CHAPTER 4

Prediction, Goodness-of-Fit, and Modeling Issues

LEARNING OBJECTIVES

Based on the material in this chapter, you should be able to

1. Explain how to use the simple linear regression model to predict the value of y for a given value of x.
2. Explain, intuitively and technically, why predictions for x values further from $\bar{x}$ are less reliable.
3. Explain the meaning of *SST*, *SSR*, and *SSE*, and how they are related to R^2.
4. Define and explain the meaning of the coefficient of determination.
5. Explain the relationship between correlation analysis and R^2.
6. Report the results of a fitted regression equation in such a way that confidence intervals and hypothesis tests for the unknown coefficients can be constructed quickly and easily.
7. Describe how estimated coefficients and other quantities from a regression equation will change when the variables are scaled. Why would you want to scale the variables?
8. Appreciate the wide range of nonlinear functions that can be estimated using a model that is linear in the parameters.
9. Write down the equations for the log-log, log-linear, and linear-log functional forms.
10. Explain the difference between the slope of a functional form and the elasticity from a functional form.
11. Explain how you would go about choosing a functional form and deciding that a functional form is adequate.
12. Explain how to test whether the equation ''errors'' are normally distributed.
13. Explain how to compute a prediction, a prediction interval, and a goodness-of-fit measure in a log-linear model.
14. Explain alternative methods for detecting unusual, extreme, or incorrect data values.

KEYWORDS

coefficient of determination
correlation
forecast error
functional form
goodness-of-fit
growth model
influential observations
Jarque–Bera test
kurtosis
least squares predictor
linear model
linear relationship
linear-log model
log-linear model
log-log model
log-normal distribution
prediction
prediction interval
R^2
residual diagnostics
scaling data
skewness
standard error of the forecast

In Chapter 3, we focused on making statistical inferences, constructing confidence intervals, and testing hypotheses about regression parameters. Another purpose of the regression model, and the one we focus on first in this chapter, is **prediction**. A prediction is a forecast of an unknown value of the dependent variable y given a particular value of x. A **prediction interval**, much like a confidence interval, is a range of values in which the unknown value of y is likely to be located. Examining the **correlation** between sample values of y and their predicted values provides a **goodness-of-fit** measure called $\boldsymbol{R^2}$ that describes how well our model fits the data. For each observation in the sample, the difference between the predicted value of y and the actual value is a **residual**. Diagnostic measures constructed from the residuals allow us to check the adequacy of the **functional form** used in the regression analysis and give us some indication of the validity of the regression assumptions. We will examine each of these ideas and concepts in turn.

4.1 Least Squares Prediction

In Example 2.4, we briefly introduced the idea that the least squares estimates of the linear regression model provide a way to predict the value of y for any value of x. The ability to predict is important to business economists and financial analysts who attempt to forecast the sales and revenues of specific firms; it is important to government policymakers who attempt to predict the rates of growth in national income, inflation, investment, saving, social insurance program expenditures, and tax revenues; and it is important to local businesses who need to have predictions of growth in neighborhood populations and income so that they may expand or contract their provision of services. Accurate predictions provide a basis for better decision making in every type of planning context. In this section, we explore the use of linear regression as a tool for prediction.

Given the simple linear regression model and assumptions SR1–SR6, let x_0 be a given value of the explanatory variable. We want to predict the corresponding value of y, which we call y_0. In order to use regression analysis as a basis for prediction, we must assume that y_0 and x_0 are related to one another by the same regression model that describes our sample of data, so that, in particular, SR1 holds for these observations

$$y_0 = \beta_1 + \beta_2 x_0 + e_0 \tag{4.1}$$

where e_0 is a random error. We assume that $E(y_0|x_0) = \beta_1 + \beta_2 x_0$ and $E(e_0) = 0$. We also assume that e_0 has the same variance as the regression errors, $\text{var}(e_0) = \sigma^2$, and e_0 is uncorrelated with the random errors that are part of the sample data, so that $\text{cov}(e_0, e_i|\mathbf{x}) = 0$, $i = 1, 2, \ldots, N$.

The task of **predicting** y_0 is related to the problem of **estimating** $E(y_0|x_0) = \beta_1 + \beta_2 x_0$, which we discussed in Section 3.6. The outcome $y_0 = E(y_0|x_0) + e_0 = \beta_1 + \beta_2 x_0 + e_0$ is composed of two parts, the systematic, nonrandom part $E(y_0|x_0) = \beta_1 + \beta_2 x_0$, and a random

component e_0. We estimate the systematic portion using $\hat{E}(y_0|x_0) = b_1 + b_2x_0$ and add an "estimate" of e_0 equal to its expected value, which is zero. Therefore, the prediction $\hat{y}_0$ is given by $\hat{y}_0 = \hat{E}(y_0|x_0) + 0 = b_1 + b_2x_0$. Despite the fact that we use the same statistic for both $\hat{y}_0$ and $\hat{E}(y_0|x_0)$, we distinguish between them because, although $E(y_0|x_0) = \beta_1 + \beta_2x_0$ is not random, the outcome y_0 is random. Consequently, as we will see, there is a difference between the **interval estimate** of $E(y_0|x_0) = \beta_1 + \beta_2x_0$ and the **prediction interval** for y_0.

Following from the discussion in the previous paragraph, the **least squares predictor** of y_0 comes from the fitted regression line

$$\hat{y}_0 = b_1 + b_2x_0 \tag{4.2}$$

That is, the predicted value $\hat{y}_0$ is given by the point on the least squares fitted line where $x = x_0$, as shown in Figure 4.1. How good is this prediction procedure? The least squares estimators b_1 and b_2 are random variables—their values vary from one sample to another. It follows that the least squares predictor $\hat{y}_0 = b_1 + b_2x_0$ must also be random. To evaluate how well this predictor performs, we define the **forecast error**, which is analogous to the least squares residual,

$$f = y_0 - \hat{y}_0 = (\beta_1 + \beta_2x_0 + e_0) - (b_1 + b_2x_0) \tag{4.3}$$

We would like the forecast error to be small, implying that our forecast is close to the value we are predicting. Taking the conditional expected value of f, we find

$$\begin{aligned} E(f|\mathbf{x}) &= \beta_1 + \beta_2x_0 + E(e_0) - \left[E(b_1|\mathbf{x}) + E(b_2|\mathbf{x})\, x_0\right] \\ &= \beta_1 + \beta_2x_0 + 0 - \left[\beta_1 + \beta_2x_0\right] \\ &= 0 \end{aligned}$$

which means, on average, the forecast error is zero, and $\hat{y}_0$ is an **unbiased predictor** of y_0. However, unbiasedness does not necessarily imply that a particular forecast will be close to the actual value. The probability of a small forecast error also depends on the variance of the forecast error. Although we will not prove it, $\hat{y}_0$ is the **best linear unbiased predictor (BLUP)** of y_0 if assumptions SR1–SR5 hold. This result is reasonable given that the least squares estimators b_1 and b_2 are best linear unbiased estimators.

Using (4.3) and what we know about the variances and covariances of the least squares estimators, we can show (see Appendix 4A) that the variance of the forecast error is

$$\text{var}(f|\mathbf{x}) = \sigma^2\left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \tag{4.4}$$

Notice that some of the elements of this expression appear in the formulas for the variances of the least squares estimators and affect the precision of prediction in the same way that they affect the precision of estimation. We would prefer that the variance of the forecast error be small, which

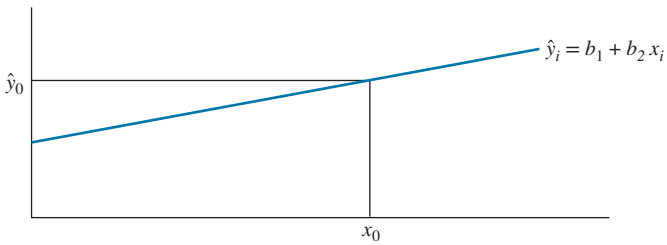

FIGURE 4.1 **A point prediction.**

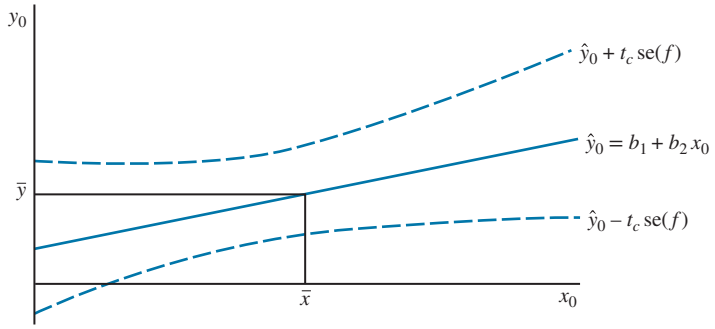

FIGURE 4.2 **Point and interval prediction.**

would increase the probability that the prediction $\hat{y}_0$ is close to the value y_0, we are trying to predict. Note that the variance of the forecast error is smaller when

- **i.** the overall uncertainty in the model is smaller, as measured by the variance of the random errors σ^2
- **ii.** the sample size N is larger
- **iii.** the variation in the explanatory variable is larger
- **iv.** the value of $(x_0 - \bar{x})^2$ is small

The new addition is the term $(x_0 - \bar{x})^2$, which measures how far x_0 is from the center of the x-values. The more distant x_0 is from the center of the sample data the larger the forecast variance will become. Intuitively, this means that we are able to do a better job predicting in the region where we have more sample information, and we will have less accurate predictions when we try to predict outside the limits of our data.

In practice we replace σ^2 in (4.4) by its estimator $\hat{\sigma}^2$ to obtain

$$\widehat{\mathrm{var}}(f|\mathbf{x}) = \hat{\sigma}^2\left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

The square root of this estimated variance is the **standard error of the forecast**

$$\mathrm{se}(f) = \sqrt{\widehat{\mathrm{var}}(f|\mathbf{x})} \tag{4.5}$$

Defining the critical value t_c to be the $100(1-\alpha/2)$-percentile from the t-distribution, we can obtain a $100(1-\alpha)\%$ **prediction interval** as

$$\hat{y}_0 \pm t_c \mathrm{se}(f) \tag{4.6}$$

See Appendix 4A for some details related to the development of this result.

Following our discussion of $\mathrm{var}(f|\mathbf{x})$ in (4.4), the farther x_0 is from the sample mean $\bar{x}$, the larger the variance of the prediction error will be, and the less reliable the prediction is likely to be. In other words, our predictions for values of x_0 close to the sample mean $\bar{x}$ are more reliable than our predictions for values of x_0 far from the sample mean $\bar{x}$. This fact shows up in the size of our prediction intervals. The relationship between point and interval predictions for different values of x_0 is illustrated in Figure 4.2. A point prediction is given by the fitted least squares line $\hat{y}_0 = b_1 + b_2 x_0$. The prediction interval takes the form of two bands around the fitted least squares line. Because the forecast variance increases the farther x_0 is from the sample mean $\bar{x}$, the confidence bands are their narrowest when $x_0 = \bar{x}$, and they increase in width as $|x_0 - \bar{x}|$ increases.

EXAMPLE 4.1 | Prediction in the Food Expenditure Model

In Example 2.4, we predicted that a household with $x_0 = \$2{,}000$ weekly income would spend \$287.61 on food using the calculation

$$\hat{y}_0 = b_1 + b_2 x_0 = 83.4160 + 10.2096(20) = 287.6089$$

Now we are able to attach a "confidence interval" to this prediction. The estimated variance of the forecast error is

$$\begin{aligned}\widehat{\text{var}}(f|\mathbf{x}) &= \hat{\sigma}^2\left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right] \\ &= \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{N} + (x_0 - \bar{x})^2\frac{\hat{\sigma}^2}{\sum(x_i - \bar{x})^2} \\ &= \hat{\sigma}^2 + \frac{\hat{\sigma}^2}{N} + (x_0 - \bar{x})^2\widehat{\text{var}}(b_2|\mathbf{x})\end{aligned}$$

In the last line, we have recognized the estimated variance of b_2 from (2.21). In Example 2.5 we obtained the values $\hat{\sigma}^2 = 8013.2941$ and $\widehat{\text{var}}(b_2|\mathbf{x}) = 4.3818$. For the food expenditure data, $N = 40$ and the sample mean of the explanatory variable is $\bar{x} = 19.6048$. Using these values, we obtain the standard error of the forecast $\text{se}(f) = \sqrt{\widehat{\text{var}}(f|\mathbf{x})} = \sqrt{8214.31} = 90.6328$. If we select $1 - \alpha = 0.95$, then $t_c = t_{(0.975,\,38)} = 2.0244$ and the 95% prediction interval for y_0 is

$$\begin{aligned}\hat{y}_0 \pm t_c\text{se}(f) &= 287.6069 \pm 2.0244(90.6328) \\ &= [104.1323,\ 471.0854]\end{aligned}$$

Our prediction interval suggests that a household with \$2,000 weekly income will spend somewhere between \$104.13 and \$471.09 on food. Such a wide interval means that our point prediction \$287.61 is not very reliable. We have obtained this wide prediction interval for the value of $x_0 = 20$ that is close to the sample mean $\bar{x} = 19.60$. For values of x that are more extreme, the prediction interval would be even wider. The unreliable predictions may be slightly improved if we collect a larger sample of data, which will improve the precision with which we estimate the model parameters. However, in this example the magnitude of the estimated error variance $\hat{\sigma}^2$ is very close to the estimated variance of the forecast error $\widehat{\text{var}}(f|\mathbf{x})$, indicating that the primary uncertainty in the forecast comes from large uncertainty in the model. This should not be a surprise, since we are predicting household behavior, which is a complicated phenomenon, on the basis of a single household characteristic, income. Although income is a key factor in explaining food expenditure, we can imagine that many other household demographic characteristics may play a role. To more accurately predict food expenditure, we may need to include these additional factors into the regression model. Extending the simple regression model to include other factors will begin in Chapter 5.

4.2 Measuring Goodness-of-Fit

Two major reasons for analyzing the model

$$y_i = \beta_1 + \beta_2 x_i + e_i \tag{4.7}$$

are to explain how the dependent variable (y_i) changes as the independent variable (x_i) changes and to predict y_0 given an x_0. These two objectives come under the broad headings of estimation and prediction. Closely allied with the prediction problem discussed in the previous section is the desire to use x_i to explain as much of the variation in the dependent variable y_i as possible. In the regression model (4.7), we call x_i the "explanatory" variable because we hope that its variation will "explain" the variation in y_i.

To develop a measure of the variation in y_i that is explained by the model, we begin by separating y_i into its explainable and unexplainable components. We have assumed that

$$y_i = E(y_i|\mathbf{x}) + e_i \tag{4.8}$$

where $E(y_i|\mathbf{x}) = \beta_1 + \beta_2 x_i$ is the explainable, "systematic" component of y_i, and e_i is the random, unsystematic, and unexplainable component of y_i. While both of these parts are unobservable to us, we can estimate the unknown parameters β_1 and β_2 and, analogous to (4.8), decompose the value of y_i into

$$y_i = \hat{y}_i + \hat{e}_i \tag{4.9}$$

where $\hat{y}_i = b_1 + b_2 x_i$ and $\hat{e}_i = y_i - \hat{y}_i$.

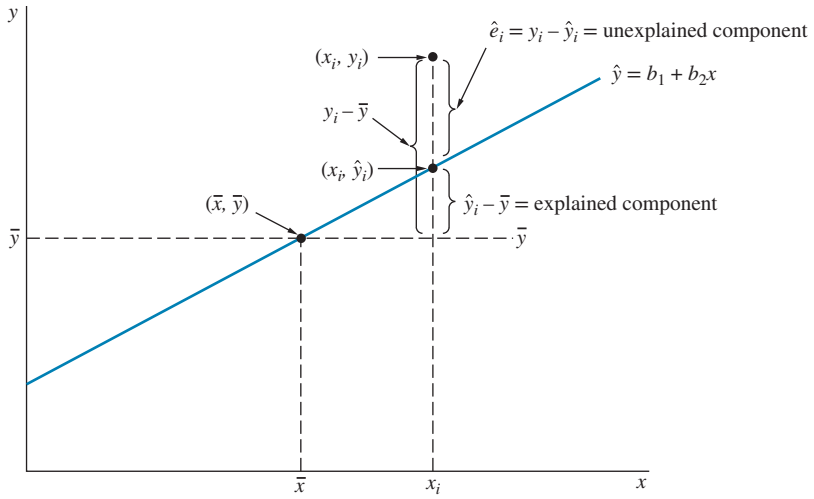

FIGURE 4.3 **Explained and unexplained components of y_i.**

In Figure 4.3, the "point of the means" $(\bar{x}, \bar{y})$ is shown, with the least squares fitted line passing through it. This is a characteristic of the least squares fitted line whenever the regression model includes an intercept term. Subtract the sample mean $\bar{y}$ from both sides of the equation to obtain

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + \hat{e}_i \tag{4.10}$$

As shown in Figure 4.3, the difference between y_i and its mean value $\bar{y}$ consists of a part that is "explained" by the regression model $\hat{y}_i - \bar{y}$ and a part that is unexplained $\hat{e}_i$.

The breakdown in (4.10) leads to a decomposition of the total sample variability in y into explained and unexplained parts. Recall from your statistics courses (see Appendix C4) that if we have a sample of observations $y_1, y_2, \ldots, y_N$, two descriptive measures are the sample mean $\bar{y}$ and the sample variance

$$s_y^2 = \frac{\sum (y_i - \bar{y})^2}{N - 1}$$

The numerator of this quantity, the sum of squared differences between the sample values y_i and the sample mean $\bar{y}$, is a measure of the total variation in the sample values. If we square and sum both sides of (4.10) and use the fact that the cross-product term $\sum (\hat{y}_i - \bar{y})\hat{e}_i = 0$ (see Appendix 4B), we obtain

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum \hat{e}_i^2 \tag{4.11}$$

Equation (4.11) gives us a decomposition of the "total sample variation" in y into explained and unexplained components. Specifically, these "sums of squares" are as follows:

1. $\sum (y_i - \bar{y})^2$ = total sum of squares = *SST*: a measure of *total variation* in y about the sample mean.
2. $\sum (\hat{y}_i - \bar{y})^2$ = sum of squares due to the regression = *SSR*: that part of total variation in y, about the sample mean, that is explained by, or due to, the regression. Also known as the "explained sum of squares."
3. $\sum \hat{e}_i^2$ = sum of squares due to error = *SSE*: that part of total variation in y about its mean that is not explained by the regression. Also known as the unexplained sum of squares, the residual sum of squares, or the sum of squared errors.

Using these abbreviations, (4.11) becomes

$$SST = SSR + SSE$$

This decomposition of the total variation in y into a part that is explained by the regression model and a part that is unexplained allows us to define a measure, called the **coefficient of determination**, or R^2, that is the proportion of variation in y explained by x within the regression model.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{4.12}$$

The closer R^2 is to 1, the closer the sample values y_i are to the fitted regression equation $\hat{y}_i = b_1 + b_2 x_i$. If $R^2 = 1$, then all the sample data fall exactly on the fitted least squares line, so $SSE = 0$, and the model fits the data "perfectly." If the sample data for y and x are uncorrelated and show no linear association, then the least squares fitted line is "horizontal," and identical to $\bar{y}$, so that $SSR = 0$ and $R^2 = 0$. When $0 < R^2 < 1$, it is interpreted as "the proportion of the variation in y about its mean that is explained by the regression model."

4.2.1 Correlation Analysis

In Appendix B.1.5, we discuss the **covariance** and **correlation** between two random variables x and y. The correlation coefficient ρ_{xy} between x and y is defined in (B.21) as

$$\rho_{xy} = \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)}\sqrt{\text{var}(y)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{4.13}$$

In Appendix B, we did not discuss *estimating* the correlation coefficient. We will do so now to develop a useful relationship between the sample correlation coefficient and R^2.

Given a sample of data pairs $(x_i, y_i), i = 1, \ldots, N$, the sample correlation coefficient is obtained by replacing the covariance and standard deviations in (4.13) by their sample analogs:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

where

$$s_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) / (N - 1)$$

$$s_x = \sqrt{\sum (x_i - \bar{x})^2 / (N - 1)}$$

$$s_y = \sqrt{\sum (y_i - \bar{y})^2 / (N - 1)}$$

The sample correlation coefficient r_{xy} has a value between -1 and 1, and it measures the strength of the linear association between observed values of x and y.

4.2.2 Correlation Analysis and R^2

There are two interesting relationships between R^2 and r_{xy} in the simple linear regression model.

