17.871, Political Science Lab
Spring 2002
Problem set # 4: Multiple regression, sampling, and hypothesis testing

Handed out: March 14, 2002
Due back: April 4, 2002 (April 9, 2002 for those scheduled to present on April 4)

When you hand back in your problem set, please estimate the number of hours (rounded to the nearest quarter hour) it took you to complete it.

Part I

Do the following Review Exercises in Freedman, 3rd edition:

Chapter 18 (pp. 327–329), # 1, 5
Chapter 20 (pp. 371–373), # 5, 6, 8
Chapter 21 (pp. 391–394), # 1, 2, 12, 14
Chapter 23 (pp. 425–428), # 10, 12
Chapter 26 (pp. 497–501), # 2, 4, 6, 7
Chapter 29 (pp. 565–567), # 1, 2, 3, 6, 8, 11

Part II

The purpose of this Part is to help you get a better understanding of probability functions, and how we use them to conduct hypothesis tests and construct confidence intervals. Recall that all of our coefficients from our regressions are random variables. We have estimated them, estimated their variances, and we know how they are distributed.

A hypothesis test asks the following question: **IF** a variable were actually equal to some value (usually 0, which means “no causal effect”), **THEN** what is the likelihood that we would observe that variable taking on a value equal to our predicted coefficient? For instance, if our coefficient is far away from 0 (relative to its standard error), then this is very unlikely. In that case, we reject the “null hypothesis”.

A confidence interval asks a different question: Given that our coefficient is a random draw from a distribution with a *true mean* and a variance which we know, how far away from our observed coefficient could that true value be if our estimated coefficient remains within the the 95% probability mass surrounding the true mean. (This probably is easier to understand if you imagine the true mean being distributed in the 95% probabilty mass around your estimated mean. Though it yields the same results, it is technically incorrect.)

Try to keep these ideas in mind as you answer the following questions.

1) Find the probability that a normally distributed random variable *X* with a mean of 0 and a standard deviation of 1:
   a. takes on a value above 0
   b. takes on a value above 0.84
   c. takes on a value above 1.96
   d. takes on a value either above 1.96 or below –1.96

1) If a variable *X* is distributed normally with a **mean** 0 and a **variance** of 1 [written *X~N(0,1)*], what is *Z* if:
   a. X takes on a value less than *Z* 97.5% of the time?
   b. X takes on a value less than *Z* 95% of the time?
   c. X takes on a value greater than –*Z* and less than *Z* 95% of the time?
   d. *X* takes on a value greater than –*Z* and less than *Z* 95% of the time?

2) a. If *X ~ N(4,9),* what is the probability that a given sample *x* is greater than 6.5?
   b. If *X ~ N(-3,4),* what is the probability that a given sample *x* is value greater than 6.5?

3) If *X~T(0,1),* with 20 degrees of freedom:
   a. What is the probability that a given sample *x* is greater than 2.09?
   b. What is the probability that a given sample *x* is less than –2.09 or greater than 2.09?
   c. Find t such that *–t < X < t* 95% of the time.
   d. Find t such that *–t < X < t* 99% of the time.

4) If *X ~ T(3,2.25)* with 20 degrees of freedom:
   a. What, approximately, is the probability that *X* takes on a value greater than 1.155?
   b. Find t such that *–t < (X-3)/1.5 < t* 99% of the time.

Part III

General directions. The following problems present you with real-life research situations and ask you to make judgements about either the data you have and what they tell you *or* the data you would need to answer the question presented you. There are no trick questions here.

Each of the questions asks you to write something to explain what you did. Please take the written assignments seriously, because you will be graded on quality of writing and substance. Accompanying most of the questions you should hand in a log file that shows the results you are talking about and a “do” file that could reproduce those results if necessary.

A. The MIT administration is interested in studying why freshmen make the dormitory choices they do. There is a debate over whether freshmen care more about the size of the rooms in various dormitories (larger rooms are better) or about the physical condition of the building (newer is better). Being empirical scientists at MIT, they decide to settle this matter using a multiple regression. They gather data about dormitory preferences, average room sizes, and age of the buildings. The (mostly fictional) data are presented on the last page of this problem set.

Using these data, do the following.

