Chapter 3

The Simple Linear Regression Model: Specification and Estimation

3.1 An Economic Model

Suppose that we are interested in studying the relationship between household income and expenditure on food. We ask the question, “How much did your household spend on food last week?” Suppose that the continuous random variable y has a probability density function, \( f(y) \), that describes the probabilities of obtaining various food expenditure values. The probability density function in this case describes how expenditures are “distributed” over the population, and it might look like one of those in Figure 3.1.
The probability distribution in Figure 3.1a is actually a conditional probability density function, since it is “conditional” upon household income. If $x =$ weekly household income, then the conditional probability density function is $f(y|x = $480). The conditional mean, or expected value, of $y$ is $E(y|x = $480) = \mu_{y|x}$ and is our population’s average weekly food expenditure. The conditional variance of $y$ is $\text{var}(y|x = $480) = \sigma^2$, which measures the dispersion of household expenditures $y$ about their mean $\mu_{y|x}$. The parameters $\mu_{y|x}$ and $\sigma^2$, if they were known, would give us some valuable information about the population we are considering. If we knew these parameters, and if we knew that the conditional distribution $f(y|x = $480) was normal, $N(\mu_{y|x}, \sigma^2)$, then we could calculate probabilities that $y$ falls in specific intervals using properties of the normal distribution.

In order to investigate the relationship between expenditure and income, we must build an economic model and then an econometric model that forms the basis for a
quantitative economic analysis. If we consider households with different levels of income, we expect the average expenditure on food to change. In Figure 3.1b we show the probability density functions of food expenditure for two different levels of weekly income, $480 and $800. Each density function $f(y|x)$ shows that expenditures will be distributed about the mean value $\mu_{y|x}$, but the mean expenditure by households with higher income is larger than the mean expenditure by lower income household.

- Our economic model of household food expenditure, depicted in Figure 3.2, is described as the simple regression function

$$E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x \quad (3.1.1)$$

The conditional mean $E(y|x)$ in Equation (3.1.1) is called a **simple regression function**. The unknown regression parameters $\beta_1$ and $\beta_2$ are the intercept and slope of the regression function, respectively.
• In our food expenditure example, the intercept, $\beta_1$, represents the average weekly household expenditure on food by a household with no income, $x = 0$. The slope $\beta_2$ represents the change in $E(y|x)$ given a $1$ change in weekly income; it could be called the marginal propensity to spend on food. Algebraically,

$$
\beta_2 = \frac{\Delta E(y|x)}{\Delta x} = \frac{dE(y|x)}{dx} 
$$

(3.1.2)

where “$\Delta$” denotes “change in” and $dE(y|x)/dx$ denotes the “derivative” of $E(y|x)$ with respect to $x$. 
3.2 An Econometric Model

The model \( E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x \) that we specified in Section 3.1 describes economic behavior, but it is an abstraction from reality. If we were to sample household expenditures at other levels of income, we would expect the sample values to be symmetrically scattered around their mean value \( E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x \). In Figure 3.3 we arrange bell-shaped figures like Figure 3.1, depicting \( f(y|x) \), along the regression line for each level of income. This figure shows that at each level of income the average value of household expenditure is given by the regression function, \( E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x \). It also shows that we assume values of household expenditure on food will be distributed around the mean value \( E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x \) at each level of income. This regression function is the foundation of an econometric model for household food expenditure.
In order to make the econometric model complete, we have to make a few more assumptions. The standard assumption about the dispersion of values $y$ about their mean is the same for all levels of income, $x$. That is, $\text{var}(y|x) = \sigma^2$ for all values of $x$. The constant variance assumption, $\text{var}(y|x) = \sigma^2$ implies that at each level of income $x$ we are equally uncertain about how far values of food expenditure, $y$, may fall from their mean value, $E(y|x) = \mu_{y|x} = \beta_1 + \beta_2 x$, and the uncertainty does not depend on income or anything else. Data satisfying this condition are said to be homoskedastic. If this assumption is violated, so that $\text{var}(y|x) \neq \sigma^2$ for all values of income, $x$, the data are said to be heteroskedastic.

Next, we have described the sample as random. By this we mean that when data are collected, they are statistically independent. If $y_i$ and $y_j$ denote the expenditures of two randomly selected households, then knowing the value of one of these (random) variables tells us nothing about what value the other might take. If $y_i$ and $y_j$ are the expenditures of two randomly selected households, then we will assume that their covariance is zero, or
\( \text{cov}(y_i, y_j) = 0 \). This is a weaker assumption than statistical independence (since independence implies zero covariance, but not vice versa); it implies only that there is no systematic linear association between \( y_i \) and \( y_j \). Refer to Chapter 2.5 for a discussion of this difference.

