CHAPTER 4

Exercise Solutions

EXERCISE 4.1

(a) $$R^2 = 1 - \frac{\sum \hat{e}_i^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{182.85}{631.63} = 0.71051$$

(b) To calculate R^2 we need $\sum (y_i - \bar{y})^2$,

$$\sum (y_i - \bar{y})^2 = \sum y_i^2 - N\bar{y}^2 = 5930.94 - 20 \times 16.035^2 = 788.5155$$

Therefore,

$$R^2 = \frac{SSR}{SST} = \frac{666.72}{788.5155} = 0.8455$$

(c) From

$$R^2 = 1 - \frac{\sum \hat{e}_i^2}{SST} = 1 - \frac{(N-K)\hat{\sigma}^2}{SST}$$

we have,

$$\hat{\sigma}^2 = \frac{SST(1-R^2)}{N-K} = \frac{552.36 \times (1 - 0.7911)}{(20-2)} = 6.4104$$

EXERCISE 4.2

(a) $\hat{y} = 5.83 + 8.69x^*$ where $x^* = \dfrac{x}{10}$
(1.23) (1.17)

(b) $\hat{y}^* = 0.583 + 0.0869x$ where $\hat{y}^* = \dfrac{\hat{y}}{10}$
(0.123) (0.0117)

(c) $\hat{y}^* = 0.583 + 0.869x^*$ where $\hat{y}^* = \dfrac{\hat{y}}{10}$ and $x^* = \dfrac{x}{10}$
(0.123) (0.117)

The values of R^2 remain the same in all cases.

EXERCISE 4.3

(a) $\hat{y}_0 = b_1 + b_2 x_0 = 1 + 1 \times 5 = 6$

(b) $\widehat{\text{var}(f)} = \hat{\sigma}^2 \left(1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) = 5.3333 \left(1 + \frac{1}{5} + \frac{(5-1)^2}{10} \right) = 14.9332$

$\text{se}(f) = \sqrt{14.9332} = 3.864$

(c) Using $\text{se}(f)$ from part (b) and $t_c = t_{(0.975,3)} = 3.182$,

$$\hat{y}_0 \pm t_c \, \text{se}(f) = 6 \pm 3.182 \times 3.864 = (-6.295, 18.295)$$

(d) Using $\text{se}(f)$ from part (b) and $t_c = t_{(0.995,3)} = 5.841$,

$$\hat{y}_0 \pm t_c \, \text{se}(f) = 6 \pm 5.841 \times 3.864 = (-16.570, 28.570)$$

(e) Using $\bar{x} = x_0 = 1$, the prediction is $\hat{y}_0 = 1 + 1 \times 1 = 2$, and

$$\widehat{\text{var}(f)} = \hat{\sigma}^2 \left(1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) = 5.3333 \left(1 + \frac{1}{5} + \frac{(1-1)^2}{10} \right) = 6.340$$

$$\text{se}(f) = \sqrt{6.340} = 2.530$$

$$\hat{y}_0 \pm t_c \, \text{se}(f) = 2 \pm 3.182 \times 2.530 = (-6.050, 10.050)$$

Width in part (c) $= 18.295 - (-6.295) = 24.59$

Width in part (e) $= 10.050 - (-6.050) = 16.1$

The width in part (e) is smaller than the width in part (c), as expected. Predictions are more precise when made for *x* values close to the mean.

EXERCISE 4.4

(a) When estimating $E(y_0)$, we are estimating the average value of *y* for all observational units with an *x*-value of x_0. When predicting y_0, we are predicting the value of *y* for one observational unit with an *x*-value of x_0. The first exercise does not involve the random error e_0; the second does.

(b) $E(b_1 + b_2 x_0) = E(b_1) + E(b_2)x_0 = \beta_1 + \beta_2 x_0$

$$\begin{aligned}
\text{var}(b_1 + b_2 x_0) &= \text{var}(b_1) + x_0^2 \,\text{var}(b_2) + 2x_0 \,\text{cov}(b_1, b_2) \\
&= \frac{\sigma^2 \sum x_i^2}{N\sum (x_i - \bar{x})^2} + \frac{\sigma^2 x_0^2}{\sum (x_i - \bar{x})^2} - \frac{2\sigma^2 x_0 \bar{x}}{\sum (x_i - \bar{x})^2} \\
&= \frac{\sigma^2 \left(\sum (x_i - \bar{x})^2 + N\bar{x}^2\right)}{N\sum (x_i - \bar{x})^2} + \frac{\sigma^2 (x_0^2 - 2x_0\bar{x})}{\sum (x_i - \bar{x})^2} \\
&= \sigma^2 \left(\frac{1}{N} + \frac{x_0^2 - 2x_0\bar{x} + \bar{x}^2}{\sum (x_i - \bar{x})^2} \right) = \sigma^2 \left(\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right)
\end{aligned}$$

(c) It is not appropriate to say that $E(\hat{y}_0) = y_0$ because y_0 is a random variable.

$$\left[E(\hat{y}_0) = \beta_1 + \beta_2 x_0\right] \neq \left[\beta_1 + \beta_2 x_0 + e_0 = y_0\right]$$

We need to include y_0 in the expectation so that

$$E(\hat{y}_0 - y_0) = E(\hat{y}_0) - E(y_0) = \beta_1 + \beta_2 x_0 - \left(\beta_1 + \beta_2 x_0 + E(e_0)\right) = 0.$$

EXERCISE 4.5

(a) If we multiply the x values in the simple linear regression model $y = \beta_1 + \beta_2 x + e$ by 10, the new model becomes

$$y = \beta_1 + \left(\frac{\beta_2}{10}\right)(x \times 10) + e$$

$$= \beta_1 + \beta_2^* x^* + e \qquad \text{where} \quad \beta_2^* = \beta_2/10 \quad \text{and} \quad x^* = x \times 10$$

The estimated equation becomes

$$\hat{y} = b_1 + \left(\frac{b_2}{10}\right)(x \times 10)$$

Thus, β_1 and b_1 do not change and β_2 and b_2 becomes 10 times smaller than their original values. Since e does not change, the variance of the error term $\text{var}(e) = \sigma^2$ is unaffected.

(b) Multiplying all the y values by 10 in the simple linear regression model $y = \beta_1 + \beta_2 x + e$ gives the new model

$$y \times 10 = (\beta_1 \times 10) + (\beta_2 \times 10)x + (e \times 10)$$

or

$$y^* = \beta_1^* + \beta_2^* x + e^*$$

where

$$y^* = y \times 10, \qquad \beta_1^* = \beta_1 \times 10, \qquad \beta_2^* = \beta_2 \times 10, \qquad e^* = e \times 10$$

The estimated equation becomes

$$\hat{y}^* = \hat{y} \times 10 = (b_1 \times 10) + (b_2 \times 10)x$$

Thus, both β_1 and β_2 are affected. They are 10 times larger than their original values. Similarly, b_1 and b_2 are 10 times larger than their original values. The variance of the new error term is

$$\text{var}(e^*) = \text{var}(e \times 10) = 100 \times \text{var}(e) = 100\sigma^2$$

Thus, the variance of the error term is 100 times larger than its original value.

