Faculté des Sciences
Aix Marseille Université

Master 1 MAS – DS
Modèle linéaire
Compte rendu du projet

«Sélection de variables en régression et résolution du problème de la multi-colinéarité»

Réalisé par :

- Saïda GUEZOUI
- Benoit GILLES

Encadré par :

- Fabienne CASTELL
- Pierre PUDLO

2020/2021

Table des matières

I. Introduction

Le but de ce projet est d'étudier et de modéliser par un modèle de régression linéaire un jeu de données nommé Ozone. En statistiques, en économétrie et en apprentissage automatique, un modèle de régression linéaire est un modèle de régression qui cherche à établir une relation linéaire entre une variable, dite à expliquer et une ou plusieurs variables, dites explicatives. Pour arriver à cela, nous allons séparer notre étude en plusieurs étapes.

Dans un premier temps, nous décrirons de manière simple les données contenues dans la table Ozone. Nous effectuerons par la suite une analyse des corrélations et les estimations des coefficients régresseurs à l'aide d'une régression, puis nous verrons comment détecter et traiter un problème de colinéarité. Enfin nous choisirons un sous modèle adapté afin de représenter nos données de manière la plus synthétique et sans perte d'informations et nous effectuerons une analyse de la régression sur ce modèle.

II. Présentation des données

La table SAS utilisée au long de ce projet est nommée Ozone. En effet, en utilisant la procédure CONTENTS qui affiche le dictionnaire, on en déduit que cette table est composée de 112 lignes (observations) qui ont été recueillies à Rennes durant l'été 2001. Elle dispose aussi de 12 colonnes (variables) dont les variables T9, T12 et T15 correspondant aux températures observées à 9h, à 12h et à 15h. Ensuite, les variables (Ne9, Ne12, Ne15) et (Vx9, Vx12, Vx15) représentant respectivement la nébulosité observée et les composante E-O du vent aux heures précédentes. Enfin, MaxO3v est la teneur maximum en ozone observée la veille et la variable obs est un identifiant. On cherche donc à expliquer la variable MaxO3 à l'aide de ces différentes variables en utilisant une régression linéaire multiple.

III. Analyse des données

A. Analyse des corrélations

A l'aide de la procédure CORR, on peut étudier les corrélations entre les différentes variables explicatives. Cela va nous aider à déterminer si les variables régresseurs peuvent expliquer la variabilité de Max03. En effet, une corrélation forte entre deux variables explicatives signifie qu'on peut expliquer la variable d'intérêt à l'aide d'une de ces deux variables uniquement, car elles nous apportent une information similaire. D'où l'importance de cette étape.

Le tableau ci-dessous, correspond à la matrice de corrélation entre les variables : max03, T9, T12, T15, Ne9, Ne12, Ne15, Vx9, Vx12 et Vx15, donnée par le logiciel SAS.

Pearson Correlation Coefficients, N = 112 Prob > |r| under H0: Rho=0											
	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
maxO3	1.00000	0.69939 <.0001	0.78426 <.0001	0.77457 <.0001	-0.62170 <.0001	-0.64075 <.0001	-0.47830 <.0001	0.52762 <.0001	0.43080 <.0001	0.39190 <.0001	0.68452 <.0001
T9	0.69939 <.0001	1.00000	0.88297 <.0001	0.84645 <.0001	-0.48386 <.0001	-0.47225 <.0001	-0.32514 0.0005	0.25069 0.0077	0.22239 0.0184	0.17032 0.0726	0.58225 <.0001
T12	0.78426 <.0001	0.88297 <.0001	1.00000	0.94619 <.0001	-0.58427 <.0001	-0.66010 <.0001	-0.45810 <.0001	0.43010 <.0001	0.31263 0.0008	0.27068 0.0039	0.56363 <.0001
T15	0.77457 <.0001	0.84645 <.0001	0.94619 <.0001	1.00000	-0.58617 <.0001	-0.64923 <.0001	-0.57468 <.0001	0.45309 <.0001	0.34375 0.0002	0.28660 0.0022	0.56789 <.0001
Ne9	-0.62170 <.0001	-0.48386 <.0001	-0.58427 <.0001	-0.58617 <.0001	1.00000	0.78834 <.0001	0.55025 <.0001	-0.49764 <.0001	-0.52878 <.0001	-0.49390 <.0001	-0.27655 0.0032
Ne12	-0.64075 <.0001	-0.47225 <.0001	-0.66010 <.0001	-0.64923 <.0001	0.78834 <.0001	1.00000	0.70987 <.0001	-0.49266 <.0001	-0.51032 <.0001	-0.43227 <.0001	-0.36192 <.0001
Ne15	-0.47830 <.0001	-0.32514 0.0005	-0.45810 <.0001	-0.57468 <.0001	0.55025 <.0001	0.70987 <.0001	1.00000	-0.40147 <.0001	-0.43186 <.0001	-0.37829 <.0001	-0.30848 0.0009
Vx9	0.52762 <.0001	0.25069 0.0077	0.43010 <.0001	0.45309 <.0001	-0.49764 <.0001	-0.49266 <.0001	-0.40147 <.0001	1.00000	0.75018 <.0001	0.68226 <.0001	0.34032 0.0002
Vx12	0.43080 <.0001	0.22239 0.0184	0.31263 0.0008	0.34375 0.0002	-0.52878 <.0001	-0.51032 <.0001	-0.43186 <.0001	0.75018 <.0001	1.00000	0.83717 <.0001	0.22368 0.0178
Vx15	0.39190 <.0001	0.17032 0.0726	0.27068 0.0039	0.28660 0.0022	-0.49390 <.0001	-0.43227 <.0001	-0.37829 <.0001	0.68226 <.0001	0.83717 <.0001	1.00000	0.18992 0.0449
maxO3v	0.68452 <.0001	0.58225 <.0001	0.56363 <.0001	0.56789 <.0001	-0.27655 0.0032	-0.36192 <.0001	-0.30848 0.0009	0.34032 0.0002	0.22368 0.0178	0.18992 0.0449	1.00000

Concernant la variables maxO3, elle est positivement corrélée avec les variables de température et maxO3v, négativement corrélée avec les variables de nébulosité. Il n'y a aucun indice d'une dépendance non linéaire en regardant les nuages des points des corrélations donnée par Sas, ce qui signifie qu'un modèle de régression linéaire parait adapté.

