SCRIPT MOD4S1C: MEASUREMENT ERROR AND TSLS

INSTRUCTOR: KLAUS MOELTNER

Load and Prepare Data

We will use Mroz's (1987) wage data described in Greene (6^{th} edition, p. 53. The data set contains 753 observations of labor supply behavior of married women. Of these, 428 were active labor market participants at the time of the survey.

```
R> data<- read.table('c:/Klaus/AAEC5126/R/data/laborsupply.txt', sep="\t", header=FALSE)
R> #
R> #assign variable names
R> names(data)[1]<-"lfp"
R> names(data)[2]<-"whrs"
R> names(data)[3]<-"kl6"
R> names(data)[4]<-"k618"
R> names(data)[5]<-"wa"
R> names(data)[6]<-"we"
R> names(data)[7]<-"ww"
R> names(data)[8]<-"rpwg"
R> names(data)[9]<-"hhrs"
R> names(data)[10]<-"ha"
R> names(data)[11]<-"he"
R> names(data)[12]<-"hw"
R> names(data)[13]<-"faminc"
R> names(data)[14]<-"wmed"
R> names(data)[15]<-"wfed"
R> names(data)[16]<-"un"
R> names(data)[17]<-"cit"
R> names(data)[18]<-"ax"
R> #
R> save(data, file = "c:/Klaus/AAEC5126/R/data/laborsupply.rda")
```

Variable definitions:

```
% Contents of Data (columns)
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%1       LFP = A dummy variable = 1 if woman worked in 1975, else 0
%2       WHRS = Wife's hours of work in 1975
%3       KL6 = Number of children less than 6 years old in household
%4       K618 = Number of children between ages 6 and 18 in household
%5       WA = Wife's age
%6       WE = Wife's educational attainment, in years
%7       WW = Wife's average hourly earnings, in 1975 dollars
%8       RPWG = Wife's 1976 reported wage (not = 1975 estimated wage)
%9       HHRS = Husband's hours worked in 1975
%10      HA = Husband's age
```

```
%11      HE = Husband's educational attainment, in years
%12      HW = Husband's wage, in 1975 dollars
%13      FAMINC = Family income, in 1975 dollars
%14      WMED = Wife's mother's educational attainment, in years
%15      WFED = Wife's father's educational attainment, in years
%16      UN = Unemployment rate in county of residence, in percentage points.
%17      CIT = Dummy variable = 1 if live in large city (SMSA), else 0
%18      AX = Actual years of wife's previous labor market experience


R> #select only cases for labor market participants
R> data<-subset(data,lfp==1)
R> attach(data)
R> #
R> #define variables of interest
R> y=log(whrs*ww) #log of annual earnings
R> n<-length(y)
R> wa2<-wa^2
```

You are interested in how a woman's education affects her annual earnings. Thus, "wife's educational attainment(we)" is one of your key regressors. However, you are worried that this variable is measured with error since most questionnaires were answered by husbands who are often ill-informed about the schooling history of their spouse.

Your strategy is to first run a basic OLS model, then an IV model using "husband's educational attainment (he)" and wife's parents' educational attainment as instruments for "we", arguing that spouses generally match along education levels and kids "inherit" education levels from their parents. Moreover, neither of these instruments ought to be directly related to the wife's annual earnings. The sample correlation between "we" and the 3 instruments is between 0.4 and 0.6 - strong enough to make this worth a try. After running OLS and TSLS, you will perform a Hausman and a Wu test to check if your worries were justified.

Simple OLS

```
R> X<-cbind(rep(1,n),wa, wa2, kl6, k618, hw, ax,we)
R> k<-ncol(X)
R> #
R> bols<-solve((t(X)) %*% X) %*% (t(X) %*% y)
R> e<-y-X%*%bols
R> SSR<-(t(e)%*%e)
R> s2<-(t(e)%*%e)/(n-k)
R> s2ols<-s2 #needed for Hausman test below
R> Vb<-s2[1,1]*solve((t(X))%*%X)
R> se=sqrt(diag(Vb))
R> tval=bols/se
R> #
R> tt<-data.frame(col1=c("constant","wa","wa2","kl6","k618","hw","ax","we"),
                  col2=bols,
                  col3=se,
                  col4=tval)
R> colnames(tt)<-c("variable","estimate","s.e.","t")
```

