Curso Análisis de datos

31 de mayo de 2022

Regresión Logística

Módulo 2

—

Modelado estadístico

Agenda de hoy

1. Regresión Logística
2. Estimación de la ecuación de regresión
3. Prueba de significancia
4. Interpretación de la ecuación de regresión
5. Transformación Logit

Regresión Logística (1/)

Hasta ahora hemos utilizado modelos de regresión en los que la variable dependiente es continua: ventas mensuales, número de homicidios, ...

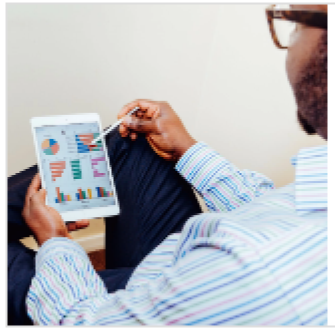
¿Qué pasa en el caso en el que la variable dependiente es discreta?
(género de una persona, cliente paga o no paga el próximo mes, email es spam o no, banco aprueba tarjeta de crédito o no, candidato por el que Sara va a votar en segunda vuelta)

Regresión Logística (1/)

La Regresión Logística nos permite, dado un conjunto particular de valores de las variables independientes, estimar la probabilidad de pertenencia a cada categoría de la variable dependiente

Regresión Logística (1/)

Utilicemos una base de datos que contiene información de empleados de una compañía para predecir cuál es la probabilidad de que estos empleados renuncien



IBM HR Analytics Employee Attrition & Performance

Predict attrition of your valuable employees

kaggle.com

Modelo de Regresión Múltiple (3/4)

El Análisis de Regresión Múltiple estudia la relación de una variable dependiente con dos o más variables independientes

Modelo de Regresión Múltiple

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon \quad \textbf{(1)}$$

Variable conocida como término del error

Modelo de Regresión Múltiple (4/4)

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

Ecuación de Regresión Múltiple Estimada

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p \quad (2)$$

Método de Mínimos Cuadrados (1/7)

¿Cómo funciona el Método de Mínimos Cuadrados?

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

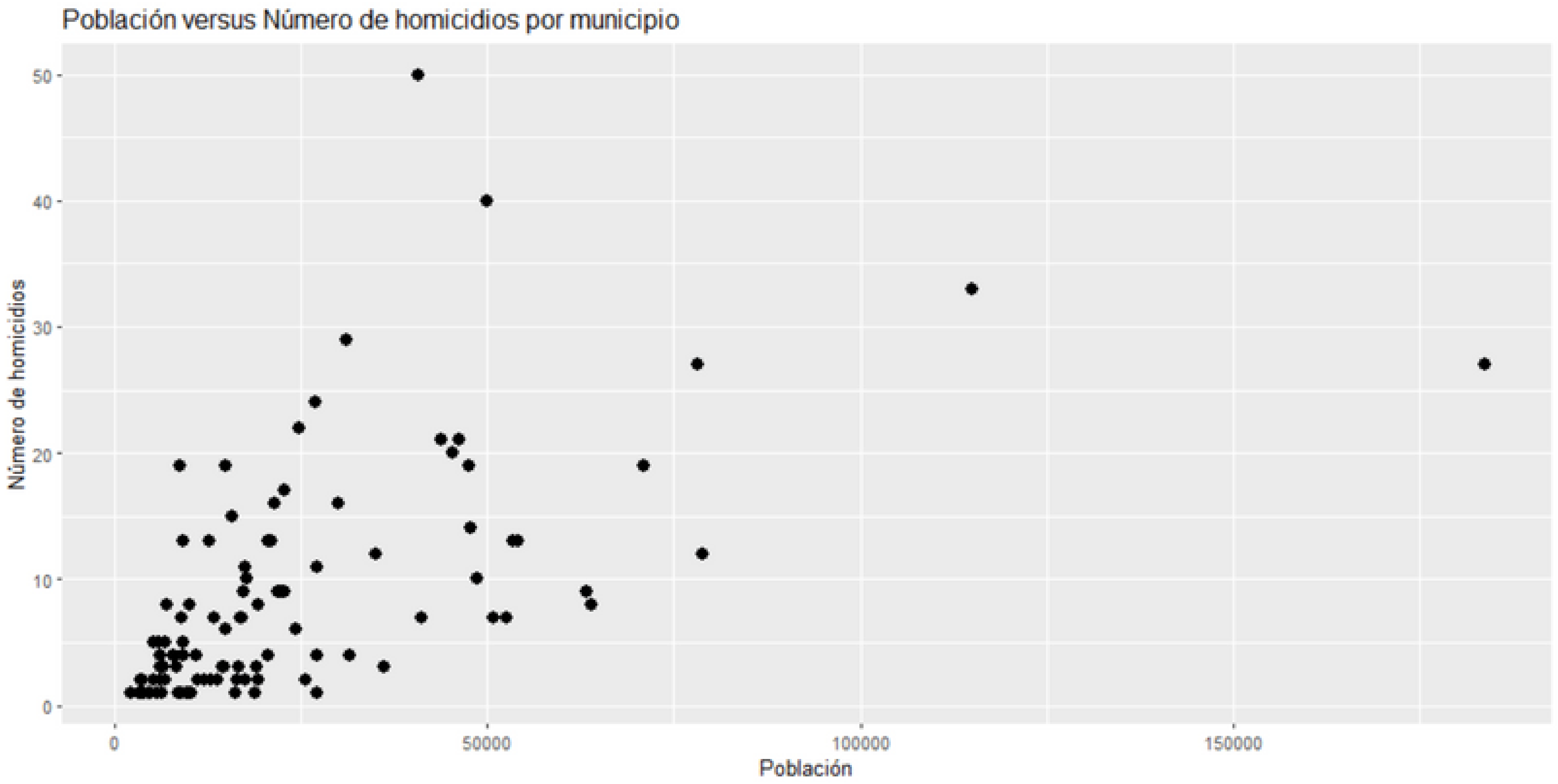
$$min \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \qquad (3)$$