Nous remarquons ainsi que les variables de température T9, T12 et T15 sont fortement corrélées avec un coefficient de corrélation supérieur à 0.80. Ce qui montre un lien entre les 3 températures de la journée.

Ensuite, les variables de nébulosité Ne9, N12 et Ne15 sont corrélées. En effet, le coef de corrélation entre Ne9 et Ne15 est de 0.55 ce qui est petit par rapport à celui de Ne9 et Ne12 qui vaut 0.78.

Enfin, les variables Vx19, Vx12 et Vx15 sont également très corrélées avec un coef supérieur à 0.60.

Pour ces 3 catégories de variables la pvaleur du test est strictement inférieure à notre alpha choisi (5%), ce qui signifie qu'il y a bien un lien entre les variables de chaque catégorie. Cela peut poser un problème de multi-colinéarité que nous allons traiter par la suite en prenant par exemple une seule variable de chaque catégorie en partant du principe que les autres variables restantes donneront des informations redondantes.

Le signe du coefficient de corrélation entre un régresseur et MaxO3 joue un rôle dans la détection du problème de colinéarité dans le cas où ce signe est différent de celui du coefficient régresseurs.

Cette procédure permet également de calculer les statistiques simples comme la moyenne, l'écart type, le minimum et le maximum.

B. Régression linéaire

La procédure REG permet d'obtenir la régression de la variable Max03 en fonction des variables explicatives.

Le coefficient de détermination (R-Square) est défini par $\frac{SSR}{SST}$. Ici R^2 = 0.7638. Le modèle de régression explique 76.38% de la variabilité de Max03.

Root MSE	14.36002	R-Square	0.7638
Dependent Mean	90.30357	Adj R-Sq	0.7405
Coeff Var	15.90194		

Le tableau suivant correspond à l'analyse de variance, permettant de déterminer si le modèle de régression est pertinent.

Nous rappelons qu'il y a 10 variables explicatives. Maintenant, nous voulons tester les deux hypothèses du test (H_0) et (H_1):

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	67364	6736.44502	32.67	<.0001
Error	101	20827	206.21018		
Corrected Total	111	88192			

(H_0) : $\beta_1 = 0,\ldots\ldots, \beta_{10} = 0$ (régression non pertinente)

(H_1) : il existe un i ∈ {1,…,10} tel que $\beta_i \neq 0$ (régression pertinente)

A partir de ce tableau, nous remarquons que la statistique du test utilisée est la statistique de Ficher définie par F = $\frac{SSR/10}{SSE/101} = \frac{MSR}{MSE}$ avec SSR est la somme des carrés de régression de degré de liberté 10, SSE est la somme des carrés des erreurs de ddl égal à 101. Ces sommes sont calculées par le logiciel et données dans la colonne (Sum of Squares).

Sous l'hypothèse (H_0) , F suit une loi de Fisher F(10, 101). Soit t le quantile d'ordre 1- α de F(10,101) tel que α= $IP_{(H_0)}$ (F > t). Rappelons que si F > t, signifie que F est grand donc on rejette (H_0). Si F ≤ t, veut dire que F est petit et qu'on ne peut pas rejeter (H_0).

Nous avons aussi que α = $pval_{obs}$ = 5% = $IP_{(H_0)}$ (rejeter (H_0))= $IP_{(H_0)}$ (Z > F_{obs}) où Z suit une loi de Fisher F(10,101). Ici F_{obs} = 32.67 et $pval_{obs} < 10^{-4} < \alpha$. Nous pouvons donc rejeter (H_0) et conclure que le modèle de régression est pertinent avec une probabilité de se tromper de 5%.

Ce troisième tableau donné par la procédure Reg, représente les valeurs estimées des paramètres β_i ainsi que leur pvaleur et leur intervalle de confiance à l'aide des options de l'instruction MODEL.

Nous rappelons qu'un paramètre nul signifie que la variable explicative correspondante n'explique pas la variable MaxO3 tandis qu'un coef élevé (resp proche de 0) désigne qu'il y a une forte (resp faible) dépendance entre la variable à expliquer et la variable explicative correspondante à ce paramètre.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	50% Confidence Limits	
Intercept	1	12.24442	13.47190	0.91	0.3656	3.12493	21.36391
T9	1	-0.01901	1.12515	-0.02	0.9866	-0.78066	0.74263
T12	1	2.22115	1.43294	1.55	0.1243	1.25115	3.19115
T15	1	0.55853	1.14464	0.49	0.6266	-0.21630	1.33337
Ne9	1	-2.18909	0.93824	-2.33	0.0216	-2.82421	-1.55398
Ne12	1	-0.42102	1.36766	-0.31	0.7588	-1.34682	0.50479
Ne15	1	0.18373	1.00279	0.18	0.8550	-0.49508	0.86255
Vx9	1	0.94791	0.91228	1.04	0.3013	0.33036	1.56545
Vx12	1	0.03120	1.05523	0.03	0.9765	-0.68311	0.74551
Vx15	1	0.41859	0.91568	0.46	0.6486	-0.20125	1.03844
maxO3v	1	0.35198	0.06289	5.60	<.0001	0.30941	0.39455

D'après l'analyse de ce tableau, nous en déduisons que les variables T12 et Vx9 ont respectivement une pvaleur de 12%, 30%. Ensuite, seules les variables Ne9 et maxO3v, ont pvaleur inférieure à 5 %, ce qui signifie que ces deux variables sont nécessaires pour notre modèle. Si nous prenons en compte ces résultats, nous pensons que tous les autres paramètres sont nuls car la pvaleur est largement supérieure à 5%, mais ceci parait non probable et ce résultat est certainement lié à la forte corrélation entre les variables explicatives. En effet, Concernant les coefficients régresseurs de T9 et Ne15 sont de signe opposé de celui du

coefficient de corrélation r(maxO3, T9) et r(MaxO3, Ne15) respectivement. Ce qui confirme l'existence d'un problème de multi colinéarité.

