AMS 315 - Project 02

TUAN BUI

SBU ID: 113141951

11/28/2021

Report

Introduction:

The objective is to find the correlation between the dependent variable and independent variables, and to determine the correct regression model. The project is about find the interactions between gene and environment with the provided dataset. In this report, I am using the R statistical package to conduct the necessary functions to draw my conclusions.

Methods:

The original datasets were supplied with a dataset in .csv file. There are one dependent variable, labeled Y, four independently environmental variables, labeled E1 to E4, and twenty independently genetic variables, labeled G1 to G20. There are totally 1001 observations in the given dataset. I used the "cor" function in R statistic package to calculate the correlation between the dependent variable and independent variables. The result of the correlation calculation attached in the Result session. I used the "lm" function in R statistic package to fit a regression model between the dependent variable and the independently environmental variables. I also assumed for this project that I had up to 2 order interactions, then fitted the regression model between the dependent variable with all the independent variables by using "lm" function in R statistic package. My next task was to create and examine the residual plot. I saw a slight pattern in this residual plot. Therefore, I used the Box-Cox transformation to transform my dependent variable that had apparently homoscedastic residuals by using the "boxcox" function from MASS package in

R. After transformation, I used the “regsubsets” function from leaps package in R to define the stepwise regression model.

Results:

The adjusted R-square in the regression model with only environmental variables is

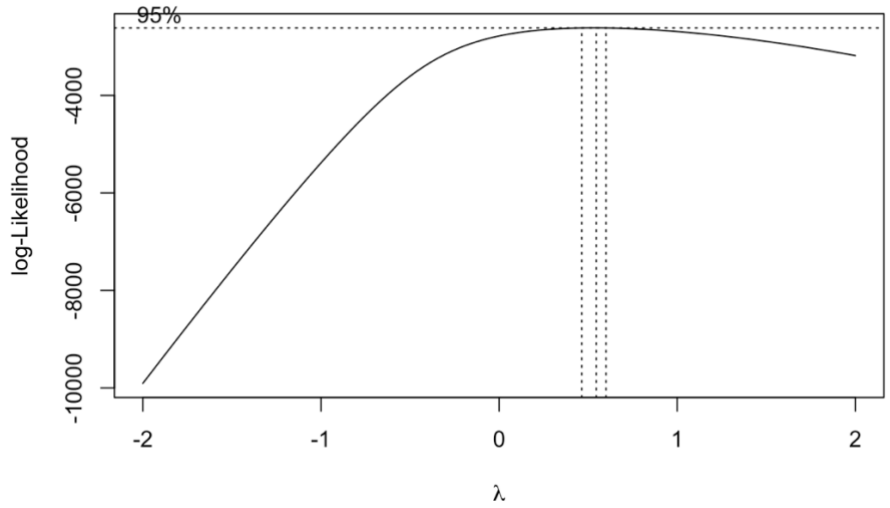
The correlation between the dependent variable and independent variables is low.

Correlation table between dependent variable, Y,

and independent variables, E1 to E4 and G1 to G20.

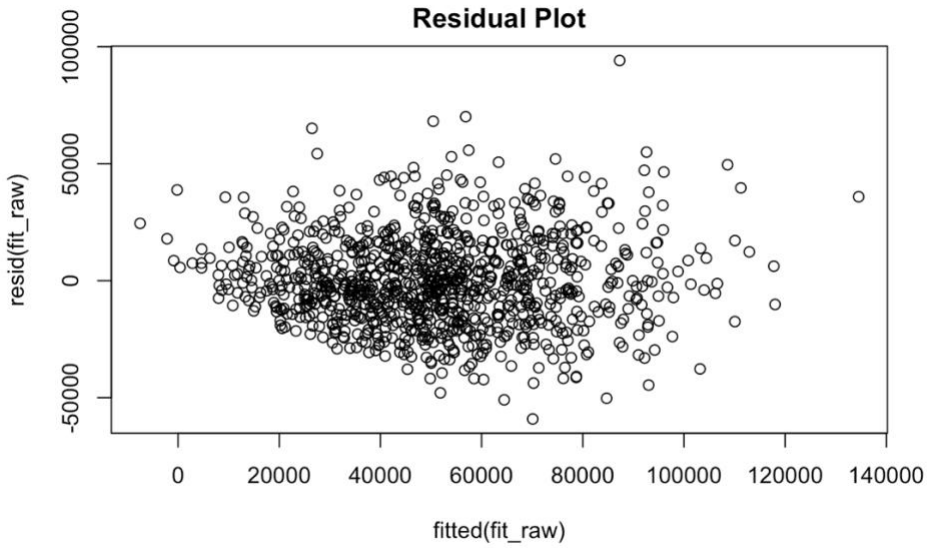
	Y
Y	1
E1	0.43903107
E2	0.02477005
E3	0.00832269
E4	0.43784038
G1	-0.0400559
G2	0.03024168
G3	0.01452073
G4	-0.079021
G5	0.00211275
G6	0.13086358
G7	0.02413605
G8	0.0255381
G9	-0.0002052
G10	0.0134103
G11	0.02414156
G12	-0.0270083
G13	0.02148256
G14	-0.0446983
G15	-0.0117108
G16	0.1046745
G17	0.01230418
G18	-0.011041
G19	0.00789784
G20	-0.0230328