1. The first is that $r_{xy}^2 = R^2$. That is, the square of the sample correlation coefficient between the sample data values x_i and y_i is algebraically equal to R^2 in a simple regression model. Intuitively, this relationship makes sense: r_{xy}^2 falls between zero and one and measures the strength of the linear association between x and y. This interpretation is not far from that of R^2: the proportion of variation in y about its mean explained by x in the linear regression model.
2. The second, and more important, relation is that R^2 can also be computed as the square of the sample correlation coefficient between y_i and $\hat{y}_i = b_1 + b_2 x_i$. That is, $R^2 = r_{y\hat{y}}^2$. As such, it measures the linear association, or goodness-of-fit, between the sample data and their predicted values. Consequently, R^2 is sometimes called a measure of "goodness-of-fit." This result is valid not only in simple regression models but also in multiple regression models

that we introduce in Chapter 5. Furthermore, as you will see in Section 4.4, the concept of obtaining a goodness-of-fit measure by predicting y as well as possible and finding the squared correlation coefficient between this prediction and the sample values of y can be extended to situations in which the usual R^2 does not strictly apply.

EXAMPLE 4.2 | Goodness-of-Fit in the Food Expenditure Model

Look at the food expenditure example, Example 2.4, and in particular, the data scatter and fitted regression line in Figure 2.8, and the computer output in Figure 2.9. Go ahead. I will wait until you get back. The question we would like to answer is "How well does our model fit the data?" To compute R^2, we can use the sums of squares

$$SST = \sum(y_i - \bar{y})^2 = 495132.160$$

$$SSE = \sum(y_i - \hat{y}_i)^2 = \sum \hat{e}_i^2 = 304505.176$$

Then

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{304505.176}{495132.160} = 0.385$$

We conclude that 38.5% of the variation in food expenditure (about its sample mean) is explained by our regression model, which uses only income as an explanatory variable. Is this a good R^2? We would argue that such a question is not useful. Although finding and reporting R^2 provides information about the relative magnitudes of the different sources of variation, debates about whether a particular R^2 is "large enough" are not particularly constructive. Microeconomic household behavior is very difficult to explain. With cross-sectional data, R^2 values from 0.10 to 0.40 are very common even with much larger regression models. Macroeconomic analyses using time-series data, which often trend together smoothly over time, routinely report R^2 values of 0.90 and higher. You should *not* evaluate the quality of the model based only on how well it predicts the sample data used to construct the estimates. To evaluate the model, it is as important to consider factors such as the signs and magnitudes of the estimates, their statistical and economic significance, the precision of their estimation, and the ability of the fitted model to predict values of the dependent variable that were not in the estimation sample. Other model diagnostic issues will be discussed in the following section.

Correlation analysis leads to the same conclusions and numbers, but it is worthwhile to consider this approach in more detail. The sample correlation between the y and x sample values is

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{478.75}{(6.848)(112.675)} = 0.62$$

The correlation is positive, indicating a positive association between food expenditure and income. The sample correlation measures the strength of the linear association, with a maximum value of 1. The value $r_{xy} = 0.62$ indicates a non-negligible but less than perfect fit. As expected $r_{xy}^2 = 0.62^2 = 0.385 = R^2$.

EXAMPLE 4.3 | Reporting Regression Results

In any paper where you write the results of a simple regression, with only one explanatory variable, these results can be presented quite simply. The key ingredients are the coefficient estimates, the standard errors (or t-values), an indication of statistical significance, and R^2. Also, when communicating regression results, avoid using symbols like x and y. Use abbreviations for the variables that are readily interpreted, defining the variables precisely in a separate section of the report. For the food expenditure example, we might have the variable definitions:

FOOD_EXP = weekly food expenditure by a household of size 3, in dollars

INCOME = weekly household income, in $100 units

Then the estimated equation results are as follows:

$$\begin{array}{lccc} FOOD\_EXP & = & 83.42 + 10.21\ INCOME & R^2 = 0.385 \\ (\text{se}) & & (43.41)^{*} \quad (2.09)^{***} & \end{array}$$

Report the standard errors below the estimated coefficients. The reason for showing the standard errors is that an approximate 95% interval estimate (if the degrees of freedom $N - 2$ are greater than 30) is $b_k \pm 2(\text{se})$. If desired, the reader may divide the estimate by the standard error to obtain the value of the t-statistic for testing a zero null hypothesis. Furthermore, testing other hypotheses is facilitated by having the standard error present. To test the null hypothesis $H_0: \beta_2 = 8.0$, we can quickly construct the t-statistic $t = [(10.21 - 8)/2.09]$ and proceed with the steps of the test procedure.

Asterisks are often used to show the reader the statistically significant (i.e., significantly different from zero using a two-tail test) coefficients, with explanations in a table footnote:

* indicates significant at the 10% level

** indicates significant at the 5% level

*** indicates significant at the 1% level

The asterisks are assigned by checking the p-values from the computer output, as shown in Figure 2.9.

4.3 Modeling Issues

4.3.1 The Effects of Scaling the Data

Data we obtain are not always in a convenient form for presentation in a table or use in a regression analysis. When the *scale* of the data is not convenient, it can be altered without changing any of the real underlying relationships between variables. For example, the real personal consumption in the United States, as of the second quarter of 2015, was \$12228.4 *billion* annually. That is, \$12,228,400,000,000 written out. While we *could* use the long form of the number in a table or in a regression analysis, there is no advantage to doing so. By choosing the units of measurement to be "billions of dollars," we have taken a long number and made it comprehensible. What are the effects of scaling the variables in a regression model?

Consider the food expenditure model. In Table 2.1 we report weekly expenditures in *dollars* but we report income in \$100 units, so a weekly income of \$2,000 is reported as $x = 20$. Why did we scale the data in this way? If we had estimated the regression using income in dollars, the results would have been

$$\begin{aligned} FOOD\_EXP &= 83.42 + 0.1021\ INCOME(\$) \quad R^2 = 0.385 \\ (\text{se}) \qquad &\quad (43.41)^{*}\ (0.0209)^{***} \end{aligned}$$

There are two changes. First, the estimated coefficient of income is now 0.1021. The interpretation is "If weekly household income increases by \$1 then we estimate that weekly food expenditure will increase by about 10 cents." There is nothing mathematically wrong with this, but it leads to a discussion of changes that are so small as to seem irrelevant. An increase in income of \$100 leads to an estimated increase in food expenditure of \$10.21, as before, but these magnitudes are more easily discussed.

The other change that occurs in the regression results when income is in dollars is that the standard error becomes smaller, by a factor of 100. Since the estimated coefficient is smaller by a factor of 100 also, this leaves the t-statistic and all other results unchanged.

Such a change in the units of measurement is called *scaling the data*. The choice of the scale is made by the researcher to make interpretation meaningful and convenient. The choice of the scale does not affect the measurement of the underlying relationship, but it does affect the interpretation of the coefficient estimates and some summary measures. Let us list the possibilities:

1. **Changing the scale of x**: In the linear regression model $y = \beta_1 + \beta_2 x + e$, suppose we change the units of measurement of the explanatory variable x by dividing it by a constant c. In order to keep intact the equality of the left- and right-hand sides, the coefficient of x must be multiplied by c. That is, $y = \beta_1 + \beta_2 x + e = \beta_1 + (c\beta_2)(x/c) + e = \beta_1 + \beta_2^* x^* + e$, where $\beta_2^* = c\beta_2$ and $x^* = x/c$. For example, if x is measured in dollars, and $c = 100$, then x^* is measured in hundreds of dollars. Then β_2^* measures the expected change in y given a \$100 increase in x, and β_2^* is 100 times larger than β_2. When the scale of x is altered, the only other change occurs in the standard error of the regression coefficient, but it changes by the same multiplicative factor as the coefficient, so that their ratio, the t-statistic, is unaffected. All other regression statistics are unchanged.

2. **Changing the scale of *y***: If we change the units of measurement of y, but not x, then all the coefficients must change in order for the equation to remain valid. That is, $y/c = (\beta_1/c) + (\beta_2/c)x + (e/c)$ or $y^* = \beta_1^* + \beta_2^* x + e^*$. In this rescaled model, β_2^* measures the change we expect in y^* given a 1-unit change in x. Because the error term is scaled in this process, the least squares residuals will also be scaled. This will affect the standard errors of the regression coefficients, but it will not affect t-statistics or R^2.
3. If the scale of y and the scale of x are changed by the same factor, then there will be no change in the reported regression results for b_2, but the estimated intercept and residuals will change; t-statistics and R^2 are unaffected. The interpretation of the parameters is made relative to the new units of measurement.

4.3.2 Choosing a Functional Form

In our ongoing example, we have assumed that the mean household food expenditure is a linear function of household income. That is, we assumed the underlying economic relationship to be $E(y|\mathbf{x}) = \beta_1 + \beta_2 x$, which implies that there is a linear, straight-line relationship between $E(y|\mathbf{x})$ and x. Why did we do that? Although the world is not "linear," a straight line is a good approximation to many nonlinear or curved relationships over narrow ranges. Moreover, in your principles of economics classes, you may have begun with straight lines for supply, demand, and consumption functions, and we wanted to ease you into the more "artistic" aspects of econometrics.

The starting point in all econometric analyses is economic theory. What does economics really say about the relation between food expenditure and income, holding all else constant? We expect there to be a positive relationship between these variables because food is a normal good. But nothing says the relationship must be a straight line. In fact, we do *not* expect that as household income rises, food expenditures will continue to rise indefinitely at the same constant rate. Instead, as income rises, we expect food expenditures to rise, but we expect such expenditures to increase at a decreasing rate. This is a phrase that is used many times in economics classes. What it means graphically is that there is not a straight-line relationship between the two variables. For a curvilinear relationship like that in Figure 4.4, the **marginal effect** of a change in the explanatory variable is measured by the slope of the tangent to the curve at a particular point. The marginal effect of a change in x is greater at the point (x_1, y_1) than it is at the point (x_2, y_2). As x increases, the value of y increases, but the slope is becoming smaller. This is the meaning of "increasing at a decreasing rate." In the economic context of the food expenditure model, the marginal propensity to spend on food is greater at lower incomes, and as income increases the marginal propensity to spend on food declines.

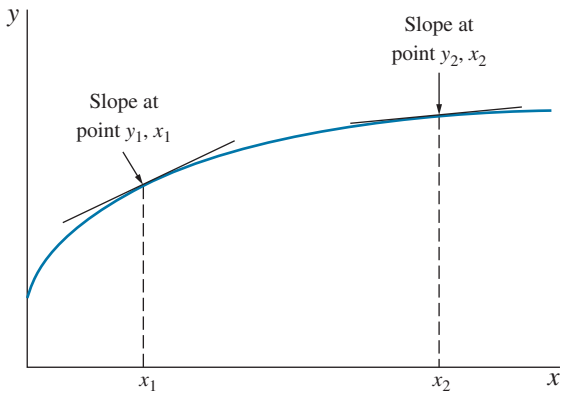

FIGURE 4.4 **A nonlinear relationship between food expenditure and income.**

The simple linear regression model is much more flexible than it appears at first glance. By *transforming* the variables y and x we can represent many curved, nonlinear relationships and still use the linear regression model. In Section 2.8, we introduced the idea of using **quadratic** and **log-linear** functional forms. In this and subsequent sections, we introduce you to an array of other possibilities and give some examples.

Choosing an algebraic form for the relationship means choosing *transformations* of the original variables. This is not an easy process, and it requires good analytic geometry skills and some experience. It may *not* come to you easily. The variable transformations that we begin with are as follows:

1. Power: If x is a variable, then x^p means raising the variable to the power p; examples are quadratic $\left(x^2\right)$ and cubic $\left(x^3\right)$ transformations.
2. The natural logarithm: If x is a variable, then its natural logarithm is $\ln(x)$.

Using just these two algebraic transformations, there are amazing varieties of "shapes" that we can represent, as shown in Figure 4.5.

A difficulty introduced when transforming variables is that regression result interpretations change. For each different functional form, shown in Table 4.1, the expressions for both the slope and elasticity change from the **linear relationship** case. This is so because the variables are related nonlinearly. What this means for the practicing economist is that great attention must be given to

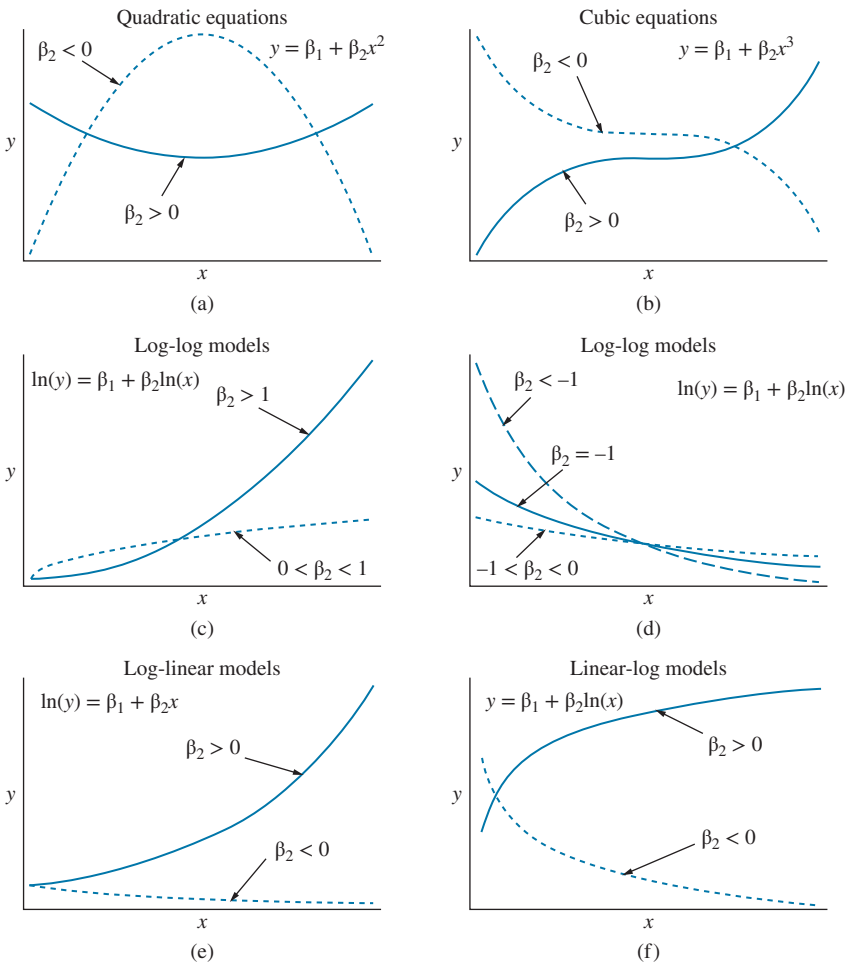

FIGURE 4.5 **Alternative functional forms.**

TABLE 4.1 Some Useful Functions, Their Derivatives, Elasticities, and Other Interpretation

Name	Function	Slope = dy/dx	Elasticity
Linear	$y = \beta_1 + \beta_2 x$	β_2	$\beta_2 \frac{x}{y}$
Quadratic	$y = \beta_1 + \beta_2 x^2$	$2\beta_2 x$	$(2\beta_2 x)\frac{x}{y}$
Cubic	$y = \beta_1 + \beta_2 x^3$	$3\beta_2 x^2$	$(3\beta_2 x^2)\frac{x}{y}$
Log-log	$\ln(y) = \beta_1 + \beta_2 \ln(x)$	$\beta_2 \frac{y}{x}$	β_2
Log-linear	$\ln(y) = \beta_1 + \beta_2 x$	$\beta_2 y$	$\beta_2 x$
	or, a 1-unit change in x leads to (approximately) a $100\beta_2$% change in y		
Linear-log	$y = \beta_1 + \beta_2 \ln(x)$	$\beta_2 \frac{1}{x}$	$\beta_2 \frac{1}{y}$
	or, a 1% change in x leads to (approximately) a $\beta_2/100$ unit change in y		

result interpretation whenever variables are transformed. Because you may be less familiar with logarithmic transformations, let us summarize the interpretation in three possible configurations.

1. In the **log-log model**, both the dependent and independent variables are transformed by the "natural" logarithm. The model is $\ln(y) = \beta_1 + \beta_2 \ln(x)$. In order to use this model, both y and x must be greater than zero because the logarithm is defined only for positive numbers. The parameter β_2 is the elasticity of y with respect to x. Referring to Figure 4.5, you can see why economists use the constant elasticity, log-log model specification so frequently. In panel (c), if $\beta_2 > 1$, the relation could depict a supply curve, or if $0 < \beta_2 < 1$, a production relation. In panel (d), if $\beta_2 < 0$, it could represent a demand curve. In each case, interpretation is convenient because the elasticity is constant. An example is given in Section 4.6.
2. In the **log-linear model** $\ln(y) = \beta_1 + \beta_2 x$, only the dependent variable is transformed by the logarithm. The dependent variable must be greater than zero to use this form. In this model, a 1-unit increase in x leads to (approximately) a $100\beta_2$% change in y. The log-linear form is common; it was introduced in Sections 2.8.3–2.8.4 and will be further discussed in Section 4.5. Note its possible shapes in Figure 4.5(e). If $\beta_2 > 0$, the function increases at an increasing rate; its slope is larger for larger values of y. If $\beta_2 < 0$, the function decreases, but at a decreasing rate.
3. In the **linear-log model** $y = \beta_1 + \beta_2 \ln(x)$ the variable x is transformed by the natural logarithm. See Figure 4.5(f). We can say that a 1% increase in x leads to a $\beta_2/100$-*unit* change in y. An example of this functional form is given in the following section.

Remark

Our plan for the remainder of this chapter is to consider several examples of the uses of alternative functional forms. In the following section we use the linear-log functional form with the food expenditure data. Then we take a brief detour into some diagnostic measures for data and model adequacy based on the least squares residuals. After discussing the diagnostic tools we give examples of polynomial equations, log-linear equations, and log-log equations.

4.3.3 A Linear-Log Food Expenditure Model

Suppose that in the food expenditure model, we wish to choose a functional form that is consistent with Figure 4.4. One option is the linear-log functional form. A linear-log equation has a

linear, untransformed term on the left-hand side and a logarithmic term on the right-hand side, or $y = \beta_1 + \beta_2 \ln(x)$. Because of the logarithm, this function requires $x > 0$. It is an increasing or decreasing function, depending on the sign of β_2. Using Derivative Rule 8, Appendix A, the slope of the function is β_2/x, so that as x increases, the slope decreases in absolute magnitude. If $\beta_2 > 0$, then the function increases at a decreasing rate. If $\beta_2 < 0$, then the function decreases at a decreasing rate. The function shapes are depicted in Figure 4.5(f). The elasticity of y with respect to x in this model is $\varepsilon = slope \times x/y = \beta_2/y$.

There is a convenient interpretation using approximations to changes in logarithms. Consider a small increase in x from x_0 to x_1. Then $y_0 = \beta_1 + \beta_2 \ln(x_0)$ and $y_1 = \beta_1 + \beta_2 \ln(x_1)$. Subtracting the former from the latter, and using the approximation developed in Appendix A, equation (A.3), gives

$$\begin{aligned} \Delta y = y_1 - y_0 &= \beta_2 \left[\ln(x_1) - \ln(x_0) \right] \\ &= \frac{\beta_2}{100} \times 100 \left[\ln(x_1) - \ln(x_0) \right] \\ &\cong \frac{\beta_2}{100} (\%\Delta x) \end{aligned}$$

The change in y, represented in its units of measure, is approximately $\beta_2/100$ times the percentage change in x.

EXAMPLE 4.4 | Using the Linear-Log Model for Food Expenditure

Using a linear-log equation for the food expenditure relation results in the regression model

$$FOOD\_EXP = \beta_1 + \beta_2 \ln(INCOME) + e$$

For $\beta_2 > 0$ this function is increasing but at a decreasing rate. As *INCOME* increases the slope $\beta_2/INCOME$ decreases. In this context, the slope is the marginal propensity to spend on food from additional income. Similarly, the elasticity, $\beta_2/FOOD\_EXP$, becomes smaller for larger levels of food expenditure. These results are consistent with the idea that at high incomes, and large food expenditures, the effect of an increase in income on food expenditure is small.

The estimated linear-log model using the food expenditure data is

$$\begin{aligned} \widehat{FOOD\_EXP} &= -97.19 + 132.17 \ln(INCOME) \quad R^2 = 0.357 \\ (se) \qquad & (84.24) \quad (28.80)^{***} \end{aligned} \tag{4.14}$$

The fitted model is shown in Figure 4.6.

