Note on Bivariate Regression: Connecting Practice and Theory

Konstantin Kashin

Fall 2012

This note will explain - in less theoretical terms - the basics of a bivariate linear regression, including testing and confidence intervals. Specifically, we will use the `Leinhardt.RData` dataset as an example. Let's treat the entire dataset of 101 countries as the population, and then take a sample of size $n = 40$. Then, the sample is all that we actually observe. Our goal is to learn something about the population from the sample we observe.

Let's suppose we're interested in estimating the population linear conditional expectation function $\mathrm{E}[Y|X = x] = \beta_0 + \beta_1 x$, where Y is the log of infant mortality and X is the log of income per capita. Furthermore, let's make the typical assumptions that we make in the context of regression: random sampling, linearity of the population conditional expectation function, constant variance, and normality.

Now, when we run a linear regression of log of infant mortality on the log of income per capita for our sample in R, this is the output we obtain from the `lm` object:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.44624    0.41046  18.141  < 2e-16 ***
lincome     -0.56600    0.06412  -8.827 9.72e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5465 on 38 degrees of freedom
Multiple R-squared: 0.6722, Adjusted R-squared: 0.6635
F-statistic: 77.91 on 1 and 38 DF,  p-value: 9.719e-11
```

What do these outputs mean and how does R calculate them?

Estimating Regression Coefficients

We estimate β_0 (the intercept) and β_1 (the slope) of the population LCEF using the following two estimators (which are unbiased for the population quantities)[1]:

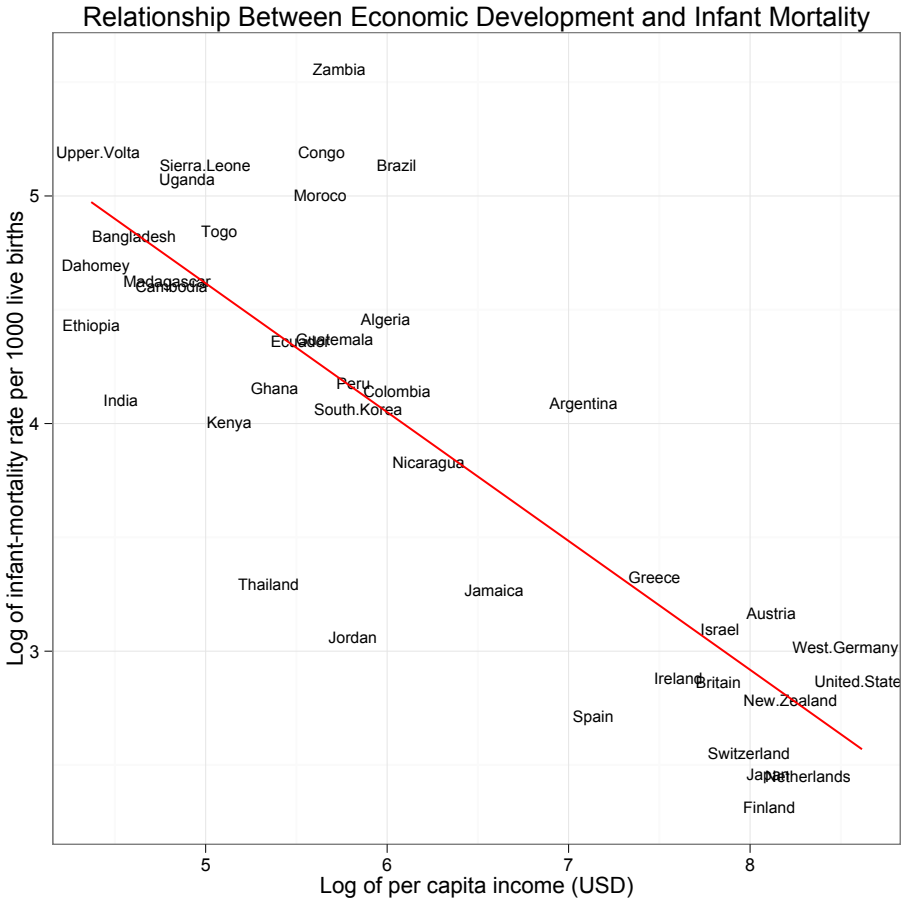
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Given the sample we drew, we obtain the following estimates: $\hat{\beta}_0 = 7.446$ and $\hat{\beta}_1 = -0.566$.

