

### Auxiliar Regresión Lineal

Se busca describir un conjunto de variables  $X_1, \dots, X_p$  llamadas variables explicativas o exógenas que influyen sobre otra variable llamada variable a explicar o endógena ( $y$ )., mediante una relación lineal del tipo

$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 \dots + \beta_p x_i^p + \epsilon_i .$$

El modelo lineal obtenido de una muestra de tamaño  $n$ , normalmente no es exacto, por lo que debe considerarse el error  $\epsilon_i$  asociado al modelo para la observación  $i$ . Por otro lado, se busca minimizar los errores con el criterio de los mínimos cuadrados  $\min \sum \epsilon_i^2 = \epsilon^t \epsilon$

Matricialmente tenemos que  $y = X\beta + \epsilon$  y con el criterio anterior nos queda que los coeficientes que minimizan el error al cuadrado son de la forma

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

Para el caso de un modelo simple  $y = \beta_0 + \beta_1 x$ , la estimación de los coeficientes viene

dada por: 
$$\hat{\beta}_1 = \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum x_i^2 - n \bar{x}^2} \quad \text{y} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

#### Propiedades de $\hat{\beta}$

- El estimador  $\hat{\beta}$  es insesgado
- El estimador  $\hat{\beta}$  es consistente
- El estimador  $\hat{\beta}$  tiene mínima varianza
- La estimación de  $\hat{\sigma}^2$  obtenida por el método de máxima verosimilitud es sesgada.

- $\hat{\sigma}^2 = \frac{\sum \epsilon_i^2}{n - r}$  con  $r$  = rango de la matriz  $X$  = nro de variables  $x + 1$
- $Var(\hat{\beta}_j) = \hat{\sigma}^2 (X^t X)^{-1}_{jj}$

#### Calidad del modelo

Los residuos  $\epsilon_i$  dan la calidad del ajuste para cada observación. Un índice que evita el problema de que  $\epsilon_i$  dependa de cada observación es

$$\frac{\sum \varepsilon_i^2}{\sigma^2} \rightarrow \chi_{n-r}^2 \Rightarrow \frac{n-r}{\sigma^2} \hat{\sigma}^2 \rightarrow \chi_{n-r}^2$$

### Coeficiente de correlación múltiple

Compara la varianza explicada con la varianza total.

$R = \sqrt{R^2}$  = coeficiente de correlación lineal entre el verdadero valor con lo que estamos estimando.

$$R = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

- Si  $R=0$ , el modelo es la media muestral de los valores de  $y$ .
- Si  $R$  cercano a 1, el modelo es bueno siendo los valores observados cercanos a los estimados.
- Si  $R=1$ , existe un modelo lineal que permite escribir las observaciones  $y_i$  como combinación lineal de las variables explicativas.

### Test de hipótesis

Por un lado vimos como calcular los coeficientes del modelo y que tan cercano a una recta es con el coeficiente de correlación  $R$ . Sin embargo, para decidir si una variable aporta o no al modelo de manera significativa estadísticamente, es decir, si el coeficiente  $B$  es distinto de 0, debemos hacer un test de hipótesis global y luego general.

### Test global

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

Para decidir si aceptamos o rechazamos  $H_0$ , se construye el estadístico  $F$  que sigue una distribución de Fisher.

$$F = \frac{R^2 / (r-1)}{(1-R^2) / (n-r)} \rightarrow F_{(r-1)(n-r)}$$

Buscamos  $C$  tal que  $P(F_{(r-1)(n-r)} > C) = \alpha$ , Si  $F > C \Rightarrow$  rechazamos  $H_0$  ( lo que indica que existe al menos una variable que aporta al modelo)

### Test local

Con el test global probamos si existe al menos una variable que debería ir en el modelo, pero no sabemos cual. Por tanto para cada coeficiente debemos hacer el siguiente test

$$H_0 : \beta_j = 0$$

Pero  $\hat{\beta}_j \rightarrow N(\beta_j, \sigma^2 (x^t x)^{-1}_{jj}) \Rightarrow \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\beta_j}} \rightarrow t_{n-r}$

Buscamos  $P(|t_{n-r}| > Q) \leq \alpha$ ,

Si  $t_{n-r} > Q$  para  $\alpha = 0.05$  o si  $P(|t_{n-r}| > Q) \leq 0.05$  rechazamos  $H_0$  ( lo que indica que existe al menos una variable que aporta al modelo)

## Ejercicios

### Problema 1

El ministerio de educación quiere estudiar de qué depende el gasto anual en educación de un hogar, para ello, recolecta información en 100 hogares y plantea el modelo lineal:

$$E(y) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 \quad (1)$$

Donde  $x_1$  es el ingreso del hogar (en miles de pesos),  $x_2$  el número de hijos,  $x_3$  la talla del jefe de hogar y  $x_4$  el número de perros en la casa.

