Rapport TP1

Candice AUBERTIN - Ewen EXPUESTO

29 Septembre 2025

Contents

1 Application: study your own data using a linear model with transformed data

Pour cet exercice, nous avons collecté les données de la valeur en bourse d'Amazon. Nous les avons sélectionnées sur les 21 dernières années. Plusieurs catégories de données pour représenter notre étude nous étaient proposées, nous avons alors gardé la plus pertinente, le prix de clôture ajusté (Adj_Close)

```
tab <- read.table("~/MRR21/Files/TP1/Donnees_amazon.txt",header=TRUE, sep=";")
tab
```

Année	Adj.Close
2005	2.16
2006	2.24
2007	1.88
2008	3.88
2009	2.94
2010	6.27
2011	8.48
2012	9.72
2013	13.27
2014	17.93
2015	17.73
2016	29.35
2017	41.17
2018	72.54
2019	85.94
2020	100.44
2021	160.31
2022	149.57
2023	103.13
2024	155.2
2025	237.68

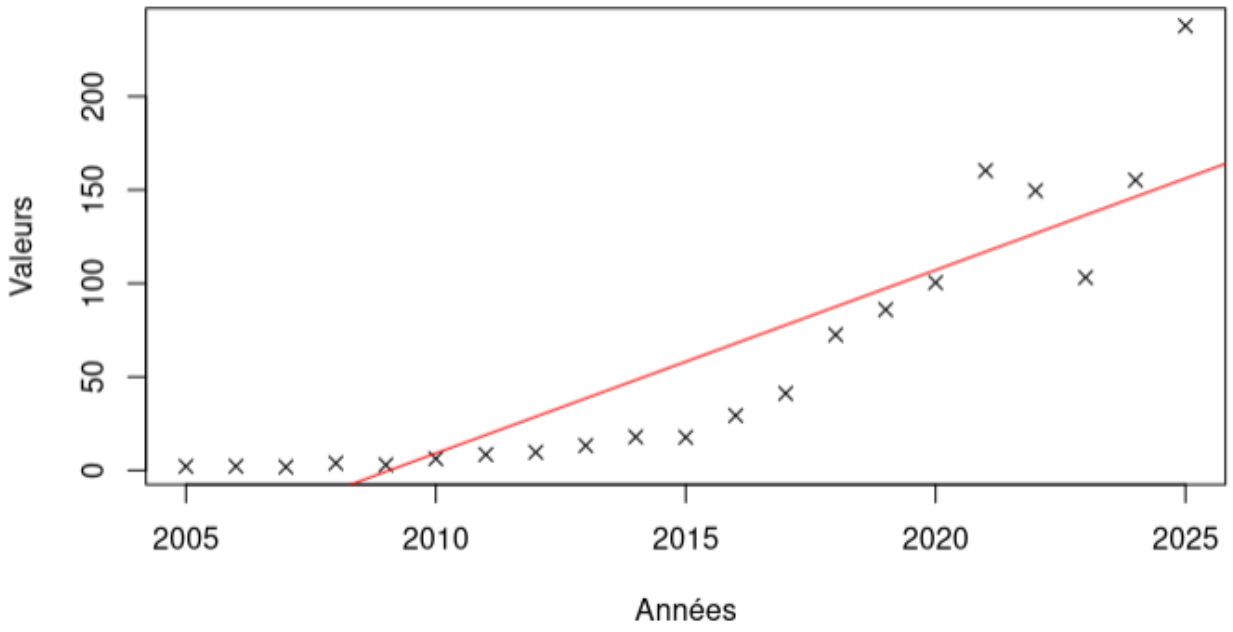
On étudie alors le modèle linéaire pour les données:

```
modreg=lm(Adj_Close ~ Date ,tab)
summary(modreg)
```

```
Call:
lm(formula = Adj_Close ~ Date, data = tab)

Residuals:
    Min      1Q  Median      3Q     Max
-40.452 -25.322  -6.718  22.059  81.546

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -19679.030   2394.846  -8.217 1.12e-07 ***
Date            9.795      1.189   8.242 1.07e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.98 on 19 degrees of freedom
Multiple R-squared:  0.7814,    Adjusted R-squared:  0.7699
F-statistic: 67.92 on 1 and 19 DF,  p-value: 1.074e-07
```

