Problem Set 8
Economics 30331
(Due by noon on Friday, May 4th to my office)
[questions with a * are questions used on previous finals for this class]

Bill Evans
Spring 2018

1. In the following three problems are samples sizes, covariates and Durbin-Watson statistics. In each case, decide whether you can reject or not reject the null of no auto correlation or are the results inconclusive.
   a. N=25, k=5, $\hat{d} = 1.80$
   b. N=60, k=9, $\hat{d} = 0.23$
   c. N=45, k=2, $\hat{d} = 1.40$

2. *Consider a simple bivariate regression of the form $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$. A researcher obtains estimates of $\hat{\beta}_0$ *and* $\hat{\beta}_1$ through "ordinary least squares." For each question below, answer true or false about these estimates.

   a) _____ Suppose the model above is subject to first order autocorrelation, $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$. In the presence of autocorrelation, $\hat{\beta}_1$ is a **biased estimate,** $E[\hat{\beta}_1] \neq \beta_1$.

   b) _____ Suppose the model above is subject to first order autocorrelation, $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$ where $\rho > 0$ and we anticipate that x_t is positively correlated over time. In the presence of positive autocorrelation, the estimated variance of $\hat{\beta}_1$ **is too small.**

   c)_____ Suppose the model above is subject to first order autocorrelation, $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$. The model is also estimated using the Cochrane-Orcutt procedure to correct for autocorrelation. The Cochrane-Orcutt procedure will produce an estimate for $\hat{\beta}_1$ that is identical to the estimate generated from an OLS procedure that ignores autocorrelation.

   d) _____ Suppose that the dependent variable y_i in the model above is subject to classical measurement error where y_i^* is the true value, v_i is a random error and the measured value y_i equals $y_i = y_i^* + v_i$. The OLS estimate of $\hat{\beta}_1$ in this case will be a **biased estimate,** $E[\hat{\beta}_1] \neq \beta_1$.

   e) _____ Suppose that the dependent variable y_i in the model above is subject to classical measurement error where y_i^* is the true value, v_i is a random error and the measured value y_i equals $y_i = y_i^* + v_i$. The OLS estimate of $\hat{\beta}_1$ in this case will produce a larger variance on $\hat{\beta}_1$ than if y_i is measured without error.

f) _____ Suppose that the independent variable x is subject to classical measurement error where x_i^* is the true value, w_i is a random error and the measured value x_i equals $x_i = x_i^* + w_i$. The OLS estimate of $\hat{\beta}_1$ in this case will be **an inconsistent estimate.**

3. In 1993, Michigan voters passed a referendum eliminating local property taxes, which are the main source of revenues for schools. To make up for lost revenue, the Michigan legislature raised the cigarette tax from 25 to 75 cents per pack. The higher tax rate went into effect on May 1, 1994. The Surgeon General of the US estimates that smoking during pregnancy doubles the chance a baby will be born with a low birth weight (<2500 grams). Although smoking rates among pregnant women have fallen considerably over the past 20 years, roughly 17 percent of births are to women who smoked during their pregnancy during this period. In recent years, a number of public health officials have suggested that higher cigarette taxes can be used as way to improve birth outcomes. We will use the data from the Michigan "experiment" to evaluate this whether higher taxes reduce smoking among pregnant women.

   The data for this project are taken from the Natality Detail File, which is an annual census of births in the US. I have taken a 20% random sample of births for the state of Michigan for the 2 years months prior and 12 months after the tax hike. I have also include a 20% random sample of data over the same period for a Midwestern state that had no nominal change in their state cigarette tax rates over this period: Iowa.

   The data set is names michigan_tax_hike.dta. The data set has 101,676 observations. There are only three variables in the data set and variable definitions are listed below.

