Project 2

Group 50
Yi Pang, Kunlang Li, Yutong Guo

Introduction

The goal of this project was to estimate the function used to generate the set given data. The data set was provided in a single xlsx containing 19 independent variables (IV), one dependent variable (DV) and 1917 observations. This data was presented in the correct order and there were no missing values, making a list-wise process unnecessary. The 15 independent variables consisted of 4 environmental variables Ei and 15 genetic variables Gi. It should be noted that the value of Gi (for i=1,2,3…..15) was given as either 0 or 1. Due to these conditions, further analysis must be done to determine whether or not there is a strong correlation between the environmental variables, Ei and the genetic variables Gi.

Methods

Correlation

By using SAS, we were able to find the correlation between DV and E1 to E4, and the 4 correlation between DV and G1 to G15. By referring to code provided by Professor Finch (Appendix 1), two tables were generated (Appendix 2). Among the correlation between DV and environmental variables, the results were: E2 ($r_{dv \cdot e2}$ =0.09021), E3 ($r_{dv \cdot e2}$ =0.23777) which are relatively high and significant correlations. Among correlation between DV and gene variables, G1 ($r_{dv \cdot g6}$=0.92955) had a significant correlation for our model. Therefore, this one variable should be able to provide a reference for the model.

Box Cox Transformation

The Box Cox Transformation was applied in order to find all of the potential nonlinear transformations. Regarding to the sample code given by Professor Finch (Appendix 1), we ran the Box Cox Transformation in SAS. The result is shown below in Figure 1. Additional charts of progress are listed in Appendix 3.

