Instrumental Variable Regression

Yalin Yang

2020-05-02

Contents

Instrumental Variable Regression

Quickview of Dataset

```
library(AER)
mroz <- foreign::read.dta("http://fmwww.bc.edu/ec-p/data/wooldridge/mroz.dta")
## Drop records without wage information
mrozwage <- subset(mroz, !is.na(wage))
```

OLS slope $\beta_1 = Cov(y,x)/Var(x)$

```
with(mrozwage, cov(log(wage),educ) / var(educ))
```

```
## [1] 0.1086487
```

OLS with linear mode

```
summary(lm(log(wage)~educ, data=mrozwage))
```

```
##
## Call:
## lm(formula = log(wage) ~ educ, data = mrozwage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.10256 -0.31473  0.06434  0.40081  2.10029
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1852     0.1852  -1.000    0.318
## educ          0.1086     0.0144   7.545 2.76e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.68 on 426 degrees of freedom
## Multiple R-squared:  0.1179, Adjusted R-squared:  0.1158
## F-statistic: 56.93 on 1 and 426 DF,  p-value: 2.761e-13
```

The correlation between the disturbuance and endogenous variable (One type of Heteroscedasticity)

In this situation,we focous on the relationship between disturbuance and independent(endogenous) variable not dependent variable

```
cor(log(mrozwage$wage) - mean(log(mrozwage$wage)), mrozwage$educ)
```

```
## [1] 0.3433404
```

Instrumental regression

```
summary(ivreg(log(wage)~educ | fatheduc, data=mrozwage))
```

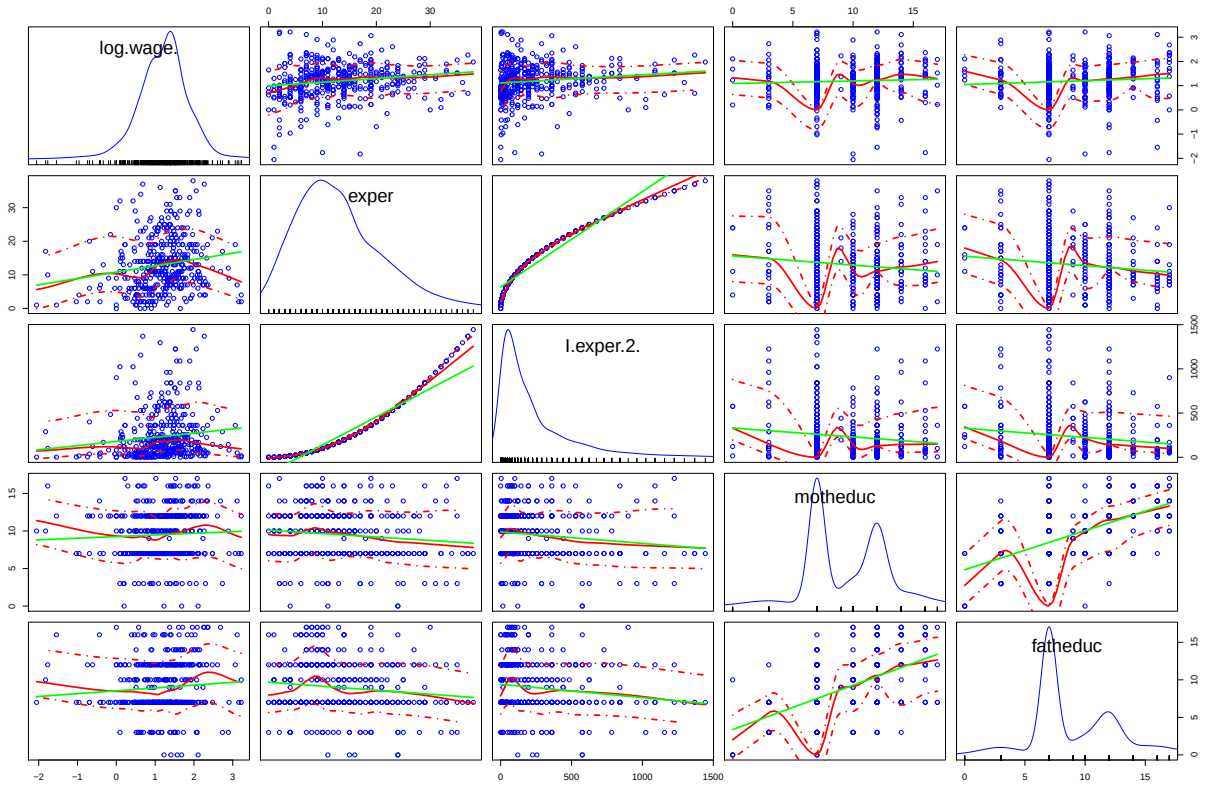
```
##
## Call:
## ivreg(formula = log(wage) ~ educ | fatheduc, data = mrozwage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0870 -0.3393  0.0525  0.4042  2.0677
##
```

```
## 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
```

