STT5100 - Automne 2018 - Examen Intra

Arthur Charpentier

Examen **A**

Les calculatrices sont autorisées. Les documents sont en revanche interdits, sauf une page d'aide mémoire.

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

- vous gagnez 1 points 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. 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 réponse est au dos de cette page : merci de décrocher la feuille et de ne rendre que cette dernière, après avoir indiqué votre code permanent en haut.

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

Formulaire

Pour la loi normale, centrée et réduite ou une loi de Student, on utilisera 1.96 comme valeur du quantile à 97.5%, et 1.64 pour le quantile à 95%.

On notera $x \mapsto \mathbf{1}_A(x)$ la fonction indicatrice vérifiant $\mathbf{1}_A(x) = 0$ si $x \notin A$ et $\mathbf{1}_A(x) = 1$ si $x \in A$. Par extention, si $A = \{a\}$, on notera $x \mapsto \mathbf{1}_a(x)$ la fonction qui vérifie $\mathbf{1}_a(x) = 0$ si $x \neq a$ et $\mathbf{1}_a(x) = 1$ si $x = a$.

La notation $\underset{(0.021)}{0.072}$ pour $\widehat{\beta}_j$ signifie que l'estimateur de β_j vaut 0.072 et que l'écart-type de cet estimateur vaut 0.021.

Pour les questions 1 à 11, on considère le modèle suivant

$$\widehat{\text{poids}}_i^{\,\text{kg}} = \widehat{\beta}_0 + \widehat{\beta}_2 \text{ cigarettes}_i^{\text{nb}} + \widehat{\beta}_3 \text{ revenu}_i^{\$} \quad (1)$$

où, pour une naissance i, $\text{cigarettes}^{\text{nb}}$ indique le nombre de cigarettes fumées par jour par la mère, $\text{revenu}^{\$}$ le revenu de la mère (en dollars), et $\widehat{\text{poids}}^{\,\text{kg}}$ est le poids à la naissance (en kilogrammes) prédit par le modèle linéaire (1). On décide de changer d'unités : le poids est en livres (1 livre = 0.454 kg) et le revenu est en milliers de dollars. On a alors le modèle

$$\breve{\text{poids}}_i^{\,\text{lb}} = \breve{\beta}_0 + \breve{\beta}_2 \text{ cigarettes}_i^{\text{nb}} + \breve{\beta}_3 \text{ revenu}_i^{\text{k\$}} \quad (2)$$

Un troisième modèle est considéré

$$\widetilde{\text{poids}}_i^{\,\text{kg}} = \widetilde{\beta}_0 + \widetilde{\beta}_1 \, \mathbf{1}_0(\text{cigarettes}_i^{\text{nb}}) + \widetilde{\beta}_2 \text{ cigarettes}_i^{\text{nb}} + \widetilde{\beta}_3 \text{ revenu}_i^{\$} \quad (3)$$

On suppose qu'il existe des personnes qui ne fument pas dans la base.

Les trois modèles sont estimés ici par la méthode des moindres carrés.

1 Comment peut-on interpréter $\widehat{\beta}_0$?

A) c'est le nombre de cigarettes par jour qu'il faut pour réduire le poids à la naissance de 1 kg, en moyenne

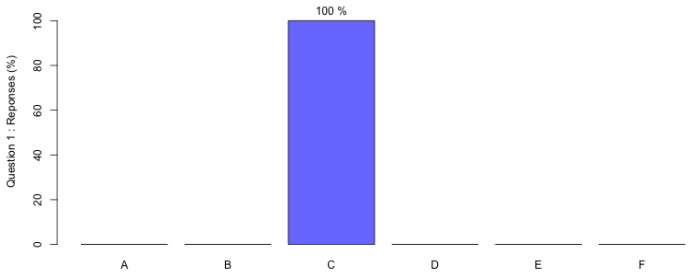
B) c'est de combien, une cigarette supplémentaire par jour diminue le poids à la naissance, en moyenne

C) c'est le poids qu'aurait, en moyenne, un bébé dont la mère ne fume pas, et n'ayant aucun revenu

D) c'est l'écart-type des résidus de la régression

E) aucune des réponses proposées

dans ce cas, $\text{cigarettes}^{\text{nb}} = 0$ et $\text{revenu}^{\$} = 0$ donc si on substitue dans $\widehat{\text{poids}}^{\,\text{kg}} = \widehat{\beta}_0 + \widehat{\beta}_2 \text{ cigarettes}^{\text{nb}} + \widehat{\beta}_3 \text{ revenu}^{\$}$ on obtient $\widehat{\text{poids}}^{\,\text{kg}} = \widehat{\beta}_0$. Les autres réponses sont bien entendu fausses. Ça correspond à la réponse C.



2 Quel est le rapport entre $\breve{\beta}_0$ et $\widehat{\beta}_0$?

A) $\breve{\beta}_0 = 2.202 \cdot \widehat{\beta}_0$

B) $\breve{\beta}_0 = 0.454 \cdot \widehat{\beta}_0$

C) $\breve{\beta}_0 = 1000 \cdot \widehat{\beta}_0$

D) $\breve{\beta}_0 = \widehat{\beta}_0$

E) aucune des réponses proposées

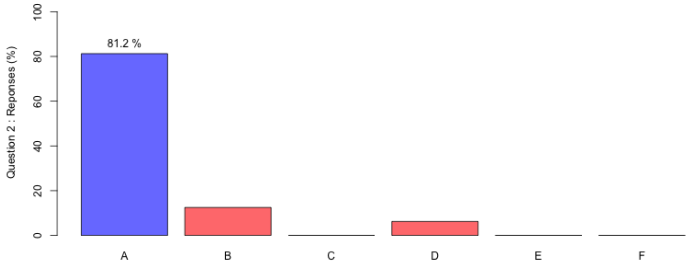
on sait que $\widehat{\text{poids}}^{\text{kg}} = \widehat{\beta}_0 + \widehat{\beta}_2 \ \text{cigarettes}^{\text{nb}} + \widehat{\beta}_3 \ \text{revenu}^{\$}$ donc on peut écrire

$$\widetilde{\text{poids}}_i^{\text{lb}} = \frac{\widehat{\text{poids}}^{\text{kg}}}{0.454} = \underbrace{\frac{\widehat{\beta}_0}{0.454}} + \frac{\widehat{\beta}_2 \ \text{cigarettes}^{\text{nb}}}{0.454} + \frac{\widehat{\beta}_3 \ \text{revenu}^{\$}}{0.454}$$

Par identification, on peut s'arrêter là pour affirmer

$$\breve{\beta}_0 = \frac{\widehat{\beta}_0}{0.454} = 2.202 \cdot \widehat{\beta}_0$$

qui correspond à la réponse A.



3 Quel est le rapport entre $\breve{\beta}_2$ et $\widehat{\beta}_2$?

A) $\breve{\beta}_2 = 2.202 \cdot \widehat{\beta}_2$
B) $\breve{\beta}_2 = 0.454 \cdot \widehat{\beta}_2$
C) $\breve{\beta}_2 = 2022 \cdot \widehat{\beta}_2$
D) $\breve{\beta}_2 = 454 \cdot \widehat{\beta}_2$
E) aucune des réponses proposées

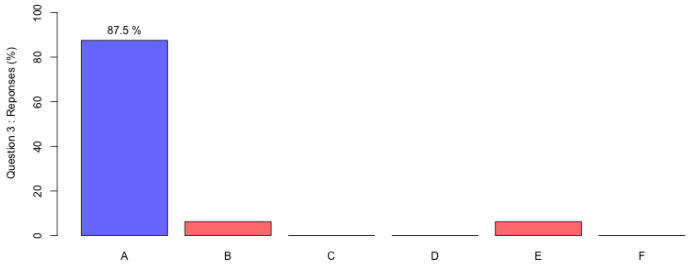
on vient de voir que

$$\widecheck{\text{poids}}_i^{\text{lb}} = \frac{\widehat{\text{poids}}^{\text{kg}}}{0.454} = \frac{\widehat{\beta}_0}{0.454} + \underbrace{\frac{\widehat{\beta}_2}{0.454}}\ \text{cigarettes}^{\text{nb}} + \frac{\widehat{\beta}_3\ \text{revenu}^{\$}}{0.454}$$

donc

$$\breve{\beta}_2 = \frac{\widehat{\beta}_2}{0.454} = 2.202 \cdot \widehat{\beta}_2$$

qui correspond à la réponse A.



