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…

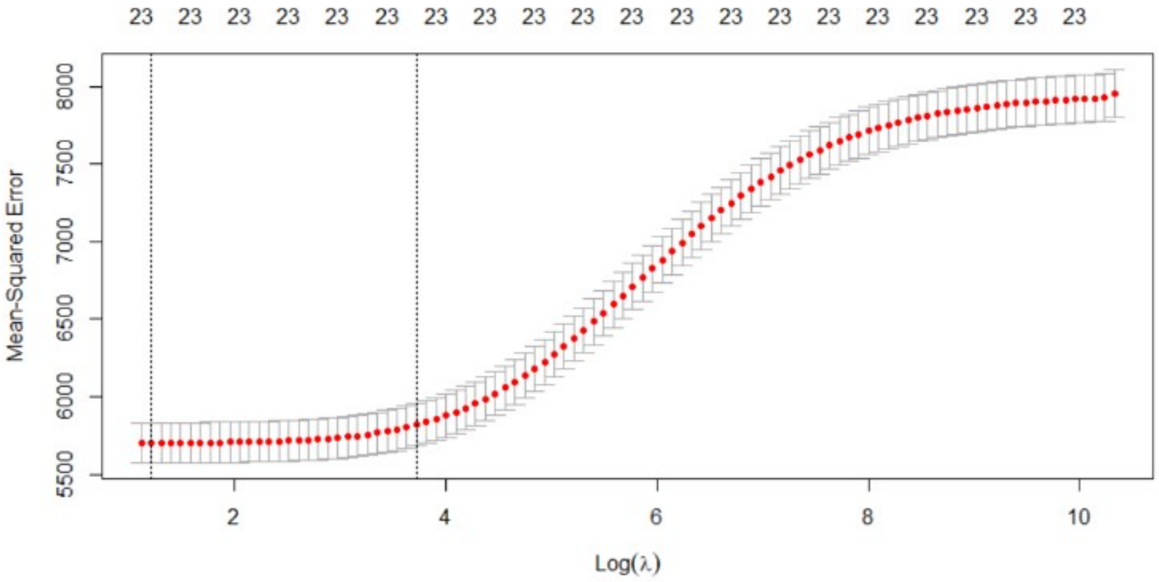
a. ***(10 points) Use Ridge regression and present your model along with appropriate outputs.***

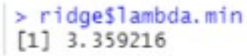
i. ***Discuss how this technique handles multicollinearity.***

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.

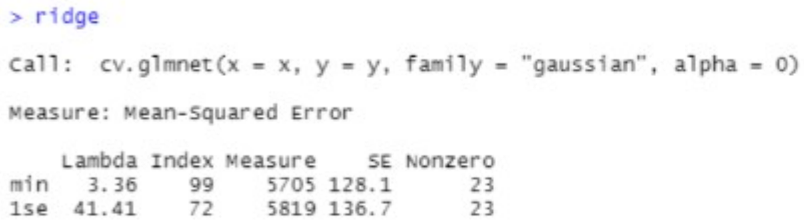
On the other hand, ordinary least squares estimates can be very unstable in several circumstances, such as in situations where the independent variables are very multicollinear.

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.

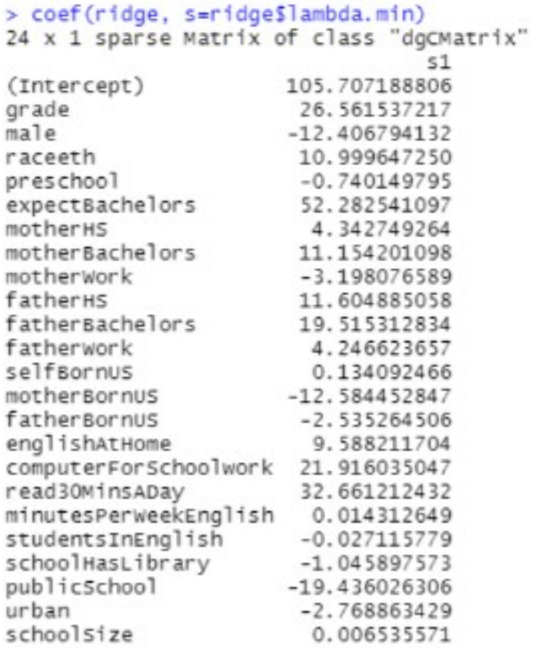




```
> ridge$lambda.min
[1] 3.359216
```



```
> ridge

Call:  cv.glmnet(x = x, y = y, family = "gaussian", alpha = 0)

Measure: Mean-Squared Error

    Lambda Index Measure    SE Nonzero
min   3.36    99    5705 128.1      23
1se  41.41    72    5819 136.7      23
```

