Modèles Linéaires

TD 1 : Régression linéaire simple
Licence 3 MIASHS - Informatique

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

Cette première séance aborde généralement le modèle linéaire gaussien simple, et plus précisément les points suivants :

- Rappel sur les hypothèses du modèle linéaire gaussien,
- Estimation des paramètres du modèle par MCO, dans le cas modèle linéaire simple,
- Ecriture du modèle sous forme matricielle
- Estimation par maximum de vraisemblance

Modèle de régression simple

On rappelle que le modèle linéaire gaussien simple s'écrit sous la forme

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

où Y est la variable à expliquer, X est la variable explicative et ε est une variable aléatoire représentant les erreurs du modèle.

Dans ce TD, on se concentre sur l'estimation des paramètres du modèle à l'aide de plusieurs méthodes différentes et nous appliquerons ensuite cela sur les données présentées ci-dessous, pour obtenir la droite de régression associée, présentée en Figure 1.

Y : Score examen 2	3.5	4	5	1	2	1.5	2.5	5.5	6	6.5
X : Score examen 1	4	3	3.5	1	1.5	1	1.5	4	3.5	4.5

1. Rappeler les hypothèses du modèle de régression linéaire gaussien.

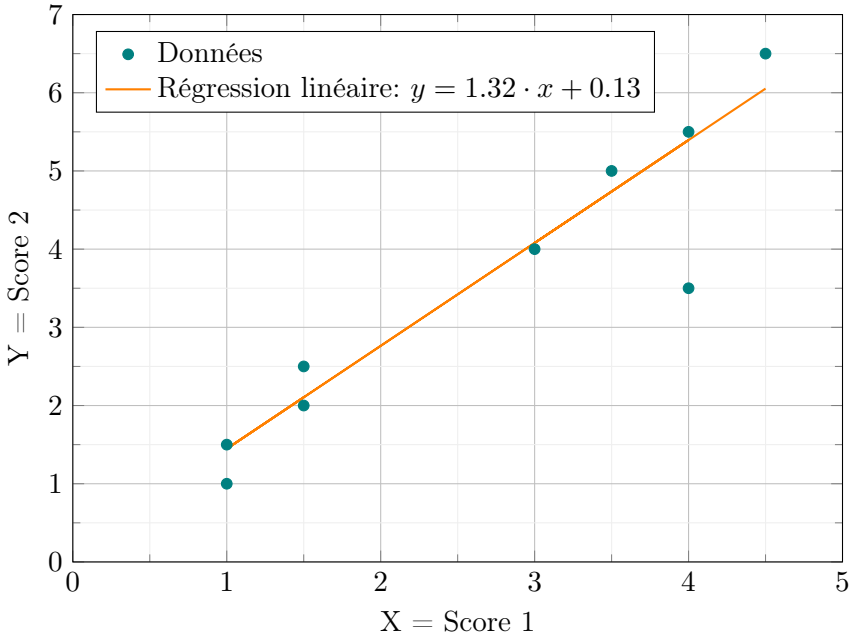

FIGURE 1 – Application de la régression linéaire simple gaussien sur les données présentées dans la table associée. On cherche alors à expliquer le score obtenu au deuxième examen en fonction du score obtenu au premier examen.

Estimation par méthode des moindres carrés

La méthode des moindres carrés ordinaires consiste à minimiser le carré des résidus du modèle de régression, *i.e.* minimiser l'écart quadratique entre la prédiction effectuée par le modèle $\hat{\beta}_0 + \hat{\beta}_1 x_i$ et la valeur observée y_i.
Pour cela on disposera d'un échantillon de données $S = \{(x_i, y_i)\}_{i=1}^m$ et on s'intéresse au problème de minimisation :

$$\min_{\beta_0,\beta_1 \in \mathbb{R}} \sum_{i=1}^{n} \varepsilon_i = \min_{\beta_0,\beta_1 \in \mathbb{R}} \sum_{i=1}^{n} y_i - \hat{y}_i = \min_{\beta_0,\beta_1 \in \mathbb{R}} \sum_{i=1}^{n} y_i - (\beta_0 + \beta_1 x_i) \tag{1}$$

2. Dans la résolution de ce problème, l'hypothèse de normalité sur les résidus ε_i est-elle importante ?
3. Montrer que les solutions du problème 1 sont données par

$$\hat{\beta}_1 = \frac{\mathrm{Cov}[X, Y]}{\mathrm{Var}[X]} \quad \text{et} \quad \hat{\beta}_0 = \mathbb{E}[Y] - \hat{\beta}_1 \, \mathbb{E}[X].$$

4. Effectuer l'application numérique avec les données de l'énoncé.

Cette façon de résoudre le problème convient très bien lorsque la modélisation implique une seule variable explicative mais est moins pratique lorsque l'on dispose de

plusieurs descripteurs. Il convient alors de passer sous une écriture matricielle de ce même problème de minimisation.

5. Montrer que le problème d'optimisation 1 peut se réécrire sous la forme

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

   où $\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}$ sont des objets dont on explicitera les définition et dimension.
6. Montrer que ce problème d'optimisation est convexe
7. Déterminer les solutions du problème et préciser à quelle condition cette solution existe.
8. Retrouver l'expression des solutions obtenues en question 3. à l'aide de cette expression matricielle.
9. Effectuer l'application numérique avec les données de l'énoncé.

Etant donnée l'hypothèse fomulée sur la distribution de la variable aléatoire Y, nous pouvons également estimer les paramètres du modèle par maximum de vraisemblance.

Estimation parpar maximum de vraisemblance

Les données y_i suivent une distribution de normale de moyenne $\beta_0 + \beta_1 x_i$ et de variance inconnue σ^2.

1. Donner la valeur de la densité de la variable aléatoire Y.
2. Si on considère à nouveau notre échantillon S, donner l'expression de la vraisemblance de cet échantillon.
3. Estimer les valeurs de β_0 et de β_1 par maximum de vraisemblance.
4. Estimer la variance des résidus σ^2 à l'aide de cette même méthode.