IV worked examples[1]

A) Simulated exercise

B) Estimating the return to education for married women

C) Estimating the effect of smoking on birth weight

D) College Proximity as IV

[1] Wooldridge, Chapter 15.

Simulated exercise

Let us suppose the "true" data generation process is

$$y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + u \qquad u \sim N(0, \omega^2)$$

where

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right],$$

for $\rho \in [-1, 1]$ and $\rho \neq 0$. It follows that

$$\begin{aligned} x_1 | x_2 &\sim N(\rho x_2, (1 - \rho^2)) \\ x_2 | x_1 &\sim N(\rho x_1, (1 - \rho^2)) \end{aligned}$$

If the fitted model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

then OLS works beautifully!

Omitting x_2

What happens if instead the fitted model is

$$y = \beta_0 + \beta_1 x_1 + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2).$$

In this case,

$$\varepsilon = \gamma_2 x_2 + u$$

and

$$\begin{aligned} Cov(x_1, \varepsilon) &= Cov(x_1, \gamma_2 x_2 + u) \\ &= Cov(x_1, \gamma_2 x_2) + Cov(x_1, u) \\ &= \gamma_2 Cov(x_1, x_2) \\ &= \gamma_2 \rho \neq 0, \end{aligned}$$

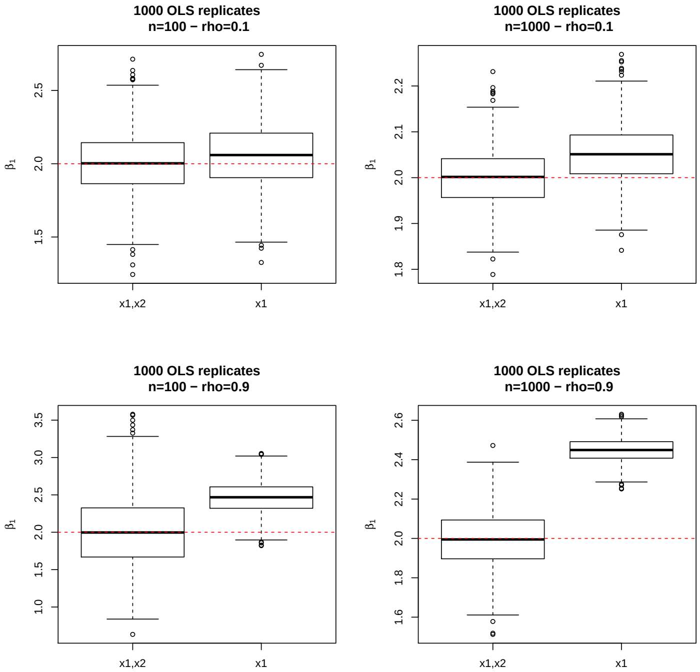
unless $\gamma_2 = 0$ and/or $\rho = 0$.

R CODE

```
install.packages("mvtnorm")
library("mvtnorm")

set.seed(234325)
omega   = 2
gamma0=1
gamma1=2
gamma2=0.5

rhos = c(0.1,0.9)
ns    = c(100,1000)
niter = 1000
coef = matrix(0,niter,2)
par(mfrow=c(2,2))
for (rho in rhos){
  V = matrix(c(1,rho,rho,1),2,2)
  for (n in ns){
    for (iter in 1:niter){
      x = rmvnorm(n,rep(0,2),V)
      error = rnorm(n,0,omega)
      y = gamma0+gamma1*x[,1]+gamma2*x[,2]+error
      coef[iter,1] = lm(y~x)$coef[2]
      coef[iter,2] = lm(y~x[,1])$coef[2]
    }
    boxplot.matrix(coef,names=c("x1,x2","x1"),ylab=expression(beta[1]))
    abline(h=gamma1,col=2,lty=2)
    title(paste("1000 OLS replicates\n n=",n," - rho=",rho,sep=""))
  }
}
```


1000 OLS replicates
n=100 − rho=0.1
1000 OLS replicates
n=1000 − rho=0.1
1000 OLS replicates
n=100 − rho=0.9
1000 OLS replicates
n=1000 − rho=0.9
β1
x1,x2
x1

B) Return to education

mroz.csv: 753 observations and 22 variables

1.	inlf	=1 if in labor force, 1975
2.	hours	hours worked, 1975
3.	kidslt6	# kids < 6 years
4.	kidsge6	# kids 6-18
5.	age	woman's age in yrs
6.	educ	years of schooling
7.	wage	estimated wage from earns., hours
8.	repwage	reported wage at interview in 1976
9.	hushrs	hours worked by husband, 1975
10.	husage	husband's age
11.	huseduc	husband's years of schooling
12.	huswage	husband's hourly wage, 1975
13.	faminc	family income, 1975
14.	mtr	fed. marginal tax rate facing woman
15.	motheduc	mother's years of schooling
16.	fatheduc	father's years of schooling
17.	unem	unem. rate in county of resid.
18.	city	=1 if live in SMSA
19.	exper	actual labor mkt exper
20.	nwifeinc	(faminc - wage*hours)/1000
21.	lwage	log(wage)
22.	expersq	exper^2

