1. Hypothesis Testing: Big Picture

We went from talking about assumptions for unbiased estimates, regression results, and R^2 to suddenly reverting back to basic statistics and hypothesis testing. Why? The reason is because up until now, we haven't really needed to use statistics to calculate a fitted regression line. Now, we need to use statistics to tell us how much confidence we have that our results aren't just random noise. A point estimate from a particular sample does not, by itself, provide enough information for testing economic theories or for informing policy discussions. A point estimate may be the researcher's best guess at the population value, but it provides no information about how close the estimate is "likely" to be to the population parameter. So OLS gives us the best possible fit of our model to our sample data, but it may not be capturing the true relationship.

For example, suppose you receive some data on a random sample of workers from New York. You find that workers who receive on-the-job training have higher hourly wages (5%). Can you say anything about whether or not this is close to the effect we would detect in the population of workers who could have been trained? Not as is, no. However, we can make statements involving probabilities with confidence intervals and hypothesis testing to figure out how far our estimator is likely to be from the true population value.

2. Hypothesis Testing: Nuts and Bolts

Last time, we discussed how to construct/interpret confidence intervals. Sometimes though we will want to test the validity of a hypothesis we have about the population. We might want to test something simple like "Is the average morning commute time from Berkeley to SF greater than 45 minutes" (a sample mean) or "Whether vocational training schools are effective at increasing employment rates" (a regression coefficient) for example. Enter hypothesis testing. Very broadly, we will use our sample of data to test whether some hypothesis we have about the true population is likely or not.

a. How to test a hypothesis

1. State the **null hypothesis,** H_0: this is the hypothesis we assume to be true until the data strongly suggests otherwise. We must also state the **alternative hypothesis,** H_1 that corresponds.

   - For example, we can test the null that the parameter of interest is zero $H_0 : \gamma = 0$ against the alternative that $H_1 : \gamma \neq 0$ for some parameter γ. This could be the population mean μ, or a population coefficient β.

2. To test the null hypothesis against the alternative, we need to calculate a test statistic. A **test statistic** is some function of the random sample.

   - As the test statistic (t) is a function of the random sample, it takes different values in different samples and we can plot its distribution. The specific value of the test statistic that we calculate for our given sample will give us some measure of how far our estimate of γ is from the value we give in null hypothesis. More formally, it measures the distance of $\hat{\gamma}$ (the value of γ we calculate for our given sample) to γ relative to the standard deviation of our estimate $\hat{\gamma}$ (also known as the standard error).
   - By the CLT, we know that if a parameter γ has mean μ and sample estimates have variance σ^2, then we know the distribution of sample estimates is $\hat{\gamma} \sim N(\mu, \sigma^2)$, and equivalently $(\hat{\gamma} - \mu)/\sigma \sim N(0,1)$. If we don't know the estimator variance σ^2 but have an estimator

$\hat{\sigma}^2$, we can still say $(\hat{\gamma} - \mu)/\hat{\sigma} \sim t_{n-k-1}$, where k is the number of parameters used in estimating your model. Because we can easily calculate critical values for these distributions, we use these functions $(\hat{\gamma} - \mu)/\hat{\sigma}$ to calculate our test statistics.

- It is important to note that in this case the σ^2 we refer to is the variance for the *estimator*: for example, for a sample mean it is σ_x^2/n, and for a coefficient β it is σ_u^2/SST_x.

3. The value of the test statistic is compared to the **critical value** that we choose ()in order to reject H_0 in favor of H_1, we must have evidence "beyond reasonable doubt" against H_0 - and the following criteria is how we define "beyond a reasonable doubt"). The critical value is determined once we choose the **significance level,** α of our test. The significance level of a test is the probability that we reject the null when the null is true. By choosing a particular value of α, we are essentially quantifying our tolerance for rejecting the null when the null is true. $\alpha = .05$, is the most common value, but $\alpha = .01$ and $\alpha = .1$ are also common.

    - c_α, the critical value associated with α, is determined by the distribution of the test statistic, assuming that the null is true. The three distributions we'll see in this course are t, z, and F (which we'll get to a bit later) and you can always look these critical values up in the corresponding tables.

4. The rejection rule (when we will reject the null) is then

    - For two sided tests:[1] $|t| > |c_{\alpha/2}|$
    - For positive one-sided tests: $t > c_\alpha$
    - For negative one-sided tests: $t < c_\alpha$ (where c is negative)

