STT5100 - Hiver 2019 - Examen Intra (OLS)

Arthur Charpentier

Examen **A**

Les calculatrices sont autorisées. Les documents sont en revanche interdits, sauf une page d'aide mémoire. L'examen dure 3 heures, mais toute sortie avant midi est autorisée, et sera définitive.

Dans les feuilles qui suivent, il y a 30 questions relatives au cours sur les modèles linéaires. Pour chaque question (sauf deux), cinq réponses sont proposées. Une seule est valide, et vous ne devez en retenir qu'une,

- vous gagnez 1 point par bonne réponse
- vous ne perdez pas de points pour une mauvaise réponse
- vous ne gagnez pas de point pour plusieurs réponses

Aucune justification n'est demandée. Deux questions reposent sur un graphique qu'il faudra tracer sur la feuille de réponses (au dos). Votre note finale est le total des points (sur 30). Il y a une 31ème question, bonus. Une prédiction parfaite (sur 30) donnera un point bonus qui s'ajoutera à la note.

La page de réponses est au dos de celle que vous lisez présentement : merci de décrocher ladite feuille et de ne rendre que cette dernière, après avoir indiqué votre code permanent en haut à gauche.

Merci de cocher le carré en bleu ou en noir. En cas d'erreur, vous pouvez cocher une autre case en rouge. Seule cette dernière sera alors retenue.

Le surveillant ne répondra à aucune question durant l'épreuve : en cas de soucis sur une question (interprétation possiblement fausse, typo, etc), vous pouvez mettre un court commentaire sur la feuille de réponses.

Formulaire : Quantiles de lois usuelles. Exemple pour une loi normale - $Z \sim \mathcal{N}(0,1)$, $\mathbb{P}(Z \leqslant 2.326) = 99\%$.

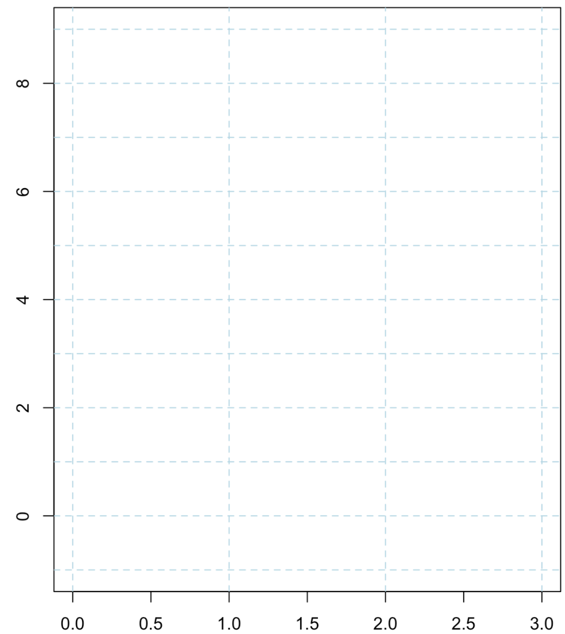
	75%	90%	95%	97.5%	99%	99.5%	99.9%	99.95%
Loi normale	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291
Student (50)	0.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496
Student (30)	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
Student (20)	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.849
Student (15)	0.691	1.341	1.753	2.131	2.602	2.947	3.733	4.073
Student (10)	0.700	1.372	1.812	2.228	2.764	3.169	4.143	4.587
Student (9)	0.703	1.383	1.833	2.262	2.821	3.250		
Student (8)	0.706	1.397	1.860	2.306	2.896	3.355		
Student (7)	0.711	1.415	1.895	2.365	2.998	3.499		
Student (6)	0.718	1.440	1.943	2.447	3.143	3.707		
Student (5)	0.727	1.476	2.015	2.571	3.365	4.032		
Student (4)	0.741	1.533	2.132	2.776	3.747	4.604		
Student (3)	0.765	1.638	2.353	3.182	4.541	5.841		

Code permanent :

