Capítulo 13: Regresión Lineal y Correlación

Estadística Inferencial

Roberto S. Villamarín Guevara

2 de julio de 2025

Carrera de PCEMYF

Tabla de contenidos

Introducción

Overview

Lo que sabemos hasta ahora:

- Estadística Decriptiva
- Teoría de la probabilidad
- Inferencia estadística (muestra para estimar parámetros poblacionales: Media o proporción)
- utilizamos muestra para probar hipótesis
  - Media o proporción poblacional
  - Diferencia entre medias poblacionales
  - Si varias medias poblacionales son iguales
- Todas estas pruebas se realizan con variables de **intervalo o razón**
- Estudio de la relación entre dos variables de nivel de intervalo o de Razón

Análisis de la Correlación

¿Qués es el análisis de la correlación?

- Estudia la relación entre dos variables en escala de intervalo (o de razón), es usual comenzar con un diagrama de dispersión
- Este procedimiento proporciona una representación visual de la relación entre las variables
- Caclular el **Coeficiente de Correlación**, que es **una medida cuantitativa de la fuerza de la relación entre dos variables**

Análisis de la correlación
Grupo de Técnicas para medir la asociación entre dos variables

- La idea básica del análisis de correlación es reportar la asociación entre dos variables

Ejemplo

Se considera que el número de llamadas de ventas tiene relación con el número de copiadoras vendidas. Ver cuadro 1

Vendedor	N. llamadas	No. Copiadoras vendidas
Tom	20	30
Jef	40	60
Brian	20	40
Greg	30	60
Susan	10	30
Carlos	10	40
Rich	20	40
Mike	20	50
Mark	20	30
Soni	30	70

Cuadro 1: Datos Ejemplo 1

Importante

Variables en estudios Correlacionales

Cuando se realiza un estudio correlacional **NO EXISTE** variables independientes o dependientes.

Son simplemente VARIABLES [1]

[1] Aunque en algunos textos, suelen indicar que la variable independiente es aquella que se coloca en el eje X y la dependiente aquella que se ubica en el eje Y. Las variables se pueden cambiar de eje indistintamente por lo tanto, dicha afirmación no es correcta.

Algunas consideraciones

- La variable independiente proporciona la base para la estimación. Variable **Predictora**
- La variable independiente no es un número aleatorio.
- La variable dependiente es la variable que se desea **predecir o estimar**.
- La variable dependiente es **aleatoria**

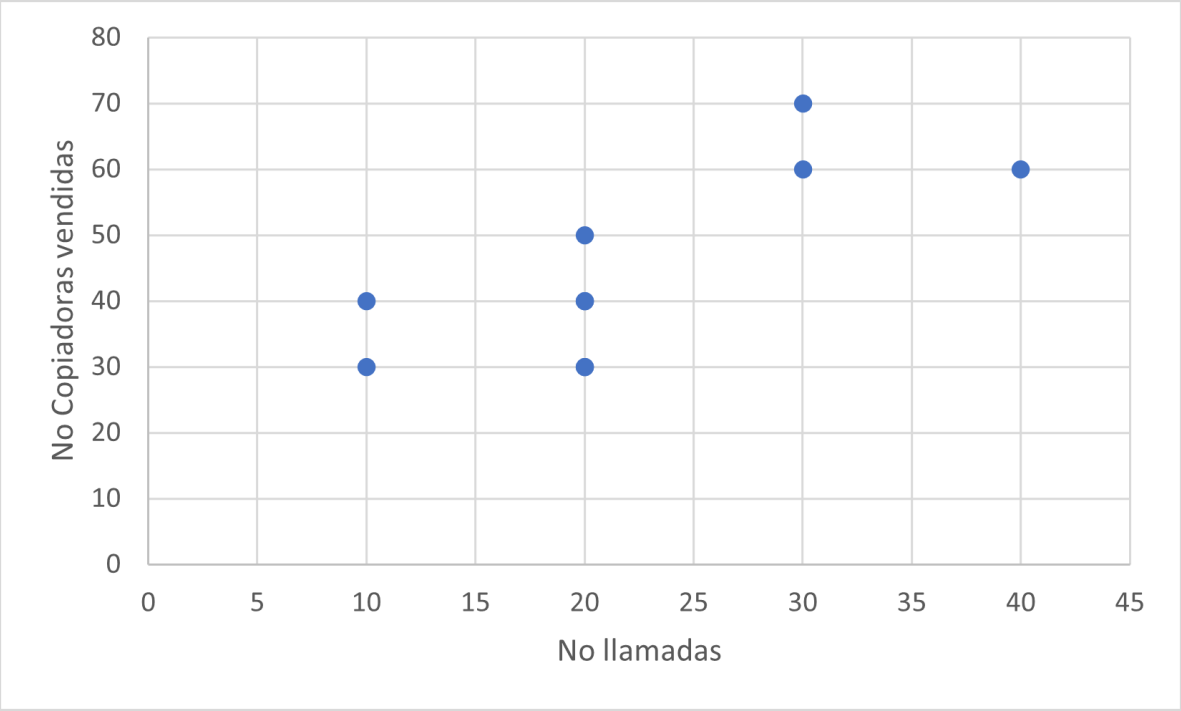

Figura 1: Grafico de Dispersión

- Parece evidente: Más llamadas $--> $ Más ventas
- Observe que, aunque parece haber una relación positiva entre las dos variables, no todos los puntos se encuentran en una recta

