

UNIDAD 2

Regresión simple

PAPIME
PE300520

Coordinador

Ing. Alejandro Augusto Pérez Pascual

Participantes

Lic. Claudia Jacqueline Bribiesca Silva,
Mtro. Ignacio Cruz López
Mtra. Hortensia Martínez Valdez
Dr. Jorge Pablo Rivas Díaz



Esta obra está bajo una Licencia Creative
Commons Atribución-NoComercial-
CompartirIgual 4.0 Internacional

CONTENIDO

1. EVALUACIÓN DIAGNÓSTICA.....	3
2. PRESENTACIÓN DE LA UNIDAD	4
3. UNIDAD 2 - ACTIVIDAD 1	7
4. UNIDAD 2 - ACTIVIDAD 2	9
5. UNIDAD 2 - ACTIVIDAD 3	10
6. UNIDAD 2 - ACTIVIDAD 4	12
7. FORMULARIO	13
8. NOTAS DE APOYO.....	14
9. AUTOEVALUACIÓN.....	21

1. EVALUACIÓN DIAGNÓSTICA

Lee con atención y responde si los siguientes enunciados son verdaderos o falsos.

1. Un modelo de regresión simple se emplea para estudiar la relación entre 2 variables.

Verdadero

Falso

2. En un modelo de regresión línea simple, la variable explicada se identifica con una X.

Verdadero

Falso

3. En la función de regresión $Y_i = \beta_0 + \beta_1 X_i + \mu$, el símbolo μ es una variable que representa factores distintos a X, que afectan el comportamiento de Y.

Verdadero

Falso

4. Se dice que un modelo de regresión es lineal en los parámetros porque el exponente de la variable X es 1.

Verdadero

Falso

5. El procedimiento de mínimos cuadrados ordinarios (MCO) consiste en minimizar la suma de todas las distancias verticales entre los valores de los datos y los de una regresión estimada.

Verdadero

Falso

2. PRESENTACIÓN DE LA UNIDAD

OBJETIVO PARTICULAR

Al finalizar el estudio de la Unidad 2, el estudiante será capaz de aplicar y analizar el procedimiento de Mínimos Cuadrados Ordinarios para obtener la función de regresión, para hacer predicciones de variables económicas, apoyado en el uso de software especializado.

TEMARIO

- 2.1 Método de momentos
- 2.2 Método de mínimos cuadrados
- 2.3 Pruebas de significancia de los coeficientes
- 2.4 Coeficiente de determinación R ajustada
- 2.5 Intervalos de confianza para los coeficientes β
- 2.6 Predicción
- 2.7 Alcances y limitaciones: análisis de resultados
- 2.8 Aplicaciones a la economía

PRESENTACIÓN

En esta unidad se estudiará el método de Mínimos Cuadrados Ordinarios (MCO), este procedimiento se atribuye al matemático alemán Carl Friedrich Gauss. A partir de ciertos supuestos, el método de mínimos cuadrados presenta propiedades estadísticas muy atractivas que lo han convertido en uno de los más eficaces y populares del análisis de regresión.

El primer tema corresponde al estudio del método de mínimos cuadrados, aquí se revisará el procedimiento para calcular los estimadores de mínimos cuadrados ordinarios (MCO) y sus propiedades a través del teorema de Gauss-Markov. Además, se revisaran las pruebas para determinar la confiabilidad o precisión de los estimadores de MCO, para ello se revisará el procedimiento para calcular $var(\hat{\beta}_1)$, $ee(\hat{\beta}_1)$, $var(\hat{\beta}_2)$, $ee(\hat{\beta}_2)$, y $cov(\hat{\beta}_1, \hat{\beta}_2)$; así como el error estándar de la regresión $\hat{\sigma}$, medida que resume la bondad de ajuste de la línea de regresión. Esta sección cierra con el estudio de dos coeficientes, el primero de ellos es el de determinación (r^2), se trata de una medida que nos indica que tan bien se ajusta la línea de regresión muestral a los datos, es decir, nos indica la proporción o el porcentaje de la variación en Y que es explicada por el modelo de regresión. El segundo se refiere al de correlación (r), se trata de una medida que nos indica el grado de asociación entre las variables del modelo.

El siguiente tema analiza las pruebas de hipótesis, aquí se estudiará que dichas pruebas son una regla que especifica si se puede rechazar o no una afirmación acerca de una población dependiendo de la evidencia proporcionada por una muestra de datos. Para esto es necesario establecer que una prueba de hipótesis examina dos hipótesis opuestas sobre una población: la hipótesis nula (H_0) y la hipótesis alternativa (H_1). Posteriormente se revisan dos métodos para el diseño de pruebas de hipótesis, el primero es el método de intervalos de confianza, se trata de construir un intervalo de confianza para β_2 a $(1 - \alpha)$. Si el β_2 en H_0 se encuentra dentro de este intervalo de confianza, no se rechaza H_0 , pero si está fuera del intervalo, entonces se rechaza H_0 . El segundo se refiere al enfoque de la prueba de

significancia, es un procedimiento que consiste en calcular el valor de un estadístico de prueba (un estimador) y su distribución muestral según la hipótesis nula. La decisión de aceptar o rechazar H_0 se toma con base en el valor del estadístico de prueba obtenido con los datos disponibles. Los estadísticos que se estudian son t y χ^2 . La sección cierra con el análisis de la varianza, para ello se revisa el procedimiento para construir la tabla ANOVA y el estimador F .

