1 Recap of OLS Estimator Properties

1.1 Assumptions

1. $Y_i = \mathbf{X}_i'\beta + \epsilon_i$ and $\mathbb{E}\left[\epsilon_i | X_i\right] = 0$ (Key Concept 4.3.1 in Text, p.131)
2. $\{X_i, Y_i\}_{i=1}^n$ is an i.i.d. sample (Key Concept 4.3.2)
3. The matrix $\mathbb{E}\left[\mathbf{X}_i\mathbf{X}_i'\right]$ is invertible (non-singular) (Implicitly assumed in text)
4. The matrix $\mathbb{E}\left[\mathbf{X}_i\epsilon_i^2\mathbf{X}_i'\right]$ is non-singular (Implicitly assumed in text)
5. The random variables (X_i, Y_i) have finite fourth moments.

1.2 Properties

1. Under assumptions 1-3, $\mathbb{E}[\hat{\beta}_{OLS}] = \beta$, that is, $\hat{\beta}_{OLS}$ is an unbiased estimator of β.
2. Under assumptions 1-3, $\hat{\beta}_{OLS} \xrightarrow[n\to\infty]{p} \beta$, and thus the OLS estimator $\hat{\beta}_{OLS}$ is consistent for β. In large samples, the OLS estimator should be close to β with high probability.
3. Under all 5 assumptions, $\sqrt{n}\left(\hat{\beta} - \beta\right)$ will in large samples be approximately bivariate normal with mean vector equal to $[0, 0]'$ and variance matrix $\hat{\mathbf{V}} = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i e_i^2\mathbf{X}_i'\right)\left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i'\right)^{-1}$, i.e., $\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(\mathbf{0}, \hat{\mathbf{V}}\right)$.

Exercise 1: ***Understanding Consistency***

1. Suppose that $\{X_n\}_{n=1}^\infty$ is a sequence of estimators (random variables). Each X_n is normally distributed with mean 0 and variance equal to $\frac{1}{n}$. Show that each $\sqrt{n}X_n$ is normally distributed with mean zero and variance one. What is the limit distribution of the sequence $\sqrt{n}X_n$?

2. Consider the example in Question 3 above. Pick any very small number, call it τ. Show that

$$\lim_{n\to\infty} P\left(X_n > \tau\right) = 0$$

so that X_n is consistent for 0.

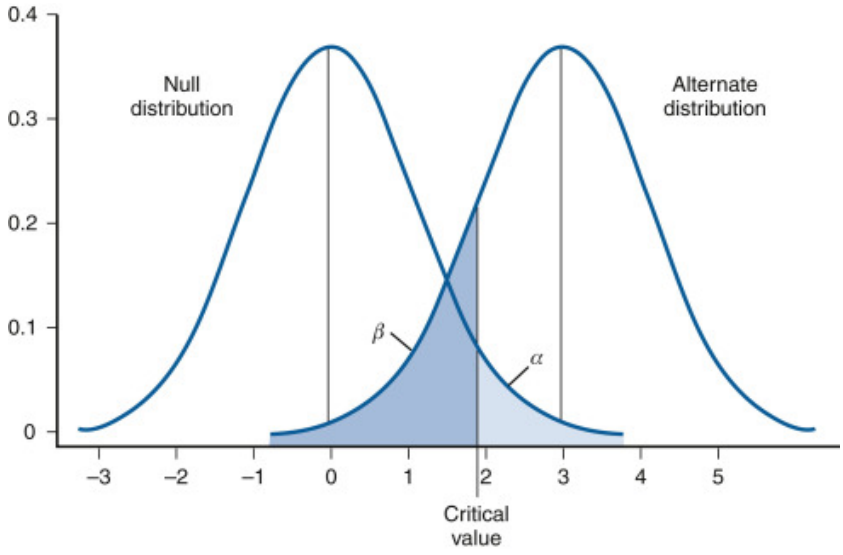
2 Review of Hypothesis Tests and Confidence Intervals

Statistical tools help us answer questions about unknown characteristics of distributions in populations of interest. For example, what is the mean of the distribution of earnings of recent college graduates? Do mean earnings differ for men and women and, if so, by how much? The key insight of statistics is that one can learn about a population distribution by selecting a random sample from that population. Three types of statistical methods are used throughout econometrics:

- Estimation: compute a "best guess" numerical value for an unknown characteristic of a population distribution.
- Hypothesis testing: formulate a specific hypothesis about the population and then using sample evidence to decide whether it is true.
- Confidence intervals: use a set of data to estimate an interval or range for an unknown population characteristic.

