Problem Set 5, ECON 30331
(Due at the start of class, Wednesday, February 28, 2018)
(Problems marked with a * are former test questions)

Bill Evans
Spring 2018

1. *On the next page are STATA results for two OLS models constructed from a sample of 30 observations:

Model 1: $y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4 + x_{5i}\beta_5 + \varepsilon_i$
Model 2: $y_i = \beta_0 + x_{1i}\beta_1 + x_{5i}\beta_5 + \varepsilon_i$

On the printout, I have "whited-out" some of the results. Please use the results on the next page to answer the following questions. Please show all work.

A) What is the R^2 for model 1 in this case?
B) What is the estimate for $\hat{\sigma}_\varepsilon^2$ from Model 1?
C) Using the results from Model 1 construct a **99% confidence interval** for the coefficient on $\mathbf{x_1}$. What are the appropriate degrees of freedom and the critical value of the t-distribution used in this case? Using this confidence interval, can you reject or not reject the null hypothesis that the true coefficient on $\mathbf{x_1}$ is zero, H_o: $\beta_1=0$?
D) Using the results from Model 1 and a **90%** confidence level, use the p-value to test the null hypothesis that the coefficient on $\mathbf{x_2}$ is zero, H_0: $\beta_2=0$. Can you reject or not reject the null?
E) Using the results from Model 1 and a **95%** confidence level, use **a t-test** to test the null hypothesis that the coefficient on $\mathbf{x_5}$ is zero, H_0: $\beta_5=0$. What are the appropriate degrees of freedom and the critical value of the t-distribution used in this case? Can you reject or not reject the null?
F) Using the results from models (1) and (2) and a 95% confidence level, test the null hypothesis that H_o: $\beta_2 = \beta_3 = \beta_4 = 0$. What is the estimate of the F test statistic $(\hat{F})$? Specify the degrees of freedom used in the test and the critical value of the F-distribution used in this test? Can you reject or not reject the null?
G) Using the reported results from models (1) and a 95% confidence level, test the null hypothesis that all coefficients on the x's are equal to zero, H_o: $\beta_1=\beta_2=\beta_3=\beta_4=\beta_5=0$. Explain your answer in detail.

2. *To test a set of q restrictions in a linear regression model, we use the F-statistic which is constructed as

$$\hat{F} = \frac{(SSE_r - SSE_u)/q}{SSE_u/(n-k-1)}$$

Show that the test statistic can be calculated as

$$\hat{F} = \frac{(R_u^2 - R_r^2)/q}{(1-R_u^2)/(n-k-1)}$$

Where R_u^2 *and* R_r^2 and the R^2's from the restricted and unrestricted models, respectively.

Results for Question 1
Model 1

```
. reg y x1 x2 x3 x4 x5

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  5,    24) =    4.02
       Model |  3.05903525     5  .611807051           Prob > F      =  0.0086
    Residual |  3.64992413    24                       R-squared     =
-------------+------------------------------           Adj R-squared =
       Total |                                         Root MSE      =  .38997

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .0928424   .0335796
          x2 |    .330204   .1766384     1.87   0.074    -.0343597    .6947677
          x3 |   .0118367   .0062711
          x4 |   .2273021    .308032
          x5 |   .3627782   .1643337
       _cons |   3.893954   .5648577
------------------------------------------------------------------------------
```

Model 2

```
. reg y x1 x5

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  2,    27) =    5.17
       Model |  1.85889311     2  .929446553           Prob > F      =  0.0125
    Residual |  4.85006628    27  .179632084           R-squared     =
-------------+------------------------------           Adj R-squared =
       Total |                                         Root MSE      =  .42383

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   .0978555   .0344212
          x5 |   .4063172   .1687102
       _cons |   4.566091   .5165097
------------------------------------------------------------------------------
```

3. *Listed below are results from STATA where using 24 observations, y is regressed on x_1 x_2 x_3 and a constant. I have "whited out" some of the results. Using the results in panel a, answer the following questions:

   A) Construct a 95% confidence interval for the parameter on x_1? Using this confidence interval, can you reject or not reject the null hypothesis that H_o: β_1=0?

   B) Using a t-test and a 95% confidence interval, test the null hypothesis H_o: β_1=0? What is the appropriate value of the t-statistic in this case?

   C) How do your results in part b) change if you change the confidence level to 99%?

