Econometrics I - Finite sample properties of the least squares estimator

Ryan T. Godwin

University of Manitoba

Finite sample properties of the least squares estimator

We'll derive some of the finite sample properties for the LS estimator. Finite sample properties $\rightarrow$ statistical properties of $\boldsymbol{b}$ for some finite value of n. Later, we'll consider properties of $\boldsymbol{b}$ for when $n \rightarrow \infty$ (i.e. asymptotic properties). Recall that the population model is:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

and that by assumptions A.3, A.4 and A.6, the error term is:

$$\boldsymbol{\epsilon} \sim N\left[\mathbf{0}, \sigma^2 I_n\right]$$

and also recall that our LS estimator is:

$$\boldsymbol{b} = \left(X'X\right)^{-1} X'\boldsymbol{y} = f(\boldsymbol{y})$$

That is, $\boldsymbol{b}$ is a function of the random sample data, so is itself random! This is a very important point.

$$\boldsymbol{\epsilon} \text{ is random} \longrightarrow \boldsymbol{y} \text{ is random} \longrightarrow \boldsymbol{b} \text{ is random}$$

- $\boldsymbol{b}$ is an estimator of $\boldsymbol{\beta}$. It is a function of the *random* sample data.
- $\boldsymbol{b}$ is a "statistic".
- $\boldsymbol{b}$ has a probability distribution – called its *sampling distribution*.

Interpretation of the sampling distribution:

- Repeatedly draw all possible samples of size n.
- Calculate values of $\boldsymbol{b}$ each time.
- Construct a relative frequency distribution for the $\boldsymbol{b}$ values and probability of occurrence.
- It is a hypothetical construct.

Question: Why is the sampling distribution a hypothetical construct?

The *sampling distribution* offers one basis for answering the question: "How good is $\boldsymbol{b}$ as an estimator of $\boldsymbol{\beta}$?"

We will be assessing the quality of the estimator in terms of its performance in *repeated samples*. The finite sample properties that we derive in this chapter tell us nothing about the quality of the estimator for *one particular sample*.

We will explore some of the properties of the LS estimator, $\boldsymbol{b}$, and build up its sampling distribution. We'll introduce some general results, and then apply them to our specific problem. Before we do so, let's look at a *simulation experiment* using R.

For the experiment, I will estimate the simple model:

$$\boldsymbol{y} = \beta_1 + \beta_2 \boldsymbol{x} + \boldsymbol{\epsilon}$$

where I will maintain all of the usual assumptions (A.1-A.6) and choose:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 2 \\ -4 \end{bmatrix} \quad ; \quad \sigma^2 = 1 \quad ; \quad n = 100$$

In addition I will use the unrealistic assumption that x is "fixed in repeated samples", and will generate the x variable from a $N(0, 1)$ distribution (I will use the same distribution for ϵ as well).

In R, run the following code to set the parameters of the experiment:

```
beta1 <- 2
beta2 <- -4
n <- 100
x <- rnorm(n)
```

and now run the following code several times:

```
epsilon <- rnorm(n)
y <- beta1 + beta2 * x + epsilon
lm(y ~ x)
```

```
Coefficients:
(Intercept)            x
      1.886       -4.060
```

Question: How can you use the above code to *simulate* the sampling distribution of b_2?

To simulate the sampling distribution, we need to run the above code many times, and collect the value of b_2 each time.

