Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

OLS #8 (cas Gaussien & tests)

Rappelons que, le paramètre $\boldsymbol{\beta}$ sous un certain nombre d'hypothèses (sans faire d'hypothèses sur la loi de $\boldsymbol{\varepsilon}$) est estimé par

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

et la variance résiduelle est estimée par

$$\hat{\sigma}^2 = \frac{1}{n-p} \|\hat{\boldsymbol{\varepsilon}}\|^2 = \frac{SCR}{n-p}$$

Régression Linéaire

$\mathcal{H}_1$: La matrice de design $\mathbf{X}$ est de plein rang.
$\mathcal{H}_2$: $\Leftrightarrow \mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ et $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbb{I}_n$.
$\mathcal{H}_{2'}$: $\Leftrightarrow \forall n$, $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ et $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbb{I}_n$, $\mu_4 = \mathbb{E}(\varepsilon_i^4) < \infty$.
$\mathcal{H}_3$: 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 Q est une matrice définie positive.

- $\mathcal{H}_1$: permet de démontrer l'existence de $\hat{\boldsymbol{\beta}}^{\text{MCO}}$.
- $\mathcal{H}_2$: permet de démontrer des propriétés pour $\hat{\boldsymbol{\beta}}^{\text{MCO}}$ (sans biais, calcul de variance).
- $\mathcal{H}_3$: permet de démontrer la convergence en moyenne quadratique de $\hat{\boldsymbol{\beta}}^{\text{MCO}}$.
- $\mathcal{H}_{2'}$: permet de démontrer la convergence en moyenne quadratique de $\hat{\sigma}^2$.

$\mathcal{H}_2^{\text{Gauss}}$: $\boldsymbol{\varepsilon}$ est un vecteur gaussien centré de matrice de covariance $\sigma^2\mathbb{I}_n$, i.e. $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbb{I}_n)$

Sans hypothèses gaussienne: dans ce cas, deux cas de figure seront envisagés,

- soit on dispose de suffisamment d'observations: on utilisera alors des résultats asymptotiques (hypothèse $\mathcal{H}_3$ sera enrichie).
- soit la taille d'échantillon n'est pas assez grande: on utilisera alors des techniques de rééchantillonnage pour estimer la distribution des $\hat{\beta}_j$ notamment.

Régression Linéaire

Soit un modèle linéaire homoscédastique vérifiant les hypothèses $\mathcal{H}_1 - \mathcal{H}_2$ [Gauss], alors

- l'estimateur du MV de $\boldsymbol{\beta}$ vaut $\hat{\boldsymbol{\beta}}^{\text{MV}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$;
- l'estimateur du MV de σ^2 vaut $\hat{\sigma}^2_{\text{MV}} = \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{n}$.
- $\hat{\boldsymbol{\beta}}^{\text{MV}} = \hat{\boldsymbol{\beta}}^{\text{MCO}}$ (la dépendance $_{\text{MV}, \text{MCO}}$ sera parfois omise).
- $\hat{\sigma}^2_{\text{MV}}$ est un estimateur biaisé de σ^2; sa version sans biais sera notée $\hat{\sigma}^2 = \|\hat{\boldsymbol{\varepsilon}}\|^2/(n-p)$.

Régression Linéaire

Soit un modèle linéaire homoscédastique vérifiant les hypothèses $\mathcal{H}_1 - \mathcal{H}_2$ [Gauss], alors $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ est une statistique complète et $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ est de variance minimimum dans la classe des estimateurs sans biais.

Sous les hypothèses $\mathcal{H}_1 - \mathcal{H}_2$ [Gauss], nous pouvons établir que

- $\hat{\boldsymbol{\beta}}$ est un vecteur gaussien centré en $\boldsymbol{\beta}$ et de variance $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.
- $(n-p)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-p}$.
- $\hat{\boldsymbol{\beta}}$ et $\hat{\sigma}^2$ sont indépendants.

Sous les hypothèses $\mathcal{H}_1$ et $\mathcal{H}_2$ [Gauss], on a

- Pour $j = 1, \ldots, p$

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathcal{S}td_{n-p} \quad \text{où } \hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}.$$

- Soit $\mathbf{R}$ une matrice de taille (q, p) (avec $q \leq p$) alors

$$\frac{1}{q\hat{\sigma}^2}(\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^\top(\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top)^{-1}\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \mathcal{F}_{q,n-p}.$$

Régression Linéaire

Sous les hypothèses $\mathcal{H}_1$ et $\mathcal{H}_2$ [Gauss], on a, pour $j = 1, \ldots, p$

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathcal{S}td_{n-p} \quad \text{où } \hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{(\mathbf{X}^\top\mathbf{X})^{-1}_{jj}}.$$

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

Au niveau de confiance $1-\alpha$:

- Un intervalle de confiance bilatéral du paramètre β_j, $j = 1, \ldots, p$ est donné par

$$\mathrm{IC}_{1-\alpha}(\beta_j) = [\hat{\beta}_j + t_1 \hat{\sigma}_{\hat{\beta}_j}; \hat{\beta}_j + t_2 \hat{\sigma}_{\hat{\beta}_j}],$$

où $\hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{((\mathbf{X}^\top \mathbf{X})^{-1})_{jj}}$ et où $\mathrm{P}(t_1 \leq T_{n-p} \leq t_2) = 1-\alpha$.

