Econ 140 - Spring 2016
Section 11

GSI: Fenella Carpena

April 21, 2016

1 IV Regression: Review the Concepts

Exercise 1.1. (Stock & Watson, Review the Concepts 12.1) Consider the problem of estimating the elasticity of demand for butter using the demand equation

$$ln(Q_i^{butter}) = \beta_0 + \beta_1 ln(P_i^{butter}) + u_i.$$

In this regression model, is $ln(P_i^{butter})$ positively or negatively correlated with the error u_i? If β_1 is estimated by OLS, would you expect the estimated value to be larger or smaller than the true value of β_1? Explain.

Exercise 1.2. (Stock & Watson, Review the Concepts 12.2) Consider the problem of estimating the elasticity of demand for cigarettes using the demand equation

$$ln(Q_i^{cigs}) = \beta_0 + \beta_1 ln(P_i^{cigs}) + u_i.$$

Suppose that we used as an instrument for $ln(P_i^{cigs})$ the number of trees per capita in the state. Is this instrument relevant? Is it exogenous? Is it a valid instrument?

Exercise 1.3. (Adapted from Stock & Watson, Review the Concepts 12.3) Consider a study on the effect of incarceration (imprisonment) on crime rates. Specifically, we want to examine whether putting criminals in jail reduces crime.

(a) One strategy for estimating this effect is to regress *crime_rates* (crimes per 100,000 member of the general population against *incarceration_rate* (prisoners per 100,000) using annual data from U.S. states. Explain why this regression is subject to bias.

(b) Suppose that we use the number of lawyers per capita as an instrument for *incarceration_rate*. Would this instrument be relevant? Would it be exogenous? Would it be a valid instrument?

Exercise 1.4. (Adapted from Stock & Watson, Review the Concepts 12.4) Does a new medical procedure (in this example, cardiac catheterization) prolong lives?

(a) Suppose we answer the above question by comparing patients who received the treatment to those who did not. This leads to regressing the length of survival of the patient, *months_survived*, on a binary variable for whether the patient received the procedure, *got_treatment*. Explain why this regression is subject to bias.

(b) Suppose that we use as an instrument for *got_treatment*, the difference between the distance from the patient's home to the nearest cardiac catheterization hospital, and the distance to the nearest hospital of any sort (which did not offer the treatment); this instrument takes on the value zero if the nearest hospital is a cardiac catheterization hospital, and otherwise it is positive. How could you determine whether this instrument is relevant? How could you determine whether this instrument is exogenous?

Exercise 1.5. (Stock & Watson, Exercise 12.7) In an IV regression model with one regressor, X_i, and two instruments, Z_{1i} and Z_{2i}, the value of the J-statistic is $J = 18.2$.

(a) Does this suggest that $E(u_i|Z_{1i}, Z_{2i}) \neq 0$? Explain.

(b) Does this suggest that $E(u_i|Z_{1i}) \neq 0$? Explain.

Exercise 1.6. (Stock & Watson, Exercise 12.9) A researcher is interested in the effect of military service on human capital. He collects data from a random sample of 4000 workers aged 40 and runs the OLS regression $Y_i = \beta_0 + \beta_1 X_i + u_i$, where Y_i is the worker i's annual earnings and X_i is a binary variable that is equal to 1 if the person served in the military and 0 otherwise.

(a) Explain why the OLS estimates are likely to unreliable. (*Hint:* Which variables are omitted from the regression? Are they correlated with military service?)

(b) During the Vietnam War there was a draft, where priority for the draft was determined by a national lottery. (The days of the year were randomly reordered 1 through 365. Those with birthdates ordered first were drafted before those with birthdates ordered second, and so forth.) Explain how the lottery might be used as an instrument to estimate the effect of military service on earnings.

2 IV Regression: Stata Example

Exercise 2.1. (Adapted from Stock & Watson, Empirical Exercise 12.1) In this exercise, we will use Stata to estimate the effect of fertility on women's labor supply. The data set `fertility.dta` contains information on married women aged 21-35 with two or more children. The variables we will use are:

- `weeksworked`: mom's weeks worked in 1979
- `morekids`: dummy variable equal to 1 if mom had more than 2 kids
- `twoboys`: dummy variable equal to 1 if mom's first two kids are boys
- `twogirls`: dummy variable equal to 1 if mom's first two kids are girls
- `age`: age of mom at 1980 census

Using these data, we are interested in understanding the following question: how much does a woman's labor supply fall when she has an additional child? Hence, the regression we want to estimate is:

$$weeksworked = \beta_0 + \beta_1 morekids + \beta_2 age + u$$

(a) Using Stata, implement an IV regression of `weeksworked` on `morekids` and `age`, and using `twoboys` and `twogirls` as instruments for `morekids`, by estimating each of the two stages using OLS. What are Y, W, X, and Z here?

