Lecture 9: Hypothesis Testing

1. The Sampling Distribution of $\hat{\beta}$

It's intuitive to think about our observations $\{X_i, Y_i\}_{i=1}^n$ being random draws from probability distributions. For example, $X_i$ might be weight, which might be normally distributed in the population. Perhaps $X_i$ is the income of individual $i$, which has a log-normal population distribution. Similarly, the unobserved stochastic error term has some (unknown and unobservable) distribution. In much the same way, our estimates $\hat{\beta}_k$ for each $k = 0, 1, \ldots, K$ can be thought of as random variables with certain distributions.

An estimator is a function which, given a dataset, produces a number (or set of numbers) called an estimate. So, ordinary least squares (OLS) is an estimator, since given a dataset, we have a formula with which we can produce actual numbers for, say, $\hat{\beta}_0$ and $\hat{\beta}_1$. These numbers are the estimates. Because the observations in each dataset are random, so are functions of those datasets, including the estimates themselves.

So, the distribution underlying the data implies some distribution of the coefficients. For a fixed sample size $n$, we call the distribution of $\hat{\beta}_k$ (for any $k = 0, 1, \ldots, K$) the sampling distribution. Each sample of size $n$ results in a different value for $\hat{\beta}_k$, and the sampling distribution is the distribution of $\hat{\beta}_k$ over all possible samples of size $n$.

The concept of the sampling distribution is important because it makes clear that the estimated coefficients are themselves random variables. Being functions of other random variables, the variables which we observe in the sample, makes them random as well.

Let's consider a simple example involving estimates of a population mean $\mu$ of some variable $X$ which is uniformly distributed between 1 and 10. Suppose we take five samples of size $n = 4$ and use the sample average $\hat{\mu} = \bar{x}$ as an estimate of $\mu$ for each sample.

<table>
<thead>
<tr>
<th>Sample</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
<th>$\hat{\mu}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>5</td>
<td>5</td>
<td>9</td>
<td>6.50</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>3</td>
<td>10</td>
<td>7</td>
<td>7.50</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>3</td>
<td>9</td>
<td>4</td>
<td>6.25</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>10</td>
<td>3</td>
<td>5.00</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>2</td>
<td>2</td>
<td>4.25</td>
</tr>
</tbody>
</table>

| Average $\hat{\mu}$ | 5.90 |

Table 1. Five estimates of $\mu$ for $n = 4$

The previous example concerned a simple statistic, a sample mean, but the concept of the sampling distribution also applies to our OLS coefficients. The following example illustrates this idea in the context of regression analysis.

**Example 1.** For a more concrete example, suppose that I stand on the corner of High St. and Lane Ave. each day for a week and survey random passersby. Each day I ask 25 random students
to tell me their GPA (GPA$_i$) and how many hours per week they study (STUDY$_i$). At the end of each day, I run a regression of GPA$_i$ on STUDY$_i$ and write down the values of $\hat{\beta}_0$ and $\hat{\beta}_1$. Let’s call the estimates for day $t$ $\hat{\beta}_0^t$ and $\hat{\beta}_1^t$. The values of $\hat{\beta}_1^t$ will be slightly different each day because I have surveyed seven different sets of 25 students, each which will have slightly different responses. The distribution of GPA$_i$ and STUDY$_i$ in the student population determines the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$. What I end up with at the end of the week is a sample of seven draws (one for each day) from the sampling distributions of $\hat{\beta}_0$ and $\hat{\beta}_1$ (see, for example, Table 2).

<table>
<thead>
<tr>
<th>Day</th>
<th>$\hat{\beta}_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.15</td>
</tr>
<tr>
<td>2</td>
<td>0.23</td>
</tr>
<tr>
<td>3</td>
<td>0.13</td>
</tr>
<tr>
<td>4</td>
<td>0.20</td>
</tr>
<tr>
<td>5</td>
<td>0.27</td>
</tr>
<tr>
<td>6</td>
<td>0.21</td>
</tr>
<tr>
<td>7</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Average 0.1971
Variance 0.0746

Table 2. Seven realizations of $\hat{\beta}_1$

The fact that the OLS coefficients are themselves random means that there is uncertainty in our estimates. We would like to think that our particular estimates are representative of the overall distribution somehow, but we can only make probabilistic statements about them. This is important when trying to evaluate whether some effect is present or not by looking at the regression coefficients. For example, is the correlation between height and weight significant? What about the coefficient on FEMALE in a wage regression? Is the difference between the wages of men and women statistically significant, controlling for other explanatory variables?

