WHAT IS REGRESSION?

WHAT DO WE USE REGRESSION FOR?

WHAT ABOUT THOSE PARAMETERS?

WHAT IS A NULL HYPOTHESIS?

HOW DO I ASSESS SIGNIFICANCE?

Regression flash slides

What is regression?

- Regression is a data-driven method that we use to estimate relationships between variables
- Simple regression estimates relationships of the form: $y = \alpha + \beta x + \varepsilon$
- Multivariate regression estimates relationships of the form:

    $y = \alpha + \beta x + \gamma z + \varepsilon$

- We use it to explain things and sometimes make predictions***
- To make certain decisions, it is nice to know how variables are related
- Be careful about how you interpret regression coefficients!

What about those coefficients?

- Regression estimates the parameters of your model (β), yielding coefficients($\hat{\beta}$)
- Those coefficients are calculated from the data
- Be careful about how you interpret them!
  - They are estimations so they have error
  - They tell us nothing about causality

What are hypotheses?

- Because our coefficients are measured with error, we want to measure how sure we are about their values

- In other words, we could have gotten this specific number by chance, so we want confidence intervals around them

- It is very common to test the hypothesis of a coefficient being zero

- To tests hypotheses, we take a conservative initial position: null hypothesis

- Then we go to our data to see if there is evidence that contradicts the null hypothesis

The null hypothesis:
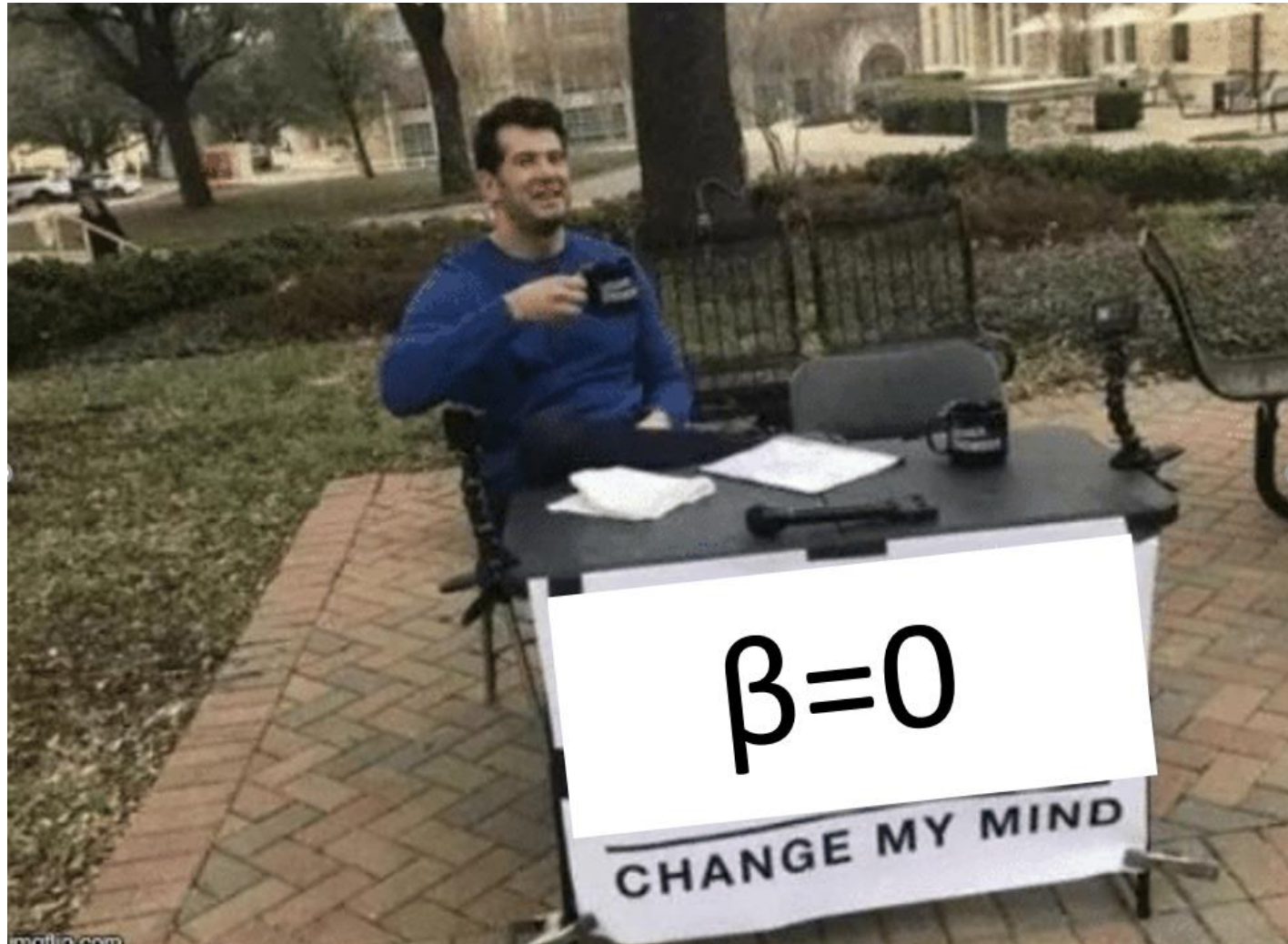
There are several ways we can test the null hypothesis that a coefficient is zero

- T stat: compare your calculated t-statistic to the corresponding critical value (on tables). This value will depend on the degrees of freedom and the confidence level.
  - The bigger the t-stat, the more evidence that contradicts the null
- Confidence intervals: is the null hypothesis value contained in the interval? If yes: no evidence that contradicts the null.
- P value: the probability that you got this coefficient by chance. If the p-value of your coefficient is lower than 1-confidence level, then this is evidence that contradicts the null.

The following are equivalent (for coefficients)

"The coefficient is significant at the 95% confidence level"

"The coefficient is statistically different from zero at the 95% confidence level"

"I reject the null hypothesis at a 95% confidence level"

What about the F statistic?

- The F statistic is a measure of joint significance

- It tells you how good your model is at explaining Y, beyond what you would get from just taking the average of Y
  - **Null hypothesis**: The fit of the intercept-only model and your model are equal

- It has its own tables, but its distribution depends on two degrees of freedom: regression, and residual
  - Regression: k-1 (numerator)
  - Residual: N-k (denominator)

- Compare it to the corresponding critical value
  - **if your calculated F value in a test is larger than your F statistic, you can reject the null hypothesis**.

- The larger the F-stat of your regression, the more likely you did not get the results by chance

- Excel also calculates the significance of the F test, which is kind of a p-value for the F statistic