Project 2

Introduction

In the data set, it has 4 environmental (E) and 20 genetic (G) variables. The genetic variables are represented as 0 and 1. The purpose of project 2 is to find out the association between the outcome values, E, and G. To find out the association, I will find the equation of the relationship between those variables.

Methods

After converting the CSV file, I named it Project2. I analyzed the data set to find the model with E variables by using lm(Y ~ E1 + E2 + E3 + E4, data = Project2) and named it as Enviroment_Model. I could get the r squared value of the Enviroment_Model by using summary ()$adj.r.squared, and I got 0.5136595. After that, I figured out the contribution of G values with the E values by using lm() with the square of all E and G values and named it Model_raw. And to see it easily, I plotted the Model_raw data with plot(resid (Model_raw) ~ fitted(Model_raw), main = 'Residual Plot') function. To get a transformation, I used boxcox() with Model_raw in the MASS library, and I got a 0.9 value. So, I implied the value to get a new transformed data set, Model_trans, by using the same function with the Model_raw but changing the Y value as Y^.9. I checked the r squared value of the transformed data set with summary()$adj.r.square function with Model_trans, and used a plot to see the data set easily by using the plot() function with Model_trans data. So, I got a 0.5145507 r squared value. After that, I used the regsubsets() function with Model_trans data in the 'leaps' library to perform stepwise regression and named the data Model. I saved the summary of the Model as a temp. To extract the x column of Model_trans data, I used the colnames() function and named the data as Var. And I used the apply() function with temp and Var[x] data and named the data as Model_select. Also, I used kable() function with Model_select to find an obvious increase in adjR^2. After finding the obvious increase, I made a table with the values which have an obvious increase by using the kable() function. Besides this, to find the other main effects in the data set, I made Model_main data with lm(I(Y^.9) ~ E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16+G17+G18+G19+G20, data = Project2) function. And I saved the summary(Model_main) in the temp and used the kable() function with 4 coefficients. According to the previous step, I could observe 3 variables that I would use as candidate variables and made 2nd stage model based on the observation. I named it as Model_2stage with lm(I(Y^.9) ~ (E3+E4+G16)^2, data = Project2) function and saved the summary() of Model_2stage in the temp. After that, I used the kable() function with 3 coefficients. Finally, I got the final model based on the previous step and made Model_final with lm(I(Y^.9) ~ E3 + E4, data = Project2) function.

Results

In the data file, I had 1002 observations with 4 environmental variables and 20 gene variables. After I used the boxcox() function with the data, I got the Figure 1 plot, and the estimated λ was about 0.9 from the plot. I applied the λ to get transformed data and got the Figure 2. The r square value of raw data was 0.5136595 and the value of transformed data was 0.5145507. After I got the transformed data, I used the stepwise regression method to select major independent variables. I used the kable() function with the transformed data and got Table 1. Based on the table, I chose the 2nd model as candidates; E3 and E4 because the difference between the 1st and 2nd model is the biggest one. Besides that, I wanted to make sure of the other main effects in the data. So, I used the kable() function based on the main model data with 4 coefficients and got Table 2. Interestingly, the second table contained the G16 variable. Therefore, I made Model_2stage with the G16, E3, and E4 and made Table 3 by using the kable() function based on the Model_2stage with 3 coefficients. More interestingly, the G16 variable was dropped again in the table. Thus, I made final model with E3 and E4, and the final model is: $Y^{0.9} = \beta_0 + \beta_1 E_3 + \beta_2 E_4 + \varepsilon$. In addition, the p-value of the final model is 2.2e-16 which much lower than 0.01.

Conclusion and Discussion

My data set has originally low r squared value even though I used Box-Cox transformation. If I got a higher r squared value from the raw data, I think it would have been a more interesting data analysis.

Appendix

Codes

```
Project2 <- read.csv('378390_project2.csv', header=TRUE)
Enviroment_Model <- lm(Y ~ E1 + E2 + E3 + E4, data = Project2)
summary(Enviroment_Model)
summary(Enviroment_Model)$adj.r.squared
Model_raw <- lm(Y ~
(E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16
+G17+G18+G19+G20)^2, data = Project2)
View(Model_raw)
plot(resid(Model_raw) ~ fitted(Model_raw), main = 'Residual Plot')
library(MASS)
boxcox(Model_raw)
Model_trans <- lm( I(Y^.9) ~
(E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16
+G17+G18+G19+G20)^2, data = Project2)
summary(Model_raw)$adj.r.square
summary(Model_trans)$adj.r.square
plot(resid(Model_trans) ~ fitted(Model_trans), main = 'Transfomed Residual Plot')
install.packages("leaps")
library(leaps)
Model <- regsubsets( model.matrix(Model_trans)[,-1], I((Project2$Y)^.9), nbset = 1, nvmax = 5,
method = 'forward', intercept = TRUE)
temp <- summary(Model)
library("knitr")
Var <- colnames(model.matrix(Model_trans))
Model_select <- apply(temp$which, 1, function(x) paste0(Var[x], collapse = '+'))
kable(data.frame(cbind(model = Model_select, adjR2 = temp$adjr2, BIC = temp$bic)), caption =
'Model Summary')
Model_main <- lm( I(Y^.9) ~
E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16+
G17+G18+G19+G20, data = Project2)
temp <- summary(Model_main)
kable(temp$coefficients[abs(temp$coefficients[,4]) <= 0.01, ], caption = 'Sig Coefficients')
Model_2stage <- lm( I(Y^.9) ~ (E3+E4+G16)^2, data = Project2)
temp <- summary(Model_2stage)
kable(temp$coefficients[ abs(temp$coefficients[, 3]) >= 4, ])
Model_fianl <- lm(I(Y^.9) ~ E3 + E4, data = Project2)
summary(Model_fianl)
```

