STT5100 - Automne 2019 - Examen Intra (OLS)

Arthur Charpentier

Examen **B**

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. Une question repose 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 : B

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	☐ A	☐ B	☐ C	☐ D	☐ E
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	☐ A	☐ B	☐ C	☐ D	☐ E
question 30	☐ A	☐ B	☐ C	☐ D	☐ E
question 31	*Combien de bonnes réponses pensez vous avoir ?*				

_____

1 On estime un modèle écrit sous la forme matricielle $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, où $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. On nous donne

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

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

et

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

Calculer le résidu associé à la 4ème observation

A) moins de -1

B) entre -1 et 0

C) entre 0 et 1

D) entre 1 et 2

E) plus de 2

2 On observe n observations, suivant un modèle $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. On pose $z_i = x_i^2$ et

$$b_1 = \frac{\sum(z_i - \overline{z})(y_i - \overline{y})}{\sum(z_i - \overline{z})(x_i - \overline{x})}$$

Quelle affimation parmi les cinq suivantes est vraie

A) b_1 est un estimateur non-linéaire de β_1

B) b_1 est un estimateur quadratique de β_1

C) b_1 est un estimateur linéaire biaisé de β_1

D) b_1 est un estimateur linéaire non-biaisé de β_1, mais il n'est pas BLUE (*best linear unbiased*)

E) b_1 est un estimateur BLUE de β_1 (*best linear unbiased*)

3 On observe n observations, suivant un modèle $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$. On nous donne les estimateurs par moindre carré,

$$\widehat{\beta}_0 = 2.5, \widehat{\beta}_1 = 2 \text{ et } \widehat{\beta}_2 = 3,$$

avec de plus

$$\sum (y_i - \widehat{y}_i)^2 = 225 \text{ et } (\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \begin{pmatrix} 0.0100 & 0 & 0 \\ 0 & 0.0200 & 0 \\ 0 & 0 & 0.0349 \end{pmatrix}$$

Déterminer la longueur du plus petit intervalle de prédiction à 95% pour y_i lorsque $x_{1,i} = 5$ et $x_{2,i} = 10$,

A) 20

B) 23

C) 25

D) 28

E) 30

4 On considère la régression suivante, pour expliquer l'impact de l'éducation et du nombre d'enfants sur le salaire d'une femme,

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + \varepsilon_i$$

où y est le logarithme du salaire

$$x_1 = \begin{cases} +1 & \text{si la femme n'a pas terminé le secondaire} \\ 0 & \text{si la femme a terminé le secondaire mais pas le collège} \\ -1 & \text{si la femme a terminé le collège} \end{cases}$$

$$x_2 = \begin{cases} 0 & \text{si la femme n'a pas terminé le secondaire} \\ 1 & \text{si la femme a terminé le secondaire mais pas le collège} \\ -1 & \text{si la femme a terminé le collège} \end{cases}$$

$$x_3 = \begin{cases} +1 & \text{si la femme a 0 enfant} \\ 0 & \text{si la femme a 1 ou 2 enfants} \\ -1 & \text{si la femme a 3 enfants, ou plus} \end{cases}$$

$$x_4 = \begin{cases} 0 & \text{si la femme a 0 enfant} \\ +1 & \text{si la femme a 1 ou 2 enfants} \\ -1 & \text{si la femme a 3 enfants, ou plus} \end{cases}$$

Quelle serait la différence entre les log(salaire) prédits, pour une femme ayant terminé le collège et ayant 3 enfants (ou plus), et la moyenne de toutes les femmes (les 9 catégories ont la même taille).

