



Modèles Linéaires

Devoir Maison Licence 3 MIASHS (2024 - 2025)

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


Ce devoir maison comprend une partie théorique et une partie pratique sur le modèle linéaire multiple mais aussi sur quelques calculs d'intégrales pour celles et ceux qui souhaitent aller plus loin.

Il n'est pas demandé de réaliser l'ensemble des exercices, il est là pour que vous vous exerciez mais aussi pour que vous puissiez vous challenger sur des exercices plus difficiles.

Le travail est à rendre pour le jeudi 6 mars.

Modèle linéaire multiple

La première partie de cet exercice est à effectuer à la main, les résultats seront arrondis à deux chiffres après la virgule. Pour être prise en compte, vous devrez indiquer l'ensemble des valeurs qui servent à effectuer les calculs. Par exemple, on pensera à effectuer le calcul des résidus qui peuvent servir à évaluer la qualité du modèle. Vous pourrez vérifier vos calculs à l'aide du logiciel .

La deuxième partie est à effectuée directement sous . Il est demandé de décrire les étapes suivies sur votre copie et préciser les valeurs clés utilisées pour conclure.

Etude d'un jeu de données synthétiques

On considère le jeu de données suivant :

Y	-3	5	7	6	-2	4	2	6	-1	10
X_1	2	-1	3	-2	1	-1	2	-2	1	-1
X_2	1	-1	2	-2	3	-3	-1	1	-2	2


Y représente la variable à prédire/dépendante et X_1 et X_2 sont les deux variables indépendantes du modèle.

On considère le modèle de régression multiple suivant

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

1. Rappeler les hypothèses du modèle linéaire gaussien.
2. Dans le cas présent, définir les objets \mathbf{y} , \mathbf{X} , $\boldsymbol{\beta}$ et $\boldsymbol{\varepsilon}$ ainsi que leurs dimensions.
3. On considère le problème d'optimisation


$$\min_{\boldsymbol{\beta} \in \mathbb{R}^3} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2.$$

- (a) Donner l'expression littérale de $\hat{\boldsymbol{\beta}}$, l'estimateur de $\boldsymbol{\beta}$.
 - (b) Calculer sa valeur numérique, à la main. On donnera un résultat précis au centième près, soit deux chiffres après la virgule.
4. Estimer la variance des résidus du modèle.
 5. A l'aide du logiciel .
 - (a) Tester la significativité des paramètres du modèle. On prendra le soin de définir le test employé, la définition et la valeur de la statistique de test ainsi que la distribution associée à ce test.

- (b) Comment tester si le modèle est globalement significatif? Définir le test, calculer sa valeur et conclure quant à la significativité du modèle.
- 6. On s'intéresse maintenant à la qualité du modèle.
 - (a) Calculer le coefficient de détermination R^2 du modèle de régression construit.
 - (b) En déduire la valeur du coefficient de détermination ajusté R^2_{aj} .
 - (c) Calculer le BIC du modèle.

Construction d'un modèle

Dans cette étude, on travaillera à l'aide du fichier *insurance.csv* (ce dernier se trouve dans le répertoire *Datasets*). Une assurance étudie le coût associé à chaque assuré (**expenses**) en fonction de diverses caractéristiques afin de déterminer une nouvelle valeur du *premium* (une franchise) à appliquer à ses nouveaux assurés lors de la nouvelle souscription d'un contrat.

1. Effectuer une régression linéaire et décrire les résultats obtenus : (i) Significativité globale du modèle, (ii) qualité du modèle et (iii) les variables significatives du modèle.
2. En utilisant la méthodologie présentée en cours, construire le meilleur modèle possible sur la base d'un critère de performance que vous préciserez.
Vous pourrez regarder les fonctions suivantes sur  qui vous permettront de faire quelques étapes pour vous : **vif** et **step**.
3. Analyser les résidus du modèle construit afin de vous assurer que les hypothèses du modèle linéaire sont bien vérifiées.

Autour des lois de probabilités

Exercice 1 : Calcul de l'intégrale de Gauss.

L'objectif de cet exercice est de calculer l'intégrale $I = \int_{-\infty}^{+\infty} e^{-x^2} dx$.

1. Montrer que l'on a

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-x^2+y^2} dx dy = I^2.$$

2. On considère le changement de variable polaire $x = r \cos(\theta)$ et $y = r \sin(\theta)$, *i.e.*, on considère l'application φ définie par

$$\varphi(r, \theta) = (r \cos(\theta), r \sin(\theta)).$$

Déterminer l'image de départ et d'arrivée de la fonction φ afin que la fonction soit bijective.

3. On considère maintenant la matrice $J_{\varphi(r,\theta)} = \begin{pmatrix} \frac{\partial x}{\partial r} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial r} & \frac{\partial y}{\partial \theta} \end{pmatrix}$.

En utilisant le fait que l'on a

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} e^{-x^2-y^2} dx dy = \int_{I_r} \int_{I_\theta} e^{-r^2(\cos^2(\theta)+\sin^2(\theta))} |\det(J_{\varphi(r,\theta)})| dr d\theta.$$

Calculer la valeur de l'intégrale I^2 et en déduire la valeur de I . Il faudra d'abord déterminer les bornes des intervalles sur lesquels on intègre, *i.e.*, I_r et I_θ .

Exercice 2 : Loi du χ^2

Dans cet exercice, on souhaite montrer que la variable aléatoire $X = \sum_{j=1}^n X_j^2$ où $X_j \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$, suit une loi du χ^2 à n degrés de libertés.

