CentraleSupélec 2A : Statistique avancées

Avril - Mai 2022

DM : Compte rendu

Analyse Statistique avec R

Auteurs :
Marouane BATTACH
Youssef BENSSY

Professeur :
Laura VUDUC

Résumé

Un des enjeux majeurs de la santé publique est la pollution de l'air. En effet, de nombreuses études épidémiologiques mettent en évidence l'influence sur la santé de certains composés chimiques présents dans l'atmosphère.Des associations de surveillance de la qualité de l'air mesurent la concentration des polluants ainsi que les conditions météorologiques comme la température, la nébulosité, le vent, les chutes de pluie. L'une des missions de ces associations est de construire des modèles de prévision de la pollution du lendemain à partir des données disponibles du jour. L'objectif de cette étude est de mettre en évidence l'influence de certains paramètres sur la pollution et différentes variables observées ou leur prévision.

Mots clés : Statistiques, R, Pollution

Table des matières

Table des figures

1 Question 1 : représentations unidimensionnelles des données

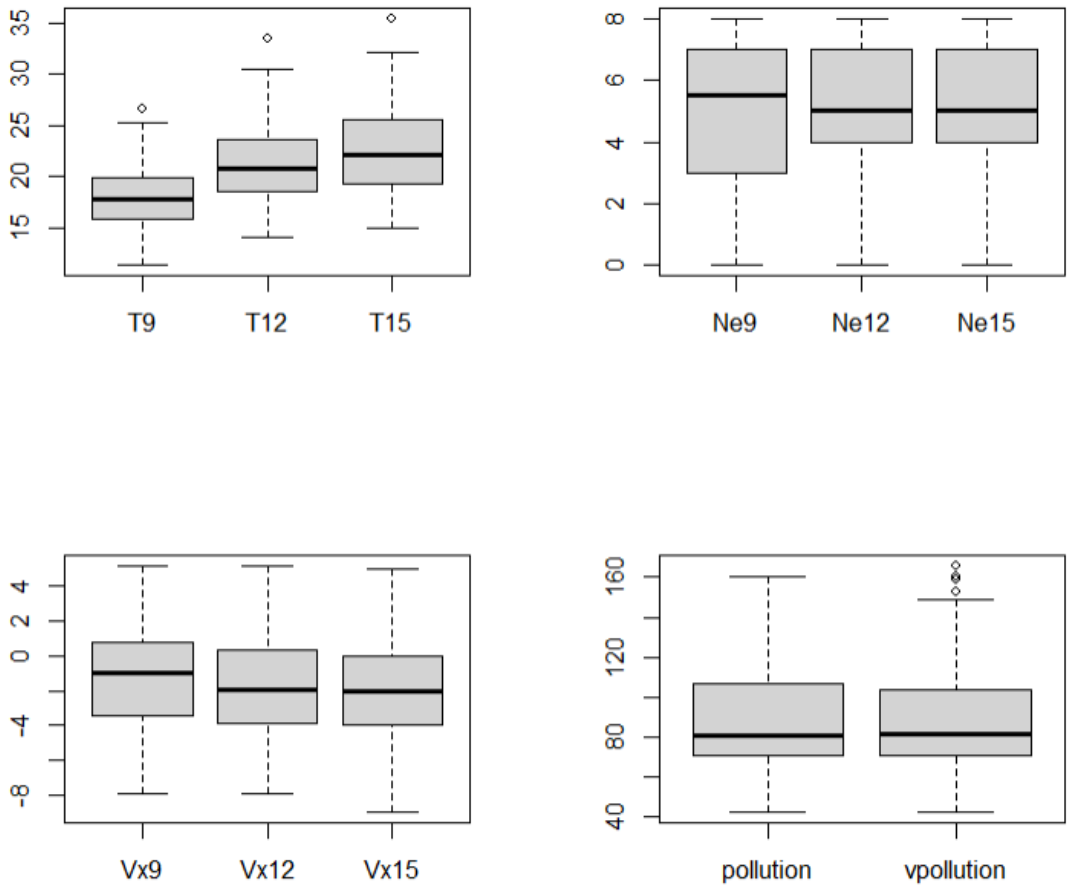

FIGURE 1 – boites à moustaches des variables explicatives

Commentaire :

— la moyenne et l'écart_type des variables de température sont augmentent avec le temps

— la nébulosité mesurée à 9 :00 a une moyenne et un écart_type plus grands que les autres heures de la journée

— La composante du vent E-O n'a pas une moyenne fixe pendant les heures de la journée

— la pollution et la pollution mesurée la veille ont des caractéristiques statistiques très similaires.

