Problem Set 4, ECON 30331
(Due at the start of class, Wednesday, February 14, 2014)
(Questions marked with a * are old test questions)

Bill Evans
Spring 2018

1. Consider a multivariate regression model of the form $y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i$. Write the 1st order conditions for the optimization problem where one is interested in minimizing the sum of squared errors SSE = $\sum_{i=1}^{n} \hat{\varepsilon}_i^2$.

   Suppose in a sample if 25 observations, the following facts are presented about the model above.

   $$\sum_{i=1}^{n} x_{1i}^2 = 40 \quad \sum_{i=1}^{n} x_{2i}^2 = 80 \quad \sum_{i=1}^{n} x_{1i}x_{2i} = 0 \quad \sum_{i=1}^{n} x_{1i} = 0 \quad \sum_{i=1}^{n} x_{2i} = 0 \quad \sum_{i=1}^{n} y_i = 0$$
   $$\sum_{i=1}^{n} x_{1i}y_i = 120 \quad \sum_{i=1}^{n} x_{2i}y_i = 160$$

   Using the first order conditions (or normal equations) and these facts, provide the estimates for $\hat{\beta}_0, \hat{\beta}_1 \text{ and } \hat{\beta}_2$.
   HINT: Solve for $\hat{\beta}_0$ first.

2. Download the data cps87.dta. Generate two new variables. The first is the natural log of weekly earnings. The second is age squared. Next, run a regression of the natural log of weekly earnings on age, age squared and years of education. We can write this model as

   $$\ln(weekly\ earn) = \beta_0 + age_i\beta_1 + age_i^2\beta_2 + educ\beta_3 + \varepsilon_i$$

   Provide a mathematical expression that defined $\dfrac{\partial \ln(weekly\ earn)}{\partial age}$. Using the results from the regression, what is $\dfrac{\partial \ln(weekly\ earn)}{\partial age}$ at age 21? Age 35? Age 50?

3. Consider a multivariate regression model of the form $y_i = \beta_0 + x_{1i}\beta_1 + \varepsilon_i$. Suppose the R^2 from this model is R_a. True, False, or Uncertain and explain. The R^2 can never fall below R_a when additional variables are added to the model? (Think of a special case where someone adds completely irrelevant variables to the model – what will happen to the R^2?)

4. On the class web page is a STATA data set called house_price.dta. It has data on 114 homes sold in 1998 in a small town in New England. The data set contains information on the sales price of the house (measured on thousands of dollars), the number of bedrooms, bathrooms, other rooms, square feet of living space and age of the home,

   Download the data and initially estimate a regression with house prices as the outcome of interest and four covariates: age in years, # number of bedrooms, # of bath rooms, # of other rooms. Call this model 1.

   a. Interpret the coefficient on age in years and # of bedrooms by providing a numeric example.

   Now, estimate a second model and add to the original regression the square feet of living space. Call this model 2.

b. What happens to the coefficient on # of rooms, # of bedrooms and # of other rooms in this new model compared to the previous one? Why have the coefficients on these three variables changed so dramatically?

c. Interpret the coefficient on square feet of living space.

Now estimate a third model with the same dependent variable but include only two covariates: age in years and square feet.

d. Compare the R^2 from this model and that in Model #2. Provide an intuitive explanation for why the difference is so small.

5. On the class web page is a data set named senior_medical_exp.dta which has information on age, the number of chronic conditions and the total medical expenses for a sample of senior citizens aged 65 to 84. I select seniors for this example because all of them have health insurance through the Medicare program. The three variables in the data set are

Variable	Label
totalexp	total expenditures on medical care, 2002
chronic	number of chronic conditions (0-5)
age	age in years

Load the data set into STATA, then construct two new variables:

Regress totalexp on age and chronic. (`reg totalexp age chronic`)

a) Interpret the coefficient on age – provide a numeric example of the magnitude of the coefficient on this variable?

b) Interpret the coefficient on chronic -- provide a numeric example of the magnitude of the coefficient on this variable?

c) Now regress totalexp on age (`reg totalexp age`). What has happened to the coefficient on age compared to the results in part a? Does this make sense? Why or why not.

d) Regress age on chronic. What is the coefficient on chronic? Does this make sense?

e) After this regression, output the residuals from the regression

```
predict res_age, residual
```

Next, regress totalexp on res_age. How does this number compare to the estimates in a)?

