Lectures 5 & 6

Simple Linear Regression

Oscar Volpe

10/11/2021 & 10/13/2021

Motivation

Suppose that we have data $\{X_i, Y_i\}_{i=1}^n$ on Y and X. We may want to:

- predict Y_i from X_i
- understand how X_i causes Y_i

In either case, we call X_i the independent variable (*regressor*). We call Y_i the dependent variable (*regressand*). A simple linear model is:

$$Y_i = \beta_0 + \beta_1 X_i + U_i,$$

where β_0 is the *intercept* and β_1 is the *slope coefficient* for this model.

The error term U_i exists because (X_i, Y_i) do not lie on a straight line.

- Why not? Omitted regressors, mis-measurement, nonlinearities, etc.
- How we interpret coefficients (β_0, β_1) and error U_i depends on how we define the linear model, i.e. is it causal or purely predictive?

Best Linear Predictor

Suppose we want the *best linear predictor* of Y given X. We minimize:

$$\text{MSE}(b_0, b_1) = E\big([Y - (b_0 + b_1 X)]^2\big)$$

Since this problem is convex in b_0 and b_1, we take first order conditions:

$$\frac{\partial \text{MSE}(b_0, b_1)}{\partial b_0} = -2E(Y - b_0 - b_1 X) = 0$$
$$\frac{\partial \text{MSE}(b_0, b_1)}{\partial b_1} = -2E(X[Y - b_0 - b_1 X]) = 0$$

The solution (β_0, β_1) to this problem corresponds to the intercept and slope of the *best linear predictor* of Y given X. See the next slide!

- *Note:* we do not assume that $E(Y|X)$ is *linear*. The solution does give us the *best linear approximation* to the conditional expectation.

Solving for (β_0, β_1)

We have two optimality conditions:

$$\frac{\partial \text{MSE}(b_0, b_1)}{\partial b_0} = -2E(Y - b_0 - b_1 X) = 0$$
$$\frac{\partial \text{MSE}(b_0, b_1)}{\partial b_1} = -2E(X[Y - b_0 - b_1 X]) = 0$$

Solving the first equation, we obtain an expression for β_0:

$$\beta_0 = E(Y) - \beta_1 E(X)$$

Plugging this into the second equation, we can solve for β_1:

$$E(X[Y - E(Y) - \beta_1(X - E(X))]) = 0$$
$$\Rightarrow \beta_1 = \frac{E(X[Y - E(Y)])}{E(X[X - E(X)])} = \frac{E(XY) - E(X)E(Y)}{E(X^2) - E(X)E(X)} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

Error Restrictions

Noting that $U = Y - \beta_0 - \beta_1 X$, our first order conditions imply:

$$E(U) = E(Y - \beta_0 - \beta_1 X) = 0$$
$$E(XU) = E(X[Y - \beta_0 - \beta_1 X]) = 0$$

So, if we interpret $\beta_0 + \beta_1 X$ as the best linear predictor (BLP) of Y, then:

$$E(U) = 0 \quad \text{and} \quad E(XU) = 0$$

So, X and U are uncorrelated: $\text{Cov}(X, U) = E(XU) - E(X)E(U) = 0$.

- Under these assumptions, we say $\beta_0 + \beta_1 X = \text{BLP}(Y|X)$.
- Importantly, BLP does not imply *best predictor* of Y given X, which would come from minimizing the mean squared error $E([Y - g(X)]^2)$.

Special Case: Linear Conditional Expectation

What if $E(Y|X)$ is actually a linear function of X? In this case, we write:

$$E(Y|X) = \beta_0 + \beta_1 X$$

Note: this is a far stronger requirement than best linear predictor. The implication of this second interpretation would be that:

$$E(U|X) = E(Y - [\beta_0 + \beta_1 X]|X) = E(Y|X) - E(Y|X) = 0$$

Using the Law of Iterated Expectations, we can show that:

$$E(U) = 0 \quad \text{and} \quad E(XU) = 0$$

The conditional moment restriction $E(U|X) = 0$ is stronger than both unconditional moment restrictions for the best linear predictor case.

