Régression linéaire multiple (suite)

Semaine 10

Exercice 2 : Application

Introduction

L'Institut de Recherche en Glaciologie Balnéaire d'Apualculco a conduit une étude statistique afin d'examiner les effets de l'exposition au rayonnement solaire (`solar_exposure`) sur deux variables :

- la consommation de desserts glacés (`icecream_consumption`), mesurée en centilitres par personne,
- et le degré de brûlure superficielle de la peau (`sunburn_grade`).

Les données issues de cette étude ont été préalablement préparées et peuvent être chargées en mémoire avec la commande suivante :

```
load("donnees_exercice_2.rda")
```

Question 1

Représenter le nuage de points illustrant la relation entre la consommation de crème glacée (`icecream_consumption`) et le degré de brûlures superficielles de la peau (`sunburn_grade`).

Ajuster ensuite un modèle de régression linéaire simple de la forme :

$$\text{sunburn grade} = a_0 + a_1 \times \text{icecream consumption} + \epsilon\,.$$

Superposer au graphique la droite de régression ajustée et ajouter un encadré indiquant le coefficient de détermination, exprimé en pourcentage et arrondi à l'unité.

Solution. Compléter le code R suivant.

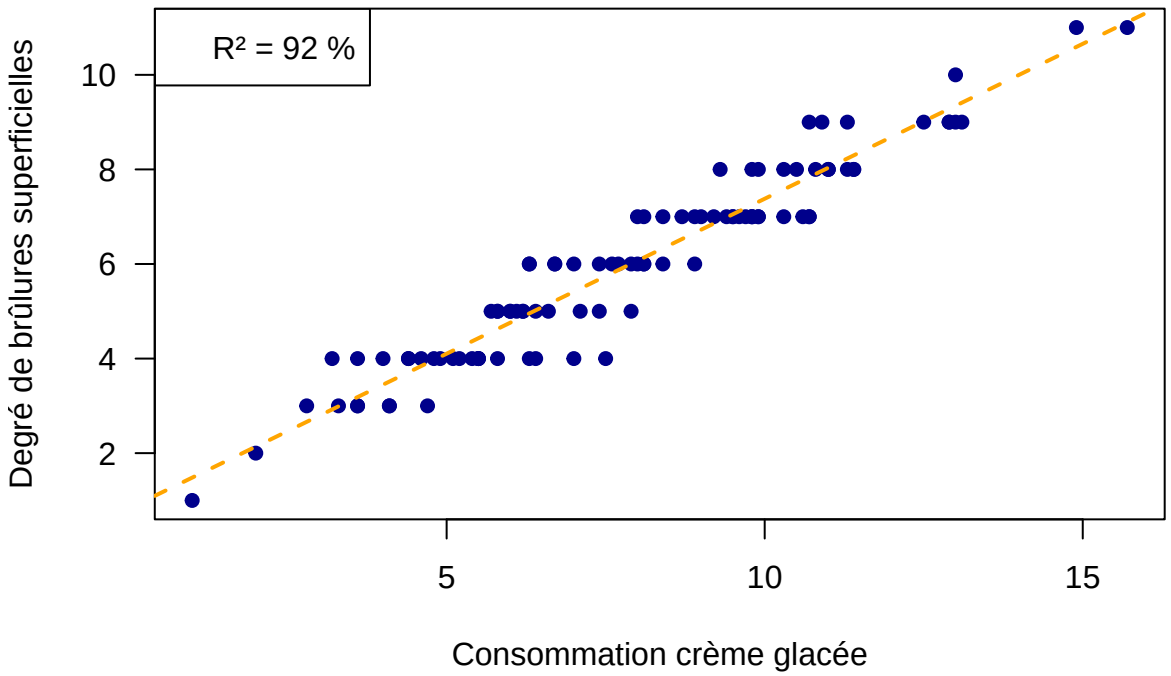
```
# Remplacer eval = FALSE par eval = TRUE

# tracé du graphique sur lequel on va superposer la droite de régression ajustée
plot(icecream_consumption, sunburn_grade,
     las = 1, pch = 19,
     cex=0.9,
     col = "darkblue",
     xlab = "Consommation crème glacée",
     ylab = "Degré de brûlures superficielles")

# Ajuster un modèle linéaire simple
mod_sunburn_simple <- lm(sunburn_grade ~ icecream_consumption)

# Calculer le cofficient de détermination
resume_glaces=summary(mod_sunburn_simple)
det=round(resume_glaces$r.squared,2)*100
# On écrit la légende
legend(x = "topleft", legend = c(paste("R² =", det,"%")))
```

```r
# Tracer la droite de régression ajustée en orange avec des pointillés
abline(a=resume_glaces$coefficients[1,1], b=resume_glaces$coefficients[2,1], col = "orange", lty = 2, lw
```