6. *Return to problem 5 on problem set 3. A pharmaceutical company is investigating the cholesterol lowering benefits of a new drug. In a sample of n subjects the company randomly assigns milligrams of active ingredients (label this as x_{1i}) and the outcome of interest, labeled as y_i, is the change in cholesterol from the start until the end of the trial. Initially, the researchers estimate a model of the form $y_i = \beta_0 + x_{1i}\beta_1 + \varepsilon_i$. However, a colleague mentions that as part of the experiment, they also collected detailed data on characteristics of survey participants that predict y_i like their weight at the start of the trial, age, sex, ethnicity/race, plus other variables. The colleague asks whether on should include these covariates (label them as x_{2i}, x_{31}, …x_{ki}) into the basic regression?

a) By estimating a model of $y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 ...... x_{ki}\beta_k + \varepsilon_i$, do you anticipate that the estimate on $\hat{\beta}_1$ will change?

b) In a multivariate model, the estimated variance of $\hat{\beta}_1$ is given as $\hat{V}(\hat{\beta}_1) = \dfrac{\hat{\sigma}_\varepsilon^2}{(1-R_1^2)\sum_{i=1}^{n}(x_{1i} - \overline{x}_1)^2}$

What is the likely consequence of adding these additional covariates (x_{2i}, x_{31}, …x_{ki}) to the estimated variance of $\hat{\beta}_1$? Explain your answer.

7. *On the next page are the results from two regression models: In model (1), I regress Y on X_1, and note that the standard error on the coefficient on X_1 is very small and the t-statistic on the coefficient on $\hat{\beta}_1$ is over 23. Note that in model (2), when I add X_2 to the model, the standard error on $\hat{\beta}_1$ increases by a factor of 3 and the t-statistic on this parameter falls to 1.39. Using the information given, provide an intuitive explanation for why the standard increases so much on $\hat{\beta}_1$ when X_2 is added to the model. To get full credit, you must provide the proper equation.

8. *Many people get their health insurance through their job and because of high health insurance costs, many employers are considering offering free on-site exercise classes as a way of encouraging healthy behaviors and hopefully reducing medical care costs. The evidence for subsidized exercise classes comes primarily from research in the field of public health. In these models the authors collect data from an employer and estimate a regression of the form $y_i = \beta_0 + x_{1i}\beta_1 + \varepsilon_i$ where y_i is annual spending on health care for employee i and x_{1i} is a dummy variable that equals 1 if the person uses the on-site health care services. Call this model (1). Let $\tilde{\beta}_1$ be the estimate for β_1 from model (1) and in this case, the author gets the expected result that $\tilde{\beta}_1 < 0$ – people that use on-site exercise classes have lower health care spending. Model (1) has been criticized because it does not control for the fact that the least healthy employees are the ones the least likely to enroll in these classes. Consider a simple extension to the model where the author has detailed data on the health of employees *prior to the exercise classes opening.* Let x_{2i} be a simple index that equals the number of chronic health conditions a person has (e.g., a person with high blood pressure, obesity, and diabetes has a count of three whereas a healthy person has a count of zero). Now consider estimating model (2) which is of the form $y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i$. If Model (2) is the true model, do you anticipate that, the estimate $\tilde{\beta}_1$ from model (1), is biased up or down? Explain your answer and to get full credit, you must provide an appropriate equation.

