Econometrics II

Fabian Waldinger (LMU Munich)

Lecture Structure

1. Recap from last lectures
2. Dummy variables
3. Violations of GM2: Dummy variable trap
4. Violations of GM3:
    1. Omitted variable bias
    2. Measurement error in X (next lecture)
    3. Simultaneity (next lecture)

Hiring Research Assistants

- We are currently hiring research assistants.
- Please apply with a CV, a transcript, and a short cover letter at: office.waldinger@econ.lmu.de

Recap from First 2 Weeks of the Course

- GM-assumptions:
  1. The true model is linear in parameters: $y = \boldsymbol{X}\beta + \varepsilon$
  2. No Perfect Collinearity: the matrix $\boldsymbol{X}$ has rank k
  3. Zero Conditional Mean: $E(\varepsilon|\boldsymbol{X}) = 0$
  4. $Var(\varepsilon|\boldsymbol{X}) = \sigma^2\boldsymbol{I}$
  5. $\varepsilon|\boldsymbol{X} \sim N(0, \sigma^2\boldsymbol{I})$
- Under GM1-GM3: OLS is LUE
- Under GM1-GM4: OLS is BLUE
- Under GM1-GM5 we can carry hypothesis tests and construct confidence intervals

Recap from Last Lecture

- t-test:
  1. GM-assumptions
  2. hypotheses
  3. test statistic
  4. critical value
  5. decision rule

Violations of the GM-Assumptions

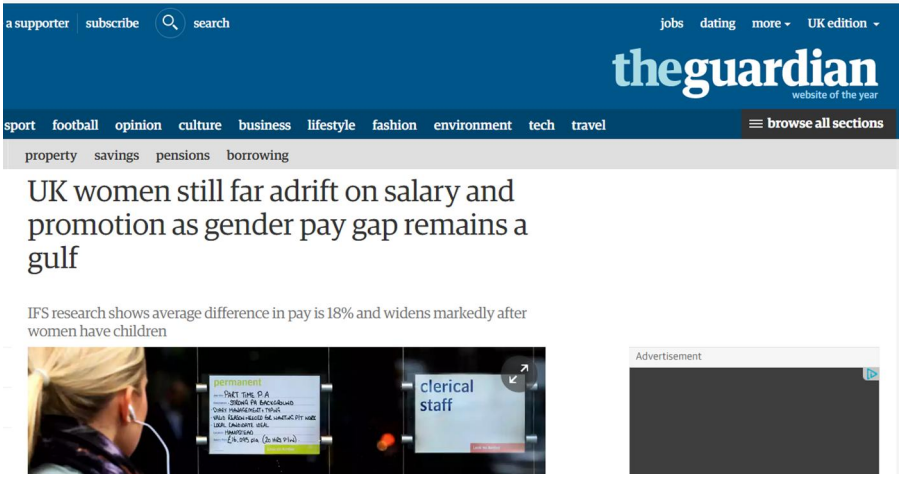
- GM Assumptions:
  1. The true model is linear in parameters: $y = \boldsymbol{X}\beta + \varepsilon$
  2. No Perfect Collinearity: the matrix X has rank k
  3. Zero Conditional Mean: $E(\varepsilon|\boldsymbol{X}) = 0$
  4. $Var(\varepsilon|\boldsymbol{X}) = \sigma^2 I$
  5. $\varepsilon|\boldsymbol{X} \sim N(0, \sigma^2 \boldsymbol{I})$
- Now and in the following lectures we will explore what happens when the GM assumptions are violated
- We start with GM2

Dummy Variables

- Some characteristics can only be coded as variables that take 2 values. Or the dataset will only give information that takes two values E.g.:
  - Female
  - College degree
  - Married
- Dummy variables only take two values 0 or 1, e.g.:

- Also called categorical variables or indicator variables
- We can use such variables as both dependent or explanatory variables, let's start by using them as Xs

Dummy Variables - Example Gender Pay Gap



Example. Data from the ACS 2015

- We investigate how schooling and gender affect log wages
- Let's start with a very simple regression:

$$ln(wage) = \beta_1 + \beta_2 S + \varepsilon$$

ln_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
education_yr	.1479137	.0004984	296.79	0.000	.1469369	.1488905
_cons	8.127956	.0069453	1170.28	0.000	8.114343	8.141569

- Where S measures years of schooling
- What does the coefficient on education_yr mean?

