STT5100 - Automne 2022 - Examen Intra (OLS)

Arthur Charpentier

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. Il est toutefois recommandé de lire attentivement les questions avant de tenter d'y répondre. 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 votre 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 présent document contient 20 pages, inclutant 4 pages de tables de lois usuelles (Gaussienne, Student, chi-deux et Fisher) à la fin.

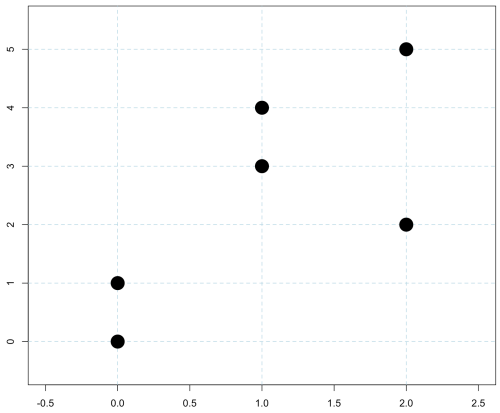
Le surveillant ne répondra à aucune question durant l'épreuve : en cas de soucis sur une question (interprétation possiblement fausse, typo, etc), vous pouvez mettre un court commentaire sur la feuille de réponses.

Code permanent : _____

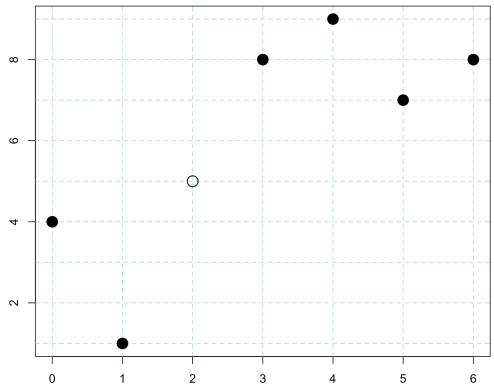
énoncé A

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	Figure à droite (à compléter)				
question 6	☐ A	☐ B	☐ C	☐ D	☐ E
question 7	Figure à droite (à compléter)				
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 ?*				

question 5 :



question 7 :



_____

1 On a estimé un modèle de régression simple, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, par moindres carrés et on a obtenu $\widehat{\beta}_1 = 0$. Alors

A) $R^2 = \overline{y}$

B) $R^2 = 1$

C) $R^2 = 0$

D) $R^2 = \text{Var}[y]$

E) aucune des affirmations ci-dessus

Comme $\widehat{\beta}_1 = 0$, toutes les prédictions sont identiques, $\widehat{y}_i = \widehat{\beta}_0$ (ainsi que la moyenne des $\widehat{y}_i$, et la moyenne des y_i, $\overline{y}$) et donc SCE $= \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 = \sum_{i=1}^{n} (\widehat{\beta}_0 - \widehat{\beta}_0)^2 = 0$, et donc $R^2 = 0$ (puisque $R^2 = \text{SCE/SCT}$) C'est la réponse C.

2 Pour obtenir l'estimateur de la pente, dans une régression simple, en utilisant le principe des moindres carrés, vous divisez

A) la variance d'échantillon de x par la variance d'échantillon de y

B) la covariance d'échantillon de x et y par la variance d'échantillon de y

C) la covariance d'échantillon de x et y par la variance d'échantillon de x.

D) la variance d'échantillon de x par la covariance d'échantillon de x et y.

E) la moyenne d'échantillon de y par la moyenne d'échantillon de x

Je renvoie au cours

3 On a estimé un modèle $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, sur un premier échantillon. On a obtenu

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = 50, \sum_{i=1}^{n} x_i = 10, \sum_{i=1}^{n} x_i^2 = 100, \widehat{\beta}_0 = 2 \text{ et } \widehat{\beta}_1 = 1$$

Sur un autre échantillon de même taille, on a obtenu

$$\sum_{i=1}^{n} \tilde{\varepsilon}_i^2 = 80, \sum_{i=1}^{n} x_i = 10, \sum_{i=1}^{n} x_i^2 = 100, \tilde{\beta}_0 = 2 \text{ et } \tilde{\beta}_1 = 1$$

Que peut-on dire sur les statistiques de test t pour nos différents estimateurs (estimés par moindres carrés)

A) $t_{\widehat{\beta}_0} \leqslant t_{\tilde{\beta}_0}$ et $t_{\widehat{\beta}_1} \leqslant t_{\tilde{\beta}_1}$

B) $t_{\widehat{\beta}_0} \leqslant t_{\tilde{\beta}_0}$ et $t_{\widehat{\beta}_1} \geqslant t_{\tilde{\beta}_1}$

C) $t_{\widehat{\beta}_0} \geqslant t_{\tilde{\beta}_0}$ et $t_{\widehat{\beta}_1} \leqslant t_{\tilde{\beta}_1}$

D) $t_{\widehat{\beta}_0} \geqslant t_{\tilde{\beta}_0}$ et $t_{\widehat{\beta}_1} \geqslant t_{\tilde{\beta}_1}$

E)$t_{\widehat{\beta}_0} = t_{\tilde{\beta}_0}$ et $t_{\widehat{\beta}_1} = t_{\tilde{\beta}_1}$

Comme les estimateurs sont identiques ($\widehat{\beta}_1 = \tilde{\beta}_1$ et $\widehat{\beta}_0 = \tilde{\beta}_0$), comparer les statistiques de test revient à comparer les variances : comme les statistiques sont positives (elles sont toujours du même signe que les estimateurs, qui sont ici positifs), une statistique plus grande correspond à une variance plus petite.

Si on utilise une écriture matricielle, on sait que la variance de nos estimateurs s'écrit $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. Or comme la somme des X_i et la somme des X_i^2 est la même, le terme de droite $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ ne change pas. Aussi, ce qu'il faut comparer, ce sont juste les estimations de variance des résidus, σ^2. Comme les résidus sont centrés, et que n est la même, on compare la somme des carrés des résidus. Qui est plus grande pour le second modèle que pour le premier. Aussim

$$\text{Var}(\tilde{\beta}_0) > \text{Var}(\hat{\beta}_0) \text{ et } \text{Var}(\tilde{\beta}_1) > \text{Var}(\hat{\beta}_1)$$

de telle sorte que

$$t_{\hat{\beta}_0} \geqslant t_{\tilde{\beta}_0} \text{ et } t_{\hat{\beta}_1} \geqslant t_{\tilde{\beta}_1}$$

qui est la réponse D.

4 On ajuste un modèle $y = \beta_0 + \beta_1 x + \varepsilon$ sur $n = 100$ observations, où x est une variable prenant les valeurs 0 ou 1. Dans 40% des cas, x_i a pris la valeur 1. On nous dit que

$$\widehat{\beta}_1 = 1.4 \text{ et } \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = 920.$$

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

A) 1.15

B) 1.78

C) 2.26

D) 2.46

E) 3.51

L'estimateur (classique) de σ^2 est

$$\widehat{\sigma}^2 = \frac{1}{100-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{920}{98} \sim 9.2$$

Pour avoir l'écart type de notre estimateur $\widehat{\beta}_1$, il nous manque le terme

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

Comme ici x_i prend les valeurs 0 ou 1, $x_i^2 = x_i$. Donc la relation précédante se simplifie,

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i - \overline{x}^2 = n\overline{x} - n\overline{x}^2 = 40 - 16 = 24$$

(qui est juste la variance d'une loi binomiale, $n\overline{x}(1-\overline{x})$). Aussi, l'écart type de notre estimateur $\widehat{\beta}_1$ s'écrit

$$\sqrt{\text{Var}(\widehat{\beta}_1)} = \sqrt{\frac{\widehat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \sqrt{\frac{9.2}{24}} = 0.6191.$$

Aussi, la statistique de test est

$$t = \frac{\widehat{\beta}_1}{\sqrt{\text{Var}(\widehat{\beta}_1)}} = \frac{1.4}{0.6191} = 2.2612$$

qui correspond à la réponse C.

5 Sur la Figure de la page 2, tracez très exactement la droite de régression (estimée par moindres carrés), sachant qu'elle passe par (au moins) un des points.

$\boldsymbol{x}$	0	0	1	1	2	2
$\boldsymbol{y}$	0	1	3	4	2	5

On sait (c'est dans le cours) que la droite estimée par moindres carrés passe forcément par $(\overline{x}, \overline{y}$ (c'est la condition du premier ordre obtenue en dérivant par rapport à β_0). Ici, calculer les deux moyennes était facile

$$\boldsymbol{x} = \{0, 0, 1, 1, 2, 2\} \text{ donc } \overline{x} = 1$$

$$\boldsymbol{x} = \{0, 1, 2, 3, 4, 5\} \text{ (dans le désordre) donc } \overline{x} = \frac{5}{2} = 2.5$$

