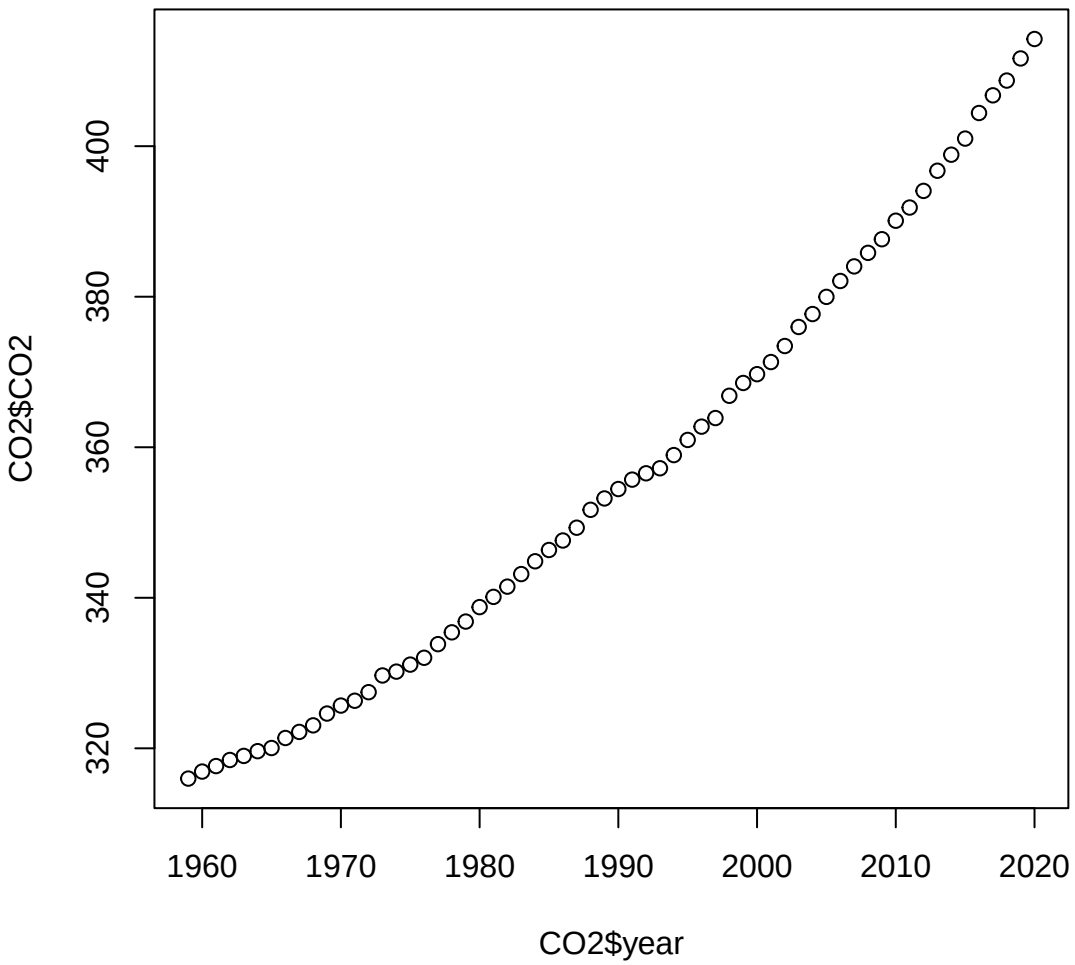
```
CO2 = read.table(file = "Data/climat_CO2.txt", header = TRUE)
head(CO2)
```

```
##   year    CO2 gistemp
## 1 1959 315.98  0.0240
## 2 1960 316.91 -0.0227
## 3 1961 317.64  0.0544
## 4 1962 318.45  0.0389
## 5 1963 318.99  0.0694
## 6 1964 319.62 -0.1960
```

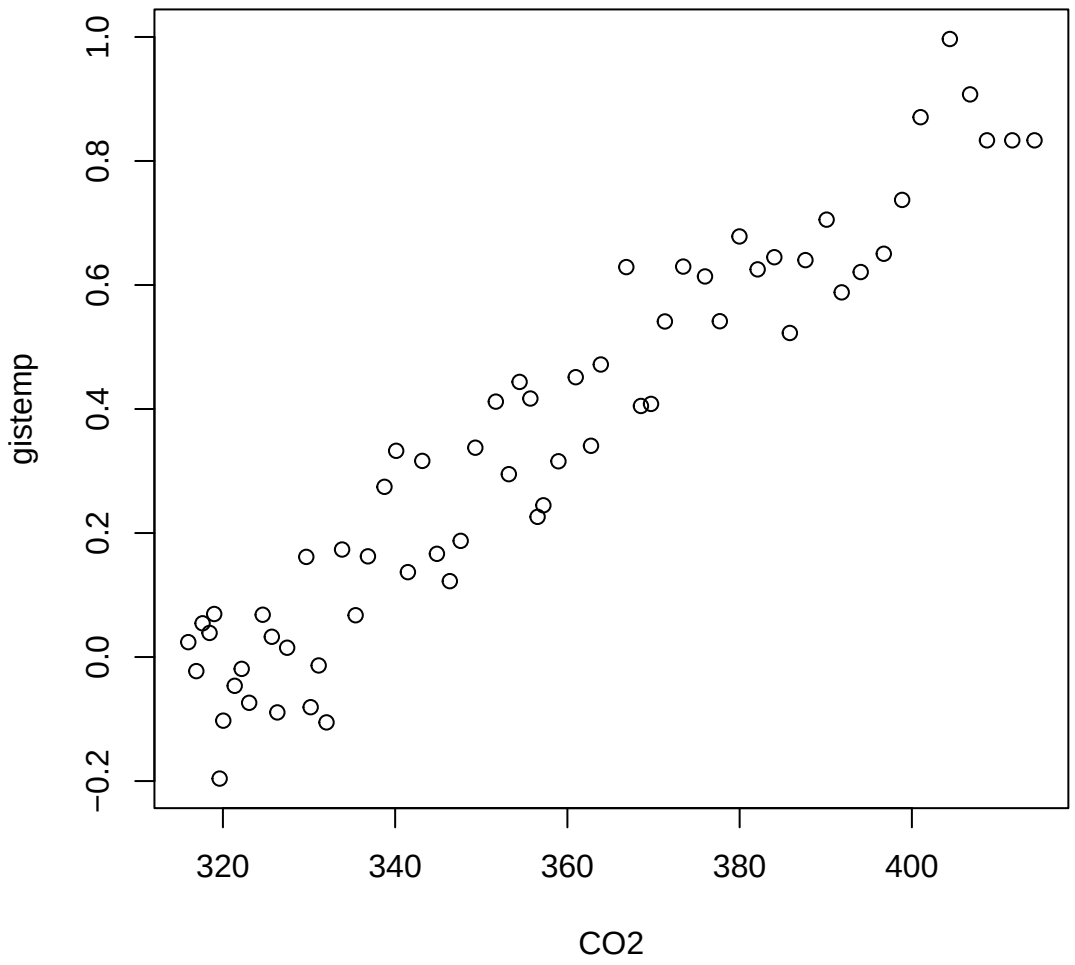
```
pairs(CO2)
```




```
plot(CO2$year,CO2$CO2)
```




```
plot(gistemp~CO2, data = CO2)
```



On remarque que l'objet CO2 est une structure de données. L'accès aux différents champs se fait avec le $. Mais la table CO2 peut aussi être utilisée comme une matrice. Les commandes pour accéder à une colonne, à une ligne, à un set d'indices sont alors les suivantes :

```
CO2[,3]
```

```
##  [1]  0.0240 -0.0227  0.0544  0.0389  0.0694 -0.1960 -0.1026 -0.0464 -0.0191
## [10] -0.0738  0.0682  0.0327 -0.0893  0.0150  0.1615 -0.0809 -0.0136 -0.1054
## [19]  0.1734  0.0674  0.1624  0.2745  0.3326  0.1369  0.3163  0.1665  0.1224
## [28]  0.1874  0.3376  0.4119  0.2949  0.4437  0.4169  0.2261  0.2447  0.3157
## [37]  0.4513  0.3407  0.4718  0.6289  0.4049  0.4082  0.5412  0.6297  0.6138
## [46]  0.5416  0.6783  0.6252  0.6448  0.5226  0.6401  0.7054  0.5881  0.6209
## [55]  0.6504  0.7372  0.8706  0.9967  0.9074  0.8333  0.8333  0.8333
```

```
CO2[30,]
```

```
##    year    CO2 gistemp
```

```
## 30 1988 351.69  0.4119
```

```
indices = CO2$gistemp>0.5
CO2[indices,1]
```

```
##  [1] 1998 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014
## [16] 2015 2016 2017 2018 2019 2020
```

```
CO2[c(12, 35, 50),]
```

```
##    year    CO2 gistemp
## 12 1970 325.68  0.0327
## 35 1993 357.21  0.2447
## 50 2008 385.83  0.5226
```

