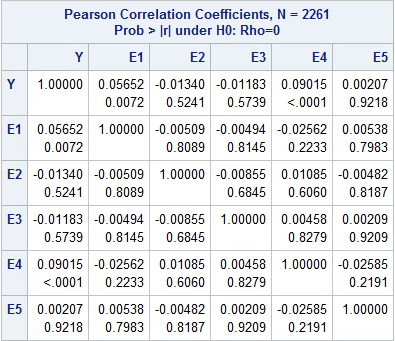
**Introduction**

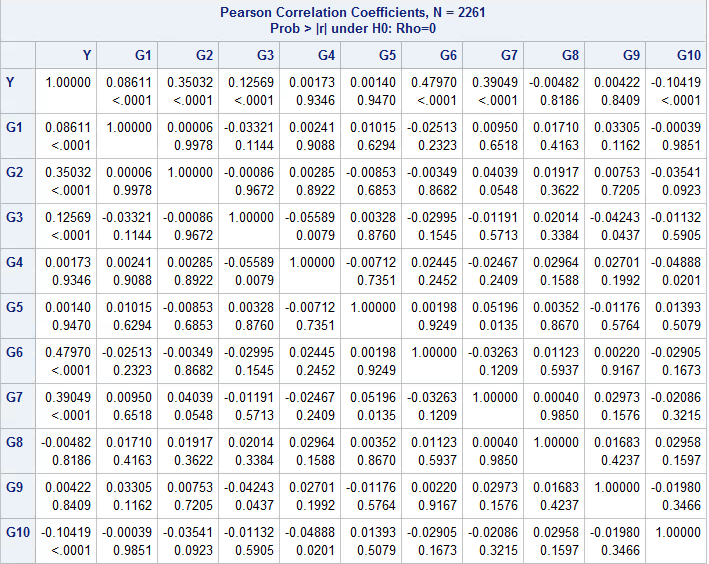
This report illustrated the process used in estimating the function through its use of given data. The information regarding the problem was provided in 1 data file, Group13.csv, with 16 variables and 2,261 observations. Amongst these variables was DV (the Dependent Variable), E1-E5 (Environmental variables 1-5), and G1-10 (Genes). The possible values for G1-G10 are 1 or 0, either the gene is present or not present. By definition, a gene-by-environment interaction is when an individual’s response to different environments affected by their DNA or genetic make-up. In order to find these interactions, studies need to be done using people with the genes that are in question and specific environments to determine if there is any correlation between the two.

**Methods**

**Correlation**:

The SAS code supplied by Professor Finch (Technical Appendix 1) was run to determine the correlation between the dependent variable and the independent variables. Two tables were generated by SAS(Technical Appendix 2), one concerning the correlation between DV and Environment Variables (E1, E2, E3, E4, E5), the other concerning the correlation between DV and the Gene Variables (G1, G2, G3, G4, G5, G6, G7, G8, G9, G10). The correlations between the dependent variable and a few independent variables(IV) were large enough to be significant, and only these IV’s were considered for our model. These IVs include the independent environmental variables E1 (rdv•e1 =0.05652), E4 (rdv•e4 =0.09015), While these numbers may appear to be too small to be correlated, when compared to the other environmental IV correlations they are actually quite large. In regards to the Gene Independent Variables, there are 6 with relatively large correlations, G1(rdv•g1= 0.08611) ,G2(rdv•g2=0.35032) , G3(rdv•g3 =0.12569), G6(rdv•g6 = 0.47970) , G7(rdv•g7 = 0.39049) and G10(rdv•g10 = -0.10419) . These six variables also need to be considered for the model. Levels of significance for each of the selected variables were also checked and found to be larger than .95, offering a conclusion that those variables would be important in the model.

