Assignment 3

Alex Teboul

Due: 2/27/2019 by 11:59pm

Class

DSC 423 - Data Analysis and Regression
Total points: 43pts

Problem 1 [33 points]

This problem asks you to build a model for the college dataset (college.csv) that contains the following variables: * school: School name * Private: public/private indicator. YES if university is privare, NO if university is public. * Accept.pct: percentage of applicants accepted * Elite10: Elite schools with majority of students from the top 10% of their high school class * F.Undergrad: number of full-time undergraduate students * P.Undergrad: number of part-time undergraduate students * Outstate: Out-of-state tuition * Room.Board: room and board costs * vBooks: estimated book costs * Personal: Estimated personal spending * PhD: Percent of faculty with terminal degrees * vS.F.Ratio: Student/faculty ratio * perc.alumni: Percent of alumni who donate * vExpend: Instructional expenditure per student * vGrad.Rate: Graduation rate in 4 years **Apply regression analysis techniques to analyze the relationship among the observed variables and build a model to predict Graduation Rates (Grad.Rate). Answer the following questions:**

Problem 1 a)

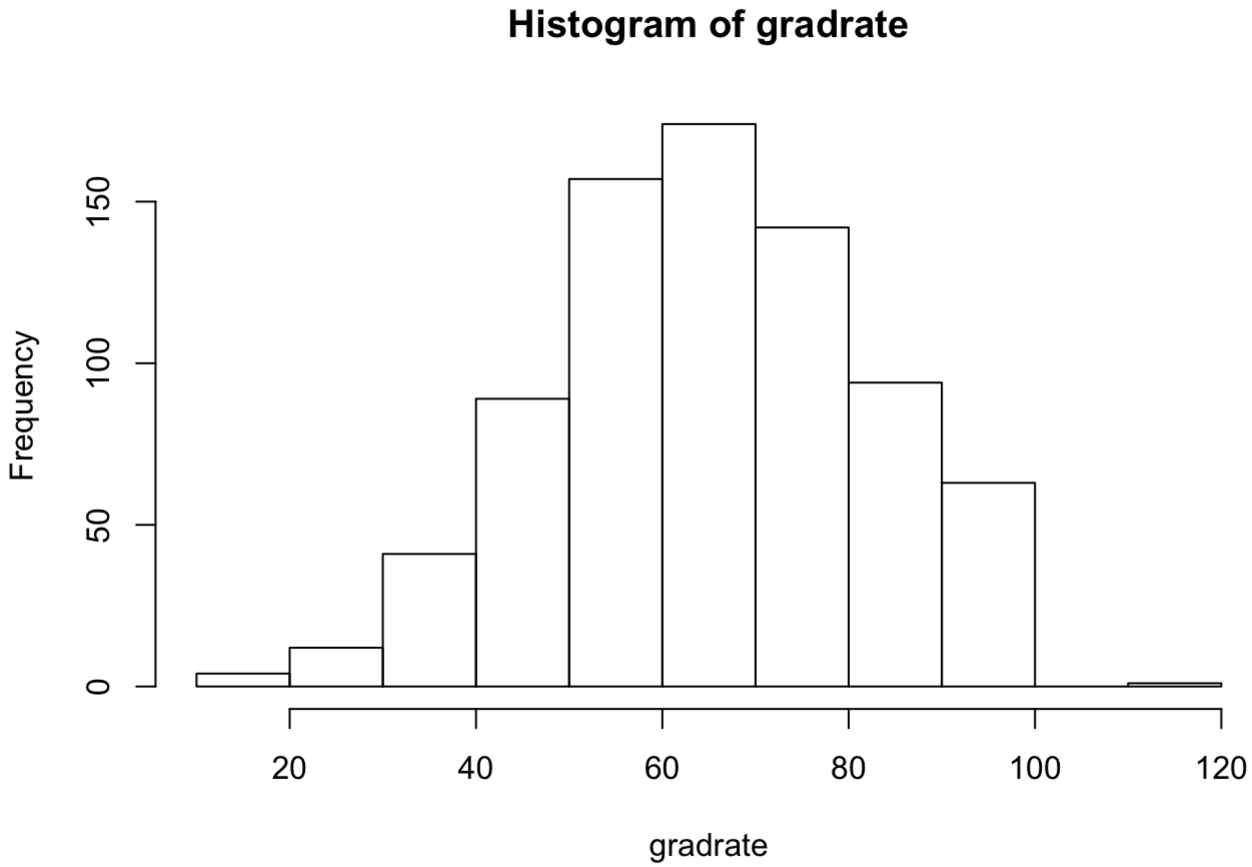
a) Analyze the distribution of Grad.Rate and discuss if the distribution is symmetric, or if you need to apply any transformation. [1 pt R code, 1 pt distribution plot, 1 pt answer = 3 pts]

```
# load in the data from file
myd=read.csv("college.csv", header=T)

myd[1,]
```

```
##                          school Private Accept.pct Elite10 F.Undergrad
## 1 Abilene Christian University     Yes  0.7421687       0        2885
##   P.Undergrad Outstate Room.Board Books Personal PhD Terminal S.F.Ratio
## 1         537     7440        3300   450     2200  70       78      18.1
##   perc.alumni Expend Grad.Rate
## 1          12   7041        60
```

```r
attach(myd)
#create dummy variable for Private;
private=(Private=='Yes')*1
myd=cbind(myd, private)

# get the variables
#school = myd$school #IGNORE SCHOOL
private = myd$private
acceptpct = myd$Accept.pct
elite = myd$Elite10
fundergrad = myd$F.Undergrad
pundergrad = myd$P.Undergrad
outstate = myd$Outstate
roomboardcosts = myd$Room.Board
bookcosts = myd$Books
personal = myd$Personal
phd = myd$PhD
terminal = myd$Terminal
sfratio = myd$S.F.Ratio
percalumni = myd$perc.alumni
expend = myd$Expend

gradrate = myd$Grad.Rate

#gradrate distribution
# plot the histogram of gradrate
hist(gradrate)

# compute descriptive statistics
library(psych)
```

```
## Warning: package 'psych' was built under R version 3.5.2
```



```
describe(gradrate)
```

```
##    vars   n  mean    sd median trimmed   mad min max range  skew kurtosis
## X1    1 777 65.46 17.18     65    65.6 17.79  10 118   108 -0.11    -0.22
##      se
## X1 0.62
```

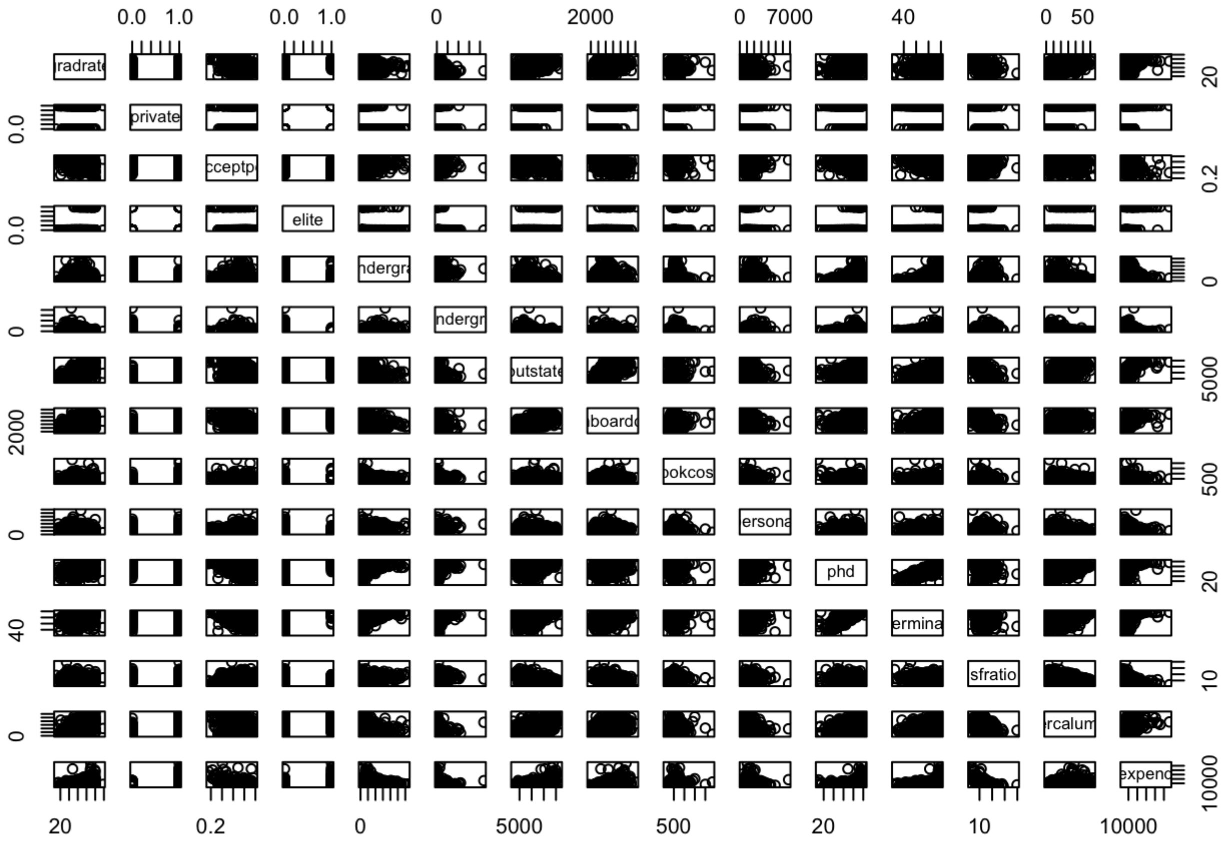
Answer: The distribution of Grad.Rate appears to be symmetric and normal. Grad.Rate has a mean of 65.46 and a median of 65, standard deviation of 17.18, skew of -0.11, on a sample of 777 schools. I do not believe that a transformation is necessary here.

It should be noted that Cazenovia College has a Grad.Rate of 118, which could be an error. Perhaps they graduated more students than they took in, yeilding a number over 100. I will keep it in the dataset, because I do not have the information to say whether or not it is an error.

Problem 1 b)

b) Create scatterplots for Grad.Rate vs each of the independent variables. What conclusions can you draw about the relationships between Grad.Rate and the independent variables? (No need to include the scatterplots in your submission, but you can use correlation analysis) [1 pt R code, 2 pts answer = 3 pts]

```
# general scatterplot matrix for quantitative variables
pairs(gradrate ~ private+acceptpct+elite+fundergrad+pundergrad+outstate+roomboardcosts+b
ookcosts+personal+phd+terminal+sfratio+percalumni+expend)
```