In panel b) of the results, I report the estimates of a model where y is regressed on x_1 and constant.

D) Using the results from panels a) and b), use and F-test and a 95% confidence level to test the null hypothesis that H_o: $\beta_2 = \beta_3 = 0$. What are the degrees of freedom of the critical value of the F in this context and can you reject or not reject the null?

Panel A

```
. reg y x1 x2 x3

      Source |       SS       df       MS              Number of obs =      24
-------------+------------------------------           F(  3,    20) =    2.89
       Model |  407067.668     3  135689.223           Prob > F      =  0.0608
    Residual |  938379.666    20  46918.9833           R-squared     =  0.3026
-------------+------------------------------           Adj R-squared =  0.1979
       Total |  1345447.33    23  58497.7101           Root MSE      =  216.61

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   34.78137   13.24421
          x2 |  -8.757655   30.83438
          x3 |    .161071   .3664861
       _cons |   83.06376   627.1563
------------------------------------------------------------------------------
```

Panel B

```
. reg y x1

      Source |       SS       df       MS              Number of obs =      24
-------------+------------------------------           F(  1,    22) =    6.80
       Model |  317743.343     1  317743.343           Prob > F      =  0.0161
    Residual |  1027703.99    22  46713.8177           R-squared     =  0.2362
-------------+------------------------------           Adj R-squared =  0.2014
       Total |  1345447.33    23  58497.7101           Root MSE      =  216.13

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |    34.4568   13.21172
       _cons |   24.97369    182.131
------------------------------------------------------------------------------
```

4. Suppose you have a regression of the form $y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + x_{4i}\beta_4 + \varepsilon_i$

a) What would the restricted model look like if one were to test the null hypothesis, H_o: $\beta_1 = (1/2)\beta_2 = 3\beta_3$

b) What would the restricted model look like if one were to test the null hypothesis, H_o: $\beta_4 = 1 - 4\beta_1 - \beta_2 - 2\beta_3$

5. On the class web page is a data set named meps_2005.dta. The data set contains 3167 observations on total annual medical expenditures for US adults aged 65 and older. The data set has 9 variables and detailed definitions for these variables are listed below.

Variable	Definition
Totalexp	Annual total expenditures on medical care
Income	Annual family income
Age	Age in years
Educ	Years of education
Male	Dummy variable, =1 if male, =0 otherwise
Bmi	Body mass index (weight in kg/height in cm^2
Srhealth	Self reported health, =1 if excellent, 2=very good, 3=good, 4=fair, and 5=poor
Region	Region of the country, 1= northeast, 2=Midwest, 3=south, 4=west
Race	Categorical variable, 1=white, non-Hispanic, 2=black, non-Hispanic, 3=other race, 4=Hispanic

Generate the following 12 variables:
3 dummy variables for white, black and other race, respectively
4 dummy variables for very good, good, fair and poor health, respectively
The natural log of income (ln_income)
The natural log of total medical expenditures, ln_totalexp
3 dummy variables for Midwest, south and west of the country, respectively.

Run a regression with the dependent variable being ln_totalexp and include 15 covariates plus the constant: age, educ, ln_income, bmi, male, 4 self reported health dummies, 3 race dummies, and 3 region dummies. From this regression, answer the following questions

a) What is the SSE and the R^2 for this model?
b) Provide interpretations (a one unit change in x will produce….) for the following coefficients: male, bmi and ln_income?
c) Using a t-statistic and a 95% confidence level, can you reject or not reject the null that $\beta_{ln\_income}=0$? What is the critical value for the t-test in this case?
d) Using a 95% confidence level, test the null hypothesis that the regional effects are all zero, $H_o:\beta_{region2}=\beta_{region3}=\beta_{region4}=0$. What is the critical value of the F-distribution in this case? Can you reject or not reject the null.
e) Interpret the coefficients on poor and fair health, respectively.

6. *Listed below are regression results explaining the retail price for a sample of 177 motor vehicles sold in the US in 2002. The dependent variable is the **manufactures suggested retail price (msrp).** In the regression, there are 7 covariates plus the constant. The first four covariates are defined as: **horse** (the horse power of the car) **ln_mpg** (the natural log of miles per gallon), **awd** (a dummy variable that equals 1 if the vehicle is "all wheel drive" and 0 otherwise). Among these cars, there are four body types: sedans, minivans, SUVs and trucks. In the model, I've include dummy variables for the first three types Provide a verbal description of how one would interpret the following coefficients in the model:

a) The coefficient on "horse"
b) The coefficient on "ln_mpg"
c) The coefficient on "awd"
d) The coefficient on "sedan"?
e) The coefficient on "suv"?