EXERCISE 4.6

(a) The least squares estimator for β_1 is $b_1 = \bar{y} - b_2\bar{x}$. Thus, $\bar{y} = b_1 + b_2\bar{x}$, and hence $(\bar{y}, \bar{x})$ lies on the fitted line.

(b) Consider the fitted line $\hat{y}_i = b_1 + x_i b_2$. Averaging over *N*, we obtain

$$\bar{\hat{y}} = \frac{\sum \hat{y}_i}{N} = \frac{1}{N}\sum\left(b_1 + x_i b_2\right) = \frac{1}{N}\left(b_1 N + b_2 \sum x_i\right) = b_1 + b_2 \frac{\sum x_i}{N} = b_1 + b_2\bar{x}$$

From part (a), we also have $\bar{y} = b_1 + b_2\bar{x}$. Thus, $\bar{y} = \bar{\hat{y}}$.

EXERCISE 4.7

(a) $\hat{y}_0 = b_2 x_0$

(b) Using the solution from Exercise 2.4 part (f)

$$SSE = \sum \hat{e}_i^2 = (2.0659^2 + 2.1319^2 + 1.1978^2 + (-0.7363)^2 + (-0.6703)^2 + (-0.6044)^2 = 11.6044$$

$$\sum y_i^2 = 4^2 + 6^2 + 7^2 + 7^2 + 9^2 + 11^2 = 352$$

$$R_u^2 = 1 - \frac{11.6044}{352} = 0.967$$

(c) $$r_{y\hat{y}}^2 = \frac{\hat{\sigma}_{y\hat{y}}^2}{\hat{\sigma}_y^2 \hat{\sigma}_{\hat{y}}^2} = \frac{\left[\sum (\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})\right]^2}{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{\hat{y}})^2} = \frac{(42.549)^2}{65.461 \times 29.333} = 0.943$$

The two alternative goodness of fit measures R_u^2 and $r_{y\hat{y}}^2$ are not equal.

(d) $SST = 29.333, \quad SSR = 67.370$

$\{SSR + SSE = 67.370 + 11.6044 = 78.974\} \neq \{SST = 29.333\}$

The decomposition does not hold.

EXERCISE 4.8

(a) Simple linear regression results:

$$\hat{y}_t = 0.6776 + 0.0161t \qquad R^2 = 0.4595$$
$$\text{(se)}\ (0.0725)^{***}\ (0.0026)^{***}$$

Linear-log regression results:

$$\hat{y}_t = 0.5287 + 0.1855\ln(t) \qquad R^2 = 0.2441$$
$$\text{(se)}\ (0.1472)^{***}\ (0.0481)^{***}$$

Quadratic regression results:

$$\hat{y}_t = 0.7914 + 0.000355t^2 \qquad R^2 = 0.5685$$
$$\text{(se)}\ (0.0482)^{***}\ (0.000046)^{***}$$

(b) (i) (ii)

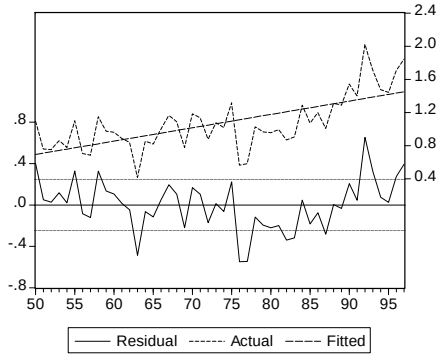

Figure xr4.8(a) Fitted line and residuals for the simple linear regression

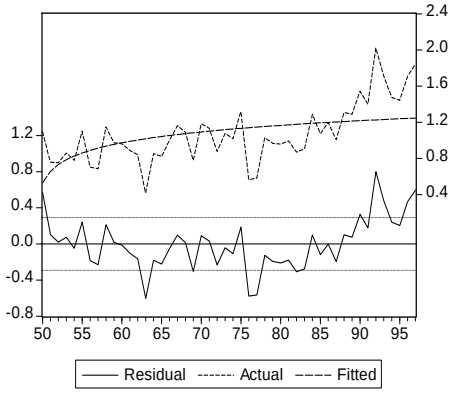

Figure xr4.8(b) Fitted line and residuals for the linear-log regression

Exercise 4.8(b) continued

(b)

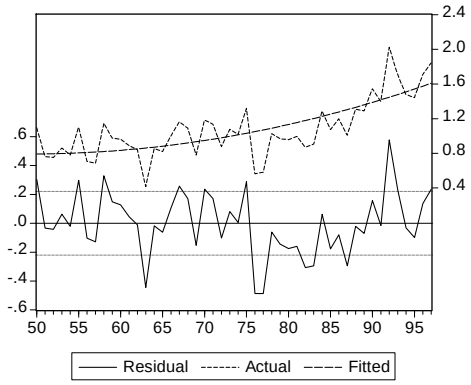

Figure xr4.8(c) Fitted line and residuals for the quadratic regression

(iii) Error normality tests Jarque-Bera:

Simple linear:	$JB = 0.279$	p-value = 0.870
Linear log:	$JB = 1.925$	p-value = 0.382
Quadratic:	$JB = 0.188$	p-value = 0.910

(iv) Values of R^2 are given in part (a)

To choose the preferred equation we consider the following.

1. The signs of the response parameters β_2, α_2 and γ_2: We expect them to be positive because we expect yield to increase over time as technology improves. The signs of the estimates of β_2, α_2 and γ_2 are as expected.
2. R^2: The value of R^2 for the third equation is the highest, namely 0.5685.
3. The plots of the fitted equations and their residuals: The upper parts of the figures display the fitted equation while the lower parts display the residuals. Considering the plots for the fitted equations, the one obtained from the third equation seems to fit the observations best. In terms of the residuals, the first two equations have concentrations of positive residuals at each end of the sample. The third equation provides a more balanced distribution of positive and negative residuals throughout the sample.

The third equation is preferable.

