المندوبية السامية للتخطيط
HAUT-COMMISSARIAT AU PLAN

INSEA
Institut National de Statistique
et d'Economie Appliquée

PROJET DE REGRESSION EN GRANDE DIMENSION

Sujet :

Analyse parcimonieuse de la vente du BIOBLANC

Préparé par :

DAOUAJI Soukaina

EZ-ZOUINE Amina

SIDIBE Moussa

Sous la direction de :

M. MOUSSANIF Ahmed

Année universitaire :

2021/2022

Abstract

This paper is the result of a four-week assignment conducted as part of an educational project. The objective of this project is to establish sales forecasts for a new liquid detergent marketed under the name BIOBLANC. In this project, we focused on parsimonious regression to solve the problem of multicollinearity in classical linear regression. First, we constructed two latent variables from our dataset based on criteria to maximize both the correlation of these new regressors with the variable of interest and the explained inertia of the original regressors. This type of method, called partial least squares method, combines PCA and regression to overcome the limitations of classical regressions. Second, we used penalized regressions (Ridge, LASSO) to solve the multicollinearity problem, but LASSO goes further and eliminates some variables to build a consistent model. Finally, we compared these three models on test sets based on the root mean square error (RMSE) measure.

Résumé

Ce document est le fruit d'un travail de quatre semaines, réalisé dans le cadre d'un projet pédagogique. L'objectif de ce projet est d'établir des prévisions de ventes d'un nouveau détergent liquide commercialisé sous le nom de BIOBLANC. Dans ce projet, nous nous sommes concentrés sur la régression parcimonieuse pour résoudre le problème de la multicolinéarité dans la régression linéaire classique. Tout d'abord, nous avons construit deux variables latentes à partir de notre ensemble de données sur des critères visant à maximiser à la fois la corrélation de ces nouveaux régresseurs avec la variable d'intérêt et la part d'inertie expliquée des régresseurs originaux. Ce type de méthode, appelée méthode des moindres carrés partiels, combine l'ACP et la régression pour surmonter les limites des régressions classiques. Deuxièmement, nous avons utilisé des régressions pénalisées (Ridge, LASSO) pour résoudre le problème de la multicolinéarité, mais LASSO va plus loin et élimine certaines variables pour construire un modèle cohérent. Enfin, nous avons comparé ces trois modèles sur des ensembles de tests en se basant sur la mesure de l'erreur quadratique moyenne (RMSE).

Sommaire

Introduction

Dans une régression, la multi-colinéarité est un problème qui survient lorsque certaines variables de prévision du modèle mesurent le même phénomène. Une multi-colinéarité peut augmenter la variance des coefficients de régression et les rendre instables et difficiles à interpréter, ce qui peut poser un problème dans l'estimation et l'interprétation d'un modèle, on risque par la suite d'avoir un modèle qui n'est pas robuste.

Dans le cadre de notre projet, on traitera le problème de multi-colinéarité dans un cas d'usage de l'analyse et la prévision du produit d'une entreprise industrielle. On créera dans un premier temps un modèle de régression linéaire de Y par rapport aux régresseurs et on observera qu'il y a un problème de multi-colinéarité dans les données causées par la redondance des informations, ce qui nous mènera à utiliser la régression PLS (Partial Least Square), qui prend en compte le problème de multi-colinéarité. Par la suite, on passera à la création d'un modèle avec la régression pénalisée (Ridge et Lasso) pour enfin choisir le modèle le plus optimal pour résoudre le problème traité, en comparant les 3 modèles crées, tout en se basant sur l'erreur quadratique moyenne (RMSE).

I. Etude préliminaire :

1. Présentation des variables de la table de données :

Variable à expliquer :

Y : La vente du produit BIOBLANC.

Variables explicatives (les régresseurs) :

$\boldsymbol{X_1}$: Le prix (en francs) du flacon BIOBLANC sur la période de vente

$\boldsymbol{X_2}$: La moyenne des prix (en francs) sur la période de vente des produits liquides concurrents

$\boldsymbol{X_3}$: Le budget de la publicité (en 10.000 francs) pour promouvoir BIOBLANC sur la période de vente.

$\boldsymbol{X_4}$: Différence de prix ($X_4 = X_2 - X_1$).

$\boldsymbol{X_5}$: Différence relative de prix en % ($X_5 = (X_4/X_1).100$)

$\boldsymbol{X_6}$: Pub carré ($X_6 = X_3^2$).

$\boldsymbol{X_7}$: Variables croisée ($X_7 = X_4 \,.\, X_3$)

$\boldsymbol{X_8}$: Variables croisée ($X_8 = X_5 \,.\, X_3$)

Aperçu sur la base de données :

Obs.	ID	Y	X1	X2	X3	X4	X5	X6	X7	X8
1	1	7.38	38.5	38	55	-0.5	-1.2987	3025	-27.5	-71.43
2	2	8.51	37.5	40	67.5	2.5	6.6667	4556.25	168.75	450
3	3	9.52	37	43	72.5	6	16.2162	5256.25	435	1175.68
4	4	7.5	37	37	55	0	0	3025	0	0
5	5	9.33	36	38.5	70	2.5	6.9444	4900	175	486.11

Ci-dessus un aperçu sur la base de données utilisée dans notre étude, contenant 30 observations et 10 variables : la variable à expliquer (Y) et les variables explicatives ($\boldsymbol{X_1,\ldots, X_8}$), qui sont tous des variables quantitatives, ainsi que l'identifiant des achats (ID).

2. Analyse descriptive :

On effectue une analyse descriptive sur la variable ventes et les 3 variables principales $\boldsymbol{X_1}, \boldsymbol{X_2}$ et $\boldsymbol{X_3}$:

La procédure MEANS

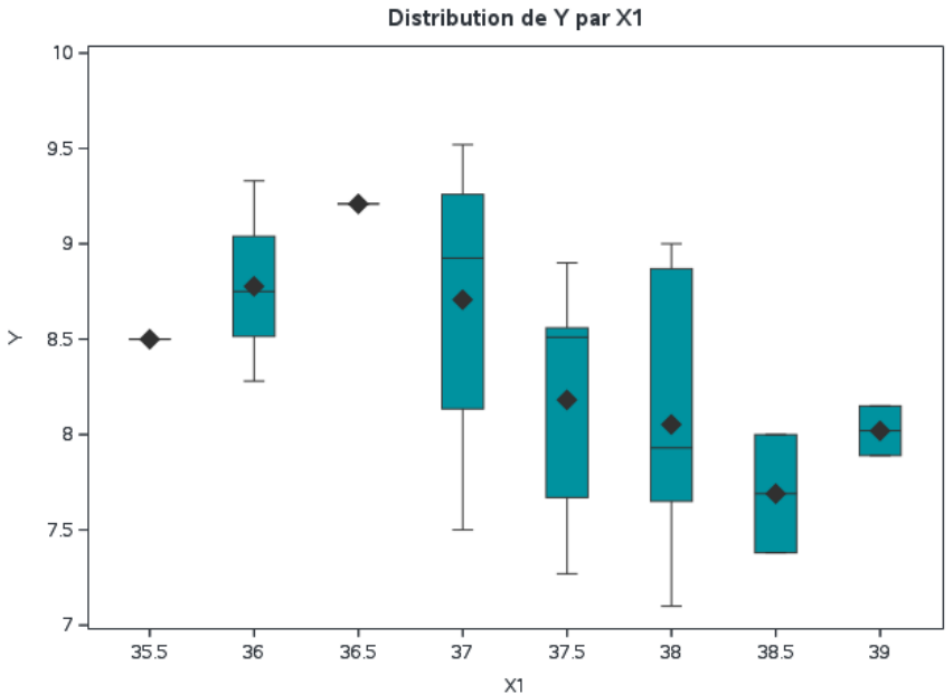
Variable	N	Moyenne	Ec-type	Minimum	Maximum
Y	30	8.3726667	0.6761755	7.1000000	9.5200000
X1	30	37.3500000	0.9016269	35.5000000	39.0000000
X2	30	39.4833333	2.1595152	36.5000000	43.0000000
X3	30	64.5166667	5.7407847	52.5000000	72.5000000

On voit dans ce tableau les différentes statistiques de nos variables, comme le nombre d'observations (n=30), la moyenne, l'écart-type, le min et le max.

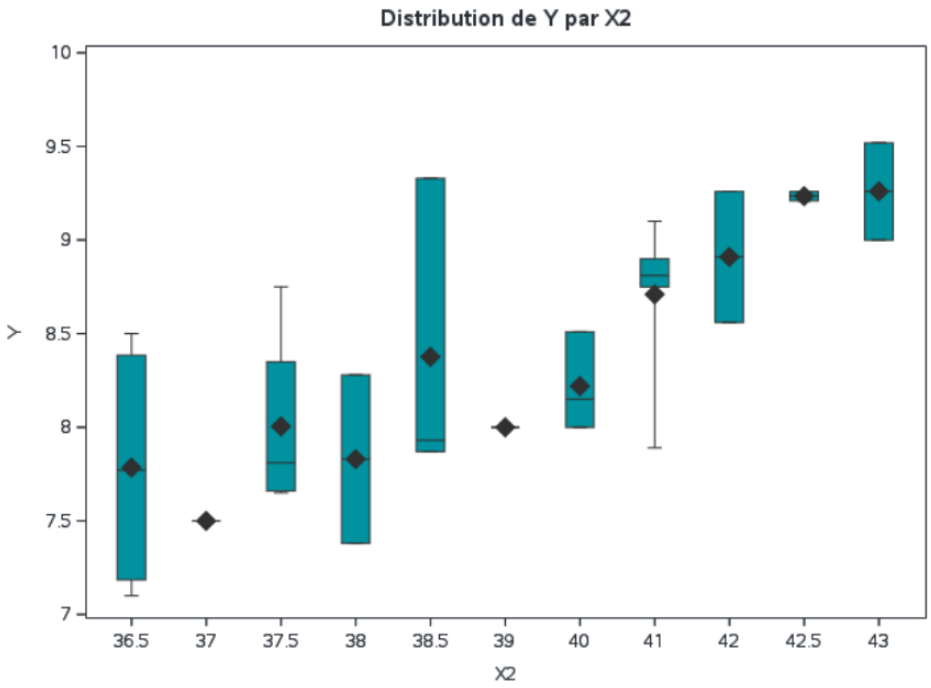
En occurrence, on voit une forte variance pour la variable $\boldsymbol{X_3}$ selon les périodes de vente, et aussi pour $\boldsymbol{X_2}$ qui s'ajuste par rapport au concurrent.

