Endogeneity

2025/2026, Semester 1

Ralf A. Wilke
Copenhagen Business School

- Text: Wooldridge (2010), Econometric Analysis of Cross Section and Panel Data, Chapter 5, parts of Chapter 6
- Less formal: Several (sub) chapters of Wooldridge (2025), Introductory Econometrics, 8th Edition, Cengage, mainly Chapter 15

- Suppose we have a linear regression model:

$$y = \mathbf{x}\beta + u$$

- Definition: Exogeneity and Endogeneity of Independent Variables.
  - x_j is exogenous if it is uncorrelated with *u*.
  - x_j is endogenous if it is correlated with *u*.

- OLS estimation of the linear regression model requires exogeneity of x_j .

- Endogeneity can be caused by many things.
  - An important variable that is not observed and omitted
  - Functional form misspecification
  - Simultaneity
  - Measurement error in the regressors
  - ...
- Endogeneity is present in most applications in applied economic research.

Omitted Variable Bias

- What happens if we omit variables that actually belong in the true model?

  Let $K_1 + K_2 = K$ with $1 \leq K_2 < K$ and

$$y = \mathbf{x}_1\boldsymbol{\beta}_1 + \mathbf{x}_2\boldsymbol{\beta}_2 + u$$

  [Full regression: $y = \mathbf{x}\beta + u$]

  Regress y on $x_1, ..., x_{K_1}$ only: $y = \mathbf{x}_1\beta_1 + v$

  The estimator is:

$$\begin{aligned}
\hat{\beta}_1 &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'y \\
&= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + u) \\
&= \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'u
\end{aligned}$$

- Then because of $E(\mathbf{x}_1'u) = 0$:

$$E(\hat{\beta}_1|\mathbf{X}) = \beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2$$

there is an omitted variable bias if:

- $\mathbf{x}_1'\mathbf{x}_2 \neq 0$, i.e. the two regressors sets are not orthogonal.
- $\beta_2 \neq \mathbf{0}$, i.e. the omitted variables play a role.

- The magnitude of the bias depends on the magnitude of the elements of β_2 and on how strongly the independent variables are correlated.

- Solutions:

  Instrumental Variables, Proxy Variables, Panel Data

Using a Proxy Variable for Unobserved Explanatory Variables

- A more difficult problem arises when a model excludes a key variable, usually because of data unavailability.
- Example: Return to Education
  - the population model is:

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u$$

  - Suppose we do not observe the ability. Ignoring *abil* would generally give biased and inconsistent estimates of the return to education.
    - We expect an upward bias for the estimated return to education. Why?
  - How can we solve or at least mitigate this omitted variable problem?

- One possibility is to use a proxy variable for the omitted variable.
  - Something that is related to the unobserved variable.

- In the wage equation one could use the intelligence quotient, or IQ as a proxy for ability. IQ and ability do not need to be the same, but they need to be correlated.
- Suppose we have the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

with x_3^* being unobserved. We have a proxy variable, which we call x_3

- What do we require of x_3?
  - When we would run the regression $x_3^* = \delta_0 + \delta_3 x_3 + \nu_3$, we should obtain $\delta_3 > 0$. Otherwise the proxy is not good.

- The proposal is just to regress *y* on x_1, x_2, x_3 as x_3 and x_3^* would be the same. This is called the **plug-in solution** to the omitted variables problem.
- Since x_3 and x_3^* are not the same: when this procedure does in fact give consistent estimators for β_1 and β_2?

- The assumptions are with respect to *u* and ν_3 :
  1. In addition to assuming that u and x_1, x_2, x_3^* are uncorrelated, we need that u and x_3 are uncorrelated. This means that x_3 is irrelevant in the population model once x_1, x_2, x_3^* are included.

  2. The error ν_3 is uncorrelated with x_1, x_2 and x_3. This means that x_3 is a good proxy for x_3^*: $E(x_3^*|x_1, x_2, x_3) = E(x_3^*|x_3)$

- From the latter assumption follows, that $E(x_3^*|x_3) = \delta_0 + \delta_3 x_3$.

