Lecture 6: The Simple Linear Regression Model: Interval Estimation and a Monte Carlo Experiment

by
Professor Scott H. Irwin

Required Readings:

Griffiths, Hill and Judge. “Interval Estimation,” Section 7.1 and "A Monte Carlo Experiment to Demonstrate the Sampling Performance of the Least Squares Estimator," Section 6.5 in Learning and Practicing Econometrics
Review of Least Squares Estimators and Sampling Properties

In the previous two lectures, we specified the simple linear regression model,

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

In compact form, the assumptions of the simple linear regression model are,

SR1. $y_t = \beta_1 + \beta_2 x_t + e_t, \quad t = 1, \ldots, T$

SR2. $E(e_t) = 0 \iff E(y_t) = \beta_1 + \beta_2 x_t$

SR3. $\text{var}(e_t) = \text{var}(y_t) = \sigma^2$

SR4. $\text{cov}[e_t, e_s] = \text{cov}[y_t, y_s] = 0 \quad t \neq s$

SR5. The variable $x_t$ is not random and must take on at least two different values

SR6. $e_t \sim N(0, \sigma^2) \iff y_t \sim N(\beta_1 + \beta_2 x_t, \sigma^2)$
The SLR model has three unknown parameters, $\beta_1$, $\beta_2$, and $\sigma^2$

Since the model contains unknown parameters, we

1. Used the concept of least squares to develop the rules $b_1$ and $b_2$ to estimate $\beta_1$ and $\beta_2$

2. Noted that since $b_1$ and $b_2$ were linear functions of the random sample observations $y_i$, the least squares estimators $b_1$ and $b_2$ also were random variables

3. Found that the least squares estimators $b_1$ and $b_2$ were unbiased, or that $E(b_1) = \beta_1$ and $E(b_2) = \beta_2$

4. Noted that intercept and slope estimates could vary substantially from sample to sample and some estimates could miss the mark ($\beta_1$ and $\beta_2$) quite badly

5. Developed the sampling distributions of the least squares estimators $b_1$ and $b_2$ in terms of means, variances, and covariance
6. Discovered that \( \text{var}(b_1), \text{var}(b_2), \text{and } \text{cov}(b_1, b_2) \) were directly related to the unknown parameter \( \sigma^2 \)

7. Derived \( \hat{\sigma}^2 \) as an unbiased estimator of \( \sigma^2 \) and used this rule to obtain estimates of the variances and covariance of \( b_1 \) and \( b_2 \) [\( \hat{\text{var}}(b_1), \hat{\text{var}}(b_2), \text{and } \hat{\text{cov}}(b_1, b_2) \)]

8. Noted that out of the class of linear unbiased estimation rules, the least squares estimators \( b_1 \) and \( b_2 \) are best in the sense of having a minimum sampling variability
Key Formulas

Least squares estimators of $\beta_1$ and $\beta_2$,

$$b_1 = \bar{y} - b_2 \bar{x}$$
$$b_2 = \frac{T \sum_{t=1}^{T} y_t x_t - \sum_{t=1}^{T} x_t \sum_{t=1}^{T} y_t}{T \sum_{t=1}^{T} x_t^2 - \left( \sum_{t=1}^{T} x_t \right)^2}$$

Variances and covariance of the sampling distributions of $b_1$ and $b_2$,

$$\text{Var}(b_1) = \sigma^2 \left[ \frac{\sum_{t=1}^{T} x_t^2}{T \sum_{t=1}^{T} (x_t - \bar{x})^2} \right]$$
$$\text{Var}(b_2) = \sigma^2 \left[ \frac{1}{\sum_{t=1}^{T} (x_t - \bar{x})^2} \right]$$

$$\text{Cov}(b_1, b_2) = \sigma^2 \left[ \frac{-\bar{x}}{\sum_{t=1}^{T} (x_t - \bar{x})^2} \right]$$
Estimator of $e_t$,

$$\hat{e}_t = y_t - b_1 - b_2 x_t$$

Estimator of $\sigma^2$,

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{T} \hat{e}_t^2}{T - 2}$$

Estimators of the variances and covariance of the sampling distributions of $b_1$ and $b_2$,

$$\text{vâr}(b_1) = \hat{\sigma}^2 \left[ \frac{\sum_{t=1}^{T} x_t^2}{T \sum_{t=1}^{T} (x_t - \bar{x})^2} \right]$$
$$\text{vâr}(b_2) = \hat{\sigma}^2 \left[ \frac{1}{\sum_{t=1}^{T} (x_t - \bar{x})^2} \right]$$
$$\text{ôcov}(b_1, b_2) = \hat{\sigma}^2 \left[ \frac{-\bar{x}}{\sum_{t=1}^{T} (x_t - \bar{x})^2} \right]$$
**Interval Estimation**

In the food expenditure example we reported regression estimates of the unknown $\beta_1$ and $\beta_2$ parameters as,

\[
b_1 = 7.3832 \quad b_2 = 0.2323
\]

These estimates are known as "point estimates"

- Point estimates alone do not give any sense of the reliability of the estimates

- Estimated variances for $b_1$ and $b_2$ do provide measures of the sampling variability of the least squares estimators

Interval estimators combine the information from the point estimator with the sampling variance to provide an indication of the reliability of the estimate
Interval Estimation when $\sigma^2$ is Known

We will begin by considering interval estimation of $b_2$ when $\sigma^2$ is assumed to be known

- Assuming $\sigma^2$ to be known is unrealistic, but helpful in understanding interval estimation

- Interval estimation for $b_1$ is the same, just substitute the appropriate formulas

To begin, note that the sampling distribution for $b_2$ is,

$$b_2 \sim N \left( \beta_2, \sigma^2 \frac{1}{\sum_{t=1}^{T} (x_t - \bar{x})^2} \right)$$

or,

$$b_2 \sim N(\beta_2, \text{var}(b_2))$$
FIGURE 11.3 A thought experiment helps explain the nature of the sampling distributions of the OLS estimates of $\beta_0$ and $\beta_1$. The same three $X$ values are used in each sample, but different sets of $Y$ values are produced because different values for the disturbances occur in each sample. The $\hat{\beta}_0$ and $\hat{\beta}_1$ values computed in each sample are collected, and their frequency distributions are constructed. The sampling distribution of an estimator is the limiting form of the frequency distribution as $N$ approaches infinity. The actual $\hat{\beta}_0$ and $\hat{\beta}_1$ that we calculate from a set of data are thought of as just one pair of outcomes from these sampling distributions.