A) $\beta_0 - \beta_1 - \beta_2$

B) $\beta_1 + \beta_2$

C) $-\beta_1 - \beta_2$

D) $a - \beta_1 - \beta_2 + \beta_4$

E) $-\beta_1 - \beta_2 - \beta_3 - \beta_4$

5 On estime un modèle linéaire (A) en utilisant deux variables catégorielles, chacune prenant 2 modalités

$$x_1 = \begin{cases} 1 & \text{si l'assuré a plusieurs contrats} \\ 0 & \text{si l'assuré a un seul contrat} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{si l'assuré a plusieurs voitures} \\ 0 & \text{si l'assuré à une seule voiture} \end{cases}$$

On a alors le modèle de régression

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{1,i} x_{2,i} + \varepsilon_i.$$

L'estimation par moindres carrés donne

$$\widehat{\beta}_0 = -0.10,\ \widehat{\beta}_1 = -0.25,\ \widehat{\beta}_2 = 0.58 \text{ et } \widehat{\beta}_3 = -0.20.$$

Un second modèle (B) est estimé, en utilisant deux variables catégorielles, chacune prenant 2 modalités

$$z_1 = \begin{cases} 0 & \text{si l'assuré a plusieurs contrats} \\ 1 & \text{si l'assuré a un seul contrat} \end{cases}$$

$$z_2 = \begin{cases} 0 & \text{si l'assuré a plusieurs voitures} \\ 1 & \text{si l'assuré à une seule voiture} \end{cases}$$

On a alors le modèle de régression

$$y_i = \alpha_0 + \alpha_1 z_{1,i} + \alpha_2 z_{2,i} + \alpha_3 z_{1,i} z_{2,i} + \varepsilon_i.$$

On obtient alors les estimateurs $\widehat{\alpha}_j$ par moindre carrés. Considérons les 4 paires $(\widehat{\beta}_j, \widehat{\alpha}_j)$. On se demande combien sont identiques

A) 0 paires sont strictement identiques, et 1 paire est identique au signe près

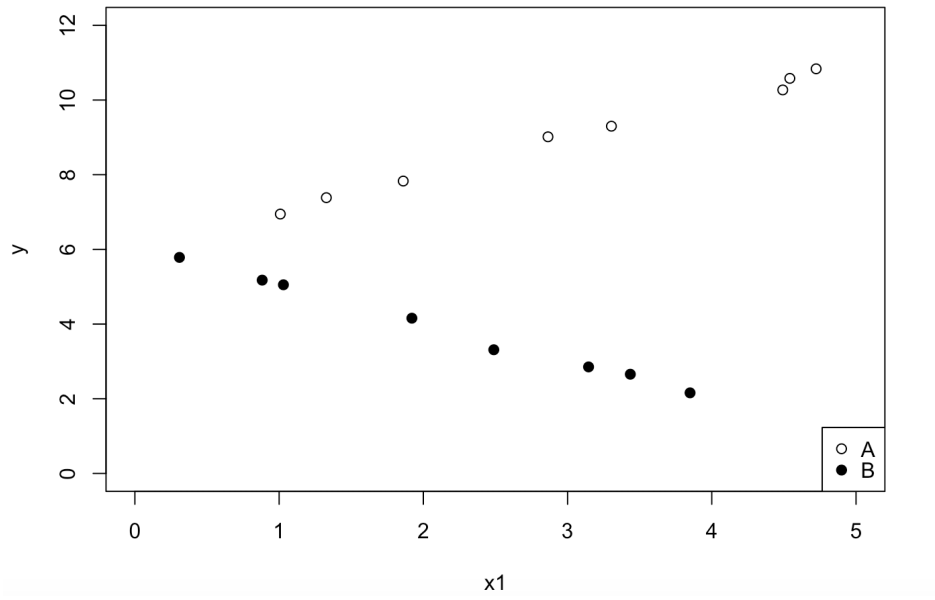
B) 1 paire est strictement identique, et 2 paires sont identiques au signe près

C) 0 paires sont strictement identiques, et 2 paires sont identiques au signe près

D) 1 paire est strictement identique, et 3 paires sont identiques au signe près

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

6 On étudie le taux d'humidité de deux aliments (y), en fonction du temps (x_1) pour deux aliments $\{A, B\}$ ($x_2 = \mathbf{1}_A$ - qui prend la valeur 0 pour l'aliment B et 1 pour A). On a le graphique suivant pour $(x_{1,i}, y_i)$