- In terms of our wage equation this means:

$$E(abil|educ, exper, IQ) = E(abil|IQ) = \delta_0 + \delta_3 IQ$$

  thus the average value of ability only changes with *IQ*.

- More formally, what are the implications of the two assumptions?

- By combining

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$
$$x_3^* = \delta_0 + \delta_3 x_3 + \nu_3$$

we obtain:

$$y = (\beta_0 + \beta_3 \delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \delta_3 x_3 + u + \beta_3 \nu_3$$

  - now let us denote $e = u + \beta_3 \nu_3$ as the composite error.
  - And note that *u* and ν_3 have both zero mean and each is uncorrelated with x_1, x_2 and x_3. Then *e* has also zero mean and is uncorrelated with x_1, x_2 and x_3 .
- For this reason, we can write

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e$$

  - This gives unbiased (or consistent) estimates of $\alpha_0, \beta_1, \beta_2$ and α_3
  - We do not get unbiased estimates for β_0, β_3.
  - In an application α_3 may even be of more interest than β_3.

- Example: Return to education
  - Wage2.dta
  - We estimate a log wage equation without *IQ* (1) and with *IQ* (2).
  - Our primary interest is in what happens to the estimated return to education.

	log(wage)	
Indep. Variables	*(1)*	(2)
educ	0.065 (0.006)	0.054 (0.007)
exper	0.014 (0.003)	0.014 (0.003)
tenure	0.012 (0.002)	0.011 (0.002)
…	…	…
IQ	-	0.0036 (0.0010)
intercept	5.395 (0.113)	5.176 (0.128)
Observations *R-Squared*	935 0.253	935 0.263

- In model (1) the estimated return to education is 6.5%, while in model (2) it is just 5.4%. This corresponds to our beliefs about omitted variable bias.

- In particular the estimate decreases but it is still large.

- In the data wage2.dta there are also other measures of ability, such as *Knowledge of the World of Work (KWW)* test.

Functional form misspecification

- Special case: omission of a relevant variable x_1^2.
- Suppose
$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + u \\ y &= \beta_0 + \beta_1 x_1 + \tilde{u} \end{aligned}$$
  with $E(u|x_1, x_1^2) = 0$ and $\tilde{u} = \beta_2 x_1^2 + u$.

- Now, since $cov(x_1, x_1^2) \neq 0$ and if $\beta_2 \neq 0$, we do not have $E(\tilde{u}|x_1) = 0$ and we would have a bias due to functional form misspecification.
- Solution: Test for functional form (RESET), Non- and semiparametric methods.

RESET test for model specification

$$y = \mathbf{x}\beta + u$$

- How do we know whether we have assumed the correct functional form?
  - For example: have we included all relevant quadratics and interaction terms?
- By noting that y^2 and y^3 are highly nonlinear functions of all regressors and their interactions, we could use the fitted values of the model above to compute $\hat{y}^2$ and $\hat{y}^3$.
- Then we estimate

$$y = \mathbf{x}\beta + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + u$$

and perform an F-test for joint significance of $\hat{y}^2$ and $\hat{y}^3$:

$$H_0 : \delta_1 = \delta_2 = 0$$

Simultaneity

- If an explanatory variable is determined simultaneously with the dependent variable, it is generally correlated with the error terms.
- In this case OLS is biased and inconsistent.
- Will be done in Part B2 “Simultaneous Equation Models” of the course.

Measurement error in an explanatory variable

- We consider the simple regression model:

$$y = \beta_0 + \beta_1 x_1^* + u$$

  and assume that it satisfies the Gauss Markov assumptions.
- We do not observe x_1^* but x_1 (e.g. actual and reported income).
- The measurement error in the population is: $e_1 = x_1 - x_1^*$
- We assume: $E(e_1) = 0$
- Moreover, we assume that *u* is uncorrelated with x_1 and x_1^* :

$$E(y|x_1, x_1^*) = E(y|x_1^*)$$

- The model can be written as: $y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$

- The **classical errors-in-variables** (**CEV**) assumption is that the measurement error is uncorrelated with the unobserved explanatory variable: $cov(x_1^*, e_1) = 0$

  - This has the meaning that the observed measure x_1 consists of two uncorrelated components: $x_1 = x_1^* + e_1$
  - (We still assume that *u* is uncorrelated with x_1 and x_1^*.)
  - The above assumption implies that x_1 and e_1 must be correlated:

$$cov(x_1, e_1) = E(x_1 e_1) = E(x_1^* e_1) + E(e_1^2) = 0 + \sigma_{e_1}^2 = \sigma_{e_1}^2$$

  - This correlation causes problems for the OLS estimation.

