Instrumental Variables| R programming language

Preliminaries

Data: MROZ.csv

Script Editor

```
# setup
rm(list = ls())
directory <- "C:/Users/amalz/OneDrive/Desktop/"

# Install packages
PackageNames <- c("tidyverse", "stargazer", "magrittr", "haven", "AER")
for(i in PackageNames){
  if(!require(i, character.only = T)){
    install.packages(i, dependencies = T)
    require(i, character.only = T)
  }
}

MROZ <- read.csv(paste0(directory, 'MROZ.csv'))

# keep only working women
MROZ %<>% filter(inlf == 1)

MROZ %>%
  select(lwage, educ, exper, expersq, fatheduc, motheduc) %>%
  stargazer(type = "text")

MROZ %>%
  select(lwage, educ, exper, expersq, fatheduc, motheduc) %>%
  head(10)

# Regression model
model1 <- lm(lwage ~ educ, MROZ)
summary(model1)
coef(model1)["educ"]
```


IV estimation

- *Dependent variable y, endogenous variable x, instrument z*
- *Coefficient_ols = sum((x-xbar)*(y-ybar))/sum((x-xbar)*(x-xbar))*
- *Coefficient_iv = sum((z-zbar)*(y-ybar))/sum((z-zbar)*(x-xbar))*

Script Editor

```
# Calculating means
attach(MROZ) # attach MROZ so the variable names can be used directly
mean_fatheduc <- mean(fatheduc)
mean_educ <- mean(educ)
mean_lwage <- mean(lwage)

# OLS coefficient on educ
numerator_ols <- (educ - mean_educ) * (lwage-mean_lwage)
denominator_ols <- (educ - mean_educ) * (educ - mean_educ)
sum_numerator_ols <- sum(numerator_ols)
sum_denominator_ols <- sum(denominator_ols)

(coeff_ols <- sum_numerator_ols/sum_denominator_ols)

# IV coefficient on educ
numerator_iv <- (fatheduc - mean_fatheduc) * (lwage - mean_lwage)
denominator_iv <- (fatheduc - mean_fatheduc) * (educ - mean_educ)

sum_numerator_iv <- sum(numerator_iv)
sum_denominator_iv <- sum(denominator_iv)

(coeff_iv <- sum_numerator_iv/sum_denominator_iv)
detach(MROZ)
```

Instrumental Variables| R programming language

Console

```
> # Calculating means
> attach(MROZ) # attach MROZ so the variable names can be used directly
> mean_fatheduc <- mean(fatheduc)
> mean_educ <- mean(educ)
> mean_lwage <- mean(lwage)
>
> # OLS coefficient on educ
> numerator_ols <- (educ - mean_educ) * (lwage-mean_lwage)
> denominator_ols <- (educ - mean_educ) * (educ - mean_educ)
> sum_numerator_ols <- sum(numerator_ols)
> sum_denominator_ols <- sum(denominator_ols)
>
> (coeff_ols <- sum_numerator_ols/sum_denominator_ols)
[1] 0.1086487
>
> # IV coefficient on educ
> numerator_iv <- (fatheduc - mean_fatheduc) * (lwage - mean_lwage)
> denominator_iv <- (fatheduc - mean_fatheduc) * (educ - mean_educ)
>
> sum_numerator_iv <- sum(numerator_iv)
> sum_denominator_iv <- sum(denominator_iv)
>
> (coeff_iv <- sum_numerator_iv/sum_denominator_iv)
[1] 0.05917348
> detach(MROZ)
```

2SLS (two stage least squares): Simple regression model with one instrument

Steps:-

1. *2SLS - first stage :- Regression of endogenous variable educ on instrument fatheduc*
2. *2SLS - second stage: Replace educ with predicted value educ_hat*

Script Editor

```
model2 <- ivreg(formula = lwage ~ educ | . - educ + fatheduc, data = MROZ)
summary(model2, diagnostics = TRUE)

model3 <- lm(educ ~ fatheduc, MROZ)
summary(model3)
MROZ <- MROZ %>% mutate(educhat = as.numeric(fitted(model3)))

model4 <- lm(lwage ~ educhat, MROZ)
summary(model4)
```

Console

```
Call:
ivreg(formula = lwage ~ educ | . - educ + fatheduc, data = MROZ)

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 .

Diagnostic tests:
                 df1 df2 statistic p-value
Weak instruments   1 426     88.84  <2e-16 ***
Wu-Hausman         1 425      2.47   0.117
Sargan             0  NA        NA      NA
---
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
```

Instrumental Variables| R programming language

Console ctd...

```
Call:
lm(formula = educ ~ fatheduc, data = MROZ)

Residuals:
    Min      1Q  Median      3Q     Max
-8.4704 -1.1231 -0.1231  0.9546  5.9546

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.23705    0.27594  37.099   <2e-16 ***
fatheduc     0.26944    0.02859   9.426   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.081 on 426 degrees of freedom
Multiple R-squared:  0.1726,    Adjusted R-squared:  0.1706
F-statistic: 88.84 on 1 and 426 DF,  p-value: < 2.2e-16
Call:
lm(formula = lwage ~ educhat, data = MROZ)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2126 -0.3763  0.0563  0.4173  2.0604

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.44110    0.46711   0.944    0.346
educhat      0.05917    0.03680   1.608    0.109

Residual standard error: 0.7219 on 426 degrees of freedom
Multiple R-squared:  0.006034,  Adjusted R-squared:  0.003701
F-statistic: 2.586 on 1 and 426 DF,  p-value: 0.1086
```

