Chapter 6

The Simple Linear Regression Model: Reporting the Results and Choosing the Functional Form

To complete the analysis of the simple linear regression model, in this chapter we will consider

- How to measure the variation in $y_t$ explained by the model
- How to report the results of a regression analysis
- Some alternative functional forms that may be used to represent possible relationships between $y_t$ and $x_t$
6.1 The Coefficient of Determination

Two major reasons for analyzing the model

\[ y_t = \beta_1 + \beta_2 x_t + e_t \]  \hspace{1cm} (6.1.1)

are

1. to explain how the dependent variable \( y_t \) changes as the independent variable \( x_t \) changes, and

2. to predict \( y_0 \) given an \( x_0 \).

- Closely allied with the prediction problem is the desire to use \( x_t \) to explain as much of the variation in the dependent variable \( y_t \) as possible.
• In Equation (6.1.1) we introduce the “explanatory” variable $x_t$ in hope that its variation will “explain” the variation in $y_t$.

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

$$y_t = E(y_t) + e_t$$  \hspace{1cm} (6.1.2)

where $E(y_t) = \beta_1 + \beta_2 x_t$ is the explainable, “systematic” component of $y_t$, and $e_t$ is the random, unsystematic, unexplainable noise component of $y_t$.

• We can estimate the unknown parameters $\beta_1$ and $\beta_2$ and decompose the value of $y_t$ into

$$y_t = \hat{y}_t + e_t$$  \hspace{1cm} (6.1.3)
where \( \hat{y}_t = b_1 + b_2 x_t \) and \( e_t = y_t - \hat{y}_t \).

- In Figure 6.1 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

\[
y_t - \bar{y} = (\hat{y}_t - \bar{y}) + e_t \quad (6.1.4)
\]

- As shown in Figure 6.1, the difference between \( y_t \) and its mean value \( \bar{y} \) consists of a part that is “explained” by the regression model, \( \hat{y}_t - \bar{y} \), and a part that is unexplained, \( e_t \).

- A measure of the “total variation” in \( y \) is to square the differences between \( y_t \) and its mean value \( \bar{y} \) and sum over the entire sample. If we square and sum both sides of Equation (6.1.4) we obtain
\[ \sum (y_i - \bar{y})^2 = \sum [(\hat{y}_i - \bar{y}) + e_i]^2 \]

\[ = \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^2 + 2 \sum (\hat{y}_i - \bar{y})e_i \]  

(6.1.5)

\[ = \sum (\hat{y}_i - \bar{y})^2 + \sum e_i^2 \]

The cross-product term \( \sum (\hat{y}_i - \bar{y})e_i = 0 \) and drops out. Please see the solution of Exercise 6.5 for details.

- Equation (6.1.5) is a decomposition of the “total sample variation” in \( y \) into explained and unexplained components. Specifically, these “sums of squares” are:
  1. \( \sum (y_i - \bar{y})^2 = \text{total sum of squares} = \texttt{SST} \): a measure of total variation in \( y \) about its sample mean.
2. $\sum (\hat{y}_i - \bar{y})^2 = \text{explained sum of squares} = \text{SSR}$: that part of total variation in $y$ about its sample mean that is explained by the regression.

3. $\sum \hat{e}_i^2 = \text{error sum of squares} = \text{SSE}$: that part of total variation in $y$ about its mean that is not explained by the regression.

Thus,

$$\text{SST} = \text{SSR} + \text{SSE}$$  \hspace{1cm} (6.1.6)

This decomposition accompanies virtually every regression analysis.

- This decomposition is usually presented in what is called an “Analysis of Variance” table with general format of Table 6.1. This table provides a basis for summarizing the decomposition in Equation (6.1.5). It gives $\text{SSR}$, the variation explained by $x$, $\text{SSE}$, the unexplained variation and $\text{SST}$, the total variation in $y$. 

Table 6.1  Analysis of Variance Table

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explained</td>
<td>1</td>
<td>$SSR$</td>
<td>$SSR/1$</td>
</tr>
<tr>
<td>Unexplained</td>
<td>$T - 2$</td>
<td>$SSE$</td>
<td>$SSE/(T - 2)$ [ = $\sigma^2$ ]</td>
</tr>
<tr>
<td>Total</td>
<td>$T - 1$</td>
<td>$SST$</td>
<td></td>
</tr>
</tbody>
</table>

- The **degrees of freedom (DF)** for these sums of squares are:
  1. $df = 1$ for $SSR$ (the number of explanatory variables other than the intercept);
  2. $df = T - 2$ for $SSE$ (the number of observations minus the number of parameters in the model); and
  3. $df = T - 1$ for $SST$ (the number of observations minus 1, which is the number of parameters in a model containing only $\beta_1$).
• In the column labeled “Mean Square” are (i) the ratio of SSR to its degrees of freedom, SSR/1, and (ii) the ratio of SSE to its degrees of freedom, SSE/(T−2) = \(\hat{\sigma}^2\).