As anticipated, the fitted function is not a straight line. The fitted linear-log model is consistent with our theoretical model that anticipates declining marginal propensity to spend additional income on food. For a household with \$1,000 weekly income, we estimate that the household will spend an additional \$13.22 on food from an additional

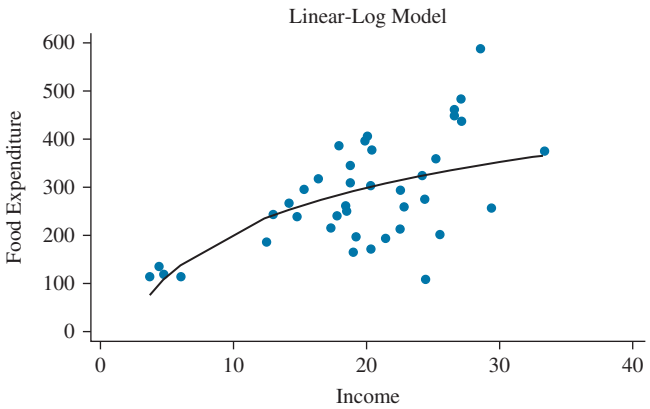

FIGURE 4.6 **The fitted linear-log model.**

$100 income, whereas we estimate that a household with $2,000 per week income will spend an additional $6.61 from an additional $100 income. The marginal effect of income on food expenditure is smaller at higher levels of income. This is a change from the linear, straight-line relationship we originally estimated, in which the marginal effect of a change in income of $100 was $10.21 for all levels of income.

Alternatively, we can say that a 1% increase in income will increase food expenditure by approximately $1.32 per week or that a 10% increase in income will increase food expenditure by approximately $13.22. Although this interpretation is conveniently simple to state, the diminishing marginal effect of income on food expenditure is somewhat disguised, though still implied. At $1,000 per week income, a 10% increase is $100, while at $2,000 income a 10% increase is $200. At higher levels of income, a larger dollar increase in income is required to elicit an additional $13.22 expenditure on food.

In terms of how well the model fits the data, we see that $R^2 = 0.357$ for the linear-log model, as compared to $R^2 = 0.385$ for the linear, straight-line relationship. Since these two models have the same dependent variable, *FOOD_EXP*, and each model has a single explanatory variable, a comparison of R^2 values is valid. However, there is a very small difference in the fit of the two models, and in any case, a model should not be chosen only on the basis of model fit with R^2 as the criterion.

Remark

Given alternative models that involve different transformations of the dependent and independent variables, and some of which have similar shapes, what are some guidelines for choosing a functional form?

1. Choose a shape that is consistent with what economic theory tells us about the relationship.
2. Choose a shape that is sufficiently flexible to "fit" the data.
3. Choose a shape so that assumptions SR1–SR6 are satisfied, ensuring that the least squares estimators have the desirable properties described in Chapters 2 and 3.

Although these objectives are easily stated, the reality of model building is much more difficult. You must recognize that we **never** know the "true" functional relationship between economic variables; also, the functional form that we select, no matter how elegant, is only an approximation. Our job is to choose a functional form that satisfactorily meets the three objectives stated above.

4.3.4 Using Diagnostic Residual Plots

When specifying a regression model, we may inadvertently choose an inadequate or incorrect functional form. Even if the functional form is adequate, one or more of the regression model assumptions may not hold. There are two primary methods for detecting such errors. First, examine the regression results. Finding an incorrect sign or a theoretically important variable that is not statistically significant may indicate a problem. Second, evidence of specification errors can reveal themselves in an analysis of the least squares residuals. We should ask whether there is any evidence that assumptions SR3 (homoskedasticity), SR4 (no serial correlation), and SR6 (normality) are violated. Usually, heteroskedasticity might be suspected in cross-sectional data analysis, and serial correlation is a potential time-series problem. In both cases, diagnostic tools focus on the least squares residuals. In Chapters 8 and 9, we will provide formal tests for homoskedasticity and serial correlation. In addition to formal tests, residual plots of all types are useful as diagnostic tools. In this section, residual analysis reveals potential heteroskedasticity and serial correlation problems and also flawed choices of functional forms.

We show a variety of residual plots in Figure 4.7. If there are no violations of the assumptions, then a plot of the least squares residuals versus x, y, or the fitted value of y, $\hat{y}$, should reveal no patterns. Figure 4.7(a) is an example of a random scatter.

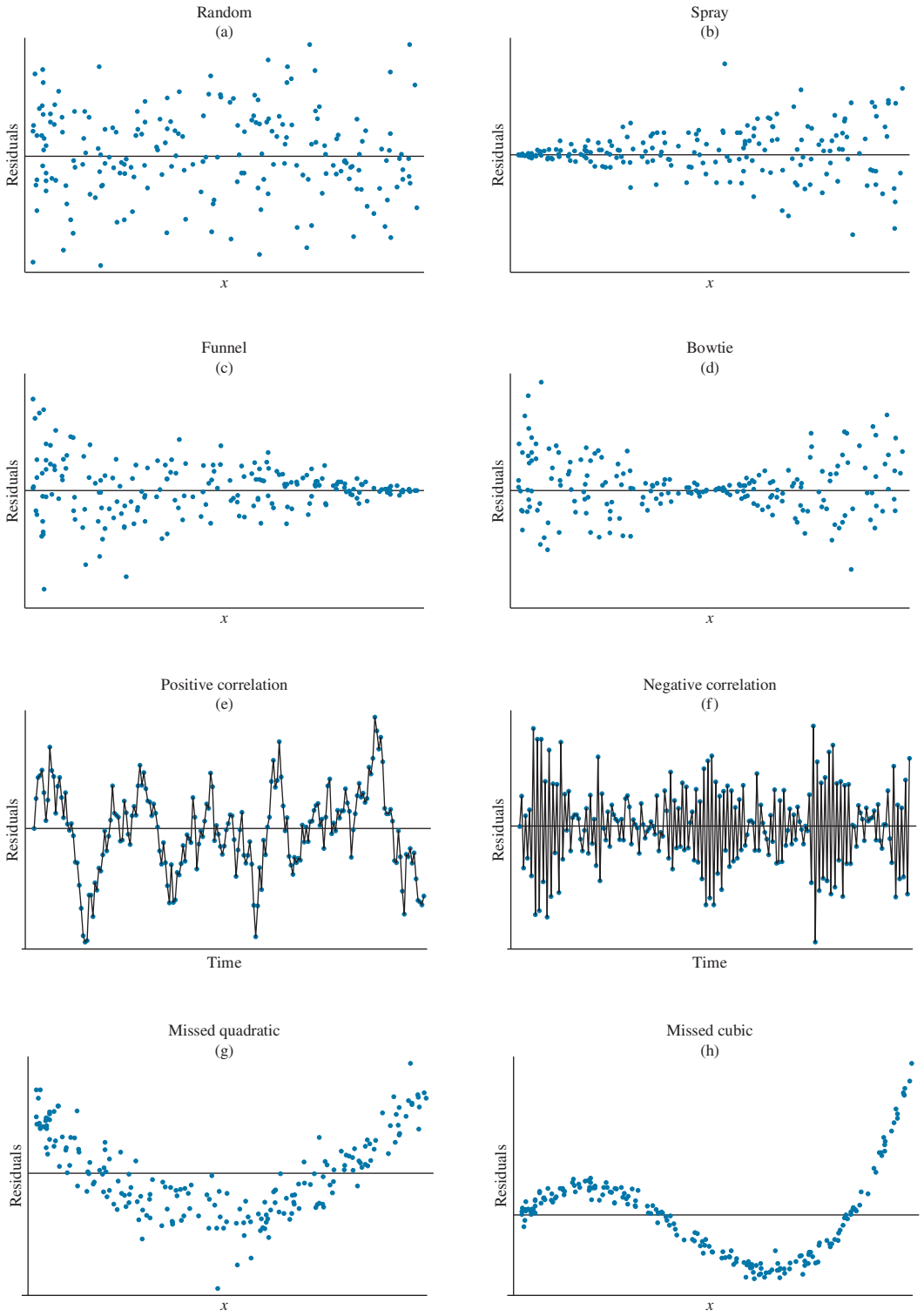

FIGURE 4.7 **Residual patterns.**

Figures 4.7(b)–(d) show patterns associated with heteroskedasticity. Figure 4.7(b) has a "spray-shaped" residual pattern that is consistent with the variance of the error term increasing as x-values increase; Figure 4.7(c) has a "funnel-shaped" residual pattern that is consistent with the variance of the error term decreasing as x-values increase; and Figure 4.7(d) has a "bow-tie" residual pattern that is consistent with the variance of the error term decreasing and then increasing as x-values increase.

Figure 4.7(e) shows a typical pattern produced with time-series regression when the error terms display a positive correlation, $\text{corr}(e_t, e_{t-1}) > 0$. Note that there are sequences of positive residuals followed by sequences of negative residuals, and so on. If assumption SR4 holds there should be no such sign patterns. Figure 4.7(f) shows a typical pattern produced with time-series regression when the error terms display a negative correlation, $\text{corr}(e_t, e_{t-1}) < 0$. In this case, each positive residual tends to be followed by a negative residual, which is then followed by a positive residual and so on. The sequence of residuals tends to alternate in sign.

If the relationship between y and x is curvilinear, such as a U-shaped quadratic function, like an average cost function, and we mistakenly assume that the relationship is linear, then the least squares residuals may show a U-shape like in Figure 4.7(g). If the relationship between y and x is curvilinear, such as a cubic function, like a total cost function, and we mistakenly assume that the relationship is linear, then the least squares residuals may show a serpentine shape like Figure 4.7(h).

The bottom line is that when least squares residuals are plotted against another variable there should be no patterns evident. Patterns of the sorts shown in Figure 4.7, except for panel (a), indicate that there may be some violation of assumptions and/or incorrect model specification.

EXAMPLE 4.5 | Heteroskedasticity in the Food Expenditure Model

The least squares residuals from the linear-log food expenditure model in (4.14) are plotted in Figure 4.8. These exhibit an expanding variation pattern with more variation in the residuals as *INCOME* becomes larger, which may suggest heteroskedastic errors. A similar residual plot is implied by Figure 2.8.

We must conclude that at this point we do not have a satisfactory model for the food expenditure data. The linear and linear-log models have different shapes and different implied marginal effects. The two models fit the data equally well, but both models exhibit least squares residual patterns consistent with heteroskedastic errors. This example will be considered further in Chapter 8.

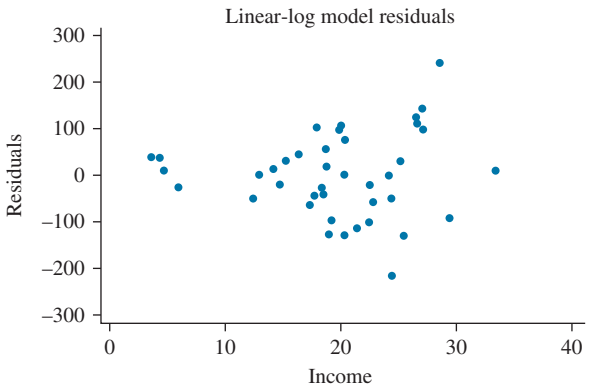

FIGURE 4.8 **Residuals from linear-log food expenditure model.**

4.3.5 Are the Regression Errors Normally Distributed?

Recall that hypothesis tests and interval estimates for the coefficients rely on SR6 assumption, that given $\mathbf{x}$, the errors, and hence the dependent variable y, are normally distributed. Though our tests and confidence intervals are valid in large samples whether the data are normally distributed or not, it is nevertheless desirable to have a model in which the regression errors are normally distributed, so that we do not have to rely on large sample approximations. If the errors are not normally distributed, we might be able to improve our model by considering an alternative functional form or transforming the dependent variable. As noted in the last "Remark," when choosing

a functional form, one of the criteria we might examine is whether a model specification satisfies regression assumptions, and in particular, whether it leads to errors that are normally distributed (SR6). How do we check out the assumption of normally distributed errors?

We cannot observe the true random errors, so we must base our analysis of their normality on the least squares residuals, $\hat{e}_i = y_i - \hat{y}_i$. Substituting for y_i and $\hat{y}_i$, we obtain

$$\begin{aligned}\hat{e}_i &= y_i - \hat{y}_i = \beta_1 + \beta_2 x_i + e_i - (b_1 + b_2 x_i) \\ &= (\beta_1 - b_1) + (\beta_2 - b_2)x_i + e_i \\ &= e_i - (b_1 - \beta_1) - (b_2 - \beta_2)x_i\end{aligned}$$

In large samples, $(b_1 - \beta_1)$ and $(b_2 - \beta_2)$ will tend toward zero because the least squares estimators are unbiased and have variances that approach zero as $N \to \infty$. Consequently, in large samples, the difference $\hat{e}_i - e_i$ is close to zero, so that these two random variables are essentially the same and thus have the same distribution.

A histogram of the least squares residuals gives us a graphical representation of the empirical distribution.

EXAMPLE 4.6 | Testing Normality in the Food Expenditure Model

The relevant EViews output for the food expenditure example, using the linear relationship with no transformation of the variables, appears in Figure 4.9. What does this histogram tell us? First, notice that it is centered at zero. This is not surprising because the mean of the least squares residuals is always zero if the model contains an intercept, as shown in Appendix 4B. Second, it seems symmetrical, but there are some large gaps, and it does not really appear bell shaped. However, merely checking the shape of the histogram, especially when the number of observations is relatively small, is not a statistical "test."

There are many tests for normality. The **Jarque–Bera test** for normality is valid in large samples. It is based on two measures, **skewness** and **kurtosis**. In the present context, **skewness** refers to how symmetric the residuals are around zero. Perfectly symmetric residuals will have a skewness of zero. The skewness value for the food expenditure residuals is −0.097. **Kurtosis** refers to the "peakedness" of the distribution. For a normal distribution, the kurtosis value is 3. For more on skewness and kurtosis, see Appendices B.1.2 and C.4.2. From Figure 4.9, we see that the food expenditure residuals have a kurtosis of 2.99. The skewness and kurtosis values are close to the values for the normal distribution. So, the question we have to ask is whether 2.99 is sufficiently different from 3, and −0.097 is sufficiently different from zero, to conclude that the residuals are not normally distributed. The Jarque–Bera statistic is given by

$$JB = \frac{N}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

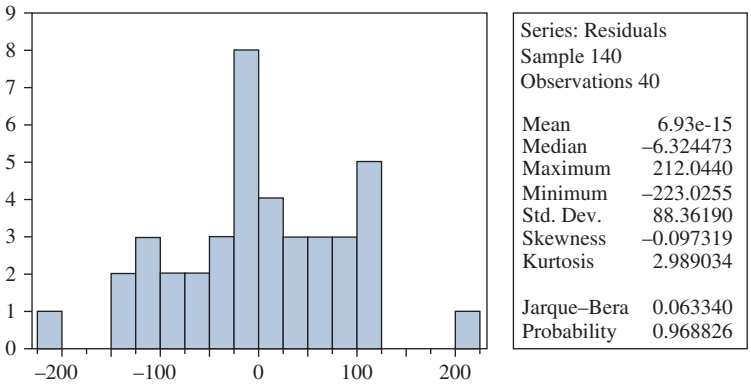

FIGURE 4.9 **EViews output: residuals histogram and summary statistics for food expenditure example.**

where N is the sample size, S is skewness, and K is kurtosis. Thus, large values of the skewness, and/or values of kurtosis quite different from 3, will lead to a large value of the Jarque–Bera statistic. When the residuals are normally distributed, the Jarque–Bera statistic has a chi-squared distribution with two degrees of freedom. We reject the hypothesis of normally distributed errors if a calculated value of the statistic exceeds a critical value selected from the chi-squared distribution with two degrees of freedom. Using Statistical Table 3, the 5% critical value from a χ^2-distribution with two degrees of freedom is 5.99, and the 1% critical value is 9.21.

Applying these ideas to the food expenditure example, we have

$$JB = \frac{40}{6}\left((-0.097)^2 + \frac{(2.99-3)^2}{4}\right) = 0.063$$

Because $0.063 < 5.99$, there is insufficient evidence from the residuals to conclude that the normal distribution assumption is unreasonable at the 5% level of significance. The same conclusion could have been reached by examining the p-value. The p-value appears in Figure 4.9 described as "Probability." Thus, we also fail to reject the null hypothesis on the grounds that $0.9688 > 0.05$.

For the linear-log model of food expenditure reported in Example 4.4, the Jarque–Bera test statistic value is 0.1999 with a p-value of 0.9049. We cannot reject the null hypothesis that the regression errors are normally distributed, and this criterion does not help us choose between the linear and linear-log functional forms for the food expenditure model.

In these examples, we should remember that the Jarque–Bera test is strictly valid only in large samples. Applying tests that are valid in large samples to smaller samples, such as $N = 40$, is not uncommon in applied work. However, we should remember in such applications that we should not give great weight to the test significance or nonsignificance.

4.3.6 Identifying Influential Observations

One worry in data analysis is that we may have some unusual and/or **influential observations**. Sometimes, these are termed "outliers." If an unusual observation is the result of a data error, then we should correct it. If an unusual observation is not the result of a data error, then understanding how it came about, the story behind it, can be informative. One way to detect whether an observation is influential is to delete it and reestimate the model, comparing the results to the original results based on the full sample. This "delete-one" strategy can help detect the influence of the observation on the estimated coefficients and the model's predictions. It can also help us identify unusual observations.

The delete-one strategy begins with the least squares parameter estimates based on the sample with the ith observation deleted. Denote these as $b_1(i)$ and $b_2(i)$. Let $\hat{\sigma}^2(i)$ be the delete-one estimated error variance. The residual $\hat{e}(i) = y_i - \left[b_1(i) + b_2(i)\,x_i\right]$ is the actual value of y for the ith observation, y_i, minus the fitted value that uses estimates from the sample with the ith observation deleted. It is the forecast error (4.3) with y_i taking the place of y_0 and x_i taking the value of x_0 and using the estimates $b_1(i)$ and $b_2(i)$. Modifying the variance of the forecast error (4.4), we obtain the variance of $\hat{e}(i)$ (and its estimator) as

$$\widehat{\text{var}}\left[\hat{e}(i)|\mathbf{x}\right] = \hat{\sigma}^2(i)\left[1 + \frac{1}{(N-1)} + \frac{\left(x_i - \bar{x}(i)\right)^2}{\sum_{j\neq i}\left(x_j - \bar{x}(i)\right)^2}\right]$$

where $\bar{x}(i)$ is the delete-one sample mean of the x-values. The ratio

$$\hat{e}_i^{\text{stu}} = \frac{\hat{e}(i)}{\left\{\widehat{\text{var}}\left[\hat{e}(i)|\mathbf{x}\right]\right\}^{1/2}}$$

is called a **studentized residual**. It is the standardized residual based on the delete-one sample. The rule of thumb is to calculate these values and compare their values to ± 2, which is roughly

a 95% interval estimate. If the studentized residual falls outside the interval, then the observation is worth examining because it is "unusually" large.

After considerable algebra, the studentized residual can also be written as

$$\hat{e}_i^{\text{stu}} = \frac{\hat{e}_i}{\hat{\sigma}(i)\left(1 - h_i\right)^{1/2}}$$

where

$$h_i = \frac{1}{N} + \frac{\left(x_i - \bar{x}\right)^2}{\sum\left(x_i - \bar{x}\right)^2}$$

The term h_i is called the **leverage** of the ith observation, with values $0 \le h_i \le 1$. If the leverage value is high, then the value of the studentized residual is inflated. The second component of h_i is $\left(x_i - \bar{x}\right)^2 / \sum\left(x_i - \bar{x}\right)^2$. Recall that the sample variance of the x_i-values is estimated by $s_x^2 = \sum\left(x_i - \bar{x}\right)^2 / (N-1)$ so that $\sum\left(x_i - \bar{x}\right)^2$ is a measure of the total variation in the sample x_i-values about their mean. If one observation's contribution $\left(x_i - \bar{x}\right)^2$ to the total is large, then that observation may have a strong effect on the least squares estimates and fitted values. The sum of the leverage terms h_i is K, the number of parameters in the regression model. Thus, the average value in the simple regression model is $\bar{h} = K/N = 2/N$. When checking data, it is a common rule of thumb to examine observations with leverage greater than two or three times the average.

