REGRESION LINEAL MULTIPLE

El modelo de regresión lineal múltiple

El modelo de regresión lineal múltiple con p variables predictoras y basado en n observaciones tomadas es de la forma:

$$y_i = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots\ldots \beta_p x_{pi} + e_i$$

para i=1,2,… .n. Escribiendo el modelo para cada una de las observaciones, éste puede ser considerado como un sistema de ecuaciones lineales de la forma:

$$\begin{gathered}
y_1 = \beta_o + \beta_1 x_{11} + \beta_2 x_{21} + \ldots\ldots \beta_p x_{p1} + e_1 \\
y_2 = \beta_o + \beta_1 x_{12} + \beta_2 x_{22} + \ldots\ldots \beta_p x_{p2} + e_2 \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
y_n = \beta_o + \beta_1 x_{1n} + \beta_2 x_{2n} + \ldots\ldots \beta_p x_{pn} + e_n
\end{gathered}$$

que puede ser escrita en forma matricial como:

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & . & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & . & x_{2p} \\ . & . & . & . & . & . \\ .. & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

O sea,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

donde **Y** es un vector columna n dimensional**, X** es una matriz n x p’, con p’=p+1, **b** es el vector de coeficientes de regresión a ser estimados, su dimensión es p’ y **e** es un vector columna aleatorio de dimensión n
Por ahora, las únicas suposiciones que se requieren son que E(**e**)=**0** y que la matriz de varianza y covarianzas de los errores está dada por Var(**e**)=σ^2**I_n**, donde **I_n** es la matriz identidad de orden n.

Vector de parámetros

$$\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

aqui **$(X'X)^{-1}$** representa la matriz inversa de **(X’X).** Notar que **X’X** es simétrica, pues su transpuesta da la misma matriz.

En la regresión lineal simple, p=1 y el modelo puede ser escrito en forma matricial como

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ & \\ & \\ & \\ 1 & x_{n1} \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

$$S_R^2 = \frac{SCE}{n-p-1}$$

$$S_R = \sqrt{S_R^2}$$

$$VAR\left(\hat{\beta}_i\right) = S_R^2\left(q_{jj}\right)$$

$$q_{jj} = diag\left[\left(X^{'}X\right)^{-1}\right]$$

Intervalos de confianza para los β_i

$$\left[\hat{\beta}_i \mp t_{(1-\alpha/2)(n-p-1)}\sqrt{VAR\left(\hat{\beta}_i\right)}\right]$$

$$SCT = \sum\left(y_i - \overline{y}\right)^2$$
$$SCR = \sum\left(\hat{y}_i - \overline{y}\right)^2$$
$$SCE = \sum\left(y_i - \hat{y}_i\right)^2$$

Coeficiente de determinación

$$R^2 = \frac{SCR}{SCT}$$

ANOVA para el contraste sobre bondad de ajuste lineal

$$CMT = \frac{SCT}{n-1}$$
$$CMR = \frac{SCR}{P}$$
$$CME = \frac{SCE}{n-p-1}$$

$$F = \frac{CMR}{CME} \sim F_{p,n-p-1}$$

Intervalos de confianza a nivel $(1-\alpha)$ para la media y para la predicción en $X = X_0$

$$\hat{y}_0 \mp t_{(1-\alpha/2, n-p-1)} \sqrt{S_R^2 X_0^{'} \left(X^{'} X \right)^{-1} X_0}$$

$$\hat{y}_0 \mp t_{(1-\alpha/2, n-p-1)} \sqrt{S_R^2 \left(1 + X_0^{'} \left(X^{'} X \right)^{-1} X_0\right)}$$

Ejemplo:

Se quiere ajustar un modelo que permita estimar los gastos de alimentación de una familia (Y) con base a la información que proporcionan las variables regresoras X_1 = Ingresos mensuales y X_2 = Número de miembros de la familia. Para ello se recoge una muestra aleatoria simple de 15 familias cuyos resultados son los de la tabla adjunta. (El gasto e ingreso está dado en cientos de miles de pesos).