La unidad cierra con la aplicación del análisis de regresión para predecir el comportamiento de determinadas variables y con la presentación del informe de resultados.

Método de mínimos cuadrados ordinarios

Introducción

La relación económica más sencilla que se puede establecerse entre dos variables es la que se establece a través de una ecuación lineal, por ejemplo: $y = \beta_0 + \beta_1 x + u$

Donde y es la variable dependiente, x es la variable explicativa, β_0 representa el intercepto, β_1 la pendiente y u es el término de perturbación estocástica.

El método de mínimos cuadrados ordinarios (MCO) establece que, para estimaciones numéricas específicas $\hat{\beta}_0$ y $\hat{\beta}_1$, el residuo en la observación i se determina de la siguiente manera:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \beta_0 - \beta_1 x_i \quad i = 1, 2, 3, \dots, n$$

Dadas n observaciones de x e y , el método se interesa por determinar una función de regresión muestral (FRM) que esté lo más cercada posible al valor y observado. Para ello, se adopta el criterio de minimizar el error total.

$$\sum \hat{u}_i = \sum (y_i - \hat{y}_i)$$

Algunos residuos serán positivos y otros negativos, una forma de eliminar los signos es elevar al cuadrado los residuos. Por tal motivo, el criterio para seleccionar la función de regresión muestral es considerar que la suma de los errores al cuadrado sea mínima, esta acción es la razón por la que se utiliza la terminología mínimos cuadrados.

$$\sum \hat{u}_i^2 = \sum (y_i - \hat{y}_i)^2$$

Por lo tanto, a partir de una muestra, el método de mínimos cuadrados sirve para determinar los estimadores cuyo error cuadrático es mínimo. Dichos estimadores se utilizan para formar la función de regresión muestral:

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$

La confiabilidad o precisión de los estimadores de MCO se mide por su error estándar (ee), y este se obtiene al sacar la raíz cuadrada de la varianza (var). Como $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores, no sólo variarán de una muestra a otra, sino que también lo harán dentro de una muestra dada, es probable que dependan entre sí, esta dependencia se mide por la covarianza entre ellos.

Para establecer que tan bien se ajusta la línea de regresión muestral a los datos se cuenta con el coeficiente de determinación r^2 , a través de este valor se establece la proporción (porcentaje) de la variación total en la variable y que es explicada por el modelo de regresión.

Para comprender estos temas es necesario que estudies los siguientes materiales que se encuentran disponibles en la BIDI de la UNAM:

- Gujarati, D. & Porter, D. (2010) Capítulo 3 Modelo de regresión con dos variables: problema de estimación. En *Econometría*. México: McGraw-Hill.
- Wooldridge, Jeffrey M. (2015) Capítulo 2 El Modelo de regresión simple. En *Introducción a la Econometría*. Quinta Edición. Cengage.

Posteriormente realiza las actividades 1 y 2 siguiendo al pie de la letra las instrucciones.

3. UNIDAD 2 - ACTIVIDAD 1

Actividad de aprendizaje 1. Método de Mínimos Cuadrados Ordinarios

En esta actividad se describirá el método de Mínimos Cuadrados Ordinarios MCO como herramienta para el análisis econométrico.

Objetivo: Explicar los procedimientos para obtener los parámetros y coeficientes del modelo de MCO.

Instrucciones: con base en los siguientes textos:

- Gujarati, D. & Porter, D. (2010) Capítulo 3 Modelo de regresión con dos variables: problema de estimación. En *Econometría*. México: McGraw-Hill.
- Wooldridge, Jeffrey M. (2015) Capítulo 2 El Modelo de regresión simple. En *Introducción a la Econometría*. Quinta Edición. Cengage.

1. Elabora un **texto descriptivo** con la siguiente estructura:

A. Introducción: redacta en un párrafo el tema que será descrito.

B. Desarrollo: en esta sección responde las siguientes preguntas:

- i. Cómo se determinan los estimadores de MCO; explica sus propiedades (Teorema de Gauss-Markov).
- ii. Cómo se determina la línea de regresión muestral y cuáles son sus propiedades.
- iii. Cómo se determina la confiabilidad o precisión de los estimadores de MCO:
 $var(\hat{\beta}_1), ee(\hat{\beta}_1), var(\hat{\beta}_2), ee(\hat{\beta}_2), \hat{\sigma}^2$ y $cov(\hat{\beta}_1, \hat{\beta}_2)$
- iv. Cómo se determina y para qué sirve el coeficiente de determinación (r^2)
- v. Cómo se determina y para qué sirve el coeficiente de correlación (r).

C. Conclusión: con tus propias palabras explica para qué sirve el método de Mínimos Cuadrados Ordinarios (MCO).

2. El texto debe tener la estructura de un trabajo académico: carátula, Introducción, desarrollo, conclusiones, citas, notas, referencias y fuentes de consulta en formato APA.