- This implies for our model $y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$ that since *u* and x_1 are uncorrelated, the covariance between x_1 and the composite error $u - \beta_1 e_1$ is:

$$cov(x_1, u - \beta_1 e_1) = -\beta_1 cov(x_1, e_1) = -\beta_1 \sigma^2_{e_1}$$

- Note also that $var(x_1) = var(x_1^*) + var(e_1)$
- Then one can show:

$$\begin{aligned} plim(\hat{\beta}_1) &= \beta_1 + \frac{cov(x_1, u - \beta_1 e_1)}{var(x_1)} \\ &= \beta_1 - \frac{\beta_1 \sigma^2_{e_1}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}} = \beta_1 \left(1 - \frac{\sigma^2_{e_1}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}}\right) \\ &= \beta_1 \left(\frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}}\right) \end{aligned}$$

- This equation is very interesting: $plim(\hat{\beta}_1)$ is always closer to zero than β_1: **attenuation bias**

- OLS is biased in the classical error in variables model:
  - If β_1 is positive, it will underestimate β_1 and vice versa.
- Things are more complicated if we look at the multiple regression model but again OLS will be biased and inconsistent.

- Solution: Instrumental Variable estimation, ...

IV Estimation of the Multiple Regression Model

- The model is: $y = \mathbf{x}\beta + u$

  with $E(u) = 0$, β is *Kx1* and $\mathbf{x} = (1, x_2, ..., x_K)$ and x_K is endogenous with $cov(x_K, u) \neq 0$.
- We call the above equation **structural equation** as we are interested in the coefficients.
- We will use an instrument for x_K to obtain consistent estimates.
- We need another exogenous variable z_1 with

$$cov(z_1, u) = 0 \; .$$

- Then $E(\mathbf{z}'u) = \mathbf{0}$ with $\mathbf{z} = (1, x_2, ..., x_{K-1}, z_1)$.

- A variable z_1 is a candidate for an instrument for a variable x_K if it satisfies:

$$cov(z_1, u) = 0$$

- Some remarks on the choice of an instrument:
  - It is often difficult to find a good instrument.
  - A proxy variable does not make a good instrument as it is supposed to be correlated with the error term.
  - Example: Ability is not observed and IQ is highly correlated with ability. Then it is a candidate for a proxy but clearly violates the condition.
  - If one is not sure about the quality of an instrument, it may be better to use a proxy variable (if available).

- The **reduced form equation** is

$$x_K = \delta_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + r_K$$

with $E(r_K) = 0$ and r_K is uncorrelated with all exogenous variables.

  - Rank condition: We require a non-zero partial correlation between x_K and z_1: $\theta_1 \neq 0$. (use t-test to check this)
  - This means x_K and z_1 need to be partially related.
  - This implies $\text{rank } E(\mathbf{z}'\mathbf{x}) = K$ which makes it invertible.
  - We obtain:

$$\begin{aligned} y &= \mathbf{x}\beta + u \\ \mathbf{z}'y &= \mathbf{z}'\mathbf{x}\beta + \mathbf{z}'u \\ E[\mathbf{z}'y] &= E[\mathbf{z}'\mathbf{x}\beta] + E[\mathbf{z}'u] \\ E[\mathbf{z}'y] &= E[\mathbf{z}'\mathbf{x}]\beta \\ \beta &= [E(\mathbf{z}'\mathbf{x})]^{-1} E(\mathbf{z}'y) \end{aligned}$$

  - $E[\mathbf{z}'\mathbf{x}]$ and $E[\mathbf{z}'y]$ can be consistently estimated.

