

DATA311_Project

Parsa

2019-03-07

```
test <- read.csv("Admission_Predict_Ver1.1.csv")
summary(test)
```

```
##      Serial.No.      GRE.Score      TOEFL.Score      University.Rating
## Min.   : 1.0      Min.   :290.0      Min.   : 92.0      Min.   :1.000
## 1st Qu.:125.8      1st Qu.:308.0      1st Qu.:103.0      1st Qu.:2.000
## Median :250.5      Median :317.0      Median :107.0      Median :3.000
## Mean   :250.5      Mean   :316.5      Mean   :107.2      Mean   :3.114
## 3rd Qu.:375.2      3rd Qu.:325.0      3rd Qu.:112.0      3rd Qu.:4.000
## Max.   :500.0      Max.   :340.0      Max.   :120.0      Max.   :5.000
##      SOP      LOR      CGPA      Research
## Min.   :1.000      Min.   :1.000      Min.   :6.800      Min.   :0.00
## 1st Qu.:2.500      1st Qu.:3.000      1st Qu.:8.127      1st Qu.:0.00
## Median :3.500      Median :3.500      Median :8.560      Median :1.00
## Mean   :3.374      Mean   :3.484      Mean   :8.576      Mean   :0.56
## 3rd Qu.:4.000      3rd Qu.:4.000      3rd Qu.:9.040      3rd Qu.:1.00
## Max.   :5.000      Max.   :5.000      Max.   :9.920      Max.   :1.00
## Chance.of.Admit
## Min.   :0.3400
## 1st Qu.:0.6300
## Median :0.7200
## Mean   :0.7217
## 3rd Qu.:0.8200
## Max.   :0.9700
```

```
head(test)
```

```
##      Serial.No. GRE.Score TOEFL.Score University.Rating SOP LOR CGPA Research
## 1           1      337       118           4 4.5 4.5 9.65         1
## 2           2      324       107           4 4.0 4.5 8.87         1
## 3           3      316       104           3 3.0 3.5 8.00         1
## 4           4      322       110           3 3.5 2.5 8.67         1
## 5           5      314       103           2 2.0 3.0 8.21         0
## 6           6      330       115           5 4.5 3.0 9.34         1
##      Chance.of.Admit
## 1           0.92
## 2           0.76
## 3           0.72
## 4           0.80
## 5           0.65
## 6           0.90
```

```
attach(test)
```

```
linear.full <- lm(Chance.of.Admit ~., data=test)
linear.null <- lm(Chance.of.Admit ~ 1, data=test)
```

```
linear.rank.full <- lm(University.Rating ~., data=test)
linear.null.full <- lm(University.Rating ~ 1, data=test)
```

Linear Regression and some plots

Here's a linear model with a few plots.

```
linear <- lm(Chance.of.Admit ~., data=test)
summary(linear)
```