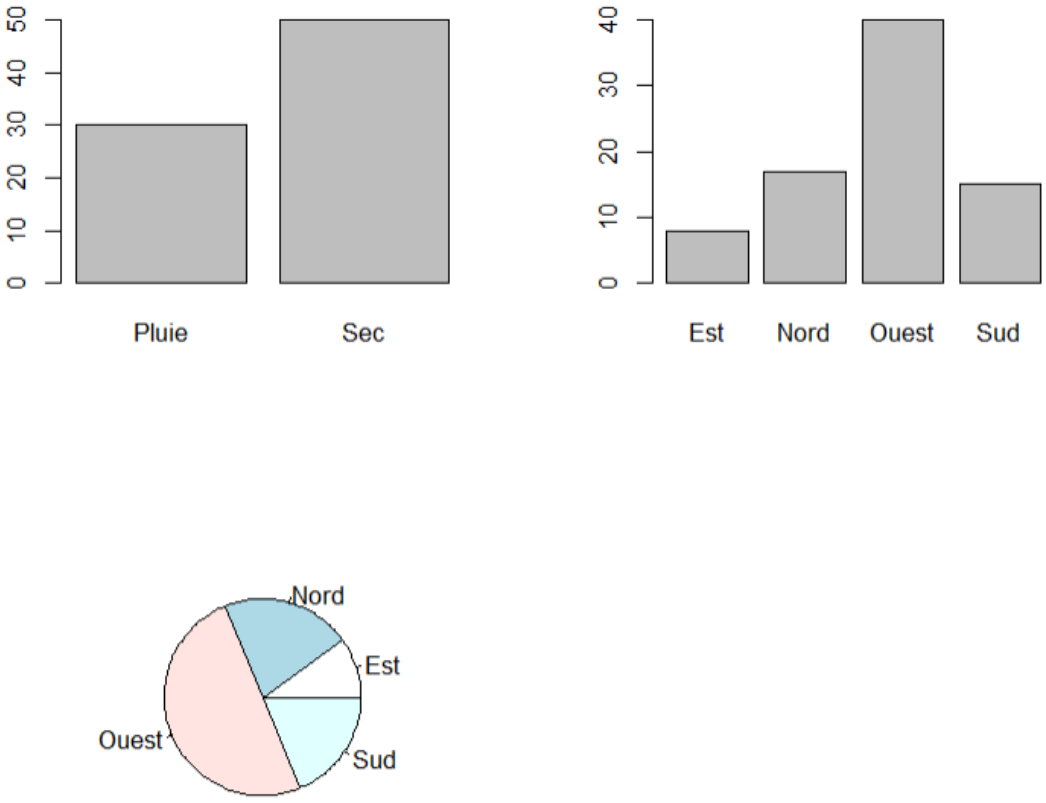

FIGURE 2 – Fréquences des modalités des variables qualitatives.

Commentaire :
— Ceux ci sont des variables catégoriques ou des facteurs
— le premier facteur pluie a deux modalités : Pluie et Sec
— le modalité Sec dans le facteur pluie est plus fréquente que la modalité Pluie
— le deuxième facteur Vent a 4 modalités : Est, Nord, Ouest, Sud
— la modalité Ouest dans le facteur Vent est plus fréquente que les autres modalités

2 Question 2 : corrélations entre les variables

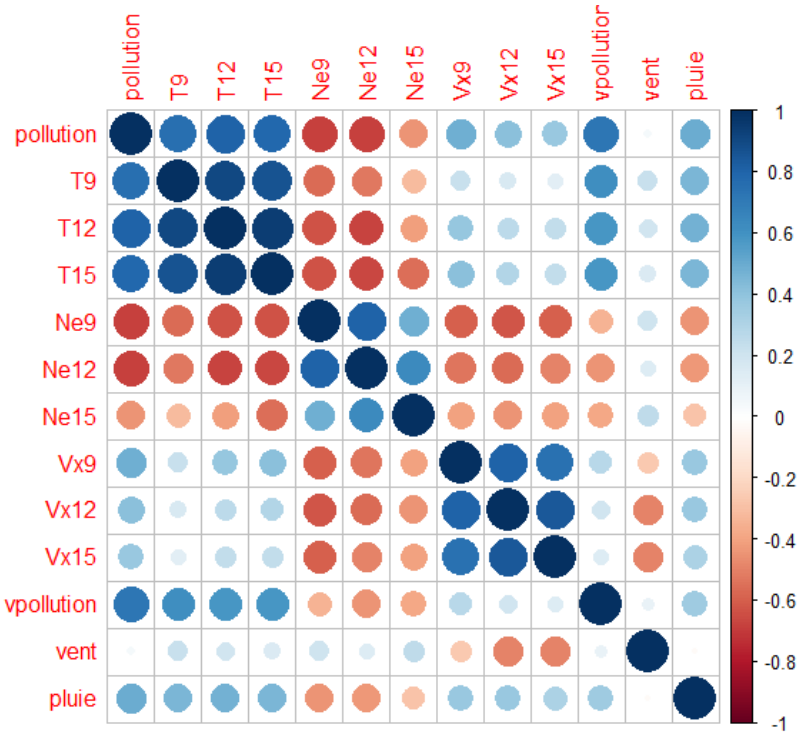

FIGURE 3 – Matrice de corrélation

Commentaire :

— On remarque une corrélation positive entre la pollution et les variables de la température et la pollution de la veille
— Une corrélation négative entre la pollution et les variables de la nébulosité
— une corrélation positive entre les variables de la température.
— une corrélation positive entre les variables de la nébulosité