Questions such as these, concerning whether a particular estimated coefficient $\hat{\beta}_k$ is nonzero in a statistical sense, are addressed by hypothesis tests of regression coefficients. Because the estimated coefficients are random, even if we obtain a positive or negative value with our particular sample (which is one of many possible samples), it doesn’t necessarily mean that the effect is statistically significant. It might be the case that our particular estimate was negative but falls in the tail of the distribution (which could have a positive mean).

2. Standard Errors

The square root of the estimated variance of $\hat{\beta}_k$ is called the standard error and is denoted SE($\hat{\beta}_k$). This is a measure of the relative precision of the estimate. Larger values indicate more uncertainty about the estimate while smaller values indicate more precise estimates.

The variance of the OLS coefficients is driven by the randomness of the data, which is in turn driven by the variance of the independent variables, $X_i$, and the stochastic error term, $\varepsilon_i$. We can estimate the variance of $X_i$ since the independent variables are observable. To calculate the variance of $\hat{\beta}_0$ or $\hat{\beta}_1$ we first need to estimate the variance of $\varepsilon_i$. Since the residuals are the
empirical analogs of the stochastic error terms, we can approximate the variance of $\varepsilon_i$ using the sample variance of $e_i$.

Software packages such as Gretl calculate and report the standard errors for all of the coefficients automatically. For example, for the regression

$$PRATE_i = \beta_0 + \beta_1 MRATE_i + \varepsilon_i,$$

given a dataset of 1000 observations, the Gretl output might be as follows:

```
Model 1: OLS estimates using the 1000 observations 1-1000
Dependent variable: PRATE

<table>
<thead>
<tr>
<th>VARIABLE</th>
<th>COEFFICIENT</th>
<th>STDERR</th>
<th>T STAT</th>
<th>P-VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>const</td>
<td>83.3404</td>
<td>0.696968</td>
<td>119.576</td>
<td>&lt;0.00001***</td>
</tr>
<tr>
<td>MRATE</td>
<td>5.49723</td>
<td>0.644386</td>
<td>8.531</td>
<td>&lt;0.00001***</td>
</tr>
</tbody>
</table>
```

Here, we have $\text{SE}(\hat{\beta}_1) = 0.644386$.

Smaller standard errors for a given coefficient indicate that the estimate is more precise. However, standard errors are not comparable across variables because they are relative to the unit of measurement.

The standard error of a particular coefficient will change with the sample size ($n$), the relative variation of $\varepsilon_i$ (which can be measured by $\sum e_i^2$), and the variance of $X_i$. Specifically:

- $\text{SE}(\hat{\beta}_k)$ will decrease as $n$ increases. As we obtain more data, the estimates will become more precise, reducing their variance and therefore, the standard deviation.

- Holding the sample size constant, $\text{SE}(\hat{\beta}_k)$ will increase as the variation in $\varepsilon_i$ increases. An error term with more variance makes it harder to obtain precise estimates of $\beta_k$.

- Finally, holding the sample size constant, when the variance of $X = (X_1, \ldots, X_K)$ increases, $\text{SE}(\hat{\beta}_k)$ will decrease. The additional variation in $X$ provides more information about the relationships of interest with which we can obtain a more accurate estimate of $\beta_k$.

### 3. Hypothesis Testing

There are two primary methods of inference about parameters of probability distributions: point estimation and hypothesis testing. Until now, this course has been concerned with the former.

The purpose of estimation is to make an informed guess about which value, out of the entire parameter space, is most likely. For estimation, given a sample $(X_i, Y_i)_{i=1}^n$ the best estimator of $\beta_1$ in the linear regression model is $\hat{\beta}_1$, as given by the OLS estimator. A point estimate is a single number such as $\hat{\beta}_1 = 49.8$.

On the other hand, hypothesis testing involves comparing particular subsets of the parameter space to determine which is more likely. For example, we might compare the set $\beta_1 = 50$ with the set $\beta_1 \neq 50$, or even $\beta_1 \leq 50$ with $\beta_1 > 50$.