2.1 The Terminology of Hypothesis Testing

The starting point of statistical hypotheses testing is specifying the hypothesis to be tested, called the **null hypothesis**. Hypothesis testing entails using data to compare the null hypothesis to a second hypothesis, called the **alternative hypothesis**, that holds if the null does not. For example, suppose a country raises its minimum wage. Economists might be interested in understanding whether the increase in the minimum wage has an effect on the employment rate in the affected region. You regress the employment rate in each country on the corresponding minimum wage and a constant term to get the slope estimator, $\hat{\beta}_1$. Suppose now all the five assumptions above are satisfied. The null hypothesis H_0 can be the increase in the minimum wage has no effect on the employment rate. Mathematically, this null hypothesis can be represented as: $\beta_1 = 0$. The alternative hypothesis (H_1) can be the increase in the minimum wage has an effect on the employment rate. Note that this is a two-sided alternative hypothesis because we're not specifying the direction of the effect (it could be positive or negative). Mathematically, this alternative hypothesis can be represented as: $\beta_1 \neq 0$.



A statistical hypothesis test can make two types of mistakes: a **type I error**, in which the null hypothesis is rejected when in fact it is true; and a **type II error**, in which the null hypothesis is not rejected when in fact it is false.

The prespecified rejection probability of a statistical hypothesis test when the null hypothesis is true, that is, the prespecified probability of a type I error—is the **significance level** of the test.

The **critical value** of the test statistic is the value of the statistic for which the test just rejects the null hypothesis at the given significance level.

The set of values of the test statistic for which the test rejects the null hypothesis is the **rejection region**, and the set of values of the test statistic for which it does not reject the null hypothesis is the **acceptance region**.

The probability that the test actually incorrectly rejects the null hypothesis when it is true is the **size of the test**, and the probability that the test correctly rejects the null hypothesis when the alternative is true is **the power of the test**.

The **p-value** is the probability of obtaining a test statistic, by random sampling variation, at least as adverse as the t-statistic actually observed, assuming that the null hypothesis is correct. Large p-values indicate that, under the null hypothesis being correct, it is very likely to obtain t-statistics at least as extreme as the t-statistic actually observed. Small p-values indicate that, under the null hypothesis being correct, it is very unlikely to obtain t-statistics at least as extreme as the t-statistic actually observed. Thus, the p-value is the smallest significance level at which you can reject the null hypothesis.

To state the definition of the p-value mathematically, take the above example again, where we have a true β_1 unknown to us, a slope estimator β_{1act} estimated from the sample we actually drew, a $\hat{\beta}_1$ representing the slope estimator that we would get from any given random sample. Let Pr_{H_0} denote the probability computed under the null hypothesis (that is, computed assuming that $\beta_1 = 0$). The p-value is

$$\text{p-value} = Pr_{H_0}[|\hat{\beta}_1 - 0| > |\beta_1^{act} - 0|]$$

If the p-value is small, then the observed value β_1^{act} is extremely unlikely under the null hypothesis. To compute the p-value, it is necessary to know the sampling distribution of $\hat{\beta}_1$ under the null hypothesis.

2.2 Two-Sided Hypotheses Concerning β_1

Under the five assumptions stated in Section 1.1, it is reasonable to pretend that $\sqrt{n}(\hat{\beta} - \beta)$ is distributed exactly as $\mathcal{N}\left(\mathbf{0}, \hat{\mathbf{V}}\right)$. The null and alternative hypotheses need to be stated precisely before they can be tested. Testing the hypothesis $H_0 : \beta_1 = b$ against the alternative $H_1 : \beta_1 \neq b$:

- Step 1: Compute the standard error of $\hat{\beta}_1$, $SE(\hat{\beta}_1)$. Specifically, $SE(\hat{\beta}_1) = \sqrt{\frac{\hat{\mathbf{V}}_{22}}{n}}$.
- Step 2: Compute the t-statistic as $t = \frac{\hat{\beta}_1 - b}{SE(\hat{\beta}_1)}$.
- Step 3: Compute the p-value, which is the probability of observing a value of $\hat{\beta}_1$ at least as different from b as the estimate actually computed ($\hat{\beta}_1^{act}$), assuming that the null hypothesis is correct. Stated mathematically,

$$\begin{aligned}
\text{p-value} &= Pr_{H_0}\left[\left|\hat{\beta}_1 - b\right| > \left|\hat{\beta}_1^{act} - b\right|\right] \\
&= Pr_{H_0}\left[\left|\frac{\hat{\beta}_1 - b}{SE(\hat{\beta}_1)}\right| > \left|\frac{\hat{\beta}_1^{act} - b}{SE(\hat{\beta}_1)}\right|\right] \\
&= Pr_{H_0}\left[|t| > \left|t^{act}\right|\right]
\end{aligned}$$

  where Pr_{H_0} denotes the probability computed under the null hypothesis and t^{act} is the value of the t-statistic actually computed. Because $\hat{\beta}_1$ is approximately normally distributed in large samples, under the null hypothesis the t-statistic is approximately distributed as a standard normal random variable, so in large samples

$$\text{p-value} = Pr[|Z| > |t^{act}|] = 2\Phi(-|t^{act}|)$$

  A p-value of less than 0.05 provides evidence against the null hypothesis in the sense that, under the null hypothesis, the probability of obtaining a value of $\hat{\beta}_1$ at least as far from the null as that actually observed is less than 5%. If so, the null hypothesis is rejected at the $\alpha = 0.05$ significance level. Alternatively, the hypothesis can be tested at the 0.05 significance level simply by comparing the absolute value of the t-statistic to $Z_{1-\alpha/2} = 1.96$, the critical value for a two-sided test, and rejecting the null hypothesis at the 0.05 level if $|t^{act}| > 1.96$.

2.3 Multi-parameter Testing

We need to introduce the chi-squared distribution and a related theorem in order to talk about hypothesis testing for multiple regression coefficients.

chi-squared distribution: Suppose that W is a normally distributed random vector of dimension $k \times 1$ with mean vector equal to μ and variance matrix V. Then, the distribution of the scalar random variable

$$T = (W - \mu)' V^{-1} (W - \mu)$$

is known as the **chi-squared distribution with k degrees of freedom**.

Exercise 2: ***Chi-squared Distribution***

Suppose that X_k is normally distributed with mean zero and variance $\frac{1}{k^2}$ for $k = 1, 2$ and that the $\{X_k\}_{k=1}^2$ are mutually independent. Then, what is the distribution of $\sum_{k=1}^2 k^2 X_k^2$?

Theorem: Suppose that $\{\mathbf{S}_n\}_{i=1}^\infty$ is a sequence of random variables of dimension $k \times 1$ such that for some constant vector μ and variance matrix $\mathbf{V}$

$$\sqrt{n} (\mathbf{S}_n - \mu) \xrightarrow[n\to\infty]{d} \mathcal{N} (\mathbf{0}, \mathbf{V})$$

and suppose that we have available an estimator $\hat{\mathbf{V}}_n$ such that $\hat{\mathbf{V}}_n \xrightarrow[n\to\infty]{p} \mathbf{V}$. Then, the scalar random variable

$$T_n = \sqrt{n} (\mathbf{S}_n - \mu)' \hat{\mathbf{V}}_n^{-1} \sqrt{n} (\mathbf{S}_n - \mu)$$

converges in distribution to a chi-squared random variable with k degrees of freedom which we sometimes write as χ_k^2.

Suppose our null hypothesis is $H_0 : \mathbf{R}\beta = \mathbf{r}$, where $\mathbf{R}$ is a known matrix and $\mathbf{r}$ is a know vector. Following the results from class and the asymptotic distribution of $\hat{\beta}$ under the null hypothesis, we have

$$\sqrt{n} \left(\mathbf{R}\hat{\beta} - \mathbf{r}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}_{\#r} \left(\mathbf{0}, \mathbf{RVR}'\right)$$

where $\#r$ is the number of (linearly independent) rows of $\mathbf{R}$.

