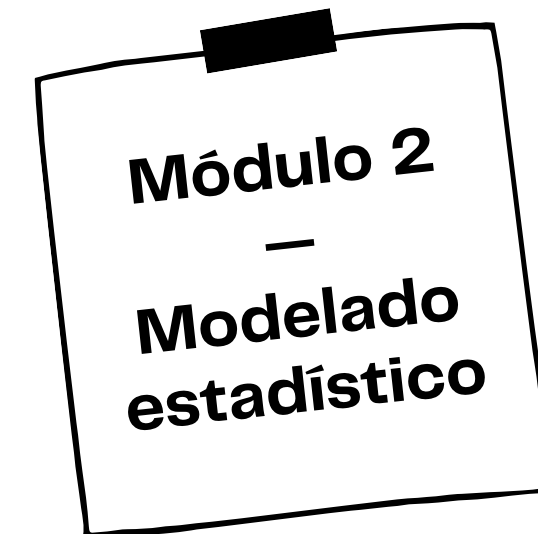


Regresión Lineal Múltiple



Q Agenda de hoy

- | | | | |
|---|---------------------------------------|---|---------------------------------------|
| 1 | Modelo de Regresión Múltiple | 5 | Pruebas de significancia |
| 2 | El método de Mínimos Cuadrados | 6 | Multicolinealidad |
| 3 | Coeficiente de Determinación Múltiple | 7 | Estimación puntual y por intervalo |
| 4 | Suposiciones del modelo | 8 | Variables independientes cualitativas |

Modelo de Regresión Múltiple (1/4)



**Cadena de
restaurantes de pizza**

**Campus
Universitarios**

¿Qué variables podrían influir en mis ventas mensuales?

Modelo de Regresión Múltiple (2/4)



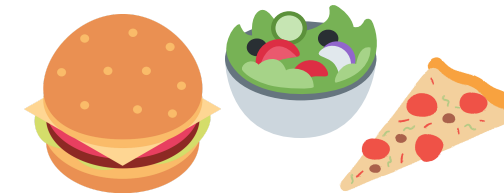
Ventas Mensuales



Población del campus



**Capacidad adquisitiva
de la población**



Comida favorita

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


$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon \quad (1)$$

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

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


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

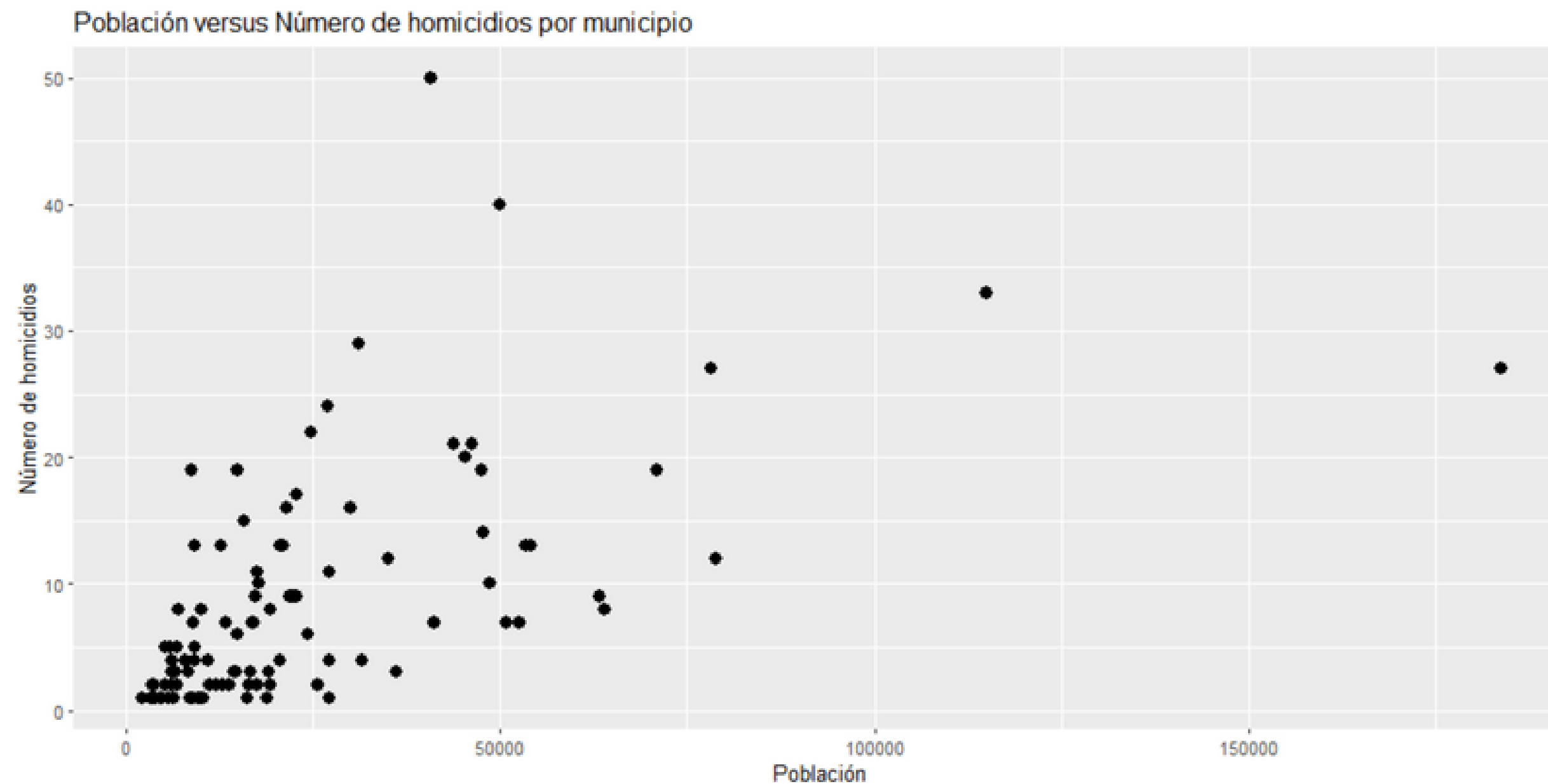
$$\min \sum_{i=1}^n \left(y_i - \hat{y}_i \right)^2 \quad (3)$$

