Projet sur la régression linéaire multiple

Clément Weinreich - Emma Marqueton - Antoine Barthas
ENSC 2A Groupe 4

June 15, 2022

1 Introduction

Dans le cadre du module de modélisation statistique, nous avons été amené à réaliser une étude sur les ventes de n=45 stations essence et sur les facteurs pouvant les influencer. Pour se faire, nous disposons d'un jeu de données contenant les informations suivantes:

- les ventes de la station exprimées en milliers de litres (variable ventes),
- le nombre de pompes de la station (variable nbpompes),
- le nombre de concurrents dans la zone desservie par la station (variable nbconc),
- le trafic quotidien exprimé en milliers de voitures (variable trafic).

Le but de ce projet est donc de modéliser les ventes d'une station en fonction des variables disponibles dans le jeu de données.

2 Description du jeu de données

Dans un premier temps, récupérons les donnnés à analyser.

[1]:
```
dataset = read.table("../input/station/station.txt",header = TRUE) # Lecture des␣
↪données
attach(dataset) # Pour pouvoir utiliser les noms de colonnes en tant que variable
```

[2]:
```
head(dataset,5) # On affiche une une partie des données
```

A data.frame: 5 × 4

	ventes <int>	nbpompes <int>	nbconc <int>	trafic <int>
1	203	4	4	13
2	262	18	21	18
3	247	16	19	10
4	239	11	12	15
5	241	10	11	19

On a donc pour chaque station les ventes en milliers de litres, le nombre de pompes disponibles, le nombre de concurrents dans la zone desservie par la station et le trafic alentour en milliers de véhicules.

[3]:
```
summary(dataset) # résumé de l'échantillon contenant les informations sur les stations
```

```
     ventes          nbpompes        nbconc          trafic
 Min.   :203.0   Min.   : 3.00   Min.   : 2.00   Min.   : 8.0
 1st Qu.:231.0   1st Qu.: 9.00   1st Qu.: 9.00   1st Qu.:13.0
 Median :241.0   Median :10.00   Median :11.00   Median :17.0
 Mean   :241.2   Mean   :11.36   Mean   :12.71   Mean   :16.4
```

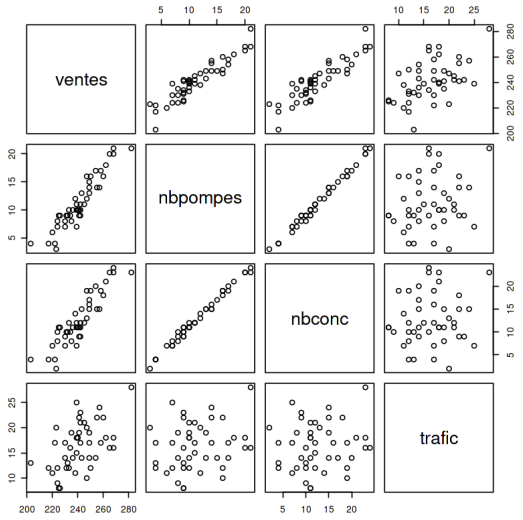
```
 3rd Qu.:249.0   3rd Qu.:14.00   3rd Qu.:16.00   3rd Qu.:19.0
 Max.   :282.0   Max.   :21.00   Max.   :24.00   Max.   :28.0
```

Si on étudie plus précisément la variables ventes, on voit qu'elle est comprise entre 203 et 282. En moyenne, les stations vendent 241.2 milliers de litres, avec une médiane de 241. Il y a donc autant de stations qui vendent en dessous et au dessus de 241 mille litres d'essence.

[4]:
```
round(apply(dataset,2,sd),2) # écart type de cet échantillon
```

ventes: 15.53 **nbpompes:** 4.65 **nbconc:** 5.53 **trafic:** 4.65

[5]:
```
plot(dataset) # on représente toutes les données
```



On observe une forte corrélation entre nbpompes et nbconc, entre ventes et nbpompes et enfin entre ventes et nbconc. D'après ces nuages de point, il n'y a visiblement pas de corrélation linéaire entre le trafic et les autres variables. On remarque tout de même une "structure" entre trafic et ventes, ce qui laisse penser que ces deux variables ne sont pas complètement indépendantes.

