Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

OLS #6 (le modèle linéaire multiple - 1)

Régression Linéaire

Nous supposons que les données collectées suivent le modèle suivant

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i,$$

où

- x_{ij} sont des nombres déterministes (connus). β_0 représente la constante (`intercept` dans les logiciels). On notera souvent $x_{i0} = 1$.
- β_j, $j = 0, 1, \ldots, k$ paramètres réels à estimer. On pose $p = k + 1$
- les variables ε_i sont des fluctuations aléatoires (erreur de mesures, mauvaise spécification du modèle,...).

On peut reformuler le modèle en:

$$\underbrace{\mathbf{Y}}_{(n\times 1)} = \underbrace{\mathbf{X}}_{(n\times p)} \underbrace{\boldsymbol{\beta}}_{(p\times 1)} + \underbrace{\boldsymbol{\varepsilon}}_{(n\times 1)}$$

Régression Linéaire

On peut reformuler le modèle en: $\underbrace{\mathbf{Y}}_{(n\times 1)} = \underbrace{\mathbf{X}}_{(n\times p)} \underbrace{\boldsymbol{\beta}}_{(p\times 1)} + \underbrace{\boldsymbol{\varepsilon}}_{(n\times 1)}$ où

$\mathbf{1 \leq p = 1 + k \leq n}$ et

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{nk} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$\boldsymbol{\beta} \in \mathbb{R}^p$ et $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^\top$. Lorsque $p = 2$, ce modèle s'appelle un modèle de régression linéaire **simple**. On notera

- $\mathbf{x}_j$ le vecteur de taille $(n \times 1)$ des n observations de la $j$ème covariable.
- $\mathbf{x}'_i$ le vecteur de taille $(1 \times p)$ des valeurs des p covariables pour l'individu i.
- $\mathbf{Y}$ vecteur réponse; $\boldsymbol{\varepsilon}$ vecteur aléatoire (centré sans perte de généralité).

Ainsi, la matrice $\mathbf{X}$ appelée matrice de design (ou de plan d'expérience) s'écrit également: $\mathbf{X} = (\mathbf{x}_0, \ldots, \mathbf{x}_k) = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix}$.

$\mathcal{H}_1$: La matrice de design $\mathbf{X}$ est de plein rang.
Puisque $p \leq n$, $\mathcal{H}_1 \Rightarrow \mathrm{rg}(\mathbf{X}) = p$, $\mathbf{X}^\top\mathbf{X}$ de taille (p, p) est symétrique, définie positive et donc **inversible**.

Comment estimer $\boldsymbol{\beta}$?

L'idée consiste à minimiser une certaine norme de $\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$.

Plusieurs méthodes pourraient être envisagées: tenter de minimiser

- $\displaystyle\sum_{i=1}^{n}(Y_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$.
  (méthode des moindres carrés)
- $\displaystyle\sum_{i=1}^{n}\rho(Y_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})$, $\rho(\cdot) \geq 0$, $\rho(0) = 0$, e.g. $\rho(t) = |t|, |t|\mathbf{1}(|t| \leq M), \ldots$.
  (M-estimation)
- $\displaystyle\sum_{i=1}^{n} d_i^2$, où d_i est la distance la plus proche du point $(x_{i1}, \ldots, x_{ip}, y_i)$ sur le plan d'équation $\beta_1 x_1 + \cdots + \beta_p x_p$
  (régression orthogonale)

On appelle estimateur des moindres carrés ordinaires (noté MCO ou OLS) $\widehat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ la valeur suivante:

$$\widehat{\boldsymbol{\beta}} = \underset{\beta_0,\beta_1,\ldots,\beta_k}{\text{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ik})^2 = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{argmin}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

Sous l'hypothèse $\mathcal{H}_1$, l'estimateur des MCO existe et vaut

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}.$$

Par calcul matriciel, en calculant le gradient et la matrice hessienne de $S(\boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2$:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\top\mathbf{Y} + 2\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \quad \text{et} \quad \frac{\partial^2 S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^\top} = 2\mathbf{X}^\top\mathbf{X}.$$

Régression Linéaire Simple

L'estimateur des MCO du modèle de régression linéaire simple $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (sous l'hypothèse que $\widehat{\text{Var}}(\mathbf{x}) \neq 0$) est donné par

$$\hat{\beta}_1 = \frac{\widehat{\text{Cov}}(\mathbf{x}, \mathbf{Y})}{\widehat{\text{Var}}(\mathbf{x})} \quad \text{et} \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{x}$$

où $\widehat{\text{Cov}}(\mathbf{a}, \mathbf{b}) = (n-1)^{-1} \sum (a_i - \overline{a})(b_i - \overline{b})$

Régression Linéaire

On suppose que l'on a un modèle homoscédastique

$\mathcal{H}_2$: Les erreurs sont centrées, de même variance et non corrélées
$\Leftrightarrow \mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ et $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbb{I}_n$.