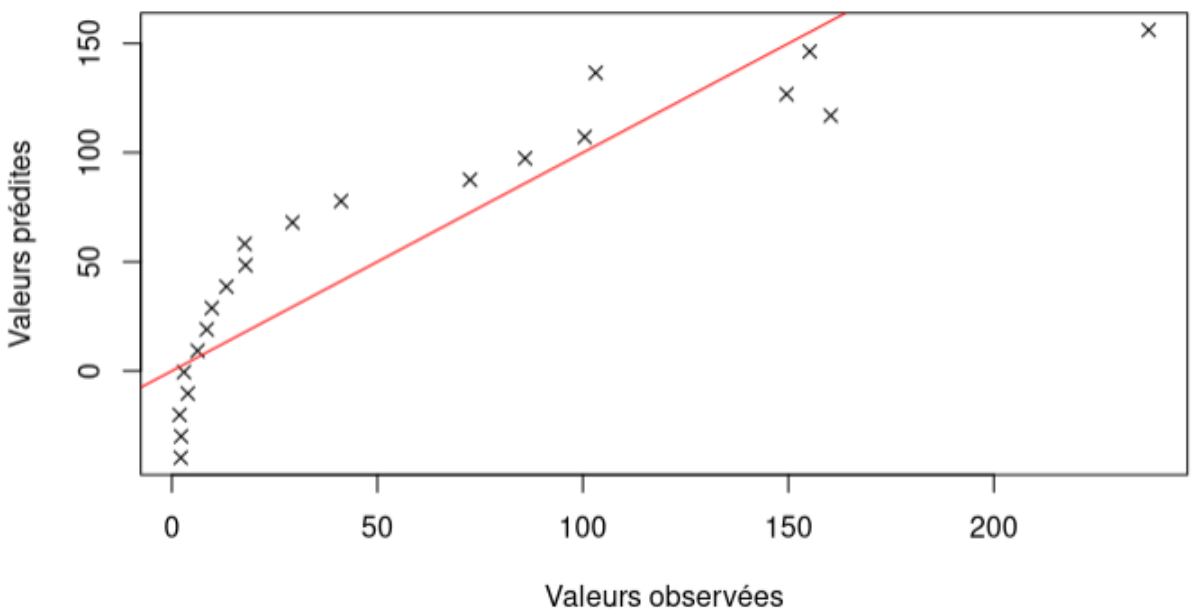
Figure 1: Ajustement du modèle linéaire sur les données d'Amazon

```
plot(tab$Date,tab$Adj\_Close,xlab="Années",ylab="Valeurs",pch=4)
abline(modreg,col="red")
```