On ajuste le modèle linéaire

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{1,i} x_{2,i} + \varepsilon_i$$

où on suppose les ε_i indépendants, et suivant une loi $\mathcal{N}(0, 1)$. On estime le modèle par moindres carrés. Au vu du graphique, quelle affirmation vous semble la plus valide

A) $\widehat{\beta}_1 < 0$ et $\widehat{\beta}_3 < 0$

B) $\widehat{\beta}_1 < 0$ et $\widehat{\beta}_3 > 0$

C) $\widehat{\beta}_1 = 0$ et $\widehat{\beta}_3 = 0$

D) $\widehat{\beta}_1 > 0$ et $\widehat{\beta}_3 < 0$

E) $\widehat{\beta}_1 > 0$ et $\widehat{\beta}_3 > 0$

7 On considère un modèle général, de la forme $y_i = f(\boldsymbol{x}_i, \boldsymbol{\beta}) + \varepsilon_i$, avec

$$\text{Var}[\varepsilon_i] = \frac{\sigma^2}{f(\boldsymbol{x}_i, \boldsymbol{\beta})^2}$$

où σ est une constante positive. Quelle transformation de la variable y permettra de stabiliser la variance (i.e. la rendre constante) ?

A) $y_i \to \log y_i$

B) $y_i \to \exp y_i$

C) $y_i \to 1/y_i$

D) $y_i \to \sqrt{y_i}$

E) $y_i \to y_i^2$

8 On dispose de la sortie suivante

```
Call:
lm(formula = rating ~ complaints + privileges + learning + raises + critical)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.011
complaints     0.692
privileges    -0.104
learning       0.249
raises        -0.033
critical       0.015
---

Residual standard error: 7.139 on *** degrees of freedom
```

Si on regarde, la première ligne $i = 1$ de la base est

```
ID rating complaints privileges learning raises critical
 1     43         51         30       39     61       92
```

et on nous dit que le premier élément de la matrice 'chapeau' $\boldsymbol{H}$ est $\boldsymbol{H}_{1,1} = 0.3234$. Quelle est la distance de Cook de cette observation?

A) moins de 0.1

B) entre 0.1 et 0.2

C) entre 0.2 et 0.3

D) entre 0.3 et 0.4

E) plus de 0.4

9 Si y_i et x_i sont deux variables continues, on appelle *élasticité* le ratio d'une variation relative de y par rapport à une variation relative de x, pour un individu i

$$e_{y|x}(i) = \frac{\partial y_i / y_i}{\partial x_i / x_i} = \frac{\partial \log y_i}{\partial \log x_i}$$

On dispose du modèle suivant

$$\log y_t = \beta_0 + \beta_1 \log x_{1,t} + \beta_2 \log x_{2,t} + + \beta_3 [\log x_{1,t} - \log x_{1,t_0}] \mathbf{1}_{>t_0}(t) + \beta_4 [\log x_{2,t} - \log x_{2,t_0}] \mathbf{1}_{>t_0}(t) + \varepsilon_t$$

où t désigne une annéee, entre 1979 et 2015, t_0 correspond à l'année 2000. Une estimation par moindres carrés donne

$$\widehat{\boldsymbol{\beta}} = (4.00, 0.60, -0.10, -0.07, -0.01)^\top$$

Quelle est l'élasticité de y par rapport à x_1 pour l'année 2010 ?

A) -0.11

B) 0.53

C) 0.60

D) 0.90

E) 1.70

10 On modèlise le coût de sinistres incendie, y, avec un modèle de la forme

$$\widehat{y}_i = 8 + 5x_{i,1} + 2(x_{i,1} - 4)x_{2,i} + 9x_{i,3} - 2x_{1,i}x_{3,i}$$

où x_1 est la distance à la plus proche caserne de pompiers (en kilomètres), $x_2 = \mathbf{1}_{\geqslant 4}(x_1)$ et x_3 prend la valeur 1 dans la ville A, 0 dans la ville B. Pour les incendies qui se déclarent à plus de 4 kilomètres d'une caserne de pompier, quelle est la distance à la plus proche caserne pour laquelle le coût moyen d'un incendie dans la ville A et dans la ville B sont identiques ?