- Given a random sample *i=1,....,N* the instrumental variables estimator of β is

$$
\begin{aligned}
\hat{\beta} &= \left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{x}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'y_i\right) \\
&= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y}
\end{aligned}
$$

with ***Z*** and ***X*** are *N x K* and ***Y*** is *N x 1*.

$\mathbf{x}_i$ and $\mathbf{z}_i$ are the *i*'th row of ***X*** and ***Z*** respectively.

- <u>Example</u>: College Proximity as IV for Education
  - Data: Card.dta
  - Log(wage) is dependent variable, several controls
  - Instrument for education: dummy if someone grew up near a four year old college (*nearc4*).
  - We assume that *nearc4* is uncorrelated with the error. Moreover, to be a valid instrument it has to be partially correlated with *educ.*
  - We can test this by estimating the reduced form equation:

$$
\begin{aligned}
\widehat{educ} &= 16.64 + 0.320 nearc4 + \ldots \\
& \quad (0.24) \quad (0.088) \\
& \quad n = 3,010, \quad R^2 = 0.477
\end{aligned}
$$

  - The *t*-statistic is 3.64 and therefore if *nearc4* is uncorrelated with the error term, we can use it as IV for *educ*.

- The following table reports OLS and IV estimates.

Dependent Variable: log(wage)		
Independent Variable	*(1) OLS*	*(2) IV*
Educ	*0.075* *(0.003)*	*0.132* *(0.055)*
Exper	*0.085* *(0.007)*	0.108 (0.024)
Exper^2	-0.0023 (0.0003)	*-0.0023* *(0.0003)*
…other controls	…	…
Observations *R-squared*	3,010 0.300	3,010 0.238

- IV estimate is almost twice as large as the OLS estimate.
- SE of the IV estimate is 18 times larger. This is the price we have to pay if we use an instrument to obtain a consistent estimator.

Two Stage Least Squares (2SLS)

- Sometimes there are multiple valid IVs for an endogenous explanatory variable.
- Suppose the variables $z_1, ..., z_M$ satisfy

$$\mathrm{cov}(z_h, u) = 0 \text{ for } h = 1, ..., M$$

- We could simply use both of them as instruments and obtain multiple IV estimators.
- The idea is to use both together to obtain a more efficient estimator:
  - Let $\mathbf{z} = (1, x_2, ..., x_{K-1}, z_1, ..., z_M)$ be *1x L* with *L=K-1+M*.
  - As each element of ***z*** is uncorrelated with *u*, any linear combination is also uncorrelated with *u*.

- The linear combination of ***z*** which is most highly correlated with x_K is the linear projection of x_K on ***z***. The reduced form equation is

$$x_K = \delta_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + ... + \theta_M z_M + r_K$$

  where $E(r_K) = 0$ and r_K is uncorrelated with all elements of ***z***.
- r_K is correlated with *u* if x_K is endogenous.
- x_K^* is not correlated with *u* with

$$x_K^* = \delta_1 + \delta_2 x_2 + ... + \delta_{K-1} x_{K-1} + \theta_1 z_1 + ... + \theta_M z_M$$

- Consistently estimate the parameters by OLS and use fitted values as an estimator for x_{iK}^*

$$\hat{x}_{iK} = \hat{\delta}_1 + \hat{\delta}_2 x_{i2} + ... + \hat{\delta}_{K-1} x_{i,K-1} + \hat{\theta}_1 z_{i1} + ... + \hat{\theta}_M z_{iM}$$

- We require that at least one θ_j is non-zero. Use F-test.