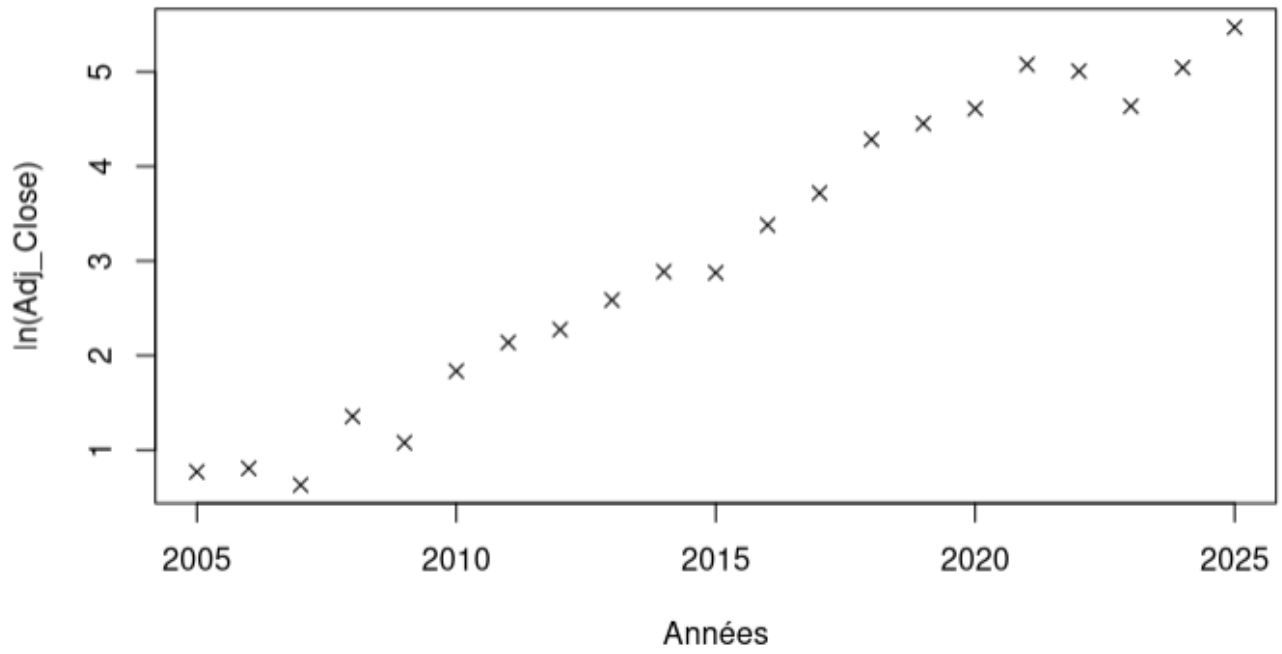

```
plot(tab$Adj_Close,modreg$fit,xlab="Valeurs observées",ylab="Valeurs prédites",pch=4)
abline(0,1,col="red")
```

On remarque sur ces graphiques que le modèle linéaire ne convient pas, notamment en vue du coefficient R^2 qui vaut 0.78. L'observation des données nous amène à penser qu'un modèle exponentiel pourrait convenir.



Nous allons donc travailler avec ln(Adj_Close):

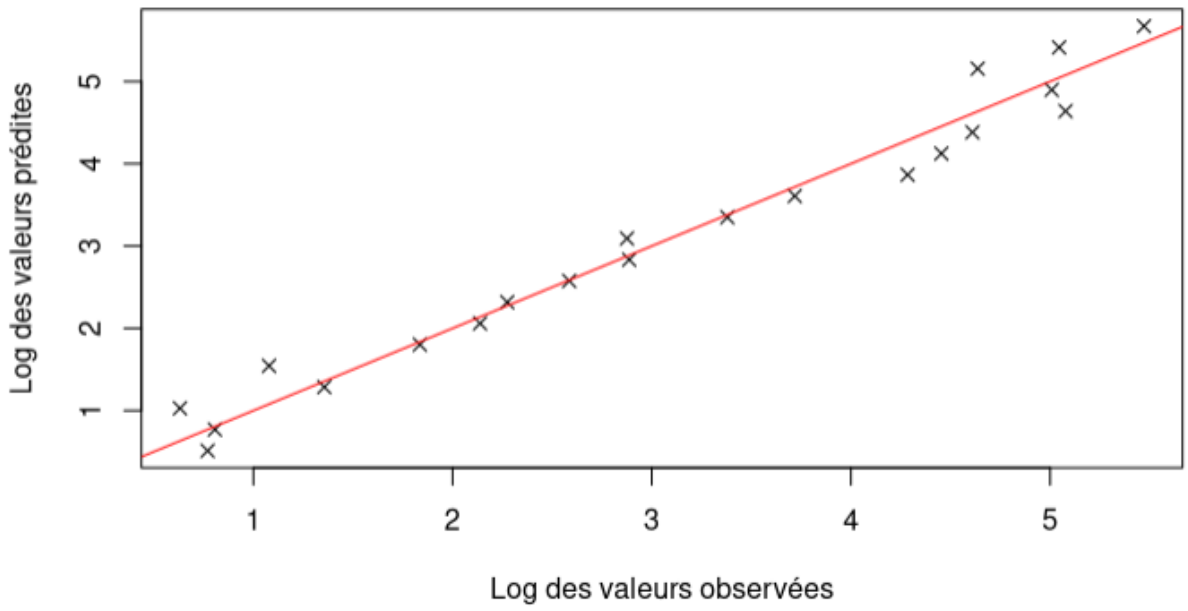
```
tab$Log_Adj_Close<-log(tab$Adj_Close)
plot(tab$Date,tab$Log_Adj_Close,xlab="Années",ylab="ln(Adj_Close)",pch=4)
abline(0,1,col="red")
```



```r
logmodreg=lm(Log_Adj_Close~ Date,tab)
summary(logmodreg)
```

