

Modèles Linéaires

Correction Séance 4 Licence 3 MIASHS (2022-2023)

Guillaume Metzler, Francesco Amato
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr ; francesco.amato@univ-lyon2.fr

Résumé

Cette deuxième se concentre sur la comparaison de modèles et des intervalles de confiance sur le modèle et les prédictions :

- Rappels sur le modèle linéaire gaussien,
- Etude des propriétés des estimateurs de $\hat{\beta}$ obtenus par MCO,
- Applications sur un jeu de données à travers les exercices 4 et 5 du TD 2.

Dans cette séance, nous allons traiter les exercices 4 et 5 de la fiche de TD 2.

L'objectif est de se familiariser avec les propriétés de l'estimateur pour le modèle linéaire multiple et de regarder ces propriétés, ainsi que quelques tests statistiques sur un exemple.

1 Propriétés de l'estimateur

Notre modèle linéaire multiple se présente sous la forme

$$Y_i = \beta_0 + X\beta + \varepsilon_i,$$

$$\text{où } \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}.$$

Ce que l'on peut encore écrire, en définissant une matrice \tilde{X} comme étant une colonne de 1 ajoutée à la matrice X , mais aussi un vecteur $\tilde{\beta}$ comme étant le vecteur β auquel on ajoute la valeur β_0 , comme :

$$Y_i = \tilde{X}\tilde{\beta} + \varepsilon_i.$$

Dans la suite, on renomme la matrice \tilde{X} par X et le vecteur $\tilde{\beta}$ par β pour simplifier les notations.

La première question nous suggère que l'on peut réécrire le modèle sous la forme suivante :

$$Y_i = \beta_*^T X_{i,*} + \varepsilon_i,$$

où

$$\beta_* = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \quad \text{et} \quad X_{i,*} = \begin{pmatrix} 1 \\ x_{i,1} \\ \vdots \\ x_{i,p} \end{pmatrix}.$$

Enfin, on rappelle le résultat suivant présenté lors de la deuxième séance de TD.

Proposition 1.1: Solution du modèle linéaire (multiple)

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

$$Y = \beta X + \varepsilon.$$

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

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

est donnée par

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

On va maintenant étudier les propriétés de cette estimateur, *i.e.* nous allons regarder l'espérance et la variance de cet estimateur $\hat{\beta}$.

Espérance de l'estimateur : $\mathbb{E}[\hat{\beta}]$ On repart de la définition de l'estimateur dans la Proposition 1.1 et on garde l'esprit que X est **déterministe** (ce n'est pas une variable aléatoire), la seule partie aléatoire dans l'expression de $\hat{\beta}$ vient de la variable aléatoire Y_i via la variable aléatoire ε_i .

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

Variance de l'estimateur $Var[\hat{\beta}]$ On garde à l'esprit ce que nous avons utilisés précédemment, à savoir que seule ε_i , *i.e.* les erreurs sont aléatoires ainsi que les propriétés de la variance.

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

Retour au modèle linéaire simple On peut repartir de cette dernière expression et retrouver la distribution des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ que nous avons travaillé lors du premier TD, *i.e.*

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}\right) \quad \text{et} \quad \hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

En ce qui concerne l'espérance de ces deux estimateurs, c'est une conséquence directe du fait que $\mathbb{E}[\hat{\beta}] = \beta$. Il faut donc maintenant se concentrer sur la variance des estimateurs. Regardons alors concrètement la matrice de variance-covariance dans le cadre du modèle linéaire simple. Nous avons :

$$\text{Var}[\hat{\beta}] = \sigma^2 (X^T X)^{-1} = \frac{\sigma^2}{n^2(\bar{x}^2 - \bar{x}^2)} \begin{pmatrix} n\bar{x}^2 & -n\bar{x} \\ -n\bar{x} & n \end{pmatrix} = \frac{\sigma^2}{n(\bar{x}^2 - \bar{x}^2)} \begin{pmatrix} \bar{x}^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}.$$

Réécrivons ces éléments là de façon complète, on utilisera le fait que $\bar{x}^2 - \bar{x}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

$$\text{Var}[\hat{\beta}] = \frac{\sigma^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{pmatrix}.$$

Si on se concentre maintenant sur les éléments diagonaux de cette matrice, on trouve respectivement, en développant le produit

$$\text{Var}[\hat{\beta}_0] = \frac{\sigma^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sum_{i=1}^n x_i^2 \quad \text{et} \quad \text{Var}[\hat{\beta}_1] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Généralisation On pourra montrer, de façon générale, pour le modèle linéaire multiple, que les coefficients de régression ont la distribution suivante

$$\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2 h_{j+1,j+1}), \quad (1)$$

où H désigne la matrice $(X^T X)^{-1}$ et $h_{j+1,j+1}$ désigne le $j+1$ -ème élément sur la diagonale de cette matrice.

Attention : il y a un décalage entre l'indice du coefficient et l'indice dans la matrice H . Cela vient du fait que nous initions notre modèle de régression à β_0 et non à β_1 .

2 Un exemple d'application

On considère maintenant le jeu de données présentés dans l'Exercice 5 du TD 2.