4 Quel est le rapport entre $\check{\beta}_3$ et $\widehat{\beta}_3$?

A) $\check{\beta}_3 = 2.202 \cdot \widehat{\beta}_3$

B) $\check{\beta}_3 = 0.454 \cdot \widehat{\beta}_3$

C) $\check{\beta}_3 = 2022 \cdot \widehat{\beta}_3$

D) $\check{\beta}_3 = 454 \cdot \widehat{\beta}_3$

E) aucune des réponses proposées

Il y a ici deux bonnes réponses : la bonne réponse devait être la réponse C, sauf que ma dysléxie créatrice m'aura joué des tours : j'ai proposé comme réponse 2022 alors que la bonne réponse était 2202. Toutes mes excuses.

on vient de voir que

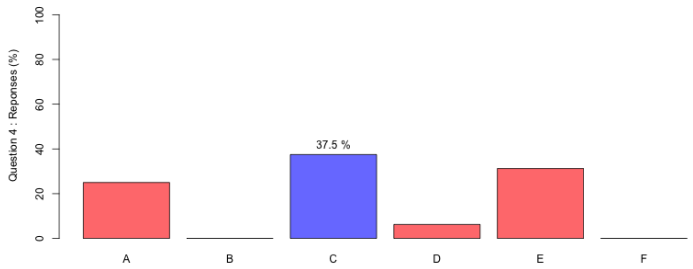
$$\widetilde{\text{poids}}_i^{\text{lb}} = \frac{\widehat{\text{poids}}^{\text{kg}}}{0.454} = \frac{\widehat{\beta}_0}{0.454} + \frac{\widehat{\beta}_2}{0.454}\,\text{cigarettes}^{\text{nb}} + \frac{\widehat{\beta}_3}{0.454}\,\text{revenu}^{\$}$$

or $\text{revenu}^{\$} = 1000 \cdot \text{revenu}^{\text{k\$}}$ donc

$$\widetilde{\text{poids}}_i^{\text{lb}} = \frac{\widehat{\text{poids}}^{\text{kg}}}{0.454} = \frac{\widehat{\beta}_0}{0.454} + \frac{\widehat{\beta}_2}{0.454}\,\text{cigarettes}^{\text{nb}} + \underbrace{\frac{1000\widehat{\beta}_3}{0.454}}\,\text{revenu}^{\text{k\$}}$$

$$\check{\beta}_3 = \frac{1000}{0.454}\widehat{\beta}_3 = 2202 \cdot \widehat{\beta}_3$$

Ça correspond à la réponse C.



5 Quel est le rapport entre $\widetilde{\beta}_0$ et $\widehat{\beta}_0$?

A) $\widetilde{\beta}_0 = \widehat{\beta}_0$ quel que soit $\widetilde{\beta}_1$

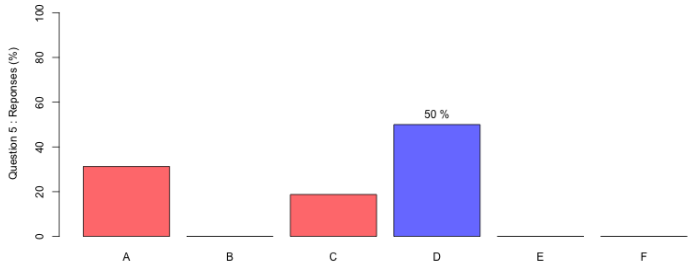
B) $\widetilde{\beta}_0 > \widehat{\beta}_0$ quel que soit $\widetilde{\beta}_1$

C) $\widetilde{\beta}_0 < \widehat{\beta}_0$ quel que soit $\widetilde{\beta}_1$

D) ça dépend de $\widetilde{\beta}_1$

E) aucune des réponses proposées

Il n'y a pas de résultat général.



6 Quel est le rapport entre $\widetilde{\beta}_3$ et $\widehat{\beta}_3$?

A) on a toujours $\widetilde{\beta}_3 = \widehat{\beta}_3$

B) on a $\widetilde{\beta}_3 = \widehat{\beta}_3$ si le revenu et le nombre de cigarettes fumées sont deux variables indépendantes

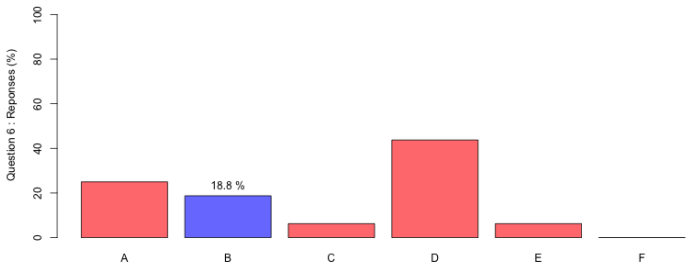
C) on a $\widetilde{\beta}_3 = \widehat{\beta}_3$ si le revenu et le revenu de la mère sont deux variables indépendantes

D) on a $\widetilde{\beta}_3 = \widehat{\beta}_3$ si le revenu et le fait de fumer sont deux variables indépendantes

E) aucune des réponses proposées

Il y avait une coquille ici pour la réponse C, qui n'a pas de sens. Copier/coller malencontreux. Même si on esaye de donner du sens, une variable ne peut pas être indépendante avec elle même, donc ça n'aurait aucun sens de la retenir. Cette question, c'est le théorème de Frish-Waugh : les variables explicatives sont indépendantes, avec l'estimation par moindres carrés sur le

modèle global correspond à l'estimation par moindres carrés sur chacunes des composantes. La réponse A est donc fausse ! Plus formellement, si on a $y_i = \beta_0 + \boldsymbol{x}_1^\mathsf{T}\boldsymbol{\beta}_1 + \boldsymbol{x}_2^\mathsf{T}\boldsymbol{\beta}_2 + v_i$ avec les variables de $\boldsymbol{x}_2$ indépendantes des variables de $\boldsymbol{x}_1$, alors $\boldsymbol{\beta}_2$ peut être obtenu en faisant *simplement* une régression sur les variables $\boldsymbol{\beta}_2$: considérons $y_i = \alpha_0 + \boldsymbol{x}_2^\mathsf{T}\boldsymbol{\alpha}_2 + u_i$, alors $\widehat{\boldsymbol{\beta}}_2 = \widehat{\boldsymbol{\alpha}_2}$. La condition est que les variables explicatives $\boldsymbol{x}_2$ soient soient indépendantes de $\boldsymbol{x}_1$. Ici, $\boldsymbol{x}_2$ c'est la variable de revenu (de la mère), et $\boldsymbol{x}_1$ ça inclu deux variables : l'indicatrice, correspondant au fait du fumer, et le nombre de cigarettes fumées. Si x_1 est le nombre de cigarettes fumées, $\boldsymbol{x}_1 = (\mathbf{1}_0(x), x$. Pour l'affirmation D, on nous dit juste que $\boldsymbol{x}_2$ est indépendante de $\mathbf{1}_0(x)$ (et pas de x) : ça ne suffit pas ! Pour l'affirmation B, on nous dit que $\boldsymbol{x}_2$ est indépendante de x, et donc a fortiori de n'importe quelle transformation de x (cf cours de probabilité). Aussi, $\boldsymbol{x}_2$ est indépendante de x, mais aussi de $\mathbf{1}_0(x)$ (et de $\mathbf{1}_A(x)$ pour tout A en fait). Bref, on a alors l'indépendance entre $\boldsymbol{x}_2$ et $\boldsymbol{x}_1$. Et dans ce cas, le théorème de Frish-Waugh s'applique. C'est la réponse B.



7 Comment interpréter le fait que $\widetilde{\beta}_1 < 0$?

A) il y a moins de mère qui ne fument pas que de mère qui fument dans la base de données

B) toutes choses égales par ailleurs, les bébés de poids faibles ont plus de chance de fumer quand ils seront adultes

C) toutes choses égales par ailleurs, le fait ne pas fumer fait que le poids à la naissance est plus faible

D) le revenu et le nombre de cigarettes fumées sont deux variables corrélées négativement

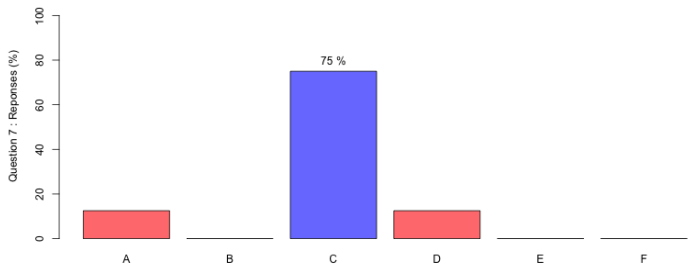
E) aucune des réponses proposées

On va aller assez vite, car les réponses A, B et D sont fausses. Reste à établir clairement que C est vraie A FAIRE

En faisant l'estimation sur $n = 100$ naissances, on obtient

$$\widehat{\text{poids}}_i^{\text{kg}} = \underset{(0.245)}{2.104} + \underset{(0.002)}{0.072}\ \text{cigarettes}_i^{\text{nb}} + \underset{(0.041)}{0.028}\ \text{revenu}_i^{\text{k\$}} \qquad (4)$$

(avec les notations usuelles).



