5. How Good Is the Least Squares Estimation Procedure?

Given the estimate b = 17.158 we are inclined to ask "How good an estimate is 17.158?" By that we mean how close is 17.158 to the true population mean, β. Unfortunately this is an ill-posed question, in the sense that it can never be answered. In order to answer it, we would have to know β, in which case we would never had been trying to estimate it in the first place! There is no such thing as a good estimate.

What we must do instead of asking about the estimate is to ask about the estimation procedure. How good is the least squares estimation procedure? This is a question we can answer. To distinguish the estimation procedure from the estimate ($\bar{y}$ =17.158) we call the estimation procedure an **estimator**. The least squares estimate of the population mean β is obtained using (4.9) no matter what the sample values y_i turn out to be. Thus we can write the **least squares estimator** as

$$b = \sum_{i=1}^{T} Y_i / T \qquad (5.1)$$

In (5.1) we have used Y_i instead of y_i to indicate that this general formula is used what ever the sample values turn out to be. In this context the Y_i are random variables, and thus the least squares estimator b given in (5.1) is random too. Since b is random we do not know its value until a sample is collected, and different samples will lead to different values of b. To illustrate we collect 10 more samples of size T = 50 and calculate the average hip size, as given in Table 2.

Table 2 Sample means from 10 samples

Sample	*b*
1	17.355
2	16.821
3	17.412
4	17.166
5	16.901
6	16.996
7	16.837
8	16.754
9	17.098
10	16.877

The estimates are differ from sample to sample because b is a random variable. This variation due to collection of different random samples is called **sampling variation**. It is an inescapable fact of statistical analysis that the least squares estimation procedure b, and indeed all statistical estimation

procedures, are subject to **sampling variability**. Because of this terminology, the least squares estimator's (estimation procedure's) probability density function is called its **sampling distribution**.

How good is the least squares estimation procedure? We can answer this question be examining the properties of the random variable *b* and the characteristics of its probability density function (its sampling distribution.)

(5.1) The Mean or Expected Value of *b*

To evaluate the mean of *b* we can write out the formula (5.1) fully as

$$b = \sum_{i=1}^{T} Y_i / T = \frac{1}{T} Y_1 + \frac{1}{T} Y_2 + \ldots + \frac{1}{T} Y_T \tag{5.2}$$

In Chapter 2.5.1 we established that the expected value of a weighted sum like (5.2) is the sum of the expected values of the summations terms,

$$\begin{aligned} E[b] &= E\left[\frac{1}{T} Y_1\right] + E\left[\frac{1}{T} Y_2\right] + \ldots + E\left[\frac{1}{T} Y_T\right] \\ &= \frac{1}{T} E[Y_1] + \frac{1}{T} E[Y_2] + \quad + \frac{1}{T} E[Y_T] \\ &= \frac{1}{T}\beta + \frac{1}{T}\beta + \quad + \frac{1}{T}\beta \\ &= \beta \end{aligned} \tag{5.3}$$

The expected value of the least squares estimator *b* is the population mean β that we are trying to estimate. What does this mean? The expectation of a random variable is its average value in many repeated trials of an experiment, which amounts to collection of a large number of random samples from the population. Thus (5.3) says that if we did obtain many samples of size *T*, and obtained their average values, like those in Table 2, then the average of all those values would equal the true population mean β. This property is deemed a good one for estimators to have. Estimators with this property are called **unbiased estimators**. The least squares estimator *b* is an unbiased estimator of the population mean β if we follow the random sampling process for collecting the data.

Unfortunately while unbiasedness is a good property for an estimator to have, it does not tell us anything about whether our estimate *b* = 17.158, based on a single sample of data, is close to the true population mean value β or not. To assess how far the estimate *b* might be from β we will determine its variance.

