Régression linéaire

13 novembre 2008

Exercice 1 : Régression linéaire simple

Considérons le cas de la régression linéaire simple $\hat{y} = w_0 + w_1 x$, avec $x, y, w_0, w_1 \in \mathbb{R}$. Les observations initiales sont $\{(x_i, y_i)\}_{1 \leq i \leq n}$, liées par $y_i = w_0 + w_1 x_i + \epsilon_i$. Supposons que $\bar{x}$ est la moyenne des $\{x_i\}$, $\bar{y}$ la moyenne des $\{y_i\}$, σ_x l'écart-type des $\{x_i\}$, $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$, et σ_y l'écart-type des $\{y_i\}$, $\sigma_y^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \bar{y})^2$. On note par

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{1}$$

la covariance empirique entre les variables X et Y.

1. Montrer que $\sigma_{xy} = \frac{1}{n}\sum_{i=1}^n x_i y_i \; - \; \bar{x}\bar{y}$.
2. Déterminer les expressions de $\hat{w}_0$ et $\hat{w}_1$ qui minimisent la somme des carrés des résidus. Montrer que $\bar{\hat{y}} = \bar{y}$.
3. Si $\sigma_{\hat{y}}^2 = \frac{1}{n}\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2$ et $\sigma_\epsilon^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{y}_i)^2$, obtenir l'*équation d'analyse de la variance* (la variance totale est la somme entre la variance expliquée et la variance résiduelle)
$$\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_\epsilon^2 \tag{2}$$
4. Montrer que le coefficient de détermination, défini par
$$r^2 = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2} \tag{3}$$
est égal au carré du coefficient de corrélation linéaire
$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \, \sigma_y} \tag{4}$$
5. Déterminer les expressions de $\hat{a}_0, \hat{a}_1 \in \mathbb{R}$ qui minimisent la somme des carrés des résidus pour la régression $\hat{x} = a_0 + a_1 y$.

Solution

1. En développant (1) on obtient

$$\begin{aligned}
\sigma_{xy} &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{\bar{x}}{n}\sum_{i=1}^{n} y_i - \frac{\bar{y}}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} \bar{x}\bar{y} \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{\bar{x}}{n} \times n\bar{y} - \frac{\bar{y}}{n} \times n\bar{x} + \bar{x}\bar{y} \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - 2\bar{x}\bar{y} + \bar{x}\bar{y} \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}
\end{aligned}$$

2. La somme des carrés des résidus est

$$\begin{aligned}
s(w_0, w_1) &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (5) \\
&= \sum_{i=1}^{n} (y_i - w_0 - w_1 x_i)^2 \qquad (6)
\end{aligned}$$

Comme $w_0, w_1 \in \mathbb{R}$ et $s(w_0, w_1)$ est différentiable, les arguments $\hat{w}_0, \hat{w}_1$ pour lesquels $s(w_0, w_1)$ prend des valeurs qui sont des extrêma locaux doivent satisfaire

$$\frac{\partial s}{\partial w_0}(\hat{w}_0, \hat{w}_1) = 0 \text{ et } \frac{\partial s}{\partial w_1}(\hat{w}_0, \hat{w}_1) = 0 \qquad (7)$$

ce qui mène, après calcul, à

$$\sum_{i=1}^{n} (y_i - \hat{w}_0 - \hat{w}_1 x_i) = 0 \text{ et } \sum_{i=1}^{n} (y_i - \hat{w}_0 - \hat{w}_1 x_i) x_i = 0 \qquad (8)$$

ou encore

$$\hat{w}_0 = \bar{y} - \hat{w}_1 \bar{x} \text{ et } \sum_{i=1}^{n} (y_i - \bar{y} - \hat{w}_1 \bar{x} - \hat{w}_1 x_i) x_i = 0 \qquad (9)$$

d'où on obtient

$$\hat{w}_0 = \bar{y} - \hat{w}_1 \bar{x} \text{ et } \hat{w}_1 = \frac{\sigma_{xy}}{\sigma_x^2} \qquad (10)$$

Il est facile de montrer que l'extrêmum correspondant de $s(w_0, w_1)$ est un minimum.

Ensuite, nous avons

$$\begin{aligned}
\bar{\hat{y}} &= \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{w}_0 + \hat{w}_1 x_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\underbrace{\bar{y} - \hat{w}_1\bar{x}}_{\hat{w}_0} + \hat{w}_1 x_i) \\
&= \bar{y} + \frac{\hat{w}_1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \\
&= \bar{y} + \hat{w}_1(\bar{x} - \bar{x}) \\
&= \bar{y}
\end{aligned}$$

3. En développant σ_y^2 on obtient

$$\begin{aligned}
\sigma_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\sigma_{\hat{y}}^2} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\sigma_\epsilon^2} + \frac{2}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})
\end{aligned}$$

En poursuivant le calcul du dernier terme,

$$\begin{aligned}
(y_i - \hat{y}_i) &= y_i - \hat{w}_0 - \hat{w}_1 x_i \\
&= (y_i - \bar{y}) - \hat{w}_1(x_i - \bar{x}) \\
(\hat{y}_i - \bar{y}) &= \hat{w}_0 + \hat{w}_1 x_i - \bar{y} \\
&= \hat{w}_1(x_i - \bar{x}) \\
\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \hat{w}_1 \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}_{\sigma_{xy}} + \hat{w}_1^2 \underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}_{\sigma_x^2} \\
&= \hat{w}_1\left(\sigma_{xy} - \frac{\sigma_{xy}}{\sigma_x^2}\sigma_x^2\right) \\
&= 0
\end{aligned}$$

ce qui permet d'arriver à (2).

