Ronaldlee Ejalu

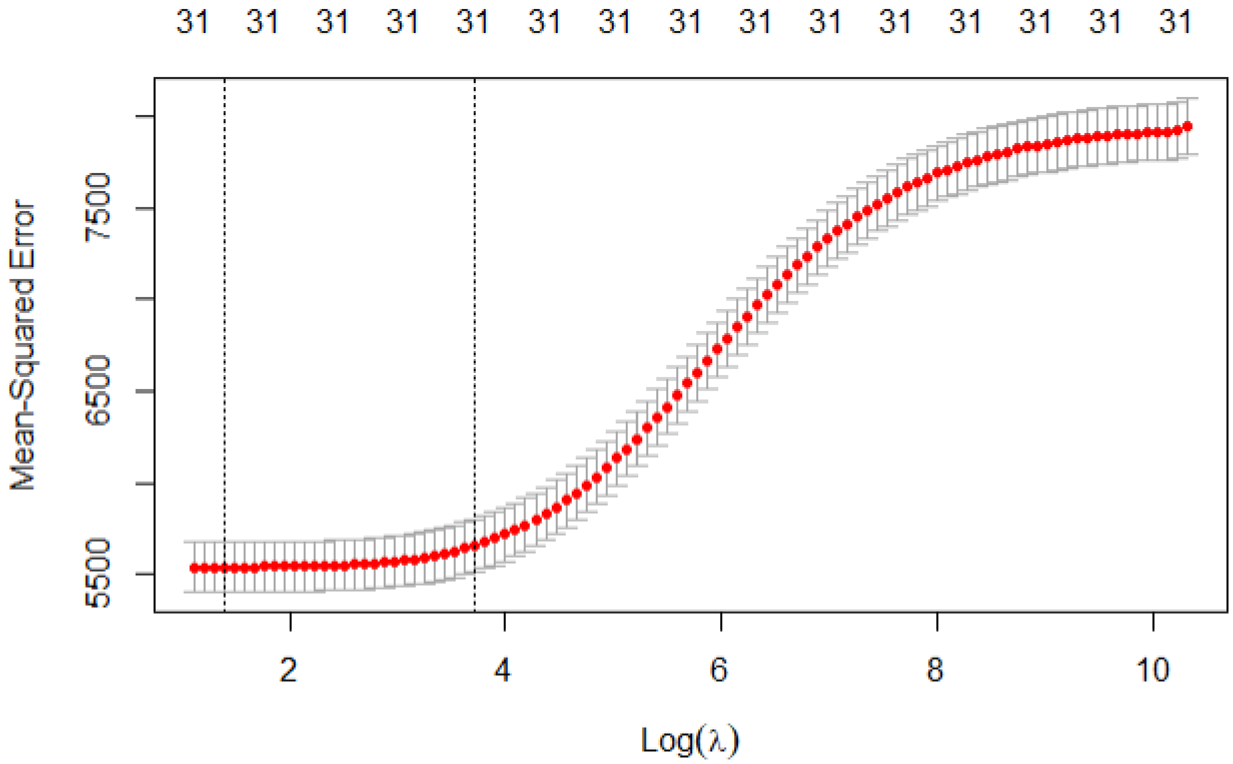
Assignment 4

Based on Modules 8 and 9

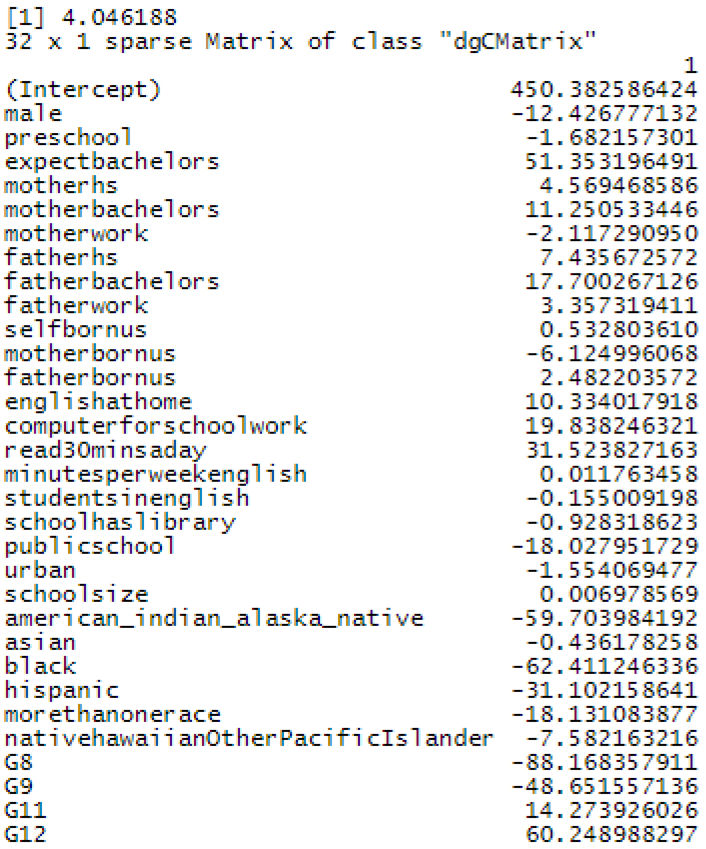
Student Id: 2020637

"I have completed this work independently. The solutions given are entirely my own work."

