Linear Regression - II

Professor Ji-Woong Chung
Korea University

This lecture note benefit greatly from the lecture notes of Prof. Todd Gormley.

Outline

Outline

Hypothesis testing

- Before getting to hypothesis testing, which allows us to say something like "our estimate is statistically significant," it is helpful to first look at OLS variance.
- Understanding it and the assumptions made to get it can help us get the right standard errors for our later hypothesis tests.

Variance of OLS Estimators

- Homoskedasticity implies $Var(u|X) = \sigma^2$
  - I.e., Variance of disturbances u does not depend on the level of observed X.
- Heteroskedasticity implies $Var(u|X) = f(X)$
  - I.e., Variance of disturbances u does depend on the level of X in some way.

Which assumption is more realistic?

- In investment regression, which is more realistic, homoskedasticity or heteroskedasticity?

$$Investment = \alpha + \beta Q + u$$

- Answer: Heteroskedasticity seems like a much safer assumption to make; not hard to produce stories on why homoskedasticity is violated.

Heteroskedasticity (HEK) and bias

Does heteroskedasticity cause bias?

- Answer: No! $E(u|X) = 0$ (which is what we need for unbiased estimates) is something entirely different. Heteroskedasticity just affects SEs!
- Heteroskedasticity just means that the OLS estimate may no longer be the most efficient (i.e., precise) linear estimator.

So, why do we care about HEK?

Default is homoskedastic (HOK) SEs

Default standard errors reported by programs like Stata assume HOK.

- If standard errors are heteroskedastic, statistical inferences made from these standard errors might be incorrect

How do we correct for this?

Robust standard errors (SEs)

Use "robust" option to get standard errors (for hypothesis testing) that are robust to heteroskedasticity.

- Typically increases SE, but usually won't make that big of a deal in practice.
- If standard errors go down, could have a problem; use the larger standard errors!
- We will talk about this later in the course.

Using WLS to deal with HEK

- Weighted least squares (WLS) is sometimes used when worried about heteroskedasticity.[1]
  - WLS basically weights the observation of X using an estimate of the variance at that value of X.
  - Done correctly, can improve the precision of estimates.

[1] In reality, we don't know the true variance function. **Feasible WLS (FWLS)** involves a two-step procedure:

1. Run OLS and obtain the residuals $\hat{u}_i$.
2. Model the variance, for example by regressing $\log(\hat{u}_i^2)$ on the predictors X, to get predicted variances $\hat{\sigma}_i^2$.
3. Use $1/\hat{\sigma}_i$ as weights in a WLS regression.

WLS continued a recommendation

- Recommendation of Angrist-Pischke [See Section 3.4.1]: **don't bother with WLS**
  - The efficiency improvements from using Weighted Least Squares (WLS) are often minor and depend on a correct specification of the variance function.
  - If the variance model is incorrectly specified, WLS can lead to biased estimates in finite samples.
  - Ordinary Least Squares (OLS) with robust standard errors is a more reliable alternative because it provides consistent estimates and valid inference without requiring the user to model the variance.

Outline

Hypothesis tests

- This type of phrase is common: "The estimate $\hat{\beta}$ is statistically significant."
  - What does this mean?
  - Answer: "Statistical significance" is generally meant to imply an estimate is statistically different than zero.
- But where does this come from?

Hypothesis tests [Part 2]

- When thinking about significance, it is helpful to remember a few things
  - Estimates of β_1, β_2, etc. are functions of random variables; thus they are random variables with variances and covariances with each other.
  - These variances & covariances can be estimated.
  - Standard error is just the square root of an estimate's estimated variance.

Hypothesis tests [Part 3]

- Reported t-stat is just telling us how many standard deviations our sample estimate, $\hat{\beta}$, is from zero.
  - I.e., it is testing the null hypothesis: $H_0 : \beta = 0$.
  - p-value is just the likelihood that we would get an estimate different from zero by luck if the true $\beta = 0$.

Outline

Statistical vs. Economic Significance

These are not the same!

- Coefficient might be statistically significant but economically small.
  - You can get this in large samples or when you have a lot of variation in X (or outliers).
- Coefficient might be economically large but statistically insignificant.
  - Might just be small sample size or too little variation in X to get a precise estimate.