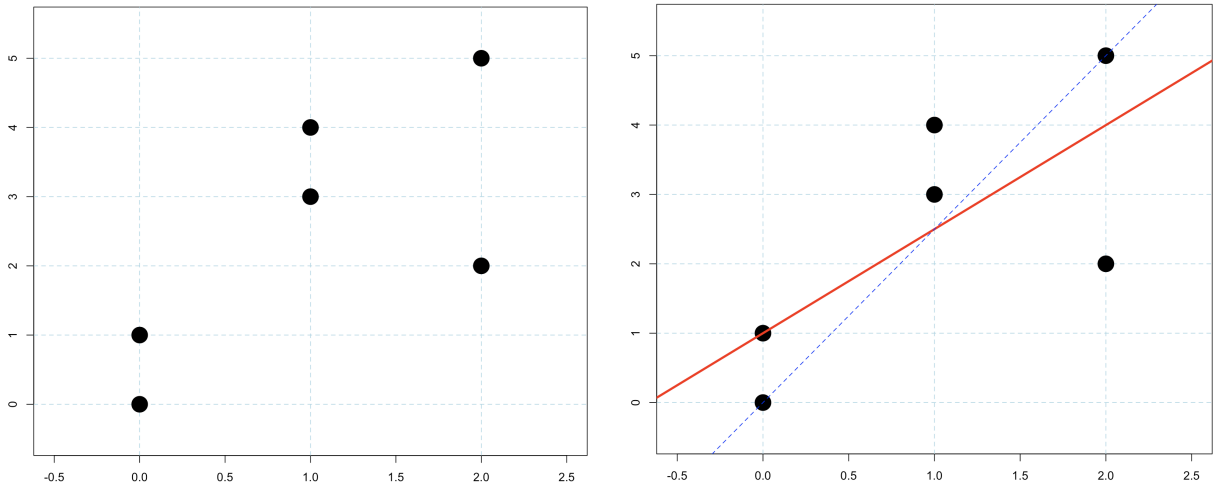
Comme la droite passe par $(1; 2.5)$, il est exclus qu'elle passe par les deux points au centre. On a alors 3 possibilités

(i) passer par le point en bas à droite, $(2, 2)$: ce cas est exclus, les résidus sont trop importants, et la pente est négatif, ce qui n'a pas trop de sens... (mais les laisses les sceptiques faire les calculs);

(ii) passer par le point en haut à gauche, $(0, 1)$;

(iii) passer par les deux autres points, $(0, 0)$ et $(2, 5)$ (car ces points sont alignés avec $(1; 2.5)$);

Ces deux derniers cas sont visualisables ci-dessous



La première chose qu'on peut noter est que, quelle que soit la droite retenue, les résidus aux centres sont identiques. La seconde chose qu'on peut noter est que, quelle que soit la droite retenue, les carrés des résidus à gauche sont identiques, avec soit 0 et $(+1)^2$, soit 0 et $(-1)^2$. Donc le meilleur modèle sera celui qui a les résidus dont la somme des carrés sera la plus faible...

Pour les deux points, si on retient la droite *rouge*, la prévision (en $x = 2$) sera $\widehat{y} = 4$. Donc les résidus sont respectivement $+1$ et -2. Si on retient la droite *bleue*, la prévision (en $x = 2$) sera $\widehat{y} = 5$. Donc les résidus sont respectivement 0 et -3. Or

$$\underbrace{(+1)^2 + (-2)^2 = 5}_{\textit{rouge}} < \underbrace{(0)^2 + (-3)^2 = 9}_{\textit{bleue}}$$

donc la courbe rouge est celle qui correspond à la plus petite somme des carrés des résidus, on va donc retenir ce modèle. L'estimation de la courbe de régression par moindres carrés donne le modèle rouge.

6 On dispose d'un jeu de données $\{(x_1, y_1), \dots, (x_n, y_n)\}$. On considère deux modèles

$$y_i = bx_i + u_i \text{ et } x_i = ay_i + v_i$$

avec les conditions usuelles pour les deux modèles (en particulier $\mathcal{H}_2$). On consière les estimateurs par moindres carrés de a et b. Quelle condition vérifient-ils ?

A) $\widehat{a} \cdot \widehat{b} = 1$

B) $\widehat{b} \sum_{i=1}^{n} x_i = \widehat{a} \sum_{i=1}^{n} x_i$

C) $\widehat{b} \sum_{i=1}^{n} x_i = \widehat{a} \sum_{i=1}^{n} y_i$

D) $\widehat{b} \sum_{i=1}^{n} y_i^2 = \widehat{a} \sum_{i=1}^{n} x_i^2$

E) $\widehat{b} \sum_{i=1}^{n} x_i^2 = \widehat{a} \sum_{i=1}^{n} y_i^2$

Pour la première équation ($y_i = bx_i + u_i$), l'estimateur par moindres carrés de b est solution du problème

$$\widehat{b} = \text{argmin} \left\{ \sum_{i=1}^{n} (y_i - bx_i)^2 \right\},$$

dont la condition du premier ordre est

$$\frac{\partial}{\partial b} \sum_{i=1}^{n} (y_i - bx_i)^2 \Big|_{b=\widehat{b}} = \sum_{i=1}^{n} \frac{\partial (y_i - bx_i)^2}{\partial b} \Big|_{b=\widehat{b}} = \sum_{i=1}^{n} 2x_i (y_i - \widehat{b} x_i) = 0$$

donc

$$\widehat{b} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i,$$

et on va s'arrêter là. Parce que si on regarde le second modèle, et qu'on cherche l'estimateur par moindres carrés de a, on va aboutir à l'équation

$$\widehat{a} \sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} y_i x_i,$$

et donc, en égalisant, on obtient

$$\widehat{b} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i = \widehat{a} \sum_{i=1}^{n} y_i^2,$$

qui est l'expression E.

7 On dispose de la base de données suivantes

$\boldsymbol{x}$	0	1	2	3	4	5	6
$\boldsymbol{y}$	4	1	5	8	9	7	8

La régression correspond à la sortie suivante

```
> summary(lm(y~x))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0000     1.6437   1.825   0.1420
x             1.0000     0.4317   2.317   0.0814 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.236 on 4 degrees of freedom
Multiple R-squared:  0.573,Adjusted R-squared:  0.4662
```

On nous demande d'enlever la troisième observation, $(x_i, y_i) = (2, 5)$, et de refaire la régression. Tracez sur la figure la nouvelle droite de régression (obtenue sur la base sans la 3ème observation).

La réponse rapide est simple: la troisième observation est sur la droite de régression estimée par moindres carrés, puisque $y_i = 5 = 3 + 1 \cdot 2 = 3 + 1 \cdot x_i$, donc enlever ce point ne change strictement rien, la droite de régression restera la même. On peut s'en convaincre mathématiquement en notant que cette nouvelle régression sera obtenue en résolvant le problème

$$\min \left\{ \sum_{i \in \{1,2,4,5,6,7\}} \left(y_i - \beta_0 - \beta_1 x_i\right)^2 \right\} \qquad (A)$$

or la première régression était obtenue en résolvant

$$\min \left\{ \sum_{i \in \{1,2,3,4,5,6,7\}} \left(y_i - \beta_0 - \beta_1 x_i\right)^2 \right\} = \sum_{i \in \{1,2,3,4,5,6,7\}} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \qquad (B)$$

(car $\min\{f(x)\} = f(x^\star)$), or

$$(B) = \underbrace{\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2}_{} = 0 + \min \left\{ \sum_{i \in \{1,2,4,5,6,7\}} \left(y_i - \beta_0 - \beta_1 x_i\right)^2 \right\} = (A)$$

8 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 (avec un effet croisé)

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

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 moindres 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

Ce n'est pas moi qui ait écrit cette question, il s'agit de la question 37 de l'examen du printemps 2018 CAS MAS-I. Notons ici que $z_1 = 1 - x_1$ et $z_2 = 1 - x_2$, ou réciproquement, que $x_1 = 1 - z_1$ et $x_2 = 1 - z_2$. Aussi, si on remplace dans le premier modèle

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{1,i} x_{2,i} + \varepsilon_i \\ &= \beta_0 + \beta_1(1 - z_{1,i}) + \beta_2(1 - z_{2,i}) + \beta_3(1 - z_{1,i})(1 - z_{2,i}) + \varepsilon_i \\ &= [\beta_0 + \beta_2 + \beta_2 + \beta_3] - (\beta_1 + \beta_3) z_{1,i} - (\beta_2 + \beta_3) z_{2,i} + \beta_3 z_{1,i} z_{2,i} + \varepsilon_i \end{aligned}$$

On peut alors identifier simplement les 4 termes,

$$\begin{cases} \widehat{\alpha}_0 = \beta_0 + \beta_2 + \beta_2 + \beta_3 = -0.10 - 0.25 + 0.58 - 0.20 = 0.03 \\ \widehat{\alpha}_1 = -(\beta_1 + \beta_3) = -(-0.25 - 0.20) = 0.45 \\ \widehat{\alpha}_2 = -(\beta_2 + \beta_3) = -(0.58 - 0.20) = -0.38 \\ \widehat{\alpha}_3 = \beta_3 = -0.20 \end{cases}$$

Comparons maintenant nos paires d'estimateurs

$$\begin{cases} \widehat{\alpha}_0 = +0.03 & \widehat{\beta}_0 = -0.10 & \textit{différents} \\ \widehat{\alpha}_1 = +0.45 & \widehat{\beta}_1 = -0.25 & \textit{différents} \\ \widehat{\alpha}_2 = -0.38 & \widehat{\beta}_2 = +0.58 & \textit{différents} \\ \widehat{\alpha}_3 = -0.20 & \widehat{\beta}_3 = -0.20 & \textit{identiques} \end{cases}$$

autrement dit, seule la paire $(\widehat{\alpha}_3, \widehat{\beta}_3)$ ne change pas, les autres étant différents, indépendamment du signe. C'est la réponse E.

9 En utilisant 143 observations, on a estimé une fonction de régression simple. L'estimation de la pente vaut 0.04, avec un écart-type de 0.01. Laquelle des décisions possibles suivantes est la seule correcte ?

A) le coefficient est petit et qu'il est donc très probablement nul dans la population

B) la pente est statistiquement significative puisqu'elle est éloignée de zéro par quatre fois l'écart-type.