EXERCISE 4.9

(a) Equation 1: $\hat{y}_0 = 0.6776 + 0.0161 \times 49 = 1.467$

Equation 2: $\hat{y}_0 = 0.5287 + 0.1855 \ln(49) = 1.251$

Equation 3: $\hat{y}_0 = 0.7914 + 0.0003547 \times (49)^2 = 1.643$

(b) Equation 1: $\widehat{\frac{dy_t}{dt}} = \hat{\beta}_1 = 0.0161$

Equation 2: $\widehat{\frac{dy_t}{dt}} = \frac{\hat{\alpha}_1}{t} = \frac{0.1855}{49} = 0.0038$

Equation 3: $\widehat{\frac{dy_t}{dt}} = 2\hat{\gamma}_1 t = 2 \times 0.0003547 \times 49 = 0.0348$

(c) Evaluating the elasticities at $t = 49$ and the relevant value for $\hat{y}_0$ gives the following results.

Equation 1: $\widehat{\frac{dy_t}{dt}\frac{t}{y_t}} = \hat{\beta}_1 \frac{t}{\hat{y}_0} = 0.0161 \times \frac{49}{1.467} = 0.538$

Equation 2: $\widehat{\frac{dy_t}{dt}\frac{t}{y_t}} = \frac{\hat{\alpha}_1}{\hat{y}_t} = \frac{0.1855}{1.251} = 0.148$

Equation 3: $\widehat{\frac{dy_t}{dt}\frac{t}{y_t}} = \frac{2\hat{\gamma}_1 t^2}{\hat{y}_0} = \frac{2 \times 0.0003547 \times 49^2}{1.643} = 1.037$

(d) The slopes $\frac{dy_t}{dt}$ and the elasticities $\frac{dy_t}{dt}\frac{t}{y_t}$ give the marginal change in yield and the percentage change in yield, respectively, that can be expected from technological change in the next year. The results show that the predicted effect of technological change is very sensitive to the choice of functional form.

EXERCISE 4.10

(a) For households with 1 child

$$\widehat{WFOOD} = 1.0099 - 0.1495\ln(TOTEXP)$$
$$\text{(se)} \quad (0.0401) \quad (0.0090) \qquad R^2 = 0.3203$$
$$(t) \quad (25.19) \quad (-16.70)$$

For households with 2 children:

$$\widehat{WFOOD} = 0.9535 - 0.1294\ln(TOTEXP)$$
$$\text{(se)} \quad (0.0365) \quad (0.0080) \qquad R^2 = 0.2206$$
$$(t) \quad (26.10) \quad (-16.16)$$

For β_2 we would expect a negative value because as the total expenditure increases the food share should decrease with higher proportions of expenditure devoted to less essential items. Both estimations give the expected sign. The standard errors for b_1 and b_2 from both estimations are relatively small resulting in high values of t ratios and significant estimates.

(b) For households with 1 child, the average total expenditure is 94.848 and

$$\hat{\eta} = \frac{b_1 + b_2\left[\ln\left(\overline{TOTEXP}\right)+1\right]}{b_1 + b_2\ln\left(\overline{TOTEXP}\right)} = \frac{1.0099 - 0.1495\times\left[\ln(94.848)+1\right]}{1.0099 - 0.1495\times\ln(94.848)} = 0.5461$$

For households with 2 children, the average total expenditure is 101.168 and

$$\hat{\eta} = \frac{b_1 + b_2\left[\ln\left(\overline{TOTEXP}\right)+1\right]}{b_1 + b_2\ln\left(\overline{TOTEXP}\right)} = \frac{0.9535 - 0.12944\times\left[\ln(101.168)+1\right]}{0.9535 - 0.12944\times\ln(101.168)} = 0.6363$$

Both of the elasticities are less than one; therefore, food is a necessity.

Exercise 4.10 (continued)

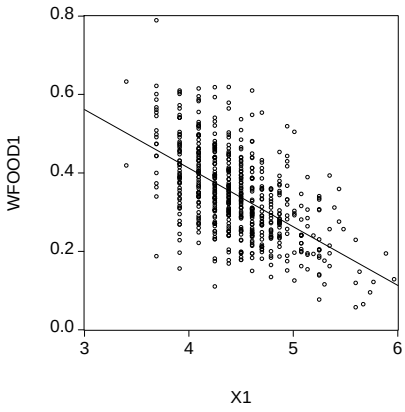

Figure xr4.10(a)

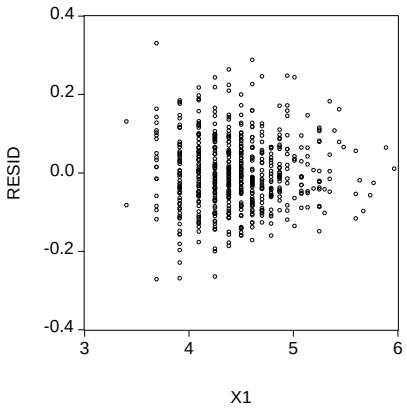

Figure xr4.10(b)

(c) Figures xr4.10 (a) and (b) display the fitted curve and the residual plot for households with 1 child. The function linear in *WFOOD* and ln(*TOTEXP*) seems to be an appropriate one. However, the observations vary considerably around the fitted line, consistent with the low R^2 value. Also, the absolute magnitude of the residuals appears to decline as ln(*TOTEXP*) increases. In Chapter 8 we discover that such behavior suggests the existence of heteroskedasticity.

Figures xr4.10 (c) and (d) are plots of the fitted equation and the residuals for households with 2 children. They lead to similar conclusions to those made for the one-child case.

The values of *JB* for testing H_0: the errors are normally distributed are 10.7941 and 6.3794 for households with 1 child and 2 children, respectively. Since both values are greater than the critical value $\chi^2_{(0.95,2)} = 5.991$, we reject H_0. The *p*-values obtained are 0.0045 and 0.0412, respectively, confirming that H_0 is rejected. We conclude that for both cases the errors are not normally distributed.

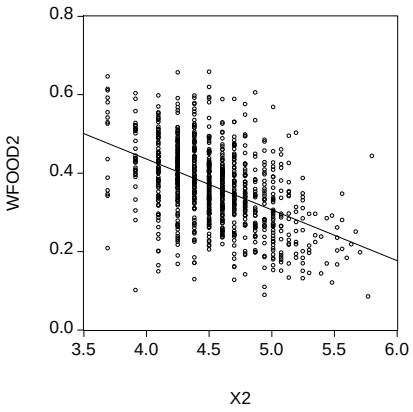

Figure xr4.10(c)

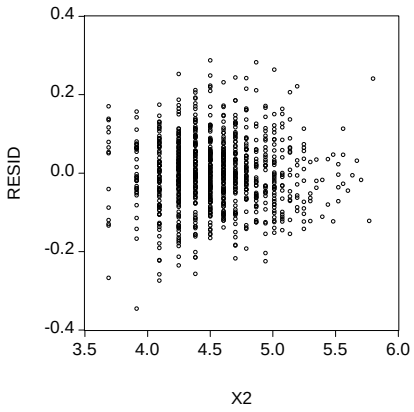

Figure xr4.10(d)

