Modèles Linéaires

Contrôle - 1h30
Correction
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.

   Le modèle linéaire gaussien se présente sous la forme suivante :

   $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon,$$

   où β_j représentent les coefficients de la régression, *i.e.* l'impact de chaque variable X_j sur les valeurs de la variable Y et ε représente un bruit ou les erreurs de modélisation.
   Nous formulons les hypothèses suivantes pour notre modèle linéaire gaussien :

   (a) $(Y_i, X_i)_{i=1}^n$ doivent être *i.i.d*, *i.e.* indépendantes et identiquement distribuées,
   (b) $Y_i \underset{i.i.d.}{\sim} \mathcal{N}(\beta_0 + \beta_1 X_i +, \sigma^2)$
   (c) $\varepsilon_i \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$: hypothèse d'homoscédasticité.

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 :

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

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

   L'estimateur $\hat{\boldsymbol{\beta}}$, obtenu par maximum de vraisemblance est donné par

   $$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}.$$

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

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

$$
\begin{aligned}
\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top Y\right], \\
&\quad \downarrow \text{ seule } \mathbf{y} \text{ est aléatoire} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathbb{E}\left[\mathbf{y}\right], \\
&\quad \downarrow \text{ par définition de } \mathbf{y} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathbb{E}\left[\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right], \\
&\quad \downarrow \text{ seule } \boldsymbol{\varepsilon} \text{ est aléatoire} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top X\boldsymbol{\beta}+\mathbb{E}\left[\boldsymbol{\varepsilon}\right], \\
&\quad \downarrow \text{ car } \mathbb{E}\left[\boldsymbol{\varepsilon}\right]=\mathbf{0}\text{, hypothèse sur les erreurs du modèle} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}, \\
\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] &= \boldsymbol{\beta}.
\end{aligned}
$$

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

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

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

$$
\begin{aligned}
\mathrm{Var}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}\left[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top\right], \\
&\quad \downarrow \text{ On repart de la définition de } \hat{\beta} \\
&= \mathbb{E}\left[((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}-\boldsymbol{\beta})((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}-\boldsymbol{\beta})^\top\right], \\
&\quad \downarrow \text{ définition de } Y \\
&= \mathbb{E}\left[((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})-\boldsymbol{\beta})((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon})-\boldsymbol{\beta})^\top\right], \\
&\quad \downarrow \text{ on développe et on simplifie les expressions} \\
&= \mathbb{E}\left[((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon})((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon})^\top\right], \\
&\quad \downarrow \text{ par définition de la transposition} \\
&= \mathbb{E}\left[(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\right], \\
&\quad \downarrow \text{ seule la partie en } \boldsymbol{\varepsilon} \text{ est aléatoire} \\
&= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right]\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}, \\
&\quad \downarrow \text{ or } \mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right]=\sigma^2\mathbf{I}_n \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1},
\end{aligned}
$$

$$\downarrow \text{ par simplification}$$
$$\mathrm{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}.$$

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 $\hat{\boldsymbol{\varepsilon}}$.

On se rappelle que les résidus sont définis par

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}.$$

Il ne reste plus qu'à faire le calcul en utilisant la définition de $\hat{\boldsymbol{\beta}}$.

$$\begin{aligned}
\mathbf{X}^\top\hat{\boldsymbol{\varepsilon}} &= \mathbf{X}^\top(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \\
&\qquad \downarrow \text{ on utilise la définition de } \hat{\boldsymbol{\beta}} \\
&= \mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}, \\
&= \mathbf{X}^\top\mathbf{y} - \mathbf{X}^\top\mathbf{y}, \\
&= \mathbf{0}
\end{aligned}$$

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 $\hat{\boldsymbol{\beta}}$:

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

On rappelle que

$$\mathrm{Cov}[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\varepsilon}}] = \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \mathbb{E}\,\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\varepsilon}} - \mathbb{E}\,\hat{\boldsymbol{\varepsilon}})^\top\right].$$

On rappelle aussi que l'on a

$$\hat{\boldsymbol{\varepsilon}} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{y} = \mathbf{L}\mathbf{y},$$

où $\mathbf{L}$ est une matrice de projection. Il suffit maintenant de repartir de la définition de la covariance et de faire la calcul.

