EC 200

Exam Review

Upcoming:

(Optional) rough draft	**Nov 30**
Exam Q & A	Nov 30
Exam	**Dec 2**
Presentation due	**Dec 4**
Paper due	**Dec 8**

WOW!

Research proposal feedback

Introduction


Keith Head

Home Publications Research Blog

The Introduction Formula

When I arrived at UBC, my colleague John Ries, who had been hired the year before, explained to me that Jim Brander had given him a formula for writing introductions. I'm afraid I didn't pay much attention at the time because I thought it would stifle my creative juices (is that a mixed metaphor?). Finally, I think I ended up internalizing the rules and now I thought I should make them explicit because they have served us well and I wish I could referee more papers that follow them.

1. **Hook:** Attract the reader's interest by telling them that this paper relates to something interesting. What makes a topic interesting? Some combination of the following attributes makes Y something worth looking at.

- As you move to your paper, you'll need an introduction.
- Introduction should **stand alone** (no surprises at the end!)
- A few suggestions from Head, Sahm, and Evans
- Hook >> Question(s) >> Approach >> Results >> Contribution

Literature review/background

- **CONDENSE**

NO

In the paper “Here is the title of my paper,” the authors Benjamin and Locke conduct a study to measure the impact of sunshine on ice cream viscosity. They find that sunshine increases the rate of change of the ice cream viscosity.”

BETTER

Benjamin and Locke (2019) find that sunshine increases the rate of change of ice cream viscosity

BEST

Sunshine melts ice cream (Benjamin and Lock, 2019)

EC 200

Literature review

- **SYNTHESIZE**
- What does the body of evidence collectively tell you?
- Where does your work fit in?

Level 1: There are many papers on ice cream. Hock and Jam (2015) find that ice cream is a delicious food. Ruddiger and Patel (2012) find that chocolate is a good flavor of ice cream.

Level 2: Hock and Jam (2015) find that ice cream is a delicious food. Recent studies find that chocolate ice cream is especially delicious (Ruddiger and Patel 2012). I will extend Ruddiger and Patel's analysis by considering pistachio ice cream.

Level 3: Several papers find that ice cream is delicious (Hock and Jam 2015; Tyrone and Pumba 2001), but recent studies have questioned these findings (Smith and Smithy 2019; Smitty and Smith 2020). I will examine the tastiness of ice cream with newer data and examine the role of air temperature, an potentially important mediating factor (Yang and Dobbles 2018).

EC 200

Other notes

- Write a population model: appropriate subscripts, error terms
- Active voice, don't be afraid of **I**
- Remove personal motivation

You've got this!

EC 200

Exam Review

- Coverage
  - Chapter 8: Non-linear regression
  - Chapter 9: Internal/External validity
  - Chapter 10: Panel Data
  - Chapter 12: Instrumental variables

- Format
  - Same as last exam (on BB)

EC 200

Big picture

- **What can go wrong with our regressions?**
  - Omitted variable bias (Always)
  - Erroneous functional form (Chapter 8)
  - Measurement error (Chapter 9)
  - Reverse causality (Chapter 9/12)
- **How can we solve these problems?**
  - Add more controls (always)
  - Add higher-order terms and/or interactions (Chapter 8)
  - Difference-in-differences model (Chapter 10)
  - First-differences model (Chapter 10)
  - Fixed effects model (Chapter 10)
  - Instrumental variables model (Chapter 12)

EC 200

What you need to know how to do

- What can go wrong with our regressions?
  - Omitted variable bias (Always)
  - Erroneous functional form (Chapter 8)
  - Measurement error (Chapter 9)
  - Reverse causality (Chapter 12)
- **Based on descriptions of regressions, questions, data sets**
  - Identify when these problems are likely to occur
  - Provide specific examples of what these problems look like
  - Discuss the impact this will have on your estimated regression coefficients
  - Discuss the impact this will have on your ability to determine causal relationships