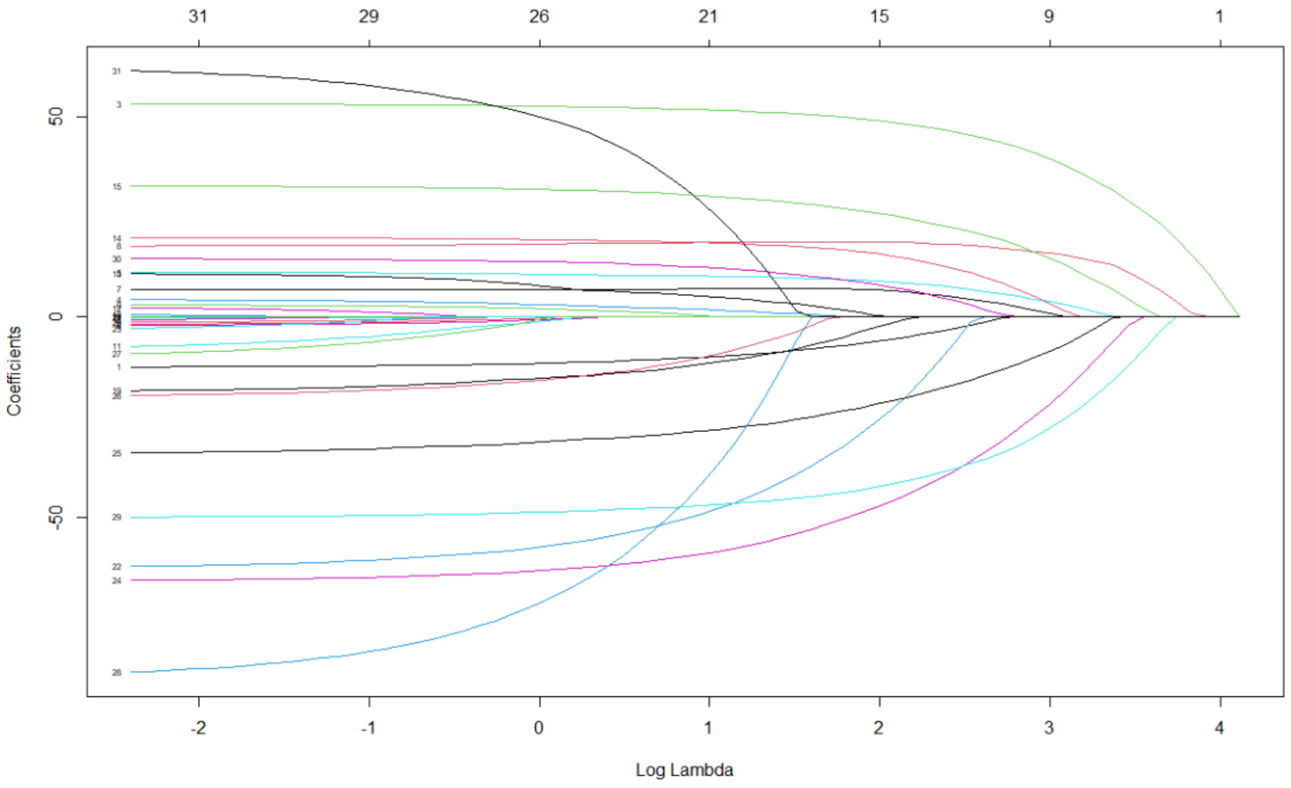
Use Ridge regression and present your model along with appropriate outputs.



On the x- axis you have lambda, then you have the Mean-Square Error on the y-axis



```
[1] 4.046188
32 x 1 sparse Matrix of class "dgCMatrix"
                                                  1
(Intercept)                            450.382586424
male                                   -12.426777132
preschool                               -1.682157301
expectbachelors                         51.353196491
motherhs                                 4.569468586
motherbachelors                         11.250533446
motherwork                              -2.117290950
fatherhs                                 7.435672572
fatherbachelors                         17.700267126
fatherwork                               3.357319411
selfbornus                               0.532803610
motherbornus                            -6.124996068
fatherbornus                             2.482203572
englishathome                           10.334017918
computerforschoolwork                   19.838246321
read30minsaday                          31.523827163
minutesperweekenglish                    0.011763458
studentsinenglish                       -0.155009198
schoolhaslibrary                        -0.928318623
publicschool                           -18.027951729
urban                                   -1.554069477
schoolsize                               0.006978569
american_indian_alaska_native          -59.703984192
asian                                   -0.436178258
black                                  -62.411246336
hispanic                               -31.102158641
morethanonerace                        -18.131083877
nativehawaiianOtherPacificIslander      -7.582163216
G8                                     -88.168357911
G9                                     -48.651557136
G11                                     14.273926026
G12                                     60.248988297
```



The above graph shows a ridge trace of coefficients.

