Compléments sur la Régression linéaire

Matthieu KOWALSKI

1 Estimation par maximum de vraisemblance des paramètres

Rappel du modèle :

$$Y_i = bX_i + a + \varepsilon_i \ \forall \ i \in \{1, \dots, N\} \ , \tag{1}$$

où les ε_i sont des V.A. i.i.d tels que $\forall i \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. On en déduit

$$Y_i \sim \mathcal{N}(a + bX_i, \sigma^2) \ \forall i \ .$$

On a donc

$$f(Y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{\sigma\sqrt{2\pi}(Y_i - a - bX_i)^2} \right) \ \forall \ i \ ,$$

ce qui permet d'écrire la vraisemblance du système :

$$L(a, b, \sigma^2, Y_1, \dots, Y_N) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{\sigma\sqrt{2\pi}(Y_i - a - bx_i)^2} \right)$$

d'où l'on déduit la log-vraisemblance :

$$\ln\left(L(a, b, \sigma^2, Y_1, \dots, Y_N)\right) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - a - bX_i)^2 \ .$$

Maximiser la vraisemblance du système revient donc à minimiser le critère des moindres carrés :

$$Q(a, b) = \sum_{i=1}^{N}(Y_i - a - bX_i)^2 \ .$$

Pour cela, on cherche à annuler les dérivées partielles :

$$
\begin{aligned}
&\begin{cases} \frac{\partial Q}{\partial a}(a,b) &= -2\sum_{i=1}^N Y_i - a - bX_i = 0 \\ \frac{\partial Q}{\partial b}(a,b) &= -2\sum_{i=1}^N X_i(Y_i - a - bX_i) = 0 \end{cases} \\
\Leftrightarrow &\begin{cases} \bar{Y} = a + b\bar{X} \\ \sum_{i=1}^N X_i(Y_i - a - bX_i) = 0 \end{cases} \\
\Leftrightarrow &\begin{cases} a = \bar{Y} - b\bar{X} \\ \sum_{i=1}^N X_iY_i - (\bar{Y} - b\bar{X})n\bar{X} - b\sum_{i=1}^N X_i^2 = 0 \end{cases} \\
\Leftrightarrow &\begin{cases} a = \bar{Y} - b\bar{X} \\ \sum_{i=1}^N X_iY_i - n\bar{Y}\bar{X} + nb\bar{X}^2 - b\sum_{i=1}^N X_i^2 = 0 \end{cases} \\
\Leftrightarrow &\begin{cases} a = \bar{Y} - b\bar{X} \\ b = \frac{\sum_{i=1}^N X_iY_i - n\bar{Y}\bar{X}}{\sum_{i=1}^N X_i^2 - n\bar{X}^2} = 0 \end{cases}
\end{aligned}
$$

avec $\bar{Y}$ (resp. $\bar{X}$) la moyenne empirique des Y_i (resp. X_i).

On note S_X^2 la variance empirique des X_i et $S_{XY} = \frac{1}{n}\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{n}\sum_{i=1}^N (X_iY_i) - \bar{X}Y$ la covariance empirique des Y_i et X_i. Les estimateurs au sens des moindres carrés de a et b sont donnés par :

$$
\hat{a} = \bar{Y} - \frac{S_{XY}^2}{S_X^2}\bar{X} \quad \text{et} \quad \hat{b} = \frac{S_{XY}^2}{S_X^2} \ .
$$

