5.2 Hypothesis testing

We'll begin this section by looking at the variance of the OLS slope estimator ($\mathrm{Var}[b_1]$). There are three reasons to get this formula:

1. Looking at it will provide insight into what determines the accuracy (a smaller variance) of the estimator.
2. It is required to prove that OLS is an efficient estimator, and therefore is BLUE.
3. It is needed for hypothesis testing.

- In Chap. 3, derived the var. of $\bar{y}$
- Similarly, b_1 is a random variable, it has a variance
- Too difficult to derive for this course.

$$\mathrm{Var}[b_1] = \frac{\sigma_\epsilon^2}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}$$

- $\text{Var}[b_1]$ decreases as n increases.
- $\text{Var}[b_1]$ decreases as the sample variation in X increases.
- $\text{Var}[b_1]$ decreases as variation in ϵ decreases.

We want our estimator to have as low a variance as possible! A lower variance means that, on average, we have a higher probability of being close to the "rights answer" (provided the estimator is unbiased). These factors that lead to a lower $\text{Var}[b_1]$ make sense:

- If we have more information (larger n), it should be "easier" to pick the right regression line.
- Since we are using changes in X to try to explain changes in Y, the bigger changes in X that we observe, the easier it is to pick the regression line.
- The less unobservable changes there are (in ϵ that are causing changes in Y, the easier it is to pick the regression line.

Gauss-Markhov Theorem

OLS is efficient. G-M theorem says it has lowest variance among all possible linear unbiased estimators for β. That is, OLS is

B.L.U.E.

The G-M theorem is not highlighted as much in the text as it should be.

It is very important!

Test-stats and CIs

$$H_0 : \beta_1 = \beta_{1,0}$$
$$H_A : \beta_1 \neq \beta_{1,0}$$

A common hypothesis in economics is where the marginal effect is zero (X does not cause Y), so that the above null and alternative become:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

As in chapter 3, we will begin with the z-test. In general, the z-statistic is determined by:

$$z\text{–statistic} = \frac{\text{estimate} - \text{value of } H_0}{\sqrt{\text{Var}\left[\text{estimator}\right]}} \tag{5.8}$$

Population mean (chapter 3):

$$z = \frac{\bar{y} - \mu_{Y,0}}{\sqrt{\sigma_Y^2 / n}}$$

Slope estimator, β_1:

$$z = \frac{b_1 - \beta_{1,0}}{\sqrt{\mathrm{Var}\left[b_1\right]}}$$

Recall that the problem with the z-test (chap. 3) was that the variance of Y was unknown. Now, we have a similar problem, the variance of ϵ is unknown in the equation:

$$\mathrm{Var}\left[b_1\right] = \frac{\sigma_\epsilon^2}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}$$

How to estimate it?

Recall that the population model is:

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

and that the estimated model is:

$$Y_i = b_0 + b_1 X_i + e_i$$

Each unobservable part in the population model (β_0, β_1, ϵ_i) has an observable counter-part in the estimated model. So, if we want to know something about ϵ we can use e. In fact, an estimator for the variance of ϵ is the *sample variance* of the OLS residuals:

$$s_\epsilon^2 = \frac{1}{n-2} \sum_{i=1}^{n} (e_i - \bar{e})^2 = \frac{1}{n-2} \sum_{i=1}^{n} e_i^2 \tag{5.9}$$

Why is the -2 in the denominator of equation 5.9? Recall that, in chapter 3, when we wanted to estimate σ_y^2 we used the sample variance of y:

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

The estimator for the variance of b_1 is now:

$$\hat{\text{Var}}[b_1] = \frac{s_\epsilon^2}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}$$

And now, the t-statistic for testing β_1 is obtained by substituting $\hat{\text{Var}}[b_1]$ for $\text{Var}[b_1]$ in the z-statistic formula:

$$t = \frac{b_1 - \beta_{1,0}}{\sqrt{\hat{\text{Var}}[b_1]}} \tag{5.10}$$

The denominator of 5.10 is often called the *standard error* of b_1 (like a standard deviation), and equation 5.10 is often written instead as:

$$t = \frac{b_1 - \beta_{1,0}}{\text{s.e.}[b_1]} \tag{5.11}$$

If the null hypothesis is true, the t-statistic in equation 5.11 follows a t-distribution with degrees of freedom $(n-k)$, where k is the number of βs we have estimated (two). To obtain a p-value we should use the t-distribution, however, if n is large, then the t-statistic follows the standard Normal distribution. For the purposes of this course, we shall always assume that n is large enough such that $t \sim N(0,1)$. To obtain a p-value, we can use the same table that we used at the end of chapter 3 (see Table 3.2).

5.2.3 Confidence intervals

Confidence intervals are obtained very similarly to how they were in chapter 3. The 95% confidence interval for b_1 is:

$$b_1 \pm 1.96 \times \text{s.e.}\,[b_1] \tag{5.12}$$

The 95% confidence interval can be interpreted as follows: (i) if we were to construct many such intervals (hypothetically), 95% of them would contain the true value of β_1; (ii) all of the values that we could choose for $\beta_{1,0}$ that we would fail to reject at the 5% significance level.

We can get the 90% confidence interval by changing the 1.96 in equation 5.12 to 1.65, and the 99% C.I. by changing it to 2.58, for example.