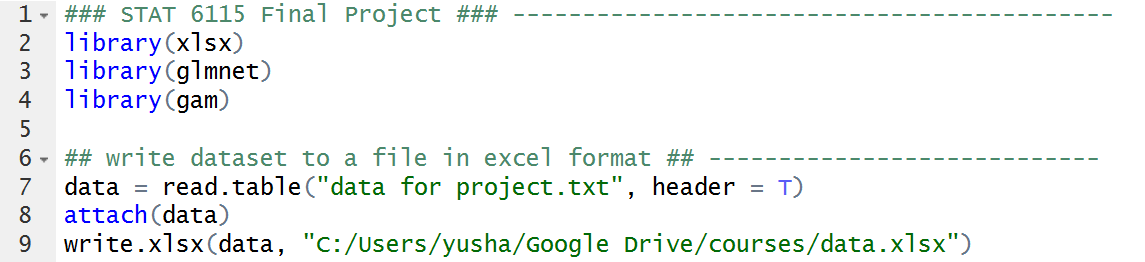
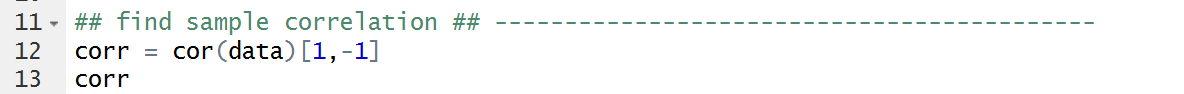
**STAT 6115 Final Project**

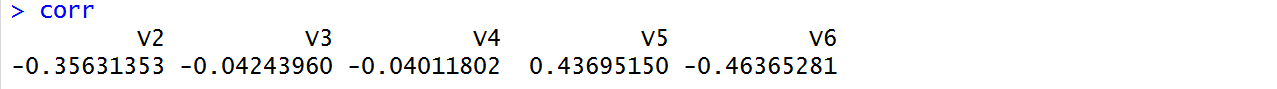
**Sha Yu**

1. Write the dataset to a file in excel form.



1. Find sample correlation between the response and each of the covariates.

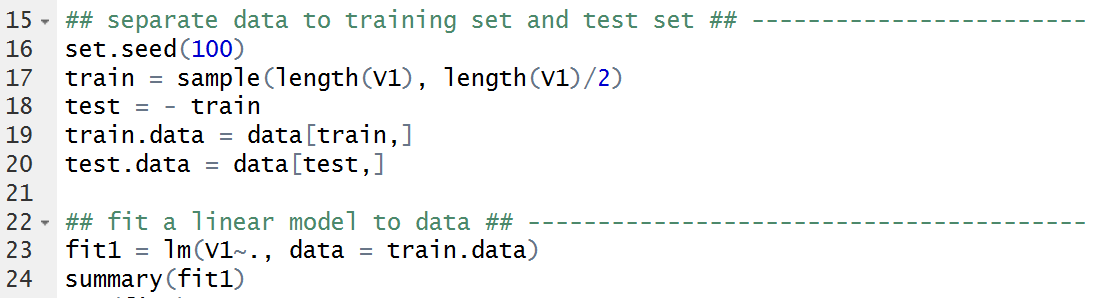


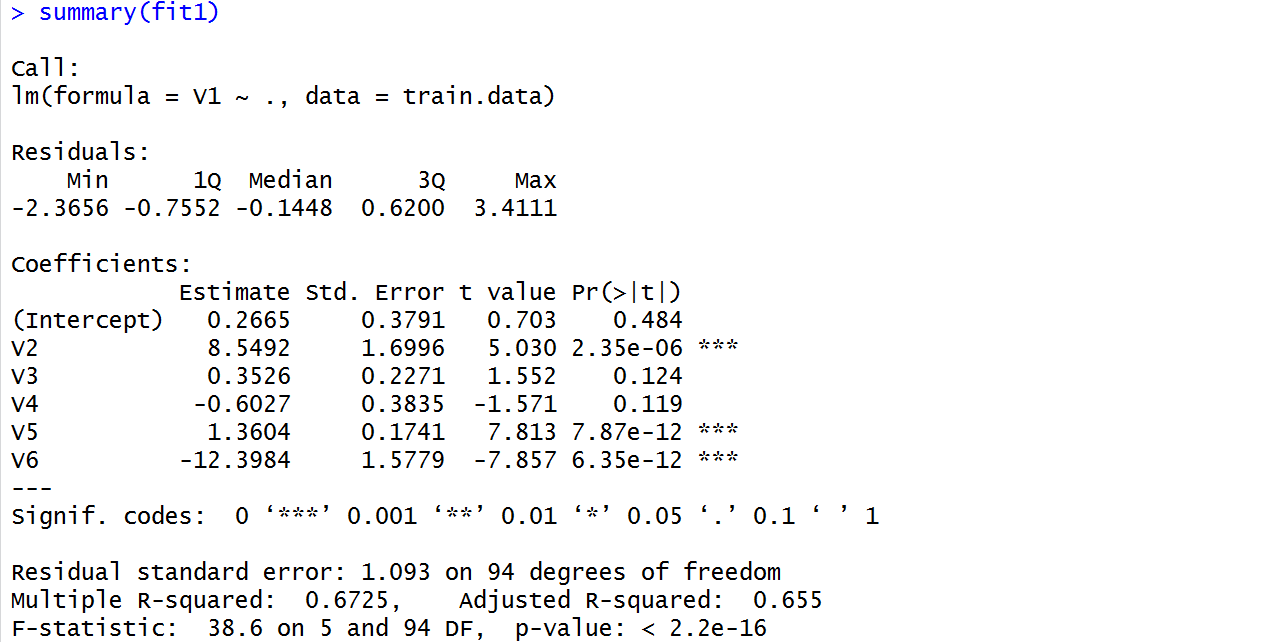


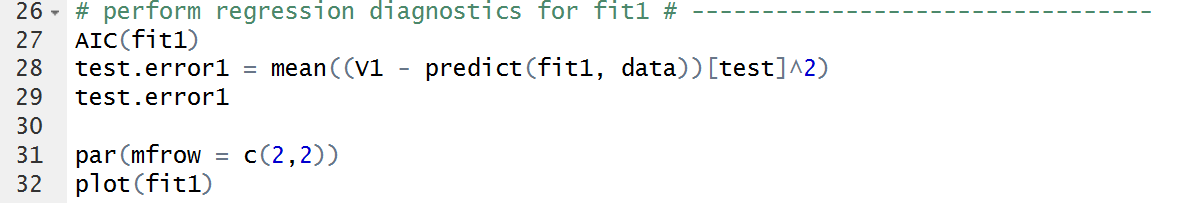
* We can tell from the output that there is little correlation between V1 and V3, V1 and V4.

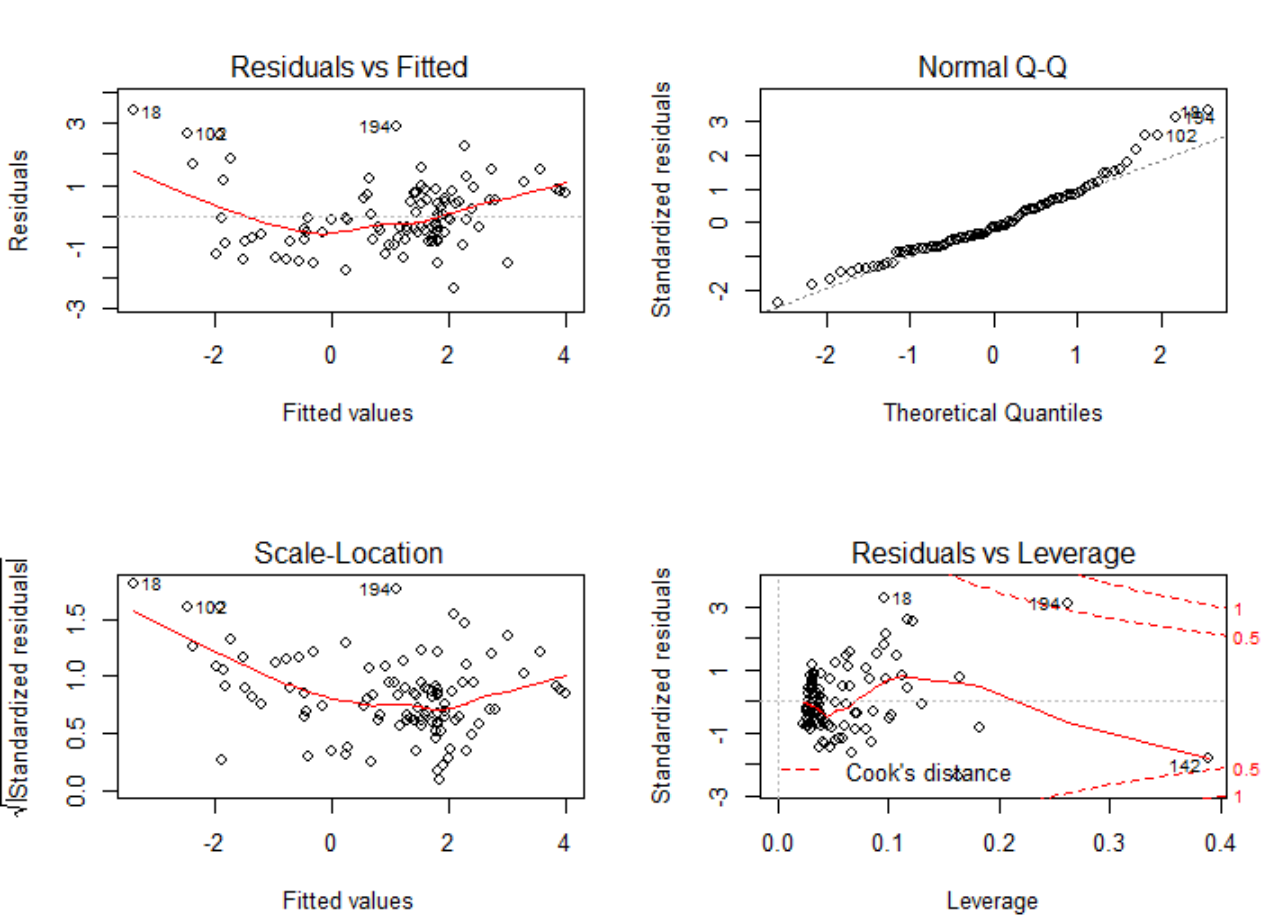
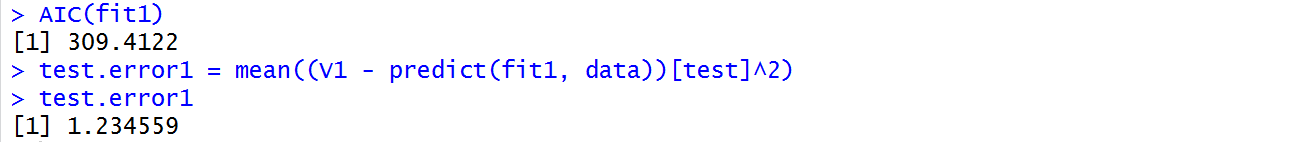
1. Propose an initial model to fit the dataset and check its appropriateness using regression diagnostics.