This is what our sample looks like, along with the sample conditional expectation function defined by the estimators for β_0 and β_1 presented above.

Figure 1: Sample Linear Conditional Expectation Function



[1]Note that we can alternatively express the estimator for the slope as as weighted average of the outcome variable: $\hat{\beta}_1 = \sum_{i=1}^{n} W_i Y_i$. This highlights the linearity of the estimator. The weights are given by $W_i = \frac{X_i - \bar{X}}{\sum_{j=1}^{n}(X_j - \bar{X})^2}$.

Variance of Regression Coefficients

We are likely not content just with reporting the estimated regression coefficients. Instead, we want to be able to give some sort of measure of how certain we are with regard to our estimates. Philosophically, under the frequentist framework of inference, we are going to conceptualize uncertainty in terms of how much our estimates would change if we were to draw a different sample, or specifically many different samples! The distributions of the estimated regression coefficients across different samples drawn from the population are called the *sampling distributions* of the regression coefficients. Note that this is a philosophical exercise, and in reality, we only ever observe one sample. However, our notions of uncertainty and everything that stems from it (such as testing and confidence intervals) are rooted in thinking about sample in the context of one of many possible samples that could have been drawn from the population.

Even though in reality, given one sample, we cannot simulate a sampling distribution for the estimated regression coefficients by drawing repeated samples from the population, we can still characterize the true variances of the estimated regression coefficients using theory:

$$\boxed{\mathrm{V}[\hat{\beta}_1] = \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

$$\boxed{\mathrm{V}[\hat{\beta}_0] = \frac{\sigma_\epsilon^2 \sum_{i=1}^n X_i}{n \sum_{i=1}^n (X_i - \bar{X})^2}}$$

Note the two quantities that make up the expression for variance for $\hat{\beta}_1$. The σ_ϵ^2 is the the variance of the true, population errors around the population linear conditional expectation function. $\sum_{i=1}^n (X_i - \bar{X})^2$ gives the squared deviation of the explanatory variable from its mean for a sample. While this quantity is actually defined on the sample - not the population - it changes across samples that we can possibly draw. As a result, these are random variables! Since we expressed the variances of the estimated regression coefficients above in the unconditional form (not conditioning on a particular sample that we obtained), we actually do not know $\sum_{i=1}^n (X_i - \bar{X})^2$, just like we do not know σ_ϵ^2.[2] Note also that $\sum_{i=1}^n (X_i - \bar{X})^2$ is closely related to the sample variance of X. Specifically $\sum_{i=1}^n (X_i - \bar{X})^2 = (n-1) S_X^2$.

The square root of the variances is the standard error of the estimated regression coefficients, which we can write respectively as $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$ for the two coefficients.

Theoretical Distributions of Estimated Regression Coefficients

We have already mentioned that $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased for their population counterparts. Moreover, we have expressed the variance of these estimators above. What is the distribution of these estimators? It turns out that both estimators are normally distributed given that we made the assumption that the outcome variable is normally distributed. Note that even if we didn't make this assumption, the estimators would tend towards a normal distribution asymptotically because of the Central Limit Theorem.[3]

[2] The difference between unconditional variances of the regression coefficients and variances conditional upon the sample we obtain are theoretically different quantities. While interesting, this fine point is rather advanced and beyond the scope of a basic introduction to regression.

[3] Loosely speaking, this follows from the fact that the estimators for β_1 and β_0 can be expressed as sequence of independent and identically distributed random variables.

As a result, we can express the theoretical distributions of the estimated regression coefficients as:

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma_\epsilon^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma_\epsilon^2 \sum_{i=1}^n X_i^2}{n \cdot \sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

Estimated Variances

Just given our sample, we do not know σ_ϵ^2, nor do we know $(X_i - \bar{X})^2$ - a quantity defined for a general sample, not the specific one we drew.

