Université de Paris 1 Panthéon-Sorbonne
Cours de Marc-Arthur Diaye/Alexis Bogroff
Sorbonne Data Analytics, Session 4, Juin 2024

Évaluation Statistique
Éléments de théorie

Exercice 1. QCM (Vrai/Faux)

(+0,25 point si bonne réponse, 0 point si pas de réponse, -0,25 point si mauvaise réponse)

Question 1. Lorsque les éléments non diagonaux de la matrice de variance-covariance des termes d'erreur sont nuls, on a de l'hétéroscédasticité.

[FAUX] Lorsque les éléments non diagonaux de la matrice de variance-covariance des termes d'erreur sont nuls, cela signifie qu'il n'y a pas de corrélation entre les erreurs, indiquant l'absence d'autocorrélation, et non pas la présence d'hétéroscédasticité.

Question 2. Lorsque les éléments non diagonaux de la matrice de variance-covariance des termes d'erreur sont nuls, on a de la non-autocorrélation.

[VRAI] Si les éléments non diagonaux de la matrice de variance-covariance des termes d'erreur sont nuls, cela signifie qu'il n'y a pas d'autocorrélation entre les erreurs.

Question 3. Si les éléments diagonaux de la matrice de variance-covariance des termes d'erreur sont égaux à l, alors il y a de l'hétéroscédasticité.

[FAUX] Si les éléments diagonaux de la matrice de variance-covariance des termes d'erreur sont constants (égaux à l), cela indique une homoscédasticité (variance constante des erreurs), et non une hétéroscédasticité.

Question 4. Lorsqu'une variance est négative alors on a de l'autocorrélation négative.

[FAUX] La variance ne peut jamais être négative. L'autocorrélation négative se réfère à la corrélation entre les termes d'erreur, pas à une variance négative.

Question 5. L'estimateur des moindres carrés ordinaires est obtenu en minimisant la somme des carrés des résidus.

[VRAI] L'estimateur des moindres carrés ordinaires (MCO) est obtenu en minimisant la somme des carrés des résidus.

Question 6. La statistique de Student associée à une variable explicative est égale au paramètre estimé associé à cette variable explicative plus par l'écart-type estimé du paramètre estimé.

[FAUX] La statistique de Student est égale au paramètre estimé divisé par l'écart-type estimé du paramètre, pas à la somme.

Question 7. Minimiser la somme des carrés des erreurs est équivalent à minimiser la distance au carré entre Y (le vecteur des valeurs observe concernant la variable dépendante) et le sous-espace vectoriel engendré par les vecteurs colonnes de la matrice X.

[VRAI] Minimiser la somme des carrés des erreurs est équivalent à minimiser la distance au carré entre YY et le sous-espace vectoriel engendré par les colonnes de la matrice XX.

Question 8. On peut trouver la valeur estimée de Y en la multipliant par un projecteur orthogonal sur le sous-espace vectoriel engendré par les vecteurs colonnes de la matrice X.

[VRAI] La valeur estimée de YY peut être obtenue en multipliant YY par un projecteur orthogonal sur le sous-espace engendré par les colonnes de XX.

Question 9. Le fait que l'espérance du terme d'erreur est nulle est un résultat alors que le fait que la moyenne des résidus est nulle est une hypothèse.

[FAUX] L'espérance du terme d'erreur nulle est une hypothèse (non-biaisée), tandis que la moyenne des résidus nulle est un résultat découlant des MCO.

Question 10. La matrice de projection orthogonale utilisée dans les moindres carrées ordinaires est $X(X^{!}X)^{"\#}X'$.

[VRAI] La matrice de projection orthogonale dans les MCO est X(X'X)−1X'X(X'X)−1X'.

Question 11. Dans le modèle linéaire général, Y (le vecteur des valeurs observés concernant la variable dépendante) est égal à la somme de la valeur estimée de Y et du vecteur des résidus.

[VRAI] Dans le modèle linéaire général, YY est égal à la somme de la valeur estimée de YY et du vecteur des résidus.

Question 12. La moyenne de la variable dépendante est égale à la moyenne de la variable estimée de la variable dépendante.

[VRAI] La moyenne de la variable dépendante est égale à la moyenne de la variable estimée dans le modèle MCO.

Question 13. Un estimateur sans biais de la variance du terme d'erreur est égal à la somme des carrés des résidus divisée par le nombre de degré de liberté.