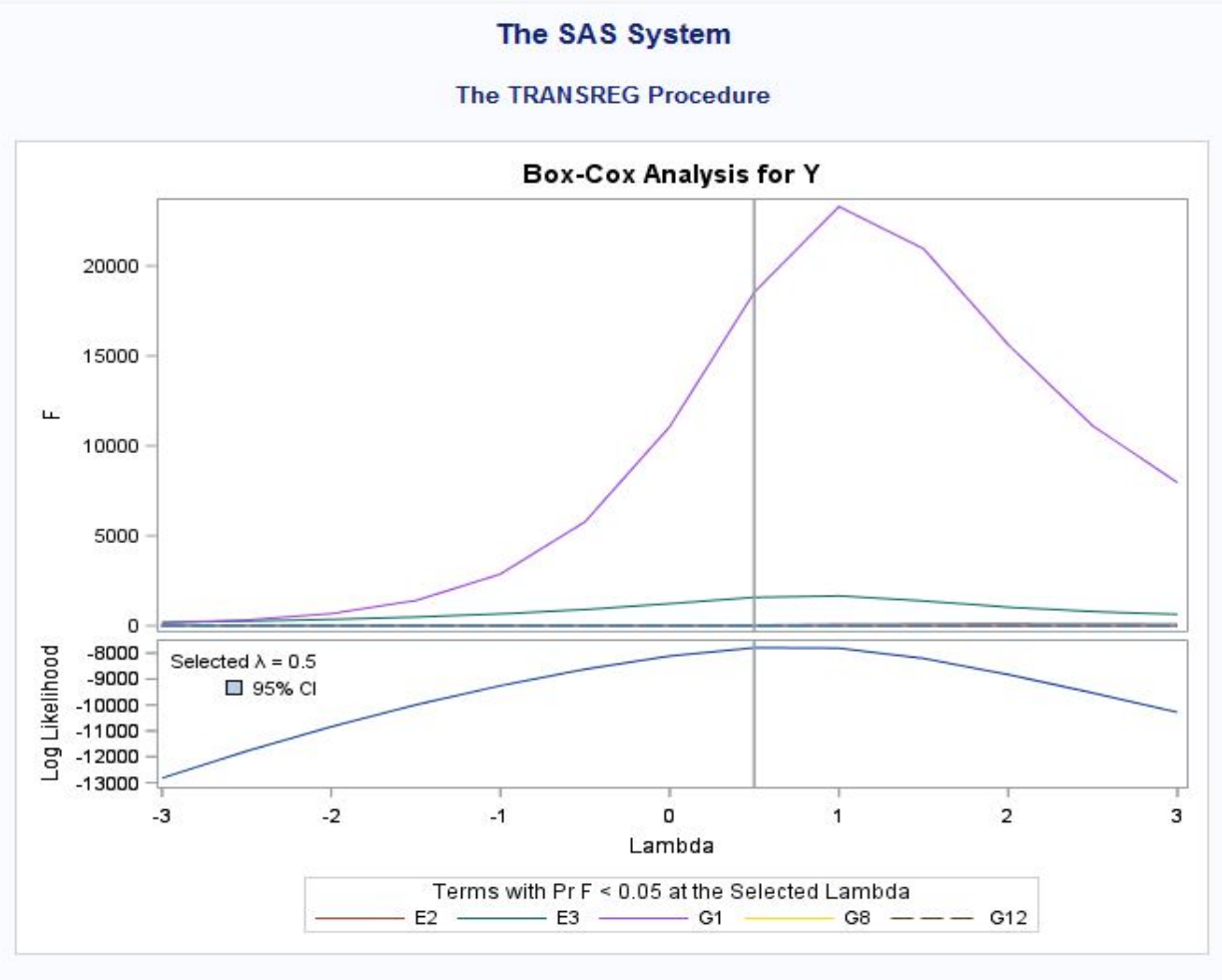


Figure 1

The graph shows the regression line of DV^λ, and we are able to see that the graph reaches its peak at the value λ=0.5; so DV is proved to be the dependent variable in this case.

Stepwise Regression

After the Box-Cox Analysis, we again used SAS to calculate the one-way and two-way interactions of the independent variables. Also by using SAS, we ran the Stepwise Regression for deciding significant independent variables. The result is shown below in Table 1 and Table2.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	e3g1		1	0.8663	0.8663	1228.45	12405.1	<.0001
2	E3		2	0.0282	0.8945	566.952	512.49	<.0001
3	e2g1		3	0.0198	0.9143	103.477	442.47	<.0001
4	G1		4	0.0038	0.9182	15.3134	89.68	<.0001
5	e2g8		5	0.0003	0.9185	10.0145	7.28	0.0070

Table 1

Univariate Regression Table Based on the Usual Degrees of Freedom						
Variable	DF	Coefficient	Type II Sum of Squares	Mean Square	F Value	Liberal p
Intercept	1	11.7675907	1500	1500	158.37	>= <.0001
Identity(E1)	1	-0.0418731	21	21	2.24	>= 0.1350
Identity(E2)	1	0.1963922	429	429	45.29	>= <.0001
Identity(E3)	1	1.1233188	14978	14978	1581.57	>= <.0001
Identity(E4)	1	-0.0168330	3	3	0.36	>= 0.5507
Identity(G1)	1	20.6948166	175574	175574	18539.9	>= <.0001
Identity(G2)	1	0.1230626	6	6	0.65	>= 0.4219
Identity(G3)	1	0.0548559	1	1	0.13	>= 0.7205
Identity(G4)	1	0.1601500	11	11	1.13	>= 0.2884
Identity(G5)	1	0.0690492	2	2	0.22	>= 0.6394
Identity(G6)	1	-0.0987677	4	4	0.42	>= 0.5164
Identity(G7)	1	0.1634416	11	11	1.21	>= 0.2724
Identity(G8)	1	0.3883590	62	62	6.58	>= 0.0104
Identity(G9)	1	0.0343408	0	0	0.05	>= 0.8226
Identity(G10)	1	-0.0704052	2	2	0.21	>= 0.6498
Identity(G11)	1	-0.0148955	0	0	0.01	>= 0.9201
Identity(G12)	1	-0.3497113	53	53	5.62	>= 0.0178
Identity(G13)	1	0.2336129	21	21	2.27	>= 0.1324
Identity(G14)	1	-0.2499203	25	25	2.62	>= 0.1056
Identity(G15)	1	-0.0805004	3	3	0.27	>= 0.6014

Table 2

In Figure 2 we can see that the largest value of R-square appears in Step 1, which is 0.8663.

