Asad Imam

Dr. Stephen Finch

AMS 315 – Data Analysis

05/10/19

Introduction

The objective of this study is to find the model used to generate the given data. The database was given in a single .csv file containing the data. This file contained one dependent variable and twenty-four independent variables. The value of the dependent variable was shown in the Y column (the left most column). The values of the twenty-four dependent variables (which included 4 environmental variables and 20 genetic variables) were presented in columns E1 to E4 and G1 to G20. The data file contained no missing values. The TA's used a simulation program to generate this data.

Methodology

To find the correlation between the independent variables and dependent variable, the statistical package R was used. The data was imported into RStudio using the *import* function provided by the library "rio". To analyze the data, the environmental variables were modelled using the function lm () and the summary was used to explain the outcome. The adjusted R^2 for this model was found to be 0.4995314.

The contribution of the genetic variables after controlling the environmental variables was then assessed using the functions lm () and plot. It was found that raising the sum of the environmental variables and the genetic variables to the second power produced the best residual plot. To find the potential nonlinear transformations of a dependent variable, the Box-Cox transformation provided by the MASS library was used. The adjusted R^2 for the data before the transformation was found to be 0.5050352. The adjusted R^2 after the transformation was found to be 0.5050352. Then the New Residual Plot was plotted.

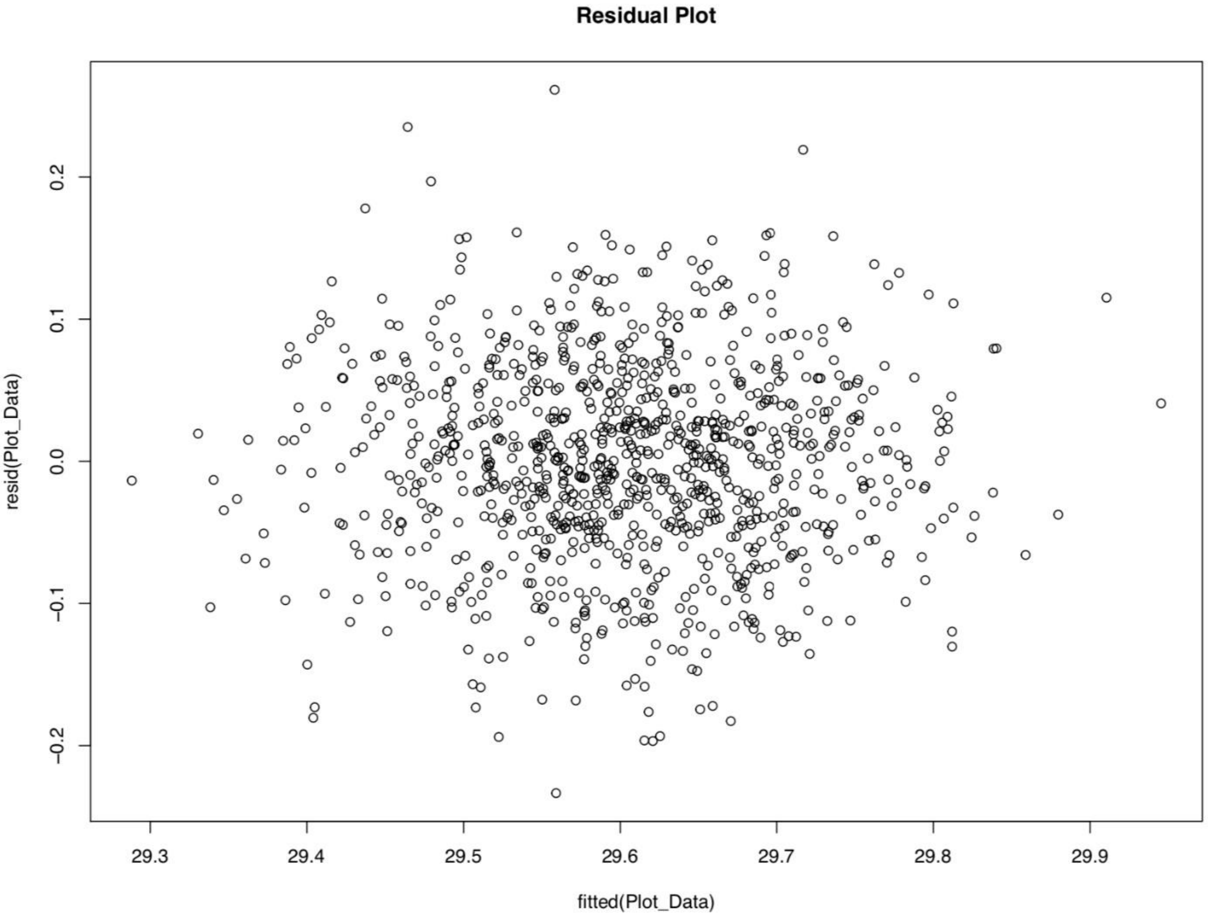
To select the significant independent variables Stepwise Regression technique was used. The function regsubset () provided by the library leaps was used to perform the stepwise regression. Then the function kable () provided by the knitr library was used to find the Model Summary. To assess this model, we used Bayesian Information Criterion (BIC). To find the effects of the most significant variables, we used the (*). Then the effects of the second order interactions were considered by raising the (.) to the second power.

```
(*)
Data_main <- lm(I(log(Y)) ~., data=Var_Import_Data)
temp <- summary(Data_main)
kable (temp$coefficients [ abs (temp$coefficients [,4]) <= 0.001,],
caption='Sig Coefficients')
```

Finally, the most significant independent variables namely E1, E2 and E4 were modelled using a similar code as (*).

