Endogeneity and Instrumental Variables

Isabel Casas
Office: V5-206a-2
icasas@sam.sdu.dk

Outline

- Omitted variables
  - Consequences: Biased and inconsistent OLS estimator
  - Solution: Proxy variables or instrumental variables (IV)
- Measument errors
  - Consequences: Biased and inconsistent estimators
  - Solution: IV
- Simultaneity
  - Consequences: Biased and inconsistent estimators
  - Solution: IV
- A more general solution:
  - 2SLS estimator and its properties
  - 2SLS pitfalls

Source of Endogeneity: Measurement Error

- The dependent variable is erratically measured.
  - Because this error is uncorrelated with the regressors $\Rightarrow$ OLS is fine
- One of the regressors is erratically measured.
  - If the error is uncorrelated with true variable $\Rightarrow$ OLS is fine
  - If the error is correlated with the true variable, then we need an IV

Measurement error in the dependent variable

Let define the *correctly* specified model:

$$\mathbf{y}^{\star} = \beta_0 + \beta_1 \mathbf{X}_1 + \ldots \beta_k \mathbf{X}_k + \boldsymbol{\epsilon}$$

We only observe $\mathbf{y} = \mathbf{y}^{\star} + \mathbf{e}_0$, then the *incorrect* model:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{X}_1 + \ldots \beta_k \mathbf{X}_k + \underbrace{\boldsymbol{\epsilon} + \mathbf{e}_0}_{\eta}$$

If $E(\mathbf{X}_i \boldsymbol{\eta}) = 0$ for $i = 1, \ldots, k$
(No correlation between regressors and committed error, $\mathbf{e}_0$)
$\Rightarrow$ the OLS estimators are consistent.

Measurement error in the dependent variable

Example (Young and Bielinska-Kwapisz, 2002) on alcohol consumption in the US (different states).

$$consumption = \beta_0 + \beta_1\ price + \text{ other variables} + \epsilon$$

- The price recorded for a six pack of Heineken, 750 ml of J&B Scotch and 1.5 l bottle of Gallo or Livingston Cellars Chablis.
- The error made in the price is correlated with consumption because this price variable does not represent the type of drinks used for alcohol abuse.
- Also simultaneity (higher demand -> higher price)
- OLS inconsistent and biased, underestimating how consumption depends on price.
- Instruments for price: beer taxes, distilled spirit taxes and state markups

Measurement error in a regressor

Let us assume that $\mathbf{X}_k^\star$ is the true value of this regressor. However, we observe $\mathbf{X}_k = \mathbf{X}_k^\star + \mathbf{e}_k$ (price).

The *correctly* specified model:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{X}_1 + \ldots \beta_k \mathbf{X}_k^\star + \boldsymbol{\epsilon}$$

The *incorrect* model:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{X}_1 + \ldots \beta_k \mathbf{X}_k + \underbrace{\boldsymbol{\epsilon} - \beta_k \mathbf{e}_k}_{\eta}$$

Measurement error in a regressor

Assume $E(e_k) = 0$. Otherwise $\beta_k E(\mathbf{e}_k)$ is added to the intercept.

Case 1 : $\text{cov}(\mathbf{X}_k, \mathbf{e}_k) = 0$, **no endogeneity in our model**.

- The OLS estimators are consistent with greater variance due to the error

Case 2 : $\text{cov}(\mathbf{X}_k, \mathbf{e}_k) \neq 0$ although $\text{cov}(\mathbf{X}_k^\star, \mathbf{e}_k) = 0$

- The true variable is uncorrelated with the error, but the variable we use it is:

$$\begin{aligned}\text{cov}(\mathbf{X}_k, \boldsymbol{\eta}) = &- \beta_k E(\mathbf{X}_k \mathbf{e}_k) = -\beta_k E(\mathbf{X}_k^\star \mathbf{e}_k) - \beta_k E(\mathbf{e}_k^2) \\ &= -\beta_k \sigma^2_{\mathbf{e}_k} \neq 0\end{aligned}$$

- Therefore, there is **endogeneity** in our model.
- The OLS estimators are biased and inconsistent
- Solution: Instrumental Variables