TABLE 1. OLS output

variable	estimate	s.e.	t
constant	4.287	1.683	2.547
wa	0.151	0.079	1.911
wa2	-0.002	0.001	-2.326
kl6	-0.620	0.148	-4.192
k618	-0.116	0.047	-2.475
hw	0.005	0.016	0.328
ax	0.052	0.008	6.614
we	0.067	0.025	2.673

TSLS, 3 INSTRUMENTS

```
R> # Build instrument matrix
R> Z<-cbind(rep(1,n),wa, wa2, kl6, k618, hw, ax, he, wmed, wfed)
R> int<- solve(t(Z) %*% Z) %*% t(Z) %*% X    #interim regression results - capture
R> # to gain more insights on quality of instruments, int will be 10 by 8
R> Xhat<-Z %*% int    #will be n by 8, same as X
R> k<-ncol(Xhat)  #Don't forget to update k!
R> #
R> btsls<-solve((t(Xhat)) %*% Xhat) %*% (t(Xhat) %*% y)
R> e<-y-X%*%btsls        #NOTE: use original X to compute residuals
R> SSR<-(t(e)%*%e)
R> s2<-(t(e)%*%e)/n
R> Vb<-s2[1,1]*solve((t(Xhat))%*%Xhat)
R> se=sqrt(diag(Vb))
R> tval=btsls/se
R> #
R> ttz<-data.frame(col1=c("constant","wa","wa2","kl6","k618","hw","ax","he","wmed","wfed"),
                col2=int[,ncol(int)])
R> colnames(ttz)<-c("Z regressors","we")
R> #
R> tt<-data.frame(col1=c("constant","wa","wa2","kl6","k618","hw","ax","we"),
                col2=btsls,
                col3=se,
                col4=tval)
R> colnames(tt)<-c("variable","estimate","s.e.","t")
```

TABLE 2. TSLS first-stage output

Z regressors	we
constant	1.613
wa	0.199
wa2	-0.002
kl6	0.529
k618	-0.129
hw	0.054
ax	0.018
he	0.347
wmed	0.121
wfed	0.099

TABLE 3. TSLS output

variable	estimate	s.e.	t
constant	4.826	1.722	2.803
wa	0.150	0.079	1.908
wa2	-0.002	0.001	-2.344
kl6	-0.590	0.149	-3.966
k618	-0.125	0.047	-2.659
hw	0.014	0.017	0.815
ax	0.053	0.008	6.719
we	0.023	0.041	0.568

HAUSMAN TEST, 3 INSTRUMENTS

```
R> d<-btsls-bols
R> W<-solve(t(Xhat) %*% Xhat)- solve(t(X) %*% X)
R> H<-(t(d) %*% pseudoinverse(W) %*% d)/s2ols[1,1] #note use of OLS s2
R> J<-1
R> pval=1-pchisq(H,J)
```

The Hausman test statistic is 1.7428. The corresponding p-value is 0.1868.

WU TEST, 3 INSTRUMENTS

```
R> # Step 1: regress dpi on Z and capture predicted values
R> wehat<- Z %*% solve(t(Z) %*% Z) %*% t(Z) %*% we
R> #
R> # Step 2: add predicted values to original regression
R> XWu<-cbind(X,wehat)
R> k<-ncol(XWu)
R> bwu<-solve((t(XWu)) %*% XWu) %*% (t(XWu) %*% y)# compute OLS estimator
R> e<-y-XWu%*%bwu # Get residuals.
```

```
R> s2<-(t(e)%*%e)/(n-k) #get the regression error (estimated variance of "eps").
R> Vb<-s2[1,1]*solve((t(XWu))%*%XWu) # get the estimated VCOV matrix of bols
R> #
R> # Step 3: Perform F-test
R> Rmat<-matrix(c(0,0,0,0,0,0,0,0,1),nrow=1)
R> q<- 0
R> J<-nrow(Rmat)
R> b<-bwu
R> Fstat<-(1/J)* t(Rmat %*% b-q) %*% solve(Rmat%*%Vb%*%t(Rmat))%*%(Rmat%*%b-q)
R> pval<-1-pf(Fstat,J,n-k)
```

The Wu test statistic is 1.7459. The corresponding p-value is 0.1871.