(b) Perform the same IV regression as in part (a) but using Stata's `ivregress` command. Verify that the coefficients are the same you obtained are the same as in part (a), but the standard errors are different; why is this the case?

(c) For the Stata command in part (b):

   (i) Explain each part of this Stata command.

   (ii) How many instruments do we have? How many endogenous regressors? Is the regression overidentified, underidentified, or exactly identified?

(d) Using Stata, implement the over-identifying restrictions test. Specifically:

   (i) Calculate the J-statistic.

   (ii) Assuming a 5% significance level, what is the relevant critical value?

   (iii) What does the test conclude?

3 Final Exam Review: Spring 2014 Final, Question 4

In honor of Mother's Day last weekend, this question concerns a dataset consisting of more than a quarter of a million moms between the ages of 21 and 35 drawn from the 1980 Census. We focus on five variables in that dataset:

Variable	Description
`weeksworked`	number of weeks worked by the mom in 1979
`morekids`	= 1 if mom had more than 2 children
`samesex`	= 1 if 2 or more children and first two are same sex
`hispan`	= 1 if mom is Hispanic
`age`	age of mom in the 1980 census

A labor economist is interested in answering how the number of children affects mothers' labor supply decisions, and specifically whether there is a causal effect of the dummy indicator `morekids` on `weeksworked`. The researcher's population regression is: $weeksworked_i = \beta_0 + \beta_1 morekids_i + \beta_2 age_i + u_i$. Below you will find the log of a series of Stata program commands executed on this dataset by the researcher, and below that is a list of questions to be answered using this output.

```
. summ weeksworked morekids samesex hispan age

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
 weeksworked |    254654    19.01833    21.86728          0         52
    morekids |    254654    .3805634    .4855263          0          1
     samesex |    254654    .5055683      .49997          0          1
      hispan |    254654    .0742066    .2621073          0          1
         age |    254654    30.39327    3.386447         21         35

. correlate weeksworked morekids samesex hispan age
(obs=254654)
             | weeksw~d morekids  samesex   hispan      age
-------------+---------------------------------------------
 weeksworked |   1.0000
    morekids |  -0.1196   1.0000
     samesex |  -0.0097   0.0695   1.0000
      hispan |  -0.0104   0.0777  -0.0002   1.0000
         age |   0.1111   0.0999  -0.0031  -0.0657   1.0000

. regress weeksworked morekids age , robust

Linear regression                                      Number of obs =  254654
                                                       F(  2,254651) = 4252.27
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0296
                                                       Root MSE      =  21.541

------------------------------------------------------------------------------
             |               Robust
 weeksworked |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -5.946385   .0865748   -68.68   0.000     -6.11607   -5.776701
         age |   .8028323   .0121562    66.04   0.000     .7790064    .8266582
       _cons |  -3.119385   .3674612    -8.49   0.000    -3.839599   -2.399171
------------------------------------------------------------------------------

. regress morekids samesex hispan , robust

Linear regression                                      Number of obs =  254654
                                                       F(  2,254651) = 1368.81
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0109
                                                       Root MSE      =  .48288

------------------------------------------------------------------------------
             |               Robust
    morekids |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     samesex |   .0675405   .0019132    35.30   0.000     .0637907    .0712902
      hispan |   .1439261   .0037629    38.25   0.000      .136551    .1513013
       _cons |   .3357368   .0013596   246.93   0.000      .333072    .3384017
------------------------------------------------------------------------------


. test samesex hispan

 ( 1)  samesex = 0
 ( 2)  hispan = 0

       F(  2,254651) = 1368.81
```

```
           Prob > F =    0.0000

. ivregress 2sls weeksworked (morekids = samesex) age , robust

Instrumental variables (2SLS) regression              Number of obs =  254654
                                                      Wald chi2(2)  = 3520.60
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.0296
                                                      Root MSE      =  21.541

------------------------------------------------------------------------------
             |               Robust
 weeksworked |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |   -6.06062   1.258803    -4.81   0.000    -8.527828   -3.593413
         age |   .8044685   .0217279    37.02   0.000     .7618825    .8470544
       _cons |  -3.125639   .3739658    -8.36   0.000    -3.858599    -2.39268
------------------------------------------------------------------------------
Instrumented:  morekids
Instruments:   age samesex

. ivregress 2sls weeksworked (morekids = hispan) age , robust

Instrumental variables (2SLS) regression              Number of obs =  254654
                                                      Wald chi2(2)  = 3469.70
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.0205
                                                      Root MSE      =  21.642

------------------------------------------------------------------------------
             |               Robust
 weeksworked |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |   -1.63085   1.037662    -1.57   0.116     -3.66463    .4029299
         age |   .7410219   .0192269    38.54   0.000     .7033379    .7787059
       _cons |    -2.8831   .3736879    -7.72   0.000    -3.615515   -2.150685
------------------------------------------------------------------------------
Instrumented:  morekids
Instruments:   age hispan

. ivregress 2sls weeksworked (morekids = samesex hispan) age , robust

Instrumental variables (2SLS) regression              Number of obs =  254654
                                                      Wald chi2(2)  = 3506.36
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.0265
                                                      Root MSE      =  21.575

------------------------------------------------------------------------------
             |               Robust
 weeksworked |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    morekids |  -3.430894   .7992682    -4.29   0.000    -4.997431   -1.864357
         age |   .7668035   .0166953    45.93   0.000     .7340813    .7995257
       _cons |  -2.981656   .3707483    -8.04   0.000     -3.70831   -2.255003
------------------------------------------------------------------------------
Instrumented:  morekids
Instruments:   age samesex hispan

. predict u2slshat , resid

. regress u2slshat samesex hispan age

      Source |       SS       df       MS              Number of obs =  254654
-------------+------------------------------           F(  3,254650) =    2.44
       Model |  3410.54014     3  1136.84671           Prob > F      =  0.0622
    Residual |   118536486254650  465.487869           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared =  0.0000
       Total |   118539896254653  465.495778           Root MSE      =  21.575

------------------------------------------------------------------------------
    u2slshat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     samesex |  -.1783145   .0855142    -2.09   0.037      -.34592   -.0107091
      hispan |   .2819951   .1634709     1.73   0.085    -.0384035    .6023937
         age |   .0013517   .0126525     0.11   0.915    -.0234469    .0261504
       _cons |   .0281409   .3904391     0.07   0.943    -.7371093    .7933912
------------------------------------------------------------------------------

. test samesex hispan

 ( 1)  samesex = 0
 ( 2)  hispan = 0

       F(  2,254650) =    3.66
            Prob > F =    0.0256
```