C) comme la pente est très faible, le R^2 de régression doit l'être aussi

D) si la constante est proche de 0, comme la pente est très faible, le R^2 de régression doit l'être aussi

E) comme la pente est positive, la constante sera aussi positive

A) est fausse : ce n'est pas la valeur du coefficient qui est importante, mais sa valeur relative. En effet, en changeant d'échelle pour x, par exemple en prenant une unité 10 fois plus grande, la valeur du coefficient devient mécaniquement 10 fois plus petite, sans pour tant changer la significativité de la variable. C) est fausse, pour la même raison que précédemment : changer l'unité de x ne change pas la valeur du R^2 D) on ne parle pas de la constante et E) n'a pas grand sens.

En revanche, B) a du sens : la statistique du test de Student est $t = \widehat{\beta}/\sqrt{\widehat{\text{Var}}(\widehat{\beta})}$, soit ici $t = 4$. Effectivement, $t = 4$ revient à rejeter l'hypothèse $H_0 : \beta = 0$ car la p-value est très très faible : se tromper en rejetant à tort H_0 surviendra avec une probabilité très très faible. Donc B est juste.

10 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

Là encore, j'ai repris une question des examens professionnels (ici, c'était la question 4 de SOA Course 120 Study Notes (210-83-98)). Rappelons que la variance de $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$ est

$$\text{Var}[\widehat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \text{ et donc } \widehat{\text{Var}}[\widehat{\boldsymbol{\beta}}] = \widehat{\sigma}^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}.$$

Aussi, la variance de $\widehat{\beta}_1 - \widehat{\beta}_2$ est

$$\text{Var}[\widehat{\beta}_1 - \widehat{\beta}_2] = \text{Var}[\widehat{\beta}_1] - 2\text{Cov}[\widehat{\beta}_1, \widehat{\beta}_2] + \text{Var}[\widehat{\beta}_2]$$

dont un estimateur naturel est

$$\widehat{\text{Var}}[\widehat{\beta}_1 - \widehat{\beta}_2] = \widehat{\text{Var}}[\widehat{\beta}_1] - 2\widehat{\text{Cov}}[\widehat{\beta}_1, \widehat{\beta}_2] + \widehat{\text{Var}}[\widehat{\beta}_2],$$

soit, numériquement

$$\widehat{\text{Var}}[\widehat{\beta}_1 - \widehat{\beta}_2] = 280.1167 \times \big(0.0087 - 2 \times (-0.0020) + 0.0087\big) = 5.9944974,$$

donc l'écart-type (estimé) de $\widehat{\beta}_1 - \widehat{\beta}_2$ sera

$$\sqrt{\widehat{\text{Var}}[\widehat{\beta}_1 - \widehat{\beta}_2]} = \sqrt{5.9944974} \sim 2.4484$$

dont la valeur la plus proche est 2.45.

11 De manière générale, on considère le modèle $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, où on suppose ε_i centré, de variance constante, et indépendants les uns des autres. On propose plusieurs estimateurs pour β_1,

$$\widehat{\beta}_1^{(1)} = \frac{\overline{y}}{\overline{x}}, \widehat{\beta}_1^{(2)} = \frac{y_2 - y_1}{x_2 - x_1} \text{ et } \widehat{\beta}_1^{(3)} = \frac{\max\{y_i\} - \min\{y_i\}}{\max\{x_i\} - \min\{x_i\}}.$$

A) $\widehat{\beta}_1^{(1)}$ est un estimateur sans biais de β_1

B) $\widehat{\beta}_1^{(2)}$ est un estimateur sans biais de β_1

C) $\widehat{\beta}_1^{(3)}$ est un estimateur sans biais de β_1

D) les trois sont des estimateurs sans biais de β_1

E) aucun n'est un estimateur sans biais de β_1

Montrons que A) est fausse (ce qui exclue A et D) et que B) est correcte (ce qui exclura E... et C car il faut une seule réponse).

Pour le premier estimateur,

$$\mathbb{E}(\widehat{\beta}_1^{(1)}) = \mathbb{E}\left(\frac{\overline{Y}}{\overline{x}}\right) = \mathbb{E}\left(\frac{\beta_0 + \beta_1\overline{x} + \overline{\varepsilon}}{\overline{x}}\right) = \frac{\beta_0}{\overline{x}} + \beta_1 + 0 \neq \beta_1$$

donc non : $\widehat{\beta}_1^{(1)}$ n'est pas un estimateur sans biais de β_1 (dans le cas général où $\beta_0 \neq 0$).

On nous dit que $\widehat{\beta}_1^{(2)} = \frac{y_2 - y_1}{x_2 - x_1}$ et il faut calculer son biais, donc on va voir $\widehat{\beta}_1^{(2)}$ comme une variable aléatoire (et remplace y_i par Y_i), i.e.

$$\mathbb{E}(\widehat{\beta}_1^{(2)}) = \mathbb{E}\left(\frac{Y_2 - Y_1}{x_2 - x_1}\right) = \mathbb{E}\left(\frac{(\beta_0 + \beta_1 x_2 + \varepsilon_2) - (\beta_0 + \beta_1 x_1 + \varepsilon_1)}{x_2 - x_1}\right)$$

bref

$$\mathbb{E}(\widehat{\beta}_1^{(2)}) = \beta_1 \frac{x_2 - x_1}{x_2 - x_1} + \underbrace{\mathbb{E}\left(\frac{\varepsilon_2 - \varepsilon_1}{x_2 - x_1}\right)}_{=0} = \beta_1$$

donc oui : $\widehat{\beta}_1^{(2)}$ est un estimateur sans biais de β_1.

12 On fait un test de Student sur une des variables explicatives dans une régression multiple, $H_0 : \beta_j = 0$. Une très large p-value (de l'ordre de 95%) signifie

A) qu'on rejette H_0

B) que la valeur absolue de la statistique de test est grande

C) que $|\widehat{\beta}_j|$ est grand

D) que $|\widehat{\beta}_j|$ est plus petit que 10% de $\sqrt{\text{Var}(\widehat{\beta}_j)}$

E) que si on supprime la j-ième variable de la régression, le R^2 augmentera

L'affirmation D) se traduit par le fait que t_j est plus petit que 10%. En effet,

$$t = \frac{\widehat{\beta}_1}{\sqrt{\text{Var}(\widehat{\beta}_1)}}$$

et on nous dit que $\mathbb{P}(|T| > |t|) \sim 0.95$ or $\mathbb{P}(|T| > |t|) = 2 \cdot \mathbb{P}(T > |t|)$, soit $\mathbb{P}(T > |t|) = 0.475$ ou $\mathbb{P}(T \leqslant |t|) = 1 = 0.475 = 0.525$ où T suit une loi normale, donc Or le quantile de niveau 0.525 est de l'ordre de 0.13 qui est la valeur de $|t|$. Autrement dit le ratio de $|\widehat{\beta}_1|$ sur $\sqrt{\text{Var}(\widehat{\beta}_j)}$ est environ de l'ordre de 0.13, ce qui est de l'ordre de 10%. Donc D serait acceptable, et je vous laisse vérifier que les autres affirmation sont fausses.

13 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 $\widehat{\beta}_0 = \dfrac{1}{n}\sum y_i - \widehat{\beta}_1 \dfrac{1}{n}\sum x_i$

Il s'agissait de la question 3 du Study Note 120-81-95 du cours 120 de la SOA. Le cours s'appelle *modèles linéaires appliqués* donc on va rendre ce modèle linéaire :

$$y_i = \exp\big[-(\beta_0 + \beta_1 x_i + \varepsilon_i)\big] \text{ ou } \underbrace{-\log[y_i]}_{\tilde{y}_i} = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Il suffit alors d'appliquer le cours : l'estimateur de la pente est

$$\widehat{\beta}_1 = \frac{\text{Cov}[x, \tilde{y}]}{\text{Var}[x]} = \frac{\sum(x_i - \overline{x})(\tilde{y}_i - \overline{\tilde{y}})}{\sum(x_i - \overline{x})^2} = \frac{\sum(x_i - \overline{x})(-\log[y_i] - \overline{-\log[y]})}{\sum(x_i - \overline{x})^2}$$

notons ici que

$$\sum_{i=1}^{n}(x_i - \overline{x}) \cdot \overline{-\log[y]} = \overline{-\log[y]} \cdot \underbrace{\sum_{i=1}^{n}(x_i - \overline{x})}_{=0} = 0$$

donc on peut simplifier

$$\widehat{\beta}_1 = -\frac{\sum(x_i - \overline{x})\log[y_i]}{\sum(x_i - \overline{x})^2}$$

On retrouve cette expression dans C et E. Pour la constante, on a comme toujours la condition du premier ordre qui garantit que

$$\overline{\tilde{y}} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x}$$

donc

$$\widehat{\beta}_0 = \overline{\tilde{y}} - \widehat{\beta}_1 \overline{x} = -\frac{1}{n}\sum_{i=1}^{n} \log y_i - \widehat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i$$

que l'on retrouve dans C et D. Aussi, la bonne réponse est C.

14 On considère un modèle estimé par moindres carrés $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. On veut faire une prévision pour un nouveau point $x^\star$, et on pose

$$\widehat{y}^\star = \widehat{\beta}_0 + \widehat{\beta}_1\ x^\star.$$

Pour quel valeur de $x^\star$ la variance de $\widehat{Y}^\star$ sera-t-elle minimale ?

