Projet : Régression linéaire avec R

Farès Fadili

Part I : Theoretical questions

1)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

β1 est calculée à partir des xi déterministes et des yi suivant une loi normale. Par conséquent, $\widehat{\boldsymbol{\beta}}_1$ suit une loi normale.
Par ailleurs, $\beta_0 = \bar{y} - \beta_1 \bar{x}$ qui est une forme linéaire de $\widehat{\boldsymbol{\beta}}_1$, suit lui-même une loi normale. Ainsi, le couple $(\hat{\beta}_0, \hat{\beta}_1)$ suit une loi normale.

2) Y =β0+β1X+ε avec ε~N(0,σ2) => étant donné que X n'est pas aléatoire, et que ε suit une loi normale centrée réduite, alors Y est aléatoire et suit une loi normale.
Par conséquent, les variables Yi sont indépendantes, et sont distribuées selon une loi normale.

3) On pourrait utiliser ce modèle pour prédire la tension artérielle nourriture en fonction d'un individu. On souhaite savoir sous quelle forme cette influence peut être exprimée. Le but est d'expliquer comment la tension artérielle varie en fonction de l'âge, et éventuellement prédire la tension à partir de l'âge.
La variable cible Y correspond ici à la variable aléatoire tension, c'est la variable à expliquer, ou à régresser, ou variable réponse, ou variable dépendante.
La variable X correspond à la variable âge, c'est la variable explicative, régresser, ou variable indépendante.
A partir d'un échantillon représentatif, on peut estimer les coefficients $\widehat{\boldsymbol{\beta}}_0$ et $\widehat{\boldsymbol{\beta}}_1$ de la régression correspondante, qui permettront de prédire la tension à partir d'un âge donné.
Comme vu dans le cours, les deux coefficients s'obtiennent par la minimisation de l'erreur quadratique.

Part II : Practical applications

Exercice 1

```
R 4.1.1 · ~/Desktop/Mathematics for Data Scientists/Lab/
> setwd("~/Desktop/Mathematics for Data Scientists/Lab")
> data <- read.csv("ozone.txt",sep="")
> print(data)
   maxO3  T12 Ne12 maxO3v
1   63.6 13.4    7   95.6
2   89.6 15.0    4  100.2
3   79.0  7.9    8  105.6
4   81.2 13.1    7   95.2
```

```
ozone    50 obs. of 4 variables
```

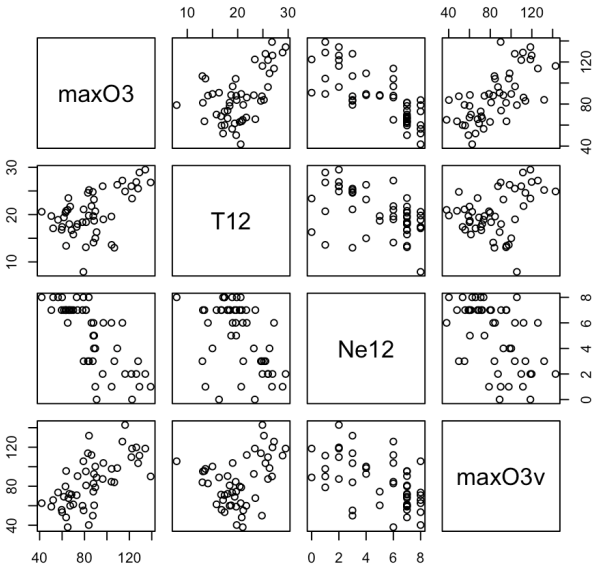
Le dataset contient 50 observations et 4 variables.

Exercice 2

```
> summary(data$maxO3)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   41.8    66.6    83.9    86.3   102.2   139.0
```

Ci-dessus le minimum, le 1er quartile, la médiane, la moyenne, le 3e quartile et le maximum de la quantité d'ozone ce jour-ci.

```
> plot(data)
```



plot(data) affiche tous les graphiques des variables les unes par rapport aux autres (Vi, Vj) avec i, j allant de 1 à 4.
On peut remarquer sur le graphique que certaines variables sont corrélées, et d'autres moins ou pas du tout.
Ainsi, nous pouvons approximativement dire que les variables corrélées positivement sont : maxO3 & T12, maxO3 & maxO3v, T12 & max03v (et toutes inversement).