Pour cela, on admettra que la densité d'une loi du χ^2 à n degrés de liberté est donnée, pour tout $x > 0$, par :

$$f_n(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2},$$

où Γ est la fonction d'Euler définie par $\Gamma(t) = \int_0^{+\infty} x^{t-1} e^{-x} dx$.

En probabilités, pour montrer que deux variables aléatoires U et V ont la même loi, on peut montrer qu'elles ont la même fonction caractéristique, *i.e.*, on montre que

$$\forall t \in \mathbb{R}, \mathbb{E}[e^{itU}] = \mathbb{E}[e^{itV}].$$

On rappelle que l'espérance d'une variable Z aléatoire à densité f , dans le cas où Z admet un moment d'ordre 1, est donnée par

$$\mathbb{E}[Z] = \int_{-\infty}^{+\infty} z f(z) dz.$$

1. Pour tout $j \in \llbracket 1, n \rrbracket$, montrer que l'on a

$$\mathbb{E}[e^{itX_j^2}] = (1 - 2it)^{-1/2}.$$

2. En utilisant le fait que la fonction caractéristique de la somme de n variables aléatoires indépendantes est égale à la somme de ces n fonctions caractéristiques, montrer que l'on

$$\mathbb{E}[e^{itX}] = (1 - 2it)^{-n/2}.$$

3. En déduire qu'une variable aléatoire suivant une loi du U suivant une loi du χ^2 et la variable aléatoire ont la même loi.
4. Déterminer l'espérance d'une loi du χ^2 à 1 puis à n degrés de liberté.

Exercice 3 : Résultats complémentaires sur le modèle multiple

L'objectif de cet exercice est de montrer plusieurs petits résultats portant sur le modèle linéaire simple ou multiple.

1. On considère le modèle linéaire simple

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

A l'aide des définitions du coefficient de corrélation linéaire $\hat{\rho}$ et du coefficient de détermination R^2 , montrez que l'on a

$$R^2 = \hat{\rho}^2.$$

2. Considérons deux modèles : un modèle M_1 qui utilise des variables X_1, \dots, X_p et un modèle M_2 qui utilise les variables X_1, \dots, X_p et X_{p+1} . Expliquer pourquoi on a

$$R^2(M_1) \leq R^2(M_2).$$

Théorème de Cochran

Le but de cet exercice est de démontrer une version simplifiée du Théorème de Cochran, en travaillant avec un vecteur gaussien centré et réduit.

Théorème: S

Soit $X = (X_1, X_2, \dots, X_p)$ un vecteur gaussien centré et réduit. Soit F un sous-espace vectoriel de \mathbb{R}^n de dimension p , on note \mathbf{P}_F et \mathbf{P}_{F^\perp} les projections orthogonales sur les sous-espaces F et F^\perp respectivement. Alors les vecteurs aléatoires $\mathbf{P}_F X$ et $\mathbf{P}_{F^\perp} X$ sont gaussiens indépendants de lois :

$$\mathbf{P}_F X \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_F) \quad \text{et} \quad \mathbf{P}_{F^\perp} X \sim \mathcal{N}(\mathbf{0}, \mathbf{P}_{F^\perp}).$$

De plus, les variables aléatoires $\|\mathbf{P}_F X\|^2$ et $\|\mathbf{P}_{F^\perp} X\|^2$ sont indépendantes de lois :

$$\|\mathbf{P}_F X\|^2 \sim \chi_p^2 \quad \text{et} \quad \|\mathbf{P}_{F^\perp} X\|^2 \sim \chi_{n-p}^2.$$

On peut voir une vision du théorème de Pythagore quant à ce théorème en ce qui concerne la norme des objets mais surtout ... au niveau des lois !

- Supposons que l'on dispose d'une base orthonormée $(\mathbf{u}_1, \dots, \mathbf{u}_p)$ de F et $(\mathbf{u}_{p+1}, \dots, \mathbf{u}_n)$ de F^\perp et notons \mathbf{U} la matrice de passage de la base canonique vers la base $(\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n)$. Donner les expressions de \mathbf{P}_F et \mathbf{P}_{F^\perp} .
- Posons $Y = \mathbf{U}^\top X$. Montrer que le vecteur gaussien X est toujours de moyenne nulle et de matrice de variance égale à l'identité, *i.e.*, qu'il est gaussien centré et réduit¹.
- En déduire que $\mathbf{I}_p Y$ et $\mathbf{I}_{n-p} Y$ sont indépendants, ainsi que leur lois, sachant que :

$$\mathbf{I}_p = \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 0 \end{pmatrix} \leftarrow p \quad \text{et} \quad \mathbf{I}_{n-p} = \begin{pmatrix} 0 & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ \vdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & 0 & 1 \end{pmatrix} \leftarrow p+1$$

- En utilisant la définition d'une loi de χ^2 , montrer alors que

1. On utilisera le fait que $\text{Var}[\mathbf{A}X] = \mathbf{A}X X^\top \mathbf{A}^\top$.

$$\|\mathbf{I}_p Y\| \sim \chi_p^2 \quad \text{et} \quad \|\mathbf{I}_{n-p} Y\| \sim \chi_{n-p}^2.$$

5. En exploitant le lien entre les vecteurs gaussiens X et Y , en déduire

$$\|\mathbf{P}_F X\| \sim \chi_p^2 \quad \text{et} \quad \|\mathbf{P}_{F^\perp} X\| \sim \chi_{n-p}^2.$$