18 de mayo de 2022

Regresión Lineal Simple

Módulo 2

—

Modelado estadístico

Agenda de hoy

1. Bibliografía Módulo 2
2. Análisis de Regresión
3. Regresión Lineal Simple
4. El método de Mínimos Cuadrados

Bibliografía Módulo 2

1. **Probabilidad y Estadística. Aplicaciones y Métodos:** https://gsosa61.files.wordpress.com/2008/03/10-canavos-g-probabilidad-y-estadistica-aplicaciones-y-metodos.pdf
2. **Modelos Estadísticos en Lenguaje R:** https://rid.unrn.edu.ar/bitstream/20.500.12049/5789/2/garibaldi_lenguajeR_eunrn.pdf
3. **Estadística para administración y economía:** https://www.upg.mx/wp-content/uploads/2015/10/LIBRO-13-Estadistica-para-administracion-y-economia.pdf
4. **Probabilidad y Estadística para Ingeniería y Ciencias:** https://vereniciafunez94hotmail.files.wordpress.com/2014/08/8va-probabilidad-y-estadistica-para-ingenier-walpole_8.pdf

Análisis de Regresión

El Análisis de Regresión es un procedimiento estadístico que nos permite obtener una ecuación que indica cuál es la relación entre ciertas variables

- **Variable dependiente:** variable que se va a predecir
- **Variable independiente:** variables que se usan para predecir el valor de la variable dependiente

Regresión Lineal Simple (1/5)

Tipo de Análisis de Regresión en el cual intervienen una variable independiente y una variable dependiente y en el que la relación entre estas dos variables es aproximada mediante una línea recta

Regresión Lineal Simple (2/5)

Cadena de restaurantes de pizza

Campus Universitarios

¿Qué variables podrían influir en mis ventas mensuales?

Regresión Lineal Simple (3/5)

Los restaurantes que están cerca o dentro de campus universitarios que tienen una población estudiantil grande generan más ventas que los restaurantes situados cerca o dentro de campus con una población estudiantil pequeña

Ventas mensuales (y) están directamente relacionadas con la Población estudiantil (x)

Regresión Lineal Simple (4/5)

Mediante el Análisis de Regresión podemos obtener una **ecuación** que muestre cuál es la relación entre la variable dependiente **y** y la variable independiente **x**

Modelo de Regresión Lineal Simple

$$y = \beta_0 + \beta_1 x + \epsilon \qquad \textbf{(1)}$$

Parámetros del modelo

Variable conocida como término del error

Regresión Lineal Simple (5/5)

En la práctica, los valores de los parámetros del modelo no se conocen y es necesario estimarlos usando **datos muestrales**