Coeficiente de Correlación

Coeficiente de Correlación de Pearson - 1900

- Describe la fuerza de la relación entre dos conjuntos de variables en escala de intervalo o de razón
- Se designa con la letra *r*, y con frecuencia se le conoce como **r de Pearson** y *coeficiente de correlación producto-momento.*

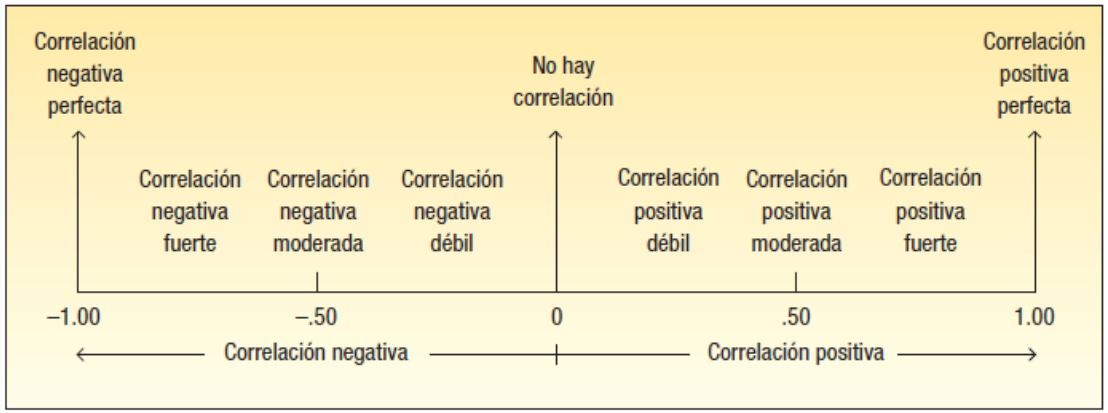

Figura 2: Valores de Correlación de Pearson

Coeficiente de Correlación
Medida de la fuerza de **relación lineal** entre dos variables

Características de la Correlación

1. El coeficiente de correlación de la muestra se identifica con la letra minúscula r.
2. Muestra la dirección y fuerza de la relación lineal (recta) entre dos variables en escala de intervalo o en escala de razón
3. Varía de -1 hasta $+1$, inclusive. $[-1, +1]$
4. Un valor cercano a 0 indica que hay poca asociación entre las variables
5. Un valor **cercano a 1** indica una asociación **directa o positiva** entre las variables.
6. Un valor cercano a **-1** indica una asociación **inversa o negativa** entre las variables

¿Cómo se determina el coeficiente de correlación?

- Calcular la media de las dos variables
- Dibujar perpendiculares a los ejes en dichos valores $\overline{X} = 22$ $\overline{Y} = 45$
- Note los valores por encima y debajo de sus respectivas medias
- Relación positiva mayor cantidad de puntos (I, III)
- Relación positiva mayor cantidad de puntos (II, IV)

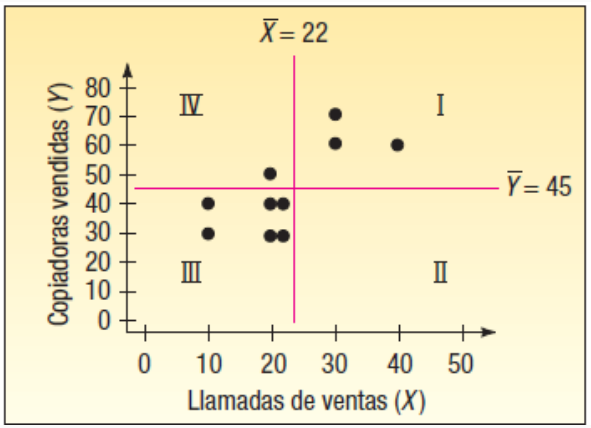

Figura 3: Cálculo del coeficiente de correlación

Representante de ventas	Llamadas, X	Ventas, Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
Tom Keller	20	30	−2	−15	30
Jeff Hall	40	60	18	15	270
Brian Virost	20	40	−2	−5	10
Greg Fish	30	60	8	15	120
Susan Welch	10	30	−12	−15	180
Carlos Ramírez	10	40	−12	−5	60
Rich Niles	20	40	−2	−5	10
Mike Kiel	20	50	−2	5	−10
Mark Reynolds	20	30	−2	−15	30
Soni Jones	30	70	8	25	200
					900

Figura 4: Desviaciones de la media y sus productos

Coeficiente de Correlación: Fórmula

Fórmula:

$$r = \frac{\sum(X - \overline{X}) \cdot (Y - \overline{Y})}{(n-1) \cdot S_x \cdot S_y} \tag{1}$$

- Calcule los valores de S_x y S_y y reemplace en la fórmula 1
- $$r = \frac{\sum(X - \overline{X}) \cdot (Y - \overline{Y})}{(n-1) \cdot S_x \cdot S_y}$$
$$r = \frac{900}{(10-1) \cdot 9{,}189 \cdot 14{,}337} = 0{,}759$$
- ¿Cómo interpretar el valor de r?
  Solución: Ver figura 2
- ¿Más llamadas, más ventas? **NO!!!.** No se ha demostrado la **cuasa-efecto**, solo que hay una relación entre la variables.