3. Etude de la dépendance au niveau des variables :

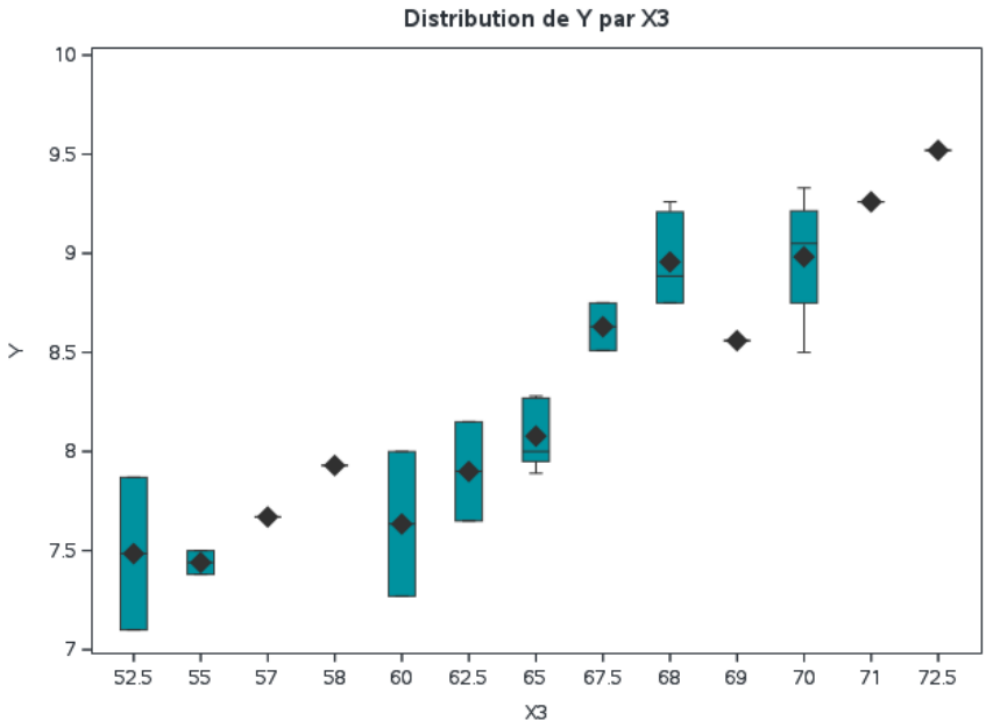
Construction des box-plots entre Y et les $\boldsymbol{X_i}$:



❖ On remarque qu'il y'a une concentration de ventes dans l'intervalle du prix [37, 38] à savoir 66% des demandes ; mais il y'a une diminution des ventes dans les deux extrêmes.



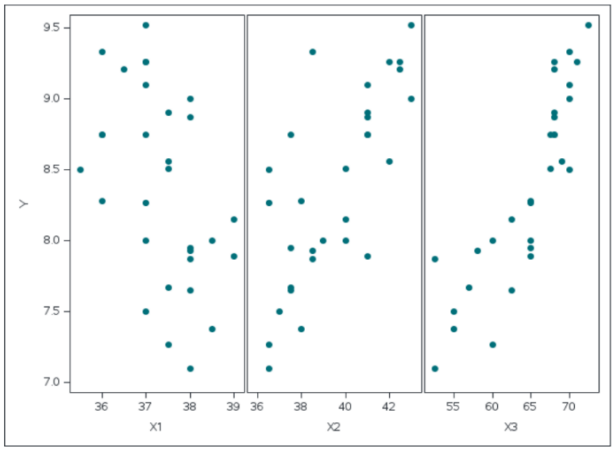
❖ Malgré les prix bas du concurrent, 44% des demandes ont été effectuées pour des prix de concurrent inférieurs à 38,5. Mais, pour les prix qui sont supérieurs à 38,5 les demandes périodiques augmentent.



❖ Nous voyons que plus le budget de publicité augmente nous avons une augmentation de la demande périodique.

4. Analyse des corrélations :

Coefficients de corrélation de Pearson, N = 30 Proba > |r| sous H0: Rho=0				
	Y	**X1**	**X2**	**X3**
Y	1.00000	-0.47528 0.0079	0.72867 <.0001	0.87037 <.0001
X1	-0.47528 0.0079	1.00000	0.07837 0.6806	-0.46751 0.0092
X2	0.72867 <.0001	0.07837 0.6806	1.00000	0.60499 0.0004
X3	0.87037 <.0001	-0.46751 0.0092	0.60499 0.0004	1.00000



➢ Il existe une corrélation négative entre Y et X_1 (-47%). En effet, l'élasticité prix-demande est négative, car plus le prix augmente la demande diminue et vice-versa.

- On remarque qu'il y a une forte corrélation entre Y et X_2 (72%), car lorsque le prix du produit du concurrent (X_2) augmente, les consommateurs vont orienter leur achat vers le produit de l'entreprise BIOCHIMIE, ce qui augmentera les ventes (Y) de cette dernière ce qui est logique du point de vue économique d'un consommateur raisonnable.
- On remarque également une forte corrélation entre Y et X_3 (87%), lorsque l'entreprise investi dans la publicité les ventes augmente. Evidemment lorsque le prix de budget publicitaire augmente plus de consommateurs sont touchés par les informations, et par conséquent, l'entreprise gagne de la notoriété ce qui crée un retour sur investissement dans leur chiffre d'affaires.

II. Régression linéaire de Y par rapport aux variables explicatives :

1. Régression linéaire simple :

1.1. Régression linéaire simple de Y par rapport à chaque variable X_i :

Dans cette partie, on effectuera une régression linéaire simple des ventes par rapport aux 8 variables explicatives en incluant une constante :

- Régression linéaire simple de Y par rapport à X_1 :

Analyse de variance					
Source	**DDL**	**Somme des carrés**	**Moyenne quadratique**	**Valeur F**	**Pr > F**
Modèle	1	2.99514	2.99514	8.17	0.0079
Erreur	28	10.26405	0.36657		
Total sommes corrigées	29	13.25919			

Root MSE	0.60545	**R carré**	0.2259
Moyenne dépendante	8.37267	**R car. ajust.**	0.1982
Coeff Var	7.23130		

Paramètres estimés					
Variable	**DDL**	**Valeur estimée des paramètres**	**Erreur type**	**Valeur du test t**	**Pr > |t|**
Intercept	1	21.68559	4.65873	4.65	<.0001
X1	1	-0.35644	0.12470	-2.86	0.0079

Dans cette régression, nous avons tous les coefficients significatifs avec un modèle significatif au niveau global au sens de Fisher.

Mais, on voit que le modèle explique seulement environ 20% de la variabilité de Y, aussi, nous voyons une forte instabilité pour la constante de la régression, ce qui pourrait impacter la robustesse du modèle.

- Normalité

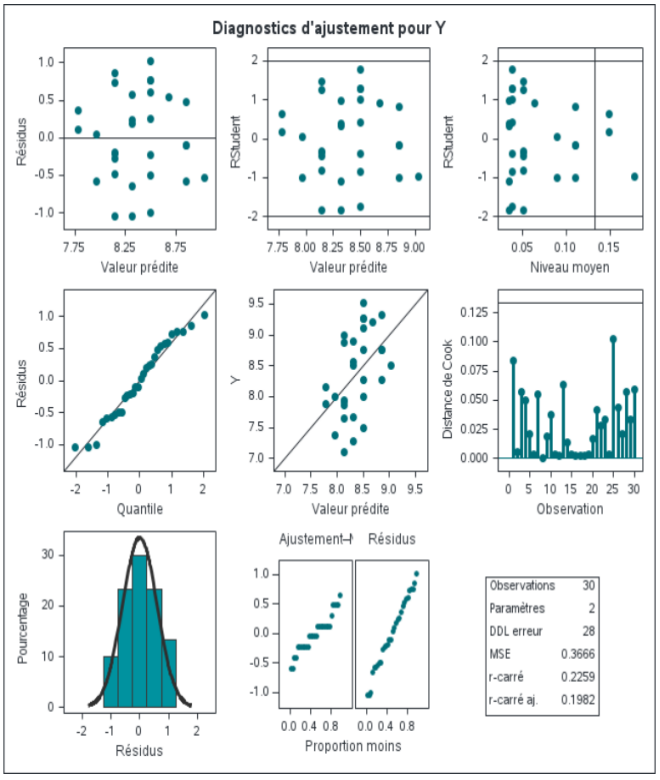
Les résidus sont approximables par une droite, donc, ils suivent une loi normale.

- Observations influentes

En se référant à RStudent et à la Distance de Cook on peut dire qu'il n'y a pas d'observation influente dans ce modèle.

- Effet de levier

Parmi les résidus, il y en a qui sont presque nuls et d'autres très grands, ce qui reflète le problème de leverage et les valeurs atypiques.



