EXAMEN INTRA, ACT6420

ARTHUR CHARPENTIER

Les calculatrices sont autorisées. Les documents sont en revanche interdits.

Dans les feuilles qui suivent, il y a

- 15 questions générales sur les modèles de régression sur donnés individuelles
- 15 questions portant sur la modélisation de la note obtenue le semestre passé aux choix multiples (sorties en annexes)

Pour chaque question, quatre réponses sont proposées, une seule est valide, et vous ne devez en retenir qu'une (au maximum),

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

Aucune justification n'est demandée. Votre note finale est le total des points (sur 30).

Formulaire

Pour la loi normale, centrée et réduite ou une loi de Student, on utilisera 2 comme valeur du quantile à 97.5%, et 1.65 pour le quantile à 95%.

Des éléments de corrections sont rajoutés en bleu, sous la questions, avec les statistiques des réponses.

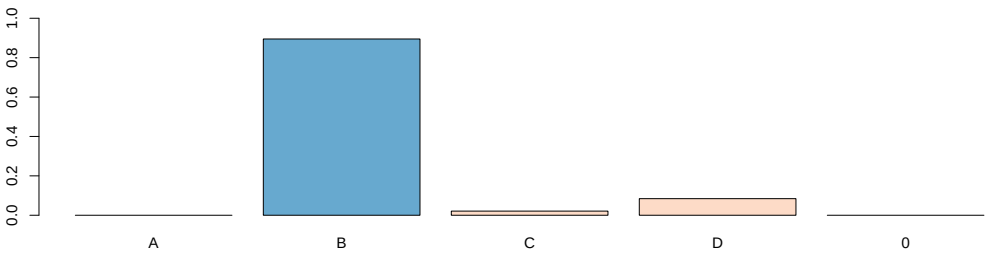
1. Régressions sur données individuelles, compréhension du cours

Considérons pour commencer un modèle de régression de la forme $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ où Y et X sont deux variables continues. On dipose de n triplets d'observations (X_i, Y_i, Z_i) indépendantes.

Question 1. Le coefficient β_0 dans la régression représente

A. la variation de Y si X augmente d'une unité
B. la valeur moyenne de Y lorsque X est nul,
C. la valeur moyenne de Y lorsque X vaut $\overline{X}$
D. la qualité de l'ajustement par une droite

Comme $Y = \beta_0 + \beta_1 X + \varepsilon$, $\mathbb{E}(Y|X=0) = \beta_0 + \mathbb{E}(\varepsilon|X=0) = \beta_0$ car le bruit ε est suppsoé être la partie non expliquée, i.e. $\mathbb{E}(\varepsilon|X=0) = 0$. C'est donc la rponse B. Je passe ici le fait que les autres sont fausses.



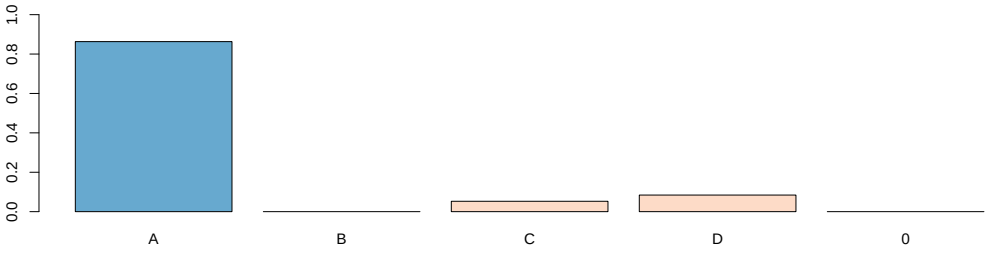
Question 2. Le coefficient β_1 dans la régression représente

A. la variation de Y si X augmente d'une unité
B. la valeur moyenne de Y lorsque X est nul,
C. la valeur moyenne de Y lorsque X vaut $\overline{X}$
D. la qualité de l'ajustement par une droite

Comme $Y = \beta_0 + \beta_1 X + \varepsilon$, $\mathbb{E}(Y|X=x) = \beta_0 + \beta_1 x$ alors que $\mathbb{E}(Y|X=x+1) = \beta_0 + \beta_1(x+1)$. Aussi,

$$\mathbb{E}(Y|X=x+1) - \mathbb{E}(Y|X=x) = \beta_1$$

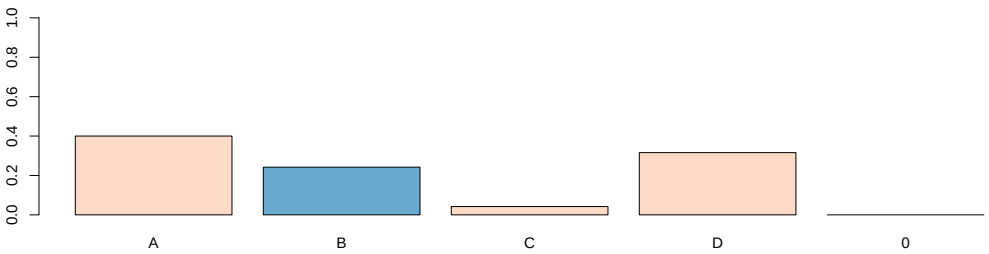
en faisant la différence. Aussi, β_1 est la variation (moyenne) de Y si X augmente de 1. On retient la réponse A.



Question 3. Dans un document trouvé sur internet, plusieurs notations sont utilisés, pour décrire le modèle. Parmi celles-ci laquelle est valide ?

A. $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \varepsilon_i$
B. $Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\varepsilon}_i$
C. $\widehat{Y}_i = \beta_0 + \beta_1 X_i + \widehat{\varepsilon}_i$
D. $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\varepsilon}_i$

Les trois notations valides sont les suivantes: $Y = \beta_0 + \beta_1 X + \varepsilon$ (c'est le vrai modèle, mais à la fois les coefficients β_0 et β_1 et le bruit ε sont non-observables), $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X$ (où $\widehat{Y}$ est la prédiction, qui est complètement déterminée), et $Y = \widehat{\beta}_0 + \widehat{\beta}_1 X + \widehat{\varepsilon}$ (une fois estimé le modèle, on construit les erreurs, induites par le modèle). Bref, la bonne réponse est B. (dans A. il n'y a plus de bruit, dans C la pédiction est construite avec $\widehat{\beta}_0$ et $\widehat{\beta}_1$, et dans D il n'y a pas de terme d'erreur, comme dans dans A car on parle de $\widehat{Y}$).



Question 4. On décide de simplifier le modèle, et de considérer le modèle suivant $Y_i = \alpha_0 + u_i$. L'estimateur par moindres carrés de α_0 est

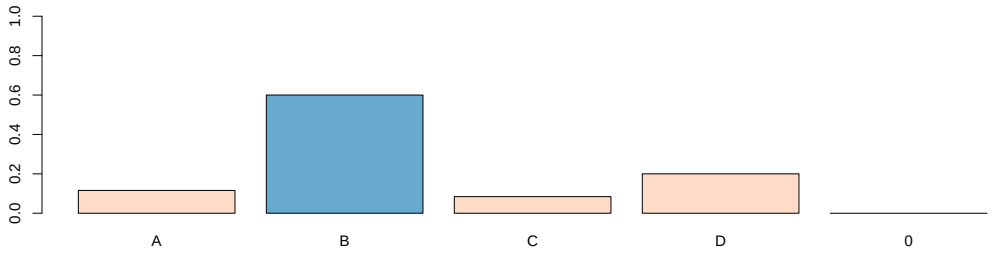
A. $\widehat{\alpha}_0 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$

B. $\widehat{\alpha}_0 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} Y_i$

C. $\widehat{\alpha}_0 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n} X_i$

D. $\widehat{\alpha}_0 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n} Y_i$

$Y = \alpha_0 + u_i$, où on suppose le bruit centré. Donc α_0 est l'espérance de Y. Aussi, $\widehat{\alpha_0} = \overline{Y}_n$, c'est donc la réponse B. Je laisse le calcul à faire en guise d'exercice.



