Assignment 3

Erik Pak

August 06, 2024

PROBLEM 1

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
- Books: estimated book costs
- Personal: Estimated personal spending
- PhD: Percent of faculty with PhD degrees
- Terminal: Percent of faculty with terminal degrees
- S.F.Ratio: Student/faculty ratio
- perc.alumni: Percent of alumni who donate
- Expend: Instructional expenditure per student
- Grad.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).

Libraries

```
library(psych)        # used for describe
library(ggplot2)      # used for ggplot
library(ggpubr)       # combine scatter plots
library(QuantPsyc)    # normalize coefficients
library(equatiomatic) # equation for a model
library(car)          # VIF for a model
library(corrplot)     # correlation plot
library(dplyr)        # using slice
library(leaps)        # variable selection
library(DAAG)         # Cross Validation
```

Import text file

```
# set working directory
setwd("~/Downloads/Data")

# header in the college.cvs
college <- read.csv(file = 'college.csv', header = TRUE)

# display using head()
head(college)
```

```
##                         school Private Accept.pct Elite10 F.Undergrad
## 1 Abilene Christian University     Yes  0.7421687       0        2885
## 2           Adelphi University     Yes  0.8801464       0        2683
## 3               Adrian College     Yes  0.7682073       0        1036
## 4          Agnes Scott College     Yes  0.8369305       1         510
## 5    Alaska Pacific University     Yes  0.7564767       0         249
## 6            Albertson College     Yes  0.8160136       0         678
##   P.Undergrad Outstate Room.Board Books Personal PhD Terminal S.F.Ratio
## 1         537     7440       3300   450     2200  70       78      18.1
## 2        1227    12280       6450   750     1500  29       30      12.2
## 3          99    11250       3750   400     1165  53       66      12.9
## 4          63    12960       5450   450      875  92       97       7.7
## 5         869     7560       4120   800     1500  76       72      11.9
## 6          41    13500       3335   500      675  67       73       9.4
##   perc.alumni Expend Grad.Rate
## 1          12   7041        60
## 2          16  10527        56
## 3          30   8735        54
## 4          37  19016        59
## 5           2  10922        15
## 6          11   9727        55
```

Descriptive Statistics

```
# descriptive statistics
describe(college)
```

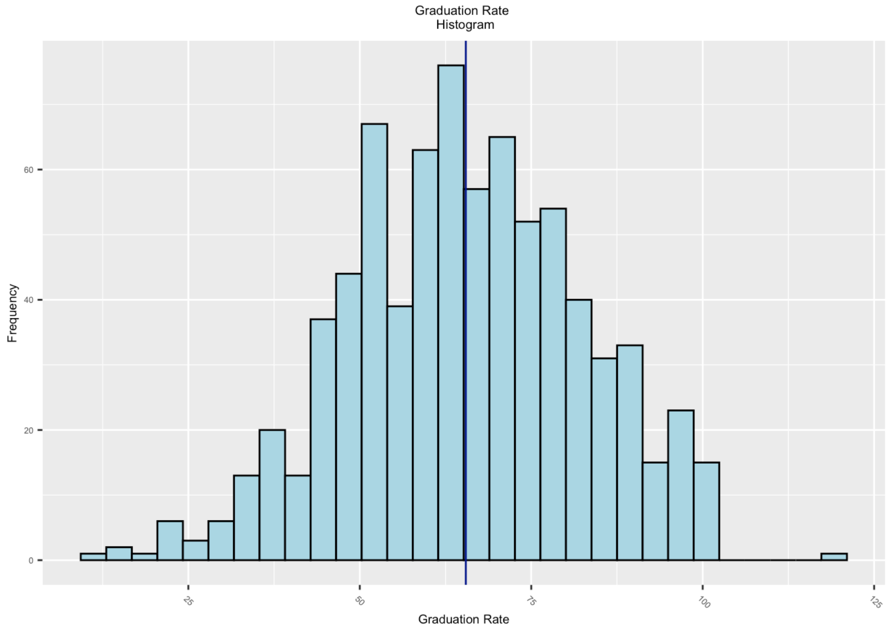
```
##             vars   n     mean      sd  median  trimmed     mad     min     max
## school*        1 777   389.00  224.44  389.00   389.00  287.62    1.00   777.0
## Private*       2 777     1.73    0.45    2.00     1.78    0.00    1.00     2.0
## Accept.pct     3 777     0.75    0.15    0.78     0.76    0.12    0.15     1.0
## Elite10        4 777     0.10    0.30    0.00     0.00    0.00    0.00     1.0
## F.Undergrad    5 777  3699.91 4850.42 1707.00  2574.88 1441.09  139.00 31643.0
## P.Undergrad    6 777   855.30 1522.43  353.00   536.36  449.23    1.00 21836.0
## Outstate       7 777 10440.67 4023.02 9990.00 10181.66 4121.63 2340.00 21700.0
## Room.Board     8 777  4357.53 1096.70 4200.00  4301.70 1005.20 1780.00  8124.0
## Books          9 777   549.38  165.11  500.00   535.22  148.26   96.00  2340.0
## Personal      10 777  1340.64  677.07 1200.00  1268.35  593.04  250.00  6800.0
## PhD           11 777    72.66   16.33   75.00    73.92   17.79    8.00   103.0
## Terminal      12 777    79.70   14.72   82.00    81.10   14.83   24.00   100.0
## S.F.Ratio     13 777    14.09    3.96   13.60    13.94    3.41    2.50    39.8
## perc.alumni   14 777    22.74   12.39   21.00    21.86   13.34    0.00    64.0
## Expend        15 777  9660.17 5221.77 8377.00  8823.70 2730.95 3186.00 56233.0
## Grad.Rate     16 777    65.46   17.18   65.00    65.60   17.79   10.00   118.0
##                range  skew kurtosis     se
## school*       776.00  0.00    -1.20   8.05
## Private*        1.00 -1.02    -0.96   0.02
## Accept.pct      0.85 -1.06     1.24   0.01
## Elite10         1.00  2.65     5.05   0.01
## F.Undergrad 31504.00  2.60     7.61 174.01
## P.Undergrad 21835.00  5.67    54.52  54.62
## Outstate    19360.00  0.51    -0.43 144.32
## Room.Board   6344.00  0.48    -0.20  39.34
## Books        2244.00  3.47    28.06   5.92
## Personal     6550.00  1.74     7.04  24.29
## PhD            95.00 -0.77     0.54   0.59
## Terminal       76.00 -0.81     0.22   0.53
## S.F.Ratio      37.30  0.66     2.52   0.14
## perc.alumni    64.00  0.60    -0.11   0.44
## Expend      53047.00  3.45    18.59 187.33
## Grad.Rate     108.00 -0.11    -0.22   0.62
```

```
# for grad rate
describe(college$Grad.Rate)
```

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

Hisrogram

```
# Grad.Rate histogram
plot_hist_coll <- ggplot(college, aes(x=Grad.Rate)) +
                  geom_histogram(bins = 30, color="black", fill="lightblue") +
                  geom_vline(aes(xintercept=mean(Grad.Rate)), col="darkblue") +
                  labs(title = "Graduation Rate \n Histogram",
                       x = "Graduation Rate", y = 'Frequency') +
                  # move the title text to the middle
                  theme(plot.title=element_text(hjust=0.5)) +
                  theme(text = element_text(size = 6))  +
                  theme(axis.title = element_text(size = 7)) +
                  theme(axis.text.x = element_text(angle = -45, hjust = .1))

plot_hist_coll
```



a. Analyze the distribution of Grad.Rate and discuss if the distribution is symmetric, or if you need to apply any transformation.

The Grad.Rate is symmetric according to the histogram. Also, the mean of 65.46 and median of 65 are almost identical; therefore, there is no need to transform the dataset.

