CHAPTER 2

Exercise Solutions

EXERCISE 2.1

(a)

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
$\sum x_i =$ 5	$\sum y_i =$ 10	$\sum(x_i-\bar{x})=$ 0	$\sum(x_i-\bar{x})^2=$ 10	$\sum(y-\bar{y})=$ 0	$\sum(x-\bar{x})(y-\bar{y})=$ 8

$\bar{x}=1, \quad \bar{y}=2$

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

(b)

b_2 is the estimated slope of the fitted line.

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

b_1 is the estimated value of $E(y)$ when $x=0$; it is the intercept of the fitted line.

(c)

$$\sum_{i=1}^{5}x_i^2=3^2+2^2+1^2+(-1)^2+0^2=15$$

$$\sum_{i=1}^{5}x_iy_i=3\times 4+2\times 2+1\times 3+(-1)\times 1+0\times 0=18$$

$$\sum_{i=1}^{5}x_i^2-N\bar{x}^2=15-5\times 1^2=10=\sum_{i=1}^{5}(x_i-\bar{x})^2$$

$$\sum_{i=1}^{5}x_iy_i-N\bar{x}\bar{y}=18-5\times 1\times 2=8=\sum_{i=1}^{5}(x_i-\bar{x})(y_i-\bar{y})$$

(d)

x_i	y_i	$\hat{y}_i$	$\hat{e}_i$	$\hat{e}_i^2$	$x_i\hat{e}_i$
3	4	3.6	0.4	0.16	1.2
2	2	2.8	−0.8	0.64	−1.6
1	3	2	1	1	1
-1	1	0.4	0.6	0.36	−0.6
0	0	1.2	−1.2	1.44	0
$\sum x_i =$ 5	$\sum y_i =$ 10	$\sum\hat{y}_i =$ 10	$\sum\hat{e}_i =$ 0	$\sum\hat{e}_i^2 =$ 3.6	$\sum x_i\hat{e}_i =$ 0

Exercise 2.1 (continued)

$$s_y^2 = \sum_{i=1}^{N}(y_i - \bar{y})^2 \Big/ (N-1) = 10/4 = 2.5$$
$$s_x^2 = \sum_{i=1}^{N}(x_i - \bar{x})^2 \Big/ (N-1) = 10/4 = 2.5$$
$$s_{xy} = \sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x}) \Big/ (N-1) = 8/4 = 2$$
$$r_{xy} = s_{xy} \Big/ (s_x s_y) = 2 \Big/ \left(\sqrt{2.5}\sqrt{2.5}\right) = 0.8$$
$$CV_x = 100(s_x/\bar{x}) = 100\sqrt{2.5}/1 = 158.11388$$
$$\text{median}(x) = 1$$

(e)

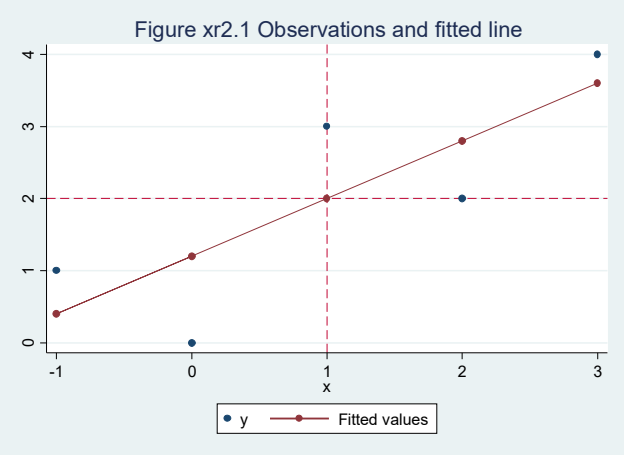


Figure xr2.1 Observations and fitted line

(f) See figure above. The fitted line passes through the point of the means, $\bar{x} = 1, \quad \bar{y} = 2$.

(g) Given $b_1 = 1.2$, $b_2 = 0.8$, $\bar{x} = 1$, $\bar{y} = 2$ and $\bar{y} = b_1 + b_2\bar{x}$, we have
$\bar{y} = 2 = b_1 + b_2\bar{x} = 1.2 + 0.8(1) = 2$

(h) $\bar{\hat{y}} = \sum \hat{y}_i / N = (3.6 + 2.8 + 2 + 0.4 + 1.2)/5 = 2 = \bar{y}$

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

(j) $$\widehat{\text{var}}(b_2) = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2} = \frac{1.2}{10} = 0.12$$ and $se(b_2) = \sqrt{\widehat{\text{var}}(b_2)} = \sqrt{0.12} = 0.34641$

EXERCISE 2.2

(a)

$$P(200 < X < 215) = P\left(\frac{200 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} < \frac{X - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} < \frac{215 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}}\right)$$

$$= P\left(\frac{200 - 220}{\sqrt{121}} < Z < \frac{215 - 220}{\sqrt{121}}\right)$$

$$= P(-1.8181 < Z < -0.4545)$$

$$= 0.2902$$

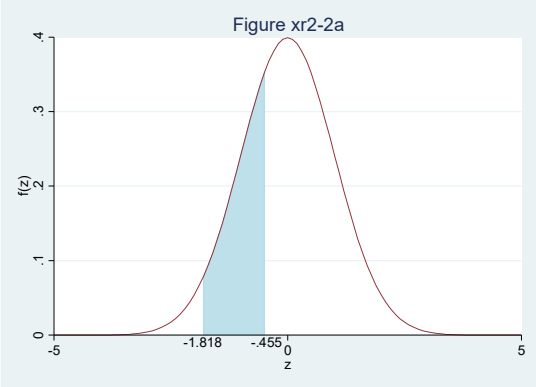


Figure xr2.2(a) Sketch of solution

(b)

$$P(X > 250) = P\left(\frac{X - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} > \frac{190 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}}\right)$$

$$= P\left(Z > \frac{250 - 220}{\sqrt{121}}\right) = P(Z > 2.7273)$$

$$= 1 - P(Z \le 2.7273)$$

$$= 0.00319301$$

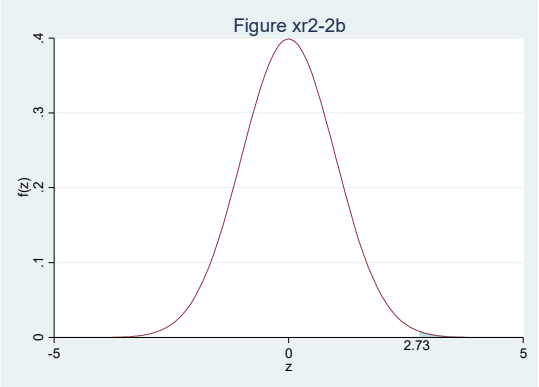


Figure xr2.2(b) Sketch of solution

Exercise 2.2 (continued)

(c)

$$P(200 < X < 215) = P\left(\frac{200 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} < \frac{X - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} < \frac{215 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}}\right)$$

$$= P\left(\frac{200 - 220}{\sqrt{144}} < Z < \frac{215 - 220}{\sqrt{144}}\right)$$

$$= P(-1.667 < Z < -0.4167)$$

$$= 0.2907$$

(d)

$$P(X > 250) = P\left(\frac{X - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} > \frac{190 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}}\right)$$

$$= P\left(Z > \frac{250 - 220}{\sqrt{144}}\right) = P(Z > 2.5)$$

$$= 1 - P(Z \le 2.5)$$

$$= 0.0062$$

$$P(X > 190) = P\left(\frac{X - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} > \frac{190 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}}\right)$$

$$= P\left(Z > \frac{190 - 200}{\sqrt{81}}\right)$$

$$= 1 - P(Z \le -1.1111)$$

$$= 0.8667$$

EXERCISE 2.3

(a) The observations on y and x and the estimated least-squares line are graphed in part (b). The line drawn for part (a) will depend on each student's subjective choice about the position of the line. We show the least squares fitted line.

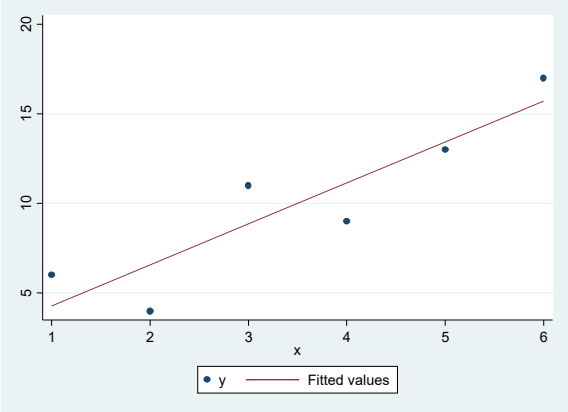


Figure xr2.3(a) Observations and line through data

(b) Preliminary calculations yield:

$$\sum x_i = 21 \qquad \sum y_i = 60 \qquad \sum (x_i - \bar{x})(y_i - \bar{y}) = 40 \qquad \sum (x_i - \bar{x})^2 = 17.5$$

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

The least squares estimates are:

$$b_2 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{40}{17.5} = 2.285714$$

$$b_1 = \bar{y} - b_2 \bar{x} = 10 - (2.285714) \times 3.5 = 2$$

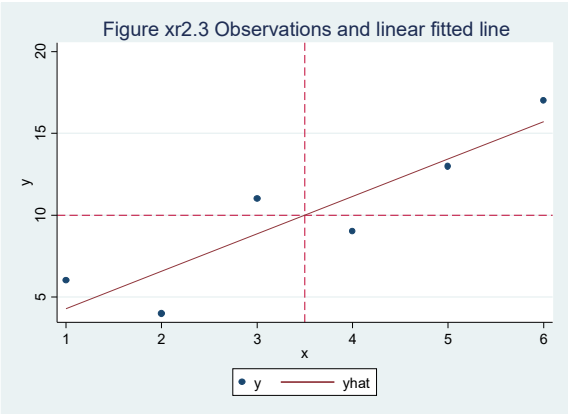


Figure xr2.3(b) Observations and fitted line

Exercise 2.3 (continued)

(c) $\bar{y} = \sum y_i / N = 60/6 = 10$ and $\bar{x} = \sum x_i / N = 21/6 = 3.5$

The predicted value for y at $x = \bar{x}$ is

$$\hat{y} = b_1 + b_2 \bar{x} = 2 + 2.285714 \times 3.5 = 10$$

We observe that $\hat{y} = b_1 + b_2 \bar{x} = \bar{y}$. That is, the predicted value at the sample mean $\bar{x}$ is the sample mean of the dependent variable $\bar{y}$. This implies that the least-squares estimated line passes through the point $(\bar{x}, \bar{y})$. This point is at the intersection of the two dashed lines plotted on the graph in part (b) .

(d) The values of the least squares residuals, computed from $\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i$, are:

x_i	y_i	$\hat{e}_i$
1	6	1.71429
2	4	−2.57143
3	11	2.14286
4	9	−2.14286
5	13	−0.42857
6	17	1.28571

(e) Their sum is $\sum \hat{e}_i = 0$ and their sum of squares is $\sum \hat{e}_i^2 = 20.57143$

(f) $\sum x_i \hat{e}_i = 1.71429 - 5.14286 + 6.42857 - 8.57143 - 2.14286 + 7.71429 = 0$

EXERCISE 2.4

(a) If $\beta_1 = 0,$ the simple linear regression model becomes

$$y_i = \beta_2 x_i + e_i$$

(b) Graphically, setting $\beta_1 = 0$ implies the mean of the simple linear regression model $E(y_i) = \beta_2 x_i$ passes through the origin (0, 0).

(c) To save on subscript notation we set $\beta_2 = \beta.$ The sum of squares function becomes

$$S(\beta) = \sum_{i=1}^{N}(y_i - \beta x_i)^2 = \sum_{i=1}^{N}(y_i^2 - 2\beta x_i y_i + \beta^2 x_i^2) = \sum y_i^2 - 2\beta\sum x_i y_i + \beta^2 \sum x_i^2$$

$$= 712 - 2 \times 250\beta + 91\beta^2 = 712 - 500\beta + 91\beta^2$$

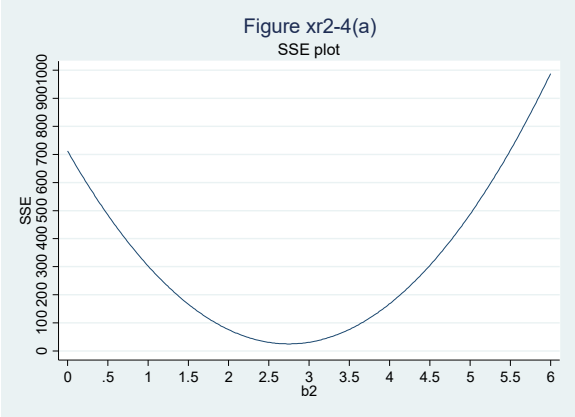


Figure xr2.4(a) Sum of squares for β_2

The minimum of this function is approximately 25 and occurs at approximately $\beta_2 = 2.7$ The significance of this value is that it is the least-squares estimate.

(d) To find the value of β that minimizes $S(\beta)$ we obtain

$$\frac{dS}{d\beta} = -2\sum x_i y_i + 2\beta\sum x_i^2$$

Setting this derivative equal to zero, we have

$$b\sum x_i^2 = \sum x_i y_i \quad \text{or} \quad b = \frac{\sum x_i y_i}{\sum x_i^2}$$

Exercise 2.4 (Continued)

Thus, the least-squares estimate is

$$b_2 = \frac{250}{91} = 2.747253$$

which agrees with the approximate value of 2.7 that we obtained geometrically.