Another measure of the influence of a single observation on the least squares estimates is called **DFBETAS**. For the slope estimate in the simple regression model, we calculate

$$\text{DFBETAS}_{2i} = \frac{b_2 - b_2(i)}{\hat{\sigma}(i) \Big/ \sqrt{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}}$$

The effect of the ith observation on the slope estimate is measured by the change in the coefficient estimate by dropping the ith observation and then standardizing. The magnitude of DFBETAS_{2i} will be larger when leverage is larger and/or the studentized residual is larger. A common rule of thumb for identifying influential observations in the simple regression model is $\left|\text{DFBETAS}_{2i}\right| > 2/\sqrt{N}$.

The effect of the ith observation on the fitted value from the least squares regression is again a measurement using the delete-one approach. Let $\hat{y}_i = b_1 + b_2 x_i$ and $\hat{y}(i) = b_1(i) + b_2(i)\, x_i$ with $\hat{y}_{(i)}$ being the fitted value using parameter estimates from the delete-one sample. The measure called **DFFITS** is

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}(i)}{\hat{\sigma}(i)\, h_i^{1/2}} = \left(\frac{h_i}{1 - h_i}\right)^{1/2} \hat{e}_i^{\text{stu}}$$

This measure will be larger when leverage is larger and/or the studentized residual is larger. A rule of thumb to identify unusual observations is $\left|\text{DFFITS}_i\right| > 2(K/N)^{1/2}$ or $\left|\text{DFFITS}_i\right| > 3(K/N)^{1/2}$ where $K = 2$ is the number of parameters in the simple regression model.

These constructs may look difficult to compute, but modern software usually computes some or all of these measures. We are **not** suggesting that you toss out unusual observations. If these measures lead you to locate an observation with an error, you can try to fix it. By looking at unusual observations, ones that have a high leverage, a large studentized residual, a large DFBETAS, or a large DFFITS, you may learn something about which data characteristics are important. All data analysts should examine their data, and these tools may help organize such an examination.

EXAMPLE 4.7 | Influential Observations in the Food Expenditure Data

Examining the influential observation measures for the food expenditure data, using a linear relationship and no transformations of the variables, reveals few real surprises. First the leverage values have the average $\bar{h} = 2/40 = 0.05$. Isolating observations with leverage more than twice the average, we have

obs	*h*	*FOOD_EXP*	*INCOME*
1	0.1635	115.22	3.69
2	0.1516	135.98	4.39
3	0.1457	119.34	4.75
4	0.1258	114.96	6.03
40	0.1291	375.73	33.4

The observations with the greatest leverage are those with the four lowest incomes and the highest income. The mean of *INCOME* is 19.6.

The observations with studentized residuals, *EHATSTU*, larger than two in absolute value are

obs	*EHATSTU*	*FOOD_EXP*	*INCOME*
31	−2.7504	109.71	24.42
38	2.6417	587.66	28.62

These two observations are interesting because the food expenditures for these two households are the minimum and maximum, despite both incomes being above the mean. In fact, the income for household 31 is the 75th percentile value, and the income for household 38 is the third largest. Thus, household 31 is spending significantly less on food than we would predict, and household 38 more than we would predict, based on income alone. These might be observations worth checking to ensure they are correct. In our case, they are.

The DFBETAS values greater than $2/\sqrt{N} = 0.3162$ in absolute value are

obs	*DFBETAS*	*FOOD_EXP*	*INCOME*
38	0.5773	587.66	28.62
39	−0.3539	257.95	29.40

Again household 38 has a relatively large influence on the least squares estimate of the slope. Household 39 shows up because it has the second highest income but spends less than the mean value (264.48) on food.

Finally, DFFITS values larger than $2(K/N)^{1/2} = 0.4472$ are as follows:

obs	*DFFITS*	*FOOD_HAT*	*FOOD_EXP*	*INCOME*
31	−0.5442	332.74	109.71	24.42
38	0.7216	375.62	587.66	28.62

The observations with a high influence of the least squares fitted values are the previously mentioned households 31 and 38, which also have large studentized residuals.

4.4 Polynomial Models

In Sections 2.8.1–2.8.2, we introduced the use of quadratic polynomials to capture curvilinear relationships. Economics students will have seen many average and marginal cost curves (U-shaped) and average and marginal product curves (inverted-U shaped) in their studies. Higher order polynomials, such as cubic equations, are used for total cost and total product curves. A familiar example to economics students is the total cost curve, shaped much like the solid curve in Figure 4.5(b). In this section, we review simplified quadratic and cubic equations and give an empirical example.

4.4.1 Quadratic and Cubic Equations

The general form of a quadratic equation $y = a_0 + a_1x + a_2x^2$ includes a constant term a_0, a linear term a_1x, and a squared term a_2x^2. Similarly, the general form of a cubic equation is $y = a_0 + a_1x + a_2x^2 + a_3x^3$. In Section 5.6, we consider multiple regression models using the

general forms of quadratic and cubic equations. For now, however, because we are working with "simple" regression models that include only one explanatory variable, we consider the simple quadratic and cubic forms, $y = \beta_1 + \beta_2 x^2$ and $y = \beta_1 + \beta_2 x^3$, respectively. The properties of the simple quadratic function are discussed in Section 2.8.1.

The simple cubic equation $y = \beta_1 + \beta_2 x^3$ has possible shapes shown in Figure 4.5(b). Using Derivative Rules 4 and 5 from Appendix A, the derivative, or slope, of this cubic equation is $dy/dx = 3\beta_2 x^2$. The slope of the curve is always positive if $\beta_2 > 0$, except when $x = 0$, yielding a direct relationship between y and x like the solid curve shown in Figure 4.5(b). If $\beta_2 < 0$, then the relationship is an inverse one like the dashed curve shown in Figure 4.5(b). The slope equation shows that the slope is zero only when $x = 0$. The term β_1 is the y-intercept. The elasticity of y with respect to x is $\varepsilon = slope \times x/y = 3\beta_2 x^2 \times x/y$. Both the slope and elasticity change along the curve.

EXAMPLE 4.8 | An Empirical Example of a Cubic Equation

Figure 4.10 is a plot of average wheat yield (in tonnes per hectare—a hectare is about 2.5 acres, and a tonne is a metric ton that is 1000 kg or 2205 lb—we are speaking Australian here!) for the Greenough Shire in Western Australia, against time. The observations are for the period 1950–1997, and time is measured using the values 1, 2, …, 48. These data can be found in the data file *wa_wheat*. Notice in Figure 4.10 that wheat yield fluctuates quite a bit, but overall, it tends to increase over time, and the increase is at an increasing rate, particularly toward the end of the time period. An increase in yield is expected because of technological improvements, such as the development of varieties of wheat that are higher yielding and more resistant to pests and diseases. Suppose that we are interested in measuring the effect of technological improvement on yield. Direct data on changes in technology are not available, but we can examine how wheat yield has changed over time as a consequence of changing technology. The equation of interest relates *YIELD* to *TIME*, where $TIME = 1, \ldots, 48$. One problem with the linear equation

$$YIELD_t = \beta_1 + \beta_2 TIME_t + e_t$$

is that it implies that yield increases at the same constant rate β_2, when, from Figure 4.10, we expect this rate to be increasing. The least squares fitted line (standard errors in parentheses) is

$$\widehat{YIELD}_t = 0.638 + 0.0210\, TIME_t \qquad R^2 = 0.649$$
$$\text{(se)} \qquad (0.064)\ (0.0022)$$

The residuals from this regression are plotted against time in Figure 4.11. Notice that there is a concentration of positive residuals at each end of the sample and a concentration of negative residuals in the middle. These concentrations are caused by the inability of a straight line to capture the fact that yield is increasing at an increasing rate. Compare the residual pattern in Figure 4.11 to Figures 4.7(g) and (h). What alternative can we try? Two possibilities are $TIME^2$ and $TIME^3$. It turns out that $TIME^3$ provides the better fit, and so we consider the functional form

$$YIELD_t = \beta_1 + \beta_2 TIME_t^3 + e_t$$

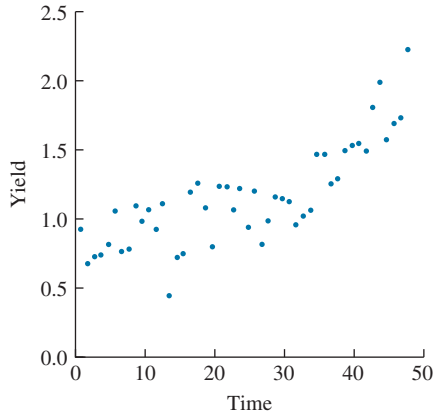

FIGURE 4.10 Scatter plot of wheat yield over time.

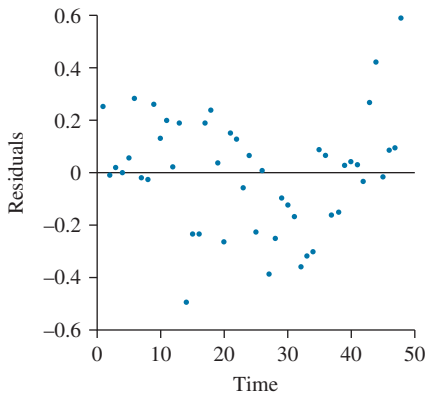

FIGURE 4.11 Residuals from a linear yield equation.

The slope of the expected yield function is $3\beta_2 TIME^2$. Thus, so long as the estimate of β_2 turns out to be positive, the function will be increasing. Furthermore, the slope is increasing as well. Thus, the function itself is "increasing at an increasing rate." Before estimating the cubic equation, note that the values of $TIME^3$ can get very large. This variable is a good candidate for scaling. If we define $TIMECUBE_t = (TIME_t/100)^3$, the estimated equation is

$$\widehat{YIELD}_t = 0.874 + 9.682\,TIMECUBE_t \qquad R^2 = 0.751$$
$$\text{(se)} \qquad (0.036)\ (0.822)$$

The residuals from this cubic equation are plotted in Figure 4.12. The predominance of positive residuals at the ends and negative residuals in the middle no longer exists. Furthermore, the R^2 value has increased from 0.649 to 0.751, indicating that the equation with *TIMECUBE* fits the data better than the one with just *TIME*. Both these equations have the same dependent variable and the same number of explanatory variables (only 1). In these circumstances, the R^2 can be used legitimately to compare goodness-of-fit.

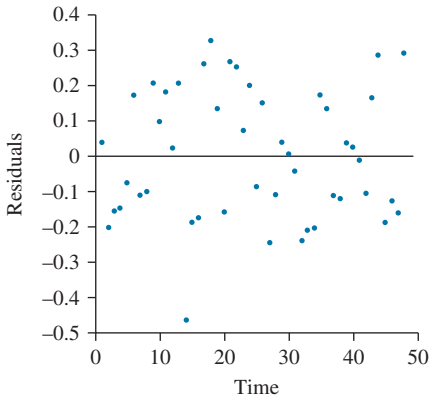

FIGURE 4.12 **Residuals from a cubic yield equation.**

What lessons have we learned from this example? First, a plot of the original dependent variable series y against the explanatory variable x is a useful starting point for deciding on a functional form in a simple regression model. Secondly, examining a plot of the residuals is a useful device for uncovering inadequacies in any chosen functional form. Runs of positive and/or negative residuals can suggest an alternative. In this example, with time-series data, plotting the residuals against time was informative. With cross-sectional data, using plots of residuals against both independent and dependent variables is recommended. Ideally, we will see no patterns, and the residual histogram and Jarque–Bera test will not rule out the assumption of normality. As we travel through the book, you will discover that patterns in the residuals, such as those shown in Figure 4.7, can also mean many other specification inadequacies, such as omitted variables, heteroskedasticity, and autocorrelation. Thus, as you become more knowledgeable and experienced, you should be careful to consider other options. For example, wheat yield in Western Australia is heavily influenced by rainfall. Inclusion of a rainfall variable might be an option worth considering. Also, it makes sense to include *TIME* and $TIME^2$ in addition to $TIME^3$. A further possibility is the constant growth rate model that we consider in the following section.

4.5 Log-Linear Models

Econometric models that employ natural logarithms are very common. We first introduced the log-linear model in Section 2.8.3. Logarithmic transformations are often used for variables that are monetary values, such as wages, salaries, income, prices, sales, and expenditures, and, in general, for variables that measure the "size" of something. These variables have the characteristic that they are positive and often have distributions that are positively skewed, with a long tail to the right. Figure P.2 in the Probability Primer is representative of the income distribution in the United States. In fact, the probability density function $f(x)$ shown is called the "log-normal" because $\ln(x)$ has a normal distribution. Because the transformation $\ln(x)$ has the effect of making larger values of x less extreme, $\ln(x)$ will often be closer to a normal distribution for variables of this kind. The **log-normal distribution** is discussed in Appendix B.3.9.

The log-linear model, $\ln(y) = \beta_1 + \beta_2 x$, has a logarithmic term on the left-hand side of the equation and an untransformed (linear) variable on the right-hand side. Both its slope and elasticity change at each point and are the same sign as β_2. Using the antilogarithm, we obtain $\exp[\ln(y)] = y = \exp(\beta_1 + \beta_2 x)$, so that the log-linear function is an exponential function. The function requires $y > 0$. The slope at any point is $\beta_2 y$, which for $\beta_2 > 0$ means that the marginal effect increases for larger values of y. An economist might say that this function is increasing at an increasing rate. The shapes of the log-linear model are shown in Figure 4.5(e), and its derivative and elasticity given in Table 4.1. To make discussion relevant in a specific context, the slope can be evaluated at the sample mean $\bar{y}$, or the elasticity $\beta_2 x$ can be evaluated at the sample mean $\bar{x}$, or other interesting values can be chosen.

Using the properties of logarithms, we can obtain a useful approximation. Consider an increase in x from x_0 to x_1. The change in the log-linear model is from $\ln(y_0) = \beta_1 + \beta_2 x_0$ to $\ln(y_1) = \beta_1 + \beta_2 x_1$. Subtracting the first equation from the second gives $\ln(y_1) - \ln(y_0) = \beta_2(x_1 - x_0) = \beta_2 \Delta x$. Multiply by 100, and use the approximation introduced in Appendix A, equation (A.3) to obtain

$$100\left[\ln(y_1) - \ln(y_0)\right] \cong \%\Delta y = 100\beta_2(x_1 - x_0) = (100\beta_2) \times \Delta x$$

A 1-unit increase in x leads approximately to a $100\beta_2\%$ change in y.

In the following two examples, we apply the familiar concept of **compound interest** to derive a log-linear economic model for growth arising from technology, and a model explaining the relation between an individual's wage rate and their years of schooling. Recall the compound interest formula. If an investor deposits an initial amount V_0 (the principal amount) into an account that earns a rate of return r, then after t periods the value V of the account is $V_t = V_0(1 + r)^t$. For example, if $r = 0.10$, so that the rate of return is 10%, and if $V_0 = \$100$, after one period the account value is $V_1 = \$110$; after two periods, the account value is $V_2 = \$121$, and so on. The compound interest formula also explains the account growth from year to year. The accumulated value earns the rate r in each period so that $V_t = V_0(1 + r)^t = (1 + r)V_{t-1}$.

EXAMPLE 4.9 | A Growth Model

Earlier in this chapter, in Example 4.8, we considered an empirical example in which the production of wheat was tracked over time, with improvements in technology leading to wheat production increasing at an increasing rate. We observe wheat production in time periods $t = 1, \ldots, T$. Assume that in each period *YIELD* grows at the constant rate g due to technological progress. Let the *YIELD* at time $t = 0$, before the sample begins, be $YIELD_0$. This plays the role of the initial amount. Applying the compound interest formula we have $YIELD_t = YIELD_0(1 + g)^t$. Taking logarithms, we obtain

$$\begin{aligned}\ln(YIELD_t) &= \ln(YIELD_0) + [\ln(1 + g)] \times t \\ &= \beta_1 + \beta_2 t\end{aligned}$$

This is simply a log-linear model with dependent variable $\ln(YIELD_t)$ and explanatory variable t, or time. We expect growth to be positive, so that $\beta_2 > 0$, in which case the plot of *YIELD* against time looks like the upward-sloping curve in Figure 4.5(c), which closely resembles the scatter diagram in Figure 4.11.

Estimating the log-linear model for yield, we obtain

$$\begin{aligned}\widehat{\ln(YIELD_t)} &= -0.3434 + 0.0178t \\ (\text{se}) \qquad & (0.0584) \quad (0.0021)\end{aligned}$$

The estimated coefficient $b_2 = \widehat{\ln(1 + g)} = 0.0178$. Using the property that $\ln(1 + x) \cong x$ if x is small [see Appendix A, equation (A.4) and the discussion following it], we estimate that the growth rate in wheat yield is approximately $\hat{g} = 0.0178$, or about 1.78% per year, over the period of the data.

EXAMPLE 4.10 | A Wage Equation

The relationship between wages and education is a key relationship in labor economics (and, no doubt, in your mind). Suppose that the rate of return to an extra year of education is a constant r. Let $WAGE_0$ represent the wage of a person with no education. Applying the compound interest formula to the investment in human capital, we anticipate that the wage of a person with one year of education will be $WAGE_1 = WAGE_0(1 + r)$. A second year of education will compound the human capital so that $WAGE_2 = WAGE_1(1 + r) = WAGE_0(1 + r)^2$. In general, $WAGE = WAGE_0(1 + r)^{EDUC}$, where $EDUC$ is years of education. Taking logarithms, we have a relationship between $\ln(WAGE)$ and years of education ($EDUC$)

$$\begin{aligned}\ln(WAGE) &= \ln\left(WAGE_0\right) + \left[\ln(1 + r)\right] \times EDUC \\ &= \beta_1 + \beta_2 EDUC\end{aligned}$$

An additional year of education leads to an approximate $100\beta_2\%$ increase in wages.

Data on hourly wages, years of education, and other variables are in the file *cps5_small*. These data consist of 1200 observations from the May 2013 Current Population Survey (CPS). The CPS is a monthly survey of about 50000 households conducted in the United States by the Bureau of the Census for the Bureau of Labor Statistics. The survey has been conducted for more than 50 years. Using these data, the estimated log-linear model is

$$\begin{aligned}\widehat{\ln(WAGE)} &= 1.5968 + 0.0988 \times EDUC \\ \text{(se)} \qquad & (0.0702) \quad (0.0048)\end{aligned}$$

We estimate that an additional year of education increases the wage rate by approximately 9.9%. A 95% interval estimate for the value of an additional year of education is 8.9% to 10.89%.

4.5.1 Prediction in the Log-Linear Model

You may have noticed that when reporting regression results in this section, we did not include an R^2 value. In a log-linear regression, the R^2 value automatically reported by statistical software is the percentage of the variation in $\ln(y)$ explained by the model. However, our objective is to explain the variations in y, not $\ln(y)$. Furthermore, the fitted regression line predicts $\widehat{\ln(y)} = b_1 + b_2 x$, whereas we want to predict y. The problems of obtaining a useful measure of goodness-of-fit and prediction are connected, as we discussed in Section 4.2.2.

How shall we obtain the predicted value of y? A first inclination might be to take the antilog of $\widehat{\ln(y)} = b_1 + b_2 x$. The exponential function is the antilogarithm for the natural logarithm, so that a natural choice for prediction is

$$\hat{y}_n = \exp\left(\widehat{\ln(y)}\right) = \exp\left(b_1 + b_2 x\right)$$

In the log-linear model, this is not necessarily the best we can do. Using properties of the log-normal distribution it can be shown (see Appendix B.3.9) that an alternative predictor is

$$\hat{y}_c = \widehat{E(y)} = \exp\left(b_1 + b_2 x + \hat{\sigma}^2/2\right) = \hat{y}_n e^{\hat{\sigma}^2/2}$$

If the sample size is large, the "corrected" predictor $\hat{y}_c$ is, on average, closer to the actual value of y and should be used. In small samples (less than 30), the "natural" predictor may actually be a better choice. The reason for this incongruous result is that the estimated value of the error variance $\hat{\sigma}^2$ adds a certain amount of "noise" when using $\hat{y}_c$, leading it to have increased variability relative to $\hat{y}_n$ that can outweigh the benefit of the correction in small samples.