Question 2

Pour le modèle de la question 1, décrire de manière scientifiquement rigoureuse la relation entre le degré de brûlures superficielles de l'épiderme et la consommation de crème glacée.

Pour cela, on procèdera en trois étapes :

1. Estimer le coefficient de corrélation entre les deux variables, en précisant son intervalle de confiance à 95 % (`cor.test`).
2. Tester la significativité de ce coefficient (hypothèse nulle : absence de corrélation) en indiquant la valeur de la statistique de test, le nom exact du test utilisé (y compris les degrés de liberté), ainsi que la p-valeur obtenue (ou son logarithme décimal).
3. Résumer les résultats en deux phrases, en présentant des valeurs numériques arrondies.
4. Répondre aux items précédents en remplaçant le terme *coefficient de corrélation* par le terme *coefficient de régression*. Utiliser uniquement `lm` pour déterminer l'intervalle de confiance et le test. Qu'est-ce qui est identique ou différent dans les deux analyses ?

Réponse. Pour savoir comment présenter correctement un résultat d'estimation statistique dans un rapport scientifique, vous pouvez poser la question suivante à un moteur de recherche ou à un modèle de langage :

"Comment reporter un résultat statistique dans un article scientifique ?"

Les éléments de réponse que vous trouverez seront utiles pour l'ensemble de vos travaux futurs.

Attention : le point 3 de la question est le plus important - c'est celui qui sera lu, compris et évalué. Il détermine la qualité de la communication scientifique de vos résultats.

Exemple de rédaction attendue : L'analyse de corrélation met en évidence une association **la décrire comme positive/negative forte/modérée/faible** entre la consommation de crème glacée et des brûlures superficielles de l'épiderme (r = **compléter** ; IC à 95 % : **compléter**). Cette relation est statistiquement significative (-log10(p) = **compléter**, correspondant à **statistique** = **compléter** ; test **nom du test**).

```r
cor.test(sunburn_grade,icecream_consumption)
```

```
##
##  Pearson's product-moment correlation
##
## data:  sunburn_grade and icecream_consumption
## t = 33.535, df = 96, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9405908 0.9729781
## sample estimates:
##       cor
## 0.9598687
```

```r
mod=lm(sunburn_grade ~ icecream_consumption)
resume=summary(mod)
resume
```

```
##
## Call:
## lm(formula = sunburn_grade ~ icecream_consumption)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74124 -0.31388 -0.03611  0.34067  1.16190
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.82672    0.16465   5.021 2.37e-06 ***
## icecream_consumption  0.65527    0.01954  33.535  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5696 on 96 degrees of freedom
## Multiple R-squared:  0.9213, Adjusted R-squared:  0.9205
## F-statistic:  1125 on 1 and 96 DF,  p-value: < 2.2e-16
```

```r
# Coefficient de régression (pente)
coef(mod)[2]
```

```
## icecream_consumption
##            0.6552689
```

```r
# Intervalle de confiance à 95 %
confint(mod)[2, ]
```

```
##     2.5 %    97.5 %
## 0.6164821 0.6940558
```

```r
# calcul de la p-valeur
t_obs <- resume$coefficients["icecream_consumption", "t value"]
# Degré de liberté
df=df.residual(mod)
p_valeur <- 2 * pt(-abs(t_obs), df)
p_valeur
```

```
## [1] 8.340592e-55
```

