Modèles Linéaires

Correction Séance 2
Licence 3 MIASHS (2022-2023)

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 deuxième séance se concentre sur le modèle linéaire gaussien multiple :

- Ecriture du modèle sous forme matricielle,
- Explication de l'estimation par MCO pour cette version plus générale (au sens de la dimension du modèle),
- Exercice 1 du TD 1 sous forme matricielle,
- Estimation des paramètres du modèle par maximum de vraisemblance et évaluation de la qualité du modèle à l'aide du BIC.

Exercice 1 : Transition vers le modèles multiple

Notations matricielles

L'objectif de cet exercice est d'opérer la transition entre le modèle linéaire simple et le modèle linéaire multiple par le biais de la manipulation de matrices et de vecteurs. Pour cela on rappelle que notre modèle linéaire simple peut s'écrire sous la forme

$$\forall i \in [\![1, n]\!],\ y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

où

- y est la variable réponse (ou la variable à prédire)
- x est la variable prédictive
- β_0 et β_1 sont les paramètres du modèles
- ε est un bruit blanc gaussien qui représente l'erreur de modélisation.

.

Ce modèle peut facilement se réécrire sous la forme suivante

$$Y = X\boldsymbol{\beta} + \varepsilon,$$

avec les notations vectorielles/matricielles suivantes

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n,\ \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \in \mathbb{R}^2,\ X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \in \mathscr{M}_{n,2}(\mathbb{R}),\ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \in \mathbb{R}^n.$$

Les objets gardent la même signification que dans l'écriture "standard". En revanche, cette notation sera beaucoup plus simple à manipuler dans le cadre du modèle linéaire multiple (multiple car il fera intervenir plusieurs variables prédictives !) que l'on peut donc écrire

$$\begin{aligned} \forall i \in [\![1, n]\!],\ y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i, \\ &= \sum_{k=0}^{p} \beta_k x_{ki} + \varepsilon_i. \end{aligned}$$

Ou encore, de façon plus condensée, à l'aide des notations vectorielles/matricielles

$$Y = X\boldsymbol{\beta} + \varepsilon.$$

Il faut admettre que cette dernière écriture est bien plus élégante.

Expression des solutions

Notre objectif est de démontrer le résultat suivant

Proposition 0.1: Solution du modèle linéaire (multiple)

Placons-nous sous les hypothèses du modèle linéaire gaussien vues lors de la précédente séance et considérons le modèle :

$$Y = X\boldsymbol{\beta} + \varepsilon.$$

Si on dispose d'un n-échantillon $(y_i, x_i)_{i=1}^n$ d'individus indépendants alors la solution du problème des *moindres carrés ordinaires*, *i.e.* du problème

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \|Y - \hat{Y}\|_2^2.$$

est donnée par

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T Y.$$

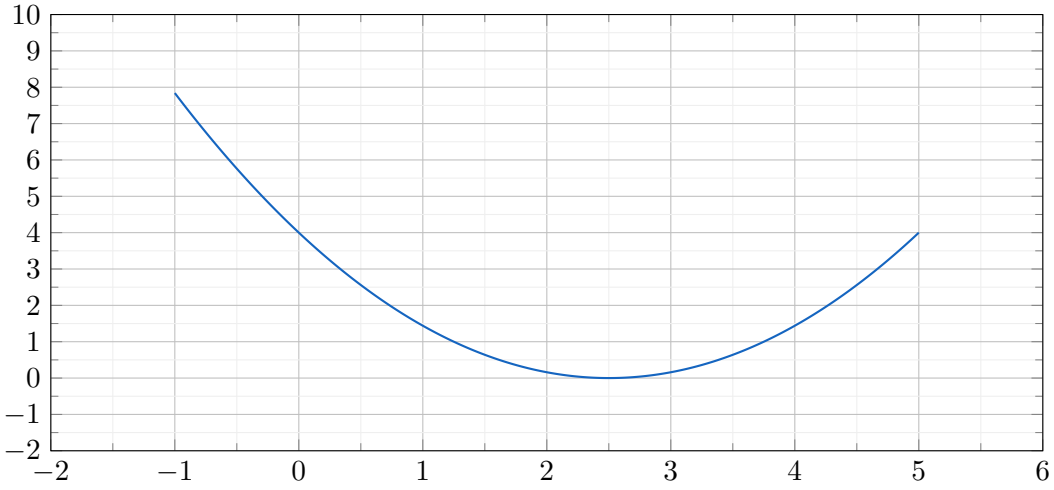
Existence de la solution Avant de montrer cette proposition, *i.e.* l'expression de l'estimateur, nous allons expliquer pourquoi ce problème a une unique solution en se focalisant sur le modèle linéaire simple.