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...

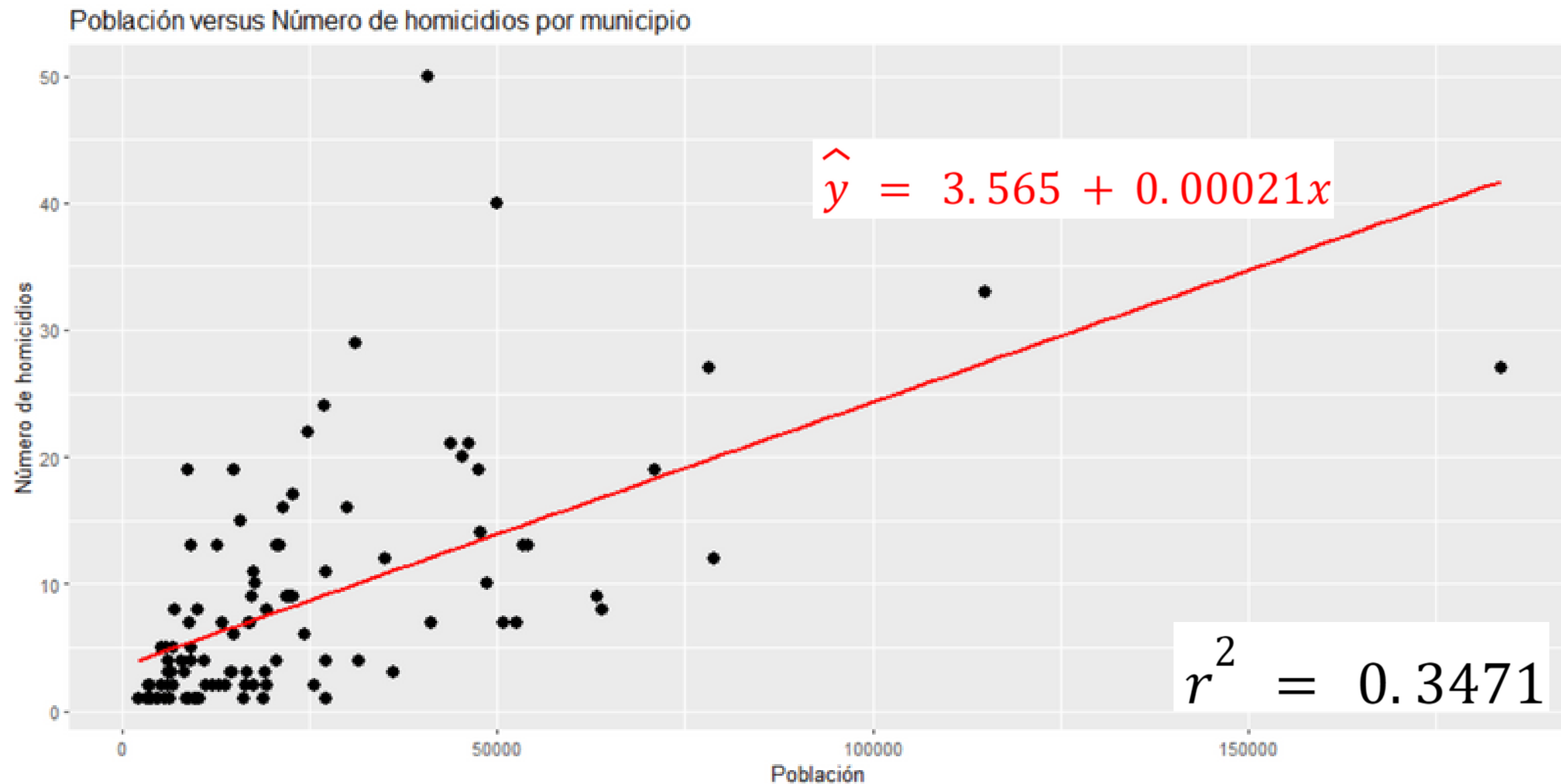


Se puede observar que en general, a medida que aumenta la población del municipio, aumenta el número de homicidios

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

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...

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



$x_1 = \text{Población}$

$x_2 = \text{Desetores escolares}$

Apliquemos el Método de Mínimos Cuadrados en R...

Ecuación de Regresión Estimada...

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

$$r^2 = 0.4566$$

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

Coeficiente de determinación múltiple (1/4)


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

$$r^2 = \frac{SCR}{STC} \quad (4)$$

Coeficiente de determinación múltiple (2/4)


$$SCE = \sum_{i=1}^n \left(y_i - \hat{y}_i \right)^2 \quad (6)$$

$$(5) \quad STC = SCR + SCE \quad \rightarrow \quad STC = \sum_{i=1}^n \left(y_i - \bar{y} \right)^2 \quad (7)$$


$$SCR = \sum_{i=1}^n \left(\hat{y}_i - \bar{y} \right)^2 \quad (8)$$

Coeficiente de determinación múltiple (3/4)

Interpretación del coeficiente

40.66% de la variabilidad en el número de homicidios es explicada por la ecuación de regresión estimada en la que las variables independientes son número de habitantes (población) y número de desertores escolares.

Coeficiente de determinación múltiple (4/4)

Coeficiente de determinación múltiple ajustado

Cuando se agrega una variable nueva al modelo, el coeficiente de determinación se vuelve más grande, incluso cuando esta variable no es estadísticamente significativa. El **coeficiente de determinación múltiple ajustado** compensa el número de variables independientes en el modelo