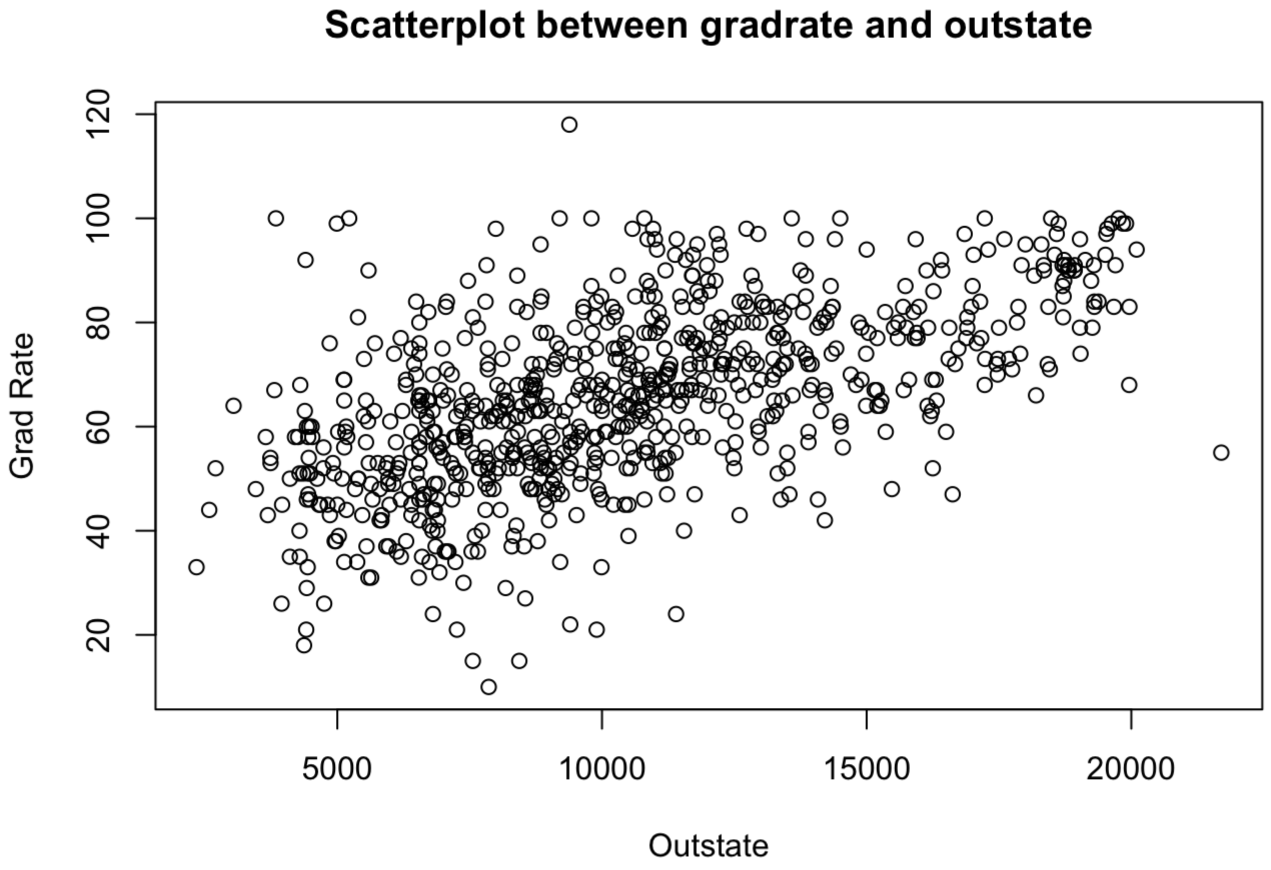


```
cor(myd[,3:17])
```

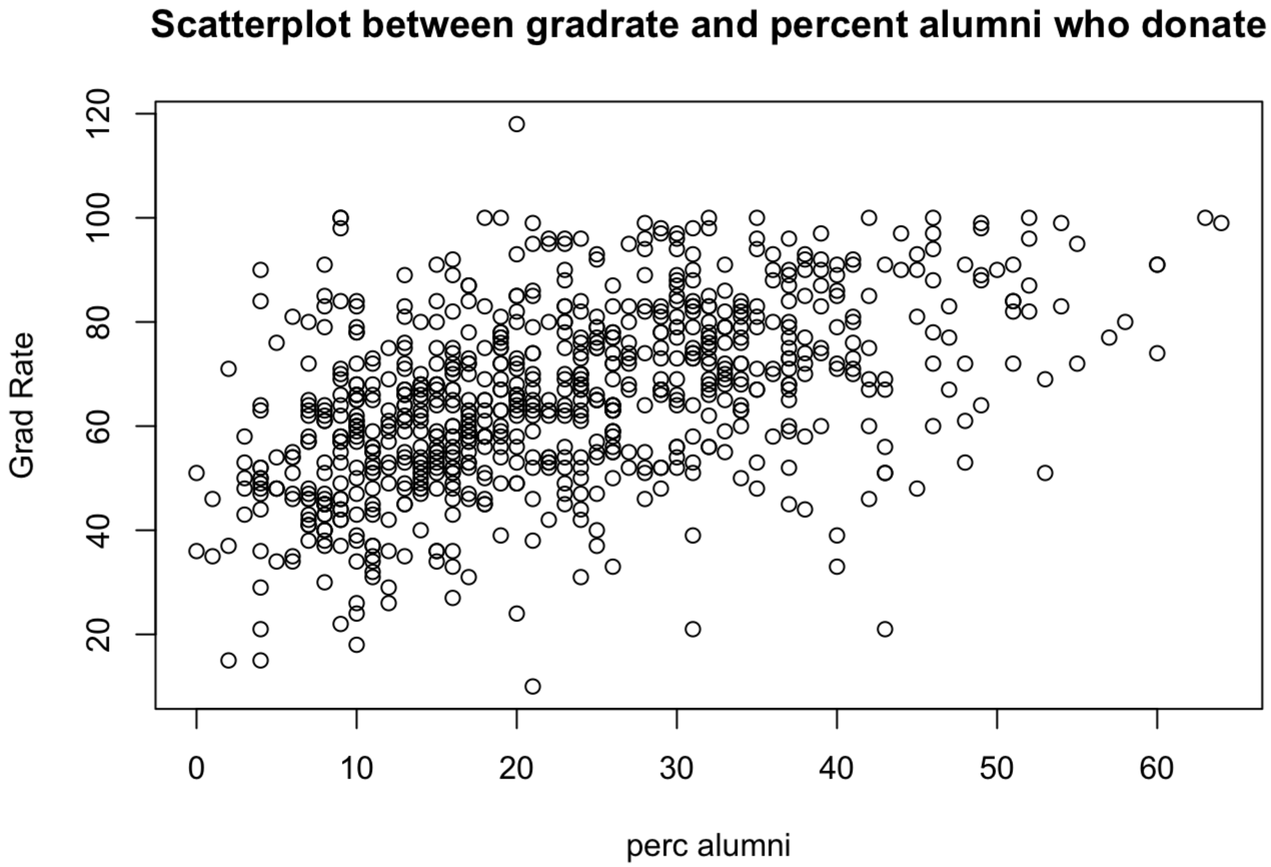
```
##              Accept.pct     Elite10 F.Undergrad P.Undergrad    Outstate
## Accept.pct   1.00000000 -0.46245330 -0.15565379 -0.09228664 -0.24095073
## Elite10     -0.46245330  1.00000000  0.06083999 -0.11644570  0.39947675
## F.Undergrad -0.15565379  0.06083999  1.00000000  0.57051219 -0.21574200
## P.Undergrad -0.09228664 -0.11644570  0.57051219  1.00000000 -0.25351232
## Outstate    -0.24095073  0.39947675 -0.21574200 -0.25351232  1.00000000
## Room.Board  -0.31030204  0.29847208 -0.06889039 -0.06132551  0.65425640
## Books       -0.17407288  0.09217607  0.11554976  0.08119952  0.03885487
## Personal     0.01997851 -0.07526924  0.31719954  0.31988162 -0.29908690
## PhD         -0.31833394  0.34106219  0.31833697  0.14911422  0.38298241
## Terminal    -0.30379999  0.32664984  0.30001894  0.14190357  0.40798320
## S.F.Ratio    0.10998188 -0.29349738  0.27970335  0.23253051 -0.55482128
## perc.alumni -0.13210402  0.30259090 -0.22946222 -0.28079236  0.56626242
## Expend      -0.40862232  0.55977784  0.01865162 -0.08356842  0.67277862
## Grad.Rate   -0.28697150  0.34873255 -0.07877313 -0.25700099  0.57128993
## private      0.08499047  0.07962886 -0.61556054 -0.45208775  0.55264990
##              Room.Board        Books    Personal         PhD    Terminal
## Accept.pct  -0.31030204 -0.174072883  0.01997851 -0.31833394 -0.30379999
## Elite10      0.29847208  0.092176073 -0.07526924  0.34106219  0.32664984
## F.Undergrad -0.06889039  0.115549761  0.31719954  0.31833697  0.30001894
## P.Undergrad -0.06132551  0.081199521  0.31988162  0.14911422  0.14190357
## Outstate     0.65425640  0.038854868 -0.29908690  0.38298241  0.40798320
## Room.Board   1.00000000  0.127962970 -0.19942818  0.32920228  0.37453955
## Books        0.12796297  1.000000000  0.17929476  0.02690573  0.09995470
## Personal    -0.19942818  0.179294764  1.00000000 -0.01093579 -0.03061311
## PhD          0.32920228  0.026905731 -0.01093579  1.00000000  0.84958703
## Terminal     0.37453955  0.099954700 -0.03061311  0.84958703  1.00000000
## S.F.Ratio   -0.36262774 -0.031929274  0.13634483 -0.13053011 -0.16010395
## perc.alumni  0.27236345 -0.040207736 -0.28596808  0.24900866  0.26713029
## Expend       0.50173942  0.112409075 -0.09789189  0.43276168  0.43879922
## Grad.Rate    0.42494154  0.001060894 -0.26934396  0.30503785  0.28952723
## private      0.34053206 -0.018548975 -0.30448505 -0.15671437 -0.12961994
##               S.F.Ratio perc.alumni      Expend    Grad.Rate     private
## Accept.pct   0.10998188 -0.13210402 -0.40862232 -0.286971504  0.08499047
## Elite10     -0.29349738  0.30259090  0.55977784  0.348732550  0.07962886
## F.Undergrad  0.27970335 -0.22946222  0.01865162 -0.078773129 -0.61556054
## P.Undergrad  0.23253051 -0.28079236 -0.08356842 -0.257000991 -0.45208775
## Outstate    -0.55482128  0.56626242  0.67277862  0.571289928  0.55264990
## Room.Board  -0.36262774  0.27236345  0.50173942  0.424941541  0.34053206
## Books       -0.03192927 -0.04020774  0.11240908  0.001060894 -0.01854897
## Personal     0.13634483 -0.28596808 -0.09789189 -0.269343964 -0.30448505
## PhD         -0.13053011  0.24900866  0.43276168  0.305037850 -0.15671437
## Terminal    -0.16010395  0.26713029  0.43879922  0.289527232 -0.12961994
## S.F.Ratio    1.00000000 -0.40292917 -0.58383204 -0.306710405 -0.47220474
## perc.alumni -0.40292917  1.00000000  0.41771172  0.490897562  0.41477493
## Expend      -0.58383204  0.41771172  1.00000000  0.390342696  0.25846068
## Grad.Rate   -0.30671041  0.49089756  0.39034270  1.000000000  0.33616229
## private     -0.47220474  0.41477493  0.25846068  0.336162290  1.00000000
```

Answer: My first conclusion is that Grad.Rate is not strongly correlated with any of the variables. It is most strongly correlated with Outstate (0.571), perc.alumni (0.491), and Room.Board (0.425). Some variables also have negative correlations, but they are weak correlations as well. Below, I explore the plots for these most highly correlated variables.