In OLS, the parameter estimates depend on the following equation:

$$\hat{\beta} = (x'x)^{-1}xy$$

$x'x$ represents a correlation matrix of all predictors; x represents a matrix of dimensions nxp, where n is the number of observations and p is the number of predictors in the regression model; y represents the a vector of outcomes that is length n, and x' represents the transpose of x.

Ridge regression allows the retention of all explanatory

variables of interest, even if they are highly collinear. Ridge regression provides information regarding which coefficients are the most sensitive to multicollinearity. Ridge regression modifies the correlation matrix of all predictors such that its determinant does not equal to zero. This ensures that the (X`X) – 1 is calculable. Modifying the matrix in this way effectively eliminates collinearity, leading to more precise, and therefore more interpretable, parameter estimates, however, there is a trade off between variance and bias. This results into a cost of decrease in variance and an increase in bias. Nonetheless, the bias introduced by ridge regression is almost always toward the null. This makes ridge regression to be considered a shrinkage method since it typically shrinks the beta coefficients towards zero.

The modification of the correlation matrix starts when either lambda or k is introduced in the model. The value of k determines hoe much the ridge parameters differ from the parameters obtained using the Ordinary Least Square (OLS) model, and it takes any values greater than or equal to 0. When K = 0, this is equivalent to using OLS. The parameter K is incorporated into the following equation:

$$\hat{\beta}_{r_i d}ge = (x'x \ + kl)^{-1}x^1y$$

The above equation should look familiar since it is equivalent to the OLS formula for estimating regression parameters except for the addition of kl to the matrix $x'x$. In this equation I represents the identity matrix and K is the ridge parameter so multiplying k by I and adding this product to $x'x$ is equivalent to adding the value of k to the diagonal elements of $x'x$.

When there is multicollinearity, the columns of a correlation matrix are not independent of one another. This is always problematic, because a matrix with non-independent columns has a determinant of 0. Therefore, the dependence between columns must be broken down so the inverse of $x'x$ can be calculated. Adding a positive value k to the diagonal elements of $x'x$ will break up any dependency between these columns. This will cause the estimated regression coefficients to shrink toward the null; the higher the value of k, the greater the shrinkage. The intercept is the only coefficient, which is not penalized in this way.

The ridge trace plot facilitates the tradition means of choosing k. Estimated coefficients and VIF are plotted against the range of specified values of k.

From this plot, you select of value of k such that:

Stabilizes the system such that it reflects an orthogonal system.

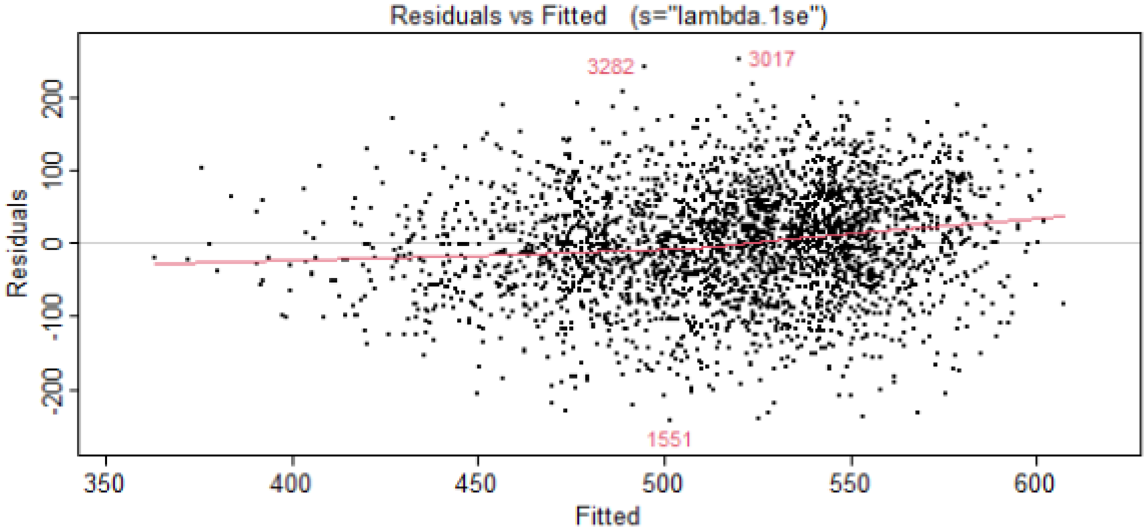
Leads to coefficients with reasonable values

Ensure that coefficients with improper signs at k=0 have switched to the proper sign

Ensures that the residual sum of squares is not inflated to an unreasonable value.