A l'inverse, celles corrélées négativement sont : maxO3 & Ne12, T12 & Ne12 (et toutes inversement).
Enfin, celles pas corrélées du tout sont : Ne12 & T12, maxO3v & T12, maxO3 et Ne12, max O3v & Ne12.

```
> cor(data)
             max03        T12        Ne12      max03v
max03    1.0000000  0.5282814 -0.7701171  0.6643758
T12      0.5282814  1.0000000 -0.4880260  0.3353368
Ne12    -0.7701171 -0.4880260  1.0000000 -0.5293615
max03v   0.6643758  0.3353368 -0.5293615  1.0000000
```

Ne12 est la variable la plus corrélée avec maxO3 : le coefficient de corrélation en valeur absolue est égal à 0.770 (c'est le plus grand de tous).
Il s'agit d'une corrélation négative assez forte, c'est-à-dire que :
Ne12i < Ne12j => maxO3i > mac03j, et ceci pour plus de 77% des cas.

Exercice 3

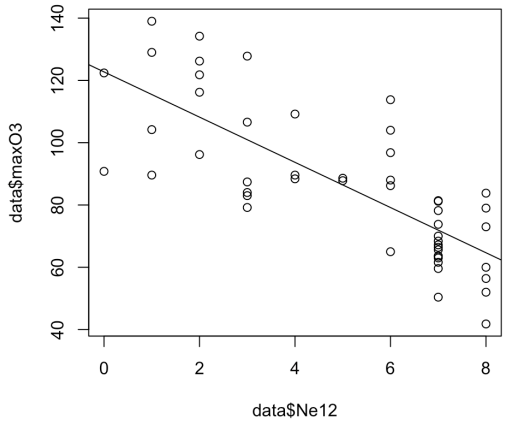
```
> oz.regsimple <- lm(data$max03~data$Ne12)
> summary(oz.regsimple)

Call:
lm(formula = data$max03 ~ data$Ne12)

Residuals:
    Min      1Q  Median      3Q     Max
-31.944 -11.045  -2.726   9.424  34.615

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  122.744      4.872  25.197  < 2e-16 ***
data$Ne12     -7.260      0.868  -8.364 6.24e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.4 on 48 degrees of freedom
Multiple R-squared:  0.5931,    Adjusted R-squared:  0.5846
F-statistic: 69.96 on 1 and 48 DF,  p-value: 6.236e-11
```

Les coefficients « estimates » (en l'occurrence ceux de « (Intercept) » et « data$Ne12 ») correspondent respectivement à :
- l'ordonnée à l'origine β0, c'est-à-dire la valeur de max03 quand Ne12 = 0,
- à la pente de la régression, c'est-à-dire le changement ou bien le delta de max03 pour un changement d'une unité de Ne12.

Le coefficient « estimate » $\widehat{\beta 1}$ est égal à -7.260 ; il correspond au fait que si j'augmente Ne12 d'une unité, maxO3 diminue de 7.260 unités.

Il confirme la tendance de la corrélation calculée dans **l'Exercice 2** (-0.770) en termes du signe négatif et de la valeur. Cette valeur exprime le fait que plus la nébulosité augmente, la quantité maximale d'ozone mesurée dans la journée diminue.

```
> plot(data$Ne12 , data$maxO3)
> abline(122.744 , -7.260)
```



Plot() et abline() confirment bien une bonne estimation de la régression de notre dataset.

Exercice 4

Formule théorique générale :

(i) Un IC de b_0 au niveau $1-\alpha$ est donné par :

$$\left[\hat{b}_0 - t\ \hat{\sigma}_{\hat{b}_0}, \hat{b}_0 + t\ \hat{\sigma}_{\hat{b}_0}\right]$$

où t représente le quantile de niveau $(1-\alpha/2)$ d'une loi de Student $n-2$.

(ii) Un IC de $b_1$au niveau $1-\alpha$ est donné par :

$$\left[\hat{b}_1 - t\ \hat{\sigma}_{\hat{b}_1}, \hat{b}_1 + t\ \hat{\sigma}_{\hat{b}_1}\right].$$

Interprétation simple de la formule théorique générale
L'intervalle de confiance IC à 1- α (niveau de risque α) pour β1est défini par :
P(β1 ∈ IC) = 1- α

```
> confint(oz.regsimple, level=0.80)
                   10 %       90 %
(Intercept) 116.414173 129.074480
data$Ne12    -8.387696  -6.131956
```

Dans notre cas, l'intervalle de confiance à 90% de la pente β1 correspond à [-8.387 ; -6.131] ; on a donc une probabilité de 90% que cet intervalle contienne la vraie pente β1.

