AMS 578 Spring 2021
Multiple Regression Computing Project
Preliminary Report

Kai Li*

Stony Brook University — April 19, 2021

1 Introduction

Depression is one of the top five leading causes of disability and disease burden worldwide [3]. Researchers such as Caspi et al. [3] and Risch et al. [9] have already conducted meta-analyses of the interaction between the serotonin transporter gene (5-HTTLPR) and stressful life events on the risk of depression using regression techniques. This multiple regression computing project aims to analyze a given synthetic dataset to fit a model using statistical software `R` given the background of related studies.

There are a few steps to perform a complete regression analysis. The first step is data summarizing and cleaning. Then, since variables can be nonlinear, checking if transformations are required is a key to ensure proper analysis and conclusion. An examination of important independent variables for the model becomes the following procedure. Especially, given a lack of consensus from the background, gene-by-environment and gene-by-gene interaction require additional inspection [3, 9]. It is also given that at most four-way interactions of independent variables will appear. Finally, compare and determine if regression results are viable.

In this preliminary report, a summary statistics table will be provided as the first step of the analysis. Then, it is necessary to explain and use a methodology to deal with missing values if they exist in the data. Finally, a check on multicollinearity after coping with missing values enables to show if multicollinearity exists between variables. A complete model selection and analysis procedure will appear on the final report.

2 Summary Statistics

To better understand and interpret a set of numerical data, it is easy to summarize the data by a few statistics representing its major characteristics, such as measures of location and dispersion [10]. Table 1 includes nine basic statistics for each variable. Note that there are 32 variables. Y is the quantitative independent variable measuring depressogenic effect; six environmental variables, denoted E1 through E6, quantify stressful life events; R1 through R25 are binary gene variables that are the candidate genes for depression from the serotonin system [3, 9]. There are either 20 or 30 missing values for some of the variables in the dataset, about 1.08% and 1.62%, respectively, of the total observations. Hence, an analysis of the missing data is needed.

*Department of Applied Mathematics and Statistics, Stony Brook University, email: kai.li@stonybrook.edu

Table 1: Summary Statistics Table

Variable	n	Min	Q_1	$\widetilde{x}$	$\bar{x}$	Q_3	Max	s	#NA
Y	1848	6.11e+09	1.24e+10	1.44e+10	1.50e+10	1.66e+10	2.95e+10	3.86e+09	30
E1	1848	551.82	954.73	1072.67	1064.29	1174.55	1692.26	163.99	20
E2	1848	311.29	677.75	782.26	785.66	889.94	1298.12	158.01	30
E3	1848	215.87	576.02	683.49	683.07	785.42	1190.05	157.05	20
E4	1848	-24.07	415.54	518.60	515.53	617.49	1053.83	158.21	20
E5	1848	223.00	710.97	822.30	822.87	932.60	1330.38	164.12	30
E6	1848	42.48	518.15	622.52	620.51	723.58	1225.55	159.18	20
R1	1848	0.00	0.00	0.00	0.49	1.00	1.00	0.50	30
R2	1848	0.00	0.00	1.00	0.52	1.00	1.00	0.50	30
R3	1848	0.00	0.00	1.00	0.52	1.00	1.00	0.50	30
R4	1848	0.00	0.00	1.00	0.52	1.00	1.00	0.50	30
R5	1848	0.00	0.00	1.00	0.51	1.00	1.00	0.50	0
R6	1848	0.00	0.00	1.00	0.51	1.00	1.00	0.50	0
R7	1848	0.00	0.00	1.00	0.51	1.00	1.00	0.50	0
R8	1848	0.00	0.00	0.00	0.50	1.00	1.00	0.50	0
R9	1848	0.00	0.00	1.00	0.51	1.00	1.00	0.50	0
R10	1848	0.00	0.00	0.00	0.50	1.00	1.00	0.50	0
R11	1848	0.00	0.00	0.00	0.50	1.00	1.00	0.50	30
R12	1818	0.00	0.00	0.00	0.50	1.00	1.00	0.50	30
R13	1848	0.00	0.00	0.00	0.50	1.00	1.00	0.50	0
R14	1848	0.00	0.00	0.00	0.48	1.00	1.00	0.50	0
R15	1848	0.00	0.00	0.00	0.49	1.00	1.00	0.50	0
R16	1848	0.00	0.00	1.00	0.51	1.00	1.00	0.50	0
R17	1848	0.00	0.00	0.00	0.50	1.00	1.00	0.50	30
R18	1848	0.00	0.00	1.00	0.51	1.00	1.00	0.50	0
R19	1848	0.00	0.00	1.00	0.52	1.00	1.00	0.50	0
R20	1848	0.00	0.00	1.00	0.50	1.00	1.00	0.50	0
R21	1848	0.00	0.00	1.00	0.51	1.00	1.00	0.50	30
R22	1848	0.00	0.00	1.00	0.50	1.00	1.00	0.50	0
R23	1848	0.00	0.00	0.00	0.49	1.00	1.00	0.50	30
R24	1848	0.00	0.00	1.00	0.50	1.00	1.00	0.50	0
R25	1848	0.00	0.00	1.00	0.50	1.00	1.00	0.50	30