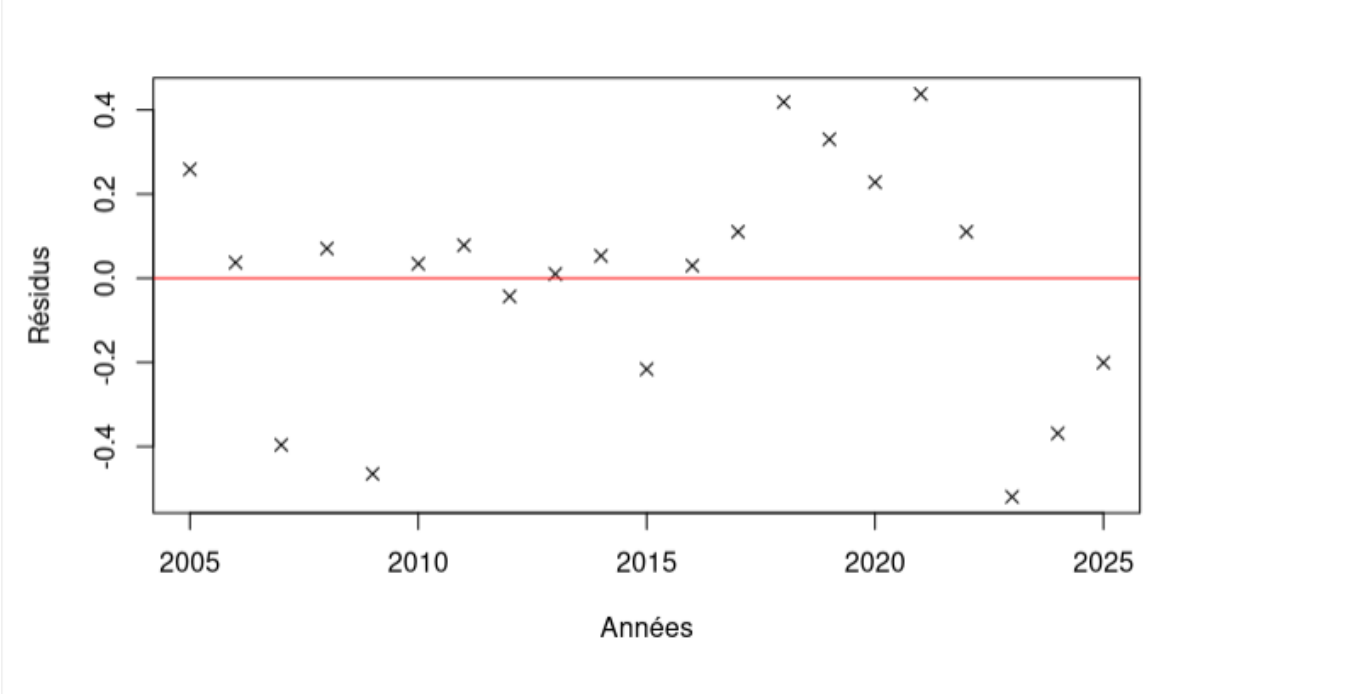
```
Call:
lm(formula = Log_Adj_Close ~ Date, data = tab)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51939 -0.20047  0.03721  0.11039  0.43774

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -516.79315   20.43333  -25.29 4.30e-16 ***
Date           0.25801    0.01014   25.44 3.85e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2814 on 19 degrees of freedom
Multiple R-squared:  0.9715,    Adjusted R-squared:   0.97
F-statistic: 647.4 on 1 and 19 DF,  p-value: 3.85e-16
```

```r
plot(tab$Log_Adj_Close,logmodreg$fit,xlab="Log des valeurs observées",ylab="Log des
↪  valeurs prédites",pch=4)
abline(0,1,col="red")
```



