CHAPTER 2

Exercise Solutions

EXERCISE 2.1

(a)

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
0	6	-2	4	3.6	−7.2
1	2	-1	1	−0.4	0.4
2	3	0	0	0.6	0
3	1	1	1	−1.4	−1.4
4	0	2	4	−2.4	−4.8
$\sum x_i =$ 10	$\sum y_i =$ 12	$\sum(x_i-\bar{x}) =$ 0	$\sum(x_i-\bar{x})^2 =$ 10	$\sum(y-\bar{y}) =$ 0	$\sum(x-\bar{x})(y-\bar{y}) =$ −13

$\bar{x}=2, \quad \bar{y}=2.4$

(b) $$b_2=\frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}=-\frac{13}{10}=-1.3$$

b_2 is the estimated slope of the fitted line.

$$b_1=\bar{y}-b_2\bar{x}=2.4-(-1.3)\times 2=5$$

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

(c) $$\sum_{i=1}^{5}x_i^2=0^2+1^2+2^2+3^2+4^2=30$$

$$\sum_{i=1}^{5}x_iy_i=0\times 6+1\times 2+2\times 3+3\times 1+4\times 0=11$$

$$\sum_{i=1}^{5}x_i^2-N\bar{x}^2=30-5\times 2^2=10=\sum_{i=1}^{5}(x_i-\bar{x})^2$$

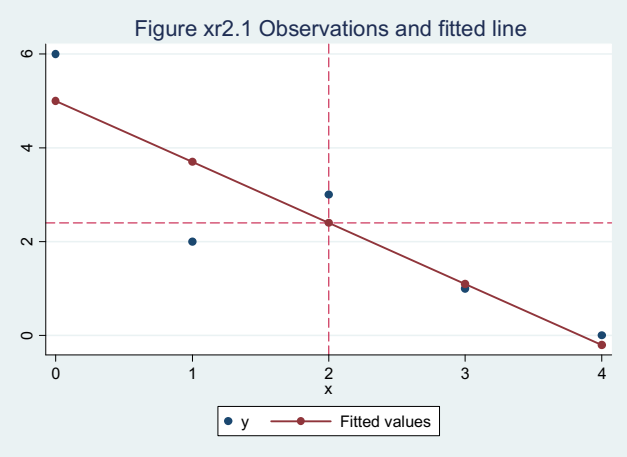
$$\sum_{i=1}^{5}x_iy_i-N\bar{x}\bar{y}=11-5\times 2\times 2.4=-13=\sum_{i=1}^{5}(x_i-\bar{x})(y_i-\bar{y})$$

(d)

x_i	y_i	$\hat{y}_i$	$\hat{e}_i$	$\hat{e}_i^2$	$x_i\hat{e}_i$
0	6	5	1	1	0
1	2	3.7	−1.7	2.89	−1.7
2	3	2.4	0.6	0.36	1.2
3	1	1.1	−0.1	0.01	−0.3
4	0	−0.2	0.2	0.04	0.8
$\sum x_i =$ 10	$\sum y_i =$ 12	$\sum\hat{y}_i =$ 12	$\sum\hat{e}_i =$ 0	$\sum\hat{e}_i^2 =$ 4.3	$\sum x_i\hat{e}_i =$ 0

Exercise 2.1 (continued)

(e)



(f) See figure above. The fitted line passes through the point of the means, $\bar{x} = 2, \quad \bar{y} = 2.4$.

(g) Given $b_1 = 5,\ b_2 = -1.3$ and $\bar{y} = b_1 + b_2\bar{x}$, we have $\bar{y} = 2.4 = b_1 + b_2\bar{x} = 5 - 1.3(2) = 2.4$

(h) $\bar{\hat{y}} = \sum \hat{y}_i / N = 12/5 = 2.4 = \bar{y}$

(i) $$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2} = \frac{4.3}{3} = 1.433\dot{3}$$

(j) $$\widehat{\text{var}}(b_2) = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2} = \frac{1.4333}{10} = 0.14333$$

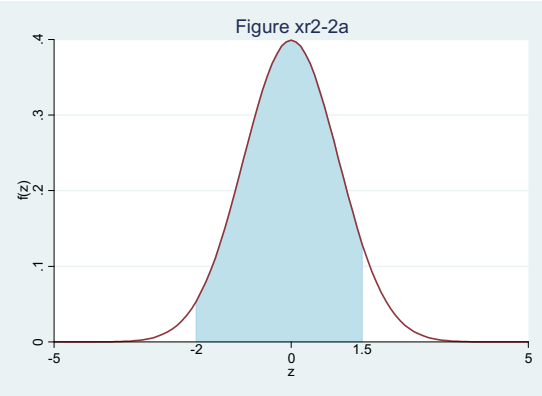
EXERCISE 2.2

(a) $$P(180 < X < 215) = P\left(\frac{180 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} < \frac{X - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} < \frac{215 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}}\right)$$

$$= P\left(\frac{180 - 200}{\sqrt{100}} < Z < \frac{215 - 200}{\sqrt{100}}\right)$$

$$= P(-2 < Z < 1.5)$$

$$= 0.9104$$

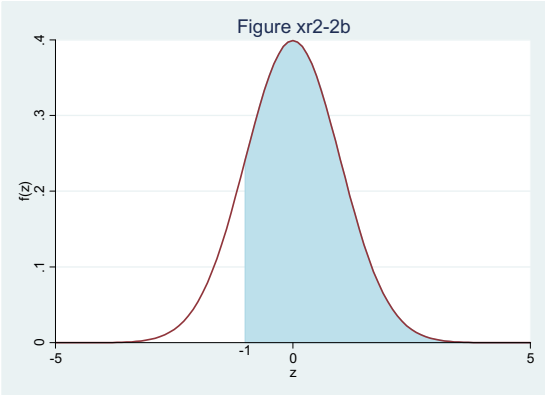


(b) $$P(X > 190) = P\left(\frac{X - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} > \frac{190 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}}\right)$$

$$= P\left(Z > \frac{190 - 200}{\sqrt{100}}\right)$$

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

$$= 0.8413$$



Exercise 2.2 (continued)

(c) $$P(180 < X < 215) = P\left(\frac{180 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} < \frac{X - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}} < \frac{215 - \mu_{y|x=\$2000}}{\sqrt{\sigma^2_{y|x=\$2000}}}\right)$$

$$= P\left(\frac{180 - 200}{\sqrt{81}} < Z < \frac{215 - 200}{\sqrt{81}}\right)$$

$$= P(-2.2222 < Z < 1.6666)$$

$$= 0.9391$$

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

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

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

$$= 0.8667$$

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. For this reason, it has been omitted.

(b) Preliminary calculations yield:

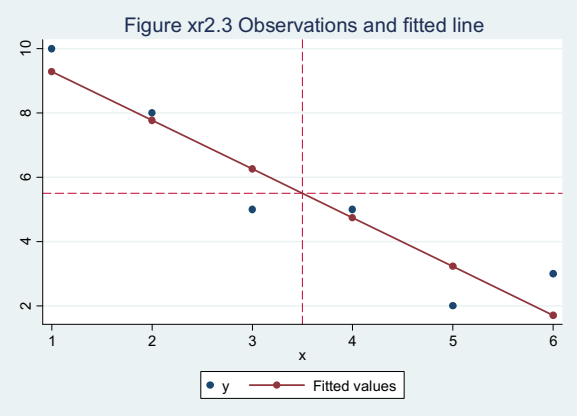
$$\sum x_i = 21 \qquad \sum y_i = 33 \qquad \sum (x_i - \bar{x})(y_i - \bar{y}) = -26.5 \qquad \sum (x_i - \bar{x})^2 = 17.5$$

$$\bar{y} = 5.5 \qquad \bar{x} = 3.5$$

The least squares estimates are:

$$b_2 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} = \frac{-26.5}{17.5} = -1.514286$$

$$b_1 = \bar{y} - b_2 \bar{x} = 5.5 - (-1.514286) \times 3.5 = 10.8$$



(c) $$\bar{y} = \sum y_i / N = 33/6 = 5.5$$

$$\bar{x} = \sum x_i / N = 21/6 = 3.5$$

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

$$\hat{y} = b_1 + b_2 \bar{x} = 10.8 - 1.514286 \times 3.5 = 5.5$$

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) .

