|  |  |
| --- | --- |
| Espoir BADOHOUN  Marion BOIVIN | Projet Logiciels statistiques |
|  |  |
| 01/05/2017 | Analyse du jeu de données : Facture mensuelle d’électricité d’un ménage en dollars sur 10 ans |
|  | Analyse du jeu de données : Facture mensuelle d’électricité d’un ménage en dollars sur 10 ans |

Projet Logiciels statistiques

Analyse du jeu de données : Facture mensuelle d’électricité d’un ménage en dollars sur 10 ans

1. **INTRODUCTION**

Le jeu de données nous présente la facture mensuelle d’électricité d’un ménage en dollars sur 10 ans (Janvier 1991 à Décembre 2000). On dispose donc de 120 observations et de 13 variables de base avec comme unité d’analyse un mois.

Le tableau suivant récapitule les informations concernant l’ensemble des variables. Nous en avons renommé certaines pour clarifier les données.

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| Numéro | Nominale | Observation |
| Année | Discrète | Année |
| Mois | Nominale | Mois |
| Facture | Continue | Montant |
| Tempe | Continue | Fahrenheit |
| HDD | Continue | Chauffage |
| CDD | Continue | Clim |
| Taille | Discrète | Personnes |
| Mètre | Binaire | Indicateur |
| Pmp1 | Binaire | Indicateur1 |
| Pmp2 | Binaire | Indicateur2 |
| Rider | Continue | Montant |
| Conso | Continue | Energie |

\*Les degrés jours mesurent la différence entre la température moyenne d’un jour par rapport à une température de la référence et expriment les besoins en chauffage ou réfrigération. Si la température du jour est inférieure à la valeur de référence, il s’agira de degrés jours de chauffage. A l’inverse si la température du jour est supérieure à la valeur de référence, il s’agira de degrés jours de réfrigération.

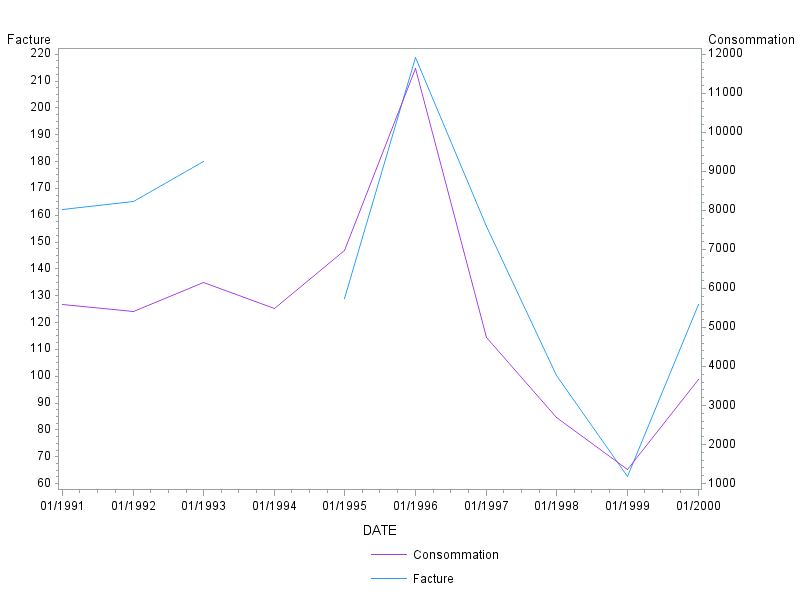
\*\*Les Riders sont des montants approuvés par la commission d’Etat pour l’électricité pour rembourser certains produits tels que les couts de combustibles ou les problèmes environnementaux.

L’objectif de notre étude est d’expliquer le montant de la facture en fonction des autres variables de la base de données à l’aide du logiciel R et SAS. Dans un premier temps nous analyserons le jeu de données afin de traiter les valeurs manquantes et aberrantes, puis nous ferons des analyses univariées et bivariées des différentes variables de la base. Enfin nous construirons des modèles de régressions simples et multiples afin d’expliquer au mieux le montant de la facture. Nous terminerons l’étude par une analyse en composantes principales et une classification hiérarchique afin de créer des groupes selon les variables présentes dans les données.

Tout d’abord on constate que le jeu de données comporte deux factures pour décembre 1993 et aucune en décembre 1992. En effet il y’a eu une erreur dans le processus de saisie ou de collecte de données.

Nous corrigeons dans un premier temps cette erreur-là. Par ailleurs on remarque d’une part qu’il y’a une valeur manquante en janvier 1994, et d’autre part que le montant de la facture est nul en aout 1999.

Discutons tout d’abord de la valeur manquante de janvier 1994. Pour ce mois-ci, aucune autre variable ne parait anormale, il n’y a eu aucun changement de pompe à chaleur ou de compteur, la température est relativement normale pour ce mois de l’année comparés aux autres années, tout comme la consommation d’énergie, comme on peut le voir sur le graphique ci-dessous qui représente la consommation d’énergie, et la facture en fonction de la date pour les mois de janvier de chaque année.



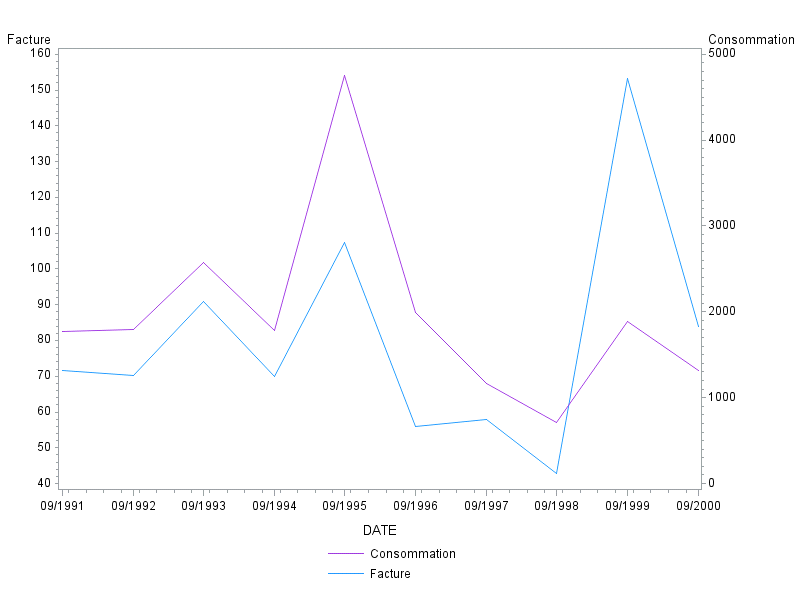
**Représentation de la facture et de la consommation pour les mois de janvier**