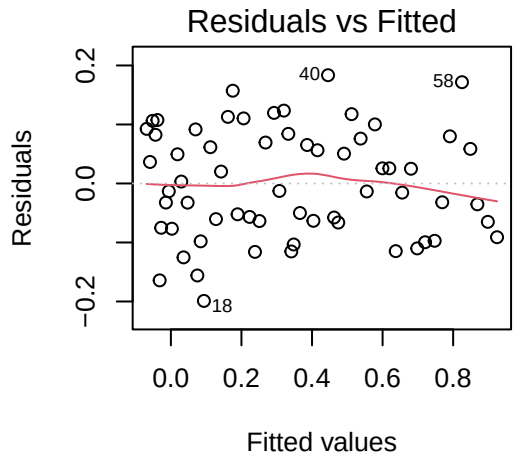
2 - On met en place un modèle de régression de la température en fonction du CO2

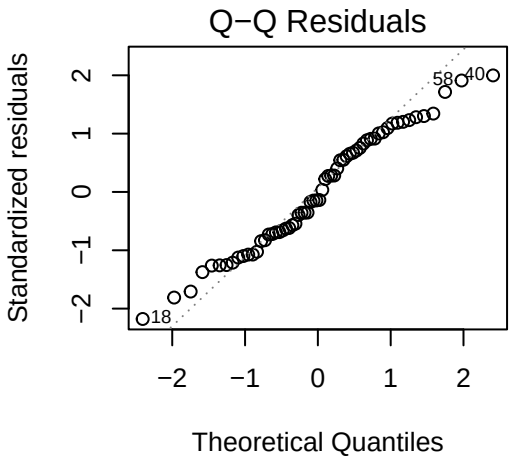
La commande summary permet d'avoir accès aux estimations des paramètres et aux tests statistiques. La commande plot permet l'analyse des résidus.

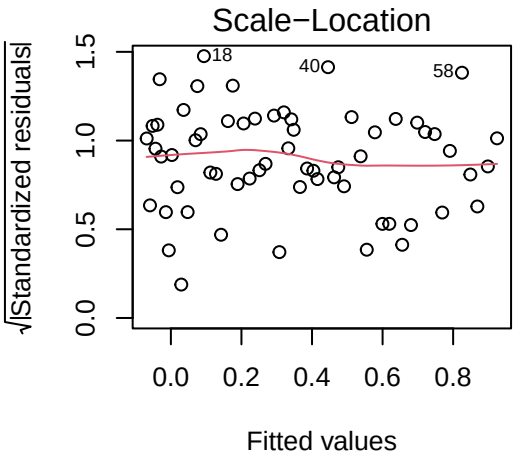
```
mod1 = lm(gistemp~CO2, data = CO2)
summary(mod1)
```

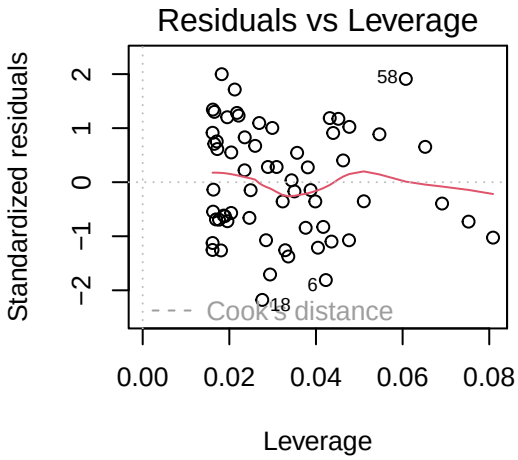
```
##
## Call:
## lm(formula = gistemp ~ CO2, data = CO2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19904 -0.06591 -0.01294  0.07893  0.18349
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.2616369  0.1457453  -22.38   <2e-16 ***
## CO2          0.0101053  0.0004076   24.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09268 on 60 degrees of freedom
## Multiple R-squared:  0.9111, Adjusted R-squared:  0.9096
## F-statistic: 614.6 on 1 and 60 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(mod1)
```

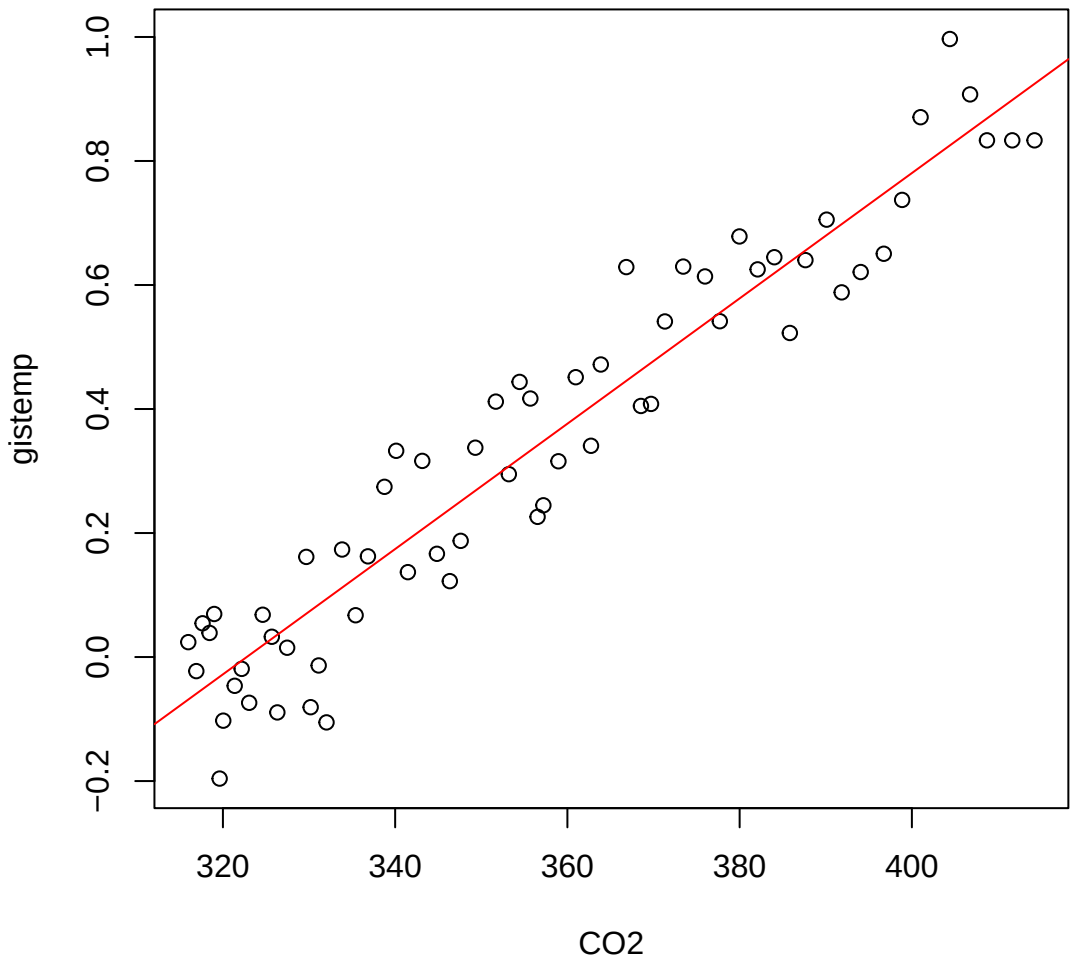









```
plot(gistemp~CO2, data = CO2)
abline(mod1$coefficients, col = 'red')
```