1) Le coefficient de corrélation entre les deux variables est de environ 0.96 (arondi au centième). Son intervalle de confiance à 95% est [0.94,0.97] (à nouveau arondi au centième).
2) Pour tester la significativité du coefficient, on effectue un test de Fisher (qui donne la même p valeur que celui de student puisque on se trouve dans une régression linéaire simple et non multiple). La statistique de test a une valeur de 1125 et des degrés de liberté qui sont 1 et 96. La p valeur est de $8.34 * 10^{-55}$environ donc inférieure à 0.05. Ainsi, on rejette l'hypothèse H0 : $\rho^2 = 0$
3) L'analyse de corrélation met en évidence une association positive et forte entre la consommation de crème glacée et des brûlures superficielles de l'épiderme (R = 0.96 ; IC à 95 % : [0.94,0.97]). Cette relation est statistiquement significative (-log10(p) = 54, correspondant à F = 1125 ; test de Fisher).
4) Le coefficient de régression entre les deux variables est de environ 0.66 (arondi au centième). Son intervalle de confiance à 95% est [0.62,0.69] (à nouveau arondi au centième). Pour tester la significativité du coefficient, on effectue un test de Student. La statistique de test a une valeur de 33.54 (arrondie au centième) et de degré de liberté 96.La p valeur est également de $8.34 * 10^{-55}$ environ donc inférieure à 0.05. Ainsi, on rejette l'hypothèse H0 : $b1 = 0$. L'analyse du coefficient de régression met en évidence une association positive et forte entre la consommation de crème glacée et des brûlures superficielles de l'épiderme (b1 = 0.66 ; IC à 95 % : [0.62,0.69]). Cette relation est statistiquement significative (-log10(p) = 54, correspondant à t=33.56 ; test de Student).

Question 3

Peut-on conclure, à partir de cette relation, que la consommation de crème glacée cause des brûlures superficielles de l'épiderme ? Si non, quelle explication alternative ou quel facteur de confusion pourrait expliquer l'association observée ?

Réponse. Facile ? Encore faudra-il prouver ce que l'on dit dans la suite.

On ne peut pas conclure à partir de cette relation que la consommation de crème glacée cause des brûlures superficielles de l'épiderme. Corrélation ne signifie pas forcément causalité. On peut expliquer l'association observée par le fait que les gens qui mangent de la crème glacée sont au bord de la mer et sont donc beaucoup exposés au soleil et ainsi sont plus victimes de brulures de la peau.

Question 4

Justifier pourquoi l'exposition au soleil (`solar_exposure`) peut être considérée comme un facteur de confusion dans la relation entre la consommation de crème glacée (`icecream_consumption`) et le degré de brûlures superficielles de l'épiderme (`sunburn_grade`).

1. Vérifier s'il existe une association entre la consommation de crème glacée et l'exposition au soleil.
2. Vérifier s'il existe une association entre l'exposition au soleil et le degré de brûlures superficielles.
3. Rédiger une brève conclusion : ces résultats suggèrent-ils un biais d'estimation de l'effet de la consommation de crème glacée sur le degré de brûlures superficielles ?

Réponse.

```
cat("Association entre la consommation de crème glacée et l'exposition au soleil : \n")
```

```
## Association entre la consommation de crème glacée et l'exposition au soleil :
```

```
cor.test(icecream_consumption,solar_exposure)
```

```
##
##  Pearson's product-moment correlation
##
## data:  icecream_consumption and solar_exposure
## t = 49.796, df = 96, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  0.9720031 0.9873771
## sample estimates:
##       cor
## 0.9811867
```

```
cat("Association entre le degré de brûlures superficielles et l'exposition au soleil : \n")
```

```
## Association entre le degré de brûlures superficielles et l'exposition au soleil :
```

```
cor.test(sunburn_grade,solar_exposure)
```

```
##
##  Pearson's product-moment correlation
##
## data:  sunburn_grade and solar_exposure
## t = 39.659, df = 96, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9566733 0.9803817
## sample estimates:
##      cor
## 0.970811
```

```
# Régression linéaire simple
mod_c_exp <- lm(icecream_consumption ~ solar_exposure)
resume_c_exp <- summary(mod_c_exp)
resume_c_exp
```

```
##
## Call:
## lm(formula = icecream_consumption ~ solar_exposure)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0905 -0.3890  0.0132  0.4166  1.4116
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.99933    0.24581  -16.27   <2e-16 ***
## solar_exposure  2.99649    0.06018   49.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5744 on 96 degrees of freedom
## Multiple R-squared:  0.9627, Adjusted R-squared:  0.9623
## F-statistic:  2480 on 1 and 96 DF,  p-value: < 2.2e-16
```

```
# Calcul de la vraie p-valeur correspond à la corrélation entre l'exposition au soleil
# et la consommation de crème glacée.
t_obs_ice <- resume_c_exp$coefficients["solar_exposure", "t value"]
# Degré de liberté
df <- df.residual(mod_c_exp)
p_value_ice <- 2 * pt(-abs(t_obs_ice), df)
p_value_ice
```

```
## [1] 2.210594e-70
```

```r
# Calcul de la vraie p-valeur correspond à la corrélation entre l'exposition au soleil
# et le degré de brûlures superficielles
mod_b_exp <- lm(sunburn_grade ~ solar_exposure)
resume_b_exp <- summary(mod_b_exp)

t_obs_burn <- resume_b_exp$coefficients["solar_exposure", "t value"]
df_burn <- df.residual(mod_b_exp)
p_value_burn <- 2 * pt(-abs(t_obs_burn), df_burn)
p_value_burn
```

```
## [1] 2.48752e-61
```