[ ]:
```
require(PCAmixdata)# permet de charger le package "PCAmixdata"
```

[7]:
```
# Mise en oeuvre de l'ACP et Choix du nombre d'axes à retenir
ACPStations = PCAmix(X.quanti=dataset, graph=FALSE) # Stocke les calculs de l'ACP dans
 ↪l'objet ACPStations
round(ACPStations$eig,digits=2) # Valeurs propres et pourcentage de variance expliqué
 ↪pour chaque axe
```

A matrix: 4 × 3 of type dbl

	Eigenvalue	Proportion	Cumulative
dim 1	3.01	75.16	75.16
dim 2	0.98	24.49	99.65
dim 3	0.01	0.32	99.97
dim 4	0.00	0.03	100.00

Cette ACP est centrée réduite et donne le même poids à tous les individus de l'étude. Pour choisir les axes retenus, on utilise le critère de Kaiser : on conserve seulement les axes factoriels associés à une valeur propre (eigenvalue) plus grande que 1. Ici on retient donc les 2 premiers axes: le 1er axe factoriel explique 75,16% de l'inertie (ou de la variance). Le deuxième axe explique 24.49% d'inertie supplémentaire. Ainsi, en considérant l'espace 1-2, on récupère 99.65% de l'information

[8]:
```
# Sorties numeriques pour les variables
round(ACPStations$quanti$cos2,digit=3) # Cosinus carrés associés aux variables
```

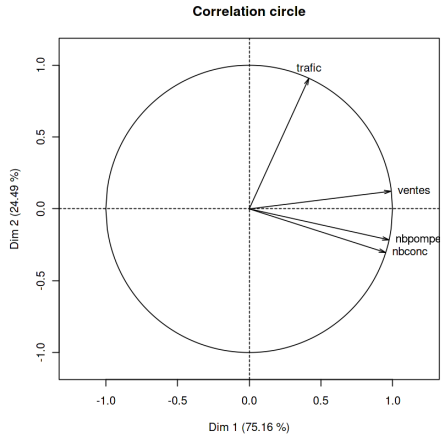
A matrix: 4 × 4 of type dbl

	dim 1	dim 2	dim 3	dim 4
ventes	0.976	0.015	0.008	0.000
nbpompes	0.951	0.047	0.001	0.001
nbconc	0.905	0.092	0.002	0.001
trafic	0.173	0.826	0.001	0.000

La qualité de représentation des variables sur les axes est donnée par les cosinus carrés. On a donc sur l'axe 1, ventes, nbpompes et nbconc qui sont bien représentés. Le trafic est très mal représenté (seulement 17%). L'axe 2 représente très bien le trafic. Il complète bien l'axe 1 avec une bonne représentation des 3 autres variables. Les axes 3 et 4 représentent extrêmement mal l'ensemble des variables (>1%).

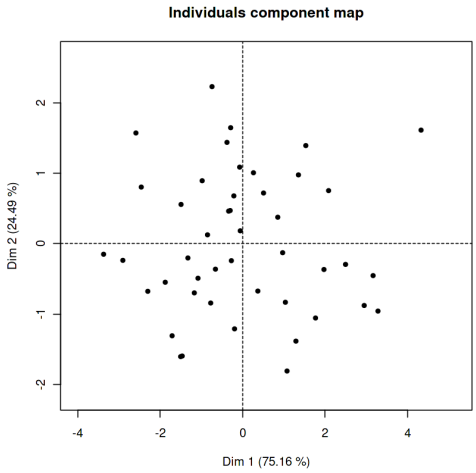
Finalement, en étudiant les axes 1 et 2 nous pouvons raisonnablement discuter de toutes les informations concernant les stations car elles sont tous représentés à au moins 99%.