Exercise 2.3 (Continued)

(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	10	0.714286
2	8	0.228571
3	5	−1.257143
4	5	0.257143
5	2	−1.228571
6	3	1.285714

Their sum is $\sum \hat{e}_i = 0.$

(e) $$\begin{aligned}\sum x_i \hat{e}_i &= 1\times 0.714286 + 2\times 0.228571 + 3\times(-1.257143) + 4\times 0.257143 \\ &\quad +5\times(-1.228571) + 6\times 1.285714 \\ &= 0\end{aligned}$$

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

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

$$= 352 - 2\times 176\beta + 91\beta^2 = 352 - 352\beta + 91\beta^2$$

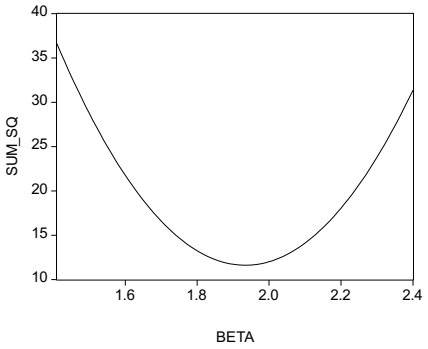


Figure xr2.4(a) Sum of squares for β_2

The minimum of this function is approximately 12 and occurs at approximately $\beta_2 = 1.95$. 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

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

Thus, the least-squares estimate is

$$b_2 = \frac{176}{91} = 1.9341$$

which agrees with the approximate value of 1.95 that we obtained geometrically.

Exercise 2.4 (Continued)

(e)

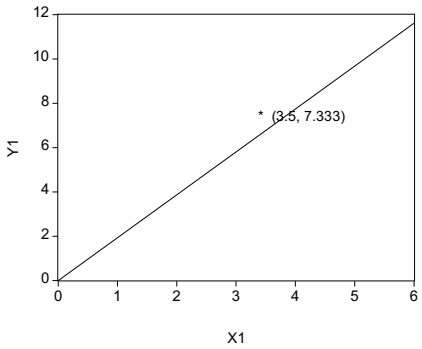


Figure xr2.4(b) Fitted regression line and mean

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:

$$\hat{e}_1 = 2.0659 \qquad \hat{e}_2 = 2.1319 \qquad \hat{e}_3 = 1.1978$$

$$\hat{e}_4 = -0.7363 \qquad \hat{e}_5 = -0.6703 \qquad \hat{e}_6 = -0.6044$$

Their sum is $\sum \hat{e}_i = 3.3846$. Note this value is not equal to zero as it was for $\beta_1 \neq 0$.

(g) $$\sum x_i \hat{e}_i = 2.0659 \times 1 + 2.1319 \times 2 + 1.1978 \times 3$$
$$- 0.7363 \times 4 - 0.6703 \times 5 - 0.6044 \times 6 = 0$$

EXERCISE 2.5

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

$$b_1 + 500b_2 = 10000$$

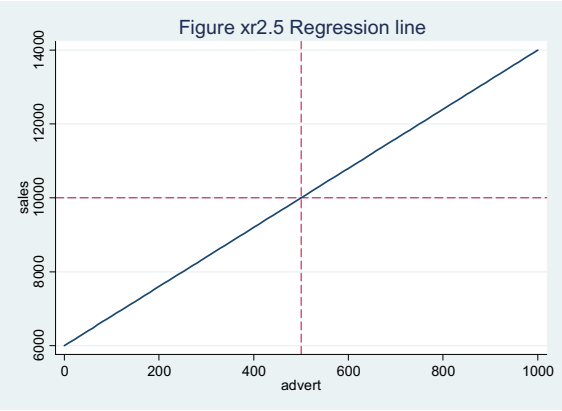
$$b_1 + 750b_2 = 12000$$

Solving these two equations yields

$$250b_2 = 2000 \;\Rightarrow\; b_2 = \frac{2000}{250} = 8 \qquad b_1 = 6000$$

Therefore, the estimated regression used by the consultant is:

$$\widehat{SALES} = 6000 + 8 \times ADVERT$$



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 $b_2 = 8$ 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

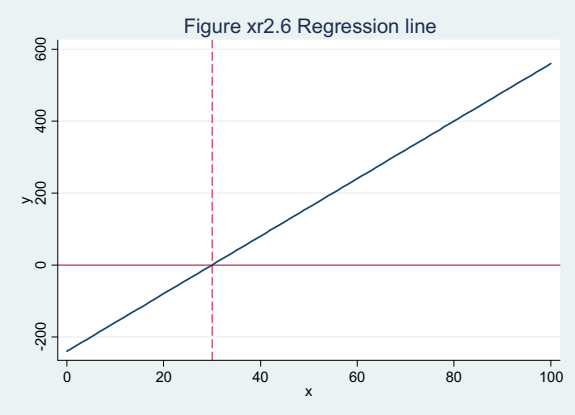
$$\hat{y} = -240 + 8 \times 80 = 400$$

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

$$0 = -240 + 8x \quad \text{or} \quad x = 30$$

Thus, she predicts no sodas will be sold below 30°F.

(d) A graph of the estimated regression line:



EXERCISE 2.7

(a) Since

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i^2}{N-2} = 2.04672$$

it follows that

$$\sum \hat{e}_i^2 = 2.04672\,(N-2) = 2.04672 \times 49 = 100.29$$

(b) The standard error for b_2 is

$$\text{se}(b_2) = \sqrt{\widehat{\text{var}(b_2)}} = \sqrt{0.00098} = 0.031305$$

Also,

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

Thus,

$$\sum \left(x_i - \bar{x}\right)^2 = \frac{\hat{\sigma}^2}{\widehat{\text{var}\left(b_2\right)}} = \frac{2.04672}{0.00098} = 2088.5$$

(c) The value $b_2 = 0.18$ suggests that a 1% increase in the percentage of males 18 years or older who are high school graduates will lead to an increase of \$180 in the mean income of males who are 18 years or older.

(d) $$b_1 = \bar{y} - b_2\bar{x} = 15.187 - 0.18 \times 69.139 = 2.742$$

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

$$\sum x_i^2 = \sum \left(x_i - \bar{x}\right)^2 + N\bar{x}^2 = 2088.5 + 51 \times 69.139^2 = 245{,}879$$

(f) For Arkansas

$$\hat{e}_i = y_i - \hat{y}_i = y_i - b_1 - b_2 x_i = 12.274 - 2.742 - 0.18 \times 58.3 = -0.962$$

EXERCISE 2.8

(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

$$\begin{aligned} E(b_{EZ}) &= E\left[\frac{y_2 - y_1}{x_2 - x_1}\right] = \frac{1}{x_2 - x_1} E(y_2) - \frac{1}{x_2 - x_1} E(y_1) \\ &= \frac{1}{x_2 - x_1}(\beta_1 + \beta_2 x_2) - \frac{1}{x_2 - x_1}(\beta_1 + \beta_2 x_1) \\ &= \frac{\beta_2 x_2}{x_2 - x_1} - \frac{\beta_2 x_1}{x_2 - x_1} = \beta_2 \left(\frac{x_2}{x_2 - x_1} - \frac{x_1}{x_2 - x_1}\right) = \beta_2 \end{aligned}$$

Thus, b_{EZ} is an unbiased estimator.