However, these criteria are very subjective. Therefore, it is best to use another method in addition to the ridge trace plot, which is the generalized cross validation like how we did. Cross validation simply entails looking at subsets of data and calculating the coefficients estiamtes for each subset of data, using the same value of k across subsets. This is then repeated multiple times with different values of k. The value of k, which minimizes the differences in coefficient estimates across theses data subsets is then selected. The model with the smallest prediction errors can be obtained by simply selecting the value of K that minimizes the generalized cross validation equation. The value of k that minimizes the equation can be computed using R like we did and it was 4.04

Evaluate the residual plots. Present the appropriate plots, describe them and draw appropriate conclusions.

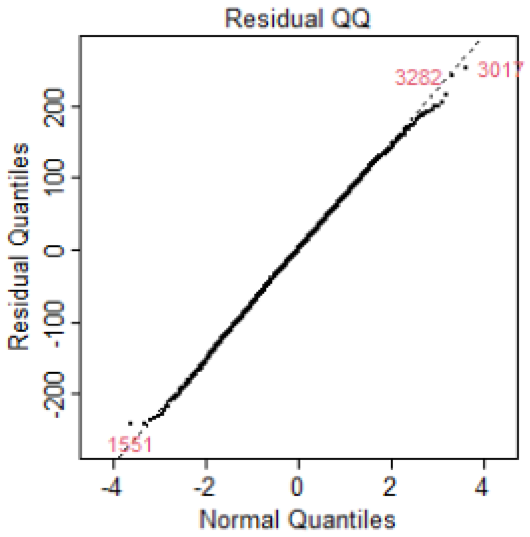


The black horizontal line which cuts through the points in the graph is the ordinary least square model, which minimizes the sum of the square of errors.

In settings with many explanatory variables in Ordinary least square when n is large because of the sampling variability, the estimates for Betas tend to be much larger in magnitude than the true values for the Betas so this means we have over fit the data. This tendency is exacerbated when we keep statisticall significant variables in the model. When shrinkage is introduced, it tends to move the estimated Betas closer towards zero.

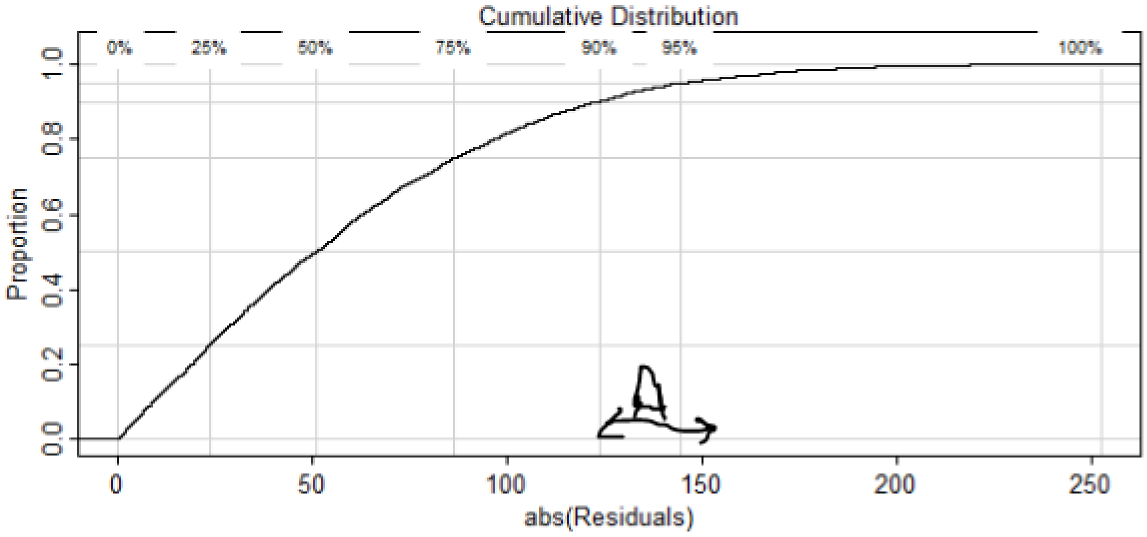
For the estimation model, colored red, we are predicting y given some estimation of Beta zero plus some estimatin of Beta one times X1. If we shrink or tried to push the Betas towards zero, the estimated Betas tend to be close to their true values. So, the graph shows how it is trying to reduce the impact of over fitting so that the Betas don't grow incorrectly by shrinking them towards zero. By shrinking the estimated coefficients, we reduce the variance at the cost of a negligible increase in bias, which improves the accuracy of prediction for future observations.

This graph also shows some outliers, which need to be looked at in our data set.

