Lucy Lin
Professor Finch
AMS 315
4 December 2020

AMS 315 Project 2

Introduction

The task for this project is to estimate the function from the data set given. The data set consists of 4 environmental variables (E1-E4) and 20 gene variables (G1-G20) that affect the outcome of a clinical trial with a total of 1325 observations with no missing information. Per the recommendations on the handout by Songzhu Zheng, third and second order interactions were tested, such that a second order interaction was modeled between the independent variables. Stepwise regression was performed, and the final variables for the model were chosen.

Methods

The data was read and analyzed using the R program. Two residual plots were created to model a possible second order interaction between the variables, resulting in a flat ellipse in Figure 1 and a regular ellipse after a transformation of sqrt(Y) in Figure 2. A correlation between the variables was determined and a Box-Cox transformation was performed (see Appendix), and the sqrt(Y) transformation was established. A stepwise regression was performed using leaps from the MASS R package. From the tables and figures produced by the kable command, four models were shown. The third and fourth model had similar R^2 and BIC values. The fourth model was eventually chosen since there were no significant coefficients generated by the third model and the remaining coefficients in the fourth model were significant at the 0.001 level.

Results

The original adjusted R^2value was 0.5079178. After the transformation of sqrt(Y), the adjusted R^2value became 0.5661488. From table 3, the model only included E4 and G4 but since the t-values are larger than 4, these variables are valid. The model becomes

$\sqrt{y} = \beta_0 + \beta_1 E_4 + \beta_2 G_4$ or as an estimate:

$\sqrt{y} = -531.8008 + \beta_1(247.0719) + \beta_2(134.7519)$. There is no second order interaction between the E_4environmental variable and G_4genetic variable. The F values of both

E_4 and G_4 are significantly higher than those of the other variables (ANOVA table in Appendix) and thus it is reasonable to conclude that they are significant variables in the model.

Conclusion/Discussion

The estimated model of the data given is approximately $\sqrt{y} = \beta_0 + \beta_1 E_4 + \beta_2 G_4$ or $\sqrt{y} = -531.8008 + \beta_1(247.0719) + \beta_2(134.7519)$. Since the R^2 value had a notable increase and the F values are significant, this is a possible fit for the data.

Appendix (Figures, Tables, R code)

