Inference: Confidence Intervals and Hypothesis Testing

Marcio Santetti | Fall 2023

Table of contents

Introduction

So far, we have studied in detail the nuts and bolts of linear regression, as well as the assumptions such models have to undertake in order to make OLS the "best" method among all linear estimators. Now, it is about time to draw **inferences** from our models. You for sure know by now how to interpret OLS coefficients, and this is a crucial quality. However, recall that we work with *samples*, and we want to use them with our models to describe *population parameters* as precisely as possible. How to know whether our OLS coefficients are valid for this purpose? Only **inference** will tell us that.

The two main tools for statistical inference are **confidence intervals** and **hypothesis testing**. For econometric purposes, usually the second is more widely used, so we will not spend too much time on confidence intervals. But we will keep doing different statistical tests throughout the semester. Moreover, after this week's content you will be able to *fully* interpret what your statistical software of use informs you regarding your estimated model. The last bit of information that we need to cover regards different tests of hypotheses about our coefficients. The applied lecture will complement the theory we will go over now.

Confidence intervals

A **confidence interval** is a *range* which contains the *true* value of an estimate a specified percentage of time, assuming a sampling distribution of that estimate. Such *percentage* is the **level of confidence**, denoted by $(1 - \alpha)$, with α being the **significance level**. This terminology must have been present in your Stats courses, and it remains the same, in case you still have some memories of that time.

For confidence intervals, we use **Student's t-distribution**. The good news is that we will not use tables with critical values

anymore. The computer will handle that for us, but we need to ask the right questions. Recall that for standard sample statistics, such as a sample mean ($\bar{x}$), we would calculate a confidence interval (CI) for the population mean in the following way:

$$CI = \bar{x} \pm t_c \cdot \sigma$$

where t_c is the *critical value* given by the t-distribution's table, and σ is the population standard deviation, which is assumed to be known. In case we do not know the population's standard deviation for that variable, we use

$$CI = \bar{x} \pm t_c \cdot \frac{s}{\sqrt{n}}$$

where s is the *sample standard deviation*, divided by the square root of the sample size, n. The critical value given by the t-table depends on two factors: the significance level (α), and the number of degrees-of-freedom, which in this case is given by $n - 1$.

For our purposes, however, we are not interested in *sample statistics* anymore, such as the sample mean. Our objects of interest are β coefficients, estimated through OLS regression. The CI "analog" to the above is the following:

$$CI = \hat{\beta}_k \pm t_c \cdot SE(\hat{\beta}_k)$$

where now we replace the sample statistic by our estimated β_k coefficient ($k = 0, 1, 2, ..., k$), and $SE(\hat{\beta}_k)$ is the **standard error** of our estimate.

Deriving the mathematical expression for the **standard error** would require some detour, and you should feel free to let me know if you want a formal presentation of it, since I'd be happy to present it to you. However, for now, the most important thing is to present the intuition of this measure: the $SE(\cdot)$ of a regression estimate basically tells us how **precise** it is. Therefore,

it is an analog for the *standard deviation* of any sample statistic that you are used to.

Figure 1 aims to refresh your memory about how the t-distribution looks. It is very similar to the standard "bell-shaped" curve from the Normal distribution, and, as the sample size increases, these two distributions are basically indistinguishable. The t-distribution is also centered around 0, and the figure shows the two *tails* in pink, each one having an area of $\alpha/2$, and the area in gray is denoted by $1 - \alpha$. Recall that the total area under the curve of *any* probability distribution is equal to 1 (or 100%), and we partition this total area for inference purposes. Thus, for example, if our significance level is $\alpha = 10\%$, the area in gray will be equal to $1 - .1 = .9$, and each tail will have an area of $\alpha/2 = .1/2 = 0.05$.

The next bit of information we need in order to calculate t_c is the number of *degrees-of-freedom*. Remember that we are in the *regression* world now, so our DOF equals $n - k - 1$, that is, we subtract the number of slope coefficients (k) and the intercept (1) from our total sample size (n). Suppose we have a large sample, say $n = 400$, and we estimate a regression with 3 independent variables, thus $k = 3$. Then, our degrees-of-freedom are $400 - 3 - 1 = 396$. Feel free to check on the t-table, and it wil give you $t_c = 1.645$ for the two-tailed case.