In order to carry out a regression analysis we must make an assumption about the values of the variable \( x \). The idea of regression analysis is to measure the effect of changes in one variable, \( x \), on another, \( y \). In order to do that, \( x \) must take several different values, at least two in this case, within the sample of data. For now we simply state the additional assumption that \( x \) must take at least two different values. Furthermore we assume that \( x \) is not random, which means that we can control the values of \( x \) at which we observe \( y \).

Finally, it is sometimes assumed that the distribution of \( y \), \( f(y|x) \), is known to be normal. It is reasonable, sometimes, to assume that an economic variable is normally distributed about its mean.
Assumptions of the Simple Linear Regression Model—I

• The average value of $y$, for each value of $x$, is given by the linear regression

$$E(y) = \beta_1 + \beta_2 x$$

• For each value of $x$, the values of $y$ are distributed about their mean value, following probability distributions that all have the same variance, i.e.,

$$\text{var}(y) = \sigma^2$$

• The values of $y$ are all uncorrelated, and have zero covariance, implying that there is no linear association among them.
\text{cov}(y_i, y_j) = 0 \text{ for all } i \text{ and } j

This assumption can be made stronger by assuming that the values of \( y \) are all \textit{statistically independent}.

- The variable \( x \) is not random and must take at least two different values.

- \textit{(optional)} The values of \( y \) are \textit{normally distributed} about their mean for each value of \( x \),

\[ y \sim N[(\beta_1 + \beta_2 x), \sigma^2] \]
3.2.1 Introducing the Error Term

The essence of regression analysis is that any observation on the dependent variable $y$ can be decomposed into two parts: a systematic component and a random component. The systematic component of $y$ is its mean $E(y)$, which itself is not random, since it is a mathematical expectation. The random component of $y$ is the difference between $y$ and its mean value $E(y)$. This is called a **random error term**, and it is defined as

$$e = y - E(y) = y - \beta_1 - \beta_2x$$

(3.2.1)

If we rearrange Equation (3.2.1) we obtain the simple linear regression model

$$y = \beta_1 + \beta_2x + e$$

(3.2.2)
The dependent variable $y$ is explained by a component that varies systematically with the independent variable $x$ and by the random error term $e$.

Equation (3.2.1) shows that $y$ and error term $e$ differ only by constant $E(y)$, and since $y$ is random, so is the error term $e$. Given what we have already assumed about $y$, the properties of the random error, $e$, can be derived directly from Equation (3.2.1). Since $y$ and $e$ differ only by a constant, $\beta_1 + \beta_2 x$, the probability density functions for $y$ and $e$ are identical except for their location, as shown in Figure 3.4.
Assumptions of the Simple Linear Regression Model—II

It is customary in econometrics to state the assumptions of the regression model in terms of the random error $e$. For future reference the assumptions are named SR1-SR6, “SR” denoting “simple regression.”

SR1. The value of $y$, for each value of $x$, is

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

SR2. The average value of the random error $e$ is

$$E(e) = 0$$

since we assume that

$$y = \beta_1 + \beta_2 x$$

SR3. The variance of the random error $e$ is

$$\text{var}(e) = \sigma^2 = \text{var}(y)$$

since $y$ and $e$ differ only by a constant, which doesn’t change the variance.

SR4. The covariance between any pair of random errors, $e_i$ and $e_j$ is

...
\[ \text{cov}(e_i, e_j) = \text{cov}(y_i, y_j) = 0 \]

If the values of \( y \) are *statistically independent*, then so are the random errors \( e \), and vice versa.

**SR5.** The variable \( x \) is not random and must take at least two different values.

**SR6.** (*optional*) The values of \( e \) are *normally distributed* about their mean

\[ e \sim N(0, \sigma^2) \]

if the values of \( y \) are normally distributed, and vice versa.

The random error \( e \) and the dependent variable \( y \) are both random variables, and the properties of one can be determined from the properties of the other. There is one interesting difference between these random variables, however, and that is that \( y \) is “observable” and \( e \) is “unobservable.” If the regression parameters \( \beta_1 \) and \( \beta_2 \) were *known*, then for any value of \( y \) we could calculate \( e = y - (\beta_1 + \beta_2x) \). This is illustrated in Figure 3.5. Knowing the regression function \( E(y|x) = \beta_1 + \beta_2x \), we could separate \( y \) into its fixed
and random parts. Unfortunately, since \( \beta_1 \) and \( \beta_2 \) are \textbf{never known}, it is \textbf{impossible} to calculate \( e \), and no bets on its true value can ever be collected.