• The “mean square error” is our unbiased estimate of the error variance, which we first developed in Chapter 4.5.

• One widespread use of the information in the Analysis of Variance table is to define a measure of the proportion of variation in \(y\) explained by \(x\) within the regression model:

\[
R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}
\]

(6.1.7)

• The measure \(R^2\) is called the coefficient of determination. The closer \(R^2\) is to one, the better the job we have done in explaining the variation in \(y_t\) with \(\hat{y}_t = b_1 + b_2x_t\); and the greater is the predictive ability of our model over all the sample observations.
• 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 percentage of the variation in $y$ about its mean that is explained by the regression model.”

**Remark:** $R^2$ is a *descriptive* measure. By itself it does not measure the *quality* of the regression model. It is *not* the objective of regression analysis to find the model with the highest $R^2$. Following a regression strategy focused solely on maximizing $R^2$ is not a good idea.
6.1.1 Analysis of Variance Table and $R^2$ for Food Expenditure Example

The analysis of variance table for the food expenditure example appears in Table 6.2. From this table, we find that

$$SST = \sum (y_i - \bar{y})^2 = 79523$$

$$SSR = \sum (\hat{y}_i - \bar{y})^2 = 25221$$

$$SSE = \sum e_i^2 = 54311$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 0.317$$

$$SSE / (T - 2) = \sigma^2 = 1429.2455$$
### Table 6.2 Analysis of Variance Table

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explained</td>
<td>1</td>
<td>25221.2229</td>
<td>25221.2229</td>
</tr>
<tr>
<td>Unexplained</td>
<td>38</td>
<td>54311.3314</td>
<td>1429.2455</td>
</tr>
<tr>
<td>Total</td>
<td>39</td>
<td>79532.5544</td>
<td></td>
</tr>
</tbody>
</table>

- R-square = 0.3171

- The value $R^2 = 0.317$ says that about 32 percent of the variation in food expenditure about its mean is explained by variations in income. Alternatively, we can say that the regression model explains 32 percent of the variation in food expenditure about its mean, leaving 68 percent of the variation unexplained.
• Although this \( R^2 \) value sounds low, it is typical in regression studies using cross-sectional data, in which a sample of individuals, or other economic units, are observed at the point in time. Studies using time-series data, in which one individual is observed over time, usually have much higher \( R^2 \) value.

• The lesson here is that the success of a model cannot be completely judged on the magnitude of its \( R^2 \). Even if this number is low, the estimated parameters may contain useful information. Attempting to summarize the entire worth of a model by this one number is an error that should be avoided.

6.1.2 Correlation Analysis

The correlation coefficient \( \rho \) between \( X \) and \( Y \) is defined in Equation (2.5.4) to be
\[ \rho = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X) \text{var}(Y)}} \] (6.1.8)

- Given a sample of data pairs \((x_t, y_t), t = 1,\ldots,T\), the sample correlation coefficient is obtained by replacing the covariance and variances in Equation (6.1.8) by their sample analogues:

\[ r = \frac{\hat{\text{cov}}(X,Y)}{\sqrt{\hat{\text{var}}(X) \text{var}(Y)}} \] (6.1.9)

where

\[
\hat{\text{cov}}(X,Y) = \sum_{t=1}^{T} (x_t - \bar{x})(y_t - \bar{y})/(T-1)
\] (6.1.10a)
\[ \text{vâr}(X) = \sum_{t=1}^{T} (x_t - \bar{x})^2 / (T - 1) \]  

(6.1.10b)

The sample variance of \( Y \) is defined like \( \text{vâr}(X) \). Thus, the sample correlation coefficient \( r \) can be written as

\[ r = \frac{\sum_{t=1}^{T} (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{T} (x_t - \bar{x})^2 \sum_{t=1}^{T} (y_t - \bar{y})^2}} \]  

(6.1.11)

• The sample correlation coefficient \( r \) has a value between \(-1\) and \(1\), and it measures the strength of the linear association between observed values of \( X \) and \( Y \).
6.1.3 Correlation Analysis and $R^2$

- There are two interesting relationships between $R^2$ and $r$ in the simple linear regression model.

