STT5100 - Hiver 2019 - Examen Intra (OLS)

Arthur Charpentier

Examen **A**

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

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

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

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

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

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

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

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

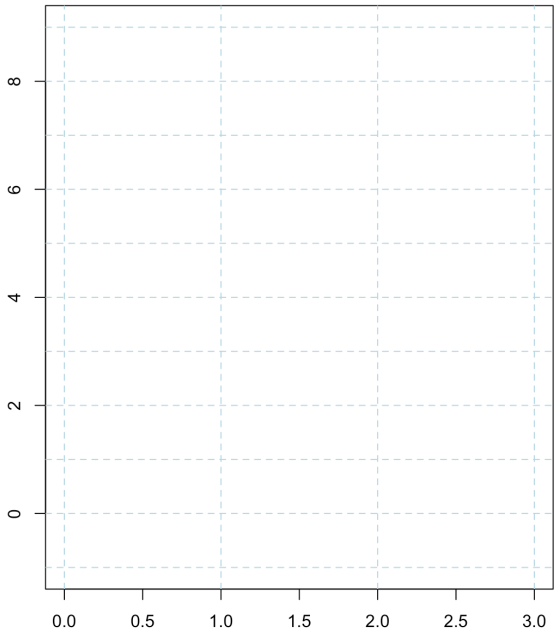
	75%	90%	95%	97.5%	99%	99.5%	99.9%	99.95%
Loi normale	0.674	1.282	1.645	1.960	2.326	2.576	3.090	3.291
Student (50)	0.679	1.299	1.676	2.009	2.403	2.678	3.261	3.496
Student (30)	0.683	1.310	1.697	2.042	2.457	2.750	3.385	3.646
Student (20)	0.687	1.325	1.725	2.086	2.528	2.845	3.552	3.849
Student (15)	0.691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

Code permanent :

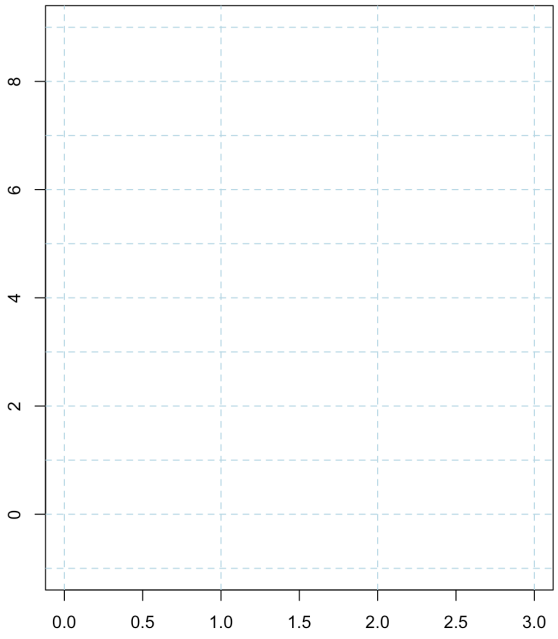
Sujet : A

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

question 7 :



question 29 :



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

A) 0

B) $\overline{y}$

C) $\overline{\varepsilon}$

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

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

C'était une question tirée d'ACTEX-SRM. L'estimateur de β_0 par la méthode des moindres carrés est

$$\widehat{\beta}_0 = \underset{b\in\mathbb{R}}{\text{argmin}} \left\{ \sum_{i=1}^{n}(y_i - b)^2 \right\}$$

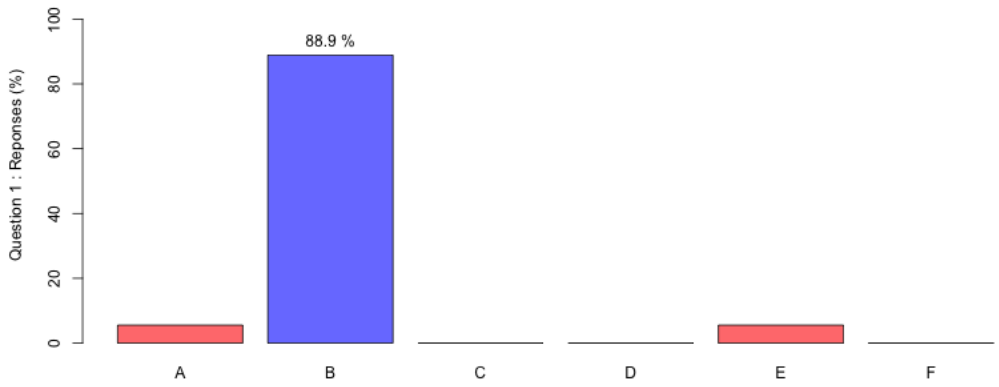
Comme on a un problème strictement convexe, on utilise la condition du premier ordre

$$\frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - b)^2\Big|_{b=\widehat{\beta}_0} = -\sum_{i=1}^{n} 2(y_i - \widehat{\beta}_0) = 0 \text{ soit } \sum_{i=1}^{n} y_i = n\widehat{\beta}_0$$

et donc

$$\widehat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}$$

(qui était une relation rappelée dans le tout premier cours). On a donc la réponse B.



2 On considère un modèle linéaire $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ estimé sur $n = 10$ observations. La variable x_i est ici binaire, et vaut 1 si l'individu i possède une carte de crédit, et 0 sinon. On sait que 40% des individus dans l'étude possèdent une carte de crédit. L'estimateur par moindre carrés de β_1 est 4. Enfin, on sait que

$\sum_{i=1}^{10}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 = 92$. Que vaut la statistique de Student du test $H_0 : \beta_1 = 0$? (*on retiendra la valeur la plus proche*)

A) 0.9

B) 1.2

C) 1.5

D) 1.8

E) 2.1

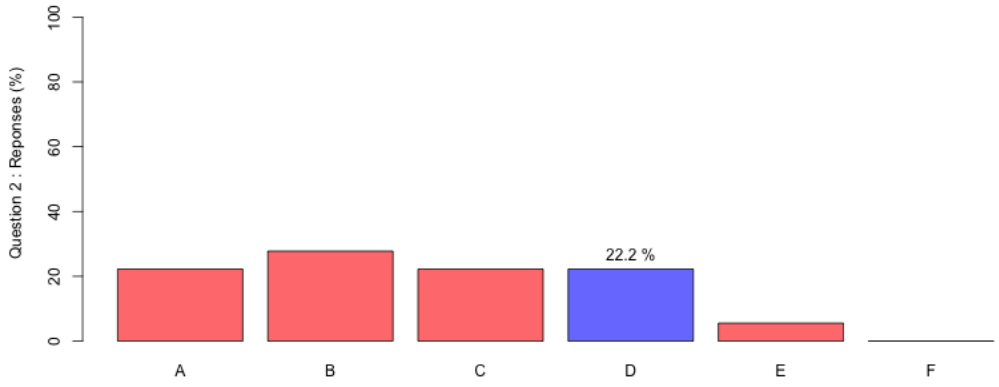
Il s'agit encore d'un exercice de la SOA, 'Course 4' d'octobre 2003. La statistique de test T est basée sur la variance estimée de $\widehat{\beta}_1$. Rappelons que

$$\widehat{\mathsf{Var}}[\widehat{\beta}_1] = \frac{\widehat{\sigma}^2}{S_x^2} = \left(\frac{1}{10-2}\sum_{i=1}^{10}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2\right) / \left(6\cdot(0-\overline{x})^2 + 4\cdot(1-\overline{x})^2\right) = \frac{115}{24}$$

avec $\overline{x} = 0.4$. La statistique de test est alors

$$T = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_1]}} = \frac{4}{\sqrt{115/24}} \sim 1.8273$$

On a donc la réponse D.



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

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

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

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

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

Residual standard error: 1.423 on 49 degrees of freedom
Multiple R-squared:  0.7913,Adjusted R-squared:  0.7828
F-statistic: 92.88 on 2 and 49 DF,  p-value: < 2.2e-16
```

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

A) environ 0.001

B) environ 0.002

C) environ 0.005

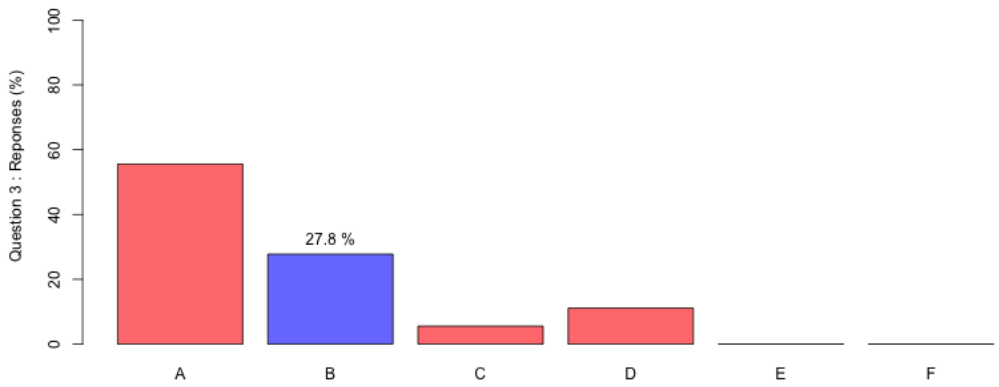
D) environ 0.010

E) environ 0.020

La statistique de test est ici

$$T = \frac{\widehat{\beta}_0}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_0]}} = \frac{1.4956}{0.4572} \sim 3.271140$$

En utilisant les valeurs indiquées sur la première page, on a une valeur de l'ordre de 0.2% (plus précisément 0.1965221%) qui correspond à la réponse B.



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

A) moins de 6

B) entre 6 et 7

C) entre 7 et 8

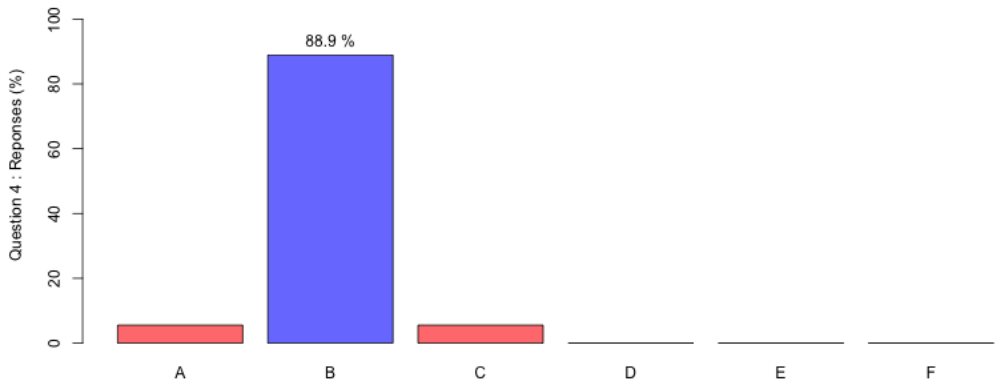
D) entre 8 et 9

E) plus que 9

On veut estimer $\widehat{y} = \mathbb{E}[Y|x_1 = 1.5, x_2 = \text{Homme}]$ et pour cela on utilise l'estimation obtenue :

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \underbrace{x_1}_{=1.5} + \widehat{\beta}_2 \underbrace{\mathbf{1}(x_2 = \text{Homme})}_{=1} = \widehat{\beta}_0 + 1.5\widehat{\beta}_1 + \widehat{\beta}_2 = 1.495648 + 1.658419 \cdot 1.5 + 2.508923 \sim 6.492199$$

On a donc la réponse B.



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

A) moins de 6

B) entre 6 et 7

C) entre 7 et 8

D) entre 8 et 9

E) plus que 9

On nous demande d'estimer $\widehat{\text{Var}}[\widehat{Y}|x_1 = 1.5, x_2 = \text{Homme}]$. Or comme $\widehat{Y} = \widehat{\beta}_0 + 1.5 \cdot \widehat{\beta}_1 + \widehat{\beta}_2$, on peut en déduire que l'on cherche

$$\text{Var}[\widehat{\beta}_0 + 1.5 \cdot \widehat{\beta}_1 + \widehat{\beta}_2]$$

soit

$$\text{Var}[\widehat{\beta}_0] + 1.5^2 \cdot \text{Var}[\widehat{\beta}_1] + \text{Var}[\widehat{\beta}_2] + 2 \cdot 1.5 \cdot \text{Cov}[\widehat{\beta}_0, \widehat{\beta}_1] + 2 \cdot \text{Cov}[\widehat{\beta}_0, \widehat{\beta}_2] + 2 \cdot 1.5 \cdot \text{Cov}[\widehat{\beta}_1, \widehat{\beta}_2]$$

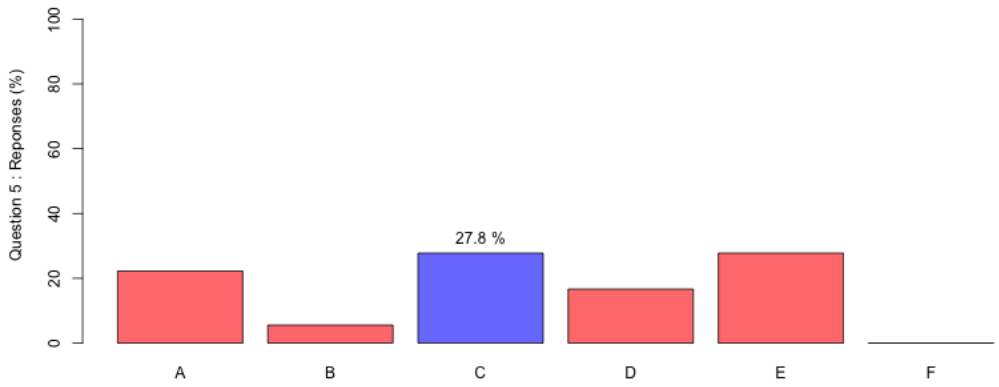
Malheureusement, la matrice de variance-covariance n'étant pas fournie (*oubli de ma part*) en première approximation, on va se contenter des 3 premiers termes,

$$\mathsf{Var}[\widehat{\beta}_0] + 1.5^2 \cdot \mathsf{Var}[\widehat{\beta}_1] + \mathsf{Var}[\widehat{\beta}_2] = 0.20905495 + 1.5^2 \cdot 0.02206977 + 0.16263106 \sim 0.421343$$

Un intervalle de confiance à 95% pour $\widehat{y}$ sera de la forme

$$\left[\widehat{y} \pm 2\sqrt{\widehat{\mathsf{Var}}[\widehat{Y}]}\right] = \left[6.492199 + 2\sqrt{0.421343}\right] \sim [5.1939, 7.7904]$$

On a donc la réponse C. On notera qu'il est étonnant de répondre A si on a répondu B à la question précédante.



Pour s'entrainer correctement, je peux remettre la matrice de variance covariance que j'avais oubliée