Next, consider subtracting the mean of the sampling distribution from $b_2$ and dividing by the standard error of $b_2$ to create a new "standardized" random variable,

$$Z_{b_2} = \frac{b_2 - \beta_2}{se(b_2)} = \frac{b_2 - \beta_2}{\sqrt{\text{var}(b_2)}} = \frac{b_2 - \beta_2}{\sqrt{\frac{1}{T} \sum_{t=1}^{T} (x_t - \bar{x})^2}}$$

When we do this, we know that the sampling distribution of the new random variable is,

$$Z_{b_2} \sim N(0,1)$$

In general, we know that we can pick critical values $c$ and $-c$ such that,

$$P[Z \geq c] = P[Z \leq -c] = \frac{\alpha}{2}$$

or,

$$P[Z \geq c] + P[Z \leq -c] = \alpha$$

and,

$$P[-c \leq Z \leq c] = 1 - \alpha$$

where $\alpha$ is the probability
From a $Z$-table, we know that when $c=1.96$ and $-c=-1.96$ the probability $\alpha$ is 0.05

This can be stated in mathematical form as,

$$P[Z \geq 1.96] = P[Z \leq -1.96] = \frac{0.05}{2} = 0.025$$

or,

$$P[Z \geq 1.96] + P[Z \leq -1.96] = 0.025 + 0.025 = 0.05$$

and,

$$P[-1.96 \leq Z \leq 1.96] = 1 - 0.05 = 0.95$$

Since $Z_{b_2}$ is a standard, normal random variable, we can write,

$$P[-1.96 \leq Z_{b_2} \leq 1.96] = 0.95$$

Substituting for $Z_{b_2}$,

$$P[-1.96 \leq \frac{b_2 - \beta_2}{\sqrt{\text{var}(b_2)}} \leq 1.96] = 0.95$$
Figure 4.1  $\alpha = .05$ critical values for the $N(0, 1)$ distribution.
Now, we want to "work backwards" to derive an interval estimator that incorporates both the point estimate and the standard error of the sampling distribution.

Multiply the inequality in the brackets by $\sqrt{\text{var}(b_2)}$,

$$P[-1.96 \cdot \sqrt{\text{var}(b_2)} \leq b_2 - \beta_2 \leq 1.96 \cdot \sqrt{\text{var}(b_2)}] = 0.95$$

Now, subtract $b_2$ from each term,

$$P[-b_2 -1.96 \cdot \sqrt{\text{var}(b_2)} \leq -\beta_2 \leq -b_2 + 1.96 \cdot \sqrt{\text{var}(b_2)}] = 0.95$$

Next, multiply the inequality by -1 to reverse the direction of the inequality,

$$P[b_2 + 1.96 \cdot \sqrt{\text{var}(b_2)} \geq \beta_2 \geq b_2 - 1.96 \cdot \sqrt{\text{var}(b_2)}] = 0.95$$

Finally, "flip" the entire inequality as follows,

$$P[b_2 -1.96 \cdot \sqrt{\text{var}(b_2)} \leq \beta_2 \leq b_2 + 1.96 \cdot \sqrt{\text{var}(b_2)}] = 0.95$$

This is the interval estimator for $\beta_2$ at the 95 percent confidence level $[(1-0.05)100]$.
We can generalize the interval estimator to any desired confidence level as follows,

\[ P[b_2 - Z_{\alpha/2} \cdot \sqrt{\text{var}(b_2)} \leq \beta_2 \leq b_2 + Z_{\alpha/2} \cdot \sqrt{\text{var}(b_2)}] = 1 - \alpha \]

where \( Z_{\alpha/2} \) is the appropriate critical value from a \( Z \)-table for \( \alpha \) percent tail probability (\( \alpha \) is the sum of both lower and upper tail probability)

**Interpretation of Interval Estimator**

Interpretation is subtle and widely abused in practice

Interval is random because \( b_2 \) is a random variable before sampling

- **Width** of interval is constant in the case when \( \sigma^2 \) is known

- "Location" of interval is variable due to the uncertainty about the particular estimate, \( b_2 \), that will be generated in the random sampling process

Correct meaning of interval estimator
• In **repeated sampling**, we expect 95% of interval estimates to **contain** $\beta_2$.

• If we use the interval estimator to compute a “large” number of interval estimates like $b_2 \pm 1.96\sqrt{\text{var}(b_2)}$, 95% of these intervals will contain (“cover”) $\beta_2$.

Incorrect meaning of interval estimator

• There is a **0.95** probability of $\beta_2$ being in the interval $b_2 \pm 1.96\sqrt{\text{var}(b_2)}$.

• This implies $\beta_2$ is **random** and the interval is not, when in reality just the opposite is true!

_**Remember, our confidence is in the procedure used to construct the interval estimator, not in the interval estimate from one particular sample!**_
Sampling Distribution of $b_2$

$(b_2 \sim N(\beta_2, 4))$

$\alpha = 0.025$

$Z_{\alpha/2}\text{var}(b_2) = 1.96 \cdot 4 = 7.84$

$\beta_2$

$\pm 7.84$

$[b_2^1, b_2^2]$ Interval based on $b_2 = 7.84$

$[b_2^3, b_2^4]$ Interval based on $b_2^1, b_2^2$

$[b_2^N]$

Interval Estimates Based on Selected Sample Estimates of $\beta_2$ when $\sigma^2$ is Known
Interval Estimation when $\sigma^2$ is Unknown

In practice, $\sigma^2$ is rarely known and it is not usually possible to construct interval estimators as we did in the previous section.

