



Modèles Linéaires

Contrôle - 1h30 Licence 3 MIASHS (2024-2025)

Guillaume Metzler

Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

L'usage de matériel électronique (téléphone, ordinateur), des notes de cours ou des notes personnelles n'est pas autorisé pendant la durée de cette épreuve

En revanche, l'usage de la calculatrice est autorisé.

Résumé

Les exercices de cet examen sont indépendants. Comme dans tout examen, une grande importance sera accordée à la qualité de la rédaction des réponses. Une réponse avec un résultat sans justification ne pourra prétendre à l'obtention de la totalité des points.

Exercice 1 : Questions théoriques

1. Rappeler l'expression d'un modèle linéaire gaussien multiple où l'on cherchera à prédire les valeurs d'une variable dépendante Y en fonction d'un ensemble de p variables indépendantes.
On prendra le soin d'énoncer les hypothèses relatives au modèle linéaire multiple en utilisant les notations introduites pour répondre à la question.
2. Supposons que l'on dispose d'un échantillon de n observations stockées dans une matrice de *design* $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$ avec les valeurs associées de Y dans un vecteur $\mathbf{y} \in \mathbb{R}^n$.
Donner l'expression de l'estimateur, que l'on notera $\hat{\boldsymbol{\beta}}$, des paramètres du modèle obtenu en résolvant le problème d'optimisation suivant :

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2. \quad (1)$$

Remarque : il n'est pas demandé de démontrer le résultat.

3. Montrer que l'estimateur $\hat{\boldsymbol{\beta}}$ est non biaisé, *i.e.*, que l'on a

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}.$$

4. Montrer que la variance de cet estimateur est égale à

$$\text{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1},$$

où σ^2 est la variance des résidus/erreurs du modèle.

5. Soit $\hat{\boldsymbol{\beta}}$ l'estimateur obtenu en résolvant le problème (1). Montrer que les résidus $\hat{\boldsymbol{\varepsilon}}$ du modèle ne sont pas corrélés aux variables indépendantes :

$$\mathbf{X}^\top \hat{\boldsymbol{\varepsilon}} = \mathbf{0}.$$

Indication : on commencera par utiliser la définition de $\hat{\boldsymbol{\varepsilon}}$.

6. **(Bonus)** Soit $\hat{\boldsymbol{\beta}}$ l'estimateur obtenu en résolvant le problème (1). Montrer que les résidus $\hat{\boldsymbol{\varepsilon}}$ du modèle ne sont pas corrélés au vecteur $\hat{\boldsymbol{\beta}}$:

$$\text{Cov}[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\varepsilon}}] = \mathbf{0}.$$

Exercice 2 : Un modèle multiple

Au cours d'une étude, on cherche à savoir s'il est possible de prédire le score Y obtenu à un examen par un étudiant en fonction de plusieurs informations : (i) le nombre d'heures d'études X_1 , (ii) le nombre d'heures passées sur les réseaux sociaux X_2 et (iii)

le nombre d'heures passées à résoudre des problèmes X_3 .

A l'aide d'un échantillon S et en résolvant le problème (1), on trouve le modèle suivant :

$$\hat{Y} = 10 + 2X_1 - 3X_2 + 0.5X_3.$$

1. Interpréter les coefficients associés aux variables X_1 et X_2 .
2. Prédire le score obtenu à l'examen par étudiant qui aurait étudié 5 heures, passé 4 heures sur les réseaux sociaux et 20 heures à résoudre des problèmes.

On s'intéresse maintenant à l'analyse des coefficients du modèle et on aboutit à la table suivante :

Paramètre	Estimation	Erreur standard	t-value	p-value
Intercept	10	1.2	3.75	0.002
X_1	2.0	0.5	4.00	0.001
X_2	-3	1.2	-2.50	0.03
X_3	0.5	0.5	1	0.32

3. On souhaite tester la significativité des paramètres. Énoncer les hypothèses, la statistique de test et sa distribution ainsi que la façon de conclure à ce test au seuil de significativité $\alpha \in [0, 1]$.
4. Au seuil de significativité $\alpha = 0.05$, déterminer les variables significatives du modèle.
5. Serait-il légitime de supprimer les variables non significatives ? Quel serait l'impact du retrait d'une telle variable sur le coefficient de détermination R^2 ? Même chose mais sur sa version ajustée R_{aj}^2 ?

On cherche maintenant à étudier la présence de multi-colinéarité dans notre modèle de régression.

