Data Project 2

Introduction

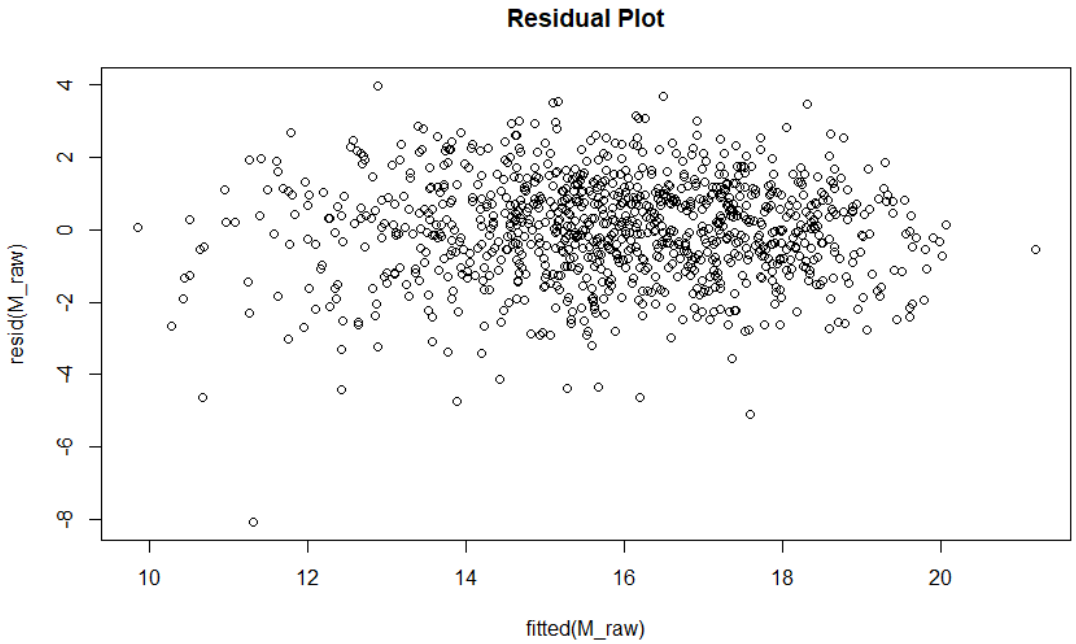
This report details the process and results of estimating the function that was used to generate the data provided for this assignment. This function (or model) was sourced from research by Caspi et al (2003) that reported a significant gene-environment (G x E) interaction implying that depression due to stressful life events is more prevalent amongst subjects who possess a short allele for the 5-HTT gene. This G x E interaction, which can be described in general as an influence on the expression of an observed trait due to interplay between genes and the environment, was tested by Caspi et al using a moderated regression framework on a set of 847 subjects of the same age, sorting them for genotype. They looked for association between depression and (i) 5-HTTLPR genotype, (ii) stressful life events, and (iii) interaction between 5-HTTLPR and the stressful life events for each subject (Caspi et al., 2003). In similar fashion, this report incorporates multiple regression analysis and will look at several independent environmental and genetic variables for any significant associations between each of them and the outcome variable, as well as associations between their interaction terms and the outcome variable. Research questions include: What is the original function? What transformations of variables may be needed? Which independent variables are considered significant?

Methods

For this report, the R programming language is used via RStudio to analyze the data, generate relevant statistics and plots, and to arrive at the original model. The methods and related code used here largely follow the steps laid out in "Multiple Regression Handout" by Benjy Hechtman and Jin Huang (2021). First, the assigned dataset containing 1,010 observations of 25 variables (the outcome variable and 24 independent variables, 4 of which are environmental and 20 genetic) was imported to RStudio. An initial model was then fitted using only the environmental variables (E1, E2, E3, and E4) to see how well they would explain the outcome alone, based on the model's adjusted r^2 value. After this control for the environmental variables, a model including all independent variables (environmental and genetic) along with all their possible 2nd order interactions was fitted and observed. Using the command "plot" to create a residuals vs fitted plot of this latter model, some heteroscedasticity was immediately evident, shown in Figure 1 below.

Figure 1

Residuals Plot for Raw Data Model Shows Clear Undesired Pattern



Note: Optimally, these points should exhibit a flat, more evenly-spaced ellipse.