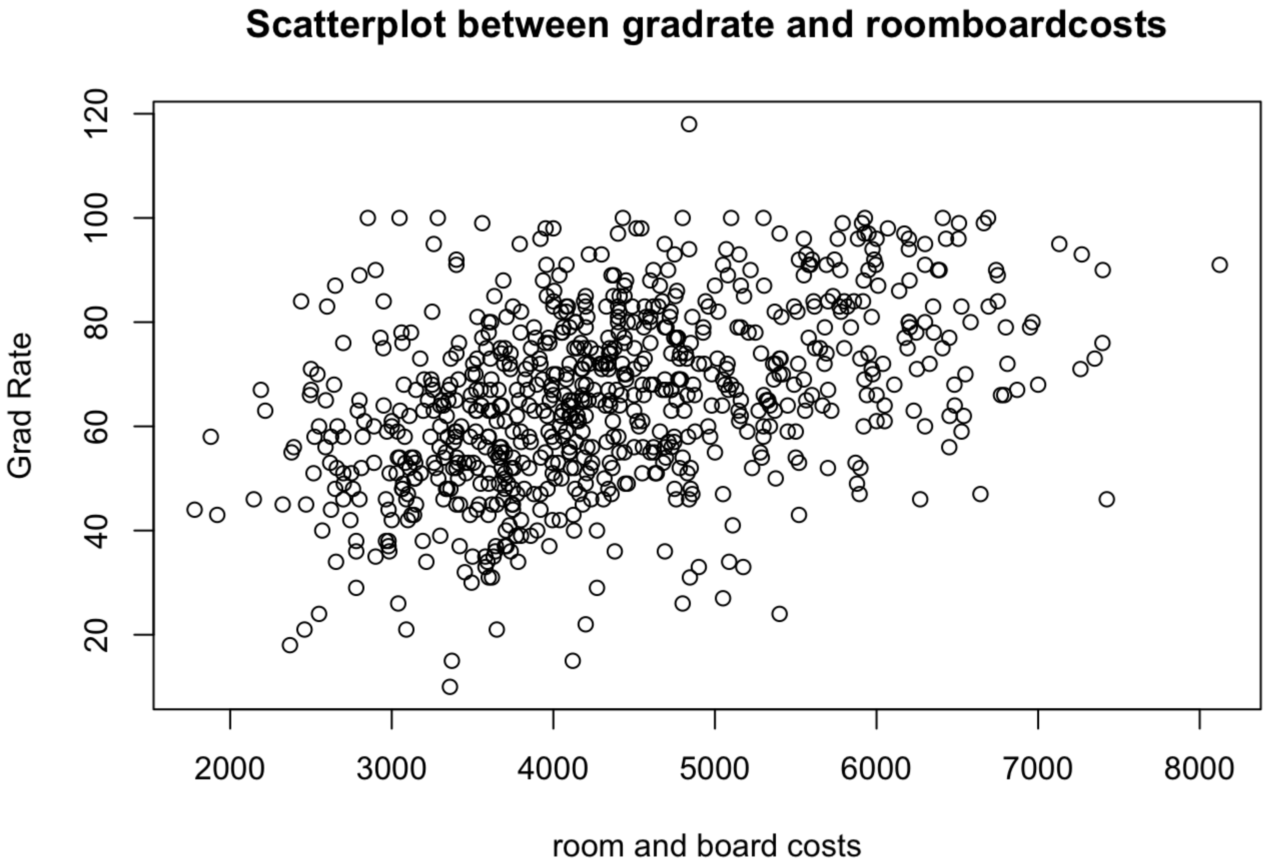
```
#scatterplot between gradrate and outstate, perc.alumni, and room.board
plot(outstate, gradrate, main="Scatterplot between gradrate and outstate", xlab="Outstate", ylab="Grad Rate")
```



```
plot(percalumni , gradrate, main="Scatterplot between gradrate and percent alumni who donate", xlab="perc alumni", ylab="Grad Rate")
```



```
plot(roomboardcosts, gradrate, main="Scatterplot between gradrate and roomboardcosts", xlab="room and board costs", ylab="Grad Rate")
```

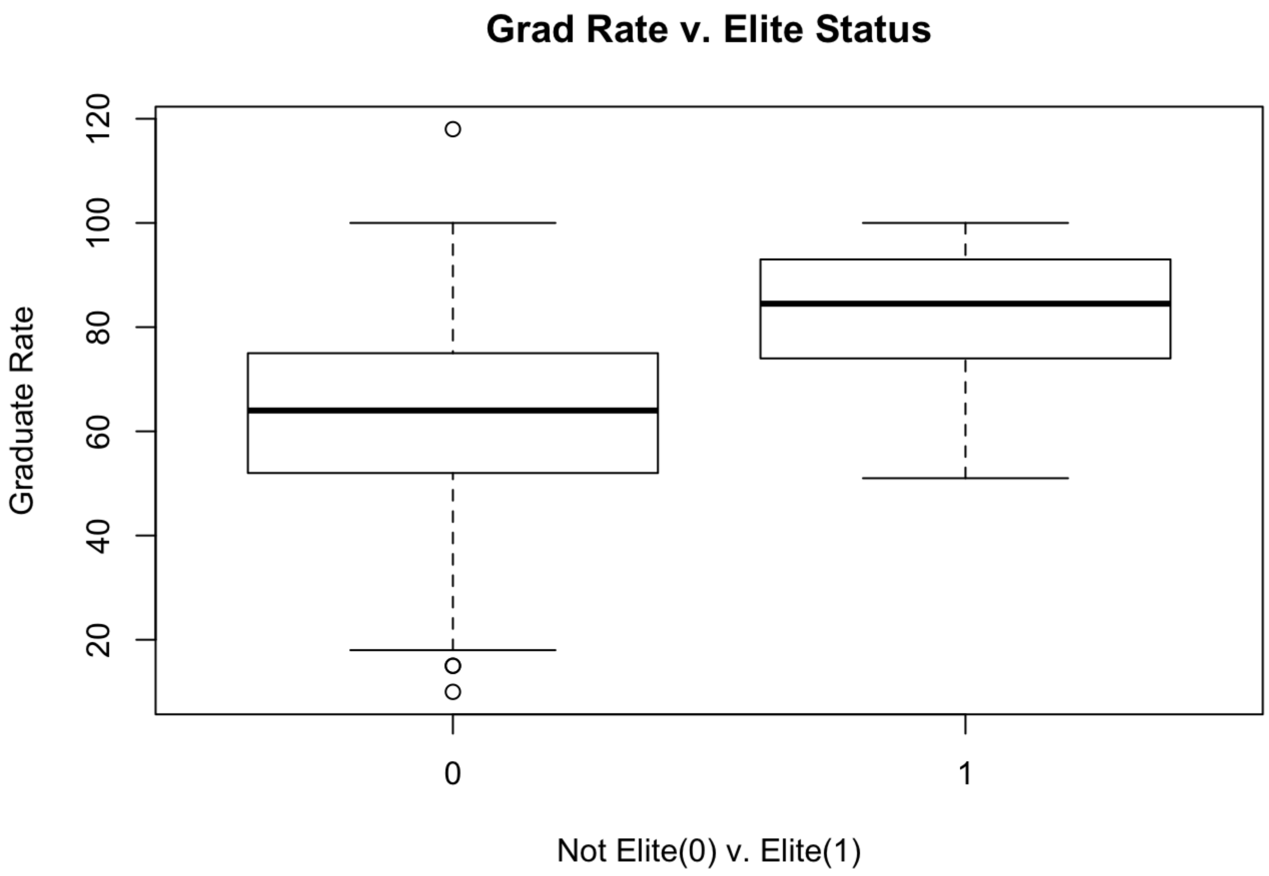


Answer: The plots illustrate the loose positive correlation between Grad Rate and these most strongly correlated three variables.

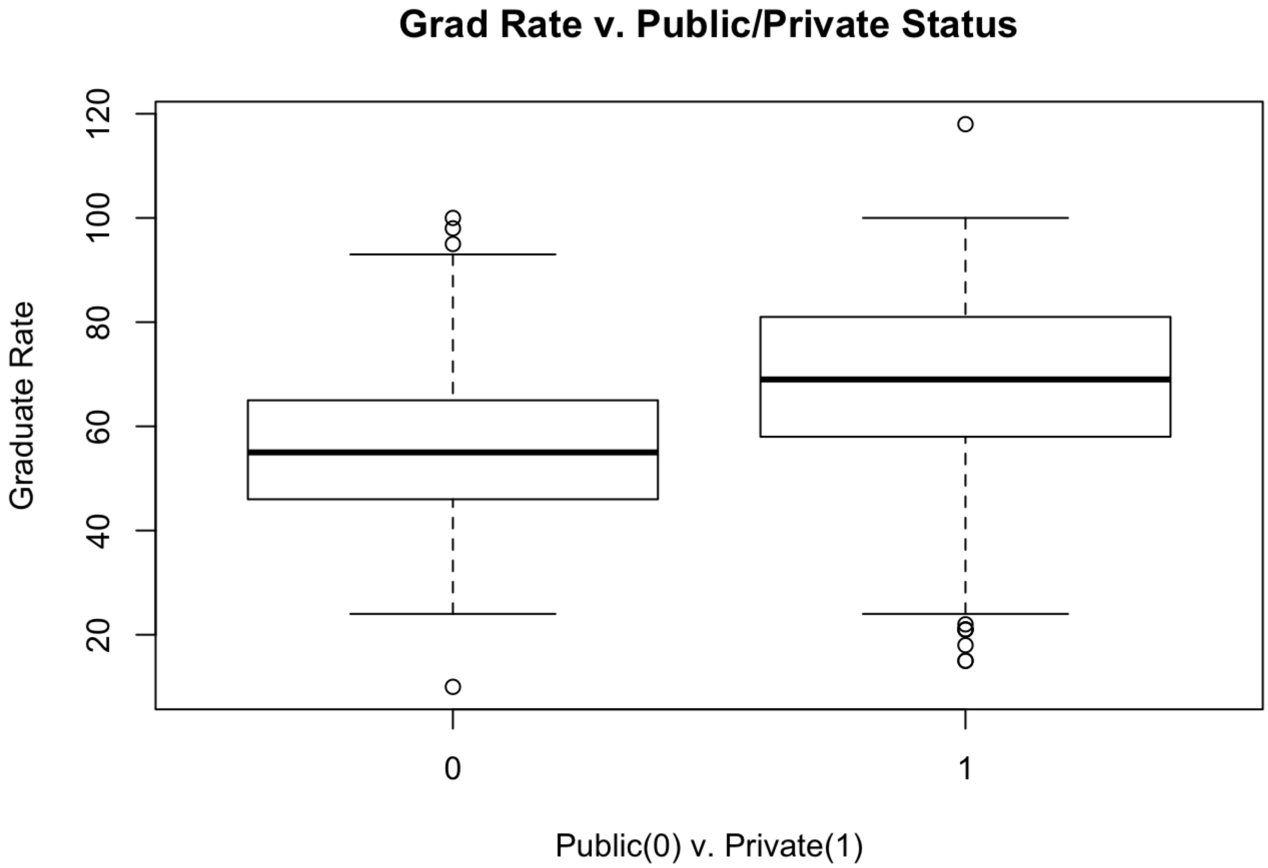
Problem 1 c)

c) Build boxplots to evaluate if graduation rates vary by university type (private vs public) and by status (elite vs not elite). Discuss your findings. [1 pt R code, 1 pt boxplots, 1 pt answer = 3 pts]

```
#public v private
boxplot(gradrate~elite,data=myd, main="Grad Rate v. Elite Status", xlab="Not Elite(0) v.
 Elite(1)",ylab="Graduate Rate")
```



```
#elite v not elite
boxplot(gradrate~private,data=myd, main="Grad Rate v. Public/Private Status", xlab="Publ
ic(0) v. Private(1)",ylab="Graduate Rate")
```



Answer: Without going into the tests of significance, one can clearly see that both “Elite” schools and “Private” schools have higher graduation rates. This is indicated by the higher Q1,Q3, and median values displayed by these schools in the boxplot visualizations.

