Name: ______________________________ Fall 2016

EC200 Econometrics and Applications

Unit 3 Quiz

You have 2 hours to complete this quiz. There are 52 total points. Please show all your work to receive full credit.

1. *(10 points)* For each of the following terms, provide a definition (one to two sentences). You may find it helpful to use an example.

Term	Definition *[2 points each]*
Repeated cross-section	
Non-classical measurement error	
External validity	
Instrumental variable	
~~Proxy variable~~	

2. *(12 points)* Suppose you have detailed self-reported survey data and want to estimate the determinants of depression using the following linear probability model:

$$Pr(Depress) = \beta_0 + \beta_1 Exercise + \beta_2 Female + \beta_3 TVhours + \beta_4 Age + \beta_5 Educ + u,$$

where *Depress* is a binary variable equal to one if the person is experiencing a major depressive episode, *Exercise* is number of hours of exercise per week, *Female* is a binary variable equal to one the person is female, *TVhours* is the number of hours of TV watched per week, *Age* is age in years, and *Educ* is years of completed education.

(a) Provide a real-world example of *classical* measurement error in one of the independent variables. What assumption(s) do you make for it to be classical? If the error really is classical, what is the impact on your estimate of that variable's $\widehat{\beta}$? *[6 points]*

Example:

Assumption(s):

Impact on estimate if classical:

(b) Provide an example of reverse causality that might arise in this model. *[3 points]*

(c) Yolanda hypothesizes that there is a non-linear relationship between age and the likelihood of depression. Explain how you would test whether her hypothesis is correct. *[3 points]*

3. *(12 points)* The recent legalization of recreational marijuana in Massachusetts may provide an interesting policy experiment for researchers! Dr. Ong is interested in the impact of marijuana use on high-school drop-out rates. In Massachusetts, possession of marijuana will become legal in 2017, and licenses to sell marijuana will be available in 2018. Suppose that the supply of marijuana will be greatest in counties that already have at least one medicinal marijuana clinic (these clinics will be given preference when applying for licenses). Dr. Ong has the following data from Massachusetts:

- County-level data on high-school drop-out rates in 2016 and 2018 ($Dropout_{c,y}$, where c is county and y is year)
- Number of medicinal marijuana clinics by county, as of 2016. ($Clinics_c$)

(a) Write a difference-in-differences population model to measure the impact of marijuana legalization on high-school drop-out rates. If you use any new variables, make sure to define them clearly. *[6 points]*

(b) What assumption(s) do you need to make for your difference-in-differences model to reflect the *causal* impact of legalization on drop-out rates? Explain what each assumption means. You can include a picture if it is helpful. *[3 points]*

(c) Suppose that counties with more marijuana clinics are also poorer, and that poorer areas have higher drop-out rates. How would this affect your estimate of the impact of marijuana legalization, if at all? Explain. *[3 points]*

4. *(6 points)* Vella and Veerbeek (1998) use longitudinal panel data from the U.S. National Longitudinal Survey of Youth (NSLY) to track working-age men from 1980-1987. They estimate the following fixed-effects model of the impact of being in a union on wages:

$$lwages_{it} = \beta_0 + \beta_1 union_{it} + a_i + u_{it}$$

where $lwages_{it}$ is the log of real hourly wages for individual i in year t and $union_{it}$ is a binary variable equal to 1 if individual i in year t is a member of a labor union.

They get the following results:

```
. xtreg lrent y90 lpop lavginc pctstu,fe

Fixed-effects (within) regression               Number of obs      =       128
Group variable: city                            Number of groups   =        64

R-sq:  within  = 0.9765                         Obs per group: min =         2
       between = 0.2173                                        avg =       2.0
       overall = 0.7597                                        max =         2

                                                F(4,60)            =    624.15
corr(u_i, Xb)  = -0.1297                        Prob > F           =    0.0000

------------------------------------------------------------------------------
       lrent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         y90 |   .3855214   .0368245    10.47   0.000     .3118615    .4591813
        lpop |   .0722456   .0883426     0.82   0.417    -.104466     .2489571
     lavginc |   .3099605   .0664771     4.66   0.000     .1769865    .4429346
      pctstu |   .0112033   .0041319     2.71   0.009     .0029382    .0194684
       _cons |   1.409384   1.167238     1.21   0.232    -.9254394    3.744208
-------------+----------------------------------------------------------------
     sigma_u |  .15905877
     sigma_e |  .06372873
         rho |   .8616755   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(63, 60) =     6.67             Prob > F = 0.0000
```