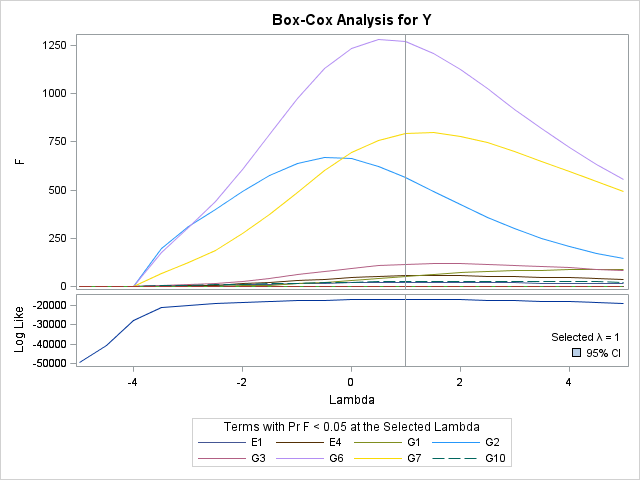




**Box Cox Transformation:**

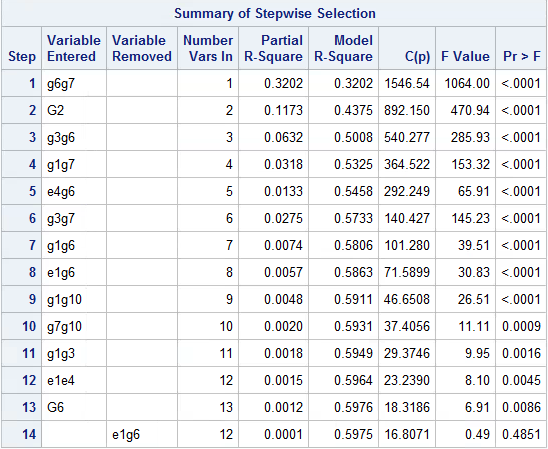
The Box Cox Transformation was then used to find potentially nonlinear transformations of the dependent variable. The graph shown in Figure 1 shows a regression of DV^λ. Actually the graph doesn’t peak at λ=1, we still determined that the transformation of the dependent variable is DV^1.

Figure 1:



**Stepwise Regression:**

After selecting the transformation of the dependent variable, the stepwise regression method was used to select the important independent variables. The Summary of Stepwise Selection is shown below. From this table it can be seen that, this model contains 12 independent variables, g6g7, G2, g3g6, g1g7, e4g6, g3g7, g1g6, g1g10, g7g10, g1g3, e1e4 and G6.



**Results**

The F-Test for this ANOVA can be said to be testing the null hypothesis that the coefficients of the model are equal to zero. The probability that we would get an extreme value of F if the null hypothesis was true is < 0.0001, thus we can reject the null hypothesis and say none of the model’s coefficients are zero. We can use fit diagnostics shown in Figure 1 aid in validating the model. The quantile vs. residual plot show the residuals are follow a normal distribution. The plot of predicted vs. actual value of Y shows a very linear correlation which means the model is close to the actual function.

Fit Diagnostics

Based on the analysis given above, the following model was obtained: DV^1=4836.2+2261.5G2+992.11G6+0.0003e1e4+0.2712e4g6-353.31g1g3+805.35g1g6+893.1g1g7-206.31g1g10+436.67g3g6+1592g3g7-419.08g7g10

The Analysis of Variance Table can be seen in Table 2. The value of the F-Test is equal to 276.81, which is very large, therefore the null hypothesis (that the constants of each variable are equal to 0) is rejected at every significance level.

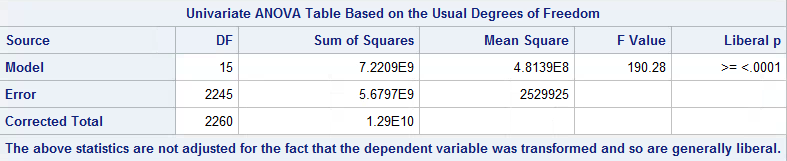
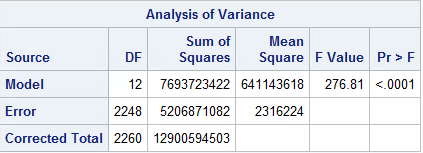


Table 2: 

**Conclusions and Discussions**

In conclusion, it was found that the generating model was: DV^1=4836.2+2261.5G2+992.11G6+0.0003e1e4+0.2712e4g6-353.31g1g3+805.35g1g6+893.1g1g7-206.31g1g10+436.67g3g6+1592g3g7-419.08g7g10. The final model found is the best fit found using the methods we used. The methods are not perfect; we did not explicitly explore the possibility of variables being independently transformed by different powers and operators as suggest in the original project description. However, the assumption that data was not transformed in this manner appears to have given a reasonable model.

Technical Appendix

1) SAS Code:

[Code Provided by Professor Finch]

/\* Importing the data\*/

**PROC** **IMPORT** OUT= WORK.Y

DATAFILE= "\\mysbfiles.campus.stonybrook.edu\~\Group13.csv"

DBMS=CSV REPLACE;

GETNAMES=YES;

DATAROW=**2**;

**RUN**;

/\* Proc Corr procedure is usually used for finding the correlation between varibles.\*/

**proc** **corr** data=y;

var DV E1-E5;

**run**;

**proc** **corr** data=y;

var DV G1-G10;

**run**;

/\*Proc Transreg procedure ﬁts linear models, optionally with spline and other nonlinear transformations, and it can be used to code experimental designs prior to their use in other analyses, especially Box-Cox transformations.\*/

**proc** **transreg** data=y ss2 detail;

model BoxCox(DV/lambda=-**3** to **3** by **0.5**)=identity (E1-E5 G1-G10);

output;

**run**;

/\*after selecting the necessary transformations, transform the dependent variable in the data step. \*/

**data** new;

set y;

Y=DV\*\***1**;/\*Here function of DV means a possible transformation of the original dependent variable, such as log(DV), exp(DV), sqrt(DV), DV^1, DV^2, DV^3, 1/sqrt(DV)\*/

