4.4 Intervalos de confianza y pruebas de hipótesis para los parámetros del modelo

A continuación nos planteamos la siguiente pregunta: la variable x realmente sirve para predecir los valores de y?

Supongamos un caso extremo en que $\beta = 0$. En esta situación

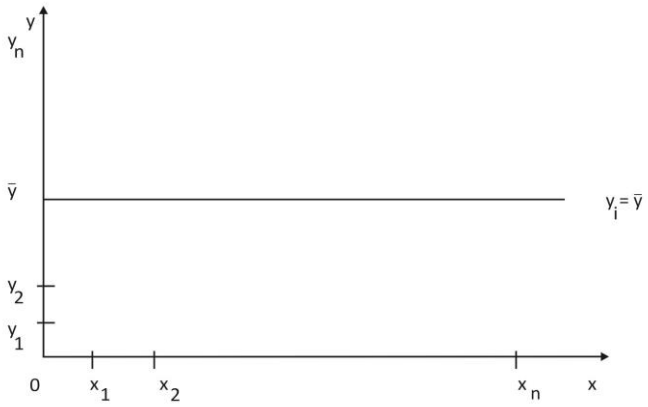
$$y_i = \alpha + 0\, x_i = \alpha$$

Pero ya hemos determinado que el estimador de α es a, siendo $a = y - b\, x$

Entonces, si $b = 0$, $a = y - 0\, x = y$

En consecuencia $y = y$ cualquiera sea el valor que toma la variable x. En este caso es fácil ver que x no sirve para explicar y_i.

Gráficamente, tenemos:



Pero, a medida que el valor de β se va alejando de 0 (o su estimador b se aleja de 0) los valores x se hacen más significativos en la predicción de y .

Para tomar una decisión objetiva en cuanto a la verdadera significaciónde β, es lógico pensar que tenemos que aplicar un procedimiento de test de hipótesis, tema que desarrollaremos a continuación.

En primer lugar desarrollaremos un test de hipótesis acerca del parámetro β (pendiente de la recta de regresión poblacional).

La hipótesis nula que interesa probar es la siguiente:

H_0) $\beta = 0$
H_1) $\beta \neq 0$

Si se rechaza la hipótesis nula estaremos diciendo que la pendiente de la recta que establece la relación entre las variables y y x es distinta de 0 y, en consecuencia, el conocimiento de x mejora la estimación de los valores de la variable y.

Es lógico pensar que el estimador puntual de β es b. Sin demostrar diremos que b es un estimador insesgado de β.

Entonces: $E(b) = \beta$
Debemos ahora calcular la varianza de b.

La varianza de b se calcula de la siguiente manera:

$$V(b) = S_e^2 = \cdot \frac{1}{\sum (x_i - \bar{x})^2}$$

y su correspondiente desviación estándar es:

$$S_b = S_e^2 \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}}$$

Estandarizando el valor de b, tenemos:

$$\frac{b - \beta}{S_b}$$

y, si se cumple el supuesto de distribución normal de e , este estadístico se distribuye como una variable t de Student con n - 2 grados de libertad.

El estadístico debe ser comparado con los siguientes valores críticos pues estamos utilizando un test bilateral:

$$y \quad t_1^* = t_{n-2;\, \alpha/2} \qquad t_2^* = t_{n-2;\, 1-\alpha/2}$$

La regla de decisión será, entonces:

$$\text{Si } t_1 < t^* \quad o \quad \text{si } t > t_2^* \quad \text{se rechaza } H_0$$

$$\text{Si } t_1^* < t < t_2^* \quad \text{no se rechaza } H_0$$

Aplicaremos el test propuesto en el ejemplo de las universidades y para ello debemos calcular S .

$$S_b = 0{,}1321\sqrt{\frac{1}{425317485}} = 0{,}1321 \cdot \frac{1}{20623{,}227} = 0{,}1321 \bullet 0{,}0000484 = 0{,}0000054$$

A continuación estandarizamos el valor de b.

$$t = \frac{0{,}0001216}{0{,}0000064} = 19$$

Buscamos luego los valores críticos:

$$t_1^* = t_{18;\,0{,}025} = -2{,}101$$
$$t_2^* = t_{18;\,0{,}975} = 2{,}101$$

Evidentemente, el valor de t calculado cae en la zona de rechazo de la hipótesis nula. Concluimos que β no es 0, la regresión es significativa. La variable x contribuye significativamente en la explicación de y.

