Rapport de TP - Modèles de Régression Linéaire

Tommy MORALES & Aslan MARTIN

28 septembre 2025

Table des matières

1 Introduction

Ce rapport présente l'analyse de deux jeux de données dans le cadre du TP 1 du cours de Modèles de Régression Linéaire. Le premier exercice nous a laissé une certaine liberté quant aux données étudiées. Notre dévolu s'est porté sur l'évolution du revenu de Microsoft. Le second nous a cependant imposé l'explication de la consommation d'électricité au Mexique, et nous a fourni un jeu de données dans ce but.

2 IV : Croissance d'une entreprise

2.1 Analyse Exploratoire et Transformation des Données

2.1.1 Problématique et présentation des données

L'objectif est de modéliser l'évolution du revenu annuel de l'entreprise Microsoft sur une période de 20 ans (2005-2025). Les données brutes, extraites de Microsoft, sont composées de deux variables : l'année et le revenu correspondant en milliards de dollars.

2.1.2 Visualisation et nécessité d'une transformation

Une première visualisation du revenu en fonction de l'année montre une relation clairement non-linéaire. La croissance semble être de nature exponentielle.

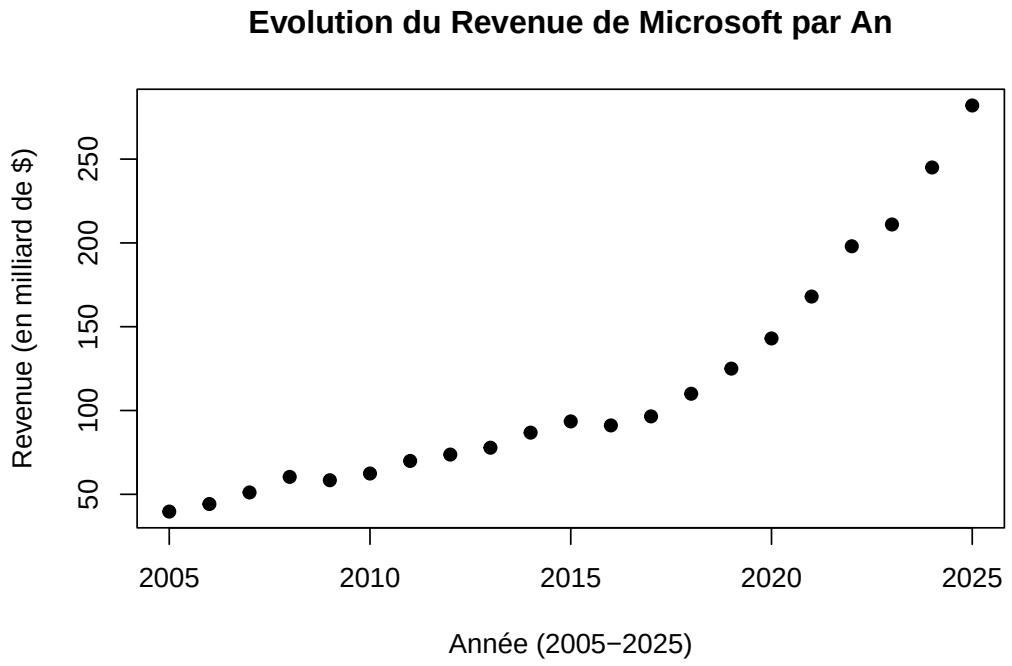

FIGURE 1 – Évolution du revenu brut de Microsoft (2005-2025).

Pour utiliser un modèle de régression linéaire, il est nécessaire de linéariser cette relation. Nous appliquons une transformation logarithmique sur la variable *Revenue*. Le nouveau nuage de points montre une tendance bien plus linéaire, justifiant l'approche de modélisation.

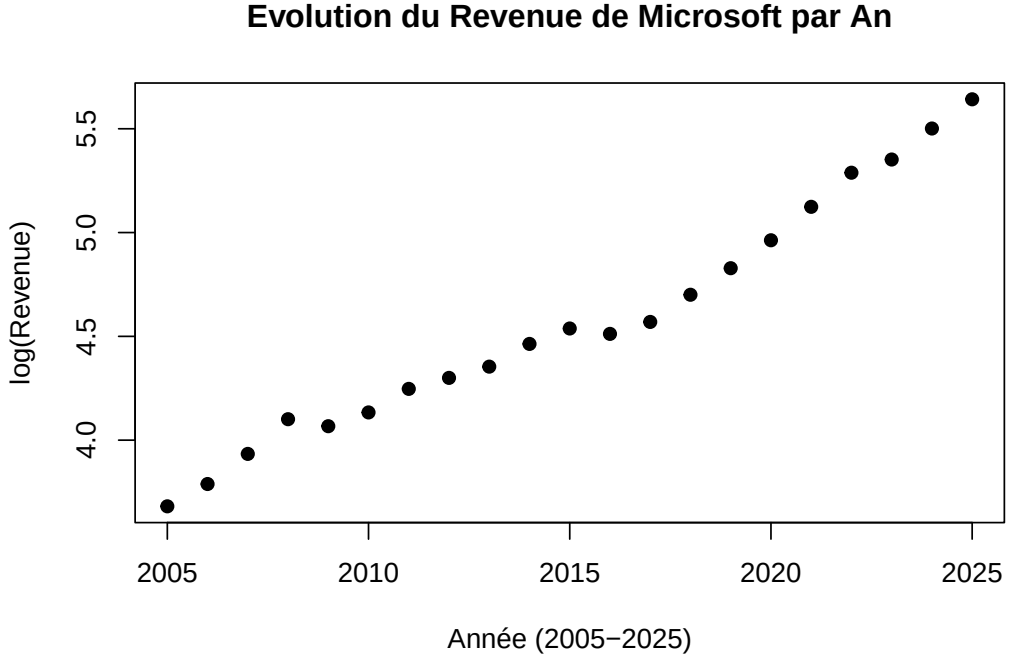

FIGURE 2 – Évolution du logarithme du revenu de Microsoft.

