Chapter 15: Instrumental Variables Estimation and Two Stage Least Squares

Introductory Econometrics: A Modern Approach

Outline

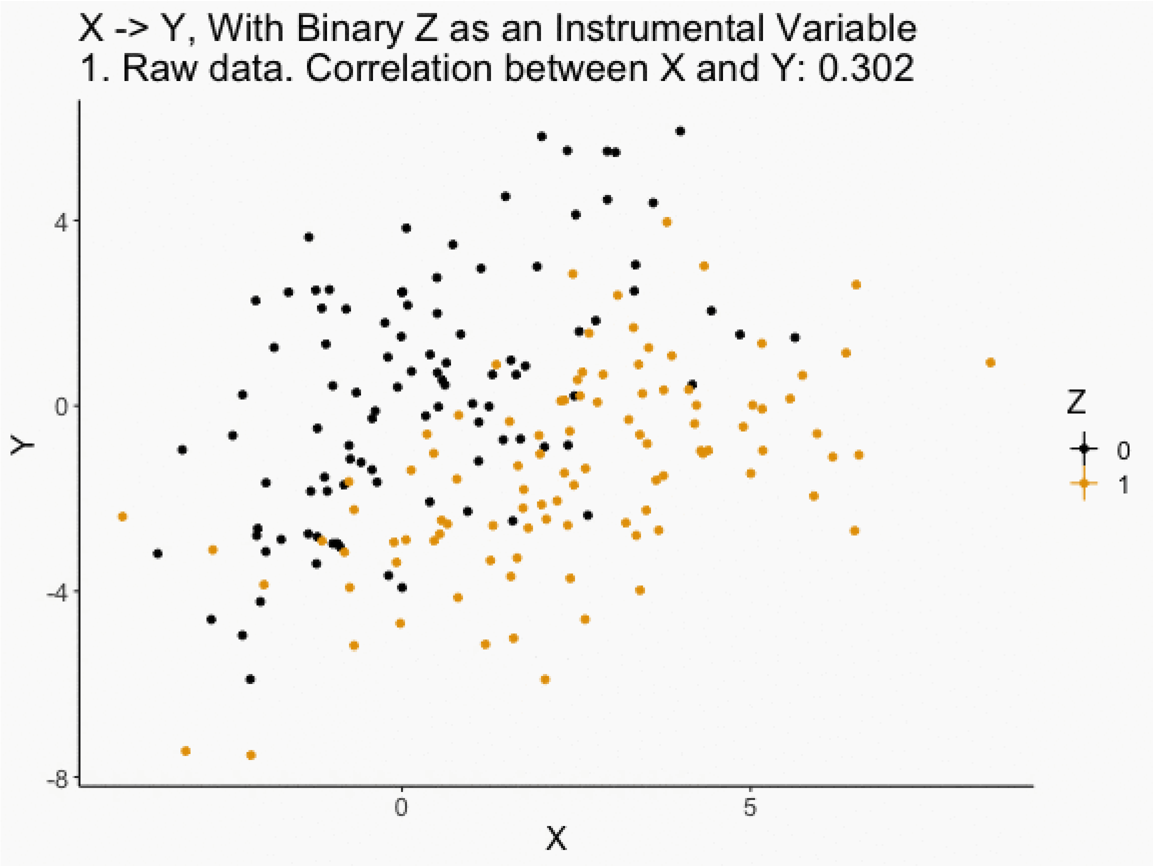
15.1 Motivation

- Omitted variables in a simple regression model

15.2 IV estimation of the multiple regression

15.3 Two-stage least squares

- A single endogenous explanatory variable 15.5 Testing for Endogeneity
- a) Testing for endogeneity
- b) Testing for overidentifying restrictions


X -> Y, With Binary Z as an Instrumental Variable
1. Raw data. Correlation between X and Y: 0.302
Y
X
Z
0
1
4
0
-4
-8
0
5

15.1 Motivation: Omitted Variables in a Simple Regression Model

The endogeneity problem is endemic in social sciences/economics:

- In many cases important personal variables cannot be observed
- The explanatory variable is caused by the dependent variable (**reverse causality**)
- In addition, measurement error of variables may also lead to endogeneity

Solutions to endogeneity problems considered so far:

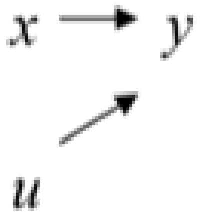
- Ignore the problem and suffer the consequences of biased and inconsistent estimators
- Proxy variables method for omitted regressors
- Fixed effects methods the following three conditions are met:
  - 1) panel data is available
  - 2) endogeneity is time-constant
  - 3) independent variables are not time-constant

Today, we will discuss the **Instrumental Variables Method (IV)**

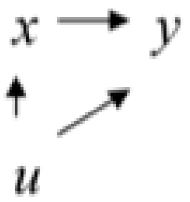
- IV is the most well-known method to address endogeneity problems.