Comme on remarque également que les valeurs de la facture sont normales pour les mois de décembre 1993 et février 1994, le montant de la facture de janvier ne peut donc pas être dû à une erreur qui a différé le payement de janvier sur le mois précédent ou suivant. On peut supposer que cette valeur manquante est due à une erreur d’enregistrement ou de lecture du compteur. Il est dans tous les cas clairs que cette valeur manquante est anormale et qu’il est nécessaire de la remplacer par une valeur qui est normale ; en effet notre graphique révèle que la consommation entre janvier 1993, janvier 1994 et janvier 1995 est relativement similaire, de ce fait la valeur manquante de janvier 1994 est remplacée par la moyenne entre janvier 1993 et janvier 1995 (c’est-à-dire 154,57). Dans la suite de notre étude nous utiliserons cette valeur de référence.

Par contre le montant nul de la facture en août 1999, il coïncide avec le changement du compteur, en effet celui-ci a été changé le mois juste après. On remarque également sur le graphique ci-dessous que le montant de la facture le mois qui suit, c’est-à-dire en septembre 1999, est particulièrement élevé alors qu’on constate notamment que la consommation d’énergie est normale. De même on peut remarquer que la consommation d’électricité est exactement la même entre les mois d’août et de septembre 1999.

On peut donc supposer que le montant nul de la facture en Août est dû à un mauvais enregistrement d’un compteur défectueux et que le montant d’août a été payé en septembre.

La solution que l’on pourrait apporter est donc de diviser en deux la facture du mois de septembre et de la répartir à nouveau sur les mois d’août et de septembre. La facture de septembre étant de 153,32 dollars, on aura alors une facture 76,66 dollars pour chacun des mois.

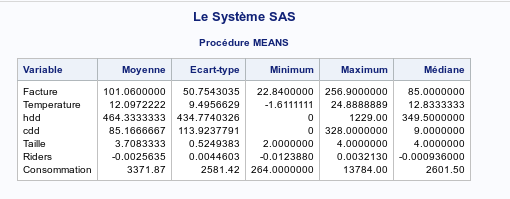


Représentation de la facture et de la consommation pour les mois de Septembre

**I) STATISTIQUE DESCRIPTIVES UNIVARIEES**

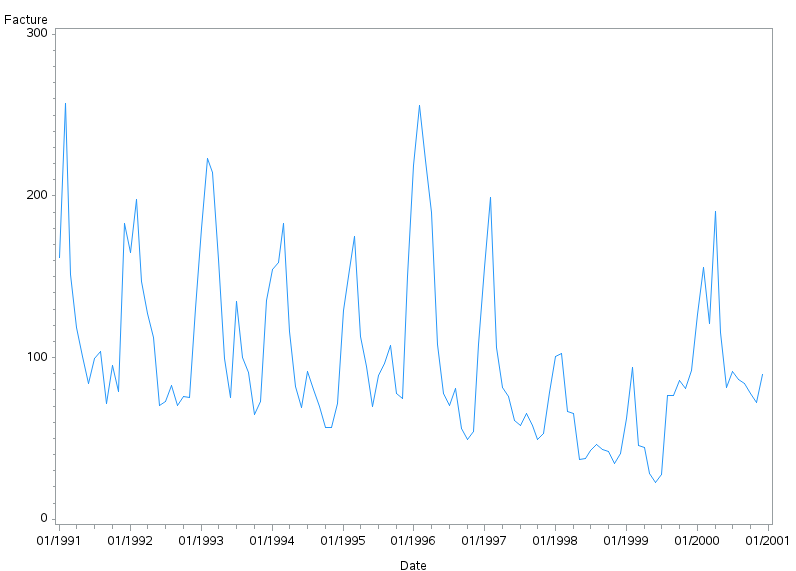
Dans cette partie, on s’intéresse à des statistiques descriptives des différentes variables quantitatives afin de résumer, synthétiser l’information contenue dans ces derniers.

Le tableau ci-dessous récapitule les indicateurs statistiques principaux (centrale et dispersion) des différentes variables quantitatives.



-On observe que le montant de la facture est varié entre 23$ (les mois d’été) et 257$ (les mois d’hiver) et en moyenne d’environ 101,06$, avec une variabilité importante vue que l’écart type est de 50,75. La médiane est de 85, c’est-à-dire que la moitié du temps la facture payée est en dessous de 85$, l’autre moitié au-dessus, ce qui indique pour la facture des valeurs élevées ayant donc pour impact d’augmenter la moyenne.

Sur le graphique ci-dessous qui représente le montant de la facture en dollars en fonction de la date à laquelle elle a été payée, on remarque une forte saisonnalité : la facture oscille entre des montants très importants les mois d’hiver et des montants plus faibles les mois d’été.

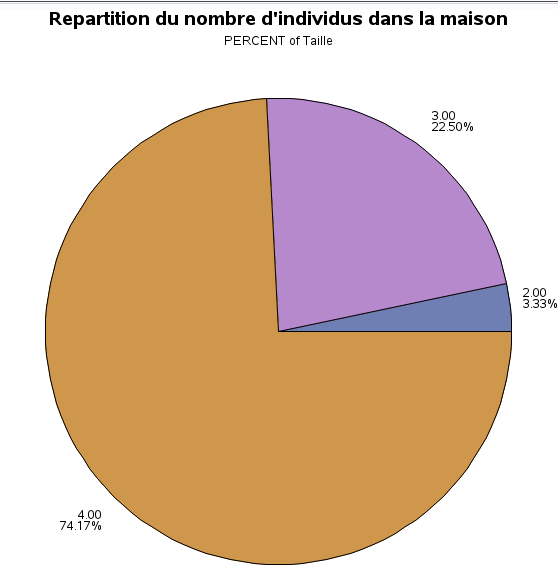


**Représentation de la facture en fonction de la date**

On a également la température que nous avons préalablement converti en degré Celsius oscille entre -2°C et 25°C avec une moyenne 12,1°C ce qui indique un climat relativement tempéré. La taille de la famille varie entre 2 et 4 personnes mais la majorité du temps la maison est habitée par 4 personnes (la moyenne étant de 3,7 avec un écart type faible et une médiane égale au maximum).