* First, I separated the data into training dataset and test dataset. And then I fit a linear model to the training dataset. From the summary(fit), p value of V3 and V4 are greater 0.05, which indicated that these two covariates are not significant.
* Next, I calculated AIC and the test error on test dataset. Also checked the performance by doing regression diagnostics. The AIC of fit1 is 309.4122 and the test error is 1.234559.
* The plot of residuals versus fitted values indicates the presence of nonlinearity in the data. The plot of standardized residuals versus leverage indicates the presence of a few outliers and a few high leverage points.



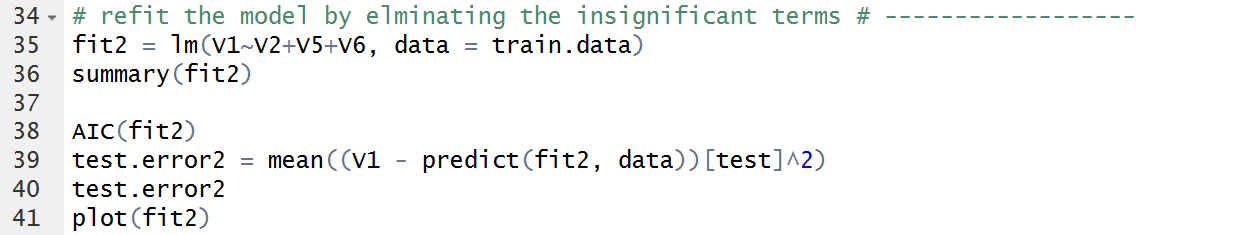


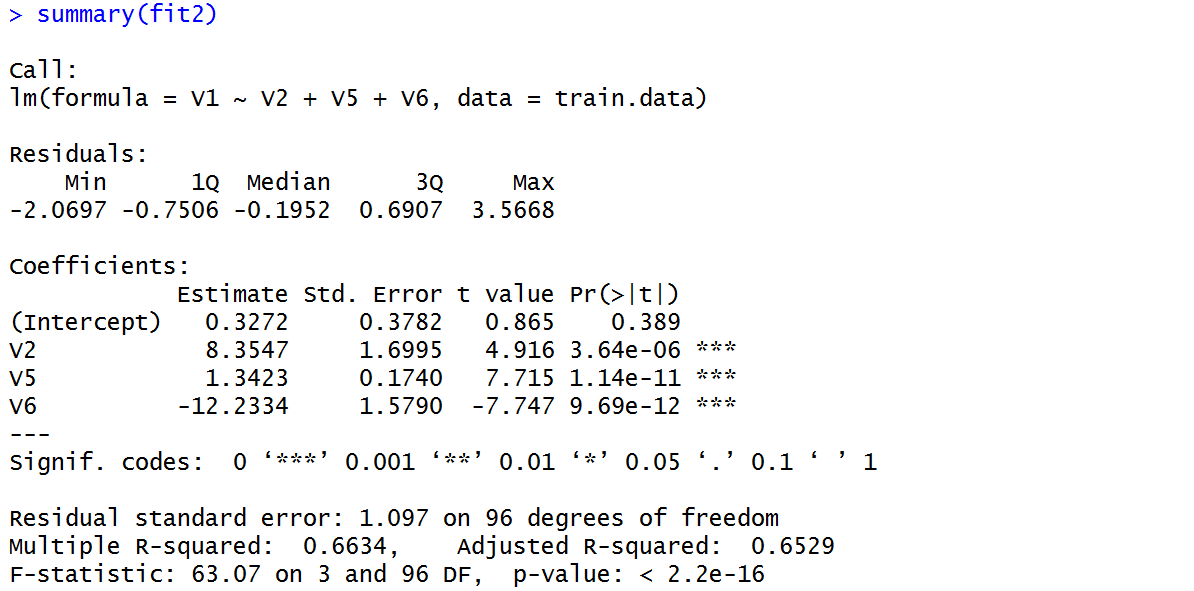


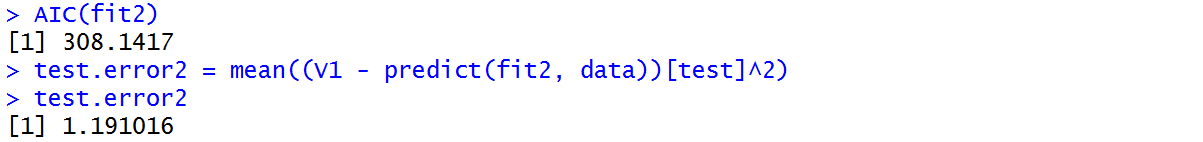


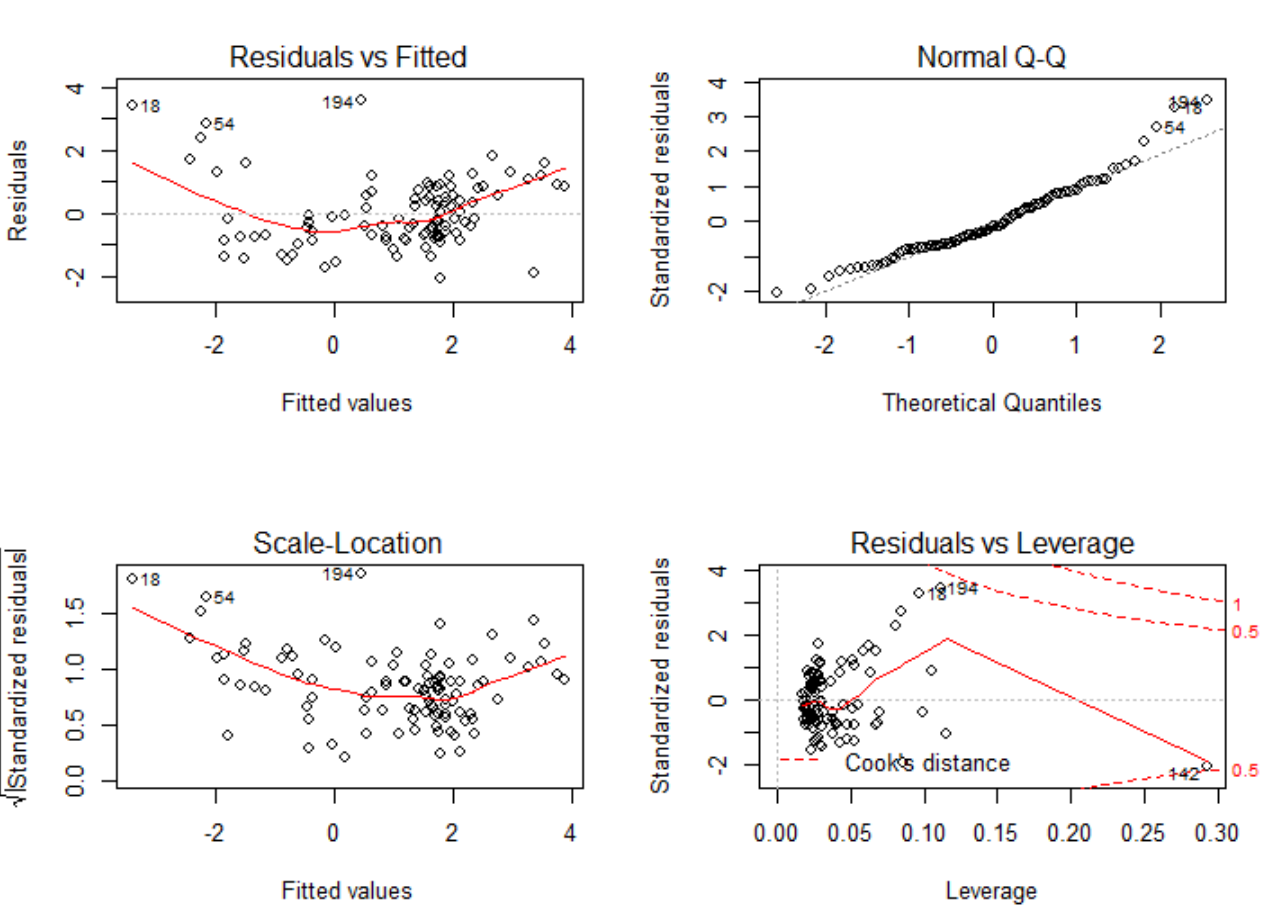
1. Refine your model based on your discovery in step 3 and check if it is appropriate.

* I refit the model by eliminating the two insignificant covariates V3 and V4. Now all variables are significant according to the summary.
* The AIC is 308.1417, which has decreased. The new test error is 1.191016 and also is smaller compared to previous test error 1.234559. Thus by eliminating the insignificant terms, the model has been refined.
* But by checking the residual plot, there is still nonlinearity. Also there are still some outliers.