TSLS, 1 Instrument

After inspecting the first stage results you conclude that "wmed" and, especially, "wfed" are rather poor instruments, explainig little of the observed variability in "we". You try TSLS again, this time using only "he" as instrument.

```
R> # Build instrument matrix
R> Z<-cbind(rep(1,n),wa, wa2, kl6, k618, hw, ax, he)
R> int<- solve(t(Z) %*% Z) %*% t(Z) %*% X    #interim regression results - capture
R> # to gain more insights on quality of instruments, int will be 10 by 8
R> Xhat<-Z %*% int    #will be n by 8, same as X
R> k<-ncol(Xhat)  #Don't forget to update k!
R> #
R> btsls<-solve((t(Xhat)) %*% Xhat) %*% (t(Xhat) %*% y)
R> e<-y-X%*%btsls
R> SSR<-(t(e)%*%e)
R> s2<-(t(e)%*%e)/n
R> Vb<-s2[1,1]*solve((t(Xhat))%*%Xhat)
R> se=sqrt(diag(Vb))
R> tval=btsls/se
R> #
R> ttz<-data.frame(col1=c("constant","wa","wa2","kl6","k618","hw","ax","he"),
                col2=int[,ncol(int)])
R> colnames(ttz)<-c("Z regressors","we")
R> #
R> tt<-data.frame(col1=c("constant","wa","wa2","kl6","k618","hw","ax","we"),
                col2=btsls,
                col3=se,
                col4=tval)
R> colnames(tt)<-c("variable","estimate","s.e.","t")
```

TABLE 4. TSLS first-stage output

Z regressors	we
constant	3.691
wa	0.175
wa2	-0.002
kl6	0.479
k618	-0.168
hw	0.057
ax	0.015
he	0.416

TABLE 5. TSLS output

variable	estimate	s.e.	t
constant	4.688	1.740	2.694
wa	0.151	0.079	1.915
wa2	-0.002	0.001	-2.347
kl6	-0.598	0.149	-4.003
k618	-0.123	0.047	-2.600
hw	0.012	0.018	0.664
ax	0.053	0.008	6.705
we	0.034	0.046	0.741

HAUSMAN TEST, 1 INSTRUMENT

```
R> d<-btsls-bols
R> W<-solve(t(Xhat) %*% Xhat)- solve(t(X) %*% X)
R> H<-(t(d) %*% pseudoinverse(W) %*% d)/s2ols[1,1]    #note use of OLS s2
R> J<-1
R> pval=1-pchisq(H,J)
```

The Hausman test statistic is 0.6653. The corresponding p-value is 0.4147.

WU TEST, 1 INSTRUMENT

```
R> # Step 1: regress dpi on Z and capture predicted values
R> wehat<- Z %*% solve(t(Z) %*% Z) %*% t(Z) %*% we
R> #
R> # Step 2: add predicted values to original regression
R> X<-cbind(X,wehat)
R> k<-ncol(X)
R> bwu<-solve((t(X)) %*% X) %*% (t(X) %*% y)# compute OLS estimator
R> e<-y-X%*%bwu # Get residuals.
R> s2<-(t(e)%*%e)/(n-k) #get the regression error (estimated variance of "eps").
R> Vb<-s2[1,1]*solve((t(X))%*%X) # get the estimated VCOV matrix of bols
R> #
```

```
R> # Step 3: Perform F-test
R> Rmat<-matrix(c(0,0,0,0,0,0,0,0,1),nrow=1)
R> q<- 0
R> J<-nrow(Rmat)
R> b<-bwu
R> Fstat<-(1/J)* t(Rmat %*% b-q) %*% solve(Rmat%*%Vb%*%t(Rmat))%*%(Rmat%*%b-q)
R> pval<-1-pf(Fstat,J,n-k)
```

The Wu test statistic is 0.6647. The corresponding p-value is 0.4154.

```
R> proc.time()-tic
   user  system elapsed
   0.23    0.14    0.38
```