Gene Interaction Model Analysis

April 6, 2025

1 Introduction

In this project, the dataset contained 1080 observations, comprising one dependent variable and 24 independent variables. Among the independent variables, there are 4 environmental variables (E1–E4) and 20 gene variables (G1–G20). The main objective is to determine the function used to generate the data. Additional research questions include examining whether there is an association of the outcome variable Y with the gene variables (G), gene-environment interactions (G×E), or gene-gene interactions (G×G). The background of this research is to apply multiple regression techniques to assess the association between the outcome variable and one or more genetic variables while controlling for the environmental variable.

2 Methods

The analysis was conducted using the R programming language. Initially, the correlation between Y and the independent variables was examined. It was found that almost no correlation existed between Y and each independent variable, except for E_4, which showed a moderately strong positive linear relationship.

A Box-Cox transformation was then used to identify a suitable transformation for Y. Based on the Box-Cox plot, a lambda value of 0.35 was chosen. Thus, the transformed variable $Y^{0.35}$ was used for the subsequent multivariable regression analysis.

Stepwise regression was performed using the `leaps` and `knitr` libraries in R. By examining the adjusted R^2 column from the output table, a model was selected when a notable increase was observed with the addition of an interaction variable. The chosen test model was:

$$(\text{Intercept}) + E_4 + (G_5 : G_{13}),$$

where the interaction term $G_5 : G_{13}$ represents a gene-gene (G×G) interaction. To address multiple testing issues, the Bonferroni correction was applied.

3 Results

Based on the analysis, the final model incorporates the variables E_4, G_5, and G_{13}. The final regression model is:

$$Y^{0.35} = 24.8394 + 14.5468\, E_4 + 16.6106\, (G_5 : G_{13}) + \epsilon,$$

with a residual standard error of 32.97. An F-statistic of 449.1 allowed us to reject the null hypothesis that the regression coefficients are zero.

3.1 Model Coefficients

Coefficient	Estimate	Std. Error	t-value	p-value
Intercept	24.8394	4.8810	5.089	~ 0
E_4	14.5468	0.5016	28.999	~ 0
$G_5 : G_{13}$	16.6106	2.7586	6.021	~ 0

Table 1: Final Model Summary

4 Conclusion/Discussion

The best model for the data is:

$$Y^{0.35} = 24.8394 + 14.5468\, E_4 + 16.6106\, (G_5 : G_{13}) + \epsilon.$$

This result indicates that while there is no direct association of Y with the gene variables on their own, a significant gene-gene interaction exists between G_5 and G_{13}. No significant gene-environment (G×E) interaction was found. The Bonferroni correction was used to reduce the likelihood of false positives, balancing the identification of true effects with the control of Type I errors.

Limitations include potential issues from multiple testing and the interpretation of the statistical methods. The Box-Cox plot aided in selecting an appropriate transformation for Y, and the stepwise regression method was useful for identifying the model with the largest increase in adjusted R^2.