However, we can estimate the variances of the estimated regression coefficients by estimating σ_ϵ^2 with $\hat{\sigma}_\epsilon^2$ and using the X's from our sample as follows:

$$\hat{\mathbb{V}}[\hat{\beta}_1] = \frac{\hat{\sigma}_\epsilon^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\mathbb{V}}[\hat{\beta}_0] = \frac{\hat{\sigma}_\epsilon^2 \sum_{i=1}^n X_i}{n \sum_{i=1}^n (X_i - \bar{X})^2}$$

$\hat{\sigma}_\epsilon^2$ is an unbiased estimator of σ_ϵ^2 and is calculated as the sum of squared residuals from the regression on the sample we obtained, divided by $n - 2$, the degrees of freedom in the regression:

$$\hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{n-2} = \frac{SSR}{n-2},$$

where $\hat{\epsilon}_i$ is the observed residual on the ith observation in the sample.

The square root of $\hat{\sigma}_\epsilon^2$ is simply called the **standard error of the regression (SER)**. The square roots of the estimated variances for the estimated regression coefficients are the estimated standard errors for the estimated regression coefficients. This is the quantity that R reports and is an estimate of the variability in the estimated regression coefficients that captures how much they would change across different samples.

Note that for our regression example:
$n - 2 = 38$ (degrees of freedom)
$\hat{\sigma}_\epsilon^2 = 0.5465$ (residual standard error)
$\widehat{SE}(\hat{\beta}_0) = 0.4105$
$\widehat{SE}(\hat{\beta}_1) = 0.06412$

Testing

We are now going to turn to hypothesis testing, where we essentially want to see how much bearing our data has on making a decision between two hypotheses. Let's focus on hypotheses regarding β_1, the slope of the population LCEF. Let's set up a general hypothesis in the form:

$H_0 : \beta_1 = c$
$H_1 : \beta_1 \neq c$

What are we saying in words? The null hypothesis - H_0 - is postulating that the true population slope is equal to c. The alternative hypothesis - H_1 - postulates that β_1 is not equal to c. Our goal is to see whether or not the data we have allows us to reject the null hypothesis (and accept the alternative), or whether we cannot reject the null. To do so, we construct a test statistic as follows:

$$T = \frac{\hat{\beta}_1 - c}{SE(\hat{\beta}_1)}$$

The test statistic is a random variable since it's just a function of random variables. We can therefore ask how it's distributed *assuming the null hypothesis is true*. That is, assuming that β_1 is c, what would the distribution of the test statistic be? We know, from before, that in this case $\mathrm{E}[\hat{\beta}_1] = c$ and $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \mathrm{V}(\hat{\beta}_1))$. The test statistic therefore represents a standardization of the normal distribution. By subtracting off the mean of $\hat{\beta}_1$ and dividing by its standard deviation, we know that the resultant distribution will be a standard normal.

Therefore:

$$\boxed{T = \frac{\hat{\beta}_1 - c}{SE(\hat{\beta}_1)} \sim \mathcal{N}(0,1)}$$

However, note that we have to estimate $SE(\hat{\beta}_1)$ using $\widehat{SE}(\hat{\beta}_1)$. This estimation uses up 2 degrees of freedom, leaving us with $n-2$ degrees of freedom. To appropriately capture the lower certainty that we have due to this estimation, the test statistic is distributed as a t-distribution with $n-2$ degrees of freedom:

$$\boxed{T = \frac{\hat{\beta}_1 - c}{\widehat{SE}(\hat{\beta}_1)} \sim t_{n-2}}$$

We call the distribution of the test statistic under the null the **null distribution**.

Now, we can calculate the observed test statistic - T_{obs} for our sample by actually plugging in the values for $\hat{\beta}_1$ and $\widehat{SE}(\hat{\beta}_1)$ that we obtained from the formulas above. The goal is then to compare T_{obs} to the sampling distribution of the test statistic under the null. Intuitively, we want to know if our observed test statistic is a likely value given the distribution of the test statistic under the null. If it's a highly unlikely value, it gives us reason to believe that the null is not likely to be true, and therefore we reject it. Of course, what we specifically mean by "highly unlikely" is governed by our tolerance for Type I error - the probability of rejecting the null hypothesis when it is in fact true (often denoted as α). Once we select an α, we can obtain critical values for the null distribution that serve as thresholds for

rejecting or failing to reject the null (and appropriately capture the amount of Type I error we're comfortable with). If the observed test statistic falls outside the critical values, we say that we can reject the null with a significance level of α. If the observed test statistic falls within these thresholds, we say that we fail to reject the null at a significance level of α. Just to reiterate, we observe only one value of the test statistic and we compare it against a hypothesized distribution - the null distribution - that we would expect if the null was true.