Definitions

Exogeneity: regressors x and the error term u are independent causes of the dependent variable y



Endogeneity: the error u is affecting the regressors x and therefore indirectly affecting y



Instrumental variables: instruments z are associated with x but not with the error term u

Definitions

A regressor is endogenous when it is correlated with the error term

Example: Education in a wage equation

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 ability_i + u_i$$

If no proxy available for `ability`, then the regression:

$$\log(\text{ wage }_i) = \beta_0 + \beta_1 educ_i + u_i$$

Error term contains factors such as innate ability which are correlated with education --> estimate is biased

We can still use the equation above if we can find an **instrumental variable** (`z`) for `educ`.

Definition of a instrumental variable (z):

Requirements for instrument z:

- 1) z is not a direct cause of the dependent variable y
  - $\operatorname{cov}(y, z \mid x) = 0$($z$ is not in the y equation)

$$y = \beta_0 + \beta_1 x + u$$

- 2) **Relevance** z is correlated with the endogeneous variable x
  - $\operatorname{cov}(z, x) \neq 0$($z$ predicts or causes x)
  - we can test it by estimating a simple regression between the endogeneous variable and the instrument

$$x = \delta_0 + \delta_1 z + v$$

- 3) **Exogeneity Assumption** z is uncorrelated with the error term u
  - $\operatorname{cov}(z, u) = 0$ (z is not endogenous)
  - we *cannot* test this assumption

The instrumental variable `z` for `educ` must be

- (1) correlated with education
- (2) uncorrelated with ability (and any other unobserved factors affecting wage)

Instrument 1: `Social Security Number`

- (1) because of the randomness of the last digit of the SSN that it is not correlated with education
- (2) it is uncorrelated with ability because it is determined randomly

Instrument 2: `IQ score`

- (1) highly correlated with `educ`
- (2) but it is also highly correlated with `ability`

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 16 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

Example

score $= \beta_0 + \beta_1$ skipped $+u$

where `score` is the final exam score and `skipped` is the total number of lectures missed during the semester.

What might be a good IV for skipped?

- Instrument: distance between living quarters and classrooms

A simple consistency proof for OLS under exogeneity:

$\operatorname{Cov}(x_i, u_i) = 0$ (Exogeneity)

$$
\begin{aligned}
&\Leftrightarrow 0 = \operatorname{Cov}(x_i, y_i - \beta_0 - \beta_1 x_i) = \operatorname{Cov}(x_i, y_i) - \beta_1 \operatorname{Var}(x_i) \\
&\Leftrightarrow \beta_1 = \operatorname{Cov}(x_i, y_i) / \operatorname{Var}(x_i) \\
&\Rightarrow \widehat{\beta}_1 = \widehat{\operatorname{Cov}}(x_i, y_i) / \widehat{\operatorname{Var}}(x_i) \to \operatorname{Cov}(x_i, y_i) / \operatorname{Var}(x_i) = \beta_1
\end{aligned}
$$

This holds as long as the data are such that sample variances and covariances converge to their theoretical counterparts as n goes to infinity; i.e. if the LLN holds. OLS will basically be consistent if, and only if, exogeneity holds.

Assume existence of an instrumental variable z:

$\operatorname{Cov}(z_i, u_i) = 0 \quad (\text{ but } \operatorname{Cov}(x_i, u_i) \neq 0)$

$\Leftrightarrow 0 = \operatorname{Cov}(z_i, y_i - \beta_0 - \beta_1 x_i) = \operatorname{Cov}(z_i, y_i) - \beta_1 \operatorname{Cov}(z_i, x_i) \Leftrightarrow \beta_1 = \operatorname{Cov}(z_i, y_i) / \operatorname{Cov}(z_i, x_i)$