Le coefficient de corrélation entre l'exposition au soleil et la consommation de crème glacée est de environ 0.98 (arondi au centième). Son intervalle de confiance à 95% est [0.97,0.99] (à nouveau arondi au centième). Pour tester la significativité du coefficient, on effectue un test de Fisher (qui donne la même p valeur que celui de student puisque on se trouve dans une régression linéaire simple et non multiple). La statistique de test a une valeur de 2480 et des degrés de liberté qui sont 1 et 96. La p-valeur est inférieure à $2.21 * 10^{-70}$ donc inférieure à 0.05. Ainsi, on rejette l'hypothèse H0 : $\rho^2 = 0$. L'analyse de corrélation met en évidence une association positive et forte entre la consommation de crème glacée et l'exposition au soleil.

Le coefficient de corrélation entre l'exposition au soleil et le degré de brûlures superficielles est de environ 0.97 (arondi au centième). Son intervalle de confiance à 95% est [0.96,0.98] (à nouveau arondi au centième). Pour tester la significativité du coefficient, on effectue un test de Fisher (qui donne la même p valeur que celui de student puisque on se trouve dans une régression linéaire simple et non multiple). La statistique de test a une valeur de 1573 et des degrés de liberté qui sont 1 et 96. La p-valeur est inférieure à $2.48 * 10^{-61}$ donc inférieure à 0.05. Ainsi, on rejette l'hypothèse H0 : $\rho^2 = 0$. L'analyse de corrélation met en évidence une association positive et forte entre le degré de brûlures superficielles et l'exposition au soleil.

Question 5

Ajuster un modèle de régression linéaire multiple expliquant le degré de brûlures superficielles de l'épiderme (`sunburn_grade`) à partir de deux variables explicatives : la consommation de crème glacée (`icecream_consumption`) et l'exposition au soleil (`solar_exposure`).

1. Estimer l'effet de la consommation de crème glacée, après ajustement pour l'exposition au soleil, et tester la significativité de cet effet.
2. Présenter une conclusion synthétique de l'ensemble des résultats obtenus dans cet exercice, en précisant ce qu'ils suggèrent sur la nature des relations entre les variables étudiées.

Réponse. On peut utiliser le code suivant.

```r
# Remplacer eval = TRUE par eval = TRUE
mod_complet <- lm(formula = sunburn_grade ~ icecream_consumption + solar_exposure)
summary(mod_complet)
```

```
##
## Call:
## lm(formula = sunburn_grade ~ icecream_consumption + solar_exposure)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67227 -0.27625 -0.00162  0.29606  1.04482
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -1.49759    0.40111  -3.734 0.000322 ***
```

```
## icecream_consumption  0.13410    0.08592   1.561 0.121891
## solar_exposure        1.62213    0.26239   6.182 1.58e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4835 on 95 degrees of freedom
## Multiple R-squared:  0.9439, Adjusted R-squared:  0.9427
## F-statistic: 799.4 on 2 and 95 DF,  p-value: < 2.2e-16
```

Le coefficient de régression associé à la consommation de crème glacée dans la régression linéaire multiple, qui vise à expliquer le degré de brûlures de la peau, est d'environ 0.13, contre environ 1.6 pour celui associé à l'exposition au soleil. Pour tester la significativité de ce coefficient, on effectue un test de Student. La p-valeur associée au coefficient de régression de la consommation de crème glacée est d'environ 0.12, donc supérieure à 0.05. Ainsi, on ne rejette pas l'hypothèse nulle H0 : $coeff = 0$.