(c) The variance is given by

$$\begin{aligned} \text{var}(b_{EZ}) &= \text{var}(\sum k_i y_i) = \sum k_i^2 \, \text{var}(e_i) = \sigma^2 \sum k_i^2 \\ &= \sigma^2 \left(\frac{1}{(x_2 - x_1)^2} + \frac{1}{(x_2 - x_1)^2}\right) = \frac{2\sigma^2}{(x_2 - x_1)^2} \end{aligned}$$

(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 2.8 (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

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

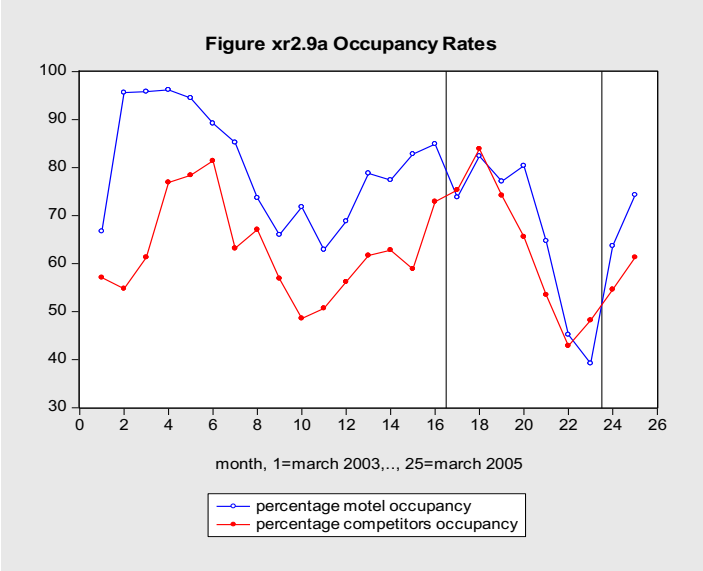
or that

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

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.

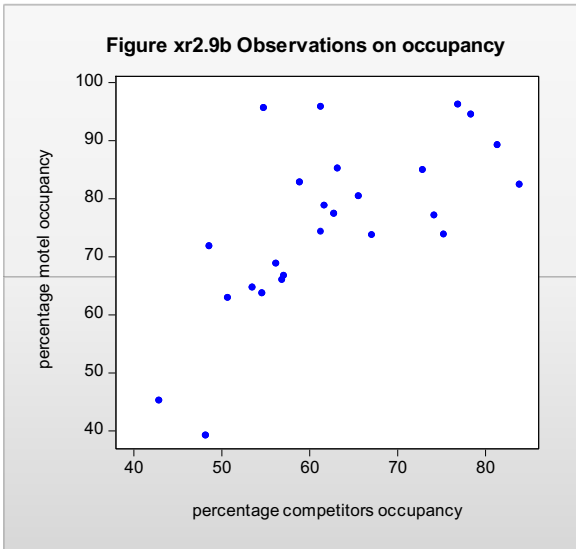
EXERCISE 2.9

(a) Plots of the occupancy rates for the motel and its competitors for the 25-month period are given in the following figure.



The repair period comprises those months between the two vertical lines. The graphical evidence suggests that the damaged motel had the higher occupancy rate before and after the repair period. During the repair period, the damaged motel and the competitors had similar occupancy rates.

(b) A plot of *MOTEL_PCT* against *COMP_PCT* yields:



There appears to be a positive relationship the two variables. Such a relationship may exist as both the damaged motel and the competitor(s) face the same demand for motel rooms. That is, competitor occupancy rates reflect overall demand in the market for motel rooms.

Exercise 2.9 (continued)

(c) The estimated regression is $\widehat{MOTEL\_PCT} = 21.40 + 0.8646 \times COMP\_PCT$.

The competitors' occupancy rates are positively related to motel occupancy rates, as expected. The regression indicates that for a one percentage point increase in competitor occupancy rate, the damaged motel's occupancy rate is expected to increase by 0.8646 percentage points.

(d)

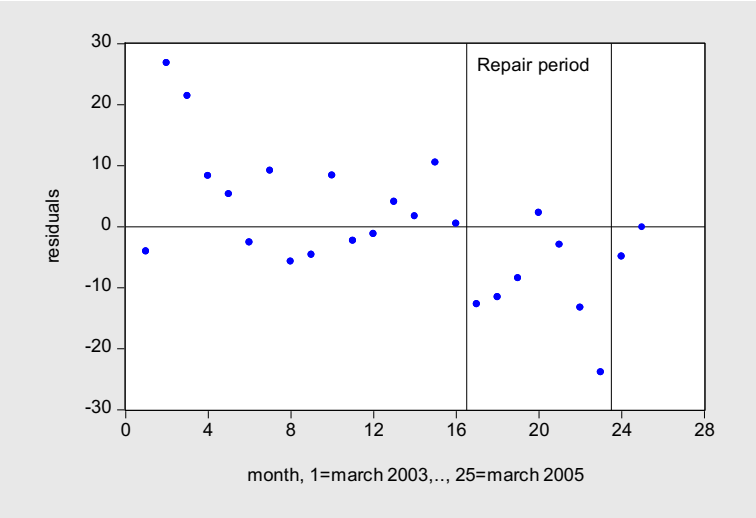


Figure xr2.9(d) Plot of residuals against time

The residuals during the occupancy period are those between the two vertical lines. All except one are negative, indicating that the model has over-predicted the motel's occupancy rate during the repair period.

(e) We would expect the slope coefficient of a linear regression of *MOTEL_PCT* on *RELPRICE* to be negative, as the higher the relative price of the damaged motel's rooms, the lower the demand will be for those rooms, holding other factors constant.

The estimated regression is:

$$\widehat{MOTEL\_PCT} = 166.66 - 122.12 \times RELPRICE$$

The sign of the estimated slope is negative, as expected.

(f) The linear regression with an indicator variable is:

$$MOTEL\_PCT = \beta_1 + \beta_2 REPAIR + e$$

From this equation, we have that:

$$E(MOTEL\_PCT) = \beta_1 + \beta_2 REPAIR = \begin{cases} \beta_1 + \beta_2 & \text{if } REPAIR = 1 \\ \beta_1 & \text{if } REPAIR = 0 \end{cases}$$

Exercise 2.9(f) (continued)

The expected occupancy rate for the damaged motel is $\beta_1 + \beta_2$ during the repair period; it is β_1 outside of the repair period. Thus β_2 is the difference between the expected occupancy rates for the damaged motel during the repair and non-repair periods.

The estimated regression is:

$$\widehat{MOTEL\_PCT} = 79.3500 - 13.2357 \times REPAIR$$

In the non-repair period, the damaged motel had an estimated occupancy rate of 79.35%. During the repair period, the estimated occupancy rate was 79.35−13.24 = 66.11%. Thus, it appears the motel did suffer a loss of occupancy and profits during the repair period.

(g) From the earlier regression, we have

$$\overline{MOTEL}_0 = b_1 = 79.35\%$$

$$\overline{MOTEL}_1 = b_1 + b_2 = 79.35 - 13.24 = 66.11\%$$

For competitors, the estimated regression is:

$$\widehat{COMP\_PCT} = 62.4889 + 0.8825 \times REPAIR$$

Thus,

$$\overline{COMP}_0 = b_1 = 62.49\%$$

$$\overline{COMP}_1 = b_1 + b_2 = 62.49 + 0.88 = 63.37\%$$

During the non-repair period, the difference between the average occupancies was:

$$\overline{MOTEL}_0 - \overline{COMP}_0 = 79.35 - 62.49 = 16.86\%$$

During the repair period it was

$$\overline{MOTEL}_1 - \overline{COMP}_1 = 66.11 - 63.37 = 2.74\%$$

This comparison supports the motel's claim for lost profits during the repair period. When there were no repairs, their occupancy rate was 16.86% higher than that of their competitors; during the repairs it was only 2.74% higher.

