Análisis de regresión simple

El modelo lineal en una única variable independiente es

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

ε_i es una variable aleatoria que satisface los siguientes supuestos:

i) $E[\varepsilon_i] = 0 \; \forall i$

ii) $V(\varepsilon_i) = \sigma_\varepsilon^2 \; \forall i \; (homocedasticidad)$

iii) $Cov(\varepsilon_i, \varepsilon_j) = 0 \; \forall i, j \; (errores \; no \; correlacionados)$

β_0: coeficiente de posición y representa el valor de la ordenada cuando x=0 (no siempre es interpretable)

β_1: pendiente del modelo, es decir, cambio en la variable dependiente y_i cuando x_i se incrementa en una unidad

β_0 y β_1 se llaman parámetros del modelo

¿cómo estimamos β_0 y β_1?

Para estimar β_0 y β_1 empleamos el método de los Mínimos Cuadrados Ordinarios (MCO) que consiste en minimizar la suma de los errores al cuadrado, dado por

$$SCE = \sum_{i=1}^{n} \varepsilon_i^2$$

pero $\varepsilon_i = \text{valor real} - \text{valor estimado} = y_i - \beta_0 - \beta_1 x_i$

Si el modelo subestima

$$\text{valor real} > \text{valor estimado} \Rightarrow \varepsilon_i > 0$$

Si el modelo sobre estima

$$\text{valor real} < \text{valor estimado} \Rightarrow \varepsilon_i < 0$$

Por lo tanto, $min \; SCE = min \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

Del cálculo para encontrar los valores de β_0 y β_1 debemos resolver el sistema de ecuaciones

$$\frac{\partial SCE}{\partial \beta_0} = 0$$

$$\frac{\partial SCE}{\partial \beta_1} = 0$$

$$\frac{\partial SCE}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial SCE}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

$$\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0$$

$$\sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} {x_i}^2 = 0$$

Ecuaciones normales

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} {x_i}^2 = \sum_{i=1}^{n} x_i y_i$$

Además, sabemos que

$$\sum_{i=1}^{n} x_i = n\bar{x} \Rightarrow \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

$$\sum_{i=1}^{n} y_i = n\bar{y}$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2$$

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

Resolviendo para β_0 y β_1 se tiene

$$\beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\beta_0 = \frac{1}{n}\left(\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i\right)$$

Teorema de Gauss-Markov

Bajos los supuestos del modelo $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores insesgados y de mínima varianza

En efecto

$$\hat{\beta}_1 = \sum_{i=1}^{n} k_i \hat{y}_i$$

Siendo

$$k_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Así

$$\hat{\beta}_1 = \sum_{i=1}^{n} k_i \hat{y}_i = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sum_{i=1}^{n} k_i (y_i - \bar{y}) = \sum_{i=1}^{n} k_i y_i - \sum_{i=1}^{n} k_i \bar{y}$$

$$= \sum_{i=1}^{n} k_i y_i - \bar{y} \sum_{i=1}^{n} k_i$$

Con esto se tiene

i) $\sum_{i=1}^{n} k_i = 0$

ii) $\sum_{i=1}^{n} x_i k_i = 1$

i) $\sum_{i=1}^{n} k_i = 0$

$$\sum_{i=1}^{n} k_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sum_{i=1}^{n}(x_i - \bar{x})$$

$$\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\left[\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}\right] = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}[n\bar{x} - n\bar{x}] = 0$$

ii) $\sum_{i=1}^{n} x_i k_i = 1$

$$\sum_{i=1}^{n} x_i k_i = \sum_{i=1}^{n} \frac{(x_i - \bar{x})x_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \sum_{i=1}^{n}(x_i - \bar{x})x_i$$

$$= \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\left[\sum_{i=1}^{n} {x_i}^2 - \sum_{i=1}^{n} \bar{x}x_i\right] = \frac{\sum_{i=1}^{n} {x_i}^2 - \sum_{i=1}^{n} \bar{x}x_i}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} {x_i}^2 - \bar{x}n\bar{x}}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2} = 1$$

Retomando

$$E[\hat{\beta}_1] = E\left[\sum_{i=1}^{n} k_i \hat{y}_i\right] = \sum_{i=1}^{n} k_i E[\hat{y}_i] = \sum_{i=1}^{n} k_i E[\beta_0 + \beta_1 x_i]$$

$$= \sum_{i=1}^{n} k_i E[\beta_0] + \sum_{i=1}^{n} k_i E[\beta_1 x_i] = \beta_0 \sum_{i=1}^{n} k_i + \beta_1 \sum_{i=1}^{n} k_i x_i = \beta_1$$