Question 5. On décide de simplifier le modèle, mais cette fois en considérant le modèle suivant $Y_i = \gamma_1 X_i + v_i$. L'estimateur par moindres carrés de γ_1 est

A. $\widehat{\gamma}_1 = \dfrac{\overline{Y}}{\overline{X}}$

B. $\widehat{\gamma}_1 = \dfrac{1}{\overline{X}}$

C. $\widehat{\gamma}_1 = \dfrac{\displaystyle\sum_{i=1}^{n} X_i Y_i}{\displaystyle\sum_{i=1}^{n} X_i^2}$

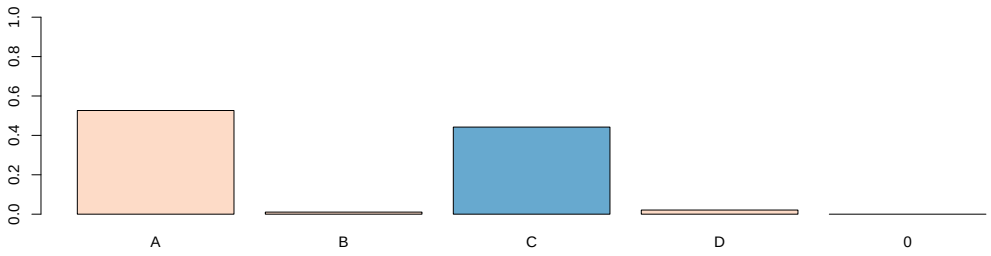
D. $\widehat{\gamma}_1 = 1$

Cette fois, on enlève la constante, $Y = \gamma_1 X + v$. on veut minimiser la fonction

$$\gamma \mapsto \sum_{i=1}^{n}(Y_i - \gamma_1 X_i)^2 = \sum_{i=1}^{n} Y_i^2 - 2\gamma X_i Y_i + \gamma^2 X_i^2$$

si on écrit la condition du premier ordre, on va dériver la fonction précédante, et chercher une valeur pour laquelle la dérivée s'annule, i.e.

$$2\sum_{i=1}^{n} X_i Y_i = 2\gamma \sum_{i=1}^{n} X_i \text{ i.e. } \gamma = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$$



La question était vicieuse, car tout le monde a retenu que la droite de régression passe par le centre de gravité du nuage, i.e. par le point $(\overline{X}, \overline{Y})$... à condition d'avoir une constante dans la régression. Sans constante, cette règle ne marche plus...

Question 6. Dans le modèle initial $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, on propose d'estimer β_1 par trois estimateurs

$$\widehat{\beta}_1^{(1)} = \frac{\overline{Y}}{\overline{X}}, \widehat{\beta}_1^{(2)} = \frac{Y_2 - Y_1}{X_2 - X_1} \text{ et } \widehat{\beta}_1^{(3)} = \frac{\max\{Y_i\} - \min\{Y_i\}}{\max\{X_i\} - \min\{X_i\}}$$

A. $\widehat{\beta}_1^{(1)}$ est un estimateur sans biais de β_1
B. $\widehat{\beta}_1^{(2)}$ est un estimateur sans biais de β_1
C. $\widehat{\beta}_1^{(3)}$ est un estimateur sans biais de β_1
D. ces trois estimateurs sont des estimateurs sans biais de β_1

Je ne vais pas montrer les propriétés de ces estimateurs... La bonne réponse était B. En fait il fallait noter que

$$\widehat{\beta}_1 = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{[\beta_0 + \beta_1 X_2 + \varepsilon_2] - [\beta_0 + \beta_1 X_1 + \varepsilon_1]}{X_2 - X_1}$$

i.e.

$$\widehat{\beta}_1 = \beta_1 + \frac{\varepsilon_2 - \varepsilon_1}{X_2 - X_1}$$

donc

$$\mathbb{E}(\widehat{\beta}_1) = \beta_1 + \underbrace{\frac{\mathbb{E}(\varepsilon_2 - \varepsilon_1)}{X_2 - X_1}}_{=0} = \beta_1.$$

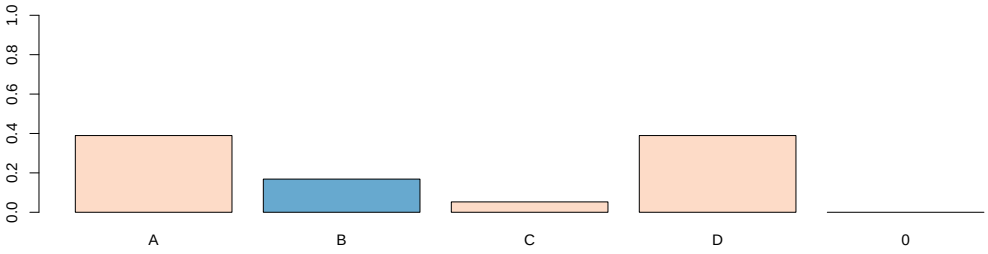
Pour les deux autres,

$$\widehat{\beta}_1 = \frac{\overline{Y}}{\overline{X}} = \frac{\beta_0 + \beta_1 \overline{X} + \overline{\varepsilon}}{\overline{X}}$$

aussi

$$\mathbb{E}(\widehat{\beta}_1) = \frac{\beta_0}{\overline{X}} + \beta_1 \neq \beta_1,$$

mais pour l'utilisation des max et des min, c'est plus compliqué à montrer. Mais comme on a vu que le premier était biaisé, D. ne pouvait être valide non plus...

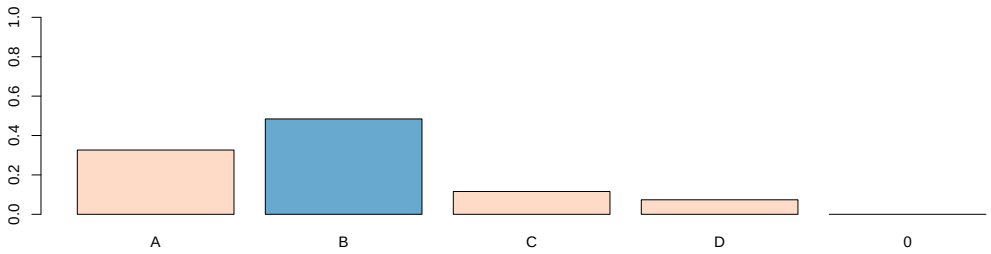


Question 7. On estime β_0 et β_1 par moindres carrés. On note les résidus obtenus. Parmi les affirmations suivantes laquelle (ou lesquelles) sont vraies ?

(i) la somme des résidus est toujours nulle,
(ii) si $R^2 = 0$ la droite de régression est horizontale
(iii) si $R^2 = 0$ la droite de régression est verticale

A. (i) seulement
B. (i) et (ii)
C. (i) et (iii)
D. (i) à condition que la variable Y soit de moyenne nulle

La condition (i) est toujours vraie car on a mis une constante dans la régression. C'est la condition du premier ordre: $\overline{Y} = \beta_0 + \beta_1 \overline{X}$. Et ce, quelle que soit la valeur de $\overline{Y}$. Si $R^2 = 0$, c'est que $\widehat{\beta}_1 = 0$, ou encore $\widehat{Y}_i = \widehat{Y}$ pour tout i, ou encore $\text{cor}(X, Y) = 0$. Bref, la droite de régression serait alors horizontale. Donc (ii) est vraie aussi. On retient alors B.



