Regresión lineal simple

El **método de los mínimos cuadrados** proporciona los coeficientes que minimizan el error SSE

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}.$$

$$\mathrm{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

1. $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$, donde $S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$ y $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$
2. $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$.

Propiedades de los estimadores de mínimos cuadrados

1. Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ son insesgados, es decir, $E(\hat{\beta}_i) = \beta_i$, para $i = 0, 1$.
2. $V(\hat{\beta}_0) = c_{00}\sigma^2$, donde $c_{00} = \sum x_i^2/(nS_{xx})$.
3. $V(\hat{\beta}_1) = c_{11}\sigma^2$, donde $c_{11} = \dfrac{1}{S_{xx}}$.
4. $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = c_{01}\sigma^2$, donde $c_{01} = \dfrac{-\overline{x}}{S_{xx}}$.
5. Un estimador insesgado de σ^2 es $S^2 = \text{SSE}/(n-2)$, donde $\text{SSE} = S_{yy} - \hat{\beta}_1 S_{xy}$ y $S_{yy} = \sum(y_i - \overline{y})^2$.

Si, además, el ε_i, para $i = 1, 2, \ldots, n$ está distribuido normalmente,

6. $\hat{\beta}_0$ y $\hat{\beta}_1$ están distribuidas normalmente.
7. La variable aleatoria $\dfrac{(n-2)S^2}{\sigma^2}$ tiene una distribución χ^2 con $n-2$ grados de libertad.
8. El estadístico S^2 es independiente de $\hat{\beta}_0$ y $\hat{\beta}_1$.

Inferencias respecto a los parámetros β_i

$$H_0 : \beta_i = \beta_{i0}.$$

$$H_a : \begin{cases} \beta_i > \beta_{i0} & \text{(región de rechazo de cola superior),} \\ \beta_i < \beta_{i0} & \text{(región de rechazo de cola inferior),} \\ \beta_i \neq \beta_{i0} & \text{(región de rechazo de dos colas).} \end{cases}$$

Estadístico de prueba:

$$T = \frac{\hat{\beta}_i - \beta_{i0}}{S\sqrt{c_{ii}}}$$

$$c_{00} = \frac{\sum x_i^2}{nS_{xx}} \quad c_{11} = \frac{1}{S_{xx}}$$

$$\text{Región de rechazo:} \begin{cases} t > t_\alpha & \text{(alternativa de cola superior),} \\ t < -t_\alpha & \text{(alternativa de cola inferior),} \\ |t| > t_{\alpha/2} & \text{(alternativa de dos colas),} \end{cases}$$

t_α está basada en $(n - 2)$ grados de libertad

Un intervalo de confianza $100(1 - \alpha)\%$ para β_i $\quad \hat{\beta}_i \pm t_{\alpha/2} S\sqrt{c_{ii}}$

Inferencias respecto a funciones lineales de los parámetros del modelo: regresión lineal simple

Una prueba para $\theta = a_0\beta_0 + a_1\beta_1$

$$H_0 : \theta = \theta_0,$$

$$H_a : \begin{cases} \theta > \theta_0, \\ \theta < \theta_0, \\ \theta \neq \theta_0. \end{cases}$$

$$\text{Estadístico de prueba: } T = \frac{\hat{\theta} - \theta_0}{S\sqrt{\left(\dfrac{a_0^2 \dfrac{\sum x_i^2}{n} + a_1^2 - 2a_0 a_1 \overline{x}}{S_{xx}}\right)}}.$$

$$\text{Región de rechazo: } \begin{cases} t > t_\alpha, \\ t < -t_\alpha, \\ |t| > t_{\alpha/2}. \end{cases}$$

Aquí, t_α y $t_{\alpha/2}$ están basados en $n - 2$ grados de libertad.

Un intervalo de confianza $100(1-\alpha)\%$ para $\theta = a_0\beta_0 + a_1\beta_1$

$$\hat{\theta} \pm t_{\alpha/2} S \sqrt{\left(\frac{a_0^2 \dfrac{\sum x_i^2}{n} + a_1^2 - 2a_0 a_1 \overline{x}}{S_{xx}}\right)},$$

Un intervalo de confianza $100(1-\alpha)\%$ para $E(Y) = \beta_0 + \beta_1 x^*$

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S \sqrt{\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}},$$

donde la $t_{\alpha/2}$ tabulada está basada en $n - 2$ grados de libertad.

Predicción de un valor particular de Y

Intervalo de predicción de $100(1-\alpha)\%$ para Y cuando $x = x^*$

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}}.$$

Coeficiente de correlación r, y de determinación r^2

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \hat{\beta}_1 \sqrt{\frac{S_{xx}}{S_{yy}}}$$

$$r^2 = \left(\frac{\hat{\beta}_1 S_{xy}}{S_{yy}} \right) = \frac{S_{yy} - \mathrm{SSE}}{S_{yy}} = 1 - \frac{\mathrm{SSE}}{S_{yy}}$$

Es la proporción de la variación total en las y_i que es explicada por la variable x en un modelo de regresión lineal simple

Pruebas de hipótesis del coeficiente de correlación para muestras grandes

$$t = \frac{\hat{\beta}_1 - 0}{S/\sqrt{S_{xx}}} \Longleftrightarrow \boxed{t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}}$$

distribución t con n – 2 grados de libertad

Probar $H_0: \rho = 0$ contra $H_a: \rho > 0$ es equivalente a probar $H_0: \beta_1 = 0$ contra $H_a: \beta_1 > 0$. Del mismo modo, $H_a: \rho < 0$ es equivalente a $H_a: \beta_1 < 0$ y $H_a: \rho \neq 0$ es equivalente a $H_a: \beta_1 \neq 0$.

$$Z = \frac{\left(\frac{1}{2}\right)\ln\left(\frac{1+r}{1-r}\right) - \left(\frac{1}{2}\right)\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\frac{1}{\sqrt{n-3}}}$$

$H_a: \rho > \rho_0, \quad \text{RR}: z > z_\alpha,$

$H_a: \rho < \rho_0, \quad \text{RR}: z < -z_\alpha,$

$H_a: \rho \neq \rho_0, \quad \text{RR}: |z| > z_{\alpha/2}$

α es la probabilidad de cometer un error tipo I