Linear Regression

At last, we use R studio to compute the final linear regression function. The result is shown below (Table 5). The ANOVA Table is attached as Appendix 5.

```
formula = Y1 ~ E3 + E3:G1 + E2:G1 + G1 + E2:G8
Residuals:
     Min       1Q   Median       3Q      Max
-6.5230 -0.8811  0.0593  0.9598  4.3960

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.930487   0.258804  30.643  < 2e-16 ***
E3          0.541134   0.016612  32.575  < 2e-16 ***
G1          5.451741   0.565966   9.633  < 2e-16 ***
E3:G1       0.074955   0.029506   2.540  0.01115 *
G1:E2       0.294188   0.025132  11.706  < 2e-16 ***
E2:G8       0.015413   0.005711   2.699  0.00702 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.501 on 1911 degrees of freedom
Multiple R-squared:  0.9185,    Adjusted R-squared:  0.9183
F-statistic:  4306 on 5 and 1911 DF,  p-value: < 2.2e-16
```

Table from R

Results

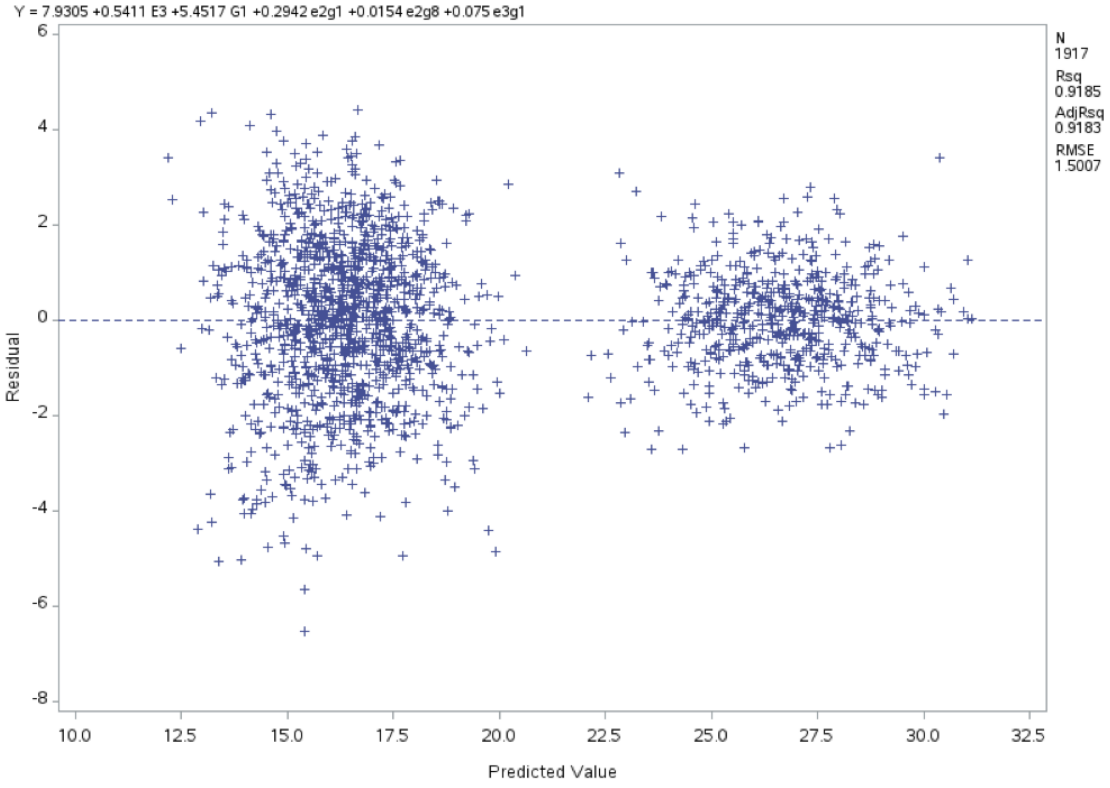
To conclude, the equation of this model is:

sqrt(Y)=7.93+0.5411E3+5.4517G1+0.0750E3G1+0.2942G1E2+0.0154E2G8

Conclusions and Discussions

A major limitation of this analysis is that only one-way or two-way interaction were considered. If there were more than two interactions occurred. This analysis would not be able to find it.