EXAMPLE 4.11 | Prediction in a Log-Linear Model

The effect of the correction can be illustrated using the wage equation. What would we predict the wage to be for a worker with 12 years of education? The predicted value of ln(*WAGE*) is

$$\widehat{\ln(WAGE)} = 1.5968 + 0.0988 \times EDUC$$
$$= 1.5968 + 0.0988 \times 12 = 2.7819$$

Then the value of the natural predictor is $\hat{y}_n = \exp\left(\widehat{\ln(y)}\right) = \exp(2.7819) = 16.1493$. The value of the corrected predictor, using $\hat{\sigma}^2 = 0.2349$ from the regression output, is

$$\hat{y}_c = \widehat{E(y)} = \hat{y}_n e^{\hat{\sigma}^2/2} = 16.1493 \times 1.1246 = 18.1622$$

We predict that the wage for a worker with 12 years of education will be \$16.15 per hour if we use the natural predictor and \$18.16 if we use the corrected predictor. In this case, the sample is large ($N = 1200$), so we would use the corrected predictor. Among the 1200 workers, there are 307 with 12 years of education. Their average wage is \$17.31, so the corrected predictor is consistent with the sample of data.

How does the correction affect our prediction? Recall that $\hat{\sigma}^2$ must be greater than zero and $e^0 = 1$. Thus, the effect of the correction is always to increase the value of the prediction because $e^{\hat{\sigma}^2/2}$ is always greater than one. The natural predictor tends to systematically underpredict the value of y in a log-linear model, and the correction offsets the downward bias in large samples. The "natural" and "corrected" predictions are shown in Figure 4.13.

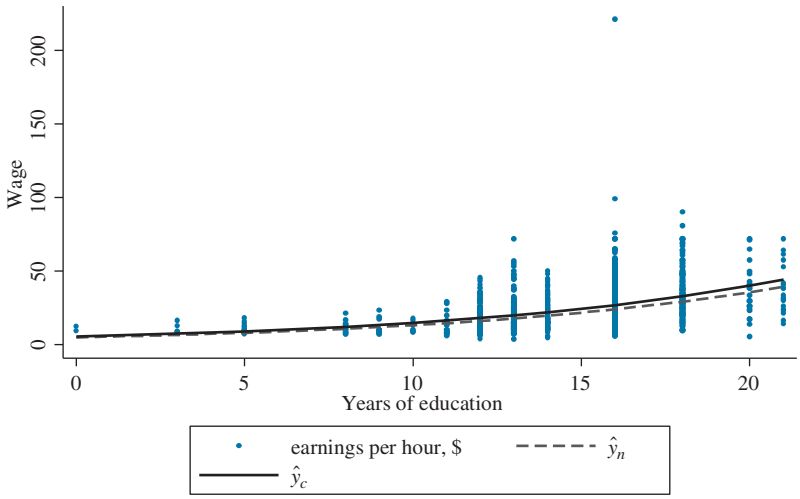

FIGURE 4.13 **The natural and corrected predictors of wage.**

4.5.2 A Generalized R^2 Measure

It is a general rule that the squared simple correlation between y and its fitted value $\hat{y}$, where $\hat{y}$ is the "best" prediction one can obtain, is a valid measure of goodness-of-fit that we can use as an R^2 in many contexts. As we have seen, what we may consider the "best" predictor can change depending on the model under consideration. That is, a general goodness-of-fit measure, or general R^2, is

$$R_g^2 = \left[\text{corr}\left(y, \hat{y}\right)\right]^2 = r_{y\hat{y}}^2$$

In the wage equation $R_g^2 = \left[\text{corr}\left(y, \hat{y}_c\right)\right]^2 = 0.4647^2 = 0.2159$, as compared to the reported $R^2 = 0.2577$ from the regression of ln(*WAGE*) on *EDUC*. (In this case since the corrected and natural predictors differ only by a constant factor, the correlation is the same for both.) These R^2 values are small, but we repeat our earlier message: R^2 values tend to be small with microeconomic, cross-sectional data because the variations in individual behavior are difficult to fully explain.

4.5.3 Prediction Intervals in the Log-Linear Model

We have a corrected predictor $\hat{y}_c$ for y in the log-linear model. It is the "point" predictor, or point forecast, that is relevant if we seek the single number that is our best prediction of y.

If we prefer a prediction or forecast interval for y, then we must rely on the natural predictor $\hat{y}_n$.[1] Specifically, we follow the procedure outlined in Section 4.1 and then take antilogs. That is, compute $\widehat{\ln(y)} = b_1 + b_2 x$ and then $\widehat{\ln(y)} \pm t_c \text{se}(f)$, where the critical value t_c is the $100(1-\alpha/2)$-percentile from the t-distribution and $\text{se}(f)$ is given in (4.5). Then, a $100(1-\alpha)\%$ prediction interval for y is

$$\left[\exp\left(\widehat{\ln(y)} - t_c \text{se}(f)\right), \exp\left(\widehat{\ln(y)} + t_c \text{se}(f)\right)\right]$$

EXAMPLE 4.12 | Prediction Intervals for a Log-Linear Model

For the wage data, a 95% prediction interval for the wage of a worker with 12 years of education is

$$\begin{aligned}&\left[\exp(2.7819 - 1.96 \times 0.4850), \exp(2.7819 + 1.96 \times 0.4850)\right] \\ &= [6.2358, 41.8233]\end{aligned}$$

The interval prediction is \$6.24–\$41.82, which is so wide that it is basically useless. What does this tell us? Nothing we did not already know. Our model is not an accurate predictor of individual behavior in this case. In later chapters, we will see if we can improve this model by adding additional explanatory variables, such as experience, that should be relevant. The prediction interval is shown in Figure 4.14.

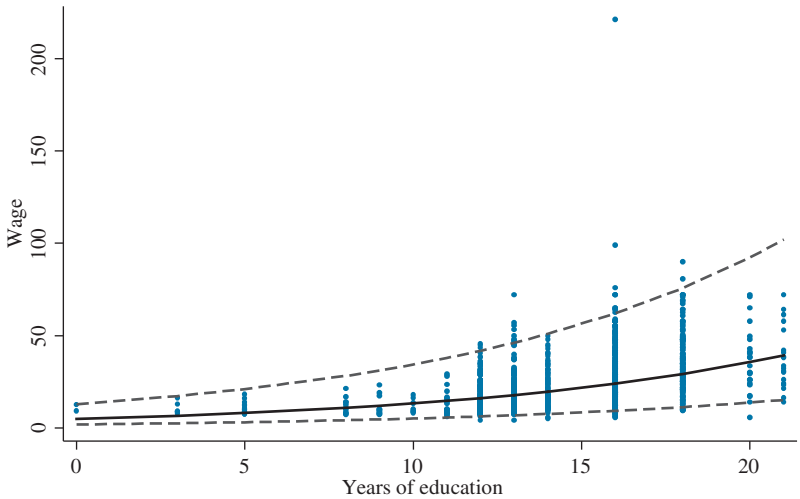

FIGURE 4.14 **The 95% prediction interval for wage.**

4.6 Log-Log Models

The log-log function, $\ln(y) = \beta_1 + \beta_2 \ln(x)$, is widely used to describe demand equations and production functions. The name "log-log" comes from the fact that the logarithm appears on both sides of the equation. In order to use this model, all values of y and x must be positive. Using the

[1] See Appendix 4A. The corrected predictor includes the estimated error variance, making the t-distribution no longer relevant in (4A.1).

properties of logarithms, we can see how to interpret the parameter of a log-log model. Consider an increase in x from x_0 to x_1. The change in the log-log model is from $\ln(y_0) = \beta_1 + \beta_2 \ln(x_0)$ to $\ln(y_1) = \beta_1 + \beta_2 \ln(x_1)$. Subtracting the first equation from the second gives $\ln(y_1) - \ln(y_0) = \beta_2[\ln(x_1) - \ln(x_0)]$. Multiply by 100, and use the approximation introduced in Appendix A, equation (A.3) to obtain $100[\ln(y_1) - \ln(y_0)] \cong \%\Delta y$ and $100[\ln(x_1) - \ln(x_0)] \cong \%\Delta x$, so that $\%\Delta y = \beta_2 \%\Delta x$, or $\beta_2 = \%\Delta y / \%\Delta x = \varepsilon_{yx}$. That is, in the log-log model, the parameter β_2 is the elasticity of y with respect to a change in x, and it is constant over the entire curve.

A useful way to think about the log-log model comes from a closer inspection of its slope. The slope of the log-log model changes at every point, and it is given by $dy/dx = \beta_2(y/x)$. Rearrange this so that $\beta_2 = (dy/y)/(dx/x)$. Thus, the slope of the log-log function exhibits constant *relative* change, whereas the linear function displays constant absolute change. The log-log function is a transformation of the equation $y = Ax^{\beta_2}$, with $\beta_1 = \ln(A)$. The various shape possibilities for log-log models are depicted in Figure 4.5(c), for $\beta_2 > 0$, and Figure 4.5(d), for $\beta_2 < 0$.

If $\beta_2 > 0$, then y is an increasing function of x. If $\beta_2 > 1$, then the function increases at an increasing rate. That is, as x increases the slope increases as well. If $0 < \beta_2 < 1$, then the function is increasing, but at a decreasing rate; as x increases, the slope decreases.

If $\beta_2 < 0$, then there is an inverse relationship between y and x. If, for example, $\beta_2 = -1$, then $y = Ax^{-1}$ or $xy = A$. This curve has "unit" elasticity. If we let y = quantity demanded and x = price, then A = total revenue from sales. For every point on the curve $xy = A$, the area under the curve A (total revenue for the demand curve) is constant. By definition, unit elasticity implies that a 1% increase in x (price, for example) is associated with a 1% decrease in y (quantity demanded), so that the product xy (price times quantity) remains constant.

EXAMPLE 4.13 | A Log-Log Poultry Demand Equation

The log-log functional form is frequently used for demand equations. Consider, for example, the demand for edible chicken, which the U.S. Department of Agriculture calls "broilers." The data for this exercise are in the data file *newbroiler*, which is adapted from the data provided by Epple and McCallum (2006).[2] The scatter plot of Q = per capita consumption of chicken, in pounds, versus P = real price of chicken is shown in Figure 4.15 for 52 annual observations, 1950–2001. It shows the characteristic hyperbolic shape that was displayed in Figure 4.5(d).

The estimated log-log model is

$$\widehat{\ln(Q)} = 3.717 - 1.121 \times \ln(P) \quad R_g^2 = 0.8817 \qquad (4.15)$$
$$\text{(se)} \quad (0.022) \quad (0.049)$$

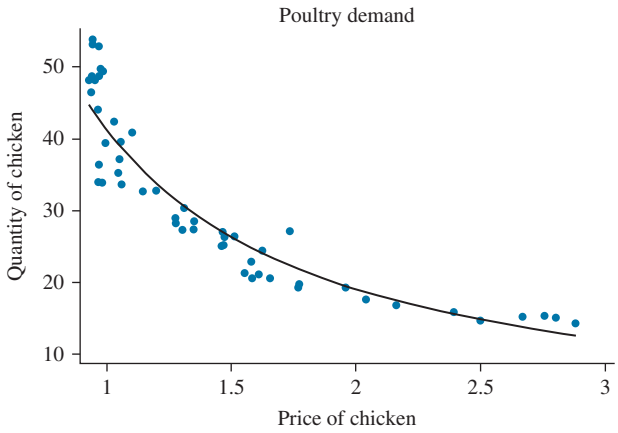

FIGURE 4.15 **Quantity and price of chicken.**

[2] "Simultaneous Equation Econometrics: The Missing Example," *Economic Inquiry*, 44(2), 374–384.

We estimate that the price elasticity of demand is 1.121: a 1% increase in real price is estimated to reduce quantity consumed by 1.121%.

The fitted line shown in Figure 4.15 is the "corrected" predictor discussed in Section 4.5.3. The corrected predictor $\hat{Q}_c$ is the natural predictor $\hat{Q}_n$ adjusted by the factor $\exp\left(\hat{\sigma}^2/2\right)$. That is, using the estimated error variance $\hat{\sigma}^2 = 0.0139$, the predictor is

$$\hat{Q}_c = \hat{Q}_n e^{\hat{\sigma}^2/2} = \exp\left(\widehat{\ln(Q)}\right) e^{\hat{\sigma}^2/2}$$
$$= \exp\left(3.717 - 1.121 \times \ln(P)\right) e^{0.0139/2}$$

The goodness-of-fit statistic $R_g^2 = 0.8817$ is the generalized R^2 discussed in Section 4.5.4. It is the squared correlation between the predictor $\hat{Q}_c$ and the observations Q

$$R_g^2 = \left[\text{corr}\left(Q, \hat{Q}_c\right)\right]^2 = [0.939]^2 = 0.8817$$

4.7 Exercises

4.7.1 Problems

4.1 Answer each of the following:

a. Suppose that a simple regression has quantities $N = 20$, $\sum y_i^2 = 7825.94$, $\bar{y} = 19.21$, and $SSR = 375.47$, find R^2.

b. Suppose that a simple regression has quantities $R^2 = 0.7911$, $SST = 725.94$, and $N = 20$, find $\hat{\sigma}^2$.

c. Suppose that a simple regression has quantities $\sum\left(y_i - \bar{y}\right)^2 = 631.63$ and $\sum \hat{e}_i^2 = 182.85$, find R^2.

4.2 Consider the following estimated regression equation (standard errors in parentheses):

$$\hat{y} = 64.29 + 0.99x \quad R^2 = 0.379$$
$$\text{(se)} \quad (2.42) \quad (0.18)$$

Rewrite the estimated equation, including coefficients, standard errors, and R^2, that would result if

a. All values of x were divided by 10 before estimation.

b. All values of y were divided by 10 before estimation.

c. All values of y and x were divided by 10 before estimation.

4.3 We have five observations on x and y. They are $x_i = 3, 2, 1, -1, 0$ with corresponding y values $y_i = 4, 2, 3, 1, 0$. The fitted least squares line is $\hat{y}_i = 1.2 + 0.8x_i$, the sum of squared least squares residuals is $\sum_{i=1}^5 \hat{e}_i^2 = 3.6$, $\sum_{i=1}^5 \left(x_i - \bar{x}\right)^2 = 10$, and $\sum_{i=1}^5 \left(y_i - \bar{y}\right)^2 = 10$. Carry out this exercise with a hand calculator. Compute

a. the predicted value of y for $x_0 = 4$.

b. the se(f) corresponding to part (a).

c. a 95% prediction interval for y given $x_0 = 4$.

d. a 99% prediction interval for y given $x_0 = 4$.

e. a 95% prediction interval for y given $x = \bar{x}$. Compare the width of this interval to the one computed in part (c).

4.4 The general manager of a large engineering firm wants to know whether the experience of technical artists influences their work quality. A random sample of 50 artists is selected. Using years of work experience (*EXPER*) and a performance rating (*RATING*, on a 100-point scale), two models are estimated by least squares. The estimates and standard errors are as follows:

Model 1:

$$\widehat{RATING} = 64.289 + 0.990EXPER \quad N = 50 \quad R^2 = 0.3793$$
$$\text{(se)} \qquad (2.422) \quad (0.183)$$

Model 2:

$$\widehat{RATING} = 39.464 + 15.312\ln(EXPER) \quad N = 46 \quad R^2 = 0.6414$$
$$\text{(se)} \qquad (4.198) \quad (1.727)$$

a. Sketch the fitted values from Model 1 for $EXPER = 0$ to 30 years.
b. Sketch the fitted values from Model 2 against $EXPER = 1$ to 30 years. Explain why the four artists with no experience are not used in the estimation of Model 2.
c. Using Model 1, compute the marginal effect on *RATING* of another year of experience for (i) an artist with 10 years of experience and (ii) an artist with 20 years of experience.
d. Using Model 2, compute the marginal effect on *RATING* of another year of experience for (i) an artist with 10 years of experience and (ii) an artist with 20 years of experience.
e. Which of the two models fits the data better? Estimation of Model 1 using just the technical artists with some experience yields $R^2 = 0.4858$.
f. Do you find Model 1 or Model 2 more reasonable, or plausible, based on economic reasoning? Explain.

4.5 Consider the regression model $WAGE = \beta_1 + \beta_2 EDUC + e$. *WAGE* is hourly wage rate in U.S. 2013 dollars. *EDUC* is years of education attainment, or schooling. The model is estimated using individuals from an urban area.

$$\widehat{WAGE} = -10.76 + 2.461965EDUC, \quad N = 986$$
$$\text{(se)} \qquad (2.27)\ (0.16)$$

a. The sample standard deviation of *WAGE* is 15.96 and the sum of squared residuals from the regression above is 199,705.37. Compute R^2.
b. Using the answer to (a), what is the correlation between *WAGE* and *EDUC*? [*Hint*: What is the correlation between *WAGE* and the fitted value $\widehat{WAGE}$?]
c. The sample mean and variance of *EDUC* are 14.315 and 8.555, respectively. Calculate the leverage of observations with $EDUC = 5$, 16, and 21. Should any of the values be considered large?
d. Omitting the ninth observation, a person with 21 years of education and wage rate \$30.76, and reestimating the model we find $\hat{\sigma} = 14.25$ and an estimated slope of 2.470095. Calculate DFBETAS for this observation. Should it be considered large?
e. For the ninth observation, used in part (d), DFFITS $= -0.0571607$. Is this value large? The leverage value for this observation was found in part (c). How much does the fitted value for this observation change when this observation is deleted from the sample?
f. For the ninth observation, used in parts (d) and (e), the least squares residual is -10.18368. Calculate the studentized residual. Should it be considered large?

4.6 We have five observations on x and y. They are $x_i = 3, 2, 1, -1, 0$ with corresponding y values $y_i = 4, 2, 3, 1, 0$. The fitted least squares line is $\hat{y}_i = 1.2 + 0.8x_i$, the sum of squared least squares residuals is $\sum_{i=1}^{5} \hat{e}_i^2 = 3.6$ and $\sum_{i=1}^{5} (y_i - \bar{y})^2 = 10$. Carry out this exercise with a hand calculator.

a. Calculate the fitted values $\hat{y}_i$ and their sample mean $\bar{\hat{y}}$. Compare this value to the sample mean of the y values.
b. Calculate $\sum_{i=1}^{5} (\hat{y}_i - \bar{y})^2$ and $\sum_{i=1}^{5} (\hat{y}_i - \bar{y})^2 / \sum_{i=1}^{5} (y_i - \bar{y})^2$.
c. The least squares residuals are $\hat{e}_i = 0.4, -0.8, 1, 0.6$, and -1.2. Calculate $\sum_{i=1}^{5} (\hat{y}_i - \bar{y})\hat{e}_i$.
d. Calculate $1 - \sum_{i=1}^{5} \hat{e}_i^2 / \sum_{i=1}^{5} (y_i - \bar{y})^2$ and compare it to the results in part (b).
e. Show, algebraically, that $\sum_{i=1}^{5} (\hat{y}_i - \bar{y})(y_i - \bar{y}) = \sum_{i=1}^{5} \hat{y}_i y_i - N\bar{y}^2$. Calculate this value.
f. Using $\sum_{i=1}^{5} (x_i - \bar{x})^2 = 10$, and previous results, calculate

$$r = \left[\sum_{i=1}^{5} (\hat{y}_i - \bar{y})(y_i - \bar{y})\right] \Big/ \left[\sqrt{\sum_{i=1}^{5} (\hat{y}_i - \bar{y})^2} \sqrt{\sum_{i=1}^{5} (y_i - \bar{y})^2}\right]$$

What statistic is r? Calculate r^2 and compare this value to the values in parts (d) and (b).

4.7 We have data on 2323 randomly selected households consisting of three persons in 2013. Let *ENTERT* denote the monthly entertainment expenditure (\$) per person per month and let *INCOME* (\$100) be monthly household income. Consider the regression model

$$ENTERT_i = \beta_1 + \beta_2 INCOME_i + e_i, \quad i = 1, \ldots, 2323$$

Assume that assumptions SR1–SR6 hold. The OLS estimated equation is $\widehat{ENTERT}_i = 9.820 + 0.503INCOME_i$. The standard error of the slope coefficient estimator is $\text{se}(b_2) = 0.029$, the standard

error of the intercept estimator is $\text{se}(b_1) = 2.419$, and the estimated covariance between the least squares estimators b_1 and b_2 is -0.062. From the summary statistics, we find

$$\sum_{i=1}^{2323}\left(ENTERT_i - \overline{ENTERT}\right)^2 = 8691035,\ \sum_{i=1}^{2323}\left(INCOME_i - \overline{INCOME}\right)^2 = 3876440$$
$$\overline{ENTERT} = 45.93,\ \overline{INCOME} = 71.84$$

a. From the estimated regression, the sum of squared least squares residuals is 7711432. How well does the regression model fit the data? How much of the household variation in entertainment expenses have we explained using this regression model? Explain your answer.

b. The Jones household has income of \$10,000 per month. Predict their per person household expenditure on entertainment.

c. Calculate a 95% prediction interval for the Jones household's per person expenditure on entertainment. Show your work.

d. Calculate a 95% prediction interval for the Jones household's total household expenditure on entertainment. Show your work.

