honor statement: "I have completed this work independently.

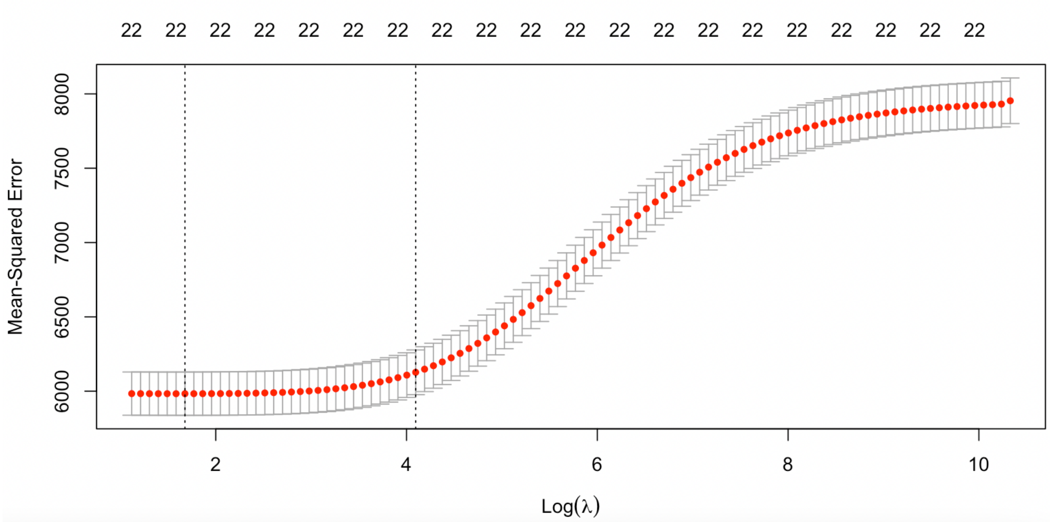
1) Previously you created a model using the PISA dataset. Build a model again, this time…

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

i. Discuss how this technique handles multicollinearity.

In the ordinary least square regression model , the estimates for the betas tend to be larger in magnitude than the true value which would cause overfitting because of sampling variability.

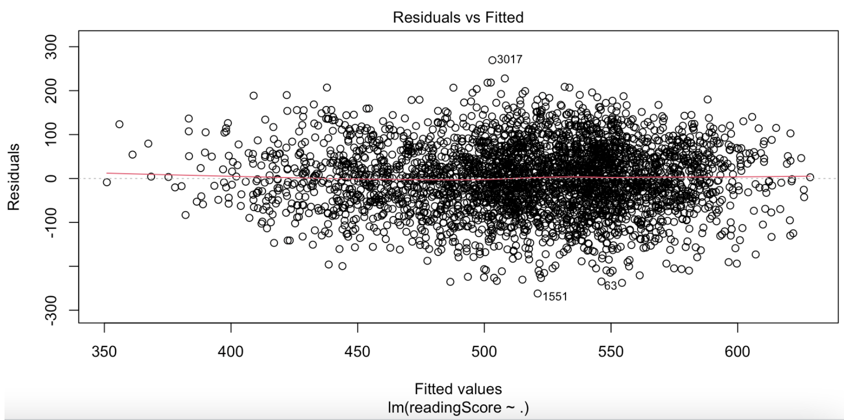
Ridge regression model tries to minimize the sum of the squares of the errors and adds on an additional term which is controlled by lambda. We iterate over the Betas and minimize the vector for the Beta which is to pull Beta back down to zero, we do not want Beta to grow too large, so ridge regression estimated tend to be stable because they are usually little affected by small changes in the data on which the fitted regression is based, so ridge regression is a way to combat multicollinearity in the data.

