NYU STERN
NEW YORK UNIVERSITY LEONARD N. STERN SCHOOL OF BUSINESS

IOMS Department

Statistics and Data Analysis

Professor William Greene
Phone: 212.998.0876
Office: KMC 7-90
Home page: http://people.stern.nyu.edu/wgreene
Email: wgreene@stern.nyu.edu
Course web page: http://people.stern.nyu.edu/wgreene/Statistics/Outline.htm

Assignment 4
Linear Regression Model

Part I. Linear Regression Analysis

1. The data contained in the data file EconGrades.mpj (it is on the course website), is a sample of 32 observations on high school students. (The data were examined in a study in the *Journal of Economic Education*.) The variables in the data set are

GPA = the student's grade point average
TUCE = the student's grade on a test of economic literacy
PSI = 1 if the student participated in a special economics course, 0 if not.

Worksheet 1 ***

↓	C1	C2	C3	C4	C5	C6
	GPA	TUCE	PSI			
1	2.66	20	0	2.75	25	0
2	2.89	22	0	2.83	19	0
3	3.28	24	0	3.12	23	1
4	2.92	12	0	3.16	25	1
5	4.00	21	0	2.06	22	1
6	2.86	17	0	3.62	28	1
7	2.76	17	0	2.89	14	1
8	2.87	21	0	3.51	26	1
9	3.03	25	0	3.54	24	1
10	3.92	29	0	2.83	27	1
11	2.63	20	0	3.39	17	1
12	3.32	23	0	2.67	24	1
13	3.57	23	0	3.65	21	1
14	3.26	25	0	4.00	23	1
15	3.53	26	0	3.10	21	1
16	2.74	19	0	2.39	19	1

a. Construct a scatter plot of **TUCE** (on the vertical axis) against **GPA** on the horizontal axis.
b. Does the scatter plot suggest that there is a relationship between **GPA** and **TUCE**? Explain.

c. Now, produce a scatter plot that separates the two groups. [Graph → Scatter Plot → With Groups: Then specify the variables and **PSI** as the categorical variable for the grouping.] Does the plot suggest that there is a different relationship for the two groups? (We will pursue this below.)

2. For the n=32 observations in problem 1, the basic statistics are:

$$\overline{GPA} = \tfrac{1}{n}\Sigma_{i=1}^{n} GPA_i = 3.1172$$

$$\overline{TUCE} = \tfrac{1}{n}\Sigma_{i=1}^{n} TUCE_i = 21.938$$

$$S^2_{GPA} = \tfrac{1}{n-1}\Sigma_{i=1}^{n}(GPA_i - \overline{GPA})^2 = 0.217821$$

$$S^2_{TUCE} = \tfrac{1}{n-1}\Sigma_{i=1}^{n}(TUCE_i - \overline{TUCE})^2 = 15.221774$$

$$S_{GPA,TUCE} = \tfrac{1}{n-1}\Sigma_{i=1}^{n}(GPA_i - \overline{GPA})(TUCE_i - \overline{TUCE}) = 0.704657$$

a. Compute the correlation coefficient between GPA and TUCE. Interpret the result.
b. Compute the constant term, *a*, and the slope, *b*, in the linear least squares regression of TUCE (the dependent variable) on GPA (the independent variable). The result should help to confirm the conclusion you drew in part 1.b. above.
c. Use the Stat → Regress feature in Minitab to confirm the computations in part b.
d. Compute the residual standard deviation, S_e using the statistics given above.

3. To pursue the question raised in part 1.c. earlier, produce a scatter plot that contains the two regressions in it for the two groups of students. (Scatterplot – With Regression and Groups.) Does this enable you to reach a conclusion about the difference in the relationship between GPA and TUCE for the two groups of students?

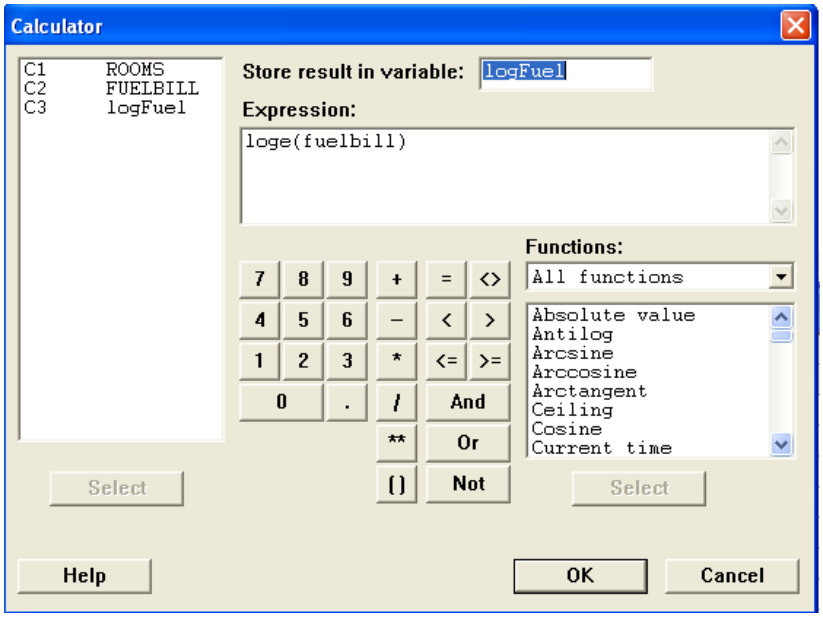
4. A set of basic statistics for the two variables GPA and TUCE in a sample of *n* = 32 students appears in problem 2. One of the sets of results that would be produced from the regression would be an analysis of variance table, containing the results shown below.