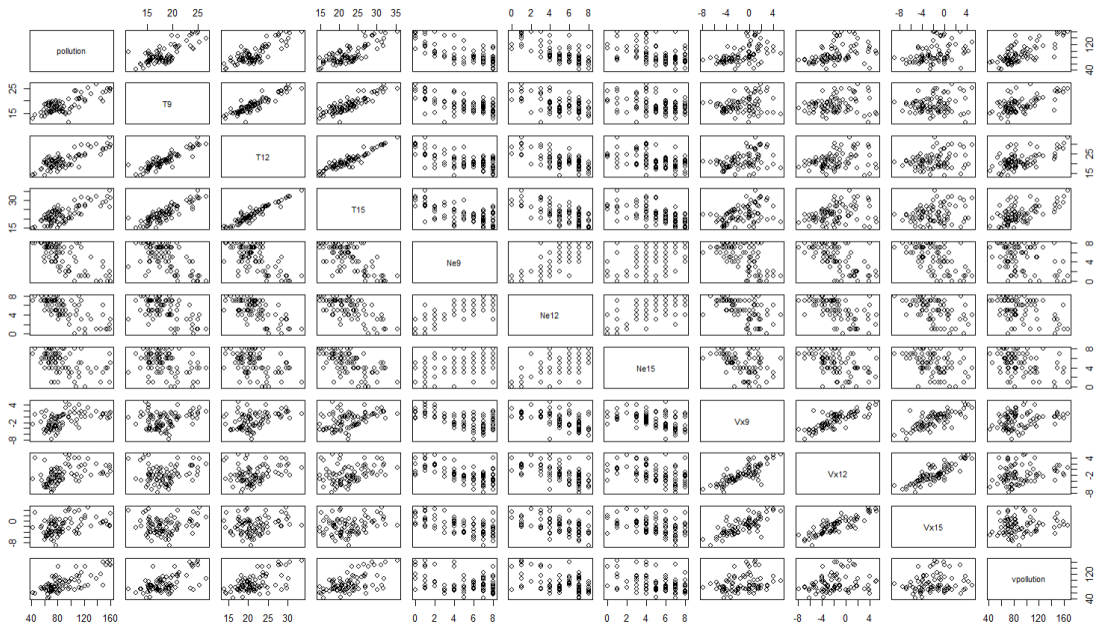

FIGURE 4 – Représentation bidimensionnelle

Commentaire : cette figure présente d'une autre façon la corrélation les variables.

3 Question 3 : pollution ~Vx12 * Vent

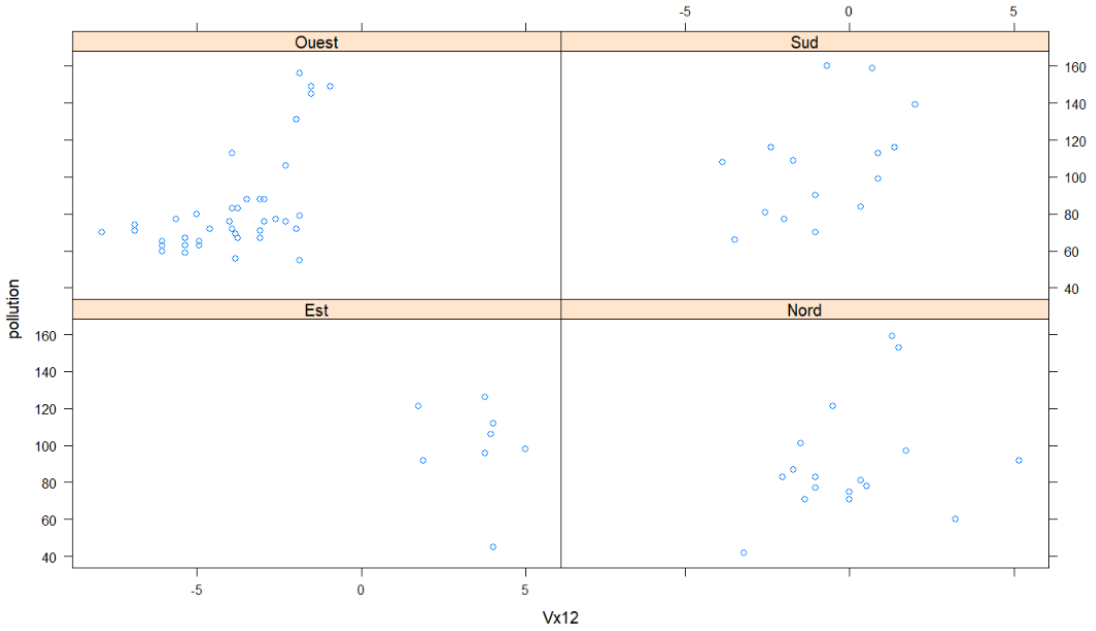

FIGURE 5 – La pollution en fonction de la variable Vx12 et en fonction de la direction du vent.

Commentaire : La variable Vx12 contient une partie d'information qui se trouve dans le facteur vent.

4 Question 4 : modèle de régression

On utilise pour la régression le modèle ANCOVA à deux facteurs sous sa forme singulière :

$$Y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \sum_{l=1}^{p} (\beta_l + a_{lj} + b_{lk} + (ab)_{ljk}) * x_l + \varepsilon_{ijk}$$

où les ε_{ijk} sont i.i.d suivent une normale et $\quad i \in [\![1, n_{jk}]\!], \quad j \in [\![1,4]\!], \quad k \in [\![1,2]\!]$

— Y est la variable quantitative "pollution"
— x_l pour $l \in [\![1, p = 11]\!]$ sont les variables quantitatives suivantes :

$$T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 + Vx12 + Vx15 + vpollution$$

— $n = \sum_{k=1}^{2} \sum_{j=1}^{4} n_{jK} = 80$ est la taille de l'échantillon
— R met automatiquement le premier niveau à 0 pour les deux facteurs
— nombre de paramètres à estimer est 1+3+1+3+10*(1+3+1+3) =88.
— $\varepsilon_{i,j,k}$ erreur associée à l'observation numéro i du niveau j du facteur, $\varepsilon_{i,j,k} \overset{\text{i.i.d}}{\sim} \mathcal{N}\left(0, \sigma^2\right), \sigma^2$ à estimer (modèle homoscédastique),

