Suggested Answers
Problem Set 7

Bill Evans
Spring 2018

1. The program that generates these results is called measurement_error_example.do. Below are a few tables that summarize the results for this problem. Please note that the variables v2 and v3 that are used to construct the new variables are produced from draws to a random number generator. Each time you run a program, the computer will generate a different sequence of random numbers so your results will differ slightly. Given the sample sizes, the mean and standard deviation of v2 should close to 0 and 1 respectively, while the same values for v3 should be 0 and 2.

Means, Standard Deviations, and Variances of Key Variables

Variable	Mean	Std. deviation	Variance
years_educ	13.16	2.80	7.84
v2	-0.000	0.995	0.99
v3	-0.0006	2.006	4.02
educ2	13.16	2.96	8.76
educ3	13.16	3.46	11.97
ln(weekly_earn)	6.067	0.513	0.26
y2	6.066	1.12	1.25
y2	6.068	2.07	4.28

OLS Estimates

Problem	Dependent Variable	Independent Variable	Parameter on Independent	Std error on independent
1a	Ln(weekly_earn)	years_educ	0.0741	0.0012
1c	Ln(weekly_earn)	educ2	0.0655	0.0011
1d	Ln(weekly_earn)	educ3	0.0486	0.0010
1e	y2	years_educ	0.0743	0.0028
1e	y3	years_educ	0.081	0.0052

a) When education does not have measurement error, the coefficient on that variable is 0.0741 indicating that each additional year of education increases earnings by 7.4 percent.

b) The random variable z2 has a mean of roughly zero and a variance of approximately 1. Therefore, the new constructed variable educ2 has a mean of 13.16 which is exactly the mean of years_educ, but now the variance of years_educ has increased by approximately 1, from 7.84 to 8.82.

c) When ln(weekly earn) is regressed on educ2, notice that the coefficient on the education variable falls to 0.0655. Notice also that the ratio of this estimate to the one without measurement error is simply 0.0655/0.0741=0.884. Is this to be expected? Yes. Recall two facts. First, in large samples, when there is random measurement error in x, the coefficient on x falls by the size of the reliability ratio. Notice that the variance of educ2 is simply the variance of years_educ plus the variance of v2. Therefore, the reliability ratio is $\theta = \dfrac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2}$ =7.84/8.76 = 0.89. The reliability ratio suggests that the coefficient on educ2 should be about 12 percent lower and it is roughly 12 percent lower.

d) Educ3=years_educ+v3 and notice that the variance for educ3 is about 4 larger than the variance of years_educ. Therefore, the reliability ratio in this context is $\theta = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2}$ = 7.84/11.97=0.65. This suggests that using educ3 instead of years_educ should reduce the coefficient on education by 33 percent. Notice that the ratio of the new to the original estimate is 0.0486/0.0741=0.655 or about 35 percent lower.

e) On problems set 3, we demonstrated that with random measurement error in the dependent variable, the estimate for β_1 is still unbiased (problem set 2) but the standard error should rise considerably. In this problem, notice that the two variables with measurement error (y2 and y3) have essentially the same mean as ln_weekly_earn but the variance increases by 1 and 4 respectively over the initial value. Therefore, when we replace ln_weekly_earn with y2 and y3, we see aomw change in the coefficient estimate for β_1 but a large change in the estimate for the standard error of β_1.

2. Because we must use the noisy value of y in our estimates, the parameter estimate for β_1 in this case will be

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^{n}(y_i^* - \bar{y}^*)(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Note that we can write the numerator as $\sum_{i=1}^{n} y_i^*(x_i - \bar{x})$ so the estimate would reduce to

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^{n} y_i^*(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Recall that $y_i^* = y_i + v_i$ *and* $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$ so substitute these values into the equation above

$$\hat{\beta}_1^* = \frac{\sum_{i=1}^{n} y_i^*(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(\beta_0 + x_i\beta_1 + \varepsilon_i + v_i)(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Note that by construction $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ and $\sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n}(x_i - \bar{x})^2$

So the estimate simplifies to

$$\hat{\beta}_1^* = \beta_1 + \frac{\sum_{i=1}^{n}\varepsilon_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{\sum_{i=1}^{n}v_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Taking expectations and remembering that $E[\beta_1] = \beta_1$ we get

$$E[\hat{\beta}_1^*] = \beta_1 + E\left[\frac{\sum_{i=1}^{n}\varepsilon_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] + E\left[\frac{\sum_{i=1}^{n}v_i(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

It is still the case that $E\left[\sum_{i=1}^{n}\varepsilon_i(x_i - \bar{x})\right] = 0$ so the middle term drops out. The final term however is problematic.

