AMS 578 Spring 2021
Multiple Regression Computing Project
Final Report

Kai Li[*]

Stony Brook University — May 3, 2021

1 Introduction

Depression is one of the top five leading causes of disability and disease burden worldwide [2]. Researchers such as Caspi et al. [2] and Risch et al. [6] 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 [2, 6]. It is also given that at most four-way interactions of independent variables will appear. Finally, compare and determine if regression results are viable.

The preliminary report shows the summary statistics table for the given dataset, the methodology used to deal with the missing data, and a close inspection of multicollinearity of the complete imputed datasets. The final report is a continuation of the preliminary report, which investigates regression model assumptions, the candidate regressors added in the model, model selection, analyses of candidate regression models, and pooling of parameters. The final report is divided into methods, results, conclusions and discussion, and appendix sections.

2 Methods

After summarizing and cleaning data, it is essential to check assumptions for linear regression. Before that, recall from the preliminary report that the methodology used to cope with the missing values in the original dataset is the multiple imputation technique. To generate unbiased estimates and correct confidence intervals utilizing this method, analyses of the two imputed datasets individually before pooling the parameters are necessary [1, 5]. Therefore, assumption verifications should be done for both imputed datasets, similarly for other applied linear regression analysis procedures, such as variable selection.

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

Kutner et al. [3] and Montgomery et al. [4] generalize the following major assumptions in linear regression theory:

1. The relationship between the dependent and independent variables should be at least approximately linear.
2. The error term has zero mean.
3. The error term has constant variance.
4. The errors are uncorrelated.
5. The errors follow a normal distribution.

Residual plots, including residuals vs. fitted, normal Q-Q (normal probability), scale-location, and residuals vs. leverage plots, can help inspect the assumptions and detect model inadequacies [3, 4]. In practice, Kutner et al. [3] and Montgomery et al. [4] suggest performing residual analysis of every regression model to ensure model adequacy, including the models obtained from transformations and stepwise regression procedures. If there are issues with nonlinearity, distribution of error terms, and unequal error variances, transformations of the dependent variable and independent variables are the appropriate next steps [3]. In particular, data transformation enables response and regressors to be in the correct scale or unit of measurement, where violations of the five basic regression assumptions are no longer exist [4]. Several techniques provide guidelines on fixing the issue of assumption violations, such as variance stabilizing, model linearization, and nonnormality correction. In this paper, transformations in the response variable will first be considered. Then, residual analysis on the transformed imputed datasets will be conducted to see if the underlying assumptions are met. If not, additional transformations on the regressor variables will be considered, again performing residual analysis after that. Also, transformations to linearize the regression model can be attempted if transformations on the dependent variable are insufficient.

When all the assumptions are satisfied for both datasets, exploring the candidate independent variables and their interaction terms in the model follows. It is essential to keep the order of the regression model as low as possible [3, 4]. Variable selection based on stepwise regression procedures will first be performed for the two imputed datasets. After that, fit linear regression models on those regressors with eight criteria examining the candidate models: SSE_p, MSE_p, R_p^2, $R_{a,p}^2$, AIC_p, BIC_p, PRESS_p, and max(VIF). Next, it is appropriate to inspect several related possible regression models provided the insights on the subject area for this project and additional candidate variables selected, such as interaction terms. The first trivial model for a dataset concerns the regression with the intercept only. Then, candidate variables without interaction terms are fitted to evaluate their appropriateness. After that, given that at most four-way interaction terms present, high-order interactions will be fitted to observe if the interaction terms are significant in stepwise procedures. All candidate models will be assessed and compared by the eight criteria. This part corresponds to the first part of model adequacy verification, using well-developed criterion. The most preferred candidate model for each imputed dataset will be determined.

The second part of model adequacy checking verifies the candidate models' assumptions and inspects influential and high-leverage observations. Unsurprisingly, residual plots are still the most powerful tools. Analysis of Variance (ANOVA) inspects the overall and individual effects of regressors in a model compared to the intercept-only model. If the regressors are significant and residual plots indicate an expected pattern, the model adequacy to the goal and the background for the project will be validated to reach the objective for this project. Finally, pooling of the parameters from the two candidate models will be done to output the final fitted model.