5 Question 5 : Le modèle linéaire sur R

Après avoir exécuté le calcul de la régression linéaire, on obtient la sortie suivante :

```
Call:
lm(formula = pollution ~ (T9 + T12 + T15 + Ne9 + Ne12 + Ne15 +
    Vx9 + Vx12 + Vx15 + vpollution) * pluie * vent, data = data_poll)

Residuals:
    Min      1Q  Median      3Q     Max
-21.191  -2.505   0.000   2.149  27.373

Coefficients: (30 not defined because of singularities)
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           17.8712   257.7989   0.069   0.9454
T9                    -1.7187    24.0140  -0.072   0.9436
T12                   19.4053    64.3996   0.301   0.7660
T15                  -11.1839    78.6083  -0.142   0.8882
Ne9                   -4.6119    19.4676  -0.237   0.8149
Ne12                   6.7041    31.0687   0.216   0.8311
Ne15                   0.6995    40.4607   0.017   0.9864
Vx9                  -15.5362    35.5701  -0.437   0.6665
Vx12                 -19.7204    11.3846  -1.732   0.0972 .
Vx15                   6.4718     7.1413   0.906   0.3746
vpollution            -0.2075     0.3877  -0.535   0.5978
pluieSec              69.4040   160.2400   0.433   0.6691
ventNord             187.0248   360.7162   0.518   0.6093
ventOuest             22.1016   258.4357   0.086   0.9326
ventSud              -65.6435   264.1123  -0.249   0.8060
T9:pluieSec          -16.1445     9.7194  -1.661   0.1109
T12:pluieSec          17.8762     8.1321   2.198   0.0387 *
T15:pluieSec          -7.3666     6.8624  -1.073   0.2947
Ne9:pluieSec          -7.1710     4.6162  -1.553   0.1346
Ne12:pluieSec          2.9041     5.5980   0.519   0.6091
Ne15:pluieSec         -3.9644     5.9497  -0.666   0.5121
Vx9:pluieSec          -1.4280     4.1428  -0.345   0.7336
Vx12:pluieSec         12.0568     5.4663   2.206   0.0382 *
Vx15:pluieSec         -6.8908     4.3987  -1.567   0.1315
vpollution:pluieSec    0.6580     0.3193   2.061   0.0513
T9:ventNord         -204.8623   156.0356  -1.313   0.2027
T9:ventOuest          1.7970    23.9171   0.075   0.9408
T9:ventSud           19.0894    21.7434   0.878   0.3895
T12:ventNord       -169.9244    94.8931  -1.791   0.0871 .
T12:ventOuest       -21.6481    64.7029  -0.335   0.7411
T12:ventSud         -39.7045    65.5124  -0.606   0.5507
T15:ventNord        316.0218   191.7877   1.648   0.1136
T15:ventOuest        13.7661    78.9520   0.174   0.8632
T15:ventSud          21.7585    79.3889   0.274   0.7866
Ne9:ventNord         10.2826    23.9542   0.429   0.6719
Ne9:ventOuest         6.5712    19.4146   0.338   0.7382
Ne9:ventSud          12.5147    19.5289   0.641   0.5283
Ne12:ventNord       -11.0149    33.6757  -0.327   0.7467
Ne12:ventOuest       -8.1094    31.1195  -0.261   0.7968
Ne12:ventSud        -20.3061    35.6905  -0.569   0.5752
Ne15:ventNord         3.7797    39.5628   0.096   0.9248
Ne15:ventOuest       -1.4550    40.0655  -0.036   0.9714
Ne15:ventSud          8.1776    41.3785   0.198   0.8451
Vx9:ventNord         18.8268    36.9172   0.510   0.6151
Vx9:ventOuest        13.1958    35.5174   0.372   0.7138
Vx9:ventSud          26.7009    38.1160   0.701   0.4909
Vx12:ventNord         5.0609    10.6109   0.477   0.6381
Vx12:ventOuest       19.3283    10.9315   1.768   0.0909 .
Vx12:ventSud             NA         NA      NA       NA
Vx15:ventNord         0.6991     9.0452   0.077   0.9391
Vx15:ventOuest       -6.3437     6.7131  -0.945   0.3549
Vx15:ventSud             NA         NA      NA       NA
vpollution:ventNord   0.2661     0.8462   0.314   0.7561
vpollution:ventOuest  0.3734     0.3249   1.149   0.2628
vpollution:ventSud       NA         NA      NA       NA
pluieSec:ventNord   -268.5605   317.6661  -0.845   0.4070
pluieSec:ventOuest    65.3732   159.2594   0.410   0.6854
pluieSec:ventSud         NA         NA      NA       NA
T9:pluieSec:ventNord  236.5279   153.9964   1.536   0.1388
T9:pluieSec:ventOuest  -1.5842     8.7485  -0.181   0.8580
T9:pluieSec:ventSud         NA         NA      NA       NA
T12:pluieSec:ventNord  128.7102    66.7296   1.929   0.0668 .
T12:pluieSec:ventOuest      NA         NA      NA       NA
T12:pluieSec:ventSud        NA         NA      NA       NA
T15:pluieSec:ventNord -303.7091   174.0226  -1.745   0.0949 .
T15:pluieSec:ventOuest      NA         NA      NA       NA
T15:pluieSec:ventSud        NA         NA      NA       NA
Ne9:pluieSec:ventNord       NA         NA      NA       NA
Ne9:pluieSec:ventOuest      NA         NA      NA       NA
Ne9:pluieSec:ventSud        NA         NA      NA       NA
Ne12:pluieSec:ventNord      NA         NA      NA       NA
Ne12:pluieSec:ventOuest     NA         NA      NA       NA
Ne12:pluieSec:ventSud       NA         NA      NA       NA
Ne15:pluieSec:ventNord      NA         NA      NA       NA
Ne15:pluieSec:ventOuest     NA         NA      NA       NA
Ne15:pluieSec:ventSud       NA         NA      NA       NA
Vx9:pluieSec:ventNord       NA         NA      NA       NA
Vx9:pluieSec:ventOuest      NA         NA      NA       NA
Vx9:pluieSec:ventSud        NA         NA      NA       NA
Vx12:pluieSec:ventNord      NA         NA      NA       NA
Vx12:pluieSec:ventOuest     NA         NA      NA       NA
Vx12:pluieSec:ventSud       NA         NA      NA       NA
Vx15:pluieSec:ventNord      NA         NA      NA       NA
Vx15:pluieSec:ventOuest     NA         NA      NA       NA
Vx15:pluieSec:ventSud       NA         NA      NA       NA
vpollution:pluieSec:ventNord  NA       NA      NA       NA
vpollution:pluieSec:ventOuest NA       NA      NA       NA
vpollution:pluieSec:ventSud   NA       NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.79 on 22 degrees of freedom
Multiple R-squared:  0.9457,    Adjusted R-squared:  0.8051
F-statistic: 6.726 on 57 and 22 DF,  p-value: 4.237e-06
```