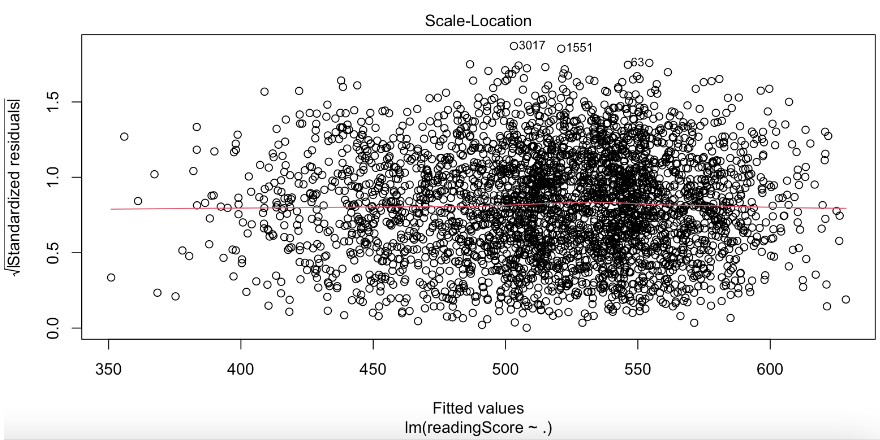


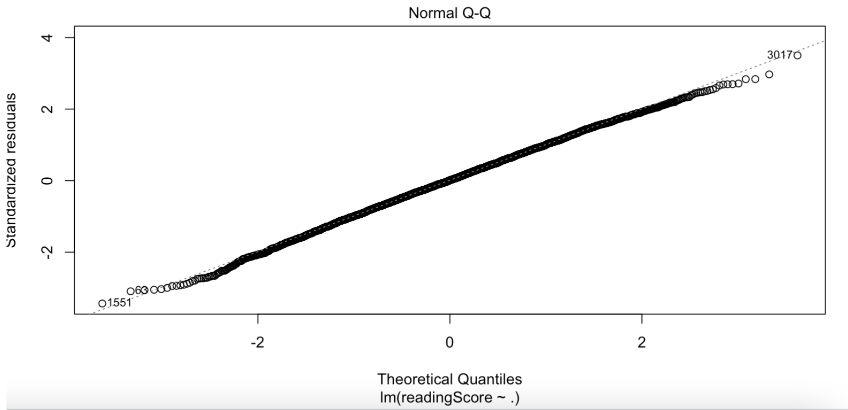

```
> Pisa2009_1 <-Pisa2009_1 %>% drop_na()
> x<-as.matrix(Pisa2009_1[,1:22])
> y<-as.double(Pisa2009_1[,23])
> set.seed(123)
> ridge <- cv.glmnet(x, y, family="gaussian", alpha=0)
```

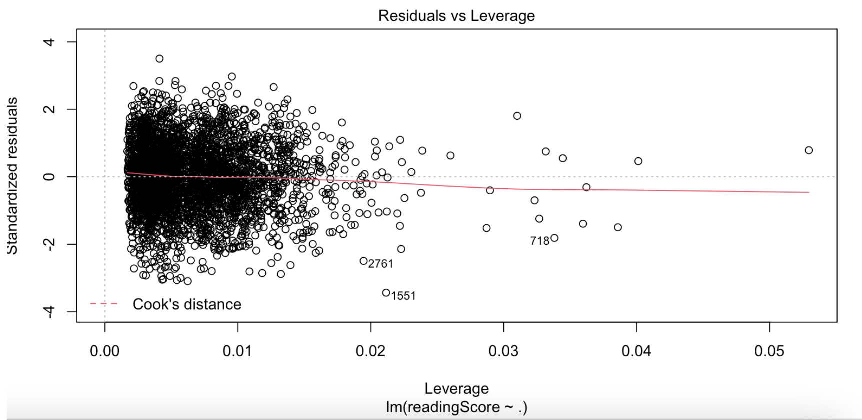
```
> plot(lasso)
> lridge$lambda.min
[1] 5.348822
> coef(lasso, s=ridge$lambda.min)
23 x 1 sparse Matrix of class "dgCMatrix"
                                s1
(Intercept)          153.118605567
grade                 26.378995390
male                 -12.084388012
preschool             -1.701141140
expectBachelors       51.420522656
motherHS               3.675078705
motherBachelors       12.069607441
motherWork            -3.195539742
fatherHS              12.224295298
fatherBachelors       22.802289022
fatherWork             8.420586488
selfBornUS            -0.238029976
motherBornUS           0.044573521
fatherBornUS           6.250526433
englishAtHome         11.532658337
computerForSchoolwork 25.979627894
read30MinsADay        31.415017807
minutesPerWeekEnglish  0.015011896
studentsInEnglish      0.013651886
schoolHasLibrary      -3.008442600
publicSchool         -24.388352147
urban                 -9.318370773
schoolSize             0.006092463
```

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.


