Christina Hartnett

Final Report AMS 578

Spring 2020

Introduction:

Synthetically generated data with 26 independent and one dependent variable was provided by the TA. The purpose of this assignment is to find the model the TA used to generate this data. We will be testing to find if there is an association between the variables and providing the fitted model for the data. This data consisted of 275 missing data points. Multiple Imputation was used in order to fill in these values with accurate data. As per instructions, the removal of outliers was not allowed in the data set. The Multiple Imputations by Chained Equations (MICE) package in R was used to perform Multiple Imputation. The acronym MICE stands for Multiple Imputations by Chained Equations. This package is a common method used while assuming the missing data is missing at random. This function will impute the data variable by variable by specifying a model for imputation for each variable. Within this package, the data is imputed using the PPM method and the seed was set to "123". This method helps make sure the imputed data, creates data that is similar to the real values. After using the MICE package in R, the simulated data has 27 columns with 1000 patients. There are 6 environmental variables and 20 genetic indicator variables.

In the article Gene-Environment Interaction, Caspi et al. tested how genetic makeup and environmental non-genetic pathogens, such as stress, can affect the onset of depression. The authors point out that a known fact for many years was that these disorders were affected by one's environment and stressful experiences. The study dives into why depression develops in some who face stress, but not all. Caspi et al. set out to prove the interaction of genes and the environment. Caspi et al. proved that the 5-HTT gene will have a large effect on the onset of depression. Using a prospective-longitudinal study, it is important to prove why experiences of stress might lead to depression in some and why it might not in others. When deciding the model, it is necessary to pick as many regressors as possible. The variance also must be low which means not including too many regressors for the final model. In other words, the right amount is needed. There is no one, unique method that can help pick the best option for the model, which is the overall task of this assignment.

Methods:

The data assigned was distributed in three separate files. File one contained the environmental variables, E; file two contained the gene indicator variables, G; the third file contained the dependent variable, Y. These files were then merged together on their index, claiming this was patient ID. For this study, all analyses were carried out using R Programming. As stated above, the missing values in the data set are filled using multiple imputation. While other methods of imputation exist, such as mean imputation and median imputation, these methods create a bias. For example, mean imputation would cause the variance to decrease while keeping the overall mean the same. Median imputation can only be used with numeric data and does not take into account any uncertainty in the imputations. Using the Cor function in R, the data was also checked to see if any values were correlated. It was discovered that there was no association.

After imputing the data set, a check was completed to see if results were normal. The p-value for this was 0.3265, thus the supplied data was approximately normal. A **box-cox** test was performed to see if the data could be made more normal. The lambda value found was 0.4646465 which is very close to .5. Because of this, a square root transformation was performed on the **y** values. After preparing the data and performing a box-cox transformation, the normality assumption was checked again. Under the Null Hypothesis that the data is normally distributed, using the Shapiro test, the p-value was 0.633. This value is much larger than .01, meaning the data is approximately normally distributed as the Null Hypothesis cannot be rejected. Additionally, after plotting the QQPlot and the histogram these plots look relatively normally distributed (Figure 1 Normal distribution). To determine if any of the environmental variables needed a transformation, a test of the linear relationships between the explanatory variable and Y was performed. The only explanatory variable with signs of linearity was E5. This variable could be transformed but transformations also come with many limitations and issues. There is not a specific recipe on how to transform data and it can take many different trials.

Methods of model selection:

- Significant variables with p-values less than .01
- If the Adjusted R-squared value increased from the previous model
- If the BIC value decreased from the previous model
- If the Residuals vs Fitted Plot is relatively straight

The first test conducted was to test the significance of the environmental variables, E1-E6. The significant environmental variables are E2, E4, and E5 and these should not be removed from the model. This model might not be the best fit because no other interactions have been tested yet. The adjusted R-squared value is 0.5737, which is relatively high. The BIC for this model is -271.0755. For the final model, the BIC should be the lowest possible value (Figure 2). The next test performed was on the independent variables, E1-E6 and G1-G25, ignoring the interactions. From the step function for step-wise regression it can be seen that E5, E2, E4, G5, G2, G15, E1, G14, G11 are significant. This had an adjusted R-squared of 0.5784 and a BIC of -246.7219. This BIC is higher than the first model which alludes to the fact that this might not be the best model. This model also has flaws within the residual's vs fitted plot, this plot shows signs of nonlinearity (Figure 3).

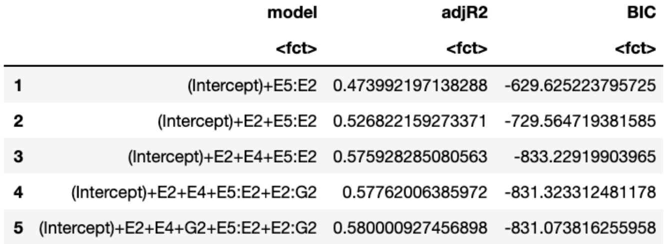
Next, including interactions, the summary and step function were performed. Testing interactions was very important in the Caspi et al. article. The significant variables from the interaction model using the step function were E5, E2, E4, G5, G2, E1, G14, G11, E5:E2, E2:G2, E4:G2, E5:E4, G5:G14, E2:G5, these had p-values lower than .05. The model created from the significant variables had an adjusted R-squared of 0.5868 and BIC of -237.4191. The BIC is much higher than the prior model, but the lower the BIC the better the model. The adjusted R-squared has increased which helps allude to the fact that the final model will have interaction variables (Figure 4).

Continuing on, a third test was performed including the third order interactions. The significant variables from the step function that were produced were E5, E2, E4, G5, G2, E1, G14, G11, E5:E2, E2:G2, E4:G2, E5:E4, G5:G14, E2:G5. This helps prove that the final model does not include any third order interactions. The adjusted R-squared and BIC values are the same as the two-way interaction model. The highly significant values from the summary on this model are E2, E5:E2, E4:E5, with an adjusted R-squared value of 0.5772 and BIC of -279.3625. Testing the ANOVA table of the third and second order interactions produces the significant values: E5, E2, E4, G5, E5:E2, E4:G2, E2:G2. Using the summary function on this model, it has an Adjusted R-squared value of 0.5784 and BIC of -258.6177. The significant values from this model creates a model with a very low Adjusted R-squared (0.5257). This has a lower adjusted R-squared than the previous model so it does not fit our model adequately and therefore will probably be rejected (Figure 5).