FIGURE 6 – Sortie de la régression linéaire sur R

Commentaire :
— il y a bien 88 paramètres.
— Certains estimations ont la valeur "NA".Cela indique que deux variables prédictives ou plus dans le modèle ont une relation linéaire parfaite et

que, par conséquent, tous les coefficients de régression du modèle ne peuvent pas être estimés. Par exemple dans le graphe suivant qui trace la corrélation entre les variable pour le niveaux 'Sud' il y a une forte corrélation entre T15 e T12

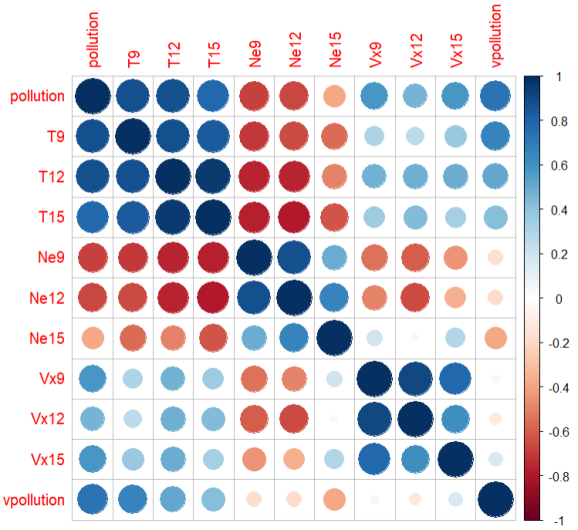

FIGURE 7 – Matrice de corrélation pour le niveau "Sud" du facteur vent

6 Question 6 : La variable Vx12

Dans la figure suivante, on extrait les estimations liées la variable Vx12 Toutes les informations demandées se trouvent la figure.

```
Call:
lm(formula = pollution ~ (T9 + T12 + T15 + Ne9 + Ne12 + Ne15 +
    Vx9 + Vx12 + Vx15 + vpollution) * pluie * vent, data = data_poll)

                            Estimate Std. Error t value Pr(>|t|)

Vx12                       -19.7204    11.3846  -1.732   0.0972 .
Vx12:pluieSec               12.0568     5.4663   2.206   0.0382 *
Vx12:ventNord                5.0609    10.6109   0.477   0.6381
Vx12:ventOuest              19.3283    10.9315   1.768   0.0909 .
Vx12:ventSud                    NA         NA      NA       NA
Vx12:pluieSec:ventNord          NA         NA      NA       NA
Vx12:pluieSec:ventOuest         NA         NA      NA       NA
Vx12:pluieSec:ventSud           NA         NA      NA       NA
```

FIGURE 8 – Les estimations liées à la variable Vx12

Le test de significavité est construit de la façon suivante :
Pour tout $j \in [\![1, p]\!]$, $(H_0)\ \beta_j = 0$ contre $(H_1)\ \beta_j \neq 0$. La statistique de test de significativité et la région de rejet s'écrivent :

$$T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{\left[(X^T X)^{-1}\right]_{j,j}}} \sim T(n-p) \quad \text{et} \quad R_\alpha = \left\{ |T| > t_{n-p}\left(1 - \frac{\alpha}{2}\right) \right\}$$

Par symétrie de la loi de Student,

$$\begin{aligned} p - value &= \mathbb{P}\left(|T(n-p)| > |t_{obs}|\right) = \mathbb{P}\left(T(n-p) > |t_{obs}| \text{ ou } T(n-p) < -|t_{obs}|\right) \\ &= 2\mathbb{P}\left(T(n-p) < -|t_{obs}|\right) \\ &= 2\mathbb{P}\left(T(n-p) > |t_{obs}|\right) \end{aligned}$$

La p-valeur pour pour Vx12 :pluieSec est inférieur à 5% donc on rejette l'hypothèse nulle pour cette variable Les autre paramètres ont une p-valeur plus grande que 5%, don on garde l'hypothèse nulle pour ces paramètres.