➢ Régression linéaire simple de Y par rapport à X_2 :

Analyse de variance					
Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	7.04009	7.04009	31.70	<.0001
Erreur	28	6.21910	0.22211		
Total sommes corrigées	29	13.25919			

Root MSE	0.47129	R carré	0.5310
Moyenne dépendante	8.37267	R car. ajust.	0.5142
Coeff Var	5.62887		

Paramètres estimés					
Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	-0.63573	1.60240	-0.40	0.6946
X2	1	0.22816	0.04053	5.63	<.0001

➢ Dans cette régression, nous avons seulement le coefficient associé au régresseur qui est significatif, avec un modèle significatif au niveau global au sens de Fisher.

➢ On voit que le modèle explique plus de 50% de la variabilité de Y.

- Normalité

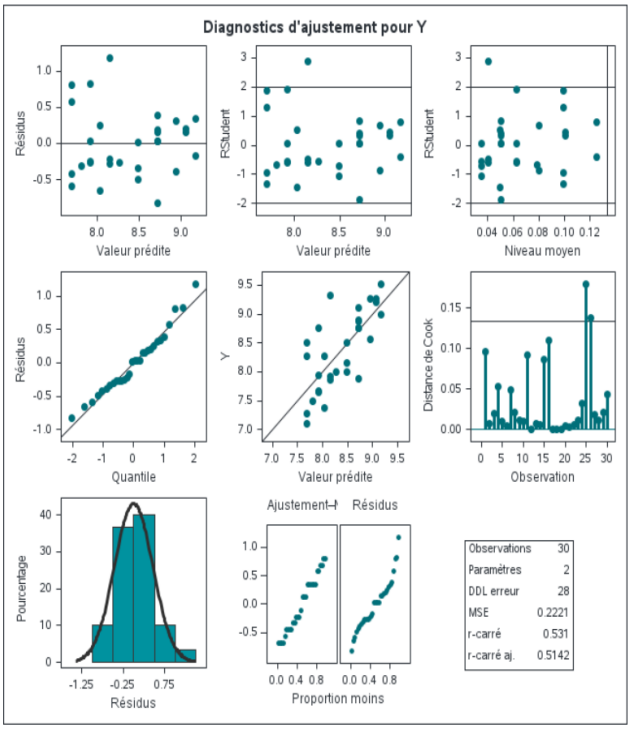
Les résidus sont approximables par une droite, donc, ils suivent une loi normale.

- Observations influentes

En se référant à RStudent et à la Distance de Cook on peut dire qu'il y a une observation influente dans ce modèle.

- Effet de levier

Parmi les résidus, il y en a qui sont presque nuls et d'autres très grands, ce qui reflète le problème de leverage et les valeurs atypiques.



➢ Régression linéaire simple de Y par rapport à X_3 :

Analyse de variance					
Source	**DDL**	**Somme des carrés**	**Moyenne quadratique**	**Valeur F**	**Pr > F**
Modèle	1	10.04437	10.04437	87.48	<.0001
Erreur	28	3.21482	0.11482		
Total sommes corrigées	29	13.25919			

Root MSE	0.33884	**R carré**	0.7575
Moyenne dépendante	8.37267	**R car. ajust.**	0.7489
Coeff Var	4.04702		

Paramètres estimés					
Variable	**DDL**	**Valeur estimée des paramètres**	**Erreur type**	**Valeur du test t**	**Pr > |t|**
Intercept	1	1.75869	0.70983	2.48	0.0195
X3	1	0.10252	0.01096	9.35	<.0001

➢ Dans cette régression, nous avons tous les coefficients significatifs avec un modèle significatif au niveau global au sens de Fisher.

➢ On voit que le modèle explique environ 75% de la variabilité de Y.

- Normalité

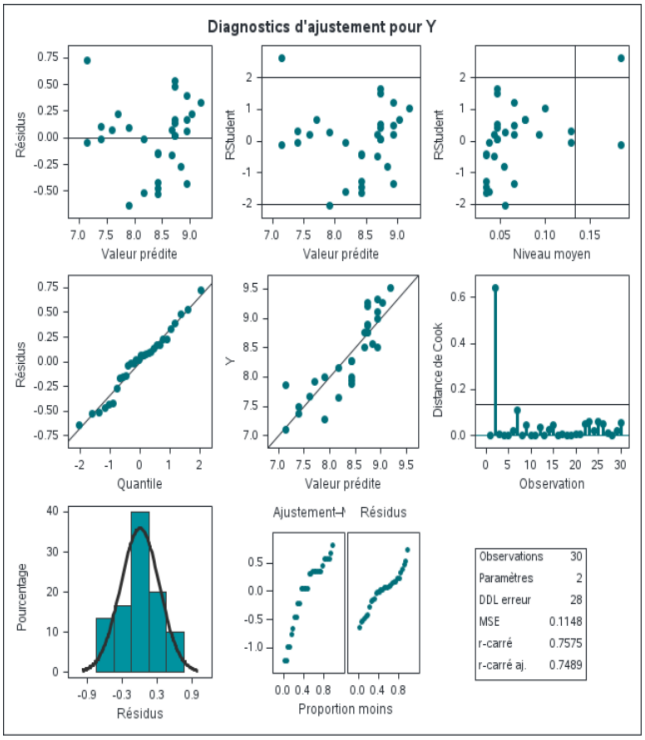
Les résidus sont approximables par une droite, donc, ils suivent une loi normale.

- Observations influentes

En se référant à RStudent et à la Distance de Cook on peut dire qu'il y a une seule observation influente dans ce modèle.

- Effet de levier

Parmi les résidus, il y en a qui sont presque nuls et d'autres très grands, ce qui reflète le problème de leverage et les valeurs atypiques.



➢ Régression linéaire simple de Y par rapport à X_4 :

Analyse de variance					
Source	**DDL**	**Somme des carrés**	**Moyenne quadratique**	**Valeur F**	**Pr > F**
Modèle	1	10.27759	10.27759	96.52	<.0001
Erreur	28	2.98160	0.10649		
Total sommes corrigées	29	13.25919			

Root MSE	0.32632	**R carré**	0.7751
Moyenne dépendante	8.37267	**R car. ajust.**	0.7671
Coeff Var	3.89746		

Paramètres estimés					
Variable	**DDL**	**Valeur estimée des paramètres**	**Erreur type**	**Valeur du test t**	**Pr > |t|**
Intercept	1	7.81419	0.08235	94.89	<.0001
X4	1	0.26179	0.02665	9.82	<.0001

➢ Dans cette régression, nous avons tous les coefficients significatifs avec un modèle significatif au niveau global au sens de Fisher.

➢ On voit que le modèle explique environ 77% de la variabilité de Y.

- Normalité

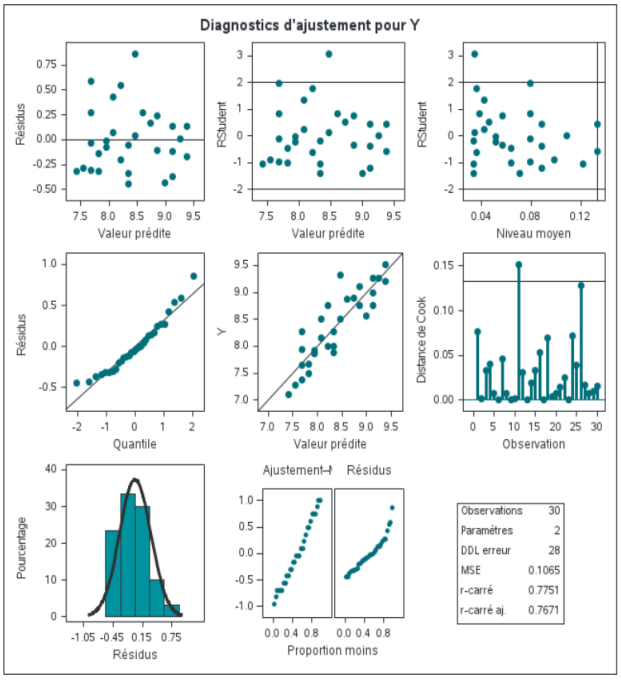
Les résidus sont approximables par une droite, donc, ils suivent une loi normale.

- Observations influentes

En se référant à RStudent et à la Distance de Cook on peut dire qu'il y a une seule observation influente dans ce modèle.

- Effet de levier

Parmi les résidus, il y en a qui sont presque nuls et d'autres très grands, ce qui reflète le problème de leverage et les valeurs atypiques.



➢ Régression linéaire simple de Y par rapport à X_5 :

Analyse de variance					
Source	**DDL**	**Somme des carrés**	**Moyenne quadratique**	**Valeur F**	**Pr > F**
Modèle	1	10.31279	10.31279	98.00	<.0001
Erreur	28	2.94639	0.10523		
Total sommes corrigées	29	13.25919			

Root MSE	0.32439	**R carré**	0.7778
Moyenne dépendante	8.37267	**R car. ajust.**	0.7698
Coeff Var	3.87438		

Paramètres estimés					
Variable	**DDL**	**Valeur estimée des paramètres**	**Erreur type**	**Valeur du test t**	**Pr > |t|**
Intercept	1	7.81334	0.08185	95.46	<.0001
X5	1	0.09710	0.00981	9.90	<.0001

➢ Dans cette régression, nous avons tous les coefficients significatifs avec un modèle significatif au niveau global au sens de Fisher.

➢ On voit que le modèle explique environ 77% de la variabilité de Y.