So, by the theorem above, under H_0,

$$W_n = \sqrt{n}\left(\mathbf{R}\hat{\beta} - \mathbf{r}\right)' \left(\mathbf{RVR}'\right)^{-1} \sqrt{n}\left(\mathbf{R}\hat{\beta} - \mathbf{r}\right)$$

converges to a chi-squared distribution. The degrees of freedom of this distribution are equal to the rank of R or put another way, to the number of restrictions in H or #r.

This statistic is known as the **Wald Statistic** and the associated test is known as the **Wald Test**.

We then ask if it is likely that our observed value of W_n comes from a $\chi^2_{\#r}$ distribution. If our observed value of W_n is much larger than we would expect if the distribution were $\chi^2_{\#r}$, we will reject H_0. As before, to find a precise definition of large, we find a critical value c_1 such that the probability of making a Type 1 error is equal to some small α

$$Pr_{H_0}\left[W_n > c_1\right] = \alpha$$

This critical value is the $(1-\alpha)^{th}$ quantile of a chi-squared distribution with #r degrees of freedom and which we denote by $c^{\chi^2_{\#r}}_{1-\alpha}$. We can use software to find this critical value.

2.4 Confidence Interval

In Intro Stats, you learned to construct confidence regions for scalar estimators. For instance, knowing that $\hat{\beta}_1 - \beta_1$ is normally distributed with mean zero and variance $\frac{\hat{v}_{22}}{n}$ you would construct the confidence interval

$$C_{1n} = \left[\hat{\beta}_1 - z_{1-\alpha/2}\sqrt{\frac{\hat{v}_{22}}{n}}, \hat{\beta}_1 + z_{1-\alpha/2}\sqrt{\frac{\hat{v}_{22}}{n}}\right]$$

This set consists of all candidate parameter values for which the null hypothesis was not rejected at level α, and the confidence interval satisfies the property that

$$Pr\left[\beta_1 \in C_{1n}\right] \geq 1 - \alpha$$

So with probability at least $1-\alpha$ the true value of the parameter will lie in the confidence region.

Exercise 3: ***(Simulation Exercise illustrating Large Sample Distribution based Hypothesis Testing)***

Consider the following model. We observe an i.i.d. sample $\{Y_i, X_i\}_{i=1}^n$ where Y_i is generated according to

$$Y_i = 1 - 2X_i + \epsilon_i \tag{1}$$

ϵ_i is an unobserved error term and is independent of X_i and distributed uniformly on the interval $[-3, 3]$. Each X_i is drawn from normal distribution with mean zero and variance one.

1. Under the conditions stated above, will $\mathbb{E}\left(\epsilon_i | X_i\right) = 0$? What is $\mathbb{E}\left(Y_i | X_i\right)$ and will it vary by i? Justify your answers.

2. ***For this part only.*** Suppose now that the error term is generated as follows. First, X_i is drawn from the stated distribution and suppose the realized value is x. Then ϵ_i is drawn from a normal distribution with mean x and variance x^2. If the error term is generated in this way, will $\mathbb{E}(\epsilon_i|X_i) = 0$? What is $\mathbb{E}(Y_i|X_i)$ in this case and what is $Var(Y_i|X_i)$?

3. Suppose that I generate five thousand data sets of size $n = 5$ based on this model: That is to say that for the i^{th} observation in the k^{th} data set I first draw $\epsilon_i^{(k)}$ from the uniform distribution stated above; then I draw $X_i^{(k)}$ from the standard normal distribution and finally, I use the equation (1) to generate the outcome variable $Y_i^{(k)}$.

   This is carried out in the file `PS1.ipynb` that you will need to pull into your Jupyter Lab environment using this link.[1] The code generates 5,000 data-sets (each data set has a sample size of 5) and for each data set estimates equation 1 using OLS. It stores the OLS estimates and their associated standard errors in the `simbetas` data frame.

   The data frame contains 5000 observations (one for each data set) on four variables `beta0`, `sebeta0`, `beta1` and `sebeta1`. `beta0` records the OLS estimate of the intercept and `sebeta0` its associated standard error. Similarly, `beta1` is the OLS estimate of the slope in (1) and `sebeta1` its associated standard error.