GASTO	INGRESO	TAMAÑO
0,43	2,1	3
0,31	1,1	4
0,32	0,9	5
0,46	1,6	4
1,25	6,2	4
0,44	2,3	3
0,52	1,8	6
0,29	1,0	5
1,29	8,9	3
0,35	2,4	2
0,35	1,2	4
0,78	4,7	3
0,43	3,5	2
0,47	2,9	3
0,38	1,4	4

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

$$Y = \begin{pmatrix} 0.43 \\ 0.31 \\ 0.32 \\ 0.46 \\ 1.25 \\ 0.44 \\ 0.52 \\ 0.29 \\ 1.29 \\ 0.35 \\ 0.35 \\ 0.78 \\ 0.43 \\ 0.47 \\ 0.38 \end{pmatrix} = X\vec{\beta} + \vec{\varepsilon} = \begin{pmatrix} 1 & 2.1 & 3 \\ 1 & 1.1 & 4 \\ 1 & 0.9 & 5 \\ 1 & 1.6 & 4 \\ 1 & 6.2 & 4 \\ 1 & 2.3 & 3 \\ 1 & 1.8 & 6 \\ 1 & 1.0 & 5 \\ 1 & 8.9 & 3 \\ 1 & 2.4 & 2 \\ 1 & 1.2 & 4 \\ 1 & 4.7 & 3 \\ 1 & 3.5 & 2 \\ 1 & 2.9 & 3 \\ 1 & 1.4 & 4 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \vec{\varepsilon}$$

$$\hat{\mathbf{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

$$X'X = \begin{pmatrix} 15 & 42 & 55 \\ 42 & 188.8 & 140.8 \\ 55 & 140.8 & 219 \end{pmatrix} \qquad X'Y = \begin{pmatrix} 8,07 \\ 32,063 \\ 28,960 \end{pmatrix}$$

$$\left(X'X\right)^{-1} = \begin{pmatrix} 1,359850298 & -0,092545445 & -0,282015378 \\ -0,092545445 & 0,016548704 & 0,012602475 \\ -0,282015378 & 0,012602475 & 0,067289577 \end{pmatrix}$$

$$\hat{\beta} = \left(X'X\right)^{-1} X'Y = \begin{pmatrix} -0.16 \\ 0.149 \\ 0.077 \end{pmatrix}$$

$$\hat{\beta}_0 = -0.16 \qquad \hat{\beta}_1 = 0.149 \qquad \hat{\beta}_2 = 0.077$$

PLANTEAMIENTO DEL MODELO:

$$\hat{Y}_i = -0.160 + 0.149 X_{i1} + 0.077 X_{i2}$$

RESIDUALES:

$$e_i = Y_i - \hat{Y}_i$$

$\hat{Y}_i$	e_i	e_i^2
0,38261429	0,05	0,00224541
0,31080246	0,00	6,4394E-07
0,35797225	-0,04	0,00144189
0,38516597	0,07	0,00560013
1,06931028	0,18	0,03264878
0,41235969	0,03	0,00076399
0,56874176	-0,05	0,00237576
0,37284495	-0,08	0,00686329
1,39395804	-0,10	0,01080727
0,3503172	0,00	1,0062E-07
0,32567516	0,02	0,0005917
0,76930455	0,01	0,00011439
0,51391693	-0,08	0,00704205
0,50159591	-0,03	0,0009983
0,35542057	0,02	0,00060415
SUMA	**0,00**	**0,07209785**

$$S_R^2 = CME = \frac{SCE}{n-p-1} \frac{0,072097848}{15\text{-}2\text{-}1} = 0,006008154$$

$$S_R = \sqrt{0,006008154} = 0,077512283$$

$$q_{jj} = diag\left[\left(X^{'}X\right)^{-1}\right]$$

$$q_{11} = 1,359850298$$
$$q_{22} = 0,016548704$$
$$q_{33} = 0,067289577$$

$$VAR\left(\hat{\beta}_i\right) = S_R^2\left(q_{jj}\right)$$

$$VAR\left(\hat{\beta}_0\right) = 0,006008154 \times 1,359850298 = 0,00817019$$

$$VAR\left(\hat{\beta}_1\right) = 0,006008154 \times 0,016548704 = 9,94272\text{E-}05$$

$$VAR\left(\hat{\beta}_1\right) = 0,006008154 \times 0,067289577 = 0,000404286$$

Intervalos de confianza para los β_i

$$\left[\hat{\beta}_i \mp t_{(1-\alpha/2)(n-p-1)} \sqrt{VAR\left(\hat{\beta}_i\right)}\right]$$