On remarque alors que les données suivent un modèle linéaire, avec R^2=0.97 De plus,les résidus correspondants sont:

```
plot(tab$Date,logmodreg$res,xlab="Années",ylab="Résidus",pch=4)
abline(h=0, col="red")
```



Ainsi, le modèle exponentiel correspond davantage aux données proposées. Cependant, on remarque des données légèrement moins cohérentes avec le modèle, correspondant à une expansion plus importante autour de l'année 2020. Cela peut notamment s'expliquer par l'impact de la crise sanitaire lors de cette période. Néanmoins le modèle choisit pourrait gagner en précision si on y ajoutait davantage de données sur des périodes plus courtes.

2 Application: Modèle linéaire pour la production d'électricité au Mexique

Le but est d'expliquer la production totale quotidienne d'électricité au Mexique à l'aide des variables contenues dans `Mexico_data.csv`.

2.1 Description des variables

- X0 : jour de l'année
- RH : humidité relative (%)
- SSRD : rayonnement solaire incident à la surface ($\mathrm{J.m^{-2}}$)
- STRD : rayonnement thermique incident à la surface ($\mathrm{J.m^{-2}}$)
- T2M : température moyenne quotidienne à 2m (°C)
- T2Mmax : température maximale quotidienne à 2m (°C)
- T2Mmin : température minimale quotidienne à 2m (°C)
- Covid : indice de rigueur COVID-19
- Holidays : jours fériés, 1 = jour férié, 0 sinon
- DOW : jour de la semaine, 0 = lundi, 1 = mardi, ...
- TOY : jour de l'année (1 à 366)
- Total : production quotidienne d'électricité (GWh)

2.2 Formule du modèle linéaire

Le modèle linéaire multiple choisi est :

$$\begin{aligned}\text{Total} = \theta_0 &+ \theta_1 \text{RH} + \theta_2 \text{SSRD} + \theta_3 \text{STRD} + \theta_4 \text{T2M} \\ &+ \theta_5 \text{T2Mmax} + \theta_6 \text{T2Mmin} + \theta_7 \text{Covid} + \theta_8 \text{Holidays} \\ &+ \theta_9 \text{DOW} + \theta_{10} \text{TOY} + \varepsilon\end{aligned}$$

où θ_0 est l'ordonnée à l'origine, $\theta_1, \ldots, \theta_{10}$ sont les coefficients associés à chaque variable explicative, et ε le terme d'erreur.

2.3 Code R utilisé

```r
# Charger les données
data <- read.csv("Mexico_data.csv")

# Identifier la variable cible et la première colonne (X0)
target_col <- "Total"
first_col <- names(data)[1]

# Exclure la cible et la première colonne des prédicteurs
predictors <- setdiff(names(data), c(target_col, first_col))
```

```r
# Construire la formule du modèle
form <- as.formula(paste(target_col, "~", paste(predictors, collapse = " + ")))

# Ajuster le modèle linéaire
model <- lm(form, data = data, na.action = na.omit)
summary(model)

# Extraire valeurs observées, ajustées et résidus
obs <- model$y
fit <- fitted(model)
res <- residuals(model)

# Index pour représenter les jours (pour éviter les erreurs de type)
x_index <- seq_len(nrow(data))

# Graphique Observé vs Ajusté
plot(x_index, data[[target_col]],
     xlab = "n° du jour",
     ylab = "Production totale d'électricité (GWh)",
     main = "Données observées vs modèle ajusté")
lines(x_index, fit, col = "red")

# Graphique des résidus vs valeurs ajustées
plot(fit, res,
     xlab = "Valeurs ajustées",
     ylab = "Résidus",
     main = "Résidus du modèle")
abline(h = 0, col = "red")
```

