STT5100 - Automne 2019 - Examen Intra (OLS)

Arthur Charpentier

Examen **B**

Il s'agit du sujet B ! Les sujets A et C contenaient les mêmes questions dans un ordre différent.

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

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

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

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

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

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

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

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

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

Code permanent :

Sujet : B

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

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

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

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

et

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

Calculer le résidu associé à la 4ème observation

A) moins de -1

B) entre -1 et 0

C) entre 0 et 1

D) entre 1 et 2

E) plus de 2

Il s'agit d'un exercice CAS-MAS-I de l'automne 2018. Rappelons (c'est la première question) que $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ est l'estimateur par moindres carrés de $\widehat{\boldsymbol{\beta}}$, et la quatrième observation a pour variables explicatives $\boldsymbol{x}_4$ correspondantn à la troisième ligne de la matrice $\boldsymbol{X}$. Aussi $\widehat{y}_4$ vaut

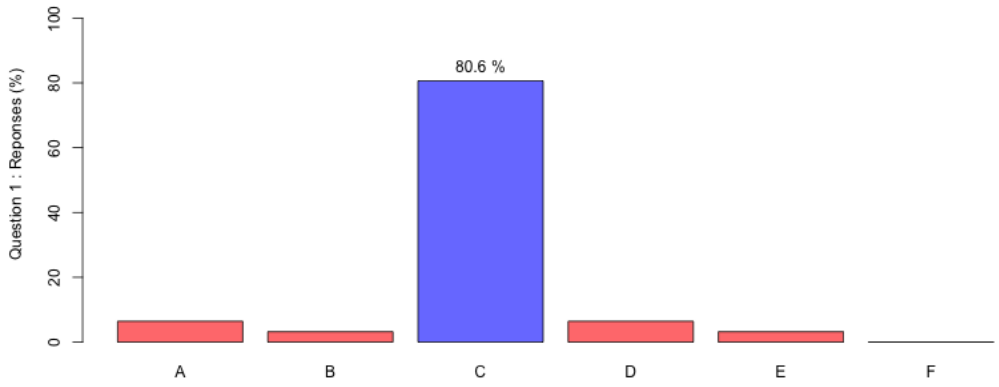
$$\widehat{y}_4 = \boldsymbol{x}_4^\top\widehat{\boldsymbol{\beta}} = \begin{pmatrix} 1 & 1 & 0 & 7 \end{pmatrix} \begin{pmatrix} 2.335 \\ 0.297 \\ -0.196 \\ 1.968 \end{pmatrix} = 2.335 + 0.297 + 7 \times 1.968 = 16.408$$

Or la quatrième observation était $y_4 = 17$, ce qui correspond à un résidu

$$\widehat{\varepsilon}_4 = y_4 - \widehat{y}_4 = 17 - 16.408 = +0.592$$

autrement dit, la bonne réponse est ici C.

Pour toutes les questions, je donne des statistiques de réponses (pour cette question, i.e. 1 pour le sujet B, 27 pour le sujet A et 22 pour le sujet C). La bonne réponse est en bleu, et la colonne F correspond à une non-réponse (pour rappel, dans un examen à choix multiple, dès lors qu'on ne perd pas de point par mauvaise réponse, il n'est pas rationnel de laisser une question sans réponse....) :



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

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

Quelle affimation parmi les cinq suivantes est vraie

A) b_1 est un estimateur non-linéaire de β_1

B) b_1 est un estimateur quadratique de β_1

C) b_1 est un estimateur linéaire biaisé de β_1

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

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

Il s'agit d'un exercice SOA course 4 de l'automne 2001. Un estimateur est dit *linéaire* s'il peut s'écrire comme une combinaison linéaire des y_i. Ici,

$$b_1 = \frac{\sum(z_i - \overline{z})(y_i - \overline{y})}{\sum(z_i - \overline{z})(x_i - \overline{x})} = \sum_{i=1}^{n} \underbrace{\frac{(z_i - \overline{z})}{\sum(z_i - \overline{z})(x_i - \overline{x})}}_{=\omega_i} y_i,$$

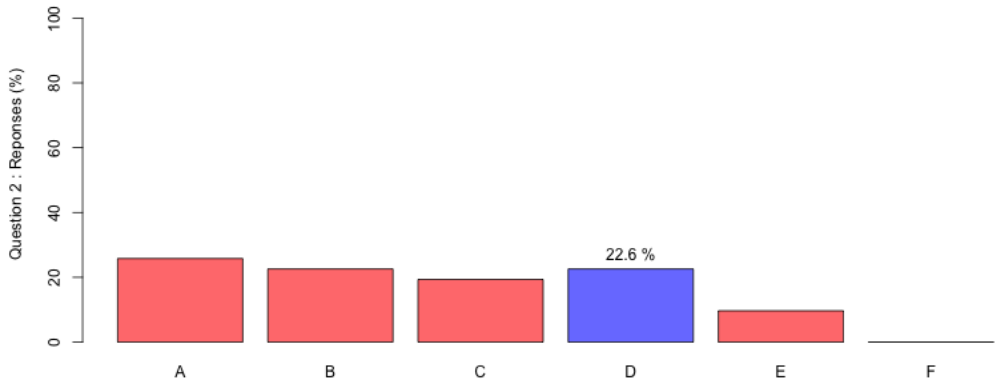
donc b_1 est un estimateur linéaire (donc A est fausse, ainsi que B - pour rappel, 'quadratique' veut dire '*qui est du second degré, élevé au carré*', selon le dictionnaire). De plus,

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

or $\mathbb{E}[y_i] = \beta_0 + \beta_1 x_i$ et $\mathbb{E}[\overline{y}] = \beta_0 + \beta_1 \overline{x}$, de telle sorte que

$$\mathbb{E}[b_1] = \frac{\sum(z_i - \overline{z})(\beta_1[x_i - \overline{x}])}{\sum(z_i - \overline{z})(x_i - \overline{x})} = \beta_1$$

donc b_1 est un estimateur sans biais de β_1 (donc C est fausse). On sait (c'est le théorème de Gauss-Markov) que $\widehat{\boldsymbol{\beta}}$, estimateur par moindre carré est l'estimateur BLUE, le *meilleur* estimateur linéaire sans biais. Or $b_1 \neq \widehat{\beta}_1$, donc b_1 n'est pas BLUE. Donc E est fausse, et D est juste.



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

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

avec de plus

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

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

A) 20

B) 23

C) 25

D) 28

E) 30

Il s'agit d'un exercice SOA course 120, Study Note (120-81-95). Ici, l'estimateur de σ^2 est

$$\widehat{\sigma}^2 = \frac{1}{n-3} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \frac{225}{28-5} = 9.$$

Pour la prévision, le plus petite intervalle de de prédiction à 95% pour y_i est celui centré sur y_i, de la forme

$$\left[y_i \pm t_{25,97.5\%} \sqrt{\widehat{\sigma}^2 [1 + \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_i]} \right]$$

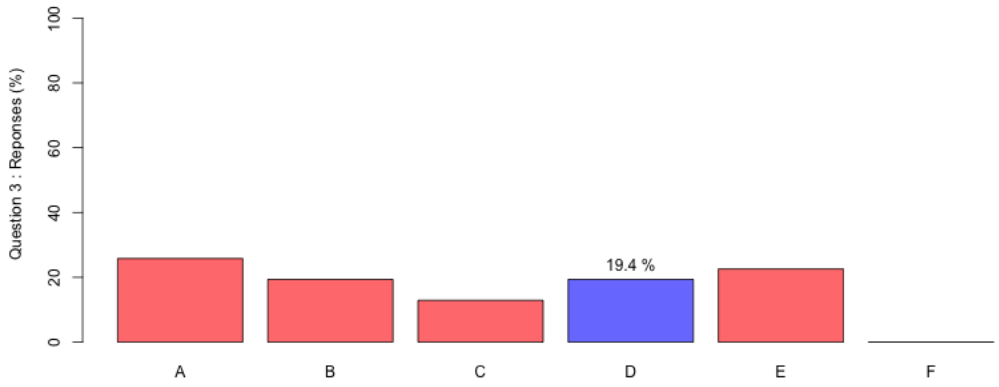
Or on ici considère $\boldsymbol{x}_i^\top = \begin{pmatrix} 1 & 5 & 10 \end{pmatrix}$, de telle sorte que

$$\boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_i = \begin{pmatrix} 1 & 5 & 10 \end{pmatrix} \begin{pmatrix} 0.0100 & 0 & 0 \\ 0 & 0.0200 & 0 \\ 0 & 0 & 0.0349 \end{pmatrix} \begin{pmatrix} 1 \\ 5 \\ 10 \end{pmatrix} = 1^2 \times 0.0100 + 5^2 \times 0.0200 + 10^2 \times 0.0349 = 4,$$

donc l'intervalle de confiance est ici

$$\left[y_i \pm 2.0595 \sqrt{9[1+4]} \right]$$

donc la longueur est $2 \times (2.0595\sqrt{9[1+4]}) = 6 \times 2.0595\sqrt{5} \sim 27.6311$. Si on arrondit, on obtient 28, qui est la réponse D.