(5.2) The Variance of *b*

The variance of *b* is easily obtained using the procedure for finding the variance of a weighted sum in (2.5.7). If random variables are uncorrelated (zero covariance) then the variance of the sum is the sum of

the variances. We can apply this rule if our data are obtained by random sampling, because with random sampling the observations are statistically independent, and thus have zero covariance. Furthermore, we have assumed that $\text{var}(Y_i) = \sigma^2$ for all observations. Therefore,

$$\begin{aligned}
\text{var}(b) &= \text{var}\left(\frac{1}{T}Y_1 + \frac{1}{T}Y_2 + \ldots + \frac{1}{T}Y_T\right) \\
&= \text{var}\left(\frac{1}{T}Y_1\right) + \text{var}\left(\frac{1}{T}Y_2\right) + \cdots + \text{var}\left(\frac{1}{T}Y_T\right) \\
&= \frac{1}{T^2}\text{var}(Y_1) + \frac{1}{T^2}\text{var}(Y_2) + \cdots + \frac{1}{T^2}\text{var}(Y_T) \qquad (5.4) \\
&= \frac{1}{T^2}\sigma^2 + \frac{1}{T^2}\sigma^2 + \cdots + \frac{1}{T^2}\sigma^2 \\
&= \sigma^2\left(T/T^2\right) = \sigma^2/T \qquad /
\end{aligned}$$

The other rule we have used in (5.4) is that $\text{var}(aY) = a^2\,\text{var}(Y) = a^2\sigma^2$, where $a = 1/T$, as shown in (2.3.5).

(5.3) The Sampling Distribution of *b*

If the random variable Y_i follows a normal distribution, then the least squares estimator b also follows a normal distribution. In (2.6.4) it is noted that weighted sums of normal random variables are normal themselves. From (5.2) we know that b is a weighted sum of the Y_i. If $Y_i \sim N(\beta, \sigma^2)$ then b also is normally distributed, and based on (5.3) and (5.4) we know the mean and variance of b. If $Y_i \sim N(\beta, \sigma^2)$, then $b \sim N(\beta, \sigma^2/T)$.

We can gain some intuition about the meaning of $b \sim N(\beta, \sigma^2/T)$ if we examine Figure 3.

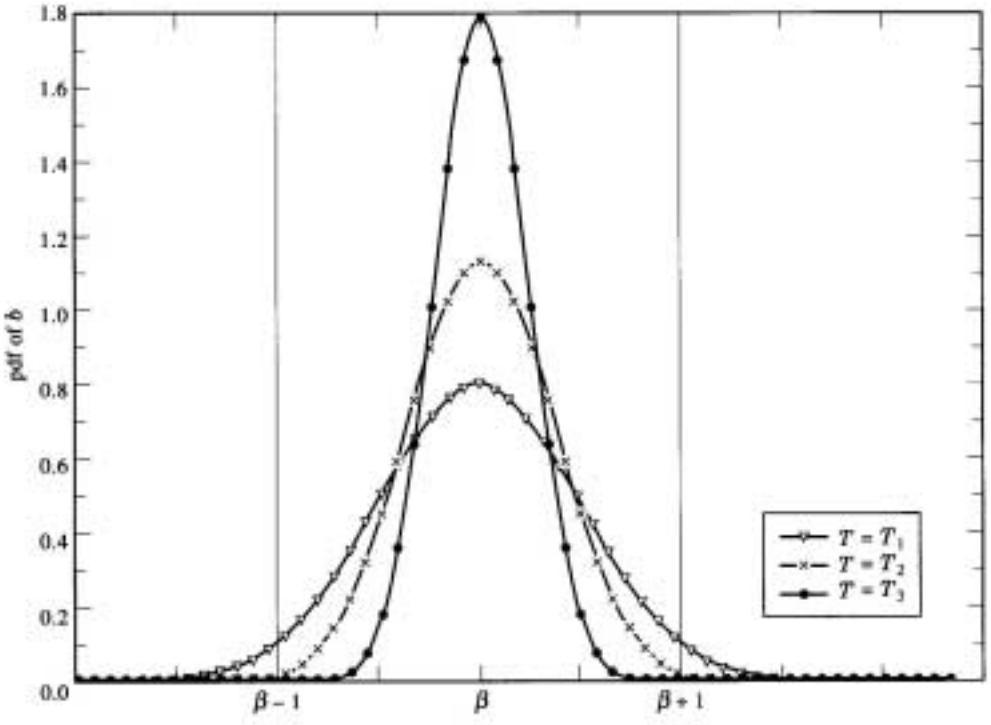

Figure 3: Increasing Sample Size and Sampling Distributions of *b*

Each of the normal distributions in Figure 3 is the sampling distribution of the least squares estimator $b = \bar{y}$. The differences between them are the samples sizes used in estimation. The sample size $T_3 > T_2 > T_1$. Increasing the sample size increases the probability that the sample mean will be "close" to the true population parameter β. Recall that the area under a probability density function measures the probability of events. If ε is any constant, the probability that $b = \bar{y}$ falls in the interval between $\beta + \varepsilon$ and $\beta - \varepsilon$ is greater for larger samples. The lesson here is that more data is better than less data, because it increases the probability of obtaining an estimate "close," within $\beta + \varepsilon$ and $\beta - \varepsilon$, of the true population parameter β.