One limitation in model building is the multicollinearity of variables. This is a situation when multiple (2 or more) dependent variables are highly related. For example, when there is a dependent variable that is dependent on another or when a dependent variable could be used to predict another. This becomes a problem in a model because it can undermine the significance of the independent variable and can create redundancy. Some sources of multicollinearity are: data collection methods, over defined model, model specification, and constraints placed on the model. To tell if there is multicollinearity, the variance inflation factor is checked. This is done in R using the VIF function from the CAR function. The rule of thumb is that there is multicollinearity when the VIF value is above 10. To deal with multicollinearity it is important to select subsets. Two of the most common ways to combat multicollinearity are Ridge and Lasso regression. Ridge regression will shrink the coefficients to prevent overfitting and reduce complexity in the model. Lasso regression can help select the model while also preventing overfitting. Lasso regression is the preferred method over Ridge regression because Ridge regression will only shrink the coefficients but will not make them zero. This is what is needed to help select the model. While Lasso regression is the preferred method there are limitations to this as well. Lasso regression only keeps the most significant variables in the final model and forces coefficients of non-contributing variables to be zero. This may lead to a result that is not the best model. It can make variables with larger lambdas equal zero, which can eliminate variables that would be kept for significance reasons. Lasso regression often has trouble with highly correlated variables which will be taken into account for the final model selection.

When using **Lasso** regression on 2nd and 3rd order interactions the significant variables are: E1, E5, E6, E2:E6 for 2nd order and E1, E5, E6, E2:E6, E1:E2: E6 for 3rd order. The 2nd order has an adjusted R-squared of 0.5258 and the 3rd order has an adjusted R-squared of 0.5253, which are both much lower than any other adjusted R-squared's tested. Testing the fourth order interactions, the significant variables were E1, E5, E6 and E2:E6. This model had an Adjusted R-squared of .5253 which is also low. This could be due to the model's complexity. It is best to reject the Lasso method due to its low adjusted R-square values (Figure 6 Lasso Method).

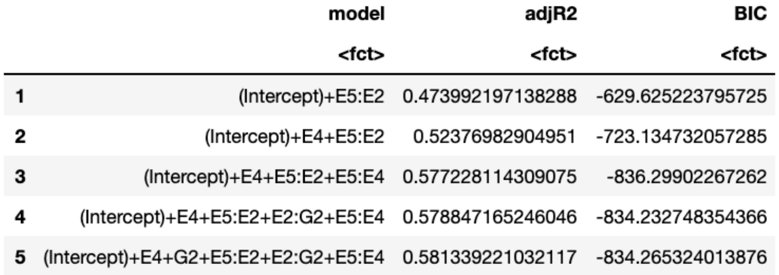
The final method that was used was the **regsubsets** function in R. This is a function in the **leaps** package that selects a model by exhaustive search, forward stepwise, backward stepwise, or sequential replacement. This function will print out the best models to describe the data. Backward selection starts with the full model and will eliminate to find significant variables. Forward selection starts with the intercept only model and will add to the current model. When looking at the output, it is important that BIC and Adjusted R-squared are taken into account. The models from this function that had the lowest BIC while having the highest Adjusted R-squared from **forward selection** were:



	model	adjR2	BIC
	<fct>	<fct>	<fct>
1	(Intercept)+E5:E2	0.473992197138288	-629.625223795725
2	(Intercept)+E2+E5:E2	0.526822159273371	-729.564719381585
3	(Intercept)+E2+E4+E5:E2	0.575928285080563	-833.22919903965
4	(Intercept)+E2+E4+E5:E2+E2:G2	0.57762006385972	-831.323312481178
5	(Intercept)+E2+E4+G2+E5:E2+E2:G2	0.580000927456898	-831.073816255958

Looking at common outputs the variables **E2, E4, and E5:E2** seem to be very important to the model. This has an adjusted R-squared of 0. 5759 and a BIC of -276.2927. This model is also a potential good fit because of the relatively flat residual plot.

The models from this function that had the lowest BIC while having the highest Adjusted R-squared from **backward selection** were:



	model	adjR2	BIC
	<fct>	<fct>	<fct>
1	(Intercept)+E5:E2	0.473992197138288	-629.625223795725
2	(Intercept)+E4+E5:E2	0.52376982904951	-723.134732057285
3	(Intercept)+E4+E5:E2+E5:E4	0.577228114309075	-836.29902267262
4	(Intercept)+E4+E5:E2+E2:G2+E5:E4	0.578847165246046	-834.232748354366
5	(Intercept)+E4+G2+E5:E2+E2:G2+E5:E4	0.581339221032117	-834.265324013876

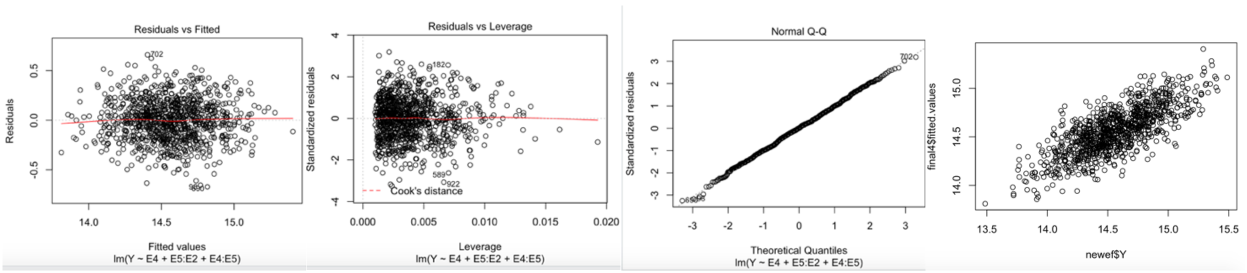
The best model from this appears to be **E4, E5:E2, E5:E4.** This model has an adjusted R-squared of **.5772** which is higher than the best model from the forward selection. When using exhaustive selection to find the best model, it produced the same result as the backward selection (E4, E5:E2, E5:E4). This model has a BIC of **-279.363.**

