Practice 2SLS with Artificial Data Part 1

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 1.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:

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

Selection into Diet - the Simplest Case

- 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: $Date_{it} = 1$
- We observe neighter $vdate_{it}$ nor ε_{it}. Yet we observe whether the person is on diet and his weight. We know that person i is on diet if:

$$D_{it} = \max\left[\left(\beta_0 + \theta_i + \varepsilon_{it} - 220\right), Date_{it}, \varepsilon_{it}\right] > 0. \tag{4}$$

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        Ma
-------------+---------------------------------------------------------
          id |     10,000        50.5    28.86751         1        10
        time |     10,000        50.5    28.86751         1        10
         PWi |     10,000       189.5    28.86751       140        23
         Eit |     10,000   -1.49e-07    13.81276    -60.54      63.4
        date |     10,000       .5004    .5000248         0
-------------+---------------------------------------------------------
       vdate |     10,000   -.4969684    .5217212  -1.257895    .263157
        Date |     10,000       .5004    .5000248         0
        Diet |     10,000       .7144    .4517223         0
        PWit |     10,000     182.356    30.97509     84.08     268.0
```

Estimating the Regression Model

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

$$Y_{it} = b_0 + b_D D_{it} + e_{it}. \tag{5}$$

```
> eststo: reg PWit Diet ;
      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(1, 9998)      =    256.56
       Model |  240026.875         1  240026.875   Prob > F        =    0.0000
    Residual |  9353574.21     9,998   935.54453   R-squared       =    0.0250
-------------+----------------------------------   Adj R-squared   =    0.0249
       Total |  9593601.08     9,999  959.456054   Root MSE        =    30.587

------------------------------------------------------------------------------
        PWit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Diet |    10.84626   .6771461    16.02   0.000      9.51892    12.17361
       _cons |    174.6074   .5723387    305.08   0.000     173.4855    175.7293
```

- According to the OLS estimate b_D^{OLS} diet leads to a gain of 10 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):

$$Y_{it} = b_0 + b_D D_{it} + \theta_i + n_{it}. \tag{6}$$

```
eststo: areg PWit Diet, absorb(id) ;
Linear regression, absorbing indicators          Number of obs     =     10,000
                                                 F(   1,   9899)   =     106.28
                                                 Prob > F          =     0.0000
                                                 R-squared         =     0.8353
                                                 Adj R-squared     =     0.8336
                                                 Root MSE          =    12.6344

------------------------------------------------------------------------------
        PWit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Diet |   2.946265    .285791    10.31   0.000     2.386056    3.506473
       _cons |   180.2512   .2400997   750.73   0.000     179.7805    180.7218
-------------+----------------------------------------------------------------
          id |       F(99, 9899) =    491.888   0.000         (100 categories)
```

- Accounting for person fixed effects matter! the bias is smaller, yet still large enough (3 rather than -10).

Estimating the Regression Model using 2SLS

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

1. The first stage model:

$$D_{it} = a_0 + a_D Date_{it} + v_{it}, \tag{7}$$

where V_{it} is the error term.

2. The second tage model:

$$Y_{it} = b_0 + b_D \hat{D}_{it} + e_{it}, \tag{8}$$

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

- We present the results in next silde.

First Stage

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

```
> reg Diet Date ;

      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(1, 9998)      =   6676.90
       Model |  816.979723         1  816.979723   Prob > F        =    0.0000
    Residual |  1223.34668     9,998   .12235914   R-squared       =    0.4004
-------------+----------------------------------   Adj R-squared   =    0.4004
       Total |   2040.3264     9,999  .204053045   Root MSE        =     .3498

------------------------------------------------------------------------------
        Diet |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Date |   .5716573    .006996    81.71   0.000     .5579438    .5853708
       _cons |   .4283427   .0049489    86.55   0.000     .4186419    .4380435
------------------------------------------------------------------------------
```

Second Stage (without correcting SE).

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

```
> eststo:  reg PWit Diethat  ;
      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(1, 9998)      =     68.61
       Model |  65389.6299         1  65389.6299   Prob > F        =    0.0000
    Residual |  9528211.45     9,998  953.011748   R-squared       =    0.0068
-------------+----------------------------------   Adj R-squared   =    0.0067
       Total |  9593601.08     9,999  959.456054   Root MSE        =    30.871

------------------------------------------------------------------------------
        PWit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Diethat |   -8.94641   1.080049    -8.28   0.000    -11.06352   -6.829296
       _cons |   188.7473   .8310522   227.12   0.000     187.1183    190.3763
------------------------------------------------------------------------------
```

Estimating the 2SLS using "ivreg"

- 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:  ivreg PWit (Diet=Date) ;
Instrumental variables (2SLS) regression
      Source |       SS           df       MS      Number of obs   =    10,000
-------------+----------------------------------   F(1, 9998)      =     64.39
       Model | -559270.763         1 -559270.763   Prob > F        =    0.0000
    Residual |  10152871.8     9,998  1015.49028   R-squared       =         .
-------------+----------------------------------   Adj R-squared   =         .
       Total |  9593601.08     9,999  959.456054   Root MSE        =    31.867

------------------------------------------------------------------------------
        PWit |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Diet |   -8.94641   1.114891    -8.02   0.000    -11.13182      -6.761
       _cons |   188.7473   .8578613   220.02   0.000     187.0657    190.4289
```

Controlling for Fixed Effects

- We can take advantage of the panel data and further control for omitted person fixed effects.

1. The first stage model controls for person fixed effects (γ_i):

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

where V_{it} is the error term.

2. The second stage model also controls for person fixed effects (δ_i):

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

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

- We present the results in next silde.

Estimating the 2SLS controlling for person fixed effects using "xtivreg"

- Person fixed effects are part of the causal model. While our instrument provides a source of exogenous variation - controlling for persons' fixed effects does not heart.

```
> eststo:  xtivreg PWit (Diet=Date) ;
G2SLS random-effects IV regression              Number of obs     =     10,000
Group variable: id                              Number of groups  =        100
R-sq:                                           Obs per group:
     within  = 0.0106                                         min =        100
     between = 0.3901                                         avg =      100.0
     overall = 0.0250                                         max =        100
                                                Wald chi2(1)      =     385.69
corr(u_i, X)       = 0 (assumed)                Prob > chi2       =     0.0000
------------------------------------------------------------------------------
       PWit |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
       Diet |  -9.548751    .486211   -19.64   0.000    -10.50171   -8.595795
      _cons |   189.1776   2.624326    72.09   0.000      184.034    194.3212
```

Estimating the Model using OLS, FE and 2SLS

	OLS	FE	OLS2ND	2SLS	2SLS w/FE
	(1)	(2)	(3)	(4)	(5)
Dependent variable: "Weight"					
Diet	**10.846*** **	**2.946*** **	**-8.946*** **	**-8.946*** **	**-9.549*** **
	(0.677)	(0.286)	(1.080)	(1.115)	(0.486)
Constant	175***	180***	189***	189***	189***
	(0.572)	(0.240)	(0.831)	(0.858)	(2.624)
First Stage: Dependent variable "Diet"					
Date			0.572***	0.572***	0.571***
			(0.007)	(0.007)	(0.007)
			0.428***	0.428***	0.429***
			(0.005)	(0.005)	(0.005)
R-square			0.400	0.400	0.390
Observations	10000	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**.