STEPWISE REGRESSION ANALYSIS

Caspi et. al. used a regression framework to test any associations between depression and various other factors in different age groups by looking for a gene-by-environment (G x E) interaction. This simulated project used 4 environmental (E) variables and 20 genetic (G) variables. We are using a regression analysis to see if there is any association between the outcome and genetic variables. Using a multiple regression analysis, I found that the model used to generate the sample data for this project is:

$$Y^{3/2} = \beta_0 + \beta_1 + \beta_2 E_2 + \beta_3 E_4 + \epsilon$$

First, we found the environmental model with interaction terms up to the 2nd order to explain the outcome of the results. Next, we created a model for the genetic variables, while holding the environmental variables constant. Since the raw model seemed inadequate (that is, the variances of the residuals were different across most or all elements), a transformation of the independent variable was required. Thus, applying a transformation of Y using a Box-Cox Transformation from the MASS package in R (Figure 7), the estimated $\lambda = 1.5$ inferred that for the sample data, we should apply a cubed-root transformation to the response variable. Lastly, utilizing stepwise regression, we simplified the model further to pinpoint the exact variables that the models used. Using this model, we were able to conclude that the environmental factors, namely E2 and E4, were the most influential in our findings compared to the genetic factors. We found that there was little association between the outcome variables with the genetic variables while holding the environmental variables constant. Specifically, as mentioned above, the model used was $Y^{3/2} = \beta_0 + \beta_1 + \beta_2 E_2 + \beta_3 E_4 + \epsilon$.

The found adjusted R^2 of the environmental variables was 0.5195, while the found adjusted R^2 of the genetic variables was 0.5206. For this sample data, we chose terms that are significant above the 0.001 level for our coefficients. Our final fit looked at t-values larger than 3. The t-value we received were 3.171 and 4.637. Our analysis found that the genetic variables G1 through G20 were not used in constructing the model. Some limitations include stating only the estimates of variance explained and using a regression model for global tests.

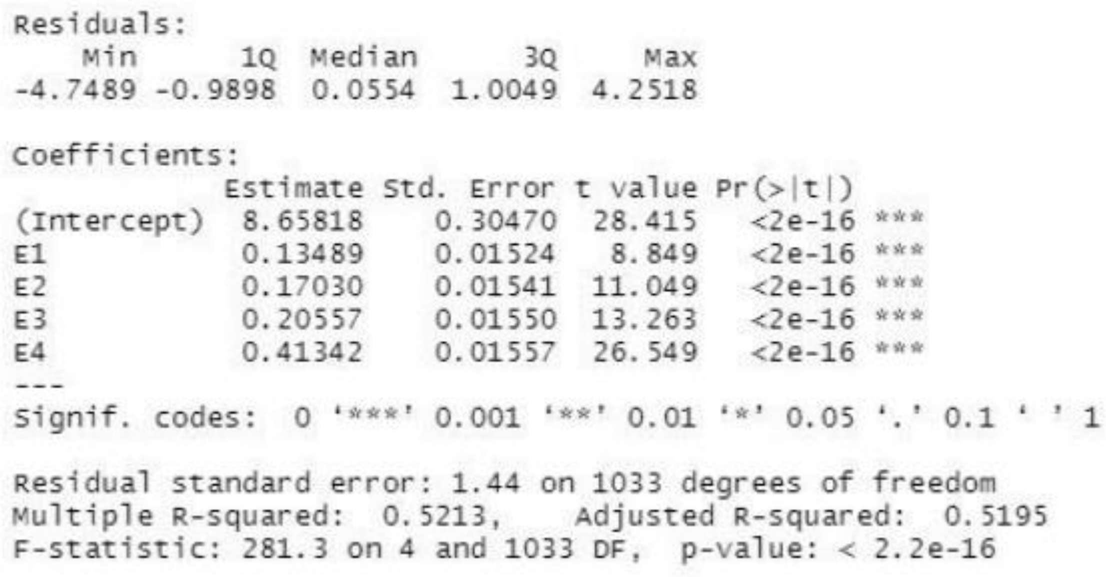
ENVIRONMENTAL AND RAW MODELING

```
getwd()
workdir <- "C:/Users/16318/Documents/SBU Classes - Fall 2021/AMS 315 Project 2/Project 2"
setwd(workdir)
project2 <- read.csv('project2.csv', header=TRUE)
env_model <- lm(Y ~ E1 + E2 + E3 + E4, data = project2)
summary(env_model)
summary(env_model)$adj.r.squared
raw_model <- lm(Y ~
(E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16+G17+G18+G19+G20)^2, data =
project2)
plot(resid(raw_model) ~ fitted(raw_model), main = 'Residual Plot')
library(MASS)
boxcox(raw_model)
trans_model <- lm(I(Y ^ 1.5) ~
(E1+E2+E3+E4+G1+G2+G3+G4+G5+G6+G7+G8+G9+G10+G11+G12+G13+G14+G15+G16+G17+G18+G19+G20)^2,
data=project2)
plot(resid(transformed_model) ~ fitted(transformed_model), main='New Residual Plot')
plot(resid(trans_model) ~ fitted(trans_model), main='New Residual Plot')
summary(trans_model)$adj.r.square
summary(raw_model)$adj.r.square
```