[VRAI] Un estimateur sans biais de la variance des erreurs est la somme des carrés des résidus divisée par le nombre de degrés de liberté.

Question 14. L'autocorrélation veut dire que la variance des termes d'erreur varie selon les individus.

[FAUX] L'autocorrélation concerne la corrélation entre les termes d'erreur consécutifs, pas la variation de la variance des erreurs selon les individus.

Question 15. Minimiser la somme des carrés des erreurs est équivalent à minimiser la distance au carré entre Y (le vecteur des valeurs observés concernant la variable dépendante) et le sous-espace vectoriel engendré par les vecteurs colonnes de la matrice X.

[VRAI] Voir la réponse à la question 7.

Question 16. En présence d'autocorrélation, l'estimateur des moindres carrés ordinaires reste à variance minimale.

[FAUX] En présence d'autocorrélation, l'estimateur des moindres carrés ordinaires n'est plus à variance minimale.

Question 17. Lorsqu'on rejette l'hypothèse de base dans un test de Student, cela veut que la variable explicative associée est aléatoire.

[FAUX] Rejeter l'hypothèse nulle dans un test de Student signifie que la variable explicative est significative, pas aléatoire.

Question 18. On peut calculer la statistique de Fisher à partir du coefficient de détermination.

[VRAI] La statistique de Fisher peut être calculée à partir du coefficient de détermination R2R2.

Question 19. Lorsqu'on rejette l'hypothèse de base dans le test de nullité globale de Fisher, cela veut dire qu'aucune variable explicative du modèle n'influence la variable dépendante.

[FAUX] Rejeter l'hypothèse de nullité globale de Fisher signifie qu'au moins une variable explicative influence la variable dépendante.

Question 20. La statistique de Fisher dans le test de nullité globale est nécessairement positive.

[VRAI] La statistique de Fisher est une somme de carrés et donc nécessairement positive.

Question 21. L'estimateur des moindres carrés généralisés s'écrit $\boldsymbol{\beta^{\&}_{MCG} = (X^{!}\Omega^{"1}X)^{"1}X'\Omega^{"1}Y}$.

[VRAI] L'estimateur des moindres carrés généralisés (MCG) est β^MCG=(X'Ω−1X)−1X'Ω−1Yβ^MCG =(X'Ω−1X)−1X'Ω−1Y.

Question 22. Dans un test de nullité globale de Fisher, si la p-value est de 0,1 et que le risque de première espèce que l'on est prêt à accepter est de 1% alors on rejette l'hypothèse de base.

[FAUX] Avec une p-value de 0,1 et un seuil de 1%, on ne rejette pas l'hypothèse nulle.

Question 23. La différence principale entre l'estimateur des moindres carrés généralisés et l'estimateur des moindres carrés quasi-généralisés est que l'estimateur des moindres carrés quasi-généralisés est BLUE.

[FAUX] La principale différence est que l'estimateur des moindres carrés généralisés (MCG) est BLUE, tandis que l'estimateur des moindres carrés quasi-généralisés ne l'est pas.

Question 24. Le théorème de Gauss-Markov dit que l'estimateur des moindres carrés ordinaires est convergent.

[FAUX] Le théorème de Gauss-Markov stipule que l'estimateur des moindres carrés ordinaires est BLUE, pas nécessairement convergent.

Question 25. Si la matrice X n'est pas de plein-rang colonne alors la matrice X'X admet une valeur propre nulle.

[VRAI] Si la matrice XX n'est pas de plein rang, alors $X'XX'X$ a une valeur propre nulle.

Question 26. En cas de corrélation forte entre deux variables explicatives du modèle, l'estimateur des moindres carrés ordinaires est biaisé.

[FAUX] La multicolinéarité n'introduit pas de biais, mais elle augmente la variance des estimateurs.

Question 27. En cas de corrélation forte entre des variables explicatives du modèle, on ne peut faire confiance aux tests de Student.

[VRAI] En cas de forte corrélation entre variables explicatives, les tests de Student peuvent devenir peu fiables.

Question 28. Lorsque je fais une régression linéaire et que je rejette l'hypothèse de nullité globale de Fisher sans rejeter, dans les tests de Student, la nullité des paramètres associés aux variables explicatives, alors il y a suspicion d'autocorrélation des résidus.

[VRAI] Cela peut indiquer une autocorrélation des résidus si la nullité globale est rejetée mais pas les tests individuels.

