Practice 2SLS with Artificial Data Part 2

Yona Rubinstein

July 2016

Practice with Artificial Data

- In this note we use artificial data to illustrate how to approach a selection problem using 2SLS.
- The "constructed" data set "MG4A4 2SLS DIET EXAMPLE PART 2.dta" can be found on my website under "Practice 2SLS using Artificial Data".
- The data contains 1000 individuals each individual observed over 100 periods (days, weeks).
- We have information regarding their weight and on whether they are dating.
- Specifically the data contains their (1) permanent weight, (2) change in weight if diet is not taken, (3) whether they are on diet and (iv) whether they received an invitation to date.
- People are on diet for three reasons: (1) if their weight exceeds 220 pounds; (2) if they gained 5 pounds; (3) if they receive an invitation to date.
- The latter is exogenous to fluctuations in their current weight.

The Model: The Production Function of Peorsons' Weight

- The casual model exhibits the following form:

$$Y_{it} = \beta_0 + \beta_D D_{it} + U_{it}. \tag{1}$$

- The variable Y_{it} is person's i weight (in pounds) in time t and D_{it} is a binary indicator which equals 1 if person i is on diet.
- The error term (U_{it}) is a composition of person's i permanent weight (relative to the population mean, that is θ_i) and person's i specific time varying fluctuations to his weight:

$$U_{it} = \theta_i + \varepsilon_{it}. \tag{2}$$

- The parameters that I used to impute weights (Y_{it}) are:

$$Y_{it} = 0 - 10 \cdot D_{it} + U_{it}. \tag{3}$$

- People are on diet for three reasons:

1. If their weight, without diet, exceeds 220 pounds (100kg): $\beta_0 + \theta_i + \varepsilon_{it} \geq 220$.
2. If they gained 5 pounds (or more): $\varepsilon_{it} >= 5$
3. If they receive an invitation to date: $Date_{it} = 1$

- The value of a date - from the perspective of the other person, the person that invites (or accepts a standing invitation) - is a function of person's i permanent weight and a random component η_{it}:

$$vdate_{it} = \alpha_0 - \alpha_U \theta_i + \eta_{it}, \tag{4}$$

where in the data $\alpha_U = 1/190$ where 190 is the median weight in the population sample.

- Person i receives an invitation to date if $vdate_{it} > 0$ that is:

$$date_{it} = 1\left(\alpha_0 + \eta_{it} > \alpha_U \theta_i\right). \tag{5}$$

- Hence, the *value* of a date is also determined by persons' weight which means that dates are endogenous to peoples' weight.
- We observe neither $vdate_{it}$ nor ε_{it}.
- Yet we observe whether person i dates and whether he is on diet and his weight.
- We further know that person i is on diet if:

$$D_{it} = \max\left[\left(\beta_0 + \theta_i + \varepsilon_{it} - 220\right), date_{it}, \varepsilon_{it}\right] > 0. \tag{6}$$

- Note that weight affects in **offsetting directions** on the likelihood to be on diet! I will come back to that.

Data

- Describe the data set

```
                 storage   display    value
variable name     type     format     label      variable label
------------------------------------------------------------------
id                byte     %8.0g                 person id number
time              byte     %8.0g
PWi               float    %9.0g                 Person's permanent weig
Eit               float    %9.0g                 episilon it
date              byte     %8.0g                 shock to date value
vdate             float    %9.0g                 date - PWi/190
Date              float    %9.0g                 1 if date==1
Diet              float    %9.0g                 1 if on diet
PWit              float    %9.0g                 PWi + Eit - 10*Diet
```

Data (cont.)

- Summary statistics

```
> sum ;
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          id |     10,000        50.5    28.86751          1        100
        time |     10,000        50.5    28.86751          1        100
         PWi |     10,000       189.5    28.86751        140        239
         Eit |     10,000   -1.49e-07    13.81276     -60.54      63.48
        date |     10,000       .5004    .5000248          0          1
-------------+---------------------------------------------------------
       vdate |     10,000   -.4969684    .5217212  -1.257895   .2631579
        Date |     10,000       .2474    .4315227          0          1
        Diet |     10,000       .5904    .4917845          0          1
        PWit |     10,000     183.596    31.43375      84.08     268.06
```

Estimating the Regression Model

- We next estimate the model in equation (1) using OLS

$$Y_{it} = b_0 + b_D D_{it} + e_{it}. \tag{7}$$

```
> eststo: reg PWit Diet ;
      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(1, 9998)      =     14.73
       Model |   14537.399         1   14537.399   Prob > F        =    0.0001
    Residual |  9865282.69     9,998  986.725614   R-squared       =    0.0015
-------------+----------------------------------   Adj R-squared   =    0.0014
       Total |  9879820.09     9,999  988.080817   Root MSE        =    31.412

------------------------------------------------------------------------------
        PWit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Diet |   2.451829   .6387708     3.84   0.000      1.19971    3.703949
       _cons |   182.1484   .4908155   371.11   0.000     181.1863    183.1105
------------------------------------------------------------------------------
```

- According to the OLS estimate b_D^{OLS} diet leads to a gain of 2.45 pounds in weight!!