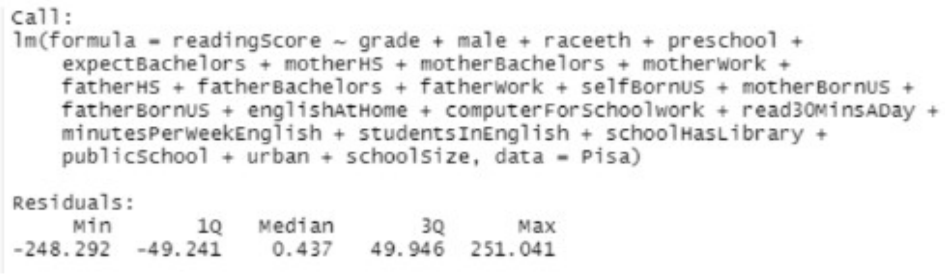


```
> coef(ridge, s=ridge$lambda.min)
24 x 1 sparse Matrix of class "dgCMatrix"
                               s1
(Intercept)          105.707188806
grade                 26.561537217
male                 -12.406794132
raceeth               10.999647250
preschool             -0.740149795
expectBachelors       52.282541097
motherHS               4.342749264
motherBachelors       11.154201098
motherWork            -3.198076589
fatherHS              11.604885058
fatherBachelors       19.515312834
fatherWork             4.246623657
selfBornUS             0.134092466
motherBornUS         -12.584452847
fatherBornUS          -2.535264506
englishAtHome          9.588211704
computerForSchoolwork 21.916035047
read30MinsADay        32.661212432
minutesPerWeekEnglish  0.014312649
studentsInEnglish     -0.027115779
schoolHasLibrary      -1.045897573
publicSchool         -19.436026306
urban                 -2.768863429
schoolSize             0.006535571
```

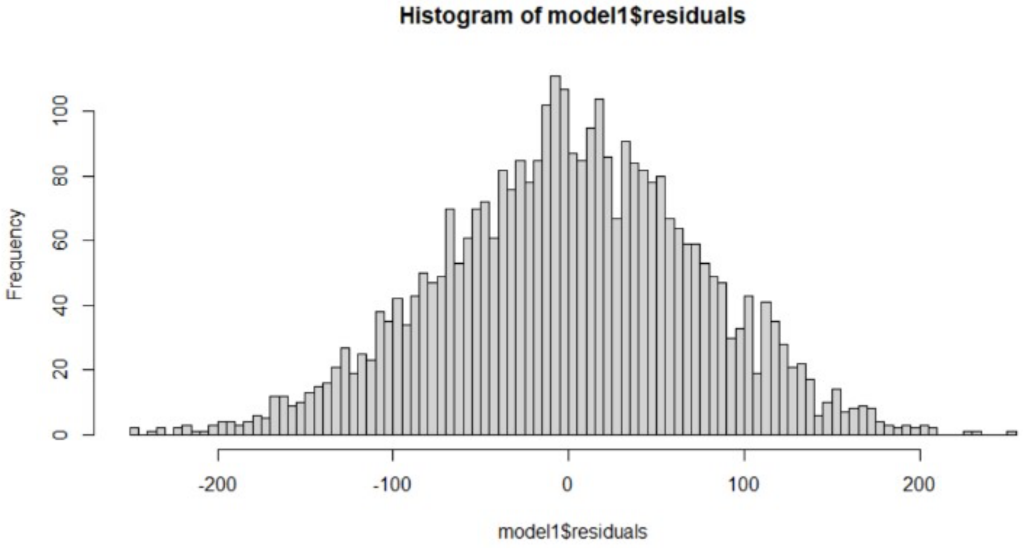
ii. ***Evaluate the residual plots. Present the appropriate plots, describe them, and draw appropriate conclusions. Note: to look at the residual plots you can - after selecting variables with ridge regression - build a model using lm and plot the model.***

