FNCE 926
Empirical Methods in CF
Lecture 1 – Linear Regression I

Professor Todd Gormley
Today’s Agenda

- Introduction
- Discussion of Syllabus
- Review of linear regressions
About Me

- PhD from MIT economics
- Undergraduate at Michigan St. Univ.
- Research on bank entry & corporate topics involving risk and governance
Today’s Agenda

- Introduction… about me
- Discussion of Syllabus
- Review of linear regressions
Course Objectives

- Provide toolbox & knowledge of cross-sectional & panel data empirical methods
- Course will have three-pronged approach
  - Lectures will provide you econometric intuition behind each method discussed
  - Course readings expose you to examples of these tools being used in recent research
  - Exercises will force you to use the methods taught in actual data
Reading Materials *Part 1*

- My lecture notes will be your primary source for each econometric tool
- But, please read background texts before lecture [see syllabus for relevant sections]
  - Angrist & Pischke’s *Mostly Harmless*… book
  - Roberts & Whited (2010) paper
  - Greene’s textbook on econometrics
  - Wooldridge’s textbook on panel data
Reading Materials [Part 2]

- We will also be covering 35+ empirical papers; obtain these using Econlit or by going to authors’ SSRN websites for working papers [I’ve provided links]

- Sorry, for copyright reasons, I can’t post the papers to Canvas…

- Just let me know if you have any problem finding a particular paper
Study Groups

- 3 study groups will do in-class presentations
  - Choose own members; can change later if need to
  - Try to split yourself somewhat equally into groups
  - Choose initial groups during today’s break; first group presentations will be in next class!

[More about group presentations in a second…]
Course Structure

- Total of 150 possible points
  - In-class exam [50 points]
  - Five data exercises [25 points]
  - In-class presentations/participation [25 points]
  - Research proposal
    - Rough draft [15 points]
    - Final proposal [35 points]
Exam

- Done in last class, April 26
- More details when we get closer..., but a practice exam is already available on Canvas
Data exercises

- Exercises will ask you to download and manipulate data within Stata
  - E.g. will need to estimate a triple-diff
  - To receive credit, you will send me your DO files; I will then run them on my own dataset to confirm the coding is correct
  - More instructions in handouts [which will be available on Canvas website]
Turning in exercises

- Please upload both DO file and typed answers to Canvas; i.e., we won’t be handing them in during class
  - They will be graded & returned on Canvas
  - Deadline to submit is noon
    \[Canvas \text{ tracks when the file is uploaded}\]
In-class presentations & participation

- In every class (except today), students will present three papers in second half
  - Each study group does one presentation (this is why there needs to be three study groups)
    - But, only one student for each group actually presents
    - Rotate the presenter each week; doing this basically guarantees everyone full participation points
  - Assign papers for next class at end of class
    [*all papers are listed in the syllabus*]
PowerPoint Presentations \textit{[Part 1]}

- Should last for 10 min., no more than 12 min.
  - Summarize \textit{[2-3 minutes]}
  - Analytical discussion which should focus on identification and causality \textit{[6-7 minutes]}
  - Conclusion \textit{[1 minute]}

- Presentations followed by 5-10 minutes discussion; students must read all three papers

- See handout on Canvas for more details
Each student must also type up 2-3 sentence concern for each paper their group does NOT present and turn it in at start of class

- I will randomly select one after each student presentation to further facilitate class discussion
- Write your comment with one of these goals in mind…
  - Write down your own view of “biggest concern”
  - Or, write a concern you think presenter might miss!
Goal of Presentations

- Help you think critically about empirical tools discussed in previous class
- Allow you to see and learn from papers that actually use these techniques
- Gives you practice on presenting; this will be important in the long run
Research Proposal

- You will outline a possible empirical paper that uses tools taught in this course
  - Rough draft due March 22
  - Final proposal due exam week, May 3
- If you want, think of this as a jump start on a possible 2\textsuperscript{nd} or 3\textsuperscript{rd} year paper
- See handout on Canvas for more details
Office Hours & E-mail

- My office hours will be…
  - Thursdays, 1:30-3:00 p.m.
  - Or, by appointment

- Office location: 2458 SH-DH

- Email: tgormley@wharton.upenn.edu
The TA for this course will be…

- Tetiana Davydiuk
  - davydiuk@wharton.upenn.edu

She will be grading the exercises and answering any questions you might have about them.

All other questions can be directed to me!
Tentative Schedule

- See syllabus…
- While exam date & final research proposal deadline are fixed, topics covered and other case due dates may change slightly if there is a sudden and unexpected class cancellation
How the course is structured…

- We will have a 1-2 lectures per ‘tool’
  - I will lecture in first half (except today) on the ‘tool’
  - In the second half of the following class, students will present papers using that particular tool
Canvas

- [https://wharton.instructure.com](https://wharton.instructure.com)

- Things available to download
  - Exercises & solutions *after turned in*
  - Lecture notes
  - Handouts that provide more details on what I expect for presentations & research proposal, including grading templates
  - Practice exam
  - Student presentations *to help study for exam*
Lecture Notes

- I will provide a copy of lecture notes on Canvas before the start of each class

- I strongly encourage printing these out and bringing them with you to class!
Structural estimation lecture

- Prof. Taylor has agreed to give this lecture
- Tuesday, April 19… the usual time
Remaining Items

- 3 hours is long! We’ll take one 10 minute break or two 5 minute breaks
- Read rest of syllabus for other details about the course including:
  - Class schedule or assigned papers are subject to change; I’ll keep you posted of changes
  - Limitation of course; I won’t have time to cover everything you should know, but it will be a good start
Questions