Instead, we can replace $\sigma^2$ by its unbiased estimator $\hat{\sigma}^2$.

As a result, the estimator of $\text{var}(b_2)$ is,

$$\text{var}(b_2) = \hat{\sigma}^2 \left[ \frac{1}{T} \sum_{i=1}^{T} (x_i - \bar{x})^2 \right]$$

Based on this, the next step would appear to be simple:

- Just "standardize" the sampling distribution of $b_2$ by replacing $\text{var}(b_2)$ with $\text{var}(b_2)$

$$Z_{b_2} = \frac{b_2 - \beta_2}{\sqrt{\text{var}(b_2)}}$$

and proceed as before to "work backward" and generate an interval estimator.
• Unfortunately, the simple act of replacing \( \text{var}(b_2) \) with \( \hat{\text{var}}(b_2) \) does not generate a standard, normal random variable.

When \( \sigma^2 \) is known, and hence, \( \text{var}(b_2) \) is known, the standardized random variable,

\[
Z_{b_2} = \frac{b_2 - \beta_2}{\sqrt{\text{var}(b_2)}}
\]

is a function of only one random variable, \( b_2 \), and distributed as \( \text{N}(0,1) \).

When we replace \( \sigma^2 \) by its unbiased estimator \( \hat{\sigma}^2 \), and hence, replace \( \text{var}(b_2) \) with \( \hat{\text{var}}(b_2) \), the resulting standardized random variable,

\[
? = \frac{b_2 - \beta_2}{\sqrt{\hat{\text{var}}(b_2)}}
\]

is a function of the ratio of two random variables, \( b_2 \) and \( \hat{\text{var}}(b_2) \), and is not distributed \( \text{N}(0,1) \).
Deriving the Sampling Distribution of $t_{b_2}$

To begin, we introduce new notation to denote the new, standardized random variable,

$$t_{b_2} = \frac{b_2 - \beta_2}{\sqrt{\text{vâr}(b_2)}}$$

We want to derive the sampling distribution of $t_{b_2}$

- First worked out by W.S. Gossett in 1919
- An employee of the Guiness Brewery in Scotland
- A very clever derivation

To derive the distribution of $t_{b_2}$, we need to first review chi-square random variables
First, let $Z_1, Z_2, \ldots, Z_m$ denote $m$ independent standard normal random variables, where each is distributed $\mathcal{N}(0,1)$

Next, form a new random variable as

$$V = Z_1^2 + Z_2^2 + \ldots + Z_m^2$$

We say $V$ is distributed as a chi-square with $m$ degrees of freedom

Now, consider the ratio of another $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(m)$, where $Z$ and $V$ are independent

$$t = \frac{Z}{\sqrt{V/m}} \sim t(m)$$

We say that $t$ follows a $t$-distribution with $m$ degrees of freedom

Note that

$$E[t] = 0 \quad \text{var}[t] = E[t]^2 = \frac{m}{m-2}$$

Our challenge is to show that $t_{b_2}$ is of the form,
The first piece of the puzzle is to define the numerator in the above expression as,

$$Z_{b_2} = \frac{b_2 - \beta_2}{\sqrt{\text{var}(b_2)}} \sim N(0,1)$$

The second piece of the puzzle is to define the denominator $\sqrt{V_{b_2}/T-2}$

We begin by re-stating earlier results about deriving the sampling distribution of $\hat{\sigma}^2$,

$$e_t \sim N(0, \sigma^2) \quad t = 1, \ldots, T$$

Consequently,

$$\frac{e_t}{\sigma} \sim N(0,1) \quad t = 1, \ldots, T$$
and,

$$\left(\frac{e_i}{\sigma}\right)^2 \sim \chi_1 \quad i = 1, \ldots, T$$

If we sum over the $T$ transformed random errors,

$$\sum_{i=1}^{T} \left(\frac{e_i}{\sigma}\right)^2 \sim \chi_T$$

Next, replace the unobserved population errors with their sample estimates $\hat{e}_t = y_t - b_1 - b_2 x_t$, and define this random variable as $V_{b_2}$,

$$V_{b_2} = \sum_{i=1}^{T} \left(\frac{\hat{e}_t}{\sigma}\right)^2 \sim \chi_{T-2}$$

The previous equation can be re-written as,

$$V_{b_2} = \frac{1}{\sigma^2} \sum_{i=1}^{T} \hat{e}_t^2$$

Note that,

$$\hat{\sigma}^2 = \frac{\sum_{t=1}^{T} \hat{e}_t^2}{T - 2} \quad \text{or} \quad (T - 2)\hat{\sigma}^2 = \sum_{t=1}^{T} \hat{e}_t^2$$
Combining the previous two expressions,

\[ V_{b_2} = \frac{1}{\sigma^2} (T - 2)\hat{\sigma}^2 = \frac{(T - 2)\hat{\sigma}^2}{\sigma^2} \sim \chi_{T-2} \]

We have now derived expressions for the two random variables, \( Z_{b_2} \) and \( V_{b_2} \), and determined that they are distributed as, respectively, \( N(0,1) \) and \( \chi_{T-2} \).

The final step is to substitute the expressions for \( Z_{b_2} \) and \( \sqrt{V_{b_2} / (T - 2)} \) into the \( t \)-distribution formula and see if we can arrive back at the original definition of \( t_{b_2} \).