Question 29. Supposons une équation de salaire avec comme variable explicative, la variable Femme (égale à 1 si l'individu est une femme et 0 sinon), la variable Jeune (égale à 1 si l'individu est Jeune et 0 sinon) et le croisement de Femme et de Jeune, alors la constante du modèle représente le salaire moyen des jeunes femmes.

[FAUX] La constante représente le salaire moyen des individus pour lesquels toutes les variables explicatives binaires sont nulles (non-jeunes hommes).

Question 30. Pour corriger de l'autocorrélation, il faut multiplier le modèle initial $Y=Xb + e$, par une matrice P, permettant au nouveau terme d'erreur Pe, d'avoir une matrice de variance-covariance avec tous ses éléments non-diagonaux nuls.

[VRAI] La multiplication par une matrice PP permet d'obtenir une matrice de variance-covariance des erreurs avec des éléments non diagonaux nuls, corrigeant l'autocorrélation.

Question 31. Soit le modèle linéaire général Y=Xb + e, avec Var(e) = s2W où W 1 I_n la matrice identité. Alors l'estimateur des moindres carrés ordinaires est biaisé.

[FAUX] Si Var(ϵ)=σ2WVar(ϵ)=σ2W avec W≠IW⍰=I, les MCO peuvent être biaisés, sauf si WW est connue et correctement ajustée.

Question 32. En présence d'autocorrélation, l'estimateur des moindres carrés ordinaires reste convergent.

[VRAI] En présence d'autocorrélation, l'estimateur des moindres carrés ordinaires reste convergent.

Question 33. Pour tester l'autocorrélation d'ordre 1, on peut utiliser le test de Durbin-Wu-Hausman.

[FAUX] Le test de Durbin-Watson est couramment utilisé pour tester l'autocorrélation d'ordre 1.

Question 34. Le test de White permet de tester l'hétéroscédasticité des résidus.

[VRAI] Le test de White permet de tester l'hétéroscédasticité des résidus.

Question 35. La méthode du Condition Index permet de détecter un problème de presque colinéarité entre des variables explicatives du modèle.

[VRAI] La méthode de l'indice de condition permet de détecter la colinéarité entre variables explicatives.

Question 36. Le test de Jarque-Bera permet de détecter la convergence des résidus.

[FAUX] Le test de Jarque-Bera teste la normalité des résidus, pas leur convergence.

Question 37. En cas de multicolinéarité parfaite, l'estimateur des moindres carrés ordinaires reste BLUE.

[FAUX] En cas de multicolinéarité parfaite, les estimateurs MCO ne sont pas définis, donc pas BLUE.

Question 38. On dispose de données pour n individus où : C'est la consommation et R est le revenu disponible.Soitlemodèlesuivant(1): $C_i = b_0 + b_1 R_i + e_i$, i=1àn;avecVar(e_i)=s2logR_i et Cov(e_i, $e_{i'}$) = 0, " i 1 i'. Alors le modèle (1) viole l'hypothèse de normalité des termes d'erreur.

[FAUX] Var(eiei) = σ2log⍰Riσ2logRi viole l'homoscédasticité, mais pas nécessairement l'hypothèse de normalité des erreurs.

Question 39. Une distribution qui suit une loi normale correspond à un Skewness=0 et un Kurtosis=3.

[VRAI] Une distribution normale a un skewness de 0 et un kurtosis de 3.

Question 40. Soit l'estimation suivante :
$Budget_i = 19{,}44 + 0{,}018X_{1i} - 0{,}284X_{2i} + 1{,}343X_{3i} + 6{,}331X_{4i}$ i = 1 à 2000 ;
(3,406) (0,006) (0,457) (0,259) (3,029)
Au seuil de 5%, la valeur dans la table de student est 1,96. Selon Zeina, la variable X1 n'est pas explicative.

[FAUX] Le coefficient estimé de X1X1 est significatif si sa valeur absolue dépasse 1,96, donc il est explicatif.

Question 41. Dans un modèle logit simple, au moins l'une des variables explicatives est binaire.

[FAUX] Dans un modèle logit, les variables explicatives ne doivent pas nécessairement être binaires.

Question 42. Dans un modèle probit simple, la variable dépendante est binaire.

[FAUX] Dans un modèle logit, les variables explicatives ne doivent pas nécessairement être binaires.