Ecuación de Regresión Lineal Simple Estimada

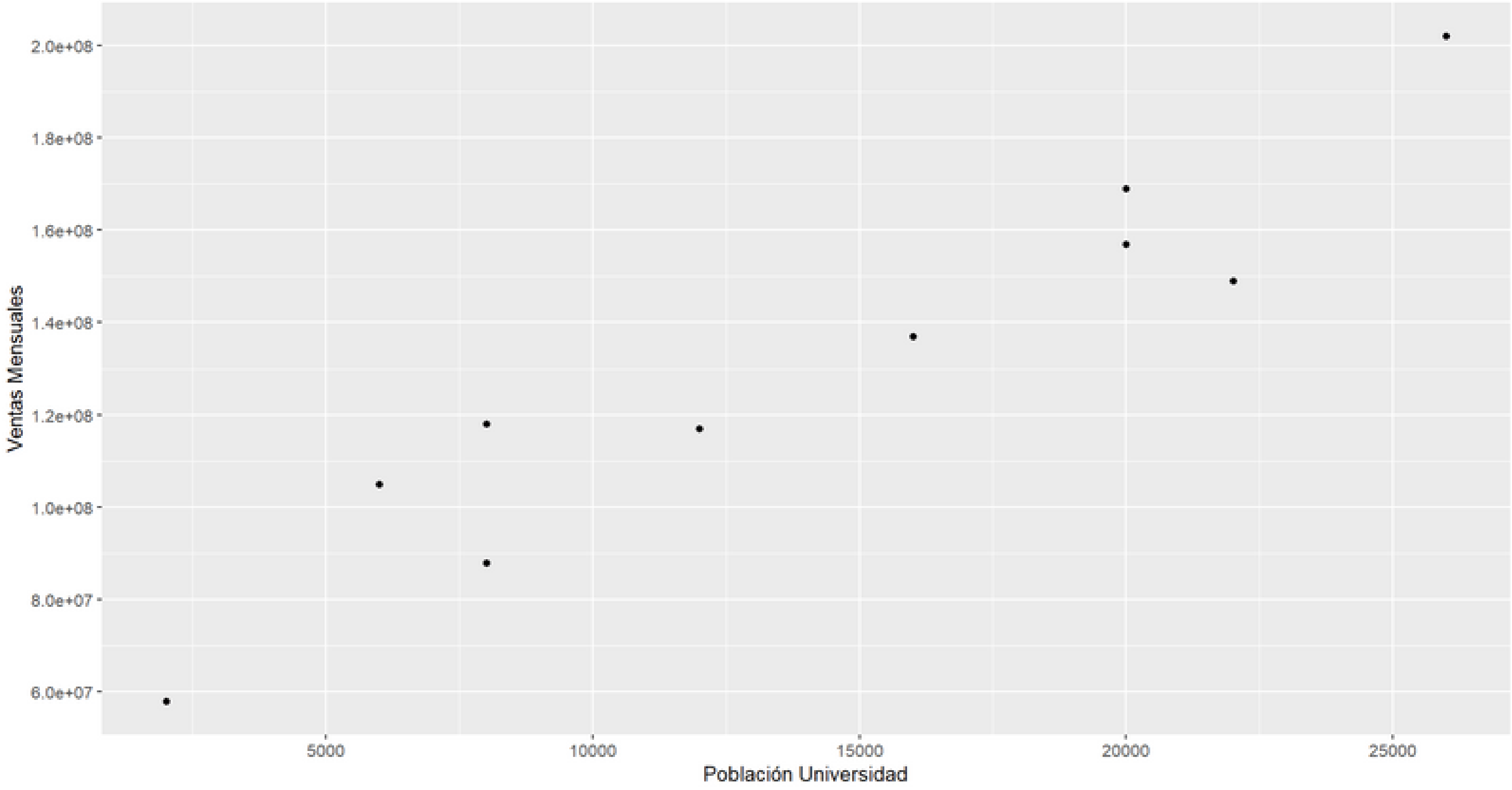
$$\hat{y} = \underbrace{b_0}_{\text{Intersección con eje y}} + \underbrace{b_1}_{\text{Pendiente}} x \qquad (2)$$

Datos muestrales recolectados...

Restaurante i	Población de estudiantes	Ventas mensuales
1	2.000	58.000.000
2	6.000	105.000.000
3	8.000	88.000.000
4	8.000	118.000.000
5	12.000	117.000.000
6	16.000	137.000.000
7	20.000	157.000.000
8	20.000	169.000.000
9	22.000	149.000.000
10	26.000	202.000.000

Datos muestrales recolectados de diferentes restaurantes

Grafiquemos los datos muestrales...



Se puede observar que a medida que aumenta la Población Estudiantil, las Ventas Mensuales aumentan. Además, la relación entre estas dos variables parece poder aproximarse mediante una línea recta

Método de Mínimos Cuadrados (1/7)

Es un método en el que se usan los datos muestrales para hallar la ecuación de regresión estimada

$$\underbrace{\hat{y}_i}_{\text{valor estimado de las ventas mensuales del restaurante i}} = b_0 + b_1 \underbrace{x_i}_{\text{tamaño de la población de estudiantes del restaurante i}} \quad (3)$$

Método de Mínimos Cuadrados (2/7)

Restaurante i

$y_i = \ ventas\ observadas\ reales$

$\hat{y}_i = \ ventas\ estimadas$

Para que la recta de regresión estimada proporcione un buen ajuste a los datos, las diferencias entre los valores observados y los valores estimados deben ser pequeñas [3]