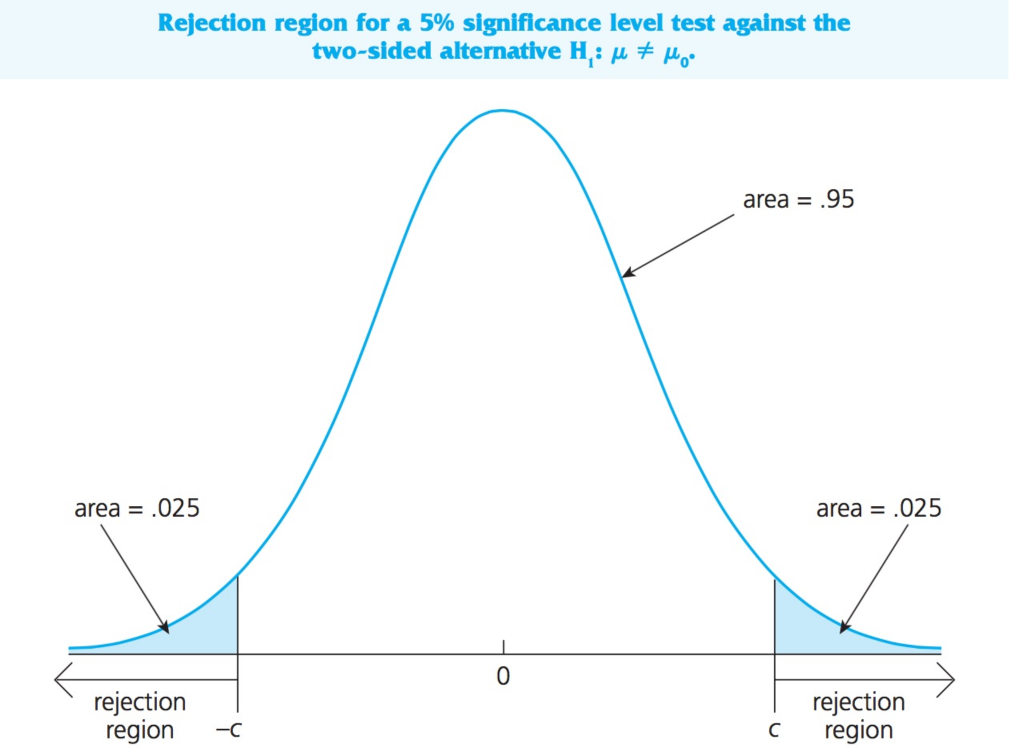
    Intuition: if $\hat{\gamma}$ is sufficiently far from our hypothesized value of γ (which we see when it is more than c standard deviations away from γ, i.e., we satisfy the rejection rule) then we have evidence against H_0 in favor of H_1. This means that if the null hypothesis were true, we would only observe a test statistic larger than c_α with probability α.
    Note: You can generally just reject if $|t| > |c|$. In only a few rare cases would the rejection rule for one-sided tests lead to a different conclusion.

5. Interpret

    - Reject the null (satisfy the rejection rule based on your calculated t and your critical value c) OR
    - Fail to reject the null

    Note that we never *accept* the null hypothesis.

[1] Our test is two-sided when $H_1 : \mu \neq 0$. It's postitive one-sided if $H_1 : \mu > 0$ and negative one-sided if $H_1 : \mu < 0$.


Rejection region for a 5% significance level test against the two-sided alternative H_1: $\mu \neq \mu_0$.
area = .95
area = .025
area = .025
0
rejection region
−c
c
rejection region

b. The formula through an example

We will review these series of steps that are necessary to conduct hypothesis testing through the use of an example. Let's think again about Michigan State University undergraduates. We collect a random sample of the MSU student body and ask them their GPA. The dean of MSU firmly believes the true average GPA of her students is 3.1. We want to test this hypothesis.

Step 1: Define hypotheses:

Here we write down the null and alternative hypotheses in precise mathematical language. The dean's hypothesis is that the true average GPA is 3.1. That is what we're trying to test. Thus, our null hypothesis is that the true MSU average GPA, which we call μ, is 3.1:

$$H_0 : \mu = 3.1$$

Since we don't have a good reason to think the average GPA should be higher or lower than 3.1, our alternative hypothesis is:

$$H_1 : \mu \neq 3.1$$

Step 2: Compute test statistic:

To compute the t-statistic, we need the sample estimate of the mean $\bar{x}$, the sample variance s_x^2, and the sample size n: Suppose we collect data from 101 students, with a mean GPA of 2.984 and a standard deviated of 0.3723

$$\begin{aligned} \bar{x} &= 2.984 \\ s_x &= 0.3723 \\ n &= 101 \\ \Rightarrow t &= \frac{\bar{x} - \mu}{\frac{s_x}{\sqrt{n}}} = \frac{2.984 - 3.1}{\sqrt{\frac{0.13861}{101}}} = -3.13 \end{aligned}$$

The **formula for the t-stat** comes from what we learned about the distribution of sample estimators last section. We know that $\frac{\bar{x}-\mu}{\sigma_x/\sqrt{n}} \xrightarrow{d} N(0,1)$, where we use $\sigma_x/\sqrt{n}$ because this is the standard deviation for a sample mean. After replacing σ_x with the sample estimator s_x, we have $\frac{\bar{x}-\mu}{s_x/\sqrt{n}} \xrightarrow{d} t_{n-1}$. The intuition behind the t-stat: we want to reject H_0 in favor of H_1 when the value of the sample average $\bar{x}$ is sufficiently greater than 3.1. However, rather than working directly with $\bar{x}$, we convert to standard deviation units, assuming the null hypothesis is true. t measures how many standard deviations away from the null hypothesis value of μ our estimate $\bar{x}$ is.

