AMS 315

Data Analysis, Spring 2020

First Computing Assignment

The first report is due on Thursday, March 26, but can be submitted without penalty by March 31. This report is worth 60 examination points. Please remember that there is a second project coming, so that you should finish the first project as soon as possible. Please submit your project as instructed on the Class Blackboard. Please submit your report of Project 1 in docx or pdf form. Each student has one chance to resubmit the report before the deadline. Detailed submission information is online.

Project 1 has two parts. There are three files for this project. Two of the files are for part A, and one file is for part B. The files are labeled with the last four digits of your Stony Brook ID number.

Part A

The model for the Part A assignment is a first data and statistical processing task that a newly hired statistician might be given. Your report should address the issues that your future supervisor would want to know about. Part A is worth 20 points. The two files for part A each contain a column for subject ID and a column for either the dependent variable value or the independent variable value. Your first task is to sort the two files by subject ID and merge them. You should not just use "cut and paste" to merge your data. Second, you are expected to deal with missing data. Your report should contain the count of the number of subject IDs that had at least one independent variable value or dependent variable value. It should also include the count of the number of subject IDs that had an independent variable value, the count of the number of subject IDs that had a dependent variable value, the count of the number of subject IDs that had both an independent and dependent variable value, and the count of the number of subject IDs that had at least one independent variable value or dependent variable value.

Your second task is to impute the missing values. There are a number of missing data procedures. Often a statistical package has imputation algorithms in the software. For example, R has a package called MICE that has a number of options. You may not choose listwise deletion

or mean imputation or its equivalent (such as median imputation). Specify your choice in your report. Often, the choice of imputation method has little effect on the results if the fraction of missing data is 30% or less.

Part B

Part B is worth 40 points. The data file for part B contains one line for each subject ID. The line will contain the subject ID, the value of the independent variable, and the value of the dependent variable. A transformation of either IV or DV or both may be required. You should read the text for suggestions on fitting a model. An approximate lack of fit (LOF) test should be applied. It is your responsibility to find repeated (or near repeated) independent variable values. That is, you will have very few exact repeats of an independent variable value. You should bin near repeated data into one level. For example, suppose that $x_1 = 1.01, x_2 = 1.02, x_3 = 1.03$ and $y_1 = 2, y_2 = 3, y_3 = 4$. While there are not exactly repeated *x* values, you could bin these points into one group of nearly repeated points. That is, choose the average x-value as the value of x after binning. Then your binned data would be $x_1 = 1.02, x_2 = 1.02, x_3 = 1.02$ and $y_1 = 2, y_2 = 3, y_3 = 4$. Now perform a LOF test on the data set after binning all near repeated values.

You must submit a one-page report on Problem A and a one-page report on Problem B. Each report should have four sections. The introduction should contain a statement of the problem and the objective of the paper. This part is easy: your problem is to recover the function that was used to generate the dependent variable value based on the value of the independent variable. The data you receive will be generated by a simulation program. The second section should describe your methodology. Specifically, how the files were merged, the program used to perform the statistical analysis, whether you used linear regression and additional procedures such as an approximate lack of fit test, how much missing data was present in the data, and the procedure for dealing with missing data. The third section should contain your results: what fraction of the variation of the dependent variable was explained, the analysis of variance table, the fitted function, confidence intervals for slope and test of the null hypothesis that the slope was zero. The fourth section should be conclusions and discussion. This section should focus on "big picture" issues. Was there an association between the variables? How important was it? That is, what was the r-squared value. What is your fitted function? You may submit a longer appendix of computer work and programs.

Grading of last semester's Project 1:

These are the grading penalties for Project 1, Fall 2019, presented in order of point deduction

Part A

-20 no report other than compilation of computer code

-20 no reported function or statistics

-20 inconsistent reported functions or statistics

-20 incorrect missing data report

-20 used only complete data points (used listwise deletion)

-20 results not consistent with assigned data

-15 used median imputation (or mean or other single valued imputation method)

-15 no specification of imputation method

-15 incorrect report of significance of association

-15 incomplete missing data report

-15 incorrect number of observations in analysis

-10 "99.9% of variance explained";

-10 99.9% independent variable

-10 "linear regression represents 99% of data";

-10 incorrect lof test

-10 incomplete specification of imputation method

-5 incorrect interpretation of CI

-5 low r-squared does not mean that transformation will help

-5 inconsistent reports of number of observations (792 vs. 791)

-2 no r or r-squared reported

Part B

-40 no report

-40 no report of function or function parameter estimates

-25 correct transformation but no report of function parameter estimates

-20 incorrect transformation selection--the r-squared for your selected transformation was 0.3707 one of the lowest values obtained

-20 incorrect interpretation of lof results;

-20 incorrect number of observations

-20 did not pick a final model

-20 incorrect report of corr(IV,DV)=0.008; correlation values reported are too small in absolute value for this data set

Important note:

Simply submitting your computer output is not acceptable and will receive a grade of 0. You must submit a formal report to get non-zero credit for this assignment.

How a student should submit the project 1 reports

1. The report should be uploaded as a pdf or docx (Word) file and submitted via the link for the first project assignment on blackboard.

2. (Not recommended) An alternative way to submit your report is to send an email (attaching your report) to TA. The file must be named with the last five digits of your Stony Brook ID_your last name_Project1.pdf/doc/docx. The email address is shuqian.xie@stonybrook.edu.

3. The report should be in a single file. Both the one page report for Part A and the one page report for Part B should be submitted in the same file.

Warnings of Plagiarism in Your Report

1. Plagiarism is a serious issue. Our expectation of you is that the work that you present in your report is yours alone.

2. Results: If you analyze the wrong data set, the grade for your report will be 0, whether or not plagiarism is involved. If you have been working jointly with other students, compare your results with their results. If they are same, then there may be a plagiarism problem.

2. Codes. You are encouraged to attach your computer code in an appendix to your report. If two students have the same codes, there may be a plagiarism problem.

3. Content of the report. You have been given example reports on the class blackboard. Your report should be in your own words rather than a minor modification of the examples. Two students who submit the same report except for statistical results have engaged in plagiarism.