*AMS 578*
*Regression Analysis, Spring 2021*
*Multiple Regression Computing Project*

## Introduction

The final report is due on Thursday, May 6, 2021, the last day of class. A preliminary report on the data is due on Tuesday, April 20 and is worth 100 points. The data for the project is in three separate files. Each file name ends with six numeric characters. Your files are the ones with last six digits that are the same as the last six digits of your Stony Brook ID Number. Each student must analyze the correct data set. Failure to use the correct dataset will lead to a grade of zero.

One file contains the patient identifier and the dependent variable value. The second file contains the patient identifier and values of six environment variables called E1 to E6. The third file contains the patient identifier and the twenty five independent indicator variables called G1 to G20. The records may not be in correct order in each file, and cases may be missing in one or more of the files. You can process the data with VMLOOKUP or other data merging software.

## Preliminary Report

Your preliminary report (due April 120 should contain counts of missing data and summary statistics on each of your variables. These summary statistics for a variable before imputation should include at least the number of observations for that variable, the mean, median, standard deviation, lower quartile point, upper quartile point, minimum, maximum, and the number of missing values. The report should include your choice of methodology for dealing with missing data. You may not use listwise deletion, mean imputation, median imputation (or any other related technique). You may not delete "outliers." You should also report the correlation matrix of your variables.

## Background

The class blackboard 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

$$Y_i = (500 + 5E_{1i} + 25G_{2i} + 50E_{8i}G_{4i} + 100G_{5i}G_{6i}G_{15i}G_{20i} + 2Z_i)^2 .$$

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). The class blackboard contains a recent paper about the genetics of mental illness and a technical appendix giving the specifics. Together these papers are an example of the response of the research community to studying the genetics of mental illness, which is a notoriously difficult research area.

**Final Report**

The final report is worth 250 points. Your report should be in standard scientific report format and should be less than 2,500 words. 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 chose the model validation settings and what your model validation procedure was. 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) that discusses reporting statistical information. The report that your group submits 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. It should include your computer programs or describe your procedures for computation. Your group should include whatever additional material it feels is necessary to report your results. 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.

**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, use model building methods such as stepwise regression to select the important independent variables. The TA will use at most four-way interactions of the independent variables (that is, terms like $E_1E_2G_2G_{17}$ or

$G_3 G_4 G_{10} G_{19}$ ) in generating your data. There may also be non-linear environmental variables, such as $E_3^2$ or $E_4^{0.5}$.

**Hints**

Remember to consider multiple testing issues. The p-value for the variables that you select should be much smaller than 0.01. Remember that you have 6 environmental variables, 25 genes, 150 gene-environment variables, 300 gene-gene interaction variables, and so on.

End of Project Assignment