1. Run the multiple regression necessary to answer the above debate.

2. Run two separate bivariate regressions on the same data, first with room size as the independent variable and then building age. Compare the coefficients you get with the multiple regression with those you get from the two bivariate regressions.

3. Write a (clearly worded) paragraph that explains why the coefficients in the two analyses are so different. Explain which sets of coefficients you trust and why.

B. You are interested in knowing what happens to the electoral support for congressional candidates when they get redistricted. To explore this topic, you gather the election returns for Jim Courter, a Republican member of the U.S. House of Representatives from New Jersey, in his run for reelection in 1982, after his House district had been redrawn to reflect the 1980 census.

The election returns in New Jersey are reported at the town level. The data file with these data is located in this Athena file: /mit/17.871/Examples/courter82.dta. This data set has data for each town in the district. For each town, the variables record the percentage of the vote Courter received in that town in 1982 (cvote82), the vote George Bush received for president in that town in 1980 (pvote80), and whether the town was new to Courter's district in 1982 (newtown=1) or whether it had been in Courter's district in prior years (newtown=0). (The newtown variable obviously is the one of interest here. The cvote82 variable is intended to control for the partisanship of the town.)

1. How much more poorly did Courter do in the new towns of his district in 1982, controlling for each town's partisanship? (Turn in a log file of the analysis, circle where the answer is. No need to write a sentence.)

2. If you do the regression of cvote82 on pvote80, you will see that Courter's vote in a town is related to the partisanship of the town. You know from studying American elections that partisanship is the strongest predictor of how someone votes. Armed with this piece of information, and assuming that the presidential vote in a town is a perfect measure of that town's partisan leanings, answer this question: To what extent did Courter's support in the towns deviate from a purely partisan explanation? (Hint: I am trying to get you to interpret the regression coefficients from this regression.) How does the answer to this question vary if

you restrict yourself to (a) just the new towns in the district and (b) just the old towns in the district. (Write a paragraph that answers these questions, attaching a log file of any statistical procedures you ran to produce the answer.)

3. Construct a single-equation (i.e., one regression) approach to this problem: Consider the regression of cvote82 on pvote80 for each of the two subsets of the data separately. (By subsets I mean new towns *versus* old towns.) How do the intercepts and the slopes of the two regressions differ? What do you make of these differences substantively? (Hint: I am trying to get you to think about the use of *interaction terms* in your regression.) Write a paragraph that answers these questions, attaching a log file of any statistical procedures you ran to produce the answer. If you generate any new variables, show how the new variables were constructed by turning in a "do" file that produces the new variables.

C. You are interested in what makes for a really great Mechanical Engineering department. You know that the National Academy of Sciences did a study of this issue a number of years ago, and so you enter the data they gathered into the computer. The codebook is available on the web at http://web.mit.edu/17.871/Examples/Codebook.html. The data are available at /mit/17.871/Examples/MechEng.dta. The three primary measures of program quality are called rate93q, rate93e, and rate93c. These measures were derived through surveys of department heads in Mechanical Engineering. These variables are basically mean values of those ratings for each department. The rest of the data set consists of characteristics of the graduate program, including things about the faculty and the students.

1. Run a regression that predicts the scholarly quality of the program faculty as a function of the number of faculty and the number of students. Decompose the total effect of these two independent variables on the dependent variable into two components: the direct effect and the indirect effect. Show your calculations.

2. What factors lead to a graduate program in Mechanical Engineering being considered effective? Limit yourself to four independent variables. Write a paragraph in which you speculate why each of these factors *should* effect effectiveness ratings, another one discussing how the factors are measured (i.e., deal with any transformation issues, recoding, etc.), and write a paragraph or two summarizing your findings.

3. Redo the previous analysis, this time reporting the standardized coefficients of the regression. Do the standardized coefficients tell you anything new about the question of Mechanical Engineering department quality?

Data set for problem II-A, Dormitory preference (mostly fictitious data)

Dormitory name	# of first choices in freshman housing lottery, 1999	Year building was built	Average sq. ft. of dormitory room space per resident
Baker	120	1949	145
Bexley	31	1920	107
Burton-Connor	50	1940	135
East Campus	59	1930	127
MacGregor	125	1970	150
McCormick	81	1965	200
New	56	1976	175
Next	98	1981	185
Random	27	1910	97
Senior	75	1916	125