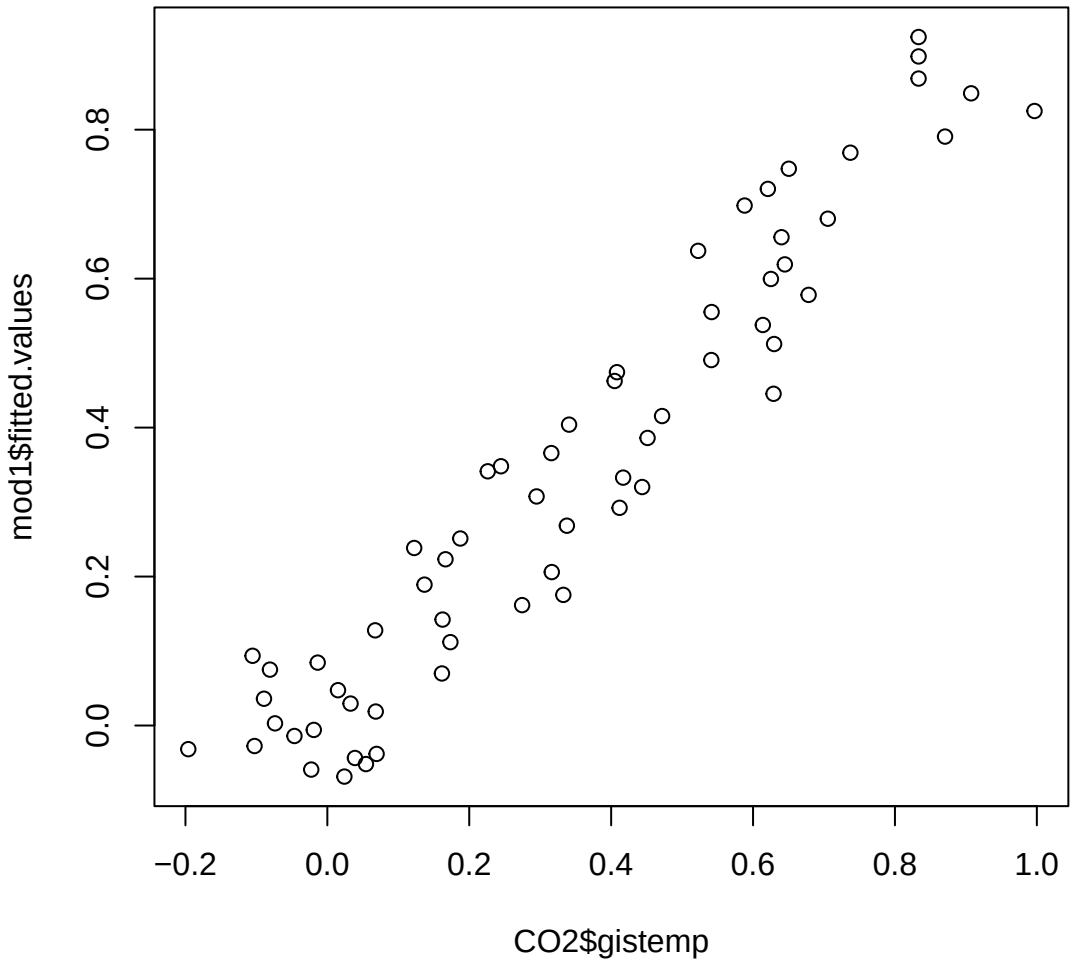


3 - Prédiction du modèle

La prédiction aux points d'observation

Le champ *fitted.values* de l'objet *CO2* contient les valeurs de $\hat{y}_i, \forall i \in \{1, ..., n\}$.

```
plot(CO2$gistemp ,mod1$fitted.values)
```



On considère que les résidus traduisent un bruit de mesure. Question : quelle est la donnée ayant la plus grande erreur de mesure ?

```
ecarts = abs(CO2$gistemp -mod1$fitted.values)
indice = ecarts == max(ecarts)
CO2[indice,]
```

```
##    year    CO2 gistemp
## 18 1976 332.03 -0.1054
```

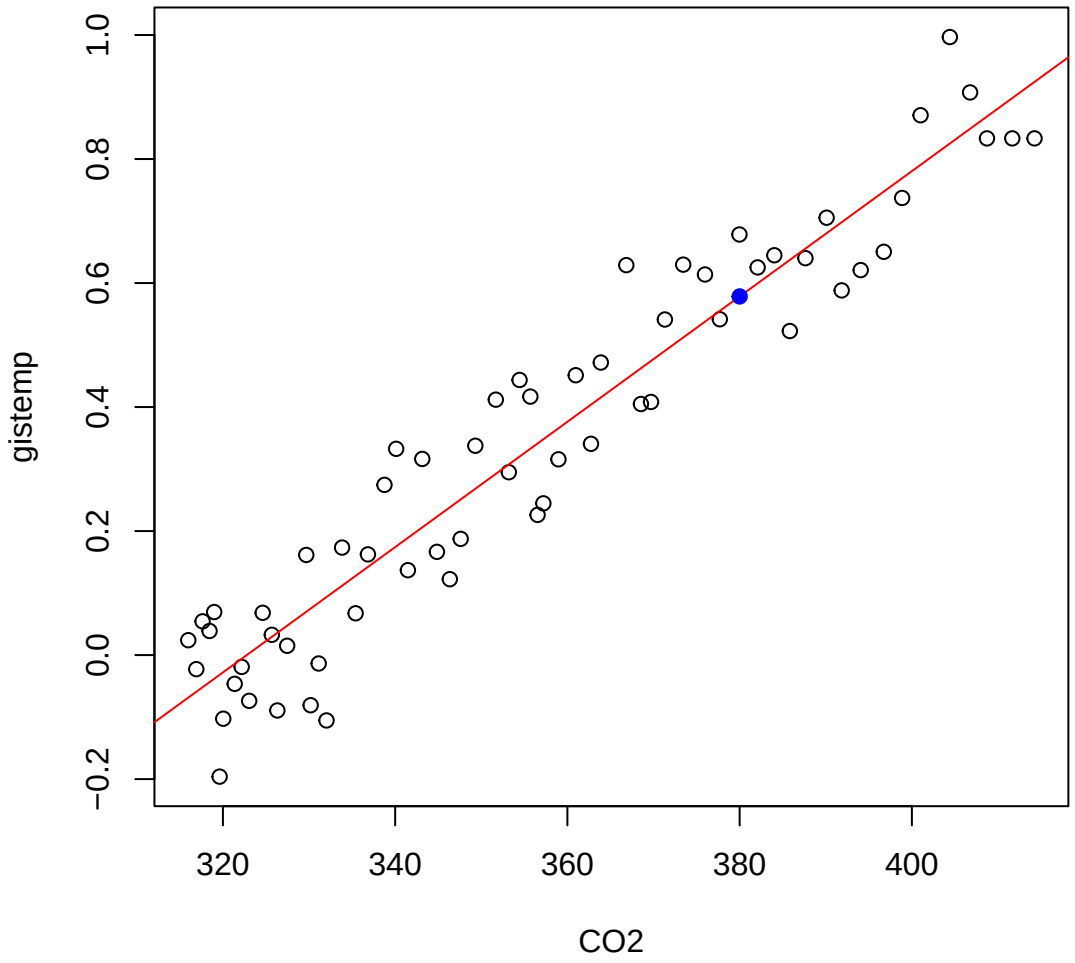
```
c(CO2$gistemp[indice] ,mod1$fitted.values[indice])
```

```
##                      18
## -0.10540000  0.09363919
```

La prédiction en dehors des points d'observation

On utilise la fonction *predict* sur un seul point :

```
nouvel_individu = data.frame(CO2 = 380)
prediction = predict(mod1, newdata = nouvel_individu)
plot(gistemp~CO2, data = CO2)
abline(mod1$coefficients, col = 'red')
points(nouvel_individu$CO2,prediction,col ="blue", pch = 19)
```



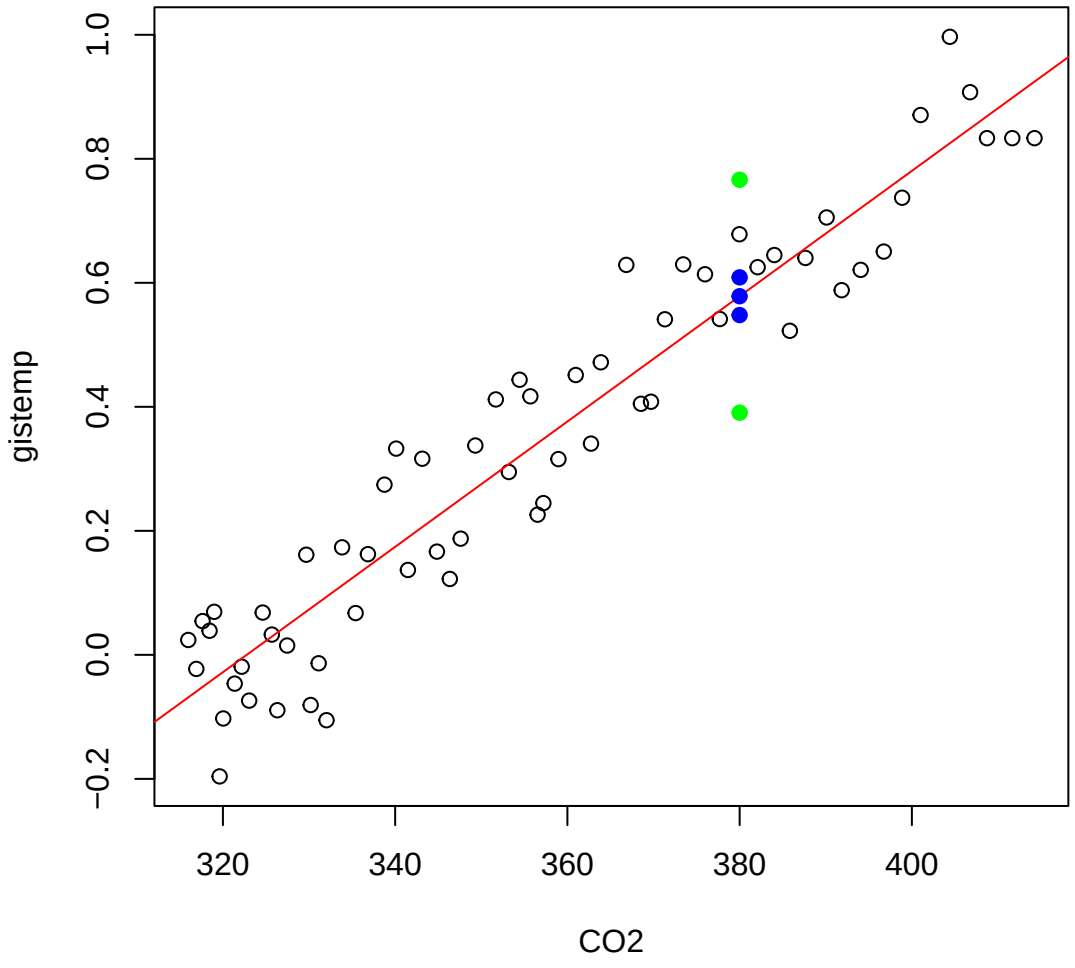
On peut aussi obtenir les intervalles de confiance et de prédiction :

```
prediction_IC = data.frame(predict(mod1, newdata = nouvel_individu, interval = 'confidence', level = 0.9
prediction_IC
```

```
##         fit       lwr       upr
## 1 0.5783924 0.5479782 0.6088065
```

```r
plot(gistemp~CO2, data = CO2)
abline(mod1$coefficients, col = 'red')
points(nouvel_individu$CO2,prediction_IC$fit,col ="blue", pch = 19)
points(nouvel_individu$CO2,prediction_IC$lwr,col ="blue", pch = 19)
points(nouvel_individu$CO2,prediction_IC$upr,col ="blue", pch = 19)

prediction_IP = data.frame(predict(mod1, newdata = nouvel_individu, interval = 'prediction', level = 0.9
points(nouvel_individu$CO2,prediction_IP$lwr,col ="green", pch = 19)
points(nouvel_individu$CO2,prediction_IP$upr,col ="green", pch = 19)
```



hhhhh partie exercice

Exercice 1 (Prise en main de R) – Évolution de la température moyenne de la surface de la planète

