Endogenous Regressors and Instrumental Variables

Ketevani Kapanadze

Brno, 2020

Non-Linear Specification

- There is not always a linear relationship between dependent variable and explanatory variables:
  - The use of OLS requires that the model be linear in parameters!
  - There is a wide variety of functional forms that are linear in coefficients while being non-linear in variables
- We have to choose carefully the functional form of the relationship between the dependent variable and each explanatory variable:
  - The choice of a functional form should be based on the underlying economic theory and/or intuition;
  - Do we expect a curve instead of a straight line? Does the effect of a variable peak at some point and then start to decline?

Linear Form

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Assumes that the effect of the explanatory variable on the dependent variable is constant;
- Interpretation: if X_i increases by 1 unit (in which X_i is measured), then Y_i will change by β_1 units (in which Y_i is measured)
- The linear form is used as a default functional form until strong evidence that it is inappropriate is found.

Log-log Form

$$lnY_i = \beta_0 + \beta_1 lnX1_i + \beta_2 lnX2_i + u_i$$

- Assumes that the elasticity of the dependent variable with respect to the explanatory variable is constant;
- Interpretation: if X_k increases by 1%, then Y_i will change by β_1 %;
- Before using a log-log model, make sure that there are no negative or zero observations in the data set!

Log-log Form

- Example: $$\widehat{\ln Q} = 2.70 + \underset{(0.14)}{0.59} \ln L + \underset{(0.17)}{0.33} \ln K$$

$$\begin{array}{lll} Q & \ldots & \text{output} \\ L & \ldots & \text{labor} \\ K & \ldots & \text{capital employed} \end{array}$$

- Interpretation: if we increase the amount of labor by 1%, the production of sugar will increase by 0.59%, ceteris paribus.

Log-linear Form

- Linear log form: $Y_i = \beta_0 + \beta_1 lnX1_i + \beta_2 lnX2_i + u_i$
  - Interpretation: if X_k increases by 1 %, then Y_i will increase by β_k /100 units (k = 1; 2);
- Log linear form: $lnY_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + u_i$
  - Interpretation: if X_k increases by 1 unit, then Y_i will change by β_k*100 %.

Log-linear Form

- Example: $$\underset{}{\widehat{Y}} = -6.94 - \underset{(0.19)}{0.57}\ PC + \underset{(0.11)}{0.25}\ PB + \underset{(2.81)}{12.2}\ \ln YD$$

Y	...	annual chicken consumption (kg.)
PC	...	price of chicken
PB	...	price of beef
YD	...	annual disposable income

- Interpretation: An increase in the annual disposable income by 1% increases chicken consumption by 0.12 kg per year, ceteris paribus.

Log-linear Form

- Example:

$$\widehat{\ln wage} = \underset{(0.008)}{0.217 + 0.098}\ educ + \underset{(0.002)}{0.010}\ exper$$

wage	...	annual wage (USD)
educ	...	years of education
exper	...	years of experience

- Interpretation: An increase in education by one year increases annual wage by 9.8%, ceteris paribus. An increase in experience by one year increases annual wage by 1%, ceteris paribus.

Polynomial Form

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i$$

- To calculate the effect of X_i on Y_i, we need to calculate the derivative;
- Clearly, the effect of X_i on Y_i is not constant, rather it changes with the level of X_i;
- We might also have higher order polynomials;

Choice of correct functional form

- The functional form has to be correctly specified in order to avoid biased estimates:
  - One of the OLS assumptions is that the model is correctly specified!
- Ideally, the specification is given by underlying theory of the eq.;
- In reality, theory does not give precise functional form;
- In most cases, either linear form is adequate, or common sense will point out an easy choice from among the alternatives

Choice of correct functional form

- Nonlinearity of explanatory variables:
  - Often approximated by polynomial form;
  - Missing higher powers of a variable can be detected as ommited variables;
- Nonlinearity of dependent variable:
  - Harder to detect base on statistical fit of the regression;
  - R-squared is incomparable across models where Y is transformed!
  - Dependent variables are often transformed to log-form in order to make their distribution closer to the normal distribution.