Step 3: Choose significance level (α) and find the critical value:

Next we need to choose a significance level. Again, the significance level is equivalent to the probability of rejecting the null when the null is true. Generally we want to make this error fairly small. Let's choose significance level $\alpha = 0.05$. Check the t-table for the critical value when our significance level for a 2-tailed test is $\alpha = 0.05$ and we have $n - 1 = 100$ degrees of freedom:

$$c = 1.984$$

Step 4: Reject the null hypothesis or fail to reject it:

Intuitively we want to reject H_0 if $\bar{x}$ is far enough away from 3.1. Recall that this is a two-sides or two-tail test.

- If $|t| > |c|$, then we reject the null hypothesis. We established above that *if the null hypothesis were true* and the true mean is 3.1, then the probability of observing a sample mean with a t-stat of $|t| > |c|$ is equal to the significance level, or 5%. How do we interpret this? Either we obtained a very unusual sample or the null hypothesis should be rejected. So if we do observe a sample mean where the associated t-stat is such that $|t| > |c|$, then we say we reject the null hypothesis.

- If $|t| < |c|$, then we fail to reject the null hypothesis. Note again that we never "accept" the null hypothesis, we simply do or do not find evidence against the null (more on this in Section 7). In our example:
$$|t| = |-3.13| = 3.13 > 1.987 = c \quad \Rightarrow \text{Reject the null}$$

Step 5: Interpret

- If we *reject* the null hypothesis: There is statistical evidence at the 5% level that the average GPA at MSU is different from 3.1. It's unlikely that the dean's assertion is correct.

- If we *fail to reject* the null hypothesis: There is no statistical evidence at the 5% level that the average GPA at MSU is different from 3.1. It's possible that the dean's assertion is correct.

Note on the **relationship between confidence intervals and hypothesis testing**: once you construct a confidence interval you can conduct hypothesis testing. Ex: we construct a 95% confidence interval for μ based on our sample and find that our hypothesis for μ, $(\mu = 3.1)$ is not in the confidence interval. Then we reject the null.

3. Hypothesis Testing: Two Variations

a. Hypothesis testing and binary variables

i. Conceptually

When $\bar{X}$ is actually a proportion, i.e. the average of a binary variable. Hypothesis tests for binary variables vary from the steps outlined above in two ways:

1. Under the null hypothesis, we know the true proportion, p. Recall the formula given for the mean and variance of a binary variable—knowing the mean implies that you know the variance. Thus, when we use the null hypothesis to compute the test statistic, we don't use the sample variance. Instead we use the variance specified by the null hypothesis.

$$\mathbb{E}[X] = p$$
$$Var(X) = p(1-p)$$

2. Remember that if we know the true mean *and* the true variance, we use a normal distribution instead of a t-distribution. So to find the critical value, we use the z-table instead of the t-table.

ii. Example

Suppose that a military dictator in an unnamed country holds a plebiscite (a yes/no vote of confidence) and claims that he was supported by 65% of the voters. Call X the binary voting variable. An NGO suspects that the dictator is lying and contracts you, a skilled econometrician, to investigate this claim. Specifically you are asked to investigate if the dictator received less than 65%. You have a small budget, so you can only collect a random sample of 200 voters in the country. From your sample of 200, you find that 115 supported the dictator, so the sample proportion $\hat{p} = 0.575$. You also decide to be very conservative and choose the 1% significance level

Step 1: State the null and the alternative hypothesis

$$H_0 : p = 0.65$$
$$H_1 : p < 0.65$$

(The reason why we're using a 1-tailed test here is because we suspect the proportion supporting the dictator is below 65%.)

Step 2: Calculate the z-statistic

We start by calculating the standard deviation for a sample mean:

$$\sqrt{\sigma_x^2/n} = \sqrt{p(1-p)/n} = \sqrt{.65(1-.65)/200} = 0.033727$$

Therefore our test statistic: means that our test statistic:

$$z = \frac{\hat{p} - p}{\sigma_x / \sqrt{n}} \sim N(0,1)$$
$$z = \frac{\hat{p} - p}{\sigma_x / \sqrt{n}} = \frac{.575 - .65}{0.033727} = -2.224$$

Step 3: Choose a significance level and a critical value

We choose a 1% significance level , so $\alpha = .01$. Then using the normal distribution tables, we know that $c = -2.32$

Step 4: Reject the null hypothesis or fail to reject it

Recall, that we reject null hypothesis if $z < c$. Therefore, here we *fail to reject* since $-2.224 > -2.32$.