6. Expliquer comment vous pourriez détecter la multi-colinéarité dans un modèle de régression multiple.
7. Supposons que les variables X_2 et X_3 ont un VIF supérieur à 20. Après avoir donné la définition du VIF et rappeler son usage, expliquer la démarche que vous entreprendriez afin de corriger ce problème de multi-colinéarité.

Exercice 3 :

Un chercheur étudie la relation entre le **chiffre d'affaires mensuel** (Y) d'une entreprise et deux variables explicatives :

- **Budget publicitaire** (X_1), en milliers d'euros,
- **Nombre de représentants du service client** (X_2).

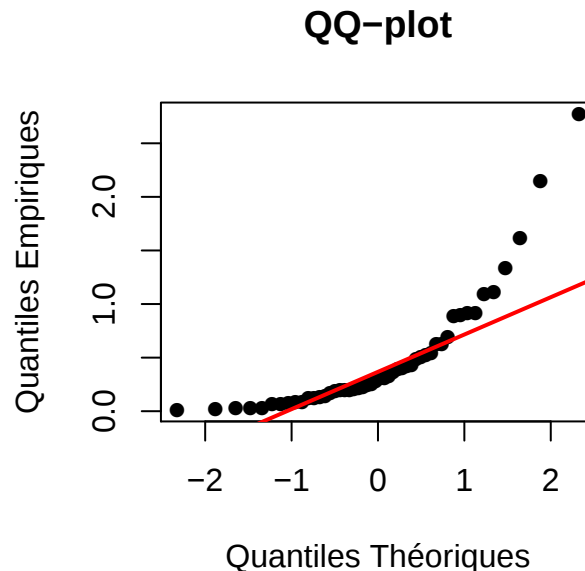
Le modèle considéré est le suivant :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \text{où } \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

On suppose que les résidus doivent suivre une loi normale avec variance constante. L'étude est menée sur un échantillon de $n = 50$ observations.

Normalité des résidus On commence par s'intéresser à la normalité des résidus.

1. Le chercheur trace un **QQ-plot** des résidus. Comment peut-il vérifier si les résidus suivent une distribution normale ?
2. Supposons que le QQ-plot ait la forme suivante



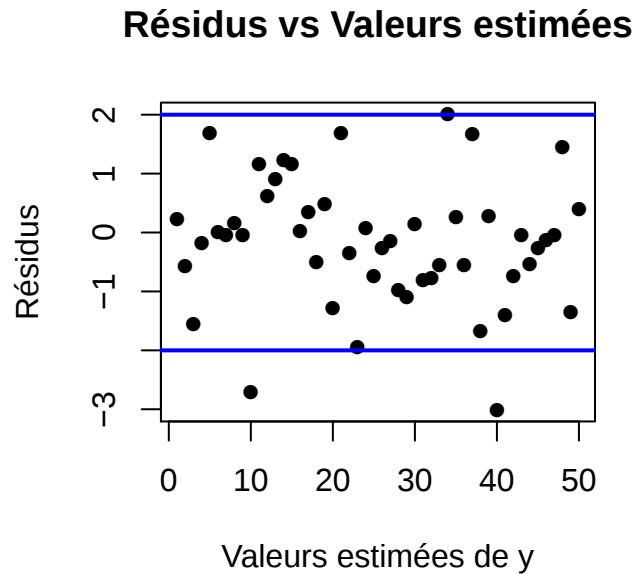
Peut-on dire que l'hypothèse de normalité des résidus est vérifiée. Expliquer votre réponse.

3. Quel test statistique permettrait de vérifier la normalité des résidus ?

Homoscedasticité (variance constante) On cherche maintenant à vérifier l'hypothèse d'homoscédasticité.

4. Le chercheur trace un graphique des **résidus en fonction des valeurs estimée**. Quelle structure indiquerait une violation de l'hypothèse d'homoscédasticité ?

5. Le chercheur obtient le graphique ci-dessous :



Est-ce que l'hypothèse d'homoscédasticité est vérifiée ?

6. Quelle transformation ou méthode peut être utilisée si une hétéroscédasticité est détectée ?

Détection d'observations influentes via les *hat-values* La matrice de projection $\mathbf{H}$ (*hat-matrix*) est donnée par :

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$$

où les éléments diagonaux h_{ii} mesurent l'influence de chaque observation.

7. Le chercheur utilise la règle $h_{ii} > \frac{2p}{n}$ pour identifier les valeurs aberrantes.
- Déterminez le **seuil** pour détecter les valeurs aberrantes si $p = 2$ et $n = 50$.
 - Une observation a $h_{ii} = 0.15$. Est-elle influente ? Justifiez.
8. Quel autre graphique vous permettrait de détecter d'éventuelles valeurs aberrantes ?