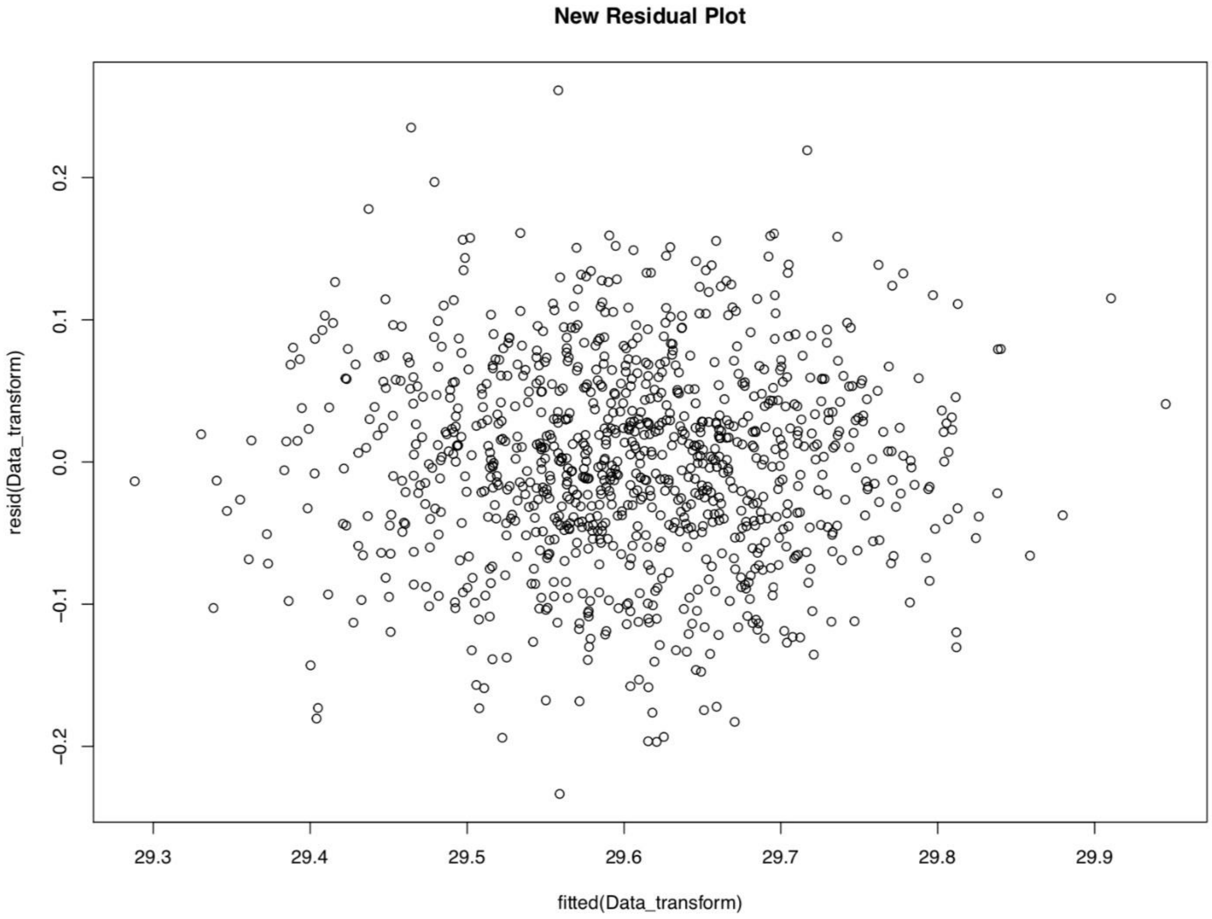
Results

The objective of this simulated study was to find the model used to generate the given data. The adjusted R^2 for the Environmental Variables was found to be 0.4995314. The adjusted R^2 of the genetic variables after controlling the environmental variables was found to be 0.5050352. This shows that by adding the genetic variable the adjusted R^2 value increased. The plotted result is shown below



This model appears adequate in that most of the data in centered in the middle with a few outliers at the extremes.

The adjusted R^2 after the Box- Cox transformation was found to be 0.5050352. The New Residual Plot after this transformation is shown below



The New Residual Plot appears to be identical to the original Residual Plot with no difference between the adjusted R^2 values. This shows that there is no notable change between the genetic variables after the logarithmic transformation of the outcome variable to the original genetic variable.