2.2 Modélisation et Interprétation

2.2.1 Ajustement du modèle par la méthode des Moindres Carrés Ordinaires (OLS)

Nous cherchons à entraîner un modèle de régression linéaire de la forme :

$$\log(\text{Revenue}_i) = \beta_0 + \beta_1 \times \text{Année}_i + \varepsilon_i$$

Les coefficients $\hat{\beta} = (\hat{\beta_0}, \hat{\beta_1})^T$ sont estimés en minimisant la somme des carrés des résidus. La solution générale est donnée par :

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

où X est la matrice formée par une colonne de 1 et une colonne contenant les années afin de former notre système d'équations, et y est le vecteur des valeurs de log(Revenue).

```
# Préparation de la matrice X et du vecteur y (y est gardé dans df_log[[2]] pour
# simplifier certains calculs par la suite)
df_log <- df
df_log[[2]] <- log(df[[2]])
X <- cbind(rep(1, 21), df_log[[1]])

# Calcul manuel des coefficients beta (On aurait pu utiliser lm, mais cette méthode
# permet d'accentuer les différentes approches explorées)
beta <- solve(t(X) %*% X) %*% t(X) %*% df_log[[2]]
```

```
beta
```

```
##                [,1]
## [1,] -176.42304469
## [2,]    0.08982566
```

Ce calcul manuel nous fournit les estimations des coefficients $\hat{\beta_0}$ et $\hat{\beta_1}$, ainsi qu'un début de profilage par le calcul de résidu qui a inspiré la méthode suivante.

2.3 Validation du Modèle et Analyse des Résidus

La figure suivante présente le graphe de l'erreur de notre modèle, normalisé et ramené à l'échelle de notre approximation, puis supperposé avec celle-ci. Ainsi, on peut voir quelles sections de notre exponentielle est la moins bien approchée.

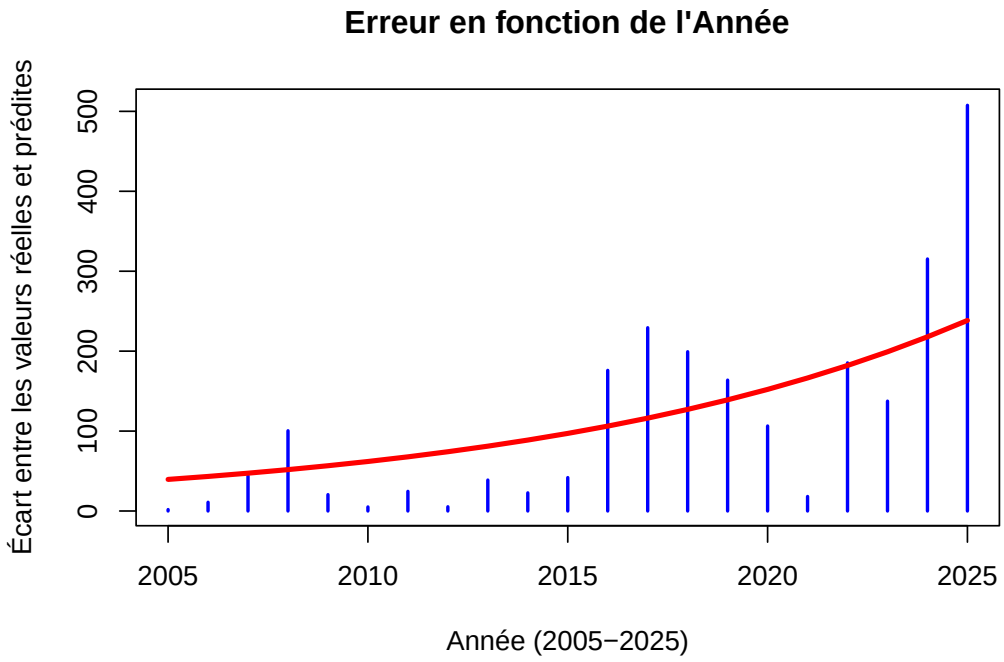

FIGURE 3 – Graphiques de diagnostic pour le modèle log-linéaire.

Notre plot démontre un modèle fidèle au départ, mais dont la validité diverge au fur et à mesure du temps, ce que nous confirme la supperposition de la prédiction avec les données.

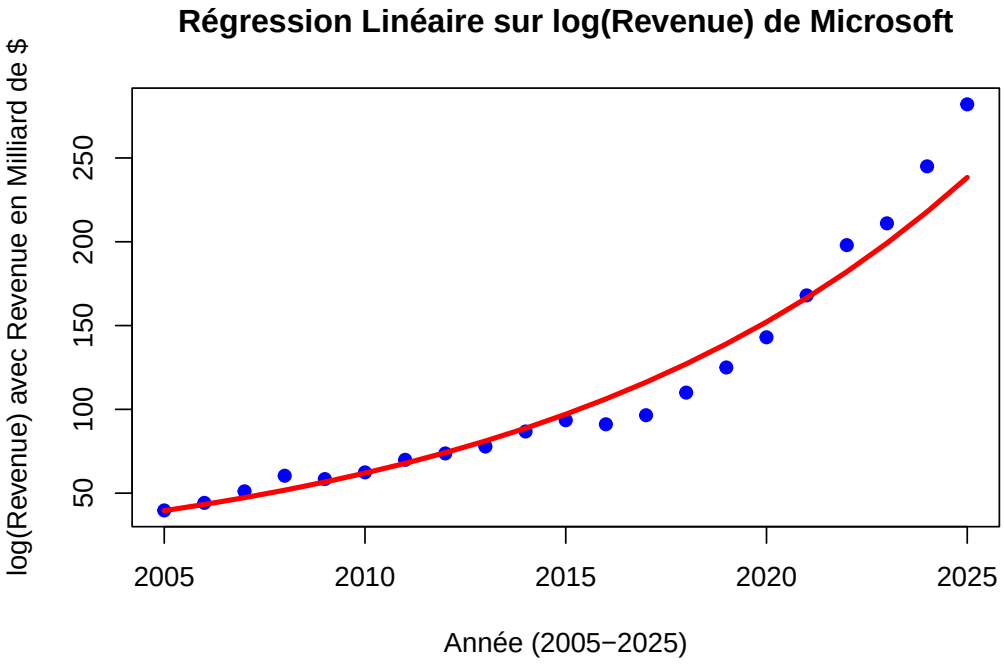

FIGURE 4 – Superposition des prédictions du modèle log-linéaire avec les données.