- *Note:* if X is binary, then $E(Y|X)$ can be written as a linear function. In general, though, $E(Y|X)$ is not linear, so $E(Y|X) \neq \text{BLP}(Y|X)$.

Defining Causal Relationships

Assume that $Y = g(X, U)$, where X is some observed determinant of Y. If we assume the relationship is linear, i.e. $g(X, U) = \beta_0 + \beta_1 X + U$, then:

$$\frac{\partial g(X, U)}{\partial X} = \beta_1,$$

in which case β_1 is interpreted as the *causal effect* of X on Y.

Here, $E(U)$ need not equal zero, but we can normalize it so that it is zero:

$$\beta_0^{(\text{new})} = \beta_0 + E(U) \quad \text{and} \quad U^{(\text{new})} = U - E(U)$$

Do we need to assume something about $E(XU)$, $E(U)$, or $E(U|X)$? *No.*

- Defining a causal relationship between Y and X is a mental exercise.
- Writing down the causal model $Y = g(X, U)$ is a thought experiment.

Three Steps of Causal Inference

Step 1: Write Down a Model

- Define the causal relationship of interest. This requires you, the researcher, to specify a counterfactual question ("What if. . . ?"). No data needed here.
- Under your model, *causal effects* become target parameters.

Step 2: Identification

- Given your model, what can you learn about the target parameters using observed data? *Identification* maps the model and data to information about target parameters. Essentially, what can you recover from data?
- We say that a parameter is *identified* if, under the model assumptions, alternative values of the parameter imply different distributions of the data.

Step 3: Estimation

- In practice, we see finite samples drawn from the population distribution.
- How can we use these samples to estimate the target parameters?

Solving for the BLP

Suppose that we have an i.i.d. sample $\{X_i, Y_i\}_{i=1}^n$ of Y and X. Using this data, we solve a sample analogue of the least-squares problem:

$$(\hat{\beta}_0, \hat{\beta}_1) \in \underset{b_0, b_1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 \tag{1}$$

Solving this minimization problem gives us an estimator for β_1:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n X_i(Y_i - \bar{Y}_n)}{\frac{1}{n}\sum_{i=1}^n X_i(X_i - \bar{X}_n)} = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2} = \frac{\widehat{\text{Cov}(X_i, Y_i)}}{\widehat{\text{Var}(X_i)}}$$

The corresponding estimator for β_0 is $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$.

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the *ordinary least squares* (OLS) estimators.
- These estimators satisfy the first order conditions of problem (1).

Residuals

The optimality conditions from the ordinary least squares problem are:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\frac{1}{n}\sum_{i=1}^{n}X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

We define $\hat{U}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ to be the ith *residual*. It follows that:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{U}_i = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}X_i\hat{U}_i = 0$$

Define the *predicted value* (or *fitted value*) of Y_i to be $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.

- *Note:* the residuals $\{\hat{U}_i\}_{i=1}^n$ are given by $\hat{U}_i = Y_i - \hat{Y}_i$.
- We can plot the fitted regression line against the realizations of Y_i.

Interpreting OLS Coefficients

Notice that β_1 is proportional to the correlation between X and Y:

$$\beta_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \sqrt{\frac{\text{Var}(Y)}{\text{Var}(X)}} \times \rho(X, Y)$$

The more correlated X and Y are, the larger the slope β_1 will be.

Example

Suppose Y_i is income and X_i is years of schooling. You estimate:

$$Y_i = \beta_0 + \beta_1 X_i + U_i$$

under the BLP assumptions. You obtain the OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

- A one unit increase in X_i is associated with an estimated $\hat{\beta}_1$ increase in Y_i. Importantly, $\hat{\beta}_1$ does not estimate a causal effect of X_i on Y_i.
- If $\beta_1 > 1$, then the correlation between X_i and Y_i should be positive.