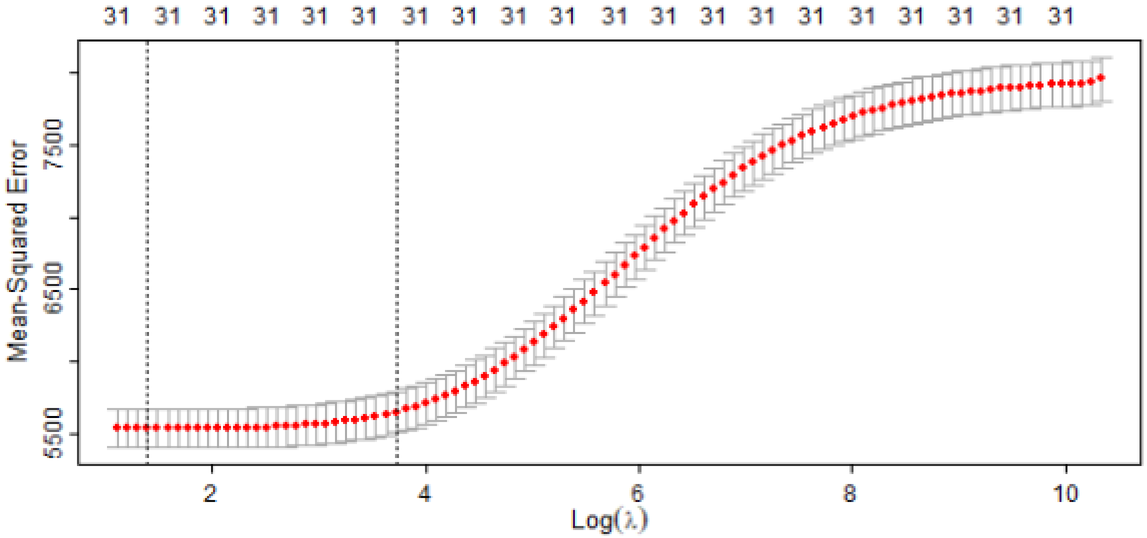


The Q-Q plot above shows how the residuals lay roughly on a nice straight diagonal line. It is plotting the residuals against their normal quantiles.

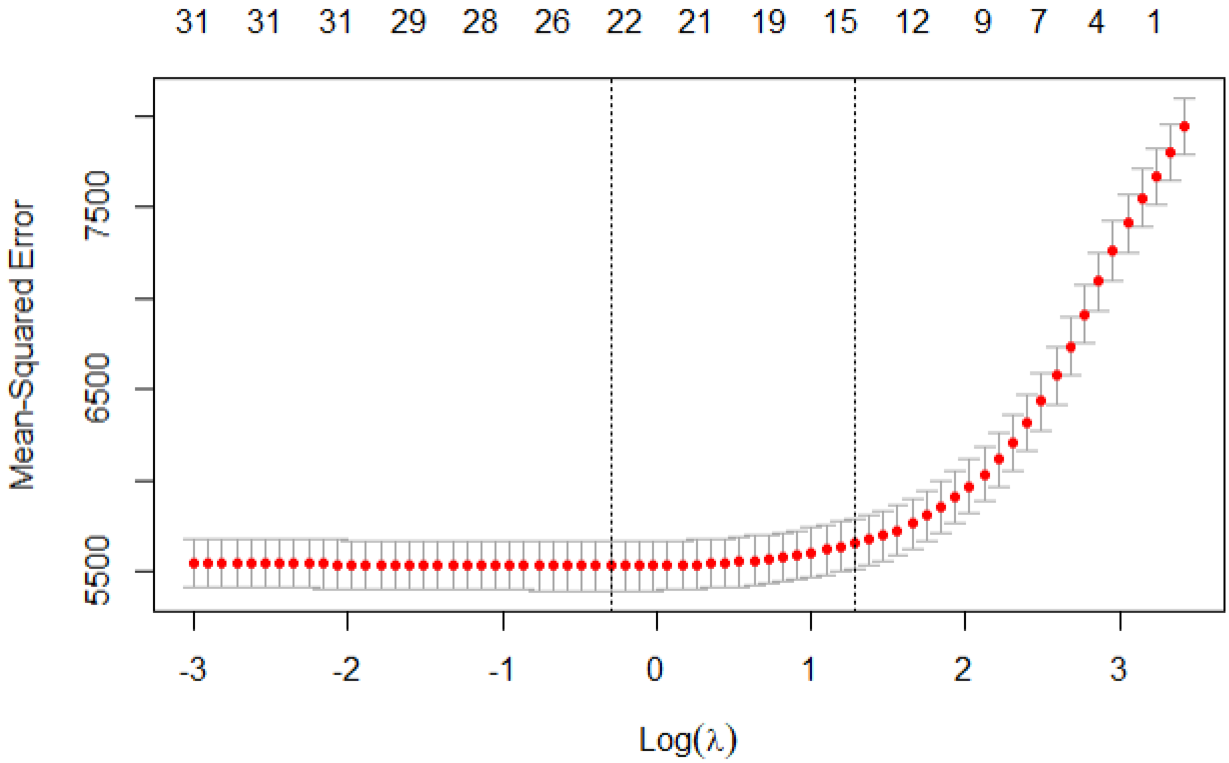
This is a normal plot which shows that the residuals are quite normal.