2.4 Exploration de la Seconde Méthode

Bien que notre premier modèle soit aussi fidèle que possible via une OLS de degré 1, il est évident que l'étude de nos données vise à prédire l'évolution future de l'entreprise, pas son passé. Pour cette raison, pour limiter la saturation causée par les données qui précèdent l'inflexion de l'exponentielle, nous avons supprimé une valeur sur deux du premier tiers des données.

```
df_treated <- df
for (i in 1:(nrow(df)/3)) {
    df_treated <- df_treated[-i, ]
}
```

Les résultats démontrent une amélioration de la prédiction au niveau des dernière valeurs, comme le démontre l'écart entre l'erreur des deux modèles.

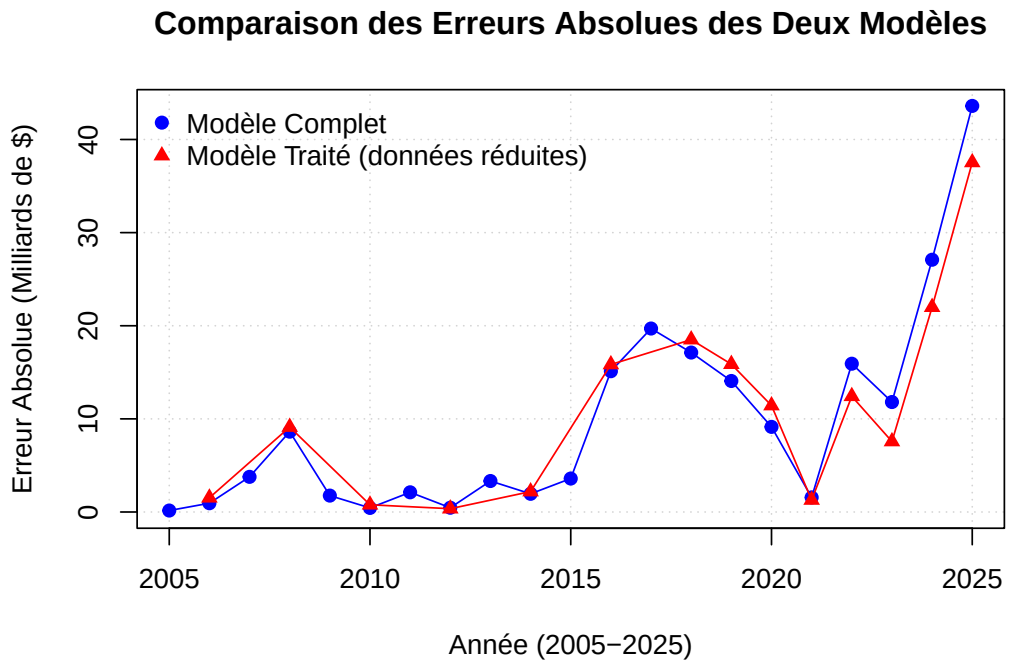

FIGURE 5 – Erreurs des deux modèles supperposées

Cependant, si la différente est visible, elle n'est pas nonplus remarquable et démontre que le problème principal vient des limites d'une approximation linéaire.

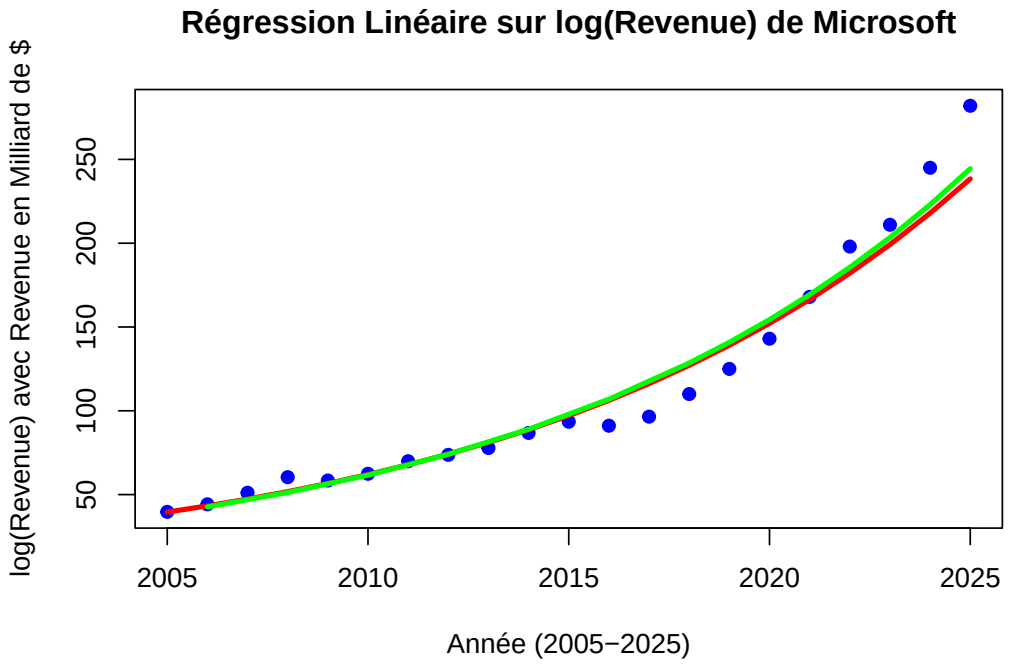

FIGURE 6 – Prédiction des deux modèles supperposées (Première version en rouge, seconde en vert) avec les données d'origine

2.5 Exploration de la Troisième Méthode

Si les limites du modèle précédents sont son manque de degrés de liberté qui ne permet par de représenter un système complexe, cette nouvelle méthode semble assez évidente: Nous allons cette fois-ci approximer grâce à un modèle parabolique plutôt que linéaire.

```
# Préparation de la matrice X et du vecteur y (y est gardé dans df_log[[2]] pour simplifier certains ca
Annee_normalisée = df_log[[1]] - mean(df_log[[1]])
X <- cbind(rep(1, 21), Annee_normalisée, Annee_normalisée^2)

# Calcul manuel des coefficients beta (On aurait pu utiliser lm, mais cette méthode
# permet d'accentuer les différentes approches explorées)
beta <- solve(t(X) %*% X) %*% t(X) %*% df_log[[2]]
```

```
beta_p
```

```
##                         [,1]
##                  4.501967347
## Annee_normalisee 0.089825661
##                  0.002009855
```

Le recentrage des données est essentiel. Sans cela, le compilateur détecte une corrélation linéaire entre les données, et une corrélation linéaire implique une infinité de solutions possibles, rendant $X^T * X$ non-invesible.

