HOMEWORK 1

Hint 1: the first few questions can be answered using the "sum" command. Reminder:

- Type "sum *varname*" to obtain summary statistics for the variable *varname*, including the average and standard deviation.
- Type "sum *varname* if *var2name*==X" to obtain summary statistics for the variable *varname*, restricting your attention to observations where some other variable *var2name* is equal to some value X.

Hint 2: Once you load each dataset into Stata, the the label window to find the variables the question is asking you about.

Question 1

Use the data in BWGHT.DTA to answer this question.

1. How many women are in the sample, and how many report smoking during pregnancy?
2. What is the average number of cigarettes smoked per day? Is the average a good measure of the "typical" woman in this case? Explain.
3. Among women who smoked during pregnancy, what is the average number of cigarettes smoked per day? How does this compare with your answer from part 2, and why?
4. Report the average family income and its standard deviation in dollars.

Question 2

The data in MEAP01.DTA are for the state of Michigan in the year 2001. Use these data to answer the following questions.

1. Find the largest and smallest values of math4. Does the range make sense? Explain.
2. How many schools have a perfect pass rate on the math test? What percentage is this of the total sample?
3. How many schools have math pass rates of exactly 50%?
4. Compare the average pass rates for the math and reading scores. Which test is harder to pass?

5. Find the correlation between math4 and read4. What do you conclude?

Question 3

The data in JTRAIN2.DTA come from a job training experiment conducted for low-income men during 1976?1977; see Lalonde (1986).

1. Use the indicator variable train to determine the fraction of men receiving job training.

2. The variable *re78* is earnings from 1978, measured in thousands of 1982 dollars. Find the averages of *re78* for the sample of men receiving job training and the sample not receiving job training. Is the difference economically large?

3. The variable *unem78* is an indicator of whether a man is unemployed or not in 1978. What fraction of the men who received job training are unemployed? What about for men who did not receive job training? Comment on the difference.

4. From parts 2 and 3, does it appear that the job training program was effective?

Question 4

The data in 401K.DTA are a subset of data analyzed by Papke (1995) to study the relationship between participation in a 401(k) pension plan and the generosity of the plan. The variable prate is the percentage of eligible workers with an active account; this is the variable we would like to explain. The measure of generosity is the plan match rate, *mrate*. This variable gives the average amount the firm contributes to each worker's plan for each \$1 contribution by the worker. For example, if *mrate*=0.50, then a \$1 contribution by the worker is matched by a 50 cent contribution by the firm.

1. Find the average participation rate and the average match rate in the sample of plans.

2. Now, estimate the simple regression equation

$$prate = \beta_0 + \beta_1 mrate + \epsilon$$

   and report the results along with the sample size and R-squared.

3. Interpret the intercept in your equation. Interpret the coefficient on *mrate*.

4. Find the predicted *prate* when *mrate*=3.5. Is this a reasonable prediction? Explain what is happening here.

5. How much of the variation in prate is explained by *mrate*? Is this a lot in your opinion?

Question 5

1. Use the clear command to clear all data from Stata.

2. Use the "set obs 500" command to create a new dataset with 500 observations.

3. Use the "gen" command to generate a variable x_n equal to a random draw from the uniform distribution with range [0,10]. (Reminder: you can do this by typing "gen x=10*runiform()". in Stata, the "runiform()" command generates a random draw from the unfiform distribution with range [0,1]; we just need to multiply each draw by 10.)

4. What are the sample mean and sample standard deviation of the x_i?

5. Randomly generate 500 errors from the Normal[0,36] distribution. (The procedure is similar to the one you used in step 3, except you give a different name to the variable you are generating ("u" instead of "x"), use the "rnormal()" command instead of the "runiform()" command, and multiply "rnormal()" by 6 instead of multiplying by "runiform()" by 10.)

6. Is the sample average of the u_i exactly zero? Why or why not? What is the sample standard deviation of the u_i?

7. Now generate the y_i as

$$y_i = 1 + 2x_i + u_i = \beta_0 + \beta_1 x_i + u_i$$

   Use the data to run the regression of y_i on x_i. What are your estimates of β_0 and β_1? Are they equal to the population values in the above equation? Explain.