IV when there are measurement errors

- $\text{cov}(\mathbf{X}_i, \boldsymbol{\epsilon}) = 0$ for $i = 1, \ldots, k-1$
- $\text{cov}(\mathbf{e}_k, \mathbf{X}_k^\star) = 0 \Rightarrow \text{cov}(\mathbf{e}_k, \mathbf{X}_k) \neq 0.$
- Therefore we need an IV for $\mathbf{X}_k$.
- Find an instrument $\mathbf{z}$ and construct:

$$\mathbf{Z} = \begin{pmatrix} 1 & X_{1,1} & \ldots & X_{k-1,1} & z_1 \\ 1 & X_{1,2} & \ldots & X_{k-1,2} & z_2 \\ \vdots & & & \vdots & \\ 1 & X_{1,n} & \ldots & X_{k-1,n} & z_n \end{pmatrix}$$

- The consistent IV estimator is:

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

Source of Endogeneity: Simultaneous Equations

- The dependent and independent variables are entwined
- Solution: Instrumental variables

Simultaneity

$$\text{homocides} = \alpha_1 + \beta_{1,0}\ \text{police} + \beta_{1,1}\ \ \text{family\_rent} + \epsilon_1$$

$$\text{police} = \alpha_2 + \beta_{2,0}\ \text{homocides} + \beta_{2,1}\ \text{other factors} + \epsilon_2$$

- An increase in the number of $homocides$ will affect the number of $police$ agents in the streets
- Therefore, there is correlation between ϵ_1 and $police$ $\Rightarrow$ Endogeneity
- Solution: Instrumental Variables

IV with simultaneous equations

We have the structural model of two equations:

$$\mathbf{y}_1 = \alpha_1 \mathbf{y}_2 + \beta_1 w_1 + \boldsymbol{\epsilon}_1$$
$$\mathbf{y}_2 = \alpha_2 \mathbf{y}_1 + \beta_2 w_2 + \boldsymbol{\epsilon}_2$$

- w_1, w_2 are exogenous
- We consider the intercept zero for simplicity
- Parameters $\alpha_1, \alpha_2, \beta_1, \beta_2$ are called *structural parameters*
- $\boldsymbol{\epsilon}_1$, $\boldsymbol{\epsilon}_2$ are called *structural errors*

IV with simultaneous equations

We want to estimate $\mathbf{y}_2$. We substitute $\mathbf{y}_1$ in the second equation and regress on y_2.

$$\begin{aligned}\mathbf{y}_2 =& \alpha_2(\alpha_1\ \mathbf{y}_2 + \beta_1\ w_1 + \boldsymbol{\epsilon}_1) + \beta_2\ w_2 + \boldsymbol{\epsilon}_2 \\ &\Downarrow \\ (1-\alpha_2\alpha_1)\mathbf{y}_2 =& \alpha_2\beta_1\ w_1 + \beta_2 w_2 + \alpha_2\boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_2\end{aligned}$$

We have to assume that $\alpha_1\alpha_2 \neq 1$, so:

$$\mathbf{y}_2 = \theta_{21}\ w_1 + \theta_{22}\ w_2 + \boldsymbol{\nu}_2$$

IV with simultaneous equations

$$\mathbf{y}_2 = \theta_{21}\ w_1 + \theta_{22}\ w_2 + \boldsymbol{\nu}_2 \quad \text{reduced eq}$$

where

- $\theta_{21} = \alpha_2\beta_1/(1-\alpha_2\alpha_1)$
- $\theta_{22} = \beta_2/(1-\alpha_2\alpha_1)$
- $\boldsymbol{\nu}_2 = (\alpha_2\boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_2)/(1-\alpha_2\alpha_1)$
- $\boldsymbol{\nu}_2$ is a linear function of $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$, then $\boldsymbol{\nu}_2$ is uncorrelated with w_1 and w_2.
- We can estimate θ_{12} y θ_{22} by OLS.
- If $\alpha_2 = 0$ then there is no simultaneity (test this).
- If $\alpha_2 \neq 0$, then there is endogeneity and we need IV for $\mathbf{y}_1$.

IV with simultaneous equations

A rank condition is necessary for consistency:

- At least one of the exogenous variables of the second equation is not in the first equation.
- At least one of the exogenous variables of the first equation should have a nonzero coefficient.