[9]:
```
plot(ACPStations,axes=c(1,2),choice="cor") # Affichage du cercle des corrélations des␣
↪variables (plan 1-2)
```



Comme nous l'avons discuté précédemment, les variables sont bien représentées dans ce plan: tous les vecteurs de projection sont proches de la circonférence du cercle. On voit que les variables ventes, nbpompes et nbconc semblent positivement corrélées (vecteurs de projection dans le même sens). Cela signifie que plus le nombre de pompes et de concurrents est élevé, plus les ventes sont importantes. En revanche, le trafic est décorrélé de nbconc et nbconc (vecteurs de projection orthogonaux), mais légèrement corrélé positivement aux ventes. Cela signifie qu'à priori, le trafic n'influe pas sur le nombre de pompes et de concurents d'une station, mais il influe légèrement ses ventes. Cela paraît cohérent: plus il y a de monde, plus il y a de chance d'avoir de client.

[10]:
```
plot(ACPStations,axes=c(1,2),choice="ind",label=TRUE) # Affichage du graphique des␣
↪individus (plan 1-2)
```



Ce nuage de points nous montre que les stations occupent équitablement le plan et il n'existe à priori pas d'individu marginal.

L'ensemble de ces données nous permettent de dire que les ventes des stations sont principalement influencées par le nombre de pompes et le nombre de concurrents. De plus, on peut noter que ces deux variables sont fortement corrélées. On peut expliquer cela par le fait que plus une station a de concurrents, plus elle va se rendre attractive en augmentant le nombre de pompes qu'elle met à disposition.

3 Régression linéaire multiple

3.1 Estimation du modèle de régression linéaire multiples

Construction d'un premier modèle de régression linéaire multiple avec les 3 variables explicatives.

[11]:
```
modele1=lm(ventes~nbpompes+nbconc+trafic,data=dataset)
summary(modele1)
```

```
Call:
lm(formula = ventes ~ nbpompes + nbconc + trafic, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-13.1412  -0.2876   0.1360   0.7434   2.0179

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 189.7673     1.6530 114.804  < 2e-16 ***
nbpompes      2.5507     1.3888   1.837   0.0735 .
nbconc        0.2755     1.1504   0.239   0.8119
trafic        1.1592     0.1464   7.920 8.55e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.213 on 41 degrees of freedom
Multiple R-squared:  0.9811,        Adjusted R-squared:  0.9797
F-statistic: 709.1 on 3 and 41 DF,  p-value: < 2.2e-16
```

Lors du test de Fisher de significativité du modèle, la p-value vaut 2.2e-16, ce qui est inférieur à α = 5% donc le modèle est utile. Le R^2 ajusté est une quantité comprise entre 0 et 1, qui représente le pourcentage de variabilité de ventes expliqué par le modèle. Ici, 97.97% de la variabilité des ventes est expliquée par le modèle, et 2.03% de la variabilité ne sont pas expliquées par le modèle. L'estimation sans biais de l'écart-type vaut 2.21, or les valeurs des ventes oscillent entre 203 et 282 donc ce taux d'erreur reste acceptable.

Concernant les coefficients, les p-values de $\beta 0$ (ordonnée à l'origine) et $\beta 3$ (trafic) sont inférieures à 5% donc on ne peut pas simplifier le modèle en les supprimant. En revanche, les p-values de $\beta 1$ (nbpompes) et $\beta 2$ (nbconc) sont supérieures à 5% donc il est possible de ne pas les prendre en compte pour simplifier le modèle.