Question 43. Le vecteur des coefficients estimés dans un modèle logit simple est solution d'un programme de maximisation d'une fonction de vraisemblance.

[VRAI] Les coefficients estimés dans un modèle logit sont obtenus par maximisation de la fonction de vraisemblance.

Question 44. Les coefficients estimés d'une régression logistique simple s'interprètent comme des élasticités.

[FAUX] Les coefficients dans une régression logistique sont interprétés comme des variations log-odds, pas des élasticités.

Question 45. Dans un modèle logit, le terme aléatoire suit une loi normale.

[FAUX] Dans un modèle logit, le terme aléatoire suit une loi logistique.

Question 46. Dans un modèle probit, le terme aléatoire suit une loi logistique.

[FAUX] Dans un modèle probit, le terme aléatoire suit une loi normale.

Question 47. Les coefficients estimés des régressions logit et probit peuvent parfois avoir des signes opposés.

[VRAI] Les signes des coefficients estimés peuvent parfois différer entre les modèles logit et probit.

Question 48. Dans un modèle logit, la significativité d'une variable explicative se fait via un test dit de Thomas.

[FAUX] La significativité dans un modèle logit est généralement testée via le test de Wald, pas de Thomas.

Question 49. Dans la pratique, l'estimation des coefficients estimés d'un modèle logit, peut se faire par l'algorithme de Newton-Raphson.

[VRAI] L'estimation des coefficients d'un modèle logit peut se faire par l'algorithme de Newton-Raphson.

Question 50. Le AIC d'une régression permet de voir si la variable dépendante est vraiment qualitative.

[FAUX] Le AIC évalue la qualité du modèle, pas la nature qualitative ou quantitative de la variable dépendante.

Question 51. La principale différence entre un modèle logit multinomial et un modèle simple réside dans leurs termes d'erreur qui sont différents.

[FAUX] La différence principale réside dans la structure des erreurs, mais aussi dans la nature des choix (multinomial vs binaire).

Question 52. Le signe d'une Log Vraisemblance n'est pas déterminé a priori.
[VRAI] Le signe d'une log-vraisemblance dépend de la définition et des données, donc il n'est pas déterminé a priori.

Question 53. Le modèle de régression multinomiale est basé sur la théorie du choix probabiliste.

[VRAI] Le modèle de régression multinomiale est basé sur la théorie du choix probabiliste.

Question 54. Dans le modèle de choix probabiliste, le concept important est que la rationalité de l'agent est probabiliste.

[VRAI] Dans le modèle de choix probabiliste, la rationalité de l'agent est interprétée de manière probabiliste.

Question 55. Les modèles traditionnels respectent l'IIA.

[FAUX] Les modèles traditionnels (logit, probit) respectent généralement l'hypothèse d'indépendance des alternatives non pertinentes (IIA), mais des extensions comme le logit multinomial peuvent ne pas le faire.

Question 56. Dans un modèle de régression multinomiale ordonnée, il y a autant de paramètres estimés que de modalités à la variable dépendante.

[FAUX] Dans un modèle de régression multinomiale ordonnée, le nombre de paramètres estimés est lié aux modalités de la variable dépendante mais pas égal à leur nombre.

Question 57. Il y a trois types de modèles de régression multinomiale non-ordonnée.

[FAUX] Il existe principalement deux types de modèles multinomiaux : logit multinomial et probit multinomial.

Question 58. Le pseudo-R2 de McFadden peut être négatif.
[FAUX] Le pseudo-R^2 de McFadden est compris entre 0 et 1, donc il ne peut pas être négatif.

Exercice 2.

On dispose de données (fictives) concernant 1500 individus et comportant trois types d'informations, à savoir : la prise de vitamines de type 1 **(vita1)** et de type 2 **(vita2)** *du 1er janvier de l'année en cours au 31 décembre de la même année,* ainsi que la hausse de taille observée **(taille)** sur cette période.

Question 1. Compléter (en justifiant rigoureusement) la sortie logicielle ci-dessous.

(a) = Nombre de prédicteurs (2 si vita1 et vita2 sont considérés comme distincts)
a = 2

(b) = n−a−1=1500−a−1n−a−1=1500−a−1
b = 1497

(c) SSModel = R2 × SSTotal = 0,32431

(d) SS_Error = SS_Total - SS_Model = 78,77656

(e) MS_Model = SS_Model / DF_Model = 29,60

(f) = MS_Model / MS_Error = 3.08.