8 Une femme ne fumant pas et ayant un revenu de 57,000$ annuel va accoucher. Quel serait le poids attendu de son bébé ?

A) 2.104 kg

B) 3.700 kg

C) 3.256 kg

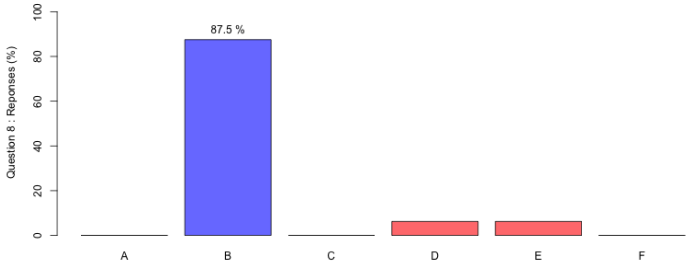
D) 4.321 kg

E) aucune des réponses proposées

En utilisant l'équation (4),

$$\widehat{\text{poids}}_i^{\text{kg}} = \underset{(0.245)}{2.104} + \underset{(0.002)}{0.072} \cdot 0 + \underset{(0.041)}{0.028} \cdot 57 = 2.104 + 0.028 \cdot 57 \approx 3.7 \text{ kg}$$

qui correspond à la réponse B.



9 La variable de revenu est-elle significative (au seuil classique de 95%) ?

A) oui

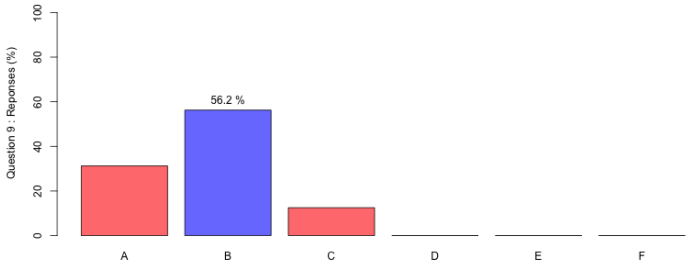
B) non

C) on ne peut pas savoir

D) ——

E) ——

Sous hypothèses de normalité, on a un intervalle de confiance à 95% de la forme $[0.028 \pm 2 \cdot 0.041]$ soit $[-0.054; 0.11]$. Comme 0 est dans cet intervalle, on rejette l'hypothèse H_1 : $\beta_{\text{revenu}} \neq 0$. La variable n'est pas significative. Pour rappel, "*pas significitive*" ou "*non significitive*" est un raccourcit pour "*pas significitivement non-nulle*", désolé pour toutes les négations, mais c'est la formulation juste. La version courte (mais logiquement fausse) serait "*significitivement nulle*" autrement dit on se demande si la vraie valeur (inconnue) peut être nulle. Ici, oui, donc on va dire qu'elle n'est "*pas significitivement non-nulle*".



On se rend compte que la variable "genre" (du bébé) a été observée, et qu'il y a 50 garçons ("H") et 50 filles ("F"). On suppose que le genre est indépendant du revenu de la mère, et du fait qu'elle soit fumeuse (ou pas). On estime

$$\widehat{\text{poids}}_i^{\,\text{kg}} = \widehat{\alpha}_0 + \widehat{\alpha}_1 \ \mathbf{1}_{\text{H}}(\text{genre}_i) + \underset{(0.002)}{0.072} \ \text{cigarettes}_i^{\text{nb}} + \underset{(0.041)}{0.028} \ \text{revenu}_i^{\text{k\$}}$$

On suppose que les bébés garçons sont, en moyenne et toutes choses étant égales par ailleurs, (strictement) plus lourds que les bébés filles.

10 Que peut-on dire sur $\widehat{\alpha}_0$ et $\widehat{\alpha}_1$?

A) $\widehat{\alpha}_0 < \widehat{\alpha}_1$

B) $\widehat{\alpha}_0 \neq \widehat{\alpha}_1$

C) $\widehat{\alpha}_1 > 0$

D) $\widehat{\alpha}_1 > 2.104$

E) aucune des réponses proposées

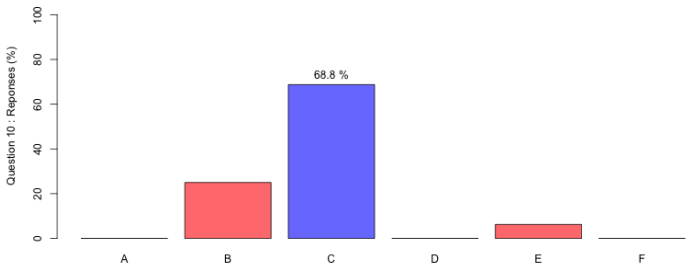
Oublions les deux dernières variables explicatives un instant, et considérons le modèle

$$\widehat{\text{poids}}_i^{\,\text{kg}} = \widehat{a}_0 + \widehat{a}_1 \ \mathbf{1}_{\text{H}}(\text{genre}_i)$$

Dans ce cas, $\widehat{a}_0$ correspond au poids moyen des filles (qui est ici la modalité de référence), et $\widehat{a}_0 + \widehat{a}_1$ est le poids moyen des garçons. Autrement dit, $\widehat{a}_1$ est la différence de poids à la naissance des garçons, par rapport aux filles. Or on nous dit "on suppose que les bébés garçons sont, en moyenne et toutes choses étant égales par ailleurs, (strictement) plus lourds que les bébés filles", autrement dit $\widehat{a}_1 > 0$. On utilise maintenant pour conclure un autre argument vu en cours, le théorême de Frish-Waugh : comme "le genre est indépendant du revenu de la mère, et du fait qu'elle soit fumeuse (ou pas)", la variable $\mathbf{1}_{\text{H}}(\text{genre})$ est indépendante du couple $(\text{cigarettes}^{\text{nb}}, \text{revenu}^{\text{k\$}})$, et donc $\widehat{a}_1 = \widehat{\alpha}_1$. Aussi, $\widehat{\alpha}_1 > 0$. La réponse C est donc correcte. Comme il n'y en a qu'une on pourrait s'arrêter là. Mais comme je présente des éléments de correction, je pense qu'il expliquer au moins sommairement pourquoi les autres sont faux. Pour le premier, on ne peut pas savoir. Dans l'exemple précédant, si on avait écrit

$$\widehat{\text{poids}}_i^{\,\text{kg}} = \widehat{a}_0 \mathbf{1}_{\text{F}}(\text{genre}_i) + \widehat{a}_1 \ \mathbf{1}_{\text{H}}(\text{genre}_i)$$

alors effectivement, on aurait $\widehat{a}_0 < \widehat{a}_1$. Mais ce n'est pas ce qu'on a. Et nulle part il n'est dit que le poids des garçons est plus du double de celui des filles. Pour la réponse B, on n'a pas l'information pour conclure. Et D, on ne sait pas non plus, encore une fois $\widehat{\alpha}_1$ est un différentiel de poids moyen, entre les garçons et les filles, à la naissance.