- If you have a question, ask! 😊
  - If you’re confused, you’re probably not alone
  - I don’t mind being interrupted
  - If I’m going too fast, just let me know

- I may not always have an immediate answer, but all questions will be answered eventually

- Any questions?
Today’s Agenda

- Introduction
- Discussion of Syllabus
- Review of linear regressions

My expectation is that you’ve seen most of this before; but it is helpful to review the key ideas that are useful in practice (without all the math)

Despite trying to do much of it without math; today’s lecture likely to be long and tedious… (sorry)
Linear Regression – Outline

- The CEF and causality (very brief)
- Linear OLS model
- Multivariate estimation
- Hypothesis testing
- Miscellaneous issues

We will cover the latter two in the next lecture
Background readings

- Angrist and Pischke
  - Sections 3.1-3.2, 3.4.1

- Wooldridge
  - Sections 4.1 & 4.2

- Greene
  - Chapter 3 and Sections 4.1-4.4, 5.7-5.9, 6.1-6.2
Motivation

- Linear regression is arguably the most popular modeling approach in corporate finance
  - Transparent and intuitive
  - Very robust technique; easy to build on
  - Even if not interested in causality, it is useful for describing the data

Given importance, we will spend today & next lecture reviewing the key ideas
Motivation continued…

- As researchers, we are interested explaining how the world works
  - E.g. how are firms’ choices regarding leverage explained by their investment opportunities
  - I.e., if investment opportunities suddenly jumped for some random reason, how would we expect firms’ leverage to respond on average?
  - More broadly, how is $y$ explained by $x$, where both $y$ and $x$ are random variables?
Linear Regression – Outline

- The CEF and causality (very brief)
  - Random variables & the CEF
  - Using OLS to learn about the CEF
  - Briefly describe “causality”

- Linear OLS model
- Multivariate estimation
- Hypothesis testing
- Miscellaneous issues
A bit about random variables

- With this in mind, it is useful know that any random variable $y$ can be written as

$$y = E(y \mid x) + \varepsilon$$

where $(y, x, \varepsilon)$ are random variables and $E(\varepsilon \mid x) = 0$

- $E(y \mid x)$ is expected value of $y$ given $x$
- In words, $y$ can be broken down into part ‘explained’ by $x$, $E(y \mid x)$, and a piece that is mean independent of $x$, $\varepsilon$
Conditional expectation function (CEF)

- \( \mathbb{E}(y | x) \) is what we call the CEF, and it has very desirable properties
  - Natural way to think about relationship between \( x \) and \( y \)
  - And, it is best predictor of \( y \) given \( x \) in a minimum mean-squared error sense
    - I.e. \( \mathbb{E}(y | x) \) minimizes \( \mathbb{E}[(y-m(x))^2] \), where \( m(x) \) can be any function of \( x \).
CEF visually...

- $E(y|\, x)$ is fixed, but unobservable

- Intuition: for any value of $x$, distribution of $y$ is centered about $E(y|\, x)$

Our goal is to learn about the CEF
Linear Regression – Outline

- The CEF and causality (very brief)
  - Random variables & the CEF
  - Using OLS to learn about the CEF
  - Briefly describe “causality”
- Linear OLS model
- Multivariate estimation
- Hypothesis testing
- Miscellaneous issues
Linear regression and the CEF

- If done correctly, a linear regression can help us uncover what the CEF is
- Consider linear regression model, $y = \beta x + u$
  - $y = \text{dependent variable}$
  - $x = \text{independent variable}$
  - $u = \text{error term (or disturbance)}$
  - $\beta = \text{slope parameter}$
Some additional terminology

- Other terms for $y$...
  - Outcome variable
  - Response variable
  - Explained variable
  - Predicted variable
  - Regressand

- Other terms for $x$...
  - Covariate
  - Control variable
  - Explanatory variable
  - Predictor variable
  - Regressor
Details about $y = \beta x + u$

- $(y, x, u)$ are random variables
- $(y, x)$ are observable
- $(u, \beta)$ are unobservable
  - $u$ captures everything that determines $y$ after accounting for $x$ ['This might be a lot of stuff!']
  - We want to estimate $\beta$
Ordinary Least Squares (OLS)

- Simply put, OLS finds the $\beta$ that minimizes the mean-squared error

$$\beta = \arg\min_b E[(y - bx)^2]$$

- Using first order condition: $E[x(y - \beta x)] = 0$, we have $\beta = E(xy)/E(x^2)$

- **Note:** by definition, the residual from this regression, $y - \beta x$, is uncorrelated with $x$
What great about this linear regression?

- It can be proved that...
  - $\beta x$ is best* linear prediction of $y$ given $x$
  - $\beta x$ is best* linear approximation of $E(y|x)$
    - *‘best’ in terms of minimum mean-squared error

- This is quite useful. I.e. even if $E(y|x)$ is nonlinear, the regression gives us the best linear approximation of it
Linear Regression – Outline

- The CEF and causality (very brief)
  - Random variables & the CEF
  - Using OLS to learn about the CEF
  - Briefly describe “causality”

- Linear OLS model
- Multivariate estimation
- Hypothesis testing
- Miscellaneous issues
What about causality?

