EEP/IAS 118 - Introductory Applied Econometrics, Lecture 7

Pierre Biscaye and Jed Silver

October 2021

Today's section

- F-test recap
- Scaling & standardized effects
- Functional form and adjusted R^2
- Interactions

F-Test

Suppose we want to test whether two (or more!) parameters are *jointly* different from zero. To do this we use an F-test. Why not just test the two parameters separately?

- You might have a group of variables that are highly correlated. High correlations (multicollinearity) among related variables will increase the standard errors, making it less likely any one of them is individually statistically significant. But this does not mean the set of variables is not helping to explain part of the variation in y!
- You might have a quadratic term - testing whether there is any effect requires an F test with both terms including your variable of interest.
- You might have an interaction term - certain tests will require an F test with both interacted and uninteracted forms of your variable.

F-Test Steps

See Section 6 notes for more detail. Suppose we want to test

whether working in/near a large city in general ($lgcity,\ sublg$) matters for wage.

1. Define hypotheses:

$$H_0 : \beta_{lgcity} = 0 \quad \& \quad \beta_{sublg} = 0$$

$$H_1 : \beta_{lgcity} \neq 0 \ or \ \beta_{sublg} \neq 0 \ or \ both$$

F-Test Steps

2. Write down the two models the hypotheses imply:

   Unrestricted model (UR):

   $$lwage = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exp + \beta_4 sublg + \beta_5 lgcity + u$$

   Restricted model (R):

   $$lwage = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exp + u$$

- We call the regression with the variables we are testing the **unrestricted model**.
- The regression without these variables (whose coefficients are 0 under the null) is the **restricted model**
- Estimate both these models separately

F-Test Steps

3. Write our F-stat from the two regression outputs:

$$F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(n - k_{UR} - 1)} = \frac{\left(R^2_{UR} - R^2_R\right)/q}{(1 - R^2_{UR})/(n - k_{UR} - 1)}$$

Where

- q is the number of restrictions (in this case, 2)
- k_{UR} is the number of variables in the unrestricted model

4. Compare the F-stat to the correct critical value c found in the F-table for a particular significance level. You will need to keep track of **both** numerator degrees of freedom (q) and denominator degrees of freedom $(n - k_{UR} - 1)$
5. Reject null if F$>$c

Scaling Variables

Often times the units that variables come in are not the most useful for interpretation or analysis.

- Rescaling monetary units - $ thousands, $ billions, etc.
- Distance per second into distance per hour

Example:

$$\widehat{sleep} = 3315.574 - 12.189educ + 2.7454age$$

Where sleep is measured in minutes per week. How is $\hat{\beta}_{educ}$ interpreted here?

Scaling Variables

Often times the units that variables come in are not the most useful for interpretation or analysis.

- Rescaling monetary units - $ thousands, $ billions, etc.
- Distance per second into distance per hour

Example:

$$\widehat{sleep} = 3315.574 - 12.189educ + 2.7454age$$

Where sleep is measured in minutes per week. How is $\hat{\beta}_{educ}$ interpreted here?

- One more year of education is estimated to decrease predicted sleep by 12.189 minutes per week, holding age constant

Scaling Variables

$$\widehat{sleep} = 3315.574 - 12.189educ + 2.7454age$$

Lets say we instead want to change the dependent variable to be measured in hours rather than minutes.

- Do this simply by changing our y variable into $\tilde{y} = \frac{y}{60}$

How would this change our $\hat{\beta}$?

Scaling Variables

$$\widehat{sleep} = 3315.574 - 12.189educ + 2.7454age$$

Lets say we instead want to change the dependent variable to be measured in hours rather than minutes.

- Do this simply by changing our y variable into $\tilde{y} = \frac{y}{60}$

How would this change our $\hat{\beta}$?

- The new β_{educ} estimate would be $\frac{12.189}{60} = 0.2$ hours per night

The entire regression result changes to this:

$$\widehat{sleep} = 55.260 - .2032educ + .0458age$$

Scaling Variables

In general, when we re-scale the outcome variable by α

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + ... + \tilde{\beta}_k x_k + u$$
$$\alpha y = \alpha\beta_0 + \alpha\beta_1 x_1 + ... + \alpha\beta_k x_k + u$$

In the above example, $\alpha = \frac{1}{60}$, so the new $\hat{\beta}$s will be divided by 60 too.

- **Note:** nothing else about the regression will change (R^2, t-stats, p-values, etc.)