Results for Question 6

```
. reg msrp horse ln_mpg awd minivan suv sedan

      Source |       SS       df       MS              Number of obs =     177
-------------+------------------------------           F(  6,   170) =   89.22
       Model |  1.2390e+10     6  2.0651e+09           Prob > F      =  0.0000
    Residual |  3.9347e+09   170  23145195.2           R-squared     =  0.7590
-------------+------------------------------           Adj R-squared =  0.7505
       Total |  1.6325e+10   176  92755889.4           Root MSE      =  4810.9

------------------------------------------------------------------------------
        msrp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       horse |   125.8846   8.413873    14.96   0.000     109.2754    142.4937
      ln_mpg |   6364.463   3226.799     1.97   0.050    -5.292338    12734.22
         awd |   469.2566   1581.476     0.30   0.767    -2652.604    3591.117
     minivan |  -2191.638    1723.98    -1.27   0.205    -5594.802    1211.527
         suv |   674.2011   1388.999     0.49   0.628    -2067.706    3416.108
       sedan |  -1053.886   852.2187    -1.24   0.218     -2736.18    628.4082
       _cons |  -16982.28   7993.087    -2.12   0.035    -32760.77   -1203.793
------------------------------------------------------------------------------
```

7. Many states run lotteries. When first proposed, lotteries always face vocal opposition. In some cases, in order for states to get the lottery passed the legislature, they must "earmark" lottery profits for a good cause. The most popular destination for lottery profits is K-12 education. Simple economic models suggest that because money is fungible, earmarking should not change spending more than a change in income. The argument goes as follows: If I were to give you $100 more in income – you would spend a fraction of that on food. If your mom thinks you are looking a little thin and gives you $100 to spend on food, you would treat that $100 as a change in income and spend the same amount on food as you would if you got $100 unrestricted. In this example, we will test whether earmarking lottery profits for K-12 education increases spending dollar for dollar.

   Download the data set lottery_example.dta. This includes data from 31 states that run lotteries over the 1977-1998 period so there are 22 years*31 states=682 observation. The data set has the following variables

Variable	Definition
Fips	State fips code, 2 digit number 1-56
Stated	2 character postal code, (AL for Alabama, IN for INDIANA)
exp_pupil	K-12 expenditures per pupil in real 1995 dollars
lottery_profit_pupil	State lottery profits per pupil in real 1995 dollars
k12_share	The share of lottery profits that are earmarked to K-12 education. Goes from 0 to 1.
inc_pupil	State aggregate income per pupil in real 1995 dollars.
Time	Time trend that equals 1 in 1977, 2 in 1978, etc.

All spending variables must be denominated by the same value – in this case, we denominate by the number of K-12 pupils in a state.

Construct two variables – the first is the amount of lottery money earmarked to K-12 education – the other is the amount of lottery profits not set aside for education

```
gen K12_earmark_pupil=k12_share*lottery_profit_pupil
gen not_earmark_pupil=(1-k12_share)*lottery_profit_pupil
```

Next, run a regression of expenditures on income, where lottery profits are earmarked and the time trend

```
reg exp_pupil inc_pupil k12_earmark_pupil not_earmark_pupil time
```

a) Interpret the coefficients on inc_pupil K12_earmark_pupil not_earmark_pupil
b) If earmarking works, it should be the case that spending on education went up dollar for dollar with money earmarked for that cause. Using a t-test and a 95% confidence level (α=0.05) test the null hypothesis that the coefficient on K12_earmark_pupil is 1. H_o: $\beta_{K12\_earmark\_pupil}=1$. Can you reject or not reject the null hypothesis?
c) Redo the test in b) but use a t-test. What is the $\hat{t}$ on the null hypothesis that H_o: $\beta_{K12\_earmark\_pupil}=1$? Can you reject or not reject the null hypothesis?
d) If the economic model that money is fungible is correct, it should be the case that the marginal spending on K-12 education from an earmarked lottery dollar should equal the same amount from an increase in income. Using an F-test and a 95% confidence level (α=0.05), test the null hypothesis that the coefficients on K12_earmark_pupil and inc_pupil are the same H_o: $\beta_{K12\_earmark\_pupil}=\beta_{inc\_pupil}$. Can you reject or not reject the null hypothesis?
e) Using an F test and a 95% confidence level, test the null hypothesis that the impact of earmarked lottery money has the same impact on spending as non-earmarked lottery spending H_o: $\beta_{K12\_earmark\_pupil}=\beta_{not\_earmark\_pupil}$. Can you reject or not reject the null hypothesis?
f) How does your answer change for part d) if the confidence level is set at 90% (α=0.1)?