(h) The estimated regression is:

$$\widehat{MOTEL\_PCT - COMP\_PCT} = 16.8611 - 14.1183 \times REPAIR$$

The intercept estimate in this equation (16.86) is equal to the difference in average occupancies during the non-repair period, $\overline{MOTEL}_0 - \overline{COMP}_0$. The sum of the two coefficient estimates $(16.86 + (-14.12) = 2.74)$ is equal to the difference in average occupancies during the repair period, $\overline{MOTEL}_1 - \overline{COMP}_1$.

This relationship exists because averaging the difference between two series is the same as taking the difference between the averages of the two series.

EXERCISE 2.10

(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)

Firm	Microsoft	General Electric	General Motors	IBM	Disney	Exxon-Mobil
$b_2 = \hat{\beta}_j$	1.3189	0.8993	1.2614	1.1882	0.8978	0.4140

The stocks Microsoft, General Motors and IBM are aggressive with Microsoft being the most aggressive with a beta value of $b_2 = 1.3189$. General Electric, Disney and Exxon-Mobil are defensive with Exxon-Mobil being the most defensive with a beta value of $b_2 = 0.4140$.

(c)

Firm	Microsoft	General Electric	General Motors	IBM	Disney	Exxon-Mobil
$b_1 = \hat{\alpha}_j$	0.0061	−0.0012	−0.0116	0.0059	−0.0011	0.0079

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 xr2.10.

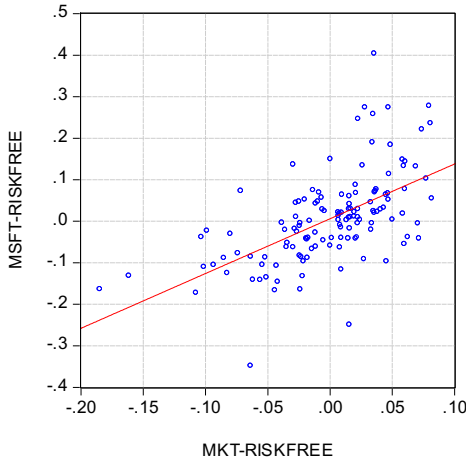


Fig. xr2.10 Scatter plot of Microsoft and market rate

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

Firm	Microsoft	General Electric	General Motors	IBM	Disney	Exxon-Mobil
$\hat{\beta}_j$	1.3185	0.8993	1.2622	1.1878	0.8979	0.4134

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 2.11

(a)

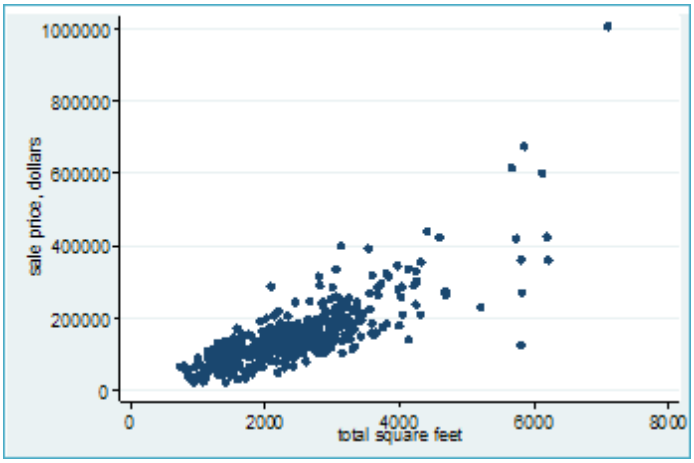
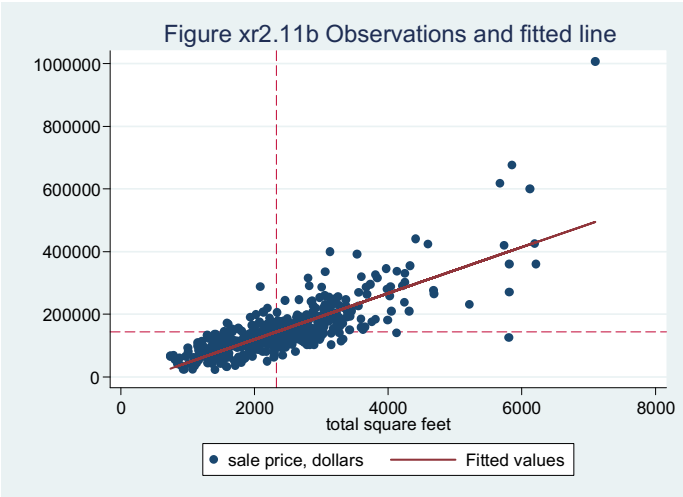


Figure xr2.11(a) Price against square feet for houses of traditional style

(b) The estimated equation for traditional style houses is:

$$\widehat{PRICE} = -28408 + 73.772\,SQFT$$

The slope of 73.772 suggests that expected house price increases by approximately \$73.77 for each additional square foot of house size. The intercept term is $-28{,}408$ which would be interpreted as the dollar price of a traditional house of zero square feet. Once again, this estimate should not be accepted as a serious one. A negative value is meaningless and there is no data in the region of zero square feet.



Exercise 2.11 (continued)

(c) The estimated equation for traditional style houses is:

$$\widehat{PRICE} = 68710 + 0.012063\,SQFT^2$$

The marginal effect on price of an additional square foot is:

$$\widehat{\text{slope}} = \frac{d\left(\widehat{PRICE}\right)}{dSQFT} = 2\left(0.012063\right)SQFT$$

For a home with 2000 square feet of living space, the marginal effect is:

$$\frac{d\left(\widehat{PRICE}\right)}{dSQFT} = 2\left(0.012063\right)\left(2000\right) = 48.25$$

That is, an additional square foot of living space for a traditional home of 2000 square feet is expected to increase its price by \$48.25.

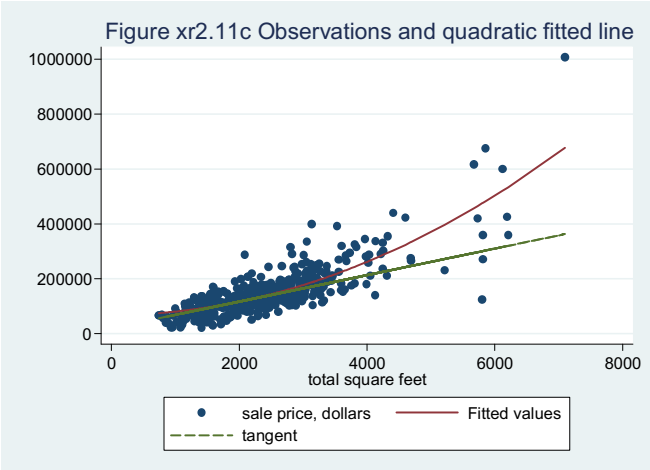
To obtain the elasticity, we first need to compute an estimate of the expected price when $SQFT = 2000$:

$$\widehat{PRICE} = 68710 + 0.0120632\left(2000\right)^2 = 116963$$

Then, the elasticity of price with respect to living space for a traditional home with 2000 square feet of living space is:

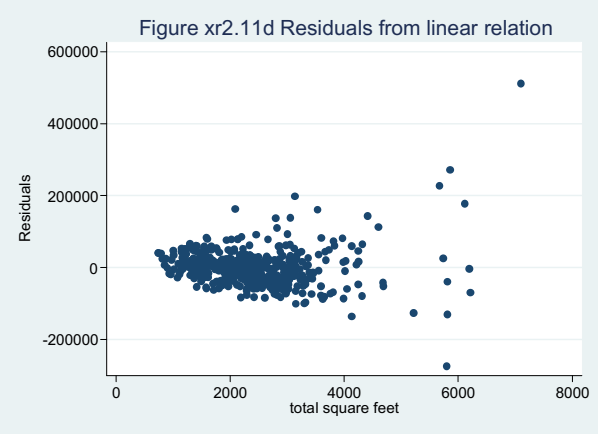
$$\hat{\varepsilon} = \widehat{\text{slope}} \times \frac{SQFT}{PRICE} = \frac{d\left(\widehat{PRICE}\right)}{dSQFT} \times \frac{SQFT}{PRICE} = 2\left(0.0120632\right)\left(2000\right)\left(\frac{2000}{116963}\right) = 0.825$$

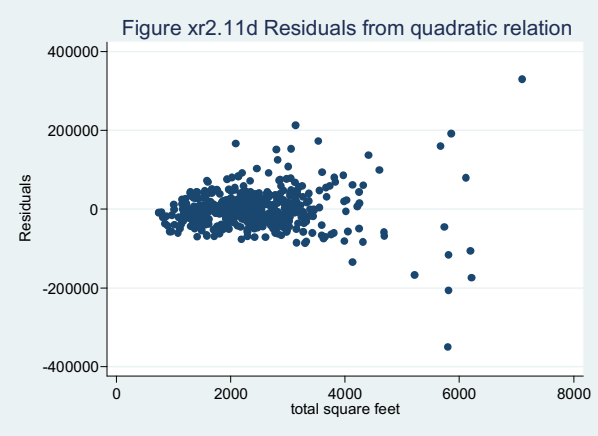
That is, for a 2000 square foot house, we estimate that a 1% increase in house size will increase price by 0.825%.



Exercise 2.11 (continued)

(d) Residual plots:





The magnitude of the residuals tends to increase as housing size increases suggesting that SR3 $\text{var}(e \mid x) = \sigma^2$ [homoskedasticity] could be violated. The larger residuals for larger houses imply the spread or variance of the errors is larger as *SQFT* increases. Or, in other words, there is not a constant variance of the error term for all house sizes.

(e) *SSE* of linear model, (b): $SSE = \sum \hat{e}_i^2 = 1.37 \times 10^{12}$

SSE of quadratic model, (c): $SSE = \sum \hat{e}_i^2 = 1.23 \times 10^{12}$

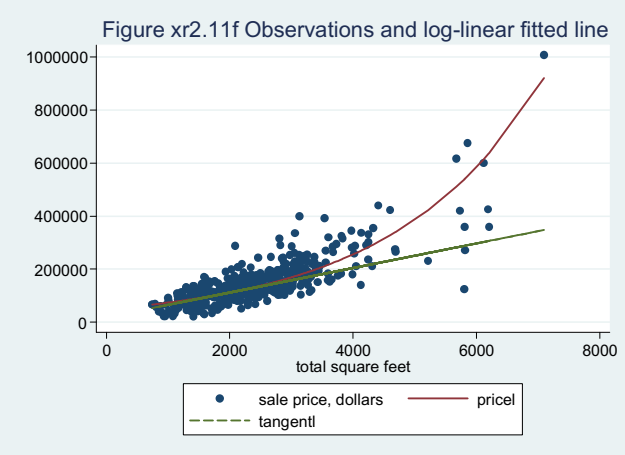
The quadratic model has a lower *SSE*. A lower *SSE*, or sum of squared residuals, indicates a lower value for the squared distance between a regression line and data points, indicating a line that better fits the data.

Exercise 2.11 (continued)

(f) The estimated equation for traditional style houses is:

$$\widehat{\ln(PRICE)} = 10.79894 + 0.000413235\,SQFT$$

The fitted line, with a tangent line included, is



(g) The *SSE* from the log-linear model is based on how well the model fits $\ln(PRICE)$. Since the log scale is compressed, the *SSE* from this specification is not comparable to the *SSE* from the models with *PRICE* as the dependent variable. One way to correct this problem is to obtain the predicted values from the log-linear model, then take the antilogarithm to make predictions in terms of *PRICE*. Then a residual can be computed as

$$\hat{e} = PRICE - \exp\left[\widehat{\ln(PRICE)}\right]$$

Using this approach the *SSE* from log-linear model is 1.31×10^{12}. This is smaller than the *SSE* from the fitted linear relationship, but not as small as the *SSE* from the fitted quadratic model.

EXERCISE 2.12

(a) The scatter plot in the figure below shows a positive relationship between selling price and house size.

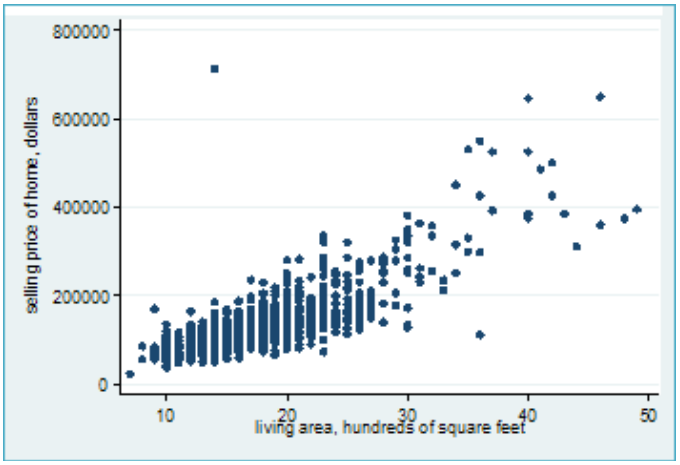
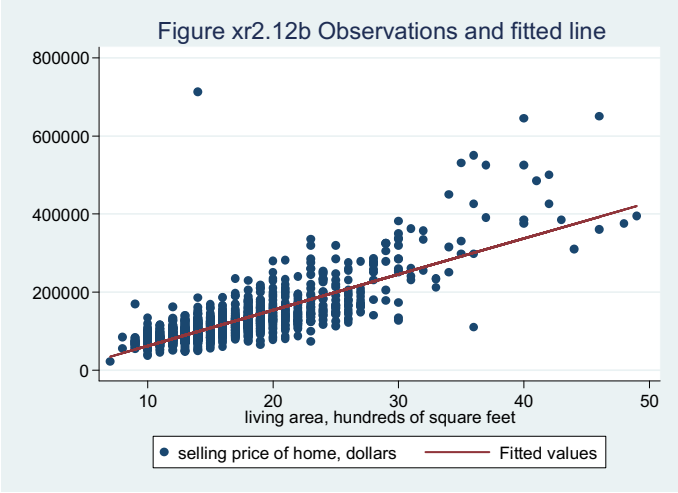


Figure xr2.12(a) Scatter plot of selling price and living area

(b) The estimated equation for all houses in the sample is

$$\widehat{SPRICE} = -30069 + 9181.7\,LIVAREA$$

The coefficient 9181.7 suggests that selling price increases by approximately \$9182 for each additional 100 square foot in living area. The intercept, if taken literally, suggests a house with zero square feet would cost $-\$30{,}069$, a meaningless value. The model should not be accepted as a serious one in the region of zero square feet.