Dummy Variables

- Dummy variable - takes on the values of 0 or 1, depending on a qualitative attribute;
- Examples of dummy variables are:

$$Male = \begin{cases} 1 & \text{if the person is male} \\ 0 & \text{if the person is female} \end{cases}$$

$$Weekend = \begin{cases} 1 & \text{if the day is on weekend} \\ 0 & \text{if the day is a work day} \end{cases}$$

$$NewStadium = \begin{cases} 1 & \text{if the team plays on new stadium} \\ 0 & \text{if the team plays on old stadium} \end{cases}$$

Intercept Dummy

- Dummy variable included in a regression alone (not interacted with other variables) is an intercept dummy;
- It changes the intercept for the subset of data defined by a dummy variable condition:

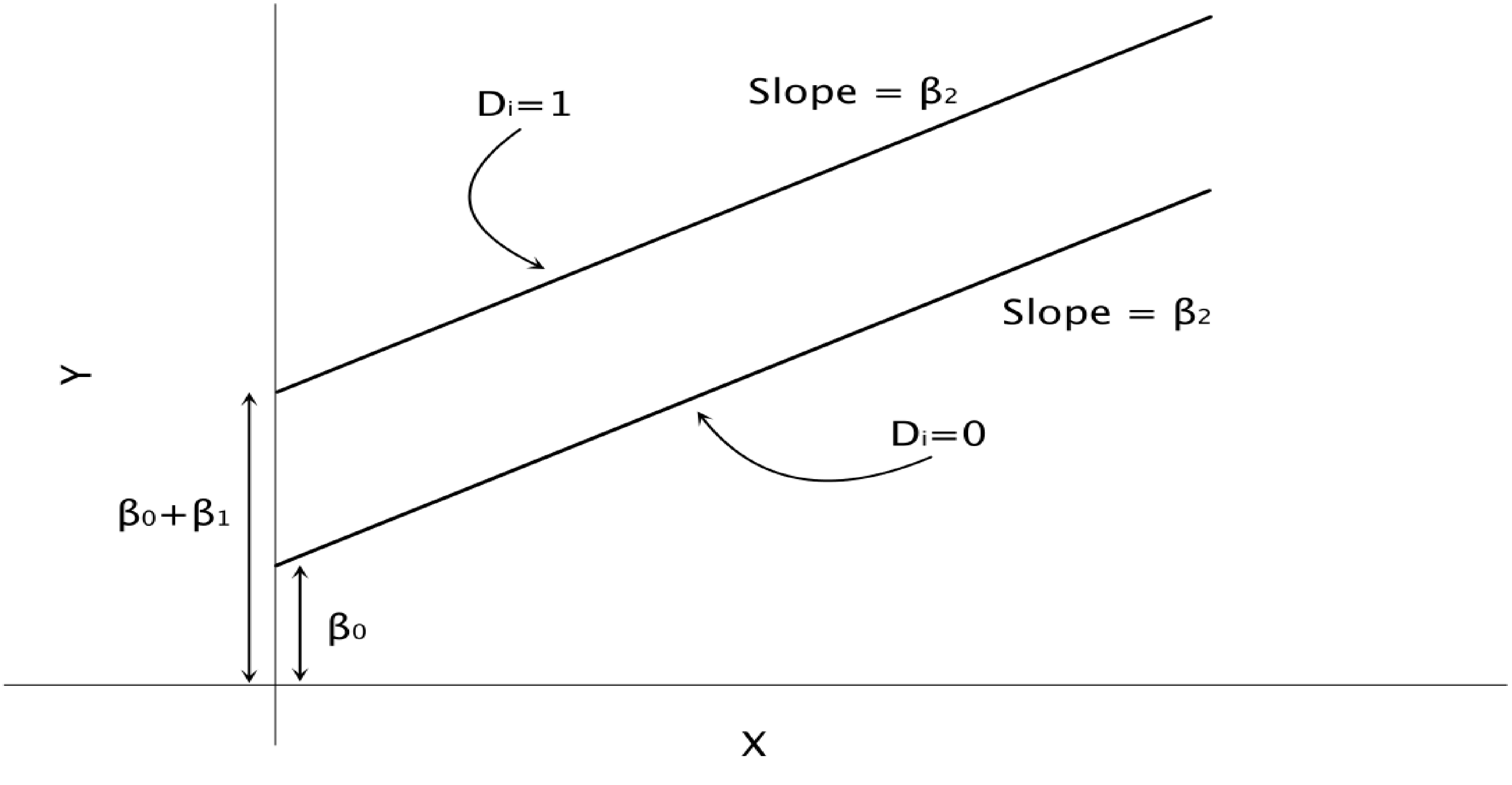
$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_\mathrm{i} + u_i$$