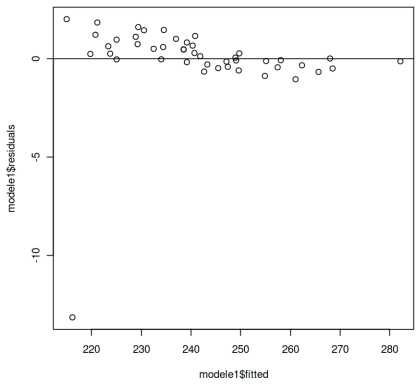
On ne peut pas retenir le modèle à 3 variables. En effet, nbpompes et nbconc sont fortement collinéaires comme nous avons vu dans la partie précédente: nous ne sommes pas capable de faire ressortir leur utilité de manière individuelle. Il ne faut pas les faire sortir du modèle simultanément. Après avoir évalué la qualité du modèle en vérifiant la normalité et le comportement aléatoire des résidus, il faudra donc supprimer du modèle la variable la moins utile (avec la p-value la plus haute), soit nbconc. Continuons donc d'étudier la qualité du modèle.

[12]:
```
# Vérifions la normalité des résidus
shapiro.test(modele1$residuals)
```

```
        Shapiro-Wilk normality test
data:  modele1$residuals
W = 0.44175, p-value = 7.354e-12
```

La p-value est inférieure à 5%, on rejette donc H0: normalité des résidus. Représentons les résidus pour essayer d'en trouver la cause.

[13]:
```
# Vérifions le comportement aléatoire des résidus
plot(modele1$fitted, modele1$residuals)
abline(h=0)
```



On observe une structure dans les résidus et une station anormalement petite en -13 environ. Ce point pourrait expliquer le rejet de la normalité. Identifions la station qui pose problème.

[14]:
```
# Identifions la station qui pose problème
modele1$residuals[which(abs(modele1$residuals)>10)]
dataset[which(abs(modele1$residuals)>10),]
```

1: -13.1412416821542

A data.frame: 1 × 4

	ventes <int>	nbpompes <int>	nbconc <int>	trafic <int>
1	203	4	4	13

On observe que cette station est celle qui fait le moins de ventes dans notre dataset. Étudions rapidement les données sans cette station :

[15]:
```
summary(dataset[-which(abs(modele1$residuals)>10),])
```

```
     ventes          nbpompes         nbconc          trafic
 Min.   :217.0   Min.   : 3.00   Min.   : 2.00   Min.   : 8.00
 1st Qu.:231.8   1st Qu.: 9.00   1st Qu.: 9.75   1st Qu.:12.75
 Median :241.0   Median :10.00   Median :11.50   Median :17.00
 Mean   :242.1   Mean   :11.52   Mean   :12.91   Mean   :16.48
 3rd Qu.:249.2   3rd Qu.:14.25   3rd Qu.:16.25   3rd Qu.:19.25
 Max.   :282.0   Max.   :21.00   Max.   :24.00   Max.   :28.00
```

On remarque que le nombre minimal de ventes est maintenant à 217.

Supprimons la station qui pose problème afin de voir si les résidus suivent une loi normale sans celle-ci.

[15]:
```
modele2=lm(ventes~nbpompes+nbconc+trafic,data=dataset[-which(abs(modele1$residuals)>10),])
summary(modele2)
```

```
Call:
lm(formula = ventes ~ nbpompes + nbconc + trafic, data =␣
 ↪dataset[-which(abs(modele1$residuals) >
    10), ])

Residuals:
    Min      1Q  Median      3Q     Max
-0.6747 -0.4380 -0.1948  0.4885  0.8659

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 192.48643    0.38182 504.135  < 2e-16 ***
nbpompes      3.47407    0.31198  11.136 7.94e-14 ***
nbconc       -0.57848    0.25877  -2.236    0.031 *
trafic        1.03561    0.03299  31.393  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4942 on 40 degrees of freedom
Multiple R-squared:  0.9989,        Adjusted R-squared:  0.9988
F-statistic: 1.243e+04 on 3 and 40 DF,  p-value: < 2.2e-16
```