1. Can you improve the model you fit in step 4? Why?

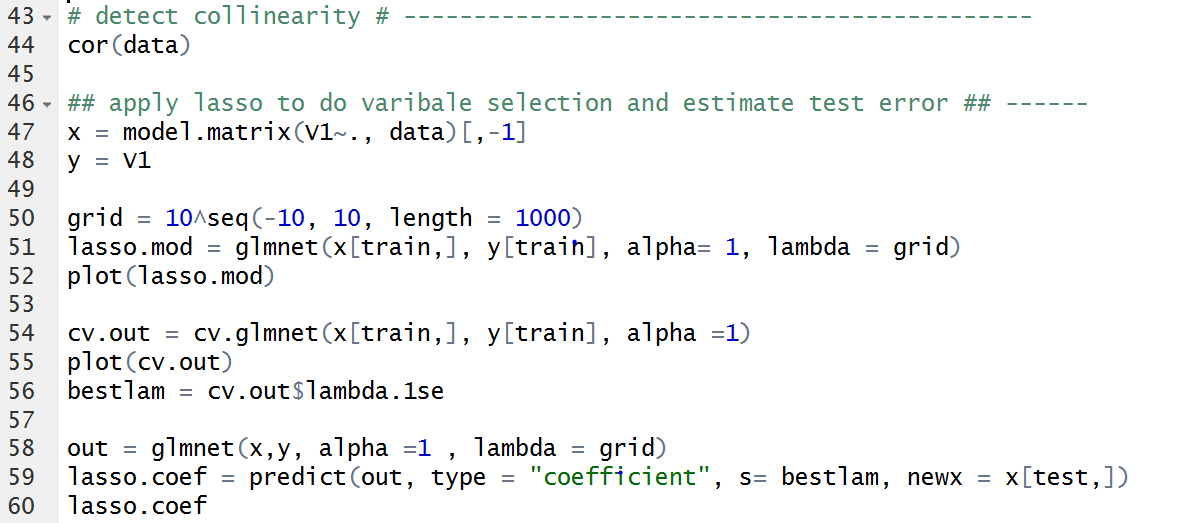
Part(1) :refine in linear models

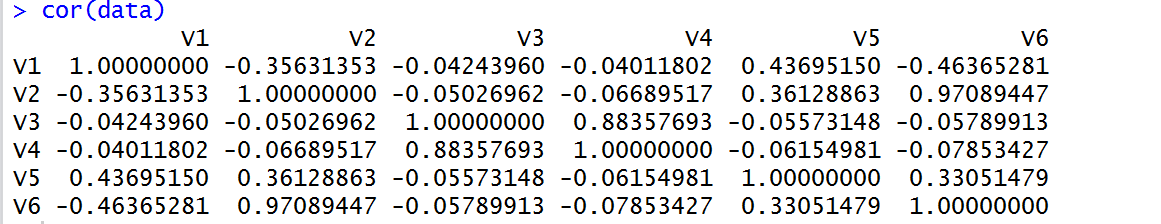
* Although step 4 has refined the model. The model is still not very appropriate. So I calculated the correlation between all covariates. I found that the correlation between V3 and V4 is 0.88357693, the correlation between V2 and V6 is 0.97089447. So there is collinearity among covariates.
* I performed lasso to do variable selection, hoping to solve the collinearity problem. I used 10-fold cross validation to choose lambda in lasso regression.

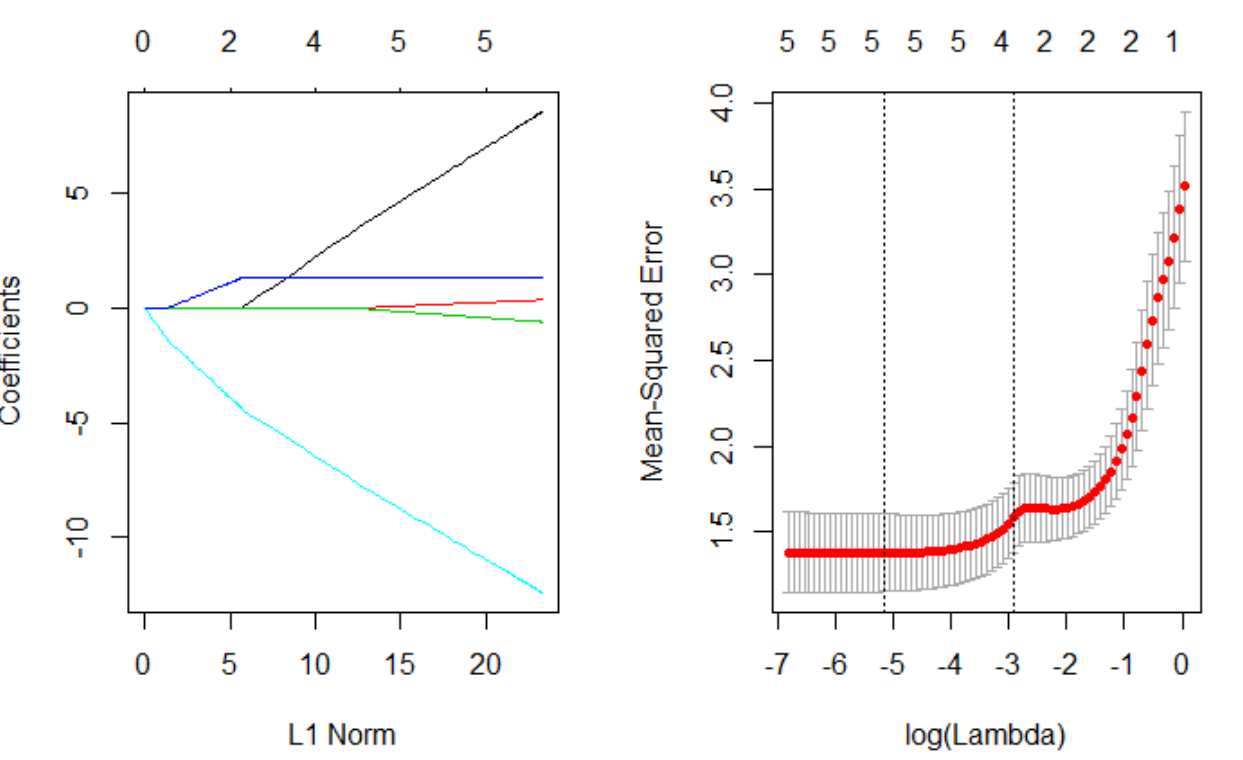
Lasso method chose covariates V5 and V6. It eliminates the insignificant variable and also only pick on among the two highly correlated covariates.

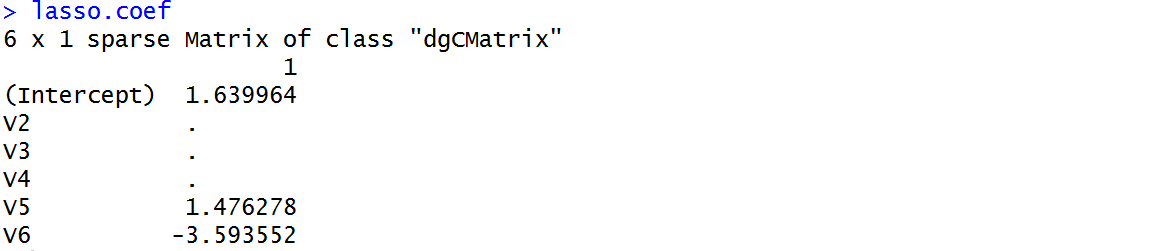
* I refit the linear model just using the variable selected by lasso. According to summary(fit), all covariates are significant. However, the new AIC is 328.594, which has increased. Also the test error is 1.500241, which is larger than the test error of the model with V2, V5 and V6.
* Check the plots, there are still nonlinearity and outliers.
* Thus by applying lasso method and eliminating V2, the model could not be refined anymore.

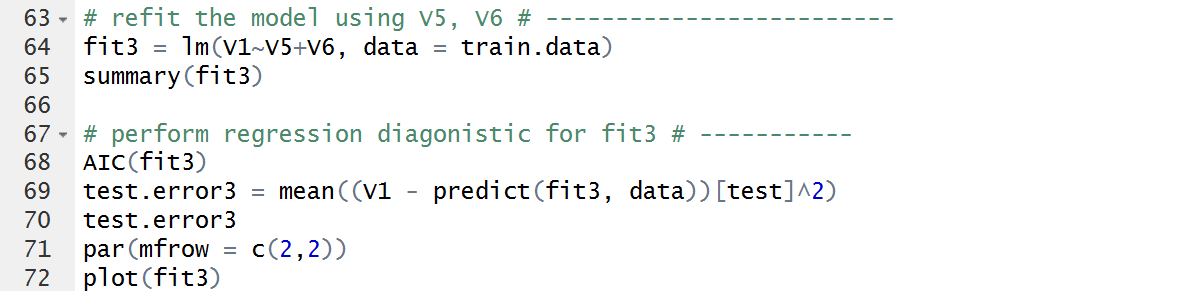
I summarized the results of using different variables to fit linear regression as below.