Step 5: Interpret

We fail to reject the null hypothesis (at the 1% significance level) that the proportion of people who support the dictator is 65%. In this sample and significance level, we do not have the statistical evidence to conclude that the dictator is fixing the vote results.

b. Hypothesis testing and the difference in two means

i. Conceptually

Hypothesis tests for a difference in means vary from the steps outlined above in one way:

1. To compute the test statistic, we use the standard deviation of the *difference*.

ii. Example, assuming population variances are not equal

Suppose we want to know whether income per capita in rural areas equal to that in urban areas? Call $I = in_u - in_r$ the true difference in income between rural and urban. An estimator for I is $\bar{I} = \overline{in_u} - \overline{in_r}$.

We have the following data:

	Urban Income	Rural Income
Mean	7061.576	3661.307
Standard Error	8973.852	5197.585
N Observations	1112	1112

Step 1: State the null and the alternative hypothesis

$$H_0 : in_u - in_r = I = 0$$
$$H_1 : in_u - in_r = I \neq 0$$

Step 2: Calculate the t-statistic

How do we compute this test statistic? Whenever we're testing a difference of means, remember the formula: $Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y})$.

So applying the formula, we have that:

$$Var(\overline{in_u} - \overline{in_r}) = Var(\overline{in_u}) + Var(\overline{in_r}) = \frac{\sigma_u^2}{n_u} + \frac{\sigma_r^2}{n_u}$$

$$\widehat{Var(\overline{in_u})} = \frac{s_u^2}{n_u} = \frac{5198^2}{1112}$$

$$\widehat{Var(\overline{in_r})} = \frac{s_r^2}{n_r} = \frac{8974^2}{1112}$$

$$\longrightarrow se(\overline{in_u} - \overline{in_r}) = \sqrt{\frac{s_u^2}{n_u} + \frac{s_r^2}{n_r}} = \sqrt{\frac{5198^2}{1112} + \frac{8974^2}{1112}} = 310$$

Moreover,

$$\bar{I} = \overline{in_u} - \overline{in_r} = 7061.6 - 3661.3 = 3400.3$$

Now we're ready to calculate our t-statistic:

$$\Rightarrow \quad t = \frac{3400.3 - 0}{310} = 10.97 \sim t_{n_u + n_r - 2 = 1112 + 1112 - 2}$$

Note that the degrees of freedom is equal to the sum of the urban and rural samples (the total sample) minus 2, the number of estimators we are considering (the means for the rural and urban samples).

Step 3: Choose a significance level and a critical value

By the null hypotheses we chose, we're doing a two-sided test. Let's choose the 1% significance level as this. Check the t table to find that $c = 2.576$

Step 4: Reject the null hypothesis or fail to reject it

$10.97 > 2.576$: reject H_0 at the 1% significance level that urban and rural per capita incomes are equal in favor of urban income per capita being larger than rural income per capita.

Step 5: Interpret

Interpret: At the 1% significance level, there is statistical evidence that urban income per capita is larger than rural income per capita

iii. Example

Now let's consider an example from actual data for a poverty alleviation program in Mexico. In 1997, 24,059 households in rural Mexico were randomly allocated between treatment and control groups for a conditional cash transfer program called Oportunidades to keep kids in school. When analyzing the results of a randomized experiment, the first step is to verify that the control group is, on average, very much like the treatment group in terms of characteristics that we observe and have data for. For example, data was collected on household assets. The data reveal that while 14.47% of the 14,846 treatment households have a refrigerator, and 16.53% of the 9,213 control households have one. In order to confirm that about the same proportion of households in each group have a refrigerator, we need to perform a hypothesis test.

Call the sample proportion of households with a refrigerator in the treatment group $\hat{p}_t$, the true treatment proportion p_t, the sample proportion of households with a refrigerator in the control group $\hat{p}_c$, and the true control proportion p_c. Also, call the whole sample proportion of households (in either treatment or control) with a refrigerator $\hat{p}$.

Step 1. Note we don't have a null about *one* mean, we have a null about the difference in means.

$$\begin{aligned} H_0 : & p_t - p_c = D = 0 \\ H_1 : & p_t - p_c = D \neq 0 \end{aligned}$$

Recall that with Bernoulli variables when we have a hypothesis about the mean, we also have a hypothesis about the variance. Here we are testing the null that means are the same, and it follows that under this hypothesis the variance for both groups is the same as in the population.

Step 2. How do we compute this test statistic? We know that the null hypothesis specifies $\mathbb{E}[p_t - p_c] = 0$, so what's left is the standard deviation. Whenever we're testing a difference of means, remember the formula: $Var(\bar{x} - \bar{y}) = Var(\bar{x}) + Var(\bar{y})$.