Concernant l'intervalle de confiance, il varie entre des valeurs négatives pour le paramètre β_4, entre des valeurs négatives et positives pour les paramètres β_1, β_3, β_5, β_6, β_8 et β_9 et le reste des paramètres ont des intervalles variant entre des valeurs positives. Mais nous remarquons également que 0 appartient à tous les intervalles des paramètres associés aux variables T9,T15,Ne12, Ne15, Vx12 et Vx15.

IV. Prévision

A. Erreur de prédiction et d'ajustement

En utilisant la procédure REG et les options de l'instruction MODEL, nous pouvons calculer pour chaque observation, la valeur prédite, la prédictions moyenne qui permet de savoir si la valeur prédite est proche ou éloignée de la valeur observée et enfin l'intervalle de confiance associé aux valeurs prédites moyennes (le tableau à droite correspond aux 15 premières observations).

Nous remarquons que certaines observations ont des résidus très élevés comme les observations 15, 11, 7 et 4.

Obs	maxO3	pred_moy	lc_inf_95	lc_sup_95	pred_b_95	pred_h_95	residu	press
1	87	84.226	74.885	93.567	54.247	114.205	2.7742	3.1084
2	82	76.049	69.159	82.938	46.741	105.357	5.9512	6.3210
3	92	88.128	79.225	97.031	58.283	117.973	3.8720	4.2911
4	114	98.903	89.768	108.039	68.988	128.819	15.0966	16.8274
5	94	86.993	79.215	94.771	57.464	116.522	7.0069	7.5713
6	80	76.807	69.116	84.498	47.301	106.313	3.1930	3.4440
7	79	60.097	51.720	68.473	30.404	89.789	18.9033	20.6926
8	79	83.304	77.414	89.193	54.215	112.392	-4.3035	-4.4957
9	101	88.184	81.537	94.831	58.933	117.436	12.8158	13.5538
10	106	96.376	88.111	104.641	66.715	126.037	9.6240	10.5085
11	101	83.824	75.949	91.699	54.269	113.379	17.1761	18.5974
12	90	86.222	76.817	95.628	56.223	116.222	3.7775	4.2398
13	72	77.824	67.369	88.278	47.480	108.168	-5.8238	-6.7303
14	70	64.230	55.864	72.596	34.540	93.920	5.7700	6.3146
15	83	59.912	53.249	66.574	30.656	89.167	23.0885	24.4245

Nous pouvons aussi calculer la somme des carrés des résidus (tableau à gauche). Ici elle vaut 20827.23 tandis que la somme des erreurs des prévisions (tableau à droite) est égale à 26294.38.

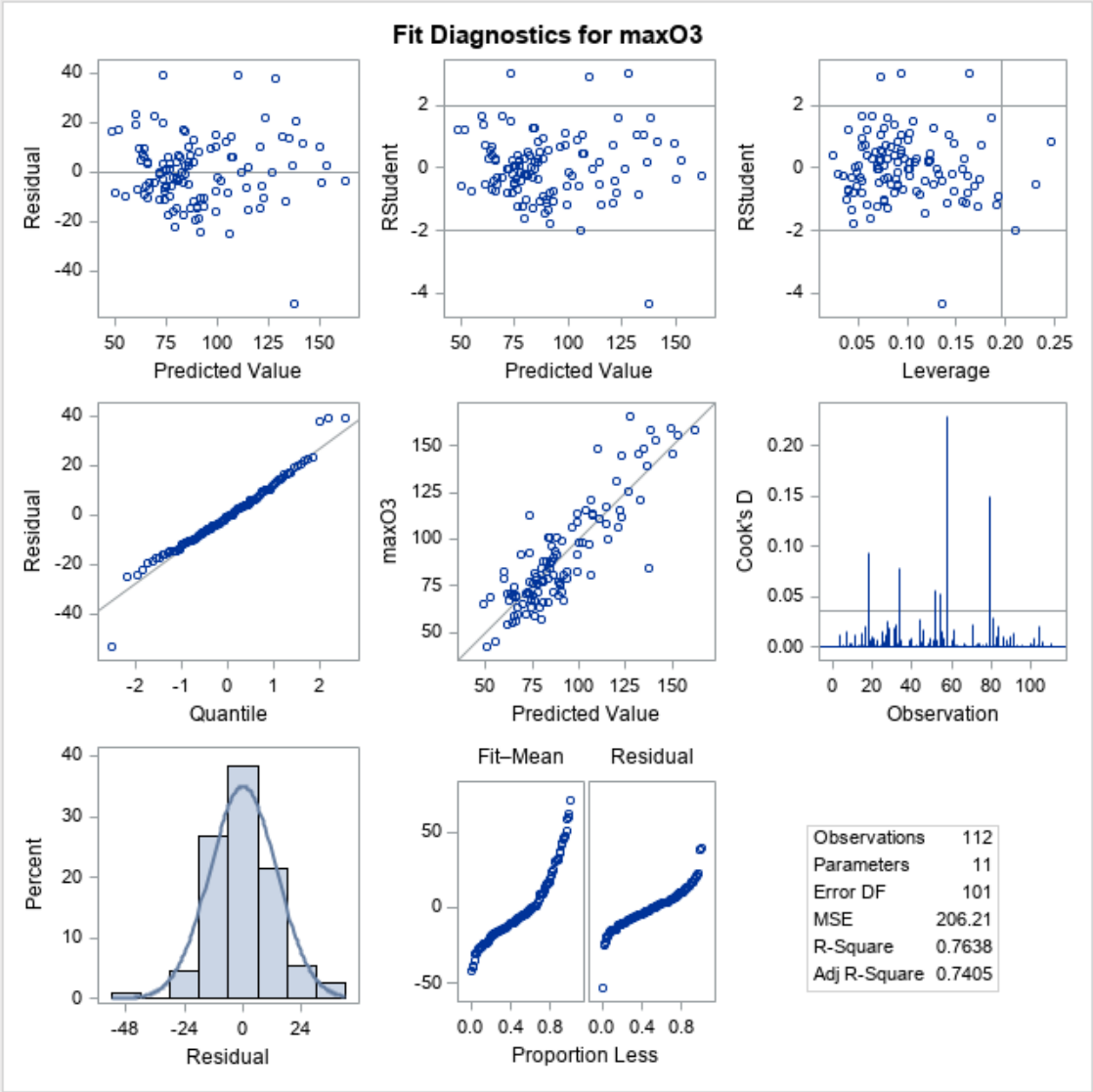
Analysis Variable : residu2		
N	Mean	Sum
112	185.9573962	20827.23