Exercise 2.12 (continued)

(c) The estimated quadratic equation for all houses in the sample is

$$\widehat{SPRICE} = 57728 + 212.611 LIVAREA^2$$

The marginal effect of an additional 100 square feet is:

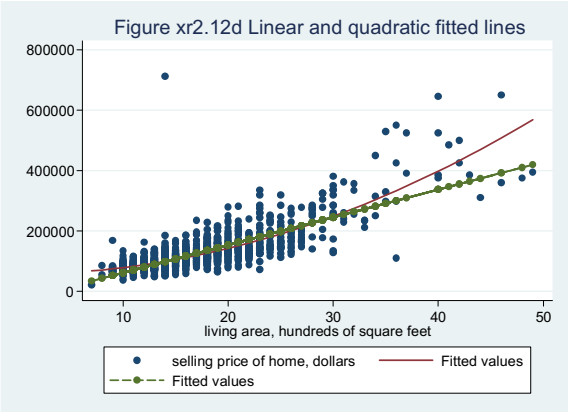
$$\widehat{\text{slope}} = \frac{d\left(\widehat{SPRICE}\right)}{dLIVAREA} = 2\left(212.611\right) LIVAREA$$

For a home with 1500 square feet of living space, the marginal effect is:

$$\frac{d\left(\widehat{SPRICE}\right)}{dLIVAREA} = 2\left(212.611\right)\left(15\right) = \ 6378.33$$

That is, adding 100 square feet of living space to a house of 1500 square feet is estimated to increase its expected price by approximately $6378.

(d)



The quadratic model appears to fit the data better; it is better at capturing the proportionally higher prices for large houses.

SSE of linear model, (b): $SSE = \sum \hat{e}_i^2 = 2.23 \times 10^{12}$

SSE of quadratic model, (c): $SSE = \sum \hat{e}_i^2 = 2.03 \times 10^{12}$

The *SSE* of the quadratic model is smaller, indicating that it is a better fit.

(e) The estimated equation for houses that are on large lots in the sample is:

$$\widehat{SPRICE} = 113279 + 193.83 LIVAREA^2$$

The estimated equation for houses that are on small lots in the sample is:

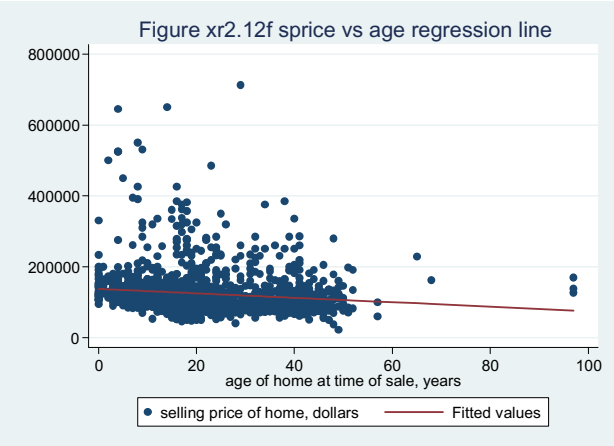
$$\widehat{SPRICE} = 62172 + 186.86 LIVAREA^2$$

Exercise 2.12(e) (continued)

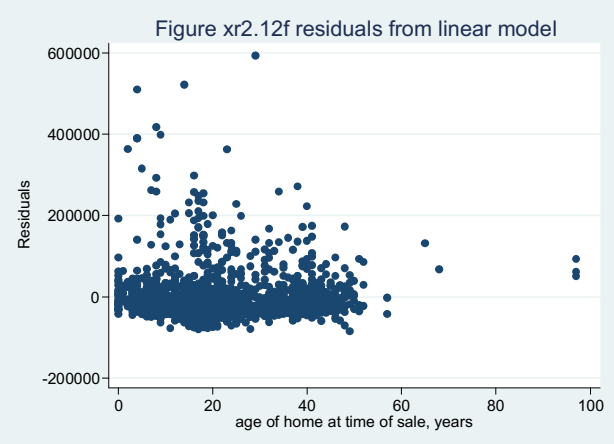
The intercept can be interpreted as the expected price of the land – the selling price for a house with no living area. The coefficient of *LIVAREA* has to be interpreted in the context of the marginal effect of an extra 100 square feet of living area, which is $2\beta_2 LIVAREA$. Thus, we estimate that the mean price of large lots is \$113,279 and the mean price of small lots is \$62,172. The marginal effect of living area on price is $\$387.66 \times LIVAREA$ for houses on large lots and $\$373.72 \times LIVAREA$ for houses on small lots.

(f) The following figure contains the scatter diagram of *PRICE* and *AGE* as well as the estimated equation which is

$$\widehat{SPRICE} = 137404 - 627.16 AGE$$



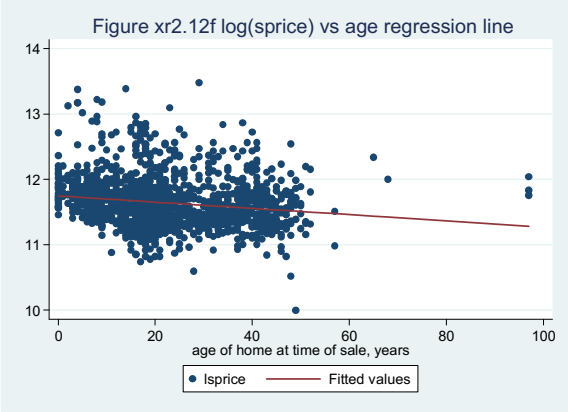
We estimate that the expected selling price is \$627 less for each additional year of age. The estimated intercept, if taken literally, suggests a house with zero age (i.e., a new house) would cost \$137,404. The model residuals plotted below show an asymmetric pattern, with some very large positive values. For these observations the linear fitted model under predicts the selling price.



Exercise 2.12(f) (continued)

The following figure contains the scatter diagram of ln(*PRICE*) and *AGE* as well as the estimated equation which is

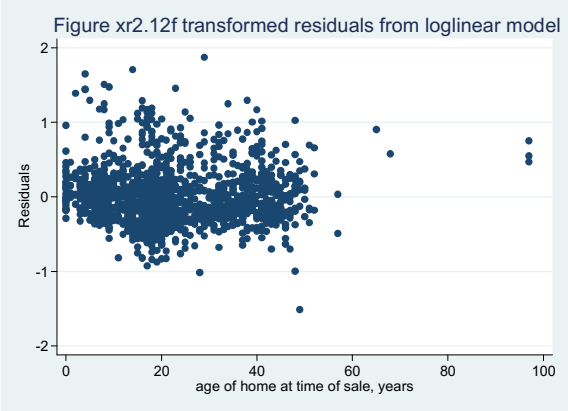
$$\widehat{\ln(SPRICE)} = 11.746 - 0.00476\,AGE$$



In this estimated model, each extra year of age reduces the selling price by 0.48%. To find an interpretation from the intercept, we set $AGE = 0$, and find an estimate of the price of a new home as

$$\exp\left[\widehat{\ln(SPRICE)}\right] = \exp(11.74597) = \$126,244$$

The following residuals from the fitted regression of ln(*SPRICE*) on *AGE* show much less of problem with under-prediction; the residuals are distributed more symmetrically around zero. Thus, based on the plots and visual fit of the estimated regression lines, the log-linear model is preferred.



(g) The estimated equation for all houses is:

$$\widehat{SPRICE} = 115220 + 133797\,LGELOT$$

The estimated expected selling price for a house on a large lot (*LGELOT* = 1) is 115220+133797 = \$249017. The estimated expected selling price for a house not on a large lot (*LGELOT* = 0) is \$115220.