The instrumental variable is correlated with the explanatory variable

$$\hat{\beta}_{IV} = \frac{\widehat{\operatorname{Cov}}(z_i, y_i)}{\widehat{\operatorname{Cov}}(z_i, x_i)}$$

IV- estimator

$$\widehat{\beta}_{IV} = \frac{\sum_{i=1}^{n} (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n} (z_i - \bar{z})(x_i - \bar{x})}$$

Example: Father's education as an IV for education

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

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

Return to education probably overestimated

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

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

<u>Is the education of the father a good IV?</u>

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

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

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

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

It is also much less precisely estimated

Instrumental Variable in R

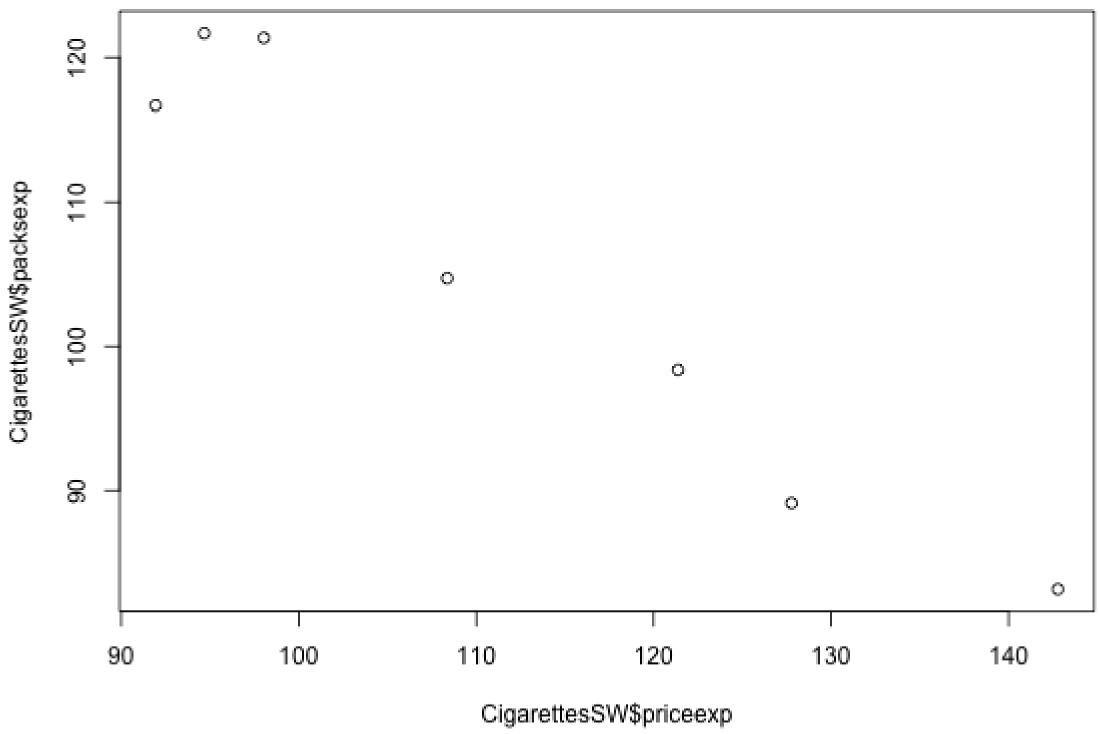
```r
library(AER)
data(CigarettesSW)

CigarettesSW <- CigarettesSW %>%
  mutate(cigtax = taxs-tax) %>%
  mutate(price = price/cpi,
         cigtax = cigtax/cpi) %>%
  group_by(cut(cigtax,breaks=7)) %>%
  summarize(priceexp = mean(price),
         packsexp = mean(packs)) %>%
  ungroup()

cor(CigarettesSW$priceexp,CigarettesSW$packsexp)
```

```
## [1] -0.9711096
```

```
plot(CigarettesSW$priceexp,CigarettesSW$packsexp)
```



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

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

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

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

IV worse than OLS if:

$$\frac{\text{Corr}(z,u)}{\text{Corr}(z,x)} > \text{Corr}(x,u)$$

e.g. $\frac{0.03}{0.2} > 0.1$

15.2 IV Estimation of the Multiple Regression Model

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

- y_2 endogeneous variable
- $z_i, i = 1 \ldots k - 1$ exogeneous variables

Conditions for instrumental variable zk:

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

This is the so called "reduced form regression"