Economic Significance

You should always check the economic significance of coefficients.

- E.g., how large is the implied change in Y for a standard deviation change in X?
- And importantly, is that plausible? If not, you might have a specification problem.

Outline

Outline

Irrelevant regressors

- What happens if we include a regressor that should not be in the model?
  - We estimate $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$
  - However, the real model is $Y = \beta_0 + \beta_1 X_1 + u$
  - Answer: We still get a consistent estimate of all the β where $\beta_2 = 0$ but our standard errors might go up (making it harder to find statistically significant effects)... see next few slides.

Variance and of OLS estimators

- Greater variance in your estimates increases your standard errors, $\hat{\beta}_j$, making it harder to find statistically significant estimates.
- So it's useful to know what increases $Var(\hat{\beta}_j)$.

Variance formula

- Sampling variance of OLS slope is

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{N}(X_{ij} - \bar{X}_j)^2(1 - R_j^2)}$$

for $j = 1, ..., k$ where R_j^2 is the R^2 from regressing X_j on all other independent variables including the intercept and σ^2 is the variance of the regression error u.

Variance formula – Interpretation

- How will more variation in X affect SE? Why?
- How will higher σ^2 affect SE? Why?
- How will higher R_j^2 affect SE? Why?

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^{N}(X_{ij} - \bar{X}_j)^2(1 - R_j^2)}$$

Variance formula – Variation in X_j

- More variation in X_j is good; smaller SE!
  - Intuitive; more variation in X_j helps us identify its effect on Y!
  - This is why we always want larger samples; it will give us more variation in X_j.

Variance formula – Effect of σ^2

- More error variance means bigger SE.
  - Intuitive; a lot of the variation in Y is explained by things you didn't model.
  - Can add variables that affect Y (even if not necessary for identification) to improve fit.

Variance formula – Effect of R_j^2

- **However**, more variables can also be bad if they are highly collinear.
  - Gets harder to disentangle the effect of the variables that are highly collinear.
  - This is why we don't want to add variables that are "irrelevant" (i.e., they don't affect Y).

Should we include variables that do explain Y and are highly correlated with our X of interest?

Multicollinearity [Part 1]

Highly collinear variables can inflate SEs

- But it does not cause a bias or inconsistency!
- Like a problem of a small sample; with a larger sample, one could get more variation in the independent variables and get more precise estimates

Multicollinearity [Part 2]

Consider the following model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

where X_2 and X_3 are highly correlated.

- $Var(\hat{\beta}_2)$ and $Var(\hat{\beta}_3)$ may be large, but correlation between X_2 and X_3 has no direct effect on $Var(\hat{\beta}_1)$.
- If X_1 is uncorrelated with X_2 and X_3, then $R_1^2 = 0$ and $Var(\hat{\beta}_1)$ is unaffected.

Multicollinearity – Key Takeaways

- It doesn't cause bias.
- Don't include controls that are highly correlated with independent variables of interest if they aren't needed for identification (i.e., $E(u|X) = 0$ without them).
  - But obviously, if $E(u|X) \neq 0$ without these controls, you need them!
  - A larger sample will help increase precision.

Outline

Models with interactions

Sometimes it is helpful for identification to add interactions between X's.

- E.g., theory suggests firms with a high value of X_1 should be more affected by some change in X_2.

The model will look something like ...

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u$$

Interactions – Interpretation [Part 1]

According to this model, what is the effect of increasing X_1 on Y holding all else equal?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + u$$

- Answer: $\frac{\partial Y}{\partial X_1} = \beta_1 + \beta_3 X_2$

Interactions – Interpretation [Part 2]

If $\beta_3 < 0$, how does a higher X_2 affect the partial effect of X_1 on Y?

$$\frac{\partial Y}{\partial X_1} = \beta_1 + \beta_3 X_2$$

- Answer: The increase in Y for a given change in X_1 will be smaller in levels (not necessarily in absolute magnitude) for firms with a higher X_2.

