EEP/IAS 118 - Introductory Applied Econometrics, Section 6

Pierre Biscaye and Jed Silver

October 2021

Agenda

- Review: $\hat{\beta}$ interpretation
- Review: OVB
- Model specification
- Hypothesis testing and p-values
- Linear combination tests
- F tests

Interpret $\hat{\beta}$

Many students understand the core concept, but consistently leave out parts in your description of $\hat{\beta}$. Suppose

$$\widehat{fdhome} = 1035.537 + 0.1025totexp + 0.302urban$$

where $fdhome$ is expenditure on food at home in USD, $totexp$ is total expenditure in USD, and $urban$ is a dummy for living in an urban setting. Interpret $\hat{\beta}_1$.

Interpret $\hat{\beta}$

$$\widehat{fdhome} = 1035.537 + 0.1025totexp + 0.302urban$$

"Our estimate of β_1 tells us that if total expenditures increase by 1 dollar, the model predicts that expenditure on food at home increases by 10 cents, holding all else constant"

- **Units!** The variables here are in dollars, so say this. It's easy to just say "one unit increase" but what does this mean?
  - Make sure you know when you're interpreting a dummy variable! Here, it doesn't make sense to say "a one unit increase in urban."
- Interpreting **logs**: think **percents!** Linear-log: divide coefficient by 100 for unit effect on Y of 1% change in X. Log-linear: multiply coefficient by 100 for % change in Y for 1 unit change in X. Log-log: coefficient is % change in Y for 1% change in X
- Holding all else constant (multivariate regression).

Interpret $\hat{\beta}$

We've also now introduced statistical significance of $\hat{\beta}$ estimates: can we say with some confidence that they are not equal to 0?

```
Call:
lm(formula = lwage ~ female + tenure, data = wage_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.00085 -0.28200 -0.06232  0.31200  1.57325

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.688842   0.034368  49.141  < 2e-16 ***
female      -0.342132   0.042267  -8.095 4.06e-15 ***
tenure       0.019265   0.002925   6.585 1.11e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4747 on 523 degrees of freedom
Multiple R-squared:  0.2055,    Adjusted R-squared:  0.2025
F-statistic: 67.64 on 2 and 523 DF,  p-value: < 2.2e-16
```

- t-value column in R output is t-statistic for null hypothesis that $\beta = 0$
- $Pr(> |t|)$ column is p-value, probability of observing a larger t-statistic under the null $\Rightarrow$ more discussion coming up!
- These give you information on whether you can reject the null that $\beta = 0$ at a particular confidence level. Higher t-stat $\Rightarrow$ smaller p-value $\Rightarrow$ higher confidence level!

Omitted Variables Bias - Review Section 4!

Take the same regression:

$$\widehat{fdhome} = 1035.537 + 0.1025totexp + 0.302urban$$

- What are some omitted variables that you think might be relevant here?
- When do we need to worry about SLR 4 being violated?

Omitted Variables Bias - Review Section 4!

$$\widehat{fdhome} = 1035.537 + 0.1025totexp + 0.302urban$$

Say you now include another relevant variable *restaurants*, the number of restaurants within 10 miles. You now get

$$\widehat{fdhome} = 1035.537 + 0.523totexp + 0.726urban - 20.43restaurants$$

How are restaurants and total expenditures correlated?

OVB Review

$$\widehat{fdhome} = 1035.537 + 0.1025totexp + 0.302urban$$

$$\widehat{fdhome} = 1035.537 + 0.523totexp + 0.726urban - 20.43restaurants$$

How are restaurants and total expenditures correlated?

$$E[\tilde{\beta}_1] = \beta_1 + \beta_3\rho$$

$$\tilde{\beta}_1 - \beta_1 = \beta_3\rho$$

- Here: $\tilde{\beta}_1 = 0.1025$, $\beta_1 = 0.523$, $\beta_3 = -20.43$.
- $\tilde{\beta}_1 - \beta_1 = 0.1025 - 0.523 = -0.4205 \implies$ **downward bias**.
- $\beta_3 = -20.43$
- Follows that $\rho = cov(totexp, restaurants) > 0$. In fact, can calculate $\rho = \frac{-0.4025}{-20.43} = 0.021$

Model Specification

Suppose MLR 2 and 3 hold but we are unsure what the specification of the "true" model is

- e.g. Model A: $y_i = \beta_0 + \beta_1 \log(x_{1i}) + \beta_2 \log(x_{2i}) + u_i$ vs Model B: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$
- Need to get this right to satisfy MLR 1.
- What if we run Model A and get highly significant coefficents β_1 and β_2 but not for Model B? Does this suggest Model A is the correct model?