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

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

où y est le logarithme du salaire

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

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

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

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

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

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

B) $\beta_1 + \beta_2$

C) $-\beta_1 - \beta_2$

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

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

Il s'agit de la question 20 de l'examen de l'automne 2002 SOA Course 4. On a ici 2 variables explicatives (le niveau d'études et le nombre d'enfants) prenant ici 3 modalités. On a alors 9 prévisions possibles. On peut retrouver cela sur le tableau suivant

éducation	enfants	x_1	x_2	x_3	x_4	$\widehat{y}$
pas secondaire	0	1	0	1	0	$\beta_0+\beta_1+\beta_3$
pas secondaire	$\{1,2\}$	1	0	0	1	$\beta_0+\beta_1+\beta_4$
pas secondaire	3+	1	0	-1	-1	$\beta_0+\beta_1-\beta_3-\beta_4$
secondaire	0	0	1	1	0	$\beta_0+\beta_2+\beta_3$
secondaire	$\{1,2\}$	0	1	0	1	$\beta_0+\beta_2+\beta_4$
secondaire	3+	0	2	-1	-1	$\beta_0+\beta_2-\beta_3-\beta_4$
collégial	0	-1	-1	1	0	$\beta_0-\beta_1-\beta_2+\beta_3$
collégial	$\{1,2\}$	-1	-1	0	1	$\beta_0-\beta_1-\beta_2+\beta_4$
collégial	3+	-1	-1	-1	-1	$\beta_0-\beta_1-\beta_2-\beta_3-\beta_4$

Si on fait la moyenne (équipondérée - on suppose ici les 9 catégories de même taille), on obtient β_0. Dans le premier cas (*pour une femme ayant terminé le collège et ayant 3 enfants (ou plus)*) on prédit (c'est la dernière ligne du tableau)

$$\beta_0-\beta_1-\beta_2-\beta_3-\beta_4$$

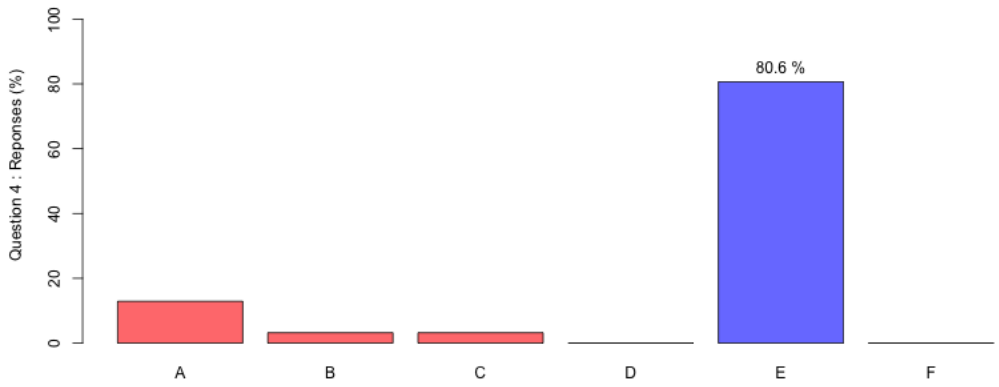
alors que pour le second (*la moyenne de toutes les femmes*) on prédit

$$\beta_0.$$

La différence est alors

$$-\beta_1-\beta_2-\beta_3-\beta_4$$

qui correspond à la réponse E.



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

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

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

On a alors le modèle de régression

$$y_i=\beta_0+\beta_1x_{1,i}+\beta_2x_{2,i}+\beta_3x_{1,i}x_{2,i}+\varepsilon_i.$$

L'estimation par moindres carrés donne

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

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

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

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

On a alors le modèle de régression

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

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

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

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

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

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

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

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 premplace dans le premier modèle

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

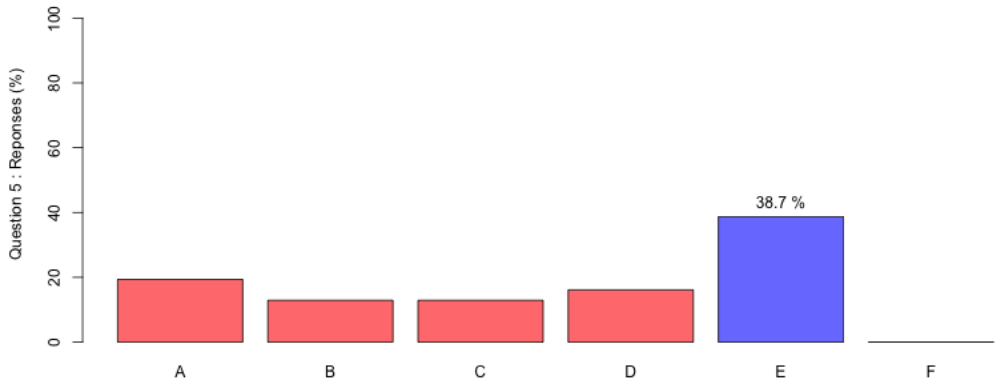
On peut alors identifier simplement les 4 termes,

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

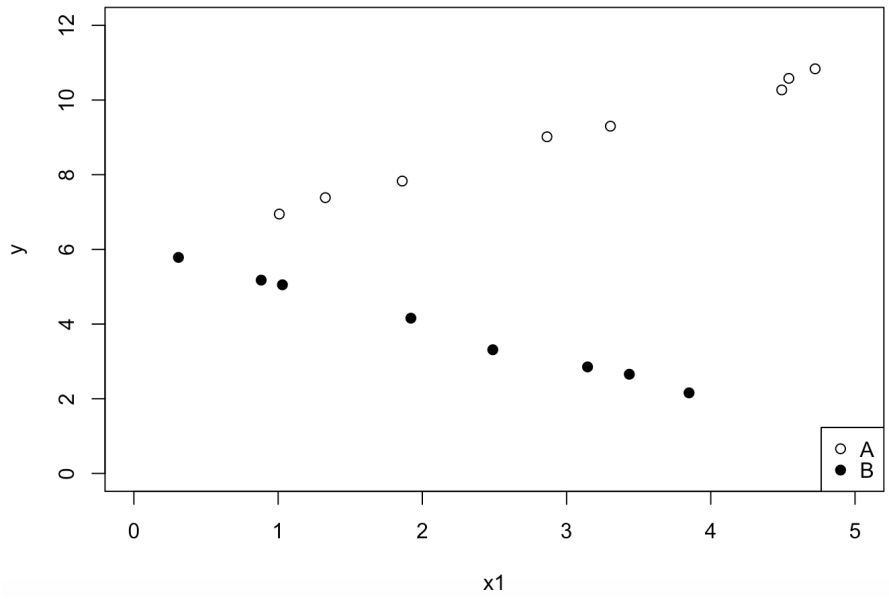
Comparons maintenant nos paires d'estimateurs

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

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.



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



On ajuste le modèle linéaire

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

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

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

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

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

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

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

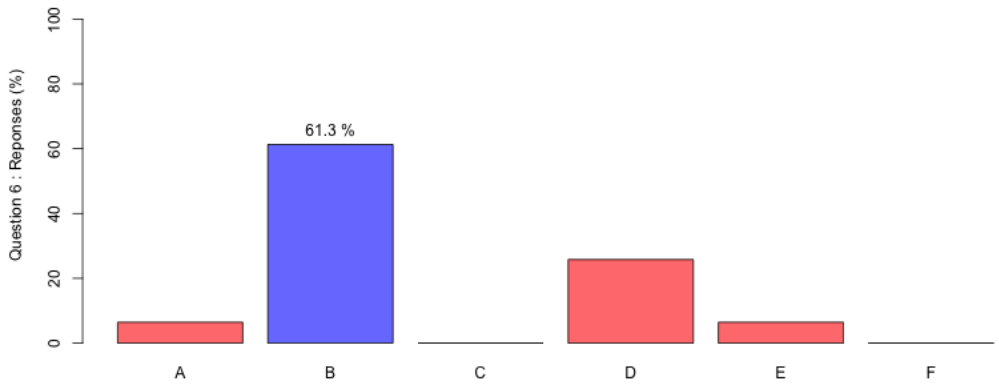
Il s'agit de la question 12 de du SOA Course 120 Study Note 120-82-94. Conditionnellement à x_2, on a les équations suivantes

$$\widehat{y} = \begin{cases} \widehat{\beta}_0 + \widehat{\beta}_1 x_1 & \text{si } x_2 = 0 \text{ (type B)} \\ (\widehat{\beta}_0 + \widehat{\beta}_2) + (\widehat{\beta}_1 + \widehat{\beta}_3) x_1 & \text{si } x_2 = 1 \text{ (type A)} \end{cases}$$

Le graphique suggère une valeur identique pour la constante ($x_1 = 0$), une pente positive pour le cas $x_2 = 1$ et négative si $x_2 = 0$,

$$\begin{cases} \widehat{\beta}_0 = (\widehat{\beta}_0 + \widehat{\beta}_2) \\ \widehat{\beta}_1 < 0 \\ (\widehat{\beta}_1 + \widehat{\beta}_3) > 0 \end{cases}$$

aussi, on a forcément $\widehat{\beta}_3 > 0$ (en plus d'avoir $\widehat{\beta}_1 < 0$), ce qui correspond à la réponse B.



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

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

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

A) $y_i \to \log y_i$

B) $y_i \to \exp y_i$

C) $y_i \to 1/y_i$

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

E) $y_i \to y_i^2$

Il s'agit de l'exercice SOA Course 120 (120-82-97) question 10. On fait ici un développement de Taylor (cf rappels du tout premier cours), en notant g la transformation de la variable y que l'on va tenter de faire

$$g(Y) \sim g(\mu) + g'(\mu) \cdot [Y - \mu], \text{ avec ici } \mu = f(\boldsymbol{x}, \boldsymbol{\beta}).$$

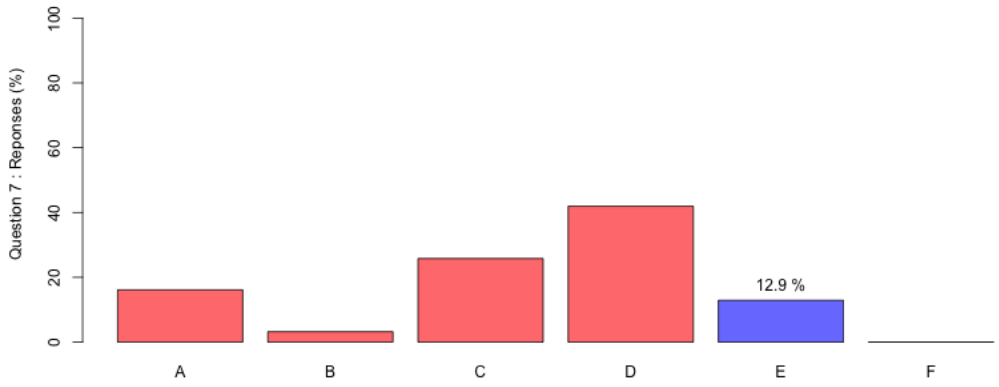
(car que la notation $f(\boldsymbol{x}, \boldsymbol{\beta})$ sert juste à faire peur ici, c'est juste la moyenne). En prenant la variance à gauche et à droite,

$$\text{Var}[g(Y)] \sim \text{Var}[g(\mu) + g'(\mu) \cdot [Y - \mu]] = \text{Var}[g'(\mu) \cdot [Y - \mu]],$$

car le terme (μ) est ici constant. Compte tenu de la formule $\text{Var}[aX] = a^2\text{Var}[X]$,

$$\text{Var}[g(Y)] \sim g'(\mu)^2 \cdot \text{Var}[[Y - \mu]] = g'(\mu)^2 \cdot \frac{\sigma^2}{\mu^2},$$

Pour avoir une variance constante, il faut que $g'(\mu)^2/\mu^2$ soit constant, soit $g'(\mu) \propto \mu$, i.e. $g(\mu) \propto \mu^2$, qui correspond à la réponse E.