2.4 Résumé des données

Après avoir chargé les données et exécuté `summary(data)`, on obtient :

```
X0                  RH              SSRD               STRD
 Length:1461        Min.   :28.37   Min.   : 407045   Min.   :1019376
 Class :character   1st Qu.:50.12   1st Qu.: 719323   1st Qu.:1156350
 Mode  :character   Median :57.66   Median : 870071   Median :1235306
                    Mean   :57.55   Mean   : 869055   Mean   :1244835
                    3rd Qu.:64.55   3rd Qu.:1022083   3rd Qu.:1353442
                    Max.   :80.54   Max.   :1217895   Max.   :1431876
      T2M             T2Mmax          T2Mmin            Covid
 Min.   :11.23   Min.   :17.82   Min.   : 5.536   Min.   : 0.00
 1st Qu.:18.09   1st Qu.:25.27   1st Qu.:12.064   1st Qu.: 0.00
 Median :22.32   Median :28.72   Median :16.220   Median :33.33
 Mean   :21.27   Mean   :27.93   Mean   :15.623   Mean   :33.00
 3rd Qu.:24.63   3rd Qu.:30.83   3rd Qu.:19.747   3rd Qu.:63.89
 Max.   :27.39   Max.   :33.83   Max.   :21.841   Max.   :82.41
    Holidays            DOW          TOY             Total
 Min.   :0.00000   Min.   :0   Min.   :  1.0   Min.   : 578.5
 1st Qu.:0.00000   1st Qu.:1   1st Qu.: 92.0   1st Qu.: 816.0
 Median :0.00000   Median :3   Median :183.0   Median : 871.4
 Mean   :0.03012   Mean   :3   Mean   :183.1   Mean   : 880.7
```

```
 3rd Qu.:0.00000   3rd Qu.:5   3rd Qu.:274.0   3rd Qu.: 958.1
 Max.   :1.00000   Max.   :6   Max.   :366.0   Max.   :1107.4
```

2.5 Résumé du modèle linéaire

Après avoir ajusté le modèle linéaire multiple (`lm(form, data)`), on obtient :

```
Call:
lm(formula = form, data = data, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-170.43  -33.41    4.57   35.44  191.36

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.289e+02  9.101e+01   3.614 0.000312 ***
RH          -7.449e-01  4.701e-01  -1.585 0.113241
SSRD         1.738e-04  2.251e-05   7.722 2.12e-14 ***
STRD         4.174e-04  9.395e-05   4.442 9.58e-06 ***
T2M          2.277e-01  8.496e+00   0.027 0.978622
T2Mmax      -4.920e+00  4.348e+00  -1.131 0.258037
T2Mmin       6.313e+00  5.842e+00   1.081 0.280094
Covid       -2.867e-01  4.964e-02  -5.776 9.35e-09 ***
Holidays    -8.732e+01  8.057e+00 -10.838  < 2e-16 ***
DOW         -1.305e+01  6.835e-01 -19.090  < 2e-16 ***
TOY          5.072e-02  1.716e-02   2.955 0.003172 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51.95 on 1450 degrees of freedom
Multiple R-squared:  0.7035, Adjusted R-squared:  0.7014
F-statistic:   344 on 10 and 1450 DF,  p-value: < 2.2e-16
```