The starting point is,

\[ t = \frac{Z_{b_2}}{\sqrt{V_{b_2} / m}} \]
Substituting and simplifying,

\[ t_{b_2} = \frac{b_2 - \beta_2}{\sqrt{\text{var}(b_2)}} = \frac{b_2 - \beta_2}{\sqrt{(T-2)\hat{\sigma}^2}} = \frac{b_2 - \beta_2}{\sqrt{\frac{\hat{\sigma}^2}{\sigma^2}}(T-2)} \]

We have now established that,

\[ t = \frac{b_2 - \beta_2}{\sqrt{\text{var}(b_2)}} \cdot \frac{\sigma}{\hat{\sigma}} \]

Now substitute for \( \text{var}(b_2) \) and simplify,

\[ t = \frac{b_2 - \beta_2}{\sqrt{\sigma^2 \sum_{t=1}^{T}(x_t - \bar{x})^2}} \cdot \frac{\sigma}{\hat{\sigma}} \]

\[ t = \frac{b_2 - \beta_2}{\sigma \sqrt{\sum_{t=1}^{T}(x_t - \bar{x})^2}} \cdot \frac{\sigma}{\hat{\sigma}} \]
Finally,

\[ t = \frac{b_2 - \beta_2}{\hat{\sigma} \sqrt{\frac{1}{T} \sum_{t=1}^{T} (x_t - \bar{x})^2}} \]

We have now demonstrated that \( t = t_{b_2} \), and hence, \( t_{b_2} \) is distributed as a t-distribution with \( T-2 \) degrees of freedom.

We can state this result more compactly as,

\[ t_{b_2} = \frac{b_2 - \beta_2}{\sqrt{\text{var}(b_2)}} \sim t_{T-2} \]
It is interesting to consider the mean of and variance of $t_{b_2}$,

$$E[t_{b_2}] = E\left[ \frac{b_2 - \beta_2}{\sqrt{\text{var}(b_2)}} \right] = 0$$

$$\text{var}[t_{b_2}] = \text{var}\left[ \frac{b_2 - \beta_2}{\sqrt{\text{var}(b_2)}} \right]$$

$$= \frac{m}{m-2}$$

$$= \frac{T-2}{(T-2)-2}$$

$$= \frac{T-2}{T-4}$$

Which can be summarized as

$$t_{b_2} \sim t_{T-2} \left( 0, \frac{T-2}{T-4} \right)$$

**How do these differ from the mean and variance of $Z_{b_2}$?**
Variance Comparison for the Standard Normal and t Distributions

Sample Observations (T)

Variance

0.00
0.50
1.00
1.50
2.00
2.50
3.00
3.50

0.00
0.50
1.00
1.50
2.00
2.50
3.00
3.50

Sample Observations (T)

Variance

t-distribution variance

Standard normal distribution variance
Figure 4.6 The standard normal and $t(3)$ probability density functions.

Interval Estimator Based on $t_{b_2}$ ($\sigma^2$ unknown)

In general, we know that we can pick critical values $c$ and $-c$ such that,

$$P[t_{T-2} \geq c] = P[t_{T-2} \leq -c] = \frac{\alpha}{2}$$

or,

$$P[t_{T-2} \geq c] + P[t_{T-2} \leq -c] = \alpha$$

and,

$$P[-c \leq t_{T-2} \leq c] = 1 - \alpha$$

where $\alpha$ is the probability

From a $t$-table, we know that when $T=40$ and $\alpha=0.05$, then $c=2.024$ and $-c=-2.024$ the probability

This can be stated in mathematical form as,

$$P[t_{T-2} \geq 2.024] = P[t_{T-2} \leq -2.024] = \frac{0.05}{2} = 0.025$$

or,

$$P[t_{T-2} \geq 2.024] + P[t_{T-2} \leq -2.024] = 0.025 + 0.025 = 0.05$$

and,

$$P[-2.024 \leq t_{T-2} \leq 2.024] = 1 - 0.05 = 0.95$$
Since $t_{b_2}$ is a $t$-distributed random variable, we can write,

$$P[-2.024 \leq t_{b_2} \leq 2.024] = 0.95$$

Substituting for $t_{b_2}$,

$$P[-2.024 \leq \frac{b_2 - \beta_2}{\sqrt{\text{var}(b_2)}} \leq 2.024] = 0.95$$

Multiply the inequality in the brackets by $\sqrt{\text{var}(b_2)}$,

$$P[-2.024 \cdot \sqrt{\text{var}(b_2)} \leq b_2 - \beta_2 \leq 2.024 \cdot \sqrt{\text{var}(b_2)}] = 0.95$$

Now, subtract $b_2$ from each term,

$$P[-b_2 - 2.024 \cdot \sqrt{\text{var}(b_2)} \leq -\beta_2 \leq -b_2 + 2.024 \cdot \sqrt{\text{var}(b_2)}] = 0.95$$

Next, multiply the inequality by -1 to reverse the direction of the inequality,

$$P[b_2 + 2.024 \cdot \sqrt{\text{var}(b_2)} \geq \beta_2 \geq b_2 - 2.024 \cdot \sqrt{\text{var}(b_2)}] = 0.95$$
Finally, "flip" the entire inequality as follows,

\[ P[b_2 - 2.024 \cdot \sqrt{\text{var}(b_2)} \leq \beta_2 \leq b_2 + 2.024 \cdot \sqrt{\text{var}(b_2)}] = 0.95 \]

This is the interval estimator for \( \beta_2 \) at the 95 percent confidence level [\((1-0.95)100\)] when \( \sigma^2 \) must be estimated.

We can generalize the interval estimator to any desired confidence level as follows,

\[ P[b_2 - t_{\alpha/2, T-2} \cdot \sqrt{\text{var}(b_2)} \leq \beta_2 \leq b_2 + t_{\alpha/2, T-2} \cdot \sqrt{\text{var}(b_2)}] = 1 - \alpha \]

where \( t_{\alpha/2, T-2} \) is the appropriate critical value from a \( t \)-table for \( \alpha \) percent tail probability and \( T-2 \) dof (\( \alpha \) is the sum of lower and upper tail probability).