Results:

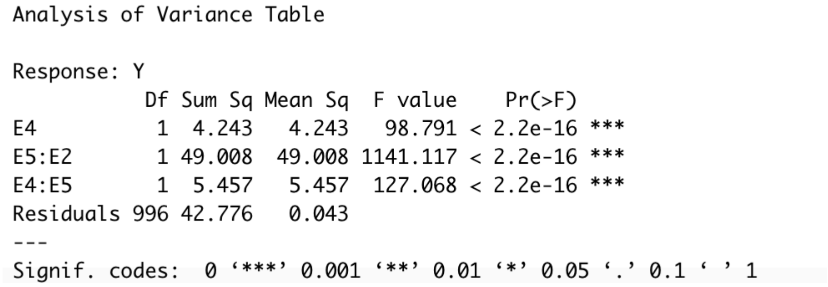
The same variables were found using the summary function in R (E4, E5:E2, E5:E4). The variance inflation factor (VIF) for these variables are all below 5. This shows that the variables are not correlated with each other, which is a necessity (Figure 7 Final Model). Using the **summary** function, **ANOVA** function and the **regsubset** function, the same results are obtained which points to the fact that this model will be a very good fit for the data provided. The plots provided below fit all the criteria and are the best of the model options that have been tested. The residuals vs fitted plot was linear and had constant variance.

The final model:

$$Y = B_0 + B_1 \mathbf{E_2} + B_2 \mathbf{E_4 E_5} + B_3 E_2 \mathbf{E_5}$$



Overall many methods were used to pick the best model for the given data set. The final model that fits the data the best is **Y= 14.25110112 - 0.17740590*E4 + 0.01024896*E5*E2 + 0.01044796*E4*E5** (Figure 7 Final Model). Based on the negative coefficient for E4, there is a negative relationship between E4 and Y. There is a positive relationship between both E5:E2 and Y and also E4:E5 and Y. Below is the Analysis of Variance table which shows our highly significant variables.



```
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq  F value    Pr(>F)    
E4          1  4.243   4.243   98.791 < 2.2e-16 ***
E5:E2       1 49.008  49.008 1141.117 < 2.2e-16 ***
E4:E5       1  5.457   5.457  127.068 < 2.2e-16 ***
Residuals 996 42.776   0.043                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion:

Many different linear models were used, including first and second interactions, testing all the options, to make sure the best model was obtained. Some of these models did not increase the Adjusted R-squared value or decrease the BIC, which is why they were rejected. A model is adequate when it meets the homoscedasticity and normality assumptions. When picking a final model, it can be expected to find some limitations of the procedures. This will ultimately lead to an issue with the accuracy of the overall model. One major limitation is the amount and quality of data provided. With missing data, it is hard to ensure the correct information is used. Although the MICE function was used to try to make the missing data as accurate as possible, only real data would provide the necessary information. Having more data could have also helped with the imputation because then the MICE function would have more data to base its imputation off of. Another limitation is the extent of interactions. While including further interactions and variables in the model showed a higher Adjusted R-squared value, they were not significant enough to be retained as the p-values were too high. Limitations are also seen/observed while using stepwise regression because it does not always provide the best option for the model. This method did not work for higher order interactions which was why LASSO regression is necessary. As described above, there are also many limitations to using LASSO regression. Multicollinearity is another limitation, when given a large number of variables this will often occur. In the end, the final model was chosen by removing p-values that were too large. Occam's Razor Principle was also applied to decide between complex and simpler models. This principle requires the selector to choose an explanation with the smallest number of assumptions. It is important that the model is not too complex when a simpler model results in a similar result. As seen above, only the environmental variables affected the dependent variable, Y. This model seems to disagree with Caspi et al.'s theory and proves that only environment variables and their interactions will affect the outcome variable. This result shows that is it very important and necessary to include the interactions between environmental variables to ensure predictions are accurate. Overall, the big picture take away from this data set is that there is no relationship between the response variable, Y, and the genetic effects. The onset of depression is brought on only by environmental effects, like stress, with no gene-environment interactions.

Appendix:

Figure 1 Normal distribution:

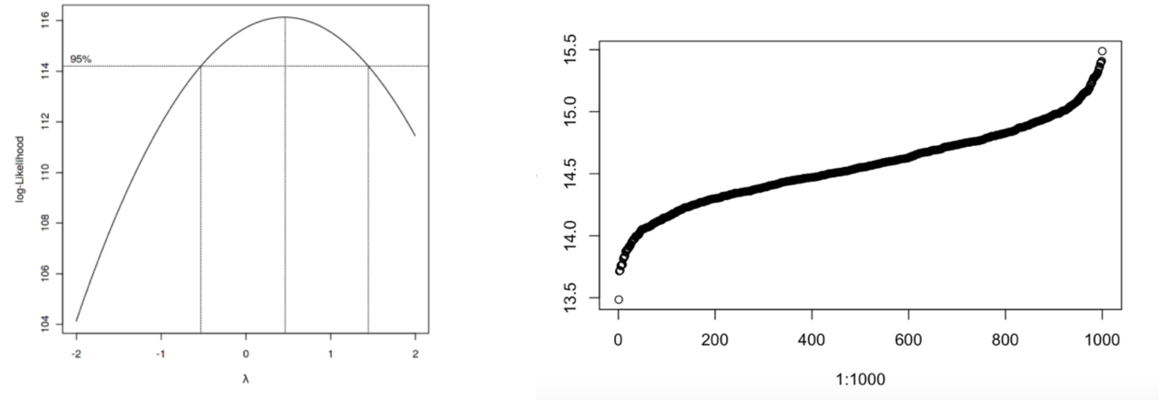
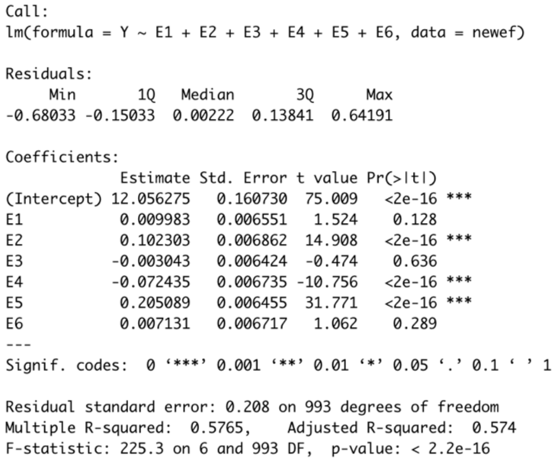
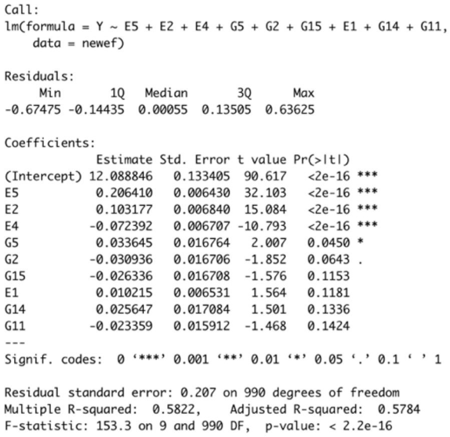


Figure 2:



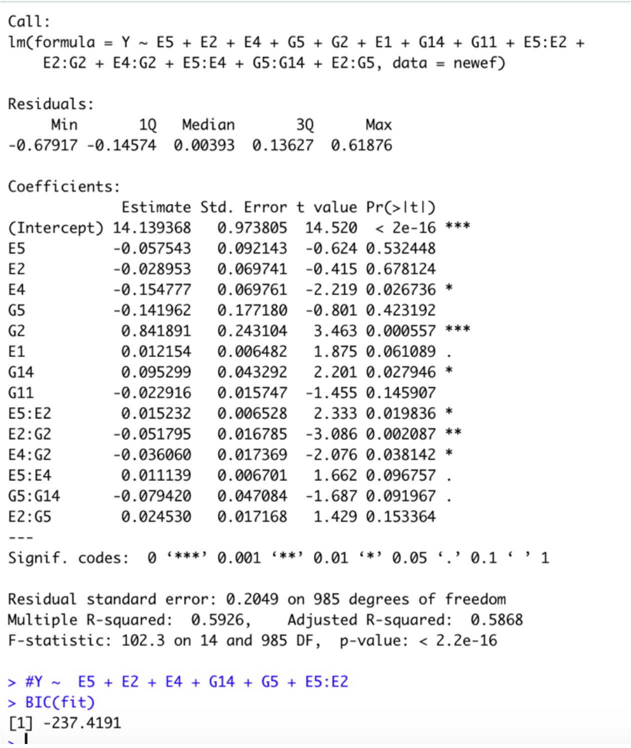
```
Call:
lm(formula = Y ~ E1 + E2 + E3 + E4 + E5 + E6, data = newef)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.68033 -0.15033  0.00222  0.13841  0.64191 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.056275   0.160730  75.009   <2e-16 ***
E1           0.009983   0.006551   1.524    0.128    
E2           0.102303   0.006862  14.908   <2e-16 ***
E3          -0.003043   0.006424  -0.474    0.636    
E4          -0.072435   0.006735 -10.756   <2e-16 ***
E5           0.205089   0.006455  31.771   <2e-16 ***
E6           0.007131   0.006717   1.062    0.289    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.208 on 993 degrees of freedom
Multiple R-squared:  0.5765,	Adjusted R-squared:  0.574 
F-statistic: 225.3 on 6 and 993 DF,  p-value: < 2.2e-16
```

Figure 3:



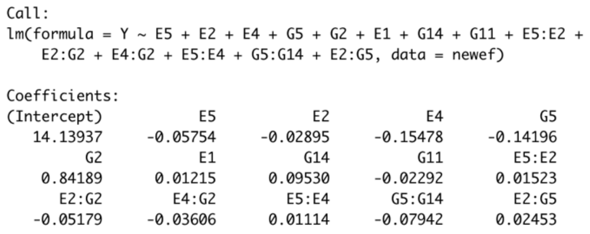
```
Call:
lm(formula = Y ~ E5 + E2 + E4 + G5 + G2 + G15 + E1 + G14 + G11, 
    data = newef)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.67475 -0.14435  0.00055  0.13505  0.63625 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.088846   0.133405  90.617   <2e-16 ***
E5           0.206410   0.006430  32.103   <2e-16 ***
E2           0.103177   0.006840  15.084   <2e-16 ***
E4          -0.072392   0.006707 -10.793   <2e-16 ***
G5           0.033645   0.016764   2.007   0.0450 *  
G2          -0.030936   0.016706  -1.852   0.0643 .  
G15         -0.026336   0.016708  -1.576   0.1153    
E1           0.010215   0.006531   1.564   0.1181    
G14          0.025647   0.017084   1.501   0.1336    
G11         -0.023359   0.015912  -1.468   0.1424    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.207 on 990 degrees of freedom
Multiple R-squared:  0.5822,	Adjusted R-squared:  0.5784 
F-statistic: 153.3 on 9 and 990 DF,  p-value: < 2.2e-16
```

Figure 4:

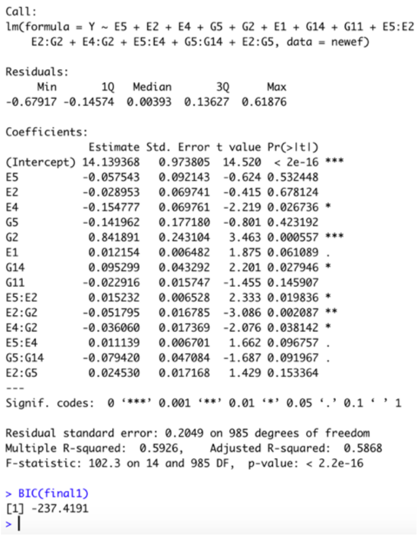