Coefficient of Determination

Suppose we want to measure how well $\{\hat{Y}_i\}_{i=1}^n$ approximates $\{Y_i\}_{i=1}^n$. The *coefficient of determination* (or R-squared) is defined to be:

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2} = 1 - \frac{\frac{1}{n}\sum_{i=1}^n \hat{U}_i^2}{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}$$

We can also write $R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}}$, where:

- $\text{TSS} = \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$
- $\text{ESS} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2$
- $\text{SSR} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{U}_i^2$

Intuitively, if the model fits the data well, then much of the variation in Y_i is captured by the variation in $\hat{Y}_i$. In this case, the R-squared is large.

Decomposing the *TSS*

Note that we can decompose the total sum of squares (TSS) as:

$$\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_n + \hat{U}_i)^2$$

$$= \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y}_n)^2}_{\text{ESS}} + 2\sum_{i=1}^{n}\hat{U}_i(\hat{Y}_i - \bar{Y}_n) + \underbrace{\sum_{i=1}^{n}\hat{U}_i^2}_{\text{SSR}}$$

Note that the middle term equals zero under the BLP assumptions, since:

$$\sum_{i=1}^{n}\hat{U}_i(\hat{Y}_i - \bar{Y}_n) = \hat{\beta}_0\sum_{i=1}^{n}\hat{U}_i + \hat{\beta}_1\sum_{i=1}^{n}X_i\hat{U}_i - \bar{Y}_n\sum_{i=1}^{n}\hat{U}_i = 0$$

It follows that TSS = ESS + SSR, which implies: $R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}}$.

Interpreting the R-Squared Term

In the simple linear regression model, $0 \leq R^2 \leq 1$.

- $R^2 = 1$ if SSR $= 0$, i.e. all data points lie on a line.
- $R^2 = 0$ if ESS $= 0$, i.e. X_i does not help us to predict Y_i.
  - $R^2 = 0 \implies \hat{\beta}_1 = 0$, i.e. the sample correlation between X and Y is zero.

Importantly, R-squared does not tell us anything about the causal relationship between X and Y. It simply measures goodness of fit.

- Recall that causality is entirely based on assumptions that you make.
- We should be very careful when interpreting the R-squared term. particularly if there is concern about the BLP assumptions holding.

Example: Regression through the Origin

Given data $\{X_i, Y_i\}_{i=1}^n$, consider the model without an intercept:

$$Y_i = \beta X_i + U_i$$

To solve for β under the least-squares interpretation, minimize:

$$\text{MSE}(b) = E([Y - bX]^2)$$

You can show $\beta = \frac{E(XY)}{E(X^2)}$. A method of moments (MoM) estimator is:

$$\hat{\beta}_n = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i}{\frac{1}{n}\sum_{i=1}^n X_i^2} = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$$

It is possible that this model fits worse than the "constant only" model, where $Y_i = \beta + U_i$. So, we can have $R^2 < 0$ if we measure R-squared by:

$$R^2 = 1 - \frac{\text{SSR}}{\text{TSS}} = 1 - \frac{\frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\beta}_n X_i)^2}{\frac{1}{n}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}$$

Unbiasedness of $(\hat{\beta}_0, \hat{\beta}_1)$

Consider our ordinary least squares (OLS) estimators for β_1 and β_0:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - \bar{Y}_n)}{\frac{1}{n}\sum_{i=1}^{n} X_i(X_i - \bar{X}_n)} \quad \text{and} \quad \hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$$

When should we expect that $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased estimators?

- In general, $\hat{\beta}_1$ and $\hat{\beta}_0$ are *not* unbiased for β_1 and β_0 (respectively).
- If $E(U_i|X_i) = 0$, then we can show $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased estimators.
  - *Note:* $E(U_i|X_i) = 0$ is implied by assuming $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$.