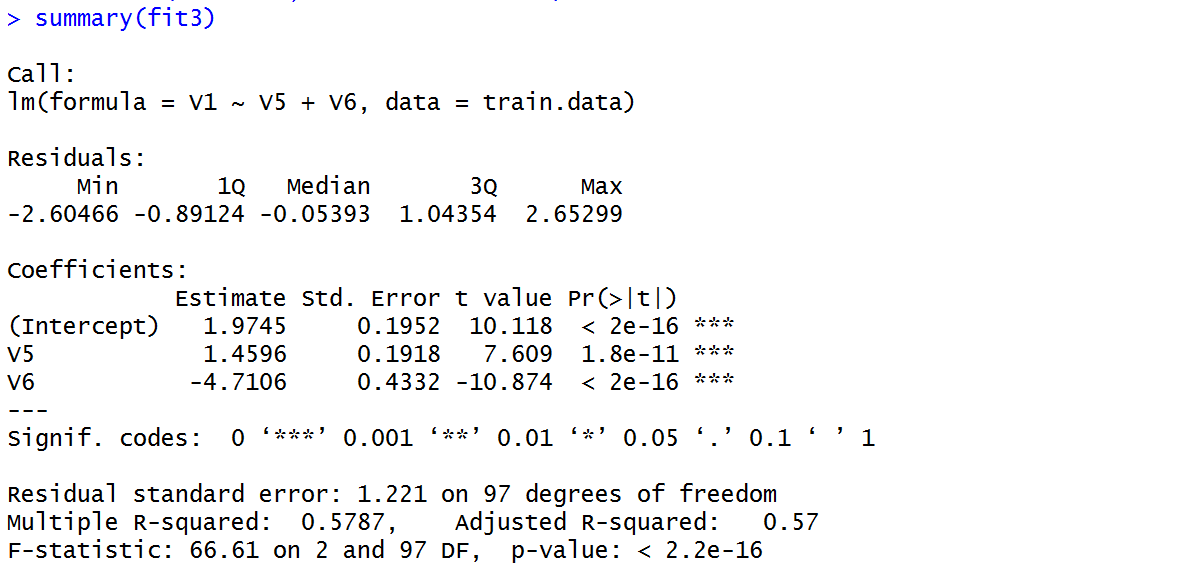


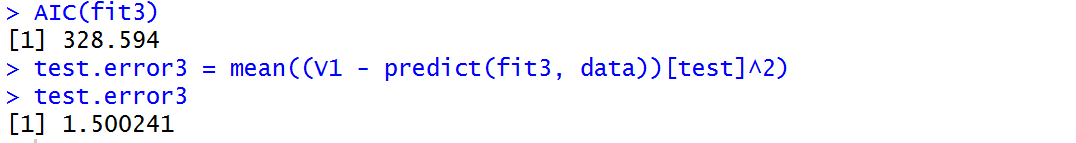


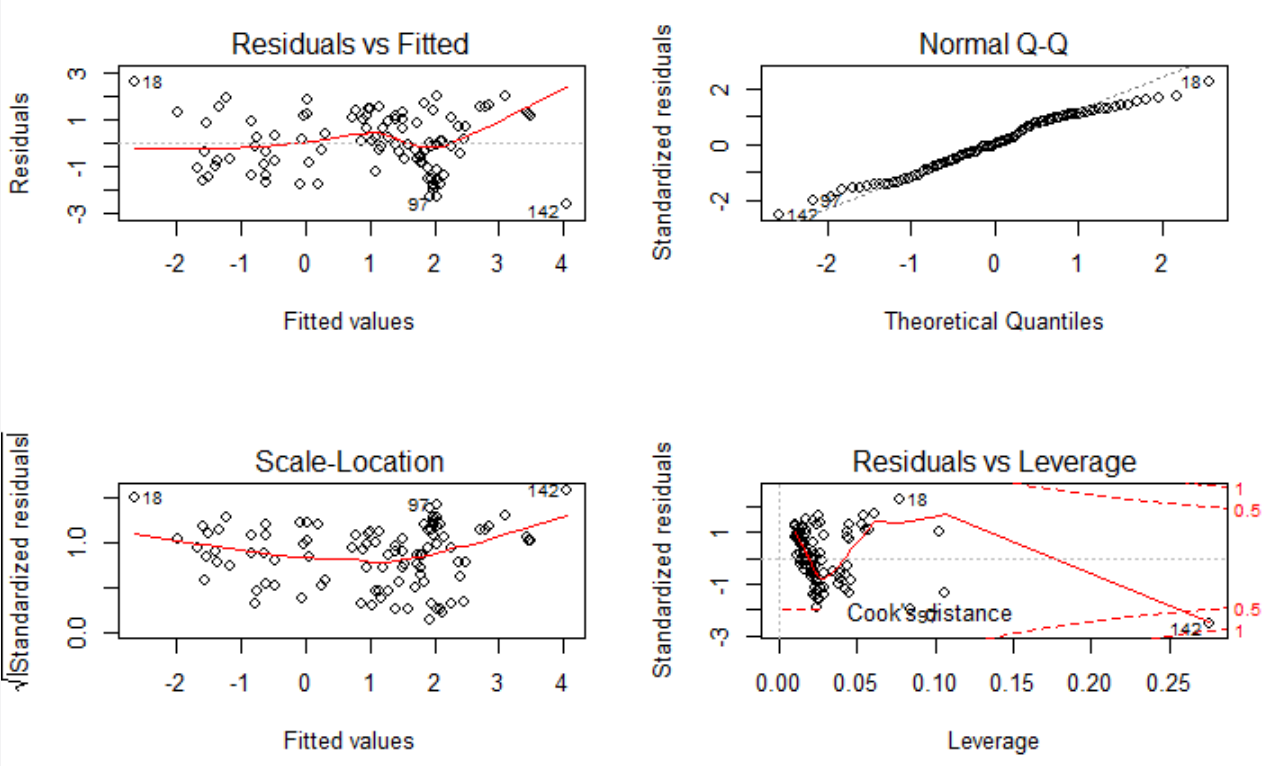






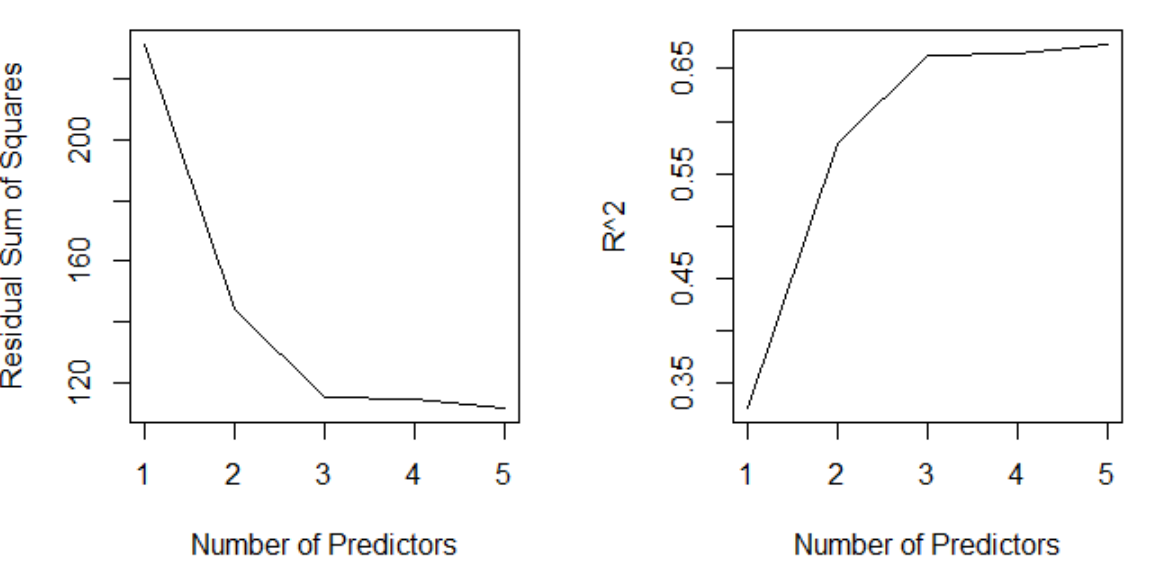
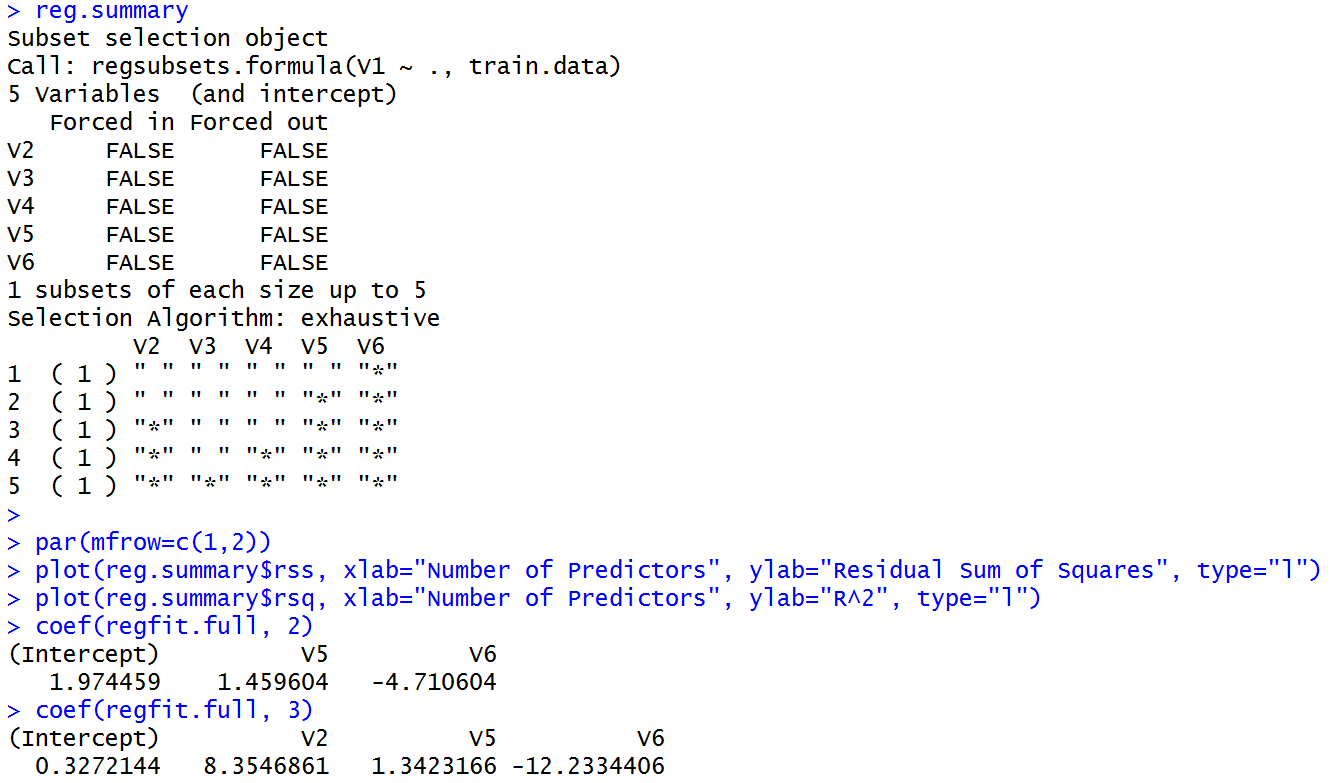
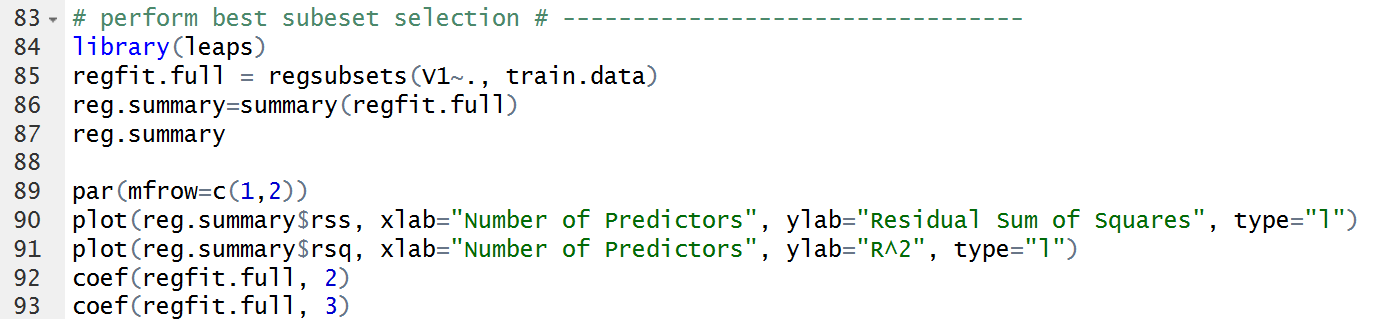