3. La extensión máxima del texto es de 12 cuartillas.

4. Revisa los criterios de evaluación para considerarlos en el desarrollo de tu actividad.

Instrumento de evaluación:

#	Criterio	Puntaje	Cumple	
			si	no
1	Describe las fórmulas para estimar $\hat{\beta}_1$ y $\hat{\beta}_2$; explica sus propiedades (Teorema de Gauss-Markov)	2		
2	Explica cómo se integra la línea de regresión muestral y describe sus propiedades.	2		
3	Explica cómo se determina la confiabilidad o precisión de los estimadores de MCO: $var(\hat{\beta}_1)$, $ee(\hat{\beta}_1)$, $var(\hat{\beta}_2)$, $ee(\hat{\beta}_2)$, $\hat{\sigma}^2$ y $cov(\hat{\beta}_1, \hat{\beta}_2)$	2		
4	Explica cómo se determina y para qué sirve el coeficiente de determinación (r^2)	1.5		
5	Explica cómo se determina y para qué sirve el coeficiente de correlación (r)	1.5		
6	Recupera información de fuentes adicionales	0.35		
7	La actividad cumple el formato solicitado (máximo 12 diapositivas)	0.35		
8	Se envía la actividad en tiempo	0.30		
Puntaje total		10		

4. UNIDAD 2 - ACTIVIDAD 2

Actividad de aprendizaje 2. Ejercicio con el método de Mínimos Cuadrados Ordinarios

En esta actividad se aplicará el método de Mínimos Cuadrados Ordinarios (MCO).

Objetivo: Aplicar el procedimiento de MCO para obtener la función de regresión lineal e interpretar los parámetros y estimadores obtenidos de dicho proceso

Instrucciones:

1. Instala el programa de procesamiento econométrico E-Views, R, R Studio o en su defecto utiliza Excel.
2. Descarga el archivo titulado “Ejercicio1-MCO-U2”, considera la tabla que proporciona datos sobre los índices de producción por hora (X) y la compensación real por hora (Y) del sector de negocios y del sector de negocios no agrícolas de la economía de Estados Unidos de 1960 a 2005.
 - a) Grafica por separado Y respecto de X para los dos sectores.
 - b) Estima la regresión MCO de Y sobre X .
 - c) Calcula el coeficiente de determinación r^2
 - d) Calcula los valores de los siguientes estimadores: $\hat{\sigma}^2, \hat{\sigma}, var(\hat{\beta}_1), ee(\hat{\beta}_1), var(\hat{\beta}_2), ee(\hat{\beta}_2), cov(\hat{\beta}_1, \hat{\beta}_2)$
3. Explica el significado de $\hat{\beta}_1, \hat{\beta}_2, r^2$ y r
4. Revisa los criterios de evaluación para considerarlos en el desarrollo de tu actividad.

Instrumento de evaluación:

#	Criterio	Puntaje	Cumple	
			si	no
1	Grafica correctamente la información de Y respecto de X para los dos sectores.	1.5		
2	Obtiene la función de regresión de Y sobre X .	2		
3	Calcula los valores de los estimadores: $\hat{\sigma}^2, \hat{\sigma}, var(\hat{\beta}_1), ee(\hat{\beta}_1), var(\hat{\beta}_2), ee(\hat{\beta}_2)$	3		
4	Explica el significado de $\hat{\beta}_1, \hat{\beta}_2, r^2$ y r	3		
5	Se envía la actividad en tiempo	0.5		
Puntaje total		10		

Nota: Deberás guardar los resultados para completar las siguientes actividades.

5. UNIDAD 2 - ACTIVIDAD 3

Actividad de aprendizaje 3. Estimación por intervalos y pruebas de hipótesis

En esta actividad se estudiará el método de intervalos de confianza y el enfoque de pruebas de significancia para aceptar o rechazar hipótesis sobre los coeficientes de regresión.

Objetivo: Conocer los criterios del método de intervalos de confianza y del enfoque de pruebas de significancia para aceptar o rechazar hipótesis.

Instrucciones: con base en el siguiente texto:

- Gujarati, D. & Porter, D. (2010) Capítulo 5 Regresión con dos variables: estimación por intervalos y pruebas de hipótesis. En *Econometría*. México: McGraw-Hill.

1. Elabora un **texto descriptivo** con la siguiente estructura:

A. Introducción: redacta en un párrafo el tema que será descrito.

B. Desarrollo: esta sección contempla los siguientes aspectos:

- a) Elabora un cuadro donde se contraste el criterio del método de intervalos de confianza y el enfoque de pruebas de significancia para aceptar o rechazar hipótesis.
- b) Explica en que consiste el procedimiento de pruebas de significancia de σ^2 para aceptar o rechazar hipótesis.
- c) Explica en qué consiste el análisis de varianza (ANOVA).
- d) Explica en que consiste las pruebas de normalidad: histograma de residuos, gráfica de probabilidad normal (GPN) y la prueba Jarque-Bera

C. Conclusión: con tus propias palabras explica para qué sirven las pruebas de hipótesis

2. El texto debe tener la estructura de un trabajo académico: carátula, Introducción, desarrollo, conclusiones, citas, notas, referencias y fuentes de consulta en formato APA.