Return to education

We use the data on married working women to estimate the return to education in the simple regression model

$$\log(\texttt{wage}) = \beta_0 + \beta_1 \texttt{educ} + u.$$

OLS estimates (for comparison):

$$\widehat{\log(\texttt{wage})} = \underset{(0.185)}{-0.185} + \underset{(0.014)}{0.109}\ \texttt{educ}$$

where $n = 428$ and $R^2 = 0.118$.

The estimate for β_1 implies an almost 11% return for another year of education.

Father's education as IV

We have to maintain that:

- `fatheduc` is uncorrelated with u, and
- `educ` and `fatheduc` are correlated.

Simple regression of `educ` on `fatheduc`:

$$\widehat{\texttt{educ}} = \underset{(0.28)}{10.24} + \underset{(0.029)}{0.269}\ \texttt{fatheduc}$$

where $n = 428$ and $R^2 = 0.173$.

IV regression

Using `fatheduc` as an IV for `educ` gives

$$\widehat{\log(\texttt{wage})} = \underset{(0.446)}{0.441} + \underset{(0.035)}{0.059}\ \texttt{educ}$$

where $n = 428$ and $R^2 = 0.093$.

The IV estimate of the return to education is 5.9%, which is barely more than one-half of the OLS estimate.

This suggests that the OLS estimate is too high and is consistent with omitted ability bias.

R CODE

```
install.packages("ivpack")
library(ivpack)


data = read.csv("mroz.csv",header=TRUE)

attach(data)

n = nrow(data)

reg1 = lm(lwage~educ)

reg2 = lm(educ~fatheduc)

reg3 = ivreg(lwage ~ educ | fatheduc)
```

IV REGRESSION[2]

```
Call:
ivreg(formula = lwage ~ educ | fatheduc)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0870 -0.3393  0.0525  0.4042  2.0677

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.44110    0.44610   0.989   0.3233
educ         0.05917    0.03514   1.684   0.0929 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.6894 on 426 degrees of freedom
Multiple R-Squared: 0.09344,Adjusted R-squared: 0.09131
Wald test: 2.835 on 1 and 426 DF,  p-value: 0.09294
```

[2] This is done by two-stage least squares.

C) Effect of smoking on birth weight

bwght.csv: 1388 observations and 14 variables.

1.	faminc	1988 family income, $1000s
2.	cigtax	cig. tax in home state, 1988
3.	cigprice	cig. price in home state, 1988
4.	bwght	birth weight, ounces
5.	fatheduc	father's yrs of educ
6.	motheduc	mother's yrs of educ
7.	parity	birth order of child
8.	male	=1 if male child
9.	white	=1 if white
10.	cigs	cigs smked per day while preg
11.	lbwght	log of bwght
12.	bwghtlbs	birth weight, pounds
13.	packs	packs smked per day while preg
14.	lfaminc	log(faminc)

Effect of smoking

Suppose we estimated the effect of cigarette smoking on child birth weight:

$$\log(\texttt{bwght}) = \beta_0 + \beta_1 \texttt{packs} + u$$

where `packs` is the number of packs smoked by the mother per day.

`packs` and u might be correlated
We might worry that `packs` is correlated with other health factors or the availability of good prenatal care.

Instrument

Possible instrumental for `packs`:
Average price of cigarettes in the state of residence, `cigprice`.

We will assume that `cigprice` and u are uncorrelated (even though state support for health care could be correlated with cigarette taxes).

If cigarettes are a typical consumption good, basic economic theory suggests that `packs` and `cigprice` are negatively correlated, so that `cigprice` can be used as an IV for `packs`.

To check this, we regress `packs` on `cigprice`:

$$\widehat{\texttt{packs}} = \underset{(0.103)}{0.067} + \underset{(0.0008)}{0.0003}\ \texttt{cigprice}$$

where $n = 1,388$ and $R^2 = 0.0000$.

This indicates no relationship between smoking during pregnancy and cigarette prices, which is perhaps not too surprising given the addictive nature of cigarette smoking.

Because packs and cigprice are not correlated, we should not use cigprice as an IV for packs.

But what happens if we do? The IV results would be

$$\widehat{\log(\texttt{bwght})} = \underset{(0.91)}{4.45} + \underset{(8.70)}{2.99}\ \texttt{packs}.$$

(the reported R-squared is negative). The coefficient on packs is huge and of an unexpected sign.

The standard error is also very large, so packs is not significant.

But the estimates are meaningless because cigprice fails the one requirement of an IV that we can always test.