For the final term, divide the numerator and denominator by (n-1), so

$$E\left[\frac{\sum_{i=1}^{n}v_i(x_i - \bar{x})/(n-1)}{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}\right] = E\left(\frac{\hat{\sigma}_{xv}}{\hat{\sigma}_x^2}\right) = \left(\frac{\sigma_{xv}}{\hat{\sigma}_x^2}\right)$$

and hence

$$E[\hat{\beta}_1^*] = \beta_1 + \frac{\sigma_{xv}}{\hat{\sigma}_x^2}$$

we assumed that cov(v_i, x_i)<0 so σ_{xv} <0 and therefore $E[\hat{\beta}_1^*] < \beta_1$. We anticipate that regulations reduce drinking so we expect β_1 <0 . However is students are LESS likely to respond in situations with more regulation, then it will appear the regulations are more effective and the estimate for $\hat{\beta}_1^*$ will be biased down.

3. Below are the estimates for the three samples. Notice that the mean beta is pretty similar across the three samples, especially for the last two larger samples (10% and 50%). However, notice the range of outcomes is incredibly large for samples of only 20 (sample 0.1). The range in values across the 20 draws is anywhere from -0.026 to 0.157 which is enormous. Moving to a 50% random sample from the original data, the range now only varies from 0.071 to 0.077. Note moving from a sample of .1 to 10 increases the sample size by a factor of 100 which means the standard deviation in the beta should fall by the square root of 100 or 10 – and it does – moving from 0.039 to 0.0037.

Sample 0.1

Variable	Obs	Mean	Std. Dev.	Min	Max
beta1	20	.0807878	.0397993	-.0257787	.1574895

Sample 10

Variable	Obs	Mean	Std. Dev.	Min	Max
beta1	20	.0729107	.003679	.0643373	.0782484

Sample 50

Variable	Obs	Mean	Std. Dev.	Min	Max
beta1	20	.0742017	.0015058	.0713476	.077194

.

4. Recursively substitute. Note that $y_t = y_{t-1} + \alpha + \varepsilon_t$ and also $y_{t-1} = y_{t-2} + \alpha + \varepsilon_{t-1}$

So $y_t = y_{t-2} + 2\alpha + \varepsilon_{t-1} + \varepsilon_t$. Note also that $y_{t-2} = y_{t-3} + \alpha + \varepsilon_{t-2}$ $y_t = y_{t-3} + 3\alpha + \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t$.

Doing this many times, we will eventually find that

$$y_t = y_0 + t\alpha + \varepsilon_1 + \varepsilon_2 + \ldots \varepsilon_{t-2} + \varepsilon_{t-1} + \varepsilon_t$$

$$y_t = y_0 + \alpha t + \sum_{j=1}^{t} \varepsilon_j$$

$$E[y_t] = y_0 + \alpha t + E\left[\sum_{j=1}^{t} \varepsilon_j\right]$$

We know the last term equals zero so $E[y_t] = y_0 + \alpha t$ meaning the expected value is time dependent – as the period increases, the mean of the dependent variable is changing.

For variance, we know that by definition $V(y_t) = E[(y_t - E(y_t))^2]$. For this case, note that we can write

$y_t = y_0 + \alpha t + \sum_{j=1}^{t} \varepsilon_j$ so

$$V(y_t) = E[(y_t - E(y_t))^2] = E[(y_0 + \alpha t + \sum_{j=1}^{t} \varepsilon_j - y_0 - \alpha t] = E\left[\left(\sum_{j=1}^{t} \varepsilon_j\right)^2\right]$$

And we showed in class that

$$E\left[\left(\sum_{j=1}^{t} \varepsilon_j\right)^2\right] = t\sigma_\varepsilon^2$$

In this case, both the mean and the variance are changing over time so this is a non-stationary series.

5. If $y_t = y_{t-1} + \delta t + \varepsilon_t$ the $y_t - y_{t-1} = \Delta y_t = y_{t-1} - y_{t-1} + \delta t + \varepsilon_t = \delta t + \varepsilon_t$ and $E[\Delta y_t] = E[\delta t] + E[\varepsilon_t] = \delta t$

In this case, the differencing does not produce a stationary series because the mean is still a function of time. Note that

$$Var[\Delta y_t] = E[(\Delta y_t - E[\Delta y_t])^2] = E[(\delta t + \varepsilon_t - \delta t)^2] = E[(\varepsilon_t)^2] = \sigma_\varepsilon^2$$

$$\begin{aligned} Cov[\Delta y_t, \Delta y_{t-1}] &= E[(\Delta y_t - E[\Delta y_t])(\Delta y_{t-1} - E[\Delta y_{t-1}])] \\ &= E[(\delta t + \varepsilon_t - \delta t)(\delta(t-1) + \varepsilon_{t-1} - \delta(t-1))] = E[\varepsilon_t \varepsilon_{t-1}] = 0 \end{aligned}$$

