AMS 315

Data Analysis, Spring 2024

First Computing Assignment

One Predictor Linear Regression

Introduction

This assignment is due on Thursday, April 4. This report is worth 100 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 on the Class Brightspace 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 before the deadline. Detailed submission information is given below.

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

Part A

Part A is worth 40 points. 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.

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.

Part B

Part B is worth 60 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 textbook (Chapter 11.1) 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. Often a transformation does not improve the apparent extent to which the data satisfies the assumptions of Chapter 11. Please check the r-squared of the data as given to the r-squared of the data after you transform it. Also, please check the residual plot of the data. It may be helpful to apply these checks to the data in part A.

Report

You must submit a one-page report on Problem A and a one-page report on Problem B, both parts in one single pdf format file. Each report should have four sections.

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, how were the files were merged? What was the program used to perform the statistical analysis? What were the statistical techniques used? Did you use linear regression Did you use additional procedures such as an approximate lack of fit test? How much missing data was present in the data? What procedure did you use to deal with missing data.
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.

You are allowed an appendix to your report and there is no page limit on the appendix. If you include a table or figure, you must discuss it. Tables and figures should be numbered and titled.

End of Project Assignment

Practical Advice for Data Analysis Project

Data is available on the Class Brightspace.

- Choose your software—see file "Obtaining Statistical Computing Resources"

Choices of software for computing project:

- R is free and will increase in prevalence as time goes on, but it requires some computing sophistication.
- SAS is popular in the pharmaceutical industry and in banking. Federal regulators insist on SAS programming. It is not a forgiving package and requires some computing sophistication.
- Minitab is user friendly, menu driven, and well documented. It has many quite sophisticated techniques.
- SPSS is user friendly, menu driven, and well documented. It is popular in market research and social science related industries.
- Excel does not have a good reputation for statistical work at this point. It is very good at data processing issues. Since it is extremely popular, it is to your professional advantage to know the basics of the problem. Most students in past semesters have used Excel to merge the data files for the first part of the first project.

Data Processing

- Use computer programs to edit data rather than cut and paste or deleting rows with missing data one by one. Also, it might be helpful to add the date of your editing to the file name.
- Make sure that you understand how your statistical package will deal with missing data. That is, check whether your package has a default option of listwise deletion.
- It is recommended that you set a plateau of creating a data file. R has software that makes merging data files easy. Excel is also popular for this. Many students have found the VLOOKUP function in Excel to be helpful in merging the data sets. Create files with a relatively small number of cases with the problems that you are dealing with and work to create software that handles the problem. Then use the perfected software on your larger files.
- Do an internet search on "missing data" and the package that you wish. Also, the package may have an internet site with information about how to deal with problems. For example, Minitab has documentation on getting started. The help menu has an option for "minitab on the web." One can choose that option and enter "missing data" in the search area. It will return a selection that is titled "Remove rows with missing values." This is not an acceptable choice for dealing with missing data.
- Check your work. Make sure that the final data file is correct. The grading in part A of project 1 is largely determined by whether you have the correct number of cases.
- Do not let data processing issues stop you from starting the analysis of part B.

Analyzing data:

- Calculate summary statistics on each variable (i.e., mean, standard deviation, median, quartile points, maximum, minimum, number of cases). The histogram of the variable may also be helpful.
- Examine the scatterplot.
- Calculate the bivariate statistics (i.e., correlation coefficients). Most of the correlations that one sees in practice are minimal in size. Consequently, a small fraction of the part B data sets has dependent variables that have no association with the independent variable.
- Calculate the Chapter 11 regression statistics. Plot the residuals against the predicted variable values. One hopes for a patternless residual plot. If not, then try transformations of both the independent and dependent variables. If there are repeated values of the independent variable, one can perform a lack of fit test.

Writing the report

- You must have a report. If you hand in a lot of computer output with no explanation, you will get 0 points.
- The classic format for scientific reports is: 1. Statement of problem and introduction; 2. Methods; 3. Results; 4. Conclusions and discussion.
- Do not plagiarize—either from the example report or from your fellow students. It is acceptable to quote another paper or book if you put quotation marks around the quoted material and identify the source. The quoted material must be a small fraction of the text. Cutting and pasting material from someone else's paper that has errors in it will be obvious in grading.

Grading of a past semester's Project 1:

These are the grading penalties for Project 1 from a past semester presented in order of point deduction

Part A

-40 no report other than compilation of computer code
-40 no reported fitting 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

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. The report should be uploaded as a pdf file and submitted via the link for the first project assignment on Brightspace.
2. The report should be in a one single file. Both the one-page report for Part A and the one age report for Part B should be submitted in the same file.

Signs of Plagiarism in Your Report

1. Plagiarism is a serious issue. It is expected of you that the work 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.
3. 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.
4. Two students who submit the same report except for statistical results have engaged in plagiarism. The enabler (originator of the paper) is more guilty in my eyes than the plagiarizer.