(g) = $\sqrt{Error} = \sqrt{3,60200} = 1,89789$

(h) = $\frac{29,60595}{78,77656} = 0,3758$

(i) Non donné, mais t = -0.01040 / 0.00425 ≈ -2.45 (on peut vérifier si p ≈ 0.014).
i ≈ -0.01040

(j) Non donné, mais la p-valeur est 0.2262.
j ≈ 0.00425

(k) = $\frac{0.01040}{0,00425} \approx -2,45$

Question 2. Que conclue le test de nullité globale de Fisher au seuil de 1%.

La p-value du test de Fisher est 0.0166, ce qui est supérieur à 0.01. On ne rejette donc pas l'hypothèse nulle au seuil de 1%. Cela signifie que le modèle global n'est pas significatif à ce seuil.

Question 3. Que concluent les tests de nullité des paramètres au seuil de 1%.

- **Intercept**: Très significatif (p < 0.0001), donc on rejette l'hypothèse nulle.
- **vita1**: Non significatif (p = 0.014), supérieur à 0.01.
- **vita2**: Non significatif (p = 0.2262), supérieur à 0.01.

Aucun des paramètres vita1 et vita2 n'est significatif au seuil de 1%.

Question 4. Interpréter le résultat de l'estimation. Vous semble-t-elle satisfaisante ?

- **Intercept**: La valeur de l'intercept (2.27839) est significative, indiquant une hausse moyenne de taille sans consommation de vitamines.
- **vita1**: La prise de vita1 a un effet négatif très léger sur la taille, mais cet effet n'est pas significatif au seuil de 1%.
- **vita2**: L'effet de vita2 sur la taille n'est pas significatif.

Dependent Variable: taille

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	**a**	29.60595	**e**	**f**	0.0166
Error	**b**	**c**	3.60200		
Corrected Total	1499	**d**			

Root MSE	**g**	R-Square	**h**
Dependent Mean	2.39933	Adj R-Sq	0.0041
Coeff Var	79.10087		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.27839	0.12633	**k**	<.0001
vita1	1	**i**	0.00425	2.63	0.0087
vita2	1	-0.01040	**j**	-1.21	0.2262

Exercice 3 : Autocorrélation d'ordre 1

On dispose de données temporelles avec une variable dépendante Y et sur une variable explicative X1.

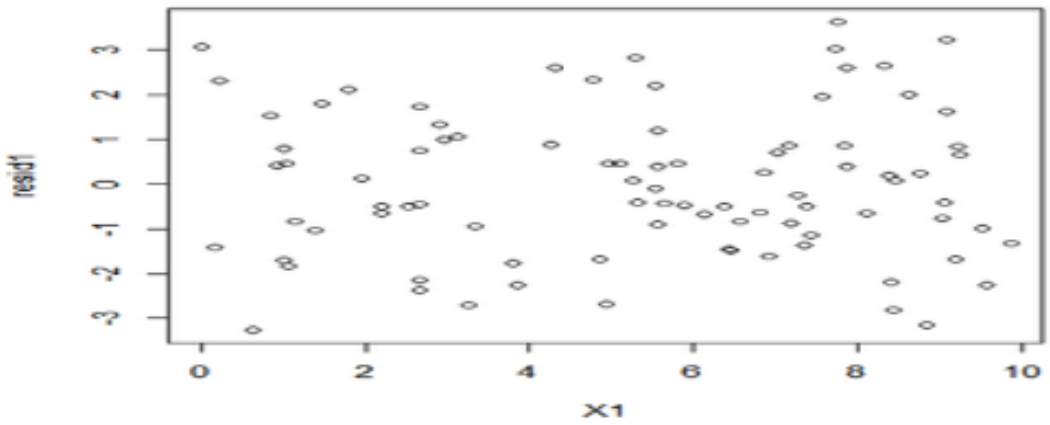
1) Commenter l'estimation ci-dessous :

```
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)  2.06845    0.37239      5.555     3.05e-07 ***
X1           0.08322    0.06148      1.354     0.179
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.632 on 86 degrees of freedom
Multiple R-squared:  0.02086,   Adjusted R-squared:  0.009477
F-statistic: 1.832 on 1 and 86 DF,  p-value: 0.1794
```

L'estimation montre que lorsque la variable explicative X1 augmente d'une unité, la variable dépendante Y augmente en moyenne de 0.08322 unités, tout en tenant compte de l'intercept qui est de 2.06845. Cependant, la faible valeur du coefficient de détermination (R^2 ajusté = 0.009477) indique que la variable explicative X1 explique peu la variation de Y.

