EISTI -Département Mathématiques
Corrigé de l'Examen Modèle Linéaire : MI
Décembre 2018- durée 2h00

La consultation et l'échange de documents et de calculatrices sont interdits. Deux feuilles manuscrites recto-verso format A4 et les calculatrices sont autorisées.

Exercice 1. Régression logistique [14 pt].

L'objectif de cet exercice est de traiter un problème de défaut bancaire à partir d'un jeu de données composé d'un échantillon de taille 10000 et **trois variables explicatives student, balance et income**.

```
              Table 1
default student   balance    income
 0       0        729.5265   44361.625
 0       1        817.1804   12106.135
 0       0        1073.5492  31767.139
 0       0        529.2506   35704.494
 .       .        ........   ........
 .       .        ........   ........
```

Nous cherchons à déterminer quels clients seront en défaut sur leur dette de carte de crédit.

- default : Yes (ou 1) si le client fait défaut sur sa dette et No (ou 0) sinon.
- student : Yes (ou 1) si le client est un étudiant et No (ou 0) sinon
- balance : montant moyen mensuel d'utilisation de la carte de crédit
- income : revenu du client.

1) Modèle 1. Nous avons ajusté un modèle logistique simple où on cherche à expliquer **default** en fonction de **balance**. Si dessous les résultats obtenus:

```
> summary(Modele1)
Call:
glm(formula = default ~ balance, family = binomial(link = "logit"),
data = Default)
Coefficients:
             Estimate    Std.Error   z value   Pr(>|z|)
(Intercept)  -1.065e+01  3.612e-01    -29.49    2e-16 ***
balance       5.499e-03  2.204e-04    24.95     2e-16 ***

Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1596.5  on 9998  degrees of freedom
AIC: 1600.5
```

(a) Donner l'équation du modèle logistique ajusté en notant a la constante et b la pente.

(b) Calculez l'odds en fonction de la balance. En déduire l'odds pour une balance de 1000. Qu'est-ce cela signifie ?

(c) Comment interprétez-vous le coefficient $\widehat{b}$?

(d) Écrire les matrices X, matrice du modèle, et V, matrice des variances, permettant d'estimer les variances des coefficients. On écrira les premières lignes correspondant à l'extrait du jeu de données de Table 1.

(e) Calculez un intervalle de confiance pour l'odds-ratio.

(f) Effectuez un test pour déterminer si b est significativement non nul.

(g) Nous avons relevé les valeurs estimées de la proportion de **default** pour une balance de 1000 et de 2000. À quelle classe appartiennent xnew=1000 et xnew=2000 ? Justifier votre réponse.

```
> xnew=data.frame(balance=c(1000,2000))
> xnew
balance
1    1000
2    2000
> predict.glm(Modele1,xnew,type="response")
1           2
0.005752145 0.585769370
```

2) Modèle 2. Nous cherchons ici à expliquer **default** en fonction de **student**

```
> summary(Modele2)
Call:
glm(formula = default ~ student, family = binomial, data = Default)
Coefficients:
             Estimate   Std.Error   z value  Pr(>|z|)
(Intercept)  -3.50413    0.07071   -49.55    2e-16 ***
student       0.40489    0.11502    3.52     0.000431 ***

Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 2908.7  on 9998  degrees of freedom
AIC: 2912.7
```

a) Donner l'équation du modèle logistique ajusté en notant a la constante et b la pente.

b) Estimer P(default=Yes / student = Yes) et P (default=Yes / student = Non)

c) Calculez l'odds des étudiants et l'odds des autres clients.

d) Comment interprétez-vous le coefficient $\widehat{b}$?

3) Comment comparer les deux modèles ?

Exercice 2. Régression linéaire multiple: Théorie [6 pt]

1ère Partie. On considère le modèle de régression linéaire multiple: $Y = X\beta + \epsilon$ où le vecteur $Y = (y_1, \ldots, y_n)^t \in \mathbb{R}^n$ représente la variable à expliquer, X est une matrice réelle de taille $n \times (p+1)$ de rang $p+1$, $\beta = (\beta_1, \ldots, \beta_p)^t \in \mathbb{R}^p$. Le vecteur $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^t \in \mathbb{R}^n$ a pour variance

$$\text{Var}(\epsilon) = \sigma^2 \text{diag}(w_1, \ldots, w_n) = \sigma^2 W^{-1} \quad \text{où}$$

$$W = \begin{pmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_n \end{pmatrix} = \text{diag}(w_1, \ldots, w_n).$$

est une matrice de poids (matrice symétrique de taille $n \times n$). Soit

$$S(\beta) = \sum_{i=1}^{n} w_i (Y_i - x_i^t \beta)^2 = (Y - X\beta)^t W (Y - X\beta)$$