A) 4.5 km

B) 6.5 km

C) 8.0 km

D) 28 km

E) 29 km

11 On considère un modèle avec deux variables explicatives,

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$$

On dispose de

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \begin{pmatrix} 6.1333 & -0.0733 & -0.1933 \\ -0.0733 & 0.0087 & -0.0020 \\ -0.1933 & -0.0020 & 0.0087 \end{pmatrix} \text{ et } \widehat{\sigma}^2 = 280.1167$$

Quel est l'écart-type de $\widehat{\beta}_1 - \widehat{\beta}_2$ (on retiendra la valeur la plus proche) ?

A) 1.92

B) 2.23

C) 2.45

D) 2.87

E) 3.11

12 On considère un modèle avec trois variables explicatives,

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i$$

et la sortie (partielle) de la régression sur $n = 49$ observations donne

```
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
```

Quelles variables peuvent être considérer comme significatives, pour un niveau de significativité $\alpha = 10\%$?

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 un modèle avec cinq variables explicatives,

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + \beta_5 x_{5,i} + \varepsilon_i$$

estimé sur $n = 20$ observations. La sortie (partielle) de la régression est

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)
x1          0.020000   0.012000   1.661
x2         -0.004950   0.008750  -0.565
x3          0.216000   0.043200   5.000
x4         -0.034600   0.115000  -0.301
x5         -0.000294   0.000141  -2.090
```

Combien de variables explicatives ne sont pas significatives pour un niveau de significativité $\alpha = 10\%$?

A) 1 variable n'est pas significative

B) 2 variables ne sont pas significatives

C) 3 variables ne sont pas significatives

D) 4 variables ne sont pas significatives

E) les 5 variables ne sont pas significatives

Le problème suivant est utilisé pour les questions 14 à 18.

On cherche une relation entre la taille des enfants et celle de leurs parents. Pour un individu i, de taille y_i (en pouces), on a un modèle

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$$

où $x_{1,i}$ est la taille de sa mère, et $x_{2,i}$ est la taille de son père. On dispose de deux sorties informatiques (partielles)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.2600    8.99400    ****
x1            0.52830    *******    3.88
x2            0.25485    0.09890    ****
```

et

```
            Df   Sum Sq Mean Sq F value   Pr(>F)
Regression  **  *******
Residuals   **   813.26
Total       20  1157.25
```

14 Quelle(s) variable(s) doit-on enlever de la régression (pour un niveau de significativité de 5%)

A) aucune

B) la constante

C) la taille de la mère

D) la taille du père

E) toutes

15 On cherche à tester (à l'aide de la sortie dont on dispose) $H_0 : \beta_2 = 0$ contre $H_1 : \beta_2 > 0$. Quelle serait la p-value de ce test ?

A) environ 0.5%

B) environ et 1%

C) environ 2%

D) environ 5%

E) environ 10%

16 On cherche à tester (à l'aide de la sortie dont on dispose) $H_0 : \beta_1 = \beta_2 = 0$ contre $H_1 : \beta_1 \neq 0$ ou $\beta_2 \neq 0$. Quelle est la statistique de Fisher de ce test ?

A) 2.5

B) 3.2

C) 3.8

D) 5.1

E) 7.6

17 Le père de Peter mesure 76 pouces, et sa mère 69. Quelle taille devrait faire Peter, selon notre modèle ? (on retiendra la valeur la plus proche)

A) 70

B) 71

C) 72

D) 73

E) 74

18 On rajoute dans notre modèle le genre de l'individu i, $x_{3,i} = 1$ si l'individu est un garçon, 0 pour une fille. Parmi les modèles suivant, lequel permet de tester si la taille globale des deux parents a un impact différent sur la taille de leur enfant suivant qu'ils ont un garçon ou une fille ?