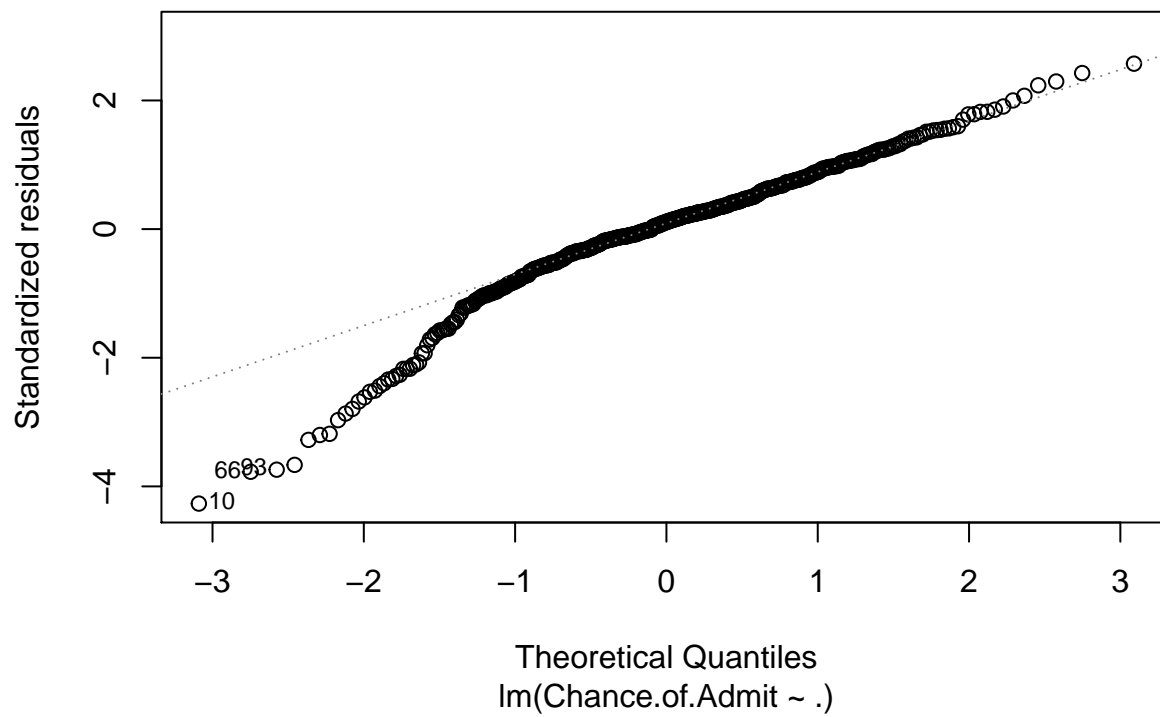
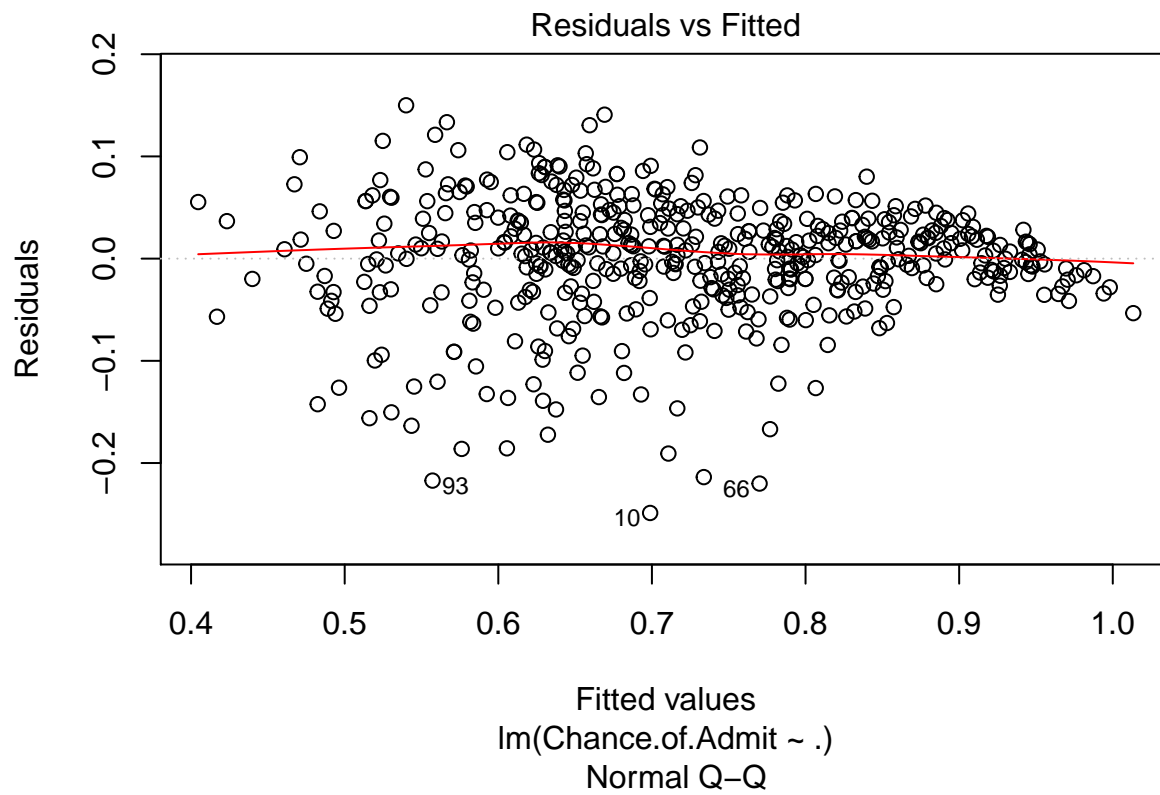
```
##
## Call:
## lm(formula = Chance.of.Admit ~ ., data = test)
##
## Residuals:
```