A) quand $x^\star = 0$

B) quand $x^\star = \overline{x}$

C) quand $x^\star = \min\{x_1, \cdots, x_n\}$

D) quand $x^\star$ est l'abscisse du point $(x^\star, y^\star)$ où $y^\star = \min\{y_1, \cdots, y_n\}$

E) aucune des réponses proposées

C'est un point qui a été discuté en cours (et je ne demandais pas de le prouver). Mais on peut (re)faire le calcul. On sait que la variance de $\widehat{Y}^\star$ est

$$\text{var}[\widehat{Y}^\star] = \text{var}[\widehat{\beta}_0 + \widehat{\beta}_1\ x^\star] = \text{var}[\widehat{\beta}_0] + 2\text{cov}[\widehat{\beta}_0, \widehat{\beta}_1\ x^\star] + \text{var}[\widehat{\beta}_1\ x^\star]$$

On cherche alors le minimum de la fonction

$$x^\star \mapsto \text{var}[\widehat{\beta}_0] + 2\text{cov}[\widehat{\beta}_0, \widehat{\beta}_1]x^\star + \text{var}[\widehat{\beta}_1]x^{\star 2}$$

que l'on peut simplement dériver (condition du premier ordre)

$$2\text{cov}[\widehat{\beta}_0, \widehat{\beta}_1] + 2\text{var}[\widehat{\beta}_1]x^\star$$

autrement dit, le minimum est obtenu pour

$$x^\star = -\frac{\text{cov}[\widehat{\beta}_0, \widehat{\beta}_1]}{\text{var}[\widehat{\beta}_1]}$$

Or on sait que

$$\text{Var}[\widehat{\boldsymbol{\beta}}] = \sigma^2(\boldsymbol{X}\,\boldsymbol{X})^{-1} = \frac{\sigma^2}{\sum(x_i - \overline{x})^2}\begin{pmatrix} \frac{1}{n}\sum x_i^2 & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix} = \begin{pmatrix} \text{var}[\widehat{\beta}_0] & \text{cov}[\widehat{\beta}_0, \widehat{\beta}_1] \\ \text{cov}[\widehat{\beta}_0, \widehat{\beta}_1] & \text{var}[\widehat{\beta}_1] \end{pmatrix}$$

donc

$$x^\star = -\frac{-\overline{x}}{1} = \overline{x}$$

qui est la réponse B.

Le problème suivant sert de base aux questions 15 et 16

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

15 Que vaut le R^2 pour le modèle (6) ?

A) moins de 60%

B) entre 60% et 70%

C) entre 70% et 80%

D) entre 80% et 90%

E) plus de 90%

Le modèle (1) nous permet d'avoir la somme des carrés totaux, $\sum_{i=1}^{n}(y_i - \overline{y})^2 = 423.48 = SCR_1 = SCT$. Or

$$R^2 = 1 - \frac{SCR}{SCT} \text{ soit pour le modèle 6 } 1 - \frac{71.96}{423.48} = 83\%$$

qui est la réponse D.

16 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

Tester "*si l'impact de l'expérience sur le salaire est identique pour les hommes et les femmes*" revient à se demander si $\beta_2 = 0$. Classiquement on pourrait faire un test de Student (c'est un test simple) mais on nous demande ici d'utiliser un test de Fisher. Le modèle non-contraint est (bien entendu) le modèle (0) alors que le modèle contraint est

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

qui correspond au modèle (6). La statistique de test de $H_0 : \beta_3 = 0$ est alors ici

$$F = \frac{SCR_6 - SCR_0}{1} \cdot \frac{30 - 4}{SCR_0} = \frac{71.96 - 55.06}{1} \cdot \frac{26}{55.06} = 7.9771$$

1 car on teste ici une seule valeur (β_3) et $30 - 4$ car on a ici 30 observations, et 4 variables dans le modèle (0). En arrondissant, on obtient la statistique de test de la réponse E.

17 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

La question là encore n'est pas de moi, il s'agissait de la question 9 de l'examen SOA Course 4 du printemps 2000. Pour le modèle (A) la somme des carrés expliqués est

$$\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{\beta}_1(x_i - \overline{x}))^2$$

(en écrivant le modèle sous la forme $y_i - \overline{y} = \beta_1(x_{1,i} - \overline{x}) + \varepsilon$). Aussi

$$\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 = \widehat{\beta}_1^2 \cdot \sum_{i=1}^{n}(x_i - \overline{x})^2 = (-2)^2 \cdot 10 = 40 = \text{SCE}_A.$$

Pour le modèle (B), pour calculer la somme des carrés expliqués, on utilise le R^2, en rappelant que

$$R^2 = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \text{ donc } \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 = R^2 \cdot \sum_{i=1}^{n}(y_i - \overline{y})^2 = 0.7 \cdot 160 = 112 = \text{SCE}_B.$$

On peut alors calculer la statistique de Fisher,

$$F = \frac{\text{SCR}_B - \text{SCR}_A}{2} \cdot \frac{30-3-1}{\text{SCR}_B} = \frac{\text{SCE}_A - \text{SCE}_B}{2} \cdot \frac{30-3-1}{\text{SCT}_B - \text{SCE}_B} = \frac{112-40}{2} \cdot \frac{26}{160-112} = 19.5$$

ce qui correspond à la réponse A.

18 Toujours pour ce modèle linéaire simple, on suppose que β_0 est connue, vaut 2, et on estime le modèle suivant

$$y_i = 2 + \beta_1 x_i + \varepsilon_i \tag{1}$$

par moindres carrés. On note $\widetilde{\beta}_1$ l'estimateur de β_1. Que vaut $\widetilde{\beta}_1$?

A) $\widetilde{\beta}_1 = \dfrac{\sum(x_i - \overline{x})y_i}{\sum(x_i - \overline{x})^2}$

B) $\widetilde{\beta}_1 = \dfrac{\sum(y_i - 2)}{\sum(x_i - \overline{x})}$

C) $\widetilde{\beta}_1 = \dfrac{\sum x_i(y_i - 2)}{\sum(x_i - \overline{x})^2}$

D) $\widetilde{\beta}_1 = \dfrac{\sum x_i(y_i - 2)}{\sum x_i^2}$

E) aucune des réponses proposées

On commence par écrire la fonction objectif, qui est la somme des carrés des erreurs

$$\beta_1 \mapsto \sum_i (y_i - 2 - \beta_1 x_i)^2$$

La condition du premier ordre s'écrit alors

$$\sum_i -2x_i(y_i - 2 - \widetilde{\beta}_1 x_i) = 0 \text{ soit } \sum_i x_i(y_i - 2) = \sum_i \widetilde{\beta}_1 x_i^2$$

soit $\widetilde{\beta}_1 = \dfrac{\sum x_i(y_i-2)}{\sum x_i^2}$ qui est la réponse D.

19 On considère le modèle suivant,

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

Donnez l'estimateur par moindres carrés de β

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

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

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

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

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

Il s'agissait de la question 5 du l'examen *applied statistics* de la CAS de l'été 2005. Comme souvent, il faut faire un peu de réécriture

$$y_i = \beta + \beta x_i + \varepsilon_i = y_i = \beta \underbrace{(1+x_i)}_{\tilde{x}_i} + \varepsilon_i$$

On a un modèle sans constante, et l'estimateur par moindres carrés est alors obtenu en regardant la condition du premier ordre du problème

$$\min\left\{\sum_{i=1}^n (y_i - \beta\tilde{x}_i)^2\right\}$$

soit

$$2\sum_{i=1}^n(-\tilde{x}_i)(y_i - \widehat{\beta}\tilde{x}_i) = 0 \text{ soit } \widehat{\beta} == \frac{\sum \tilde{x}_i y_i}{\sum \tilde{x}_i^2}$$

on peut alors remplacer $\tilde{x}_i$ par $1 + x_i$, et on obtient

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

qui correspond à la réponse D.

20 Toujours sur le modèle $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, où on suppose ε_i centré, de variance constante, et indépendants les uns des autres, on estime les coefficients par moindres carrés. Quelles affirmations parmi les suivantes sont justes ?

i) la somme des résidus estimés est toujours nulle

ii) la somme des résidus estimés est nulle à condition que $\overline{y} = 0$

iii) si $R^2 = 0$, $\widehat{\beta}_1 = 0$ (et la droite de régression est horizontale)

iv) la droite de régression $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$ passe par le point $(\overline{x}, \overline{y})$ à condition que ce point soit un point de l'échantillon

A) (i) seulement

B) (ii) seulement

C) (i) et (iii)

D) (ii) et (iii)

E) (i) et (iv)

(i) oui, on l'a vu plusieurs fois en cours, c'est simplement la condition du premier ordre lorsqu'on différencie la somme des carrés des erreurs par rapport à β_0, i.e. $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i$, ce qui se traduit aussi par $\sum_{i=1}^{n} \widehat{\varepsilon}_i$.

A fortiori (ii) est fausse.

(iii) si $\widehat{\beta}_1 = 0$ alors $R^2 = 0$ (on reverra ce point dans la question 16 - la seconde sur les sorties de régression des annexes). Mais là, on nous demande la réciproque. Si $R^2 = 0$ alors $\sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 = 0$. La seule possibilité pour qu'une somme de termes positifs soit nulle est que *tous* les termes sont nuls. Donc pour tout i, $\widehat{y}_i - \overline{y}$, autrement dit, le modèle estimé est tout simplement un modèle constant, prenant la valeur $\overline{y}$, soit $\widehat{\beta}_0 = \overline{y}$ et $\widehat{\beta}_1 = 0$. Donc oui, (iii) est juste.

(iv) La droite de régression passe *toujours* par $(\overline{x}, \overline{y})$ - au risque de me répéter, c'est une conséquence directe de la condition du premier ordre $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i$. En effet

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i = \sum_{i=1}^{n} \widehat{\beta}_0 + \widehat{\beta}_1 x_i = n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_i$$

soit, en divisant par n,

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} y_i}_{\overline{y}} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{\overline{x}}$$

ce qui signifie que la droite de régression passe par $(\overline{x}, \overline{y})$. Donc (iv) est fausse.