**run**;

/\*Then we need to computer the two way interaction of the independent variables.\*/

**data** new1;

set new;

array one[\*] E1-E5 G1-G10;

array two[\*]

e1e2 e1e3 e1e4 e1e5 e1g1 e1g2 e1g3 e1g4 e1g5 e1g6 e1g7 e1g8 e1g9 e1g10

e2e3 e2e4 e2e5 e2g1 e2g2 e2g3 e2g4 e2g5 e2g6 e2g7 e2g8 e2g9 e2g10

e3e4 e3e5 e3g1 e3g2 e3g3 e3g4 e3g5 e3g6 e3g7 e3g8 e3g9 e3g10

e4e5 e4g1 e4g2 e4g3 e4g4 e4g5 e4g6 e4g7 e4g8 e4g9 e4g10

e5g1 e5g2 e5g3 e5g4 e5g5 e5g6 e5g7 e5g8 e5g9 e5g10

g1g2 g1g3 g1g4 g1g5 g1g6 g1g7 g1g8 g1g9 g1g10

g2g3 g2g4 g2g5 g2g6 g2g7 g2g8 g2g9 g2g10

g3g4 g3g5 g3g6 g3g7 g3g8 g3g9 g3g10

g4g5 g4g6 g4g7 g4g8 g4g9 g4g10

g5g6 g5g7 g5g8 g5g9 g5g10

g6g7 g6g8 g6g9 g6g10

g7g8 g7g9 g7g10

g8g9 g8g10

g9g10

;

n=**0**;

do i=**1** to dim(one);

do j=i+**1** to dim(one);

n=n+**1**;

two(n)=one(i)\*one(j);

end;

end;

**run**;

/\*Then we use the stepwise option in SAS procedure Proc Reg to select the reasonable independent variables at significance level of 0.01\*/

**proc** **reg** data=new1;

model Y= E1-E5 G1-G10

e1e2 e1e3 e1e4 e1e5 e1g1 e1g2 e1g3 e1g4 e1g5 e1g6 e1g7 e1g8 e1g9 e1g10

e2e3 e2e4 e2e5 e2g1 e2g2 e2g3 e2g4 e2g5 e2g6 e2g7 e2g8 e2g9 e2g10

e3e4 e3e5 e3g1 e3g2 e3g3 e3g4 e3g5 e3g6 e3g7 e3g8 e3g9 e3g10

e4e5 e4g1 e4g2 e4g3 e4g4 e4g5 e4g6 e4g7 e4g8 e4g9 e4g10

e5g1 e5g2 e5g3 e5g4 e5g5 e5g6 e5g7 e5g8 e5g9 e5g10

g1g2 g1g3 g1g4 g1g5 g1g6 g1g7 g1g8 g1g9 g1g10

g2g3 g2g4 g2g5 g2g6 g2g7 g2g8 g2g9 g2g10

g3g4 g3g5 g3g6 g3g7 g3g8 g3g9 g3g10

g4g5 g4g6 g4g7 g4g8 g4g9 g4g10

g5g6 g5g7 g5g8 g5g9 g5g10

g6g7 g6g8 g6g9 g6g10

g7g8 g7g9 g7g10

g8g9 g8g10

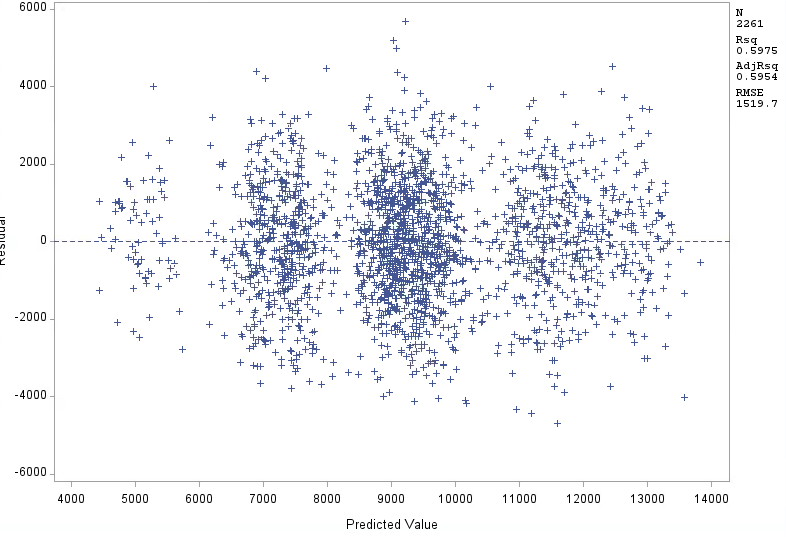
g9g10

/selection=stepwise SLENTRY=**0.01**;

plot residual.\*predicted.;

**run**;

The REG Procedure Plot



---The End---