4. Par définition, $\hat{y}_i = \hat{w}_0 + \hat{w}_1 x_i$, or $\hat{w}_0 = \bar{y} - \hat{w}_1 \bar{x}$, donc

$$\begin{aligned} \hat{y}_i - \bar{y} &= \hat{w}_1 (x_i - \bar{x}) \\ (\hat{y}_i - \bar{y})^2 &= \hat{w}_1^2 (x_i - \bar{x})^2 \\ \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 &= \frac{\hat{w}_1^2}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ \sigma_{\hat{y}}^2 &= \hat{w}_1^2 \, \sigma_x^2 \\ \sigma_{xy}^2 &= \sigma_{\hat{y}}^2 \, \sigma_x^2 \quad \left(\text{car } \hat{w}_1 = \frac{\sigma_{xy}}{\sigma_x^2} \right) \end{aligned}$$

ce qui implique

$$\rho_{xy}^2 = \frac{\sigma_{xy}^2}{\sigma_x^2 \, \sigma_y^2} = \frac{\sigma_{\hat{y}}^2 \, \sigma_x^2}{\sigma_x^2 \, \sigma_y^2} = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2} = r^2 \tag{11}$$

5. On montre (comme pour $\hat{y} = w_0 + w_1 x$) que pour la régression $\hat{x} = a_0 + a_1 y$ les valeurs des paramètres qui minimisent la somme des carrés des résidus sont

$$\hat{a}_0 = \bar{x} - \hat{a}_1 \bar{y} \text{ et } \hat{a}_1 = \frac{\sigma_{xy}}{\sigma_y^2} \tag{12}$$

donc

$$x_i = \hat{a}_0 + \hat{a}_1 y_i + \xi_i \tag{13}$$

Pour la solutions des moindres carrés de la régression $\hat{y} = w_0 + w_1 x$,

$$y_i = \hat{w}_0 + \hat{w}_1 x_i + \epsilon_i \tag{14}$$

donc en exprimant x_i on obtient

$$x_i = -\frac{\hat{w}_0}{\hat{w}_1} + \frac{1}{\hat{w}_1} x_i - \frac{\epsilon_i}{\hat{w}_1} \tag{15}$$

Pour que le coefficient de x_i soit le même, il faudrait que $\frac{1}{\hat{w}_1} = \hat{a}_1$, ou encore $\frac{\sigma_x^2}{\sigma_{xy}} = \frac{\sigma_{xy}}{\sigma_y^2}$, donc que $\sigma_{xy}^2 = \sigma_x^2 \, \sigma_y^2$. Mais nous avions constaté que $\sigma_{xy}^2 = \sigma_x^2 \, \sigma_{\hat{y}}^2$ et l'équation d'analyse de la variance nous indique que $\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_\epsilon^2$, donc $\sigma_{xy}^2 = \sigma_x^2 \, \sigma_y^2$ ne peut être vrai que si $\sigma_\epsilon^2 = 0$, c'est à dire si pour la régression $\hat{y} = w_0 + w_1 x$ (et implicitement pour $\hat{x} = a_0 + a_1 y$) tous les résidus sont nuls.

Exercice 2 : Dépendances non linéaires et régression linéaire

La régression linéaire peut être employée pour déterminer les paramètres d'une dépendance non linéaire (entre la variable expliquée et les paramètres, entre la variable

expliquée et la variable explicative). Bien entendu, la forme de la dépendance doit être connue.

Considérons la loi de Laplace pour une transformation adiabatique (sans échange de chaleur avec le milieu extérieur) d'un gaz parfait :

$$PV^\gamma = C \tag{16}$$

Les constantes (paramètres) γ et C dépendent du gaz choisi. On suppose que la V est la variable explicative et P la variable expliquée. A partir d'un ensemble d'observations $\{(V_i, P_i)\}_{1\leq i\leq n}$, nous souhaitons estimer des valeurs de γ et C telles que la somme des carrés des écarts de la variable expliquée soit minimale. Indication : pour utiliser la régression linéaire il faut définir de nouvelles variables telles que la dépendance devienne linéaire.

Solution

En appliquant le logarithme à la loi $PV^\gamma = C$ nous obtenons la dépendance suivante :

$$\ln P + \gamma \ln V = \ln C \tag{17}$$

En définissant les nouvelles variables $X = \ln V$ et $Y = \ln P$, nous pouvons mettre (17) sous une forme qui correspond directement à la régression linéaire

$$Y = \ln C - \gamma X \tag{18}$$

On note $w_0 = \ln C$ et $w_1 = -\gamma$. Il est maintenant facile d'appliquer la régression linéaire pour déterminer les paramètres optimaux $\hat{w}_0, \hat{w}_1$, à partir des observations $\{(x_i, y_i)\}_{1\leq i\leq n} = \{(V_i, P_i)\}_{1\leq i\leq n}$, ce qui nous donne ensuite les estimations $\hat{C} = e^{\hat{w}_0}$ et $\hat{\gamma} = -\hat{w}_1$.

Remarque importante : les transformations $X = \ln V$ et $Y = \ln P$ peuvent amplifier fortement le "bruit de mesure" de V ou de P pour des valeurs très proches de 0 (car le logarithme diverge lorsque l'argument tend vers 0), avec un impact négatif sur la qualité d'estimation des paramètres. Il peut être nécessaire d'éliminer de l'ensemble $\{(V_i, P_i)\}_{1\leq i\leq n}$ les observations pour lesquelles soit P, soit V sont proches de 0.