$$R_a^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1} \quad (9)$$

Suposiciones del modelo

- $E(\epsilon) = 0$
- La varianza de ϵ que se denota σ^2 es la misma para todos los valores de las variables independientes x_1, x_2, \dots, x_p
- Los valores de ϵ son independientes.
- $\epsilon \sim N(0, \sigma^2) \rightarrow y \sim N(\mu, \sigma^2)$

Prueba de Significancia (1/7)

Prueba F

La prueba F se usa para determinar si existe una relación de significancia entre la variable dependiente y las variables independientes
(Prueba de significancia global)

Prueba t

Si la prueba F indica que hay significancia global, posteriormente se usa la prueba t para determinar si cada una de las variables independientes es significativa
(Prueba de significancia individual)


Prueba de Significancia (2/7)

Prueba F

1 Definir las hipótesis

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$$

H_a = uno o más de los parámetros es distinto de 0



Si se rechaza la hipótesis nula, hay suficiente evidencia estadística para concluir que uno o más de los parámetros es diferente de cero y que **la relación global entre la variable dependiente y las variables independientes es significativa**

Prueba de Significancia (3/7)

Prueba F

② Definir el estadístico de prueba

$$F = \frac{CMR}{CME}$$



$$CMR = \frac{SCR}{p}$$



$$CME = \frac{SCE}{n - p - 1}$$

Prueba de Significancia (4/7)

Prueba F

③ Defino la regla de rechazo

Valor P

Rechazo H_0 si:

$$valor\ p \leq \alpha$$



Siempre que se realiza una prueba de hipótesis se debe seleccionar un nivel de significancia α

Prueba de Significancia (5/7)

Prueba t

1 Definir las hipótesis

$$H_0: \beta_i = 0$$

$$H_a: \beta_i \neq 0$$



Si se rechaza la hipótesis nula, se concluirá que el parámetro es estadísticamente significativo

Prueba de Significancia (6/7)