Example: Include Female Dummy

- We now add a dummy variable for female to the regression:

$$ln(wage) = \beta_1 + \beta_2 S + \beta_3 Female + \varepsilon$$

ln_wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
education_yr	.1523815	.0004917	309.91	0.000	.1514178	.1533452
female	-.4296507	.0027837	-154.34	0.000	-.4351067	-.4241947
_cons	8.275142	.0069065	1198.17	0.000	8.261605	8.288678

Dummy Variables: The Mathematics

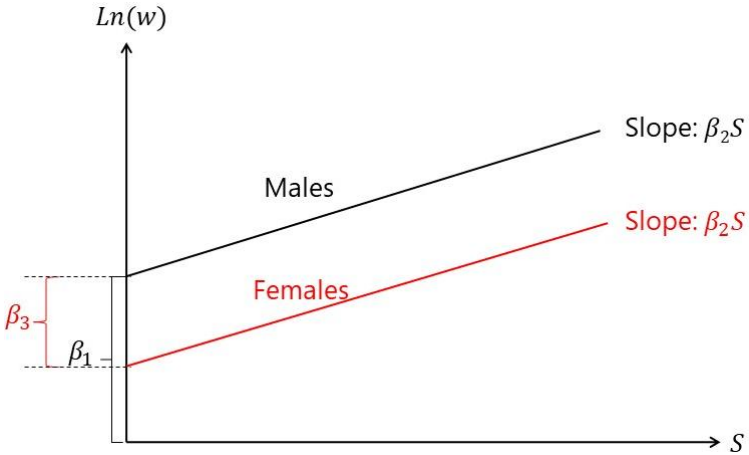
- Dummy variables as explanatory variables shift the intercept of the regression line for certain groups

$$ln(wage) = \beta_1 + \beta_2 S + \beta_3 Female + \varepsilon$$

- For males ($Female = 0$) the regression equation becomes:

- For females ($Female = 1$) the regression equation becomes:

Dummy Variables: Graphical Representation



Dummy Variable Trap: Violation of GM2

- How does the $\boldsymbol{X}$ matrix look like with a dummy variable? e.g.(constant, S, *Female*)

- Now suppose we also include the dummy variable male in this model. How would X look like?

Dummy Variable Trap: Violation of GM2

- If we sum columns 3 and 4 we get a column vector of 1's:

$$Female + Male = constant$$

- This would be an example of perfect multicollinearity -> Violation of GM2 -> OLS could not be estimated
- If we wanted to estimate coefficients on both male and female we would have to omit the constant, i.e. estimate the following model:

$$ln(wage) = \beta_2 S + \beta_3 Female + \beta_4 Male + \varepsilon$$

- Where would be the βs in the figure above?

Interacting Continuous Variables with Dummies

- Sometimes we want to understand how the effect of a certain X variable varies by group characteristics
- This can be achieved by interacting X variables. E.g.

$$ln(wage) = \beta_1 + \beta_2 S + \beta_3 Female + \beta_4 S * Female + \varepsilon$$

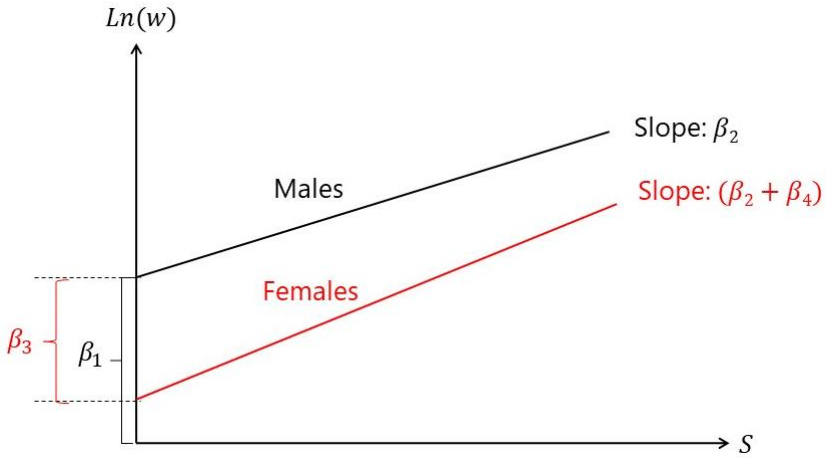
- This does not only estimate different intercepts for each group but also different slopes

Interactions

- For males ($Female = 0$) the regression equation becomes:

- For females ($Female = 1$) the regression equation becomes:

Dummy Variables with Interactions - Graph



Violation of GM3: Omitted Variable Bias

- What happens if you omit a relevant variable from the regression model?
- Suppose the true model is:

- But you estimate the following model:

- In that case the error term is:
- And hence if x_3 is correlated with x_2 we have a violation of GM3 (if $\beta_3 \neq 0$)

Derivation of Omitted Variable Bias Formula

- If we omit x_3 the estimated model is:

- Note: we use $\sim$ instead of $\wedge$ to emphasize that $\tilde{\beta}$ comes from an underspecified model
- The OLS estimator of the underspecified model is:

Derivation Omitted Variable Bias Formula

- Substituting the true model for y:

Derivation Omitted Variable Bias Formula

- We would like to take expectations to see whether the estimator is biased or not
- However both $\sum(x_{2i} - \bar{x}_2)(\varepsilon_i - \bar{\varepsilon})$ and $\sum(x_{2i} - \bar{x}_2)^2$ depend on X and -> take plims

$$= \beta_2 + \beta_3 \frac{\sum(x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3)}{\sum(x_{2i} - \bar{x}_2)^2} + plim\left(\frac{\sum(x_{2i} - \bar{x}_2)(\varepsilon_i - \bar{\varepsilon})}{\sum(x_{2i} - \bar{x}_2)^2}\right)$$

$$= \beta_2 + \beta_3 \frac{\sum(x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3)}{\sum(x_{2i} - \bar{x}_2)^2} + \frac{plim[\sum(x_{2i} - \bar{x}_2)(\varepsilon_i - \bar{\varepsilon})]}{plim[\sum(x_{2i} - \bar{x}_2)^2]}$$

$$= \beta_2 + \beta_3 \frac{\sum(x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3)}{\sum(x_{2i} - \bar{x}_2)^2} + \frac{\sum(x_{2i} - \bar{x}_2) plim[(\varepsilon_i - \bar{\varepsilon})]}{\sum(x_{2i} - \bar{x}_2)^2}$$

$$= \beta_2 + \beta_3 \frac{\sum(x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3)}{\sum(x_{2i} - \bar{x}_2)^2} + 0$$

- The last term is 0 by GM2

Omitted Variable Bias

- The omitted variable bias is therefore:

- This is only equal to β_2 if either:
  - $\beta_3 = 0$
    i.e. x_3 does not belong in the true model
  - $Cov(x_2, x_3) = 0$
    i.e. the omitted variable is uncorrelated with the variable that is included in the model
- The formula does not only tell us whether the coefficient is biased, but also tells us the sign of the bias

Sign of the Omitted Variable Bias

- The sign of the bias depends on the sign of β_3 (which is an unknown parameter but one can make educated guesses about it) and the sign of $Cov(x_2, x_3)$(which is again unknown)

	$Cov(x_2, x_3) > 0$	$Cov(x_2, x_3) < 0$
$\beta_3 > 0$	+	-
$\beta_3 < 0$	-	+

Example Omitted Variable Bias: Ability Bias

- One of the most famous examples of omitted variable bias in economics is the ability bias when estimating returns to education
- More able people could get more schooling and at the same time earn more not because of the additional schooling but just because they are more able:
  - $y =$ earnings
  - $S =$ years of schooling
  - $A =$ ability
- Most datasets do not contain measures of ability; we therefore estimate

Ability Bias

- What is the expected value of $\tilde{\beta}_2$?

- Is this likely positive or negative?
- Under which circumstances would the ability bias be 0?

Solution to Avoid Ability Bias

- The simplest way of solving ability bias would be finding a variable that measures ability
- In many cases that is not possible but one can find a variable that is at least correlated with the omitted variable
- Such variables are called proxy variables
- Suppose the true model has three explanatory variables two of which are observed:

- In the Mincer earnings regression the variables could be:

Proxy Variables

- The explanatory variable x_4^* is unobserved
- But we have a proxy variable x_4 (e.g. IQ score)
- What do we require for this proxy?
- There should be a relationship, e.g.

- Because of the error ν_4, x_4^* and x_4 are not exactly related

Proxy Variables

- Plugging in the equation of the proxy variable into the true model:
- Rearranging gives:
- If ν_4 and ε are uncorrelated with the Xs, $\hat{\beta}_2$ and $\hat{\beta}_3$ will be unbiased

Proxy Variables

- Griliches (1977) illustrates what happens if you get a measure for ability in U.S. data
  1. $ln(y) = \beta_1 + 0.068S + experience$
  2. $ln(y) = \beta_1 + 0.059S + 0.0028IQ + experience$
- Is IQ only a proxy or the ideal measure of ability?

Omitted Variable Bias: Multiple Regressors

- Above we derived the omitted variable bias formula for the case of two explanatory variables, one of which was omitted from the model
- If the true model has more than one regressor the formula would be more complicated
- In general, if there is a correlation of a single x variable and the error term ALL coefficients will be biased
- Suppose the true model is:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

- But we omit x_4 and estimate the model as:

$$\tilde{y} = \tilde{\beta}_1 + \tilde{\beta}_2 x_2 + \tilde{\beta}_3 x_3$$

- The omitted variable bias formula for this model would depend on the correlations of all x variables

Omitted Variable Bias: Multiple Regressors

- Now suppose x_2 is correlated with x_4 but x_3 is not
- Even in that case, both β_2 and β_3 would be biased
- If additionally x_2 and x_3 were uncorrelated the intuition of the omitted variable bias would go through (see also ability bias example above)