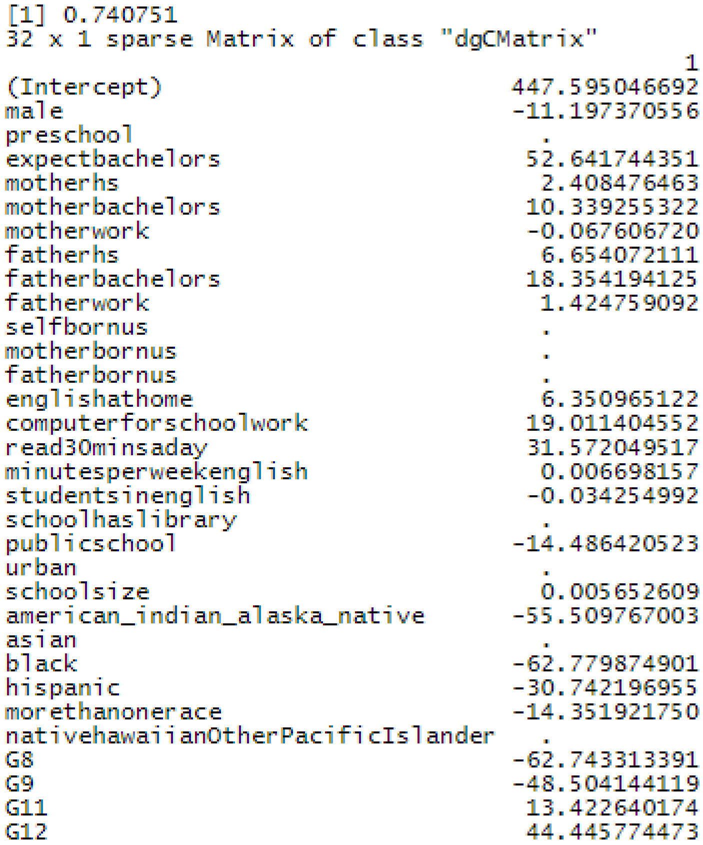


If we have a unbiased estimator as indicated by A, the above graph is what we would get, this is what we usually have in a least square regression model, The plot is trying to show whether or not we have correctly estimated the Beta, we will get some sort of distribution. We believe that Beta should, and it could be quite a large variance. With the help of regularization, we shrunk the variance at the cost of introducing a little bit of bias.



a. Use LASSO regression and present your model along with appropriate outputs.





```
[1] 0.740751
32 x 1 sparse Matrix of class "dgCMatrix"
                                             1
(Intercept)                      447.595046692
male                             -11.197370556
preschool                          .
expectbachelors                   52.641744351
motherhs                           2.408476463
motherbachelors                   10.339255322
motherwork                        -0.067606720
fatherhs                           6.654072111
fatherbachelors                   18.354194125
fatherwork                         1.424759092
selfbornus                         .
motherbornus                       .
fatherbornus                       .
englishathome                      6.350965122
computerforschoolwork             19.011404552
read30minsaday                    31.572049517
minutesperweekenglish              0.006698157
studentsinenglish                 -0.034254992
schoolhaslibrary                   .
publicschool                     -14.486420523
urban                              .
schoolsize                         0.005652609
american_indian_alaska_native    -55.509767003
asian                              .
black                            -62.779874901
hispanic                         -30.742196955
morethanonerace                  -14.351921750
nativehawaiianOtherPacificIslander  .
G8                               -62.743313391
G9                               -48.504144119
G11                               13.422640174
G12                               44.445774473
```

i. LASSO is a form of feature selection. Discuss how it reduced the feature space.