7 Question 7 : Vx15

La figure suivante présente un extrait de la sortie de la régression linéaire sur R liée à la variables Vx15

```
Call:
lm(formula = pollution ~ (T9 + T12 + T15 + Ne9 + Ne12 + Ne15 +
    Vx9 + Vx12 + Vx15 + vpollution) * pluie * vent, data = data_poll)

                              Estimate Std. Error t value Pr(>|t|)
Vx15                            6.4718     7.1413   0.906   0.3746
Vx15:pluieSec                  -6.8908     4.3987  -1.567   0.1315
Vx15:ventNord                   0.6991     9.0452   0.077   0.9391
Vx15:ventOuest                 -6.3437     6.7131  -0.945   0.3549
Vx15:ventSud                       NA         NA      NA       NA
Vx15:pluieSec:ventNord             NA         NA      NA       NA
Vx15:pluieSec:ventOuest            NA         NA      NA       NA
Vx15:pluieSec:ventSud              NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

FIGURE 9 – Les estimations liées à la variables Vx15

les p-values sont tous grands à 5%, donc on garde les hypothèses nulles

8 Question 8 : significativité globale de la régression

Test de significativité globale de la régression : $M_p = M_1$ est le modèle i.i.d. (sans covariables) et M_p le modèle de régression complet

le test de significativité globale de la régression est construit de la façon suivante :

- SCRH0 $= \left\|Y - \hat{\beta}_0 \mathbb{1}\right\|^2 = \|Y - \mathbb{E}(Y)\|^2$ (modèle iid : $Y = \beta_0 \mathbb{1} + \varepsilon$) et $SCRH1 = \|Y - X\hat{\beta}\|^2$ (modèle complet)
- $F = \frac{(SCRH0 - SCRH1)/(p-q)}{SCRH1/(n-p)}$ suit la loi de Fisher
- quantile de la loi $F_{p-q,n-p}(1-\alpha)$: $f_{p-q,n-p}(1-\alpha)4$
- $R_\alpha = \{F > f_{p-q,n-p}(1-\alpha)\}$
- $p - value = \mathbb{P}\left(F_{p-q,n-p} > f_{obs}\right)$

```
Analysis of Variance Table

Model 1: pollution ~ 1
Model 2: pollution ~ (T9 + T12 + T15 + Ne9 + Ne12 + Ne15 + Vx9 + Vx12 +
    Vx15 + vpollution) * pluie * vent
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1     79 66260
2     22  3596 57     62664 6.7256 4.237e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FIGURE 10 – Test de significativité globale de la régression

la p-valeur=4.237e-06 $<<$ 5% $\Rightarrow$ on rejette (H_0) et on garde le grand modèle pour l'instant.

9 Question 9 : Sélection des variables

On utilise la méthode ascendante (forward). Dans la sélection ascendante avec des p-values, nous commençons avec un modèle qui n'a pas de prédicteurs, puis nous ajustons un modèle pour chaque prédicteur possible, en identifiant le modèle où la valeur de p du prédicteur correspondant est la plus petite. Si cette p-value est inférieure à $= 0.05$ Nous l'ajoutons au modèle et répétons le processus,

en considérant s'il faut ajouter plus de variables une à la fois. Lorsqu'aucun des prédicteurs restants ne peut être ajouté au modèle et qu'il n'a pas une p-value inférieure à 0,05, nous arrêtons d'ajouter des variables et le modèle actuel serait notre modèle le mieux ajusté.

On utlise la package "olsrr" qui nous donne directement les étapes de la sélection

la figure suivante montre les étapes de la sélection :

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	T12	0.6541	0.6496	64.2259	685.6542	17.1425
2	vpollution	0.7526	0.7462	26.2859	660.8359	14.5909
3	Ne9	0.8095	0.8019	5.2374	641.9446	12.8888

FIGURE 11 – régression ascendante utilisant la p-value

Commentaire :

— Première étape : on prend la variable T12
— Deuxième étape : On prend la variable vpollution
— Troisième étape : On prend la variable Ne9
— Après, il n y aucune variable avec une p-value inférieure à 0.05 donc on arrête l'ajustement.

10 Question 10 : coefficient de détermination

Le coefficient de détermination est défini par :

$$R^2 = \frac{SCM\left(M_0, M_p\right)}{SCR\left(M_0\right)} = \frac{\|\widehat{Y} - \bar{Y}\mathbb{1}\|^2}{\|Y - \bar{Y}\mathbb{1}\|^2}$$

Le coefficient de détermination R^2 s'interprète comme la part de variance expliquée par les régresseurs supplémentaires.

— $R^2 = 1 \Leftrightarrow \widehat{Y} = Y \rightarrow$ bon ajustement du modèle
— $R = 0 \Leftrightarrow \widehat{Y} = \bar{Y}\mathbb{1}$.

Le coefficient de détermination donne uniquement une indication sur la linéarité de la relation :

— si les observations sont très nombreuses, on ne peut pas espérer un grand R^2 ;
— R^2 est d'autant plus grand que le modèle comporte beaucoup de paramètres.

Le coefficient de détermination augmente donc avec le nombre de variables : si on doit choisir entre un modèle avec p variables et le même modèle mais avec une variable de plus en se fiant au R^2, on choisira toujours le modèle avec le plus de variables :

— ne prend en compte que l'ajustement du modèle et pas le prix de cet ajustement, i.e. le nombre de paramètres à estimer
— coefficient de détermination ajusté R^2_{adj} : version du R^2, ajusté au nombre de paramètres à estimer.

$$R^2_{adj} = 1 - \frac{n-1}{n-p-1}\frac{\|\hat{\varepsilon}\|^2}{\|Y - \bar{Y}\mathbb{1}\|^2} = 1 - \frac{n-1}{n-p-1}\left(1 - R^2\right)$$

Dans la figure 11, on remarque que le R_{adj} augmente au fur et au mesure qu'on avance dans les étapes de la sélection. Ceci montre que notre stratégie de sélection est efficace vu que le R_{adj} ne dépend pas du nombre de variables choisies.