on a

$$\begin{aligned}
\min_{\boldsymbol{\beta}\in\mathbb{R}^2} \|Y - \hat{Y}\|_2^2 &= \min_{\boldsymbol{\beta}\in\mathbb{R}^2} \|Y - (X\boldsymbol{\beta})\|_2^2, \\
&\quad \downarrow \text{ on se rappelle que } \hat{Y} = \beta X \\
&= \min_{\boldsymbol{\beta}\in\mathbb{R}^2} \|Y - (\beta_0 + \boldsymbol{\beta}_1 \mathbf{x})\|_2^2, \\
&\quad \downarrow \text{ on se rappelle que pour tout vecteur } \mathbf{x},\ \langle \mathbf{x}, \mathbf{x}\rangle = \mathbf{x}^T\mathbf{x} = \langle \mathbf{x} \mid \mathbf{x}\rangle = \|\mathbf{x}\|_2^2 \\
&= \min_{\boldsymbol{\beta}\in\mathbb{R}^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.
\end{aligned}$$

On voit bien que cette fonction est convexe. En effet, on se rappelle qu'une fonction $x \mapsto (a-x)^2$ est une fonction *quadratique* donc convexe, et la somme de fonctions convexes reste convexe.

On peut représenter cela graphiquement (dans le cas unidimensionnel) par le graphe ci-dessous.



Solution La fonction étant convexe, ce que l'on peut vérifier aussi calculant la matrice hessienne, les minima de cette fonction sont données en cherchant l'endroit où la dérivée de la fonction $\boldsymbol{\beta} \mapsto \|Y - X\boldsymbol{\beta}\|_2^2$ s'annule. On va donc chercher les valeurs de β telles que

$$\frac{\partial}{\partial \boldsymbol{\beta}} \|Y - X\boldsymbol{\beta}\|_2^2 = 0 \iff -2X^T(Y - X\boldsymbol{\beta}) = 0 \tag{1}$$

En dérivant à nouveau fonction, on trouve

$$\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \|Y - X\boldsymbol{\beta}\|_2^2 = 2X^T X \succ 0,$$

i.e. la matrice hessienne est définie positive, ce qui est le cas ici car il s'agit de la matrice de variance-covariance des données, cette convexité nous permettra de dire que la vecteur β vérifiant l'équation (1) est bien solution de notre problème de minimisation. Or

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} \|Y - X\boldsymbol{\beta}\|_2^2 = 0 &\iff -2X^T(Y - X\boldsymbol{\beta}) = 0, \\
&\downarrow \text{ on peut diviser par } -2 \\
&\iff X^T(Y - X\boldsymbol{\beta}) = 0, \\
&\iff X^TY - X^TX\boldsymbol{\beta} = 0, \\
&\iff X^TX\boldsymbol{\beta} = X^TY, \\
&\downarrow \text{ à condition que la matrice } X^TX \text{ soit inversible} \\
&\iff \boldsymbol{\beta} = (X^TX)^{-1}X^TY.
\end{aligned}$$

Solution du problème linéaire simple On va maintenant chercher à retrouver les solutions du modèle linéaire gaussien simple, *i.e.* les expression de β_0 et β_1 trouvées lors de la précédente séance.

A partir de maintenant, on va introduire les notations suivantes

$$\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i,\ s_u^2 = \frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})^2,\ s_{u,v}^2 = \frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})(v_i - \bar{v}) = \bar{xy} - \bar{x}\bar{y}$$

Commençons par calculer les produits matriciels X^TY et X^TX

$$X^TY = \begin{pmatrix} n\bar{y} \\ n\bar{xy} \end{pmatrix} \quad \text{et} \quad X^TX = \begin{pmatrix} n & n\bar{x} \\ n\bar{x} & n\bar{x^2} \end{pmatrix}$$

Nous devons ensuite calculer l'inverse de la matrice X^TX, on se rappelle que cet inverse, pour une matrice carrée d'ordre 2, est donné par

$$(X^TX)^{-1} = \frac{1}{n^2(\bar{x^2} - \bar{x}^2)}\begin{pmatrix} n\bar{x^2} & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} = \frac{1}{ns_x^2}\begin{pmatrix} \bar{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}.$$