1. The first is that $r^2 = R^2$. That is, the square of the sample correlation coefficient between the sample data values $x_t$ and $y_t$ is algebraically equal to $R^2$. Intuitively, this relationship makes sense: $r^2$ falls between 0 and 1 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. $R^2$ can also be computed as the square of the sample correlation coefficient between $y_t$ and $\hat{y}_t = b_1 + b_2 x_t$. 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.”
6.2 Reporting the Results of a Regression Analysis

One way to summarize the regression results is in the form of a “fitted” regression equation:

\[ \hat{y}_t = 40.7676 + 0.1283x_i \quad R^2 = 0.317 \]  
\[ \text{(s.e.)} \quad (22.1387)(0.0305) \]  

- The value \( b_1 = 40.7676 \) estimates the weekly food expenditure by a household with no income; \( b_2 = 0.1283 \) implies that given a $1 increase in weekly income we expect expenditure on food to increase by $.13; or, in more reasonable units of measurement, if income increases by $100 we expect food expenditure to rise by $12.83.
- The \( R^2 = 0.317 \) says that about 32% of the variation in food expenditure about its mean is explained by variations in income.
• The numbers in parentheses underneath the estimated coefficients are the *standard errors* of the least squares estimates. Apart from critical values from the *t*-distribution, (R6.6) contains all the information that is required to construct interval estimates for $\beta_1$ or $\beta_2$ or to test hypotheses about $\beta_1$ or $\beta_2$. Another conventional way to report results is to replace the standard errors with the *t*-values, given in the computer output. These values arise when testing $H_0$: $\beta_1 = 0$ against $H_1$: $\beta_1 \neq 0$ and $H_0$: $\beta_2 = 0$ against $H_1$: $\beta_2 \neq 0$. Using these *t*-values we can report the regression results as

$$\hat{y}_i = 40.7676 + 0.1283x_i \quad R^2 = 0.317$$

$(t) \quad (1.84) \quad (4.20)$ (6.2.2)
When reporting the results this way, we recognize that the null hypothesis $H_0: \beta_2 = 0$ is an important one, since, if $b_2$ is not statistically significantly different from 0, then we cannot conclude that $x$ influences $y$.

6.2.1 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, suppose we are interested in the variable $x =$ U.S. total real disposable personal income. In 1999 the value of $x =$ $93,491,400,000,000$.

- As written the number is very cumbersome. We might divide the variable $x$ by 1 trillion and use instead the scaled variable $x^* = x/1,000,000,000,000 = $93.4914 trillion dollars.
• What, if any, are the effects of scaling the variables in a regression model? Consider the food expenditure model. We interpret the least squares estimate \( b_2 = 0.1283 \) as the expected increase in food expenditure, in dollars, given a $1 increase in weekly income.

• It may be more convenient to discuss increases in weekly income of $100. Such a change in the units of measurement is called *scaling the data*. The choice of the scale is made by the investigator so as 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 summarize the possibilities:
  1. Changing the scale of \( x \): Consider the estimated food expenditure equation
\[
\hat{y}_t = 40.77 + 0.1283 x_t
\]

\[
= 40.77 + (100 \times 0.1283) \left( \frac{x_t}{100} \right) 
\]

\[
= 40.77 + 12.83 x_t^*
\]  

(R6.8)

In the food expenditure model \( b_2 = 0.1283 \) measures the effect of a change in income of $1 while \( 100b_2 = 12.83 \) measures the effect of a change in income of $100. When the scale of \( x \) is altered the only other change occurs in the standard error of the regression coefficient, which 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 \( y \): For example, if food expenditure is measured in cents rather than dollars, we multiply \( y_t \) by 100 to get
100\(\hat{y}_i\) = (100 \times 40.77) + (100 \times 0.1283)x_i
\[
\hat{y}_i^* = 4077 + 12.83x_i
\] (R6.9)

In this rescaled model \(\beta_2^*\) measures the change we expect in \(y^*\) given a one-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.
6.3 Choosing a Functional Form