I refer this site for the codes:
https://blackboard.stonybrook.edu/bbcswebdav/pid-1724221-dt-content-rid-13925836_1/courses/1224-AMS-315-SEC01-48518/AMS-315-Multiple-Regression-Handout-Updated-S22.html

Figures

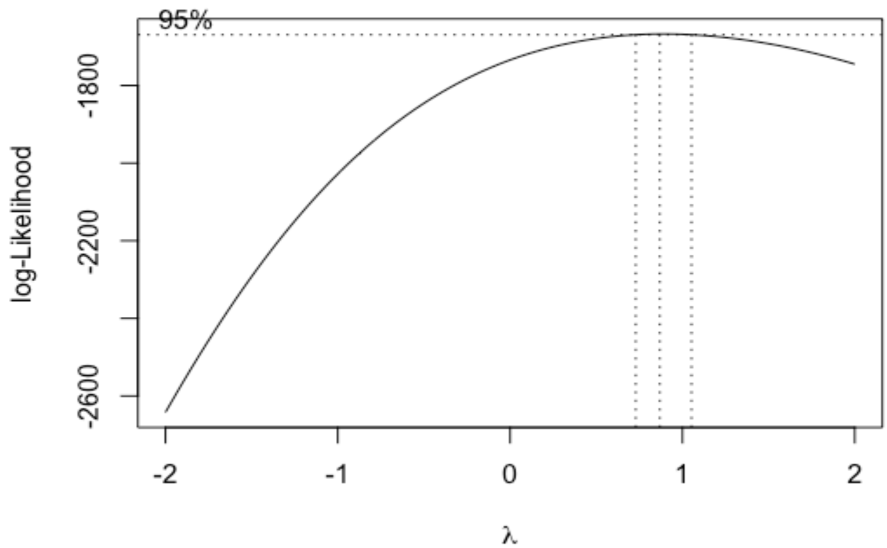

Figure 1

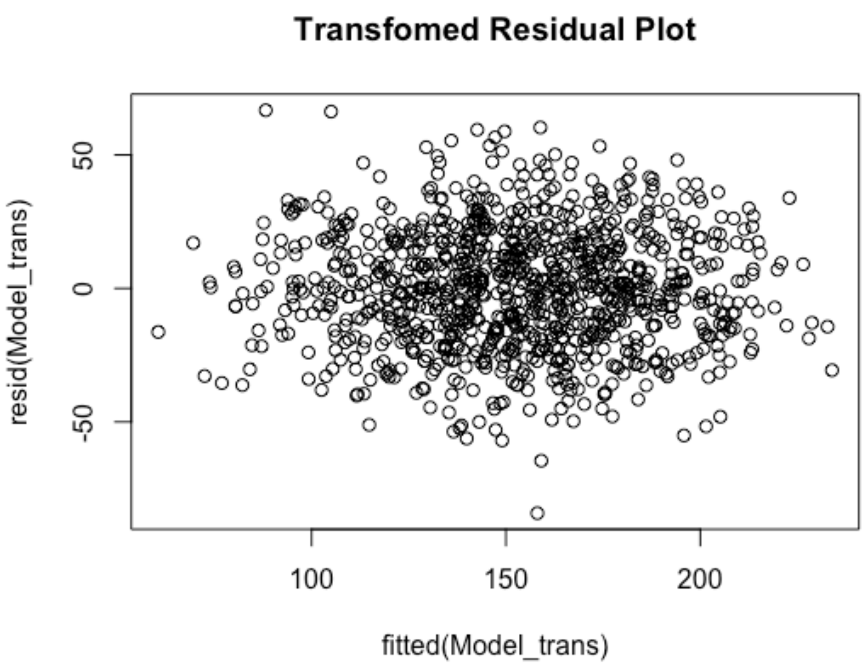

Figure 2

Tables

Table 1

Model Summary		
Model	adjR2	BIC
(Intercept)+E3:E4	0.476050599242237	-634.834875859014
(Intercept)+E4+E3:E4	0.503917563484779	-683.689930565582
(Intercept)+E3+E4+E3:E4	0.513998821498067	-698.355811527038
(Intercept)+E3+E4+E1:G16+E3:E4	0.524803840846936	-714.978899707699
(Intercept)+E3+E4+E1:G16+E3:E4+G9:G11	0.527380642461561	-714.522902396564

Table 2

Sig Coefficients				
	Estimate	Std. Error	t value	Pr (>|t|)
(Intercept)	17.635997	6.7400449	2.616599	0.0090182
E3	5.251617	0.2954004	17.777963	0.0000000
E4	7.884195	0.2841125	27.750258	0.0000000
G16	8.515699	1.7916282	4.753050	0.0000023

Table 3

	Estimate	Std. Error	t value	Pr (>|t|)
E3	5.271140	1.051350	5.013688	6e-07
E4	8.013501	1.047197	7.652334	0e+00