3 Results

First, assumptions of linear regression for both imputed datasets are verified. Both imputed datasets violate some of the linear regression assumptions, but the patterns of the residual plots are very similar. The residual vs. fitted values plots of the datasets indicate a violation of linearity and independence assumptions. The homoscedasticity assumption is violated by inspecting the scale-location plots. Based on the normal Q-Q plots, it is reasonable to have the normality assumption for both datasets. Hence, transformations are required for both imputed datasets to continue with multiple linear regression procedures.

The first transformation procedure considers transforming the response variable. Box-Cox transformations are used here. For both imputed datasets, the parameter λ obtained using the Box-Cox method returns value 0. In other words, if y is denoted to be the dependent variable values, then the transformations are $y' = \ln y$ for the two imputed datasets. By examining the residuals vs. fitted plots, normal Q-Q plots, and scale-location plots for both transformed datasets again, the transformations provide a satisfactory solution to handle assumption violations of the imputed datasets. That is, for both datasets, the relationship between the dependent and independent variables is approximately linear; the error term has a mean close to 0 and approximately constant variance; the errors can be assumed to be uncorrelated; the errors are normally distributed. Hence, the five basic regression assumptions are considered met after transforming the response.

Based on the analysis of the residual plots, a first-order regression model is sufficient and appropriate for both datasets. Stepwise regression is ready to run. Table 1 shows the selected possible regression models and the eight criteria statistics based on the stepwise regression procedures and insights given the background of the original dataset. It is crucial to remark that the stepwise regression candidate variable outputs are the same for both imputed datasets. Also, the statistics obtained from the eight criteria are very similar. Therefore, the conclusion for the model selection process should be the same. The report will mainly discuss detailed results for the first imputed dataset.

Table 1: SSE$_p$, MSE$_p$, R^2_p, $R^2_{a,p}$, AIC$_p$, BIC$_p$, PRESS$_p$, and max(VIF) Values for Selected Possible Regression Models for the First Imputed Dataset

X Variables in Model	**(1)** **p**	**(2)** **SSE$_p$**	**(3)** **MSE$_p$**	**(4)** **R^2_p**	**(5)** **$R^2_{a,p}$**	**(6)** **AIC$_p$**	**(7)** **BIC$_p$**	**(8)** **PRESS$_p$**	**(9)** **max (VIF)**
None	1	110.6	0.060	0.000	0.000	44.60	55.65	4.4×10^{23}	NA
E4, E5, R1, R7, R16	6	69.76	0.038	0.369	0.368	−797.1	−758.4	4.4×10^{23}	1.003
E4, E5, R1, R7, R16 R1×R7, R1×R16, R7×R16	9	47.38	0.026	0.572	0.570	−1506	−1451	4.4×10^{23}	3.091
E4, E5, R1, R7, R16 R1×R7, R1×R16, R7×R16, E4×E5	10	46.73	0.025	0.577	0.575	−1529	−1469	4.4×10^{23}	38.55
E4, E5, R1, R7, R16 R1×R7, R1×R16, R7×R16, R1×R7×R16	10	42.05	0.023	0.620	0.618	−1724	−1664	4.4×10^{23}	7.259

Note that interactions between the environment candidate variables and the gene candidate variables are not significant and therefore not considered the candidates of regressors. Moreover, the four-way interaction of candidate gene variables is not significant and thus omitted as well. The fourth model with an environment-by-environment interaction term yields a significant multicollinearity problem among the five selected models. Hence, any model with an environment-by-environment interaction is excluded. Also, notice that the fifth model with a three-way interaction term of candidate gene variables has a potential multicollinearity issue as well, depending on the threshold of potential multicollinearity. The preliminary report suggests a multicollinearity issue for a maximum of VIF greater than 10 [3]. Montgomery et al. [4] recommend that a VIF greater than 5 or 10 suggests the problem with potential multicollinearity. Given that other criteria do not deviate much, the three-way-interaction model will be excluded due to robustness and simplicity. After that, by comparing and analyzing the interpretation of the criterion of the top three models, the third candidate model with three one-way interactions is selected as the preferred candidate model for now. The same decision applies to the second imputed dataset.