The random error \( e \) represents all factors affecting \( y \) rather than \( x \). Moreover, the random error \( e \) also captures any approximation error that arises, because the linear functional form we have assumed may be only an approximation to reality.
3.3 Estimating the Parameters for the Expenditure Relationship

The economic and statistical models we developed in the previous section are the basis for using a sample of data to estimate the intercept and slope parameters, $\beta_1$ and $\beta_2$.

3.3.1 The Least Squares Principle

- The least squares principle asserts that to fit a line to the data values we should fit the line so that the sum of the squares of the vertical distances from each point to the line is as small as possible. The distances are squared to prevent large positive distances from being canceled by large negative distances. The intercept and slope of this line, the line that best fits the data using the least squares principle, are $b_1$ and $b_2$, the least squares estimates of $\beta_1$ and $\beta_2$. The fitted regression line is then
\[ \hat{y}_t = b_1 + b_2 x_t \] (3.3.1)

- The vertical distances from each point to the fitted line are the *least squares residual*. They are given by

\[ \varepsilon_t = y_t - \hat{y}_t = y_t - b_1 - b_2 x_t \] (3.3.2)

These residuals are depicted in Figure 3.7a.

- Now suppose we fit another line, *any other fitted line*, to the data, say

\[ \hat{y}_t^* = b_1^* + b_2^* x_t \] (3.3.3)
where $b_1^*$ and $b_2^*$ are any other intercept and slope values. The residuals for this line, $\varepsilon_t^* = y_t - y_t^*$, are shown in Figure 3.7b.

- The least squares estimates $b_1$ and $b_2$ have the property that the sum of their squared residuals is less than the sum of squared residuals for any other line,

$$\sum \varepsilon_t^2 = \sum (y_t - y_t)^2 \leq \sum \varepsilon_t^{*2} = \sum (y_t - y_t^*)^2$$

no matter how the other line might be drawn through the data. The least squares principle says that the estimates $b_1$ and $b_2$ of $\beta_1$ and $\beta_2$ are the ones to use, since the line using them as intercept and slope fits are the data best.

- Given the sample observations on $y$ and $x$, least squares estimates of the unknown parameters $\beta_1$ and $\beta_2$ are obtained by minimizing the sum of squares function
Since the points \((y_t, x_t)\) have been observed, the sum of squares function \(S\) is a function of the unknown parameters \(\beta_1\) and \(\beta_2\). This function, which is a quadratic in terms of the unknown parameters \(\beta_1\) and \(\beta_2\), is a “bowl-shaped surface” like the one depicted in Figure 3.8.

- Given the data, our task is to find, out of all the possible values that \(\beta_1\) and \(\beta_2\) can take, the point \((b_1, b_2)\) at which the sum of squares function \(S\) is minimum. The partial derivatives of \(S\) with respect to \(\beta_1\) and \(\beta_2\) are
\[ \frac{\partial S}{\partial \beta_1} = 2T \beta_1 - 2 \sum y_t + 2 \sum x_i \beta_2 \]  
\[ \frac{\partial S}{\partial \beta_2} = 2 \sum x_i^2 \beta_2 - 2 \sum x_t y_t + 2 \sum x_i \beta_1 \]  

(3.3.5)

- Algebraically, to obtain the point \((b_1, b_2)\) we set the derivatives of Equation (3.35) to zero, and replace \(\beta_1\) and \(\beta_2\) by \(b_1\) and \(b_2\), respectively, to obtain

\[ 2(\sum y_t - Tb_1 - \sum x_i b_2) = 0 \]  
\[ 2(\sum x_t y_t - \sum x_i b_1 - \sum x_i^2 b_2) = 0 \]  

(3.3.6)

- Rearranging Equation (3.3.6) leads to two equations usually known as the normal equations,
These two equations comprise a set of two linear equations in two unknowns $b_1$ and $b_2$.

- To solve for $b_2$, multiply the first Equation (3.3.7a) by $\sum x_t$; multiply the second Equation (3.3.7b) by $T$; subtract the second equation from the first and then isolate $b_2$ on the left-hand side. To solve for $b_1$, divide both sides of Equation (3.3.7a) by $T$.

The formulas for the least squares estimates $\beta_1$ and $\beta_2$ are

$$b_2 = \frac{T \sum x_t y_t - \sum x_t \sum y_t}{T \sum x_t^2 - (\sum x_t)^2} \quad (3.3.8a)$$

$$b_1 = \bar{y} - b_2 \bar{x} \quad (3.3.8b)$$
where \( \bar{y} = \sum y_i / T \) and \( \bar{x} = \sum x_i / T \) are the sample means of the observations on \( y \) and \( x \).