3. La extensión máxima del texto es de 10 cuartillas.

4. Revisa los criterios de evaluación para considerarlos en el desarrollo de tu actividad.

Instrumento de evaluación:

#	Criterio	Puntaje	Cumple	
			si	no
1	Contraste el criterio del método de intervalos de confianza y el enfoque de pruebas de significancia para aceptar o rechazar hipótesis.	3		
2	Explica en que consiste el procedimiento de pruebas de significancia de σ^2 para aceptar o rechazar hipótesis	2		
3	Explica qué es el análisis ANOVA	2		
4	Explica en qué consisten las pruebas de normalidad: histograma de residuos, gráfica de probabilidad normal (GPN) y la prueba Jarque-Bera	2		
5	Recupera información de fuentes adicionales	0.35		
6	La actividad cumple el formato solicitado (máximo 10 diapositivas)	0.35		
7	Se envía la actividad en tiempo	0.30		
Puntaje total		10		

6. UNIDAD 2 - ACTIVIDAD 4

Actividad de aprendizaje 4. Ejercicio de pruebas de hipótesis

En esta actividad se utilizarán los intervalos de confianza y las pruebas de significancia para rechazar o no hipótesis.

Objetivo: Aplicar el criterio del método de confianza y de la prueba de significancia a un ejercicio práctico.

Instrucciones:

1. Instala el programa de procesamiento econométrico E-Views, R, R Studio o en su defecto utiliza Excel.
2. Considera la información que trabajaste en el “Ejercicio1-MCO-U2” y desarrolla lo siguiente:
 - a) Con $\alpha = 5\%$ y $(n - 2)gl$, determina los intervalos de confianza para β_1 y β_2 para ambos sectores.
 - b) Para el sector de negocios, valora $H_0: \beta_2 = 0.4$
 - c) Para el sector de negocios no agrícolas, valora $H_0: \beta_2 = 0.35$
 - d) Para el sector de negocios, aplica la prueba t para valorar $H_0: \beta_2 = 0.4$
 - e) Para el sector de negocios no agrícolas, aplica la prueba t para valorar $H_0: \beta_2 = 0.35$
 - f) Para ambos sectores, calcula el valor de t cuando $H_0: \beta_2 = 0$
3. Construye la tabla ANOVA para obtener el *valor F* del ejercicio
4. Interpreta los resultados obtenidos.
5. Revisa los criterios de evaluación para considerarlos en el desarrollo de tu actividad.

Instrumento de evaluación:

#	Criterio	Puntaje	Cumple	
			si	no
1	Determina los intervalos de confianza para β_1 y β_2 para ambos sectores	1		
2	Realiza las pruebas correspondientes al método de intervalos de confianza y el enfoque de pruebas de significancia para el sector de negocios ($H_0: \beta_2 = 0.4$)	3		
3	Realiza las pruebas correspondientes al método de intervalos de confianza y el enfoque de pruebas de significancia para el sector de negocios no agrícolas ($H_0: \beta_2 = 0.35$)	3		
4	Obtiene el valor de t cuando $H_0: \beta_2 = 0$ para ambos sectores	2		
5	Construye la tabla ANOVA	1		
Puntaje total		10		

7. FORMULARIO

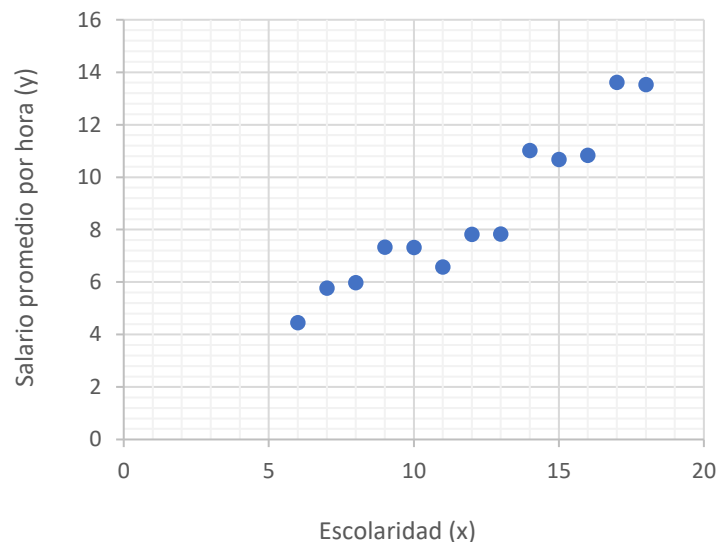
1. Función de regresión poblacional: $Y_i = \beta_1 + \beta_2 X_i + u_i$
2. Función de regresión muestral: $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$
3. Estimador $\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum x_i y_i}{\sum x_i^2}$
4. Estimador $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$
5. $\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}$
6. $\sigma^2 = \sqrt{\frac{\sum \hat{u}_i^2}{n-2}}$
7. $var(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2}$
8. $ee(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}}$
9. $var(\hat{\beta}_1) = \left(\frac{\sum X_i^2}{n \sum x_i^2} \right) \sigma^2$
10. $var(\hat{\beta}_1) = \sqrt{\left(\frac{\sum X_i^2}{n \sum x_i^2} \right)} \sigma^2$
11. $r^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{SCR}{SCT}$
12. $r = \sqrt{1 - \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2}} = \sqrt{1 - \frac{SCR}{SCT}}$
13. $t = \frac{\hat{\beta}_2 - \beta_2}{ee(\hat{\beta}_2)}$
14. Intervalos de confianza: $\left(\hat{\beta}_2 \pm t_{\alpha/2} ee(\hat{\beta}_2) \right)$ y $\left(\hat{\beta}_1 \pm t_{\alpha/2} ee(\hat{\beta}_1) \right)$
15. $\chi^2 = (n-2) \frac{\hat{\sigma}^2}{\sigma^2}$
16. Prueba de significancia: $Pr \left[-t_{\alpha/2} \leq \frac{\hat{\beta}_2 - \beta_2^*}{ee(\hat{\beta}_2)} \leq t_{\alpha/2} \right]$
17. $F = \frac{\hat{\beta}_2^2 \sum x_i^2}{\hat{\sigma}^2}$

8. NOTAS DE APOYO

Ejemplo del método de Mínimos Cuadrados Ordinarios.