Prueba t

- 2 Definir el estadístico de prueba

$$t = \frac{b_i}{S_{b_i}}$$

Prueba de Significancia (7/7)

Prueba t

③ Defino la regla de rechazo

Valor P

Rechazo H_0 si:

$$valor\ p \leq \alpha$$



Siempre que se realiza una prueba de hipótesis se debe seleccionar un nivel de significancia α

Multicolinealidad (1/2)

El término variable independiente se usa para referirse a cualquier variable que se usa para explicar el valor de la variable dependiente



El término variable independiente no implica que las variables independientes sean independientes entre ellas

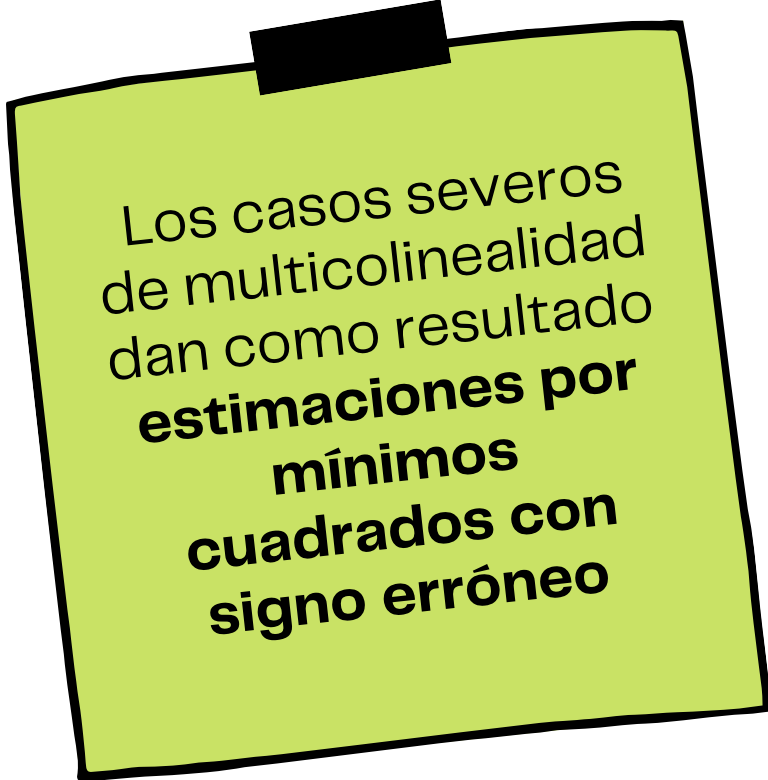


En un problema de regresión múltiple, la mayoría de variables independientes están correlacionadas unas con otras

Multicolinealidad (2/2)

La multicolinealidad es un problema potencial si el valor absoluto del coeficiente de correlación muestral es **mayor a 0.7** para un par de variables independientes

Siempre que sea posible se debe evitar incluir en el modelo variables independientes que estén fuertemente correlacionadas



Los casos severos de multicolinealidad dan como resultado **estimaciones por mínimos cuadrados con signo erróneo**

**Veamos si tenemos problemas de
multicolinealidad en nuestro modelo...**

Uso de la ecuación de regresión para estimaciones (1/2)

Estimación puntual

Estimar el número de homicidios en municipios con una población de 30.000 habitantes y un número de desertores escolares igual a 450

$$\hat{y} = 2.930 - 0.00002 (30.000) + 0.0434 (450) = 21.84$$

Uso de la ecuación de regresión para estimaciones (2/2)

Estimación por intervalo

Construir un intervalo del 95% confianza para la media del número de homicidios en municipios con una población de 30.000 habitantes y un número de desertores escolares igual a 450

(16.35, 27.33)

Variables independientes cualitativas (1/11)

Las variables independientes pueden ser cuantitativas (población, número de desertores escolares) o cualitativas (género, estrato, región)

Cuando se agrega una variable cualitativa que tenga **k** niveles al modelo, se deben crear **k-1** variables ficticias o indicadoras

Variables independientes cualitativas (2/11)

Estudiante i	Género
1	Mujer
2	Mujer
3	Hombre
4	Hombre
5	Mujer
6	Hombre
7	Mujer
8	Hombre
9	Hombre
10	Mujer
11	Hombre




Estudiante i	Género
1	1
2	1
3	0
4	0
5	1
6	0
7	1
8	0
9	0
10	1
11	0

Variables independientes cualitativas (3/11)

Estudiante i	Género
1	1
2	1
3	0
4	0
5	1
6	0
7	1
8	0
9	0
10	1
11	0

$$x_1 = \begin{cases} 1 & \text{si el género del estudiante es Mujer} \\ 0 & \text{si el género del estudiante es Hombre} \end{cases}$$