D) College Proximity as IV

Card (1995)[3] used wage and education data for a sample of men in 1976 to estimate the return to education.

He used a dummy variable for whether someone grew up near a four-year college (**nearc4**) as an instrumental variable for education.

In a log(wage) equation, he included other standard controls: experience, a black dummy variable, dummy variables for living in an Standard Metropolitan Statistical Area (SMSA) and living in the South, and a full set of regional dummy variables and an SMSA dummy for where the man was living in 1966.

[3] Card (1995) Using Geographic Variation in College Proximity to Estimate the Return to Schooling. In *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*, ed. Christophides, Grant and Swidinsky, 201-222. Toronto: University of Toronto Press.

card.csv: 3010 observations and 31 variables.

1.	id	person identifier
2.	nearc2	=1 if near 2 yr college, 1966
3.	nearc4	=1 if near 4 yr college, 1966
4.	educ	years of schooling, 1976
5.	age	in years
6.	fatheduc	father's schooling
7.	motheduc	mother's schooling
8.	weight	NLS sampling weight, 1976
9.	momdad14	=1 if live with mom, dad at 14
10.	sinmom14	=1 if with single mom at 14
11.	step14	=1 if with step parent at 14
12.	reg661	=1 for region 1, 1966
13.	reg662	=1 for region 2, 1966
14.	reg663	=1 for region 3, 1966
15.	reg664	=1 for region 4, 1966
16.	reg665	=1 for region 5, 1966
17.	reg666	=1 for region 6, 1966
18.	reg667	=1 for region 7, 1966
19.	reg668	=1 for region 8, 1966
20.	reg669	=1 for region 9, 1966
21.	south66	=1 if in south in 1966
22.	black	=1 if black
23.	smsa	=1 in in SMSA, 1976
24.	south	=1 if in south, 1976
25.	smsa66	=1 if in SMSA, 1966
26.	wage	hourly wage in cents, 1976
27.	enroll	=1 if enrolled in school, 1976
28.	KWW	knowledge world of work score
29.	IQ	IQ score
30.	married	=1 if married, 1976
31.	libcrd14	=1 if lib. card in home at 14
32.	exper	age - educ - 6

In order for `nearc4` to be a valid instrument, it must be uncorrelated with the error term in the wage equation – we assume this – and it must be partially correlated with `educ`.

Regression of `educ` on `nearc4` and exogenous variables:

$$\widehat{\texttt{educ}} = \underset{(0.24)}{16.64} + \underset{(0.088)}{0.320}\ \texttt{exper} - \underset{(0.034)}{0.413}\ \texttt{nearc4} + \cdots$$

where $n = 3,010$ and $R^2 = 0.477$.

In 1976, other things being fixed (experience, race, region, and so on), people who lived near a college in 1966 had, on average, about one-third of a year more education than those who did not grow up near a college.

If `nearc4` is uncorrelated with unobserved factors in the error term, we can use `nearc4` as an IV for `educ`.

TABLE 15.1 Dependent Variable: log(*wage*)		
Explanatory Variables	OLS	IV
educ	.075 (.003)	.132 (.055)
exper	.085 (.007)	.108 (.024)
$exper^2$	−.0023 (.0003)	−.0023 (.0003)
black	−.199 (.018)	−.147 (.054)
smsa	.136 (.020)	.112 (.032)
south	−.148 (.026)	−.145 (.027)
Observations	3,010	3,010
R-squared	.300	.238
Other controls: *smsa66*, *reg662*, ..., *reg669*		

Note: $\hat{\beta}_{\texttt{iv}}^{\texttt{educ}} \approx 2\hat{\beta}_{\texttt{ols}}^{\texttt{educ}}$ and $\text{se}(\hat{\beta}_{\texttt{iv}}^{\texttt{educ}}) \approx 18\text{se}(\hat{\beta}_{\texttt{ols}}^{\texttt{educ}})$.

The presence of larger confidence intervals is a price we must pay to get a consistent estimator of the return to education when we think `educ` is endogenous.

```
data = read.csv("card.csv",header=TRUE)

attach(data)

n = nrow(data)


reg1 = lm(lwage~educ+exper+expersq+black+smsa+south+smsa66+
                reg662+reg663+reg664+reg665+reg666+reg667+reg668+reg669)

reg2 = lm(educ~nearc4+exper+expersq+black+smsa+south+smsa66+
               reg662+reg663+reg664+reg665+reg666+reg667+reg668+reg669)

reg3 = ivreg(lwage ~ educ+exper+expersq+black+smsa+south+smsa66+
                     reg662+reg663+reg664+reg665+reg666+reg667+reg668+reg669 |
                     nearc4+exper+expersq+black+smsa+south+smsa66+
                     reg662+reg663+reg664+reg665+reg666+reg667+reg668+reg669)
```