- Normalité

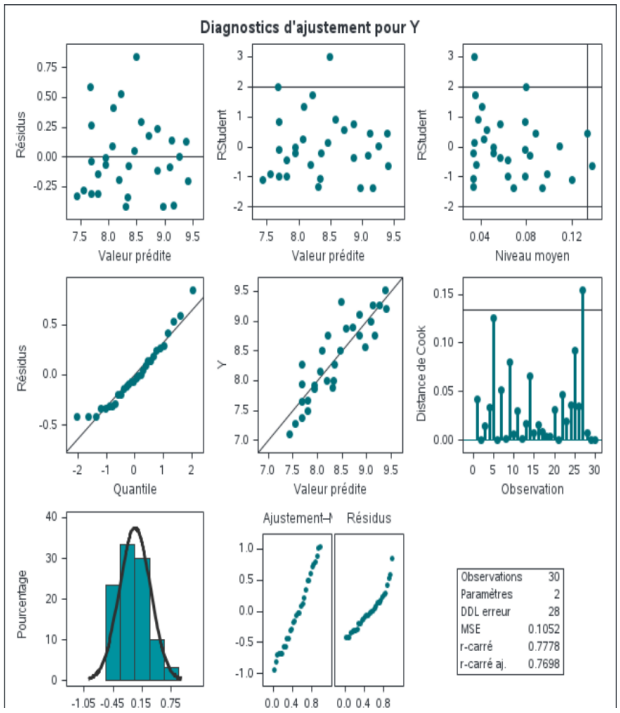
Les résidus sont approximables par une droite, donc, ils suivent une loi normale.

- Observations influentes

En se référant à RStudent et à la Distance de Cook on peut dire qu'il y a une seule observation influente dans ce modèle.

- Effet de levier

Parmi les résidus, il y en a qui sont presque nuls et d'autres très grands, ce qui reflète le problème de leverage et les valeurs atypiques.



➢ Régression linéaire simple de Y par rapport à X_6 :

Analyse de variance					
Source	**DDL**	**Somme des carrés**	**Moyenne quadratique**	**Valeur F**	**Pr > F**
Modèle	1	10.28025	10.28025	96.63	<.0001
Erreur	28	2.97894	0.10639		
Total sommes corrigées	29	13.25919			

Root MSE	0.32618	**R carré**	0.7753
Moyenne dépendante	8.37267	**R car. ajust.**	0.7673
Coeff Var	3.89572		

Paramètres estimés					
Variable	**DDL**	**Valeur estimée des paramètres**	**Erreur type**	**Valeur du test t**	**Pr > |t|**
Intercept	1	4.88389	0.35988	13.57	<.0001
X6	1	0.00083180	0.00008462	9.83	<.0001

➢ Dans cette régression, nous avons tous les coefficients significatifs avec un modèle significatif au niveau global au sens de Fisher.

➢ On voit que le modèle explique environ 77% de la variabilité de Y.

- Normalité

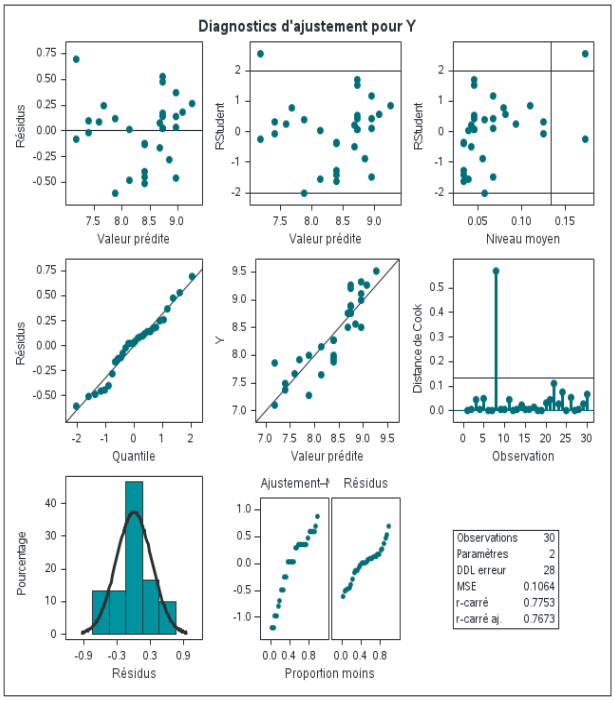
Les résidus sont approximables par une droite, donc, ils suivent une loi normale.

- Observations influentes

En se référant à RStudent et à la Distance de Cook on peut dire qu'il y a une seule observation influente dans ce modèle.

- Effet de levier

Parmi les résidus, il y en a qui sont presque nuls et d'autres très grands, ce qui reflète le problème de leverage et les valeurs atypiques.



➢ Régression linéaire simple de Y par rapport à X_7 :

Analyse de variance					
Source	**DDL**	**Somme des carrés**	**Moyenne quadratique**	**Valeur F**	**Pr > F**
Modèle	1	9.23780	9.23780	64.32	<.0001
Erreur	28	4.02138	0.14362		
Total sommes corrigées	29	13.25919			

Root MSE	0.37897	**R carré**	0.6967
Moyenne dépendante	8.37267	**R car. ajust.**	0.6859
Coeff Var	4.52632		

Paramètres estimés					
Variable	**DDL**	**Valeur estimée des paramètres**	**Erreur type**	**Valeur du test t**	**Pr > |t|**
Intercept	1	7.75218	0.10379	74.69	<.0001
X7	1	0.00391	0.00048768	8.02	<.0001

➢ Dans cette régression, nous avons tous les coefficients significatifs avec un modèle significatif au niveau global au sens de Fisher.

➢ On voit que le modèle explique presque 70% de la variabilité de Y.

- Normalité

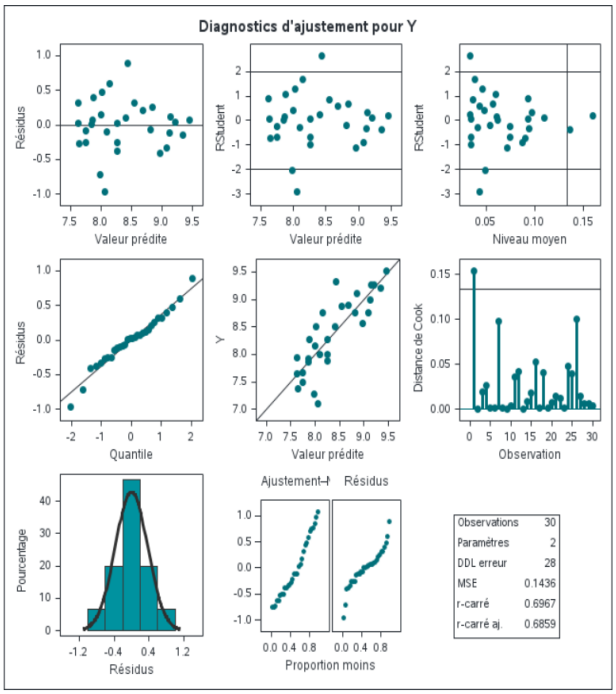
Les résidus sont approximables par une droite, donc, ils suivent une loi normale.

- Observations influentes

En se référant à RStudent et à la Distance de Cook on peut dire qu'il y a une seule observation influente dans ce modèle.

- Effet de levier

Parmi les résidus, il y en a qui sont presque nuls et d'autres très grands, ce qui reflète le problème de leverage et les valeurs atypiques.



➢ Régression linéaire simple de Y par rapport à X_8 :

Analyse de variance					
Source	**DDL**	**Somme des carrés**	**Moyenne quadratique**	**Valeur F**	**Pr > F**
Modèle	1	10.33205	10.33205	98.83	<.0001
Erreur	28	2.92713	0.10454		
Total sommes corrigées	29	13.25919			

Root MSE	0.32333	**R carré**	0.7792
Moyenne dépendante	8.37267	**R car. ajust.**	0.7714
Coeff Var	3.86170		

Paramètres estimés					
Variable	**DDL**	**Valeur estimée des paramètres**	**Erreur type**	**Valeur du test t**	**Pr > |t|**
Intercept	1	7.81074	0.08173	95.57	<.0001
X8	1	0.00141	0.00014218	9.94	<.0001

➢ Dans cette régression, nous avons tous les coefficients significatifs avec un modèle significatif au niveau global au sens de Fisher.

➢ On voit que le modèle explique environ 77% de la variabilité de Y.

- Normalité

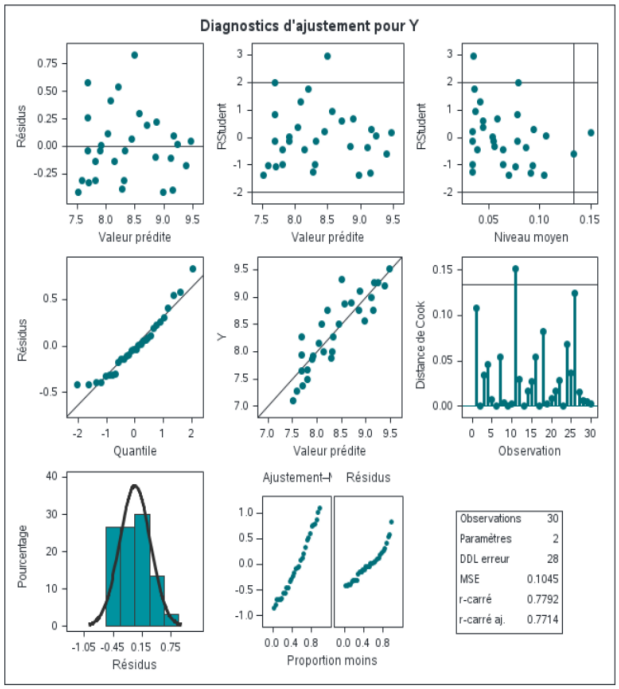
Les résidus sont approximables par une droite, donc, ils suivent une loi normale.

- Observations influentes

En se référant à RStudent et à la Distance de Cook on peut dire qu'il y a une seule observation influente dans ce modèle.