Manually apply 2SLS with augmented model

- educ is an endogenous regressor
- exper is an exogenous regressor
- mother's and father's education are instruments for ability

```
scatterplotMatrix(~log(wage)+exper+I(exper^2)+motheduc+fatheduc, data=mrozwage,
                  pch=1, smooth=list(span = 0.35,lty.smooth=1, col.smooth="red", col.var="red"),
                  regLine=list(col="green"))
```



1st stage

Regression between endogenous and exogenous plus instruments [endogenous should be independent with exogenous but not instruments]

```r
stage1 <- lm(educ~exper+I(exper^2)+motheduc+fatheduc, data=mrozwage)
summary(stage1)
```

```
##
## Call:
## lm(formula = educ ~ exper + I(exper^2) + motheduc + fatheduc,
##     data = mrozwage)
##
## 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
## I(exper^2)  -0.001009   0.001203  -0.839    0.402
## motheduc     0.157597   0.035894   4.391 1.43e-05 ***
## fatheduc     0.189548   0.033756   5.615 3.56e-08 ***
## ---
## 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
```

2nd stage

With incorrect standard errors

```r
preEduc <- fitted(stage1)
# preEduc <- residuals(stage1)
stage2 <-lm(log(wage)~preEduc+exper+I(exper^2), data=mrozwage)
summary(stage2)
```

```
##
## Call:
## lm(formula = log(wage) ~ preEduc + exper + I(exper^2), data = mrozwage)
##
## 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
## preEduc      0.0613966  0.0329624   1.863  0.06321 .
## exper        0.0441704  0.0140844   3.136  0.00183 **
## I(exper^2)  -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
```

Biased OLS estimates

```
summary(lm(log(wage)~educ+exper+I(exper^2), data=mrozwage))
```

```
##
## Call:
## lm(formula = log(wage) ~ educ + exper + I(exper^2), data = mrozwage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.08404 -0.30627  0.04952  0.37498  2.37115
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.5220406  0.1986321  -2.628  0.00890 **
## educ         0.1074896  0.0141465   7.598 1.94e-13 ***
## exper        0.0415665  0.0131752   3.155  0.00172 **
## I(exper^2)  -0.0008112  0.0003932  -2.063  0.03974 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6664 on 424 degrees of freedom
## Multiple R-squared:  0.1568, Adjusted R-squared:  0.1509
## F-statistic: 26.29 on 3 and 424 DF,  p-value: 1.302e-15
```

IV Reg model

```
aut.2SLS<-ivreg(log(wage)~educ+exper+I(exper^2) |
                          motheduc+fatheduc+exper+I(exper^2) , data=mrozwage)
summary(aut.2SLS, diagnostics = TRUE)
```

```
##
## Call:
## ivreg(formula = log(wage) ~ educ + exper + I(exper^2) | motheduc +
##     fatheduc + exper + I(exper^2), data = mrozwage)
##
## 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 **
```

```
## I(exper^2)  -0.0008990  0.0004017  -2.238  0.02574 *
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   2 423    55.400  <2e-16 ***
## Wu-Hausman         1 423     2.793  0.0954 .
## Sargan             1  NA     0.378  0.5386
## ---
## 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
```

Partial F-test

Test relevance of instruments (weak instruments) with partial F-test

```
stage1.aux <- lm(educ~exper+I(exper^2), data=mrozwage)
anova(stage1.aux,stage1)
```

```
## Analysis of Variance Table
##
## Model 1: educ ~ exper + I(exper^2)
## Model 2: educ ~ exper + I(exper^2) + motheduc + fatheduc
##   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
```