(a) Interpret the coefficient on $\widehat{\beta}_1$. *[3 points]*

(b) Richard suggests that you add race/ethnicity controls to your model of wages and union membership because African-Americans are more likely to be members of labor unions. Should you add these controls? Explain. *[3 points]*

5. *(12 points)* At the Fulton Fish Market in New York City, sellers bring in just-caught fish to sell and negotiate prices with buyers. As a result, the average price and quantity sold fluctuate daily. Graddy (1995) collected data on individual transactions over time at the Fulton Fish Market. Consider the following model of demand for fish.

$$lavgprc = \beta_0 + \beta_1 lavgqty + u$$

where *lavgprc* is the log of the daily average price of fish sold and *lavgqty* is the log of the daily average quantity of fish sold. The regression results follow:

```
. regress children educ age agesq
      Source |       SS       df       MS              Number of obs =    4361
-------------+------------------------------           F(  3,  4357) = 1915.20
       Model |  12243.0295     3  4081.00985           Prob > F      =  0.0000
    Residual |  9284.14679  4357  2.13085765           R-squared     =  0.5687
-------------+------------------------------           Adj R-squared =  0.5684
       Total |  21527.1763  4360  4.93742577           Root MSE      =  1.4597

------------------------------------------------------------------------------
    children |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -.0905755   .0059207   -15.30   0.000     -.102183   -.0789679
         age |   .3324486   .0165495    20.09   0.000     .3000032     .364894
       agesq |  -.0026308   .0002726    -9.65   0.000    -.0031652   -.0020964
       _cons |  -4.138307   .2405942   -17.20   0.000    -4.609994    -3.66662
------------------------------------------------------------------------------
```

(a) Interpret the magnitude of $\widehat{\beta_1}$. That is, what does -0.138 mean? *[3 points]*

(b) Graddy estimates a two-stage least squares (2SLS) model by using weather as an instrument for the quantity of fish sold. Specifically, her instrument is the maximum height of waves averaged over the past two days, *wave*2. Explain why this might be a reasonable instrument. *[3 points]*

(c) Using the estimated 2SLS results below, interpret the coefficient on *ltotqty*. That is, what does −1.176 mean? *[3 points]*

```
. ivregress 2sls children (educ = frsthalf) age agesq
Instrumental variables (2SLS) regression              Number of obs =    4361
                                                      Wald chi2(3)  = 5300.22
                                                      Prob > chi2   =  0.0000
                                                      R-squared     =  0.5502
                                                      Root MSE      =    1.49
```

children	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	-.1714989	.0531553	-3.23	0.001	-.2756813	-.0673165
age	.3236052	.0178514	18.13	0.000	.2886171	.3585934
agesq	-.0026723	.0002796	-9.56	0.000	-.0032202	-.0021244
_cons	-3.387805	.5478988	-6.18	0.000	-4.461667	-2.313943

```
Instrumented:  educ
Instruments:   age agesq frsthalf
```

(d) Graddy also reports her first-stage results below. Do they raise any concerns about the validity of her instrumental variables strategy? Why or why not? *[3 points]*

```
. reg lrent y90 lpop lavginc pctstu
```

Source	SS	df	MS
Model	12.1080112	4	3.02700281
Residual	1.9501234	123	.015854662
Total	14.0581346	127	.110693974

```
Number of obs =     128
F(  4,   123) =  190.92
Prob > F      =  0.0000
R-squared     =  0.8613
Adj R-squared =  0.8568
Root MSE      =  .12592
```

lrent	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
y90	.2622267	.0347632	7.54	0.000	.1934151	.3310384
lpop	.0406863	.0225154	1.81	0.073	-.0038815	.0852541
lavginc	.5714461	.0530981	10.76	0.000	.4663417	.6765504
pctstu	.0050436	.0010192	4.95	0.000	.0030262	.007061
_cons	-.5688069	.5348808	-1.06	0.290	-1.627571	.4899568