• In the household food expenditure function the dependent variable, household food expenditure, has been assumed to be a linear function of household income. That is, we represented the economic relationship as \( E(y_t) = \beta_1 + \beta_2 x_t \), which implies that there is a linear, straight-line relationship between \( E(y) \) and \( x \). The econometric model that corresponds to this economic model is

\[
y_t = \beta_1 + \beta_2 x_t + e_t \tag{6.3.1}
\]

• What if the relationship between \( y_t \) and \( x_t \) is not linear? Fortunately, all we have done is not lost. One of the important features of the simple linear regression model is that it is much more flexible than it appears at first glance.
Remark: The term *linear* in “simple linear regression model” means not a linear relationship between the variables, but a model in which the *parameters* enter in a linear way. That is, the model is “linear in the parameters,” but it is not, necessarily, “linear in the *variables.*”

- By “linear in the parameters” we mean that the *parameters* are *not* multiplied together, divided, squared, cubed, etc.
- The variables, however, can be *transformed* in any convenient way, *as long as the resulting model satisfies assumptions SR1-SR5 of the simple linear regression model.*
- The motivation for this discussion is that economic theory does not always imply that there is a linear relationship between the variables.
- In the food expenditure model we do *not* expect that as household income rises that 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.

![Graph showing a nonlinear relationship between food expenditure and income.](Image)

**Figure 6.2** A Nonlinear Relationship between Food Expenditure and Income

- If we believe that the relationship between $E(y)$ and $x$ looks like Figure 6.2, then specifying a linear relationship between the variables may not produce a satisfactory approximation.
6.3.1 Some Commonly Used Functional Forms

- Choosing an algebraic form for the relationship means choosing *transformations* of the original variables. The variable transformations that we begin with are:

  1. The natural logarithm: If $x$ is a variable then its natural logarithm is $\ln(x)$.
  2. The reciprocal: If $x$ is a variable then its reciprocal is $1/x$.

- In Table 6.3 we provide six commonly used statistical models that employ the original variables, $y$ and $x$, their logarithmic transformations, their reciprocal transformations, or some combination.
- In Figure 6.3 we illustrate the shapes that these models (without the random errors $e_i$) can take.
- Let us examine each of the functional forms in Table 6.3, the shapes they can take, and some economic implications of their use.
<table>
<thead>
<tr>
<th>Type</th>
<th>Statistical Model</th>
<th>Slope</th>
<th>Elasticity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Linear</td>
<td>$y_t = \beta_1 + \beta_2 x_t + e_t$</td>
<td>$\beta_2$</td>
<td>$\beta_2 \frac{x_t}{y_t}$</td>
</tr>
<tr>
<td>2. Reciprocal</td>
<td>$y_t = \beta_1 + \beta_2 \frac{1}{x_t} + e_t$</td>
<td>$-\beta_2 \frac{1}{x_t^2}$</td>
<td>$-\beta_2 \frac{1}{x_t y_t}$</td>
</tr>
<tr>
<td>3. Log-Log</td>
<td>$\ln(y_t) = \beta_1 + \beta_2 \ln(x_t) + e_t$</td>
<td>$\beta_2 \frac{y_t}{x_t}$</td>
<td>$\beta_2$</td>
</tr>
<tr>
<td>4. Log-Linear</td>
<td>$\ln(y_t) = \beta_1 + \beta_2 x_t + e_t$</td>
<td>$\beta_2 y_t$</td>
<td>$\beta_2 x_t$</td>
</tr>
<tr>
<td>(Exponential)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Linear-Log</td>
<td>$y_t = \beta_1 + \beta_2 \ln(x_t) + e_t$</td>
<td>$\beta_2 \frac{1}{x_t}$</td>
<td>$\beta_2 \frac{1}{y_t}$</td>
</tr>
<tr>
<td>(Semi-log)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Log-inverse</td>
<td>$\ln(y_t) = \beta_1 - \beta_2 \frac{1}{x_t} + e_t$</td>
<td>$\beta_2 \frac{y_t}{x_t^2}$</td>
<td>$\beta_2 \frac{1}{x_t}$</td>
</tr>
</tbody>
</table>
1. The model that is *linear in the variables* describes fitting a straight line to the original data, with slope $\beta_2$ and point elasticity $\beta_2 x_t / y_t$. The slope of the relationship is constant but the elasticity changes at each point.