Figure A

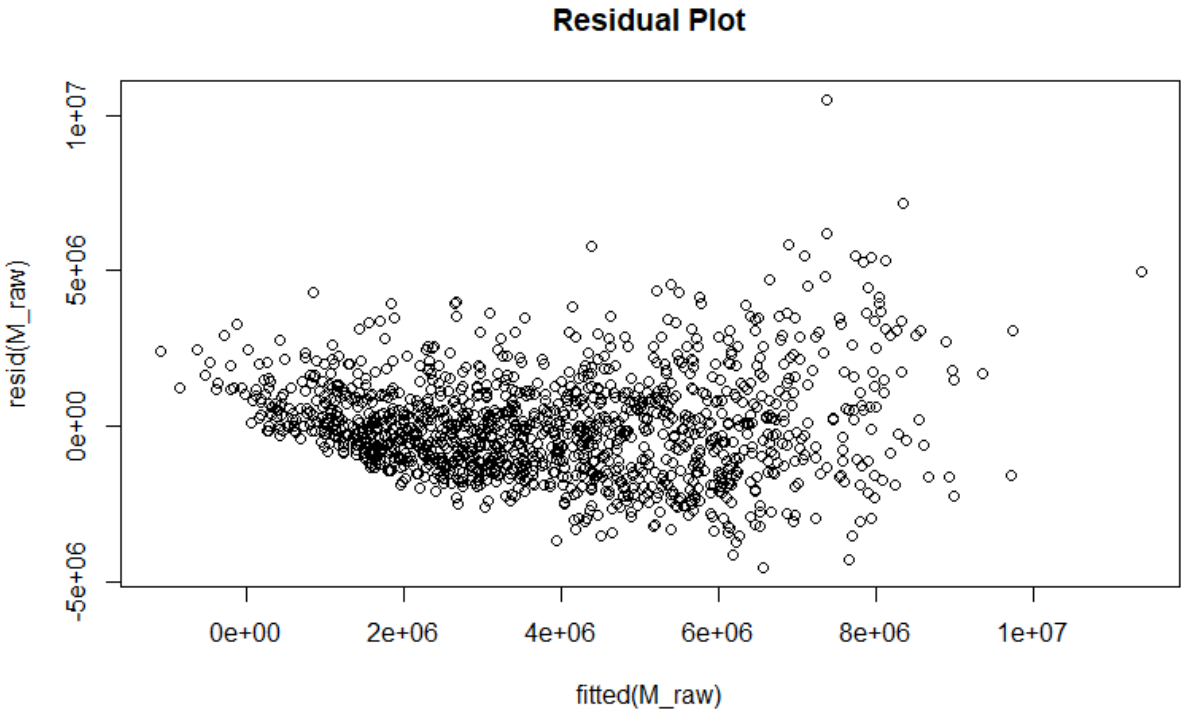

Figure B

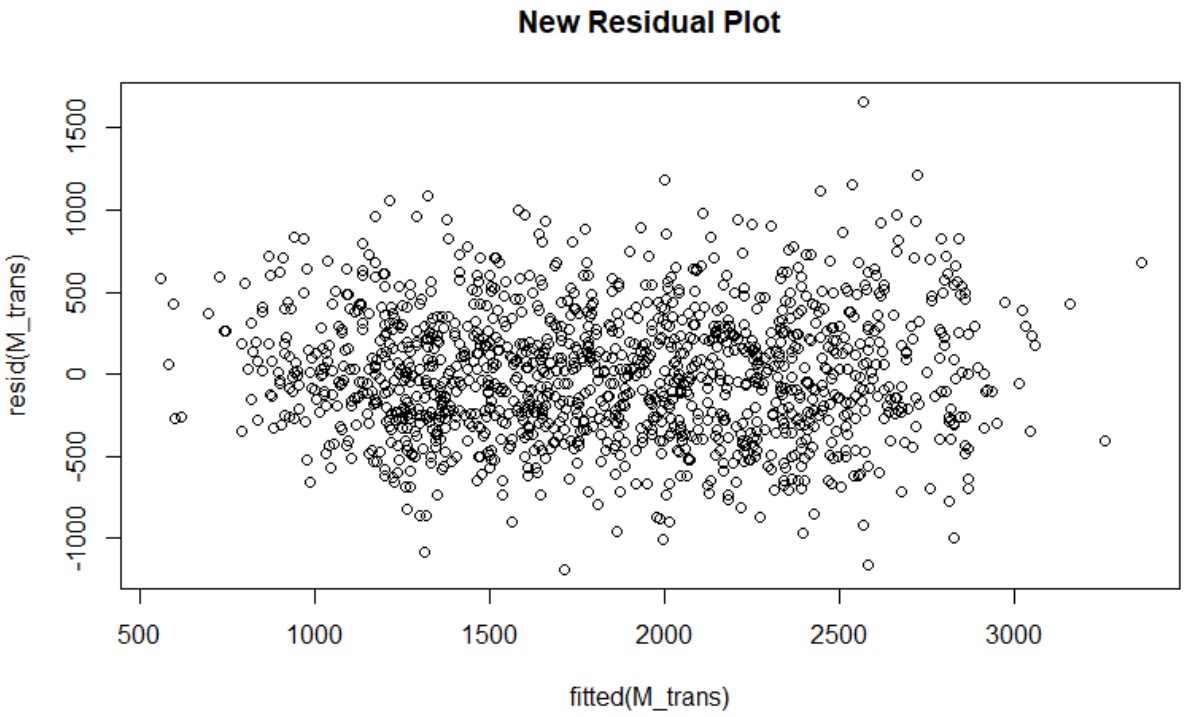

Table 1

Table: Sig Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-531.8008	116.605564	-4.560681	5.6e-06
E4	247.0719	6.108471	40.447424	0.0e+00
G4	134.7519	27.614633	4.879728	1.2e-06

Table 2

Table: Model Summary

model	adjR2	BIC
(Intercept)+E4	0.555459792853152	-1060.81987261196
(Intercept)+E4+G4:G14	0.566969493384017	-1089.39042728018
(Intercept)+E4+G1:G8+G4:G14	0.569422747507069	-1090.73179427346
(Intercept)+E4+G1:G8+G4:G14+G7:G8	0.571419292591046	-1090.70421977321