Recentrer les données amplifie l'aspect parabolique en accentuant l'écart entre chaque valeur.

Les résultat obtenus avec ce modèle sont nettement meilleurs, tant sur l'aspect visuel que par l'étude des résidus.

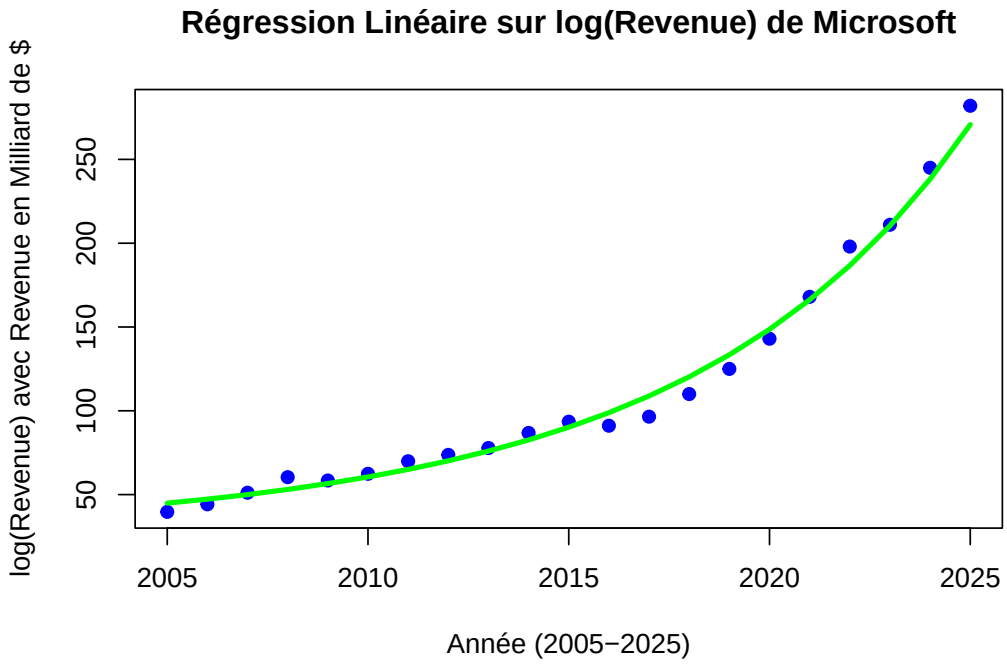

FIGURE 7 – Prédiction du nouveau modèle

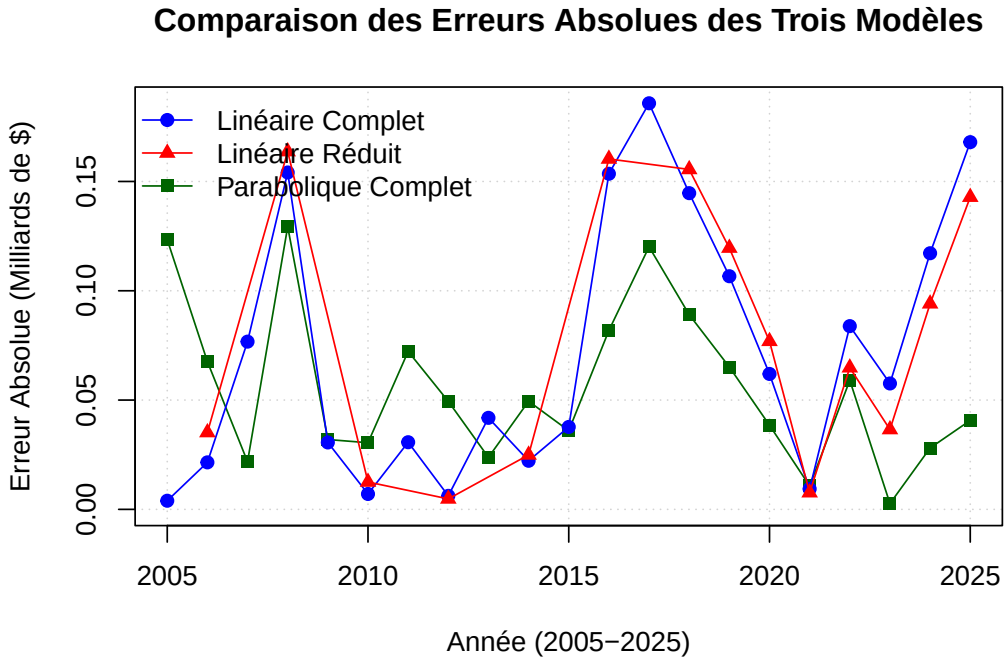

FIGURE 8 – Erreurs des différents modèles

Comme attendu, le modèle parabolique est nettement meilleur que prédécesseurs.

On pourrait continuer à rajouter des degrés de liberté, cependant, même si les résultats s'amélioreraient continuellement, cela se ferait au prix de beaucoup d'overfitting.

3 Étude V : Production d'Électricité au Mexique

3.1 Analyse Exploratoire des Données (EDA)

3.1.1 Problématique et présentation des données

L'objectif est d'expliquer la production d'électricité journalière (*Total*) au Mexique en utilisant un ensemble de variables météorologiques et temporelles. Une rapide analyse confirme qu'aucune donnée n'est manquante, qu'il n'y a pas d'outliers, et que les formats sont consistants. Le csv contient les données de 12 variables collectées pendant 1461 jours.