where

$$D_i = \begin{cases} 1 & \text{if the } i\text{-th observation meets a particular condition} \\ 0 & \text{otherwise} \end{cases}$$

- We have: (on the board)

Intercept Dummy



Example

- Estimating the determinant of wages:

$$wage_i = -3.89 + \underset{(0.270)}{2.156}\, M_i + \underset{(0.051)}{0.603}\, educ_i + \underset{(0.064)}{0.010}\, exper_i$$

- Interpretation of the dummy variable M: men earn on average $2.156 per hour more than women, ceteris paribus

Slope Dummy

- If a dummy variable is interacted with another variable (x), it is a slope dummy;
- It changes the relationship between x and y for a subset of data defined by a dummy variable condition:

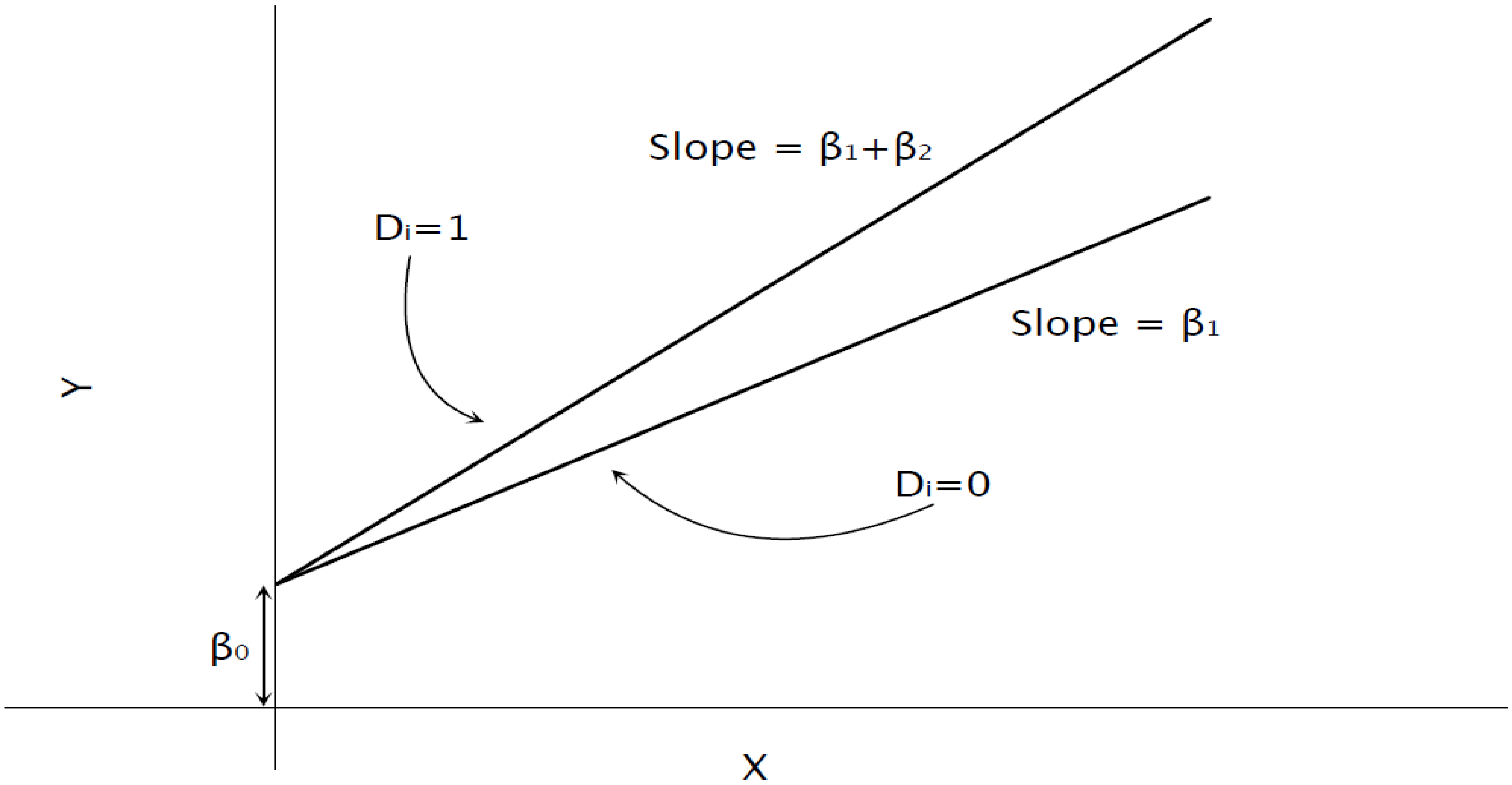
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i * D_i) + u_i$$