Before concluding the models to be the final fits for the two imputed datasets, it is important to confirm the regressors are significant in the fitted candidate models through ANOVA. In particular, the first step is to inspect whether the joint effect of all regressors is significant. The ANOVA statistics for the first imputed dataset are shown in Table 2.

Table 2: Analysis of Variance for Significance of Joint Regression for the First Imputed Dataset

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-Value
Regression	63.225	8	7.903	306.75	$< 2.2 \times 10^{-16}$
Residual	47.379	1839	0.026		
Total	110.60	1847			

The P-value is very small, suggesting that the log of the dependent variable is related to at least one of the candidate regressors. Thus, at least one of the regressors is important, meaning that the candidate independent variables in the candidate model improve the fit. The conclusion is the same for the second imputed dataset. A logical question next becomes which one(s) are crucial. The results for hypothesis testing of the individual regressor effect are shown in Table 3.

Table 3: Analysis of Variance for Significance of Individual Regression for the First Imputed Dataset

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-Value
E4	15.93	1	15.93	618.16	$< 2.2 \times 10^{-16}$
E5	5.909	1	5.909	229.37	$< 2.2 \times 10^{-16}$
R1	8.192	1	8.192	317.96	$< 2.2 \times 10^{-16}$
R7	5.503	1	5.503	213.60	$< 2.2 \times 10^{-16}$
R16	5.311	1	5.311	206.16	$< 2.2 \times 10^{-16}$
R1×R7	6.732	1	6.732	261.28	$< 2.2 \times 10^{-16}$
R1×R16	9.095	1	9.095	353.00	$< 2.2 \times 10^{-16}$
R7×R16	6.557	1	6.557	254.50	$< 2.2 \times 10^{-16}$
Residual	47.379	1839	0.026		
Total	110.60	1847			

The P-values of all the candidate regressors are small, meaning that the candidate variables are all significant. Thus, including every candidate regressor can improve the fit for the first imputed data, and the same conclusion can be drawn for the second imputed dataset.

For model adequacy checking and validation, the residuals vs. fitted plot Figure 1 provides a satisfactory output because the red line is satisfactorily straight except for the boundary where some underlying influential points present. Expectedly, the normal probability and scale-location plots behave naturally, and therefore the elementary assumptions are not considered violated in the candidate model for the first imputed dataset, similarly for the second.

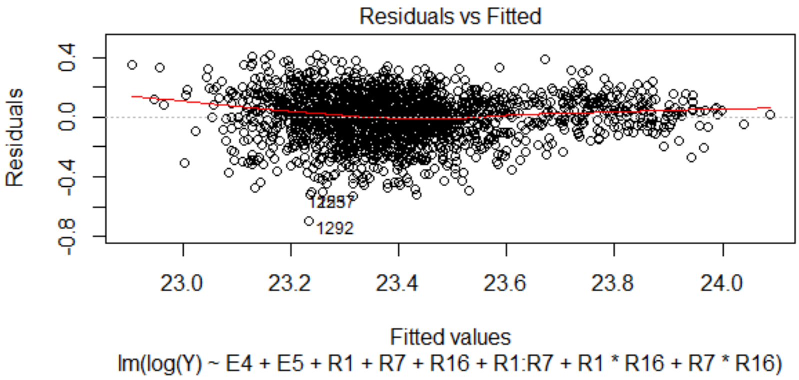

Figure 1: Residuals vs. Fitted Plot

The extreme values have to be investigated as well. The points can be leverage or influential observations. Figure 2 is a residuals vs. leverage plot that depicts Cook's distances comparing fit. From the plot, several leverage observations exist, but most data points behave naturally in a cluster. Moreover, only a few points have high Cook's distance scores compared to a large number of observations. Thus, the extreme points will not drastically affect the coefficients of the candidate regressors for the first imputed dataset. The second imputed dataset candidate model yields a similar residuals vs. leverage plot as the first imputed dataset.