Analysis Variable : press2		
N	Mean	Sum
112	234.7712378	26294.38

La colonne Mean de residu2 (resp press2) correspond à l'erreur d'ajustement (resp l'erreur de prédiction). Nous voyons que ces deux valeurs sont très différentes et éloignées (erreur des prédiction est largement supérieure à l'erreur d'ajustement). Il faut donc analyser les résidus pour expliquer cette différence.

B. Analyse des résidus

Nous rappelons que dans une régression linéaire, les résidus doivent être indépendants de même loi à savoir la loi normale de moyenne nulle et de variance σ^2. Analysons le graphique ci-dessous pour vérifier ces hypothèses afin de valider le modèle de régression.



Nous remarquons que le premier graphe correspondant au nuage de points de la dispersion des résidus en fonction des valeurs prédites n'a pas de forme particulière. Ceci signifie qu'il n'y a pas de dépendance entre les résidus et les variables régresseurs. Concernant le graphique des résidus en fonction des quantiles, les points sont presque parfaitement alignés sur une droite sauf quelques points représentant les observations influentes. Quant à l'histogramme, il montre que les résidus suivent une loi normale. Les hypothèses de départ sont donc bien vérifiées. Ce qui signifie que le modèle de régression est pertinent.

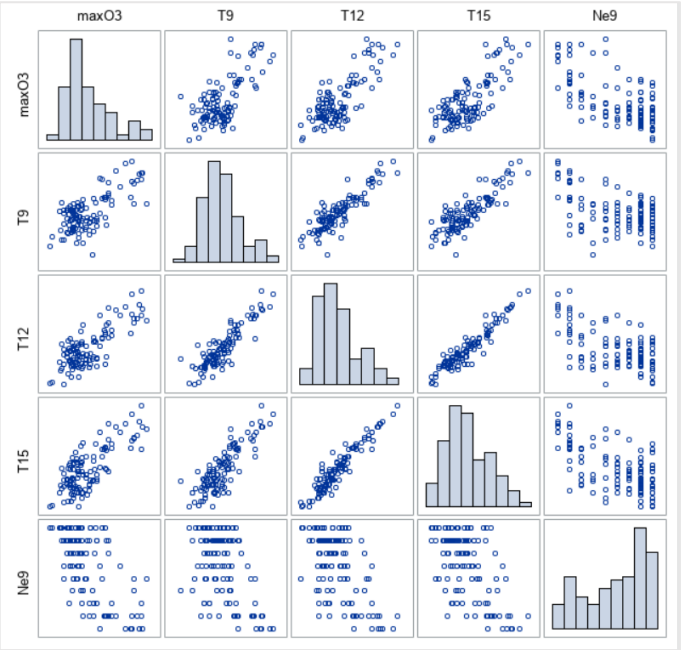
Enfin, le graphique de Cook's D en fonction des observations, correspondant à la distance de Cook de chaque observation. Elle est définie par :

$$DCOOK_i = \frac{\left\|\hat{\beta} - \hat{\beta}_{(i)}\right\|^2_{X'X}}{(p+1)\hat{\sigma}^2}$$

En général, si cette distance est supérieure à 4/n, on dit que l'individu i est influent. Ici nous voyons qu'il y a 6 observations supérieures 4/111≈ 0.036 (barre horizontale), notamment les 4 pics les plus élevés. Ces points peuvent influencer sur l'estimation des β_i et donc sur la régression.

V. Multi-Colinéarité des variables explicatives

A. Explication du diagnostic de la multi-colinéarité



Nous rappelons qu'un problème de multi colinéarité pourrait remettre en cause la significativité des variables et les signes des coefficients régresseurs. Il se pose lorsque les variables régresseurs sont linéairement dépendantes ou proches de l'être. Dans ce cas la matrice (X'X) n'est pas inversible ou proche de ne pas l'être, le calcul des estimations de β_i est instable et les coefficients régresseurs ont une grande variance.

En effet, d'après la matrice de corrélation, nous remarquons que les nuages de points (T9, T12), (T12, T15)et (T9,T15) ressemblent à une droite linéaire y = x, ce qui confirme qu'il y a un lien entre la température à 9h, à 12H et à 15h.

Nous concluons de la même façon pour les variables de nébulosité et de composante E-O du vent.

A.1. Régression avec diagnostic de colinéarité par valeurs propres

Dans cette partie, nous cherchons à vérifier l'existence du problème de la multi colinéarité à l'aide de l'analyse des valeurs propres de la matrice de corrélation. Le tableau ci-dessous, nous donne les 10 valeurs propres de cette matrice. Un indice de multi colinéarité consiste à étudier l'indice de conditionnement de la matrice de corrélation des variables régresseurs ($x^1, \cdots, x^p$) qu'on définit par la formule ci-dessous avec $\lambda_{min} = \lambda_1$ et $\lambda_{max} = \lambda_{10}$ dans notre cas.

$$K = \frac{\lambda_{min}(Cor(x^1, \ldots, x^p))}{\lambda_{max}(Cor(x^1, \ldots, x^p))}$$

Ici, K = 5.49443/ 0.03215 = 170.899844. Cette valeur est largement plus grande que 100, ce qui confirme l'existence du problème de la multi-colinéarité.