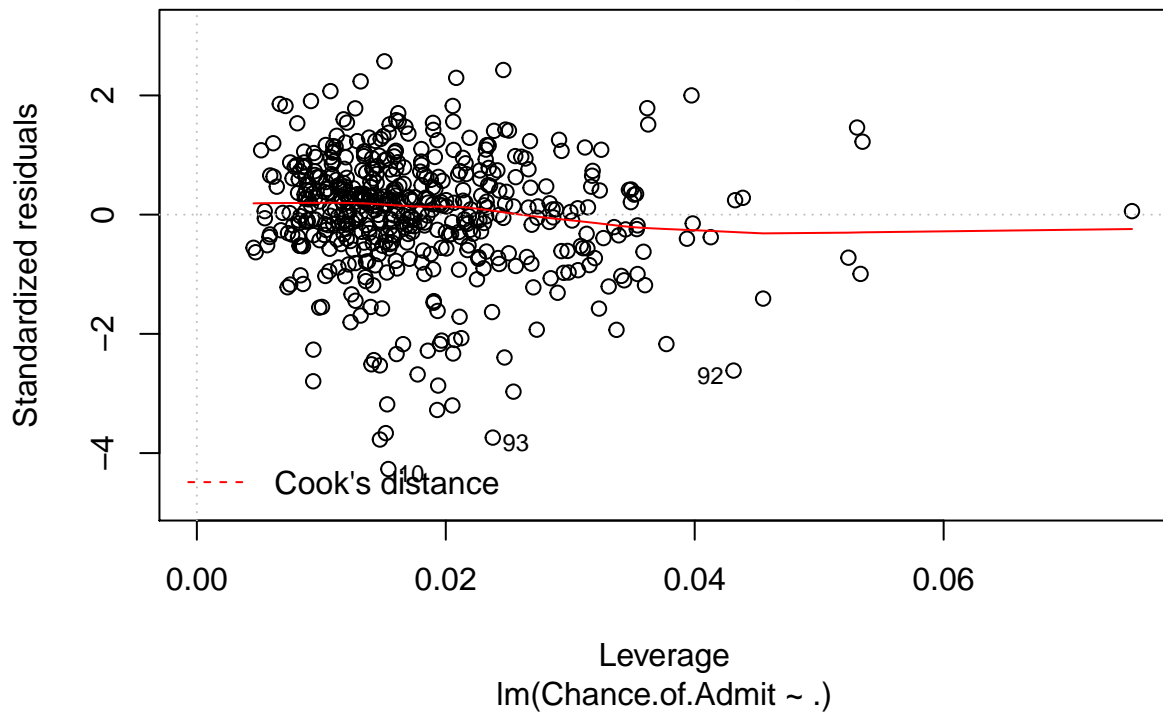
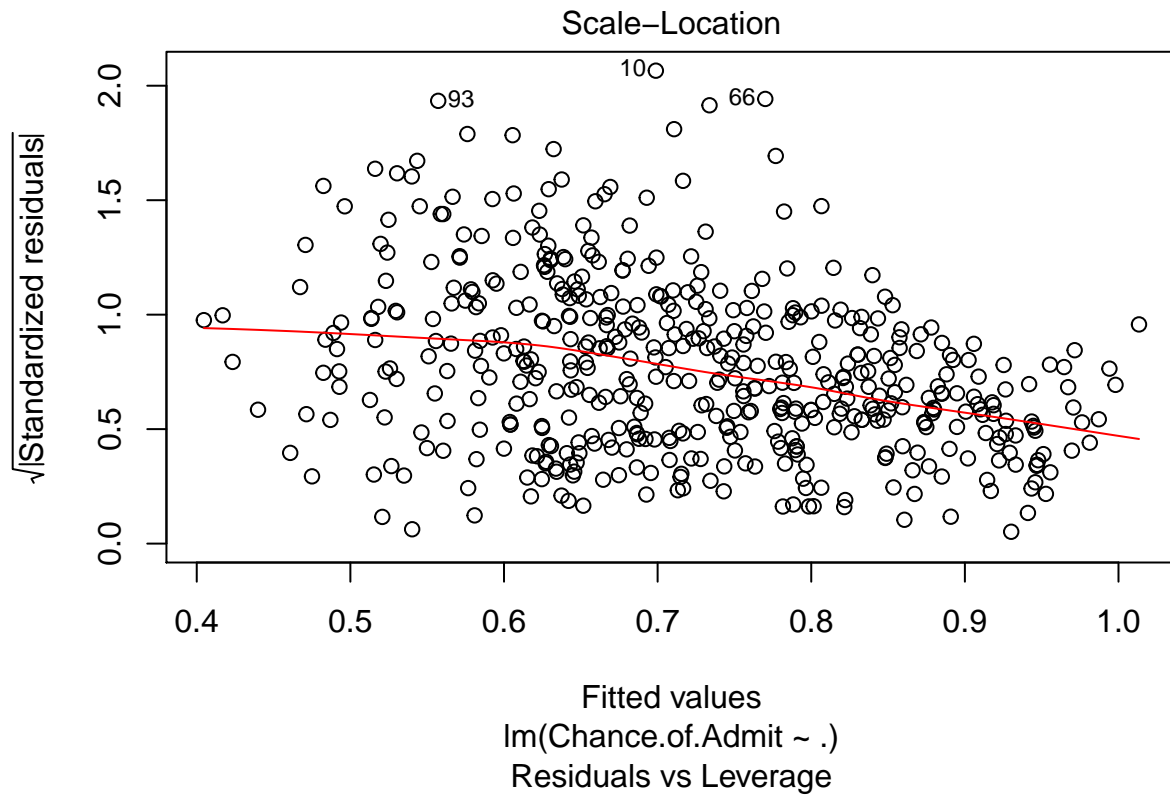
	Min	1Q	Median	3Q	Max
	-0.248847	-0.025984	0.006627	0.036671	0.150015

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.3379983	0.1030617	-12.982	< 2e-16 ***
Serial.No.	0.0000868	0.0000187	4.641	4.44e-06 ***
GRE.Score	0.0019217	0.0004923	3.903	0.000108 ***
TOEFL.Score	0.0031928	0.0008594	3.715	0.000227 ***
University.Rating	0.0053164	0.0037273	1.426	0.154405
SOP	0.0045661	0.0045161	1.011	0.312489
LOR	0.0149151	0.0040757	3.660	0.000280 ***
CGPA	0.1155561	0.0095282	12.128	< 2e-16 ***
Research	0.0225254	0.0064834	3.474	0.000557 ***