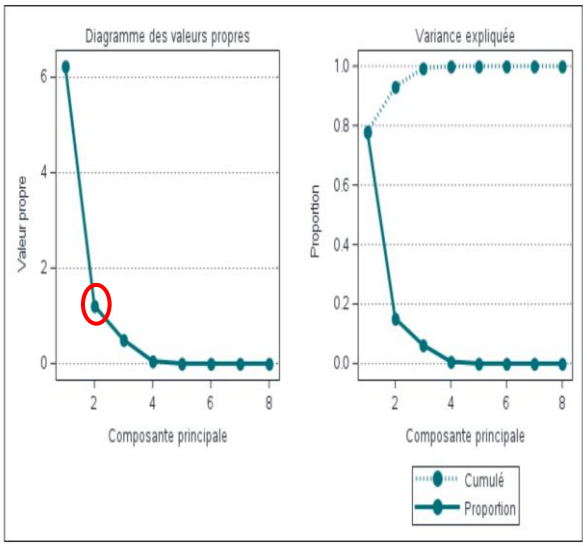
- Effet de levier

Parmi les résidus, il y en a qui sont presque nuls et d'autres très grands, ce qui reflète le problème de leverage et les valeurs atypiques.



1.2. Analyse en composantes principales (ACP) :

Valeurs propres de la matrice de corrélation				
	Valeur propre	Différence	Proportion	Cumulé
1	6.23484531	5.02283223	0.7794	0.7794
2	1.21201308	0.71320902	0.1515	0.9309
3	0.49880406	0.44632924	0.0624	0.9932
4	0.05247483	0.05101754	0.0066	0.9998
5	0.00145729	0.00111576	0.0002	0.9999
6	0.00034152	0.00027761	0.0000	1.0000
7	0.00006391	0.00006391	0.0000	1.0000
8	0.00000000		0.0000	1.0000



Le problème de multicolinéarité rencontré lors de la régression linéaire multiple est confirmé par la nullité de la 8ème valeur propre, de plus, les autres valeurs à part les deux premières apportent une très faible contribution dans l'explication de l'inertie totale.

⇨ Les composantes principales à faible inertie sont considérées comme **des bruits**.

2. Régression linéaire multiple :

Note: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

X4 = 545E-12 * Intercept - 1 * X1 + 1 * X2 - 466E-13 * X3 - 162E-12 * X5 - 399E-15 * X6 + 586E-15 * X8

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	B	28.36590	8.95198	3.17	0.0044
X1	B	0.83363	1.53324	0.54	0.5921
X2	B	-0.81383	1.44506	-0.56	0.5790
X3	B	-0.76119	0.34036	-2.24	0.0358
X4	0	0	.	.	.
X5	B	0.70711	0.69961	1.01	0.3231
X6	B	0.00672	0.00284	2.37	0.0270
X7	1	-0.00049289	0.00180	-0.27	0.7865
X8	B	-0.00523	0.00457	-1.15	0.2643

On observe qu'il y a un problème de multi-colinéarité dans les données causées par la redondance des informations. Notamment, X_4 est une combinaison linéaire des autres variables, ce qui rend la matrice X'X non inversible. Alors, il existe plusieurs valeurs possibles pour les coefficients de régression. C'est pourquoi SAS a marqué les coefficients comme étant biaisés.

Alors, malgré l'estimation de ces derniers, les hypothèses de base d'une régression ne sont pas satisfaites, d'où les statistiques classiques (Test de significativité de Student, de Fisher, ...) ne sont pas applicables. Nous atteignons ainsi la limite d'une régression multiple ce qui nous amène à utiliser la régression parcimonieuse PLS (Partial Least Square).

III. Régression PLS :

1. Introduction à la régression PLS :

L'importance de la régression PLS :

❖ La prise en compte de la multi-colinéarité.

En utilisant la régression PLS, nous créons d'autres variables dites latentes étant la combinaison linéaire des variables originales, ces variables sont construites en maximisant sa corrélation avec la variable d'intérêt et simultanément maximiser la variance expliquée des régresseurs originaux ainsi cette méthode combine à la fois ACP et la régression pour explorer des horizons inaccessibles à la régression ordinaire.

❖ La gestion de tableaux écrasés.

Lorsque le nombre d'observations est faible et le nombre de variables est élevé, la régression ordinaire ne peut plus être appliquée. La régression PLS permet de traiter ce type de données du fait de la réduction de dimension induite par l'utilisation de composantes.

❖ La possibilité d'avoir plus d'une variable dépendante.

Avec l'introduction de l'algorithme PLS2, on peut expliquer un groupe de variables Y par un groupe de variables X.

❖ Le traitement des données manquantes.

La régression PLS offre une méthodologie originale pour le traitement des données manquantes. En effet, la régression PLS est basée sur un algorithme itératif pouvant travailler sur des données incomplètes. On utilisera donc les données disponibles pour construire les composantes et il n'y aura pas besoin de compléter les données manquantes.

2. Cas pratique :

2.1. Construction de T_1 :

	Y
Y	1.00000
X1	-0.43448 0.0339
X2	0.69050 0.0002
X3	0.87446 <.0001
X4	0.88516 <.0001
X5	0.89022 <.0001
X6	0.88730 <.0001
X7	0.81358 <.0001
X8	0.88908 <.0001

A partir de la matrice de corrélation de Y par rapport aux X_i, il parait que toutes les variables explicatives sont significativement corrélées à Y, donc elles vont toutes entrées dans la construction de T_1.

Formule de T_1:

$$T_1 = \frac{\sum_{j=1}^{p} cov(Y,X_j)X_j}{\sqrt{\sum_{j=1}^{p} cov(Y,X_j)^2}}$$

$$T_1 = \frac{-0.43*X_1+0.69*X_2+0.87*X_3+0.88*X_4+0.89*X_5+0.88*X_6+0.81*X_7+0.88*X_8}{\sqrt{(0.43\text{^}2+0.69\text{^}2+0.87\text{^}2+0.88\text{^}2+0.89\text{^}2+0.88\text{^}2+0.81\text{^}2+0.88\text{^}2)}}$$

2.2. Construction de T_2 :

Afin de sélectionner les variables explicatives pouvant intervenir dans la construction de $\boldsymbol{T_2}$ nous régressons Y sur les $\boldsymbol{X_i}$ sachant $\boldsymbol{T_1}$ et on garde les variables significatives :

$\boldsymbol{Y = \beta_1 T_1 + \alpha_i X_i}$ avec i = 1,…,8

Finalement, on a trouvé que la variable X_7 est la seule variable significative, et par la suite, elle va entrer dans la construction de T_2.

Ensuite, on récupère le résidu X_{17} de la régression de X_7 sur T_1. Cela étant, on normalise :

$$X_{17\_cn} = \frac{X_{17}}{var(X_{17})}$$

Nous vérifions par la suite la significativité du résidu dans la régression de Y sur X_{17_cn} sachant T_1 :

$$\boldsymbol{Y = \beta_{17} T_1 + \alpha_{17} X_{17\_cn}}$$

⇨ On trouve que X_{17_cn} est significative, alors, on passe à la construction de T_2.

Formule de T_2 :

$$\boldsymbol{T_2} = \frac{\sum_{j=1}^{p} \boldsymbol{cov(Y,X_{1j})X_{1j}}}{\sqrt{\sum_{j=1}^{p} \boldsymbol{cov(Y,X_{1j})^2}}} \Rightarrow T_2 = \frac{\boldsymbol{\alpha_{17}X_{17}}}{\boldsymbol{|\alpha_{17}|}} = \frac{\boldsymbol{-0{,}05531 \times X_{17}}}{\boldsymbol{0{,}05531}}$$

2.3. Construction de T_3 :

Afin de sélectionner les variables explicatives pouvant intervenir dans la construction de $\boldsymbol{T_3}$ nous régressons Y sur les $\boldsymbol{X_i}$ sachant $\boldsymbol{T_1}$ et $\boldsymbol{T_2}$ et on garde les variables significatives :

$$\boldsymbol{Y = \beta_1 T_1 + \beta_2 T_2 + \alpha_i X_i} \text{ avec i = 1,…,8}$$

Finalement, nous constatons qu'aucune variable n'entre dans la construction de T_3.

⇨ Donc, il faut retenir que les deux composantes PLS $\boldsymbol{T_1}$ et $\boldsymbol{T_2}$.

2.4. Etape finale :

Nous régressons Y sur $\boldsymbol{T_1}$ et $\boldsymbol{T_2}$ pour construire le modèle avec les variables d'origines :

$$\boldsymbol{Y} = 8.33708 + 0.25659 * \boldsymbol{T_1} + 0.32888 * \boldsymbol{T_2}$$

Avec :

$$\boldsymbol{T_1} = -5.697 - 0.210 * X_1 + 0.139 * X_2 + 0.066 * X_3 + 0.169 * X_4 + 0.063 * X_5 + 0.0005 * X_6 + 0.002 * X_7 + 0.0009 * X_8$$

$$\boldsymbol{T_2} = -1.062 - 0.079 * X_1 + 0.052 * X_2 + 0.025 * X_3 + 0.064 * X_4 + 0.024 * X_5 + 0.0002 * X_6 - 0.005 * X_7 + 0.0003 * X_8$$

Pour la lisibilité, nous avons choisi de retenir seulement les coefficients significatifs à 10^{-3} près.

$$\boldsymbol{Y} = 6.525 - 0.080 * X_1 + 0.053 * X_2 + 0.025 * X_3 + 0.064 * X_4 + 0.024 * X_5 + 0.0002 * X_6 - 0.001 * X_7 + 0.0003 * X_8$$

Interprétation du point de vue marketing :

- En moyenne, on reçoit une demande de 6.735 par période de vente en absence de contrainte liée au prix de la vente, du prix de concurrent et de la dépense publicitaire.
- La variation de la demande est principalement due au prix de ventes (34%), à la différence de prix de ventes et celui du concurrent (25%), au prix du concurrent (22%), la publicité (10%), enfin la différence relative de prix (9%).