Si se rechaza H , podemos estimar por intervalos el verdadero valor del parámetro β. El intervalo es:

$$P(b - t_{1-\alpha/2;\,n-2} \cdot S_b \leq \beta \leq b + t_{1-\alpha/2,\,n-2} \cdot S_b) = 1 - \alpha$$

Construimos, entonces, el intervalo de confianza.

$$P(0{,}0001216 - 2{,}101 \cdot 0{,}0000064 \leq \beta \leq 0{,}0001216 + 2{,}101 \cdot 0{,}0000064) = 0{,}95$$

$$P(0{,}0001216 - 0{,}0000134 \leq \beta \leq 0{,}0001216 + 0{,}0000134) = 0{,}95$$

$$P(0{,}000108 \leq \beta \leq 0{,}000135) = 0{,}95$$

4.4.1. Tabla de análisis de la varianza en regresión

Existe otra manera de probar la hipótesis nula que hemos establecido acerca del parámetro β.

Este test se efectúa construyendo una tabla de análisis de la varianza considerando las sumas de cuadrados que hemos calculado al hablar de las fuentes de variabilidad que intervienen en un análisis de regresión.

La tabla es la siguiente:

Fuentes de variación	Sumas de cuadrados	Grados de libertad	Cuadrados medios	Fobs.
Regresión	SCR	1	SCR/1	CMR/CMe
Error	SCe	n - 2	SCe/n - 2	
Total	**SCTotal**	**n - 1**		

El estadístico para probar la hipótesis nula referida al parámetro β es F = CMR/CMe.

Si la variabilidad explicada por la regresión es grande, tal situación estaría diciendo que una gran parte de la variabilidad ha sido explicada por la regresión de y sobre x. En este caso, la relación que liga a las variables en estudio es significativa y se llegaría al rechazo de H .

En consecuencia, β no es 0 por lo cual la pendiente de la recta es significativa.
El valor F se compara con un valor F crítico con 1 y n - 2 grados de libertad.

Construimos a continuación la tabla de análisis de la varianza con los datos del ejemplo que venimos desarrollando.

Fuentes de variación	Sumas de cuadrados	Grados de libertad	Cuadrados medios	Fobs.
Regresión	6,28699	1	6,28699	360,04713
Error	0,31431	18	0,01746	
Total	**6,60130**	**19**		

Si comparamos el valor de Fobs. con $F^{1;\ 38;\ 0,95} = 4,45$, debemos tomar la decisión de rechazar la hipótesis nula. En consecuencia β no es igual a 0 y la regresión es significativa. Se puede utilizar la relación lineal que existe entre la cantidad de alumnos y el costo por alumno para predecir otros costos no contemplados en la muestra.

4.4.2. Predicción

Uno de los objetivos más importantes del análisis de regresión consiste en obtener predicciones de la variable y para determinados valores de x.

El modelo de regresión sirve para estimar valores de y utilizando tanto la estimación puntual como por intervalos de confianza.

En primer lugar veremos cómo efectuamos una estimación puntual.

Dado un valor x , reemplazando a x por este valor en la ecuación de regresión podemos estimar el valor de y.

Por ejemplo, si quisiéramos conocer cuál será el costo estimado cuando una universidad tenga 15.000 alumnos, reemplazamos x por 15.000 y obtenemos el costo.

$$\hat{y}_i = - 0,0185 + 0,0001216\ (15.000) = - 0,0185 + 1,824 = 1,81$$

El costo estimado para una universidad que cuenta con 15.000 alumnos es de $ 1,81 mensual.

También podemos estimar y a través de una estimación por intervalos.

Estimación de Valores Medios

En este caso consideramos estimar un valor promedio de la variable dependiente y en función de un valor particular de x.

El intervalo para efectuar esta estimación es:

$$\hat{Y}_i \pm t_{n-2}\, S_{y/x} \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

t^{n-2} = valor de la distribución t de Student con n - 2 grados de libertad.

$S_{Y/X}$ = error estándar de estimación.

(x - x) : la amplitud del intervalo varía de acuerdo a cuan distante esta el valor de x de la media.