Problem 1 d)

d) Fit a full model (with all independent variables) to predict Grad.Rate [1 pt R code, 1 pt full model equation = 2 pts]

```
# Fit a regression model (M1) gradrate v. the other variables (SchoolName is ignored)
M1 <- lm(gradrate ~ private+acceptpct+elite+fundergrad+pundergrad+outstate+roomboardcost
s+bookcosts+personal+phd+terminal+sfratio+percalumni+expend, data=myd)

# Model parameters
summary(M1)
```

```
##
## Call:
## lm(formula = gradrate ~ private + acceptpct + elite + fundergrad +
##     pundergrad + outstate + roomboardcosts + bookcosts + personal +
##     phd + terminal + sfratio + percalumni + expend, data = myd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.200  -6.777  -0.707   7.217  57.907
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.140e+01  6.124e+00   8.393 2.30e-16 ***
## private         4.620e+00  1.722e+00   2.683  0.00746 **
## acceptpct      -1.811e+01  3.843e+00  -4.712 2.91e-06 ***
## elite           4.017e+00  2.003e+00   2.005  0.04527 *
## fundergrad      6.809e-04  1.429e-04   4.767 2.24e-06 ***
## pundergrad     -1.956e-03  3.904e-04  -5.009 6.80e-07 ***
## outstate        1.235e-03  2.286e-04   5.401 8.88e-08 ***
## roomboardcosts  1.667e-03  5.944e-04   2.805  0.00517 **
## bookcosts      -2.524e-03  2.966e-03  -0.851  0.39511
## personal       -1.718e-03  7.781e-04  -2.208  0.02753 *
## phd             1.306e-01  5.621e-02   2.324  0.02037 *
## terminal       -7.284e-02  6.257e-02  -1.164  0.24469
## sfratio         1.003e-03  1.619e-01   0.006  0.99506
## percalumni      3.092e-01  4.839e-02   6.390 2.89e-10 ***
## expend         -4.365e-04  1.518e-04  -2.875  0.00415 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.92 on 762 degrees of freedom
## Multiple R-squared:  0.4448, Adjusted R-squared:  0.4346
## F-statistic: 43.61 on 14 and 762 DF,  p-value: < 2.2e-16
```

Answer: Full Model Equation:

GradRate = 5.140e+01(Intercept) + 4.620e+00(private) - 1.811e+01(acceptpct) + 4.017e+00(elite) + 6.809e-04(fundergrad) - 1.956e-03(pundergrad) + 1.235e-03(outstate) + 1.667e-03(roomboardcosts) - 2.524e-03(bookcosts) - 1.718e-03(personal) + 1.306e-01(phd) - 7.284e-02(terminal) + 1.003e-03(sfratio) + 3.092e-01(percalumni) - 4.365e-04(expend)

Adjusted R-squared: 0.4346

Insignificant Variables:1) bookcosts: p(0.39511) 2) terminal: p(0.24469) 3)sfratio: p(0.99506)

```
#Removing Insignficant Variables
M1_removevars <- lm(gradrate ~ private + acceptpct + elite + fundergrad + pundergrad + o
utstate + roomboardcosts + personal + phd + percalumni + expend, data=myd)
summary(M1_removevars)
```

```
##
## Call:
## lm(formula = gradrate ~ private + acceptpct + elite + fundergrad +
##     pundergrad + outstate + roomboardcosts + personal + phd +
##     percalumni + expend, data = myd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.085  -6.932  -0.775   7.325  57.598
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.840e+01  4.621e+00  10.475  < 2e-16 ***
## private         4.770e+00  1.689e+00   2.824  0.00486 **
## acceptpct      -1.778e+01  3.797e+00  -4.683 3.34e-06 ***
## elite           4.022e+00  2.002e+00   2.009  0.04492 *
## fundergrad      6.631e-04  1.411e-04   4.699 3.10e-06 ***
## pundergrad     -1.963e-03  3.901e-04  -5.031 6.09e-07 ***
## outstate        1.215e-03  2.270e-04   5.352 1.15e-07 ***
## roomboardcosts  1.534e-03  5.878e-04   2.610  0.00924 **
## personal       -1.820e-03  7.638e-04  -2.383  0.01742 *
## phd             8.424e-02  3.706e-02   2.273  0.02329 *
## percalumni      3.060e-01  4.806e-02   6.367 3.32e-10 ***
## expend         -4.465e-04  1.390e-04  -3.211  0.00138 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.91 on 765 degrees of freedom
## Multiple R-squared:  0.4431, Adjusted R-squared:  0.4351
## F-statistic: 55.33 on 11 and 765 DF,  p-value: < 2.2e-16
```

Answer: Full Model Equation when insignificant variables are removed:

GradRate = 4.840e+01(Intercept) + 4.770e+00 (private) - 1.778e+01(acceptpct) + 4.022e+00(elite) + 6.631e-04(fundergrad) - 1.963e-03(pundergrad) + 1.215e-03(outstate) + 1.534e-03(roomboardcosts) - 1.820e-03(personal) + 8.424e-02(phd) + 3.060e-01(percalumni) - 4.465e-04(expend)

Adjusted R-squared: 0.4351 (barely up from 0.4346)

Problem 1 e)

e) Does multi-collinearity seem to be a problem here? What is your evidence? Compute and analyze the VIF statistics. [1 pt R code, 1 pt VIF statistics, 2 pts answer = 4 pts]

```r
# Compute the VIF statistics
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
## The following object is masked from 'package:psych':
##
##     logit
```

```
vif(M1)
```

```
##        private      acceptpct          elite     fundergrad     pundergrad
##       2.739521       1.486633       1.687903       2.233117       1.643393
##       outstate roomboardcosts      bookcosts       personal            phd
##       3.935059       1.976762       1.115823       1.290983       3.917716
##       terminal        sfratio     percalumni         expend
##       3.946581       1.909722       1.672367       2.922643
```

Answer: No, multicollinearity does not seem to be a problem here. Specifically, after calculating the VIF statistics, none of the variables had VIF values over 10.

Problem 1 f)

f) Apply TWO variable selection procedures to find an optimal subset of independent variables to predict Grad.Rate. You can choose any two procedures among the ones we learned in class: backward selection, forward selection, adj-R2, Cp, stepwise, press. [2 pts R code for the 2 variable selection procedures, 1 pt answer = 3 pts]

```
# Variable Selection Procedure 1: Backward Selection
step(M1, direction=c("backward"), trace=F)
```

```
##
## Call:
## lm(formula = gradrate ~ private + acceptpct + elite + fundergrad +
##     pundergrad + outstate + roomboardcosts + personal + phd +
##     percalumni + expend, data = myd)
##
## Coefficients:
##    (Intercept)         private       acceptpct           elite
##      4.840e+01       4.770e+00      -1.778e+01       4.022e+00
##     fundergrad      pundergrad        outstate  roomboardcosts
##      6.631e-04      -1.963e-03       1.215e-03       1.534e-03
##       personal             phd      percalumni          expend
##     -1.820e-03       8.424e-02       3.060e-01      -4.465e-04
```

```
#Backward Slection Model
M1_backward <- lm(formula = gradrate ~ private + acceptpct + elite + fundergrad + punder
grad + outstate + roomboardcosts + personal + phd + percalumni + expend, data = myd)
summary(M1_backward)
```

```
##
## Call:
## lm(formula = gradrate ~ private + acceptpct + elite + fundergrad +
##     pundergrad + outstate + roomboardcosts + personal + phd +
##     percalumni + expend, data = myd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.085  -6.932  -0.775   7.325  57.598
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.840e+01  4.621e+00  10.475  < 2e-16 ***
## private         4.770e+00  1.689e+00   2.824  0.00486 **
## acceptpct      -1.778e+01  3.797e+00  -4.683 3.34e-06 ***
## elite           4.022e+00  2.002e+00   2.009  0.04492 *
## fundergrad      6.631e-04  1.411e-04   4.699 3.10e-06 ***
## pundergrad     -1.963e-03  3.901e-04  -5.031 6.09e-07 ***
## outstate        1.215e-03  2.270e-04   5.352 1.15e-07 ***
## roomboardcosts  1.534e-03  5.878e-04   2.610  0.00924 **
## personal       -1.820e-03  7.638e-04  -2.383  0.01742 *
## phd             8.424e-02  3.706e-02   2.273  0.02329 *
## percalumni      3.060e-01  4.806e-02   6.367 3.32e-10 ***
## expend         -4.465e-04  1.390e-04  -3.211  0.00138 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.91 on 765 degrees of freedom
## Multiple R-squared:  0.4431, Adjusted R-squared:  0.4351
## F-statistic: 55.33 on 11 and 765 DF,  p-value: < 2.2e-16
```

Answer: Backward Selection : lm(formula = gradrate ~ private + acceptpct + elite + fundergrad + pundergrad + outstate + roomboardcosts + personal + phd + percalumni + expend, data = myd)

```
# Variable Selection Procedure 2: Forward Selection
Base = lm(gradrate~1)
step(Base, scope=list( upper=M1, lower=~1 ), direction=c("forward"), trace=F)
```

```
##
## Call:
## lm(formula = gradrate ~ outstate + percalumni + acceptpct + pundergrad +
##     fundergrad + roomboardcosts + expend + personal + private +
##     phd + elite)
##
## Coefficients:
##    (Intercept)        outstate      percalumni       acceptpct
##      4.840e+01       1.215e-03       3.060e-01      -1.778e+01
##     pundergrad      fundergrad  roomboardcosts          expend
##     -1.963e-03       6.631e-04       1.534e-03      -4.465e-04
##       personal         private             phd           elite
##     -1.820e-03       4.770e+00       8.424e-02       4.022e+00
```

```
M1_forward <- lm(formula = gradrate ~ private + acceptpct + elite + fundergrad + punderg
rad + outstate + roomboardcosts + personal + phd + percalumni + expend, data = myd)

summary(M1_forward)
```

```
##
## Call:
## lm(formula = gradrate ~ private + acceptpct + elite + fundergrad +
##     pundergrad + outstate + roomboardcosts + personal + phd +
##     percalumni + expend, data = myd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.085  -6.932  -0.775   7.325  57.598
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.840e+01  4.621e+00  10.475  < 2e-16 ***
## private         4.770e+00  1.689e+00   2.824  0.00486 **
## acceptpct      -1.778e+01  3.797e+00  -4.683 3.34e-06 ***
## elite           4.022e+00  2.002e+00   2.009  0.04492 *
## fundergrad      6.631e-04  1.411e-04   4.699 3.10e-06 ***
## pundergrad     -1.963e-03  3.901e-04  -5.031 6.09e-07 ***
## outstate        1.215e-03  2.270e-04   5.352 1.15e-07 ***
## roomboardcosts  1.534e-03  5.878e-04   2.610  0.00924 **
## personal       -1.820e-03  7.638e-04  -2.383  0.01742 *
## phd             8.424e-02  3.706e-02   2.273  0.02329 *
## percalumni      3.060e-01  4.806e-02   6.367 3.32e-10 ***
## expend         -4.465e-04  1.390e-04  -3.211  0.00138 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.91 on 765 degrees of freedom
## Multiple R-squared:  0.4431, Adjusted R-squared:  0.4351
## F-statistic: 55.33 on 11 and 765 DF,  p-value: < 2.2e-16
```

Answer: Both Backward and Forward Selection lead to the linear model equation below. They both have adj-r2 values of 0.4531, and removed the same three variables bookcosts, terminal, and sfratio.