Question 8. Dans le modèle initial $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ estimé par moindres carrés, quelle relation parmi les quatre suivantes est valide

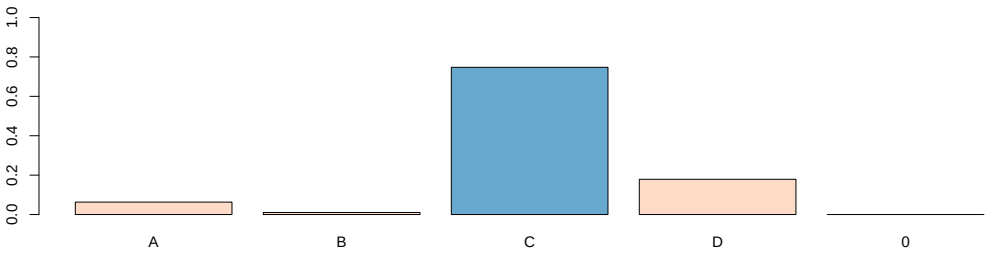
A. pour tout $i \in \{1, \cdots, n\}$ $(Y_i - \overline{Y})^2 = (Y_i - \widehat{Y}_i)^2 + (\widehat{Y}_i - \overline{Y})^2$

B. pour tout $i \in \{1, \cdots, n\}$ $(Y_i - \widehat{Y}_i)^2 = (Y_i - \overline{Y})^2 + (\widehat{Y}_i - \overline{Y})^2$

C. $\displaystyle\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 + \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2$

D. $\displaystyle\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 + \sum_{i=1}^{n} (\widehat{Y}_i - \overline{Y})^2$

On l'a vu en cours. On a que $Y_i - \overline{Y} = (Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \overline{Y})$. Mais si on passe au carré, i.e. $(a + b)^2$ on aurait $a^2 + b^2 + 2ab$. Les termes croisés ne s'annulent pas. On n'a donc ni A, ni B. En revanche, en sommant, on retrouve l'égalité de Pythagore (ou *formule de décomposition de la variance* comme on dit dans les livres savants). La bonne réponse est C.



Question 9. Dans le modèle initial $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ estimé par moindres carrés, quelle(s) relation(s) parmi les suivantes sont valides

(i) $\sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2 = 0$

(ii) $\sum_{i=1}^{n} (Y_i - \widehat{Y}_i) = 0$

(iii) $\sum_{i=1}^{n} (Y_i - \widehat{Y}_i) X_i = 0$

(iv) $\sum_{i=1}^{n} (Y_i - \widehat{Y}_i) Y_i = 0$

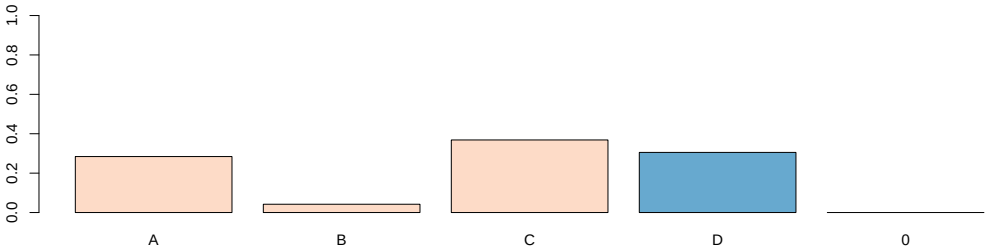
A. (i) seulement

B. (iii) et (iv)

C. (ii) seulement

D. (ii) et (iii).

La première condition signifie que $\sum \widehat{\varepsilon}_i^2$ (i.e. la variance des résidues, à un facteur n près) serait nulle. (i) est fausse. On a vu que (ii) était vérifiée, c'est la première condition du premier ordre (obtenue en écrivant que ∂ somme des carrés des résidus $/\partial\beta_0$ est nulle). La condition (iii) est la seconde condition du premier ordre (obtenue en écrivant que ∂ somme des carrés des résidus $/\partial\beta_1$ est nulle). La traduction est que $\varepsilon \perp X$, i.e. les résidus sont non-expliqués par la variable explicative. On va retenir ces deux réponses, i.e. D.



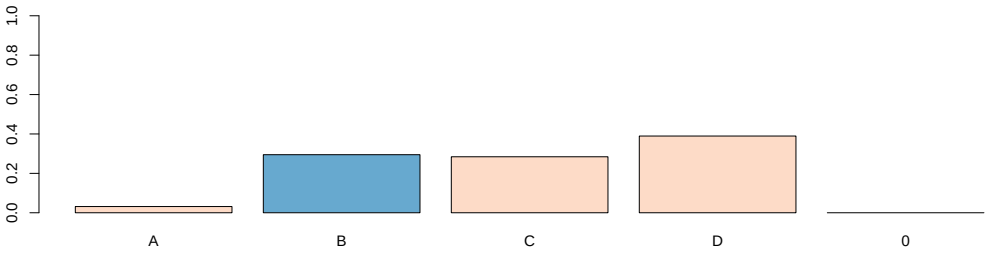
Question 10. On se rends compte que l'on a oublié de prendre en compte une variable explicative, Z. On suppose que $\sum_{i=1}^{n} X_i Z_i = \sum_{i=1}^{n} X_i \cdot \sum_{i=1}^{n} Z_i$. On construit alors le modèle

$$Y_i = \delta_0 + \delta_1 X_i + \delta_2 Z_i + \eta_i$$

On note $\widehat{\delta}_k$ les estimateurs par moindres carrés ordinaires. Que peut-on dire que $\widehat{\delta}_1$?

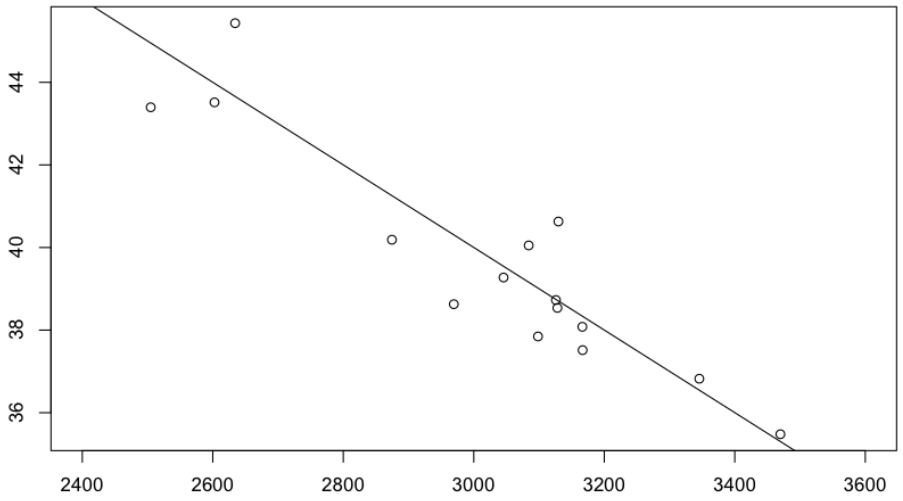
A. $\widehat{\delta}_1 > \widehat{\delta}_2$
B. $\widehat{\delta}_1 = \widehat{\beta}_1$
C. $\widehat{\delta}_1 < \widehat{\beta}_1$
D. aucune des propositions A, B ou C.

On l'a vu en cours (et cet exemple est détailé sur le blog): c'est une conséquence du thérème de Frisch-Waugh. La bonne réponse est B.