Residuals vs Fitted
Residuals
Fitted values
lm(readingScore ~ .)
3017
1551
63
300
200
100
0
-100
-300
350
400
450
500
550
600


Scale-Location
√|Standardized residuals|
Fitted values
lm(readingScore ~ .)
3017
1551
63
0.0
0.5
1.0
1.5
350
400
450
500
550
600



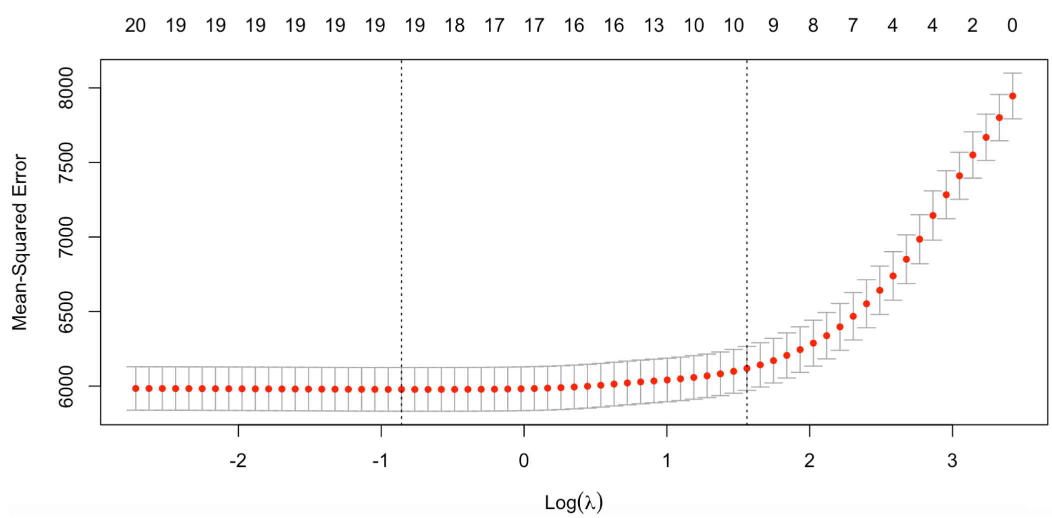


b. (10 points) Use LASSO regression and present your model along with appropriate outputs.

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

When running the LASSO regression model, we are trying to regularize or shrink the beta to zero, because of the LASSO equation, many of the Beta will actually be equal to zero, so it's a form of feature selection. If X value is associated with Beta and it can be removed from the model. It is continuous because it changes the lambda a little bit at a time in LASSO Regression model.

LASSO shrinkage causes the estimates of non - zero coefficients to be biased toward zero. LASSO regression could identify the set of non- zero coefficients and then fit an unrestricted linear model to the selected set of features.




```
 Pisa2009_1<-Pisa2009[,-c(1,4)]
> Pisa2009_1 <-Pisa2009_1 %>% drop_na()
> x<-as.matrix(Pisa2009_1[,1:22])
> y<-as.double(Pisa2009_1[,23])
> set.seed(123)
> lasso <- cv.glmnet(x, y, family="gaussian", alpha=1)
> plot(lasso  )
> lasso $ lambda.min
[1] 0.423885
> coef(lasso  , s=lasso $ lambda.min)
23 x 1 sparse Matrix of class "dgCMatrix"
                            s1
(Intercept)          143.303778659
grade                27.070798340
male                -11.539260227
preschool            -0.796973678
expectBachelors       53.305947532
motherHS              2.303585364
motherBachelors       11.424579415
motherWork           -2.270384941
fatherHS             11.809511561
fatherBachelors       23.574176177
fatherWork            7.476046816
selfBornUS             .
motherBornUS           .
fatherBornUS          5.895578819
englishAtHome         11.411897733
computerForSchoolwork  25.814744819
read30MinsADay        32.326905190
```

```
minutesPerWeekEnglish   0.012820450
studentsInEnglish       .
schoolHasLibrary       -0.639076633
publicSchool         -23.163597239
urban                 -8.724870570
schoolSize             0.005628654
```

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

They are not the same,ridge regression requires a separate strategy for finding a parsimonious model, because all explanatory variables remain in the model, however, LASSO yields sparse models that involve only a subset of the variables which are generally much easier to interpret.