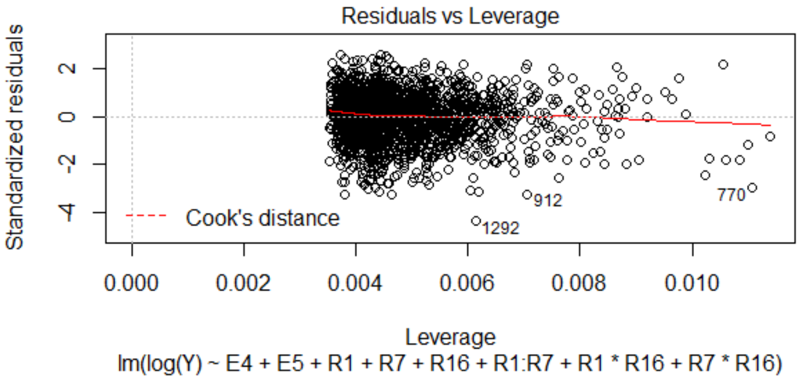

Figure 2: Residuals vs. Leverage Plot

Finally, model validation requires an analysis of whether the model will function successfully in the given background. Research conducted by Caspi et al. [2] and Risch et al. [6] illustrated the relationship between life stress events and the risk of depression, with polymorphism in the 5-HTTLTR serotonin transporter gene moderating the influence of the relationship. Here, the quantitative measurement of depression is the response variable Y; E represents the quantitative measurement of different life stress events; R consists of indicator variables denoting specific genes. The Risch et al. [6] meta-analysis results demonstrated a potential relationship between stressful life events and the risk of depression. In the candidate model, the effect of the environment on depression is significant, though the interaction between environment candidate variables is excluded from multicollinearity. Moreover, the interaction terms between E and R are all not significant, and therefore not appeared in the candidate models. However, three lines of experimental research reported a gene-by-environment interaction [2]. Also, Caspi et al. [2] showed a significant interaction between the environment candidate variables and the gene candidate variables. On the other hand, Risch et al. [6] claimed no evidence of the interaction between E and R, also shown in other meta-analyses studies in Risch et al.'s [6] paper. Due to limitations of meta-analysis, the research framework of Caspi et al. [2] may have led to different findings compared to the originally reported findings, which may result in a false positive report [6]. What is more, Caspi et al. [2] suggested the possibility of misidentifying a gene-by-environment interaction but, in fact, a gene-by-gene interaction. In this project, interactions between gene candidate variables are significant. A valid confirmation of Caspi et al. [2] results is needed [6]. Hence, the candidate model's lack of gene-by-environment interactions and the presence of gene-by-gene interactions are plausible and appropriate for the project. To summarize, the candidate model is validated with enough supportive research evidence.

Last but not least, the final model for the original dataset can be obtained by pooling the parameters using multiple imputation techniques. The model is shown in Equation 1.

$$\begin{aligned}\ln \hat{\mathrm{Y}} = \; & 22.857 + 0.000571\,\mathrm{E4} + 0.000330\,\mathrm{E5} - 0.1353\,\mathrm{R1} - 0.1384\,\mathrm{R7} - 0.1515\,\mathrm{R16} \\ & + 0.2500\,\mathrm{R1}{\times}\mathrm{R7} + 0.2732\,\mathrm{R1}{\times}\mathrm{R16} + 0.2398\,\mathrm{R7}{\times}\mathrm{R16}. \end{aligned} \tag{1}$$

4 Conclusions and Discussion

The project aims at finding an appropriate linear regression model for a given synthetic dataset with related background studies. The research started with summarizing and cleaning datasets so that a general understanding of the data is acquired. Multiple imputation was used to compensate for the missing information from the original data. Then, analysis on assumptions verification and transformation procedures for imputed datasets enabled to perform an unbiased and correct regression. The best candidate model for each imputed dataset was evaluated and chosen by the eight criteria and residual plots through a series of attempts on fitting candidate models with candidate variables selected from stepwise regression. Next, the candidate models chosen were verified viable because they matched the expectations given from the mental illness subject background. Finally, the model for this project was selected through parameters pooling.