Model Specification

Suppose MLR 2 and 3 hold but we are unsure what the specification of the "true" model is

- e.g. Model A: $y_i = \beta_0 + \beta_1 \log(x_{1i}) + \beta_2 \log(x_{2i}) + u_i$ vs Model B: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$
- Need to get this right to satisfy MLR 1.
- What if we run Model A and get highly significant coefficents β_1 and β_2 but not for Model B? Does this suggest Model A is the correct model?
  - Not really. If we believe the true model is in logs, then we have evidence that β_1 and β_2 are significant. But if we don't have the model correctly specificied, this is a meaningless correlation.
  - We need MLR1 for βs to be interpreted as causal. The fact that something is precisely measured doesn't tell us whether something is causal (which relies on assumptions)

Model Specification

How can we tell whether MLR 1 holds?

- Theory
  - We might have some prior knowledge that tells us whether a specification should be in logs or levels
  - e.g. Estimating Cobb-Douglas or CES production functions, gravity equations in Trade lend themselves to log-linear specifications
- Goodness of fit
  - R^2 tells us which model fits the data better. Assuming we have random sample (MLR 2), this can help with model selection.

TLDR: Significant coefficients are only useful for showing causality if you believe all the MLR assumptions. Otherwise they just show whether a correlation that exists in the data is precise or not.

Hypothesis Testing Review

Test Type	Test Statistic	Distribution
Population mean e.g. $H_0 : \mu = \mu_0$	$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$	$t \sim t_{n-1}$
Difference in population means e.g. $H_0 : \mu_1 - \mu_2 = \mu_0$	$t = \frac{(\bar{x}_1 - \bar{x}_2) - \mu_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$t \sim t_{n_1+n_2-2}$
Population proportion e.g. $H_0 : p = p_0$	$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	$z \sim N(0,1)$
Difference in population proportions e.g. $H_0 : p_1 - p_2 = p_0$	$z = \frac{(\hat{p}_1 - \hat{p}_2) - p_0}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}}$	$z \sim N(0,1)$
True regression parameter (k variables) e.g. $H_0 : \beta = \beta_0$	$t = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})}$	$t \sim t_{n-k-1}$
Multiple restrictions in regression (q restrictions, k total variables in UR model)	$F = \frac{(SSR_R - SSR_{UR})/q}{(SSR_{UR})/(n-k_{UR}-1)}$	$F \sim F_{q,n-k-1}$

Using R output for other β hypothesis tests

R output gives t-statistics and p-values for two-sided $H_0 : \beta = 0$.
What if you want to test a different hypothesis?

```
Call:
lm(formula = lwage ~ female + tenure, data = wage_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.00085 -0.28200 -0.06232  0.31200  1.57325

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.688842   0.034368  49.141  < 2e-16 ***
female      -0.342132   0.042267  -8.095 4.06e-15 ***
tenure       0.019265   0.002925   6.585 1.11e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4747 on 523 degrees of freedom
Multiple R-squared:  0.2055,    Adjusted R-squared:  0.2025
F-statistic: 67.64 on 2 and 523 DF,  p-value: < 2.2e-16
```

You have all the information you need!

