Suggested Answers
Problem Set 8

Bill Evans
Spring 2018

1. a. N=25, k=5, $\hat{d}=1.80$, lower=0.953, upper=1.886, since lower< $\hat{d}$ <upper the test is inconclusive
   b. N=60, k=9, $\hat{d}=0.23$, lower=1.260, upper=1.939, since $\hat{d}$ <lower, we can reject the null that ρ=0
   c. N=45, k=2, $\hat{d}=1.40$, lower=1.430, upper=1.615, since $\hat{d}$ <lower, we can reject the null that ρ=0

2. a. False -- $\hat{\beta}_1$ is still unbiased even in the presence of autocorrelation
   b. True -- $Var(\hat{\beta}_1)$ is too small in this situation
   c. False – Although the OLS and the AR(1) corrected estimates are both unbiased, because we are working with finite samples, there is little chance these two estimates will produce identical results
   d. False – Measurement error in y such as this only increases variance, it does not produce biased estimates – which means part e is TRUE
   e. True --
   f. True – with classical measurement error, the OLS estimated tends to be attenuated towards zero

3. The answers for this question are contained in the program titled michigan_dnd.do.

   The results for models (1) – (3) are below

   The means for the 2 x 2 table are as follows:

```
.
. * get means of smoked for 2 x 2 table
. by michigan after:  sum smoked

-------------------------------------------------------------------------------
-> michigan = 0, after = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      smoked |     15152    .1855861    .3887851          0          1

-------------------------------------------------------------------------------
-> michigan = 0, after = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      smoked |      7304    .1856517     .388852          0          1

-------------------------------------------------------------------------------
-> michigan = 1, after = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      smoked |     53232    .1922904    .3941037          0          1

-------------------------------------------------------------------------------
-> michigan = 1, after = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
```

```
-------------+--------------------------------------------------------
      smoked |     25988    .1786979    .3831065          0          1
```

Putting these means into the 2x2 table, we obtain a difference in difference estimate of -0.0137 or, the tax hike reduce smoking rates among pregnant women in Michigan by 1.36 percentage points

	Before (1)	After (2)	Difference (2) – (1)
Michigan (1)	0.1923	0.1787	-0.0136
Iowa (2)	0.1856	0.1857	0.0001
Difference (1) – (2)			-0.0137

Now, estimating the model within a regression, we obtain the following:

```
      Source |       SS       df       MS              Number of obs =  101676
-------------+------------------------------           F(  3,101672) =    7.25
       Model |   3.3128883     3   1.1042961           Prob > F      =  0.0001
    Residual |  15476.2314101672  .152217242           R-squared     =  0.0002
-------------+------------------------------           Adj R-squared =  0.0002
       Total |  15479.5443101675  .152245333           Root MSE      =  .39015

------------------------------------------------------------------------------
      smoked |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    michigan |   .0067043   .0035924     1.87   0.062    -.0003368    .0137454
       after |   .0000656   .0055575     0.01   0.991    -.0108271    .0109583
   treatment |  -.0136581   .0062931    -2.17   0.030    -.0259925   -.0013237
       _cons |   .1855861   .0031695    58.55   0.000     .1793738    .1917983
------------------------------------------------------------------------------
```

d) The primary assumption of the difference in difference model is that the comparison state provides an estimate of the time path of outcomes that would have occurred in the absence of the intervention. The data has two years pre tax hike (years 1 and 2) and one year post (year 3). We can test this assumption by running a fake difference in difference model. You are to estimate a difference in difference model assuming the tax hike occurred in year 2 and delete the data for year 3. In this case, the coefficient on the "treatment" effect should be zero, which it is.

```
. reg smoked michigan after2 treatment2 if year<=2

      Source |       SS       df       MS              Number of obs =   68384
-------------+------------------------------           F(  3, 68380) =    8.43
       Model |  3.90186034     3  1.30062011           Prob > F      =  0.0000
    Residual |  10554.4762 68380   .15435034           R-squared     =  0.0004
-------------+------------------------------           Adj R-squared =  0.0003
       Total |  10558.3781 68383  .154400627           Root MSE      =  .39287

------------------------------------------------------------------------------
      smoked |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    michigan |   .0093967   .0050651     1.86   0.064    -.0005309    .0193244
      after2 |   -.010017   .0063859    -1.57   0.117    -.0225334    .0024995
  treatment2 |  -.0049776   .0072374    -0.69   0.492    -.0191628    .0092076
       _cons |   .1904517   .0044507    42.79   0.000     .1817284    .1991751
------------------------------------------------------------------------------
```

This is easier to see in a 2 x 2 box. Using only data from year 2 and 1, we see that the both states experienced a drop in smoking of roughly the same size between year 2 and 1. From the regression estimate, we cannot reject the null that the drop is the same across the two states.