8. Obtain the OLS residuals $\hat{u}_i$ and verify that the following equations hold (subject to rounding error):

$$\sum_i \hat{u}_i = 0, \qquad \sum_i x_i \hat{u}_i = 0$$

9. Repeat parts 1-7 with a new sample of data, starting with generating the x_i. Now what do you obtain for $\hat{\beta}_0$ and $\hat{\beta}_1$? Why are these different from what you obtained in part 7?

Question 6

1. Consider the following dataset:

Country	A	B	C	D	E	F	G	H	I	J
Per-capita income	6	8	8	7	7	12	9	8	9	10
% in agriculture	9	10	8	7	10	4	5	.5	6	7

   - The top row is per-capita income in thousands of dollars
   - The bottom row is percentage of the country's labor force engaged in agricultural work

   Estimate the following model *without* using the computer:

$$PercapitaIncome = \beta_0 + \beta_1 PercentageAgriculture + \epsilon$$

   SHOW YOUR WORK! (You can use the formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ derived in class.)

2. Interpret the coefficient on the agriculture variable.

3. Calculate R^2.

4. If the percent of the labor force in agriculture in another developed country was 8 percent, what level of per capita income (in thousands of U.S. dollars) would you guess that country had?

Question 6

Consider the following sample: $\{1, 5, 2, 4, 1\}$. Assume that the sample comes from a normal distribution and that the the standard deviation of $\bar{X}$ is equal to 1 (so that a z-test can be used). Let μ denote the population mean.

1. Use a z-test to test the null hypothesis $H_0 : \mu = 5$. What is the z-score? Do your reject the null hypothesis at a 5% level? At a 1% level? Explain.

2. Repeat the exercise above with the null hypothesis $H_0 : \mu = 2$.

3. Use the sample mean, the standard deviation of $\bar{X}$, and the critical values of the z-distribution to construct a 95% confidence interval for μ.

4. Repeat 3 to construct a 99% confidence interval for μ.

Question 7

Consider the following sample: $\{4, 4, 5, 6\}$.

1. Compute the sample mean, sample variance, sample standard deviation, and the standard error of the mean. Show your work!

2. What are the degrees of freedom?

3. Use the t-table to find the critical values for a 5% test and a 1% test (two-sided).

4. Compute the t-statistic for the null hypothesis $H_0 : \mu = 7$.

5. Do you reject the null hypothesis at a 5% level of significance? A 1% level of significance? Explain by comparing the t-statistic to the critical vaues.

6. Repeat 3-4 with $H_0 : \mu = 5$ as the null hypothesis.

7. Now consider a one-sided test with $H_A : \mu < 7$ as the alternative hypothesis. Use the t-table to find the critical values for a 5% test and a 1% test.

8. Do you reject the null hypothesis at a 5% level of significance? A 1% level of significance? Explain by comparing the t-statistic to the critical values.

9. Now consider a one-sided test with $H_A : \mu < 5$ as the alternative hypothesis. Do you reject the null hypothesis at a 5% level of significance? A 1% level of significance? Explain by comparing the t-statistic to the critical values.

10. Use information from the sample and the critical values of the t-distribution to construct 95% and 99% confidence intervals for μ.

Question 8

You have a dataset with N=50 observations measuring the effectiveness of a job-training program. The explanatory variable (on a zero to one scale) is weeks spent in job training, and the dependent variable is wages measured in terms of dollars per hour. Assume you ran a regression of wages on job training and found a coefficient of 3.61 on the job-training variable. Assume that you computed the standard error of the coefficient and found that it is equal to 1.62.

1. Is the coefficient on job training significantly different from 0 according to a 5% test? What about a 1% test? Explain by computing the t-statistic and comparing it to the critical values.

2. Suppose that you have reason to rule out the possibility that the effect of job training on wages is negative (i.e., you are allowed to use a one sided test with the alternative hypothesis that the coefficient on job training is positive). Is the coefficient on job training significantly different from 0 according to a one-sided 5% test? What about a one-sided 1% test? Explain by comparing the t-statistic to the critical values.

3. Construct a 95% confidence interval for the effect of job training on wages.