Sous les hypothèses $\mathcal{H}_1 - \mathcal{H}_2$, l'estimateur $\hat{\boldsymbol{\beta}}$ des MCO est un estimateur sans biais de $\boldsymbol{\beta}$ et $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\Big((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}\Big) = \mathbb{E}\Big((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}]\Big)$$

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \underbrace{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}}_{=\mathbb{I}}\boldsymbol{\beta} + \mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\underbrace{\mathbb{E}(\boldsymbol{\varepsilon})}_{=0} = \boldsymbol{\beta}.$$

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\Big((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top[\boldsymbol{\varepsilon}][\boldsymbol{\varepsilon}]^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\Big)$$

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\underbrace{\mathbb{E}\Big(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\Big)}_{=\sigma^2\mathbb{I}_n}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

Régression Linéaire, Gauss-Markov

Si $p = 2$, les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont sans biais et

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \overline{x})^2},\ \text{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{n\sum(x_i - \overline{x})^2},$$

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \overline{x}}{\sum(x_i - \overline{x})^2}$$

Sous les hypothèses $\mathcal{H}_1 - \mathcal{H}_2$, l'estimateur des MCO est optimal (sans biais et de variance minimale) dans la classe des estimateurs linéaires sans biais de $\boldsymbol{\beta}$.

Si n augmente, $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ dépend de $\mathbf{X}^\top\mathbf{X}$, donc sans hypothèse supplémentaire, il est impossible d'établir une convergence pour $\hat{\boldsymbol{\beta}}$.
$\mathcal{H}_3$: La matrice de design $\mathbf{X}$ est telle que lorsque $n \to \infty$

$$\frac{1}{n}(\mathbf{X}^\top\mathbf{X}) \to \mathbf{Q}$$

où $\mathbf{Q}$ est une matrice définie positive.
Sous les hypothèses, $\mathcal{H}_1 - \mathcal{H}_3$ l'estimateur des MCO $\hat{\boldsymbol{\beta}}$ est convergent en moyenne quadratique

Aspects Computationnels

```
> data(apartments, package = "DALEX")
> n = nrow(apartments)
> X = matrix(c(rep(1,n), apartments$construction.year,
    apartments$surface,
apartments$no.rooms),n,4)
> Y = matrix(apartments$m2.price,n,1)
> head(cbind(Y,X))
     [,1] [,2] [,3] [,4] [,5]
[1,] 5897    1 1953   25    1
[2,] 1818    1 1992  143    5
[3,] 3643    1 1937   56    2
[4,] 3517    1 1995   93    3
[5,] 3013    1 1992  144    5
[6,] 5795    1 1926   61    2
```

Ici $n = 1,000$, $\mathbf{Y} \in \mathbb{R}^n$ et $\mathbf{X}$ est une matrice $(n \times (1+3))$.

Aspects Computationnels

$\widehat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ est l'estimateur par moindres carrés

```
> solve(t(X)%*%X)%*%(t(X)%*%Y)
             [,1]
[1,] 6295.7094722
[2,]   -0.8829108
[3,]   -9.3826890
[4,]  -80.6138913
```

$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}$ et $\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$. On note $\hat{\sigma}^2 = \dfrac{1}{n-p}\sum \hat{\varepsilon}_i^2$

```
> ypred = X %*% solve(t(X)%*%X)%*%(t(X)%*%Y)
> residus = Y-ypred
> (sigma2 = sum(residus^2)/(n-4))
[1] 611175.7
> sqrt(sum(residus^2)/(n-4))
[1] 781.7772
```

Aspects Computationnels

$\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$ peut être estimée par $\hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}$

```
> (sum(residus^2)/(n-4))*solve(t(X)%*%X)
        [,1]          [,2]         [,3]         [,4]
[1,] 3550207 -1.807629e+03 152.91300074 -3277.950843
[2,]   -1807  9.214743e-01  -0.07991224     1.170977
[3,]     152 -7.991224e-02   2.56217569   -64.046489
[4,]   -3277  1.170977e+00 -64.04648859  1922.299635
> sqrt(diag((sum(residus^2)/(n-4))*solve(t(X)%*%X)))
[1] 1884.19946    0.95993    1.60067   43.84403
```

Comme auparavant, on obtient directement ces informations via

```
> reg = lm(m2.price~construction.year+surface+no.rooms
    , data=apartments)
> summary(reg)
```

Aspects Computationnels

On retrouve les $\hat{\boldsymbol{\beta}}_j$, et un estimateur des $\text{Var}(\hat{\boldsymbol{\beta}}_j)$

```
> summary(reg)

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) 6295.7095  1884.1995   3.341 0.000865 ***
const.year    -0.8829     0.9599  -0.920 0.357920
surface       -9.3827     1.6007  -5.862 6.22e-09 ***
no.rooms     -80.6139    43.8440  -1.839 0.066264 .

Residual standard error: 781.8, 996 degrees of freedom
Multiple R-squared: 0.2588, Adjusted R-squared: 0.2566
F-statistic: 115.9 on 3 and 996 DF, p-value: < 2.2e-16
```

et un estimateur de σ, l'écart-type des résidus ε_i.

Note On a un estimateur de $\text{Var}(\hat{\boldsymbol{\beta}})$ avec

```
> vcov(reg)
             (Intercept) const.year   surface   no.rooms
(Intercept)    3550207  -1.80e+03 152.91300 -3277.9508
construc.year    -1807   9.21e-01  -0.07991     1.1709
surface            152  -7.99e-02   2.56217   -64.0464
no.rooms         -3277   1.17e+00 -64.04648  1922.2996
```