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

$$y_t = \beta_1 + \beta_2 x_t + e_t \tag{6.1.1}$$

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 (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.

$$y_t = E(y_t) + e_t \qquad (6.1.2)$$

- $E(y_t) = \beta_1 + \beta_2 x_t$ 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 β_1 and β_2 and decompose the value of y_t into

$$y_t = \text{涔}y_t + e_t \qquad (6.1.3)$$

where $\text{涔}y_t = b_1 + b_2 x_t$ and $e_t = y_t - \text{?}y_t$.

[Figure 6.1 goes here]

- Subtract the sample mean $\bar{y}$ from both sides of the equation to obtain

$$y_t - \bar{y} = (y\text{洚}_t - \bar{y}) + e_t \qquad (6.1.4)$$

- 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, $\hat{e}_t$.
- A measure of the “total variation” y is to square the differences between y_t and its mean value $\bar{y}$ and sum over the entire sample.

$$\sum (y_t - \overline{y})^2 = \sum [(\hat{y}_t - \overline{y}) + e_t]^2$$

$$= \sum (\hat{y}_t - \overline{y})^2 + \sum e_t^2 + 2\sum (\hat{y}_t - \overline{y})e_t \qquad (6.1.5)$$

$$= \sum (\hat{y}_t - \overline{y})^2 + \sum e_t^2$$

- The cross-product term $\sum (\hat{y}_t - \overline{y})e_t = 0$ and drops out.

1. $\sum (y_t - \overline{y})^2$ = total sum of squares = *SST*: a measure of *total variation* in y about its sample mean.
2. $\sum (\hat{y}_t - \overline{y})^2$ = explained sum of squares = *SSR*: that part of total variation in *y* about its sample mean that is explained by the regression.
3. $\sum \hat{e}_t^2$ = error sum of squares = *SSE*: that part of total variation in *y* about its mean that is not explained by the regression.

Thus,

$$SST = SSR + SSE \tag{6.1.6}$$

- This decomposition is usually presented in what is called an “Analysis of Variance” table with general format

Table 6.1 Analysis of Variance Table

Source of		Sum of	Mean
Variation	DF	Squares	Square
Explained	1	SSR	$SSR/1$
Unexplained	$T–2$	SSE	$SSE/(T–2)$ $[=\hat{\sigma}^2]$
Total	$T–1$	SST	

- 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);
3. *df* = T–1 for *SST* (the number of observations minus 1, which is the number of parameters in a model containing only β_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.
- 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:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{6.1.7}$$

- 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 $\hat{y}_t = b_1 + b_2 x_t$; 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 computer output usually contains the Analysis of Variance, Table 6.1. For the food expenditure data it is:

Table 6.3 Analysis of Variance Table

Source	DF	Sum of Squares	Mean Square
Explained	1	25221.2229	25221.2229
Unexplained	38	54311.3314	1429.2455
Total	39	79532.5544	
		R-square	0.3171

From this table we find that:

$$SST = \sum (y_t - \overline{y})^2 \quad = 79532.$$

$$SSR = \sum (\hat{y}_t - \overline{y})^2 \quad = 25221.$$

$$SSE = \sum \hat{e}_t^2 \quad = 54311.$$

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

$$SSE/(T-2) = \hat{\sigma}^2 = 1429.2455$$

6.1.2 Correlation Analysis

The correlation coefficient ρ between *X* and *Y* is

$$\rho = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}} \qquad (6.1.8)$$

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

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

where

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

$$\hat{\text{var}}(X)=\sum_{t=1}^{T}(x_t-\bar{x})^2/(T-1) \qquad (6.1.10b)$$

- The sample variance of *Y* is defined like $\hat{\text{var}}(X)$.
- The sample correlation coefficient *r* is

$$r=\frac{\sum_{t=1}^{T}(x_t-\bar{x})(y_t-\bar{y})}{\sqrt{\sum_{t=1}^{T}(x_t-\bar{x})^2\sum_{t=1}^{T}(y_t-\bar{y})^2}} \qquad (6.1.11)$$

- 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
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:

$$\begin{array}{lll} \hat{y}_t & = 40.7676 + 0.1283 x_t & R^2 = 0.317 \\ \text{(s.e.)} & (22.1387)\,(0.0305) & \end{array} \qquad \text{(R6.6)}$$

- 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 β_1 or β_2 or to test hypotheses about β_1 or $\beta_{2.}$
- Another conventional way to report results is to replace the standard errors with the "*t*-values"
- 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

$$\begin{array}{lll} \hat{y}_t = 40.7676 + 0.1283x_t & & R^2 = 0.317 \\ (t) \quad (1.84) \qquad (4.20) & & \end{array} \tag{6.2.2}$$