Modified Hausman test

Test educ for endogeneity (see the coefficient for resid(stage1))

```
res.2SLS <- lm(log(wage)~educ+exper+I(exper^2)+resid(stage1), data=mrozwage)
coeftest(res.2SLS)
```

```
##
## t test of coefficients:
##
##                  Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)    0.04810030  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
```

Sargan test

Test of exogeneity of instruments

```
res.aux <- lm(resid(aut.2SLS) ~ motheduc+fatheduc+exper+I(exper^2) , data=mrozwage)
(r2 <- summary(res.aux)$r.squared)
```

```
## [1] 0.0008833444
```

```
(n <- nobs(res.aux))
```

```
## [1] 428
```

```
(teststat <- n*r2)
```

```
## [1] 0.3780714
```

```
(pval <- 1-pchisq(teststat, df=1))
```

```
## [1] 0.5386372
```

Chi-Square Test

χ^2 Test is very sensitive to the sample size if the sample size go larger, the significance for the Chi-Square Test would increase substantially (p-value would decrese)

Small sample size test

```
(lowCount <- matrix(c(6,8,7,10),nrow=2))
```

```
##      [,1] [,2]
## [1,]    6    7
## [2,]    8   10
```

```
(low.test <- chisq.test(lowCount, correct=F))
```

```
##
##  Pearson's Chi-squared test
##
## data:  lowCount
## X-squared = 0.0089061, df = 1, p-value = 0.9248
```

Low expected counts

```
low.test$expected
```

```
##          [,1]     [,2]
## [1,] 5.870968 7.129032
## [2,] 8.129032 9.870968
```

Get p-value through simulation

```
chisq.test(lowCount, simulate.p.value=T)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  lowCount
## X-squared = 0.0089061, df = NA, p-value = 1
```

Rescaling the sample size

```
(hiCount <- lowCount*1000)
```

```
##      [,1]  [,2]
## [1,] 6000  7000
## [2,] 8000 10000
```

chi-square increased by 1000

```
(hi.test <- chisq.test(hiCount, correct=F))
```

```
##
##  Pearson's Chi-squared test
##
## data:  hiCount
## X-squared = 8.9061, df = 1, p-value = 0.002842
```

```
hi.test$expected
```

```
##          [,1]     [,2]
## [1,] 5870.968 7129.032
## [2,] 8129.032 9870.968
```

```
chisq.test(hiCount, simulate.p.value=T)
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  hiCount
## X-squared = 8.9061, df = NA, p-value = 0.003498
```