(e)

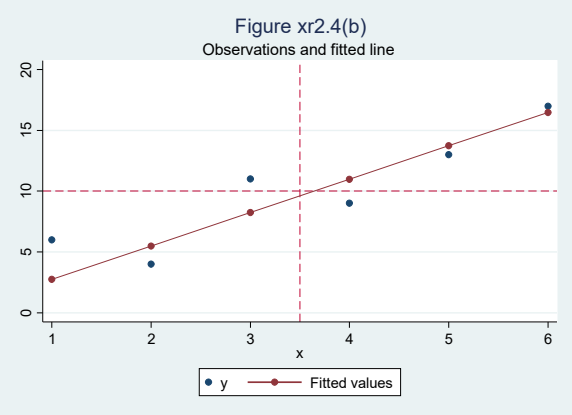


Figure xr2.4(b) Observations and fitted line

The fitted regression line is plotted in Figure xr2.4 (b). Note that the point $(\bar{x}, \bar{y})$ does not lie on the fitted line in this instance.

(f) The least squares residuals, obtained from $\hat{e}_i = y_i - b_2 x_i$ are:

$$\hat{e}_1 = 3.25275 \qquad \hat{e}_2 = -1.49451 \qquad \hat{e}_3 = 2.75824$$

$$\hat{e}_4 = -1.98901 \qquad \hat{e}_5 = -0.73626 \qquad \hat{e}_6 = 0.51648$$

Their sum is $\sum \hat{e}_i = 2.307692.$ Note this value is not equal to zero as it was for $\beta_1 \neq 0.$

(g) $$\sum x_i \hat{e}_i = 3.25275 - 2.98901 + 8.27473 - 7.95604 - 3.68132 + 3.09890 = 0$$

EXERCISE 2.5

(a) The consultant's report implies that the least squares estimates satisfy the following two equations

$$b_1 + 1500b_2 = 10000$$

$$b_1 + 2000b_2 = 12000$$

Solving these two equations yields

$$500b_2 = 2000 \Rightarrow b_2 = \frac{2000}{500} = 4 \qquad b_1 = 4000$$

Therefore, the estimated regression used by the consultant is:

$$\widehat{SALES} = 4000 + 4 \times ADVERT$$

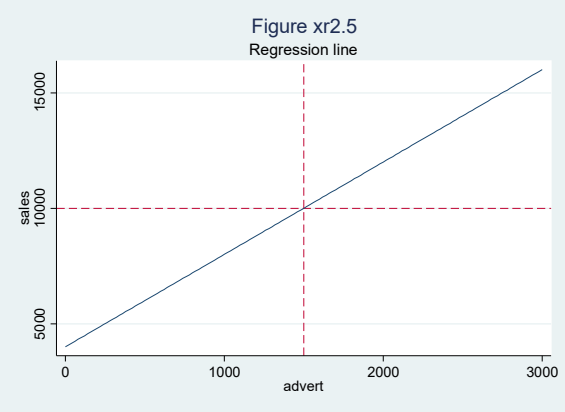


Figure xr2.5 Fitted regression line and mean

EXERCISE 2.6

(a) The intercept estimate $b_1 = -240$ is an estimate of the number of sodas sold when the temperature is 0 degrees Fahrenheit. A common problem when interpreting the estimated intercept is that we often do not have any data points near $x = 0$. If we have no observations in the region where temperature is 0, then the estimated relationship may not be a good approximation to reality in that region. Clearly, it is impossible to sell –240 sodas and so this estimate should not be accepted as a sensible one.

The slope estimate $b_2 = 20$ is an estimate of the increase in sodas sold when temperature increases by 1 Fahrenheit degree. This estimate does make sense. One would expect the number of sodas sold to increase as temperature increases.

(b) If temperature is 80°F, the predicted number of sodas sold is

$$\hat{y} = -240 + 20 \times 80 = 1360$$

(c) If no sodas are sold, $y = 0,$ and

$$0 = -240 + 20x \quad \text{or} \quad x = 12$$

Thus, she predicts no sodas will be sold below 12°F.

(d) A graph of the estimated regression line:

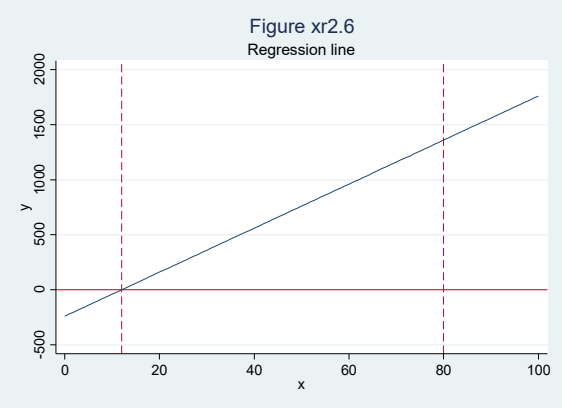


Figure xr2.6 Fitted regression line

EXERCISE 2.7

(a) Since

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2} = 14.24134$$

it follows that

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

(b) The standard error for b_2 is

$$\text{se}(b_2) = \sqrt{\widehat{\text{var}}(b_2)} = \sqrt{0.009165} = 0.09573401$$

Also,

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

Thus,

$$\sum (x_i - \bar{x})^2 = \frac{\hat{\sigma}^2}{\widehat{\text{var}}(b_2)} = \frac{14.24134}{0.009165} = 1553.8833$$

(c) The value $b_2 = 1.02896$ suggests that a 1% increase in the percentage of the population with a bachelor's degree or more will lead to an increase of \$1028.96 in the mean income per capita.

(d) $$b_1 = \bar{y} - b_2\bar{x} = 39.66886 - 1.02896 \times 27.35686 = 11.519745$$

(e) Since $\sum (x_i - \bar{x})^2 = \sum x_i^2 - N\bar{x}^2$, we have

$$\sum x_i^2 = \sum (x_i - \bar{x})^2 + N\bar{x}^2 = 1553.8833 + 51 \times 27.35686^2 = 39722.17$$

(f) For Georgia

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i = 34.893 - 11.519745 - 1.02896 \times 27.5 = -4.9231453$$

EXERCISE 2.8

(a) The sample means from the two data parts are

$$\bar{y}_1 = \sum_{i=1}^{3} y_i / 3 = 7, \bar{x}_1 = \sum_{i=1}^{3} x_i / 3 = 2 \text{ and } \bar{y}_2 = \sum_{i=4}^{6} y_i / 3 = 13, \bar{x}_2 = \sum_{i=4}^{6} x_i / 3 = 5$$

Using these values, we find $\hat{\beta}_{2,mean} = (7-13)/(2-5) = 2$ and $\hat{\beta}_{1,mean} = 10 - 2(3.5) = 3$. The fitted line is shown in Figure xr2.8.

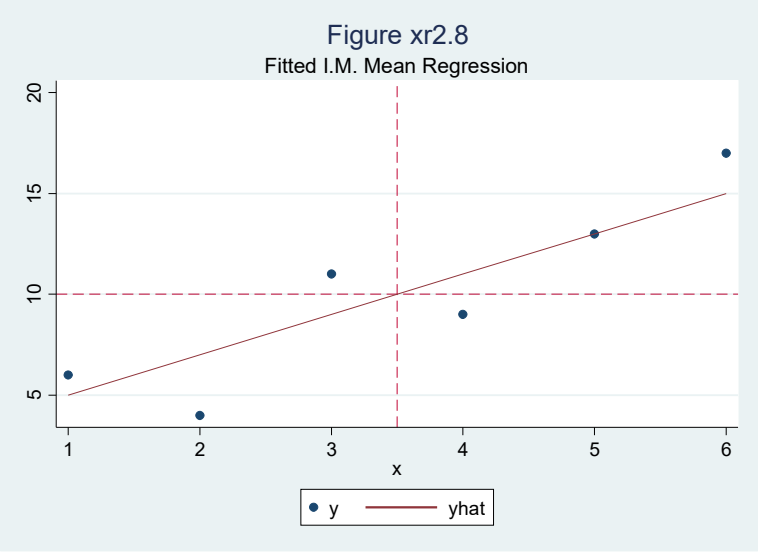


Figure xr2.8 Fitted regression line and mean

(b) The values of the residuals, computed from $\hat{e}_{i,mean} = y_i - \hat{y}_{i,mean} = y_i - \left(\hat{\beta}_{1,mean} + \hat{\beta}_{2,mean} x_i\right)$, are:

x_i	y_i	$\hat{y}_{i,mean}$	$\hat{e}_{i,mean}$	$x_i \hat{e}_{i,mean}$
1	6	6	1	1
2	4	4	−3	−6
3	11	11	2	6
4	9	9	−2	−8
5	13	13	0	0
6	17	17	2	12

The required sums are $\sum_{i=1}^{6} \hat{e}_{i,mean} = 0$ $\sum_{i=1}^{6} x_i \hat{e}_{i,mean} = 5$,

Exercise 2.8 (continued)

(c) The least squares estimates are

$$b_2 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{40}{17.5} = 2.285714$$

$$b_1 = \bar{y} - b_2\bar{x} = 10 - (2.285714) \times 3.5 = 2$$

For the least squares residuals $\sum \hat{e}_i = 0$, $\sum x_i \hat{e}_i = 0$.

(d) The sum of squared residuals from the *mean* regression is $\sum_{i=1}^{6} \hat{e}_{i,mean}^2 = 22$. The sum of the least squares residuals is $\sum \hat{e}_i^2 = 20.57143$. The least squares estimator is designed to provide the smallest value.

EXERCISE 2.9

(a) $E\left(\hat{\beta}_{2,mean} \mid \mathbf{x}\right) = E\left[\left(\bar{y}_2 - \bar{y}_1\right)/\left(\bar{x}_2 - \bar{x}_1\right) \mid \mathbf{x}\right] = \left[1/\left(\bar{x}_2 - \bar{x}_1\right)\right] E\left[\left(\bar{y}_2 - \bar{y}_1\right) \mid \mathbf{x}\right]$

$$E\left[\left(\bar{y}_2 - \bar{y}_1\right) \mid \mathbf{x}\right] = E\left[\bar{y}_2 \mid \mathbf{x}\right] - E\left[\bar{y}_1 \mid \mathbf{x}\right]$$

$$E\left[\bar{y}_2 \mid \mathbf{x}\right] = E\left[\frac{1}{3}\sum_{i=4}^{6} y_i \mid \mathbf{x}\right] = \frac{1}{3}\sum_{i=4}^{6} E\left(y_i \mid \mathbf{x}\right) = \frac{1}{3}\sum_{i=4}^{6}\left(\beta_1 + \beta_2 x_i\right)$$
$$= \frac{1}{3}\left[3\beta_1 + \beta_2 \sum_{i=4}^{6} x_i\right] = \beta_1 + \beta_2 \frac{1}{3}\left(\sum_{i=4}^{6} x_i\right) = \beta_1 + \beta_2 \bar{x}_2$$

Similarly, $E\left[\bar{y}_1 \mid \mathbf{x}\right] = \beta_1 + \beta_2 \bar{x}_1$. Then

$$E\left[\left(\bar{y}_2 - \bar{y}_1\right) \mid \mathbf{x}\right] = E\left[\bar{y}_2 \mid \mathbf{x}\right] - E\left[\bar{y}_1 \mid \mathbf{x}\right] = \left(\beta_1 + \beta_2 \bar{x}_2\right) - \left(\beta_1 + \beta_2 \bar{x}_1\right) = \beta_2\left(\bar{x}_2 - \bar{x}_1\right)$$

Finally,

$$E\left(\hat{\beta}_{2,mean} \mid \mathbf{x}\right) = E\left[\left(\bar{y}_2 - \bar{y}_1\right)/\left(\bar{x}_2 - \bar{x}_1\right) \mid \mathbf{x}\right] = \left[1/\left(\bar{x}_2 - \bar{x}_1\right)\right] E\left[\left(\bar{y}_2 - \bar{y}_1\right) \mid \mathbf{x}\right]$$
$$= \left[1/\left(\bar{x}_2 - \bar{x}_1\right)\right]\beta_2\left(\bar{x}_2 - \bar{x}_1\right) = \beta_2$$

We have shown that conditional on **x** the estimator $\hat{\beta}_{2,mean}$ is unbiased.

(b) Use the law of iterated expectations. $E\left(\hat{\beta}_{2,mean}\right) = E_{\mathbf{x}}\left[E\left(\hat{\beta}_{2,mean} \mid \mathbf{x}\right)\right] = E_{\mathbf{x}}\left(\beta_2\right) = \beta_2$

Because the estimator is conditionally unbiased it is unconditionally unbiased also.