1) Analyser le Notebook "climat.rmd"

fait

2-0) Réaliser la régression linéaire du CO2 en fonction de year

```
mod2 = lm(CO2 ~ year, data = CO2)
summary(mod2)
```

```
##
## Call:
## lm(formula = CO2 ~ year, data = CO2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.775 -2.837 -1.301  2.381  9.062
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2826.2821    54.1163  -52.23   <2e-16 ***
## year            1.5997     0.0272   58.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.833 on 60 degrees of freedom
## Multiple R-squared:  0.983,  Adjusted R-squared:  0.9827
## F-statistic:  3459 on 1 and 60 DF,  p-value: < 2.2e-16
```

2-a) Donner l'équation du modèle

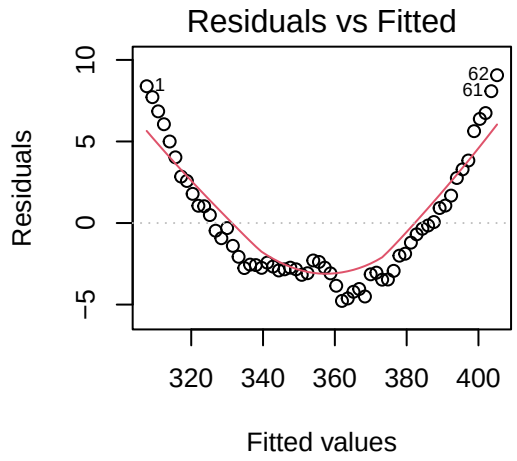
l'équation du modèle est : pred_CO2 = b0 + b1 * Year et d'après le summary : b0 = -2826.2821, b1 = 1.5997 (Approximations)

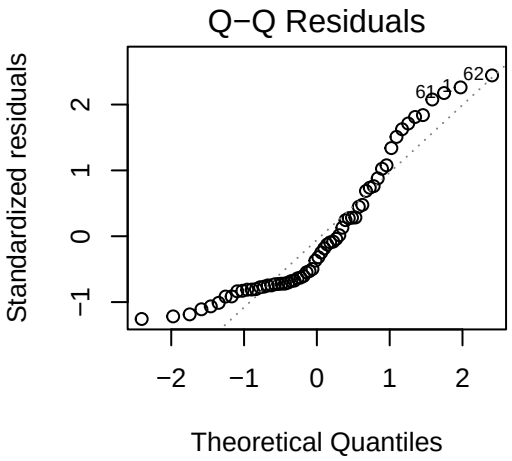
2-b) Part de variance expliquée par le modèle

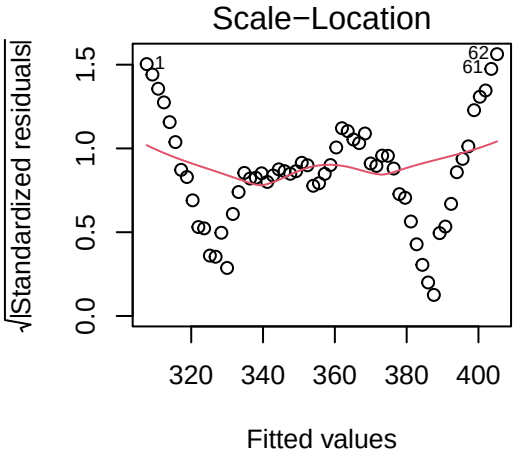
Le coefficient de détermination R2 mesure la part de la variance du CO2 expliquée par la variable year. Le modèle explique donc environ 98.3 % de la variabilité observée dans les concentrations de CO2

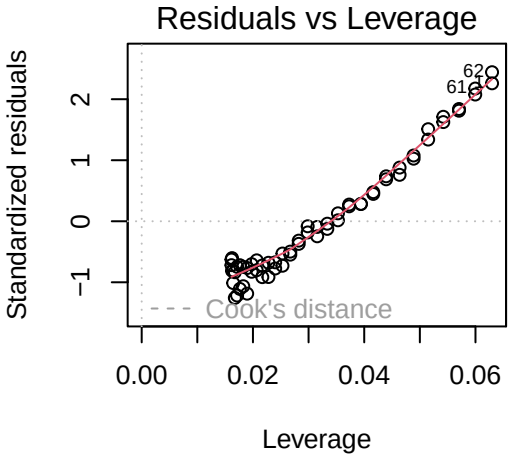
2-c) étude des résidus

```
par(mfrow = c(2,2))
plot(mod2)
```

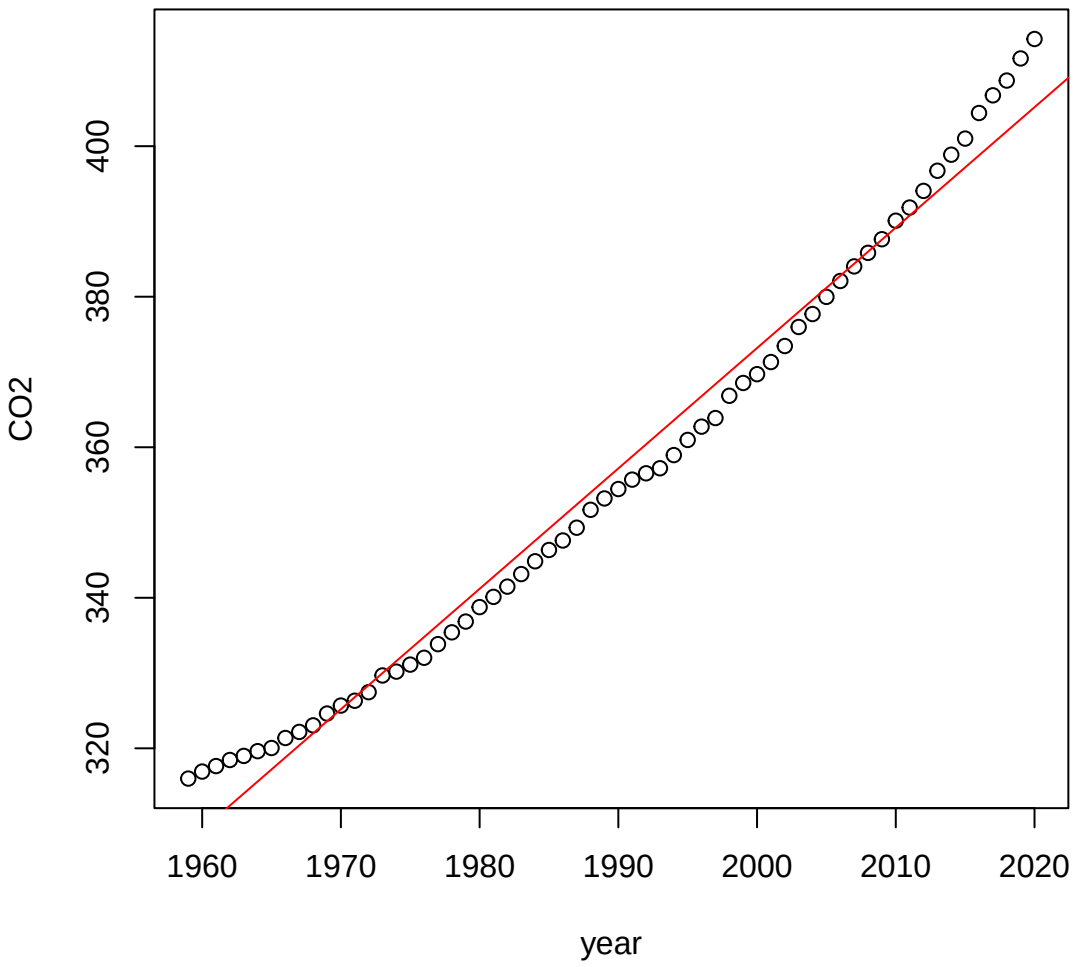









```
plot(CO2 ~ year, data = CO2)
abline(mod2$coefficients, col = "red")
```