EXERCISE 2.13

(a) The estimated equation using a sample of small and regular classes is:

$$\widehat{TOTALSCORE} = 918.043 + 13.899SMALL$$

Comparing a sample of small and regular classes, we find students in regular classes achieve an average total score of 918.0 while students in small classes achieve an average of 918.0 + 13.9 = 931.9. This is a 1.50% 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:

$$\widehat{READSCORE} = 434.733 + 5.819SMALL$$

$$\widehat{MATHSCORE} = 483.310 + 8.080SMALL$$

Students in regular classes achieve an average reading score of 434.7 while students in small classes achieve an average of 434.73 + 5.82 = 440.6. This is a 1.34% increase. In math students in regular classes achieve an average score of 483.31 while students in small classes achieve an average of 483.31 + 8.08 = 491.4. This is a 1.67% increase. These results suggests that small class sizes also have a positive impact on learning math and reading.

(c) The estimated equation using a sample of regular classes and regular classes with a full-time teacher aide is:

$$\widehat{TOTALSCORE} = 918.043 + 0.314AIDE$$

Students in regular classes without a teacher aide achieve an average total score of 918.0 while students in regular classes with a teacher aide achieve an average total score of 918.04 + 0.31 = 918.4. These results suggest that having a full-time teacher aide has little 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:

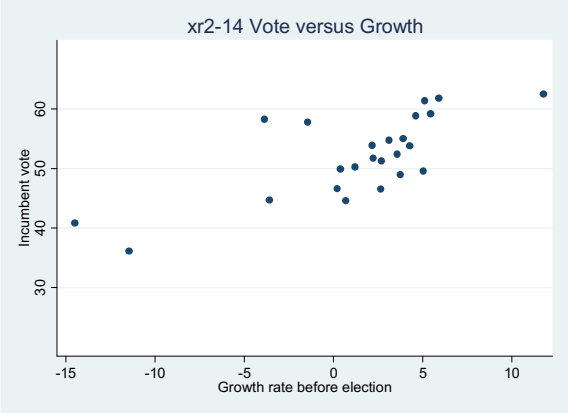
$$\widehat{READSCORE} = 434.733 + 0.705AIDE$$

$$\widehat{MATHSCORE} = 483.310 - 0.391AIDE$$

The effect of having a teacher aide on learning, as measured by reading and math scores, is negligible. This result does not differ from the case using total scores.

EXERCISE 2.14

(a)



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

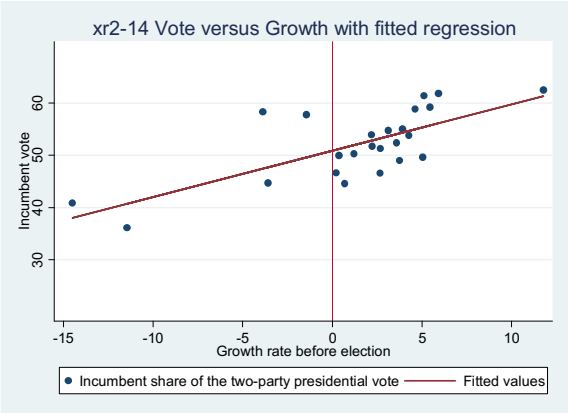
(b) The estimated equation for 1916 to 2008 is

$$\widehat{VOTE} = 50.848 + 0.88595GROWTH$$

The coefficient 0.88595 suggests that for a 1 percentage point increase in the growth rate of *GDP* in the 3 quarters before the election there is an estimated increase in the share of votes of the incumbent party of 0.88595 percentage points.

We estimate, based on the fitted regression intercept, that that the incumbent party's expected vote is 50.848% when the growth rate in *GDP* is zero. This suggests that when there is no real *GDP* growth, the incumbent party will still maintain the majority vote.

A graph of the fitted line and data is shown in the following figure.



(c) The estimated equation for 1916 - 2004 is

$$\widehat{VOTE} = 51.053 + 0.877982GROWTH$$

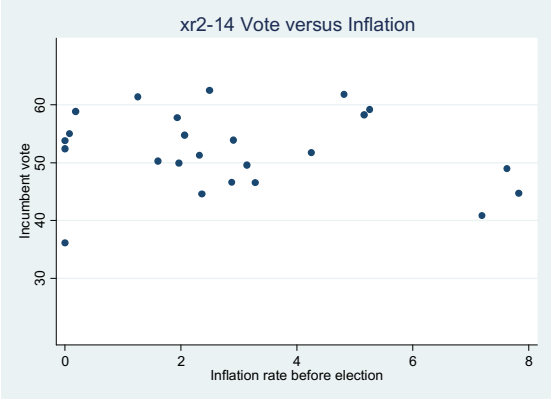
The actual 2008 value for growth is 0.220. Putting this into the estimated equation, we obtain the predicted vote share for the incumbent party:

Exercise 2.14(c) (continued)

$$\widehat{VOTE}_{2008} = 51.053 + 0.877982 GROWTH_{2008} = 51.053 + 0.877982(0.220) = 51.246$$

This suggests that the incumbent party will maintain the majority vote in 2008. However, the actual vote share for the incumbent party for 2008 was 46.60, which is a long way short of the prediction; the incumbent party did not maintain the majority vote.

(d) The figure below shows a plot of *VOTE* against *INFLATION*. There appears to be a negative association between the two variables.

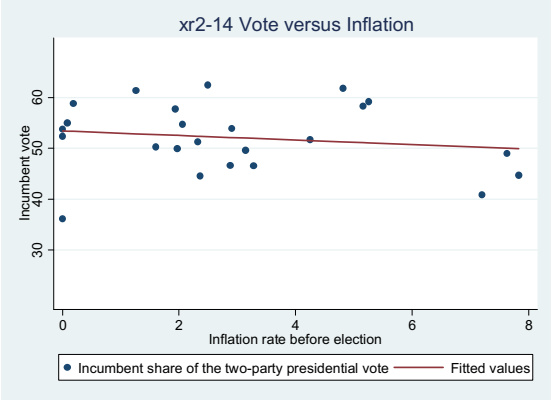


The estimated equation (plotted in the figure below) is:

$$\widehat{VOTE} = 53.408 - 0.444312 INFLATION$$

We estimate that a 1 percentage point increase in inflation during the incumbent party's first 15 quarters reduces the share of incumbent party's vote by 0.444 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 incumbent party is 53.4%; the incumbent party is predicted to maintain the majority vote when inflation, during its first 15 quarters, is at 0%.



EXERCISE 2.15

(a)

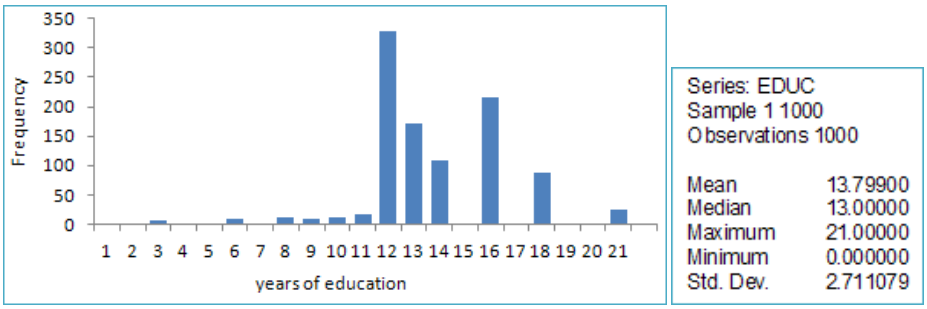


Figure xr2.15(a) Histogram and statistics for *EDUC*

Most 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 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.