a) **[5]** Why did the researcher include the `age` of the mother in the OLS regression? What would you expect would happen to the coefficient on `morekids` if `age` was excluded?

b) **[4]** Give one reason why you would suspect that the coefficient on `morekids` would not be an unbiased and consistent estimate of the population coefficient.

c) **[4]** What is the interpretation of the coefficient on `morekids` in the OLS regression. Is it economically significant? Is it statistically significant?

From the above output, the researcher thinks that two of the other variables in the dataset, `samesex` and `hispan` , are potential instruments for `morekids`. The variable `samesex` takes the value of 1 when the mother has had 2 or more children *and* the first two were both boys *or* both girls, and 0 otherwise.

d) **[5]** Why might the variable `samesex` be a "relevant" instrument for the endogenous regressor `morekids`? Explain why there is empirical evidence in the Stata output supporting the relevance of both `samesex` and `hispan`.

e) **[4]** Assuming that both candidates are valid instruments, why does the researcher have a case of "over identification"? What evidence do you see in the Stata output that confirms that there is an issue with overidentification?

f) **[4]** For these variables, what does it mean for the candidate instruments to meet the second criterion of a valid instrument, i.e., to be "exogenous"? What is in the Stata output that should make you suspicious that the one or the other or both of these candidate instruments are not exogenous?

g) **[7]** How did the researcher attempt to determine whether the candidate instruments are exogenous? Describe the steps she took. What should the researcher conclude about exogeneity of these instruments given the evidence?

TABLE 3 Critical Values for the χ^2 Distribution

	Significance Level		
Degrees of Freedom	**10%**	**5%**	**1%**
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.72
12	18.55	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.59	33.41
18	25.99	28.87	34.81
19	27.20	30.14	36.19
20	28.41	31.41	37.57
21	29.62	32.67	38.93
22	30.81	33.92	40.29
23	32.01	35.17	41.64
24	33.20	36.41	42.98
25	34.38	37.65	44.31
26	35.56	38.89	45.64
27	36.74	40.11	46.96
28	37.92	41.34	48.28
29	39.09	42.56	49.59
30	40.26	43.77	50.89

This table contains the 90th, 95th, and 99th percentiles of the χ^2 distribution. These serve as critical values for tests with significance levels of 10%, 5%, and 1%.