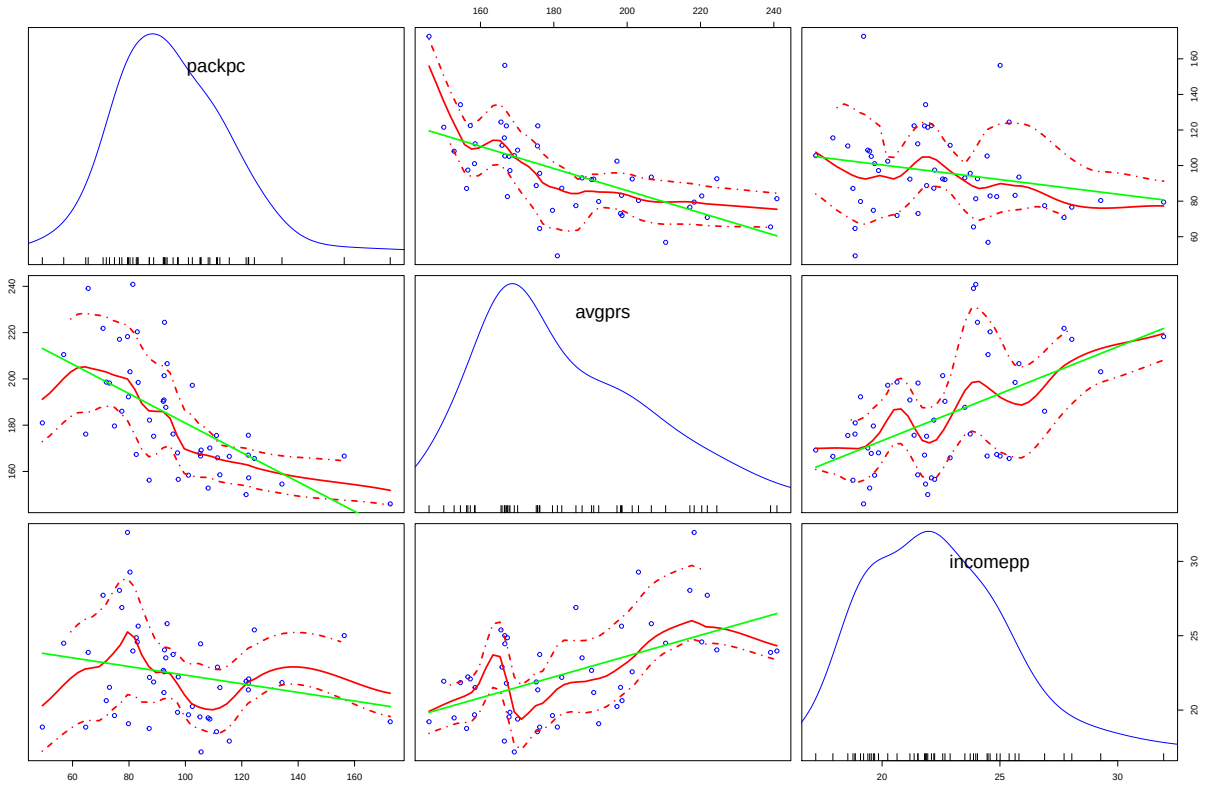
Stock & Watson Smoking Dataset Modeling

- Cross-sectional data for 48 contiguous U.S. states from 1985-1995
- packpc: average number of packs sold per capita in a year
- avgprs: average annual expenditure per person
- income: average income/pop in $1000
- taxs: proportion of sales tax on each package
- tax: general sales tax rate. May depend on incomepp