```
              (Intercept)           x1 x2 == "2"TRUE
(Intercept)    0.20905495 -0.055889268  -0.046796591
x1            -0.05588927  0.022069768  -0.008183933
x2 == "2"TRUE -0.04679659 -0.008183933   0.162631057
```

On peut alors montrer qu'avec les termes de covariances, la variance de $\widehat{Y}$ est de 0.1355302 ce qui donne un intervalle de confiance à 95% de la forme $[5.752385, 7.232012$. La réponse C reste valide (même si on voit que notre approximation était *très* grossière...

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

A) plus de 12.5%

B) environ 10%

C) environ 5%

D) environ 1%

E) moins de 0.5%

Rappelons que $Y|X_1, X_2)$ suit une loi normale $\mathcal{N}(\beta_0 + \beta_1 x_1 + \beta_2 \mathbf{1}(x_2 = \text{Homme}), \sigma^2)$ On cherche ici

$$\mathbb{P}\left[\beta_0 + 1.5 \cdot \beta_1 + \beta_2 + \sigma Z > 9\right]$$

où $Z \sim \mathcal{N}(0,1)$, ou encore

$$\mathbb{P}\left[Z > \frac{9-(\beta_0+1.5\cdot\beta_1+\beta_2)}{\sigma}\right]$$

En première approximation, on remplaçant les grandeurs inconnues par leurs estimateurs, on cherche

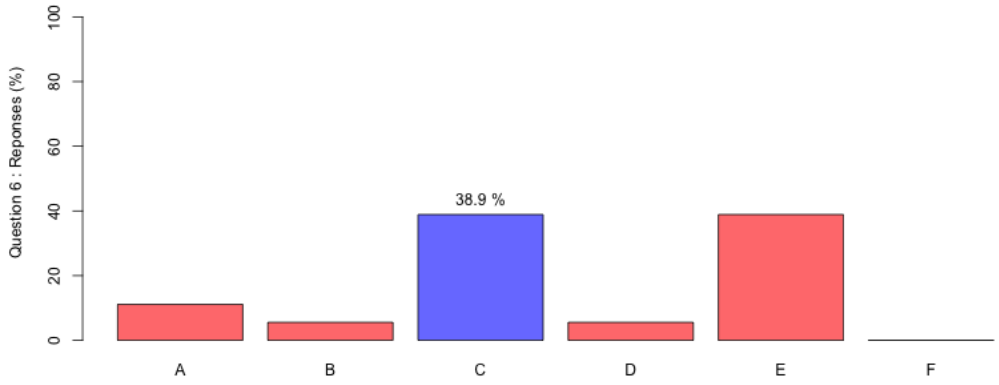
$$\mathbb{P}\left[Z > \frac{9-(\widehat{\beta}_0+1.5\cdot\widehat{\beta}_1+\widehat{\beta}_2)}{\widehat{\sigma}}\right] = \mathbb{P}\left[Z > \frac{9-(1.4956+1.5\cdot 1.6584+2.5089)}{1.423}\right] = \mathbb{P}[Z > 1.762403]$$

Dans la table de la page 1, on nous dit que

$$\mathbb{P}[Z > 1.645] = 5\% \text{ et } \mathbb{P}[Z > 1.96] = 2.5\%$$

On est donc entre 5% et 2.5%, la seule réponse possible est donc C. En fait, la vraie probabilité est de l'ordre de 4% :

```
> pnorm(9,6.492199,1.423)
[1] 0.9609935
```



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

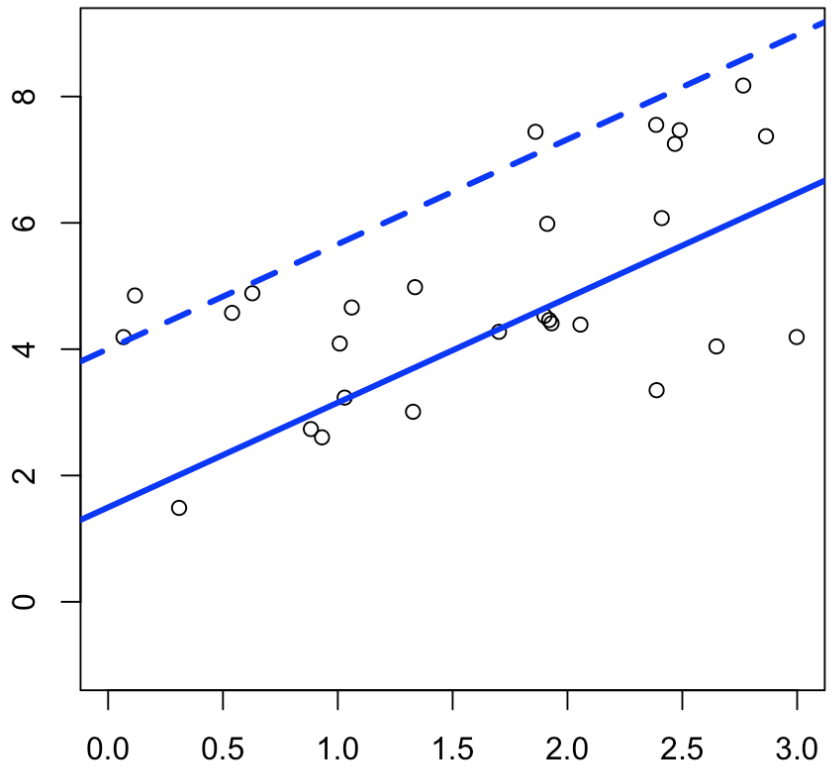
- tracez avec un trait plein l'estimateur de $\mathbb{E}[Y|x_1, x_2 = \text{Femme}]$ lorsque x varie entre 0 et 3
- tracez avec un trait en pointillés l'estimateur de $\mathbb{E}[Y|x_1, x_2 = \text{Homme}]$ lorsque x varie entre 0 et 3

Pour chaque des droites de régression, on a les équations :

$$\begin{cases} 1.4956 + 1.6584\cdot x_1 & \text{pour les femmes} \\ (1.4956+2.5089) + 1.6584\cdot x_1 = 4.0045 + 1.6584\cdot x_1 & \text{pour les hommes} \end{cases}$$

On observe que les deux droites sont parallèles, et séparées de 2.5089. On peut aussi se contenter de tracer deux points, puis de les relier par une droite : le point pour $x_1 = 0$ et on peu aussi se place à droite de la figure, en $x_1 = 3$ (par exemple). Autrement dit, les droites de régression passent par les points

$$\begin{cases} (0, 1.4956) \sim (0, 1.5) \text{ et } (3, 6.4708) \sim (3, 6.5) & \text{pour les femmes} \\ (0, 4.0045) \sim (0, 4) \text{ et } (3, 8.9797) \sim (3, 9) & \text{pour les hommes} \end{cases}$$



Pour avoir tous les points, il fallait avoir les droites parallèles, dans le bon ordre, et passsant (approximativement) par les bons points.

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

A) moins que -80

B) entre -80 et -60

C) entre -60 et -40

D) entre -40 et -20

E) plus que -20

Rappelons que pour une régression linéaire, la log-vraisemblance est

$$\log \mathcal{L} = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n} \underbrace{[y_i - (\beta_0 + \beta_1 x_1 + \beta_2 \mathbf{1}(x_2 = \text{homme})]^2}_{=\varepsilon_i^2}$$

L'idée est de remplacer les termes inconnus par leur estimateur. Autrement dit, on a ici

$$\log \mathcal{L} \sim -\frac{n}{2}\log(2\pi) - n\log(\widehat{\sigma}) - \frac{1}{2\widehat{\sigma}^2}\cdot n\widehat{\sigma}^2 = -\frac{n}{2}[1 + \log(2\pi)] - n\log(\widehat{\sigma})$$

soit numériquement

$$\log \mathcal{L} \sim -\frac{52}{2}2.837877 - 52 \cdot \log(1.423252) \sim -92.1379$$

ce qui est la réponse A.

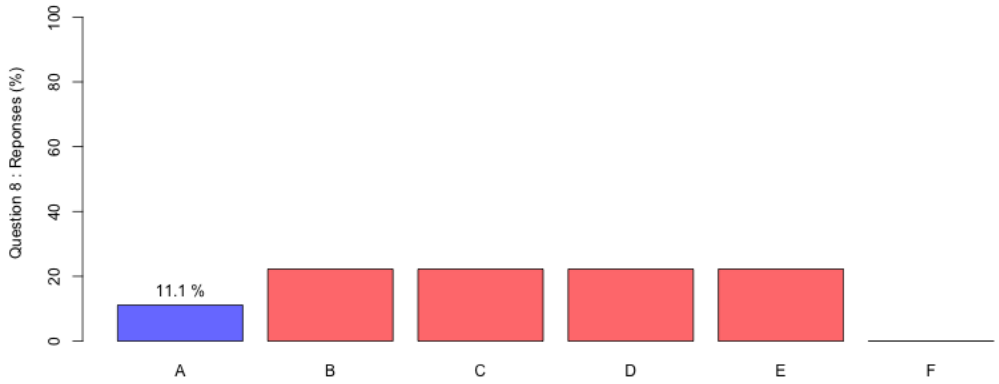
En fait, il s'agit d'un calcul approché, à cause des histoires de $(n-3)$ au lieu des n qui devraient apparaître quand on utilise un estimateur de l'écart-type. En faisant *vraiment* le calcul sous R, on obtient

