JACQUES VIRGINIE
MOULIN SARAH
VILLAINES FLORIANE

SAE EXPLIQUER/PREDIRE UNE VAR. QUANTI. AVEC DES FACTEURS

Table des matières

Introduction

En 2019, la pollution atmosphérique est considérée comme une des principales menaces sur la santé, selon l'Organisation Mondiale de la Santé. La qualité de l'air peut être modifiée par des polluants qui peuvent être d'origine naturelle ou d'origine anthropique, c'est-à-dire liés à l'activité humaine. Les PM2,5 sont des particules en suspension de diamètre inférieur à 2,5 micromètres. Elles pénètrent profondément dans l'appareil respiratoire jusqu'aux alvéoles pulmonaires et peuvent passer dans la circulation sanguine. Les particules primaires sont directement émises dans l'atmosphère. Elles sont majoritairement issues de toutes les combustions incomplètes liées aux activités industrielles ou domestiques, ainsi qu'aux transports. Elles sont aussi émises par l'agriculture (épandage, travail du sol, etc). Elles peuvent également être d'origine naturelle (érosion des sols, pollens, feux de biomasse, etc.). Les particules secondaires sont formées dans l'atmosphère suite à des réactions physico-chimiques. Les particules sont particulièrement nocives pour la santé. Elles provoquent des irritations et des problèmes respiratoires chez les personnes sensibles et sont associées à une augmentation de la mortalité (affections respiratoires, maladies cardiovasculaires, cancers…). En France, Santé publique France estime chaque année que près de 40 000 décès seraient attribuables à une exposition des personnes âgées de 30 ans et plus aux particules fines.

Le rapport est rédigé comme suit : nous présentons dans un premier temps les données, puis dans un deuxième temps la régression multiple réalisée, ainsi que les résultats obtenus suite à cette régression. Nous proposons dans un troisième temps de réaliser une analyse de la variance (ANOVA). Pour finir, nous réaliserons une régression multiple avec sélection de facteurs afin d'améliorer le modèle, que ce soit au niveau de la modélisation ou des prévisions.

Présentation des données

Le jeu de données provient des archives disponibles en ligne de l'université d'Irvine en Californie **"**UC Irvine Machine Learning Repository"**.** Nous l'avons importé sur Rstudio. Il contient 52 584 observations et 17 variables :

- No: numéro de la ligne
- year: l'année
- month: le mois
- day: le jour
- hour: l'heure
- season: la saison
- PM_Caotangsi: la concentration en PM2.5 (ug/m^3) du capteur de Caotangsi
- PM_Shahepu: la concentration en PM2.5 (ug/m^3) du capteur de Shahepu
- PM_USpost: la concentration en PM2.5 (ug/m^3) du capteur de USpost
- DEWP: le point de rosée (°C)
- TEMP: la temperature (°C)
- HUMI: l'humidité (%)
- PRES: la pression (hPa)
- cbwd: la direction du vent
- Iws: la vitesse du vent (m/s)
- precipitation: la précipitation dans l'heure (mm)
- Iprec: la precipitation cumulée (mm)

Les données proviennent de mesures faites dans la ville de Chengdu, chaque heure, chaque jour et chaque mois entre 2010 et 2015.

Chengdu est une ville située au centre-ouest de la Chine, c'est la capitale de la province de Sichuan. Il s'agit de la quatrième ville la plus peuplée de Chine; elle recensait plus de 14 millions d'habitants en 2010, et en compte maintenant 20 millions. Chengdu possède une culture gastronomique très populaire, de superbes jardins historiques et est connu pour l'élevage de pandas géants. C'est une ville où se côtoient traditions et modernité, avec temples, monuments, maisons traditionnelles ainsi que des commerces artisanaux.

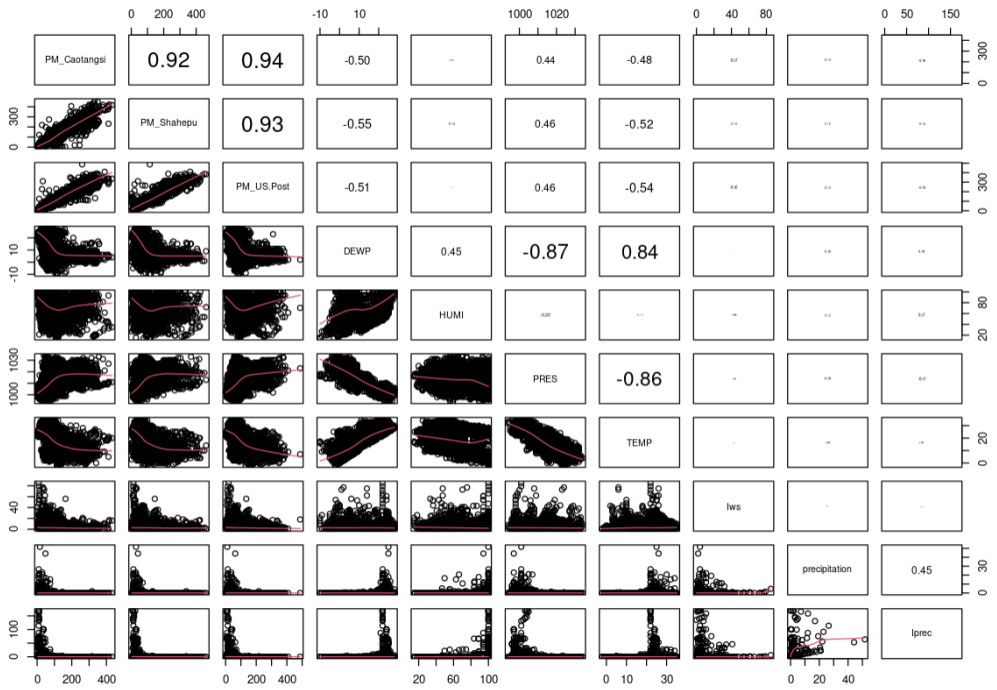
Le but principal de cette SAE est d'analyser les variations de la pollution de Chengdu en fonction d'autres variables, grâce à l'application de modèles linéaires.