	Before (year 1)	After (year 2)	Difference (year 2) – (year 1)
Michigan (1)	0.1998	0.1848	-0.0150
Iowa (2)	0.1905	0.1805	-0.0100
Difference (1) – (2)			-0.0050

4. A program to answer this question is in the program titled smoked_dnd. The results for models (1) through (3) are below.

Model 1

```
      Source |       SS       df       MS              Number of obs =    1020
-------------+------------------------------           F(  2,  1017) =  369.06
       Model |  32.8291921     2  16.414596           Prob > F      =  0.0000
    Residual |  45.2330044  1017  .044476897           R-squared     =  0.4206
-------------+------------------------------           Adj R-squared =  0.4194
       Total |  78.0621965  1019   .07660667           Root MSE      =   .2109

------------------------------------------------------------------------------
  packs_pc_l |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       rpcil |  -.0647348   .0436796    -1.48   0.139    -.1504473    .0209777
    real_tax |  -.0093986   .0004151   -22.64   0.000    -.0102132   -.0085841
       _cons |    5.81216   .4291643    13.54   0.000     4.970011    6.654309
------------------------------------------------------------------------------
```

Model 2

```
. * add state effects
. reg packs_pc_l rpcil real_tax _Is*

      Source |       SS       df       MS              Number of obs =    1020
-------------+------------------------------           F( 52,   967) =  176.38
       Model |  70.6170713    52   1.3580206           Prob > F      =  0.0000
    Residual |  7.44512522   967  .007699199           R-squared     =  0.9046
-------------+------------------------------           Adj R-squared =  0.8995
       Total |  78.0621965  1019   .07660667           Root MSE      =  .08775

------------------------------------------------------------------------------
  packs_pc_l |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       rpcil |  -.7787655   .0332236   -23.44   0.000    -.8439642   -.7135668
    real_tax |  -.0074131   .0002477   -29.92   0.000    -.0078993   -.0069269
   _Istate_2 |  -.3447585   .0301021   -11.45   0.000    -.4038315   -.2856855
   _Istate_3 |  -.2408913   .0313047    -7.70   0.000    -.3023244   -.1794583
   _Istate_4 |  -.4076549   .0293303   -13.90   0.000    -.4652132   -.3500966

             delete some results
  _Istate_49 |  -.1535768   .0290644    -5.28   0.000    -.2106134   -.0965401
  _Istate_50 |  -.3799918   .0310106   -12.25   0.000    -.4408476    -.319136
  _Istate_51 |  -.2216287   .0286039    -7.75   0.000    -.2777615   -.1654958
       _cons |     13.107   .3355447    39.06   0.000     12.44852    13.76548
------------------------------------------------------------------------------
```

Model 3

```
. * add year effects
. reg packs_pc_l rpcil real_tax _Is* _Iy*

      Source |       SS       df       MS              Number of obs =    1020
-------------+------------------------------           F( 71,   948) =  226.24
       Model |  73.7119499    71  1.03819648           Prob > F      =  0.0000
    Residual |  4.35024662   948  .004588868           R-squared     =  0.9443
-------------+------------------------------           Adj R-squared =  0.9401
       Total |  78.0621965  1019   .07660667           Root MSE      =  .06774

------------------------------------------------------------------------------
  packs_pc_l |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       rpcil |   .2818674   .0585799     4.81   0.000     .1669061    .3968287
    real_tax |  -.0062409   .0002227   -28.03   0.000    -.0066779   -.0058039
   _Istate_2 |   .0926469   .0321122     2.89   0.004     .0296277     .155666
            delete some results
  _Istate_50 |   .1260896   .0350074     3.60   0.000     .0573885    .1947906
  _Istate_51 |   .0543776   .0264025     2.06   0.040     .0025635    .1061916
 _Iyear_1982 |  -.0180335    .013415    -1.34   0.179      -.04436     .008293
            delte some results
 _Iyear_1999 |  -.3664177   .0232861   -15.74   0.000     -.412116   -.3207194
 _Iyear_2000 |   -.373204   .0255011   -14.63   0.000    -.4232492   -.3231589
       _cons |   2.294338   .5966798     3.85   0.000     1.123372    3.465304
------------------------------------------------------------------------------
```