Table 3

Table: ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
E1	1	9.797708e+05	9.797708e+05	5.0597156	0.0246550
E2	1	1.393628e+04	1.393628e+04	0.0719695	0.7885331
E3	1	1.688241e+06	1.688241e+06	8.7183830	0.0032067
E4	1	3.275105e+08	3.275105e+08	1691.3240669	0.0000000
G1	1	4.017946e+05	4.017946e+05	2.0749411	0.1499765
G2	1	1.923581e+05	1.923581e+05	0.9933725	0.3191050
G3	1	9.705627e+01	9.705627e+01	0.0005012	0.9821420
G4	1	4.268269e+06	4.268269e+06	22.0421203	0.0000029
G5	1	1.148164e+05	1.148164e+05	0.5929331	0.4414270
G6	1	2.387188e+05	2.387188e+05	1.2327874	0.2670710
G7	1	7.289416e+05	7.289416e+05	3.7643880	0.0525710
G8	1	1.240556e+05	1.240556e+05	0.6406457	0.4236233
G9	1	8.412284e+04	8.412284e+04	0.4344257	0.5099420
G10	1	2.120380e+05	2.120380e+05	1.0950028	0.2955599
G11	1	2.086602e+05	2.086602e+05	1.0775594	0.2994372
G12	1	2.208347e+05	2.208347e+05	1.1404310	0.2857595
G13	1	7.117796e+04	7.117796e+04	0.3675760	0.5444344
G14	1	1.079080e+06	1.079080e+06	5.5725647	0.0183910
G15	1	3.453967e+05	3.453967e+05	1.7836920	0.1819304
G16	1	4.736980e+05	4.736980e+05	2.4462631	0.1180484

G17	1	2.039825e+05	2.039825e+05	1.0534028	0.3049156
G18	1	1.929718e+04	1.929718e+04	0.0996542	0.7522957
G19	1	4.982610e+04	4.982610e+04	0.2573111	0.6120596
G20	1	4.660404e+03	4.660404e+03	0.0240672	0.8767382
Residuals	1300	2.517339e+08	1.936415e+05	NA	NA

R Code

```
getwd()
P2 <- read.csv('FA20_P2_33301.csv', header=TRUE)
M_E <- lm(Y ~ E1+E2+E3+E4, data=P2)

View(P2)
summary(M_E)
summary(M_E)$adj.r.squared
##[1] 0.5079178
M_raw <- lm( Y ~
(E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16
+G17+G18+G19+G20)^2, data=P2)
plot(resid(M_raw) ~ fitted(M_raw), main='Residual Plot')
boxcox(M_raw)
M_trans <- lm( I(sqrt(Y)) ~ (.)^2, data=P2)
summary(M_raw)$adj.r.square;
##[1] 0.5209232
```

```
summary(M_trans)$adj.r.square
##[1] 0.5661488
plot(resid(M_trans) ~ fitted(M_trans), main='New Residual Plot')

install.packages("leaps")
library(leaps)
M <- regsubsets( model.matrix(M_trans)[,-1], I(sqrt(P2$Y)),
                 nbest = 1 , nvmax=4,
                 method = 'forward', intercept = TRUE )
temp <- summary(M)
Var <- colnames(model.matrix(M_trans))
M_select <- apply(temp$which, 1,
                 function(x) paste0(Var[x], collapse='+'))
kable(data.frame(cbind( model = M_select, adjR2 = temp$adjr2, BIC = temp$bic)),
      caption='Model Summary')
M_main <- lm( I(sqrt(Y)) ~ ., data=P2)
temp <- summary(M_main)
kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.001, ], caption='Sig Coefficients')

M_2stage <- lm( I(sqrt(Y)) ~ (E4+G4)^2, data=P2)
temp <- summary(M_2stage)
temp$coefficients[ abs(temp$coefficients[,3]) >= 4, ]
citation()
```

References

Zheng, Songzhu. (n.d.). “Multiple Regression Handout.” *Applied Mathematics and Statistics | Stony Brook University*, blackboard.stonybrook.edu/bbcswebdav/pid-5786100-dt-content-rid-49205933_1/courses/1208-AMS-315-SEC01-89219/Multiple%20Regression%20Handout%20F2020.html.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

(Additional credit to various AMS 315 TAs for the continuous help during this project)