EC 200

What you need to know how to do

- How can we solve these problems?
  - Difference-in-differences model (Chapter 10)
  - First-differences model (Chapter 10)
  - Fixed effects model (Chapter 10)
  - Instrumental variables model (Chapter 12)
- Write population models of these models
- Write step-by-step how to implement these models
- Review results of estimation of these models, interpret coefficients, and “big picture” interpretation.
- Compare results from these models with OLS and discuss which is more appropriate and why

EC 200

General skills you need

- Look at Stata output and/or formatted tables
    - Interpret coefficients (put numbers with them, and units!)
    - Interpret statistical significance (practice with those p-values)
    - Set up hypotheses and determine results
        - That a regression coefficient = 0
        - That multiple exclusion restrictions hold
        - Remember:
            - Set up a null
            - Set up an alternative
            - Compute a test statistic or p-value
            - Make a conclusion

EC 200

Non-linear functions

- Polynomials
  - Compute effects by derivative (approximate) or by calculating for each value and taking the difference (exact)
- Logs
- Interaction terms
  - Binary-binary
  - Continuous-binary
  - Continuous-continuous

EC 200

Using logs to compute percentage changes

- **We do not take logs of percents/etc.**
  - **If LFP is 75% → easy to think about 5pp increase (levels)**
    - **→ harder to think about about 5% increase → 0.05/0.75 = 6.7pp increase**
- Suppose we want to model hourly wages (wage) as a function of years of education (educ)

wage = 10.5 + 3educ

Level-level: A 1-year increase in years of education is associated with a $3 increase in wages (unit-unit)

log(wage) = 10.5 + 3log(educ)

Log-log (elasticity): A 1% increase in years of education is associated with a 3% increase in wages

EC 200

Using logs to compute percentage changes

$\log(wage) = 10.5 + 3educ$

Log-level (semi-elasticity): A 1-year increase in years of education is associated with a 300% increase in wages

(approximation)

$wage = 10.5 + 3\log(educ)$

Level-log: A 1% increase in years of education is associated with a 3/100 = $0.03 increase in wages

(approximation)

Example

`assaults` = number of assaults in a particular weekend across a subset of US counties

`attend` = total weekend movie attendance (millions)

```
. sum assaults attend
```

Variable	Obs	Mean	Std. Dev.	Min	Max
assaults	516	4352.663	2120.995	683	8719
attend	516	18.86187	4.906061	9.8085	36.5028

EC 200

Interpret the coefficient on `attend`

```
. regress assaults attend

      Source |       SS           df       MS      Number of obs   =       516
-------------+----------------------------------   F(1, 514)       =     28.40
       Model |  121306939         1   121306939   Prob > F        =    0.0000
    Residual |  2.1955e+09       514  4271367.88   R-squared       =    0.0524
-------------+----------------------------------   Adj R-squared   =    0.0505
       Total |  2.3168e+09       515  4498621.42   Root MSE        =    2066.7

------------------------------------------------------------------------------
    assaults |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      attend |   98.92505   18.56295     5.33   0.000     62.45647    135.3936
       _cons |   2486.752   361.7598     6.87   0.000     1776.042    3197.461
------------------------------------------------------------------------------
```

1 million more attendees associated w/ 98 more weekend assaults.

EC 200

Interpret the coefficient on `ln_attend`

```
. regress ln_assaults ln_attend
```

Source	SS	df	MS
Model	15.652297	1	15.652297
Residual	189.04063	514	.367783327
Total	204.692927	515	.397461994

Number of obs	=	516
F(1, 514)	=	42.56
Prob > F	=	0.0000
R-squared	=	0.0765
Adj R-squared	=	0.0747
Root MSE	=	.60645

ln_assaults	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ln_attend	.6788489	.1040591	6.52	0.000	.4744154	.8832824
_cons	6.244118	.3033823	20.58	0.000	5.648096	6.84014

