**AMS 315**

**Data Analysis, Spring 2018**

**First Computing Assignment**

The first report is due on Thursday, March 29th  but can be submitted without penalty by April 5. 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 via e-mail as instructed on the Class Blackboard. Detailed submission information is online.

For the first project, there are two parts: A and B. You and your partner should receive a total of three files. Two of the files are for part A, and one file is for part B.

For part A, each file will contain a column for subject ID and a column for either the dependent variable value or the independent variable value. First, you and your partner are expected to sort the two files by subject ID and merge them. You are not encouraged to just “cut and paste” to merge your data. Second, you and your partner are expected to deal with entries with missing data using “listwise deletion”. Although there are more advanced methods used currently, the focus of this assignment is on data processing. For listwise deletion, you simply delete any line of data that has either missing IV or DV. Many statistical packages will do this for you automatically by choosing listwise delete as an option for missing data. Then, you and your partner are expected to use statistical packages to find the fitted linear model.

The data file for part B will contain 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. A lack of fit (LOF) test should be applied if there are repeated values in the data sets. It is your group’s responsibility to find repeated (or near repeated) independent variable values. That is, your group should bin near repeated data into one level. For example, suppose that and . While there are not exactly repeated *x* values, your group 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 and . Then perform a LOF test on the data set after binning all near repeated values.

Each group should 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 a lack of fit test, how much missing data was present in the data, and the procedure for dealing with missing data (here listwise deletion). 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. You may submit a longer appendix of computer work and programs.

***Important note:***

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

**Example Report**

Here is a sample report. Keep in mind, this is just a general idea of what should the first project looks like. ***You must not copy and paste it to submit as your report with the values of the numbers changed. Such activity is plagiarism and you will receive a grade of 0****.*

*Introduction*

The objective is to find the model describing the data in Problem A. A simulation program using an unknown linear function was used to generate the data.

*Methodology*

In order to solve problem A, we used the statistics package SPSS and Microsoft Excel spreadsheet program. The original data files were supplied with two data sheets in Excel. One data sheet had the ID of an observation and its associated independent variable value, and the other had the ID and associated dependent variable value. The independent variable data file had a total of 710 independent variable values with ID# ranging from 1 to 729. The dependent variable value had a total of 690 dependent variable values with ID # ranging from 1 to 730. We first sorted data in both files in ascending ID# order and then used Excel to merge the files. We next used listwise deletion to remove 40 entries that were missing either the independent variable value or the dependent variable value. Finally, we merged the two files into one file with three columns: ID, IV and DV. There were 670 entries with both values, with ID# ranging from 1 to 729. The data was then imported into SPSS. We assume linear regression for our data, but in order to find a better fit, we also transformed dependent variable into DV^2, Sqrt(DV) and independent variables into IV^2, Sqrt(IV), 1/IV, and ln(IV).

*Results*

The fitted function for the model Y= B+B1 X was DV=20.966IV+2123.719 with 99.9% fraction of variance was explained. The 95% confidence interval for the slope was [20.914 , 21.019]. The 95% confidence interval for the intercept was [2068.988 , 2178.450]. The analysis of variance table is shown below and the association between the independent variable and dependent variable was highly significant (p=0.000).

Table 1

Analysis of Variance Table

DV regressed on IV

(n=670)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | Df | Mean Square | F | Sig. |
| 1 | Regression | 25021381100.435 | 1 | 25021381100.435 | 617186.738 | .000b |
| Residual | 27081402.664 | 668 | 40541.022 |  |  |
| Total | 25048462503.099 | 669 |  |  |  |
| a. Dependent Variable: DV | | | | | | |
| b. Predictors: (Constant), IV | | | | | | |

*Conclusion*

For problem A, the association between independent variables and dependent variables was highly significant (p=0.000), with 99.9% of the dependent variable variation explained. The plot of residual versus predicted value confirmed the validity of this model.

Note: For question B, please report transformation you have performed and the model in transformed format.

End of Report