I - Régression multiple

Le modèle de régression linéaire multiple est une généralisation du modèle de régression simple, mais avec plusieurs facteurs. Elle a pour but d'identifier les relations existantes entre plusieurs variables indépendantes ou prédictives avec une variable explicative (dite dépendante). La régression multiple nous permettra d'estimer le vecteur des paramètres, d'estimer la variance des erreurs, d'étudier les résidus afin de valider le modèle de régression, de construire des tests, des intervalles de confiance et de prévision et enfin de sélectionner des facteurs.

Avant de procéder à notre première analyse, nous devons supprimer nos variables qualitatives. Parmi nos 17 variables, une seule l'est, la direction du vent. Nous avons également supprimé les variables de mesures temporelles (l'année, le mois, le jour, l'heure, la saison) ainsi que la colonne indiquant le numéro de ligne des données (No). En effet, ces variables n'étaient pas pertinentes pour notre analyse et nous avons choisi de ne pas les conserver. Le nombre d'observations étant beaucoup trop conséquent, nous choisissons de ne sélectionner qu'une partie de celles-ci. Nous allons donc poursuivre notre analyse avec 3000 observations de nos 52 584 observations initiales.

Nous effectuons une première analyse graphique du jeu de données avec la fonction *pairs* afin d'apercevoir d'éventuelles liaisons linéaires entre paires de variables. Nous obtenons alors les graphiques de corrélations ci-dessous :



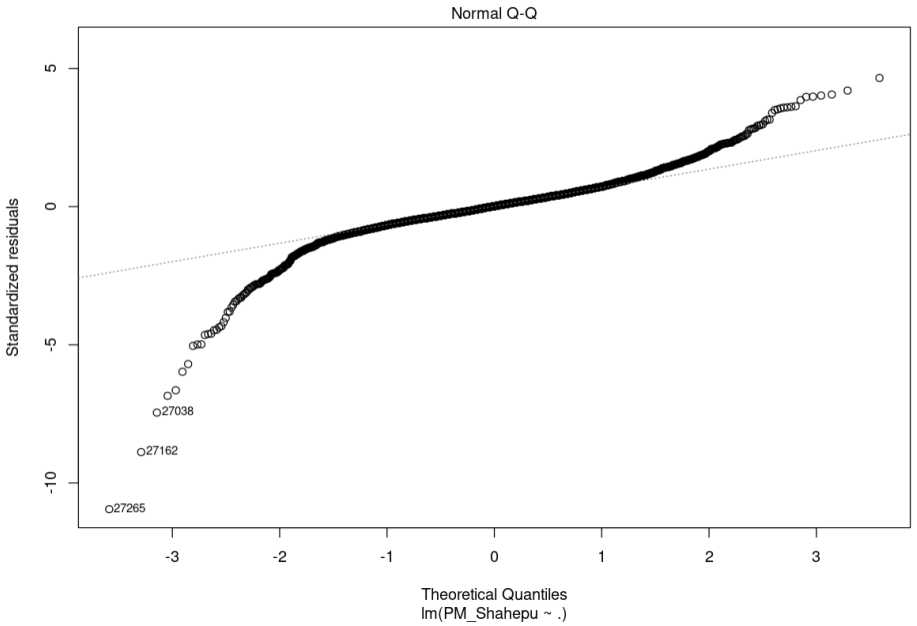
Nous observons des corrélations très fortes entre les 3 capteurs de pollution, avec un coefficient de corrélation supérieur à 0.9. Il y a donc une forte relation entre les 3 variables de particules PM. Nous observons également une forte corrélation négative entre le point de rosée et la pression et la température et la pression, ainsi qu'une forte corrélation positive entre le point de rosée et la température, respectivement inférieure à -0.8 et supérieure à 0.8.

On conclut alors grâce à cette première analyse qu'il existe des liaisons linéaires entre paires de variables.

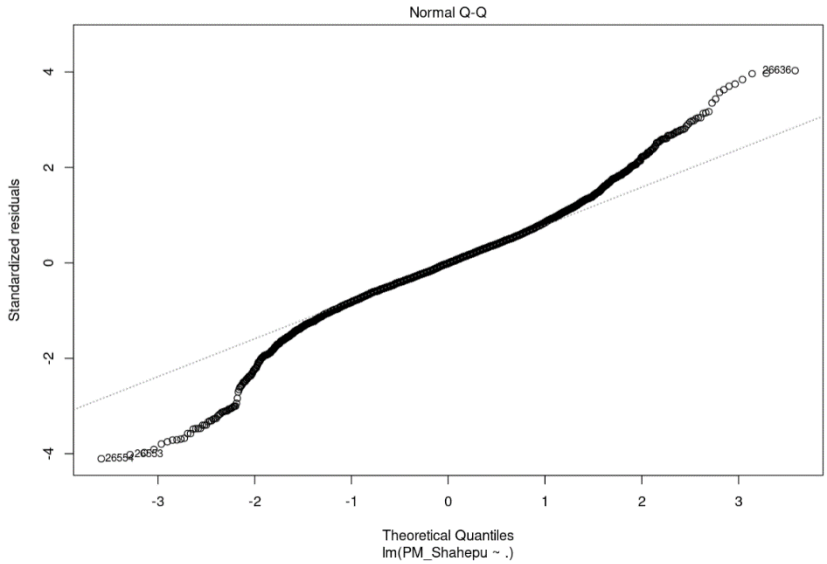
Parmi les trois colonnes de mesures de la pollution nous sélectionnons le capteur PM_Shahepu. Cette variable sera notre variable à expliquer. Les autres feront partie des variables explicatives. Nos facteurs seront le point de rosée, la température, la pression, la précipitation, l'humidité et la vitesse du vent.