3 Missing Data Methodology

In this section, a methodology to deal with the missing data will be discussed. Before that, the pattern of the missing values needs to be identified. In general, there are three types of missing data based on the patterns of missingness. Missing completely at random (MCAR) is defined as when the data missing probability is the same for all cases; missing at random (MAR) is defined as when the data missing probability is the same within the observed data alone; missing not at random (MNAR) is defined to be the data which is not MCAR or MAR [2, 8]. There are available tools to visualize missing pattern. One is called the missing value specification plot, shown in Figure 1. No systematic pattern of the missing values is observed from the plot. That is, the plot describes a general pattern with no specific structure, and hence the MCAR/MAR assumption is more likely to be true. Hypothesis testing can also be used to check for associations between missing and observed data. The null hypothesis is that the missing data is MCAR/MAR, versus the alternative hypothesis that it is MNAR. As shown in Table 2, at the 0.01 level of significance, the test result is not significant, with Y being the dependent variable, environmental and gene as independent variables. In this project, the missing data will be considered MAR instead of MCAR because, for robustness, MCAR is an ideal but unreasonable assumption [4].

After diagnosing the missing value pattern, it is appropriate to find a solution to deal with them. In many existing fields, multiple imputation is now accepted as the best general method to deal with datasets with missing values [2]. One significant advantage of multiple imputation, as opposed to single imputation, is that it preserves the natural variability of the missing data and

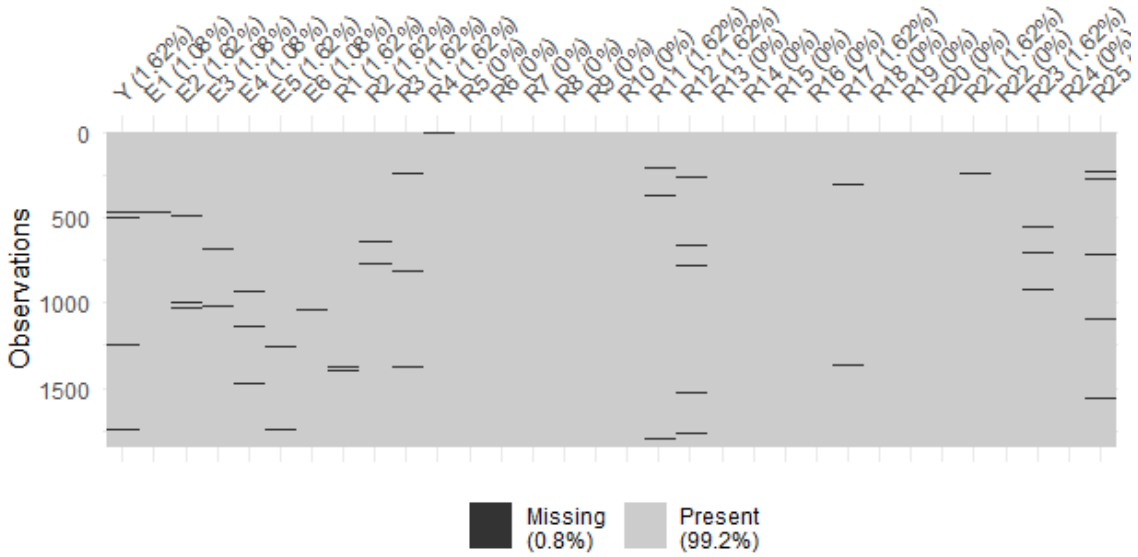

Figure 1: Missing Value Specification Plot

incorporates the uncertainty due to the nature of MAR, which enables to perform a valid statistical inference [4]. For this project, multivariate imputation by chained equations (MICE), a particular type of multiple imputation method, will be selected to impute the missing data. Because the dataset contains both continuous and binary variables, the chained equations approach is utilized due to its flexibility with variable types [1].

A package in `R` called `mice` corresponds directly to MICE techniques. The number of imputations created and the number of iterations for each imputation should be considered to use the method properly. Buuren [2] mentions that multiple imputation can generate unbiased estimates and correct confidence intervals with the number of imputed datasets as small as two. Raghunathan et al. [7] suggest 10 iterations per imputation. Given the size of the dataset, the amount of missing information, and the available computer resources, two imputed datasets and ten iterations for each will be outputted in this report [1].

4 Multicollinearity Diagnostics

Multiple regression models are used for a wide variety of applications. One common serious issue in regression analysis that may lead to a considerable decrease in the usefulness of a model is multicollinearity or near-linear dependence among regression variables [6]. Currently, there are techniques for detecting multicollinearity.

One informal measure of multicollinearity is the examination of the correlation matrix. Large coefficients of simple correlation coefficients between pairs of independent variables indicate a presence of serious multicollinearity [5]. That is, if the absolute value of a non-diagonal element, also called pairwise correlation, is close to 1, a nearly linearly dependent relationship between regressors is detected [6]. The correlation matrices for the two imputed datasets are shown in Tables 3 and 4. All non-diagonal absolute pairwise correlation coefficients are less than or equal to 0.07, suggesting that the variables are not nearly linearly dependent.

