

Econometrics II

Fabian Waldinger (LMU Munich)

Lecture Structure

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

Recap from First 2 Weeks of the Course

- GM-assumptions:
 - ① The true model is linear in parameters: $y = \mathbf{X}\beta + \varepsilon$
 - ② No Perfect Collinearity: the matrix \mathbf{X} has rank k
 - ③ Zero Conditional Mean: $E(\varepsilon|\mathbf{X}) = 0$
 - ④ $\text{Var}(\varepsilon|\mathbf{X}) = \sigma^2\mathbf{I}$
 - ⑤ $\varepsilon|\mathbf{X} \sim N(0, \sigma^2\mathbf{I})$
- 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: $y = \mathbf{X}\beta + \varepsilon$
 - ② No Perfect Collinearity: the matrix X has rank k
 - ③ Zero Conditional Mean: $E(\varepsilon|\mathbf{X}) = 0$
 - ④ $Var(\varepsilon|\mathbf{X}) = \sigma^2 I$
 - ⑤ $\varepsilon|\mathbf{X} \sim N(0, \sigma^2 I)$
- 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.:

$$female = \begin{cases} 1 & \text{if woman} \\ 0 & \text{if man} \end{cases}$$

- 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 X s

Dummy Variables - Example Gender Pay Gap

a supporter | subscribe | search | jobs | dating | more ▾ | UK edition ▾

theguardian

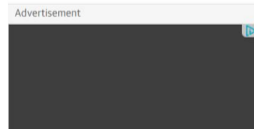
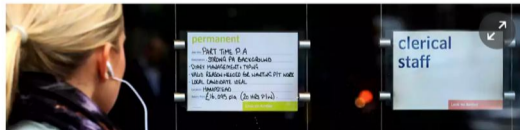
website of the year

sport | football | opinion | culture | business | lifestyle | fashion | environment | tech | travel | **≡ browse all sections**

property | savings | pensions | borrowing

UK women still far adrift on salary and promotion as gender pay gap remains a gulf

IFS research shows average difference in pay is 18% and widens markedly after women have children



Example. Data from the ACS 2015

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

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

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

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

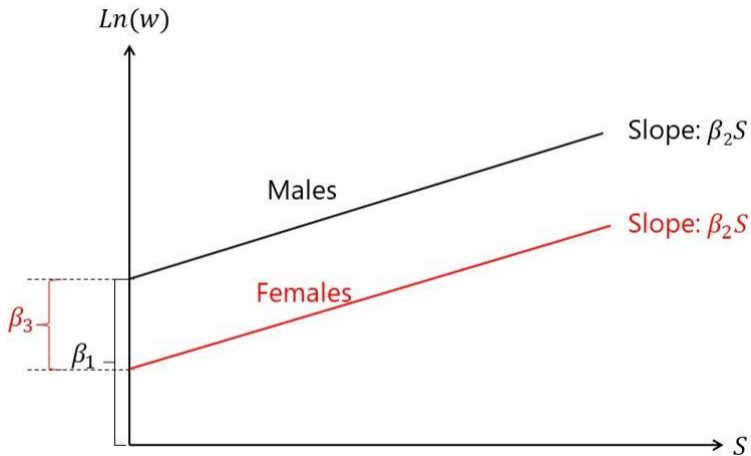
- For males ($Female = 0$) the regression equation becomes:

$$E[\ln(wage)|Female = 0, S] = \beta_1 + \beta_2 S + \varepsilon$$

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

$$\begin{aligned} E[\ln(wage)|Female = 1, S] &= \beta_1 + \beta_2 S + \beta_3 1 + \varepsilon \\ &= (\beta_1 + \beta_3) + \beta_2 S + \varepsilon \end{aligned}$$

Dummy Variables: Graphical Representation



Dummy Variable Trap: Violation of GM2

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

$$\mathbf{X} = \begin{pmatrix} 1 & 15 & 1 \\ 1 & 12 & 0 \\ 1 & 10 & 0 \\ 1 & 12 & 1 \\ 1 & 18 & 0 \end{pmatrix}$$

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

$$\mathbf{X} = \begin{pmatrix} 1 & 15 & 1 & 0 \\ 1 & 12 & 0 & 1 \\ 1 & 10 & 0 & 1 \\ 1 & 12 & 1 & 0 \\ 1 & 18 & 0 & 1 \end{pmatrix}$$

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 β s 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.

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

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

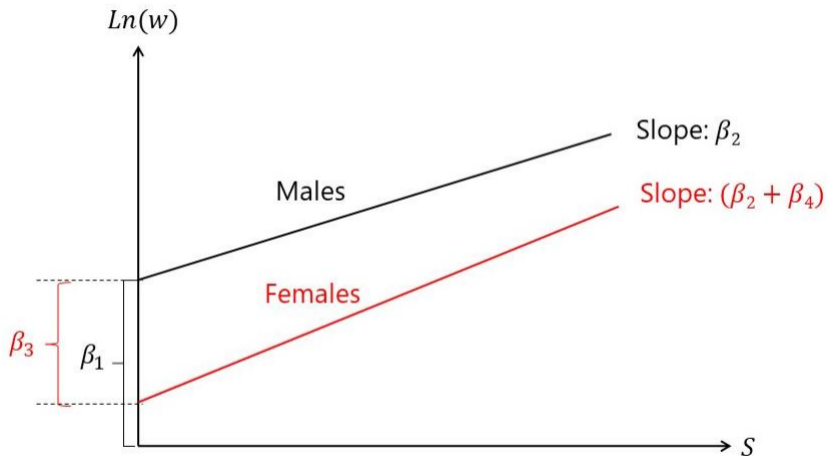
- For males ($Female = 0$) the regression equation becomes:

$$E[\ln(wage)|Female = 0, S] = \beta_1 + \beta_2 S + \varepsilon$$

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

$$\begin{aligned} E[\ln(wage)|Female = 1, S] &= \beta_1 + \beta_2 S + \beta_3 \mathbf{1} + \beta_4 S * \mathbf{1} + \varepsilon \\ &= (\beta_1 + \beta_3) + (\beta_2 + \beta_4) S + \varepsilon \end{aligned}$$

Dummy Variables with Interactions - Graph



Omitted Variable Bias

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

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

- But you estimate the following model:

$$y = \beta_1 + \beta_2 x_2 + u$$

- In that case the error term is: $u = \beta_3 x_3 + \varepsilon$
- 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:

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

- 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:

$$\tilde{\beta}_2 = \frac{\text{Cov}(x_2, y)}{\text{Var}(x_2)} = \frac{\sum (x_{2i} - \bar{x}_2)(y_i - \bar{y})}{\sum (x_{2i} - \bar{x}_2)^2}$$

Derivation Omitted Variable Bias Formula

- Substituting the true model for y :

$$\begin{aligned} &= \frac{\sum (x_{2i} - \bar{x}_2) [(\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i) - (\beta_1 + \beta_2 \bar{x}_2 + \beta_3 \bar{x}_3 + \bar{\varepsilon})]}{\sum (x_{2i} - \bar{x}_2)^2} \\ &= \frac{\sum \beta_2 (x_{2i} - \bar{x}_2)^2 + \beta_3 (x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3) + (x_{2i} - \bar{x}_2)(\varepsilon_i - \bar{\varepsilon})}{\sum (x_{2i} - \bar{x}_2)^2} \\ &= \beta_2 + \beta_3 \frac{\sum (x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3)}{\sum (x_{2i} - \bar{x}_2)^2} + \frac{\sum (x_{2i} - \bar{x}_2)(\varepsilon_i - \bar{\varepsilon})}{\sum (x_{2i} - \bar{x}_2)^2} \end{aligned}$$

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 \rightarrow have to take plims

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

$$= \beta_2 + \beta_3 \frac{\sum(x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3)}{\sum(x_{2i} - \bar{x}_2)^2} + \frac{\text{plim}[\frac{1}{N} \sum(x_{2i} - \bar{x}_2)(\varepsilon_i - \bar{\varepsilon})]}{\text{plim}[\frac{1}{N} \sum(x_{2i} - \bar{x}_2)^2]}$$

$$= \beta_2 + \beta_3 \frac{\sum(x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3)}{\sum(x_{2i} - \bar{x}_2)^2} + \frac{\sum(x_{2i} - \bar{x}_2) \text{plim}[\frac{1}{N}(\varepsilon_i - \bar{\varepsilon})]}{\sum(x_{2i} - \bar{x}_2)^2}$$

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

Omitted Variable Bias

- The omitted variable bias is therefore:

$$E(\tilde{\beta}_2) = \beta_2 + \beta_3 \frac{\sum (x_{2i} - \bar{x}_2)(x_{3i} - \bar{x}_3)}{\sum (x_{2i} - \bar{x}_2)^2} = \beta_2 + \beta_3 \frac{\text{Cov}(x_2, x_3)}{\text{Var}(x_2)}$$

- This is only equal to β_2 if either:
 - $\beta_3 = 0$
i.e. x_3 does not belong in the true model
 - $\text{Cov}(x_2, x_3) = 0$
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:

$$\log(y) = \beta_1 + \beta_2 S + \beta_3 A + \varepsilon$$

- y = earnings
- S = years of schooling
- A = ability
- Most datasets do not contain measures of ability; we therefore estimate

$$\log(y) = \beta_1 + \beta_2 S + \varepsilon$$

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

$$E(\tilde{\beta}_2|\mathbf{X}) = \beta_2 + \beta_3 \frac{\text{Cov}(S, A)}{\text{Var}(S)}$$

- 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:

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

- In the Mincer earnings regression the variables could be:

$$\ln(y) = \beta_1 + \beta_2 S + \beta_3 Exp + \beta_4 A + \varepsilon$$

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.

$$x_4^* = \delta_1 + \delta_4 x_4 + \nu_4$$

- 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:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 (\delta_1 + \delta_4 x_4 + \nu_4) + \varepsilon$$

- Rearranging gives:

$$y = (\beta_1 + \beta_4 \delta_1) + \beta_2 x_2 + \beta_3 x_3 + \beta_4 \delta_4 x_4 + \underbrace{\beta_4 \nu_4 + \varepsilon}_e$$

- If ν_4 and ε are uncorrelated with the X s, $\hat{\beta}_2$ and $\hat{\beta}_3$ will be unbiased

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