Often \( t_{b_2} \) is expressed in the following form,

\[ t_{b_2} = \frac{b_2 - \beta_2}{s\hat{e}(b_2)} \]

and the resulting interval estimator is,

\[ P[b_2 - t_{\alpha/2, T-2} \cdot s\hat{e}(b_2) \leq \beta_2 \leq b_2 + t_{\alpha/2, T-2} \cdot s\hat{e}(b_2)] = 1 - \alpha \]
Interpretation of Interval Estimator when $\sigma^2$ is Unknown

Interval is random because $b_2$ and $\text{vâr}(b_2)$ are random variables before sampling

- **Location** and **width** of interval are variable

- **Width** of interval is variable because $\text{vâr}(b_2)$ is a random variable

- **Location** of interval is variable because $b_2$ is a random variable

Correct meaning of interval estimator

- In repeated sampling, we expect 95% of interval estimates to contain $\beta_2$

- If we use the interval estimator to compute a “large” number of interval predictions like $b_2 \pm 2.024 \sqrt{\text{vâr}(b_2)}$, 95% of these intervals will contain $\beta_2$
Interval Estimator vs. Interval Estimate

Before sampling takes place, the interval estimator, or rule, is

\[ b_2 \pm t_{\alpha/2, T-2} \sqrt{\text{var}(b_2)} \]

When \( b_2 \) and \( \text{var}(b_2) \) are estimates for a single sample,

- The interval \( b_2 \pm t_{\alpha/2, T-2} \sqrt{\text{var}(b_2)} \) is a \( 100(1-\alpha)\% \) interval estimate of the true parameter \( \beta_2 \)

- An interval estimate is a pair of numbers, e.g. (0.10, 0.25), computed for a particular sample, with the numbers indicating the estimated endpoints of the interval
Interval Estimate for the Household Expenditure Example

For the sample of 40 households,

\[ b_2 = 0.2323 \quad \text{vár}(b_2) = 0.00306 \]

Assuming a 95% confidence level (\(\alpha = 0.05\)), the critical value for the \(t\)-distribution is \(t_{0.025, 38} = 2.024\)

Based on this information, the 95% confidence interval estimate for \(\beta_2\) is,

\[ 0.2323 \pm 2.024 \cdot \sqrt{0.00306} \]

or,

\[ (0.12 \leq \beta_2 \leq 0.34) \]
Sample Regression Output from Excel

SUMMARY OUTPUT

<table>
<thead>
<tr>
<th>Regression Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
<td>0.563096017</td>
</tr>
<tr>
<td>R Square</td>
<td>0.317077125</td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.29910547</td>
</tr>
<tr>
<td>Standard Error</td>
<td>6.844922384</td>
</tr>
<tr>
<td>Observations</td>
<td>40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANOVA</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td></td>
</tr>
<tr>
<td>MS</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Significance F</td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>1</td>
</tr>
<tr>
<td>Residual</td>
<td>38</td>
</tr>
<tr>
<td>Total</td>
<td>39</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>t Stat</td>
<td></td>
</tr>
<tr>
<td>P-value</td>
<td></td>
</tr>
<tr>
<td>Lower 95%</td>
<td></td>
</tr>
<tr>
<td>Upper 95%</td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>7.383217543</td>
</tr>
<tr>
<td>X Variable 1</td>
<td>0.23225333</td>
</tr>
</tbody>
</table>
**Interpretation Guidelines:**

In repeated sampling, we expect 95% of interval estimates to contain $\beta_2$.

If we use the interval estimator to compute a “large” number of interval estimates like $0.2323 \pm 2.024 \cdot \sqrt{0.00306}$, 95% of these intervals will contain $\beta_2$.

It is incorrect to state,

"There is a 0.95 probability that $\beta_2$ falls in the interval (0.12, 0.34)."

- This implies $\beta_2$ is random and the interval is not, when in reality just the opposite is true!

It is also incorrect to state,

"There is a 0.95 probability that the interval (0.12, 0.34) contains $\beta_2$."

- We will never know whether $\beta_2$ falls in the estimated interval (0.12, 0.34)
- Remember that our confidence is in the estimator not the particular estimate
Compromise language,

“We are 95% confident that the interval (0.12, 0.34) contains $\beta_2$.”

where “confident” is understood to apply to the interval estimator in repeated sampling, not the (0.12, 0.34) interval estimate

HGWJ in *Undergraduate Econometrics* take a more conservative approach and argue it would only be appropriate to state,

*Based on our one sample of data, given the reliability of the interval estimate procedure, we would be “surprised” if $\beta_2$ did not fall in the interval (0.12, 0.34)*

They argue that confidence interval estimates are best understood as giving a general notion of reliability
• A “wide” interval suggests there is not much information in the sample about $\beta_2$

• A “narrow” interval suggests we have learned more about the true value of $\beta_2$

Whether an interval estimate should be considered “wide” or “narrow” depends on

• Research problem under study

• Economic magnitude of interval
Specific Interpretation

The 95% interval estimate for $\beta_2$, the response of food expenditure to income, is given by

$$(0.12, 0.34)$$

- Indicates we are 95% confident that increasing income $1$ per week will lead to an increase in food expenditure between $12$ and $34$ cents per week.

- From an economic perspective, this is a relatively narrow interval and it is informative.

- Another way of describing this situation is to say that the point estimate of $b_2 = 0.2323$ is reliable.
A Monte Carlo Experiment Using the Simple Linear Regression Model

We noted earlier that it is often difficult to gain an intuitive understanding of the sampling properties of an estimator.

Sampling experiments, or Monte Carlo simulations, can be very helpful in understanding sampling properties.