- Now, let $\hat{\mathbf{x}}_i = (1, x_{i1}, ..., x_{i,K-1}, \hat{x}_{iK})$ and use it as the instruments for $\mathbf{x}_i$:

$$\begin{aligned}\hat{\beta} &= \left(N^{-1}\sum_{i=1}^{N}\hat{\mathbf{x}}_i'\mathbf{x}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\hat{\mathbf{x}}_i' y_i\right) \\ &= (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y}\end{aligned}$$

- It can be shown that $\hat{\mathbf{X}}'\mathbf{X} = \hat{\mathbf{X}}'\hat{\mathbf{X}}$ and thus

$$\hat{\beta} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}$$

- This estimator is consistent under the conditions:

$E(\mathbf{z}'u) = \mathbf{0}$, $\text{rank } E(\mathbf{z}'\mathbf{x}) = K$, $\text{rank } E(\mathbf{z}'\mathbf{z}) = L$, $L \geq K$

- The last condition suggests that we need at least as many instruments as explanatory variables in the model (**order condition**)

- By noting $\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ the 2SLS estimator can be written as:

$$
\begin{aligned}
\hat{\beta} &= (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y} \\
&= (\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} \\
&= \left[\left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{z}_i\right)\left(\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{x}_i\right)\right]^{-1} \\
&\quad \times \left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{z}_i\right)\left(\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{z}_i'y_i\right)
\end{aligned}
$$

(AB)'=B'A'

  - One step estimator.
  - In the case of L=K (just identified): Replace $\hat{\mathbf{X}} = \mathbf{Z}$ in the equation for $\hat{\beta}$.

- By using $y_i = \mathbf{x}_i\beta + u_i$, we obtain:

$$\hat{\beta} = \beta + \left[\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{z}_i\right)\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{x}_i\right)\right]^{-1}$$
$$\times\left(N^{-1}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{z}_i\right)\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'\mathbf{z}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{z}_i'u_i\right)$$

  - By application of a law of large numbers to the various terms and the Slutsky theorem, we obtain consistency provided $E(\mathbf{z}_i'u_i) = 0$.

- Under homoscedasticity $E(u^2\mathbf{z}'\mathbf{z}) = \sigma^2 E(\mathbf{z}'\mathbf{z})$, which is slightly weaker than $E(u^2|\mathbf{z}) = \sigma^2$, it is possible to show that asymptotically:

$$\sqrt{N}(\hat{\beta} - \beta) \to N\left(\mathbf{0}, \sigma^2([E(\mathbf{x}'\mathbf{z})][E(\mathbf{z}'\mathbf{z})]^{-1}E(\mathbf{z}'\mathbf{x})])^{-1}\right)$$

  - The more unrelated (orthogonal) ***x*** and ***z***, the smaller is $E(\mathbf{x}'\mathbf{z})$, and the larger the variance of $\hat{\beta}$ becomes.

- Under homoskedasticity the 2SLS estimator is asymptotically efficient.

- The variance matrix can be estimated by $\hat{\sigma}^2(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$ with

$$\hat{\sigma}^2 = (N-K)^{-1}\sum_{i=1}^{N}\hat{u}_i^2 = (N-K)^{-1}\sum_{i=1}^{N}(y_i - \mathbf{x}_i\hat{\beta})^2$$

  - The more instruments we use the more variation we will have in $\hat{\mathbf{X}}$ and the smaller the variance of $\hat{\beta}$.

Slide 32

RW5 it is equivalent to u^2 is uncorrelated with all x_j, x_j^2 and cross products x_j*x_k

Ralf Wilke; 12-03-2018

- The IV estimator with multiple instruments is called two stage least squares (2SLS) estimator:
  - One can show that the IV estimates are identical to OLS estimates from the regression of y on $1, x_2, ..., x_{K-1}$ and $\hat{x}_K$. This is the second stage.
  - The first stage is the regression of

$$x_K \text{ on } 1, x_2, ..., x_{K-1}, z_1, ..., z_M.$$

- 2SLS standard errors are larger than for OLS. This is because:
  - $\hat{x}_K$ has less variation than x_K.
  - $\hat{x}_K$ has more correlation with $x_2, ...x_{K-1}$ than x_K. (multicollinearity in the second stage)

IV Estimation with a poor Instrumental Variable

- Simple regression: $y = \beta_0 + \beta_1 x + u$, instrument: *z*
- IV estimates can have large standard errors if *x* and *z* are only weakly correlated. (Don't use IV in this case.)
- IV estimates can have a large asymptotic bias if *z* and *u* are only weakly correlated:
  - For illustration: model with one regressor (*x*) and one instrument (*z*):

$$plim \hat{\beta}_1 = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)} \frac{\sigma_u}{\sigma_x}$$