Avertissement: Ci-après, nous utiliserons *Total* autant comme "Production Energétique" que comme "Consommation Energétique" car la demande s'adapte toujours à l'offre.

3.1.2 Analyse univariée et bivariée

```
head(df_mexico)
```

```
##           X0       RH     SSRD    STRD      T2M   T2Mmax    T2Mmin Covid
## 1 2019-01-01 54.88692 556483.6 1095579 13.61629 20.25525  8.480851     0
## 2 2019-01-02 61.36585 551500.0 1148553 13.64353 19.71474  9.270146     0
## 3 2019-01-03 60.81884 643224.1 1086882 13.25186 20.81150  7.662988     0
## 4 2019-01-04 55.19776 661235.1 1069402 14.09160 22.81910  7.496520     0
## 5 2019-01-05 54.77641 609757.8 1110888 15.24040 23.64965  8.615337     0
## 6 2019-01-06 63.39149 548361.9 1207957 16.86465 23.43387 12.072207     0
##   Holidays DOW TOY    Total
## 1        1   1   1 588.3452
## 2        0   2   2 736.1832
## 3        0   3   3 793.0117
## 4        0   4   4 800.6797
## 5        0   5   5 762.7347
## 6        0   6   6 690.9642
```

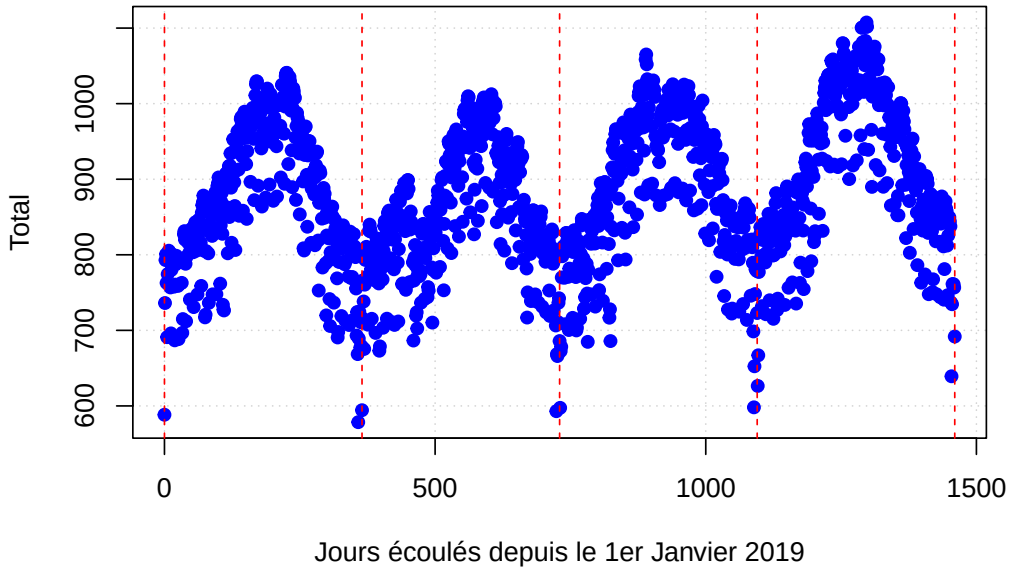

FIGURE 9 – Consommation totale au fil du temps.

Une tendance saisonnière claire se dégage, avec une consommation qui augmente jusqu'à un pic estival avant

de redescendre en hiver. Ce cycle s'explique probablement par l'usage accru de la climatisation en été, qui l'emporte sur les besoins en chauffage et éclairage en hiver, surtout dans le contexte climatique du Mexique.

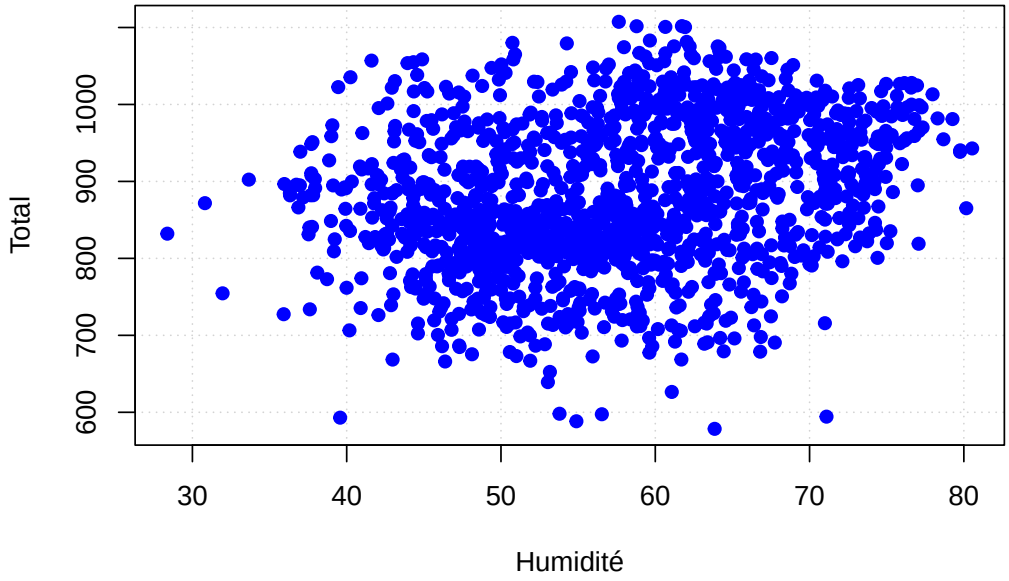

FIGURE 10 – Consommation vs. Humidité.

La relation entre l'humidité et la consommation est plus diffuse, bien que deux clusters de points puissent être distingués.