EXERCISE 4.11

(a) Regression results:

$$\widehat{VOTE} = 51.9387 + 0.6599GROWTH \qquad R^2 = 0.3608$$
$$\text{(se)} \quad (0.9054)\ (0.1631)$$
$$(t) \quad (57.3626)(4.4060)$$

Predicted value of *VOTE* in 2000:

$$\widehat{VOTE}_0 = 51.9387 + 0.6599 \times 1.603 = 52.9965$$

Least squares residual:

$$VOTE_0 - \widehat{VOTE}_0 = 50.2650 - 52.9965 = -2.7315$$

(b) Estimated regression:

$$\widehat{VOTE} = 52.0281 + 0.6631GROWTH$$
$$\text{(se)} \quad (0.931) \quad (0.1652)$$

Predicted value of *VOTE* in 2000:

$$\widehat{VOTE}_0 = 52.0281 + 0.6631 \times 1.603 = 53.0910$$

Prediction error in forecast:

$$f = VOTE_0 - \widehat{VOTE}_0 = 50.2650 - 53.0910 = -2.8260$$

This prediction error is larger in magnitude than the least squares residual. This result is expected because the estimated regression in part (b) does not contain information about *VOTE* in the year 2000.

(c) 95% prediction interval:

$$\widehat{VOTE}_0 \pm t_{(0.975,28)} \times \text{se}(f) = 53.091 \pm 2.048 \times 5.1648 = (42.513,\ 63.669)$$

(d) The non-incumbent party will receive 50.1% of the vote if the incumbent party receives 49.9% of the vote. Thus, we want the value of *GROWTH* for which

$$49.9 = 52.0281 + 0.6631 \times GROWTH$$

Solving for *GROWTH* yields

$$GROWTH = -3.209$$

Real per capita GDP would have had to decrease by 3.209% in the first three quarters of the election year for the non-incumbent party to win 50.1% of the vote.

EXERCISE 4.12

(a) Estimated regression:

$$\widehat{STARTS}_0 = 2992.739 - 194.2334 \times FIXED\_RATE_0$$

In May 2005: $\widehat{STARTS} = 2992.739 - 194.2334 \times 6.00 = 1827$

In June 2005: $\widehat{STARTS} = 2992.739 - 194.2334 \times 5.82 = 1862$

(b) Prediction error for May 2005:

$$f = STARTS_0 - \widehat{STARTS}_0 = 2041 - 1827 = 214$$

Prediction error for June 2005:

$$f = STARTS_0 - \widehat{STARTS}_0 = 2065 - 1862 = 203$$

(c) Prediction interval for May 2005:

$$\widehat{STARTS}_0 \pm t_{(0.975,182)} \times \text{se}(f) = 1827 \pm 1.973 \times 159.58 = (1512, 2142)$$

Prediction interval for June 2005:

$$\widehat{STARTS}_0 \pm t_{(0.975,182)} \times \text{se}(f) = 1862 \pm 1.973 \times 159.785 = (1547, 2177)$$

Both prediction intervals contained the true values.

EXERCISE 4.13

(a) Regression results:

$$\ln(PRICE) = 10.5938 + 0.000596\,SQFT$$
$$(\text{se}) \qquad (0.0219)(0.000013)$$
$$(t) \qquad (484.84)\ \ (46.30)$$

The intercept 10.5938 is the value of ln(*PRICE*) when the area of the house is zero. This is an unrealistic and unreliable value since there are no prices for houses of zero area. The coefficient 0.000596 suggests an increase of one square foot is associated with a 0.06% increase in the price of the house.

To find the slope $d(PRICE)/d(SQFT)$ we note that

$$\frac{d\ln(PRICE)}{dSQFT} = \frac{d\ln(PRICE)}{dPRICE} \times \frac{dPRICE}{dSQFT} = \frac{1}{PRICE} \times \frac{dPRICE}{dSQFT} = \beta_2$$

Therefore

$$\frac{dPRICE}{dSQFT} = \beta_2 \times PRICE$$

At the mean

$$\frac{dPRICE}{dSQFT} = \beta_2 \times \overline{PRICE} = 0.00059596 \times 112810.81 = 67.23$$

The value 67.23 is interpreted as the increase in price associated with a 1 square foot increase in living area at the mean.

The elasticity is calculated as:

$$\beta_2 \times SQFT = \frac{1}{PRICE} \times \frac{dPRICE}{dSQFT} \times SQFT = \frac{dPRICE/PRICE}{dSQFT/SQFT} = \frac{\%\Delta PRICE}{\%\Delta SQFT}$$

At the mean,

$$\text{elasticity} = \beta_2 \times \overline{SQFT} = 0.00059596 \times 1611.9682 = 0.9607$$

This result tells us that, at the mean, a 1% increase in area is associated with an approximate 1% increase in the price of the house.

Exercise 4.13 (continued)

(b) Regression results:

$$\ln(PRICE) = 4.1707 + 1.0066\ln(SQFT)$$
$$(\text{se}) \quad (0.1655) \quad (0.0225)$$
$$(t) \quad (25.20) \quad (44.65)$$

The intercept 4.1707 is the value of ln(*PRICE*) when the area of the house is 1 square foot. This is an unrealistic and unreliable value since there are no prices for houses of 1 square foot in area. The coefficient 1.0066 says that an increase in living area of 1% is associated with a 1% increase in house price.

The coefficient 1.0066 is the elasticity since it is a constant elasticity functional form.

To find the slope $d(PRICE)/d(SQFT)$ note that

$$\frac{d\ln(PRICE)}{d\ln(SQFT)} = \frac{SQFT}{PRICE}\frac{dPRICE}{dSQFT} = \beta_2$$

Therefore,

$$\frac{dPRICE}{dSQFT} = \beta_2 \times \frac{PRICE}{SQFT}$$

At the means,

$$\frac{dPRICE}{dSQFT} = \beta_2 \times \frac{\overline{PRICE}}{\overline{SQFT}} = 1.0066 \times \frac{112810.81}{1611.9682} = 70.444$$

The value 70.444 is interpreted as the increase in price associated by a 1 square foot increase in living area at the mean.

(c) From the linear function, $R^2 = 0.672$.

From the log-linear function in part(a),

$$R_g^2 = [\text{corr}(y, \hat{y})]^2 = \frac{[\text{cov}(y, \hat{y})]^2}{\text{var}(y)\,\text{var}(\hat{y})} = \frac{\left[1.99573 \times 10^9\right]^2}{2.78614 \times 10^9 \times 1.99996 \times 10^9} = 0.715$$

From the log-log function in part(b),

$$R_g^2 = [\text{corr}(y, \hat{y})]^2 = \frac{[\text{cov}(y, \hat{y})]^2}{\text{var}(y)\,\text{var}(\hat{y})} = \frac{\left[1.57631 \times 10^9\right]^2}{2.78614 \times 10^9 \times 1.32604 \times 10^9} = 0.673$$

The highest R^2 value is that of the log-linear functional form. The linear association between the data and the fitted line is highest for the log-linear functional form. In this sense the log-linear model fits the data best.