Variable indicadora


$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

Variables independientes cualitativas (4/11)

Incorporemos a nuestro modelo de regresión múltiple la variable cualitativa **Región...**



Como la variable Región tiene **9 niveles**, debemos crear **8 variables ficticias** o indicadoras

Variables independientes cualitativas (5/11)

$$x_1 = \begin{cases} 1 & \text{si la región es Valle de Aburra} \\ 0 & \text{si no es así} \end{cases}$$

$$x_5 = \begin{cases} 1 & \text{si la región es Norte} \\ 0 & \text{si no es así} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{si la región es Uraba} \\ 0 & \text{si no es así} \end{cases}$$

$$x_6 = \begin{cases} 1 & \text{si la región es Occidente} \\ 0 & \text{si no es así} \end{cases}$$

$$x_3 = \begin{cases} 1 & \text{si la región es Magdalena Medio} \\ 0 & \text{si no es así} \end{cases}$$

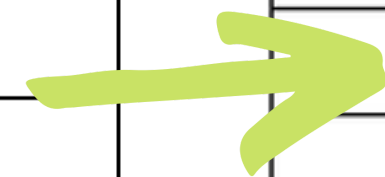
$$x_7 = \begin{cases} 1 & \text{si la región es Oriente} \\ 0 & \text{si no es así} \end{cases}$$

$$x_4 = \begin{cases} 1 & \text{si la región es Nordeste} \\ 0 & \text{si no es así} \end{cases}$$

$$x_8 = \begin{cases} 1 & \text{si la región es Suroeste} \\ 0 & \text{si no es así} \end{cases}$$

Variables independientes cualitativas (6/11)

Municipio i	Desertores escolares	Región
1	105	Valle de Aburra
2	761	Bajo Cauca
3	175	Nordeste
4	30	Magdalena Medio
5	276	Valle de Aburra
6	70	Urabá
7	15	Occidente
8	25	Norte
9	35	Suroeste
10	89	Oriente
11	22	Oriente
12	515	Bajo Cauca



Municipio i	Desertores escolares	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
1	105	1	0	0	0	0	0	0	0
2	761	0	0	0	0	0	0	0	0
3	175	0	0	0	1	0	0	0	0
4	30	0	0	1	0	0	0	0	0
5	276	1	0	0	0	0	0	0	0
6	70	0	1	0	0	0	0	0	0
7	15	0	0	0	0	0	1	0	0
8	25	0	0	0	0	1	0	0	0
9	35	0	0	0	0	0	0	0	1
10	89	0	0	0	0	0	0	1	0
11	22	0	0	0	0	0	0	1	0
12	515	0	0	0	0	0	0	0	0

Variables independientes cualitativas (7/11)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \epsilon$$

$x_9 = \text{Número de desertores escolares}$

La categoría de referencia en este caso es Bajo Cauca

Ajustemos este modelo de regresión múltiple usando R...

Variables independientes cualitativas (8/11)

$$\hat{y} = 1.93 - 0.85 x_1 - 7.46 x_2 - 3.31 x_3 + 3.09 x_4 + 0.61 x_5 \\ + 1.11 x_6 - 0.90 x_7 + 2.6 x_8 + 0.04 x_9$$

Variables independientes cualitativas (9/11)

Interpretación de parámetros: Variables cuantitativas




$$b_9 = 0.04$$

Se espera que el número de homicidios aumente en 0.04 por cada aumento de una persona por cada aumento de una persona en el número de desertores escolares cuando las demás variables permanecen constantes

Variables independientes cualitativas (10/11)

Interpretación de parámetros: Variables cualitativas

b_i es la diferencia entre la media del número de homicidios en la región i y la media del número de homicidios en la región de referencia (Bajo Cauca)



La interpretación de las variables cualitativas siempre se hace con **respecto a la categoría de referencia**

Variables independientes cualitativas (11/11)

Interpretación de parámetros: Variables cualitativas



$$b_1 = -0.85$$

- La diferencia entre la media del número de homicidios en un municipio de la Región del Valle de Aburra y en un municipio de la región del Bajo Cauca es de -0.85
- **Los municipios pertenecientes a la región del Valle de Aburra presentan un promedio de homicidios menor que los municipios pertenecientes a la región del Bajo Cauca**

Análisis residual

Es una herramienta que nos permite determinar si el modelo de regresión múltiple es apropiado **confirmando las suposiciones del modelo**

Si se encuentra que una o más de las suposiciones son dudosas, habrá que considerar otro modelo o una transformación de los datos