Por ejemplo, si queremos estimar el costo promedio estimado para una universidad que tiene 15.000 alumnos, se tiene:

$$1{,}81 \pm 0{,}1321 \times 2{,}101 \sqrt{\frac{1}{20} + \frac{(15000 - 6524.45)^2}{425317485}}$$

$$1{,}81 \pm 0{,}1321 \times 2{,}101 \sqrt{0{,}05 + 0{,}168}$$

$$1{,}81 \pm 0{,}278 \times 0{,}467$$

$$1{,}81 \pm 0{,}130$$

$$1{,}68 \quad 1{,}94$$

El costo promedio mensual para una universidad con 15.000 alumnos es un valor comprendido en el intervalo $ 1,68 y $ 1,94.

4.4.3. Predicción de Valores Individuales

Cuando se trata de estimar un valor individual de la variable dependiente, la estimación es menos precisa. En lugar de hablar de un intervalo de confianza se habla de un intervalo de predicción.

El intervalo de predicción para estimar un valor individual de y es:

$$\hat{Y}_i \pm t_{n-2} S_{y/x} \sqrt{1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

Aplicaremos este intervalo de predicción al problema del costo anterior:

$$1{,}81 \pm 2{,}101 \times 0{,}1321 \sqrt{1 + \frac{1}{20} + \frac{(15000 - 6524.45)^2}{425317485}}$$

$$1{,}81 \pm 0{,}277 \sqrt{1 + 0{,}05 + 0{,}168}$$

$$1{,}81 \pm 0{,}277 \times 1{,}10$$

$$1{,}81 \pm 0{,}305$$

$$1{,}51 \quad 2{,}11$$

El costo de una universidad que tiene 15.000 es un valor comprendido en el intervalo $ 1,51 y $ 2,11.

Se puede apreciar que la última estimación es menos precisa que en caso de estimar un promedio ya que el intervalo de predicción tiene mayor amplitud.

Una pregunta que no nos hemos planteado hasta ahora es la siguiente: ¿el modelo de regresión lineal produce un buen ajuste a los datos del ejemplo? ¿No nos habremos equivocado al suponer una relación lineal solamente observando el diagrama de dispersión?

Para responder esta pregunta, existe un coeficiente que precisamente mide la bondad del ajuste que hemos realizado. Este coeficiente se denomina **coeficiente de determinación** y lo estudiaremos a continuación.

4.4.4. Coeficiente de determinación

Explicaremos la lógica a seguir para Interpretar un coeficiente de determinación presentándolo a través de las SCTotal, SCRegresión y de la SCerror.

Debido a la aditividad que presentan estas sumas de cuadrados, podemos escribir:
SCT = SCR + SCe

Si dividimos cada término por SCT, tenemos:

$$\frac{SCT}{SCT} = \frac{SCr}{SCT} + \frac{SCe}{SCT}$$

De acuerdo a lo que hemos expresado cuando comentamos las componentes de variabilidad en un modelo de regresión, podemos decir que:

SCe/SCT = proporción de la variación total que no ha sido explicada por la recta de regresión.

SCR/SCT = proporción de la variación total que ha sido explicada por la recta de regresión.

Este último cociente será utilizado para medir la bondad del ajuste de los puntos realmente observados a la recta de regresión.

Entonces, el coeficiente de determinación, que simbolizaremos con R será:

$R = SCR/SCT$

El campo de variación del coeficiente de determinación es el siguiente:

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

Cuanto más se acerca a 0 el coeficiente de determinación, peor es el ajuste. Por el contrario, cuanto más se acerca R^2 a 1, mejor es el ajuste.

Cuando el coeficiente de determinación es igual a 1, quiere decir que la suma de cuadrados explicada por la regresión es igual a la variación total. Esto implica que la suma de cuadrados de error es 0.

Ahora bien, si la suma de cuadrados de error es igual a 0, esto implica que todos los puntos observados están ubicados exactamente sobre la recta de regresión. En esta situación el ajuste es perfecto.

Cuando el coeficiente de determinación es igual a 0, se produce el hecho de que la suma de cuadrados debida a la regresión es 0. En este caso, toda la variabilidad está representada por la suma de cuadrados de error. Esta situación implica que la regresión no ha explicado nada y, en consecuencia, el ajuste es malo.

Estas situaciones no se dan en la práctica, pero sí podemos juzgar la bondad del ajuste en función de si el coeficiente de determinación se cerca a 0 o a 1.

Calcularemos ahora el coeficiente de determinación con los datos del ejemplo que venimos desarrollando.

Este alto valor del R^2 indica que el ajuste ha sido muy bueno. El 95% de la variabilidad total ha sido explicada por la regresión de y en x.