Si on conclue, (i) et (iii) sont valides (et ce sont les seules), donc la bonne réponse est C.

Les questions 21 à 30 portent sur les sorties de régressions suivantes. En 1886, Galton avait obtenu des tailles (en pouces) de 934 enfants, dans 205 familles, à l'âge adulte, avec la taille de l'enfant (`childHeight`, y) en pouces, la taille de la mère (`mother`, x_1) en pouces, la taille du père (`father`, x_2) en pouces, la taille moyenne des parents (`midparentHeight`, x_3) en pouces, les versions en centimètres, `childHeightcm`, `mothercm` `fathercm`, `midparentHeightcm` (pour rappel, un pouce correspond à 2.54 cm). On a aussi le sexe de l'enfant (`gender`, x_4), variable binaire prenant les modalités `male` et `female`. On sait aussi s'il est enfant unique (`unique = TRUE`, x_5)

Je vais mettre ici les sorties R complètes, avec en bleu les valeurs qui avaient été cachées dans l'énoncé

• modèle (A)

```
> summary(lm(childHeight ~ midparentHeight, data=GaltonFamilies))

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     22.63624    4.26511   5.307 1.39e-07 ***
midparentHeight  0.63736    0.06161  10.345  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.392 on 932 degrees of freedom
Multiple R-squared:  0.103,Adjusted R-squared:  0.102
```

• modèle (B)

```
> summary(lm(childHeight ~ father, data=GaltonFamilies, subset = (gender == "male")))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 38.36258    3.30837  11.596   <2e-16 ***
father       0.44652    0.04783   9.337   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.416 on 479 degrees of freedom
Multiple R-squared:  0.154,Adjusted R-squared:  0.1522
```

• modèle (C)

```
> summary(lm(childHeight ~ father+mother, data=GaltonFamilies, subset = (gender == "male")))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.31281    4.09503   4.716 3.16e-06 ***
father       0.41756    0.04561   9.154  < 2e-16 ***
mother       0.32877    0.04530   7.258 1.61e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.295 on 478 degrees of freedom
Multiple R-squared:  0.2379,Adjusted R-squared:  0.2347
```

• modèle (D)

```
> summary(lm(childHeight ~ 0 + gender + midparentHeight, data=GaltonFamilies))

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
genderfemale    16.51410    2.73392   6.040 2.22e-09 ***
gendermale      21.72921    2.72893   7.963 4.89e-15 ***
midparentHeight  0.68702    0.03944  17.419  < 2e-16 ***

Residual standard error: 2.17 on 931 degrees of freedom
Multiple R-squared:  0.9989,Adjusted R-squared:  0.9989
```

• modèle (E)

```
> summary(lm(childHeight ~  gender + midparentHeight, data=GaltonFamilies))

Coefficients:
               Estimate Std. Error  t value Pr(>|t|)
(Intercept)     16.51410    2.73392     6.04 2.22e-09 ***
gendermale       5.21511    0.14216    36.69  < 2e-16
midparentHeight  0.68702    0.03944    17.42  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.17 on 931 degrees of freedom
Multiple R-squared:  0.6332,        Adjusted R-squared:  0.6324
```

• modèle (F)

```
> summary(lm(childHeight ~  gender + father+mother, data=GaltonFamilies))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.52124    2.72720   6.058    2e-09 ***
gendermale   5.21499    0.14181  36.775   <2e-16 ***
father       0.39284    0.02868  13.699   <2e-16 ***
mother       0.31761    0.03100  10.245   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.165 on 930 degrees of freedom
Multiple R-squared:  0.6354, Adjusted R-squared:  0.6342
```

• modèle (G)

```
> summary(lm(childHeight ~  gender + midparentHeight + unique , data=GaltonFamilies))

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     16.49150    2.73647   6.027 2.41e-09 ***
gendermale       5.21514    0.14223  36.667  < 2e-16 ***
midparentHeight  0.68729    0.03947  17.412  < 2e-16 ***
uniqueTRUE       0.10737    0.38492   0.279     0.78
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.171 on 930 degrees of freedom
Multiple R-squared:  0.6332, Adjusted R-squared:  0.6321
```

• modèle (H)

```
> summary(lm(childHeightcm ~  gender + midparentHeightcm , data=GaltonFamilies))

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       41.94582    6.94415    6.04 2.22e-09 ***
gendermale        13.24637    0.36108   36.69  < 2e-16 ***
midparentHeightcm  0.68702    0.03944   17.42  < 2e-16 ***

Residual standard error: 5.512 on 931 degrees of freedom
Multiple R-squared:  0.6332, Adjusted R-squared:  0.6324
```

• modèle (I)

```
> summary(lm(childHeight ~  gender + midparentHeight +I(father-mother), data=GaltonFamilies))

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        16.52124    2.72720   6.058    2e-09 ***
gendermale          5.21499    0.14181  36.775   <2e-16 ***
midparentHeight     0.68313    0.03938  17.348   <2e-16 ***
I(father - mother)  0.05128    0.02169   2.364   0.0183 *

Residual standard error: 2.165 on 930 degrees of freedom
Multiple R-squared:  0.6354, Adjusted R-squared:  0.6342
```

21 Un homme de 64 pouces a un père mesurant 62 pouces, une mère mesurant 66 pouces. Pour le modèle (F), quelle serait l'erreur associée à cette nouvelle observation, $\boldsymbol{x}_{n+1}$, avec $\widehat{\varepsilon}_{n+1} = \widehat{y}_{n+1} - y_{n+1}$?

A) moins de -2

B) entre -2 et -1

C) entre -1 et +1

D) entre +1 et +2

E) plus que +2

Le modèle (F) nous dit que $\widehat{y} = 16.52124 + 5.21499 \cdot \mathbf{1}(x_4 = \text{homme}) + 0.39284 \cdot x_2 + 0.31761 \cdot x_1$, où x_4 est le sexe de la personne, x_2 la taille du père et x_1 la taille de la mère. Or on nous parle de "un homme ($\mathbf{1}(x_4 = \text{homme}) = 1$) de 64 pouces ($y = 64$)" ayant "un père mesurant 62 pouces ($x_2 = 62$), une mère mesurant 66 pouces ($x_1 = 66$)". Donc ici

$$\widehat{y} = 16.52124 + 5.21499 \cdot 1 + 0.39284 \cdot 62 + 0.31761 \cdot 66 = 67.05478$$

et comme $y = 64$, on en déduit que $\widehat{\varepsilon} = 67.05478 - 64 = 3.05478$ (on notera ici que les résidus sont définis "à l'envers", la définition usuelle étant $\widehat{\varepsilon}_{n+1} = y_{n+1} - \widehat{y}_{n+1}$ - mais c'est juste une convention, plus intuitive que celle de l'énoncé, puisqu'une résidu positif signifie normalement que la personne est réellement plus grande que ce à quoi on s'attend). Bref, il fallait répondre E.

22 Dans la régression (E), $y = \beta_0 + \beta_{\text{male}}\mathbf{1}(x_4 = \text{“male”}) + \beta_3 x_3 + \varepsilon$, que vaut l'estimateur par la méthode des moindres carrés de β_{male}

A) moins de 5

B) entre 5 et 10

C) entre 10 et 15

D) entre 15 et 20

E) plus de 20

Les régressions (D) et (E) sont rigoureusement équivalentes, cf discussion sur les variables (explicatives) factorielles - c'est juste de l'algèbre linéaire, $\mathbf{1}(x_4 = \text{“female”}) + \mathbf{1}(x_4 = \text{“male”}) = \mathbf{1}$, qui est la constante de la régression (E). En particulier les prévisions sont identiques avec ces deux modèles,

$$\widehat{y} = 16.51410 \cdot \mathbf{1}(x_4 = \text{“female”}) + 21.72921 \cdot \mathbf{1}(x_4 = \text{“male”}) + 0.68702 \cdot x_3 = 16.51410 + \widehat{\beta}_{\text{male}} \cdot \mathbf{1}(x_4 = \text{“male”}) + 0.68702 \cdot x_3$$

(comme on s'y attend, les coefficients pour x_3 sont identiques). Comme $\mathbf{1}(x_4 = \text{“female”}) + \mathbf{1}(x_4 = \text{“male”}) = 1$, on peut écrire $\mathbf{1}(x_4 = \text{“female”}) = 1 - \mathbf{1}(x_4 = \text{“male”})$, et on a alors, par substitution

$$16.51410 \cdot \big(1 - \mathbf{1}(x_4 = \text{“male”})\big) + 21.72921 \cdot \mathbf{1}(x_4 = \text{“male”}) = 16.51410 + \widehat{\beta}_{\text{male}} \cdot \mathbf{1}(x_4 = \text{“male”})$$

soit

$$\big(21.72921 - 16.51410\big) \cdot \mathbf{1}(x_4 = \text{“male”})\big)\widehat{\beta}_{\text{male}} \cdot \mathbf{1}(x_4 = \text{“male”})$$

donc, par identification, $\widehat{\beta}_{\text{male}} = 21.72921 - 16.51410 = 5.21511$, qui correspond à la réponse B.