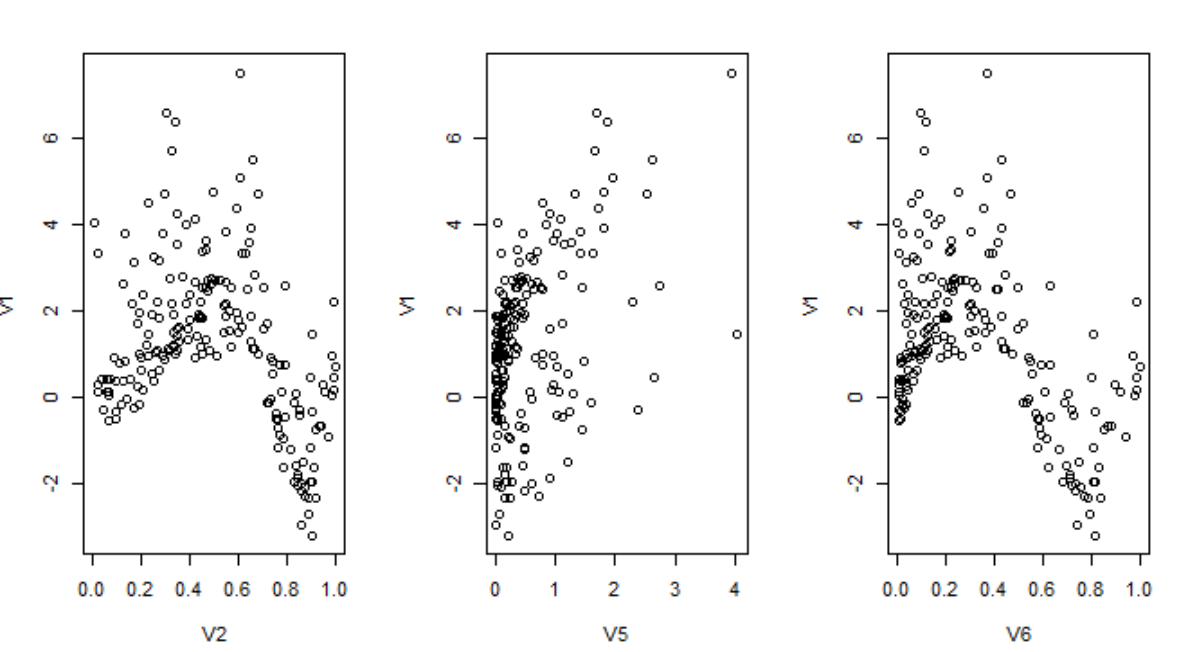


Part(2): extend to nonlinear models.

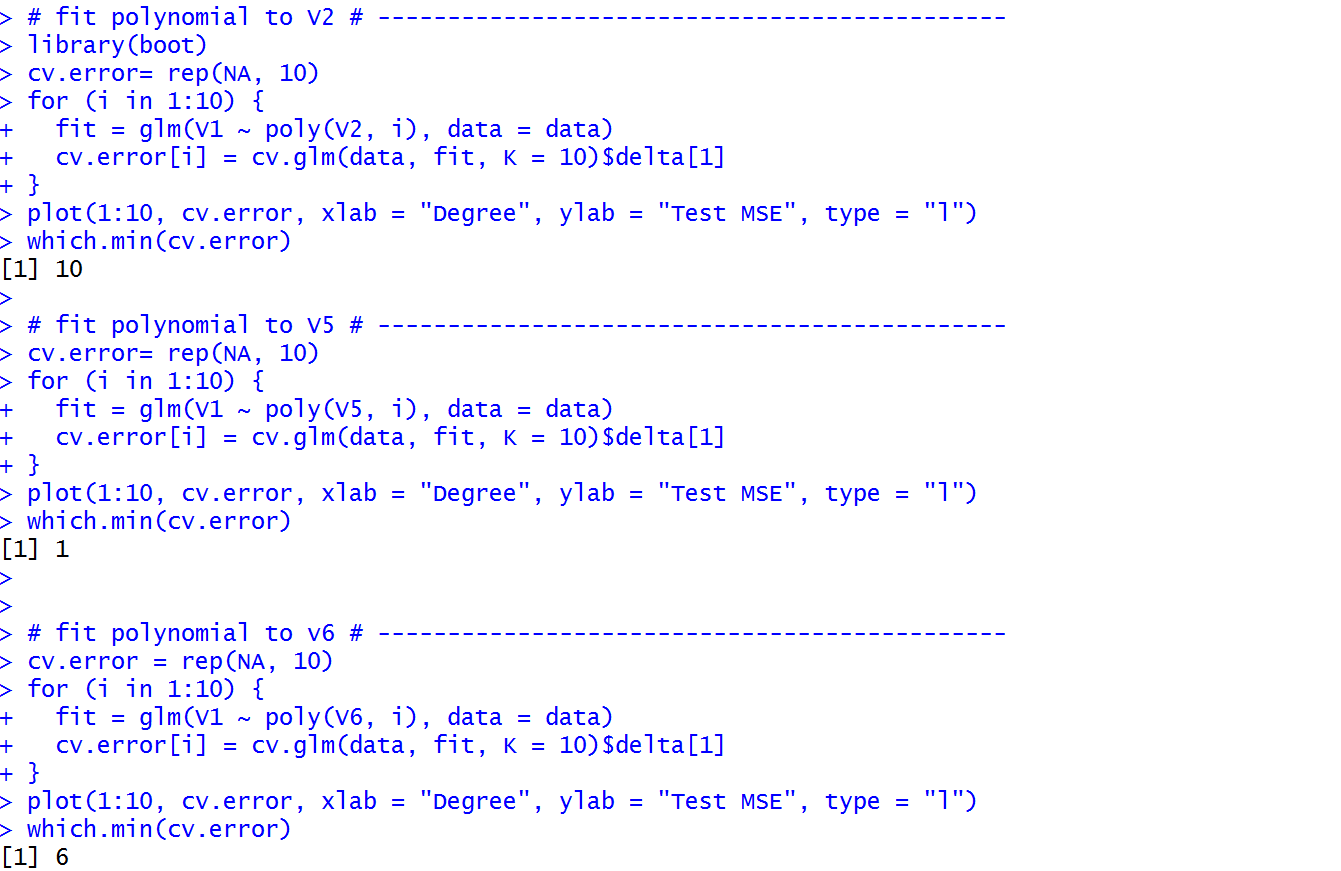
* Recall there exits nonlinear patter between response and covariates. Thus the next I tried some nonlinear models.
* Recall that V3 and V4 have little correlation with V1 and there is also collinearity among variables, so I first performed best subset selection on the training set in order to identify a satisfactory model that used just a subset of the predictors.
* According to the best subset selection output, the best 2-variable model is the model with covariates V5 and V6. The best 3-variable model is the model with covariates V2, V5 and V6.

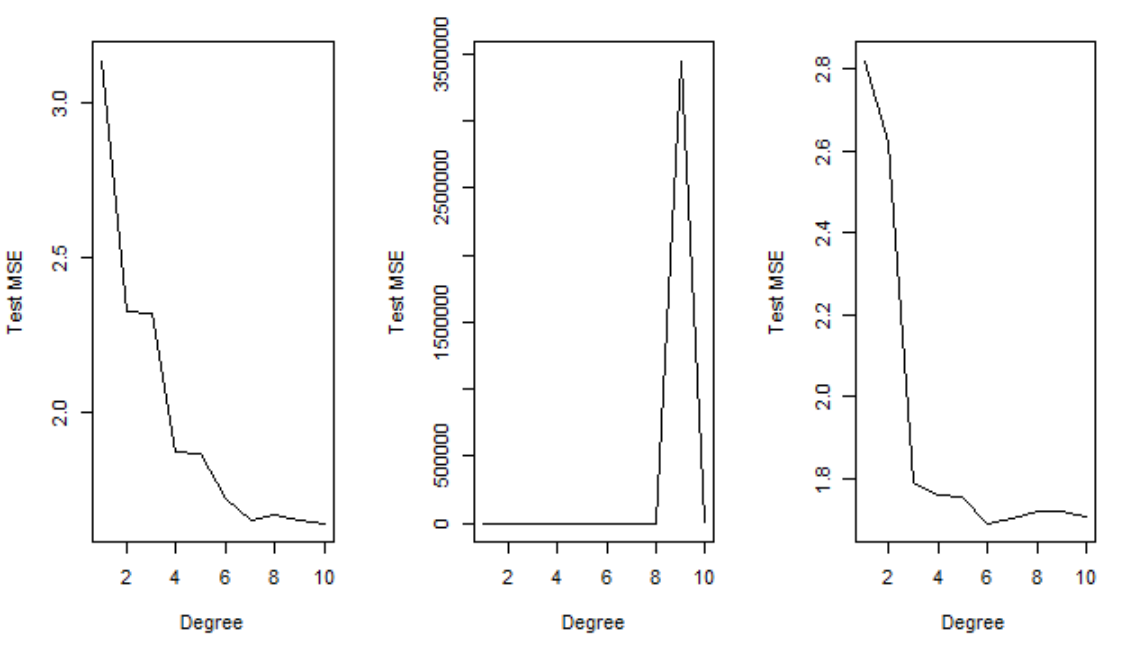


* To explore more information between V1 and V2, V5 and V6, I first did scatter plots. We can tell from the plots, there is nonlinear patter between V1 and V2, also between V1 and V6.



