Modèles Linéaires

Devoir Maison
Licence 3 MIASHS (2025 - 2026)

Guillaume Metzler

Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Résumé

Pour rappel, le devoir maison est facultatif et il a pour but de vous exercer en prévision des différents examens dans le cadre de cet enseignement.
Il n'est pas demandé de réaliser l'ensemble des exercices, je vous demande simplement de faire ce que vous pouvez faire.
Ce devoir maison comprend une partie théorique et une partie pratique à effectuer sur R. Vous pourrez ainsi me rendre votre travail en partie sur papier ainsi que votre code R que vous me transmettrez par mail.

Modèle Linéaire Multiple

Une entreprise de distribution exploite plusieurs centres de distribution régionaux et souhaite comprendre les facteurs expliquant les retards de livraison hebdomadaires observés dans son réseau logistique.

Pour chaque centre de distribution, les variables suivantes sont collectées :

- Y : **retard moyen hebdomadaire de livraison** (en heures)

Dans cette étude, on considère les variables explicatives suivantes :

- X_1 : **volume moyen de commandes quotidiennes** (en centaines de commandes)
- X_2 : **temps moyen d'arrêt des machines** par semaine (en heures)
- X_3 : **heures supplémentaires moyennes des employés** par semaine (en heures)
- X_4 : **taux de fiabilité des stocks** (en %, entre 0 et 100)
- X_5 : **nombre moyen de références (SKU) traitées** par semaine (en centaines)
- X_6 : **ancienneté moyenne des employés d'entrepôt** (en années)

À partir d'un échantillon de centres de distribution et en utilisant la méthode des moindres carrés ordinaires, on obtient le modèle de régression linéaire multiple suivant :

$$\hat{Y} = 18 + 1.5X_1 + 2.8X_2 + 0.6X_3 - 0.25X_4 + 1.2X_5 - 0.9X_6$$

Généralités et tests statistiques

1. Interpréter les coefficients associés à X_2 (temps d'arrêt des machines) ; X_4 (fiabilité des stocks) ; X_6 (expérience des employés).
2. Prédire le retard moyen hebdomadaire de livraison pour un centre de distribution ayant les caractéristiques suivantes : 6 centaines de commandes traitées par jour, 3 heures d'arrêt des machines par semaine, 15 heures supplémentaires par semaine, un taux de fiabilité des stocks de 96%, 4 centaines de références (SKU) traitées, une ancienneté moyenne des employés de 5 ans.

Nous nous intéressons maintenant aux coefficients estimés et à leurs propriétés présentées dans la Table 1.

4. Pour un coefficient générique β_j :
   — formuler les hypothèses nulle et alternative, définir la statistique de test ainsi que sa loi, et expliquer comment la p-valeur est utilisée pour conclure.
5. Compléter la Table 1 avec les valeurs manquantes de la colonne *t-value* et, pour la colonne *p-value*, indiquer si elle est inférieure ou supérieure à 0.05 en utilisant le fait qu'un quantile pertinent est 1.80.

Variable	Estimation	Erreur standard	t-value	p-value
Constante	18.0	4.5	4.00	0.001
X_1	1.5	0.6	...	0.028
X_2	2.8	0.7	4.00	...
X_3	0.6	0.4	1.50	...
X_4	-0.25	0.08	...	0.009
X_5	1.2	0.9	1.33	...
X_6	-0.9	0.5	...	0.09

TABLE 1 – Résultats de la régression linéaire

Indiquer quelles variables sont significatives et pourquoi au seuil $\alpha = 0.05$.

6. Un coefficient non significatif signifie-t-il nécessairement que la variable n'a aucune importance opérationnelle ?

7. Donner les définitions et interpréter le coefficient de détermination R^2 et le coefficient de détermination ajusté R^2_{adj}. Pourquoi R^2_{adj} est-il généralement préféré pour comparer des modèles ayant des nombres différents de variables explicatives ?

8. L'entreprise envisage de simplifier le modèle. Décrire une méthode permettant d'obtenir un tel modèle en se basant sur R^2_{adj}.