- $t = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})}$
- Plug in $\hat{\beta}$ and $SE\left(\hat{\beta}\right)$ from the regression output, and put in whatever your β_0 is under H_0
- Compare to the appropriate critical value, whether two-sided or one-sided

P-Values

- Recall that the significance level (α) is the probability of *falsely* rejecting the null
- Selecting α can be arbitrary, so instead we can look at the p-value to get a better sense of how strong the evidence is against the null
    - For example, a result with a p-value of 0.049 and a result with a p-value of 0.00001 both result in a rejection at the 5% significance level, but one is much stronger evidence against the null

Definition of p-value

Here are three definitions of p-values (all three are right and essentially mean the same thing):

- The p-value is the largest significance level at which we could carry out the test and still fail to reject the null hypothesis.
- The p-value is the smallest significance level at which the null hypothesis would be rejected.
- The p-value is the probability of obtaining a value of the test statistic as extreme or more extreme than the one actually obtained from the sample under the null (i.e., if the null is true).

Key takeaway: If your p-value is less than your significance level, your reject the null and vice versa.

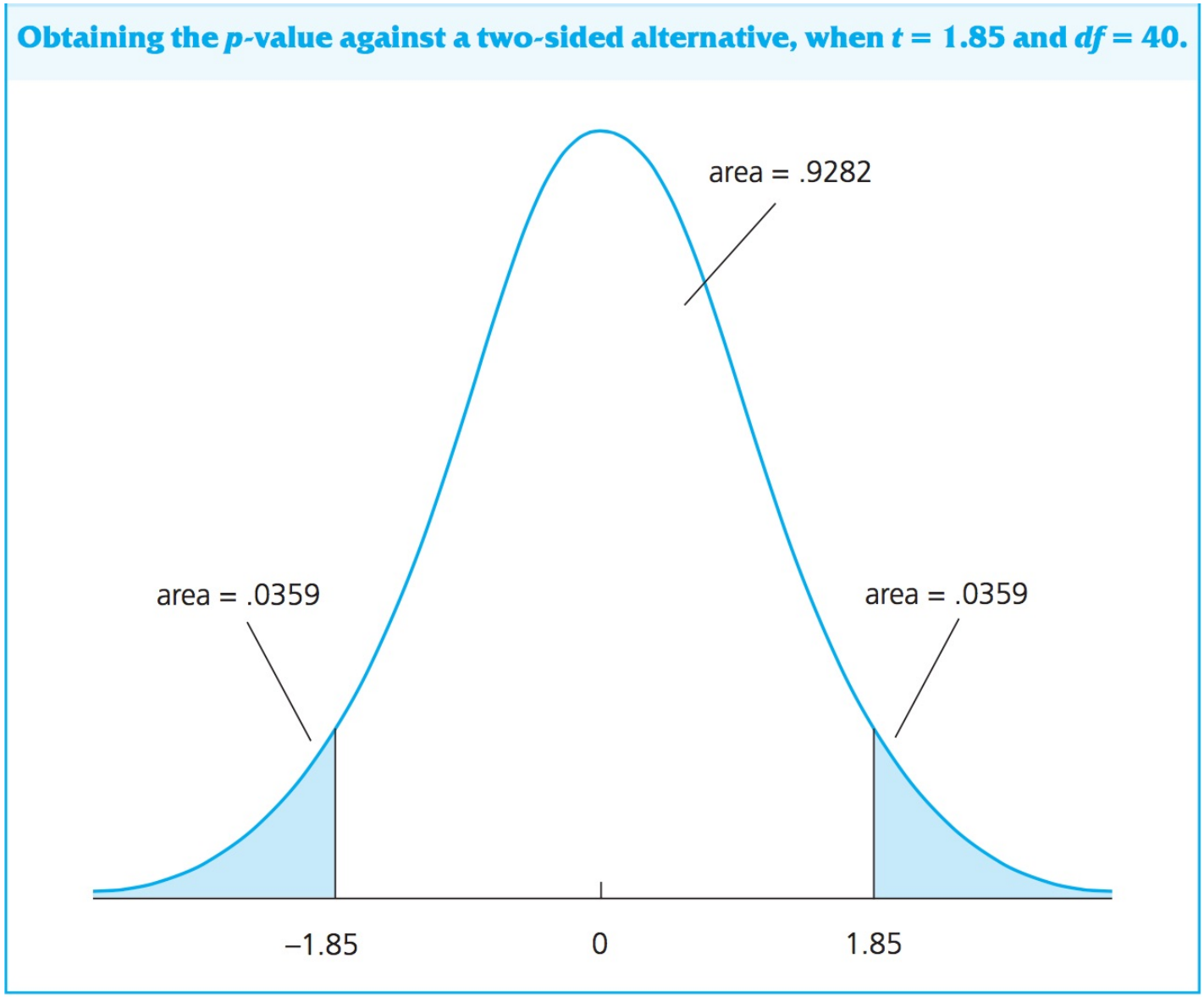
Calculating p-values

- P-values close to zero constitute strong evidence against the null H_0
- Large p-values ()closer to one) constitute weak evidence against the null H_0.

Example: Suppose we calculate a test statistic of t=1.85, with 40 degrees of freedom (two sided). We can find the p-value

$$\text{p-value} = P(|T| > 1.85|H_0) = 2P(T > 1.85) = 2(0.0359) = 0.0718$$

P-value of 0.0718:



Interpretation: Example with reject

- Say the null and alternative hypotheses are $H_0 : \beta_j = 0$ and $H_1 : \beta_j \neq 0$, respectively, and I reject H_0 based on the p-value, and find $\beta_j > 0$.
- Conclude: I reject H_0 in favor of the alternative H_1, and $\hat{\beta}_j$ is statistically greater than zero at the $\alpha\%$ significance level (when we reject a null such as $H_0 : \beta_j = 0$, we usually say our variable x_j is statistically significant).