4.8 Consider a log-linear regression for the weekly sales (number of cans) of a national brand of canned tuna ($SAL1$ = target brand sales) as a function of the ratio of its price to the price of a competitor, $RPRICE3 = 100$(price of target brand ÷ price competitive brand #3), $\ln(SAL1) = \gamma_1 + \gamma_2 RPRICE3 + e$. Using $N = 52$ weekly observations, the OLS estimated equation is

$$\widehat{\ln(SAL1)} = 11.481 - 0.031 RPRICE3$$
$$\text{(se)} \qquad (0.535) \quad (0.00529)$$

a. The sample mean of $RPRICE3$ is 99.66, its median is 100, its minimum value is 70.11, and its maximum value is 154.24. What do these summary statistics tell us about the prices of the target brand relative to the prices of its competitor?

b. Interpret the coefficient of $RPRICE3$. Does its sign make economic sense?

c. Using the "natural" predictor, predict the weekly sales of the target brand if $RPRICE3$ takes its sample mean value. What is the predicted sales if $RPRICE3$ equals 140?

d. The estimated value of the error variance from the regression above is $\hat{\sigma}^2 = 0.405$ and $\sum_{i=1}^{52}\left(RPRICE3_i - \overline{RPRICE3}\right)^2 = 14757.57$. Construct a 90% prediction interval for the weekly sales of the target brand if $RPRICE3$ takes its sample mean value. What is the 90% prediction interval for sales if $RPRICE3$ equals 140? Is one interval wider? Explain why this happens.

e. The fitted value of $\ln(SAL1)$ is $\widehat{\ln(SAL1)}$. The correlation between $\ln(SAL1)$ and $\widehat{\ln(SAL1)}$ is 0.6324, the correlation between $\widehat{\ln(SAL1)}$ and $SAL1$ is 0.5596, and the correlation between $\exp\left[\widehat{\ln(SAL1)}\right]$ and $SAL1$ is 0.6561. Calculate the R^2 that would normally be shown with the fitted regression output above. What is its interpretation? Calculate the "generalized-R^2." What is its interpretation?

4.9 Consider the weekly sales (number of cans) of a national brand of canned tuna ($SAL1$ = target brand sales) as a function of the ratio of its price to the price of a competitor, $RPRICE3 = 100$(price of target brand ÷ price competitive brand #3). Using $N = 52$ weekly observations, and for this exercise scaling $SAL1/1000$ so that we have sales measured as thousands of cans per week, we obtain the following least squares estimated equations, the first being a linear specification, the second a log-linear specification, and the third a log-log specification.

$$\widehat{SAL1} = 29.6126 - 0.2297 RPRICE3 \qquad \widehat{\ln(SAL1)} = 4.5733 - 0.0305 RPRICE3$$
$$\text{(se)} \quad (4.86) \quad (4.81) \qquad\qquad \text{(se)} \quad (0.54) \quad (0.0053)$$

$$\widehat{\ln(SAL1)} = 16.6806 - 3.3020 \ln(RPRICE3)$$
$$\text{(se)} \qquad (2.413) \quad (0.53)$$

a. For the linear specification, the sum of squared residuals is 1674.92, the estimated skewness and kurtosis of the residuals are 1.49 and 5.27, respectively. Calculate the Jarque–Bera statistic and test the hypothesis that the random errors in this specification are normally distributed, at the 5% level of significance. Specify the distribution of the test statistic if the null hypothesis of normality is true and the rejection region.

b. For the log-linear specification, the estimated skewness and kurtosis of the residuals are 0.41 and 2.54, respectively. Calculate the Jarque–Bera statistic and test the hypothesis that the random errors in this specification are normally distributed, at the 5% level of significance.

c. For the log-log specification, the estimated skewness and kurtosis of the residuals are 0.32 and 2.97, respectively. Calculate the Jarque–Bera statistic and test the hypothesis that the random errors in this specification are normally distributed, at the 5% level of significance.

d. For the log-linear and log-log specifications, define a residual as $SAL1 - \exp\left(\widehat{\ln(SAL1)}\right)$. For the two models, the sum of the squared residuals as defined are 1754.77 for the log-linear model and 1603.14 for the log-log model. Based on these values, and comparing them to the sum of squared residuals from the linear specification, which model seems to fit the data best?

e. Table 4.2 reports correlations between the regression model variables and predictions from the linear relationship (*YHAT*), predictions from the log-linear relationship $\left(YHATL = \exp\left[\widehat{\ln(SAL1)}\right]\right)$, and predictions from the log-log model $\left(YHATLL = \exp\left[\widehat{\ln(SAL1)}\right]\right)$.

- **i.** Why is the correlation between *SAL*1 and *RPRICE*3 the same as the correlation between *YHAT* and *SAL*1 (except for the sign)?
- **ii.** What is the R^2 from the linear relationship model?
- **iii.** Why is the correlation between *YHAT* and *RPRICE*3 a perfect—1.0?
- **iv.** What is the generalized-R^2 for the log-linear model?
- **v.** What is the generalized-R^2 for the log-log model?

f. Given the information provided in parts (a)–(e) which model would you select as having the best fit to the data?

TABLE 4.2 Correlations for Exercise 4.9

	*RPRICE*3	*SAL*1	*YHAT*	*YHATL*	*YHATLL*
*RPRICE*3	1.0000				
*SAL*1	−0.5596	1.0000			
YHAT	−1.0000	0.5596	1.0000		
YHATL	−0.9368	0.6561	0.9368	1.0000	
YHATLL	−0.8936	0.6754	0.8936	0.9927	1.0000

4.10 Using data on 76 countries, we estimate a relationship between the growth rate in prices, *INFLAT*, and the rate of growth in the money supply, *MONEY*. The least squares estimates of the model are as follows:

$$\begin{aligned} INFLAT &= -5.57 + 1.05MONEY \quad R^2 = 0.9917 \\ (\text{se}) \quad &\quad (0.70)\ (0.11) \end{aligned}$$

The data summary statistics are as follows:

	Mean	Median	Std. Dev.	Min	Max
INFLAT	25.35	8.65	58.95	−0.6	374.3
MONEY	29.59	16.35	56.17	2.5	356.7

Table 4.3 contains the data and some diagnostics for several observations.

a. Determine observations for which *LEVERAGE* is large. What is your rule?

b. Determine observations for which *EHATSTU* (the studentized residual) is large. What is your rule?

c. Determine observations for which *DFBETAS* is large. What is your rule?

d. Determine observations for which *DFFITS* is large. What is your rule?

e. Sketch the fitted relationship. On the graph locate the point of the means, and medians, and the nine data points in Table 4.3. Which observations are remote, relative to the center of the data, the point of the means and medians?

TABLE 4.3 Diagnostics for Selected Observations for Exercise 4.10

ID	INFLAT	MONEY	LEVERAGE	EHATSTU	DFBETAS	DFFITS
1	374.3	356.7	0.4654	1.8151	1.6694	1.6935
2	6.1	11.5	0.0145	−0.0644	0.0024	−0.0078
3	3.6	7.3	0.0153	0.2847	−0.0131	0.0354
4	187.1	207.1	0.1463	−5.6539	−2.2331	−2.3408
5	12.3	25.2	0.0132	−1.5888	0.0144	−0.1840
6	4.0	3.1	0.0161	1.1807	−0.0648	0.1512
7	316.1	296.6	0.3145	2.7161	1.8007	1.8396
8	13.6	17.4	0.0138	0.1819	−0.0046	0.0215
9	16.4	18.5	0.0137	0.4872	−0.0112	0.0574

4.11 Consider the regression model $WAGE = \beta_1 + \beta_2 EDUC + e$ where *WAGE* is hourly wage rate in U.S. 2013 dollars, *EDUC* is years of education attainment. The model is estimated twice, once using individuals from an urban area, and again for individuals in a rural area.

$$\text{Urban} \quad \underset{\text{(se)}}{\widehat{WAGE} = -10.76 + 2.46EDUC,} \quad \underset{(2.27)\ (0.16)}{N = 986}$$

$$\text{Rural} \quad \underset{\text{(se)}}{\widehat{WAGE} = -4.88 + 1.80EDUC,} \quad \underset{(3.29)\ (0.24)}{N = 214}$$

a. For the rural regression, compute a 95% prediction interval for *WAGE* if $EDUC = 16$, and the standard error of the forecast is 9.24. The standard error of the regression is $\hat{\sigma} = 9.20$ for the rural data.

b. For the urban data, the sum of squared deviations of *EDUC* about its sample mean is 8435.46 and the standard error of the regression is $\hat{\sigma} = 14.25$. The sample mean wage in the urban area is \$24.49. Calculate the 95% prediction interval for *WAGE* if $EDUC = 16$. Is the interval wider or narrower than the prediction interval for the rural data? Do you find this plausible? Explain.

4.12 Consider the share of total household expenditure (*TOTEXP*) devoted to expenditure on food (*FOOD*). Specify the log-linear relationship $FOOD/TOTEXP = \beta_1 + \beta_2 \ln(TOTEXP)$.

a. Show that the elasticity of expenditure on food with respect to total expenditure is

$$\varepsilon = \frac{dFOOD}{dTOTEXP} \times \frac{TOTEXP}{FOOD} = \frac{\beta_1 + \beta_2\left[\ln(TOTEXP) + 1\right]}{\beta_1 + \beta_2 \ln(TOTEXP)}$$

[*Hint*: Solve the log-linear relationship as $FOOD = \left[\beta_1 + \beta_2 \ln(TOTEXP)\right] TOTEXP$ and differentiate to obtain $dFOOD/dTOTEXP$. Then multiply by $TOTEXP/FOOD$ and simplify.]

b. The least squares estimates of the regression model $FOOD/TOTEXP = \beta_1 + \beta_2 \ln(TOTEXP) + e$, using 925 observations from London, are as follows:

$$\underset{(t)}{\widehat{\frac{FOOD}{TOTEXP}} = 0.953 - 0.129 \ln(TOTEXP)} \quad \underset{(26.10)\,(-16.16)}{R^2 = 0.2206,} \quad \hat{\sigma} = 0.0896$$

Interpret the estimated coefficient of $\ln(TOTEXP)$. What happens to the share of food expenditure in the budget as total household expenditures increase?

c. Calculate the elasticity in part (a) at the 5th percentile, and the 75th percentile of total expenditure. Is this a constant elasticity function? The 5th percentile is 500 UK pounds, and the 75th percentile is 1200 UK pounds.

d. The residuals from the model in (b) have skewness 0.0232 and kurtosis 3.4042. Carry out the Jarque–Bera test at the 1% level of significance. What are the null and alternative hypotheses for this test?

e. In $FOOD/TOTEXP = \beta_1 + \beta_2 \ln(TOTEXP)$, take the logarithm of the left-hand side and simplify the result to obtain $\ln(FOOD) = \alpha_1 + \alpha_2 \ln(TOTEXP)$. How are the parameters in this model related to the budget share relation?

f. The least squares estimates of $\ln(FOOD) = \alpha_1 + \alpha_2 \ln(TOTEXP) + e$ are as follows:

$$\widehat{\ln(FOOD)} = 0.732 + 0.608 \ln(TOTEXP) \quad R^2 = 0.4019 \quad \hat{\sigma} = 0.2729$$
$$(t) \qquad (6.58) \ (24.91)$$

Interpret the estimated coefficient of $\ln(TOTEXP)$. Calculate the elasticity in this model at the 5th percentile and the 75th percentile of total expenditure. Is this a constant elasticity function?

g. The residuals from the log-log model in (e) show skewness $= -0.887$ and kurtosis $= 5.023$. Carry out the Jarque–Bera test at the 5% level of significance.

h. In addition to the information in the previous parts, we multiply the fitted value in part (b) by *TOTEXP* to obtain a prediction for expenditure on food. The correlation between this value and actual food expenditure is 0.641. Using the model in part (e) we obtain $\exp\left[\widehat{\ln(FOOD)}\right]$. The correlation between this value and actual expenditure on food is 0.640. What if any information is provided by these correlations? Which model would you select for reporting, if you had to choose only one? Explain your choice.

4.13 The linear regression model is $y = \beta_1 + \beta_2 x + e$. Let $\bar{y}$ be the sample mean of the y-values and $\bar{x}$ the average of the x-values. Create variables $\tilde{y} = y - \bar{y}$ and $\tilde{x} = x - \bar{x}$. Let $\tilde{y} = \alpha \tilde{x} + e$.

a. Show, algebraically, that the least squares estimator of α is identical to the least square estimator of β_2. [*Hint*: See Exercise 2.4.]

b. Show, algebraically, that the least squares residuals from $\tilde{y} = \alpha \tilde{x} + e$ are the same as the least squares residuals from the original linear model $y = \beta_1 + \beta_2 x + e$.

4.14 Using data on 5766 primary school children, we estimate two models relating their performance on a math test (*MATHSCORE*) to their teacher's years of experience (*TCHEXPER*).

Linear relationship

$$\widehat{MATHSCORE} = 478.15 + 0.81 TCHEXPER \quad R^2 = 0.0095 \quad \hat{\sigma} = 47.51$$
$$(se) \qquad (1.19) \ (0.11)$$

Linear-log relationship

$$\widehat{MATHSCORE} = 474.25 + 5.63 \ln(TCHEXPER) \quad R^2 = 0.0081 \quad \hat{\sigma} = 47.57$$
$$(se) \qquad (1.84) \ (0.84)$$

a. Using the linear fitted relationship, how many years of additional teaching experience is required to increase the expected math score by 10 points? Explain your calculation.

b. Does the linear fitted relationship imply that at some point there are diminishing returns to additional years of teaching experience? Explain.

c. Using the fitted linear-log model, is the graph of *MATHSCORE* against *TCHEXPER* increasing at a constant rate, at an increasing rate, or at a decreasing rate? Explain. How does this compare to the fitted linear relationship?

d. Using the linear-log fitted relationship, if a teacher has only one year of experience, how many years of extra teaching experience is required to increase the expected math score by 10 points? Explain your calculation.

e. 252 of the teachers had no teaching experience. What effect does this have on the estimation of the two models?

f. These models have such a low R^2 that there is no statistically significant relationship between expected math score and years of teaching experience. True or False? Explain your answer.

4.15 Consider a **log-reciprocal model** that relates the logarithm of the dependent variable to the reciprocal of the explanatory variable, $\ln(y) = \beta_1 + \beta_2(1/x)$. [Note: An illustration of this model is given in Exercise 4.17].

a. For what values of y is this model defined? Are there any values of x that cause problems?

b. Write the model in exponential form as $y = \exp\left[\beta_1 + \beta_2(1/x)\right]$. Show that the slope of this relationship is $dy/dx = \exp\left[\beta_1 + (\beta_2/x)\right] \times (-\beta_2/x^2)$. What sign must β_2 have for y and x to have a positive relationship, assuming that $x > 0$?

c. Suppose that $x > 0$ but it converges toward zero from above. What value does y converge to? What does y converge to as x approaches infinity?

d. Suppose $\beta_1 = 0$ and $\beta_2 = -4$. Evaluate the slope at the x-values 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0. As x increases, is the slope of the relationship increasing or decreasing, or both?

e. Show that the second derivative of the function is

$$\frac{d^2y}{dx^2} = \left(\frac{\beta_2^2}{x^4} + \frac{2\beta_2}{x^3}\right)\exp\left[\beta_1 + (\beta_2/x)\right]$$

Assuming $\beta_2 < 0$ and $x > 0$, set the equation to zero, and show that the x-value that makes the second derivative zero is $-\beta_2/2$. Does this result agree with your calculations in part (d)? [*Hint*: $\exp\left[\beta_1 + (\beta_2/x)\right] > 0$. You have solved for what is called an *inflection point*.]

4.7.2 Computer Exercises

4.16 In Section 4.6, we considered the demand for edible chicken, which the U.S. Department of Agriculture calls "broilers." The data for this exercise are in the file *newbroiler*.

a. Using the 52 annual observations, 1950–2001, estimate the **reciprocal model** $Q = \alpha_1 + \alpha_2(1/P) + e$. Plot the fitted value of Q = per capita consumption of chicken, in pounds, versus P = real price of chicken. How well does the estimated relation fit the data?

b. Using the estimated relation in part (a), compute the elasticity of per capita consumption with respect to real price when the real price is its median, \$1.31, and quantity is taken to be the corresponding value on the fitted curve.

[*Hint*: The derivative (slope) of the reciprocal model $y = a + b(1/x)$ is $dy/dx = -b(1/x^2)$.] Compare this estimated elasticity to the estimate found in Section 4.6 where the log-log functional form was used.

c. Estimate the poultry consumption using the linear-log functional form $Q = \gamma_1 + \gamma_2 \ln(P) + e$. Plot the fitted values of Q = per capita consumption of chicken, in pounds, versus P = real price of chicken. How well does the estimated relation fit the data?

d. Using the estimated relation in part (c), compute the elasticity of per capita consumption with respect to real price when the real price is its median, \$1.31. Compare this estimated elasticity to the estimate from the log-log model and from the reciprocal model in part (b).

e. Estimate the poultry consumption using a log-linear model $\ln(Q) = \phi_1 + \phi_2 P + e$. Plot the fitted values of Q = per capita consumption of chicken, in pounds, versus P = real price of chicken. How well does the estimated relation fit the data?

f. Using the estimated relation in part (e), compute the elasticity of per capita consumption with respect to real price when the real price is its median, \$1.31. Compare this estimated elasticity to the estimate from the previous models.

g. Evaluate the suitability of the alternative models for fitting the poultry consumption data, including the log-log model. Which of them would you select as best, and why?

4.17 McCarthy and Ryan (1976) considered a model of television ownership in the United Kingdom and Ireland using data from 1955 to 1973. Use the data file *tvdata* for this exercise.

a. For the United Kingdom, plot the rate of television ownership (*RATE_UK*) against per capita consumer expenditures (*SPEND_UK*). Which models in Figure 4.5 are candidates to fit the data?

b. Estimate the linear-log model $RATE\_UK = \beta_1 + \beta_2 \ln(SPEND\_UK) + e$. Obtain the fitted values and plot them against *SPEND_UK*. How well does this model fit the data?

c. What is the interpretation of the intercept in the linear-log model? Specifically, for the model in (b), for what value of *SPEND_UK* is the expected value $E(RATE\_UK|SPEND\_UK) = \beta_1$?

d. Estimate the linear-log model $RATE\_UK = \beta_1 + \beta_2 \ln(SPEND\_UK - 280) + e$. Obtain the fitted values and plot them against *SPEND_UK*. How well does this model fit the data? How has the adjustment (−280) changed the fitted relationship? [Note: You might well wonder how the value 280 was determined. It was estimated using a procedure called *nonlinear least squares*. You will be introduced to this technique later in this book.]

e. A competing model is the log-reciprocal model, described in Exercise 4.15. Estimate the log-reciprocal model $\ln(RATE\_UK) = \alpha_1 + \alpha_2(1/SPEND\_UK) + e$. Obtain the fitted values and plot them against *SPEND_UK*. How well does this model fit the data?

f. Explain the failure of the model in (e) by referring to Exercise 4.15(c).

g. Estimate the log-reciprocal model $\ln(RATE\_UK) = \alpha_1 + \alpha_2(1/[SPEND\_UK - 280]) + e$. Obtain the fitted values and plot them against *SPEND_UK*. How well does this model fit the data? How has this modification corrected the problem identified in part (f)?

h. Repeat the above exercises for Ireland, with correcting factor 240 instead of 280.

