Elements of Econometrics.
Lecture 29.
Revision 3

FCS, 2022-2023

Model B vs Model A

- In the Model A we assumed:
  - X is nonstochastic
    - Assume values of regressors are nonrandom (**<u>in fact</u>**, no endogeneity)
- In the Model B:
  - X is stochastic
    - Realization of random variable

Why we need Model A at all?

- Analytical simplicity

Why we need Model B?

- More realistic

Model B vs Model A

A.1 The model is linear in parameters and correctly specified.

$$Y = \beta_1 + \beta_2 X + u$$

A.2 There is some variation in the regressor in the sample and no exact linear relationship bw regressors in the sample.

A.3(G-M 1)The disturbance term has zero expected value in each observation: $E(u_i) = 0$

A.4 (G-M 2)The disturbance term is homoscedastic

$$\sigma^2_{u_i} = \sigma^2_u$$

A.5(G-M 3)The values of the disturbance term have independent distributions (u_i and u_j are independent for all $j \neq i$)

$$\sigma_{u_i u_j} = E[(u_i - \mu_u)(u_j - \mu_u)] = E(u_i u_j)$$

$$= E(u_i)E(u_j) = 0$$

A.6 The disturbance term has a normal distribution

B.1 The model is linear in parameters and correctly specified.(A.1)

B.2 Values of regressors are drawn randomly from fixed populations. New!

B.3 No exact linear relationship among the regressors.(A.2)

B.4 (G-M 1) The disturbance term has zero expectation.(A.3)

B.5 (G-M2) The disturbance term is homoscedastic.(A.4)

B.6 (G-M 3) The values of the disturbance term have independent distributions.(A.5)

B.7 (G-M 4) The disturbance term is distributed independently of regressors for each i (in addition, assume cross-section independency bw observations). New!

B.8 The disturbance term has a normal distribution.(A.6)

UNDER MODEL B ASSUMPTIONS OLS GIVES BLUE and CONSISTENT ESTIMATES

Model B: Consistency

$$\boldsymbol{Y_i = \beta_1 + \beta_2 X_i + u_i}$$

$$\hat{\beta}_2 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \beta_2 + \frac{\sum (X_i - \bar{X})(u_i - \bar{u})}{\sum (X_i - \bar{X})^2} = \beta_2 + \sum a_i u_i$$

$$\operatorname{plim} \hat{\beta}_2 = \beta_2 + \operatorname{plim}\left(\frac{\sum (X_i - \bar{X})(u_i - \bar{u})}{\sum (X_i - \bar{X})^2} \right) = \beta_2 + \operatorname{plim}\left(\frac{\frac{1}{n}\sum (X_i - \bar{X})(u_i - \bar{u})}{\frac{1}{n}\sum (X_i - \bar{X})^2} \right)$$

$$= \beta_2 + \frac{\operatorname{plim}\left(\frac{1}{n}\sum (X_i - \bar{X})(u_i - \bar{u}) \right)}{\operatorname{plim}\left(\frac{1}{n}\sum (X_i - \bar{X})^2 \right)} = \beta_2 + \frac{\operatorname{cov}(X, u)}{\operatorname{var}(X)} = \beta_2$$

Violation of Model B Assumptions

B.1 The model is linear in parameters and correctly specified.(A.1)

Same as Model A

- Violation 1: omitted variable
- Violation 2: irrelevant variable

B.2 Values of regressors are drawn randomly from fixed populations.

B.3 No exact linear relationship among the regressors. (A.2)

Same as Model A: see Review lecture 1 (perfect multicollenearity)

B.4 (G-M 1) The disturbance term has zero expectation.A.3)

Same as Model A (satisfied if intercept included)

B.5 (G-M2) The disturbance term is homoscedastic.(A.4)

Violation 3: heteroscedasticity

B.6 (G-M 3) The values of the disturbance term have independent distributions.(A.5)

Same as Model A (satisfied for cross-section data)

B.7 (G-M 4) The disturbance term is distributed independently of regressors. New!