```
> logLik(lm(y~x1+x2))
'log Lik.' -90.59289 (df=4)
```

que l'on peut comparer à

```
> reg = lm(y~x1+x2)
> sigma = summary(modlm)$sigma
> sum(log(dnorm(x = y, mean = predict(reg), sd = sigma)))
[1] -90.6379
> n = length(y)
> sigma.ML = sigma*sqrt((n-dim(model.matrix(reg))[2])/n)
> sum(log(dnorm(x = y, mean = predict(modlm), sd = sigma.ML)))
[1] -90.59289
```

Bref, dans tous les cas, on a environ -90, qui correspond à la réponse A.



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

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

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

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

II. les deux modèles ont le même R^2

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

Parmi les affirmations, lesquelles sont vraies :

A) aucune

B) I et II seulement

C) I et III seulement

D) II et III seulement

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

Question ACTEX SRM. L'affirmation I est fausse, cela vient de la géométrie du problème : les résidus ne sont du tout liés. En effet, dans un grahiques x (sur l'axe des abscisses) et y (sur l'axe des ordonnées), les résidus de la régression de y sur x se lisent *verticalement* alors que ceux de la régression de y sur x se lisent *horizonalement*. Pour s'en convaincre, regardons les pentes des deux droites :

$$\widehat{\beta}_{1:y\sim x} = \text{corr}[x,y]\frac{S_y}{S_x} \text{ et } \widehat{\beta}_{1:x\sim y} = \text{corr}[y,x]\frac{S_x}{S_y}.$$

Aussi, si on fait le produit des deux pentes, on obtient $\widehat{\beta}_{1:y\sim x}\cdot\widehat{\beta}_{1:x\sim y} = \text{corr}[x,y]^2 = R^2$. Autrement dit, à moins que les points (x_i, y_i) soient alignés suivant une droite (et dans ce cas, il n'y a pas de résidus) alors $R^2 < 1$ et donc la pente de l'un n'est *jamais* l'inverse de l'autre.

L'affirmation II est valide. En fait, on l'utilisait dans la question précédante, car $R^2 = \text{corr}[x,y]^2 = \text{corr}[y,x]^2$, par esymmétrie de l'opérateur de corrélatation (et de covariance en fait).

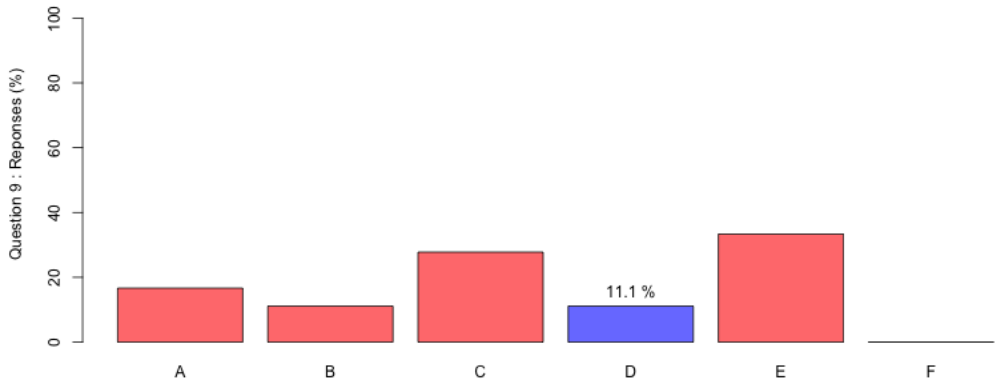
L'affirmation III est valide. En fait, c'est lié a la question précédante. En effet, comme on a le même R^2 on a la même statistique F pour le test de significativité globale. Or quand on a une seule variable explicative, la significativité globale est la même chose que la significativité de la variable explicative. Certes, on n'utilise pas la même statistique de test, mais les tests sont *équivalents* au sens où les p-value sont les mêmes. Et donc les deux statistiques de test sont identiques... On peut s'en convaincre en regardant les sorties suivantes :

```
Call:
lm(formula = speed ~ dist, data = cars)

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  8.28391    0.87438   9.474 1.44e-12 ***
dist         0.16557    0.01749   9.464 1.49e-12 ***

Residual standard error: 3.156 on 48 degrees of freedom
Multiple R-squared:  0.6511,        Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12

Call:
lm(formula = dist ~ speed, data = cars)

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,        Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Par contre, ce n'est pas le cas pour la constante, qui peut etre significative dans un cas, et pas dans l'autre (mais c'est géométriquement intuitive : dans un cas, on regarde l'intersection de la droite de régression avec l'axe $x = 0$

(axe des ordonnées) et dans l'autre l'intersection de la droite de régression avec l'axe $y = 0$ (axe des abscisses) les deux n'ayant rien à voir - sauf peut-être si les deux variables étaient centrées).

On a donc la réponse D, comme II et III sont justes.



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

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

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

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

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

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

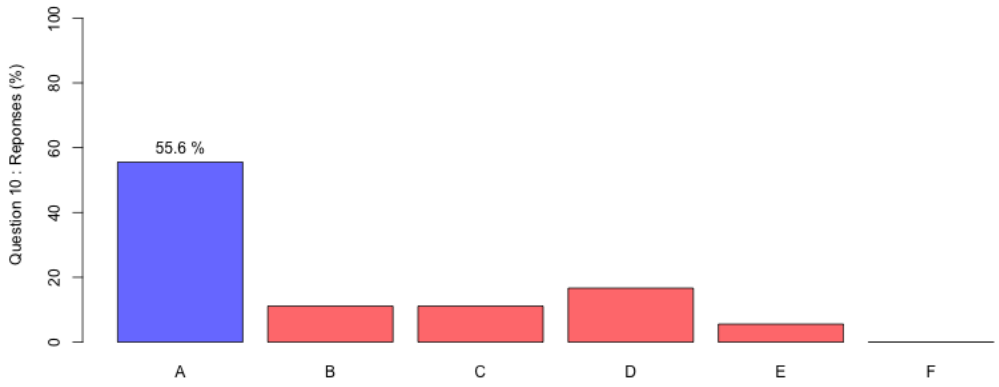
B) $F_{6,77} = \dfrac{(SCR_C - SCR_A - SCR_B)/6}{(SCR_A + SCR_B)/77}$

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

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

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

Question de l'examen 'Course 4' de l'automne 2000. C'est juste une analyse de variance dans le context du test de Chow, cf wikipedia (ou le cours en ligne). Pour rappel, les degrés de liberté sont, pour le numérateur $(n-3) - (n/2-3+n/2-3)$, soit 3, et pour le dénominateur $(n/2-3+n/2-3) = n-6$ soit 76. On a donc la réponse A.



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

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

On veut tester $H_0 : \beta_1 = 1$ contre l'hypothèse alternative $H_1 : \beta_1 \neq 1$. Quel serait la plus petite valeur de α pour que H_0 soit rejetée avec un niveau de significativité α ?

A) moins de 0.01

B) entre 0.01 et 0.02

C) entre 0.02 et 0.05

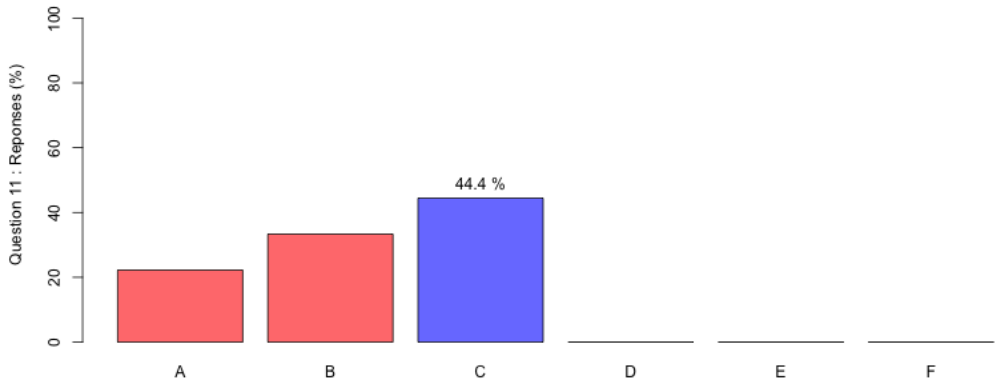
D) entre 0.05 et 0.10

E) plus de 0.10

Question de l'examen ST d'avril 2014. On utilise le test de Student,

$$T = \frac{\widehat{\beta}_1 - 1}{\sqrt{\widehat{\text{Var}}[\widehat{\beta}_1]}} = \frac{1.15 - 1}{0.043} \sim 3.488$$

car on cherche à tester $H_0 : \beta_1 = 1$ (et non pas 0 comme classiquement pour un test de significativité). On a ici 6 observations, donc sous H_0, on sait que T suit une loi $\mathcal{S}td(n-2)$ - i.e. $\mathcal{S}td(4)$. Or on nous dans la table page 1 que $\mathbb{P}[T > 3.747] = 1\%$ donc $\mathbb{P}[T > 3.488] = 1\%^+$ (pour dire que ça sera un peu plus de 1% (disons 1.25% pour fixer les idées). Mais comme on fait un test *bilatéral*, on a plutôt $\mathbb{P}[|T| > 3.488] = 2 \times 1\%^+ = 2\%^+$. On est donc un peu au dessus de 2%, ce qui correspond à la réponse C.