$$
\begin{aligned}
\mathrm{Cov}[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\varepsilon}}] &= \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}])(\hat{\boldsymbol{\varepsilon}} - \mathbb{E}[\hat{\boldsymbol{\varepsilon}}])^\top\right], \\
&\quad \downarrow\ \hat{\boldsymbol{\beta}} \text{ est un estimateur non biaisé et } \mathbb{E}\,\hat{\boldsymbol{\varepsilon}} = \mathbf{0} \\
&= \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}])(\hat{\boldsymbol{\varepsilon}} - \mathbb{E}[\hat{\boldsymbol{\varepsilon}}])^\top\right], \\
&= \mathbb{E}[\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\varepsilon}}^\top] - \underbrace{\mathbb{E}[\boldsymbol{\beta}\hat{\boldsymbol{\varepsilon}}^\top]}, \\
&\quad \downarrow\ \mathbb{E}[\hat{\boldsymbol{\varepsilon}}] = \mathbf{0} \\
&= \mathbb{E}[\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\varepsilon}}^\top] - \mathbf{0}, \\
&\quad \downarrow\ \text{définition de } \hat{\boldsymbol{\beta}} \text{ et } \mathbf{L} \text{ est une projection} \\
&= \mathbb{E}[((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)\mathbf{y}\mathbf{y}^\top\mathbf{L}], \\
&\quad \downarrow\ \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\
&= \mathbb{E}[((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})^\top\mathbf{L}], \\
&= \mathbb{E}[((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)(\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{L} + \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\mathbf{L})], \\
&\quad \downarrow\ \text{il reste à simplifier l'expression et on utilise } \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \sigma^2\mathbf{I}_n \\
&= \boldsymbol{\beta}\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{L} + \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{L}, \\
&\quad \downarrow\ \text{on utilise la propriété de la transposition} \\
&= \underbrace{\boldsymbol{\beta}\boldsymbol{\beta}^\top(\mathbf{L}\mathbf{X})^\top} + \underbrace{(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{L}\mathbf{X})^\top}, \\
&\quad \downarrow\ \text{car } \mathbf{L}\mathbf{X} = \mathbf{0} \\
\mathrm{Cov}[\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\varepsilon}}] &= \mathbf{0}.
\end{aligned}
$$

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.

   Ces deux coefficients représentent l'impact de ces deux variables sur la variable Y. plus précisément :

- Variable X_1 : on note un impact positif de cette variable sur Y, plus précisément, pour chaque heure d'études, le score obtenu à l'examen augmente de deux points.
- Variable X_2 : cette variable est corrélée négativement à la variable Y. On peut même dire que chaque passé sur les réseaux sociaux entraîne une chute du score de trois points.

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 va simplement appliquer le modèle à cette nouvelle observation

$$\hat{Y} = 10 + 2 \times 5 - 3 \times 4 + 0.5 \times 20 = 18.$$

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. Enoncer les hypothèses, la statistique de test et sa distribution ainsi que la façon de conclure à ce test au seuil de significativité $\alpha \in [0, 1]$.

   Pour tester la significativité d'un paramètre β_j du modèle de régression, on formule les hypothèses suivantes :

$$H_0 : \ \beta_j = 0 \text{ v.s. } H_1 : \ \beta_j \neq 0$$

   .

   Il s'agit d'un test bilatéral qui repose sur la statistique de test est la suivante

$$t_{\text{test}} = \frac{\hat{\beta}_j}{\hat{\sigma}_{\beta_j}},$$

   où $\hat{\sigma}_{\beta_j} = \dfrac{(\mathbf{X}^\top \mathbf{X})^{-1}_{j+1,j+1}}{n} \sum_{i=1}^{n} \varepsilon_i^2$.

   Cette statistique de test suit une loi de Student à $n - p - 1$ degrés de libertés. On rejette en suite l'hypothèse H_0 si

(a) $|t_{\text{test}}| \geq t_{1-\alpha/2}(n-p-1)$, qui est la quantile d'ordre $1-\alpha/2$ d'une loi de Student à $n-p-1$ degrés de libertés.
OU

(b) si $2\mathbb{P}[T \geq |t_{\text{test}}|] \leq \alpha$, ou T est une variable aléatoire qui suit une loi de Student à $n-p-1$ degrés de libertés.

4. Au seuil de significativité $\alpha = 0.05$, déterminer les variables significatives du modèle.

   D'après le tableau fourni, seules les variables X_1 et X_2 sont significatives. Mais, avec une p-value égale à $0.32 > \alpha = 0.05$, la variable X_3 n'est pas significative.

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^2_{aj} ?

   Oui, il est légitime de retirer la variable non significative du modèle dans la mesure ou c'est la seule variable non significative. En revanche, si plusieurs variables sont non signficative, il vaut mieux les retirer une par une et relancer une régression à chaque fois. En effet, il est possible qu'une variable initialement non significative, devienne significative lorsque l'on en supprime une autre.
   En terme de R^2, un modèle avec moins de variable ne pourra pas avoir un R^2 plus élevé, ce dernier sera au plus égal au R^2 du modèle complet. En revanche, le R^2 peut vraisemblablement augmenter.

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.

   La présence de multi-colinéarité dans un modèle de régression se fait à l'aide du calcul du VIF. Le VIF se calcule uniquement à partir des variables indépenddantes du modèle en faisant des modèles de régression de la forme

   $$X_j \sim xX_1 + X_2 + \ldots X_{j-1} + X_{j+1} + \ldots + X_p + \varepsilon.$$

   On calcule ensuite le R^2, associé au modèle cherchant à prédire les valeurs de la variable X_j, noté R^2_j, par

   $$VIF(X_j) = \frac{1}{1-R^2_j}.$$

Un seuil de 5 (ou 10) est ensuite appliqué pour déterminer si la variable est colinéaire à une autre.