Collinearity Diagnostics (intercept adjusted)												
			Proportion of Variation									
Number	Eigenvalue	Condition Index	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
1	5.49443	1.00000	0.00263	0.00132	0.00171	0.00656	0.00442	0.00550	0.00526	0.00325	0.00355	0.00670
2	1.84496	1.72571	0.01369	0.00314	0.00327	0.00125	0.00009365	0.00069078	0.02023	0.02458	0.03477	0.02265
3	0.89758	2.47415	0.00616	0.00033086	0.00004122	0.04273	0.03919	0.09421	0.03365	0.00952	0.01647	0.14079
4	0.57844	3.08199	0.02294	0.00565	0.00122	0.04705	0.00126	0.18156	0.00110	0.00157	0.00570	0.43144
5	0.39394	3.73461	0.00310	0.00412	0.01898	0.28337	0.03657	0.16530	0.01337	0.00285	0.00178	0.30835
6	0.33153	4.07097	0.03514	0.00191	0.00135	0.00049111	0.01027	0.02556	0.55053	0.03092	0.20086	0.01646
7	0.18171	5.49879	0.00742	0.01114	0.00241	0.47870	0.50356	0.14917	0.05760	0.08791	0.00529	0.00094281
8	0.15904	5.87779	0.07883	0.01567	0.00478	0.01501	0.02187	0.01511	0.00087340	0.59796	0.67300	0.00117
9	0.08621	7.98341	0.71354	0.02294	0.23505	0.07941	0.09774	0.07018	0.30263	0.19550	0.03035	0.06159
10	0.03215	13.07307	0.11657	0.93379	0.73120	0.04543	0.28502	0.29273	0.01475	0.04592	0.02821	0.00989

L'indice de conditionnement donné par SAS se situe dans la dernière ligne de la colonne (Condition Index) sous forme d'une racine. En effet, en calculant le carré de 13.07, nous retrouvons la valeur précédemment calculée par la formule ci-dessus.

A.2. Régression avec diagnostic de colinéarité par VIF

Rappelons que l'indice de conditionnement ne permet pas de savoir quelles sont les variables régresseurs à l'origine du problème. C'est pour cela que nous allons étudier la structure de corrélation des variables de prédiction et examiner les facteurs d'inflation de la variance (VIF).

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	12.24442	13.47190	0.91	0.3656	.	0
T9	1	-0.01901	1.12515	-0.02	0.9866	0.15049	6.64512
T12	1	2.22115	1.43294	1.55	0.1243	0.05537	18.06065
T15	1	0.55853	1.14464	0.49	0.6266	0.06907	14.47800
Ne9	1	-2.18909	0.93824	-2.33	0.0216	0.31341	3.19068
Ne12	1	-0.42102	1.36766	-0.31	0.7588	0.19074	5.24264
Ne15	1	0.18373	1.00279	0.18	0.8550	0.33964	2.94433
Vx9	1	0.94791	0.91228	1.04	0.3013	0.32205	3.10515
Vx12	1	0.03120	1.05523	0.03	0.9765	0.21346	4.68465
Vx15	1	0.41859	0.91568	0.46	0.6486	0.28056	3.56427
maxO3v	1	0.35198	0.06289	5.60	<.0001	0.58748	1.70220

VIF (Variance Inflation Factor) signifie Facteur d'Inflation de la Variance. C'est un critère qui permet d'évaluer si les facteurs sont corrélés les uns aux autres (multi-colinéarité). Il est défini par :

$$VIF(j) = \frac{1}{1 - R_j^2}$$

Plus R^2_j est proche de 1 plus VIF(j) est important. En général, dans le cas où VIF est supérieur à 10, on dit qu'il y a une multi-colinéarité élevée. Le but est d'identifier des groupes de variables avec un VIF élevé.

La colonne (Variation Inflation) du tableau ci-dessus, montre que le VIF des variables T12 et T15 sont largement supérieurs à 10 et par rapport aux VIF des autres variables, ce qui signifie que ces variables pourraient être à l'origine du problème de la multi colinéarité.

La tolérance d'une variable se définit comme 1 moins la corrélation multiple au carré de cette variable avec toutes les autres variables indépendantes de l'équation de régression. Par conséquent, plus la tolérance d'une variable donnée est faible, plus cette variable a une contribution redondante pour la régression. Dans notre cas, les variables T12 et T15 ont une faible tolérance de 0.05 et 0.06 respectivement, ce qui confirme l'analyse précédente avec le VIF sachant que VIF = 1/ tolérance.

B. Traitement de la multi-colinéarité

Après avoir déterminé les variables qui sont à l'origine du problème de la multi colinéarité, nous procédons au traitement. En effet, il existe plusieurs méthodes comme la régression RIDGE, régression sur composantes principales, sélection de variables,…

A l'aide de l'option RIDGE de la procédure REG qui consiste à évaluer les VIF de chaque paramètre quand λ croit dans un intervalle proche de 0, on souhaite déterminer une valeur de λ pour laquelle les VIF se stabilisent à une valeur faible.

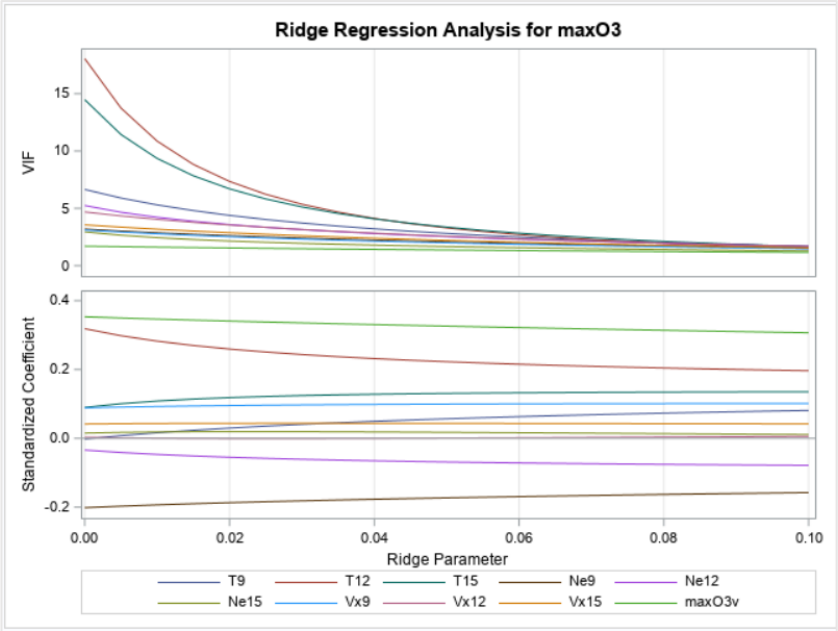
Plus précisément, nous cherchons à estimer β par :

$$\tilde{\beta} = (k\mathrm{Id} + X'X)^{-1}X'Y$$

C'est ce k strictement positif qu'on cherche à déterminer et pour lequel la matrice $k\mathrm{Id} + X'X$ est toujours inversible. Nous obtenons donc un nouvel estimateur biaisé mais de variance faible. Autrement dit un k qui minimise le risque quadratique estimé.

