Introduction to Econometrics
Problem Set 3
Due date: February 28 at the beginning of class

You may collaborate with other students. Please hand in your own work. At the top of your answer sheet, please identify the students with whom you have consulted to complete this problem set.
The data for this problem set are available on the class webpage. Click on the link to the data file and save the data on your local disk.

HELPFUL STATA COMMANDS:
In order to calculate the z-stat or t-stat that gives you some α% of the distribution
use the following commands:
To calculate and display z^* so that α% of dist. is to left of z^*
display invnormal(α/100) (e.g. for 5%: **display invnormal(.05)**)
To calculate and display t^* so that α% of dist. is to right of t^*
display invttail(df, α/100) (e.g. for 5% w/100 degrees of freedom: **display invttail(100, .05)**)

Question 1

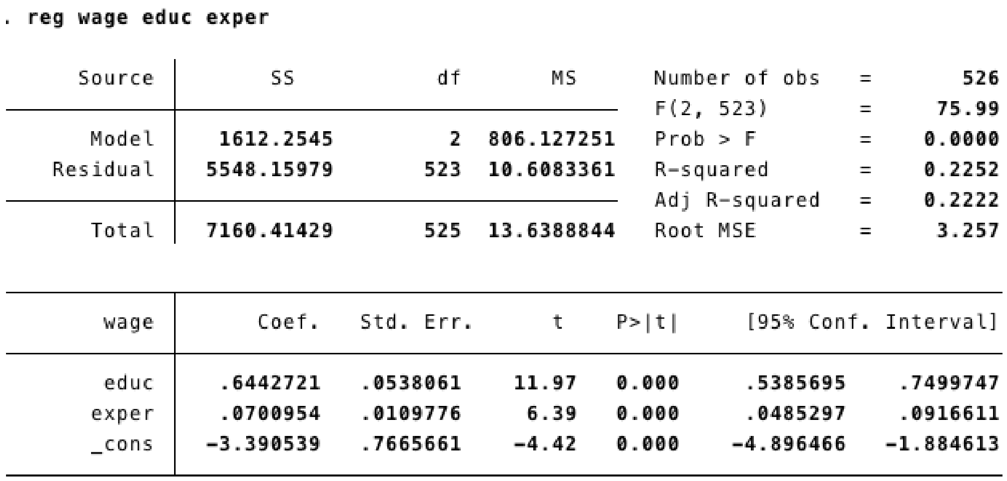
Suppose you estimate the following econometric model of wages in \$1000 (*wage*) as a function of years of education (*educ*), and years of experience (*exper*) using OLS:

$$wage_i = \alpha_0 + \alpha_1 educ_i + \alpha_2 exper_i + u_i \tag{1}$$

You also run the following econometric model of hourly wages (in dollars) wages in (*wage*) as a function of years of education (*educ*) using OLS:

$$wage_i = \beta_0 + \beta_1 educ_i + e_i \tag{2}$$

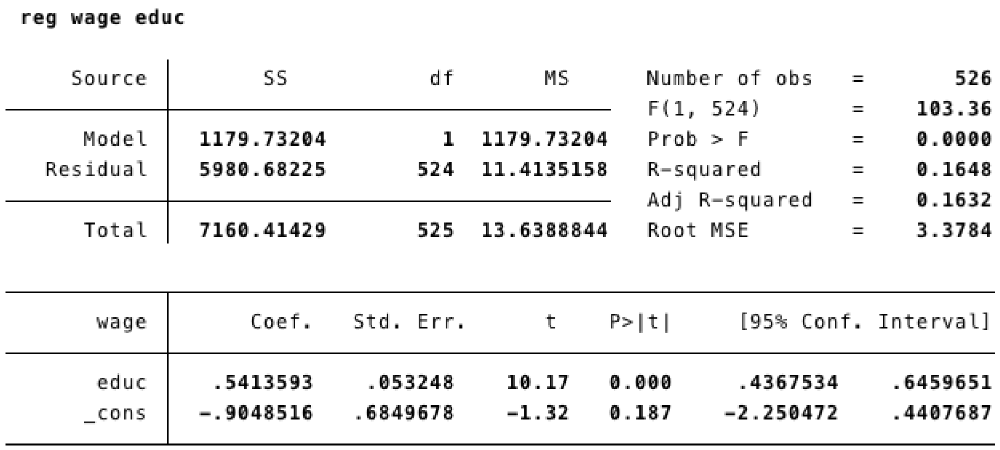
and get the following output in STATA:



```
. reg wage educ exper

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(2, 523)       =     75.99
       Model |   1612.2545         2  806.127251   Prob > F        =    0.0000
    Residual |  5548.15979       523  10.6083361   R-squared       =    0.2252
-------------+----------------------------------   Adj R-squared   =    0.2222
       Total |  7160.41429       525  13.6388844   Root MSE        =     3.257

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .6442721   .0538061    11.97   0.000     .5385695    .7499747
       exper |   .0700954   .0109776     6.39   0.000     .0485297    .0916611
       _cons |  -3.390539   .7665661    -4.42   0.000    -4.896466   -1.884613
------------------------------------------------------------------------------
```



```
reg wage educ

      Source |       SS           df       MS      Number of obs   =       526
-------------+----------------------------------   F(1, 524)       =    103.36
       Model |  1179.73204         1  1179.73204   Prob > F        =    0.0000
    Residual |  5980.68225       524  11.4135158   R-squared       =    0.1648
-------------+----------------------------------   Adj R-squared   =    0.1632
       Total |  7160.41429       525  13.6388844   Root MSE        =    3.3784

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5413593    .053248    10.17   0.000     .4367534    .6459651
       _cons |  -.9048516   .6849678    -1.32   0.187    -2.250472    .4407687
------------------------------------------------------------------------------
```

