

Tarea 1

NOMBRE: JESUS ALEXIS SANCHEZ MORENO

MATRÍCULA: 224470329

2.1 Consider the following five observations. You are to do all the parts of this exercise using only a calculator.

$x$	$y$	$x - \bar{x}$	$(x - \bar{x})^2$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
3	4				
2	2				
1	3				
-1	1				
0	0				
$\sum x_i =$	$\sum y_i =$	$\sum (x_i - \bar{x}) =$	$\sum (x_i - \bar{x})^2 =$	$\sum (y_i - \bar{y}) =$	$\sum (x_i - \bar{x})(y_i - \bar{y}) =$

- a. Complete the entries in the table. Put the sums in the last row. What are the sample means  $\bar{x}$  and  $\bar{y}$ ?
- b. Calculate  $b_1$  and  $b_2$  using (2.7) and (2.8) and state their interpretation.
- c. Compute  $\sum_{i=1}^5 x_i^2$ ,  $\sum_{i=1}^5 x_i y_i$ . Using these numerical values, show that  $\sum (x_i - \bar{x})^2 = \sum x_i^2 - N\bar{x}^2$  and  $\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - N\bar{x}\bar{y}$ .
- d. Use the least squares estimates from part (b) to compute the fitted values of  $y$ , and complete the remainder of the table below. Put the sums in the last row.  
Calculate the sample variance of  $y$ ,  $s_y^2 = \sum_{i=1}^N (y_i - \bar{y})^2 / (N - 1)$ , the sample variance of  $x$ ,  $s_x^2 = \sum_{i=1}^N (x_i - \bar{x})^2 / (N - 1)$ , the sample covariance between  $x$  and  $y$ ,  $s_{xy} = \sum_{i=1}^N (y_i - \bar{y})(x_i - \bar{x}) / (N - 1)$ , the sample correlation between  $x$  and  $y$ ,  $r_{xy} = s_{xy} / (s_x s_y)$  and the coefficient of variation of  $x$ ,  $CV_x = 100(s_x / \bar{x})$ . What is the median, 50th percentile, of  $x$ ?

$x_i$	$y_i$	$\hat{y}_i$	$\hat{e}_i$	$\hat{e}_i^2$	$x_i \hat{e}_i$
3	4				
2	2				
1	3				
-1	1				
0	0				
$\sum x_i =$	$\sum y_i =$	$\sum \hat{y}_i =$	$\sum \hat{e}_i =$	$\sum \hat{e}_i^2 =$	$\sum x_i \hat{e}_i =$

- e. On graph paper, plot the data points and sketch the fitted regression line  $\hat{y}_i = b_1 + b_2 x_i$ .
- f. On the sketch in part (e), locate the point of the means  $(\bar{x}, \bar{y})$ . Does your fitted line pass through that point? If not, go back to the drawing board, literally.
- g. Show that for these numerical values  $\bar{y} = b_1 + b_2 \bar{x}$ .
- h. Show that for these numerical values  $\bar{y} = \bar{y}$ , where  $\bar{y} = \sum \hat{y}_i / N$ .
- i. Compute  $\hat{\sigma}^2$ .
- j. Compute  $\widehat{\text{var}}(b_7 | \mathbf{x})$  and  $\text{se}(b_7)$ .

2.1 EJERCICIOS

a. Medias muestrales y tabla completada:

$$\bar{x} = \frac{3 + 2 + 1 + (-1) + 0}{5} = 1$$

$$\bar{y} = \frac{4 + 2 + 3 + 1 + 0}{5} = 2$$

$x$	$y$	$x - \bar{x}$	$(x - \bar{x})^2$	$y - \bar{y}$	$(x - \bar{x})(y - \bar{y})$
3	4	2	4	2	4
2	2	1	1	0	0
1	3	0	0	1	0
-1	1	-2	4	-1	2
0	0	-1	1	-2	2
$\sum = 5$	$\sum = 10$	$\sum = 0$	$\sum = 10$	$\sum = 0$	$\sum = 8$

b. Coeficientes de regresión:

$$b_2 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \frac{8}{10} = 0.8$$

$$b_1 = \bar{y} - b_2 \bar{x} = 2 - 0.8(1) = 1.2$$

Interpretación:

- $b_0 = 1.2$ : Valor esperado de  $y$  cuando  $x = 0$
  - $b_1 = 0.8$ : Por cada unidad que aumenta  $x$ ,  $y$  aumenta 0.8 unidades

c. Comparación de sumas:

$$\sum x_i^2 = 15$$

$$\sum x_i y_i = 18$$

$$\sum x_i^2 - N\overline{x}^2 = 15 - 5(1)^2 = 10$$

$$\sum x_i y_i - N\overline{x}\overline{y} = 8$$

**d. Valores ajustados y estadísticas:**

$x$	$y$	$\hat{y}$	$e_i$	$e_i^2$
3	4	3.6	0.4	0.16
2	2	2.8	-0.8	0.64
1	3	2.0	1.0	1.00
-1	1	0.4	0.6	0.36
0	0	1.2	-1.2	1.44
$\sum y = 10$		$\sum \hat{y} = 10$	$\sum e_i = 0$	$\sum e_i^2 = 3.6$

$$s_y^2 = \frac{10}{4} = 2.5$$

$$s_x^2 = \frac{10}{4} = 2.5$$

$$s_{xy} = \frac{8}{4} = 2$$

$$r_{xy} = 0.8$$

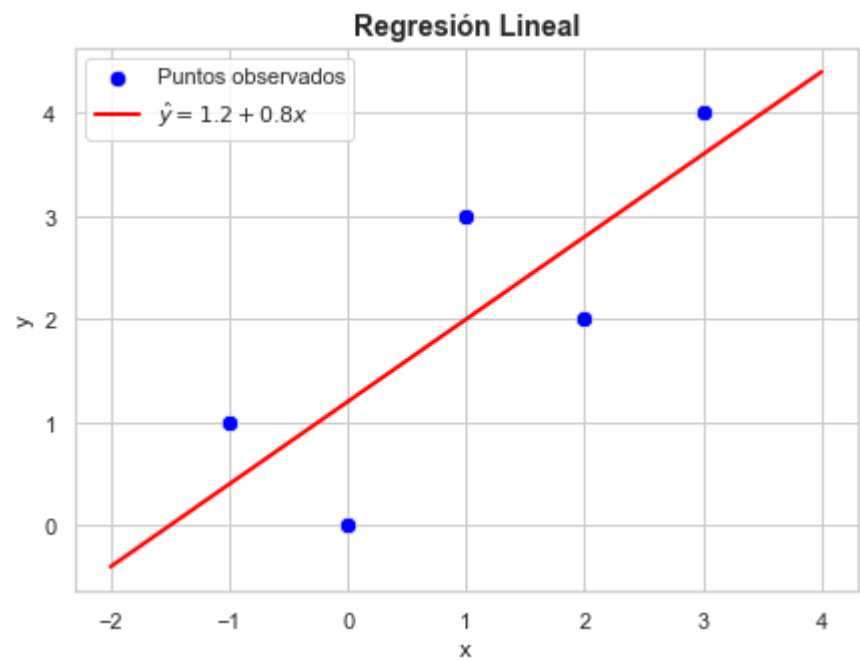
$$CV_x = 158.1\%$$

Mediana de  $x = 1$ , Percentil 50 de  $x = 1$

**e. Gráfico:**

Puntos: (3, 4), (2, 2), (1, 3), (−1, 1), (0, 0)

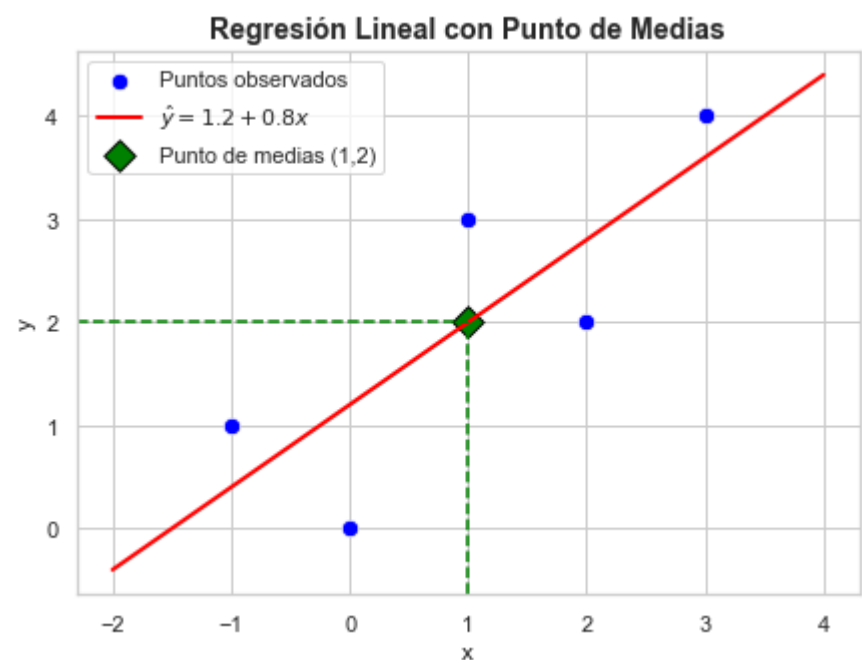
Recta:  $\hat{y} = 1.2 + 0.8x$



**f. Punto de medias:**

$$(\overline{x}, \overline{y}) = (1, 2)$$

$$\hat{y} = 1.2 + 0.8(1) = 2$$



### g. Demostración:

$$\bar{y} = b_0 + b_1\bar{x} = 1.2 + 0.8(1) = 2$$

### h. Demostración:

$$\bar{\hat{y}} = \frac{3.6 + 2.8 + 2.0 + 0.4 + 1.2}{5} = 2 = \bar{y}$$

### i. Varianza del error:

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N - 2} = \frac{3.6}{3} = 1.2$$

### j. Varianzas de los estimadores:

$$\widehat{\text{var}}(b_2) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} = \frac{1.2}{10} = 0.12$$

$$\widehat{\text{var}}(b_2) = \sqrt{0.12} = 0.36$$

**2.2** A household has weekly income of \$2000. The mean weekly expenditure for households with this income is  $E(y|x = \$2000) = \mu_{y|x=\$2000} = \$220$ , and expenditures exhibit variance  $\text{var}(y|x = \$2,000) = \sigma^2_{y|x=\$2,000} = \$121$ .

- Assuming that weekly food expenditures are normally distributed, find the probability that a household with this income spends between \$200 and \$215 on food in a week. Include a sketch with your solution.
- Find the probability that a household with this income spends more than \$250 on food in a week. Include a sketch with your solution.
- Find the probability in part (a) if the variance of weekly expenditures is  $\text{var}(y|x = \$2,000) = \sigma^2_{y|x=\$2,000} = 144$ .
- Find the probability in part (b) if the variance of weekly expenditures is  $\text{var}(y|x = \$2,000) = \sigma^2_{y|x=\$2,000} = 144$ .

## 2.2 EJERCICIOS

#### Datos:

Ingreso semanal = \$2000

Media del gasto:  $\mu = E(y|x = 2000) = \$220$

Varianza inicial:  $\sigma^2 = 121$  (desviación  $\sigma = 11$ )

### a. Probabilidad entre \$200 y \$215

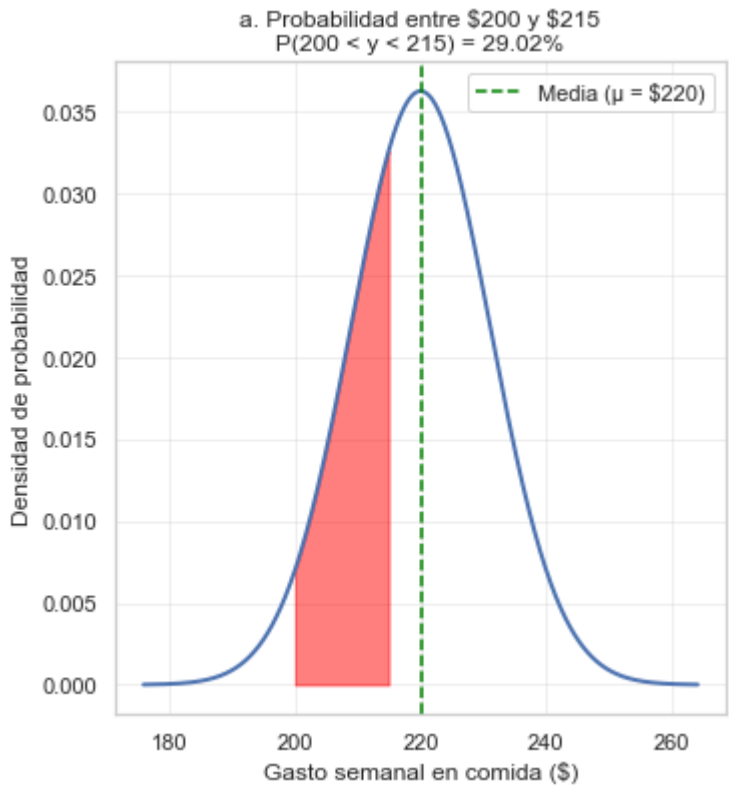
$P(200 < y < 215)$  con  $\mu = 220$ ,  $\sigma = 11$

$$z_1 = \frac{200-220}{11} = -1.818$$

$$z_2 = \frac{215-220}{11} = -0.455$$

$$P(200 < y < 215) = P(-1.818 < z < -0.455) = P(z < -0.455) - P(z < -1.818) = 0.3247 - 0.0345 = 0.2902$$