2. The reciprocal model takes shapes shown in Figure 6.3(a). As $x$ increases $y$ approaches the intercept, its asymptote, from above or below depending on the sign of $\beta_2$. The slope of this curve changes, and flattens out, as $x$ increases. The elasticity also changes at each point and is opposite in sign to $\beta_2$. In Figure 6.3(a), when $\beta_2 > 0$, the relationship between $x$ and $y$ is an inverse one and the elasticity is negative: a 1% increase in $x$ leads to a reduction in $y$ of $-\beta_2 / (x_t y_t)\%$.

3. The log-log model is a very popular one. 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. The shapes that this equation can take are shown in Figures 6.3(b) and 6.3(c). Figure 6.3(b) shows cases in which $\beta_2 > 0$, and Figure 6.3(c) shows cases when $\beta_2 < 0$. The slopes of these curves change at every
point, but the elasticity is constant and equal to $\beta_2$. This constant elasticity model is very convenient for economists, since we like to talk about elasticities and are familiar with their meaning.

4. The log-linear model (“log” on the left-hand-side of the equation and “linear” on the right) can take the shapes shown in Figure 6.3(d). Both its slope and elasticity change at each point and are the same sign as $\beta_2$.

5. The linear-log model has shapes shown in Figure 6.3(e). It is an increasing or decreasing function depending upon the sign of $\beta_2$.

6. The log-inverse model (“log” on the left-hand-side of the equation and a reciprocal on the right) has a shape shown in Figure 6.3(f). It has the characteristic that near the origin it increases at an increasing rate (convex) and then, after a point, increases at a decreasing rate (concave).
Remark: Given this array of models, some of which have similar shapes, what are some guidelines for choosing a functional form? We must certainly choose a functional form that is sufficiently flexible to “fit” the data. Choosing a satisfactory functional form helps preserve the model assumptions. That is, a major objective of choosing a functional form, or transforming the variables, is to create a model in which the error term has the following properties:

1. $E(e_t) = 0$
2. $\text{var}(e_t) = \sigma^2$
3. $\text{cov}(e_i, e_j) = 0$
4. $e_t \sim N(0, \sigma^2)$

If these assumptions hold, then the least squares estimators have good statistical properties and we can use the procedures for statistical inference that we have developed in Chapters 4 and 5.
6.3.2 Examples Using Alternative Functional Forms

In this section we will examine an array of economic examples and possible choices for the functional form.

6.3.2a The Food Expenditure Model

- From the array of shapes in Figure 6.3 two possible choices that are similar in some aspects to Figure 6.2 are the reciprocal model and the linear-log model.
- The reciprocal model is

\[
y_t = \beta_1 + \beta_2 \frac{1}{x_t} + e_t
\]  

(6.3.2)
• For the food expenditure model we might assume that $\beta_1 > 0$ and $\beta_2 < 0$. If this is the case, then as income increases, household consumption of food increases at a decreasing rate and reaches an upper bound $\beta_1$.
• This model is *linear in the parameters* but it is *nonlinear in the variables*. If the error term $e_t$ satisfies our usual assumptions, then the unknown parameters can be estimated by least squares, and inferences can be made in the usual way.
• Another property of the reciprocal model, ignoring the error term, is that when $x < -\beta_2/\beta_1$ the model predicts expenditure on food to be negative. This is unrealistic and implies that this functional form is inappropriate for small values of $x$.
• When choosing a functional form one practical guideline is to consider how the dependent variable changes with the independent variable. In the reciprocal model the slope of the relationship between $y$ and $x$ is
If the parameter $\beta_2 < 0$ then there is a positive relationship between food expenditure and income, and, as income increases this “marginal propensity to spend on food” diminishes, as economic theory predicts.

- For the food expenditure relationship an alternative to the reciprocal model is the linear-log model

$$y_t = \beta_1 + \beta_2 \ln(x_t) + e_t$$

(6.3.6)

which is shown in Figure 6.3(e).

- For $\beta_2 > 0$ this function is increasing, but at a decreasing rate. As $x$ increases the slope $\beta_2/x_t$ decreases. Similarly, the greater the amount of food expenditure $y$ the smaller the
elasticity, $\beta_2/y_t$. 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.