As a numerical example, suppose we want to be within 1 inch of the true population hip width. Let us compute the probability of getting an estimate "close" to β, that is, within the interval $[\beta - 1, \beta + 1]$. For the purpose of illustration let us assume that $\sigma^2 = 10$ and that the population is normal. If for example T = 40, then $b \sim N\left(\beta, \sigma^2/T = 10/40 = .25\right)$. We can compute the probability that *b* is within 1 inch of β by computing the probability

$$P[\beta - 1 \le b \le \beta + 1]$$

As shown in Chapter 2.6, to compute this probability we convert it into one involving a "standard normal" random variable,

$$Z = \frac{b-\beta}{\sqrt{\sigma^2/T}} = \frac{b-\beta}{\sigma/\sqrt{T}} \tag{5.5}$$

Using the table of normal probabilities at the end of the book, we compute

$$\begin{aligned} P[\beta - 1 \le b \le \beta + 1] &= P\left[\frac{-1}{\sigma/\sqrt{T}} \le \frac{b-\beta}{\sigma/\sqrt{T}} \le \frac{1}{\sigma/\sqrt{T}}\right] \\ &= P\left[\frac{-1}{\sqrt{.25}} \le Z \le \frac{1}{\sqrt{.25}}\right] \\ &= P[-2 \le Z \le 2] = .95 \end{aligned} \tag{5.6}$$

Thus, if we draw a random sample of size T = 40 from a normal population with variance 10, the least squares estimator will provide an estimate within 1 inch of the true value about 95% of the time. It T = 80, or T = 200, the probability of an estimate being within 1 inch approaches *one*.

(5.4) Best Linear Unbiased Estimation

One of the powerful findings about the least squares estimator is that it is the best of all possible estimators that are both *linear* and *unbiased*. The fact that it the "best" linear unbiased estimator (BLUE) accounts for its wide use. In this context we mean by "best" that it is the estimator with the smallest variance of all linear and unbiased estimators. In the previous section we demonstrated why it is better to have an estimator with a smaller variance rather than a larger one; it increases the chances of getting an estimate close to the true population mean β. This important result about the least squares estimator is true *if* the sample values $Y_i \sim (\beta, \sigma^2)$ are independent and identically distributed. It does not depend on the population being normally distributed.

Proof: The sample mean is a weighted average of the sample values,

$$\begin{aligned} b &= \sum_{i=1}^{T} Y_i/T = \frac{1}{T}Y_1 + \frac{1}{T}Y_2 + \ldots + \frac{1}{T}Y_T \\ &= a_1Y_1 + a_2Y_2 + \quad + a_TY_T. \\ &= \sum_{i=1}^{T} a_iY_i \end{aligned} \tag{5.7}$$

where the weights $a_i = 1/T$. Weighted averages are also called **linear combinations**, thus we call the least squares estimator a linear estimator. In fact any estimator that can be written like (5.7) is a linear estimator. For example, suppose the weights a_i^* are constants different from $a_i = 1/T$. Then we can define another linear estimator of β as

$$b^* = \sum_{i=1}^{T} a_i^* Y_i \tag{5.8}$$

To ensure that b^* is different from b let us define

$$a_i^* = a_i + c_i = \frac{1}{T} + c_i \tag{5.9}$$

where c_i are constants that are not all zero. Thus

$$\begin{aligned} b^* &= \sum_{i=1}^{T} a_i^* Y_i = \sum_{i=1}^{T} \left(\frac{1}{T} + c_i \right) Y_i \\ &= \sum_{i=1}^{T} \frac{1}{T} Y_i + \sum_{i=1}^{T} c_i Y_i \\ &= b + \sum_{i=1}^{T} c_i Y_i \end{aligned} \tag{5.10}$$