Scaling Variables

Let's say instead we rescale an independent x variable:

- Rescale education to be in units of half-years (6 months) - i.e. we multiply $educ$ by 2
- The new regression would give us:

$$\widehat{sleep} = 3315.574 - 6.095educ + 2.7454age$$

- Only the coefficient on the independent variable we modified has changed

Scaling Variables

In general, if we scale x by α, the equation becomes:

$$\begin{aligned} y =& \beta_0 + \tilde{\beta}_1 \tilde{x}_1 + ... + \beta_k x_k + u \\ =& \beta_0 + \frac{\beta_1}{\alpha}(\alpha x_1) + ... + \beta_k x_k + u \end{aligned}$$

- In the above example, we had $\alpha = 2$, which meant we had to scale our estimate of $\hat{\beta}_{educ}$ by $\frac{1}{2}$.

Standardizing Variables

Up until now we've been considering cases where we want to change the units of a variable into units that are more useful

- But what if we don't want units at all? Why would we want this?
- Want to compare the relative effects of two variables that don't have the same unit - e.g. education and SAT score on income
- Can be unclear how big of a change a unit increase in education is relative to a unit increase in SAT score
- But we might be interested in in comparing sizes of effects in a variety of contexts, such as determining which of two inputs to focus on to increase some output

Standardize Variables

Standardizing means we will compare how a one standard deviation increase in x_1 affects y to how a one s.d. increase in x_2 affects y

We do this by transforming all our variables by subtracting their mean and dividing by the standard deviation:

$$\tilde{y} = \left(\frac{y - \bar{y}}{\hat{\sigma}_y} \right)$$

$$\tilde{x} = \left(\frac{x_1 - \bar{x}_1}{\hat{\sigma}_{x_1}} \right)$$

This should look familiar - this is what we do with our t-stats! Idea is that we put all the variables on the same scale. Then we can compare relative effects.

Standardize Variables

Once we standardize all the units, re-running the regression produces:

$$\left(\frac{y-\bar{y}}{\hat{\sigma}_y}\right) = \frac{\hat{\sigma}_{x_1}}{\hat{\sigma}_y}\hat{\beta}_1\left(\frac{x_1-\bar{x}}{\hat{\sigma}_{x_1}}\right) + \frac{\hat{\sigma}_{x_2}}{\sigma_y}\hat{\beta}_2\left(\frac{x_2-\bar{x}_2}{\hat{\sigma}_{x_2}}\right)$$