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

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

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

we use: z_1 to indicate that this variable is exogenous y_2 to indicate that this variable is endogeneous

$$\lg(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + u_1$$

$\mathrm{y}_1 = \log(\text{ wage }), \mathrm{y}_2 =$ educ, and $\mathrm{z}_1 =$ exper

We need an instrumental variable z_2 and that the following conditions are satisfied:

$E(u_1) = 0$

$\operatorname{Cov}(z_1, u_1) = 0$

$\operatorname{Cov}(z_2, u_1) = 0$

Given the zero mean assumption, the last two assumptions are equivalent to $E(z_1 u_1) = E(z_2 u_1) = 0$

This is a set of three linear equations in the three unknowns $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\beta}_2$, and it is easily solved given the data on y_1, y_2, z_1, and z_2. The estimators are called **instrumental variables estimators**.

$$\sum_{i=1}^{n}\left(y_{i1}-\hat{\beta}_0-\hat{\beta}_1 y_{i2}-\hat{\beta}_2 z_{i1}\right)=0$$

$$\sum_{i=1}^{n} z_{i1}\left(y_{i1}-\hat{\beta}_0-\hat{\beta}_1 y_{i2}-\hat{\beta}_2 z_{i1}\right)=0$$

$$\sum_{i=1}^{n} z_{i2}\left(y_{i1}-\hat{\beta}_0-\hat{\beta}_1 y_{i2}-\hat{\beta}_2 z_{i1}\right)=0$$

We still need: $\pi_2 \neq 0$

$$y_2=\pi_0+\pi_1 z_1+\pi_2 z_2+v_2$$

15.3 Two Stage Least Squares

In the previous sections we had a single endogeneous variable (y_2) along with one instrumental variable for y_2.

It might be possible to have several instrumental variables for a single endogeneous variable.

It turns out that the IV estimator is equivalent to the following procedure, which has a much more intuitive interpretation:

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

First stage (reduced form regression):

- The endogenous explanatory variable y_2 is predicted using only exogenous information

$$\hat{y}_2 = \widehat{\pi}_0 + \widehat{\pi}_1 z_1 + \ldots + \widehat{\pi}_k z_{k-1} + \widehat{\pi}_k z_k$$

Second stage (OLS with y_2 replaced by its prediction from the first stage)

$$y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 z_1 + \ldots + \beta_k z_{k-1} + \text{ error}$$

Why does Two Stage Least Squares (2SLS) work?

All variables in the second stage regression are exogenous because y_2 was replaced by a prediction based on only exogenous information.

By using the prediction based on exogenous information, y_2 is purged of its endogenous part (the part that is related to the error term).

- If there is one endogenous variable and one instrument, then the $\mathrm{2SLS}$ estimates (replacing x with $\hat{x}$ based on z) will be the same as the IV estimates $(\operatorname{cov}(z, y) / \operatorname{cov}(z, x))$.

- The `2SLS` estimation can also be used if there is more than one endogenous variable and at least as many instruments.

2SLS in R

Common 2SLS estimators: `ivreg` in `AER`, `iv_robust` in `estimatr`, and `feols()` in `fixest`. We'll use the latter since it's fast easy to combine with fixed effects and all kinds of error adjustments

Practice:

- Reload the cigarette data and skip the summarize step
- Run our cigarette analysis first doing 2SLS by hand - use lm() to run the first stage, then replace price with predict(m) in the second stage
- Then use feols() to do the same (use 1 to indicate no controls). Coefficients should be the same but the standard errors will be corrected in the feols() version!
- Show both results in msummary()

```r
data(CigarettesSW)

# instrumental variable
CigarettesSW <- CigarettesSW %>%
  mutate(cigtax = taxs-tax) %>%
  mutate(price = price/cpi,
         cigtax = cigtax/cpi)
first_stage <- lm(price~cigtax, data = CigarettesSW)
second_stage <- lm(packs ~ predict(first_stage), data = CigarettesSW)

# 2sls
package <- feols(packs ~ 1 | price ~ cigtax, data = CigarettesSW)
```

```r
msummary(list(second_stage, package), stars = TRUE, gof_omit = 'AIC|BIC|Lik|F|R2')
```