where

$$D_i = \begin{cases} 1 & \text{if the } i\text{-th observation meets a particular condition} \\ 0 & \text{otherwise} \end{cases}$$

- We have: (on the board)

Slope Dummy



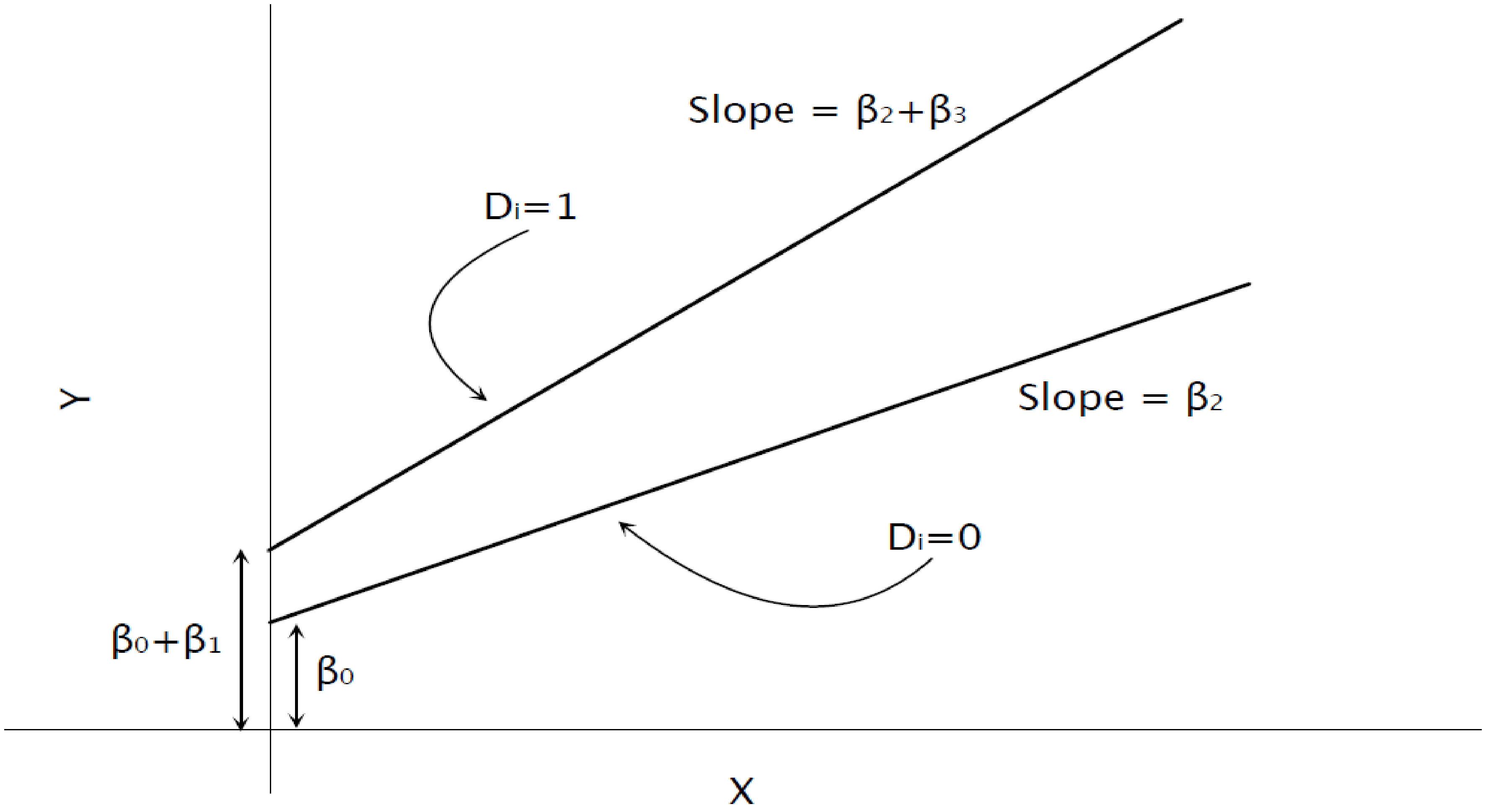
Example

- Estimating the determinant of wages:

$$\underset{}{wage_i} = -2.620 + \underset{(0.054)}{0.450}\ educ_{\mathrm{i}} + \underset{(0.021)}{0.17}\ M_i * educ_{\mathrm{i}} + \underset{(0.065)}{0.010}\ exper_{\mathrm{i}}$$

- Interpretation: men gain on average 17 cents per hour more than women for each additional year of education, ceteris paribus

Slope and intercept Dummies



Multiple categories

- What if a variable defines three or more qualitative attributes?
- Example: level of education - elementary school, high school, and college;
- Define and use a set of dummy variables:

$$H = \begin{cases} 1 & \text{if high school} \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad C = \begin{cases} 1 & \text{if college} \\ 0 & \text{otherwise} \end{cases}$$

- Should we include also a third dummy in the regression, which is equal to 1 for people with elementary education?
  - No, unless we exclude the intercept!
  - Using full set of dummies leads to perfect multicollinearity (dummy variable trap)

Endogeneity Problem

- An ***endogenous*** variable is one that is correlated with u
- An ***exogenous*** variable is one that is uncorrelated with u

- In IV regression, we focus on the case that X is endogenous and there is an instrument, Z, which is exogenous.