Compte tenu de l'échec massif à la question, et que nous n'avions par revu cette technique en cours, je supprime la question... Désolé pour ceux qui y auraient passé du temps.

8 On dispose de la sortie suivante

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

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

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

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

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

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

A) moins de 0.1

B) entre 0.1 et 0.2

C) entre 0.2 et 0.3

D) entre 0.3 et 0.4

E) plus de 0.4

La valeur prédite pour l'observation i est ici $\widehat{y}_i = \boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}}$, soit

$$\widehat{y}_1 = 11.011 + 0.692 \times 51 - 0.104 \times 30 + 0.249 \times 39 - 0.033 \times 61 + 0.015 \times 92 = 52.261$$

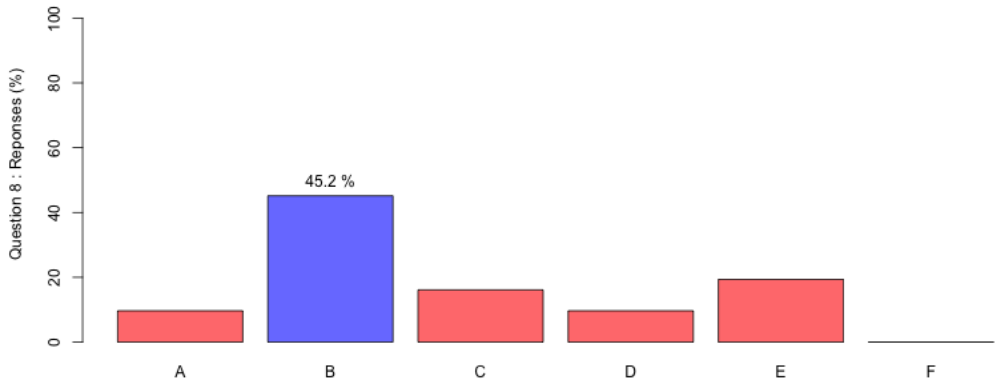
On nous donne ici $\boldsymbol{H}_{1,1} = 0.3234$, qui va nous permettre de calculer le résidu studentisé,

$$e_1 = \frac{y_1 - \widehat{y}_1}{\widehat{\sigma}\sqrt{1 - \boldsymbol{H}_{1,1}}} = \frac{43 - 52.261}{7.139\sqrt{1 - 0.3234}} \sim -1.5771$$

La distance de Cook est

$$d_1 = \frac{e_1^2}{k+1} \frac{\boldsymbol{H}_{1,1}}{1 - \boldsymbol{H}_{1,1}} = \frac{(-1.5771)^2}{5+1} \frac{0.3234}{1 - 0.3234} \sim 0.1981$$

qui est compris entre 0.1 et 0.2, c'est donc la réponse B



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

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

On dispose du modèle suivant

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

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

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

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

A) -0.11

B) 0.53

C) 0.60

D) 0.90

E) 1.70

Il s'agissait de la question 27 de l'examen 4 de la SOA d'automne 2004 (dans le sujet, l'élasticité n'était pas définie, et supposée connue). En $t = 2000$, on est après t_0, donc $\mathbf{1}_{>t_0}(t) = 1$, et donc le modèle s'écrit

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

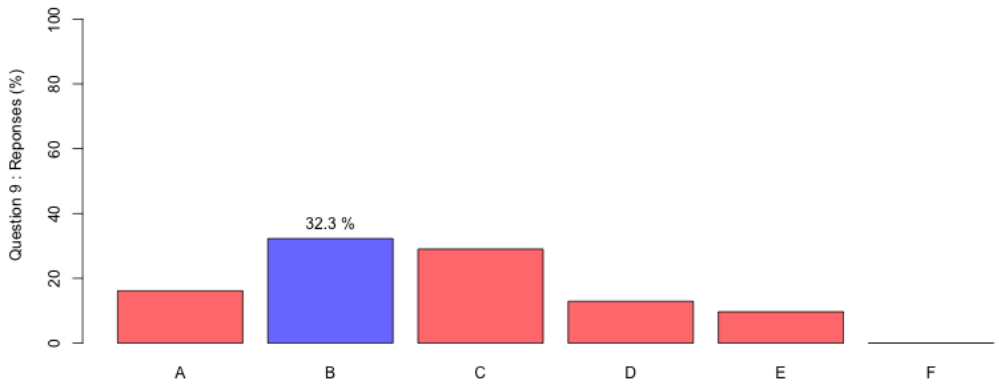
soit

$$\log y_t = \big(\beta_0 - \beta_3 \log x_{1,t_0} - \beta_5 \log x_{2,t_0}\big) + \big(\beta_1 + \beta_3\big)\log x_{1,t} + \big(\beta_2 + \beta_4\big) \log x_{2,t} + \varepsilon_t$$

Si on reprend la définition, on cherche

$$e_{y|x_1} = \frac{\partial \log y_i}{\partial \log x_{1,i}} = \beta_1 + \beta_3 = 0.60 - 0.07 = 0.53$$

ce qui correspond à la réponse B.



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

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

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

A) 4.5 km

B) 6.5 km

C) 8.0 km

D) 28 km

E) 29 km

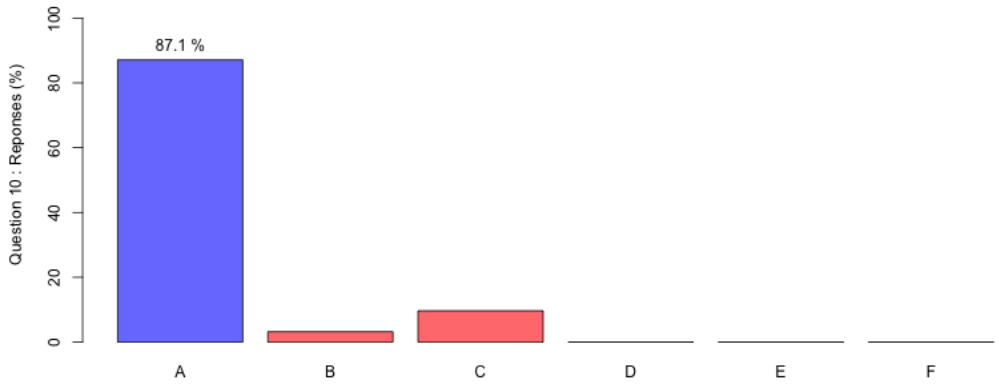
Question 7 de SOA Course 120 Study Notes (210-83-98). Comme le feu se déclare à plus de 4 kilomètre, $x_1 \geqslant 4$, et $x_2 = 1$. Dans la ville A, $x_3 = 1$, de telle sorte que

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

alors que dans la ville B, $x_3 = 0$,

$$\widehat{y}_i = 8 + 5x_{i,1} + 2(x_{i,1} - 4) = 7x_{1,i}$$

Les deux coûts moyens sont alors identiques si $9 + 5x_{1,i} = 7x_{1,i}$, soit $x_{1,i} = 4.5$ km, ce qui est correspond à la réponse A.



11 On considère un modèle avec deux variables explicatives,

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

On dispose de

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

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

A) 1.92

B) 2.23

C) 2.45

D) 2.87

E) 3.11

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

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

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

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

dont un estimateur naturel est

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

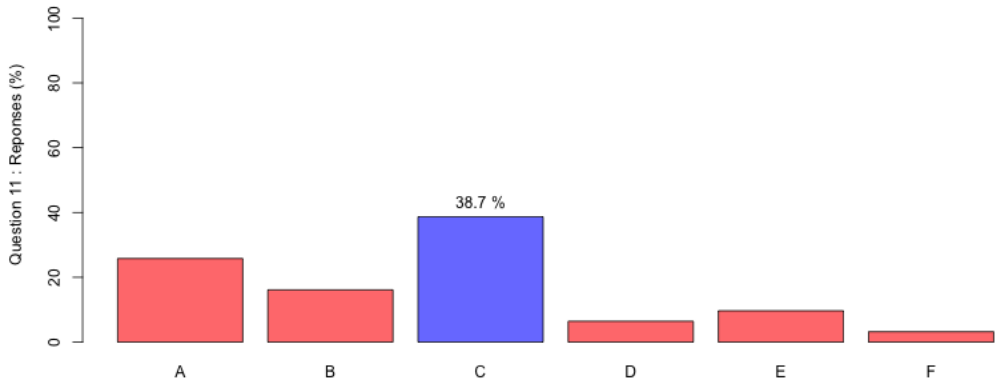
soit, numériquement

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

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

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

dont la valeur la plus proche est 2.45.