(où A^t désigne le transposé de A).

a) Démontrer en dérivant $S(\beta)$, par rapport à β, que les équations normales à résoudre pour obtenir l'estimateur des moindres carrés pondérés de β sont:

$$(X^t W X)\widehat{\beta} = X^t W y \tag{1}$$

puis en déduire cet estimateur.

NB. On pourra utiliser le théorème suivant. Soit u un vecteur dans $\mathbb{R}^n$ et A une matrice carrée. Si $f(u)$ est une fonction quelconque du vecteur u, alors

$$\frac{d}{du} f(u)^t A f(u) = 2 \left(\frac{d}{du} f(u) \right)^t A f(u) \tag{2}$$

b) Démontrer que

$$Var(\widehat{\beta}) \quad = \quad \sigma^2 (X^t W X)^{-1} \tag{3}$$

On rappel que: $Var(y) = Var(\epsilon)$.

[On pourra utiliser : $Var[AX] = AVar[X]A^t$]

2ème Partie. Dans cette question on considère le modèle de régression linéaire simple passant par l'origine (modèle sans constante)

$$Y_i = \beta X_i + \epsilon_i, \quad 1 \leq i \leq n, \quad \beta \in \mathbb{R}. \tag{4}$$

a) Dans ce cas précis, quand on écrit

$$Y = \beta X + \epsilon$$

à quoi correspond la matrice X ?

b) Déduire $\widehat{\beta}$ de (1).

c) Déduire $Var(\widehat{\beta})$ de (3).

En déduire $Var(\widehat{\beta})$ dans les cas suivants:

d) $Var(\epsilon_i) = \sigma^2$.

e) $Var(\epsilon_i) = \sigma^2 X_i$.

f) $Var(\epsilon_i) = \sigma^2 X_i^2$.

Exercice 3. Régression Linéaire Simple [11pt]. On dispose des données issues du rapport publié par l'OMS en février 2011 sur la consommation d'alcool dans le monde en projection pour l'année 2008 et de l'espérance de vie à la naissance en 2009 pour 188 pays. On a les résultats numériques suivants:

$$\sum_{i=1}^{188} x_i = 1250.77, \quad \sum_{i=1}^{188} x_i^2 = 12699.04, \quad \sum_{i=1}^{188} x_i y_i = 88858.02, \quad \sum_{i=1}^{188} y_i = 12935, \quad \sum_{i=1}^{188} y_i^2 = 907647.$$

On considère le modèle de régression linéaire simple avec constante

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad 1 \leq i \leq n, \quad \epsilon_i \hookrightarrow \mathcal{N}(0, \sigma^2) \tag{5}$$

(ϵ_i) est une suite de variables aléatoires indépendantes et de même loi.

a) Écrire le modèle sous forme matricielle. Donner les expressions des estimateurs des MCO $\widehat{\beta}_0$ et $\widehat{\beta}_1$. Donner les valeurs de ces estimateurs calculés sur les observations.

b) Écrire l'équation d'analyse de la variance et calculer le coefficient de détermination R^2.

c) Donner l'expression de l'estimateur sans biais $\widehat{\sigma^2}$ de σ^2. Calculer sa valeur sur les observations.

d) Déterminer les lois des variables aléatoires $\widehat{\beta}_0$ et $\widehat{\beta}_1$.

e) Donner les expressions des estimateurs $\widehat{\sigma}(\widehat{\beta}_0)$ et $\widehat{\sigma}(\widehat{\beta}_1)$ et des écart-types $\sigma(\widehat{\beta}_0)$ et $\sigma(\widehat{\beta}_1)$ de $\widehat{\beta}_0$ et $\widehat{\beta}_1$. Donner les valeurs de ces estimateurs calculés sur les observations.

f) Déterminer un intervalle de confiance à 95% pour β_1.

g) Tester l'hypothèse nulle $(H0)$: $\beta_1 = 0$ contre l'alternative (H1) : $\beta_1 \neq 0$ au niveau 5%. Commenter.

h) Pour un modèle de régression linéaire simple sans constante [Modèle (4), Exercice 2 avec $W = I$ (matrice identité)], déterminer les valeurs de l'estimateur des moindres carrés ordinaires du coefficient de régression et du coefficient de détermination calculés sur ces données.

i) Que constate-t-on ? Faut-il pour autant préférer le modèle de régression linéaire simple sans constante au modèle de régression linéaire simple avec constante ?