Nous procédons ensuite à la régression de la variable pollution sur les autres variables à l'aide de la fonction *lm*. Les sorties générées par la commande *summary* nous indiquent les différentes valeurs de la régression effectuée. On obtient une valeur de $R^2 = 0.8985$, cela signifie que la modélisation par la droite des moindres carrés explique 89% de la variation totale, ce qui est un très bon résultat, le modèle est donc bien adapté. La variance estimée des erreurs est de 744. La dernière ligne du summary "F-statistic" représente le test simultané, celui-ci permet de tester la significativité du modèle. Si ce test a une grande p-value, cela signifie que la régression linéaire est totalement inadaptée pour nos données. Ici la p-value est < 2.2e-16, le modèle linéaire est donc bien pertinent pour l'étude de nos données.

Nous allons à présent procéder à l'analyse des résidus. L'examen des résidus se fait sur le graphique ci-dessous :



Nous observons la présence d'écarts par rapport à la droite de Henry en haut à droite et surtout en bas à gauche (avec les points 27265, 27162, 27038). Il se pourrait donc que la loi des erreurs ne soit pas une loi normale. Nous repérons alors ces valeurs aberrantes et nous les supprimons une à une, en recommençant une nouvelle régression à chaque fois. Nous supprimons un total de 36 valeurs aberrantes. La suppression de ces 36 valeurs nous permet d'obtenir le graphique suivant :



Nous venons de supprimer l'ensemble des valeurs dont leur résidu standardisé n'était pas compris entre - 4 et 4. Nous observons alors que ce nouveau jeu de données suit davantage la droite d'Henry.

Nous décidons d'effectuer une nouvelle analyse avec un *summary* afin de noter les éventuelles améliorations. On remarque dans ce nouveau résumé une amélioration au niveau du R^2, qui vaut maintenant 0.9272. Il est supérieur au R^2 de notre précédente analyse, celle comprenant toutes les données, mêmes aberrantes. On s'intéresse ensuite aux intervalles de confiance des EMC, qui sont les suivants :

	2.5 %	97.5 %
(Intercept)	437.72150532	869.4645172
PM_Caotangsi	0.27935003	0.3417700
PM_US.Post	0.67974712	0.7447727
DEWP	0.40273579	2.0346331
HUMI	-0.97776156	-0.5717449
PRES	-0.78024437	-0.3645471
TEMP	-2.90227744	-1.3183071
Iws	0.02073419	0.2477540
precipitation	-0.11958553	0.8057390
Iprec	0.07367289	0.2270446

Les hypothèses de ce test sont H0 : $\beta = 0$ contre H1 : $\beta \neq 0$. La règle de décision de ce test nous indique que l'on conserve l'hypothèse H0 quand 0 est compris dans l'intervalle. D'après cette règle, nous pouvons alors dire que nous conservons H0 seulement pour la variable précipitation ([−0.119, 0.805]). L'hypothèse H0 est rejetée pour toutes les autres variables.

Pour préciser ces résultats, nous allons maintenant nous intéresser aux p-valeurs, fournies par le résumé (commande *summary*).

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   681.17199  131.28816   5.188 2.26e-07 ***
PM_Caotangsi    0.39261    0.01829  21.461  < 2e-16 ***
PM_US.Post      0.62198    0.01893  32.856  < 2e-16 ***
DEWP            1.06706    0.49529   2.154  0.03129 *
HUMI           -0.70438    0.12322  -5.717 1.19e-08 ***
PRES           -0.60483    0.12640  -4.785 1.79e-06 ***
TEMP           -1.98783    0.48126  -4.130 3.72e-05 ***
Iws             0.12089    0.06917   1.748  0.08061 .
precipitation   0.33542    0.28247   1.187  0.23513
Iprec           0.14098    0.04680   3.012  0.00262 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.28 on 2990 degrees of freedom
Multiple R-squared:  0.8985,    Adjusted R-squared:  0.8982
F-statistic:  2942 on 9 and 2990 DF,  p-value: < 2.2e-16
```

Les estimateurs des moindres carrés sont sur la première colonne, on peut également trouver leur variance sur la deuxième colonne.

Ainsi pour le test H0 : β8 = 0 (β correspondant à la variable precipitation), on obtient une p-valeur de 0.23513, ce qui signifie qu'on ne rejette pas H0 au niveau 5%. On ne rejette H0 qu'à partir du niveau 23.5%, ce qui est beaucoup trop élevé. On en déduit que, au niveau 5%, l'hypothèse H0 : β8 = 0 est vraie. Alors que pour le test H0 : β1 = 0, on obtient une p-valeur égale à 2e − 16, ce qui signifie qu'on rejette H0 à partir d'un niveau de l'ordre de 10−16, ce qui est très proche de 0. Donc on rejette H0 pour quasiment tous les niveaux, et donc aussi au niveau 5%. On en déduit que, au niveau 5%, β1 est différent de 0. On peut en déduire de même pour les β2, β3, β4, β5, β6, β7 et β9.

Au niveau 5%, une variable est dite significative si sa p-valeur est inférieure à 0.05. Le nombre d'étoiles présent à côté des p-valeurs nous permet également de déterminer le taux de significativité. Ici, les variables les plus significatives sont : les deux variables de pollution en PM2,5 PM_Caotangsi et PM_USpost, l'humidité (humi), la pression (pres) et la température (temp). En revanche, on peut noter que seule la variable precipitation n'est pas significative puisque sa p-valeur, qui vaut 0.23513, est supérieure à 0.05.

Nous cherchons à présent à effectuer des prévisions. Pour cela nous avons récupéré dans le jeu de données complet les 10 valeurs qui suivent le bloc des 3000 données que nous avions choisi. Grâce à la commande *fitted* on obtient les prévisions suivantes:

```
   31560    31561    31562    31563    31564    31565    31566    31567    31568    31569
12.97575 23.02425 18.06260 19.99946 21.94037 24.91035 25.98870 26.08722 18.01373 20.99757
```

A partir de ces nouvelles valeurs, nous avons déterminé les intervalles de confiance au niveau 95 % pour l'espérance (à gauche) ainsi que les intervalles de prévision (à droite) :

