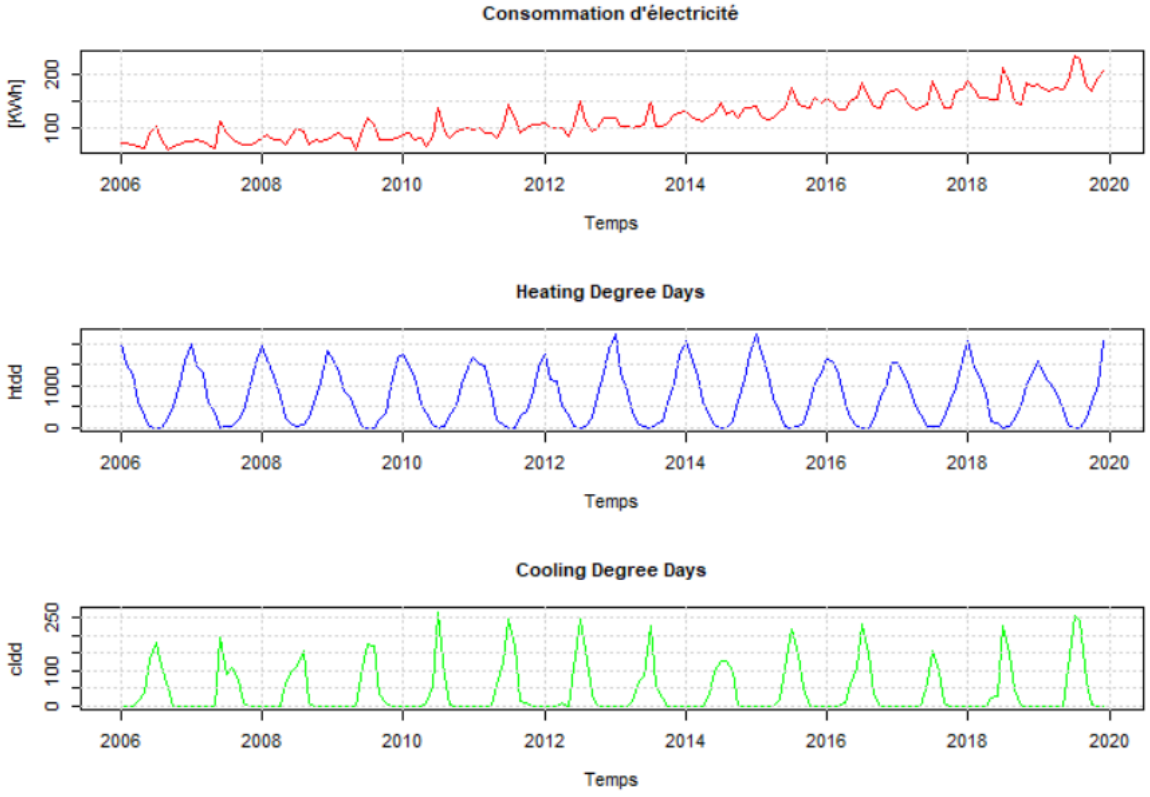
TP1 UP1 : Probabilités Avancées

Le problème posé est de prédire la consommation d'énergie pour l'année 2020 sur la base de prédicteurs possibles. Les données sur la consommation d'énergie de 2006 à 2019 sont fournies, ainsi que les mesures des degrés-jours de chauffage (**htdd**, *Heating Degree Days*) et des degrés-jours de refroidissement (**cldd**, *Cooling Degree Days*). À partir de ces données, une analyse en quatre étapes est proposée : examen graphique général des données, conception de modèles prédictifs sans tenir compte du temps, examen des résidus et enfin conception d'un modèle qui tient compte du temps et améliore la qualité de l'ajustement.

1ère partie : examen graphique des données

Les trois variables données sont les suivantes :



On peut voir que les deux variables explicatives possibles, **htdd** et **cldd**, ont un comportement stationnaire qui ne fluctue pas sur l'horizon temporel, alors que la variable dépendante a une tendance croissante dans le temps, ce qui montre qu'il n'y aurait pas assez d'informations avec les deux variables indépendantes pour prédire cette tendance. Cependant, il est nécessaire de corroborer cette hypothèse de manière pratique avec les données, et en outre d'avoir une idée de l'ampleur avec laquelle la consommation d'électricité peut être expliquée à partir ces variables.