Let us look at an applied example. Consider the following regression output. Here, the *standard errors* are presented in parentheses, right below the actual $\hat{\beta}$ coefficients. This how usually such results are presented in papers and books.

$$\hat{S}_i = \underset{}{102,192} - \underset{(2,053)}{9.075}N_i + \underset{(.0727)}{0.3547}P_i - \underset{(.543)}{1.288}I_i$$

$$n = 33 \qquad \bar{R}^2 = .579$$

where S_i is the gross sales volume at each i^{th} location of a restaurant chain (in thousands); N_i is the number of close

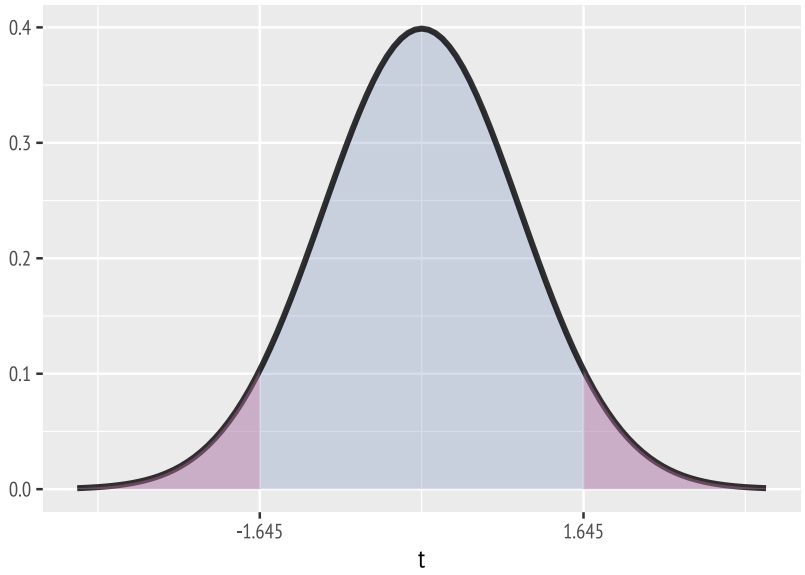

Figure 1: The t distribution.

competitors; P_i is the population within a 3-mile radius (in hundreds), and I_i is the average income per household.

Before we exercise our new content, make sure you are comfortable interpreting each $\hat{\beta}$ coefficient, as well as the adjusted R^2, displayed below the regression output. After that, let's do new business.

Say we want to construct a **90% confidence interva**l for the coefficient on P_i, that is, the effect of a one-hundred increase in the population that lives close by the restaurant on its gross sales volume. To do it, we need 3 pieces of information: $\hat{\beta}_P$, $SE(\hat{\beta}_P)$, and t_c.

$\hat{\beta}_P$ and $SE(\hat{\beta}_P)$ are already given by the regression output: .3547 and .0727, respectively. Let us consult the table once again—do not worry, you will say farewell to these tables after you learn how to do this with the computer—, with $\alpha = 10\%$, and $DOF = 33 - 3 - 1 = 29$. This is a two-tailed procedure, so $t_{.1,29} = 1.699$.[1] Once again, make sure you feel comfortable looking for critical values in the table. Any issues, please reach

[1] Recall that the *t*-table already distinguishes between one- and two-tailed procedures. In case you need to refresh your memory, you can take a look at one here.

out.

Now, we just plug in these three pieces into the formula:

$$CI = .3547 \pm 1.699 \cdot (.0727)$$

Our confidence interval for $\hat{\beta}_P$ is [.2312; .4782]. This means that, *90% of the time, the "true" coefficient for* P_i *will fall between .2312 and .4782*. And our estimated coefficient, whose value is .3547, is included in this interval.

Informally, you can reproduce the latter interpretation out loud with no problem. However, if we want to be *technically precise*, the correct interpretation for this confidence interval is: *there is a 90% probability that the estimated coefficient will be equal to a value such that the interval* [.2312; .4782] *will include the "true" parameter* β_P. For exams and assignments, feel free to use any of these interpretations. Just be aware of the technical interpretation, in case someday you need it. Who knows?

As stated in the Introduction, confidence intervals tend to be secondary, at least for our purposes in this course. We will see some applications of it in the future, but our inference bread-and-butter will be **hypothesis testing**.

Hypothesis testing

Hypothesis testing determines what we can learn about the real world from sample data, through simple statistical tests. Even though we would like to *prove* that a given hypothesis about a theory being supported by empirical estimation is *correct*, what hypothesis testing provides is that we can often **reject** a given hypothesis with a certain level of significance.