Theorem (Unbiasedness of the OLS Estimator)

Let $\{X_i, Y_i\}_{i=1}^{n}$ be an i.i.d. sample, and let $Y_i = \beta_0 + \beta_1 X_i + U_i$ be the model under consideration. If there is variation in X_i within the sample and if $E(U_i|X_i) = 0$, then the OLS estimators $(\hat{\beta}_0, \hat{\beta}_1)$ are unbiased.

Deriving the Bias in $\hat{\beta}_1$ (*Part 1*)

To show that $E(U_i|X_i) = 0$ guarantees unbiasedness for $\hat{\beta}_1$, we write:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - \bar{Y}_n) &= \frac{1}{n}\sum_{i=1}^{n} X_i\Big([\beta_0 + \beta_1 X_i + U_i] - \frac{1}{n}\sum_{j=1}^{n}[\beta_0 + \beta_1 X_i + U_i]\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i\Big(\beta_0 + \beta_1 X_i + U_i - \beta_0 - \beta_1 \bar{X}_n - \bar{U}_n\Big) \\
&= \frac{1}{n}\sum_{i=1}^{n} \beta_1 X_i(X_i - \bar{X}_n) + \frac{1}{n}\sum_{i=1}^{n} X_i(U_i - \bar{U}_n)
\end{aligned}$$

Rewriting the numerator of $\hat{\beta}_1$ in this way, we have:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - \bar{Y}_n)}{\frac{1}{n}\sum_{i=1}^{n} X_i(X_i - \bar{X}_n)} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n} X_i(U_i - \bar{U}_n)}{\frac{1}{n}\sum_{i=1}^{n} X_i(X_i - \bar{X}_n)}$$

Deriving the Bias in $\hat{\beta}_1$ (*Part 2*)

Take the conditional expectation $E(\hat{\beta}_1|X_1,\ldots,X_n)$ as:

$$\begin{aligned}
E(\hat{\beta}_1|X_1,\ldots,X_n) &= \beta_1 + E\left(\frac{\frac{1}{n}\sum_{i=1}^{n} X_i(U_i-\bar{U}_n)}{\frac{1}{n}\sum_{i=1}^{n} X_i(X_i-\bar{X}_n)}\middle|X_1,\ldots,X_n\right) \\
&= \beta_1 + \frac{E\left(\frac{1}{n}\sum_{i=1}^{n} X_i(U_i-\bar{U}_n)\middle|X_1,\ldots,X_n\right)}{\frac{1}{n}\sum_{i=1}^{n} X_i(X_i-\bar{X}_n)} \\
&= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n} X_i E\left((U_i-\bar{U}_n)\middle|X_1,\ldots,X_n\right)}{\frac{1}{n}\sum_{i=1}^{n} X_i(X_i-\bar{X}_n)} = \beta_1,
\end{aligned}$$

where the last equality holds because our sample is *i.i.d.* and $E(U_i|X_i)=0$. Finally, by the Law of Iterated Expectations, we write:

$$E(\hat{\beta}_1) = E(E(\hat{\beta}_1|X_1,\ldots,X_n)) = \beta_1$$

Deriving the Bias in $\hat{\beta}_0$

To show that $E(U_i|X_i) = 0$ guarantees unbiasedness for $\hat{\beta}_0$, we write:

$$\begin{aligned}
E(\hat{\beta}_0|X_1,\ldots,X_n) &= E(\bar{Y}_n - \hat{\beta}_1\bar{X}_n|X_1,\ldots,X_n) \\
&= E(\bar{Y}_n|X_1,\ldots,X_n) - E(\hat{\beta}_1|X_1,\ldots,X_n)\bar{X}_n \\
&= E(\beta_0 + \beta_1\bar{X}_n + \bar{U}_n|X_1,\ldots,X_n) - \beta_1\bar{X}_n \\
&= \beta_0 + \beta_1\bar{X}_n + E(\bar{U}_n|X_1,\ldots,X_n) - \beta_1\bar{X}_n \\
&= \beta_0 + E(\bar{U}_n|X_1,\ldots,X_n) = \beta_0,
\end{aligned}$$