- Need to be careful here…
  - How $x$ explains $y$, which this regression helps us understand, is not the same as learning the **causal** effect of $x$ on $y$
  - For that, we need more assumptions…
The basic assumptions \([Part 1]\)

- **Assumption #1:** \(E(u) = 0\)
  - With intercept, this is totally innocuous
  - Just change regression to \(y = \alpha + \beta x + u\), where \(\alpha\) is the intercept term
  - Now suppose, \(E(u) = k \neq 0\)
    - We could rewrite \(u = k + w\), where \(E(w) = 0\)
    - Then, model becomes \(y = (\alpha + k) + \beta x + w\)
    - Intercept is now just \(\alpha + k\), and error, \(w\), is mean zero
    - I.e. Any non-zero mean is absorbed by intercept
The basic assumptions [Part 2]

- Assumption #2: \( E(u \mid x) = E(u) \)
  - In words, average of \( u \) (i.e. unexplained portion of \( y \)) does not depend on value of \( x \)
  - This is “conditional mean independence” (CMI)
    - True if \( x \) and \( u \) are independent of each other
    - Implies \( u \) and \( x \) are uncorrelated

This is the key assumption being made when people make causal inferences
CMI Assumption

- Basically, assumption says you’ve got correct CEF model for causal effect of $x$ on $y$
  - CEF is causal if it describes differences in average outcomes for a change in $x$
  - i.e. increase in $x$ from values $a$ to $b$ is equal to $E(y|x=b) - E(y|x=a)$ [In words?]
  - Easy to see that this is only true if $E(u|x) = E(u)$ [This is done on next slide...]
**Example of why CMI is needed**

- With model \( y = \alpha + \beta x + u \),
  - \( \mathbb{E}(y | x=a) = \alpha + \beta a + \mathbb{E}(u | x=a) \)
  - \( \mathbb{E}(y | x=b) = \alpha + \beta b + \mathbb{E}(u | x=b) \)
  - Thus, \( \mathbb{E}(y | x=b) - \mathbb{E}(y | x=a) = \beta (b-a) + \mathbb{E}(u | x=b) - \mathbb{E}(u | x=a) \)
  - This only equals what we think of as the ‘causal’ effect of \( x \) changing from \( a \) to \( b \) if \( \mathbb{E}(u | x=b) = \mathbb{E}(u | x=a) \)… i.e. CMI assumption holds
Tangent – CMI versus correlation

- CMI (which implies \( x \) and \( u \) are uncorrelated) is needed for no bias
  \[ \text{[which is a finite sample property]} \]

- But, we only need to assume a zero correlation between \( x \) and \( u \) for consistency
  \[ \text{[which is a large sample property]} \]

- More about bias vs. consistency later; but we typically care about consistency, which is why
  I’ll often refer to correlations rather than CMI
Is it plausible?

- Admittedly, there are many reasons why this assumption might be violated
  - Recall, $u$ captures all the factors that affect $y$ other than $x$… It will contain a lot!
  - Let’s just do a couple of examples…
Ex. #1 – Capital structure regression

Consider following firm-level regression:

\[ Leverage_i = \alpha + \beta Profitability_i + u_i \]

- CMI implies average \( u \) is same for each profitability
- Easy to find a few stories why this isn’t true…
  - #1 – unprofitable firms tend to have higher bankruptcy risk, which by tradeoff theory, should mean a lower leverage
  - #2 – unprofitable firms have accumulated less cash, which by pecking order means they should have more leverage
Ex. #2 – Investment

Consider following firm-level regression:

\[ \text{Investment}_i = \alpha + \beta Q_i + u_i \]

- CMI implies average \( u \) is same for each Tobin’s Q
- Easy to find a few stories why this isn’t true…
  - #1 – Firms with low Q might be in distress & invest less
  - #2 – Firms with high Q might be smaller, younger firms that have a harder time raising capital to fund investments
Is there a way to test for CMI?

- Let $\hat{y}$ be the predicted value of $y$, i.e. $\hat{y} = \alpha + \beta x$, where $\alpha$ and $\beta$ are OLS estimates.
- And, let $\hat{u}$ be the residual, i.e. $\hat{u} = y - \hat{y}$.
- Can we prove CMI if residuals if $E(\hat{u}) = 0$ and if $\hat{u}$ is uncorrelated with $x$?

- **Answer:** No! By construction these residuals are mean zero and uncorrelated with $x$. See earlier derivation of OLS estimates.
Identification police

- What people call the “identification police” are those that look for violations of CMI.
  - I.e. the “police” look for a reason why the model’s disturbance is correlated with $x$.
  - Unfortunately, it’s not that hard…
  - Trying to find ways to ensure the CMI assumption holds and causal inferences can be made will be a key focus of this course.
A side note about “endogeneity”

- Many “police” will criticize a model by saying it has an “endogeneity problem” but then don’t say anything further…

- But what does it mean to say there is an “an endogeneity problem”? 
A side note about “endogeneity”

- **My view:** such vague “endogeneity” critics suspect something is potentially wrong, but don’t really know why or how

  - Don’t let this be you! Be specific about what the problem is!

- Violations to CMI can be roughly categorized into three bins… which are?
Three reasons why CMI is violated

- Omitted variable bias
- Measurement error bias
- Simultaneity bias

- We will look at each of these in much more detail in the “Causality” lecture
What “endogenous” means to me

- An “endogenous” $x$ is when its value depends on $y$ (i.e. it determined jointly with $y$ such that there is simultaneity bias).