Consideremos otro ejemplo...

Region	Municipio	Poblacion	Robos	Accidentes_trafico	Homicidios	Desertores_escolares	Escenarios_deportivos	Extorsiones	Lesiones_personales
Valle de Aburra	Barbosa	50836	115	19	7	105	63	4	74
Valle de Aburra	Caldas	78756	141	12	12	276	66	1	85
Valle de Aburra	Copacabana	71035	385	18	19	254	81	5	143
Valle de Aburra	Girardota	55.49	174	15	8	129	53	4	76
Valle de Aburra	La Estrella	63335	264	6	9	172	76	6	68
Valle de Aburra	Sabaneta	52554	506	12	7	225	54	4	132
Bajo Cauca	Caceres	38.85	17	10	22	228	38	1	16
Bajo Cauca	Caucasia	114902	212	22	33	761	93	9	132
Bajo Cauca	El Bagre	49913	17	4	40	515	79	2	44
Bajo Cauca	Nechi	27238	5	1	1	332	42	1	9
Bajo Cauca	Taraza	43856	28	14	21	453	46	4	20
Bajo Cauca	Zaragoza	31129	11	3	29	367	41	2	16
Magdalena Medio	Caracoli	4569	8	1	1	30	37	1	4
Magdalena Medio	Maceo	6775	10	4	2	54	34	1	17

¿Qué variables podrían influir en número de homicidios que ocurren en un municipio?

Grafiquemos los datos muestrales...