Violation 5: Endogeneity ($cov(X,u) \neq 0$)

B.8 The disturbance term has a normal distribution.(A.6)

Violation of B.7 assumption: measurement errors in *X*

- Definition:

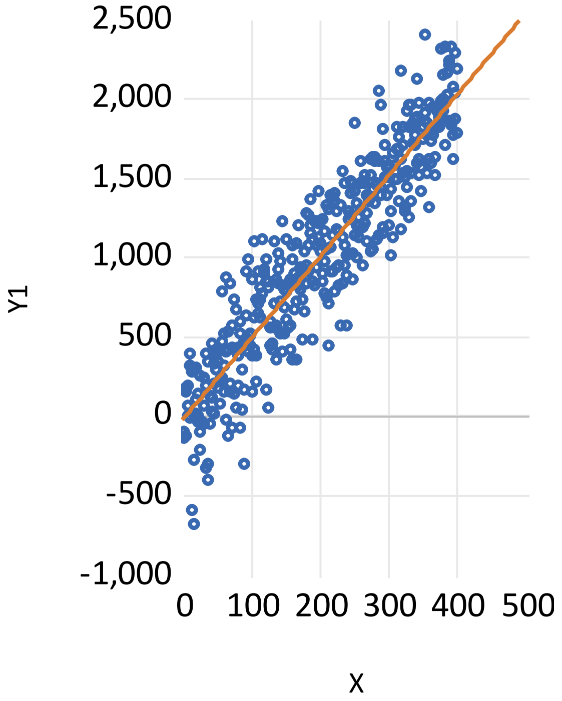
True model: $Y_i = \beta_1 + \beta_2 Z_i + v_i$
Observe regressor with measurement error: $X_i = Z_i + w_i$

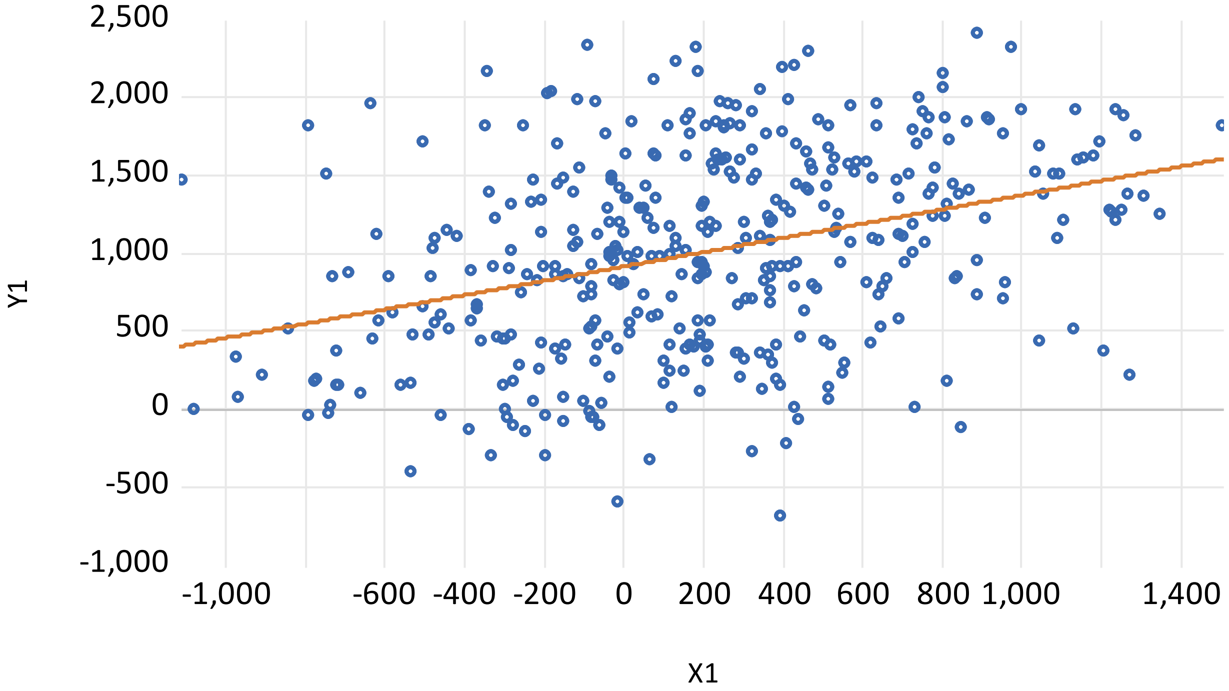
$$\begin{aligned} Y_i &= \beta_1 + \beta_2 (X_i - w_i) + v_i \\ &= \beta_1 + \beta_2 X_i + v_i - \beta_2 w_i \\ &= \beta_1 + \beta_2 X_i + u_i \end{aligned}$$

Consider covariance between X and u:

$$\begin{aligned} Cov(X, u) &= Cov(Z + w, v - \beta_2 w) \\ &= Cov(Z, v) - \beta_2 Cov(Z, w) + Cov(w, v) - \beta_2 Cov(w, w) \\ &= -\beta_2 Cov(w, w) = -\beta_2 \sigma_w^2 \end{aligned}$$

MEASUREMENT ERRORS IN EXPLANATORY VARIABLE: GRAPHICAL ILLUSTRATION.