La question suivante porte sur le graphique et la sortie ci-dessous,

```
Coefficients:
```




```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.004440   3.493117  20.044 3.70e-11 ***
X            0.010005   0.001151   8.777 7.99e-07 ***

Residual standard error: 1.157 on 13 degrees of freedom
Multiple R-squared: 0.8556,Adjusted R-squared: 0.8445
F-statistic: 77.04 on 1 and 13 DF,  p-value: 7.988e-07
```

Question 11. Une nouvelle arrive: $X_{16} = 3500$ et $Y_{16} = 44$. Si on refaisait la régression (sur les 16 observations), laquelle parmi les affirmations suivantes, est vraies

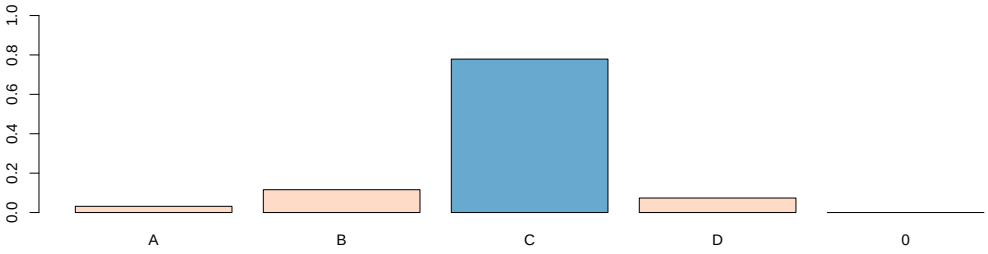
A. cette nouvelle observation n'aura aucun impact sur le R^2 car le R^2 est insensible aux nouvelles observations

B. cette nouvelle observation devrait augmenter le R^2 car il s'agit d'une valeur aberrante

C. cette nouvelle observation devrait diminuer le R^2 car il s'agit d'une valeur aberrante

D. cette nouvelle observation devrait augmenter le R^2 le point est en dessus de la droite de régression

Désolé pour l'erreur dans la sortie, il manquait deux signes '-'. On rajoute ici une observation en haut à droite (i.e. $X > \overline{X}$). Cette observation va déteriorer le modèle. L'erreur sera importante, et donc la somme des carrés des erreurs va augmenter, ce qui aura pour effet de faire baisser (a priori) le R^2. La bonne réponse est la réponse C.



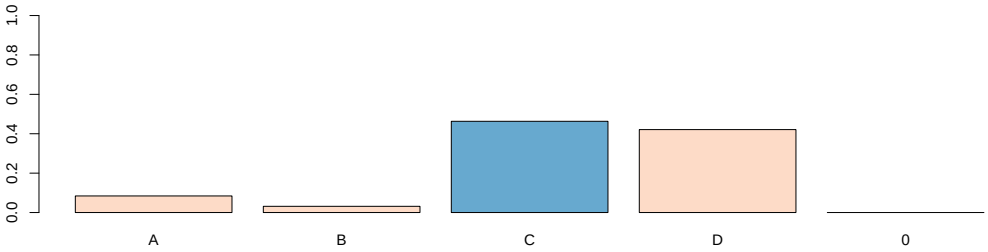
Question 12. Dans un rapport, on lit la sortie suivante

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.03070    0.10306   0.298    0.767
X           -0.08156    0.09949  -0.820    0.415

Residual standard error: 0.8845 on 73 degrees of freedom
Multiple R-squared: 0.89122,Adjusted R-squared: -0.004452
F-statistic: 0.672 on 1 and 73 DF,  p-value: 0.415
```

La sortie a été falsifiée car les chiffres ne sont pas cohérents entre eux,

A. $\widehat{\beta}_1$ devrait être plus proche de 0 car il n'est pas significatif
B. la p-value associée à β_1 devrait être négative car $\widehat{\beta}_1 < 0$
C. le R^2 devrait être proche de 0 car ni $\widehat{\beta}_0$ ni $\widehat{\beta}_1$ ne sont significativement non nuls
D. le R^2 ajusté ne peut pas être négatif

La bonne réponse est C. Une forte proportion a répondu D... et malheureusement le R^2 ajusté *peut* être négatif. Rien dans sa définition ne l'interdit



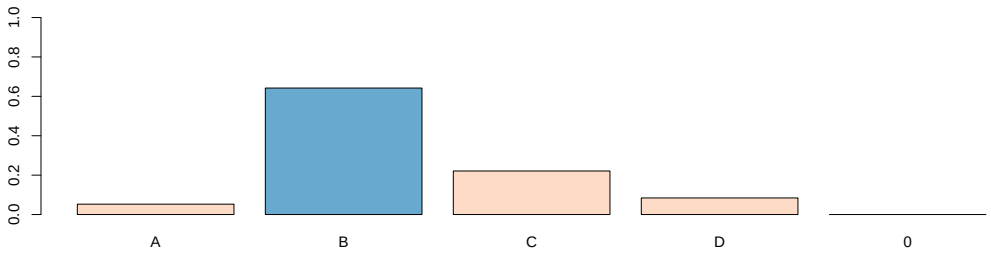
Question 13. Dans un autre rapport, on lit la sortie suivante

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.1212     0.1340  38.215  < 2e-16 ***
X            -1.2406     0.1340  -0.920  0.361

Residual standard error: 1.102 on 73 degrees of freedom
Multiple R-squared: 0.5371,Adjusted R-squared: 0.5308
F-statistic:  84.7 on 1 and 73 DF,  p-value: 7.75e-14
```

Là encore, la sortie a été à nouveau falsifié car les chiffres ne sont pas cohérents entre eux,

A. $\widehat{\beta}_1$ devrait être positif car $\widehat{\beta}_0 > 0$ et le R^2 ajusté aussi
B. le test de Student de β_1 devrait être plus élevé (et conclure à une significativité du coefficient)
C. $\widehat{\beta}_0$ et $\widehat{\beta}_1$ ne peuvent pas avoir le même écart-type car la constante est 4 fois plus grande que la pente (et l'écart-type devrait être 2 fois plus grand)
D. le R^2 et le R^2 ajusté sont trop proches

La bonne réponse est B. En fait, en faisant le calcul, on obtient une valeur proche de -9.258.



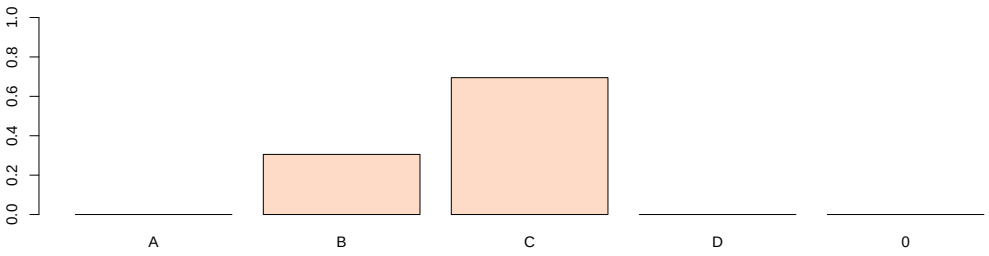
Question 14. L'estimation d'un modèle $\log(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$ donne la sortie suivante

```
lm(formula = log(Y) ~ X, data = B)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.013856   0.245255   4.134 0.000109 ***
X           0.005060   0.002266   2.233 0.029192 *

Residual standard error: 1.01 on 62 degrees of freedom
Multiple R-squared: 0.07442,Adjusted R-squared: 0.05949
F-statistic: 4.985 on 1 and 62 DF,  p-value: 0.02919
```

Quelle serait votre prédiction pour Y si X valait 100 ?

