STAT 8003, Homework 10

Group # ... (Replace this)
Members: ... (Replace this)

November 18, 2013

Due at 5:30pm on class on Thu., Nov. 21. Please submit one and only one pdf file for your group via blackboard. Each sup-problem is 10 points (Total points = 70).

Problem 1. Criminologists are interested in the effect of punishment regimes on crime rates. This has been studied using aggregate data on 47 states of the USA for 1960. The data set contains the following columns:

Variable	Description
M	percentage of males aged 1424 in total state population
So	indicator variable for a southern state
Ed	mean years of schooling of the population aged 25 years or over
Po1	per capita expenditure on police protection in 1960
Po2	per capita expenditure on police protection in 1959
LF	labour force participation rate of civilian urban males in the age-group 14-24
M.F	number of males per 100 females
Pop	state population in 1960 in hundred thousands
NW	percentage of nonwhites in the population
U1	unemployment rate of urban males 1424
U2	unemployment rate of urban males 3539
Wealth	wealth: median value of transferable assets or family income
Ineq	income inequality: percentage of families earning below half the median income
Prob	probability of imprisonment: ratio of number of commitments to number of offenses
Time	average time in months served by offenders in state prisons before their first release
Crime	crime rate: number of offenses per 100,000 population in 1960

For the following problems, you should NOT use "lm" function or any existing function

which yields the results immediately. Write your own code to solve the problem.

a) Plot the scatter plot matrix between these variables. Decribe the pattern you observe. We are interested in studying the adjusted effect of punishment (Time or Prob) on Crime. Is it a good idea to inlude both variables in the model?

b) Construct a linear model to study the relationship between Crime (Y) and Prob, adjusting for the effect of the 13 characteristcs variables (M, So, ..., Ineq). Note that we don't want Time to be included in the model. Is the effect of Prob significant on Crime after adjusting for other characteristic variables? What is the p-value? Please use R to show the results.

c) Perform the same analysis to study the adjusted effect of Time on Crime, adjusting for 13 characterist variables. Compare to model b), which variable do you think has a stronger adjusted effect on Crime? Why?

d) If the purpose is to only study the adjusted effect of Time on Crime, is it a serious problem that we include 13 characteristic variables which might be correlated in the model? If the purpose is to study the effect of all these covariates (Time and 13 characteristic variables) on Crime, is it a serious problem that we include in total 14 variables in the model? Use the data to justify your answer.

e) Now suppose we take the first four principal components of $X = (X_1, \ldots, X_{13})$ (13 characteristic variables) to be the variables included in the model, together with Prob. Please write out the model. What is the LSE of the coefficient for Prob? What is its variance?

f) Let's get back to Model b). Now suppose the investigator is not only interested in the adjusted effect of Prob, but also the effect of other variables. The investigator noticed that the variables Po1 and Po2 are highly correlated. It is reasonable to include (Po1 + Po2)/2 in the model to replace these two variables. How to test whether such replacement is fine? Formulate this problem and use hypothesis testing to solve it.

g) How to test whether the effect of the first five characteristic variables (M, So, ...,Po2) are all zero? What's your test result?

Remark:

Denote the outcome random variable by Y, and the available covariates by $X_1, X_2, \ldots, X_p$. Suppose the researcher's main interest is to study the relationship between Y and X_1. In many situations, X_1 are correlated with $X_2, \ldots, X_p$. To study the relationship between Y and X_1, it is better to adjust for other covariates. (Please figure out why. You don't need to include this part in the homework though.)

Suppose the true model is

$$Y = \beta_0 + \sum_{i=1}^{p} X_i \beta_i.$$

We say β_1 is the effect of X_1 adjusting for the effect of $X_2, \ldots, X_p$.