Source	Degrees of Freedom	Sum of Squares	Mean Square	F Ratio	P Value
Regression	1	$\Sigma_{i=1}^{n}(\hat{y}_i - \overline{y})^2$	$\frac{\Sigma_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{1}$	$\frac{(n-2)\Sigma_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\Sigma_{i=1}^{n} e_i^2}$	$2P[z \geq \sqrt{F}]^*$
Residual	n-2	$\Sigma_{i=1}^{n} e_i^2$	$\frac{\Sigma_{i=1}^{n} e_i^2}{n-2}$	Empty	Empty
Total	n-1	$\Sigma_{i=1}^{n}(y_i - \overline{y})^2$	$\frac{\Sigma_{i=1}^{n}(y_i - \overline{y})^2}{n-1}$	Empty	Empty

Using the basic statistics given in problem 2, fill in the numbers for this analysis of variance table. (Hint: There is another very useful slide in your class notes.) Now that you have filled in the table, what is the R^2 in this regression?

5. The file heating.mtp deals with the heating bill for dwelling units of variousnumbers of rooms.
(a) What are the names of the variables in this data set?
(b) Obtain a scatterplot of the two variables. Which variable should be on thehorizontal axis?
(c) Find the linear regression equation resulting from the regression of FUELBILL onROOMS.
(d) Examine the residual-fitted-plot from this regression. Is there a tendency forlarger dwelling units to have more variable heating bills?

Minitab HINT: Use **Calc →Calculator** and then set up the resulting panel asfollows:



In Minitab, thefunction name for natural logs is LOGE. In the **Functions** list, this is identified as "Natural log." (You can just type the "loge," you need not use the menu.) You get to select the name for **Store result in variable**. Here logFUEL seems areasonable choice. The new variable will be placed in the next open column.Instead of giving a name to the new variable, you can select a column.

6. The increasing variation of the residuals as the number of rooms increases that you see in the residual-versus-fitted plot of the previous problem suggest that the constant variance assumption of the regression model is wrong for these data. A common corrective action is to replace thedependent variable, here FUELBILL, with its logarithm. Take this action and repeat theregression. Note the revised fitted model equation and give the new residual-versusfittedplot. Finally, can you compare the value of R^2 for the two versions?

7. A number that many researchers have analyzed in the study of health care is the elasticity of health care expenditure with respect to per capita income. You can obtain an estimate of this important parameter by using a log-on-log regression of the log of health care expenditure (PCHEXP) on the log of per capita income (GDPC) using the WHO health care data, WHO-HealthStudy.mpj. (Note, you will have to obtain the logs of the two variables first. Use the hint in Problem 8.) Carry out the regression. What is your estimate of the elasticity? Is the provision of health care expenditure *elastic* (elasticity > 1) or *inelastic* (elasticity < 1)?

Assignment 4

8. The following data appeared in the Wall Street Journal article (December 30, 1986) that is mentioned in Slide 44 of Session 16 in your class notes. These are height in inches and monthly income data for a sample of MBAs.

a. Produce a scatter plot of Income (Y axis) against Height (X axis).
b. Compute the intercept and slope of the linear regression of Income on Height.
c. The slope can be interpreted as the marginal value (in $/month) of an additional inch in height. What is the value?
d. What is the R^2 in your regression?
e. Test the hypothesis that there is no relationship between height and income in the populatioin from which these data were drawn.

Height	Income
70	2990
68	2910
75	3150
67	2870
66	2840
68	2860
69	2950
71	3180
69	2930
70	3140
68	3020
76	3210
65	2790
73	3220
71	3180
73	3230
73	3370
66	2670
64	2880
70	3180
69	3050
70	3140
71	3340
65	2750
69	3000
69	2970
67	2960
73	3170
73	3240
70	3050