AMS 315 Project 2
Professor: Stephen Finch
Group 87: Ivan Tinov, Lipeng Feng

Introduction:

In this report we aim to analyze a database file that contains various data on gene-by-environmental interaction pertaining to when an organism responds to different environments and the way their genetic make-up affects that response. This study was carried out by Caspi te al and we are trying to carry out our own study based on his research. There is a single dependent variable (DV) and twenty independent variables – five labeled as E1-E5, and fifteen labeled as G1-G15. The variables E1-E5 represent the environmental variables, while the variables G1-G15 represent the various gene variables. It can be seen in the data set that G1-G15 are labeled as either a 1 for the gene being present or a 0 for the gene being not present. In order to find out how a specific gene or set of genes affects the way an organism reacts in an environment, we have to have different cases with organisms that have some of the genes in question, none at all, or all of the genes present and see how each of them react to the environment. We tried to simulate this with the data set analyzed below.

Methodology:

For our analysis of the data, we used SAS as our program of choice. We downloaded a virtual machine, as well as the SAS program image file and loaded it onto the machine to boot up SAS. After SAS was set up, we imported our Group_87.csv excel file into the program and made a new SAS program. We used the code provided below in Appendix 1 to analyze our data and create the charts that can be seen in this report. Our data set contained a total of 2000 rows of data and since there was no missing independent or dependent variables, we were able to process the data right away without any transforms.

Below can be seen the table for the Pearson Correlation Coefficient analysis between DV and each of the environmental variables (E1-E5). Variables E2, E3, and E5 show significant p-values.

Table 1a Pearson Correlation Coefficient Table (N = 2000) for E1-E5

Pearson Correlation Coefficients, N = 2000
Prob > |r| under H0: Rho=0

	DV	E1	E2	E3	E4	E5
DV	1.00000	0.00250 0.9112	0.26681 <.0001	0.43850 <.0001	0.00786 0.7253	0.25195 <.0001
E1	0.00250 0.9112	1.00000	-0.06484 0.0037	0.01306 0.5595	0.01362 0.5426	-0.03747 0.0939
E2	0.26681 <.0001	-0.06484 0.0037	1.00000	-0.02238 0.3171	-0.03664 0.1014	-0.00628 0.7790
E3	0.43850 <.0001	0.01306 0.5595	-0.02238 0.3171	1.00000	0.02309 0.3020	-0.01247 0.5772
E4	0.00786 0.7253	0.01362 0.5426	-0.03664 0.1014	0.02309 0.3020	1.00000	-0.02088 0.3507
E5	0.25195 <.0001	-0.03747 0.0939	-0.00628 0.7790	-0.01247 0.5772	-0.02088 0.3507	1.00000

Below is the table for the Pearson Correlation Coefficient analysis between DV and each of the gene variables (G1-G15). Variables G1, G7, and G13 show significant p-values.