Montgomery et al. [6] also claim that when more than two independent variables are involved in a near-linear dependence relationship, it is not guaranteed that any of the pairwise correlations will be large. Thus, a formal way to diagnose multicollinearity is using variance inflation factor

(VIF). If the largest VIF value among the regressors exceeds 10, it indicates multicollinearity, which can unduly influence model estimates [5]. The maximum VIFs in both datasets are 1.025 and 1.026, respectively, which suggests no multicollinearity issue in both imputed datasets.

A Technical Implementation in R

In the technical appendix, code implementation will be provided to show how results are obtained. At the beginning of the project, three files containing the unmerged datasets with variables `ID`, `Y`, `E`, and `R` are provided. The initial dataset is generated by combining the three files based on variable `ID`. Function `merge()` in `R` will work. Because there are numerous existing software/methods to merge datasets, the code for merging is omitted. The focus is the implementation of the combined data. The combined data is named variable `data` in the code below.

A.1 Summary Statistics

Table 1 is created using package `reporttools`, which directly outputs the summary statistics table in LaTeX. Parts of summary statistics are omitted for simplicity purposes.

project.r

```
library(reporttools)
vars0 <- with(data, data.frame(data[, -1]))
tableContinuous(vars = vars0, prec = 2, longtable = FALSE)
```

A.2 Missing Value Specification Plot

Figure 1 is a plot generated by package `naniar`. It provides functions to facilitate the plottings of missing values. `vis_miss()` returns the missing value specification plot.

project.r

```
library(naniar)
vis_miss(data)
```

A.3 Missing Data Pattern Analysis

`finalfit` includes functions to ensure missing data is correctly identified. `missing_compare()` performs the hypothesis testing of whether the missing data follows an MCAR/MAR or MNAR by comparing missing data in the dependent variable across explanatory variables. Table 2 shows the p-values in the return output.

project.r

```
library(finalfit)
missing_compare(data, "Y", colnames(data)[-1])
```

Table 2: Missing Data Hypothesis Testing: Y

Variable		Not missing	Missing	p	Variable		Not missing	Missing	p
E1	$\bar{x}$ (s)	1063.2 (162.9)	1134.3 (214.7)	0.023	E4	$\bar{x}$ (s)	515.1 (158.6)	541.6 (131.2)	0.371
E2	$\bar{x}$ (s)	785.4 (157.8)	801.0 (169.9)	0.591	E5	$\bar{x}$ (s)	823.0 (164.2)	815.7 (160.3)	0.816
E3	$\bar{x}$ (s)	683.6 (156.6)	650.3 (183.6)	0.257	E6	$\bar{x}$ (s)	620.5 (159.3)	622.2 (153.2)	0.954
R1	0	910 (98.5)	14 (1.5)	0.929	R14	0	936 (98.3)	16 (1.7)	0.987
	1	879 (98.3)	15 (1.7)			1	882 (98.4)	14 (1.6)	
R2	0	855 (98.5)	13 (1.5)	0.762	R15	0	937 (98.5)	14 (1.5)	0.730
	1	933 (98.2)	17 (1.8)			1	881 (98.2)	16 (1.8)	
R3	0	851 (98.4)	14 (1.6)	1.000	R16	0	893 (98.0)	18 (2.0)	0.318
	1	937 (98.3)	16 (1.7)			1	925 (98.7)	12 (1.3)	
R4	0	862 (98.6)	12 (1.4)	0.479	R17	0	900 (98.7)	12 (1.3)	0.443
	1	926 (98.1)	18 (1.9)			1	889 (98.1)	17 (1.9)	
R5	0	885 (98.1)	17 (1.9)	0.494	R18	0	886 (97.9)	19 (2.1)	0.161
	1	933 (98.6)	13 (1.4)			1	932 (98.8)	11 (1.2)	
R6	0	893 (98.1)	17 (1.9)	0.525	R19	0	870 (98.2)	16 (1.8)	0.681
	1	925 (98.6)	13 (1.4)			1	948 (98.5)	14 (1.5)	
R7	0	890 (98.7)	12 (1.3)	0.430	R20	0	905 (98.4)	15 (1.6)	1.000
	1	928 (98.1)	18 (1.9)			1	913 (98.4)	15 (1.6)	
R8	0	908 (98.2)	17 (1.8)	0.585	R21	0	877 (98.7)	12 (1.3)	0.424
	1	910 (98.6)	13 (1.4)			1	911 (98.1)	18 (1.9)	
R9	0	898 (98.6)	13 (1.4)	0.635	R22	0	902 (98.3)	16 (1.7)	0.826
	1	920 (98.2)	17 (1.8)			1	916 (98.5)	14 (1.5)	
R10	0	906 (97.9)	19 (2.1)	0.200	R23	0	914 (98.5)	14 (1.5)	0.910
	1	912 (98.8)	11 (1.2)			1	875 (98.3)	15 (1.7)	
R11	0	903 (98.6)	13 (1.4)	0.552	R24	0	906 (98.2)	17 (1.8)	0.577
	1	885 (98.1)	17 (1.9)			1	912 (98.6)	13 (1.4)	
R12	0	896 (98.1)	17 (1.9)	0.469	R25	0	886 (98.0)	18 (2.0)	0.342
	1	893 (98.7)	12 (1.3)			1	902 (98.7)	12 (1.3)	
R13	0	919 (98.5)	14 (1.5)	0.812					
	1	899 (98.3)	16 (1.7)						