Ainsi, même si la corrélation entre les brulures et la consommation de crème glacée est haute cela ne signifie pas qu'il y a corrélation. En effet, en mettant une autre variable en jeu, l'exposition au soleil, la consommation de crème n'est plus du tout significative. On peut supposer que le lien de causalité existe plutôt entre exposition au soleil et brûlures de la peau mais pour bien le valider il faudrait faire de multiples expériences dans des contextes différents.

Exercice 3

Introduction

On considère le modèle de régression linéaire multiple suivant :

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \epsilon \,,$$

où les prédicteurs X_1 et X_2 sont des variables aléatoires indépendantes du terme d'erreur ϵ.

On suppose en outre que les variables aléatoires X_1 et X_2 **sont non-corrélées**.

Question 1

Démontrer que la part de variance de Y expliquée par $X = (X_1, X_2)$, définie par

$$\rho^2 = \frac{\mathrm{Var}(\mathbb{E}[Y \mid X])}{\mathrm{Var}(Y)}$$

s'écrit, sous les hypothèses du modèle précédent

$$\rho^2 = \rho_1^2 + \rho_2^2 \,,$$

où $\rho_1 = \mathrm{Cor}(X_1, Y)$ désigne le coefficient de corrélation linéaire entre X_1 et Y, et $\rho_2 = \mathrm{Cor}(X_2, Y)$ le coefficient de corrélation linéaire entre X_2 et Y.

Solution. Commencer par exprimer l'espérance conditionnelle sous la forme suivante

$$\mathbb{E}[Y \mid X] = b_0 + b_1 X_1 + b_2 X_2 \,.$$

Remplacer ensuite les coefficients b_1 et b_2 par leurs valeurs théoriques déterminées lors de la semaine 9 (CM et CTD), puis simplifier l'expression obtenue.

On considère le modèle :

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \epsilon$$

avec $\mathbb{E}[\epsilon] = 0$ et X_1, X_2 non-corrélés et indépendants de ϵ.

L'espérance conditionnelle est

$$\mathbb{E}[Y \mid X_1, X_2] = b_0 + b_1 X_1 + b_2 X_2$$

et sa variance

$$\mathrm{Var}(\mathbb{E}[Y \mid X_1, X_2]) = \mathrm{Var}(b_1 X_1 + b_2 X_2) = b_1^2 \mathrm{Var}(X_1) + b_2^2 \mathrm{Var}(X_2)$$

car X_1 et X_2 sont non-corrélés.

Les coefficients théoriques de régression sont

$$b_1 = \frac{\mathrm{Cov}(X_1, Y)}{\mathrm{Var}(X_1)}, \quad b_2 = \frac{\mathrm{Cov}(X_2, Y)}{\mathrm{Var}(X_2)}.$$

On a donc

$$b_1^2 \mathrm{Var}(X_1) = \frac{\mathrm{Cov}(X_1, Y)^2}{\mathrm{Var}(X_1)}.$$

Or, par définition du coefficient de corrélation linéaire :

$$\rho_1 = \frac{\mathrm{Cov}(X_1, Y)}{\sqrt{\mathrm{Var}(X_1)\mathrm{Var}(Y)}} \quad \Rightarrow \quad \mathrm{Cov}(X_1, Y)^2 = \rho_1^2 \mathrm{Var}(X_1)\mathrm{Var}(Y).$$

En remplaçant, on obtient

$$b_1^2 \mathrm{Var}(X_1) = \frac{\rho_1^2 \mathrm{Var}(X_1)\mathrm{Var}(Y)}{\mathrm{Var}(X_1)} = \rho_1^2 \mathrm{Var}(Y).$$