There are certain limitations and weaknesses in this project. First, since the synthetic dataset is observational, the given dataset and the final model do not provide adequate information about cause-and-effect relationships [3, 4]. Moreover, since multiple imputation is not an easy technical fix for the missing values in a dataset, extra caution is needed so that the likelihood of getting nonsensible or even misleading results is small [1]. Especially, Buuren [1] clarifies the danger of solely using computer software to obtain outputs for using multiple imputation. Furthermore, VIF cannot distinguish between several simultaneous multicollinearities, which means additional care and insight to detect multicollinearity is needed [3, 4]. Thus, the nature of the multicollinearity may not be identified. Additionally, residual plots may not correctly show the nature behind the marginal effect, given the other regressors in the model [3]. A more sophisticated method called partial regression plots or added-variable plots can identify the marginal effects in a regression model [3, 4]. Regardless, the final model obtained in this report is still meaningful and satisfactory.

A Technical Implementation in R

Because multiple imputation technique, with two imputed datasets outputted, is used, each dataset should be inspected and analyzed on its own before pooling the parameters across the two analyses [1]. The technical implementations in R for the procedures in the two imputed datasets are identical except for changing all data1 to data2 in the code. Therefore, the implementation details for the second imputed dataset are omitted.

A.1 Regression Analysis Assumptions Verification

The validity of regression analysis assumptions should always be examined by inspecting the adequacy of the tentative model [4]. Residual plots are powerful tools for detecting violations of the five basic regression assumptions [4]. In R, function plot() outputs the four plots for a given linear regression lm(), which provide tremendous help in examining the assumptions. Readability can be enhanced if function par() is utilized to arrange the four plots in two rows and two columns.

project.r

```
fit1 <- lm(Y ~ ., data = data1)
par(mfrow = c(2, 2))
plot(fit1)
```

The residuals vs. fitted plot indicates the need for a curvilinear regression function, which means the presence of nonlinearity [3, 4]. Moreover, the zero error mean assumption can be assumed given that the intercept term is added to the model. The scale-location plot can check the constancy of the error variance. The curvilinear red line indicates a possibly nonconstant variance. Furthermore, the uncorrelated error assumption can be checked in a method similar to the checking linearity assumption. In other words, the nonstraight red line shows a possible correlation between the error terms. Finally, the normal Q-Q plot indicates a satisfactory deviation from the normal line. Thus, it is reasonable to assume that the errors are normally distributed.

A.2 Box-Cox Transformations

Box-Cox transformations are useful in transforming the dependent variable to correct nonnormality or nonconstant variance. In particular, The Box-Cox procedure uses the method of maximum likelihood (ML) to estimate the parameter λ [3]. In the MASS package, boxcox() computes the log-likelihood given λ values from -2 to 2 in steps of 0.1 by default. In the following implementation, x represents λ values, and y represents the log-likelihood values in the output of the function boxcox().

project.r

```
library(MASS)
bc1 <- boxcox(Y ~ ., data = data1, plotit = FALSE)
lambda1 <- bc1$x[which.max(bc1$y)]
```

The λ value returned by the Box-Cox method is 0 for both imputed datasets, corresponding to the transformations that $y' = \ln y$ in the response. Then, it is crucial to inspect whether a violation of assumptions exists after the transformations.

project.r

```
fit2 <- lm(log(Y) ~ ., data = data1)
plot(fit2)
```

Residual analyses for both imputed datasets provide the same conclusion. The residuals vs. fitted plot shows a far less curvilinear line. The line is almost straight, which gives a satisfactory transformation in terms of linearity and correlation of the errors. The scale-location plot also suggests a significant improvement in stabilizing the variance of both datasets. The error term has zero mean assumption is assumed because the model has an intercept term. Visually speaking, though the normal probability plot already illustrates the normality without transformations, the datasets fit the normal line more perfectly after the transformations. To conclude, the transformations are effective.