Rappel.

On note $\mathbf{1} = (1, 1, \ldots, 1)^t$ le vecteur colonnes formé que des 1.

La somme des carrés résiduelle : $SCR = ||\widehat{\epsilon}||^2$ (où $||\cdot||$ désigne la norme euclidienne).

A) Régression avec Constante.

- La somme des carrés totale : $SCT = ||Y - \overline{Y}\mathbf{1}||^2$.

-somme des carrés expliquée : $SCE = ||\widehat{Y} - \overline{Y}||^2$

B) Régression sans Constante.

- la somme des carrés totale: $SCT_{sc} = ||Y||^2$

- la somme des carrés expliquée : $SCE_{sc} = ||\widehat{Y}||^2$

Corrigé de l'exercice 1

Modèle 1.
a) **[1 pt]**. Le modèle logistique s'écrit:

$$logit(\widehat{\pi}(x)) = \widehat{a} + \widehat{b}x = -1,065 \times 10 + 5,499 \times 10^{-3}x, \quad \text{où} \quad x = \text{balance} \in \mathbb{R}^+.$$

Ce qui implique que

$$\widehat{\pi}(x) = \frac{e^{\widehat{a}+\widehat{b}x}}{1+e^{\widehat{a}+\widehat{b}x}} = \frac{1}{1+e^{-\widehat{a}-\widehat{b}x}} = \frac{1}{1+e^{10,65-0.005499x}}$$

b) **[1.5 pt]** Calcul de oddd(x):

$$odds(x) = \frac{\widehat{\pi}(x)}{1-\widehat{\pi}(x)} = e^{\widehat{a}+\widehat{b}x} = e^{-10,65+0.005499x}$$

et

$$odds(1000) = e^{-10,65+0.005499\times 1000} = e^{-5.151} = 0.005793608 = 5.793608 \times 10^{-3}$$

$$5.793608 \times 10^{-3} = \frac{\mathbb{P}(Y=1|X=1000)}{\mathbb{P}(Y=0|X=1000)}$$

un individu avec une balance de 1000 a 5.793608×10^{-3} *fois plus de chance d'avoir default=1 (donc d'être en défaut) que de ne pas être en défaut.*

c) **[1.5pt]** Si on augmente la balance de x_1 à x_2 , on a l'odds-ratio

$$OR_{balance} = OR(x_1, x_2) = \frac{odds(x_2)}{odds(x_1)} = e^{\widehat{b}(x_2-x_1)}$$

ce qui équivaut que

$$\ln(OR(x_1, x_2)) = \widehat{b}(x_2 - x_1) = 5.499^{-3}(x_2 - x_1)$$

$\widehat{b}$ mesure l'influence de la variable balance sur default. Il est positif, donc plus le client utilise ça carte, plus il a de chances d'être en défaut.

d) **[1 pt]** Écriture des matrices:

$$X = \begin{pmatrix} 1 & 729,5265 \\ 1 & 817,1804 \\ 1 & 1073,292 \\ 1 & 529,2506 \end{pmatrix}$$

La matrice des variances

$$V = diag(\widehat{\pi}(x)(1-\widehat{\pi}(x)) = diag(\widehat{\pi}(729,5265)(1-\widehat{\pi}(729,5265)), \ldots)$$

e) **[1.5 pt]**. Intervalle de confiance

$$IDC(OR) = [e^{\widehat{b}-z_{1-\alpha/2}\widehat{\sigma}_{\widehat{b}}}, \quad e^{\widehat{b}+z_{1-\alpha/2}\widehat{\sigma}_{\widehat{b}}}]$$

où $z_{1-\alpha/2}$ est le quantile de normale centré réduite. Avec $\alpha = 5\%$ on a $z_{1-\alpha/2} = 1,96$ et

$$\widehat{\sigma}_{\widehat{b}} = 2.204 \times 10^{-4}$$

$$e^{5.499\times10^{-3}-1.96\times2.204\times10^{-4}} = 1.00508$$
$$e^{5.499\times10^{-3}+1.96\times2.204\times10^{-4}} = 1.005949$$

$$IDC(OR) = [1.00508, \quad 1.005949]$$

f) [1.5 pt] teste de la pente b. On veut tester

$$H_0 : b = 0 \qquad H_1 : n \neq 0$$

On utilise la statistique

$$T = \frac{|\widehat{b}|}{\widehat{\sigma}_{\widehat{b}}} \hookrightarrow \mathcal{N}(0,1)$$

