Econometrics I - Simple hypothesis testing

Ryan T. Godwin

University of Manitoba

$$H_0 : \beta_j = \beta_{j,0}$$
$$H_A : \beta_j \neq \beta_{j,0}$$

The decision to "reject" or "fail to reject" H_0 may begin by the researcher *subjectively* deciding on a *significance level* and then doing one or more of the following:

- Calculating a (p-value) and comparing it to the significance level.
- Seeing whether or not $\beta_{j,0}$ is contained in a confidence interval.
- Calculating a test statistic and seeing if it exceeds a critical value.

We need an *estimator* for the *standard error of the estimator* being used to assess the hypothesis.

For example, the t-test statistic for testing:

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

is:

$$t = \frac{b_i}{\hat{se}.(b_i)}$$

We need this quantity $\hat{se}.(b_i)$, which is called the *estimated standard error*.

Estimating σ^2

Know a lot about estimating $\boldsymbol{\beta}$. Another parameter in the model: σ^2 – the variance of each ϵ_i. We need to estimate σ^2 so that we can get an estimate for the covariance matrix of the LS estimator: $V(\boldsymbol{b}) = \sigma^2 (X'X)^{-1}$.

Let's derive an estimator for σ^2. Begin by noting that

$$\sigma^2 = \text{var}(\epsilon_i) = \text{E}\left[(\epsilon_i - \text{E}(\epsilon_i))^2\right] = \text{E}(\epsilon_i^2),$$

due to assumption A.3.

The sample counterpart to this population parameter (σ^2) is the sample average of the "residuals":

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\boldsymbol{e}'\boldsymbol{e},$$

which is the method of moments, and the maximum likelihood estimator. However, there is a distortion in this estimator of σ^2. Although the mean of the e_i's is zero (if there is an intercept in the model), not all of e_i's are independent of each other: only $(n-k)$ of them are.

We should consider what properties $\hat{\sigma}^2$ has as an estimator of σ^2, before we use it. Is this a *good* estimator? What properties of the LS estimator did we evaluate? We will write $\boldsymbol{e}'\boldsymbol{e}$ in terms of only $\boldsymbol{\epsilon}$ (which we have made assumptions about), and then derive its expected value:

$$\boldsymbol{e} = (\boldsymbol{y} - \widehat{\boldsymbol{y}}) = (\boldsymbol{y} - X\boldsymbol{b}) = \boldsymbol{y} - X\left(X'X\right)^{-1}X'\boldsymbol{y} = M\boldsymbol{y}$$

where

$$M = I_n - X\left(X'X\right)^{-1}X' \quad ; \quad \text{idempotent, and } MX = \mathbf{0}$$

So,

$$\boldsymbol{e} = M\boldsymbol{y} = M(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) = M\boldsymbol{\epsilon}$$

and

$$\boldsymbol{e}'\boldsymbol{e} = (M\boldsymbol{\epsilon})'(M\boldsymbol{\epsilon}) = \boldsymbol{\epsilon}'M\boldsymbol{\epsilon} \quad ; \quad \text{a scalar}$$

From this, it can be shown that:

$$\begin{aligned} E\left(\boldsymbol{e}'\boldsymbol{e}\right) = E\left[\boldsymbol{\epsilon}' M \boldsymbol{\epsilon}\right] &= E\left[\operatorname{tr}\left(\boldsymbol{\epsilon}' M \boldsymbol{\epsilon}\right)\right] = E\left[\operatorname{tr}\left(M \boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right)\right] \\ &= \operatorname{tr}\left[M E\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right)\right] = \operatorname{tr}\left[M\sigma^2 I_n\right] = \sigma^2 \operatorname{tr}(M) \\ &= \sigma^2(n-k) \end{aligned}$$

We won't cover the trace operator (tr), we will not discuss why $\operatorname{tr}(M) = \sigma^2(n-k)$. However, you need to be aware that an important step in considering whether an estimator should be used is to examine its *bias*, and in the case of $\hat{\sigma}^2$:

$$E\left[\hat{\sigma}^2\right] = E\left[\frac{1}{n}\boldsymbol{e}'\boldsymbol{e}\right] = \frac{1}{n}(n-k)\sigma^2 < \sigma^2$$

The method of moments and maximum likelihood estimator, $\hat{\sigma}^2$, is *biased*.