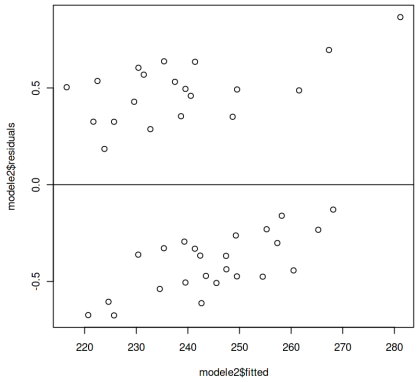
Le test de Fisher de significativité du modèle nous permet d'affirmer que le modèle reste utile après avoir supprimé la station problématique. Le taux de variabilité des ventes est expliquée à 99.88% par le modèle. L'estimation sans biais de l'écart-type vaut 0.49, or les valeurs des ventes oscillent entre 217 et 282 donc ce taux d'erreur est très acceptable. On observe déjà que le modèle 2 est meilleur que le modèle 1 de part un R^2 plus élevé, et un écart-type plus faible.

Les p-values des 4 coefficients sont inférieures à 5%, elles sont donc utiles au modèle.

Étudions maintenant la normalité, et le comportement aléatoire desrésidus.

[16]:
```
# Vérifions la normalité, et le comportement aléatoire des résidus
shapiro.test(modele2$residuals)
plot(modele2$fitted, modele2$residuals)
abline(h=0)
```

```
        Shapiro-Wilk normality test
data:  modele2$residuals
W = 0.88682, p-value = 0.0004388
```



On remarque que la p-value du test de shapiro est inférieure à 5%, on rejette donc H0: la normalité des résidus. On constate également des structures dans les résidus, ils n'ont donc pas un comportement aléatoire. Cela pourrait venir du fait de la collinéarité des variables nbconc et nbpompes.

Lors de l'étude des coefficients du modèle 2, nous avons pu constater que la p-value de nbconc était relativement proche de 5%. De plus, nous avons précédemment constaté que cette variable n'était pas utile au modèle 1. Il est donc pertinent de recréer un modèle à deux variables, ne prenant en compte que le nombre de pompes et le trafic, tout en supprimant la station problématique observée précédemment.

3.2 Création d'un modèle de régression linéaire multiple à 2 variables

[17]:
```
modele3=lm(ventes~nbpompes+trafic,data=dataset[-which(abs(modele1$residuals)>10),])
summary(modele3)
```

```
Call:
lm(formula = ventes ~ nbpompes + trafic, data = dataset[-which(abs(modele1$residuals) >
    10), ])

Residuals:
     Min       1Q   Median       3Q      Max
-0.85703 -0.44413 -0.04286  0.37860  0.89185

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 191.98974    0.32530  590.20   <2e-16 ***
nbpompes      2.77766    0.01763  157.53   <2e-16 ***
trafic        1.09956    0.01721   63.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.5178 on 41 degrees of freedom
Multiple R-squared:  0.9988,        Adjusted R-squared:  0.9987
F-statistic: 1.699e+04 on 2 and 41 DF,  p-value: < 2.2e-16
```

Lors du test de Fisher de significativité du modèle, la p-value vaut 2.2e-16, ce qui est inférieur à $\alpha = 5\%$ donc le modèle reste utile avec 2 variables. 99.87% du taux de variabilité des ventes est expliqué par le modèle. L'estimation sans biais de l'écart-type vaut 0.51, ce qui reste très acceptable au vu de la plage de valeur des ventes (217-282).

Les p-values de $\beta 0$ (ordonnée à l'origine), $\beta 1$ (nbpompes) et $\beta 2$ (trafic) sont inférieures à 5%, tous les coefficients sont donc utiles au modèle.

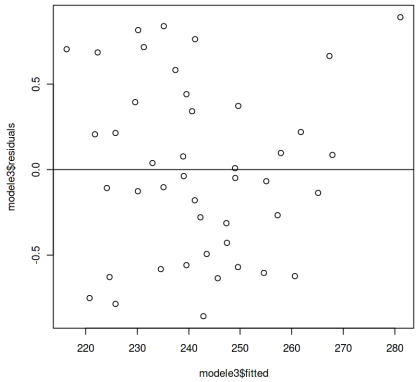
Étudions maintenant la normalité, et le comportement aléatoire des résidus.

