AMS 315
Data Analysis, Spring 2023
Multiple Regression Computing Project

Introduction

This assignment is due on Tuesday, May 2, 2023, at 11:59 pm Stony Brook time. Extensions will be granted, but remember the last class is May 4. The assignment is worth 250 points, and there is a 25 point bonus for on time submission.

Each student is assigned to an individual database, with a single file containing the data. Please go to the assignments section, which contains the data files. Your file is the one with the last six digits of your Stony Brook identification number. Each file contains one dependent variable and twenty-four independent variables. The values of the dependent variable are in the Y column (first column on the left). The values of the twenty-four independent variables are in the columns with names of E1 to E4 and G1 to G20. There should be no missing values; that is, the data file is complete and needs no further processing. This project is worth up to 200 points. Failure to use the correct dataset will lead to a grade of zero. Again, the data sets are identified by the last six digits of your Stony Brook University ID as a csv file. The datasets are posted in a zip format on the class blackboard.

Background

The class blackboard (Chapter 12 material) has a pdf file of a paper by Caspi et al. that reports a finding of a gene-environment interaction. This paper used multiple regression techniques as the methodology for its findings. You should read it for background, as it is the genesis of the models that you will be given. The data that you are analyzing is synthetic. That is, the TA used a model to generate the data. Your task is to find the model that the TA used for your data. For example, one possible model is

$$Y_i = (500 + 5E_{1i} + 25G_{2i} + 50E_{3i}G_{4i} + 100G_{5i}G_{6i} + 2Z_i)^2 .$$

The class blackboard also contains a paper by Risch et al. that uses a larger collection of data to assess the findings in Caspi et al. These researchers confirmed that Caspi et al. calculated their results correctly but that no other dataset had the relation reported in Caspi et al. That is, Caspi et al. seem to have reported a false positive (Type I error).

Report

The report that you submit should be no more than 2500 words with no more than 3 tables and 2 figures. It should include references (which do not count in the 2500 words). The report may have a technical appendix. The appendix could include your computer programs or describe your procedures for computation. You should include

whatever additional material you feel is necessary to report your results in the technical appendix. There are no length restrictions on the appendix. A submission of only computer output without a report is not sufficient and will receive a grade of zero. Analyses that report an incorrect number of observations will also receive a grade of zero.

Your report should be in standard scientific report format. It should contain an introduction, methods section, results section, and a section with conclusions and discussion. You may add whatever other material you wish in a technical appendix. The introduction should contain the statement of your problem (namely estimating the function that the TA used to generate your data). It should discuss the context of finding GxE interactions, as given by Caspi et al. and others. The methods section should discuss how you performed your statistical calculations, what independent variables that you considered, and other methodological issues, such as how you dealt with interaction variables. The results section should contain an objective statement of your findings. That is, it should contain the statement of the model that your group proposes for the data, the analysis of variance table for this model, and other key summary results. The discussion and conclusion section should include the limitations of your procedures. The class blackboard has an editorial (by Cummings, Reporting Statistical Information) that discusses reporting statistical information.

Guidelines for analysis

The first task for this problem is to use the statistical package of your choice to find the correlations between the independent variables and the dependent variable. Transformations of variables may be necessary. The Box-Cox transformation may find potentially nonlinear transformations of a dependent variable. After selecting the transformations of the dependent variable, you may use stepwise regression methods to select the important independent variables. The Lasso technique was helpful to many groups in past semesters. The TA will usually use at most two-way interactions of the independent variables (that is, terms like E_1G_2 or G_3G_4) in generating your data. There may also be non-linear environmental variables, such as E_3^2 or $E_4^{0.5}$. The TA may well have used three factor interactions in the models for a few of the groups.

Hints

Chapter 12 and Chapter 13 in your text contain important information, especially Chapter 12. Also remember to consider multiple testing issues (as described in Chapter 9). The p-value for the variables that you select should be much smaller than 0.01. Remember that you have 4 environmental variables, 20 gene indicator variables, 80 gene-environment variables, 190 gene-gene interaction variables, and a very large number of three gene interaction variables. The class blackboard has a handout describing one approach to analyzing a data set like the one in this assignment.

Your technical appendix may include:
(a) Your SAS or R script (If you are using SAS or R)
(b) Additional information that you want to report.
(c) Any comments or suggestions

End of Project Assignment