11 Que peut-on dire sur $\widehat{\alpha}_0$ et $\widehat{\alpha}_1$?

A) $\widehat{\alpha}_0 + \widehat{\alpha}_1 = 2.104$

B) $\widehat{\alpha}_0 \cdot \widehat{\alpha}_1 = 2.104$

C) $\widehat{\alpha}_1 = 0$ car il y a autant de garçons que de filles dans l'échantillon

D) $\widehat{\alpha}_0 < 2.104$

E) aucune des réponses proposées

Revenons sur notre écriture précédante. On a écrit ici

$$\widehat{\text{poids}}_i^{\,\text{kg}} = \widehat{\alpha}_0 + \widehat{\alpha}_1 \, \mathbf{1}_{\text{H}}(\text{genre}_i) + \text{quelque chose}_i$$

or l'équation de départ - équation (4) - était

$$\widehat{\text{poids}}_i^{\,\text{kg}} = 2.104 + \text{même quelque chose}_i$$

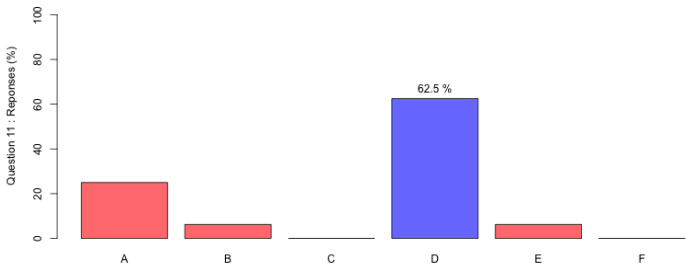
Si on oublie le terme de droite, 2.104 est le poids à la naissance d'un bébé (si on n'oublie pas le terme de droite, ça serait $2.104 + 0.072 \, \overline{\text{cigarette}}^{\text{nb}} + 0.028 \, \overline{\text{revenu}}^{\text{k\$}}$) peu importe le genre. Donc

$$2.104 = \text{proportion fille} \cdot \widehat{\alpha}_0 + \text{proportion garçon} \cdot [\widehat{\alpha}_0 + \widehat{\alpha}_1]$$

soit

$$2.104 = \widehat{\alpha}_0 + \underbrace{\text{proportion garçon} \cdot \widehat{\alpha}_1}_{>0} > \widehat{\alpha}_0,$$

qui correspond à la réponse D. Les autres réponses sont fausses...



Pour les questions 12 à 14, on considère le (vrai) modèle suivant

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i \tag{5}$$

où les observations sont des réalisation de variables aléatoires, avec en particulier $\mathbb{E}[\varepsilon] = 0$. On suppose de plus que $\text{cov}[X_1, X_2] > 0$. On estime trois modèles. Le premier est correctement spécifié

$$\widetilde{y}_i = \widetilde{\beta}_0 + \widetilde{\beta}_1 x_{1,i} + \widetilde{\beta}_2 x_{2,i} \text{ et } \widetilde{\varepsilon}_i = y_i - \widetilde{y}_i \tag{6}$$

Le second est mal spécifié (avec une sous-spécification)

$$\breve{y}_i = \breve{\beta}_0 + \breve{\beta}_1 x_{1,i} \text{ et } \breve{\varepsilon}_i = y_i - \breve{y}_i \tag{7}$$

ainsi que le troisième (avec une sur-spécification)

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1,i} + \widehat{\beta}_2 x_{2,i} + \widehat{\beta}_3 x_{3,i} \text{ et } \widehat{\varepsilon}_i = y_i - \widehat{y}_i \tag{8}$$

On suppose que la troisième variable vérifie $\text{cov}[X_1, X_3] < 0$.

12 Laquelle des affirmations ci-dessous est juste ?

A) comme le modèle (6) est correctement spécifié $\widetilde{\varepsilon}_i = \varepsilon_i$

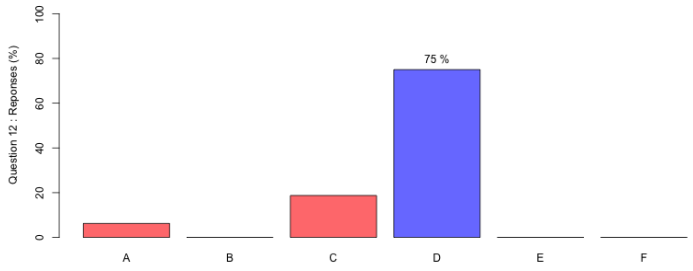
B) comme le modèle (7) est sous-spécifié et (8) est sur-spécifié, pour tout i, $\widehat{\varepsilon}_i \geqslant \breve{\varepsilon}_i$

C) si $R^2_{(j)}$ est le R^2 du modèle (j), $R^2_{(7)} \leqslant R^2_{(6)}$ et $R^2_{(8)} \leqslant R^2_{(6)}$

D) si $R^2_{(j)}$ est le R^2 du modèle (j), $R^2_{(7)} \leqslant R^2_{(6)} \leqslant R^2_{(8)}$

E) aucune des réponses proposées

Cette fois, on peut y aller par élimination. La réponse A est fausse, ce n'est pas ça la notion de "correctement spécifié", ça ressemble plutôt à un modèle parfait en fait... Pareil pour la réponse B. Pour les deux suivantes, la bonne réponse est la réponse D, c'est une propriété des modèles dits "imbriqués' : si on a (1) $y_i = \boldsymbol{x}_1^\intercal \boldsymbol{\alpha}_1 + u_i$ et (2) $y_i = \boldsymbol{x}_1^\intercal \boldsymbol{\beta}_1 + \boldsymbol{x}_2^\intercal \boldsymbol{\beta}_2 + v_i$ alors forcément $R^2_{(2)} \geqslant R^2_{(1)}$.



13 Laquelle des affirmations ci-dessous est juste ?

A) $\check{\beta}_1 = \widetilde{\beta}_1$

B) $\check{\beta}_1$ est un estimateur sans biais de β_1

C) comme le modèle (7) est sous-spécifié, alors pour tout i, $\check{\varepsilon}_i < \varepsilon_i$

D) si on ordonne les observations suivant la valeur des y_i (avec $y_i \leqslant y_{i+1}$), $\check{\varepsilon}_i \leqslant \check{\varepsilon}_{i+1}$ pour tout i.

E) aucune des réponses proposées

Pour la première affirmation, on l'avait vu en cours. Revenons sur l'écriture matricielle classique,

$$\begin{bmatrix} \check{\beta}_0 \\ \check{\beta}_1 \end{bmatrix} = (\boldsymbol{X}_1^\intercal \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\intercal \boldsymbol{y}$$

$$\begin{bmatrix} \widetilde{\beta}_0 \\ \widetilde{\beta}_1 \\ \widetilde{\beta}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1^\intercal \boldsymbol{X}_1 & \boldsymbol{X}_1^\intercal X_2 \\ X_2^\intercal \boldsymbol{X}_1 & X_2^\intercal X_2 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X}_1^\intercal \boldsymbol{y} \\ X_2^\intercal \boldsymbol{y} \end{bmatrix}$$

aussi

$$\widetilde{\beta}_1 = \underbrace{(\boldsymbol{X}_1' \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\intercal \boldsymbol{y}}_{\check{\beta}_1} + \underbrace{(\boldsymbol{X}_1' \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^\intercal X_2 \check{\beta}_2}_{\check{\beta}_{12}}$$

autrement dit $\widetilde{\beta}_1 = \breve{\beta}_1 + \breve{\beta}_{12} \neq \breve{\beta}_1$, à moins que $\boldsymbol{X}_1 \perp \boldsymbol{X}_2$. Or justement, il est précisé que $\text{cov}[X_1, X_2] > 0$, donc A est fausse. Et pareil pour l'affirmation B :

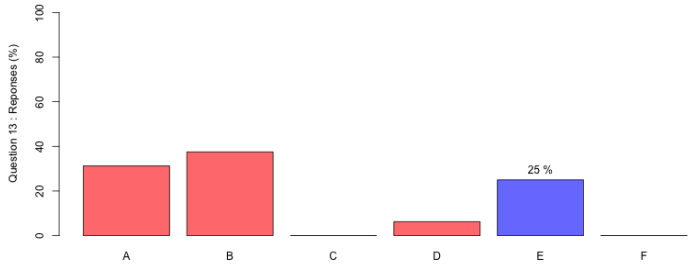
$$\mathbb{E}[\breve{\boldsymbol{\beta}}] = \mathbb{E}\big[(\boldsymbol{X}_1^\intercal \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\intercal \boldsymbol{y}\big] = \mathbb{E}\big[(\boldsymbol{X}_1^\intercal \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\intercal \begin{bmatrix}\boldsymbol{X}_1 & X_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta} \\ \beta_2\end{bmatrix} + \varepsilon\big] \neq \boldsymbol{\beta}$$

En fait, on aurait pu se limiter à la discussion précédante : on a vu que $\widetilde{\beta}_1 \neq \breve{\beta}_1$, donc a fortiori, $\mathbb{E}[\widetilde{\beta}_1] \neq \mathbb{E}[\breve{\beta}_1]$. Or $\widetilde{\beta}_1$ est l'estimateur par moindres carrés sur le bon modèle : on sait que c'est un estimateur sans biais

$$\mathbb{E}[\widetilde{\boldsymbol{\beta}}] = \mathbb{E}\big[(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \boldsymbol{y}\big] = \mathbb{E}\big[(\boldsymbol{X}^\intercal \boldsymbol{X})^{-1}\boldsymbol{X}^\intercal \boldsymbol{X}\boldsymbol{\beta} + \varepsilon\big] = \boldsymbol{\beta}$$

Bref, cette réponse n'était pas valide non plus.

