TP4 : Modèles de régression linéaire

Analyse du jeu de données « Ozone »

Le jeu de données que nous allons utiliser dans le cadre de ce TP concerne la qualité de l'air.

Contexte : Les problématiques touchant à la pollution de l'air représentent un enjeu majeur en termes de santé publique car certains composés chimiques comme le dioxyde de souffre (SO2), le dioxyde d'ozone (NO2) ou l'ozone(O3) ont un impact négatif sur la santé des individus, et en particulier des citadins.
Des associations de surveillance de la qualité de l'air existent sur tout le territoire français et mesurent quotidiennement la concentration de différents polluants, ainsi que certains paramètres météorologiques (température, nébulosité, intensité et sens du vent, précipitation...).

Présentation du jeu de données :

Le jeu de données dont nous disposons est composé de 38 séries d'observations collectées dans la ville de Rennes entre début juin et fin octobre 2001 (l'identifiant unique de chaque observation correspond à la date de mesure).

Les 6 variables qu'il contient sont :

- La concentration maximale en ozone (en µg/m^3) enregistrée le jour J : maxO3.
- La température (T) enregistrée à 12h : T12.
- La nébulosité (Ne) = couverture nuageuse enregistrée à 12h : Ne12.
- La projection du vent sur l'axe Est-Ouest (Vx) à 9h : Vx9.
- La direction du vent (vent), variable à 4 modalités : Nord, Sud, Est, Ouest.
- Une variable (pluie) indiquant s'il a plu le jour J, avec 2 modalités : Pluie ou Sec.

Importation du jeu de données :

Télécharger le jeu de données « data_ozone.txt » disponible sur Moodle, placez le dans votre répertoire de travail, puis importez sous R.

```
data_ozone <- read.table("data_ozone.txt", sep=";", dec=".", header=TRUE)
```

Régression linéaire simple

On souhaite étudier la relation entre le maximum journalier de concentration en Ozone et la température à 12h.
Est-ce que la concentration max d'ozone varie significativement en fonction de la température à midi ?
Si oui, comment et de combien ?
Peut-on prédire ce pic d'ozone journalier uniquement en fonction de la température à midi ?

1. Visualisez les données.

Créez un diagramme de dispersion pour visualiser le nuage de point correspondant aux 2 variables d'intérêt.

Un modèle de régression linéaire simple semble-t-il adapté à ces données ?

2. Créez le modèle de régression linéaire simple, visant à évaluer la significativité du lien linéaire entre ces 2 variables numériques continues, en utilisant la fonction **lm()** (pour linear model).

La syntaxe est la suivante : lm (**Y ~ X**, data= **jeu_de_données**)

Créez le modèle en remplaçant les arguments en gras/gris par les éléments appropriés, et stockez les résultats de ce modèle dans un objet nommé **reg_s**.

3. Evaluation des hypothèses de validité du modèle de régression linéaire simple

3.1. Linéarité de la relation entre VD et VI

Pour être valide, la régression linéaire simple pré-suppose l'existence d'une relation de forme linéaire (au moins grossièrement) entre les 2 variables quantitatives. La forme de cette relation n'est pas toujours simple à évaluer en utilisant un nuage de point classique.

La fonction **scatterplot()** du package "car" permet de grandement simplifier cette étape. Utilisez la ligne de code suivante en remplaçant les arguments en gras par les éléments appropriés.

```
scatterplot ( VD ~ VI, data= jeu_de_données, col = "#f22b11", regLine = list(col="red"), cex = 0.8,
   smooth = list ( col.smooth="blue", col.spread="darkcyan" ) )
```

Interprétation : La ligne en trait plein est la droite de régression linéaire (définie par la méthode des moindres carrés) entre les deux variables.
La ligne centrale en pointillé est la courbe de régression locale de type lowess (ou loess = méthode de régression non-paramétrique)**,** indiquant la tendance globale de la relation entre les deux variables.
Les deux lignes extérieures en pointillé représentent l'intervalle de confiance (IC) à 95% de la courbe lowess.
L'hypothèse de linéarité est jugée acceptable si la droite de régression est incluse dans l'IC de la courbe lowess.

Que concluez-vous ?