The Model Summary after doing the Stepwise Regression to find important independent variable is shown below

MODEL SUMMARY

Model	Adjusted R^2	BIC
$(Intercept) + E2{:}\,E4$	0.415833427381835	-524.755101516648
(Intercept)+E1:E4+E2:E4	0.499093760050405	-672.617084070194
(Intercept)+E1:E4+E2:E4+G1:G10	0.502124205865152	-672.781142380976
(Intercept)+E1:E4+E2:E4+G1:G10+G11:G16	0.50376686666851	-670.182700902286
(Intercept)+E1:E4+E2:E4+G1:G10 + G4:G19+G11:G16	0.505360517631481	-667.497140531699

There is a significant increase in R_a^2 from the 1st model to the 2nd model. There is only a very small increase in the models following the 2nd, this indicates insignificant changes. Similarly, by looking at the Bayesian Information Criterion (BIC) we can see that there is a significant decrease in the BIC value from the 1st model to the 2nd model. Therefore, the 2nd model was selected as candidate: namely E1, E2 and E4.

The significant coefficients for the model are given below

Significant Coefficients

	Estimate	Std. Error	T value	Pr$(> \lvert t \rvert)$
(Intercept)	2.9725743	0.0163002	182.36444	0
E1	0.0009054	0.0000813	11.13969	0
E2	0.0012734	0.0000814	15.65109	0
E4	0.0020152	0.0000812	24.81668	0

All variables of the 2nd model has a significant main effect.

The second order interactions by using the second power in the model request had the same variables as the significant coefficients table.

Finally, the summary for the candidate variable E1, E2 and E4 is given below

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.9725742809	1.630019e-02	182.36444	0.000000e+00
E1	0.0009054498	8.128140e-05	11.13969	3.258731e-27
E2	0.0012734329	8.136385e-05	15.65109	2.000157e-49
E4	0.0020152096	8.120382e-05	24.81668	9.032847e-106

Conclusions and Discussions

It was found that the model that the TA's used to generate this function was:

$$ln(y) = 2.9725742809 + 0.0009054498E1 + 0.0012734329E2 + 0.0020152096E4$$

These variables were chosen because their adjusted R^2 values had the biggest change of any other model. Also, because their T-values were bigger than 4 and their P-values were significant.

References

https://blackboard.stonybrook.edu/bbcswebdav/pid-4795848-dt-content-rid-34989443_1/courses/1194-AMS-315-SEC01-49559/Multiple_Regression_Handout_S2019_Blackboard.html

Technical Appendix

My Code:

```
# this is the second project for AMS 315  by Asad Imam
# initializing/installing the libraries
install.packages("rio")
library(rio)
library(knitr)
library(MASS)

# importing the data
Var_Import_Data <- import("P2 95732.CSV")

names(Var_Import_Data) <- c('Y', paste0('E', 1:4), paste0('G', 1:20))
#Modeling the Environmental Variables
Model_Environmental <- lm(Y ~ E1+E2+E3+E4, data=Var_Import_Data)
summary(Model_Environmental)
summary(Model_Environmental)$adj.r.squared

#Modeling the Additional Contribution of the Genetic Variables
Plot_Data <- 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=Var_Import_Data )
#Plotting the Data
plot(resid(Plot_Data) ~ fitted(Plot_Data), main='Residual Plot')

#BOX-COX Transformation
boxcox(Plot_Data)
Data_transform <- lm( Y ~ (.)^2, data=Var_Import_Data )
summary(Plot_Data)$adj.r.square
summary(Data_transform)$adj.r.square
plot(resid(Data_transform) ~ fitted(Data_transform), main='New Residual Plot')

#Stepwise Regression
install.packages("leaps")
library(leaps)
s_reg <- regsubsets( model.matrix(Data_transform)[,-1], I(Var_Import_Data$Y),
                nbest = 1 , nvmax=5,
                method = 'forward', intercept = TRUE )
temp <- summary(s_reg)

#The software produces a proposed model :
Var <- colnames(model.matrix(Data_transform))
Data_select <- apply(temp$which, 1,
               function(x) paste0(Var[x], collapse='+'))
kable(data.frame(cbind( model = Data_select, adjR2 = temp$adjr2, BIC = temp$bic)),
    caption='Model Summary')

# Significant Coefficients
Data_main <- lm( I(log(Y)) ~ ., data=Var_Import_Data)
# . here means include all variable from E1 to E5 and from G1 to G15 to the model
temp <- summary(Data_main)
kable(temp$coefficients[ abs(temp$coefficients[,4]) <= 0.001, ], caption='Sig Coefficients')

# Second Order Interactions
Data_2nd <- lm( I(log(Y)) ~ (.)^2, data=Var_Import_Data)
temp1  <- summary(Data_2nd)
kable(temp1$coefficients[ abs(temp1$coefficients[,4]) <= 0.001, ], caption='2nd Interaction')

#Final Step
Data_2stage <- lm( I(log(Y)) ~ (E1+E2+E4)^3, data=Var_Import_Data)
temp2 <- summary(Data_2stage)
temp$coefficients[ abs(temp$coefficients[,3]) >= 4, ]
```