2) Le graphe ci-dessous décrit les résidus en fonction de X1. Au vu du graphe, y a-t-il hétéroscédasticité ou homoscédasticité ?



Le graphe ci-dessus exprime les résidus en fonction de X1. Pour déterminer s'il y a homoscédasticité ou hétéroscédasticité, il faudrait examiner la dispersion constante ou variable des résidus en fonction de X1. A première vue, il semblerait y avoir de l'hétéroscédasticité.

3) Quelle est l'hypothèse de base du test de Breusch-Pagan ?

Le test de Breusch-Pagan teste l'hypothèse nulle que la variance des résidus est constante (homoscédasticité) contre l'alternative qu'elle est différente (hétéroscédasticité).

4) Le test de Breusch-Pagan donne le résultat ci-dessous :

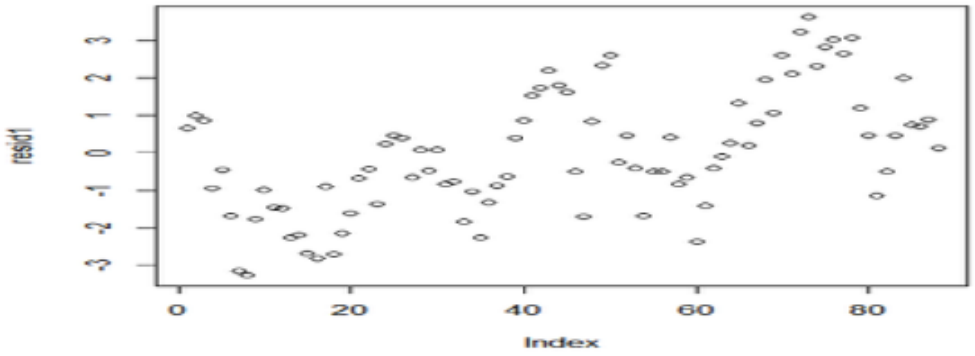
studentized Breusch-Pagan test
data: reg1
BP = 0.00026223, df = 1, p-value = 0.9871

Que concluez-vous quant à hétéroscédasticité ou homoscédasticité ?

Le test donne un p-value de 0.9871. Avec une valeur de p aussi élevée, on ne rejette pas l'hypothèse nulle d'homoscédasticité. Ainsi, il semble que les résidus présentent une homoscédasticité.

5) Le graphe ci-dessous donne la distribution des résidus. Au vu de ce graphe, y a-t-il autocorrélation des résidus ?

Pour déterminer s'il y a autocorrélation des résidus, il faudrait visualiser le graphe B qui donne la distribution des résidus. Sans cette visualisation, il est difficile de conclure sur la présence d'autocorrélation.



6) Quelle est l'hypothèse de base du test de Durbin et Watson ?

Le test de Durbin-Watson teste l'autocorrélation des résidus en vérifiant si les erreurs sont indépendantes les unes des autres (hypothèse nulle) ou si elles sont corrélées (hypothèse alternative).

7) Le test de Durbin et Watson donne le résultat ci-dessous :

Durbin-Watson test
data: reg1
DW = 0.39904, p-value < 2.2e-16

Que concluez-vous quant à l'existence de l'autocorrélation ?

Le test donne un DW de 0.39904 avec une très faible p-value (< 2.2e-16). Cela suggère fortement la présence d'autocorrélation positive des résidus.

8) L'estimation en moindres carrés généralisés donne :

Coefficients:

	Value	*Std.Error*	*t-value*	*p-value*
(Intercept)	*1.9475699*	*0.5495166*	*3.544151*	*6e-04*
X1	*0.1123282*	*0.0272659*	*4.119724*	*1e-04*

Commentez cette estimation, notamment en la comparant avec l'estimation de la question 1).

Les t-values sont plus élevées, indiquant une plus grande significativité des coefficients.

Cette estimation suggère une relation plus forte entre X1 et Y comparée à la première estimation, avec des coefficients plus précis et significatifs, malgré des valeurs légèrement différentes.

Ces réponses sont basées sur les informations fournies. Pour une analyse plus détaillée, des graphiques et des données supplémentaires seraient nécessaires.