Exercise 4.13 (continued)

(d)

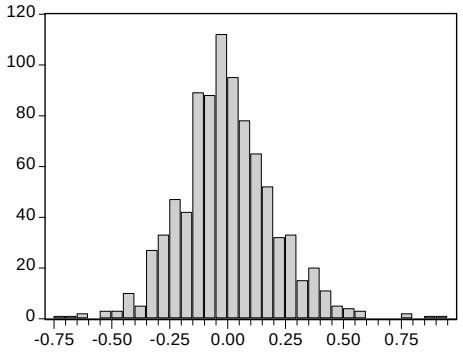

Figure xr4.13(a) Histogram of residuals for log-linear model

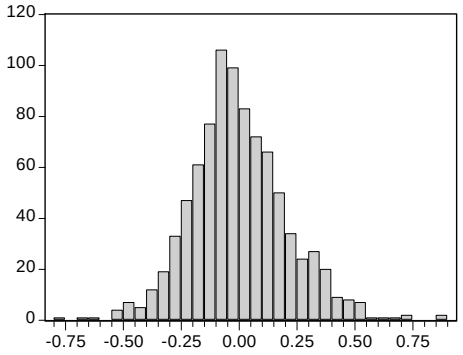

Figure xr4.13(b) Histogram of residuals for log-log model

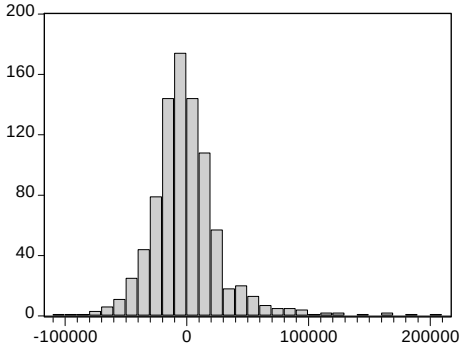

Figure xr4.13(c) Histogram of residuals for simple linear model

Log-linear:	Jarque-Bera = 78.85,	p -value = 0.0000
Log-Log:	Jarque-Bera = 52.74,	p -value = 0.0000
Simple linear:	Jarque-Bera = 2456,	p -value = 0.0000

All Jarque-Bera values are significantly different from 0 at the 1% level of significance. We can conclude that the residuals are not normally distributed.

Exercise 4.13 (continued)

(e)

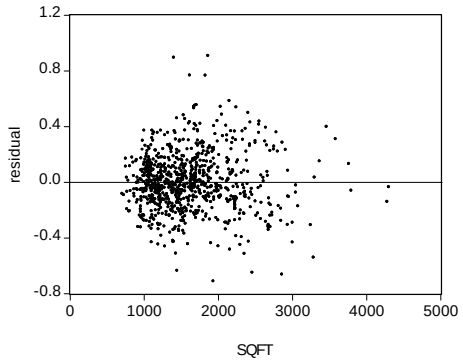

Figure xr4.13(d) Residuals of log-linear model

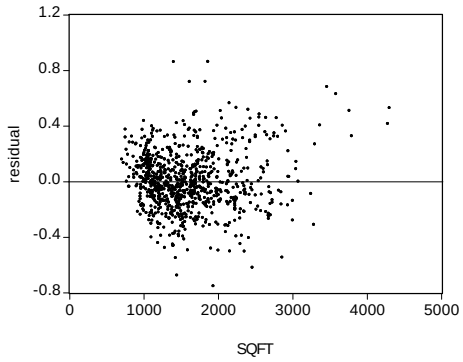

Figure xr4.13(e) Residuals of log-log model

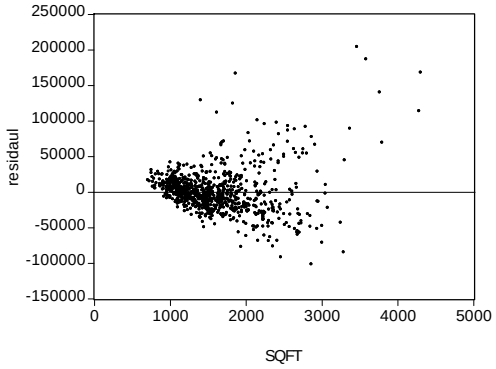

Figure xr4.13(f) Residuals of simple linear model

The residuals appear to increase in magnitude as *SQFT* increases. This is most evident in the residuals of the simple linear functional form. Furthermore, the residuals in the area around 1000 square feet of the simple linear model are all positive indicating that perhaps the functional form does not fit well in this region.

Exercise 4.13 (continued)

(f) Prediction for log-linear model:

$$\widehat{PRICE} = \exp\left(b_1 + b_2 SQFT + \hat{\sigma}^2/2\right)$$
$$= \exp\left(10.59379 + 0.000595963 \times 2700 + 0.20303^2/2\right)$$
$$= 203,516$$

Prediction for log-log model:

$$\widehat{PRICE} = \exp\left(4.170677 + 1.006582 \times \log(2700) + 0.208251^2/2\right)$$
$$= 188,221$$

Prediction for simple linear model:

$$\widehat{PRICE} = -18385.65 + 81.3890 \times 2700 = 201,365$$

(g) The standard error of forecast for the log-linear model is

$$\text{se}(f) = \sqrt{\hat{\sigma}^2\left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]}$$
$$= 0.203034\sqrt{1 + \frac{1}{880} + \frac{(2700 - 1611.968)^2}{248768933.1}} = 0.20363$$

The 95% confidence interval for the prediction from the log-linear model is:

$$\exp\left(\widehat{\ln(y)} \pm t_{(0.975,878)}\text{se}(f)\right)$$
$$= \exp\left(10.59379 + 0.000595963 \times 2700 \pm 1.96267 \times 0.20363\right)$$
$$= [133,683;\ 297,316]$$

The standard error of forecast for the log-log model is

$$\text{se}(f) = 0.208251\sqrt{1 + \frac{1}{880} + \frac{(7.90101 - 7.3355)^2}{85.34453}} = 0.20876$$

The 95% confidence interval for the prediction from the log-log model is

$$\exp\left(\widehat{\ln(y)} \pm t_{(0.975,878)}\text{se}(f)\right)$$
$$= \exp\left(4.170677 + 1.006582 \times \log(2700) \pm 1.96267 \times 0.20876\right)$$
$$= [122,267;\ 277,454]$$

Exercise 4.13(g) (continued)

(g) The standard error of forecast for the simple linear model is

$$\text{se}(f) = 30259.2\sqrt{1+\frac{1}{880}+\frac{(2700-1611.968)^2}{248768933.1}} = 30348.26$$

The 95% confidence interval for the prediction from the simple linear model is

$$\hat{y}_0 \pm t_{(0.975,878)}\text{se}(f) = 201,364.62 \pm 1.96267 \times 30,348.26$$
$$= (141,801;\ 260,928)$$

(h) The simple linear model is not a good choice because the residuals are heavily skewed to the right and hence far from being normally distributed. It is difficult to choose between the other two models – the log-linear and log-log models. Their residuals have similar patterns and they both lead to a plausible elasticity of price with respect to changes in square feet, namely, a 1% change in square feet leads to a 1% change in price. The log-linear model is favored on the basis of its higher R_g^2 value, and its smaller standard deviation of the error, characteristics that suggest it is the model that best fits the data.