23 Dans la régression (E), que vaut le R^2 ?

A) moins de 55%

B) entre 55% et 60%

C) entre 60% et 65%

D) entre 65% et 70%

E) plus de 70%

Pour rappel, le R^2 est la part de variance expliquée par le modèle, donc $1 - R^2$ est la variance des résidus, divisée par la variance de la variable d'intérêt, donc $R^2 = 1 - \text{Var}[\varepsilon]^2/\text{Var}[y]$. Or la variance $\text{Var}[y]$ est constante, donc comme $\text{Var}[\varepsilon]$ vaut ici 2.17, on devrait avoir un R^2 proche de celui de (G), où $\text{Var}[\varepsilon]$ vaut 2.171. Mais on peut le faire plus rigoureusement en prenant n'importe quel modèle (sauf (D)), par exemple (A) :

$$R^2 = 0.103 = 1 - \frac{3.392^2}{\text{Var}[y]} \text{ donc } \text{Var}[y] = \frac{3.392^2}{1 - .103} = 12.82683$$

ou (F) (car (B) et (C) sont construits sur des sous-bases)

$$R^2 = 0.154 = 1 - \frac{2.416^2}{\text{Var}[y]} \text{ donc } \text{Var}[y] = \frac{2.416^2}{1 - .154} = 12.8558$$

la différence numérique est expliquée par des histoires de $n - p$ avec des p différents. Aussi, on obtient numériquement $R^2 = 1 - 2.17^2/12.8558 \approx 0.633714$ (la vraie valeur était 0.6324), qui correspond à la réponse C.

24 Pour la régression (G), où la taille de l'enfant est expliquée par le sexe (x_4), la taille moyenne des parents (x_3) et le fait que l'enfant soit unique ou pas (x_5), donnez un ordre de grandeur pour la p-value du test de Student de $H_0 : \beta_5 = 0$ (paramètre associé à x_5, `uniqueTRUE`)

A) moins de 10%

B) entre 10% et 25%

C) entre 25% et 50%

D) entre 50% et 75%

E) plus que 75%

Sur la base de la sortie du modèle (G), on peut calculer

$$t = \frac{\widehat{\beta}_5}{\sqrt{\mathrm{Var}(\widehat{\beta}_5)}} = \frac{0.10737}{0.38492} = 0.2789411$$

la seconde colonne est l'écart-type, directement (et pas la variance). La p-value est alors la probabilité pour que $|T|$ dépasse 0.2789411, où T suit une loi normale (on a plus de $n = 900$ observations). Or d'après la table de la loi normale (fonction de répartition de T), $\mathbb{P}[T \leqslant 0.28] = \Phi(0.28) = 0.61$, donc $\mathbb{P}[T > 0.28] = 1 - 0.61 = 0.39$. Or $\mathbb{P}[|T| > 0.28] = \mathbb{P}[T \leqslant -0.28] + \mathbb{P}[T \leqslant 0.28] = 2 \cdot \mathbb{P}[T \leqslant 0.28]$ par symmétrie de la loi de T (par rapport à 0). Donc $\mathbb{P}[|T| > 0.28] = 2 \cdot 0.39 = 0.78$. C'était la réponse E.

25 On considère un jeune homme, dont la mère mesure 67.0 pouces (x_1), et le père 78.5 pouces (x_2). On considère (1) le modèle de régression sur x_1 et x_2 estimé uniquement sur les jeunes homme (2) le modèle de régression sur x_1, x_2 et en rajoutant le sexe, x_4. On note $\widehat{y}^{(1)}$ la prédiction avec le modèle (1) et $\widehat{y}^{(2)}$ la prédiction avec le modèle (2). Que vaut $\widehat{y}^{(1)} - \widehat{y}^{(2)}$?

A) moins de -1 pouce

B) entre -1 pouce et 0

C) exactement 0 pouce

D) entre 0 et 1 pouce

E) plus de +1 pouce

Pour les deux questions suivantes, j'avais fait une coquille, en intervertissant β_3^{cm} (qui aurait du être pour la question 27) et β_4^{cm} (qui aurait du être pour la question 26). Je vais donc accepter les réponses que j'attendais (certains avaient noté sur la copie que je m'étais mélangé) et la réponse logique à la question. Je vais expliquer ici comment répondre aux deux questions.

On s'appuie ici sur la régression (E), qui est la même, aux unités près ((E) est en pouces, et (H) en centimètres). On a (cf question 22)

$$\widehat{y}^{pouces} = 16.5140 + 5.21511 \cdot \mathbf{1}(\text{“male”}) + 0.68702 \cdot x_3^{pouces}$$

Or $\widehat{y}^{pouces} = \widehat{y}^{cm}/2.54$ (comme rappelé plus haut), et pareil pour x_3^{pouces}. Bref,

$$\frac{\widehat{y}^{cm}}{2.54} = 16.5140 + 5.21511 \cdot \mathbf{1}(\text{“male”}) + 0.68702 \cdot \frac{{x_3}^{cm}}{2.54}$$

soit

$$\widehat{y}^{cm} = 2.54 \cdot 16.5140 + 2.54 \cdot 5.21511 \cdot \mathbf{1}(\text{“male”}) + 0.68702 \cdot {x_3}^{cm}$$

On retrouve bien la constante ($2.54 \cdot 16.5140 = 41.945$) et les deux autres valeurs que l'on cherche sont 13.24 pour β_4^{cm} et 0.68702 pour β_3^{cm} (qui étaient les valeurs que j'attendais respectivement aux deux prochaines questions, mais je me suis emmêlé.)

26 Pour la régression (H), $y^{cm} = \beta_0^{cm} + \beta_4^{cm}\mathbf{1}(x_4 = \text{“male”}) + \beta_3^{cm}x_3^{cm} + \varepsilon^{cm}$, que vaut l'estimateur par moindres carrés de β_3^{cm}

A) moins de 3

B) entre 3 et 5

C) entre 5 et 8

D) entre 8 et 12

E) plus que 12

27 Pour la régression (H), $y^{cm} = \beta_0^{cm} + \beta_4^{cm}\mathbf{1}(x_4 = \text{“male”}) + \beta_3^{cm}x_3^{cm} + \varepsilon^{cm}$, que vaut l'estimateur par moindres carrés de β_4^{cm}

A) moins de 0.3

B) entre 0.3 et 0.5

C) entre 0.5 et 0.8

D) entre 0.8 et 1.5

E) plus que 1.5

28 Pour la régression (H), $y^{cm} = \beta_0^{cm} + \beta_4^{cm}\mathbf{1}(x_4 = \text{“male”}) + \beta_3^{cm}x_3^{cm} + \varepsilon^{cm}$, que vaut l'estimateur de l'écart-type de $\widehat{\beta}_0^{cm}$, estimé par moindres carrés ?

A) moins de 1

B) entre 1 et 3.5

C) entre 3.5 et 6

D) entre 6 et 7.5

E) plus que 7.5

6.94415 qui correspond à la réponse D.

29 Pour la régression (H), $y^{cm} = \beta_0^{cm} + \beta_4^{cm}\mathbf{1}(x_4 = \text{“male”}) + \beta_3^{cm}x_3^{cm} + \varepsilon^{cm}$, que vaut l'estimateur de la variance des $\widehat{\varepsilon}_i^{cm}$ (résidus de la régression estimée par moindres carrés).

A) moins de 10

B) entre 10 et 15

C) entre 15 et 20

D) entre 20 et 25

E) plus que 25

5.512001^2 soit un peu plus de 25, qui est la réponse E.

30 Dans le modèle (I), on tente d'expliquer la taille de l'enfant par la taille moyenne des parents x_3 et l'écart de taille entre le père et la mère, $x_6 = x_2 - x_1$. Que vaut l'estimateur par moindres carrés du paramètre associé à la constante, β_0 ? C(on retiendra la valeur la plus proche)

A) 8.26

B) 16.52

C) 19.31

D) 22.63

E) 38.36

16.52124, qui est la réponse B.