1% increase in attendance associated with 0.67% increase in assaults

EC 200

Interpret the coefficient on `attend`

```
. regress ln_assaults attend ,robust

Linear regression                               Number of obs     =        516
                                                F(1, 514)         =      29.21
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0633
                                                Root MSE          =     .61077
```

ln_assaults	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
attend	.0323187	.0059794	5.40	0.000	.0205716	.0440659
_cons	7.606019	.1207744	62.98	0.000	7.368746	7.843291

0.032 → When attendance increases by 1 million, assaults increase by 3.2%

EC 200

Interpret the coefficient on `ln_attend`

```
. regress assaults ln_attend

      Source |       SS           df       MS      Number of obs   =       516
-------------+----------------------------------   F(1, 514)       =     33.54
       Model |   141908410         1   141908410   Prob > F        =    0.0000
    Residual |  2.1749e+09       514   4231287.2   R-squared       =    0.0613
-------------+----------------------------------   Adj R-squared   =    0.0594
       Total |  2.3168e+09       515  4498621.42   Root MSE        =      2057

------------------------------------------------------------------------------
    assaults |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   ln_attend |   2044.034   352.9558     5.79   0.000     1350.621    2737.448
       _cons |   -1583.56   1029.036    -1.54   0.124    -3605.194    438.0734
------------------------------------------------------------------------------
```

EC 200

When 1% increase in attendance leads to 20.44 assautls

Interaction terms

```
. reg sleepdef male hrstotwrk yngkid marr maleXmarr maleXyngkid maleXhrs
```

Source	SS	df	MS
Model	8.81324949	7	1.25903564
Residual	105.222161	698	.150748082
Total	114.035411	705	.161752356

Number of obs =	706
F(7, 698) =	8.35
Prob > F =	0.0000
R-squared =	0.0773
Adj R-squared =	0.0680
Root MSE =	.38826

sleepdef	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
male	-.1301368	.1032824	-1.26	0.208	-.3329181	.0726445
hrstotwrk	.0053348	.0014888	3.58	0.000	.0024119	.0082578
yngkid	.1116302	.0787625	1.42	0.157	-.0430096	.2662699
marr	-.1240539	.0521929	-2.38	0.018	-.2265278	-.02158
maleXmarr	.004336	.0795358	0.05	0.957	-.1518221	.1604941
maleXyngkid	-.0827995	.0953117	-0.87	0.385	-.2699315	.1043325
maleXhrstotwrk	.0023426	.0020265	1.16	0.248	-.0016361	.0063213
_cons	.126543	.0671915	1.88	0.060	-.0053787	.2584647

What is the predicted probability of being sleep deficient for a married woman with young kids who works 40 hours/week? For an equivalent man?

EC 200

Chapter 9

- Internal Validity
  - OBV → correlation between x and u non-zero → endogeneity
  - Errors in measurements!
  - Simultaneous causality bias
  - Functional form error
  - Selection bias
- External validity → we know what we set out to find out, but is it valid/applicable to tother populations/setting

EC 200

Internal/External Validity

Internal Validity (5 threats)

Do we measure what we meant to measure?

- Omitted variable bias
- Bad functional form
- Missing data/sample selection
- Measurement error
- Simultaneity

External validity

Do the results generalize?

- What if we change the setting?
- What if we change the population?