On obtient un R_{adj} à celui du modèle qui prend en compte les interactions avec les variables catégoriques.

11 Question 11 : sélection de modèle par StepAIC

le critère AIC (Akaike Information Criterion) est défini par

$AIC\left(M_q\right) = -2\log L\left(M_q\right) + 2q$. (où L est la vraisemblance du modèle)

La sélection du modèle par stepAIC cherchera à minimiser le critère AIC. Après avoir exécuté la commande StepAIC sur le modèle linéaire précédant, on obtient la sortie suivante :

```
Call:
lm(formula = pollution ~ T9 + T12 + T15 + Ne9 + Ne15 + Vx9 +
    Vx12 + Vx15 + vpollution + pluie + vent + T9:pluie + T12:pluie +
    T15:pluie + Ne9:pluie + Vx12:pluie + Vx15:pluie + vpollution:pluie +
    T9:vent + T12:vent + T15:vent + Vx9:vent + Vx12:vent + Vx15:vent +
    pluie:vent + T9:pluie:vent + T12:pluie:vent + T15:pluie:vent,
    data = data_poll)

Residuals:
   Min     1Q Median     3Q    Max
-22.293  -2.730   0.000   3.052  25.747

Coefficients: (5 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          32.5819    86.4481   0.377  0.70859
T9                   -0.1587     5.0178  -0.032  0.97495
T12                  23.8693    16.7032   1.429  0.16212
T15                 -13.4429    14.4421  -0.931  0.35851
Ne9                   2.0682     2.6441   0.782  0.43952
Ne15                 -1.6437     1.6130  -1.019  0.31537
Vx9                  21.1526    17.4200   1.214  0.23301
Vx12                -45.0407    18.4668  -2.439  0.02010 *
Vx15                  8.8313    10.0932   0.875  0.38772
vpollution            0.1629     0.1769   0.921  0.36358
pluieSec            -41.5502    84.8873  -0.489  0.62765
ventNord             28.8210   164.2893   0.175  0.86178
ventOuest            -0.9002    87.5482  -0.010  0.99186
ventSud             -53.4213    60.0645  -0.889  0.38004
T9:pluieSec         -14.1259     6.6667  -2.119  0.04149 *
T12:pluieSec         10.2829     5.1498   1.997  0.05391 .
T15:pluieSec         -2.9398     3.7496  -0.784  0.43844
Ne9:pluieSec         -4.7998     2.9985  -1.601  0.11869
Vx12:pluieSec        11.3089     4.1359   2.734  0.00985 **
Vx15:pluieSec        -4.2253     2.9821  -1.417  0.16561
vpollution:pluieSec   0.4279     0.2070   2.068  0.04636 *
T9:ventNord        -90.4742    82.7725  -1.093  0.28206
T9:ventOuest        -0.1220     5.2353  -0.023  0.98154
T9:ventSud          12.6576     4.7073   2.689  0.01102 *
T12:ventNord      -128.2524    48.5305  -2.643  0.01235 *
T12:ventOuest      -25.3728    16.7697  -1.513  0.13952
T12:ventSud        -35.5812    18.5737  -1.916  0.06385 .
T15:ventNord       186.5546   102.4199   1.821  0.07734 .
T15:ventOuest       15.7954    14.2217   1.111  0.27451
T15:ventSud         18.7473    14.9231   1.256  0.21759
Vx9:ventNord       -17.9562    17.1552  -1.047  0.30263
Vx9:ventOuest      -24.0960    17.3725  -1.387  0.17446
Vx9:ventSud        -19.7601    19.2603  -1.026  0.31216
Vx12:ventNord       30.4796    17.8693   1.706  0.09719 .
Vx12:ventOuest      44.7782    18.2356   2.456  0.01934 *
Vx12:ventSud        36.1755    18.9761   1.906  0.06508 .
Vx15:ventNord       -5.3625    10.0720  -0.532  0.59790
Vx15:ventOuest      -8.4529     9.9525  -0.849  0.40164
Vx15:ventSud        -0.9511    10.1480  -0.094  0.92588
pluieSec:ventNord   -2.7036   164.2180  -0.016  0.98696
pluieSec:ventOuest 115.2710    93.8872   1.228  0.22797
pluieSec:ventSud   142.5445    79.4350   1.794  0.08163 .
T9:pluieSec:ventNord 118.4876  82.7723   1.431  0.16142
T9:pluieSec:ventOuest  3.6758   5.7722   0.637  0.52851
T9:pluieSec:ventSud       NA       NA      NA       NA
T12:pluieSec:ventNord 92.5391   46.0089   2.011  0.05227 .
T12:pluieSec:ventOuest    NA       NA      NA       NA
T12:pluieSec:ventSud      NA       NA      NA       NA
T15:pluieSec:ventNord -178.1435 101.7552  -1.751  0.08902 .
T15:pluieSec:ventOuest    NA       NA      NA       NA
T15:pluieSec:ventSud      NA       NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.24 on 34 degrees of freedom
Multiple R-squared:  0.9352,    Adjusted R-squared:  0.8494
F-statistic:  10.9 on 45 and 34 DF,  p-value: 6.256e-11
```

FIGURE 12 – sélection de modèle avec la commande stepAIC

Commentaire :

— La minimisation du critère AIC supprime plusieurs interactions des variables quantitatives avec les deux facteurs.