It is easy to convert this biased estimator to an *unbiased* one:

$$s^2 = \frac{1}{(n-k)} e'e$$

Some notes:

- $(n-k)$ is the "degrees of freedom" – number of independent sources of information in the n residuals (the e_i's).
- We can use s as an estimator of σ, but it is a biased estimator. Even though it is biased, s is typically used in practice as the bias is small.
- s is called the "standard error of the regression", or the "standard error of estimate".
- s^2 is a statistic. It has its own sampling distribution, etc.

Let's see one immediate application of s^2 and s. Recall the sampling distribution for the LS estimator, $\boldsymbol{b}$:

$$\boldsymbol{b} \sim N\left[\boldsymbol{\beta}, \sigma^2 \left(X'X\right)^{-1}\right]$$

So, the variance of the i^{th} LS estimator is the i^{th} diagonal of the covariance matrix of $\boldsymbol{b}$: $\text{var}\left(b_i\right) = \sigma^2 \left[\left(X'X\right)^{-1}\right]_{ii}$, but σ^2 is *unobservable*. If we want to report the variability associated with b_i as an estimator of $\boldsymbol{\beta}_i$, we need to use an estimator of σ^2. The estimated variance of the i^{th} LS estimator is then:

$$\widehat{\text{var}\left(b_i\right)} = s^2 \left[\left(X'X\right)^{-1}\right]_{ii}$$

The square-root of the above is called the "standard error" of b_i. This quantity will be very important when it comes to constructing *interval estimates* of our regression coefficients, and when we construct *tests of hypotheses* about these coefficients.

Standard errors in R.

Start by setting the random seed and sample size, and then generate some data:

```
set.seed(7010)
n <- 10
x <- rnorm(n)
y <- rnorm(n)
```

Estimate and summarize a model:

```
mod <- lm(y ~ x)
summary(mod)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2472     0.3734   0.662    0.526
x            -0.2455     0.6348  -0.387    0.709

Residual standard error: 1.178 on 8 degrees of freedom
Multiple R-squared:  0.01835, Adjusted R-squared:  -0.1044
F-statistic: 0.1495 on 1 and 8 DF,  p-value: 0.7091
```

R is reporting the standard errors of b_1 and b_2 as 0.3734 and 0.6348 respectively. We will see one way how statistical packages can calculate these numbers. Start by getting the estimate s^2:

```
s2 <- sum(mod$residuals ^ 2) / (n - 2)
```

Note that $k = 2$ above. If we take the square root, we get the "residual standard error" reported in the R output above:

```
sqrt(s2)
```

```
[1] 1.178469
```

Next we will calculate the $V(\boldsymbol{b})$ matrix. Start by arranging the x data into a matrix (and take a look at the X matrix):

```
X <- matrix(c(rep(1, n), x), n, 2)
X
```

```
      [,1]       [,2]
 [1,]    1 -0.2214732
 [2,]    1  0.6051370
 [3,]    1  0.7208573
 [4,]    1 -0.2230900
 [5,]    1 -0.2662395
 [6,]    1 -1.0890823
 [7,]    1 -0.5655553
 [8,]    1  0.5330395
 [9,]    1  0.6225352
[10,]    1 -0.4755066
```

The $s^2(X'X)^{-1}$ matrix is then:

```
s2 * solve(t(X) %*% X)
```

```
              [,1]        [,2]
[1,] 0.13939930 0.01448197
[2,] 0.01448197 0.40297318
```

I have used the `solve()` function to find the inverse. Taking the square root of any of the diagonal elements gives the standard error reported in the `summary()` output above:

```
sqrt(s2 * solve(t(X) %*% X))[1, 1]
```

```
[1] 0.3733622
```

In order to assess the variability of the estimator in relation to a point estimate, we'll need the full sampling distributions of both $\boldsymbol{b}$ and s^2. Note that assumption A.6 will be particularly important in what follows. Recall that:

$$\boldsymbol{b} \sim N\left[\boldsymbol{\beta}, \sigma^2 \left(X'X\right)^{-1}\right]$$

and that because the marginal distribution from the multivariate-Normal distribution is still Normal:

$$b_i \sim N\left[\beta_i, \sigma^2 \left(\left(X'X\right)^{-1}\right)_{ii}\right]$$

Suppose that your null hypothesis is $H_0 : \beta_{2,0} = 0$[1] (you think that x_2 has *zero effect* on y), but that the value you estimate for β_2 is $b_2 = 2,100,000$. What can guide you in your decision to reject or fail-to-reject H_0? One possibility is to use a p-value.

p-value The probability of obtaining an estimate more adverse to the null hypothesis, compared to the estimate just obtained, provided the null hypothesis is true.