Notice that as we add more control variables, the coefficient on real_tax increases along the number line (falls in absolute value) from -0.0094, to -0.0074, -0.0062. This suggests that the model we initially estimated at the start of class (model 1) is biased for two reasons. First, the fact that the coefficient increased along the number lines when we added state effects suggests that high cigarette consuming states tend to also be low taxing states. This is not surprise – states that produce tobacco tend to not tax the product much and residents in these states smoke a lot. Not controlling for the fact that states with greater than average propensity to smoke are less likely to tax will seriously bias down the parameter estimates. Note also that because the coefficient increases along the number line when we ad in time effects, it must be the case that there is persistent negative correlation between consumption and taxes over time. Over the past 40 years, smoking in the US has fallen considerably as people have learned about the dangers of smoking. At the same time, states have found it easier to raise taxes cigarettes (it is easy to raise taxes on a product when a minority of the population consume it.). Therefore, no controlling for the fact that over time that consumption levels have declined and taxes have increases will also bias down the estimates.

Notice as well that massive increase in the R^2 as we add state and year effects. Most of the variation in smoking rates is between states (Utah is always lower than Nevada) than within a state over time.

In model (3), a 10 cent increase in real taxes will reduce per capita cigarette consumption by 6.2%.

5. a) Start with $\text{var}(\hat{\beta}_1) = \dfrac{\sigma_\varepsilon^2 \sum_{i=1}^{n}(z_i - \bar{z})^2}{\left[\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})\right]^2}$ and place $\sum_{i=1}^{n}(z_i - \bar{z})^2$ in the denominator

$$\text{var}(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{\dfrac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(z_i-\bar{z})\right]^2}{\sum_{i=1}^{n}(z_i-\bar{z})^2}}.$$

Divide the numerator and denominator by $\sum_{i=1}^{n}(x_i-\bar{x})^2$ which produces

$$\text{var}(\hat{\beta}_1^{2SLS}) = \frac{\sigma_\varepsilon^2}{\dfrac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(z_i-\bar{z})\right]^2}{\sum_{i=1}^{n}(z_i-\bar{z})^2}}\left[\frac{\dfrac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}{\dfrac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\right] = \frac{\dfrac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}{\dfrac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(z_i-\bar{z})\right]^2}{\sum_{i=1}^{n}(z_i-\bar{z})^2\sum_{i=1}^{n}(x_i-\bar{x})^2}}.$$

Notice that the numerator is nothing more than $Var(\hat{\beta}_1^{OLS})$ where $Var(\hat{\beta}_1^{OLS}) = \dfrac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$. In the denominator, divide the numerator and denominator by (n-1)^2

$$\frac{\left[\sum_{i=1}^{n}(x_i-\bar{x})(z_i-\bar{z})\right]^2/(n-1)^2}{\left(\dfrac{\sum_{i=1}^{n}(z_i-\bar{z})^2}{n-1}\right)\left(\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}\right)} = \frac{\hat{\sigma}_{xy}^2}{\hat{\sigma}_z^2\hat{\sigma}_x^2} = \left(\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_z\hat{\sigma}_x}\right)^2 = \hat{\rho}_{xy}^2 \text{ so therefore}$$

$$\text{var}(\hat{\beta}_1^{2SLS}) = \frac{Var(\hat{\beta}_1^{OLS)}}{\hat{\rho}_{xz}^2}$$

b) If Z does a poor job of explaining Z then $\hat{\rho}_{xz} \to 0$ and $\text{var}(\hat{\beta}_1^{2SLS})$ blows up because the denominator approaches zero.

6.a) There is room for concern that this condition is not met in this case. It is clear that BMI of siblings will be correlated. However, there is some concern that having heavier siblings may signal something about an individual's earnings capacity. Suppose that some obese people have a number of habits that lead to their obesity: lack of discipline, impulsivity, inability to sacrifice today (e.g. diet, exercise) for goals in the future. It is likely that these same traits are negatively rewarded in the job market. This is why the OLS estimates are subject to an omitted variable bias. However, suppose these traits are transmitted to children through the parents – either through genetics or nurture. If this is the case, then a sibling's obesity would contain some of the information about people from this family having these negative traits as well.

a) We know that $p\lim(\hat{\beta}_1^{2SLS}) = \beta_1 + \frac{\sigma_{z\varepsilon}}{\sigma_{zx}}$ and from the text above, this suggests that $\sigma_{z\varepsilon} < 0$, and hence, $p\lim(\hat{\beta}_1^{2SLS})$ is biased down.