(c)

$$\text{var}\left(\hat{\beta}_{2,mean} \mid \mathbf{x}\right) = \left[1/\left(\bar{x}_2 - \bar{x}_1\right)\right]^2 \text{var}\left[\left(\bar{y}_2 - \bar{y}_1\right) \mid \mathbf{x}\right] = \left[1/\left(\bar{x}_2 - \bar{x}_1\right)\right]^2 \left\{\text{var}\left[\bar{y}_2 \mid \mathbf{x}\right] + \text{var}\left[\bar{y}_1 \mid \mathbf{x}\right]\right\}$$

$$\text{var}\left[\bar{y}_2 \mid \mathbf{x}\right] = \text{var}\left[\frac{1}{3}\sum_{i=4}^{6} y_i \mid \mathbf{x}\right] = \frac{1}{9}\left[\sum_{i=4}^{6} \text{var}\left(y_i \mid \mathbf{x}\right)\right] = \frac{1}{9}\left(3\sigma^2\right) = \sigma^2/3$$

Similarly, $\text{var}\left[\bar{y}_1 \mid \mathbf{x}\right] = \sigma^2/3$. So that

$$\text{var}\left(\hat{\beta}_{2,mean} \mid \mathbf{x}\right) = \left[1/\left(\bar{x}_2 - \bar{x}_1\right)\right]^2 \left\{\text{var}\left[\bar{y}_2 \mid \mathbf{x}\right] + \text{var}\left[\bar{y}_1 \mid \mathbf{x}\right]\right\} = \left[1/\left(\bar{x}_2 - \bar{x}_1\right)\right]^2 \left[\frac{\sigma^2}{3} + \frac{\sigma^2}{3}\right] = \frac{2\sigma^2}{3\left(\bar{x}_2 - \bar{x}_1\right)^2}$$

Exercise 2.9(c) (continued)

We know that $\text{var}\left(\hat{\beta}_{2,mean} \mid \mathbf{x}\right)$ is larger than the variance of the least squares estimator because $\hat{\beta}_{2,mean}$ is a linear estimator. To show this note that

$$\hat{\beta}_{2,mean} = \left(\bar{y}_2 - \bar{y}_1\right)/\left(\bar{x}_2 - \bar{x}_1\right) = \frac{1}{\left(\bar{x}_2 - \bar{x}_1\right)}\left[\left(\frac{\sum_{i=4}^{6} y_i}{3}\right) - \left(\frac{\sum_{i=1}^{3} y_i}{3}\right)\right] = \left[\frac{\sum_{i=4}^{6} y_i}{3\left(\bar{x}_2 - \bar{x}_1\right)} - \frac{\sum_{i=1}^{3} y_i}{3\left(\bar{x}_2 - \bar{x}_1\right)}\right]$$

$$= \sum_{i=1}^{6} a_i y_i$$

Where $a_1 = a_2 = a_3 = \dfrac{-1}{3\left(\bar{x}_2 - \bar{x}_1\right)}$ and $a_4 = a_5 = a_6 = \dfrac{1}{3\left(\bar{x}_2 - \bar{x}_1\right)}$

Furthermore $\hat{\beta}_{2,mean}$ is an unbiased estimator. From the Gauss-Markov theorem we know that the least squares estimator is the "best" linear unbiased estimator, the one with the smallest variance. Therefore, we know that $\text{var}\left(\hat{\beta}_{2,mean} \mid \mathbf{x}\right)$ is larger than the variance of the least squares estimator.

EXERCISE 2.10

(a) If $\beta_2 = 0$ the model reduces to

$$y_i = \beta_1 + e_i$$

(b) Graphically, setting $\beta_2 = 0$ implies the regression model is a horizontal line when plotted against x_i at the height β_1.

(c)

$$S(\beta_1) = \sum_{i=1}^{N}(y_i - \beta_1)^2 = \sum_{i=1}^{N}\left(y_i^2 + \beta_1^2 - 2y_i\beta_1\right) = \sum_{i=1}^{N} y_i^2 + N\beta_1^2 - 2\beta_1\sum_{i=1}^{N} y_i$$
$$= 712 + 6\beta_1^2 - 2(60)\beta_1$$

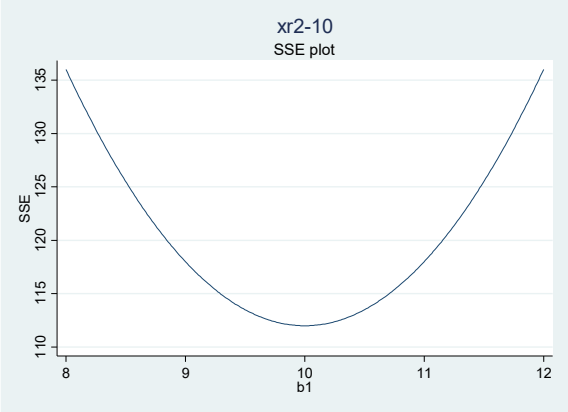


Figure xr2.10 Sum of squares for β_1

The minimum appears to be at $b_1 = 10$

(d) To find the minimum, we find the value of β_1 such that the slope of the sum of squares function is zero.

$$dS(\beta_1)/d\beta_1 = 2N\beta_1 - 2\sum_{i=1}^{N} y_i = 0$$

Solving, we find

$$\hat{\beta}_1 = \left(\sum_{i=1}^{N} y_i\right)/N = \bar{y}$$

To ensure that this is a minimum the second derivative must be positive. $d^2S(\beta_1)/d\beta_1^2 = 2N > 0$ as long as $N > 0$, so that we have at least one data point.

Exercise 2.10 (Continued)

(e) The least-squares estimate is

$$\hat{\beta}_1 = \left(\sum_{i=1}^{N} y_i\right) / N = 60/6 = 10$$

It is the same given the accurate graph.

$$S\left(\hat{\beta}_1\right) = \sum_{i=1}^{N}\left(y_i - \hat{\beta}_1\right)^2 = \sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2 = 112$$

(f) Since . The sum of squared residuals from the least squares regression including the explanatory variable is $S\left(b_1, b_2\right) = \sum_{i=1}^{N}\left(y_i - b_1 - b_2 x_i\right)^2 = 20.5714$. We are able to "fit" the model to the data much better by including the explanatory variable.

EXERCISE 2.11

(a) We estimate that each additional \$100 per month income is associated with an additional 52 cents per person expenditure, on average, on food away from home. If monthly income is zero, we estimate that household will spend an average of \$13.77 per person on food away from home.

(b) $\hat{y} = 13.77 + 0.52(20) = 24.17$. We predict that household with \$2000 per month income will spend on average \$24.17 per person on food away from home.

(c) In this linear relationship, the elasticity is $\hat{\varepsilon} = b_2\left(x/\hat{y}\right) = 0.52(20/24.17) = 0.43$. We estimate that a 1% increase in income will increase expected food expenditure by 0.43% per person.

(d) In this log-linear relationship, the elasticity is $\hat{\varepsilon} = 0.007(20) = 0.14$.

(e)

$$\hat{y} = \exp\left(3.14+0.007(20)\right) = 26.58, \quad d\hat{y}/dx = \exp\left(3.14+0.007(20)\right)(0.007) = 0.1860$$
$$\hat{y} = \exp\left(3.14+0.007(30)\right) = 28.50, \quad d\hat{y}/dx = \exp\left(3.14+0.007(30)\right)(0.007) = 0.1995$$

It is increasing at an increasing rate. This is shown on Figure xr2.11. Also, the second derivative, the rate of change of the first derivative is $d^2\hat{y}/dx^2 = \exp(3.14+0.007x)(0.007)^2 > 0$. A positive second derivative means that the function is increasing at an increasing rate for all values of x.

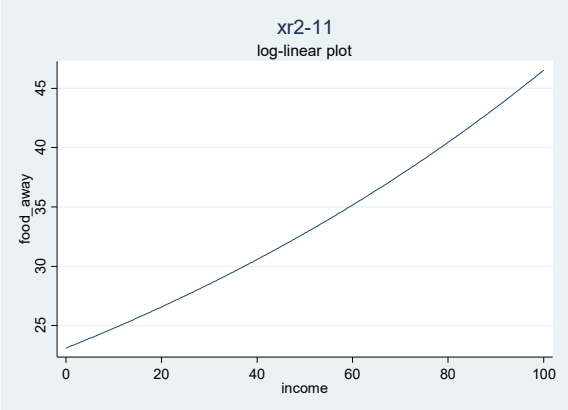


Figure xr2.11 Log-linear plot

(f) The number of zeros is 2334 – 2005 = 329. The reason for the reduction in the number of observations is that the logarithm of zero is undefined and creates a missing data value. The software throws out the row of data when it encounters a missing value when doing its calculations.

EXERCISE 2.12

(a) The model estimates for the two values of x are

$$\hat{y} = \begin{cases} 44.96 + 30.41 = 75.37 & \text{if} \quad x = 1 \\ 44.96 & if \quad x = 0 \end{cases}$$

We estimate that a household without an advanced degree holder will spend on average \$44.96 per month on food away per person. We estimate that a household with an advanced degree holder will spend on average \$75.37 per month on food away per person. The coefficient on x is the difference between the average expenditures per month on food away for households with an advanced degree holder and households without an advanced degree holder. The intercept is the average expenditure per month on food away for a household without an advanced degree holder.

(b) In this sample, for households with a member having an advanced degree, their average expenditure on food away from home is \$75.37 per person.

(c) In this sample, for households without a member having an advanced degree, their average expenditure on food away from home is \$44.96 per person.

EXERCISE 2.13

(a) We estimate that each additional 1000 *FTE* students increase real total academic cost per student by \$266, holding all else constant. The intercept suggests if there were no students the real total academic cost per student would be \$14,656. This is meaningless in the pure sense because there are no universities with zero students. However, it is true that many of the costs of a university, related to research and the functioning of hospitals, etc., carry on and are "fixed costs" with respect to student population.

(b) $\widehat{ACA\_LSU} = 14.656 + 0.266(27.950) = 22.0907$. We predict the total cost per student at LSU in 2011 to be \$21,403.

(c) The least squares residual for LSU is $\hat{e} = 21.403 - 22.0907 = -0.6877$. The regression prediction is too high, an over-prediction of \$687.70.

(d) The least squares regression passes through the point of the means, so that $\overline{ACA} = 14.656 + 0.266(22.84577) = 20.732975$. The average *ACA* is \$20,732.98 for these 141 universities.

EXERCISE 2.14

(a) The elasticity at a point on the fitted regression line is $\hat{\varepsilon} = b_2(\bar{x}/\bar{y})$. We are given the estimate of the slope and the mean wage in the non-urban area. The fitted least squares line passes through the point of the means, so that $\overline{WAGE} = -4.88 + 1.80\overline{EDUC} \Rightarrow \overline{EDUC} = \left[\overline{WAGE} - (-4.88)\right]/1.80 = 13.678$. The elasticity at the means is then $\hat{\varepsilon} = b_2(\bar{x}/\bar{y}) = 1.80(13.678/19.74) = 1.247$.

(b) We are given the mean level of *EDUC*. Therefore $\overline{WAGE} = -10.76 + 2.46\overline{EDUC} = 22.8928$. The elasticity is then $\hat{\varepsilon} = b_2(\bar{x}/\bar{y}) = 2.46(13.68/22.8928) = 1.47$. The variance of the elasticity is $\widehat{\mathrm{var}}(\hat{\varepsilon} \mid \mathbf{x}) = \widehat{\mathrm{var}}\left[b_2(\bar{x}/\bar{y}) \mid \mathbf{x}\right] = (\bar{x}/\bar{y})^2 \widehat{\mathrm{var}}(b_2 \mid \mathbf{x})$. The standard error of the elasticity is then $\mathrm{se}(\hat{\varepsilon}) = \sqrt{\widehat{\mathrm{var}}(\hat{\varepsilon} \mid \mathbf{x})} = (\bar{x}/\bar{y})\sqrt{\widehat{\mathrm{var}}(b_2 \mid \mathbf{x})} = (\bar{x}/\bar{y})\mathrm{se}(b_2)$. The standard error of the estimated slope is 0.16, so the standard error of the elasticity is $\mathrm{se}(\hat{\varepsilon}) = (\bar{x}/\bar{y})\mathrm{se}(b_2) = (13.68/22.8928)0.16 = 0.0956$.

(c) For the urban area $\widehat{WAGE} = -10.76 + 2.46EDUC$. Given *EDUC* = 12 the predicted wage is $\widehat{WAGE} = -10.76 + 2.46(12) = 18.76$. Given *EDUC* = 16 the predicted wage is $\widehat{WAGE} = -10.76 + 2.46(16) = 28.60$.

For the non-urban area, $\widehat{WAGE} = -4.88 + 1.80EDUC$. Given *EDUC* = 12 the predicted wage is $\widehat{WAGE} = -4.88 + 1.80(12) = 16.72$. Given *EDUC* = 16 the predicted wage is $\widehat{WAGE} = -4.88 + 1.80(16) = 23.92$.

EXERCISE 2.15

(a) The EZ estimator can be written as

$$b_{EZ} = \frac{y_2 - y_1}{x_2 - x_1} = \left(\frac{1}{x_2 - x_1}\right) y_2 - \left(\frac{1}{x_2 - x_1}\right) y_1 = \sum k_i y_i$$

where

$$k_1 = \frac{-1}{x_2 - x_1}, \quad k_2 = \frac{1}{x_2 - x_1}, \quad \text{and} \quad k_3 = k_4 = \ldots = k_N = 0$$

Thus, b_{EZ} is a linear estimator.

(b) Taking expectations yields

$$E(b_{EZ}) = E\left[\frac{y_2 - y_1}{x_2 - x_1}\right] = \frac{1}{x_2 - x_1} E(y_2) - \frac{1}{x_2 - x_1} E(y_1)$$

$$= \frac{1}{x_2 - x_1}(\beta_1 + \beta_2 x_2) - \frac{1}{x_2 - x_1}(\beta_1 + \beta_2 x_1)$$

$$= \frac{\beta_2 x_2}{x_2 - x_1} - \frac{\beta_2 x_1}{x_2 - x_1} = \beta_2 \left(\frac{x_2}{x_2 - x_1} - \frac{x_1}{x_2 - x_1}\right) = \beta_2$$

Thus, b_{EZ} is an unbiased estimator.