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.



```
Call:
lm(formula = readingScore ~ grade + male + raceeth + preschool +
    expectBachelors + motherHS + motherBachelors + motherWork +
    fatherHS + fatherBachelors + fatherWork + selfBornUS + motherBornUS +
    fatherBornUS + englishAtHome + computerForSchoolwork + read30MinsADay +
    minutesPerWeekEnglish + studentsInEnglish + schoolHasLibrary +
    publicSchool + urban + schoolSize, data = Pisa)

Residuals:
     Min       1Q   Median       3Q      Max
-248.292  -49.241    0.437   49.946  251.041
```

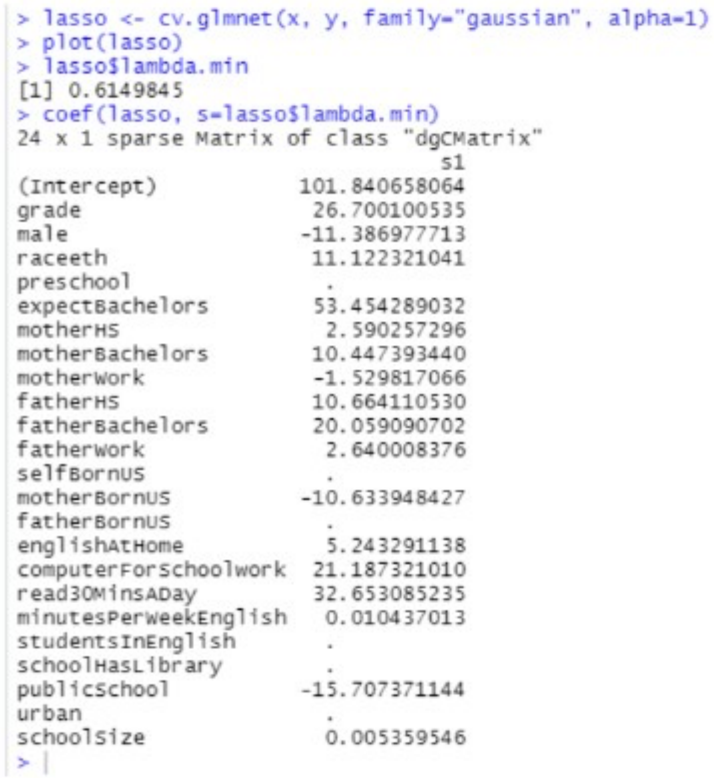


b. ***(10 points) Use LASSO regression and present your model along with appropriate outputs.***