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05877 on 491 degrees of freedom
## Multiple R-squared:  0.8294, Adjusted R-squared:  0.8266
## F-statistic: 298.4 on 8 and 491 DF,  p-value: < 2.2e-16
plot(linear)
```



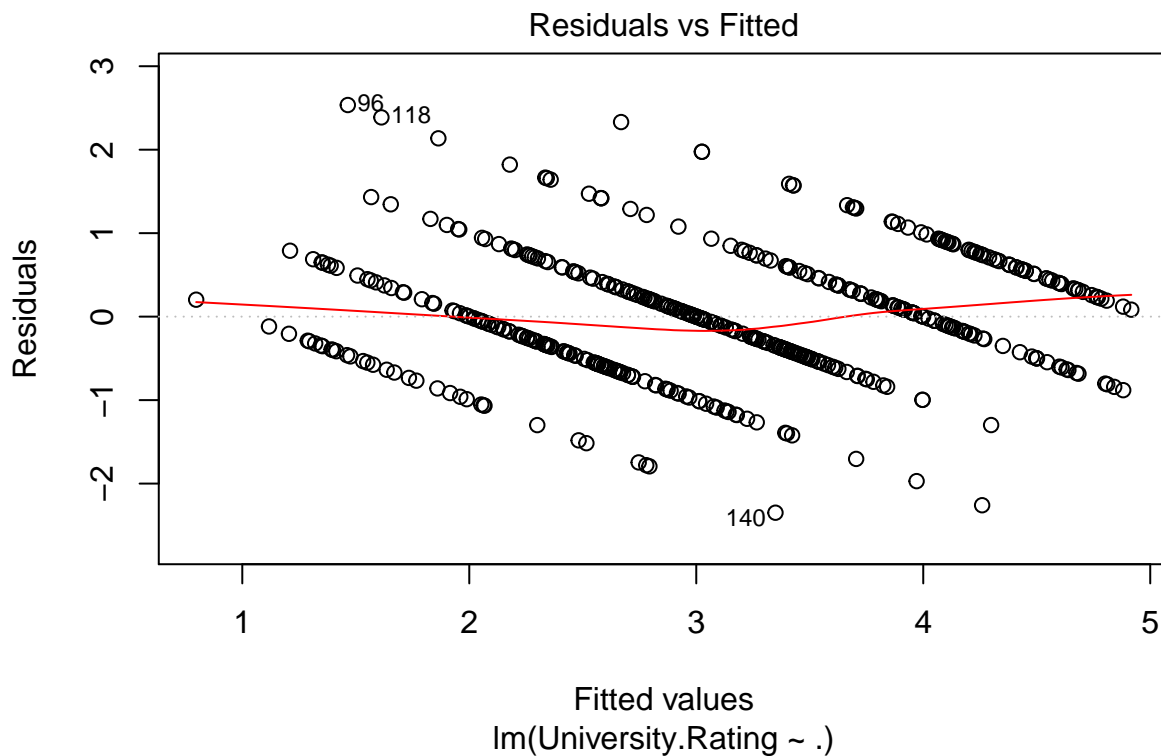


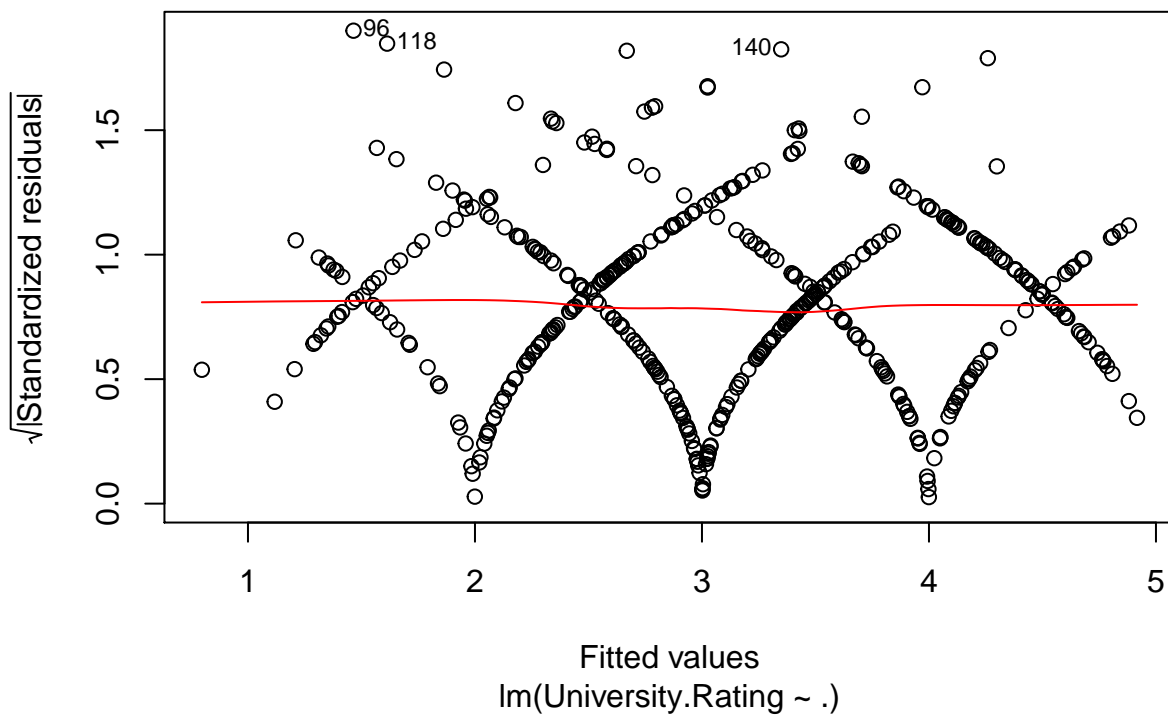
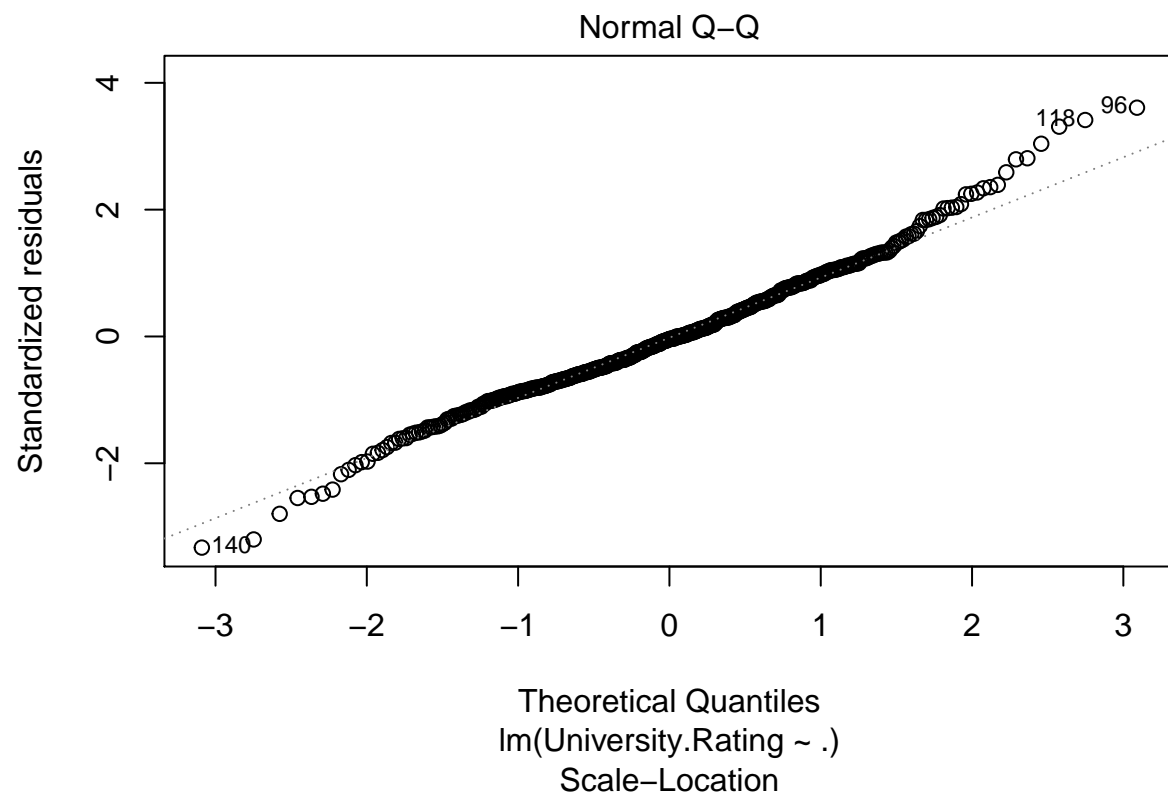
```
linear <- lm(University.Rating ~ ., data=test)
summary(linear)
```

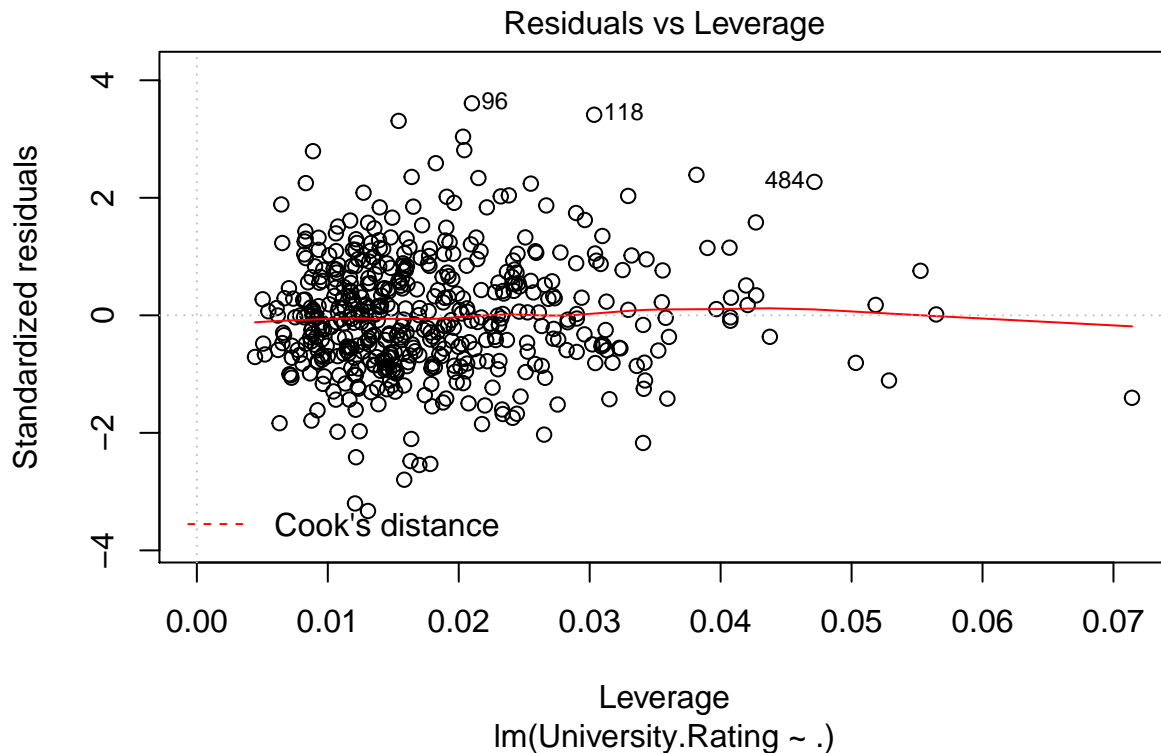
```
##
## Call:
## lm(formula = University.Rating ~ ., data = test)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34889 -0.46404 -0.02909  0.43638  2.53513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.3520556   1.4229030   -3.761 0.000189 ***
## Serial.No.      0.0001131   0.0002308    0.490 0.624275
## GRE.Score      0.0050723   0.0060361    0.840 0.401135
## TOEFL.Score    0.0184033   0.0104963    1.753 0.080172 .
## SOP           0.4420126   0.0508516    8.692 < 2e-16 ***
## LOR            0.1376178   0.0495241    2.779 0.005665 **
## CGPA           0.2666732   0.1306889    2.041 0.041833 *
## Research       0.0744728   0.0792227    0.940 0.347657
## Chance.of.Admit 0.7761573   0.5441596    1.426 0.154405
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7101 on 491 degrees of freedom
## Multiple R-squared:  0.6205, Adjusted R-squared:  0.6144
## F-statistic: 100.4 on 8 and 491 DF,  p-value: < 2.2e-16
```