  This implies that the bias can be large if the population correlation between *z* and *x* is small even if the population correlation between *z* and *u* is small.
- For this reason IV can be worse in terms of consistency than OLS even if *Corr(z,u)* is small (provided that *Corr(z,x)* is also small).
- One can show that IV is only superior in terms of asymptotic bias if

$$Corr(z,u)/Corr(z,x) < Corr(x,u)$$

R-Squared and IV Estimation

$$R^2 = 1 - SSR/SST$$

- SSR (sum of squared IV residuals) can be larger than SST. For this reason the R-squared can become negative and it is smaller than for OLS.
- It is not clear whether the IV R-squared should be reported after IV estimation.
- If you try to maximise the R-squared, use OLS as IV tries to improve the quality of ceteris paribus effects.

Some Remarks:

- If the R-squared of $\hat{x}_K$ on the exogenous variables appearing in the structural equation (without instruments) is very large, the standard error of 2SLS explodes. Can be verified with data at hand.
- 2SLS can be also used in models with more than one endogenous variable.
  - We need more candidates for instruments to achieve identification.
  - The sufficient condition for identification is the **rank condition**.
- Since R-squared after 2SLS cannot be compared to OLS R-squared we must be careful when using the F-test.
- It is possible to derive a statistic with an approximate F-distribution in large samples. Use econometric packages to test multiple hypothesis after 2SLS as commands are available.

IV Solutions to Errors in Variables Problems

- Suppose we have the model

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + u$$

with x_1^* is not observed but $x_1 = x_1^* + e_1$, with e_1 is the measurement error.

  - x_1 and e_1 are correlated and therefore OLS when regressing *y* on x_1 and x_2 is biased and inconsistent.

- In the case of the classical errors-in-variable model, we have seen that the OLS estimator is biased towards zero.
- It is possible to use an IV procedure to overcome the measurement error problem.

- After plugging in the model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (u - \beta_1 e_1)$$

- We assume that e_1 is uncorrelated with x_1^* and x_2.
- We also assume that *u* is uncorrelated with x_1, x_1^* and x_2.
- Therefore x_2 is exogenous, but x_1 is correlated with e_1.
- We need an instrument for x_1 that is correlated with x_1, but uncorrelated with u and e_1.
  - One possibility could be a second measurement of x_1^*:

    $$z_1 = x_1^* + a_1$$

    where we need that a_1 is uncorrelated with e_1 and *u*. This means the two measurement errors need to be uncorrelated.
  - Another possibility is to use another exogenous variable as IV for x_1 as with the usual IV procedure.

- Example: Wage regression with two erroneous measures of ability. (wage2.dta,IV_ErrorsInVariables.R)
- Continued example from proxy variable model.
- We use *IQ* as a mismeasured observed variable for ability.
  - But now *IQ* is endogenous. Given that *IQ* is correlated with *educ,* the estimate for the return to education might be biased as well.
- We use *KWW* as the IV for IQ.
  - KWW is another mismeasured ability variable.
- Resulting IV estimate for *educ* is smaller and insignificant.
  - Statistically not different from OLS estimate.
  - Large standard errors due to multicollinearity in second stage regression.

Testing for Endogeneity

- 2SLS is less efficient than OLS and can have large standard errors.
- It is therefore useful to have a test for endogeneity of explanatory variables to show whether 2SLS is even necessary:

1. Regression based test (Hausman, 1978)

2. Durbin, Wu and Hausman suggest a test which directly compares OLS and 2SLS estimates and determines whether differences are statistically significant (DWH Test):

   H0: all regressors are exogeneous

Regression based test (Hausman, 1978)

- Suppose we have the structural equation

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$

  with y_2 being endogenous and there are two exogenous variables z_3, z_4 which are not included in the model.