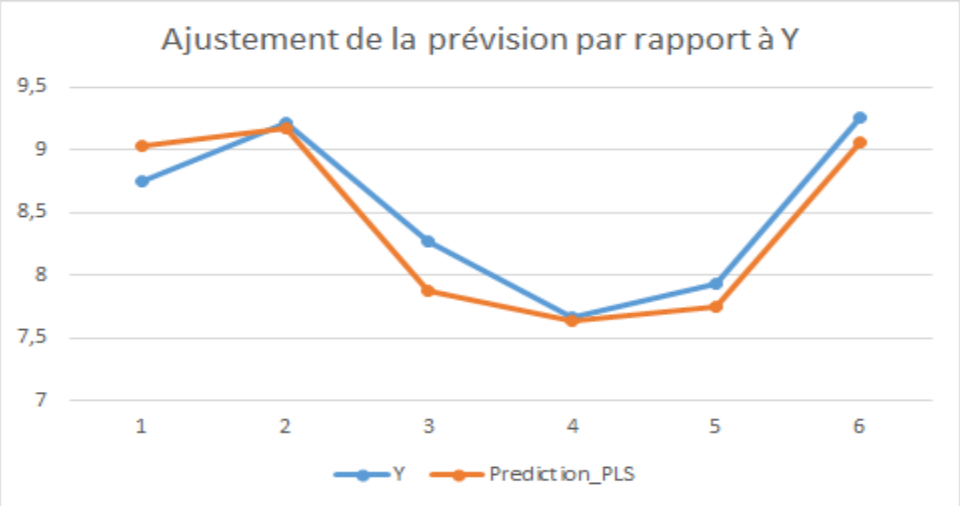
Après la construction du modèle on passe à la partie de la prédiction.

IV. Prédiction :

Dans cette partie, en utilisant le modèle construit précédemment, on va générer des prévisions de vente à l'aide du tableur Excel, on obtient les résultats affichés dans le tableau ci-dessous :

Y	X1	X2	X3	X4	X5	X6	X7	X8	Prédiction_PLS	Erreur_PLS
8,75	36	41	68	5	13,8889	4624	340	944,44	9,033791972	0,28379197
9,21	36,5	42,5	68	6	16,4384	4624	408	1117,81	9,169506561	0,04049344
8,27	37	36,5	65	-0,5	-1,3514	4225	32,5	-87,84	7,881550266	0,38844973
7,67	37,5	37,5	57	0	0	3249	0	0	7,629707679	0,04029232
7,93	38	38,5	58	0,5	1,3158	3364	29	76,32	7,743972786	0,18602721
9,26	37	42,5	68	5,5	14,8649	4624	374	1010,81	9,067070216	0,19292978
									RMSE	0,55364456

Graphiquement on obtient :



V. Régression pénalisée :

1. Introduction :

Nous considérons toujours le modèle linéaire :

$$\boldsymbol{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d} + \varepsilon$$

Lorsque d est grand ou que les variables sont linéairement dépendantes, les estimateurs des moindres carrées peuvent être mis en défaut. Les méthodes pénalisées ou sous contraintes consistent alors à restreindre l'espace sur lequel on minimise ce critère. On va alors chercher le vecteur β qui minimise

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{d} X_{ij}\beta_j\right)^2$$

sous la contrainte $\sum_{j=1}^{d} \beta_j^{\,2} \leq t$

ou de façon équivalente (dans le sens où il existe une équivalence entre t et λ)

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{d} X_{ij}\beta_j\right)^2 + \lambda \sum_{j=1}^{d} \beta_j^{\,2}$$

Les estimateurs obtenus sont les estimateurs **Ridge**. Les estimateurs **Lasso** s'obtiennent en remplaçant la contrainte ou la pénalité par une norme 1 ($\sum_{j=1}^{d}|\beta_j|$) .

2. La régression Ridge :

L'estimateur Ridge s'écrit :

$$\widehat{\beta_{ridge}} = (X'X + \lambda I_p)^{-1}X'y$$

• On peut avoir une estimation même si (X'X) n'est pas inversible

• On voit bien que λ = 0, alors on a l'estimateur des MCO.

Détermination de la valeur de λ :

• Fixer une plage de valeurs de λ

• Construire le modèle sur un échantillon d'apprentissage

• L'évaluer sur un échantillon test

• Choisir λ qui minimise un critère d'erreur en test

Application de Ridge :

Pour le choix de λ, nous choisissons une plage de 99 valeurs entre 0,1 et 10, et on retient celle permettant de minimiser l'erreur GCV (Generalised Cross validation). Pour ce cas nous avons trouvé λ = 4.5.

Y = 9.273352-0.1137110*X1+2.575409*10^(-2)*X2+1.936221*10^(-2)*X3+4.852966*10^(-2)*X4+1.906539*10^(-2)*X5+1.774839*10^(-4)*X6+6.428905*10^(-5)*X7+2.684253*10^(-4)*X8

Prédiction :

Y	X1	X2	X3	X4	X5	X6	X7	X8	Prediction ridge	Erreur_ridge
8,75	36	41	68	5	13,8889	4624	340	944,44	9,155804986	0,40580499
9,21	36,5	42,5	68	6	16,4384	4624	408	1117,81	9,285626043	0,07562604
8,27	37	36,5	65	-0,5	-1,3514	4225	32,5	-87,84	7,94296353	0,32703647
7,67	37,5	37,5	57	0	0	3249	0	0	7,655259036	0,01474096
7,93	38	38,5	58	0,5	1,3158	3364	29	76,32	7,735632156	0,19436784
9,26	37	42,5	68	5,5	14,8649	4624	374	1010,81	9,143598987	0,11640101
									RMSE	0,57349384

3. La régression Lasso :

Application de lasso :

En prenant de même le hyper paramètre λ minimisant l'erreur de Cross Validation, nous retenons que la variable X2, X3, X5, X6 et X7. Mais, seules les variables X5 et X6 sont significatives dans la régression de Y sur les régresseurs retenus. Le modèle est particulièrement bien ajusté vu qu'il explique 87% de la variation de Y autour de sa moyenne et le test global selon Fisher est significatif.

Y = 6.09+5.514*10^(-2) *X5+4.674*10^(-4)*X6

Prédiction :

Y	X1	X2	X3	X4	X5	X6	X7	X8	Prediction_LASSO	Erreur LASSO
8,75	**36**	41	68	5	13,8889	4624	340	944,44	**9,017091546**	0,267091546
9,21	**36,5**	42,5	68	6	16,4384	4624	408	1117,81	**9,157670976**	0,052329024
8,27	**37**	36,5	65	-0,5	-1,3514	4225	32,5	-87,84	**7,990248804**	0,279751196
7,67	**37,5**	37,5	57	0	0	3249	0	0	**7,6085826**	0,0614174
7,93	**38**	38,5	58	0,5	1,3158	3364	29	76,32	**7,734886812**	0,195113188
9,26	**37**	42,5	68	5,5	14,8649	4624	374	1010,81	**9,070908186**	0,189091814
									RMSE	0,479514254

Conclusion

L’objectif de ce projet, et dont le présent document décrit le déroulement, était d’établir des prévisions de ventes d’un nouveau détergent liquide commercialisé sous le nom de BIOBLANC.

Ce projet a été axé sur 3 missions principales. En effet, La première mission consistait à la construction des variables latentes à partir de notre ensemble de données en utilisant la régression PLS, la deuxième mission consistait à résoudre le problème de la multi-colinéarité en utilisant des régressions pénalisées (Ridge, LASSO) et la troisième mission consistait à la comparaison de ces trois modèles sur des ensembles de tests en se basant sur la mesure de l'erreur quadratique moyenne (RMSE).

Cette expérience était très enrichissante et formatrice. Nous avons pu accumuler beaucoup d’informations précieuses et donc pu élargir nos connaissances sur tout ce qui concerne la régression parcimonieuse et d’acquérir une sagesse professionnelle capable d’être au futur un outil indispensable pour un data scientist.

Annexes

Annexe 1 :

Afficher la statistique descriptive sur le dataset bioblanc :

```
PROC MEANS DATA=BIOBLANC;
VAR Y X1 X2 X3 X4 X5 X6 X7 X8;
RUN;
```

Afficher les boxplots sur les variables d'origines Y X1 X2 X3 :

```
PROC SORT DATA=BIOBLANC;
BY X1;
PROC BOXPLOT DATA=BIOBLANC;
PLOT X2*X1;

PROC BOXPLOT DATA=BIOBLANC;
PLOT Y*X1;

PROC SORT DATA=BIOBLANC;
BY X2;

PROC BOXPLOT DATA=BIOBLANC;
PLOT Y*X2;

PROC SORT DATA=BIOBLANC;
BY X3;

PROC BOXPLOT DATA=BIOBLANC;
PLOT Y*X3;
```

Afficher la matrice de corrélation :

```
PROC CORR DATA=BIOBLANC;
TITLE "Matrice de corrélation";
VAR Y X1 X2 X3;
RUN;
```

Afficher la dispersion des variables de l'une autour des l'autres :

```
PROC SGSCATTER DATA=BIOBLANC;
MATRIX Y X1 X2 X3;
TITLE "DIAGRAMME DE DISPERSION";
RUN;
```

Annexe 2 : Régression linéaire simple

Afficher la corrélation entre les variables :

```
PROC CORR DATA=BIOBLANC;
TITLE "Matrice de corrélation";
VAR Y X1 X2 X3 X4 X5 X6 X7 X8;
RUN;
```

Afficher la dispersion :

```
PROC sgscatter DATA=BIOBLANC;
TITLE "Dispersion des variables par rapport à Y";
COMPARE X=Y Y=(X1 X2 X3 X4 X5 X6 X7 X8);
RUN;
```