4.18 Do larger universities have lower cost per student or a higher cost per student? Use the data on 141 public universities in the data file *pubcoll* for 2010 and 2011. A university is many things and here we only focus on the effect of undergraduate full-time student enrollment (*FTESTU*) on average total cost per student (*ACA*). Consider the regression model $ACA_{it} = \beta_1 + \beta_2 FTESTU_{it} + e_{it}$ where the subscripts i and t denote the university and the time period, respectively. Here, e_{it} is the usual random error term.

a. Estimate the model above using 2010 data only, again using 2011 data only, and again using both years of data together. What is the estimated effect of increasing enrollment on average cost per student? Base your answer on both point and 95% interval estimates.

b. There are certainly many other factors affecting average cost per student. Some of them can be characterized as the university "identity" or "image." Let us denote these largely unobservable individual attributes as u_i. If we could add this feature to the model, it would be $ACA_{it} = \beta_1 + \beta_2 FTESTU_{it} + (\theta u_i + e_{it})$. We place it in parentheses with e_{it} because it is another unobservable random error, but it is different because the character or identify of a university does not change from one year to the next. Do you suppose that our usual exogeneity assumptions hold in light of this new class of omitted variables? Might some unobservable characteristics of a university be correlated with student enrollment? Give some examples.

c. With our two years of data, we can take "first differences," by subtracting the model in 2010 from the model in 2011, $\Delta ACA_i = \beta_2 \Delta FTESTU_i + \Delta e_i$, where

$$\Delta ACA_i = ACA_{i,2011} - ACA_{i,2010}$$
$$\Delta FTESTU_i = FTESTU_{i,2011} - FTESTU_{i,2010}$$
$$\Delta e_i = e_{i,2011} - e_{i,2010}$$

Explain why the intercept and θu_i drop from the model. Explain how the exogeneity assumptions might now hold.

d. Estimate $\Delta ACA_i = \beta_2 \Delta FTESTU_i + \Delta e_i$ and also $\Delta ACA_i = \delta + \beta_2 \Delta FTESTU_i + \Delta e_i$. What now is the estimated effect of increasing enrollment on average cost per student? Base your answer on both point and 95% interval estimates. Does adding an intercept to the model make any fundamental difference in this case?

e. Estimate the model $\Delta\ln(ACA_i) = \alpha + \gamma \Delta\ln(FTESTU_i) + \Delta e_i$ where

$$\Delta\ln(ACA_i) = \ln(ACA_{i,2011}) - \ln(ACA_{i,2010})$$

and

$$\Delta\ln(FTESTU_i) = \ln(FTESTU_{i,2011}) - \ln(FTESTU_{i,2010})$$

Interpret the estimated coefficient of $\Delta\ln(FTESTU_i)$.

[*Hint*: See equation (A.3) in Appendix A.]

4.19 The data file *wa_wheat* contains wheat yield for several shires in Western Australia from 1950 to 1997.

a. If the variable *YIELD* is "average wheat yield" in tonnes per hectare what is the interpretation of $RYIELD = 1/YIELD$?

b. For Northampton and Mullewa shires, plot $RYIELD = 1/YIELD$ against $YEAR = 1949 + TIME$. Do you notice any anomalies in the plots? What years are most unusual? Using your favorite search engine discover what conditions may have affected wheat production in these shires during these years.

c. For Northampton and Mullewa shires, estimate the reciprocal model $RYIELD = \alpha_1 + \alpha_2 TIME + e$. Interpret the estimated coefficient. What does the sign tell us?

d. For the estimations in part (c), test the hypothesis that the coefficient of *TIME* is greater than or equal to zero against the alternative that it is negative, at the 5% level of significance.

e. For each of the estimations in part (c), calculate studentized residuals, and values for the diagnostics *LEVERAGE*, *DFBETAS*, and *DFFITS*. Identify the years in which these are "large" and include your threshold for what is large.

f. Discarding correct data is hardly ever a good idea, and we recommend that you not do it. Later in this book, you will discover other methods for addressing such problems—such as adding

additional explanatory variables—but for now experiment. For each shire, identify the most unusual observation. What grounds did you use for choosing?

g. Drop the most unusual observation for each shire and reestimate the model. How much do the results change? How do these changes relate to the diagnostics in part (e)?

4.20 In the log-linear model $\ln(y) = \beta_1 + \beta_2 x + e$, the corrected predictor $\hat{y}_c = \exp(b_1 + b_2 x) \times \exp\left(\hat{\sigma}^2/2\right)$ is argued to have a lower mean squared error than the "normal" predictor $\hat{y}_n = \exp(b_1 + b_2 x)$. The correction factor $\exp\left(\hat{\sigma}^2/2\right)$ depends on the regression errors having a normal distribution.

a. In exponential form, the log-linear model is $y = \exp(\beta_1 + \beta_2 x)\exp(e)$. Assuming that the explanatory variable x and the random error e are statistically independent, find $E(y)$.

b. Use the data file *cps5_small* for this exercise. [The data file *cps5* contains more observations and variables.] Estimate the model $\ln(WAGE) = \beta_1 + \beta_2 EDUC + e$ using the first 1000 observations. Based on this regression, calculate the correction factor $c = \exp\left(\hat{\sigma}^2/2\right)$. What is this value?

c. Obtain the 1000 least squares residuals $\hat{e}$ from the regression in (b). Calculate the correction factor $d = \sum_{i=1}^{1000} \exp(\hat{e}_i)/1000$. What is this value?

d. Using the estimates from part (b), obtain the predictions for observations 1001–1200, using $\hat{y}_n = \exp(b_1 + b_2 x)$, $\hat{y}_c = c\hat{y}_n$, and $\hat{y}_d = d\hat{y}_n$. Calculate the mean (average) squared forecast errors $MSE_n = \sum_{i=1001}^{1200} (\hat{y}_{ni} - y_i)^2/200$, $MSE_c = \sum_{i=1001}^{1200} (\hat{y}_{ci} - y_i)^2/200$, and $MSE_d = \sum_{i=1001}^{1200} (\hat{y}_{di} - y_i)^2/200$. Based on this criterion, which predictor is best?

4.21 The data file *malawi_small* contains survey data from Malawi during 2007–2008 on total household expenditures in the prior month (in Malawian Kwacha) as well as expenditures on categories of goods such as food, clothes, and fuel.

a. Locate Malawi and its neighboring countries on a map. Find the exchange rate between US $1 and the Malawian Kwacha. What is the population size of Malawi? Which industry drives the Malawi economy?

b. Define the proportion of expenditure on food as $PFOOD = FOOD/TOTEXP$. Estimate the linear-log regression model $PFOOD = \beta_1 + \beta_2 \ln(TOTEXP) + e$ and report the estimation results. What happens to the share of total expenditure devoted to food as total expenditure rises. Construct a 95% interval estimate for β_2. Have we estimated this coefficient relatively precisely or not? Does the model fit the data well? Is there a problem?

c. The elasticity of expenditure on food with respect to total expenditure is

$$\varepsilon = \frac{dFOOD}{dTOTEXP} \times \frac{TOTEXP}{FOOD} = \frac{\beta_1 + \beta_2\left[\ln(TOTEXP) + 1\right]}{\beta_1 + \beta_2 \ln(TOTEXP)}$$

This result is derived in Exercise 4.12. Calculate the elasticity at the 5th percentile and the 75th percentile of total expenditure. Is this a constant elasticity function? If your software permits, calculate a standard error for the elasticity.

d. Calculate the least squares residuals from the model in (b). Construct a histogram of these residuals and plot them against $\ln(TOTEXP)$. Are any patterns evident? Find the sample skewness and kurtosis of the least squares residuals. Carry out the Jarque–Bera test at the 1% level of significance. What are the null and alternative hypotheses for this test?

e. Take the logarithm of the left-hand side of $FOOD/TOTEXP = \beta_1 + \beta_2 \ln(TOTEXP)$ and simplify the result, and add an error term, to obtain $\ln(FOOD) = \alpha_1 + \alpha_2 \ln(TOTEXP) + v$. Estimate this model. Interpret the estimated coefficient of $\ln(TOTEXP)$. What is the estimated elasticity of expenditure on food with respect to total expenditure?

f. Calculate the residuals from the model in (e). Construct a histogram of these residuals and plot them against $\ln(TOTEXP)$. Are any patterns evident? Find the sample skewness and kurtosis of the least squares residuals. Carry out the Jarque–Bera test at the 1% level of significance.

g. Estimate the linear-log model $FOOD = \gamma_1 + \gamma_2 \ln(TOTEXP) + u$. Discuss the estimation results. Calculate the elasticity of food expenditure with respect to total expenditure when food expenditure is at its 50th percentile and at its 75th percentile. Is this a constant elasticity function, or is elasticity increasing or decreasing?

h. Calculate the residuals from the model in (g). Construct a histogram of these residuals and plot them against $\ln(TOTEXP)$. Are any patterns evident? Find the sample skewness and kurtosis of the least squares residuals. Carry out the Jarque–Bera test at the 1% level of significance.

i. Calculate predicted values of expenditure on food from each model. Multiply the fitted value from the model in part (b) to obtain a prediction for expenditure on food. Using the model in part (e) obtain $\exp\left[\widehat{\ln(FOOD)}\right]$. For the model in part (g), obtain fitted values. Find the correlations between the actual value of *FOOD* and the three sets of predictions. What, if any, information is provided by these correlations? Which model would you select for reporting, if you had to choose only one? Explain your choice.

4.22 The data file *malawi_small* contains survey data from Malawi during 2007–2008 on total household expenditures in the prior month (in Malawian Kwacha) as well as expenditures on categories of goods such as food, clothes, and fuel.

a. Define the proportion of expenditure on food consumed away from home as $PFOODAWAY = FOODAWAY/TOTEXP$. Construct a histogram for *PFOODAWAY* and its summary statistics. What percentage of the sample has a zero value for *PFOODAWAY*. What does that imply about their expenditures last month?

b. Create the variable $FOODAWAY = PFOODAWAY \times TOTEXP$. Construct a histogram for *FOODAWAY* and another histogram for *FOODAWAY* if $FOODAWAY > 0$. Compare the summary statistics for *TOTEXP* for households with $FOODAWAY > 0$ to those with $FOODAWAY = 0$. What differences do you observe?

c. Estimate the linear regression model $FOODAWAY = \beta_1 + \beta_2 TOTEXP + e$ twice, once for the full sample, and once using only households for whom $FOODAWAY > 0$. What differences in slope estimates do you observe? How would you explain these differences to an audience of noneconomists?

d. Calculate the fitted values from each of the estimated models in part (c) and plot the fitted values, and *FOODAWAY* values, versus *TOTEXP*. Think about how the least squares estimation procedure works to fit a line to data. Explain the relative difference in the two estimations based on this intuition.

4.23 The data file *malawi_small* contains survey data from Malawi during 2007–2008 on total household expenditures in the prior month (in Malawian Kwacha) as well as expenditures on categories of goods such as food, clothes, and fuel. Consider the following models.

i. Budget share: $PTELEPHONE = \beta_1 + \beta_2 \ln(TOTEXP) + e$

ii. Expenditure: $\ln(PTELEPHONE \times TOTEXP) = \alpha_1 + \alpha_2 \ln(TOTEXP) + e$

iii. Budget share: $PCLOTHES = \beta_1 + \beta_2 \ln(TOTEXP) + e$

iv. Expenditure: $\ln(PCLOTHES \times TOTEXP) = \alpha_1 + \alpha_2 \ln(TOTEXP) + e$

v. Budget share: $PFUEL = \beta_1 + \beta_2 \ln(TOTEXP) + e$

vi. Expenditure: $\ln(PFUEL \times TOTEXP) = \alpha_1 + \alpha_2 \ln(TOTEXP) + e$

a. Estimate each of the models (i) to (vi). Interpret the estimated coefficients of $\ln(TOTEXP)$. Is each item a necessity, or a luxury?

b. For each commodity equation (ii), (iv), and (vi), calculate the expenditure elasticity with respect to total expenditure at the 25th and 75th percentiles of *TOTEXP*.

c. For the budget share equations, (i), (iii), and (v), find the elasticities that are given by $\varepsilon = \dfrac{\beta_1 + \beta_2\left[\ln(TOTEXP) + 1\right]}{\beta_1 + \beta_2 \ln(TOTEXP)}$ (see Exercise 4.12). Are the changes in elasticities between the two percentiles, noticeable? [A standard log-log expenditure model can be obtained using the data, by creating a dependent variable that is the logarithm of the budget share times total expenditure. That is, for example, $\ln(TELEPHONE) = \ln(PTELEPHONE \times TOTEXP)$.]

4.24 Reconsider the presidential voting data (*fair5*) introduced in Exercises 2.23 and 3.24.

a. Using all the data from 1916 to 2012, estimate the regression model $VOTE = \beta_1 + \beta_2 GROWTH + e$. Based on these estimates, what is the predicted value of *VOTE* in favor of the Democrats in 2012? At the time of the election, a Democrat, Barack Obama, was the incumbent. What is the least squares residual for the 2012 election observation?

b. Estimate the regression in (a) using only data up to 2008. Predict the value of *VOTE* in 2012 using the actual value of *GROWTH* for 2012, which was 1.03%. What is the prediction error in this forecast? Is it larger or smaller than the error computed in part (a).

c. Using the regression results from (b), construct a 95% prediction interval for the 2012 value of *VOTE* using the actual value of $GROWTH = 1.03\%$.

d. Using the estimation results in (b), what value of *GROWTH* would have led to a prediction that the nonincumbent party [Republicans] would have won 50.1% of the vote in 2012?

e. Use the estimates from part (a), and predict the percentage vote in favor of the Democratic candidate in 2016. At the time of the election, a Democrat, Barack Obama, was the incumbent. Choose several values for *GROWTH* that represent both pessimistic and optimistic values for 2016. Cite the source of your chosen values for *GROWTH*.

4.25 The file *collegetown* contains data on 500 houses sold in Baton Rouge, LA during 2009–2013. Variable descriptions are in the file *collegetown.def*.

a. Estimate the log-linear model $\ln(PRICE) = \beta_1 + \beta_2 SQFT + e$. Interpret the estimated model parameters. Calculate the slope and elasticity at the sample means, if necessary.

b. Estimate the log-log model $\ln(PRICE) = \alpha_1 + \alpha_2 \ln(SQFT) + e$. Interpret the estimated parameters. Calculate the slope and elasticity at the sample means, if necessary.

c. Compare the R^2 value from the linear model $PRICE = \delta_1 + \delta_2 SQFT + e$ to the "generalized" R^2 measure for the models in (b) and (c).

d. Construct histograms of the least squares residuals from each of the models in (a)–(c) and obtain the Jarque–Bera statistics. Based on your observations, do you consider the distributions of the residuals to be compatible with an assumption of normality?

e. For each of the models in (a)–(c), plot the least squares residuals against *SQFT*. Do you observe any patterns?

f. For each model in (a)–(c), predict the value of a house with 2700 square feet.

g. For each model in (a)–(c), construct a 95% prediction interval for the value of a house with 2700 square feet.

h. Based on your work in this problem, discuss the choice of functional form. Which functional form would you use? Explain.

4.26 The file *collegetown* contains data on 500 houses sold in Baton Rouge, LA during 2009–2013. Variable descriptions are in the file *collegetown.def.*

a. Estimate the log-linear model $\ln(PRICE) = \beta_1 + \beta_2 SQFT + e$ for houses close to Louisiana State University [$CLOSE = 1$] and again for houses that are not close to Louisiana State University. How similar are the two sets of regression estimates. For each find the "corrected" predictor for a house with 2700 square feet of living area. What do you find?

b. Using the sample of homes that are not close to LSU [$CLOSE = 0$], find any observations on house sales that you would classify as unusual, based on the studentized residuals, *LEVERAGE*, *DFBETAS,* and *DFFITS*. Can you identify any house characteristics that might explain why they are unusual?

c. Estimate the log-linear model $\ln(PRICE) = \beta_1 + \beta_2 SQFT + e$ for houses for which $AGE < 7$ and again for houses with $AGE > 9$. Note that *AGE* is not the actual age of the house, but a category. Examine the file *collegetown.def* for the specifics. How similar are the two sets of regression estimates. For each find the "corrected" predictor of a house with 2700 square feet of living area. What do you find?

d. Using the sample of homes with $AGE > 9$, find any observations on house sales that you would classify as unusual, based on the studentized residuals, *LEVERAGE*, *DFBETAS*, and *DFFITS*. Can you identify any house characteristics that might explain why they are unusual?

4.27 Does the return to education differ by race and gender? For this exercise use the file *cps5*. [This is a large file with 9799 observations. If your software is a student version, you can use the smaller file *cps5_small* if your instructor permits]. In this exercise, you will extract subsamples of observations consisting of (i) white males, (ii) white females, (iii) black males, and (iv) black females.

a. For each sample partition, obtain the summary statistics of *WAGE*.

b. A variable's **coefficient of variation** (*CV*) is 100 times the ratio of its sample standard deviation to its sample mean. For a variable *y,* it is

$$CV = 100 \times \frac{s_y}{\bar{y}}$$

It is a measure of variation that takes into account the size of the variable. What is the coefficient of variation for *WAGE* within each sample partition?

c. For each sample partition, estimate the log-linear model

$$\ln(WAGE) = \beta_1 + \beta_2 EDUC + e$$

What is the approximate percentage return to another year of education for each group?

d. Create 95% interval estimates for the coefficient β_2 in each partition. Identify partitions for which the 95% interval estimates of the rate of return to education *do not* overlap. What does this imply about the population relations between wages and education for these groups? Are they similar or different? For the nonoverlapping pairs, test the null hypothesis that the parameter β_2 in one sample partition (the larger one, for simplicity) equals the estimated value in the other partition, using the 5% level of significance.

e. Create 95% interval estimates for the intercept coefficient in each partition. Identify partitions for which the 95% interval estimates for the intercepts *do not* overlap. What does this imply about the population relations between wages and education for these groups? Are they similar or different? For the nonoverlapping pairs, test the null hypothesis that the parameter β_1 in one sample partition (the larger one, for simplicity) equals the estimated value in the other partition, using the 5% level of significance.

f. Does the model fit the data equally well for each sample partition?

4.28 The file *wa-wheat.dat* contains observations on wheat yield in Western Australian shires. There are 48 annual observations for the years 1950–1997. For the Northampton shire, consider the following four equations:

$$
\begin{aligned}
YIELD_t &= \beta_0 + \beta_1 TIME + e_t \\
YIELD_t &= \alpha_0 + \alpha_1 \ln(TIME) + e_t \\
YIELD_t &= \gamma_0 + \gamma_1 TIME^2 + e_t \\
\ln(YIELD_t) &= \phi_0 + \phi_1 TIME + e_t
\end{aligned}
$$

a. Estimate each of the four equations. Taking into consideration (i) plots of the fitted equations, (ii) plots of the residuals, (iii) error normality tests, and (iii) values for R^2, which equation do you think is preferable? Explain.

b. Interpret the coefficient of the time-related variable in your chosen specification.

c. Using your chosen specification, identify any unusual observations, based on the studentized residuals, *LEVERAGE*, *DFBETAS*, and *DFFITS*.

d. Using your chosen specification, use the observations up to 1996 to estimate the model. Construct a 95% prediction interval for *YIELD* in 1997. Does your interval contain the true value?

4.29 Consider a model for household expenditure as a function of household income using the 2013 data from the Consumer Expenditure Survey, *cex5_small*. The data file *cex5* contains more observations. Our attention is restricted to three-person households, consisting of a husband, a wife, plus one other. In this exercise, we examine expenditures on a staple item, food. In this extended example, you are asked to compare the linear, log-log, and linear-log specifications.

a. Calculate summary statistics for the variables: *FOOD* and *INCOME*. Report for each the sample mean, median, minimum, maximum, and standard deviation. Construct histograms for both variables. Locate the variable mean and median on each histogram. Are the histograms symmetrical and "bell-shaped" curves? Is the sample mean larger than the median, or vice versa? Carry out the Jarque–Bera test for the normality of each variable.

b. Estimate the linear relationship $FOOD = \beta_1 + \beta_2 INCOME + e$. Create a scatter plot *FOOD* versus *INCOME* and include the fitted least squares line. Construct a 95% interval estimate for β_2. Have we estimated the effect of changing income on average *FOOD* relatively precisely, or not?

c. Obtain the least squares residuals from the regression in (b) and plot them against *INCOME*. Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. Is it more important for the variables *FOOD* and *INCOME* to be normally distributed, or that the random error e be normally distributed? Explain your reasoning.

d. Calculate both a point estimate and a 95% interval estimate of the elasticity of food expenditure with respect to income at $INCOME = 19, 65$, and 160, and the corresponding points on the fitted line, which you may treat as not random. Are the estimated elasticities similar or dissimilar? Do the interval estimates overlap or not? As *INCOME* increases should the income elasticity for food increase or decrease, based on Economics principles?

e. For expenditures on food, estimate the log-log relationship $\ln(FOOD) = \gamma_1 + \gamma_2 \ln(INCOME) + e$. Create a scatter plot for $\ln(FOOD)$ versus $\ln(INCOME)$ and include the fitted least squares line. Compare this to the plot in (b). Is the relationship more or less well-defined for the log-log model

relative to the linear specification? Calculate the generalized R^2 for the log-log model and compare it to the R^2 from the linear model. Which of the models seems to fit the data better?

f. Construct a point and 95% interval estimate of the elasticity for the log-log model. Is the elasticity of food expenditure from the log-log model similar to that in part (d), or dissimilar? Provide statistical evidence for your claim.

g. Obtain the least squares residuals from the log-log model and plot them against ln(*INCOME*). Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. What do you conclude about the normality of the regression errors in this model?

h. For expenditures on food, estimate the linear-log relationship $FOOD = \alpha_1 + \alpha_2 \ln(INCOME) + e$. Create a scatter plot for *FOOD* versus ln(*INCOME*) and include the fitted least squares line. Compare this to the plots in (b) and (e). Is this relationship more well-defined compared to the others? Compare the R^2 values. Which of the models seems to fit the data better?

i. Construct a point and 95% interval estimate of the elasticity for the linear-log model at $INCOME = 19, 65$, and 160, and the corresponding points on the fitted line, which you may treat as not random. Is the elasticity of food expenditure similar to those from the other models, or dissimilar? Provide statistical evidence for your claim.

j. Obtain the least squares residuals from the linear-log model and plot them against ln(*INCOME*). Do you observe any patterns? Construct a residual histogram and carry out the Jarque–Bera test for normality. What do you conclude about the normality of the regression errors in this model?

k. Based on this exercise, do you prefer the linear relationship model, or the log-log model or the linear-log model? Explain your reasoning.