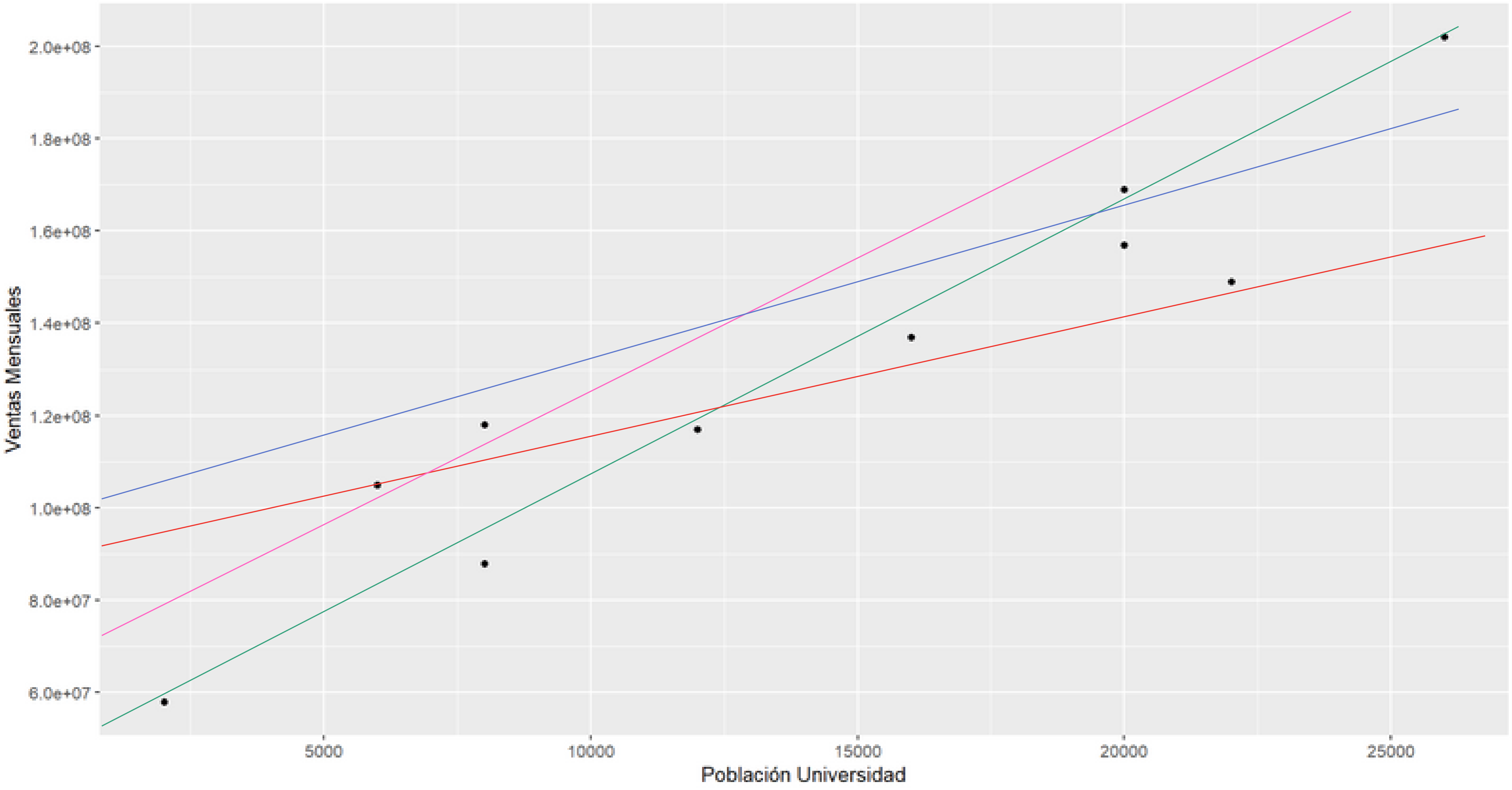
Método de Mínimos Cuadrados (3/7)

¿Cómo funciona el Método de Mínimos Cuadrados?