La régression de Y par rapport à chaque variable Xi :

```
Proc reg data=Work.bioblanc;
model Y= X1 /R INFLUENCE;
title "La régression de la vente Y sur X1";
run;

Proc reg data=Work.bioblanc;
model Y= X2 /R INFLUENCE;
title "La régression de la vente Y sur X2";
run;

Proc reg data=Work.bioblanc;
model Y= X3 /R INFLUENCE;
title "La régression de la vente Y sur X3";
run;

Proc reg data=Work.bioblanc;
model Y= X4 /R INFLUENCE;
title "La régression de la vente Y sur X4";
run;

Proc reg data=Work.bioblanc;
model Y= X5 /R INFLUENCE;
title "La régression de la vente Y sur X5";
run;
```

```
Proc reg data=Work.bioblanc;
model Y= X6 /R INFLUENCE;
title "La régression de la vente Y sur X6";
run;

Proc reg data=Work.bioblanc;
model Y= X7 /R INFLUENCE;
title "La régression de la vente Y sur X7";
run;

Proc reg data=Work.bioblanc;
model Y= X8 /R INFLUENCE;
title "La régression de la vente Y sur X8";
run;
```

Annexe 3 : Régression linéaire multiple

La régression de Y sur toutes les régresseurs de la base des données :

```
Proc reg data=Work.bioblanc;
model Y= X1 X2 X3 X4 X5 X6 X7 X8/R INFLUENCE;
title "La régression de la vente Y sur X1 X2 X3 X4 X5 X6 X7 X8";
run;
```

Avec la procédure suivante nous tentons de supprimer la multi-colinéarité dans notre modèle mais à la fin nous remarquons qu'il y a une incohérence au niveau des signes coefficients de régression car cette sélection s'effectue séquentiellement et le choix est au niveau individuel.

```
Proc reg data=Work.bioblanc;
model Y= X1 X2 X3 X4 X5 X6 X7 X8/ selection=Backward;
title "La régression de la vente Y sur X1 X2 X3 X4 X5 X6 X7 X8";
run;
```

Annexe 4 : Régression PLS (Partial least squares)

Normalisation des données pour supprimer l'effet d'échelle :

```
proc standard data=Work.bioblanc mean=0 std=1
out= Bioblanc_stan;
var Y X1 X2 X3 X4 X5 X6 X7 X8;
run;
```

Cette méthode a pour objectif de maximiser la corrélation entre les variables construites et la variable cible Y tout en réduisant la dimension avec la combinaison linéaire des variables originales, ainsi, nous nous basons sur les variables significativement corrélées avec la vente Y pour construire T1(Première variable latente).

Construction de T1 :

```
proc sql;
create table table1 as
(select ID,  -0.47528*X1+0.72867*X2+0.87037*X3+0.88041*X4+0.88192*X5+0.88053*X6+0.83469*X7+0.88274*X8 as
create table table2 as
(select ID, nume/sqrt(5.3185700232999995) as T1 from work.table1);
run;
```

Vérifier si les modifications ont été effectuées :

```
proc print data=table2; run;
```

On rassemble la table contenant les variables standardisées avec la table contenant les valeurs de T1 pour chaque observation dans une table nommée final_table afin de l'utiliser pour les opérations suivantes :

```
proc sql;
create table final_table as
select * from table2 natural join bioblanc_stan;
run;
```

Affichage de la table obtenue :

```
proc print data=final_table;
```

Comme cité ci-dessus, nous cherchons le minimum de variables possibles pour construire notre modèle mais qui soit robuste, ainsi, ayant déjà construit T1 on s'intéresse à construire une autre composante à partir des résidus qui paraissent très grandes ou significatives pour apporter des informations supplémentaires pour expliquer Y. Ainsi, nous effectuons ces régressions sur les Xi sachant T1 pour repérer les variables significatives :

Régression de Y sur X1 sachant T1 :

```
proc reg data=final_table;
model Y=T1 X1;
Title "Regression de Y sur X1 sachant T1";
run;
```

Régression de Y sur X2 sachant T1 :

```
proc reg data=final_table;
model Y=T1 X2;
Title "Regression de Y sur X2 sachant T1";
run;
```

Régression de Y sur X3 sachant T1 :

```
proc reg data=final_table;
model Y=T1 X3;
Title "Regression de Y sur X3 sachant T1";
run;
```

Régression de Y sur X4 sachant T1 :

```
proc reg data=final_table;
model Y=T1 X4;
Title "Regression de Y sur X4 sachant T1";
run;
```

Régression de Y sur X5 sachant T1 :

```
proc reg data=final_table;
model Y=T1 X5;
Title "Regression de Y sur X5 sachant T1";
run;
```

Régression de Y sur X6 sachant T1 :

```
proc reg data=final_table;
model Y=T1 X6;
Title "Regression de Y sur X6 sachant T1";
run;
```

Régression de Y sur X7 sachant T1 :

```
proc reg data=final_table;
model Y=T1 X7;
Title "Regression de Y sur X7 sachant T1";
run;
```

Régression de Y sur X8 sachant T1 :

```
proc reg data=final_table;
model Y=T1 X8;
Title "Regression de Y sur X8 sachant T1";
run;
```

Pour la prochaine étape nous ne retenons que X7 car elle est la seule présentant une part importante non emmagasinée par T1 alors nous récupèrerons la part non emmagasinée par T1 à travers son résidu de la régression de X7 sur T1

On procède à la récupération du résidu de la régression de X7 sur T1 :

```
proc reg data=Work.final_table;
model X7= T1/R Influence;
output out=final_table residual=residual17;
run;

proc print data=final_table; run;
```

Normalisation du résidu et son ajout au final_table :

```
proc sql;
create table final_table as
select *,
residual17/var(residual17) as residual17_cn
from final_table;


proc print data=final_table;
run;
```

Nous confirmons effectivement que le résidu est significatif sachant T1 dans l'explication de Y :

```
proc reg data=final_table;
model Y=T1 residual17_cn;
run;
```

A partir de ce résidu nous construisons la deuxième composante T2 comme suit :

```
proc sql;
create table table as
(select ID,-0.04746*residual17 as nume2
from final_table);
create table table as
(select ID, nume2/0.04746 as T2 from table);

proc print data=table; run;

/*On l'ajoute dans la table final_table*/

proc sql;
create table final_table as
select * from table natural join final_table;

proc print data=final_table; run;
```

Nous cherchons une troisième composante permettant d'apporter encore une information supplémentaire mais au final aucune variable n'apporte de nouvelles informations :

La régression de Y sur X1 sachant T1 et T2 :

```
proc reg data=final_table;
title "La régression de Y sur X1 sachant T1 et T2";
model Y=T1 T2 X1/R influence;
run;
```

La régression de Y sur X2 sachant T1 et T2 :

```
proc reg data=final_table;
title "La régression de Y sur X2 sachant T1 et T2";
model Y=T1 T2 X2/R influence;
run;
```

La régression de Y sur X3 sachant T1 et T2 :

```
proc reg data=final_table;
title "La régression de Y sur X3 sachant T1 et T2";
model Y=T1 T2 X3/R influence;
run;
```

La régression de Y sur X4 sachant T1 et T2 :

```
proc reg data=final_table;
title "La régression de Y sur X4 sachant T1 et T2";
model Y=T1 T2 X4/R influence;
run;
```

La régression de Y sur X5 sachant T1 et T2 :

```
proc reg data=final_table;
title "La régression de Y sur X5 sachant T1 et T2";
model Y=T1 T2 X5/R influence;
run;
```

La régression de Y sur X6 sachant T1 et T2 :

```
proc reg data=final_table;
title "La régression de Y sur X6 sachant T1 et T2";
model Y=T1 T2 X6/R influence;
run;
```

La régression de Y sur X7 sachant T1 et T2 :

```
proc reg data=final_table;
title "La régression de Y sur X7 sachant T1 et T2";
model Y=T1 T2 X7/R influence;
run;
```

La régression de Y sur X7 sachant T1 et T2 :

```
proc reg data=final_table;
title "La régression de Y sur X8 sachant T1 et T2";
model Y=T1 T2 X8/R influence;
run;
```

La régression de Y sur X8 sachant T1 et T2 :

```
proc reg data=final_table;
title "La régression de Y sur X8 sachant T1 et T2";
model Y=T1 T2 X8/R influence;
run;
```

On va trouver qu'aucune des variables X1...X8 n'est significative au niveau de risque 5%. Il faut retenir que les deux composantes PLS T1 et T2.

Enfin, nous récupérons dans la table_utile les variables qu'on aura besoin à savoir Y, T1 et T2 :

```
proc sql;
create table temp1 as select ID,T1,T2 from final_table;
create table temp2 as select ID,Y from bioblanc;
create table table_utile as select * from temp1 natural join temp2;
run;
```

La régression multiple de la vente sur les deux variables latentes T1 et T2 :

```
proc reg data=table_utile;
title "La régression de Y sur T1 et T2";
model Y=T1 T2/R influence;
run;

proc print data=table_utile ; run;
```

Nous appliquons ACP pour confirmer le résultat des deux composants retenus, avec ces dernières nous pouvons expliquer plus de 94% de l'inertie totales de la population :

```
PROC PRINCOMP DATA=WORK.bioblanc_stan N=4 out=coordon
 plots=all;
var X1 X2 X3 X4 X5 X6 X7 X8;
run;
```

Validation des résultats avec PLS automatisé :

```
ods graphics on;
proc pls data=work.bioblanc_stan plots=(ParmProfiles VIP) cv=block;
model Y = X1 X2 X3 X4 X5 X6 X7 X8;
run;
ods graphics off;
```

Annexe 4 : Version R

Importation des library et du fichier CSV :

```
library(data.table)
library(Hmisc)
library(caTools)
path2data<-file.path('C:','Users',"pc","Desktop","SA","projet_moussanif")

setwd("C:/Users/pc/Desktop/SA/projet_moussanif")


bioblanc <- fread(file.path(path2data, "Donnees_projet1.csv"))
bioblanc <-bioblanc[,2:10]
head(bioblanc)
```

Division des données en données d'entrainement et données de test :

```
train_set<-bioblanc[seq(1,24)]
test_set<-bioblanc[seq(25,30)]
```