```
> predict.lm(reg_prev,interval= "confidence"
           fit      lwr      upr
31560 12.97575 11.00606 14.94544
31561 23.02425 21.05456 24.99394
31562 18.06260 16.23452 19.89069
31563 19.99946 18.00583 21.99309
31564 21.94037 20.09631 23.78443
31565 24.91035 23.27421 26.54649
31566 25.98870 24.00023 27.97716
31567 26.08722 24.42999 27.74445
31568 18.01373 16.02774 19.99973
31569 20.99757 19.00416 22.99097
```

```
> predict.lm(reg_prev,interval= "prediction"
           fit      lwr      upr
31560 12.97575 10.17320 15.77830
31561 23.02425 20.22170 25.82680
31562 18.06260 15.35770 20.76751
31563 19.99946 17.18003 22.81889
31564 21.94037 19.22464 24.65610
31565 24.91035 22.33128 27.48941
31566 25.98870 23.17292 28.80448
31567 26.08722 23.49472 28.67972
31568 18.01373 15.19970 20.82777
31569 20.99757 18.17830 23.81684
```

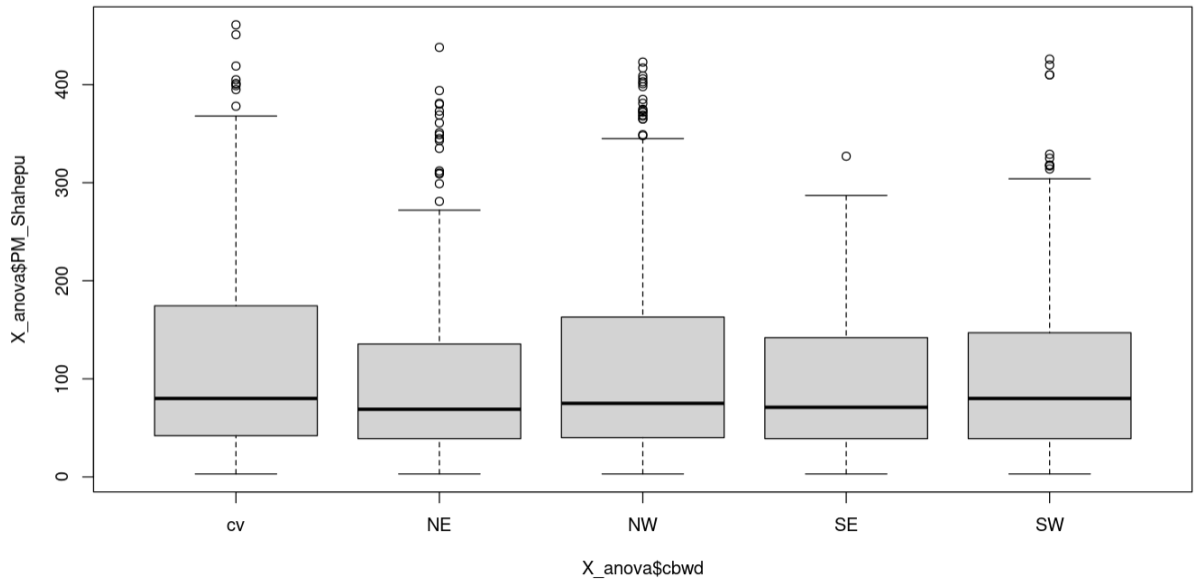
La première colonne ”fit” donne les valeurs des prévisions calculées par le modèle. La deuxième colonne ”lwr” donne les valeurs des bornes inférieures des intervalles de confiance pour chaque donnée, et la troisième colonne ”upr” donne les valeurs des bornes supérieures.

Nous terminons notre première partie par l'utilisation d'une MSE (Mean Squared Error) afin de quantifier la qualité de nos prévisions, pour laquelle on obtient une valeur de 300.

II - Anova

Dans cette partie, nous étudierons les variations de la pollution en fonction de la direction du vent, afin de voir si cette dernière a une influence ou non.

Nous repartons de nos 3000 données initiales et nous construisons dans un premier temps un data-frame contenant la colonne de la pollution et celle de la direction du vent. Nous effectuons une première analyse graphique à travers des boîtes à moustaches.



On observe sur les boxplots ci-dessus que la moyenne de pollution est la plus élevée quand il n'y a pas de vent. C'est également le boxplot qui a le premier et troisième quartile le plus élevé, et les valeurs extrêmes les plus grandes. On constate également grâce à cette première analyse graphique que la moyenne de pollution est la plus faible quand le vent vient du nord-est.

Nous définissons dans un deuxième temps les quantités importantes du jeu de données. Le facteur explicatif, la direction du vent, comporte 5 modalités. Sur l’image de gauche ci-dessous, on trouve d’abord la moyenne générale, qui est de 106.8353, puis la première ligne du tableau nous donne les moyennes partielles et enfin la ligne « rep » nous donne les effectifs pour chaque modalités. Sur l’image de droite, qui représente la table des effets, on trouve sur la première ligne les estimations des effets puis on retrouve en deuxième ligne les effectifs partiels.

```
Tables of means
Grand mean

106.8353

 cbwd
       cv     NE    NW     SE    SW
    111.1  95.69 110.6  98.83 102.3
rep 1231.0 363.00 706.0 190.00 510.0
```

