Introduction to Computational Statistics INSH 5301

Homework 08

03/09/2020

PLEASE copy and paste the whole question number and text into submission so I can grade easily.
When I grade easier, you might get better grade!
For this homework you'll need to use some real world data to answer some research questions using multiple regression. The data, along with the data description, can be downloaded from the course material section in Blackboard.

Problem 1: Seat belts
1. For this problem we are trying to test the effectiveness of mandatory seat belts usage laws in reducing traffic mortality. The independent variable (Y) is fatalityrate. Run all your regressions using the lm parameter. You need to download the seatbelts dataset to complete this part.
1.1 Run an interpret the bivariate regression of fatalityrate on primary (this is a binary variable that indicates the primary enforcement of seat belt laws).
1.2 Create a correlation matrix for the entire dataset using the cor command - exclude non-numeric variables -. Do you think that the exogeneity assumption may not be satisfied for the previous regression? (Explain)
1.3 Using the dataset provided, run a set of 3 additional multiple regressions by sequentially adding other variables that you think are relevant in the model. For each regression (1) Argue why you add the particular additional variable, (2) interpret the parameters, (3) the R^2 and adjusted R^2, and, (4) the F-statistic.

2. College on educational attainment.
For this problem we are going to explore the effect of distance from college on educational attainment. The independent variable (Y) is years of completed education ed. All the estimated regression parameters for this part should be computed using linear algebra - see lesson 8.2 -. Also, any statistic (F-statistic or R^2) should be computed manually and without the use of the lm command - you can use the command to verify your work -. You need to download the collegeDistance dataset to complete this part.
2.1 Run an interpret the bivariate regression of ed on dist (distance to college). What's the estimated slope?
2.2 Now, run a multiple regression of ed on dist but also include: bytest, female, black, hispanic, incomehi, ownhome, dadcoll, momcoll, cue80, and, stwmfg80. What is the estimated effect of ed on dist? Compare your result to the previous estimation. Explain why the effects may differ.
2.3 Compute the R^2 and the adjusted R2 for both regressions and interpret its significance. Which measure of goodness of fit you prefer in each regression?
2.4 Bob is a non-hispanic black male. His high school was 20 miles from the nearest college. His base year composite score (bytest) was 58. His family income in 1980 was \$26, 000, and his family owned a house. His mother attended college, but his father did not. The unemployment rate in his county was 7.5%, and the state average manufacturing hourly wage was \$9.75. Predict Bob's years of completed schooling using both regressions and compare the results. Which result you prefer? (Explain)
2.5 Test if all the parameters of the model are simultaneously equal to zero.