GradRate = 4.840e+01(Intercept) + 4.770e+00(private) - 1.778e+01(acceptpct) + 4.022e+00(elite) + 6.631e-04(fundergrad) - 1.963e-03(pundergrad) + 1.215e-03(outstate) + 1.534e-03(roomboardcosts) - 1.820e-03(personal) + 8.424e-02(phd) + 3.060e-01(percalumni) - 4.465e-04(expend)

Problem 1 g)

g) Fit a final regression model M1 for Grad.Rate based on the results in f). Explain your choice. Write down the expression of the estimated model M1. [1 pt R code for final model, 1 pt explanation, 1 pt expression = 3 pts]

```
M1_final <- lm(formula = gradrate ~ private + acceptpct + elite + fundergrad + pundergra
d + outstate + roomboardcosts + personal + phd + percalumni + expend, data = myd)
summary(M1_final)
```

```
##
## Call:
## lm(formula = gradrate ~ private + acceptpct + elite + fundergrad +
##     pundergrad + outstate + roomboardcosts + personal + phd +
##     percalumni + expend, data = myd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.085  -6.932  -0.775   7.325  57.598
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.840e+01  4.621e+00  10.475  < 2e-16 ***
## private         4.770e+00  1.689e+00   2.824  0.00486 **
## acceptpct      -1.778e+01  3.797e+00  -4.683 3.34e-06 ***
## elite           4.022e+00  2.002e+00   2.009  0.04492 *
## fundergrad      6.631e-04  1.411e-04   4.699 3.10e-06 ***
## pundergrad     -1.963e-03  3.901e-04  -5.031 6.09e-07 ***
## outstate        1.215e-03  2.270e-04   5.352 1.15e-07 ***
## roomboardcosts  1.534e-03  5.878e-04   2.610  0.00924 **
## personal       -1.820e-03  7.638e-04  -2.383  0.01742 *
## phd             8.424e-02  3.706e-02   2.273  0.02329 *
## percalumni      3.060e-01  4.806e-02   6.367 3.32e-10 ***
## expend         -4.465e-04  1.390e-04  -3.211  0.00138 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.91 on 765 degrees of freedom
## Multiple R-squared:  0.4431, Adjusted R-squared:  0.4351
## F-statistic: 55.33 on 11 and 765 DF,  p-value: < 2.2e-16
```

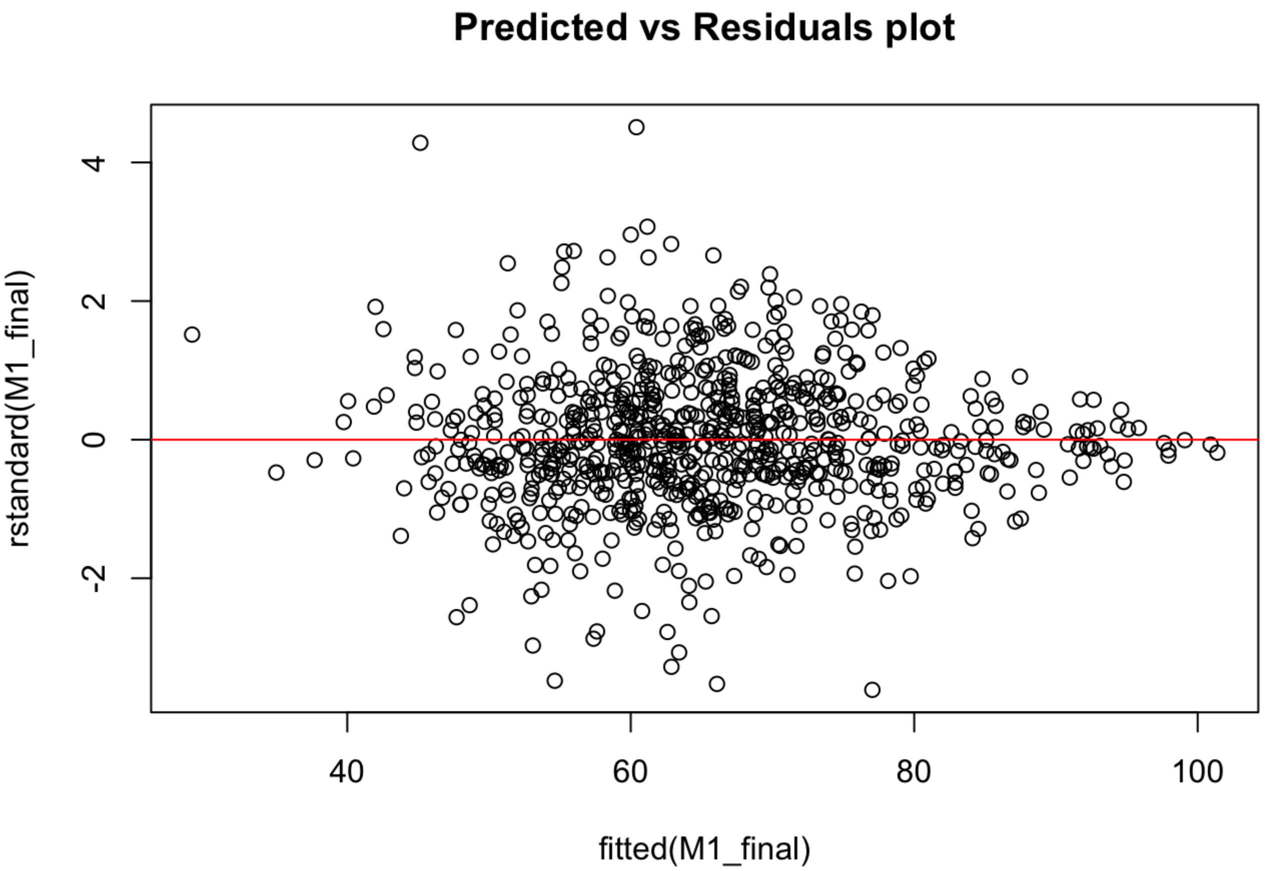
Answer: As explained in part f, M1_final removed variables bookcosts, terminal, and sfratio. The adj-r2 of M1_final is 0.4351. All included variables are significant at at least the 0.01 level. The parameters for this model were chose by backward and forward selection, as well as the standard, remove insignificant variables technique. All methods led to this model, which slightly increased the adj-r2 of the model to 0.4351 from 0.4346. The model equation once again is:

GradRate = 4.840e+01(Intercept) + 4.770e+00(private) - 1.778e+01(acceptpct) + 4.022e+00(elite) + 6.631e-04(fundergrad) - 1.963e-03(pundergrad) + 1.215e-03(outstate) + 1.534e-03(roomboardcosts) - 1.820e-03(personal) + 8.424e-02(phd) + 3.060e-01(percalumni) - 4.465e-04(expend)

Problem 1 h)

h) Draw a scatter plot of the studentized residuals against the predicted values. Does the plot show any striking pattern indicating problems in the regression analysis? [1 pt R code, 1 pt answer = 2 pts]

```
#residual plots
#Plot residuals vs predicted values
plot( fitted(M1_final), rstandard(M1_final), main="Predicted vs Residuals plot")
abline(a=0, b=0, col='red') #add zero line
```

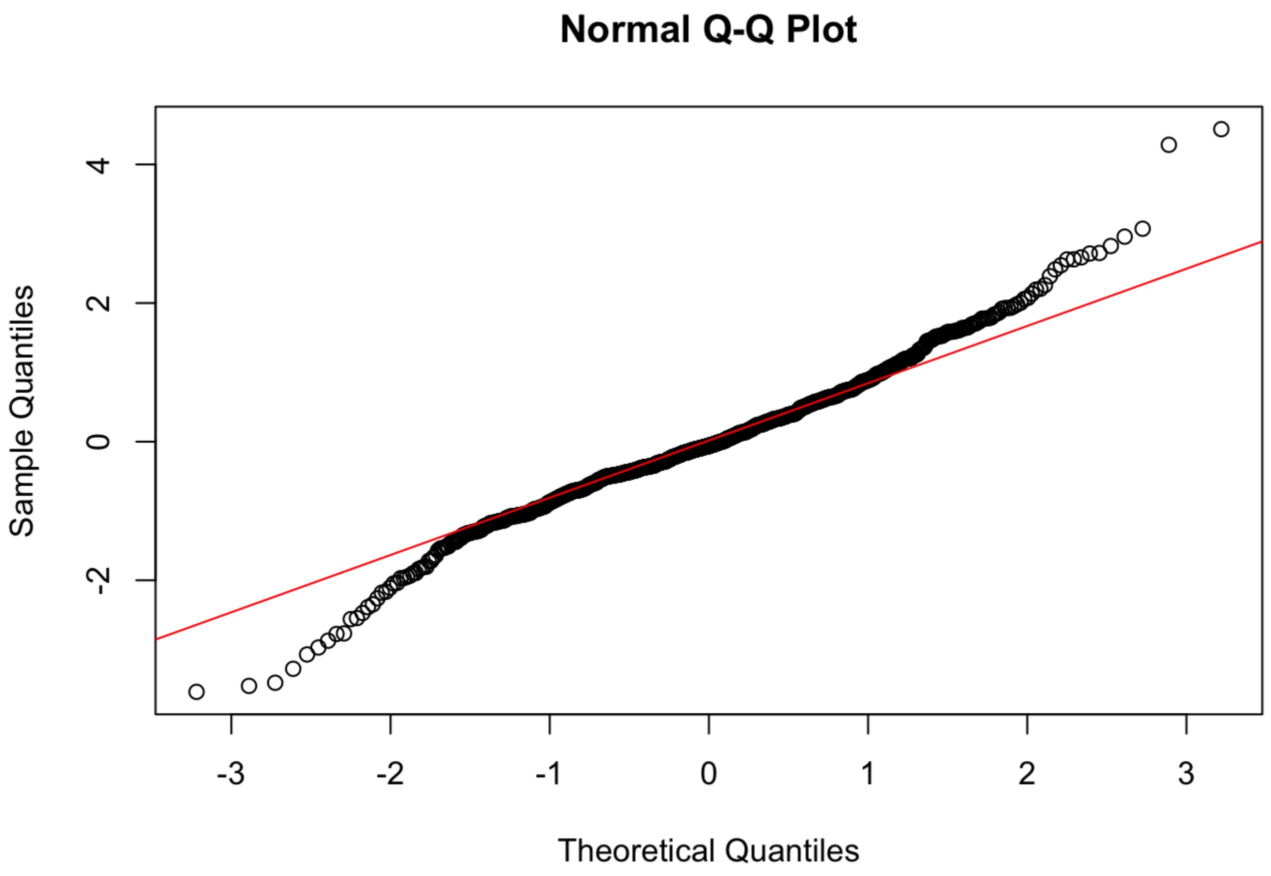


Answer: This plot, predicted v. residuals, does appear to show a random scatter for the most part, so linearity seems to be mostly satisfied. Though constant variance cannot be fully guaranteed. There may be a slight variance problem given the way the plot tapers in a V shape as fitted(M1_final) approaches 100.

Problem 1 i)

i) Analyze normal probability plot of residuals. Is there any evidence that the assumption of normality is not satisfied? [1 pt R code, 1 pt answer = 2 pts]

```
#normal probability plot of residuals
qqnorm(rstandard(M1_final))
qqline(rstandard(M1_final), col = 2)
```

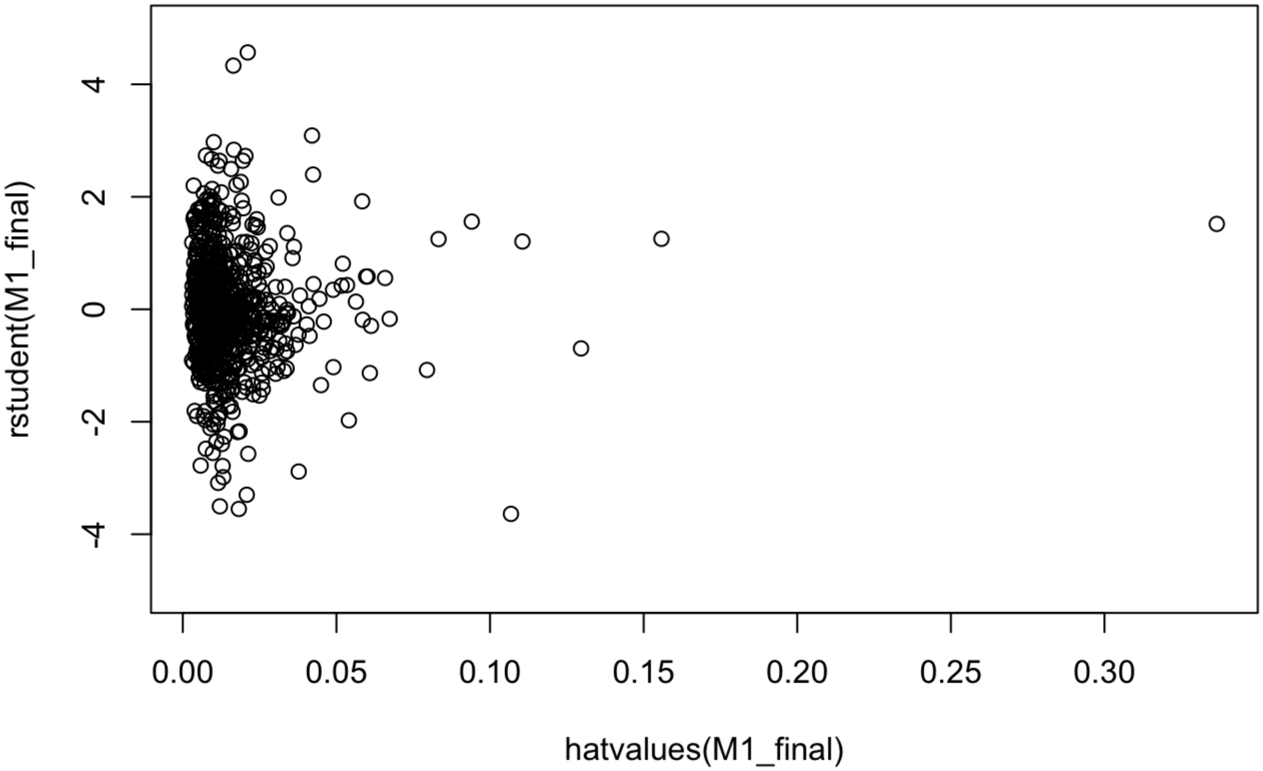


Answer: I would say that the assumption of normality is ‘somewhat’ met. The majority of the points are close to the line indicating normal distribution of errors. That said, beyond the -2 and 2 theoretical quantiles, points are quite far off the line.