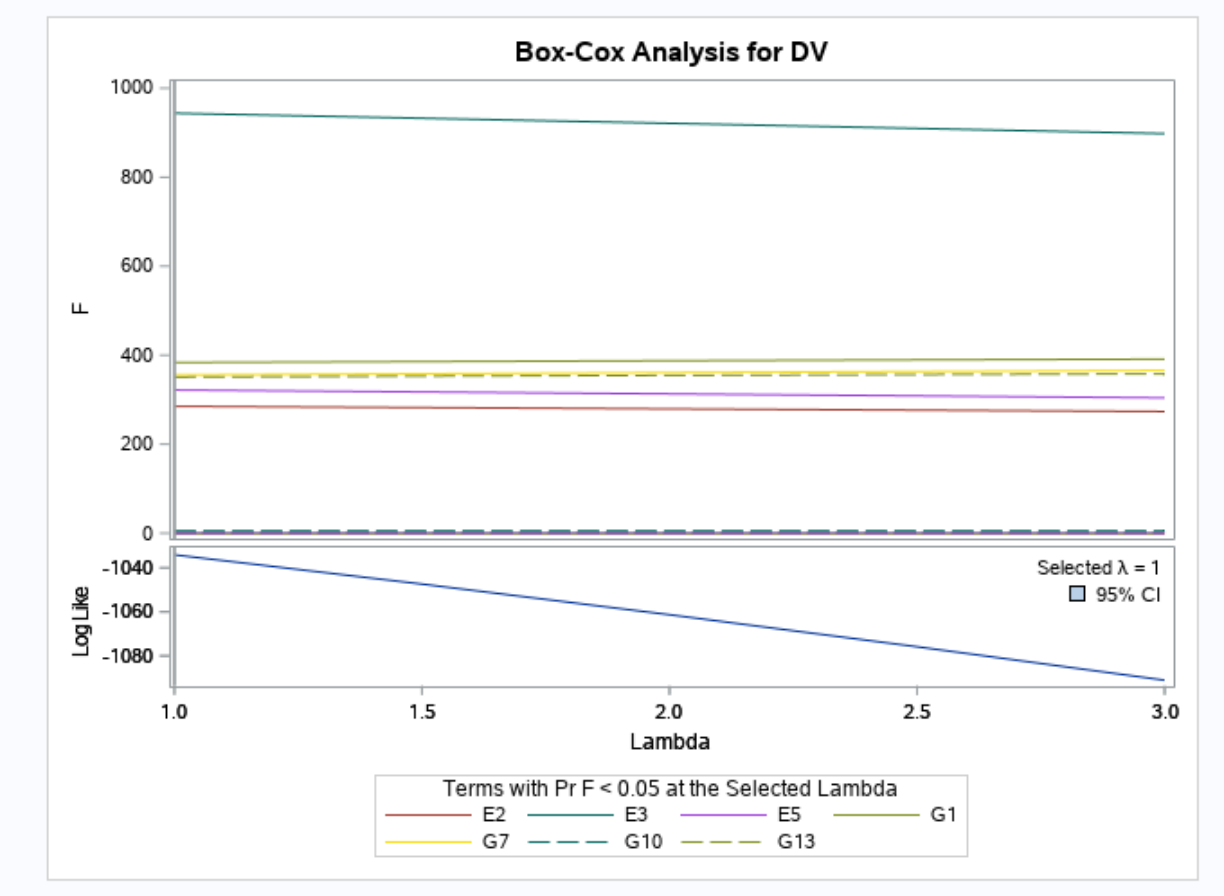
Table 1b Pearson Correlation Coefficient Table (N = 2000) for G1-G15