3.2. Indépendance des résidus

En général, l'hypothèse d'indépendance des résidus est validée ou rejetée en fonction du protocole expérimental. Un exemple fréquent de non indépendance se rencontre lorsque la variable prédictive (Xi) est une variable indiquant le temps (ex : jours, mois, années...) ou une position géographique (ex : coordonnées longitudinales, latitudinales, altitude). Dans ce cas, on observe souvent une autocorrélation (corrélation d'une variable avec elle-même) temporelle ou spatiale : 2 points proches dans le temps ou l'espace auront tendance à plus à se ressembler que des points éloignés.

On parle d'auto-corrélation des résidus lorsque, le résidu d'un point quelconque est liée à celui d'un autre point dans le tableau de données. Le décalage entre les lignes du tableau correspondant à ces points est nommé le "lag" (ex: pour un décalage d'1 ligne → lag = 1, décalage de 2 lignes → lag = 2...).

Parmi les méthodes disponibles pour mettre en évidence une autocorrélation entre les observations ou résidus d'un jeu de données, on peut utiliser une représentation graphique, le « lag plot ».

Pour **construire un lag plot**, on utilise la fonction **acf()** que l'on applique aux résidus du modèle de regression. (rappel: utilisez la fonction **residuals()** sur l'objet contenant les résultats du modèle de regression).
Construisez ce graphique et évaluez s'il existe une auto-correlation entre les résidus du modèle.

Interprétation :
- L'axe des abscisses correspond au lag (= décallage en nombre de lignes) entre les observations du jeu de données.
- L'axe des ordonnées correspond au coefficient de corrélation entre les résidus de chaque point et ceux correspondant à un lag donné.
- La hauteur des segments verticaux renseigne sur l'intensité et le sens de la correlation entre le résidu d'un point quelconque et celui correspondant à un lag donné (de 0 à n).
- Les pointillées horizontaux représentent les intervalles de confiance à 95% du test de la significativité du coefficient de corrélation (H0 : le coefficient de corrélation est égal à 0).

NB: il y aura toujours une correlation de r = 1 pour un lag de 0 car cela revient à comparer les observations avec elles mêmes. Ce que l'on ne souhaite pas voir, c'est une correlation significative (= sortant de l'IC à 95%) pour les lags ≥ 1.

Que concluez-vous ?

3.3. Normalité des résidus

A l'aide des représentations graphiques et test habituels, **déterminez si les résidus suivent bien une loi Normale**.

3.4. Homogénéité de la variance des résidus (homoscédasticité)

Pour vérifier la condition d'homoscédasticité des résidus du modèle, on peut à nouveau combiner des méthodes visuelles et des tests statistiques.

3.4.1. Graphique des résidus Studentisés (réduits) en fonction des les valeurs prédites.

Pour cela, utilisez les lignes de code suivantes :

