Problem Set 3
ECON 30331
(Due by the start of class, Wednesday, February 7, 2018)
(Problems marked with a * are former test questions)

Bill Evans
Spring 2018

1. On the Stata page for this class is a copy of the data for the cigarette tax problem we have been using as an example in class. The file is called state_cig_data.dta. Download the data set then load that data set up into Stata and construct two new variables: The natural log of per capita consumption ln(Q) and the natural log of the real retail price (ln(P)). For this second variable, use the statement, gen ln_r_price=ln(retail_price/cpi). Next, run a regression of ln(Q) on ln(P), or the model $\ln(Q_i)=\beta_0+\ln(P_i)\beta_1+\varepsilon_i$.

   a. What are the estimates for $\hat{\beta}_1$ and $\hat{\beta}_0$ and what is the R^2 for the model?

   b. Next, interpret the estimate for $\hat{\beta}_1$. Be precise, explain the units of measure on the variable and give a numeric example.

   After you run the regression, output the residuals from the regression. To do this, right after the "reg" statement, type

   ```
   predict error, residuals
   ```

   c. What is the mean value of the residuals.

   Next, calculate the correlation coefficient between the error and ln_r_price. You can get this with the command **corr error ln_r_price**

   d. What is that value?

1. With the same data in problem 1, keep data for 1985 only by typing

   ```
   keep if year==1985
   ```

   This will produce a data set of 51 observations. Then, run a regression of pc_packs on the federal_tax. What do you get as the coefficient on this variable? Provide an explanation using appropriate equations why you cannot obtain an estimate in this case.

2. A dean is trying to evaluate the quality of instruction of the professors in her college. At the end of every semester students evaluate the professors on a scale of 1 to 10 with 10 being outstanding and 1 being horrible. Label this score Y_i. The dean gives an award to the 5 professors with the highest scores. A faculty member with some statistical training notes that grades tend to be higher in classes that are graded more easily. Let x_i be the class average GPA and the professor runs a regression of the form $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$ and finds $\hat{\beta}_1 > 0$ – professors that give higher grades tend to have much higher teacher evaluations. The professor suggests that awards should be adjusted for how easily a professor grades. Within the context of a regression model, explain how one could adjust teaching scores to reflect the fact that some teachers are easier graders than others in awarding the teaching prize.

3. In a simple bivariate regression model, $Y_i=\beta_0 + X_i\beta_1 +\varepsilon_i$, using the three key assumptions we make about ε_i, show that the OLS estimate for $\hat{\beta}_0$ is unbiased.

4. *There are dozens of sites on the internet where one could illegally download movies to watch on their device. A researcher is interested in examining the impact of illegal movie downloads on commercial movie sales such as blueray disks or download on places like iTunes or Amazon. The author collects data on commercial sales of the top 200

movies from 2007 (Y) and the number of downloads from a web site that allows 'file sharing' (X). The author estimates an OLS model of the form $Y_i=\beta_0 +X_i \beta_1+\varepsilon_i$ and the author gets a large positive coefficient on the estimated parameter $\hat{\beta}_1$. The author concludes these results demonstrate that downloads actually spur on movie sales. Is this an unbiased estimate of the impact of illegal download on sales? Why or why not? Do you expect the estimate to overstate or understate the true relationship between Y and X? Please provide all appropriate equations.

5. *A pharmaceutical company is interested in estimating the impact of a new drug on cholesterol levels. They enroll 200 people in a clinical trial. People are randomly assigned dosage levels or they are randomly assigned into the control group. Half of the people are given dosages of the new drug and half the people are given a sugar pill with no active ingredient. To examine the impact of dosage on reductions in cholesterol levels, the authors of the study regress change in cholesterol levels (Y) on dosage level (X). For people in the control group, x=0 and for people in the treatment group, x measures milligrams of active ingredient and the model they estimate is of the form $Y_i=\beta_0 +X_i\beta_1+\varepsilon_i$. In this case, the authors find a statistically significant and negative coefficient on the estimate for β_1 – larger dosages reduce cholesterol levels. Is this an unbiased estimate of the impact of dosage on change in cholesterol level? Why or why not?