Pour $\alpha = 5\%$ le quantile associé est $q_{1-\alpha/2} = 1,96$
et la valeur de la statitque observée est

$$T_{obs} = \frac{5.499^{-3}}{2.204^{-4}} = 24.95009 > 1.96$$

On rejette l'hypothèse H_0. La balance a de l'influence sur le défaut de la dette avec une erreur de 5% de se tromper.

g) [1pt]. Prédiction et classement.
Pour $xnew = 1000$ la valeur de probabilité prédite $\widehat{\pi}(xnew) = 0.0058 < 0.5$ donc $xnew = 1000$ appartient à la classe $(Y = 0)$, c'st-à-dire à la classe des clients qui ne font pas défaut
Pour $xnew = 2000$, $\widehat{\pi} = 0.59 > 0.5$ donc l'individu appartient à la classe de ceux qui font défaut.

2-Modèle 2.
a) [1 pt]

$$logit(\widehat{\pi}(x) = \widehat{a} + \widehat{b}1_{x=1} = -3.50413 + 0.40489 \times 1_{\{x=1\}}$$

b) [1.5 pt]

$$\widehat{\pi}(x) = \frac{e^{\widehat{a}+\widehat{b}1_{\{x=1\}}}}{1 + e^{\widehat{a}+\widehat{b}1_{\{x=1\}}}} = \frac{1}{1 + e^{-\widehat{a}-\widehat{b}1_{\{x=1\}}}}$$

avec $x =$ student.
Si $x = 1$, on a:

$$\mathbb{P}(\text{default = Yes} \mid \text{student = Yes}) = \exp(-3.50 + 0.40 \times 1)/(1 + \exp(-3.50 + 0.40 \times 1)) = 0.0431,$$

Si si $x = 0$, on a:

$$\mathbb{P}(\text{default = Yes} \mid \text{student = Non}) = \exp(-3.50 + 0.40 \times 0)/(1 + \exp(-3.50 + 0.40 \times 0)) = 0.0292$$

c) [1.5 pt]
Pour sutdent=Yes, $x = 1$

$$odds(1) = e^{-3,50413+0.40489} = 0.04508345 = \frac{\mathbb{P}(Y = 1|X = 1)}{\mathbb{P}(Y = 0|X = 1)}$$

Pour Student=No, $x = 0$ autre client

$$odds(0) = e^{-3,50413} = 0.03007293 = \frac{\mathbb{P}(Y=1|X=0)}{\mathbb{P}(Y=0|X=0)}$$

d) [1pt] Pour comparer les deux modèles on regarde l'AIC, on choisit celui avec qui un l'AIC plus petit. Il faut donc choisir le Modèle 1.

Corrigé de l'exercice 2.

$1^{ère}$ **Partie.**
a) [1 pt] On choisit $A = W$ et $f(\beta) = y - X\beta$. On obtient

$$\begin{aligned} \frac{d}{d\beta}S(\beta) &= 2\left(\frac{d}{d\beta}(y - X\beta)\right)^t W(y - x\beta) \\ &= -2X^tW(y - X\beta) \\ &= -2(X^tWy - X^tWX\beta) \end{aligned}$$

Ainsi, les équations normales à résoudre pour trouver l'estimateur $\widehat{\beta}$ minimisant la somme de carrés pondérés $S(\beta)$ sont

$$(X^tWX)\widehat{\beta} = X^tWy$$

D'où

$$\widehat{\beta} = (X^tWX)^{-1}X^tWy.$$

b) [1 pt]
En utilisant le fait que $Var[AX] = AVar[X]A^t$ et que $Var(y) = var(\epsilon) = \sigma^2W^{-1}$ et la symétrie de X, on obtient

$$\begin{aligned} Var(\widehat{\beta}) &= (X^tWX)^{-1}X^tWVar(y)W^tX(X^tWX)^{-1} \\ &= \sigma^2(X^tWX)^{-1}X^tWW^{-1}WX(X^tWX)^{-1} \\ &= \sigma^2(X^tWX)^{-1} \end{aligned}$$

puisque les matrices W et X^tWX sont symétriques.