In R and Stata, the default hypothesis test for the regression coefficients is testing whether the true regression coefficient is equal to zero. So, for β_1:

$H_0 : \beta_1 = 0$
$H_1 : \beta_1 \neq 0$

The observed test statistic we obtain in our example is:

$$T_{obs} = \frac{-0.5660 - 0}{0.06412} = -8.827$$

Now, suppose that we take the common significance level of $\alpha = 0.05$. We can calculate the critical values for the null distribution - which is a t distribution with $n - 2$ degrees of freedom - using the inverse CDF of the t-distribution such that there is $\alpha/2$ mass in each of the tails of the t-distribution (since this is a two-sided hypothesis test).

The critical values are thus: $t_{\alpha/2} = 2.0243$ and $-t_{\alpha/2} = -2.0243$.

Since T_{obs} falls outside the thresholds established by the critical values (it falls into the rejection region), we can reject the null hypothesis at a significance level of $\alpha = 0.05$. Note that we can see this graphically in Figure 2.

An alternative way to present the results of a hypothesis test is through a p-value. A p-value is defined as the probability of obtaining a test statistic at least as extreme that as the one that we observed for our sample assuming the null is true. Since the form of our hypothesis is that of a two-sided test, the corresponding p-value will need to convey notions of extremeness in both directions (both greater than $|T_{obs}|$ and less than $-|T_{obs}|$). More formally, a p-value is:

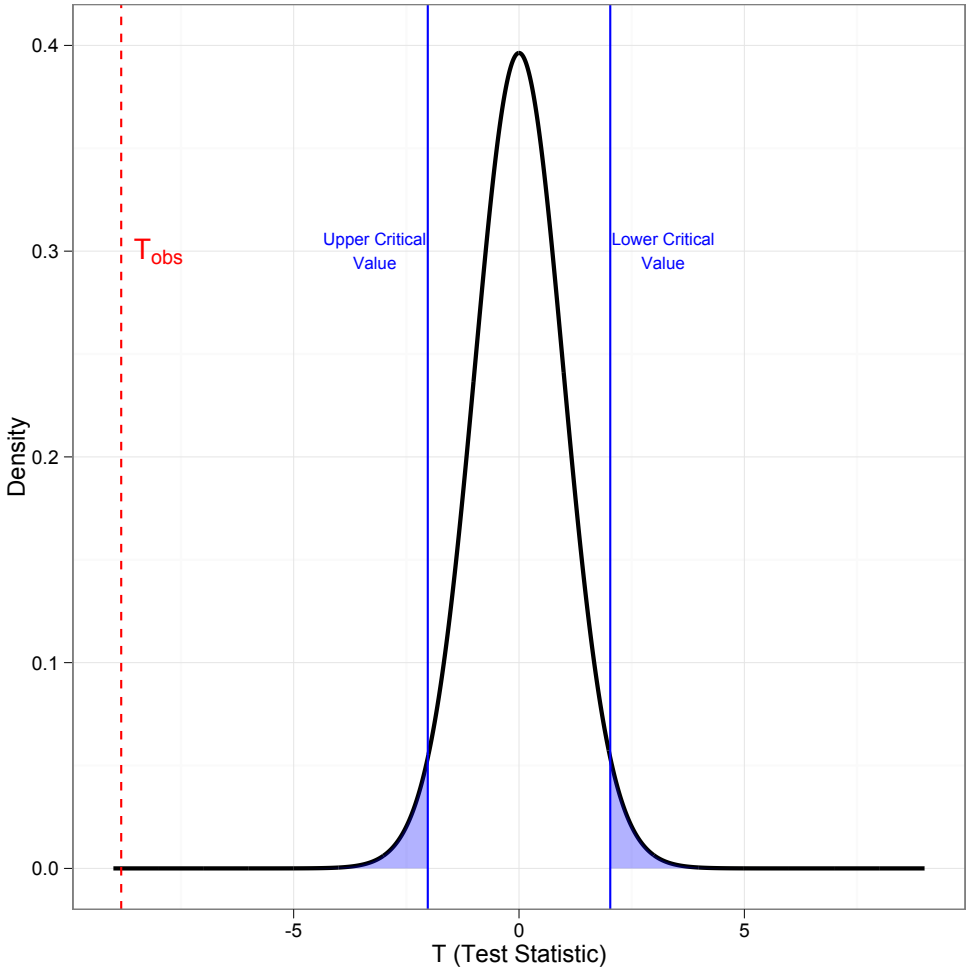
$$\boxed{p = P(|T| \geq |T_{obs}|) = 2 \cdot P(T \geq |T_{obs}|)}$$

In our example:
$p = P(|T| \geq 8.827) = P(T \geq 8.827) + P(T \leq -8.827)$

By symmetry of the t-distribution:
$p = 2 \cdot P(T \leq -8.827) = 9.72e - 11$

Since our p-value is less than $\alpha = 0.001$, we can state that we can reject the null hypothesis at a significance level of

Figure 2: Depiction of Hypothesis Test for $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. Test statistic is distributed as a t-distribution with $n - 2 = 38$ degrees of freedom. Significance level is $\alpha = 0.05$. Since this is a two-sided test, critical values are calculated so that there is $\alpha/2 = 0.025$ density in each of the tails (shaded in blue). Values outside of the critical values (corresponding to the tails shaded in blue) define the rejection region. The observed test statistic - T_{obs} - is also plotted. Since it falls in the rejection region, we reject the null hypothesis at the 0.05 significance level.



0.001.

Testing Non-Default Hypotheses

Note that even though the default hypothesis test in R or Stata is that the population parameter is equal to 0, there is no reason we can't do a hypothesis test where c is another value. For example:

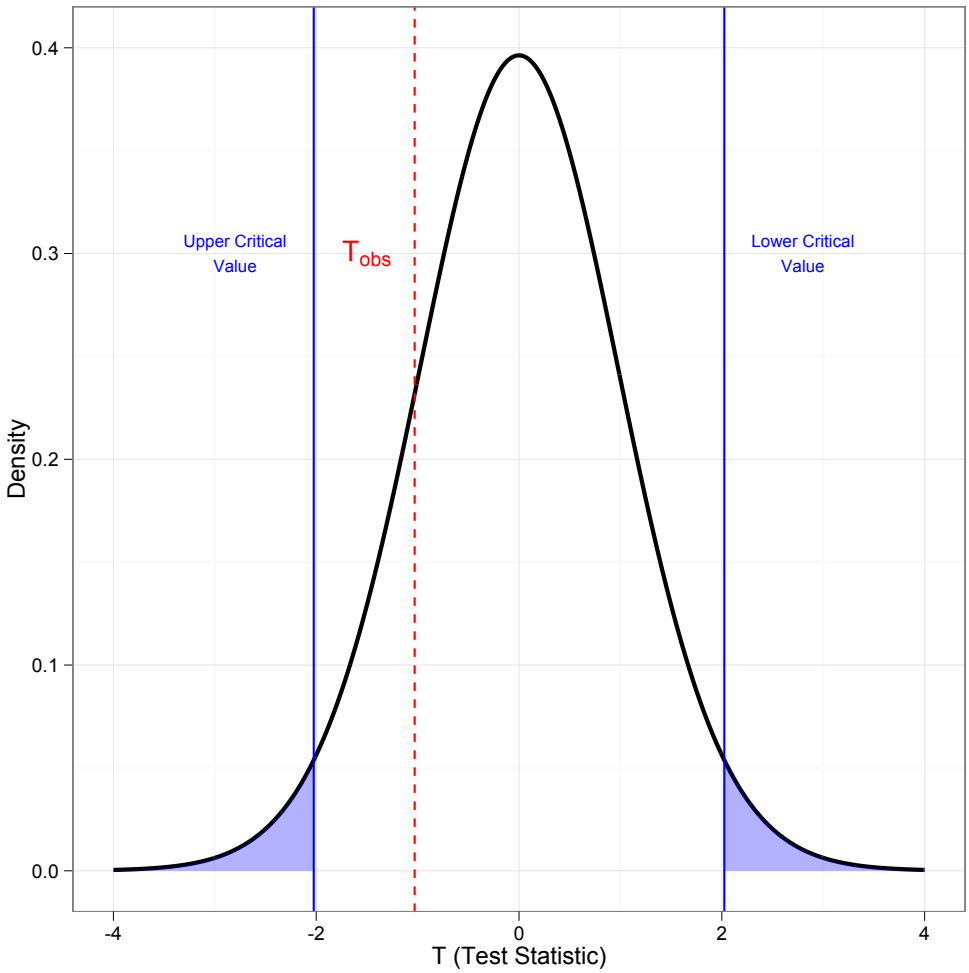
$H_0 : \beta_1 = -0.50$
$H_1 : \beta_1 \neq -0.50$

The observed test statistic we obtain in this example is:

$$T_{obs} = \frac{-0.5660 - (-0.50)}{0.06412} = -1.03$$

The observed test statistic does not fall within the rejection region of the test, and thus we fail to reject the null hypothesis at the $\alpha = 0.05$ significance level (note that we cannot say that we accept the null). This is presented graphically in Figure 3.