EXERCISE 4.14

(a)

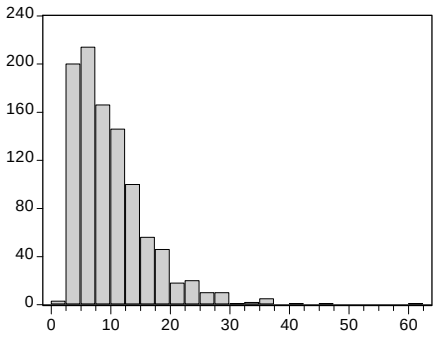

Figure xr4.14(a) Histogram of *WAGE*

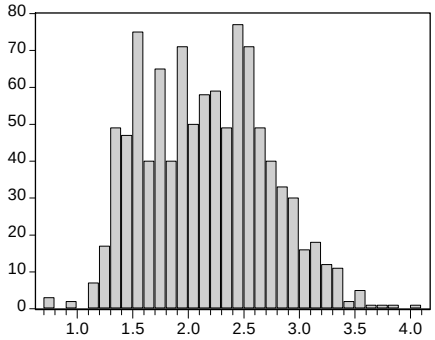

Figure xr4.14(b) Histogram of ln(*WAGE*)

Neither *WAGE* nor ln(*WAGE*) appear normally distributed. The distribution for *WAGE* is positively skewed and that for ln(*WAGE*) is too flat at the top. However, ln(*WAGE*) more closely resembles a normal distribution. This conclusion is confirmed by the Jarque-Bera test results which are $JB = 2684$ (p-value = 0.0000) for *WAGE* and $JB = 17.6$ (p-value = 0.0002) for ln(*WAGE*).

(b) The regression results for the linear model are

$$\widehat{WAGE} = -4.9122 + 1.1385EDUC \qquad R^2 = 0.2024$$
$$(\text{se}) \quad (0.9668)\ (0.0716)$$

$$\text{The estimated return to education at the mean} = \frac{b_2}{\overline{WAGE}} \times 100 = \frac{1.1385}{10.2130} \times 100 = 11.15\%$$

The results for the log-linear model are

$$\ln\left(\widehat{WAGE}\right) = 0.7884 + 0.1038EDUC \qquad R^2 = 0.2146$$
$$(\text{se}) \quad (0.0849)\ (0.0063)$$

$$\text{The estimated return to education} = b_2 \times 100 = 10.38\%.$$

Exercise 4.14 (continued)

(c)

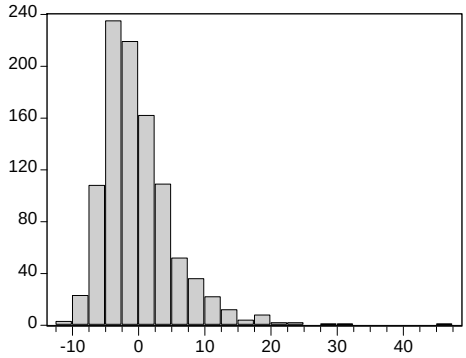

Figure xr4.14(c) Histogram of residuals from simple linear regression

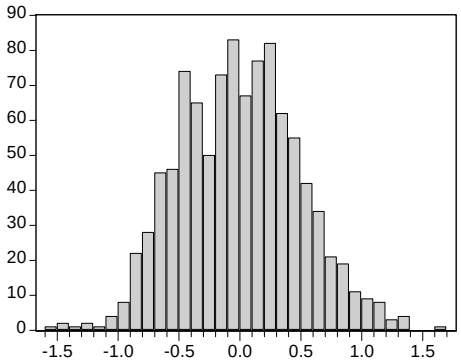

Figure xr4.14(d) Histogram of residuals from log-linear regression

The Jarque-Bera test results are $JB = 3023$ (p-value = 0.0000) for the residuals from the linear model and $JB = 3.48$ (p-value = 0.1754) for the residuals from the log-linear model.

Both the histograms and the Jarque-Bera test results suggest the residuals from the log-linear model are more compatible with normality. In the log-linear model a null hypothesis of normality is not rejected at a 10% level of significance. In the linear regression model it is rejected at a 1% level of significance.

(d) Linear model: $R^2 = 0.2024$

Log-linear model:

$$R_g^2 = [\text{corr}(y, \hat{y})]^2 = \frac{[\text{cov}(y, \hat{y})]^2}{\text{var}(y)\,\text{var}(\hat{y})} = \frac{6.87196^2}{38.9815 \times 5.39435} = 0.2246$$

Since, $R_g^2 > R^2$ we conclude that the log-linear model fits the data better.

Exercise 4.14 (continued)

(e)

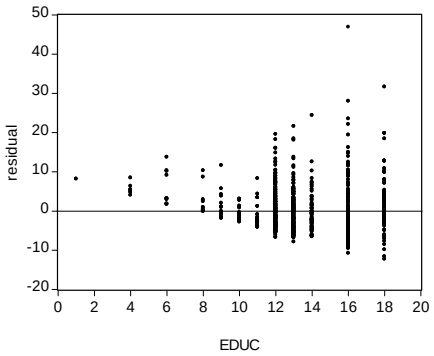

Figure xr4.14(e) Residuals of the simple linear model

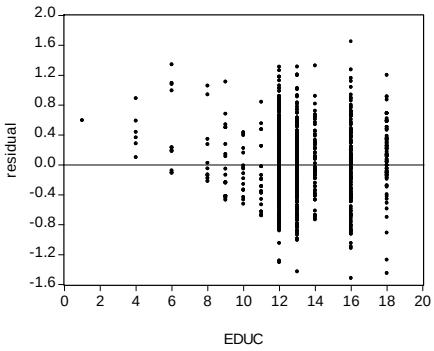

Figure xr4.14(f) Residuals of the log-linear model

The absolute value of the residuals increases in magnitude as *EDUC* increases, suggesting heteroskedasticity which is covered in Chapter 8. It is also apparent, for both models, that there are only positive residuals in the early range of *EDUC*. This suggests that there might be a threshold effect – education has an impact only after a minimum number of years of education. We also observe the non-normality of the residuals in the linear model; the positive residuals tend to be greater in absolute magnitude than the negative residuals.

