Name: Adarsh Shankar

ID: 2117611

**DSC 423: Data Analytics and Regression**

**Assignment 09**

***Honor Statement: “I have completed this work independently. The solutions given are entirely my own work.”***

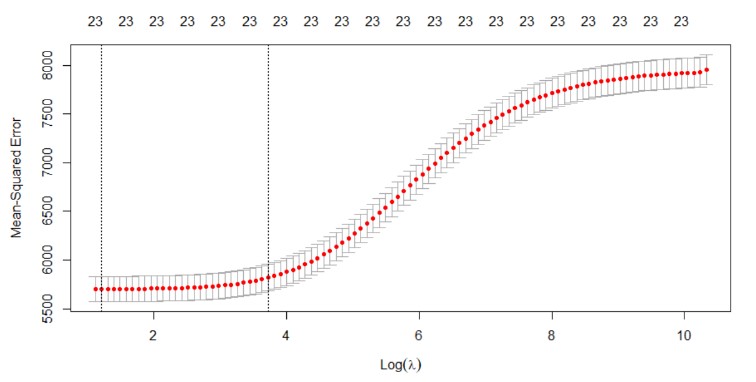
***1) Previously you created a model using the PISA dataset. Build a model again, this time…***

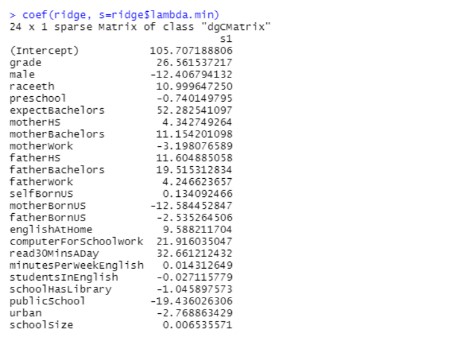
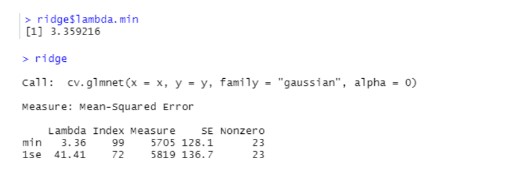
1. ***(10 points) Use Ridge regression and present your model along with appropriate outputs.*** 
   1. ***Discuss how this technique handles multicollinearity.***

Ridge regression can be used to lessen multicollinearity. Moreover, their estimates are typically stable, meaning that they are unaffected by minor changes in the data that the fitted regression is based on.

On the other hand, ordinary least squares estimates can be very unstable in several circumstances, such as in situations where the independent variables are very multicollinear.

The mean-squared error is displayed versus the lambda. The mean-squared error is decreased since the lambda is lowered to 3.359216. Moreover, the beta coefficients are presented.

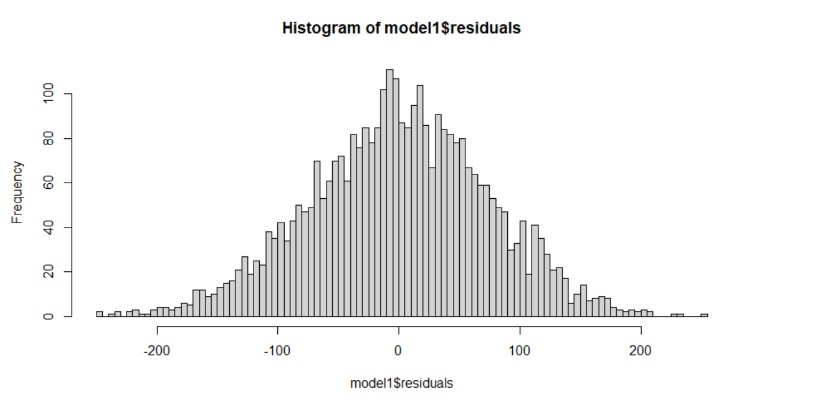
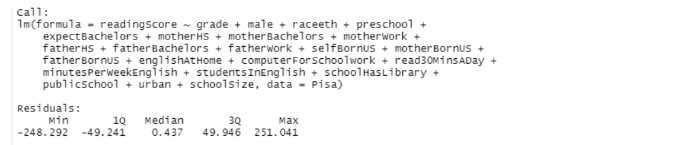




* 1. ***Evaluate the residual plots. Present the appropriate plots, describe them, and draw appropriate conclusions. Note: to look at the residual plots you can - after selecting variables with ridge regression - build a model using lm and plot the model.***

Following the selection of variables with ridge regression, we create a model with lm and visualize it as shown below.

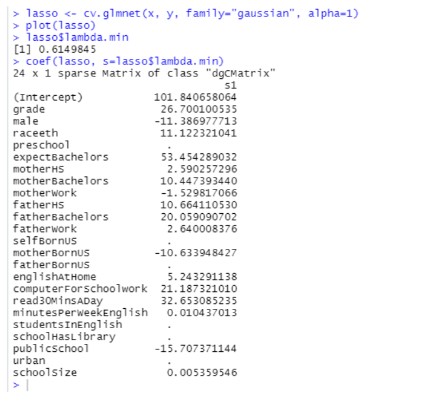
Looking at the residuals in the histogram, we can observe that the graph is normal distributed, reasonably symmetrical, and unbiased. There are no exceptionally extreme outliers.