	Model 1	Model 2
(Intercept)	219.576***	219.576***
	(20.863)	(16.989)
predict(first_stage)	−1.019***	
	(0.191)	
fit_price		−1.019***
		(0.156)
Num.Obs.	96	96
RMSE	22.80	
Std.Errors		IID
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001		

```r
library(AER)
#US income and consumption data 1950-1993
data(USConsump1993)
USC93 <- as.data.frame(USConsump1993)
#lag() gets the observation above; here the observation above is last year
IV <- USC93 %>% mutate(lastyr.invest = lag(income) - lag(expenditure))
```

```r
m <- feols(expenditure ~ 1 | income ~ lastyr.invest, data = IV, se = 'hetero')
```

Example: Returns to education for working women

Model:

$$lwage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u_1$$

- here $educ$ is endogenous and exper and $exper^2$ are exogenous
- find two instruments $fatheduc$ and $motheduc$ for $educ$

$\underline{2SLS}$ First Stage - estimate the reduced form equation:

$$educ = \delta_0 + \delta_1 exper + \delta_2 exper^2 + \delta_3 fatheduc + \delta_4 motheduc + v_2$$

Obtain the predicted values educ, which contain only exogenous information.

$\underline{2SLS}$ Second Stage - estimate the structural equation replacing educ with the predicted value $\widehat{\text{educ}}$:

$$lwage = \beta_0 + \beta_1 \widehat{\text{educ}} + \beta_2 exper + \beta_3 exper^2 + u_1$$

```r
data(mroz, package = "wooldridge")
ols<- feols(lwage ~ educ+ exper+exper^2 , data = mroz, se = 'hetero')
first_stage<- feols(educ ~ exper+exper^2 +motheduc+fatheduc, data = mroz, se = 'hetero')
#second_stage <- lm(lwage ~ predict(first_stage), data = mroz, se = 'hetero')
two_sls <- feols(lwage ~ exper+exper^2  | educ ~ motheduc+fatheduc, data = mroz, se = 'hetero')

msummary(list(ols,first_stage,two_sls), stars = TRUE, gof_omit = 'AIC|BIC|Lik|F|R2')
```

	Model 1	Model 2	Model 3
(Intercept)	−0.522**	8.367***	0.048
	(0.202)	(0.280)	(0.430)
educ	0.107***		
	(0.013)		
exper	0.042**	0.085**	0.044**
	(0.015)	(0.026)	(0.016)
I(exper^2)	−0.001+	−0.002*	−0.001*
	(0.000)	(0.001)	(0.000)
motheduc		0.186***	

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

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.

Example: 2SLS in a wage equation using two instruments

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

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

Education is significantly partially correlated with the education of the parents

Two Stage Least Squares estimation results:

$$\widehat{\log}(wage) = \underset{(.400)}{.048} + \underset{(.031)}{.061}\ educ + \underset{(.013)}{.044}\ exper - \underset{(.0004)}{.0009}\ exper^2$$

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

Detecting Weak Instruments

A small correlation between the instrument and error can lead to very large inconsistency (and therefore bias) if the instrument, z, also has little correlation with the explanatory variable x

$$\operatorname{plm} \hat{\beta}_{1, \mathrm{IV}}=\beta_{1}+\frac{\operatorname{Corr}(z, u)}{\operatorname{Corr}(z, x)} \cdot \frac{\sigma_{u}}{\sigma_{x}}$$

Staiger and Stock (1997), Stock and Yogo (2005) (SY for short) proposed methods for detecting situations where weak instruments will lead to substantial bias and distorted statistical inference.

- Single bias calculations for the instrumental variables estimator, SY recommend that one can proceed with the usual IV inference if the first-stage `t statistic` has absolute value larger than 3.2
- `2SLS`: SY rule is `F statistic` >10

15.5 Testing for Endogeneity and Testing Overidentifying Restrictions

a) Testing for Endogeneity

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

- Testing for endogeneity of y_2.
- Find two instruments z_3 and z_4 for y_2.

Estimate the reduced form equation:

$$y_2 = \delta_0 + \delta_1 z_1 + \delta_2 z_2 + \delta_3 z_3 + \delta_4 z_4 + v_2$$

- Obtain the residuals $\hat{v}_2$, which would contain the endogenous information.
  - The predicted values $\hat{y}_2$ only contains the exogenous information.
  - So the endogenous variable is broken down in exogenous part $\hat{y}_2$ and endogenous part $\hat{v}_2$, $y_2 = \hat{y}_2 + \hat{v}_2$.