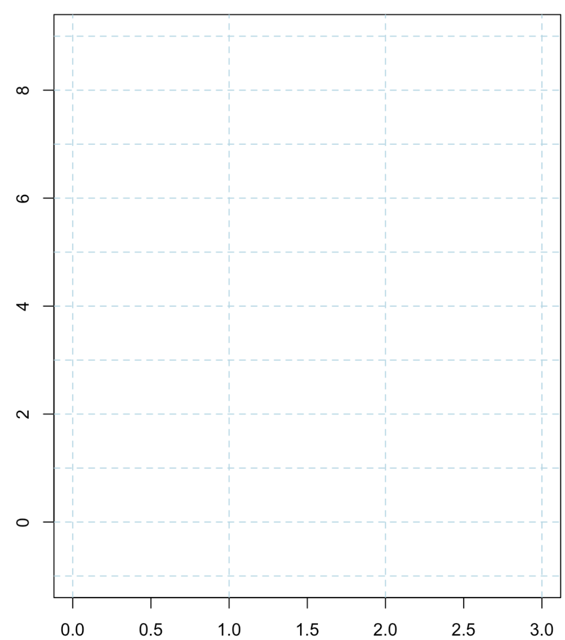
Sujet : A

question 1	☐ A	☐ B	☐ C	☐ D	☐ E
question 2	☐ A	☐ B	☐ C	☐ D	☐ E
question 3	☐ A	☐ B	☐ C	☐ D	☐ E
question 4	☐ A	☐ B	☐ C	☐ D	☐ E
question 5	☐ A	☐ B	☐ C	☐ D	☐ E
question 6	☐ A	☐ B	☐ C	☐ D	☐ E
question 7	Figure à droite (en haut)				
question 8	☐ A	☐ B	☐ C	☐ D	☐ E
question 9	☐ A	☐ B	☐ C	☐ D	☐ E
question 10	☐ A	☐ B	☐ C	☐ D	☐ E
question 11	☐ A	☐ B	☐ C	☐ D	☐ E
question 12	☐ A	☐ B	☐ C	☐ D	☐ E
question 13	☐ A	☐ B	☐ C	☐ D	☐ E
question 14	☐ A	☐ B	☐ C	☐ D	☐ E
question 15	☐ A	☐ B	☐ C	☐ D	☐ E
question 16	☐ A	☐ B	☐ C	☐ D	☐ E
question 17	☐ A	☐ B	☐ C	☐ D	☐ E
question 18	☐ A	☐ B	☐ C	☐ D	☐ E
question 19	☐ A	☐ B	☐ C	☐ D	☐ E
question 20	☐ A	☐ B	☐ C	☐ D	☐ E
question 21	☐ A	☐ B	☐ C	☐ D	☐ E
question 22	☐ A	☐ B	☐ C	☐ D	☐ E
question 23	☐ A	☐ B	☐ C	☐ D	☐ E
question 24	☐ A	☐ B	☐ C	☐ D	☐ E
question 25	☐ A	☐ B	☐ C	☐ D	☐ E
question 26	☐ A	☐ B	☐ C	☐ D	☐ E
question 27	☐ A	☐ B	☐ C	☐ D	☐ E
question 28	☐ A	☐ B	☐ C	☐ D	☐ E
question 29	Figure à droite (en bas)				
question 30	☐ A	☐ B	☐ C	☐ D	☐ E
question 31	*Combien de bonnes réponses pensez vous avoir ?*				

question 7 :



question 29 :



1 On considère un modèle linéaire $y = \beta_0 + \varepsilon$ où ε est un bruit Gaussien, centré, de variance σ^2. Quel est l'estimateur par moindres carrés du paramètre β_0 ?

A) 0

B) $\overline{y}$

C) $\overline{\varepsilon}$

D) $\overline{y}/\overline{\varepsilon}$

E) $\dfrac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$

2 On considère un modèle linéaire $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ estimé sur $n = 10$ observations. La variable x_i est ici binaire, et vaut 1 si l'individu i possède une carte de crédit, et 0 sinon. On sait que 40% des individus dans l'étude possèdent une carte de crédit. L'estimateur par moindre carrés de β_1 est 4. Enfin, on sait que $\sum_{i=1}^{10}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 = 92$. Que vaut la statistique de Student du test $H_0 : \beta_1 = 0$? (*on retiendra la valeur la plus proche*)

A) 0.9

B) 1.2

C) 1.5

D) 1.8

E) 2.1

Les questions 4 à 8 portent sur la sortie de la question 3.

3 On a obtenu la sortie de régression suivante, où une variable continue y a été régressée sur une autre variable continue x_1 et une variable catégorielle $x_2 \in \{\text{femme}, \text{homme}\}$,

```
Call
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8129 -1.0718  0.1338  0.8287  3.0995

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     1.4956     0.4572
x1              1.6584     0.1486  11.163 4.60e-15 ***
x2Homme         2.5089     0.4033   6.221 1.07e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.423 on 49 degrees of freedom
```

```
Multiple R-squared:  0.7913,Adjusted R-squared:  0.7828
F-statistic: 92.88 on 2 and 49 DF,  p-value: < 2.2e-16
```

Quelle est la p-value du test de significatité de la constante β_0 ?