De même :

$$b_2^2 \mathrm{Var}(X_2) = \rho_2^2 \mathrm{Var}(Y).$$

Ainsi, la part de variance expliquée par $X = (X_1, X_2)$ est

$$\rho^2 = \frac{\mathrm{Var}(\mathbb{E}[Y \mid X_1, X_2])}{\mathrm{Var}(Y)} = \frac{b_1^2 \mathrm{Var}(X_1) + b_2^2 \mathrm{Var}(X_2)}{\mathrm{Var}(Y)} = \rho_1^2 + \rho_2^2.$$

Question 2

On considère un échantillon de taille $n = 1000$, contenant les variables `y`, `x_1` et `x_2`, chargées en mémoire de la manière suivante :

```
load("donnees_exercice_3.rda")
```

Ajuster un modèle linéaire aux données de l'exercice, en considérant `y` comme variable réponse et `x_1` et `x_2` comme variables explicatives :

1. Vérifier si les variables explicatives sont corrélées.
2. Calculer le coefficient de détermination empirique, d'abord à partir de la formule de la question 1, puis selon sa définition à partir des sommes RSS et TSS.
3. Expliquer pourquoi les résultats peuvent légèrement différer.
4. Vérifier vos calculs à l'aide du résumé du modèle (`summary`).

Solution. Ajuster un modèle linéaire aux données de l'exercice de la manière suivante

```
# Remplacer eval = TRUE par eval = TRUE
# Ajuster un modèle linéaire aux données
mod = lm(formula = y ~ x_1 + x_2)
summary(mod)
```

```
##
## Call:
## lm(formula = y ~ x_1 + x_2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.07821 -0.67536  0.03575  0.67866  2.77055
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.16312    0.03089   37.66   <2e-16 ***
## x_1          1.98275    0.03048   65.06   <2e-16 ***
## x_2         -2.88310    0.02988  -96.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9767 on 997 degrees of freedom
## Multiple R-squared:  0.9325, Adjusted R-squared:  0.9323
## F-statistic:  6883 on 2 and 997 DF,  p-value: < 2.2e-16
```

Répondre à l'ensemble des questions posées.

```
#Vérification si les variables explicatives sont corrélées
cor.test(x_1,x_2)
```

```
##
##  Pearson's product-moment correlation
##
## data:  x_1 and x_2
## t = -0.55665, df = 998, p-value = 0.5779
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.07952402  0.04442381
## sample estimates:
##        cor
## -0.0176178
```

```
#Méthode 1 : Calcul de coefficient de détermination empirique en utilisant
#la formule de la question 1
p1=cor(x_1,y)
p2=cor(x_2,y)
p_empi=p1^2+p2^2
cat("Le  coefficient de détermination empirique en utilisant
    la formule de la question 1: R2=",p_empi,"\n")
```

```
## Le  coefficient de détermination empirique en utilisant
##     la formule de la question 1: R2= 0.9477398
```

```
#Méthode 2 : Calcul de R2 à partir des sommes RSS et TSS
tss=sum((y-mean(y))^2)
b0=mod$coefficients[1]
b1=mod$coefficients[2]
b2=mod$coefficients[3]
rss=sum((y-b0-b1*x_1-b2*x_2)^2)
R2=1-(rss/tss)
cat("Le coefficient de détermination selon sa définition
    à partir des sommes RSS et TSS: R2=",R2,"\n")
```

```
## Le coefficient de détermination selon sa définition
##     à partir des sommes RSS et TSS: R2= 0.9324695
```

```
cat("Vérification à l'aide du résumé du modèle: R2=",summary(mod)$r.squared,"\n")
```

```
## Vérification à l'aide du résumé du modèle: R2= 0.9324695
```

1.À l'aide du test de corrélation de Pearson (cor.test). On obtient une p-valeur supérieure à 0.05, donc on ne rejette pas l'hypothèse nulle selon laquelle $cor(X_1, X_2) = 0$. On ne peut toutefois pas exclure que les variables explicatives X_1 et X_2 soient non corrélées.

4.Deux méthodes différentes peuvent donner des résultats d'estimation différents. De plus, cette méthode nécessite que les variables explicatives soient parfaitement non corrélées. Or ce n'est pas exactement le cas, puisque $cor(X_1, X_2) = -0.0176$. On peut donc dire que c'est pour cette raison que les résultats obtenus diffèrent légèrement.