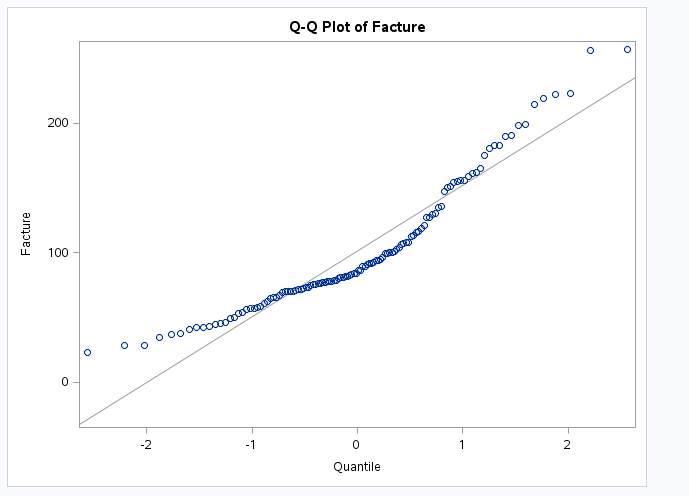
Pour les Degree Day, on remarque que dans les mois les plus froids il est nécessaire de réchauffer la maison de 1229 degrés par mois pour atteindre la température de référence journalière. A l’inverse dans les mois les plus chauds il faut rafraichir la maison de 328 degrés par mois pour atteindre la température de référence journalière. Globalement, le climat fait que cette maison nécessite plus de chauffage que de climatisation. La consommation varie entre 264 kWh et 13784 kWh avec un écart type de 2581. La moyenne est assez faible (plus proche du minimum que du maximum) et tirée vers le haut par de fortes valeurs de consommation tout comme pour la facture.

Le jeu de données comporte trois autres variables qui sont binaires : le changement du compteur et des deux pompes à chaleur. La première pompe a été changée en janvier 1997, la deuxième en octobre 1998 et le compteur en septembre 1999.



Sur le diagramme circulaire représentant la répartition du nombre d’individus dans la maison, on remarque que 3,33% du temps 2 personnes sont absentes. 22,5% du temps, une seule personne est absente, et le reste du temps 4 personnes occupent l’habitation.

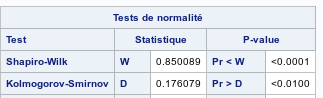
Effectuons maintenant des tests d’adéquations. On décide maintenant de tester la normalité du montant de la Facture. Comme on peut le voir ci-contre sur le QQ-plot sur sas, nous obtenons ceci :



**QQ-plot de la Facture**

Nous remarquons que la courbe de la facture n'est pas linéaire ne s’alignent pas parfaitement aux quantiles théoriques de la loi normale.

De plus, en faisant des tests de normalité, nous obtenons :



Nous pouvons voir les p-valeurs des tests de normalité de Shapiro-Wilk et de Kolmogorov-Smirnov sont très faibles, inférieur à 5%. A ce seuil-là, on rejette donc l’hypothèse nulle qui est la normalité de la variable Facture. On en conclu que le montant de la facture n’est pas réparti selon une loi normale.

On a également effectué des tests sur la valeur de la moyenne de la facture. On utilise pour cela le test de Student avec tout d’abord les hypothèses {H0 : mu=0 contre H1 : mu≠0}.

Les résultats sont résumés dans le tableau ci-contre.

C:\Users\BCSI\Desktop\pastedImage (3).png

On remarque que la p-valeur est très faible, donc le test est significatif. On décide de rejeter l’hypothèse nulle au seuil de 5%, c’est-à-dire de rejeter l’hypothèse que la moyenne de la facture est de 0 $.

On teste également les hypothèses H0 : mu=100 contre H1 : mu≠100.

On obtient les résultats suivants :

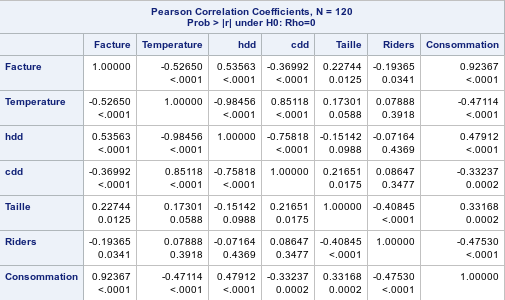
C:\Users\BCSI\Desktop\pastedImage (5).png

Le test n’est pas significatif pour le seuil de 5% :

on décide donc de ne pas rejeter l’hypothèse nulle c’est -à- dire d’accepter le fait que la moyenne de la facture soit proche de 100$.

**II Liaison entre les variables du jeu de données**

Dans le but d'anticiper et de pouvoir estimer les factures d'électricité on dresse un tableau de corrélations en étudiant le coefficient de corrélation linéaire entre les différentes variables quantitatives.



Lorsque la p-valeur est supérieure à 5% on statue sur la non corrélation des variables. Dans le cas contraire on statue sur une corrélation, de plus si la p-valeur est inférieure à 0,01 % on note une corrélation élevée

En récapitulatif on remarque :

- une corrélation élevée entre les variables facture et températures ainsi qu'entre le HDD et la consommation.

- une corrélation entre les variables Taille et Riders malgré un coefficient plus faible ce qui est logique vu que le Riders dépend de la taille de la famille

- pour toutes les variables on rejette l’hypothèse nulle d’absence de corrélation linéaire entre la facture et la variable concernée. Au risque de 5% on accepte donc l’existence d’un lien linéaire entre la facture et chacune de ces variables.