A.4 Multiple Imputation

Package `mice`, written by Karin Groothuis-Oudshoorn and Van Buuren, can perform multiple imputation computations [2]. The imputation method selected for each dataset is classification and regression trees (CART). Function `mice()` generates multiple imputation datasets, and function `complete()` returns the complete data in a specified format. After that, each complete dataset is ready to perform analysis.

project.r

```
library(mice)
imp <- mice(data, method = "cart", m = 2, maxit = 10,
            seed = 123, print = FALSE)
data1 <- complete(imp, 1)
data2 <- complete(imp, 2)
```

A.5 Multicollinearity Diagnostics

Correlation matrices in Tables 3 and 4 can be obtained using function `cor()`. The correlation coefficients are rounded to 2 decimals to reduce the unnecessary digits. VIFs can be calculated under package `car` for linear models. Function `vif()` outputs VIF values for all variables in linear regression model `lm()`. Because only the maximums are considered, return using function `max()` is sufficient.

Table 3: Correlation Matrix for Dataset 1

	E1	E2	E3	E4	E5	E6	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20	R21	R22	R23	R24	R25
E1	1.00	-0.01	0.00	0.00	0.02	-0.01	0.04	0.03	0.02	-0.04	0.01	0.01	-0.00	-0.01	-0.00	-0.03	-0.00	0.01	-0.01	-0.01	0.00	-0.02	-0.01	0.01	-0.01	0.01	0.03	-0.02	0.05	0.00	-0.01
E2	-0.01	1.00	0.05	0.01	-0.03	-0.01	0.00	0.02	-0.01	-0.05	-0.02	0.02	0.01	-0.02	-0.02	-0.01	0.01	0.01	-0.01	-0.02	0.02	0.00	0.00	-0.01	0.01	-0.01	-0.02	0.03	-0.00	-0.01	0.04
E3	0.00	0.05	1.00	-0.03	-0.01	-0.02	-0.00	-0.01	0.02	-0.02	0.02	-0.01	-0.00	0.00	-0.00	0.02	-0.01	0.04	0.03	0.03	-0.00	0.04	0.01	0.03	-0.02	-0.01	0.02	0.00	0.00	0.00	0.01
E4	0.00	0.01	-0.03	1.00	0.01	-0.01	0.02	0.01	0.02	-0.03	-0.04	-0.01	0.00	0.00	-0.01	0.02	0.04	-0.04	-0.05	0.00	-0.02	-0.02	0.01	-0.01	0.03	0.03	-0.00	0.01	0.02	0.01	0.01
E5	0.02	-0.03	-0.01	0.01	1.00	0.00	0.02	0.04	0.04	0.01	0.02	0.01	-0.01	-0.04	0.02	-0.01	-0.06	0.03	0.06	0.01	0.02	0.02	-0.04	-0.04	-0.03	-0.01	0.01	-0.00	0.01	0.00	0.06
E6	-0.01	-0.01	-0.02	-0.01	0.00	1.00	-0.04	-0.01	0.01	-0.03	0.01	-0.01	0.01	-0.01	0.03	0.03	0.02	-0.01	0.00	0.00	-0.03	-0.00	-0.02	-0.06	-0.03	0.04	0.01	-0.02	-0.01	-0.06	-0.03
R1	0.04	0.00	-0.00	0.02	0.02	-0.04	1.00	0.06	-0.02	0.01	0.01	0.00	0.03	-0.02	0.04	0.04	0.03	0.01	0.03	-0.01	-0.03	-0.03	0.04	0.00	0.04	-0.01	0.01	0.01	0.01	0.02	-0.01
R2	0.03	0.02	-0.01	0.01	0.04	-0.01	0.06	1.00	-0.01	0.04	-0.01	0.01	0.01	-0.02	-0.00	0.04	0.04	0.04	0.01	-0.06	-0.00	-0.03	0.01	-0.04	0.04	-0.00	-0.00	-0.02	0.03	0.01	-0.00
R3	0.02	-0.01	0.02	0.02	0.04	0.01	-0.02	-0.01	1.00	0.04	0.01	-0.00	0.01	-0.03	-0.05	-0.05	-0.00	-0.02	-0.03	-0.01	-0.03	-0.01	0.02	0.01	-0.06	0.03	0.01	0.02	0.01	0.02	-0.01
R4	-0.04	-0.05	-0.02	-0.03	0.01	-0.03	0.01	0.04	0.04	1.00	-0.01	-0.00	-0.01	0.03	-0.00	0.03	-0.03	-0.01	-0.03	0.00	-0.01	0.01	0.01	-0.00	0.02	0.01	-0.01	-0.02	0.01	0.01	0.01
R5	0.01	-0.02	0.02	-0.04	0.02	0.01	0.01	-0.01	0.01	-0.01	1.00	0.00	0.02	0.01	-0.01	-0.01	-0.04	-0.02	-0.01	0.04	0.00	-0.01	-0.04	-0.03	0.01	-0.01	0.01	-0.02	-0.02	0.02	0.04
R6	0.01	0.02	-0.01	-0.01	0.01	-0.01	0.00	0.01	-0.00	-0.00	0.00	1.00	0.02	-0.01	-0.00	0.03	0.01	0.02	0.04	-0.03	0.01	0.02	-0.01	0.01	0.00	-0.02	-0.02	-0.01	-0.01	0.02	-0.03