Estimating the Regression Model Controlling for Fixed Effects

- Next next turn to estimate the model using controlling for person fixed effects (θ_i):

$$Y_{it} = b_0 + b_D D_{it} + \theta_i + n_{it}. \tag{8}$$

```
> eststo: areg PWit Diet, absorb(id) ;
Linear regression, absorbing indicators         Number of obs     =     10,000
                                                F(   1,   9899)   =     938.81
                                                Prob > F          =     0.0000
                                                R-squared         =     0.8729
                                                Adj R-squared     =     0.8716
                                                Root MSE          =    11.2651

------------------------------------------------------------------------------
        PWit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Diet |   7.471088   .2438343    30.64   0.000     6.993124    7.949053
       _cons |   179.1851   .1827971   980.24   0.000     178.8268    179.5434
-------------+----------------------------------------------------------------
          id |      F(99, 9899) =    685.247   0.000           (100 categories)
```

- Accounting for person fixed effects increases the bias. Why?

Estimating the Regression Model using 2SLS

- Next we turn to estimate the model using 2SLS using the following equations:

1. The first stage model:

$$D_{it} = a_0 + a_D Date_{it} + \gamma_i + v_{it}, \tag{9}$$

   where V_{it} is the error term.

2. The second tage model:

$$Y_{it} = b_0 + b_D \hat{D}_{it} + \delta_i + e_{it}, \tag{10}$$

   where $\hat{D}_{it} = a_0^{OLS} + a_D^{OLS} Date_{it} + \gamma_i^{OLS}$.

- We present the results in next silde.

First Stage

$$D_{it} = a_0 + a_D Date_{it} + \gamma_i + v_{it}, \tag{11}$$

```
> areg Diet Date, absorb(id) ;
Linear regression, absorbing indicators           Number of obs     =     10,000
                                                  F(   1,    9899)  =    3215.55
                                                  Prob > F          =     0.0000
                                                  R-squared         =     0.3338
                                                  Adj R-squared     =     0.3271
                                                  Root MSE          =     0.4034
------------------------------------------------------------------------------
        Diet |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Date |    .650042   .0114634    56.71   0.000     .6275714    .6725126
       _cons |   .4295796   .0049314    87.11   0.000     .4199131    .4392461
-------------+----------------------------------------------------------------
          id |      F(99, 9899) =     15.868   0.000          (100 categories)
```

Second Stage (without correcting SE).

$$Y_{it} = b_0 + b_D \hat{D}_{it} + \delta_i + e_{it}, \tag{12}$$

```
> eststo:  areg PWit Diethat, absorb(id)  ;
Linear regression, absorbing indicators          Number of obs     =     10,000
                                                 F(   1,    9899)   =     321.98
                                                 Prob > F          =     0.0000
                                                 R-squared         =     0.8652
                                                 Adj R-squared     =     0.8638
                                                 Root MSE          =    11.6001
------------------------------------------------------------------------------
        PWit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Diethat |  -9.098722   .5070707   -17.94   0.000    -10.09268   -8.104761
       _cons |   188.9679   .3210628   588.57   0.000     188.3385    189.5972
-------------+----------------------------------------------------------------
          id |       F(99, 9899) =    435.351   0.000         (100 categories)
```

Estimating the 2SLS using "xtivreg"

- Note that we obtain identical point estimates. The standard errors were corrected to account for using a projected variable $\hat{D}_{it}$.
- Using dates as an instrument allows to correct of selection on "LHS" variable.

```
> eststo:  xtivreg PWit (Diet=Date), fe i(id) ;
Fixed-effects (within) IV regression            Number of obs     =     10,000
Group variable: id                              Number of groups  =        100
R-sq:                                           Obs per group:
     within  =      .                                        min =        100
     between = 0.0416                                        avg =      100.0
     overall = 0.0015                                        max =        100
                                                Wald chi2(1)      =   1.81e+06
corr(u_i, Xb)  = 0.0523                         Prob > chi2       =     0.0000
------------------------------------------------------------------------------
        PWit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Diet |  -9.098722   .5963281   -15.26   0.000    -10.2675    -7.92994
       _cons |   188.9679   .3775781   500.47   0.000    188.2278    189.7079
-------------+----------------------------------------------------------------
     sigma_u |  29.034412
     sigma_e |  13.642005
         rho |  .81915843   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F  test that all u_i=0:     F(99,9899) =   449.25          Prob > F    = 0.0000
```

Estimating the Model using OLS, FE and 2SLS

	OLS	FE	OLS2ND	2SLS
	(1)	(2)	(3)	(4)
Dependent variable: "Weight"				
Diet	2.452***	7.471***	-9.099***	-9.099***
	(0.639)	(0.244)	(0.507)	(0.596)
Constant	182***	179***	189***	189***
	(0.491)	(0.183)	(0.321)	(0.378)
R-square	0.001	0.873	0.865	0.865
First Stage: Dependent variable "Diet"				
Date			0.650***	
			(0.011)	
			0.430***	
			(0.005)	
R-square			0.334	
Observation:	10000	10000	10000	10000

Take Home Message

- We can *identify* the causal impact of a *treatment* on an *outcome of interest* – **accounting for selection into treatment** – if we have a variable that

1. Is *uncorrelated* with the **unobserved component** in the *outcome equation* (the error term);
2. Does **not** affect *directly* the outcome of interest

- Controlling for subjects **fixed effects** might eliminate some of the bias but not all as long as subject self-sort into treatment on **time varying unobservables**. Yet, it might **magnify** the *bias* when these factors affect other related choices.
- Therefore, we should be careful with the source of variation in treatment status we utilize to *identify* **causal impacts**.