2) REMISSION

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

i. Present your model.

1. I would make a logistic model first and check the t-value of each variable.
2. After checking the t-value in each variable, I would say the t-values are too bad to reject the null hypothesis.So I would use backward elimination to decide which variables should be included in my model.
3. The I would include the final model temp/cell/li in my final model.

```
> remission$remiss<- factor(remission$remiss)
> model<- glm(remiss~cell+infil +li+blast +temp, data = remission, family = "binomial")
> summary(model)

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

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.88165  -0.66603  -0.07206   0.78546   1.71792
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  70.09136   63.80360   1.099   0.2720
cell          9.22784    8.80720   1.048   0.2947
infil         0.95518    3.78107   0.253   0.8006
li            3.93020    2.26615   1.734   0.0829 .
blast        -0.04828    2.20111  -0.022   0.9825
temp        -84.82414   66.97814  -1.266   0.2054
---
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: 21.869  on 21  degrees of freedom
AIC: 33.869

Number of Fisher Scoring iterations: 7

> step<- stepAIC(model, direction="backward")
Start:  AIC=33.87
remiss ~ cell + infil + li + blast + temp

        Df Deviance    AIC
- blast  1   21.869 31.869
- infil  1   21.933 31.933
- cell   1   23.404 33.404
<none>       21.869 33.869
- temp   1   23.901 33.901
- li     1   26.878 36.878

Step:  AIC=31.87
remiss ~ cell + infil + li + temp

        Df Deviance    AIC
- infil  1   21.953 29.953
- cell   1   23.776 31.776
<none>       21.869 31.869
- temp   1   24.302 32.302
- li     1   30.490 38.490

Step:  AIC=29.95
remiss ~ cell + li + temp
```

```
       Df Deviance    AIC
<none>      21.953 29.953
- temp  1   24.341 30.341
- cell  1   24.648 30.648
- li    1   30.829 36.829
)
```

```
> model_1<- glm(remiss~cell +li +temp, data = remission, family = "binomial")
> summary(model_1)

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

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.02043  -0.66313  -0.08323   0.81282   1.65887

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   67.634     56.888   1.189   0.2345
cell           9.652      7.751   1.245   0.2130
li             3.867      1.778   2.175   0.0297 *
temp         -82.074     61.712  -1.330   0.1835
---
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: 21.953  on 23  degrees of freedom
AIC: 29.953

Number of Fisher Scoring iterations: 7
```

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

i. Explain how this differs from lm.

lm is used to fit linear models, including multivariate ones. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance

glm is used to fit generalized linear models, specified by giving a symbolic description of the linear predictor and a description of the error distribution

c. (5 points) Evaluate the model particularly the independent variables

Every unit change in cell, the log odd of remission changed by 965%
Every unit change in il, the log odd of remission changed by 386%
Every unit change in temp, the log odd of remission changed by -8207%

```
> model_1<- glm(remiss~cell +li +temp, data = remission, family = "binomial")
> summary(model_1)

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

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-2.02043  -0.66313  -0.08323   0.81282   1.65887

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   67.634     56.888   1.189   0.2345
cell           9.652      7.751   1.245   0.2130
li             3.867      1.778   2.175   0.0297 *
temp         -82.074     61.712  -1.330   0.1835
---
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: 21.953  on 23  degrees of freedom
AIC: 29.953


> exp(coef(model_1))-1
  (Intercept)         cell           li         temp
```

2.360653e+29 1.555423e+04 4.680357e+01 -1.000000e+00