R7	-0.00	0.01	-0.00	0.00	-0.01	0.01	0.03	0.01	0.01	-0.01	0.02	0.02	1.00	0.01	0.01	0.04	-0.01	-0.01	-0.00	0.01	-0.05	0.00	-0.01	-0.02	-0.01	-0.02	0.00	0.03	-0.01	-0.01	-0.01
R8	-0.01	-0.02	0.00	0.00	-0.04	-0.01	-0.02	-0.02	-0.03	0.03	0.01	-0.01	0.01	1.00	-0.03	0.00	0.05	0.02	0.01	-0.01	-0.01	0.05	0.01	0.01	-0.02	-0.02	-0.00	-0.00	-0.01	-0.04	-0.04
R9	-0.00	-0.02	-0.00	-0.01	0.02	0.03	0.04	-0.00	-0.05	-0.00	-0.01	-0.00	0.01	-0.03	1.00	0.06	0.00	0.01	-0.01	-0.01	-0.00	-0.04	0.05	0.01	-0.00	0.01	0.01	0.00	-0.00	-0.01	0.02
R10	-0.03	-0.01	0.02	0.02	-0.01	0.03	0.04	0.04	-0.05	0.03	-0.01	0.03	0.04	0.00	0.06	1.00	0.02	-0.04	-0.02	0.02	-0.02	-0.02	0.02	-0.00	0.05	-0.02	-0.01	0.01	-0.01	-0.02	0.03
R11	-0.00	0.01	-0.01	0.04	-0.06	0.02	0.03	0.04	-0.00	-0.03	-0.04	0.01	-0.01	0.05	0.00	0.02	1.00	-0.02	-0.04	0.02	-0.02	-0.01	0.01	0.01	-0.01	-0.02	0.06	-0.03	-0.00	-0.02	-0.02
R12	0.01	0.01	0.04	-0.04	0.03	-0.01	0.01	0.04	-0.02	-0.01	-0.02	0.02	-0.01	0.02	0.01	-0.04	-0.02	1.00	0.01	-0.00	0.01	-0.02	-0.01	0.02	0.01	-0.01	-0.01	0.02	-0.03	0.01	-0.00
R13	-0.01	-0.01	0.03	-0.05	0.06	0.00	0.03	0.01	-0.03	-0.03	-0.01	0.04	-0.00	0.01	-0.01	-0.02	-0.04	0.01	1.00	0.00	-0.01	0.04	-0.01	-0.06	0.00	-0.01	-0.02	0.02	-0.02	-0.03	0.00
R14	-0.01	-0.02	0.03	0.00	0.01	0.00	-0.01	-0.06	-0.01	0.00	0.04	-0.03	0.01	-0.01	-0.01	0.02	0.02	-0.00	0.00	1.00	0.01	0.02	-0.00	0.01	-0.00	-0.03	0.00	0.04	0.03	0.01	-0.00
R15	0.00	0.02	-0.00	-0.02	0.02	-0.03	-0.03	-0.00	-0.03	-0.01	0.00	0.01	-0.05	-0.01	-0.00	-0.02	-0.02	0.01	-0.01	0.01	1.00	0.01	0.01	0.01	0.00	0.02	-0.01	-0.02	0.05	0.01	0.01
R16	-0.02	0.00	0.04	-0.02	0.02	-0.00	-0.03	-0.03	-0.01	0.01	-0.01	0.02	0.00	0.05	-0.04	-0.02	-0.01	-0.02	0.04	0.02	0.01	1.00	0.04	-0.03	-0.01	0.03	-0.03	-0.02	0.02	-0.05	-0.00
R17	-0.01	0.00	0.01	0.01	-0.04	-0.02	0.04	0.01	0.02	0.01	-0.04	-0.01	-0.01	0.01	0.05	0.02	0.01	-0.01	-0.01	-0.00	0.01	0.04	1.00	-0.01	0.00	0.00	-0.01	0.00	0.00	0.01	0.06
R18	0.01	-0.01	0.03	-0.01	-0.04	-0.06	0.00	-0.04	0.01	-0.00	-0.03	0.01	-0.02	0.01	0.01	-0.00	0.01	0.02	-0.06	0.01	0.01	-0.03	-0.01	1.00	0.05	0.04	0.03	-0.01	0.02	-0.01	-0.03
R19	-0.01	0.01	-0.02	0.03	-0.03	-0.03	0.04	0.04	-0.06	0.02	0.01	0.00	-0.01	-0.02	-0.00	0.05	-0.01	0.01	0.00	-0.00	0.00	-0.01	0.00	0.05	1.00	0.00	-0.01	-0.00	0.00	-0.03	-0.03
R20	0.01	-0.01	-0.01	0.03	-0.01	0.04	-0.01	-0.00	0.03	0.01	-0.01	-0.02	-0.02	-0.02	0.01	-0.02	-0.02	-0.01	-0.01	-0.03	0.02	0.03	0.00	0.04	0.00	1.00	0.06	-0.03	-0.04	-0.01	0.01
R21	0.03	-0.02	0.02	-0.00	0.01	0.01	0.01	-0.00	0.01	-0.01	0.01	-0.02	0.00	-0.00	0.01	-0.01	0.06	-0.01	-0.02	0.00	-0.01	-0.03	-0.01	0.03	-0.01	0.06	1.00	-0.01	-0.02	0.01	-0.00
R22	-0.02	0.03	0.00	0.01	-0.00	-0.02	0.01	-0.02	0.02	-0.02	-0.02	-0.01	0.03	-0.00	0.00	0.01	-0.03	0.02	0.02	0.04	-0.02	-0.02	0.00	-0.01	-0.00	-0.03	-0.01	1.00	-0.01	-0.03	0.03
R23	0.05	-0.00	0.00	0.02	0.01	-0.01	0.01	0.03	0.01	0.01	-0.02	-0.01	-0.01	-0.01	-0.00	-0.01	-0.00	-0.03	-0.02	0.03	0.05	0.02	0.00	0.02	0.00	-0.04	-0.02	-0.01	1.00	0.02	0.04
R24	0.00	-0.01	0.00	0.01	0.00	-0.06	0.02	0.01	0.02	0.01	0.02	0.02	-0.01	-0.04	-0.01	-0.02	-0.02	0.01	-0.03	0.01	0.01	-0.05	0.01	-0.01	-0.03	-0.01	0.01	-0.03	0.02	1.00	0.02
R25	-0.01	0.04	0.01	0.01	0.06	-0.03	-0.01	-0.00	-0.01	0.01	0.04	-0.03	-0.01	-0.04	0.02	0.03	-0.02	-0.00	0.00	-0.00	0.01	-0.00	0.06	-0.03	-0.03	0.01	-0.00	0.03	0.04	0.02	1.00