Resultado: 29.02%



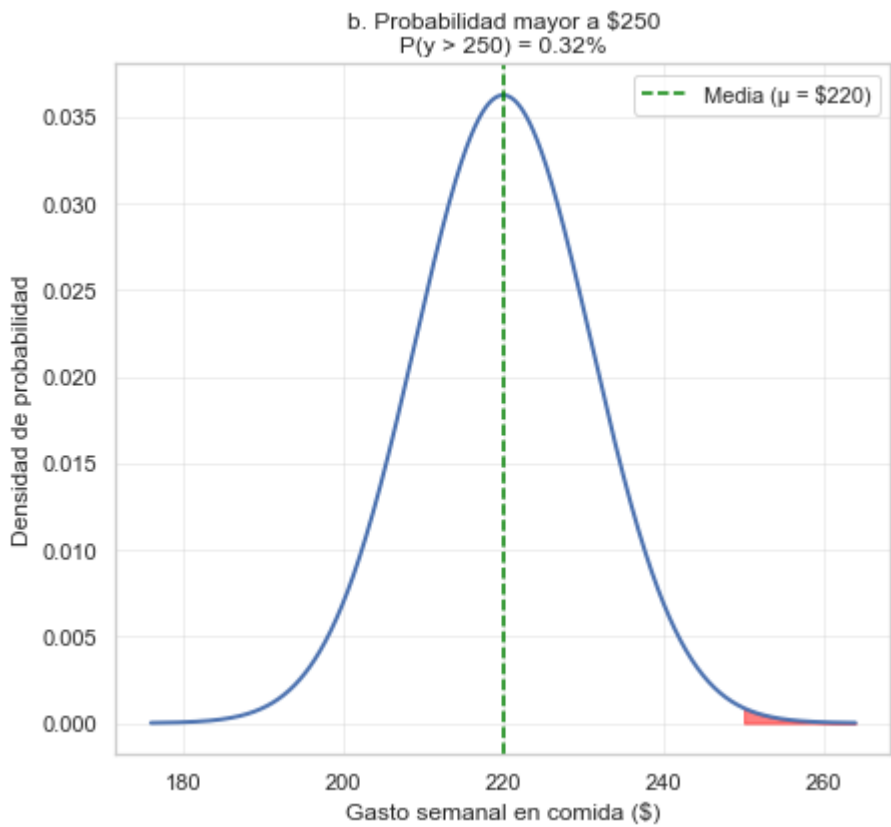
b. Probabilidad mayor a \$250

$P(y > 250)$  con  $\mu = 220$ ,  $\sigma = 11$

$$z = \frac{250-220}{11} = 2.727$$

$$P(y > 250) = P(z > 2.727) = 1 - P(z < 2.727) = 1 - 0.9968 = 0.0032$$

Resultado: 0.32%



c. Probabilidad entre \$200 y \$215 con  $\sigma^2 = 144$

Nueva desviación:  $\sigma = 12$

$$z_1 = \frac{200-220}{12} = -1.667$$

$$z_2 = \frac{215-220}{12} = -0.417$$

$$P(200 < y < 215) = P(-1.667 < z < -0.417) = P(z < -0.417) - P(z < -1.667) = 0.3383 - 0.0478 = 0.2905$$

Resultado: 29.05%

d. Probabilidad mayor a \$250 con  $\sigma^2 = 144$

$\sigma = 12$

$$z = \frac{250-220}{12} = 2.5$$

$$P(y > 250) = P(z > 2.5) = 1 - P(z < 2.5) = 1 - 0.9938 = 0.0062$$

Resultado: 0.62%



2.4 We have defined the simple linear regression model to be  $y = \beta_1 + \beta_2 x + e$ . Suppose, however, that we knew, for a fact, that  $\beta_1 = 0$ .

- What does the linear regression model look like, algebraically, if  $\beta_1 = 0$ ?
- What does the linear regression model look like, graphically, if  $\beta_1 = 0$ ?
- If  $\beta_1 = 0$ , the least squares “sum of squares” function becomes  $S(\beta_2) = \sum_{i=1}^N (y_i - \beta_2 x_i)^2$ . Using the data in Table 2.4 from Exercise 2.3, plot the value of the sum of squares function for enough values of  $\beta_2$  for you to locate the approximate minimum. What is the significance of the value of  $\beta_2$  that minimizes  $S(\beta_2)$ ? [Hint: Your computations will be simplified if you algebraically expand  $S(\beta_2) = \sum_{i=1}^N (y_i - \beta_2 x_i)^2$  by squaring the term in parentheses and carrying through the summation operator.]
- Using calculus, show that the formula for the least squares estimate of  $\beta_2$  in this model is  $b_2 = \sum x_i y_i / \sum x_i^2$ . Use this result to compute  $b_2$  and compare this value with the value you obtained geometrically.
- Using the estimate obtained with the formula in (d), plot the fitted (estimated) regression function. On the graph locate the point  $(\bar{x}, \bar{y})$ . What do you observe?
- Using the estimate obtained with the formula in (d), obtain the least squares residuals,  $\hat{e}_i = y_i - b_2 x_i$ . Find their sum.
- Calculate  $\sum x_i \hat{e}_i$ .

2.3 Graph the following observations of  $x$  and  $y$  on graph paper.

TABLE 2.4		Exercise 2.3 Data				
$x$	1	2	3	4	5	6
$y$	6	4	11	9	13	17

## 2.4 EJERCICIOS

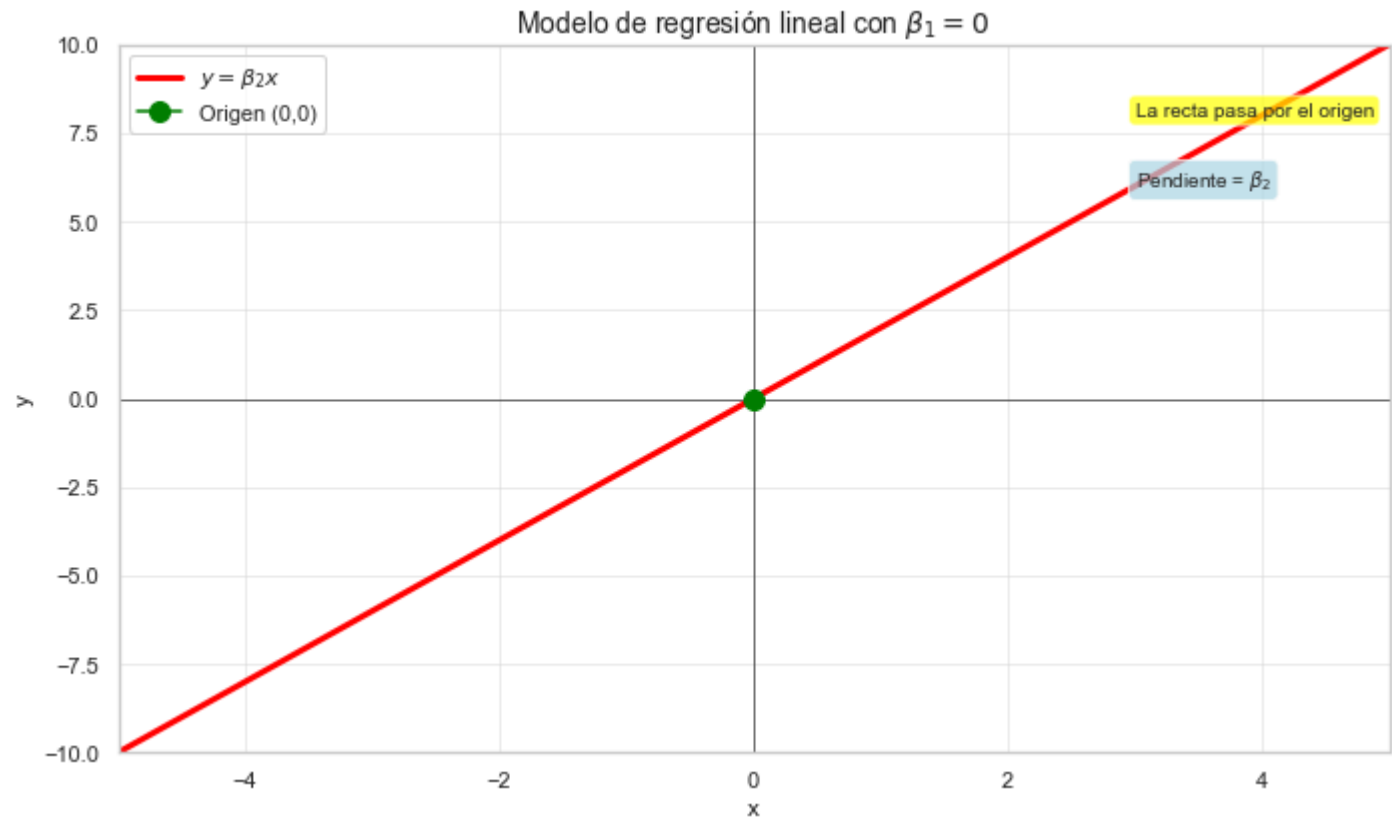
### a. Modelo algebraico:

Si  $\beta_1 = 0$ , el modelo se reduce a:

$$y = \beta_2 x + \epsilon$$

### b. Modelo gráfico:

La línea de regresión pasa por el origen  $(0, 0)$  con pendiente  $\beta_2$ .



### c. Suma de cuadrados y gráfica:

Datos de la Tabla 2.4:

$x$	1	2	3	4	5	6
$y$	6	4	11	9	13	17

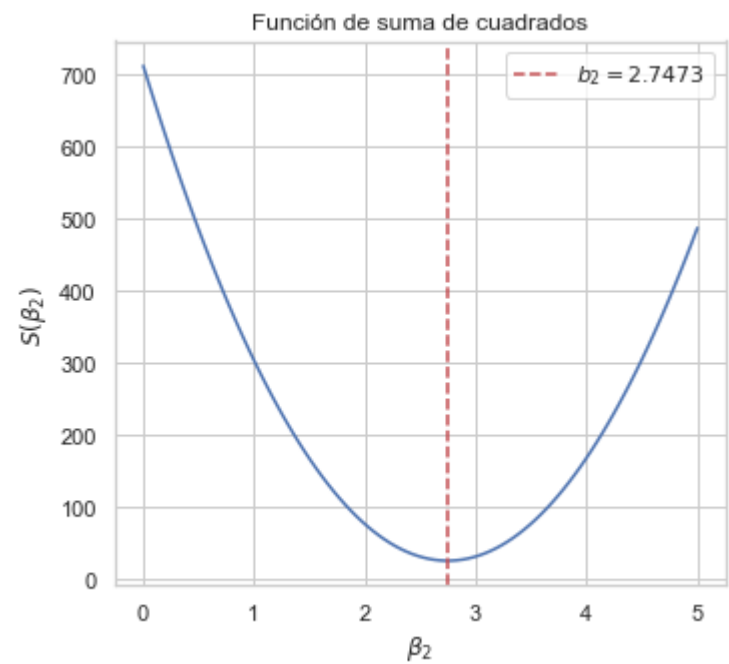
Desarrollo algebraico:

$$S(\beta_2) = \sum (y_i - \beta_2 x_i)^2 = \sum y_i^2 - 2\beta_2 \sum x_i y_i + \beta_2^2 \sum x_i^2$$

Cálculos:

- $\sum x_i = 21$
- $\sum y_i = 60$
- $\sum x_i y_i = 1 \times 6 + 2 \times 4 + 3 \times 11 + 4 \times 9 + 5 \times 13 + 6 \times 17 = 6 + 8 + 33 + 36 + 65 + 102 = 250$
- $\sum x_i^2 = 1 + 4 + 9 + 16 + 25 + 36 = 91$
- $\sum y_i^2 = 36 + 16 + 121 + 81 + 169 + 289 = 712$

$$S(\beta_2) = 712 - 2\beta_2(250) + \beta_2^2(91)$$



**Valor que minimiza:**  $\beta_2$  que minimiza  $S(\beta_2)$  es el estimador de mínimos cuadrados.

### d. Estimador por cálculo:

Sacamos la derivada de la función e igualamos a 0 para obtener el valor minimo:

$$\frac{dS(\beta_2)}{d\beta_2} = -2 \sum x_i y_i + 2\beta_2 \sum x_i^2 = 0$$

Despejando:

$$b_2 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{250}{91} = 2.7473$$

### e. Gráfica de regresión ajustada:

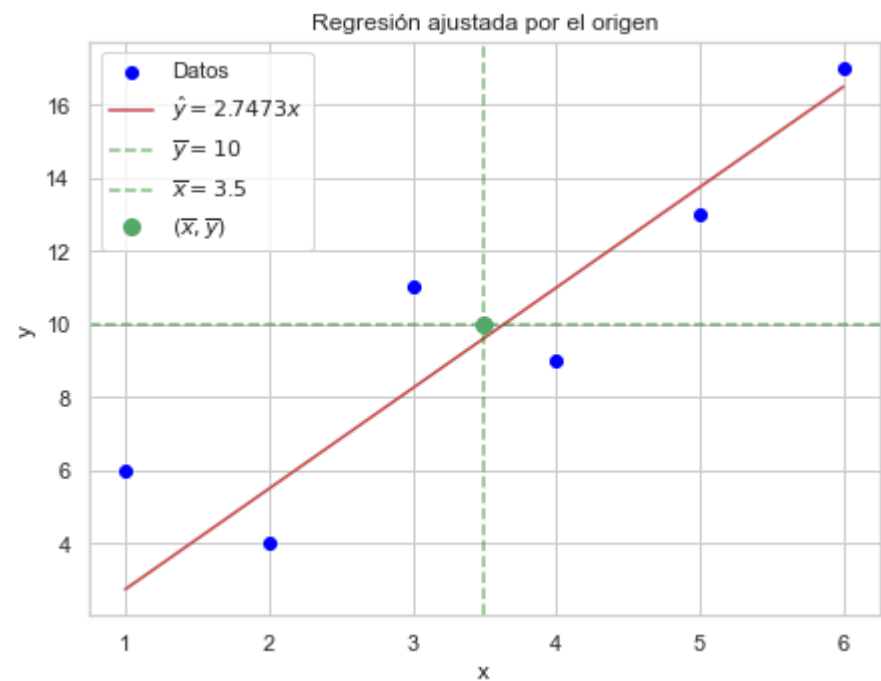
**Ecuación ajustada:**  $\hat{y} = 2.7473x$