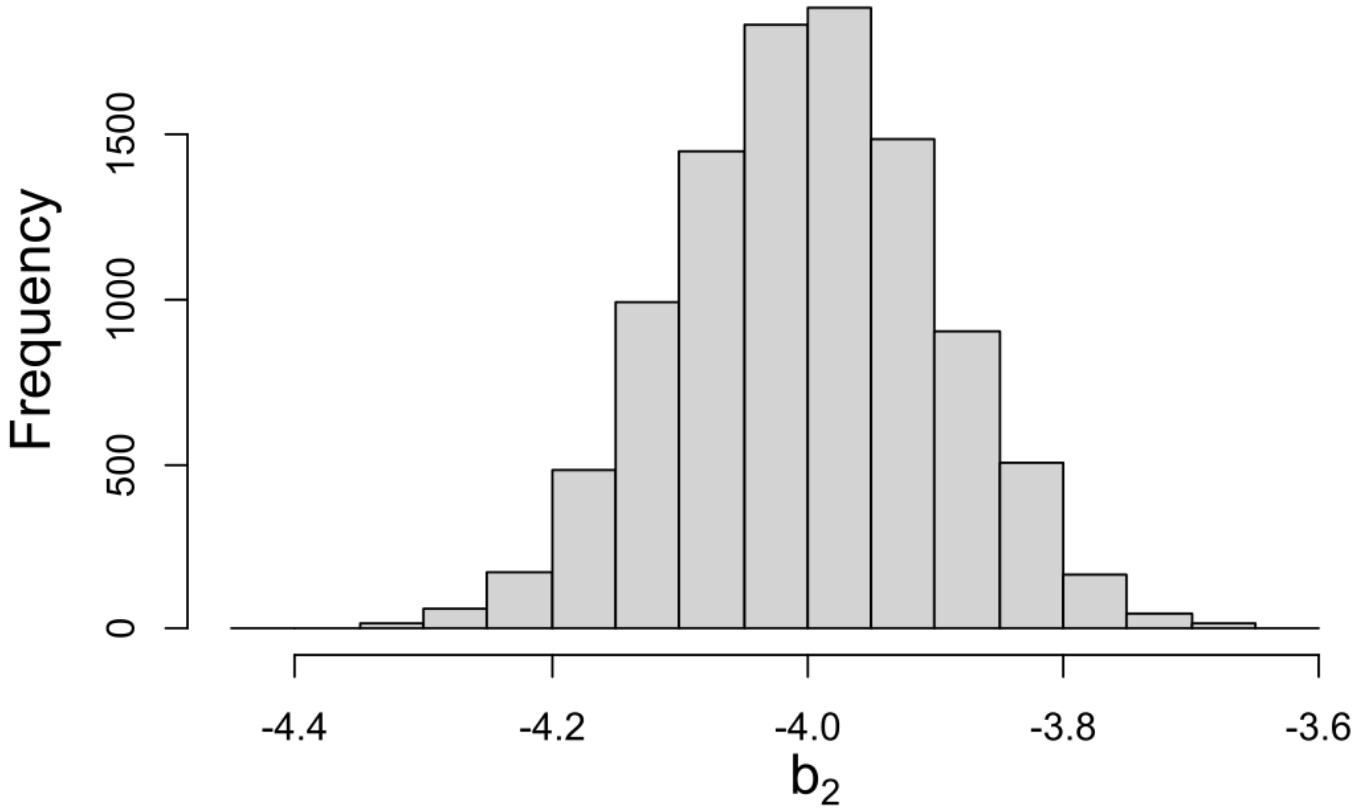
Then, we can plot b_2 in a histogram, take mean of b_2, take the var$[b_2]$, etc. I set the *random seed* so that when you run the code on your computer, you get the same results as me. I use a `for` loop to repeat the experiment `nrep` $= 10000$ times.

```
set.seed(7010)
nrep <- 10000
n <- 100
x <- rnorm(n)
beta1 <- 2
beta2 <- -4
b2 <- numeric(nrep)
for(i in 1:nrep){
  epsilon <- rnorm(n, mean=0, sd=1)
  y <- beta1 + beta2 * x + epsilon
  b2[i] <- lm(y ~ x)$coefficients[2]
}
mean(b2)
var(b2)
hist(b2)
```

```
> mean(b2)
[1] -4.000452
> var(b2)
[1] 0.01061148
```

The histogram of all 10,000 b_2 values is shown in Figure 1.

Figure: A simulated sampling distribution.



Questions:

1. What distribution appears to describe Figure 1?
2. Is the average value of the *estimator* close to the true *population* value?

Unbiased

Unbiased estimator. An estimator, $\hat{\boldsymbol{\theta}}$, is an *unbiased* estimator of the parameter vector, $\boldsymbol{\theta}$, if:

$$E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$$

That is, if $E[\hat{\boldsymbol{\theta}}(\boldsymbol{y})] = \boldsymbol{\theta}$, or $\int \hat{\boldsymbol{\theta}}(\boldsymbol{y}) p(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{y} = \boldsymbol{\theta}$. The "Bias" of $\hat{\boldsymbol{\theta}}$ is the quantity:

$$\text{Bias}(\boldsymbol{\theta}, \boldsymbol{y}) = E[\widehat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}]$$

In words, an estimator is *unbiased* if it gives the right answer on average.