N. llamada	
Media	22
Error típico	2,9059326
Mediana	20
Moda	20
Desviación estándar	9,1893658
Varianza de la muestra	84,444444
Curtosis	0,3962208
Coeficiente de asimetría	0,6013816
Rango	30
Mínimo	10
Máximo	40
Suma	220
Cuenta	10
Mayor (1)	40
Menor(1)	10
Nivel de confianza(95,0%)	6,5736763

N. ventas	
Media	45
Error típico	4,5338235
Mediana	40
Moda	30
Desviación estándar	14,337209
Varianza de la muestra	205,55556
Curtosis	-1,0011479
Coeficiente de asimetría	0,5655291
Rango	40
Mínimo	30
Máximo	70
Suma	450
Cuenta	10
Mayor (1)	70
Menor(1)	30
Nivel de confianza(95,0%)	10,256221

	N. llamada	*N. ventas*
N. llamada	1,000000	
N. ventas	0,759014	1,000000

Figura 5: Estadísticas de los datos y valor de r, obtenidos en Excel

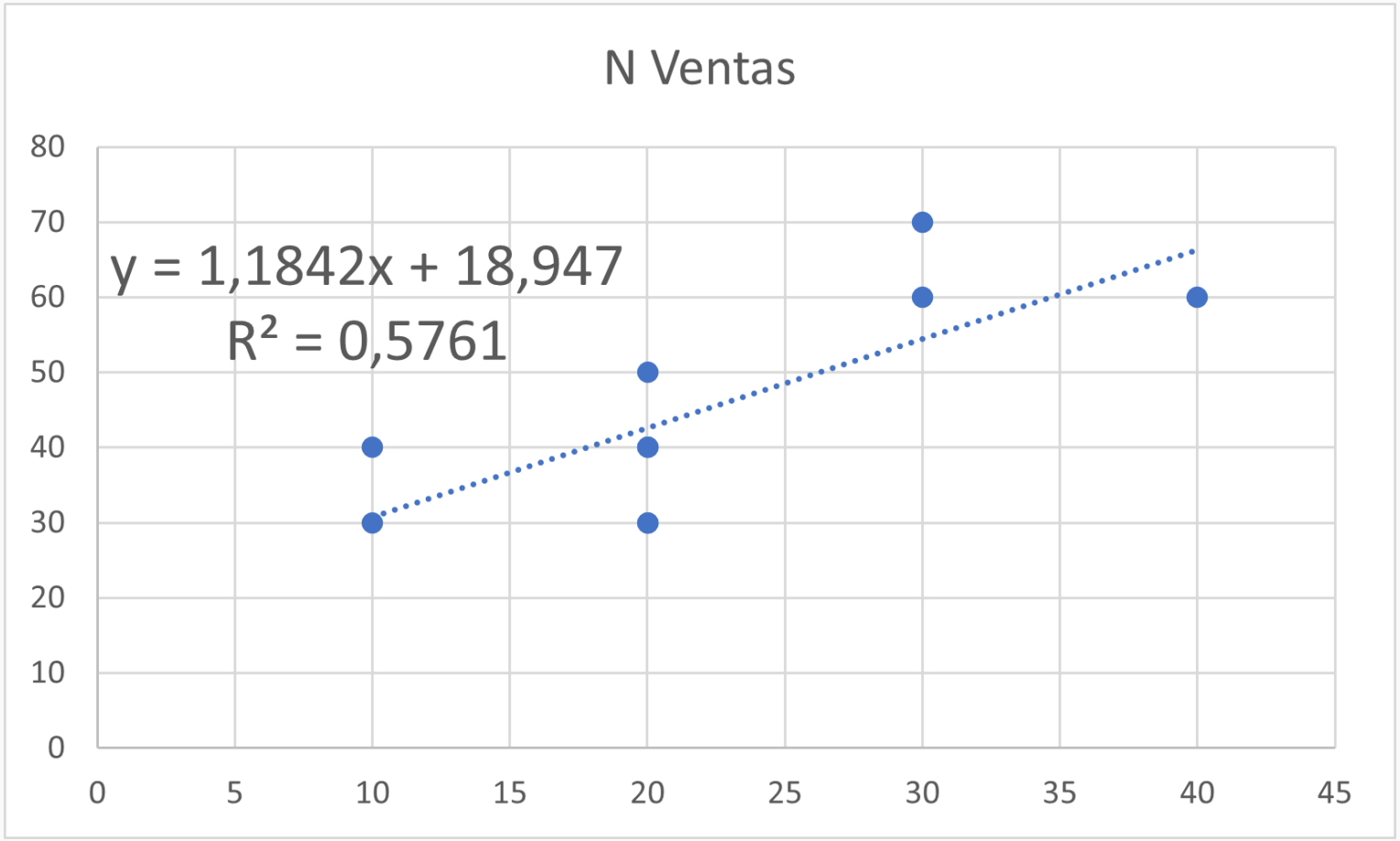

Figura 6: Gráfica de la correlación y r^2

Actividades de Aprendizaje

1. Autoevaluación $13-1$
2. Ejercicios $13-1$ al $13-6$

Texto Guía.