The expected value of the new estimator b^* is

$$\begin{aligned} E\left[b^* \right] &= E\left[b + \sum_{i=1}^{T} c_i Y_i \right] \\ &= \beta + \sum_{i=1}^{T} c_i E\left[Y_i \right] \\ &= \beta + \beta \sum_{i=1}^{T} c_i \end{aligned} \tag{5.11}$$

The estimator b^* is not unbiased unless $\sum c_i = 0$. We want to compare the least squares estimator to other linear and unbiased estimators, so we will assume $\sum c_i = 0$ holds. Now we find the variance of b^*. The linear unbiased estimator with the smaller variance will be best.

$$\text{var}(b^*) = \text{var}\left(\sum_{i=1}^{T} a_i^* Y_i\right) = \text{var}\left(\sum_{i=1}^{T}\left(\frac{1}{T} + c_i\right) Y_i\right)$$

$$\text{var}() \quad \sum_{i=1}^{T}\left(\frac{1}{T} + c_i\right)^2 \qquad Y_i$$

$$= \sigma^2 \sum_{i=1}^{T}\left(\frac{1}{T} + c_i\right)^2$$

$$= \sigma^2 \sum_{i=1}^{T}\left(\frac{1}{T^2} + \frac{2}{T} c_i + c_i^2\right) \qquad (5.12)$$

$$= \sigma^2 \left(\frac{1}{T} + \frac{2}{T}\sum_{i=1}^{T} c_i + \sum_{i=1}^{T} c_i^2\right)$$

$$= \sigma^2 \quad T + \sigma^2 \sum_{i=1}^{T} c_i^2 \qquad \sum_{i=1}^{T} c_i =$$

$$\text{var}() \quad b \quad + \sigma^2 \sum_{i=1}^{T} c_i^2$$

It follows that the variance of b^* must be greater than the variance of *b*, unless all the c_i values are zero, in which $b^* = b$.

To numerically illustrate the importance of this finding, let us use the example developed at the end of Section 5.3. We assume that the population is normal and the population variance is $\sigma^2 = 10$. Suppose T = 20, so that for the least squares estimator the weights $a_i = 1/T = 1/20$. As an alternative estimator suppose we define b^*

$$b^* = \sum_{i=1}^{T} a_i^* Y_i = \sum_{i=1}^{10}\left(\frac{1}{10}\right) Y_i \qquad (5.13)$$

This estimator uses only the first 10 sample values, so $a_1^* = a_2^* = \cdots = a_{10}^* = 1/10$ and $a_{11}^* = a_{12}^* = \cdots = a_{20}^* = 0$. The probability density functions of the two estimators are shown in Figure 4.

The probability that *b* is within 1 inch of β is about 88%, and the probability that b^* is within 1 inch of β is about 64%. Thus the least squares estimator has a higher probability of yielding an estimate close to the true parameter β than b^*. This is true whether the population is normal or not, and no matter what the values of a_i^* might be.

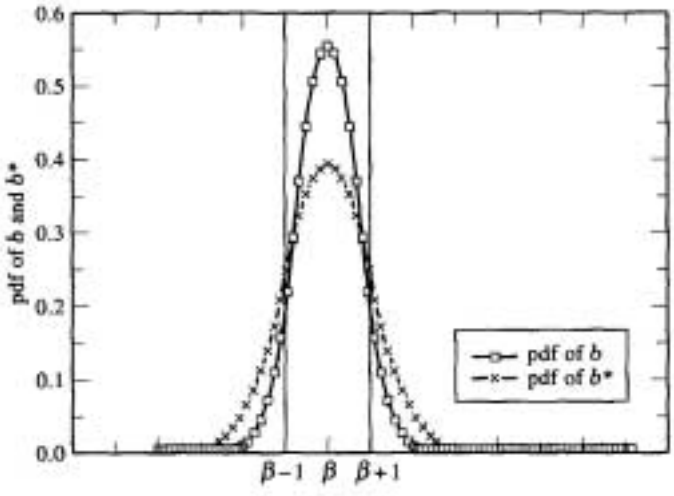

Figure 4 The probability densities of two estimators