TP1 UP1 : Probabilités Avancées

Examen graphique des données

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.

Conception de régressions linéaires

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
```

Aussi, on a le 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.

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

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
```

Le RMSE obtenue pour 2019 est supérieure à celle du modèle précédent :

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

Modèle 3 : Régression linéaire multiple avec interaction et termes quadratiques

On ajoute des termes quadratiques des variables explicatives initiales.

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

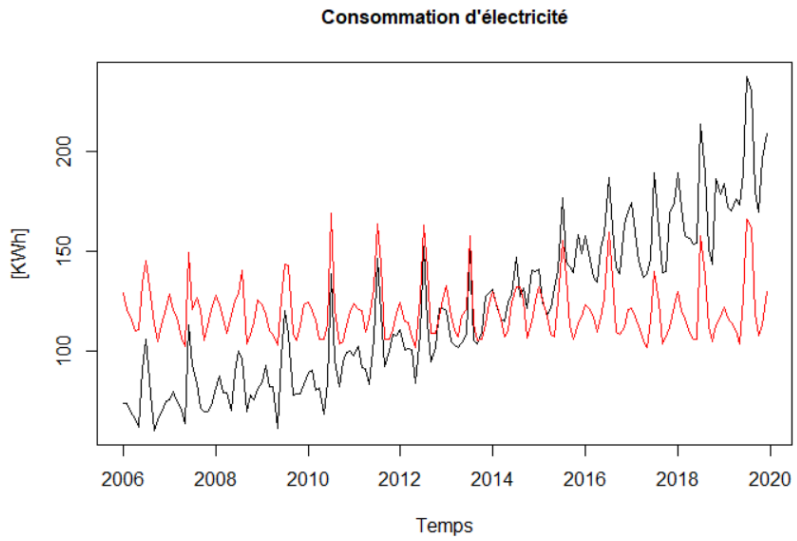
Ce nouveau modèle a permis d'obtenir les résultats suivants :

```
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
```

Le RMSE pour l'année 2019 a encore augmenté :

```
> 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 :



Modèle prenant en compte le temps

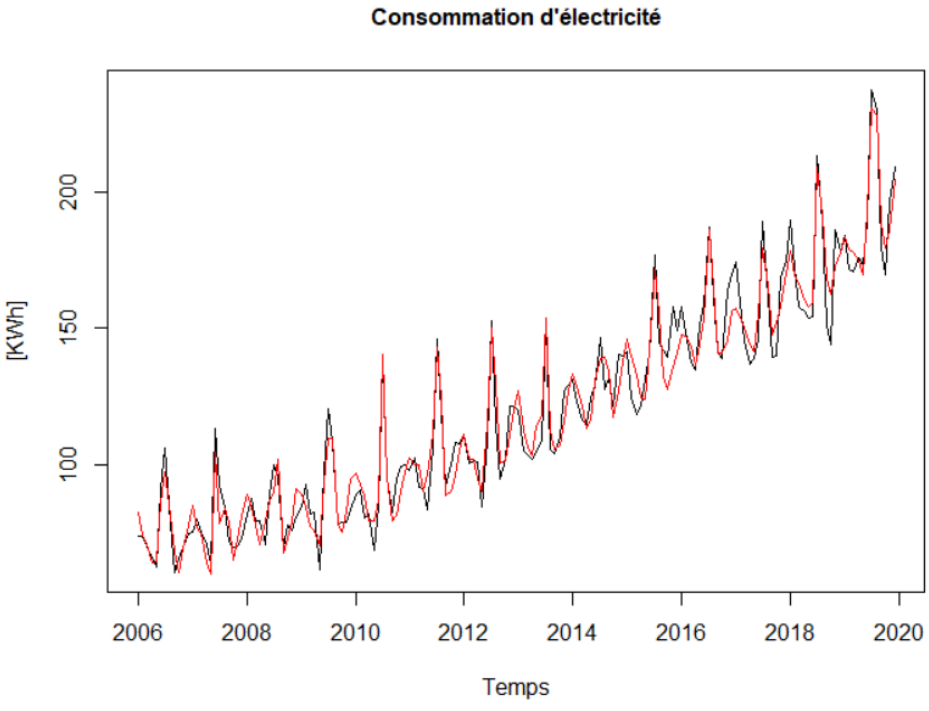
Il est nécessaire de prendre en compte le temps dans le modèle de prédiction pour améliorer l'ajustement. Alors, on peut ajouter pas seulement le terme du temps, mais de temps carré pour améliorer l'ajustement. Ces termes sont ensuite ajoutés à la régression finale de la partie 2 pour tenir compte du nouveau modèle :

$$kWh = \beta_0 + \beta_1 htdd + \beta_2 cldd + \beta_3 temps + \beta_4 temps^2 + \varepsilon, \qquad (4)$$

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

```
Residuals:
     Min       1Q   Median       3Q      Max
-18.7762  -5.4129   0.1527   4.6711  22.1115

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.608e+06  1.627e+05   9.882   <2e-16 ***
dataTot$htdd  1.623e-02  1.206e-03  13.459   <2e-16 ***
dataTot$cldd  2.583e-01  1.140e-02  22.665   <2e-16 ***
dataTot$date -1.606e+03  1.616e+02  -9.935   <2e-16 ***
tempsCarrée   4.010e-01  4.015e-02   9.989   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.599 on 163 degrees of freedom
Multiple R-squared:  0.9633,    Adjusted R-squared:  0.9624
F-statistic:  1070 on 4 and 163 DF,  p-value: < 2.2e-16
```

On obtient ainsi les valeurs ajustées des séries temporelles suivantes :



Repository : https://github.com/dasierra2021/TP-1-UP-1-Fondaments-Probabilistes