Prueba de la importancia de la Correlación

Aspectos a considerar:

1. ¿Puede ser que la correlación entre la población sea 0? *Esto significaría que la correlación de 0.759 se debió a la casualidad*
2. ¿puede haber una correlación *cero* entre la población de la cual se seleccionó la muestra? o; ¿proviene el valor r calculado de una población de observaciones pareadas con correlación cero?
3. Utilizaremos ρ para referirnos a la correlación entre la **población**

Prueba t para r

1º Hipótesis

- $H_0 : \rho = 0$ La correlación entre la población **es cero**.
- $H_1 : \rho \neq 0$ La correlación entre la población **es distinta cero**. Prueba a dos colas.

Prueba t del coeficiente de correlación

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} \qquad \text{con n-2 gl} \tag{2}$$

2º Significancia: $\alpha = 0{,}05$

3º Estadístico de Prueba: t

4º Valor crítico y regla de decisión:

- $gl = n - 2 = 10 - 2 = 8$
- $t_t = \pm 2{,}306$ (a dos colas)
- Regla de decisión: **Rechace** H_0 **si:** $t_c < -t_t$ **ó si** $t_c > +t_t$

5º Toma de decisión

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}759 \cdot \sqrt{10-2}}{\sqrt{1-0{,}759^2}} = 3{,}297$$

Dado que $t_c > t_t$ Se rechaza H_0, es decir; **La correlación entre la población no es cero**

Hay una correlación entre el numero de llamadas y el número de ventas realizadas.

Ejemplo 2

El coeficiente de relación entre la ganancia en la venta de un vehículo, y la edad del comprador es de 0.262. Por lo tanto existe una relación directa (baja) entre las dos variables, por lo que no existe garantías en un campaña de publicidad dirigida a personas mayores, generen una ganancia más grande. (n=180) Ver. Ejercicio pag 470

¿Significa esto no existe relación entre las variables?

Observe que:
ρ se utiliza para la relación entre la **población**
r se utiliza para la relación entre la **muestra**

Prueba de Hipótesis con r

1. $H_0 : \rho \leq 0$ La correlación entre la población es cero
   $H_1 : \rho > 0$ La correlación entre la población es positiva.
2. $\alpha = 0{,}05$
3. t para r , gl=n-2=180-2=178 (Se utilizará 180, valor más cercano)
   $t_t = 1{,}653$
4. Rechace H_0 si $t_c > t_t$
5. $t_c = \dfrac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{0{,}262 \cdot \sqrt{180-2}}{\sqrt{1-0{,}62^2}} = 3{,}622$
6. Dado que $t_c > t_t$ se rechaza H_0. Se comprueba que el coeficiente de correlación de la muestra 0.262 es demasiado grande como para provenir de una población sin correlación, es decir, si existe una correlación positiva entre la ganacia y la edad de la población.

Actividades de Aprendizaje

1. Autoevaluación $13-2$
2. Ejercicios $13-7$ al $13-12$

Texto Guía.

Análisis de la Regresión

Ecuación de la Regresión

- Se desarrollaron medidas para expresar la fuerza y la dirección de la relación entre dos variables.
- Ahora se desarrolla una ecuación para expresar la relación lineal entre dos variables
- Además, se desea estimar el valor de la **variable dependiente** Y con base en un valor seleccionado de la **variable independiente** X.

Ecuación de la Regresión
Ecuación que expresa la relación lineal entre dos variables

Del ejemplo 1, sabemos que:

- $r = 0{,}759$
- Existe una relación significativa entre las variables
- *Ahora se busca desarrollar una ecuación lineal que exprese la relación entre el número de llamadas de ventas*
- A la ecuación de la recta para estimar Y con base en X se le denomina **ecuación de regresión**.

Principio de los mínimos cuadrados

- **Objetivo:** Utilizar los datos para trazar una línea que represente mejor la relación entre las dos variables
- Realizar un diagrama de dispersión, con una recta que une los puntos para ilustrar que una recta probablemente ajustaría los datos.
- utilizar un método que resulte en una sola y mejor línea de regresión

PRINCIPIO DE LOS MÍNIMOS CUADRADOS

Determina una ecuación de regresión al minimizar la suma de los cuadrados de las distancias verticales entre los valores reales de Y y los valores pronosticados de Y.

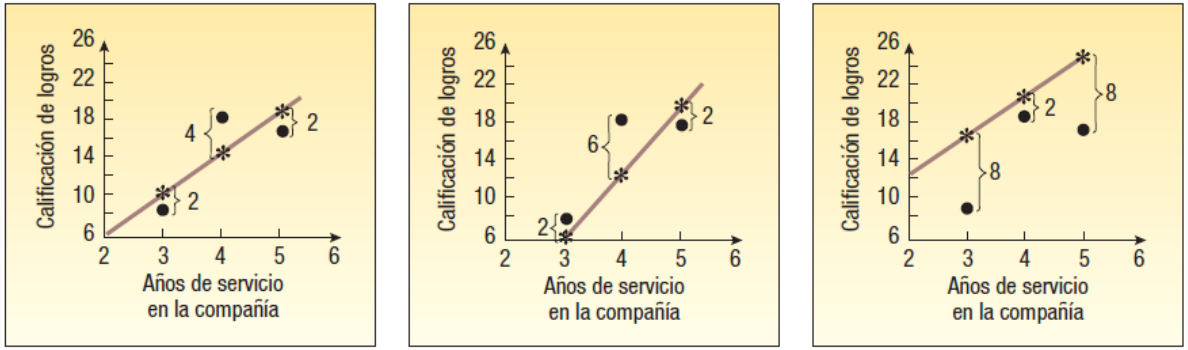

Figura 7: Aplicación del método

FORMA GENERAL DE LA ECUACIÓN DE REGRESIÓN LINEAL

$$\widehat{Y} = a + bX \tag{3}$$

donde:

- $\widehat{Y}$: (Y prima) es el valor de la estimación de la variable Y para un valor X seleccionado
- a : es la intersección Y. Valor estimado de $\widehat{Y}$ cuando $X = 0$.
- b : m de la recta, o el cambio promedio de $\widehat{Y}$, por cada cambio de unidad en X
- X : es la variable independiente que se seleccione.