Para β_0

$$\left[\hat{\beta}_0 \mp t_{(1-\alpha/2)(n-p-1)} \sqrt{VAR\left(\hat{\beta}_0\right)}\right]$$

$$t_{(12\ ,\ 0.05)} = 1.78$$

$$\left[-0.16 \mp 1.78\sqrt{0,00817019}\right] = \left[-0.16 \mp 0,160892604\right]$$

$$L_i = -0.160 - 0,160892604 = -0,321350647$$
$$L_s = -0.160 + 0,160892604 = 0,000434561$$

$$P\left[-0,321350647 < \beta_0 < 0,000434561\right] = 90\%$$

De la misma forma se realizan los intervalos para β_1 y β_2

$\hat{Y}_i$	e_i	e_i^2	$(Y_i - \bar{Y})^2$	$(\hat{Y}_i - \bar{Y})^2$
0,38261429	0,047385712	0,00224541	0,011664	0,02414472
0,31080246	-0,000802459	6,4394E-07	0,051984	0,05161872
0,35797225	-0,037972249	0,00144189	0,047524	0,03240999
0,38516597	0,074834029	0,00560013	0,006084	0,02335824
1,06931028	0,180689725	0,03264878	0,506944	0,28229061
0,41235969	0,027640308	0,00076399	0,009604	0,01578549
0,56874176	-0,048741764	0,00237576	0,000324	0,00094506
0,37284495	-0,082844951	0,00686329	0,061504	0,02727619
1,39395804	-0,103958043	0,01080727	0,565504	0,73266417
0,3503172	-0,0003172	1,0062E-07	0,035344	0,03522483
0,32567516	0,024324838	0,0005917	0,035344	0,04508184
0,76930455	0,010695453	0,00011439	0,058564	0,05350179
0,51391693	-0,083916926	0,00704205	0,011664	0,00057999
0,50159591	-0,031595906	0,0009983	0,004624	0,00132526
0,35542057	0,024579434	0,00060415	0,024964	0,03333525
SUMA	**1,31006E-14**	**0,07209785**	**1,43164**	**1,35954215**

$$SCT = \sum\left(y_i - \bar{y}\right)^2 = 1,43164$$
$$SCR = \sum\left(\hat{y}_i - \bar{y}\right)^2 = 1,35954215$$
$$SCE = \sum\left(y_i - \hat{y}_i\right)^2 = 0,07209785$$

$$R^2 = \frac{SCR}{SCT} = \frac{1,35954215}{1,43164} = 0,94963968$$

$$CMT = \frac{SCT}{n-1} = \frac{1,43164}{14} = 0,10226$$

$$CMR = \frac{SCR}{P} = \frac{1,359542152}{2} = 0,679771076$$

$$CME = \frac{SCE}{n-p-1} = \frac{0,072097848}{12} = 0,006008154$$

$$F_c = \frac{CMR}{CME} = \frac{0,679771076}{0,006008154} = 113,1414203$$

Pruebas de hipótesis individuales para los β_i

Para β_0 **:**

1. $H_0 : \beta_0 = 0$

2. $H_a : \beta_0 \neq 0$

3. $\alpha = 0.05$

4. a) $E.P \Rightarrow t_c = \dfrac{\hat{\beta}_0 - \beta_0}{S_{\beta_0}}$

4. b)

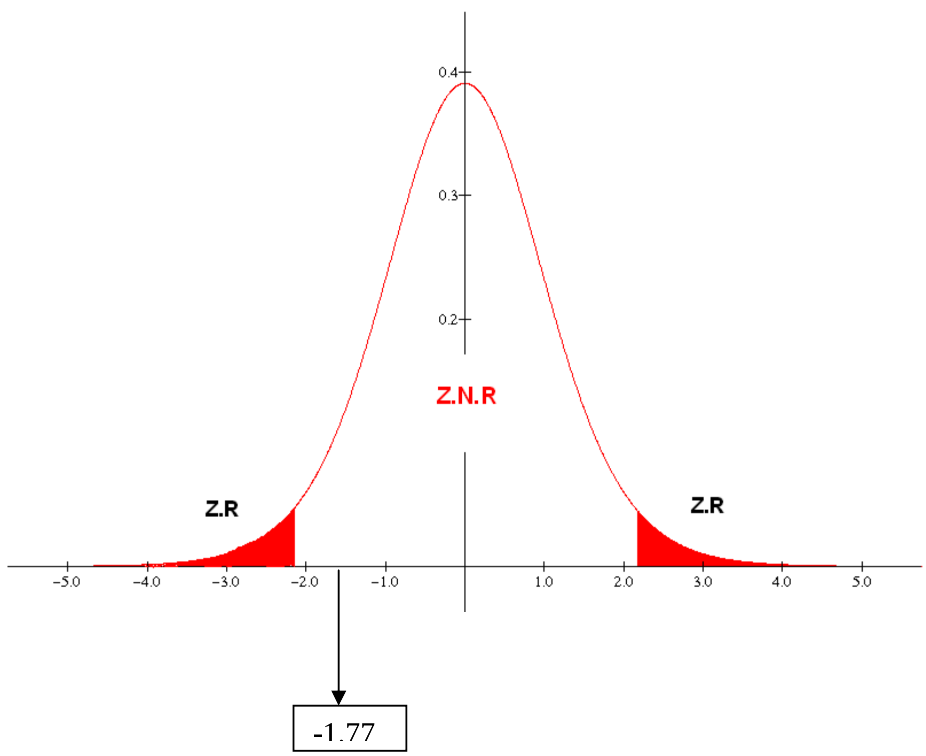
$$R.C. \Rightarrow \text{se rechaza } \mathrm{H}_0 \text{ si } \mathrm{t_c} < -t_{(1-\alpha/2)(n-p-1)} \text{ o } \mathrm{t_c} > t_{(1-\alpha/2)(n-p-1)}$$