Counter example: House expenses and savings

Let us assume that house $expenses$ and $savings$ of a random family are determined simultaneously by:

$$expenses = \alpha_1\ savings + \beta_{10} + \beta_{11}\ salary + \beta_{12}\ educ + \beta_{13}\ age + \epsilon_1$$
$$savings = \alpha_2\ expenses + \beta_{20} + \beta_{21}\ salary + \beta_{12}\ educ + \beta_{13}\ age + \epsilon_2$$

2 Stages Least Squares

- When each endogenous variable has more than one IV
- Statistical properties of the 2SLS
- Example with simultaneous equations

Two stage least squares (2SLS)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{with } \mathbf{X}_k \text{ endogenous}$$

Assuming that we have valid instruments: $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m$ for $\mathbf{X}_k$, i.e:

- $\text{cov}(\mathbf{z}_i, \boldsymbol{\epsilon}) = 0$ for $i = 1, \ldots, m$
- Each $\mathbf{z}_i$ is partially correlated with $\mathbf{X}_k$
- Out of all linear combinations of $\mathbf{z}_i$ the 2SLS method used the most highly correlated with $\mathbf{X}_k$
- $\mathbf{X}_k = \delta_0 + \delta_1 \mathbf{X}_1 + \ldots + \delta_{k-1}\mathbf{X}_{k-1} + \theta_1\ \mathbf{z}_1 + \ldots + \theta_m\ \mathbf{z}_m + \eta$
- As z is uncorrelated with $\boldsymbol{\epsilon}$ then $\hat{\mathbf{X}}_k = \hat{\delta}_0 + \hat{\delta}_1 \mathbf{X}_1 + \ldots + \hat{\delta}_{k-1}\mathbf{X}_{k-1} + \hat{\theta}_1\ \mathbf{z}_1 + \ldots + \hat{\theta}_m\ \mathbf{z}_m$ isn't either.
- So $\hat{\mathbf{X}}_k$ can be used as an instrument of $\mathbf{X}_k$

Two steps least squares (2SLS)

We could estimate the parameters of interest with two regressions:

[Stage 1] Estimate $\mathbf{X}_k$ by OLS:

- $\mathbf{X}_k = \delta_0 + \delta_1 \mathbf{X}_1 + \ldots + \delta_{k-1} \mathbf{X}_{k-1} + \theta_1 \mathbf{z}_1 + \ldots + \theta_m \mathbf{z}_m + \boldsymbol{\eta}$
- $\hat{\mathbf{X}}_k = \hat{\delta}_0 + \hat{\delta}_1 \mathbf{X}_1 + \ldots + \hat{\delta}_{k-1} \mathbf{X}_{k-1} + \hat{\theta}_1 \mathbf{z}_1 + \ldots + \hat{\theta}_m \mathbf{z}_m$

[Stage 2] Substitute $\mathbf{X}$ by $\hat{\mathbf{X}}$ in the original equation:

- $\hat{\boldsymbol{\beta}}^{2SLS} = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y}$

The 2SLS estimator is the same than the IV estimator if we have only one IV.

If there are more than one endogenous variable, we have more regressions in Stage 1.

Two steps least squares (2SLS)

We test for validity of instruments and rank condition in the first stage:

- Even if they are not individually significant, they should be jointly significant.
- If $\boldsymbol{\eta}$ is homokedastic then we test joint significance $H_0 : \theta = \theta_1 = \ldots = \theta_m = 0$ with an F-test
- Otherwise, use the Wald test or LM test.
- If we cannot reject H_0, we should not use the 2SLS.

2SLS in R

- First time, install the package sem.

  ```
  > install.packages("sem")
  ```

- Include this library with

  ```
  > library(sem)
  ```

- Look at the help of function tsls

  ```
  > ?tsls
  ```

Statistical properties of 2SLS

Let have the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\mathbf{X} = (\mathbf{1}, \mathbf{X}_1, \ldots, \mathbf{X}_k)'$.

- There might be several endogenous variables amongst the regressors (correlated with $\boldsymbol{\epsilon}$).
- We have one or more IV for each endogenous variable
- There exists $\mathbf{Z} = (\mathbf{1}, \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_l)'$
- Any exogenous elements of $\mathbf{X}$ are included in $\mathbf{Z}$, plus the instrumental variables of the endogenous variables.

Statistical properties of 2SLS

Using this notation, the 2SLS estimator can also be written as:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{2SLS} &= (\hat{X}'\hat{X})^{-1}(\hat{X}'Y) \\ &= \left[(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X})\right]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{y})\end{aligned}$$

This notation is handier to proof consistency and asymptotic normality.