OJO!!!

El propósito de la regresión es hallar los valores de a y b para desarrollar la ecuación lineal que mejor se ajuste a los datos.

Pendiente de la recta de regresión

Pendiente de la recta de regresión

$$b = r \cdot \frac{S_y}{S_x} \tag{4}$$

Donde:

- r : es el coeficiente de correlación
- S_y y S_x es la desviación estándar de Y y X, respectivamente.

Intersección con el eje Y

$$a = \overline{Y} - b \cdot \overline{X} \tag{5}$$

Donde:

- $\overline{Y}$ y $\overline{X}$ son las medias de Y y X respectivamente.

Ejemplo

Datos ejemplo 1.
Se desea presentar información específica acerca de la relación entre el número de llamadas y el número de ventas. Con el método de los mínimos cuadrados, determine una ecuación lineal que exprese la relación entre ambas variables.
¿Cuál es el número esperado de copiadoras vendidas de un representante de ventas que hizo 20 llamadas?

Solución

1. Hay que hallar el valor de b
2. $r = 0{,}759$ $S_x = 9{,}189$; $S_y = 14{,}337$ reemplazamos en (4)

$$b = r\frac{S_y}{S_x} = 0{,}759\frac{14{,}337}{9{,}189} = 1{,}1842$$

3. Hallamos el valor de a aplicando la Ecuación (5) Ver datos en la figura 5

$$a = \overline{Y} - b \cdot \overline{X} = 45 - 1{,}184(22) = 18{,}9476$$

4. **Ecuación de Regresión**:

$$\widehat{Y} = 18{,}9476 + 1{,}1842 \cdot X$$

5. Con 20 llamadas debería vender 42.6316 copiadoras.
6. Cada llamada genera 1,2 copiadoras vendidas (valor de b)
7. ¿Cuántas copiadoras se venderán con **cero** llamadas? Argumente su respuesta.

Solución-continuación

1. Resp: 18.94.
2. $X = 0$ no esta dentro del rango de valores incluidos en la muestra. **NO debe utilizarse** para estimar el número de copiadoras vendidas.
3. El número de llamadas varían de 10 a 40, por lo que las estimaciones deben realizarse en ese rango!!!!

Trazo de la recta de regresión

1. Recuerde la ecuación $\widehat{Y} = 18{,}9476 + 1{,}1842 \cdot X$
2. Trace el diagrama de dispersión. Ver (Fig. 1)
3. Halle la tabla de valores.
4. Haga la gráfica [2] Ver figuras (8 y 9)

[2] Los valores de X han sido ordenados previamente

Trazo de la recta de regresión

N. llamada	N. ventas	Ventas Estimadas
10	30	30,7896
10	40	30,7896
20	30	42,6316
20	40	42,6316
20	40	42,6316
20	50	42,6316
20	30	42,6316
30	60	54,4736
30	70	54,4736
40	60	66,3156

Figura 8: Tabla de valores de la Regresión

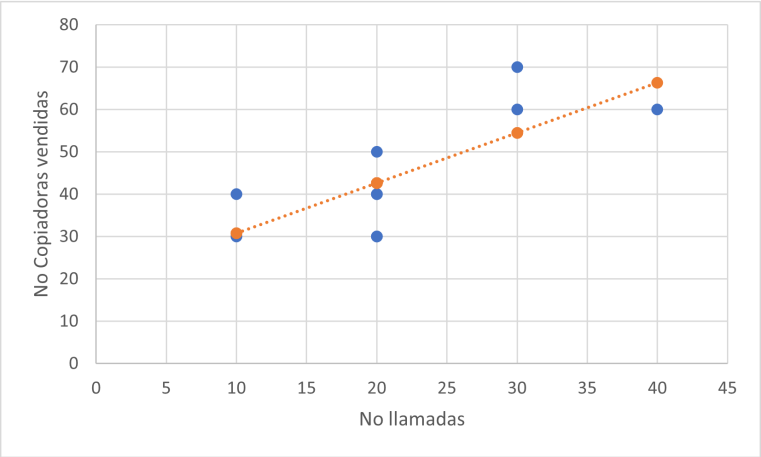

Figura 9: Gráfica de la recta de regresión

Características de la recta de regresión

1. Siempre pasa por el punto $(\overline{X}, \overline{Y})$ Demuéstrelo!
2. El número estimado de copiadoras vendidas es exactamente igual al número medio de copiadoras vendidas. Demuéstrelo!
3. No hay otra recta que pase por los datos donde la suma de las desviaciones al cuadrado es menor. Demuéstrelo Utilizando Excel.

Actividades de Aprendizaje

1. Autoevaluación $13-3$
2. Ejercicios $13-13$ al $13-20$

Texto Guía.

Probar la significancia de la pendiente

Recuerde que....