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

B) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

C) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 x_1 + \varepsilon$

D) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 x_2 + \varepsilon$

E) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 x_1 + \beta_4 x_3 x_2 + \varepsilon$

19 On construit deux modèles, que l'on va estimer à l'aide de $n = 30$ observations

$$\text{modèle (A):} \quad y = \beta_0 + \beta_1 x_1 + \varepsilon$$

et

$$\text{modèle (B):} \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \eta$$

On nous donne

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = 160 \text{ et } \sum_{i=1}^{n}(x_{1,i} - \overline{x}_1)^2 = 10.$$

De plus, pour le modèle (A), $\widehat{\beta}_1 = -2$ alors que pour le modèle (B), $R^2 = 0.7$. Quelle est la valeur de la statistique de test F du test $H_0 : \beta_2 = \beta_3 = 0$?

A) moins de 22

B) entre 22 et 25

C) entre 25 et 27

D) entre 27 et 30

E) plus de 30

20 On considère un modèle linéaire simple, de la forme $y = \beta_0 + \beta_1 x + \varepsilon$. On suppose que la variance conditionnelle de ε est de la forme $\text{Var}[\varepsilon|x] = \sigma^2 x^{-1/2}$. Quel modèle permet de corriger de cette hétéroscédasticité ?

A) $yx^{1/4} = \beta_0 x^{1/4} + \beta_1 x^{5/4} + \eta$

B) $yx^{1/4} = \beta_0 + \beta_1 x^{5/4} + \eta$

C) $yx^{1/2} = \beta_0 x^{1/2} + \beta_1 x^{3/2} + \eta$

D) $yx^{-1/4} = \beta_0 x^{-1/4} + \beta_1 x^{3/4} + \eta$

E) $yx^{-1/2} = \beta_0 x^{-1/2} + \beta_1 x^{1/2} + \eta$

21 On considère le modèle suivant,

$$y_i = \exp\big[-(\beta_0 + \beta_1 x_i + \varepsilon_i)\big]$$

Donnez les estimateurs par moindres carrés de β_1 et β_0

A) $\widehat{\beta}_1 = \dfrac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$ et $\widehat{\beta}_0 = \dfrac{1}{n}\sum y_i - \widehat{\beta}_1 \dfrac{1}{n}\sum x_i$

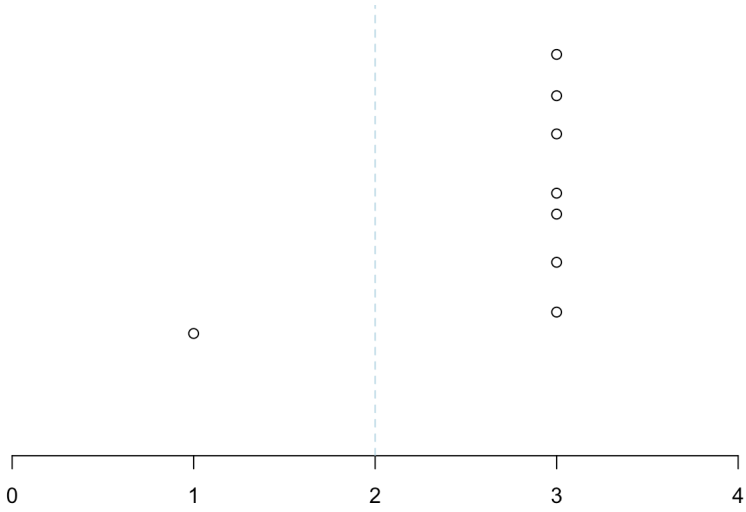
B) $\widehat{\beta}_1 = \dfrac{\sum(x_i - \overline{x})\log(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$ et $\widehat{\beta}_0 = \dfrac{1}{n}\sum y_i - \widehat{\beta}_1 \dfrac{1}{n}\sum x_i$

C) $\widehat{\beta}_1 = \dfrac{-\sum(x_i - \overline{x})\log(y_i)}{\sum(x_i - \overline{x})^2}$ et $\widehat{\beta}_0 = -\dfrac{1}{n}\sum \log(y_i) - \widehat{\beta}_1 \dfrac{1}{n}\sum x_i$