So applying the formula (and plugging in the sample variance as our estimate of the population variance, since in this case we don't have a null hypothesis about the population value of p), we have that:

$$\begin{aligned} Var(\hat{p}_t - \hat{p}_c) &= Var(\hat{p}_t) + Var(\hat{p}_c) \\ \widehat{Var(\hat{p}_t)} &= \frac{\hat{p}(1-\hat{p})}{n_t} \\ \widehat{Var(\hat{p}_c)} &= \frac{\hat{p}(1-\hat{p})}{n_c} \\ \rightarrow se(\hat{p}_t - \hat{p}_c) &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n_t} + \frac{\hat{p}(1-\hat{p})}{n_c}} \end{aligned}$$

Now we're ready to calculate our t-statistic:

$$\begin{aligned} \hat{D} &= \hat{p}_t - \hat{p}_c = -.0206 \\ \hat{p} &= \frac{14846}{24059}(.1447) + \frac{9213}{24059}(.1653) = .1526 \\ se(\hat{D}) =& \sqrt{\frac{.1526(1-.1526)}{14846} + \frac{.1526(1-.1526)}{9213}} = .00477 \\ \Rightarrow z &= \frac{-.0206 - 0}{.00477} = -4.32 \end{aligned}$$

Step 3. By the null hypotheses we chose, we're doing a two-sided test. Let's choose the 5% significance level as this is the most common test that economists evaluate. Here our null hypothesis does not specify a value for p, and thus does not specify a value for the variance. Since we are using an estimator of the variance, we use the t-distribution. Check the t-distribution table to find that $c = 1.96$

Step 4. Reject: —-4.32—¿1.96

Step 5. Interpret: At the 5% significance level, there is statistical evidence that the proportion of households with a refrigerator in the control group is not the same as the proportion of households with a refrigerator in the treatment group.

What does this mean for the study? *Probably not much. In randomized experiments such as this one, many household characteristics are checked for "balance" across treatment and control. Statistically, we* *expect that some of our hypothesis tests will reject the null simply because a 5% significance level indicates that 5% of the time we will reject the null even though it's true. So if we checked for differences on 100 variables, we would mechanically expect to reject the null that the means are equal with 95% confidence for 5 of them.*

4. OLS Estimators (Lecture 10)

We now transition back to the population model we discussed previously:

$$y = \beta_0 + \beta_1 x + \mu$$

Using the language from previous sections, $\hat{\beta}_0$ and $\hat{\beta}_1$ are **estimators** for the parameters β_0 and β_1. Then, for the given sample of data we work with, we obtain particular **estimates** $\hat{\beta}_0$ and $\hat{\beta}_1$. Using $\hat{\beta}_0$ and $\hat{\beta}_1$ we can calculate the fitted values $\hat{y}_i$ for each observation from the equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Recall $\hat{y}_i$ is the fitted regression line. It can be thought of as our best guess for y given a certain value of x.

Next, recall that the linear regression model includes of a set of **assumptions** about how a data set will be produced by an underlying "data-generating process" :

1. Linear in Parameters
2. Random Sampling
3. Sample Variation in the Explanatory Variable, and No Perfect Multicollinarity in a multivariate regression
4. Zero Conditional Mean Errors
5. Homoskedasticity

We use these assumptions to obtain the expected value and the variance of the OLS estimators: [2]

$$\begin{aligned} E[\hat{\beta}_j] &= \beta_j & \forall j = 0, 1, \cdots, k \\ Var[\hat{\beta}_j] &= \frac{\sigma_u^2}{SST_j(1 - R_j^2)} & \forall j = 0, 1, \cdots, k \end{aligned}$$

In order to perform statistical inference we need to know the full sampling distribution of the $\hat{\beta}_j$. This requires a new assumption that we present here:

MLR 6: The population error u is normally distributed with mean 0 and variance σ^2: $u \sim N(0, \sigma^2)$ Wooldridge provides some intuition for why we assume the errors are normally distributed "because u is the sum of many different unobserved factors affecting y, we can invoke the Central Limit Theorem (CLT) to conclude that u has an approximate normal distribution". Note that this is a very strong assumption, since it automatically assumes MLR4 and MLR5.

Next, normality of the error term implies the OLS estimators are also normally distributed.

[2] See proofs p.114-115 in Woolridge

Theorem 1 *Under assumptions 1-6, conditional on the sample values of the independent variables,*

$$\hat{\beta}_j \sim Normal[\beta_j, Var(\hat{\beta}_j)]$$

Therefore

$$(\hat{\beta}_j - \beta_j)/sd(\hat{\beta}_j) \sim Normal(0,1)$$

A corrolary is that $(\hat{\beta}_j - \beta_j)/se(\hat{\beta}_j) \sim t_{n-k-1}$, where k is the number of parameters estimated in the regression.

In class we showed that $\hat{\beta}_j$ could be expressed as a linear combination of the true population parameters and the sum of the error terms u_i. Therefore we assume that if the u_i's are normally distributed, then the $\hat{\beta}_j$ is normally distributed as well.