- El método para encontrar la ecuación se basa en el **principio de los mínimos cuadrados.**
- El propósito de la ecuación de regresión es **cuantificar una relación lineal entre dos variables**.
- Siguiente paso: **es analizar la ecuación de regresión mediante una prueba de hipótesis para ver si la pendiente de la recta de regresión es distinta a cero**
- ¿Por qué es importante esto?
  Si es posible demostrar que **la pendiente** de la recta de la población es distinta de cero, entonces se puede concluir que al utilizar la ecuación de regresión **aumenta la capacidad de predecir** o pronosticar la variable dependiente basándose en la variable independiente.
- Si *no podemos demostrar que la pendiente de la recta es distinta de cero, podríamos utilizar la media de la variable dependiente como factor de predicción*, en vez de usar la ecuación de regresión.

Prueba

Hipótesis

- $H_0 : \beta = 0$ [3] [4]
- $H_1 : \beta \neq 0$

Análisis de los datos en Excel

Estadísticas de la regresión								
Coeficiente de correlación múltiple	0,7590141							
Coeficiente de determinación R^2	0,5761024							
R^2 ajustado	0,5231152							
Error típico	9,900824							
Observaciones	10							
ANÁLISIS DE VARIANZA								
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*			
Regresión	1	1065,7895	1065,789474	10,8724832	0,0109019			
Residuos	8	784,21053	98,02631579					
Total	9	1850						
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*
Intercepción	18,947368	8,4988186	2,22941204	0,05634865	-0,6509423	38,545679	-0,6509423	38,545679
N. llamada	1,1842105	0,3591406	3,297344875	0,01090193	0,3560307	2,0123903	0,3560307	2,0123903

Figura 10: Análsis de la regresión en Excel

[3] β es la pendiente de la población de la ecuación de regresión

[4] Si la pendiente de la recta es cero, es una recta horizontal

Análisis de los datos en Excel

1. $r = 0{,}7590141$ que ya se calculó.
2. **ANOVA** Detalles más adelante
3. **Coeficentes** Resaltado en amarillo y verde. Ligeramente diferentes a los ya obtenidos
4. **Error estándar**. Muy similar al error estándar de la media. Recuerde que el *error estándar* de la media reporta la variación entre las medias muestrales.
   El error estándar del coeficiente de la pendiente es 0,35914.
   Reportan la posible variación de los valores de la pendiente y de la intersección.

Para probar la hipótesis nula, utilizamos la distribución t con (n–2 grados de libertad) y la siguiente fórmula:

Prueba de la pendiente

$$t = \frac{b - 0}{S_b} \tag{6}$$

con $gl = n - 2$

Donde:

- b : es la estimación de la pendiente de la recta de regresión, calculada a partir de la información de la muestra.
- S_b es el error estándar de la estimación de la pendiente, determinado también a partir de la información de la muestra.

Prueba de la pendiente

1º Hipótesis

- $H_0 : \beta \leq 0$
- $H_1 : \beta > 0$ [5]
  - Si no rechazamos la hipótesis nula, se concluye que la pendiente de la recta de regresión entre la población podría ser cero.
  - Si rechazamos la hipótesis nula y aceptamos la alternativa, se concluye que la pendiente de la recta es mayor a cero.
  - Por lo tanto, la variable independiente es una ayuda para predecir la variable dependiente.

2º Nivel de significancia $\alpha = 0{,}05$

3º Estadístico: t

- $gl = n - 2 = 8$
- $\alpha = 0,05$
- $t_t = 1{,}860$

[5] Prueba a una cola.

Prueba de la pendiente

4º Regla de Decisión: Rechazar la H_0 si $t_c > t_t$

5º Toma de decisión:

- $$t = \frac{b - 0}{S_b} = \frac{1{,}18421 - 0}{0{,}35814} = 3{,}297$$
- Dado que $t_c > t_t$ se rechaza H_0
- **Concluimos que la pendiente de la recta es mayor a cero**. *La variable independiente, que se refiere al número de llamadas de venta, es útil para obtener una mejor estimación de las ventas*.
- Se verifica lo mismo con el $p\_valor = 0{,}01090$ [6]
- Los valores obtenidos de $t = 3{,}297$ y con la Ec. (2) son los mismos para t y p_valor. Son pruebas equivalentes.

[6] los valores p que se reportan en el software estadístico suelen ser para una prueba de dos colas.

Actividades de Aprendizaje

1. Autoevaluación $13-4$
2. Ejercicios $13-21$ al $13-24$

Texto Guía.

Evaluación de la capacidad predictora de una ecuación de regresión

Error estándar de estimación

En relación al problema de las copiadoras la ecuación puede reescribirse de la siguiente manera

$$\#copiadoras\ vendidas = 18,9476 + 1,1842 * \#dellamadas$$

La ecuación sirve estimar el número de copiadoras vendidas por cada **número de llamadas de ventas** dentro del rango de los datos.

Pregunta: **¿La ecuación de regresión es un buen predictor del "Número de copiadoras vendidas"?**
Es necesario contar con una medida para describir cuán preciso es el pronóstico de Y con base en X, o a la inversa, qué tan inexacta puede ser la estimación. Esta medida se denomina **error estándar de estimación.**

Error estándar de estimación
Medida de la dispersión de los valores observados respecto de la recta de regresión para un valor dado de X.