12 On considère un modèle avec trois variables explicatives,

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

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

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

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

A) la constante

B) la constante et x_1

C) la constante et x_2

D) la constante, x_1 et x_3

E) la constante, x_2 et x_3

Il s'agissait de la question 21 de l'examen CAS ST de l'automne 2015. On va commencer par un test de Student, pour chacune des variables (y compris la constante),

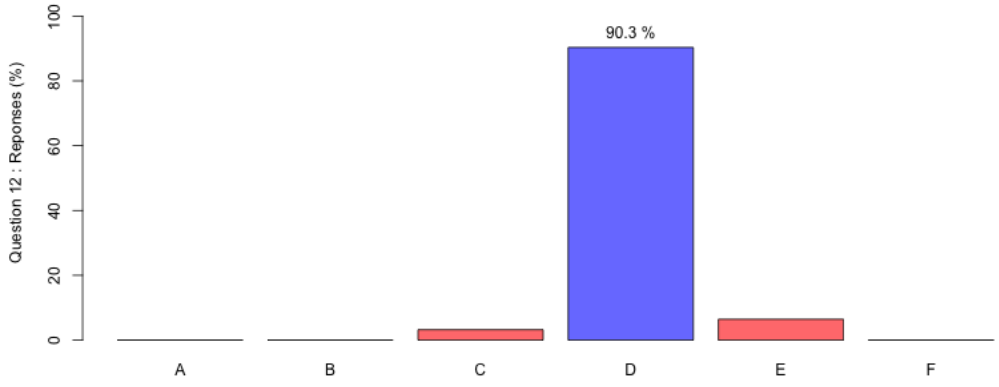
$$t_0 = \frac{\widehat{\beta}_0}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_0]}} = \frac{44.2}{5.96} \sim 7.4161$$

$$t_1 = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_1]}} = \frac{-0.295}{0.118} \sim -2.5$$

$$t_2 = \frac{\widehat{\beta}_2}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_2]}} = \frac{9.11}{6.86} \sim 1.328$$

$$t_3 = \frac{\widehat{\beta}_3}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_3]}} = \frac{-8.7}{1.2} \sim -7.25$$

Ces statistiques permettent de tester $H_0 : \beta_k = 0$, et sous hypothèse, les statistiques de test doivent suivre une loi de Student à 28 degrés de liberté. Il y a 90% de chances qu'une telle loi soit comprise dans l'intervalle $[\pm 1.7]$ en utilisant la table donnée sur la page de couverture (et en arrondissant 28 à 30). Seule t_2 est dans l'intervalle $[\pm 1.7]$, autrement dit, toutes les variables sont statistiquement significatives, sauf x_2. La réponse D est la bonne.



13 On considère un modèle avec cinq variables explicatives,

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

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

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

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

A) 1 variable n'est pas significative

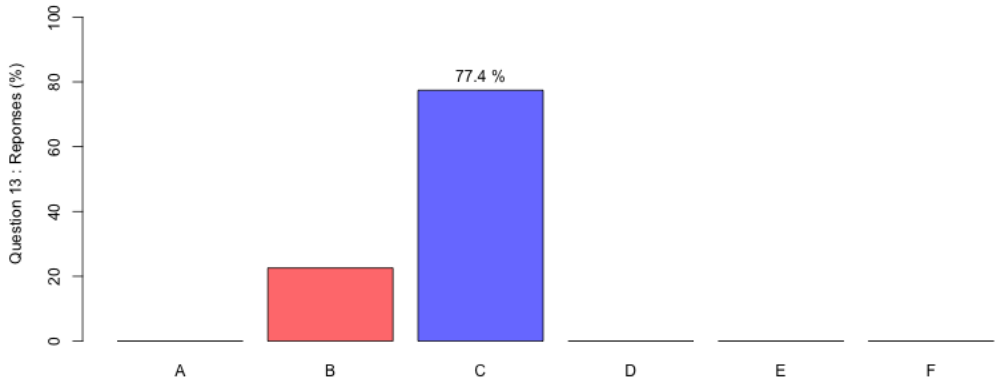
B) 2 variables ne sont pas significatives

C) 3 variables ne sont pas significatives

D) 4 variables ne sont pas significatives

E) les 5 variables ne sont pas significatives

Question 20 de l'examen CAS ST du printemp 2016. C'est encore plus facile que la question d'avant puisqu'ici les statistiques de Student nous sont données. Ici, sous H_0 (pour rappel, $H_0 : \beta_k = 0$), la statistique T doit suivre une loi de Student à $n - k = 20 - 6 = 14$ degrés de liberté. Pour simplifier, on utilisera 15 dans la table. Le quantile à 95% est ici 1.75, donc à avec une probabilité de 90%, la statistique de test est dans l'intervalle ± 1.75 si H_0 est vraie. Ici, seules x_3 et x_5 sont significatives, et 3 ne le sont pas (x_1, x_2 et x_4). Il fallait retenir la réponse C. Maintenant, si on esssaye de faire les choses proprement, l'affirmation C correspond à un test multiple, et on n'a pas assez d'éléments pour répondre... Mais bon, on utilise les informations à notre disposition, et la réponse *on ne peut pas répondre* n'était pas proposée ! Donc il fallait répondre C (ou laisser une note sur la copie indiquant qu'on n'avait pas assez d'éléments pour répondre, j'aurais accepté !).



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

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

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

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

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

et

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

Il s'agissait d'un exercice du guide ACTEX de préparation à l'examen SRM.

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

A) aucune

B) la constante

C) la taille de la mère

D) la taille du père

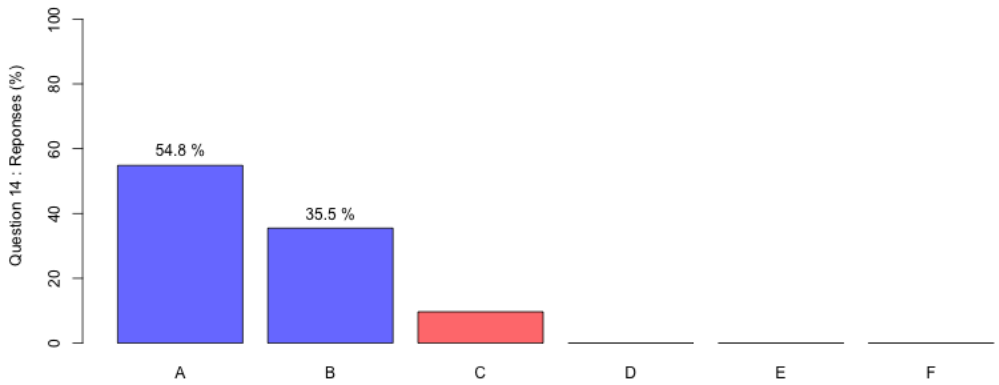
E) toutes

On va utiliser ici la statistique de Student, qu'il faut calculer pour β_0 et β_2,

$$t_0 = \frac{\widehat{\beta}_0}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_0]}} = \frac{17.26}{8.994} \sim 1.9191 \text{ et } t_2 = \frac{\widehat{\beta}_2}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_2]}} = \frac{0.25485}{0.0989} \sim 2.5768$$

Pour un seuil de significativité à 5%, on utilise le quantile de niveau 97.5%, ici avec 20-2=18 degrés de liberté, soit environ 20. Autrement dit la région d'acceptation du test de l'hypothèse $H_0 : \beta_k = 0$ est $[\pm 2.086]$ (la vraie valeur est à 2.1). Seule la constante n'est pas significative, ici. C'est la réponse B, et c'était la réponse attendue dans l'examen.

En regardant les réponses, je vois que beaucoup on répondu A, et effectivement, je l'ai dit en cours, on n'enlève pas la constante dans un modèle de régression... donc je retiens aussi la réponse A.



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

A) environ 0.5%

B) environ et 1%

C) environ 2%

D) environ 5%

E) environ 10%

L'hypothèse H_0 est classique (test simple) par contre, l'hypothèse alternative est peu usuelle, avec ici un test unilatéral. La p-value est alors définie par

$$p = \mathbb{P}[T > t_{\text{observé}}] \text{ où } T \sim \text{Student à 18 degrés de liberté}$$

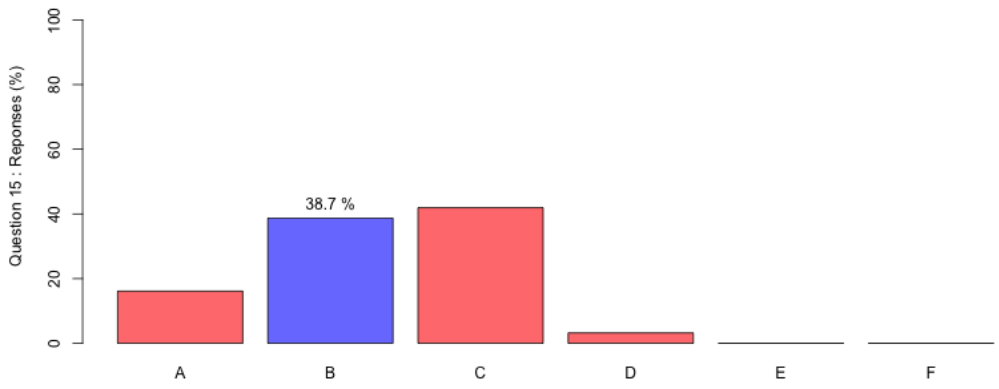
autrement dit, on cherche

$$p = \mathbb{P}[T > 2.5768]$$

Or si on arrondit 18 à 15 ou 20, dans la table on nous dit que

$$\mathbb{P}[T_{20} > 2.528] = .99\% \text{ alors que } \mathbb{P}[T_{15} > 2.602] = .99\%$$

donc la p-value est de l'ordre de 1%, qui est la réponse B. En fait, la vraie valeur est 0.9501%, qui vaut presque 1% effectivement.