7. The program that generates these results is called twin1st.do and the log is twin1st.log

a) 60.4% of women worked last year, average weeks worked is 23 weeks and median labor income was $1005.

b) The answers for part b are below. The coefficient on second is -6.8 meaning that among women with one or more kids, the presence of the second child reduces weeks worked by an average of 6.8 weeks/year.

```
 . ************* part b
. * run OLS of weeks on second
. reg weeks second

      Source |       SS       df       MS              Number of obs =   12500
-------------+------------------------------           F(  1, 12498) =  140.68
       Model |  71801.5838     1  71801.5838           Prob > F      =  0.0000
    Residual |   6378669.1 12498  510.375188           R-squared     =  0.0111
-------------+------------------------------           Adj R-squared =  0.0111
       Total |  6450470.68 12499  516.078941           Root MSE      =  22.591

------------------------------------------------------------------------------
       weeks |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      second |  -6.813862   .5744749   -11.86   0.000    -7.939921   -5.687803
       _cons |   28.98838    .531307    54.56   0.000     27.94694    30.02983
------------------------------------------------------------------------------
```

c) In part c, the presence of a twin increases the probability of having a second child by 27.5 percentage points. Why is this coefficient not 1? At the time of the birth, the presence of the twin increases family size from 1 to 2. However, many of the women who had a twin on the 1st birth would have had a second one anyway so that is the reason the twin1st coefficient is less than 1.

Notice that in the reduced form regression (weeks worked on twin1st) produces a coefficient of -0.99. Women assigned a twin on the first birth are working 1 week fewer per ye. Notice that -0.99/0.2746 = -3.605 which is exactly the 2SLS estimate below.

According to the OLS model, the presence of the 2nd kid reduces work by almost 7 weeks per year. In the 2SLS model, however, this number reduces to -3.6. The OLS estimate is too large by a factor of 2 suggesting large omitted variables problems in the OLS model.

```
.
. ************ part c
. * run the first stage, does having
. * a twin (z) increase the kids in the home (x)?
. reg second twin1st

      Source |       SS       df       MS              Number of obs =   12500
-------------+------------------------------           F(  1, 12498) = 2239.20
       Model |  234.976907     1  234.976907           Prob > F      =  0.0000
    Residual |  1311.51397 12498  .104937908           R-squared     =  0.1519
-------------+------------------------------           Adj R-squared =  0.1519
       Total |  1546.49088 12499  .123729169           Root MSE      =  .32394
```

```
------------------------------------------------------------------------------
      second |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     twin1st |   .2746051   .0058031    47.32   0.000     .2632301    .2859801
       _cons |   .7253949   .0039923   181.70   0.000     .7175694    .7332204
------------------------------------------------------------------------------

.
. * run the reduced form, impact of twins (z)
. * on weeks worked (y)
. reg weeks twin1st

      Source |       SS       df       MS              Number of obs =   12500
-------------+------------------------------           F(  1, 12498) =    5.92
       Model |  3054.30028     1  3054.30028           Prob > F      =  0.0150
    Residual |  6447416.38 12498  515.875851           R-squared     =  0.0005
-------------+------------------------------           Adj R-squared =  0.0004
       Total |  6450470.68 12499  516.078941           Root MSE      =  22.713

------------------------------------------------------------------------------
       weeks |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     twin1st |   -.990038   .4068821    -2.43   0.015     -1.78759   -.1924865
       _cons |   23.62865    .279916    84.41   0.000     23.07997    24.17732
------------------------------------------------------------------------------
.
* run the 2sls model (Wald estimate)
. * ivregress 2sls y (x=z)
. ivregress 2sls weeks (second=twin1st)

Instrumental variables (2SLS) regression               Number of obs =   12500
                                                       Wald chi2(1)  =    5.97
                                                       Prob > chi2   =  0.0145
                                                       R-squared     =  0.0087
                                                       Root MSE      =  22.618

------------------------------------------------------------------------------
       weeks |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      second |  -3.605315   1.475498    -2.44   0.015    -6.497239   -.7133917
       _cons |   26.24392   1.278193    20.53   0.000     23.73871    28.74913
------------------------------------------------------------------------------
Instrumented:  second
Instruments:   twin1st

..
```