Variable	**Definition**
state	2-digit state FIPS code. Michigan is state 26, Iowa is 19.
smoked	Dummy variable, =1 if a mother self-reported that she smoked during her pregnancy, =0 otherwise.
year	The years in the data set. =1 is the data is from 2 years before the tax hike, =2 the year before the tax hike, =3 the first year after the tax hike.

   a. Construct two dummy variables, one called michigan, (`gen michigan=state==26`) and another for after the tax hike goes into effect (`gen after=year==3`). Now, get the means of "smoked" for the four boxes necessary to construct a simple difference-in-difference estimate.

   b. Construct the 2 x 2 table necessary to generate the difference-in-difference estimate. What is this estimate?

```
sort michigan after
by michigan after:  sum smoked
```

   c. Now, estimate the difference in difference estimate part a) by running a regression. Construct the "treatment" effect variable which equals 1 in Michigan after the tax hike

```
gen treatment=michigan*after
```

   Now run a regression of smoked on michigan, after and treatment.

```
reg smoked michigan after treatment
```

Compare the coefficient on "treatment" with the difference-in-difference estimate you generated in part a). Are they the same? At the 95% confidence level, can you reject the null that the tax hike had no impact on smoking rates among pregnant women?

d. What is the key assumption in difference-in-difference models? We can never test this assumption directly but what we can do is provide some data that suggests the time series of outcomes is similar before the intervention. Note that the data set has information for the two years prior to the tax hike. In this portion of the problem, you are to run a difference-in-difference model using only data from years 1 and 2, assuming the treatment occurred in year 2 and deleting data from year 3. So, construct a near variable after2 that equals 2 in years >=2 and a new treatment variable that equals michigan*after2. Next, run a regression of smoked on michigan, after2 and treatment2 but only for years<=2.

```
gen after2=year>=2
gen treatment2=after2*michigan
reg smoked michigan after2 treatment2 if year<=2
```

What is the coefficient on treatment2 and can we reject the null that the coefficient on treatment2 equals 0? What does this say about using Iowa as a control group for Michigan in this context?

4. Throughout the year, we have been using a data set that has data on per capita cigarette consumption for all states and DC over the period 1981-2000. This data set is called state_cig_data.dta. This data has information on per capita packs, state and federal taxes, per capita income, the consumer price index (cpi), plus variables that identify the state and year. In this problem, we will examine what happens to the estimate of the real tax coefficient when we progressively add more variables to the regression.

Initially, do three things, construct real taxes, real ln(per capita income) and ln(per capita sales)

```
gen real_tax=(state_tax+federal_tax)/cpi
gen rpcil=ln(pci/cpi)
gen packs_pc_l=ln(packs_pc)
```

Next, generate dummies for all state and year effects except for the reference year

```
xi i.state i.year
```

Next, run three models. Model 1 is a regression of packs_pc_l on real_tax and rpcil. In model 2, add state effects, and in model 2, add year effects.

```
reg packs_pc_l rpcil real_tax
reg packs_pc_l rpcil real_tax _Is*
reg packs_pc_l rpcil real_tax _Is* _Iy*
```

What is happening to the coefficient on real_tax as more variables are added to the model? In model (3), can we reject the null that the coefficient is 0? Interpret the coefficient on real_tax in model 3. What is the impact of a 10 cent/pack increase in taxes on cigarette consumption?

5. If the simple bivariate equation $y_i=\beta_0+x_i\beta_1+\varepsilon_i$ is estimated by 2SLS and the only instrument is the variable z_i, one can show that

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

a) Show that $\text{var}(\hat{\beta}_1)$ can be re-written to read

$$\text{var}(\hat{\beta}_1^{2sls}) = \frac{\text{var}(\hat{\beta}_1^{OLS})}{\hat{\rho}(z,x)^2}$$

Where $\text{var}(\hat{\beta}_1^{OLS})$ is the variance of the OLS estimate of β_1.

b) Suppose that the instrument z_i does a poor job of explaining the variation in x. In this case, what happens to $\text{var}(\hat{\beta}_1^{2sls})$?

6. *Recent research has demonstrated that the obese tend to have lower wages than lighter people. The typical model is generated through an OLS model of the form $y_i=\beta_0 +x_i\beta_1+\varepsilon_i$ where y is the natural log of wages of earnings and x is a measure of weight such as body mass index (BMI) which is weight in kilograms divided by height in centimeters squared. Most models find that $\hat{\beta}_1 < 0$ indicating that heavier people have lower wages. The concern is obviously that the obese are not a random selection of the population and as a result, this estimate for $\hat{\beta}_1$ is subject to an omitted variables bias or that $\text{cov}(x_i, \varepsilon_i) \neq 0$. To address this concern, an author uses a 2SLS model and as an instrument (z_i) for x_i the author uses the BMI of a sibling. The author argues that the weights of siblings should be correlated but that the fact that a sibling is heavy should not be correlated with ε_i.
   a. Is this a reasonable assumption, do you think that $\text{cov}(z_i, \varepsilon_i) = 0$ in this instance? Why might the instrument **not** satisfy the exclusion restriction in this context?
   b. If not, use the results from the consistency of 2SLS to sign the direction of the bias in the 2SLS estimate in this case?