Quick view of dataset

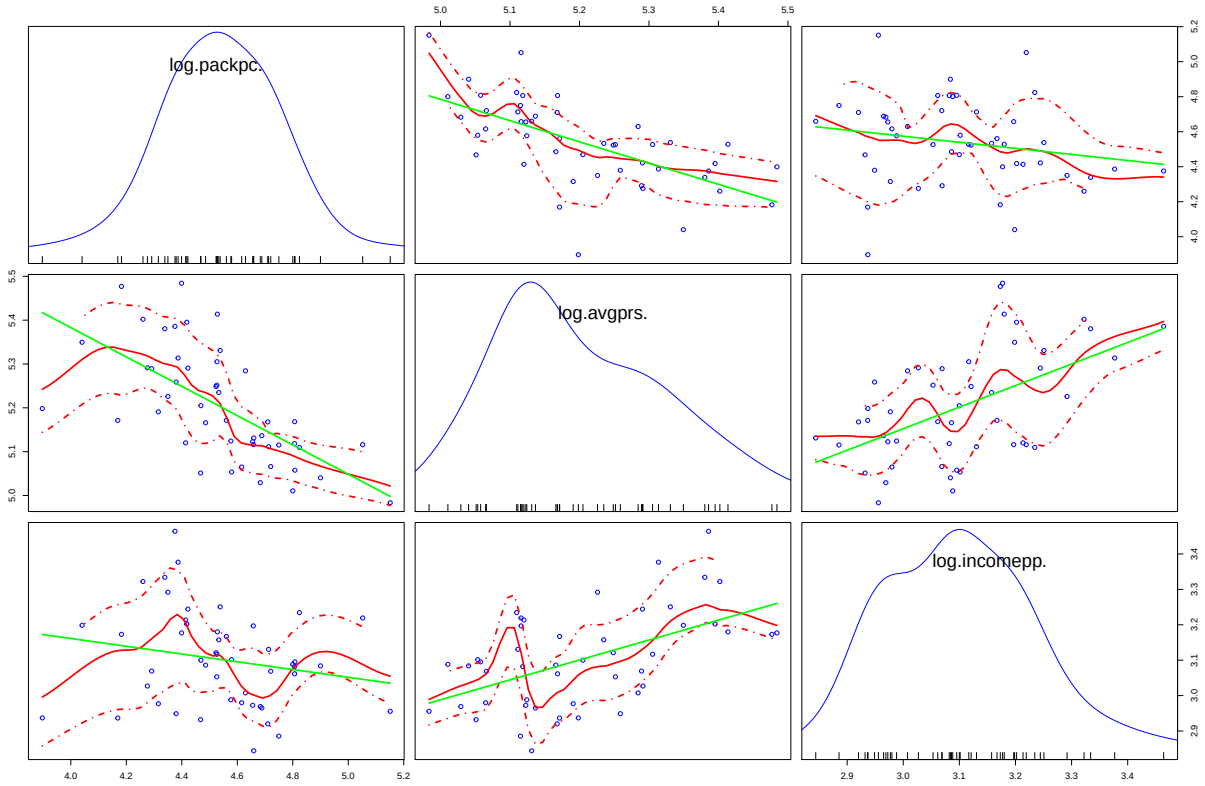
```
cig <- foreign::read.dta("http://fmwww.bc.edu/ec-p/data/stockwatson/cig85_95.dta")
cig$incomepp <- cig$income/cig$pop
cig <- cig[cig$year==1995,]

scatterplotMatrix(~packpc+avgprs+incomepp, data=cig,
                  pch=1, smooth=list(span = 0.35,lty.smooth=1, col.smooth="red", col.var="red"),
                  regLine=list(col="green"))
```