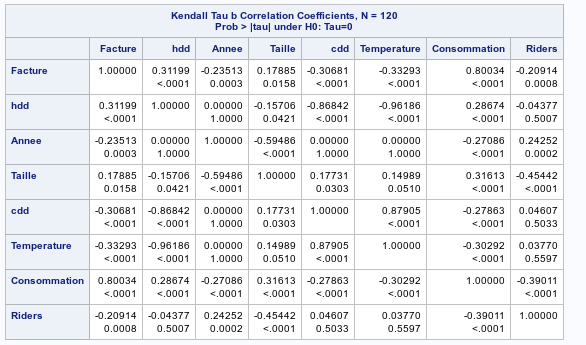
Dans ce tableau on remarque que le test de Pearson est non significatif au risque 5% entre les variables :

-Taille et température ainsi que taille et HDD

-riders avec les variables HDD, CDD et température

Pour ces couples de variables on ne rejette donc pas l’hypothèse nulle : On accepte l’absence de corrélation linéaire entre elles.

On s'intéresse maintenant au test de Kendall pour les mêmes variables :



Ce test confirme les conclusions faites après la lecture des résultats du test de Pearson.

On effectue une analyse de la variance à un facteur contrôlé sur la variable année par rapport aux saisons.

Auparavant il est nécessaire de vérifier que les années suivent une loi normale sur chacun des sous échantillons selon les différentes saisons, et que ces sous-échantillons ont tous même variance. Le test de Shapiro-Wilk nous donne les résultats suivants.

|  |  |  |
| --- | --- | --- |
|  | **Statistique de test** | **p-valeur** |
| **Printemps** | 0,94 | 0,09 |
| **Été** | 0,94 | 0,09 |
| **Automne** | 0,94 | 0,09 |
| **Hiver** | 0,94 | 0,09 |

Les p-valeurs ne sont pas significatives pour chacun des 4 différents tests au niveau 5%. On ne rejette donc pas l’hypothèse nulle dans les 4 cas, c’est-à-dire qu’on accepte l’hypothèse de normalité de la variable Année pour les 4 sous échantillons.

Le test de Bartlett sur l’égalité des variances nous fournit les résultats suivants.

|  |  |  |  |
| --- | --- | --- | --- |
| **Test de Bartlett** | **DDL** | **Statistique de test** | **p-valeur** |
| **Saison** | 3 | 0 | 1 |

Le test n’est pas significatif au risque de 5%, on ne rejette donc pas l’hypothèse nulle qui est l’égalité des variances. Au risque de 5% on en conclut que les variances de la variable année sont égales sur les 4 sous échantillons.

On peut à présent effectuer une analyse de la variance sur la variable année par rapport aux saisons. On obtient les résultats suivants (la statistique de test suit une loi de Fisher).

|  |  |  |  |
| --- | --- | --- | --- |
| **Test de Fisher** | **DDL** | **Statistique de test** | **p-valeur** |
| **Saison** | 3 | 0,31 | 0,82 |

Le test n’est pas significatif, on ne rejette donc pas l’hypothèse nulle qui est l’indépendance de la variable Année et des Saisons. Au risque de 5%, il n’existe donc pas de différence significative de la distribution entre les différents sous-échantillons.

On effectue ensuite un test du Khi² entre la saison et le nombre de personnes dans la maison. On remarque que les effectifs observés ne sont pas très éloignés des effectifs attendus. Par exemple en hiver si les deux variables étaient indépendantes, on attendrait qu’il y ait 4 personnes dans la maison 22,25 fois sur 30, or on en compte 20 dans le jeu de données. Cependant la p-valeur observée pour ce test est de 0,06 ce qui nous amène à rejeter l’hypothèse nulle d’indépendance des deux variables au seuil de 5%. On suppose donc que les deux variables sont dépendantes avec un risque de 5% de se tromper. Cependant si on avait pris un seuil un peu plus grand, 10% par exemple, on aurait accepté l’indépendance des variables.

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistique** | **DDL** | **Valeur** | **p-valeur** |
| **Khi-2** | 6 | 11,86 | 0,06 |

**III) Régression linéaire simple**

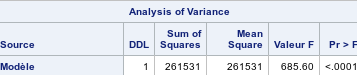
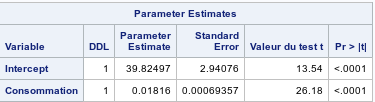
La **régression** linéaire est une modélisation linéaire qui permet d'établir des estimations dans le futur à partir d'informations provenant du passé. Dans ce modèle de **régression** linéaire, on a plusieurs variables dont une qui est une variable explicative et une autre qui est la variable à expliquer.

Dans cette partie, on s’intéresse à un modèle de régression simple entre la variable à expliquer, la facture et la variable explicative, la consommation.

On obtient alors le modèle suivant :

**Facture = 13,54 + 26,18 \* Consommation**

SAS donne les estimations des paramètres (coefficients du modèle) et estime la pertinence des variables dans le modèle. La procédure SAS « proc reg » permet les tests de Fisher et de Student sur la nullité des paramètres du modèle :

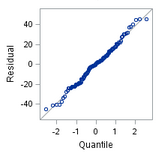
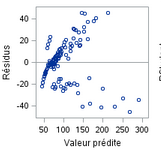
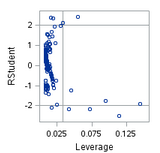
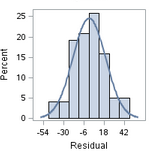


Dans ces résultats, on regarde avant tout la p-value, c’est la probabilité que les variables ne soient pas liées, Ici , toutes les p-valeur sont inférieur à 0,001 ,

Les p-valeurs étant donc toutes significatives, on rejette donc l’hypothèse de la nullité de l’ensemble des paramètres, et de la nullité de chacun des deux paramètres : la consommation a une influence sur le montant de la facture.