- The idea behind the test is as follows:
  - We have:
    $$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$
    and
    $$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + \nu_2$$
  - Each z_j is uncorrelated with u_1.
  - y_2 is uncorrelated with u_1 if and only if, ν_2 is uncorrelated with u_1 and has zero mean. This is what we want to test.
  - Write $u_1 = \delta_1 \nu_2 + e_1$, where e_1 is uncorrelated with ν_2 and $E(e_1) = 0.$
  - Then u_1 and ν_2 are uncorrelated if and only if $\delta_1 = 0$.
  - Simply plug this into the structural equation and do a *t* test.
  - Since ν_2 is not observed, use instead the residuals from the reduced form equation as a regressor:
    $$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{\nu}_2 + error$$

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{\nu}_2 + error$$

- We then test: $H_0 : \delta_1 = 0$ using a t-test.
- if we reject it at a small significance level, we conclude y_2 is endogenous because u_1 and ν_2 are correlated.

- Practical guideline for the **Hausman test**:

1. Estimate the reduced form for y_2 and obtain $\hat{\nu}_2$.
2. Add $\hat{\nu}_2$ to the structural regression and estimate it by OLS. You may want to use a heteroscedasticity robust version of the t-test for testing whether the coefficient on $\hat{\nu}_2$ is significant. If it is statistically significant from zero, we conclude that y_2 is indeed endogenous.

- This test requires the availability of valid instruments.

- It can be easily extended to the case of multiple endogenous variables:
  - The reduced form of step 1 is then estimated for each endogenous regressor.
  - The regression in step 2 then includes the residuals obtained by all regressions of step 1. The test is then to test for joint significance of all residuals using F- or LM test in this regression.

Durbin-Wu-Hausman (DWH) test

- If under the H0 all regressors are exogenous but some are endogenous under H1, we can base a test directly on the difference between 2SLS and OLS estimators.

- The DWH statistic

$$(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})'[(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}]^{-}(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})/\hat{\sigma}^2_{OLS}$$

  is asymptotically $\chi^2_{(G_1)}$ distributed with G_1 is the number of endogenous regressors and $\mathbf{H}^{-}$ is a generalised RW6 inverse of **H**.

- Statistic may be cumbersome to compute. Wald test.

RW6 Generalised inverse A^- is such that: A A^- A=A.
A useful concept in the case the regular inverve A^-1 does not exist. If A^-1 exists it is the unique generalised inverve A^-.
Ralf Wilke; 12-03-2018

Testing Overidentifying restrictions

- If we have more instruments for one endogenous explanatory variable, we can test whether at least some of them are not correlated with u_1 (validity of the instrument). We need that at least one of the IVs is exogenous and we need to know which one.
- Then we can test the overidentifying restrictions that are used in 2SLS:
  - E(**z**'u)=**0** is Lx1
  - Suppose we estimate the same model by 2SLS but with a different number of instruments. Say model 1 is just identified and model 2 is overidentified. Then L in the second model is bigger than in the first and thus we have imposed additional moment conditions. These conditions can be tested.
  - Under H0: E(**z**'u)=**0** is true for model 2.
- Regression based version (LM test): ***Sargan Test***

More remarks on IV estimation:

- Heteroscedasticity in the context of 2SLS raises the same issues as with OLS.
- There are standard errors and test statistics available which are robust with respect to heteroscedasticity. R: iv_robust in `estimatr`
- There are also tests for heteroscedasticity available.

- Lasso IV variable Selection for IV estimation:

Belloni et al. (2012) "Sparse models and methods for optimal instruments with an application to eminent domain." Econometrica 80: 2369-2429.

- In more general regression models, IV models are not identified (set estimation).
  - Due to support restrictions on the dependent variable, ordered data, interval censoring of regressors etc.
  - Compare Chesher/Rosen (2017), Generalised Instrumental Variable Models, Econometrica, 959-989 or Chesher/Rosen (2013), What Do Instrumental Variable Models Deliver with Discrete Dependent Variables? American Economic Review.
  - Check for potential issues before you apply IV methods outside the standard linear regression world.

Summary

- We have seen the method of instrumental variables as a way to consistently estimate the parameters in a linear model when there are endogenous explanatory variables.
- When instruments are poor, IV estimates can be worse than OLS.
- 2SLS is routinely used in economics and social sciences alike.
- Tests for endogeneity.