Digression on terminology: "Endogenous" literally means "determined within the system." If X is jointly determined with Y, then a regression of Y on X is subject to simultaneous causality bias. But this definition of endogeneity is too narrow because IV regression can be used to address OV bias and errors-in-variable bias. Thus we use the broader definition of endogeneity above.

Endogeneity Problem

- *Omitted variable bias* from a variable that is correlated with X but is unobserved and for which there are inadequate control variables (LS 2 p. 17);
- *Measurement error bias* (X is measured with error)
- *Simultaneous causality bias* (X causes Y, Y causes X);

All three problems cause X to be **endogenous,** $E(u|X) \neq 0$

Endogeneity Problem

- **The endogeneity problem is endemic in social sciences/economics**
  - In many cases important personal variables cannot be observed (examples?)
  - These are often correlated with observed explanatory information
  - In addition, measurement error may also lead to endogeneity
  - Solutions to endogeneity problems:
    - *Proxy variables method for omitted regressors*
    - *Fixed effects methods if: 1) panel data is available, 2) endogeneity is time-constant, and 3) regressors are not time-constant*
- **<u>Instrumental variables method (IV)</u>**
  - IV is the most well-known method to address endogeneity problems

Instrumental Variables

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- IV regression breaks X into two parts: a part that might be correlated with u, and a part that is not. By isolating the part that is not correlated with u, it is possible to estimate β_1.
- This is done using an ***instrumental variable***, Z_i, which is correlated with X_i but uncorrelated with u_i.

