Bilal Haider (ID: 109317390)
Yun Joon Soh (ID: 108256259)
Shengyuan Luo (ID: 110598918) AMS 315 – Project II – Group 1–Spring 2017

Introduction

This project is for AMS 315 Spring 2017. The influence of genetic variables and environmental variables is of great importance as demonstrated in the depression study of Caspi et al. with subsequent meta-analysis studies confirming that environmental factors have an influence (Caspi et al., 2003; Risch et al., 2009). We received a matrix of data values with 2282 values with 5 potential environmental influencing independent variables, 15 potential gene-influencing variables, and the dependent variable. Our main purpose is to find the model used during the simulation to generate the data using various multiple regression technique.

Methods

Overview

After multiple tests, we transformed our dependent variable, y, using the exponential function. We then used Minitab and SAS to study different regression models with different interactions. We concluded that SAS had the best model. The details of why are explained in the methods section.

Transformation of Data

We first correlated the independent variables against the dependent by importing the data into SAS. Plotting the independent variable against the dependent variable also helped in decision-making. If there appeared to be a non-linear association between an independent variable and the dependent variable, a transformation may need to be considered. We also considered various non-linear transformation of the independent variable in R.

Normality of Y

The residual plot for Y was tested for normality. If the residuals of Y were found to be non-normal, we need to try a multitude of transformations. We then approached Professor Finch for advice, who suggested to use the exponential function as the transformation. We did the transformation $e^{(y-45.7)}$ where the constant 45.7 is close to our minimum value. We used this constant to reduce the exponential term so we do not have a large output.

Multiple Regression

We then used SAS and Minitab to perform a stepwise regression with a maximum of two-way interactions. We also used Minitab to establish three way interactions. With R, we considered quadratic and/or square root of a variable with two interaction variables before stepwise regression (Appendix R Code). Given a model, first thing we did was to investigate residual versus fitted value plot and make sure it does not have any pattern. We defined our α value upon entry and removal to be .01. In our comparison between the models in R, Minitab and SAS, we took into factor the p-value, F statistic, t statistic, adjusted R^2, and Mallow's C_p value. We asked graduate TA Chen for his advice in choosing the best model.

Final Decision

At last, we used Occam's razor to finalize our model. We discuss various limitations on the discussion part. R was our main package to confirm the results created by other packages and to compare between the models to come up with the final decision. R was also used to find the **confidence intervals** of the coefficients.

Results

Data Overview

Nothing seemed wrong with the data set provided. The correlation of the environmental and experimental variables on the dependent variable is listed in **Table 1**. No variable that seemed to have a significant correlation (>.4) nor did they have any patterns or relationship that appeared non-linear. We doubted that non-linear transformation to the variable was needed. However, because the dependent variable y failed to pass the normality test, we had to make the transformation on it. Results were delineated below. The simple statistics of the environmental variables were depicted in the **Appendix Table S1**. The simple statistics of the gene variables were depicted in the **Appendix Table S2**.

Transformation of Dependent Variable

At first, we were trying the Box-Cox transformation on the data set y in both SAS and Minitab. However, our optimal λ value was greater than 5. In fact, when we forced Minitab to output a definite value, we got a λ value above 40! Typical box cox values are less than 5, and since the transformation would be y^{λ}, this did not seem to be an appropriate transformation. Thus, we resorted to $e^{(y-45.7)}$as the transformation. The results of the residual plot and test for normality calculated from Minitab is seen in the **Appendix Figure S1**.

Transformation of Independent Variable

We tested a variety of transformations of the independent variable in R (for example E_1^2) and no regression output after performing a stepwise regression seemed to incorporate these nonlinear terms.

Final Decision

Minitab and SAS had different stepwise regression outputs, so we needed to determine the best regression model. It was difficult to discern with solely adjusted R^2, as they were the same. We then compared the t values, which was given in Minitab and calculated by $\sqrt{F}$ in SAS as advised by the graduate TA. Any values smaller than 3 were rejected. The p-value results were also evaluated. We also sought whether the C_p value outputted in Step 11 in SAS was adequate. From the model, C_p had a negative value, less than p, so no bias (Best Subset Regression, PSU). From here, we determined the SAS model was the best.

$$e^{y-45.7} = \beta_0 + \beta_1 E4 + \beta_2 G6 + \beta_3 G10 + \beta_4\, E3G13 + \beta_5 G1G6$$

For the best fit, the fit diagnostics for Y were depicted in **Figure 1**. The last step (Step 11) in the stepwise regression output as well as the summary of the stepwise selection was shown in the **Table 2 and Appendix Table S3** respectively. The residuals by regressors for Y was depicted in the **Appendix Figure S2**. **The residual plot was depicted in the Appendix Figure S3**.

The Minitab output was shown in the **Appendix Text S1**. The SAS code is shown in the **Appendix Text S2**.

Confidence Interval

We then used R to verify this best model in SAS. We used R to calculate the confidence intervals of the associated coefficients and constants β_k. The results of the programming calculations were listed in **Table 3**. The final equation with coefficient values is listed in the conclusion.

a

Pearson Correlation Coefficients, N = 2282
Prob > |r| under H0: Rho=0

	Y	E1	E2	E3	E4	E5
Y	1.00000	0.01243 0.5527	0.01344 0.5210	0.66053 <.0001	0.42005 <.0001	0.03200 0.1265
E1	0.01243 0.5527	1.00000	-0.00195 0.9260	-0.00185 0.9297	0.02073 0.3222	0.02331 0.2657
E2	0.01344 0.5210	-0.00195 0.9260	1.00000	0.01433 0.4938	0.00503 0.8102	0.00658 0.7534
E3	0.66053 <.0001	-0.00185 0.9297	0.01433 0.4938	1.00000	0.03746 0.0736	0.04358 0.0374
E4	0.42005 <.0001	0.02073 0.3222	0.00503 0.8102	0.03746 0.0736	1.00000	0.01181 0.5728
E5	0.03200 0.1265	0.02331 0.2657	0.00658 0.7534	0.04358 0.0374	0.01181 0.5728	1.00000