```
Tables of effects

 cbwd
          cv     NE      NW      SE      SW
       4.248 -11.14   3.779  -8.004  -4.573
rep 1231.000 363.00 706.000 190.000 510.000
```

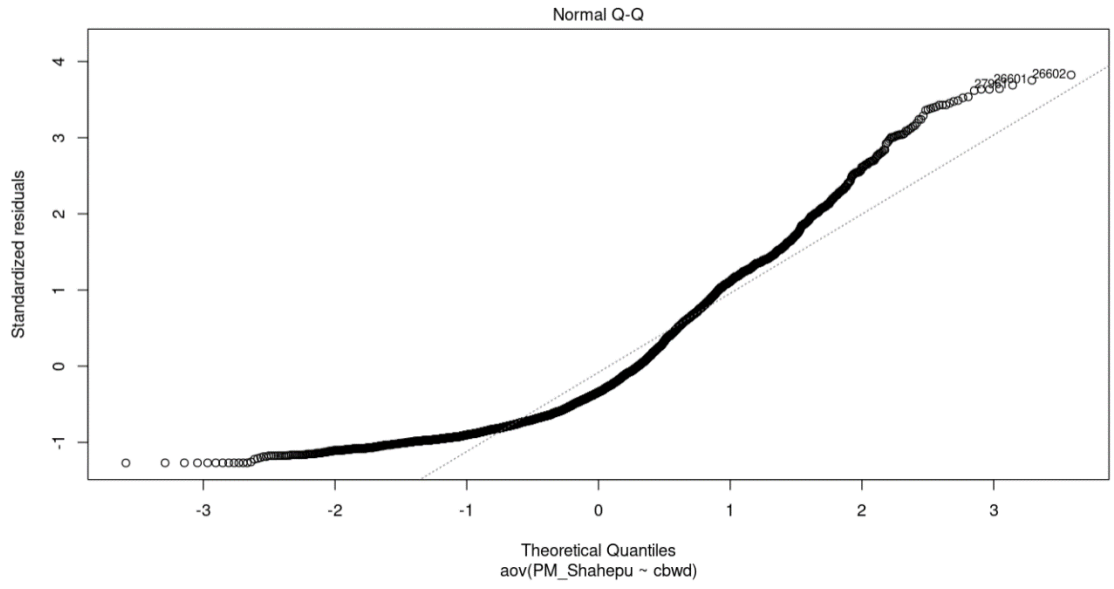
Enfin, voici l'affichage de la table d'analyse de la variance :

```
              Df   Sum Sq Mean Sq F value  Pr(>F)
cbwd           4   100216   25054   3.437 0.00825 **
Residuals   2995 21834528    7290
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Pour la colonne « Df », la première ligne correspond à K-1. Ici, cette case est égale à 4, on retrouve alors bien les 5 modalités. La deuxième ligne correspond à n-K. La colonne « Sum Sq » correspond à la somme des carrés. La première ligne correspond à la somme des carrés factorielle (SC fac), la deuxième ligne correspond à la somme des carrés résiduelle (SC rec). La colonne « Mean Sq » correspond au carré moyen. La première ligne indique la valeur des carrés moyens factorielle (CM fac). La deuxième ligne indique la valeur des carrés moyens résiduelle (CM res).

On teste l'égalité des moyennes, la p-valeur étant proche de 0, on peut rejeter H0 (au niveau 0.05 et 0.01). Les moyennes sont donc différentes, ce qui rejoint notre première analyse graphique des boxplot.

Nous avons ensuite effectué l'analyse des résidus. Comme pour notre première partie, nous avons réalisé un graphique nous permettant d’identifier les valeurs aberrantes. Après les avoir supprimées, voici le graphique que nous obtenons :

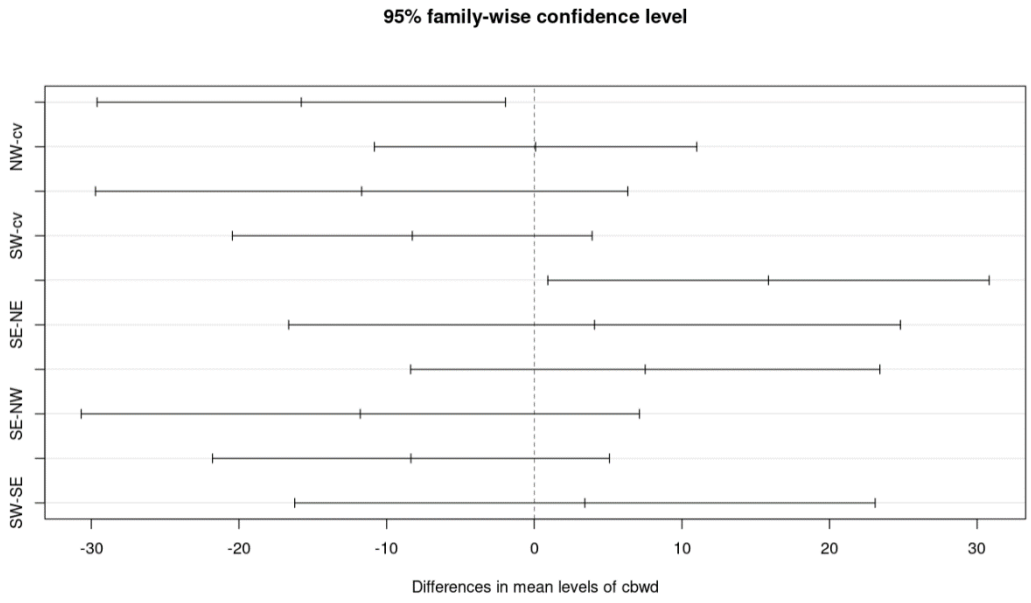


Afin d'affiner notre analyse et de localiser les différences, nous décidons d'utiliser 2 méthodes de comparaisons multiples.

Nous utilisons tout d'abord la méthode de Tukey. Ce test cherche à établir les directions où la pollution en PM2.5 est significativement différente. La méthode de Tukey garantit un test avec un niveau global ne dépassant pas la valeur α qu'on s'est fixé.

```
$cbwd
             diff         lwr       upr     p adj
NE-cv -15.77651956 -29.6058487 -1.947190 0.0159989
NW-cv   0.09235496 -10.8285454 11.013255 0.9999999
SE-cv -11.69079697 -29.7182639  6.336670 0.3913750
SW-cv  -8.25963082 -20.4406888  3.921427 0.3445716
NW-NE  15.86887452   0.9194236 30.818325 0.0310415
SE-NE   4.08572259 -16.6316513 24.803096 0.9833531
SW-NE   7.51688874  -8.3764493 23.410227 0.6968828
SE-NW -11.78315193 -30.6835567  7.117253 0.4330318
SW-NW  -8.35198578 -21.7912533  5.087282 0.4363763
SW-SE   3.43116615 -16.2243015 23.086634 0.9894824
```