- This suggests that the variable x_j has a statistically significant and positive relationship with y **holding all other variables** constant.

Interpretation: Example with fail to reject

- Say the null and alternative hypothesis are $H_0 : \beta_j = -1$ and $H_1 : \beta_j \neq -1$ where β_j is the effect of air pollution on housing prices. We estimate the effect to be $\hat{\beta}_j = -0.954$. Doing all the math, we get a t-stat of 0.393, and we see that we cannot reject H_0.
- **But** there are many other values for β_j that we cannot reject. For example, if our null was $H_0 : \beta_j = -0.9$, we would get a t-statistic of -0.462, which cannot be rejected either. But $\beta_j = -1$ and $\beta_j = -0.9$ can't both we true. So it makes no sense to say that we "accept" either of these hypotheses.
- So to conclude, you would say"I fail to reject the null in favor of the alternative. There is no statistical evidence (implicitly, at a 10% significance level or lower) that the impact of air pollution on $prices$ is not -1 **holding all other variables** constant."

Linear Combination Test

Let's say we want to test that two $\hat{\beta}$ parameters are equal to one another. How do we go about that?

Take the following population model we saw in class:

$$log(wage) = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exper + u$$

where:

$$\begin{aligned} jc &= \text{number of years attending a two-year college} \\ univ &= \text{number of years at a four year college} \\ exper &= \text{months in the workforce} \end{aligned}$$

We want to test whether one year of junior college is worth one year at a university. So our null:

$$H_0 : \beta_1 = \beta_2 \qquad H_1 : \beta_1 \neq \beta_2$$

Linear Combination Test

- We can rewrite the hypotheses:

$$H_0 : \beta_1 - \beta_2 = 0 \quad H_1 : \beta_1 - \beta_2 \neq 0$$

  Then our test-statistic becomes:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{se(\hat{\beta}_1 - \hat{\beta}_2)}$$

- Very similar to process of testing the equality of two estimated means:
- But we don't know how to calculate $se(\hat{\beta}_1 - \hat{\beta}_2)$

Linear Combination Test

That's why we have R! Two options in R to test equality of parameters:

1. Use the "car" package to run the command: "linearHypothesis" like so: linearHypothesis(modelr, "jc = univ")
2. Change variables in the regression so that the output tests equality directly.