5. Hypothesis tests β: Example

i. Statistical inference

We will hypothesize certain values for β_j and then use statistical inference to test our hypothesis. We can apply the same formula and steps that we used for sample means to find confidence intervals and test hypotheses for regression parameters. Data from the Indian DHS survey from 2006 that include a measure of autonomy of the women surveyed (a scale from 0-10, 10 being the most autonomous) that's based on decision-making in the household and domestic violence, the age when married, current age, a dummy for husband's education greater than primary school, and a dummy for an urban location. We can estimate the following model:

$$autonomy = \beta_0 + \beta_1 mrage + \beta_2 crage + \beta_3 husbedu + \beta_4 urban + u$$

Here are the results (Note that this is not R output, but looks somewhat similar):

```
. reg autonomy marr_age curr_age husb_educ urban

      Source |       SS       df       MS              Number of obs =     976
-------------+------------------------------           F(  4,   971) =   14.58
       Model |  949.690268     4  237.422567           Prob > F      =  0.0000
    Residual |  15809.3097   971  16.2814724           R-squared     =  0.0567
-------------+------------------------------           Adj R-squared =  0.0528
       Total |       16759   975  17.1887179           Root MSE      =   4.035

------------------------------------------------------------------------------
    autonomy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    marr_age |   .0329975   .0400423     0.82   0.410
    curr_age |   .1129321   .0265058     4.26   0.000
   husb_educ |   .5356327   .2834984     1.89   0.059
       urban |   1.192164   .2735485     4.36   0.000
       _cons |   .3844601   .7484315     0.51   0.608
------------------------------------------------------------------------------
```

Now, let's test the hypothesis that age at marriage has no effect on female autonomy

Step 1: Define hypotheses:

$$H_0 : \beta_{marrage} = 0$$
$$H_1 : \beta_{marrage} \neq 0$$

Under the null hypothesis, the true population parameter $\beta_{marrage} = 0$, which is saying that marriage age doesn't have any effect on female autonomy when we control for current age, husband education and urban location. When this null hypothesis holds, we know that

$$t = \frac{\hat{\beta}_{marrage} - \beta_{marrage}}{SE\left(\hat{\beta}_{marrage}\right)} \sim t_{n-k-1}$$

Step 2: Compute test statistic:

Stata gives us into the formula above for the t-stat.

$$t = \frac{\hat{\beta}_{marrage} - \beta_{marrage}}{SE\left(\hat{\beta}_{marrage}\right)} = \frac{.033 - 0}{.040} = .825$$

(Notice how this is the same as the number in the t column of the Stata output for age at marriage.)

Step 3: Choose significance level (α) and find the critical value: To shake it up a little, let's choose 10%. We need three things to find c:

1. Significance level : here $\alpha = .10$
2. Two sided/One sided test: here we chose a two sided
3. Degrees of freedom $= n - k - 1 = 976 - 4 - 1 = 971 \approx 1000$

Check the t-table to see that $c_{.10} = 1.645$

Step 4: Reject the null hypothesis or fail to reject it:

Here

$$|t| = 0.825 < |c| = 1.645$$

We fail to reject the null hypothesis.

Step 5: Interpret[3]

At the 10% significance level, we fail to reject the null hypothesis that the age of marriage has no impact on expected female autonomy controlling for current age, husband education, and urban location. So we don't find statistical evidence that suggests delaying marriage will result in more female autonomy in India.

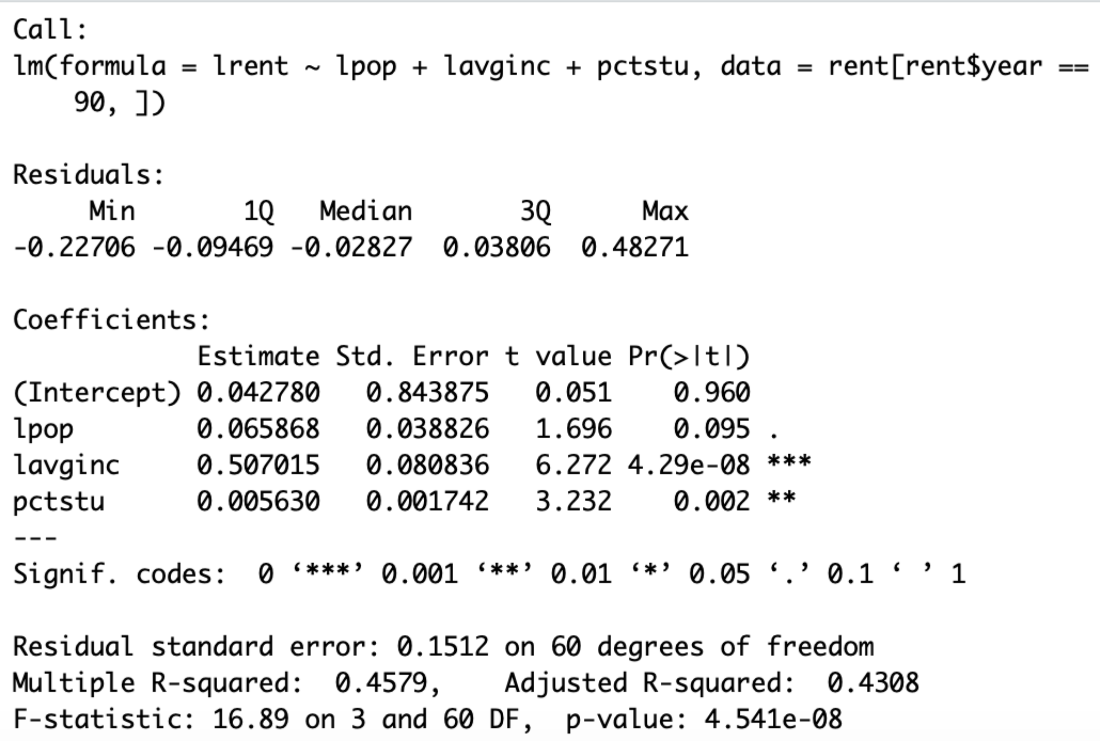
[3] As discussed in class, when we fail to reject we have a hard time saying anything. Indeed we made 6 assumptions to start. Finding evidence against the null is one way to interpret our result; but maybe we mispecified the model, maybe $E[u|x] = 0$ fails here, maybe we didn't have a random sample, etc.