Le R carré (coefficient de détermination) détermine à quel point l’équation de régression est adaptée pour décrire le modèle, Le R-carré est de 85%, on en déduit que le modèle explique 85% de la variabilité du montant de la facture

On remarque sur le graphique des résidus observés en fonction des valeurs prédites que les résidus sont plutôt centrés, et sont plutôt répartis de façon homogène ce qui nous permet de valider l’hypothèse d’homoscédasticité des résidus. Enfin on peut valider l’hypothèse de normalité des résidus grâce au QQ-plot observé ci-dessous. De plus, la répartition des variables résidus sont indépendantes identiquement distribuées et suivent une loi normale Aussi, le graphique « leverage » (influence) nous montre l’impact des résidus sur le modèle de la régression linéaire.





Toutes ces observations nous permettent de valider la qualité du modèle linéaire. On peut noter par ailleurs que ce modèle n’a pas réellement de sens, en effet le montant de la facture est utilisé dans les données afin de calculer la consommation.

On a ensuite le modèle suivant : **Température = 12,5 + 0,86 \* riders**.

Les tests de Fisher et de student sur la nullité des paramètres du modèle sont récapitulés dans le tableau suivant :

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **Ddl** | **Test valeur** | **Valeur** | **Pr > F** | **Estimation du paramètre** |
| **Modèle** | 1 | Ficher | 0,74 | 0,39 | -- |
| **Intercept** | 1 | Student | 12,5 | <0,001 | 12,53 |
| **riders** | 1 | Student | 0,86 | 0,39 | 197,93 |

On remarque que la p-valeur du test de Fisher estimant la nullité de l’ensemble des paramètres du modèle n’est pas significative : on ne rejette donc pas H0 qui est la nullité de l’ensemble des paramètres. De même pour le test de Fisher concernant le coefficient de Riders, celui-ci n’est en effet pas significatif avec une p-valeur de 39%. On en déduit donc que Riders ne parait pas avoir d’effet explicatif linéaire sur la température.

On remarque par ailleurs que le R carré est de 0,006 : 0,6% de la variabilité de la température est expliquée par le modèle ce qui est extrêmement faible.

Le modèle est donc de très faible qualité, en effet la température et les Riders ne sont pas (ou très faiblement) corrélées linéairement.

1. **Régression linéaire multiple**

Dans cette partie, on se propose d’expliquer la variable facture en fonction des autres variables. Il s’agit d’un modèle linéaire multivarié. La variable facture est la variable à expliquer, les autres variables de la base de données sont des variables explicatives.

Nous avons en préambule testé plusieurs modèles de régression pour aboutir à une synthèse des plus pertinents. Voici celui que nous avons construit :

**Facture = - 93,69 + 70,42 \* Février + 68,91 \* Mars + 94,17 \* Avril - 87,98 \* Juin - 169,21 \* Juillet - 110,94 \* Août + 31,88 \* Octobre + 0,13 \* HDD + 0,81 \* CDD + 22,16 \* Taille + 73,93 \* Mètre - 34,18 \*Pompe1 - 23,45 \* Pompe2**

Pour arriver à ce modèle nous avons mené plusieurs réflexions préliminaires en examinant les données.

Tout d’abord nous avions remarqué que les variables factures et consommation sont très liées d’après les tests effectués, en effet le montant de la facture découle directement de la quantité d’énergie consommée.

D’autre part on note que dans le jeu de données il nous est explicité que la variable consommation est calculée à partir de la facture, des Riders et d’un autre indicateur. Par conséquent intégrer cette variable à la régression biaiserait notre modèle, en effet lorsque nous cherchons anticiper le montant de la facture il suffit juste d’anticiper le niveau de consommation d’énergie, grâce à la relation des deux variables.

Dans la même logique, on décide de ne pas intégrer la température, car la facture est d’un montant élevé pour des valeurs extrêmes de la température (soit basses, soit hautes) et un montant faible pour les valeurs moyennes de la température, par conséquent la relation n’est pas linéaire. On préfère donc intégrer les HDD et CDD à la place qui ont un lien linéaire plus marqué avec le montant de la facture, en effet quand les HDD et CDD prennent des valeurs élevées alors la facture aussi, ceux sont les périodes de température extrême.

En outre, les variables HDD et CDD sont très corrélées : en effet plus il est nécessaire de chauffer la maison dans le mois, moins il est nécessaire de la refroidir et inversement. Nous avons essayé d’optimiser le modèle avec une variable qui ferait la somme de HDD et CDD seulement cela ne donnait pas un résultat satisfaisant vis à vis de la qualité du modèle, ainsi nous avons donc intégré ces deux variables dans le modèle.

Pour la division dans le temps, nous avons essayé la variable Saison utilisée précédemment, seulement le résultat n’était adéquat compte tenu de la nécessité de conserver une courbe des températures plus précise que 4 moyennes sur une année. Nous décidons donc de conserver la variable mois pour notre modèle.

En testant nous-même des modèles selon les intensités des corrélations entre les différentes variables ainsi qu’en effectuant des sélections automatiques (backward, forward et stepwise) nous avons abouti à plusieurs modèles pertinents. Le modèle final est issu d’une sélection automatique pas à pas (stepwise).

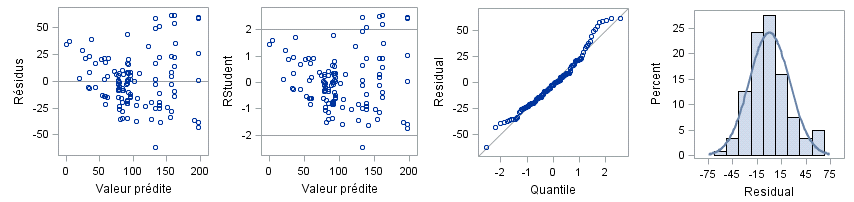
Voici le résumé des tests effectués sur le modèle :