Appendix A: Graphs and Tables

Graphs

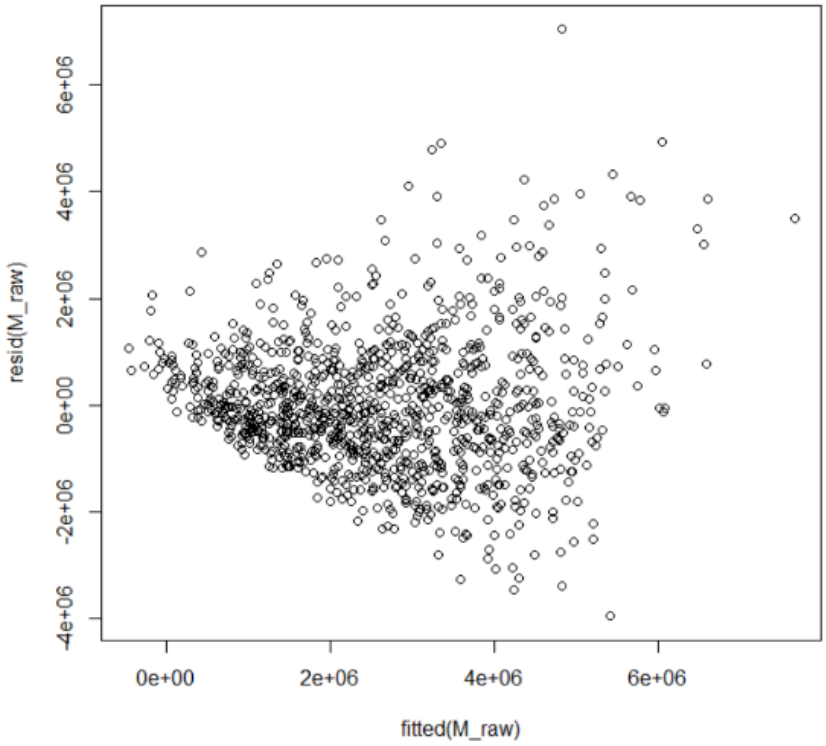

Figure 1: Original Residual Plot

Tables

Model Summary Table for Environmental Variables

Coefficient	Estimate	Std. Error	t-value	p-value
Intercept	-3311665	430651	-7.69	~ 0
E_1	-1994	22018	-0.091	0.928
E_2	29484	22208	1.328	0.185
E_3	15410	21829	0.706	0.48
E_4	580497	22015	26.369	~ 0

Table 2: Environmental Variables Model Summary. Residual standard error: 144600 on 1075 degrees of freedom. Multiple R^2: 0.3932, Adjusted R^2: 0.391, F-statistic: 172.4 on 4 and 1075 DF, p-value: ~ 0.

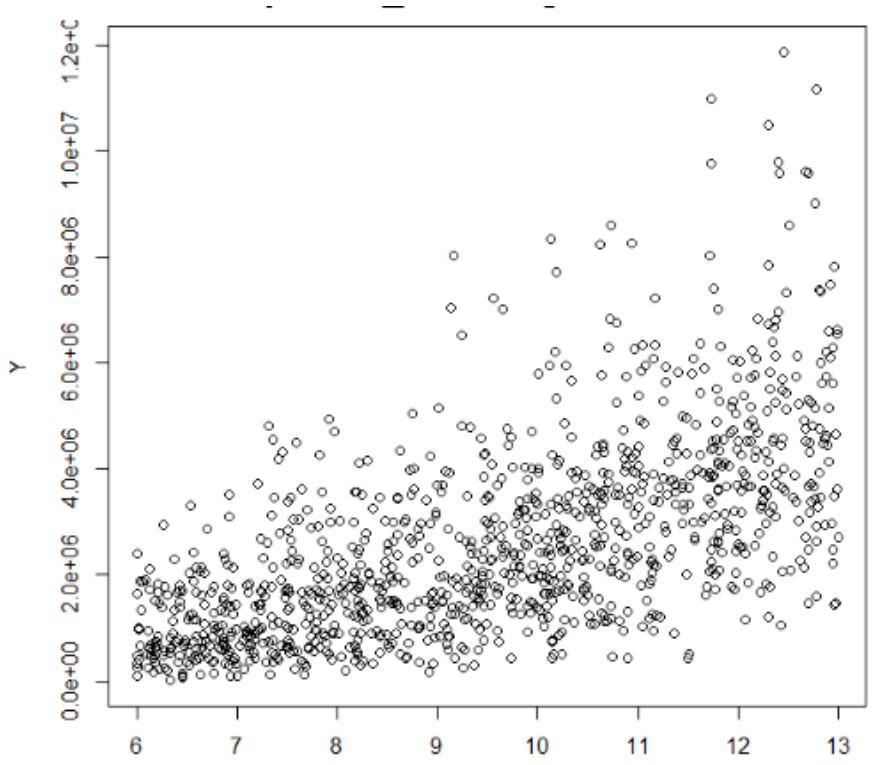

Figure 2: Y vs. E_i Scatterplot

Model Summary Table for Stepwise Regression

Model	Variables	Adjusted R^2	Notes
Model 1	...	...	...
Model 2	Intercept, E_4, $G_5 : G_{13}$	...	Selected model

Table 3: Stepwise Regression Model Summary

Significant Coefficients Table

Coefficient	Estimate	Std. Error	t-value	p-value
Intercept	24.8394	4.8810	5.089	~ 0
E_4	14.5468	0.5016	28.999	~ 0
$G_5 : G_{13}$	16.6106	2.7586	6.021	~ 0