Le premier graphique ci-dessous, correspond à l'évolution de la stabilité des coefficients ridge en fonction de l'évolution de k tandis que le deuxième graphique représente les facteurs d'inflation de la variance en fonction de k. En pratique nous allons déterminer k le plus petit possible pour que ces courbes se stabilisent.

Nous en déduisons donc qu'à partir de $\mathrm{k} = 0.015$, les VIF sont inférieurs à 10 et commencent à se stabiliser à partir de $\mathrm{k} = 0.06$ autour d'une valeur très faible.



Nous pourrons également utiliser la méthode de la régression sur les composantes principales pour traiter ce problème. Elle se déroule en trois étapes : analyse en composantes principales sur les variables explicatives, régression linéaire de la variable indépendante sur la partie des composantes la plus corrélée à celle-ci. Nous choisissons de faire une régression sur les composantes principales car elles ne sont pas décorrélées et il n'y aura plus de problème de multi colinéarité. Enfin, le calcul des paramètres de la régression en fonction des variables d'origines en exprimant les composantes principales comme des combinaisons linéaires des variables initiales :

$$\hat{y} = \sum_{l=1}^{k} \hat{\gamma}_l \, c^l = \sum_{l=1}^{k} \hat{\gamma}_l \left(\sum_{j=1}^{p} C_{lj} \, x^j \right) = \sum_{j=1}^{p} \left(\sum_{l=1}^{k} C_{lj} \, \hat{\gamma}_l \right) x^j$$

Le résultat de cette étape est donné par SAS. Maintenant cherchons à faire une sélection de variables afin de construire un sous modèle ayant un nombre réduit de variables explicatives.

VI. Sélection des variables

A. Méthode du R^2

Model Index	Number in Model	R-Square	C(p)	Variables in Model
1	1	0.6151	56.6274	T12
2	2	0.7012	21.7729	T12 maxO3v
3	3	0.7520	2.0744	T12 Ne9 maxO3v
4	4	0.7622	-0.3065	T12 Ne9 Vx9 maxO3v
5	5	0.7631	1.3341	T12 Ne9 Vx9 Vx15 maxO3v
6	6	0.7636	3.1131	T12 T15 Ne9 Vx9 Vx15 maxO3v
7	7	0.7638	5.0344	T12 T15 Ne9 Ne12 Vx9 Vx15 maxO3v
8	8	0.7638	7.0010	T12 T15 Ne9 Ne12 Ne15 Vx9 Vx15 maxO3v
9	9	0.7638	9.0003	T12 T15 Ne9 Ne12 Ne15 Vx9 Vx12 Vx15 maxO3v
10	10	0.7638	11.0000	T9 T12 T15 Ne9 Ne12 Ne15 Vx9 Vx12 Vx15 maxO3v

La méthode du R^2 est une méthode qui consiste à maximiser le R^2 et à minimiser d'autres critères comme C(p) de Mallows qui correspond à une statistique souvent utilisée comme règle d'arrêt pour diverses formes de régression pas et le critère d'Akaike. D'après le tableau ci-dessus, nous voyons que le modèle contenant les variables T12, Ne9, Vx9 et maxO3v a un R^2 élevé égal à 76% et un C(p) le plus petit. Nous remarquons également qu'à partir de ce modèle le R^2 augmente faiblement.

B. Méthode FORWARD (ascendante)

La méthode FORWARD consiste à faire un test de Fisher pour chaque modèle estimé, afin de vérifier si les coefficients régresseurs sont nuls ou pas.
Généralement, on peut tester le sous modèle W ayant un nombre petit de régresseurs contre le modèle total V. En effet, dans W, le vecteur des moyennes m, appartient à un sous espace vectoriel W de V. On définit donc les deux hypothèses du test :
(H_0) : m ∈ W contre (H_1) : m ∈ V \ W
La statistique du test F est définie par :

$$F = \frac{\frac{\|P_V(Y) - P_W(Y)\|^2}{\dim(V) - \dim(W)}}{\frac{\|Y - P_V(Y)\|^2}{n - \dim(V)}}$$

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.24442	13.47190	0.91	0.3656
T9	1	-0.01901	1.12515	-0.02	0.9866
T12	1	2.22115	1.43294	1.55	0.1243
T15	1	0.55853	1.14464	0.49	0.6266
Ne9	1	-2.18909	0.93824	-2.33	0.0216
Ne12	1	-0.42102	1.36766	-0.31	0.7588
Ne15	1	0.18373	1.00279	0.18	0.8550
Vx9	1	0.94791	0.91228	1.04	0.3013
Vx12	1	0.03120	1.05523	0.03	0.9765
Vx15	1	0.41859	0.91568	0.46	0.6486
maxO3v	1	0.35198	0.06289	5.60	<.0001

Sous (H_0) F suit une F(6,106) dans notre cas. En effet, la méthode FORWARD consiste à ajouter les variables une par une dans le modèle. A l'étape 0, on supprime la variable ayant la pvaleur la plus élevée si elle est supérieure à 5% tout en respectant la condition sur R^2 qui ne doit pas être inférieure à un seuil éventuellement fixé.
Pour les autres étapes, on réitère le même procédé jusqu'à ce qu'aucune variable n'ait une pvaleur supérieure au niveau de significativité.