A) environ 0.001

B) environ 0.002

C) environ 0.005

D) environ 0.010

E) environ 0.020

4 (*suite*) On souhaite faire une prévision pour une nouvelle observation, ($x_1 = 1.5, x_2 =$ Homme). Quelle serait la prévision pour $\widehat{y} = \mathbb{E}[Y|x_1 = 1.5, x_2 = \text{Homme}]$ obtenue par moindres carrés ?

A) moins de 6

B) entre 6 et 7

C) entre 7 et 8

D) entre 8 et 9

E) plus que 9

5 (*suite*) Compte tenu de l'erreur d'échantillonnage, quel serait la borne supérieure de l'intervalle de confiance à 95% pour $\widehat{y}$?

A) moins de 6

B) entre 6 et 7

C) entre 7 et 8

D) entre 8 et 9

E) plus que 9

6 (*suite*) Quelle est la probabilité que Y dépasse 9 (donnez un estimateur de $\mathbb{P}[Y > 9|x_1 = 1.5, x_2 = \text{Homme}]$) ?

A) plus de 12.5%

B) environ 10%

C) environ 5%

D) environ 1%

E) moins de 0.5%

7 (*suite*) Sur la Figure en haut de la feuille de réponses

- tracez avec un trait plein l'estimateur de $\mathbb{E}[Y|x_1, x_2 = \text{Femme}]$ lorsque x varie entre 0 et 3
- tracez avec un trait en pointillés l'estimateur de $\mathbb{E}[Y|x_1, x_2 = \text{Homme}]$ lorsque x varie entre 0 et 3

8 (*suite*) Quelle est la log-vraisemblance maximale d'un modèle linéaire Gaussien de la forme $y = \beta_0 + \beta_1 x_1 + \beta_2 \mathbf{1}(x_2 = \text{homme}) + \varepsilon$?

A) moins que -80

B) entre -80 et -60

C) entre -60 et -40

D) entre -40 et -20

E) plus que -20

9 Deux actuaires tentent de construire un modèle à partir d'un jeu de données contenant deux variables x et y.

- l'actuaire (1) fait la régression linéaire de y sur x
- l'actuaire (2) fait la régression linéaire de x sur y

Ils estiment leurs modèles par la méthode des moindres carrés. Considérons les trois affirmations suivantes

I. la pente (estimée) dans le modèle (2) est l'inverse de la pente (estimée) dans le modèle (1)

II. les deux modèles ont le même R^2

III. les statistiques du test de Student pour la significativité de la pente sont identiques dans les deux modèles

Parmi les affirmations, lesquelles sont vraies :

A) aucune

B) I et II seulement

C) I et III seulement

D) II et III seulement

E) ni A, ni B, ni C, ni D

10 On dispose d'observations $(y_i, x_{1,i}, x_{2,i})$ avec $i = 1, \cdots, 80$. On estime trois modèles :

(A) $y = \widehat{\beta}_0^A + \widehat{\beta}_0^A x_1 + \widehat{\beta}_0^A x_2 + \widehat{\varepsilon}$ estimé sur les observations $i \in \{1, \cdots, 30\}$

(B) $y = \widehat{\beta}_0^B + \widehat{\beta}_0^B x_1 + \widehat{\beta}_0^B x_2 + \widehat{\varepsilon}$ estimé sur les observations $i \in \{31, \cdots, 50\}$

(C) $y = \widehat{\beta}_0^C + \widehat{\beta}_0^C x_1 + \widehat{\beta}_0^C x_2 + \widehat{\varepsilon}$ estimé sur les observations $i \in \{1, \cdots, 80\}$

On note $R^2_\bullet$ et $SCR_\bullet$ le R^2 et la somme des carrés des résidus des trois modèles. On souhaite utiliser le test de Fisher pour tester si les modèles (A) et (B) sont identiques. Quelle statistique doit-on utiliser ?