Commençons par reprendre notre jeu de données

```
# Variable réponse
y = c(125,158,207,182,196,175,145,144,160,175,
      151,161,200,173,175,162,155,230,162,153)
x0 = rep(1,length(y))
x1 = c(13,39,52,29,50,64,11,22,30,51,27,41,51,37,23,43,38,62,28,30)
x2 = c(18,18,50,43,37,19,27,23,18,11,15,22,52,36,48,15,19,56,30,25)
x3 = c(25,59,62,50,65,79,17,31,34,58,29,53,75,44,27,65,62,75,36,41)
x4 = c(11,30,53,29,56,49,14,17,22,40,31,39,36,27,20,36,37,50,20,33)
```

On rappelle que notre modèle de régression se présente sous la forme

$$Y = X\beta + \varepsilon,$$

où

- Y est le vecteur des variables réponses
- X est notre matrice de design dont les colonnes sont formées des vecteurs $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ et \mathbf{x}_4 .
- $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ est le vecteur des paramètres du modèle de régression
- ε est le vecteur des erreurs du modèle (bruit gaussien)

On rappelle également que le vecteur $\hat{\beta}$ solution des paramètres du modèle est défini par


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

On peut alors déterminer sa valeur directement à partir des données introduites précédemment.

```
# Matrice de design X
X = cbind(x0,x1,x2,x3,x4)

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

```
##          [,1]
## x0 102.7439113
## x1   1.2540154
## x2   1.0642913
## x3  -0.3713815
## x4   0.2338960
```

On pourra ensuite comparer cela aux résultats de la fonction `lm` de  et vérifier que ces derniers coïncident.

```
# Résolution en utilisant R
data = data.frame(y, x1, x2, x3, x4)
mymodel = lm(y~., data)
mymodel$coefficients

## (Intercept)          x1          x2          x3          x4
## 102.7439113   1.2540154   1.0642913  -0.3713815   0.2338960
```

On se propose maintenant d'estimer la variance du vecteur des paramètres $\hat{\beta}$. Pour cela, on rappelle, d'après ce qui a été vu en Section 1, que

$$\text{Var}[\hat{\beta}] = \mathbb{E}(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T = \sigma^2(X^T X)^{-1}.$$

Or σ^2 est inconnu mais nous pouvons l'estimer à partir des résidus ε_i de notre modèle

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^n \varepsilon_i^2.$$

```
n = nrow(X)
p = ncol(X)

# Estimation de sigma2
sigma2 = (1/(n-p))*sum((y - X%%beta_hat)^2)
```

On en déduit alors la variance de l'estimateur $\hat{\beta}$

```

# Variance estimateur beta_hat
var_beta_hat = sigma2*solve(t(X)%*%X)
var_beta_hat

##           x0           x1           x2           x3           x4
## x0 21.4102077  0.1956510354 -0.2178526911 -0.2954384348 -0.177860934
## x1  0.1956510  0.1240232098 -0.0005394934 -0.0612044395 -0.053986850
## x2 -0.2178527 -0.0005394934  0.0112474326 -0.0002625021 -0.002353997
## x3 -0.2954384 -0.0612044395 -0.0002625021  0.0513147016  0.001179114
## x4 -0.1778609 -0.0539868495 -0.0023539966  0.0011791140  0.067334946

# On peut extraire les éléments diagonaux de la matrice
diag_beta_hat = diag(var_beta_hat)
std_beta_hat = sqrt(diag_beta_hat)
std_beta_hat

##           x0           x1           x2           x3           x4
## 4.6271166 0.3521693 0.1060539 0.2265275 0.2594898

```

On va finir en testant la nullité des paramètres β_j pour les valeurs de $j = 0, 1$ et 2 . Pour cela, nous utiliserons les propriétés des estimateurs présentés dans l'Equation (1)

On en déduit que les statistiques de tests à employer sont définies par

$$t_j = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{h_{j+1,j+1}}} \sim T_{n-p}.$$

```

# Pour j = 0
t0 = beta_hat[1]/sqrt(var_beta_hat[1,1])
t0

## [1] 22.20474

# Calcul de la p-value
p0 = 2*(1-pt(abs(t0),n-p))
p0

## [1] 6.898926e-13

```

```


# Pour j = 1
t1 = beta_hat[2]/sqrt(var_beta_hat[2,2])
# Calcul de la p-value
p1 = 2*(1-pt(abs(t1),n-p))
p1

## [1] 0.002845303

# Pour j = 2
t2 = beta_hat[3]/sqrt(var_beta_hat[3,3])
# Calcul de la p-value
p2 = 2*(1-pt(abs(t2),n-p))
p2

## [1] 4.769925e-08

```

Remarque Pour vérifier que l'on ne s'est pas trompé, on pourra toujours comparer nos résultats avec les sorties de la fonction *lm* de .

```

summary(mymodel)

##
## Call:
## lm(formula = y ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3442  -3.7614  -0.1699   3.3459   8.9112
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  102.7439     4.6271   22.205 6.90e-13 ***
## x1           1.2540     0.3522    3.561 0.00285 **
## x2           1.0643     0.1061   10.035 4.77e-08 ***
## x3          -0.3714     0.2265   -1.639 0.12192
## x4           0.2339     0.2595    0.901 0.38164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.26 on 15 degrees of freedom

```



```
## Multiple R-squared:  0.9485, Adjusted R-squared:  0.9348  
## F-statistic: 69.12 on 4 and 15 DF,  p-value: 1.76e-09
```