b

Pearson Correlation Coefficients, N = 2282
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	0.02240 0.2847	-0.01315 0.5301	-0.01501 0.4737	-0.01373 0.5121	0.03630 0.0830	0.03352 0.1094	-0.00848 0.6856	0.05392 0.0100	0.00317 0.8799	0.41061 <.0001	-0.00102 0.9613	-0.00947 0.6511	0.19577 <.0001	-0.04092 0.0507	-0.01583 0.4499
G1	0.02240 0.2847	1.00000	-0.00023 0.9914	0.00378 0.8568	0.00386 0.8538	0.00986 0.6377	-0.02839 0.1751	-0.01718 0.4120	0.00634 0.7621	0.01619 0.4395	-0.01691 0.4194	-0.01314 0.5305	-0.02686 0.1996	-0.01341 0.5218	-0.00749 0.7205	-0.00708 0.7355
G2	-0.01315 0.5301	-0.00023 0.9914	1.00000	0.03474 0.0970	0.01758 0.4013	-0.04394 0.0358	0.00788 0.7067	-0.00499 0.8116	-0.04032 0.0541	-0.01196 0.5681	0.00694 0.7405	0.00365 0.8618	-0.02229 0.2873	0.00958 0.6475	-0.01199 0.5671	-0.00003 0.9990
G3	-0.01501 0.4737	0.00378 0.8568	0.03474 0.0970	1.00000	-0.02862 0.1718	-0.00405 0.8466	-0.00572 0.7849	-0.00917 0.6614	-0.05175 0.0134	0.04221 0.0438	-0.01857 0.3752	0.03857 0.0654	0.00238 0.9096	0.00086 0.9674	0.00838 0.6891	-0.01775 0.3967
G4	-0.01373 0.5121	0.00386 0.8538	0.01758 0.4013	-0.02862 0.1718	1.00000	-0.02324 0.2670	0.01374 0.5119	-0.01108 0.5969	-0.01113 0.5953	-0.01837 0.3803	-0.04391 0.0360	0.02028 0.3330	-0.01364 0.5148	-0.00007 0.9974	-0.03311 0.1138	-0.01212 0.5627
G5	0.03630 0.0830	0.00986 0.6377	-0.04394 0.0358	-0.00405 0.8466	-0.02324 0.2670	1.00000	0.00799 0.7029	0.01109 0.5966	-0.01232 0.5565	0.00585 0.7800	0.02696 0.1979	0.02266 0.2791	0.03577 0.0876	0.00496 0.8126	-0.00856 0.6828	0.01125 0.5911
G6	0.03352 0.1094	-0.02839 0.1751	0.00788 0.7067	-0.00572 0.7849	0.01374 0.5119	0.00799 0.7029	1.00000	0.01442 0.4910	0.00343 0.8699	-0.00543 0.7955	-0.04083 0.0511	-0.00967 0.6443	-0.01808 0.3879	0.01938 0.3547	-0.01291 0.5376	0.01998 0.3401
G7	-0.00848 0.6856	-0.01718 0.4120	-0.00499 0.8116	-0.00917 0.6614	-0.01108 0.5969	0.01109 0.5966	0.01442 0.4910	1.00000	0.04110 0.0496	0.02982 0.1544	-0.02155 0.3034	0.00856 0.6828	0.06789 0.0012	0.00410 0.8449	-0.01547 0.4601	-0.03145 0.1332
G8	0.05392 0.0100	0.00634 0.7621	-0.04032 0.0541	-0.05175 0.0134	-0.01113 0.5953	-0.01232 0.5565	0.00343 0.8699	0.04110 0.0496	1.00000	-0.02510 0.2307	0.00467 0.8237	-0.00157 0.9402	0.00009 0.9965	0.00765 0.7149	0.01555 0.4579	0.00172 0.9344
G9	0.00317 0.8799	0.01619 0.4395	-0.01196 0.5681	0.04221 0.0438	-0.01837 0.3803	0.00585 0.7800	-0.00543 0.7955	0.02982 0.1544	-0.02510 0.2307	1.00000	-0.01888 0.3673	0.00103 0.9609	0.00982 0.6391	0.00182 0.9308	0.02833 0.1761	0.01900 0.3642
G10	0.41061 <.0001	-0.01691 0.4194	0.00694 0.7405	-0.01857 0.3752	-0.04391 0.0360	0.02696 0.1979	-0.04083 0.0511	-0.02155 0.3034	0.00467 0.8237	-0.01888 0.3673	1.00000	0.00774 0.7119	0.01345 0.5208	0.01327 0.5264	0.01158 0.5804	-0.01424 0.4967
G11	-0.00102 0.9613	-0.01314 0.5305	0.00365 0.8618	0.03857 0.0654	0.02028 0.3330	0.02266 0.2791	-0.00967 0.6443	0.00856 0.6828	-0.00157 0.9402	0.00103 0.9609	0.00774 0.7119	1.00000	-0.01472 0.4820	-0.00242 0.9080	0.00421 0.8406	-0.02378 0.2561
G12	-0.00947 0.6511	-0.02686 0.1996	-0.02229 0.2873	0.00238 0.9096	-0.01364 0.5148	0.03577 0.0876	-0.01808 0.3879	0.06789 0.0012	0.00009 0.9965	0.00982 0.6391	0.01345 0.5208	-0.01472 0.4820	1.00000	0.02496 0.2333	-0.03422 0.1022	-0.01068 0.6100
G13	0.19577 <.0001	-0.01341 0.5218	0.00958 0.6475	0.00086 0.9674	-0.00007 0.9974	0.00496 0.8126	0.01938 0.3547	0.00410 0.8449	0.00765 0.7149	0.00182 0.9308	0.01327 0.5264	-0.00242 0.9080	0.02496 0.2333	1.00000	-0.00778 0.7105	-0.00890 0.6709
G14	-0.04092 0.0507	-0.00749 0.7205	-0.01199 0.5671	0.00838 0.6891	-0.03311 0.1138	-0.00856 0.6828	-0.01291 0.5376	-0.01547 0.4601	0.01555 0.4579	0.02833 0.1761	0.01158 0.5804	0.00421 0.8406	-0.03422 0.1022	-0.00778 0.7105	1.00000	0.02117 0.3122
G15	-0.01583 0.4499	-0.00708 0.7355	-0.00003 0.9990	-0.01775 0.3967	-0.01212 0.5627	0.01125 0.5911	0.01998 0.3401	-0.03145 0.1332	0.00172 0.9344	0.01900 0.3642	-0.01424 0.4967	-0.02378 0.2561	-0.01068 0.6100	-0.00890 0.6709	0.02117 0.3122	1.00000

Table 1 - Correlation table of environmental (a) and genetic variables (b). Generated in SAS.

a

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	133.73891	26.74778	2303.07	<.0001
Error	2276	26.43337	0.01161		
Corrected Total	2281	160.17228			

b

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.06853	0.04511	0.02681	2.31	0.1288
E4	0.00010901	0.00000230	26.09735	2247.07	<.0001
G6	0.00021624	0.00002907	0.64278	55.35	<.0001
G10	0.00142	0.00002905	27.89491	2401.84	<.0001
e3g13	2.928249E-7	3.609636E-9	76.43101	6580.96	<.0001
g1g6	1.542276E-7	2.142172E-8	0.60200	51.83	<.0001

Table 2 - ANOVA table (a) and statistics of variables used in regression model (b). Stepwise regression used, where this is the last step (Step 11). Generated in SAS.