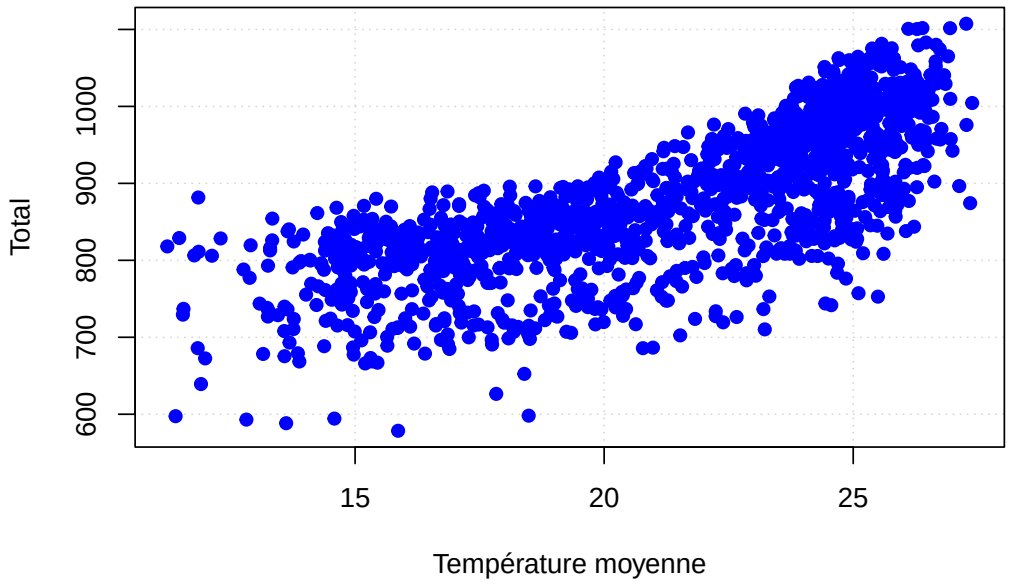

FIGURE 11 – Consommation vs. Température moyenne.

Comme attendu, la consommation est fortement corrélée avec la température. La relation semble même de nature exponentielle, ce qui pourrait indiquer un seuil de tolérance à la chaleur au-delà duquel l'usage de la climatisation devient systématique.

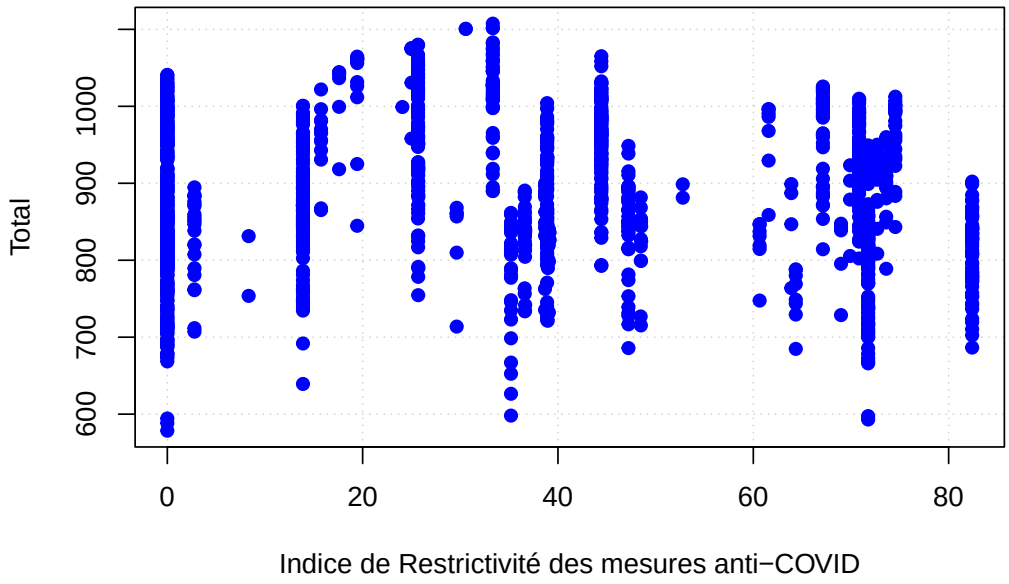

FIGURE 12 – Consommation vs. Indice COVID.

De manière inattendue, la relation globale avec l'indice de restrictivité COVID ne montre pas de tendance claire. L'impact pourrait cependant varier selon la saison.

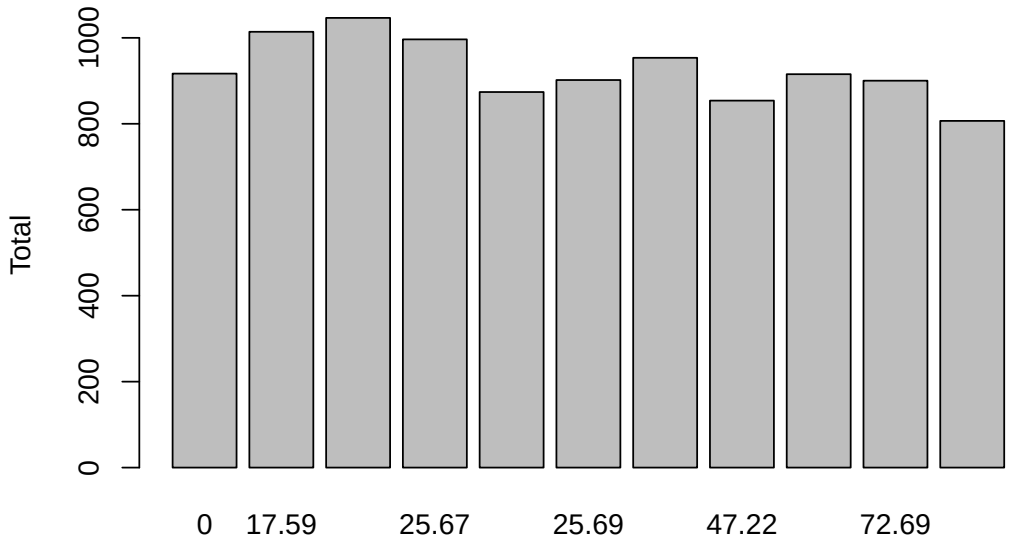

FIGURE 13 – Consommation vs. COVID en Été.