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.
- 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.
- 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*:

$$\begin{aligned}\hat{y}_t &= 40.77 + 0.1283x_t \\ &= 40.77 + (100 \times 0.1283)\left(\frac{x_t}{100}\right) \\ &= 40.77 + 12.83x_t^*\end{aligned} \tag{R6.8}$$

- 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, 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*:

$$\begin{aligned} 100\hat{y}_t &= \left(100 \times 40.77\right) + \left(100 \times 0.1283\right)\left(\frac{x_t}{100}\right) \\ \hat{y}_t^* &= 4077 + 12.83 x_t \end{aligned} \qquad \text{(R6.9)}$$

- 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.

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.
- What if the relationship between y_t and x_t is not linear?

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.*
- 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.

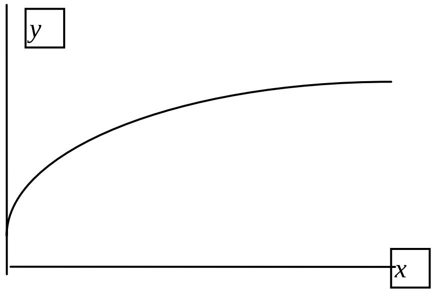

Figure 6.2 A Nonlinear Relationship between Food Expenditure and Income

6.3.1 Some Commonly Used Functional Forms

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 $1/x$.

Type	Statistical Model	Slope	Elasticity
1. Linear	$y_t = \beta_1 + \beta_2 x_t + e_t$	β_2	$\beta_2 \frac{x_t}{y_t}$
2. Reciprocal	$y_t = \beta_1 + \beta_2 \frac{1}{x_t} + e_t$	$-\beta_2 \frac{1}{x_t^2}$	$-\beta_2 \frac{1}{x_t y_t}$
3. Log-Log	$\ln(y_t) = \beta_1 + \beta_2 \ln(x_t) + e_t$	$\beta_2 \frac{y_t}{x_t}$	β_2
4. Log-Linear (Exponential)	$\ln(y_t) = \beta_1 + \beta_2 x_t + e_t$	$\beta_2 y_t$	$\beta_2 x_t$
5. Linear-Log (Semi-log)	$y_t = \beta_1 + \beta_2 \ln(x_t) + e_t$	$\beta_2 \frac{1}{x_t}$	$\beta_2 \frac{1}{y_t}$
6. Log-inverse	$\ln(y_t) = \beta_1 - \beta_2 \frac{1}{x_t} + e_t$	$\beta_2 \frac{y_t}{x_t^2}$	$\beta_2 \frac{1}{x_t}$

[Figure 6.3 goes here]

1. The model that is *linear in the variables* describes fitting a straight line to the original data, with slope β_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 β_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 β_2. In Figure 6.3(a), when β_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 β_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*** β_2. This <u>constant</u> <u>elasticity</u> model is very convenient for economists, since we like to talk about elasticites 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 β_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 β_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. $\mathrm{var}(e_t)=\sigma^2$
3. $\mathrm{cov}(e_i,e_j)=0$
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

$$y_t = \beta_1 + \beta_2 \frac{1}{x_t} + e_t \qquad (6.3.2)$$

- 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 β_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

$$\frac{dy}{dx} = -\beta_2 \frac{1}{x_t^2}$$

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

$$y_t = \beta_1 + \beta_2 \ln(x_t) + e_t \qquad (6.3.3)$$

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 β_2/x_t decreases.
- Similarly, the greater the amount of food expenditure y the smaller the elasticity, β_2/y_t.

6.3.2b Some Other Economic Models and Functional Forms

1. Demand Models: 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, $\ln(y_t^d) = \beta_1 + \beta_2 \ln(x_t) + e_t$, 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 β_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, $\ln(y_t^s) = \beta_1 + \beta_2 \ln(x_t) + e_t$, can be used

3. Production Functions: 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

$$y_t = \beta_1 + \beta_2 x_t^2 + e_t \tag{6.3.4}$$

  - 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

$$(y_t / x_t) = \beta_1 / x_t + \beta_2 x_t + e_t / x_t \qquad (6.3.5)$$

which is consistent with the quadratic total cost curve.

5. The Phillips Curve: If we let w_t be the wage rate in time t, then the percentage change in the wage rate is

$$\%\Delta w_t = \frac{w_t - w_{t-1}}{w_{t-1}} \qquad (6.3.6)$$

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

$$\%\Delta w_t = \gamma d_t \qquad (6.3.7)$$

where γ 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

$$d_t = \alpha + \eta \frac{1}{u_t} \tag{6.3.8}$$

where α and η 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*.