6. The major of a particular city is trying to encourage the hotels in the area to reduce their rates as a way to encourage tourism – if they reduce rates, more people will come and spend money in restaurants, shops, local attractions, etc. The local chamber of commerce asks how much of a demand boost would they get if hotels dropped prices. The city hires DA Consultants, LLC, a local company, to estimate the demand for hotel rooms. As it turns out, you are friends with the consultant and you know they can talk a good game but got a C in econometrics. They collect data on 100 cities and regress the number of hotel rooms rented per year (Y_i) on the average price of a hotel room (X_i) and estimate the equation $Y_i=\beta_0 +X_i\beta_1+\varepsilon_i$. To their surprise, the coefficient on $\hat{\beta}_1$ is actually positive. Your friend the consultant, knowing you are better in econometrics than they are, comes and ask "Shouldn't demand curves slope down – why am I getting a positive coefficient on my demand equation?" Please provide some free advice to DA Consultants and explain why their model might be getting $\hat{\beta}_1>0$. [This is a true story].

7. Consider a regression that attempts to generate a demand elasticity from an older data set. The model is of the form $\ln(Q_i)=\beta_0 +\ln(P_i)\beta_1+\varepsilon_i$ where P is measured in 1990 dollars. The researcher obtains an estimate for $\hat{\beta}_1$ of -0.50. Someone comments that they should use constant 2017 dollars in the regression to reflect the current time, so the researcher replaces P_i with $P_i^* = P_i C$ where C is the price index adjustment that is multiplied by all observations. Show that when the author switches to using P_i^* instead of P_i in the regression, the estimate for $\hat{\beta}_1$ will not change (HINT: use the properties of natural logs – you also have a data set you can use to verify your results).

8. (Pretty tricky) *A researcher estimates a bivariate regression of the form $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$ but confides to a colleague that she believes $\text{cov}(\varepsilon_i,x_i)\neq 0$ and therefore, $\hat{\beta}_1$ is a biased estimate. The colleague then asks whether one can test whether $\text{cov}(\varepsilon_i,x_i)\neq 0$. The colleague suggests that the researcher construct $\hat{\varepsilon}_i = y_i - \hat{\beta}_0 - x_i\hat{\beta}_1$ then run a regression of $\hat{\varepsilon}_i$ on x_i, that is, a regression of the form $\hat{\varepsilon}_i = \gamma_0 + x_i\gamma_1 + v_i$, then test the null H_o: $\gamma_1 = 0$ to see whether ε_i, and x_i are correlated. Is this a good idea or not?

HINT: The OLS estimate of $\hat{\gamma}_1$ would be $\hat{\gamma}_1 = \dfrac{\sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ --- and don't over think the problem.

9. (Pretty hard) An author wants to estimate a model of the form $y_i=\beta_0 + x_i\beta_1+\varepsilon_i$. Unfortunately, the author is unable to find exactly the variable for y but instead, they find a variable that has some 'measurement error', that is, the variable they use in the regression is y_i^* and $y_i^* = y_i + v_i$ where v_i is a random error with $E[v_i]=0$, $V(v_i)=\sigma^2_v$ and $Cov(v_i,\varepsilon_i)= Cov(v_i,x_i)=0$. Think of the problem this way. A survey asks people for their usual weekly earnings and instead of responding with their exact earnings, (y_i) they give an approximation y_i^* that varies randomly where the error in their response is given by the random variable (v_i). Suppose the researcher estimates the model with y_i^* instead of y, gets and estimate for $\hat{\beta}_1$ that equals the following

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

In this example, will the use of y_i^* instead of the true y_i generate biased estimates? HINT: Write the numerator as $\sum_{i=1}^{n} y_i^*(x_i - \bar{x})$ and recall that $y_i^* = y_i + v_i$. Substitute in the true value for y_i^* into the model and then take expectations.

10. (Pretty hard) Consider a bivariate regression model of the form $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$. Show that the square of the correlation coefficient between y_i *and* $\hat{y}_i$ is equal to the R^2.

HINT: Start with the definition of the squared correlation coefficient between y_i *and* $\hat{y}_i$ – then re-write this to read $R^2=SSM/SST$. One "trick" you can use is something we showed when generating the R^2 – that $\sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{y}_i - \bar{\hat{y}}) = 0.$