2.6 Visualisation des résultats

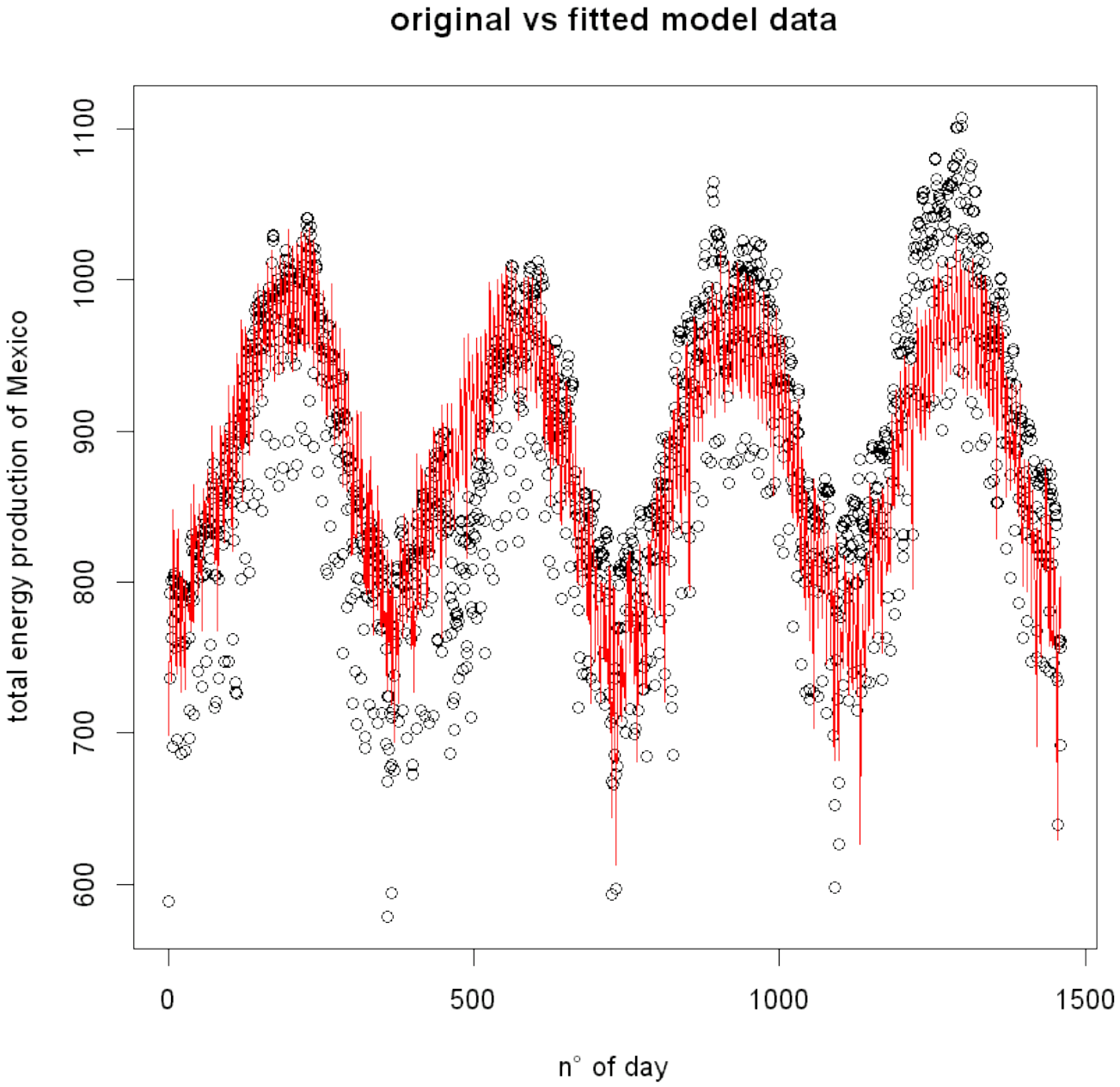


Figure 2: Production d'électricité observée et ajustée par le modèle

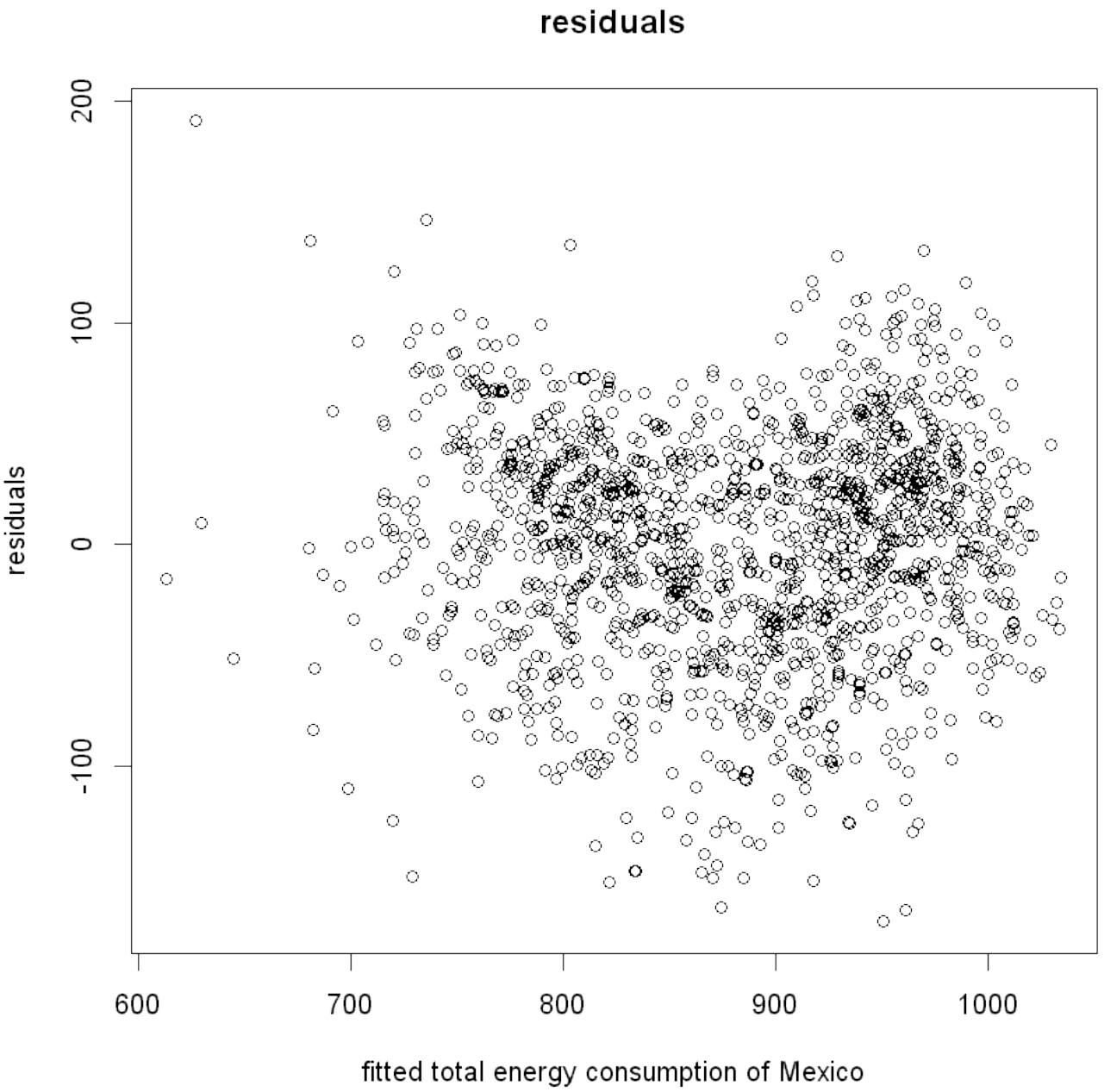


Figure 3: Résidus du modèle linéaire en fonction des valeurs ajustées

2.7 Commentaires et interprétation des résultats

- Le modèle linéaire multiple permet d'expliquer la production quotidienne d'électricité en fonction de plusieurs variables climatiques et sociales.
- D'après les valeurs de t et les p-values, les variables les plus significatives pour expliquer la production d'électricité sont **DOW** (jour de la semaine), **Holidays** (jours fériés), **SSRD** (rayonnement solaire) et **Covid** (indice de rigueur COVID-19). Ces variables ont des coefficients significatifs avec $p < 0.001$ et des valeurs absolues de t élevées.
- Les coefficients positifs ou négatifs indiquent le sens de l'effet sur la production totale : par exemple, **Holidays** et **DOW** ont un effet négatif important, indiquant que la production diminue pendant les jours fériés et certains jours de la semaine, tandis que **SSRD** et **STRD** ont un effet positif, montrant que la production augmente avec le rayonnement solaire et thermique.

- Les variables **T2M**, **T2Mmax** et **T2Mmin** ne semblent pas significatives dans ce modèle ($p > 0.1$), suggérant que la température quotidienne a moins d'impact direct sur la production totale dans cette période.
- Le modèle explique environ 70% de la variance totale de la production ($R^2 = 0.7035$), ce qui indique un ajustement raisonnable mais laisse encore de la variance inexpliquée, possiblement liée à d'autres facteurs non inclus dans le dataset.