Y1=10+X*5+250*nrnd
X1=X+500*nrnd

Measurement errors

- Consequences:

Variance of regressor with measurement error:

$$
\begin{aligned}
Var(X) &= Var(Z + w) \\
&= Var(Z) + Var(w) + 2Cov(Z, w) \\
&= \sigma_Z^2 + \sigma_w^2
\end{aligned}
$$

Probability limit of OLS estimator b_2:

$$plim(b_2) = \beta_2 + \frac{cov(X, u)}{Var(X)} = \beta_2 - \beta_2 \frac{\sigma_w^2}{\sigma_Z^2 + \sigma_w^2}$$

$\rightarrow$ OLS estimator b_2 inconsistent

-Inconsistent estimator

Measurement errors in Y

True model: $Q_i = \beta_1 + \beta_2 X_i + v_i$
Observe dependent variable with with measurement error: $Y_i = Q_i + r_i$

$$Y_i - r_i = \beta_1 + \beta_2 X_i + v_i$$
$$Y_i = \beta_1 + \beta_2 X_i + v_i + r_i$$
$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

- Consequences:
- B.7 holds
- S.E. valid but higher (Lower precision of estimates)
- <u>Inefficiency</u>

How to overcome errors in X and Y:
1) Use precise data
2) Use instruments to get consistent estimates in case of measurement errors in X

Simultaneous causality

- In our standard model we assume that *X* influence *Y* but sometimes (virtually very often) *Y* can influence X too (for example, drinking and health).
- Thus in model $Y_i = \alpha + \beta X_i + u_i$ with simultaneous causality we need to add equation with dependence of *X* on *Y*.
- We can estimate system of equation or use instruments. In fact, these approaches are similar

Endogeneity

- Why do we talk so much about endogeneity?

  If we have endogeneity we lose consistency, thus our models are incorrect and useless.
- There are 5 causes of endogeneity (and inconsistency):
  - Omitted variables
  - Incorrect functional form
  - Measurement errors in X
  - Simultaneous causality
  - Sample bias (we did not discuss this point before, but if our sample in not representative we call it biased)
- How to overcome problems with endogeneity?
  - General approach in practice – use instruments. Instrument – variable which is correlated with endogeneous regressor, but is uncorrelated with error term (so it is exogenous).

DURBIN–WU–HAUSMAN TEST. TESTING FOR RELATIONSHIP OF EXPLANATORY VARIABLES AND DISTURBANCE TERM.

H_0: Assumption B.7 is valid (The disturbance term is distributed independently of the regressors).

Estimator b - consistent under H_0 and H_1

Estimator B - inconsistent under H_1, efficient under H_0

The IV estimator b is consistent under both the H_0 and H_1

The OLS estimator B is consistent (and unbiased), and more efficient than the IV estimator under the H_0, but it is inconsistent under H_1.

Test: H_0: difference in coefficients is not systematic

$$\chi^2(k) = (b-B)'[(V\_b-V\_B)^{-1}](b-B)$$

(Here V_b, V_B – estimated covariance matrices).

Under the null hypothesis, the test statistic has a *χ2* (chi-squared) distribution with degrees of freedom equal to the number of regressors tested for endogeneity.

Two Stage Least Squares (TSLS)

- At the first step estimate equation (OLS) for *X* using instruments, $X = \gamma_1 + \gamma_2 Z + exogenous\ controls + v$, and get predicted values of *X*, $\hat{X}$
- At the second step estimate (OLS) the main model