Figure 3: Depiction of Hypothesis Test for $H_0 : \beta_1 = -0.50$ versus $H_1 : \beta_1 \neq -0.50$. Test statistic is distributed as a t-distribution with $n - 2 = 38$ degrees of freedom. Significance level is $\alpha = 0.05$. Since this is a two-sided test, critical values are calculated so that there is $\alpha/2 = 0.025$ density in each of the tails (shaded in blue). Values outside of the critical values (corresponding to the tails shaded in blue) define the rejection region. The observed test statistic - T_{obs} - is also plotted. Since it does not fall in the rejection region, we cannot reject the null hypothesis at the 0.05 significance level.

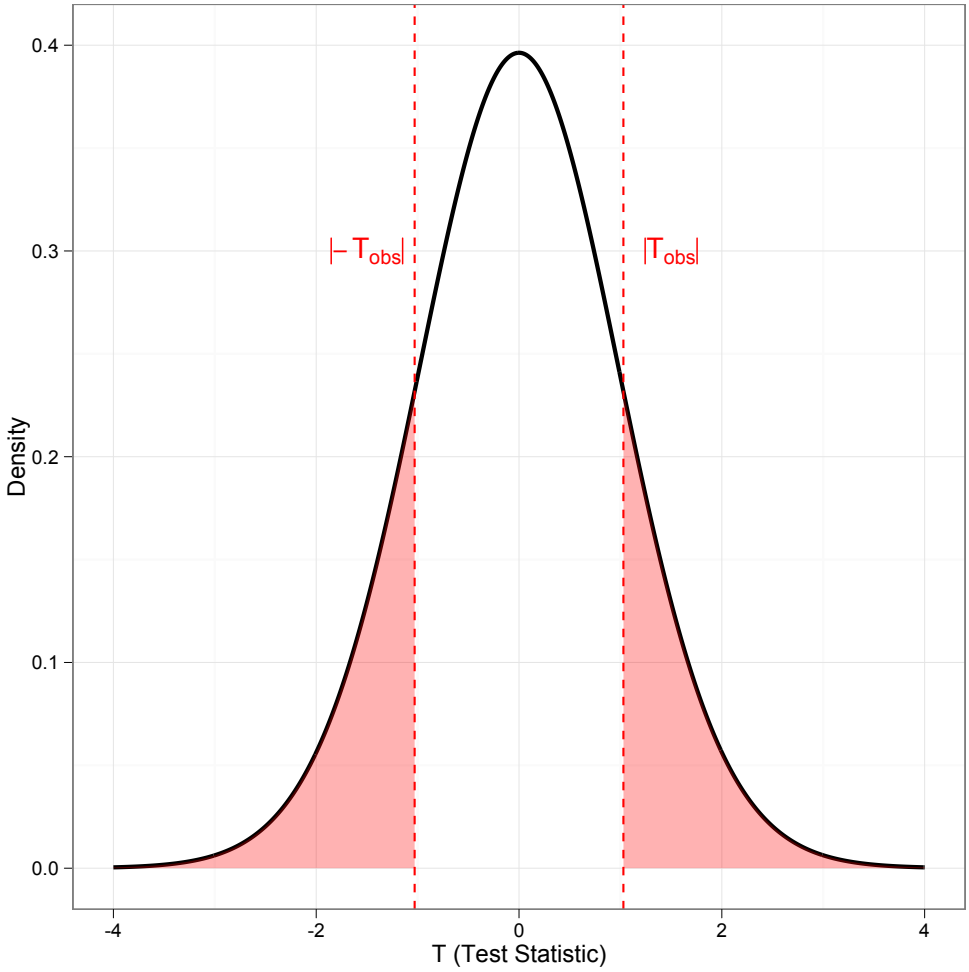


Finally, we can also present a p-value for this hypothesis test (intuition behind the p-value is presented graphically in Figure 4):

$p = 2 \cdot P(T \leq -1.03) = 0.31$

Since the p-value is greater than $\alpha = 0.05$, this confirms that we cannot reject the null at that significance level.