L'analyse saisonnière révèle une corrélation inverse en été : des restrictions plus fortes sont associées à une consommation plus faible. Ceci suggère que la baisse de l'activité économique (tourisme, industries, commerces) a un impact plus important que l'augmentation de la consommation domestique.

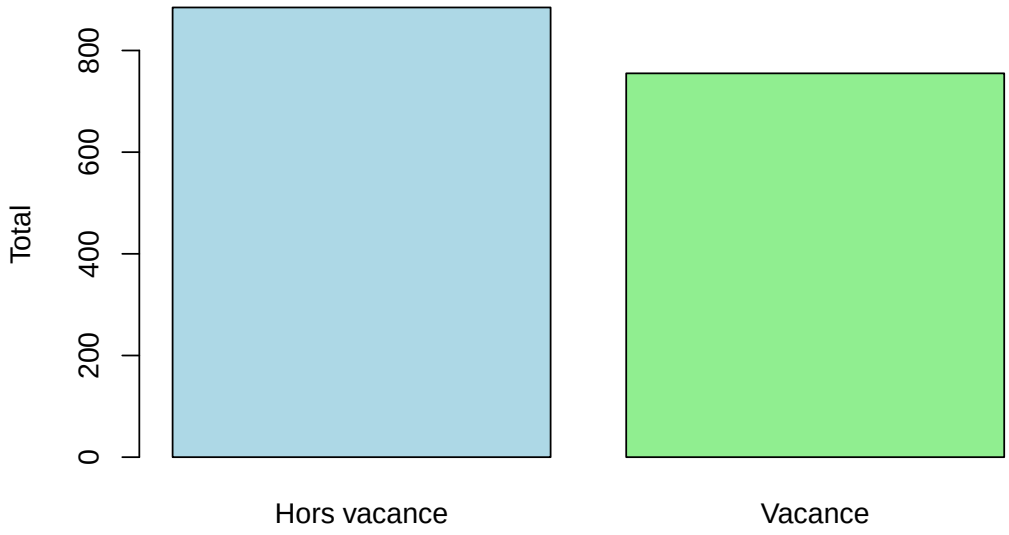

FIGURE 14 – Consommation moyenne pendant et hors vacances.

La consommation moyenne est plus faible durant les jours fériés que les jours ouvrés.

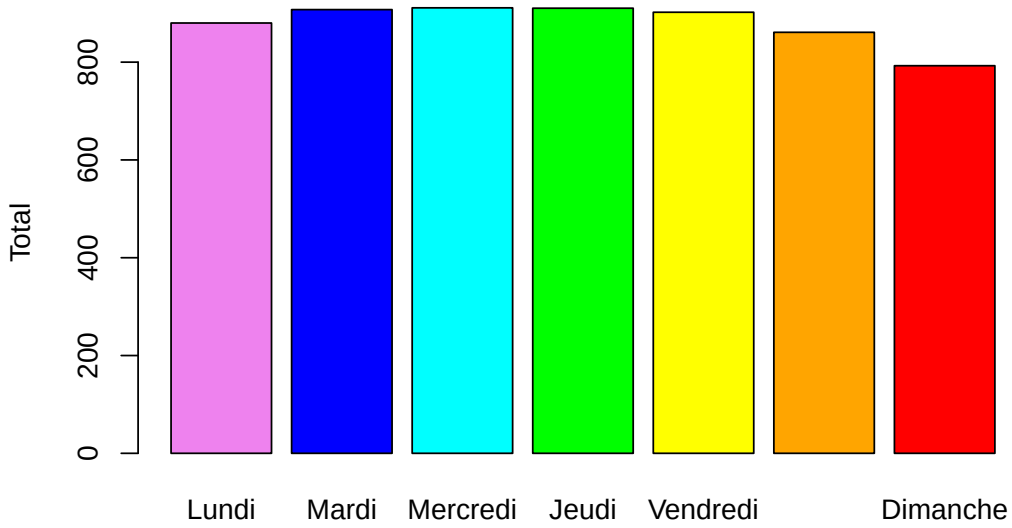

FIGURE 15 – Consommation moyenne par jour de la semaine.

Confirmant la tendance précédente, la consommation est plus élevée durant les jours de la semaine (lundi-vendredi) que le week-end, soulignant le poids de l'activité industrielle et commerciale.

3.1.3 Détection de la Multicolinéarité

```
# On ne prend pas la première colonne (dates) qui n'est pas numérique
cor(df_mexico[, -1])
```

```
##                  RH         SSRD        STRD         T2M       T2Mmax
## RH        1.000000000 -0.369741473  0.65033745  0.20927601  0.004801816
## SSRD     -0.369741473  1.000000000  0.35845077  0.73602599  0.828985571
## STRD      0.650337452  0.358450773  1.00000000  0.85875721  0.727680798
```

```
## T2M       0.209276006  0.736025989  0.85875721  1.00000000  0.971378355
## T2Mmax    0.004801816  0.828985571  0.72768080  0.97137836  1.000000000
## T2Mmin    0.422447721  0.587983694  0.95097300  0.97047525  0.889710714
## Covid    -0.115982225  0.138992727  0.06739498  0.14572241  0.163118527
## Holidays -0.007021789 -0.110012494 -0.09649889 -0.13061251 -0.140456069
## DOW      -0.018972929  0.007923398 -0.01478416 -0.01125942 -0.012348564
## TOY       0.428759293 -0.299716343  0.30201298  0.14392373  0.058712339
## Total     0.272554421  0.531909579  0.71562794  0.75636145  0.707368524
##               T2Mmin        Covid     Holidays          DOW          TOY
## RH        0.42244772 -0.115982225 -0.007021789 -0.018972929  0.428759293
## SSRD      0.58798369  0.138992727 -0.110012494  0.007923398 -0.299716343
## STRD      0.95097300  0.067394979 -0.096498890 -0.014784163  0.302012981
## T2M       0.97047525  0.145722408 -0.130612512 -0.011259420  0.143923735
## T2Mmax    0.88971071  0.163118527 -0.140456069 -0.012348564  0.058712339
## T2Mmin    1.00000000  0.113783648 -0.116875463 -0.016114345  0.227593952
## Covid     0.11378365  1.000000000 -0.026655244  0.001350033  0.040192537
## Holidays -0.11687546 -0.026655244  1.000000000 -0.044091342 -0.013465392
## DOW      -0.01611434  0.001350033 -0.044091342  1.000000000 -0.003547436
## TOY       0.22759395  0.040192537 -0.013465392 -0.003547436  1.000000000
## Total     0.76518565  0.010384629 -0.233082247 -0.273101633  0.123642205
##                Total
## RH        0.27255442
## SSRD      0.53190958
## STRD      0.71562794
## T2M       0.75636145
## T2Mmax    0.70736852
## T2Mmin    0.76518565
## Covid     0.01038463
## Holidays -0.23308225
## DOW      -0.27310163
## TOY       0.12364221
## Total     1.00000000
```