A.3 Model Selection

Model selection is a process that chooses the best fitted linear regression model among all possible models [3, 4]. Before model selections can be made, variable selection procedures need to be performed first. Because evaluating all possible regression models (including interaction terms) are computationally heavy, stepwise regression is chosen to perform the variable selection. In `R`, variable selection can be performed using function `step()` by specifying the full model containing the regressors and the argument parameter `k`, the multiple of the number of degrees of freedom used for the penalty. AIC and BIC criteria correspond to `k` equals 2 and the log of the number of observations. Here, the BIC criterion is used to generate stepwise regression results. The argument parameter `scope` in function `step()` indicates the number of ways of the interaction terms among the candidate variables. After obtaining the candidate variables from stepwise regression, it is appropriate to fit the candidate variables.

project.r

```
step(fit2, k = log(nrow(data1)))
fit3 <- lm(log(Y) ~ E4 + E5 + R1 + R7 + R16, data = data1)
step(fit3, scope = . ~ .^2, k = log(nrow(data1)))
fit4 <- lm(log(Y) ~ E4 + E5 + R1 + R7 + R16 + R1*R7 + R1:R16 +
           R7*R16 + E4*E5, data = data1)
fit5 <- lm(log(Y) ~ E4 + E5 + R1 + R7 + R16 + R1*R7 + R1:R16 +
           R7*R16, data = data1)
step(fit3, scope = . ~ .^3, k = log(nrow(data1)))
fit6 <- lm(log(Y) ~ E4 + E5 + R1 + R7 + R16 + R1*R7 + R1:R16 +
           R7*R16 + R1*R7*R16, data = data1)
step(fit3, scope = . ~ .^4, k = log(nrow(data1)))
fit7 <- lm(log(Y) ~ 1, data = data1)
```

Note that `fit5` is generated because `fit4` may result in a multicollinearity issue based on insights on the subject field. Also, the stepwise regression for the four-way interaction term yields the same regression function as the three-way interactions stepwise regression function, which means that four-way interaction is not significant.

The eight criteria shown in Table 1 can be generated with several functions in `R`. The sum of squares error, denoted SSE, is the sum of the squared residuals. The mean square for error, denoted MSE, is the sum of squares error divided by the number of observations minus the number of parameters p in the model. The determination coefficient and adjusted determination coefficient denoted R^2 and R_a^2, respectively, can be obtained directly from function `summary()` statistics. The PRESS statistics can be obtained from function `PRESS()` under `qpcR` package. Finally, VIF implementation is the same as the one in the preliminary report. The following code shows the implementation for the second candidate model in Table 1. The others are the same except for changing the fitted model number and the number of parameters in the criterion functions.

project.r

```r
sum(fit3$residuals^2)
sum(fit3$residuals^2)/(nrow(data1)-6)
summary(fit3)$r.squared
summary(fit3)$adj.r.squared
AIC(fit3)
AIC(fit3, k = log(nrow(data1)))
library(qpcR)
PRESS(fit3)$stat
library(car)
max(vif(fit3))
```

A.4 Analysis of Variance

The analysis of variance table statistics can be obtained using function `anova()`. Function `anova()` accepts both one and two arguments as the parameters. If two arguments are inputted, with the first being the intercept-only model and the second being the fitted model, the ANOVA statistics for joint regressors' effect on the fitted model will be returned. The statistics are summarized in Table 2. If only one argument is inputted, individual effect of the regressors to the fitted model will be generated, shown in Table 3.

project.r

```r
anova(fit7, fit5)
anova(fit5)
```

A.5 Model Adequacy Checking

The procedure for model adequacy checking is mainly examining the residual plots of the best candidate model. Therefore, the function `plot()` creates the four plots for visualization, similar to the implementation mentioned before.

project.r

```
plot(fit5)
```

A.6 Parameter Pooling

Several classes in the `mice` package help to perform multiple imputation procedures. Function `mice()`, used in the preliminary report to generate imputed datasets, is the foundation for generating a pool of multiple imputed parameters. Then, the model of interest on each imputed dataset is fitted by function `with()`. Finally, the final pooled model is obtained using function `pool()`.

project.r

```
imp <- mice(data, method = "cart", m = 2, maxit = 10,
            seed = 123, print = FALSE)
fit <- with(data = imp,
            exp = lm(log(Y) ~ E4 + E5 + R1 + R7 + R16 +
                     R1:R7 + R1*R16 + R7*R16))
summary(pool(fit))
```

References

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

[2] 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.

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

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

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

[6] 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.