Pour la commande *TukeyHSD*, la première colonne donne la différence entre les moyennes partielles (sans valeur absolue) pour les K(K-1)/2 = 11 combinaisons possibles, selon les modalités du facteur. Les colonnes "lwr" et "upr" donnent un intervalle de confiance pour cette "diff". Si la valeur 0 n'est pas dans cette intervalle, c'est qu'on peut considérer la différence entre les 2 moyennes comme significatives et donc on choisit de rejeter H0 : mu_k = mu_k'. Le graphique ci-dessous permet de mieux visualiser les intervalles :



La dernière colonne donne une p-value ajustée du test H0 : muk = muk' contre H1. Ici, on fixe un seuil (par exemple 5 %) et on rejette H0 si la p-value ajustée est inférieure au seuil qu'on s'est fixé (ici 5% donc 0.05). Si la p-value est supérieure au seuil, alors on choisit H0 muk = muk' pour les deux facteurs

concernés. On rejette H0 pour NE-cv et NW-NE car 0 n'est pas compris dans leur intervalle. Leur p-valeur, inférieure à 0.05 (seuil fixé à 5%), confirme cette affirmation.

La moyenne de NE est donc différente de celle de cv et de NW. Cela rejoint le fait que la moyenne de NE est la plus faible.

Nous utilisons ensuite la méthode de Scheffé, une méthode d'ajustement des niveaux de signification dans une analyse de régression linéaire pour tenir compte de comparaisons multiples. Tout comme la méthode de Tukey, on est assuré que le niveau global de la décision ne dépasse pas le risque α fixé au départ.

```
$cbwd
              diff     lwr.ci     upr.ci   pval
NE-cv -15.77651956 -31.392905 -0.1601345 0.0461 *
NW-cv   0.09235496 -12.239768 12.4244777 1.0000
SE-cv -11.69079697 -32.047812  8.6662183 0.5359
SW-cv  -8.25963082 -22.014752  5.4954899 0.4895
NW-NE  15.86887452  -1.012377 32.7501259 0.0784 .
SE-NE   4.08572259 -19.308795 27.4802404 0.9905
SW-NE   7.51688874 -10.430221 25.4639983 0.7968
SE-NW -11.78315193 -33.125908  9.5596041 0.5755
SW-NW  -8.35198578 -23.527905  6.8239330 0.5786
SW-SE   3.43116615 -18.764224 25.6265560 0.9940
```

Nous procédons de façon similaire pour tester l'égalité des moyennes. Ce test de Scheffé confirme les résultats trouvés précédemment avec le test de Tukey.

Dans cette partie, une première analyse graphique nous a permis de déterminer les moyennes de pollution en fonction de la direction du vent et de remarquer que ces dernières étaient différentes. L'application des tests de Tukey et Scheffé nous ont permis de confirmer ces résultats et d'affirmer plus précisément que la moyenne de pollution est la plus faible pour un vent venant du nord-est.

III - Régression multiple - Sélection de facteurs

Nous souhaitons à présent effectuer une sélection de facteurs afin d'améliorer le modèle, que ce soit au niveau de la modélisation ou des prévisions.

Nous reprenons les tableaux de données avec les variables quantitatives et les données que nous avons nettoyées dans notre partie I.

En régression linéaire, plus le nombre p de facteurs explicatifs est important, plus le R^2 a tendance à être proche de 1. Mais si p est trop grand, l'estimation des paramètres peut être détériorée : il y a en effet un risque de colinéarité entre facteurs, ce qui entraîne une augmentation de la variance des estimateurs des paramètres du modèle. Ce phénomène peut remettre gravement en cause la validité des estimations et des prévisions calculées sur ce modèle.

Pour détecter les problèmes de colinéarité, on utilise le facteur d'inflation de la variance (VIF : Variance Inflation Factor). Le VIF mesure la colinéarité entre les facteurs. On veut le moins de colinéarité

possible entre les facteurs. On rappelle qu'on soupçonne qu'un facteur est colinéaire avec d'autres si son VIF est supérieur à 4, et qu'il est fortement corrélé si son VIF est supérieur à 10.

```
 PM_Caotangsi    PM_US.Post          DEWP          HUMI          PRES          TEMP
     8.890352      9.471206     93.266623     28.450701      5.675507     77.026352
             Iws precipitation         Iprec
     1.079638      1.265345      1.305126
```

Les résultats de la commande *vif* nous permettent de soupçonner une colinéarité pour les facteurs de pollution PM_Caotangsi et PM_US.Post ainsi que la variable PRES car ils ont tous les trois un VIF supérieur à 4. Nous remarquons également une forte colinéarité pour les facteurs DEWP, HUMI et TEMP qui ont un VIF supérieur à 10. Il y a donc 6 facteurs ayant des problèmes de colinéarité.