Testing for Endogeneity

Estimate the structural equation with the residuals $\hat{v}_2$ included:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \gamma_1 \hat{v}_2 + u_1$$

- $\mathrm{H}_0 : \gamma_1 = 0$ (exogeneity)
- $\mathrm{H}_\mathrm{a} : \gamma_1 \neq 0$ (endogeneity)

Testing for endogeneity example

Structural equation model: lwage $= \beta_0 + \beta_1$ educ $+\beta_2$ exper $+\beta_3$ exper $^2 + u_1$

- Testing for endogeneity of `educ`
- Find two instruments `fatheduc` and `motheduc` for `educ`.

Estimate the reduced form equation: $$educ $=\delta*{0}+\delta*{1}exper+\delta*{2} exper { }^{2}+\delta*{3} fatheduc +\delta*{4} motheduc +v*{2}$$

- Obtain the residuals $\hat{v}_2$, which would contain the endogenous information.
- The predicted values $\widehat{educ}$ only contains the exogenous information.

Estimate the structural equation with the residuals $\hat{v}_2$ included:

$$lwage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \gamma_1 \hat{v}_2 + u_1$$

- $\mathrm{H}_0 : \gamma_1 = 0$ (exogeneity)
- $\mathrm{H}_\mathrm{a} : \gamma_1 \neq 0$ (endogeneity)

```r
library (AER)
library (lmtest)
data (mroz, package='wooldridge')
# restrict to non-missing wage observations
oursample <- subset (mroz, !is.na (wage))
# 1st stage : reduced form
stage1<-lm (educ ~exper+I(exper^2) +motheduc+fatheduc, data=oursample)
# 2nd stage
stage2<-lm(lwage~ educ+exper+I(exper^2) +resid(stage1), data=oursample)
# results including t tests
coeftest (stage2)
```

```
##
## t test of coefficients:
##
##                   Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)     0.04810031  0.39457526  0.1219 0.9030329
## educ            0.06139663  0.03098494  1.9815 0.0481824 *
## exper           0.04417039  0.01323945  3.3363 0.0009241 ***
## I(exper^2)     -0.00089897  0.00039591 -2.2706 0.0236719 *
## resid(stage1)   0.05816661  0.03480728  1.6711 0.0954406 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b) Testing Overidentifying Restrictions

1) Estimate the structural equation by $2SLS$ and obtain the $2SLS$ residuals, $\hat{u}_1$

2) Regress $\hat{u}_1$ on all exogenous variables. Obtain the R-squared, say, R_1^2

3) Under the null hypothesis that all IVs are uncorrelated with u_1, $nR_1^2 \overset{\text{a}}{\sim} \chi_q^2$, where q is the number of instrumental variables from outside the model minus the total number of endogenous explanatory variables.

- If nR_1^2 exceeds (say) the 5% critical value in the χ_q^2 distribution, we reject H_0 and conclude that at least some of the IVs are not exogenous

```r
library (AER)
library (lmtest)
data (mroz, package='wooldridge')
# restrict to non-missing wage observations
oursample <- subset (mroz, !is.na (wage))

res_iv <-feols(lwage ~ exper+exper^2  | educ ~ motheduc+fatheduc+huseduc, data = oursample, se = 'he

oursample <- oursample%>%
              mutate(residuals=residuals(res_iv))

reg<- feols( residuals~ exper+exper^2+motheduc+fatheduc+huseduc  , data = oursample, se = 'hetero')

#summary(reg)
```

```r
summary(reg)
```

```
## OLS estimation, Dep. Var.: residuals
## Observations: 428
## Standard-errors: Heteroskedasticity-robust
##                 Estimate Std. Error   t value Pr(>|t|)
## (Intercept)  0.00860634   0.196215  0.043862  0.96504
## exper        0.00005603   0.015314  0.003659  0.99708
## I(exper^2)  -0.00000888   0.000421 -0.021120  0.98316
## motheduc    -0.01038516   0.011393 -0.911573  0.36251
## fatheduc     0.00067344   0.010773  0.062515  0.95018
## huseduc      0.00678106   0.010973  0.618000  0.53691
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.665294   Adj. R2: -0.009212
```

$R_1^2 = .0009$ Therefore, $n_1^2 = 428(.0009) = 3852$ which is a very small value in a χ_1^2 distribution $(p - value = 535)$

Therefore, the `parents’ education` variables pass the overidentification test