Consistency of 2SLS

Assumption 2SLS.1: IV is exogenous

- $E(\mathbf{Z}'\boldsymbol{\epsilon}) = 0$.

Assumption 2SLS.2: Multicollinearity

- rank $E(\mathbf{Z}'\mathbf{Z}) = l$: this is automatically satisfy because all the variables in $\mathbf{Z}$ are lineally independent
- rank $E(\mathbf{Z}'\mathbf{X}) = k + 1$: It is necessary $l \geq k$ and $\mathbf{Z}$ and $\mathbf{X}$ are appropriately correlated

Assumption 2SLS.3: Homokedasticity, $E(\boldsymbol{\epsilon}^2|Z) = \sigma^2$

- $E(\boldsymbol{\epsilon}^2\mathbf{Z}'\mathbf{Z}) = \sigma^2 E(\mathbf{Z}'\mathbf{Z})$

2SLS with simultaneous equations

To make sure that Assumption 2SLS.2 is satisfied:

- At least one of the exogenous variables of the second equation is not in the first equation.
- At least one of the exogenous variables of the first equation should have a nonzero coefficient.

Counter example: House expenses and savings

Let us assume that house $expenses$ and $savings$ of a random family are determined simultaneously by:

$$expenses = \alpha_1\ savings + \beta_{10} + \beta_{11}\ salary + \beta_{12}\ educ + \beta_{13}\ age + \epsilon_1$$
$$savings = \alpha_2\ expenses + \beta_{20} + \beta_{21}\ salary + \beta_{12}\ educ + \beta_{13}\ age + \epsilon_2$$

Asymptotic normality of 2SLS

Under Assumptions 2SLS.1–2SLS.2, $\hat{\boldsymbol{\beta}}^{2SLS}$ is consistent.

Theorem

Under Assumptions 2SLS.1–2SLS.3,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^{2SLS} - \beta) \to^{d} N\left(0, \frac{\sigma^2}{E(\mathbf{X}'\mathbf{Z})[E(\mathbf{Z}'\mathbf{Z})]^{-1}E(\mathbf{Z}'\mathbf{X})}\right)$$

as $n \to \infty$

Residuals of 2SLS

- The 2SLS residuals are $\hat{\epsilon}_j = y_j - \mathbf{X}_j \hat{\boldsymbol{\beta}}^{2SLS}$ for $j = 1, 2, \ldots, n$.
- If you do the two stages by yourself, you will get $y_j - \hat{\mathbf{X}}_j \hat{\boldsymbol{\beta}}^{2SLS}$ instead, which are wrong.
- We need them to estimate σ^2:

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum \hat{\epsilon}_j^2$$

- The variance–covariance matrix is

$$\hat{\mathbf{V}}^{2SLS} = \widehat{AVar}(\hat{\boldsymbol{\beta}}^{2SLS}) = \hat{\sigma}^2 (\hat{\mathbf{X}}' \hat{\mathbf{X}})^{-1}$$

  where $\hat{\mathbf{X}}$ is estimated in Stage 1.
- The standard error is the square root of the diagonal of $\hat{\mathbf{V}}^{2SLS}$.

Covariance with heteroskedasticity

If $E(\boldsymbol{\epsilon}^2|\mathbf{X}) \neq \sigma^2$, the robust variance–covariance matrix (White) is

$$\hat{\mathbf{V}}^{2SLS} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1} \sum_{j=1}^{n} (\hat{\mathbf{X}}_j \hat{\epsilon}_j^2 \hat{\mathbf{X}}'_j)(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$$

There is also a expression with Z.

You can use the *coeftest* R function to account for this in the tests.

Hypothesis test of 2SLS estimates

- Confidence interval and t-statistics on single variables are obtained as usual, using the standard errors or robust standard errors as necessary
- Multiple restrictions of the form $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ are tested with the Wald statistics or LM-test using the White estimator of $\hat{\mathbf{V}}^{2SLS}$.

Pitfalls with 2SLS

- In practice, the 2SLS estimator is never unbiased.
- For example in a simple model with only one explanatory variable $\mathbf{X}_1$ whose instrument is z, the asymptotic bias is:

$$\text{plim } \hat{\beta}_1^{2SLS} = \beta_1 + \frac{\text{cov}(\mathbf{z}, \boldsymbol{\epsilon})}{\text{cov}(\mathbf{z}, \mathbf{X}_1)}$$

- If $\text{cov}(\mathbf{z}, \boldsymbol{\epsilon}) = 0 \Rightarrow$ consistent estimator
- Otherwise...