In conclusion, by using SAS and R, we calculate the final function to be sqrt(Y)=7.93+0.5411E3+5.4517G1+0.0750E3G1+0.2942G1E2+0.0154E2G



Technical Appendix

1）Code provided by Professor Finch

```
/******** Instructions on SAS     **********************************/

/* Importing the data*/
PROC IMPORT OUT= WORK.Y
            DATAFILE= "C:\Users\Han\Desktop\data2\group1.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=(function of DV);/*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
```

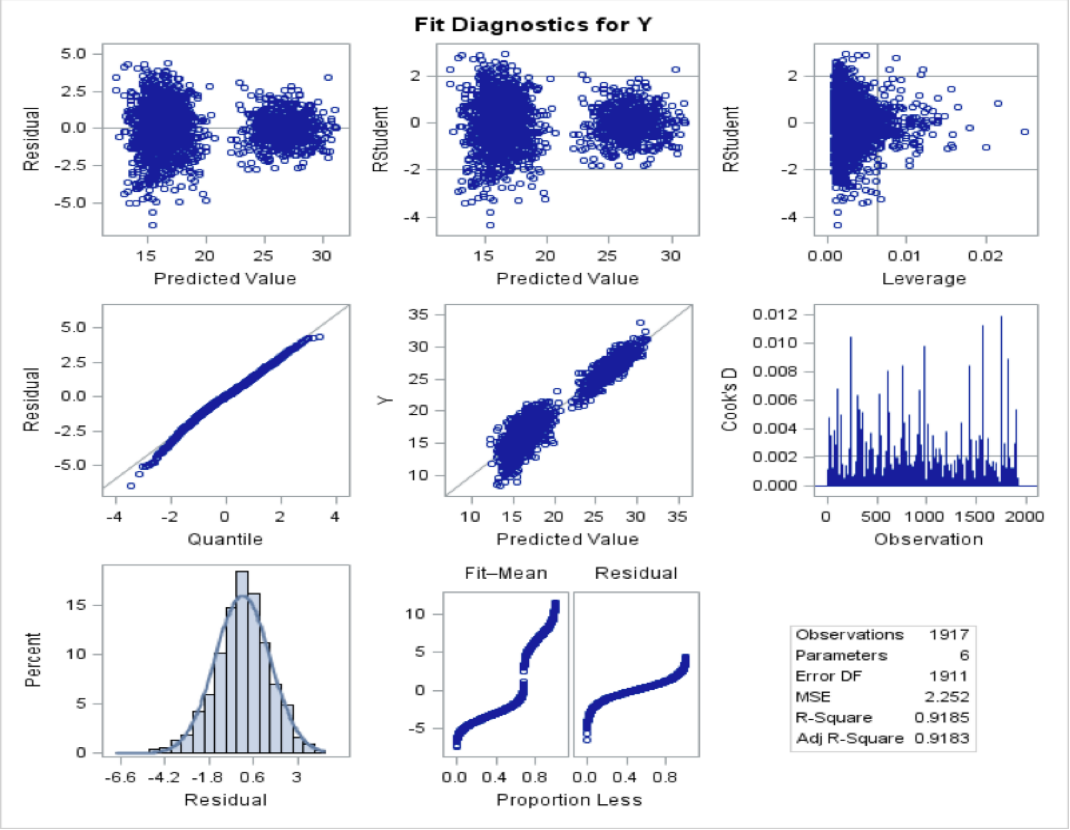
```
/selection=stepwise SLENTRY=0.01;
plot residual.*predicted.;
run;
```

2）Correlation

Pearson Correlation Coefficients, N = 1917 Prob > |r| under H0: Rho=0					
	Y	E1	E2	E3	E4
Y	1.00000	-0.00734 0.7482	0.09021 <.0001	0.23777 <.0001	0.00475 0.8352
E1	-0.00734 0.7482	1.00000	0.03601 0.1150	0.02231 0.3289	-0.01307 0.5675
E2	0.09021 <.0001	0.03601 0.1150	1.00000	0.00389 0.8649	-0.02534 0.2674
E3	0.23777 <.0001	0.02231 0.3289	0.00389 0.8649	1.00000	0.00188 0.9346
E4	0.00475 0.8352	-0.01307 0.5675	-0.02534 0.2674	0.00188 0.9346	1.00000