Problem 1 j)

j) Are there any outliers or Influential Points? Compute appropriate statistics. [1 pt answer, 1 pt R code for statistics = 2 pts]

```
# plot of deleted studentized residuals vs hat values
plot(rstudent(M1_final)~hatvalues(M1_final),ylim=c(-5, 5))
```



```
#rstudent(M1_final)

# print out only observations that may be influential
summary(influence.measures(M1_final))
```

```
## Potentially influential observations of
##   lm(formula = gradrate ~ private + acceptpct + elite + fundergrad +      pundergrad
+ outstate + roomboardcosts + personal + phd +      percalumni + expend, data = myd) :
##
##     dfb.1_ dfb.prvt dfb.accp dfb.elit dfb.fndr dfb.pndr dfb.otst dfb.rmbr
## 5    0.01  -0.16     0.01     0.03     0.08    -0.02     0.13     0.02
## 17   0.00   0.00     0.01     0.00     0.00     0.00     0.00     0.00
## 21   0.01   0.02    -0.06     0.09     0.02     0.00     0.04     0.06
## 24   0.02  -0.02    -0.02    -0.01    -0.04    -0.05    -0.01    -0.01
## 38   0.00   0.00     0.00    -0.01     0.00     0.00     0.00    -0.01
## 48  -0.23   0.17     0.10     0.10    -0.02    -0.02    -0.36     0.19
## 67  -0.02  -0.16     0.00    -0.04     0.01    -0.02     0.10     0.11
## 70   0.05  -0.74     0.08     0.10    -1.06_*   0.33     0.60    -0.09
## 96   0.17  -0.09     0.13     0.06     0.12    -0.01    -0.01     0.20
## 99  -0.04   0.06    -0.07    -0.01     0.00    -0.03    -0.02    -0.14
## 101 -0.08   0.04     0.02     0.02    -0.02     0.01    -0.01    -0.10
## 107  0.05  -0.01    -0.05    -0.08     0.01    -0.02     0.04    -0.01
## 114 -0.22  -0.23     0.24     0.08    -0.02     0.03     0.26     0.15
## 127  0.07   0.00    -0.02     0.01     0.00     0.08    -0.02     0.05
## 170 -0.03   0.05     0.06     0.02    -0.02    -0.04     0.09    -0.15
## 198 -0.13   0.13     0.03     0.02     0.09     0.01    -0.05     0.13
## 199 -0.03  -0.05    -0.02    -0.01    -0.01     0.04    -0.02     0.00
## 202  0.04  -0.01    -0.05     0.03    -0.15     0.41     0.07    -0.14
## 216 -0.02  -0.09    -0.03    -0.03    -0.02     0.00     0.06    -0.02
## 224  0.00   0.00     0.01     0.00     0.03    -0.06    -0.01     0.02
## 239  0.23   0.20    -0.16     0.32     0.04     0.02    -0.14    -0.09
## 251  0.01   0.00    -0.04    -0.01     0.01    -0.01    -0.03     0.00
## 265 -0.47  -0.11     0.44     0.12    -0.03     0.06    -0.04     0.17
## 266  0.05   0.04     0.06     0.07     0.01     0.02     0.03     0.02
## 273 -0.09   0.08     0.00     0.01    -0.02    -0.04    -0.05     0.09
## 275  0.00   0.00     0.00     0.00    -0.02     0.01     0.00     0.00
## 276 -0.20  -0.08     0.17     0.03    -0.05     0.00    -0.03     0.07
## 285 -0.02  -0.02     0.05    -0.07    -0.06     0.04    -0.18     0.01
## 318 -0.02   0.12    -0.21    -0.01    -0.02    -0.14     0.07     0.09
## 320  0.23   0.16    -0.22    -0.07     0.02    -0.04    -0.21    -0.04
## 355  0.01   0.00    -0.02     0.00     0.00    -0.01     0.00    -0.01
## 367  0.00   0.01     0.01    -0.01     0.06     0.00     0.01    -0.02
## 369 -0.01   0.00     0.01     0.00     0.00     0.00     0.00     0.00
## 378  0.24  -0.24     0.09     0.10    -0.14     0.18    -0.03    -0.06
## 379 -0.12  -0.01     0.04    -0.04    -0.05     0.02     0.02    -0.14
## 385 -0.03  -0.04    -0.03     0.01     0.03     0.06    -0.03     0.07
## 395 -0.12   0.08    -0.04    -0.02    -0.04    -0.03    -0.04     0.00
## 419  0.07  -0.06    -0.02    -0.03     0.02    -0.27     0.00    -0.09
## 427 -0.04   0.02    -0.04    -0.04    -0.02    -0.02     0.00    -0.10
## 431  0.02   0.00    -0.01    -0.03     0.01     0.00    -0.01     0.00
## 446 -0.05  -0.04     0.07     0.06    -0.25     0.11    -0.06     0.03
## 460  0.00   0.00     0.00     0.00     0.00     0.00     0.00     0.00
## 462  0.00   0.00     0.00     0.00     0.01     0.00     0.00     0.00
## 498 -0.17   0.04     0.05    -0.05    -0.04    -0.05     0.00     0.03
## 499 -0.04   0.00     0.03    -0.04    -0.01    -0.03     0.02     0.00
## 507  0.15   0.02    -0.02    -0.01     0.03    -0.01     0.08    -0.07
## 543 -0.01  -0.02     0.03     0.01     0.00     0.01     0.02    -0.01
## 582  0.00  -0.01     0.00     0.01    -0.04     0.01     0.01     0.00
## 586 -0.11   0.23     0.02    -0.04     0.13    -0.18    -0.16     0.06
## 591  0.02   0.00    -0.03    -0.04     0.01    -0.01     0.01    -0.01
## 606  0.00   0.00     0.00    -0.01    -0.01     0.00     0.00     0.00
## 610 -0.01   0.00     0.01     0.01    -0.01     0.00    -0.02     0.00
## 620  0.01  -0.01    -0.01    -0.03    -0.03     0.00     0.01     0.00
## 624 -0.03   0.02     0.04     0.08     0.11    -0.05    -0.01     0.02
## 638  0.00  -0.01     0.01     0.02     0.03    -0.01     0.02     0.00
## 641 -0.03   0.05     0.02     0.05    -0.30     1.04_*   0.02    -0.13
## 645 -0.01  -0.01     0.00     0.00    -0.02     0.02     0.01     0.00
## 677  0.01   0.01    -0.01     0.00    -0.03     0.13     0.00    -0.02
## 685  0.00   0.00    -0.01    -0.02     0.03    -0.08    -0.01     0.02
## 686 -0.01   0.04    -0.01    -0.01     0.08     0.00    -0.01    -0.01
## 688 -0.02   0.01     0.01     0.00    -0.01     0.00    -0.02     0.01
## 692  0.02   0.00    -0.02    -0.01     0.01    -0.04     0.00     0.00
## 701  0.00   0.00     0.00     0.00     0.02    -0.01     0.00     0.00
## 721  0.03   0.04    -0.01    -0.01    -0.02     0.00    -0.15    -0.07
## 729 -0.09  -0.01     0.14     0.01    -0.04     0.04    -0.11    -0.03
## 732  0.25   0.17    -0.27    -0.03     0.00    -0.05    -0.03    -0.13
## 736  0.03  -0.02     0.03    -0.01     0.03    -0.01    -0.03     0.07
## 766  0.01   0.13     0.00     0.01    -0.01    -0.03    -0.04    -0.08
## 776  0.01  -0.01    -0.03    -0.02     0.00    -0.01    -0.03     0.00
## 777  0.08   0.22    -0.15    -0.01     0.01     0.08    -0.22    -0.01
##     dfb.prsn dfb.phd dfb.prcl dfb.expn dffit   cov.r   cook.d hat
## 5    0.01    -0.15    0.21    -0.13    -0.34    0.90_*  0.01   0.01
## 17   0.00     0.00   -0.01     0.00    -0.01    1.05_*  0.00   0.03
## 21   0.03     0.04    0.01    -0.26    -0.27    1.16_*  0.01   0.13_*
## 24   0.01     0.00    0.01     0.02    -0.09    1.05_*  0.00   0.04
## 38   0.00     0.00    0.00     0.01    -0.02    1.05_*  0.00   0.03
## 48   0.05     0.31    0.03    -0.03    -0.47_*  1.01    0.02   0.05_*
## 67  -0.05    -0.08    0.07    -0.01    -0.27    0.94_*  0.01   0.01
## 70   0.19    -0.03   -0.38    -0.07    -1.26_*  0.93_*  0.13   0.11_*
## 96  -0.20    -0.56    0.04     0.08     0.67_*  0.75_*  0.04   0.02
## 99   0.08     0.24    0.01    -0.04    -0.33    0.89_*  0.01   0.01
## 101  0.04     0.19    0.05    -0.06    -0.23    1.05_*  0.00   0.05_*
## 107  0.00    -0.03   -0.06     0.01    -0.12    1.05_*  0.00   0.04
## 114  0.08    -0.12   -0.13    -0.06    -0.48_*  0.88_*  0.02   0.02
## 127 -0.01    -0.17    0.15    -0.04     0.26    0.92_*  0.01   0.01
## 170  0.21     0.00   -0.05    -0.02     0.31    0.94_*  0.01   0.02
## 198  0.10    -0.03    0.08    -0.03    -0.27    0.95_*  0.01   0.01
## 199 -0.04     0.07    0.12    -0.02    -0.21    0.91_*  0.00   0.01
## 202  0.09     0.01    0.07    -0.04     0.48_*  1.02    0.02   0.06_*
## 216  0.13    -0.04    0.09     0.02    -0.22    0.93_*  0.00   0.01
## 224 -0.02    -0.01    0.00     0.00    -0.08    1.08_*  0.00   0.06_*
## 239 -0.16     0.04   -0.06    -0.15     0.50_*  0.97    0.02   0.04
## 251  0.02    -0.01    0.02     0.07     0.10    1.07_*  0.00   0.05_*
## 265 -0.06     0.23    0.15    -0.06    -0.57_*  0.93_*  0.03   0.04
## 266  0.01    -0.15   -0.12    -0.02     0.27    0.93_*  0.01   0.01
## 273  0.24     0.03    0.04    -0.07     0.29    0.92_*  0.01   0.01
## 275  0.00     0.01    0.00     0.00    -0.02    1.05_*  0.00   0.04
## 276  0.05     0.10    0.05     0.04    -0.25    0.94_*  0.01   0.01
## 285 -0.06    -0.05   -0.03     0.51     0.54_*  1.17_*  0.02   0.16_*
## 318  0.55     0.03   -0.06    -0.23     0.65_*  0.91_*  0.03   0.04
## 320 -0.09    -0.02   -0.02     0.12     0.37    0.93_*  0.01   0.02
## 355  0.01     0.00   -0.01     0.05     0.07    1.05_*  0.00   0.03
## 367 -0.03     0.00   -0.02     0.00     0.08    1.07_*  0.00    0.05_*
## 369  0.03    -0.01    0.01     0.00     0.03    1.08_*  0.00    0.06_*
## 378 -0.34    -0.17   -0.04     0.06     0.56_*  0.77_*  0.03    0.02
## 379  0.05     0.22    0.00     0.07    -0.32    0.91_*  0.01    0.01
## 385 -0.22     0.13    0.14    -0.14    -0.38_*  0.94_*  0.01    0.02
## 395 -0.02     0.36   -0.30     0.03    -0.48_*  0.85_*  0.02    0.02
## 419  0.01     0.04   -0.01     0.04    -0.32    1.08_*  0.01    0.08_*
## 427  0.08     0.15   -0.06     0.07    -0.25    0.93_*  0.01    0.01
## 431 -0.03    -0.01    0.00     0.02    -0.05    1.08_*  0.00    0.06_*
## 446 -0.03     0.07   -0.01     0.04    -0.29    1.06_*  0.01    0.06_*
## 460  0.00     0.00    0.00     0.00     0.00    1.05_*  0.00    0.03
## 462  0.00     0.00    0.00     0.00     0.01    1.06_*  0.00    0.04
## 498  0.40     0.04    0.01     0.07     0.42_*  1.12_*  0.02    0.11_*
## 499  0.05     0.01    0.03     0.02     0.13    0.95_*  0.00    0.00
## 507 -0.13    -0.15   -0.16     0.05     0.30    0.89_*  0.01    0.01
## 543 -0.02    -0.01    0.00     0.00    -0.05    1.05_*  0.00    0.03
## 582  0.00     0.00   -0.01     0.00    -0.05    1.09_*  0.00    0.07_*
## 586 -0.03     0.11   -0.09     0.14    -0.39_*  0.85_*  0.01    0.01
## 591  0.00    -0.01   -0.01     0.01    -0.05    1.05_*  0.00    0.03
## 606  0.00     0.00    0.00     0.00    -0.01    1.05_*  0.00    0.03
## 610  0.00     0.00   -0.01     0.08     0.09    1.06_*  0.00    0.04
## 620  0.01     0.00    0.00     0.00    -0.05    1.06_*  0.00    0.05
## 624  0.00    -0.01   -0.01    -0.03     0.15    1.08_*  0.00    0.06_*
## 638  0.00    -0.01    0.00    -0.01     0.04    1.06_*  0.00    0.04
## 641 -0.03    -0.07    0.22     0.08     1.08_*  1.48_*  0.10    0.34_*
## 645  0.04     0.00    0.01    -0.01     0.05    1.06_*  0.00    0.04
## 677 -0.01     0.00    0.00     0.01     0.15    1.08_*  0.00    0.07_*
## 685 -0.01     0.00    0.00     0.01    -0.10    1.06_*  0.00    0.04
## 686  0.02     0.00    0.00    -0.01     0.10    1.07_*  0.00    0.05_*
## 688  0.01     0.03    0.01    -0.01    -0.04    1.05_*  0.00    0.03
## 692 -0.01    -0.01    0.00     0.00    -0.06    1.06_*  0.00    0.04
## 701  0.00     0.00    0.00     0.00     0.02    1.05_*  0.00    0.03
## 721 -0.03     0.01   -0.01     0.33     0.38_*  1.08_*  0.01    0.08_*
## 729 -0.02    -0.05   -0.06     0.45     0.50_*  1.08_*  0.02    0.09_*
## 732  0.08     0.00   -0.14    -0.04     0.39_*  0.92_*  0.01    0.02
## 736 -0.03    -0.14   -0.02     0.10     0.19    1.06_*  0.00    0.05_*
## 766  0.02     0.09   -0.14     0.02     0.24    0.91_*  0.00    0.01
## 776  0.03    -0.02    0.02     0.11     0.15    1.07_*  0.00    0.06_*
## 777 -0.03     0.13    0.11    -0.06     0.37    0.91_*  0.01    0.02
```