To reduce the presence of this pattern, a Box-Cox transformation was applied to the model using the command "boxcox" from the R library MASS. As shown in Figure 3 in the appendix, log-likelihood is maximized with $\lambda \approx 2$. Thus, a third model was fitted, again with all independent variables and 2nd order interactions, but now with the dependent (outcome) variable, Y, raised to the 2nd power. This transformation slightly improved the adjusted $r\text{^}2$ value of the model, and made the residuals plot virtually without any pattern (see Figure 4 in Appendix). Stepwise regression was then conducted on this transformed model with the "regsubsets" command of the leaps R package to find the most influential independent variables, and a set of proposed models was produced by the software. From this set, the 4th model, including variables E3 and E4 as well as the interaction terms E2:G8 and E3:E4, was chosen as a candidate for the final model since it maximized adjusted r^2 while minimizing the Bayesian Information Criterion (BIC) value (there was not a significant difference in either of these values going to the 5th model) (see Figure 5 in Appendix). To observe which variables had a significant main effect in the data, a 1st order model with Y^2 and all independent variables (no interaction terms) was analyzed. Only E3, E4, and G8 proved to be highly significant with p-values < 0.001, meaning

they were each highly likely to be part of the original model used to generate the data. However, with E2 absent from this list of significant coefficients, it was possible that the candidate model chosen was not accurate. Thus, several models were tested, including all combinations of the candidate model, looking for t values > 4 for all coefficients and maximization of adjusted r^2. Finally, a seemingly accurate model was reached, with all coefficients significant.

Results

Adjusted r^2 for the model containing solely environmental variables was 0.4998917. With the addition of genetic variables and all 2nd order interactions, this increased to 0.5097335, and after transformation of Y to the 2nd power, adjusted r^2 was equivalent to 0.5177034. The final model arrived at was $Y^2 = \beta_0 + \beta_1 E_3 + \beta_2 E_4 + \beta_3 E_2 G_8 + \epsilon$. Although E3:E4 was a candidate term, it was not found to be significant, and though E2 was not significant, the interaction term E2:G8 was. Then, plugging in the estimates provided by R for β_i , $i = 1, 2, 3$, the exact final model would be: $Y^2 = 5.7115 + (9.7444)E_3 + (15.1887)E_4 + (1.4098)E_2 G_8 + \epsilon$. See Figure 2 below for these respective estimates, along with other summary statistics and the model's analysis of variance table. With this final model, the adjusted r^2 increased once more, reaching 0.523715. Also shown below, all coefficients included were sufficiently significant in their main effects, each with t-values > 4.

Figure 2

Analysis of Variance Table and Summary Statistics for Final Model

Table: ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
E3	1	821265.62	821265.616	325.33167	0.00e+00
E4	1	1940384.57	1940384.573	768.65334	0.00e+00
E2:G8	1	46687.82	46687.825	18.49466	1.87e-05
Residuals	1006	2539541.27	2524.395	NA	NA

Final Model ANOVA Table

```
> summary(M_final)

Call:
lm(formula = I(Y^2) ~ E3 + E4 + E2:G8, data = projdata)

Residuals:
     Min       1Q   Median       3Q      Max
-178.534  -33.763    1.029   35.290  135.020

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.7115     7.9188   0.721    0.471
E3            9.7444     0.5544  17.577  < 2e-16 ***
E4           15.1887     0.5480  27.717  < 2e-16 ***
E2:G8         1.4098     0.3278   4.301 1.87e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 50.24 on 1006 degrees of freedom
Multiple R-squared:  0.5251,    Adjusted R-squared:  0.5237
F-statistic: 370.8 on 3 and 1006 DF,  p-value: < 2.2e-16
```

Final Model Summary Results

Conclusions and Discussion

With a final adjusted r^2 value of 0.523715, roughly 52.37% of variation in the outcome variable is explained by the independent variables that affect it. Thus, there is a modest association between the squared outcome variable and these select independent variables. Although the final model described was the best found, there were limitations inherent to this report's procedures that leave the exact model originally used unknown. One such limitation was the necessity to test several different final models because inclusion or exclusion of terms and interactions of terms would change the list of significant coefficients. This left finalization somewhat up to guess and check. Furthermore, transformations of independent variables may have been needed, and even the transformation of Y chosen here was a rough estimate. Nonetheless, this report reflects a best effort. In conclusion, this analysis found a significant G x E interaction (E2:G8), and including the chosen variables increased the adjusted r-squared value.