Scatter Plots

```
# scatter plots
plot_acc <- ggplot(college, aes(x = Accept.pct, y = Grad.Rate)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Acceptance Rate (%)",
            x="Acceptance Rate (%)", y = "Graduation Rate") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_fug <- ggplot(college, aes(x = F.Undergrad, y = Grad.Rate)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Full-time Undergrad",
            x="Full-time Undergrad", y = "Graduation Rate") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_pug <- ggplot(college, aes(x = P.Undergrad, y = Grad.Rate)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Part-time Undergrad",
            x="Part-time Undergrad", y = "Graduation Rate") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_out <- ggplot(college, aes(x = Outstate, y = Grad.Rate)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Out of State",
            x="Out of State", y = "Graduation Rate") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_rnb <- ggplot(college, aes(x = Room.Board, y = Grad.Rate)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Room & Board",
            x="Room & Board", y = "Graduation Rate") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_bok <- ggplot(college, aes(x = Books, y = Grad.Rate)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Books",
            x="Books", y = "Graduation Rate") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_per <- ggplot(college, aes(x = Personal, y = Grad.Rate)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Personal Spending",
            x="Personal Spending", y = "Graduation Rate") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_phd <- ggplot(college, aes(x = PhD, y = Grad.Rate)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Phd degrees (%)",
            x="Faculty with terminal degrees (%)", y = "Graduation Rate") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
```

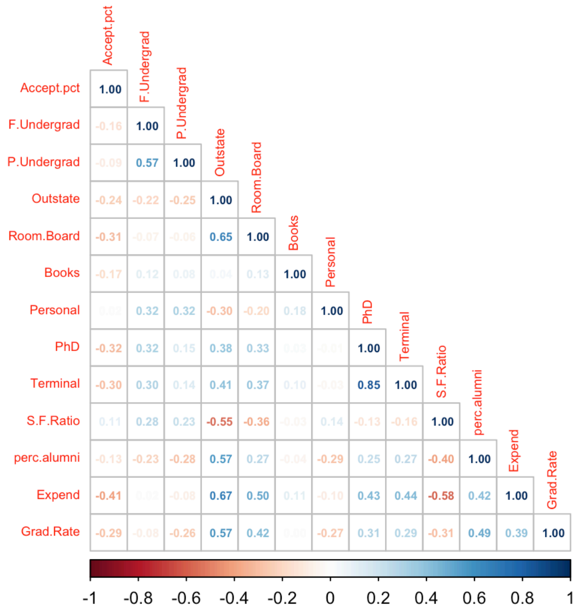
```
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_ter <- ggplot(college, aes(x = Terminal, y = Grad.Rate)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Terminal degree (%)",
            x="Terminal", y = "Graduation Rate") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_sfr <- ggplot(college, aes(x = S.F.Ratio, y = Grad.Rate)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Student/Faculty Ratio",
            x="Student/Faculty Ratio", y = "Graduation Rate") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_pal <- ggplot(college, aes(x = perc.alumni, y = Grad.Rate)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="% Alumni Donate",
            x="% Alumni Donate", y = "Graduation Rate") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))

plot_exp <- ggplot(college, aes(x = Expend, y = Grad.Rate)) +
            geom_point() +
            geom_smooth(method=lm, se=FALSE) +
            labs(title="Instruction \n Expenditure per Student",
            x="Instruction \n Expenditure per Student",
            y = "Graduation Rate") +
            # move the title text to the middle
            theme(plot.title=element_text(hjust=0.5)) +
            theme(text = element_text(size = 6))  +
            theme(axis.title = element_text(size = 7))
```

Dummy Variable

```
# new column college$Elite10 where 1 = Yes & 0 = No for Box plot purpose
# making copy of the variable for Box Plots
college$Private.1 <- ifelse(college$Private == "Yes", "Yes","No")
college$Elite10.1 <- ifelse(college$Elite10 == 1, "Yes","No")
# update column Private where Yes = 1 & No = 0
college$Private   <- ifelse(college$Private == "Yes", 1,0)
```

Correlation

```
# select all the columns except Private & college
college.numeric <- college[, !names(college) %in% c("Private", "school", "Elite10.1", "Private.1", "Elite10")]

# correlation values
corr.college <- cor(college.numeric)

# plot correlation
corr_plot <- corrplot(corr.college, method = 'number', addCoef.col = 'green',
                      type = 'lower', number.cex = 0.55, tl.cex = 0.6,)
```

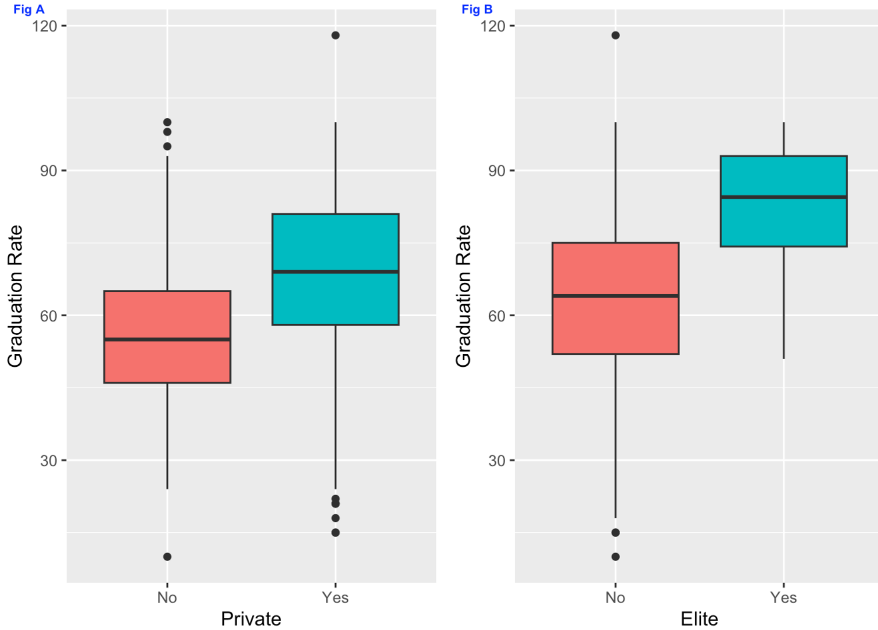


b. Create scatterplots for Grad.Rate vs each of the independent variables. What conclusions can you draw about the relationships between Grad.Rate and the independent variables? (No need to include the scatterplots in your submission, but you can use correlation analysis)

Examining the correlation plot (`PhD & Terminal`) has a relatively strong positive relationship at 0.85 correlation value, and we should consider removing one of the values from our regression model. We have (`P.Undergrad & F.Undergrad`), (`Outstate , Grad.Rate`), (`Outstate , Expend`), (`Outstate & perc.alumni`), (`Outstate & Room.Board`), (`Expend & Room.Board`), and (`Grad.Rate & perc.alumni`) have a moderate positive relationship. We also have (`S.F.Ratio & Outstate`), (`Expend , S.F.Ratio`) with a moderate negative relationship.

There are slightly positive relationships in (`Terminal & Outstate`), (`PhD & Outstate`), (`PhD & Room.Board`), (`PhD & Room.Board`), (`Expend & PhD`), (`Expend & Terminal`), (`Expend & perc.alumni`), (`Expend & Grad.Rate`). Conversely, slightly negative relationships are in (`Expend & Accept.pct`), (`perc.alumni & S.F.Ratio`). The rest are all negligibly correlated, and some are not correlated at all.

```
# Box plot Private & Graduate rate
p_box_private <- ggplot(college, aes(x=Private.1, y=Grad.Rate, fill=Private.1)) +
                 geom_boxplot() + labs(x="Private", y = "Graduation Rate") +
                 theme(legend.position="none")

p_box_elite10 <- ggplot(college, aes(x=Elite10.1, y=Grad.Rate, fill=Elite10.1)) +
                 geom_boxplot() + labs(x="Elite", y = "Graduation Rate") +
                 theme(legend.position="none")

# combine box plots
box_com_plot <- ggarrange(p_box_private, p_box_elite10,
                 labels = c("Fig A", "Fig B"),
                 font.label = list(size = 7, color = "blue"))
# plot all
box_com_plot
```



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.

Fig A: - This box plot provides graduation rates vary by university type (private vs. public), and the median graduation rate in four years is higher in private universities, which tells a different distribution of graduation rates between private and public. The lowest graduation rate, excluding outliers, is about the same. The highest score, excluding outliers, could contain some data errors since some values are higher than a 100% graduation rate and would require investigation. Since the inter-quartile range (IQR) is slightly wider tells us there is more variability in the private school data from the median, and the public school data has a bit longer upper whisker tells us it is slightly positively skewed. The private school has more negativity skewed due to the longer whisker at the bottom. The private school's four-year graduation rate is significant than in public schools.