Answer: Yes there appear to be a number of potentially influenciel points / outliers. Specifically, at least 5 points fall outside of the (-3,3) bound for studentized residuals. Hat values are within acceptable ranges, as well as cook's distances for the most part. The summary of these measures identified a number of other points, but at lower 'tolerances'/significant levels. So, there are a minimum of 5 points that are likely outliers, and at least influential points, though likely more.

Problem 1 k)

k) Analyze the R2 value for the final model and discuss how well the model explains the variation in graduation rates among the universities. [1 pt R2 value, 1 pt answer = 2 pts]

```
#summary of the model for reference in the answer below
summary(M1_final)
```

```
##
## Call:
## lm(formula = gradrate ~ private + acceptpct + elite + fundergrad +
##     pundergrad + outstate + roomboardcosts + personal + phd +
##     percalumni + expend, data = myd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.085  -6.932  -0.775   7.325  57.598
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.840e+01  4.621e+00  10.475  < 2e-16 ***
## private         4.770e+00  1.689e+00   2.824  0.00486 **
## acceptpct      -1.778e+01  3.797e+00  -4.683 3.34e-06 ***
## elite           4.022e+00  2.002e+00   2.009  0.04492 *
## fundergrad      6.631e-04  1.411e-04   4.699 3.10e-06 ***
## pundergrad     -1.963e-03  3.901e-04  -5.031 6.09e-07 ***
## outstate        1.215e-03  2.270e-04   5.352 1.15e-07 ***
## roomboardcosts  1.534e-03  5.878e-04   2.610  0.00924 **
## personal       -1.820e-03  7.638e-04  -2.383  0.01742 *
## phd             8.424e-02  3.706e-02   2.273  0.02329 *
## percalumni      3.060e-01  4.806e-02   6.367 3.32e-10 ***
## expend         -4.465e-04  1.390e-04  -3.211  0.00138 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.91 on 765 degrees of freedom
## Multiple R-squared:  0.4431, Adjusted R-squared:  0.4351
## F-statistic: 55.33 on 11 and 765 DF,  p-value: < 2.2e-16
```

Answer: The coefficient of determination R2 represents the amount of variation in Y explained by the regression model. So for this analysis, the M1_final R2 is about 0.4431, which means that about 44% of the variation in graduation rates at the universities is explained by the relationship with the variables used in the model. The Adj-R2 further explains this relationship and provides a metric that is not improved merely by adding terms. The Adj-R2 for this model is 0.4351, which also shows about 44% of the variation explained by the model.

In terms of the goodness of fit of the model, we see that the F-statistic is 55.33 with p-value of 2.2e-16 < 0.001. Therefore, we can reject the null hypothesis that there is no significant effect of the model's included variables on graduation rates. So we can accept the alternative hypothesis (at least one of the Beta parameters is not 0) and conclude that there is at least one variable that has a significant effect on balance.

Problem 1 l)

l) Draw conclusions on graduation rates based on your regression analysis.

GradRate = 4.840e+01(Intercept) + 4.770e+00(private) - 1.778e+01(acceptpct) + 4.022e+00(elite) + 6.631e-04(fundergrad) - 1.963e-03(pundergrad) + 1.215e-03(outstate) + 1.534e-03(roomboardcosts) - 1.820e-03(personal) + 8.424e-02(phd) + 3.060e-01(percalumni) - 4.465e-04(expend)

Outstate (0.571), perc.alumni (0.491), and Room.Board (0.425)

a. What are the most important predictors in your model? The most important predictors in the model are the variables that have the greatest impact on the value of graduation rate in the linear model. Specifically, the variables with the greatest beta values are the most important predictors. In my final model, the top three predictors were: acceptpct(- 1.778e+01), private(+ 4.770e+00), and elite(4.022e+00). For acceptpct, the impact on graduation rate is negative. So when a university accepts a higher percentage of students who apply (maybe it's easier to get in), the graduation rate of those schools will tend to trend down as a result of this variable. For private, this again makes sense, as you might expect private schools to have higher graduation rates than public schools. Finally, for elite, which meant that the majority of students came from the top 10% of students in their high school, the impact on graduation rate in the model is again positive, as one might expect. Similar logic applies to determining the impact of the other predictors in the model.

b. Does your model show a significant difference in graduation rates between private and public universities? Yes, there is a significant difference in graduation rates between private and public universities. The test of signicance is not necessary in this case as we can tell from the boxplots that private schools have higher Q1,Q3, and Median values for Graduation Rates compared to public universities. The beta value of 4.77 in the regression model for the private variable indicates that the relationship is positive (going to a private school has a positive impact on graduation rates), but it does not necessarily mean that going to a public school would have a negative impact, the impact would just be smaller or negative depending on the model. Another way of saying this is that private schools in general may have graduation rates that are about 4.77% higher.

c. Do "elite" universities have higher graduation rates? [1 pt conclusion, 1 pt predictors, 1 pt significant difference, 1 pt answer = 4 pts] Yes, elite universities do generally have higher graduation rates, or at least going to an "elite" increases graduation rates compared to not going to an elite university. Based on the regression model beta value of 4.02 for the elite variable, it shows that an increase to 1 from 0 (elite from not elite - qualitative) increases the graduation rate in general by about 4%. The boxplots also confirm this line of reasoning. See boxplots in part 1 c) for reference.

Part 2: Only for Graduate Students

Interaction Terms [8 pts]:

Part 2 a)

a) You are asked to build a new regression model that includes the following independent variables: Elite10, Accept.pct, Outstate, perc.alumni and Expend, together with the interaction effects of elite10 with each independent variable. Fit the model and analyze if the interaction terms are significant. [1 pt fitted regression model with R code, 1 pt answer = 2 pts]

```
# load in the data from file
#myd=read.csv("college.csv", header=T)
#myd[1,]
#create dummy variable for Private;
#attach(myd)
d1 = (myd$Elite10==1)*1 #when it is elite d1
#myd=cbind(myd,d1)


#fit model with main effects and interaction terms only.
new_M <- lm (gradrate ~ (acceptpct + outstate + percalumni + expend + d1)^2, data=myd)
summary(new_M)
```

```
##
## Call:
## lm(formula = gradrate ~ (acceptpct + outstate + percalumni +
##     expend + d1)^2, data = myd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.384  -7.793   0.274   7.560  57.038
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.698e+01  9.094e+00   5.166 3.05e-07 ***
## acceptpct            -1.573e+01  9.930e+00  -1.584 0.113673
## outstate              3.799e-03  1.355e-03   2.804 0.005183 **
## percalumni            5.232e-01  3.704e-01   1.413 0.158180
## expend               -1.436e-03  9.098e-04  -1.579 0.114866
## d1                    3.524e+01  1.057e+01   3.333 0.000901 ***
## acceptpct:outstate   -1.083e-03  1.461e-03  -0.741 0.458660
## acceptpct:percalumni -1.732e-01  3.892e-01  -0.445 0.656448
## acceptpct:expend      1.480e-03  9.472e-04   1.563 0.118541
## acceptpct:d1         -2.876e+01  1.229e+01  -2.340 0.019526 *
## outstate:percalumni  -5.319e-06  1.432e-05  -0.371 0.710419
## outstate:expend      -6.064e-08  3.942e-08  -1.538 0.124350
## outstate:d1          -2.267e-03  6.272e-04  -3.614 0.000321 ***
## percalumni:expend     2.098e-06  1.536e-05   0.137 0.891420
## percalumni:d1        -1.428e-01  1.716e-01  -0.832 0.405469
## expend:d1             1.821e-03  4.763e-04   3.824 0.000142 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.12 on 761 degrees of freedom
## Multiple R-squared:  0.4281, Adjusted R-squared:  0.4169
## F-statistic: 37.98 on 15 and 761 DF,  p-value: < 2.2e-16
```

```
#backward selection
#step(new_M, direction = "backward", trace=F)
```

Answer: Yes, 5 of the "terms" are significant. This is indicated in the Pr(>|t|) column. The significant terms are outstate, d1, acceptpct:d1, outstate:d1, and expend:d1. I keep the simplified ":" notation for clarity. The rest of the terms are insignificant and can be removed.