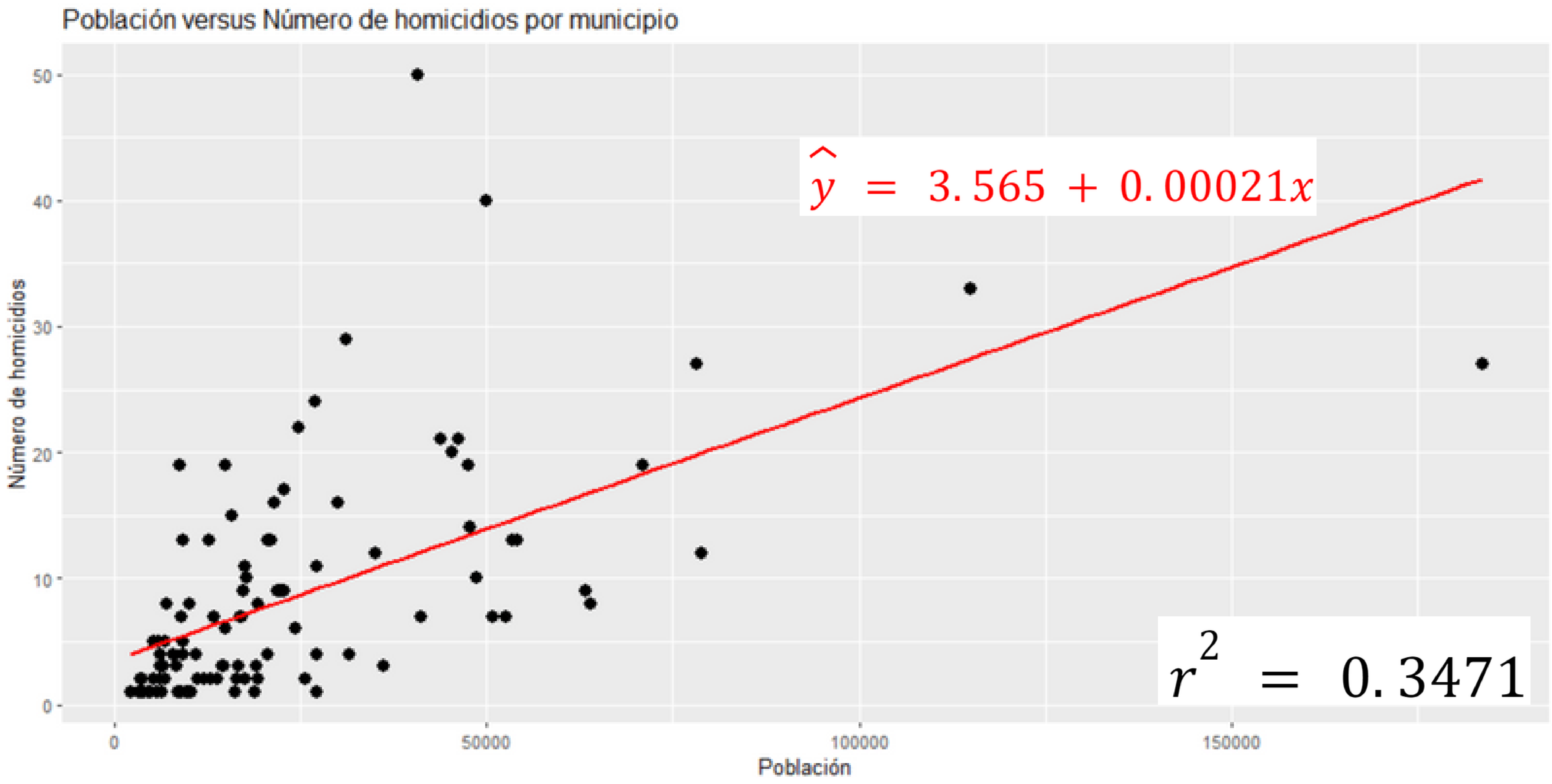


Se puede observar que en general, a medida que aumenta la población del municipio, aumenta el número de homicidios

$$\hat{y} = b_0 + b_1 x$$

Apliquemos el Método de Mínimos Cuadrados en R...

Apliquemos el Método de Mínimos Cuadrados...



Interpretación de parámetros y estimación

Como la pendiente de la ecuación es positiva podemos concluir:
Se espera que el número de homicidios aumente en 0.00021 unidades por cada aumento de una persona en la población

Después de estimar la ecuación de regresión podemos usarla para estimar el valor de **y** dado un valor de **x:**
Vamos a estimar el número de homicidios en un municipio con una población de 50.836 habitantes

$$\hat{y} = 3.565 + 0.00021(50836) = 14.11$$

Agreguemos una nueva variable...

$$\boxed{\hat{y} = b_0 + b_1 x_1 + b_2 x_2}$$

$$x_1 = Población$$

$$x_2 = Desetores\ escolares$$

Apliquemos el Método de Mínimos Cuadrados en R...

Ecuación de Regresión Estimada...

$$\hat{y} = 2.930 - 0.00002\,x_1 + 0.0434\,x_2$$

$$\boxed{r^2 = 0.4566}$$

Interpretación de parámetros

b_i representa la estimación del cambio en y debido a un cambio en una unidad en x_i mientras todas las demás variables independientes permanecen constantes

- $b_1 = -0.00002$

  Se espera que el número de homicidios disminuya en -0.00002 por cada aumento de una persona en la población cuando el número de desertores escolares permanece constante

- $b_2 = 0.0434$

  Se espera que el número de homicidios aumente en 0.0434 por cada aumento de una persona en el número de desertores escolares cuando el número habitantes permanece constante