A. 10.0
B. 1.5
C. 4.6
D. 7.6

La question était vicieuse car nous n'avions pas eu le temps de revoir cette propriété de la loi lognormale en cours (j'ai noté sur 29 l'examen). Pour rappel, Si Y suit une loi normale $\mathcal{N}(\mu, \sigma^2)$, i.e. Y a pour moyenne μ, $\exp(Y)$ n'a pas pour moyenne $\exp(\mu)$ (à méditer).



Question 15. On a estimé un modèle suivant

```
lm(formula = Y ~ X)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.1349     0.1386  37.045  < 2e-16 ***
X             1.3100     0.1619   8.094 9.39e-12 ***

Residual standard error: 1.167 on 73 degrees of freedom
Multiple R-squared: 0.473,Adjusted R-squared: 0.4657
F-statistic: 65.51 on 1 and 73 DF,  p-value: 9.391e-12
```

On fait la même régression en enlevant une observation

```
lm(formula = Y[-21] ~ X[-21])

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.1734     0.1364  37.942  < 2e-16 ***
X[-21]        1.3361     0.1583   8.439 2.32e-12 ***

Residual standard error: 1.139 on 72 degrees of freedom
Multiple R-squared: 0.4973,Adjusted R-squared: 0.4903
F-statistic: 71.22 on 1 and 72 DF,  p-value: 2.318e-12
```

Où se trouvait l'observation (X_{21}, X_{21}) dans le graphique des (X_i, Y_i)

A. on ne peut pas savoir
B. proche de $(\overline{X}, \overline{Y})$
C. probablement *au dessus* de la droite de régression, dans la partie supérieure gauche
D. probablement *en dessous* de la droite de régression, dans la partie inférieure droite

Comme la majorité n'a pas donné la réponse, on va prendre un peu de temps. Car le question était difficile... et pas forcément intuitive. En dimension 1, si la moyenne diminue en rajoutant une observation, c'est que l'observation, c'est que l'observation rajoutée était plus petite que l'observation. C'est la formule de mise à jour de la moyenne. On a une formule similaire en régression. Posons

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_{n-2} \\ 1 & X_{n-1} \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_{-n} \\ \boldsymbol{X}'_n \end{bmatrix}$$

Rappelons que $\widehat{\boldsymbol{\beta}}_{-n} = [\boldsymbol{X}'_{-n}\boldsymbol{X}_{-n}]^{-1}\boldsymbol{X}'_{-n}\boldsymbol{Y}_{-n}$. Pour passer de $\widehat{\boldsymbol{\beta}}_{-n}$ à $\widehat{\boldsymbol{\beta}}$ (prenant en compte la $n$ème observation), on peut écrire une formule de mise à jour (formule 4.59 dans le Greene

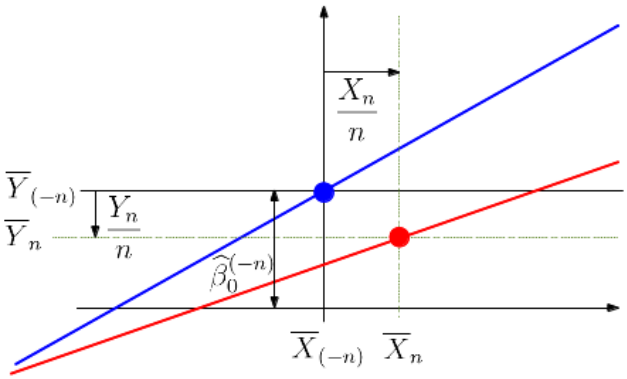
(2012))

$$\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{-n} = \frac{1}{1 + \boldsymbol{X}_n'[\boldsymbol{X}_{-n}'\boldsymbol{X}_{-n}]^{-1}\boldsymbol{X}_n}[\boldsymbol{X}_{-n}'\boldsymbol{X}_{-n}]^{-1}\boldsymbol{X}_{-n}(Y_n - \boldsymbol{X}_n'\widehat{\boldsymbol{\beta}}_{-n})$$

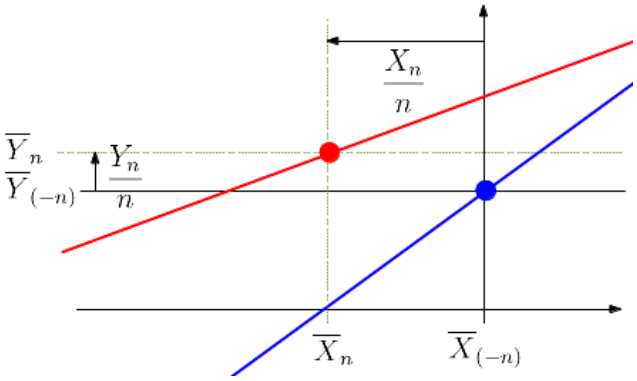
On peut utiliser le fait que l'on est en dimension 2, et on peut alors écrire (cf cours)

$$(\boldsymbol{X}_{-n}'\boldsymbol{X}_{-n})^{-1} = \frac{1}{\sum_{i=1}^{n-1} X_i^2 - n\overline{X}_{-n}^2}\begin{pmatrix} \frac{1}{n-1}\sum_{i=1}^{n-1} X_i^2 & -\overline{X}_{-n} \\ -\overline{X}_{-n} & 1 \end{pmatrix}$$

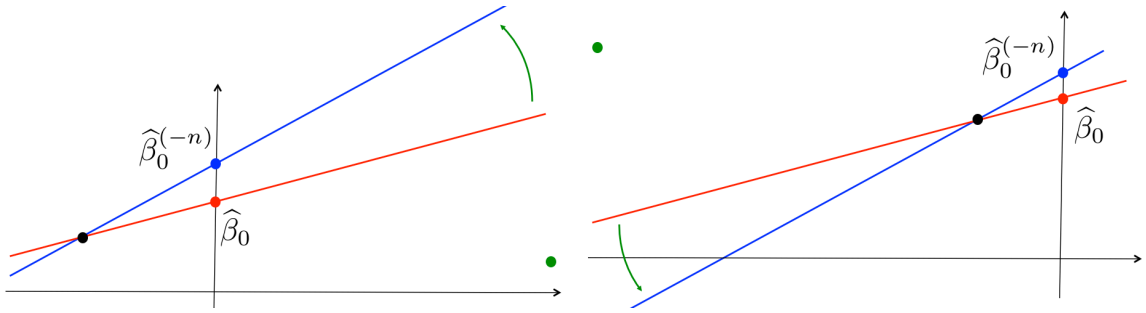
Je passe les calculs. On souhaiterait pouvoir passer du faite que les deux composantes du terme de gauche sont négatifs (i.e. $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{-n}$) pour dire quelque chose sur le signe de $(Y_n - \boldsymbol{X}_n'\widehat{\boldsymbol{\beta}}_{-n})$. Par exemple, on aurait envie de dire que l'on est sur des valeurs positives. C'est d'ailleurs ce que l'on a sur la figure suivante (centrée sur $\overline{X}_{-n}$). En rajoutant une observation en bas à droite, le barycentre se décale vers la droite, sous la première droite de régression. Ca serait donc la réponse D.