Pearson Correlation Coefficients, N = 1917 Prob > |r| under H0: Rho=0																
	Y	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15
Y	1.00000	0.92955 <.0001	-0.00717 0.7538	-0.04471 0.0503	-0.00689 0.7630	-0.03157 0.1670	-0.03540 0.1213	0.02770 0.2254	0.01403 0.5392	0.00186 0.9352	0.04384 0.0550	0.00142 0.9506	-0.00520 0.8201	-0.01819 0.4260	-0.07042 0.0020	-0.00697 0.7603
G1	0.92955 <.0001	1.00000	-0.00600 0.7929	-0.04640 0.0422	-0.01051 0.6456	-0.03761 0.0997	-0.03378 0.1392	0.02194 0.3371	-0.01922 0.4003	0.00849 0.7104	0.05543 0.0152	-0.00164 0.9426	0.01775 0.4373	-0.02265 0.3217	-0.05519 0.0157	-0.00880 0.7002
G2	-0.00717 0.7538	-0.00600 0.7929	1.00000	0.01782 0.4356	0.01046 0.6473	0.00104 0.9637	0.00857 0.7076	0.01085 0.6348	-0.00558 0.8070	-0.03007 0.1882	-0.03528 0.1226	-0.02491 0.2758	0.00723 0.7518	0.00203 0.9292	0.03856 0.0915	-0.03303 0.1483
G3	-0.04471 0.0503	-0.04640 0.0422	0.01782 0.4356	1.00000	-0.01778 0.4367	-0.02192 0.3374	0.00460 0.8405	0.01873 0.4124	0.00498 0.8276	0.02502 0.2735	-0.05651 0.0133	0.01159 0.6121	0.02448 0.2840	-0.02430 0.2876	0.03704 0.1050	0.03704 0.1050
G4	-0.00689 0.7630	-0.01051 0.6456	0.01046 0.6473	-0.01778 0.4367	1.00000	-0.00112 0.9610	-0.00052 0.9819	0.00689 0.7630	-0.01508 0.5093	0.00552 0.8093	0.00351 0.8781	-0.00017 0.9940	0.00272 0.9053	0.00402 0.8603	-0.01660 0.4677	0.00534 0.8152
G5	-0.03157 0.1670	-0.03761 0.0997	0.00104 0.9637	-0.02192 0.3374	-0.00112 0.9610	1.00000	0.01541 0.5001	0.01971 0.3884	0.00424 0.8530	-0.01304 0.5684	0.00503 0.8259	-0.02163 0.3440	0.03014 0.1872	-0.00324 0.8872	0.02765 0.2263	-0.00089 0.9690
G6	-0.03540 0.1213	-0.03378 0.1392	0.00857 0.7076	0.00460 0.8405	-0.00052 0.9819	0.01541 0.5001	1.00000	0.03029 0.1849	0.00763 0.7385	-0.00627 0.7838	0.02585 0.2580	0.00183 0.9362	-0.01611 0.4809	-0.00594 0.7950	0.02286 0.3171	-0.01152 0.6141
G7	0.02770 0.2254	0.02194 0.3371	0.01085 0.6348	0.01873 0.4124	0.00689 0.7630	0.01971 0.3884	0.03029 0.1849	1.00000	0.05231 0.0220	0.00378 0.8686	-0.01189 0.6030	0.01493 0.5135	0.02590 0.2571	0.04707 0.0393	0.03746 0.1010	0.00155 0.9460