Table de la fonction de répartition de la loi normale $\Phi(u)$

u	$0,00$	$0,01$	$0,02$	$0,03$	$0,04$	$0,05$	$0,06$	$0,07$	$0,08$	$0,09$
$0,0$	$0,5000$	$0,5040$	$0,5080$	$0,5120$	$0,5160$	$0,5199$	$0,5239$	$0,5279$	$0,5319$	$0,5359$
$0,1$	$0,5398$	$0,5348$	$0,5478$	$0,5517$	$0,5557$	$0,5596$	$0,5636$	$0,5675$	$0,5714$	$0,5753$
$0,2$	$0,5793$	$0,5832$	$0,5871$	$0,5910$	$0,5948$	$0,5987$	$0,6026$	$0,6064$	$0,6103$	$0,6141$
$0,3$	$0,6179$	$0,6217$	$0,6255$	$0,6293$	$0,6331$	$0,6368$	$0,6406$	$0,6443$	$0,6480$	$0,6517$
$0,4$	$0,6554$	$0,6591$	$0,6628$	$0,6664$	$0,6700$	$0,6736$	$0,6772$	$0,6808$	$0,6844$	$0,6879$
$0,5$	$0,6915$	$0,6950$	$0,6985$	$0,7019$	$0,7054$	$0,7088$	$0,7123$	$0,7157$	$0,7190$	$0,7224$
$0,6$	$0,7257$	$0,7290$	$0,7324$	$0,7357$	$0,7389$	$0,7422$	$0,7454$	$0,7486$	$0,7517$	$0,7549$
$0,7$	$0,7580$	$0,7611$	$0,7642$	$0,7673$	$0,7704$	$0,7734$	$0,7764$	$0,7794$	$0,7823$	$0,7852$
$0,8$	$0,7881$	$0,7910$	$0,7939$	$0,7967$	$0,7995$	$0,8023$	$0,8051$	$0,8078$	$0,8106$	$0,8133$
$0,9$	$0,8159$	$0,8186$	$0,8212$	$0,8238$	$0,8264$	$0,8289$	$0,8315$	$0,8340$	$0,8365$	$0,8389$
$1,0$	$0,8413$	$0,8438$	$0,8461$	$0,8485$	$0,8508$	$0,8531$	$0,8554$	$0,8577$	$0,8599$	$0,8621$
$1,1$	$0,8643$	$0,8665$	$0,8686$	$0,8708$	$0,8729$	$0,8749$	$0,8770$	$0,8790$	$0,8810$	$0,8830$
$1,2$	$0,8849$	$0,8869$	$0,8888$	$0,8907$	$0,8925$	$0,8944$	$0,8962$	$0,8980$	$0,8997$	$0,9015$
$1,3$	$0,9032$	$0,9049$	$0,9066$	$0,9082$	$0,9099$	$0,9115$	$0,9131$	$0,9147$	$0,9162$	$0,9177$
$1,4$	$0,9192$	$0,9207$	$0,9222$	$0,9236$	$0,9251$	$0,9265$	$0,9279$	$0,9292$	$0,9306$	$0,9319$
$1,5$	$0,9332$	$0,9345$	$0,9357$	$0,9370$	$0,9382$	$0,9394$	$0,9406$	$0,9418$	$0,9429$	$0,9441$
$1,6$	$0,9452$	$0,9463$	$0,9474$	$0,9484$	$0,9495$	$0,9505$	$0,9515$	$0,9525$	$0,9535$	$0,9545$
$1,7$	$0,9554$	$0,9564$	$0,9573$	$0,9582$	$0,9591$	$0,9599$	$0,9608$	$0,9616$	$0,9625$	$0,9633$
$1,8$	$0,9641$	$0,9649$	$0,9656$	$0,9664$	$0,9671$	$0,9678$	$0,9686$	$0,9693$	$0,9699$	$0,9706$
$1,9$	$0,9713$	$0,9719$	$0,9726$	$0,9732$	$0,9738$	$0,9744$	$0,9750$	$0,9756$	$0,9761$	$0,9767$
$2,0$	$0,9772$	$0,9779$	$0,9783$	$0,9788$	$0,9793$	$0,9798$	$0,9803$	$0,9808$	$0,9812$	$0,9817$
$2,1$	$0,9821$	$0,9826$	$0,9830$	$0,9834$	$0,9838$	$0,9842$	$0,9846$	$0,9850$	$0,9854$	$0,9857$
$2,2$	$0,9861$	$0,9864$	$0,9868$	$0,9871$	$0,9875$	$0,9878$	$0,9881$	$0,9884$	$0,9887$	$0,9890$
$2,3$	$0,9893$	$0,9896$	$0,9898$	$0,9901$	$0,9904$	$0,9906$	$0,9909$	$0,9911$	$0,9913$	$0,9916$
$2,4$	$0,9918$	$0,9920$	$0,9922$	$0,9925$	$0,9927$	$0,9929$	$0,9931$	$0,9932$	$0,9934$	$0,9936$
$2,5$	$0,9938$	$0,9940$	$0,9941$	$0,9943$	$0,9945$	$0,9946$	$0,9948$	$0,9949$	$0,9951$	$0,9952$
$2,6$	$0,9953$	$0,9955$	$0,9956$	$0,9957$	$0,9959$	$0,9960$	$0,9961$	$0,9962$	$0,9963$	$0,9964$
$2,7$	$0,9965$	$0,9966$	$0,9967$	$0,9968$	$0,9969$	$0,9970$	$0,9971$	$0,9972$	$0,9973$	$0,9974$
$2,8$	$0,9974$	$0,9975$	$0,9976$	$0,9977$	$0,9977$	$0,9978$	$0,9979$	$0,9979$	$0,9980$	$0,9981$
$2,9$	$0,9981$	$0,9982$	$0,9982$	$0,9983$	$0,9984$	$0,9984$	$0,9985$	$0,9985$	$0,9986$	$0,9986$

Table de quantile de Student $F_\nu^{-1}(p)$

ν \ p	0.60	0.70	0.80	0.90	0.95	0.975	0.990	0.995	0.999	0.9995
1	0.325	0.727	1.376	3.078	6.314	12.71	31.82	63.66	318.3	636.6
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925	22.33	31.60
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	10.22	12.94
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032	5.893	6.859
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	4.785	5.405
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878	3.611	3.922
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807	3.485	3.767
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750	3.385	3.646
32	0.256	0.530	0.853	1.309	1.694	2.037	2.449	2.738	3.365	3.622
34	0.255	0.529	0.852	1.307	1.691	2.032	2.441	2.728	3.348	3.601
36	0.255	0.529	0.852	1.306	1.688	2.028	2.434	2.719	3.333	3.582
38	0.255	0.529	0.851	1.304	1.686	2.024	2.429	2.712	3.319	3.566
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	0.255	0.528	0.849	1.298	1.676	2.009	2.403	2.678	3.261	3.496
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660	3.232	3.460
70	0.254	0.527	0.847	1.294	1.667	1.994	2.381	2.648	3.211	3.435
80	0.254	0.527	0.846	1.292	1.664	1.990	2.374	2.639	3.195	3.415
90	0.254	0.526	0.846	1.291	1.662	1.987	2.368	2.632	3.183	3.402
100	0.254	0.526	0.845	1.290	1.660	1.984	2.365	2.626	3.174	3.389
200	0.254	0.525	0.843	1.286	1.653	1.972	2.345	2.601	3.131	3.339
500	0.253	0.525	0.842	1.283	1.648	1.965	2.334	2.586	3.106	3.310
∞	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Table de quantile du chi-deux $F_\nu^{-1}(p)$

ν \ P	0.001	0.005	0.010	0.025	0.05	0.10	0.50	0.90	0.95	0.975	0.990	0.995	0.999
1	–	–	–	0.001	0.004	0.016	0.455	2.71	3.84	5.02	6.63	7.88	10.8
2	0.002	0.010	0.020	0.051	0.103	0.211	1.39	4.61	5.99	7.38	9.21	10.6	13.8
3	0.024	0.072	0.115	0.216	0.352	0.584	2.37	6.25	7.81	9.35	11.3	12.8	16.3
4	0.091	0.207	0.297	0.484	0.711	1.06	3.36	7.78	9.49	11.1	13.3	14.9	18.5
5	0.210	0.412	0.554	0.831	1.15	1.61	4.35	9.24	11.1	12.8	15.1	16.7	20.5
6	0.381	0.676	0.872	1.24	1.64	2.20	5.35	10.6	12.6	14.4	16.8	18.5	22.5
7	0.598	0.989	1.24	1.69	2.17	2.83	6.35	12.0	14.1	16.0	18.5	20.3	24.3
8	0.857	1.34	1.65	2.18	2.73	3.49	7.34	13.4	15.5	17.5	20.1	22.0	26.1
9	1.15	1.73	2.09	2.70	3.33	4.17	8.34	14.7	16.9	19.0	21.7	23.6	27.9
10	1.48	2.16	2.56	3.25	3.94	4.87	9.34	16.0	18.3	20.5	23.2	25.2	29.6
11	1.83	2.60	3.05	3.82	4.57	5.58	10.3	17.3	19.7	21.9	24.7	26.8	31.3
12	2.21	3.07	3.57	4.40	5.23	6.30	11.3	18.5	21.0	23.3	26.2	28.3	32.9
13	2.62	3.57	4.11	5.01	5.89	7.04	12.3	19.8	22.4	24.7	27.7	29.8	34.5
14	3.04	4.07	4.66	5.63	6.57	7.79	13.3	21.1	23.7	26.1	29.1	31.3	36.1
15	3.48	4.60	5.23	6.26	7.26	8.55	14.3	22.3	25.0	27.5	30.6	32.8	37.7
16	3.94	5.14	5.81	6.91	7.96	9.31	15.3	23.5	26.3	28.8	32.0	34.3	39.3
17	4.42	5.70	6.41	7.56	8.67	10.1	16.3	24.8	27.6	30.2	33.4	35.7	40.8
18	4.90	6.26	7.01	8.23	9.39	10.9	17.3	26.0	28.9	31.5	34.8	37.2	42.3
19	5.41	6.84	7.63	8.91	10.1	11.7	18.3	27.2	30.1	32.9	36.2	38.6	43.8
20	5.92	7.43	8.26	9.59	10.9	12.4	19.3	28.4	31.4	34.2	37.6	40.0	45.3
21	6.45	8.03	8.90	10.3	11.6	13.2	20.3	29.6	32.7	35.5	38.9	41.4	46.8
22	6.98	8.64	9.54	11.0	12.3	14.0	21.3	30.8	33.9	36.8	40.3	42.8	48.3
23	7.53	9.26	10.2	11.7	13.1	14.8	22.3	32.0	35.2	38.1	41.6	44.2	49.7
24	8.08	9.89	10.9	12.4	13.8	15.7	23.3	33.2	26.4	29.4	43.0	45.6	51.2
25	8.65	10.5	11.5	13.1	14.6	16.5	24.3	34.4	37.7	40.6	44.3	46.9	52.6
26	9.22	11.2	12.2	13.8	15.4	17.3	25.3	35.6	38.9	41.9	45.6	48.3	54.1
27	9.80	11.8	12.9	14.6	16.2	18.1	26.3	36.7	40.1	43.2	47.0	49.6	55.5
28	10.4	12.5	13.6	15.3	16.9	18.9	27.3	37.9	41.3	44.5	48.3	51.0	56.9
29	11.0	13.1	14.3	16.0	17.7	19.8	28.3	39.1	42.6	45.7	49.6	52.3	58.3
30	11.6	13.8	15.0	16.8	18.5	20.6	29.3	40.3	43.8	47.0	50.9	53.7	59.7
32	12.8	15.1	16.4	18.3	20.1	22.3	31.3	42.6	46.2	49.5	53.5	56.3	62.5
34	14.1	16.5	17.8	19.8	21.7	24.0	33.3	44.9	48.6	52.0	56.1	59.0	65.2
36	15.3	17.9	19.2	21.3	23.3	25.6	35.3	47.2	51.0	54.4	58.6	61.6	68.0
38	16.6	19.3	20.7	22.9	24.9	27.3	37.3	49.5	53.4	56.9	61.2	64.2	70.7
40	17.9	20.7	22.2	24.4	26.5	29.1	39.3	51.8	55.8	59.3	63.7	66.8	73.4
50	24.7	28.0	29.7	32.4	34.8	37.7	49.3	63.2	67.5	71.4	76.2	79.5	86.7
60	31.7	35.5	37.5	40.5	43.2	46.5	59.3	74.4	79.1	83.3	88.4	92.0	99.6
70	39.0	43.3	45.4	48.8	51.7	55.3	69.3	85.5	90.5	95.0	100.4	104.2	112.3
80	46.5	51.2	53.5	57.2	60.4	64.3	79.3	96.6	101.9	106.6	112.3	116.3	124.8
90	54.2	59.2	61.8	65.6	69.1	73.3	89.3	107.6	113.1	118.1	124.1	128.3	137.2
100	61.9	67.3	70.1	74.2	77.9	82.4	99.3	118.5	124.3	129.6	135.8	140.2	149.4