Il ne reste plus qu'à déterminer le vecteur β en utilisant sa définition

$$\beta = (X^TX)^{-1}X^TY = \frac{1}{s_x^2}\begin{pmatrix} \bar{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}\begin{pmatrix} \bar{y} \\ \bar{xy} \end{pmatrix} = \frac{1}{s_x^2}\begin{pmatrix} \bar{y}\bar{x^2} - \bar{x}\bar{xy} \\ \bar{xy} - \bar{x}\bar{y} \end{pmatrix}.$$

On trouve donc

$$\beta_1 = \frac{\bar{xy} - \bar{x}\bar{y}}{s_x^2} = \frac{s_{x,y}^2}{s_x^2}.$$

et

$$\beta_0 = \frac{\bar{y}\bar{x^2} - \bar{x}^2\bar{y} + \bar{y}\bar{x}^2 - \bar{x}\bar{xy}}{s_x^2} = \bar{y} - \frac{s_{x,y}^2}{s_x^2}\bar{x}.$$

On peut alors reprendre un exercice précédent et vérifier que l'on obtient bien les mêmes estimations avec le résultat de la Proposition 0.1.

On considère le jeu de données ci-dessous et on affiche les estimations du coefficient directeur et de l'ordonnée à l'origine de la droite.

```r
# Un vecteur y qui représente la taille de l'individu
y = c(96,104.8,110.3,115.3,121.9,127.4,130.8,136,139.7,144.5)
# Un vecteur x qui représentera l'âge de l'individu
x = c(3:12)
# Mise sous format data.frame
data <- data.frame(age = x, taille = y)
# Modèle
mymodel <- lm(y~x,data)
mymodel$coefficients

## (Intercept)           x
##       83.52        5.22
```

On fait maintenant de même mais avec les notations matricielles. On commence par définir nos objets X et Y qui serviront à faire l'estimation

```r
# Création des objets X et Y
Y = matrix(y, ncol = 1, nrow = length(y))
X = cbind(rep(1,length(x)),x)

# Estimation des paramètres
beta_hat = solve(t(X)%*%X)%*%t(X)%*%Y
beta_hat

##    [,1]
##   83.52
## x  5.22
```

Autour de l'estimation par maximum de vraisemblance et BIC

Nous proposons d'étudier une autre façon d'estimer les paramètres du modèles. A savoir, les coefficients du paramètre $\boldsymbol{\beta}$ et la variance des erreurs σ^2.

On rappelle que le modèle étudié est de la forme :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon = X\boldsymbol{\beta} + \varepsilon$$

où $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ est une variable normale avec une moyenne nulle et une variance de σ^2.

Pour notre estimation, nous allons procéder par maximum de vraisemblance. Considérons un échantillon *i.i.d.* $S = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$ de n exemples. On rappelle (c'est une des hypothèses du modèle linéaire gaussien) que les données y_i sont distribuées selon une loi gaussienne, *i.e.* $y_i \sim \mathcal{N}(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2)$. De plus, la vraisemblance d'un échantillon *i.i.d* est défini comme le produit des valeurs de la densité en chaque valeurs de l'échantillon et la densité d'une loi normale de moyenne μ et de variance σ^2 est donnée par

$$f(t, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)}.$$

Ainsi la vraisemblance L de notre échantillon S est donnée par

$$\begin{aligned}
L(S, \boldsymbol{\beta}, \sigma) &= \prod_{i=1}^n f(S, \boldsymbol{\beta}, \sigma), \\
&\quad \downarrow \text{ densité de la loi gaussienne} \\
&= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right)}, \\
&\quad \downarrow \text{ propriété de l'exponentielle} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\left(-\sum_{i=1}^n \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right)}, \\
&\quad \downarrow \text{ or } \sum_{i=1}^n \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2} = \frac{\|Y - X\boldsymbol{\beta}\|_2^2}{2\sigma^2} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\left(-\frac{\|Y - X\boldsymbol{\beta}\|_2^2}{2\sigma^2}\right)},
\end{aligned}$$

où les notations X et Y sont celles employées dans la partie précédente.