Coefficients are correct but the standard errors are not correct

2SLS (two stage least squares): Multiple regression model with several independent variables and two instruments

Steps:-

1. *2SLS - first stage : Regression of endogenous variable educ on instruments fatheduc and motheduc*
2. *Testing whether educ and fatheduc and motheduc are correlated*
3. *Predicted values for educ_hat1*
4. *2SLS - second stage: Replace endogenous variable educ with predicted value educ_hat1*

Script Editor

```
model6 <- ivreg(lwage ~ educ + exper + expersq | . - educ + fatheduc + motheduc,
                data = MROZ)
summary(model6)

model7 <- lm(educ ~ exper + expersq + fatheduc + motheduc, MROZ)
summary(model7)
car::linearHypothesis(model7, c("fatheduc = 0", "motheduc = 0"))

MROZ %<>% mutate(educ_hat1 = fitted(model7))
model8 <- lm(lwage ~ educ_hat1 + exper + expersq, MROZ)
summary(model8)
```

Instrumental Variables| R programming language

Console

```
> summary(model6)

Call:
ivreg(formula = lwage ~ educ + exper + expersq | . - educ + fatheduc +
    motheduc, data = MROZ)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0986 -0.3196  0.0551  0.3689  2.3493

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0481003  0.4003281   0.120  0.90442
educ         0.0613966  0.0314367   1.953  0.05147 .
exper        0.0441704  0.0134325   3.288  0.00109 **
expersq     -0.0008990  0.0004017  -2.238  0.02574 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6747 on 424 degrees of freedom
Multiple R-Squared: 0.1357,     Adjusted R-squared: 0.1296
Wald test: 8.141 on 3 and 424 DF,  p-value: 2.787e-05
> summary(model7)

Call:
lm(formula = educ ~ exper + expersq + fatheduc + motheduc, data = MROZ)

Residuals:
    Min      1Q  Median      3Q     Max
-7.8057 -1.0520 -0.0371  1.0258  6.3787

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.102640   0.426561  21.340  < 2e-16 ***
exper        0.045225   0.040251   1.124    0.262
expersq     -0.001009   0.001203  -0.839    0.402
fatheduc     0.189548   0.033756   5.615 3.56e-08 ***
motheduc     0.157597   0.035894   4.391 1.43e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.039 on 423 degrees of freedom
Multiple R-squared:  0.2115,    Adjusted R-squared:  0.204
F-statistic: 28.36 on 4 and 423 DF,  p-value: < 2.2e-16
> car::linearHypothesis(model7, c("fatheduc = 0", "motheduc = 0"))

Linear hypothesis test:
fatheduc = 0
motheduc = 0

Model 1: restricted model
Model 2: educ ~ exper + expersq + fatheduc + motheduc

  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1    425 2219.2
2    423 1758.6  2    460.64 55.4 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(model8)

Call:
lm(formula = lwage ~ educ_hat1 + exper + expersq, data = MROZ)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1631 -0.3539  0.0326  0.3818  2.3727

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0481003  0.4197565   0.115  0.90882
educ_hat1    0.0613966  0.0329624   1.863  0.06321 .
exper        0.0441704  0.0140844   3.136  0.00183 **
expersq     -0.0008990  0.0004212  -2.134  0.03338 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7075 on 424 degrees of freedom
Multiple R-squared:  0.04978,   Adjusted R-squared:  0.04306
F-statistic: 7.405 on 3 and 424 DF,  p-value: 7.615e-05
```

Instrumental Variables| R programming language

Testing for endogeneity

Steps:-

1. *Estimate reduced form model for education*
2. *Predict the residuals vhat*
3. *Structural equation for log wage that includes residuals vhat*

H0: coeff on vhat=0 (exogeneity) and H1: coeff on vhat ne 0 (endogeneity)

Script Editor

```
#Structural equation
model5 <- lm(lwage ~ educ + exper + expersq, MROZ)
model9 <- lm(educ ~ exper + expersq + fatheduc + motheduc, MROZ)
MROZ %<>% mutate(vhat = resid(model9))
model10 <- update(model5, ~ . + vhat)
summary(model10)
```

Console

```
Call:
lm(formula = lwage ~ educ + exper + expersq + vhat, data = MROZ)

Residuals:
     Min       1Q   Median       3Q      Max
-3.03743 -0.30775  0.04191  0.40361  2.33303

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0481003  0.3945753   0.122 0.903033
educ         0.0613966  0.0309849   1.981 0.048182 *
exper        0.0441704  0.0132394   3.336 0.000924 ***
expersq     -0.0008990  0.0003959  -2.271 0.023672 *
vhat         0.0581666  0.0348073   1.671 0.095441 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.665 on 423 degrees of freedom
Multiple R-squared:  0.1624,    Adjusted R-squared:  0.1544
F-statistic:  20.5 on 4 and 423 DF,  p-value: 1.888e-15
```

The coefficient on vhat is significant so education is endogenous.