6. Hypothesis tests: Example 4.4 p.160 Woolridge

Are rent rates influenced by the student population in a college town? Let rent be the average monthly rent paid on rental units in a college town in the United States. Let pop denote the total city population, avginc the average city income, and pctstu the student population as a percentage of the total population. One model to test for a relationship is

$$log(rent) = \beta_0 + \beta_1 log(pop) + \beta_2 log(avginc) + \beta_3 pctstu + u$$

1. State the null hypothesis that size of the student body relative to the population has no ceteris paribus effect on monthly rents. State the alternative that there is an effect.
2. What signs do you expect for β_1, β_2.
3. We estimate the population regression equation using 1990 data from RENTAL.DTA for 64 college towns. The R output is:



```
Call:
lm(formula = lrent ~ lpop + lavginc + pctstu, data = rent[rent$year ==
    90, ])

Residuals:
     Min       1Q   Median       3Q      Max
-0.22706 -0.09469 -0.02827  0.03806  0.48271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.042780   0.843875   0.051    0.960
lpop        0.065868   0.038826   1.696    0.095 .
lavginc     0.507015   0.080836   6.272 4.29e-08 ***
pctstu      0.005630   0.001742   3.232    0.002 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1512 on 60 degrees of freedom
Multiple R-squared:  0.4579,    Adjusted R-squared:  0.4308
F-statistic: 16.89 on 3 and 60 DF,  p-value: 4.541e-08
```

   (a) Write out the equation estimated by R: replace the estimated coefficients by their numerical value from R, and below the variable, write the corresponding standard error

   (b) What is the R^2 ? Adjusted R^2?

   (c) What is the t-stat and p-value associated with lnpop? What is the null that these are calculated off of?

4. What is wrong with the statement: "A 10% increase in population is associated with about a 6.5% increase in log rent"?
5. Test the hypothesis stated in part 1) at the 1% level

Answers:

1. $H0 : \beta_3 = 0$, $H1 : \beta_3 \neq 0$.

2. Other things equal, a larger population increases the demand for rental housing, which should increase rents. The demand for overall housing is higher when average income is higher, pushing up the cost of housing, including rental rates.

3. We estimate the equation

    (a) We write the equation as:

    $$\widehat{logrent} = \underset{(0.8439)}{0.0428} + \underset{(0.03883)}{0.0659} log(pop) + \underset{(0.0808)}{0.507} log(avginc) + \underset{(.00174)}{0.0056} pctstu$$

    (b) $R^2 = 0.4579$, Adjusted $R^2 = 0.4308$

    (c) The t-stat is 1.696, and the p-value is 0.095 - these are based on the null hypothesis that $\beta_1 = 0$

4. The coefficient on log(pop) is an elasticity. A correct statement is that "a 10% increase in population increases rent by .065(10) = .65%."

5. 5 steps of hypothesis testing:

    (a) $H0 : \beta_3 = 0$, $H1 : \beta_3 \neq 0$.

    (b) The t-statistic is about 3.23 (can get this both from the table and from calculating it using the formula):

    $$t = \frac{0.005630 - 0}{0.001742} = 3.23$$

    (c) With df = 64 - 3 -1 = 60, the 1% critical value for a two-tailed test is 2.660.

    (d) T-statistic is well above the critical value $\Rightarrow$ reject the null

    (e) At the 1% significance level, we can reject the null that the size of the student body relative to the population has no effect on expected monthly rents holding population and average income in the city constant.