* First, I tried to fit univariate polynomial model between response and each covariate V2, V5, V6. I performed the 10-fold cross validation to choose the best degree for each covariate. Then I fit the generalized additive model to the training data by setting each basis function as the best polynomial. I fit a model with best 2-variable along with the model with best 3-variable.
* From the output, the best degree for V2 is 10, best degree for V5 is 1 and best degree for V6 is 6.

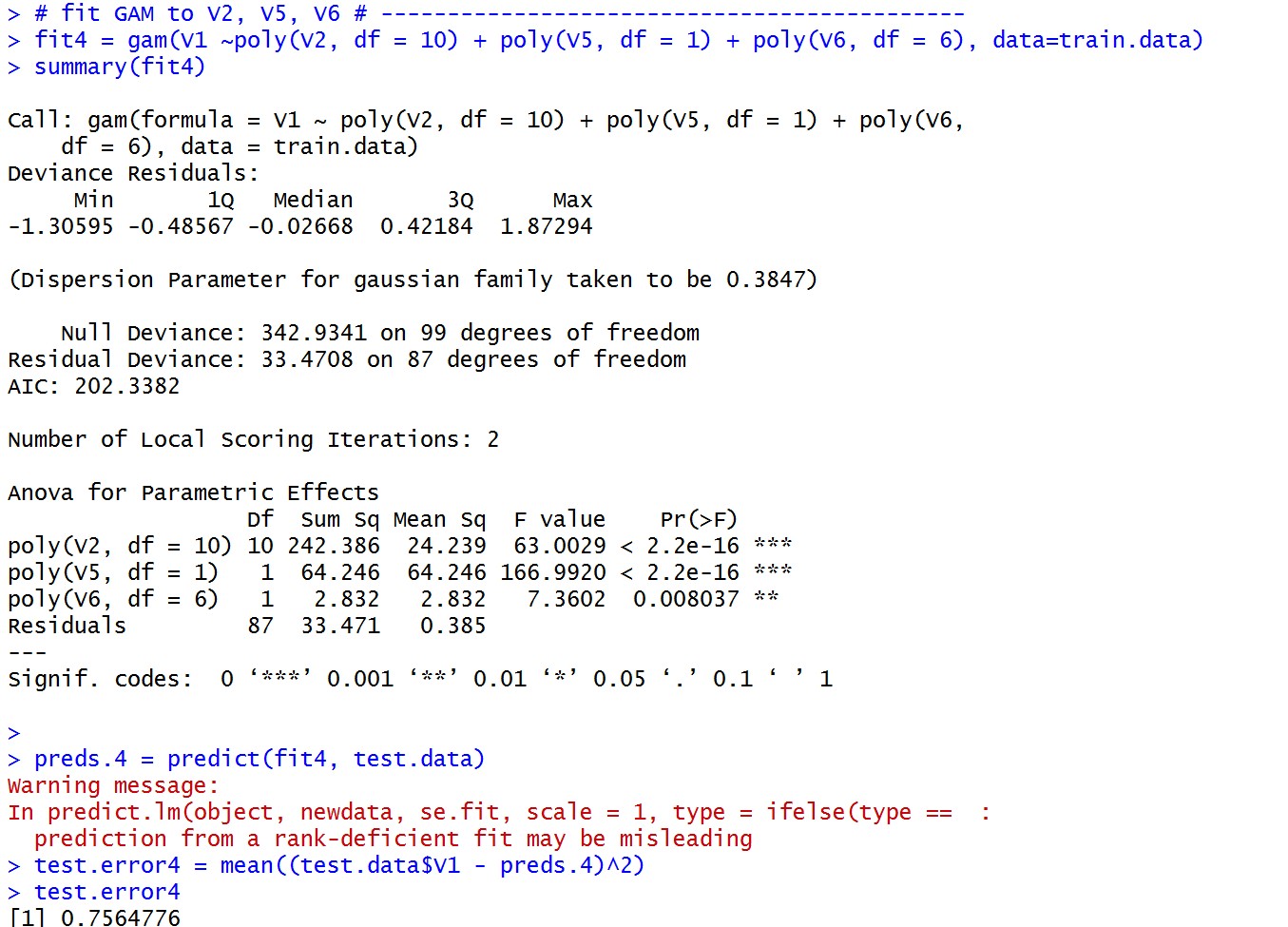




* For best 3-variable model, fit a GAM to V2, V5 and V6.

The test error is 0.7564776, which is much smaller compared to any of the test error by fitting linear model.

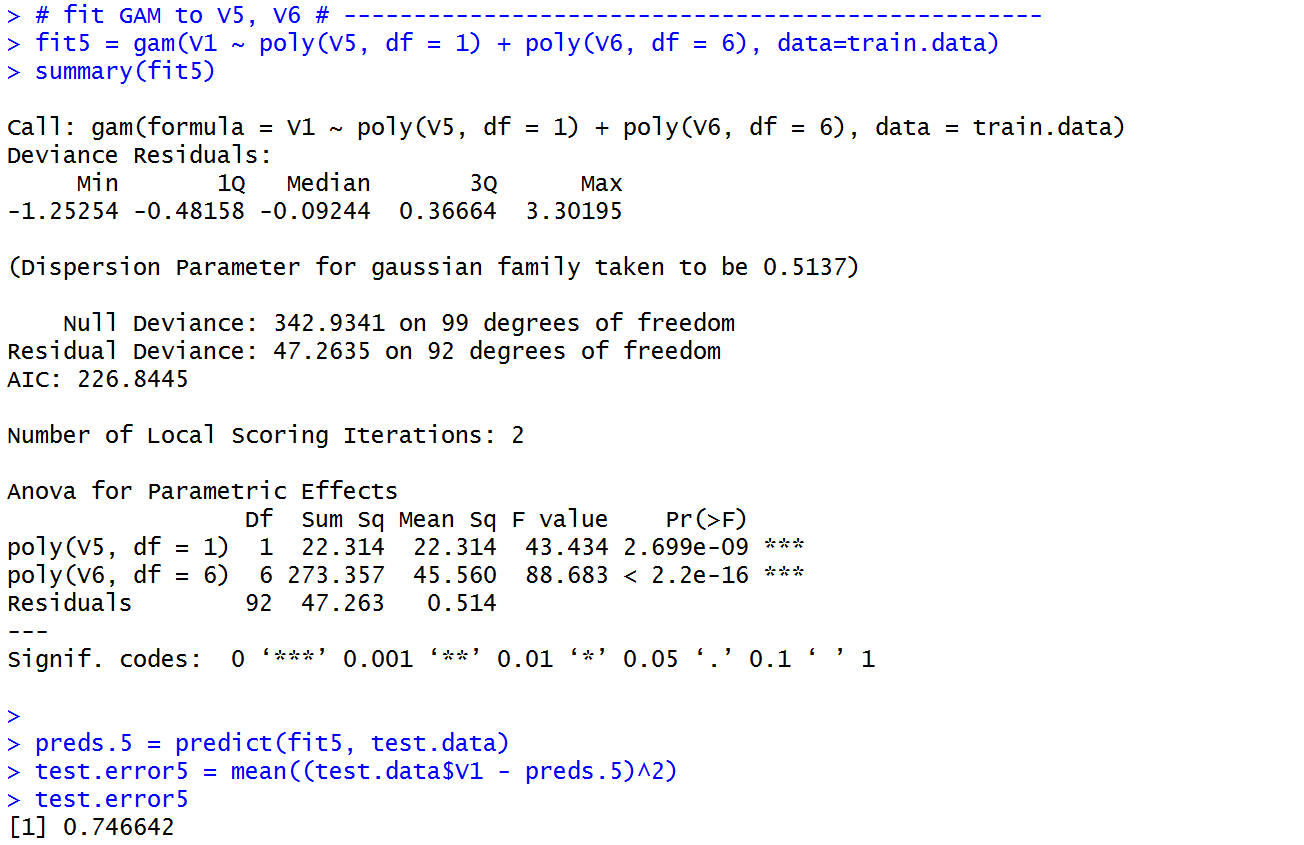
* Thus this new model significantly improves the performance.



* For best 2-variable model, fit a GAM to V5 and V6.

The test error is 0.746642, which is much smaller compared to any of the test error by fitting linear model. And it is also smaller than the test error of GAM with variable V2, V5 and V6.

* So far, GAM models fit the data much better than multiple linear models. And it is the best model so far.