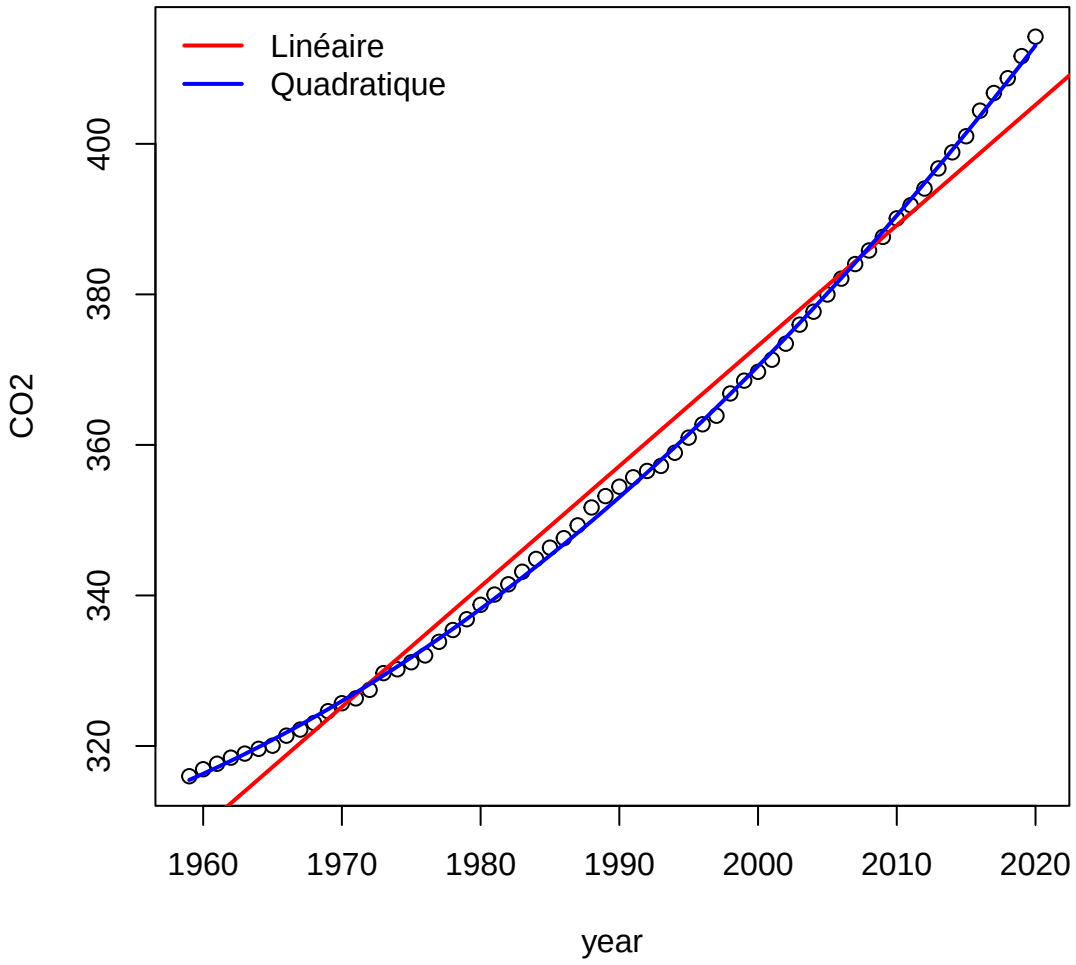
2-d) Autre modèle : Modèle quadratique

```
# Modèle avec un terme quadratique
mod2_quad = lm(CO2 ~ poly(year, 2), data = CO2)
summary(mod2_quad)
```

```
##
## Call:
## lm(formula = CO2 ~ poly(year, 2), data = CO2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0906 -0.5787 -0.1941  0.5301  1.8152
##
## Coefficients:
```

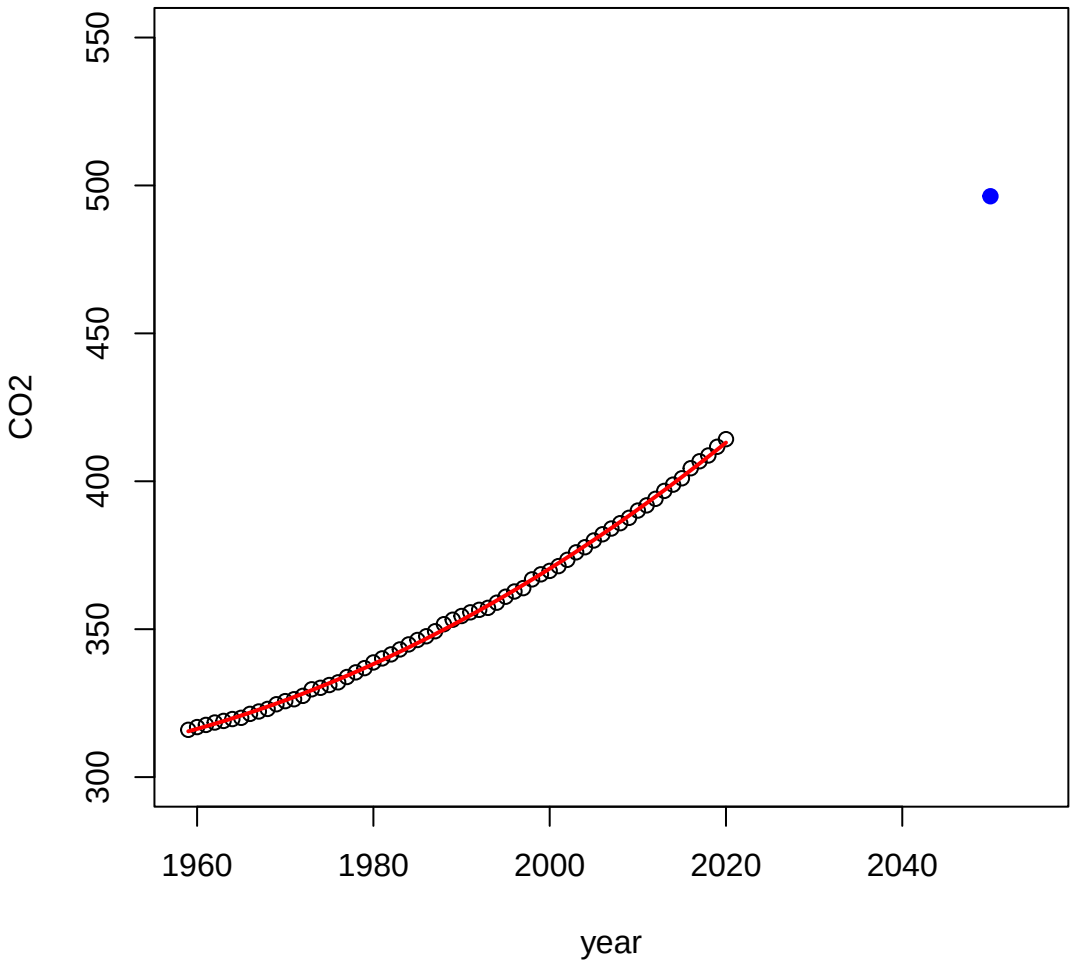
```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    356.38581    0.09315 3826.02   <2e-16 ***
## poly(year, 2)1 225.41759    0.73345  307.34   <2e-16 ***
## poly(year, 2)2  29.14867    0.73345   39.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7334 on 59 degrees of freedom
## Multiple R-squared:  0.9994, Adjusted R-squared:  0.9994
## F-statistic: 4.802e+04 on 2 and 59 DF,  p-value: < 2.2e-16
```

```r
plot(CO2 ~ year, data = CO2)
abline(mod2$coefficients, col = "red", lwd = 2)
lines(CO2$year, fitted(mod2_quad), col = "blue", lwd = 2)
legend("topleft", legend = c("Linéaire", "Quadratique"),
       col = c("red", "blue"), lwd = 2, bty = "n")
```



3-a) Prédiction du niveau de CO2 Moyen en 2050

```
nouvelle_annee = data.frame(year = 2050)
prediction = predict(mod2_quad, newdata = nouvelle_annee)
plot(CO2~year, data = CO2, xlim = c(1959, 2055),ylim = c(300,550))
lines(CO2$year, fitted(mod2_quad), col = "red", lwd = 2)
points(nouvelle_annee$year,prediction,col ="blue", pch = 19)
```



3-b) prédiction de l'anomalie de température en 2050

```
nouvelle_annee = data.frame(year = 2050)
predict_CO2_2050 = predict(mod2_quad, newdata = nouvelle_annee, interval = "confidence")
predict_CO2_2050
```

```
##        fit      lwr      upr
## 1 496.3518 494.0805 498.6232
```

```
CO2_2050 = predict_CO2_2050[1]  # valeur centrale
prediction_temp_2050 = predict(mod1, newdata = data.frame(CO2 = CO2_2050), interval = "prediction")
prediction_temp_2050
```

```
##        fit      lwr      upr
## 1 1.754167 1.535204 1.973131
```

3-c) prévision optimiste et pessimiste de la température

Exercice 2 – Étude de la densité de peuplement des chenilles

On utilise les données du fichier `Data/chenilles.txt`, qui contiennent 32 parcelles et 10 variables explicatives [X1, . . . , X10] comme décrit.