Measurement error

- Dependent variable (if uncorrelated with x)
  - Reduces precision
  - Does not affect coefficients

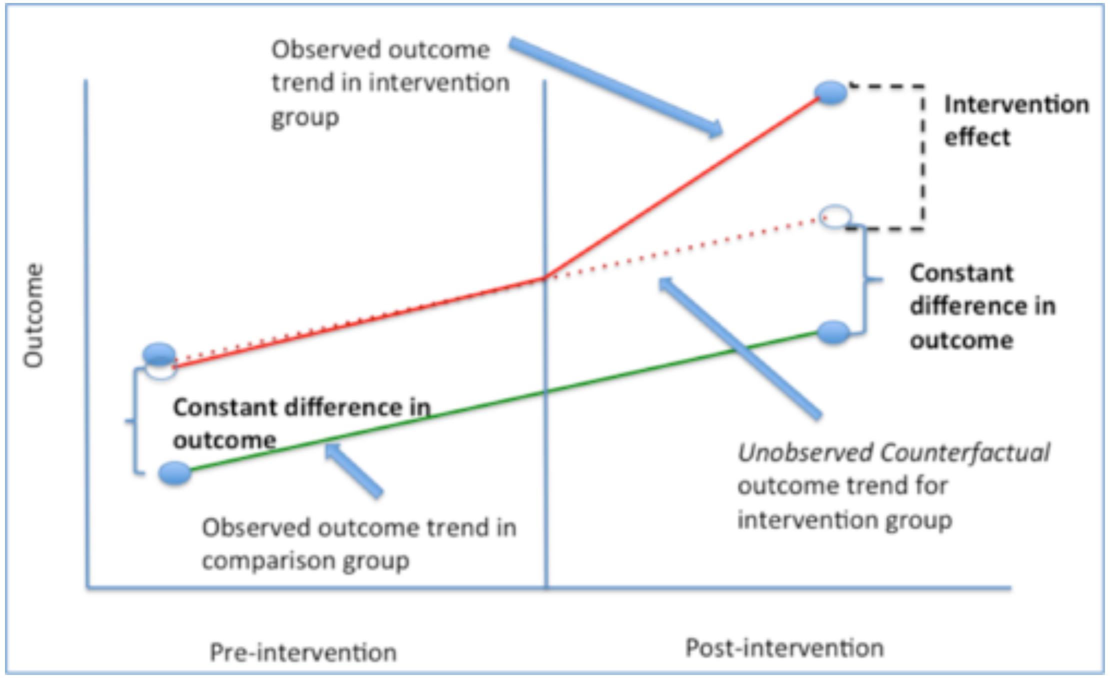
- Independent variable
  - Classical (at random)
    - Attenuation bias

$$\widehat{\beta_1} \overset{p}{\rightarrow} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \beta_1$$

  - Non-classical (not at random)
    - Bias!

Panel data methods

- Difference-in-differences
  - Requires ”natural experiment”
  - For our purposes, before and after, “treatment” and “control”
  - Assumption of parallel trends



$$y_{it} = \beta_0 + \beta_1 Post_t + \beta_2 Treat_i + \beta_3 Post_t X Treat_i + u_{it}$$

Panel data methods

- First differences:
  - Measure impact of change in x on change in y!
  - Subtract out any time-invariant characteristics

$$\Delta y_i = \beta_0 + \beta_1 \Delta x_i + u_i$$

- Fixed effects
  - Control specifically for individual/unit-specific effects!
  - Control specifically for time-invariant effects
  - Still assume no omitted variables

$$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + b_t + u_{it}$$

Instrumental variables

- Find an instrument: something that manipulates Y *only through* manipulating X
  - That is, corr(z,x)> 0 but corr(z,u) = 0!

Good instruments are…

- **Powerful:** (First stage F-stat > 10)
- **Excludable:** Not correlated with y directly
- **Exogenous:** Not correlated with other unobserved factors

Instrumental variables

- First stage

$$x_1 = \alpha_0 + \alpha_1 z + \alpha_2 x_2 + v$$

$$\rightarrow \widehat{x_1} = \widehat{\alpha_0} + \widehat{\alpha_1} z + \widehat{\alpha_2} x_2$$

- Second stage

$$y = \beta_0 + \beta_1 \widehat{x_1} + \beta_2 x_2 + u$$

- β_1 is causal impact of x on y among those who responded to z
  - Local average treatment effect
- Covariates (like x_2) can help meet our identification assumptions