Unbiasedness of $\bar{y}$. Let $\{y_1, y_2, \ldots, y_n\}$ be a random sample from population with a finite mean, μ, and a finite variance, σ^2. Consider the *statistic*:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Then,

$$E[\bar{y}] = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left(y_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mu = \left(\frac{1}{n} n\mu\right) = \mu$$

So, $\bar{y}$ is an *unbiased estimator* of the parameter, μ. Here, there are lots of possible unbiased estimators of μ. So, we will need to consider additional characteristics of estimators to help us choose from among them.

Return to our LS problem:

$$\boldsymbol{b} = (X'X)^{-1} X'\boldsymbol{y}$$

Recall assumption A.5. We will use the strongest version of this assumption - that X is non-random. We could use the weaker versions of the assumption, and we'll get the same results, but the notation will be more cumbersome. Now, take the expected value of the random estimator $\boldsymbol{b}$:

$$\begin{aligned} E(\boldsymbol{b}) &= E\left[(X'X)^{-1} X'\boldsymbol{y}\right] = (X'X)^{-1} X'E(\boldsymbol{y}) \\ &= (X'X)^{-1} X'E[X\boldsymbol{\beta} + \boldsymbol{\epsilon}] = (X'X)^{-1} X'[X\boldsymbol{\beta} + E(\boldsymbol{\epsilon})] \\ &= (X'X)^{-1} X'[X\boldsymbol{\beta} + \mathbf{0}] = (X'X)^{-1} X'X\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned} \tag{1}$$

This proves that the **LS estimator of $\boldsymbol{\beta}$ is Unbiased**.

Linear

Linear estimator. Any estimator that is a *linear function* of the random sample data is called a *Linear Estimator*.

Sample average is linear. Let $\{y_1, y_2, \ldots, y_n\}$ be a random sample from population with a finite mean, μ, and a finite variance, σ^2. Consider the statistic:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\left[y_1 + y_2 + \cdots + y_n\right]$$

This statistic is a linear estimator of μ (note that the "weights" are non-random).

Return to our LS problem:

$$\underset{(k\times 1)}{\boldsymbol{b}} \quad = \quad (X'X)^{-1}\,X'\boldsymbol{y} \quad = \quad \underset{(k\times n)(n\times 1)}{A\boldsymbol{y}}$$

Note that, under our (strictest form of the) assumptions, A is a *non-random* matrix. So,

$$\begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \dots & \vdots \\ a_{k1} & \dots & a_{kn} \end{bmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

For example, $b_1 = [a_{11}y_1 + a_{12}y_2 + \cdots + a_{1n}y_n]$, etc.

Thus: **the LS estimator, $\boldsymbol{b}$, is a linear (and unbiased) estimator of $\boldsymbol{\beta}$.**

Efficient

Now let's consider the dispersion (variability) of $\boldsymbol{b}$, as an estimator of $\boldsymbol{\beta}$.

Suppose we have an $(n \times 1)$ random vector, $\boldsymbol{x}$. Then the *covariance matrix* of $\boldsymbol{x}$ is defined as the $(n \times n)$ matrix:

$$V(\boldsymbol{x}) = E\left[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))'\right]$$

- Diagonal elements of $V(\boldsymbol{x})$ are $\text{var}(x_1), \ldots, \text{var}(x_n)$.
- Off-diagonal elements are $\text{cov}(x_i, x_j)$; $i, j = 1, \ldots, n$; $i \neq j$.
- We have already made use of the *covariance matrix* when we made assumption A.4.

Return to our LS problem. We have a $(k \times 1)$ random vector, $\boldsymbol{b}$, and we know that $E(\boldsymbol{b}) = \boldsymbol{\beta}$. The *covariance matrix* of $\boldsymbol{b}$, $V(\boldsymbol{b})$, is:

$$\begin{aligned} V(\boldsymbol{b}) &= E\left[(\boldsymbol{b} - E(\boldsymbol{b}))(\boldsymbol{b} - E(\boldsymbol{b}))'\right] \\ &= E\left[(\boldsymbol{b} - \boldsymbol{\beta})(\boldsymbol{b} - \boldsymbol{\beta})'\right] \end{aligned}$$

Now,

$$\begin{aligned} \boldsymbol{b} &= (X'X)^{-1} X'\boldsymbol{y} = (X'X)^{-1} X'(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= (X'X)^{-1} (X'X)\boldsymbol{\beta} + (X'X)^{-1} X'\boldsymbol{\epsilon} \\ &= I\boldsymbol{\beta} + (X'X)^{-1} X'\boldsymbol{\epsilon} \end{aligned}$$

So,

$$(\boldsymbol{b} - \boldsymbol{\beta}) = (X'X)^{-1} X'\boldsymbol{\epsilon} \tag{2}$$