Petite parenthèse (cf cours de statistique): avec un test plus classique - i.e. bilatéral, avec comme hypothèse alternative $H_1 : \beta_1 \neq 0$ - la p-value est

$$p = \mathbb{P}[|T| > |t_{\text{observé}}|] \text{ où } T \sim \text{Student à 18 degrés de liberté}$$

autrement dit

$$p = \mathbb{P}\big[\{T > |t_{\text{observé}}|\} \cup \{T < |-t_{\text{observé}}|\}\big] = \mathbb{P}[T > |t_{\text{observé}}|] + \mathbb{P}[T < -|t_{\text{observé}}|] = 2\mathbb{P}[T > |t_{\text{observé}}|]$$

par symmétrie de la loi de Student (par rapport à 0), ce qui correspond au double de la valeur que l'on a calculée ici, avec le test unilatéral (car $\widehat{\beta}_2 > 0$ et l'hypothèse alternative est dans le même sens, $H_1 : \beta_2 > 0$... là encore, je renvoie au cours de statistique). J'imagine que ça explique la forte proportion de C...

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

A) 2.5

B) 3.2

C) 3.8

D) 5.1

E) 7.6

Pour rappel, on nous donne la table d'ANOVA suivante

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

On va chercher à la compléter. Tout d'abord, on note qu'on a 18 degrés de liberté, et 2 variables. Ça nous donne la première colonne. Ensuite, la somme des carrés des résidus pour la première ligne est la part expliqué, soit 1157.25 -813.26 =343.99.

```
            Df   Sum Sq Mean Sq F value   Pr(>F)
Regression   2   343.99
Residuals   18   813.26
Total       20  1157.25
```

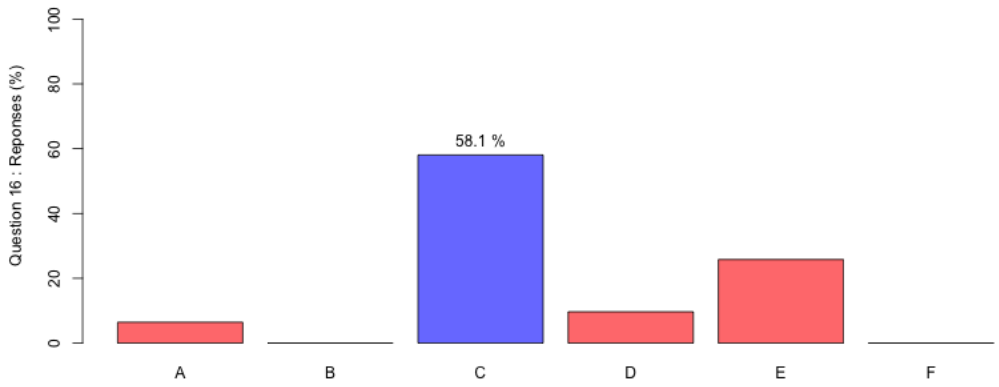
La statistique de Fisher est

$$F = \frac{\texttt{Sum Sq de Regression}}{\texttt{Df de Regression}} \cdot \frac{\texttt{Df de Residuals}}{\texttt{Sum Sq de Residuals}}$$

soit

$$F = \frac{343.99}{2} \cdot \frac{18}{813.26} = \frac{171.995}{45.1811} \sim 3.81$$

qui est la réponse C.



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

A) 70

B) 71

C) 72

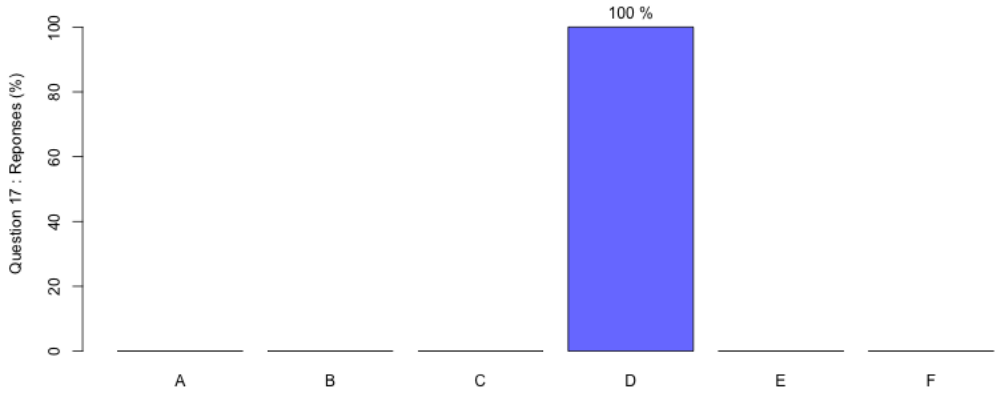
D) 73

E) 74

Il suffit de remplacer :

$$\widehat{y} = 17.26 + 0.5283 \times 69 + 0.25485 \times 76 \sim 73.0813$$

ce qui correspond à la réponse D.



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

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

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

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

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

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

Regardons le modèle E :

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

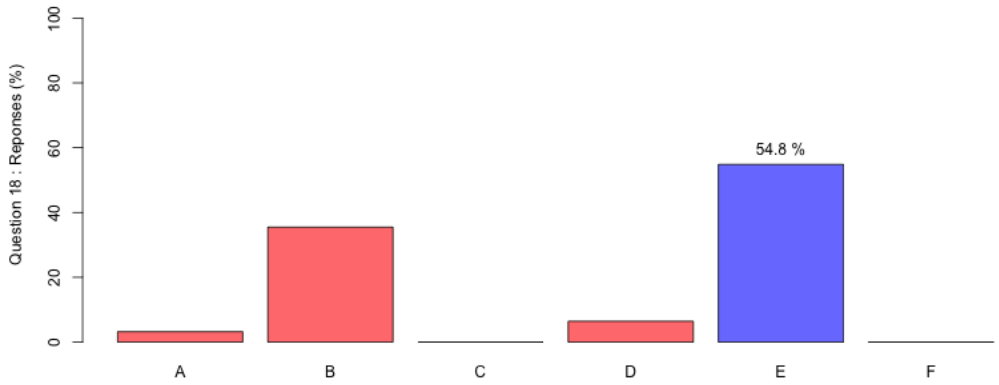
qui peut se réécrire

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon & \text{si } x_3 = 0 \text{ (femme)} \\ \beta_0 + (\beta_1 + \beta_3) x_1 + (\beta_2 + \beta_4) x_2 + \varepsilon & \text{si } x_3 = 1 \text{ (homme)} \end{cases}$$

Ce modèle permet d'avoir des effets (totalement) différents pour x_1 et x_2 en fonction du genre, et c'est le seul. Par exemple le modèle D peut se réécrire

$$y = \begin{cases} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon & \text{si } x_3 = 0 \text{ (femme)} \\ \beta_0 + \beta_1 + x_1 + (\beta_2 + \beta_3) x_2 + \varepsilon & \text{si } x_3 = 1 \text{ (homme)} \end{cases}$$

autrement dit, pour ce modèle l'effet de x_1 est identique. Pour tester si l'effet est identique, ou pas, on utiliser un test (multiple) de Fisher de l'hypothèse $H_0 : \beta_3 = \beta_4 = 0$. Il fallait répondre E.



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

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

et

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

On nous donne

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

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

A) moins de 22

B) entre 22 et 25

C) entre 25 et 27

D) entre 27 et 30

E) plus de 30

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

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

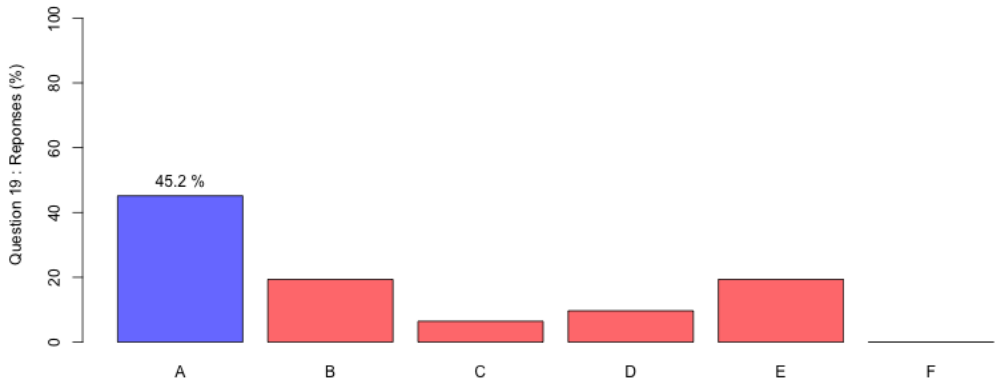
Pour le modèle (B), pour calculer la somme des carrés expliqués, on utilise le R^2, en rappelant que

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

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.