G8	0.01403 0.5392	-0.01922 0.4003	-0.00558 0.8070	0.00498 0.8276	-0.01508 0.5093	0.00424 0.8530	0.00763 0.7385	0.05231 0.0220	1.00000	-0.00568 0.8038	-0.01506 0.5100	0.04760 0.0372	0.00107 0.9628	-0.01461 0.5226	0.00144 0.9499	-0.01561 0.4945
G9	0.00186 0.9352	0.00849 0.7104	-0.03007 0.1882	0.02502 0.2735	0.00552 0.8093	-0.01304 0.5684	-0.00627 0.7838	0.00378 0.8686	-0.00568 0.8038	1.00000	-0.03808 0.0956	0.01094 0.6320	0.01212 0.5958	-0.02072 0.3647	0.03096 0.1755	-0.00365 0.8730
G10	0.04384 0.0550	0.05543 0.0152	-0.03528 0.1226	-0.05651 0.0133	0.00351 0.8781	0.00503 0.8259	0.02585 0.2580	-0.01189 0.6030	-0.01506 0.5100	-0.03808 0.0956	1.00000	0.01695 0.4583	-0.00310 0.8921	0.00374 0.8700	-0.01385 0.5446	-0.05365 0.0188
G11	0.00142 0.9506	-0.00164 0.9426	-0.02491 0.2758	0.01159 0.6121	-0.00017 0.9940	-0.02163 0.3440	0.00183 0.9362	0.01493 0.5135	0.04760 0.0372	0.01094 0.6320	0.01695 0.4583	1.00000	0.00751 0.7423	0.00119 0.9586	0.00155 0.9460	-0.00563 0.8053
G12	-0.00520 0.8201	0.01775 0.4373	0.00723 0.7518	0.02448 0.2840	0.00272 0.9053	0.03014 0.1872	-0.01611 0.4809	0.02590 0.2571	0.00107 0.9628	0.01212 0.5958	-0.00310 0.8921	0.00751 0.7423	1.00000	-0.00664 0.7713	-0.01135 0.6194	-0.03515 0.1240
G13	-0.01819 0.4260	-0.02265 0.3217	0.00203 0.9292	-0.02430 0.2876	0.00402 0.8603	-0.00324 0.8872	-0.00594 0.7950	0.04707 0.0393	-0.01461 0.5226	-0.02072 0.3647	0.00374 0.8700	0.00119 0.9586	-0.00664 0.7713	1.00000	0.02612 0.2530	0.00611 0.7890
G14	-0.07042 0.0020	-0.05519 0.0157	0.03856 0.0915	0.03704 0.1050	-0.01660 0.4677	0.02765 0.2263	0.02286 0.3171	0.03746 0.1010	0.00144 0.9499	0.03096 0.1755	-0.01385 0.5446	0.00155 0.9460	-0.01135 0.6194	0.02612 0.2530	1.00000	0.03800 0.0962
G15	-0.00697 0.7603	-0.00880 0.7002	-0.03303 0.1483	0.03704 0.1050	0.00534 0.8152	-0.00089 0.9690	-0.01152 0.6141	0.00155 0.9460	-0.01561 0.4945	-0.00365 0.8730	-0.05365 0.0188	-0.00563 0.8053	-0.03515 0.1240	0.00611 0.7890	0.03800 0.0962	1.00000