2ème partie : Conception de régressions linéaires

Sur la base des modèles appris en classe, trois options avec des prédicteurs différents sont proposées. Il convient de noter que pour les trois modèles, l'horizon temporel de 2006 à 2018 a été utilisé pour entraîner le modèle, puis l'année 2019 a été utilisée pour vérifier si le modèle s'adapte aux données (en utilisant le RMSE).

Modèle 1 : Régression linéaire multiple sans interaction

Le premier modèle proposé est le modèle de régression multiple le plus simple, avec les variables explicatives données :

$$kWh = \beta_0 + \beta_1 htdd + \beta_2 cldd + \varepsilon, \qquad (1)$$

En exécutant un tel modèle de régression linéaire sur R, on obtient les résultats suivants :

```
Residuals:
    Min      1Q  Median      3Q     Max
-49.679 -27.743  -7.042  30.121  77.211

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       95.508844   6.329032  15.091  < 2e-16 ***
dataTotTrain$htdd  0.012965   0.005413   2.395   0.0178 *
dataTotTrain$cldd  0.231872   0.052948   4.379  2.2e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.85 on 153 degrees of freedom
Multiple R-squared:  0.1139,    Adjusted R-squared:  0.1023
F-statistic: 9.829 on 2 and 153 DF,  p-value: 9.641e-05
```

Le modèle obtenu est positif dans la mesure où les coefficients sont significatifs pour le modèle, bien que le R-carré ne soit pas assez bon car il est très éloigné de 1, ce qui montre que les erreurs entre les valeurs ajustées et les valeurs réelles sont trop importantes. Ceci est visible dans la valeur de la RMSE obtenue pour les données ajustées de 2019 :

```
> print(RMSE.reg21)
[1] 72.6775
```

Modèle 2 : Régression linéaire multiple avec interaction

Un terme d'interaction entre htdd et cldd est ajouté au modèle précédent pour voir si cette interaction permet un meilleur ajustement du modèle.

$$kWh = \beta_0 + \beta_1 htdd + \beta_2 cldd + \beta_3 htdd \cdot cldd + \varepsilon, \qquad (2)$$

Sur la base de ce nouveau modèle, les résultats suivants sont obtenus :

```
Residuals:
    Min      1Q  Median      3Q     Max
-48.994 -27.467  -6.302  29.075  74.894

Coefficients:
                                   Estimate Std. Error t value Pr(>|t|)
(Intercept)                       1.033e+02  7.607e+00  13.582  < 2e-16 ***
dataTotTrain$htdd                 7.707e-03  6.099e-03   1.264   0.2083
dataTotTrain$cldd                 2.268e-01  5.263e-02   4.310 2.92e-05 ***
dataTotTrain$htdd:dataTotTrain$cldd -1.581e-03  8.684e-04  -1.820   0.0707 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.61 on 152 degrees of freedom
Multiple R-squared:  0.1328,    Adjusted R-squared:  0.1156
F-statistic: 7.756 on 3 and 152 DF,  p-value: 7.454e-05
```

Bien que le R-carré se soit légèrement amélioré par rapport au modèle précédent, le coefficient de htdd n'est plus significatif, et la réduction de l'erreur standard résiduelle est marginale ; cette modification indique donc que l'ajout de l'interaction n'entraîne pas une réponse significativement meilleure. Ceci est corroboré par la RMSE obtenue pour 2019, qui est supérieure à celle du modèle précédent :

```
> print(RMSE.reg22)
[1] 73.45474
```

Modèle 3 : Régression linéaire multiple avec interaction et termes quadratiques

Lorsqu'aucune réponse positive n'est obtenue, on ajoute des termes quadratiques des variables explicatives initiales qui peuvent expliquer un pourcentage plus élevé des données.

$$kWh = \beta_0 + \beta_1 htdd + \beta_2 cldd + \beta_3 htdd \cdot cldd + \beta_4 htdd^2 + \beta_5 cldd^2 + \varepsilon, \qquad (3)$$