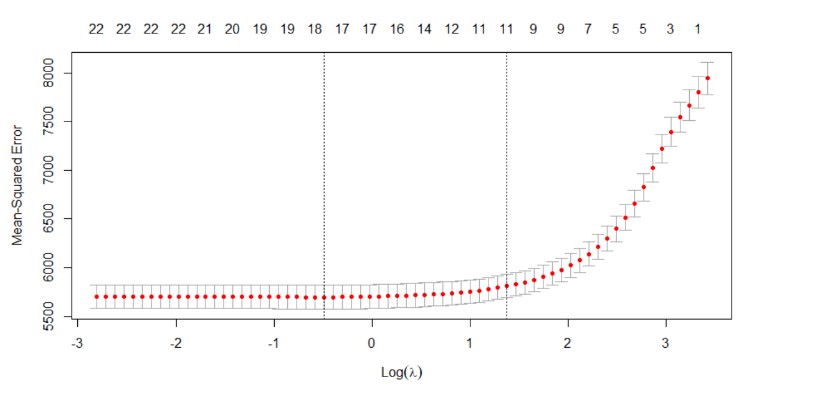


1. ***(10 points) Use LASSO regression and present your model along with appropriate outputs.*** 
   1. ***LASSO is a form of feature selection. Discuss how it reduced the feature space***

A continuous feature selection technique that can be used to choose features is LASSO regression. To carry out LASSO, separate structures for the dependent and independent variables must be established. The number of features kept in LASSO is determined by the penalty factor, and choosing the penalty factor through cross-validation ensures that the model will generalize well to new data sets.

The model appears to be missing preschool, selfBornUS, fatherBornUS, studentsInEnglish, schoolHasLibrary, and urban.





***c. (10 points) Are the two models the same? Explain.***

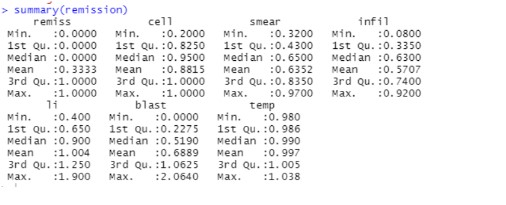
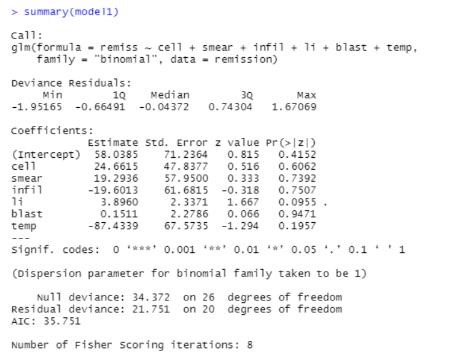
The two models are not interchangeable. The LASSO model seems to have been modified to exclude preschool, selfBornUS, fatherBornUS, studentsInEnglish, schoolHasLibrary, and urban.

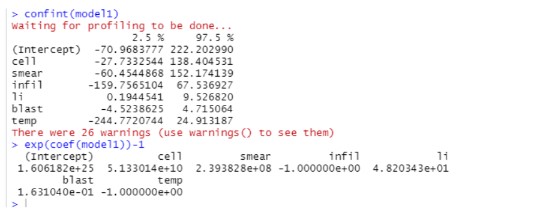
A scientifically sound method for reducing the amount of features in a model is LASSO. We may not need to employ feature selection at all and may instead rely on ridge regression to keep track of all the variables in the model if our primary goal is prediction and obtaining data on all features isn't too expensive. LASSO is an excellent option if we need to restrict the number of predictors for practical reasons. Yet all it does is provide us a useful selection of picky predictions, which aren't always the most crucial in the broad sense.

***2) REMISSION***

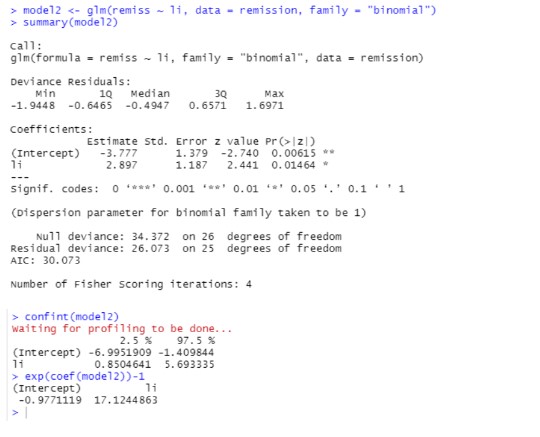
1. ***(10 points) Download "remission" and create a logistic model to predict remission.***

***i. Present your model.***





After dropping irrelative variables:



***b. (5 points) Notice that you are using the glm function.***

***i. Explain how this differs from lm.***

Generalized linear regression models are fitted with glm, whereas linear regression models are fitted with lm. Complex models like logistic regression and poison regression can also be fitted using it.

The dependant variable in logistic regression is the log probability of an event occurring. Rational regression can be used to evaluate a logistic model's recall, precision, specificity, and accuracy.

1. ***(5 points) Evaluate the model particularly the independent variables.***

Initial Model: model1 <- glm(remiss ~ cell + smear + infil + li + blast + temp, data = remission, family = "binomial")

After dropping irrelative variables:

model2 <- glm(remiss ~ li, data = remission, family = "binomial")

In the final model, we can de-log the coefficients, exp(coef(model2))-1 the probability of remiss changes.