2.1 Complete los resultados de la regresión lineal (1) dados en las tablas n° 4 y 5.

2.2 Interprete los resultados.

2.3 Se plantea un modelo con el ingreso y el n° de hijos solamente:

$$E(y) = b_0 + b_1 x_1 + b_2 x_2 \quad (2)$$

Se propone resolver el test: de hipótesis  $H_0 : E(y) = b_0 + b_1 x_1 + b_2 x_2$  contra

$$H_1 : E(y) = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4.$$

Para esto, se resuelve el modelo (2) obteniéndose el conjunto de resultados presentados en las tablas 6 a 7.

Comente el cambio en la suma de los cuadrados de los residuos  $SSR$  del modelo (1) al modelo (2) y calcule la variación porcentual.

Se propone como estadístico para medir la significación del cambio en la suma residual a:

$$\frac{(SSE_2 - SSE_1)/(k_2 - k_1)}{SSE_1/(n - k_1)}$$

donde  $SSE_2$  y  $SSE_1$  corresponden a la suma de los cuadrados de los residuos de los modelos (2) y (1) respectivamente, y donde  $k_1$  y  $k_2$  son la cantidad de coeficientes de cada modelo. Encuentre la distribución que sigue este estadístico y concluye con un error de tipo I de 5% si las variables *n° de perros* y *talla del jefe* son significativas en el modelo (1) utilizando los resultados de las tablas 4 a 7.

2.4 Dé intervalos de confianza de nivel 95% para los tres parámetros del modelo (2).

2.5 Se tiene un nuevo hogar con un ingreso de 400 y 3 hijos. Dé una estimación de su gasto en educación.

Table n°4

Variable	Estimación	Desviación típica	t-Student	P-Valor
----------	------------	-------------------	-----------	---------

Constante	20.387	20.384	1.000	0.319
Ingreso	0.189		9.242	0.000
N° hijos	17.379	2.978	5.836	0.000
Talla jefe	8.869	6.176		0.154
N° perros		0.107	1.749	0.083

**Coefficiente de correlación múltiple  $R=0.785$**

**Estimación insesgada de la varianza del error  $\hat{\sigma} = 29.12$**

Tabla n°5

Fuente	Grados libertad	Suma cuadrados	F	p-valor
Regresión		129489.083	38.185	0.0000
Residuos	95	80539.635		
Total	99			

Table n°6

Variable	Estimación	Desviación típica	t-Student	P-Valor
Constante	54.03477	8.575475	6.301	0.000
Ingreso	0.197514	0.019715	10.019	0.000
N° hijos	17.804395	2.969696	5.995	0.000

**Coefficiente de correlación múltiple  $R=0.772$**

**Estimación insesgada de la varianza del error  $\hat{\sigma} = 29.56$**

Tabla n°7

Fuente	Grados libertad	Suma cuadrados	F	p-valor
Regresión	2	125292.851	71.713473	0.0000
Residuos	97	84735.8665		
Total	99	210028.718		

### **Solución:**

Normalmente los programas estadísticos entregan dos tablas: una tabla con los resultados del test de Fisher (5 y 7) y una tabla con los resultados del test local t-student (4 y 6). Deben tener claro que representa cada valor se esas tablas para poder completarla.

La tabla del **test t-student** entrega lo siguiente:

- Estimación: Corresponde a la estimación de los coeficientes  $\beta_i = \hat{\beta}_i$
- Desviación Típica: Corresponde a la desviación standard estimada de los coeficientes , es decir  $\hat{\sigma}_{\beta_j}$
- T-student: Valor asociado al estadístico  $t_i = \frac{\hat{\beta}_i}{\hat{\sigma}_{\beta_i}}$  (bajo hipótesis  $H_0 : \beta_i = 0$ )
- Pvalor :  $P(|t_{n-r}| > t)$  con t descrito anteriormente.