On se rappelle que les meilleurs paramètres sont ceux qui permettent de maximiser la vraisemblance de nos données. Mais cette expression est bien trop complexe à manipuler. Donc au lieu de maximiser la vraisemblance L on va chercher à maximiser la log-vraisemblance ℓ, définie par $L = \ln(L)$ soit

$$\ell(S, \boldsymbol{\beta}, \sigma) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{\|Y - X\boldsymbol{\beta}\|_2^2}{2\sigma^2}.$$

Les valeurs de σ^2 et $\boldsymbol{\beta}$ qui maximisent la vraisemblance sont données en résolvant le système défini par

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}}(S, \sigma^2, \boldsymbol{\beta}) = 0 \quad \text{et} \quad \frac{\partial \ell}{\partial \sigma^2}(S, \sigma^2, \boldsymbol{\beta}) = 0.$$

Concentrons nous sur la première équation, ce qui nous donne

$$\begin{aligned}
&\frac{\partial \ell}{\partial \boldsymbol{\beta}}(S, \sigma^2, \boldsymbol{\beta}) = 0, \\
&\quad \downarrow \text{ on dérive notre norme} \\
\iff &-\frac{X^T(Y - X\boldsymbol{\beta})}{2\sigma^2} = 0, \\
\iff &-X^TY - X^TX\boldsymbol{\beta} = 0, \\
&\quad \downarrow \text{ on isole le vecteur } \boldsymbol{\beta} \\
\iff &\boldsymbol{\beta} = \left(X^TX\right)^{-1} X^TY.
\end{aligned}$$

Faisons de même avec la deuxième équation en utilisant l'estimateur $\hat{\boldsymbol{\beta}}$ précédemment obtenu et en notant que

$$\|Y - X\hat{\boldsymbol{\beta}}\|_2^2 = \|Y - \hat{Y}\|_2^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = RSS.$$

où RSS signifie *Residual Sum of Squares*. On a ainsi,

$$\begin{aligned}
&\frac{\partial \ell}{\partial \sigma^2}(S, \sigma^2, \boldsymbol{\beta}) = 0, \\
&\quad \downarrow \text{ on en utilisant les notations précédentes} \\
\iff &-\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{RSS}{(\sigma^2)^2} = 0, \\
&\quad \downarrow \text{ en multipliant par } 2\sigma^2 \\
\iff &-n\sigma^2 + RSS = 0, \\
&\quad \downarrow \text{ on isole } \sigma^2 \\
\iff &\sigma^2 = \frac{RSS}{n}.
\end{aligned}$$

Ainsi les estimateurs obtenus par maximum de vraisemblance sont donnés par

$$\hat{\boldsymbol{\beta}} = \left(X^TX\right)^{-1} X^TY \quad \text{et} \quad \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}) = \frac{RSS}{n}.$$

Remarque L'estimateur de σ^2 ainsi obtenu est biaisé ! Il faudrait corriger cet estimateur pour le rendre non biaisé.

Ainsi, en ces points là, notre log-vraisemblance ℓ est maximale, et son maximum $m\ell$ est donné par, en notant $\hat{Y}$ le vecteur des prédictions,

$$\begin{aligned} m\ell(S, \hat{\boldsymbol{\beta}}, \hat{\sigma^2}) = & -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma^2}) - \frac{\|Y-\hat{Y}\|_2^2}{2\hat{\sigma^2}}, \\ & \downarrow \text{ en utilisant les estimateurs} \\ = & -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\frac{1}{n}\|Y-\hat{Y}\|_2^2\right) - \frac{\|Y-\hat{Y}\|_2^2}{\frac{2}{n}\|Y-\hat{Y}\|_2^2}, \\ & \downarrow \text{ après simplification} \\ = & -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\frac{RSS}{n}\right) - \frac{n}{2}, \end{aligned}$$

car $\|Y-\hat{Y}\|_2^2 = \sum_{i=1}^n (y_i - \hat{y}_i) = RSS$.

Cette quantité joue un rôle important dans la définition du BIC qui permet de comparer deux modèles en statistiques et donc de faire de la sélection de modèles.
Ce critère est défini par

$$BIC = -2m\ell + (p+1)\ln(n),$$

où p représente le nombre de paramètres à estimer au cours de la régression. Le "+1" fait référence à l'estimation de la variance σ^2 des erreurs. En utilisant l'expression du maximum de vraisemblance, on obtient

$$BIC = n(\ln(2\pi)+1) + n\ln\left(\frac{RSS}{n}\right) + (p+1)\ln(n).$$