(c) The variance is given by

$$\text{var}(b_{EZ}) = \text{var}(\sum k_i y_i) = \sum k_i^2 \, \text{var}(e_i) = \sigma^2 \sum k_i^2$$

$$= \sigma^2 \left(\frac{1}{(x_2 - x_1)^2} + \frac{1}{(x_2 - x_1)^2}\right) = \frac{2\sigma^2}{(x_2 - x_1)^2}$$

(d) If $e_i \sim N(0, \sigma^2)$, then $b_{EZ} \sim N\left[\beta_2, \frac{2\sigma^2}{(x_2 - x_1)^2}\right]$

Exercise 2.15 (continued)

(e) To convince E.Z. Stuff that var(b_2) < var(b_{EZ}), we need to show that

$$\frac{2\sigma^2}{(x_2-x_1)^2} > \frac{\sigma^2}{\sum(x_i-\bar{x})^2} \quad \text{or that} \quad \sum(x_i-\bar{x})^2 > \frac{(x_2-x_1)^2}{2}$$

Consider

$$\frac{(x_2-x_1)^2}{2} = \frac{\left[(x_2-\bar{x})-(x_1-\bar{x})\right]^2}{2} = \frac{(x_2-\bar{x})^2+(x_1-\bar{x})^2-2(x_2-\bar{x})(x_1-\bar{x})}{2}$$

Thus, we need to show that

$$2\sum_{i=1}^{N}(x_i-\bar{x})^2 > (x_2-\bar{x})^2+(x_1-\bar{x})^2-2(x_2-\bar{x})(x_1-\bar{x})$$

or that

$$(x_1-\bar{x})^2+(x_2-\bar{x})^2+2(x_2-\bar{x})(x_1-\bar{x})+2\sum_{i=3}^{N}(x_i-\bar{x})^2 > 0$$

or that

$$\left[(x_1-\bar{x})+(x_2-\bar{x})\right]^2+2\sum_{i=3}^{N}(x_i-\bar{x})^2 > 0.$$

This last inequality clearly holds. Thus, b_{EZ} is not as good as the least squares estimator. Rather than prove the result directly, as we have done above, we could also refer Professor E.Z. Stuff to the Gauss Markov theorem.

EXERCISE 2.16

(a) The model is a simple regression model because it can be written as $y = \beta_1 + \beta_2 x + e$ where $y = r_j - r_f$, $x = r_m - r_f$, $\beta_1 = \alpha_j$ and $\beta_2 = \beta_j$.

(b) The estimates are in the table below

Firm	GE	IBM	FORD	MSFT	DIS	XOM
$b_1 = \hat{\alpha}_j$	−0.000959 (0.00442)	0.00605 (0.00483)	0.00378 (0.0102)	0.00325 (0.00604)	0.00105 (0.00468)	0.00528 (0.00354)
$b_2 = \hat{\beta}_j$	1.148 (0.0895)	0.977 (0.0978)	1.662 (0.207)	1.202 (0.122)	1.012 (0.0946)	0.457 (0.0716)
N	180	180	180	180	180	180

Standard errors in parentheses

The stocks Ford, GE, and Microsoft are relatively aggressive with Ford being the most aggressive with a beta value of $b_2 = 1.662$. The others are relatively defensive with Exxon-Mobil being the most defensive with a beta value of $b_2 = 0.457$.

(c) All estimates of the α_j are close to zero and are therefore consistent with finance theory. The fitted regression line and data scatter for Microsoft are plotted in Figure xr2.15.

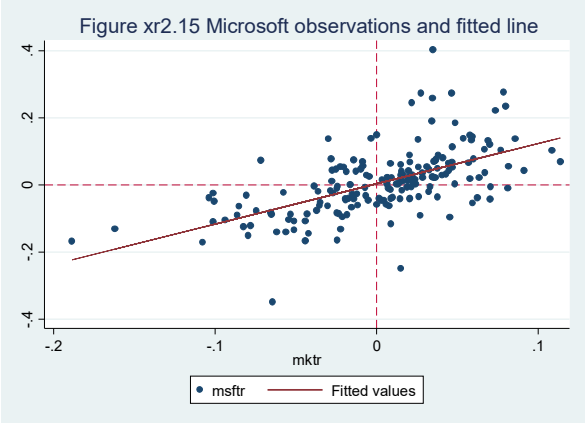


Fig. xr2.15 Scatter plot of Microsoft and market rate

(d) The estimates for β_j given $\alpha_j = 0$ are as follows.

Firm	GE	IBM	FORD	MSFT	DIS	XOM
$b_2 = \hat{\beta}_j$	1.147 (0.0891)	0.984 (0.0978)	1.667 (0.206)	1.206 (0.122)	1.013 (0.0942)	0.463 (0.0717)

Standard errors in parentheses

The restriction $\alpha_j = 0$ has led to small changes in the $\hat{\beta}_j$; it has not changed the aggressive or defensive nature of the stock.

EXERCISE 2.17

(a)

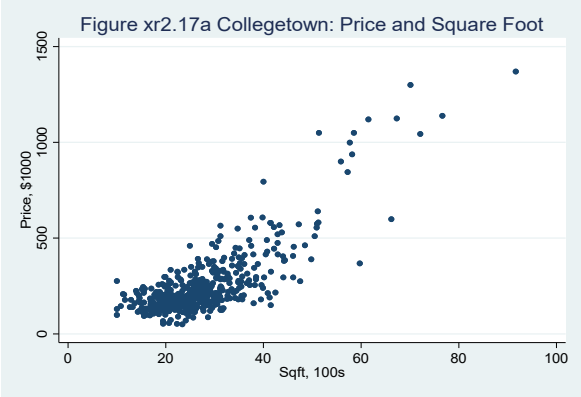


Figure xr2.17(a) Price (in $1,000s) against square feet for houses (in 100s)

(b) The fitted linear relationship is

$$\widehat{PRICE} = -115.4236 + 13.40294SQFT$$
$$(se) \qquad (13.0882) \quad (0.4492)$$

We estimate that an additional 100 square feet of living area will increase the expected home price by $13,402.94 holding all else constant. The estimated intercept −115.4236 would imply that a house with zero square feet has an expected price of $−115,423.60. This estimate is not meaningful in this example. The reason is that there are no data values with a house size near zero.

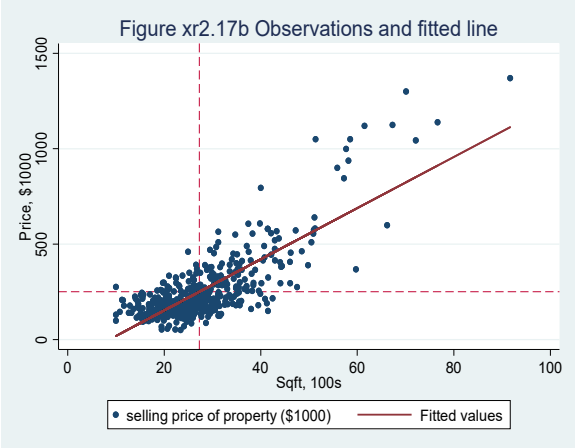


Figure xr2.17(b) Observations and fitted line

Exercise 2.17 (continued)

(c) The fitted quadratic model is

$$\widehat{PRICE} = 93.5659 + 0.1845SQFT^2$$
$$\text{(se)} \quad (6.0722) \quad (0.00525)$$

$$d\left(\widehat{PRICE}\right)\Big/dSQFT = 2\hat{\alpha}_2 SQFT$$

The marginal effect is . For a house with 2000 square feet of living area the estimated marginal effect is 2(0.1845)20 = 7.3808. We estimate that an additional 100 square feet of living area for a 2000 square foot home will increase the expected home price by $7,380.80 holding all else constant.

(d)

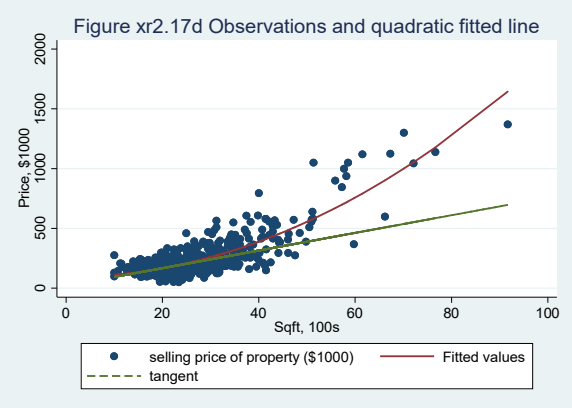


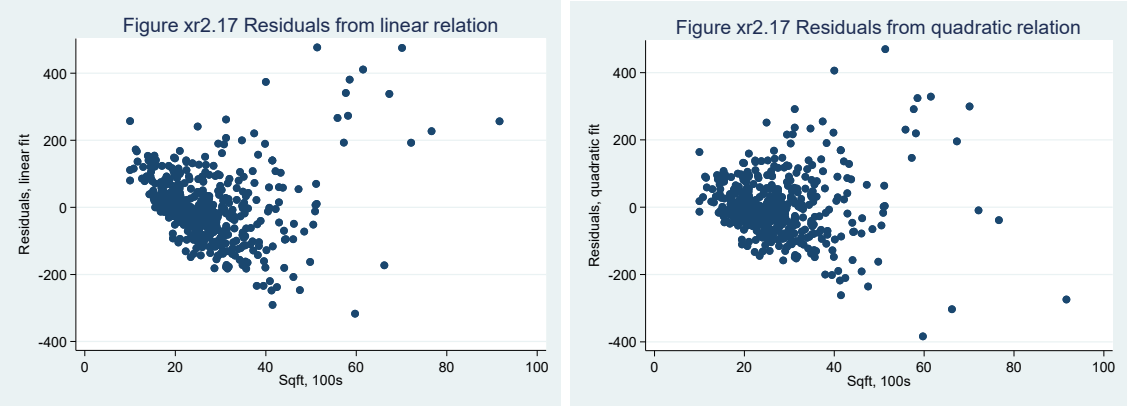
Figure xr2.17(d) Observations and quadratic fitted line

(e) The estimated elasticity is

$$\hat{\varepsilon} = slope \times \frac{SQFT}{\widehat{PRICE}} = \left(2\hat{\alpha}_2 SQFT\right) \times \frac{SQFT}{\widehat{PRICE}} = 7.3808 \times \frac{20}{167.3735} = 0.882$$

For a 2000 square foot house, we estimate that a 1% increase in house size will increase expected price by 0.882%, holding all else fixed.

(f) The residual plots are



Figures xr2.17(f) Residuals from linear and quadratic relations

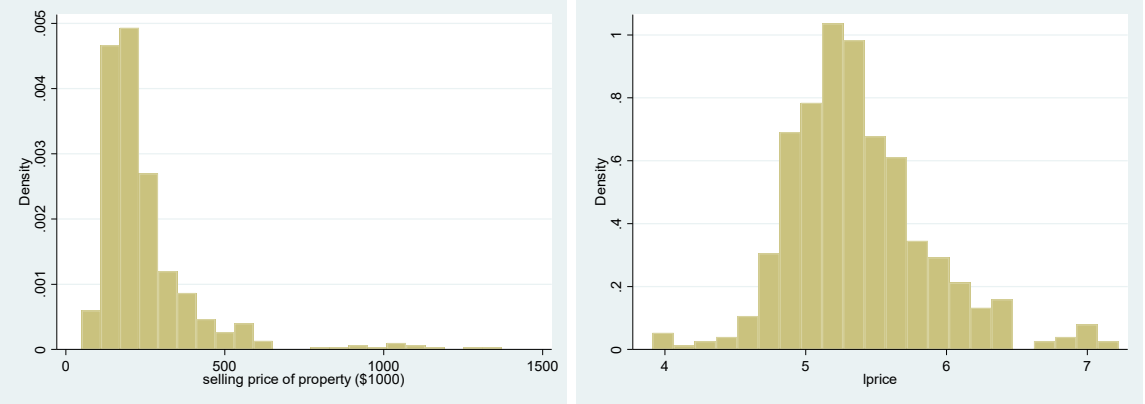
Exercise 2.17(f) (continued)

In both models, the residual patterns do not appear random. The variation in the residuals increases as *SQFT* increases, suggesting that the homoskedasticity assumption may be violated.

(g) The sum of square residuals linear relationship is 5,262,846.9. The sum of square residuals for the quadratic relationship is 4,222,356.3. In this case the quadratic model has the lower *SSE*. The lower *SSE* means that the data values are closer to the fitted line for the quadratic model than for the linear model.

EXERCISE 2.18

(a) The histograms for *PRICE* and $\ln(PRICE)$ are below. The distribution of PRICE is skewed with a long tail to the right. The distribution of $\ln(PRICE)$ is more symmetrical



Figures xr2.18(a) Histograms for *PRICE* and ln(*PRICE*)

(b) The estimated log-linear model is

$$\widehat{\ln(PRICE)} = 4.3939 + 0.0360SQFT$$
$$\text{(se)} \qquad (0.0433)\ (0.0015)$$

The estimated slope can be interpreted as telling us that a 100 square foot increase in house size increases predicted price by approximately 3.6%, holding all else fixed. The estimated intercept tells us little as is. But $\exp(4.3939) = 80.953$ suggests that the predicted price of a zero square foot house is $80,953. This estimate has little meaning because in the sample there are no houses with zero square feet of living area.