D) $\widehat{\beta}_1 = \dfrac{\sum \log(x_i - \overline{x})\log(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$ et $\widehat{\beta}_0 = -\dfrac{1}{n}\sum \log(y_i) - \widehat{\beta}_1 \dfrac{1}{n}\sum x_i$

E) $\widehat{\beta}_1 = \dfrac{-\sum(x_i - \overline{x})\log(y_i)}{\sum(x_i - \overline{x})^2}$ et $\widehat{\beta}_0 = \dfrac{1}{n}\sum y_i - \widehat{\beta}_1 \dfrac{1}{n}\sum x_i$

22 On dispose du jeu de données suivant



autrement dit on a un $(1, y_1)$ et 7 observations de la forme $(3, y_i)$. On note $\overline{y}$ la moyenne des y_i sur les 8 observations. Si on ajuste une droite de régression par moindres carrés, quelle serait la prévision pour $x = 2$?

A) $\frac{3}{8}y_1 + \frac{5}{8}\overline{y}$

B) $\frac{3}{7}y_1 + \frac{4}{7}\overline{y}$

C) $\frac{1}{2}y_1 + \frac{1}{2}\overline{y}$

D) $\frac{4}{7}y_1 + \frac{3}{7}\overline{y}$

E) $\frac{5}{8}y_1 + \frac{3}{8}\overline{y}$

23 On considère le modèle suivant,

$$y_i = \beta + \beta x_i + \varepsilon_i$$

Donnez l'estimateur par moindres carrés de β

A) $\widehat{\beta} = \frac{\sum y_i}{\sum x_i}$

B) $\widehat{\beta} = \frac{\sum y_i}{\sum(1 + x_i)}$

C) $\widehat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$

D) $\widehat{\beta} = \frac{\sum(1 + x_i) y_i}{\sum(1 + x_i)^2}$

E) $\widehat{\beta} = \frac{\sum(x_i - \overline{x}) y_i}{\sum(x_i - \overline{x})^2}$

Le problème suivant sert de base aux questions 24 et 25

On cherche à examiner le lien entre le salaire y et le nombre d'années d'expérience x_1, en fonction du genre x_2 (1 pour les hommes, 0 pour les femmes). On cherche à estimer le modèle

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{1,i} x_{2,i} + \varepsilon_i \tag{0}$$

à l'aide de 30 observations. On obtient un R^2 de 87%. On considère alors 6 modèles alternatifs plus simples,

	modèle	somme des carrés des résidus
(1)	$y_i = \beta_0 + \varepsilon_i$	423.58
(2)	$y_i = \beta_0 + \beta_1 x_{1,i} + \varepsilon_i$	75.69
(3)	$y_i = \beta_0 + \beta_2 x_{2,i} + \varepsilon_i$	381.23
(4)	$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_3 x_{1,i} x_{2,i} + \varepsilon_i$	68.74
(5)	$y_i = \beta_0 + \beta_2 x_{2,i} + \beta_3 x_{1,i} x_{2,i} + \varepsilon_i$	260.42
(6)	$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$	71.96

24 Calculez la statistique de test F pour tester si l'impact de l'expérience sur le salaire est identique pour les hommes et les femmes. (on retiendra la valeur la plus proche)

A) 2

B) 4

C) 5

D) 6

E) 8

25 Calculez la statistique de test F pour tester si la relation linéaire reliant l'expérience et le salaire est identique pour les hommes et les femmes. (on retiendra la valeur la plus proche)