Pitfalls with 2SLS

- ... if $\text{cov}(\mathbf{z}, \boldsymbol{\epsilon}) \neq 0$ and the instrument is weak then $\text{corr}(\mathbf{z}, \mathbf{X}_1)$ is small which might results in large inconsistency

$$\text{plim } \hat{\beta}_1^{2SLS} = \beta_1 + \frac{\sigma_\epsilon \text{corr}(\mathbf{z}, \boldsymbol{\epsilon})}{\sigma_{X_1} \text{corr}(\mathbf{z}, \mathbf{X}_1)}$$

- In the latter case, it is better to use the OLS estimator rather than the IV estimator because

$$\text{plim } \hat{\beta}_1^{OLS} = \beta_1 + \frac{\sigma_\epsilon \text{corr}(\mathbf{X}_1, \boldsymbol{\epsilon})}{\sigma_{X_1}}$$

Pitfalls with 2SLS

- The standard errors of the 2SLS estimator tend to be large (imprecise estimator),
- This results in statistically insignificant variables.
- The size of s.e. depends on the quality of the instrument. See (AngristKrueger_table.pdf).

Pitfalls with 2SLS

- So, the bias in small sample is going to be large if we have a weak instrument
- Therefore, it is important to test the strength of the instruments in the first state of the 2SLS

$$H_0 : \ \theta_1 = \ldots = \theta_m = 0$$

- Rule-of-thumb: F-statistics should exceed 10, otherwise a weak instruments.

How do we act with endogeneity?

1. Ignore it $\Rightarrow$ OLS biased and inconsistent parameter estimates
2. Use proxy (only works with omitted variables).
   - If imperfect $\Rightarrow$ still biased and inconsistent estimates
   - But may reduce bias and lower the variance
3. IV
   - If weak instruments or not exogenous $\Rightarrow$ biased and imprecise estimates
   - Good instruments, 2SLS is still less efficient
4. First difference (panel data) works in some cases to get rid of the endogenous variable.
5. What to do? Try it all and decide with your personal criteria.

Hausman test for endogeneity

This is a regression-based test (Hausman, 1978, 1983) which is asymptotically equivalent to the original Hausman test.

Let us do it with Example 6.1 of Wooldridge:

$$\log(wage) = \delta_0 + \delta_1 \ exper + \delta_2 \ exper^2 + \alpha_1 \ educ + \boldsymbol{\epsilon}$$

- We believe that $educ$ is endogenous with intruments *motheduc*, *fatheduc* and $huseduc$.
- We have to test for it

Hausman test for endogeneity

Three steps:

1. OLS of $educ$ over all the intruments (its intruments and the rest of exogenous variables)

$$\begin{aligned} educ =& \beta_0 + \beta_1\ exper + \beta_2\ exper^2 + \beta_3\ motheduc \\ &+ \beta_4\ fatheduc + \beta_5\ huseduc + \boldsymbol{\eta} \end{aligned}$$

   and obtain the residuals $\hat{\boldsymbol{\eta}}$

2. Include the residuals in the original model and obtain the OLS estimates

$$\log(wage) = \delta_0 + \delta_1\ exper + \delta_2\ exper^2 + \alpha_1\ educ + \rho_1\hat{\eta} + \boldsymbol{\eta}$$

3. $H_0 : Cov(educ, \boldsymbol{\epsilon}) = 0\ \Rightarrow E(\boldsymbol{\eta}, \boldsymbol{\epsilon}) = 0 \Rightarrow \rho_1 = 0$?
   - If we fail to reject H_0 then *educ* is exogenous and we should estimate with OLS
   - If we reject H_0 then $educ$ is endogenous and we should estimate with 2SLS, if we trust the instruments

Q: What if we have more than one endogenous variable?

Identification

- Let us say that we have k regressors of which p are endogenous
- We say that parameters are *exactly identified* if the number of instrumental variables m is greater than the number of endogenous variables
- We say that parameters are *overidentified* if $m > p$
- We say that parameters are *underidentified* if $m < p$. In this case, we cannot find the estimators and we need to get more instruments.