## Tabla de Fisher

- Suma cuadrados Regresión =  $\sum_i (y_i - \bar{y})^2 = SSR$
- Suma cuadrados Residuos =  $\sum_i \varepsilon_i^2 = SSE$
- Suma cuadrados Total =  $SSR + SSE$
- Grados libertad Regresión =  $r-1$  (numero de variables x)
- Grados libertad residuos =  $n-r$
- Grados de libertad totales =  $n-1$
- Suma cuadrados medio = Suma cuadrados/ gradoslibertad
- $F = \frac{R^2 / (r-1)}{(1-R^2) / (n-r)} \rightarrow F_{(r-1)(n-r)}$
- Pvalor =  $P(F_{(r-1)(n-r)} > F)$

Con las fórmulas anteriores calculamos lo que falta

- $\hat{\beta}_4 = t_4 * \hat{\sigma}_{\beta_4} = 1.749 * 0.107 = 0.188$
- $\hat{\sigma}_{\beta_2} = \hat{\beta}_2 / t_2 = 0.189 / 9.242 = 0.0204$
- $t_3 = \hat{\beta}_3 / \hat{\sigma}_{\beta_3} = 8.869 / 6.176 = 1.436$
- Grados libertad regresión =  $r-1 = 5-1=4$
- Suma Cuadrados Totales =  $SSR + SSE = 129489.083 + 80539.635 = 210028.718$

Las tablas completas son:

Table nº3

Variable	Estimación	esviación típica	t-Student	P-Valor
Constante	20.387	20.384	1.000	0.319
X <sub>1</sub>	0.189	0.020	9.242	0.000
X <sub>2</sub>	17.379	2.978	5.836	0.000
X <sub>3</sub>	8.869	6.176	1.436	0.154
X <sub>4</sub>	0.188	0.107	1.749	0.083

Tabla nº4

Fuente	Grados libertad	Suma cuadrados	Cuadrados medio	F	p-valor
Regresión	4	129489.083	32372.271	38.185	0.0000
Residuos	95	80539.635	847.786		
Total	99	210028.718			

2.2 Para interpretar los resultados del modelo, deben fijarse en 3 cosas: El pvalor del Fisher, el pvalor de la Tstudent y el coeficiente de correlación múltiple R .

El p-valor de la Fisher es menor a 0.05, por lo tanto, se rechaza que todos los coeficientes beta son nulos y existe al menos una variable x que debería ir en el modelo y que permite explicar la variable y. Entonces analizamos cada variable para ver si va o no en el modelo. De la tabla-. Tstudent tenemos que las variables x3 y x4 tienen pvalor mayor a 0.05, por lo

tanto no deberían ir en el modelo, es decir, no son estadísticamente significativas. Por otro lado, el R es bastante alto (0,785) lo que indica que el modelo estimado es bastante bueno.

2.3) Del punto anterior, vemos que las variables x3 y x4 no deberían ir en el modelo, por lo que se construye un nuevo modelo que solo depende de x1 y x2 cuyo resultado se encuentra en las tablas 6 y 7. Del pvalor del Fisher y de las t-student de esas tablas vemos que el nuevo modelo sigue siendo estadísticamente significativo. El R disminuyo un poco, pero sigue siendo bueno.

Para ver si el nuevo modelo es estadísticamente significativo se construye el siguiente test de hipótesis:

$$\begin{aligned} H_0 : E(y) &= b_0 + b_1 x_1 + b_2 x_2 && \text{contra} \\ H_1 : E(y) &= b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 \end{aligned}$$

Como regla de decisión se utiliza el siguiente estadístico  $F = \frac{(SSE_2 - SSE_1)/(k_2 - k_1)}{SSE_1/(n - k_1)}$

Con donde  $SSE_2$  y  $SSE_1$  corresponden a la suma de los cuadrados de los residuos de los modelos (2) y (1) respectivamente, y donde  $k_1$  y  $k_2$  son la cantidad de coeficientes de cada modelo.

Por otro lado ,  $SSE_2 = \sum (\hat{\varepsilon}_i^{H_0})^2 = 97 * \hat{\sigma}^2 = 97 * 29.56^2 = 84758$ ; en la tabla es 84735.8665 exactamente y  $SSE_1 = \sum (\hat{\varepsilon}_i^{H_1})^2 = 95 * \hat{\sigma}^2 = 95 * 29.12^2 = 80558$  en la tabla es exactamente 80539.635

Entonces  $F = (84735 - 80539) * 95 / 2 * 80539 = 2.4748$

Luego el estadístico es el p-valor del test es:  $Pr(F_{2,95} > 2.4748) \approx 0.09$ , entonces No se rechaza la hipótesis nula y El modelo (2) es más significativo que el modelo (1).