Table 4: Significant Coefficients from the Final Model

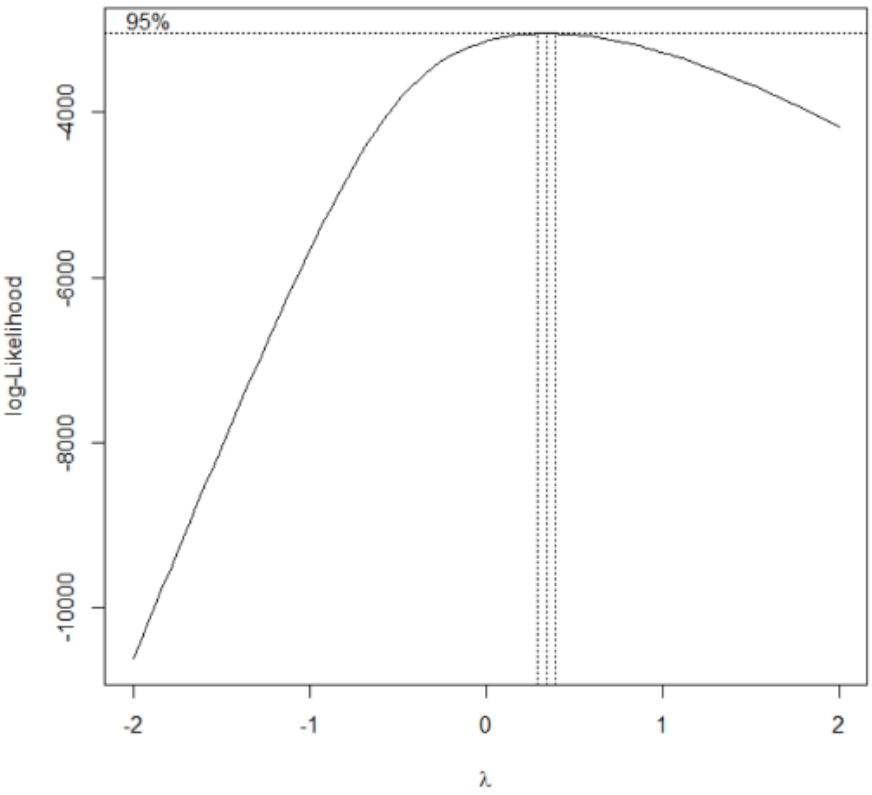

Figure 3: Box-Cox Plot

Model Summary Table for the Final Model

Coefficient	Estimate	Std. Error	t-value	p-value
Intercept	24.8394	4.8810	5.089	~ 0
E_4	14.5468	0.5016	28.999	~ 0
$G_5 : G_{13}$	16.6106	2.7586	6.021	~ 0

Table 5: Final Model Summary. Residual standard error: 32.97 on 1077 degrees of freedom. Multiple R^2: 0.4547, Adjusted R^2: 0.453, F-statistic: 499.11 on 2 and 1077 DF, p-value: ~ 0.

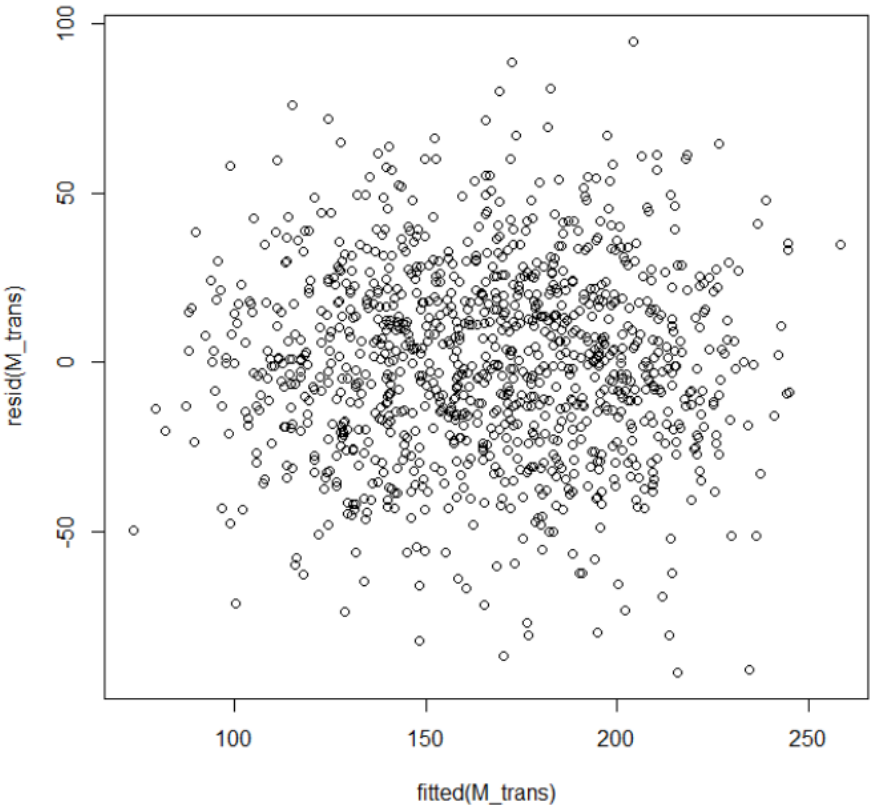

Figure 4: Transformed Residual Plot