Using the result from 2 in $V(\boldsymbol{b})$ we have:

$$\begin{aligned} V(\boldsymbol{b}) &= E\left\{ \left[(X'X)^{-1} X'\boldsymbol{\epsilon}\right] \left[(X'X)^{-1} X'\boldsymbol{\epsilon}\right]' \right\} \\ &= (X'X)^{-1} X' E\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right] X (X'X)^{-1} \end{aligned}$$

For assumption A.4, we showed earlier that because $E(\boldsymbol{\epsilon}) = \mathbf{0}$:

$$V(\boldsymbol{\epsilon}) = E\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right) = \sigma^2 I_n$$

So, we have:

$$\begin{aligned} V(\boldsymbol{b}) &= \left(X'X\right)^{-1} X' E\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right] X \left(X'X\right)^{-1} \\ &= \left(X'X\right)^{-1} X' \sigma^2 I X \left(X'X\right)^{-1} \\ &= \sigma^2 \left(X'X\right)^{-1} \left(X'X\right) \left(X'X\right)^{-1} \\ &= \sigma^2 \left(X'X\right)^{-1}. \end{aligned}$$

So, the covariance matrix of $\boldsymbol{b}$ is:

$$V(\boldsymbol{b}) = \sigma^2 \left(X'X\right)^{-1} \tag{3}$$

Question: What is the interpretation of the diagonal and off-diagonal elements of this matrix? What might the elements of this matrix be used for, in practice?

Finally, because the error term, $\boldsymbol{\epsilon}$ is assumed to be Normally distributed,

1. $y = X\boldsymbol{\beta} + \epsilon$: this implies that $\boldsymbol{y}$ is also Normally distributed.
2. $\boldsymbol{b} = (X'X)^{-1} X'\boldsymbol{y} = A\boldsymbol{y}$: this implies that $\boldsymbol{b}$ is also Normally distributed.

Question: Why does the Normality of $\boldsymbol{\epsilon}$ transfer to $\boldsymbol{b}$?

We now have the full *sampling distribution* of the LS estimator, $\boldsymbol{b}$:

$$\boldsymbol{b} \sim N\left[\boldsymbol{\beta}, \sigma^2 \left(X'X\right)^{-1}\right]$$

Note:

- This result depends on our *rigid* assumptions about the various components of the regression model.
- The Normal distribution here is a "multivariate Normal" distribution.
- As with estimation of the population mean μ, there are lots of other unbiased estimators of $\boldsymbol{\beta}$.

Question: How might we choose between the many possible linear and unbiased estimators for $\boldsymbol{\beta}$? Why is linearity desirable?

We need to consider other desirable properties that these unbiased estimators may have. One option to help discern the "best" estimator for $\boldsymbol{\beta}$ is to take into account the estimators' *precisions*.

Suppose that we have two different *unbiased* estimators, $\widehat{\theta}_1$ and $\widehat{\theta}_2$, of the (scalar) parameter, θ. Then we say that $\widehat{\theta}_1$ is **at least as efficient** as $\widehat{\theta}_2$ if $\text{var}\left(\widehat{\theta}_1\right) \leq \text{var}\left(\widehat{\theta}_2\right)$

- The variance of an estimator is just the variance of its sampling distribution.
- "Efficiency" is a relative concept.

Question: What if there are 3 or more unbiased estimators being compared?

Mean Squared Error (MSE)

If one or more of the estimators being compared is biased we can take account of both variance, and any bias, at the same time by using "mean squared error" (MSE) of the estimators.

Suppose that $\widehat{\theta}$ is an estimator of the (scalar) parameter, θ. Then the MSE of $\widehat{\theta}$ is defined as:

$$\text{MSE}(\widehat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right]$$

Note that:

$$\text{MSE}(\widehat{\theta}) = \text{var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2$$

Efficiency and MSE for a vector of estimators

If we extend all of this to the case where we have a vector of parameters, $\boldsymbol{\theta}$, then we have the following definitions:

Suppose that we have two different *unbiased* estimators, $\widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2$, of the parameter vector, $\boldsymbol{\theta}$. Then we say that $\widehat{\boldsymbol{\theta}}_1$ is **at least as efficient** as $\widehat{\boldsymbol{\theta}}_2$ if $\Delta = V\left(\widehat{\boldsymbol{\theta}}_2\right) - V\left(\widehat{\boldsymbol{\theta}}_1\right)$ is *positive semi-definite*.