Ce nouveau modèle a permis d'obtenir les résultats suivants :

```
Residuals:
    Min      1Q  Median      3Q     Max
-47.918 -27.331  -7.006  28.793  74.283

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                        1.016e+02  1.438e+01   7.066 5.58e-11 ***
dataTotTrain$htdd                  1.205e-02  2.709e-02   0.445    0.657
dataTotTrain$cldd                  2.245e-01  2.311e-01   0.972    0.333
htddCarrée                        -2.016e-06  1.149e-05  -0.175    0.861
clddCarrée                         5.446e-05  8.961e-04   0.061    0.952
dataTotTrain$htdd:dataTotTrain$cldd -1.487e-03  1.031e-03  -1.441    0.152
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.82 on 150 degrees of freedom
Multiple R-squared:  0.1331,    Adjusted R-squared:  0.1042
F-statistic: 4.606 on 5 and 150 DF,  p-value: 0.0006146
```

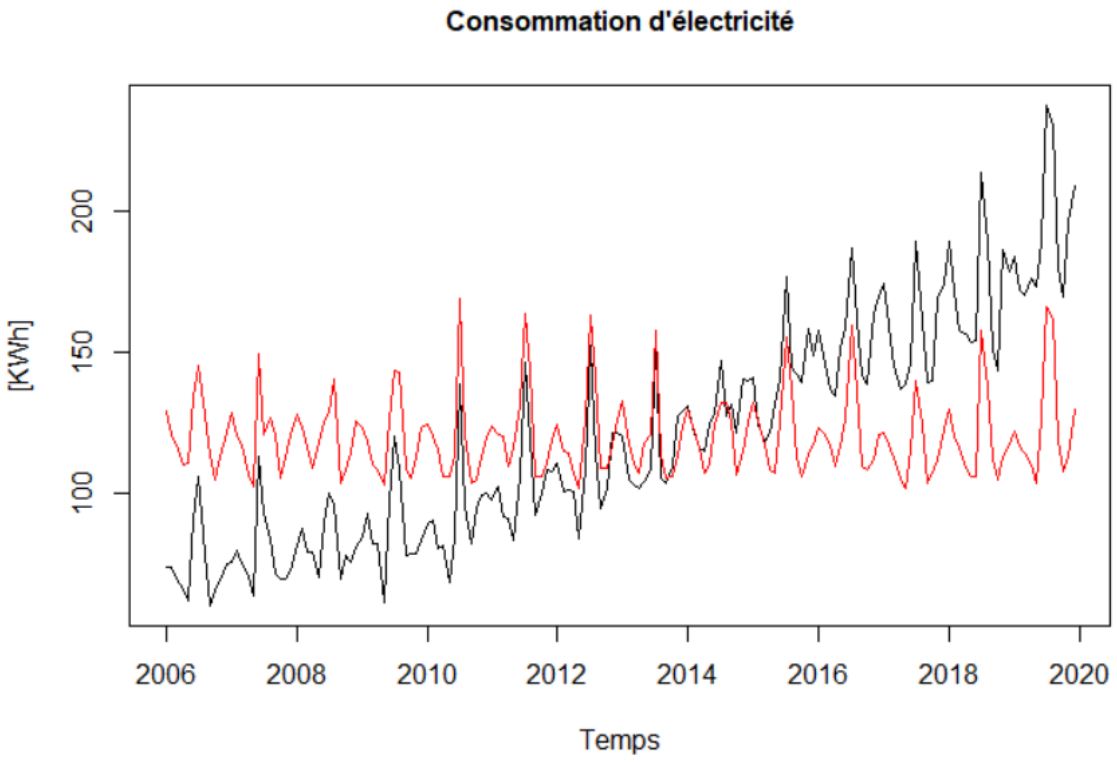
Nous constatons à nouveau une amélioration du R-carré, bien que cette fois l'effet sur les coefficients soit beaucoup plus prononcé et qu'aucun d'entre eux (à l'exception de l'ordonnée à l'origine) ne soit significatif. L'erreur standard résiduelle a augmenté et la RMSE pour l'année 2019 a encore augmenté, ce qui indique que le modèle 1 est le modèle (sans tenir compte du temps) le mieux adapté pour prédire ces données.

```
> print(RMSE.reg23)
[1] 92.8469
```

