REGRESION LINEAL SIMPLE

FORMULARIO

Jorge Galbiati Riesco

Modelo de Regresión Lineal Simple

$$\mathbf{y} = \alpha + \beta\,\mathbf{x} + \varepsilon$$

en que α y β son fijos, ε es una variable aleatoria con esperanaza $E(\varepsilon) = 0$ y varianza $V(\varepsilon) = \sigma^2$ fija. Los parámetros del modelo son α, β y σ^2.

x representa la variable *independiente*, que toma valores fijos determinados por el experimentador. **y** es la variable *dependiente*, que es aleatoria por depender de ε.

El valor esperado de **y** dado algún valor x de **x**, es

$$y = \alpha + \beta\, x$$

que se denomina *recta de regresión*.

α es el intercepto de la recta con el eje de las y.

β es la pendiente de la recta.

Para estimar los parámetros del modelo, se dispone de una muestra de *n* pares (x_i, y_i), *i=1,2,...,n,* que corresponden a observaciones de un experimento en que el experimentador asignó valores arbitrarios a la variable independiente *x* y observó los correspondientes resultados de la variable *y*, que supone se comporta de acuerdo al modelo de regresión lineal simple. En tal caso las observaciones obedecen a la relación

$$y_i = \alpha + \beta \cdot x_i + \varepsilon_i \qquad i=1,2,\ldots,n$$

en que los e_i son variables aleatorias independientes, con igual distribución, media 0 y varianza común σ^2. Se denominan *errores*.

Estimadores mínimo cuadráticos de los parámetros

Los estimadores de α, β y σ^2, son, respectivamente, *a*, *b* y $\hat{\sigma}^2$, en que

$$a = \bar{y} - b \cdot \bar{x}$$

$$b = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\sigma}^2 = \frac{S_{yy} - b \cdot S_{xy}}{n-2} = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right)$$

en que $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ e $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ son los promedios de las **x** y de las **y**, respectivamente.

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

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

$$S_{xy} = \sum_{i=1}^{n} x_i \cdot y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i \right) \cdot \left(\sum_{i=1}^{n} y_i \right)$$

Recta de regresión ajustada

o recta de regresión estimada

$$\hat{y} = a + bx$$

las observaciones ajustadas están dadas por

$$\hat{y}_i = a + bx_i \quad i = 1,2,\ldots,n$$

Residuos

Son las diferencias entre los valores observados y los valores ajustados de la variable independiente *y*

$$e_i = y_i - \hat{y}_i = y_i - a - b \cdot x_i \qquad i = 1,2,\ldots,n$$

Varianza estimada del error

$$\hat{\sigma}^2 = \sum_{i=1}^{n} e_i^2 = \frac{S_{yy} - b \cdot S_{xy}}{n-2} = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}} \right)$$

en que los e_i son las diferencias entre los valores observados y los valores ajustados de la variable independiente ***y***

$$e_i = y_i - \hat{y}_i = y_i - a - b \cdot x_i \qquad i = 1,2,\ldots,n$$

y son los residuos.

$\hat{\sigma}^2$ es un estimador insesgado de la varianza de los errores ε_i del modelo de regresión lineal simple.

Valores esperados, varianzas y covarianza de los estimadores de α y de β

$$E(a) = \alpha \qquad E(b) = \beta$$

ambos son estimadores insesgados.

$$Var(a) = \sigma^2 \cdot \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right] \qquad Var(b) = \frac{\sigma^2}{S_{xx}}$$

$$Cov(a,b) = -\sigma^2 \cdot \frac{\bar{x}}{S_{xx}}$$

Para estimar las varianzas y covarianza, se sustituye σ^2 por el estimador $\hat{\sigma}^2$.

Valor esperado del estimador de la varianza σ^2

$$E(\hat{\sigma}^2) = \sigma^2$$

también es un estimador insesgado.

Coeficiente de determinación

Es una medida de bondad de ajuste de la recta $\hat{y} = a + bx$ a los puntos (x_i, y_i), y corresponde al cuadrado del coeficiente de correlación entre los valores observados de la variable dependiente ***y*** y los valores ajustados. Es igual a

$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

Intervalos de confianza para estimar los parámetros α, β y σ^2

Para que sean válidos estos intervalos, debe cumplirse el supuesto adicional de que los errores ε_i tienen distribución normal, es decir

$$\varepsilon_i \sim N(o,\sigma^2),\ independientes \qquad para\ i=1,2,\ldots,n$$

Intervalo de coeficiente de confianza *100(1-α)* para α :

$$a \pm t_{1-\frac{\alpha}{2}}(n-2)\cdot\hat{\sigma}\cdot\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}$$

Intervalo de coeficiente de confianza 100(1-α) para β :

$$b \pm t_{1-\frac{\alpha}{2}}(n-2)\cdot\frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

Pruebas de hipótesis para los parámetros α , β y σ^2

Para efectuar estas pruebas, tambien se requiere el supuesto de normalidad precedente.

Prueba de nivel de significación α para α :

La hipótesis nula es $\alpha = \alpha_0$ ó $\alpha \leq \alpha_o$ ó $\alpha \geq \alpha$

El estadístico de prueba es

$$t = \frac{a-\alpha_o}{\hat{\sigma}\cdot\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}}}$$

su distribución dada la hipótesis nula es t de student con *n-2* grados de libertad.

Prueba de nivel de significación α para β :

La hipótesis nula es $\beta = \beta_0$ ó $\beta \leq \beta_o$ ó $\beta \geq \beta$

El estadístico de prueba es

$$t = \frac{b-\beta_o}{\hat{\sigma}/\sqrt{S_{xx}}}$$

su distribución dada la hipótesis nula es t de student con *n-2* grados de libertad.

Prueba de nivel de significación α para σ^2 :

La hipótesis nula es $\sigma^2 = \sigma^2_0$ ó $\sigma^2 \leq \sigma^2_0$ ó $\sigma^2 \geq \sigma^2_0$

El estadístico de prueba es

$$x = (n-2) \cdot \frac{\hat{\sigma}^2}{\sigma_o^2}$$

su distribución dada la hipótesis nula es ji-cuadrado con *n-2* grados de libertad.

Análisis de varianza

La prueba de hipótesis asociada al análisis de varianza en el caso de regresión lineal simple es

Ho: $\beta = 0$

H1: $\beta \neq 0$

Para efectuar esta prueba se requiere el supuesto adicional de que los errores tienen distribución normal.

La tabla de análisis de varianza es la siguiente:

Fuente de variación	Sumas de cuadrados. SC	Grados de libertad. GL	Cuadrados medios. CM	Cuociente F
Regresión	$S_{yy} - S_{xy}^2/S_{xx}$	1	SCReg	CMReg /CMErr
Error	S_{xy}^2/S_{xx}	n-2	SCErr/(n-2)	--
Total	S_{yy}	n-1	--	--

El estadístico de prueba es F = CMReg / CMErr

Tiene distribución F con 1 grado de libertad en el numerador y n-2 grados de libertad en el denominador.

Se rechaza la hipótesis nula si F es grande.

Predicción

Es la estimación del valor de la variable dependiente **Y** cuando la variable dependiente **X** toma un valor $\mathbf{x_o}$ cualquiera. El valor puntual de una predicción se obtiene reemplazando x_o en la ecuación de la recta estimada, es decir, es igual a

$$y_o = a + b \cdot x_o$$

Predicción de una observación individual mediante intervalos de confianza

Para construir un intervalo de confianza para una predicción es necesario que se verifique el supuesto de normalidad de los errores.
Un intervalo de confianza para la predicción del valor individual de una observación **Y** cuando la variable independiente **X** toma el valor $\mathbf{x_o}$ está dado por

$$(a + b \cdot x_o) \pm t_{1-\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{S_{xx}}}$$

en que $t_{1-\frac{\alpha}{2}}(n-2)$ es el cuantil de la distribución t de student que acumula una probabilidad $1-\frac{\alpha}{2}$; $\hat{\sigma}$ es la desviación estándar del error, estimada; $\bar{x}$ es el promedio de los valores observados de **X** ; S_{xx} es la suma de cuadrados centrados de las **X**.

Predicción de la respuesta media mediante intervalos de confianza

$$a + b \cdot x_o \pm t_{1-\frac{\alpha}{2}}(n-2) \cdot \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{(x_o - \bar{x})^2}{S_{xx}}}$$

Se puede ver que es similar al intervalo para un valor individual, pero es más angosto, debido a que la estimación de un promedio es más precisa que la de un valor individual.

Bandas de confianza

Si se considera x_o como una variable que recorre todo el dominio de la **X** , los extremos de los intervalos de confianza describen unas bandas con forma de hipérbolas, cuya parte más angosta está en **X** = $\bar{x}$, el promedio, y se ensanchan a medida que se alejan del centro de los valores observados de la variable **X**.