Fig B: - The range in non-elite has a larger spread from the median since IQR is slightly broader than the elite. Also, in non-elite, we have a top whisker beyond 100% graduation rate, which would require investigation for data error and is also visible in elite. Both box plots are negatively skewed, but the elite will have greater skew due to the difference between the length of the whiskers, including the median visibly bottom-heavy. This box plot is more striking when comparing the graduation rate from elite to non-elite, and the elite school's four-year graduation rate is more significant than that of non-elite schools.

```
# all the variables from the data file
graduation_model <- lm(Grad.Rate ~ Private + Accept.pct + Elite10 +
                        F.Undergrad + P.Undergrad + Outstate + Room.Board + Books +
                        Personal + PhD + Terminal + S.F.Ratio + perc.alumni +
                        Expend, data=college)

# display the actual full model equation
equatiomatic::extract_eq(graduation_model, use_coefs = FALSE)
```

$$\operatorname{Grad.Rate} = \alpha + \beta_{1}(\operatorname{Private}) + \beta_{2}(\operatorname{Accept.pct}) + \beta_{3}(\operatorname{Elite10}) + \beta_{4}(\operatorname{F.Undergrad}) + \beta_{5}(\operatorname{P.Undergrad}) + \beta_{6}(\operatorname{Outstate}) + \beta_{7}(\operatorname{Room.Board}) + \beta_{8}(\operatorname{Books}) + \beta$$

Full Model Info

```
# summary of the model
summary(graduation_model)

# variance analysis
anova(graduation_model)
```

d. Fit a full model (with all independent variables) to predict Grad.Rate.

```
# Variance Inflation Factors
vif.grad_model <- vif(graduation_model)
vif.grad_model
```

```
##     Private  Accept.pct     Elite10 F.Undergrad P.Undergrad    Outstate
##      2.7395      1.4866      1.6879      2.2331      1.6434      3.9351
##  Room.Board       Books    Personal         PhD    Terminal   S.F.Ratio
##      1.9768      1.1158      1.2910      3.9177      3.9466      1.9097
## perc.alumni      Expend
##      1.6724      2.9226
```

e. Does multi-collinearity seem to be a problem here? What is your evidence? Compute and analyze the VIF statistics.

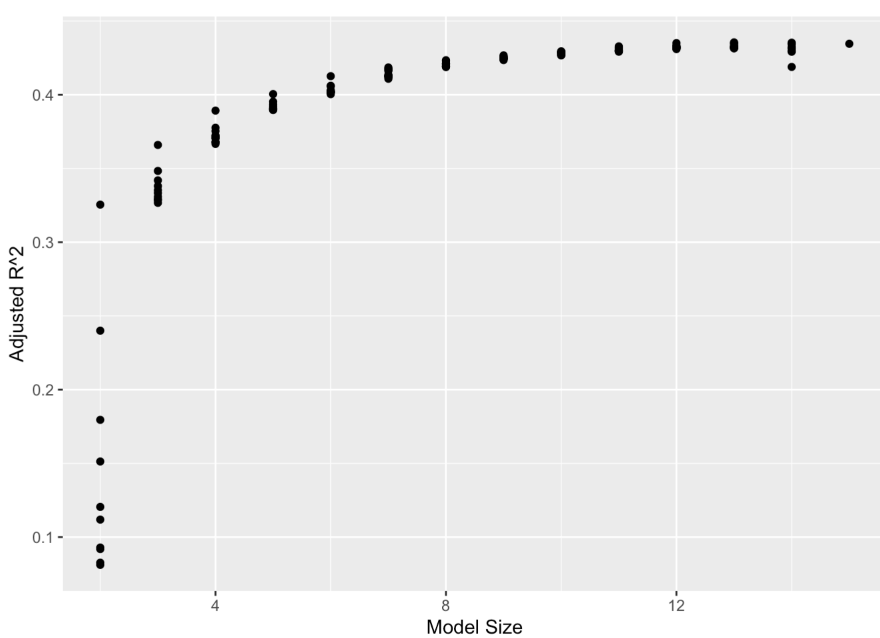
Our model has no multi-collinearity because Variance Inflation Factors analysis for all the variables does not come close to 10.

Variable Selection - adj-R^2

```
# omit missing values. This is necessary to use leaps function.
# not needed for this dataframe
# newcollege <- na.omit(college)

# collect varaibles for the model
xvarsname <- names(college[2:15])
xvars     <- college[2:15]
yvar      <- college[,16]

# new dataframe for variable selection
college_new <- college[,2:16]

# best subset model selection according to Adj-R2 statistics
leapmodels=leaps(x=xvars, y=yvar, names=xvarsname, method="adjr2")
mat=cbind(leapmodels$size,leapmodels$which, leapmodels$adjr2)

# display results in increasing order of adjr2
# first element is # of vars in the model
# values of 1 in row indicates selected variables
# last column shows adjr2 values
head(mat[order(mat[,dim(mat)[2]], decreasing=TRUE),], 5)
```

```
##       Private Accept.pct Elite10 F.Undergrad P.Undergrad Outstate Room.Board
## 12 13       1          1       1           1           1        1          1
## 13 14       1          1       1           1           1        1          1
## 12 13       1          1       1           1           1        1          1
## 11 12       1          1       1           1           1        1          1
## 13 14       1          1       1           1           1        1          1
##    Books Personal PhD Terminal S.F.Ratio perc.alumni Expend
## 12     0        1   1        1         0           1      1 0.4355531
## 13     1        1   1        1         0           1      1 0.4353498
## 12     1        1   1        0         0           1      1 0.4350850
## 11     0        1   1        0         0           1      1 0.4350763
## 13     0        1   1        1         1           1      1 0.4348133
```

```
# plot results on a scatterplot.
# Best model is selected by largest adjr2 values.
plot_lead <- data.frame(size = leapmodels$size, adjr2 = leapmodels$adjr2)
#
ggplot(plot_lead, aes(x = size, y = adjr2)) +
            geom_point() + labs(x="Model Size", y="Adjusted R^2")
```



Variable Selection - backward & both direction

```
# backward selection
# start with the full model "graduation_model"
step(graduation_model, direction = "backward")
```

```
## Start:  AIC=3990.75
## Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad + P.Undergrad +
##     Outstate + Room.Board + Books + Personal + PhD + Terminal +
##     S.F.Ratio + perc.alumni + Expend
##
##               Df Sum of Sq    RSS    AIC
## - S.F.Ratio    1       0.0 127126 3988.8
## - Books        1     120.8 127247 3989.5
## - Terminal     1     226.1 127352 3990.1
## <none>                     127126 3990.8
## - Elite10      1     671.0 127797 3992.8
## - Personal     1     813.4 127939 3993.7
## - PhD          1     901.2 128027 3994.2
## - Private      1    1200.9 128327 3996.1
## - Room.Board   1    1312.3 128438 3996.7
## - Expend       1    1379.4 128505 3997.1
## - Accept.pct   1    3704.3 130830 4011.1
## - F.Undergrad  1    3790.8 130917 4011.6
## - P.Undergrad  1    4185.8 131312 4013.9
## - Outstate     1    4866.1 131992 4017.9
## - perc.alumni  1    6812.2 133938 4029.3
##
## Step:  AIC=3988.75
## Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad + P.Undergrad +
##     Outstate + Room.Board + Books + Personal + PhD + Terminal +
##     perc.alumni + Expend
##
##               Df Sum of Sq    RSS    AIC
## - Books        1     120.8 127247 3987.5
## - Terminal     1     226.3 127352 3988.1
## <none>                     127126 3988.8
## - Elite10      1     671.0 127797 3990.8
## - Personal     1     818.0 127944 3991.7
## - PhD          1     903.9 128030 3992.3
## - Private      1    1227.8 128354 3994.2
## - Room.Board   1    1312.3 128438 3994.7
## - Expend       1    1642.3 128768 3996.7
## - Accept.pct   1    3734.1 130860 4009.2
## - F.Undergrad  1    3854.2 130980 4010.0
## - P.Undergrad  1    4186.5 131313 4011.9
## - Outstate     1    4891.0 132017 4016.1
## - perc.alumni  1    6848.2 133974 4027.5
##
## Step:  AIC=3987.49
## Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad + P.Undergrad +
##     Outstate + Room.Board + Personal + PhD + Terminal + perc.alumni +
##     Expend
##
##               Df Sum of Sq    RSS    AIC
## - Terminal     1     274.2 127521 3987.2
## <none>                     127247 3987.5
## - Elite10      1     664.8 127912 3989.5
## - Personal     1     960.8 128208 3991.3
## - PhD          1    1018.4 128265 3991.7
## - Private      1    1188.8 128436 3992.7
## - Room.Board   1    1254.2 128501 3993.1
## - Expend       1    1672.3 128919 3995.6
## - Accept.pct   1    3625.5 130872 4007.3
## - F.Undergrad  1    3781.9 131029 4008.2
## - P.Undergrad  1    4171.4 131418 4010.6
## - Outstate     1    4937.7 132185 4015.1
## - perc.alumni  1    6929.8 134177 4026.7
##
## Step:  AIC=3987.16
## Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad + P.Undergrad +
##     Outstate + Room.Board + Personal + PhD + perc.alumni + Expend
##
##               Df Sum of Sq    RSS    AIC
## <none>                     127521 3987.2
## - Elite10      1     672.6 128194 3989.3
## - PhD          1     861.3 128382 3990.4
## - Personal     1     946.5 128467 3990.9
## - Room.Board   1    1135.3 128656 3992.1
## - Private      1    1329.5 128851 3993.2
## - Expend       1    1719.0 129240 3995.6
## - Accept.pct   1    3655.7 131177 4007.1
## - F.Undergrad  1    3680.7 131202 4007.3
## - P.Undergrad  1    4219.0 131740 4010.5
```

```
## - Outstate     1    4773.9 132295 4013.7
## - perc.alumni  1    6758.1 134279 4025.3
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad +
##     P.Undergrad + Outstate + Room.Board + Personal + PhD + perc.alumni +
##     Expend, data = college)
##
## Coefficients:
## (Intercept)      Private   Accept.pct      Elite10  F.Undergrad  P.Undergrad
##   4.840e+01    4.770e+00   -1.778e+01    4.022e+00    6.631e-04   -1.963e-03
##    Outstate   Room.Board     Personal          PhD  perc.alumni       Expend
##   1.215e-03    1.534e-03   -1.820e-03    8.424e-02    3.060e-01   -4.465e-04
```

```
#Forward selection
# start with Base model that has no variables:
Base = lm(Grad.Rate ~ 1, data=college_new)

# Stepwise selection
step(Base, scope = list(upper = graduation_model, lower = ~1), direction = "both",
     trace=FALSE)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Outstate + perc.alumni + Accept.pct +
##     P.Undergrad + F.Undergrad + Room.Board + Expend + Personal +
##     Private + PhD + Elite10, data = college_new)
##
## Coefficients:
## (Intercept)     Outstate  perc.alumni   Accept.pct  P.Undergrad  F.Undergrad
##   4.840e+01    1.215e-03    3.060e-01   -1.778e+01   -1.963e-03    6.631e-04
##  Room.Board       Expend     Personal      Private          PhD      Elite10
##   1.534e-03   -4.465e-04   -1.820e-03    4.770e+00    8.424e-02    4.022e+00
```