(Dado que  $\beta_1 = 0$ )

**Punto de medias:**

$$\bar{x} = 3.5, \bar{y} = 10$$

**Observación:** La línea ajustada NO pasa por el punto de medias (3.5, 10)



### f. Residuales y su suma:

$x$	$y$	$\hat{y} = 2.7473x$	$\hat{e} = y - \hat{y}$
1	6	2.7473	3.2527
2	4	5.4946	-1.4946
3	11	8.2419	2.7581
4	9	10.9892	-1.9892
5	13	13.7365	-0.7365
6	17	16.4838	0.5162

$$\sum \hat{e}_i = 3.2527 - 1.4946 + 2.7581 - 1.9892 - 0.7365 + 0.5162 = 2.307$$

g. Suma de  $x_i\hat{e}_i$ :

$$\sum x_i\hat{e}_i = 1\times 3.2527 + 2\times (-1.4946) + 3\times 2.7581 + 4\times (-1.9892) + 5\times (-0.7365) + 6\times 0.5162 = 0$$

**Nota:**  $\sum x_i\hat{e}_i = 0$  es una propiedad de los mínimos cuadrados.

**2.7** We have 2008 data on  $y$  = income per capita (in thousands of dollars) and  $x$  = percentage of the population with a bachelor's degree or more for the 50 U.S. states plus the District of Columbia, a total of  $N = 51$  observations. We have results from a simple linear regression of  $y$  on  $x$ .

a. The estimated error variance is  $\hat{\sigma}^2 = 14.24134$ . What is the sum of squared least squares residuals?

b. The estimated variance of  $b_2$  is 0.009165. What is the standard error of  $b_2$ ? What is the value of  $\sum (x_i - \bar{x})^2$ ?

c. The estimated slope is  $b_2 = 1.02896$ . Interpret this result.

d. Using  $\bar{x} = 27.35686$  and  $\bar{y} = 39.66886$ , calculate the estimate of the intercept.

e. Given the results in (b) and (d), what is  $\sum x_i^2$ ?

f. For the state of Georgia, the value of  $y = 34.893$  and  $x = 27.5$ . Compute the least squares residual, using the information in parts (c) and (d).

2.7 EJERCICIOS

a. Suma de cuadrados de los residuales:

$\hat{\sigma}^2 = 14.24134$   
 $N = 51$

$$\sum \hat{e}_i^2 = \hat{\sigma}^2 \times (N - 2) = 14.24134 \times 49 = 697.82566$$

b. Error estándar y suma de cuadrados:

$$\widehat{\text{var}}(b_2) = 0.009165$$

$$\text{se}(b_2) = \sqrt{0.009165} = 0.09574$$

$$\widehat{\text{var}}(b_2) = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}$$

$$\sum (x_i - \bar{x})^2 = \frac{\hat{\sigma}^2}{\widehat{\text{var}}(b_2)} = \frac{14.24134}{0.009165} = 1553.72$$

c. Interpretación de la pendiente:

$b_2 = 1.02896$

**Interpretación:** Por cada punto porcentual adicional de la población con licenciatura o más, el ingreso per cápita aumenta aproximadamente \$1,029 dólares.

d. Estimación del intercepto:

$\bar{x} = 27.35686, \bar{y} = 39.66886$

$$b_1 = \bar{y} - b_2\bar{x} = 39.66886 - 1.02896 \times 27.35686 = 11.535$$

e. Cálculo de  $\sum x_i^2$ :

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - N\bar{x}^2$$

$$\begin{aligned} \sum x_i^2 &= \sum (x_i - \bar{x})^2 + N\bar{x}^2 = 1553.72 + 51 \times (27.35686)^2 \\ &= 1553.72 + 51 \times 748.396 = 1553.72 + 38168.196 = 39721.916 \end{aligned}$$

f. Residual para Georgia:

$y_{\text{GA}} = 34.893, x_{\text{GA}} = 27.5$

$\hat{y}_{\text{GA}} = b_1 + b_2x_{\text{GA}} = 11.535 + 1.02896 \times 27.5 = 39.8314$

$\hat{e}_{\text{GA}} = y_{\text{GA}} - \hat{y}_{\text{GA}} = 34.893 - 39.8314 = -4.9384$

**El estado de Georgia tiene un residual negativo de \$4,938, indicando que su ingreso per cápita es menor al predicho por el modelo.**

**2.13** Using 2011 data on 141 U.S. public research universities, we examine the relationship between academic cost per student,  $ACA$  (real total academic cost per student in thousands of dollars) and full-time enrollment  $FTESTU$  (in thousands of students).

a. The least squares fitted relation is  $\widehat{ACA} = 14.656 + 0.266FTESTU$ . What is the economic interpretation of the estimated parameters? Why isn't the intercept zero?

b. In 2011 Louisiana State University (LSU) had a full-time student enrollment of 27,950. Using the fitted related in part (a), compute the predicted value of  $ACA$ .

c. The actual value of  $ACA$  for LSU that year was 21,403. Calculate the least squares residual for LSU? Does the model overpredict or underpredict  $ACA$  for LSU?

d. The sample mean (average) full-time enrollment in U.S. public research universities in 2011 was 22,845.77. What was the sample mean of academic cost per student?

2.13 EJERCICIOS

a. Interpretación económica de los parámetros:



Ecuación estimada:  $\widetilde{ACA} = 14.656 + 0.266 \times FTESTU$

- **Intercepto** ( $b_1 = 14.656$ ): Representa el costo académico fijo por estudiante (en miles de dólares) cuando la matrícula es cero. No es cero porque existen costos fijos institucionales (infraestructura, administración) que deben cubrirse independientemente del número de estudiantes.
- **Pendiente** ( $b_2 = 0.266$ ): Indica que por cada mil estudiantes adicionales de tiempo completo, el costo académico por estudiante aumenta en \$266 dólares. Esto sugiere economías de escala limitadas o posibles costos marginales asociados con el crecimiento de la matrícula.

b. Predicción para LSU:

$FTESTU_{LSU} = 27.95$  (miles de estudiantes)

$\widetilde{ACA}_{LSU} = 14.656 + 0.266 \times 27.95 = 14.656 + 7.4347 = 22.091$

Costo académico predicho: \$22,091 por estudiante

c. Residual para LSU:

$ACA_{real} = 21.403$   
 $ACA_{predicho} = 22.091$

$\hat{e}_{LSU} = 21.403 - 22.091 = -0.688$

El modelo sobrestima el costo académico de LSU por \$688 dólares por estudiante.

d. Media muestral del costo académico:

$\overline{FTESTU} = 22.84577$  (miles de estudiantes)

$\overline{ACA} = b_0 + b_1 \times \overline{FTESTU} = 14.656 + 0.266 \times 22.84577$   
 $= 14.656 + 6.077 = 20.733$

Costo académico promedio: \$20,733 por estudiante

**2.16** The capital asset pricing model (CAPM) is an important model in the field of finance. It explains variations in the rate of return on a security as a function of the rate of return on a portfolio consisting of all publicly traded stocks, which is called the *market* portfolio. Generally, the rate of return on any investment is measured relative to its opportunity cost, which is the return on a risk-free asset. The resulting difference is called the *risk premium*, since it is the reward or punishment for making a risky investment. The CAPM says that the risk premium on security *j* is *proportional* to the risk premium on the market portfolio. That is,

$$r_j - r_f = \beta_j (r_m - r_f)$$

where  $r_j$  and  $r_f$  are the returns to security *j* and the risk-free rate, respectively,  $r_m$  is the return on the market portfolio, and  $\beta_j$  is the *j*th security's "*beta*" value. A stock's *beta* is important to investors since it reveals the stock's volatility. It measures the sensitivity of security *j*'s return to variation in the whole stock market. As such, values of *beta* less than one indicate that the stock is "defensive" since its variation is less than the market's. A *beta* greater than one indicates an "aggressive stock." Investors usually want an estimate of a stock's *beta* before purchasing it. The CAPM model shown above is the "economic model" in this case. The "econometric model" is obtained by including an intercept in the model (even though theory says it should be zero) and an error term

$$r_j - r_f = \alpha_j + \beta_j (r_m - r_f) + e_j$$

- Explain why the econometric model above is a simple regression model like those discussed in this chapter.
- In the data file *capm5* are data on the monthly returns of six firms (GE, IBM, Ford, Microsoft, Disney, and Exxon-Mobil), the rate of return on the market portfolio (*MKT*), and the rate of return on the risk-free asset (*RISKFREE*). The 180 observations cover January 1998 to December 2012. Estimate the CAPM model for each firm, and comment on their estimated *beta* values. Which firm appears most aggressive? Which firm appears most defensive?
- Finance theory says that the intercept parameter  $\alpha_j$  should be zero. Does this seem correct given your estimates? For the Microsoft stock, plot the fitted regression line along with the data scatter.
- Estimate the model for each firm under the assumption that  $\alpha_j = 0$ . Do the estimates of the *beta* values change much?

2.16 EJERCICIOS

a. Explicación del modelo econométrico

El modelo econométrico CAPM es un **modelo de regresión simple** porque puede expresarse en la forma:

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

Donde:

- $y = r_j - r_f$  (variable dependiente: exceso de retorno del activo)
- $x = r_m - r_f$  (variable independiente: exceso de retorno del mercado)
- $\beta_0 = a_j$  (intercepto)
- $\beta_1 = \beta_j$  (pendiente, coeficiente beta)
- $\epsilon = \epsilon_j$  (término de error)



b. Estimación del modelo CAPM para cada empresa

Resultados de regresión con intercepto:

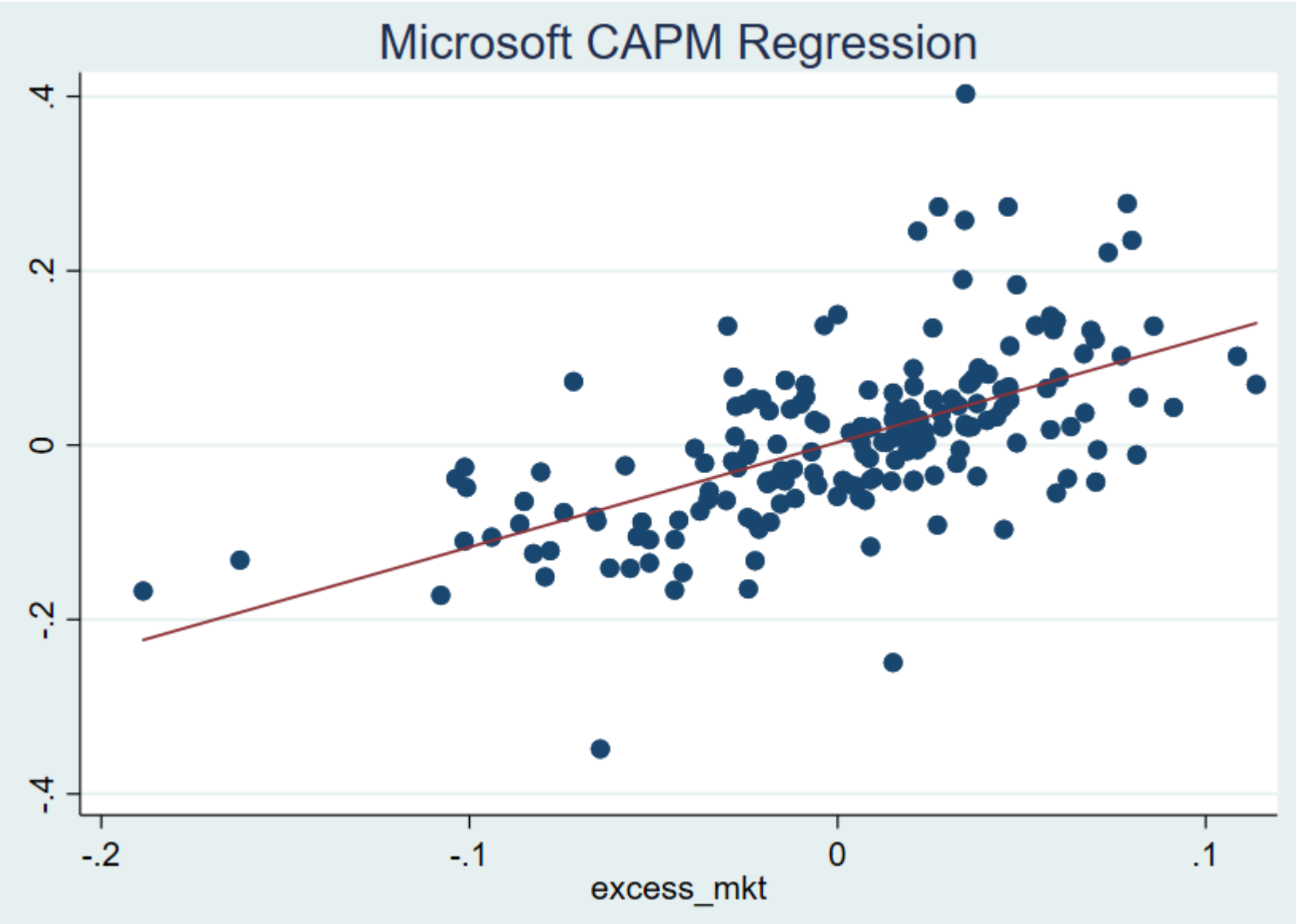


Firma	Beta ( $\beta$ )	Alpha ( $\alpha$ )	P-value $\alpha$	R <sup>2</sup>	Clasificación
Ford	1.6620	0.0038	0.712	0.2660	Muy Agresiva
Microsoft	1.2018	0.0032	0.591	0.3523	Agresiva
GE	1.1480	-0.0010	0.829	0.4801	Agresiva
Disney	1.0115	0.0010	0.823	0.3909	Neutra
IBM	0.9769	0.0061	0.212	0.3590	Defensiva
Exxon-Mobil	0.4565	0.0053	0.137	0.1861	Muy Defensiva

Conclusiones:

-  **Firma más agresiva:** Ford ( $\beta = 1.6620$ )
-  **Firma más defensiva:** Exxon-Mobil ( $\beta = 0.4565$ )

### c. Validación de la teoría financiera ( $a_j = 0$ )



Hipótesis nula ( $H_0$ ):  $\alpha_j = 0$  (Teoría financiera)

Hipótesis alternativa ( $H_1$ ):  $\alpha_j \neq 0$

"Cero estadístico: Sí" = "No tenemos pruebas suficientes para decir que NO es cero"

Firma	Intercepto ( $\alpha$ )	P-value	¿Cero estadístico?	Conclusión
Microsoft	0.0032	0.591	✅ Sí	No significativo
GE	-0.0010	0.829	✅ Sí	No significativo
Ford	0.0038	0.712	✅ Sí	No significativo
IBM	0.0061	0.212	✅ Sí	No significativo
Disney	0.0010	0.823	✅ Sí	No significativo
Exxon-Mobil	0.0053	0.137	✅ Sí	No significativo

**Conclusión:** Todos los interceptos son estadísticamente iguales a cero ( $p\text{-value} > 0.05$ ), lo que **confirma la teoría financiera** que establece  $a_j = 0$ .