[1] The 0 in the subscript (after the comma) is to denote that this is the value for the parameter under the *null hypothesis*.

Using properties of the mean and variance[2], we can *standardize* b_i:

$$z_i = \frac{(b_i - \beta_{i,0})}{\sqrt{\sigma^2 \left[(X'X)^{-1}\right]_{ii}}} \sim N(0,1)$$

All we are accomplishing here is creating a *test statistic*; altering the distribution of b_i to one that is standard Normal, **provided the null hypothesis is true**. If H_0 is false, then the mean of z_i is not 0. Calculating a standardized test statistic is a very common way of obtaining a p-value, but it is not necessary. To see that standardization is not needed to obtain the p-value, note that $Pr(z_i > 0) = Pr(b_i > \beta_{i,0})$.

Why bother standardizing?

[2] (i) $\mathrm{E}[c + Y] = c + \mathrm{E}[Y]$; (ii): $\mathrm{var}[cY] = c^2\,\mathrm{var}[Y]$, where c is a constant and Y is a random variable

Problem: z-statistic can't be calculated in practice, since the value for σ^2 is unknown. Instead, we will have to replace the unknown σ^2 with an estimator: s^2 for example.

When we introduce a random variable into the denominator of the z-statistic, we change the distribution to something that is *not* Normal. Given all of our assumptions A.1 to A.6, it turns out that the test statistics is transformed so that it follows the t-distribution. We will not fully prove this, but instead sketch out the proof.

- Definition: let $z_1, z_2, \ldots, z_m$ be independent $N(0,1)$ random variables. Then the quantity $\sum_{i=1}^{m} \left(z_i^2\right)$ has a Chi-square distribution with m degrees of freedom, χ_m^2.
- Note: $\boldsymbol{e}'\boldsymbol{e} = \boldsymbol{\epsilon}' M \boldsymbol{\epsilon}$ is a sum of squared Normal variables.
- Note: not all of the $\boldsymbol{e}$ are independent, only $(n-k)$ of them. That is, the degrees of freedom in $\boldsymbol{e}$ is $(n-k)$.
- This leads to the distribution for s^2:

$$\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{(n-k)}$$

- Definition: let $z \sim N(0,1)$, and let $x \sim \chi_v^2$, where z and x are independent. Then the statistic, $t = z/\sqrt{x/v}$ follows the Student's t-distribution, with "v" degrees of freedom.
- Notice the t-statistic has a Normal variable in the numerator (b_i) and the square root of a chi-square variable in the denominator (s^2). With some re-arranging, it can be shown that this statistic matches the definition for a t-distribution.

$$t_i = \frac{(b_i - \beta_{i,0})}{\sqrt{s^2 \left[(X'X)^{-1}\right]_{ii}}} = \frac{(b_i - \beta_{i,0})}{\widehat{s.e.(b_i)}} \sim t_{(n-k)}$$

- Finally, note that for large n, the t-distribution becomes the standard Normal distribution.

Suppose the estimated model is:

$$\begin{array}{ccccccc} \hat{y} & = & 1.4 & + & 0.25x_2 & + & 0.6x_3 \\ & & (0.7) & & (0.1) & & (1.4) \end{array}$$

with

$$H_0 : \beta_2 = 0 \quad \text{vs.} \quad H_A : \beta_2 \neq 0.$$

The t-statistic associated with this hypothesis test is:

$$t = \left[\frac{b_2 - \beta_2}{s.e.\,(b_2)} \right] = \left[\frac{0.25 - 0}{0.1} \right] = 2.5$$

Now we must determine the p-value, or compare this "2.5" to a *critical value*. Suppose that $n = 20$. Then, $t \sim t_{(17)}$. The p-value can be easily obtained from R using

```
2 * (1 - pt(2.5, 17))
```

```
[1] 0.02294781
```

What is the interpretation of 0.02295 here? If H_0 is true, there is only a 2.3% chance of obtaining a b_2 that is "further away" from 0 then what was just observed (a "distance" of 0.25). Things can't get much worse. Either: (i) we obtained a strange sample, or (ii) H_0 is false. With a *significance level* of $\alpha = 5\%$, we would reject H_0.