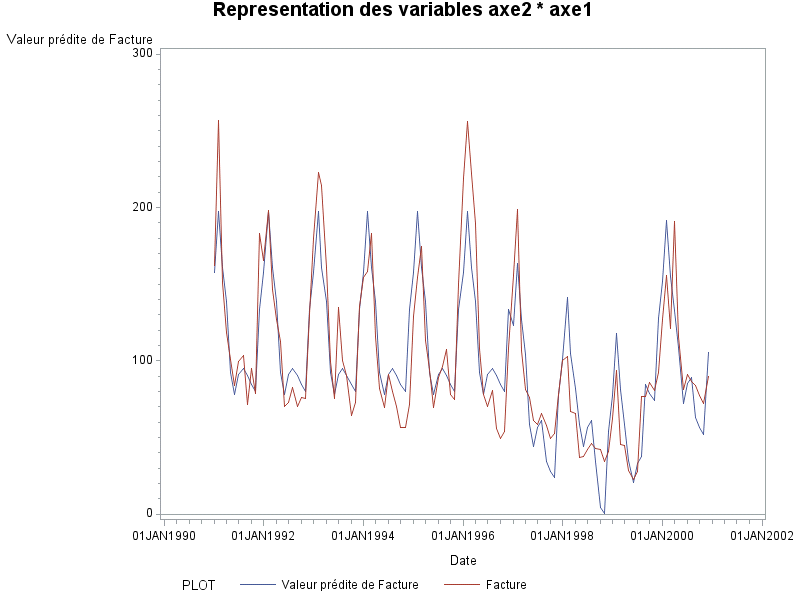
On remarque que le test de Fisher est significatif : On rejette donc l’hypothèse nulle qui est la nullité de l’ensemble des paramètres du modèle.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Ddl** | **Estimation du paramètre** | **Statistique de test** | **P-valeur** |
| **Intercept** | 1 | -93,6886 | -2,59 | 0,0109 |
| **Hdd** | 1 | 0,1321 | 6,73 | <,0001 |
| **Cdd** | 1 | 0,8086 | 4,41 | <,0001 |
| **Taille** | 1 | 22,1620 | 2,75 | 0,007 |
| **Compteur** | 1 | 73,9257 | 6,61 | <,0001 |
| **Pompe1** | 1 | -34,1803 | -4,58 | <,0001 |
| **Pompe2** | 1 | 23,4501 | -2,36 | 0,0199 |
| **Février** | 1 | 70,4198 | 7,32 | <,0001 |
| **Mars** | 1 | 68,9067 | 6,42 | <,0001 |
| **Avril** | 1 | 94,1689 | 6,17 | <,0001 |
| **Juin** | 1 | 87,9803 | -3,93 | 0,0002 |
| **Aout** | 1 | -110,9428 | -3,61 | 0,0005 |
| **Juillet** | 1 | -169,2065 | -3,98 | 0,0001 |
| **Octobre** | 1 | 31,878 | 2,38 | 0,0191 |

De plus le R carré ajusté est de 73,25%, on en déduit que 73,25% de la variabilité de la facture est expliquée par le modèle. En examinant le nuage de points des résidus on observe que ceux-ci sont bien centrés et répartis de façon homogène. De plus le qq-plot et l’histogramme des résidus nous permettent de valider l’hypothèse de normalité.



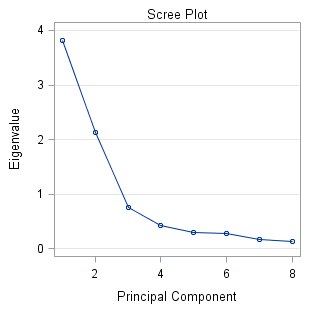
On représente ci-dessous le graphique des valeurs prédites de la facture par rapport aux valeurs réelles. On remarque que les deux courbes sont relativement similaires en effet l'ensemble de ces éléments nous permettent de valider la qualité du modèle.



1. **ACP**

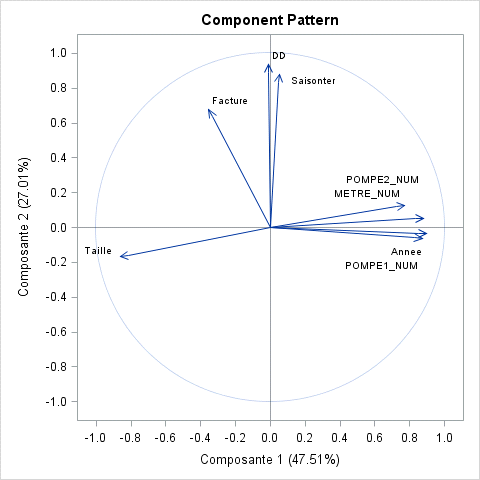
Faisons une Analyse en Composantes Principales sur le jeu de données. Pour cela nous créons une nouvelle variable saison **qui** prend 1 pour les mois de Novembre à Mars (saison froide) et 0 pour les mois d’Avril à Octobre (Saison Chaude). Ainsi on additionne HDD et CDD en une variable DD (nombre de jours mensuels pour lesquels il est nécessaire d’utiliser de l’énergie pour réchauffer ou refroidir la maison). Donc la création de cette variable augmente le pourcentage d’inertie porté par le premier axe et celui que nous sommes intéressées de représenter. En effet on centre et réduit les données. Ainsi nous avons de variables qui ont des unités différentes il est donc bien de les standardiser.

Ainsi dans l’ACP, le logiciel SAS nous fournit les informations sur les axes suivants :



|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Valeurs propres de la matrice de corrélation** | | | | | | |
| **Valeur propre** | | **Proportion** | | | **Cumulé** | |
| **1** | 3,80 | | 47% | 47% | |
| **2** | 2,16 | | 27% | 74% | |
| **3** | 0,80 | | 10% | 84% | |
| **4** | 0,47 | | 6% | 90% | |
| **5** | 0,29 | | 4% | 94% | |
| 6 | 0,20 | | 3% | 97% | |
| 7 | 0,14 | | 1% | 98% | |
| 8 | 0,13 | | 2% | 100% | |

