

Delong Zhu
Gene–Environment Interaction Modeling

Introduction:

The purpose of this project is to perform statistical and data analysis on the relationship and interactions of gene-environments. Through the implementation of R, Multiple regression testing was done, considering the 5-HTT genotype to determine the relationship between the Y variable and the controlled environment variable that were taken from a CSV file.

Methods:

A CSV dataset file was imported into RStudio, where a systematic approach to analyze the data set was implemented. Upon importing the data, `is.na()` was used to determine if there are any missing values within the database, in order to ensure that the data was complete and eligible for further analysis. The datasets' statistics were then computed to get an idea of what the dataset implies and represents.

Exploratory data analysis was then conducted on the dataset, which calculates and stats the correlations between the variables, specifically, the dependent Y variable, E1, E2, E3, E4, and other variables in the dataset. After conducting the exploratory data analysis, 4 scatter plots were generated to show the relationship between each E variable and Y respectively.

Multiple regression analyses were performed to further investigate the relationship between Y and the other variables. Setting Y as the dependent variable and E1, E2, E3, E4 as the independent variable, a multiple regression model was computed, to get the coefficient, p-values, and R-squared values of the dataset. The Box-Cox transformation was implemented due to the original residual plot showing non-normality, as such a new residual model was created with the newly transformed Y variable.

G1, G2, G3... genetic variables were then used to determine the relationship between the environments (E) and genetic (G) variables. The Bonferroni inequality was applied to adjust p-values, ensuring the reliability of the statistical test in order to mitigate multiple comparison issues. Through the incorporation of the environmental variables E1, E2, E3, E4 and the genetic variables G1, G2, G3... G12, a final regression model was constructed aligned with new tables to visualize and analyze the dataset.

Results:

Through the analysis of the Box-Cox transformation plot, it was found that there was an approximation of 0.56. The original multiple regressions has a multiple R-squared value of 0.556, with an adjusted R-squared value of 0.5545. While the final model regression reported a multiple R-squared value of 0.6622 and adjusted R-squared value of 0.5524

Conclusion:

In conclusion, from the result of the multiple regression model, there is a relationship between the dependent variable Y, and the environment variables E1, E2, E3 and E4. Furthermore, G2,G3,G4...G12 has a great association with the square root of the outcome variable, as the multiple R-squared value has increased to 0.6622 from the original 0.556.

Appendix:

Original Model Summary:

Residuals:

Min	1Q	Median	3Q	Max
-1.54860	-0.30914	0.02068	0.30525	1.35298

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	24.174225	0.393534	61.429	<2e-16 ***
E1	0.124694	0.009062	13.760	<2e-16 ***
E2	-0.012611	0.009145	-1.379	0.168
E3	0.229252	0.009022	25.411	<2e-16 ***
E4	0.224836	0.009015	24.940	<2e-16 ***

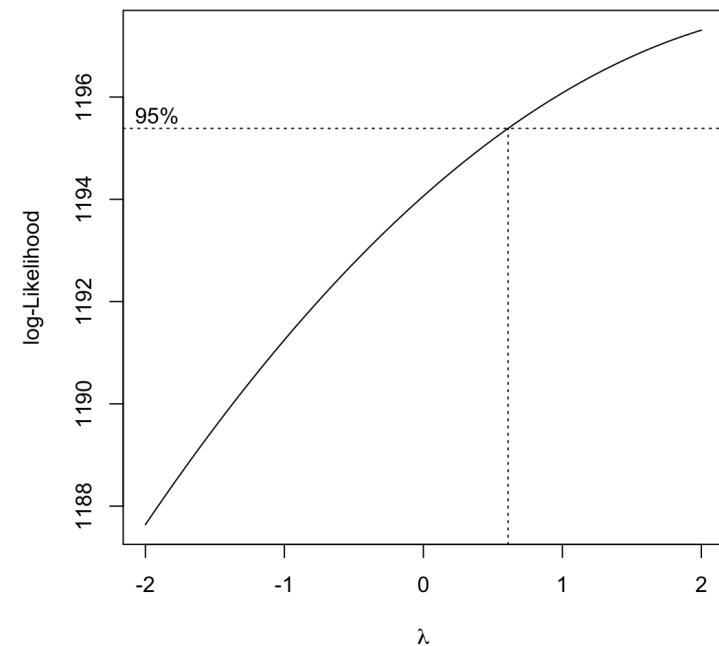
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4562 on 1219 degrees of freedom