7. *(Parts of this was on a final) To obtain a consistent estimate of the impact of kids on labor supply, some authors have suggested using whether a mother had twins on their first birth as an instrument for the number of children in the household. Twins are in many respect random and by definition, the realization of a twin increases the number of children in the household. Using data from the 1980 Public Use Micro Sample 5% Census data files, I constructed a sample of women aged 21-40 with at least one kid. The 1980 PUMS identifies a person's age at the time of then census and their quarter of birth. Because the census is taken on April 1st, we know a person's year and quarter of birth and we can infer that any two kids in the household with the same age and quarter of birth are twins. There are roughly 6,000 1st births to mothers that are twins. There are over 800,000 observations in the original data set so to make the problem manageable, I select a random sample of about 6,500 non-twin births for a total of about 12,500 observations. The STATA data file is called twins1st.dta and below are detailed descriptions of the variables.

Variable name	Description
Age	Mother's current age in years
Agefst	Mom's age when she first gave birth
Race	1=white, 2=black, 3=other race
Educ	Mother's years of education

Married	Dummy variable for current marital statue, 1= married, 0=not
Kids	Number of children ever born to the mother
boy1st	Dummy variable, =1 if first kid is a boy, =0 otherwise.
twin1st	Dummy variable, =1 if the first pregnancy ended in a twin birth
Weeks	Weeks worked in previous year (from 0-52)
Worked	Dummy variable, = 1 if the Mom worked at all in the previous year
Lincome	Labor income earned in the previous year

a. What fraction of women work? What is average weeks worked among women that work? What is median labor earnings for women who worked?

b. Construct an indicator that equals 1 for women that have a second child. Call this variable SECOND. What fraction of women had a second child? Consider a simple bivariate regression where WEEKS of work (Y) is regressed on SECOND (X), $Y = \beta_0 + \beta_1 X_i + \varepsilon_i$. What is the coefficient for β_1 in this regression and interpret the coefficient?

c. Because of the concern that X and ε are correlated, use twins on 1st birth (Z) as an instrument for X in an instrumental variables model. What is the first-stage and reduced-form estimates for this model? Interpret these coefficients, that is, what do these coefficients measure? Consider the regression of X on Z. Why is the coefficient on Z not 1 - e..g, don't twins increase the number of kids in the house by 1? What is the indirect least squares estimate for β_1 and compare the coefficient to the OLS estimate you produced above?

d. Now, estimate the basic model $Y = \beta_0 + \beta_1 X_i + \varepsilon_i$ by 2SLS using twin1st as an instrument. Using "weeks" as the outcome of interest and "second" for x. The statement in STATA will be

```
ivregress 2sls weeks (second=twin1st)
```

e. Now, expand the 2SLS model to include some other covariates. First, generate dummy variables for mothers that are black and other_race. Run a structural labor supply models with weeks worked as the outcome (y) and control for mothers age, age1st, educ, black, other race, married and whether the mother has a second child. What is the impact of a second child on weeks worked?

f. Now, use twin1st as an instrument for the second child the model above. Compare these estimates to the results in part d. Next, compare these results to the simple 2SLS estimates in b). What has happened to the labor supply impacts of having a second child? Explain why this is the case.

```
Ivregress 2sls weeks age age1st educ black other_race
(second=twin1st)
```

8. *It is easy to construct a case that the US spends way too much money on health care. The US spends twice as much as the average OECD country, 90% more than the Canada and 150% more than the UK, yet the US has one of the lowest life expectancy rates and one of the highest infant mortality rates in the developed world. Some have suggested these numbers indicate that at the margin, additional health care dollars generate little in the way of better health outcomes.

An author is interested in examining whether greater health care spending produces better outcomes for newborns. The author focuses on newborns because childbirth is the most frequent reason for a hospital admission and the average hospital stay for childbirth costs about $6000.