Puisque nous considérons le modèle complet comme celui de la régression linéaire multiple simple, nous exécutons à nouveau le modèle en considérant toutes les données, obtenant les résultats suivants :

```
Residuals:
   Min     1Q Median     3Q    Max
-55.41 -30.36  -7.33  31.86  83.52

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  96.821282   6.816420  14.204  < 2e-16 ***
dataTot$htdd  0.015920   0.005839   2.726   0.0071 **
dataTot$cldd  0.269873   0.055174   4.891 2.36e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.8 on 165 degrees of freedom
Multiple R-squared:  0.1289,    Adjusted R-squared:  0.1183
F-statistic:  12.2 on 2 and 165 DF,  p-value: 1.141e-05
```

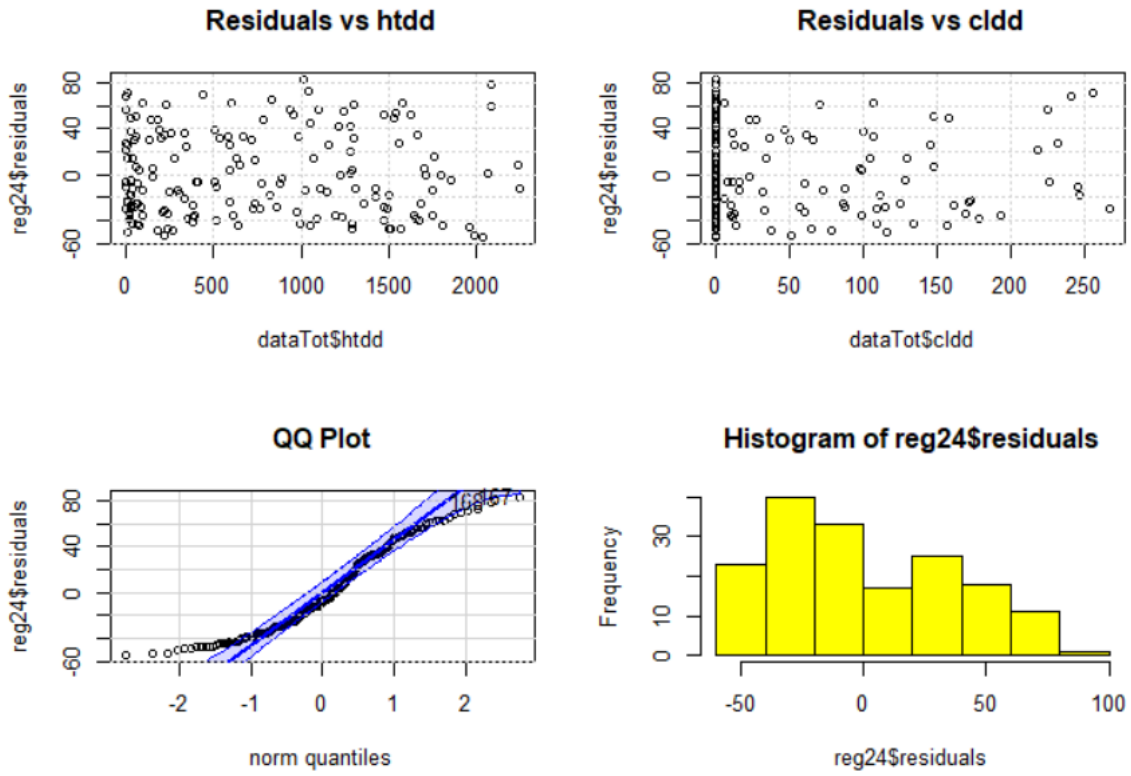
En outre, nous examinons également la comparaison entre les valeurs ajustées et les valeurs réelles :



On observe que les deux variables sont de bons prédicteurs, mais qu'elles ne tiennent pas compte de la saisonnalité des séries.