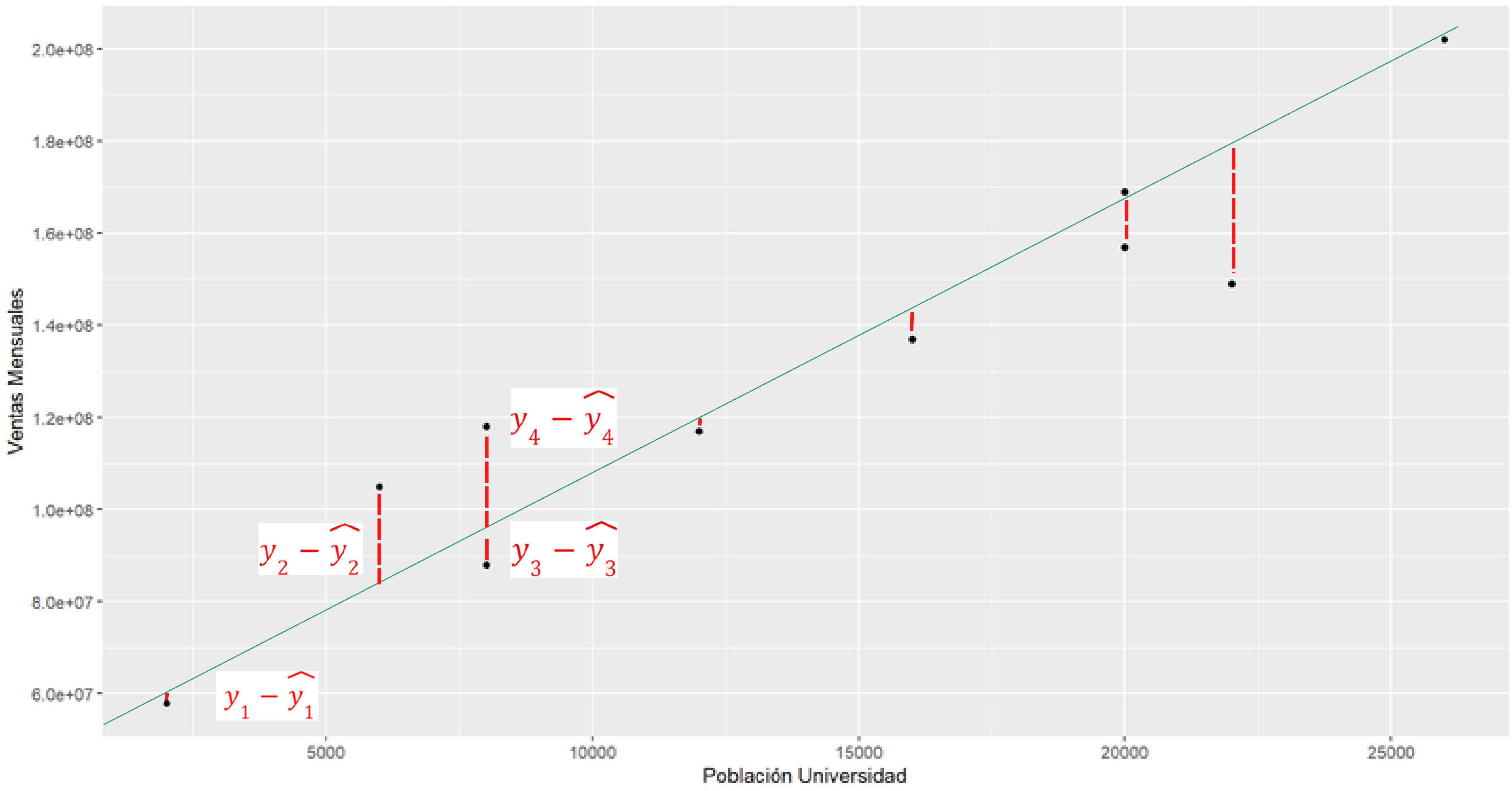
Este método usa los datos muestrales recolectados para obtener los valores de b_0 y b_1 que minimicen la suma de los cuadrados de las diferencias entre los valores observados y_ii y los valores estimados mediante la recta de regresión $\hat{y}_i$

$$min \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \qquad \textbf{(4)}$$