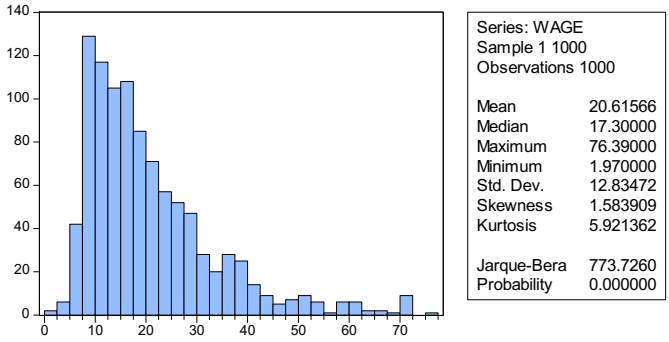


Figure xr2.15(a) Histogram and statistics for *WAGE*

The observations for *WAGE* are skewed to the right indicating that most of the observations lie between the hourly wages of 5 to 40, and that there is a smaller proportion of observations with an hourly wage greater than 40. Half of the sample earns an hourly wage of more than 17.30 dollars per hour, with the average being 20.62 dollars per hour. The maximum earned in this sample is 76.39 dollars per hour and the least earned in this sample is 1.97 dollars per hour.

(b) The estimated equation is

$$\widehat{WAGE} = -6.7103 + 1.9803EDUC$$

The coefficient 1.9803 represents the estimated increase in the expected hourly wage rate for an extra year of education. The coefficient −6.7103 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.

Exercise 2.15 (continued)

(c) The residuals are plotted against education in Figure xr2.15(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.

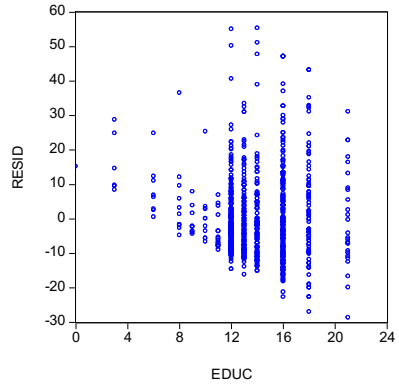


Figure xr2.15(c) Residuals against education

(d) The estimated equations are

If female: $\widehat{WAGE} = -14.1681 + 2.3575EDUC$

If male: $\widehat{WAGE} = -3.0544 + 1.8753EDUC$

If black: $\widehat{WAGE} = -15.0859 + 2.4491EDUC$

If white: $\widehat{WAGE} = -6.5507 + 1.9919EDUC$

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

(e) The estimated quadratic equation is

$$\widehat{WAGE} = 6.08283 + 0.073489EDUC^2$$

The marginal effect is therefore:

$$\widehat{\text{slope}} = \frac{d\left(\widehat{WAGE}\right)}{dEDUC} = 2\left(0.073489\right)EDUC$$

For a person with 12 years of education, the estimated marginal effect of an additional year of education on expected wage is:

$$\widehat{\text{slope}} = \frac{d\left(\widehat{WAGE}\right)}{dEDUC} = 2\left(0.073489\right)\left(12\right) = 1.7637$$

That is, an additional year of education for a person with 12 years of education is expected to increase wage by $1.76.

Exercise 2.15(e) (continued)

For a person with 14 years of education, the marginal effect of an additional year of education is:

$$\widehat{\text{slope}} = \frac{d\left(\widehat{WAGE}\right)}{dEDUC} = 2\left(0.073489\right)\left(14\right) = 2.0577$$

An additional year of education for a person with 14 years of education is expected to increase wage by $2.06.

The linear model in (b) suggested that an additional year of education is expected to increase wage by $1.98 regardless of the number of years of education attained. That is, the rate of change is 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)

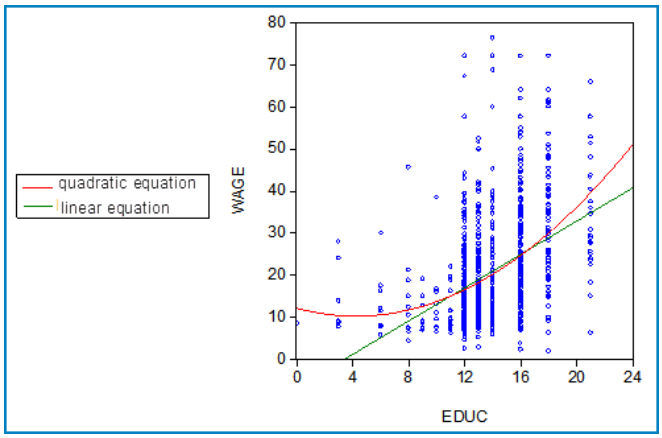


Figure xr2.15(f) Quadratic and linear equations for wage on education

The quadratic model appears to fit the data slightly better than the linear equation.

(g) The histogram of ln(*WAGE*) in the figure below is more symmetrical and bell-shaped than the histogram of *WAGE* given in part (a).

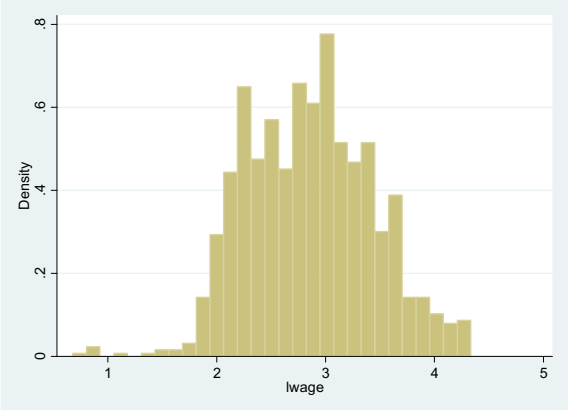


Figure xr2.15(g) Histogram for ln(*WAGE*)

Exercise 2.15 (continued)

(h) The estimated log-linear model is

$$\widehat{\ln(WAGE)} = 1.60944 + 0.090408 EDUC$$

We estimate that each additional year of education increases expected wage by approximately 9.04%.

The estimated marginal effect of education on *WAGE* is

$$\frac{dWAGE}{dEDUC} = \beta_2 \times WAGE$$

This marginal effect depends on the wage rate. For workers with 12 and 14 years of education we predict the wage rates to be

$$\widehat{WAGE}\Big|_{EDUC=12} = \exp(1.60944 + 0.090408 \times 12) = 14.796$$

$$\widehat{WAGE}\Big|_{EDUC=14} = \exp(1.60944 + 0.090408 \times 14) = 17.728$$

Evaluating the marginal effects at these values we have

$$\frac{\widehat{dWAGE}}{dEDUC} = b_2 \times WAGE = \begin{cases} 1.3377 & EDUC = 12 \\ 1.6028 & EDUC = 14 \end{cases}$$

For the linear relationship the marginal effect of education was estimated to be \$1.98. For the quadratic relationship the corresponding marginal effect estimates are \$1.76 and \$2.06.

The marginal effects from the log-linear model are lower.

A comparison of the fitted lines for the linear and log-linear model appears in the figure below.

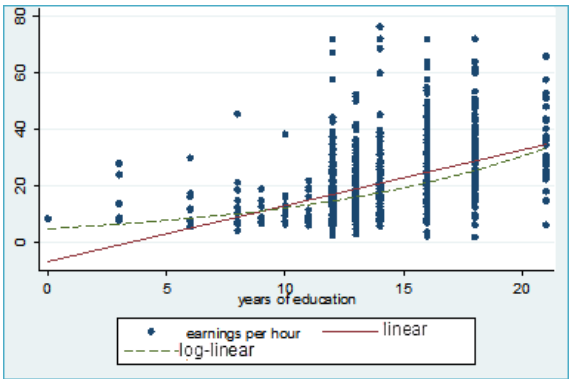


Figure xr12.15(h) Observations with linear and log-linear fitted lines