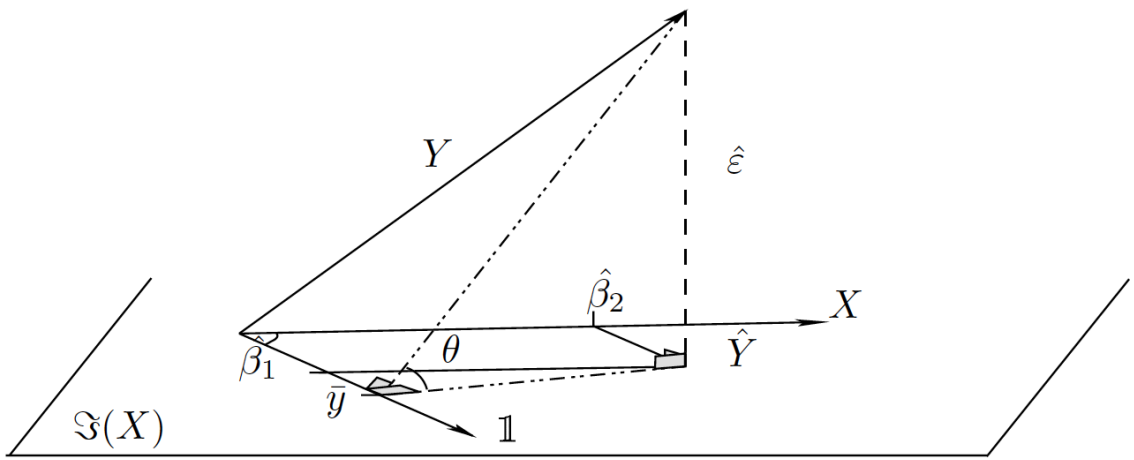
Modèles Linéaires Appliqués

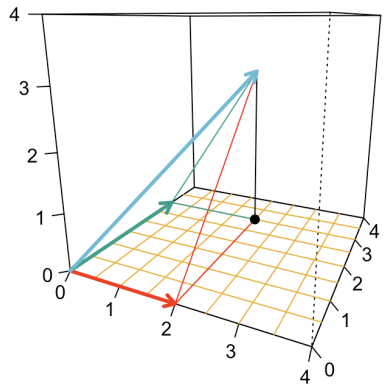
Arthur Charpentier

Automne 2020

OLS #7 (le modèle linéaire multiple - 2)

Résidus





via Cornillon & Matzner-Løber (2007)

$$\boldsymbol{\mathcal{P}}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top \text{ et } \boldsymbol{\mathcal{P}}_{\mathbf{X}^\perp} = (\mathbb{I}_n - \boldsymbol{\mathcal{P}}_{\mathbf{X}})$$

Résidus

On appelle résidus le vecteur défini par $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$. Les résidus constituent l'erreur d'approximation de $\mathbf{Y}$ par rapport à son projeté sur $\mathcal{V}(\mathbf{X})$. On a

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y} = (\mathbb{I}_n - \boldsymbol{\mathcal{P}}_{\mathbf{X}})\mathbf{Y} = \boldsymbol{\mathcal{P}}_{\mathbf{X}^\perp}\mathbf{Y} = \boldsymbol{\mathcal{P}}_{\mathbf{X}^\perp}\boldsymbol{\varepsilon}.$$

Sous les hypothèses $\boldsymbol{\mathcal{H}_1} - \boldsymbol{\mathcal{H}_2}$, on a

$$\mathbb{E}(\hat{\mathbf{Y}}) = \mathbf{X}\boldsymbol{\beta}, \quad \mathrm{Var}(\hat{\mathbf{Y}}) = \sigma^2\boldsymbol{\mathcal{P}}_{\mathbf{X}} = \sigma^2\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$$
$$\mathbb{E}(\hat{\boldsymbol{\varepsilon}}) = 0, \mathrm{Var}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2\boldsymbol{\mathcal{P}}_{\mathbf{X}^\perp} = \sigma^2(\mathbb{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top)$$
$$\mathrm{Cov}(\hat{\mathbf{Y}}, \hat{\boldsymbol{\varepsilon}}) = 0.$$

Résidus

- La statistique $\hat{\sigma}^2 = \dfrac{1}{n-p}\sum \hat{\varepsilon}_i^2 = \dfrac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{n-p}$ est un estimateur sans biais de σ^2.
- Un estimateur sans biais de $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}} = \mathrm{Var}(\hat{\boldsymbol{\beta}})$ est donné par

$$\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}$$

$\boldsymbol{\mathcal{H}_{2'}}$: Les erreurs sont centrées, de même variance, non corrélées et possèdent un moment d'ordre 4 fini pour tout n,
$\Leftrightarrow \forall n, \mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ et $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbb{I}_n$, $\mu_4 = \mathbb{E}(\varepsilon_i^4) < \infty$.