Considere los datos de la siguiente tabla que relacionan el salario promedio por hora (Y) y los años de escolaridad (X). La economía laboral básica indica que, entre otros factores, la escolaridad es un elemento muy importante en la determinación de los salarios.

Obs	Salario promedio por hora Y	Escolaridad X
1	4.4567	6
2	5.7700	7
3	5.9787	8
4	7.3317	9
5	7.3182	10
6	6.5844	11
7	7.8182	12
8	7.8351	13
9	11.0223	14
10	10.6738	15
11	10.8361	16
12	13.6150	17
13	13.5310	18



Junto a la tabla se presenta una gráfica que relaciona las dos variables, advierte que a medida que el número de años de escolaridad aumenta, el salario también lo hace. Pero, es posible determinar de qué forma se relacionan escolaridad y salarios.

El método de Mínimos Cuadrados Ordinarios (MCO) nos ayuda a determinar la relación existente entre dos o más variables, a través de una función de regresión lineal que tiene una finalidad predictiva.

Lo primero que se debe hacer es calcular el valor de los estimadores $\hat{\beta}_1$ y $\hat{\beta}_2$, las fórmulas que se van a utilizar son las siguientes:

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} \quad \text{y} \quad \hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

Las fórmulas requieren calcular las desviaciones respecto de la media para ambas variables, es decir:

$$x_i = (X_i - \bar{X}) \quad \text{y} \quad y_i = (Y_i - \bar{Y})$$

Posteriormente se elevan al cuadrado las desviaciones de ambas variables:

$$x_i^2 = (X_i - \bar{X})^2 \quad \text{y} \quad y_i^2 = (Y_i - \bar{Y})^2$$

Además, se multiplican las desviaciones para cada valor observado, es decir $x_i y_i$

$\bar{Y}$ representa el valor promedio de Y , $\bar{X}$ representa el valor promedio de la variable X . Los cálculos se muestran a continuación:

Obs	Y	X	$(X_i - \bar{X})$ x	$(Y_i - \bar{Y})$ y	$(X_i - \bar{X})(Y_i - \bar{Y})$ xy	$(X_i - \bar{X})^2$ x^2	$(Y_i - \bar{Y})^2$ y^2
1	4.4567	6	-6	-4.2180	25.3080	36	17.791589
2	5.7700	7	-5	-2.9047	14.5235	25	8.437327
3	5.9787	8	-4	-2.6960	10.7840	16	7.268457
4	7.3317	9	-3	-1.3430	4.0290	9	1.803670
5	7.3182	10	-2	-1.3565	2.7130	4	1.840113
6	6.5844	11	-1	-2.0903	2.0903	1	4.369386
7	7.8182	12	0	-0.8565	0.0000	0	0.733605
8	7.8351	13	1	-0.8396	-0.8396	1	0.704941
9	11.0223	14	2	2.3476	4.6952	4	5.511190
10	10.6738	15	3	1.9991	5.9973	9	3.996370
11	10.8361	16	4	2.1614	8.6456	16	4.671617
12	13.6150	17	5	4.9403	24.7015	25	24.406488
13	13.5310	18	6	4.8563	29.1378	36	23.583575
	8.674708	12	0		131.7856	182	105.118328

Ahora se sustituye la información en las fórmulas de los estimadores:

$$\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{131.7856}{182} = 0.724097$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 8.674708 - 0.724097(12) = -0.014453$$

Con esta información es posible determinar la línea de regresión muestral con la siguiente función:

$$\hat{Y}_i = -0.014453 + 0.724097X_i$$

Con esta expresión se puede calcular la $\hat{Y}$ estimada para cada valor observado de X_i , para posteriormente determinar los residuos $\hat{u}_i = (Y_i - \hat{Y}_i)$, y elevarlos al cuadrado $\hat{u}_i^2 = (Y_i - \hat{Y}_i)^2$, tal como marca el criterio de MCO.

La sumatoria del cuadrado de las desviaciones respecto de y , es decir $\sum y_i^2$, se conoce como suma de cuadrados total (SCT), en este caso la $SCT = 105.118328$, este dato se va a utilizar más adelante en el análisis ANOVA.