- But, some use a broader definition to mean any correlation between $x$ and $u$ [e.g. Roberts & Whited (2011)]

- Because of the confusion, I avoid using “endogeneity”; I’d recommend the same for you

- I.e. Be specific about CMI violation; just say omitted variable, measurement error, or simultaneity bias
A note about presentations…

- Think about “causality” when presenting next week and the following week
  - I haven’t yet formalized the various reasons for why “causal” inferences shouldn’t be made; but I’d like you to take a stab at thinking about it
Linear Regression – Outline

- The CEF and causality (very brief)
- Linear OLS model
  - Basic interpretation
  - Rescaling & shifting of variables
  - Incorporating non-linearities
- Multivariate estimation
- Hypothesis testing
- Miscellaneous issues
Interpreting the estimates

- Suppose I estimate the following model of CEO compensation

  \[ salary_i = \alpha + \beta ROE_i + u_i \]

  - Salary for CEO \( i \) is in $000s; ROE is a %

- If you get… \( \hat{\alpha} = 963.2 \)
  \[ \hat{\beta} = 18.50 \]

  - What do these coefficients tell us?
  - Is CMI likely satisfied?
Interpreting the estimates – Answers

\[
salary_i = 963.2 + 18.5ROE_i + u_i
\]

- What do these coefficients tell us?
  - 1 percentage point increase in ROE is associated with $18,500 increase in salary
  - Average salary for CEO with ROE = 0 was equal to $963,200

- Is CMI likely satisfied? Probably not
Linear Regression – Outline

- The CEF and causality (very brief)
- Linear OLS model
  - Basic interpretation
  - Rescaling & shifting of variables
  - Incorporating non-linearities
- Multivariate estimation
- Hypothesis testing
- Miscellaneous issues
Scaling the dependent variable

- What if I change measurement of salary from $000s to $s by multiplying it by 1,000?

  - Estimates were… \( \hat{\alpha} = 963.2 \)
    \[ \hat{\beta} = 18.50 \]
  
  - Now, they will be… \( \hat{\alpha} = 963,200 \)
    \[ \hat{\beta} = 18,500 \]
Scaling $y$ continued...

- Scaling $y$ by an amount $c$ just causes all the estimates to be scaled by the same amount $q$
  - Mathematically, easy to see why…

$$y = \alpha + \beta x + u$$
$$cy = (c\alpha) + (c\beta)x + cu$$

New intercept

New slope
Scaling $y$ continued...

- Notice, the scaling has no effect on the relationship between ROE and salary

  - I.e. because $y$ is expressed in $\$s$ now, $\hat{\beta} = 18,500$ means that a one percentage point increase in ROE is still associated with $\$18,500$ increase in salary
Scaling the *independent variable*

- What if I instead change measurement of ROE from percentage to decimal? (i.e. multiply ROE by 1/100)
  
  Estimates were…
  \[
  \hat{\alpha} = 963.2 \\
  \hat{\beta} = 18.50
  \]
  
  Now, they will be…
  \[
  \hat{\alpha} = 963.2 \\
  \hat{\beta} = 1,850
  \]
Scaling $x$ continued…

- Scaling $x$ by an amount $k$ just causes the slope on $x$ to be scaled by $1/k$.

- Mathematically, easy to see why…

\[ y = \alpha + \beta x + u \]

\[ y = \alpha + \left( \frac{\beta}{k} \right) k x + u \]

Will interpretation of estimates change?

**Answer:** Again, no!
Scaling both $x$ and $y$

- If scale $y$ by an amount $c$ and $x$ by amount $k$, then we get...
  - Intercept scaled by $c$
  - Slope scaled by $c/k$

$$y = \alpha + \beta x + u$$
$$cy = (c\alpha) + \left(\frac{c\beta}{k}\right)kx + cu$$

- When is scaling useful?
Practical application of scaling #1

- No one wants to see a coefficient of 0.000000456 or 1,234,567,890
- Just scale the variables for cosmetic purposes!
  - It will effect coefficients & SEs
  - But, it won’t affect t-stats or inference
To improve interpretation, in terms of found magnitudes, helpful to scale by the variables by their sample standard deviation

- Let $\sigma_x$ and $\sigma_y$ be sample standard deviations of $x$ and $y$ respectively
- Let $c$, the scalar for $y$, be equal to $1/\sigma_y$
- Let $k$, the scalar for $x$, be equal to $1/\sigma_x$
- I.e. unit of $x$ and $y$ is now standard deviations
With the prior rescaling, how would we interpret a slope coefficient of 0.25?

- **Answer** = a 1 s.d. increase in $x$ is associated with $\frac{1}{4}$ s.d. increase in $y$

- The slope tells us how many standard deviations $y$ changes, on average, for a standard deviation change in $x$

- Is 0.25 large in magnitude? What about 0.01?
**Shifting the variables**

- Suppose we instead add $c$ to $y$ and $k$ to $x$ (i.e. we shift $y$ and $x$ up by $c$ and $k$ respectively)

- Will the estimated slope change?
Shifting continued...

- No! Only the estimated intercept will change
  - Mathematically, easy to see why…

\[ y = \alpha + \beta x + u \]
\[ y + c = \alpha + c + \beta x + u \]
\[ y + c = \alpha + c + \beta (x + k) - \beta k + u \]
\[ y + c = (\alpha + c - \beta k) + \beta (x + k) + u \]

New intercept  Slope the same
Practical application of shifting

- To improve interpretation, sometimes helpful to demean $x$ by its sample mean
  - Let $\mu_x$ be the sample mean of $x$; regress $y$ on $x - \mu_x$
  - Intercept now reflects expected value of $y$ for $x = \mu_x$
    \[
y = (\alpha + \beta \mu_x) + \beta (x - \mu_x) + u
    \]
    \[
    E(y \mid x = \mu_x) = (\alpha + \beta \mu_x)
    \]
  - This will be very useful when we get to diff-in-diffs
Break Time

- Let’s take a 10 minute break
Linear Regression – Outline

- The CEF and causality (very brief)
- Linear OLS model
  - Basic interpretation
  - Rescaling & shifting of variables
  - Incorporating non-linearities
- Multivariate estimation
- Hypothesis testing
- Miscellaneous issues
Incorporating non-linearities