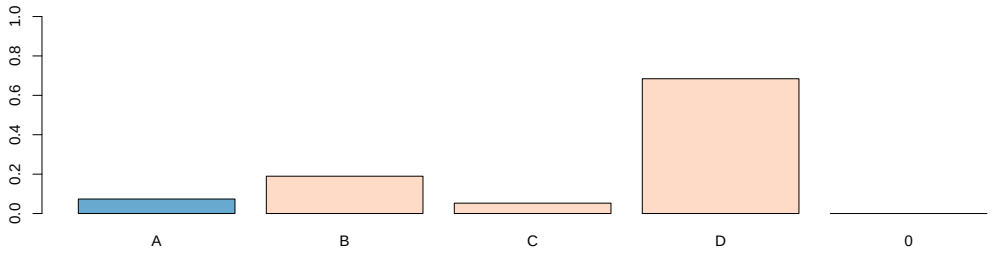
Le soucis est que l'on n'a pas pris en compte le signe de $\overline{X}_{-n}$. Regardons la figure suivante (ici centrée - pour l'instant - sur $\overline{X}_{-n}$). On note qu'en rajoutant une observation en haut, on va baisser la pente (ce qui est observé sur les sorties de régression). En supposant que $\overline{X}_{-n} = 0$, on note que la constante en revanche a augmenté (sur l'axe des ordonnéss (y), l'intersection avec la droite rouge est au dessus de l'intersection avec la droite bleue). Donc ce n'est pas tout à fait ce que l'on a sur nos sorties... Mais supposons maintenant que $\overline{X}_{-n} < 0$, et que l'axe des ordonnées se trouve beaucoup plus à droite: dans ce cas, l'intersection de la courbe bleu $\widehat{\beta}_0^{(-n)}$ sera au dessus de l'intersection de la courbe rouge $\widehat{\beta}_0$. On se retrouve dans les conditions des sorties données.



Pour une autre illustration, considérons les deux dessins ci-dessous, où dans les deux cas, on enlève le point vert (respectivement en bas à droite, et en haut à gauche). Comme les deux pentes changes, on pivote autour du point d'intersection.



Bref, tant que l'on ne sait pas où se trouve $\overline{X}_{-n}$ et X_nj, on n'a pas vraiment moyen de savoir où se trouvait le point rajouté (ou enlevé).

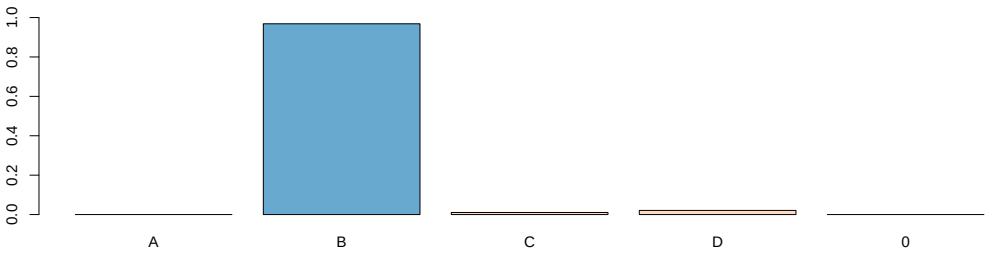


2. Régressions et prévisions: analyse de la base de données

Les questions portent sur les sorties présentées dans l'Annexe.

Question 16. Dans la sortie **R1** que vaut la statistique du test de Student associée à la constante β_0 ?

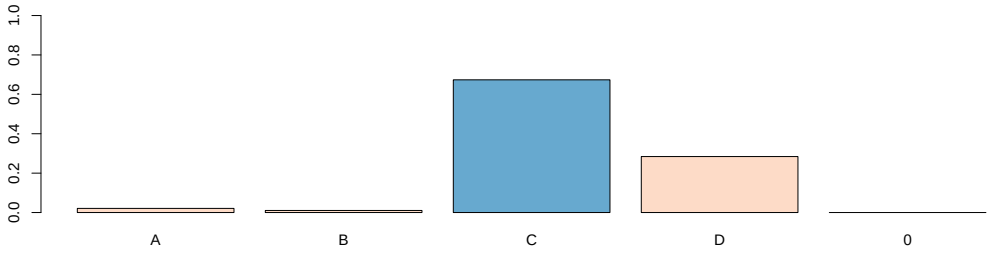
A. 0.4301914
B. 2.325
C. 4.522258
D. 0.2211285



Question 17. À l'aide de la sortie **R1** on sait que la moyenne obtenue sur l'examen MAT3080 était de 73. Quelle a été la moyenne obtenue à l'examen à choix multiples du cours ACT6420

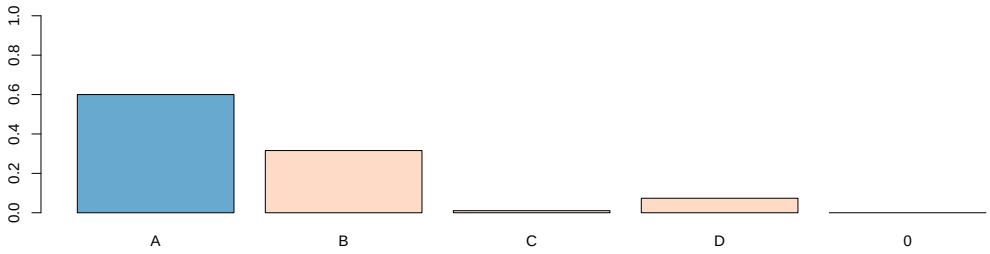
A. 73
B. 48.5
C. 55
D. on peut pas le savoir

Comme on a une constante dans la régression, on sait que la droite de régression ajustée par moindres carrés passe par le centre de gravité. Si on a la moyenne des X, on a alors la moyenne des Y.



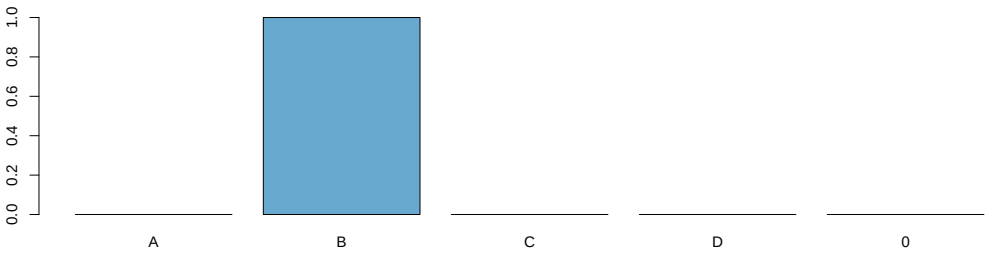
Question 18. À l'aide de la sortie **R1** quel intervalle de confiance (à 95%) pour β_1 construiriez-vous ?

A. $[0.4057 \pm 2 \cdot 0.1419]$
B. $\left[0.4057 \pm 2 \cdot \dfrac{0.1419}{\sqrt{39}}\right]$
C. $[0.4057 \pm 2 \cdot 12.67]$
D. $\left[0.4057 \pm 2 \cdot \dfrac{12.67}{\sqrt{39}}\right]$



Question 19. À l'aide de la sortie **R1** quel note prédiriez vous pour un étudiant qui a eu 80 à l'examen MAT3080 ?

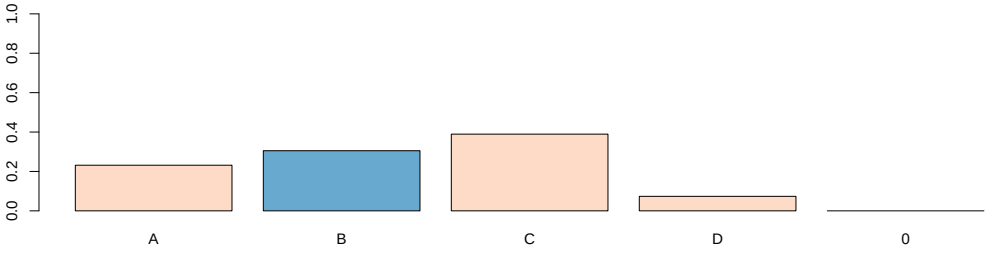
A. 24.4
B. 56.9
C. 80.0
D. 80.4



Question 20. (suite de la question 19.) Toujours pour cet étudiant, à l'aide de la sortie **R1** au lieu de donner seulement une valeur prédite, on souhaite indiquer quelle devrait être sa vraie note, avec une probabilité de 95%. Quelle devraient être les bornes (autour de la valeur prédite) ?