### d. Estimación con restricción $a_j = 0$

Comparación de coeficientes beta:

Firma	Beta con $\alpha$	Beta sin $\alpha$	Diferencia	% Cambio
Microsoft	1.2018	1.2059	+0.0041	+0.34%
GE	1.1480	1.1468	-0.0012	-0.10%
Ford	1.6620	1.6667	+0.0047	+0.28%
IBM	0.9769	0.9844	+0.0075	+0.77%
Disney	1.0115	1.0128	+0.0013	+0.13%
Exxon-Mobil	0.4565	0.4631	+0.0066	+1.45%

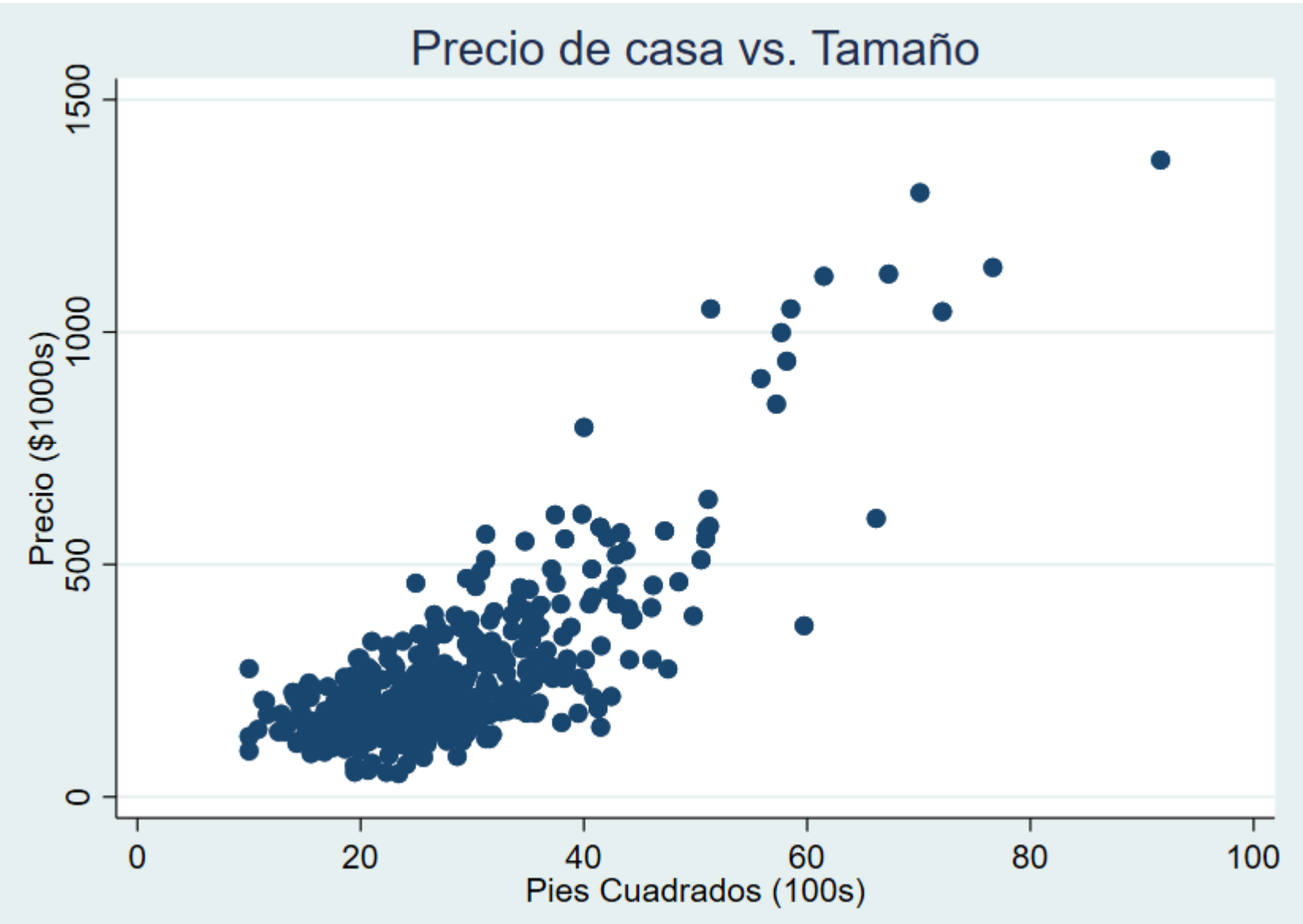
**Conclusión:** Las estimaciones de beta **cambian mínimamente** (menos del 1.5% en todos los casos) al imponer la restricción  $a_j = 0$ .

**2.17** The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

- Plot house price against house size in a scatter diagram.
- Estimate the linear regression model  $PRICE = \beta_1 + \beta_2 SQFT + e$ . Interpret the estimates. Draw a sketch of the fitted line.
- Estimate the quadratic regression model  $PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$ . Compute the marginal effect of an additional 100 square feet of living area in a home with 2000 square feet of living space.
- Graph the fitted curve for the model in part (c). On the graph, sketch the line that is tangent to the curve for a 2000-square-foot house.
- For the model in part (c), compute the elasticity of *PRICE* with respect to *SQFT* for a home with 2000 square feet of living space.
- For the regressions in (b) and (c), compute the least squares residuals and plot them against *SQFT*. Do any of our assumptions appear violated?
- One basis for choosing between these two specifications is how well the data are fit by the model. Compare the sum of squared residuals (*SSE*) from the models in (b) and (c). Which model has a lower *SSE*? How does having a lower *SSE* indicate a “better-fitting” model?

## 2.17 EJERCICIOS

### a. Diagrama de Dispersión



### b. Regresión Lineal Simple

**Modelo estimado:**  $PRICE = \beta_0 + \beta_1 SQFT + e$

**Resultados:**

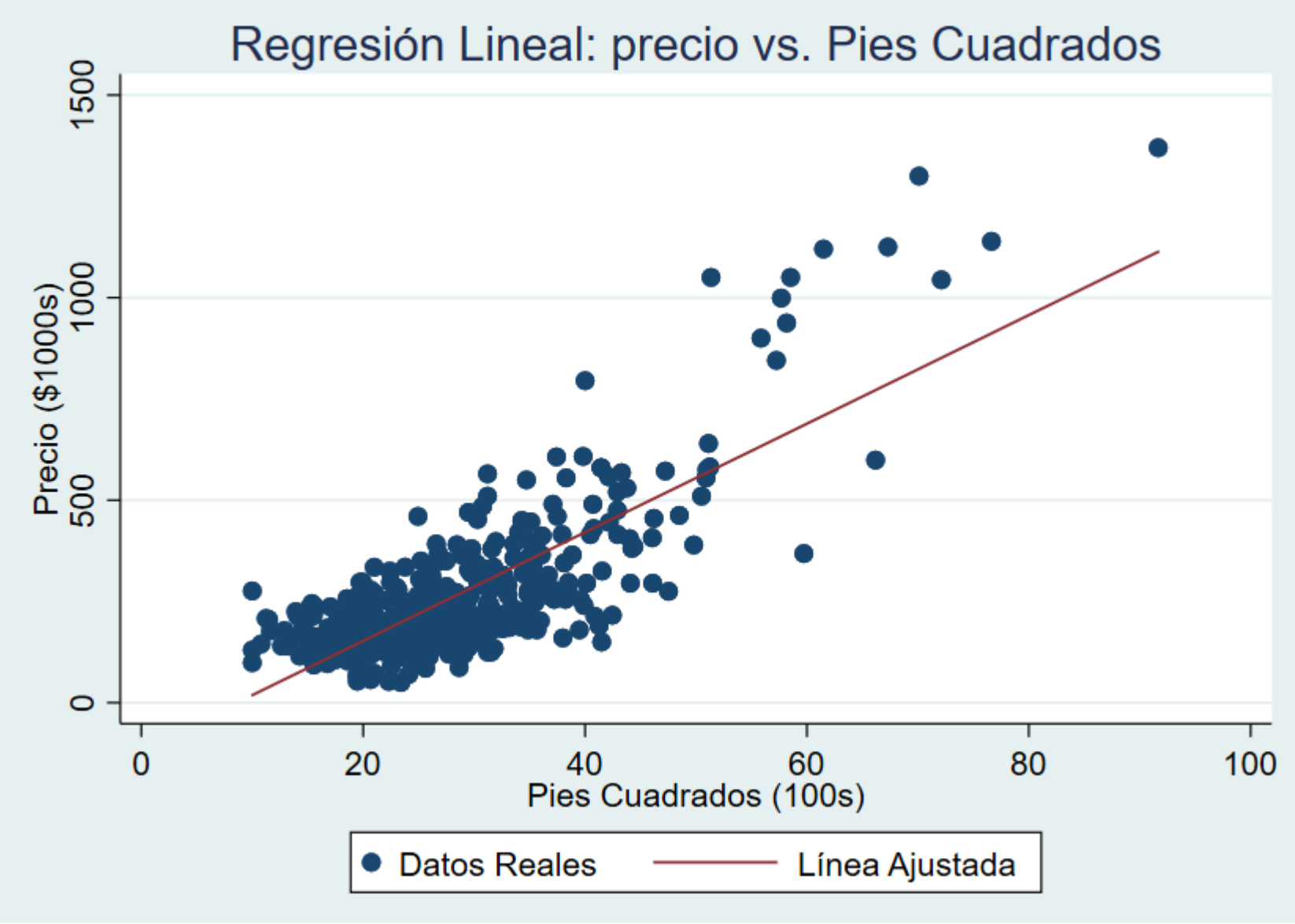
Variable	Coeficiente	Error Estándar	Estadístico t	Valor p
sqft	13.40294	.4491636	29.84	0.000
_cons	-115.4236	13.08815	-8.82	0.000

Ecuación estimada:

$$\widehat{PRICE} = -115.42 + 13.40 \cdot SQFT$$

Interpretación:

- Intercepto** ( $\beta_0 = -115.42$ ): Cuando  $SQFT = 0$ , el precio predicho es -\$115,420 (interpretación poco realista en este contexto)
- Pendiente** ( $\beta_1 = 13.40$ ): Por cada 100 pies cuadrados adicionales, el precio aumenta en \$13,400
- R-cuadrado = 0.6413**: El 64.13% de la variación en el precio es explicada por el tamaño



c. Regresión Cuadrática

Modelo estimado:  $PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$

Resultados:

Variable	Coeficiente	Error Estándar	t	P>t
sqft2	.1844281	.0052624	35.05	0.000
_cons	93.74121	6.08553	15.40	0.000

Ecuación estimada:

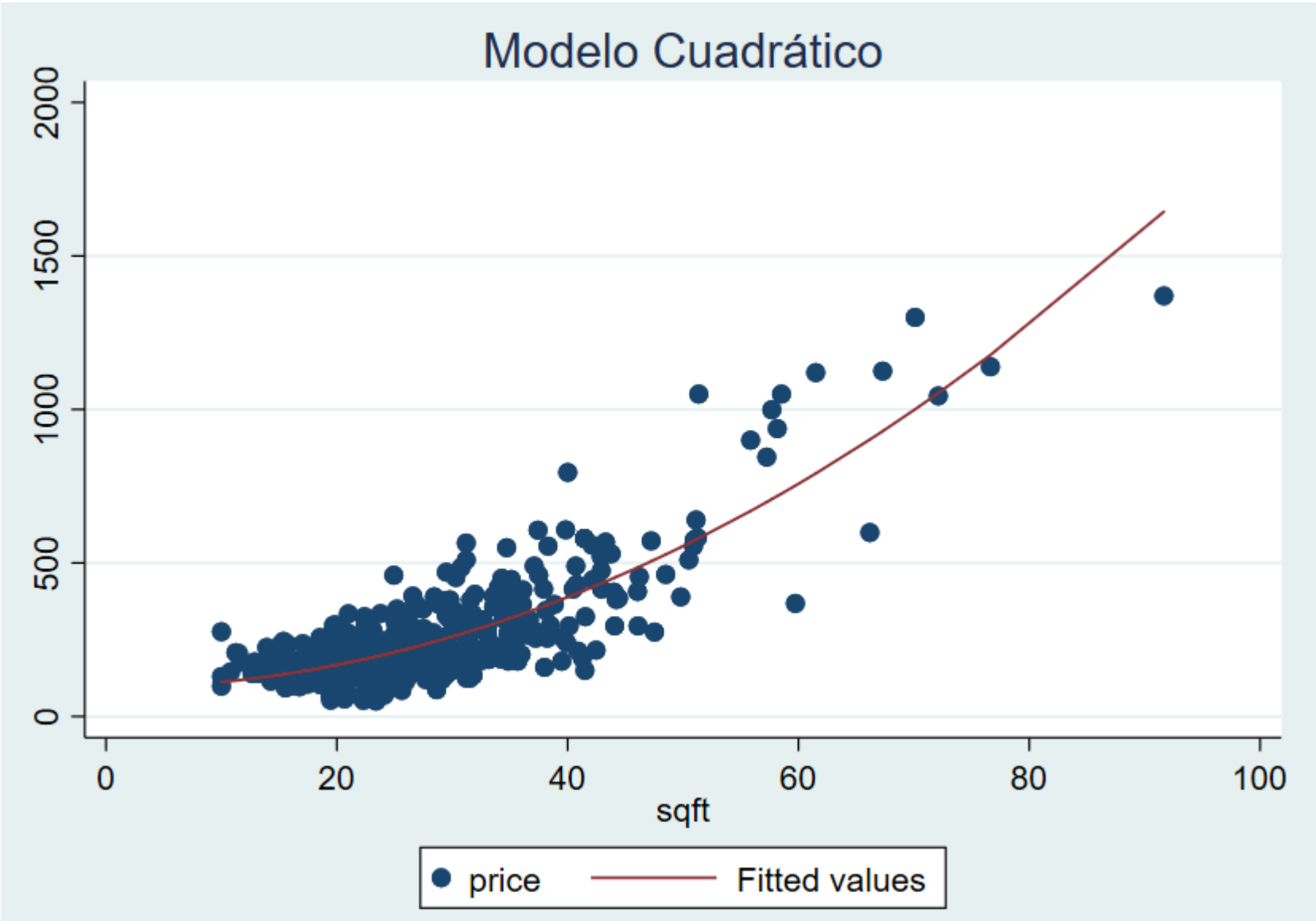
$$\widehat{PRICE} = 93.74 + 0.1844 \cdot SQFT^2$$

Efecto marginal para una casa de 2000 pies cuadrados ( $SQFT = 20$ ):

$$\frac{\partial PRICE}{\partial SQFT} = 2\alpha_2 \cdot SQFT = 2(0.208)(20) = 8.32$$

**Interpretación:** Para una casa de 2000 pies cuadrados, un aumento de 100 pies cuadrados adicionales incrementa el precio en aproximadamente \$8,320.

d. Gráfica del Modelo Cuadrático



e. Elasticidad para casa de 2000 pies cuadrados

Precio estimado para  $SQFT = 20$ :

$$\widehat{PRICE} = 93.74 + 0.1844 \cdot SQFT^2 = 93.74 + 0.1844 \cdot (20)^2 = 167.5$$

Elasticidad:

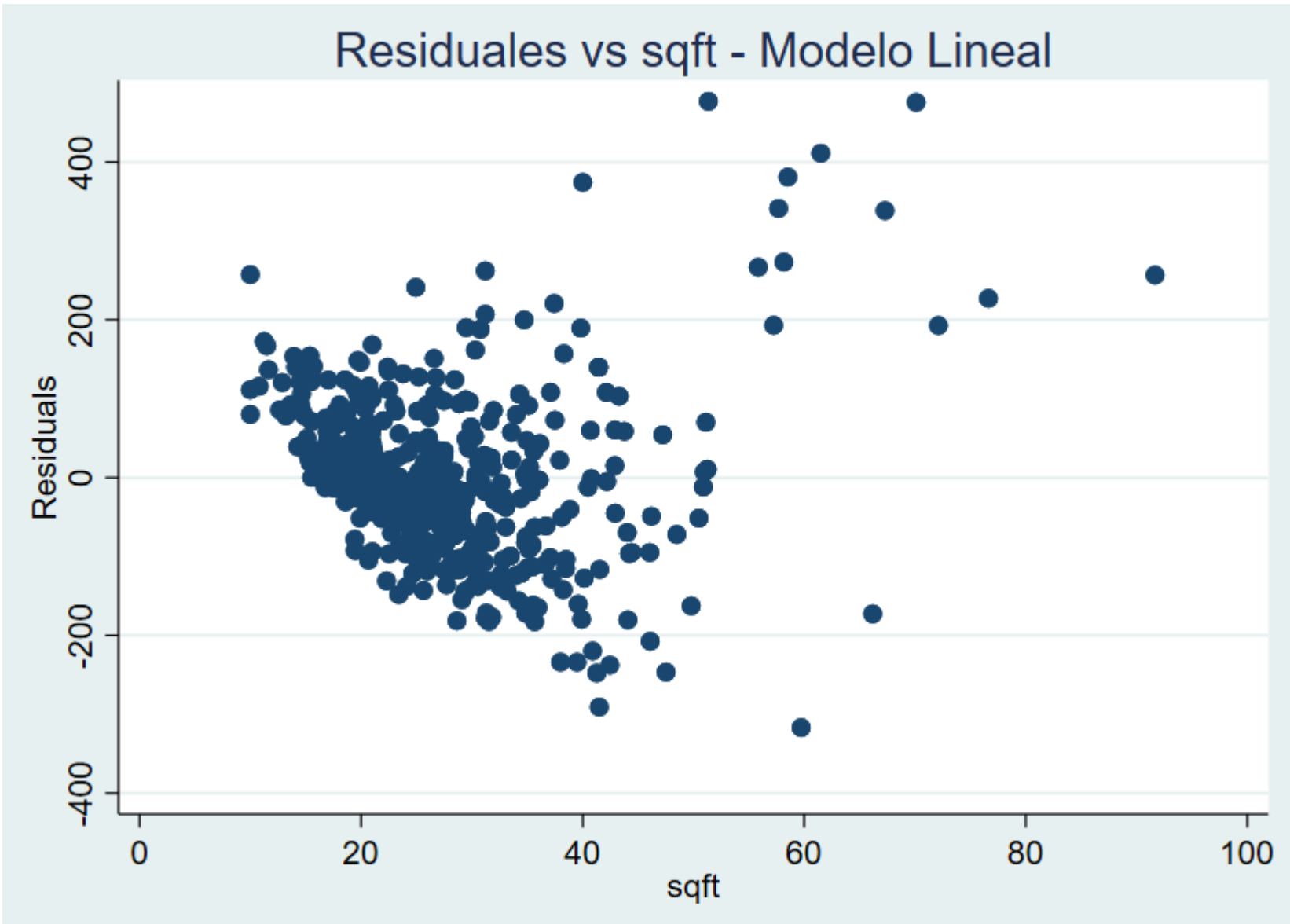
$$\eta = \left(\frac{\partial PRICE}{\partial SQFT}\right) \cdot \left(\frac{SQFT}{PRICE}\right) = \left(\frac{20}{167.5}\right) = 0.1194$$

**Interpretación:** Para una casa de 2000 pies cuadrados, un aumento del 1% en el área construida se asocia con un aumento del 0.1194% en el precio.

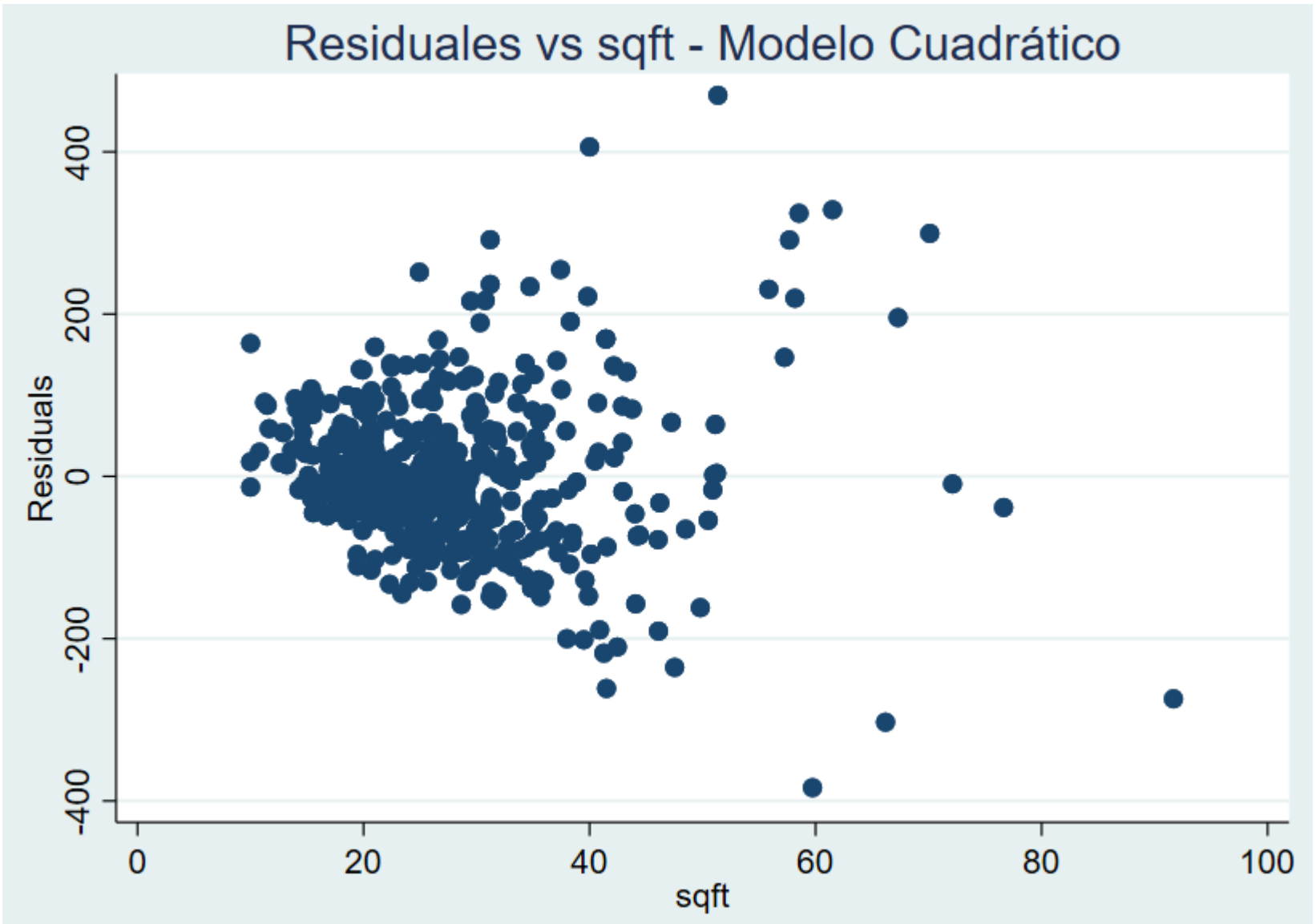
f. Análisis de Residuales

Suma de Residuales al Cuadrado (SSE):

- Modelo lineal:  $SSE = 5,262,681$



- Modelo cuadrático:  $SSE = 4,219,952.1$



**g. Comparación de Modelos**

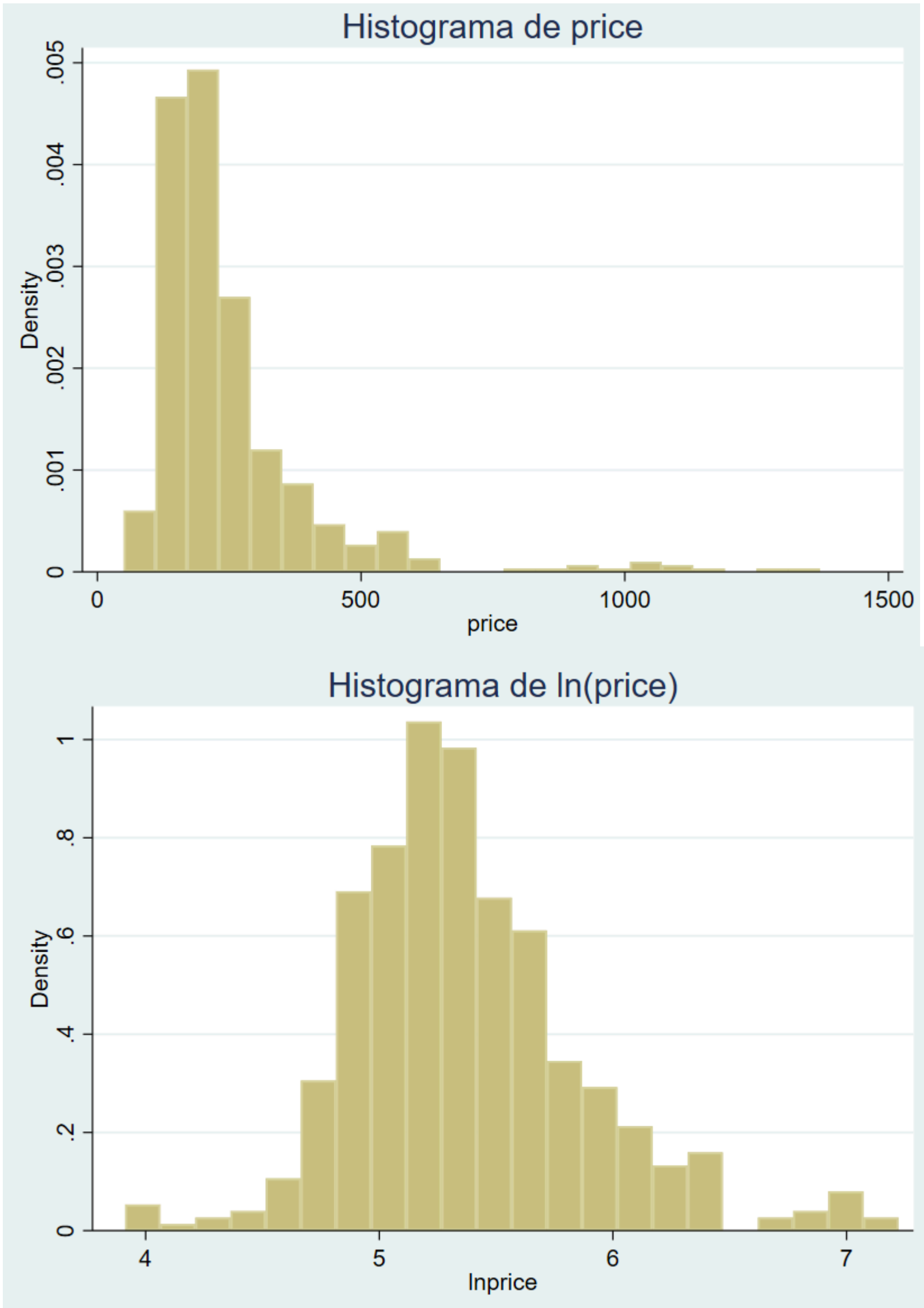
El **modelo cuadrático** tiene un mejor ajuste:

- Menor SSE (4,219,952.1 vs 5,262,681)
- Mayor R-cuadrado (0.7119 vs 0.6407)
- El modelo cuadrático explica aproximadamente 7% más de la variación en precios

- 2.18 The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.
- Create histograms for *PRICE* and  $\ln(\textit{PRICE})$ . Are the distributions skewed or symmetrical?
  - Estimate the log-linear regression model  $\ln(\textit{PRICE}) = \gamma_1 + \gamma_2 \textit{SQFT} + \varepsilon$ . Interpret the OLS estimates,  $\hat{\gamma}_1$  and  $\hat{\gamma}_2$ . Graph the fitted *PRICE*,  $\widehat{\textit{PRICE}} = \exp(\hat{\gamma}_1 + \hat{\gamma}_2 \textit{SQFT})$ , against *SQFT*, and sketch the tangent line to the curve for a house with 2000 square feet of living area. What is the slope of the tangent line?
  - Compute the least squares residuals from the model in (b) and plot them against *SQFT*. Do any of our assumptions appear violated?

## 2.18 EJERCICIOS

### a. Histogramas de PRICE y ln(PRICE)



#### Interpretación:

- El histograma de PRICE muestra una distribución sesgada a la derecha
- El histograma de  $\ln(\textit{PRICE})$  muestra una distribución más simétrica y cercana a la normal
- La transformación logarítmica mejora las propiedades distribucionales del precio



## b. Modelo Log-Lineal

Resultados de la regresión:

Variable	Coeficiente	Error Estándar	t	P>t	[95% Conf. Interval]
sqft	0.0360445	0.0014856	24.26	0.000	0.0331257 - 0.0389633
_cons	4.393866	0.0432883	101.50	0.000	4.308816 - 4.478916

Ecuación estimada:

$$\ln(\widehat{PRICE}) = 4.3939 + 0.0360 \cdot SQFT$$

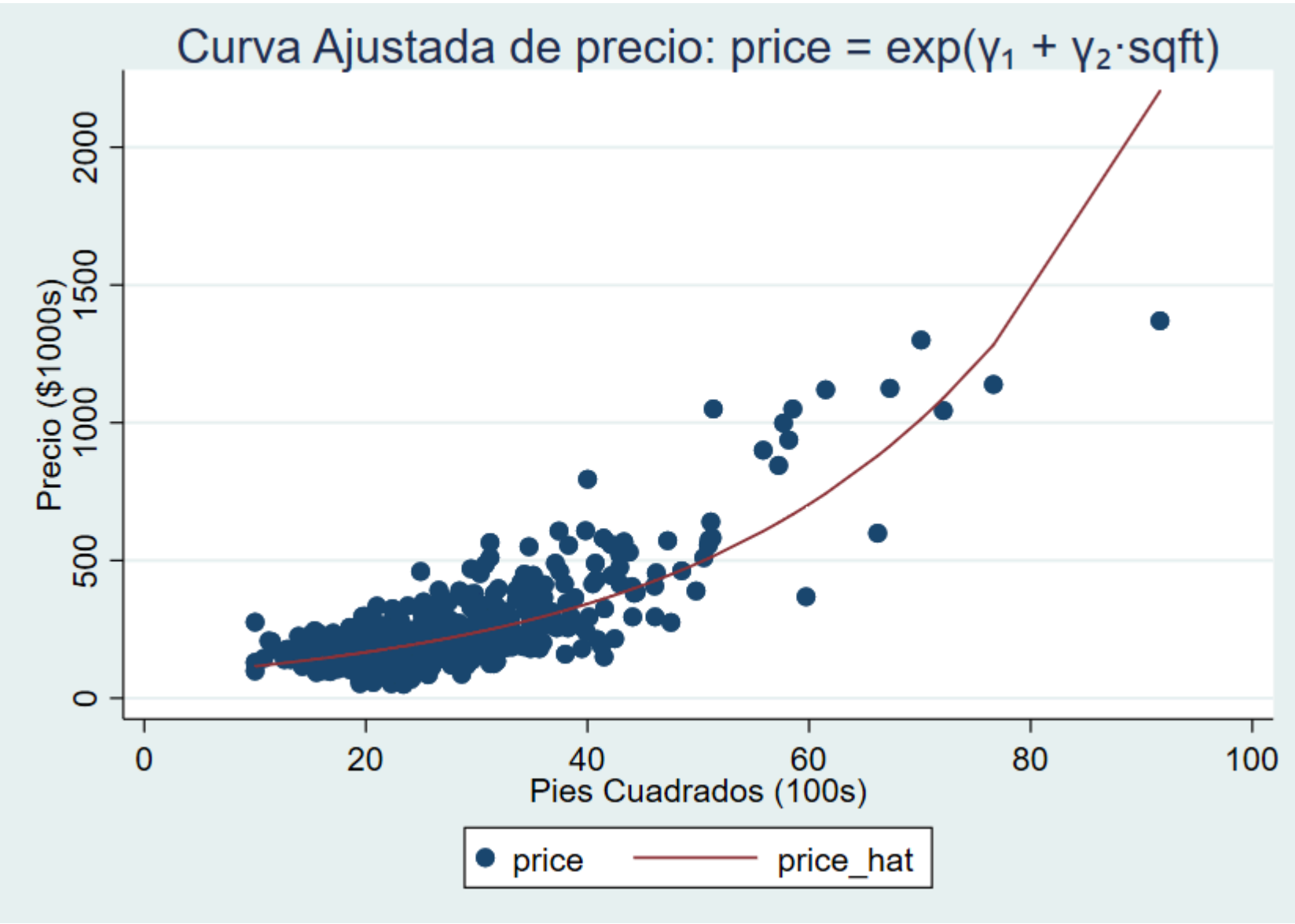
Interpretación de coeficientes:

- **$\gamma_1 = 4.3939$** : Es el log-precio cuando SQFT = 0 (interpretación limitada en contexto práctico)
- **$\gamma_2 = 0.0360$** : Representa la elasticidad precio-tamaño aproximada. Un aumento de 100 pies cuadrados incrementa el precio en aproximadamente 3.6%

Pendiente de la línea tangente para 2000 sq ft:

$$\text{Pendiente} = \hat{\gamma}_2 \cdot \exp(\hat{\gamma}_1 + \hat{\gamma}_2 \cdot 20) = 0.0360 \cdot \exp(4.3939 + 0.0360 \cdot 20) = 6.00$$

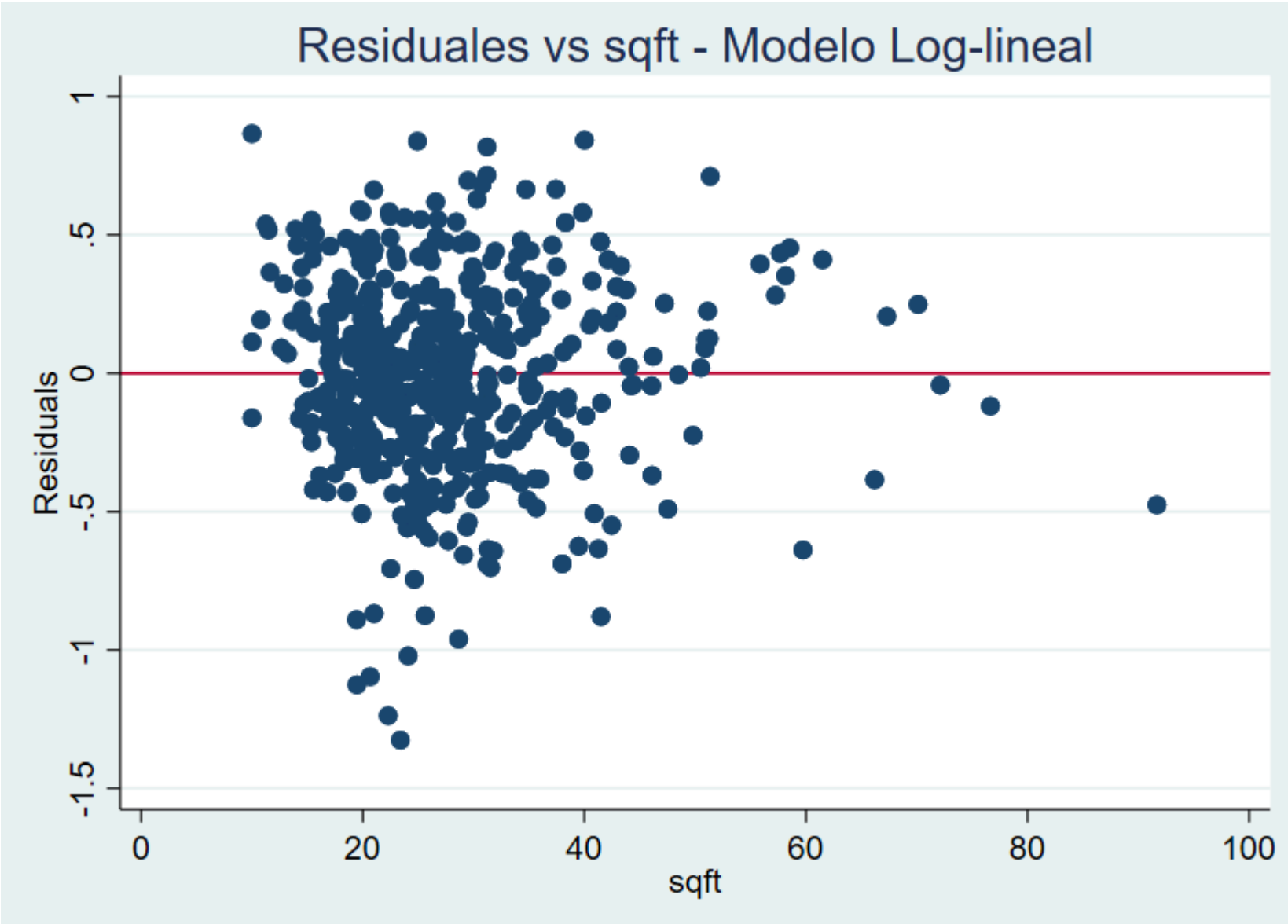
**Interpretación:** Para una casa de 2000 pies cuadrados, la pendiente de la línea tangente es 6.00, lo que significa que un aumento de 100 pies cuadrados incrementa el precio en aproximadamente \$6,000.