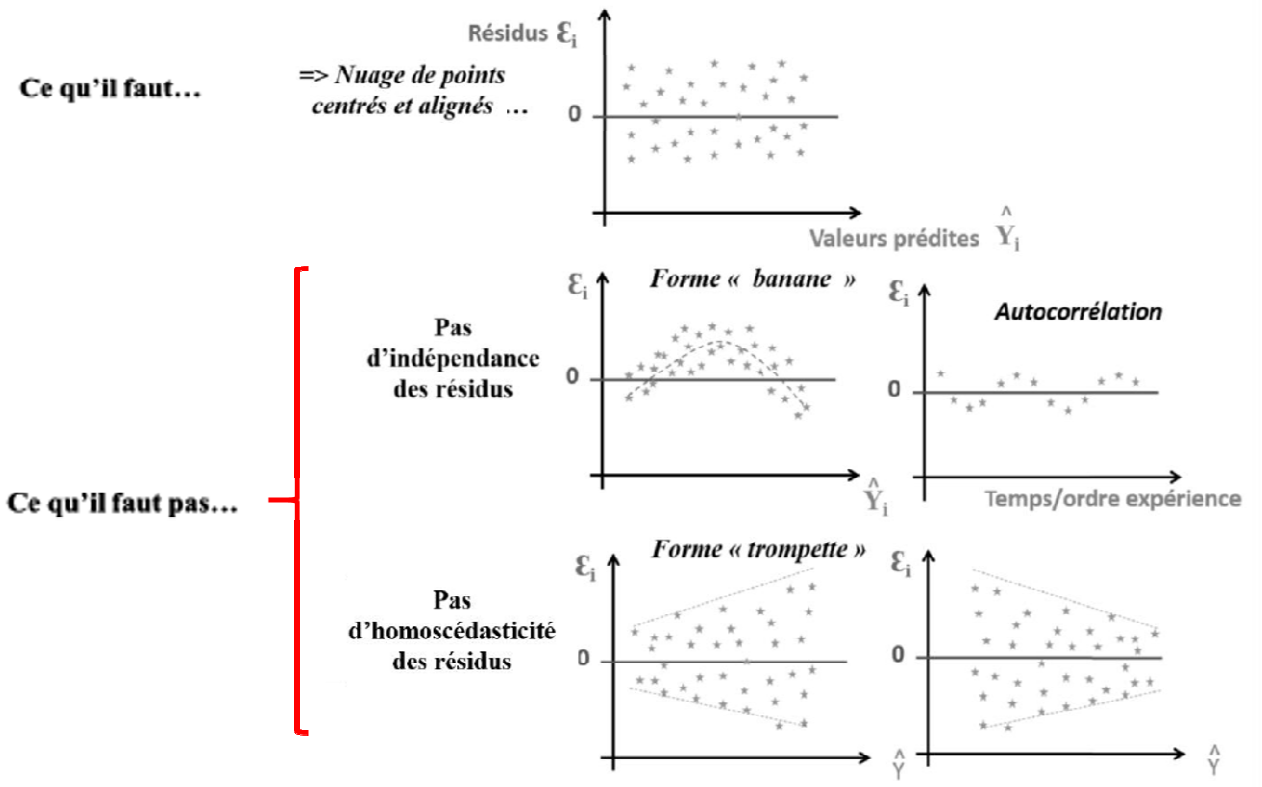
```
res_s <- rstudent ( reg_s )  # procédure de "Studentisation" : chaque résidu est divisé par l'écart type des résidus
plot (res_s, pch=1, cex=0.5, ylab="Résidus studentisés", xlab = " valeurs prédites de y")
abline(h = c ( -2, 0, 2 ), lw = c ( 2, 2, 2 ), col = c ( "purple", "red", "purple" ))
# abline() permet d'ajouter des lignes au graphique, h= c (-2, 0, 2) signifie qu'on souhaite 3 lignes « horizontales »
# et précise les 3 intercepts sur l'axe des Y, lw= permet d'indiquer une épaisseur de trait pour chaque ligne.
```

Interprétation :
Les résidus sont ainsi centrés sur 0, c'est à dire sur la droite de régression (ligne rouge) matérialisant les valeurs prédites de Y ($\widehat{y_i}$) pour les valeurs observées de la variable prédictive (les x_i).
La distance entre un point et la ligne rouge traduit l'écart (résidu = erreur commise) entre entre la valeur observée yi et la prédiction du modèle $\widehat{y_i}$ (rappel : Résidu = valeur observée – valeur prédite). Par conséquent, un point tombant sur la ligne rouge témoigne d'une bonne prédiction de l'observation correspondante, un point ayant une valeur positive sur l'axe de ordonnées correspond à un point pour lequel la valeur prédite de Y < à la valeur observée (et réciproquement).

Pour que le modèle soit valide :

- Les points doivent être répartis de façon homogène autour de la droite horizontale rouge, sur toute sa longueur (condition d'homoscédasticité).
- 95% des points doivent se trouver dans l'intervalle [-2, 2] délimité par les droites violettes (= IC à 95% de Student → condition de normalité).
- Les points sortant de cet intervalle peuvent correspondre à de potentiels outliers dont il faudra vérifier l'influence (voir section 3.5.).
- On ne doit pas non plus observer de tendance flagrante qui remettrait en cause la normalité ou l'homoscédasticité des résidus, ou la condition de linéarité de la relation entre Y et X.

Que concluez-vous quant aux résidus de notre modèle (aidez-vous de la figure page suivante) ?



3.4.2. Test de Breush-Pagan

L'hypothèse H0 correspond à l'homogénéité de la variance des résidus (homoscédasticité).
Pour réaliser ce test, utilisez la fonction **ncvTest()** directement sur l'objet contenant les résultats du modèle.

Que concluez-vous ?

3.5. Vérification supplémentaire : outliers et données influentes

Afin d'identifier les potentiels outliers du jeu de données et mesurer leur influence sur les paramètres de la régression, on peut utiliser 2 des graphiques générés automatiquement par R quand on utilise la fonction **plot()** sur l'objet contenant les résultats du modèle.

Graphique des distances de Cook (**" Cook's distance "**), qui permet d'évaluer l'influence des données (numéro des observations en abscisse) sur les paramètres de régression.