$$R.C. \Rightarrow \text{se rechaza } \mathrm{H}_0 \text{ si } \mathrm{t_c} < -2.18 \text{ o } \mathrm{t_c} > 2.18$$

5. a)

$$t_c = \frac{\hat{\beta}_0 - \beta_0}{S_{\beta_0}} = \frac{-0.16 - 0}{0.0903} = -1.77$$

5. b)



6. a) Decisión: No se rechaza H_0.

6. b) Conclusión: $\beta_0 = 0$

De la misma forma se realizan las pruebas de hipótesis para β_1 y β_2

Prueba de hipótesis para la conveniencia del modelo lineal.

$$H_0 : \beta_1 = \beta_2 = 0$$

Tabla de análisis de varianza (ANOVA):

FUENTE	SUMA DE CUADRADOS	GL	CUADRADOS MEDIOS	F CALC.
Regresión	SCR	P	CMR	F_C
Error	SCE	n-P-1	CME	
Total	SCT	n-1		

FUENTE	SUMA DE CUADRADOS	GL	CUADRADOS MEDIOS	F CALC.
Regresión	**1,35954215**	2	0,67977108	113,14142
Error	**0,07209785**	12	0,006008154	
Total	**1,43164**	14		

$$F_{p,n-p-1} = F_{2,\ 12} = 5.10 \text{ para } \alpha = 0.05$$

Como $F_C > F_T$ se rechaza H_0

Intervalos de predicción para la respuesta media y para una respuesta individual cuando: $X_1 = 6$ y $X_2 = 5$

$$\hat{Y}_i = -0.160 + 0.149 X_{i1} + 0.077 X_{i2}$$

$$\hat{Y}_0 = -0.160 + (0.149 \times 6) + (0.077 \times 5) = 1.12$$

$$X_0 = \begin{pmatrix} 1 \\ 6 \\ 5 \end{pmatrix} \qquad X_0^{'} = \begin{pmatrix} 1 & 6 & 5 \end{pmatrix}$$

$$X_0^{'} \left(X\,' X \right)^{-1} X_0 = 0{,}463292428$$

$$\hat{y}_0 \mp t_{(1-\alpha/2, n-p-1)} \sqrt{S_R^2 X_0^{'} \left(X^{'} X \right)^{-1} X_0}$$

$$\hat{y}_0 \mp t_{(1-\alpha/2, n-p-1)} \sqrt{S_R^2 \left(1 + X_0^{'} \left(X^{'} X \right)^{-1} X_0 \right)}$$

$$1.12 \mp 1.78 \sqrt{(0.006008154)(0{,}463292428)}$$

$$1.12 \mp 1.78 \sqrt{(0.006008154)\left(1 + 0{,}463292428\right)}$$

Para la respuesta media:

$$L_i = 1.12 - 0.0939 = 1.03$$

$$L_s = 1.12 + 0.0939 = 1.21$$

$$P\left[1.03 < \overline{Y}_0 < 1.21 \right] = 90\%$$

Para la respuesta individual:

$$L_i = 1.12 - 0.00879 = 1.11$$

$$L_s = 1.12 + 0.00879 = 1.13$$

$$P[1.11 < Y_0 < 1.13] = 90\%$$

- **VERIFIQUE LOS RESULTADOS MATEMATICAMENTE Y MEDIANTE UN PAQUETE ESTADISTICO**
- **INTERPRETE CADA UNO DE LOS RESULTADOS**
- **REALICE EL ANALISIS DE RESIDUOS**