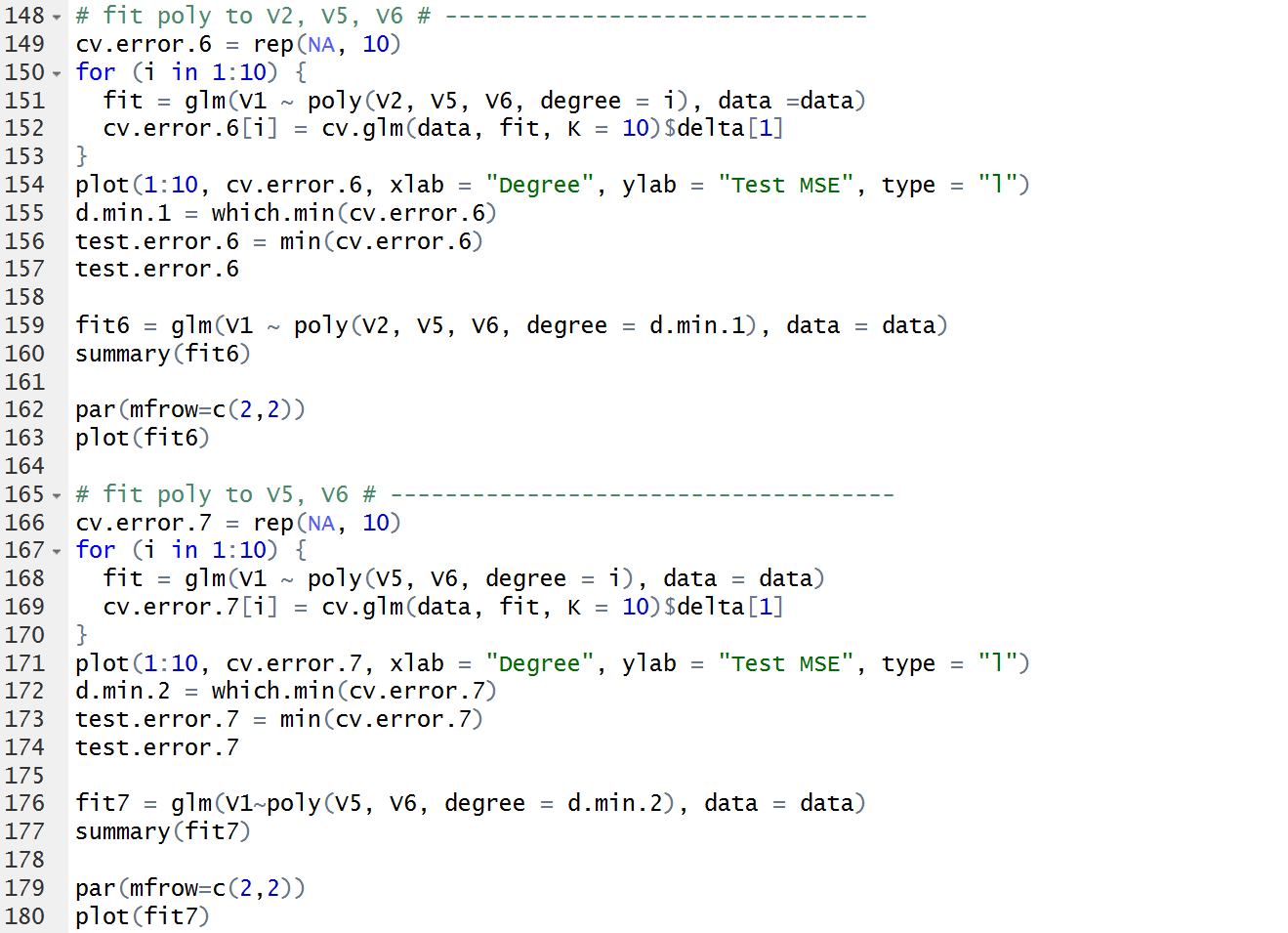


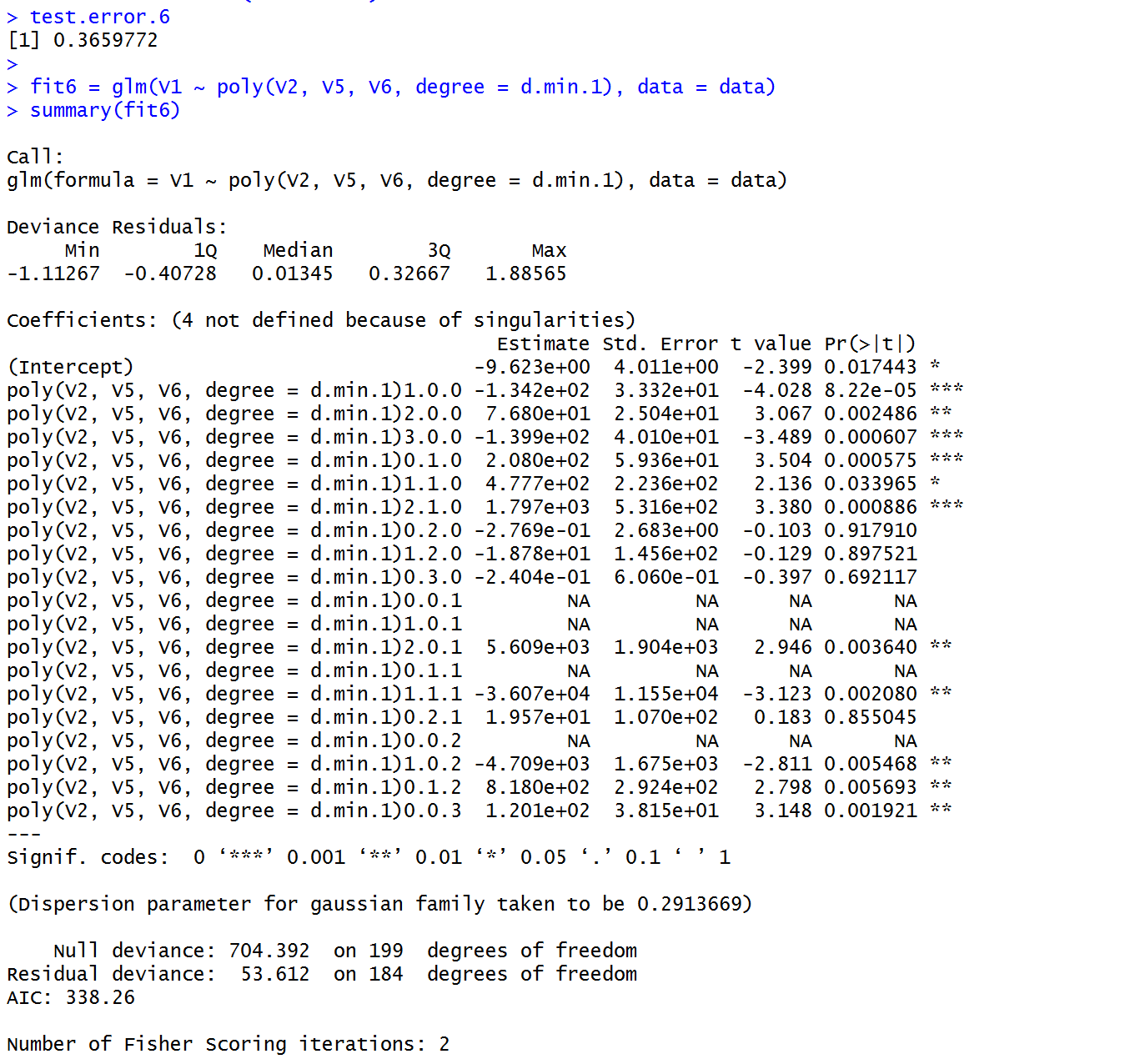
* I want to see whether I could refine the model more. Considering that GAM models do not consider the interaction between covariates. So in order to include the possible interaction between covariates, I fit polynomial to covariates by using 10-fold cross validation to choose the best degree.
* For best 3- variable model (V2, V5 and V6), the best degree is 3. According to summary, some interactions are significant. Test error is 0.3245652, both significantly reduced further. But some coefficients are not defined due to singularities. Consider together with the high correlation between V2 and V6, it is better to only include one of them in the model.

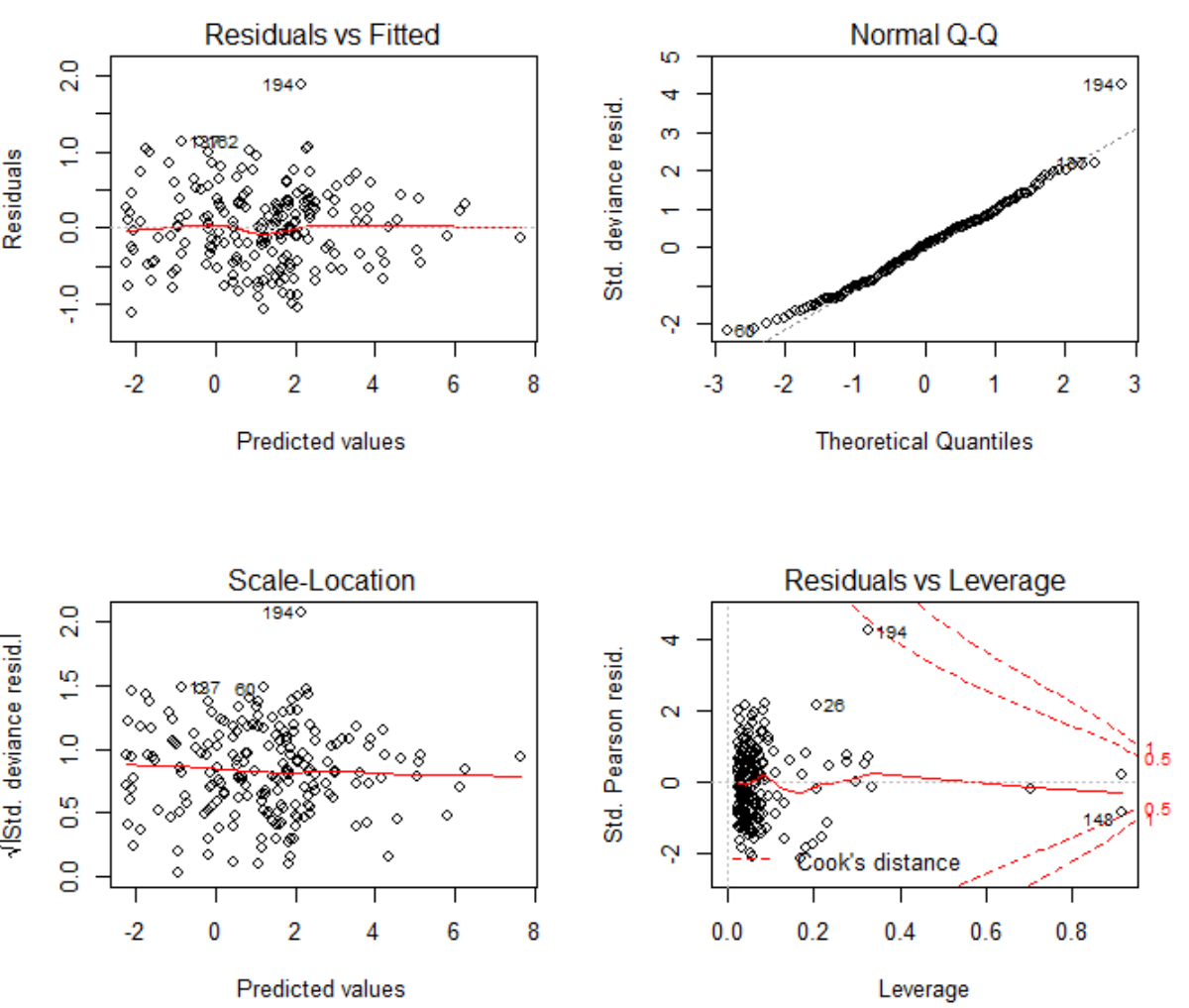
The diagnostic plots show that the model is adequate. There is almost no pattern in residual plot and also the QQ plot looks good.