On l'obtient avec la syntaxe suivante : plot (reg_s, 4, las=1)

Interprétation :
La distance de Cook (axe des ordonnées) mesure le changement dans l'estimation des paramètres de régression lorsque la donnée n'est pas prise en compte par le modèle (estimation des paramètres par la méthode des moindres carrés). L'identifiant des potentiels outliers est affichée sur le graphique. Plus la distance de Cook est élevée, plus la modification des paramètres de régression est importante. On considère généralement qu'une observation (en particulier, un outlier) est très influent si la distance de Cook qui lui correspond est ≥ 1.

Que concluez-vous par rapport aux observations du jeu de données ?

4. Interpréter des résultats du modèle de régression simple

Si les hypothèses de linéarité, normalité et d'homogénéité et d'indépendance sont acceptées, les résultats de la régression sont valides, et on peut donc les interpreter.

On consulte le tableau de resultats à l'aide de la fonction : summary (*nom_du_modele*)

Interprétation :

- La partie **"Call"** nous rappelle le modèle sur lequel portent les résultats qui suivent.
- La partie **"Residuals"** permet d'évaluer rapidement la normalité des résidus. Pour cela, la médiane doit être autour de 0, et les valeurs absolues de Q1 et Q3 doivent être proches.
- La partie **"Coefficients"** nous donne des indications sur les paramètres de la droite de régression.

- La 1ère ligne (intercept) concerne l'ordonnée à l'origine ($\mathbf{b_0}$).
- La 2 ème ligne concerne la pente ($\mathbf{b_1}$).

Concernant les **colonnes** :

- La 1ère rapporte l'estimation des coefficients des paramètres.
- La 2 ème rapporte l'estimation de leur erreur standard.
- La 3 ème correspond à la valeur observée de la statistique T.
- La 4 ème rapporte la p-value du test évaluant l'égalité à 0 de chaque coefficient $\mathbf{b_i}$.

- La **dernière partie** renseigne sur l'erreur standard des résidus et le nombre de ddl du modèle (= n -2).
- L'ajustement du modèle de régression linéaire simple est donné par le R^2 (coefficient de détermination = **Multiple R-squared**) → varie [0;1] et mesure la proportion de variation totale de Y expliquée par le modèle.
- Le coefficient de détermination ajusté (**Adjusted R-squared**) mesure la même chose mais permet, en régression multiple, de tenir compte du nombre de variables et de la taille d'échantillon (permet de comparer différents modèles entre eux).
  → Dans tous les cas, le R^2 ne peut être interprété que si le test de Fisher (ligne suivante) est significatif.
- Le **test de Fisher** test l'égalité à 0 de l'ensemble des pentes de toutes les VI du modèle (H_0 : $b_1 = b_2 = \ldots = b_p = 0$, pour "p" VI, contre H_1 : $b_1 \neq b_i \neq \ldots \neq b_p \neq 0$). Il permet de répondre à la question « est-ce-que le fait de rajouter une ou plusieurs VI permet d'augmenter le pouvoir explicatif (R^2) du modèle par rapport au modèle nul (ne contenant que la constante b_0 → aucun effet de la ou des VI) ? ». Dans le cas d'un modèle de régression simple (1 seule VI), le résultat du test est strictement équivalent à celui de test T (il donne la même p-value).

- Peut-on interpréter les coefficients (paramètres) de la régression et comment ?

- Ecrivez la formule de la droite de régression.

- Quel est le pourcentage de variance totale de Y expliquée par ce modèle ?

5. Représentation finale de la régression

Lorsque la pente est significative, on réalise généralement un graphique pour illustrer la relation linéaire liant Y à X, en ajoutant la droite de régression au diagramme de dispersion, et éventuellement son équation. (Attention : si une transformation a été utilisée dans le modèle, par exemple log10 sur Y, cela doit figurer sur le graphique).
On représente généralement aussi l'intervalle de confiance à 95% de la droite de régression (intervalle autour de la droite de régression observée ayant une probabilité de 0.95 de contenir celle (inconnue) de la population dont est issu l'échantillon).

La fonction **geom_smooth()** du package { **ggplot2** } permet de produire facilement ce graphique avec IC95% de la pente car c'est l'option par défaut.

