Problem Set 2
ECON 30331
(Due by the start of class, Wednesday, January 31, 2018)
(Problems marked with a * are former test questions)

Bill Evans
Spring 2018

1. Suppose a researcher is interested in estimating the linear regression model $Y_i = \beta_0 + X_i\beta_1 + \varepsilon_i$ and in a sample of 101 observations, the following descriptive statistics are generated:

$$\bar{x} = 60,\ \bar{y} = 40,\ \sum_{i=1}^{n}(x_i - \bar{x})^2 = 2500,\ \sum_{i=1}^{n}(y_i - \bar{y})^2 = 3600$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y - \bar{y}) = 1500,\ R^2=0.2$$

a) What is $\hat{\sigma}_x^2$?
b) What is $\hat{\rho}(x,y)$?
c) What is the OLS estimate of $\hat{\beta}_1$?
d) What is the OLS estimates of $\hat{\beta}_0$?
e) What is the SSE for this model?

2. *Using data from the 2004 baseball season, a researcher collects data on the number of wins a team had during the year and payroll in millions of dollars. The researcher wants to estimate a model to examine whether the size of the payroll alters wins, so they want to consider an OLS model of the form $wins_i = \beta_0 + payroll_i\ \beta_1 + \varepsilon_i$. The author gets as far as getting descriptive statistics and the correlation coefficient between wins and payroll (presented below), then their computer crashes. Using the data below:

a) Calculate the estimate for $\hat{\beta}_0$.
b) For $\hat{\beta}_1$.
c) Interpret the results for $\hat{\beta}_1$. According to the model estimates, by how much will wins increase if a team spends $15 million more on salary?

```
. sum wins payroll
```

Variable	Obs	Mean	Std. Dev.	Min	Max
wins	30	80.96667	13.36615	43	101
payroll	30	70.13708	27.26755	19.63	149.711

```
. corr wins payroll
(obs=30)
```

	wins	payroll
wins	1.0000	
payroll	0.4176	1.0000

3. *Below are STATA results from a CORRelation, SUMmary, and REGression statement, but some of the results have been whited-out. Please provide estimates for A, B, C, D, and E

```
. sum x y

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           x |        48          .5    .5052912          0          1
           y |        48           D    .1177725   7.674153   8.037867

. corr x y
(obs=48)

             |        x        y
-------------+------------------
           x |   1.0000
           y |   0.9285   1.0000


. reg y x

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  1,    46) =  287.76
       Model |        A        1  .562059213           Prob > F      =  0.0000
    Residual |  .08984759     46  .001953208           R-squared     =       B
-------------+------------------------------           Adj R-squared =  0.8592
       Total |  .651906803    47  .013870358           Root MSE      =       E

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |         C     .012758    16.96   0.000     .1907409    .2421021
       _cons |   7.771764   .0090213   861.49   0.000     7.753605    7.789923
------------------------------------------------------------------------------
```

4. Given an OLS model of the form $Y_i=\beta_0 + X_i\beta_1 + \varepsilon_i$, suppose that $\hat{\beta}_1 = 0$.

a) What is the estimate for $\hat{\beta}_0$?
b) What does R^2 equal in this case?

5. Given an OLS model of the form $Y_i=\beta_0 + X_i\beta_1 + \varepsilon_i$, show that $\bar{Y} = \bar{\hat{Y}}$.

6. Download the STATA data set pop_1950_2000.dta which has the US population (measured in millions of people) from 1950 through 2000. Construct a variable called timetrend which equals year-1949

```
gen timetrend=year-1949
```

Next, run a regression with population as the outcome of interest and the timetrend as the covariate.

```
reg population timetrend
```

a) What is the coefficient on the timetrend variable? Interpret what the coefficient means.
b) What is the R^2 from this model?
c) What does this value for the R^2 say about changing population values in the US?
d) Google the US population total for 2017. What is it?
e) What does the model predict population in the US will be in 2017?

7. On the assignments page for the class is a STATA data set called meps_senior.dta that has information on total medical care expenditures for a sample of elderly respondents (aged 65 and over) from the Medical Expenditure Panel Survey (MEPS). There are many variables in the data set but for now, run a regression using totalexp (total annual medical expenditures) as the dependent variable, and as the independent covariate, use age (age in years), so the regression is $totalexp_i=\beta_0 + age_i\beta_1 + \varepsilon_i$.

```
reg totalexp age
```

A) What are the estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$?
B) Interpret the coefficient for $\hat{\beta}_1$, what is $\dfrac{\partial total\exp}{\partial age}$?
C) Using the estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$ what is predicted spending for someone 70 years of age?
D) Using the estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$ what is predicted spending for someone 71 years of age?
E) Generate the difference between the answers in part D) and C) – does this make sense?

8. (Challenging problem). Continue with problem 7.
a. What is the R^2 in the previous problem?
b. Run a regression of the form $age_i=\gamma_0 + totalexp_i\gamma_1 + v_i$. What is the R^2 from this regression?
c. Show that in general, the R^2 from a regression of Y on X is the same as a regression of the R^2 from a regression of X on Y.

[Hint: $R^2 = \dfrac{\sum_i (\hat{y}_i - \bar{\hat{y}})^2}{\sum_i (y_i - \bar{y})^2}$, substitute the definition of $\hat{y}_i - \bar{\hat{y}}$. In the second model $age_i=\gamma_0 + totalexp_i\gamma_1 + v_i$ and which we can write as $x=\gamma_0 + y_i\gamma_1 + v_i$ we know that $\hat{\gamma}_1 = \dfrac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (y_i - \bar{y})^2}$ and the R^2 in this case equals

$$R^2 = \frac{\sum_i (\hat{x}_i - \bar{\hat{x}})^2}{\sum_i (x_i - \bar{x})^2}$$

9. (Challenging problem). Suppose a researcher is interested in estimating the impact of gasoline taxes (X_i) on per capital gallons of gasoline consumed per year (Y_i) Assume tax is measured in cents per gallon. The researcher has data from 51 states for a 10 year period for a total of 510 observations. The researcher estimates the linear OLS model $Y_i=\beta_0 +X_i\beta_1+\varepsilon_i$ and calculates $\hat{\beta}_1 = -0.90$.

a) Suppose instead of measuring taxes in cents per gallon, the researcher measures taxes in dollars per gallon where the new model is $Y_i= \gamma_0 +X_i^*\gamma_1+ \varepsilon_i$ and $X^*=X/100$. What will be the estimate on the coefficient on γ_1?

b) Suppose taxes are measured in cents as in the first case, but consumption is measured as gallons consumed per month, where $Y^*_i=Y_i/12$. The model now is of the form $Y_i^* =\alpha_0 + X_i\alpha_1 + \varepsilon_i$. What will be the estimate on α_1?

[You have a data set with Q and taxes but for a different product. Check your answer with the data set we have used in class.]