[18]:
```
#Vérifions une dernière fois la normalité des résidus
shapiro.test(modele3$residuals)
```

```
        Shapiro-Wilk normality test
data:  modele3$residuals
W = 0.95481, p-value = 0.08319
```

La p-value est supérieure à 5%, on ne peut pas rejeter H0, on a donc la normalité des résidus. Étudions maintenant si la suppression de la variable nbconc dans le modèle a permis d'enlever les structures observées parmi les résidus.

[19]:
```
# Vérifions le comportement aléatoire des résidus
plot(modele3$fitted, modele3$residuals)
abline(h=0)
```



Les résidus du modèle 3 ne présentent aucune structure ni en variance ni en forme. On peut donc attester que le modèle 3 est meilleur que le modèle 1 et 2.

4 Interprétation et utilisation du modèle de régression linéaire multiple à 2 variables

Interprétons le signe des coefficients de régression du modèle 3. Tous les coefficients du modèle sont positifs. Les ventes varient donc de la même manière que le trafic et le nombre de pompes. Le nombre de pompe influe 2.7 fois plus les ventes que le trafic, ce qui paraît cohérent avec la réalité. Plus une station propose de pompes, plus de gens vont s'y rendre. De plus, en général les gens se rendent spécifiquement

à une station service, même si elle est peu desservie (dépend des habitudes, du prix du carburant). Les ventes ne dépendent pas principalement du trafic alentours, même si il les influe positivement.

Testons ce modèle en prédisant les ventes des stations étant donné un couple de variables explicatives (nbpompes et trafic).

[20]:
```
# Calculons les ventes à partir des valeurs choisies comprises dans
# le support d'observation des variables nbpompes et trafic.
# Testons avec des valeurs "classiques", minimales et maximales sur le support
# d'observation.
x0 = data.frame(nbpompes = c(12,3,21), trafic = c(16,8,28))

# Intervalle de confiance des prédiction de chacune des valeurs
predict(modele3, new = x0, interval="pred", level=0.95)
```

A matrix: 3 × 3 of type dbl

	fit	lwr	upr
1	242.9145	241.8567	243.9724
2	209.1192	207.9953	210.2431
3	281.1082	279.9501	282.2662

[21]:
```
# Intervalle de confiance de l'estimation de l'esperance de chacune des valeurs
predict(modele3, new = x0, interval="conf", level=0.95)
```

A matrix: 3 × 3 of type dbl

	fit	lwr	upr
1	242.9145	242.7548	243.0743
2	209.1192	208.7073	209.5311
3	281.1082	280.6106	281.6057

On constate que les ventes correspondent aux ventes moyennes observées dans le support d'observation, le modèle paraît donc cohérent avec la réalité. Même si une variable explicative a été supprimée du modèle, les prédictions restent parfaitement cohérentes. On remarque que l'intervalle de confiance (IC) est plus restreint lors de l'estimation de l'espérance des prédictions, que dans l'estimation des prédictions directement. Cela signifie qu'il y a moins de variation lors de l'estimation de l'espérance des prédictions.

Par exemple, si on veut estimer le nombre de ventes d'une station qui a 12 pompes, et un trafic de 16 milliers de voitures par jour, on obtient 242.9145 milliers de litres vendus. Les IC associés sont :

- [241.8567 , 243.9724] pour l'IC de l'estimation des ventes
- [242.7548 , 243.0743] pour l'IC de l'estimation de l'espérance des ventes

On voit bien que l'IC de l'estimation de l'espérance des ventes est bien moins étendu que l'IC de l'estimation des ventes.

5 Comparaison du modèle à 2 variables aux 3 différents modèles de régression linéaire simple

Comparons notre modèle à 2 variable, avec 3 modèles de régression linéaire simple.