A) 2

B) 4

C) 5

D) 6

E) 8

26 On nous donne la sortie suivante, obtenu en estimant plusieurs modèles à partir de 25 observations, avec une variable (continue) y et quatre variables explicatives potentielles x_1, x_2, x_3 et x_4. La colonne de droite est la somme des carrés des résidus

variables	SCR
constante	1.70167
constante et x_1	1.29764
constante et x_2	0.64834
constante et x_3	1.70113
constante et x_4	1.62735
constante, x_1 et x_2	0.14216
constante, x_1 et x_3	1.29512
constante, x_1 et x_4	1.28927

variables	SCR
constante, x_2 et x_3	0.62097
constante, x_2 et x_4	0.59745
constante, x_3 et x_4	1.61918
constante, x_1, x_2, et x_3	0.12133
constante, x_1, x_2, et x_4	0.14191
constante, x_1, x_4, et x_4	1.28384
constante, x_2, x_3, et x_4	0.58474
constante, x_1, x_2, x_3, et x_4	0.12089

On utilise une procédure itérative descendante de choix de variable (*backward*) basée sur le C_p de Mallow. Quel sera le modèle final retenu ?

A) aucune variable explicative (juste la constante)

B) la constante et x_2 (seulement)

C) la constante, $x1$ et x_2 (seulement)

D) la constante, x_1, x_2 et x_3 (seulement)

E) la constante et les quatre variables explicatives

Le problème suivant sert de base aux questions 27 et 28

On nous donne la sortie suivante, correspondant à l'estimation de sept modèles : on considère un sous-ensemble de nos trois variables explicatives (parmi x_1, x_2 et x_3) et on nous donne la statistique de test T pour un test de Student de $H_0 : \beta_j = 0$,

variables du modèle	T pour le test $H_0 : \beta_1 = 0$	T pour le test $H_0 : \beta_2 = 0$	T pour le test $H_0 : \beta_3 = 0$
x_1	3.00	-	-
x_2	-	-2.51	-
x_3	-	-	2.79
x_1 et x_2	2.61	-2.51	-
x_1 et x_3	3.00	-	2.41
x_2 et x_3	-	-2.70	2.51
x_1, x_2 et x_3	2.00	-2.61	2.90

27 On veut construire deux modèles :

- un modèle avec 2 variables explicatives avec une sélection descendante (*backward*)
- un modèle avec 1 variable explicative avec une sélection ascendante (*forward*)

Quels sont les modèles retenus ?

	2 variables (*backward*)	1 variable (*forward*)
A)	x_1 et x_2	x_1
B)	x_1 et x_2	x_2
C)	x_1 et x_3	x_2
D)	x_1 et x_3	x_3
E)	x_2 et x_3	x_1

28 On veut construire deux modèles :

- un modèle avec 2 variables explicatives avec une sélection ascendante (*forward*)
- un modèle avec 1 variable explicative avec une sélection descendante (*backward*)

Quels sont les modèles retenus ?

	2 variables (*forward*)	1 variable (*backward*)
A)	x_1 et x_2	x_1
B)	x_1 et x_2	x_2
C)	x_1 et x_3	x_2
D)	x_1 et x_3	x_3
E)	x_2 et x_3	x_1

Le problème suivant sert de base aux questions 29 et 30

On considère le modèle noté (12) $y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$. On dispose de

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \begin{pmatrix} 25.0487 & -0.8457 & 0.2864 \\ -0.8457 & 0.0294 & -0.0104 \\ 0.2864 & -0.0104 & 0.0040 \end{pmatrix}$$

Le modèle estimé est $\widehat{y} = 34.5 - 0.304x_1 + 0.383x_2$ et on a la sortie (partielle) suivante

```
            Df   Sum Sq Mean Sq F value   Pr(>F)
Regression  **   270.09
Residuals   **   ******
Total       5    290.00
```

29 Que vaut la statistique de Student associé au test $H_0 : \beta_1 = 0$

A) moins de -2

B) entre -1.5 et -1

C) entre -1 et -0.5

D) entre -0.5 et 0

E) plus de 0

30 On considère maintenant le modèle $y_i = \beta_0 + \beta_2 x_{2,i} + \varepsilon_i$, noté modèle (2). Que vaut la variation du $\overline{R}^2$ entre (12) et (2) ?

A) il diminue de 0.025

B) il diminue de 0.015

C) il ne varie pas (différence absolue infèrieure à 0.01)

D) il augmente de 0.015

E) il augmente de 0.025