Les propositions C et D aussi - on n'a aucun résultats sur les valeurs des résidues individuels dans le cours ! La bonne réponse est la réponse E.



14 Laquelle des affirmations ci-dessous est juste ?

A) La variance de $\widehat{\beta}_1$ est plus grande que la variance de $\widetilde{\beta}_1$

B) comme le modèle (8) est sur-spécifié, alors pour tout i, $\widehat{\varepsilon}_i > \varepsilon_i$

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

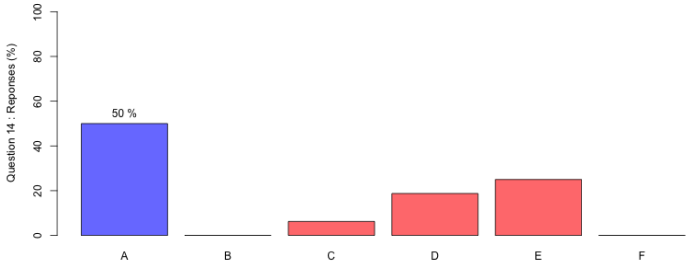
D) $\widehat{\beta}_3 \leqslant 0$ car $\text{cov}[X_1, X_3] < 0$

E) aucune des réponses proposées

On l'avait vu en cours, et les réponses B, C, et D n'ont aucune raison d'être valide !

Pour les questions 15 à 19, considérons le modèle

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \varepsilon_i \tag{9}$$



15 Que signifie l'hypothèse d'*homoscédasticité* du modèle ?

A) $\text{Var}[X_1] = \text{Var}[X_2] = \text{Var}[X_3]$

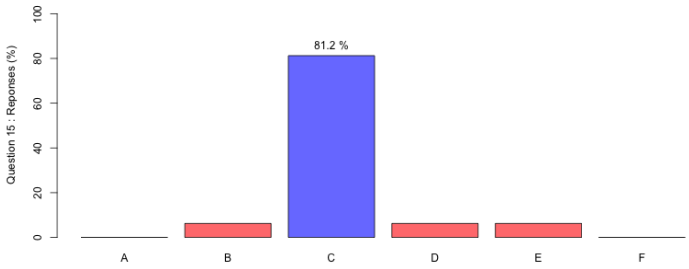
B) $\text{Var}[X_1|Y] = \text{Var}[X_2|Y] = \text{Var}[X_3|Y]$

C) $\text{Var}[\varepsilon|X_1, X_2, X_3]$ est constante

D) $\text{Var}[X_1|Y]$, $\text{Var}[X_2|Y]$ et $\text{Var}[X_3|Y]$ sont constantes (pas forcément égales)

E) aucune des réponses proposées

Dit rapidement, oralement, l'homoscédasticité, c'est une variance des résidus qui ne dépendant pas des variables explicatives. "*On parle d'homoscédasticité lorsque la variance des erreurs stochastiques de la régression est la même pour chaque observation*" (wikipedia)



16 Il n'est pas rare de supposer que $\mathbb{E}[\varepsilon|X_1, X_2, X_3] = 0$. Qu'est-ce que cette hypothèse implique ?

A) $\mathbb{E}[\varepsilon] = 0$

B) $\text{Var}[\varepsilon|X_1, X_2, X_3] = 0$

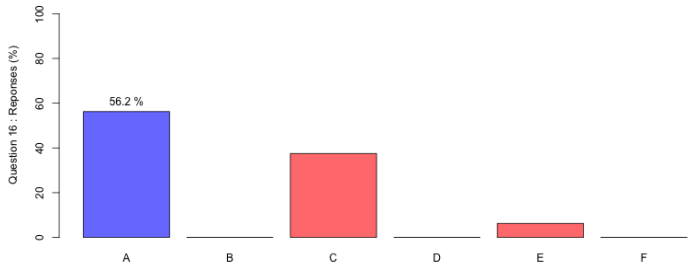
C) $\text{Var}[\varepsilon|X_1, X_2, X_3]$ est constante (pas forcément nulle)

D) $\text{Var}[\varepsilon]$ est constante

E) aucune des réponses proposées

C'est ici l'hypothèse que les résidus sont non-prédits par les variables explicatives. Ca n'implique rien du tout sur la variance des résidus ! La seule réponse possible est la réponse A). Et la preuve est simple : c'est le fait que $\mathbb{E}[\mathbb{E}[\varepsilon|\star]] = \mathbb{E}[\varepsilon]$.

Pour les deux questions suivantes, on procéde à une estimation par moindres carrés, et on note $\widehat{\beta}_j$ l'estimateur de β_j, pour $j = 0, 1, 2, 3$.



17 Quelle affirmation donnera une plus grande variance pour $\widehat{\beta}_3$?

A) si la taille de l'échantillon augmente (toutes autres choses étant égales par ailleurs)

B) si la variance de X_3 augmente (toutes autres choses étant égales par ailleurs)

C) si la valeur absolue de la corrélation entre X_1 et X_3 diminue (toutes autres choses étant égales par ailleurs)

D) si on calcule l'estimateur après avoir ordonné les observations suivant la valeur des y_i dans l'ordre croissant

E) aucune des réponses proposées

Pour commencer, revenons au cas de la régression simple (en gros, oublions X_1 et X_2. Dans ce cas, avec juste la constance et X_3,

$$\widehat{\beta}_3 = \frac{\sum x_i \sum y_i - n \sum x_i y_i}{\left(\sum x_i\right)^2 - n \sum x_i^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

et plus intéressant, sa variance est ici

$$\mathsf{Var}[\widehat{\beta}_3] = \frac{\sigma^2}{(n-1)\mathsf{Var}[X_3]}$$

autrement dit, la variance de $\widehat{\beta}_3$ augmente avec celle des résidus (σ^2, pas proposé ici), et augmente lorsque n diminue (A est donc fausse) et lorsque $\mathsf{Var}[X_3]$ diminue (donc B est fausse). La réponse D n'a pas de sens (et personne ne l'avait retenue). Reste la réponse C. Cette dernière est fausse, car heuristiquement, une hausse de la corrélation entre deux variables explicatives rend l'estimation de l'effet des variables moins précis : si les variables sont très très corrélées, on ne peut pas choisir entre les deux variables... Plus formellement, revenons alors à la régression multiple,

$$\mathsf{Var}[\widehat{\boldsymbol{\beta}}] = \sigma^2 (\boldsymbol{X}\,\boldsymbol{X})^{-1}$$

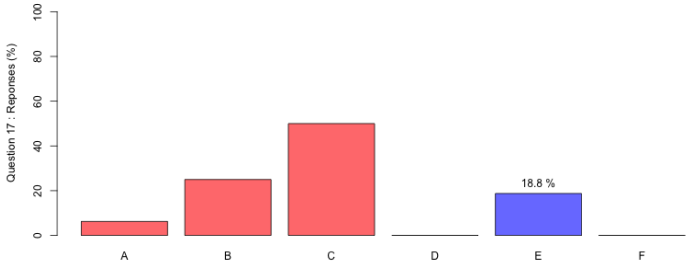
autrement dit, au 'dénominateur', on retrouve la variance de $\boldsymbol{X}$. Comme le note Uriel 2013 (pdf), équation 3-64, on peut écrire

$$\mathsf{Var}[\widehat{\beta}_3] = \frac{\sigma^2}{(n-1)\mathsf{Var}[X_3] \cdot (1 - R_3^2)}$$

où R_3^2 est le R^2 de la régression de X_3 sur toutes les variables explicatives. Ici, le "toutes choses étant égales para ailleurs" s'interprète en disant qu'on se contente de régresser sur X_1, et donc R_3^2 correspond à la corrélation entre X_1 et X_3. Bref,

$$\mathsf{Var}[\widehat{\beta}_3] = \frac{\sigma^2}{(n-1)\mathsf{Var}[X_3] \cdot (1 - \mathsf{cor}[X_1, X_3]^2)}.$$

Aussi, on retrouve ici de manière plus formelle le résultat intuité : plus de corrélation augmente la variance. Ou moins de corrélation fera baisser la variance de notre estimateur. Bref, C était aussi fausse. Une petite simulation donnée en annexe permet de tout visualiser...



