Exercise No. 8: Multiple Regression

New Stata Commands		**Old Commands Reviewed**	
regress			
estat vif			
estat hettest			
test			
predict			

{Be sure to open a log and keep it for your records}

1 Introduction

General Instructions: For this exercise you are allowed to use any reference material in your possession: textbooks, lecture notes, slides, old exercises that you created for this course, Stata Help files.

- Use the “Ec400tp.dotx” Word template to format your exercise just as you have done on the previous computer exercises.

v When you have completed the exercise, bring it **to class on Tuesday where we will use the exercise as a review of multiple regression.**

v Once having completed the exercise use it as a study aid for the final exam. You may have to answer questions relating to this exercise, or to a similar data analysis problem.

- Download the Stata data set: **multreg_2016.dta** from the course web site. These data come from the 1994 Consumer Expenditure Survey. (NB: these data are saved in Stata Version 11 format and should be usable by Stata versions from 11 onward.

- Immediately save the data set as **Ex8.dta**

Instructions for completing the exercise: The data set you will be analyzing contains information about a large number of families from all over the United States. You will perform various analyses on the data and you will answer, during class, various questions about this data set. Be sure to open a log and save it so that you can trace your work and copy information from it to your exercise, should you need to do so.

2 *Required steps*

Here you'll do various analyses. You should paste the results *and* the commands you used to get the results into your exercise: Be sure to number and identify each step so that the grader can see easily what you've done.

1. Now, *describe* the data set so that you’ll have a code book to use when performing the later analyses. Paste the results into your exercise.
2. *Summarize* all the variables in the data (don’t use the *detail* option) and paste the results into your exercise.

3. First, tabulate the number of people in the households. That is, produce a table that shows the number of households with one, two, three, ... members. There are two measures of family size in the data set; use the *integer* measure. Paste the results into your exercise.

4. Create a histogram showing the *percentage* households containing 1, 2, 3, ... etc., members, again using the integer measure. Title the histogram "Percentage of Consumer Units by Size." Paste the results into your exercise.

5. Tabulate the race of the household head and paste the results into your exercise. If the household contains a husband and a wife, then the "head" is arbitrarily assigned to be the husband. If two spouses are not present, the "head" may be either male or female. Race is coded as follows:

   1 White
   2 Black
   3 American Indian, Aleut, Eskimo
   4 Asian or Pacific Islander
   5 Other

6. Cross tabulate (i.e., tabulate together) the race of the head and the race of the spouse and paste the results into your exercise.

7. Tabulate the number of full- part-time college students in the consumer unit. Fractional numbers indicate part-year residence. Paste the results into your exercise.

8. Produce a histogram illustrating the relative frequencies (in percent) of the various *family types* in the survey, where the codes for family type are coded:

   1 Husband and wife (H/W) only
   2 H/W, own children only, oldest child < 6
   3 H/W, own children only, oldest child > 5, <= 17
   4 H/W, own children only, oldest child > 17
   5 All other H/W CU's
   6 One parent, male, own children only, at least one child age < 18
   7 One parent, female, own children only, at least one child age < 18
   8 Single persons
   9 Other CU's
   Use the *discrete* option to produce the histogram and make sure each bar is labeled with a code number. Paste the graph into your exercise with an appropriate title.

9. The data set contains a variable called *pig3* which stands for "Potential Income Group" (three categories, low medium and high). "Potential income" is created by combining education and occupational groups, and it is supposed to track the long range income of households. Using *before tax income* again, produce a series of 3 box plots arrayed from lowest pig to highest pig. You can also plot the median income as you did in the previous problem. Paste the graph into your exercise.

10. Generate a new variable *saving* which is defined as the difference between after tax income and total expenditures. Summarize the new variable (with detail) and paste the result into your exercise.

11. Sort the data by *pig3* and then **summarize** *saving* again using the *by pig3:* option to get a separate set of descriptive statistics (with detail) for each potential income group and the "missing" category. Paste your results into your exercise.

Regression and Inference

12. Let's test the following model that attempts to predict life insurance expenditures as a function of number of persons in each age/sex group, household total expenditures, and age of the head of household. The regression equation would be:

$$lifins = \beta_0 + \beta_1 as\_com1 + \beta_2 as\_com2 + \beta_3 as\_com3 + \beta_4 as\_com4 + \beta_5 as\_com5 + \beta_6 tot\exp + \beta_7 agehead$$

Paste the regression results into your exercise.

13. Next, test the hypothesis that *all* of the family composition variables (*as_com1 ... as_com5)* have coefficients equal to zero against the alternative that *at least one* of those variables has a non zero coefficient. Paste the results of the test into your exercise.
14. Test the hypothesis that *all* of the family composition variables *except as_com1 (as_com2... as_com5)* have coefficients equal to zero against the alternative that *at least one* of those variables has a non zero coefficient. Paste the results of the test into your exercise.
15. Test the hypothesis that the coefficients on variables *as_com1* and *as_com2* are equal to each other against the alternative hypothesis that they are different. Paste the results of the test into your exercise.
16. Now, we're going to create a new variable, *infant,* which will equal "1" if a child under two years of age is in the household and "0" if no child under 2 is present. (Hint: First create a new variable *infant* set equal to zero for all households and then use the "replace" command to set *infant* equal to "1" for all those cases where the variable *as_com5* is greater than zero (See, Hamilton). Then label the new variable *infant* "Infant present." *Describe* the variable infant and *tabulate* its values. Paste all these results into your exercise.
17. Regress *fdaway* (Food expenditures away from home) on the independent variables *infant, fam_siz, fincata2* and paste the results into your exercise.
18. Create an interaction variable *infam* which is equal to the product of *infant* and *fam_siz.* Label the new variable "Infant/family size interaction." Regress *fdaway* against *infant, fam_siz, fincata2, and infam.* Paste all these results into your exercise.
19. Check for multicollinearity in the last regression and paste your results into the exercise.
20. Create another interaction variable, *infinc,* by multiplying *infant* and *fincata2.* Label the new variable "Infant/family income interaction." Regress *fdaway* against *infant, fam_siz, fincata2, and infinc.* Paste all these results into your exercise.
21. Check this last regression for multicollinearity and paste your results into the exercise.
22. Check this last regression for heteroscedasticity and paste the results into your exercise.
23. Then, create a dummy variable equal to 1 if *agehead* is 75 or greater, zero otherwise. Name the new variable *oldhead.*
24. Now, we're going to test a theory of household health expenditures. Perform a multiple regression with annual health expenditures for the family *(health)* as the dependent variable, and family income after taxes *(fincata2),* number of people in each age group *(as_com1, as_com2, as_com3, as_com4, as_com5),* infant present *(infant),* and aged head of household present *(oldhead).* Include standardized regression coefficients in your regression output. Paste the result into your exercise.
25. Create an interaction term for *oldhead* and *fincata2* (call the new variable *old_fincata2)* and rerun the previous regression including the interaction term. Paste the result into your exercise.
26. ***Predict*** health expenditures using this most recent regression, create residuals for the regression, and graph the residuals against the predicted values of health expenditures. Paste the graph into your exercise.
27. Check the regression for multicollinearity. Paste the result into your exercise.
28. Check for heteroscedasticity. Paste the results into your exercise.
29. Save your working data set **Ex8.dta**

Bring to class on Tuesday:

1. One copy of your exercise. We will answers questions about the exercise in class. You are free to make notes on the exercise regarding what you did, interpretations of results, etc.