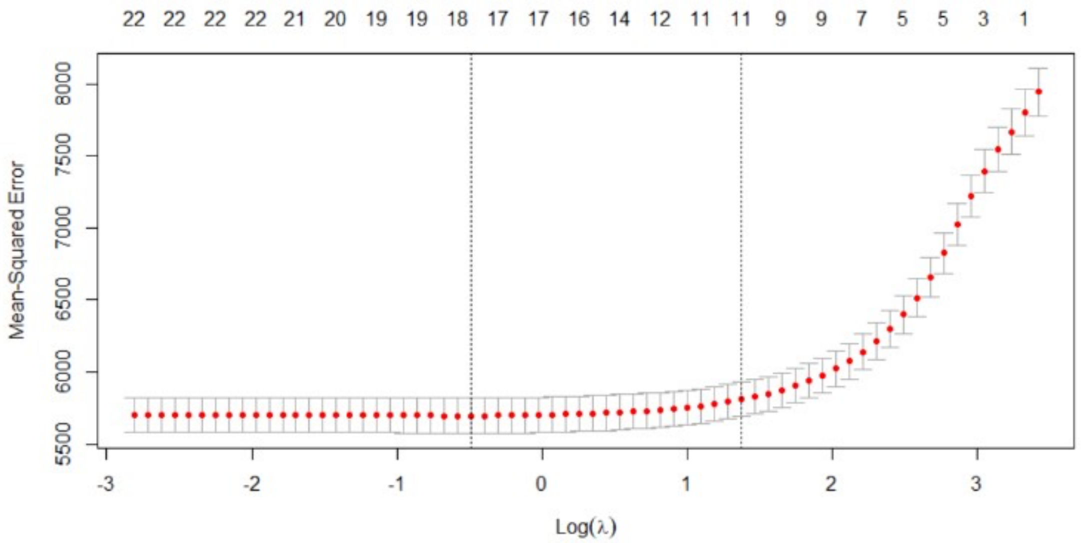
i. ***LASSO is a form of feature selection. Discuss how it reduced the feature space***

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.



```
> lasso <- cv.glmnet(x, y, family="gaussian", alpha=1)
> plot(lasso)
> lasso$lambda.min
[1] 0.6149845
> coef(lasso, s=lasso$lambda.min)
24 x 1 sparse Matrix of class "dgCMatrix"
                                  s1
(Intercept)            101.840658064
grade                   26.700100535
male                   -11.386977713
raceeth                 11.122321041
preschool                .          
expectBachelors         53.454289032
motherHS                 2.590257296
motherBachelors         10.447393440
motherWork              -1.529817066
fatherHS                10.664110530
fatherBachelors         20.059090702
fatherWork               2.640008376
selfBornUS               .          
motherBornUS           -10.633948427
fatherBornUS             .          
englishAtHome            5.243291138
computerForSchoolwork   21.187321010
read30MinsADay          32.653085235
minutesPerWeekEnglish    0.010437013
studentsInEnglish        .          
schoolHasLibrary         .          
publicSchool           -15.707371144
urban                    .          
schoolSize               0.005359546
> |
```



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

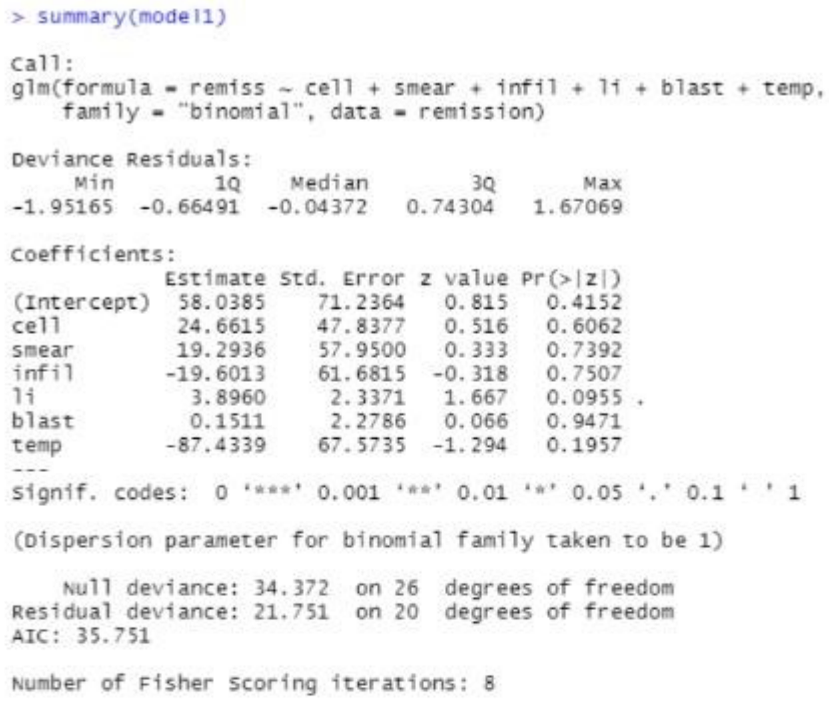
The two models are not interchangeable. The LASSO model seems to have been modified to exclude preschool, selfBornUS, fatherBornUS, studentsInEnglish, schoolHasLibrary, and urban.

