Name: ____ Fall 2016

EC200 Econometrics and Applications

Unit 3 Quiz

You have 2 hours to complete this quiz. There are 52 total points and one extra credit question worth up to 6 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	Data collected repeatedly over time, with new random sample taken each time. For example, the CPS or US census is a repeated cross-section
Non-classical measurement error	When the true value of an variable is not observed, and the difference between the true and observed value is correlated with the error term
External validity	Whether the results from a specific study can be generalized to a broader population/context
Instrumental variable	a. A variable that is exogenous to the dependent variable but correlated with the endogenous independent of interest. It must be relevant (Cov(Z,X) not equal to zero) and excludable (Cov (Z,u) = 0) For example, rainfall may be an instrument for income when estimating the impact of income on nutrition.
Proxy variable	A variable that is correlated with an unobservable variable, which we include in our OLS regression in order to account for omitted variable bias. For example, we use IQ as a proxy for ability.

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: $Exercise$ may be measured with error if people don't accurately time the number of hours spent exercising or they don't always correctly remember.

Assumption(s):To be classical, it can't be correlated with the error term.

Impact on estimate if classical: If it is classical, then the estimate $\widehat{\beta}_1$ is biased towards zero as a results of attenuation bias.

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

If depression causes people to exercise less and watch more TV, then we would see reverse causality.

(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]

You could estimate this model, adding in a term form age^2 or even age^3. To test whether the relationship is non-linear, you'd test the null hypothesis that $\beta_{age^2} = \beta_{age^3} = 0$ against the alternative hypothesis that at least one is not equal to zero. If you just added age^2, you could use the t-statistic from the regression to determine its significance. With multiple terms, you'd need to conduct an F-test of the joint significance of all higher-order terms.

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]

$$Dropout_{c,y} = \beta_0 + \beta_1 Year2018_y + \beta_2 AnyClinic2016_c + \beta_3 Year2018_y * AnyClinic2016_c + u$$

where $Year2018$ equals 1 if the year is 2018 and $AnyClinic2016_c$ is a variable equal to 1 if county c has at least one medicinal marijuana clinic in 2016.

(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]*

For estimates of β_3 to reflect the causal impact, you need to assume "parallel trends" - that is, you need to assume that between counties with and without medicinal marijuana clinics in 2016, the trend in high-school drop-out rates would be the same in the absence of marijuana legalization.

(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]*

This would not affect our ability to measure the impact of marijuana legalization, unless the trend in high-school drop-out rates is different between poorer and richer areas. So long as it's just a difference in levels, then it won't be a problem, as it will be captured by β_2.

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 lwage union ,fe

Fixed-effects (within) regression               Number of obs      =      4360
Group variable: nr                              Number of groups   =       545

R-sq:  within  = 0.0032                         Obs per group: min =         8
       between = 0.0454                                        avg =       8.0
       overall = 0.0209                                        max =         8

                                                F(1,3814)          =     12.39
corr(u_i, Xb)  = 0.1164                         Prob > F           =    0.0004

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |   .0746846   .0212205     3.52   0.000     .0330801    .1162891
       _cons |   1.630921   .0078171   208.64   0.000     1.615595    1.646248
-------------+----------------------------------------------------------------
     sigma_u |  .38625193
     sigma_e |   .3866551
         rho |  .49947837   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(544, 3814) =     7.88           Prob > F = 0.0000
```

(a) Interpret the coefficient on $\widehat{\beta_1}$. *[3 points]*

Being in a union is associated with a 7.5% increase in real hourly wages.

(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]*

There is no need to add these controls because we already include fixed effects, which capture any time-invariant individual-level characteristics, such as race.

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:

```
. reg  lavgprc ltotqty,robust

Linear regression                                      Number of obs =      97
                                                       F(  1,    95) =    9.08
                                                       Prob > F      =  0.0033
                                                       R-squared     =  0.0681
                                                       Root MSE      =  .39259

------------------------------------------------------------------------------
             |               Robust
     lavgprc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ltotqty |   -.138111   .0458421    -3.01   0.003    -.2291192   -.0471029
       _cons |   .8710151   .3773548     2.31   0.023     .1218711    1.620159
------------------------------------------------------------------------------
```

(a) Interpret the magnitude of $\widehat{\beta}_1$. That is, what does -0.138 mean? *[3 points]*

A 1% increase in the daily quantity of fish sold is associated with a 0.138% decrease in the price of fish.

(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]*

This would be a reasonable instrument because weather is likely to affect the quantity of fish caught, and therefore supplied to the market - as big waves would reduce fishers' ability to catch fish. However, since it is the average weather from previous days, it's unlikely to affect the number of people (the demand) for fish - so it is both relevant and plausibly exogenous.

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

```
. ivregress 2sls  lavgprc (ltotqty = wave2) ,robust
Instrumental variables (2SLS) regression              Number of obs =      97
                                                      Wald chi2(1)  =    4.11
                                                      Prob > chi2   =  0.0426
                                                      R-squared     =       .
                                                      Root MSE      =  .87962

------------------------------------------------------------------------------
             |               Robust
     lavgprc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ltotqty |  -1.175514   .5797975    -2.03   0.043    -2.311896    -.039132
       _cons |   9.259001    4.70206     1.97   0.049     .0431333    18.47487
------------------------------------------------------------------------------
Instrumented:  ltotqty
Instruments:   wave2

.
```

A 1 percent increase in total quantity of fish sold is associated with a 1.76-percent decrease in average prices.

(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 ltotqty  wave2 ,robust
Linear regression                                     Number of obs =      97
                                                      F(  1,    95) =    3.88
                                                      Prob > F      =  0.0519
                                                      R-squared     =  0.0456
                                                      Root MSE      =  .75093

------------------------------------------------------------------------------
             |               Robust
     ltotqty |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       wave2 |   -.091397   .0464193    -1.97   0.052    -.1835511     .000757
       _cons |   8.551025   .2408214    35.51   0.000     8.072934    9.029116
------------------------------------------------------------------------------

.
```

While it looks like wave height does predict reduced fish sales, the correlation is not very strong The F-statistic is only 3.88, indicating that this is a weak instrument. As a rule of thumb, we look for F-statistics of at least 10. As a result, our two-stage least squares estimates are likely to be biased