Testing Overidentifying restrictions

An instrument ($\mathbf{z}$) must satified:

1. Relevant: $Cov(\mathbf{X}_k, \mathbf{z}) \neq 0$
2. Exogenous: $Cov(\boldsymbol{\epsilon}, \mathbf{z}) \neq 0$

- If the model is exactly identified then we cannot test 2). However, if the model is overidentified then we can test whether any of the instruments (we do not know which) is correlated with the error.
- The number of overidentification restrictions is $m - p$
- We do not observe ϵ but we have the residuals $\hat{\epsilon}$ from the 2SLS.

Sargan test!!!

Sargan test

H_0: All instruments are exogenous

1. Obtain residuals $\hat{\epsilon}$ from the 2SLS estimation using all instrumental variables.
2. Run an OLS of $\hat{\epsilon}$ on 1, all exogenous variables and instruments of the endogenous and obtain $LM = nR^2_{\hat{\epsilon}}$.
3. $LM \sim \chi^2_{m-p}$, find the p-value and conclude whether H_0 is rejected.

- If H_0 is rejected, then we have to choose other instruments.
- If H_0 is not rejected, then we can have some confidence in our instruments.
- There is a lot of research into weak instruments nowadays.

Control function approach to endogeneity

- This methodology is very flexible and can be used in linear models, in fact, it is equivalent to 2SLS in linear models
- It also can be used in non-linear models like Probit, logit, Poisson, etc. when endogeneity is present in the underlying linear model.

Control function approach to endogeneity

Assume that we have the following model:

$$\mathbf{y}_1 = \mathbf{X}\boldsymbol{\beta} + \alpha_1 \mathbf{y}_2 + \boldsymbol{\epsilon}$$

- $\mathbf{y}_1$ is the dependent variable
- $\mathbf{X}$ are the exogenous variables including 1
- $\mathbf{y}_2$ is the endogenous variable
- $\boldsymbol{\epsilon}$ is the error term
- $\mathbf{Z} = (\mathbf{1}, \mathbf{X}, \mathbf{Z}_2)'$ contains $\mathbf{X}$ and other exogeneous variables $\mathbf{Z}_2$ which are the instruments of $\mathbf{y}_2$.
- $\mathbf{Z}_2$ must include at least one variable
- We could run 2SLS to obtain the estimates of $\boldsymbol{\beta}$ and α_1.
- We also can...

Control function approach to endogeneity

$$\mathbf{y}_2 = \mathbf{Z}\boldsymbol{\pi}_2 + \boldsymbol{\nu}_2 \qquad E(\mathbf{Z}'\nu_2) = 0$$

Endogeneity arises if $\boldsymbol{\nu}_2$ and $\boldsymbol{\epsilon}$ are correlated, i.e. if $\boldsymbol{\epsilon} = \rho_1\boldsymbol{\nu}_2 + \mathbf{e}_1$
Substituing in the equation of $\mathbf{y}_1$:

$$\mathbf{y}_1 = \mathbf{X}\boldsymbol{\beta} + \alpha_1\mathbf{y}_2 + \rho_1\boldsymbol{\nu}_2 + \mathbf{e}_1$$

Now, $\mathbf{X}, \mathbf{y}_2, \boldsymbol{\nu}_2$ are uncorrelated with $\mathbf{e}_1$ and can run and OLS
Q: Any problem?

Control function approach to endogeneity

The problem is that $\boldsymbol{\nu}_2$ is unknown but it can be estimated by the residuals of the regression of $\mathbf{y}_2$.

The algorithm:

1. OLS on $\mathbf{y}_2 = \mathbf{Z}\boldsymbol{\pi}_2 + \boldsymbol{\nu}_2$ and obtain residuals $\hat{\boldsymbol{\nu}}_2$
2. OLS on $\mathbf{y}_1 = \mathbf{X}\boldsymbol{\beta} + \alpha_1 \mathbf{y}_2 + \rho_1 \hat{\boldsymbol{\nu}}_2 + \mathbf{e}_1$

- The inclusion of the residuals control for the endogeneity of $\mathbf{y}_2$.
- $H_0 : \rho_1 = 0$ is a t-test of exogeneity of $\mathbf{y}_2$ (use robust standard errors in case of heterogeneity).
- This methodology is better than the 2SLS methodology for nonlinear models.