For a 2000 square foot house the predicted price is

$$\widehat{PRICE} = \exp\left[\widehat{\ln(PRICE)}\right] = \exp(4.3939 + 0.0360 \times 20) = 166.4601$$

The estimated slope is

$$\frac{d\left(\widehat{PRICE}\right)}{dSQFT} = \hat{\gamma}_2 \widehat{PRICE} = 0.0360 \times 166.4601 = 6.0$$

Exercise 2.18 (continued)

The predicted price of a house with 2000 square feet of living area is $166,460.10. We estimate that 100 square foot size increase for a house with 2000 square feet of living area will increase price by $6,000, holding all else fixed. This is the slope of the tangent line in the figure below.

Figure xr2.18(b) Observations and log-linear fitted line

(c) The residual plot is shown below. The residual plot is a little hard to interpret because there are few very large homes in the sample. The variation in the residuals appears to diminish as house size increases, but that interpretation should not be carried too far.

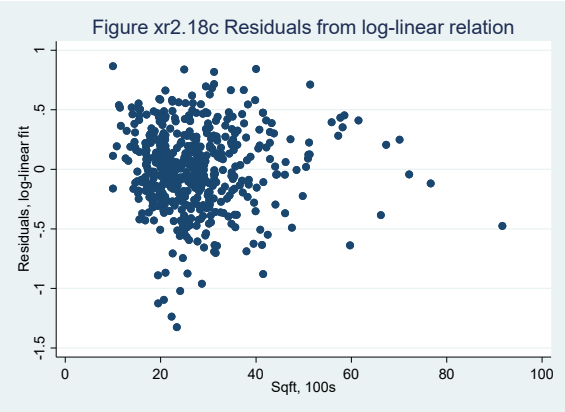


Figure xr2.18(c) Residuals from log-linear relation

(d) The summary statistics show that there are 189 houses close to LSU and 311 houses not close to LSU in the sample. The mean house price is $10,000 larger for homes close to LSU, and the homes close to LSU are slightly smaller, by about 100 square feet. The range of the data is smaller for the homes close to LSU, and the standard deviation for those homes is half the standard deviation of homes not close to LSU.

Exercise 2.18 (continued)

	CLOSE = 1		*CLOSE* = 0	
STATS	*PRICE*	*SQFT*	*PRICE*	*SQFT*
N	189	189	311	311
mean	256.6298	26.59011	246.3518	27.70267
sd	108.5878	8.735512	200.3505	11.05563
min	110	10	50	10
max	900	59.73	1370	91.67

(e) The estimates for the two sub-samples are

		C	*SQFT*	*N*	*SSE*
CLOSE = 1	Coeff	4.7637	0.0269	189	14.2563
	Std. err.	(0.0645)	(0.0023)		
CLOSE = 0	Coeff	4.2019	0.0402	311	36.6591
	Std. err.	(0.0528)	(0.0018)		

For homes close to LSU we estimate that an additional 100 square feet of living space will increase predicted price by about 2.69% and for homes not close to LSU about 4.02%.

(f) Assumption SR1 implies that the data are drawn from the same population. So the question is, are homes close to LSU and homes not close to LSU in the same population? Based on our limited sample, and using just a simple, one variable, regression model it is difficult to be very specific. The estimated regression coefficients for the sub-samples are different, the question we will be able to address later is "Are they significantly different." Just looking at the magnitudes is not a statistical test.

EXERCISE 2.19

(a)

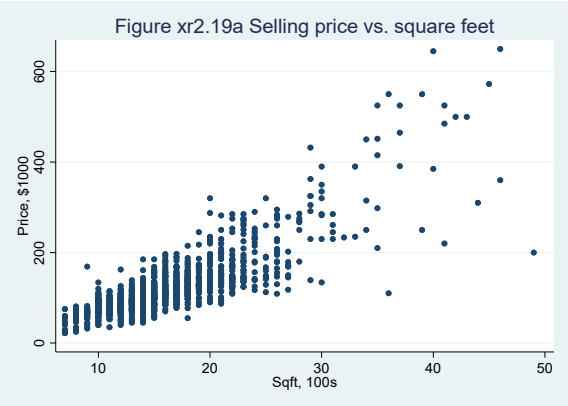


Figure xr2.19(a) Scatter plot of selling price and living area

(b) The estimated linear relationship is

$$\widehat{SPRICE} = -35.9664 + 9.8934 LIVAREA$$
$$\text{(se)} \qquad (3.3085) \quad (0.1912)$$

We estimate that an additional 100 square feet of living area will increase the expected home price by \$9,893.40 holding all else constant. The estimated intercept −35.9664 would imply that a house with zero square feet has an expected price of \$−35,966.40. This estimate is not meaningful in this example. The reason is that there are no data values with a house size near zero.

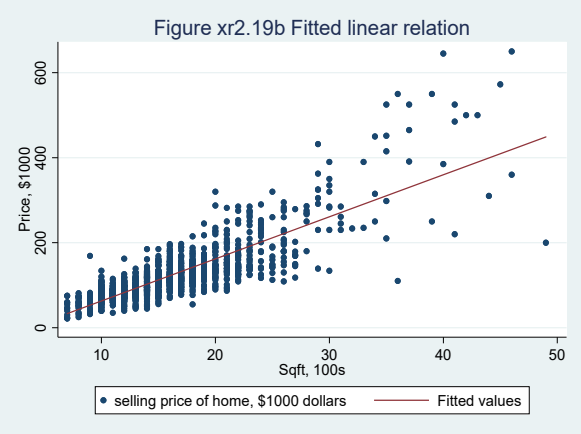


Figure xr2.19(b) Fitted linear relation

Exercise 2.19 (continued)

(c) The estimated quadratic equation is

$$\widehat{SPRICE} = 56.4572 + 0.2278 LIVAREA^2$$
$$\text{(se)} \qquad (1.6955) \quad (0.0043)$$

The marginal effect is $d\left(\widehat{SPRICE}\right)/dLIVAREA = 2\hat{\alpha}_2 LIVAREA$. For a house with 1500 square feet of living area the estimated marginal effect is 2(0.2278)15 = 6.834. We estimate that an additional 100 square feet of living area for a 1500 square foot home will increase the expected home price by $6,834 holding all else constant.

(d)

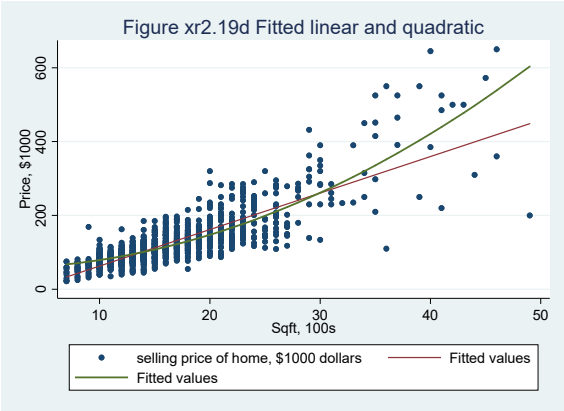


Figure xr2.19(d) Fitted linear and quadratic relations

The sum of squared residuals for the linear relation is *SSE* = 1,879,826.9948. For the quadratic model the sum of squared residuals is *SSE* = 1,795,092.2112. In this instance, the sum of squared residuals is smaller for the quadratic model, one indicator of a better fit.

(e) If the quadratic model is in fact "true," then the results and interpretations we obtain for the linear relationship are incorrect, and may be misleading.

EXERCISE 2.20

(a) The estimates are reported in the table below. Of the 1200 homes in the sample, 69 are on large lots. None of the estimated intercepts has a useful interpretation because no houses in the samples have near zero living area. The estimated slope coefficients suggest that for houses on large lots, a 100 square foot increase in house size will increase expected price by \$9,763.20, holding all else fixed. For houses not on large lots the estimate is \$9,289.70, about \$500 less than for houses on large lots. The full sample estimate is \$9,893.40, which is between the estimates for homes on large lots and not on large lots.

		C	*LIVAREA*	*N*	*SSE*
LGELOT = 1	Coeff	5.0199	9.7632	69	490972.8
	Std. err.	(25.6709)	(1.0014)		
LGELOT = 0	Coeff	−28.7476	9.2897	1131	1271831.3
	Std. err.	(3.1374)	(0.1884)		
All	Coeff	−35.9664	9.8934	1200	1879827.0
	Std. err.	(3.3085)	(0.1912)		

(b) The estimates are reported in the table below. Of the 1200 homes in the sample, 69 are on large lots. None of the estimated intercepts has a useful interpretation because no houses in the samples have near zero living area. The estimated coefficients of $LIVAREA^2$ are somewhat different for houses on large lots and those not on large lots.

		C	*LIVAREA*	*N*	*SSE*
LGELOT = 1	Coeff	120.7025	0.1728	69	538400.4
	Std. err.	(16.6150)	(0.0192)		
LGELOT = 0	Coeff	52.2575	0.2368	1131	1128980.3
	Std. err.	(1.5431)	(0.0044)		
All	Coeff	56.4572	0.2278	1200	1795092.2
	Std. err.	(1.6955)	(0.0043)		

To evaluate the differences, it is useful to calculate the slope, $2\alpha_2 LIVAREA$. For homes with 2000 square feet of living area the estimated slopes are

Large lots: 6.91128; Not Large lots: 9.471073; All lots: 9.112585

That is, we estimate that for a 2000 square foot home, 100 more square feet of living area, the expected price will increase by \$6,911 for homes on large lots, \$9,471 for homes not on large lots, and \$9,113 based on all lots. The difference between the marginal effect of house size on house price for large lots and not large lots is substantial. The estimate using all the data is close to the estimate on lots that are not large because most of the data comes from such lots.

Exercise 2.20 (continued)

(c)

$$E(SPRICE \mid LGELOT) = \eta_1 + \eta_2 LGELOT = \begin{cases} \eta_1 + \eta_2 & LGELOT = 1 \\ \eta_1 & LGELOT = 0 \end{cases}$$

In this model η_1 is the expected price of houses not on large lots, and $\eta_1 + \eta_2$ is the expected price of houses on large lots. Inserting the estimates, we obtain

$$\widehat{SPRICE} = 117.9487 + 116.2940 LGELOT = \begin{cases} 234.2428 & \text{if} \quad LGELOT = 1 \\ 117.9487 & \text{if} \quad LGELOT = 0 \end{cases}$$

That is, the expect price of houses on lots that are not large is $117,948.70 and the expected price of houses on large lots is $234,242.80. The expected price on large lots is about twice the expected price of houses on lots that are not large.

(d) Assumption SR1 requires that the data pairs in the sample are from the same population. If there are substantial differences between homes on lots and those not on large lots then SR1 will be violated meaning that estimation results on a pooled sample are not reliable. The result in part (c) indicates that there may be large differences between homes on these types of lots. What will be of interest later, in Chapter 3, is whether the difference is statistically significant.

EXERCISE 2.21

(a)

$$\widehat{SPRICE} = 152.6144 - 0.9812\,AGE$$
$$\text{(se)} \qquad (3.3473)(0.0949)$$

We estimate that a house that is new, $AGE = 0$, will have expected price \$152,614.40. We estimate that each additional year of age will reduce expected price by \$981.20, other things held constant. The expected selling price for a 30-year-old house is $\widehat{SPRICE} = 152.6144 - 0.9812(30) = \$123{,}177.70$.

(b)

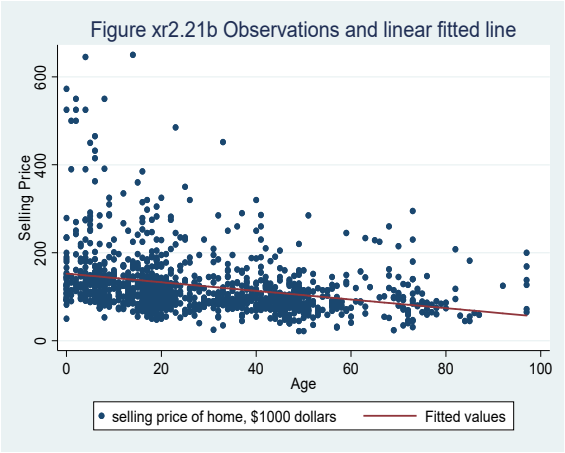


Figure xr2.21(b) Observations and linear fitted line

The data show an inverse relationship between house prices and age. The data on newer houses is not as close to the fitted regression line as the data for older homes.

(c)

$$\widehat{\ln(SPRICE)} = 4.9283 - 0.0075\,AGE$$
$$\text{(se)} \qquad (0.0205)(0.0006)$$

We estimate that each additional year of age reduces expected price by about 0.75%, holding all else constant.

Exercise 2.21 (continued)

(d)

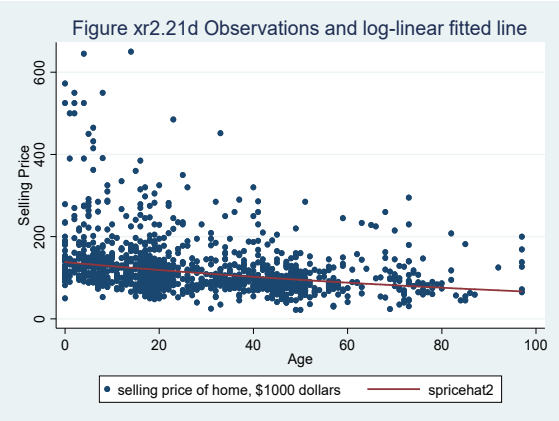


Figure xr2.21(c) Observations and log-linear fitted line

The fitted log-linear model is not too much different than the fitted linear relationship.

