Ronald Chao

Introduction

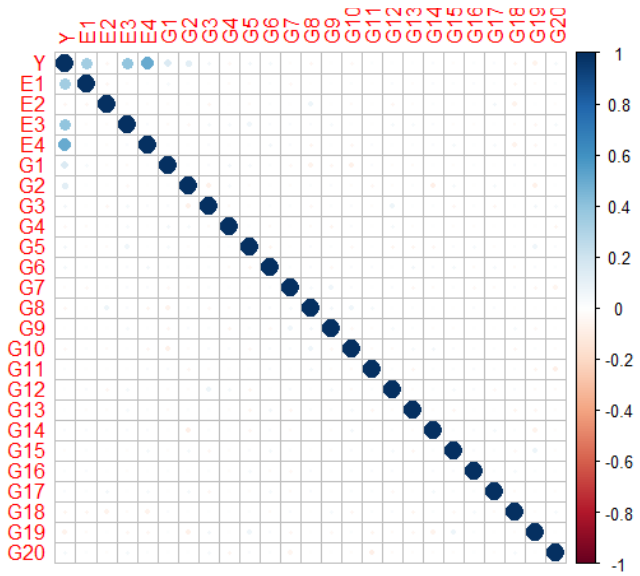
The purpose of this study is to create a model from synthetic data based off the psychological study of Caspi et al. (2003), where they measured the risk of depression from the short and long alleles of the 5HTT transporter gene and from major life events. The synthetic data consists of a response variable Y along with 4 numerical and 20 binary explanatory variables. In the scope of Caspi et al.'s study, the numerical variables represent environmental factors, and the binary variables represent genetic factors for the measure of the risk of depression, Y. As I analyze the data, I will also consider the question of finding interactions between or within the environmental and genetic variables when developing our model.

Methods

In the data, there was no missing values. The dataset consists of 1,143 entries with 25 total columns: one response variable Y, 4 Environmental variables, and 20 genetic variables. Significant variables are determined by looking at the summaries for the first and second order linear models. Then, a Box-Cox transformation is used to determine the transformation needed for the Y variable. With the transformation, forward and backward model selection methods are used to find the candidate variables for our true model. Then by looking at those variables' significance in the main models, the best model for the data is determined.

Results

The adjusted r squared value for the linear model using all variables is 0.5566, and the adjusted r square for using only the environmental data is 0.5158. The value is lower and suggests the need to include genetic variables.



Looking at the correlation matrix, there is a significantly large correlation between the Y variables and the E1, E3, E4, G1, and G2 variables. Using only those variables for the model, the adjusted

r squared value improved to 0.5557. Using the Box-Cox transformation, the likelihood of lambda is highest at around the value of 2, which suggests that a transformation of Y to Y^2.

With the help of the method used by Songzhu Zheng to create tabular output, I created a function to display the results of the 'regsubsets' function from the leaps package in R. Using this function, models of up to 6 variables are found, with first and second order interactions, and using forward and backward methods for each. The one with the highest adjusted r square is the transformed second order model with backward selection:

Model: $Y^{2\sim}$2 , method: backward

model	adjR2	BIC
(Intercept)+E4	0.2524	-319.3956
(Intercept)+E3+E4	0.3941	-553.6424
(Intercept)+E1+E3+E4	0.5181	-809.2953
(Intercept)+E1+E3+E4+G1	0.5389	-853.6904
(Intercept)+E1+E3+E4+G1+E2:G2	0.5582	-896.4729
(Intercept)+E1+E3+E4+G1+E2:G2+G4:G9	0.5652	-908.697

The Bayesian Information Criterion (BIC) did not improve much between the model with 5 variables and the model with 6 variables, so the former is used to determine candidate variables: E1, E3, E4, G1, E2, and G2. In the transformed model with only first order interactions and all variables, it is shown that E2 is not significant (even at 0.1 level) and therefore very unlikely to be in the true model. Also, by looking at the transformed model with second order interactions, no terms are significant at the 0.001 level, which suggests that it is unnecessary to include interaction terms.