Table 4: Correlation Matrix for Dataset 2

	E1	E2	E3	E4	E5	E6	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R15	R16	R17	R18	R19	R20	R21	R22	R23	R24	R25
E1	1.00	-0.02	-0.00	0.00	0.02	-0.01	0.03	0.03	0.03	-0.04	0.01	0.01	-0.00	-0.01	-0.00	-0.03	-0.00	0.01	-0.01	-0.01	0.01	-0.02	-0.01	0.01	-0.01	0.01	0.03	-0.02	0.06	0.00	-0.01
E2	-0.02	1.00	0.05	0.01	-0.02	-0.01	-0.00	0.02	-0.02	-0.05	-0.01	0.01	0.01	-0.02	-0.01	-0.02	0.00	0.02	-0.00	-0.02	0.02	0.01	-0.01	-0.01	0.01	-0.00	-0.01	0.03	-0.01	-0.01	0.04
E3	-0.00	0.05	1.00	-0.03	-0.01	-0.03	0.00	-0.01	0.03	-0.03	0.02	-0.01	-0.00	-0.00	0.00	0.02	-0.00	0.05	0.03	0.03	0.00	0.05	0.02	0.03	-0.02	-0.01	0.01	0.00	-0.00	0.00	0.01
E4	0.00	0.01	-0.03	1.00	0.01	-0.01	0.02	0.01	0.02	-0.02	-0.04	-0.01	0.00	-0.00	-0.01	0.02	0.03	-0.04	-0.05	0.01	-0.02	-0.02	0.01	-0.01	0.03	0.03	-0.01	0.01	0.03	0.01	0.01
E5	0.02	-0.02	-0.01	0.01	1.00	-0.01	0.02	0.03	0.05	0.01	0.03	0.01	-0.01	-0.04	0.02	-0.01	-0.06	0.03	0.06	0.01	0.02	0.02	-0.04	-0.04	-0.03	-0.01	0.02	-0.00	0.00	0.01	0.07
E6	-0.01	-0.01	-0.03	-0.01	-0.01	1.00	-0.05	-0.01	0.01	-0.04	0.01	-0.01	-0.00	-0.01	0.02	0.03	0.02	-0.01	0.00	0.01	-0.03	-0.00	-0.02	-0.06	-0.03	0.04	0.01	-0.03	-0.02	-0.05	-0.02
R1	0.03	-0.00	0.00	0.02	0.02	-0.05	1.00	0.06	-0.02	0.01	0.02	-0.00	0.02	-0.02	0.05	0.04	0.04	0.02	0.02	-0.01	-0.03	-0.03	0.04	0.00	0.04	-0.01	0.02	0.01	-0.01	0.02	-0.02
R2	0.03	0.02	-0.01	0.01	0.03	-0.01	0.06	1.00	-0.00	0.05	-0.00	0.01	0.01	-0.02	-0.01	0.03	0.04	0.05	0.01	-0.05	-0.00	-0.03	0.01	-0.05	0.05	-0.01	-0.00	-0.01	0.03	0.00	0.01
R3	0.03	-0.02	0.03	0.02	0.05	0.01	-0.02	-0.00	1.00	0.04	0.01	-0.01	0.01	-0.03	-0.05	-0.05	-0.00	-0.02	-0.03	0.00	-0.03	-0.01	0.01	0.01	-0.05	0.02	0.00	0.02	0.01	0.03	-0.02
R4	-0.04	-0.05	-0.03	-0.02	0.01	-0.04	0.01	0.05	0.04	1.00	-0.01	-0.00	-0.01	0.04	0.01	0.04	-0.03	-0.01	-0.03	-0.00	-0.02	0.01	0.01	0.01	0.02	0.01	-0.01	-0.02	0.01	0.01	0.01
R5	0.01	-0.01	0.02	-0.04	0.03	0.01	0.02	-0.00	0.01	-0.01	1.00	0.00	0.02	0.01	-0.01	-0.01	-0.05	-0.02	-0.01	0.04	0.00	-0.01	-0.05	-0.03	0.01	-0.01	0.01	-0.02	-0.02	0.02	0.05
R6	0.01	0.01	-0.01	-0.01	0.01	-0.01	-0.00	0.01	-0.01	-0.00	0.00	1.00	0.02	-0.01	-0.00	0.03	0.00	0.02	0.04	-0.03	0.01	0.02	-0.00	0.01	0.00	-0.02	-0.02	-0.01	-0.01	0.02	-0.03
R7	-0.00	0.01	-0.00	0.00	-0.01	-0.00	0.02	0.01	0.01	-0.01	0.02	0.02	1.00	0.01	0.01	0.04	-0.01	-0.01	-0.00	0.01	-0.05	0.00	-0.01	-0.02	-0.01	-0.02	-0.00	0.03	-0.01	-0.01	-0.01
R8	-0.01	-0.02	-0.00	-0.00	-0.04	-0.01	-0.02	-0.02	-0.03	0.04	0.01	-0.01	0.01	1.00	-0.03	0.00	0.05	0.02	0.01	-0.01	-0.01	0.05	0.01	0.01	-0.02	-0.02	-0.01	-0.00	-0.00	-0.04	-0.04
R9	-0.00	-0.01	0.00	-0.01	0.02	0.02	0.05	-0.01	-0.05	0.01	-0.01	-0.00	0.01	-0.03	1.00	0.06	0.00	0.00	-0.01	-0.01	-0.00	-0.04	0.05	0.01	-0.00	0.01	0.01	0.00	-0.01	-0.01	0.01