Instrumental Variables

- **Example: Education in a wage equation**

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + u_i$$

Error terms contains factors (such as innate ability) which are correlated with education

- **Definition of a instrumental variable:**
  - 1) It does not appear in the regression (why?)
  - 2) It is highly correlated with the endogenous variable
  - 3) It is uncorrelated with the error term
- **Reconsideration of OLS in a simple regression model**

$y_i = \beta_0 + \beta_1 x_i + u_i$ <u>and assume</u> $Cov(x_i, u_i) = 0$

Instrumental Variables

- **Example: Father's education as an IV for education**

OLS: $\widehat{\log}(wage) = -\underset{(.185)}{.185} + \underset{(.014)}{.109}\, educ$

Return to education probably overestimated

$n = 428, R^2 = .118$

Is the education of the father a good IV?

1) It doesn't appear as regressor
2) It is significantly correlated with educ
3) It is uncorrelated with the error (?)

$\widehat{educ} = -\underset{(.28)}{10.24} + \underset{(.029)}{.269}\, fatheduc$

$n = 428, R^2 = .173$

IV: $\widehat{\log}(wage) = \underset{(.446)}{.441} + \underset{(.035)}{.059}\, educ$

The estimated return to education decreases (which is to be expected)

It is also much less precisely estimated

$n = 428, R^2 = 1 - RSS_{IV}/TSS = .093$

Instrumental Variables

- **Other IVs for education that have been used in the literature:**
- The number of siblings
  - 1) No wage determinant, 2) Correlated with education because of resource constraints in hh, 3) Uncorrelated with innate ability
- College proximity when 18 years old
  - 1) No wage determinant, 2) Correlated with education because more education if lived near college, 3) Uncorrelated with error (?)
- Month of birth
  - 1) No wage determinant, 2) Correlated with education because of compulsory school attendance laws, 3) Uncorrelated with error

Instrumental Variables

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

For an instrumental variable (an "***instrument***") ***Z*** to be valid, it must satisfy the following conditions:

1. Does not appear in the regression

2. Instrument relevance: $\text{corr}(Z_i, X_i) \neq 0$

3. Instrument exogeneity: $\text{corr}(Z_i, u_i) = 0$

- **Example: effect of skipping classes on final exam score**
- Sign and magnitude of the instrument!

Instrumental Variables

- **Properties of IV with a poor instrumental variable**
  - IV may be much more inconsistent than OLS if the instrumental variable is not completely exogenous and only weakly related to x

$$plim\ \widehat{\beta}_{1,OLS} = \beta_1 + Corr(x,u) \cdot \frac{\sigma_u}{\sigma_x}$$

$$plim\ \widehat{\beta}_{1,IV} = \beta_1 + \frac{\boxed{Corr(z,u)}}{Corr(z,x)} \cdot \frac{\sigma_u}{\sigma_x}$$

There is no problem if the instrumental variable is really exogenous. If not, the asymptotic bias will be the larger the weaker the correlation with x.

IV worse than OLS if: $\frac{Corr(z,u)}{Corr(z,x)} > Corr(x,u)$ e.g. $\frac{0.03}{0.2} > 0.1$

  - **Variance of IV estimator is always (!) greater than variance of OLS estimator!**

Instrumental Variables

- **IV estimation in the multiple regression model**

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 \boxed{z_1} + \cdots + \beta_k \boxed{z_{k-1}} + u_1$$

endogenous

exogenous variables

In a regression of the endogenous explanatory variable on all exogenous variables, the instrumental variable must have a non-zero coefficient.