In practice it is nearly impossible to prove that a particular hypothesis is correct. However, it is much easier to provide evidence to support rejecting a hypothesis. The primary focus of
hypothesis testing is to provide statistical guidance to determine when a particular piece of evidence (a measurement, such as an estimated regression coefficient) is strong enough to cause us to reject a hypothesis or not.

3.1. The Null and Alternative

A hypothesis should be specified first, before the regression is carried out. This involves clearly stating what you think is true and what you think is false. These are statements about the population parameters and must be exactly opposite (one is the logical negation of the other).

The basis of a hypothesis test is some statement about the population parameters, called the null hypothesis (denoted $H_0$), that we assume is true until there is sufficient evidence from the data to show that it is false. The null hypothesis is (typically) the opposite of what the researcher believes to be true. The goal is to provide as much support as possible against the null, or in support of the alternative hypothesis, which is (typically) what the researcher believes to be true.

We either reject the null hypothesis if there is sufficient evidence against it, or else we fail to reject the null hypothesis. This terminology is deliberate, as failing to reject is fundamentally different than accepting the null hypothesis.

There is a close connection between hypothesis testing and the criminal justice system that might be useful to keep in mind. The null hypothesis is like a defendant: it is assumed to be true until sufficient evidence is presented to indicate that it is false. The researcher plays the role of a prosecutor, attempting to show that the null hypothesis is false. If the researcher cannot provide sufficient evidence that the null hypothesis is false, then the researcher fails to reject it much like the judge will rule not guilty. If the researcher provides enough evidence against the null, then the null is rejected much like a defendant is found guilty if the evidence indicates guilt beyond a “reasonable doubt.”

Example 2. Hypotheses in econometrics do not typically specify particular values (other than zero) for a coefficient. Suppose that you think that $\beta_1$ is negative in a regression. Then the null hypothesis is

$$H_0 : \beta_1 \geq 0,$$

which is the opposite of your hypothesis. The alternative hypothesis is

$$H_A : \beta_1 < 0.$$

Example 3. To test whether $\beta_1$ is significantly different from zero in either direction, the null and alternative hypotheses would be

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

3.2. Type I and Type II Errors

There are two types of errors we must consider. We could reject $H_0$ when it’s actually true, which is called a type I error, or we could fail to reject $H_0$ when it’s actually false, a type II error. See Table 3.
Null Hypothesis

<table>
<thead>
<tr>
<th>Decision</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reject $H_0$</td>
<td>Type I error</td>
<td>Correct decision</td>
</tr>
<tr>
<td>Fail to reject $H_0$</td>
<td>Correct decision</td>
<td>Type II error</td>
</tr>
</tbody>
</table>

Table 3. Types of errors

In terms of our criminal trial analogy, we have

- $H_0$: innocent
- $H_A$: guilty

A type I error is committed when an innocent defendant is sent to prison ($H_0$ is true but is rejected). A type II error is committed when a guilty defendant goes free ($H_0$ is false but is not rejected).

3.3. Components of a hypothesis test

A statistical hypothesis test consists of five parts:

- a null hypothesis, denoted $H_0$,
- an alternative hypothesis, denoted $H_A$,
- a test statistic,
- a rejection region, and
- a significance level, $\alpha$.

Specifying these five elements defines a test.

The null hypothesis is the negation of the alternative hypothesis. Again, the goal is to support the alternative hypothesis, and the easiest way to do so is by providing evidence that the null is false. If there is enough evidence against it, as determined by the significance level $\alpha$, then we reject $H_0$. Otherwise, we fail to reject $H_0$.

It is only correct to say that we “cannot reject $H_0$.” We can never “accept $H_0$.” There will always be many null hypotheses that can’t be rejected, for example we might reject both $H_0 : \beta_1 = 50.1$ and $H_0 : \beta_1 = 50.2$. We could also fail to reject both, but it wouldn’t make sense to “accept” them both because they are contradictory!