R10	-0.03	-0.02	0.02	0.02	-0.01	0.03	0.04	0.03	-0.05	0.04	-0.01	0.03	0.04	0.00	0.06	1.00	0.01	-0.04	-0.02	0.02	-0.02	-0.02	0.01	-0.00	0.05	-0.02	-0.01	0.01	-0.01	-0.02	0.03
R11	-0.00	0.00	-0.00	0.03	-0.06	0.02	0.04	0.04	-0.00	-0.03	-0.05	0.00	-0.01	0.05	0.00	0.01	1.00	-0.01	-0.04	0.01	-0.02	-0.00	0.00	0.00	-0.02	-0.02	0.06	-0.02	-0.00	-0.02	-0.01
R12	0.01	0.02	0.05	-0.04	0.03	-0.01	0.02	0.05	-0.02	-0.01	-0.02	0.02	-0.01	0.02	0.00	-0.04	-0.01	1.00	0.01	0.00	0.00	-0.02	-0.01	0.01	0.01	-0.02	-0.00	0.01	-0.03	0.01	-0.00
R13	-0.01	-0.00	0.03	-0.05	0.06	0.00	0.02	0.01	-0.03	-0.03	-0.01	0.04	-0.00	0.01	-0.01	-0.02	-0.04	0.01	1.00	0.00	-0.01	0.04	-0.02	-0.06	0.00	-0.01	-0.02	0.02	-0.02	-0.03	0.00
R14	-0.01	-0.02	0.03	0.01	0.01	0.01	-0.01	-0.05	0.00	-0.00	0.04	-0.03	0.01	-0.01	-0.01	0.02	0.01	0.00	0.00	1.00	0.01	0.02	0.00	0.01	-0.00	-0.03	-0.00	0.04	0.03	0.01	-0.01
R15	0.01	0.02	0.00	-0.02	0.02	-0.03	-0.03	-0.00	-0.03	-0.02	0.00	0.01	-0.05	-0.01	-0.00	-0.02	-0.02	0.00	-0.01	0.01	1.00	0.01	0.00	0.01	0.00	0.02	-0.01	-0.02	0.05	0.01	0.01
R16	-0.02	0.01	0.05	-0.02	0.02	-0.00	-0.03	-0.03	-0.01	0.01	-0.01	0.02	0.00	0.05	-0.04	-0.02	-0.00	-0.02	0.04	0.02	0.01	1.00	0.04	-0.03	-0.01	0.03	-0.03	-0.02	0.02	-0.05	0.00
R17	-0.01	-0.01	0.02	0.01	-0.04	-0.02	0.04	0.01	0.01	0.01	-0.05	-0.00	-0.01	0.01	0.05	0.01	0.00	-0.01	-0.02	0.00	0.00	0.04	1.00	-0.01	0.01	0.01	-0.02	0.00	0.00	0.01	0.05
R18	0.01	-0.01	0.03	-0.01	-0.04	-0.06	0.00	-0.05	0.01	0.01	-0.03	0.01	-0.02	0.01	0.01	-0.00	0.00	0.01	-0.06	0.01	0.01	-0.03	-0.01	1.00	0.05	0.04	0.03	-0.01	0.02	-0.01	-0.04
R19	-0.01	0.01	-0.02	0.03	-0.03	-0.03	0.04	0.05	-0.05	0.02	0.01	0.00	-0.01	-0.02	-0.00	0.05	-0.02	0.01	0.00	-0.00	0.00	-0.01	0.01	0.05	1.00	0.00	-0.00	-0.00	0.01	-0.03	-0.02
R20	0.01	-0.00	-0.01	0.03	-0.01	0.04	-0.01	-0.01	0.02	0.01	-0.01	-0.02	-0.02	-0.02	0.01	-0.02	-0.02	-0.02	-0.01	-0.03	0.02	0.03	0.01	0.04	0.00	1.00	0.06	-0.03	-0.04	-0.01	0.00
R21	0.03	-0.01	0.01	-0.01	0.02	0.01	0.02	-0.00	0.00	-0.01	0.01	-0.02	-0.00	-0.01	0.01	-0.01	0.06	-0.00	-0.02	-0.00	-0.01	-0.03	-0.02	0.03	-0.00	0.06	1.00	-0.01	-0.02	0.01	-0.01
R22	-0.02	0.03	0.00	0.01	-0.00	-0.03	0.01	-0.01	0.02	-0.02	-0.02	-0.01	0.03	-0.00	0.00	0.01	-0.02	0.01	0.02	0.04	-0.02	-0.02	0.00	-0.01	-0.00	-0.03	-0.01	1.00	-0.01	-0.03	0.03
R23	0.06	-0.01	-0.00	0.03	0.00	-0.02	-0.01	0.03	0.01	0.01	-0.02	-0.01	-0.01	-0.00	-0.01	-0.01	-0.00	-0.03	-0.02	0.03	0.05	0.02	0.00	0.02	0.01	-0.04	-0.02	-0.01	1.00	0.02	0.05
R24	0.00	-0.01	0.00	0.01	0.01	-0.05	0.02	0.00	0.03	0.01	0.02	0.02	-0.01	-0.04	-0.01	-0.02	-0.02	0.01	-0.03	0.01	0.01	-0.05	0.01	-0.01	-0.03	-0.01	0.01	-0.03	0.02	1.00	0.03
R25	-0.01	0.04	0.01	0.01	0.07	-0.02	-0.02	0.01	-0.02	0.01	0.05	-0.03	-0.01	-0.04	0.01	0.03	-0.01	-0.00	0.00	-0.01	0.01	0.00	0.05	-0.04	-0.02	0.00	-0.01	0.03	0.05	0.03	1.00