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

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

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

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

A) la constante

B) la constante et x_1

C) la constante et x_2

D) la constante, x_1 et x_3

E) la constante, x_2 et x_3

Il y avait une petite coquille : il fallait lire

```
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
```

Il s'agissait d'une question de l'examen ST de l'automne 2015. On va simplement remplir le tableau. Pour x_1 par exemple, on calcule la statistique du test de Student,

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

La première méthode est la suivante : on cherche alors $|t_1|$ dans la table de la page 1, avec 30 degrés de liberté; on peut voir que $\mathbb{P}[T > |t_1|] \sim 1\%$, or $\mathbb{P}[|T| > |t_1|] \sim 2\%$, qui correspond à `Pr(>|t|)` (dans la sortie de la régression). Comme β_1 est considéré comme significativement différent de zéro pour un seuil $\alpha = 10\%$, car $\alpha > 2\%$. La seconde méthode (peut-être la plus simple ici) est de revenir á la region de rejet / acceptation : si on revient sur la page 1, le quantile à 95% de la loi de Student est 1.6972. Autrement dit $\mathbb{P}[|T| > 1.6972|] = 10\%$. Autrement dit, si la statistique du test de Student est dans l'intervalle $[-1.6972, +1.6972]$, les variables sont considérées comme non-significativement différentes de 0, et sinon, elles sont significativement différente de 0. Reprenons pour toutes les variables

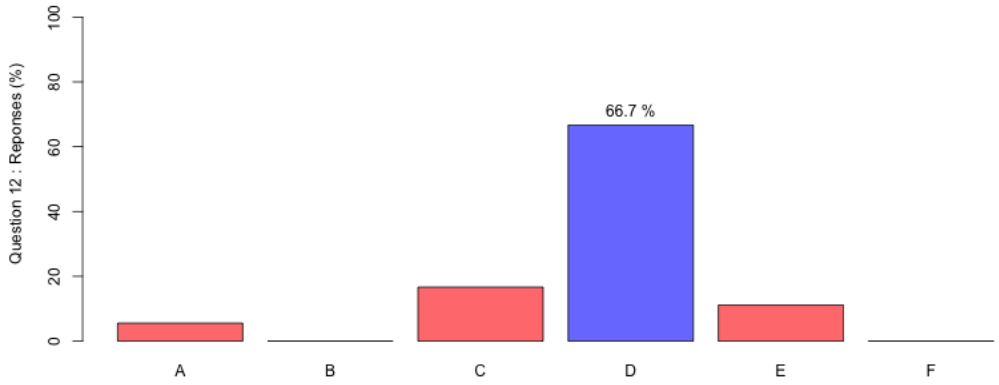
$$t_0 = \frac{\widehat{\beta}_0}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_0]}} = \frac{44.200}{5.960} \sim 7.416 \notin [-1.6972, +1.6972] : \text{ significativement non nul}$$

$$t_1 = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_1]}} = \frac{-0.295}{0.118} \sim -2.5 \notin [-1.6972, +1.6972] : \text{ significativement non nul}$$

$$t_2 = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_1]}} = \frac{9.110}{6.860} \sim 1.327 \in [-1.6972, +1.6972] : \text{ non-significativement non nul}$$

$$t_3 = \frac{\widehat{\beta}_1}{\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_1]}} = \frac{-8.700}{1.200} \sim -7.25 \notin [-1.6972, +1.6972] : \text{ significativement non nul}$$

On a donc la réponse D.



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

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

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

On sait de plus que

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

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

A) moins de 1

B) entre 1 et 3

C) entre 3 et 5

D) au moins 5

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

Il s'agit d'une question de l'examen MAS-I d'octobre 2018. Il s'agit du test standard de comparaison de variance. On peut revenir deux minutes sur la théorie. Considérons ici deux modéles, (0) et (1) avec (0) inclus dans (1) (en anglais (0) *is 'nested' within* (1)), et on suppose que p_0 et p_1 sont respectivement les nombres de variables explicatives. Par hypothèse, $p_1 > p_0$. Ici, (0) c'est le modèle sous H_0, autrement dit $y_i = \beta_0 + \varepsilon_i$ alors que le modèle (1) est celui sous H_1. Et dans ce cas, on utilise la statistique de Fisher $y_i = \beta_0+\beta_1 x_{1,i}+\beta_2 x_{2,i}+\beta_3 x_{3,i}+\varepsilon_i$

$$F = \frac{\left(\frac{\mathsf{SCR}_0-\mathsf{SCR}_1}{p_1-p_0}\right)}{\left(\frac{\mathsf{SCR}_1}{n-p_1}\right)} = \frac{n-p_1}{p_1-p_0}\cdot\frac{\mathsf{SCR}_0-\mathsf{SCR}_1}{\mathsf{SCR}_1}$$

(qui doit suivre une loi de Fisher si H_0 est vraie) où SCR correspond à la somme des carrés de résidus. Donc ici,

$$\mathsf{SCR}_0 = \sum_{i=1}^{20}(y_i-\widehat{\beta}_0)^2 = \sum_{i=1}^{20}(y_i-\overline{y})^2 = 7.62956$$

alors que

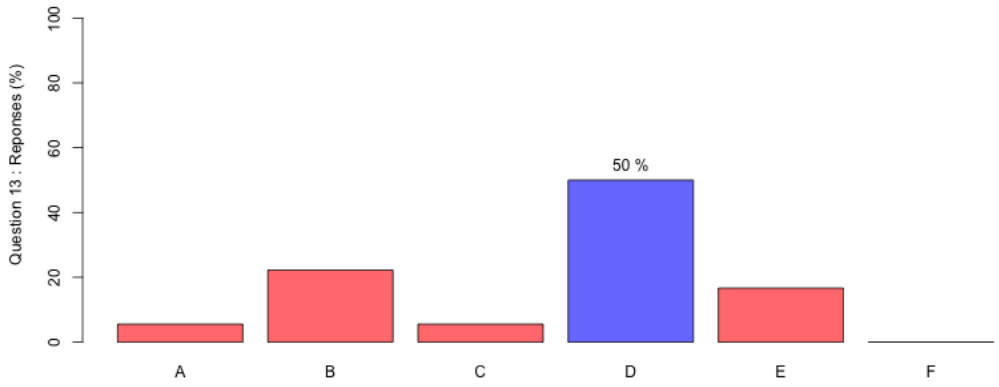
$$\mathsf{SCR}_1 = \sum_{i=1}^{20}(y_i-\widehat{\beta}_0-\widehat{\beta}_1 x_{1,i}+\widehat{\beta}_2 x_{2,i}+\widehat{\beta}_3 x_{3,i})^2 = \sum_{i=1}^{20}(y_i-\widehat{y}_i)^2 = \sum_{i=1}^{20}\widehat{\varepsilon}_i^2 = 2.65376$$

donc si on substitue

$$F = \frac{n-p_1}{p_1-p_0}\cdot\frac{\mathsf{SCR}_0-\mathsf{SCR}_1}{\mathsf{SCR}_1} = \frac{20-4}{4-1}\cdot\frac{7.62956-2.65376}{2.65376} = 10$$

On a donc la réponse D.

Le corrigé de la CAS donnait comme réponse B, mais je suis ici d'accord avec la majorité. Pour une loi de Fisher $\mathcal{F}_{20-4,4-1}$ - ou $\mathcal{F}_{16,3}$ - la moyenne est de 3 (=3/(3−2)) et le quantile à 90% est à 5 (celui à 95% est à 8.7). Autrement dit, si on avait un modèle avec 3 variables explicatives non-significatives, une statistique de 5 serait acceptable, car effectivement les trois variables ne servent à rien. Or ici, les trois variables explicatives sont significatives, avec un test de Student (comme le montre la sortie de la régression). Aussi, on doit s'attendre à avoir une statistique supérieure à 5... donc 5 a du sens.



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

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

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

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

et

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

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

A) moins de -1

B) entre -1 et 0

C) entre 0 et 1

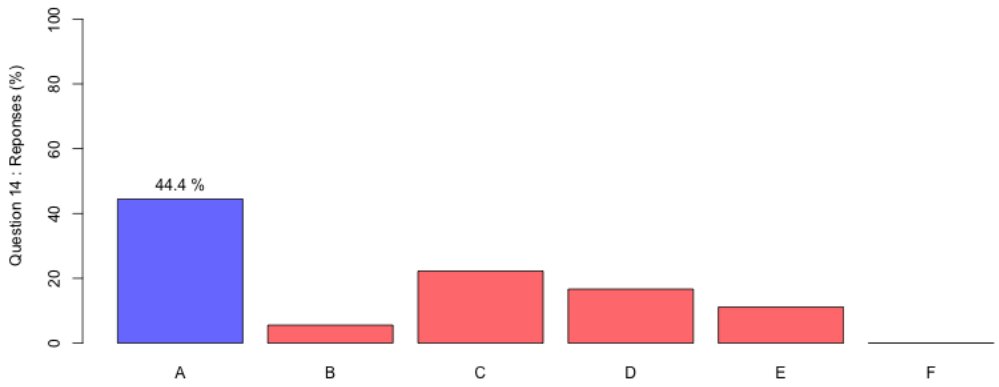
D) entre 1 et 2

E) plus de 2

Question de l'examen MAS-I d'octobre 2018. On nous donne ici $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ qui correspond à $\widehat{\boldsymbol{\beta}}$. Or le vecteur de prédictions est ici $(\boldsymbol{X}\boldsymbol{\beta}$, en sachant que la seule valeur qui nous intéresse est $\widehat{y}_5$, soit

$$\widehat{\boldsymbol{y}} = \begin{pmatrix} 1 & 0 & 1 & 9 \\ 1 & 1 & 1 & 15 \\ 1 & 1 & 1 & 8 \\ 1 & 1 & 0 & 7 \\ 1 & 1 & 0 & 6 \\ 1 & 0 & 0 & 6 \end{pmatrix} \begin{pmatrix} 2.335 \\ 0.297 \\ -0.196 \\ 1.968 \end{pmatrix} = \begin{pmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ 1 \cdot 2.335 + 1 \cdot 0.297 + 0 \cdot -0.196 + 6 \cdot 1.968 \\ \cdot \end{pmatrix} = \begin{pmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ 14.44 \\ \cdot \end{pmatrix}$$

donc le résidus associé est $\widehat{\varepsilon}_5 = y_5 - \widehat{y}_5 = 13 - 14.4 = -1.4$. On a donc la réponse A.



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

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

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

C) σ^2

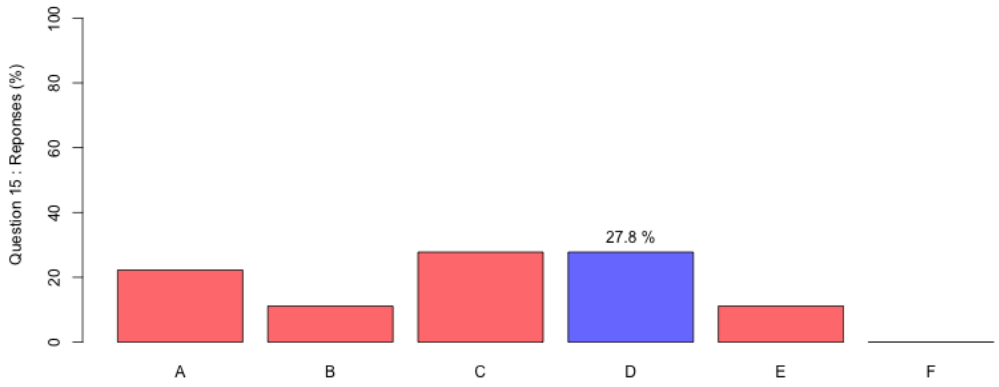
D) 0

E) 45

Pour commencer, notons que $\widehat{\boldsymbol{y}}$ et $[\boldsymbol{y} - \widehat{\boldsymbol{y}}]$ sont deux vecteurs (de dimension 6) donc le produit matriciel $\widehat{\boldsymbol{y}}^\top[\boldsymbol{y} - \widehat{\boldsymbol{y}}]$ est une somme de 6 termes : c'est une constante ! ce qui exclue d'office les deux premières réponses proposées. Si on regarde un peu, $\widehat{\boldsymbol{y}}$ est la projection orthogonale de $\boldsymbol{y}$ sur l'espace (linéaire) engendré par les variables explicatives. Et $[\boldsymbol{y} - \widehat{\boldsymbol{y}}]$ correspond au vecteur des résidus, $\widehat{\varepsilon}$. En faisant un peu de géométrie, $\widehat{\boldsymbol{y}} = \Pi_{\boldsymbol{X}}\boldsymbol{y}$ alors que $\widehat{\boldsymbol{y}} = \Pi_{\boldsymbol{X}^\perp}\boldsymbol{y}$. Bref, les résidus et la projection sont, par construction, orthogonaux. C'est en fait la traduction de la condition du premier ordre de la minimisation de la somme des carrés des erreurs. Bref,

$$\widehat{\boldsymbol{y}}^\top[\boldsymbol{y} - \widehat{\boldsymbol{y}}] =< \widehat{\boldsymbol{y}}, \widehat{\varepsilon} >= 0 \text{ car } \widehat{\boldsymbol{y}} \perp \widehat{\varepsilon}$$

On a donc la réponse D.