3）Fit Diagnostics for Y



4）Code used to calculating interactions and stepwise regression

```
data new;
set Y;
Y=sqrt(Y);
run;

data new1;
set new;
array one[*] E1-E4 G1-G15;
array two[*]
e1e2  e1e3  e1e4  e1g1  e1g2  e1g3  e1g4  e1g5  e1g6  e1g7  e1g8  e1g9  e1g10
e1g11 e1g12 e1g13 e1g14 e1g15
      e2e3  e2e4  e2g1  e2g2  e2g3  e2g4  e2g5  e2g6  e2g7  e2g8  e2g9  e2g10
e2g11       e2g12 e2g13 e2g14 e2g15
            e3e4  e3g1  e3g2  e3g3  e3g4  e3g5  e3g6  e3g7  e3g8  e3g9  e3g10
e3g11 e3g12 e3g13 e3g14 e3g15
                  e4g1  e4g2  e4g3  e4g4  e4g5  e4g6  e4g7  e4g8  e4g9  e4g10
e4g11 e4g12 e4g13 e4g14 e4g15
```

```
                        g1g2  g1g3  g1g4  g1g5  g1g6  g1g7  g1g8  g1g9  g1g10
g1g11 g1g12 g1g13 g1g14 g1g15
                              g2g3  g2g4  g2g5  g2g6  g2g7  g2g8  g2g9  g2g10
g2g11 g2g12 g2g13 g2g14 g2g15
                                    g3g4  g3g5  g3g6  g3g7  g3g8  g3g9  g3g10
g3g11 g3g12 g3g13 g3g14 g3g15
                                          g4g5  g4g6  g4g7  g4g8  g4g9  g4g10
g4g11 g4g12 g4g13 g4g14 g4g15
                                                g5g6  g5g7  g5g8  g5g9  g5g10
g5g11 g5g12 g5g13 g5g14 g5g15
                                                      g6g7  g6g8  g6g9  g6g10
g6g11 g6g12 g6g13 g6g14 g6g15
                                                            g7g8  g7g9  g7g10
g7g11 g7g12 g7g13 g7g14 g7g15
                                                                  g8g9  g8g10
g8g11 g8g12 g8g13 g8g14 g8g15
                                                                        g9g10
g9g11 g9g12 g9g13 g9g14 g9g15

      g10g11      g10g12      g10g13      g10g14      g10g15

      g11g12      g11g13      g11g14      g11g15

      g12g13      g12g14      g12g15

      g13g14      g13g15

      g14g15
;
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-E4 G1-G15
e1e2  e1e3  e1e4  e1g1  e1g2  e1g3  e1g4  e1g5  e1g6  e1g7  e1g8  e1g9  e1g10
e1g11 e1g12 e1g13 e1g14 e1g15
      e2e3  e2e4  e2g1  e2g2  e2g3  e2g4  e2g5  e2g6  e2g7  e2g8  e2g9  e2g10
e2g11       e2g12 e2g13 e2g14 e2g15
            e3e4  e3g1  e3g2  e3g3  e3g4  e3g5  e3g6  e3g7  e3g8  e3g9  e3g10
e3g11 e3g12 e3g13 e3g14 e3g15
                  e4g1  e4g2  e4g3  e4g4  e4g5  e4g6  e4g7  e4g8  e4g9  e4g10
e4g11 e4g12 e4g13 e4g14 e4g15
```

```
                        g1g2  g1g3  g1g4  g1g5  g1g6  g1g7  g1g8  g1g9  g1g10
g1g11 g1g12 g1g13 g1g14 g1g15
                              g2g3  g2g4  g2g5  g2g6  g2g7  g2g8  g2g9  g2g10
g2g11 g2g12 g2g13 g2g14 g2g15
                                    g3g4  g3g5  g3g6  g3g7  g3g8  g3g9  g3g10
g3g11 g3g12 g3g13 g3g14 g3g15
                                          g4g5  g4g6  g4g7  g4g8  g4g9  g4g10
g4g11 g4g12 g4g13 g4g14 g4g15
                                                g5g6  g5g7  g5g8  g5g9  g5g10
g5g11 g5g12 g5g13 g5g14 g5g15
                                                      g6g7  g6g8  g6g9  g6g10
g6g11 g6g12 g6g13 g6g14 g6g15
                                                            g7g8  g7g9  g7g10
g7g11 g7g12 g7g13 g7g14 g7g15
                                                                  g8g9  g8g10
g8g11 g8g12 g8g13 g8g14 g8g15
                                                                        g9g10
g9g11 g9g12 g9g13 g9g14 g9g15

      g10g11      g10g12      g10g13      g10g14      g10g15

      g11g12      g11g13      g11g14      g11g15

      g12g13      g12g14      g12g15

      g13g14      g13g15

      g14g15
/selection=stepwise SLENTRY=0.01;
plot residual.*predicted.;
run;
```

R code

```
> fit<-lm(Y1~E3+E3:G1+E2:G1+G1+E2:G8,data=data)
> summary(fit)
> anova(fit)
```

5)
The ANOVA Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
E3	1	3488	3488	1548.6931	< 2e-16 ***
G1	1	44653	44653	19828.0332	< 2e-16 ***
E3:G1	1	12	12	5.5496	0.01859 *
G1:E2	1	318	318	141.4257	< 2e-16 ***
E2:G8	1	16	16	7.2836	0.00702 **
Residuals	1911	4304	2		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1