En regardant le tableau des paramètres estimés, nous en déduisons que la pvaleur de T9 est de 0.98 qui est une valeur très élevée et largement supérieure à 5%, nous pouvons donc supprimer cette variable du modèle à l'étape 0. Nous itérons ce procédé jusqu'à obtenir uniquement des variables dont la pvaleur est inférieure à un niveau qu'on a choisi (5%). Nous obtenons alors un modèle avec les 4 variables : (T12, maxO3v, Ne9 et Vx9).

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	12.63131	11.00088	258.38172	1.32	0.2534
T12	2.76409	0.47450	6650.38921	33.93	<.0001
Ne9	-2.51540	0.67585	2714.81300	13.85	0.0003
Vx9	1.29286	0.60218	903.37445	4.61	0.0341
maxO3v	0.35483	0.05789	7363.50362	37.57	<.0001

Nous voyons que la pvaleur de chaque paramètre correspondant à chaque variable, est inférieure à 5%, ce qui signifie que nous pouvons rejeter (H_0) et conclure que ces coefficient régresseurs sont significativement différents de 0.

Variable Vx9 Entered: R-Square = 0.7622 and C(p) = -0.3065

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	67221	16805	85.75	<.0001
Error	107	20970	195.98354		
Corrected Total	111	88192			

Nous pouvons aussi remarquer que le R^2 vaut 0.7622, ce qui veut dire que ce modèle explique environ 76% de la variabilité de maxO3, ce qui est suffisant. Nous avons donc retrouvé les mêmes résultats qu'avec la méthode du R^2.

C. Analyse du modèle et comparaison avec le modèle complet

Nous rappelons que notre modèle choisi est composé des variables (T12, maxO3v, Ne9 et Vx9). Dans cette partie, nous essayerons de calculer les prévisions afin de comparer ce modèle à notre modèle complet.

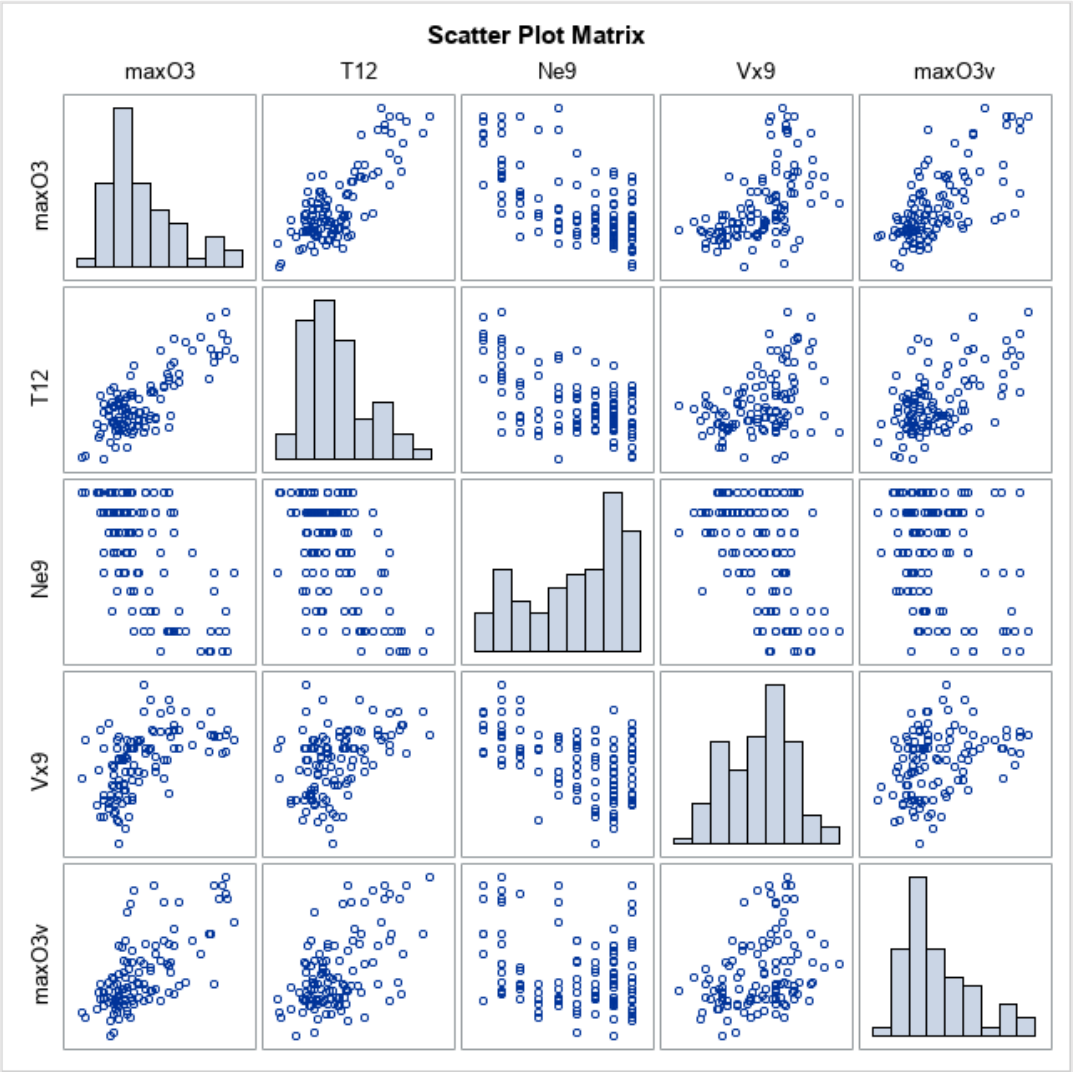
Le tableau suivant montre que la somme des carrés des erreurs de prévisions est égale à 23149.07, cette valeur est proche de 20827.23 qui représente la somme des carrés des résidus. La différence entre ces deux valeurs est de 2321,84, ce qui représente environ 11% d'erreur, tandis que cette différence était de 5467,15 avec le modèle complet, ce qui correspond à environ 26% d'erreurs. Avec ce modèle, nous pouvons réduire plus que la moitié des erreurs de prédiction du modèle. Ce qui nous confirme que ce modèle à 4 variables est plus adapté que le modèle complet.