A) $F_{3,74} = \dfrac{(SCR_C - SCR_A - SCR_B)/3}{(SCR_A + SCR_B)/74}$

B) $F_{6,77} = \dfrac{(SCR_C - SCR_A - SCR_B)/6}{(SCR_A + SCR_B)/77}$

C) $F_{77,74} = \dfrac{SCR_C/77}{(SCR_A + SCR_B)/74}$

D) $F_{3,74} = \dfrac{(R^2_C - R^2_A - R^2_B)/3}{(R^2_A + R^2_B)/74}$

E) $F_{77,74} = \dfrac{(R^2_C)/77}{(R^2_A + R^2_B)/74}$

11 On considère un modèle $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ estimé par moindres carrés sur $n = 6$ observations. On a

- $\widehat{\beta}_0 = 2.31$
- $\widehat{\beta}_1 = 1.15$
- $\widehat{\text{Var}}[\widehat{\beta}_0] = 0.057^2$
- $\widehat{\text{Var}}[\widehat{\beta}_1] = 0.043^2$

On veut tester $H_0 : \beta_1 = 1$ contre l'hypothèse alternative $H_0 : \beta_1 \neq 1$. Quel serait la plus petite valeur de α pour que H_0 soit rejetée avec un niveau de significativité α ?

A) moins de 0.01

B) entre 0.01 et 0.02

C) entre 0.02 et 0.05

D) entre 0.05 et 0.10

E) plus de 0.10

12 On considère le modele suivant, $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i$ estimé par moindres carrés sur $n = 32$ observations. On obtient la sortie (partielle) suivante

```
Call:
lm(y = x1 + x2 + x3, data = df)

Coefficients:
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)     44.200      5.960
x1              -0.295      0.118
x2               9.110      6.860
x3              -8.700      1.200
```

On souhaite juger la significativité avec un seuil $\alpha = 10\%$. Quelles sont les variables pour lesquels les coefficients sont significativement différents de zéro ?

A) la constante

B) la constante et x_1

C) la constante et x_2

D) la constante, x_1 et x_3

E) la constante, x_2 et x_3

13 On considère le modele suivant, $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i$ estimé par moindres carrés sur $n = 20$ observations. Ici y désigne le montant dépensé dans des équipements de cuisine, x_1 est le revenu, x_2 est le nombre d'années d'études et x_3 le montant d'épargne. On obtient la sortie (partielle) suivante

```
Call:
lm(kitchen = income + education + savings, data = df)

Coefficients:
              Estimate Std. Error   t value  Pr(>|t|)
(Intercept)    0.15085    0.73776   0.20447
income         0.26528    0.10127   2.61953
education      6.64357    2.01212   3.30178
savings        7.31450    2.73977   2.66975
```

On sait de plus que

$$\sum_{i=1}^{20} \widehat{\varepsilon}_i^2 = 2.65376 \text{ et } \sum_{i=1}^{20} (y_i - \overline{y})^2 = 7.62956$$

On veut utiliser un test de Fisher pour tester $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ (l'hypothèse alternative étant qu'un des coefficients - au moins - est non nul). Quelle est la valeur de la statistique de test F ?

A) moins de 1

B) entre 1 et 3

C) entre 3 et 5

D) au moins 5

E) nous n'avons pas assez d'information pour faire les calculs

Les questions 15 à 17 portent sur les informations données dans la question 14.

14 On considère le modele suivant, $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i$ estimé par moindres carrés. On note $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$. On nous donne le vecteur $\boldsymbol{y}$, la matrice $\boldsymbol{X}$ et quelques autres,

$$\boldsymbol{y} = \begin{pmatrix} 19 \\ 32 \\ 19 \\ 17 \\ 13 \\ 15 \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & 0 & 1 & 9 \\ 1 & 1 & 1 & 15 \\ 1 & 1 & 1 & 8 \\ 1 & 1 & 0 & 7 \\ 1 & 1 & 0 & 6 \\ 1 & 0 & 0 & 6 \end{pmatrix}, \boldsymbol{X}^\top \boldsymbol{X} = \begin{pmatrix} 6 & 4 & 3 & 51 \\ & 4 & 2 & 36 \\ & & 3 & 32 \\ & & & 491 \end{pmatrix},$$