Pearson Correlation Coefficients, N = 2000
Prob > |r| under H0: Rho=0

	DV	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15
DV	1.00000	0.29610 <.0001	0.04078 0.0682	0.02729 0.2225	-0.01951 0.3833	-0.02498 0.2641	-0.03006 0.1790	0.25933 <.0001	-0.02358 0.2919	-0.01012 0.6509	0.00374 0.8672	0.03619 0.1056	-0.00581 0.7952	0.28774 <.0001	-0.00895 0.6890	-0.01308 0.5587
G1	0.29610 <.0001	1.00000	-0.00413 0.8536	0.00616 0.7830	0.03217 0.1504	-0.00041 0.9854	0.00797 0.7217	0.00555 0.8039	-0.00864 0.6995	-0.01836 0.4118	-0.00768 0.7313	0.01419 0.5258	0.00026 0.9909	-0.01516 0.4981	-0.03200 0.1525	-0.05683 0.0110
G2	0.04078 0.0682	-0.00413 0.8536	1.00000	0.02808 0.2094	-0.01592 0.4767	-0.02472 0.2691	-0.00202 0.9282	0.00378 0.8659	-0.03483 0.1195	-0.03270 0.1438	0.02417 0.2801	0.00410 0.8547	-0.02188 0.3282	0.00946 0.6723	0.02000 0.3713	-0.00491 0.8262
G3	0.02729 0.2225	0.00616 0.7830	0.02808 0.2094	1.00000	-0.01610 0.4717	0.02865 0.2003	0.03202 0.1523	-0.00172 0.9387	0.00277 0.9014	0.00059 0.9789	-0.00620 0.7816	-0.01012 0.6510	0.02184 0.3289	0.00070 0.9750	-0.01000 0.6549	-0.02111 0.3453
G4	-0.01951 0.3833	0.03217 0.1504	-0.01592 0.4767	-0.01610 0.4717	1.00000	-0.01338 0.5498	-0.00198 0.9295	-0.02573 0.2501	0.00077 0.9725	-0.02543 0.2556	0.01780 0.4261	0.01388 0.5350	-0.00816 0.7153	-0.02737 0.2212	-0.03000 0.1799	0.01889 0.3984
G5	-0.02498 0.2641	-0.00041 0.9854	-0.02472 0.2691	0.02865 0.2003	-0.01338 0.5498	1.00000	0.01909 0.3936	-0.03600 0.1075	0.00115 0.9590	-0.00152 0.9459	-0.02176 0.3307	-0.00945 0.6727	0.03043 0.1737	0.02768 0.2160	-0.01501 0.5023	-0.02242 0.3162
G6	-0.03006 0.1790	0.00797 0.7217	-0.00202 0.9282	0.03202 0.1523	-0.00198 0.9295	0.01909 0.3936	1.00000	-0.00006 0.9980	-0.03096 0.1663	-0.00893 0.6900	0.02004 0.3703	-0.03198 0.1528	0.01403 0.5305	-0.01618 0.4696	-0.02400 0.2834	0.00502 0.8224
G7	0.25933 <.0001	0.00555 0.8039	0.00378 0.8659	-0.00172 0.9387	-0.02573 0.2501	-0.03600 0.1075	-0.00006 0.9980	1.00000	-0.00836 0.7086	-0.01988 0.3743	-0.01545 0.4899	-0.02167 0.3326	-0.00756 0.7356	-0.00798 0.7213	-0.03001 0.1797	0.02132 0.3407