(i) Interpret the following estimates: $\hat{\alpha}_1$, $\hat{\alpha}_2$ and $\hat{\beta}_1$.

(ii) Suppose you meet two people (Anna and Dora) who are in your sample. Both have the same level of education, but Anna has 3 more years of experience. What is your best guess for the difference in wages between Anna and Dora? Explain why this is your best guess.

(iii) Suppose you meet two more people (Charles and Bob) who are also in your sample. You don't know their years of experience but you know that Bob has 5 more years of education than Charles. What is your best guess for the difference in wages between Charles and Bob? Explain why this is your best guess.

(iv) What can you say about the correlation between the variables *educ* and *exper*?

(v) Draw a rough graph (including the intercept and slope) of the estimated relationship between education and wages for the entire sample.

(vi) Draw a rough graph (including the intercept and slope) of the estimated relationship between education and wages for people that have 10 years of experience.

(vii) Suppose you were told that the true value of α_1 is known to be 0.85. Does this prove that your estimate $\hat{\alpha_1}$ is biased? Why or why not?

(viii) Suppose that you were told that the regression equation [1] does not meet all the assumptions necessary to be considered BLUE. Does this mean that the true α_2 is definitely not equal to 0.07? Why or why not?

Question 2

Let X be a random variable drawn from some distribution with an unknown mean, μ_x, and known variance, σ_x^2, which is 400. You draw a random sample of N=100 observations.

(i) What would be the (approximate) probability distribution of $\bar{X}$? Explain.

(ii) You calculate $\bar{X}$ is 15. Construct a 95% confidence interval for μ_x. Explain intuitively how to interpret your confidence interval.

(iii) Say you do not know the true variance of the X_i, but you have calculated $\sum_i X_i^2$ to be 84,375. You estimate the standard error of $\bar{X}$ using the estimator:
$\sqrt{\frac{1}{N-1}\sum_i(X_i - \bar{X})^2}$
Construct a 95% confidence interval for μ_x explaining your choice of distribution for your calculation.

(iv) Give an intuitive explanation for why your answers in (ii) and (iii) differ.

Question 3

For the following questions test the hypotheses and state your conclusion.

i. Consider the following regression model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u_1$
With a sample of 30 people (N=30), which is enough to use a t distribution, you get the following estimates, the standard errors are in parentheses:

$$\hat{Y} = 300 + \underset{(1.0)}{10} X_1 + \underset{(25)}{200} X_2 \tag{1}$$

Test the hypothesis that

$$H_0 : \beta_2 = 160$$
$$H_A : \beta_2 \neq 160$$

at the 5% level of significance.

ii. Consider the following regression model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u_1$
With a sample of 33 people (N=33) you get the following estimates, the standard errors are in parentheses:

$$\hat{Y} = 102.19 - \underset{(2053)}{9075} X_1 + \underset{(0.073)}{0.355} X_2 + \underset{(0.543)}{1.289} X_3 \tag{2}$$

Test the hypothesis that

$$H_0 : \beta_3 = 0$$
$$H_A : \beta_3 \neq 0$$

at the 1% level of significance.

iii. Consider regression (2). Test the hypothesis that

$$H_0 : \beta_2 > 0$$
$$H_A : \beta_2 \leq 0$$

at the 5% level of significance.

Question 4

The World Health Organization claims that healthier people are more productive. Nutritionists have argued that, in low income settings, the height of an adult is a good indicator of early childhood nutrition which is, in turn, an important predictor of adult health. They claim, therefore, that adult height is a good indicator of health. They argue that investments in early child health will result in faster economic growth. In support of their claim, they present evidence that, controlling age and education, height is positively associated with wages.