where the last equality holds because our sample is *i.i.d.* and $E(U_i|X_i) = 0$. Finally, by the Law of Iterated Expectations, we write:

$$E(\hat{\beta}_0) = E(E(\hat{\beta}_0|X_1,\ldots,X_n)) = \beta_0$$

1. Interpretation of Least Squares
   - Predictive Interpretation
   - Causal Interpretation
2. Least Squares Estimators
   - OLS Estimation
   - Goodness of Fit
3. **Properties of OLS**
   - Unbiasedness
   - **Consistency**
   - Gauss-Markov Theorem
4. *Example:* Hypothesis Testing for OLS

Consistency of $(\hat{\beta}_0, \hat{\beta}_1)$

Can we show that $(\hat{\beta}_0, \hat{\beta}_1)$ converge (in a "$\xrightarrow{p}$" sense) to (β_0, β_1)?

- *Yes.* In fact, we do not even need to assume $E(U_i|X_i) = 0$.
- Consistency arguments follow from the WLLN and the CMT.

Theorem (Consistency of the OLS Estimator)

Let $\{X_i, Y_i\}_{i=1}^n$ be an i.i.d. sample, and let $Y_i = \beta_0 + \beta_1 X_i + U_i$ be the model under consideration. If there is variation in $0 < Var(X_i) < \infty$, then the OLS estimators $(\hat{\beta}_0, \hat{\beta}_1)$ are consistent for (β_0, β_1), respectively.

Proof. See the next slide.

Deriving Limits of Probability

How do we show that $\hat{\beta}_1 \xrightarrow{p} \beta_1$ and $\hat{\beta}_0 \xrightarrow{p} \beta_0$? First, write:

$$\hat{\beta}_1 = \frac{\widehat{\text{Cov}(X_i, Y_i)}}{\widehat{\text{Var}(X_i)}}, \quad \text{where:} \quad \begin{array}{c} \widehat{\text{Cov}(X_i, Y_i)} \xrightarrow{p} \text{Cov}(X_i, Y_i) \\ \widehat{\text{Var}(X_i)} \xrightarrow{p} \text{Var}(X_i) \end{array}$$

Therefore, as long as $0 < \text{Var}(X_i) < \infty$, the CMT guarantees that:

$$\hat{\beta}_1 = \frac{\widehat{\text{Cov}(X_i, Y_i)}}{\widehat{\text{Var}(X_i)}} \xrightarrow{p} \frac{\text{Cov}(X_i, Y_i)}{\text{Var}(X_i)} = \beta_1$$

Similarly, we can show consistency of $\hat{\beta}_0$ for β_0 by writing:

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n, \quad \text{where:} \quad \begin{array}{c} \bar{Y}_n \xrightarrow{p} E(Y_i) \\ \bar{X}_n \xrightarrow{p} E(X_i) \end{array} \quad \text{and} \quad \hat{\beta}_1 \xrightarrow{p} \beta_1$$

So, by the CMT, we know: $\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n \xrightarrow{p} E(Y_i) - \beta_1 E(X_i) = \beta_0$.

Homoskedasticity

Given data on X and Y, consider our simple linear regression model:

$$Y = \beta_0 + \beta_1 X + U$$

One convenient assumption to make about U is that $\text{Var}(Y|X)$ is constant.

- When $\text{Var}(Y_i|X_i) = \sigma^2$ for all i, we say the errors are *homoskedastic*.
- Intuitively, *homoskedasticity* implies that the variability in Y around the population regression line does not depend on the value of X.

Equivalently, the errors are *homoskedastic* if $\text{Var}(U|X) = \sigma^2$, since:

$$\text{Var}(Y|X) = \text{Var}(\beta_0 + \beta_1 X + U|X) = \text{Var}(U|X)$$

Note: if *homoskedasticity* fails, then we say U is *heteroskedastic*.