(e) The expected selling price of a house that is 30 years old is $\widehat{SPRICE} = \exp(4.9283 - 0.0075 \times 30) = \$110{,}370.32$. This is about $13,000 less than the prediction based on the linear relationship.

(f) Based on the plots and visual fit of the estimated regression lines it is difficult to choose between the two models. For the estimated linear relationship $\sum_{i=1}^{1200}\left(SPRICE - \widehat{SPRICE}\right)^2 = 5{,}580{,}871$. For the log-linear model $\sum_{i=1}^{1200}\left(SPRICE - \widehat{SPRICE}\right)^2 = 5{,}727{,}332$. The sum of squared differences between the data and fitted values is smaller for the estimated linear relationship, by a small margin. This is one way to measure how well a model fits the data. In this case, based on fit alone, we might choose the linear relationship rather than the log-linear relationship.

EXERCISE 2.22

(a) The regression model is $TOTALSCORE = \beta_1 + \beta_2 SMALL + e$. Under the model assumptions

$$E(TOTALSCORE \mid SMALL) = \beta_1 + \beta_2 SMALL = \begin{cases} \beta_1 + \beta_2 & \text{if} \quad SMALL = 1 \\ \beta_1 & \text{if} \quad SMALL = 0 \end{cases}$$

Thus β_1 is the expected total score in regular sized classes, and $\beta_1 + \beta_2$ is the expected total score in small classes. The difference β_2 is an estimate of the difference in performance in small and regular sized classes. The model estimates are given in Table xr2-22a, Model (1).

Table xr2-22a

		C	*SMALL*	*N*	*SSE*
(1) *TOTALSCORE*	Coeff	916.4417	12.1753	775	4300389
	Std. err.	(3.6746)	(5.3692)		
(2) *READSCORE*	Coeff	432.6650	6.9245	775	705200
	Std. err.	(1.4881)	(2.1743)		
(3) *MATHSCORE*	Coeff	483.7767	5.2508	775	1910009
	Std. err.	(2.4489)	(3.5783)		

The estimated equation using a sample of small and regular classes (where *AIDE* = 0) is

$$\widehat{TOTALSCORE} = 916.442 + 12.175 SMALL$$

Comparing a sample of small and regular classes, we find students in regular classes achieve an average total score of 916.442 while students in small classes achieve an average of $916.442 + 12.175 = 928.617$. This is a 1.33% increase. This result suggests that small classes have a positive impact on learning, as measured by higher totals of all achievement test scores.

(b) The estimated equations using a sample of small and regular classes are given in Table xr2-22a as Models (2) and (3)

$$\widehat{READSCORE} = 432.665 + 6.925 SMALL$$

$$\widehat{MATHSCORE} = 483.77 + 5.251 SMALL$$

Students in regular classes achieve an average reading score of 432.7 while students in small classes achieve an average of 439.6. This is a 1.60% increase. In math students in regular classes achieve an average score of 483.77 while students in small classes achieve an average of 489.0. This is a 1.08% increase. These results suggests that small class sizes also have a positive impact on learning math and reading.

Exercise 2.22 (continued)

(c) The estimated equations using a sample of regular classes and regular classes with a full-time teacher aide (when $SMALL = 0$) are given in Table xr2-22b

Table xr2-22b

		C	*AIDE*	*N*	*SSE*
(4) *TOTALSCORE*	Coeff	916.4417	4.3065	837	4356550
	Std. err.	(3.5586)	(4.9940)		
(5) *READSCORE*	Coeff	432.6650	2.8714	837	733335
	Std. err.	(1.4600)	(2.0489)		
(6) *MATHSCORE*	Coeff	483.7767	1.4351	837	1907234
	Std. err.	(2.3546)	(3.3043)		

$$\widehat{TOTALSCORE} = 916.442 + 4.31\,AIDE$$

Students in regular classes without a teacher aide achieve an average total score of 916.4 while students in regular classes with a teacher aide achieve an average total score of 920.7. This is an increase of 0.47%. These results suggest that having a full-time teacher aide has a small impact on learning outcomes as measured by totals of all achievement test scores.

(d) The estimated equations using a sample of regular classes and regular classes with a full-time teacher aide are

$$\widehat{READSCORE} = 432.67 + 2.87\,AIDE$$

$$\widehat{MATHSCORE} = 483.78 + 1.44\,AIDE$$

The effect of having a teacher aide on learning is 0.66% for reading and 0.30% for math. These increases are smaller than the increases provided by smaller classes.

EXERCISE 2.23

(a)

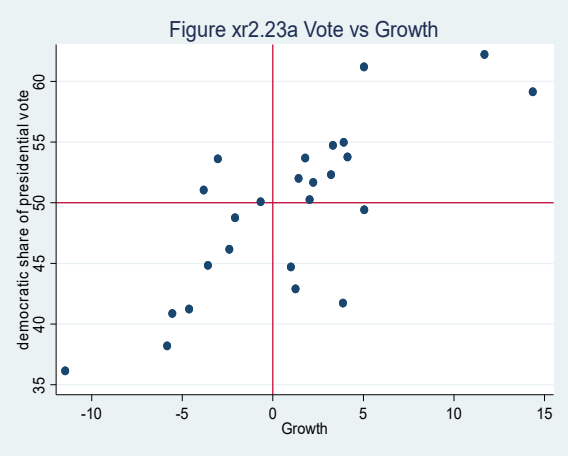


Figure xr2.23(a) Vote against Growth

There appears to be a positive association between *VOTE* and *GROWTH*.

(b) The estimated equation for 1916 to 2012 is

$$\widehat{VOTE} = 48.6160 + 0.9639GROWTH$$
$$\text{(se)} \quad (0.9043) \quad (0.1658)$$

The coefficient 0.9639 suggests that for a 1 percentage point increase in a favorable growth rate of *GDP* in the 3 quarters before the election there is an estimated increase in the share of votes of the democratic party of 0.9639 percentage points.

We estimate, based on the fitted regression intercept, that that the Democratic party's expected vote is 48.62% when the growth rate in *GDP* is zero. This suggests that when there is no real *GDP* growth, the Democratic party is expected to lose the popular vote. A graph of the fitted line and data is shown in the following figure.

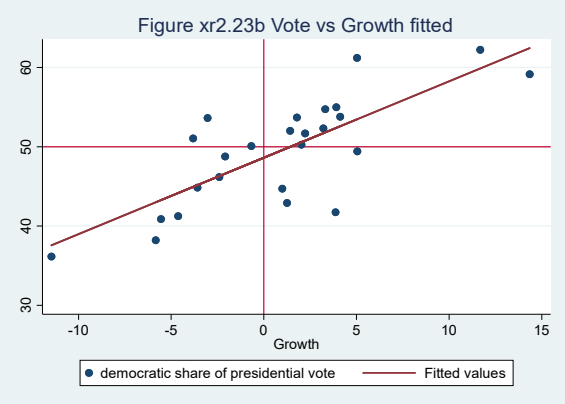


Figure xr2.23(a) Vote vs Growth fitted

Exercise 2.23 (continued)

(c) In 2016 the actual growth rate in GDP was 0.97% and the predicted expected vote in favor of the Democratic party was $\widehat{VOTE} = 48.6160 + 0.9639(0.97) = 49.55$, or 49.55%. The actual popular vote in favor of the Democratic party was 50.82%.

(d) The figure below shows a plot of *VOTE* against *INFLATION*. It is difficult to see if there is positive or inverse relationship.

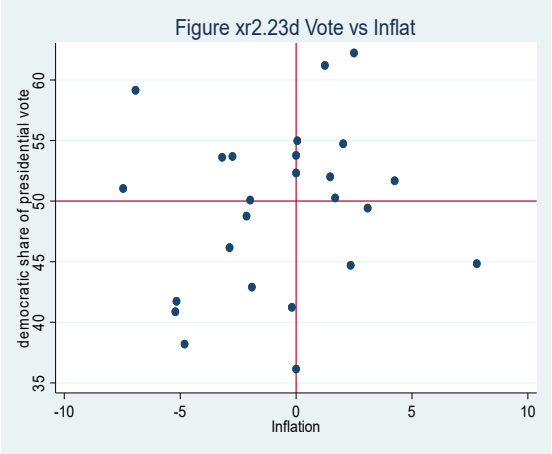


Figure xr2.23(d) Vote against Inflat

(e) The estimated equation (plotted in the figure below) is

$$\widehat{VOTE} = 49.6229 + 0.2616INFLATION$$
$$\text{(se)} \quad (1.4188) \quad (0.3907)$$

We estimate that a 1 percentage point increase in inflation during the party's first 15 quarters increases the share of Democratic party's vote by 0.2616 percentage points. The estimated intercept suggests that when inflation is at 0% for that party's first 15 quarters, the expected share of votes won by the Democratic party is 49.6%.

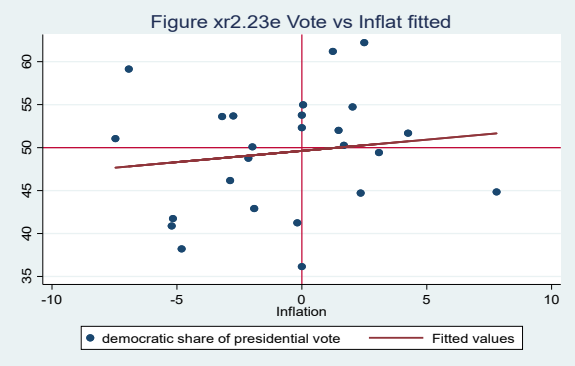


Figure xr2.23(e) Vote vs Inflat fitted

Exercise 2.23 (continued)

(f) The actual inflation value in the 2016 election was 1.42%. The predicted vote in favor of the Democratic candidate (Clinton) was $\widehat{VOTE} = 49.6229 + 0.2616(1.42) = 49.99$, or 49.99%.

EXERCISE 2.24

(a) The histogram shows a very skewed distribution

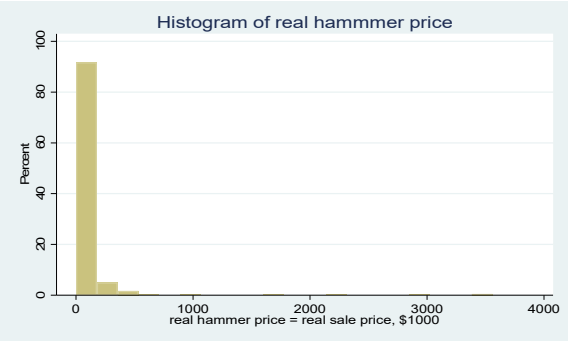


Figure xr2.24(a) Histogram of real hammer price

The sample mean, based on 422 works that sold is \$78,682. But the 25th, 50th and 75th percentiles are \$2,125, \$13,408 and \$46,102 respectively; all less than the mean which is inflated due to some extreme values. The two largest values are \$3,559,910 and \$3,560,247.

(b)

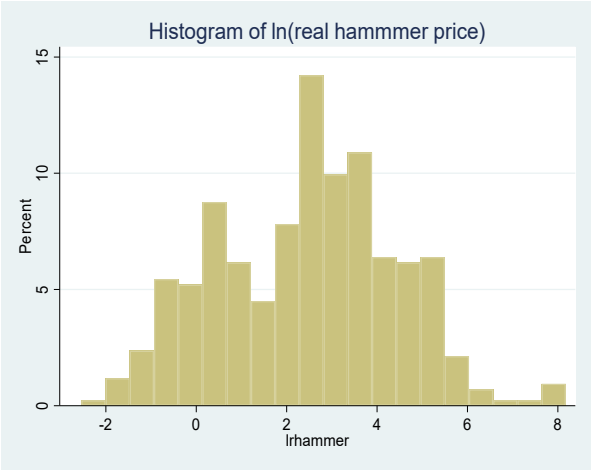


Figure xr2.24(b) Histogram of ln(real hammer price)

$\ln(RHAMMER)$ is not "bell shaped" but it is hardly skewed at all (skewness close to zero). It has been "regularized" by the transformation. This is not necessary for regression, but as you will see in Chapter 3 having data closer to normal makes analysis nice.

Exercise 2.24 (continued)

(c)

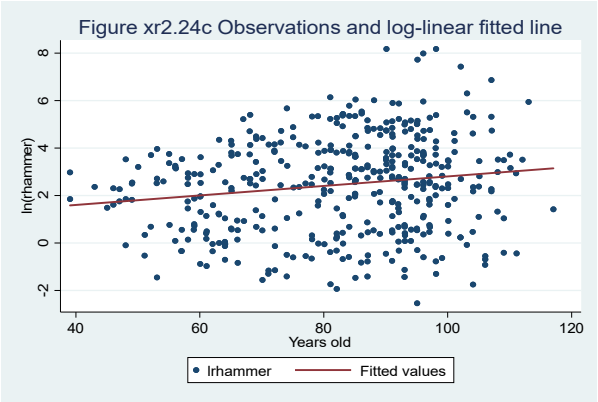


Figure xr2.24(c) Observations and log-linear fitted line

The data scatter shows a positive association between $\ln(RHAMMER)$ and the age of the painting. The fitted OLS regression line passes through the center of the data, as it is designed to do.