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.

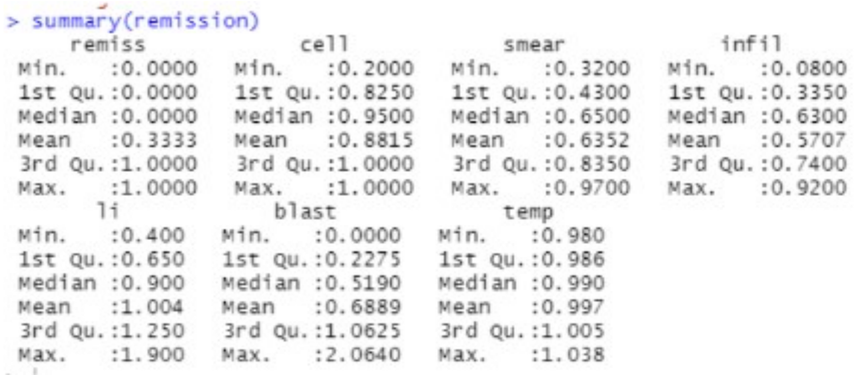
2) REMISSION

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

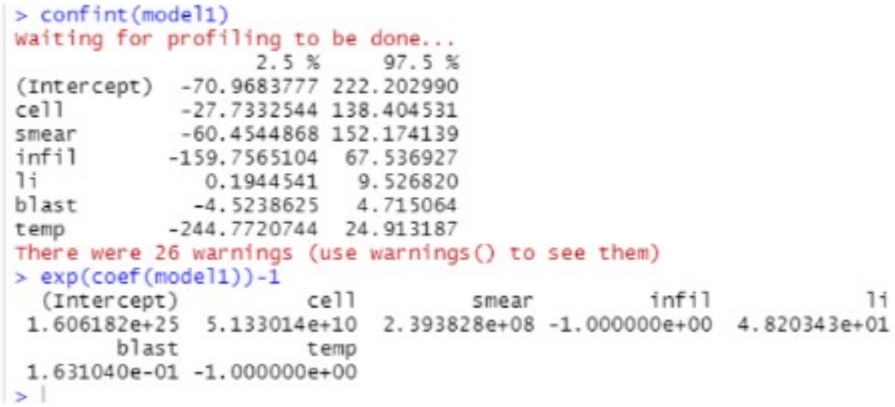
i. Present your model.



```
> summary(model1)

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
```