6.3.2b Some Other Economic Models and Functional Forms

1. **Demand Models:** statistical models of the relationship between quantity demanded ($y^d$) and price ($x$) are very frequently taken to be linear in the variables, creating a linear demand curve, as so often depicted in textbooks. Alternatively, the “log-log” form of the model, $\ln(y^d_t) = \beta_1 + \beta_2 \ln(x_t) + e_t$, is very convenient in this situation because of its “constant elasticity” property. Consider Figure 6.3(c) where several log-log models are shown for several values of $\beta_2 < 0$. They are negatively sloped, as is appropriate for demand curves, and the price-elasticity of demand is the constant $\beta_2$. 
2. **Supply Models:** if \( y^s \) is the quantity supplied, then its relationship to price is often assumed to be linear, creating a linear supply curve. Alternatively the log-log, constant elasticity form, \( \ln(y^s_t) = \beta_1 + \beta_2 \ln(x_t) + e_t \), can be used.

3. **Production Functions:** another basic economic relationship that can be investigated using the simple linear statistical model is a production function, or total product function, that relates the output of a good produced to the amount of a variable input, such as labor. One of the assumptions of production theory is that diminishing returns hold; the marginal-physical product of the variable input declines as more is used. To permit a decreasing marginal product, the relation between output (\( y \)) and input (\( x \)) is often modeled as a “log-log” model, with \( \beta_2 < 1 \). This relationship is shown in Figure 6.3(b). It has the property that the marginal product, which is the slope of the total product curve, is diminishing, as required.

4. **Cost Functions:** a family of cost curves, which can be estimated using the simple linear regression model, is based on a “quadratic” total cost curve.
Suppose that you wish to estimate the total cost \( (y) \) of producing output \( (x) \); then a potential model is given by

\[
y_t = \beta_1 + \beta_2 x_t^2 + e_t
\]  
(6.3.4)

If we wish to estimate the average cost \( (y/x) \) of producing output \( x \) then we might divide both sides of Equation (6.3.4) by \( x \) and use

\[
(y_t / x_t) = \beta_1 / x_t + \beta_2 x_t + e_t / x_t
\]  
(6.3.5)

which is consistent with the quadratic total cost curve.

5. **The Phillips Curve:** it conjectures a systematic relationship between changes in the wage rate and changes in the level of unemployment. If we let \( w_t \) be the wage rate in time \( t \), then the percentage change in the wage rate is
\[
\% \Delta w_t = \frac{w_t - w_{t-1}}{w_{t-1}} \quad (6.3.6)
\]

If we assume that \( \% \Delta w_t \) is proportional to the excess demand for labor \( d_t \), we may write

\[
\% \Delta w_t = \gamma d_t \quad (6.3.7)
\]

where \( \gamma \) is an economic parameter. Since the unemployment rate \( u_t \) is inversely related to the excess demand for labor, we could write this using a reciprocal function as

\[
d_t = \alpha + \eta (1/u_t) \quad (6.3.8)
\]
where \( \alpha \) and \( \eta \) are economic parameters. Given Equation (6.3.7) we can substitute for \( d_t \), and rearrange, to obtain

\[
\%\Delta w_t = \gamma \left( \alpha + \eta \frac{1}{u_t} \right)
\]

\[
= \gamma \alpha + \gamma \eta \frac{1}{u_t}
\]

This model is *nonlinear in the parameters* and *nonlinear in the variables*. However, if we represent \( y_t = \%\Delta w_t \) and \( x_t = 1/u_t \), and \( \gamma \alpha = \beta_1 \) and \( \gamma \eta = \beta_2 \), then the simple linear regression model \( y_t = \beta_1 + \beta_2 x_t + e_t \) represents the relationship between the rate of change in the wage rate and the unemployment rate.
6.3.3 Choosing a Functional Form: Empirical Issues

One important question is that how does one choose the best transformation for $y$ and $x$, and hence the best functional form for describing the relationship between $y$ and $x$? Unfortunately, there are no hard and fast rules that will work for all situations. One important principle is that we should choose a function such that the properties of the simple regression model hold.

- Figure 6.4 describes a plot of average wheat yield (in tones per hectare) 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.
- Notice in Figure 6.4 that wheat yield fluctuates quite a bit, but, overall, it tends to increase over time, and the increase is at an increasing rate, particularly towards the end of the time period. An increase in yield is expected because of technological improvements.
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.

Let $y = \text{yield}$ and let $x = \text{time}$, measured from 1 to 48. One problem with the linear equation

$$y_t = \beta_1 + \beta_2 x_t + e_t$$

(6.3.9)

is that it implies that yield increases at the same constant rate $\beta_2$, when, from Figure 6.4, we expect this rate to be increasing.