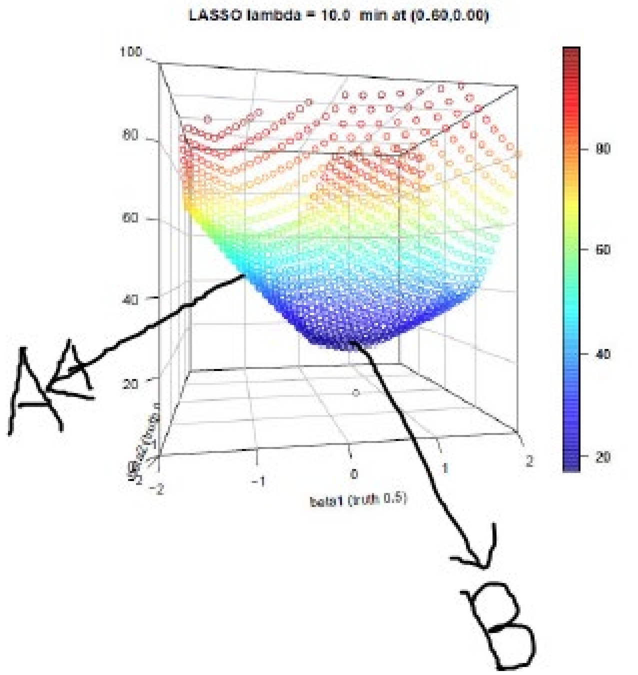
LASSO

$$\arg\min_{\beta}\left[\sum_{i=1}^{n}\left(Y_i-\beta_0-\sum_{j=1}^{p}\beta_j X_{ji}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|\right]$$

Recall that lasso regression, we are trying to minimize the sum of squares of the errors plus the sum of Betas.

λ is a tuning parameter, that is going to control how big an influence this shrinkage term is going to have on the model.

When we add the Ordinary Least Square term, the sum of squares of the errors to the manifold for Lasso, we get this stark arrow-shaped manifold, which is the manifold for this error term and they are added together we get the following shape



You will notice that the above shape has edges and points to them, curves, and that is what is causing Lasso regression to form as a method of feature selection.

Because when start getting those inflection points here **(marked A)**, that is when Beta equals zero so the arrows in the shape are projecting into the error space, and that is what is causing it to select Betas that are equal to zero since you get those little points **(marked B)** in the manifold.

b. Are the two models from a and b the same? Explain.

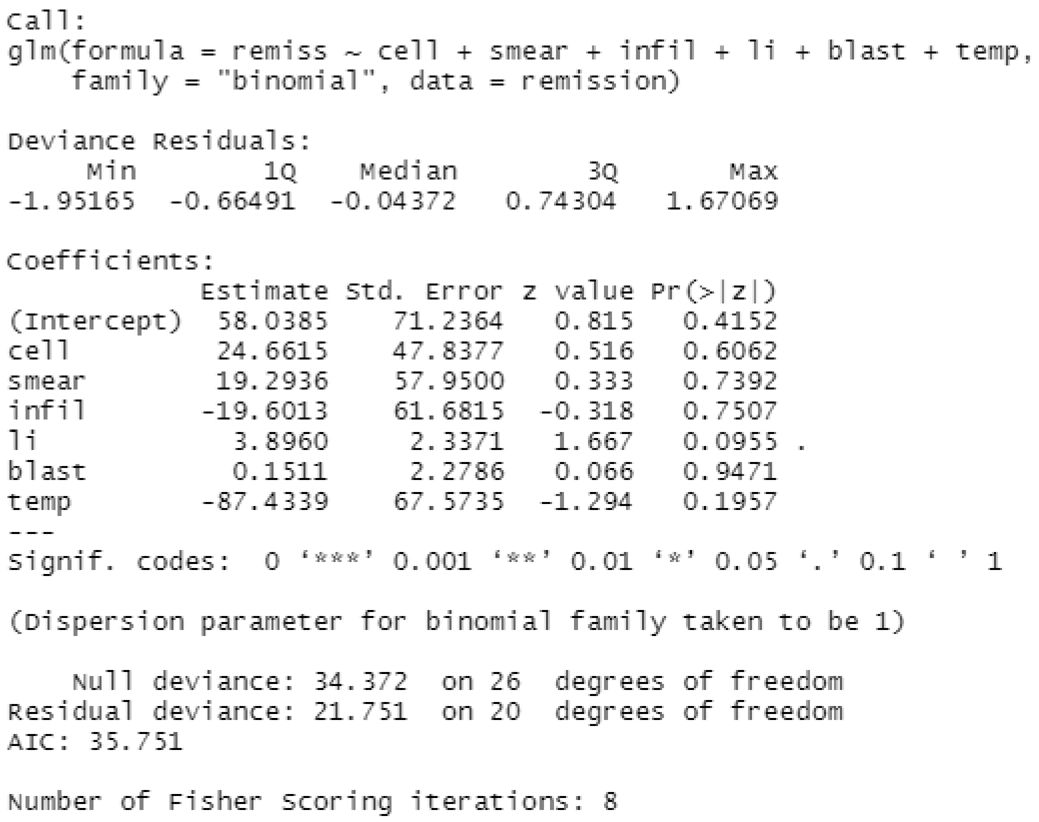
The two models are not the same because all the explanatory variables remain in the ridge regression model since ridge regression doesn't have the impact that some of the variables will be equal to zero where as in lasso regression, only a few of the estimated betas are practically different from zero

2) REMISSION (15 points)

a. Download "remission" and create a logistic model to predict remission.