- Un intervalle de confiance bilatéral du paramètre σ^2 est donné par

$$\mathrm{IC}_{1-\alpha}(\sigma^2) = [(n-p)\hat{\sigma}^2/q_2; (n-p)\hat{\sigma}^2/q_1] \text{ où } \mathrm{P}(q_1 \leq \chi^2_{n-p} \leq q_2).$$

Régression Linéaire

$$\mathrm{IC}_{1-\alpha}(\beta_j) = [\hat{\beta}_j + t_1 \hat{\sigma}_{\hat{\beta}_j}; \hat{\beta}_j + t_2 \hat{\sigma}_{\hat{\beta}_j}],$$

où $\hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{((\mathbf{X}^\top \mathbf{X})^{-1})_{jj}}$ et où $\mathrm{P}(t_1 \leq T_{n-p} \leq t_2) = 1 - \alpha$.

```
> confint(reg)
                          2.5 %       97.5 %
(Intercept)        2598.253225 9993.165719
construction.year    -2.766637    1.000815
surface             -12.523781   -6.241597
no.rooms           -166.651179    5.423396
```

Une région de confiance pour q ($q \leq p$) paramètres β_j notés $(\beta_{j_1}, \ldots, \beta_{j_q})$ est donné par

$$\mathrm{RC}_{1-\alpha}(\mathbf{R}\boldsymbol{\beta}) = \Bigg\{ \mathbf{R}\boldsymbol{\beta} \in \mathbb{R}^q, \\ \frac{1}{q\hat{\sigma}^2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top \mathbf{R}^\top (\mathbf{R}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{R}^\top)^{-1}\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq f_{1-\alpha,q,n-p} \Bigg\}$$

où $\mathbf{R}$ est la matrice $q \times p$ dont tous les éléments sont nuls sauf les R_{ij_i} qui valent 1 et où $\mathrm{P}(F_{q,n-p} \leq f_{1-\alpha}) = 1 - \alpha$.

Soit $e_{ij} = ((\mathbf{X}^\top \mathbf{X})^{-1})_{ij}$, dans le dernier cas la RC s'écrit

$$\mathrm{RC}_{1-\alpha}(\beta_1, \beta_2) = \Bigg\{ (\beta_1, \beta_2) \in \mathbb{R}^2, \frac{1}{2\hat{\sigma}^2(e_{11}e_{22} - e_{12}^2)} \times$$

$$\Big(e_{22}(\hat{\beta}_1 - \beta_1)^2 - 2e_{12}(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) + e_{11}(\hat{\beta}_2 - \beta_2)^2 \Big) \leq f_{1-\alpha,2,n-p} \Bigg\}$$

C'est une ellipse centrée en $(\hat{\beta}_1, \hat{\beta}_2)$.

Régression Linéaire

Soit $\mathbf{x}_{n+1}$ une nouvelle valeur.
On veut prédire Y_{n+1} par $\hat{Y}_{n+1} = \mathbf{x}_{n+1}^\top \hat{\boldsymbol{\beta}}$
Un "IC pour Y_{n+1}" au niveau $1-\alpha$ est donné par

$$\left[\mathbf{x}'_{n+1}\hat{\boldsymbol{\beta}} \pm t_{1-\alpha/2,n-p}\hat{\sigma}\sqrt{1+\mathbf{x}'_{n+1}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}'^{\top}_{n+1}}\right]$$

```
> predict(reg, newdata = data.frame(construction.year
    =1992, surface=80, no.rooms=3), interval = "
    prediction")
       fit      lwr      upr
1 3544.494 2008.663 5080.326
> x=c(1,1992,80,3)
> t(x)%*%reg$coefficients
          [,1]
[1,] 3544.494
> t(x)%*%reg$coefficients+qt(c(.025,.975),n-4)*sqrt(
    sum(residus^2)/(n-4))*sqrt(1+t(x)%*%solve(t(X)%*%X
    )%*%x)
[1] 2008.663 5080.326
```

Modèsles imbriqués

Soit $Q \subset \{1, \ldots, p\}$ un ensemble de dimension q, et soit $\mathbf{X}_Q$ la matrice de design correspondant aux $\mathbf{x}_j, j \in Q$ et soit $\boldsymbol{\beta}_Q = (\beta_j)_{j \in Q}$.
On souhaite tester: $H_0 : \boldsymbol{\beta}_Q = \mathbf{0}$ contre $H_1 : \exists j \in Q, \beta_j \neq 0$.
Soit $Q^c = \{1, \ldots, p\} \setminus Q$. Sous H_0, $\mathbf{Y} = \mathbf{X}_{Q^c}\boldsymbol{\beta}_{Q^c} + \boldsymbol{\varepsilon}_0$.
Sous les hypothèses $\boldsymbol{\mathcal{H}_1 - \mathcal{H}_2}$ [Gauss], soit $Q \subset \{1, \ldots, p\}$ de dimension $1 \leq q < p$, $\mathbf{X}_Q = (\mathbf{x}_j, j \in Q)$, $\boldsymbol{\beta}_Q = (\beta_j)_{j \in Q}$ et soit $p_0 = p - q$. Pour tester les hypothèses