The least squares estimated equation is

$$\hat{y}_t = 0.638 + 0.0210x_t \quad R^2 = 0.649$$

$$\text{(0.064) (0.0022)} \quad \text{(s.e.)}$$
The predicted values from this regression ($\hat{y}_t$), along with the actual values for yield, are displayed in the upper part of Figure 6.5. The lower part of the graph displays the residuals, centered around zero.

- There is a concentration of positive residuals at each end of the sample and a concentration of negative residuals in the middle. The bar chart in Figure 6.6 makes these concentrations even more apparent. They are caused by the inability of the straight line to capture the fact that yield is increasing at an increasing rate.

- Consider the alternative functional form using $x^3$ as follows:

$$y_t = \beta_1 + \beta_2(x_t)^3 + e_t \quad \text{(6.3.10)}$$

The slope of this function is $dy/dx = 3\beta_2(x_t)^2$. So, providing the estimate of $\beta_2$ turns out to be positive, the function will be increasing, and it will be increasing at the same rate as $x_t^2$ increases.
• We define $z_i^3 = x_i^3 / 1,000,000$. Then the estimated version of Equation (6.3.10) is

$$\hat{y}_i = 0.874 + 9.68z_i^3 \quad R^2 = 0.751$$

(0.036) (0.824) (s.e.)

• The fitted, actual, and residual values from this equation appear in Figure 6.7. Notice how the predicted (fitted) values are now increasing at an increasing rate. Also, the predominance of positive residuals at the ends and negative residuals in middle no longer exists. Furthermore, the $R^2$ value has increased from 0.649 to 0.751, indicating that the equation with $x_i^3$ fits the data better than the one with just $x$.

• Both these equations have the same dependent variable ($y_i$), and the same number of explanatory variables (only 1). In these circumstances the $R^2$ can be used legitimately to compare goodness-of-fit. If the dependent variables are different or the numbers of
explanatory variables in each equation are different, a direct comparison using $R^2$ is not possible.

- 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. 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.
6.4 Are the Residuals Normally Distributed?

- Recall that hypothesis tests and interval estimates for the coefficients relied on the assumption that the errors, and hence the dependent variable $y$, are normally distributed. Thus, when choosing a functional form, it is desirable to create a model in which the errors are normally distributed.

- Most computer software will create a histogram of the residuals and also give statistics that can be used to formally test a null hypothesis that the residuals come from a normal distribution. The relevant EViews output for the food expenditure example appears in Figure 6.8.

- What does Figure 6.8 tell us? First, notice that it is centered around zero, the mean of the least squares residuals. Second, it seems, on the whole, to represent an approximate normal distribution, but the low count in the cell on the right of zero, and
the empty cells near 40, may make us wonder if a formal test would reject a hypothesis of normality. A convenient formal test is one that is attributable to Jarque and Bera and has been named after them.

- The Jarque-Bera test for normality 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 residual is 0.397. Kurtosis refers to the “peakedness” of the distribution. For a normal distribution the kurtosis value is 3. From Figure 6.8, we see that the food expenditure residuals have a kurtosis of 2.87.

- So, the question we have to ask is whether 2.87 is sufficiently different from 3, and 0.397 sufficiently different from zero, to conclude the residuals are not normally distributed. The Jarque-Bera statistic is given by

\[ JB = \frac{T}{6} \left( S^2 + \left( \frac{k-3}{4} \right)^2 \right) \]
where $S$ 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 big value of the Jarque-Bera statistic.

- When the residuals are normally distributed, the Jarque-Bera statistic has a chi-squared distribution with 2 degrees of freedom. Thus, 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 2 degrees of freedom.

- Applying these ideas to the food expenditure example, we have

$$JB = \frac{40}{6} \left( 0.3969^2 + \frac{(2.874 - 3)^2}{4} \right) = 1.077$$

The 5% critical value from a $\chi^2$ distribution with 2 degrees of freedom is 5.99. Because $1.077 < 5.99$, there is insufficient evidence from the residuals to conclude that
the normal distribution assumption is unreasonable. The same conclusion could have
been reached by examining the $p$-value. The $p$-value appears in the EViews output
described as “Probability.” Thus, we also fail to reject the null hypothesis on the
grounds that $0.584 > 0.05$. 
Exercise

<table>
<thead>
<tr>
<th></th>
<th>6.1</th>
<th>6.2</th>
<th>6.3</th>
<th>6.4</th>
<th>6.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.8</td>
<td>6.12</td>
<td>6.13</td>
<td>6.14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>