6. a) Notice that since $y_t = y_{t-1} + \varepsilon_t$ then $y_{t+1} = y_t + e_{t+1}$. Taking expectations condition on observing y_t
Then $E[y_{t+1} \mid y_t] = y_t + E[e_{t+1} \mid y_t]$. Since $E[e_{t+1} \mid y_t]$=0 then $E[y_{t+1} \mid y_t] = y_t$

b) By construction, $y_{t+2} = y_{t+1} + e_{t+2}$ and since $y_{t+1} = y_t + e_{t+1}$ then
$y_{t+2} = y_{t+1} + e_{t+2} = [y_t + e_{t+1}] + e_{t+2} = y_t + e_{t+1} + e_{t+2}$. Since $E[e_{t+1} \mid y_t] = 0$ and $E[e_{t+2} \mid y_t] = 0$ then $E[y_{t+2} \mid y_t] = E[y_t \mid y_t] + E[e_{t+1} \mid y_t] + E[e_{t+2} \mid y_t] = y_t$

c) Since $E[y_{t+1} | y_t] = y_t$ and $E[y_{t+2} | y_t] = y_t$ then we can safely conclude that $E[y_{t+h} | y_t] = y_t$

7. Thus was a crowd-sourced Monte Carlo. It was clear that 64 of the 65 students did the exercise correctly. There were a total of 64*10 or 640 regressions. In the first part, a total of 593 or 92.7% of the models rejected the null. The implication is that with non-stationary series, Type I error rates are real high.

   The second part of the exercise was to make all of the series stationary by first differencing and then see how many rejections of the null we obtain. The Type I error rate should be 5% and low and behold, there were 32 rejections in all of the 640 regressions or EXACTLY 5%.

   Not bad.

8. A sample program called wilcox.do is on the class web page.

   a) Below are regression results for parts a and b. Note that the coefficient on the 1st difference in ln(OASI) is 0.075 and the lag is 0.053. These are close to the estimates in Wilcox – but not spot on. The P-value on the F test that the coefficients on d_oasi_ln and d_oasi_ln_1 are zero is 0.027 so we can reject the null.

```
. * question a -- replicate wilcox results
. reg d_retail_ln time d_oasi_ln d_oasi_ln_1

    Source |       SS       df       MS              Number of obs =     250
-------------+------------------------------           F(  3,   246) =    2.45
       Model |  .001273058     3  .000424353           Prob > F      =  0.0643
    Residual |  .042638714   246  .000173328           R-squared     =  0.0290
-------------+------------------------------           Adj R-squared =  0.0171
       Total |  .043911772   249  .000176352           Root MSE      =  .01317

------------------------------------------------------------------------------
 d_retail_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |  -5.13e-07   .0000116    -0.04   0.965    -.0000233    .0000222
   d_oasi_ln |   .0757343   .0331578     2.28   0.023      .010425    .1410437
 d_oasi_ln_1 |   .0535633   .0331576     1.62   0.108    -.0117456    .1188722
       _cons |   .0004101   .0017005     0.24   0.810    -.0029393    .0037596
------------------------------------------------------------------------------

.
. * part b -- generate the f test
. test d_oasi_ln d_oasi_ln_1

 ( 1)  d_oasi_ln = 0
 ( 2)  d_oasi_ln_1 = 0

       F(  2,   246) =    3.66
            Prob > F =    0.0272
```

c. The results when levels are used instead of 1st differences are quite different. The coefficient for on oasi_ln triples and the coefficient for oasi_ln_1 increases by a factor of 5. Recall that when you regress a non-stationary series on a non-stationary series, there is a high type I error rate due to spurious correlation. This is just that example.

```
. * part c
. * run the model but ignore the fact that
. * variables are non stationary and regress
```

```
. * levels on levels
. reg retail_ln time oasi_ln oasi_ln_1

     Source |       SS       df       MS              Number of obs =     251
-------------+------------------------------           F(  3,   247) =  119.50
       Model |  .554763472     3  .184921157           Prob > F      =  0.0000
    Residual |  .382233746   247  .001547505           R-squared     =  0.5921
-------------+------------------------------           Adj R-squared =  0.5871
       Total |  .936997218   250  .003747989           Root MSE      =  .03934

------------------------------------------------------------------------------
   retail_ln |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |  -.0007771   .0001065    -7.29   0.000     -.000987   -.0005673
     oasi_ln |   .2016764   .1006617     2.00   0.046     .0034116    .3999412
   oasi_ln_1 |   .3035478   .1010089     3.01   0.003     .1045991    .5024964
       _cons |   3.220011   .2197767    14.65   0.000     2.787135    3.652886
------------------------------------------------------------------------------

. test oasi_ln oasi_ln_1

 ( 1)  oasi_ln = 0
 ( 2)  oasi_ln_1 = 0

       F(  2,   247) =   78.48
            Prob > F =    0.0000
```