$2^{ème}$ **Partie.**
a) [0.5 pt]. On a

$$Y = \beta X + \epsilon$$

avec

$$Y = (y_1, \dots y_n)^t, \qquad X = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix}, \qquad \epsilon = (\epsilon_1, \dots, \epsilon_n)^t$$

b) [1 pt] Dans le cas de la régression linéaire simple passant par l'origine et en supposant que $W = diag(w_1, \dots, w_n)$, ces formules se réduisent en

$$\widehat{\beta} = \frac{\sum_{i=1}^n w_iX_iY_i}{\sum_{i=1}^n w_iX_i^2}$$

c) [1 pt]

$$Var(\widehat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^n w_iX_i^2}$$

d) **[0.5 pt]** Ici,on a $W = I$, d'où

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$

et

$$Var(\widehat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2}$$

e) **[0.5 pt]**) Si $Var[\epsilon_i] = \sigma^2 X_i$, alors $w_i = X_i^{-1}$. Le cas général se simplifie donc en

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i} = \frac{\overline{Y}}{\overline{X}}$$

$$Var(\widehat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n} X_i} = \frac{\sigma^2}{n\overline{X}}$$

f) **[0.5 pt]** Si $Var[\epsilon_i] = \sigma^2 X_i^2$, alors $w_i = X_i^{-2}$. On a donc

$$\widehat{\beta} = \frac{1}{n} \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}$$

$$Var(\widehat{\beta}) = \frac{\sigma^2}{n}$$

Corrigé de l'exercice 3.

a. Forme matricielle

[0.5 pt]

$$y = (y_1, \ldots, y_n)^t, \qquad X = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}, \qquad \epsilon = (\epsilon_1, \ldots, \epsilon_n)^t, \quad \beta = (\beta_0, \beta_1)^t$$

- Expression des estimateurs:**[1 pt]**

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}, \qquad \widehat{\beta}_1 = \frac{\sum_i X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_i X_i^2 - n\overline{X}^2}$$

- Calcul **[1 pt]**

$$\widehat{\beta}_1 = \frac{88858.02 - 1250.77 \times (12935/188)}{12699.04 - 188 \times (1250.77/188)^2} = \frac{2801.052}{4377.627} = 0.6398974$$

$$\widehat{\beta}_0 = 12935/188 - 0.6398974 \times 1250.77/188 = 64.54593$$

b). Équation d'Anova **[1.5 pt]**

$$SCT = SCE + SCR$$

avec

La somme des carrés totale : $SCT = ||Y - \overline{Y}\mathbf{1}||^2$.

-somme des carrés expliquée : $SCE = ||\widehat{Y} - \overline{Y}||^2$

Le coefficient de détermination R^2 est défini par :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

$$\begin{aligned} SCT &= ||Y - \overline{Y}\mathbf{1}||^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \\ &= \sum_i Y_i^2 - 2n\overline{Y}^2 + n\overline{Y}^2 \\ &= \sum_i Y_i^2 - n\overline{Y}^2 \\ &= 907647 - 188 \times (12935/188)^2 = 17677.72 \end{aligned}$$

$$\begin{aligned} SCE &= ||\widehat{Y} - \overline{Y}||^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \overline{Y})^2 = \sum_{i=1}^{n}\left(\widehat{\beta}_0 + \widehat{\beta}_1 X_i - \overline{Y}\right)^2 \\ &= \widehat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \overline{X})^2 = \widehat{\beta}_1^2[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2] \\ &= (0.6398974)^2 \times (12699.04 - 188 \times (1250.77/188)^2) \\ &= 1792.501 \end{aligned}$$

Ce qui donne

$$R^2 = \frac{SCE}{SCT} = \frac{1792.501}{17677.72} = 0.1013989$$

Donc $\rho = \sqrt{0.1013989} = 0.3184319$.