9. Expliquer comment détecter la multicolinéarité dans un modèle de régression linéaire multiple.

10. Supposons que les variables X_1 (volume de commandes) et X_5 (nombre de références) aient toutes deux un VIF supérieur à 15. Définir le facteur d'inflation de variance (VIF) et expliquer la procédure à suivre pour supprimer les variables redondantes à l'aide du VIF.

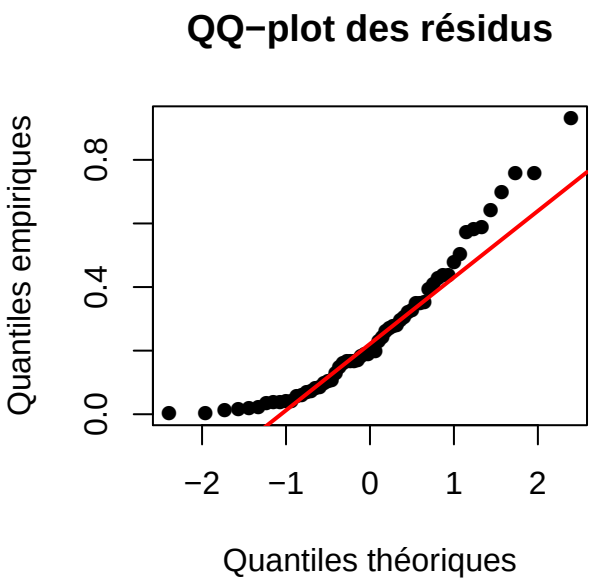
Analyse des résidus et détection des observations atypiques

L'entreprise souhaite vérifier si les hypothèses du modèle linéaire gaussien sont satisfaites.

On rappelle que le modèle explique le **retard moyen hebdomadaire de livraison** (Y) à partir de plusieurs facteurs opérationnels et humains. L'analyse est réalisée sur un échantillon de $n = 60$ centres de distribution.

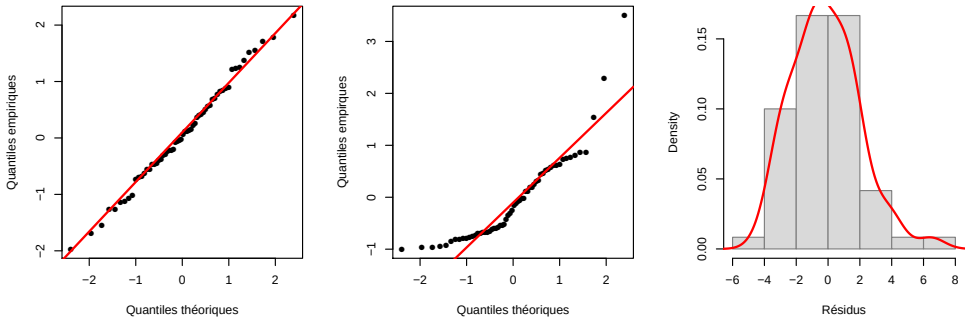
1. Un analyste trace un **QQ-plot des résidus**. Expliquer comment cet outil graphique permet de vérifier si les résidus suivent une loi normale.

2. Supposons que le QQ-plot des résidus ait la forme suivante :



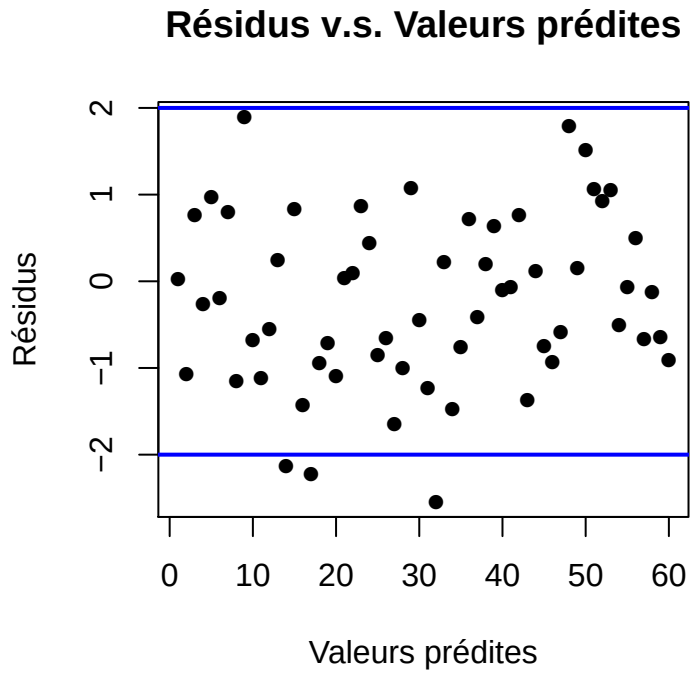
   Cette figure valide-t-elle l'hypothèse de normalité ? Justifier.