---

1) Lecture des données

```
chenilles = read.table("Data/chenilles.txt", header = TRUE)
head(chenilles)
```

```
##     X1 X2 X3  X4   X5  X6  X7  X8  X9 X10    Y
## 1 1200 22  1 4.0 14.8 1.0 1.1 5.9 1.4 1.4 2.37
## 2 1342 28  8 4.4 18.0 1.5 1.5 6.4 1.7 1.7 1.47
## 3 1231 28  5 2.4  7.8 1.3 1.6 4.3 1.5 1.4 1.13
## 4 1254 28 18 3.0  9.2 2.3 1.7 6.9 2.3 1.6 0.85
## 5 1357 32  7 3.7 10.7 1.4 1.7 6.6 1.8 1.3 0.24
## 6 1250 27  1 4.4 14.8 1.0 1.7 5.8 1.3 1.4 1.49
```

2) Visualisation des données

```
summary(chenilles)
```

```
##        X1             X2              X3              X4              X5
##  Min.   :1075   Min.   :15.00   Min.   : 0.00   Min.   :2.400   Min.   : 5.80
##  1st Qu.:1223   1st Qu.:24.00   1st Qu.: 3.50   1st Qu.:3.650   1st Qu.:11.47
##  Median :1307   Median :27.50   Median : 8.00   Median :4.400   Median :16.05
##  Mean   :1312   Mean   :28.72   Mean   :11.22   Mean   :4.447   Mean   :15.25
##  3rd Qu.:1394   3rd Qu.:32.50   3rd Qu.:17.25   3rd Qu.:5.325   3rd Qu.:18.45
##  Max.   :1575   Max.   :46.00   Max.   :32.00   Max.   :6.500   Max.   :21.80
##        X6              X7              X8              X9
##  Min.   :1.000   Min.   :1.100   Min.   : 3.600   Min.   :1.100
##  1st Qu.:1.175   1st Qu.:1.600   1st Qu.: 5.875   1st Qu.:1.500
##  Median :1.500   Median :1.700   Median : 7.100   Median :1.900
##  Mean   :1.772   Mean   :1.659   Mean   : 7.491   Mean   :1.988
```

```
##  3rd Qu.:2.325   3rd Qu.:1.800   3rd Qu.: 8.525   3rd Qu.:2.500
##  Max.   :3.300   Max.   :1.900   Max.   :13.700   Max.   :2.900
##       X10               Y
##  Min.   :1.300   Min.   :0.0300
##  1st Qu.:1.550   1st Qu.:0.1675
##  Median :1.800   Median :0.5400
##  Mean   :1.747   Mean   :0.8141
##  3rd Qu.:2.000   3rd Qu.:1.1500
##  Max.   :2.000   Max.   :3.0000
```

```
pairs(chenilles)
```

3) Equation du modèle de régression

Y = b0 + b1 * X1 + b2 * X2 + … + b10 * X10 (+ Epsilon)

Il y a donc 11 paramètres à estimer (b0 à b10)

4) Calcul de la régression

```r
# Régression linéaire multiple

modele_full = lm(Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10, data = chenilles)
summary(modele_full)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +
##     X10, data = chenilles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.06301 -0.31676  0.02223  0.21138  1.24047
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.788813   2.123588   4.139 0.000466 ***
## X1          -0.003250   0.001067  -3.047 0.006128 **
## X2          -0.042677   0.014887  -2.867 0.009234 **
## X3           0.049665   0.067846   0.732 0.472248
## X4          -0.538905   0.380108  -1.418 0.170927
## X5           0.136178   0.070675   1.927 0.067642 .
## X6          -0.422358   1.082617  -0.390 0.700371
## X7           0.045941   0.680063   0.068 0.946780
## X8          -0.038185   0.152034  -0.251 0.804133
## X9          -0.362574   0.559346  -0.648 0.523874
## X10         -0.597971   0.496597  -1.204 0.241934
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5537 on 21 degrees of freedom
## Multiple R-squared:  0.6903, Adjusted R-squared:  0.5428
## F-statistic: 4.681 on 10 and 21 DF,  p-value: 0.001411
```

5) Test de non-régression

Hypothèse nulle H0 : Tous les coefficients sauf l'intercept sont nuls (b1=b2=...=b10=0)

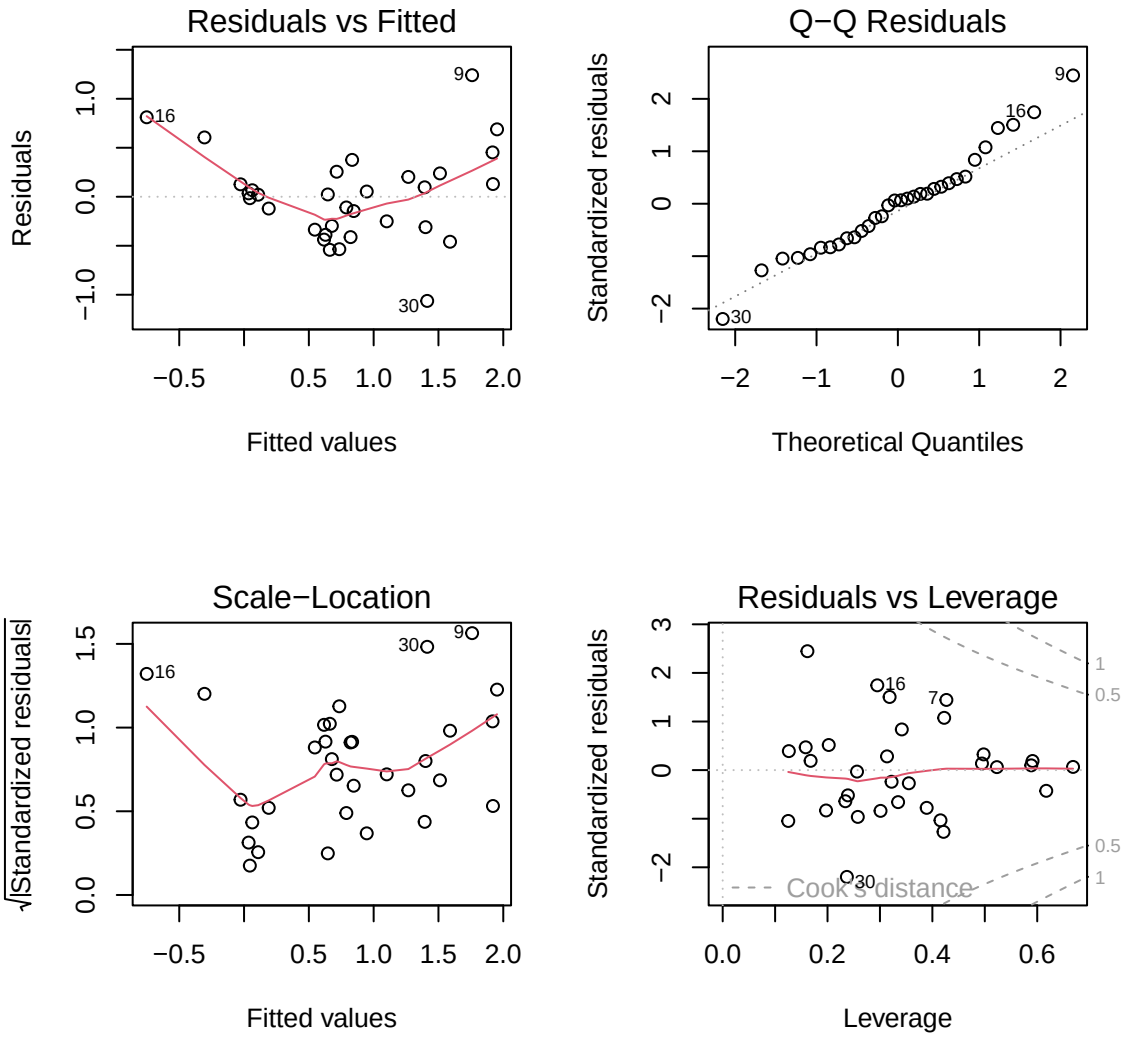
Statistique : F = (SSR/k)/(SSE/(n-k-1))

Résultat : summary(modele_full) fournit la p-value. Si la p-value est inférieure à 0.05, on rejette H0, le modèle est significatif.

Conclusion : Ici p-value = 0.001411 $<$ 0.05 donc le modèle est significatif.