Para determinar la confiabilidad o precisión de los estimadores de MCO se calcula su error estándar con ayuda de las siguientes fórmulas:

$$var(\hat{\beta}_2) = \frac{\hat{\sigma}^2}{\sum x_i^2}; \quad ee(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}}; \quad var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum x_i^2} \left(\frac{\sum x_i^2}{n} \right); \quad ee(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2} \left(\frac{\sum x_i^2}{n} \right)}$$

También será necesario calcular la varianza y el error estándar de la regresión, a través de las fórmulas:

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} \quad y \quad \hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

Los cálculos se muestran a continuación:

Obs	Y	X	$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_i$ $\hat{Y}$	X_i^2	Y_i^2	$(Y_i - \hat{Y}_i)$ $\hat{u}_i$	$\hat{u}_i^2$	$(\hat{Y}_i - \bar{Y})^2$ $\hat{y}^2$
1	4.4567	6	4.330129	36	19.862175	0.126571	0.016020	18.875364
2	5.7700	7	5.054226	49	33.292900	0.715774	0.512332	13.107888
3	5.9787	8	5.778323	64	35.744854	0.200377	0.040151	8.389044
4	7.3317	9	6.502420	81	53.753825	0.829280	0.687705	4.718834
5	7.3182	10	7.226517	100	53.556051	0.091683	0.008406	2.097256
6	6.5844	11	7.950614	121	43.354323	-1.366214	1.866541	0.524312
7	7.8182	12	8.674711	144	61.124251	-0.856511	0.733611	0.000000
8	7.8351	13	9.398808	169	61.388792	-1.563708	2.445183	0.524321
9	11.0223	14	10.122905	196	121.491097	0.899395	0.808911	2.097275
10	10.6738	15	10.847002	225	113.930006	-0.173202	0.029999	4.718863
11	10.8361	16	11.571099	256	117.421063	-0.734999	0.540224	8.389083
12	13.6150	17	12.295196	289	185.368225	1.319804	1.741883	13.107936
13	13.5310	18	13.019293	324	183.087961	0.511707	0.261844	18.875421
	8.674708	12	8.674711	2054	1083.375524	0	9.692810	95.425597

La información recién calculada se sustituye en las fórmulas anteriores.

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} = \frac{9.692810}{11} = 0.881165 \quad \hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{0.881165} = 0.938704$$

$$var(\hat{\beta}_2) = \frac{\hat{\sigma}^2}{\sum x_i^2} = \frac{0.881165}{182} = 0.004842 \quad ee(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}} = \frac{0.938704}{\sqrt{182}} = 0.069581$$

$$var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum x_i^2} \left(\frac{\sum x_i^2}{n} \right) = 0.004842 \left(\frac{2054}{13} \right) = 0.764966 \quad ee(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum x_i^2} \left(\frac{\sum x_i^2}{n} \right)} = 0.874624$$

Con estos datos es posible determinar la covarianza de los estimadores a través del siguiente cálculo:

$$cov(\hat{\beta}_1, \hat{\beta}_2) = -\bar{X} var(\hat{\beta}_2) = -12(0.004842) = -0.058104 \quad (\text{dependencia negativa})$$

La expresión $\sum \hat{u}_i^2$ se conoce como la suma de cuadrados de los residuos (SCR), en este caso la $SCR = 9.692810$; también se puede calcular la desviación de la $\hat{y}$ respecto de la media, la sumatoria del cuadrado de estas desviaciones ($\sum \hat{y}_i^2$) se conoce como suma de cuadrados explicada (SCE), en este caso la $SCE = 95.425597$; estos datos se utiliza más adelante en el análisis ANOVA.

El siguiente paso es calcular el coeficiente de determinación r^2 con la siguiente fórmula:

$$r^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2}$$

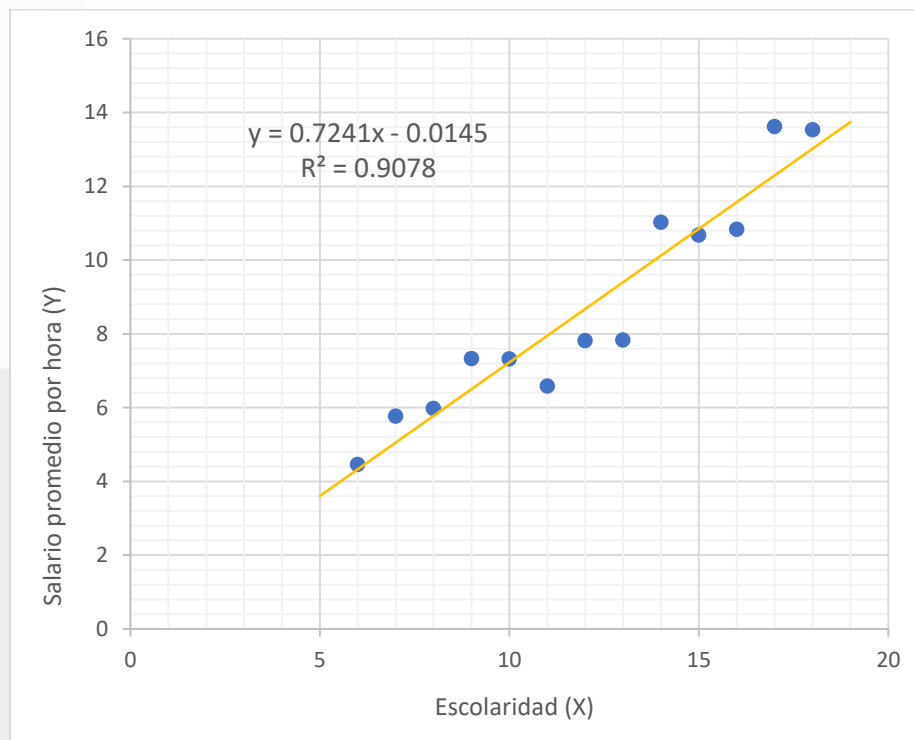
Los datos se obtienen de las tablas anteriores:

$$r^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{9.692810}{105.118328} = 0.907791$$