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

A) moins de 100

B) entre 100 et 110

C) entre 110 et 120

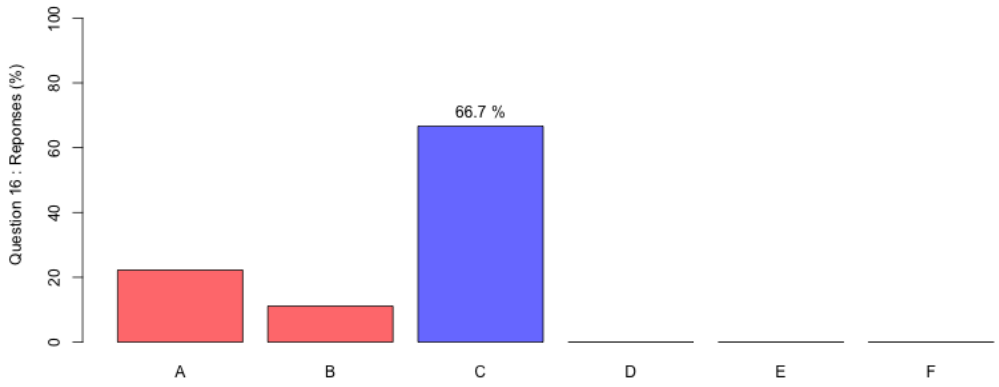
D) entre 120 et 130

E) plus de 130

On l'a vu longuement en cours : dans une régression linéaire, $\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0$, c'est la condition du premier ordre quand on dérive la somme des carrés des résidus par β_0. On dira alors que "*les résidus sont centrés*". Cette condition peut en fait se réécrire simplement $\sum_{i=1}^{n} \widehat{y}_i = \sum_{i=1}^{n} y_i$, donc ici, il suffit de calculer la somme des y_i :

$$\sum_{i=1}^{6} \widehat{y}_i = (19 + 32 + 19 + 17 + 13 + 15) = 115$$

On a donc la réponse C.



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

A) $i = 1$
B) $i = 2$
C) $i = 3$
D) $i = 4$
E) $i = 5$

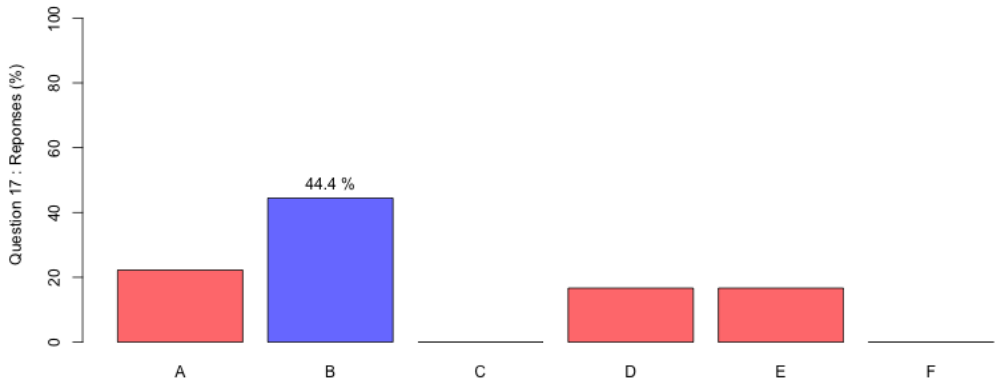
On en avait parlé au dernier cours (en parlant d'outliers et de points influents). On avait vue que si $\boldsymbol{H}$ est la *matrice chapeau* (ou *hat matrix*), alors, comme $\widehat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$. Un peu d'algèbre linéaire et de calcul différentiel permet décrire que $H_{i,i}$ est alors égal à $\frac{\partial \widehat{y}_i}{\partial y_i}$. Or cette matrice $\boldsymbol{H}$ est tout simplement $\boldsymbol{X}(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal$. En effet,

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \underbrace{\boldsymbol{X}\cdot(\boldsymbol{X}^\intercal\boldsymbol{X})^{-1}\boldsymbol{X}^\intercal}\boldsymbol{y}.$$

On regarde donc sur la diagonale de la dernière matrice.

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

On a donc la réponse B.



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

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

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

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

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

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

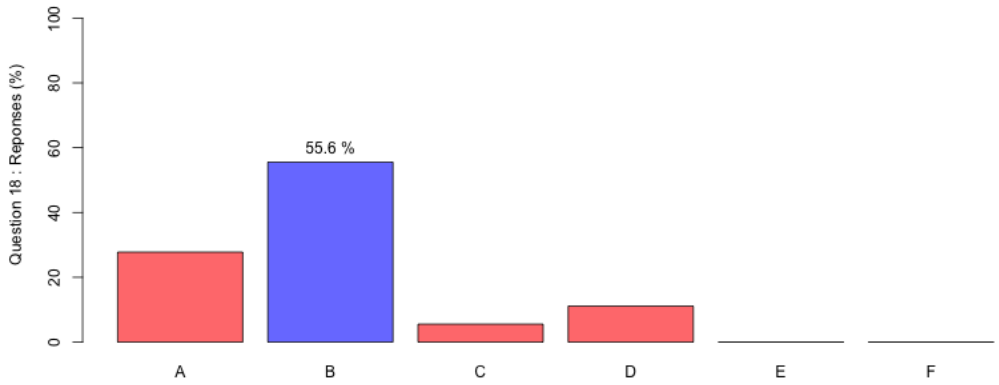
Question de l'examen SOA 'Part 3' d'octobre 1984. Faisons le tranquillement : on nous demande

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

donc (classiquement), on différencie (par rapport à β) pour obtenir la condition du premier ordre,

$$2 \sum_{i=1}^n (-x_i^2)(y_i - \widehat{\beta} x_i^2) = 0 \text{ soit } \sum_{i=1}^n -x_i^2 \cdot y_i + \widehat{\beta} \sum_{i=1}^n x_i^2 \cdot x_i^2 = 0, \text{ i.e. } \widehat{\beta} = \frac{\sum_{i=1}^n x_i^2 y_i}{\sum_{i=1}^n x_i^4}$$

On a donc la réponse B.



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

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

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

Deux actuaires adoptent des statégies différentes pour construire leur modèle

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

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

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

A) moins de 0.15

B) entre 0.15 et 0.30

C) entre 0.30 et 0.45

D) entre 0.45 et 0.60

E) plus de 0.60

Là encore, on a une question de l'examen MAS-I d'octobre 2018. Commençons par rappeler que $AIC = 2p - 2\log\mathcal{L}$, et qu'un AIC est meilleur qu'un autre s'il est plus *faible*. On préfèra un modèle avec un paramètre de moins si la différence entre log-vraisemblance associée est inférieure à 1. Le plus simple est de tout réécrire (en tous cas pour expliquer) : par exemple pour le premier modèle, $AIC/2 = 1 - \log\mathcal{L}$, etc. Et pour simplifier, on va mettre p et pas R^2

modèle	variables	p	$\log\mathcal{L}$	$AIC/2$
(1)	constante	1	0.05	0.95
(2)	constante, x_1	2	1.30	0.70
(3)	constante, x_2	2	1.40	0.60
(4)	constante, x_3	2	1.20	0.80
(5)	constante, x_4	2	1.15	0.85
(6)	constante, x_5	2	1.10	0.90
(7)	constante, x_1, x_2	3	2.50	0.50
(8)	constante, x_1, x_3	3	2.75	0.25
(9)	constante, x_1, x_4	3	2.60	0.40
(10)	constante, x_1, x_5	3	3.00	0.00
(11)	constante, x_2, x_3	3	2.50	0.50
(12)	constante, x_2, x_4	3	2.55	0.45
(13)	constante, x_2, x_5	3	2.90	0.10
(14)	constante, x_3, x_4	3	2.80	0.20
(15)	constante, x_3, x_5	3	2.75	0.25
(16)	constante, x_4, x_5	3	2.45	0.55
(17)	constante, x_1, x_2, x_3	4	3.35	0.65
(18)	constante, x_1, x_2, x_4	4	3.25	0.75
(19)	constante, x_1, x_2, x_5	4	3.30	0.70
(20)	constante, x_1, x_3, x_4	4	3.50	0.50
(21)	constante, x_1, x_3, x_5	4	3.60	0.40
(22)	constante, x_1, x_4, x_5	4	3.90	0.10
(23)	constante, x_2, x_3, x_4	4	3.70	0.30
(24)	constante, x_2, x_3, x_5	4	3.40	0.60
(25)	constante, x_2, x_4, x_5	4	3.45	0.55
(26)	constante, x_3, x_4, x_5	4	3.35	0.65
(27)	constante, x_1, x_2, x_3, x_4	5	4.20	0.80
(28)	constante, x_1, x_2, x_3, x_5	5	3.95	1.05
(29)	constante, x_1, x_2, x_4, x_5	5	4.10	0.90
(30)	constante, x_1, x_3, x_4, x_5	5	4.00	1.00
(31)	constante, x_2, x_3, x_4, x_5	5	4.05	0.95
(32)	constante, x_1, x_2, x_3, x_4, x_5	6	4.25	1.75

Pour la valeur minimale du AIC, on a ici obtenu un 0 ! Pour le AIC forward, on compare (pour commencer) le modèle sans rien avec les modèles avec une variables, et on les ordonne en fonction du AIC - croissant:

modèle	variables	p	$\log\mathcal{L}$	$AIC/2$
(3)	constante, x_2	2	1.40	0.60
(2)	constante, x_1	2	1.30	0.70
(4)	constante, x_3	2	1.20	0.80
(5)	constante, x_4	2	1.15	0.85
(6)	constante, x_5	2	1.10	0.90
(1)	constante	1	0.05	0.95

Comme le modèle (3) est le plus intéressant, on va garder x_2. On va alors comparer le modèle avec juste x_2 et les modèles avec deux variables dont x_2. Et là encore, on réordonne :