As explained in the previous section, our interest lies in the estimated coefficients from our regression, the $\hat{\beta}$'s. We use these to test hypotheses about **population** parameters, the β_{true} coefficients. Thus, in case we want to test whether the *true*

coefficient for some variable x_i is *different from* zero, we have our **null** and **alternative** hypotheses, H_0 and H_a, respectively:

- H_0: $\beta_{i,true} = 0$
- H_a: $\beta_{i,true} \neq 0$

Notice that we use the "true" parameter in our hypotheses statement, and not the estimated coefficients.

In Econometrics, a test like the one above is known as a **test for statistical significance**, or a **significance test**. This is straightforward, since if we do not reject the null hypothesis that a coefficient (statistically) equals zero, this means that the latter is **not statistically significant**, given the chosen significance level (α). In case we reject this null hypothesis, then the coefficient is *statistically significant*, and the variable may be considered as *relevant* for our regression model, given α.

As another example, consider a simple wage-education model:

$$wage_i = \beta_0 + \beta_1 educ_i + u_i$$

Our standard expectation would be that the *more* an individual is educated, the *better* they will be paid. Translating this expectation into a *test of hypothesis*, we set:

- H_0: $\beta_1 = 0$
- H_a: $\beta_1 > 0$

In case we **reject** H_0, our expectation is met, and thus our sample regression model meets the research hypothesis (at least with respect to the sign of β_1) for the entire population. Notice that

this is a *one-tailed* test, while the previous significance test is a *two-tailed* one.

However, hypothesis testing is not free from error, and being aware of the possible pitfalls is relevant for inference. From hypothesis testing, we may have two **types** of error:

- **Type I error**: when a *true* null hypothesis (H_0) is *rejected*. Its probability is given by α.
- **Type II error**: when a *false* alternative hypothesis (H_a) is *not rejected*. Its probability cannot generally be calculated, since knowing the "true" parameter would be required.[2]

[2] Of course, we can set values for population parameters, but we will not do this in our course. A useful measure concerning the probability of Type II errors is known as the **power of a hypothesis test**, which is $1 - P(\text{type II error})$. We will not study it here, but feel free to catch up with this concept in any Stats book.

The t-test

The *t-test* is usually used to test hypotheses about **individual** regression slope coefficients. Consider the following multiple regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

The *form* of the t-statistic for the k^{th} coefficient is:

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

where β_{H_0} is the value for the **population parameter** that appears in the tests's null hypothesis (H_0). The rest of the test's components you know from before. In case you are running a **significance test**, β_{H_0} is automatically 0, so t_k becomes

$$t_k = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)}$$

For our example, $k = 2$ slope coefficients. Suppose we want to test whether the coefficient on x_2 is *statistically significant* in our model. Th only information we need are $\hat{\beta}_2$ and its standard error. Thus, $t_2 = \hat{\beta}_2 / SE(\hat{\beta}_2)$.

Now, let us go back to the sales volume example introduced in the previous section. I will just reproduce its output once again, so you do not need to look back there.

$$\hat{S}_i = 102,192 - \underset{(2,053)}{9.075} N_i + \underset{(.0727)}{0.3547} P_i - \underset{(.543)}{1.288} I_i$$

$$n = 33 \qquad \bar{R}^2 = .579$$

Let us test whether the "true" parameter reflecting the change in sales volume due to population growth (β_P) is *statistically* greater than zero (i.e., positive). The estimated coefficient from the regression model is indeed positive, but it is simply a *point estimate*. Ideally, we would like to have several different estimates for β_P, using different samples of the same size. However, in reality such procedure is almost never possible. Therefore, we use hypothesis tests, based on *sampling distributions*, to evaluate this conjecture.

Our null and alternative hypotheses are:

- H_0: $\beta_P = 0$
- H_a: $\beta_P > 0$

Since $\beta_{H_0} = 0$, we have

$$t_P = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} = \frac{.3547}{.0727} = 4.88$$

Our **test statistic** is 4.88. So what? We compare this number to a *critical value*, given by the t-table. Assuming this time $\alpha = 5\%$,

we look for the corresponding critical value for $t_{0.05,29}$. Notice that this is a *one-tailed* test, so we **do not** divide α by 2.