- If we plug the sample values \( y_i \) and \( x_i \) into Equation (3.3.8), then we obtain the least squares estimates of the intercept and slope parameters \( \beta_1 \) and \( \beta_2 \). When the formulas for \( b_1 \) and \( b_2 \) are taken to be rules that are used whatever the sample data turn out to be, then \( b_1 \) and \( b_2 \) are random variables. When actual sample values are substituted into the formulas, we obtain numbers that are the observed values of random variables. To distinguish these two cases we call the rules or general formulas for \( b_1 \) and \( b_2 \) the least squares estimators. We call the numbers obtained when the formulas are used with a particular sample least squares estimates.
3.3.2 Estimates for the Food Expenditure Function

- We have used the least squares principle to derive Equation (3.3.8), which can be used to obtain the least squares estimates for the intercept and slope parameters $\beta_1$ and $\beta_2$. To illustrate the use of these formulas, we will use them to calculate the values of $b_1$ and $b_2$ for the household expenditure data given in Table 3.1. From Equation (3.3.8a) we have

\[
b_2 = \frac{T \sum x_i y_i - \sum x_i \sum y_i}{T \sum x_i^2 - (\sum x_i)^2}
\]

(3.3.9a)

\[
= \frac{(40)(3834936.497) - (27920)(5212.520)}{(40)(21020623.02) - (27920)^2} = 0.1283
\]

and from Equation (3.3.8b)
\[ b_1 = \bar{y} - b_2 \bar{x} \]
\[ = 130.313 - (0.1282886)(698.0) \]
\[ = 40.7676 \]  

A convenient way to report the values for \( b_1 \) and \( b_2 \) is to write out the estimated or fitted regression line:

\[ y = b_0 + b_1 x \]  

This line is graphed in Figure 3.9. The line’s slope is 0.1283 and its intercept, where it crosses the vertical axis, is 40.7676. The least squares fitted line passes through the middle of the data in a very precise way, since one of the characteristics of the fitted line based on the least squares parameter estimates is that it passes through the point defined by the sample means, \((\bar{x}, \bar{y}) = (698.00, 130.31)\).
3.3.3 Interpreting the Estimates

- The value $b_2 = 0.1283$ is an estimate of $\beta_2$, the amount by which weekly expenditure on food increases when weekly income increases by $1$. Thus, we estimate that if income goes up by $100$, weekly expenditure on food will increase by approximately $12.83$.

- Strictly speaking, the intercept estimate $b_1 = 40.7676$ is an estimate of the weekly amount spent on food for a family with zero income. In most economic models we must be very careful when interpreting the estimated intercept. The problem is that we usually do not have any data points near $x = 0$, which is certainty true for the food expenditure data shown in Figure 3.9. If we have no observations in the region where income is near zero, then our estimated relationship may not be a good approximation to reality in that region. So, although our estimated model suggests that a household with zero income will spend $40.7676$ per week on food, it might be risky to take this
estimate literally. You should consider this issue in each economic model that you estimate.

3.3.3a Elasticities

- The income elasticity of demand is a useful way to characterize the responsiveness of consumer expenditure to changes in income. From microeconomic principles the elasticity of any variable \( y \) with respect to another variable \( x \) is

\[
\eta = \frac{\text{percentage change in } y}{\text{percentage change in } x} = \frac{\Delta y}{\Delta x} \cdot \frac{y}{x} = \frac{\Delta y}{\Delta x} \cdot \frac{x}{y}
\]  

(3.3.11)

- In the linear economic model given by Equation (3.1.1) we have shown that
\[ \beta_2 = \frac{\Delta E(y)}{\Delta x} \]  

(3.3.12)

so the elasticity of “average” expenditure with respect to income is

\[ \eta = \frac{\Delta E(y)/E(y)}{\Delta x/x} = \frac{\Delta E(y)}{\Delta x} \cdot \frac{x}{E(y)} = \beta_2 \cdot \frac{x}{E(y)} \]  

(3.3.13)

To estimate this elasticity we replace \( \beta_2 \) by \( b_2 = 0.1283 \). We must also replace “\( x \)” and “\( E(y) \)” by something, since in a linear model the elasticity is different on each point upon the regression line. One possibility is to choose a value of \( x \) and replace \( E(y) \) by its fitted value. Another frequently used alternative is to report the elasticity at the “point of the means” \((\bar{x}, \bar{y}) = (698.00, 130.31)\) since that is a representative point on the regression line. If we calculate the income elasticity at the point of the means, we obtain
We estimate that a 1% change in weekly household income will lead, on average, to approximately a 0.7% increase in weekly household expenditure on food when $(x, y) = (\bar{x}, \bar{y}) = (698.00, 130.31)$. Since the estimated income elasticity is less than one, we would classify food as a “necessity” rather than a “luxury,” which is consistent with what we would expect for an “average” household.