## c. Análisis de Residuales

Evaluación de supuestos:

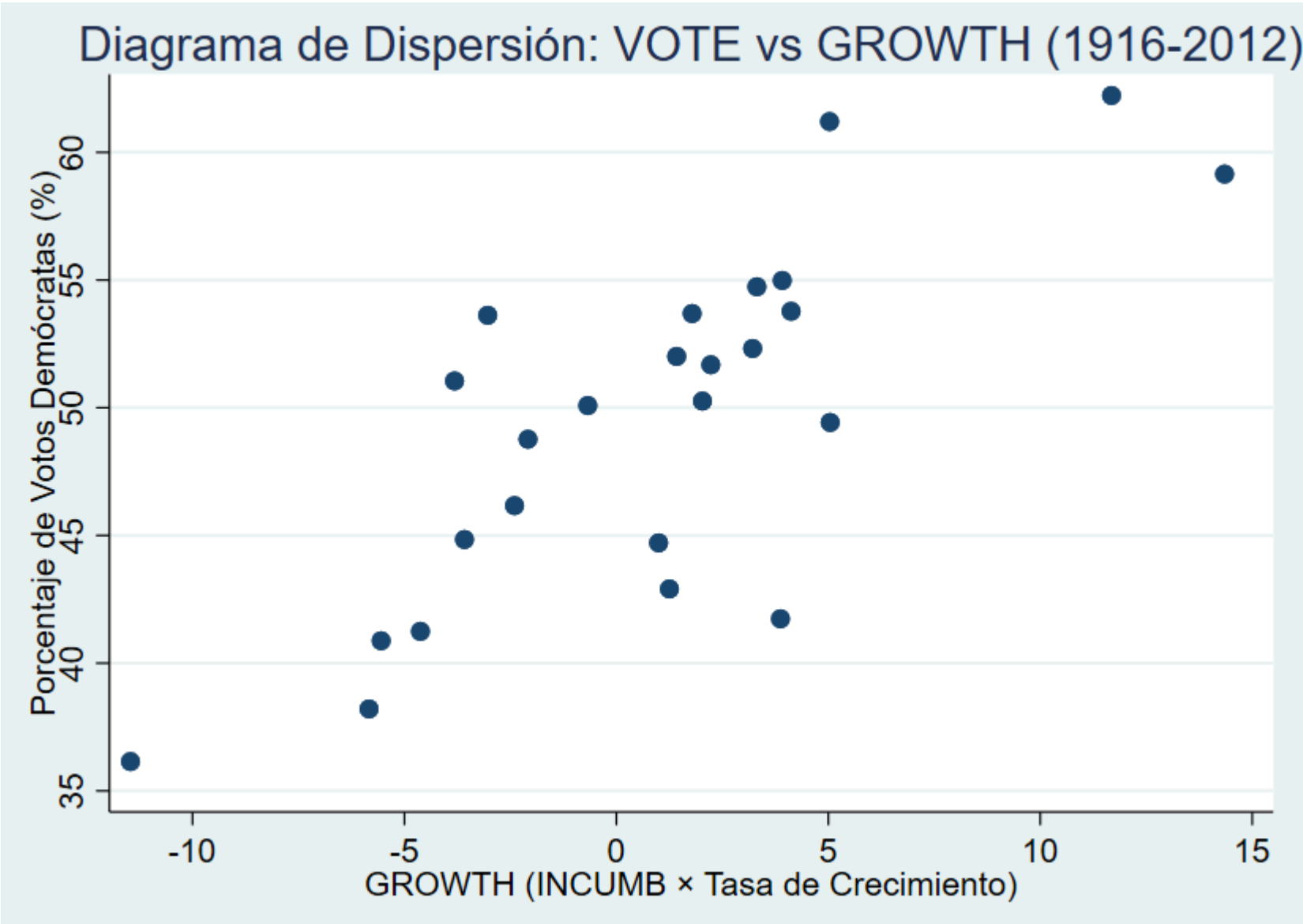
- Los residuales se distribuyen aleatoriamente alrededor de cero
- Se verificar homocedasticidad (varianza constante de los residuales)
- No hay presencia de patrones sistemáticos lo que sugeriría violación de los supuestos del modelo



- 2.23 Professor Ray C. Fair has for a number of years built and updated models that explain and predict the U.S. presidential elections. Visit his website at <https://fairmodel.econ.yale.edu/vote2016/index2.htm>. See in particular his paper entitled “Presidential and Congressional Vote-Share Equations: November 2010 Update.” The basic premise of the model is that the Democratic Party’s share of the two-party [Democratic and Republican] popular vote is affected by a number of factors relating to the economy, and variables relating to the politics, such as how long the incumbent party has been in power, and whether the President is running for reelection. Fair’s data, 26 observations for the election years from 1916 to 2016, are in the data file *fair5*. The dependent variable is *VOTE* = percentage share of the popular vote won by the Democratic Party. Consider the effect of economic growth on *VOTE*. If Democrats are the incumbent party (*INCUMB* = 1) then economic growth, the growth rate in real per capita GDP in the first three quarters of the election year (annual rate), should enhance their chances of winning. On the other hand, if the Republicans are the incumbent party (*INCUMB* = -1), growth will diminish the Democrats’ chances of winning. Consequently, we define the explanatory variable  $GROWTH = INCUMB \times \text{growth rate}$ .
- a. Using the data for 1916–2012, plot a scatter diagram of *VOTE* against *GROWTH*. Does there appear to be a positive association?
  - b. Estimate the regression  $VOTE = \beta_1 + \beta_2 GROWTH + e$  by least squares using the data from 1916 to 2012. Report and discuss the estimation result. Plot the fitted line on the scatter diagram from (a).
  - c. Using the model estimated in (b), predict the 2016 value of *VOTE* based on the actual 2016 value for *GROWTH*. How does the predicted vote for 2016 compare to the actual result?
  - d. Economy wide inflation may spell doom for the incumbent party in an election. The variable  $INFLAT = INCUMB \times \text{inflation rate}$ , where the inflation rate is the growth in prices over the first 15 quarters of an administration. Using the data from 1916 to 2012, plot *VOTE* against *INFLAT*.
  - e. Using the data from 1916 to 2012, report and discuss the estimation results for the model  $VOTE = \alpha_1 + \alpha_2 INFLAT + e$ .
  - f. Using the model estimated in (e), predict the 2016 value of *VOTE* based on the actual 2012 value for *INFLAT*. How does the predicted vote for 2016 compare to the actual result?

## 2.23 EJERCICIOS

a. Diagrama de dispersión:



**Gráfica:** El diagrama de dispersión muestra una tendencia positiva clara entre ambas variables.

**b. Regresión  $VOTE = \beta_1 + \beta_2GROWTH + \varepsilon$ :**

Resultados de la estimación MCO:

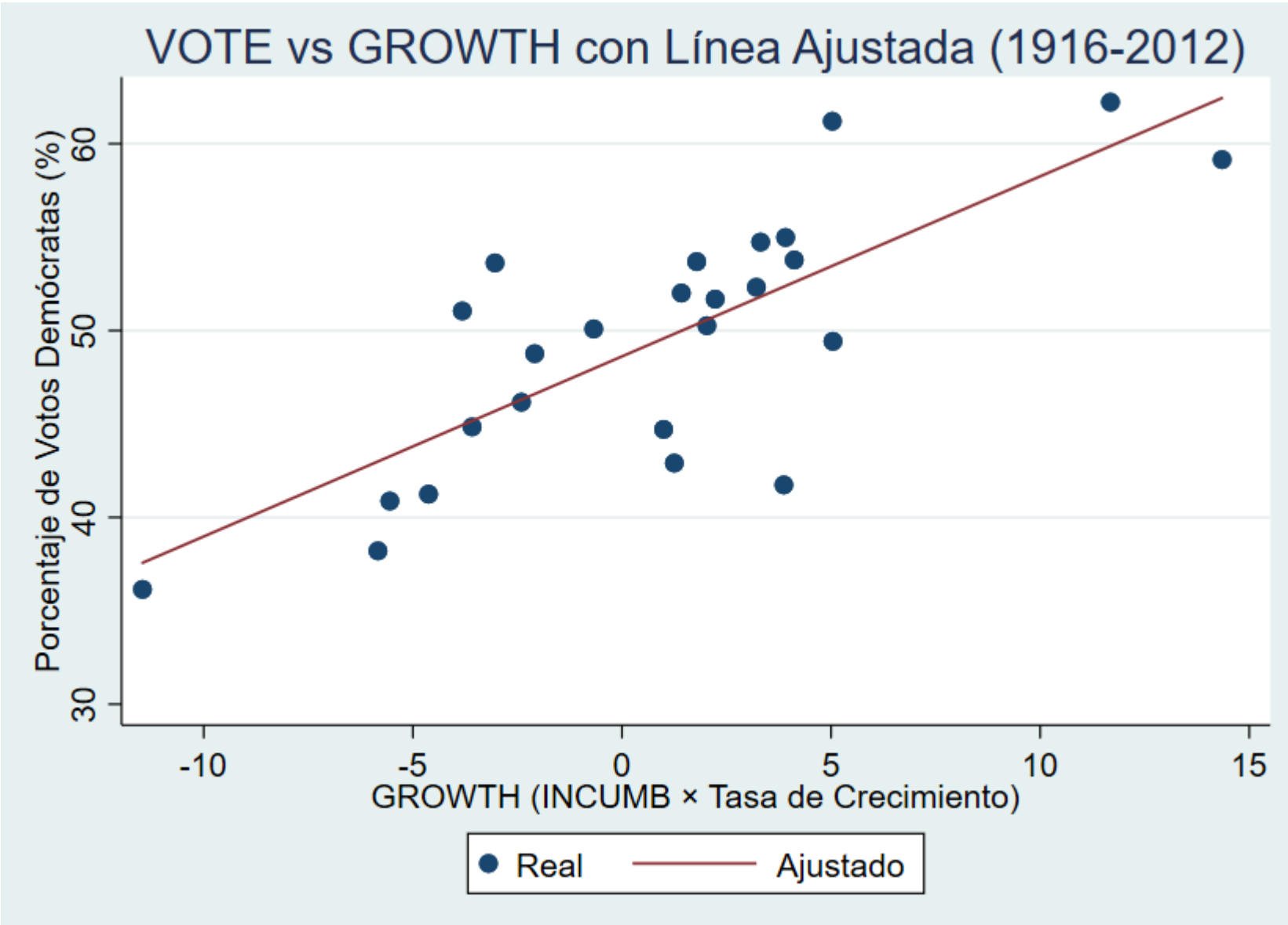
Variable	Coeficiente	Error Estándar	t	P>t	[95% Conf. Interval]
growth	0.9638972	0.1658262	5.81	0.000	0.6208596 - 1.306935
_cons	48.616	0.9042853	53.76	0.000	46.74535 - 50.48666

Ecuación estimada:

$$\widehat{VOTE} = 48.616 + 0.964 \times GROWTH$$

Interpretación:

- $R^2 = 0.595$ : El modelo explica el 59.5% de la variación en el voto demócrata
- Por cada punto porcentual de aumento en GROWTH, el voto demócrata aumenta aproximadamente 0.96 puntos porcentuales
- Ambos coeficientes son estadísticamente significativos ( $p < 0.001$ )



c. Predicción para 2016:

Cálculo:

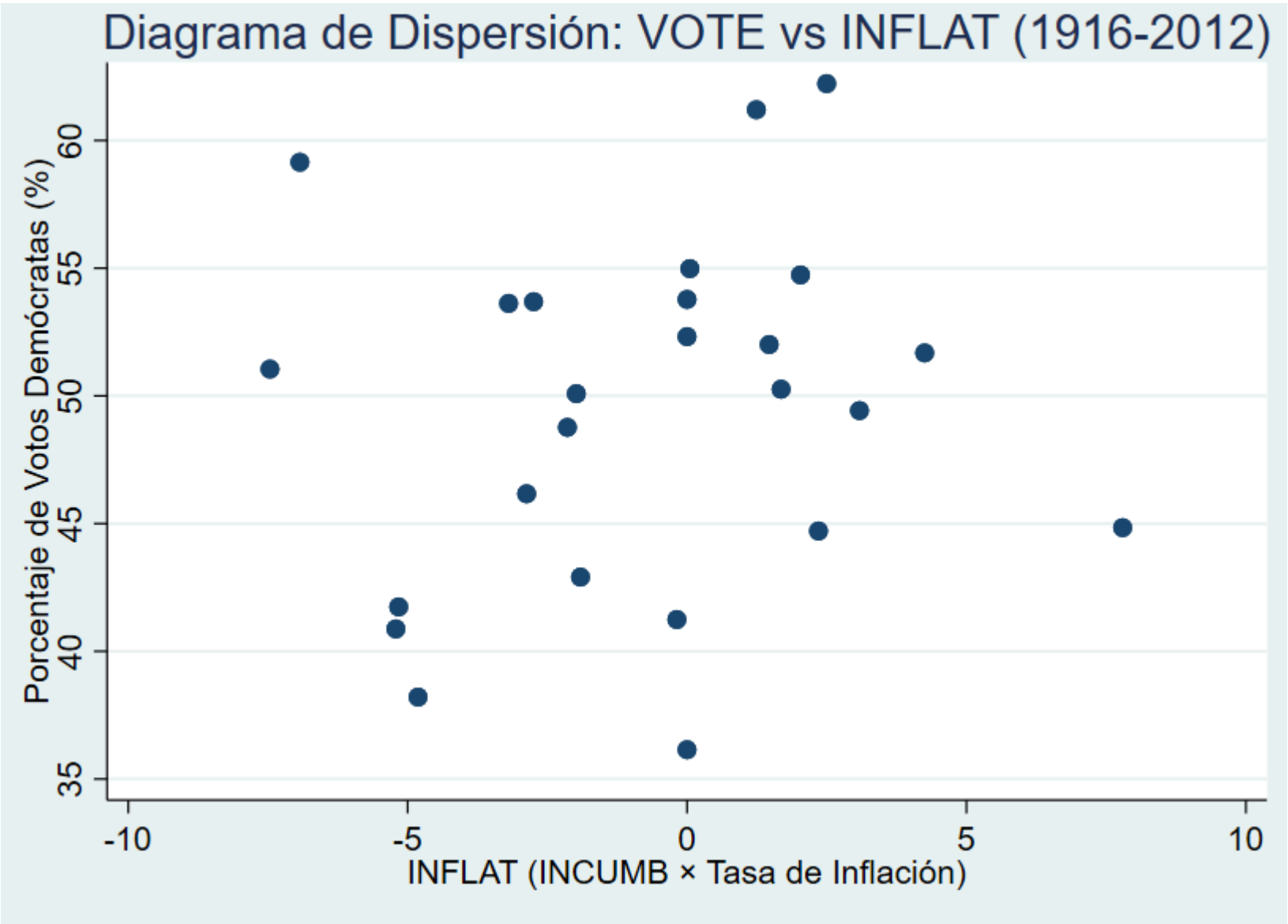
$$\widehat{VOTE}_{2016} = 48.616 + 0.964 \times GROWTH_{2016} = 49.55\%$$

Comparación:

- Voto predicho: 49.55%
- Voto real 2016: 50.82%
- **Error de predicción:** 1.27 puntos porcentuales

El modelo subestimó ligeramente el voto demócrata en las elecciones de 2016.

d. Diagrama de dispersión VOTE vs INFLAT:



**Gráfica:** El diagrama de dispersión no muestra una relación clara entre INFLAT y VOTE.

e. Regresión  $VOTE = \alpha + \gamma INFLAT + \varepsilon$ :

Resultados de la estimación MCO:

Variable	Coeficiente	Error Estándar	t	P>t	[95% Conf. Interval]
inflat	0.2616102	0.3907449	0.67	0.510	-0.5467072 - 1.069928
_cons	49.62289	1.418761	34.98	0.000	46.68796 - 52.55782

Ecuación estimada:

$$\widehat{VOTE} = 49.623 + 0.262 \times INFLAT$$

Interpretación:

- $R^2 = 0.019$ : El modelo explica solo el 1.9% de la variación
- El coeficiente de INFLAT no es estadísticamente significativo (p = 0.510)
- La inflación no tiene un efecto sistemático discernible sobre el voto demócrata

f. Predicción 2016 con modelo de inflación:

Cálculo:

$$\widehat{VOTE}_{2016} = 49.623 + 0.262 \times INFLAT_{2016} = 49.99\%$$

Comparación:

- Voto predicho: 49.99%
- Voto real 2016: 50.82%
- **Error de predicción:** 0.83 puntos porcentuales

Aunque el error absoluto es menor que con el modelo de crecimiento, el modelo de inflación carece de poder explicativo y validez estadística.