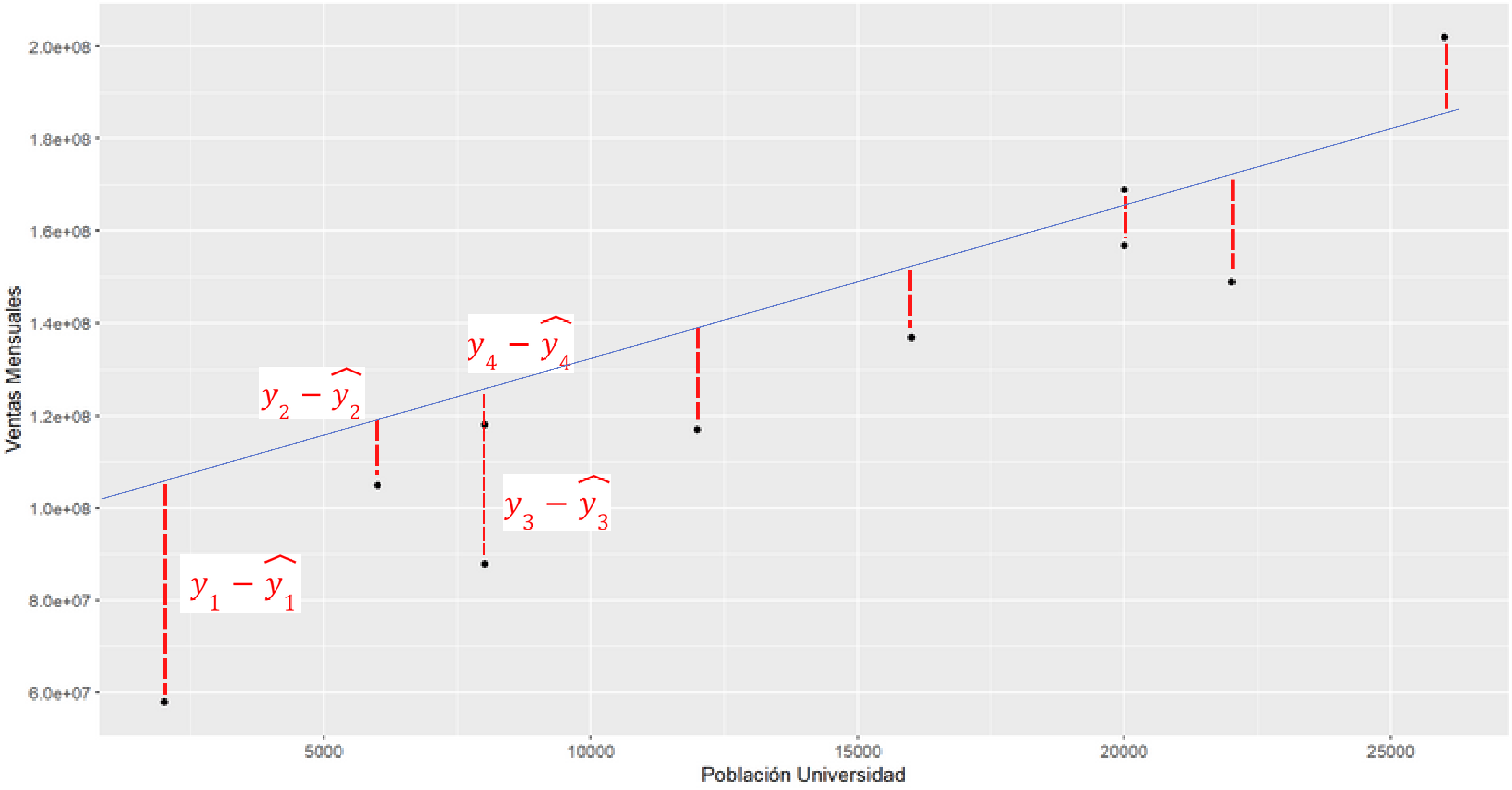
Método de Mínimos Cuadrados (4/7)



Método de Mínimos Cuadrados (5/7)



Método de Mínimos Cuadrados (6/7)



Método de Mínimos Cuadrados (7/7)

Se puede demostrar que los valores de b_0 y b_1 que minimizan la expresión **(4)** se pueden encontrando usando las siguientes ecuaciones:

(5) $$b_1 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

(6) $$b_0 = \bar{y} - b_1\bar{x}$$

x_i valor de la variable independiente en la observación i

y_i valor de la variable dependiente en la observación i

$\bar{x}$ promedio de la variable independiente

$\bar{y}$ promedio de la variable dependiente

n número total de observaciones

Apliquemos el Método de Mínimos Cuadrados...