Lorsque l'on rencontre un problème de colinéarité avec des facteurs, une sélection de facteurs est nécessaire.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   653.59301  110.09542   5.937 3.25e-09 ***
PM_Caotangsi    0.31056    0.01592  19.511  < 2e-16 ***
PM_US.Post      0.71226    0.01658  42.955  < 2e-16 ***
DEWP            1.21868    0.41614   2.929 0.003431 **
HUMI           -0.77475    0.10354  -7.483 9.53e-14 ***
PRES           -0.57240    0.10600  -5.400 7.20e-08 ***
TEMP           -2.11029    0.40392  -5.225 1.87e-07 ***
Iws             0.13424    0.05789   2.319 0.020467 *
precipitation   0.34308    0.23596   1.454 0.146063
Iprec           0.15036    0.03911   3.844 0.000123 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.79 on 2954 degrees of freedom
Multiple R-squared:  0.9272,    Adjusted R-squared:  0.927
F-statistic:  4183 on 9 and 2954 DF,  p-value: < 2.2e-16
```

La commande *summary* nous permet de remarquer que la valeur de notre R^2 est très bonne (0.9272). Pour s'assurer de la significativité du modèle, nous regardons la dernière ligne du summary, la F-statistic. Il s'agit de la statistique de test pour le test de significativité du modèle. La p-value du test doit être très petite, sinon il vaut mieux arrêter tout de suite la procédure, ou voir où est le problème. Ici la p-value est très proche de 0 (2.2e-16), il n'y a donc pas de problème.

Nous allons alors pouvoir appliquer la méthode de régression par élimination (backward regression) afin de pouvoir faire une première sélection de facteurs. Cette méthode s'effectue pas à pas, en regardant à chaque étape quel est le meilleur modèle en se basant sur les valeurs des p-value. A chaque étape, on enlève le facteur ayant la plus forte p-value (supérieure à 0.1, c'est le niveau que nous nous fixons).

Nous enlevons alors le facteur precipitation car c'est le facteur ayant la plus forte p-value (0.146063). On recommence la régression sans cette variable puis on affiche à nouveau le summary. Nous nous arrêtons là car toutes les p-valeurs sont inférieures à 0.1 (notre seuil fixé)

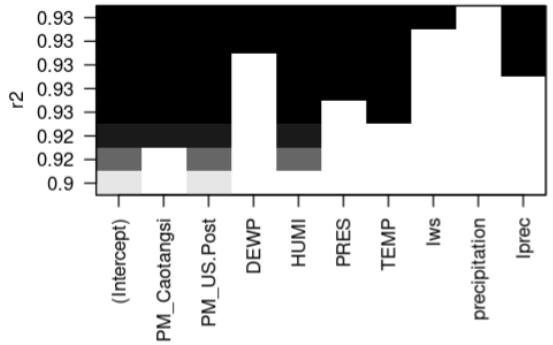
Le modèle retenu après avoir effectué cette régression par élimination serait donc un modèle contenant 8 facteurs (au lieu de 9).

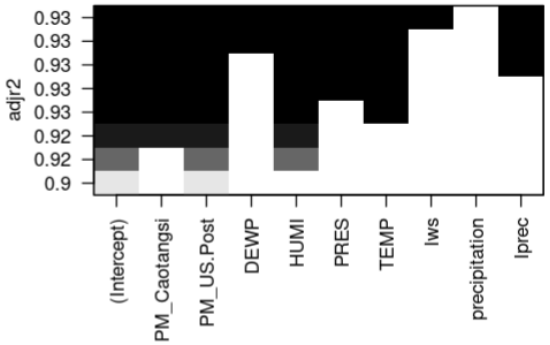
On peut comparer les R^2 entre le modèle retenu et le modèle complet. Les R^2 ont des valeurs qui se rapprochent énormément donc la variable est presque aussi bien expliquée par le modèle complet que par le modèle à 8 facteurs. On peut également comparer les R^2 ajustés des deux modèles. Le R^2 ajusté a l'avantage de prendre en compte le nombre de facteurs dans le modèle. On obtient respectivement 0,926 et 0,927 pour le modèle retenu et pour le modèle complet.

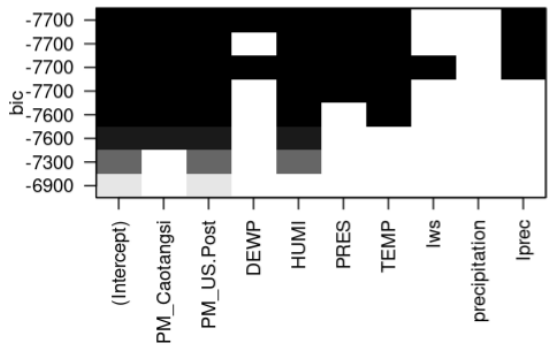
Nous avons également utilisé la méthode de la régression par ajout (forward regression). Pour ce modèle, on ajoute le facteur avec la p-value la plus petite, puis on recommence la régression jusqu'à ce que les p-valeurs des variables restantes soient supérieures à notre seuil fixé.

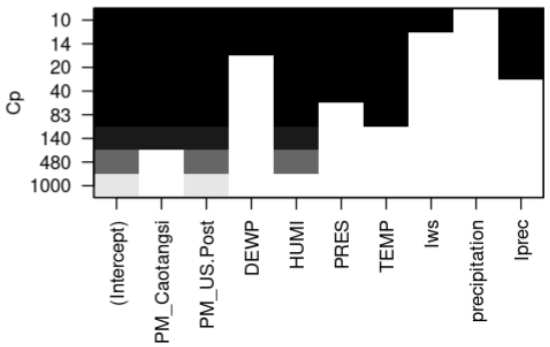
Nous obtenons alors le même modèle que nous avions obtenu avec la méthode de régression par élimination. Les R^2 (0.927191) et R^2 ajustés (0.926) sont égaux car on a obtenu les mêmes modèles avec le même nombre de facteurs.

Il peut être utile de comparer les résultats des différentes méthodes. Nous avons alors également procédé à une recherche exhaustive grâce à la commande *regsubsets.* Une procédure exhaustive consiste à examiner l'ensemble de tous les sous-modèles possibles et à sélectionner le meilleur d'entre eux selon un critère défini à l'avance. En théorie, ces procédures sont préférables aux méthodes pas à pas puisqu'elles mettent en concurrence beaucoup plus de modèles. Nous avons utilisé différents critères de sélection. Le premier critère de sélection est le critère du R^2 ajusté. Nous recherchons le modèle ayant le meilleur R^2 ajusté, il s'agit du modèle à 8 facteurs. Nous retenons alors grâce à ce premier critère de sélection un modèle à 8 facteurs. Le critère de Mallows nous pousse à préférer également un modèle à 8 facteurs. Le dernier critère de sélection que nous avons décidé d'utiliser et le critère BIC. Le modèle sélectionné au sens du critère BIC correspond au modèle à k facteurs pour lequel BIC(k) est le plus petit. Ici, il s'agit du modèle à 7 facteurs.
Ci-dessous la représentation graphique des résultats obtenus selon nos différents critères :









La plupart des critères de sélections ont retenu un modèle à 8 facteurs.

Pour pouvoir effectuer des prévisions nous avons choisi de prendre le modèle avec le moins de facteurs, en effet, cela permet de limiter les problèmes de colinéarité. Il s'agit donc de celui sélectionné par le critère de sélection BIC avec les 7 facteurs suivants : PM_Caotangsi, PM_US.Post, DEWP, HUMI, PRES, TEMP, Iprec. La variance étant quasiment égale pour tous les modèles nous n'avons pris que le critère du nombre de facteur en compte.

Pour de la modélisation nous recherchons le modèle ayant le R^2 le plus élevé. Ici le R^2 est égal pour tous les critères. Nous pouvons alors prendre n'importe quel modèle pour les modélisations.

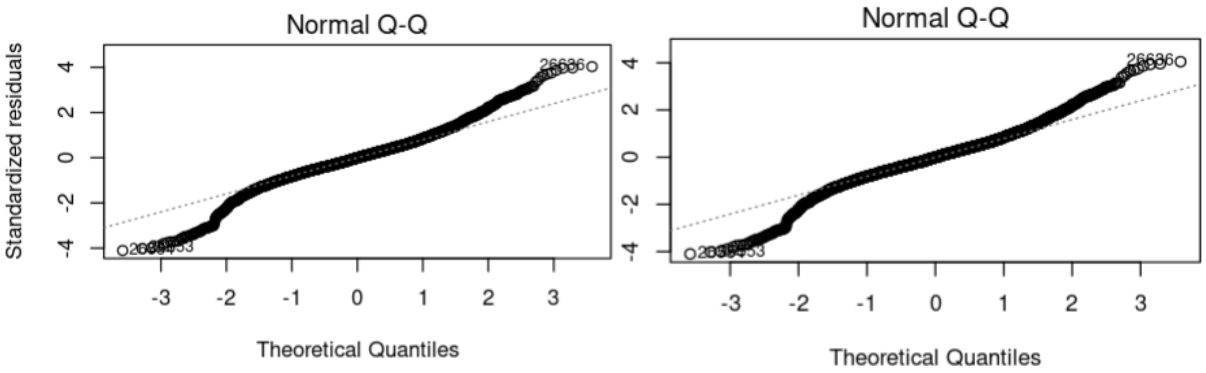
Nous effectuons à présent une nouvelle analyse des nouveaux modèles obtenus, c'est-à-dire celui du bic pour les prévisions et celui par exemple du R^2 pour la modélisation. Pour cela, nous allons tout d'abord étudier le vif de ces deux modèles.