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \begin{pmatrix} 1.75 & -0.20 & 0.54 & -0.20 \\ & 0.84 & 0.25 & -0.06 \\ & & 1.38 & -0.16 \\ & & & 0.04 \end{pmatrix}, (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \begin{pmatrix} 2.335 \\ 0.297 \\ -0.196 \\ 1.968 \end{pmatrix},$$

et

$$\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top = \begin{pmatrix} 0.684 & 0.070 & 0.247 & -0.171 & -0.146 & 0.316 \\ & 0.975 & -0.044 & 0.108 & -0.038 & -0.070 \\ & & 0.797 & 0.063 & 0.184 & -0.247 \\ & & & 0.418 & 0.411 & 0.171 \\ & & & & 0.443 & 0.146 \\ & & & & & 0.684 \end{pmatrix}.$$

Donnez la valeur de $\widehat{\varepsilon}_5$

A) moins de -1

B) entre -1 et 0

C) entre 0 et 1

D) entre 1 et 2

E) plus de 2

15 (*suite*) On note $\widehat{\boldsymbol{\beta}}$ l'estimateur par moindres carrés de $\boldsymbol{\beta}$, et $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$. Que vaut $\widehat{\boldsymbol{y}}^\top [\boldsymbol{y} - \widehat{\boldsymbol{y}}]$?

A) $\widehat{\sigma}^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$

B) $\widehat{\sigma}^2 \mathbb{I}$ où $\mathbb{I}$ est la matrice identié 6×6

C) σ^2

D) 0

E) 45

16 (*suite*) Que vaut $\sum_{i=1}^{6} \widehat{y}_i$?

A) moins de 100

B) entre 100 et 110

C) entre 110 et 120

D) entre 120 et 130

E) plus de 130

17 (*suite*) On s'interroge sur la sensitivité des prédictions aux observations, plus précisément, on note sensitivité$_i$ la grandeur $\left|\frac{\partial \widehat{y}_i}{\partial y_i}\right|$. Pour quelle observation i cette grandeur est maximale ?

A) $i = 1$

B) $i = 2$

C) $i = 3$

D) $i = 4$

E) $i = 5$

18 On considère un modèle linéaire $y = \beta_1 x^2 + \varepsilon$ où ε est un bruit Gaussien, centré, de variance σ^2. Quel est l'estimateur par moindres carrés du paramètre β_1 ?

A) $\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$

B) $\frac{\sum_{i=1}^n x_i^2 y_i}{\sum_{i=1}^n x_i^4}$

C) $\frac{\sum_{i=1}^n (x_i - \overline{x})^2 y_i^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$

D) $\frac{\sum_{i=1}^n (x_i - \overline{x})^2 y_i}{\sum_{i=1}^n (x_i - \overline{x})^4}$

E) $\frac{\sum_{i=1}^n (x_i - \overline{x})^2 y_i^2}{\sum_{i=1}^n (x_i - \overline{x})^4}$

19 On souhaite ici modéliser une variable y à partir de 5 prédicteurs possibles (x_1,x_2,x_3,x_4,x_5). En gardant toujours la constante, 32 modèles sont alors possibles. En les estimant tous, on obtient les sorties suivantes

modèle	variables	R^2	$\log \mathcal{L}$
(1)	constante	0.00	0.05
(2)	constante, x_1	0.56	1.30
(3)	constante, x_2	0.57	1.40
(4)	constante, x_3	0.55	1.20
(5)	constante, x_4	0.52	1.15
(6)	constante, x_5	0.51	1.10
(7)	constante, x_1, x_2	0.61	2.50
(8)	constante, x_1, x_3	0.64	2.75
(9)	constante, x_1, x_4	0.63	2.60
(10)	constante, x_1, x_5	0.69	3.00
(11)	constante, x_2, x_3	0.61	2.50
(12)	constante, x_2, x_4	0.62	2.55
(13)	constante, x_2, x_5	0.68	2.90
(14)	constante, x_3, x_4	0.66	2.80
(15)	constante, x_3, x_5	0.64	2.75
(16)	constante, x_4, x_5	0.60	2.45

modèle	variables	R^2	$\log \mathcal{L}$
(17)	constante, x_1, x_2, x_3	0.73	3.35
(18)	constante, x_1, x_2, x_4	0.71	3.25
(19)	constante, x_1, x_2, x_5	0.72	3.30
(20)	constante, x_1, x_3, x_4	0.75	3.50
(21)	constante, x_1, x_3, x_5	0.76	3.60
(22)	constante, x_1, x_4, x_5	0.79	3.90
(23)	constante, x_2, x_3, x_4	0.78	3.70
(24)	constante, x_2, x_3, x_5	0.74	3.40
(25)	constante, x_2, x_4, x_5	0.75	3.45
(26)	constante, x_3, x_4, x_5	0.73	3.35
(27)	constante, x_1, x_2, x_3, x_4	0.88	4.20
(28)	constante, x_1, x_2, x_3, x_5	0.80	3.95
(29)	constante, x_1, x_2, x_4, x_5	0.87	4.10
(30)	constante, x_1, x_3, x_4, x_5	0.83	4.00
(31)	constante, x_2, x_3, x_4, x_5	0.85	4.05
(32)	constante, x_1, x_2, x_3, x_4, x_5	0.90	4.25