Part 2 b)

b) Simplify the model and remove interaction terms and additive terms that are not significant. Remember that additive terms included in interaction terms should not be removed. Write down the expression of the final model M2. [1 pt simplified model, 1 pt expression = 2 pts]

```
#simplified M2
M2 <- lm(gradrate ~ outstate + d1 + acceptpct:d1 + outstate:d1 + expend:d1)
summary(M2)
```

```
##
## Call:
## lm(formula = gradrate ~ outstate + d1 + acceptpct:d1 + outstate:d1 +
##     expend:d1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.763  -8.250   0.102   7.986  55.733
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.069e+01  1.506e+00  27.024  < 2e-16 ***
## outstate      2.300e-03  1.427e-04  16.116  < 2e-16 ***
## d1            5.399e+01  9.625e+00   5.610 2.83e-08 ***
## d1:acceptpct -4.060e+01  9.196e+00  -4.415 1.15e-05 ***
## outstate:d1  -1.684e-03  4.989e-04  -3.375 0.000776 ***
## d1:expend     7.468e-05  2.102e-04   0.355 0.722474
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.72 on 771 degrees of freedom
## Multiple R-squared:  0.3657, Adjusted R-squared:  0.3616
## F-statistic: 88.92 on 5 and 771 DF,  p-value: < 2.2e-16
```

Answer: From Simplified Model M2:

GradRate = 4.069e+01(Intercept) + 2.300e-03(outstate) + 5.399e+01(d1) - 4.060e+01(acceptpct:d1) - 1.684e-03(outstate:d1) + 7.468e-05(expend:d1)

Part 2 c)

c) Analyze the parameter estimates of the fitted model and discuss how being an "Elite10" University affects the relationship between Graduation Rates and the four predictors Accept.pct, Outstate, perc.alumni and Expend. [1 pt R code, 1 pt analysis, 2 pts answer = 4 pts] Answer: I will go through each term in the model and describe what they mean.

- 2.300e-03(outstate) : Universities with out of state state tuition will generally experience higher graduation rates.
- 5.399e+01(d1) : Elite10 universities will have higher graduation rates generally compared to non-elite10.
- -4.060e+01(acceptpct:d1) : Even if you’re an Elite10 University, if you accept a higher percentage of applicants, graduation rates will in generally decrease in relation to this. This acceptpct with elite10 is the strongest predictor of graduation rate for the model.
- 1.684e-03(outstate:d1) : Elite10 Universities that have out of state tuition see a positive impact on graduation rate.
- 7.468e-05(expend:d1) : Elite10 Universities that spend more on instruction per student see a small increase in graduation rate.

OPTIONAL: Cross-validation [+6 extra credit pts]:

Part 2 d)

d) Apply cross-validation techniques (5-fold cross validation or divide dataset into a training and a testing set) to compute how well your final model M1 in 1 predicts new data. Compute the MAPE (mean absolute percentage error) statistic and discuss the results. [1 pt cross-validation, 1 pt MAPE, 1 pt answer = 3 pts]

```
myd2=read.csv("college.csv", header=T)

myd2[1,]
```

```
##                          school Private Accept.pct Elite10 F.Undergrad
## 1 Abilene Christian University     Yes  0.7421687       0        2885
##   P.Undergrad Outstate Room.Board Books Personal PhD Terminal S.F.Ratio
## 1         537     7440       3300   450     2200  70       78      18.1
##   perc.alumni Expend Grad.Rate
## 1          12   7041        60
```

```
attach(myd2)
```

```
## The following objects are masked from myd:
##
##     Accept.pct, Books, Elite10, Expend, F.Undergrad, Grad.Rate,
##     Outstate, P.Undergrad, perc.alumni, Personal, PhD, Private,
##     Room.Board, S.F.Ratio, school, Terminal
```

```
#create dummy variable for Private;
Private=(Private=='Yes')*1
myd2=cbind(myd2, Private)

# get the variables
#school = myd$school #IGNORE SCHOOL
Private = myd2$Private
Accept.pct = myd2$Accept.pct
Elite10 = myd2$Elite10
F.Undergrad  = myd2$F.Undergrad
P.Undergrad = myd2$P.Undergrad
Outstate = myd2$Outstate
Room.Board = myd2$Room.Board
Books = myd2$Books
Personal = myd2$Personal
PhD = myd2$PhD
Terminal = myd2$Terminal
S.F.Ratio = myd2$S.F.Ratio
perc.alumni = myd2$perc.alumni
Expend = myd2$Expend

Grad.Rate = myd2$Grad.Rate

# Create training and testing set
# split samples (75% for training and 25% for testing)
select.myd2 <- sample(1:nrow(myd2), 0.75*nrow(myd2))
train.myd2 <- myd2[select.myd2,] #Selecting 75% of the data for trainingpurpose
test.myd2 <- myd2[-select.myd2,] #Selecting 25% (remaining) of the data for testing purp
ose

# fit model using training set
# Model 1:
# Selected model
fit = lm(formula = Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad + P.Undergra
d + Outstate + Room.Board + Personal + PhD + perc.alumni + Expend, data=train.myd2)
summary(fit)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad +
##     P.Undergrad + Outstate + Room.Board + Personal + PhD + perc.alumni +
##     Expend, data = train.myd2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.204  -7.015  -0.125   7.859  55.860
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.781e+01  5.156e+00   9.272  < 2e-16 ***
## PrivateYes   5.754e+00  1.861e+00   3.092 0.002083 **
## Accept.pct  -2.057e+01  4.368e+00  -4.709 3.12e-06 ***
## Elite10      4.491e+00  2.249e+00   1.997 0.046335 *
## F.Undergrad  6.572e-04  1.632e-04   4.028 6.39e-05 ***
## P.Undergrad -1.861e-03  4.388e-04  -4.241 2.59e-05 ***
## Outstate     1.206e-03  2.594e-04   4.648 4.16e-06 ***
## Room.Board   1.196e-03  6.677e-04   1.791 0.073794 .
## Personal    -1.666e-03  8.268e-04  -2.015 0.044369 *
## PhD          1.497e-01  4.154e-02   3.603 0.000342 ***
## perc.alumni  2.523e-01  5.450e-02   4.630 4.53e-06 ***
## Expend      -5.081e-04  1.773e-04  -2.865 0.004327 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.52 on 570 degrees of freedom
## Multiple R-squared:  0.4468, Adjusted R-squared:  0.4361
## F-statistic: 41.84 on 11 and 570 DF,  p-value: < 2.2e-16
```

```
#Create fitted values using test.myd data
y_pred <- predict.glm(fit, test.myd2)
y_obs<-test.myd2[,"Grad.Rate"]
# Compute RMSE of prediction errors
rmse_m1 <- sqrt((y_obs - y_pred)%*%(y_obs - y_pred)/nrow(test.myd2))
rmse_m1
```

```
##          [,1]
## [1,] 14.19436
```

```
# Compute mean absolute error
mae_m1<-mean(abs(y_obs - y_pred))
mae_m1
```

```
## [1] 10.319
```

```
# Compute mean percentage absolute error
mape_m1<-mean(abs((y_obs - y_pred)/y_obs))*100
mape_m1
```

```
## [1] 18.78621
```

```
# compute cross-validated R^2_pred
r2_pred = cor(cbind(y_obs,y_pred))**2
r2_train = summary(fit)$r.squared
diffr2_m1=abs(r2_train-r2_pred)
#print difference of cross-validate R2 and R2
diffr2_m1[1,2]
```

```
## [1] 0.02619838
```

Answer: Validation statistics for Model M1:

RMSE = 12.63

MAE = 9.51

MAPE = 15.6

Difference between R2 of fitted model and testing set R2 = 0.010

Model 1 minimizes all three validation metrics, and we can conclude that it provides more accurate predictions (closer to actual values). The MAPE value for M1 indicates that on average predictions are off by about 16% of the actual value.

The values of cross-validated R2 value are very close to the model R2 value indicating that the model is not over-fitting.

Part 2 e)

e) Apply the same cross-validation procedure and compute the MAPE statistic for the interaction model M2 computed in Part 2. Compare the predictive power of the models M1 and M2 fitted in Part 1 and Part 2. [1 pt cross-validation, 1 pt MAPE, 1 pt answer = 3 pts]

```
# fit model using training set
# Model 2:
# Selected model
fit2 = lm(formula = Grad.Rate ~ Outstate + Elite10 + Accept.pct:Elite10 + Outstate:Elite
10 + Expend:Elite10, data=train.myd2)
summary(fit2)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Outstate + Elite10 + Accept.pct:Elite10 +
##     Outstate:Elite10 + Expend:Elite10, data = train.myd2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -48.626  -7.804  -0.186   8.176  50.208
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.135e+01  1.698e+00  24.351  < 2e-16 ***
## Outstate            2.198e-03  1.598e-04  13.752  < 2e-16 ***
## Elite10             5.234e+01  1.039e+01   5.037 6.34e-07 ***
## Elite10:Accept.pct -4.181e+01  1.010e+01  -4.140 4.00e-05 ***
## Outstate:Elite10   -1.545e-03  5.672e-04  -2.723  0.00666 **
## Elite10:Expend      1.336e-04  2.541e-04   0.526  0.59921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.36 on 576 degrees of freedom
## Multiple R-squared:  0.363,  Adjusted R-squared:  0.3575
## F-statistic: 65.66 on 5 and 576 DF,  p-value: < 2.2e-16
```

```
#Create fitted values using test.myd data
y_pred <- predict.glm(fit2, test.myd2)
y_obs<-test.myd2[,"Grad.Rate"]
# Compute RMSE of prediction errors
rmse_m2 <- sqrt((y_obs - y_pred)%*%(y_obs - y_pred)/nrow(test.myd2))
rmse_m2
```

```
##          [,1]
## [1,] 14.79245
```

```
# Compute mean absolute error
mae_m2<-mean(abs(y_obs - y_pred))
mae_m2
```

```
## [1] 10.81406
```

```
# Compute mean percentage absolute error
mape_m2<-mean(abs((y_obs - y_pred)/y_obs))*100
mape_m2
```

```
## [1] 20.0692
```

```
# compute cross-validated R^2_pred
r2_pred = cor(cbind(y_obs,y_pred))**2
r2_train = summary(fit2)$r.squared
diffr2_m2=abs(r2_train-r2_pred)
#print difference of cross-validate R2 and R2
diffr2_m2[1,2]
```

```
## [1] 0.009038047
```

Answer: Validation statistics for Model M1:

RMSE = 14.01

MAE = 10.51

MAPE = 20.96

Difference between R2 of fitted model and testing set R2 = 0.037

Model 2 has highest values for all three metrics, and we can conclude that it provides less accurate predictions (farther from the actual values compared to M1). The MAPE value for M2 indicates that on average predictions are off by about 21% of the actual value.

The values of cross-validated R2 value are very close to the model R2 value indicating that the model is not over-fitting at least.

Basically it would appear that Model 1 is better than Model 2…

“Reflection” Problem [2 pts]

Post a message in the “Reflection for Assignment 3” thread on the discussion board indicating which question in this assignment you found to be the easiest, the one you found to be the hardest, and why. Alex Teboul Answer:

Easiest Question: Problem 1 part 1 - analyze the distribution. This was the easiest question because it didn’t involve much. It was just generating the histogram and looking at the distribution to see if it was symmetric, which it was.

Hardest Question: Problem Part 2 Optional Extra Credit. Had trouble getting the cross validation to work. Still not sure if the numbers are accurate or if I did it correctly given that we didn’t spend much time going over this technique in class. Kept getting errors with regards to the length of my variables because I think the predict function couldn’t find the variables I used in the “headers” of each column. Either way it was an interesting assignment.