IV Estimation Using Stata – A Very Basic Introduction

The Stata dataset CARD.DTA contains data on a sample of 3010 working men aged between 24 and 34 who were part of the 1976 wave of the US National Longitudinal Survey of Young Men. This dataset was used to estimate earnings equations by D. Card (1995), "Using geographic variation in college proximity to estimate the return to schooling", published in L. Christophides, E. Grant and R. Swidinsky (eds.), *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*.

The dataset can be downloaded from

http://www.nuffield.ox.ac.uk/teaching/economics/bond/

or from

http://www.stata.com/texts/eacsap/

A working paper version of Card's paper is available at

http://emlab.berkeley.edu/users/card/papers/geo_var_schooling.pdf

The example below is considered in more detail in Problem 5.4 of J. Wooldridge (2002), *Econometric Analysis of Cross Section and Panel Data*.

The dataset includes a measure of the log of hourly wages in 1976 (lwage), and measures of years of education (educ), years of labour market experience (exper), the square of years of labour market experience (expersq), a dummy variable equal to one if the individual is married (married), a dummy variable equal to one if the individual is black (black), a dummy variable equal to one if the individual lives in the south in 1976 (south76), a dummy variable equal to one if the individual lives in a standard metropolitan statistical area in 1976 (smsa), a dummy variable equal to one if the individual lived in a standard metropolitan statistical area in 1966 (smsa66), and nine regional dummy variables indicating where the individual lived in 1966 (reg661 – reg669).

The dataset also includes a dummy variable equal to one if the individual lived close to a college that offered 4 year courses in 1966 (nearc4). Card's suggestion was to use proximity to a college as an instrumental variable for years of education – the idea being that, all else equal, individuals are less likely to choose college education if they live a long way from a suitable college.

Opening the dataset

The dataset is already in Stata format. It can be opened in various ways, for example by opening Stata and using File Open to locate and open the dataset; or by double clicking the file in My Computer. If the dataset is opened correctly, a list of the variable names should appear in the variables window.

OLS regression

regress is Stata’s basic command to compute OLS estimates.

```
        regress lwage educ exper expersq black south smsa reg661 reg662 reg663
reg664 reg665 reg666 reg667 reg668 smsa66
```

Most estimation commands in Stata have a similar format. The first variable in the variable list is treated as the dependent variable. The remaining variables are the explanatory variables. The default is to include an intercept. The default standard errors reported are not robust to heteroskedasticity.

This command should reproduce the results in Table 2, column (2) of Card (1995).

Heteroskedasticity-robust standard errors

These can be obtained using the , vce(robust) option.

```
        regress lwage educ exper expersq black south smsa reg661 reg662 reg663
reg664 reg665 reg666 reg667 reg668 smsa66, vce(r)
```

In this example the robust standard errors are similar to the default standard errors, suggesting that heteroskedasticity may not be important in this model.

2SLS estimation

ivregress is Stata’s basic command to compute IV estimates.

To use ivregress to compute 2SLS estimates, we use the 2sls option, so that the command is in effect ivregress 2sls. There are also options to compute GMM and LIML (limited information maximum likelihood) estimates.

To use nearc4 as an instrumental variable for educ in the earnings equation, treating all the remaining explanatory variables as being uncorrelated with the error term in the earnings equation, the syntax would be

```
        ivregress 2sls lwage (educ=nearc4) exper expersq black south smsa reg661
reg662 reg663 reg664 reg665 reg666 reg667 reg668 smsa66
```

Placing the educ variable in brackets tells Stata that we want to treat this variable as endogenous. Typing nearc4 after the = sign tells Stata that we want to use nearc4 as an instrumental variable for educ.

This should reproduce the coefficient on the educ variable reported in Table 3, column (5) of Card (1995).

To inspect the first stage regression, we can use

```
        reg educ exper expersq black south smsa reg661 reg662 reg663 reg664 reg665
reg666 reg667 reg668 smsa66 nearc4
```

This should reproduce the coefficient on the nearc4 variable reported in Table 3, column (1) of Card (1995). Note that individuals who lived near to a college were significantly more likely to choose additional education, even controlling for all the other explanatory variables included in this earnings equation.

Note that using OLS to estimate the second stage regression explicitly, using

```
        predict educhat

        reg lwage educhat exper expersq black south smsa reg661 reg662 reg663
reg664 reg665 reg666 reg667 reg668 smsa66
```

produces the same estimated coefficients as we obtain with the ivregress 2sls command, but not the same standard errors.

Typing a list of variable names after the = sign inside the brackets would tell Stata to use more than one variable as an instrument.

Placing more than one of the explanatory variables before the = sign inside the brackets would specify more than one of the explanatory variables to be endogenous.

See help ivregress for more details.

ivreg2 (if installed) provides a more powerful alternative to the ivregress command. As well as allowing GMM, LIML and other related estimators to be computed, ivreg2 also provides (robust) tests of over-identifying restrictions, and implements various tests for under-identification or for weak identification/weak instruments.

See help ivreg2 for more details (provided the ivreg2 command is installed).