Best Linear Unbiased Estimator

Consider the model $Y = \beta_0 + \beta_1 X + U$ and an *i.i.d.* sample $\{Y_i, X_i\}_{i=1}^n$.

- Suppose that our least squares assumptions are satisfied.
- Assume $\mathbb{E}(U|X) = 0$ and the error is homoskedastic: $\text{Var}(U|X) = \sigma^2$.

Under these assumptions, $(\hat{\beta}_0,\ \hat{\beta}_1)$ are the *best linear unbiased estimators*.

- *Interpretation:* $\hat{\beta}_{\text{OLS}} = (\hat{\beta}_0, \hat{\beta}_1)$ have the "smallest" variance in the class of estimators that are linear in X and unbiased for (β_0, β_1).

We seek to show that $\text{Var}(\hat{\beta}_{\text{OLS}}|X)$ is "smaller" than $\text{Var}(\tilde{\beta}|X)$, where:

- $\tilde{\beta}$ is linear, i.e. it can be written as $\tilde{\beta} = A(\{X_i\}_{i=1}^n)Y$.
- $\tilde{\beta}$ is unbiased, i.e. $\mathbb{E}[\tilde{\beta}_0|X] = \beta_0$ and $\mathbb{E}[\tilde{\beta}_1|X] = \beta_1$.

Gauss-Markov Assumptions

The following are collectively known as the *Gauss-Markov assumptions*.

(1) The model is $Y = \beta_0 + \beta_1 X + U$.

(2) We observe an *iid* sample $\{X_i, Y_i\}_{i=1}^{n}$.

(3) There is variation in X within the sample.

(4) Suppose $E(U|X) = 0$.

(5) The conditional variance is constant: $\text{Var}(U|X) = \sigma^2$.

Quick Review

- Even if (5) fails, the OLS estimators are *unbiased* if (1) – (4) hold.
- Even if (4) and (5) fail, the OLS estimators are *consistent* if the BLP assumptions hold, i.e. if (1) – (3) hold and if $E(XU) = E(U) = 0$.
- We need all these conditions, (1) – (5), for the next theorem to hold.

Stating the Theorem

Theorem (Gauss-Markov Theorem)

Suppose that the Gauss-Markov assumptions are satisfied. Then the OLS estimator $(\hat{\beta}_0, \hat{\beta}_1)$ *will be the best linear unbiased estimator for* (β_0, β_1).

The Gauss-Markov Theorem says that, under homoskedasticity, the OLS estimator is the *best* among those that are *linear* and *unbiased*.

- *best* means having the smallest conditional variance $\text{Var}(\tilde{\beta}|X)$
- the result only compares *linear* and *unbiased* estimators
- key assumption: homoskedasticity

Nonetheless, this theorem validates the use of OLS among a large class of estimators, and it also some suggests reasons to deviate from OLS.

Variances of $(\hat{\beta}_0, \hat{\beta}_1)$

Suppose that the five Gauss-Markov assumptions are satisfied.

- When $E(U|X) = 0$, we have $\text{Var}(U|X) = E(U^2|X)$.
- Under homoskedasticity, we know $\text{Var}(U|X) = \sigma^2$.

As first step, recall that the OLS estimators are:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2}$$
$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{X}_n$$

We derive the (conditional) variances of $(\hat{\beta}_0, \hat{\beta}_1)$ to be:

$$\text{Var}(\hat{\beta}_0|X_1, \dots, X_n) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}_n^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}\right]$$
$$\text{Var}(\hat{\beta}_1|X_1, \dots, X_n) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}$$

Estimating Var(U)

Right now, we can't test hypotheses about $(\hat{\beta}_0, \hat{\beta}_1)$, since σ^2 is unknown.

- How can we estimate the *error variance* $\sigma^2 = \text{Var}(U)$?
- *Idea:* find a consistent, unbiased estimator $\hat{\sigma}^2$ for σ^2, then use $\hat{\sigma}^2$ to estimate the variances $\text{Var}(\hat{\beta}_0|X_1, \ldots, X_n)$ and $\text{Var}(\hat{\beta}_1|X_1, \ldots, X_n)$.