(d)

$$\widehat{\ln(RHAMMER)} = 0.8000 + 0.0201YEARSOLD$$
$$\text{(se)} \qquad (0.5022)\ (0.0060)$$

We estimate that each additional year of age increases predicted hammer price by about 2%, other factors held constant.

(e)

$$E\left[\ln(RHAMMER) \mid DREC\right] = \alpha_1 + \alpha_2 DREC = \begin{cases} \alpha_1 + \alpha_2 & \text{if} \quad DREC = 1 \\ \alpha_1 & \text{if} \quad DREC = 0 \end{cases}$$

In this model, the expected $\ln(RHAMMER)$ is α_1 during non-recession and is $\alpha_1 + \alpha_2$ in a recession. The estimated regression function during a recession is $2.5547 - 1.0420 = 1.5127$. We estimate that during a non-recessionary period the average hammer price is \$12,867, using $\exp(2.5547)$, and during a recession we predict the average price to be \$4,539, using $\exp(1.5127)$, more than a 50% reduction.

EXERCISE 2.25

(a)

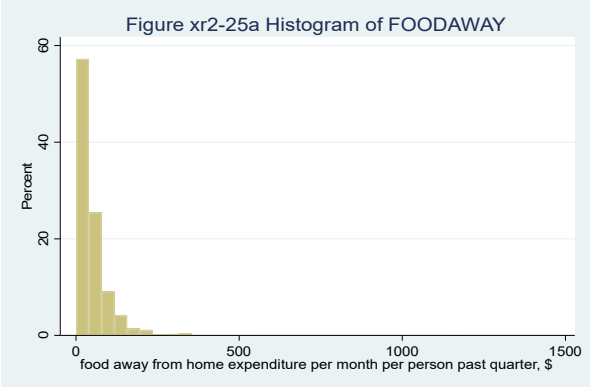


Figure xr2.25(a) Histogram of foodaway

The mean of the 1200 observations is 49.27, the 25th, 50th and 75th percentiles are 12.04, 32.56 and 67.60. The histogram figure shows a very skewed distribution, with a mean that is larger than the median. 50% of households spend $32.56 per person or less during a quarter.

(b) Households with a member with an advanced degree spend an average of about $25 more per person than households with a member with a college degree, but not advanced degree. Households with a member with a college degree, but not advanced degree, spend an average of about $9 more per person than households with no members with a college or advanced degree.

	N	Mean	Median
ADVANCED = 1	257	73.15	48.15
COLLEGE = 1	369	48.60	36.11
NONE	574	39.01	26.02

Exercise 2.25 (continued)

(c)

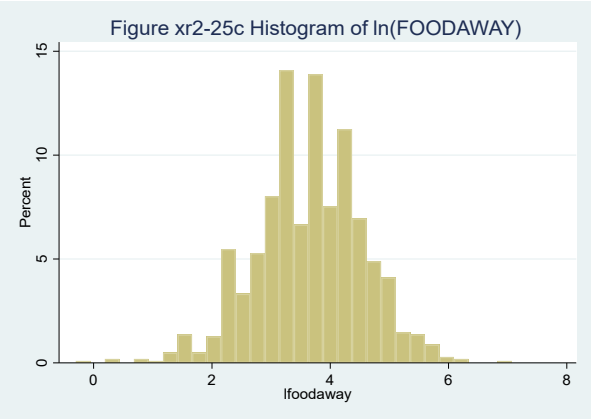


Figure xr2.25(c) Histogram of ln(foodaway)

The histogram of ln(*FOODAWAY*) is much less skewed. There are 178 fewer values of ln(*FOODAWAY*) because 178 households reported spending $0 on food away from home per person, and ln(0) is undefined. It creates a "missing value" which software cannot use in the regression. If any variable has a missing value in either y_i or x_i the entire observation is deleted from regression calculations.

(d) The estimated model is

$$\widehat{\ln(FOODAWAY)} = 3.1293 + 0.0069 INCOME$$
$$\text{(se)} \qquad (0.0566) \quad (0.0007)$$

We estimate that each additional $100 household income increases food away expenditures per person of about 0.69%, other factors held constant.

(e)

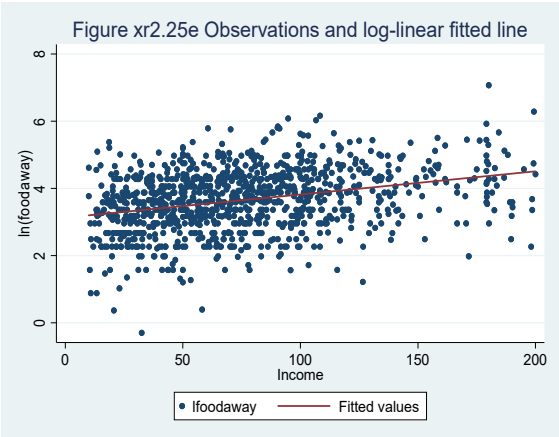


Figure xr2.25(e) Observations and log-linear fitted line

The plot shows a positive association between ln(*FOODAWAY*) and *INCOME*s.

Exercise 2.25 (continued)

(f)

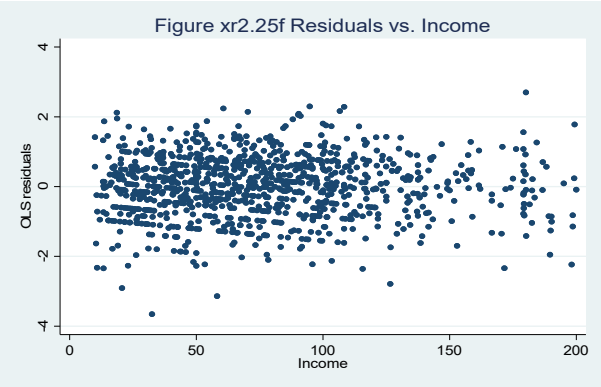


Figure xr2.25(f) Residuals vs. income

The OLS residuals do appear randomly distributed with no obvious patterns. There are fewer observations at higher incomes, so there is more "white space."

EXERCISE 2.26

(a)

$$\widehat{FOODAWAY} = 13.7138 + 0.4929 INCOME$$
$$(\text{se}) \qquad (3.5805) \quad (0.0430)$$

We estimate that a household with zero income in the past quarter will spend an average of $13.71 per member on food away from home. This estimate should not be taken too seriously because there are no households with income near zero in the sample. We estimate that each additional $100 household income increases expected food expenditure away from home by 49 cents, holding other factors fixed.

(b)

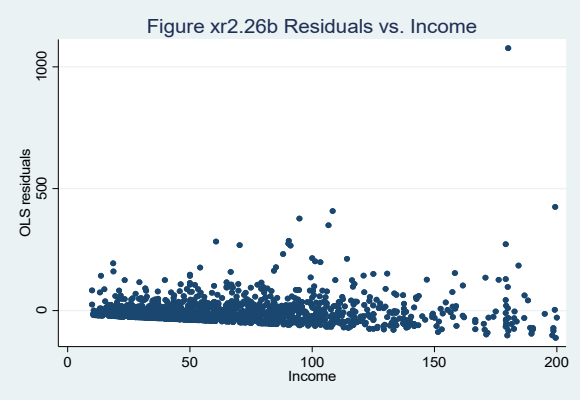


Figure xr2.25(e) Observations and log-linear fitted line

The residuals do not appear randomly distributed. There is a "spray" pattern with a concentration of observations along the lower edge.

(c)

$$\widehat{FOODAWAY} = 42.7616 + 30.3933 ADVANCED$$
$$(\text{se}) \qquad (2.0876) \quad (4.5110)$$

We estimate that the expected per person expenditure for households with no advanced degree holder is $42.76. We estimate that the expected per person expenditure for households with an advanced degree holder is $73.15, which is $30.39 higher.

(d) The sample means for the two groups are shown below. The mean of the observations with *ADVANCED* = 0 is the estimated intercept in (c), and the estimated mean of the observations with *ADVANCED* = 1 is $30.39 higher, the estimated coefficient of advanced in part (c).

	N	Mean
ADVANCED = 1	257	73.15494
ADVANCED = 0	943	42.76161

EXERCISE 2.27

(a)

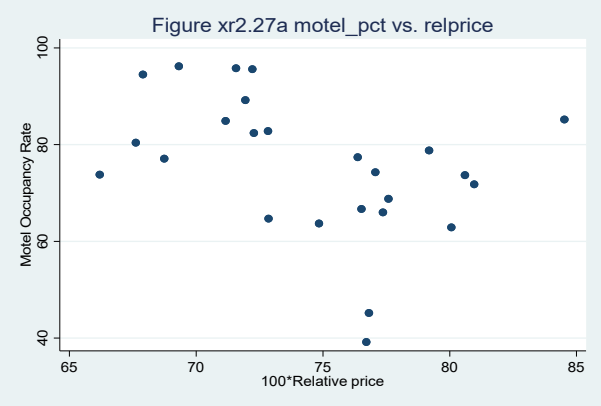


Figure xr2.27(a) Motel_pct vs. 100relprice

There seems to be an inverse association between relative price and occupancy rate.

(b)

$$\widehat{MOTEL\_PCT}_t = 166.6560 - 1.2212 RELPRICE_t$$
$$\text{(se)} \qquad (43.5709)\ (0.5835)$$

Based economic reasoning we anticipate a negative coefficient for RELPRICE. The slope estimate is interpreted as saying, the expected model occupancy rate falls by 1.22% given a 1% increase in relative price, other factors held constant.

Exercise 2.27 (continued)

(c)

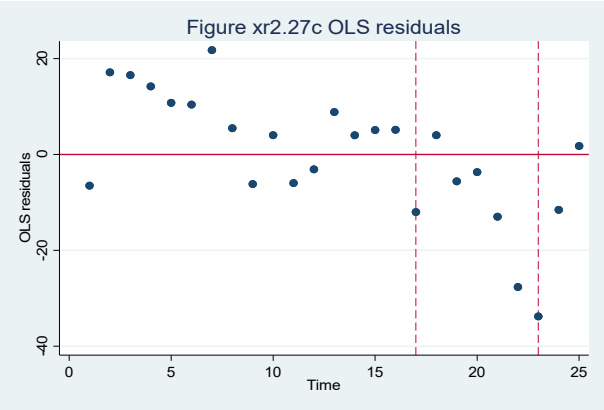


Figure xr2.27(c) OLS residuals

The residuals are scattered about zero for the first 16 observations but for observations 17-23 all but one of the residuals is negative. This suggests that the occupancy rate was lower than predicted by the regression model for these dates. Randomly scattered time series residuals should not have strings of consecutive observations with the same sign.

(d)

$$MOTEL\_PCT_t = 79.3500 - 13.2357 REPAIR_t$$
$$\text{(se)} \qquad (3.1541) \quad (5.9606)$$

We estimate that during the non-repair period the expected occupancy rate is 79.35%. During the repair period, the expected occupancy rate is estimated to fall by 13.24%, other things held constant, to 66.11%.

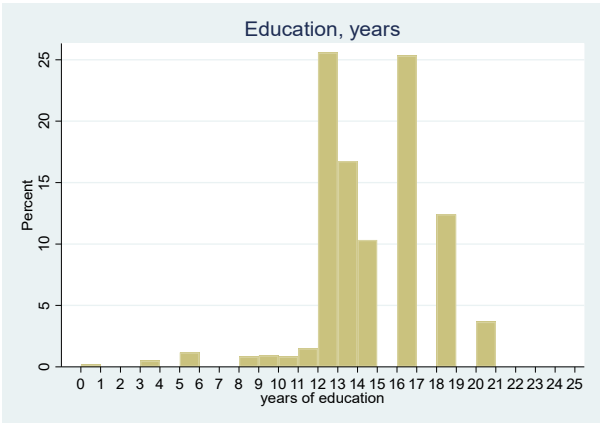
EXERCISE 2.28

(a)

variable	N	mean	median	min	max	skewness	kurtosis
WAGE	1200	23.64	19.3	3.94	221.1	2.9594	27.5787

Figure xr2.28(a1) Histogram and statistics for *WAGE*

The observations for *WAGE* are skewed to the right indicating that most of the observations lie between the hourly wages of 5 to 50, and that there is a smaller proportion of observations with an hourly wage greater than 50. Half of the sample earns an hourly wage of more than $19.30 per hour, with the average being $23.64 per hour. The maximum earned in this sample is $221.10 per hour and the least earned in this sample is $3.94 per hour.



variable	N	mean	median	min	max	skewness	kurtosis
EDUC	1200	14.20	14	0	21	−.45625	4.95745

Figure xr2.28(a2) Histogram and statistics for *EDUC*

307 people had 12 years of education, implying that they finished their education at the end of high school. There are a few observations at less than 12, representing those who did not complete high school. The spike at 16 years describes those 304 who completed a 4-year college degree, while those at 18 and 21 years represent a master's degree, and further education such as a PhD, respectively. Spikes at 13 and 14 years are people who had one or two years at college.

Exercise 2.28 (continued)

(b) The estimated model is

$$\widehat{WAGE} = -10.4000 + 2.3968EDUC$$
$$\text{(se)} \qquad (1.9624) \quad (0.1354)$$

The coefficient 2.3968 represents the estimated increase in the expected hourly wage rate for an extra year of education. The coefficient −10.4 represents the estimated wage rate of a worker with no years of education. It should not be considered meaningful as it is not possible to have a negative hourly wage rate.

