EC 421 Classwork 4: Hypothesis Testing (analytical)

In chapter 2, we determined that (given some assumptions): $\hat{\beta}_1$ is distributed $N(\beta_1, \frac{\text{Var}(u)}{\sum_i (x_i - \bar{x})^2})$. Var($u$) is unknown since u_i is unobservable, so we have to approximate it using the regression residuals e_i. Because of this, we also use the t-distribution, which is similar to the Normal distribution, but with slightly fatter tails.

We define "standard error" as our estimate of the standard deviation of the regression coefficient. The formula for a simple regression standard error of $\hat{\beta}_1$ is $\sqrt{\frac{\sum_i (e_i^2)}{(n-2)\sum_i (x_i - \bar{x})^2}}$.

1) We'd like our standard errors to be as small as possible so we can increase the precision of our estimates. If we increased the number of observations (assuming all else is held equal), will our standard errors increase or decrease?

2) All else held equal, if the sample variance of x_i decreased, should we expect that the standard errors would increase or decrease?

For the sake of building some intuition, an example of this:

Take 2 studies designed to find the effect of a blood pressure medication on health.

Study A: 9 people take the placebo; 1 person takes the medication, so X = (0, 0, 0, 0, 0, 0, 0, 0, 0, 1).

Study B: 5 people take the placebo; 5 people take the medication so X = (0, 0, 0, 0, 0, 1, 1, 1, 1, 1).

Calculate the sample variance of X in both of these studies (sample variance = $\frac{\sum_i (x_i - \bar{x})^2}{n-1}$. Which study will yield a more confident estimate of the effect, and which study has a lower Var(X)?

3) Finally, if we decreased the variance of u_i by including more explanatory variables (if we used the second model instead of the first model below), should we expect that the standard error on β_1 will increase or decrease?

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

$$wage_i = \beta_0 + \beta_1 education_i + \beta_2 sex_i + \beta_3 race_i + \beta_4 family\ wealth_i + u_i$$

Hypothesis Testing

Suppose we fit the model $y_i = \beta_0 + \beta_1 x_i + u_i$ by running this:

```
# lm(y ~ x, data) %>%
#   broom::tidy(conf.int = TRUE)
```

And suppose we got these results:

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.41	0.924	1.53	0.145	-0.53	3.35
x	0.930	0.302	3.08	0.00646	0.296	1.56

In this question, we'll practice doing hypothesis tests on OLS estimates. First, I'll do the hypothesis test on the estimate for the intercept, $\hat{\beta}_0$, and then you'll do the hypothesis test on the estimate for the slope, $\hat{\beta}_1$.

Hypothesis Test for $\hat{\beta}_0$ Our estimate for β_0 is 1.41, and our standard error is 0.924.

Formally, we'd like to test the null hypothesis: H_0: $\beta_0 = 0$ Against the alternative: H_a: $\beta_0 \neq 0$

Here are 3 equivalent ways to do hypothesis testing:

1) Compare the test statistic to the critical values. If the test statistic is more extreme than the critical values, you have evidence in favor of rejecting the null.
2) Calculate a confidence interval around the estimate. If 0 lies inside of that confidence interval, you'll fail to reject the null.
3) Calculate the p-value for the estimate. If the p-value is less than .05, you have evidence in favor of rejecting the null at the 5% level.

Test Statistic and Critical Value The *test statistic* is the number of standard deviations from 0 the estimate falls. It's calculated by:

abs(estimate/standard error)

```
abs(1.41/0.924)
```

```
## [1] 1.525974
```

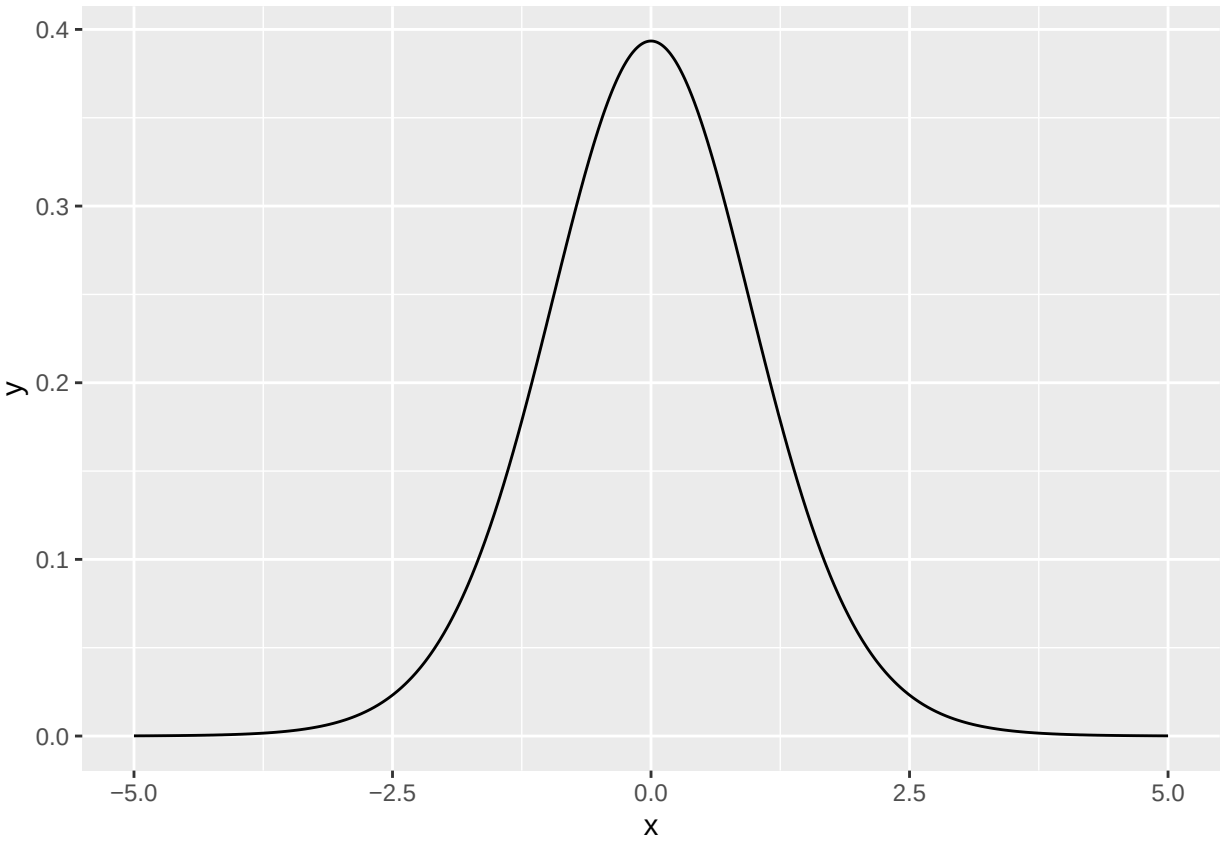
The *critical value* for a 2-tailed 95% test is $t_{.05/2,df=n-k-1}$. Suppose n is 20 and k is 1 (k is the number of explanatory variables, which is 1 here). Then we'd use $t_{.05/2,df=18} = 2.100922$. You could look in the back of a statistics textbook to find that number, but in this class, I want you to find it using R with the `dt()` family. Read the help docs to find out how:

```
?dt
```

`dt(x, df)` gives you the density (the pdf) of the t distribution. So one way to plot the t distribution would be to use `dt()` like this:

```
library(tidyverse)

tibble(
  x = seq(-5, 5, by = .01),
  y = dt(x, df = 18)
) %>%
  qplot(data = ., x = x, y = y, geom = "line")
```



`pt(q, df)` takes a `q` quantile and gives you the distribution function. For example, since the t distribution is centered on 0, `pt(0)` will return `.5` because half of the distribution is to the left of 0.

```
pt(0, df = 18)
```

```
## [1] 0.5
```

`qt(p, df)` does the inverse of `pt()`: it takes a `p` probability and gives you the quantile function. For example, `qt(.5)` returns 0 because half of the distribution is to the left of 0.

```
qt(.5, df = 18)
```

```
## [1] 0
```

Finally, `rt(n, df)` generates `n` random numbers from the t distribution. Here I generate 5 random numbers from the t distribution (with 18 degrees of freedom):

```
rt(5, df = 18)
```

```
## [1]  0.7482643  1.8298626 -1.1976996  0.7001277 -0.8331140
```

The interpretation of the critical value of 2.1 is that 95% of the density of the t-distribution with 18 degrees of freedom falls between -2.1 and 2.1.

Next, we'll compare the test statistic to the critical values: 1.53 is not outside of the interval -2.100922 and 2.100922, so we *fail to reject* the null, and β_0 may very well be equal to 0.

Confidence Interval The 95% *confidence interval* is +/- about 2 standard deviations from the estimate: 95/100 times we sample the population and estimate β_0, this confidence interval will hold the true value of β_0 (as long as all our other OLS assumptions hold, of course).

[estimate - (critical value x standard error); estimate + (critical value x standard error)]

1.41 - 2.100922*0.924 = -0.53

1.41 + 2.100922*0.924 = 3.35

Since the 95% confidence interval overlaps 0, that tells you (again) that you can't reject the null that $\beta_0 = 0$ at the 5% level.

P-value *P-value*: given sampling uncertainty, this is the probability of getting a test statistic as extreme as this under the null.

Again we'll use the `dt()` family: in particular, we can use `pt()`. We multiply by 2 to get the probability of getting a test statistic greater than 1.53 or less than -1.53. The chance of getting that extreme of a test statistic is 14.3%. If the p-value was less than 10%, we could reject the null at the 10% level. If the p-value was less than 5%, we could reject the null at the 5% level.

```
2 * (1 - pt(1.53, df = 18))
```

```
## [1] 0.1434013
```

In future projects, I'll often say something like: "Estimate this model and perform appropriate hypothesis tests." That just means run `lm()` and report the hypothesis test results for each parameter. In this case, you could simply say:

I fail to reject the null that $\beta_0 = 0$ at the 10% level.

4) Your turn! Perform a formal hypothesis test on $\hat{\beta}_1$ assuming n = 20, using the OLS results from the table above. Make sure to state the null and alternative hypotheses, show how the test statistic was calculated, compare the test statistic to the critical values, show how the confidence interval was calculated, and interpret the p-value.