4.30 Consider a model for household expenditure as a function of household income using the 2013 data from the Consumer Expenditure Survey, *cex5_small*. The data file *cex5* contains more observations. Our attention is restricted to three person households, consisting of a husband, a wife, plus one other. In this exercise, we examine expenditures on alcoholic beverages.

a. Obtain summary statistics for *ALCBEV*. How many households spend nothing on alcoholic beverages? Calculate the summary statistics restricting the sample to those households with positive expenditure on alcoholic beverages.

b. Plot *ALCBEV* against *INCOME* and include the fitted least squares regression line. Obtain the least squares estimates of the model $ALCBEV = \beta_1 + \beta_2 INCOME + e$. Obtain the least squares residuals and plot these versus *INCOME*. Does this plot appear random, as in Figure 4.7(a)? If the dependent variable in this regression model is zero ($ALCBEV = 0$), what is the least squares residual? For observations with $ALCBEV = 0$, is the least squares residual related to the explanatory variable *INCOME*? How?

c. Suppose that some households in this sample may never purchase alcohol, regardless of their income. If this is true, do you think that a linear regression including all the observations, even the observations for which $ALCBEV = 0$, gives a reliable estimate of the effect of income on average alcohol expenditure? If there is estimation bias, is the bias positive (the slope overestimated) or negative (slope underestimated)? Explain your reasoning.

d. For households with $ALCBEV > 0$, construct histograms for *ALCBEV* and ln(*ALCBEV*). How do they compare?

e. Create a scatter plot of ln(*ALCBEV*) against ln(*INCOME*) and include a fitted regression line. Interpret the coefficient of ln(*INCOME*) in the estimated log-log regression. How many observations are included in this estimation?

f. Calculate the least squares residuals from the log-log model. Create a histogram of these residuals and also plot them against ln(*INCOME*). Does this plot appear random, as in Figure 4.7(a)?

g. If we consider only the population of individuals who have positive expenditures for alcohol, do you prefer the linear relationship model, or the log-log model?

h. Expenditures on apparel have some similar features to expenditures on alcoholic beverages. You might reconsider the above exercises for *APPAR*. Think about part (c) above. Of those with no apparel expenditure last month, do you think there is a substantial portion who never purchase apparel regardless of income, or is it more likely that they sometimes purchase apparel but simply did not do so last month?

Appendix 4A Development of a Prediction Interval

The forecast error is $f = y_0 - \hat{y}_0 = (\beta_1 + \beta_2 x_0 + e_0) - (b_1 + b_2 x_0)$. To obtain its variance, let us first obtain the variance of $\hat{y}_0 = b_1 + b_2 x_0$. The variances and covariance of the least squares estimators are given in Section 2.4.4. Using them, and obtaining a common denominator, we obtain

$$\operatorname{var}(\hat{y}_0|\mathbf{x}) = \operatorname{var}\left[(b_1 + b_2 x_0)|\mathbf{x}\right] = \operatorname{var}(b_1|\mathbf{x}) + x_0^2 \operatorname{var}(b_2|\mathbf{x}) + 2x_0 \operatorname{cov}(b_1, b_2|\mathbf{x})$$

$$= \frac{\sigma^2}{N\sum(x_i - \bar{x})^2}\left[\sum x_i^2 + Nx_0^2 - 2N\bar{x}x_0\right]$$

The term in brackets can be simplified. First, factor N from the second and third terms to obtain $\sum x_i^2 + Nx_0^2 - 2N\bar{x}x_0 = \sum x_i^2 + N(x_0^2 - 2\bar{x}x_0)$. Complete the square within the parentheses by adding $\bar{x}^2$, and subtracting $N\bar{x}^2$ to keep the equality. Then the term in brackets is

$$\sum x_i^2 - N\bar{x}^2 + N\left(x_0^2 - 2\bar{x}x_0 + \bar{x}^2\right) = \sum(x_i - \bar{x})^2 + N(x_0 - \bar{x})^2$$

Finally

$$\operatorname{var}(\hat{y}_0|\mathbf{x}) = \sigma^2\left[\frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

Taking into account that x_0 and the unknown parameters β_1 and β_2 are not random, you should be able to show that $\operatorname{var}(f|\mathbf{x}) = \operatorname{var}(\hat{y}_0|\mathbf{x}) + \operatorname{var}(e_0) = \operatorname{var}(\hat{y}_0|\mathbf{x}) + \sigma^2$. A little factoring gives the result in (4.4). We can construct a standard normal random variable as

$$\frac{f}{\sqrt{\operatorname{var}(f|\mathbf{x})}} \sim N(0, 1)$$

If the forecast error variance in (4.4) is estimated by replacing σ^2 by its estimator $\hat{\sigma}^2$,

$$\widehat{\operatorname{var}}(f|\mathbf{x}) = \hat{\sigma}^2\left[1 + \frac{1}{N} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]$$

then

$$\frac{f}{\sqrt{\widehat{\operatorname{var}}(f|\mathbf{x})}} = \frac{y_0 - \hat{y}_0}{\operatorname{se}(f)} \sim t_{(N-2)} \tag{4A.1}$$

where the square root of the estimated variance is the standard error of the forecast given in (4.5). The t-ratio in (4A.1) is a pivotal statistic. It has a distribution that does not depend on $\mathbf{x}$ or any unknown parameters.

Using these results, we can construct an interval prediction procedure for y_0 just as we constructed confidence intervals for the parameters β_k. If t_c is a critical value from the $t_{(N-2)}$-distribution such that $P(t \geq t_c) = \alpha/2$, then

$$P(-t_c \leq t \leq t_c) = 1 - \alpha \tag{4A.2}$$

Substitute the t-random variable from (4A.1) into (4A.2) to obtain

$$P\left[-t_c \leq \frac{y_0 - \hat{y}_0}{\operatorname{se}(f)} \leq t_c\right] = 1 - \alpha$$

Simplify this expression to obtain

$$P\left[\hat{y}_0 - t_c \operatorname{se}(f) \leq y_0 \leq \hat{y}_0 + t_c \operatorname{se}(f)\right] = 1 - \alpha$$

A $100(1 - \alpha)\%$ confidence interval, or prediction interval, for y_0 is given by (4.6). This prediction interval is valid if $\mathbf{x}$ is fixed or random, as long as assumptions SR1–SR6 hold.

Appendix 4B The Sum of Squares Decomposition

To obtain the sum of squares decomposition in (4.11), we square both sides of (4.10)

$$(y_i - \bar{y})^2 = \left[(\hat{y}_i - \bar{y}) + \hat{e}_i\right]^2 = (\hat{y}_i - \bar{y})^2 + \hat{e}_i^2 + 2(\hat{y}_i - \bar{y})\hat{e}_i$$

Then sum

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum \hat{e}_i^2 + 2\sum (\hat{y}_i - \bar{y})\hat{e}_i$$

Expanding the last term, we obtain

$$\sum (\hat{y}_i - \bar{y})\hat{e}_i = \sum \hat{y}_i\hat{e}_i - \bar{y}\sum \hat{e}_i = \sum (b_1 + b_2 x_i)\hat{e}_i - \bar{y}\sum \hat{e}_i$$
$$= b_1 \sum \hat{e}_i + b_2 \sum x_i\hat{e}_i - \bar{y}\sum \hat{e}_i$$

Consider first the term $\sum \hat{e}_i$

$$\sum \hat{e}_i = \sum (y_i - b_1 - b_2 x_i) = \sum y_i - Nb_1 - b_2 \sum x_i = 0$$

This last expression is zero because of the first normal equation (2A.3). The first normal equation is valid *only if the model contains an intercept*. The sum of the least squares residuals is always zero *if* the model contains an intercept. It follows, then, that the *sample mean* of the least squares residuals is also zero (since it is the sum of the residuals divided by the sample size) if the model contains an intercept. That is, $\bar{\hat{e}} = \sum \hat{e}_i / N = 0$.

The next term $\sum x_i\hat{e}_i = 0$, because

$$\sum x_i\hat{e}_i = \sum x_i(y_i - b_1 - b_2 x_i) = \sum x_i y_i - b_1 \sum x_i - b_2 \sum x_i^2 = 0$$

This result follows from the second normal equation (2A.4). This result always holds for the least squares estimator and does not depend on the model having an intercept. See Appendix 2A for discussion of the normal equations. Substituting $\sum \hat{e}_i = 0$ and $\sum x_i\hat{e}_i = 0$ back into the original equation, we obtain $\sum (\hat{y}_i - \bar{y})\hat{e}_i = 0$.

Thus, if the model contains an intercept, it is guaranteed that $SST = SSR + SSE$. If, however, the model does not contain an intercept, then $\sum \hat{e}_i \neq 0$ and $SST \neq SSR + SSE$.

Appendix 4C Mean Squared Error: Estimation and Prediction

In Chapter 2, we discussed the properties of the least squares estimator. Under assumptions SR1–SR5, the least squares estimator is the **B**est **L**inear **U**nbiased **E**stimator (BLUE). There are no estimators that are both linear and unbiased that are better than the least squares estimator. However, this rules out many alternative estimators that statisticians and econometricians have developed over the years, which might be useful in certain contexts. Mean squared error (MSE) is an alternative metric for the quality of an estimator that doesn't depend on linearity or unbiasedness, and hence is more general.

In the linear regression model $y = \beta_1 + \beta_2 x + e$, suppose that we are keenly interested in obtaining an estimate of β_2 that is as close as possible to the true value. The mean squared error of an estimator $\hat{\beta}_2$ is

$$\text{MSE}\left(\hat{\beta}_2\right) = E\left[\left(\hat{\beta}_2 - \beta_2\right)^2\right] \tag{4C.1}$$

The term $\left(\hat{\beta}_2 - \beta_2\right)^2$ is the squared estimation error, that is, the squared difference or distance between the estimator $\hat{\beta}_2$ and the parameter β_2 of interest. Because the estimator $\hat{\beta}_2$ exhibits sampling variation, it is a random variable, and the squared term $\left(\hat{\beta}_2 - \beta_2\right)^2$ is also random. If we

think of "expected value" as "the average in all possible samples," then the mean squared error $E\left[\left(\hat{\beta}_2 - \beta_2\right)^2\right]$ is the average, or mean, squared error using $\hat{\beta}_2$ as an estimator of β_2. It measures how close the estimator $\hat{\beta}_2$ is on average to the true parameter β_2. We would like an estimator that is as close as possible to the true parameter and one that has a small mean squared error.

An interesting feature of an estimator's mean squared error is that it takes into account both the estimator's bias and its sampling variance. To see this we play a simple trick on equation (4C.1); we will add and subtract $E\left(\hat{\beta}_2\right)$ inside the parentheses and then square the result. That is,

$$\begin{aligned}
\mathrm{MSE}\left(\hat{\beta}_2\right) &= E\left[\left(\hat{\beta}_2 - \beta_2\right)^2\right] = E\left\{\left(\hat{\beta}_2 \underbrace{- E\left(\hat{\beta}_2\right) + E\left(\hat{\beta}_2\right)}_{=0} - \beta_2\right)^2\right\} \\
&= E\left\{\left(\left[\hat{\beta}_2 - E\left(\hat{\beta}_2\right)\right] + \left[E\left(\hat{\beta}_2\right) - \beta_2\right]\right)^2\right\} \\
&= E\left\{\left[\hat{\beta}_2 - E\left(\hat{\beta}_2\right)\right]^2\right\} + E\left\{\left[E\left(\hat{\beta}_2\right) - \beta_2\right]^2\right\} \\
&\quad + 2E\left\{\left[\hat{\beta}_2 - E\left(\hat{\beta}_2\right)\right]\left[E\left(\hat{\beta}_2\right) - \beta_2\right]\right\} \\
&= \mathrm{var}\left(\hat{\beta}_2\right) + \left[\mathrm{bias}\left(\hat{\beta}_2\right)\right]^2
\end{aligned} \tag{4C.2}$$

To go from the third to the fourth lines, we first recognize that $E\left\{\left[\hat{\beta}_2 - E\left(\hat{\beta}_2\right)\right]^2\right\} = \mathrm{var}\left(\hat{\beta}_2\right)$. Secondly, in the term $E\left\{\left[E\left(\hat{\beta}_2\right) - \beta_2\right]^2\right\}$, the outside expectation is not needed because $E\left(\hat{\beta}_2\right)$ is not random and β_2 is not random. The difference between an estimator's expected value and the true parameter is called the **estimator bias**, so $E\left(\hat{\beta}_2\right) - \beta_2 = \mathrm{bias}\left(\hat{\beta}_2\right)$. The term $\left[\mathrm{bias}\left(\hat{\beta}_2\right)\right]^2$ is the squared estimator bias. The final term in the third line of (4C.2) is zero. To see this note again that $\left[E\left(\hat{\beta}_2\right) - \beta_2\right]$ is not random, so that it can be factored out of the expectation

$$\begin{aligned}
2E\left\{\left[\hat{\beta}_2 - E\left(\hat{\beta}_2\right)\right]\left[E\left(\hat{\beta}_2\right) - \beta_2\right]\right\} &= 2\left[E\left(\hat{\beta}_2\right) - \beta_2\right]\left\{E\left[\hat{\beta}_2 - E\left(\hat{\beta}_2\right)\right]\right\} \\
&= 2\left[E\left(\hat{\beta}_2\right) - \beta_2\right]\left[E\left(\hat{\beta}_2\right) - E\left(\hat{\beta}_2\right)\right] \\
&= 2\left[E\left(\hat{\beta}_2\right) - \beta_2\right]0 = 0
\end{aligned}$$

We have shown that an estimator's mean squared error is the sum of its variance and squared bias,

$$\mathrm{MSE}\left(\hat{\beta}_2\right) = \mathrm{var}\left(\hat{\beta}_2\right) + \left[\mathrm{bias}\left(\hat{\beta}_2\right)\right]^2 \tag{4C.3}$$

This relationship is also true if we use conditional expectations. The conditional MSE is

$$\mathrm{MSE}\left(\hat{\beta}_2|\mathbf{x}\right) = \mathrm{var}\left(\hat{\beta}_2|\mathbf{x}\right) + \left[\mathrm{bias}\left(\hat{\beta}_2|\mathbf{x}\right)\right]^2 \tag{4C.4}$$

with $\mathrm{bias}\left(\hat{\beta}_2|\mathbf{x}\right) = E\left(\hat{\beta}_2|\mathbf{x}\right) - \beta_2$. Because the least squares estimator is unbiased under SR1–SR5, its mean squared error is

$$\mathrm{MSE}\left(b_2|\mathbf{x}\right) = \mathrm{var}\left(b_2|\mathbf{x}\right) + \left[\mathrm{bias}\left(b_2|\mathbf{x}\right)\right]^2 = \mathrm{var}\left(b_2|\mathbf{x}\right) + [0]^2 = \mathrm{var}\left(b_2|\mathbf{x}\right) \tag{4C.5}$$

The mean squared error concept can also be applied to more than one parameter at once. For example, the mean squared error of $\hat{\beta}_1$ and $\hat{\beta}_2$ as estimators of β_1 and β_2 is

$$\begin{aligned}\text{MSE}\left(\hat{\beta}_1, \hat{\beta}_2|\mathbf{x}\right) &= E\left\{\left[\left(\hat{\beta}_1-\beta_1\right)^2+\left(\hat{\beta}_2-\beta_2\right)^2\right]\middle|\mathbf{x}\right\}\\ &= \text{var}\left(\hat{\beta}_1|\mathbf{x}\right)+\left[\text{bias}\left(\hat{\beta}_1|\mathbf{x}\right)\right]^2+\text{var}\left(\hat{\beta}_2|\mathbf{x}\right)+\left[\text{bias}\left(\hat{\beta}_2|\mathbf{x}\right)\right]^2\end{aligned}$$

In the simple linear regression model, there are no estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ of β_1 and β_2 that have mean squared error $\text{MSE}\left(\hat{\beta}_1, \hat{\beta}_2|\mathbf{x}\right)$ smaller than the mean squared error for the least squares estimator, $\text{MSE}\left(b_1, b_2|\mathbf{x}\right)$, for any and all parameter values. This statement turns out not to be true in the multiple regression model.

We can apply the mean squared error concept to prediction situations too. Suppose that we are predicting an outcome y_0 using a predictor $\hat{y}_0(\mathbf{x})$, which is a function of the sample $\mathbf{x}$. The conditional mean squared error of the predictor is $E\left[\left(y_0-\hat{y}_0(\mathbf{x})\right)^2\middle|\mathbf{x}\right]$. We employ the same trick as in (4C.2), adding and subtracting $E\left(y_0|\mathbf{x}\right)$, the conditional expected value of y_0,

$$\begin{aligned}E\left[\left(y_0-\hat{y}_0(\mathbf{x})\right)^2\middle|\mathbf{x}\right] &= E\left[\left(y_0-E\left(y_0|\mathbf{x}\right)+E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right)^2\middle|\mathbf{x}\right]\\ &= E\left[\left(y_0-E\left(y_0|\mathbf{x}\right)\right)^2\middle|\mathbf{x}\right]+E\left[\left(E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right)^2\middle|\mathbf{x}\right]\\ &\quad+2E\left\{\left(\left[y_0-E\left(y_0|\mathbf{x}\right)\right]\left[E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right]\right)\middle|\mathbf{x}\right\}\\ &= \text{var}\left(y_0|\mathbf{x}\right)+\left\{\left[E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right]^2\middle|\mathbf{x}\right\}\end{aligned} \tag{4C.6}$$

The third line in (4C.6) is zero because conditional on $\mathbf{x}$ the term $E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})$ is not random, and it can be factored out of the expectation

$$\begin{aligned}2E\left\{\left(\left[y_0-E\left(y_0|\mathbf{x}\right)\right]\left[E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right]\right)\middle|\mathbf{x}\right\} &= 2\left(E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right)E\left\{\left(\left[y_0-E\left(y_0|\mathbf{x}\right)\right]\right)\middle|\mathbf{x}\right\}\\ &= 2\left(E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right)\left[E\left(y_0|\mathbf{x}\right)-E\left(y_0|\mathbf{x}\right)\right]\\ &= 2\left(E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right)\times 0=0\end{aligned}$$

The conditional mean squared error of our predictor is then

$$E\left[\left(y_0-\hat{y}_0(\mathbf{x})\right)^2\middle|\mathbf{x}\right]=\text{var}\left(y_0|\mathbf{x}\right)+\left[\left(E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right)^2\middle|\mathbf{x}\right] \tag{4C.7}$$

Using the law of iterated expectations

$$E\left[\left(y_0-\hat{y}_0(\mathbf{x})\right)\right]^2=E_{\mathbf{x}}\left[\text{var}\left(y_0|\mathbf{x}\right)\right]+E_{\mathbf{x}}\left\{\left[E\left(y_0|\mathbf{x}\right)-\hat{y}_0(\mathbf{x})\right]^2\right\} \tag{4C.8}$$

If we are choosing a predictor, then the one that minimizes the mean squared error is $\hat{y}_0(\mathbf{x})=E\left(y_0|\mathbf{x}\right)$. This makes the final term in (4C.8) zero. The conditional mean of y_0 is the minimum mean squared error predictor of y_0.