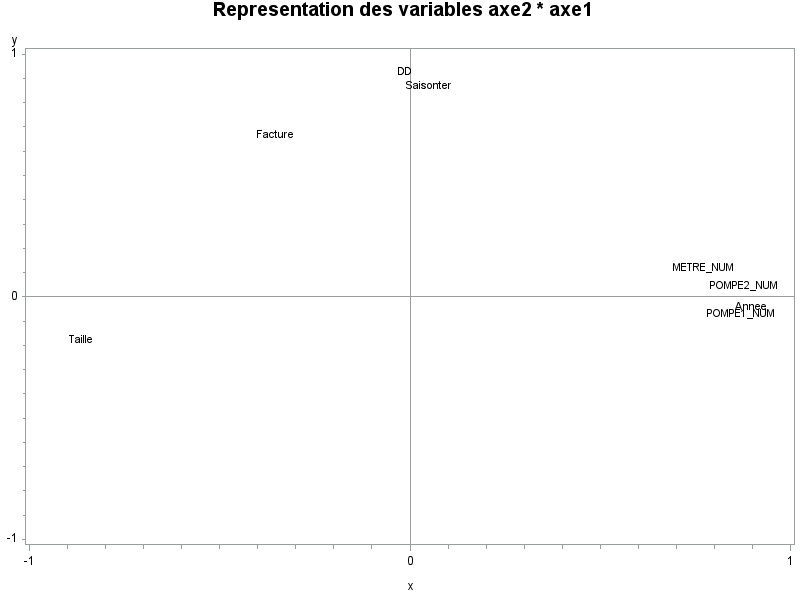
Nous constatons que les 4 premiers axes portent 90% de la variance. Sur le graphique nous remarquons que le « coude » se forme entre le 3ème et le 4ème axe. Donc nous maintenons 4 axes principaux pour notre ACP qui représentent 90% de l’inertie totale. La plupart des renseignements fournis par les données sont dû à L'ACP.



**Représentation du cercle des corrélations entre le premier et le deuxième axe :**

Nous constatons que le montant de la facture est lié aux variables DD et saisonnier : Plus le nombre de jours mensuels auxquels la consommation de l’énergie est élevée et plus on est dans une saison froide et plus le montant de la facture est élevé. Par contre les deux premiers axes n'expliquent pas la corrélation entre les autres variables et factures. Donc les vecteurs des autres variables sont orthogonaux à celui de la facture, ces variables sont donc indépendantes au montant de la facture selon les deux premiers axes. Nous constatons aussi qu'il y a une liaison entre elles. Dans les dernières années de l’étude, le compteur et les pompes sont changés et le nombre de personnes habitants dans la maison est petit. Nous constatons aussi que l’année contribue le plus à la formation du premier axe, et la variable DD contribue le plus à la formation du deuxième axe car ce sont les variables qui ont les plus grandes coordonnées en valeur absolue.

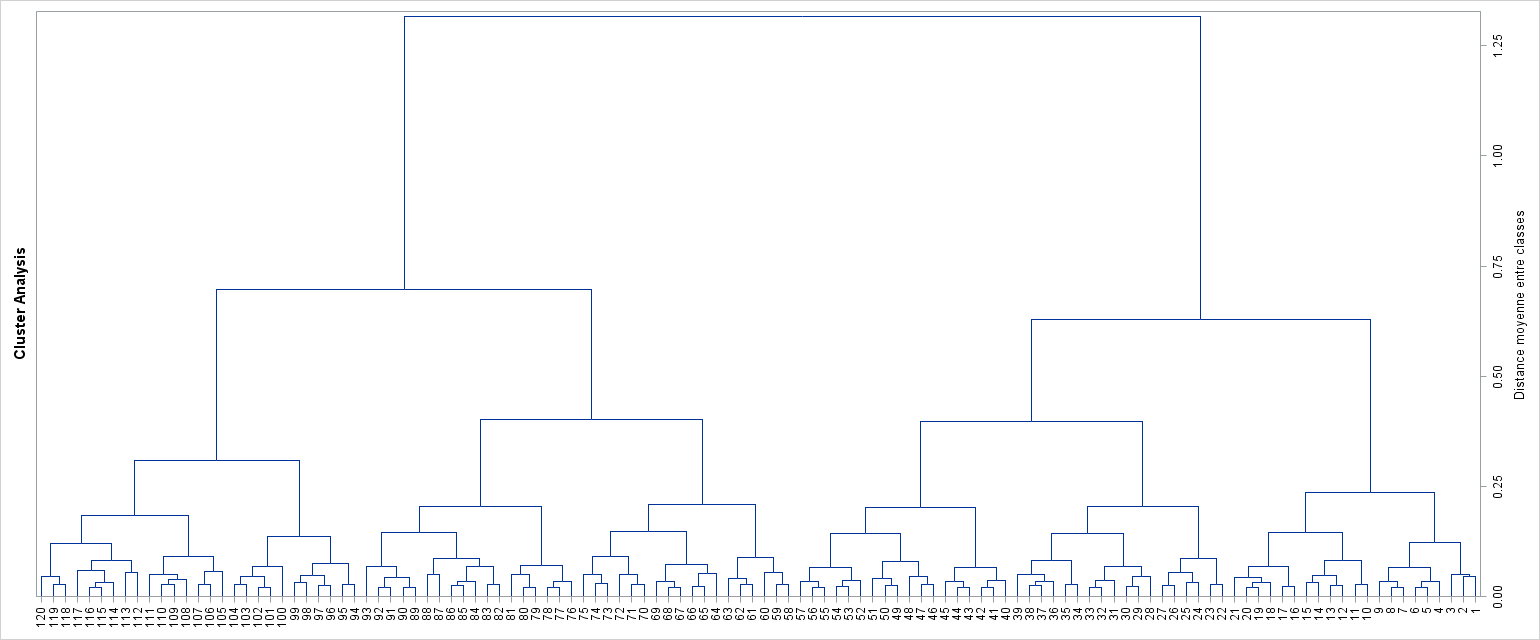
**Représentation de nuage de points des variables et des individus :**



En observant le nuage de points des individus et celui des variables, on peut déterminer les caractéristiques de certains individus. Par exemple au début de l’étude pendant la saison froide, et au moment que le compteur et les pompes n’avaient pas été changés et qu’il y avait beaucoup de personnes dans la maison avec un haut besoin chauffage, d'où la facture n°2 était élevée. En fin d’étude, lorsque le compteur et les pompes avaient été changés et que peu de personnes étaient dans la maison, la facture n°119 a été enregistrée avec un montant élevé. Nous constatons les mêmes caractéristiques pour l’individu 117 avec un montant de facture peu élevée et dans une période plus chaude. Si le numéro de l’observation est élevé, dans ce cas elle est vers la droite ; d’où la variable année indique que plus la facture est enregistrée tard dans l’étude, plus elle se situe vers la droite.