Interactions – Interpretation [Part 3]

Suppose $\beta_1 > 0$ and $\beta_3 < 0$. What is the sign of the effect of an increase in X_1 for the average firm in the population?

$$\frac{\partial Y}{\partial X_1} = \beta_1 + \beta_3 x_2$$

- Answer: It is the sign of

$$\frac{\partial Y}{\partial X_1}\Big|_{x_2=\bar{x}_2} = \beta_1 + \beta_3 \bar{X}_2$$

A very common mistake! [Part 1]

- Researcher claims that "since $\beta_1 > 0$ and $\beta_3 < 0$, an increase in X_1 increases Y for the average firm but the increase is less for firms with a high X_2."

$$\frac{\partial Y}{\partial X_1}\Big|_{X_2=\bar{x}_2} = \beta_1 + \beta_3 \bar{X}_2$$

- Wrong!!! The average effect of an increase in X_1 might be negative if $\bar{X}_2$ is very large!
- β_1 only captures partial effect when $X_2 = 0$, which might not even make sense if X_2 is never 0!

A very common mistake! [Part 2]

- To improve interpretation of β_1, you can reparameterize the model by demeaning each variable in the model and estimate

$$\tilde{Y} = \delta_0 + \delta_1 \tilde{X}_1 + \delta_2 \tilde{X}_2 + \delta_3 \tilde{X}_1 \tilde{X}_2 + u$$

where

$$\begin{aligned} \tilde{Y} &= Y - \mu_Y \\ \tilde{X}_1 &= X_1 - \mu_{X_1} \\ \tilde{X}_2 &= X_2 - \mu_{X_2} \end{aligned}$$

A very common mistake! [Part 3]

You can then show ... $\Delta Y = (\delta_1 + \delta_3 \tilde{X}_2)\Delta X_1$
and thus,

$$\left.\frac{\partial Y}{\partial X_1}\right|_{X_2=\bar{x}_2} = \delta_1 + \delta_3(X_2 - \mu_{X_2})$$
$$\left.\frac{\partial Y}{\partial X_1}\right|_{X_2=\bar{x}_2} = \delta_1$$

Now β_1 tells us the effect of X_1 for the average firm!

The main takeaway – Summary

If you want coefficients on non-interacted variables to reflect the effect of that variable for the "average" firm, demean all your variables before running the specification.

Why is there so much confusion about this? Probably because of indicator variables

Indicator (binary) variables

We will now talk about indicator variables:

- Interpretation of the indicator variables.
- Interpretation when you interact them.
- When demeaning is helpful.
- When using an indicator rather than a continuous variable might make sense.

Motivation

- Indicator variables, also known as binary variables, are quite popular these days.
  - Ex. #1 – Sex of CEO (male/female).
  - Ex. #2 – Employment status (employed/unemployed).
  - Also see in many diff-in-diff specifications.
    - Ex. #1 – Size of firm (above vs. below median).
    - Ex. #2 – Pay of CEO (above vs. below median).

How they work

Code the information using dummy variables:

- Ex.#1:$Male_i = \begin{cases} 1 \text{ if male} \\ 0 \text{ otherwise} \end{cases}$
- Ex.#2:$Large_i = \begin{cases} 1 \text{ if ln(assets)} > \text{median} \\ 0 \text{ otherwise} \end{cases}$

Choice of 0 or 1 is relevant only for interpretation.

Single dummy variable model

Consider: $Wage = \beta_0 + \delta_0 Female + \beta_1 Educ + u$

δ_0 measures the difference in wage between male and female given the same level of education.

- $E(Wage|Female = 0, Educ) = \beta_0 + \beta_1 Educ$
- $E(Wage|Female = 1, Educ) = \beta_0 + \delta_0 + \beta_1 Educ$
- Thus, $E(Wage|F = 0, Educ) - E(Wage|F = 0, Educ) = \delta_0$

Intercept for males $= \beta_0$, females $= \beta_0 + \delta_0$.
Single dummy just shifts intercept!

Single dummy example – Wages

Suppose we estimate the following wage model:

$$Wage = -1.57 - 1.8Female + 0.57Educ + 0.03Exp + 0.14Tenure$$

- Male intercept is -1.57; it is meaningless. Why?
- How should we interpret the 1.8 coefficient?
- Answer: Females earn \$1.80/hour less than men with the same education, experience, and tenure.

Log dependent variable & indicators

Nothing new; coefficient on indicator has % interpretation.