Restaurante i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2.000	58.000.000	-12.000	-72.000.000	864.000.000.000	144.000.000
2	6.000	105.000.000	-8.000	-25.000.000	200.000.000.000	64.000.000
3	8.000	88.000.000	-6.000	-42.000.000	252.000.000.000	36.000.000
4	8.000	118.000.000	-6.000	-12.000.000	72.000.000.000	36.000.000
5	12.000	117.000.000	-2.000	-13.000.000	26.000.000.000	4.000.000
6	16.000	137.000.000	2.000	7.000.000	14.000.000.000	4.000.000
7	20.000	157.000.000	6.000	27.000.000	162.000.000.000	36.000.000
8	20.000	169.000.000	6.000	39.000.000	234.000.000.000	36.000.000
9	22.000	149.000.000	8.000	19.000.000	152.000.000.000	64.000.000
10	26.000	202.000.000	12.000	72.000.000	864.000.000.000	144.000.000
Total	**140.000**	**1.300.000.000**	-	-	**2.840.000.000.000**	**568.000.000**
Promedio	**14.000**	**130.000.000**	-	-	-	-

Apliquemos el Método de Mínimos Cuadrados...

$$b_1 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = \frac{2.840.000.000.000}{568.000.000} = 5.000$$

$$b_0 = \bar{y} - b_1\bar{x} = 130.000.000 - 5.000(14.000) = 60.000.000$$

$$\hat{y} = 60.000.000 + 5.000x$$

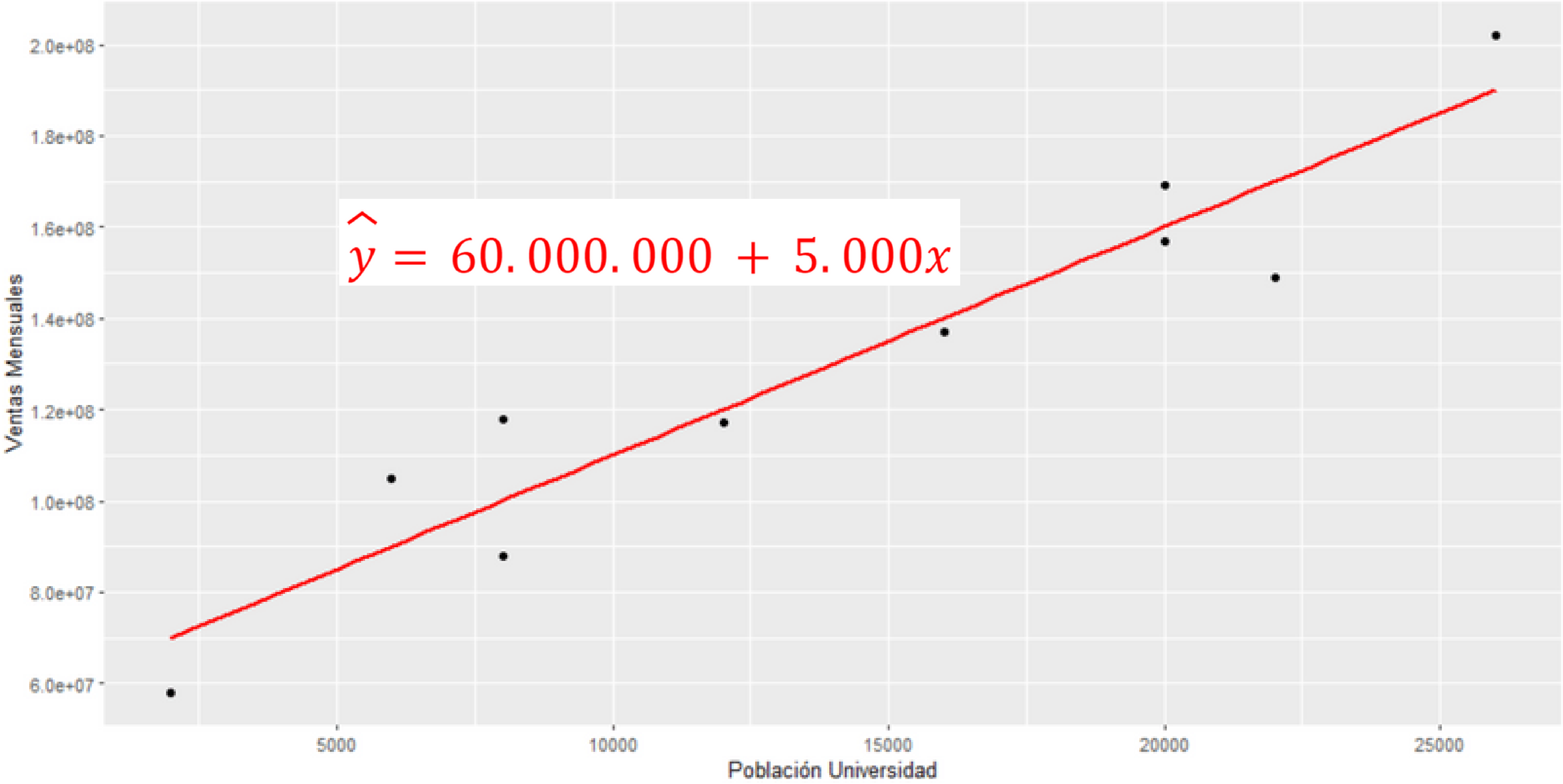
Apliquemos el Método de Mínimos Cuadrados...