1. **CAH**

Faisons une Classification Ascendante Hiérarchique. Nous avons retenu 4 axes pour l’ACP, On regroupe nos individus en 4 groupes. Sur le schéma suivant : il y’a 4 groupes qu'on numérote de 1 à 4 de gauche à droite.



Dans le tableau suivant. Les moyennes des différentes variables dans les groupes sont répertoriées (On rappelle que l’on avait standardisé les variables au préalable, une valeur supérieure à 0 indique donc une moyenne supérieure à celle de l’ensemble des observations et inversement).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Groupe** | **Fréquence** | **Facture** | **Taille** | **Année** | **DD** | **Mètre** | **Pompe1** | **Pompe2** | **Saisonter** |
| **1** | 16 | 0,01 | -1,71 | 1,48 | 0 | 2,55 | 1,22 | 1,86 | 0,13 |
| **2** | 21 | -0,83 | -0,9 | 0,92 | 0,18 | -0,39 | 1,22 | 0,72 | 0,24 |
| **3** | 57 | -0,34 | 0,56 | -0,48 | -0,66 | -0,39 | -0,49 | -0,54 | -0,58 |
| **4** | 26 | 1,41 | 0,56 | -0,60 | 1,3 | -0,39 | -0,66 | -0,54 | 1 |

Le premier groupe de 16 individus est caractérisé par un montant de facture moyen enregistré à la fin de l’étude dans les dernières années, lorsque peu d’individus étaient présents dans la maison, et également après que le compteur et les pompes aient été changés. Le deuxième groupe est caractérisé par le montant de facture le plus faible de tous les groupes, il s’agit des factures enregistrées dans la seconde moitié de l’étude, après que les pompes aient été changées mais avant que le compteur soit changé. Le troisième groupe est le plus grand, ce sont les factures qui ont été enregistrées au moment de la saison chaude et dans la première partie de l’étude lorsque, et le compteur et les pompes n’ont été changés. Au Final le dernier groupe a été enregistré au moment d'une saison froide, c’est le groupe qui a un montant de facture moyen le plus élevé. De façon globale, les groupes sont décomposés de manière chronologique : les factures du tout début d’étude sont très élevées, les factures de la première moitié d’étude sont faibles, les factures de la seconde moitié d’étude avec un montant très faible, juste avant que le compteur ne soit changé, puis les factures de la dernière partie de l’étude ont un montant moyen.

**Conclusion**

Au cours de notre étude, après avoir traité les valeurs manquantes et aberrantes dans notre base de données, nous avons été amenées à étudier le montant de la facture d’électricité du foyer par rapport aux différentes variables dont nous disposions.

Après avoir écarté les variables qui n’intervenaient pas dans l’évolution du montant de la facture, nous avons remarqué que la saisonnalité présente dans le jeu de données était particulièrement importante dans l’explication du montant de la facture. Il aurait pu être intéressant d’étudier ce phénomène en utilisant les séries temporelles. Enfin en effectuant une Analyse en composantes principales puis une classification hiérarchique nous observons que les différentes groupes sont formés essentiellement de manière chronologique : les premières factures d’électricité sont regroupées entre elles, de même que les dernières. Les autres variables permettent de caractériser ces différentes périodes

**ANNEXES CODES SAS**

data projet;

infile "/users/licence/ii12517/projet stat/electricbill.dat";

input Numero Annee Mois $ Facture Temperature hdd cdd Taille Metre Pompe1 Pompe2 Riders Consommation;

run;

/modifier les valeurs manquantes \*/

data projet; set projet;

\*transformer les degrés en C\u1e1e;

Temperature=(5/9)\*(Temperature-32);

format Temperature 4.1;

\*corriger l'erreur de saisie en décembre 92;

if Numero=24 then Annee=1992;

\*créer une var qui contient mois et année;

a=mod(Numero,12);

if a=0 then a=12;

Date=mdy(a,1,Annee);

format Date mmyys10.;

drop a;

if Numero=37 then Facture = 154.57; \*remplacement de la valeur manquante;

if Numero in (104,105) then Facture = 76.66; \*remplacement de la valeur nulle;

run;

goption i=join;

proc gplot data = projet \*On remarque une forte saisonalité ;

plot facture \* date=7 /skipmiss legend;

plot2 consommation \* date=5 /legend;

run;

/\*satat descriptives\*/

proc means data = projet mean std min max median; \*valeur manquante ...;

var facture temperature hdd cdd taille riders consommation;

run;

goption i=join;

proc gplot data = projet;

plot facture \* date=7 /skipmiss;

run;

proc gchart data=projet;

pie taille / type =PERCENT;

Title "Repartition du nombre d'individus dans la maison";

run;

/\* Etude des corrélations \*/

proc corr data = projet;

var facture temperature hdd cdd taille riders consommation;

run;

goption i=join;

proc gplot data = projet;

plot facture \* date=7 /skipmiss;

run;

goption i=join;

proc gplot data = projet \*On remarque une forte saisonalité ;

plot facture \* date=7 /skipmiss legend;

plot2 consommation \* date=5 /legend;

run;

proc gchart data = projet;

vbar taille /discrete;

run;

/\* teste kendall\*/

proc corr data = projet kendall;

var riders temperature;

run;

proc corr data = projet kendall;

var facture HDD annee taille CDD Temperature Consommation riders;

run;

/\* teste de students \*/

proc ttest plots=all;

run;

proc ttest H0=100 ;

var facture;

run;

/\* test de normalité \*/

proc univariate data = projet all normal plot;

var Facture Consommation;

histogram Facture Consommation /normal (color=red W=5);

run;

/\*regression lineaire\*/

proc reg data = projet ;

model Facture=consommation;

run;

proc reg data = projet ;

model Facture=riders;

run;

/\*regression multiple\*/

proc reg data= projet ;

model Facture =hdd cdd taille metre pompe1 pompe2 Mois$;

run;

proc glm data =projet;

model facture = hdd cdd taille metre pompe1 pompe2 Mois;

run;