The author starts the analysis by considering a very simple question: are outcomes better for newborns that receive more care? The author takes a sample of newborns and regresses whether the child died within 28 days of birth (y=1 if they dies, y=0 otherwise) on the hospital expenditures for the newborn right after their birth (x). This basic regression is defined as equation (1) $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$ and the model is estimated by OLS.

a) Is the OLS estimate of $\hat{\beta}_1$ from the simple bivariate model above an unbiased estimate of the impact of health care spending on infant mortality? Why or why not? If the estimate is biased, provide an equation that illustrates whether the estimate is biased up or down.

Consider the following Regression Discontinuity Design model constructed to estimate the impact of greater health care spending on newborn health. The author's exploit the fact that many hospitals have rules requiring greater care for newborns with particular characteristics. The average child weighs 3500 grams at birth and low birth weight is a good predictor of later outcomes. As a result, in some hospitals, newborns with very low weight infants (those < 1500 grams) are sent to directly to the neonatal intensive care units (NICU) or hospital wards with greater nurse supervision. The authors hope to exploit the difference in health care use at 1500 grams in a 'regression discontinuity' model to estimate the benefits of greater health care on child outcomes.

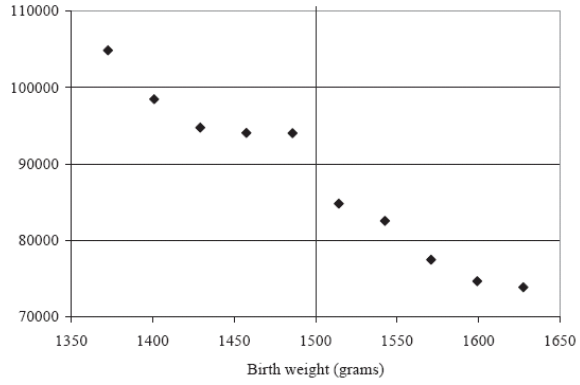
The authors collect data on a large sample of children with low birth weights (1350 to 1650 grams). The outcome of interest (y) is a dummy variable that equals 1 if a child dies within 28 days of birth. The key covariate is x (hospital spending in dollars for the newborn). In the first stage and reduced-forms, the authors control for the birth weight in grams (BW) and a dummy variable (D) that equals 1 if the newborn is less than 1500 grams in weight.

First stage: $x_i = \alpha_0 + BW_i\alpha_1 + D_i\alpha_2 + u_i$

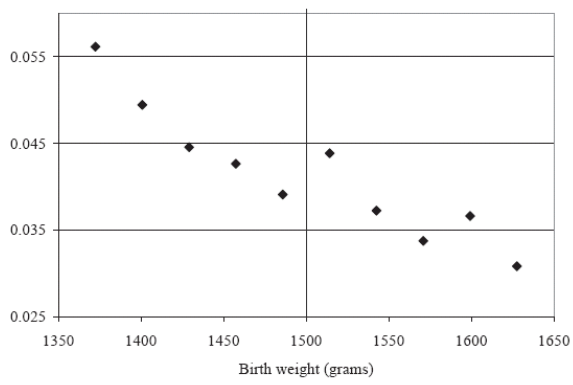
Reduced form: $y_i = \gamma_0 + BW_i\gamma_1 + D_i\gamma_2 + v_i$

Figure A below is a graphical presentation of the data for the first stage and figure B contains the reduced form. The table below has the results from the first stage and reduced forms. Please answer the questions on the next page.

A: Hospital charges



B: 28-Day mortality



Results for Question 8

	(X) Hospital Charges in dollars	(Y) The newborn died within 28 days
Constant	260,250	0.168
	(23,000)	(0.021)
BW (in grams)	-115	-0.000083
	(15.1)	(0.00002)
D (BW<1500 grams)	7670	-0.0228
	(2300)	(0.003)

b) What do the results and the two figures suggest is the relationship between greater health care spending and outcomes for low weight newborns?

c) Using the results from both the first-stage and reduced-form models, calculate the 2SLS estimate of greater hospital spending on 28-day infant mortality. Interpret this coefficient – what happens to the probability the newborn dies within 28 days if spending increases by $10,000?

d) What assumption must be correct in order for the estimate in part c) to be a consistent estimate of the impact of greater health care spending on outcomes?

e) Suppose a public health advocate uses the results in part c) to argue for more health care spending for newborns in general. Using the properties of RDD models, what word of caution would you have for this person?