```
Call:
lm(formula = Y ~ E5 + E2 + E4 + G5 + G2 + E1 + G14 + G11 + E5:E2 +
    E2:G2 + E4:G2 + E5:E4 + G5:G14 + E2:G5, data = newef)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.67917 -0.14574  0.00393  0.13627  0.61876 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 14.139368   0.973805  14.520  < 2e-16 ***
E5          -0.057543   0.092143  -0.624 0.532448    
E2          -0.028953   0.069741  -0.415 0.678124    
E4          -0.154777   0.069761  -2.219 0.026736 *  
G5          -0.141962   0.177180  -0.801 0.423192    
G2           0.841891   0.243104   3.463 0.000557 ***
E1           0.012154   0.006482   1.875 0.061089 .  
G14          0.095299   0.043292   2.201 0.027946 *  
G11         -0.022916   0.015747  -1.455 0.145907    
E5:E2        0.015232   0.006528   2.333 0.019836 *  
E2:G2       -0.051795   0.016785  -3.086 0.002087 ** 
E4:G2       -0.036060   0.017369  -2.076 0.038142 *  
E5:E4        0.011139   0.006701   1.662 0.096757 .  
G5:G14      -0.079420   0.047084  -1.687 0.091967 .  
E2:G5        0.024530   0.017168   1.429 0.153364    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2049 on 985 degrees of freedom
Multiple R-squared:  0.5926,	Adjusted R-squared:  0.5868 
F-statistic: 102.3 on 14 and 985 DF,  p-value: < 2.2e-16

> #Y ~  E5 + E2 + E4 + G14 + G5 + E5:E2
> BIC(fit)
[1] -237.4191
>
```

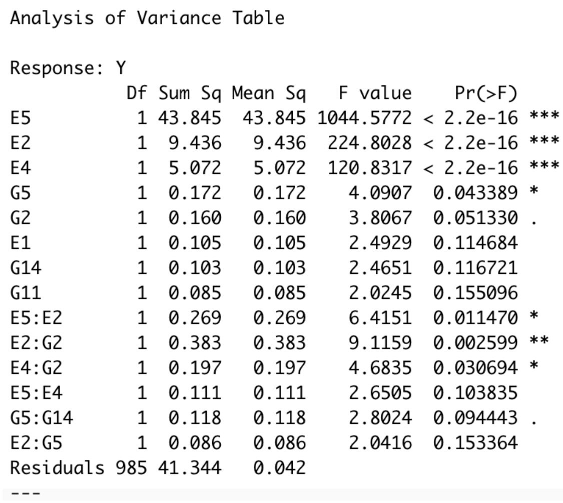
Figure 5:



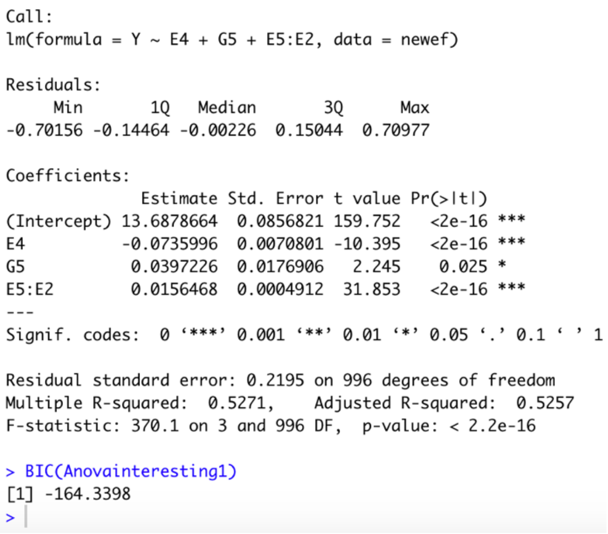
```
Call:
lm(formula = Y ~ E5 + E2 + E4 + G5 + G2 + E1 + G14 + G11 + E5:E2 +
    E2:G2 + E4:G2 + E5:E4 + G5:G14 + E2:G5, data = newef)

Coefficients:
(Intercept)           E5           E2           E4           G5  
   14.13937     -0.05754     -0.02895     -0.15478     -0.14196  
         G2           E1          G14          G11        E5:E2  
    0.84189      0.01215      0.09530     -0.02292      0.01523  
      E2:G2        E4:G2        E5:E4       G5:G14        E2:G5  
   -0.05179     -0.03606      0.01114     -0.07942      0.02453  
```



```
Call:
lm(formula = Y ~ E5 + E2 + E4 + G5 + G2 + E1 + G14 + G11 + E5:E2
    E2:G2 + E4:G2 + E5:E4 + G5:G14 + E2:G5, data = newef)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.67917 -0.14574  0.00393  0.13627  0.61876 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 14.139368   0.973805  14.520  < 2e-16 ***
E5          -0.057543   0.092143  -0.624 0.532448    
E2          -0.028953   0.069741  -0.415 0.678124    
E4          -0.154777   0.069761  -2.219 0.026736 *  
G5          -0.141962   0.177180  -0.801 0.423192    
G2           0.841891   0.243104   3.463 0.000557 ***
E1           0.012154   0.006482   1.875 0.061089 .  
G14          0.095299   0.043292   2.201 0.027946 *  
G11         -0.022916   0.015747  -1.455 0.145907    
E5:E2        0.015232   0.006528   2.333 0.019836 *  
E2:G2       -0.051795   0.016785  -3.086 0.002087 ** 
E4:G2       -0.036060   0.017369  -2.076 0.038142 *  
E5:E4        0.011139   0.006701   1.662 0.096757 .  
G5:G14      -0.079420   0.047084  -1.687 0.091967 .  
E2:G5        0.024530   0.017168   1.429 0.153364    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2049 on 985 degrees of freedom
Multiple R-squared:  0.5926,	Adjusted R-squared:  0.5868 
F-statistic: 102.3 on 14 and 985 DF,  p-value: < 2.2e-16