3.3.3b Prediction

- Suppose that we wanted to predict weekly food expenditure for a household with a weekly income of $750. This prediction is carried out by substituting $x = 750$ into our estimated equation to obtain

$$\hat{y} = b_2 \cdot \frac{\bar{x}}{\bar{y}} = 0.1283 \times \frac{698.00}{130.31} = 0.687$$

(3.3.14)
\[ \hat{y}_t = 40.7676 + 0.1283x_t = 40.7676 + 0.1283(750) = $130.98 \] (3.3.15)

We predict that a household with a weekly income of $750 will spend $130.98 per week on food.
### 3.3.3c Examining Computer Output

**Dependent Variable:** FOODEXP  
**Method:** Least Squares  
**Sample:** 1 40  
**Included observations:** 40

<table>
<thead>
<tr>
<th>Variable</th>
<th>Coefficient</th>
<th>Std. Error</th>
<th>t-Statistic</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>40.76756</td>
<td>22.13865</td>
<td>1.841465</td>
<td>0.0734</td>
</tr>
<tr>
<td>INCOME</td>
<td>0.128289</td>
<td>0.030539</td>
<td>4.200777</td>
<td>0.0002</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-squared</td>
<td>Mean dependent var</td>
<td>130.3130</td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>S.D. dependent var</td>
<td>45.15857</td>
</tr>
<tr>
<td>S.E. of regression</td>
<td>Akaike info criterion</td>
<td>10.15149</td>
</tr>
<tr>
<td>Sum squared resid</td>
<td>Schwarz criterion</td>
<td>10.23593</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>F-statistic</td>
<td>17.64653</td>
</tr>
<tr>
<td>Durbin-Watson stat</td>
<td>Prob(F-statistic)</td>
<td>0.000155</td>
</tr>
</tbody>
</table>

**Figure 3.10 EViews Regression Output**
Dependent Variable: FOODEXP

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>25221.22299</td>
<td>25221.22299</td>
<td>17.647</td>
<td>0.0002</td>
</tr>
<tr>
<td>Error</td>
<td>38</td>
<td>54311.33145</td>
<td>1429.24556</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>39</td>
<td>79532.55444</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 37.80536  R-square 0.3171
Dep Mean 130.31300  Adj R-sq 0.2991
C.V. 29.01120

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > |T| |
|----------|----|-------------------|----------------|------------------------|--------|
| INTERCEP | 1  | 40.767556         | 22.13865442    | 1.841                  | 0.0734 |
| INCOME   | 1  | 0.128289          | 0.03053925     | 4.201                  | 0.0002 |

Figure 3.11 SAS Regression Output
3.3.4 Other Economic Models

- If $y$ and $x$ are transformed in some way, then the economic interpretation of the parameters can change. A popular transformation in economics is the natural logarithm. Economic model like $\ln(y) = \beta_1 + \beta_2 \ln(x)$ are common. A nice feature of this model, if the assumptions of the regression model hold, is that the parameter $\beta_2$ is the elasticity of $y$ with respect to $x$. The derivative of $\ln(y)$ with respect to $x$ is

$$\frac{d[\ln(y)]}{dx} = \frac{1}{y} \cdot \frac{dy}{dx}$$

The derivative of $\beta_1 + \beta_2 \ln(x)$ with respect to $x$ is

$$\frac{d[\ln(y)]}{dx} = \frac{d[\beta_1 + \beta_2 \ln(x)]}{dx} = \frac{1}{x} \cdot \beta_2$$
Setting these two pieces equal to one another, and solving for $\beta_2$ gives

$$
\beta_2 = \frac{dy}{dx} \cdot \frac{x}{y} = \eta \tag{3.3.16}
$$

Equation (3.3.16) shows that in an economic model in which $\ln(y)$ and $\ln(x)$ are linearly related, the parameters $\beta_2$ is the point elasticity of $y$ with respect to $x$. 
<table>
<thead>
<tr>
<th>Exercise</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
</tr>
<tr>
<td>3.9</td>
</tr>
</tbody>
</table>