Deux actuaires adoptent des statégies différentes pour construire leur modèle

(1) le premier actuaire utilise le meilleur modèle au sens du critère d'Akaike (*best subset selection*)

(2) le second actuaire utilise une méthode pas à pas *forward* (*stepwise*)

Soit $AIC_\bullet$ le critère d'Akaike retenu par chacun des actuaires, au final. Que vaut $|AIC_{(1)} - AIC_{(2)}|$?

A) moins de 0.15

B) entre 0.15 et 0.30

C) entre 0.30 et 0.45

D) entre 0.45 et 0.60

E) plus de 0.60

20 On a obtenu la sortie de régression suivante

```
Call
lm(formula = y ~ x1 + x2)
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  -2.36500     0.22184  -10.661   < 2e-16
x1            0.47621     0.06169    7.720  3.25e-13
x2            0.08269     0.06977    1.185     0.237
---
Residual standard error: 0.3064 on 237 degrees of freedom
Multiple R-squared: 0.3438, Adjusted R-squared: 0.3382
F-statistic: 62.07 on 2 and 237 DF, p-value: < 2.2e-16
```

Quelle est la borne inférieure de l'intervalle de confiance à 95% de β_2 ?

A) moins de -0.1

B) entre -0.1 et -0.05

C) entre -0.05 et 0

D) entre 0 et 0.05

E) plus de 0.05

21 On considère un modèle, constuit à partir d'observations (y_i, x_i), de la forme

$$y_i = \underbrace{\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 (x_i - \zeta)_+^2}_{=f(x_i)} + \varepsilon_i = f(x_i) + \varepsilon_i$$

où classiquement $(x - \zeta)_+$ désigne la partie positive de $(x - \zeta)$, c'est à dire $(x - \zeta)$ si $x > \zeta$ et zéro sinon. On considère les affirmations suivantes

(1) $x \mapsto f(x)$ est continue en ζ

(2) $x \mapsto f'(x)$ est continue en ζ

(3) $x \mapsto f''(x)$ est continue en ζ

Quelles affirmations sont vraies ?

A) (1) seulement

B) (2) seulement

C) (1) et (2)

D) (1), (2) et (3)

E) aucune des affirmations

22 On veut construire un modèle $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$. On dispose d'un échantillon de n observations. Quelles sont les hypothèse habituelles faite pour estimer les paramètres du modèles ?

A) $\varepsilon_i = 0$ pour $i = 1, \cdots, n$, et β_1 et β_2 suivent une loi normale

B) $\varepsilon_i = 0$ pour $i = 1, \cdots, n$, et les variables X_1 et X_2 suivent une loi normale

C) $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$

D) $\mathbb{E}[\varepsilon] = \text{Var}[\varepsilon] = 0$

E) aucune des affirmations (A)-(D)

23 On considère le modèle $y = \beta_0 + \beta_1 x + \varepsilon$, estimé par moindres carrés sur un échantillon de $n = 6$ observations. On sait que

$$\widehat{\beta}_1 = 4, \ \sum_{i=1}^{6}(x_i - \overline{x})^2 = 50 \text{ et } \sum_{i=1}^{6}(y_i - \overline{y})^2 - \widehat{\beta}_1 \sum_{i=1}^{6}(x_i - \overline{x})(y_i - \overline{y}) = 25$$

Quelle est la borne supérieure à 95% de l'intervalle de confiance pour β_1 ?