- Assuming that the causal CEF is linear may not always be that realistic
  - E.g., consider the following regression
    \[
    \text{wage} = \alpha + \beta \text{education} + u
    \]
  - Why might a linear relationship between the number of years of education and level of wages be unrealistic? How can we fix it?
Better assumption is that each year of education leads to a constant proportionate (i.e. percentage) increase in wages

- Approximation of this intuition captured by…

\[ \ln(wage) = \alpha + \beta \text{ education} + u \]

- I.e. the linear specification is very flexible because it can capture linear relationships between non-linear variables
Common nonlinear function forms

- Regressing Levels on Logs
- Regressing Logs on Levels
- Regressing Logs on Logs

Let’s discuss how to interpret each of these
The usefulness of log

- Log variables are useful because

\[ 100 \times \Delta \ln(y) \approx \% \ \Delta y \]

- **Note:** When I (and others) say “Log”, we really mean the natural logarithm, “Ln”. E.g. if you use the “log” function in Stata, it assumes you meant “ln”
Interpreting log-level regressions

If estimate, the ln(wage) equation, $100 \beta$ will tell you the $\% \Delta$ wage for an additional year of education. To see this…

\[
\ln(\text{wage}) = \alpha + \beta \text{education} + u
\]

\[
\Delta \ln(\text{wage}) = \beta \Delta \text{education}
\]

\[
100 \times \Delta \ln(\text{wage}) = (100 \beta) \Delta \text{education}
\]

\[
\% \Delta \text{wage} \approx (100 \beta) \Delta \text{education}
\]
Log-level interpretation continued...

- The proportionate change in $y$ for a given change in $x$ is assumed constant
  - The change in $y$ is not assumed to be constant... it gets larger as $x$ increases
  - Specifically, $\ln(y)$ is assumed to be linear in $x$; but $y$ is not a linear function of $x$...

$$\ln(y) = \alpha + \beta x + u$$
$$y = \exp(\alpha + \beta x + u)$$
Example interpretation

- Suppose you estimated the wage equation (where wages are $/hour) and got…

\[ \ln(wage) = 0.584 + 0.083 \text{education} \]

- What does an additional year of education get you?
  \[ \text{Answer} = 8.3\% \text{ increase in wages.} \]

- Any potential problems with the specification?

- Should we interpret the intercept?
Interpreting log-log regressions

- If estimate the following...

\[ \ln(y) = \alpha + \beta \ln(x) + u \]

- \( \beta \) is the **elasticity** of \( y \) w.r.t. \( x \)!
  - i.e. \( \beta \) is the percentage change in \( y \) for a percentage change in \( x \)
  - **Note:** regression assumes constant elasticity between \( y \) and \( x \) regardless of level of \( x \)
Suppose you estimated the CEO salary model using logs got the following:

\[ \ln(salary) = 4.822 + 0.257 \ln(sales) \]

What is the interpretation of 0.257?

**Answer** = For each 1% increase in sales, salary increases by 0.257%
Interpreting *level-log* regressions

- If estimate the following…

\[ y = \alpha + \beta \ln(x) + u \]

- \( \beta /100 \) is the change in \( y \) for 1% change in \( x \)
Suppose you estimated the CEO salary model using logs got the following, where salary is expressed in $000s:

\[
salary = 4.822 + 1,812.5 \ln(sales)\]

What is the interpretation of 1,812.5?

**Answer** = For each 1% increase in sales, salary increases by $18,125
Summary of log functional forms

<table>
<thead>
<tr>
<th>Model</th>
<th>Dependent Variable</th>
<th>Independent Variable</th>
<th>Interpretation of $\beta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level-Level</td>
<td>$y$</td>
<td>$x$</td>
<td>$dy = \beta dx$</td>
</tr>
<tr>
<td>Level-Log</td>
<td>$y$</td>
<td>$\ln(x)$</td>
<td>$dy = (\beta/100)% dx$</td>
</tr>
<tr>
<td>Log-Level</td>
<td>$\ln(y)$</td>
<td>$x$</td>
<td>$%dy = (100\beta) dx$</td>
</tr>
<tr>
<td>Log-Log</td>
<td>$\ln(y)$</td>
<td>$\ln(x)$</td>
<td>$%dy = \beta % dx$</td>
</tr>
</tbody>
</table>

- Now, let’s talking about what happens if you change units (i.e. scale) for either $y$ or $x$ in these regressions…
Rescaling logs doesn’t matter [Part 1]

- What happens to intercept & slope if rescale (i.e. change units) of y when in log form?
- **Answer** = Only intercept changes; slope unaffected because it measures proportional change in y in Log-Level model

\[
\log(y) = \alpha + \beta x + u
\]

\[
\log(c) + \log(y) = \log(c) + \alpha + \beta x + u
\]

\[
\log(cy) = (\log(c) + \alpha) + \beta x + u
\]
Rescaling logs doesn’t matter [Part 2]

- Same logic applies to changing scale of $x$ in level-log models... only intercept changes

\[
y = \alpha + \beta \log(x) + u
\]
\[
y + \beta \log(c) = \alpha + \beta \log(x) + \beta \log(c) + u
\]
\[
y = (\alpha - \beta \log(c)) + \beta \log(cx) + u
\]
Rescaling logs doesn’t matter \textit{[Part 3]}

- **Basic message** – If you rescale a logged variable, it will not effect the slope coefficient because you are only looking at proportionate changes.
Log approximation problems

- I once discussed a paper where author argued that allowing capital inflows into country caused -120% change in stock prices during crisis periods...

- **Do you see a problem with this?**

  - Of course! A 120% drop in stock prices isn’t possible. The true percentage change was -70%. Here is where that author went wrong...
Log approximation problems [Part 1]

- Approximation error occurs because as true % $\Delta y$ becomes larger, $100 \Delta \ln(y) \approx \% \Delta y$ becomes a worse approximation.

- To see this, consider a change from $y$ to $y'$ …

  - Ex. #1: $\frac{y' - y}{y} = 5\%$, and $100\Delta \ln(y) = 4.9\%$
  - Ex. #2: $\frac{y' - y}{y} = 75\%$, but $100\Delta \ln(y) = 56\%$
Log approximation problems [Part 2]

![Graph showing the comparison between Approximation and Exact methods for % Change Y vs Delta x. The graph illustrates the increase in percentage change as Delta x increases. The Approximation line is a dashed line, while the Exact line is a solid line. The percentage change increases significantly with Delta x.]
Problem also occurs for negative changes

- Ex. #1: \( \frac{y' - y}{y} = -5\% \), and \( 100 \Delta \ln(y) = -5.1\% \)

- Ex. #2: \( \frac{y' - y}{y} = -75\% \), but \( 100 \Delta \ln(y) = -139\% \)
Log approximation problems [Part 4]

- So, if implied percent change is large, better to convert it to true % change before interpreting the estimate

\[
\ln(y) = \alpha + \beta x + u \\
\ln(y') - \ln(y) = \beta(x' - x) \\
\ln(y'/y) = \beta(x' - x) \\
y'/y = \exp(\beta(x' - x)) \\
\left[\frac{y' - y}{y}\right] \% = 100\left[\exp(\beta(x' - x)) - 1\right]
\]
We can now use this formula to see what true % change in $y$ is for $x' - x = 1$

\[
\frac{(y' - y)}{y} \% = 100\left[ \exp(\beta(x' - x)) - 1 \right]
\]

\[
\frac{(y' - y)}{y} \% = 100\left[ \exp(\beta) - 1 \right]
\]

- If $\beta = 0.56$, the percent change isn’t 56%, it is

\[
100\left[ \exp(0.56) - 1 \right] = 75\%
\]
Recap of last two points on logs

- Two things to keep in mind about using logs
  - Rescaling a logged variable doesn’t affect slope coefficients; it will only affect intercept
  - Log is only approximation for % change; it can be a very bad approximation for large changes
Usefulness of logs – Summary

- Using logs gives coefficients with appealing interpretation

- Can be ignorant about unit of measurement of log variables since they’re proportionate $\Delta s$

- Logs of $y$ or $x$ can mitigate influence of outliers
“Rules of thumb” on when to use logs

- Helpful to take logs for variables with...
  - Positive currency amount
  - Large integral values (e.g. population)

- Don’t take logs for variables measured in years or as proportions

- If \( y \in [0, \infty) \), can take \( \ln(1+y) \), but be careful... nice interpretation no longer true...
What about using $\ln(1+y)$?

- Because $\ln(0)$ doesn’t exist, people use $\ln(1+y)$ for non-negative variables, i.e. $y \in [0, \infty)$
- Be careful interpreting the estimates! Nice interpretation no longer true, especially if a lot of zeros or many small values in $y$ (Why?)
  - **Ex. #1:** What does it mean to go from $\ln(0)$ to $\ln(x>0)$?
  - **Ex. #2:** And, $\ln(x' + 1) - \ln(x+1)$ is not percent change of $x$
- In this case, might be better to scale $y$ by another variable instead, like firm size
**Tangent – Percentage Change**

- What is the percent change in unemployment if it goes from 10% to 9%?
  - This is a 10 percent drop
  - It is a 1 percentage point drop

  - **Percentage change** is \([(x_1 - x_0)/x_0]\)\(\times100\)
  - **Percentage point change** is the raw change in percentages

  Please take care to get this right in description of your empirical results
Models with quadratic terms \([\text{Part 1}]\)

- Consider \(y = \beta_0 + \beta_1 x + \beta_2 x^2 + u\)
- Partial effect of \(x\) is given by...

\[
\Delta y = (\beta_1 + 2\beta_2 x) \Delta x
\]

- What is different about this partial effect relative to everything we’ve seen thus far?

- **Answer** = It depends on the value of \(x\). So, we will need to pick a value of \(x\) to evaluation (e.g. \(\bar{x}\))
If $\hat{\beta}_1 > 0, \hat{\beta}_2 < 0$, then it has parabolic relation

- Turning point = Maximum = $\left| \frac{\hat{\beta}_1}{2\hat{\beta}_2} \right|
- Know where this turning point is! Don’t claim a parabolic relation if it lies outside range of $x$!
- Odd values might imply misspecification or simply mean the quadratic terms are irrelevant and should be excluded from the regression
Linear Regression – Outline

- The CEF and causality (very brief)
- Linear OLS model
- Multivariate estimation
  - Properties & Interpretation
    - Partial regression interpretation
    - $R^2$, bias, and consistency
- Hypothesis testing
- Miscellaneous issues
Motivation

- Rather uncommon that we have just one independent variable
  - So, now we will look at multivariate OLS models and their properties…
Basic multivariable model

- Example with constant and $k$ regressors

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

- Similar identifying assumptions as before
  - No collinearity among covariates [why?]?
  - $E(u|x_1, \ldots, x_k) = 0$

  - Implies no correlation between any $x$ and $u$, which means we have the correct model of the true causal relationship between $y$ and $(x_1, \ldots, x_k)$
Interpretation of estimates

- Estimated intercept, $\hat{\beta}_0$, is predicted value of $y$ when all $x = 0$; sometimes this makes sense, sometimes it doesn’t.

- Estimated slopes, $(\hat{\beta}_1, \ldots, \hat{\beta}_k)$, have a more subtle interpretation now…

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k + \hat{u}$$

- How would you interpret $\hat{\beta}_1$?
Interpretation – *Answer*

- Estimated slopes, \( \hat{\beta}_1, \ldots, \hat{\beta}_k \), have partial effect interpretations.

- Typically, we think about change in just one variable, e.g. \( \Delta x_1 \), holding constant all other variables, i.e. (\( \Delta x_2, \ldots, \Delta x_k \) all equal 0).
  - This is given by \( \Delta \hat{y} = \hat{\beta}_1 \Delta x_1 \).
  - I.e. \( \hat{\beta}_1 \) is the coefficient holding *all else fixed* (ceteris paribus).
Interpretation continued…

- But, can also look at how changes in multiple variables at once affects predicted value of $y$

  - I.e. given changes in $x_1$ through $x_k$, we obtain the predicted change in $y$, $\Delta y$

  \[
  \Delta \hat{y} = \hat{\beta}_1 \Delta x_1 + ... + \hat{\beta}_k \Delta x_k
  \]
Suppose we regress college GPA onto high school GPA (4-point scale) and ACT score for N = 141 university students

\[ \text{colGPA} = 1.29 + 0.453 \text{hsGPA} + 0.0094 \text{ACT} \]

- What does the intercept tell us?
- What does the slope on \( \text{hsGPA} \) tell us?
Example – Answers

- Intercept pretty meaningless… person with zero high school GPA and ACT doesn’t exist

- Example interpretation of slope…
  - Consider two students, Ann and Bob, with identical ACT score, but Ann’s GPA is 1 point higher than Bob. Best prediction of Ann’s college GPA is that it will be 0.453 higher than Bob’s
Now, what is effect of increasing high school GPA by 1 point and ACT by 1 point?

\[
\Delta \text{colGPA} = 0.453 \times \Delta \text{hsGPA} + 0.0094 \times \Delta ACT
\]

\[
\Delta \text{colGPA} = 0.453 + 0.0094
\]

\[
\Delta \text{colGPA} = 0.4624
\]
Lastly, what is effect of increasing high school GPA by 2 points and ACT by 10 points?

\[
\Delta colGPA = 0.453 \times \Delta hsGPA + 0.0094 \times \Delta ACT
\]
\[
\Delta colGPA = 0.453 \times 2 + 0.0094 \times 10
\]
\[
\Delta colGPA = 1
\]
Fitted values and residuals

- Definition of residual for observation $i$, $\hat{u}_i$

$$\hat{u}_i = y_i - \hat{y}_i$$

- Properties of residual and fitted values
  - Sample average of residuals $= 0$; implies that sample average of $\hat{y}$ equals sample average of $y$
  - Sample covariance between each independent variable and residuals $= 0$
  - Point of means $(\bar{y}, \bar{x}_1, \ldots, \bar{x}_k)$ lies on regression line
Tangent about residuals

- Again, it bears repeating…
  - Looking at whether the residuals are correlated with the $x$’s is NOT a test for causality
  - By construction, they are uncorrelated with $x$
  - There is no “test” of whether the CEF is the causal CEF; that justification will need to rely on economic arguments
The CEF and causality (very brief)
Linear OLS model
Multivariate estimation
  - Properties & Interpretation
  - Partial regression interpretation
  - $R^2$, bias, and consistency
Hypothesis testing
Miscellaneous issues
What is wrong with the following? And why?

- Researcher wants to know effect of $x$ on $y$ after controlling for $z$
- So, researcher removes the variation in $y$ that is driven by $z$ by regressing $y$ on $z$ & saves residuals
- Then, researcher regresses these residuals on $x$ and claims to have identified effect of $x$ on $y$ controlling for $z$ using this regression

We’ll answer why it’s wrong in a second...
The following is quite useful to know…

Suppose you want to estimate the following

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \]

Is there another way to get \( \hat{\beta}_1 \) that doesn’t involve estimating this directly?

**Answer:** Yes! You can estimate it by regressing the residuals from a regression of \( y \) on \( x_2 \) onto the residuals from a regression of \( x_1 \) onto \( x_2 \).
Partial regression \[Part 2\]

- To be clear, you get $\hat{\beta}_1$, by…

#1 – Regress $y$ on $x_2$; save residuals (call them $\tilde{y}$)

#2 – Regress $x_1$ on $x_2$; save residuals (call them $\tilde{x}$)

#3 – Regress $\tilde{y}$ onto $\tilde{x}$; the estimated coefficient will be the same as if you’d just run the original multivariate regression!!!
Partial regression – *Interpretation*

- Multivariate estimation is basically finding the effect of each independent variable after partialing out the effect of other variables.

  - I.e. Effect of $x_1$ on $y$ after controlling for $x_2$, (i.e. what you’d get from regressing $y$ on both $x_1$ and $x_2$) is the same as what you get after you partial out the effect $x_2$ from both $x_1$ and $y$ and then run a regression using the residuals.
This property holds more generally...

- Suppose $X_1$ is vector of independent variables
- $X_2$ is vector of more independent variables
- And, you want to know that coefficients on $X_1$ that you would get from a multivariate regression of $y$ onto all the variables in $X_1$ and $X_2$...
Partial regression – Generalized, Part 2

- You can get the coefficients for each variable in $X_1$ by...
  
  - Regress $y$ and each variable in $X_1$ onto all the variables in $X_2$ (at once), save residuals from each regression
  
  - Do a regression of residuals; i.e. regress $y$ onto variables of $X_1$, but replace $y$ and $X_1$ with the residuals from the corresponding regression in step #1
Practical application of partial regression

- Now, what is wrong with the following?
  - Researcher wants to know effect of $x$ on $y$ after controlling for $z$
  - So, researcher removes the variation in $y$ that is driven by $z$ by regressing $y$ on $z$ & saves residuals
  - Then, researcher regresses these residuals on $x$ and claims to have identified effect of $x$ on $y$ controlling for $z$ using this regression
Practical application – Answer

- It’s wrong because it didn’t partial effect of \( z \) out of \( x \)! Therefore, it is NOT the same as regressing \( y \) onto both \( x \) and \( z \)!

- Unfortunately, it is commonly done by researchers in finance [e.g. industry-adjusting]

  - We will see how badly this can mess up things in a later lecture where we look at my paper with David Matsa on unobserved heterogeneity
Linear Regression – Outline

- The CEF and causality (very brief)
- Linear OLS model
- Multivariate estimation
  - Properties & Interpretation
  - Partial regression interpretation
  - $R^2$, bias, and consistency
- Hypothesis testing
- Miscellaneous issues
Goodness-of-Fit ($R^2$)

- A lot is made of $R$-squared; so let’s quickly review exactly what it is

- **Start by defining the following:**
  - Sum of squares total (SST)
  - Sum of squares explained (SSE)
  - Sum of squares residual (SSR)
Definition of SST, SSE, SST

If $N$ is the number of observations and the regression has a constant, then

\[
SST = \sum_{i=1}^{N} (y_i - \bar{y})^2
\]

SST is total variation in $y$

\[
SSE = \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2
\]

SSE is total variation in predicted $y$

\[\text{[mean of predicted } y = \text{mean of } y]\]

\[
SSR = \sum_{i=1}^{N} \hat{u}_i^2
\]

SSR is total variation in residuals

\[\text{[mean of residual } = 0]\]
SSR, SST, and SSE continued...

- The total variation, SST, can be broken into two pieces... the explained part, SSE and unexplained part, SSR

\[ \text{SST} = \text{SSE} + \text{SSR} \]

- \( R^2 \) is just the share of total variation that is explained! In other words,

\[ R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}} \]
More about $R^2$

- As seen on last slide, $R^2$ must be between 0 and 1
- It can also be shown that $R^2$ is equal to the square of the correlation between $y$ and predicted $y$
- If you add an independent variable, $R^2$ will never go down
Adjusted $R^2$

- Because $R^2$ always goes up, we often use what is called Adjusted $R^2$

$$AdjR^2 = 1 - (1 - R^2) \left( \frac{N - 1}{N - 1 - k} \right)$$

- $k = \#$ of regressors, excluding the constant
- Basically, you get penalized for each additional regressor, such that adjusted $R^2$ won’t go up after you add another variable if it doesn’t improve fit much [it can actually go down!]
Interpreting $R^2$

- If I tell you the $R^2$ is 0.014 from a regression, what does that mean? Is it bad?
  - **Answer #1** = It means I’m only explaining about 1.4% of the variation in $y$ with the regressors that I’m including in the regression.
  - **Answer #2** = Not necessarily! It doesn’t mean the model is wrong; you might still be getting a **consistent** estimate of the $\beta$ you care about!
Unbiasedness *versus* Consistency

- When we say an estimate is unbiased or consistent, it means we think it has a causal interpretation…
  
  - I.e. the CMI assumption holds and the $x$’s are all uncorrelated with the disturbance, $u$

- **Bias** refers to finite sample property; **consistency** refers to asymptotic property
More formally...

- An estimate, $\hat{\beta}$, is **unbiased** if $E(\hat{\beta}) = \beta$
  - I.e. on average, the estimate is centered around the true, unobserved value of $\beta$
  - Doesn’t say whether you get a more precise estimate as sample size increases

- An estimate is **consistent** if $\lim_{N \to \infty} \hat{\beta} = \beta$
  - I.e. as sample size increases, the estimate converges (in probability limit) to the true coefficient
Unbiasedness of OLS

- OLS will be unbiased when…
  - Model is linear in parameters
  - We have a random sample of $x$
  - No perfect collinearity between $x$’s
  - $E(u|x_1, \ldots, x_k) = 0$
    
    [Earlier assumptions #1 and #2 give us this]

- Unbiasedness is nice feature of OLS; but in practice, we care more about consistency
Consistency of OLS

- OLS will be consistent when
  - Model is linear in parameters
  - $u$ is not correlated with any of the $x$’s,
    
    \[ \text{[CMI assumptions #1 and #2 give us this]} \]
  
- Again, this is good

- See textbooks for more information
The CEF, $E(y | x)$ has desirable properties

- Linear OLS gives best linear approx. of it
- If correlation between error, $u$, and independent variables, $x$’s, is zero it has causal interpretation

Scaling & shifting of variables doesn’t affect inference, but can be useful

- E.g. demean to give intercepts more meaningful interpretation or rescale for cosmetic purposes
Summary of Today [Part 2]

- Multivariate estimates are partial effects
  - I.e. effect of $x_1$ holding $x_2, \ldots, x_k$ constant
  - Can get same estimates in two steps by first partialing out some variables and regressing residuals on residuals in second step
Assign papers for next week…

- Angrist (AER 1990)
  - Military service & future earnings
- Angrist and Lavy (QJE 1999)
  - Class size & student achievements
- Acemoglu, et al. (AER 2001)
  - Institutions and economic development

These seminal papers in economics with clever identification strategies… i.e., what we aspire to learn about later in the course
In First Half of Next Class

- Finish discussion of the linear regression
  - Hypothesis testing
  - Irrelevant regressors & multicollinearity
  - Binary variables & interactions

- Relevant readings; see syllabus