We will use the following linear statistical model to generate samples of data,

\[ y_i = \beta_1 + \beta_2 x_i + e_i \]

Next, we are going to pretend that the sample estimates from the household expenditure problem are the true population parameters (remember, we could assume any values, these are just convenient),

\[ \beta_1 = 7.3832 \quad \beta_2 = 0.2323 \quad \sigma^2 = 46.853 \]

Finally, we will need values for \( x_t \)

- Assumed to be the same as the 40 in the original sample
The assumptions of the Monte Carlo experiment can be summarized as,

- The linear statistical model

\[ y_t = 7.3832 + 0.2323x_t + e_t \]

- \( e_t \sim N(0, 46.853) \)

- \( x_t \) from the household data sample

- Sample size of \( T=40 \)

Based on these assumptions, we know that

\[
\text{var}(b_1) = 16.0669 \\
\text{var}(b_2) = 0.00306
\]

and

\[
b_1 \sim N(7.3832, 16.0669) \\
b_2 \sim N(0.2323, 0.00306)
\]
Simulation Steps

1. Compute the expected value for $y_t$ at each level of $x_t$ using the "true" population line of $7.383 + 0.2323x_t$

2. Randomly draw 40 error term values from the distribution $e_t \sim N(0, 46.853)$

3. Add the 40 random error values to the expected value for $y_t$ computed in step (1), which will create 40 "sample" observations on $y_t$

4. Regress the 40 "sample" observations of $y_t$ on the 40 values for $x_t$

5. Save the estimates for $b_1$, $b_2$, $\hat{\sigma}^2$, $\text{vâr}(b_1)$, and $\text{vâr}(b_2)$

6. Repeat steps (2) - (5) 1,000 times
Results for One Iteration of a Sampling Experiment with the Simple Linear Regression Model

<table>
<thead>
<tr>
<th>t</th>
<th>$x_t$</th>
<th>$7.3832 + 0.2323x_t$</th>
<th>$e_t$</th>
<th>$y_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25.83</td>
<td>13.38</td>
<td>1.01</td>
<td>14.39</td>
</tr>
<tr>
<td>2</td>
<td>34.31</td>
<td>15.35</td>
<td>-1.74</td>
<td>13.62</td>
</tr>
<tr>
<td>3</td>
<td>42.50</td>
<td>17.26</td>
<td>2.23</td>
<td>19.49</td>
</tr>
<tr>
<td>4</td>
<td>46.75</td>
<td>18.24</td>
<td>-4.00</td>
<td>14.24</td>
</tr>
<tr>
<td>5</td>
<td>48.29</td>
<td>18.60</td>
<td>-11.70</td>
<td>6.90</td>
</tr>
<tr>
<td>6</td>
<td>48.77</td>
<td>18.71</td>
<td>9.36</td>
<td>28.07</td>
</tr>
<tr>
<td>7</td>
<td>49.65</td>
<td>18.92</td>
<td>-2.68</td>
<td>16.24</td>
</tr>
<tr>
<td>8</td>
<td>51.94</td>
<td>19.45</td>
<td>5.98</td>
<td>25.42</td>
</tr>
<tr>
<td>9</td>
<td>54.33</td>
<td>20.00</td>
<td>11.48</td>
<td>31.48</td>
</tr>
<tr>
<td>10</td>
<td>54.87</td>
<td>20.13</td>
<td>-20.85</td>
<td>-0.72</td>
</tr>
<tr>
<td>11</td>
<td>56.46</td>
<td>20.50</td>
<td>4.00</td>
<td>24.50</td>
</tr>
<tr>
<td>12</td>
<td>58.83</td>
<td>21.05</td>
<td>0.82</td>
<td>21.86</td>
</tr>
<tr>
<td>13</td>
<td>59.13</td>
<td>21.12</td>
<td>8.20</td>
<td>29.32</td>
</tr>
<tr>
<td>14</td>
<td>60.73</td>
<td>21.49</td>
<td>1.78</td>
<td>23.27</td>
</tr>
<tr>
<td>15</td>
<td>61.12</td>
<td>21.58</td>
<td>-5.37</td>
<td>16.21</td>
</tr>
<tr>
<td>16</td>
<td>63.10</td>
<td>22.04</td>
<td>3.59</td>
<td>25.63</td>
</tr>
<tr>
<td>17</td>
<td>65.96</td>
<td>22.71</td>
<td>8.15</td>
<td>30.86</td>
</tr>
<tr>
<td>18</td>
<td>66.40</td>
<td>22.81</td>
<td>-8.55</td>
<td>14.26</td>
</tr>
<tr>
<td>19</td>
<td>70.42</td>
<td>23.74</td>
<td>1.77</td>
<td>25.51</td>
</tr>
<tr>
<td>20</td>
<td>70.48</td>
<td>23.76</td>
<td>3.24</td>
<td>26.99</td>
</tr>
<tr>
<td>21</td>
<td>71.98</td>
<td>24.10</td>
<td>-10.81</td>
<td>13.29</td>
</tr>
<tr>
<td>22</td>
<td>72.00</td>
<td>24.11</td>
<td>2.18</td>
<td>26.29</td>
</tr>
<tr>
<td>23</td>
<td>72.23</td>
<td>24.16</td>
<td>3.98</td>
<td>28.14</td>
</tr>
<tr>
<td>24</td>
<td>72.23</td>
<td>24.16</td>
<td>9.02</td>
<td>33.19</td>
</tr>
<tr>
<td>25</td>
<td>73.44</td>
<td>24.44</td>
<td>6.84</td>
<td>31.29</td>
</tr>
<tr>
<td>26</td>
<td>74.25</td>
<td>24.63</td>
<td>-2.96</td>
<td>21.67</td>
</tr>
<tr>
<td>27</td>
<td>74.77</td>
<td>24.75</td>
<td>-7.98</td>
<td>16.77</td>
</tr>
<tr>
<td>28</td>
<td>76.33</td>
<td>25.11</td>
<td>-15.91</td>
<td>9.21</td>
</tr>
<tr>
<td>29</td>
<td>81.02</td>
<td>26.20</td>
<td>10.89</td>
<td>37.09</td>
</tr>
<tr>
<td>30</td>
<td>81.85</td>
<td>26.40</td>
<td>4.49</td>
<td>30.89</td>
</tr>
<tr>
<td>31</td>
<td>82.56</td>
<td>26.56</td>
<td>6.74</td>
<td>33.30</td>
</tr>
<tr>
<td>32</td>
<td>83.33</td>
<td>26.74</td>
<td>3.13</td>
<td>29.87</td>
</tr>
<tr>
<td>33</td>
<td>83.40</td>
<td>26.76</td>
<td>0.12</td>
<td>26.87</td>
</tr>
<tr>
<td>34</td>
<td>91.81</td>
<td>28.71</td>
<td>-2.33</td>
<td>26.38</td>
</tr>
<tr>
<td>35</td>
<td>91.81</td>
<td>28.71</td>
<td>-10.85</td>
<td>17.86</td>
</tr>
<tr>
<td>36</td>
<td>92.96</td>
<td>28.98</td>
<td>3.74</td>
<td>32.72</td>
</tr>
<tr>
<td>37</td>
<td>95.17</td>
<td>29.49</td>
<td>0.11</td>
<td>29.61</td>
</tr>
<tr>
<td>38</td>
<td>101.40</td>
<td>30.94</td>
<td>1.95</td>
<td>32.88</td>
</tr>
<tr>
<td>39</td>
<td>114.13</td>
<td>33.90</td>
<td>-15.93</td>
<td>17.97</td>
</tr>
<tr>
<td>40</td>
<td>115.46</td>
<td>34.20</td>
<td>0.14</td>
<td>34.34</td>
</tr>
</tbody>
</table>
Results for Another Iteration of a Sampling Experiment with the Simple Linear Regression Model