- **Conditions for instrumental variable**
  - 1) Does not appear in regression equation
  - 2) Is uncorrelated with error term
  - 3) Is partially correlated with endogenous explanatory variable

$$y_2 = \pi_0 + \pi_1 z_1 + \cdots + \pi_k z_{k-1} + \pi_k z_k + v_2$$

This is the so called „reduced form regression“

Two Stage Least Squares: 2SLS

As it sounds, TSLS has two stages – two regressions:

1. Isolate the part of X that is uncorrelated with u by regressing X on Z using OLS:

$$X_i = \pi_0 + \pi_1 Z_i + v_i \qquad (1)$$

- Because Z_i is uncorrelated with u_i, $\pi_0 + \pi_1 Z_i$ is uncorrelated with u_i. We don't know π_0 or π_1 but we have estimated them, so…
- Compute the predicted values of X_i,

2. Replace X_i by $\hat{X}_i$ in the regression of interest:

regress Y on $\hat{X}_i$ using OLS:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i \qquad (2)$$

Two Stage Least Squares: 2SLS

- **Because $\hat{X}_i$ is uncorrelated with u_i, the first least squares assumption holds for regression (2).** (This requires *n* to be large so that π_0 and π_1 are precisely estimated.)

- Thus, in large samples, β_1 can be estimated by OLS using regression (2)

- The resulting estimator is called the *Two Stage Least Squares (TSLS)* estimator, $\hat{\beta}_1^{TSLS}$.

Two Stage Least Squares: 2SLS

Suppose Z_i, satisfies the two conditions for a valid instrument:

1. Instrument relevance: corr(Z_i,X_i) ≠ 0

2. Instrument exogeneity: corr(Z_i,u_i) = 0

<u>Two-stage least squares:</u>

Stage 1: Regress X_i on Z_i (including an intercept), obtain the predicted values, $\hat{X}_i$

Stage 2: Regress Y_i on $\hat{X}_i$ (including an intercept); the coefficient on $\hat{X}_i$ is the TSLS estimator, $\hat{\beta}_1^{TSLS}$.

$\hat{\beta}_1^{TSLS}$ is a consistent estimator of β_1.

Two Stage Least Squares: 2SLS

- **Why does Two Stage Least Squares work?**
  - All variables in the second stage regression are exogenous because endogenous variable has been replaced by a prediction based on only exogenous information;
  - By using the prediction based on exogenous information, endog. variable is purged of its endogenous part (the part that is related to the error term)
- **Properties of Two Stage Least Squares**
  - The standard errors from the OLS second stage regression are wrong. However, it is not difficult to compute correct standard errors.
  - If there is one endogenous variable and one instrument then 2SLS = IV
  - The 2SLS estimation can also be used if there is more than one endogenous variable and at least as many instruments

Two Stage Least Squares: 2SLS

- **Example: 2SLS in a wage equation using two instruments**

First stage regression (regress educ on all exogenous variables):

$$\widehat{educ} = \underset{(.27)}{8.37} + \underset{(.026)}{.085}\ exper - \underset{(.001)}{.002}\ exper^2 + \underset{(.024)}{.185}\ \boxed{fatheduc} + \underset{(.026)}{.186}\ \boxed{motheduc}$$

Education is significantly partially correlated with the education of the parents

Two Stage Least Squares estimation results:

$$\widehat{log}(wage) = \underset{(.400)}{0.48} + \boxed{\underset{(.031)}{.061}}\ educ + \underset{(.013)}{.044}\ exper - \underset{(.0004)}{.0009}\ exper^2$$

$$n = 428, R^2 = 0.136$$

The return to education is much lower but also much more imprecise than with OLS

Two Stage Least Squares: 2SLS

- **Statistical properties of 2SLS/IV-estimation**
  - Under assumptions completely analogous to OLS, but conditioning on $\mathbf{z}_i$ rather than on $\mathbf{x}_i$ 2SLS/IV is consistent and asymptotically normal
  - 2SLS/IV is typically much less precise because there is more multicollinearity and less explanatory variation in the second stage regression
  - Corrections for heteroscedasticity analogous to OLS
  - 2SLS/IV easily extends to time series and panel data situations

Next Class

- **Qualitative and Limited Dependent Variable Models**

20.03.2020