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

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

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

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

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

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

Il s'agissait de la quetion 28 de l'examen 4 de la SOA de l'automne 2001. Pour rappel, dans un modèle hétérosédastique, avec $\text{Var}[\varepsilon_i|x_i] = \sigma^2\omega_i$, on transforme le modèle en *divisant* par $\sqrt{\omega_i}$, de telle sorte que

$$\omega_i^{-1/2} y_i = \omega_i^{-1/2}\beta_0 + \omega_i^{-1/2}\beta_1 x_i + \underbrace{\omega_i^{-1/2}\varepsilon_i}_{\eta_i}$$

qui vérifie

$$\text{Var}[\eta_i|x_i] = \text{Var}[\omega_i^{-1/2}\varepsilon_i|x_i] = \omega_i^{-1}\text{Var}[\varepsilon_i|x_i] = \omega_i^{-1}\sigma^2\omega_i = \sigma^2$$

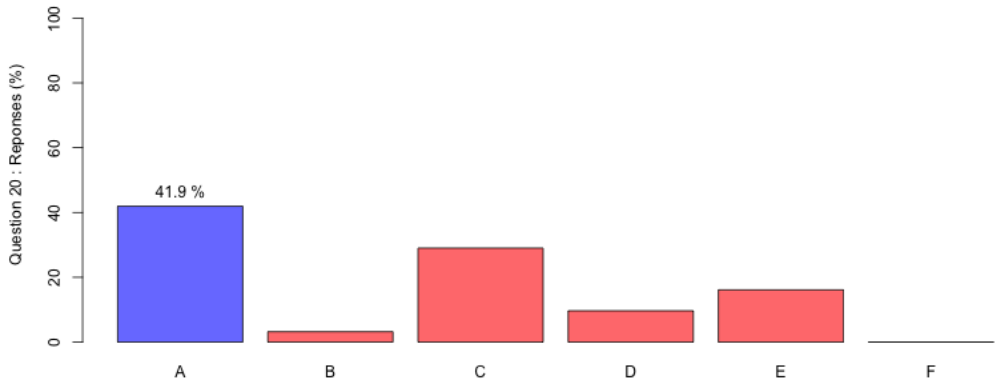
autrement dit, le modèle est maintenant homoscédastique (i.e. on a réussit à corriger de l'hétéroscédasticité). Ici $\text{Var}[\varepsilon|x] = \sigma^2 x^{-1/2}$ donc $\omega = x^{-1/2}$, ou $\omega^{-1/2} = x^{1/4}$, et le modèle corrigé devient alors

$$x^{1/4} y = \beta_0 x^{1/4} + \beta_1 x \cdot x^{1/4} + \underbrace{x^{1/4}\varepsilon_i}_{\eta}$$

soit

$$x^{1/4}y = \beta_0 x^{1/4} + \beta_1 x^{5/4} + \eta$$

qui correspond à la proposition A.



21 On considère le modèle suivant,

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

Donnez les estimateurs par moindres carrés de β_1 et β_0

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

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

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

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

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

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

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

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

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

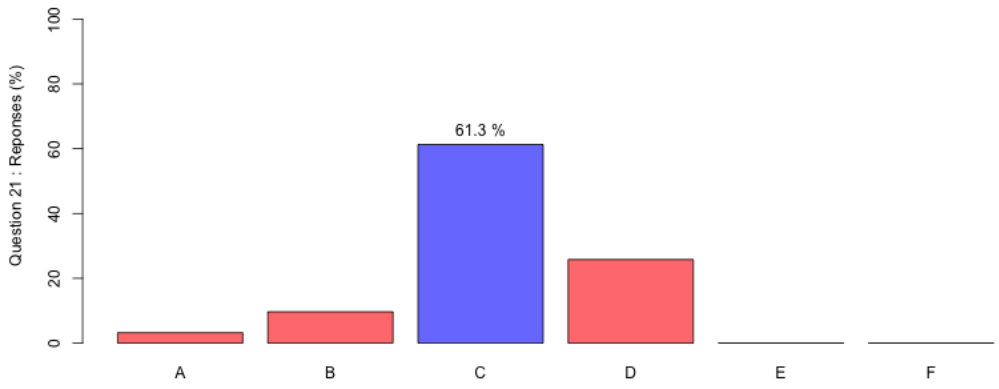
On retrouve cette expression dans C et E. Pour la constante, on a comme toujours la condition du premier ordre qui garantit que

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

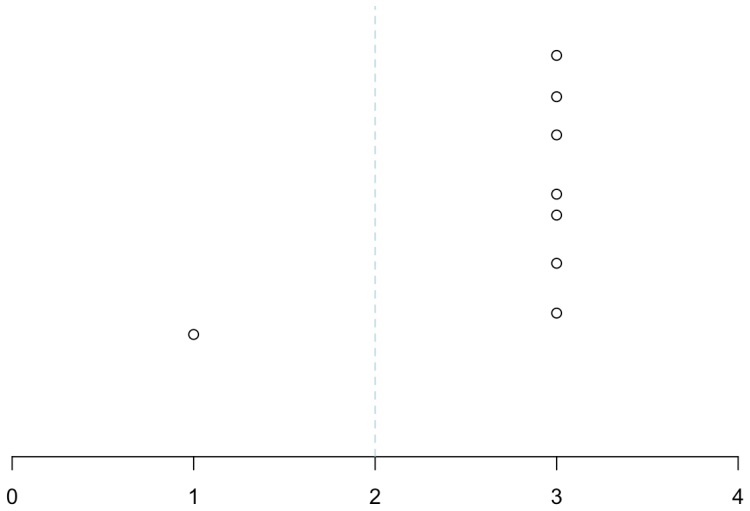
donc

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

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



22 On dispose du jeu de données suivant



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

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

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

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

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

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

Il s'agissait de la question 7 de l'examen 120 de la SOA de mai 1990. On en avait parlé en cours : la droite de régression passe ici par $(1, y_1)$ et par $(3, \overline{y}_3)$ où $\overline{y}_3$ est la moyenne des points localisés à droite,

$$\overline{y}_3 = \frac{1}{7}\sum_{i=2}^{8} y_i = \frac{1}{7}\left(\sum_{i=1}^{8} y_i - y_1\right) = \frac{8}{7}\underbrace{\frac{1}{8}\sum_{i=1}^{8} y_i}_{\overline{y}} - \frac{1}{7}y_1$$

On cherche donc une droite $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$ qui vérifier

$$\begin{cases} \widehat{\beta}_0 + \widehat{\beta}_1 1 = y_1 & (A) \\ \widehat{\beta}_0 + \widehat{\beta}_1 3 = \frac{8}{7}\overline{y} - \frac{1}{7}y_1 & (B) \end{cases}$$

soit, en faisant $(B) - (A)$ et $3 \times (A) - (B)$

$$2\widehat{\beta}_1 = \frac{8}{7}\overline{y} - \frac{1}{7}y_1 - y_1 \text{ et } 2\widehat{\beta}_0 = 3y_1 - \left(\frac{8}{7}\overline{y} - \frac{1}{7}y_1\right)$$

i.e.

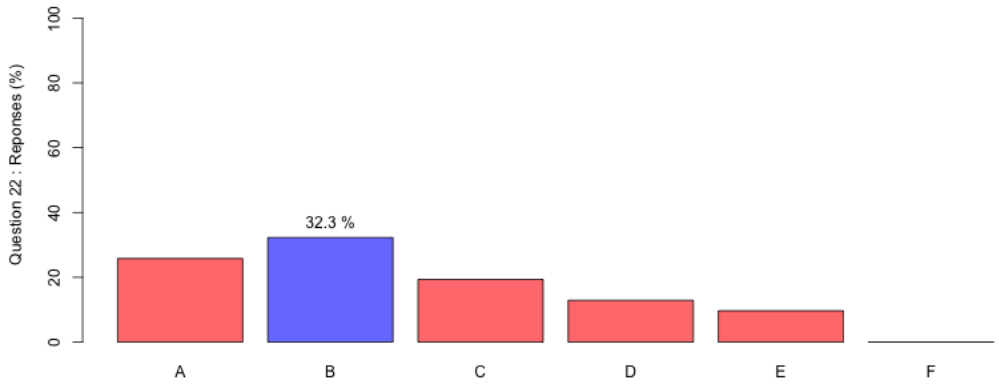
$$\widehat{\beta}_1 = \frac{4}{7}\overline{y} - \frac{4}{7}y_1 - \text{ et } \widehat{\beta}_0 = \frac{11}{7}y_1 - \frac{4}{7}\overline{y}$$

Pour obtenir la prévision en $x = 2$, on utilise

$$\widehat{\beta}_0 + 2\widehat{\beta}_1 = \frac{11}{7}y_1 - \frac{4}{7}\overline{y} + 2\left(\frac{4}{7}\overline{y} - \frac{4}{7}y_1\right) = \frac{11-8}{7}y_1 + \frac{-4+8}{7}\overline{y} = \frac{3}{7}y_1 + \frac{4}{7}\overline{y}$$

qui correspond à la réponse B. Cela dit, il y avait plus simple : 2 étant au milieu de 1 et 3, la prédiction en 2 soit au milieu de la prédiction en 1 et en 3, soit y_1 et $\overline{y}_3$, i.e.

$$\frac{1}{2}y_1 + \frac{1}{2}\overline{y}_3 = \frac{1}{2}y_1 + \frac{1}{2}\left(\frac{8}{7}\overline{y} - \frac{1}{7}y_1\right) = \frac{3}{7}y_1 + \frac{4}{7}\overline{y}$$



23 On considère le modèle suivant,

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

Donnez l'estimateur par moindres carrés de β

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

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

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

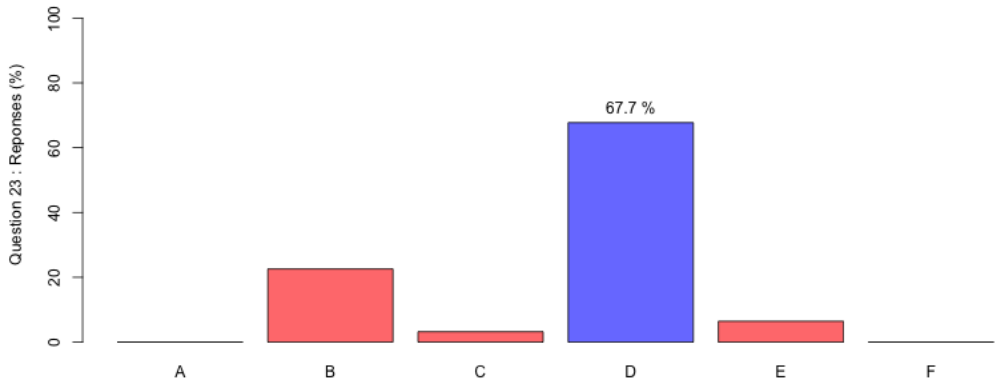
soit

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

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.



Le problème suivant sert de base aux questions 24 et 25

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

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

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

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

Il s'agissait du problème 2.5.45 du *Study Manuel* de l'examen SRM d'Actex de 2019.