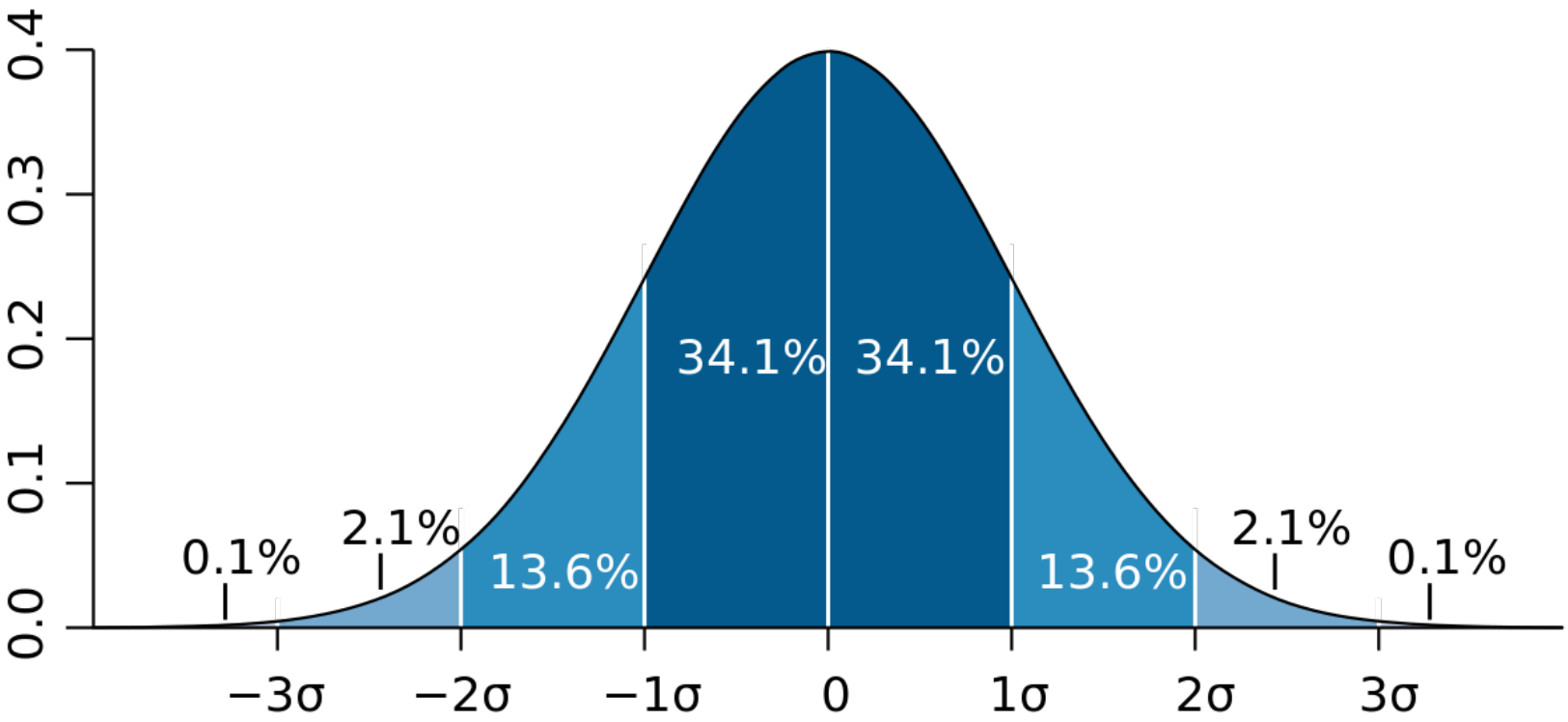
- The new parameters will be equal to the old parameters scaled by $\frac{\hat{\sigma}_{x_1}}{\hat{\sigma}_y}$
- This is called the "standardized coefficient" or the "beta coefficient"
- In R we can produce these coefficients if we standardize each variable in our regression with the "scale()" function.

Standardize Variables

```
Call:
lm(formula = scale(sleep) ~ scale(educ) + scale(age), data = sleep75)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5746 -0.5290  0.0047  0.5869  3.1118

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.909e-16  3.743e-02   0.000   1.0000
scale(educ) -7.638e-02  3.886e-02  -1.966   0.0497 *
scale(age)   7.007e-02  3.886e-02   1.803   0.0718 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9946 on 703 degrees of freedom
Multiple R-squared:  0.01359,   Adjusted R-squared:  0.01078
F-statistic: 4.842 on 2 and 703 DF,  p-value: 0.008155
```

- One SD increase in education leads to 0.076 SD decrease in sleep time, ceteris paribus
- One SD increase in age leads to a 0.07 SD increase in sleep time, ceteris paribus

What are Standard Deviations?



- 68.2% of data are within 1 SD
- 95.4% are within 2 SD
- 99.7% within 3 SD

Functional Form and Adjusted-R^2

We've gone over how to do the following:

1. Deciding if one of your x variables is significant $\Rightarrow$ t-test
2. Deciding if multiple variables *together* are significant $\Rightarrow$ F-test.

These tests compare *nested* models

- Nested models are cases where one equation is just a special case of the other (e.g. fixing β_3 and $\beta_4 = 0$)

How do we compare and choose between *non-nested* models?

- Use **Adjusted** R^2

R^2 vs. Adjusted R^2

Regular R^2 is a measure of "goodness of fit", so why not just use that?

- $R^2 = 1 - \frac{SSR}{SST}$ will always (weakly) increase when you add more variables to the regression
- Not useful to choosing which model is better, more complex one will always win

Therefore, we use Adjusted R^2 which adds a penalty for each additional variable added to the model

Adjusted R^2

The formula for adjusted R^2 is:

$$1 - \frac{SSR/(n-k-1)}{SST/(n-1)}$$

Adding variables now has two effects:

1. The SSR in the numerator will always (weakly) decrease with an additional variable
2. However, k will also increase (making the numerator larger)

Therefore, the effect on the adjusted R^2 from adding an additional variable to the regression will depend on if the extra explanatory power is larger than the penalty

Adj R^2: Comparing Non-nested Models

When does Adj R^2 come in handy:

- Choosing a functional form for the right hand side variables can be difficult
- A common example is a choice between $log(x)$ and a quadratic x and x^2
- Both can be reasonable choices and it is difficult to eyeball which is better
- The choice matters a lot for interpretations of β estimates (refer to L13 slides)
- Nothing stopping you from running both: the models will find the best fit for the data given the functional form