Inferencia sobre $\boldsymbol{\beta_1}$

Una de las inferencias más importantes respecto de β_1 es:

$$H_0: \beta_1 = 0$$

$$H_0: \beta_1 \neq 0$$

Recordemos que es estimador de β_1 es $\hat{\beta}_1$. Para determinar la distribución muestral de $\hat{\beta}_1$ haremos el supuesto distribucional sobre la variable aleatoria ε_i la cual nos dice que es Normal e independiente (NID)

$$\varepsilon_i \sim NID(0, \sigma^2)$$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Por teorema $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

Por otra parte como

$$\hat{\beta}_1 = \sum_{i=1}^{n} k_i \hat{y}_i$$

Entonces $\hat{\beta}_1 \sim N(\beta_1, \sigma^2(\hat{\beta}_1))$ *con* $\sigma^2(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

- Si σ^2 es conocido entonces el estadístico de prueba es

$$z = \frac{\hat{\beta}_1 - \beta_1}{\sigma(\hat{\beta}_1)} \sim N(0,1)$$

  Así, H_0 será rechazada si p-value< α para algún nivel de significancia dado.

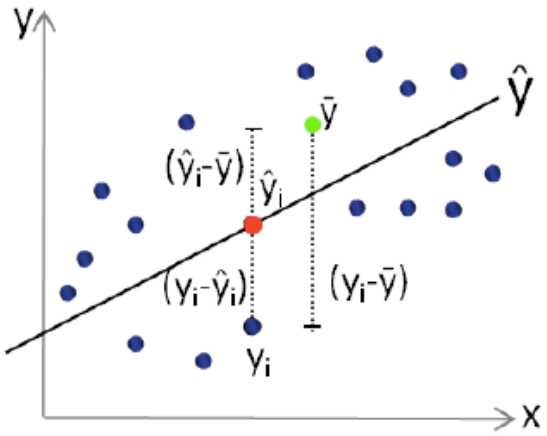
- Si σ^2 es desconocido entonces el estadístico de prueba es

$$t = \frac{\hat{\beta}_1 - \beta_1}{S(\hat{\beta}_1)} \sim t_{n-2}$$

$$S^2(\hat{\beta}_1) = \sigma^2(\hat{\beta}_1) = \frac{CME}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

  Así, H_0 será rechazada si p-value< α para algún nivel de significancia dado.

Análisis de varianza, para el análisis de regresión



$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2$$

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

Por lo tanto

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$\sum_{i=1}^{n}(y_i - \bar{y})^2$ es la suma de cuadrados totales = SCT

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ es la suma de cuadrados del error = SCE

$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ es la suma de cuadrados de la regresión = SCR

Así,

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SCT = SCR + SCE$$

Tabla ANOVA

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	F	p-value
Regresión o modelo	1	SCR	$\text{CMR} = \frac{\text{SCR}}{1}$	$F_c = \frac{CMR}{CME}$	$P(F > F_c)$
Error o residuo	$n-2$	SCE	$\text{CME} = \frac{\text{SCE}}{\text{n}-2}$		
Total	$n-1$	SCT			

Esta tabla ANOVA sirve para contrastar la hipótesis

$$H_0: \beta_1 = 0$$

$$H_0: \beta_1 \neq 0$$

Cuando se vea regresión multivariada, tomará más relevancia esta tabla.

$$SCT = SCR + SCE$$

$$1 = \frac{SCR}{SCT} + \frac{SCE}{SCT} \Rightarrow R^2 = \frac{SCR}{SCT}$$

$$1 = R^2 + \frac{SCE}{SCT}$$

$$R^2 = 1 - \frac{SCE}{SCT}$$

R^2 recibe el nombre de coeficiente de determinación y su interpretación es dada en término del porcentaje de variabilidad total que queda explicada por las variables independientes en el modelo. R^2 es considerado también como la capacidad predictiva que tiene el modelo.

$$0 \leq R^2 \leq 1$$

Mientras más cercano a uno sea este valor, el modelo de considera que representa de mejor forma los datos y por lo tanto, puede ser usado con mayor confianza para predecir los valores de la variable respuesta.