<table>
<thead>
<tr>
<th></th>
<th>t</th>
<th>x_t</th>
<th>7.3832+0.2323x_t</th>
<th>e_t</th>
<th>y_t</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25.83</td>
<td>13.38</td>
<td>-4.63</td>
<td>8.75</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>34.31</td>
<td>15.35</td>
<td>9.55</td>
<td>24.90</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>42.50</td>
<td>17.26</td>
<td>-2.35</td>
<td>14.91</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>46.75</td>
<td>18.24</td>
<td>-0.38</td>
<td>17.86</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>48.29</td>
<td>18.60</td>
<td>6.29</td>
<td>24.89</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>48.77</td>
<td>18.71</td>
<td>-7.54</td>
<td>11.17</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>49.65</td>
<td>18.92</td>
<td>0.44</td>
<td>19.36</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>51.94</td>
<td>19.45</td>
<td>4.80</td>
<td>24.25</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>54.33</td>
<td>20.00</td>
<td>-7.70</td>
<td>12.30</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>54.87</td>
<td>20.13</td>
<td>-0.76</td>
<td>19.37</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>56.46</td>
<td>20.50</td>
<td>3.60</td>
<td>24.10</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>58.83</td>
<td>21.05</td>
<td>-6.52</td>
<td>14.52</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>59.13</td>
<td>21.12</td>
<td>-11.21</td>
<td>9.91</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>60.73</td>
<td>21.49</td>
<td>-2.85</td>
<td>18.64</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>61.12</td>
<td>21.58</td>
<td>5.41</td>
<td>26.99</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>63.10</td>
<td>22.04</td>
<td>3.93</td>
<td>25.97</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>65.96</td>
<td>22.71</td>
<td>-23.65</td>
<td>-0.95</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>66.40</td>
<td>22.81</td>
<td>4.37</td>
<td>27.18</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>70.42</td>
<td>23.74</td>
<td>5.46</td>
<td>29.21</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>70.48</td>
<td>23.76</td>
<td>-4.36</td>
<td>19.40</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>71.98</td>
<td>24.10</td>
<td>-5.01</td>
<td>19.10</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>72.00</td>
<td>24.11</td>
<td>7.48</td>
<td>31.59</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>72.23</td>
<td>24.16</td>
<td>3.27</td>
<td>27.43</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>72.23</td>
<td>24.16</td>
<td>1.44</td>
<td>25.60</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>73.44</td>
<td>24.44</td>
<td>5.72</td>
<td>30.17</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>74.25</td>
<td>24.63</td>
<td>8.99</td>
<td>33.63</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>74.77</td>
<td>24.75</td>
<td>-2.83</td>
<td>21.93</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>76.33</td>
<td>25.11</td>
<td>-5.03</td>
<td>20.09</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>81.02</td>
<td>26.20</td>
<td>-9.05</td>
<td>17.16</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>81.85</td>
<td>26.40</td>
<td>-11.77</td>
<td>14.63</td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>82.56</td>
<td>26.56</td>
<td>2.74</td>
<td>29.30</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>83.33</td>
<td>26.74</td>
<td>4.18</td>
<td>30.92</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>83.40</td>
<td>26.76</td>
<td>-2.05</td>
<td>24.71</td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>91.81</td>
<td>28.71</td>
<td>-0.57</td>
<td>28.14</td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>91.81</td>
<td>28.71</td>
<td>1.41</td>
<td>30.12</td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>92.96</td>
<td>28.98</td>
<td>-7.36</td>
<td>21.61</td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>95.17</td>
<td>29.49</td>
<td>-5.13</td>
<td>24.37</td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>101.40</td>
<td>30.94</td>
<td>4.48</td>
<td>35.41</td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>114.13</td>
<td>33.90</td>
<td>-6.43</td>
<td>27.46</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>115.46</td>
<td>34.20</td>
<td>0.64</td>
<td>34.84</td>
<td></td>
</tr>
</tbody>
</table>
### Table 6.1
Point Estimates of $\beta_1$, $\beta_2$, $\sigma^2$, $\text{var}(b_1)$, and $\text{var}(b_2)$ for 20 Samples of Size 40