De forma complementaria se calcula el coeficiente de correlación para establecer el grado de asociación de las variables, el valor se obtiene de la raíz cuadrada del coeficiente de determinación r^2 :

$$r = \sqrt{r^2} = \sqrt{0.907791} = 0.953$$

Hasta este punto sabemos lo siguiente:



- Cada punto de la línea de regresión da una estimación del valor medio de Y que corresponde al valor seleccionado de X ; es decir, $\hat{Y}_i$ es una estimación de $E(Y | X_i)$.
- La línea de regresión muestra se estima por: $\hat{Y}_i = -0.014453 + 0.724097X_i$
- El valor de $\beta_2 = 0.724097$, que mide la pendiente de la línea de regresión, indica que a medida que el valor de X aumenta 1 dólar, el incremento estimado en el salario promedio por hora es cercano a 72 centavos de dólar.
- El valor de $\beta_1 = -0.014453$, que es el intercepto de la línea de regresión, indica el nivel promedio de los salarios cuando el nivel de escolaridad es cero.
- El valor de $r^2 = 0.907791$ indica que el nivel de escolaridad explica alrededor de 90 por ciento de la variación del salario por hora. Si consideramos que r^2 puede tener un valor máximo de 1, la línea de regresión se ajusta muy bien a los datos.
- El coeficiente de correlación, $r = 0.953$, muestra que los salarios y la escolaridad tienen una alta correlación positiva.

El siguiente paso involucra los intervalos de confianza y las pruebas de hipótesis.

El intervalo de confianza para β_2 a $100(1-\alpha)$ se determina con la siguiente expresión:

$$\beta_2 \pm t_{\alpha/2} ee(\beta_2)$$

De los cálculos anteriores se sabe que $\beta_2 = 0.724097$ y $ee(\hat{\beta}_2) = 0.069581$; como en la muestra del ejercicio hay 13 observaciones, los grados de libertad (gl) son 11. Si suponemos que $\alpha = 5\%$, es decir, un coeficiente de confianza a 95%, la tabla t muestra que para 11 gl el valor crítico $t_{\alpha/2} = 2.201$; al sustituir estos valores en la expresión anterior se obtiene el intervalo de confianza para β_2 a 95% de la siguiente manera:

$$0.724097 \pm 2.201(0.069581) \rightarrow 0.724097 \pm 0.153147$$

El intervalo será: $0.57095 \leq \beta_2 \leq 0.87724$

El intervalo de confianza para β_1 $\chi^2_{0.025} = 21.9200$ será: $\beta_1 \pm t_{\alpha/2} ee(\beta_1)$

$$-0.014453 \pm 2.201(0.874624) \rightarrow -0.014453 \pm 1.925047$$

El intervalo será: $-1.9395 \leq \beta_1 \leq 1.9106$

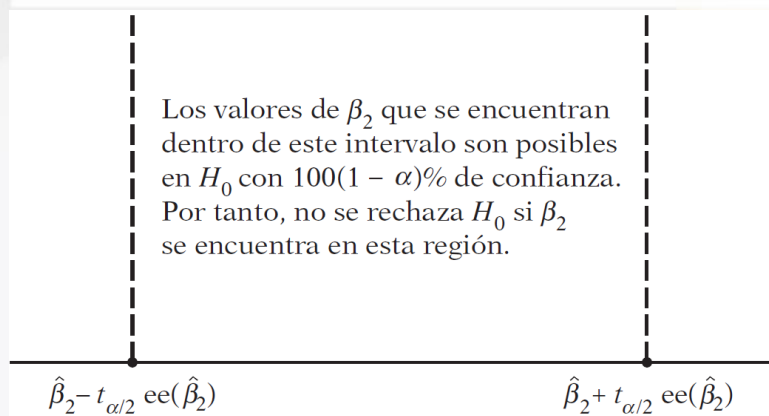
De forma análoga, el intervalo de confianza para σ^2 se determina con la siguiente expresión:

$$Pr \left[\frac{(n-2)\hat{\sigma}^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq \frac{(n-2)\hat{\sigma}^2}{\chi^2_{1-\alpha/2}} \right] = 1 - \alpha$$

Con $\hat{\sigma}^2 = 0.881165$; si seleccionamos α de 5%, la tabla ji cuadrada para 11 gl da los siguientes valores críticos: $\chi^2_{0.025} = 21.9200$ y $\chi^2_{0.975} = 3.8157$; por lo tanto, el intervalo será:

$$\left[\frac{(11)0.881165}{21.9200} \leq \sigma^2 \leq \frac{(11)0.881165}{3.8157} \right] \quad 0.44219 \leq \sigma^2 \leq 2.54025$$

Una vez que se definieron los intervalos de confianza, se pueden probar distintas hipótesis, si la regla que se establece en la hipótesis nula (H_0) cae en el intervalo, la hipótesis nula no se rechaza, de lo contrario la H_0 se rechaza. Observe la siguiente imagen:



De forma complementaria a los intervalos de confianza, se utiliza el enfoque de pruebas de significancia para probar hipótesis. El primer estadístico de prueba es se obtiene con la siguiente expresión:

$$t = \frac{\hat{\beta}_2 - \beta_2}{ee(\hat{\beta}_2)}$$

Pero también es necesario construir un intervalo de confianza con la siguiente expresión:

$$Pr \left[\beta_2^* - t_{\alpha/2} ee(\hat{\beta}_2) \leq \hat{\beta}_2 \leq \beta_2^* + t_{\alpha/2} ee(\hat{\beta}_2) \right] = 1 - \alpha$$

Donde β_2^* es el valor de β_2 en H_0 , $-t_{\alpha/2}$ y $t_{\alpha/2}$ son los valores de t (los valores críticos de t) obtenidos de la tabla t para un nivel de significancia ($\alpha/2$) con $(n - 2)$ grados de libertad (gl).