A) moins de 5.1

B) entre 5.1 et 5.3

C) entre 5.3 et 5.5

D) entre 5.5 et 5.7

E) plus que 5.7

24 On considère deux modèles

(1) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

(2) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$

et un échantillon de $n = 20$ observations. On sait aussi que

$$\text{pour le modèle (1) :} \quad \sum_{i=1}^{20}(y_i - \widehat{y}_i)^2 = 13.47 \text{ et } \sum_{i=1}^{20}(\widehat{y}_i - \overline{y})^2 = 22.75$$

$$\text{pour le modèle (2) :} \quad \sum_{i=1}^{20}(y_i - \widehat{y}_i)^2 = 10.52 \text{ et } \sum_{i=1}^{20}(\widehat{y}_i - \overline{y})^2 = 25.70$$

On veut tester l'hypothèse $H_0 : \beta_3 = \beta_4 = 0$ (contre l'hypothèse alternative qu'au moins un des deux est non-nul) par un test de Fisher. Quelle est la valeur de la statistique de test ?

A) moins de 1.7

B) entre 1.7 et 1.8

C) entre 1.8 et 1.9

D) entre 1.9 et 2

E) plus que 2

25 On considère une régression simple (sur une seule variable explicative). On sait que $R^2 = 0.64$ et $\sum_{i=1}^{25} \widehat{\varepsilon}_i^2 = 230$.

Calculer $\sum_{i=1}^{25} (y_i - \overline{y})^2$:

A) moins de 620

B) entre 620 et 650

C) entre 650 et 700

D) entre 700 et 730

E) plus de 730

26 On considère un modèle $y = \beta_0 + \beta_1 x + \varepsilon$, estimé sur un échantillon de $n = 10$ observations, tel que ε est un terme d'erreur, centré, de variance σ^2. On sait que

$$\sum_{i=1}^{10} x_i = 50, \quad \sum_{i=1}^{10} x_i^2 = 750, \quad \widehat{\sigma}^2 = 100.$$

Donnez la variance estimée de Y quand x vaut 10 (on retiendra la valeur la plus proche).

A) 100

B) 105

C) 110

D) 115

E) 120

27 On considère un modèle $y = \beta_0 + \beta_1 x + \varepsilon$, estimé par moindres carrés, tel que ε est un terme d'erreur, centré, de variance σ^2. On sait que

$$\overline{x} = 4, \quad \widehat{\beta}_0 = 6, \quad \widehat{\beta}_1 = 1.5, \quad \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2 = 2^2 \text{ et } \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \overline{y})^2 = 4^2.$$

En supposant obtenir un meilleur modèle, le modèle suivant est estimé, $x = \alpha_0 + \alpha_1 y + \eta$, par moindres carrés, tel que η est un terme d'erreur, centré, de variance κ^2

Que vaut $\widehat{\alpha}_1$?

A) moins de 0.3

B) entre 0.3 et 0.5

C) entre 0.5 et 0.7

D) entre 0.7 et 1

E) plus de 1

28 Considérons les affirmations suivantes, dans le contexte d'un modèle linéaire simple

(1) la variance de $\widehat{\beta}_0$ est fonction du nombre d'observations

(2) la variance de $\widehat{\beta}_1$ peut être réduite en rajoutant des variables explicatives

(3) la variance des prédictions est plus faible quand la variable explicative x est proche de $\overline{x}$

Quelles affirmations sont justes ?

A) (1) et (2) seulement

B) (1) et (3) seulement

C) (2) et (3) seulement

D) les trois affirmations

E) ni (A), ni (B), ni (C), ni (D)

29 On dispose des observations suivantes

x_i	1	1	1	1	2	2	2	2
y_i	-1	1	2	6	8	3	8	5

Sur la Figure de la page 2, tracer la droite de régression obtenue par moindres carrés, $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$.

30 On dispose d'une nouvelle observation $(x_9, y_9) = (1, 4)$, et on note $y = \widehat{\beta}'_0 + \widehat{\beta}'_1 x$ la nouvelle régression obtenue par moindres carrés.

A) $\widehat{\beta}'_0 = \widehat{\beta}_0$ et $\widehat{\beta}'_1 = \widehat{\beta}_1$

B) $\widehat{\beta}'_0 > \widehat{\beta}_0$ et $\widehat{\beta}'_1 > \widehat{\beta}_1$

C) $\widehat{\beta}'_0 > \widehat{\beta}_0$ et $\widehat{\beta}'_1 < \widehat{\beta}_1$

D) $\widehat{\beta}'_0 < \widehat{\beta}_0$ et $\widehat{\beta}'_1 > \widehat{\beta}_1$

E) $\widehat{\beta}'_0 < \widehat{\beta}_0$ et $\widehat{\beta}'_1 < \widehat{\beta}_1$

31 (*bonus*) Sur les 30 questions précédantes, combien de bonnes réponses pensez vous avoir ?