18 On note $[b^-_{n,\alpha\%}, b^+_{n,\alpha\%}]$ l'intervalle de confiance de $\widehat{\beta}_3$ de niveau $\alpha\%$ obtenu à partir des n observations. Quelle affirmation parmi les suivantes est **fausse** ?

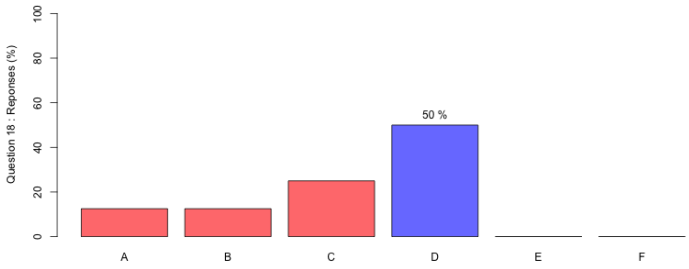
A) $[b^-_{n,90\%}, b^+_{n,90\%}]$ est toujours inclu dans $[b^-_{n,99\%}, b^+_{n,99\%}]$

B) si on suppose $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $b^+_{n,\alpha\%} - b^-_{n,\alpha\%}$ est proportionnel à σ

C) $[b^-_{n,99\%}, b^+_{n,99\%}]$ contient toujours $\widehat{\beta}_3$

D) $[b^-_{n,99\%}, b^+_{n,99\%}]$ contient toujours β_3

E) aucune des réponses proposées



La réponse C était juste : les intervalles de confiance sont de la forme $\left[\widehat{\beta}_3 \pm q_{1-\alpha/2}\sqrt{\mathsf{Var}[\widehat{\beta}_3]}\right]$ et sont donc toujours centrés sur $\widehat{\beta}_3$. Dans la formule $q_{1-\alpha/2}$ est le quantile de niveau $1-\alpha/2$ de la loi normale (ou de la loi

de Student si le nombre d'observations est trop faible). On retrouve aussi que A est juste, car $q_{99\%} > q_{90\%}$. B est aussi juste. En effet, $\text{Var}[\widehat{\beta}_3]$ est $\sigma^2\Omega_{3,3}$ (où $\Omega = (\boldsymbol{X}^\text{T}\boldsymbol{X})^{-1}$). Aussi, $\text{Var}[\widehat{\beta}_3]$ est proportionnel à σ. Bref, B est juste. Maintenant regardons D : si c'était vrai, on pourrait trouver β_3, non ? Bref, la seule réponse fausse était la réponse D.

19 Si on suppose que $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, alors $[b^-_{n,95\%}, b^+_{n,95\%}]$ s'écrit

A) $\left[\widehat{\beta}_3 \pm \dfrac{1.96}{\sqrt{n}}\sigma\right]$

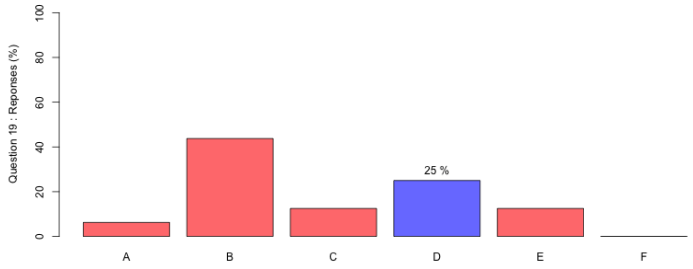
B) $\left[\widehat{\beta}_3 \pm 1.96\sigma\right]$

C) $\left[\widehat{\beta}_3 \pm \dfrac{1.96}{\sqrt{n}}s\right]$ avec $s > \sigma$

D) $\left[\widehat{\beta}_3 \pm 1.96s\right]$ avec $s > 0$, a priori différent de σ et de $\sigma/\sqrt{n}$

E) aucune des réponses proposées

La formule est $\left[\widehat{\beta}_3 \pm 1.96\sqrt{\text{Var}[\widehat{\beta}_3]}\right]$. Le terme de variance n'est pas la variance des résidus, mais la variance de l'estimateur !



20 On rajoute ici une variable X_4 pour construire le modèle

$$y_i = \alpha_0 + \alpha_1 x_{1,i} + \alpha_2 x_{2,i} + \alpha_3 x_{3,i} + \alpha_4 x_{4,i} + \eta_i \tag{10}$$

Soit $R^2_{(j)}$ le coefficient de déterminiation, R^2, pour le modèle (j). A quelle condition a-t-on $R^2_{(10)} \geqslant R^2_{(9)}$?

A) si X_4 est corrélée positivement avec *au moins* une variable X_j (avec $j = 1, 2, 3$), mais pas forcément toutes

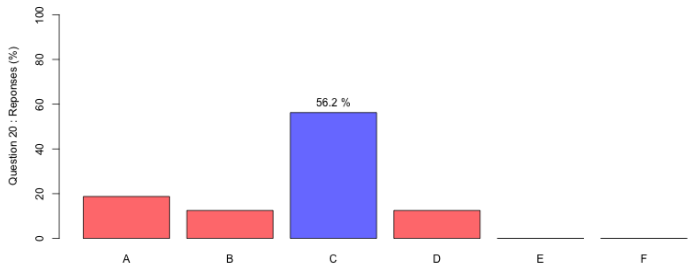
B) si X_4 est corrélée positivement avec *toutes* les variables X_j (avec $j = 1, 2, 3$)

C) on a toujours $R^2_{(10)} \geqslant R^2_{(9)}$

D) si $\beta_0 \geqslant \alpha_0$

E) aucune des réponses proposées

La différence entre le modèle (10) et (9) est qu'on a rajouté une variable explicative ! Autrement dit, comme on l'a vu en cours, le R^2 va mécaniquement augmenter. Donc $R^2_{(10)} \geqslant R^2_{(9)}$, peut importe la variable que l'on rajoute. En fait, si on rajoue x_4 qui est une des autres variables (disons x_1), alors $R^2_{(10)} = R^2_{(9)}$. Mais sinon, le R^2 augmentera !



Pour les questions 21 à 26, on considère le moèle suivant :

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \varepsilon_i$$

en utilisant l'écriture vectorielle classique. Soit $\boldsymbol{X}$ la matrice $n \times 3$ dont les lignes sont les $\boldsymbol{x}_i$. On suppose que

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{pmatrix} 25 & 0 & 0 \\ \star & 9.3 & 5.4 \\ \star & \star & 12.7 \end{pmatrix} \text{ et } (\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \begin{pmatrix} a & 0 & 0 \\ \star & 0.1428 & -0.0607 \\ \star & \star & 0.1046 \end{pmatrix}$$

où $\star$ indique une valeur supprimée.

21 Que vaut a dans le coin supérieur gauche de la matrice de droite ?

A) on ne peut pas savoir

B) 0.04

C) 0.004

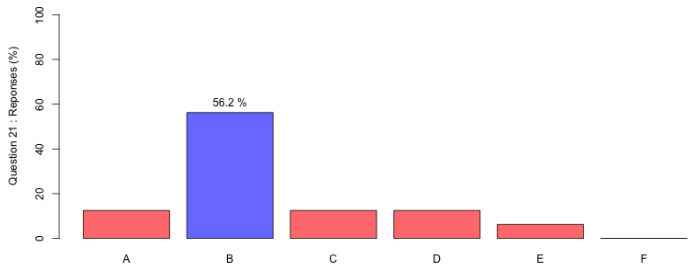
D) 0.1854

E) aucune des réponses proposées

Ca avait été fait en démo, c'est l'inverse de la valeur dans le coin supérieur gauche de la matrice $\boldsymbol{X}^\top\boldsymbol{X}$. C'est une propriété des matrices diagonales par blocs :

$$\boldsymbol{X}^\top\boldsymbol{X} = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \text{ alors } (\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \begin{pmatrix} A^{-1} & 0 \\ 0 & B^{-1} \end{pmatrix} \text{ avec ici } A = \{25\}.$$

Or $1/25 = 0.04$, donc on retient la réponse B.