> BIC(final1)
[1] -237.4191
> 
```

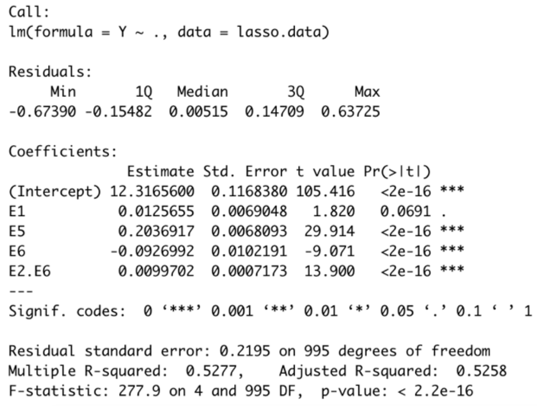


```
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq   F value    Pr(>F)    
E5          1 43.845  43.845 1044.5772 < 2.2e-16 ***
E2          1  9.436   9.436  224.8028 < 2.2e-16 ***
E4          1  5.072   5.072  120.8317 < 2.2e-16 ***
G5          1  0.172   0.172    4.0907  0.043389 *  
G2          1  0.160   0.160    3.8067  0.051330 .  
E1          1  0.105   0.105    2.4929  0.114684    
G14         1  0.103   0.103    2.4651  0.116721    
G11         1  0.085   0.085    2.0245  0.155096    
E5:E2       1  0.269   0.269    6.4151  0.011470 *  
E2:G2       1  0.383   0.383    9.1159  0.002599 ** 
E4:G2       1  0.197   0.197    4.6835  0.030694 *  
E5:E4       1  0.111   0.111    2.6505  0.103835    
G5:G14      1  0.118   0.118    2.8024  0.094443 .  
E2:G5       1  0.086   0.086    2.0416  0.153364    
Residuals 985 41.344   0.042                        
---
```

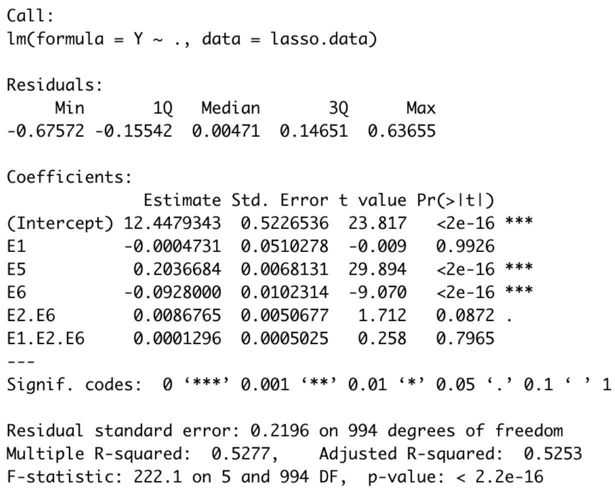


```
Call:
lm(formula = Y ~ E4 + G5 + E5:E2, data = newef)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.70156 -0.14464 -0.00226  0.15044  0.70977 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.6878664  0.0856821 159.752   <2e-16 ***
E4          -0.0735996  0.0070801 -10.395   <2e-16 ***
G5           0.0397226  0.0176906   2.245    0.025 *  
E5:E2        0.0156468  0.0004912  31.853   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2195 on 996 degrees of freedom
Multiple R-squared:  0.5271,	Adjusted R-squared:  0.5257 
F-statistic: 370.1 on 3 and 996 DF,  p-value: < 2.2e-16

> BIC(Anovainteresting1)
[1] -164.3398
> 
```

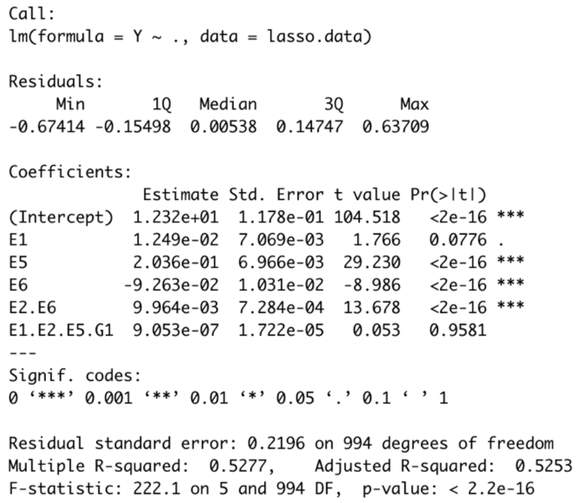
Figure 6 Lasso Method:



```
Call:
lm(formula = Y ~ ., data = lasso.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.67390 -0.15482  0.00515  0.14709  0.63725 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.3165600  0.1168380 105.416   <2e-16 ***
E1           0.0125655  0.0069048   1.820   0.0691 .  
E5           0.2036917  0.0068093  29.914   <2e-16 ***
E6          -0.0926992  0.0102191  -9.071   <2e-16 ***
E2.E6        0.0099702  0.0007173  13.900   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2195 on 995 degrees of freedom
Multiple R-squared:  0.5277,	Adjusted R-squared:  0.5258 
F-statistic: 277.9 on 4 and 995 DF,  p-value: < 2.2e-16
```



```
Call:
lm(formula = Y ~ ., data = lasso.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.67572 -0.15542  0.00471  0.14651  0.63655 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 12.4479343  0.5226536  23.817   <2e-16 ***
E1          -0.0004731  0.0510278  -0.009   0.9926    
E5           0.2036684  0.0068131  29.894   <2e-16 ***
E6          -0.0928000  0.0102314  -9.070   <2e-16 ***
E2.E6        0.0086765  0.0050677   1.712   0.0872 .  
E1.E2.E6     0.0001296  0.0005025   0.258   0.7965    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2196 on 994 degrees of freedom
Multiple R-squared:  0.5277,	Adjusted R-squared:  0.5253 
F-statistic: 222.1 on 5 and 994 DF,  p-value: < 2.2e-16
```