Régression linéaire multiple de Y par rapport aux variables explicatives :

```
lm<-lm(formula = Y ~ 0 + X1 + X2+ X3 + X4 + X5 +X6 + X7 + X8, data=train_set)
#print
print(lm)
#summary
print(summary(lm))
```

Normalisation des données :

```
DM_Matrix<-as.matrix(train_set)
```

Construction de la matrice de corrélation :

```
rcorr(DM_Matrix, type=c("pearson","spearman"))

Scale_DM<-scale(train_set)
head(Scale_DM)
colnames(Scale_DM)<- c("Y_cn","X1_cn","X2_cn","X3_cn","X4_cn","X5_cn","X6_cn","X7_cn","X8_cn")
DT_scale<- cbind(train_set,Scale_DM)
head(DT_scale)
```

Construction de la composante T1 :

```
DT_scale<-DT_scale[,':='(T1=(1/sqrt(0.43^2+0.69^2+0.87^2+0.89^2+0.89^2+0.89^2+0.81^2+0.89^2))
                         *((-0.43*X1_cn)+(0.69* X2_cn)+(0.87*X3_cn)+(0.89* X4_cn)+(0.89* X5_cn)+(0.89* X6_c
```

Régression de Y sur T1 et X_j j=1...8 pour chercher les variables qui contribuent de manière significative à la construction de T2 :

```
lm11<-lm(formula = Y_cn ~ 0 + T1 + X1_cn, data=DT_scale) #Non
print(summary(lm11))

lm12<-lm(formula = Y_cn ~ 0 + T1 + X2_cn, data=DT_scale) #Non
print(summary(lm12))
lm13<-lm(formula = Y_cn ~ 0 + T1 + X3_cn, data=DT_scale) #Non
print(summary(lm13))
lm14<-lm(formula = Y_cn ~ 0 + T1 + X4_cn, data=DT_scale) #Non
print(summary(lm14))
lm15<-lm(formula = Y_cn ~ 0 + T1 + X5_cn, data=DT_scale) #Non
print(summary(lm15))
lm16<-lm(formula = Y_cn ~ 0 + T1 + X6_cn, data=DT_scale) #Non
print(summary(lm16))
lm17<-lm(formula = Y_cn ~ 0 + T1 + X7_cn, data=DT_scale) #Oui
print(summary(lm17))
lm18<-lm(formula = Y_cn ~ 0 + T1 + X8_cn, data=DT_scale) #Non
print(summary(lm18))
```

- Seules les variables X7 est significative au risque de 5%

Calcul du résidu X17 de la régression de X7_nc sur T1 :

```
lm_R17<-lm(formula = X7_cn ~ 0 + T1  , data=DT_scale)
print(lm_R17)
```

Extraction des résidus :

```
X17<-resid(lm_R17)
DT_scale<- cbind(DT_scale,X17)
X17n<-X17/var(X17)
DT_scale<- cbind(DT_scale,X17n)
```

Effectuer la régression multiple de Y_cn sur T1 et X17n = x17/var(x17)

```
lm_Y17<-lm(formula = Y_cn ~ 0 + T1 + X17n  , data=DT_scale)
print(summary(lm_Y17))
```

Calcul de T2 :

```
T2=(-0.05527*X17)/0.05527
DT_scale<-cbind(DT_scale,T2)
DT_scale$T2
```

Construction de T3 :

Régression de Y sur T1, T2 et X_j j=1..8 pour chercher les variables qui contribuent de manière significative à la construction de T3 :

```
lm21<-lm(formula = Y_cn ~ 0 + T1 + T2 + X1_cn, data=DT_scale) #Non
print(summary(lm21))
lm22<-lm(formula = Y_cn ~ 0 + T1 + T2 + X2_cn, data=DT_scale) #Non
print(summary(lm22))
lm23<-lm(formula = Y_cn ~ 0 + T1 + T2 + X3_cn, data=DT_scale) #Non
print(summary(lm23))
lm24<-lm(formula = Y_cn ~ 0 + T1 + T2 + X4_cn, data=DT_scale) #Non
print(summary(lm24))
lm25<-lm(formula = Y_cn ~ 0 + T1 + T2 + X5_cn, data=DT_scale) #Non
print(summary(lm25))
lm26<-lm(formula = Y_cn ~ 0 + T1 + T2 + X6_cn, data=DT_scale) #Non
print(summary(lm26))

lm27<-lm(formula = Y_cn ~ 0 + T1 + T2 + X7_cn, data=DT_scale) #Non
print(summary(lm27))
lm28<-lm(formula = Y_cn ~ 0 + T1 + T2 + X8_cn, data=DT_scale) #Non
print(summary(lm28))
```

- Aucune des variables X1...X8 n'est significative au risque de 5%. Il faut retenir que les deux composantes PLS T1 T2.

Construction de l'équation de régression PLS à deux composantes :

Régression de Y sur T1 et T2 :

```
lm_PLS<-lm(formula = Y ~   T1 + T2, data=DT_scale)
print(summary(lm_PLS))
attach(train_set)
Y_pred=6.50146556837332-0.0794302243222059*X1+0.0532153618252829*X2+0.0252401648409459*X3+0.065183223465237
plot(Y,Y_pred)
abline(a=0,b=1)
```

Calcul de l'erreur quadratique moyenne :

```
attach(train_set)
fitted=6.44908162853694-0.0772729118518608*X1+0.0549065267121064*X2+0.0243873350790104*X3+0.07041806030878
attach(test_set)
prediction=6.44908162853694-0.0772729118518608*X1+0.0549065267121064*X2+0.0243873350790104*X3+0.0704180603

RMSE_pls_fitted=sqrt(mean(train_set$Y-fitted)^2)
RMSE_pls_predict=sqrt(mean(test_set$Y-prediction)^2)

library(mctest)
library(car)
imcdiag(lm_PLS,all = TRUE)
vif(lm_PLS)
```

Utilisation du PLS automatisé :

```
library(plsdepot)
train_set=train_set[,c(2:9,1)]
test_set=test_set[,c(2:9,1)]
head(train_set)
# Ceci confirme le choix du nombre de composante en haut nbre=2

modele=plsreg1(train_set[,c(1:8)],train_set[,c(9)],crosval = TRUE)
print(modele$Q2)
print(modele$R2)
```

Ci-dessous le code confirmant les résultats obtenus précédemment concernant le choix du nombre de composantes (n=2) :

```
modele=plsreg1(train_set[,c(1:8)],train_set[,c(9)],crosval = TRUE)
print(modele$Q2)
print(modele$R2)

plot(train_set$Y,modele$y.pred,type='n',xlab='Original',ylab='Predicted')
abline(a=0,b=1)
text(train_set$Y,modele$y.pred,col = "blue")
```

Annexe 5 : Régression pénalisée (Ridge et Lasso) sur R :

1- La régression Ridge :

```
library(MASS)

#Mis en place de la régression ridge avec 99 valeur possible entre 0.1 à 10

model_ridge=lm.ridge(Y~., data=train_set,lambda = seq(0.1,10,0.1))
print(model_ridge)

#Répérage de lambda permettant de minimiser de cross validation GCV

plot(seq(0.1,10,0.1),model_ridge$GCV,xlab = "lambda",ylab = "GCV")
model_ridge$lambda[which.min(model_ridge$GCV)]
help(model_ridge$GCV)
model_ridge=lm.ridge(Y~., data=train_set,lambda = 4.5)
print(model_ridge)
```

Construction de la méthode de prédiction absente dans le package MASS :

```
attach(train_set)
fitted=9.273352-0.1137110*X1+2.575409*10^(-2)*X2+1.936221*10^(-2)*X3+4.852966*10^(-2)*X4+1.906539*10^(-2)*›

attach(test_set)
prediction=9.273352-0.1137110*X1+2.575409*10^(-2)*X2+1.936221*10^(-2)*X3+4.852966*10^(-2)*X4+1.906539*10^(-
```

Ajustement du modèle et calcul du RMSE :

```
plot(my_ridge,train_set$Y)
abline(a=0,b=1)

# L'erreur commise

RMSE_ridge_fitted=sqrt(mean(train_set$Y-fitted)^2)
RMSE_ridge_predict=sqrt(mean(test_set$Y-prediction)^2)
```

2- La régression LASSO :

```
library("lars")
model_lasso=lars(as.matrix(train_set[,1:8]),train_set$Y,type="lasso",
                 trace=F,normalize=TRUE)
plot(model_lasso,xvar = 'df', plottype = 'coeff')
print(model_lasso$beta)

plot(model_lasso$df,summary(model_lasso)$Rss,
     xlab='Df',ylab='Rss',main='LASSO')
```

Utilisation de la Cross-validation pour déterminer le coefficient optimal :

```
cv=cv.lars(as.matrix(train_set[,1:8]),train_set$Y,K=10)

print(model_lasso$lambda[12])
print(model_lasso$beta[12,])
```

Modèle final après élimination de certaines variables par LASSO par manque de significativité :

```
model_lasso_final=lm(Y~X5+X6, data=train_set)
summary(model_lasso_final)

attach(train_set)
fitted=6.09+5.514*10**(-2)*X5+4.674*10**(-4)*X6
attach(test_set)
prediction=6.09+5.514*10**(-2)*X5+4.674*10**(-4)*X6
plot(fitted,train_set$Y)
abline(a=0,b=1)
```

Calcul des erreurs (RMSE) :

```
RMSE_lasso_fitted=sqrt(mean((fitted-train_set$Y)^2))
RMSE_lasso_predict=sqrt(mean((prediction-test_set$Y)^2))
```

Calcul des statistiques sur la performance des modèles :

```
compare_tools=cbind(RMSE_ridge_fitted,RMSE_lasso_fitted,RMSE_pls_fitted,RMSE_ridge_predict,
                    RMSE_lasso_predict,RMSE_pls_predict)
compare_tools
```