	0.5 %	99.5 %
(Intercept)	-4.775431e-02	1.848226e-01
E4	1.030772e-04	1.149336e-04
G6	1.413046e-04	2.911703e-04
G10	1.349029e-03	1.498835e-03
E3:G13	2.835193e-07	3.021305e-07
G6:G1	9.900256e-08	2.094526e-07

Table 3 - Confidence Interval of Coefficients in our model

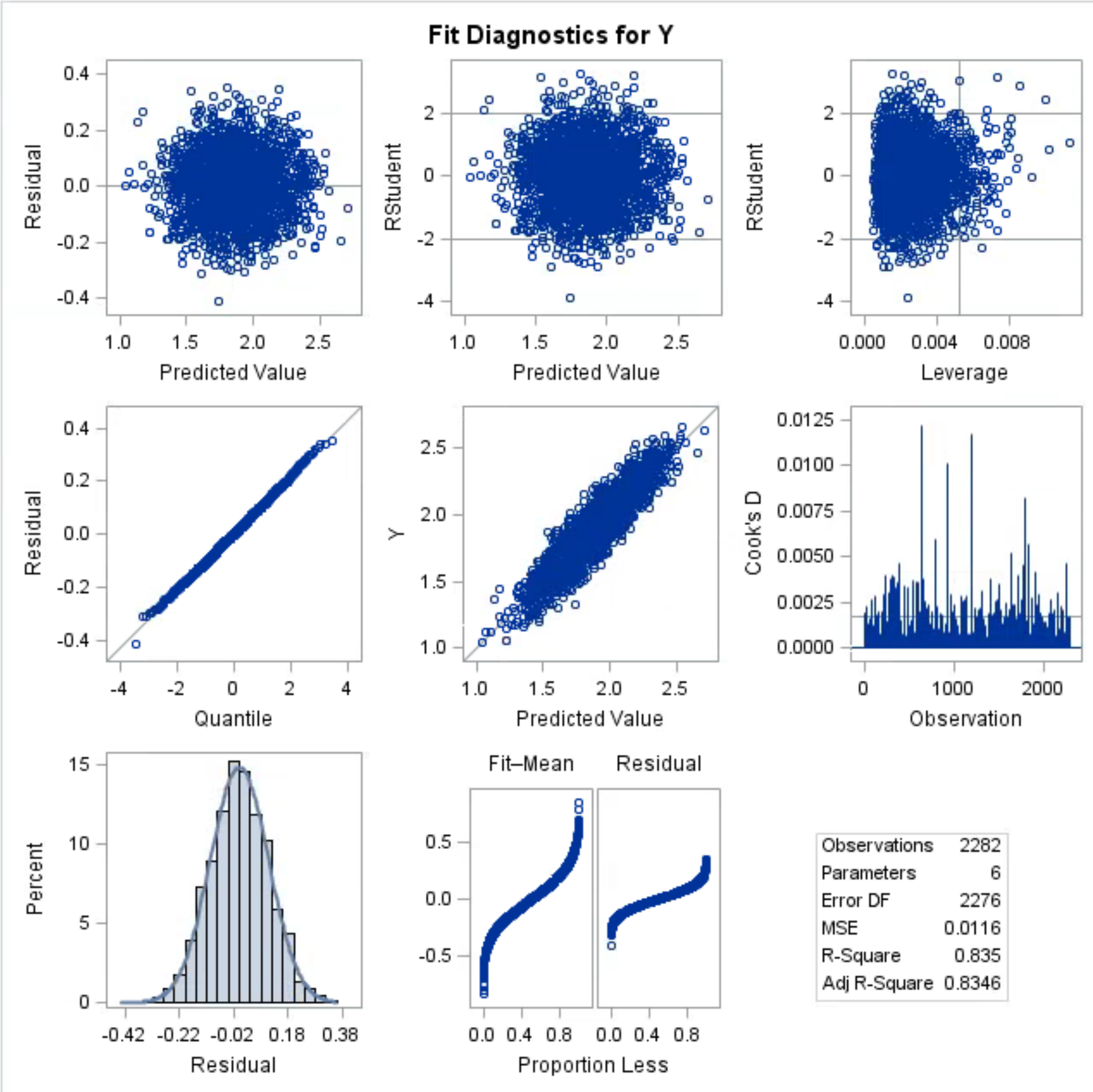


Figure 1 - Fit diagnostics for Y. Created using SAS. See Appendix for code details.

Bilal Haider (ID: 109317390)
Yun Joon Soh (ID: 108256259)
Shengyuan Luo (ID: 110598918)

Discussion and Conclusion

Discussion

We concluded non-linear independent variable terms were not needed in our model. One of the hardest decision was deciding whether G6 and G1*G6 variables are both in the model or not. The reason that it was a tough decision was that because G1*G6 has G6 in itself already affecting the value. In other words, VIF (Variance Inflation Factor) is high (higher than 10). However, we decided to leave both in the model for the following reasons.

1. Adding both variables had higher adjusted R^2 value than adding just either one of them.
2. Adding both variables would not make any of the variables' t-value below three.

However, it was a tough call up to the very last minute due to the following reason.

1. Creating a new model that splits G1*G6 to separate variable, G1 and G6, resulted in the same adjusted R^2 as the one without it.
2. Because there is no same variable appearing multiple times, VIF of all variables are below 10.

Limitation

1. One of the limitations was that we were not able to come up with a good explanation of why E3 is not in the regression model.
2. Based on E3's correlation with tY (transformed Y value) and the plot, it seemed obvious that E3 is in the model with high probability. However, we were wrong.