La matrice de corrélation confirme que les variables $T2M$, $T2Mmax$ et $T2Mmin$ sont quasiment identiques d'un point de vue informatif (cor > 0.85). Les inclure simultanément dans un modèle rendrait l'interprétation des coefficients impossible. Nous faisons donc le choix de ne conserver que $T2M$ comme représentant de la température.

3.2 Modélisation et Sélection de Variables

3.2.1 Premier modèle

Nous construisons un premier modèle linéaire multiple incluant les variables jugées pertinentes et non-colinéaires suite à notre EDA.

```
model1 <- lm(Total ~ RH + SSRD + T2M + Covid + Holidays + DOW, data = df_mexico)
summary(model1)
```

```
##
## Call:
## lm(formula = Total ~ RH + SSRD + T2M + Covid + Holidays + DOW,
##     data = df_mexico)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -164.509  -33.298    3.938   35.375  206.057
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.236e+02  1.602e+01  26.445  < 2e-16 ***
## RH           2.179e+00  2.453e-01   8.884  < 2e-16 ***
## SSRD         1.107e-04  1.887e-05   5.865 5.54e-09 ***
## T2M          1.341e+01  8.369e-01  16.018  < 2e-16 ***
## Covid       -2.556e-01  5.034e-02  -5.077 4.34e-07 ***
## Holidays    -8.420e+01  8.179e+00 -10.295  < 2e-16 ***
## DOW         -1.289e+01  6.936e-01 -18.586  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.9 on 1454 degrees of freedom
## Multiple R-squared:  0.6917, Adjusted R-squared:  0.6905
## F-statistic: 543.8 on 6 and 1454 DF,  p-value: < 2.2e-16
```

3.2.2 Interprétation du modèle final

Le modèle est globalement très significatif (p-value < 2.2e-16). Le R-squared ajusté de **0.6905** signifie que notre modèle explique environ **69%** de la variance de la consommation totale d'électricité.

Puisque toutes les variables sont significatives, aucune simplification n'est nécessaire. L'interprétation des coefficients nous apprend que, toutes choses égales par ailleurs :

- **(Intercept) :** La consommation de base, si toutes les autres variables étaient à zéro, est estimée à 423.6 GWh.
- **RH :** Une augmentation de 1% de l'humidité relative est associée à une augmentation de la consommation de 2.18 GWh.
- **SSRD :** Chaque unité supplémentaire de rayonnement solaire (en J.m-2) augmente la consommation de 1.107e-04 GWh.
- **T2M :** Une augmentation de 1°C de la température moyenne est associée à une hausse de 13.41 GWh.
- **Covid :** Chaque point d'augmentation de l'indice de restrictivité Covid est associé à une baisse de la consommation de 0.256 GWh.
- **Holidays :** Un jour férié réduit la consommation de 84.2 GWh par rapport à un jour ouvré non férié.
- **DOW :** Chaque jour qui passe dans la semaine (de Lundi=0 à Dimanche=6) est associé à une baisse moyenne de 12.89 GWh.

3.3 Validation et Conclusion de l'étude V

L'analyse des résidus du modèle final ne révèle pas de violation majeure des hypothèses de la régression linéaire (voir page suivante).

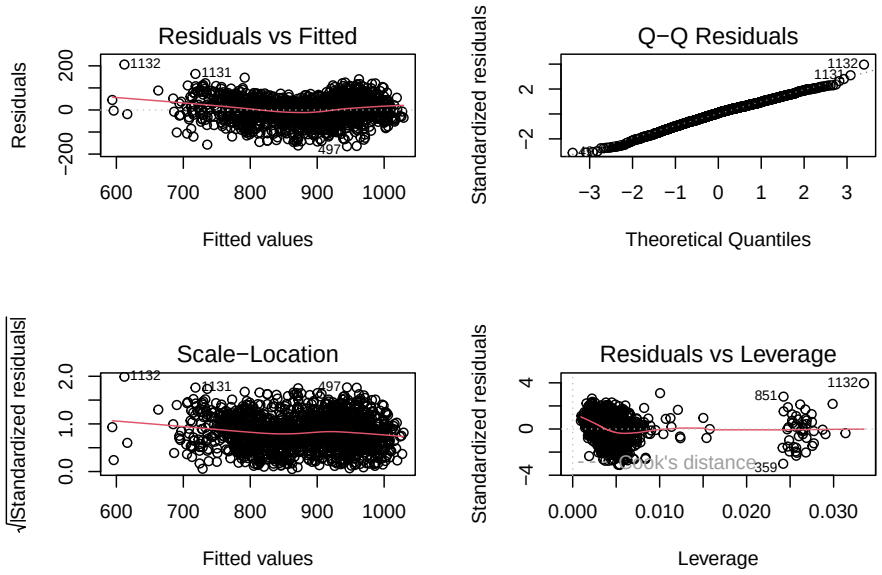

FIGURE 16 – Graphes des Résidus

En conclusion, notre démarche a permis de construire un modèle robuste expliquant la consommation électrique au Mexique. L'analyse exploratoire a été l'étape clé qui a guidé nos choix, notamment pour la gestion de la multicolinéarité. Le modèle final met en lumière que la consommation est un phénomène complexe dicté à la fois par le climat, le calendrier et le contexte socio-économique. Ce modèle simple mais puissant explique **69%** de la variance observée. "'