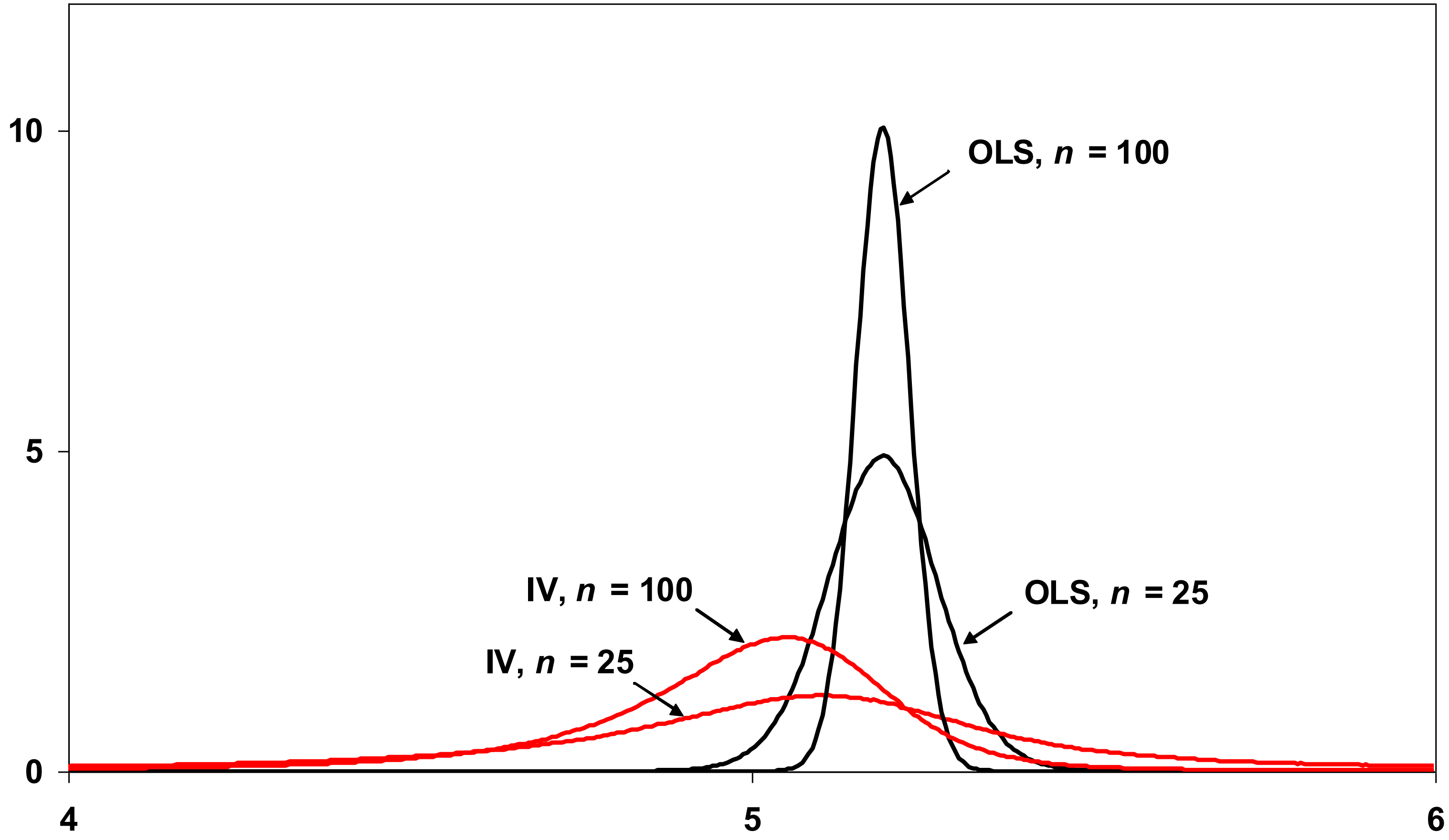
$$Y = \beta_1 + \beta_2 \hat{X} + exogenous\ controls + u$$

- TSLS estimates in the case of relevant instruments are consistent.
- There could be more than one endogenous regressor and instrument (IV is a special case for TSLS, when there are 1 regressor and 1 instrument)

FINITE-SAMPLE DISTRIBUTIONS OF THE IV ESTIMATOR: MONTE CARLO EXPERIMENT

$$Y = \beta_1 + \beta_2 X + u \qquad Y = 10 + 5X + u$$

$$X = \lambda_1 Z + \lambda_2 V + u \qquad X = 0.5Z + 2.0V + u$$



The diagram shows the distributions of the OLS and IV estimators of β_2 for n = 25 and n = 100, for 10 million samples in both cases. In this case $\text{plim}\,\hat{\beta}_2^{\text{OLS}} = 5.19$ **, and** $\text{plim}\,\hat{\beta}_2^{\text{IV}} = 5.00$