2 Intervalle de confiance sur $\mathbb{E}\{Y_0\}$

On cherche une fonction pivot. L'estimation ponctuelle de $\mathbb{E}\{Y_0\}$ donne $\mathbb{E}\{\hat{Y}_0\} = \mathbb{E}\{Y_0\}$. De plus

$$
\begin{aligned}
\mathrm{Var}(Y_0) = \mathrm{Var}(\hat{a} + \hat{b}X_0) &= \mathrm{Var}(\hat{a}) + X_0^2\mathrm{Var}(\hat{b}) + 2X_0\mathrm{Cov}(\hat{a},\hat{b}) && (2) \\
&= \frac{\sigma^2}{N}(1 + \frac{\bar{X}^2}{S_X^2}) + X_0^2\frac{\sigma^2}{NS_X^2} - 2X_0\frac{\bar{X}\sigma^2}{NS_X^2} && (3) \\
&= \sigma^2\left(\frac{1}{N} + \frac{\bar{X}^2 - 2\bar{X}X_0 + X_0^2}{NS_X^2}\right) = \sigma^2\left(\frac{1}{N} + \frac{(X_0 - \bar{X})}{NS_X^2}\right) \ . && (4)
\end{aligned}
$$

On a donc :

$$
\frac{\hat{Y}_0 - \mathbb{E}\{Y_0\}}{\sigma\sqrt{\left(\frac{1}{N} + \frac{(X_0 - \bar{X})}{NS_X^2}\right)}} \sim \mathcal{N}(0,1) \quad \text{et} \quad \frac{\hat{Y}_0 - \mathbb{E}\{Y_0\}}{\hat{\sigma}\sqrt{\left(\frac{1}{N} + \frac{(X_0 - \bar{X})}{NS_X^2}\right)}} \sim t_{N-2}(0,1)(\star) \ .
$$

On admettra la partie de droite. On rappelle que $\hat{\sigma} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - a - bX_i)^2 = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_i$.

$(\star)$ est une fonction pivotable pour $\mathbb{E}\{Y_0\}$ qui permet d'obtenir l'intervalle de confiance :

$$\left[\hat{Y}_0 - t_{N-2;1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{N} + \frac{(X_0-\bar{X})^2}{NS_X^2}}, \hat{Y}_0 + t_{N-2;1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\frac{1}{N} + \frac{(X_0-\bar{X})^2}{NS_X^2}}\right]$$

3 Intervalle de prédiction sur $\mathbb{E}\{Y_0\}$

On cherche deux bornes A et B telles que $\mathbb{P}\{A \leq Y_0 \leq B\} = 1-\alpha$ (α une erreur "petite" fixée).

On a

$$\mathbb{E}\{\hat{Y}_0 - Y_0\} = \mathbb{E}\{\hat{Y}_0\} - \mathbb{E}\{Y_0\} = 0 \ .$$

Vue que $\mathbb{E}\{Y_0\}$ et Y_0 sont indépendant, on a

$$\begin{aligned}\mathrm{Var}(\hat{Y}_0 - Y_0) &= \mathrm{Var}(\hat{Y}_0) + \mathrm{Var}(Y_0) = \sigma^2\left(\frac{1}{N} + \frac{(X_0-\bar{X})^2}{NS_x^2} + \sigma^2\right) \\ &= \sigma^2\left(1 + \frac{1}{N} + \frac{(X_0-\bar{X})^2}{NS_X^2}\right).\end{aligned}$$

On a donc

$$\frac{\hat{Y}_0 - Y_0}{\sigma^2\left(1 + \frac{1}{N} + \frac{(X_0-\bar{X})^2}{NS_X^2}\right)} \sim \mathcal{N}(0,1) \ .$$

On remplace σ par son estimation $\hat{\sigma}$, et on obtient :

$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma}^2\left(1 + \frac{1}{N} + \frac{(X_0-\bar{X})^2}{NS_X^2}\right)} \sim t_{N-2} \ .$$

D'où l'intervalle de prédiction pour $\mathbb{E}\{Y_0\}$:

$$\left[\hat{Y}_0 - t_{N-2;1-\frac{\alpha}{2}}\hat{\sigma}^2\left(1 + \frac{1}{N} + \frac{(X_0-\bar{X})^2}{NS_X^2}\right), \hat{Y}_0 + t_{N-2;1-\frac{\alpha}{2}}\hat{\sigma}^2\left(1 + \frac{1}{N} + \frac{(X_0-\bar{X})^2}{NS_X^2}\right)\right]$$