On peut également ajouter l'équation de la droite, après avoir extrait les valeurs des coefficients du modèle de la façon suivante :

```
names ( reg_s ) # liste du nom de toutes les infos récupérables dans l'objet reg_s (résultats du modèle)
## Recuperer les coefficients :
coeff <- reg_s $ coefficients

## Ecrire l'equation de la droite de regression a partir des coefficients :
eq <- paste0("y = ", round ( coeff [ 1 ], 1 ), " + ", round (coeff [ 2 ], 1 ), "*x" )

## tracer la courbe et ajouter l'équation droite et l'intervalle de confiance 95% de la pente
ggplot(data_ozone, aes(x=T12,y=maxO3))+   # Précise le nom du data frame et des variables à utiliser
    geom_point()+                                  # Précise qu'on veut un nuage de points
    geom_smooth(colour="red", method="lm", fill="red") +   # Ajouter la droite de régression et son IC 95%
    ylab("max journalier ozone (µg/m^3)")+                 # Nom axe des Y
    xlab("Température à 12h (°c)") + ggtitle(eq)          # Nom axe des X + afficher l'équation de la droite en titre
```

6. Prédictions

Parfois on souhaite utiliser le modèle de régression à des fins de prévision, c'est-à-dire pour obtenir la valeur de Y prédite par le modèle, pour une ou plusieurs valeurs non observées de la (ou des) VI.

Pour ce faire, on doit créer un data frame contenant la ou les valeurs de la (ou des) VI pour laquelle (lesquelles) on souhaite une prédiction. Attention : le nom de la (ou des) VI doit être strictement identique à celui correspondant au data frame à partir duquel on a créé le modèle !

On utilise la fonction **predict()**, et on utilise l'argument "newdata=" pour spécifier le nom du data frame contenant la (les) valeur(s) de x_i pour laquelle (lesquelles) on souhaite prédire la valeur de Y.

Prévoyons par exemple la concentration en ozone pour des journées ayant une température prévue de 22.5°c et 38°c, et fournissez les intervalles de confiance et de prédiction correspondants.

```
a_predire <- data.frame( T12=c(22.5, 38)) # Création d'un data frame contenant les valeurs de la X pour lesquelles on veut une prédiction de Y d'après le modèle de régression linéaire.

predict(reg_s, newdata =a_predire, interval="confidence") # On spécifie le modèle puis les valeurs de X à utiliser pour la prédiction, et on précise qu'on eut les intervalles de confiances à 95% (par défaut) des valeurs prédites.
```

Que remarquez-vous concernant les valeurs de X pour lesquelles on souhaite une prédiction de Y, et concernant les IC des valeurs prédites ?

Régression linéaire multiple

« Dans la vraie vie », il est extrêmement rare (voire impossible) de se satisfaire d'une seule VI pour expliquer ou prédire efficacement les variations observées sur une VD. Dans le milieu naturel, cette VD a de grandes chances d'être liées à de nombreuses autres variables/facteurs, qui vont agir isolement et/ou en combinaison pour moduler la réponse de la VD. Il est donc important de mesurer simultanément plusieurs VI potentiellement explicatives pour déterminer lesquelles sont significativement liées à la VD.

La régression (linéaire) multiple permet de répondre à différentes questions :
- Quelle est la combinaison de VI les plus pertinentes pour expliquer ou prédire la VD ?
- Quelle proportion de variance de la VD cette combinaison permet-elle d'expliquer ?
- Quelle est la contribution relative de ces VI aux variations de la VD (quantification de l'effet, effet complémentaire ou antagoniste) ? …

La régression linéaire multiple est donc une **généralisation de la régression linéaire simple** dans laquelle la VD (quantitative continue) et une combinaison linéaire (=somme) de ***p* VI quantitatives**.

Ses conditions d'application sont :

- Les mêmes que modèle linéaire simple (indépendance, normalité & homoscédasticité des résidus)
- Disposer d'un nombre d'observations n > nombre de paramètres à estimer (p +1)
  (sinon, il faut réduire le nombre de variables concurrentes!)
- Chaque VI doit être corrélée linéairement avec la VD (au moins grossièrement)
- Les VI ne doivent pas être redondantes ou fortement corrélés entre elles (= multi-colinéarité)

On souhaite ici étudier la relation entre le maximum journalier de la concentration en Ozone et les 3 variables quantitatives du jeu de données "data_ozone" pour trouver la meilleure combinaison de prédicteurs (VI) possible.

1. Visualisez les données.

Créez un diagramme de dispersion pour visualiser le nuage de point correspondant au croisement des 4 variables quantitatives d'intérêt (avec **ggpairs() { GGally }** ou **chart.Correlation() { PerformanceAnalytics }** par exemple).