Error estándar de estimación

$$S_{y \cdot x} = \sqrt{\frac{\sum (Y - \widehat{Y})^2}{n - 2}} \tag{7}$$

Para el caso del ejemplo estudiado:

$$S_{y \cdot x} = \sqrt{\frac{784,211}{10 - 2}} = 9,901$$

El valor del numerador y el $S_{y \cdot x}$ ya consta en la Figura 10 **Interpretación del Error**

- $S_{y \cdot x}$ pequeño, los datos están relativamente cercanos a la recta de regresión y la ecuación sirve para predecir los valores de $\widehat{Y}$
- $S_{y \cdot x}$ grande los datos están relativamente lejanos a la recta de regresión y la ecuación no proporcionará una estimación precisa de $\widehat{Y}$

El coeficiente de determinación o r^2

Proporciona una medida relativa de la capacidad de predicción de una ecuación de regresión.

El coeficiente de determinación o r^2
Proporción de la variación total de la variable dependiente Y que se explica, o contabiliza, por la variación de la variable dependiente X.

Es el coeficiente de Correlación al cuadrado.

$$(r)^2 = (0,79)^2 = 0,576 = 57,6\,\%$$

Interpretación de r^2
Se dice que 57,6 % de la variación del número de copiadoras vendidas se explica, o está representado por la variación del número de llamadas de ventas

Actividades de Aprendizaje

1. Autoevaluación $13-5$
2. Ejercicios $13-25$ al $13-28$

Texto Guía.

Relaciones entre el coeficiente de correlación, el coeficiente de determinación y el error estándar de estimación

- El error estándar de estimación, el cual mide la cercanía entre los valores reales y la recta de regresión.
- En el cálculo del error estándar, el término clave de la Ec. 7 es $\sum(Y - \widehat{Y})^2$, si éste valor es pequeño el error estándar de estimación, también lo será.
- El coeficiente de correlación mide la fuerza de la asociación lineal entre dos variables.
- Cuando los puntos del diagrama de dispersión aparecen cerca de la recta, se observa que el coeficiente de correlación tiende a ser grande. *Todo ello indica que el error estándar de estimación y el coeficiente de correlación están inversamente relacionados.*
- También se hizo notar que el cuadrado del coeficiente de correlación es el coeficiente de determinación, que mide el porcentaje de la variación de Y que se explica por la variación de X.

Observe la tabla ANOVA:

Estadísticas de la regresión									
Coeficiente de correlación múltiple	0,7590141								
Coeficiente de determinación R^2	0,5761024								
R^2 ajustado	0,5231152								
Error típico	9,900824								
Observaciones	10								
ANÁLISIS DE VARIANZA									
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*				
Regresión	1	1065,7895	1065,789474	10,8724832	0,0109019				
Residuos	8	784,21053	98,02631579						
Total	9	1850							
	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95,0%*	*Superior 95,0%*	
Intercepción	18,947368	8,4988186	2,22941204	0,05634865	-0,6509423	38,545679	-0,6509423	38,545679	
N. llamada	1,1842105	0,3591406	3,297344875	0,01090193	0,3560307	2,0123903	0,3560307	2,0123903	

Figura 11: ANOVA

- Tabla es similar a las estudiadas en el capítulo anterior, donde se estudió la variación debido a dos componentes: **los tratamientos** y los **los debidos a los errores aleatorios (dentro de los grupos)**
- Ahora, las variaciones son debidas a:
    1. La que explica la regresión (a su vez explicada por la variables independiente)
    2. El error o variación inexplicable.
- En la figura 11: (Aplica para cualquier ANOVA)
    - *gl* Se refiere a los grados de libertad asociados a cada categoría. $n-1$. En la regresión es 1, **solo hay una variable independiente**
    - El número de grados de libertad asociados con el término de error es $n-2$.
    - *SS* ubicado en medio de la tabla ANOVA, se refiere a la suma de los cuadrados.
    - El total de los grados de libertad es igual a la suma de los grados de libertad de la regresión y del residual (error)
    - suma total de los cuadrados es igual a la suma delos cuadrados de la suma de la regresión y el residuo (error).

La suma de los cuadrados de ANOVA, se calcula así:

- Suma de regresión de los cuadrados:
  $SSR = \sum(\widehat{Y} - \overline{Y})^2 = 1065,789$
- Suma del residual o error de los cuadrados:
  $SSE = \sum(Y - \widehat{Y})^2 = 784,211$
- Suma total de los cuadrados: $SS_t = \sum(Y - \overline{Y})^2 = 1850,00$
- Coeficiente de determinación (r^2) se define como el porcentaje de variación total SS_t explicado por la ecuación de regresión SSR.
- r^2 puede ser validado mediante la tabla ANOVA

Coeficiente de determinación

$$r^2 = \frac{SSR}{SS_t} = 1 - \frac{SSE}{SS_t} \tag{8}$$

- $r^2 = \dfrac{1065,789}{1850} = 0,576$
- $r^2 = 1 - \dfrac{SSE}{SS_t} = 1 - \dfrac{784,211}{1850} = 1 - 0,424 = 0,576$

- El coeficiente de determinación y la suma del residuo o error de los cuadrados están **inversamente relacionados.**
- 42,4 % de la variación total de la variable dependiente es una variación residual o error. $(1 - r^2 = 0,576)$

Al sustituir SSE por $\sum(Y - \widehat{Y})^2$ en la fórmula del error estándar de estimación tenemos:

Error Estándar de estimación

$$S_{y \cdot x} = \sqrt{\frac{SSE}{n-2}} \tag{9}$$

- En suma, el análisis de regresión proporciona dos estadísticos para evaluar la **capacidad de predicción** de una ecuación de regresión: el **error estándar de estimación** y el **coeficiente de determinación.**

Actividades de Aprendizaje

1. Ejercicios $13-29$ al $13-30$

Texto Guía.