(f) Prediction for simple linear model:

$$\widehat{WAGE}_0 = -4.9122 + 1.1385 \times 16 = 13.30$$

Prediction for log-linear model:

$$\widehat{WAGE}_c = \exp\left(0.7884 + 0.1038 \times 16 + (0.4902^2)/2\right) = 13.05$$

Actual average wage of all workers with 16 years of education = 13.30

(g) The log-linear function is preferred because it has a higher goodness-of-fit value and its residuals are consistent with normality. However, when predicting the average age of workers with 16 years of education, the linear model had a smaller prediction error

EXERCISE 4.15

Results using *cps_small.dat*

(a), (b)

Summary statistics for *WAGE*

Sub-sample	Mean	Std Dev	Min	Max	CV
(i) all males	11.525	6.659	2.07	60.19	57.8
(ii) all females	8.869	5.484	2.03	41.32	61.8
(iii) all whites	10.402	6.343	2.03	60.19	61.0
(iv) all blacks	8.259	4.740	3.50	25.26	57.4
(v) white males	11.737	6.716	2.07	60.19	57.2
(vi) white females	9.007	5.606	2.03	41.32	62.2
(vii) black males	9.066	5.439	3.68	25.26	60.0
(viii) black females	7.586	4.003	3.50	18.44	52.8

These results show that, on average, white males have the highest wages and black females the lowest. The wage of white females is approximately the same as that of black males. White females have the highest coefficient of variation and black females have the lowest.

(c)

Regression results

Sub-sample	Constant	*EDUC*	% return	R^2
(i) all males	1.0075	0.0967	9.67	0.2074
(se)	(0.1144)	(0.0084)		
(ii) all females	0.5822	0.1097	10.97	0.2404
(se)	(0.1181)	(0.0088)		
(iii) all whites	0.7822	0.1048	10.48	0.2225
(se)	(0.0881)	(0.0065)		
(iv) all blacks	1.0185	0.0744	7.44	0.1022
(se)	(0.3108)	(0.0238)		
(v) white males	0.9953	0.0987	9.87	0.2173
(se)	(0.1186)	(0.0087)		
(vi) white females	0.6099	0.1085	10.85	0.2429
(se)	(0.1223)	(0.0091)		
(vii) black males	1.3809	0.0535	5.35	0.0679
(se)	(0.4148)	(0.0321)		
(viii) black females	0.2428	0.1275	12.75	0.2143
(se)	(0.4749)	(0.0360)		

The return to education is highest for black females (12.75%) and lowest for black males (5.35%). It is approximately 10% for all other sub-samples with the exception of all blacks where it is around 7.5%.

Exercise 4.15 (continued)

Results using *cps_small.dat*

(d) The model does not fit the data equally well for each sub-sample. The best fits are for all females and white females. Those for all blacks and black males are particularly poor.

(e) The *t*-value for testing $H_0: \beta_2 = 0.10$ against $H_1: \beta_2 \neq 0.10$ is given by

$$t = \frac{b_2 - 0.1}{\text{se}(b_2)}$$

We reject H_0 if $t > t_c$ or $t < -t_c$ where $t_c = t_{(0.975, \text{df})}$. The results are given in the following table.

Test results for $H_0: \beta_2 = 0.10$ versus $H_1: \beta_2 \neq 0.10$

Sub-sample	*t*-value	df	t_c	*p*-value	Decision
(i) all males	– 0.394	504	1.965	0.6937	Fail to reject H_0
(ii) all females	1.103	492	1.965	0.2707	Fail to reject H_0
(iii) all whites	0.745	910	1.963	0.4563	Fail to reject H_0
(iv) all blacks	– 1.074	86	1.988	0.2856	Fail to reject H_0
(v) white males	– 0.149	464	1.965	0.8817	Fail to reject H_0
(vi) white females	0.931	444	1.965	0.3525	Fail to reject H_0
(vii) black males	– 1.447	38	2.024	0.1560	Fail to reject H_0
(viii) black females	0.764	46	2.013	0.4485	Fail to reject H_0

There are no sub-samples where the data contradict the assertion that the wage return to an extra year of education is 10%. Thus, although the estimated return to education is much lower for all blacks and black males, it is not sufficiently less to conclude conclusively it is not equal to 10%.

EXERCISE 4.15

Results using *cps.dat*

(a), (b)

Summary statistics for *WAGE*

Sub-sample	Mean	Std Dev	Min	Max	CV
(i) all males	11.315	6.521	1.05	74.32	57.6
(ii) all females	8.990	5.630	1.28	78.71	62.6
(iii) all whites	10.358	6.275	1.05	78.71	60.6
(iv) all blacks	8.626	5.387	1.57	39.35	62.5
(v) white males	11.491	6.591	1.05	74.32	57.4
(vi) white females	9.105	5.648	1.28	78.71	62.0
(vii) black males	9.307	5.274	2.76	34.07	56.7
(viii) black females	8.129	5.424	1.57	39.35	66.7

These results show that, on average, white males have the highest wages and black females the lowest. Males have higher average wages than females and whites have higher average wages than blacks. The highest wage earner is, however, a white female. Black females have the highest coefficient of variation and black males have the lowest.

(c)

Regression results

Sub-sample	Constant	*EDUC*	% return	R^2
(i) all males	0.9798	0.0982	9.82	0.1954
(se)	(0.0543)	(0.0040)		
(ii) all females	0.4776	0.1173	11.73	0.2479
(se)	(0.0579)	(0.0043)		
(iii) all whites	0.7965	0.1040	10.40	0.2030
(se)	(0.0428)	(0.0032)		
(iv) all blacks	0.6230	0.1066	10.66	0.1800
(se)	(0.1390)	(0.0106)		
(v) white males	0.9859	0.0988	9.88	0.2009
(se)	(0.0561)	(0.0042)		
(vi) white females	0.5142	0.1152	11.52	0.2453
(se)	(0.0611)	(0.0045)		
(vii) black males	1.0641	0.0798	7.98	0.1167
(se)	(0.2063)	(0.0157)		
(viii) black females	0.2147	0.1327	13.27	0.2569
(se)	(0.1820)	(0.0138)		

The return to education is highest for black females (13.27%) and lowest for black males (7.98%). It is approximately 10% for all other sub-samples with the exception of all females and white females where it is around 11.5%.

Exercise 4.15 (continued)

Results using *cps.dat*

(d) The model does not fit the data equally well for each sub-sample. The best fits are for all females, white females and black females. That for black males is particularly poor.