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.

I've tried the three-variable selection method to determine the necessary variables. My adjR^2 variable selection did not remove the Terminal independent variable but Stepped backward, and both directions stated it should be removed. I selected to remove the Terminal to simplify the model, and having a high correlation with PhD is also an indication to remove this predictor.

Variable selection:

- adjusted R^2
- backward direction
- both direction

Final Model

```
# new model using variable selection method
grad_model_1 <- lm(Grad.Rate ~ Private + Accept.pct + Elite10 +
                     F.Undergrad + P.Undergrad + Outstate + Room.Board +
                     Personal + PhD + perc.alumni + Expend, data=college)

# summary report of the grad_model_1 model
summary(grad_model_1)

# variance report of the grad_model_1 model
anova(grad_model_1)
```

Extra Model: for fun

```
# removing Room.Board due to Analysis of Variance
# because F-value is 0.0725184 which greater than 5% significance level
grad_model_2 <- lm(Grad.Rate ~ Private + Accept.pct + Elite10 +
                     F.Undergrad + P.Undergrad + Outstate + Personal + PhD + perc.alumni +
                     Expend, data=college)

# summary report of the grad_model_2 model
summary(grad_model_2)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad +
##     P.Undergrad + Outstate + Personal + PhD + perc.alumni + Expend,
##     data = college)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.227  -7.301  -0.582   7.203  59.073
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.327e+01  4.244e+00  12.552  < 2e-16 ***
## Private      5.257e+00  1.685e+00   3.120  0.00188 **
## Accept.pct  -1.950e+01  3.754e+00  -5.195 2.62e-07 ***
## Elite10      3.891e+00  2.009e+00   1.936  0.05318 .
## F.Undergrad  6.711e-04  1.416e-04   4.739 2.56e-06 ***
## P.Undergrad -1.872e-03  3.900e-04  -4.799 1.92e-06 ***
## Outstate     1.459e-03  2.076e-04   7.025 4.72e-12 ***
## Personal    -1.959e-03  7.648e-04  -2.562  0.01059 *
## PhD          9.238e-02  3.707e-02   2.492  0.01290 *
## perc.alumni  2.847e-01  4.754e-02   5.989 3.24e-09 ***
## Expend      -4.272e-04  1.394e-04  -3.065  0.00225 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.96 on 766 degrees of freedom
## Multiple R-squared:  0.4381, Adjusted R-squared:  0.4308
## F-statistic: 59.73 on 10 and 766 DF,  p-value: < 2.2e-16
```

```
# variance report of the grad_model_2 model
anova(grad_model_2)
```

```
## Analysis of Variance Table
##
## Response: Grad.Rate
##              Df Sum Sq Mean Sq  F value    Pr(>F)
## Private       1  25876 25875.6 154.0593 < 2.2e-16 ***
## Accept.pct    1  22964 22964.4 136.7266 < 2.2e-16 ***
## Elite10       1   8861  8860.7  52.7550 9.309e-13 ***
## F.Undergrad   1   2521  2521.0  15.0098 0.0001161 ***
## P.Undergrad   1   6894  6894.2  41.0468 2.594e-10 ***
## Outstate      1  22260 22259.6 132.5305 < 2.2e-16 ***
## Personal      1   2133  2133.3  12.7016 0.0003880 ***
## PhD           1   1512  1512.3   9.0037 0.0027817 **
## perc.alumni   1   5721  5721.4  34.0645 7.869e-09 ***
## Expend        1   1578  1578.3   9.3972 0.0022495 **
## Residuals   766 128656   168.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Variance Inflation Factors
vif.grad_model_2 <- vif(grad_model_2)
vif.grad_model_2
```

```
##     Private  Accept.pct     Elite10 F.Undergrad P.Undergrad    Outstate
##      2.6062      1.4088      1.6865      2.1800      1.6291      3.2237
##    Personal         PhD perc.alumni      Expend
##      1.2388      1.6923      1.6034      2.4470
```

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.

I started with a complete model, which gave me Multiple R-squared: 0.4448, Adjusted R-squared: 0.4346. Then ran, multiple variable selections to determine the model's required independent variables. I decided to remove three variables (`Book` , `S.F.Ratio` , `Terminal`) and ran a second model, which gave me Multiple R-squared: 0.4431, Adjusted R-squared: 0.4351.

extra Model: By examing the variance, and noticed that `Room.Board` 's F-value had a p-value of 0.0725184 which is greater than 0.05; therefore, I removed this independent variable and ran a new model and Multiple R-squared: 0.4381, Adjusted R-squared: 0.4308. Finally, in my last model (`grad_model_2`), I did not remove `Elite10` even though the t-value was 0.05318, greater than 0.05, because of the high graduation rate according to the box plot and felt this is an essential factor in the graduation rate.

- Beta0 = Intercept
- x1 = Private
- x2 = Accept.pct
- x3 = Elite10

- x4 = F.Undergrad
- x5 = P.Undergrad
- x6 = Outstate
- x7 = Room.Board
- x8 = Personal
- x9 = PhD
- x10 = perc.alumni
- x11 = Expend

Both qualitative variables:

- x1 = 1 if private & 0 if not private
- x3 = 1 if elite & 0 if non elite

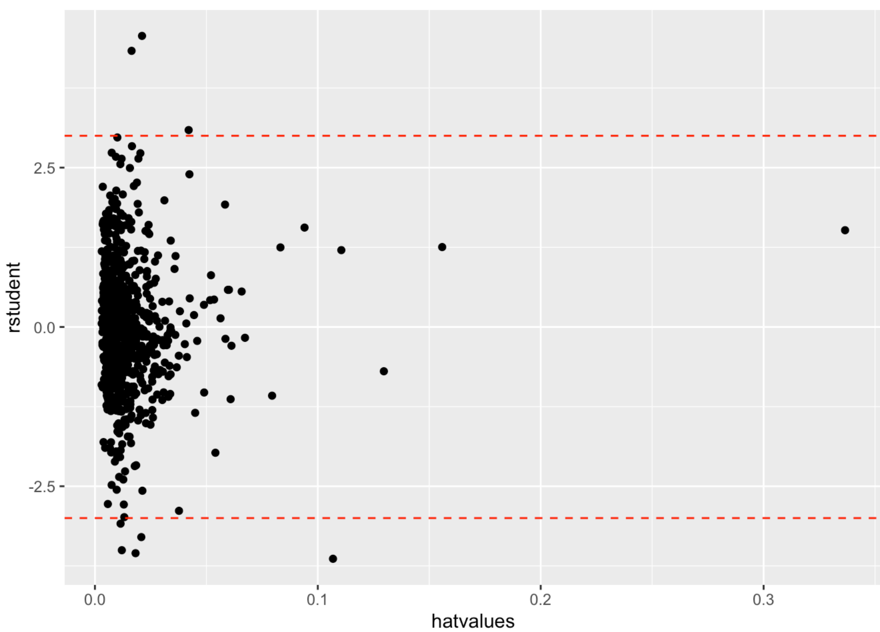
Final Model Equation: y = 4.840e+01 + 4.770e+00x1 - 1.778e+01x2 +4.022e+00x3 + 6.631e-04x4 - 1.963e-03x5 + 1.215e-03x6 + 1.534e-03x7 - 1.820e-03x8 + 8.424e-02x9 + 3.060e-01x10 - -4.465e-04x11

Final Model Genersal Equation:

$$\text{Grad. Rate} = \alpha + \beta_1(\text{Private}) + \beta_2(\text{Accept. pct}) + \beta_3(\text{Elite10}) + \beta_4(\text{F. Undergrad}) + \beta_5(\text{P. Undergrad}) + \beta_6(\text{Outstate}) + \beta_7(\text{Room. Board}) + \beta_8(\text{Personal}) -$$

Studentized Residuals

```
# plot of deleted studentized residuals vs hat values
student_hat <- data.frame(rstudent = rstudent(grad_model_1), hatvalues =  hatvalues(grad_model_1))

# plot rstudent vs hatvalues
student_hat_plot <- ggplot(student_hat, aes(x = hatvalues, y = rstudent)) + geom_point() +
                    # Change line type and color
                    geom_hline(yintercept=3, linetype="dashed", color = "red") +
                    geom_hline(yintercept=-3, linetype="dashed", color = "red")

# plot
student_hat_plot
```

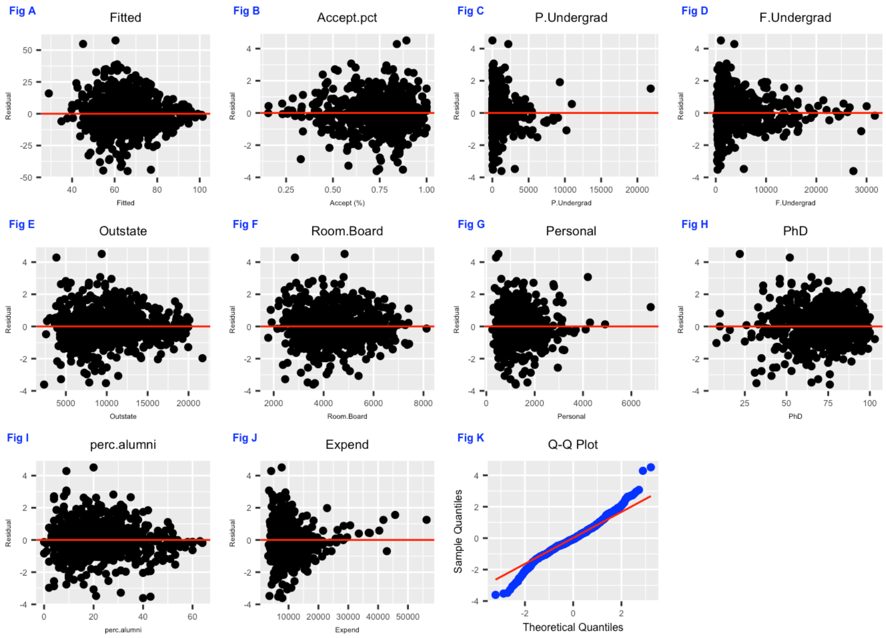


h. Draw a scatter plot of the studentized residuals against the predicted values. Does the plot show any striking pattern indicating problems in the regression analysis?

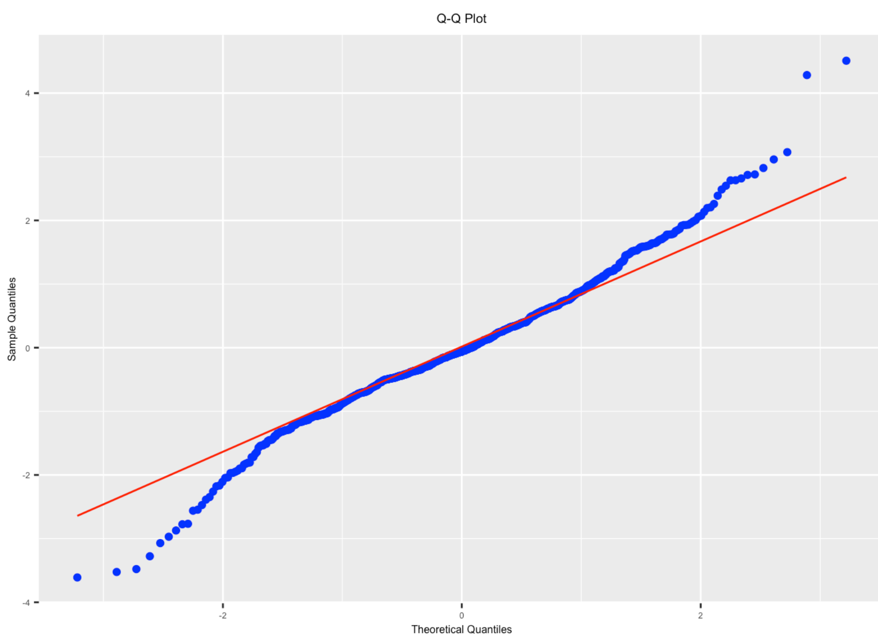
There are multiple values outside of absolute|3| on rstudentized, which indicates potential, influential point exists, and high leverage points are below 0.5, so hat values are good.

```
# residual vs fitted
residual_plot <-  ggplot(grad_model_1, aes(x = .fitted, y = .resid)) +
                  geom_point() +
                  geom_hline(yintercept = 0, col = "red") +
                  labs(title="Fitted",
                    x = "Fitted", y = "Residual") +
                  # move the title text to the middle
                  theme(plot.title=element_text(hjust=0.5))  +
                  theme(text = element_text(size = 6))  +
                  theme(axis.title = element_text(size = 4))

# Accept.pct vs residuals
accept.pct_plot <-  ggplot(college, aes(x = Accept.pct, y = rstandard(grad_model_1))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="Accept.pct",
                    x = "Accept (%)", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 4))

# P.Undergrad  vs residuals
PUndergrad_plot <-  ggplot(college, aes(x = P.Undergrad, y = rstandard(grad_model_1))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="P.Undergrad",
                    x = "P.Undergrad", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 4))

# F.Undergrad  vs residuals
FUndergrad_plot <-  ggplot(college, aes(x = F.Undergrad, y = rstandard(grad_model_1))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="F.Undergrad",
                    x = "F.Undergrad", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 4))


# Outstate  vs residuals
Outstate_plot <-  ggplot(college, aes(x = Outstate, y = rstandard(grad_model_1))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="Outstate",
                    x = "Outstate", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 4))

# Room.Board  vs residuals
Room_plot <-  ggplot(college, aes(x = Room.Board, y = rstandard(grad_model_1))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="Room.Board",
                    x = "Room.Board", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 4))

# Personal  vs residuals
Personal_plot <-  ggplot(college, aes(x = Personal, y = rstandard(grad_model_1))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="Personal",
                    x = "Personal", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 4))
```

```
# PhD  vs residuals
PhD_plot <-  ggplot(college, aes(x = PhD, y = rstandard(grad_model_1))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="PhD",
                    x = "PhD", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 4))

# perc.alumni  vs residuals
perc.alumni_plot <-  ggplot(college, aes(x = perc.alumni, y = rstandard(grad_model_1))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="perc.alumni",
                    x = "perc.alumni", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 4))


# Expend  vs residuals
Expend_plot <-  ggplot(college, aes(x = Expend, y = rstandard(grad_model_1))) +
                geom_point() +
                geom_hline(yintercept = 0, col = "red") +
                labs(title="Expend",
                    x = "Expend", y = "Residual") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 4))

#create Q-Q plot
qq_grad_plot <-  ggplot(grad_model_1, aes(sample=rstandard(grad_model_1))) +
                stat_qq(size=1.5, color='blue') +
                stat_qq_line(col = "red") +
                labs(title="Q-Q Plot",
                  x = "Theoretical Quantiles", y = "Sample Quantiles") +
                # move the title text to the middle
                theme(plot.title=element_text(hjust=0.5))  +
                theme(text = element_text(size = 6))  +
                theme(axis.title = element_text(size = 6))

combine_plot <- ggarrange(residual_plot, accept.pct_plot, PUndergrad_plot, FUndergrad_plot,
                          Outstate_plot, Room_plot, Personal_plot, PhD_plot,
                          perc.alumni_plot, Expend_plot, qq_grad_plot,
                          labels = c("Fig A", "Fig B", "Fig C", "Fig D", "Fig E", "Fig F",
                                     "Fig G", "Fig H", "Fig I", "Fig J", "Fig K"),
                           font.label = list(size = 6, color = "blue"))

# plot all
combine_plot
```




```
# separate Q-Q Plot
qq_grad_plot
```