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.

   La première partie est décrite à la question précédente. Dans le cas présent, on commence par supprimer la variable ayant le VIF le plus élevé, puis on recalcule les VIF avec les variables restantes. Le processus est répété jusqu'à ce que toutes les variables aient un VIF inférieur à 5 (ou 10).

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 :

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

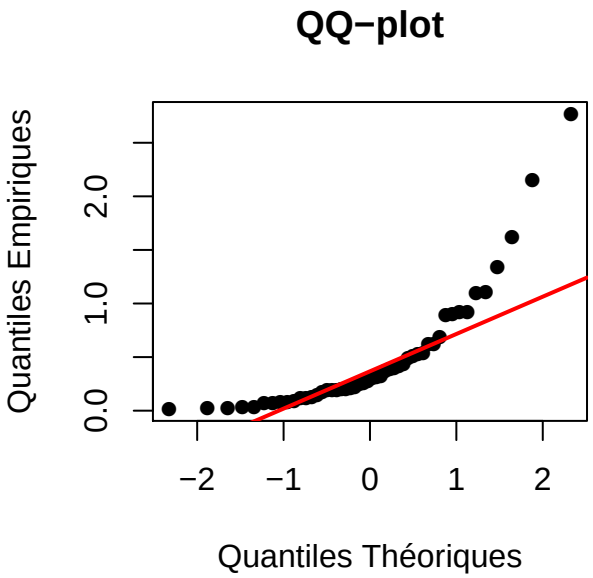
On suppose que les résidus doivent suivre une loi normale avec variance constante. L'étude est menée sur un échantillon de $n = 50$ observations.

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 ?

   Pour cela ils doivent vérifier que les points de ce graphe sont alignés sur la droite, dite Droite de Henry, si tel est le cas, l'hypothèse de normalité n'est pas rejetée. Si trop de points dévient de cette droite, on rejette l'hypothèse.

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.

La courbe des résidus empiriques vs résidus théoriques ne forme pas une droite, *i.e.*, les points sont non alignés, donc on peut conclure que l'hypothèse de normalité des résidus n'est pas vérifiée.

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

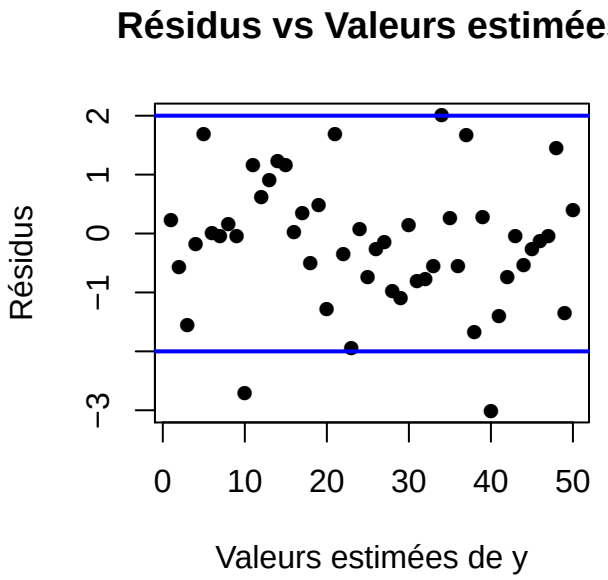
   Le test de Shapiro-Wilk.

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'homoscedasticité ?

   Une représentation graphique sous forme de cône, *i.e.*, si les valeurs des résidus studentisés ont tendance à augmenter ou à diminuer en fonction des valeurs prédites $\hat{y}$.

5. Le chercheur obtient le graphique ci-dessous :



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

Les résidus semblent avoir la même valeur peu importe la valeur de $\hat{y}$, donc cette hypothèse n'est pas contredite par le graphique.

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

Dans le cas où cette hypothèse n'est pas vérifiée, il conviendrait d'effectuer une analyse plus poussée afin de voir s'il ne faudra pas effectuer une transformation de la variable Y ou des variables X_j.

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.

(a) Déterminez le **seuil** pour détecter les valeurs aberrantes si $p = 2$ et $n = 50$.

La seuil en question est de $1/25$.

(b) Une observation a $h_{ii} = 0.15$. Est-elle influente ? Justifiez.

Comme nous avons $h_{ii} > 1/25$, la donnée considérée est donc une valeur aberrante/influente.

8. Quel autre graphique vous permettrait de détecter d'éventuelles valeurs aberrantes ?

Nous pourrions également nous concentrer sur le graphe résidus studentisés v.s. valeurs prédites $\hat{y}$. Nous pourrions considérer comme valeurs aberrantes les points pour lesquels les résidus studentisés sont en dehors de l'intervalle $[-2, 2]$.