It turns out that the estimator $\hat{\sigma}^2$ is unbiased for σ^2 when:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{U}_i^2 = \frac{\text{SSR}}{n-2}$$

- We divide by $n - 2$ to correct for bias.
- Intuitively, we have $n - 2$ in the denominator because we have two parameters (β_0 and β_1) in the regression model. More on this later!

Testing Hypotheses about β_1

Suppose n is "large". We can use asymptotic theory to test hypotheses about β_1. As a first step, recall that $\hat{\beta}_1$ can be expressed as:

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i(Y_i - \bar{Y}_n)}{\frac{1}{n}\sum_{i=1}^{n} X_i(X_i - \bar{X}_n)} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n} X_i(U_i - \bar{U}_n)}{\frac{1}{n}\sum_{i=1}^{n} X_i(X_i - \bar{X}_n)}$$

Applying the Central Limit Theorem, we find that:

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \xrightarrow{d} N\left(0, \frac{\sigma^2}{\text{Var}(X)}\right)$$

By Slutsky's theorem, we can divide by $\text{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}}$ so that:

$$T_n = \frac{(\hat{\beta}_1 - \beta_1)}{\text{se}(\hat{\beta}_1)} \xrightarrow{d} N(0,1)$$

One- and Two-Sided Tests

One-Sided Test

Suppose we want to test $H_0 : \beta_1 \leq 0$ against $H_1 : \beta_1 > 0$.

(1) Choose a significance level $\alpha \in (0, 1)$, e.g. $\alpha = 0.05$.

(2) Write down the test statistic (under H_0): $T_n = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$

(3) Reject H_0 whenever $T_n > z_{1-\alpha}$.

Two-Sided Test

Suppose we want to test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$.

(1) Choose a significance level $\alpha \in (0, 1)$, e.g. $\alpha = 0.05$.

(2) Write down the test statistic (under H_0): $T_n = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$

(3) Reject H_0 whenever $|T_n| > z_{1-\alpha/2}$.

Note that $z_{1-\alpha/2} \approx 1.96$ when $\alpha = 0.05$. We might say that "β_1 is statistically significant at the 5% level whenever $|T_n|$ lies above 1.96.

Computing p-values

Given our sample $\{X_i, Y_i\}_{i=1}^n$, test statistic T_n, and critical value $c_n(\alpha)$, the p-value is the smallest value of α at which H_0 is rejected:

$$\hat{p}_n = \inf\{\alpha \in (0,1) : T_n > c_n(\alpha)\}$$

For a two-sided test, we define $\hat{p}_n$ so that:

$$\hat{p}_n = \inf\left\{\alpha \in (0,1) : \left|\frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}\right| > z_{1-\alpha/2}\right\}$$

Idea: "shrink" α until we reach α^* satisfying $\left|\frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}\right| = z_{1-\alpha^*/2}$.

- The p-value is below 0.05 if $|T_n|$ lies above 1.96.
- *Note:* the p-value is specific to the hypothesis you are testing.

Confidence Intervals

Now suppose we want to construct a confidence interval for β_1.

$$C_n = \left[\hat{\beta}_1 - \text{se}(\hat{\beta}_1) z_{1-\alpha/2}, \hat{\beta}_1 + \text{se}(\hat{\beta}_1) z_{1-\alpha/2}\right]$$

To show that C_n is an asymptotic confidence interval for β_1, we need:

$$P(\beta_1 \in C_n) \to 1 - \alpha$$

To see why this holds, notice that:

$$\begin{aligned} P(\beta_1 \in C_n) &= P(|T_n| \leq z_{1-\alpha/2}) \\ &\to P(|Z| \leq z_{1-\alpha/2}) = 1 - \alpha \end{aligned}$$

As $n \to \infty$, the coverage probability of C_n approaches $1 - \alpha$, as desired.