Sous les hypothèses $\boldsymbol{\mathcal{H}_1}, \boldsymbol{\mathcal{H}_{2'}}$ et $\boldsymbol{\mathcal{H}_3}$, les estimateurs $\hat{\sigma}^2$ et $\widehat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}$ sont respectivement des estimateurs convergents en moyenne quadratique de σ^2 et $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$.

Prévision

- Un des buts de la régression: proposer une nouvelle valeur pour la variable à expliquer pour une ou des nouvelles valeurs des covariables.
- $\mathbf{x}'_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,p})$ vecteur de covariables pour un nouvel individu.
- On veut prédire $Y_{n+1} = \mathbf{x}'_{n+1}\boldsymbol{\beta} + \varepsilon_{n+1}$ en supposant ε_{n+1} est ind. de $\boldsymbol{\varepsilon}$ et de même loi que ε_i (ou plus spécifiquement que $\mathcal{H}_{2'}$ est vraie).

Il est naturel de proposer

$$\hat{Y}_{n+1} = \mathbf{x}_{n+1}^{\top}\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1,1} + \cdots + \hat{\beta}_p x_{n+1,p}$$

Deux types d'erreur perturbent cette prévision: (i) incertitude sur ε_{n+1}; (ii) incertitude due à l'estimation. On s'intéresse à l'erreur de prédiction

Prévision

La variable $e_{n+1} = Y_{n+1} - \hat{Y}_{n+1}$ mesure l'erreur de prédiction.
Celle-ci est centrée et vérifie

$$e_{n+1} = \mathbf{x}'_{n+1}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\varepsilon} + \varepsilon_{n+1},$$

et

$$\mathrm{Var}(e_{n+1}) = \sigma^2\left(1 + \mathbf{x}'_{n+1}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}'^{\top}_{n+1}\right).$$

La variance de l'erreur de prévision peut facilement être estimée en remplaçant σ^2 par son estimateur consistant $\hat{\sigma}^2$.
Si $p = 2$,

$$\mathrm{Var}(e_{n+1}) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum(x_i - \overline{x})^2}\right)$$

Prévision

Le théorème de Pythagore donne:

$$\begin{aligned}\|\mathbf{Y}\|^2 &= \|\hat{\mathbf{Y}}\|^2 + \|\hat{\boldsymbol{\varepsilon}}\|^2 \\ &= \|\mathbf{X}\hat{\boldsymbol{\beta}}\|^2 + \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2.\end{aligned}$$

$$\begin{aligned}\|\mathbf{Y} - \overline{Y}\mathbf{1}\|^2 &= \|\hat{\mathbf{Y}} - \overline{Y}\mathbf{1}\|^2 + \|\hat{\boldsymbol{\varepsilon}}\|^2 \\ SCT &= SCE + SCR\end{aligned}$$

(Somme des carrés Totaux = SC expliquée par le modèle + SC résiduelle).

Prévision

Le coefficient de détermination (multiple), noté R^2 est défini par

$$R^2 = \frac{\|\hat{\mathbf{Y}}\|^2}{\|\mathbf{Y}\|^2} = \cos^2(\theta) \in [0, 1].$$

On peut écrire

$$R^2 = \frac{\text{Variation expl. par modèle}}{\text{Variation totale}} = \frac{\|\hat{\mathbf{Y}} - \overline{Y}\|^2}{\|\mathbf{Y} - \overline{Y}\|^2} = 1 - \frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{\|\mathbf{Y} - \overline{Y}\|^2} = \cos^2(\theta).$$

$R^2 = 0 \Rightarrow \hat{Y}_i = \overline{Y}$, modèle de régression inadapté.
Lorsque $p = 2$

$$R^2 = \frac{\widehat{\text{Cov}}(\mathbf{x}, \mathbf{y})^2}{\widehat{\text{Var}}(\mathbf{x})\widehat{\text{Var}}(\mathbf{y})} = \widehat{\text{Corr}}(\mathbf{x}, \mathbf{y})^2.$$

Prévision

Lorsque l'on augmente le nombre de covariables, le coefficient R^2 augmente nécessairement.
Le coefficient de détermination ajusté R_a^2 est défini par

$$R_a^2 = 1 - \frac{n}{n-p}\frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{\|\mathbf{Y}\|^2}.$$

De plus

$$R_a^2 = 1 - \frac{n-1}{n-p}\frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{\|\mathbf{Y} - \overline{Y}\|^2} = 1 + \frac{n-1}{n-p}(R^2 - 1).$$

Aspects Computationnels

```
> 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
```

Les autres valeurs sont associées à des tests: il nous faut des hypothèses supplémentaires pour avoir la loi des estimateurs (et faire des tests).