6) Le modèle estimé convient-il ?

```r
par(mfrow=c(2,2))
plot(modele_full)
```



7) Algorithme backward et test de Student

Principe : On commence avec toutes les variables et on retire la variable la moins significative (p-value la plus élevée > 0.05) jusqu'à ce que toutes soient significatives.

```
modele_backward = step(modele_full, direction = "backward", trace = FALSE)
summary(modele_backward)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X4 + X5 + X9, data = chenilles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07735 -0.31540  0.01193  0.27466  1.18208
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.4550929  1.0083502   6.402 8.81e-07 ***
## X1          -0.0025453  0.0007877  -3.231  0.00333 **
## X2          -0.0412082  0.0135267  -3.046  0.00526 **
## X4          -0.5716393  0.2458661  -2.325  0.02815 *
## X5           0.1349001  0.0554544   2.433  0.02218 *
## X9          -0.3190932  0.2181035  -1.463  0.15544
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5226 on 26 degrees of freedom
## Multiple R-squared:  0.6584, Adjusted R-squared:  0.5927
## F-statistic: 10.02 on 5 and 26 DF,  p-value: 1.968e-05
```

Le modèle obtenu avec le test de Student est :

pred (Y) = 6.4550929 - 0.0025453 * X1 - 0.0412082 * X2 - 0.5716393 * X4 + 0.1349001 * X5 - 0.3190932 * X9

8-1) Critère AIC

```r
# AIC
modele_AIC = step(modele_full, direction = "both", k = 2) # k=2 correspond à AIC
```

```
## Start:  AIC=-29.31
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
##
##        Df Sum of Sq    RSS     AIC
## - X7    1   0.00140 6.4402 -31.302
## - X8    1   0.01934 6.4582 -31.212
## - X6    1   0.04667 6.4855 -31.077
## - X9    1   0.12883 6.5677 -30.674
## - X3    1   0.16430 6.6031 -30.502
## <none>              6.4388 -29.308
## - X10   1   0.44457 6.8834 -29.172
## - X4    1   0.61631 7.0551 -28.383
## - X5    1   1.13833 7.5772 -26.099
## - X2    1   2.51977 8.9586 -20.740
## - X1    1   2.84635 9.2852 -19.594
##
## Step:  AIC=-31.3
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X8 + X9 + X10
##
##        Df Sum of Sq    RSS     AIC
## - X8    1   0.02258 6.4628 -33.190
## - X6    1   0.04869 6.4889 -33.060
## - X9    1   0.14498 6.5852 -32.589
## - X3    1   0.17647 6.6167 -32.436
## <none>              6.4402 -31.302
## - X10   1   0.50959 6.9498 -30.865
## - X4    1   0.73736 7.1776 -29.833
```

```
## + X7    1   0.00140 6.4388 -29.308
## - X5    1   1.42808 7.8683 -26.893
## - X2    1   2.59823 9.0385 -22.456
## - X1    1   3.06570 9.5059 -20.842
##
## Step:  AIC=-33.19
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X9 + X10
##
##        Df Sum of Sq    RSS     AIC
## - X6    1   0.09347 6.5563 -34.730
## - X9    1   0.21889 6.6817 -34.124
## - X3    1   0.23416 6.6970 -34.051
## <none>              6.4628 -33.190
## - X10   1   0.52151 6.9843 -32.706
## + X8    1   0.02258 6.4402 -31.302
## + X7    1   0.00464 6.4582 -31.212
## - X4    1   1.41968 7.8825 -28.835
## - X5    1   1.51893 7.9817 -28.435
## - X2    1   2.66641 9.1292 -24.136
## - X1    1   3.05717 9.5200 -22.795
##
## Step:  AIC=-34.73
## Y ~ X1 + X2 + X3 + X4 + X5 + X9 + X10
##
##        Df Sum of Sq    RSS     AIC
## - X3    1   0.23363 6.7899 -35.610
## <none>              6.5563 -34.730
## - X10   1   0.42815 6.9844 -34.706
## - X9    1   0.54662 7.1029 -34.167
## + X6    1   0.09347 6.4628 -33.190
## + X8    1   0.06735 6.4889 -33.060
## + X7    1   0.00256 6.5537 -32.743
## - X4    1   1.36578 7.9221 -30.675
## - X5    1   1.44776 8.0040 -30.345
## - X2    1   2.59648 9.1528 -26.054
## - X1    1   3.02671 9.5830 -24.584
##
## Step:  AIC=-35.61
## Y ~ X1 + X2 + X4 + X5 + X9 + X10
##
##        Df Sum of Sq    RSS     AIC
## - X10   1   0.31212 7.1020 -36.171
## - X9    1   0.36883 7.1587 -35.917
## <none>              6.7899 -35.610
## + X3    1   0.23363 6.5563 -34.730
## + X6    1   0.09293 6.6970 -34.051
## + X8    1   0.02656 6.7633 -33.735
## + X7    1   0.00034 6.7896 -33.611
## - X4    1   1.67114 8.4610 -30.568
## - X5    1   1.79263 8.5825 -30.112
## - X2    1   2.40560 9.1955 -27.905
## - X1    1   3.09562 9.8855 -25.589
##
## Step:  AIC=-36.17
```

```
## Y ~ X1 + X2 + X4 + X5 + X9
##
##        Df Sum of Sq    RSS     AIC
## <none>              7.1020 -36.171
## - X9    1   0.58468 7.6867 -35.640
## + X10   1   0.31212 6.7899 -35.610
## + X3    1   0.11759 6.9844 -34.706
## + X6    1   0.08852 7.0135 -34.573
## + X7    1   0.03013 7.0719 -34.307
## + X8    1   0.01313 7.0889 -34.231
## - X4    1   1.47657 8.5786 -32.127
## - X5    1   1.61644 8.7185 -31.609
## - X2    1   2.53509 9.6371 -28.404
## - X1    1   2.85229 9.9543 -27.367
```

```
summary(modele_AIC)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X4 + X5 + X9, data = chenilles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07735 -0.31540  0.01193  0.27466  1.18208
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.4550929  1.0083502   6.402 8.81e-07 ***
## X1          -0.0025453  0.0007877  -3.231  0.00333 **
## X2          -0.0412082  0.0135267  -3.046  0.00526 **
## X4          -0.5716393  0.2458661  -2.325  0.02815 *
## X5           0.1349001  0.0554544   2.433  0.02218 *
## X9          -0.3190932  0.2181035  -1.463  0.15544
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5226 on 26 degrees of freedom
## Multiple R-squared:  0.6584, Adjusted R-squared:  0.5927
## F-statistic: 10.02 on 5 and 26 DF,  p-value: 1.968e-05
```

Le modèle obtenu avec le critère AIC est :

pred (Y) = 6.4550929 - 0.0025453 * X1 - 0.0412082 * X2 - 0.5716393 * X4 + 0.1349001 * X5 - 0.3190932 * X9

8-2) Critère BIC

```
# BIC
modele_BIC = step(modele_full, direction = "both", k = log(nrow(chenilles)))
```

```
## Start:  AIC=-13.19
```

```
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
##
##        Df Sum of Sq    RSS      AIC
## - X7    1   0.00140 6.4402 -16.6442
## - X8    1   0.01934 6.4582 -16.5551
## - X6    1   0.04667 6.4855 -16.4200
## - X9    1   0.12883 6.5677 -16.0172
## - X3    1   0.16430 6.6031 -15.8448
## - X10   1   0.44457 6.8834 -14.5146
## - X4    1   0.61631 7.0551 -13.7260
## <none>              6.4388 -13.1854
## - X5    1   1.13833 7.5772 -11.4418
## - X2    1   2.51977 8.9586  -6.0826
## - X1    1   2.84635 9.2852  -4.9368
##
## Step:  AIC=-16.64
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X8 + X9 + X10
##
##        Df Sum of Sq    RSS      AIC
## - X8    1   0.02258 6.4628 -19.9979
## - X6    1   0.04869 6.4889 -19.8688
## - X9    1   0.14498 6.5852 -19.3975
## - X3    1   0.17647 6.6167 -19.2448
## - X10   1   0.50959 6.9498 -17.6730
## <none>              6.4402 -16.6442
## - X4    1   0.73736 7.1776 -16.6411
## - X5    1   1.42808 7.8683 -13.7010
## + X7    1   0.00140 6.4388 -13.1854
## - X2    1   2.59823 9.0385  -9.2643
## - X1    1   3.06570 9.5059  -7.6506
##
## Step:  AIC=-20
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X9 + X10
##
##        Df Sum of Sq    RSS     AIC
## - X6    1   0.09347 6.5563 -23.004
## - X9    1   0.21889 6.6817 -22.398
## - X3    1   0.23416 6.6970 -22.325
## - X10   1   0.52151 6.9843 -20.980
## <none>              6.4628 -19.998
## - X4    1   1.41968 7.8825 -17.109
## - X5    1   1.51893 7.9817 -16.709
## + X8    1   0.02258 6.4402 -16.644
## + X7    1   0.00464 6.4582 -16.555
## - X2    1   2.66641 9.1292 -12.410
## - X1    1   3.05717 9.5200 -11.069
##
## Step:  AIC=-23
## Y ~ X1 + X2 + X3 + X4 + X5 + X9 + X10
##
##        Df Sum of Sq    RSS     AIC
## - X3    1   0.23363 6.7899 -25.349
## - X10   1   0.42815 6.9844 -24.445
## - X9    1   0.54662 7.1029 -23.907
```

```
## <none>              6.5563 -23.004
## - X4    1   1.36578 7.9221 -20.415
## - X5    1   1.44776 8.0040 -20.085
## + X6    1   0.09347 6.4628 -19.998
## + X8    1   0.06735 6.4889 -19.869
## + X7    1   0.00256 6.5537 -19.551
## - X2    1   2.59648 9.1528 -15.794
## - X1    1   3.02671 9.5830 -14.324
##
## Step:  AIC=-25.35
## Y ~ X1 + X2 + X4 + X5 + X9 + X10
##
##        Df Sum of Sq    RSS     AIC
## - X10   1   0.31212 7.1020 -27.377
## - X9    1   0.36883 7.1587 -27.122
## <none>              6.7899 -25.349
## + X3    1   0.23363 6.5563 -23.004
## + X6    1   0.09293 6.6970 -22.325
## + X8    1   0.02656 6.7633 -22.009
## + X7    1   0.00034 6.7896 -21.885
## - X4    1   1.67114 8.4610 -21.774
## - X5    1   1.79263 8.5825 -21.318
## - X2    1   2.40560 9.1955 -19.110
## - X1    1   3.09562 9.8855 -16.795
##
## Step:  AIC=-27.38
## Y ~ X1 + X2 + X4 + X5 + X9
##
##        Df Sum of Sq    RSS     AIC
## - X9    1   0.58468 7.6867 -28.311
## <none>              7.1020 -27.377
## + X10   1   0.31212 6.7899 -25.349
## - X4    1   1.47657 8.5786 -24.798
## + X3    1   0.11759 6.9844 -24.445
## + X6    1   0.08852 7.0135 -24.313
## - X5    1   1.61644 8.7185 -24.281
## + X7    1   0.03013 7.0719 -24.047
## + X8    1   0.01313 7.0889 -23.971
## - X2    1   2.53509 9.6371 -21.075
## - X1    1   2.85229 9.9543 -20.039
##
## Step:  AIC=-28.31
## Y ~ X1 + X2 + X4 + X5
##
##        Df Sum of Sq     RSS     AIC
## <none>               7.6867 -28.311
## + X9    1    0.5847  7.1020 -27.377
## + X10   1    0.5280  7.1587 -27.122
## + X8    1    0.4874  7.1993 -26.942
## + X6    1    0.2878  7.3989 -26.067
## + X3    1    0.2099  7.4768 -25.732
## + X7    1    0.0072  7.6795 -24.875
## - X5    1    3.0422 10.7289 -21.107
## - X2    1    3.2755 10.9622 -20.418
```

```
## - X4    1    3.4821 11.1688 -19.820
## - X1    1    3.6820 11.3687 -19.253
```

```
summary(modele_BIC)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X4 + X5, data = chenilles)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97905 -0.34390 -0.06326  0.31679  1.18260
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.6030875  1.0242331   6.447 6.58e-07 ***
## X1          -0.0028129  0.0007822  -3.596  0.00127 **
## X2          -0.0456472  0.0134575  -3.392  0.00215 **
## X4          -0.7550950  0.2159068  -3.497  0.00164 **
## X5           0.1684745  0.0515384   3.269  0.00294 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5336 on 27 degrees of freedom
## Multiple R-squared:  0.6303, Adjusted R-squared:  0.5755
## F-statistic: 11.51 on 4 and 27 DF,  p-value: 1.394e-05
```