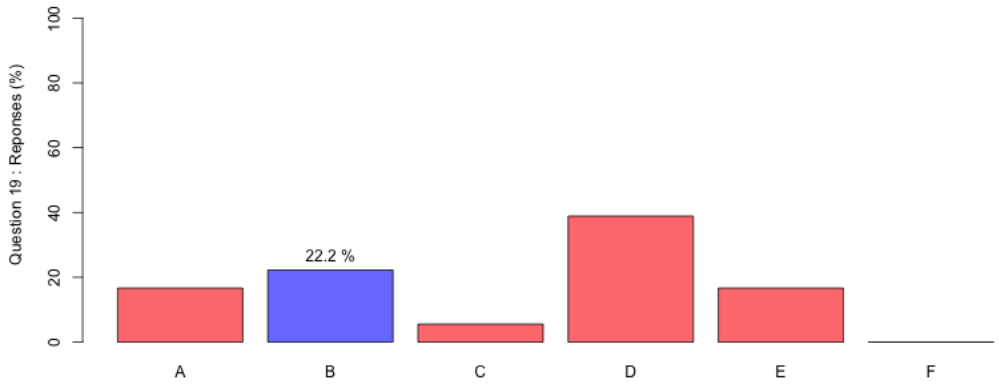
modèle	variables	p	$\log\mathcal{L}$	$AIC/2$
(13)	constante, x_2, x_5	3	2.90	0.10
(12)	constante, x_2, x_4	3	2.55	0.45
(7)	constante, x_1, x_2	3	2.50	0.50
(11)	constante, x_2, x_3	3	2.50	0.50
(3)	constante, x_2	2	1.40	0.60

Comme le modèle (14) est le plus intéressant, on va garder x_5 (en plus de x_2). On va alors comparer le modèle avec juste x_2 et x_5 et les modèles avec trois variables dont x_2 et x_5. Et là encore, on réordonne :

modèle	variables	p	$\log\mathcal{L}$	$AIC/2$
(13)	constante, x_2, x_5	3	2.90	0.10
(25)	constante, x_2, x_4, x_5	4	3.45	0.55
(24)	constante, x_2, x_3, x_5	4	3.40	0.60
(19)	constante, x_1, x_2, x_5	4	3.30	0.70

Comme le modèle avec les deux variables est meilleur que tous les autres, la procédure s'arrête : on a notre modèle. Et il a un AIC de 2 fois 0.10, soit 0.20

On a donc la réponse B. Pour information, dans l'examen initial, la procédure de choix était plus compliquée.



20 On a obtenu la sortie de régression suivante

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

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

A) moins de -0.1

B) entre -0.1 et -0.05

C) entre -0.05 et 0

D) entre 0 et 0.05

E) plus de 0.05

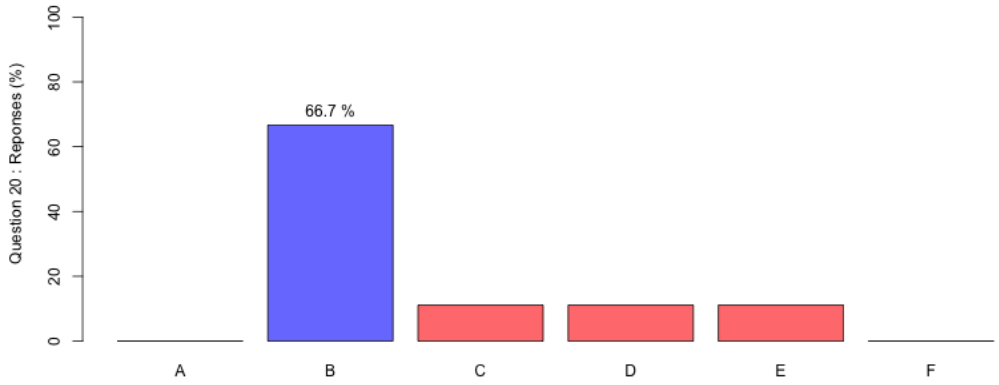
On l'a déjà fait, rappelons que sous hypothèse de normalité des résidus, ou asymptotiquement, les estimateurs sont Gaussiens. On est ici dans le second cas : l'intervalle de confiance est ici

$$\left[\widehat{\beta}_2 \pm q_{95\%}\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_2]}\right]$$

où $q_{95\%}$ est le quantile de niveau 97.5% d'une normale centrée réduite, soit 1.96. Aussi, on a

$$[-0.0540592, 0.2194392]$$

On a donc la réponse B. En fait, on *ne pouvait pas* avoir D ou E : le coefficient n'est pas significatif, pour un seuil de 5%, donc 0 appartient à l'intervalle de confiance de β_2. Comme l'estimateur est positif, c'est que la borne inférieure est négative.



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

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

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

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

Quelles affirmations sont vraies ?

A) (1) seulement
B) (2) seulement
C) (1) et (2)
D) (1), (2) et (3)
E) aucune des affirmations

On l'avait vu en cours ! Pour la première partie, $\beta_0 + \beta_1 x_i + \beta_2 x_i^2$, pas de soucis, c'est infiniment dérivable, et continu. Le soucis est donc le dernier terme, $(x-\xi)_+^2$. J'avais rapidement dit en cours que cette fonction était continue, de dérivée continue, et j'avais dit que si on voulait avoir la derivé seconde continue aussi, il fallait un exposant 3. On peut vérifier. Si on regarde à gauche de $x \mapsto (x-\xi)_+^2$ on a une fonction nulle (de dérivées toutes nulles). Notons $g(x) = (x-\xi)_+^2$, et $g(0^+)$ la valeur à droite de 0 et $g(0^-)$ la valeur à gauche :

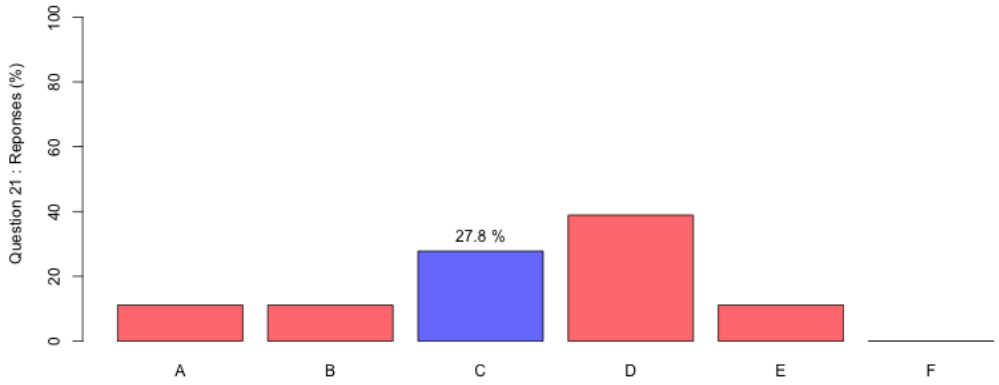
– $g(\zeta^-) = 0$ et $g(\zeta^+) = (\zeta^+ - \zeta)^2 \to 0$, donc g est continue en ζ

– $g'(\zeta^-) = 0$ et $g'(\zeta^+) = 2(\zeta^+ - \zeta) \to 0$, donc g' est continue en ζ

– $g''(\zeta^-) = 0$ et $g(\zeta^+) = 2 \nrightarrow 0$, donc g'' n'est pas continue en ζ

donc (1) et (2) sont justes : c'est la réponse C.

Pour information, cétait une question de l'examen MAS-I d'octobre 2018.



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

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

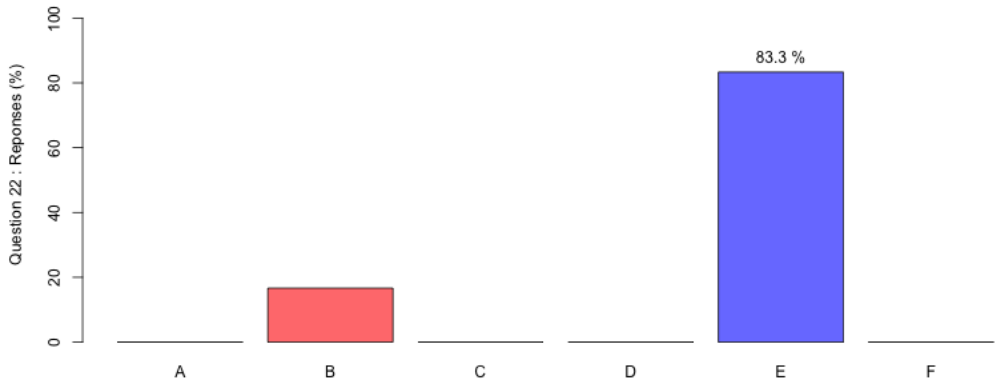
B) $\varepsilon_i = 0$ pour $i = 1, \cdots, n$, et les variables X_1 et X_2 suivent une loi normale

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

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

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

A et B sont fausses : les résidus ne sont pas nuls ! Ce qui implique d'ailleurs que D est aussi fausse : $\text{Var}[\varepsilon] = 0$ signifie que les résidus sont supposés constants, et comme $\mathbb{E}[\varepsilon] = 0$, c'est qu'ils sont nuls ! Quant à l'affirmation C, elle dit que les varaibles explicatives sont centrées. Ce qui n'est jamais une hypothèse dans la méthode des moindres carrés. Autrement dit, toutes les affirmations sont fausses. C'est la réponse E.



23 On considère le modèle $y = \beta_0 + \beta_1 x + \varepsilon$, [il y avait une coquille dans l'énoncé] estimé par moindres carrés sur un échantillon de $n = 6$ observations. On sait que

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

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

A) moins de 5.1

B) entre 5.1 et 5.3

C) entre 5.3 et 5.5

D) entre 5.5 et 5.7

E) plus que 5.7

Question de l'examen ST de l'automne 2014. Pour rappel, l'intervalle de confiance à 95% s'écrit

$$\left[\widehat{\beta}_1 \pm q_{95\%}\sqrt{\widehat{\mathsf{Var}}[\widehat{\beta}_1]}\right]$$

où, comme on a $n = 6$ et une variable explicative, $q_{95\%}$ est le quantile de niveau 97.5% d'une loi de Student à 6-2 degrés de liberté, soit 2.7764. Et rappelons que

$$\widehat{\mathsf{Var}}[\widehat{\beta}_1] = \frac{1}{6-2}\frac{\sum_{i=1}^{6}\widehat{\varepsilon}_i^2}{\sum_{i=1}^{6}(x_i - \bar{x})^2}$$

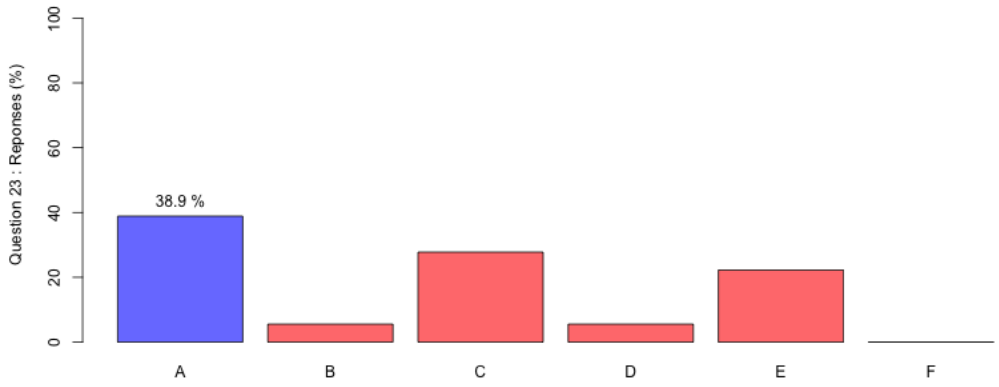
La première somme mentionnée est $S_{x^2} = 50$ et on nous dit aussi que $S_{y^2} - \widehat{\beta}_1 S_{xy} = 25$, or on sait que $S_{xy} = \widehat{\beta}_1 S_{x^2}$, donc on reconnaît ici $S_{y^2} - \widehat{\beta}_1^2 S_{x^2} = 25$, qui correspond à la somme des carrés des résidus !

$$\widehat{\mathsf{Var}}[\widehat{\beta}_1] = \frac{1}{6-2}\frac{25}{50} = \frac{1}{8}$$

donc en substituant, on obtient pour l'intervalle de confiance

$$\left[4 \pm \frac{2.7764}{\sqrt{8}}\right] = [3.018378, 4.981622]$$

donc la borne supérieure est inférience à 5 : c'est la réponse A.