Exercice 5

```
> summary(oz.regsimple)

Call:
lm(formula = data$maxO3 ~ data$Ne12)

Residuals:
    Min      1Q  Median      3Q     Max
-31.944 -11.045  -2.726   9.424  34.615

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  122.744      4.872  25.197  < 2e-16 ***
data$Ne12     -7.260      0.868  -8.364 6.24e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.4 on 48 degrees of freedom
Multiple R-squared:  0.5931,    Adjusted R-squared:  0.5846
F-statistic: 69.96 on 1 and 48 DF,  p-value: 6.236e-11
```

1) Les variances estimées des coefficients estimés (*Standard Error* dans le tableau) sont respectivement 4.872 et 0.868.
On peut remarquer que la variance de β0 > variance de β1, ainsi on peut dire que la dispersion autour de β0 est plus grande qu'autour de β1.
En termes de précision, on peut dire que la précision pour β1 est meilleure que la précision pour β0.

2) **Zero slope hypothesis test (*test de significativité)***

On fait l'hypothèse que la pente est nulle.

```
> oz.regsimple2 <- lm(data$maxO3v~data$Ne12)
> summary(oz.regsimple2)

Call:
lm(formula = data$maxO3v ~ data$Ne12)

Residuals:
    Min      1Q  Median      3Q     Max
-45.004 -13.964   0.141  10.401  46.626

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  110.435      6.764  16.326  < 2e-16 ***
data$Ne12     -5.210      1.205  -4.323 7.73e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.39 on 48 degrees of freedom
Multiple R-squared:  0.2802,    Adjusted R-squared:  0.2652
F-statistic: 18.69 on 1 and 48 DF,  p-value: 7.727e-05
```

On suppose H0 : β1 = 0
Et l'hypothèse adverse H1 : β1 différent de 0.

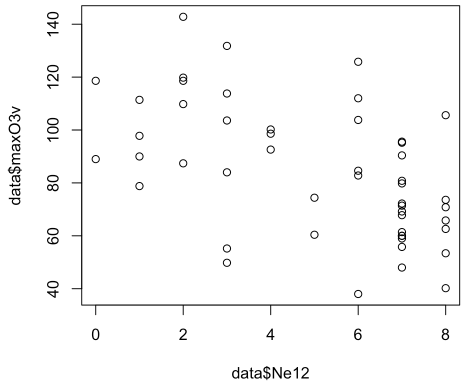
On calcule les intervalles de confiance au seuil de 5% pour l'estimateur β1.
On remarque que :

- -5,210, la valeur estimée du coefficient β1 associée à la variable Ne12 n'appartient pas à L'IC de β1 = [-8,39 ; -6,13], donc on valide l'hypothèse H0 : β1 = 0, ainsi on valide le fait que la pente soit nulle pour cette variable, et que les deux variables Ne12 et maxO3v sont indépendantes (déjà remarqué dans le plot du premier exercice)

3)

```
> summary(oz.regsimple2)$r.squared
[1] 0.2802236
```

Les variations de la variable maxO3v sont expliquées à environ 28% par la variable Ne12.



Cela confirme ce qu'on avait signalé dans **l'exercice 1**, à savoir qu'il n'y a pas de forte corrélation entre maxO3v et Ne12, contrairement à la corrélation qu'on avait montrée entre maxO3 et Ne12.
Cela est tout à fait normal dans la réalité, étant donné que la nébulosité n'a pas d'impact réel sur la quantité d'ozone du jour précédent.

Le modèle n'est donc pas de bonne qualité dans ce cas précis.

Exercice 6

```
> oz.regmult<-lm(data$maxO3 ~ data$Ne12 + data$maxO3v, data=data)
> summary(oz.regmult)

Call:
lm(formula = data$maxO3 ~ data$Ne12 + data$maxO3v, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-29.602  -8.109  -0.038   9.295  28.888

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 85.02028   11.09425   7.663 8.13e-10 ***
data$Ne12   -5.48007    0.91018  -6.021 2.50e-07 ***
data$maxO3v  0.34160    0.09248   3.694 0.000575 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.7 on 47 degrees of freedom
Multiple R-squared:  0.6846,    Adjusted R-squared:  0.6712
F-statistic: 51.02 on 2 and 47 DF,  p-value: 1.668e-12
```

Estimates :

- Intercept : la valeur de maxO3, quand Ne12 et maxO3v sont nuls, c'est-à-dire β0.
- Data$Ne12 : Indique de combien augmente en moyenne la teneur maximale en ozone lorsque la nébuleuse à 12h augmente de 1 unité, et sachant que toutes les autres variables sont fixes (maxO3v). Cela correspond à β1 et à la pente par rapport à l'axe maxO3v seul.
- Data$maxO3v : Indique de combien augmente en moyenne la teneur maximale en ozone lorsque la quantité maximale d'ozone de la veille augmente de 1 unité, et sachant que toutes les autres variables sont fixes (Ne12). Cela correspond à β2 et à la pente par rapport à l'axe Ne12 seul.