i. Analyze normal probability plot of residuals. Is there any evidence that the assumption of normality is not satisfied?

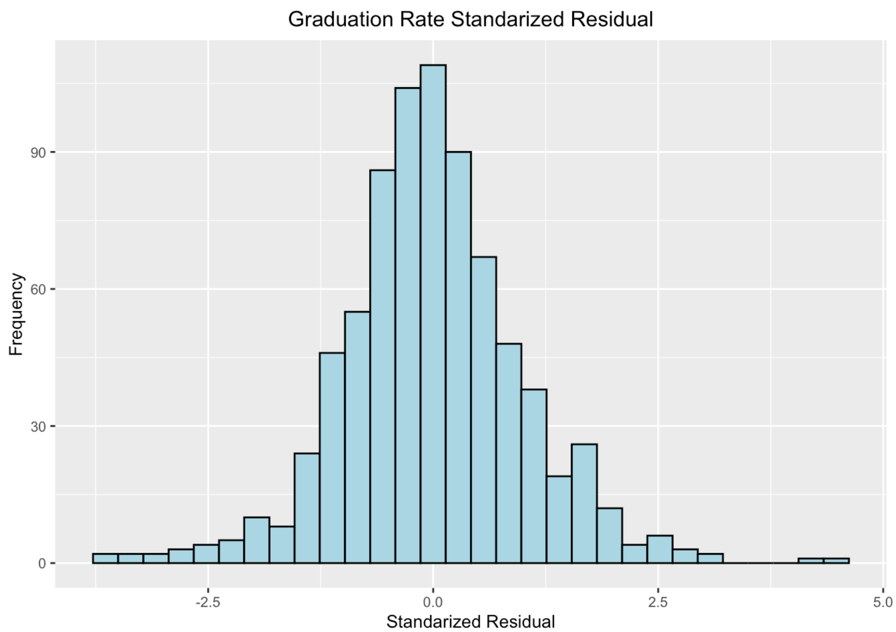
Examining the normal probability plot of residuals shows no evidence of normality not being satisfied, but signs of outliers are present.

Influential Points and Outliers

```
# outliers |standardized residuals| > 3
grad_std_residual = data.frame(residual = rstandard(grad_model_1))

# display |standardized residuals| > 3
filter(grad_std_residual, abs(residual) > 3)
```

```
##      residual
## 70  -3.609875
## 96   4.509045
## 99  -3.070339
## 114 -3.277254
## 318  3.072519
## 378  4.283390
## 395 -3.524229
## 586 -3.478561
```

```
# historgram for outliers
ggplot(grad_std_residual, aes(x = residual)) +
  geom_histogram(bins=30, color="black", fill="lightblue") +
  labs(title = "Graduation Rate Standarized Residual", x = "Standarized Residual",
       y = 'Frequency') +
      # move the title text to the middle
      theme(plot.title=element_text(hjust=0.5)) +
      theme(text = element_text(size = 10))  +
      theme(axis.title = element_text(size = 10))
```




```
# print out only observations that may be influential
summary(influence.measures(grad_model_1))
```

```
## Potentially influential observations of
##   lm(formula = Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad +      P.Undergrad + Outstate + Room.B
oard + Personal + PhD + perc.alumni +      Expend, data = college) :
##
##     dfb.1_ dfb.Prvt dfb.Acc. dfb.El10 dfb.F.Un dfb.P.Un dfb.Otst dfb.Rm.B
## 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.prc. 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
```

j. Are there any outliers or Influential Points? Compute appropriate statistics.

There are outliers and influential points in our dataset according to summary(influence.measures(grad_model_1)) and need to be investigated for the validity of the input data.

```
# summary of the model
summary(grad_model_1)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad +
##     P.Undergrad + Outstate + Room.Board + Personal + PhD + perc.alumni +
##     Expend, data = college)
##
## 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 **
## Accept.pct  -1.778e+01  3.797e+00  -4.683 3.34e-06 ***
## Elite10      4.022e+00  2.002e+00   2.009  0.04492 *
## F.Undergrad  6.631e-04  1.411e-04   4.699 3.10e-06 ***
## P.Undergrad -1.963e-03  3.901e-04  -5.031 6.09e-07 ***
## Outstate     1.215e-03  2.270e-04   5.352 1.15e-07 ***
## Room.Board   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 *
## perc.alumni  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
```

k. Analyze the R2 value for the final model and discuss how well the model explains the variation in graduation rates among the universities.

Multiple R-squared: 0.4431

The final model explains 44.31% of the variability of the model in graduation rates among the universities.

Standardized Coefficients

```
# standardized coefficients
lm.beta(grad_model_1)
```

```
##     Private  Accept.pct     Elite10 F.Undergrad P.Undergrad    Outstate
##  0.12377128 -0.15228071  0.07040430  0.18723983 -0.17395180  0.28449029
##  Room.Board    Personal         PhD perc.alumni      Expend
##  0.09794545 -0.07173299  0.08007146  0.22073143 -0.13573070
```

l. Draw conclusions on graduation rates based on your regression analysis. What are the most important predictors in your model? Does your model show a significant difference in graduation rates between private and public universities? Do “elite” universities have higher graduation rates?

Based on my regression analysis, I can state that `Outstate` has the most influence and, followed by `perc.alumni` , `F.Undergrad` , `Private` are the top three in this order. So yes, elite universities have a significantly higher graduation rate looking at the box plot (c). Still, it adds the smallest positive value according to the normalized beta coefficient in the current model.

Part 2 - Interaction Terms

```
# interaction model
interaction_model_1 <- lm(Grad.Rate ~ Elite10 + Accept.pct + Outstate +
                          perc.alumni + Expend + Elite10:Accept.pct +
                          Elite10:Outstate + Elite10:perc.alumni +
                          Elite10:Expend, data=college)

# summary report of the interaction_model
summary(interaction_model_1)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Elite10 + Accept.pct + Outstate + perc.alumni +
##     Expend + Elite10:Accept.pct + Elite10:Outstate + Elite10:perc.alumni +
##     Elite10:Expend, data = college)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.787  -7.785  -0.400   7.769  57.177
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          5.316e+01  3.592e+00  14.801  < 2e-16 ***
## Elite10              3.763e+01  1.000e+01   3.762 0.000181 ***
## Accept.pct          -1.519e+01  4.129e+00  -3.678 0.000251 ***
## Outstate             2.296e-03  1.991e-04  11.532  < 2e-16 ***
## perc.alumni          3.505e-01  5.030e-02   6.968 6.95e-12 ***
## Expend              -9.536e-04  2.073e-04  -4.601 4.93e-06 ***
## Elite10:Accept.pct  -2.274e+01  9.822e+00  -2.315 0.020881 *
## Elite10:Outstate    -2.054e-03  5.390e-04  -3.811 0.000150 ***
## Elite10:perc.alumni -1.227e-01  1.347e-01  -0.911 0.362485
## Elite10:Expend       1.050e-03  2.889e-04   3.635 0.000297 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.12 on 767 degrees of freedom
## Multiple R-squared:  0.4234, Adjusted R-squared:  0.4167
## F-statistic: 62.58 on 9 and 767 DF,  p-value: < 2.2e-16
```

```
# variance report of the interaction_model
anova(interaction_model_1)
```

```
## Analysis of Variance Table
##
## Response: Grad.Rate
##                      Df Sum Sq Mean Sq  F value    Pr(>F)
## Elite10               1  27847   27847 161.7792 < 2.2e-16 ***
## Accept.pct            1   4602    4602  26.7364 2.978e-07 ***
## Outstate              1  48983   48983 284.5689 < 2.2e-16 ***
## perc.alumni           1   8862    8862  51.4830 1.706e-12 ***
## Expend                1   1584    1584   9.2030 0.0024977 **
## Elite10:Accept.pct    1    917     917   5.3301 0.0212251 *
## Elite10:Outstate      1   1709    1709   9.9300 0.0016892 **
## Elite10:perc.alumni   1    176     176   1.0215 0.3124757
## Elite10:Expend        1   2274    2274  13.2115 0.0002968 ***
## Residuals           767 132023     172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Variance Inflation Factors for interaction_model
vif.grad_inter_model_1 <- vif(interaction_model_1)
vif.grad_inter_model_1
```

```
##             Elite10          Accept.pct            Outstate         perc.alumni
##             40.7840              1.6632              2.8927              1.7518
##              Expend  Elite10:Accept.pct    Elite10:Outstate Elite10:perc.alumni
##              5.2808             13.2160             29.7410             10.1280
##      Elite10:Expend
##             15.0050
```

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.

