

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 données 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 point 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%.

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 dispose 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 \bar{X}
- D. la qualité de l'ajustement par une droite

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 \bar{X}
- D. la qualité de l'ajustement par une droite

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

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

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. $\hat{\alpha}_0 = \frac{1}{n} \sum_{i=1}^n X_i$
- B. $\hat{\alpha}_0 = \frac{1}{n} \sum_{i=1}^n Y_i$
- C. $\hat{\alpha}_0 = \frac{1}{n-1} \sum_{i=1}^n X_i$
- D. $\hat{\alpha}_0 = \frac{1}{n-1} \sum_{i=1}^n Y_i$

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. $\hat{\gamma}_1 = \frac{\bar{Y}}{\bar{X}}$
- B. $\hat{\gamma}_1 = \frac{1}{\bar{X}}$
- C. $\hat{\gamma}_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$
- D. $\hat{\gamma}_1 = 1$

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

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

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

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

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 $i \in \{1, \dots, n\}$ $(Y_i - \bar{Y})^2 = (Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \bar{Y})^2$
- B. pour tout $i \in \{1, \dots, n\}$ $(Y_i - \hat{Y}_i)^2 = (Y_i - \bar{Y})^2 + (\hat{Y}_i - \bar{Y})^2$
- C. $\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$
- D. $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \bar{Y})^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$

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) $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 0$
- (ii) $\sum_{i=1}^n (Y_i - \hat{Y}_i) = 0$
- (iii) $\sum_{i=1}^n (Y_i - \hat{Y}_i) X_i = 0$
- (iv) $\sum_{i=1}^n (Y_i - \hat{Y}_i) Y_i = 0$

- A. (i) seulement
- B. (iii) et (iv)
- C. (ii) seulement
- D. (ii) et (iii).

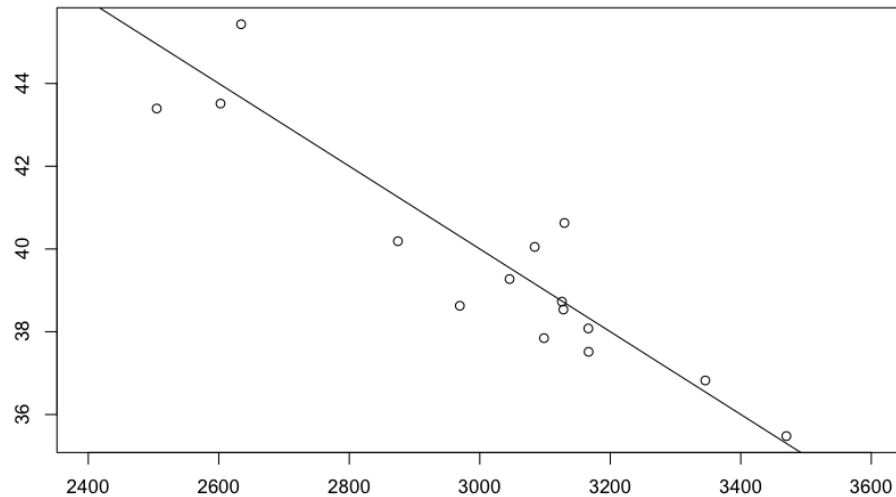
Question 10. On se rends compte que l'on a oublié de prendre en compte une variable explicative, Z . On suppose que $\sum_{i=1}^n X_i Z_i = \sum_{i=1}^n X_i \cdot \sum_{i=1}^n Z_i$. On construit alors le modèle

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

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

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

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 vraie

- 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 dessous de la droite de régression

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

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 falsifiée car les chiffres ne sont pas cohérents entre eux,

- A. $\hat{\beta}_1$ devrait être positif car $\hat{\beta}_0 > 0$ 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. $\hat{\beta}_0$ et $\hat{\beta}_1$ 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

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

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}, Y_{21}) dans le graphique des (X_i, Y_i)

- A. on ne peut pas savoir
- B. proche de (\bar{X}, \bar{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

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

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. $\left[0.4057 \pm 2 \cdot \frac{0.1419}{\sqrt{39}}\right]$
- C. $[0.4057 \pm 2 \cdot 12.67]$
- D. $\left[0.4057 \pm 2 \cdot \frac{12.67}{\sqrt{39}}\right]$

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

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'élève (41) conteste 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]

Question 26. Dans le modèle **reg3b**, il manque les écart-types des deux estimateurs. À 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 \hat{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 \hat{Y} ?

- A. ± 3.75
- B. ± 10.13
- C. ± 3.42
- D. ± 2.36

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'hypothèse 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