(e) The *t*-value for testing $H_0: \beta_2 = 0.10$ against $H_1: \beta_2 \neq 0.10$ is given by

$$t = \frac{b_2 - 0.1}{\text{se}(b_2)}$$

We reject H_0 if $t > t_c$ or $t < -t_c$ where $t_c = t_{(0.975, \text{df})}$. The results are given in the following table.

Test results for $H_0: \beta_2 = 0.10$ versus $H_1: \beta_2 \neq 0.10$

Sub-sample	*t*-value	df	t_c	*p*-value	Decision
(i) all males	– 0.444	2435	1.961	0.6568	Fail to reject H_0
(ii) all females	4.023	2294	1.961	0.0001	Reject H_0
(iii) all whites	1.276	4264	1.961	0.2019	Fail to reject H_0
(iv) all blacks	0.629	465	1.965	0.5294	Fail to reject H_0
(v) white males	– 0.296	2238	1.961	0.7669	Fail to reject H_0
(vi) white females	3.385	2024	1.961	0.0007	Reject H_0
(vii) black males	– 1.284	195	1.972	0.2005	Fail to reject H_0
(viii) black females	2.370	268	1.969	0.0185	Reject H_0

The null hypothesis is rejected for females, white females and black females. In these cases the wage return to an extra year of education is estimated as greater than 10%. In all other sub-samples, the data do not contradict the assertion that the wage return is 10%.

EXERCISE 4.16

(a) Regression results:

$$\widehat{BUCHANAN} = 65.503 + 0.003482\,BUSH \qquad R^2 = 0.7535$$
$$(\text{se}) \qquad (17.293) \quad (0.000249)$$
$$(t) \qquad (3.788) \quad (13.986)$$

The R^2 tells us that 75.35% of the variation in votes for Pat Buchanan are explained by variation in the votes for George Bush (excluding Palm Beach).

(b) The vote in Palm Beach for George Bush is 152,846. Therefore, the predicted vote for Pat Buchanan is:

$$\widehat{BUCHANAN}_0 = 65.503 + 0.003482 \times 152{,}846 = 598$$

$$\text{se}(f) = 112.2647\sqrt{1 + \frac{1}{66} + \frac{(152{,}846 - 41761.9697)^2}{2.0337296 \times 10^{11}}} = 116.443$$

The 99.9% confidence interval is

$$\hat{y}_0 \pm t_{(0.9995,\,66)} \times \text{se}(f) = 597.7 \pm 3.449 \times 116.443 = (196,\ 999)$$

The actual vote for Pat Buchanan in Palm Beach was 3407 which is not in the prediction interval. The model is clearly not a good one for explaining the Palm Beach vote. This conclusion is confirmed by the scatter diagram in part (c).

(c)

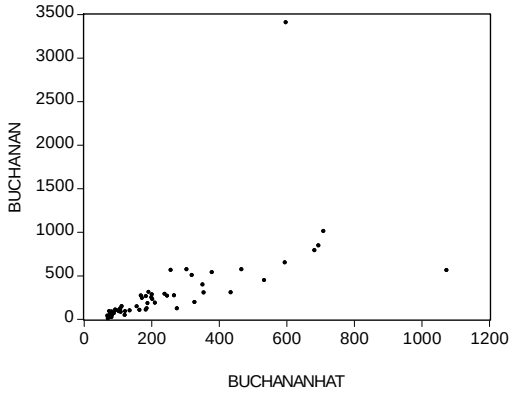

Figure xr4.16(a) Predictions versus actual observations on Buchanan vote

Exercise 4.16 (continued)

(d) Regression results:

$$\widehat{BUCHANAN} = 109.23 + 0.002544\,GORE \qquad R^2 = 0.6305$$
$$(\text{se}) \qquad (19.52) \quad (0.000243)$$
$$(t) \qquad (5.596) \quad (10.450)$$

The R^2 tells us that 63.05% of the variation in votes for Pat Buchanan are explained by variation in the votes for Al Gore (excluding Palm Beach).

The vote in Palm Beach for Al Gore is 268,945. Therefore, the predicted vote for Pat Buchanan is:

$$\widehat{BUCHANAN}_0 = 109.23 + 0.002544 \times 268945 = 793$$

$$\text{se}(f) = 137.4493\sqrt{1 + \frac{1}{66} + \frac{(268{,}945 - 39975.55)^2}{3.188628 \times 10^{11}}} = 149.281$$

The 99.9% confidence interval is

$$\hat{y}_0 \pm t_{(0.9995,\,66)} \times \text{se}(f) = 793.3 \pm 3.449 \times 149.281 = (278,\ 1308)$$

The actual vote for Pat Buchanan in Palm Beach was 3407 which is not in the prediction interval. The model is clearly not a good one for explaining the Palm Beach vote. This conclusion is confirmed by the scatter diagram below.

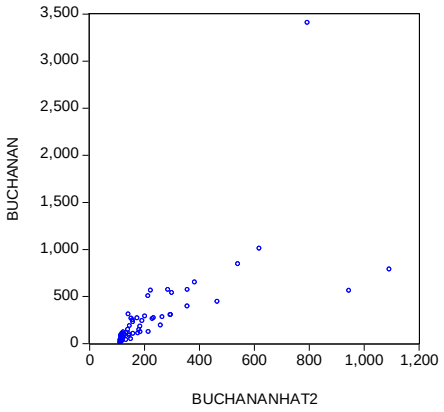

Figure xr4.16(b) Predictions versus actual observations on Buchanan vote

Exercise 4.16 (continued)

(e) Regression results:

$$\widehat{BUCHSHARE} = -0.0017 + 0.01142\,BUSHSHARE \qquad R^2 = 0.1004$$
$$(\text{se}) \qquad (0.0024)\ (0.00427)$$
$$(t) \qquad (-0.710)\ \ (2.673)$$

The share of votes for George Bush in Palm Beach was 0.354827. Therefore, the predicted share of votes in Palm Beach for Pat Buchanan is:

$$\widehat{BUCHSHARE}_0 = -0.001706 + 0.011424 \times 0.354827 = 0.002348$$

The standard error of the forecast error is

$$\text{se}(f) = 0.003078\sqrt{1 + \frac{1}{66} + \frac{(0.354827 - 0.554756)^2}{0.518621}} = 0.0032168$$

A 99.9% confidence interval is given by

$$\hat{y}_0 \pm t_{(0.9995,\,66)} \times \text{se}(f) = 0.002349 \pm 3.449 \times 0.0032168 = (-0.0087457,\ 0.0134437)$$

There were 430,762 total votes cast in Palm Beach. Multiplying the confidence interval endpoints by this figure yields $(-3767,\ 5791)$. The actual vote for Pat Buchanan in Palm Beach was 3407 which falls inside this interval.