STEPWISE REGRESSION

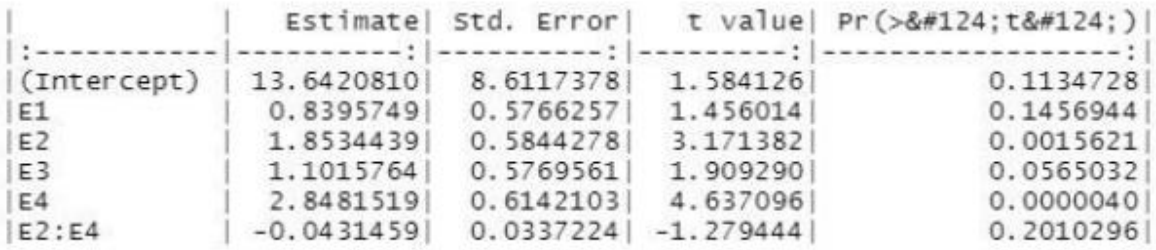
```
install.packages("leaps")
library(leaps)
model <- regsubsets( model.matrix(trans_model)[, -1], I((project2$Y)^1.5),
nbest = 1, nvmax = 5,
method = 'forward', intercept = TRUE)
summary(model)
temp <- summary(model)
install.packages("knitr")
library(knitr)
Variables <- colnames(model.matrix(trans_model))
model_select <- apply(temp$which, 1,
function(x) paste0(Variables[x], collapse='+'))
kable(data.frame(cbind( model = model_select, adjR2 = temp$adjr2, BIC = temp$bic)),
caption='Model Summary')
model_main <- lm( I(Y^1.5) ~
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]) <= 0.001, ], caption='Sig
Coefficients')
model_second <- lm( I(Y^1.5) ~ (EE2+E3+E4)^2, data=project2)
temp <- summary(model_second)
kable(temp$coefficients[ abs(temp$coefficients[,3]) >= 3, ])
```

RESIDUALS OF ENVIRONMENTAL MODEL



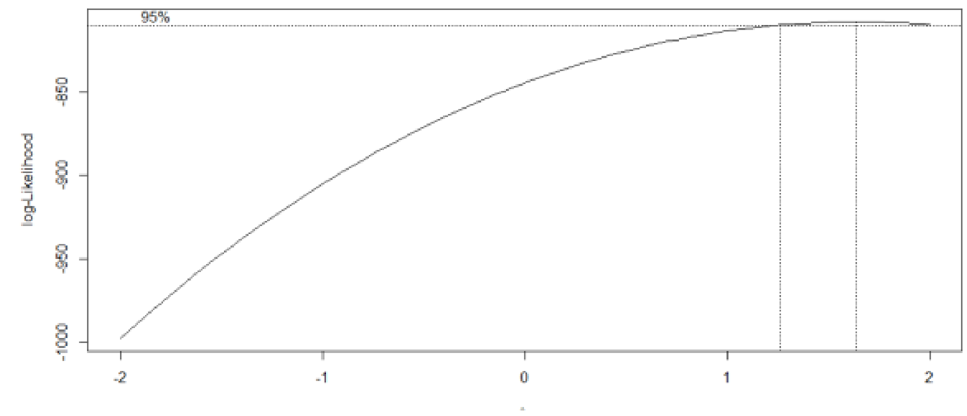
```
Residuals:
    Min      1Q  Median      3Q     Max 
-4.7489 -0.9898  0.0554  1.0049  4.2518 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.65818    0.30470  28.415   <2e-16 ***
E1           0.13489    0.01524   8.849   <2e-16 ***
E2           0.17030    0.01541  11.049   <2e-16 ***
E3           0.20557    0.01550  13.263   <2e-16 ***
E4           0.41342    0.01557  26.549   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.44 on 1033 degrees of freedom
Multiple R-squared:  0.5213,	Adjusted R-squared:  0.5195 
F-statistic: 281.3 on 4 and 1033 DF,  p-value: < 2.2e-16
```