```
plot(linear)
```







Variable Selection for Chance of Admission

By performing backwards selection, we will remove the least significant values until all values are significant.

```
linear <- lm(Chance.of.Admit ~ ., data = test)
summary(linear)
```

```
##
## Call:
## lm(formula = Chance.of.Admit ~ ., data = test)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-0.248847	-0.025984	0.006627	0.036671	0.150015

```
##
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-1.3379983	0.1030617	-12.982	< 2e-16 ***
Serial.No.	0.0000868	0.0000187	4.641	4.44e-06 ***
GRE.Score	0.0019217	0.0004923	3.903	0.000108 ***
TOEFL.Score	0.0031928	0.0008594	3.715	0.000227 ***
University.Rating	0.0053164	0.0037273	1.426	0.154405
SOP	0.0045661	0.0045161	1.011	0.312489
LOR	0.0149151	0.0040757	3.660	0.000280 ***
CGPA	0.1155561	0.0095282	12.128	< 2e-16 ***
Research	0.0225254	0.0064834	3.474	0.000557 ***

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.05877 on 491 degrees of freedom
## Multiple R-squared:  0.8294, Adjusted R-squared:  0.8266
## F-statistic: 298.4 on 8 and 491 DF,  p-value: < 2.2e-16
#Remove University Ranking because it has the highest non significant p value
linear <- lm(Chance.of.Admit~ Serial.No. + GRE.Score + TOEFL.Score + SOP +LOR + CGPA + Research , data = test)
summary(linear)

##
## Call:
## lm(formula = Chance.of.Admit ~ Serial.No. + GRE.Score + TOEFL.Score +
##     SOP + LOR + CGPA + Research, data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.249225 -0.026058  0.005588  0.037182  0.150359
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.372e+00  1.004e-01 -13.673  < 2e-16 ***
## Serial.No.   8.776e-05  1.871e-05   4.691 3.53e-06 ***
## GRE.Score    1.957e-03  4.922e-04   3.975 8.09e-05 ***
## TOEFL.Score  3.304e-03  8.568e-04   3.857 0.000130 ***
## SOP          6.945e-03  4.201e-03   1.653 0.098981 .
## LOR          1.571e-02  4.041e-03   3.888 0.000115 ***
## CGPA         1.175e-01  9.444e-03  12.437  < 2e-16 ***
## Research     2.302e-02  6.481e-03   3.551 0.000420 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05883 on 492 degrees of freedom
## Multiple R-squared:  0.8287, Adjusted R-squared:  0.8262
## F-statistic: 340 on 7 and 492 DF,  p-value: < 2.2e-16
#Remove SOP has the second highest non significant p value
linear <- lm(Chance.of.Admit~ Serial.No. + GRE.Score + TOEFL.Score +LOR + CGPA + Research , data = test)
#All variables are now significant
summary(linear)

##
## Call:
## lm(formula = Chance.of.Admit ~ Serial.No. + GRE.Score + TOEFL.Score +
##     LOR + CGPA + Research, data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.247948 -0.026442  0.005457  0.036306  0.152463
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.406e+00  9.844e-02 -14.280  < 2e-16 ***
## Serial.No.   8.348e-05  1.856e-05   4.498 8.58e-06 ***
## GRE.Score    1.941e-03  4.930e-04   3.937 9.42e-05 ***
## TOEFL.Score  3.478e-03  8.518e-04   4.083 5.18e-05 ***
## LOR          1.831e-02  3.729e-03   4.911 1.23e-06 ***
```



```
## CGPA          1.215e-01  9.132e-03  13.310  < 2e-16 ***
## Research      2.357e-02  6.484e-03   3.635  0.000307 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05894 on 493 degrees of freedom
## Multiple R-squared:  0.8277, Adjusted R-squared:  0.8256
## F-statistic: 394.8 on 6 and 493 DF,  p-value: < 2.2e-16
```