```
Call:
lm(formula = Y ~ ., data = lasso.data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.67414 -0.15498  0.00538  0.14747  0.63709 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.232e+01  1.178e-01 104.518   <2e-16 ***
E1           1.249e-02  7.069e-03   1.766   0.0776 .  
E5           2.036e-01  6.966e-03  29.230   <2e-16 ***
E6          -9.263e-02  1.031e-02  -8.986   <2e-16 ***
E2.E6        9.964e-03  7.284e-04  13.678   <2e-16 ***
E1.E2.E5.G1  9.053e-07  1.722e-05   0.053   0.9581    
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2196 on 994 degrees of freedom
Multiple R-squared:  0.5277,	Adjusted R-squared:  0.5253 
F-statistic: 222.1 on 5 and 994 DF,  p-value: < 2.2e-16
```

Figure 7 Final model:

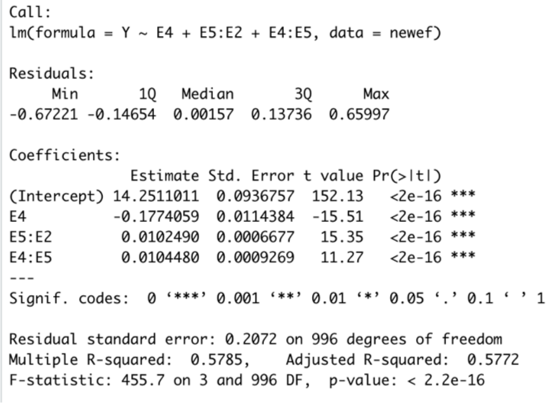
variance inflation factors for final model:

E4 2.94036435315357

E5:E2 2.07491826932905

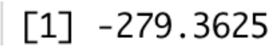
E4:E5 4.09541128899353

Summary Function for final model:



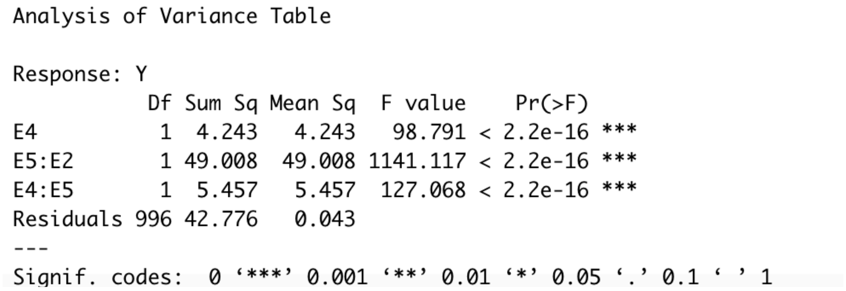
```
Call:
lm(formula = Y ~ E4 + E5:E2 + E4:E5, data = newef)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.67221 -0.14654  0.00157  0.13736  0.65997 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 14.2511011  0.0936757  152.13   <2e-16 ***
E4          -0.1774059  0.0114384  -15.51   <2e-16 ***
E5:E2        0.0102490  0.0006677   15.35   <2e-16 ***
E4:E5        0.0104480  0.0009269   11.27   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2072 on 996 degrees of freedom
Multiple R-squared:  0.5785,	Adjusted R-squared:  0.5772 
F-statistic: 455.7 on 3 and 996 DF,  p-value: < 2.2e-16
```

BIC for final model:



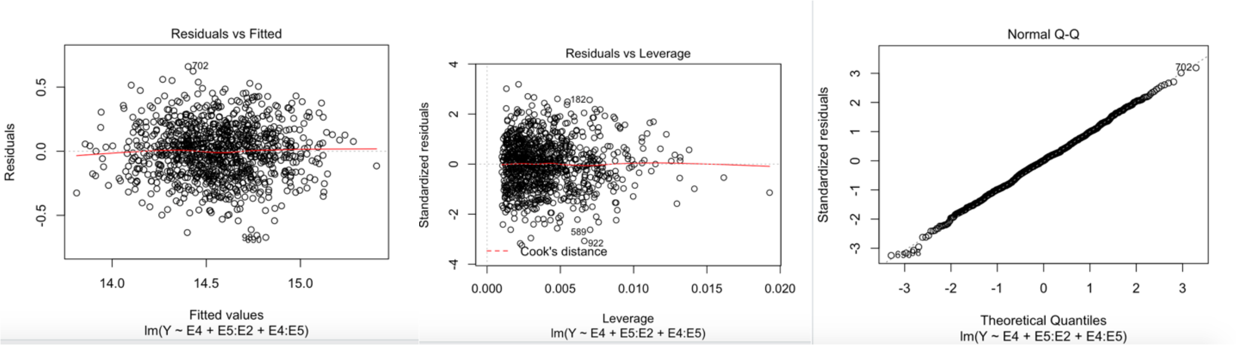
```
[1] -279.3625
```

ANOVA Function for final model:



```
Analysis of Variance Table

Response: Y
           Df Sum Sq Mean Sq  F value    Pr(>F)    
E4          1  4.243   4.243   98.791 < 2.2e-16 ***
E5:E2       1 49.008  49.008 1141.117 < 2.2e-16 ***
E4:E5       1  5.457   5.457  127.068 < 2.2e-16 ***
Residuals 996 42.776   0.043                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Plots for final model:



Works Cited:

Caspi et al. “Influence of Life Stress on Depression: Moderation by a Polymorphism in the 5-HTT Gene.” 18 July 2003.

Lumley, Thomas, and Alan Miller. “Regsubsets: Functions for Model Selection in Leaps: Regression Subset Selection.” *Regsubsets: Functions for Model Selection in Leaps: Regression Subset Selection*, 17 Jan. 2020, rdrr.io/cran/leaps/man/regsubsets.html.

“Package ‘leaps’”, https://cran.r-project.org/web/packages/leaps/leaps.pdf

“Ridge and Lasso Regression: L1 and L2 Regularization”, https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b

“6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples)”, https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779

“6.6: Ridge Regression and the Lasso.” Science.smith.edu, www.science.smith.edu/~jcrouser/SDS293/labs/lab10-r.html