A current model summary report has `Elite10:perc.alumni` t-value is `0.362485` and F-value being 0.3124757 for hypothesis testing, in which we can not reject the null hypothesis; therefore, `Elite10:perc.alumni` interaction variable should be removed. All the interaction variables having VIF greater than 10 make sense, including the `Elite10` predictor.

```
# interaction model 2:
interaction_model_2 <- lm(Grad.Rate ~ Elite10 + Accept.pct + Outstate +
                          perc.alumni + Expend + Elite10:Accept.pct +
                          Elite10:Outstate + Elite10:Expend, data=college)

# summary report of the interaction_model
summary(interaction_model_2)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Elite10 + Accept.pct + Outstate + perc.alumni +
##     Expend + Elite10:Accept.pct + Elite10:Outstate + Elite10:Expend,
##     data = college)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.724  -7.744  -0.468   7.727  57.150
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.314e+01  3.591e+00  14.797  < 2e-16 ***
## Elite10             3.585e+01  9.808e+00   3.655 0.000275 ***
## Accept.pct         -1.505e+01  4.126e+00  -3.647 0.000283 ***
## Outstate            2.322e-03  1.970e-04  11.786  < 2e-16 ***
## perc.alumni         3.334e-01  4.666e-02   7.145 2.09e-12 ***
## Expend             -9.506e-04  2.072e-04  -4.587 5.24e-06 ***
## Elite10:Accept.pct -2.164e+01  9.747e+00  -2.220 0.026705 *
## Elite10:Outstate   -2.253e-03  4.926e-04  -4.575 5.56e-06 ***
## Elite10:Expend      1.057e-03  2.888e-04   3.661 0.000268 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.12 on 768 degrees of freedom
## Multiple R-squared:  0.4228, Adjusted R-squared:  0.4168
## F-statistic: 70.32 on 8 and 768 DF,  p-value: < 2.2e-16
```

```
# variance report of the interaction_model
anova(interaction_model_2)
```

```
## Analysis of Variance Table
##
## Response: Grad.Rate
##                     Df Sum Sq Mean Sq  F value    Pr(>F)
## Elite10              1  27847   27847 161.8150 < 2.2e-16 ***
## Accept.pct           1   4602    4602  26.7423 2.968e-07 ***
## Outstate             1  48983   48983 284.6318 < 2.2e-16 ***
## perc.alumni          1   8862    8862  51.4943 1.694e-12 ***
## Expend               1   1584    1584   9.2051  0.002495 **
## Elite10:Accept.pct   1    917     917   5.3313  0.021210 *
## Elite10:Outstate     1   1709    1709   9.9322  0.001687 **
## Elite10:Expend       1   2307    2307  13.4057  0.000268 ***
## Residuals          768 132166     172
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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.

Model Equation:

- Beta0 = intercept
- Elite10 = x1
- Accept.pct = x2
- Outstate = x3
- perc.alumni = x4
- Expend = x5
- Elite10:Accept.pct = x1x2
- Elite10:Outstate = x1x3
- Elite10:Expend = x1x5

y = 5.314e+01 + 3.585e+01x1 - 1.505e+01x2 + 2.322e-03x3 + 3.334e-01x4 - 9.506e-04x5 - 2.164e+01(x1x2) - 2.253e-03(x1x3) + 1.057e-03(x1x5)

Final Model General Equation:

$$\text{Grad.Rate} = \alpha + \beta_1(\text{Elite10}) + \beta_2(\text{Accept.pct}) + \beta_3(\text{Outstate}) + \beta_4(\text{perc.alumni}) + \beta_5(\text{Expend}) + \beta_6(\text{Elite10} \times \text{Accept.pct}) + \beta_7(\text{Elite10} \times \text{Outstate}) + \beta_8$$

```
# summary report to get coefficients
summary(interaction_model_2)
```

```
##
## Call:
## lm(formula = Grad.Rate ~ Elite10 + Accept.pct + Outstate + perc.alumni +
##     Expend + Elite10:Accept.pct + Elite10:Outstate + Elite10:Expend,
##     data = college)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.724  -7.744  -0.468   7.727  57.150
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         5.314e+01  3.591e+00  14.797  < 2e-16 ***
## Elite10             3.585e+01  9.808e+00   3.655 0.000275 ***
## Accept.pct         -1.505e+01  4.126e+00  -3.647 0.000283 ***
## Outstate            2.322e-03  1.970e-04  11.786  < 2e-16 ***
## perc.alumni         3.334e-01  4.666e-02   7.145 2.09e-12 ***
## Expend             -9.506e-04  2.072e-04  -4.587 5.24e-06 ***
## Elite10:Accept.pct -2.164e+01  9.747e+00  -2.220 0.026705 *
## Elite10:Outstate   -2.253e-03  4.926e-04  -4.575 5.56e-06 ***
## Elite10:Expend      1.057e-03  2.888e-04   3.661 0.000268 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.12 on 768 degrees of freedom
## Multiple R-squared:  0.4228, Adjusted R-squared:  0.4168
## F-statistic: 70.32 on 8 and 768 DF,  p-value: < 2.2e-16
```

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.

According to the model, if Elite10 is Yes, the Graduation Rate will increase by 35.85% if all other variables are held constant. If non-elite, the beta coefficient will be zero; therefore, there is no relationship with other variables.

Optional - Model validation:

```
# make this example reproducible
set.seed(1980)

# Create training and testing set
# split samples (75% for training and 25% for testing)
select.college_new <- sample(1:nrow(college_new), 0.75*nrow(college_new))
# selecting 75% of data for training purpose
train.college <- college_new[select.college_new,]
# selecting 25% (remaining) of data for testing purpose
test.college <- college_new[-select.college_new,]

# training model with split set
p_model_1 <- lm(Grad.Rate ~ Private + Accept.pct + Elite10 +
                 F.Undergrad + P.Undergrad + Outstate + Room.Board +
                 Personal + PhD + perc.alumni + Expend, data=train.college)

# predict fitted values using test.college data M1
y_pred_M1 <- predict.glm(p_model_1, test.college)
y_obs_M1  <- test.college[,"Grad.Rate"]

# Compute RMSE of prediction errors
rmse_m1 <- sqrt((y_obs_M1 - y_pred_M1)%*%(y_obs_M1 - y_pred_M1)/nrow(test.college))

# Compute mean absolute error
mae_m1 <- mean(abs(y_obs_M1 - y_pred_M1))

# Compute mean percentage absolute error
mape_m1 <- mean(abs((y_obs_M1 - y_pred_M1)/y_obs_M1))*100

# compute cross-validated R^2_pred
r2_pred_M1  <- cor(cbind(y_obs_M1, y_pred_M1))**2
r2_train_M1 <- summary(p_model_1)$r.squared
diffr2_M1   <- abs(r2_train_M1 - r2_pred_M1)

# difference of cross-validate R2 and R2
dr2_M1 <- diffr2_M1[1,2]


# training model with test set w/o Room.Board - extra for fun
p_model_2  <- lm(Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad +
                 P.Undergrad + Outstate + Personal + PhD + perc.alumni +
                 Expend, data=train.college)

# extra: checking to see if grad_model_2 is better than the grad_model_1
# create fitted values using test.college data M1
y_pred_M2 <- predict.glm(p_model_2, test.college)
y_obs_M2  <- test.college[,"Grad.Rate"]

