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

Il n'est pas demandé de réaliser l'ensemble des exercices, je vous demande simplement de faire ce que vous pouvez.

Autour du modèle linéaire gaussien

On suppose que l'on dispose d'un échantillon $S = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$ où $y_i \in \mathbb{R}$ et $\mathbf{x}_i \in \mathbb{R}^p$, où $p > 1$ représente la dimension de notre jeu de données. Notre objectif est de déterminer une relation linéaire entre les valeurs observées y_i et les caractéristiques des individus $\mathbf{x}_i$. Pour cela, on considère le modèle suivant :

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

où $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ est la matrice de *design*, $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ et $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ est notre vecteur des résidus ou erreurs du modèle. On suppose que les nos erreurs suivent une distribution normale de moyenne nulle et de variance inconnue σ^2.

On rappelle que le vecteur $\boldsymbol{\beta}$ est solution du problème suivant :

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

Une autre façon d'obtenir cette solution est de procéder par **maximum de vraisemblance**.

1. Déterminer la vraisemblance de l'échantillon S.
2. Déterminer les expressions de $\boldsymbol{\beta}$ et σ^2 par maximum de vraisemblance dans le cas où $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.

Autour de la loi gaussienne

Les questions de cette section sont facultatives. Elles nécessitent de connaître le calcul d'intégrale simple et multiple.

Soit X une variable aléatoire distribués selon une loi normale de moyenne μ et de variance σ^2 dont notera f la densité. On rappelle que l'espérance d'une variable X est définie par

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f(x)\, dx$$

et le *moment d'ordre 2* qui sert à définir la variance est donnée par

$$\mathbb{E}[X^2] = \int_{\mathbb{R}} x^2 f(x)\, dx.$$

3. Montrer que l'on a $\mathbb{E}[X] = \mu$.
4. Montrer que l'on a $\mathbb{E}[X] = \sigma^2$.

Etude des résidus

NOus avons supposé que les erreurs de notre modèle sont centrées. Notons à présent $\hat{\varepsilon}_i$ les résidus associées à la i-ème observation.

5. En utilisant le formulation du problème par **moindres carrés ordinaires** , montrer que l'on a

$$\sum_{i=1}^{n} \hat{\varepsilon}_i = 0.$$

Variantes du modèles gaussien

Dans cette section on va regarder deux variantes du modèle linéaire gaussien : *(i)* on remet en cause l'hypothèse d'homoscédasticité et *(ii)* en supposant que les individus $\mathbf{x}_i$ n'ont pas le même poids lors de l'estimation des paramètres du modèle.

(i) Remise en cause de l'homoscédasticité On suppose que l'hypothèse $\mathrm{Var}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}$ n'est plus vérifiée mais que l'on cette fois ci $\mathrm{Var}[\boldsymbol{\varepsilon}] = \sigma^2\boldsymbol{\Sigma}$, où la matrice $\boldsymbol{\Sigma} \in \mathbb{R}n \times n$ est connue.

6. Déterminer l'estimateur obtenu par **MCO** en tenant compte de cette nouvelle hypothèse.

(ii) Pondération des individus On suppose maintenant que chque individu a un poids différent dans l'estimation des paramètres du modèle. On notera w_i la pondération de l'exemple $\mathbf{x}_i$. Notre prolbème de minimisation peut alors se réécrire

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} w_i(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}).$$

7. Déterminer l'estimateur obtenu par **MCO** en tenant compte de cette nouvelle hypothèse.