Inference Basics

ECON 490

Taylor Mackay || Email: tmackay@fullerton.edu

Slides Overview

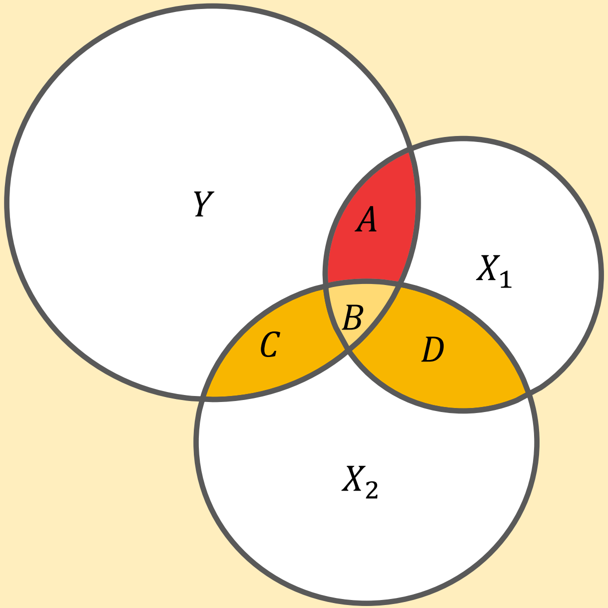
In these slides, we'll discuss:

- How regression isolates unique variation
- Inference and statistical significance in regression
- Using R for inference

Isolating Variation

Last week, we said that regression isolates ***unique*** covariance between Y and X's

- Represented this visually using Venn diagrams
- Today, we'll explore what this means further



Revisiting Residuals

We've defined residuals previously using the following:

$$Residual = Actual\ Y - Predicted\ Y = Y - \hat{Y}$$

$\hat{Y}$ is our "best guess" about the value of Y given our X variables

- In other words, everything in Y that our regression ***can*** explain is reflected in $\hat{Y}$
- The residual is all the "left over" variation in Y → it's what we ***can't*** explain

Key property of residuals – they are ***uncorrelated*** with the explanatory variables in the regression used to generate those residuals

Thinking about Residuals

Suppose we've got the following regression equation: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

We know regression isolates ***unique*** variation in our X's

- What does this mean in practice?
- Consider the following regression:

$$X_1 = \alpha_0 + \alpha_1 X_2 + \epsilon$$

Let's denote the residuals from this regression $\widetilde{X_1}$

- $\widetilde{X_1}$ represents the variation in X_1 that ***cannot*** be explained by X_2
- In other words, the residuals are the ***unique*** variation in X_1

An Example

Consider the following regression (using a sample data set I created):

$$\log(\textit{City Economic Output per Capita}) =$$
$$\beta_0 + \beta_1 \textit{Tech Share of Employment} + \beta_2 \textit{College Graduation Rate} + u$$

Let’s compare the output from two approaches:

1. Directly estimate the regression above
2. Collect $\widetilde{Tech\ Share}$ residuals from $Tech\ Share = \alpha_0 + \alpha_1 College\ Graduation + \epsilon$ then estimate the following:

$$\log(\textit{City Economic Output per Capita}) = \theta_0 + \theta_1 \widetilde{Tech\ Share} + \nu$$

Sample Data

Here's what the `city.data` sample data set we'll use for this example looks like:

city.ID	city.output	tech.share	college.grad
1	$ 68,156	48%	41%
2	$ 50,942	33%	10%
3	$ 50,000	8%	9%
4	$ 67,387	51%	21%
5	$ 72,429	84%	0%
6	$ 85,918	82%	50%
7	$ 91,717	100%	72%
8	$ 80,274	45%	20%
9	$ 85,618	49%	30%
10	$ 70,186	57%	26%

An Example

Option 1

`lm(log(city.output) ~ tech.share + college.grad, city.data)`

Gives us the following output:

```
Coefficients:
             Estimate
(Intercept)   11.0028
tech.share     0.1105
college.grad   0.4727
```

Coefficients on both tech variables are ***identical***
(intercepts differ, but that's not important here)

Option 2

Start by running: `lm(tech.share ~ college.grad, city.data)`

Store residuals $\widetilde{tech.share}$, confirm that correlation with $college.grad$ is 0

Then run the following:

`lm(log(city.output) ~ tech.share.tilde, city.data)`

```
Coefficients:
                 Estimate
(Intercept)       11.2411
tech.share.tilde   0.1105
```

R Code for Output from Previous Slide

```
# First step - run tech.share ~ college.grad regression

step.1.regression <- lm(tech.share ~ college.grad, city.data)

# Collect tech.share.tilde residuals from this regression

city.data$tech.share.tilde <- step.1.regression$residuals

# Check correlation between tech.share.tilde and college.grad - by definition,
# this is equal to 0

round(cor(city.data$tech.share.tilde, city.data$college.grad), 4)

# Second step - run log(city.output) ~ tech.share.tilde regression

step.2.regression <- lm(log(city.output) ~ tech.share.tilde, city.data)

summary(step.2.regression)
```

Setting the Stage

In most econ-related jobs, you'll work with data in some capacity

- Compare revenue growth, sales trends, user retention, etc.
- Analysis almost always entails looking at summary statistics or graphs

You might never be asked to conduct formal statistical inference

- For this class, focus on (1) design and (2) general awareness of uncertainty
- If I was a writing a referee report, discussion would be different!