2.28 How much does education affect wage rates? The data file *cps5\_small* contains 1200 observations on hourly wage rates, education, and other variables from the 2013 Current Population Survey (CPS). [Note: *cps5* is a larger version.]

- Obtain the summary statistics and histograms for the variables *WAGE* and *EDUC*. Discuss the data characteristics.
- Estimate the linear regression  $WAGE = \beta_1 + \beta_2 EDUC + e$  and discuss the results.
- Calculate the least squares residuals and plot them against *EDUC*. Are any patterns evident? If assumptions SR1–SR5 hold, should any patterns be evident in the least squares residuals?
- Estimate separate regressions for males, females, blacks, and whites. Compare the results.
- Estimate the quadratic regression  $WAGE = \alpha_1 + \alpha_2 EDUC^2 + e$  and discuss the results. Estimate the marginal effect of another year of education on wage for a person with 12 years of education and for a person with 16 years of education. Compare these values to the estimated marginal effect of education from the linear regression in part (b).

2.28 EJERCICIOS

a. Estadísticas Descriptivas e Histogramas

Resumen estadístico:

Variable	Observaciones	Media	Desviación Estándar	Mínimo	Máximo
wage	1,200	23.64	15.22	3.94	221.10
educ	1,200	14.20	2.89	0	21

Características de los datos:

- El salario promedio por hora es de \$23.64 con una dispersión considerable (DE = \$15.22)
- La educación promedio es de 14.2 años
- Existe una amplia variación en ambas variables, con salarios que van desde \$3.94 hasta \$221.10
- Los años de educación varían desde 0 hasta 21 años

b. Regresión Lineal  $WAGE = \beta_1 + \beta_2 EDUC + \varepsilon$

Resultados de la estimación MCO:

Variable	Coeficiente	Error Estándar	t	P>t	[95% Conf. Interval]
educ	2.3968	0.1354	17.70	0.000	2.1311 - 2.6624
_cons	-10.4000	1.9624	-5.30	0.000	-14.2501 - -6.5498

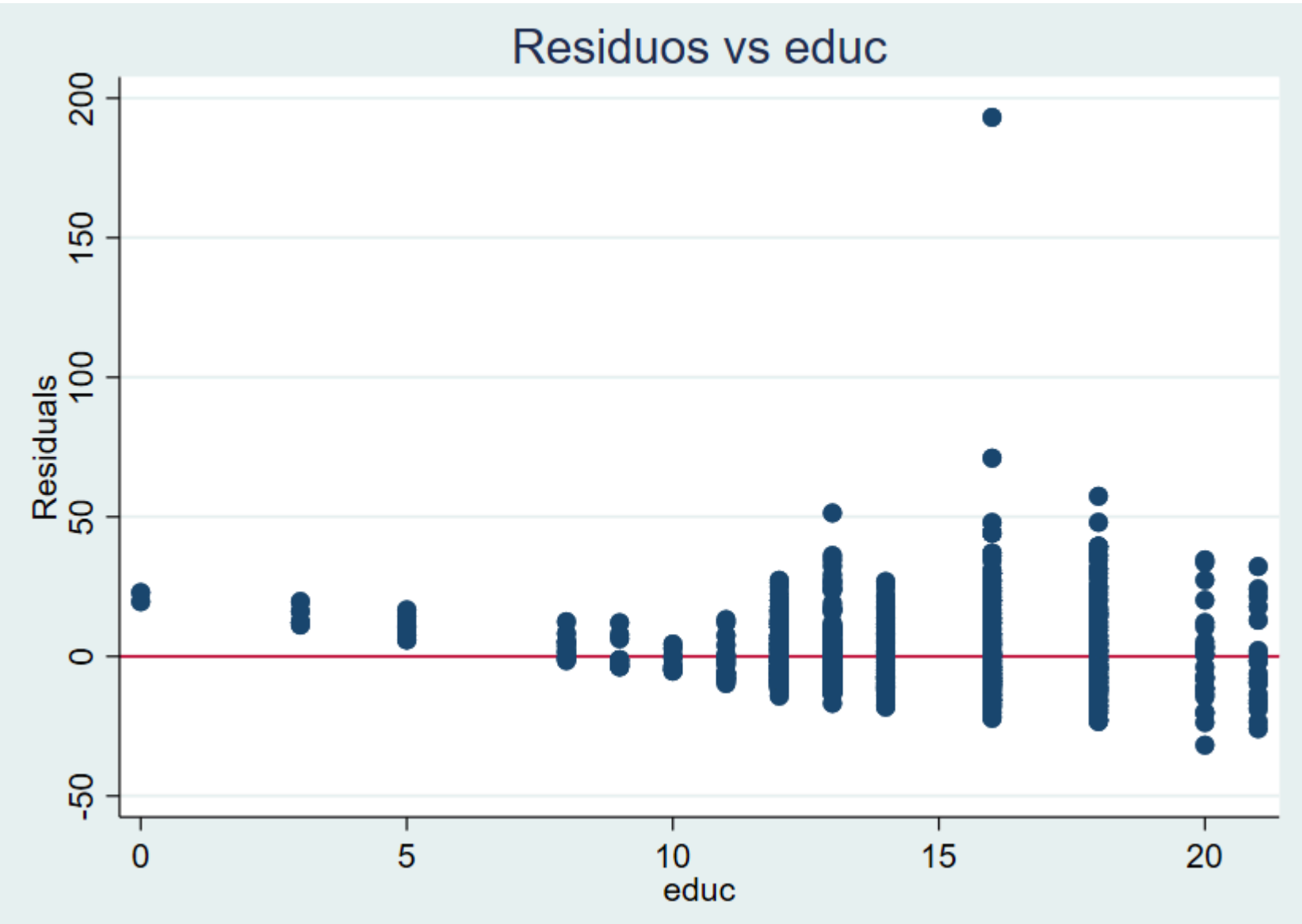
Ecuación estimada:

$$\widehat{WAGE} = -10.4000 + 2.3968 \times EDUC$$

Interpretación:

- R² = 0.2073: El modelo explica el 20.73% de la variación en los salarios
- El coeficiente de EDUC es estadísticamente significativo (p < 0.001)
- Cada año adicional de educación se asocia con un aumento de \$2.40 en el salario por hora
- El intercepto negativo sugiere que sin educación formal, el salario sería negativo (interpretación problemática)

c. Análisis de Residuales



Patrones observados:

- Los residuos muestran una dispersión heterogénea a lo largo de los niveles educativos
- La varianza de los residuos parece aumentar con mayores niveles de educación
- Existen varios valores atípicos significativos

Implicaciones:

- La heterocedasticidad observada viola el supuesto de varianza constante en los errores
- Bajo el supuesto de MCO de muestreo aleatorio simple (SRS), no deberían observarse patrones sistemáticos en los residuos
- Los patrones observados sugieren posibles problemas de especificación del modelo

d. Regresiones Separadas por Grupos

Comparación de coeficientes:

Grupo	Coeficiente EDUC	Intercepto	R²	Observaciones
Total	2.3968***	-10.4000***	0.2073	1,200
Hombres	2.3785***	-8.2849**	0.1927	672
Mujeres	2.6595***	-16.6028***	0.2764	528
Blancos	2.4178***	-10.4747***	0.2072	1,095
Negros	1.9233***	-6.2541	0.1846	105

Hallazgos principales:

- Las mujeres experimentan un mayor retorno por año de educación (\$2.66) que los hombres (\$2.38)
- La educación explica una mayor proporción de la variación salarial en mujeres (R² = 0.2764) que en hombres (R² = 0.1927)



- Los blancos tienen retornos educativos ligeramente superiores a los negros
- El intercepto para negros no es estadísticamente significativo

e. Regresión Cuadrática  $WAGE = \alpha_1 + \alpha_2 EDUC^2 + \varepsilon$

Resultados de la estimación MCO:

Variable	Coeficiente	Error Estándar	t	P>t	[95% Conf. Interval]
educ	-0.4295	0.6644	-0.65	0.518	-1.7330 - 0.8740
educ2	0.1043	0.0240	4.34	0.000	0.0572 - 0.1515
_cons	7.8220	4.6251	1.69	0.091	-1.2522 - 16.8962

Efectos marginales de la educación:

- Para persona con 12 años de educación:  $-0.4295 + 2 \times 0.1043 \times 12 = \$2.0737$  por año adicional
- Para persona con 16 años de educación:  $-0.4295 + 2 \times 0.1043 \times 16 = \$2.9081$  por año adicional

Comparación con modelo lineal:

- El modelo lineal estima un efecto marginal constante de \$2.3968 por año educativo
- El modelo cuadrático muestra que el efecto marginal aumenta con el nivel educativo:
  - A 12 años: \$2.07 (13.6% menor que el modelo lineal)
  - A 16 años: \$2.91 (21.4% mayor que el modelo lineal)
- El modelo cuadrático tiene mayor poder explicativo ( $R^2 = 0.2196$  vs  $0.2073$ )

Conclusión:

La relación entre educación y salarios parece ser no lineal, con retornos crecientes a la educación a medida que se alcanzan niveles educativos más altos.

**2.29** How much does education affect wage rates? The data file *cps5\_small* contains 1200 observations on hourly wage rates, education, and other variables from the 2013 Current Population Survey (CPS). [Note: *cps5* is a larger version with more observations and variables.]

b. Obtain the OLS estimates from the log-linear regression model  $\ln(WAGE) = \beta_1 + \beta_2 EDUC + e$  and interpret the estimated value of  $\beta_2$ .

c. Obtain the predicted wage,  $\widehat{WAGE} = \exp(b_1 + b_2 EDUC)$ , for a person with 12 years of education and for a person with 16 years of education.

d. What is the marginal effect of additional education for a person with 12 years of education and for a person with 16 years of education? [Hint: This is the slope of the fitted model at those two points.]

e. Plot the fitted values  $\widehat{WAGE} = \exp(b_1 + b_2 EDUC)$  versus *EDUC* in a graph. Also include in the graph the fitted linear relationship. Based on the graph, which model seems to fit the data better, the linear or log-linear model?

2.29 EJERCICIOS

b. Regresión Log-Lineal  $\ln(WAGE) = \beta_1 + \beta_2 EDUC + \varepsilon$

Resultados de la estimación MCO:

Variable	Coeficiente	Error Estándar	t	P>t	[95% Conf. Interval]
educ	0.0988	0.0048	20.39	0.000	0.0893 - 0.1083
_cons	1.5968	0.0702	22.75	0.000	1.4591 - 1.7345

Ecuación estimada:

$$\ln(\widehat{WAGE}) = 1.5968 + 0.0988 \times EDUC$$

Interpretación:

- $R^2 = 0.2577$ : El modelo explica el 25.77% de la variación en el logaritmo de los salarios
- El coeficiente de EDUC es altamente significativo ( $p < 0.001$ )
- **Interpretación de  $\beta_2$ :** Un año adicional de educación se asocia con un incremento aproximado del 9.88% en el salario por hora
- Esta especificación log-lineal captura mejor la relación porcentual entre educación y salarios

c. Predicciones de Salario

Salarios predichos utilizando  $\widehat{WAGE} = \exp(b_1 + b_2 \times EDUC)$ :

Años de Educación	Salario Predicho
12 años	\$16.15
16 años	\$23.97

Observaciones:

- La diferencia de 4 años de educación se traduce en una diferencia de \$7.82 en salario predicho
- El modelo predice un salario de \$16.15 para personas con educación secundaria completa
- Para personas con educación universitaria completa (16 años), predice \$23.97

d. Efectos Marginales de la Educación

El efecto marginal se calcula como:  $\frac{\partial WAGE}{\partial EDUC} = \beta_2 \times \exp(\beta_1 + \beta_2 \times EDUC)$

Años de Educación	Efecto Marginal
12 años	\$1.59
16 años	\$2.37

Interpretación:

- Para una persona con 12 años de educación, un año adicional incrementa el salario en aproximadamente \$1.59 por hora
- Para una persona con 16 años de educación, un año adicional incrementa el salario en aproximadamente \$2.37 por hora
- El efecto marginal es **creciente** con el nivel educativo: la educación tiene rendimientos marginales crecientes

e. Comparación de Modelos

Análisis gráfico de los ajustes:

- El gráfico muestra la línea de ajuste del modelo lineal (azul) y la del modelo log-lineal (rojo)
- Ambos modelos no se ajusta bien a la distribución de los datos
- Para niveles educativos altos, el modelo lineal predice salarios más altos que el modelo log-lineal

Evaluación de los modelos:

- Modelo lineal:** R² = 0.2073
- Modelo log-lineal:** R² = 0.2577
- El modelo log-lineal explica una mayor proporción de la variación en los salarios
- La forma funcional log-lineal captura mejor la relación no lineal entre educación y salarios

Comparación de Modelos: Lineal vs Log-Lineal