Analysis Variable : press2		
N	Mean	Sum
112	206.6881588	23149.07

D. Régression sur le sous modèle sélectionné

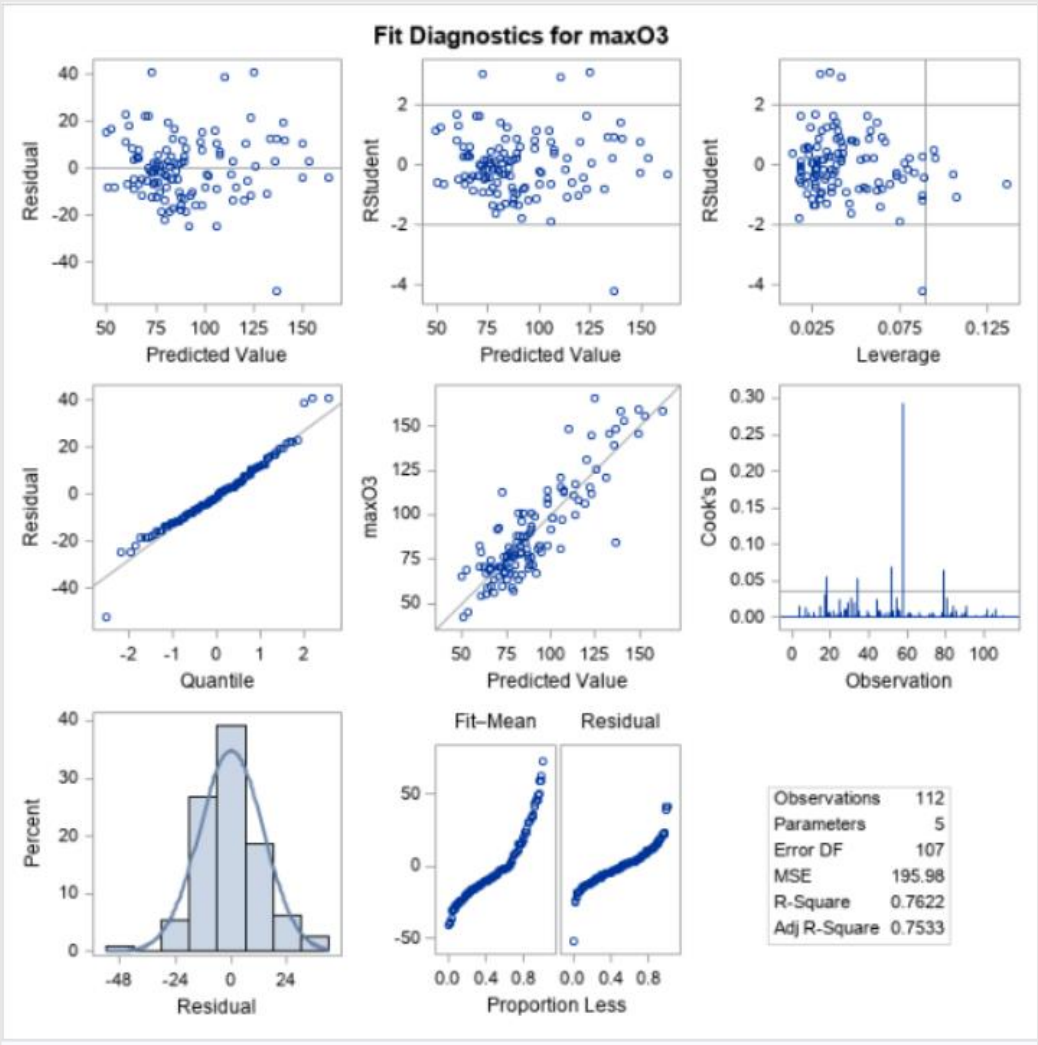
Nous avons calculé la matrice de corrélation ci-dessous en utilisant uniquement les variables explicatives T12, maxO3v, Ne9 et Vx9. En effet, le résultat montre qu'il n'y a plus de corrélation forte entre les variables explicatives sélectionnées (nuage de points ne montre pas de forte dépendance linéaire), ce qui montre qu'il n'y a plus de problème de multi colinéarité.

Nous remarquons aussi qu'il y a une dépendance linéaire entre MaxO3 et les régresseurs, donc ce modèle parait bien adapté pour expliquer la variabilité de MaxO3.



En faisant la régression sur notre sous modèle composé des variables T12, maxO3v, Ne9 et Vx9 (graphiques ci-dessous) nous remarquons que le graphique des résidus en fonction des valeurs prédites ci-dessous, n'a pas de forme particulière. L'histogramme montre une distribution d'une loi normale des résidus. Concernant le graphique correspondant aux résidus en fonction des quantiles, les points sont alignés sur une droite. Les hypothèses sur les résidus sont bien vérifiées dans ce cas.

Au sujet de la distance de Cook's D, on voit l'apparition d'un seul pic qui est vraiment élevé autour de l'observation 58 qu'on suppose comme individu influent. Mais on remarque une amélioration par rapport au modèle complet.



VII. Conclusion

En conclusion nous avons réussi à modéliser notre jeu de données à l'aide d'un modèle de régression linéaire. Dans un premier temps, l'analyse des données nous a permis de conclure qu'un modèle linéaire était bel et bien adapté, mais qu'un problème de multi colinéarité pouvait se poser. Après détection de ce problème grâce au diagnostic de colinéarité par valeurs propres et par VIF, nous avons pu le traiter en utilisant la régression Ridge. De ce fait, nous avons pu sélectionner les variables Vx9, Ne9, T12, maxO3v qui permettent d'expliquer la majorité de l'information contenue dans nos données, et nous avons enlevé les variables qui transmettent des informations redondantes dans notre modèle. Enfin, nous avons effectué une régression en ne gardant que les variables régresseurs ci-dessus, et nous avons conclu que ce modèle était bien adapté. Ainsi la teneur maximum en Ozone à Rennes lors d'une journée peut être exprimée uniquement en fonction de la teneur maximum en Ozone à Rennes la veille, la nébulosité observée à 9h, les composantes du vent à 9h et la température observée à 12h.