Adj R^2: Two Models of Sleep

```
Call:
lm(formula = sleep ~ log(age), data = sleep75)

Residuals:
     Min       1Q   Median       3Q      Max
-2452.18  -258.46    11.21   269.80  1387.55

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2821.78     209.42   13.47   <2e-16 ***
log(age)      122.92      57.72    2.13   0.0335 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 443.3 on 704 degrees of freedom
Multiple R-squared:  0.006401,  Adjusted R-squared:  0.00499
F-statistic: 4.535 on 1 and 704 DF,  p-value: 0.03355
```

```
Call:
lm(formula = sleep ~ age + agesq, data = sleep75)

Residuals:
    Min      1Q  Median      3Q     Max
-2518.1  -250.4     2.6   276.8  1390.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3608.0297   230.6457  15.643   <2e-16 ***
age          -21.4904    11.7367  -1.831   0.0675 .
agesq          0.3012     0.1401   2.150   0.0319 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 441.8 on 703 degrees of freedom
Multiple R-squared:  0.01464,   Adjusted R-squared:  0.01184
F-statistic: 5.224 on 2 and 703 DF,  p-value: 0.005598
```

Which do we prefer? Look at $Adj - R^2$ as a basic test if don't have other theoretical reasons to prefer one or the other.

Comparing Functional Forms

- Any theoretical reason to believe a particular functional form is correct?
- Which form fits the data best? Look at $Adj - R^2$
- What does the density of your X variable look like?
    - Logs place higher weight on lower values, so useful for X variables with long right tails; underfined for X$\leq$0
- Lots of functional forms you can consider, but rare to go beyond linear, logs, or quadratic (sometimes higher-order polynomial), or interaction terms

Dummy Variables

Dummy variables: these are binary variables (or zero-one variables). For example, urban/rural or female/male

How do we interpret dummies?

$$wage = \beta_0 + \beta_1 female + \beta_2 educ + u$$

- Difference between female and male wages is β_1.
- We can also think of this dummy as introducing an intercept shift between males and females:
  - The intercept for males is β_0
  - The intercept for female is $\beta_0 + \beta_1$

Interactions: Two Continuous Variables

What if we think the effect of some variable on your outcome varies depending on the level of another variable? Let's consider this model:

$$wage = \beta_0 + \beta_1 age + \beta_2 educ + \beta_3 age \times educ + u$$

What is the marginal effect of age? Re-group all the terms with age in them:

$$E[wage|educ, age] = \beta_0 + \underbrace{(\beta_1 + \beta_3 educ)}_{regrouped} age + \beta_2 educ$$

Therefore, the "age effect" i.e the marginal effect of age is $\beta_1 + \beta_3 educ$.

Interactions: Interpretation

$$\frac{\partial E[wage]}{\partial age} = \beta_1 + \beta_3 educ$$

- Marginal effect varys with $educ$. To get one value, we can plug in for $educ$ (usually with the median or mean)
- The marginal effect of an additional year of age on wages for people with 10 years of education is $\beta_1 + \beta_3 * 10$

We can follow the same intuition for education:

$$\frac{\partial E[wage]}{\partial educ} = \beta_2 + \beta_3 age$$

- The marginal effect of education on expected wage for people with 20 years of age is $\beta_2 + \beta_3 * 20$

Hypotheses with Interactions

$$wage = \beta_0 + \beta_1 female + \beta_2 educ + \beta_3 female \times educ + u$$

1. Write the null and alternative hypothesis to test that the return to education is the same for women and men
2. Write the null and alternative hypothesis to test that average wages are identical for men and women who have the same levels of education:

Hypotheses with Interactions

$$wage = \beta_0 + \beta_1 female + \beta_2 educ + \beta_3 female \times educ + u$$

1. Write the null and alternative hypothesis to test that the return to education is the same for women and men:

$$H_0 : \beta_3 = 0 \quad vs. \quad H_1 : \beta_3 \neq 0$$

2. Write the null and alternative hypothesis to test that average wages are identical for men and women who have the same levels of education:

$$H_0 : \beta_1 = 0 \ \ \& \ \ \beta_3 = 0 \quad vs. \quad H_1 : \beta_1 \neq 0 \ \ \&/or \ \ \beta_3 \neq 0$$

Q: What type of test is this?