The table gives us $t_{0.05,29} = 1.699$. Now, we compare t_P with this critical value. Is 4.88 greater than 1.699? Yes, it is. So, we **reject the null hypothesis**, and the effect of population growth on the gross volume of sales for this particular restaurant chain is statistically positive and significant (with 95% of confidence), given the sample used in our model. Visually, the procedure worked like this:

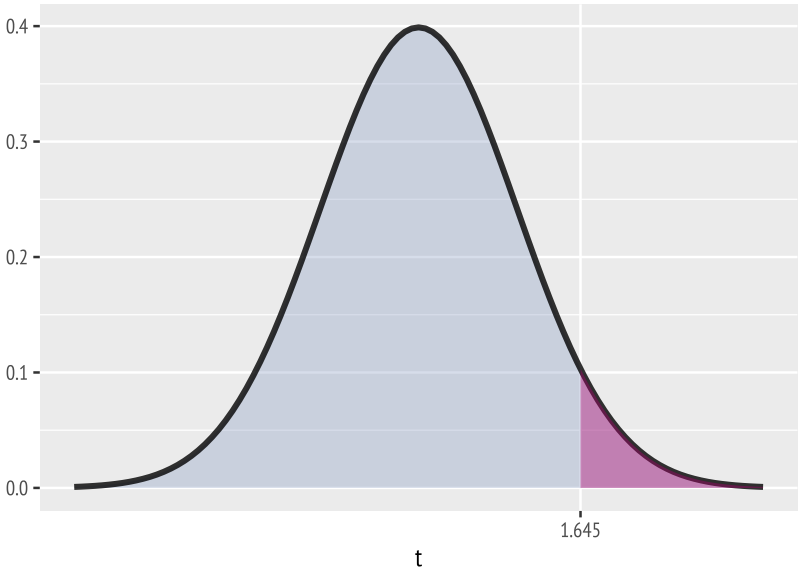

Figure 2: Rejection region for a right-tailed t-test.

Since the test statistic fell **inside** the *rejection region* (area in pink), we **reject** the null hypothesis.

As an exercise, repeat this procedure, this time changing the significance level to $\alpha = 10\%$. Does your statistical decision change? Does it stay the same?

Choosing a level of significance (α)

As you probably know, the **standard** significance level adopted by almost every applied field is $\alpha = 5\%$. Unless informed otherwise, keep this level as a standard for our course as well.

The p-value method

In most econometric applications, we may also resort to **p-values** in order to evaluate our hypothesis testing. As you have probably seen, a regression output usually returns both t-statistics and their respective p-values.

p-values can be understood as the *lowest possible significance level at which a null hypothesis cannot be rejected*. In other words, if a test's p-value falls *below* a given significance level, a null hypothesis is rejected; otherwise, we cannot reject it.

For practical purposes, the rule-of-thumb for statistical decisions using p-values is really simple:

- If the p-value is **lower** than the significance level (α): Reject H_0.
- If the p-value is **greater** than the significance level (α): Do not reject H_0.

Limitations of the t-test

By definition, the t-test is already limited, since it only evaluates hypotheses for *individual* coefficients. In case we want to test hypotheses on a *group* of coefficients, we use **F-tests**, which will appear in the next section.

However, beyond limitations imposed by its very definition, t-tests may also be *misleading* if we lose sight of what our regression procedure is about. In a nutshell, the fact that our regression coefficients are **statistically significant** does not

imply that these are *theoretically valid*. Recall: the computer accepts any regression model, and it will spit out numbers that we need to make sense of. But we need to go beyond that: are the values from our regression outputs *consistent* with our theoretical assumptions? Let me illustrate this situation with an example. Consider the following model:

$$\widehat{CPI}_t = 10.9 - \underset{(.23)}{3.2}C_t + \underset{(.02)}{.39}C_t^2$$

$$n = 21 \qquad \bar{R}^2 = .982$$

where CPI_t is the consumer price index, at time t, and C_t is the cumulative amount of rainfall, also at time t in the United Kingdom. Notice that C_t is also *squared* in this model, and we will look at these augmented models later on, but for now this is not relevant.

Notice that the adjusted R^2 for this model is almost 100%, and the coefficient on C_t is statistically significant (that is, we reject H_0: $\beta_{C_t} = 0$). However, despite this fantastic result in terms of significance and goodness-of-fit, this model does not make any theoretical sense: how can rainfall significantly affect an economy's inflation?

This does not make sense. Therefore, if we are not aware of the model's *underlying theory*, as well as our previous expectations with respect to signs and coefficients, we are just throwing numbers in a computer and asking our statistical package to do boring matrix algebra for us. Once again: **statistical significance does not imply theoretical and empirical validity**!