<table>
<thead>
<tr>
<th>Sample Number</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$\hat{\sigma}^2$</th>
<th>$\text{var}(b_1)$</th>
<th>$\text{var}(b_2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8.505</td>
<td>.2399</td>
<td>73.41</td>
<td>25.17</td>
<td>.00479</td>
</tr>
<tr>
<td>2</td>
<td>10.348</td>
<td>.2114</td>
<td>60.59</td>
<td>20.78</td>
<td>.00395</td>
</tr>
<tr>
<td>3</td>
<td>6.613</td>
<td>.2353</td>
<td>30.46</td>
<td>10.45</td>
<td>.00199</td>
</tr>
<tr>
<td>4</td>
<td>13.812</td>
<td>.1383</td>
<td>55.35</td>
<td>18.98</td>
<td>.00361</td>
</tr>
<tr>
<td>5</td>
<td>4.824</td>
<td>.2815</td>
<td>50.79</td>
<td>17.42</td>
<td>.00331</td>
</tr>
<tr>
<td>6</td>
<td>9.086</td>
<td>.1748</td>
<td>70.79</td>
<td>24.27</td>
<td>.00462</td>
</tr>
<tr>
<td>7</td>
<td>11.897</td>
<td>.1597</td>
<td>36.32</td>
<td>12.45</td>
<td>.00237</td>
</tr>
<tr>
<td>8</td>
<td>12.168</td>
<td>.1608</td>
<td>66.89</td>
<td>22.94</td>
<td>.00437</td>
</tr>
<tr>
<td>9</td>
<td>2.596</td>
<td>.2977</td>
<td>49.07</td>
<td>16.83</td>
<td>.00320</td>
</tr>
<tr>
<td>10</td>
<td>5.763</td>
<td>.2737</td>
<td>54.14</td>
<td>18.57</td>
<td>.00353</td>
</tr>
<tr>
<td>11</td>
<td>8.312</td>
<td>.2042</td>
<td>36.09</td>
<td>12.38</td>
<td>.00235</td>
</tr>
<tr>
<td>12</td>
<td>3.482</td>
<td>.2816</td>
<td>54.96</td>
<td>18.85</td>
<td>.00359</td>
</tr>
<tr>
<td>13</td>
<td>10.498</td>
<td>.1889</td>
<td>51.79</td>
<td>17.76</td>
<td>.00338</td>
</tr>
<tr>
<td>14</td>
<td>9.930</td>
<td>.1921</td>
<td>49.75</td>
<td>17.06</td>
<td>.00325</td>
</tr>
<tr>
<td>15</td>
<td>10.876</td>
<td>.2027</td>
<td>51.53</td>
<td>17.67</td>
<td>.00336</td>
</tr>
<tr>
<td>16</td>
<td>3.533</td>
<td>.2629</td>
<td>48.22</td>
<td>16.53</td>
<td>.00315</td>
</tr>
<tr>
<td>17</td>
<td>11.764</td>
<td>.1707</td>
<td>43.34</td>
<td>14.86</td>
<td>.00283</td>
</tr>
<tr>
<td>18</td>
<td>7.152</td>
<td>.2449</td>
<td>42.49</td>
<td>14.57</td>
<td>.00277</td>
</tr>
<tr>
<td>19</td>
<td>5.688</td>
<td>.2613</td>
<td>51.19</td>
<td>17.55</td>
<td>.00334</td>
</tr>
<tr>
<td>20</td>
<td>2.132</td>
<td>.3047</td>
<td>39.26</td>
<td>13.46</td>
<td>.00256</td>
</tr>
</tbody>
</table>

Figure 6.3 Frequency distribution of the estimates for $\beta_1$ from 1000 samples of size 40.

### Table 7.1  Point and Interval Estimates for $\beta_1$ and $\beta_2$ for 10 Samples of Data for the Monte Carlo Experiment Presented in Section 6.5

<table>
<thead>
<tr>
<th>Sample Number</th>
<th>$\beta_1 = 7.3832$</th>
<th>$\beta_2 = .2323$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Point</td>
<td>Interval</td>
</tr>
<tr>
<td>1</td>
<td>8.505</td>
<td>-1.650</td>
</tr>
<tr>
<td>2</td>
<td>10.348</td>
<td>1.121</td>
</tr>
<tr>
<td>3</td>
<td>6.613</td>
<td>0.071</td>
</tr>
<tr>
<td>4</td>
<td>13.812</td>
<td>4.994</td>
</tr>
<tr>
<td>5</td>
<td>4.824</td>
<td>-3.623</td>
</tr>
<tr>
<td>6</td>
<td>9.086</td>
<td>-0.886</td>
</tr>
<tr>
<td>7</td>
<td>11.897</td>
<td>4.755</td>
</tr>
<tr>
<td>8</td>
<td>12.168</td>
<td>2.474</td>
</tr>
<tr>
<td>9</td>
<td>2.596</td>
<td>-5.707</td>
</tr>
<tr>
<td>10</td>
<td>5.763</td>
<td>-2.958</td>
</tr>
</tbody>
</table>

What Have We Learned from the Monte Carlo Experiment?

• **Point estimates** vary considerably from sample to sample

• Despite variability, the point estimates on average are about equal to the true parameter value

• Sampling distributions for $b_1$ and $b_2$ are approximately **normal**

• Interval estimates vary substantially in **width** and **location** from sample to sample

• As theory would predict about 95% of interval estimates **contain** the true parameter values for $\beta_1$ and $\beta_2$
Final Point

• Monte Carlo simulations demonstrate the nature of sampling distributions

• In practice, we do not need to obtain repeated samples to work out how the estimators may perform in repeated samples!!

• Estimated variances, from just one sample, provide "good" estimates of this information

• A truly remarkable result, but remember that it depends on the statistical model being correctly specified