Un modèle de régression linéaire simple semble-t-il adapté à ces données ?

2. Créez le modèle de régression linéaire multiple, visant à expliquer et prédire les variations de "maxO3" en fonction des variables T12, Ne12 et Vx9.

La syntaxe est la suivante : lm (**Y ~ X1 + X2 + X3**, data= **jeu_de_données**)

Créez le modèle en remplaçant les arguments en gras/gris par les éléments appropriés, et stockez les résultats de ce modèle dans un objet nommé **reg_multi**.

3. Evaluation des hypothèses de validité du modèle de régression linéaire multiple

Vous savez déjà comment vérifier les conditions d'indépendance, normalité & homoscédasticité des résidus, ainsi que le fait que le nombre d'individus statistiques soit > nombre de paramètres du modèle à estimer (= 1 coefficient de pente *bi* par VI + 1 ordonné à l'origine b_0 = 4).

Nous allons donc nous concentrer sur l'évaluation de la dernière condition qui est que les VI ne doivent pas être colinéaire, c'est à dire porter une information redondante (= mesurer quasiment la même chose), ni être fortement correlées entre elles, ce qui biaiserait les résultats du modèle.

Lorsque c'est le cas, il faut réduire le nombre de VI en enlevant celles qui sont redondantes pour ne garder que les plus informatives.

Plusieurs méthodes de detection de (multi)colinéarité existent dont par exemple :

- **1e règle de Klein** : On peut soupçonner une colinéarité entre 2 VI si la valeur absolue de leur coefficient de corrélation ≥ 0.8

→ Voir les coefficients de corrélation de Pearson obtenus avec **ggpairs()** ou **chart.Correlation()**.

- On peut utiliser le « **facteur d'inflation de la variance** » (VIF pour *"variance inflation factor"*) pour mesurer le degré de (multi)colinéarité entre 1 VI et l'ensemble des autres VI. Un VIF proche de 1 indique qu'il n'y a pas de problème de colinéarité. Lorsque cette valeur est ≥ 10, cela révèle un problème de colinéarité.

→ utilisez la fonction **vif()** du package {**car**} directement sur l'objet contenant les résultats du modèle de régression linéaire multiple pour obtenir les valeurs de VIF de chaque VI.

Conclure quant à un possible problème de colinéarité des variables du modèle.

4. Interpréter des résultats du modèle de régression multiple

Si les hypothèses sont vérifiées, on peut consulter le tableau de resultats à l'aide de la fonction : Summary ( )

L'interprétation se fait de la même façon que pour le modèle linéaire simple, si ce n'est que :

- On regardera le R^2 ajusté (**Adjusted R-squared**), qui tient compte du nombre de variables et de la taille d'échantillon, plutôt que le **Multiple R-squared**. Son interprétation est la même.

- On interprétera les coefficients bi significatifs comme indiquant la variation (augmentation ou diminution) de Y quand la VI Xj concernée augmente d'une unité, et que toutes les autres sont maitenues constantes.

- En général, on part du modèle complet (avec toutes les VI) et on retire les VIs non significatives 1 à 1 pour ne conserver que celles qui sont significativement liées à la VD (= qui contribuent à ses variations) dans la version finale du modèle. Ce modèle final pourra ensuite être utilisé pour faire des prédictions des la VD pour de nouvelles valeurs des VIs. (NB: il existe une procédure automatisée sous pour obtenir ce modèle final avec uniquement les Vis significatives; voir les fonctions step() {base} ou stepAIC() {MASS}).

- Quel est le pourcentage de variance totale de Y expliquée par ce modèle ?

- Peut-on interpréter les coefficients (paramètres) de la régression et comment ?

- Quelles sont les variables à conserver dans le modèle ?

- Utilisez le modèle final ne contenant que des VI significatives pour prédire *quelle serait la concentration max en ozone pour une journée avec une température à midi de 17 °C et une projection de vent qui vaudrait 1.3 ?*

BONUS :

A) Vous pouvez utiliser ce jeu de données pour tester s'il existe une différence significative du maximum journalier de la concentration d'ozone en fonction de la direction du vent (Nord, Sud, Est, Ouest).

B) Vous pouvez également comparer le maximum journalier de la concentration d'ozone en fonction de la présence ou non de pluie.

NB : n'oubliez pas de transformer ces variables en facteur pour vous en servir pour constituer vos groupes à comparer (as.factor()).