This blog presents hilarious examples of what we call *spurious* correlations. In other words, relationships that exist numerically, but do not come from any reasonable theoretical prior. Have a look at it and you will laugh.

The F-test

The last section gave a *spoiler* on the purpose of the **F-test**: testing *multiple* hypotheses simultaneously, or testing a single hypothesis about a *group* of parameters. In Econometrics, we usually use F-tests for the second option, and such joint null hypotheses are appropriate whenever the underlying economic theory specifies values for multiple coefficients simultaneously.

As the name already reveals, the F-test uses the *F-distribution* as a reference. You have probably been introduced to it in your Stats courses, but in case you have not, no need to worry. Take a quick look at its shape, and in the applied lecture for this topic everything will be clear.

Our usual procedure in this course will be to use the F-test to evaluate **joint significance**, that is, simultaneously testing the statistical significance of two or more slope coefficients. Suppose we have the following model, with k independent variables:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_3 x_{2i} + \beta_3 x_{3i} + ... + \beta_k x_{ki} + u_i$$

Now, suppose we want to evaluate, with a single test, whether **all** coefficients for the independent variables are *jointly significant*. The null and alternative hypotheses are

- $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$
- $H_a : H_0$ is not true

Such procedure requires the computation of an F-test. The simplified formula to obtain the F-statistic is:

$$F = \frac{R^2_{unr} - R^2_{rest}}{1 - R^2_{unr}} \cdot \frac{(n - k - 1)}{q}$$

It is better to break down the components of this formula by using an example. Below, we estimate a model to predict salaries of Major League Baseball (MLB) players (in logs), controlling for the number of years in the league ($years_i$), games played per year ($gamesyr_i$), career batting average ($bavg_i$), career home-runs ($hruns_i$), and RBIs per year ($rbisyr_i$). Our sample size is 353 individuals.

$$log(salary_i) = \beta_0 + \beta_1 years_i + \beta_2 gamesyr_i + \beta_3 bavg_i + \beta_4 hruns_i + \beta_5 rbisyr_i + u_i$$

We want to run a *joint significance* test on the coefficients β_3, β_4, and β_5. In other words, we want to test whether the effects of *career batting average*, *career home-runs*, and *RBIs per year* are jointly significant to explain variations in players' salaries:

- $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$
- $H_a : H_0$ is not true

Thus, we need an **F-test**. First, we estimate this regression, and extract its R^2. For the test's purposes, this regression is an **unrestricted model**, whose R^2 will be R^2_{unr}. After running this model via OLS, we obtain $R^2_{unr} = 0.627$.

Next, we estimate a **restricted model**, that is, the regression *without* the variables whose coefficients we will run the F-test on. Then, we estimate the following model:

$$log(salary_i) = \beta_0 + \beta_1 years_i + \beta_2 gamesyr_i$$

and compute its R^2, which is $R^2_{rest} = 0.597$. To complete the F-statistic's formula, we need q and $(n-k-1)$. q is the **number of restrictions** imposed to estimate the restricted model, that is, the number of dropped β coefficients to estimate the restricted regresion. Thus, in our example, $q = 3$. And $(n-k-1)$ is simply the number of degrees-of-freedom from the *unrestricted*

model. Therefore, $n - k - 1 = 353 - 5 - 1 = 347$. Let's plug in all values in the formula:

$$F = \frac{R^2_{unr} - R^2_{rest}}{1 - R^2_{unr}} \cdot \frac{(n - k - 1)}{q} = \frac{0.627 - 0.597}{1 - 0.627} \cdot \frac{347}{3} = 9.55$$

Now, we compare this test statistic with a critical value provided by the F-distribution table. If you recall, it requires *three* bits of information: the significance level, the DOF in the numerator, and the DOF in the denominator. Let's use $\alpha = 10\%$. The number of DOF in the numerator is q, and in the denominator is $(n - k - 1)$[3]. The table gives us $F_c = 4.353$ for the right tail and $F_c = 0.023$ for the left tail[4]. Our test statistic is greater than these critical values, so we **reject** the null hypothesis. We may thus infer, with 90% of confidence, that these 3 variables are *jointly* significant to explain MLB players' salaries, based on our sample.

In our applied lecture, we will say goodbye to looking for critical values in tables. The software gives you everything, with our job being performing the appropriate inferences.

[3] The formula shown here is a simplified version of the "original" F-statistic formula. In the original, q lies in the numerator, and $(n - k - 1)$ in the denominator.

[4] Recall that we have to divide the significance level by 2, since this is a two-tailed test.