Suppose that we have two different (possibly) *biased* estimators, $\widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2$, of the parameter vector, $\boldsymbol{\theta}$. Then we say that $\widehat{\boldsymbol{\theta}}_1$ is **at least as efficient** as $\widehat{\boldsymbol{\theta}}_2$ if $\Delta = \text{MMSE}\left(\widehat{\boldsymbol{\theta}}_2\right) - \text{MMSE}\left(\widehat{\boldsymbol{\theta}}_1\right)$ is *positive semi-definite*.

Note: $\text{MMSE}(\widehat{\boldsymbol{\theta}}) = E\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\right] = V[\widehat{\boldsymbol{\theta}}] + \text{Bias}(\widehat{\boldsymbol{\theta}})\,\text{Bias}(\widehat{\boldsymbol{\theta}})'$.

Gauss-Markhov theorem

Taking account of its *linearity*, *unbiasedness*, and its *precision*, in what sense is the LS estimator, $\boldsymbol{b}$, of $\boldsymbol{\beta}$, optimal?

Gauss-Markhov Theorem: In the "standard" linear regression model, $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the LS estimator, $\boldsymbol{b}$, of $\boldsymbol{\beta}$ is **Best Linear Unbiased** (BLU). That is, it is *efficient* in the class of all linear and unbiased estimators of $\boldsymbol{\beta}$.

Question: Why is this an interesting result?

Proof: Let $\boldsymbol{b}_0$ be any other *linear* estimator of $\boldsymbol{\beta}$:

$$\boldsymbol{b}_0 = C\boldsymbol{y} \quad ; \quad \text{for some non-random } C$$

The *covariance matrix* for $\boldsymbol{b}_0$ is:

$$V\left(\boldsymbol{b}_0\right) = CV(\boldsymbol{y})C' = C\left(\sigma^2 I_n\right)C' = \sigma^2 CC'$$

Take the difference between $\boldsymbol{b}_0$ and $\boldsymbol{b}$:

$$\boldsymbol{b}_0 - \boldsymbol{b} = C\boldsymbol{y} - (X'X)^{-1} X'\boldsymbol{y} = D\boldsymbol{y},$$

where

$$D = C - (X'X)^{-1} X'$$

is the difference between how the other estimator uses the X data to "weight" the y data (C), and how the LS estimator uses the X data ($(X'X)^{-1} X'$). Now restrict $\boldsymbol{b}_0$ to be unbiased, so that:

$$E(\boldsymbol{b}_0) = E(C\boldsymbol{y}) = CX\boldsymbol{\beta} = \boldsymbol{\beta}.$$

This requires that $CX = I$, which in turn implies that:

$$DX = \left[C - (X'X)^{-1} X'\right] X = CX - I = 0 \quad (\text{and } X'D' = 0)$$

Solve for C in terms of D:

$$C = D + (X'X)^{-1} X',$$

and return to the covariance matrix of $\boldsymbol{b}_0$:

$$\begin{aligned} V(\boldsymbol{b}_0) &= \sigma^2 CC' \\ &= \sigma^2 \left[D + (X'X)^{-1} X'\right] \left[D + (X'X)^{-1} X'\right]' \\ &= \sigma^2 \left[DD' + (X'X)^{-1} X'X (X'X)^{-1}\right] \quad ; \quad DX = X'D' = 0 \\ &= \sigma^2 DD' + \sigma^2 (X'X)^{-1} \\ &= \sigma^2 DD' + V(\boldsymbol{b}) \end{aligned}$$

or:

$$[V(\boldsymbol{b}_0) - V(\boldsymbol{b})] = \sigma^2 DD' \quad ; \quad \sigma^2 > 0 \tag{4}$$

Now we just have to "sign" this (matrix) difference:

$$\boldsymbol{\eta}' (DD') \boldsymbol{\eta} = (D'\boldsymbol{\eta})' (D'\boldsymbol{\eta}) = v'v = \sum_{i=1}^{n} v_i^2 \geq 0$$

So, $\Delta = [V(\boldsymbol{b_0}) - V(\boldsymbol{b})]$ is a p.s.d. matrix, implying that $\boldsymbol{b}_0$ is relatively less efficient than $\boldsymbol{b}$. Result:

The LS estimator is the Best Linear Unbiased estimator (BLUE) of $\boldsymbol{\beta}$.

Questions:

1. What assumptions did we use, and where?
2. Were there any standard assumptions that we *didn't* use?
3. What does this suggest?