G8	-0.02358 0.2919	-0.00864 0.6995	-0.03483 0.1195	0.00277 0.9014	0.00077 0.9725	0.00115 0.9590	-0.03096 0.1663	-0.00836 0.7086	1.00000	-0.03298 0.1404	-0.00146 0.9479	-0.01128 0.6142	-0.00337 0.8803	-0.00340 0.8792	-0.02901 0.1947	0.00575 0.7972
G9	-0.01012 0.6509	-0.01836 0.4118	-0.03270 0.1438	0.00059 0.9789	-0.02543 0.2556	-0.00152 0.9459	-0.00893 0.6900	-0.01988 0.3743	-0.03298 0.1404	1.00000	-0.03786 0.0905	0.00051 0.9819	0.01035 0.6435	0.03800 0.0894	-0.00300 0.8932	0.01556 0.4867
G10	0.00374 0.8672	-0.00768 0.7313	0.02417 0.2801	-0.00620 0.7816	0.01780 0.4261	-0.02176 0.3307	0.02004 0.3703	-0.01545 0.4899	-0.00146 0.9479	-0.03786 0.0905	1.00000	-0.01224 0.5842	0.01368 0.5408	-0.04071 0.0687	0.00800 0.7206	-0.02123 0.3427
G11	0.03619 0.1056	0.01419 0.5258	0.00410 0.8547	-0.01012 0.6510	0.01388 0.5350	-0.00945 0.6727	-0.03198 0.1528	-0.02167 0.3326	-0.01128 0.6142	0.00051 0.9819	-0.01224 0.5842	1.00000	-0.01419 0.5258	0.02691 0.2290	0.02400 0.2833	-0.03114 0.1639
G12	-0.00581 0.7952	0.00026 0.9909	-0.02188 0.3282	0.02184 0.3289	-0.00816 0.7153	0.03043 0.1737	0.01403 0.5305	-0.00756 0.7356	-0.00337 0.8803	0.01035 0.6435	0.01368 0.5408	-0.01419 0.5258	1.00000	0.03321 0.1377	-0.03200 0.1525	0.01483 0.5075
G13	0.28774 <.0001	-0.01516 0.4981	0.00946 0.6723	0.00070 0.9750	-0.02737 0.2212	0.02768 0.2160	-0.01618 0.4696	-0.00798 0.7213	-0.00340 0.8792	0.03800 0.0894	-0.04071 0.0687	0.02691 0.2290	0.03321 0.1377	1.00000	0.03007 0.1788	-0.00825 0.7123
G14	-0.00895 0.6890	-0.03200 0.1525	0.02000 0.3713	-0.01000 0.6549	-0.03000 0.1799	-0.01501 0.5023	-0.02400 0.2834	-0.03001 0.1797	-0.02901 0.1947	-0.00300 0.8932	0.00800 0.7206	0.02400 0.2833	-0.03200 0.1525	0.03007 0.1788	1.00000	0.01100 0.6230
G15	-0.01308 0.5587	-0.05683 0.0110	-0.00491 0.8262	-0.02111 0.3453	0.01889 0.3984	-0.02242 0.3162	0.00502 0.8224	0.02132 0.3407	0.00575 0.7972	0.01556 0.4867	-0.02123 0.3427	-0.03114 0.1639	0.01483 0.5075	-0.00825 0.7123	0.01100 0.6230	1.00000

The above tables provided us with a rough understanding of the correlations between the various variables analyzed, however we need to further analyze the dependent variable (DV) and see how it correlates to the rest of the variables. The dependent variable can have a linear, exponential, logarithmic, or even polynomial model. As seen in Table 2, we apply a Box-Cox Analysis for the DV to transform the data so that it can be normally distributed. In this way we can find the linear model of the transformed data. In the results below, we have a box-cox transform parameter of $\lambda=1$, which is tested with a range of 1 to 3, and a step size of 0.5. The results suggest that the data doesn't need to be transformed.

Table 2 Box-Cox Table



Box-Cox Analysis table with Selected λ=1

Result:

According to the data from the Methodology section above as well as the stepwise selection (Table 3a below), we formulated an equation that best fits our data:

$$Y = 98.4710 + 0.0354E1 + 0.3389E2 + 0.2535E3 - 0.0378E4 - 0.4429E5$$

Below in Tables 3b and 3c, we provide a Univariate Analysis of Variance Table (ANOVA) on the Usual Degrees of Freedom, and an ANOVA table for a Stepwise selection of Step 9.

Table 3a ANOVA table

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	97.7799301	1131668	1131668	402383	>= <.0001
Identity(E1)	1	0.0515177	5	5	1.81	>= 0.1783
Identity(E2)	1	0.6599444	803	803	285.41	>= <.0001
Identity(E3)	1	1.1915647	2655	2655	944.02	>= <.0001
Identity(E4)	1	0.0179326	1	1	0.22	>= 0.6360
Identity(E5)	1	0.6726451	906	906	321.99	>= <.0001

ANOVA table for variables (E1-E5) with coefficients that were used to formulate the equation that best fits our data (above). Other data such as type II sum of squares(SS), mean square(MS), F-value and p-values were included in the table as well.

Table 3b ANOVA table

Univariate ANOVA Table Based on the Usual Degrees of Freedom					
Source	DF	Sum of Squares	Mean Square	F Value	Liberal p
Model	20	7458.92	372.9460	132.61	>= <.0001
Error	1979	5565.77	2.8124		
Corrected Total	1999	13024.69			
The above statistics are not adjusted for the fact that the dependent variable was transformed and so are generally liberal.					

Univariate ANOVA table with DF, SS, MS, F-value, and p-value

Table 3c ANOVA table for Stepwise Selection (Step 9)

Stepwise Selection: Step 9

Variable G13 Entered: R-Square = 0.7792 and C(p) = 9.3312

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	10149	1127.68545	780.41	<.0001
Error	1990	2875.52472	1.44499		
Corrected Total	1999	13025			

ANOVA table with DF, SS, MS, F-value, and Pr > F after stepwise selection (Step 9) was done on the data set.

Table 4 Summary of Stepwise Selection

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	g1g13		1	0.2465	0.2465	4793.41	653.68	<.0001
2	E3		2	0.1877	0.4342	3104.30	662.41	<.0001
3	g1g7		3	0.1152	0.5494	2068.48	510.16	<.0001
4	E5		4	0.0660	0.6153	1476.11	342.10	<.0001
5	g7g13		5	0.0648	0.6802	893.830	404.28	<.0001
6	E2		6	0.0573	0.7374	379.821	434.69	<.0001
7	G7		7	0.0076	0.7450	313.278	59.43	<.0001
8	G1		8	0.0122	0.7572	205.706	99.72	<.0001
9	G13		9	0.0220	0.7792	9.3312	198.44	<.0001

Summary table of the stepwise selection for all 9 steps that were on the various variables entered through the arrays created below (in the SAS code section).

Conclusion:

As a basis to test whether our fitted linear model was good, we used the generated value of R^2 to test it. The higher R^2 obtained, the more linear our data was and thus a higher R^2 value is better. As can be seen in the Fit Diagnostics for DV (Appendix 2), our R^2 value is 0.7792, with an adjusted R^2 value of 0.7782, which are both good values although they aren't exactly 1 or really close to 1. This shows that our results still give a good model based on our predictions.

We only considered single variable and two-way interactions during our procedures, which means that some variables can be missing due to only being limited to one-way and two-way interactions.

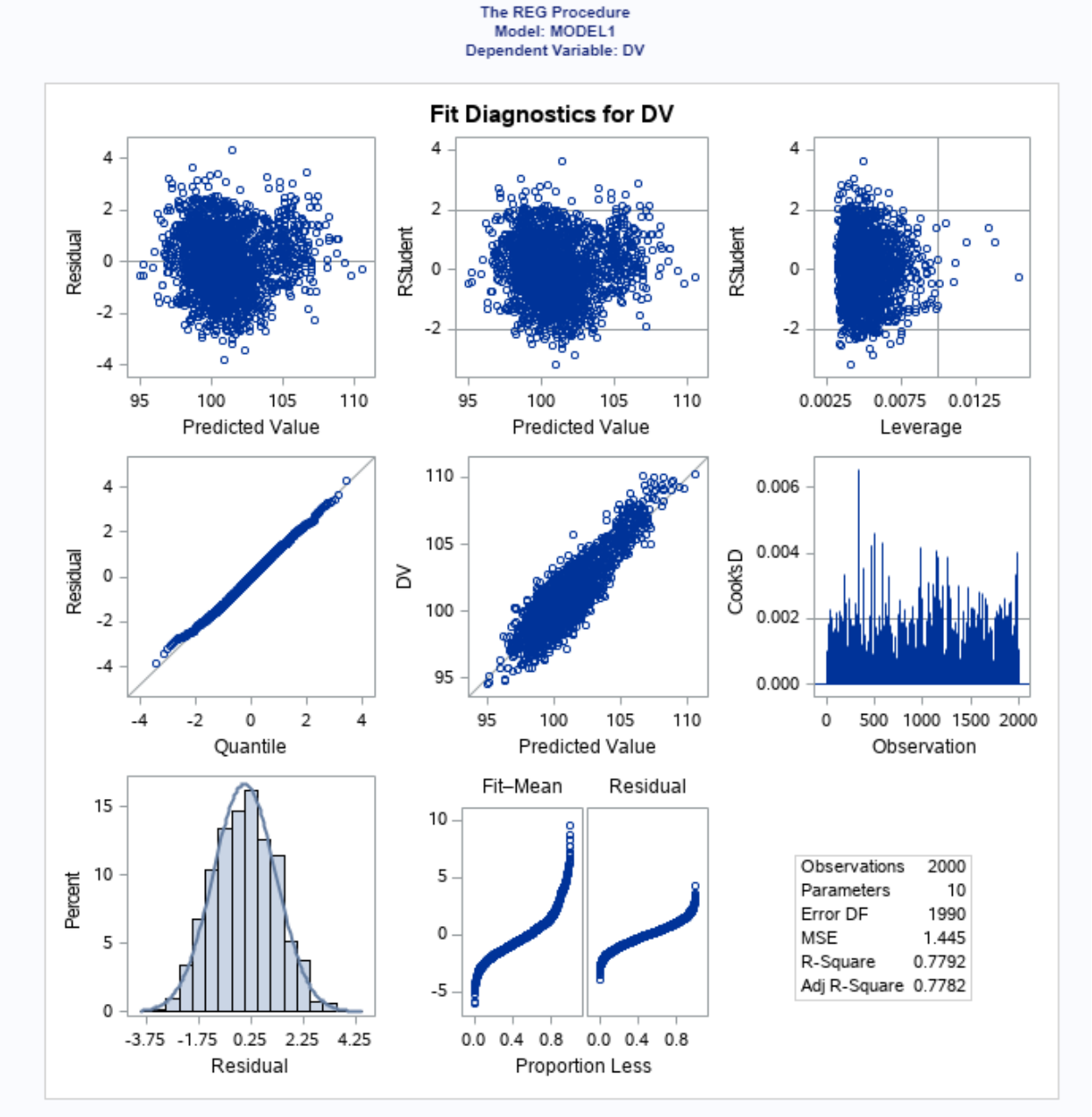
References:

- Caspi, A., Sugden, K., Moffitt, T., Taylor, A., & Craig, I. (2003). Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. Science, 301(5631), 386-389. doi:10.1126/science.1083968

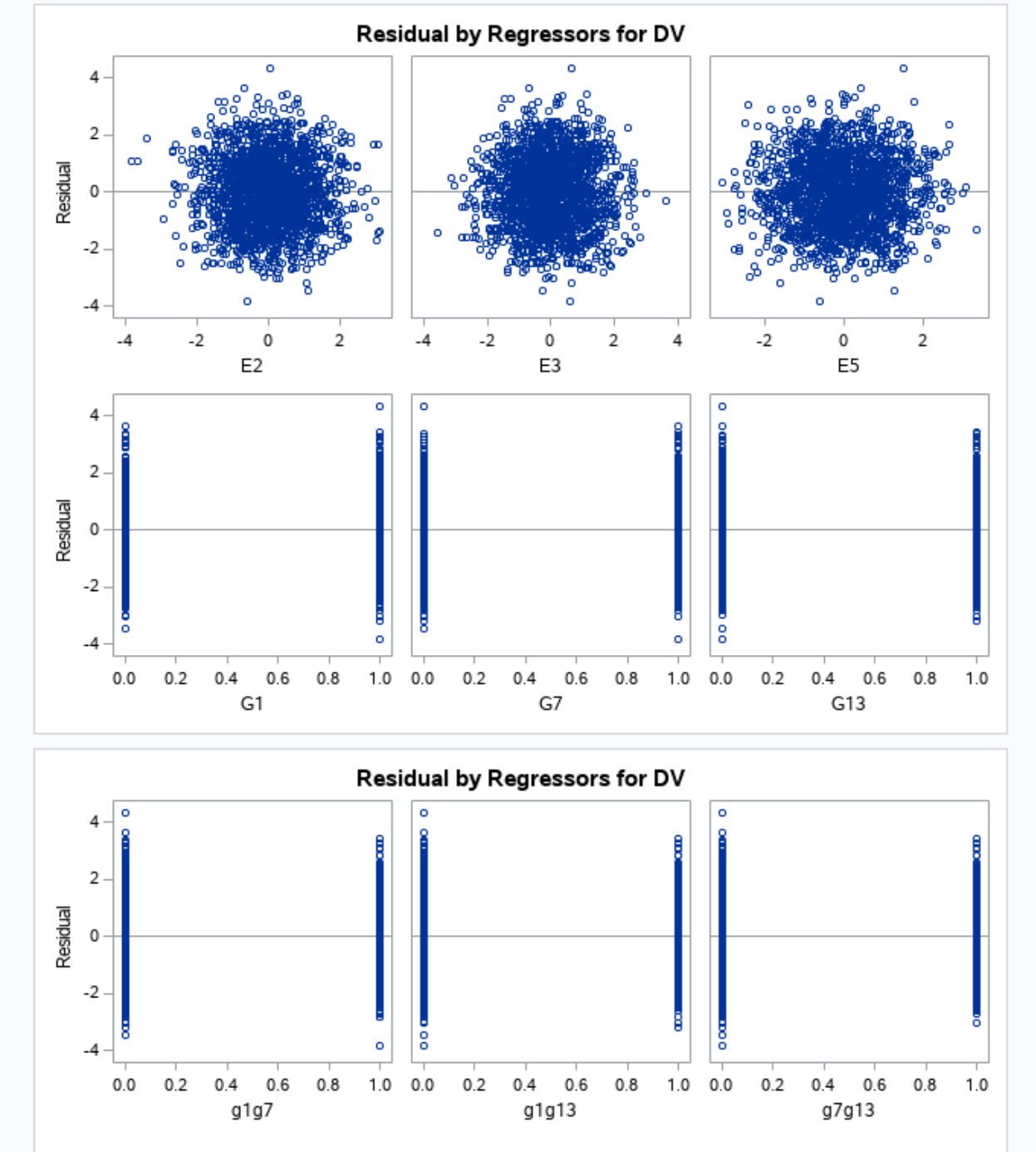
Appendix 1: ***SAS Code***

```
/* Commands to import our data set into SAS */
PROC IMPORT OUT=WORK.Y
        DATAFILE="/folders/myfolders/Group_87.csv"
        DBMS=CSV REPLACE;

        GETNAMES=YES;
        DATAROW=2;
RUN;
/* proc corr command that is used to find the correlation between the environmental and gene variables */
proc corr data=y;
        var DV E1-E5;
run;
proc corr data=y;
        var DV G1-G15;
run;

/* proc transreg command to create a model BoxCox with range 1 to 3 and step size of 0.5 */
proc transreg data=y ss2 detail;
        model BoxCox(DV/lambda=1 to 3 by 0.5)=identity(E1-E5 G1-G15);
        output;
run;

/* After the necessary transform is selected, the dependent variable will be transformed */
data new;
        set y;
        Y= exp(DV - 45.7);/* Here we use an exponential function to transform the DV */
run;

/* We create two arrays and insert the data shown below into them to compute the two way interaction
between the environmental and gene variables */
data new1;
        set new;
        array one[*] E1-E5 G1-G15;
        array two[*]
(input array values here)
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;
/* stepwise selection in proc reg to select independent variables at significance level of 0.01 */
proc reg data=new1;
        model DV= E1-E5 G1-G15
(input array values here)
/selection=stepwise SLENTRY=0.01 SLSTAY = .01;
plot residual.*predicted.;  /*  plot the residual * predicted data */
run;
```

Appendix 2: Fit Diagnostics for DV and Residual by Regressors for DV



Fit Diagnostics for the dependent variable (DV) with Mean Square Error (MSE), R-Square, and Adjusted R-Square values at the lower right-hand corner.



Residual by Regressors for DV with E2, E3, E5, as well as G1, G7, G17, g1g7, g1g13, and g7g13 that are grouped well together

End of Report