— le R_{adj} est relativement supérieur aux modèles précédents

la figure suivante montre le modèle qui minimise le critère BIC (Bayesian Information Criterion)

$$BIC\left(M_q\right) = -2\log L\left(M_q\right) + q\log(n).$$

```
Call:
lm(formula = pollution ~ T9 + T15 + Ne9 + vpollution + vent +
    T9:vent + T15:vent, data = data_poll)

Residuals:
    Min      1Q  Median      3Q     Max
-25.989  -7.971  -0.706   6.871  41.242

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    14.77103   22.94607   0.644 0.521982
T9             -3.03949    2.42524  -1.253 0.214528
T15             4.68563    2.17422   2.155 0.034808 *
Ne9            -2.95058    0.88141  -3.348 0.001350 **
vpollution      0.43893    0.07646   5.741 2.58e-07 ***
ventNord      -22.26716   27.87713  -0.799 0.427294
ventOuest      -4.78458   27.71032  -0.173 0.863444
ventSud         2.09172   30.04096   0.070 0.944700
T9:ventNord    12.59255    3.57409   3.523 0.000779 ***
T9:ventOuest    2.96522    2.93829   1.009 0.316582
T9:ventSud      4.23086    3.13121   1.351 0.181249
T15:ventNord   -9.31740    2.86613  -3.251 0.001813 **
T15:ventOuest  -2.21423    2.31560  -0.956 0.342448
T15:ventSud    -3.41608    2.47671  -1.379 0.172465
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.84 on 66 degrees of freedom
Multiple R-squared:  0.8358,    Adjusted R-squared:  0.8035
F-statistic: 25.84 on 13 and 66 DF,  p-value: < 2.2e-16
```

FIGURE 13 – sélection de modèle avec le critère BIC

On garde ce modèle car il est le modèle le plus simple avec un bon coefficient d'ajustement et un critère BIC minimal.

12 Question 12 : Validation du modèle

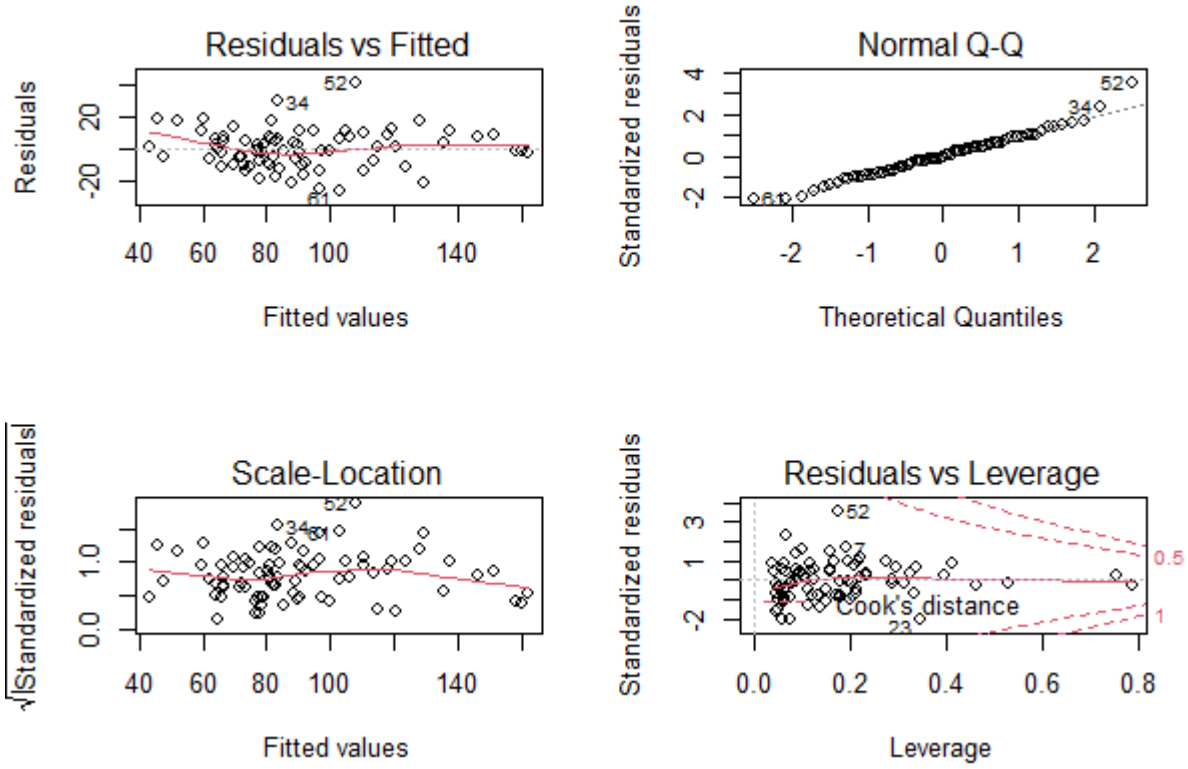

FIGURE 14 – graphes de validation du modèle

— Residuals vs Fitted :le nuage de points est relativement sans structure
— Normal Q-Q : les points sont alignés sur la première bissectrice
— Scale-Location : On a un nuage de points sans structure entre $[-2, 2]$
— Cook's distance : les points sont entre les lignes pointillées correspondant à 1.

13 Question 13 : Conclusion

— Le modèle final associe la pollution à la température de 9 :00 et de 15 :00 avec des termes d'interaction avec le vent.
— La pluie est absente dans le modèle final

— les variables Vx sont aussi absentes dans le modèle final mais la variable catégorique vent est présente. Ceci peut être du au fait que le facteur vent et les variables Vx partagent une partie d'information.
— Finalement Il peut de ne pas avoir une relation de causalité entre les variables explicatives et la variable pollution. le modèle de régression linéaire ne fait que associer statistiquement les variables entre eux.