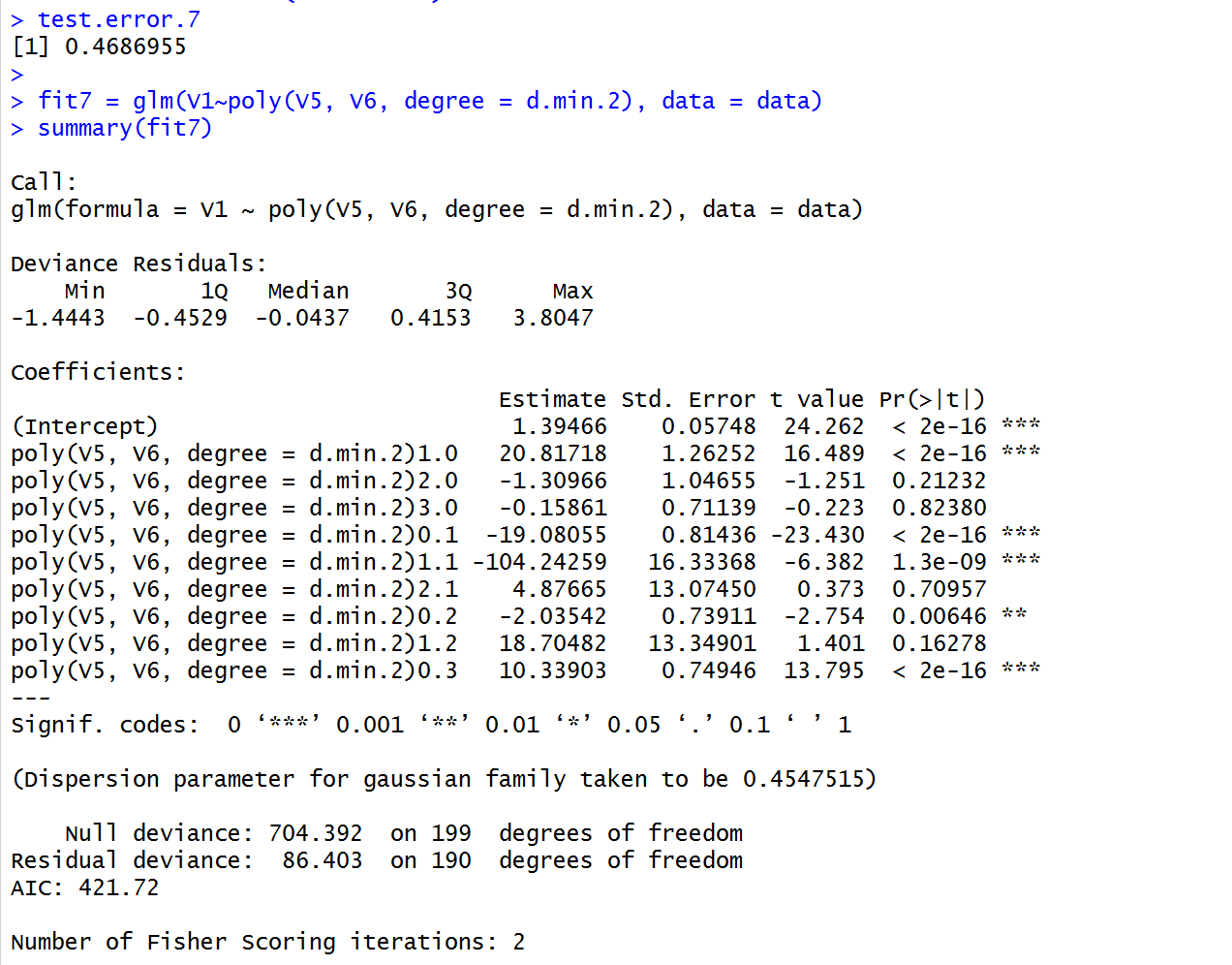
* For best 2-variable model (V5 and V6), there are also significant interaction terms. Also, the test error is 0.4686955.

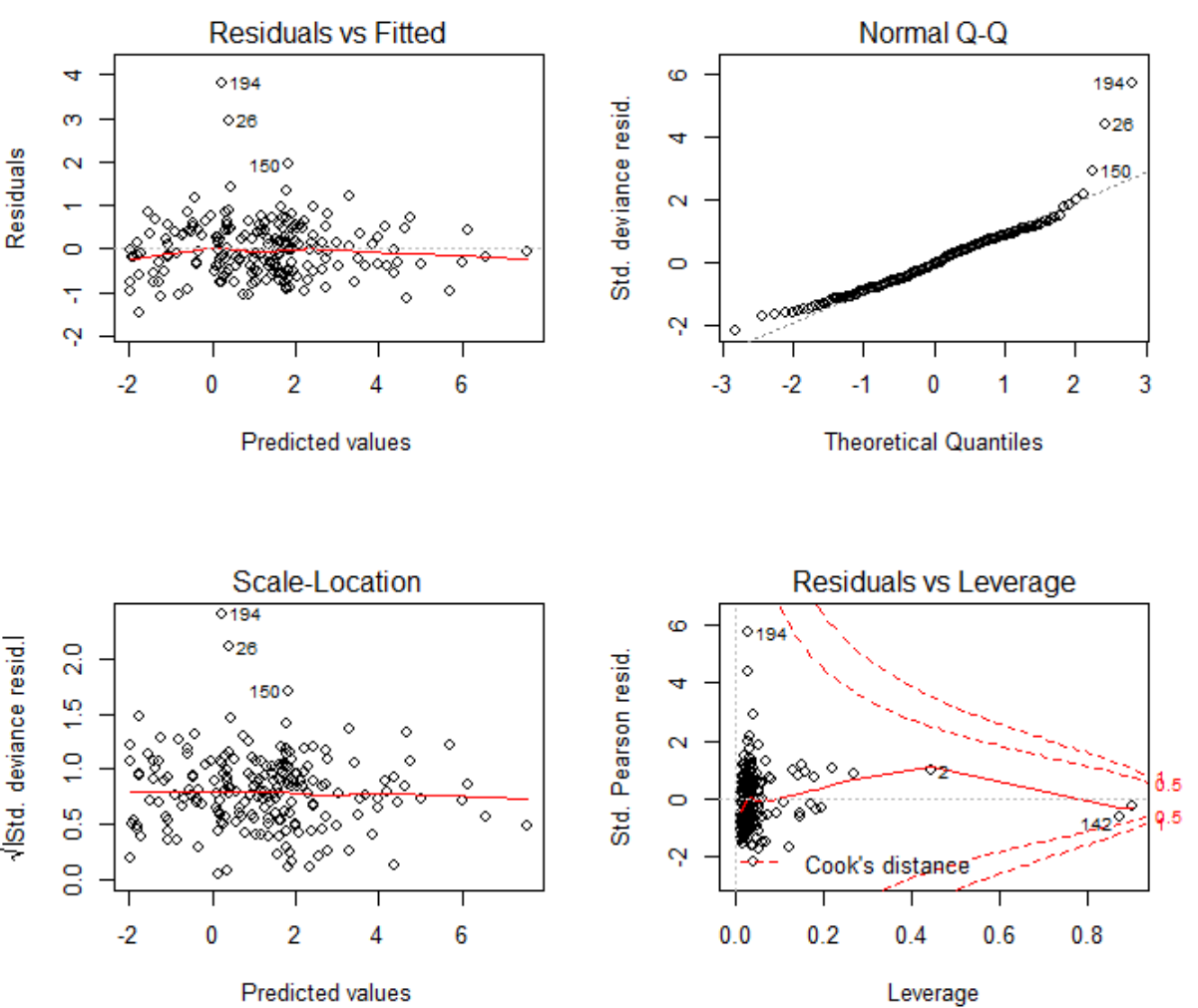
The diagnostic plots show that the model is adequate. There is almost no pattern in residual plot and also the QQ plot looks good except some outliers.











1. Write down the final model which is the best for fitting the dataset and provide the 95% confidence intervals/band s to your estimators of the parameters/curves.

* Summary of all the models fitted

|  |  |  |
| --- | --- | --- |
|  | model | Test Error |
| best 3-variable model:  V2, V5, V6 | Linear | 1.191016 |
| GAM | 0.7390142 |
| Poly | 0.3659772 |
| Best 2-variable model:  V5, V6 | Linear | 1.500241 |
| GAM | 0.7466420 |
| Poly | 0..4686955 |

* Although fit6 has smaller test error than fit7 but it has singularities. So the best model I chose fit6. Check the summary(fit7), the significant terms are V5, V6, V6^2, V6^3 and V5\*V6. So I fit the model by eliminating the insignificant terms. From the summary, this time all variables are significant. Also the diagnostics plots show that the fit is adequate although there still exits outliers.
* The final model is:
* The 95% confidence interval for parameters are:

