Regresión lineal Multiple

Se desea estudiar la relación de una variable respuesta con 2 o más variables regresoras.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$$

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$\forall i \quad i = 1, 2, 3 \dots, n.$$

y_i: Variable respuesta de observaciones i.

x_{ik}: observación i de la variable regresora k.

$\beta_0, \beta_1, \beta_2 \dots \beta_k$: Parámetros.

k: ⟶ regresoras (# de x)

p: k+1 ⟶ parámetros (# de β)

Interpretaciones:

β_0: es el valor medio de y cuando todas las variables x_k son iguales a cero. (Es, no contiene al 0 no es interpreta)

β_k ⟶ es el efecto parcial sobre la Media y por cada unidad de incremento de la variable x_k cuando las demás variables permanecen constantes. Predictora

Matriz de Varianzas y Covarianzas.

Está Matriz contiene la covarianza entre variables, la cual indica su grado de variación conjunta.

$$\sum(X) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{bmatrix}$$

Caracteristicas

- Es una Matriz simétrica $\sigma_{ij} = \sigma_{ji}$
- Matriz cuadrada
- De orden $n \times n$
- los elementos de la diagonal principal son las varianzas de las variables $\sigma_{i,i} = \sigma_i^2$
- Por Fuera de la diagonal están las covarianzas entre los pares de elementos del vector aleatorio.

Forma matricial del Modelo

$$\underset{y}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}} = \underset{X}{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1K} \\ 1 & x_{21} & \cdots & x_{2K} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nK} \end{bmatrix}} \underset{\beta}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}} + \underset{\varepsilon}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}$$

$n \times 1 \qquad n \times p \qquad p \times 1 \qquad n \times 1$

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

matriz de diseño

$$\varepsilon \sim N(\mathbf{0}_{n\times 1}, \sigma^2 \mathbf{I}_{n\times n})$$

Estimación por Mínimos cuadrados de los Parámetros del Modelo

Se buscan valores estimados de los parámetros tales que se minimice la suma de cuadrados del error:

$$S(\underline{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_K x_{iK})^2$$

$$\hat{\underline{\beta}} = (X^t X)^{-1} X^t \underline{Y}$$

Significancia de los Parámetros β_j

Hipotesis:

$H_0 : \beta_j = 0$ v $H_a : \beta_j \neq 0$

con $j = 0, 1, 2 \dots k$

Estadistico de Prueba:

$$t_0 = \frac{\hat{\beta}_j}{S_e(\hat{\beta}_j)} \sim \text{Bajo } H_0 \; t_{n-p}$$

Decisión:

Se rechaza H_0 si $|t_0| > t_{\frac{\alpha}{2}, n-p}$

Se rechaza H_0 si el valor p

$p = p(t_{\frac{\alpha}{2}, n-p} > |t_0|)$

es pequeño

Interpretación:

→ Para que β_0 sea interpretable debe de ser significativo y en el conjunto de observaciones debe estar $(x_1, x_2 \dots, x_k) = (0, 0 \dots, 0)$

→ Para interpretar los demas β_j solo es necesario que sean significativos.

interpretan como:

- Por un aumento unitario en x_j se espera que la variable respuesta (Y) aumente o disminuya en promedio β_j unidades, siempre que las demás predictoras permanezcan constantes.

ESTIMACIÓN de σ^2

Bajo los supuestos $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

el estimador insesgado para la varianza es:

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-p}$$

donde:

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

Análisis de la varianza ANOVA

Fuente de Variación	Suma de cuadrados	Grados de libertad	Cuadrados Medios	F_0
Regresión	SSR	k	MSR: $\frac{SSR}{k}$	$F_0 = \frac{MSR}{MSE}$
Error	SSE	n-p	MSE: $\frac{SSE}{n-p}$	$= \frac{SSR\,(n-p)}{SSE\,(k)}$
Total	SST	n-1		

Prueba de significancia de la regresión

Hipotesis

$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ vs H_a: Ameros o $\beta_j \neq 0$

$j = 1, 2, 3 \dots k$

Estadistico de Prueba:

$$F_0 = \frac{MSR}{MSE} \quad \text{Bajo } H_0 \quad F_{k,\,n-p}$$

Decisión:

Se rechaza H_0 si $F_0 > F_{\alpha,\,k,\,n-p}$

Se rechaza H_0 si el valor $p = p\,(F_{\alpha,\,k,\,n-p} > F_0)$ es pequeño

Coeficientes de determinación:

$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR+SSE}$	$R^2_{adj} = 1 - \frac{MSE}{MST} = \frac{1-SSE(n-1)}{SST(n-p)}$
Al ingresar la variable tiende a no decrecer, aún cuando existan dentro del grupo de variables, un subconjunto de ellas que no aportan significativamente	Disminuye cuando al modelo de regresión ingresan variables que no logran reducir el SSE. Penaliza al Modelo al ingreso de nuevas variables.

Intervalo de confianza para los Parámetros β_j:

$$\hat{B}_j \pm t_{\frac{\alpha}{2}, n-p} \cdot se(\hat{B}_j)$$

pruebas sobre subconjuntos de Coeficientes.

se quiere probar si un subconjunto de β es significativo simultaneamente:

Hipotesis:

H_0: El subconjunto de $\beta = 0$

vs

H_a: Al menos un β del grupo $\neq 0$

Planteamos 2 Modelos $\quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

MC (Modelo completo = $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$
MF

MR (Modelo Reducido) = Reducido Bajo H_0

Estadístico de Prueba:

$$F_0 = \frac{SSE(MR) - SSE(MF)}{r} \Big/ MSE(M.C) \sim F_{r, n-p}$$

Decisión:

Se rechaza H_0 si $f_0 > F_{\alpha, r, n-p}$

Se rechaza si el valor $p = P(F_{\alpha, r, n-p} > f_0)$ es pequeño.

HIPOTESIS LINEAL General.

Se quiere ver por ejemplo si el efecto de dos o más variables es igual, o si el efecto de una variables es c veces otra.

Se pueden probar varias hipotesis juntas.

Ejemplo: 5 variables regresoras:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i$$

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Probar las siguientes hipotesis:

$H_0: B_1 = B_2, B_3 = B_4, B_5 = 0$ vs $H_a: B_1 \neq B_2, B_3 \neq B_4$ o $B_5 \neq 0$

igualamos las ecuaciones a cero y tenemos:

$H_0: B_1 - B_2 = 0, B_3 - B_4 = 0, B_5 = 0$ vs $H_a: B_1 - B_2 \neq 0$ o $B_3 - B_4 \neq 0$ o $B_5 \neq 0$

Estamos probando 3 ecuaciones diferentes, entonces armamos la Matriz L de orden mxp en este caso 3X6

$$L = \begin{matrix} \\ \\ \end{matrix} \begin{matrix} B_0 & B_1 & B_2 & B_3 & B_4 & B_5 \end{matrix}$$
$$\begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Y

Planteamos el Modelo reducido bajo la hipotesis nula:

$$M.R: y_i = \beta_0 + \beta_1 x_{i1} + \beta_1 x_{i2} + \beta_3 x_{i3} + \beta_3 x_{i4} + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 (x_{i1} + x_{i2}) + \beta_3 (x_{i3} + x_{i4}) + \varepsilon_i$$
$$y_i = \beta_0 + \beta_1 z_{i1,2} + \beta_3 z_{i3,4} + \varepsilon_i$$

Estadístico de Prueba:

$$F_0 : \frac{SSE(M.R) - SSE(M.F)/m}{MSE(M.C)} \sim F_{m, n-p}$$

m = # de ecuaciones que se estan probando o # de Filas L.I de matriz L

Decisión:

Se rechaza H_0 si $F_0 > F_{\alpha, m, n-p}$

Se rechaza H_0 el valor $p = P(F_{\alpha, m, n-p} > F_0)$ es pequeño

Normalidad:

Hipotesis:

H_0 : $\varepsilon_i \sim$ Normal vs H_a: $\varepsilon_i \neq$ Normal

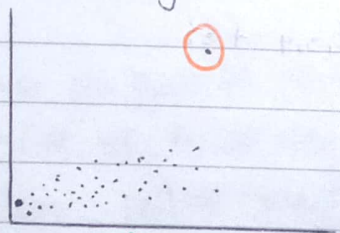
Prueba gráfica: se evalúa si los puntos en la escala normal se pueden ajustar por una línea recta.

Test de Shapiro-Wilk

Es una prueba específica de normalidad que arroja un valor p.

observaciones atípicas

Es un valor atípico en Y que está separado del resto de datos y podría afetar los resultados del ajuste del modelo



Residuales estandarizados d_i:

$$d_i = \frac{e_i}{\sqrt{MSE}}$$

si $|d_i| > 3$ indicios de atípica

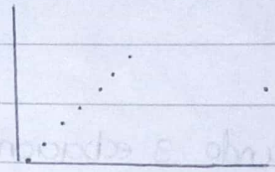
Residuales estudentizados r_i:

$$r_i = \frac{d_i}{\sqrt{1 - h_{ii}}}$$

si $|r_i| > 3$ i es una observación atípica.

Puntos de balanceo

un valor en las X alejado y puede controlar algunas propiedades del modelo como el R^2

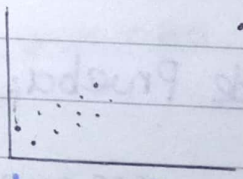


sea $H_{ij} = X_i^t (X^t X)^{-1} X_i$ es el elemento i de la Matriz hat (H) entonces:

si $h_{ii} > \frac{2p}{n}$ entonces i es un punto de balanceo.

Puntos Influyentes:

Es un valor inusual tanto en X como en Y. suelen halar el modelo en su dirección.



- Distancia de Cook:

si cooks $d_i > 1$ entonces i es un punto influyente.

- DFFITS:

si $|DFFITS_i| > \sqrt{\frac{2p}{n}}$ entonces i es un punto influyente.

Si $h_{oo} < h_{max}$ es una interpolación

Si $h_{oo} > h_{max}$ es una extrapolación

h_{max} es el valor máximo de h_{ii}
que no sea un punto de balanceo.

Intervalo de Confianza Para la respuesta Media.

$$\hat{Y}_o \pm t_{\frac{\alpha}{2}, n-p} \cdot Se(\hat{Y}_o)$$

Intervalo de Predicción Para Futuras observaciones

$$\hat{Y}_o \pm t_{\frac{\alpha}{2}, n-p} \cdot \sqrt{MSE + Se(\hat{Y}_o)^2}$$

Coeficientes de regresión estandarizados.

si las X no están en la misma escala de Medición no podemos determinar cual tiene Mayor o Menor efecto, para esto los estandarizamos asi:

$$y_i^* = \frac{Y_i - \bar{Y}}{\sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}}} \qquad x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{\sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - x_j)^2}{n-1}}}$$

$j = 1, 2, 3 \ldots k \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

modelo sin intercepto: $\beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \beta_3 x_{i3}^* + \ldots + \beta_k x_{ik}^* + \varepsilon_i$