The General Instrumental Variables Regression Model and Terminology

The general IV regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i, \tag{10.12}$$

$i = 1, \ldots, n$, where:

- Y_i is the dependent variable;
- u_i is the error term, which represents measurement error and/or omitted factors;
- $X_{1i}, \ldots, X_{ki}$ are k endogenous regressors, which are potentially correlated with u_i;
- $W_{1i}, \ldots, W_{ri}$ are r included exogenous regressors, which are uncorrelated with u_i;
- $\beta_0, \beta_1, \ldots, \beta_{k+r}$ are unknown regression coefficients;
- $Z_{1i}, \ldots, Z_{mi}$ are m instrumental variables.

The coefficients are overidentified if there are more instruments than endogenous regressors ($m > k$); they are underidentified if $m < k$; and they are exactly identified if $m = k$. Estimation of the IV regression model requires exact identification or overidentification.

Two Stage Least Squares

The TSLS estimator in the general IV regression model in Equation (10.12) with multiple instrumental variables is computed in two stages:

1. **First-stage regression(s):** Regress X_{1i} on the instrumental variables $(Z_{1i}, \ldots, Z_{mi})$ and the included exogenous variables $(W_{1i}, \ldots, W_{ri})$ using OLS. Compute the predicted values from this regression; call these $\hat{X}_{1i}$. Repeat this for all the endogenous regressors $X_{2i}, \ldots, X_{ki}$, thereby computing the predicted values $\hat{X}_{1i}, \ldots, \hat{X}_{ki}$.
2. **Second-stage regression:** Regress Y_i on the predicted values of the endogenous variables $(\hat{X}_{1i}, \ldots, \hat{X}_{ki})$ and the included exogenous variables $(W_{1i}, \ldots, W_{ri})$ using OLS. The TSLS estimators $\hat{\beta}_0^{TSLS}, \ldots, \hat{\beta}_{k+r}^{TSLS}$ are the estimators from the second-stage regression.

In practice, the two stages are done automatically within TSLS estimation commands in modern econometric software.

EViews: ***tsls Y c X1 … Xk W1 … Wr @ c Z1 … Zm W1 … Wr***

Simultaneous equations: Definition and Example

System of equation determines **set** of variables jointly. Example:

- Structural Form:

$$p = \beta_1 + \beta_2 w + u_p$$
$$w = \alpha_1 + \alpha_2 p + \alpha_3 U + u_w$$

- Endogenous variables (p,w): "whose values are determined by the interaction ... in the model"
- Exogenous variables (U): "whose values are determined externally"
- Reduced form: "expressing the endogenous variables in terms of the exogenous variables and disturbance terms"

$$p = \frac{\beta_1 + \alpha_1 \beta_2 + \alpha_3 \beta_2 U + u_p + \beta_2 u_w}{1 - \alpha_2 \beta_2}$$
$$w = \frac{\alpha_1 + \alpha_2 \beta_1 + \alpha_3 U + u_w + \alpha_2 u_p}{1 - \alpha_2 \beta_2}$$

Simultaneous equations: Asumption B7 violation

1. Reasons: Endogeneity, dependence of explanatory variable(s) and disturbance term: disturbance term → dependent variable → endogenous regressor(s) through other equation(s)

2. Consequences: Simultaneous Equations Bias: biased and inconsistent OLS estimators, standard statistics wrongly calculated, tests invalid.

3. Detection: Durbin-Wu-Hausman test

4. Remedial measures: Instrumental Variables.
Two Stage Least Squares.

Identification

- For our course we can define identification problem as “Do we have enough exogenous variables (instruments) to get consistent estimates for all coefficients before endogenous variables?”
- Underindentification: we have not got enough instruments to get consistent estimates for all coefficients
- Exact identification: number of instruments is equal to number of endogenous variables
- Overidentification: number of instruments is greater than number of endogenous variables

Identification: Order Condition

- Write down **<u>all</u>** variables from **<u>all</u>** G equations (endogenous and exogenous, both sides).
- The number of endogenous variables is G
- For each equation calculate number of missing variables, both endogenous and exogenous, d
- For each equation compare d with $G-1$
- If $d<G-1$ then number of instruments is less than number of endogenous variables (Underidentification)
- If $d=G-1$ – Exact identification
- If $d>G-1$ – Overidentification