Le modèle obtenu avec le critère BIC est :

pred (Y) = 6.6030875 - 0.0028129 * X1 - 0.0456472 * X2 - 0.7550950 * X4 + 0.1684745 * X5

9) Choix du meilleur modèle

Comparaison des modèles

```
anova(modele_backward, modele_AIC, modele_BIC)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2 + X4 + X5 + X9
## Model 2: Y ~ X1 + X2 + X4 + X5 + X9
## Model 3: Y ~ X1 + X2 + X4 + X5
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 7.1020
## 2     26 7.1020  0   0.00000
## 3     27 7.6867 -1  -0.58468 2.1405 0.1554
```

Le modèle réduit obtenu via BIC, sans la variable X9 dont la suppression ne dégrade pas significativement l'ajustement (p = 0.1554), est retenu. Ce modèle est plus simple et conserve quasiment le même pouvoir explicatif que le modèle complet.

Exercice 3 - Estimateur du maximum de vraisemblance du paramètre p

1. Fonction de vraisemblance

La fonction de masse de probabilité pour $X_i \sim B(m, p)$ est :

$$P(X_i = x_i) = \binom{m}{x_i} p^{x_i}(1-p)^{m-x_i}, \quad x_i = 0, 1, \ldots, m$$

Comme les X_i sont indépendantes, la fonction de vraisemblance pour l'échantillon est :

$$L(p) = \prod_{i=1}^{n} \binom{m}{x_i} p^{x_i}(1-p)^{m-x_i} = \left(\prod_{i=1}^{n} \binom{m}{x_i} \right) p^{\sum_{i=1}^{n} x_i}(1-p)^{nm - \sum_{i=1}^{n} x_i}$$

Posons $S = \sum_{i=1}^{n} X_i$. Alors :

$$L(p) = \left(\prod_{i=1}^{n} \binom{m}{x_i} \right) p^{S}(1-p)^{nm-S}$$

---

2. Log-vraisemblance

$$\ell(p) = \log L(p) = \sum_{i=1}^{n} \log \binom{m}{x_i} + S \log p + (nm - S) \log(1-p)$$

On maximise uniquement la partie dépendante de p :

$$\ell(p) = S \log p + (nm - S) \log(1-p)$$

---

3. Estimateur du maximum de vraisemblance

Dérivons $\ell(p)$ par rapport à p :

$$\frac{d\ell}{dp} = \frac{S}{p} - \frac{nm - S}{1-p} = 0 \implies S(1-p) = (nm - S)p \implies \hat{p} = \frac{S}{nm} = \frac{\sum_{i=1}^{n} X_i}{nm}$$

---

4. Biais

$$E[\hat{p}] = E\left[\frac{\sum_{i=1}^{n} X_i}{nm} \right] = \frac{1}{nm} \sum_{i=1}^{n} E[X_i] = \frac{1}{nm} \cdot n \cdot mp = p$$

$$\text{Biais}(\hat{p}) = E[\hat{p}] - p = 0$$

L'estimateur est **non biaisé**.

---

5. Variance

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}\left(\frac{1}{nm}\sum_{i=1}^{n} X_i\right) = \frac{1}{(nm)^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{n \cdot mp(1-p)}{(nm)^2} = \frac{p(1-p)}{nm}$$

Résumé

$$\hat{p}_{\mathrm{MLE}} = \frac{\sum_{i=1}^{n} X_i}{nm}, \quad \text{Biais} = 0, \quad \mathrm{Var}(\hat{p}) = \frac{p(1-p)}{nm}$$

Exercice 4 - Variables aléatoires, estimateurs sans biais

1. Espérance et variance de Y_j

$$P(Y_j = 1) = P(X_j \geq a) = \frac{b-a}{b}, \quad P(Y_j = -1) = \frac{a}{b}$$

Espérance

$$E[Y_j] = 1 \cdot \frac{b-a}{b} + (-1) \cdot \frac{a}{b} = \frac{b-2a}{b}$$

Variance

$$\mathrm{Var}(Y_j) = E[Y_j^2] - (E[Y_j])^2 = 1 - \left(\frac{b-2a}{b}\right)^2 = \frac{4a(b-a)}{b^2}$$

---

2. Estimateur $\hat{a} = k\sum_{j=1}^{n} Y_j$

$$E[\hat{a}] = E\left[k\sum_{j=1}^{n} Y_j\right] = kn\frac{b-2a}{b}$$

Pour que $\hat{a}$ soit sans biais :

$$a = kn\frac{b-2a}{b} \implies k = \frac{ab}{n(b-2a)}$$

Cela dépend de a, donc **il n'existe pas de $k \in \mathbb{R}$ indépendant de a** permettant d'obtenir un estimateur sans biais.

---

3. Estimateur du maximum de vraisemblance

Soit $\epsilon_j = Y_j \in \{-1, 1\}$. Alors :

$$P(\epsilon_1, \ldots, \epsilon_n; a) = \prod_{j=1}^{n} P(Y_j = \epsilon_j) = \left(\frac{b-a}{b}\right)^{\sum_{j=1}^{n} \frac{1+\epsilon_j}{2}} \left(\frac{a}{b}\right)^{\sum_{j=1}^{n} \frac{1-\epsilon_j}{2}}$$

Log-vraisemblance

$$\ell(a) = \sum_{j=1}^{n} \frac{1+\epsilon_j}{2} \log(b-a) + \sum_{j=1}^{n} \frac{1-\epsilon_j}{2} \log(a) - n \log b$$

Dérivée :

$$\frac{d\ell}{da} = -\frac{n_+}{b-a} + \frac{n_-}{a} = 0 \implies a = \frac{n_-}{n} b$$

avec $n_+ = \#\{\epsilon_j = 1\}$, $n_- = \#\{\epsilon_j = -1\}$.

Donc le **MLE est** :

$$\hat{a}_{\text{MLE}} = \frac{b}{2}\left(1 - \frac{1}{n}\sum_{j=1}^{n} Y_j\right)$$

Biais et variance du MLE

$$E[\hat{a}_{\text{MLE}}] = \frac{b}{2}(1 - E[Y_j]) = \frac{b}{2}\left(1 - \frac{b-2a}{b}\right) = a$$

$$\text{Var}(\hat{a}_{\text{MLE}}) = \left(\frac{b}{2}\right)^2 \frac{\text{Var}(Y_j)}{n} = \frac{b^2}{4} \cdot \frac{4a(b-a)}{b^2 n} = \frac{a(b-a)}{n}$$

$$\lim_{n\to\infty} \text{Var}(\hat{a}_{\text{MLE}}) = 0$$

Résumé

$$\begin{cases} E[Y_j] = \frac{b-2a}{b}, & \text{Var}(Y_j) = \frac{4a(b-a)}{b^2} \\ \hat{a} = k \sum Y_j & \text{pas de k indépendant de a pour un estimateur sans biais} \\ \hat{a}_{\text{MLE}} = \frac{b}{2}(1 - \bar{Y}) & \text{sans biais, Var} = \frac{a(b-a)}{n} \to 0 \end{cases}$$