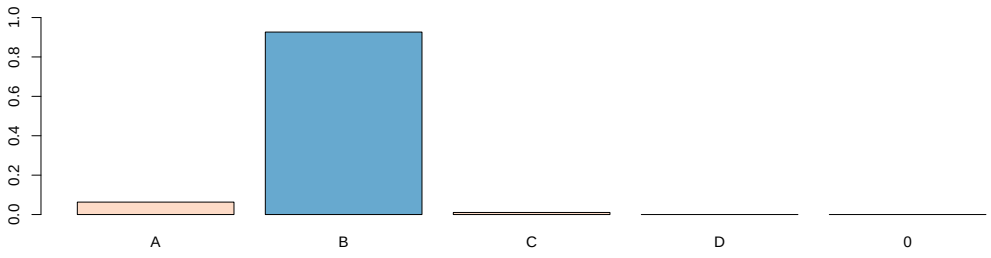
A. ± 12.7
B. ± 25.6
C. ± 4.1
D. ± 10.5

On utilise le fait que la vraie valeur est $\widehat{Y} + \widehat{\varepsilon}$. Avec une hypothèse de normalité sur les résidus, cela signifie qu'un intervalle de confiance à 95% pour Y est $\widehat{Y} \pm 1.96\widehat{\sigma}$ (en fait, en première page, on suggèrait d'utiliser 2 comme valeur du quantile de la loi $\mathcal{N}(0,1)$ pour les calculs, et pas 1.96).



Question 21. À l'aide de la sortie **R2** quelle est la note prédite par le modèle pour un élève de sexe masculin à l'aide du modèle `reg2a` ?

A. on peut pas le savoir
B. 59.3
C. 48.4
D. 10.9

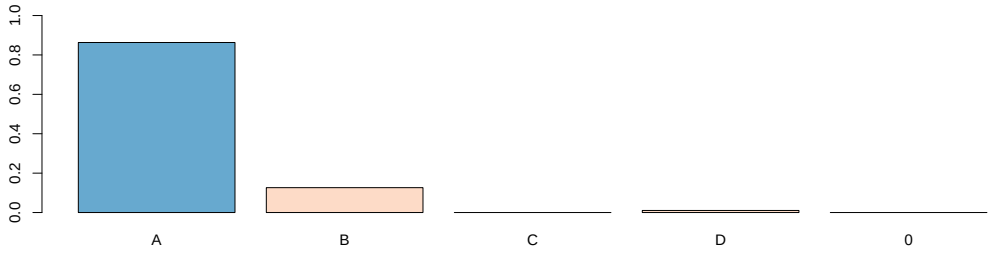


Question 22. À l'aide de la sortie **R2** quelle est l'estimation de β_{homme} dans le modèle `reg2b`?

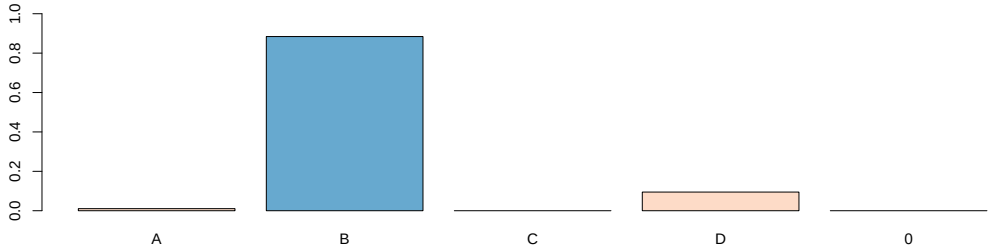
A. 59.286
B. 10.911
C. 37.464
D. 48.375

Question 23. Après avoir rendu les notes, l'éleve (41) consteste la note, car elle espérait avoir plus. Il s'agit d'une fille de 23 ans, qui avait eu 61 au cours MAT3080, et 47.5 au cours ACT6420 (au choix multiple), alors qu'elle avait prédit avoir 70. A l'aide du modèle `reg3a` de la sortie **R3** quelle note devrions nous attendre ?

A. 70.0
B. 40.9
C. 47.5



D. 54.7



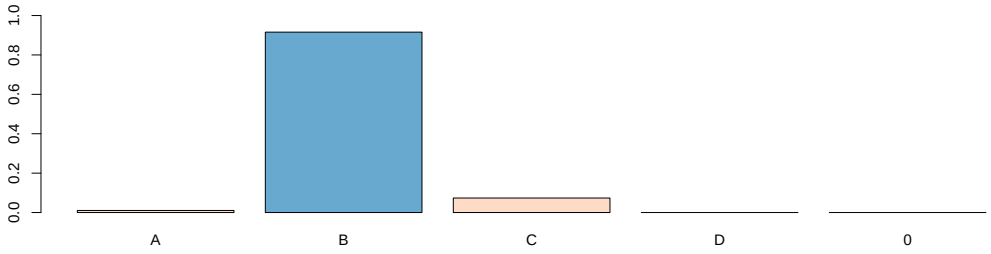
Question 24. (suite de la question 23) La constante n'étant pas significative, on décide de l'enlever. À l'aide du modèle `reg3b` de la sortie **R3** quelle note devrions nous attendre ?

A. 70.0

B. 38.5

C. 54.6

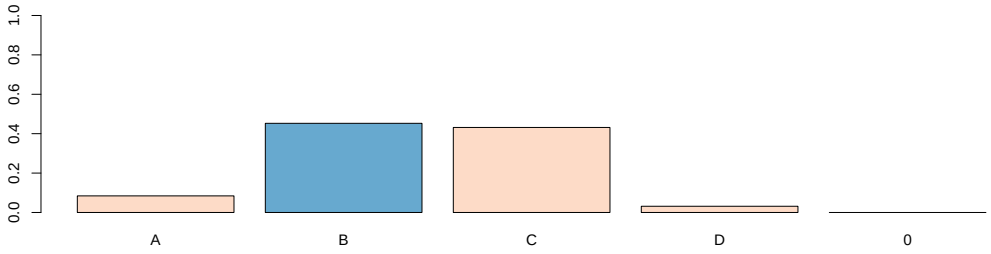
D. 47.1



Question 25. (suite des questions 23 et 24) À l'aide du modèle `reg3b`, en tenant compte de l'erreur associée au modèle linéaire, donnez un intervalle dans lequel la note devrait être avec une probabilité de 90% (sans tenir compte de l'erreur d'estimation) ?

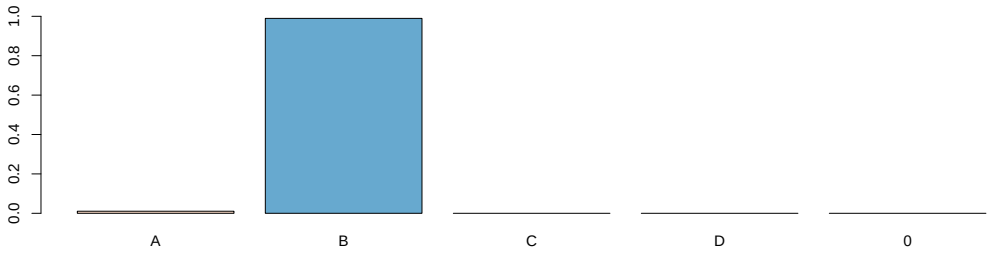
A. [52;88]
B. [21;56]
C. [36;42]
D. [67;73]

On avait vu à la question prédédante qu'on devait être centr'sur 38.5. Il faut alors rajouter $1.64\widehat{\sigma}$, soit environ 17 .5.