SIGNIFICANT COEFFICIENTS

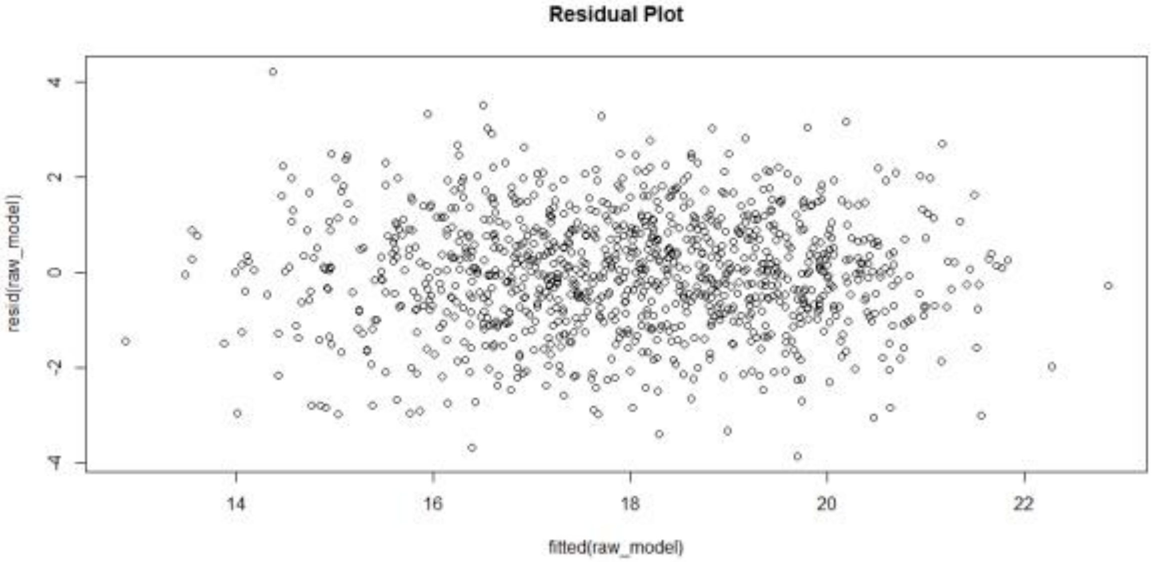


```
|            |   Estimate| Std. Error|   t value| Pr(>&#124;t&#124;)|
|:-----------|----------:|----------:|---------:|------------------:|
|(Intercept) | 13.6420810|  8.6117378|  1.584126|          0.1134728|
|E1          |  0.8395749|  0.5766257|  1.456014|          0.1456944|
|E2          |  1.8534439|  0.5844278|  3.171382|          0.0015621|
|E3          |  1.1015764|  0.5769561|  1.909290|          0.0565032|
|E4          |  2.8481519|  0.6142103|  4.637096|          0.0000040|
|E2:E4       | -0.0431459|  0.0337224| -1.279444|          0.2010296|
```

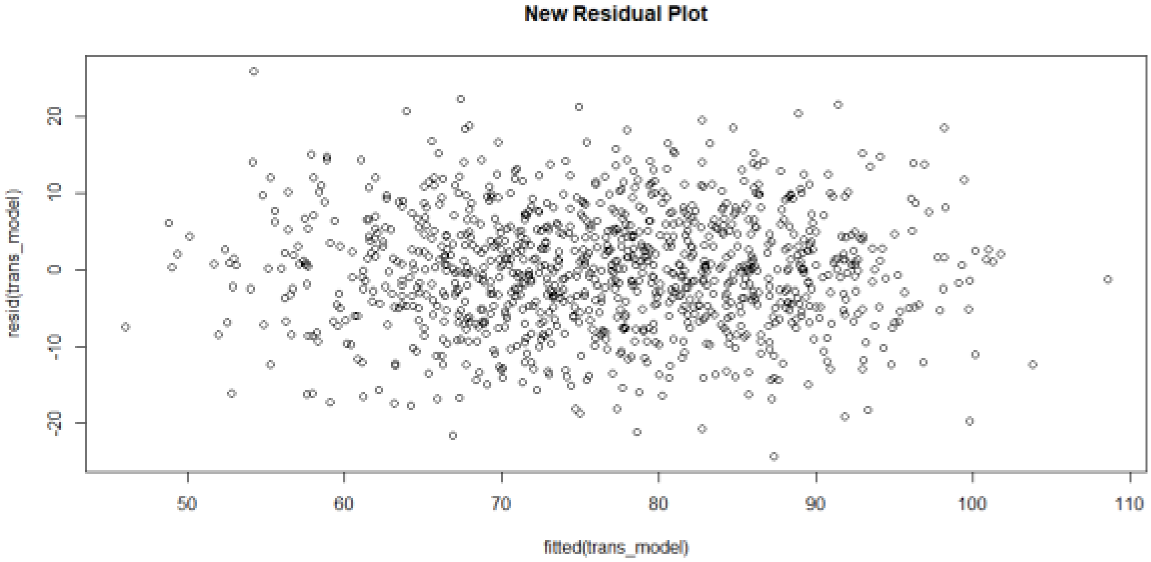
BOX COX TRANSFORMATION



OLD RESIDUAL PLOT



NEW RESIDUAL PLOT


New Residual Plot
resid(trans_model)
fitted(trans_model)