La hipótesis que se va a probar es: $H_0: \beta_2 = 0.5$ y $H_1: \beta_2 \neq 0.5$

Sabemos que $\hat{\beta}_2 = 0.724097$; $ee(\hat{\beta}_2) = 0.069581$; $gl = 11$; si suponemos $\alpha = 5\%$, entonces $t_{\alpha/2} = 2.201$

La información se sustituye para calcular t :

$$t = \frac{0.724097 - 0.5}{0.069581} = 3.220663$$

Y el intervalo de confianza:

$$\begin{aligned} 0.5 - 2.201(0.069581) &\leq \hat{\beta}_2 \leq 0.5 + 2.201(0.069581) \\ 0.5 - 0.153147 &\leq \hat{\beta}_2 \leq 0.5 + 0.153147 \\ 0.346853 &\leq \hat{\beta}_2 \leq 0.653147 \end{aligned}$$

Como el valor de t cae fuera del intervalo, la H_0 se rechaza.

Finalmente, con el análisis de la varianza (ANOVA) se obtiene la siguiente tabla:

Fuente de variación	Suma de Cuadrados (SC)	gl	Suma de Cuadrados Promedio (SCP)
Debido a la regresión	$\sum \hat{y}_i^2 = \hat{\beta}_2^2 \sum x_i^2$	1	$\hat{\beta}_2^2 \sum x_i^2$
Debido a los residuos	$\sum \hat{u}_i^2$	$n-2$	$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}$
Suma de Cuadrados Total (SCT)	$\sum y_i^2$	$n-1$	

Con la información de la tabla se obtiene el estadístico F con la siguiente fórmula:

$$F = \frac{\hat{\beta}_2^2 \sum x_i^2}{\hat{\sigma}^2}$$

La razón F constituye una prueba de la hipótesis nula $H_0: \beta_2 = 0$; sólo debe calcularse la razón F y compararla con el valor crítico F obtenido de las tablas F en el nivel de significancia seleccionado, u obtener el valor p del estadístico F calculado.

Los datos que se utilizan en la fórmula se obtienen de las tablas anteriores:

$$F = \frac{(0.724097)^2(182)}{0.881165} = \frac{95.425201}{0.881165} = 108.2943$$

El valor p del estadístico F es 0.00000050, es una probabilidad muy pequeña. Si escoge el método del nivel de significancia para la prueba de hipótesis y se fija α en 0.01, o en un nivel de 1%, se verá que la F calculada de 108.3026 es significativa en este nivel. Por tanto, si se rechaza la hipótesis nula de que $\beta_2 = 0$, la probabilidad de cometer un error tipo I es muy pequeña.

9. AUTOEVALUACIÓN

Lee con atención y responde si los siguientes enunciados son verdaderos o falsos.

1. La función de regresión poblacional muestra la relación que existe entre el promedio de Y y diferentes valores de X .

Verdadero

Falso

2. El valor de $\hat{\beta}_1$ representa la pendiente en la línea de regresión por MCO.

Verdadero

Falso

3. Un valor positivo del residual ($\hat{u}$) significa que el valor ajustado ($\hat{Y}_i$) es menor al valor observado (Y_i).

Verdadero

Falso

4. Un valor de r^2 cercano a 0, indica un muy buen ajuste de la función de regresión de MCO.

Verdadero

Falso

5. Dado el coeficiente de confianza de 95%, en 95 de cada 100 casos, el intervalo $\left(\hat{\beta}_2 \pm t_{\alpha/2} ee(\hat{\beta}_2)\right)$ contendrán al verdadero valor de β_2 .

Verdadero

Falso

6. En el lenguaje de pruebas de hipótesis, el intervalo de confianza, a $(1-\alpha)$, $Pr\left[\beta_2^* - t_{\alpha/2} ee(\hat{\beta}_2) \leq \hat{\beta}_2 \leq \beta_2^* + t_{\alpha/2} ee(\hat{\beta}_2)\right]$ se conoce como la región de rechazo (de H_0), y las regiones que quedan fuera del intervalo de confianza se llaman regiones de aceptación (de H_0) o regiones críticas.

Verdadero

Falso

7. La razón $F = \frac{\hat{\beta}_2^2 \sum x_i^2}{\hat{\sigma}^2}$ constituye una prueba de la hipótesis nula $H_0: \beta_2 = 0$. Como todas las cantidades que forman parte de esta ecuación se obtienen de la muestra disponible, la razón F constituye un estadístico de prueba para verificar la hipótesis nula de que el verdadero β_2 es igual a cero. Sólo se debe calcular la razón F y compararla con el valor crítico F obtenido de las tablas F en el nivel de significancia seleccionado, u obtener el valor p del estadístico F calculado.

Verdadero

Falso