   Run the Python code in Jupyter Lab to compute the mean and standard deviation of `beta0` and `beta1` as well as the median and the 25$^{\text{th}}$ and 75$^{\text{th}}$ percentiles.

   Next, run the Python code on Jupyter Lab that draws a separate histogram for each of `beta0` and `beta1`. In addition, the code also overlays a normal density (with mean and standard deviation equal to the mean and standard deviation of the respective `beta` columns).

   The comparison between the histogram and the normal density suggest that the histogram is less well approximated by the normal distribution at the tails of the distribution. We examine this in more detail by comparing the quantiles of the two distributions. That is to say, we compare the quantiles of the `beta0` and `beta1` columns to the quantiles of the corresponding normal distributions defined above by using a qqplot. Describe the concept of qqplot in a few sentences and then run the python code to draw the qqplots and comment on them (in particular, how do they behave at the extremes).

4. Next, replace the sample size of 5 for the `sample_size` variable with 200. Redo part (3) and comment on any differences you observe.

[1] https://datahub.berkeley.edu/hub/user-redirect/git-pull?repo=https%3A%2F%2Fgithub.com%2Fds-modules%2FENVECON-118-FA23branch=mainurlpath=lab%2Ftree%2FENVECON-118-FA23%2FProblem+Sets%2FPS1

5. Now, I hand over the 1000^{th} estimates (i.e.`beta0`,`sebeta0`,`beta1`,`sebeta0` for the 1000^{th} data set) to a student in EEP 118 and inform them that the model satisfies the assumptions collected in Key Concept 4.3 in the Stock and Watson text-book (but do not tell them the true values of the intercept and slope). I ask them to use the data to conclude whether the true value of the slope in (1) was 3 and to use a significance level of .05.[2] Suppose that the slope estimate and its associated standard error for this data set are

$$\hat{\beta} = -1.65 \ \ (.14)$$

What decision procedure would you use to test whether the true slope was 3 and what is your conclusion? Make reference to the observations you made in part (4) above.

6. This is an artificial example, but use this to think about how you do inference with real data. To make things concrete, pick one of the data sets we have worked with so far in class and discuss how hypothesis testing in that context is informed by this example here.

7. This part seeks to clarify the meaning of the significance level. Suppose we are now interested in testing the null that the true slope parameter is -2 against the two sided alternative that it is not, at a significance level $\alpha = .05$.

   The python code constructs a test statistic `teststat` for testing this null for each of the 5,000 observations (remember each observation corresponds to the OLS estimates and standard errors from one data realization). Use the simulations with $n = 200$ for this part.

   Next, the code creates a variable called `reject` that is equal to 1 if the null hypothesis is rejected for that particular observation (data set). Suppose the critical value you use is `cval`[3], the code generates a new variable `reject` equal to 1 if `|teststat| > cval`. This is given by the line

[2] The alternative hypothesis is that the true slope parameter is not equal to 3 so this is a two sided alternative.

[3] The critical value for the two-sided alternative is given by the $1 - \alpha^{th}$ quantile of the standard normal distribution and we will reject if the absolute value of the test statistic is larger than `cval`. We use the python function `norm.ppf` to compute `cval`.

```
# Generate the reject variable,
simbetas1['reject'] = (simbetas1['teststat'].abs() > cval).astype(int)
```

Next, the code displays the fraction of rejections by displaying the mean of the `reject` variable. Is it close to what you would expect it to be? Why or why not?

8. Redo the previous part but now testing the null that the true slope parameter is equal to 1 against the two-sided alternative that it is not equal to 1 (at $\alpha = .05$). Comment on the mean of the `reject` variable. Is it is what you expected it to be? Why or why not?

9. Experiment with the Python code to examine how the OLS estimates change when you change the data generating process so that the key assumptions in Key Concept 4.3 no longer hold. For instance, if the error term ϵ_i is drawn from a distribution where the relevant moments do not exist or are infinite. This can be achieved by changing the line generating the error term e to

```
    e = np.random.standard_cauchy(sample_size).
```

Comment on how your answers to part (3) above change.