3. Parmi les graphiques suivants, indiquer lesquels correspondent à des résidus gaussiens.



L'analyste trace les **résidus en fonction des valeurs ajustées**. Quel type de motif indiquerait une violation de l'hypothèse d'homoscedasticité ?

Le graphique suivant est obtenu :



L'homoscedasticité semble-t-elle vérifiée ? Justifier et donner un exemple de situation contraire.

Proposer une solution en cas d'hétéroscédasticité dans ce contexte logistique.

L'analyste utilise la règle :

$$h_{ii} > \frac{2p}{n}$$

1. Calculer le seuil pour $p = 6$ et $n = 60$.

2. Une observation a $h_{ii} = 0.25$. Est-elle influente ? Justifier.

Expliquer pourquoi il peut être pertinent de supprimer ces observations, ainsi que les limites de cette approche.

Etude des facteurs permettant d'expliquer la réponse immunitaire

Pour cette analyse pratique, à effectuer sous R, on se servira du jeu de données qui se trouvent dans le fichier **biology.csv**. A l'aide de ces données, on cherche à modéliser la réponse immunitaire, mesurée par le nombre de lymphocytes circulants (en *cellules/*μL) dans notre corps, en fonction de facteurs, physiologiques, inflammatoires, hormonaux et environnementaux.
On donne ci-dessous une brève description des différentes variables de notre jeu de données :

- LYMPHOCYTE_COUNT : nombre de lymphocytes circulants (en *cellules/*μL)
- AGE : age de l'individu en années. Le vieillissement est en général associé à une immunoscénescence progressive, *i.e.*, une baisse du système immunitaire.
- BMI : indice de masse corporelle (en kg/m) des individus.
- CRP : Protéine C-réactive (en mg/L). Il s'agit d'un marqueur systémique de l'inflammation.
- IL6 : Interleukine 6 (en pg/mL). Il s'agit d'une cytokine (substance sécrétée par les cellules immunitaires favorisant la communication entre elles, elles participent à l'activation de la réponse immunitaire) pro-inflammatoire stimulant la production de CRP
- VITAMIN_D : taux de vitamine D dans le sang (en ng/L).
- CORTISOL : hormone du stress (en μg/dL). Le stress a des effets immunosuppresseurs sur le long terme.
- SLEEP_DURATION : nombre d'heures de sommeil par nuit (heure).
- HYDRATATION : quantité d'eau absorbée par jour (en L).
- INFECTION_STATUS : type d'infection présenté par l'individu ACUTE pour infection aigüe ou CHRONIC pour une infection chronique.

L'objectif de cet exercice est de trouver le meilleur modèle possible pour expliquer la réponse immunitaire à l'aide des observations. Pour cela, vous devrez reprendre les étapes présentées en cours et TD afin de déterminer quelles sont les variables significatives les plus importantes en vous basant sur un critère étudié. On rappelle ci-dessous les éléments clefs de la construction d'un tel modèle :

- Construire un modèle utilisant l'ensemble des informations comme modèle de base.
- Détecter s'il y a des informations redondantes.
- A l'aide d'une méthode *itérative* déterminer quelles sont les variables les plus importantes à inclure dans le modèle.
- Une fois le modèle trouvé, analyser les sorties (impact des variables, qualité du modèle, *etc.*).
- Validité du modèle.
- Présence d'outliers.

Il est attendu que les étapes soient présentées avec les différentes sorties.
Dans un deuxième temps on cherchera ensuite à enrichir notre modèle à l'aide de nouvelles variables qui sont des transformations ou combinaisons de l'ensemble des variables explicatives X_j afin de voir s'il est possible d'améliorer les prédictions effectuées par notre modèle.