2.4) Para encontrar el intervalo de confianza para los coeficientes, calculamos a y b tal que

$$\begin{aligned} P(a < \beta_j < b) &= 1 - \alpha \\ \text{entonces } P\left(\frac{\hat{\beta}_j - b}{\hat{\sigma}_{\beta_j}} < \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\beta_j}} < \frac{\hat{\beta}_j - a}{\hat{\sigma}_{\beta_j}}\right) &= 1 - \alpha \\ \text{y } P(t_1 < t_{n-r} < t_2) &= 1 - \alpha \rightarrow P(|t_{n-r}| > t_{\alpha/2}) = \alpha / 2 \end{aligned}$$

$$\text{despejando } a = \hat{\beta}_j - t_{\alpha/2} * \hat{\sigma}_{\beta_j} \quad \text{y} \quad b = \hat{\beta}_j + t_{\alpha/2} * \hat{\sigma}_{\beta_j}$$

Para  $n-r = 95$  grados de libertad y  $\alpha = 0.05$  ,  $t_{n-r, \alpha/2} = 1.96$

Los intervalos de confianza para cada coeficiente son:

Variable	Estimación	Desviación típica	t-Student	Intervalo
Constante	54.03477	8.575475	6.301	[37.23, 70.84]
Ingreso	0.197514	0.019715	10.019	[0.159, 0.236]
N° hijos	17.804395	2.969696	5.995	[11.98, 23.62]

- 2.5) La estimación del gasto en educación de un hogar con un ingreso de 400 y 3 hijos es  
 $y = \beta_0 + \beta_1 * 400 + \beta_2 * 3 = 54.034 + 0.1975 * 400 + 17.8043 * 3 = 186.45$

## **Problema 2**

Consideramos un modelo de regresión lineal simple :  $y = \beta_0 + \beta_1 x + \varepsilon$

2.1) Dé la expresión de los estimadores  $\hat{\beta}_0$  y  $\hat{\beta}_1$  de los mínimos cuadrados para  $\beta_0$  y  $\beta_1$

2.2) Se hace el cambio de variable  $z=10x$ . Obtenga los estimadores del nuevo modelo.

$\hat{y} = \gamma_0 + \gamma_1 z + \varepsilon$  en función de  $\hat{\beta}_0$  y  $\hat{\beta}_1$ . Compare ambos modelos.

2.3) Dé el estadístico del test  $H_0 : \beta_1 = 0$  y explique que pasa cuando se rechaza.

### **Solución:**

2.1) Los estimadores de mínimos cuadrados de un modelo de regresión lineal simples son:

$$\hat{\beta}_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \quad y \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

2.2) Los coeficientes del nuevo modelo son

$$\hat{\gamma}_1 = \frac{\sum z_i y_i - n\bar{z}\bar{y}}{\sum z_i^2 - n\bar{z}^2} = \frac{\sum 10x_i y_i - n10\bar{x}\bar{y}}{\sum 100x_i^2 - n100\bar{x}^2} = \frac{10(\sum x_i y_i - n\bar{x}\bar{y})}{100(\sum x_i^2 - n\bar{x}^2)} = \frac{\hat{\beta}_1}{10}$$

$$\hat{\gamma}_0 = \bar{y} - \hat{\gamma}_1 \bar{z} = \bar{y} - \frac{\hat{\beta}_1}{10} * 10\bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} = \hat{\beta}_0.$$

Comparando los modelos

$\hat{y} = \hat{\gamma}_0 + \hat{\gamma}_1 z = \hat{\beta}_0 + \frac{\hat{\beta}_1}{10} * 10x = \hat{\beta}_0 + \hat{\beta}_1 x = y$ , es decir, ambos modelos tiene la misma predicción.

2.3) El estadístico del test  $H_0 : \beta_1 = 0$  es  $\frac{\hat{\beta}_1}{\hat{\sigma}_1}$  que sigue una t- student a n-2 grados de

libertad bajo la hipótesis nula. Rechazar la hipótesis nula indica que hay una cierta influencia de la variable x sobre la variable y. Se podría usar después de estimar el modelo para hacer predicciones.