d) In this model, we run an OLS model similar to that in part B) but we add additional covavariates. Notice that the estimated impact of havin a second kid increases in magnitude from -6.8 to -9.26, providing strong evidence that the observed characteristics of the mother are correlated with whether the mother had a second child.

```
. ************* part e
. * run OLS of weeks worked model with other covariates
. reg weeks second agem agefst black other_race educm married

      Source |       SS       df       MS              Number of obs =   12500
-------------+------------------------------           F(  7, 12492) =  150.56
       Model |  501874.986     7  71696.4266           Prob > F      =  0.0000
```

```
    Residual |  5948595.69 12492  476.192419           R-squared     =  0.0778
-------------+------------------------------           Adj R-squared =  0.0773
       Total |  6450470.68 12499  516.078941           Root MSE      =  21.822

------------------------------------------------------------------------------
       weeks |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      second |  -9.255974   .5768304   -16.05   0.000    -10.38665   -8.125297
        agem |   1.000666   .0462932    21.62   0.000     .9099239    1.091407
      agefst |  -1.110525    .065915   -16.85   0.000    -1.239728   -.9813213
       black |   2.722332   .6233304     4.37   0.000     1.500509    3.944156
  other_race |   2.647268   1.171034     2.26   0.024     .3518603    4.942676
       educm |   1.321557   .0847274    15.60   0.000     1.155478    1.487636
     married |  -5.520823   .5492189   -10.05   0.000    -6.597377   -4.444269
       _cons |   11.67178   1.634199     7.14   0.000       8.4685    14.87506
------------------------------------------------------------------------------

.
```

e) In this problem, we estimate the model from part e) but by 2SLS using twinst as the instrument for second. Notice that the modle with covariates produces an 2SLS estimate on second of -3.8, which is only slightly larger than the Wald estimate in part c). This means that having a twin on the 1st birth is only weakly correlated with the observed characteristics (agem, agefst, black, etc.).

```
. ************* part f
. * run the 2sls with additional covariates in the model
. ivregress 2sls weeks agem agefst black other_race educm married (second=twin1st)

Instrumental variables (2SLS) regression               Number of obs =   12500
                                                       Wald chi2(7)  =  799.03
                                                       Prob > chi2   =  0.0000
                                                       R-squared     =  0.0713
                                                       Root MSE      =  21.892

------------------------------------------------------------------------------
       weeks |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      second |  -3.840711   1.388089    -2.77   0.006    -6.561314   -1.120107
        agem |    .893219    .052759    16.93   0.000     .7898133    .9966247
      agefst |   -1.00932   .0702044   -14.38   0.000    -1.146918   -.8717218
       black |   2.761305   .6253911     4.42   0.000     1.535561    3.987049
  other_race |   2.651669   1.174782     2.26   0.024     .3491376      4.9542
       educm |   1.338171   .0850866    15.73   0.000     1.171404    1.504938
     married |  -6.005684   .5624385   -10.68   0.000    -7.108044   -4.903325
       _cons |   8.371989   1.810752     4.62   0.000     4.822981      11.921
------------------------------------------------------------------------------
Instrumented:  second
Instruments:   agem agefst black other_race educm married twin1st
```

8. a. One would anticipate that people in most need of medical care (i.e., highest risk or mortality) would also receive the greatest amount of care or cov($x_i.\varepsilon_i$) > 0. We know that $E[\hat{\beta}_1] = \beta_1 + \frac{\sigma_{x\varepsilon}}{\sigma_x^2}$. Since we anticipate that $\beta_1 < 0$ and cov($x_i.\varepsilon_i$) > 0, then $E[\hat{\beta}_1] > \beta_1$.

b) The two figures suggest that babies with slightly lower weight than 1500 grams have a lot more spent on them, which translates into better health since mortality for this group is lower.

c) The 2SLS estimate is the reduced for divided by the 1st stage or $\hat{\beta}_1 = \hat{\pi}_1 / \hat{\theta}_1 = -0.02280 / 7670 = -2.97E-6.$ this means that for every additional $10K in spending on a low weight infant, mortality falls by (10,000)(-2.97E-6) by 0.0297 or 2.97 percentage points.

d) RDD models assume that the health of infants just above and below 1500 grams are functionally the same in the absence of treatment. Therefore, we can use the stark increase in treatment intensity right below 1500 grams to estimate the impact of spending on outcomes.

e) This RDD model only estimates the impact of greater health care spending at 1500 grams – it does not say anything about increased spending on health care at other birth weights. RDD models have high internal validity – in most situations they have low external validity.