Multiple R-squared: 0.556, Adjusted R-squared: 0.5545

F-statistic: 381.6 on 4 and 1219 DF, p-value: < 2.2e-16

Box-Cox Transformation:



Final Model Summary:

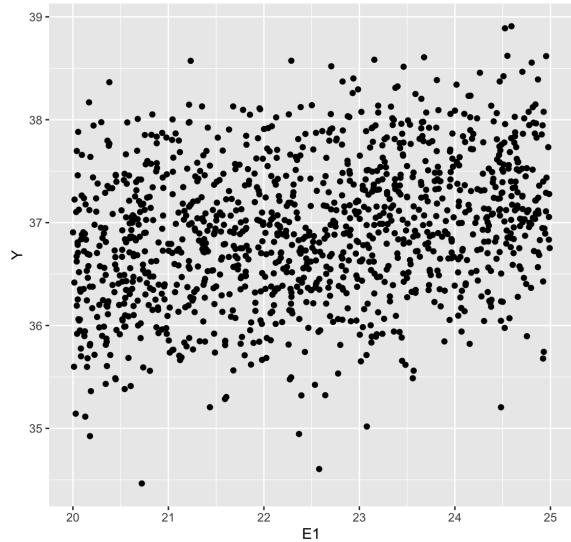
Residual standard error: 0.4573 on 923 degrees of freedom
Multiple R-squared: 0.6622, Adjusted R-squared: 0.5524
F-statistic: 6.031 on 300 and 923 DF, p-value: < 2.2e-16

Significant Coefficient Effects:

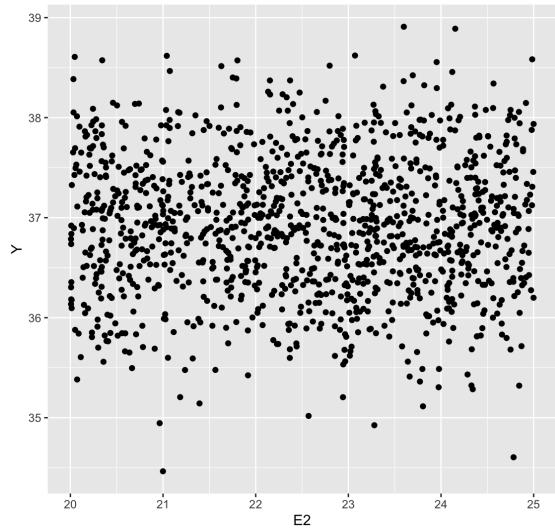
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.65278081	0.01834311	199.136421	0
E1	0.00568601	0.00041791	13.606581	0
E3	0.01046611	0.00041521	25.204571	0
E4	0.01038721	0.00041741	24.886151	0

Scatter Plots:

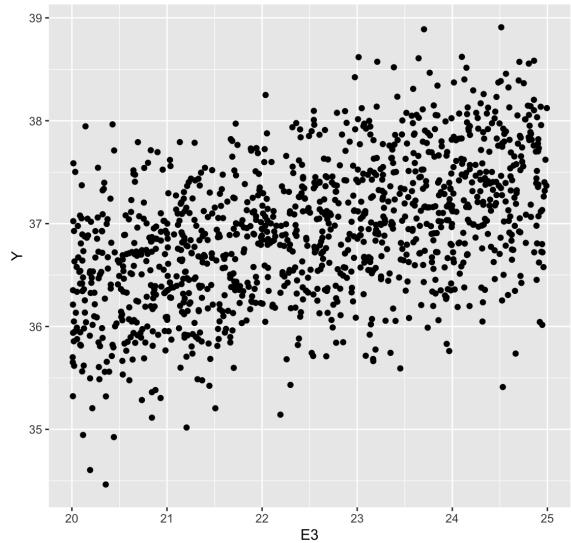
Scatter plot of Y vs E1



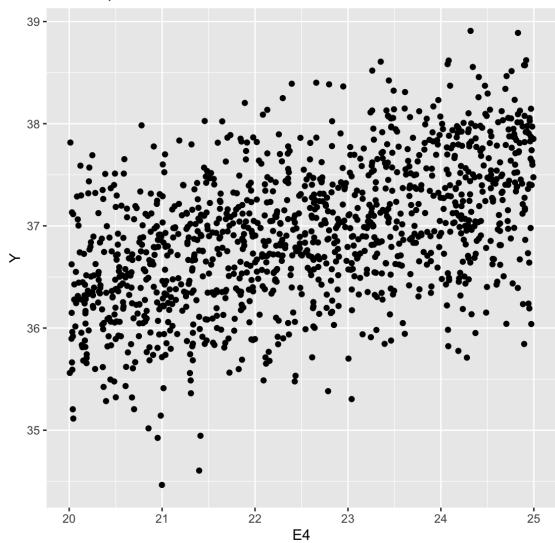
Scatter plot of Y vs E2



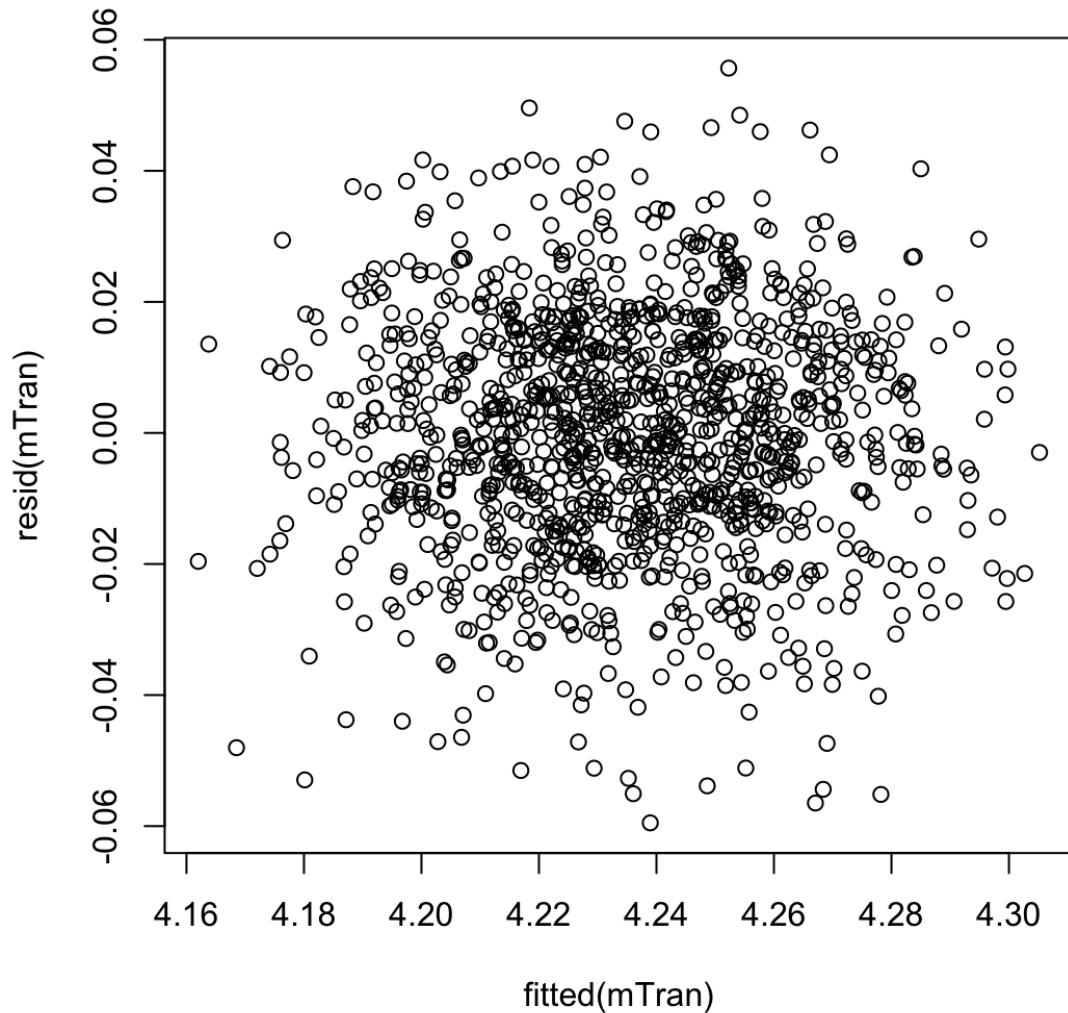
Scatter plot of Y vs E3



Scatter plot of Y vs E4



New Residual Plot



model	adjR2
:-----	:-----
(Intercept)+E3:E4	0.485265555571121
(Intercept)+E1+E3:E4	0.553608928980981
(Intercept)+E1+E3:E4+G7:G17	0.556728572077699
(Intercept)+E1+E3:E4+G7:G17+G9:G11	0.559883596008713
(Intercept)+E1+E3:E4+G5:G15+G7:G17+G9:G11	0.561311178243038