22 Combien y-a-t-il d'observations n ?

A) on ne peut pas savoir

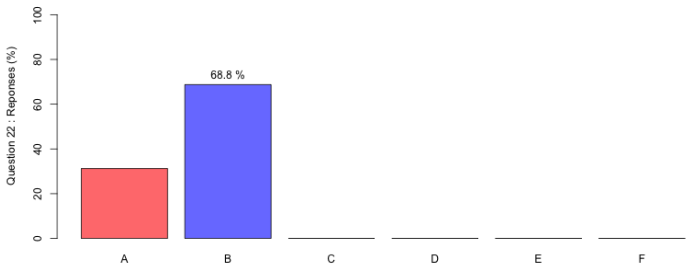
B) 25

C) 5

D) 100

E) aucune des réponses proposées

Dans la matrice $\boldsymbol{X}^\top\boldsymbol{X}$, comme la première colonne de $\boldsymbol{X}$ est un vecteur avec n fois 1, le terme en haut à gauche de $\boldsymbol{X}^\top\boldsymbol{X}$ est la somme de n 1, soit n. Donc $n = 25$.



23 Que vaut (environ) la corrélation empirique entre les vecteurs des deux variables X_1 et X_2 ?

A) on ne peut pas savoir

B) 0.5

C) 0.25

D) 0.707

E) aucune des réponses proposées

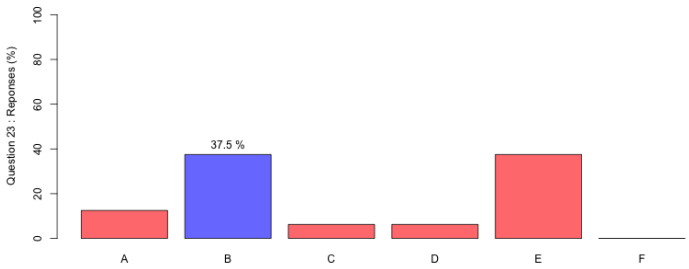
On va utiliser une propriété de la matrice $\boldsymbol{X}^\top \boldsymbol{X}$: les termes sur la première ligne et la première colonne (par symmétrie) sont nuls (sauf le premier). Or ces termes sont les sommes des x_i ! Autrement dit $\overline{x}_1$ et $\overline{x}_2$ sont nuls. Si on regarde maintenant le bloc inférieur de la matrice

$$\boldsymbol{X}^\top \boldsymbol{X} = \left(\begin{array}{c|cc} 25 & 0 & \\ \hline 0 & 9.3 & 5.4 \\ & 5.4 & 12.7 \end{array}\right)$$

on retrouve la matrice de variance empirique du couple (X_1, X_2). Donc

$$\text{corrélation}[X_1, X_2] = \frac{\text{covariance}[X_1, X_2]}{\sqrt{\text{variance}[X_1] \cdot \text{variance}[X_2]}} = \frac{5.4}{\sqrt{9.3 \cdot 12.7}}$$

soit envrion 0.4968. On retiendra la réponse B.



24 On suppose que $y_i = 1.60 + 0.61\ x_1 + 0.46\ x_2 + \widehat{\varepsilon}_i$, avec

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = 0.3$$

Que vaut $\overline{y}$ (la moyenne empirique des observations y_i)

A) on ne peut pas savoir

B) 1.6

C) -1.6

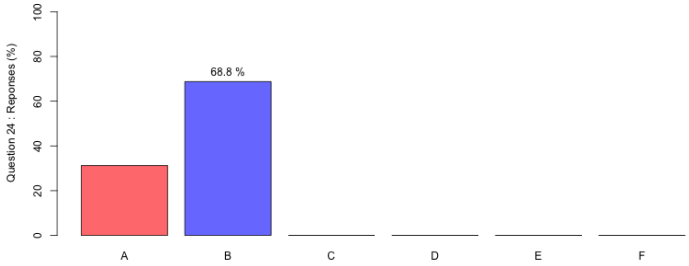
D) 0

E) aucune des réponses proposées

On a vu tantôt que $\overline{x}_1$ et $\overline{x}_2$ sont nuls. Or on sait que la droite de régression passe par les points moyens (c'est la première des condition du premier ordre - quand on dérive par rapport à la constante β_0) au sens où

$$\overline{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x}_1 + \widehat{\beta}_2 \overline{x}_2 = 1.60 + 0.61 \cdot \overline{x}_1 + 0.46 \cdot \overline{x}_2 = 1.60,$$

qui est la réponse B.



25 Que vaut $\overline{\widehat{y}}$ (la moyenne empirique des prévisions $\widehat{y}_i$)

A) on ne peut pas savoir

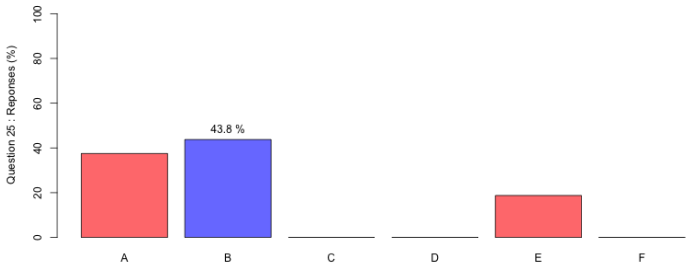
B) 1.6

C) -1.6

D) 0

E) aucune des réponses proposées

Par construction (c'est toujours la condition du premier ordre), $\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0$, et donc $\sum_{i=1}^{n} \widehat{y}_i = \sum_{i=1}^{n} y_i$ ce qui se réécrit, en divisant simplement par n, $\overline{\widehat{y}} = \overline{y}$, donc compte tenu de la réponse précédante, c'est la réponse B.



26 Que vaut le coefficient de détermination R^2

A) on ne peut pas savoir

B) 78.2%

C) 90.5%

D) 96.8%

E) aucune des réponses proposées

C'est la question la plus calculatoire du devoir. On avait vu en cours que le R^2 correspond à

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

où SCR est la somme des carrés des résidus, qu'on nous donne, et SCE est la somme des carrées "expliqués". Plus précisément, on a la formule de décomposiiton de la variance

$$\underbrace{\sum_{i=1}^{n}(y_i - \overline{y})^2}_{=SCT} = \underbrace{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}_{=SCE} + \underbrace{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}_{=SCR} = \underbrace{\sum_{i=1}^{n}(\widehat{\beta}_1 x_{1,i}\widehat{\beta}_2 x_{2,i})^2}_{=SCE} + \underbrace{\sum_{i=1}^{n}\widehat{\varepsilon}_i^2}_{=0.3}.$$

Pour le terme SCT, il va falloir calculer :

$$\sum_{i=1}^{n}(\widehat{\beta}_1 x_{1,i}\widehat{\beta}_2 x_{2,i})^2 = \widehat{\beta}_1^2\sum_{i=1}^{n}x_{1,i}^2 + 2\widehat{\beta}_1\widehat{\beta}_2\sum_{i=1}^{n}x_{1,i}x_{2,i} + \widehat{\beta}_2^2\sum_{i=1}^{n}x_{2,i}^2$$

l'avantage c'est que les termes sont dans le petit bloc de la matrice $\boldsymbol{X}^\top\boldsymbol{X}$,

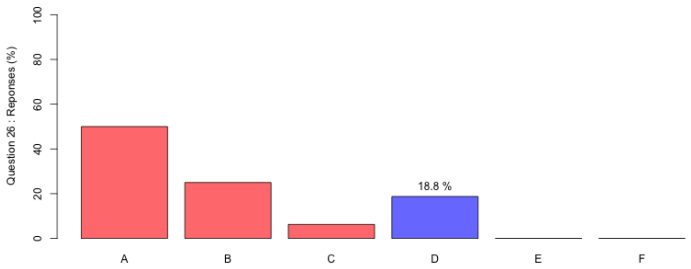
$$\sum_{i=1}^{n}(\widehat{\beta}_1 x_{1,i}\widehat{\beta}_2 x_{2,i})^2 = 0.61^2 \cdot 9.3 + 2 \cdot 0.61 \cdot 0.46 \cdot 5.4 + 0.46^2 \cdot 12.7 = 9.17833$$

On peut maintenant calculer le R^2,

$$R^2 = \frac{9.178}{9.178 + 0.3} \approx 0.968$$

autrement dit, 97% de la variance des données est ici expliquée par le modèle de régression. On pourrait même calculer le R^2-ajusté,

$$\overline{R}^2 = 1 - \frac{n-1}{n-3}[1 - R^2] \approx 0.965.$$



Pour les questions 27 à 30, On considère un modèle estime par moindres carrés

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \text{ et } \widehat{\varepsilon}_i = y_i - \widehat{y}_i \tag{11}$$