Consider the following example:

$$\ln(Price) = -1.35 + 0.17\ln(lotsize) + 0.71\ln(sqrft) + 0.03bdrms + 0.054colonial + u$$

- Again, negative intercept meaningless; all other variables are never all equal to zero
- Interpretation: Colonial style home costs **about** 5.4% more than "otherwise similar" homes.

Multiple indicator variables

Suppose you want to know how much lower wages are for married and single females.

- Now you have 4 possible outcomes:
  - Single & male
  - Married & male
  - Single & female
  - Married & female
- To estimate, create indicators for three of the variables and add them to the regression.

But which to exclude?

We must exclude one of the four because they are perfectly collinear with the intercept. But does it matter which?

- Answer: No, not really. It just affects the interpretation. Estimates of included indicators will be relative to the excluded indicator.
- For example, if we exclude "single & male," we are estimating the partial change in wage relative to that of single males.

Note: if you don't exclude one, then statistical programs like Stata will just drop one for you automatically. For interpretation, you need to figure out which one was dropped!

Multiple indicators – Example

Consider the following estimation results:

$$\begin{aligned}\ln(\mathit{Wage}) = 0.3 &+ 0.21\mathit{marriedMale} - 0.20\mathit{marriedFemale} \\ &- 0.11\mathit{singleFemale} + 0.08\mathit{Education}\end{aligned}$$

- Single male is omitted; thus intercept is for single males.
- And can interpret other coefficients as
  - Married men earn $\approx 21\%$ more than single males all else equal.
  - Married women earn $\approx 20\%$ less than single males all else equal.

Interactions with Indicators

We could also do the prior regression using interactions between indicators.

- I.e., construct just two indicators 'female' and 'married' and estimate the following:

  $$\ln(Wage) = \beta_0 + \beta_1 Married + \beta_2 Female + \beta_3 (Married \times Female) + u$$

- How will our estimates and interpretation differ from earlier estimates?

Interactions with Indicators [Part 2]

Before we had

$$\ln(Wage) = 0.3 + 0.21 marriedMale - 0.20 marriedFemale - 0.11 singleFemale + 0.08 Education$$

Now, we will have,

$$\ln(Wage) = 0.3 - 0.11 Female + 0.21 Married - 0.30(Female \times Married) + 0.08 Education$$

- Question: Before, married females had wages that were 0.20 lower; how much lower are wages of married females now?

Interactions with Indicators [Part 3]

Answer: It will be the same!

$$\begin{aligned}\ln(\mathit{Wage}) = 0.3 &- 0.11\mathit{Female} + 0.21\mathit{Married} \\ &- 0.30(\mathit{Female} \times \mathit{Married}) + 0.08\mathit{Education}\end{aligned}$$

- Difference for married female $= -0.11 + 0.21 - 0.30 = -0.20$; the same as before

Bottom line = you can do the indicators either way; inference is unaffected

Indicator Interactions – Example

Krueger (1993) found

$$\ln(\textit{Wage}) = \beta_0 + 0.18\,\textit{ComputerWork} + 0.07\,\textit{ComputerHome} + 0.02(\textit{ComputerWork} \times \textit{ComputerHome}) + \ldots$$

- Excluded category = people with no computer
- How do we interpret these estimates?
  - How much higher are wages if you have a computer at work? $\approx 18\%$
  - If you have a computer at home? $\approx 7\%$
  - If you have computers at both work and home? $\approx 27\%$

Indicator Interactions – Example [Part 2]

Remember, these are just approximate percent changes To get true change, need to convert

E.g.,% change in wages for having computers at both home and work is given by

$$100 \times [\exp(0.18 + 0.07 + 0.02) - 1] = 31\%$$

Interacting Indicators w/ Continuous Variables

Adding dummies alone will only shift intercepts for different groups.

However, if we interact these dummies with continuous variables, we can get different slopes for different groups as well.

- See next slide for an example.

Continuous Interactions – Example

Consider the following:

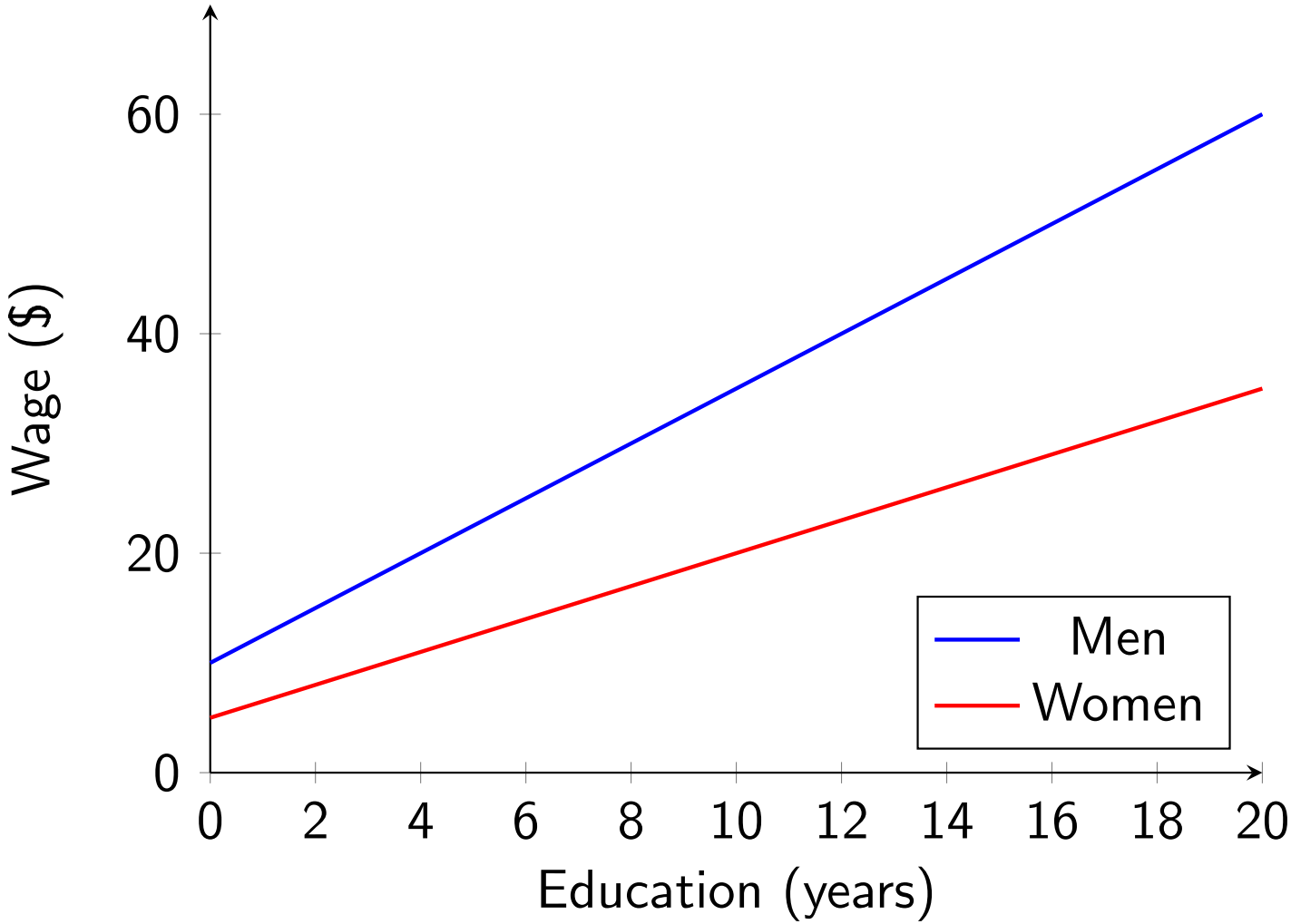
$$\ln(Wage) = \beta_0 + \delta_0 Female + \beta_1 Educ + \delta_1 (Female \times Educ) + u$$

- What is intercept for males? β_0
- What is slope for males? β_1
- What is intercept for females? $\beta_0 + \delta_0$
- What is slope for females? $\beta_1 + \delta_1$

Visual #1 of Example

$\delta_0 < 0, \delta_1 < 0$

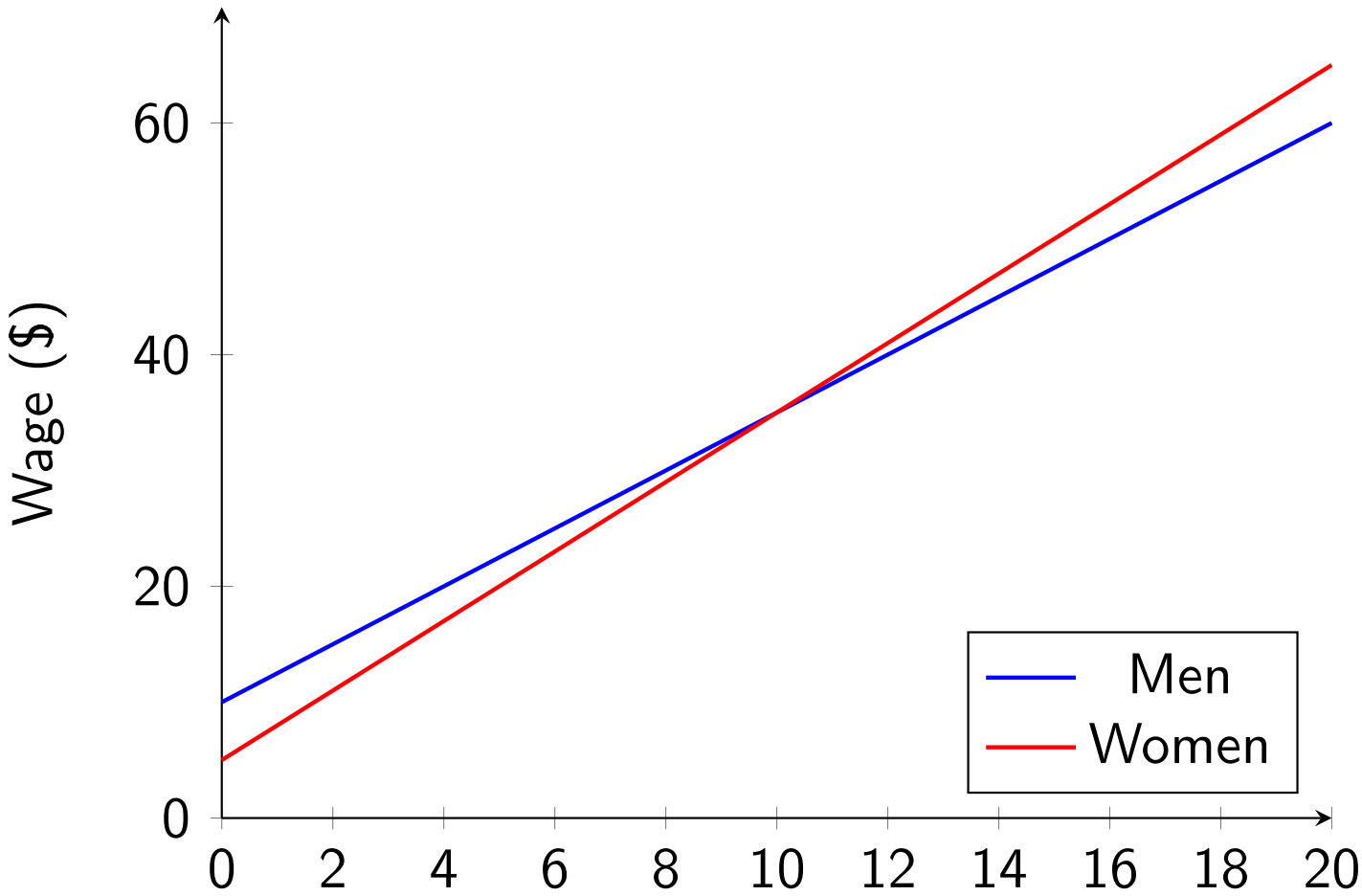
- Females earn lower wages at all levels of education.
- Avg. increase per unit of education is also lower.



Visual #2 of Example

$\delta_0 < 0, \delta_1 > 0$

- Wage is lower for females but only for lower levels of education because their slope is larger.
- Is it fair to conclude that women eventually earn higher wages with enough education?



Cautionary Note on Different Slopes!

Crossing point (where women earn higher wages) might occur outside the data (i.e., at education levels that don't exist).