8. A typical production function used in empirical work assumes that industry output (q) is produced by four inputs capital (k), labor (l), energy (e) and materials (m). In the data set klem_chem.dta, I have data from 1960 through 2005 on quantities of each input in the chemical industry for the US and industry output. The Cobb-Douglas production function in this case can be considered $q=\alpha k^{\beta_k} l^{\beta_l} e^{\beta_e} m^{\beta_m}$ and the parameters of the model can be estimated by the equation

$$\ln(q_i)=\beta_0+\ln(k_i)\beta_k+\ln(l_i)\beta_l+\ln(e_i)\beta_e+\ln(m_i)\beta_m+\varepsilon_i$$

a) Take the logs of all the relevant variables and estimate the Cobb0-Douglas production function.
b) What is the coefficient on $\ln(m_i)$ and provide an interpretation of that parameter.
c) Test the null hypothesis that $H_o:\beta_k=\beta_l=\beta_e=0$. Can you reject or not reject the null?
d) Test the null hypothesis that the production function exhibits constant returns to scale, that is
e) $H_o:\beta_k+\beta_l+\beta_e+\beta_m=1$. Can you reject or not reject the null?

9. A researcher is interested in examining whether a new drug can lower cholesterol levels in patients with high cholesterol. The author recruits 100 people into a clinical trial and randomly assigns people to treatment (the active ingredient) and control (a placebo). The dependent variable y_i is the change in cholesterol levels over the 6 month trial and the key covariate is x_i=1 is the patient is assigned to treatment and =0 of they are assigned to control. The researcher estimates a bivariate regression model of the form $y_i=\beta_0+x_i\beta_1+\varepsilon_i$. The coefficient on $\hat{\beta}_1$= -10 suggesting the drug worked but the standard error on that estimate is only 7.5

meaning $\hat{t} = -1.33$ meaning that the author cannot reject the null hypothesis $H_0 : \beta_1 = 0$ at the 95% confidence level. The research thinks this may be a Type II error – the drug works but the power of the test is low. Suppose the researcher is correct, that $\hat{\beta}_1$= -10 and the drug does work. Assuming the coefficient stays at -10 as the sample size expands, what sample size would the author need to produce an estimated t-statistic of 2? HINT: We know that $\hat{V}(\hat{\beta}_1) = \dfrac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Note that $\hat{\sigma}_x^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ which means that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = (n-1)\hat{\sigma}_x^2$.

10. You are walking with your friends one some train tracks. You hear a train whistle behind you. You can make one of two decisions – keep walking on the tracks or get off the tracks. What are the Type I and Type II errors associated with your decision? (HINT: you first have to decide: what is the null hypothesis?)

11. (Hard --- follow the hints and figure out what the denominator equals first.) Consider a regression of y_i on a dummy variable (x_i). The regression is of the form $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$ and we know that OLS estimate for β_1 is

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

Show that because x_i is a dummy variable that the OLS estimate for β_1 equal to $\hat{\beta}_1 = \bar{y}_1 - \bar{y}_0$ where $\bar{y}_1$ is the means of the y's for x=1 and $\bar{y}_0$ is the mean for the y's where x=0. All terms were defined in class. [Hint: Here is help with the denominator. Let n be the number of observations. Let n_1 be the number of observations where x_i=1 so $n_1 = \sum_{i=1}^{n} x_i$. The variable n_0 is the number of observations where x_i=0 and since $n = n_1 + n_0$ then $n_0 = n - n_1$. Note that $\bar{x} = n_1 / n.$. Note also that $\sum_{i=1}^{n}(x_i - \bar{x})^2$ can be written as $\sum_{i=1}^{n} x_i^2 - n\bar{x}^2$. You should be able to calculate the denominator as solely a function of *n, n_1 and n_0*. Note one final thing – since x=1 or 0 then in this case only, $x_i = x_i^2$.]