$$H_0 : \boldsymbol{\beta}_Q = \mathbf{0} \Leftrightarrow \mathbb{E}(\mathbf{Y}) \in \mathcal{V}(\mathbf{X}_{Q^c})$$

contre

$$H_1 : \exists j \in Q, \beta_j \neq 0 \Leftrightarrow \mathbb{E}(\mathbf{Y}) \in \mathcal{V}(\mathbf{X})$$

on s'appuie sur la statistique F qui sous H_0

$$F = \frac{\|\hat{\mathbf{Y}}_0 - \hat{\mathbf{Y}}\|^2/(p - p_0)}{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2/(n - p)} \sim \mathcal{F}_{p-p_0, n-p}$$

Au seuil $\alpha \in (0, 1)$ on rejette H_0 pour H_1si $f_{obs} > f_{1-\alpha, p-p_0, n-p}$ ou si p-value $= \mathbb{P}(F_{p-p_0, n-p} > f_{obs}) < \alpha$.

Régression Linéaire

Soit $\mathbf{R}$ une matrice de taille (q, p) et soit $\mathbf{r}_0 \in \mathbb{R}^q$. Pour tester les hypothèses $H_0 : \mathbf{R}\boldsymbol{\beta} = r_0$ contre $H_1 : \mathbf{R}\boldsymbol{\beta} \neq r_0$, on s'appuie sous la statistique de test F qui sous H_0 (et sous les hypothèses $\mathcal{H}_1 - \mathcal{H}_2$ [Gauss]) s'écrit et suit

$$F = \frac{1}{q\hat{\sigma}^2}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}_0)^\top (\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top)^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}_0) \sim \mathcal{F}_{q,n-p}.$$

```
> library(car)
> reg = lm(weight ~ reportedWeight, data=Davis)
> linearHypothesis(reg, diag(2), c(0,1))
Hypothesis:
(Intercept) = 0
reportedWeight = 1

Model 1: restricted model
Model 2: weight ~ reportedWeight
  Res.Df    RSS Df Sum of Sq    F   Pr(>F)
1    183 975.00
2    181 914.66  2    60.337 5.97 0.003085 **
```

On peut relâcher l'hypothèse $\mathcal{H}_2^{\text{Gauss}}$ en nous plaçant dans un cadre où les variables ε_i sont des v.a.i.i.d. ayant deux moments, i.e. $\mathcal{H}_2$.
on supposera ici que n est "assez" grand

$\mathcal{H}_3^{\text{tcl}}$: 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 Q est une matrice définie positive. De plus, $h_n = \max_{1 \leq i,j \leq n} (\mathcal{P}_{\mathbf{X}})_{ij} \to 0$ lorsque $n \to \infty$

Sous les hypothèses $\mathcal{H}_1$, $\mathcal{H}_2$ et $\mathcal{H}_3$ [tcl], alors lorsque $n \to \infty$

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2 \mathbf{Q}^{-1})$$

De plus $\hat{\sigma}^2 \to \sigma^2$ en probabilité, ce qui permet d'avoir

$$\hat{\sigma}^{-1}\mathbf{Q}^{1/2}\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbb{I}_p)$$

ou encore

$$\hat{\sigma}^{-1}(\mathbf{X}^\top\mathbf{X})^{1/2}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbb{I}_p)$$

soit, lorsque $n \to \infty$

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \approx \mathcal{N}\left(0, \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}\right).$$

Régression Linéaire

Sous les hypothèses $\mathcal{H}_1$, $\mathcal{H}_2$ et $\mathcal{H}_3$ [tcl]. Soit $j = 1, \ldots, p$ et soit $\mathbf{R}$ une matrice de taille (q, p) $(1 \leq q \leq p)$, alors lorsque $n \to \infty$

- $T_j = \dfrac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.
- $F = \dfrac{1}{q\hat{\sigma}^2}(\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^\top (\mathbf{R}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{R}^\top)^{-1}\mathbf{R}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathcal{L}} \chi^2_q/q$.

sous $\mathcal{H}_2$ [Gauss], $T_j \sim \mathcal{S}td_{n-p}$ et $F \sim \mathcal{F}_{q,n-p}$.
Rappel: $\mathcal{S}td_{n-p} \to \mathcal{N}(0,1)$ et $\mathcal{F}_{q,n-p} \to \chi^2_q/q$ lorsque $n \to \infty$.

```
> linearHypothesis(mod.davis, diag(2), c(0,1), test="
    Chisq")
Linear hypothesis test

  Res.Df     RSS Df Sum of Sq Chisq Pr(>Chisq)
1    183 975.00
2    181 914.66  2    60.337 11.94   0.002554 **
```