Critical values

We could also perform this hypothesis test using a *critical value*. Recall that for a significance level of 5%, the critical value from the Normal distribution is 1.96. This just means that if you obtain $|z_i| > 1.96$, you will obtain a p-value less than 5%. So, we *could* compare $t = 2.5$ to 1.96, except the sample size isn't large enough!

Confidence Intervals

We can also use our t-statistic to construct a confidence interval for β_i. Note that if H_0 is true, then the probability that the t-statistic lies within the α critical range is $(1-\alpha)$:

$$\Pr.\ [-t_c \leq t \leq t_c] = (1-\alpha).$$

For example, if n is large and $\alpha = 0.05$, then there is a 0.95 probability that t lies within the values -1.96 and 1.96, provided H_0 is true. Now, substitute the formula for t into the above probability statement:

$$\Pr.\left[-t_c \leq \left[\frac{b_i - \beta_i}{s.e.\,(b_i)}\right] \leq t_c\right] = (1-\alpha),$$

and now solve the inequality so that β_i is in the centre:

$$\begin{aligned}
&\Pr.\,[-t_c \times s.e.\,(b_i) \leq (b_i - \beta_i) \leq t_c \times s.e.\,(b_i)] = (1-\alpha)\\
&\Pr.\,[-b_i - t_c \times s.e.\,(b_i) \leq (-\beta_i) \leq -b_i + t_c \times s.e.\,(b_i)] = (1-\alpha)\\
&\Pr.\,[b_i + t_c \times s.e.\,(b_i) \geq \beta_i \geq b_i - t_c \times s.e.\,(b_i)] = (1-\alpha)\\
&\Pr.\,[b_i - t_c \times s.e.\,(b_i) \leq \beta_i \leq b_i + t_c \times s.e.\,(b_i)] = (1-\alpha)
\end{aligned}$$

Interpretation of the confidence interval

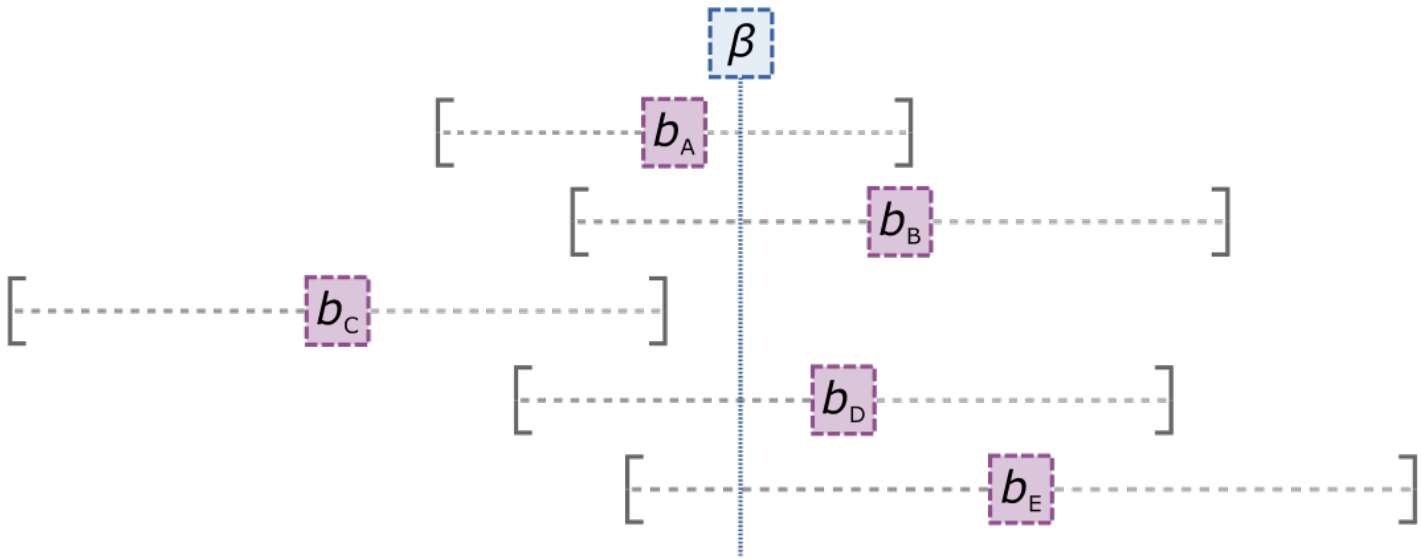
$$\left[b_i - t_c \times s.e.\left(b_i\right), \quad b_i + t_c \times s.e.\left(b_i\right)\right],$$

is *random*. The parameter, β_i is *fixed*, but unknown.