By dropping the E2 variable and looking only at first order interactions, the best model is:

$$Y^2 = \beta_0 + \beta_1 E1 + \beta_2 E3 + \beta_3 E4 + \beta_4 G1 + \beta_5 G2$$

This can be supported by looking at the stepwise selection table for transformed Y and only first order interaction (the table for forward selection is the same):

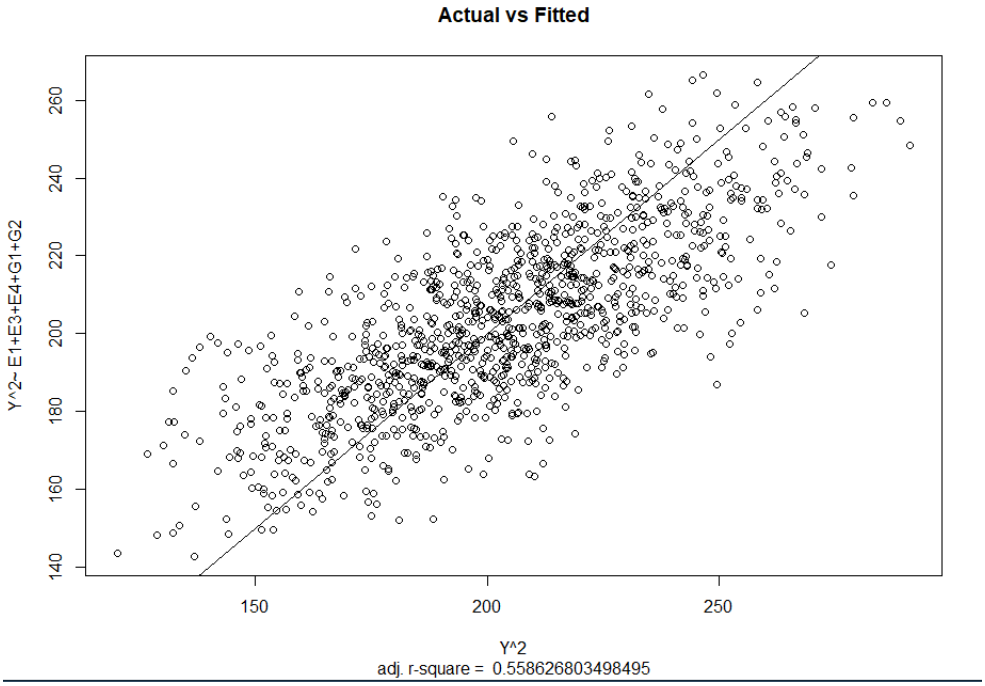
Model: Y^2~. , method: backward

model	adjR2	BIC
(Intercept)+E4	0.2524	-319.3956
(Intercept)+E3+E4	0.3941	-553.6424
(Intercept)+E1+E3+E4	0.5181	-809.2953
(Intercept)+E1+E3+E4+G1	0.5389	-853.6904
(Intercept)+E1+E3+E4+G1+G2	0.5586	-897.586
(Intercept)+E1+E3+E4+G1+G2+G4	0.5614	-898.6644

Here the model with 6 variables is not considered since the BIC barely improves and G4 is not significant at the 0.001 level. Also, the model with 5 variables has a higher adjusted r square (0.5586) than that with second order interaction (0.5582) from the previous table.

Discussion and Conclusions

What I found was that using second order interactions did not improve the r squared value, and that the best model with significant variables was $Y^2 = \beta_0 + \beta_1 E1 + \beta_2 E3 + \beta_3 E4 + \beta_4 G1 + \beta_5 G2$, with coefficients 5.192, 7.181, 8.319, 10.273, 10.181, and 9.868 respectively and an adjusted r squared of 0.5586. In the graph below it can be seen that the fitted and actual values of this model do not behave abnormally. The line in the graph is y = x.



One limitation to my analysis is that I did not analyze third order interactions within my dataset. The issue is that there would be 25*24*23 = 13800 terms to consider, which is more than the number of rows in the data. This introduces the possibility of overfitting and non-unique solutions for a model.

In the lens of the study by Caspi et al. (2003), the likelihood of depression is increased by the increase of the E1, E3, and E4 environmental variables and by the presence of the G1 and G2 genetic variables.

Reference:

Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington H, McClay J, Mill J, Martin J, Braithwaite A, Poulton R. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science. 2003 Jul 18;301(5631):386-9. doi: 10.1126/science.1083968. PMID: 12869766.