AMS 315

Data Analysis, Spring 2023

First Computing Assignment

This report is worth 100 examination points. The first report is due on Tuesday, April 11. A paper submitted before April 11 at 11:59 pm will receive an additional 20 point on-time bonus. I will grant an extension to April 19 for a student who requests it. 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 on the Class Blackboard as instructed below. Please submit your report of Project 1 (both parts) in one pdf file. Each student has one chance to resubmit the report. Detailed submission information is given below.

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 six digits of your Stony Brook ID number. You must analyze the data set that is assigned to you. Otherwise, your grade will be 0.

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: how many observations, fraction of missing data in independent variable and dependent variable, and imputation of missing data. Please remember to include the description of the data using linear regression. That is, report the OLS estimates of the parameters, the r-squared value, and the analysis of variance table. Part A is worth 40 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 many of missing data procedures. Often a statistical package has imputation algorithms in the software. For example, R has a package called MICE that has several options. You may not choose listwise deletion or mean imputation (or its equivalent 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.

Your third task is to report the OLS estimates, the r-squared value, and the anova table.

Part B

Part B is worth 60 points. The data file for part B contains one line for each subject ID. The line contains 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. There is software in R that performs an approximate lack of fit test.

Report

You must submit a one-page report on Problem A and a one-page report on Problem B in a pdf format file. Each report should have four sections and should address the standard issues.

1. *Introduction*. The introduction should contain a statement of the problem and the objective of the paper. Some of the questions that you should answer are: What is the objective of your effort? What are your research questions? What is the background of

this work? The introduction 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.

2. *Methods*. The second section should describe your methodology. Specifically, in part A, how were the files were merged? What was the program used to perform the statistical analysis? What were the statistical missing data techniques used? Did you use linear regression? What statistical package did you use? In Part B, did you use additional procedures such as an approximate lack of fit test?
3. *Results*. The third section should contain your results: What fraction of the variation of the dependent variable was explained? What was the analysis of variance table? What was the fitted function? What was the confidence interval for the slope? What was the conclusion to the test of the null hypothesis that the slope was zero.
4. *Conclusions and Discussion*. 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.

If you include a table or figure, you must discuss it. Tables and figures should be numbered and titled.

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 report

1. *Preferred method*: The report should be uploaded as a pdf 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 six digits of your Stony Brook ID as in your last name_Project1.pdf/doc/docx. The email address is wenhan.gao@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.

Signs of Plagiarism in Your Report

1. Plagiarism is a serious issue. My 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 may attach your computer code in an appendix to your report. If two students have the same codes, there may be a plagiarism problem.

3. Two students who submit the same report except for statistical results have engaged in plagiarism. The enabler (originator of paper) is more guilty in my eyes than the plagiarizer.

Grading of a past semester's Project 1:

These are the grading penalties for Project 1 for common student mistakes from a past semester presented in order of point deduction value:

Part A, 40 points

-40 no report other than compilation of computer code

-40 no reported function or statistics

-40 inconsistent reported functions or statistics

-40 incorrect missing data report

-40 used only complete data points (used listwise deletion)

-40 results not consistent with assigned data

-30 used median imputation (or mean or other related imputation method)

-30 no specification of imputation method

-30 incorrect report of significance of association

-30 incomplete missing data report

-30 incorrect number of observations in analysis

-20 "99.9% of variance explained"

-20 99.9% independent variable

-20 "linear regression represents 99% of data"

-20 incomplete specification of imputation method

-10 incorrect interpretation of CI

-10 low r-squared does not mean that transformation will help

-10 inconsistent reports of number of observations (792 vs. 791)

-5 no r or r-squared reported

Part B

-60 no report

-60 no report of function or function parameter estimates

-40 correct transformation but no report of function parameter estimates

-30 incorrect transformation selection--the r-squared for your selected transformation was one of the lowest values obtained

-30 incorrect interpretation of lof results

-30 incorrect number of observations

-30 did not pick a final model

-30 incorrect report of corr(IV,DV); correlation values reported are too small in absolute value for this data set