project.r

```
cor_mat1 <- round(cor(data1[-1, -1]), 2)
cor_mat2 <- round(cor(data2[-1, -1]), 2)

library(car)
max(vif(lm(Y ~ ., data = data1)))
max(vif(lm(Y ~ ., data = data2)))
```

References

[1] M. J. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1):40–49, 3 2011.

[2] S. V. Buuren. *Flexible imputation of missing data*. Taylor & Francis Group CRC Press, 2nd edition, 2018.

[3] A. Caspi, K. Sugden, T. E. Moffitt, A. Taylor, I. W. Craig, H. Harrington, J. McClay, J. Mill, J. Martin, A. Braithwaite, and R. Poulton. Influence of life stress on depression: Moderation by a polymorphism in the 5-htt gene. *Science*, 301(5631):386–389, 2003.

[4] H. Kang. The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5):402–406, 5 2013.

[5] J. N. Michael H. Kutner, Chris Nachtsheim. *Applied Linear Regression Models*. McGraw-Hill Education, 4th edition, 2004.

[6] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to linear regression analysis*. Wiley-Blackwell, 5th edition, 2013.

[7] T. Raghunathan, P. Solenberger, and J. Van Hoewyk. *IVEware: Imputation and Variance Estimation Software User Guide*. Ann Arbor, MI, University of Michigan, 2002.

[8] C. R. Rao, H. Toutenburg, Shalabh, and C. Heumann. *Linear Models and Generalizations: Least Squares and Alternatives*. Springer, 3rd edition, 2008.

[9] N. Risch, R. Herrell, T. Lehner, K.-Y. Liang, L. Eaves, J. Hoh, A. Griem, M. Kovacs, J. Ott, and K. R. Merikangas. Interaction Between the Serotonin Transporter Gene (5-HTTLPR), Stressful Life Events, and Risk of Depression: A Meta-analysis. *JAMA*, 301(23):2462–2471, 06 2009.

[10] A. C. Tamhane and D. D. Dunlop. *Statistics and Data Analysis: From Elementary to Intermediate*. Prentice Hall, Upper Saddle River, NJ, USA, 2000.