Dans notre cas, les estimations sont :
Intercept = 85.020 : ce coefficient n'a pas d'importance en général.
data$Ne12 = -5.480 : si on augmente Ne12 d'une unité, maxO3 diminue d'environ 5,5 unités.
data$maxO3v = 0.341 : sur cet axe, si on augmente maxO3v d'une unité, maxO3 augmente d'environ 0,3 unité (ce qui est normal, parce que maxO3v n'a pas beaucoup d'influence sur maxO3).

```
> confint(oz.regmult, level = 0.80)
                  10 %       90 %
(Intercept) 70.5996916 99.4408670
data$Ne12   -6.6631555 -4.2969939
data$max03v  0.2213915  0.4618001
```

On suppose H0 : β1 et β2 = 0
Et l'hypothèse adverse H1 : β1 et β2 différents de 0.

On calcule les intervalles de confiance au seuil de 5% pour nos 2 estimateurs β1 et β2. On remarque que :

- -5,480 la valeur estimée du coefficient β1 associée à la variable Ne12 appartient à L'IC de β1 = [-6,663 ; -4,296], donc on rejette l'hypothèse H0 : β1 = 0, ainsi on rejette le fait que la pente soit nulle pour cette variable.
- Pour β2, son coefficient associé pour la variable maxO3v est 0,34160, qui appartient à l'IC de β2 = [0,221, 0,461]. Comme précédemment, on rejette l'hypothèse H0 : = 0, ainsi on rejette le fait que la pente soit nulle pour cette variable.

On rejette donc l'hypothèse pour le modèle.

Exercice 7

Voici les prédictions pour les valeurs demandées :

```
> oz.regsimple <- lm(maxO3~Ne12, data = data)
> new.dataS <- data.frame(Ne12=6)
> predict.lm(oz.regsimple, new.dataS)
       1
79.18537
```

```
> oz.regmult = lm(maxO3 ~ Ne12 + maxO3v, data = data)
> new.dataM <- data.frame(Ne12=6, maxO3v=80)
> predict.lm(oz.regmult, new.dataM)
       1
79.4675
```

Étant donné que la moyenne de la variable maxO3 est de 86.3 (voir *summary(data)*), on remarque que la valeur obtenue prédite par la régression multiple est plus proche que celle prédite par la régression simple.

Exercice 8

1) Les variances par rapport au modèle.

R^2 (***R-squared***) mesure la proximité des données par rapport à la droite de régression calculée, c'est ce qu'on appelle coefficient de détermination (ou coefficient de détermination pour les régressions multiples).
Ce coefficient est compris entre 0 et 100%.
S'il est proche de 0, le modèle n'explique aucune variabilité des données de sa moyenne. S'il est plus élevé, cela indique que le modèle explique toute la variabilité des données. Ainsi, plus R^2 est élevé, et plus le modèle s'ajuste bien aux données.
En outre, $R^2 = 1 - \frac{SSE}{SST}$

$\bar{R}$**^2 (*Adjusted R-squared)*** permet de corriger R^2 en prenant en compte le nombre de variables utilisées dans le modèle. La formule devient donc : $R^2 = 1 - \frac{SSE\ /(n-p-1)}{SST/(n-1)}$

Selon moi, étant donné que $\bar{R}$**^2** tient compte du nombre de paramètres, je pense qu'il est plus contraignant. Ainsi, R^2 lui est préférable.
(J'ai pris connaissance de résultats de la question suivante pour confirmer cette affirmation).

2) Pour oz.regsimple

```
Residual standard error: 15.4 on 48 degrees of freedom
Multiple R-squared:  0.5931,    Adjusted R-squared:  0.5846
F-statistic: 69.96 on 1 and 48 DF,  p-value: 6.236e-11
```

Pour oz.regmult

```
Residual standard error: 13.7 on 47 degrees of freedom
Multiple R-squared:  0.6846,    Adjusted R-squared:  0.6712
F-statistic: 51.02 on 2 and 47 DF,  p-value: 1.668e-12
```

Les coefficients de détermination et de détermination ajustés sont de :

- 0.593 et 0.584 pour la régression simple,
- 0.684 et 0.671 pour la régression multiple

On peut remarquer que la régression multiple en intégrant une variable supplémentaire est plus performante que dans le cas de la régression simple. C'est tout à fait normal, parce qu'on rajoute d'autres critères qui caractérisent mieux la variable expliquée. Ceci est aussi valable pour les variances ajustées.

On peut aussi remarquer que les variances ajustées sont toujours inférieures dans les deux cas (régression simple et multiple). Ceci est aussi normal car ce coefficient est ajusté par des estimations tenant compte du nombre de paramètres du modèle, constituant une contrainte supplémentaire.

Le modèle multiple est donc préférable au modèle simple, car R^2 et $\bar{R}$^2 du modèle multiple sont plus élevés que ceux du modèle simple.