24 On considère deux modèles

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

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

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

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

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

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

A) moins de 1.7

B) entre 1.7 et 1.8

C) entre 1.8 et 1.9

D) entre 1.9 et 2

E) plus que 2

C'était une question de l'examen ST d'avril 2015. On a déjà (re)parlé longuement du test de Fisher (question 13). Pour reprendre nos notations

$$F = \frac{\left(\frac{\mathsf{SCR}_0 - \mathsf{SCR}_1}{p_1 - p_0}\right)}{\left(\frac{\mathsf{SCR}_1}{n - p_1}\right)} = \frac{n - p_1}{p_1 - p_0} \cdot \frac{\mathsf{SCR}_0 - \mathsf{SCR}_1}{\mathsf{SCR}_1}$$

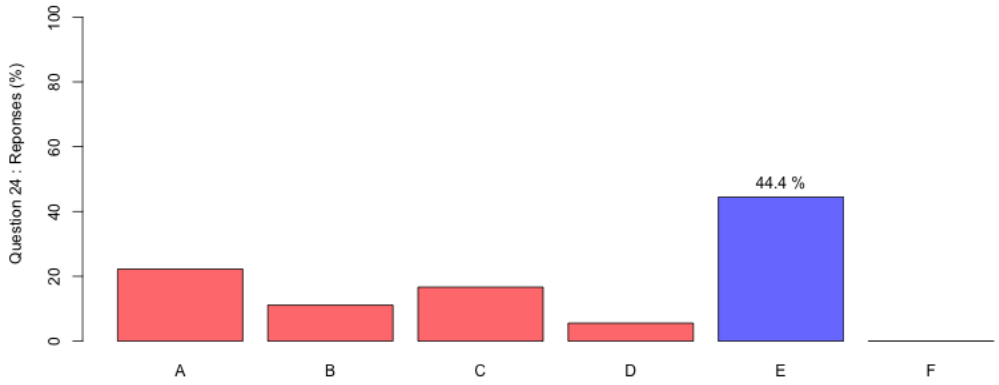
La somme des carrés des résidus est ici le premier terme,

$$\text{SCR}_0 = \sum_{i=1}^{20}(y_i - \widehat{y}_i)^2 = 13.47 \text{ et } \text{SCR}_1 = \sum_{i=1}^{20}(y_i - \widehat{y}_i)^2 = 10.52$$

avec $n = 20$, $p_0 = 3$ et $p_1 = 5$. On peut alors se lancer dans le calcul :

$$F = \frac{n - p_1}{p_1 - p_0} \cdot \frac{\text{SCR}_0 - \text{SCR}_1}{\text{SCR}_1} = \frac{20 - 5}{5 - 3} \cdot \frac{13.47 - 10.52}{10.52} = 2.103137$$

On a donc la réponse E.



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

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

A) moins de 620

B) entre 620 et 650

C) entre 650 et 700

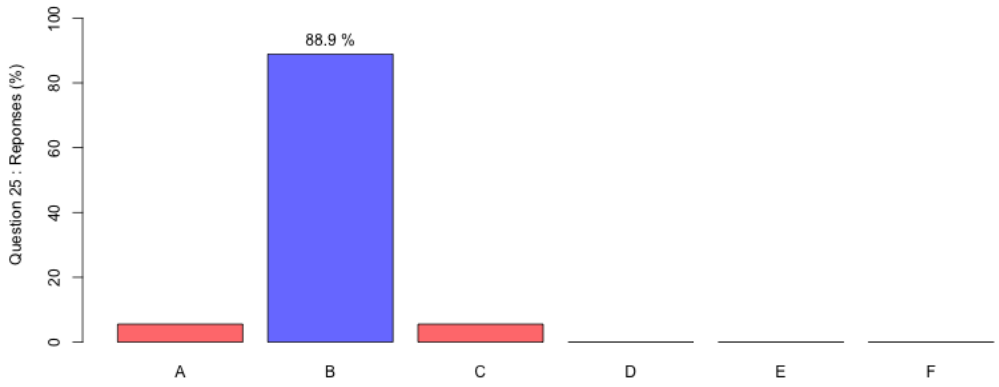
D) entre 700 et 730

E) plus de 730

Il s'agit d'une '*sample question*' de l'examen SRM. Je vais aller rapidement, on reprend la définition du R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{25}\widehat{\varepsilon}_i^2}{\sum_{i=1}^{25}(y_i - \overline{y})^2} \text{ soit } 0.64 = 1 - \frac{230}{\sum_{i=1}^{25}(y_i - \overline{y})^2}$$

donc $\sum_{i=1}^{25}(y_i - \overline{y})^2 = 230/(1 - 0.64) \sim 638.889$. On a donc la réponse B.



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

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

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

A) 100

B) 105

C) 110

D) 115

E) 120

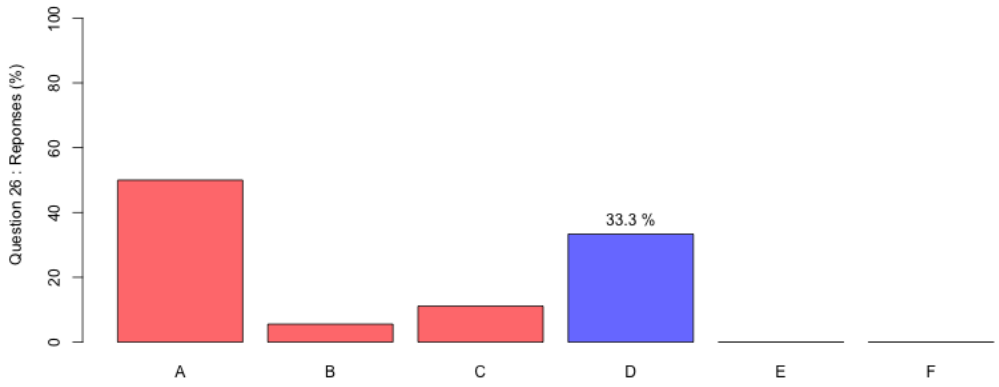
Question de l'examen SOA 'Course 120' d'octobre 1991. On cherche ici $\widehat{\text{Var}}[Y|X = x_\star]$ (où $x_\star$ vaut ici 10). On utilise ici

$$\widehat{\text{Var}}[Y|X = x_\star] = \widehat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x_\star - \overline{x})^2}{S_{x^2}}\right] \text{ où } S_{x^2} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2 = 750 - 10 \cdot 5^2 = 500,$$

aussi

$$\widehat{\text{Var}}[Y|X = x_\star] = 100 \cdot \left[1 + \frac{1}{10} + \frac{(10-5)^2}{500} = 100 + 10 + 5 = 115\right]$$

cf dans les notes de cours pour le détail du calcul. On a donc la réponse D.



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

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

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

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

A) moins de 0.3

B) entre 0.3 et 0.5

C) entre 0.5 et 0.7

D) entre 0.7 et 1

E) plus de 1

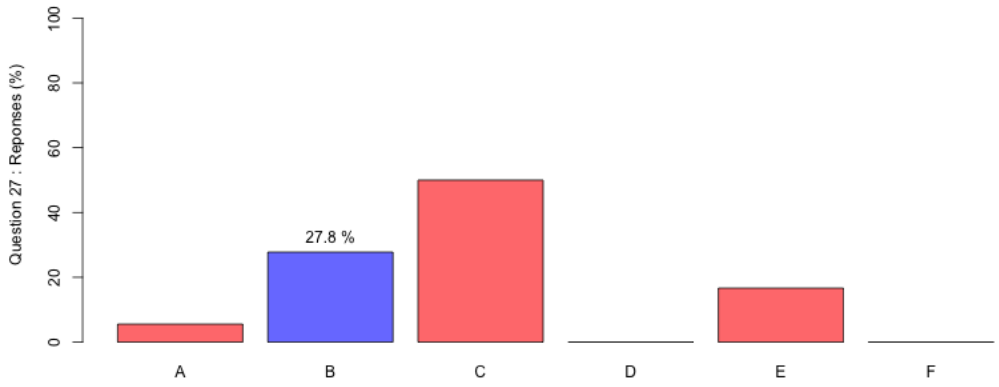
Question de l'examen SOA 'Course 120' d'octobre 1986. On avait une question proche au début de l'examen. On part de la formule dans le cas d'une régression simple

$$\widehat{\beta}_1 = \frac{S_{xy}}{S_{x^2}} \text{ et } \widehat{\alpha}_1 = \frac{S_{xy}}{S_{y^2}} = \widehat{\beta}_1 \cdot \frac{S_{x^2}}{S_{y^2}}$$

et si on remplace par les grandeurs numériques dont on dispose

$$\widehat{\alpha}_1 = 1.5 \cdot \frac{2^2}{4^2} = \frac{1.5}{4} = 0.375$$

On a donc la réponse B.



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

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

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

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

Quelles affirmations sont justes ?

A) (1) et (2) seulement

B) (1) et (3) seulement

C) (2) et (3) seulement

D) les trois affirmations

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

Question adeptée de l'examen SOA 'Course 120' de mai 1986. Pour être plus précis, les affirmations étaient originellement

(1) The variance of $\widehat{\beta}_0$ is a function of the number of observations.

(2) The variance of $\widehat{\beta}_1$ can be reduced by using a wider range of the explanatory variable.

(3) The variance associated with the predicted values of individual observations is smallest for values of the explanatory variable closest to the mean.

(1) est vraie. La réponse complète serait un peu longue, mais on a une relation du genre

$$\mathsf{Var}[\widehat{\beta}_0] = \mathsf{Var}[\widehat{\beta}_1]\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sqrt{\frac{1}{n(n-2)}\left(\sum_{i=1}^{n}\widehat{\varepsilon}_j^2\right)\frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

donc en gros, on a une variance qui décroit en $1/n$, comme classiquement en statistique.

L'affirmation (3) est vraie - et était dans l'examen de la session d'automne dernier (question 27). En effet, la variance est minimale précisément en $\overline{x}$. On sait que la variance de $\widehat{Y}^\star$ est

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

On cherche alors le minimum de la fonction

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

que l'on peut simplement dériver et écrire la condition du premier ordre

$$2\text{cov}[\widehat{\beta}_0, \widehat{\beta}_1] + 2\text{var}[\widehat{\beta}_1]x^\star = 0,$$

autrement dit, le minimum est obtenu pour

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

Or on sait que

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

donc

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

Autrement dit (3) est vraie.