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

A) 2

B) 4

C) 5

D) 6

E) 8

Le modèle (1) nous permet d'avoir la somme des carrés totaux, $\sum_{i=1}^{n}(y_i - \overline{y})^2 = 423.48 = SCR_1$. On nous donne aussi le R^2 du modèle (0), et donc on peut calcuer SCR_0 puisque

$$R^2 = 1 - \frac{SCR_0}{SCR_1} \text{ soit } SCR_0 = (1 - R^2) SCR_1 = 423.48 \times (1 - 0.87) = 55.0654.$$

Si on réécrit le modèle (0) conditionnellement à x_2,

$$y_i = \begin{cases} \beta_0 + \beta_1 x_{1,i} + \varepsilon_i & \text{si } x_2 = 0 \text{ , i.e une femme} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{1,i} + \varepsilon_i & \text{si } x_2 = 1 \text{ , i.e un hommme} \end{cases}$$

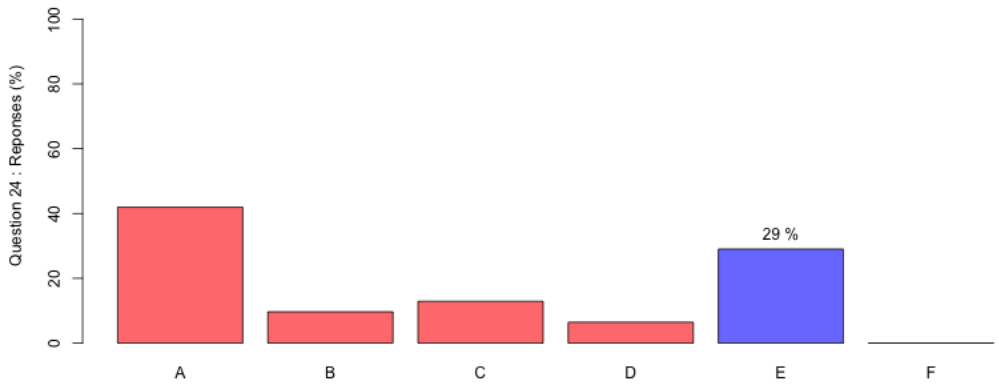
Tester "*si l'impact de l'expérience sur le salaire est identique pour les hommes et les femmes*" revient à se demander si $\beta_3 = 0$. Classiquement on pourrait faire un test de Student (c'est un test simple) mais on nous demande ici d'utiliser un test de Fisher. Le modèle non-contraint est (bien 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.



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

A) 2

B) 4

C) 5

D) 6

E) 8

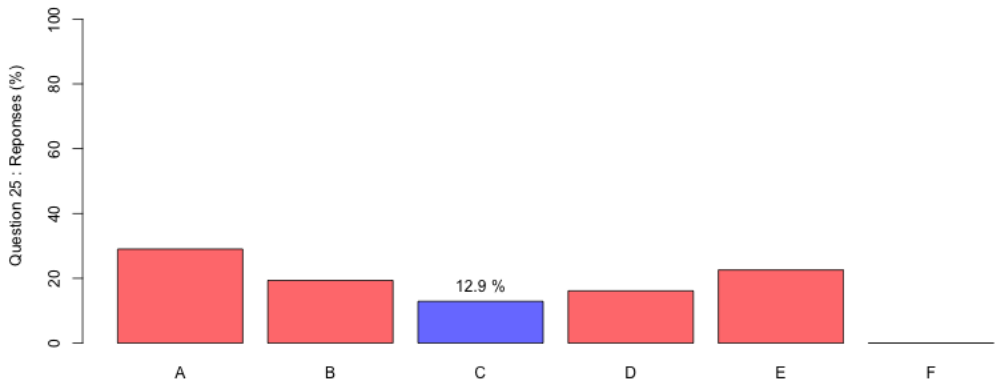
On reprend l'écriture de la question précédante,

$$y_i = \begin{cases} \beta_0 + \beta_1 x_{1,i} + \varepsilon_i & \text{si } x_2 = 0 \text{ , i.e une femme} \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{1,i} + \varepsilon_i & \text{si } x_2 = 1 \text{ , i.e un hommme} \end{cases}$$

Ici, tester "*si la relation linéaire reliant l'expérience et le salaire est identique pour les hommes et les femmes*" revient à se demander si $\beta_3 = 0$ **et** $\beta_2 = 0$. Cette fois, on se doit d'utiliser un test de Fisher ! Le modèle contraint est cette fois le modèle (2), et la statistique de test est alors

$$F = \frac{SCR_2 - SCR_0}{2} \cdot \frac{30-4}{SCR_0} = \frac{75.69 - 55.06}{2} \cdot \frac{26}{55.06} = 4.8869$$

2 car cette fois on teste ici deux valeurs (β_3 et β_2) et $30-4$ est inchangé car on a ici 30 observations, et 4 variables dans le modèle (0). Cette fois, on est plus proche de 5, donc on va retenir la réponse C.



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

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

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

il y avait une petite typo dans le tableau, mais par élimination ou peu de réflexion, on pouvait retrouver la bonne valeur...

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

A) aucune variable explicative (juste la constante)

B) la constante et x_2 (seulement)

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

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

E) la constante et les quatre variables explicatives

Il s'agissait du probème 4.5.26 des Study Notes pour l'examen SRM. Dans une approche *backward*, on la procédure sera la suivante

1. on estime le modèle complet, avec x_1, x_2, x_3 et x_4, la SCR vaut 0.12089
2. on estime 4 modèles en enlevant une seule variable parmi x_1, x_2, x_3 et x_4. Celle que l'on enlève est celle qui a la plus petite SCR. Il s'agit du modèle avec {constante, x_1, x_2, et x_3} et on a alors une SCR de 0.12133
3. on estime 3 modèles en enlevant une seule variable parmi x_1, x_2 et x_3 (puisqu'on avait retiré x_4 à l'étape précédante). Le meilleur modèle est {constante, x_1 et x_2} et on a alors une SCR de 0.14216
4. on estime 2 modèles en enlevant une seule variable parmi x_1 et x_2 (puisqu'on avait retiré x_3 à l'étape précédante). Le meilleur modèle est {constante et x_2} et on a alors une SCR de 0.64834
5. on estime 1 modèle, sans variable puisqu'il ne restait plus que x_2 à l'étape précédante, et on doit enlever la constante. Le modèle est {constante} et on a alors une SCR de 1.70167

La procédure précédante avait du sens car comparer des modèles ayant le même nombre de variables explicatives revient à comparer les SCR. Aussi, *à l'intérieur d'une étape*, notre stratégie avait du sens. Mais on n'a pas fini parce qu'on nous demande d'utiliser le C_p de Mallow pour comparer des modèles n'ayant pas le même nombre de variables explciatives. Pour rappel,

$$C_p = \frac{1}{n}\left(\text{SCR} + 2p\hat{\sigma}^2\right),$$

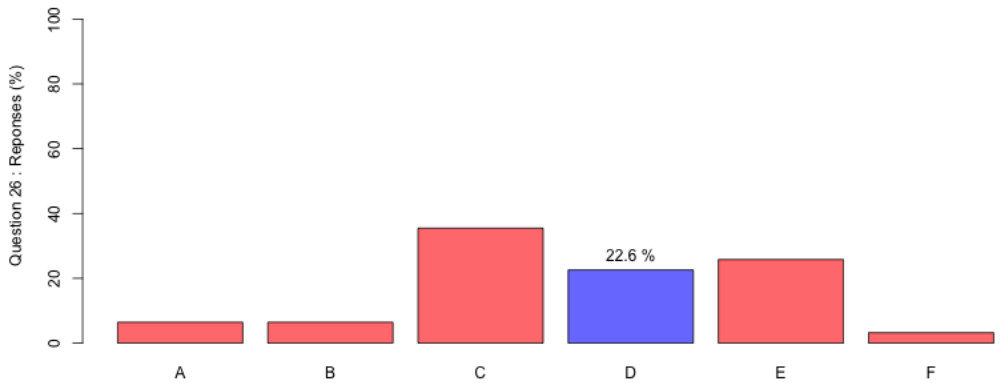
et

$$\hat{\sigma}^2 = \frac{1}{n-5}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{\text{SCR}_5}{25-5} = \frac{0.12089}{25-5} = 0.00604$$

(oui, on utilise le modèle complet pour estimer la variance des erreurs). Si on regarde les 5 meilleurs modèles obtenus, on a alors

p	modèle	SCR	C_p
0	constante	1.70167	0.068550
1	constante et x_2	0.64834	0.026901
2	constante, x_1 et x_2	0.14216	0.007137
3	constante, x_1, x_2 et x_3	0.12133	0.006787
4	constante, x_1, x_2, x_3 et x_4	0.12089	0.007253

Notons ici que si on se contente de regarder la SCR, elle diminue avec p (on s'y attendait, on l'a vu en cours), donc ce n'est pas un critère pour comparer des modèles avec des nombres de variables explicatives différents. Si on retient le *meilleur* C_p, i.e. le plus petit, on va retenir le modèle avec 3 variables explicatives, {constante, x_1, x_2 et x_3}, ce qui correspond à la réponse D.



Le problème suivant sert de base aux questions 27 et 28

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

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

Il s'agissait de l'exercice 8 du Study Note 120-81-95 de l'examen 120 de la SOA.