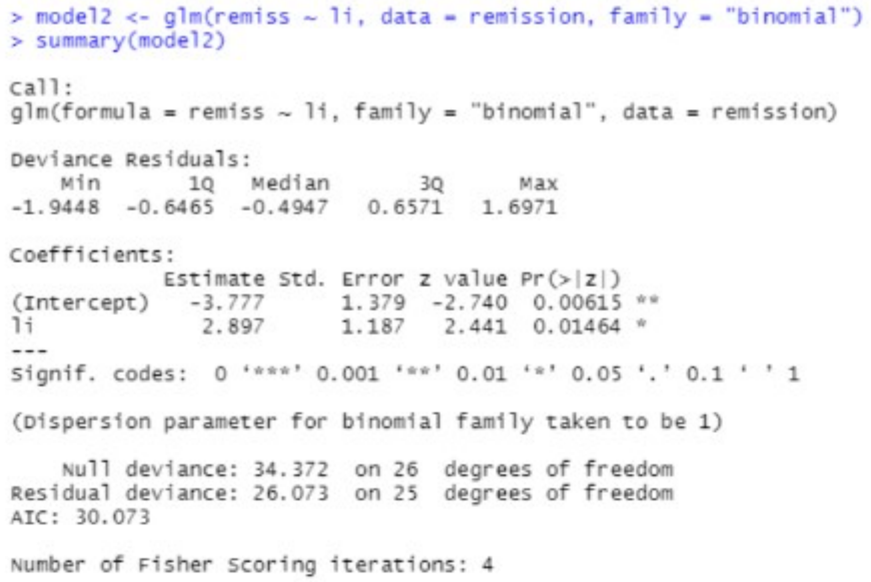


```
> summary(remission)
     remiss            cell            smear            infil
 Min.   :0.0000   Min.   :0.2000   Min.   :0.3200   Min.   :0.0800
 1st Qu.:0.0000   1st Qu.:0.8250   1st Qu.:0.4300   1st Qu.:0.3350
 Median :0.0000   Median :0.9500   Median :0.6500   Median :0.6300
 Mean   :0.3333   Mean   :0.8815   Mean   :0.6352   Mean   :0.5707
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.8350   3rd Qu.:0.7400
 Max.   :1.0000   Max.   :1.0000   Max.   :0.9700   Max.   :0.9200
       li            blast             temp
 Min.   :0.400   Min.   :0.0000   Min.   :0.980
 1st Qu.:0.650   1st Qu.:0.2275   1st Qu.:0.986
 Median :0.900   Median :0.5190   Median :0.990
 Mean   :1.004   Mean   :0.6889   Mean   :0.997
 3rd Qu.:1.250   3rd Qu.:1.0625   3rd Qu.:1.005
 Max.   :1.900   Max.   :2.0640   Max.   :1.038
```

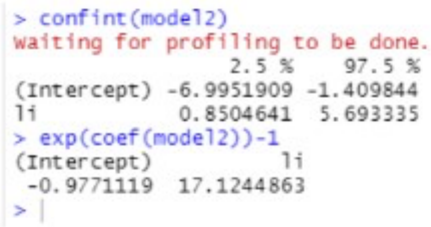


```
> confint(model1)
Waiting for profiling to be done...
                  2.5 %     97.5 %
(Intercept)  -70.9683777 222.202990
cell         -27.7332544 138.404531
smear        -60.4544868 152.174139
infil       -159.7565104  67.536927
li             0.1944541   9.526820
blast         -4.5238625   4.715064
temp        -244.7720744  24.913187
There were 26 warnings (use warnings() to see them)
> exp(coef(model1))-1
  (Intercept)          cell         smear         infil            li 
 1.606182e+25  5.133014e+10  2.393828e+08 -1.000000e+00  4.820343e+01 
        blast          temp 
 1.631040e-01 -1.000000e+00 
> |
```

After dropping irrelative variables:



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

Call:
glm(formula = remiss ~ li, family = "binomial", data = remission)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.9448  -0.6465  -0.4947   0.6571   1.6971  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)   
(Intercept)   -3.777      1.379  -2.740  0.00615 **
li             2.897      1.187   2.441  0.01464 * 
---
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: 26.073  on 25  degrees of freedom
AIC: 30.073

Number of Fisher Scoring iterations: 4
```



```
> confint(model2)
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) -6.9951909 -1.409844
li           0.8504641  5.693335
> exp(coef(model2))-1
(Intercept)          li 
 -0.9771119  17.1244863 
> |
```

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.

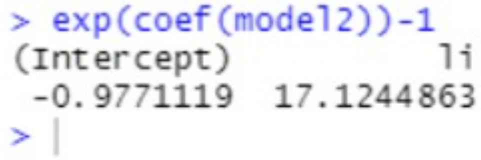
b. ***(5 points) Evaluate the model particularly the independent variables.***

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.



```
> exp(coef(model2))-1
(Intercept)          li
 -0.9771119  17.1244863
> |
```