c) Expression des estimateurs de la variance et calcul. **[1 pt]**
L'estimateur des moindres carrés de σ^2 est donné par

$$\widehat{\sigma^2} = \frac{||\widehat{\epsilon}||^2}{n-2} = \frac{SCT - SCE}{n-2} = \frac{17677.72 - 1792.501}{186} = 85.4044$$

d) Lois des estimateurs (avec justification) **[1 pt]**
Si on suppose de plus que les ϵ_i suivent une loi normale de moyenne nulle de variance σ^2, alors chacun des estimateurs suie une loin normale, car ces estimateurs sont des combinaisons linéaires Y_i donc des ϵ_i. Plus précisément

$$\widehat{\beta}_0 \hookrightarrow \mathcal{N}(\beta_0, \sigma^2_{\widehat{\beta}_0})$$
$$\widehat{\beta}_1 \hookrightarrow \mathcal{N}(\beta_1, \sigma^2_{\widehat{\beta}_1})$$

e) Expression des estimateurs de variances et calcul **[1.5 pt]**

$$\sigma^2_{\widehat{\beta}_0} = \sigma^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]$$
$$\sigma^2_{\widehat{\beta}_1} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Comme σ^2 est inconnue, on la remplace par son estimateur $\widehat{\sigma^2} = \frac{||\widehat{\epsilon}||^2}{n-2}$ dans les expressions ci-dessus pour avoir les expressions de $\widehat{\sigma}^2_{\widehat{\beta}_0}$ et $\widehat{\sigma}^2_{\widehat{\beta}_1}$

-Valeur des estimateurs

$$\sigma^2_{\widehat{\beta}_0} = \widehat{\sigma}^2\left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n} x_i^2 - n(\overline{x})^2}\right] = 85.4044\left[\frac{1}{188} + \frac{(1250.77/188)^2}{12699.04 - 188(1250.77/188)^2}\right] = 1.317815$$

$$\sigma_{\widehat{\beta}_0} = \sqrt{1.317815} = 1.147961$$

Et

$$\sigma^2_{\widehat{\beta}_1} = \widehat{\sigma}^2 \left[\frac{1}{\sum_{i=1}^n x_i^2 - n(\overline{x})^2} \right] = 85.4044 \left[\frac{1}{12699.04 - 188(1250.77/188)^2} \right] = 0.01950929$$

$$\sigma_{\widehat{\beta}_1} = \sqrt{0.01950929} = 0.1396757$$

f) Intervalle de confiance **[1 pt]**

$$IC(\beta_1) = \left[\widehat{\beta}_1 - t_{\frac{\alpha}{2}} \sigma_{\widehat{\beta}_1},\ \widehat{\beta}_1 - t_{\frac{\alpha}{2}} \sigma_{\widehat{\beta}_1} \right]$$

où $t_{\frac{\alpha}{2}}$ est le quantile de la loi de student à $n-2$ degrés de liberté. On a $t_{\frac{\alpha}{2}} = 1.96$ et

$$\begin{aligned} IC(\beta_1) &= [0.6398974 - 1.96 \times 0.1396757, \quad 0.6398974 + 1.96 \times 0.1396757] \\ &= [0.366133,\ 0.9136618] \end{aligned}$$

h) Test d'hypothèses:**[1 pt]**
La statistique de test est

$$T = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\widehat{\beta}_1}}$$

sous $H_0 : \beta_1 = 0$, T suit une loi de student à $n-2$ degrés de liberté.

On a

$$t_{obs} = \frac{0.6398974}{0.1396757} = 4.581308 > 1.96$$

On rejette donc l'hypothèse $\beta_1 = 0$ et conclut que le modèle est significatif avec une risque de 5% de se tromper.

h) **[1 pt]** On a d'après la question b) de l'exercice 2 que

$$\widehat{\beta} = \frac{\sum_{i=n} X_i Y_i}{\sum_{i=1}^n X_i^2} = \frac{88858.02}{12699.04} = 6.997223$$

Dans ce cas

$$\begin{aligned} R^2 &= \frac{SCE}{SCT} = \frac{||\widehat{Y}||^2}{||Y||^2} = \frac{\sum_{i=1}^n \widehat{y_i}^2}{\sum_{i=1}^n y_i^2} = \frac{\widehat{\beta}^2 \sum_{i=1}^n x_i^2}{\sum_{i=1}^n y_i^2} \\ &= (6.997223)^2 \times \frac{12699.04}{907647} \\ &= 0.6850233 \end{aligned}$$

i) **[0.5 pt]**