Variable Selection for Research

```
linear <- lm(Research~ Serial.No. + GRE.Score + TOEFL.Score + University.Rating + SOP +LOR + CGPA, data = test)
#summary(linear)

#Remove LOR
linear <- lm(Research~ Serial.No. + GRE.Score + TOEFL.Score + University.Rating +LOR + CGPA, data = test)
summary(linear)
```

```
##
## Call:
## lm(formula = Research ~ Serial.No. + GRE.Score + TOEFL.Score +
##      University.Rating + LOR + CGPA, data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.0861 -0.3358  0.0128  0.2852  0.9840
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.3639911   0.6526388   -9.751  < 2e-16 ***
## Serial.No.      0.0001593   0.0001284    1.240    0.215
## GRE.Score       0.0217245   0.0032763    6.631 8.79e-11 ***
## TOEFL.Score    -0.0051749   0.0059488   -0.870    0.385
## University.Rating 0.0365662   0.0240158    1.523    0.129
## LOR             0.0361827   0.0268979    1.345    0.179
## CGPA           0.0377370   0.0650001    0.581    0.562
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4084 on 493 degrees of freedom
## Multiple R-squared:  0.3325, Adjusted R-squared:  0.3243
## F-statistic: 40.92 on 6 and 493 DF,  p-value: < 2.2e-16
```

```
#Remove CGPA
linear <- lm(Research~ Serial.No. + GRE.Score + TOEFL.Score + University.Rating +LOR, data = test )
summary(linear)
```

```
##
## Call:
## lm(formula = Research ~ Serial.No. + GRE.Score + TOEFL.Score +
##      University.Rating + LOR, data = test)
##
## Residuals:
```

```
##      Min      1Q  Median      3Q      Max
## -1.0821 -0.3360  0.0127  0.2866  0.9834
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.4346661  0.6407546 -10.042  < 2e-16 ***
## Serial.No.      0.0001623  0.0001283   1.265   0.2064
## GRE.Score       0.0225289  0.0029669   7.593 1.57e-13 ***
## TOEFL.Score    -0.0041137  0.0056572  -0.727   0.4675
## University.Rating 0.0398047  0.0233433   1.705   0.0888 .
## LOR            0.0405427  0.0258109   1.571   0.1169
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4082 on 494 degrees of freedom
## Multiple R-squared:  0.332, Adjusted R-squared:  0.3253
## F-statistic: 49.11 on 5 and 494 DF,  p-value: < 2.2e-16

#Remove LOR
linear <- lm(Research~ Serial.No. + GRE.Score + TOEFL.Score + University.Rating, data = test )
summary(linear)

##
## Call:
## lm(formula = Research ~ Serial.No. + GRE.Score + TOEFL.Score +
##      University.Rating, data = test)
##
## Residuals:
##      Min      1Q  Median      3Q      Max
## -1.1057 -0.3428  0.0090  0.2871  1.0214
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.5778405  0.6351775 -10.356  < 2e-16 ***
## Serial.No.      0.0001785  0.0001280   1.394   0.1638
## GRE.Score       0.0228912  0.0029623   7.727 6.16e-14 ***
## TOEFL.Score    -0.0029714  0.0056186  -0.529   0.5971
## University.Rating 0.0536923  0.0216362   2.482   0.0134 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4088 on 495 degrees of freedom
## Multiple R-squared:  0.3287, Adjusted R-squared:  0.3233
## F-statistic: 60.59 on 4 and 495 DF,  p-value: < 2.2e-16

#Remove TOEFL
linear <- lm(Research~ Serial.No. + GRE.Score + University.Rating, data = test )
summary(linear)

##
## Call:
## lm(formula = Research ~ Serial.No. + GRE.Score + University.Rating,
##      data = test)
##
## Residuals:
##      Min      1Q  Median      3Q      Max
```

```
## -1.10835 -0.34957 0.00049 0.28952 1.02269
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.5389338  0.6304444 -10.372  <2e-16 ***
## Serial.No.      0.0001855  0.0001272   1.458  0.1455
## GRE.Score       0.0217887  0.0021030  10.361  <2e-16 ***
## University.Rating 0.0504027  0.0207077   2.434  0.0153 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4085 on 496 degrees of freedom
## Multiple R-squared:  0.3283, Adjusted R-squared:  0.3242
## F-statistic: 80.81 on 3 and 496 DF,  p-value: < 2.2e-16
```

```
#Remove Serial Number
linear <- lm(Research~ + GRE.Score + University.Rating, data = test )
summary(linear)
```

```
##
## Call:
## lm(formula = Research ~ +GRE.Score + University.Rating, data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.14033 -0.35017  0.00906  0.29255  1.00181
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.415603  0.625451 -10.258  <2e-16 ***
## GRE.Score       0.021546  0.002099  10.266  <2e-16 ***
## University.Rating 0.050337  0.020731   2.428  0.0155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4089 on 497 degrees of freedom
## Multiple R-squared:  0.3254, Adjusted R-squared:  0.3227
## F-statistic: 119.9 on 2 and 497 DF,  p-value: < 2.2e-16
```