L'affirmation (2) était différente de l'énoncé intial, qui me paraissait peu claire. L'affirmation (2) telle que je l'avais réécrite est fausse. La démonstration est compliqée mais on peut comparer quatre cas

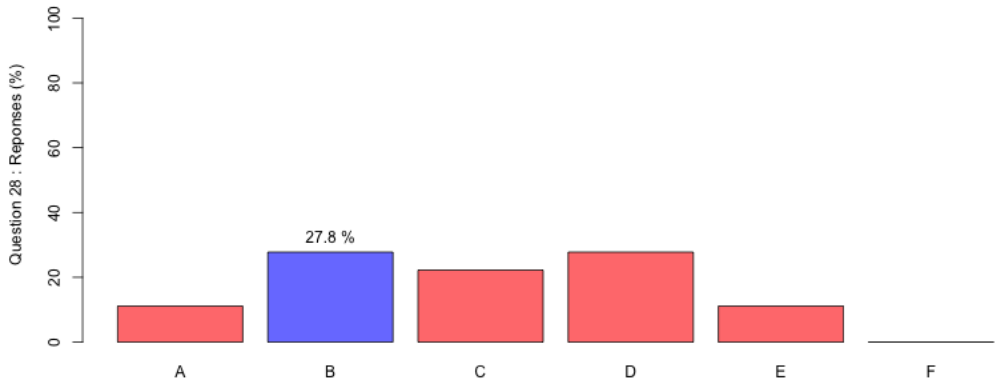
- on rajoute une variable explicative non corrélée avec la première, mais corrélé avec Y
- on rajoute une variable explicative très corrélée avec la première, et corrélé avec Y
- on rajoute une variable explicative non corrélée avec la première, et non corrélé avec Y
- on rajoute une variable explicative très corrélée avec la première, mais non corrélé avec Y

En fait, les deux derniers peut-être vu comme équivalent : rajouter une variable non-corrélée avec Y, c'est rajouter du bruit. Dans le meilleur des cas, il sera aussi orthogonal à X, ce qui ne changera pas la variance de $\widehat{\beta}_1$, mais sinon, ca augmentera sa variance.

- on rajoute une variable explicative non corrélée avec X, mais corrélé avec Y : $\text{Var}[\widehat{\beta}_1] < \text{Var}[\widehat{\beta}_1']$
- on rajoute une variable explicative très corrélée avec X, et corrélé avec Y : $\text{Var}[\widehat{\beta}_1] > \text{Var}[\widehat{\beta}_1']$
- on rajoute une variable explicative non corrélée avec X, et non corrélé avec Y $\text{Var}[\widehat{\beta}_1] \sim \text{Var}[\widehat{\beta}_1']$
- on rajoute une variable explicative très corrélée avec X, mais non corrélé avec Y : $\text{Var}[\widehat{\beta}_1] < \text{Var}[\widehat{\beta}_1']$

Autrement dit (2) est fausse.

On a donc la réponse B, car seulement (1). et (3) sont justes. L'affirmation (2) étant compliquée, j'ai donné 1/2 points pour ceux qui l'avait vue vraie (ainsi que les deux autres) et donc ont répondu D.



29 On dispose des observations suivantes

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

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

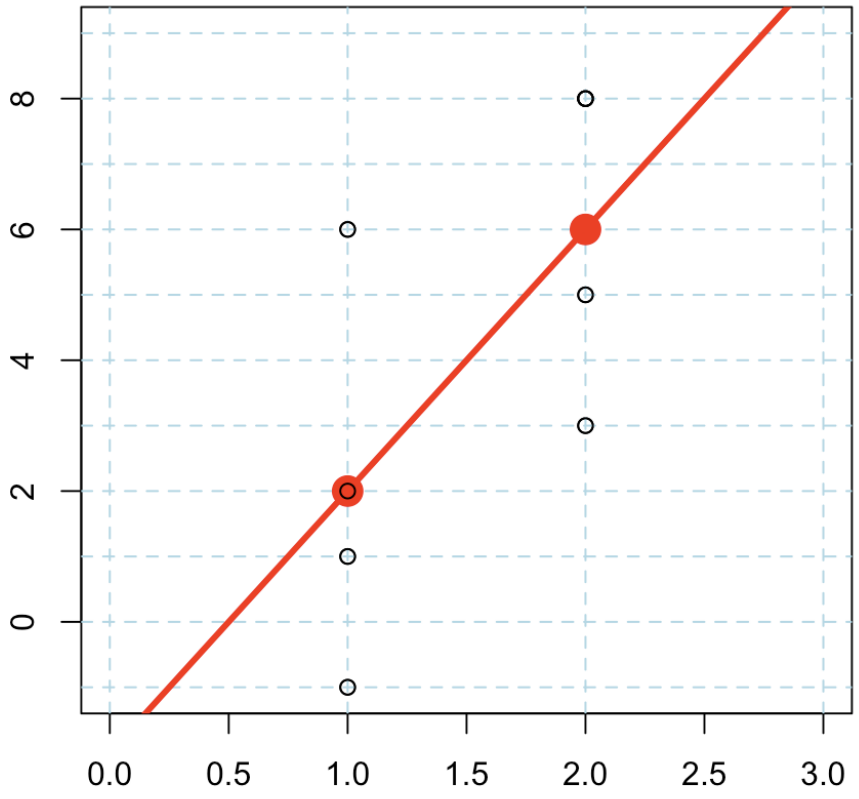
Pour la réponse rapide, notons que l'on a ici deux valeurs possibles pour x : 1 et 2. On peut imaginer avec deux facteurs. La prévision pour $x = a$ est alors la moyenne des y_i tels que $x_i = a$. Or

$$\overline{y}_{x=1} = \frac{1}{\#\{i : x_i = 1\}} \sum_{i:x_i=1} y_i \text{ et } \overline{y}_{x=2} = \frac{1}{\#\{i : x_i = 2\}} \sum_{i:x_i=2} y_i$$

soit ici

$$\overline{y}_{x=1} = \frac{(-1)+1+2+6}{4} = \frac{8}{4} = 2 \text{ et } \overline{y}_{x=2} = \frac{8+3+8+5}{4} = \frac{24}{4} = 6$$

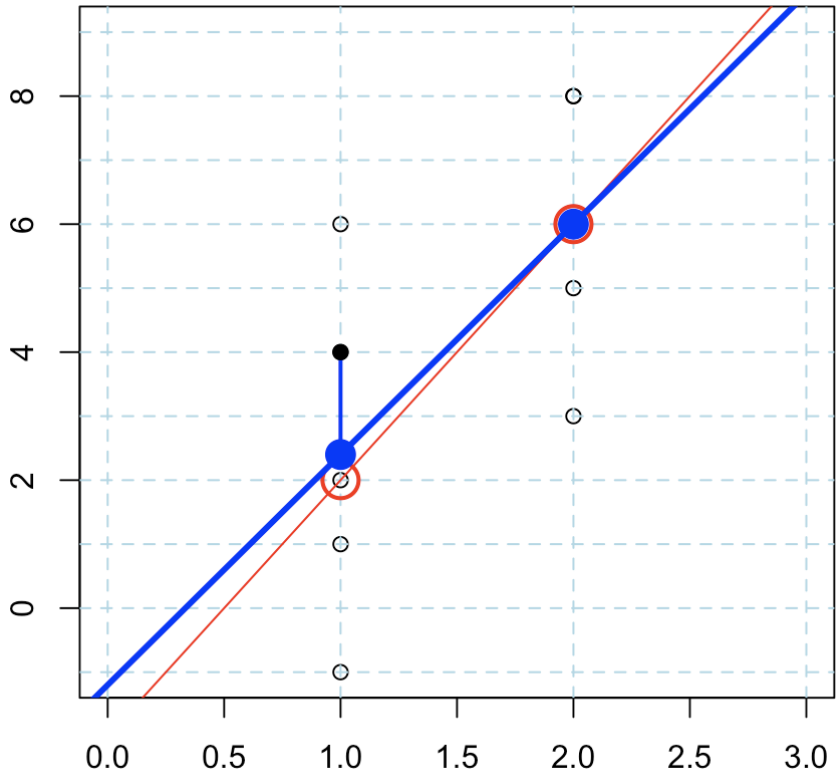
La droite de régression passe alors par $Z_1 = (1, 2)$ et $Z_2 = (2, 6)$. On reconnaît la droite d'équation $y = -2 + 4x$. C'est la droite sur la figure ci-dessous,



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

A) $\widehat{\beta}'_0 = \widehat{\beta}_0$ et $\widehat{\beta}'_1 = \widehat{\beta}_1$
B) $\widehat{\beta}'_0 > \widehat{\beta}_0$ et $\widehat{\beta}'_1 > \widehat{\beta}_1$
C) $\widehat{\beta}'_0 > \widehat{\beta}_0$ et $\widehat{\beta}'_1 < \widehat{\beta}_1$
D) $\widehat{\beta}'_0 < \widehat{\beta}_0$ et $\widehat{\beta}'_1 > \widehat{\beta}_1$
E) $\widehat{\beta}'_0 < \widehat{\beta}_0$ et $\widehat{\beta}'_1 < \widehat{\beta}_1$

C'est le dessin de la Figure ci-dessous,



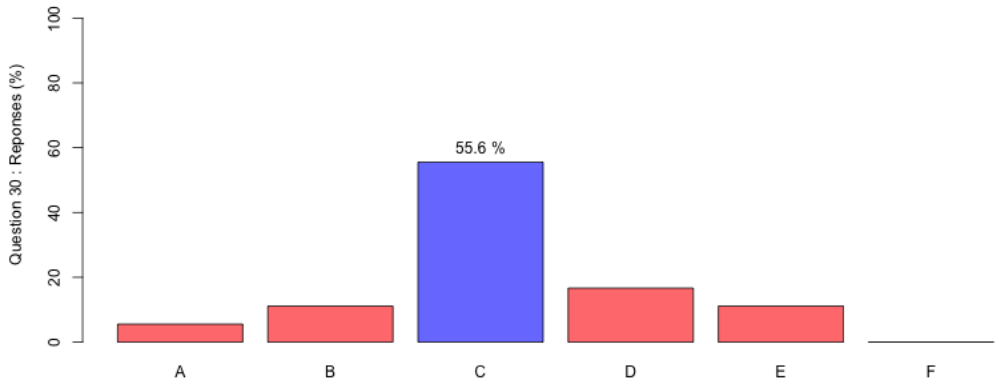
On a une nouvelle obervation, située sur la gauche (plus précisément la gauche de $\overline{x}$). La prédiction aurait été de 2 avec l'ancien modèle, et on est *au-dessus* (on a ici 4). La droite de régression aura alors tendance à pivoter dans le sens des aiguilles d'une montre (autour du nouveau barycentre $(\overline{x}, \overline{y})$). Ce pivotement a alors deux conséquences :

- la pente diminue, donc $\widehat{\beta}_1' < \widehat{\beta}_1$
- la constante est plus élevée, $\widehat{\beta}_0' > \widehat{\beta}_0$

On a donc la réponse C. Pour ceux qui veulent des chiffres, on a ici

- avec 8 observations, $y = -2 + 4x$, soit $\widehat{\beta}_0 = -2$ et $\widehat{\beta}_1 = 4$
- avec 8 observations, $y = -1.2 + 3.6x$, soit $\widehat{\beta}_0' = -1.2$ et $\widehat{\beta}_1' = 3.6$

(il existe des formules de mise à jour, mais nous parlerons pas dans le cours).


Question 30 : Reponses (%)
0
20
40
60
80
100
55.6 %
A
B
C
D
E
F