Table de quantile de la loi de Fisher (quantile à 97.5%, $F^{-1}_{\nu_n,\nu_d}(97.5\%)$)

num	den 1	2	3	4	5	6	7	8	9	10
1	161.4476	18.5128	10.1280	7.7086	6.6079	5.9874	5.5914	5.3177	5.1174	4.9646
2	199.5000	19.0000	9.5521	6.9443	5.7861	5.1433	4.7374	4.4590	4.2565	4.1028
3	215.7073	19.1643	9.2766	6.5914	5.4095	4.7571	4.3468	4.0662	3.8625	3.7083
4	224.5832	19.2468	9.1172	6.3882	5.1922	4.5337	4.1203	3.8379	3.6331	3.4780
5	230.1619	19.2964	9.0135	6.2561	5.0503	4.3874	3.9715	3.6875	3.4817	3.3258
6	233.9860	19.3295	8.9406	6.1631	4.9503	4.2839	3.8660	3.5806	3.3738	3.2172
7	236.7684	19.3532	8.8867	6.0942	4.8759	4.2067	3.7870	3.5005	3.2927	3.1355
8	238.8827	19.3710	8.8452	6.0410	4.8183	4.1468	3.7257	3.4381	3.2296	3.0717
9	240.5433	19.3848	8.8123	5.9988	4.7725	4.0990	3.6767	3.3881	3.1789	3.0204
10	241.8817	19.3959	8.7855	5.9644	4.7351	4.0600	3.6365	3.3472	3.1373	2.9782
11	242.9835	19.4050	8.7633	5.9358	4.7040	4.0274	3.6030	3.3130	3.1025	2.9430
12	243.9060	19.4125	8.7446	5.9117	4.6777	3.9999	3.5747	3.2839	3.0729	2.9130
13	244.6898	19.4189	8.7287	5.8911	4.6552	3.9764	3.5503	3.2590	3.0475	2.8872
14	245.3640	19.4244	8.7149	5.8733	4.6358	3.9559	3.5292	3.2374	3.0255	2.8647
15	245.9499	19.4291	8.7029	5.8578	4.6188	3.9381	3.5107	3.2184	3.0061	2.8450
16	246.4639	19.4333	8.6923	5.8441	4.6038	3.9223	3.4944	3.2016	2.9890	2.8276
17	246.9184	19.4370	8.6829	5.8320	4.5904	3.9083	3.4799	3.1867	2.9737	2.8120
18	247.3232	19.4402	8.6745	5.8211	4.5785	3.8957	3.4669	3.1733	2.9600	2.7980
19	247.6861	19.4431	8.6670	5.8114	4.5678	3.8844	3.4551	3.1613	2.9477	2.7854
20	248.0131	19.4458	8.6602	5.8025	4.5581	3.8742	3.4445	3.1503	2.9365	2.7740
21	248.3094	19.4481	8.6540	5.7945	4.5493	3.8649	3.4349	3.1404	2.9263	2.7636
22	248.5791	19.4503	8.6484	5.7872	4.5413	3.8564	3.4260	3.1313	2.9169	2.7541
23	248.8256	19.4523	8.6432	5.7805	4.5339	3.8486	3.4179	3.1229	2.9084	2.7453
24	249.0518	19.4541	8.6385	5.7744	4.5272	3.8415	3.4105	3.1152	2.9005	2.7372
25	249.2601	19.4558	8.6341	5.7687	4.5209	3.8348	3.4036	3.1081	2.8932	2.7298
26	249.4525	19.4573	8.6301	5.7635	4.5151	3.8287	3.3972	3.1015	2.8864	2.7229
27	249.6309	19.4587	8.6263	5.7586	4.5097	3.8230	3.3913	3.0954	2.8801	2.7164
28	249.7966	19.4600	8.6229	5.7541	4.5047	3.8177	3.3858	3.0897	2.8743	2.7104
29	249.9510	19.4613	8.6196	5.7498	4.5001	3.8128	3.3806	3.0844	2.8688	2.7048
30	250.0951	19.4624	8.6166	5.7459	4.4957	3.8082	3.3758	3.0794	2.8637	2.6996

num	den 11	12	13	14	15	16	17	18	19	20
1	4.8443	4.7472	4.6672	4.6001	4.5431	4.4940	4.4513	4.4139	4.3807	4.3512
2	3.9823	3.8853	3.8056	3.7389	3.6823	3.6337	3.5915	3.5546	3.5219	3.4928
3	3.5874	3.4903	3.4105	3.3439	3.2874	3.2389	3.1968	3.1599	3.1274	3.0984
4	3.3567	3.2592	3.1791	3.1122	3.0556	3.0069	2.9647	2.9277	2.8951	2.8661
5	3.2039	3.1059	3.0254	2.9582	2.9013	2.8524	2.8100	2.7729	2.7401	2.7109
6	3.0946	2.9961	2.9153	2.8477	2.7905	2.7413	2.6987	2.6613	2.6283	2.5990
7	3.0123	2.9134	2.8321	2.7642	2.7066	2.6572	2.6143	2.5767	2.5435	2.5140
8	2.9480	2.8486	2.7669	2.6987	2.6408	2.5911	2.5480	2.5102	2.4768	2.4471
9	2.8962	2.7964	2.7144	2.6458	2.5876	2.5377	2.4943	2.4563	2.4227	2.3928
10	2.8536	2.7534	2.6710	2.6022	2.5437	2.4935	2.4499	2.4117	2.3779	2.3479
11	2.8179	2.7173	2.6347	2.5655	2.5068	2.4564	2.4126	2.3742	2.3402	2.3100
12	2.7876	2.6866	2.6037	2.5342	2.4753	2.4247	2.3807	2.3421	2.3080	2.2776
13	2.7614	2.6602	2.5769	2.5073	2.4481	2.3973	2.3531	2.3143	2.2800	2.2495
14	2.7386	2.6371	2.5536	2.4837	2.4244	2.3733	2.3290	2.2900	2.2556	2.2250
15	2.7186	2.6169	2.5331	2.4630	2.4034	2.3522	2.3077	2.2686	2.2341	2.2033
16	2.7009	2.5989	2.5149	2.4446	2.3849	2.3335	2.2888	2.2496	2.2149	2.1840
17	2.6851	2.5828	2.4987	2.4282	2.3683	2.3167	2.2719	2.2325	2.1977	2.1667
18	2.6709	2.5684	2.4841	2.4134	2.3533	2.3016	2.2567	2.2172	2.1823	2.1511
19	2.6581	2.5554	2.4709	2.4000	2.3398	2.2880	2.2429	2.2033	2.1683	2.1370
20	2.6464	2.5436	2.4589	2.3879	2.3275	2.2756	2.2304	2.1906	2.1555	2.1242
21	2.6358	2.5328	2.4479	2.3768	2.3163	2.2642	2.2189	2.1791	2.1438	2.1124
22	2.6261	2.5229	2.4379	2.3667	2.3060	2.2538	2.2084	2.1685	2.1331	2.1016
23	2.6172	2.5139	2.4287	2.3573	2.2966	2.2443	2.1987	2.1587	2.1233	2.0917
24	2.6090	2.5055	2.4202	2.3487	2.2878	2.2354	2.1898	2.1497	2.1141	2.0825
25	2.6014	2.4977	2.4123	2.3407	2.2797	2.2272	2.1815	2.1413	2.1057	2.0739
26	2.5943	2.4905	2.4050	2.3333	2.2722	2.2196	2.1738	2.1335	2.0978	2.0660
27	2.5877	2.4838	2.3982	2.3264	2.2652	2.2125	2.1666	2.1262	2.0905	2.0586
28	2.5816	2.4776	2.3918	2.3199	2.2587	2.2059	2.1599	2.1195	2.0836	2.0517
29	2.5759	2.4718	2.3859	2.3139	2.2525	2.1997	2.1536	2.1131	2.0772	2.0452
30	2.5705	2.4663	2.3803	2.3082	2.2468	2.1938	2.1477	2.1071	2.0712	2.0391