Log transformation

```
scatterplotMatrix(~log(packpc)+log(avgprs)+log(incomepp), data=cig,
                  pch=1, smooth=list(span = 0.35,lty.smooth=1, col.smooth="red", col.var="red"),
                  regLine=list(col="green"))
```



Regression Modeling

Misspecified elasticity model without income

```
cig.lm <- lm(log(packpc)~log(avgprs), data=cig)
summary(cig.lm)
```

```
##
## Call:
## lm(formula = log(packpc) ~ log(avgprs), data = cig)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.64676 -0.09030  0.01787  0.11245  0.40779
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.8500     1.1265   9.632 1.32e-12 ***
## log(avgprs)  -1.2131     0.2164  -5.604 1.13e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1896 on 46 degrees of freedom
## Multiple R-squared:  0.4058, Adjusted R-squared:  0.3928
## F-statistic: 31.41 on 1 and 46 DF,  p-value: 1.13e-06
```

Income adjusted elasticity model

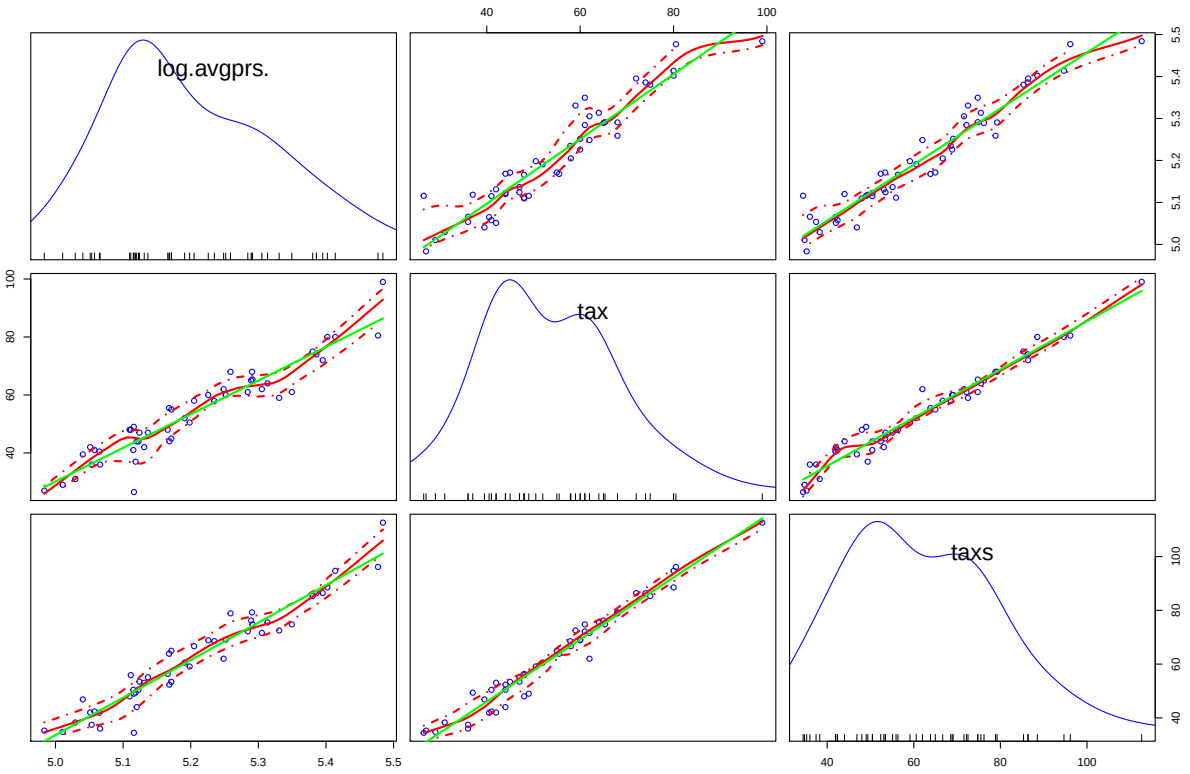
```
cig.lm <- lm(log(packpc)~log(avgprs)+log(incomepp), data=cig)
summary(cig.lm)
```

```
##
## Call:
## lm(formula = log(packpc) ~ log(avgprs) + log(incomepp), data = cig)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59077 -0.07856 -0.00149  0.11860  0.35442
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.7898     1.1135   9.690 1.38e-12 ***
## log(avgprs)    -1.4065     0.2514  -5.595 1.24e-06 ***
## log(incomepp)   0.3439     0.2350   1.463     0.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1873 on 45 degrees of freedom
## Multiple R-squared:  0.4327, Adjusted R-squared:  0.4075
## F-statistic: 17.16 on 2 and 45 DF,  p-value: 2.884e-06
```

IV Regression

Evaluate first stage of IV regression

```
scatterplotMatrix(~log(avgprs)+tax+taxs, data=cig,
         pch=1, smooth=list(span = 0.35,lty.smooth=1, col.smooth="red", col.var="red"),
                  regLine=list(col="green"))
```




```
cig.rf <- lm(log(avgprs)~tax+taxs+log(incomepp), data=cig)
summary(cig.rf)
```

```
##
## Call:
## lm(formula = log(avgprs) ~ tax + taxs + log(incomepp), data = cig)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.067411 -0.017296 -0.001123  0.023591  0.071556
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.478722   0.115423  38.803  < 2e-16 ***
## tax           -0.001009   0.001583  -0.638  0.52693
## taxs           0.007146   0.001318   5.422 2.37e-06 ***
## log(incomepp)  0.108345   0.039738   2.726  0.00916 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03226 on 44 degrees of freedom
## Multiple R-squared:  0.9403, Adjusted R-squared:  0.9363
## F-statistic: 231.1 on 3 and 44 DF,  p-value: < 2.2e-16
```

2nd estimation

```r
cig.iv <-ivreg(log(packpc)~log(avgprs)+log(incomepp)|
                    tax+taxs+log(incomepp), data=cig)
summary(cig.iv, diagnostics=T)
```

```
##
## Call:
## ivreg(formula = log(packpc) ~ log(avgprs) + log(incomepp) | tax +
##     taxs + log(incomepp), data = cig)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.6006931 -0.0862222 -0.0009999  0.1164699  0.3734227
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.3150     1.1508   8.963 1.43e-11 ***
## log(avgprs)    -1.2774     0.2632  -4.853 1.50e-05 ***
## log(incomepp)   0.2804     0.2386   1.175    0.246
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   2  44   244.734  <2e-16 ***
## Wu-Hausman         1  44     3.068  0.0868 .
## Sargan             1  NA     0.333  0.5641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1879 on 45 degrees of freedom
## Multiple R-Squared: 0.4294,  Adjusted R-squared: 0.4041
## Wald test: 13.28 on 2 and 45 DF,  p-value: 2.931e-05
```