Conclusion

Our model is $e^{y-45.7} = .0685 + .0001\boldsymbol{E4} + .0002\boldsymbol{G6} + .0014\boldsymbol{G10} + (2.93 \times 10^{-7})\boldsymbol{E3G13} + (1.54 \times 10^{-7})\boldsymbol{G1G6}$

References

Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington, H., . . . Poulton, R. (2003). Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene. *Science, 301*(5631), 386.

Risch, N., Herrell, R., Lehner, T., Liang, K.-Y., Eaves, L., Hoh, J., . . . Merikangas, K. R. (2009). Interaction Between the Serotonin Transporter Gene (5-HTTLPR), Stressful Life Events, and Risk of Depression: A Meta-analysis. *JAMA : the journal of the American Medical Association, 301*(23), 2462-2471. doi:10.1001/jama.2009.878

10-3 - Best Subsets Regression, Adjusted R-Sq, Mallows Cp. (n.d.). Pennsylvania State University. Retreived May 04, 2017 from https://onlinecourses.science.psu.edu/stat501/node/330

Appendix

6 Variables:	Y E1 E2 E3 E4 E5

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Y	2282	46.31514	0.14483	105691	45.73576	46.67792
E1	2282	1984	1049	4528410	-1898	5275
E2	2282	1669	1035	3808139	-2408	5879
E3	2282	2636	975.68947	6016242	-893.14017	5443
E4	2282	441.83632	982.69452	1008270	-2900	4536
E5	2282	-103.43382	990.24029	-236036	-3390	3363

Appendix Table S1 - Simple statistics matrix of the environmental variables. Generated in SAS.

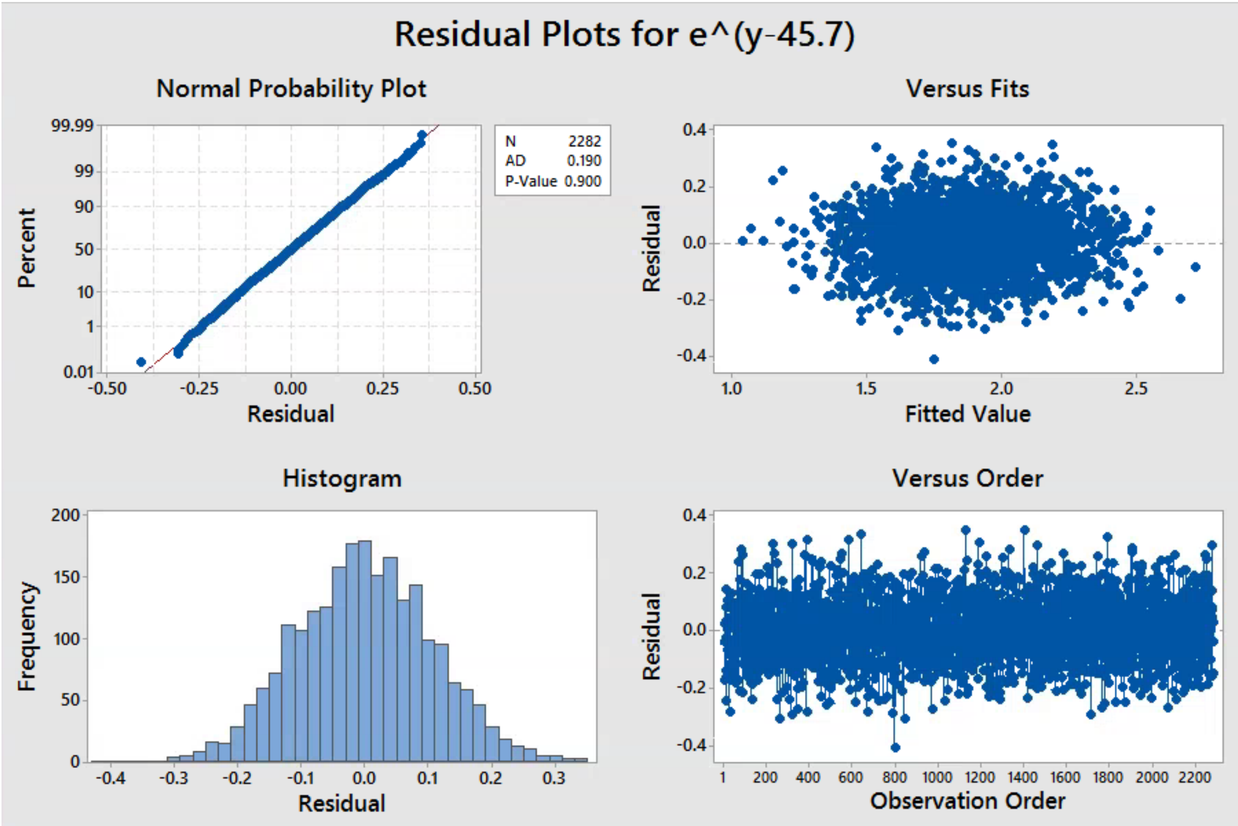
16 Variables:	Y G1 G2 G3 G4 G5 G6 G7 G8 G9 G10 G11 G12 G13 G14 G15

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Y	2282	46.31514	0.14483	105691	45.73576	46.67792
G1	2282	248.06722	76.56585	566089	-52.76352	506.92546
G2	2282	892.35867	91.10698	2036362	608.69785	1185
G3	2282	1148	71.92611	2619870	879.09231	1392
G4	2282	512.23678	84.31192	1168924	190.33798	768.13829
G5	2282	629.20979	81.58613	1435857	328.11970	873.02070
G6	2282	1376	78.66123	3139184	1106	1639
G7	2282	728.41337	80.31367	1662239	390.69878	997.87250
G8	2282	706.73229	82.30636	1612763	441.69295	1002
G9	2282	707.19671	79.95230	1613823	441.27691	1002
G10	2282	653.49667	77.76273	1491279	359.63302	891.61727
G11	2282	683.18384	74.31004	1559026	452.14030	931.75799
G12	2282	1118	84.02148	2551434	827.69603	1405
G13	2282	611.87175	75.47058	1396291	345.97996	863.40351
G14	2282	409.28451	89.66876	933987	95.78073	684.83422
G15	2282	1095	69.14702	2497989	886.09301	1353

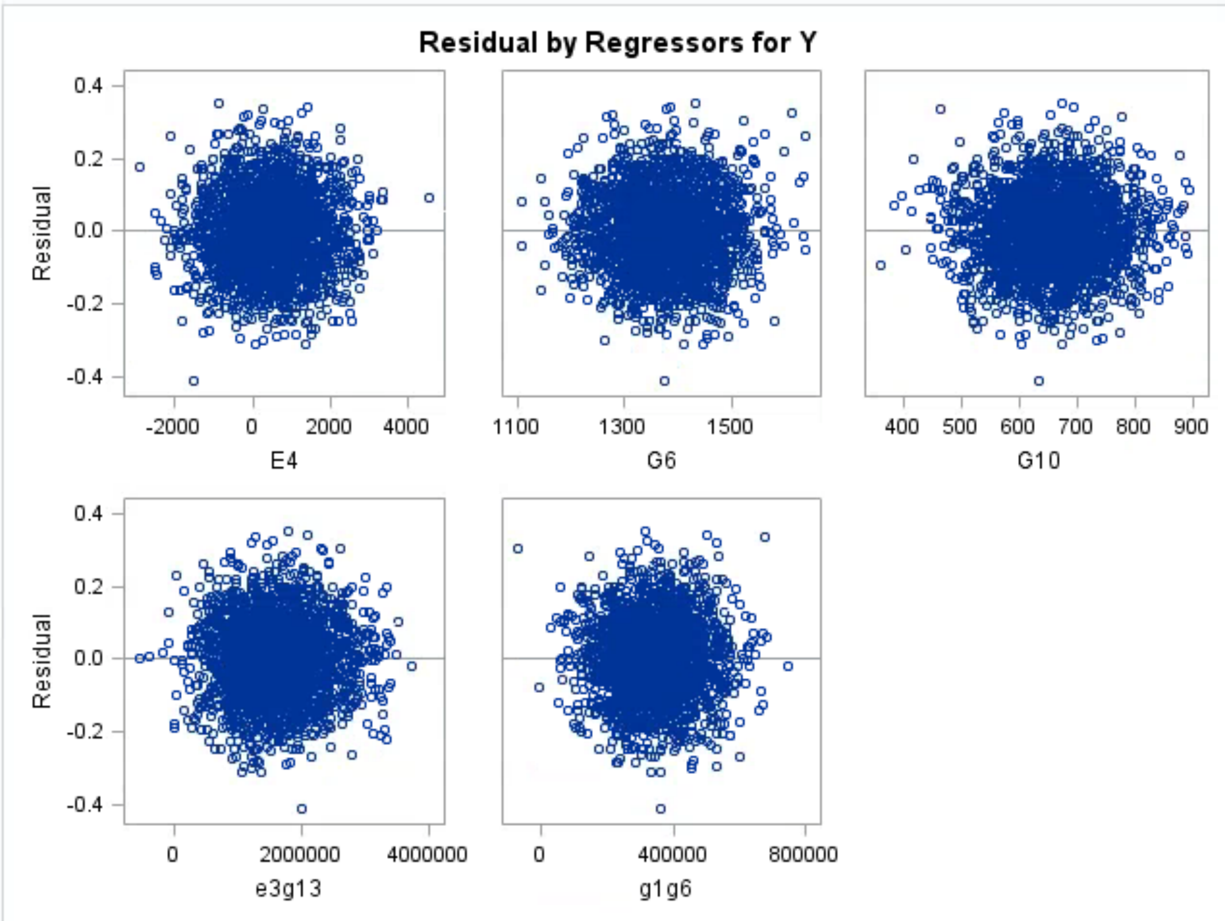
Appendix Table S2 - Simple statistics matrix of the genetic variables. Generated in SAS.

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	e3g10		1	0.5684	0.5684	3613.72	3002.48	<.0001
2	e4g13		2	0.1627	0.7311	1394.84	1378.81	<.0001
3	g10g13		3	0.0769	0.8080	347.203	912.20	<.0001
4	g1g6		4	0.0054	0.8134	275.379	65.99	<.0001
5	G6		5	0.0037	0.8171	226.967	45.95	<.0001
6	E4		6	0.0022	0.8193	199.220	27.43	<.0001
7		e4g13	5	0.0000	0.8192	197.527	0.28	0.5948
8	e3g13		6	0.0016	0.8208	177.977	20.04	<.0001
9	G10		7	0.0143	0.8351	-14.783	196.73	<.0001
10		e3g10	6	0.0001	0.8350	-15.668	1.13	0.2886
11		g10g13	5	0.0000	0.8350	-17.264	0.41	0.5231

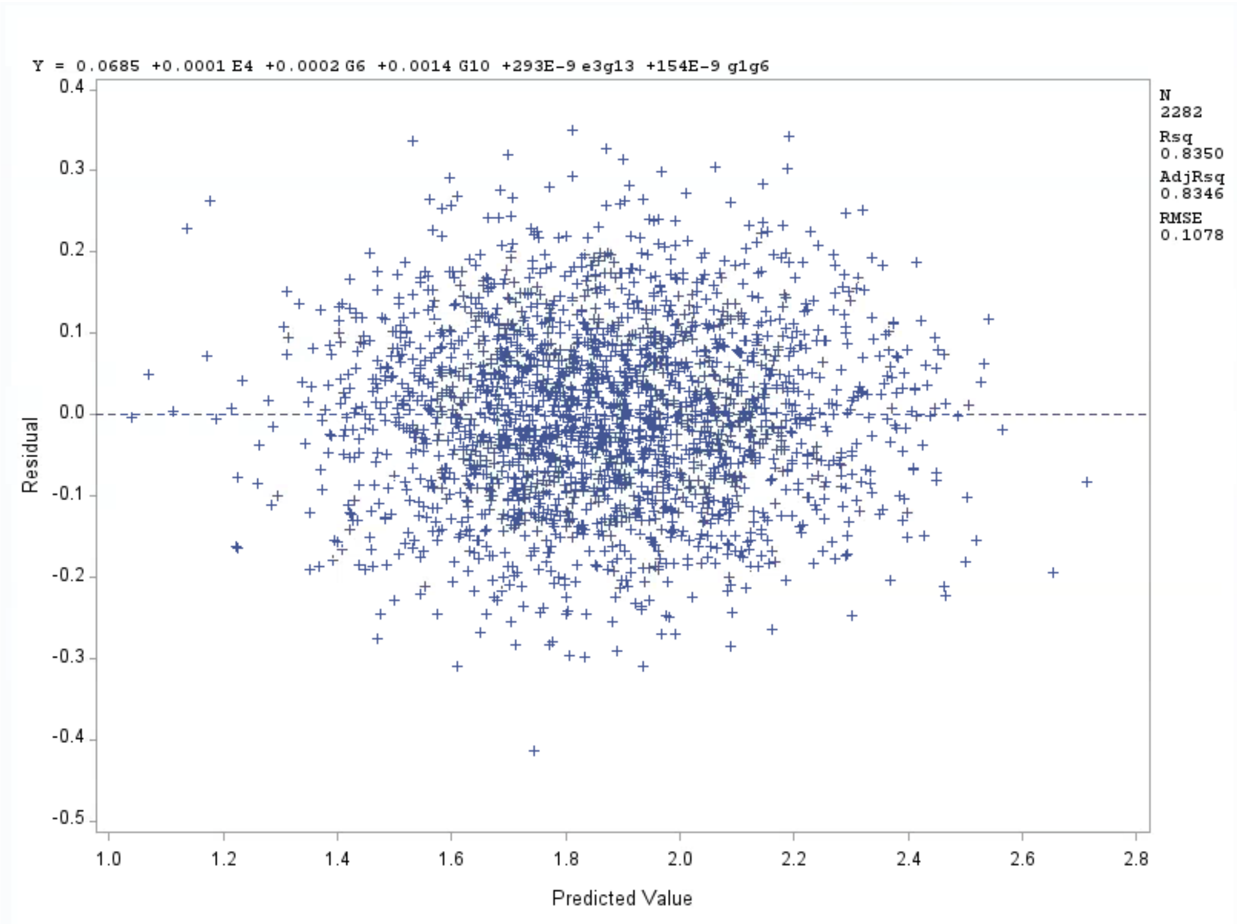
Appendix Table S3 - Summary of stepwise selection used to create regression model. Generated in SAS.



Appendix Figure S1 - Residual plots of transformed dependent variable data. Generated in Minitab.



Appendix Figure S2 - Residual by regressors plot for Y based on the multiple variables used for the plot. Generated in SAS.



Appendix Figure S3 - Residual vs Predicted Value output for our regression model. Generated in SAS.

Bilal Haider (ID: 109317390)
Yun Joon Soh (ID: 108256259)
Shengyuan Luo (ID: 110598918) AMS 315 – Project II – Group 1–Spring 2017

Appendix Text S1 - SAS code used to generate figures and tables.

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

/* Importing the data*/
PROC IMPORT OUT= WORK.Y
            DATAFILE= "\\mysbfiles.campus.stonybrook.edu\bhaider\AMS 315\group1.csv"
            DBMS=CSV REPLACE;
     GETNAMES=YES;
     DATAROW=2;
RUN;
/* Proc Corr procedure is usually used for finding the correlation between variables.*/
proc corr data=y;
        var y E1-E5;
run;
proc corr data=y;
        var y G1-G15;
run;
/*after selecting the necessary transformations, transform the dependent variable in the data step. */
data new;
        set y;
        Y= exp(y - 45.7);/*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-G15;
        array two[*]
e1e2    e1e3    e1e4    e1e5    e1g1    e1g2    e1g3    e1g4    e1g5    e1g6    e1g7    e1g8    e1g9
        e1g10   e1g11   e1g12   e1g13   e1g14   e1g15
        e2e3    e2e4    e2e5    e2g1    e2g2    e2g3    e2g4    e2g5    e2g6    e2g7    e2g8    e2g9
        e2g10   e2g11   e2g12   e2g13   e2g14   e2g15
                e3e4    e3e5    e3g1    e3g2    e3g3    e3g4    e3g5    e3g6    e3g7    e3g8    e3g9
        e3g10   e3g11   e3g12   e3g13   e3g14   e3g15
                        e4e5    e4g1    e4g2    e4g3    e4g4    e4g5    e4g6    e4g7    e4g8    e4g9
        e4g10   e4g11   e4g12   e4g13   e4g14   e4g15
                                e5g1    e5g2    e5g3    e5g4    e5g5    e5g6    e5g7    e5g8    e5g9
        e5g10   e5g11   e5g12   e5g13   e5g14   e5g15
                                        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-E5 G1-G15
e1e2    e1e3    e1e4    e1e5    e1g1    e1g2    e1g3    e1g4    e1g5    e1g6    e1g7    e1g8    e1g9
        e1g10   e1g11   e1g12   e1g13   e1g14   e1g15
        e2e3    e2e4    e2e5    e2g1    e2g2    e2g3    e2g4    e2g5    e2g6    e2g7    e2g8    e2g9
        e2g10   e2g11   e2g12   e2g13   e2g14   e2g15
                e3e4    e3e5    e3g1    e3g2    e3g3    e3g4    e3g5    e3g6    e3g7    e3g8    e3g9
        e3g10   e3g11   e3g12   e3g13   e3g14   e3g15
                        e4e5    e4g1    e4g2    e4g3    e4g4    e4g5    e4g6    e4g7    e4g8    e4g9
        e4g10   e4g11   e4g12   e4g13   e4g14   e4g15
                                e5g1    e5g2    e5g3    e5g4    e5g5    e5g6    e5g7    e5g8    e5g9
        e5g10   e5g11   e5g12   e5g13   e5g14   e5g15
                                        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  SLSTAY = .01;
plot residual.*predicted.;
Run;
```

Bilal Haider (ID: 109317390)
Yun Joon Soh (ID: 108256259)
Shengyuan Luo (ID: 110598918)

Appendix Text S2 - Minitab output.

Stepwise Selection of Terms

α to enter = 0.01, α to remove = 0.01
Stepwise selection stopped because step 10 and step 23 include identical terms.
The stepwise procedure added terms during the procedure in order to maintain a hierarchical model at each step.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	7	133.762	19.1088	1645.32	0.000
E3	1	0.018	0.0176	1.52	0.218
E4	1	26.084	26.0843	2245.92	0.000
G1	1	0.596	0.5955	51.28	0.000
G6	1	0.916	0.9160	78.87	0.000
G10	1	27.830	27.8299	2396.22	0.000
G13	1	0.020	0.0199	1.71	0.191
E3*G13	1	1.416	1.4160	121.92	0.000
Error	2274	26.410	0.0116		
Total	2281	160.172			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.107769	83.51%	83.46%	83.40%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.0829	0.0690	1.20	0.230	
E3	-0.000023	0.000018	-1.23	0.218	62.96
E4	0.000109	0.000002	47.39	0.000	1.00
G1	0.000212	0.000030	7.16	0.000	1.00
G6	0.000255	0.000029	8.88	0.000	1.00
G10	0.001423	0.000029	48.95	0.000	1.00
G13	-0.000110	0.000084	-1.31	0.191	7.87
E3*G13	0.000000	0.000000	11.04	0.000	68.67

Regression Equation

```
e^(y-45.7) = 0.0829 - 0.000023 E3 + 0.000109 E4 + 0.000212 G1 + 0.000255 G6 + 0.001423 G10
             - 0.000110 G13 + 0.000000 E3*G13
```

Fits and Diagnostics for Unusual Observations

Obs	e^(y-45.7)	Fit	Resid	Std Resid		
9	1.5553	1.8010	-0.2457	-2.28	R	
23	1.4880	1.7743	-0.2862	-2.66	R	
43	2.3909	2.3281	0.0628	0.59		X
63	2.1912	1.9531	0.2380	2.21	R	
79	2.1922	1.9101	0.2821	2.62	R	
85	1.8667	2.0843	-0.2176	-2.02	R	
87	2.1299	1.8691	0.2608	2.42	R	
126	2.0546	2.2997	-0.2452	-2.28	R	
152	2.4796	2.4746	0.0049	0.05		X
202	2.1334	1.9148	0.2186	2.03	R	
224	1.8882	1.5883	0.2999	2.79	R	
229	1.8787	1.6117	0.2670	2.48	R	
245	2.3536	2.1221	0.2315	2.15	R	
246	1.7284	1.9690	-0.2406	-2.24	R	
258	1.6265	1.9361	-0.3097	-2.87	R	
267	1.6872	1.6495	0.0377	0.35		X
273	2.2500	2.4670	-0.2170	-2.02	R	

289	1.3641	1.3622	0.0019	0.02		X
301	1.2296	1.4735	-0.2439	-2.27	R	
303	2.5948	2.5386	0.0562	0.52		X
313	1.5350	1.8330	-0.2980	-2.77	R	
318	2.4892	2.1879	0.3012	2.80	R	
338	1.4966	1.7749	-0.2783	-2.59	R	
358	2.4607	2.6624	-0.2017	-1.88		X
375	2.2017	1.9382	0.2636	2.45	R	
383	2.0178	1.7028	0.3149	2.93	R	
384	1.4298	1.7144	-0.2847	-2.64	R	
408	1.9255	1.6769	0.2486	2.31	R	
415	2.2031	1.9637	0.2395	2.22	R	
443	2.1457	1.9254	0.2203	2.05	R	
448	2.3832	2.1451	0.2382	2.21	R	
451	1.6912	1.9251	-0.2340	-2.17	R	
476	1.3821	1.6437	-0.2616	-2.43	R	
490	1.7314	1.9834	-0.2520	-2.34	R	
493	1.9740	1.7532	0.2208	2.05	R	
543	2.5378	2.2889	0.2489	2.31	R	
544	1.9730	1.7069	0.2661	2.47	R	
558	2.2742	2.2686	0.0056	0.05		X
579	2.2148	1.9006	0.3142	2.92	R	
588	1.6046	1.5151	0.0895	0.84		X
612	1.6267	1.8795	-0.2528	-2.35	R	
639	1.8674	1.5326	0.3348	3.12	R	
666	1.2704	1.5025	-0.2321	-2.16	R	
668	1.8044	2.0889	-0.2845	-2.64	R	
688	1.4497	1.3917	0.0580	0.54		X
689	1.5170	1.3992	0.1178	1.10		X
693	1.7178	1.9410	-0.2232	-2.08	R	
712	2.4530	2.2265	0.2265	2.11	R	
721	1.5125	1.7594	-0.2468	-2.29	R	
745	1.5636	1.8046	-0.2410	-2.24	R	
761	1.6956	1.9280	-0.2324	-2.16	R	
787	1.5974	1.8877	-0.2902	-2.70	R	
797	1.3307	1.7440	-0.4133	-3.84	R	
801	2.2159	2.2179	-0.0020	-0.02		X
810	2.5474	2.5780	-0.0306	-0.29		X
832	1.4236	1.6394	-0.2158	-2.00	R	
840	1.2984	1.6101	-0.3117	-2.89	R	
853	2.1525	2.2411	-0.0886	-0.83		X
919	1.4391	1.1832	0.2560	2.39	R	X
933	1.9618	1.6885	0.2733	2.54	R	
969	1.1940	1.4729	-0.2789	-2.59	R	
985	1.9686	1.7531	0.2155	2.00	R	
997	1.9826	1.9358	0.0468	0.44		X
999	2.0201	1.8044	0.2157	2.00	R	
1026	1.3297	1.5773	-0.2476	-2.30	R	
1049	1.6458	1.7675	-0.1217	-1.14		X
1092	1.6621	1.8868	-0.2247	-2.09	R	
1126	2.1599	1.8101	0.3499	3.25	R	
1129	2.2299	1.9895	0.2404	2.23	R	
1134	1.5765	1.8015	-0.2250	-2.09	R	
1185	1.9662	1.7386	0.2275	2.11	R	
1187	2.3669	2.0616	0.3053	2.85	R	
1205	2.0822	2.1528	-0.0705	-0.66		X
1224	2.3497	2.0908	0.2589	2.41	R	
1237	1.4910	1.7242	-0.2332	-2.16	R	
1255	1.4506	1.6794	-0.2288	-2.13	R	
1275	1.5270	1.5632	-0.0363	-0.34		X
1287	1.8468	2.0907	-0.2440	-2.27	R	
1289	1.4493	1.6684	-0.2191	-2.04	R	
1299	2.5708	2.3207	0.2502	2.33	R	
1317	1.3788	1.3233	0.0555	0.52		X
1329	2.0504	1.7694	0.2810	2.61	R	
1343	1.5660	1.5671	-0.0012	-0.01		X
1401	2.5336	2.1856	0.3480	3.24	R	
1415	1.5209	1.7610	-0.2401	-2.23	R	
1442	2.0042	1.7861	0.2181	2.03	R	
1461	2.2646	1.9681	0.2965	2.75	R	

1488	1.8186	1.8177	0.0009	0.01		X
1489	1.3406	1.4468	-0.1062	-0.99		X
1502	2.2277	2.0060	0.2216	2.07	R	
1524	2.4111	2.1776	0.2335	2.17	R	
1548	1.4707	1.6869	-0.2162	-2.01	R	
1582	2.1670	1.9012	0.2658	2.47	R	
1583	1.1771	1.2685	-0.0914	-0.85		X
1598	1.5435	1.4902	0.0532	0.50		X
1604	1.7944	1.5653	0.2291	2.13	R	
1626	1.3652	1.1460	0.2192	2.05	R	X
1664	1.7975	2.0220	-0.2244	-2.09	R	
1667	1.9643	1.7422	0.2221	2.06	R	
1712	1.5117	1.8085	-0.2969	-2.76	R	
1740	1.8555	1.5989	0.2566	2.38	R	
1746	1.5467	1.4437	0.1029	0.96		X
1762	1.7207	1.9907	-0.2699	-2.51	R	
1777	1.8967	2.1546	-0.2579	-2.40	R	
1784	1.2735	1.4402	-0.1668	-1.57		X
1788	2.1982	1.8729	0.3253	3.03	R	
1790	1.9042	1.6739	0.2303	2.15	R	
1811	2.0871	1.8559	0.2311	2.15	R	
1814	1.4483	1.4572	-0.0089	-0.08		X
1827	1.4138	1.6661	-0.2522	-2.35	R	
1868	1.4486	1.7017	-0.2531	-2.35	R	
1875	1.2964	1.2645	0.0319	0.30		X
1903	2.2833	2.0105	0.2728	2.53	R	
1905	2.4272	2.1422	0.2851	2.65	R	
1924	2.3795	2.1584	0.2212	2.06	R	
1944	2.2438	2.4740	-0.2301	-2.15	R	
1946	2.3662	2.1387	0.2274	2.12	R	
1949	1.8025	1.5844	0.2181	2.03	R	
1994	1.6934	1.9334	-0.2400	-2.23	R	
2010	2.0576	1.8383	0.2193	2.04	R	
2038	2.1865	1.9421	0.2444	2.27	R	
2072	1.6968	1.9663	-0.2695	-2.50	R	
2073	1.9471	1.7037	0.2433	2.26	R	
2102	1.8267	1.5654	0.2613	2.43	R	
2111	1.5917	1.8365	-0.2448	-2.27	R	
2139	1.3236	1.5445	-0.2208	-2.05	R	
2238	2.3848	2.1930	0.1918	1.79		X
2243	2.1554	1.8953	0.2601	2.42	R	
2251	1.8354	1.5863	0.2491	2.31	R	
2252	1.9350	1.8661	0.0689	0.65		X
2274	1.2803	1.4393	-0.1590	-1.49		X
2277	2.1039	1.8095	0.2944	2.73	R	

R Large residual
X Unusual X

Bilal Haider (ID: 109317390)
Yun Joon Soh (ID: 108256259)
Shengyuan Luo (ID: 110598918)

Appendix Text S3 - R code

```
####################################################################
"Function to draw plot of IV vs DV"
drawPlot <- function(){
  #column names
  cnames = colnames(input)

  par(mfrow=c(5,4))

  for(i in 1:20){
    # plot it
    plot(input[[i]], input[[21]])

    # Create title for each plot
    ti=paste("DV vs ", cnames[i])

    # set title
    title(main=ti, xlab=str(i))
  }

  par(mfrow=c(1,1))
}
####################################################################
# Data creation
# Read data, box-cox transformation, add interaction variables

# Read input
input = read.csv("Group1.csv", header = T)
df = data.frame(input)

# Box-cox transformation alternative : exp(Y-45.7)
tY = exp(df$Y -45.7)
df = cbind(df, tY)

# Shapiro test for normality
shapiro.test(df$Y) # before transformation
shapiro.test(df$tY) # after transformation

# Generate new variable columns
cnames = colnames(input)

# Gene-Env
index = 23
for(i in 1:5){
  for(j in 6:20){
    toAdd = input[[i]] * input[[j]]
    ti = paste(cnames[i], paste("*", cnames[j]))
    df = cbind(df, toAdd)
    colnames(df)[index] = ti
    index=index+1
  }
}

# Gene-Gene
for(i in 6:19){
  for(j in (i+1):20){
    toAdd = input[[i]] * input[[j]]
    ti = paste(cnames[i], paste("*", cnames[j]))
    df = cbind(df, toAdd)
    colnames(df)[index] = ti
    index=index+1
  }
}
```

```
# Quadratic (i.e., E1 raised to power of 2)
for(i in 1:20){
  toAdd = input[[i]] * input[[i]]
  ti = paste(cnames[i], paste("*", cnames[i]))
  df = cbind(df, toAdd)
  colnames(df)[index] = ti
  index=index+1
}

rm(i,j, index, toAdd, tY, ti, cnames)

# End of Data Creation
####################################################################
# Multiple Regression

# One big linear model
cnames = colnames(df)

# Generate string of the following format: tY ~ E1 + E2 + .. + G15 + E1 * G1 + ... + E5 * G15 + G1 * G2 + ... +
G14 * G15
formulaStr = "tY ~ E1"
for(i in 2:length(cnames)){
  if(!grepl(cnames[i], "Y") && !grepl(cnames[i], "tY")){ # ignore Y and tY
    formulaStr=paste(formulaStr, paste("+", cnames[i]))
  }
}

# Linear fit
fit = lm(as.formula(formulaStr), data = df)

# Reduce the number of variables
require("MASS")

# Multiple R-squared:  0.8351,    Adjusted R-squared:  0.8346
# E3 + E4 + G6 + G10 +            E3:G13 + G1 + G13
fit_mini = lm(tY~E3 + E4 + G6 + G10 + E3:G13 + G1 + G13, data=df)

# Multiple R-squared:  0.8351,    Adjusted R-squared:  0.8345
#      E4 + G6 + G10 +               + E3:G13 + G1:G6
fit_sas = lm(tY~ E4 + G6 + G10 + E3:G13 + G1:G6, data=df)

# Multiple R-squared:  0.8361,    Adjusted R-squared:  0.8353
# E3 + G10 + E4 + G13 + G6 + G1 + G2 + E3:G13 + E3:G1 + G10:G13 + E3:G2
fit_r = step(lm(tY~1, data=df), method = "forward", scope = as.formula(formulaStr))

# Multiple R-squared:  0.8351,    Adjusted R-squared:  0.8346
# E3 + E4 + G6 + G10 + G13 + G1 + E3 * G13
fit_r2 = lm(tY ~ E3 + E4 + G6 + G10 + G13 + G1 + E3:G13, data=df)

rm(i, cnames)

# Confidence Interval
confint(fit_sas, level = .99)

####################################################################
```