(i) Using data on a sample of Indonesian male workers, in ps3q4.dta, estimate the following model:

$$ln(w_i) = \beta_0 + \beta_1 ht_i + \beta_2 educ_i + \beta_3 age_i + u_i \tag{1}$$

where $ln(w_i)$ is the logarithm of the wage of male i, ht is height, $educ$ is years of completed education and age is age in years of the male.

a) Test the hypothesis that taller men earn more at a 5% level of significance.

b) Interpret the magnitude of your estimate of β_1.

c) Is it important to control age and education in the model? Provide an economic or intuitive explanation for your answer.

ii) A revisionist school of nutrition argues that height is a poor proxy for current health of adults. They argue that the biomedical evidence suggests economic success is more tightly linked to micro-nutrient deficiencies. To make their point, they re-estimate the model and also include hemoglobin (Hb) levels in the blood of the Indonesian males, a marker of iron status.

$$ln(w_i) = \alpha_0 + \alpha_1 ht_i + \alpha_2 Hb_i + \alpha_3 educ_i + \alpha_4 age_i + v_i \tag{2}$$

Test the hypothesis that men with higher levels of Hb in their blood earn more, conditional on height, age and education at a 5% level of significance.

iii) Iron has been shown to be related to work capacity with those males who are anemic ($Hb < 13g/dl$) having lower levels of work capacity and reduced endurance. Randomized trials in human and animal models have established that iron levels above this cut-off ($> 13g/dl$) are not related to work capacity. To allow for the relationship between Hb and $ln(w)$ to be nonlinear, add a quadratic term to your model. Since there may also be diminishing marginal returns to height, also include a quadratic in height:

$$ln(w_i) = \gamma_0 + \gamma_1 ht_i + \gamma_2 ht_i^2 + \gamma_3 Hb_i + \gamma_4 Hb_1^2 + \gamma_5 educ_i + \gamma_6 age_i + \epsilon_i \tag{3}$$

a) Using the output from your regression, test the hypothesis that the coefficient on ht, γ_1, is zero with a 5% size of test.

b) Calculate the p-value associated with the t-statistic calculated in the previous question and show it in a graph.

c) Define Type I error associated to question a) and show it in a graph.

d) Repeat a) and b) to test whether test for γ_2, γ_3, and γ_4 each taken one at a time are zero with a 5% size of test. For each test calculate the p-value associated with the t-statistic and show it in a graph.

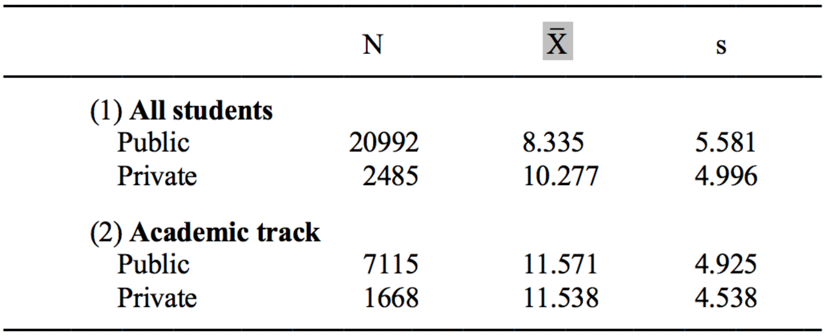
iv) What would happen if you add the natural logarithm of education $ln(educ)$ to model [3]? Explain.

Question 5

Education is perennially among the topics that voters care about most. The Coleman et al. Report provided a comprehensive discussion of the benefits of private and religious schools relative to public schools. It has been very influential in debates about school reform and, even to this day, it is frequently cited in debates about the pros and cons of vouchers and charter schools.

Table 1 contains a summary of some key results from the Coleman, et al. study based on a random sample of students who have completed a standardized math test. The table displays the number of students N, mean $\bar{X}$, and standard deviations s of math scores for students in each group.

Table 1: Math scores



	N	$\bar{X}$	s
(1) **All students**			
Public	20992	8.335	5.581
Private	2485	10.277	4.996
(2) **Academic track**			
Public	7115	11.571	4.925
Private	1668	11.538	4.538

Assume that math scores are drawn from distributions with different standard deviations for public and private school students.

(i) Examine the proposition that private school students have higher math scores than public school students by constructing a test with a significance level of 5% for the null hypothesis that $\mu_{private} \leq \mu_{public}$:

a) for all students;

b) for those students enrolled in the academic track.

(ii) Describe in a few words how your test would change if the null hypothesis was that the mean score for private school students is equal to the mean score for public school children, against the alternate hypothesis that the scores are different.

(iii) Redo part (i.a.) for all students under the assumption that $\bar{X}$ and s remain the same, but the sample sizes are reduced to 15 in each group.

(iv) What can you conclude about the value of an education in a private school relative to a public school?