Thinking about Inference

Up until this point, we've focused on the β's when we've talked about regression

- What is the estimated effect of one variable on another?

Let's imagine there's some "true" relationship that exists between two variables

- If we don't know that true relationship, what do we do?
- Use the data that's available to ***estimate*** that relationship

There's naturally variability in data – who happens to show up in the sample, etc.

- This variability in our data affects our regression estimates

Regression Inference

In our regression review notes, we considered the following regression:

$$Household\ Income = \beta_0 + \beta_1 Age + u$$

Estimating this using our ACS data in R gives us the output on the right

```
lm(formula = hhincome ~ age, data = graph.data)

Residuals:
    Min      1Q  Median      3Q     Max
-141229  -68518  -21386   51575  268930

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    10735      36505    0.29  0.76912
age             3792        989    3.83  0.00019 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9280 on 148 degrees of freedom
Multiple R-squared:  0.0903,	Adjusted R-squared:  0.0842
F-statistic: 14.7 on 1 and 148 DF,  p-value: 0.000186
```

Tonight, we want to focus on this output – start by asking, what's our standard error?

Defining Standard Errors (Pt. 1)

Suppose we have the following regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

- How does R calculate the SE for our estimate of β_1?

Key "ingredient" – unique or residual variation in our X's

- To isolate that variation, revisit the following regression:

$$\boxed{X_1 = \alpha_0 + \alpha_1 X_2 + \epsilon}$$

Remember, residuals from this regression represent variation in X_1 that ***cannot*** be explained by X_2

Defining Standard Errors (Pt. 2)

We had the following regression from last slide: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$

- We estimate this using `lm()` in R
- The formula below gives us the SE for our estimate of β_1

$SE(\hat{\beta}_1) = \frac{\sigma_e}{\sqrt{n}} \times \frac{1}{\sigma_{\tilde{X}_1}}$	• σ_e is the standard deviation of our residuals (what our model can't explain) • n is the sample size (number of observations) • $\sigma_{\tilde{X}_1}$ is the standard deviation of $\widetilde{X_1}$ = residuals from a regression of X_1 on X_2

Residuals from our $X_1 = \alpha_0 + \alpha_1 X_2 + \epsilon$ regression from last slide

Interpreting our Standard Error Formula

$SE(\hat{\beta}_1) = \frac{\sigma_e}{\sqrt{n}} \times \frac{1}{\sigma_{\tilde{X}_1}}$	• σ_e is the standard deviation of our residuals (what our model can't explain) • n is the sample size (number of observations) • $\sigma_{\tilde{X}_1}$ is the standard deviation of the residuals from a regression of X_1 on X_1

Don't need to memorize this formula! You should remember the following:

1. Adding observations can improve precision (larger $\sqrt{n}$ = smaller SEs)
2. Adding X's can help improve precision (smaller σ_e = smaller SEs)…
3. …but this isn't guaranteed (depends on how $\sigma_{\tilde{X}_1}$ changes)

For (2) and (3), what matters is ***unique*** variability in X_1 (not shared with X_2)

Sampling Variability

We started this section by talking about sampling variability

- How does this connect with our SE definition?

Imagine rewinding the clock on our data and letting things play out again

- Not just collecting another survey, but people going to work, etc.
- Big picture's the same … but the details might differ

If we re-estimate our regression, our estimated β_1 won't be exactly the same

- Our standard error is trying to give us a sense of that variability

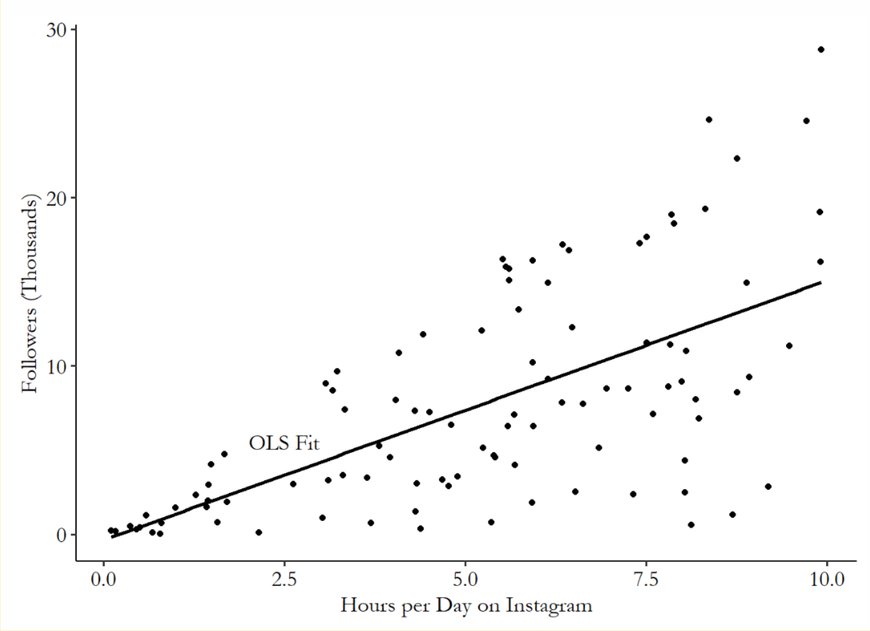
Heteroskedasticity (HK)