- If we were to take a sample of n observations, and construct such an interval, and then repeat this exercise many, many times, then $100(1-\alpha)\%$ of such intervals would cover the true value of β_i.
- Such an interval contains all the values for the parameter under the null hypothesis, that will not be rejected.

If we just construct an interval, for our given sample of data, we'll never know if this particular interval covers β_i, or not.

Figure: Each hypothetical sample of size n that we could draw (sample A, sample B, etc.) provides a 95% confidence interval that has a 95% probability of containing the true population β. In practice, we will only draw one sample from the population, and calculate one interval.



Some Properties of Tests

Classical hypothesis testing:

- Assume that H_0 is *true*
- Compute value of test statistic using random sample of data
- Determine distribution of the test statistic (when H_0 is true)
- Check if observed value of test statistic is likely to occur, if H_0 is true
- If this event is sufficiently unlikely, then reject H_0 (in favour of H_A)

Note:

1. Can never accept H_0. Why not?
2. What constitutes "unlikely" – subjective?
3. There are two types of errors we might incur with this process.

Type I and II error

- Type I Error: **Reject** H_0 when in fact it is **true**.
- Type II Error: **Do not reject** H_0 when in fact it is **false**.

Prob. type I error is:

- denoted by α
- is also the *significance* level of the test
- (sometimes also called the "size" of the test)

The significance level of the test is a predetermined max p-value before which H_0 is rejected. If p-val $< \alpha$, H_0 is rejected. In deciding on this maximum acceptable p-value, we are also determining the type I error. Even when H_0 is true, there is an $\alpha\%$ probability of drawing an "extreme" sample that will lead to a incorrect rejection of H_0.

Prob. of a type II error is sometimes denoted by β. β typically will not be known, since β will depend on *how* H_0 is false. Usually, there are many ways. For example, H_0 could be false because the true parameter value is very far away from the value under the null; or it could be that the truth is only a little different from H_0.

Type II error is theoretically useful for designing or *choosing* a testing procedure. You are likely familiar with the t-test and F-test, but why do we use these tests? Under certain assumptions, these tests have *desirable properties*. We want to use tests that are *powerful*, *unbiased*, and *consistent*. Note, however, that the properties *unbiased* and *consistent* have different meanings depending on whether we are talking about a *test* or an *estimator*.

Power

The "power" of a test is $Pr.[\text{Reject } H_0 | H_0 \text{ is false}]$. So,
Power $= 1 - Pr.[\text{Do not reject } H_0 | H_0 \text{ is false}] = 1 - \beta$.

Depending on the *way* that H_0 is false, we typically have a **Power Curve**. For a fixed value of α, this curve plots Power against parameter value(s). We want our tests to have *high power*, and we want the power of our tests to *increase* as H_0 becomes *increasingly false*. Now let's consider some desirable properties for a test.

Property 1 - UMP

Consider a fixed sample size, n, and a fixed significance level, α. Then, a test is "Uniformly Most Powerful" if its power exceeds (or is no less than) that of any other test, for all possible ways that H_0 could be false.

Property 2 - consistent (test)

Consider a fixed significance level, α. Then, a test is "consistent" if its Power $\to 1$, as $n \to \infty$, for all possible ways that H_0 is false.

Property 3 - unbiased (test)

Consider a fixed sample size, n, and a fixed significance level, α.
Then, a test is said to be "unbiased" if its power never falls below the significance level.

Property 4 - LMP

Consider a fixed sample size, n, and a fixed significance level, α. Then, a test is said to be "Locally Most Powerful" if the slope of its Power curve is greater than the slope of the power curves of all other size α tests, in a neighbourhood of H_0.

Note:

- For many testing problems, no UMP test exists. This is why LMP tests are important.
- Why do we use our "t-test" in the regression model? Because it has properties 1 - 3 against one-sided alternative hypotheses, and has properties 2 - 4 against two-sided alternatives. Similar to how the LS estimator is "best" among other estimation alternatives (given certain assumptions), so is the t-test "best" among other potential testing strategies.