[ ]:
```
modelePompes = lm(ventes~nbpompes, dataset[-which(abs(modele1$residuals)>10),])
summary(modelePompes)
```

[ ]:
```
modeleTrafic = lm(ventes~trafic, dataset[-which(abs(modele1$residuals)>10),])
summary(modeleTrafic)
```

```
[ ]: modeleNbConc = lm(ventes~nbconc, dataset[-which(abs(modele1$residuals)>10),])
     summary(modeleNbConc)
```

```
Call:
lm(formula = ventes ~ nbpompes, data = dataset[-which(abs(model
    10), ])

Residuals:
     Min       1Q   Median       3Q      Max
-9.5639 -3.5420 -0.5164  3.4471 11.5239

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 207.6297     2.1214   97.87   <2e-16 ***
nbpompes      2.9927     0.1714   17.46   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.129 on 42 degrees of freedom
Multiple R-squared:  0.8789,    Adjusted R-squared:  0.876
F-statistic: 304.7 on 1 and 42 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = ventes ~ trafic, data = dataset[-which(abs(modele1
    10), ])

Residuals:
     Min       1Q   Median       3Q      Max
-24.811  -7.598  -2.255   6.769  26.658

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 215.4648     7.0347  30.629  < 2e-16 ***
trafic        1.6173     0.4111   3.934 0.000307 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.6 on 42 degrees of freedom
Multiple R-squared:  0.2693,    Adjusted R-squared:  0.2519
F-statistic: 15.48 on 1 and 42 DF,  p-value: 0.0003073
```

```
Call:
lm(formula = ventes ~ nbconc, data = dataset[-which(abs(modele1$residuals) >
    10), ])

Residuals:
     Min       1Q   Median       3Q      Max
-12.5063  -4.1096  -0.1066   4.5370  15.5332

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 210.9592     2.5356    83.2   <2e-16 ***
nbconc        2.4134     0.1814    13.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.453 on 42 degrees of freedom
Multiple R-squared:  0.8082,    Adjusted R-squared:  0.8037
F-statistic:   177 on 1 and 42 DF,  p-value: < 2.2e-16
```

Pour mieux comparer ces modèles, on peut utiliser ce tableau récapitulatif :

Modèle		p-value du test de significativité du modèle	R^2(R^2 ajusté)	Estimation sans bias de l'écart type (sur une plage de valeurs allant de 217 à 282)
simple	ventes ∼nbpompes	<2.2e-16	0.8789 (0.876)	5.129
	ventes ∼trafic	0.0003073	0.2693 (0.2519)	12.6
	ventes ∼nbconc	<2.2e-16	0.8082 (0.8037)	6.453
multiple	ventes ∼nbpompes + trafic	<2.2e-16	0.9988 (0.9987)	0.5178

On peut voir que tous les modèles sont significatifs. Si on compare le pourcentage de variabilité des ventes expliqué par chacun des modèles, on remarque que le modèle de régression multiple est bien meilleur. Au vu du R^2, le modèle de régression linéaire simple qui explique le mieux la variabilité des ventes est celui qui utilise la variable nbpompes en variable explicative. Son R^2 est de 0.8789, ce qui reste inférieur au R^2 ajusté du modèle de régression linéaire multiple qui est à 0.9987. Le modèle qui explique le moins bien la variabilité des ventes est le modèle de régression linéaire simple avec trafic pour variable explicative (R^2 = 0.2693). Concernant l'écart type, on constate les mêmes tendances. Le modèle qui a le plus grand écart type est ventes trafic avec 12.6 . L'écart type du modèle de régression simple qui explique le mieux la variabilité des ventes est de 5.129 . L'écart type du modèle de régression linéaire multiple est bien inférieur aux modèle de régression linéaire simple : 0.5178 .

Au vu de ces comparaisons, on peut affirmer que le modèle à deux variables est préférable aux trois différents modèles de régression linéaire simple.