Appendix

Figures and Computer Code

Figure 3

Box Cox Transformation Plot

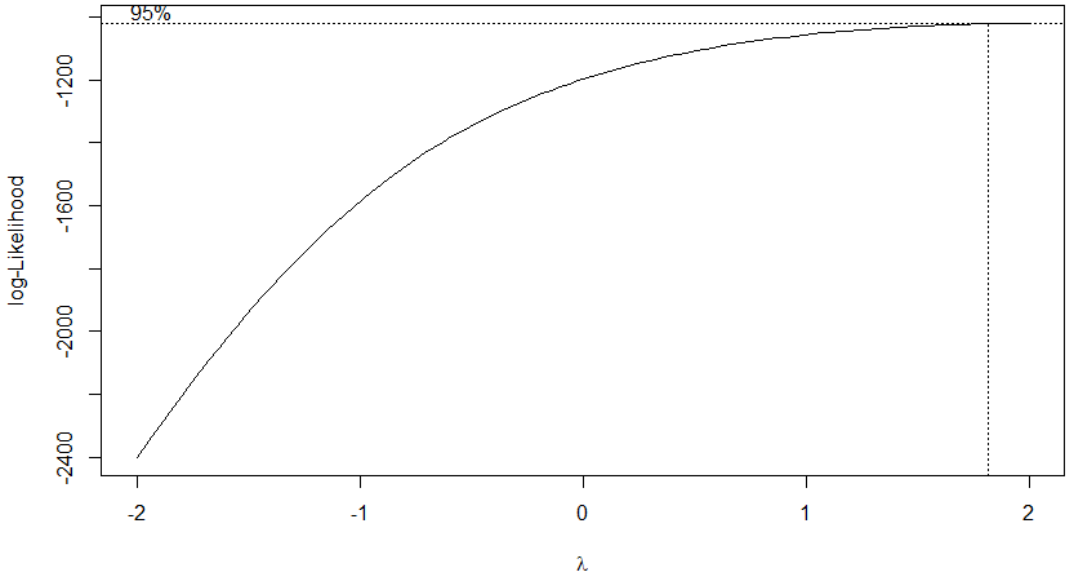


Figure 4

New Residuals vs Fitted Plot Following Transformation of Y by 2nd Power

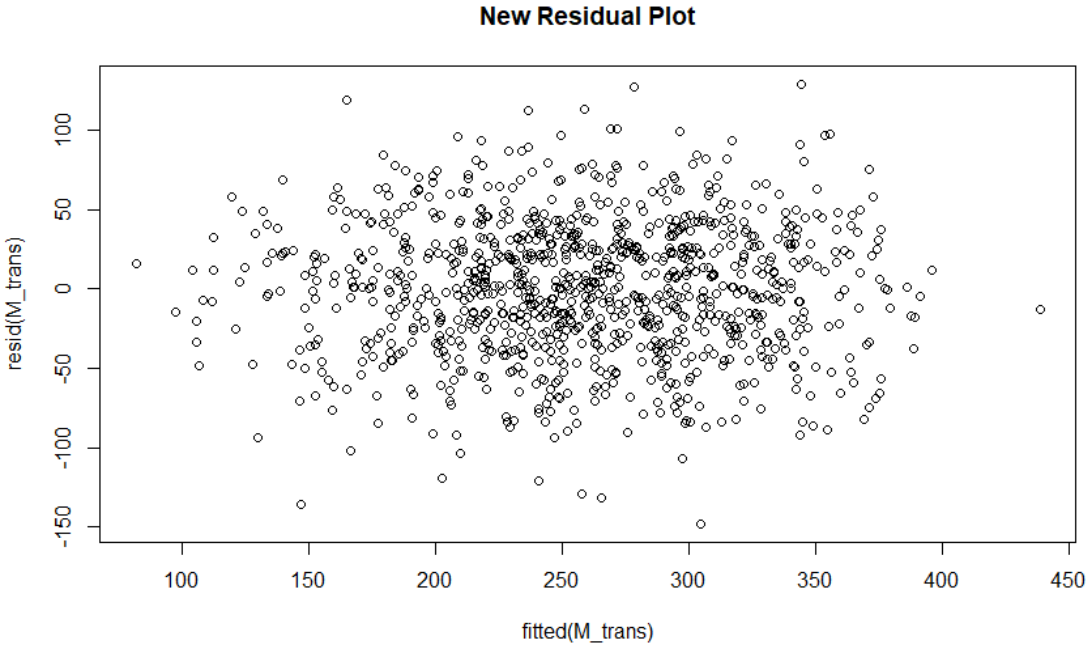


Figure 5

Potential Models as Proposed by Stepwise Regression + Significant Coefficients

Table: Model Summary

model	adjR2	BIC
(Intercept)+E3:E4	0.465606712278376	-620.055525631981
(Intercept)+E4+E3:E4	0.496812640624957	-674.911473416074
(Intercept)+E3+E4+E3:E4	0.516265149768837	-708.81715995647
(Intercept)+E3+E4+E2:G8+E3:E4	0.524781709668394	-720.844225948404
(Intercept)+E3+E4+E2:G8+E3:E4+G1:G13	0.528110124612573	-722.030895474575

Proposed Models

#4

```
> M_main <- lm( I(Y^2) ~ E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16+G17+G18+G19+G20,
> temp <- summary(M_main)
> kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.001, ], caption='Sig Coefficients')
```

Table: Sig Coefficients

	Estimate	Std. Error	t value	Pr(>|t|)
E3	9.796713	0.5611130	17.459430	0.00e+00
E4	15.151631	0.5550192	27.299294	0.00e+00
G8	14.925070	3.5146394	4.246544	2.38e-05

Significant Coefficients