7. Hypothesis tests: Example 4.4 p.131 Woolridge

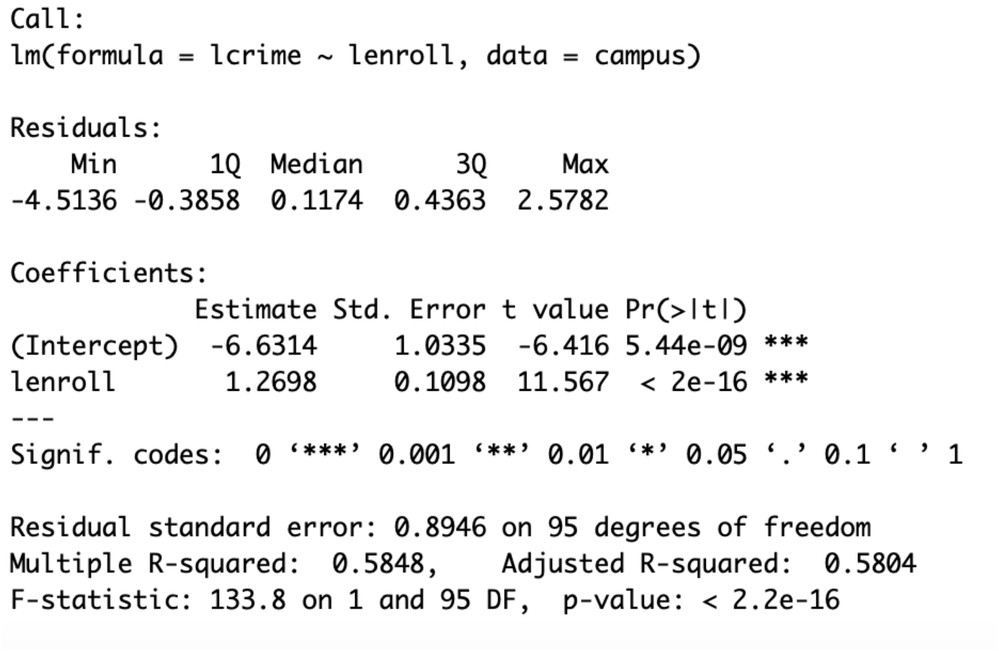
Consider a simple model relating the annual number of crimes on college campuses (crime) to student enrollment (enroll):

$$log(crime) = \beta_0 + \beta_1 log(enroll) + u$$

Here β_1 is the elelasticity of crime with respect to enrollment. An interesting hypothesis to test would be that the elasticity of crime with respect to enrollment is one: $H_0 : \beta_1 = 1$. This means that a 1% increase in enrollment leads to, on average, a 1% increase in crime. A noteworthy alternative is $H_1 : \beta_1 > 1$ which implies that a 1% increase in enrollment increases campus crime by more than 1%. If $\beta_1 > 1$ then, in a relative sense-not just an absolute sense-crime is more of a problem on larger campuses.

1. State the null hypothesis and the alternative.

2. We estimate the population regression equation using 1992 data from CAMPUS.RAW for 97 college campuses. The R output is:



```
Call:
lm(formula = lcrime ~ lenroll, data = campus)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.5136 -0.3858  0.1174  0.4363  2.5782 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -6.6314     1.0335  -6.416 5.44e-09 ***
lenroll       1.2698     0.1098  11.567  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8946 on 95 degrees of freedom
Multiple R-squared:  0.5848,	Adjusted R-squared:  0.5804 
F-statistic: 133.8 on 1 and 95 DF,  p-value: < 2.2e-16
```

   (a) Write out the equation estimated by R: replace the estimated coefficients by their numerical value from stata, and below the variable, write the corresponding standard error

   (b) What is the R^2?

   (c) What is the t-stat and p-value associated with lenroll? What is the null that these are calculated off of?

3. Is there anything wrong with the statement: “A 1% increase in population is associated with about a 1.26% increase in crime”?

4. Test the hypothesis stated in part 1) at the 5% level

Answers:

1. $H0 : \beta_1 = 1$, $H1 : \beta_1 > 1$
2. We estimate the equations
   (a) We write the equation as:
   $$\underset{}{\widehat{logcrime}} = \underset{(1.03354)}{-6.63} + \underset{(0.10977)}{1.27 log(enroll)}$$
   (b) $R^2 = 0.5848$
   (c) The t-stat is 11.57, and the p-value is approx 0.000 - these are based on the null hypothesis that $\beta_1 = 0$
3. Nothing is wrong
4. 5 steps of hypothesis testing:
   (a) $H0 := \beta_1 = 1$, $H1 : \beta_1 > 1$.
   (b) The t-statistic is about 4.27 :
   $$t = \frac{1.26976 - 1}{0.109776} = 2.45736$$
   (c) With df = 97 -1- 1 = 95, the 5% critical value for a one-tailed test is $c = 1.66$.
   (d) We clearly reject $\beta_1 = 1$ in favor of $\beta_1 > 1$ at the 5% level. In fact, the 1% critical value is about 2.37, and so we reject the null in favor of the alternative at even the 1% level.
   (e) So is statistically different from 1 at the 1% level.