Question 26. Dans le modèle `reg3b`, il manque les écart-types des deux estimateurs. A l'aide des données de la sortie **R3** indiquez l'écart type de l'estimateur associé au sexe,

A. 10.16
B. 3.187
C. 0.2605
D. 0.0307

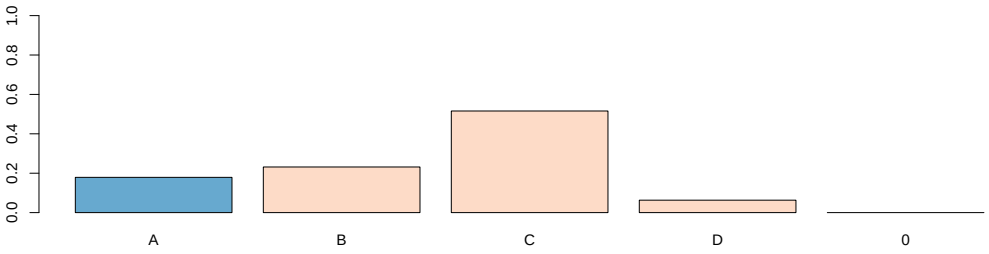


Question 27. (suite de la question 26) Dans le modèle `reg3b` de la sortie **R3**, on souhaite donner un intervalle de confiance pour la note prédite $\widehat{Y}$ pour cette élève (fille de 23 ans, qui avait eu 61 au cours MAT3080, et 47.5 au cours ACT6420). Quel intervalle de confiance à 95% aurait-on, autour de $\widehat{Y}$?

A. $\pm$ 3.75
B. $\pm$ 10.13
C. $\pm$ 3.42
D. $\pm$ 2.36

Prenons deux minutes pour refaire le calcul, comme la majorité semble avoir un résultat différent du mien... Ici, la prédiction est $\widehat{Y} = 61\beta_0$ (ici il n'a pas de constante, β_0 est la pente associée à la note à l'examen de statistique. Donc $\text{Var}(\widehat{Y}) = 61^2\text{Var}(\widehat{\beta}_0)$. Or il manque cette variance, mais on peut la calculer, car $T = \widehat{\beta}_0/\sqrt{\text{Var}(\widehat{\beta}_0)}$. Bref, ici la variance serait 0.0308^2. Aussi, l'écart-type de $\widehat{Y}$ sera 61 fois l'écart-type de $\widehat{\beta}_0$. Soit 1.87772. Pour passer à l'intervalle de confiance à 95%, on fait une hypothèse de normalité sur nos résidus,

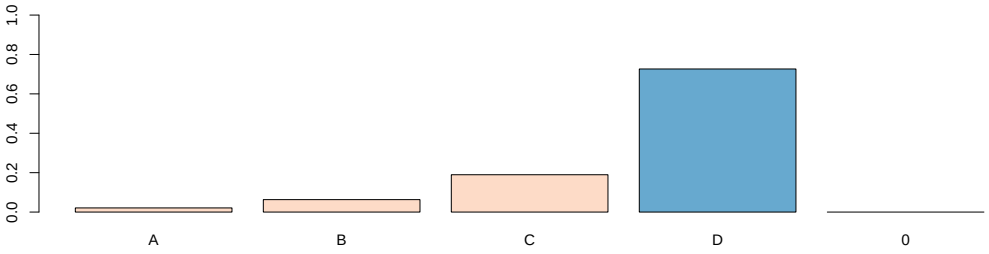
qui va induire une normalité de nos estimateurs, et donc, l'intervalle de confiance à 95% sera ± 2 fois cette valeur (utilisant 2 comme indiqué en première page.)



Question 28. On pense que la note prédite doit être révélatrice du travail fourni par l'étudiant, et on souhaite la prendre en compte. À l'aide du modèle `reg4` de la sortie **R4** on hésite sur le modèle à retenir. Parmi les affirmations suivantes, lesquelles sont vraies ?

(i) on retient `reg4` par rapport à `reg3b` sur la base du R^2 ajusté
(ii) on retient `reg4` par rapport à `reg3b` sur la base du critère d'Akaike
(iii) on retient `reg3b` par rapport à `reg4` sur la base du R^2 ajusté
(iv) on retient `reg3b` par rapport à `reg4` sur la base du critère d'Akaike

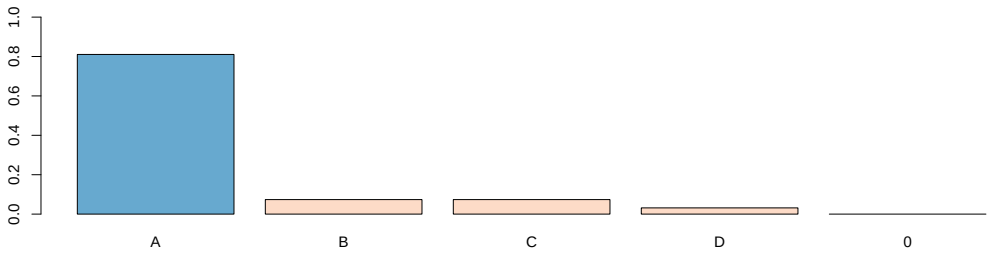
A. (i) et (ii) sont vraies
B. (i) et (iv) sont vraies
C. (iii) et (ii) sont vraies
D. (iii) et (iv) sont vraies



Question 29. Avant de finir, on souhaite comprendre qui se trompe sur la prédiction. On considère des modèles pour l'erreur de prédiction faite par les élèves. Les modèles `reg5a` et `reg5b` sont présentés dans la sortie **R5**.

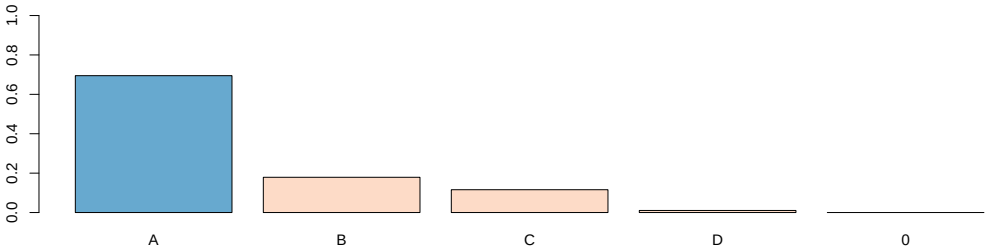
(i) on retient `reg5a` par rapport à `reg5b` sur la base du R^2 ajusté
(ii) on retient `reg5a` par rapport à `reg5b` sur la base du critère d'Akaike
(iii) on retient `reg5b` par rapport à `reg5a` sur la base du R^2 ajusté
(iv) on retient `reg5b` par rapport à `reg5a` sur la base du critère d'Akaike

A. (i) et (ii) sont vraies
B. (i) et (iv) sont vraies
C. (iii) et (ii) sont vraies
D. (iii) et (iv) sont vraies



Question 30. On regarde - avant de conclure - les résidus des modèles `reg5a` et `reg5b` (présentés dans la sortie **R5**).

A. les résidus du modèle `reg5b` peuvent être supposés Gaussiens
B. les tests rejettent l'hypothèses que les résidus de `reg5b` puissent être Gaussiens
C. les résidus du modèle `reg5a` et du modèle `reg5b` sont forcément Gaussiens car ils ont été obtenus en estimant des modèles linéaires
D. on ne peut rien dire sur les résidus car on a moins de 100 observations


0.0
0.2
0.4
0.6
0.8
1.0
A
B
C
D
0