Estimaciones de intervalos de predicción

Overview

- Recuerde que para cada valor seleccionado de la variable independiente (X), la variable dependiente (Y) es una variable aleatoria que está distribuida normalmente con una media
- Cada distribución de Y tiene una desviación estándar igual al error estándar de estimación del análisis de regresión.
- Cuando se utiliza una ecuación de regresión, se pueden hacer dos predicciones distintas para un valor seleccionado de la variable independiente.

1. Se utiliza cuando la ecuación de regresión se emplea para predecir el valor medio de Y para una X dada.

Intervalos de Confianza de la media de Y, dada X

$$\widehat{Y} \pm t \cdot (S_{y \cdot x}) \sqrt{\frac{1}{n} + \frac{(X - \overline{X})^2}{\sum (X - \overline{X})^2}} \tag{10}$$

2. Se utiliza cuando la ecuación de regresión se emplea para predecir una Y individual $(n = 1)$ para un valor dado de X

Intervalos de Confianza de Predicción de Y, dada X

$$\widehat{Y} \pm t \cdot (S_{y \cdot x}) \sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{\sum (X - \overline{X})^2}} \tag{11}$$

Ejemplo

Determine un intervalo de confianza de 95 % para todos los representantes de ventas que hacen 25 llamadas y un intervalo de predicción para Sheila Baker, representante de ventas de la Costa Oeste que hizo 25 llamadas. Emplee la fórmula 10 para determinar un intervalo de confianza.

Representante de ventas	Llamadas de ventas, (X)	Ventas de copiadoras, (Y)	$(X - \bar{X})$	$(X - \bar{X})^2$
Tom Keller	20	30	−2	4
Jeff Hall	40	60	18	324
Brian Virost	20	40	−2	4
Greg Fish	30	60	8	64
Susan Welch	10	30	−12	144
Carlos Ramírez	10	40	−12	144
Rich Niles	20	40	−2	4
Mike Kiel	20	50	−2	4
Mark Reynolds	20	30	−2	4
Soni Jones	30	70	8	64
			0	760

Figura 12: Tabla de datos

Solución. Parte 1

1. Determinar el número de copiadoras que se espera que venda un representante (cualquiera) de ventas si él o ella hacen 25 llamadas.
   $\widehat{Y} = 18,9476 + 1,184x = 18,9476 + 1,184(25) = 48,5526$
2. Hallar el valor de t
   - $gl = n - 2 = 10 - 2 = 8$
   - nivel de confianza 95 %, $\alpha = 0,05$
   - $t_t = 2,036$
3. Error estándar de estimación : $9,901$
4. $X = 25$; $\overline{X} = \dfrac{\sum X}{n} = \dfrac{220}{10} = 22$; $\sum(X - \overline{X}) = 760$
5. Reemplazando en la Ec. 10, tenemos:

$$IC = \widehat{Y} \pm t \cdot (S_{y \cdot x}) \sqrt{\frac{1}{n} + \frac{(X - \overline{X})^2}{\sum (X - \overline{X})^2}}$$

$$IC = 48,5526 \pm 2,306 \cdot (9,901) \sqrt{\frac{1}{10} + \frac{(25 - 22)^2}{760}}$$

$$IC = 48,5526 \pm 7,6356$$

Solución. Parte 1

- Por lo tanto, el intervalo de confianza de 95 % de todos los representantes de ventas que hacen 25 llamadas es de 40,9170 a 56,1882.
- **Si un representante de ventas hace 25 llamadas, debería vender 48.6 copiadoras.**
- Es probable que estas ventas varíen de 41 a 56 copiadoras. (datos discretos).

Solución. Parte 2

Se desea estimar el número de copiadoras que vendió Sheila Baker, quien hizo 25 llamadas.

El **intervalo de predicción** de 95 % se determina como sigue:

$$I_d P = \widehat{Y} \pm t \cdot (S_{y \cdot x}) \sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{\sum (X - \overline{X})^2}}$$

$$I_d P = 48,5526 \pm 2,306 \cdot (9,901) \sqrt{1 + \frac{1}{10} + \frac{(25 - 22)^2}{760}}$$

$$I_d P = 48,5526 \pm 24,0746$$

- El intervalo es de 24,478 a 72,627 copiadoras
- Se concluye que el número de copiadoras que venderá **un representante** que haga 25 llamadas estará aproximadamente entre 24 y 73

NO olvide que:

Recuerde!!!

Un **intervalo de confianza** se refiere a todos los casos con un valor dado de X y su valor se calcula por medio de la fórmula 10.

Un **intervalo de predicción** se refiere a un caso particular de un valor dado de X y su valor se determina mediante la fórmula 11.

El intervalo de predicción siempre será más ancho debido al 1 adicional debajo del radical en la segunda ecuación.

Actividades de Aprendizaje

1. Autoevaluación $13-6$
2. Ejercicios $13-31$ al $13-34$

Texto Guía.

¿Preguntas?