Linear Combination Test

Example from class of option 2:

Define $\theta_1 = \beta_1 - \beta_2$. Testing $\theta_1 = 0$.

$$\begin{aligned} lwage &= \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exp + u \\ &= \beta_0 + (\theta_1 + \beta_2) jc + \beta_2 univ + \beta_3 exper + u \\ &= \beta_0 + \theta_1 jc + \beta_2 (jc + univ) + \beta_3 exper + u \\ &= \beta_0 + \theta_1 jc + \beta_2 (totcoll) + \beta_3 exper + u \end{aligned}$$

So we can test if θ is different from zero directly. **NOTE:** we could have also replaced jc with $totcoll$ and we would also get the same test on θ

F-Test

Suppose we want to test whether two parameters are *jointly* different from zero. To do this we use an F-test. It's easiest to see how to do this from an example:

- Have data on wages, education (junior college or university), experience, and job location.
- Want to test whether working in/near a large city in general ($lgcity,\ sublg$) matters for wage

Steps:

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

2. Write down the two models the null hypothesis implies:

   Unrestricted model:

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

   Restricted model:

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

- 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

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
- k_{UR} is the number of variables in the unrestricted model

4. Compare the F-stat to the correct critical value found in the F-table. You will need to keep track of **both** numerator degrees of freedom (q) and denominator degrees of freedom ($n - k_{UR} - 1$)

Unrestricted Model

```
Call:
lm(formula = lwage ~ jc + univ + exper + sublg + lgcity, data = twoyear)

Residuals:
     Min       1Q   Median       3Q      Max
-2.09025 -0.27929  0.00423  0.28029  1.79032

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.4656944  0.0211152  69.414  < 2e-16 ***
jc          0.0657284  0.0068181   9.640  < 2e-16 ***
univ        0.0762304  0.0023071  33.041  < 2e-16 ***
exper       0.0049255  0.0001572  31.334  < 2e-16 ***
sublg       0.1008531  0.0186367   5.412 6.46e-08 ***
lgcity      0.0181336  0.0179366   1.011    0.312
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4293 on 6757 degrees of freedom
Multiple R-squared:  0.2258,    Adjusted R-squared:  0.2253
F-statistic: 394.2 on 5 and 6757 DF,  p-value: < 2.2e-16
```

Restricted Model

```
Call:
lm(formula = lwage ~ jc + univ + exper, data = twoyear)

Residuals:
     Min       1Q   Median       3Q      Max
-2.10362 -0.28132  0.00551  0.28518  1.78167

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.4723256  0.0210602  69.910   <2e-16 ***
jc          0.0666967  0.0068288   9.767   <2e-16 ***
univ        0.0768762  0.0023087  33.298   <2e-16 ***
exper       0.0049442  0.0001575  31.397   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4301 on 6759 degrees of freedom
Multiple R-squared:  0.2224,    Adjusted R-squared:  0.2221
F-statistic: 644.5 on 3 and 6759 DF,  p-value: < 2.2e-16
```

F-Test

Fill in the formula:

$$F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(n - k_{UR} - 1)}$$

```
> SSR_U<-sum(modelur$residuals^2)
> SSR_U
[1] 1245.106
> SSR_R<-sum(modelr$residuals^2)
> SSR_R
[1] 1250.544
> summary(modelur)$r.squared
[1] 0.225823
> summary(modelr)$r.squared
[1] 0.222442
> nobs(modelur)
[1] 6763
> nobs(modelr)
[1] 6763
```

F-Test

Fill in the formula:

$$F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(n - k_{UR} - 1)}$$

- $SSR_R = 1250.544$
- $SSR_{UR} = 1245.106$
- $R_R^2 = 0.22242$
- $R_{UR}^2 = 0.225823$
- $n = 6763$
- $k_{UR} = 5$
- $q = 2$

$$F = \frac{(1250.544 - 1245.106)/2}{1245.106/(6763 - 5 - 1)} = 14.75$$

F-Test

OR Fill in the alternative formula:

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

- $SSR_R = 1250.544$
- $SSR_{UR} = 1245.106$
- $R^2_R = 0.22242$
- $R^2_{UR} = 0.225823$
- $n = 6763$
- $k_{UR} = 5$
- $q = 2$

$$F = \frac{(0.225823 - 0.222442)/2}{(1 - 0.225823)/(6763 - 5 - 1)} = 14.75$$