- Como la pendiente de la ecuación es positiva podemos concluir:
  Se espera que las ventas mensuales aumenten 5.000 pesos por cada aumento de una persona en la población universitaria

- Después de estimar la ecuación de regresión podemos usarla para estimar el valor de **y** dado un valor de **x:**
  Vamos a predecir las ventas mensuales de un restaurante ubicado en un campus de 16.000 personas

$$\hat{y} = 60.000.000 + 5.000(16) = 140.000$$

Apliquemos el Método de Mínimos Cuadrados…



Apliquemos el Método de Mínimos Cuadrados en R...

¡ALERTA!

Debemos tener cuidado al hacer predicciones por fuera del rango de valores de **x** ya que por fuera de este rango no puede asegurarse que esta relación sea válida

Coeficiente de determinación (1/8)

¿Qué tan bien se ajusta a los datos la ecuación de regresión estimada?

$$r^2 = \frac{SCR}{STC} \tag{7}$$

Coeficiente de determinación (2/8)

RESIDUAL: la diferencia que existe, para la observación i, entre el valor observado y_i y el valor estimado $\hat{y}_i$ de la variable independiente

$$residual_i = y_i - \hat{y}_i \qquad \textbf{(8)}$$

Representa el error que existe al usar $\hat{y}_i$ para estimar y_i

Coeficiente de determinación (3/8)

SUMA DE CUADRADOS DEBIDA AL ERROR (SCE)

$$SCE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \quad \textbf{(9)}$$

Medida del error al utilizar la ecuación de regresión estimada para estimar los valores de y

Coeficiente de determinación (4/8)

SUMA TOTAL DE CUADRADOS (STC)

$$STC = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2 \qquad (10)$$

Medida del error al utilizar el promedio de y para estimar los valores de y

Coeficiente de determinación (5/8)

SUMA DE CUADRADOS DEBIDA A LA REGRESIÓN (SCR)