The boxcox of the regression model with all independent variable

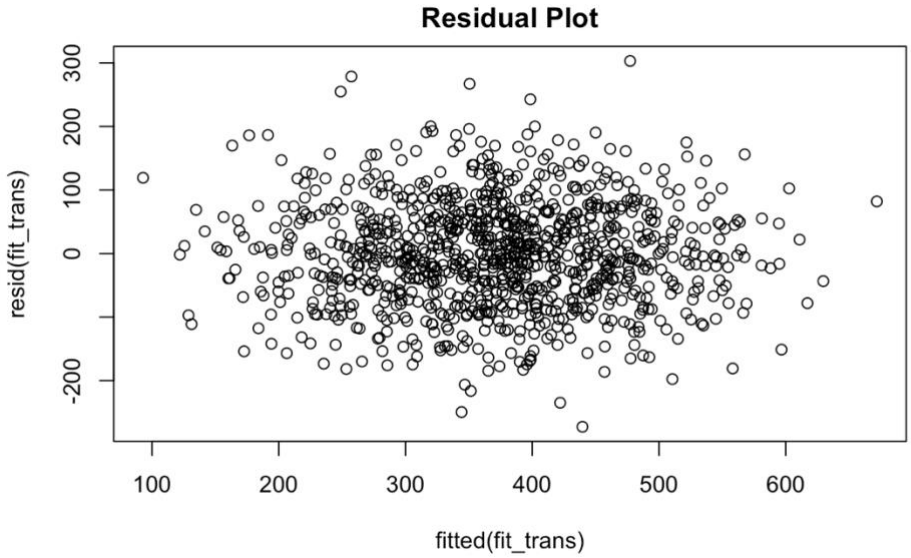


My estimated lambda was 0.55, so I applied the transformation of the dependent variable with an exponent of 0.55. My new adjusted R-square after transformation was 0.3797804, which was 0.0146336 higher than the original adjusted R-square of the regression with all independent variables, and 0.0122638 higher than the original adjusted R-square of the regression with only environmental variables.

The residual plot of the regression model with all independent variables before transformation



The residual plot of the regression model with all independent variables after transformation



Stepwise regression model summary

Model	Adjusted R-square	BIC
(Intercept)+E1:E4	0.366199718403503	-443.661400089459
(Intercept)+E1:E4+G6:G16	0.387215331138535	-471.509161059619
(Intercept)+G4+E1:E4+G6:G16	0.389863563271667	-469.939247148818
(Intercept)+G4+E1:E4+G6:G16+G10:G16	0.391916981402857	-467.409561976871
(Intercept)+G4+E1:E4+E1:G6+G6:G16+G10:G16	0.39498745394734	-466.573615662551

I saw that the adjusted R-square in the 4th model was slightly increase from the 3rd model, which might not be significant. Moreover, the different between the BIC value in model 3rd and 4th was much smaller than the other model, expected the 5th. Therefore, I chose the model 3 as candidates, namely G4, E1, E4, G6, G16. The final model is

$$Y = \beta_0 + \beta_1 G4 + \beta_2 E1E4 + \beta_3 G6G16 + \epsilon$$

Conclusion:

The final model is $Y = \beta_0 + \beta_1 G4 + \beta_2 E1E4 + \beta_3 G6G16 + \epsilon$. The adjusted R-square of the final model is 0.4111