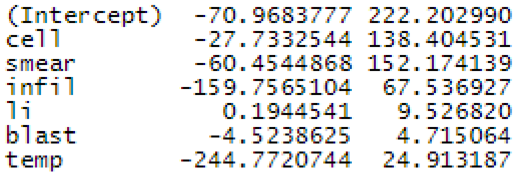
b. Perform logistic regression.

i. Submit your model.

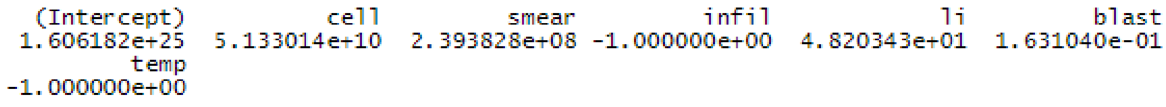


```
Call:
glm(formula = remiss ~ cell + smear + infil + li + blast + temp,
    family = "binomial", data = remission)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.95165  -0.66491  -0.04372   0.74304   1.67069

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  58.0385    71.2364   0.815   0.4152
cell         24.6615    47.8377   0.516   0.6062
smear        19.2936    57.9500   0.333   0.7392
infil       -19.6013    61.6815  -0.318   0.7507
li            3.8960     2.3371   1.667   0.0955 .
blast         0.1511     2.2786   0.066   0.9471
temp        -87.4339    67.5735  -1.294   0.1957
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 34.372  on 26  degrees of freedom
Residual deviance: 21.751  on 20  degrees of freedom
AIC: 35.751

Number of Fisher Scoring iterations: 8
```

95 % C.I



```
(Intercept)  -70.9683777 222.202990
cell         -27.7332544 138.404531
smear        -60.4544868 152.174139
infil       -159.7565104  67.536927
li             0.1944541   9.526820
blast         -4.5238625   4.715064
temp        -244.7720744  24.913187
```



```
 (Intercept)          cell         smear         infil            li         blast
1.606182e+25  5.133014e+10  2.393828e+08 -1.000000e+00  4.820343e+01  1.631040e-01
        temp
-1.000000e+00
```

c. Notice that you are using the *glm* function.

i. Explain how this differs from lm().

glm() fits generalized linear models where lm() fits linear models.

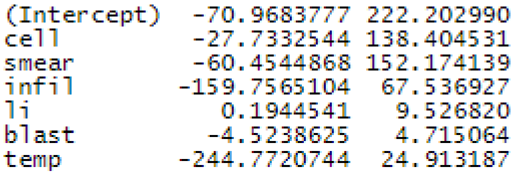
The glm() produces a model with z values for estimates, this z value is the Wald statistic that test the null hypothesis that all the Betas are equal to zero whereas lm() produces a model with p-values from t tests that also test the null hypothesis that all the Betas are equal to zero.

When we use the lm() in R we assume that our data follows a specific distribution: either normal or Gaussian distribution whereas, when we use a glm() in R, we specify a data distribution that our observations follow under for example, a family of binomial in glm().

d. Provide an analysis.

The parameter estimates indicate that cell, smear, li, and blast have positive effects on the probability of remission. If I look at my Wald Chi-Square test for temp, blast, li, infil, smear and cell, all the P values for z tests are greater than 0.05 if alpha is 0.05 meaning that we fail to reject the null hypothesis. However, if alpha is 0.1 then the Wald Chi test probability for li will be less than or equal to 0.1 meaning that we reject the null hypothesis that the beta associated with li is equal to zero then we accept the alternative that the Beta is not equal to zero and use an estimate of 3.8960. This means that temp, blast, infil, smear and cell are not statistically significant since they have high P-values as explained above. li, the only significant variable has a strong association with probability of remission. The odds of remission increase by exp(3.8960) – 1 = 48.205 times for any additional units of li.

When I look at the Confidence interval of the model this give me the 95 percent Confidence Interval for all of my variables:



```
(Intercept)  -70.9683777 222.202990
cell         -27.7332544 138.404531
smear        -60.4544868 152.174139
infil       -159.7565104  67.536927
li             0.1944541   9.526820
blast         -4.5238625   4.715064
temp        -244.7720744  24.913187
```

So, I believe cell is somewhere between -27.733 and 138.405.

i. Evaluate the independent variables?

The beta for cell of 24.6615 means that for every unit change in cell, the log odds of remission changes by 2466 percent.

The beta for smear of 19.294 means that for every unit change in cell, the log odds of remission changes by 1929.4 percent.

The beta for infill of -19.6013 means that for every unit change in infil, the log odds of remission reduces by 1960.13 percent.

The beta for li of 3.896 means that for every unit change in li, the log odds of remission changes by 389.6

The beta for blast of 0.151 means that for every unit change in blast, the log odds of remission changes by 15.1 percent.

The beta for temp of -87.434 means that for every unit change in temp, the log odds of remission reduces by -8743.4