R CODE

```
setwd("~/Desktop/College/AMS 578 Proj")
mydata578 <- cbind(read.csv("AMS578_Y_1535.csv",header =
T),read.csv("AMS578_E_1535.csv",header = T),read.csv("AMS578_G_1535.csv",header = T))
mydata578<- mydata578[2:28]
library(mice)
tempData <- mice(data=mydata578s,m=5,method='pmm', maxit=3,seed= 123)
mydata578<-complete(tempData,3)
shapiro.test(mydata578$Y)

library(MASS)
bc<-boxcox(lm(Y~(.),data=mydata578),data=mydata578)
lambda<-bc$x[which.max(bc$y)]
lambda

Y<- sqrt(mydata578$Y)
other<- mydata578[-1]
mydata578<- cbind(Y,other)
hist(mydata578$Y)
shapiro.test(mydata578$Y)
qqplot(1:1000, Y)


###testing final model
E2<- mydata578$E2
E4<- mydata578$E4
E5<- mydata578$E5
Y<- 14.25110112-0.17740590*E4+ 0.01024896*E5*E2+0.01044796*E4*E5


first<-lm(formula=Y~E1+E2+E3+E4+E5+E6,data=mydata578)
summary(first)
BIC(first)

first11<-lm(formula=Y~E2+E4+E5,data=mydata578)
summary(first11)
BIC(first11)
```

```
null<-lm(Y~1,data=mydata578)
full<-lm(Y~.,data=mydata578)
fit1<-step(null,scope=list(lower=null,upper=full),direction='both')
fit1
#E5 + E2 + E4 + G14
summary(fit1)
BIC(fit1)
#WITHOUT STEP
fullNOstep<-lm(Y~.,data=mydata578)
summary(fullNOstep,data=mydata578)

friend1<-lm(Y~E5 + E2 + E4 + G5,data=mydata578)
summary(friend1)
BIC(friend1)

null<-lm(Y~1,data=mydata578)
full<-lm(Y~.^2,data=mydata578)
fit<-step(null,scope=list(lower=null,upper=full),direction='both')
summary(fit)
fit
anova(fit)
#Y ~  E5 + E2 + E4 + G14 + G5 + E5:E2
BIC(fit)


#WITHOUT STEP
####JUST GETS ME NOTHING SIGNIFICANT IN THE END
fullNOstep2<-lm(Y~.^2,data=mydata578)
summary(fullNOstep2,data=mydata578)
model1<-lm( Y ~ E2+E4+E5+G2 +  G11 + E2:G1 +
            E2:G2 + E2:G4 + E2:G12 +
            E2:G17 + E3:G6 + E3:G15 + E3:G19 + E4:G2 + E5:G9 + E5:G18 + E6:G20+ G1:G11+ G1:G15
+G2:G6 +G2:G20,data=mydata578)
summary(model1)

model1<-lm( Y ~ E5 + E2 + E4 + G5 + G2 + E1 + G14 + G11 + E5:E2 +
            E2:G2 + E4:G2 + E5:E4 + G5:G14 + E2:G5,data=mydata578)
summary(model1)
BIC(model1)

model2<-lm(Y~E4 + G14 + G2 + E5:E2 + E5:E4 + E2:G2 ,data=mydata578)
summary(model2)
BIC(model2)

null<-lm(Y~1,data=mydata578)
full<-lm(Y~(.)^3,data=mydata578)
fit2<-step(null,scope=list(lower=null,upper=full),direction='both')
summary(fit2)
fit2
BIC(fit2)
anova(fit2)
```

```
final1<-lm( Y ~ E5 + E2 + E4 + G5 + G2 + E1 + G14 + G11 + E5:E2 +
            E2:G2 + E4:G2 + E5:E4 + G5:G14 + E2:G5,data=mydata578)
summary(final1)
BIC(final1)
plot(resid(final1) ~ fitted(final1), main='Residual Plot')
anova(final1)

Anovainteresting<- lm( Y ~ E5 + E2 + E4 + G5 + E5:E2+ E2:G2 + E4:G2,data=mydata578)
summary(Anovainteresting)
BIC(Anovainteresting)

Anovainteresting1<- lm( Y ~  E4 + G5 + E5:E2,data=mydata578)
summary(Anovainteresting1)
BIC(Anovainteresting1)

final1<-lm( Y ~ E4 + G2 + E1 + G14 + G11 + E5:E2+ E2:G2 + E4:G2 + E5:E4 +
G5:G14,data=mydata578)
summary(final1)
BIC(final1)
plot(resid(final1) ~ fitted(final1), main='Residual Plot')
anova(final1)

library(leaps)
M <- regsubsets( model.matrix(final1)[,-1], Y,
              nbest = 1 , nvmax=5,
              method = 'forward', intercept = TRUE )
temp <- summary(M)

Var <- colnames(model.matrix(final1))
M_select <- apply(temp$which, 1,
               function(x) paste0(Var[x], collapse='+'))

r<-data.frame(cbind( model = M_select, adjR2 = temp$adjr2, BIC = temp$bic))
r

final2<-lm(Y ~ E4+G2+E5:E2+E2:G2+ E4:E5+ E4:G2,data=mydata578)
summary(final2)
step(final2)
anova(final2)
BIC(final2)
par(mfrow=c(1,1))
plot(final2,1)
plot(final2,5)

final3<-lm(Y ~  E4+G2+E5:E2+E2:G2+ E4:E5,data=mydata578)
summary(final3)
par(mfrow=c(1,1))
plot(final3,1)
plot(final3,5)
```

```
vif(final3)
anava(final3)

final4<-lm(Y ~  E4+E5:E2+ E4:E5,data=mydata578)
summary(final4)
BIC(final4)
par(mfrow=c(1,1))
plot(final4,1)
plot(final4,5)
plot(final4,2)
vif(final4)
anova(final4)

coef(final4)

plot(resid(final3) ~ fitted(final3), main='Residual Plot')
anova(final1)

library('glmnet')
train.data <- mydata578[,-1]
train.data <- cbind(mydata578$Y, train.data)
colnames(train.data)[1] <- "Y"
x.train <- model.matrix(Y ~ .^3, data=train.data)[,-1]
y.train <- train.data$Y
grid <- 10^seq(10,-2,length=100)

lasso.mod <- glmnet(x.train,y.train,alpha=1,lambda=grid)
plot(lasso.mod)
set.seed(123)
cv.out<-cv.glmnet(x.train,y.train,alpha=1)
plot(cv.out)

lam <- cv.out$lambda.1se
lam
out <- glmnet(x.train, y.train, alpha=1, lambda=lam)

lasso.cf<-predict(out,type='coefficients',s=lam)
dim(Y)
lass.c<-coef(out)
lass.c

x.lasso<-x.train[,which(lass.c!=0)]
lasso.data<-data.frame(cbind(y.train,x.lasso))
colnames(lasso.data)[1]<-'Y'
fit4<-lm(Y~.,data=lasso.data)
summary(fit4)

M_E <- lm(Y ~ E5+E6+E2:E6 , data=mydata578)
summary(M_E)
```