HK occurs when our model does a better job explaining Y for some values of X than other values

HK causes our normal SE's to be wrong

In practice, this happens all the time

- As a result, we want a method of ***correcting*** for HK
- Do this by using HK-robust SE's



Inference in R

Up to this point, we've used the `lm()` function to estimate OLS regression

- This reports "vanilla" standard errors
- What if we want heteroskedasticity (HK) robust SE's?

Use `lm_robust()` function from the `estimatr` package

Two benefits to `lm_robust()`:

- Robust SEs + p-values + etc.
- Summary output is easier to work with in R

Reading Regression Results

```
Call:
lm_robust(formula = hhincome ~ sex + age + education + statefip,
    data = subset.data, clusters = statefip, se_type = "stata")

Standard error type:  stata

Coefficients:
                      Estimate Std. Error  t value  Pr(>|t|) CI Lower CI Upper D
(Intercept)           141294.4    2283.92   61.865 2.612e-04 131467.5 151121.3  2
sex                    -6767.8     995.06   -6.801 2.094e-02 -11049.2  -2486.4  2
age                     -506.8      80.09   -6.327 2.408e-02   -851.4   -162.2  2
educationHS Grad        5099.2    1020.20    4.998 3.778e-02    709.6   9488.7  2
educationCollege Grad  73217.4    7110.49   10.297 9.300e-03  42623.4 103811.3  2
statefipFlorida       -29365.7     440.63  -66.645 2.251e-04 -31261.6 -27469.9  2
statefipTexas         -25113.8     211.93 -118.501 7.121e-05 -26025.7 -24201.9  2
```

Key components:

- Coefficient estimate
- Std. Error & P-Value
- Confidence Interval (CI)

```
> round(model$p.value, 3)
          (Intercept)                   sex                   age      educationHS Grad
                0.000                 0.021                 0.024                 0.038
educationCollege Grad       statefipFlorida         statefipTexas
                0.009                 0.000                 0.000
> |
```

Use round() to check size of p-values

P-Values

Statistics like t-stats and p-values are calculated using your β estimate and SE

In practice, p-values are the easiest way to check statistical significance

- Generally, estimates with p-values less than 0.05 are statistically significant
- What does this mean?

Imagine collecting our data again and estimating the same regression

- What's the probability we observe a t-stat at least as large as our original value, *if there was actually no "true" effect*?

Confidence Intervals

95 pct. confidence interval around a coefficient:

$$\hat{\alpha}_1 \pm 1.96 * SE_{\hat{\alpha}_1}$$

Useful for characterizing the ***practical*** significance of an estimate

- What range of estimates can we rule out?
- Sometimes, precisely estimated 0's are informative

F-Tests

Regression output tells us statistical significance of individual coefficients

- Want to test the joint significance of multiple coefficients?
- Use the `linearHypothesis()` function from the `car` package

```
model <- lm_robust(data = acs.data, hhincome ~ sex + age + education)

# Joint test that both age and education are equal to 0. P-value is reported
# is the last number on second row:

linearHypothesis(model, c("age = 0", "education = 0"))
```

```
Linear hypothesis test

Hypothesis:
age = 0
education = 0

Model 1: restricted model
Model 2: hhincome ~ sex + age + education

  Res.Df Df Chisq Pr(>Chisq)
1 246964
2 246962  2 13054  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpreting Statistical Tests

P-values ***don't*** tell us the probability our estimate is correct

- Likewise, confidence intervals aren't measuring (subjective) confidence
- Instead, describing what would happen if we fired up our time machine

In practice, try closely reading the language journal articles use

- Initially formal language (reject / fail to reject, etc.), then a change of tone
- Think about distinguishing between signal and noise

Precise estimates are more likely to be signal than noise … but no guarantees!

Practical Advice (Pt. 1)

General rule of thumb – p-value less than 0.05 is “statistically significant”

- In some contexts, p-value < 0.10 is the standard
- For papers, defer to authors’ interpretations

Practical significance is what ultimately matters

- How large is your estimated effect?

For your capstone analysis, use robust SE’s via `lm_robust()` in R

Practical Advice (Pt. 2)

What "breaks" normal inference? What causes problems for regression?

- Main pitfall = having too few observations
- Context matters – how many observations is enough?

A ***very*** rough rule of thumb = you should have ***at least ~50 observations***

- What comparisons do you care about?
- If you care about estimating a difference between groups, then ideally, you'd like ***each*** group to have 40-50 observations

Sometimes, just having lots of data isn't enough

- Suppose you have GDP data for the US → lots of years… but only one country
- We'll talk about this issue more this later