3ème partie : Analyse des résidus bruts

Les deux hypothèses relatives aux résidus qui permettent d'effectuer les régressions linéaires sont l'indépendance et la normalité. Les deux sont visibles dans les graphiques suivants :



Dans les deux premiers graphiques, où les résidus sont comparés aux variables indépendantes, on peut voir qu'il n'y a pas de relation évidente entre eux et les erreurs, ce qui indique que la première hypothèse est satisfaite. En ce qui concerne la deuxième hypothèse, deux graphiques sont réalisés : le premier est un Q-Q Plot, qui compare les quantiles théoriques d'une distribution normale avec les quantiles des erreurs. Dans ce cas, on peut voir que la plupart des données se situent dans l'enveloppe bleue, ce qui indique que, bien que la distribution ne soit pas exactement normale, il est possible de s'en approcher. Ceci est vérifié avec l'histogramme adjacent, qui concentre la majorité des données au centre et la minorité dans les queues, comme une distribution (approximativement) gaussienne. Cependant, il s'agit d'une hypothèse faible qui peut être améliorée en ajustant la régression plus étroitement aux données.

4ème partie : Modèles prenant en compte le temps

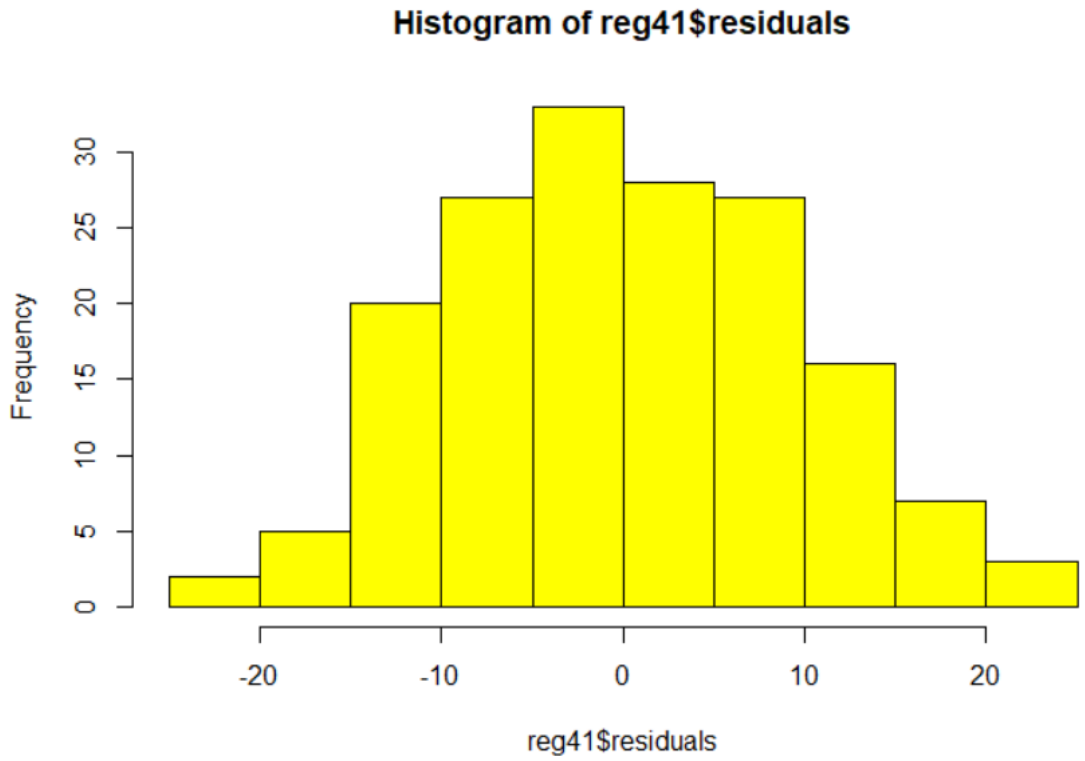
Comme nous l'avons vu dans les parties précédentes, il est nécessaire de prendre en compte le temps dans le modèle de prédiction pour améliorer l'ajustement. Le terme est ensuite ajouté à la régression finale de la partie 2 pour tenir compte du nouveau modèle :

$$kWh = \beta_0 + \beta_1 htdd + \beta_2 cldd + \beta_3 temps + \varepsilon, \qquad (4)$$