Figure 4: Depiction of p-value for $H_0 : \beta_1 = -0.50$ versus $H_1 : \beta_1 \neq -0.50$. Test statistic is distributed as a t-distribution with $n - 2 = 38$ degrees of freedom. The observed test statistic - T_{obs} - is plotted in red. P-value is the area shaded in red that represents the probability of obtaining a value of the test statistic at least as extreme as the one we actually observed.



Confidence Intervals

We often want to report the uncertainty around our estimates of the regression coefficients using confidence intervals. Recall the interpretation of the $1-\alpha$ level confidence interval for a parameter: it is constructed such that in $100 \cdot (1-\alpha)$ percent of samples you draw from the population, the confidence interval will contain the true value of the population parameter. However, a given confidence interval either contains the true population parameter or does not. Moreover, given that we only observe one sample, we don't know whether the confidence interval we construct from that sample for a parameter will contain the true value of the parameter - we know we have an $100 \cdot \alpha\%$ chance that it may not. This of course, is one of several ways in which confidence intervals are connected to hypothesis testing. α, in the hypothesis testing perspective, captures our tolerance for Type 1 error.

Suppose we want to construct a $100 \cdot (1-\alpha)\%$ confidence interval for β_1. Starting with the fact that $T = \dfrac{\hat{\beta}_1 - \beta_1}{\widehat{SE}(\hat{\beta}_1)} \sim t_{n-2}$, we can define critical values $\pm t_{1-\alpha/2}$ such that the following relationship is satisfied:

$$P\big(-t_{1-\alpha/2} \leq T \leq t_{1-\alpha/2}\big) = 1-\alpha$$

$$P\Big(-t_{1-\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}(\hat{\beta}_1)} \leq t_{1-\alpha/2}\Big) = 1-\alpha$$

Multiplying through by $\widehat{SE}(\hat{\beta}_1)$, subtracting off $\hat{\beta}_1$, and then reversing the inequality:

$$P\big(\hat{\beta}_1 - t_{1-\alpha/2} \cdot \widehat{SE}(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t_{1-\alpha/2} \cdot \widehat{SE}(\hat{\beta}_1)\big) = 1-\alpha$$

Note that $\pm t_{1-\alpha/2}$ are then the critical values such that $F_T(t_{1-\alpha/2}) = 1-\alpha/2$, where F_T is the CDF of the null distribution (t-distribution with $n-2$ degrees of freedom).

We can more concisely express the $100 \cdot (1-\alpha)\%$ confidence interval for β_1 as:

$$\boxed{\hat{\beta}_1 \pm t_{1-\alpha/2} \cdot \widehat{SE}(\hat{\beta}_1)}$$

For our example, supposing we want the commonly reported 95% confidence interval for β_1 which means that $t_{1-\alpha/2} = 1.96$:
$-0.566 \pm 1.96 \cdot (0.06412) = [-0.692, -0.440]$

Note the link between confidence intervals and hypothesis testing. The area contained of the $100 \cdot (1-\alpha)\%$ confidence interval actually provides the set of values $\mathcal{C}$ such that the null hypothesis $H_0 : \beta_1 = c$ cannot be rejected at the significance level of α for any $c \in \mathcal{C}$.