Results for Question 7

Correlation between X_1 and X_2

```
. corr x1 x2
(obs=2489)

             |       x1       x2
-------------+------------------
          x1 |   1.0000
          x2 |   0.9994   1.0000
```

Model 1: Regression of Y on X_1

```
. reg y x1

      Source |       SS       df       MS              Number of obs =    2489
-------------+------------------------------           F(  1,  2487) =  562.63
       Model |  121.044173     1  121.044173           Prob > F      =  0.0000
    Residual |  535.054756  2487  .215140634           R-squared     =  0.1845
-------------+------------------------------           Adj R-squared =  0.1842
       Total |  656.098929  2488  .263705357           Root MSE      =  .46383

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .0765488   .0032272    23.72   0.000     .0702205    .0828771
       _cons |   5.059357   .0435541   116.16   0.000     4.973951    5.144763
------------------------------------------------------------------------------
```

Model 2: Regression of Y on X_1 and X_2

```
. reg y x1 x2

      Source |       SS       df       MS              Number of obs =    2489
-------------+------------------------------           F(  2,  2486) =  281.41
       Model |  121.115811     2  60.5579054           Prob > F      =  0.0000
    Residual |  534.983118  2486  .215198358           R-squared     =  0.1846
-------------+------------------------------           Adj R-squared =  0.1839
       Total |  656.098929  2488  .263705357           Root MSE      =  .46389

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .1304861   .0935397     1.39   0.163    -.0529375    .3139098
          x2 |  -.0539557   .0935159    -0.58   0.564    -.2373328    .1294213
       _cons |   5.059738   .0435649   116.14   0.000     4.974311    5.145166
------------------------------------------------------------------------------
```

9. *Research has shown that students attending higher quality colleges and universities tend to have higher wages after graduation than those attending less selective institutions. Using a nationally representative sample of college graduates aged 30-39, researchers regress the natural log of annual earnings (y_i) on the **average SAT score from the college the respondent attended** (x_{1i}) using the simple bivariate regression model $y_i = \beta_0 + x_{1i}\beta_1 + \varepsilon_i$. Call this model (1). Let $\tilde{\beta}_1$ be the estimate for β_1 from model (1) and in this case, the author gets the expected result that $\tilde{\beta}_1$>0 – students that graduated from higher quality schools tends to have higher earnings. Someone criticizes model (1) because it does not control for differences in other characteristics of the students that are likely to be correlated with earnings. For example, the author does not have a measure of academic ability for the student like an SAT score which they argue should be included in the model. Suppose the author considers estimating model (2) which is of the form $y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \varepsilon_i$ where x_{2i} is the students **own SAT score**. If Model (2) is the true model, do you anticipate that $\tilde{\beta}_1$, the estimate from model (1), is biased up or down? Explain your answer and to get full credit, you must provide an appropriate equation.

10. A researcher regresses y on x1 and produces the results below. A colleague argues that the model should also include the covariates x2, x2, and x4, which the colleague argues are strong predictors of y. Below is a matrix that provides the correlation coefficients for the variables x1, x2, x3 and x4. Given these results, do you expect that adding x2, x3 and x4 to the model will change the results much? Assume your colleague is correct that x2, x3 and x4 are strong predictors of y.

Results for Problem 10

```
. reg y x1

      Source |       SS       df       MS              Number of obs =    3981
-------------+------------------------------           F(  1,  3979) =  795.19
       Model |  174.739778     1  174.739778           Prob > F      =  0.0000
    Residual |   874.36848  3979  .219745785           R-squared     =  0.1666
-------------+------------------------------           Adj R-squared =  0.1664
       Total |  1049.10826  3980   .26359504           Root MSE      =  .46877

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .0739822   .0026236    28.20   0.000     .0688386    .0791259
       _cons |   5.105143   .0353055   144.60   0.000     5.035925    5.174362
------------------------------------------------------------------------------

. corr x1 x2 x3 x4
(obs=3981)

             |       x1       x2       x3       x4
-------------+------------------------------------
          x1 |   1.0000
          x2 |   0.0182   1.0000
          x3 |   0.0002   0.0061   1.0000
          x4 |   0.0025   0.0075  -0.0211   1.0000
```