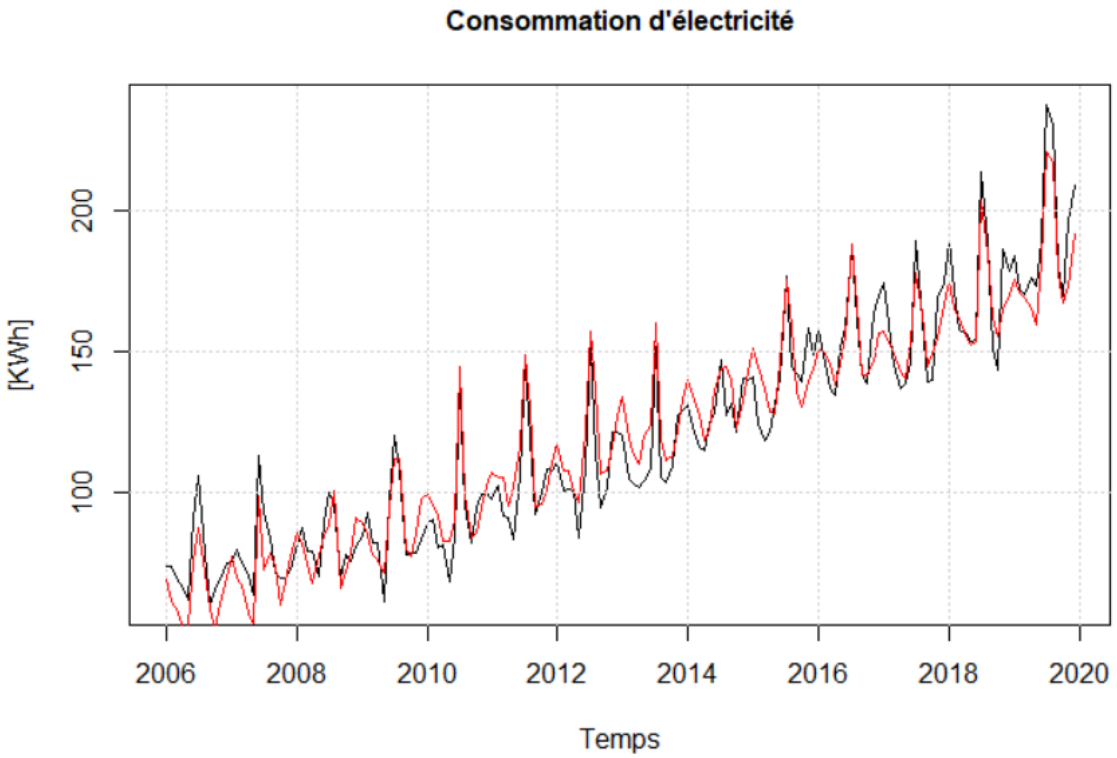
Où $temps$ est l'année [2006-2019] de mesure. Avec ces variables, on obtient les résultats suivants :

```
Residuals:
    Min      1Q  Median      3Q     Max
-21.559  -6.677  -1.151   6.776  23.061

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.745e+04  3.698e+02  -47.18   <2e-16 ***
dataTot$htdd  1.654e-02  1.527e-03   10.83   <2e-16 ***
dataTot$cldd  2.609e-01  1.442e-02   18.09   <2e-16 ***
dataTot$date  8.715e+00  1.837e-01   47.44   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.619 on 164 degrees of freedom
Multiple R-squared:  0.9408,    Adjusted R-squared:  0.9398
F-statistic: 869.3 on 3 and 164 DF,  p-value: < 2.2e-16
```

Il y a une amélioration significative par rapport aux modèles du point 2 : il y a un meilleur ajustement avec le R-carré et avec l'erreur résiduelle standard, les bêtas sont significatifs et même les erreurs s'ajustent mieux à une distribution normale :



On obtient ainsi les valeurs ajustées des séries temporelles suivantes :



Avec ce modèle, les valeurs pour l'année 2020 sont projetées selon les données fournies, obtenant les valeurs suivantes :

Mois	Valeur
Janvier 2020	187,24
Février 2020	179,30
Mars 2020	182,53
Avril 2020	171,05
Mai 2020	167,84
Juin 2020	179,41
Juillet 2020	189,49
Août 2020	209,96
Septembre 2020	174,18
Octobre 2020	173,08
Novembre 2020	182,40
Décembre 2020	191,44