(c)

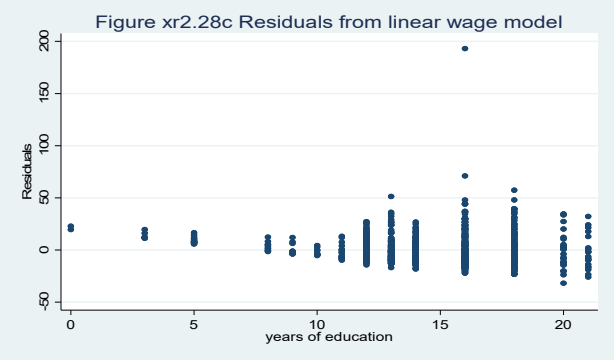


Figure xr2.28(c) Residuals from linear wage model

The residuals are plotted against education in Figure xr2.28(c). There is a pattern evident; as *EDUC* increases, the magnitude of the residuals also increases, suggesting that the error variance is larger for larger values of *EDUC*—a violation of assumption SR3. If the assumptions SR1-SR5 hold, there should not be any patterns evident in the residuals.

(b) The estimated model equations, including the one from part (b), are given in Table xr2-28

Table xr2-28

			C	*EDUC*	*N*	*SSE*
part (b)	all	Coeff	−10.4000	2.3968	1200	220062.3
		Std. err.	(1.9624)	(0.1354)		
part (c)	male	Coeff	−8.2849	2.3785	672	144901.4
		Std. err.	(2.6738)	(0.1881)		
	female	Coeff	−16.6028	2.6595	528	69610.5
		Std. err.	(2.7837)	(0.1876)		
	white	Coeff	−10.4747	2.4178	1095	207901.2
		Std. err.	(2.0806)	(0.1430)		
	black	Coeff	−6.2541	1.9233	105	11369.7
		Std. err.	(5.5539)	(0.3983)		

The white equation is obtained from those workers who are neither black nor Asian. From the results, we can see that an extra year of education increases the expected wage rate of a

white worker more than it does for a black worker. And an extra year of education increases the expected wage rate of a female worker more than it does for a male worker.

Exercise 2.28 (continued)

(e) The estimated quadratic equation is

$$\widehat{WAGE} = 4.9165 + 0.0891EDUC^2$$
$$\text{(se)} \quad (1.0919)(0.0049)$$

The marginal effect is $d(\widehat{WAGE})/dEDUC = 2\hat{\alpha}_2 EDUC$. For a person with 12 years of education, the estimated marginal effect of an additional year of education on expected wage is 2(0.0891)(12) = 2.1392. That is, an additional year of education for a person with 12 years of education is expected to increase wage by \$2.14. For a person with 16 years of education, the marginal effect of an additional year of education is 2(0.0891)(16) = 2.8523. An additional year of education for a person with 16 years of education is expected to increase wage by \$2.85. The linear model in (b) suggested that an additional year of education is expected to increase wage by \$2.40 regardless of the number of years of education attained. That is, the rate of change was constant. The quadratic model suggests that the effect of an additional year of education on wage increases with the level of education already attained.

(f)

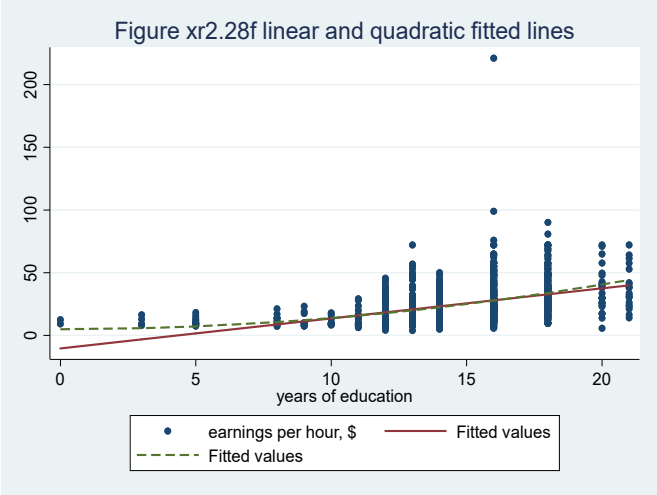
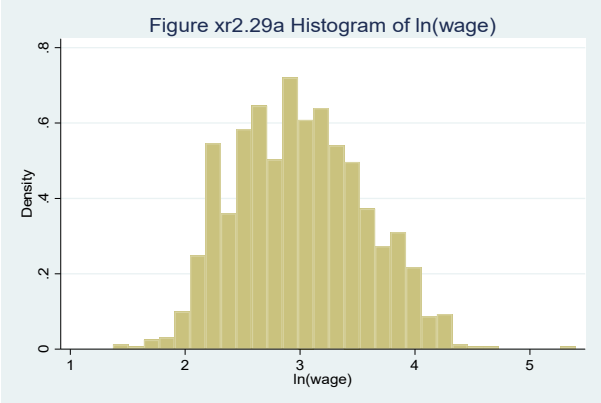


Figure xr2.28(f) Quadratic and linear equations for wage on education

The quadratic model appears to fit the data slightly better than the linear equation, especially at lower levels of education.

EXERCISE 2.29

(a)



variable	N	mean	median	min	max	skewness	kurtosis
ln(*WAGE*)	1200	2.9994	2.9601	1.3712	5.3986	0.2306	2.6846

Figure xr2.29(a) Histogram and statistics for ln(*WAGE*)

The histogram shows the distribution of ln(*WAGE*) to be almost symmetrical. Note that the mean and median are similar, which is not the case for skewed distributions. The skewness coefficient is not quite zero. Similarly, the kurtosis is not quite three, as it should be for a normal distribution.

(b) The OLS estimates are

$$\widehat{\ln(WAGE)} = 1.5968 + 0.0987 EDUC$$
$$\text{(se)} \qquad (0.0702)\ (0.0048)$$

We estimate that each additional year of education predicts a 9.87% higher wage, all else held constant.

(c) The antilogarithm is $\widehat{WAGE} = \exp\left[\widehat{\ln(WAGE)}\right] = \exp(1.5968 + 0.0987 EDUC)$. For someone with 12 years of education the predicted value is $\widehat{WAGE} = \exp(1.5968 + 0.0987 \times 12) = 16.1493$ and for someone with 16 years of education it is $\widehat{WAGE} = \exp(1.5968 + 0.0987 \times 16) = 23.9721$.

(d) The marginal effect in the log-linear model $\ln(y) = \beta_1 + \beta_2 x$, ignoring the error term, is $dy/dx = \beta_2 \exp(\beta_1 + \beta_2 x)$. For individuals with 12 and 16 years of education, respectively, these values are \$1.5948 and \$2.3673. These are the estimated marginal effects of education on expected wage in this log-linear model.

Exercise 2.29 (continued)

(e)

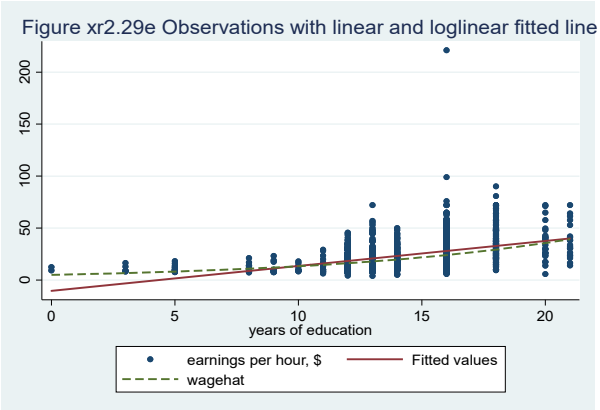


Figure xr2.29(e) Observations with linear and loglinear fitted lines

The log-linear model fits the data better at low levels of education.

(f) A more objective measure of fit is $\sum\left(WAGE_i - \widehat{WAGE}_i\right)^2$. For the log-linear model this value is 228,573.5 and for the linear model 220,062.3. Based on this measure the linear model fits the data better than the linear model.

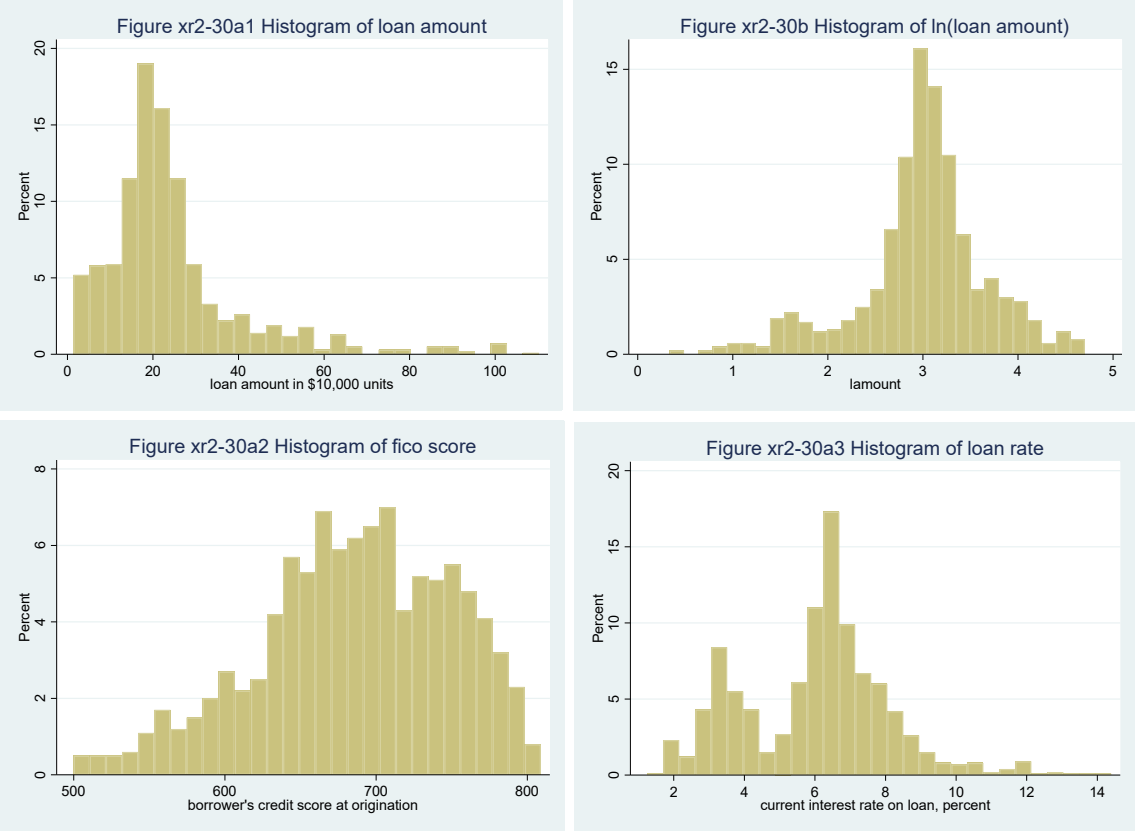
EXERCISE 2.30

(a)

variable	N	mean	p50	min	max	skewness	kurtosis	p10	p90
AMOUNT	1000	24.46	20.8	1.4	110.3	2.018	8.458	7.994	45.7
FICO	1000	686	688.5	500	809	−0.4233	2.713	596.5	767
RATE	1000	6.024	6.25	1.25	14.4	0.2543	3.454	3.125	8.387
TERM30	1000	0.853	1	0	1	−1.994	4.975	0	1

The average amount borrowed is \$244,600. The 90th percentile FICO score is 767. The median interest rate paid was 6.25%. 85.3% of the loans were for 30 years.

(b) The empirical distribution of the loan amount is skewed with a long tail to the right. The empirical distribution for ln(*AMOUNT*) is less noticeably skewed. The skewness coefficient is −0.6341 and kurtosis is 4.3028 so the distribution is far from normal. The FICO score ranges from 500 to 800 and has a bit of left skew. The loan rate is "bi-modal" (two modes) with the most common rates about 3.1% and 6.5%.



Figures xr2.30(b) Histograms

Exercise 2.30 (continued)

(c)

$$\widehat{AMOUNT} = -4.9607 + 0.0429FICO$$
$$\text{(se)} \qquad (5.5517) \quad (0.0081)$$

For each additional point on the FICO score we predict loan amount will increase by $429, holding other factors fixed.

$$\widehat{\ln(AMOUNT)} = 2.4153 + 0.0008FICO$$
$$\text{(se)} \qquad (0.2293) \quad (0.0003)$$

For each additional point on the FICO score we predict loan amount will increase by 0.08%, holding other factors fixed.

(d)

$$\widehat{AMOUNT} = 35.4844 - 1.8306RATE$$
$$\text{(se)} \qquad (1.5669) \quad (0.2459)$$

For each one percent increase in the mortgage rate we predict the amount borrowed will fall by $18,306 other factors held constant.

$$\widehat{\ln(AMOUNT)} = 3.7202 - 0.1211RATE$$
$$\text{(se)} \qquad (0.0611) \quad (0.0096)$$

For each one percent increase in the mortgage rate we predict the amount borrowed will fall by 12.11%, other factors held constant.

(e)

$$\widehat{AMOUNT} = 17.8401 + 7.7576TERM30$$
$$\text{(se)} \qquad (1.3481) \quad (1.4597)$$

There are 853 loans with 30-year terms, and the average borrowed is $255,976.40. For the 147 loans of something other than 30-year terms the average borrowed is $178,400.80. In the regression model, the estimated intercept is the average amount borrowed when *TERM30* = 0. The estimated coefficient of *TERM30* is the difference between the amounts borrowed when *TERM30* = 0 and when *TERM30* = 1.