A test statistic is some function of the data, just like our OLS estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are functions of the data. A test statistic for a particular test is chosen such that the distribution of the statistic is known. For example, to test the hypothesis $\beta_1 = 0$, we can use the estimate and the standard error to form a $t$-statistic using the estimate and the standard error:

$$t_k = \frac{\hat{\beta}_k - \beta_{H_0}}{SE(\hat{\beta}_k)}.$$  

This statistic is known to be distributed according to Student’s $t$ distribution, a distribution which is well-known and tabulated, making it easy to carry out a $t$-test.
3.4. Significance Level

The level of a test is denoted $\alpha$ and is the probability of a type I error. We must specify $\alpha$ before carrying out a test. There is a fundamental trade-off between the type I and type II error probabilities. Essentially, we could try to minimize the probability of wrongly convicting innocent people, but only by letting more guilty people go free.

Typically, researchers choose values of $\alpha$ such as 0.10, 0.05, or 0.01. These correspond to 10%, 5%, and 1% respectively.

3.5. Types of Tests

Hypothesis tests are classified as being either one-sided or two-sided tests, depending on the nature of the rejection region.

- **One-sided hypothesis test**: A test involving a “greater-than” or “less-than” alternative hypothesis. For example, $H_A : \beta_1 > 50$ or $H_A : \beta_1 < 50$. The corresponding null hypothesis is thus $H_0 : \beta_1 \leq 50$ or $H_0 : \beta_1 \geq 50$. The rejection region here is a region involving only the upper or lower tail.

- **Two-sided hypothesis test**: A test involving a “not-equal” alternative hypothesis. For example, $H_0 : \beta_1 = 50$ and $H_A : \beta_1 \neq 50$. The rejection region here is a region involving both the upper and lower tails.

3.6. Rejection Regions and Critical Values

The decision rule in hypothesis testing is to reject $H_0$ when the test statistic falls within a region called the rejection region. This region is determined according to the type of test, the sample size, and the significance level. Intuitively, the region becomes larger as the sample size increases (as our estimate of the mean becomes more precise) decreases as the level increases (as we’re more willing to be wrong). Points on the boundary of the rejection region, beyond which we reject $H_0$, are called critical values.

**Example 4.** Let’s suppose the critical value for the one-sided test $H_0 : \beta_1 = 0$, stated in terms of $\beta_1$, is 1.8. If we obtain the estimate $\hat{\beta}_1 = 2.1$, which lies past the critical value and in the rejection region, then we will reject $H_0$. If we found $\hat{\beta}_1 = 1.3$ instead, we would fail to reject $H_0$.

3.7. Summary of Statistical Hypothesis Testing

- **Null hypothesis**: A statement about the population that is assumed to be true until proven false. Denoted $H_0$.

- **Alternative hypothesis**: A statement about the population that the researcher wishes to support, in order to provide evidence against $H_0$. This is denoted $H_A$. $H_0$ is the negation of $H_A$.

- **Test statistic**: A statistic, calculated using the sample, that is used to make a decision. One such statistic is the $t$-statistic.
• **Level of significance:** We fix the probability of type I error. This is denoted by $\alpha$ and is fixed, typically at $\alpha = 0.05$, $\alpha = 0.01$, or $\alpha = 0.10$.

• **Rejection region:** Values of the test statistic for which we will reject $H_0$.

• **Critical values of the test statistic:** Values of the test statistic beyond which we reject $H_0$.

4. Examples

**Example 5** (Height and Weight). Recall our height-weight regression where we included the individual’s post office box number as an explanatory variable. Although we said that this variable was nonsensical, the OLS coefficient was actually nonzero:

$$\text{WEIGHT}_i = 102.35 + 6.36 \cdot \text{HEIGHT}_i + 0.02 \cdot \text{BOX}_i.$$  

Suppose your friend collected the data for you and suggested including $\text{BOX}_i$ in the regression. A test of the hypothesis $H_0: \beta_2 = 0$ will allow you to check your friend’s claim that $\text{BOX}_i \neq 0$ ($H_A$).

**Example 6** (Wage Discrimination). Suppose we are interested in testing whether there is a significant difference in wages for men and women in a particular occupation after controlling for age, experience, education, and some measures of ability. We can test whether the coefficient on a dummy variable MALE is significantly different from zero:

$$H_0 : \beta_k = 0$$
$$H_A : \beta_k \neq 0$$

If we reject $H_0$, then the difference in wage is said to be statistically significant at the level $\alpha$ (e.g., 5%).