```
> vif(r2)
PM_Caotangsi   PM_US.Post         DEWP         HUMI         PRES         TEMP        Iprec
    8.889348     9.471197    93.243724    28.434872     5.668815    77.011783     1.079850
         Iws
    1.078561
```

```
> vif(bic)
PM_Caotangsi   PM_US.Post         DEWP         HUMI         PRES         TEMP        Iprec
    8.859438     9.465053    93.225369    28.336981     5.532688    76.704447     1.079397
```

Nous remarquons que des problèmes de colinéarité persistent. Cependant, les vifs sont légèrement moins élevés pour le critère du bic.

Nous allons à présent étudier les résidus du R^2, à gauche, ainsi que du bic, à droite.



Sur ces graphiques, on peut voir que les résidus, peu importe le modèle, suivent plutôt bien une loi normal, malgré quelques écarts en haut et en bas.

Enfin nous terminons par effectuer à nouveaux des prévisions avec le modèle le plus adapté, le modèle bic.

```
   31560    31561    31562    31563    31564    31565    31566    31567    31568    31569
13.50303 22.49697 21.45628 17.36212 21.50702 26.23266 24.28367 24.44191 20.04175 20.67457
```

La MSE est plus faible (290 au lieu de 300). On en conclut alors que le critère du bic est meilleur pour les prévisions que le modèle complet.

Conclusion

Cette étude avait pour but d'analyser les variations de la pollution de la ville de Chengdu en fonctions d'autres variables, grâce à l'application de modèles linéaires. Nous avons dans un premier temps appliqué une régression multiple après avoir constaté que la régression linéaire était bien adaptée à nos données. Nous avons ensuite déterminé qu'il existait des liaisons linéaires entre le point de rosée, la pression et la température. Nous avons également déterminé que les deux variables de pollution en PM2,5 PM_Caotangsi et PM_US.post, l'humidité, la pression et la température étaient des variables significatives. La méthode de l'ANOVA nous a permis de conclure que les moyennes de pollution diffèrent selon la direction du vent, et plus précisément que la pollution est la plus faible quand le vent vient du nord-est. Enfin nous avons effectué une sélection de facteurs afin d'améliorer le modèle et définir les modèles les plus adaptés à la modélisation et à la prévision.

Abstract

Within the framework of our second year of study in Data Science, we had to write this scientific report that relates the hypothesis we have made, the methods we used and the results we got. We had a data set of 52584 datas and 17 variables. These data are measures made in Chengdu, China, between 2010 and 2015. The aim of our study was to analyse the variations in pollution in Chengdu depending on other variables, using linear models. First we used the ggpairs() function, which makes a matrix of plots with our data set. It produces scatter plots for each pair of variables, density plots for each variable, and also shows the Pearson Correlation Coefficients of each pair of variables. Those plots showed that there was linear links between pressure, temperature and the dew point. Residuals are important when determining the quality of a model, we used a normal QQ plot to spot errors we could find and proceeded to erase them. The summary command showed that the captors counting the concentration of Particulate Matter 2.5, humidity, pressure and temperature were significant variables. To compare the means of each wind directions we used the ANOVA test, this test showed that the mean of the pollution was smaller when the wind came from the North-East. In order to refine our analysis and spot differences we decided to use both Tukey and Scheffe test. Those two tests also proved that the pollution was smaller when the wind came from the North-East. To detect multicollinearity we used the Variance Inflation Factor. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model. The results showed that there was 6 factors that had problems of multicollinearity. When this occurs, we need to select factors to improve the model. To finish our analysis we picked models to make predictions and built intervals.