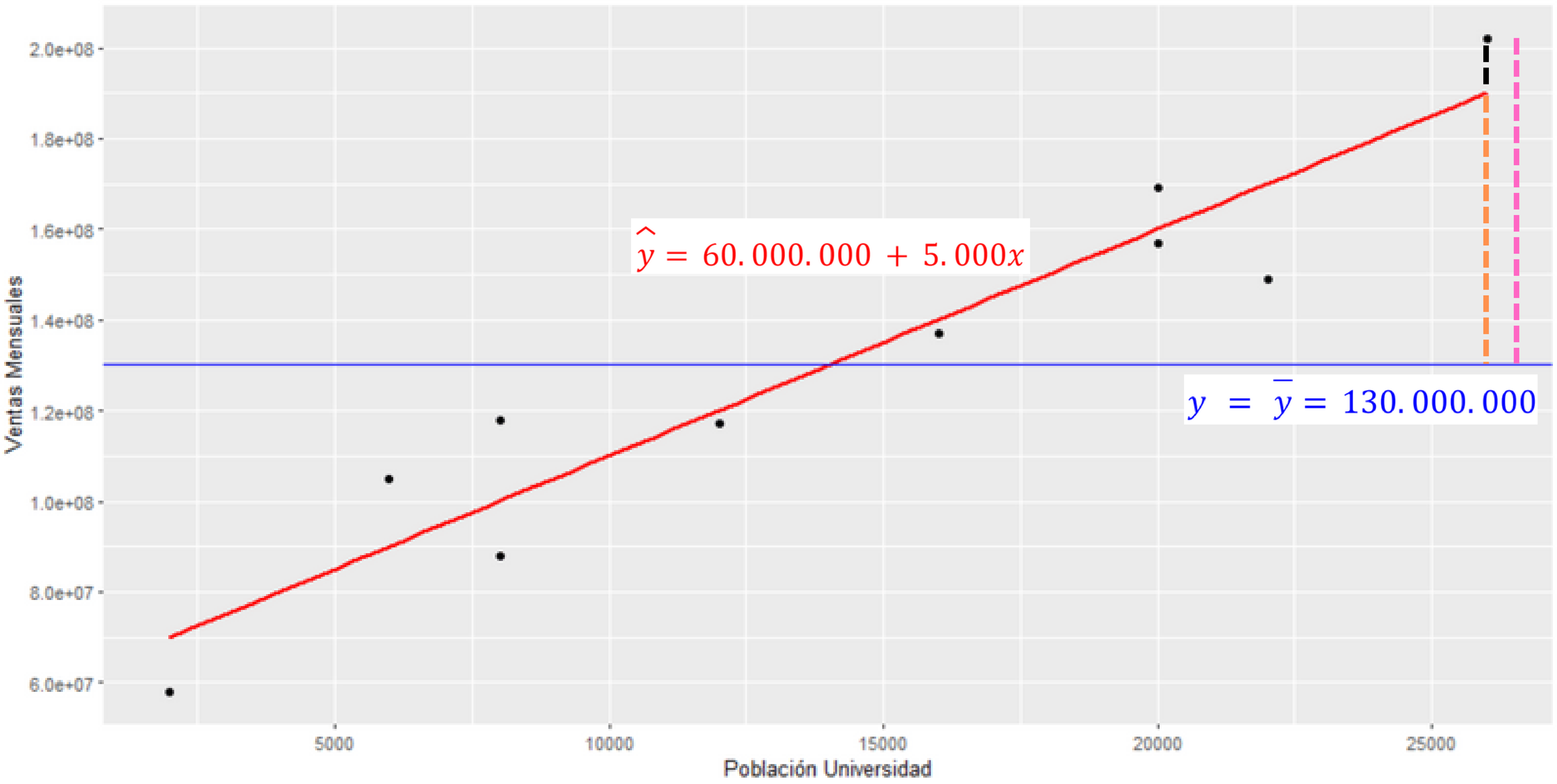
$$SCR = \sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}\right)^2 \qquad \textbf{(11)}$$

Medida de que tanto se alejan de $\bar{y}$ los valores estimados

Coeficiente de determinación (6/8)

$$STC = SCR + SCE$$

Coeficiente de determinación (7/8)



$y_{10} - \hat{y}_{10}$

$\hat{y}_{10} - \bar{y}$

$y_{10} - \bar{y}$

Coeficiente de determinación (8/8)

El ajuste perfecto se logra cuando:

$$SCE = 0 \rightarrow STC = SCR \rightarrow r^2 = \frac{SCR}{STC} = 1$$

Coeficiente de correlación

Medida de la intensidad de la relación lineal entre 2 variables **x** y **y**

$$\gamma_{xy} = (signo\ de\ b_1)\sqrt{r^2}$$

$\gamma_{xy} = 0$ — **x, y** NO están relacionadas linealmente

$\gamma_{xy} < 0$ — **x, y** tienen una relación lineal negativa

$\gamma_{xy} > 0$ — **x, y** tienen una relación lineal positiva

Calculemos estos coeficientes en R...

Interpretación de los coeficientes

Coeficiente de determinación:

90.27% de la variabilidad en las ventas se explica por la relación lineal que existe entre el tamaño de la población universitaria y las ventas.

Coeficiente de correlación:

Como el coeficiente de correlación es igual a 0.9501, se puede concluir que existe una relación lineal fuerte entre la población universitaria y las ventas. Además, esta relación es positiva.

Actividades evaluativas

- **Proyecto #1:** Introducción al software R **(20%) -> 22 mayo**
- **Proyecto #2:** Modelado Estadístico **(20%) -> 8 junio**
- **Proyecto #3:** Diseño de experimentos **(20%) -> 26 julio**
- **Proyecto #4:** Análisis multivariado **(20%) -> 18 agosto**
- **Proyecto #5:** Pronóstico **(20%) -> 8 septiembre**