On veut faire une prévision pour un nouveau point $x^\star$, et on pose

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

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

A) on ne peut pas savoir

B) quand $x^\star = 0$

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

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

E) aucune des réponses proposées

On l'avait vu plusieurs fois en cours, et aucune justification n'était demandée ici ! Mais on peut faire le calcul. On sait que la variance de $\widehat{Y}^\star$ est

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

On cherche alors le minimum de la fonction

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

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

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

autrement dit, le minimum est obtenu pour

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

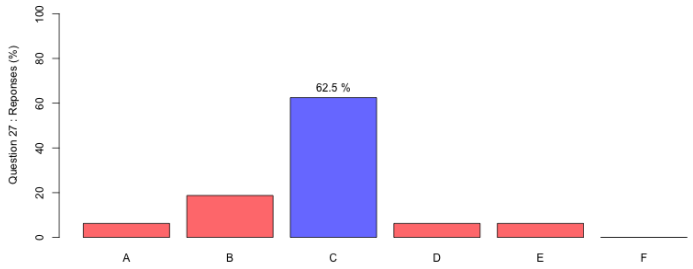
Or on sait que

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

donc

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

qui est la réponse C.



La régression informatique sur des vraies données donne la sortie suivante

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.72385    0.12974    ?     < 2e-16 ***
X            -0.27793        ?     -11.04 1.05e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.3548 on 28 degrees of freedom
Multiple R-Squared: 0.8131,     Adjusted R-squared: 0.8064
F-statistic: 121.8 on 1 and 28 DF,  p-value: ?
```

28 Que vaut la statistique de test pour le test de significativité de la constante ?

A) 10.39404

B) -36.41013

C) -280.6392

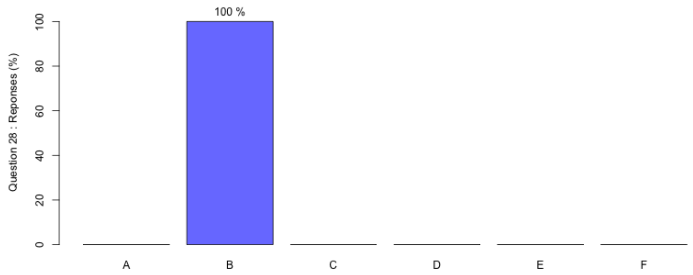
D) 0.027464

E) aucune des réponses proposées

C'est le test de Student, de l'hypothèse $H_0 : \beta_0 = 0$. On pose alors ici

$$T = \frac{\widehat{\beta}_0}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_0]}} = \frac{-0.472385}{0.12974} \approx -36.41013$$

qui est la réponse B.



29 Que vaut la p-value pour le test de Fisher (derniére ligne) ?

A) $8.21 \cdot 10^{-3}$

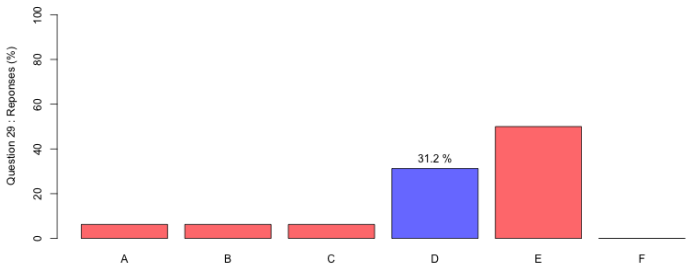
B) $3.24 \cdot 10^{-6}$

C) $2.29 \cdot 10^{-1}$

D) $1.05 \cdot 10^{-11}$

E) on ne peut pas savoir

De manière générale, le test de Fisher est un test de $H_0 : \beta_1 = \cdots = \beta_k = 0$, dans un modèle multiple. On a une seule variable explicative ici, donc on teste $H_0 : \beta_1 = 0$, ce qui est le même test que le test de significativité de Student. Si les statistiques de test sont différentes, les tests sont équivalents, et en particulier, ils ont la même p-value. Compte tenu de la seconde ligne, on en déduit que la p-value est de l'ordre de $1.05 \cdot 10^{-11}$



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

$$y_i = 2 + \beta_1 x_i + \varepsilon_i \tag{12}$$

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

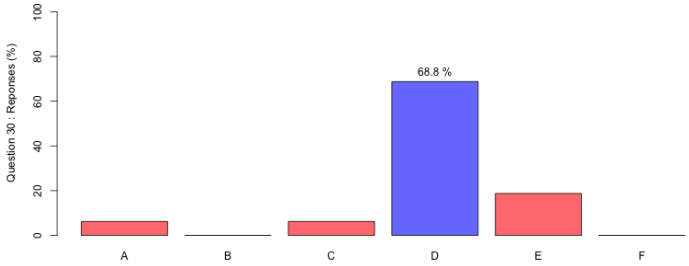
Pour éviter de dire n'importe quoi, on va tranquillement l'écrire ! On commence par écrire la fonction objectif, qui est la somme des carrés des erreurs

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

La condition du premier ordre s'écrit alors

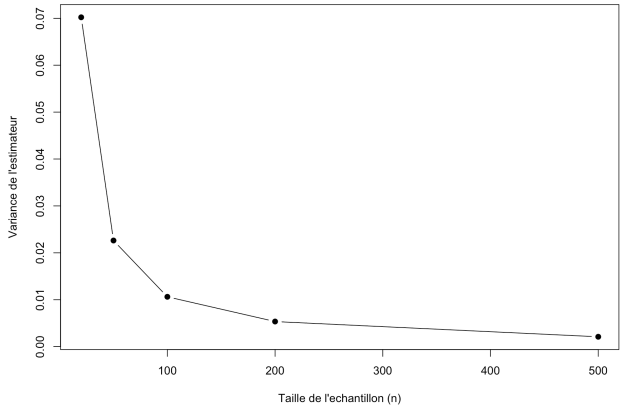
$$\sum_i -2x_i(y_i - 2 - \widetilde{\beta}_1 x_i) = 0 \text{ soit } \sum_i x_i(y_i - 2) = \sum_i \widetilde{\beta}_1 x_i^2$$

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

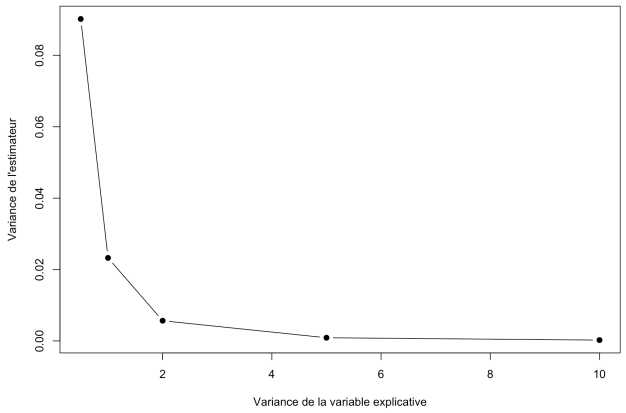


31 question bonus : quel sera votre nombre de bonnes réponses sur les 30 questions précédantes ?

```
library(mnormt)
variance=function(n=50,s3=1,r=.2){
V=rep(NA,200)
for(s in 1:200){
M=matrix(c(1,0,r*s3,0,1,0,r*s3,0,s3^2),3,3)
X=rmnorm(n,mean = rep(0,3),varcov = M)
Y=1+X[,1]/2-X[,2]+X[,3]+rnorm(n)
df=data.frame(Y=Y,X1=X[,1],X2=X[,2],X3=X[,3])
reg=lm(Y~.,data=df)
V[s]=summary(reg)$coefficients[4,2]^2
}
mean(V)}
vn=c(20,50,100,200,500)
vv=Vectorize(function(x) variance(n=x))(vn)
plot(vn,vv,type="b",pch=19,xlab="Taille de l'echantillon (n)",
ylab="Variance de l'estimateur")
```




```
vn=c(.5,1,2,5,10)
vv=Vectorize(function(x) variance(s3=x))(vn)
plot(vn,vv,type="b",pch=19,xlab="Variance de la variable explicative",
ylab="Variance de l'estimateur")
```




```
vn=c(0,.2,.5,.75,.95)
vv=Vectorize(function(x) variance(r=x))(vn)
plot(vn,vv,type="b",pch=19,xlab="Correlation entre variables explicatives",
ylab="Variance de l'estimateur")
```