- Need to solve for crossing point before making this claim about the data.

  $$\text{Women: } \ln(Wage) = \beta_0 + \delta_0 + (\beta_1 + \delta_1)Educ + u$$
  $$\text{Men: } \ln(Wage) = \beta_0 + \beta_1 Educ + u$$

- They are equal when $Educ = \frac{\delta_0}{\delta_1}$.

Cautionary Note on Interpretation!

Interpretation of non-interacted terms when using continuous variables is tricky

E.g., consider the following estimates

$$\ln(Wage) = 0.39 - 0.23 Female + 0.08 Educ - 0.01(Female \times Educ) + u$$

- Return to *Education* is 8% for men, 7% for women
- But, at the average education level, how much less do women earn?

$$[-0.23 - 0.01 \times avg.Education]\%$$

Cautionary Note [Part 2]

Again, interpretation of non-interacted variables does not equal average effect unless you demean the continuous variables

In prior example estimate the following:

$$\begin{aligned} \ln(Wage) = \beta_0 + \delta_0 Female + \beta_1 (Educ - \mu_{Educ}) \\ + \delta_1 Female \times (Educ - \mu_{Educ}) + u \end{aligned}$$

Now, δ_0 tells us how much lower the wage is of women at the average education level

Cautionary Note [Part 3]

Recall! As we discussed in prior lecture, the slopes won't change because of the shift

- Only the intercepts, β_0 and $\beta_0 + \delta_0$, and their standard errors will change

Bottom line = if you want to interpret non-interacted indicators as the effect of indicators at the average of the continuous variables, you need to demean all continuous variables

Ordinal Variables

- Consider credit ratings: $CR \in \{AAA, AA, ..., C, D\}$
- If you want to explain interest rate (IR) with ratings, we could convert CR to numeric scale, e.g., $AAA = 1, AA = 2$, and estimate:

$$\text{IR} = \beta_0 + \beta_1 CR + u$$

- However, what are we implicitly assuming and how might it be a problematic assumption?

Ordinal Variables continued

Answer: We assume a constant linear relation between interest rates and CR.

- I.e., moving from AAA to AA produces the same change as moving from BBB to BB.
- Could take log interest rate, but is a constant proportional change much better? Not really

A better route might be to convert the ordinal variable to indicator variables.

Convert ordinal to indicator variables

E.g., let $CR_{AAA} = 1$ if $CR = AAA$, 0 otherwise; $CR_{AA} = 1$ if $CR = AA$, 0 otherwise, etc.

Then run this regression:

$$IR = \beta_0 + \delta_1 CR_{AAA} + \delta_2 CR_{AA} + \delta_3 CR_A + \ldots + u$$

Remember to exclude one (e.g., "D").

This allows IR change from each rating category (relative to the excluded indicator) to be of different magnitude!

Outline

Reporting regressions

- Table of OLS outputs should generally show the following:
  - Dependent variable [clearly labeled]
  - Independent variables
  - Estimated coefficients, their corresponding standard errors (or t-stat), and ... stars indicating the level of statistical significance
  - R^2
  - # of observations in each regression

Reporting regressions [Part 2]

- In the body of the paper:
  - Focus only on variable(s) of interest.
  - Tell us their sign, magnitude, statistical & economic significance, interpretation, etc.
  - Don't waste time on other coefficients unless they are "strange" (e.g., wrong sign, huge magnitude, etc.).

Reporting regressions [Part 3]

- And last, but not least, don't report regressions in tables that you aren't going to discuss and/or mention in the paper's body
- If it's not important enough to mention in the paper, it's not important enough to be in a table

Summary of Today [Part 1]

- Irrelevant regressors and multicollinearity do not cause bias.
  - However, they can inflate standard errors.
  - So avoid adding unnecessary controls.
- Heteroskedastic variance does not cause bias.
  - Just means the default standard errors for hypothesis testing are incorrect.
  - Use 'robust' standard errors (if larger).

Summary of Today [Part 2]

- Interactions and binary variables can help us get a causal CEF.
  - However, if you want to interpret non-interacted indicators, it is helpful to demean continuous variables.
- When writing up regression results:
  - Make sure you put key items in your tables.
  - Make sure to talk about both economic and statistical significance of estimates.