Variable Selection for University Ranking

```
linear <- lm(University.Rating~ Serial.No. + GRE.Score + TOEFL.Score + SOP +LOR + CGPA + Research, data = test)
summary(linear)
```

```
##
## Call:
## lm(formula = University.Rating ~ Serial.No. + GRE.Score + TOEFL.Score +
##      SOP + LOR + CGPA + Research, data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.34352 -0.46556 -0.03557  0.44046  2.44809
##
```

```
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.4170319  1.2125399  -5.292 1.82e-07 ***
## Serial.No.   0.0001812  0.0002260   0.802  0.42307
## GRE.Score    0.0065910  0.0059476   1.108  0.26833
## TOEFL.Score  0.0209679  0.0103520   2.025  0.04336 *
## SOP          0.4474027  0.0507642   8.813 < 2e-16 ***
## LOR          0.1498125  0.0488318   3.068  0.00227 **
## CGPA         0.3578395  0.1141124   3.136  0.00182 **
## Research     0.0923371  0.0783086   1.179  0.23891
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7109 on 492 degrees of freedom
## Multiple R-squared:  0.619, Adjusted R-squared:  0.6135
## F-statistic: 114.2 on 7 and 492 DF,  p-value: < 2.2e-16

#Remove Serial Number
linear <- lm(University.Rating~ GRE.Score + TOEFL.Score + SOP +LOR + CGPA + Research, data = test )
summary(linear)

##
## Call:
## lm(formula = University.Rating ~ GRE.Score + TOEFL.Score + SOP +
##     LOR + CGPA + Research, data = test)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.36251 -0.47140 -0.04223  0.45376  2.41297
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.295220  1.202548  -5.235 2.45e-07 ***
## GRE.Score    0.006468  0.005943   1.088  0.27705
## TOEFL.Score  0.020128  0.010295   1.955  0.05114 .
## SOP         0.441757  0.050255   8.790 < 2e-16 ***
## LOR         0.154072  0.048524   3.175  0.00159 **
## CGPA        0.364222  0.113793   3.201  0.00146 **
## Research    0.096184  0.078133   1.231  0.21890
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7106 on 493 degrees of freedom
## Multiple R-squared:  0.6185, Adjusted R-squared:  0.6138
## F-statistic: 133.2 on 6 and 493 DF,  p-value: < 2.2e-16

#Remove GRE
linear <- lm(University.Rating~ TOEFL.Score + SOP +LOR + CGPA + Research, data = test )
summary(linear)

##
## Call:
## lm(formula = University.Rating ~ TOEFL.Score + SOP + LOR + CGPA +
##     Research, data = test)
##
## Residuals:
```

```
##      Min      1Q   Median      3Q      Max
## -2.37560 -0.47448 -0.03629  0.45065  2.41676
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.243653   0.715856  -7.325 9.79e-13 ***
## TOEFL.Score  0.025353   0.009109   2.783  0.00559 **
## SOP          0.440906   0.050259   8.773 < 2e-16 ***
## LOR          0.151540   0.048478   3.126  0.00188 **
## CGPA         0.414718   0.103920   3.991 7.59e-05 ***
## Research     0.120784   0.074805   1.615  0.10702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7107 on 494 degrees of freedom
## Multiple R-squared:  0.6176, Adjusted R-squared:  0.6137
## F-statistic: 159.5 on 5 and 494 DF,  p-value: < 2.2e-16

#Remove Research
linear <- lm(University.Rating~ TOEFL.Score + SOP +LOR + CGPA, data = test )
summary(linear)

##
## Call:
## lm(formula = University.Rating ~ TOEFL.Score + SOP + LOR + CGPA,
##     data = test)
##
## Residuals:
##      Min      1Q   Median      3Q      Max
## -2.46231 -0.46269 -0.04935  0.45262  2.39211
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.62010   0.67792  -8.290 1.07e-15 ***
## TOEFL.Score  0.02695   0.00907   2.971  0.00311 **
## SOP          0.44423   0.05030   8.832 < 2e-16 ***
## LOR          0.15563   0.04849   3.210  0.00142 **
## CGPA         0.44360   0.10254   4.326 1.83e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7119 on 495 degrees of freedom
## Multiple R-squared:  0.6155, Adjusted R-squared:  0.6124
## F-statistic: 198.1 on 4 and 495 DF,  p-value: < 2.2e-16
```

CV for linear model

```
set.seed(7861)

cvlm <- list()
msecv <- NA
for(i in 1:nrow(test)){
  #Fit the linear model
```


[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
## Warning in msecv[i] <- (predict(cvlm[[i]], newdata =  
## data.frame(TOEFL.Score[-i] + : number of items to replace is not a multiple  
## of replacement length  
  
## Warning in msecv[i] <- (predict(cvlm[[i]], newdata =  
## data.frame(TOEFL.Score[-i] + : number of items to replace is not a multiple  
## of replacement length  
  
#output mean of MSE  
mean(msecv)  
  
## [1] 4.549386
```