Stata Lab 3 – Instrumental Variables

The content of this section is taken directly from the Empitical Problem Set #2 of Glenn Ellison and Stephen Ryan's Graduate Industrial Organization course at MIT. The intent of this exercise is to get students familiar with Stata through estimating demand models. After finishing this exercise, students should be familiar with how to load data sets into Stata, manipulate the variables in basic ways, and perform simple instrumental variables regressions.

Demand for Broilers

Step 1: Download the data sets from your TU Moodle. The first data set is quantity, price, cost, and demographic variables on broiler chickens over 40 years in the United States, and is called "broiler.csv". The data is taken from Dennis Epple and Bennett McCallum's paper: "Simultaneous Equation Econometrics: The Missing Example." The data is in comma delimited format, a common format that is supported by just about every data processing application out there. It is completely portable across operating systems, and has the additional benefit of being human-readable.

Step 2: Use excel to open the data. You will see the following column headers: year, q, y, pchick, pbeef, pcor, pf, cpi, qproda, pop, meatex, and time. The cryptic names are common in empirical work, even appearing in a dataset that is intended to be used for instructional purposes. Here are the variable definitions:

y	is percapita real disposable income
pchick	is price of chicken
pbeef	is price of beef
pcor	is price of corn
pf	is price of feed
cpi	is consumer price index
proda	is aggregate production of chicken in pounds
pop	is population of the US
ex	is export of beef, veal and pork in pounds

Step 3: Upload the this data file to Stata. Select File -> Import -> ASCII data created by a spreadsheet -> browse (then direct the program to the location of the file broiler.csv) -> OK.

or you can type

```
insheet using "C:\Users\user\Downloads\broiler.csv"
```

Step 4: Explore the data by typing the following commands in the Command window.

```
summarize
tabulate year
histogram q
histogram y
histogram pchick
correlate
(Give pairwise correlation of each pair of variables)
correlate q pchick
gen log_pchick = log(pchick)
gen log_q = log(q)
regress q pchick
```

(This estimates coefficients in $q = \beta_0 + \beta_1 pchick + \varepsilon$)

regress q pchick pbeef

(This estimates coefficients in $q = \beta_0 + \beta_1 pchick + \beta_2 pbeef + \varepsilon$. The first variable is always the depedent (y) variable.)

Step 5: Try estimating a few different specifications of demand model. What should be included as explanatory (X) variables?

- Price of Chicken
- Other demand shifters

Step 6: Try estimating the coefficients in this model

$$q = \beta_0 + \beta_1 pchick + \varepsilon$$

```
. reg q pchick
```

Source	SS	df	MS	Number of obs =	40
				F(1, 38) =	667.75
Model	3616.86534	1	3616.86534	Prob > F =	0.0000
Residual	205.825702	38	5.41646586	R-squared =	0.9462
				Adj R-squared =	0.9447
Total	3822.69105	39	98.0177191	Root MSE =	2.3273

q	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pchick	.2515834	.0097358	25.84	0.000	.2318742	.2712926
_cons	10.19811	.9859519	10.34	0.000	8.202155	12.19407

The reason why the coefficient β_1 is positive is because we have not included other variables that explains q. Coincedently, those variables must be correlating with *pchick*. Thus, β_1 picks up effects of those omitted variables. This is called the "omitted variable bias" (see more in Wooldridge (2002), Wooldrige (2008) or any standard econometrics textbook.)

Step 7: We discussed in class that most empirical demand estimations would suffer from the simultaneity bias between quantity and price. Now, try fixing the simultaneity problem using "pcor" and "pf" as instrumental variables for "pchick". Why do you think "pcor" and "pf" are appropriate instrumental variables?

```
. ivreg q (pchick = pcor pf)

Instrumental variables (2SLS) regression
```

Source	SS	df	MS	Number of obs =	40
				F(1, 38) =	572.71
Model	3616.01136	1	3616.01136	Prob > F =	0.0000
Residual	206.679685	38	5.43893907	R-squared =	0.9459
				Adj R-squared =	0.9445
Total	3822.69105	39	98.0177191	Root MSE =	2.3322

q	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pchick	.2477176	.0103511	23.93	0.000	.2267628	.2686724
_cons	10.56131	1.040077	10.15	0.000	8.455786	12.66684

Instrumented: pchick
Instruments: pcor pf

Step 8: The pchick coefficient is still positive – we have not fixed the omitted variable bias. Perhaps, adding price of other food products would help. Now, try adding price of beef to the model.

```
. ivreg q (pchick = pcor pf) pbeef

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  2,    37) =  369.13
       Model |  3642.69994     2  1821.34997            Prob > F      =  0.0000
    Residual |  179.991107    37  4.86462451            R-squared     =  0.9529
-------------+------------------------------           Adj R-squared =  0.9504
       Total |  3822.69105    39  98.0177191            Root MSE      =  2.2056

------------------------------------------------------------------------------
           q |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      pchick |   .1361941   .1482062     0.92   0.364    -.1641002    .4364883
       pbeef |   .1134975   .1405136     0.81   0.424    -.1712102    .3982051
       _cons |   11.55708   2.357234     4.90   0.000     6.780868    16.33329
------------------------------------------------------------------------------
Instrumented:  pchick
Instruments:   pbeef pcor pf
------------------------------------------------------------------------------
```

Step 9: The pchick coefficient is still positive – we have not added enough control variables. Now, try adding population, income and CPI to the model.

```
. ivreg q (pchick = pcor pf) pbeef pop y cpi

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =      40
-------------+------------------------------           F(  5,    34) =  136.22
       Model |  3634.1562      5  726.831241            Prob > F      =  0.0000
    Residual |  188.534843    34  5.54514245            R-squared     =  0.9507
-------------+------------------------------           Adj R-squared =  0.9434
       Total |  3822.69105    39  98.0177191            Root MSE      =  2.3548

------------------------------------------------------------------------------
           q |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      pchick |  -.4139693   .2261444    -1.83   0.076      -.87355    .0456115
       pbeef |  -.0191511   .0614266    -0.31   0.757     -.143985    .1056827
         pop |   .2276305   .1516255     1.50   0.143    -.0805096    .5357705
           y |   .0007828   .0011096     0.71   0.485    -.0014722    .0030377
         cpi |   .3532133   .1602303     2.20   0.034     .0275862    .6788404
       _cons |  -19.89407   21.34878    -0.93   0.358    -63.28002    23.49188
------------------------------------------------------------------------------
Instrumented:  pchick
Instruments:   pbeef pop y cpi pcor pf
------------------------------------------------------------------------------
```

What do you think of this results? Why do you think we now have the right sigh for pchick's coefficient?

Step 10: to perform the overidentifying restrictions, we type
ivregress 2sls q (pchick = pcor pf) pbeef pop y cpi
estat overid

Step 11: We can check for the endogeneity of "pchick"
ivregress 2sls q (pchick = pcor pf) pbeef pop y cpi
estat endogenous pchick