27 On veut construire deux modèles :

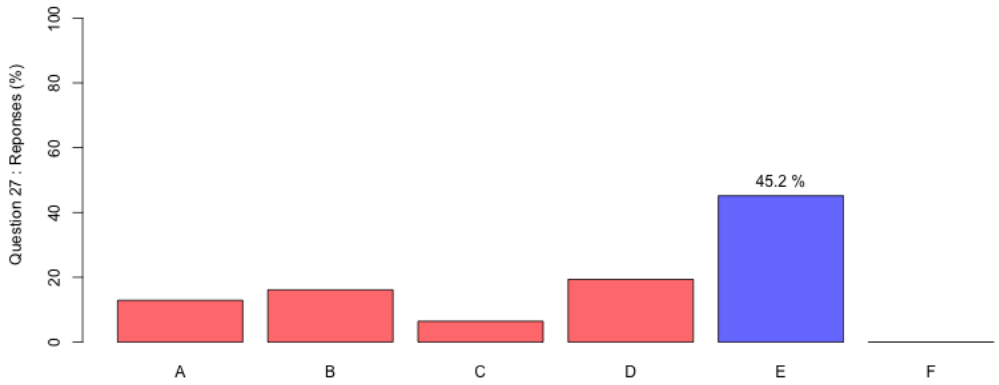
- un modèle avec 2 variables explicatives avec une sélection descendante (*backward*)
- un modèle avec 1 variable explicative avec une sélection ascendante (*forward*)

Quels sont les modèles retenus ?

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

Commençons par l'approche *forward*: on part du modèle avec seulement la constante, et on veut rajouter une variable, et une seule. On va retenir la variable avec la plus statistique de Student, en valeur absolue. Ici, c'est $t = 3$ obtenue avec la variable x_1. Le meilleur modèle avec une variable explicative est ici x_1.

Regardons maintenant l'approche *backward*: on part du modèle le plus complet possible (avec la constante et les trois variables), et on enlève cette avec la plus faible statistique de Student, en valeur absolue. Ici, c'est $t = 2$ obtenue avec la variable x_1. Le meilleur modèle avec deux variables explicatives est ici construit sur x_2 et x_3. On retient donc la réponse E.



28 On veut construire deux modèles :

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

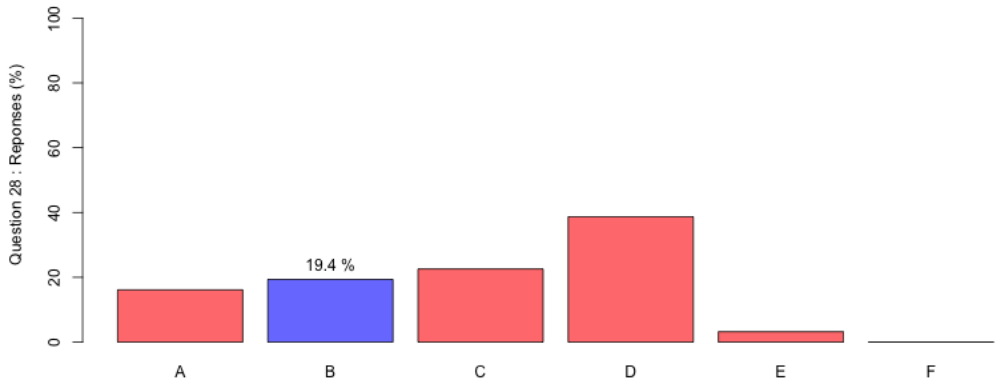
Quels sont les modèles retenus ?

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

On est très proche de l'exercice précédant, sauf que dans les deux cas, on va maintenant un cran plus loin !

Commençons par l'approche *forward*: Le meilleur modèle avec une variable explicative était ici x_1 (cf question précédante). Mais à l'étape suivant, on alors le choix : soit on retient la paire (x_1, x_2), soit la paire (x_1, x_3). En mettant x_2 on a une statistique de Student $|T| = 2.51$ alors qu'en mettant x_3 on a une statistique de Student $|T| = 2.41$. Comme $2.51 > 2.41$, on préfèrera rajouter x_2. Le meilleur modèle avec cette procédure itérative est alors x_1 et x_2.

Regardons maintenant l'approche *backward*: Le meilleur modèle avec deux variables explicatives était ici construit sur x_2 et x_3. On va chercher à en enlever une. Or dans ce modèle, les statistiques de test sont respectivement -2.70 et 2.51. Comme $|-2.70| > |2.51|$ on va préférer supprimer la seconde, c'est à dire x_3. Le modèle que l'on garde est alors construit uniquement sur x_2. Aussi, ici, on retient la réponse B.



Le problème suivant sert de base aux questions 29 et 30

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

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

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

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

Il s'agissait d'une version simplifiée du problème 4.5.28 des Study Notes pour l'examen SRM.

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

A) moins de -2

B) entre -1.5 et -1

C) entre -1 et -0.5

D) entre -0.5 et 0

E) plus de 0

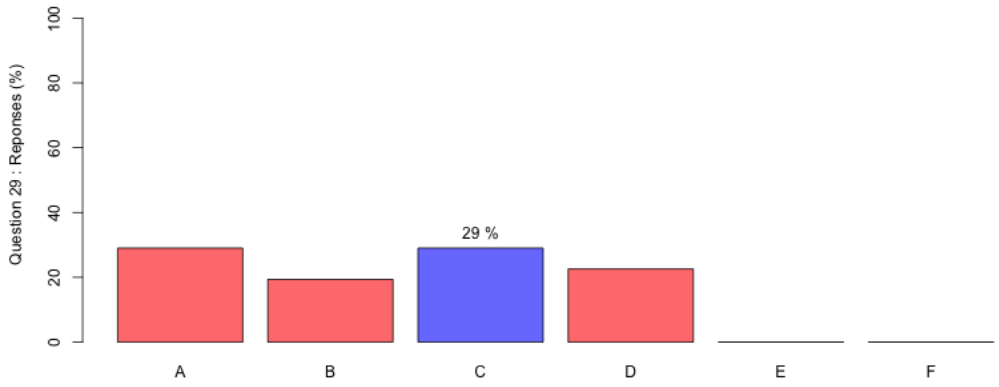
On va commencer par remplir la sortie de l'ANOVA (oui, j'ai longuement dit en cours que faire une régression avec 2 variables explicatives et 5 observations n'avait pas forcément de sens. Mais les organisations professionnelles aiment ça...)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	2	270.09			
Residuals	3	19.91	6.6367		
Total	5	290.00			

6.6367 nous intéresse car il s'agit de $\widehat{\sigma}^2$. Or on sait que l'estimateur de la matrice de variance de $\widehat{\boldsymbol{\beta}}$ est $\widehat{\sigma}^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$, donc la variance de $\widehat{\beta}_1$ est $6.6367 \cdot 0.0294 = 0.19512$. Aussi, la statistique de Student associée au test de $H_0 : \beta_1 = 0$ est

$$T = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_1]}} = \frac{-0.304}{\sqrt{0.19512}} = -0.6882$$

On retient la réponse C.



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

A) il diminue de 0.025

B) il diminue de 0.015

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

D) il augmente de 0.015

E) il augmente de 0.025

Il va falloir calculer ici deux $\overline{R}^2$. Avec les 2 variables explicatives, on peut utiliser la sortie de l'ANOVA qu'on a complétée :

$$\overline{R}^2_{(12)} = 1 - \frac{SCR/3}{SCT/5} = 1 - \frac{19.91/3}{290/5} = 0.8855$$

Pour le modèle une fois exclu la variable x_1, il faut ruser un peu. Mais la question précédante demandait de calculer la statistique de Student relative à la significativité de x_1 justement. Or la statistique de Fisher est le carré de cette statistique,

$$F = T^2 = 0.6882^2 = 0.4736 = \frac{SCR_2 - SCR_{12}}{1} \cdot \frac{1}{\widehat{\sigma}^2} = \frac{SCR_2 - 19.91}{6.6367}$$

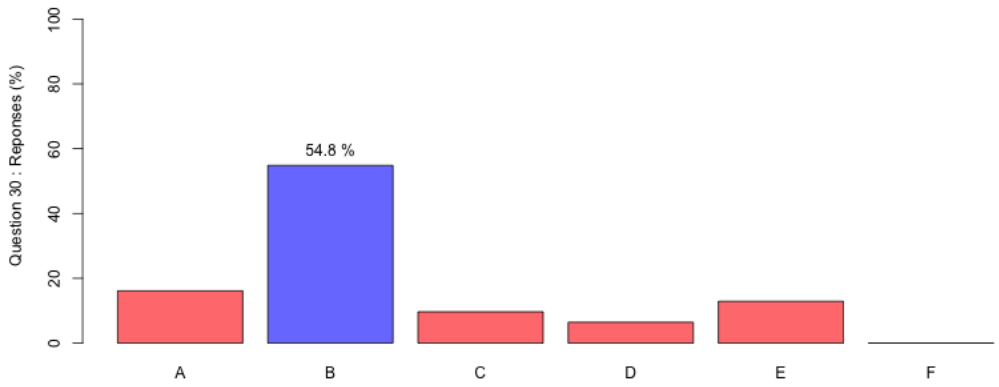
donc $SCR_2 = 6.6367 \times 0.4736 + 19.91 = 23.0534$. On peut alors maintenant calculer le $\overline{R}^2$:

$$\overline{R}^2_{(2)} = 1 - \frac{SCR_2/4}{SCT/5} = 1 - \frac{23.0534/4}{290/5} = 0.90006$$

(il faut faire attention à la petite correction avec 4 - soit 5-1 - car on a juste une variable dans ce modèle). Notons que

$$\overline{R}^2_{(12)} - \overline{R}^2_{(2)} = 0.8855 - 0.90006 = -0.015$$

ce qui correspond à une baisse de 0.015, qui est la réponse B. (Pour information, l'exercice 4.5.28 posait seulement la question de la variation de $\overline{R}^2$: j'ai préféré donner un indice avec la question précédante en forçant à calculer la statistique de Student avant)



Un petit commentaire rapide sur la dernière question, bonus, demandant une auto-évaluation... Je suis surpris d'avoir eu aussi peu de réponse. Pour vous, c'est la possibilité de gagner un point. Pour moi, ça permet de voir si (et qui) se sous-évalue, ou se sur-évalue. Ce qui me permet aussi de faire mon cours en fonction : dans une classe où beaucoup de monde se sur-évalue, je ne prends pas pour argent comptant la réponse à une question du genre *vous avez déjà vu cette notion dans un cours précédant ?*