# Compute RMSE of prediction errors
rmse_m2 <- sqrt((y_obs_M2 - y_pred_M2)%*%(y_obs_M2 - y_pred_M2)/nrow(test.college))

# Compute mean absolute error
mae_m2 <- mean(abs(y_obs_M2 - y_pred_M2))

# Compute mean percentage absolute error
mape_m2 <- mean(abs((y_obs_M2 - y_pred_M2)/y_obs_M2))*100

# compute cross-validated R^2_pred
r2_pred_M2  <- cor(cbind(y_obs_M2, y_pred_M2))**2
r2_train_M2 <- summary(p_model_2)$r.squared
diffr2_M2   <- abs(r2_train_M2 - r2_pred_M2)

#print difference of cross-validate R2 and R2
dr2_M2 <- diffr2_M2[1,2]


# create dataframe
Model      <-  c("Model 1", "Model 2")
Model_Type <-  c("p_model_1", "p_model_2")
RMSE       <-  c(rmse_m1, rmse_m2)
MAE        <-  c(mae_m1, mae_m2)
MAPE       <-  c(mape_m1, mape_m2)
Diff_R2    <-  c(dr2_M1, dr2_M2)

df <- data.frame(Model, Model_Type, RMSE, MAE, MAPE, Diff_R2)

# print M! & M2 Model Info
df
```

```
##     Model Model_Type     RMSE      MAE     MAPE    Diff_R2
## 1 Model 1  p_model_1 11.83692 9.067492 15.01750 0.02433098
## 2 Model 2  p_model_2 11.89292 9.117195 15.11857 0.02473060
```

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 Part 1 predicts new data. Compute the MAPE (mean absolute percentage error) statistic and discuss the results.

The Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are metrics used to evaluate a Regression Model. These metrics tell us how accurate our predictions are and what is the amount of deviation from the actual values. The RMSE is a quadratic scoring rule that measures the average magnitude of the error. The MAE is a linear score, and all the individual differences are weighted equally in the average. The MAE and the RMSE can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or equal to the MAE; the more significant difference between them, the greater the variance in the individual errors in the sample. If the RMSE = MAE, all the errors are of the same magnitude.

The mean absolute percentage error (MAPE) is the mean or average of the absolute percentage errors of forecasts. For example, our M1 MAPE is 15.02%, then our predictions are, on average, 15.02% away from the actual values. Our M1 RMSE is greater than MAE because RMSE uses squared differences in its formula and the squared difference between the observed value and the predicted value; therefore, outliers will create a more considerable difference.

The cross-validated R^2 value for the M1 model is 0.02433098, indicating that the M1 is not over-fitting due to the difference between R^2 values.

Compare Model 1 & 2 from Prob 1 & Prob 2 & using training set

```
# Problem 1 Final Model
p2_model_1 <- lm(Grad.Rate ~ Private + Accept.pct + Elite10 +
                   F.Undergrad + P.Undergrad + Outstate + Room.Board +
                   Personal + PhD + perc.alumni + Expend, data=train.college)


# create fitted values using test.college data M1
y_pred_cM1 <- predict.glm(p2_model_1, test.college)
y_obs_cM1  <- test.college[,"Grad.Rate"]

# Compute RMSE of prediction errors
rmse_cm1 <- sqrt((y_obs_cM1 - y_pred_cM1)%*%(y_obs_cM1 - y_pred_cM1)/nrow(test.college))

# Compute mean absolute error
mae_cm1 <- mean(abs(y_obs_cM1 - y_pred_cM1))

# Compute mean percentage absolute error
mape_cm1 <- mean(abs((y_obs_cM1 - y_pred_cM1)/y_obs_cM1))*100

# compute cross-validated R^2_pred
r2_pred_cM1  <- cor(cbind(y_obs_cM1, y_pred_cM1))**2
r2_train_cM1 <- summary(p2_model_1)$r.squared
diffr2_cM1   <- abs(r2_train_cM1 - r2_pred_cM1)

#print difference of cross-validate R2 and R2
dr2_cM1 <- diffr2_cM1[1,2]

# Problem 2 Final Model
p2_model_2 <- lm(Grad.Rate ~ Elite10 + Accept.pct + Outstate +
                   perc.alumni + Expend + Elite10:Accept.pct +
                   Elite10:Outstate + Elite10:Expend, data=train.college)

# create fitted values using test.college data M2
y_pred_cM2 <- predict.glm(p2_model_2, test.college)
y_obs_cM2  <- test.college[,"Grad.Rate"]

# Compute RMSE of prediction errors
rmse_cm2 <- sqrt((y_obs_cM2 - y_pred_cM2)%*%(y_obs_cM2 - y_pred_cM2)/nrow(test.college))

# Compute mean absolute error
mae_cm2 <- mean(abs(y_obs_cM2 - y_pred_cM2))

# Compute mean percentage absolute error
mape_cm2 <- mean(abs((y_obs_cM2 - y_pred_cM2)/y_obs_cM2))*100

# compute cross-validated R^2_pred
r2_pred_cM2  <- cor(cbind(y_obs_cM2, y_pred_cM2))**2
r2_train_cM2 <- summary(p2_model_2)$r.squared
diffr2_cM2   <- abs(r2_train_cM2 - r2_pred_cM2)

# difference of cross-validate R2 and R2
dr2_cM2 <- diffr2_cM2[1,2]

# create dataframe
Model   <-  c("Model 1", "Model 2")
RMSE    <-  c(rmse_cm1, rmse_cm2)
MAE     <-  c(mae_cm1, mae_cm2)
MAPE    <-  c(mape_cm1, mape_cm2)
Diff_R2 <-  c(dr2_cM1, dr2_cM2)

df <- data.frame(Model, RMSE, MAE, MAPE, Diff_R2)

# print M! & M2 Model Info
df
```

```
##     Model     RMSE      MAE     MAPE    Diff_R2
## 1 Model 1 11.83692 9.067492 15.01750 0.02433098
## 2 Model 2 12.13696 9.558507 15.86089 0.03209292
```

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.

Model 1 minimizes three out of three validation metrics. As a result, model 1 provides more accurate predictions (closer to actual values). The MAPE value for M1 indicates that, on average, predictions are off by about 15.02% of the actual value. In addition, the cross-validated R^2 value is 0.02433098 for the M1, indicating that the model is not over-fitting.

note: CV five-fold cross-validation also chose the same model as train and test.

Cross Validation: extra work for fun

```
# cross validation for the Prob 1 comparing grad_model_1
results_cvP1 = cv.lm(data = college, form.lm = grad_model_1, m = 5, seed=11,
                     plotit = FALSE, printit = FALSE)

# validation statistics
# Root Mean Squared Error (RMSE)  value
rmse_cv_P1 <- sqrt((results_cvP1$Grad.Rate - results_cvP1$cvpred)%*%(results_cvP1$Grad.Rate - results_cvP1$cvpre
d)/nrow(results_cvP1))

# create cross validated values using full data
y_pred_cvP1 <- results_cvP1$Predicted
y_obs_cvP1  <- results_cvP1$Grad.Rate

# Compute mean absolute error
mae_cv_P1 <- mean(abs(y_obs_cvP1 - y_pred_cvP1))

# Compute mean percentage absolute error
mape_cvP1 <- mean(abs((y_obs_cvP1 - y_pred_cvP1)/y_obs_cvP1))*100

# compute cross-validated R^2_pred
r2_pred_cvP1  <- cor(cbind(y_obs_cvP1, y_pred_cvP1))**2
r2_train_cvP1 <- summary(grad_model_1)$r.squared
diffr2_cvP1   <- abs(r2_train_cvP1 - r2_pred_cvP1)

# difference of cross-validate R2 and R2
dr2_cvP1 <- diffr2_cvP1[1,2]

# cross validation for the Prob 1 comparing grad_model_1
results_cvP2 = cv.lm(data = college, form.lm = grad_model_2, m = 5, seed=11,
                     plotit = FALSE, printit = FALSE)

# validation statistics
# Root Mean Squared Error (RMSE)  value
rmse_cv_P2 <- sqrt((results_cvP2$Grad.Rate - results_cvP2$cvpred)%*%(results_cvP2$Grad.Rate - results_cvP2$cvpre
d)/nrow(results_cvP2))

# create cross validated values using full data
y_pred_cvP2 <- results_cvP2$Predicted
y_obs_cvP2  <- results_cvP2$Grad.Rate

# Compute mean absolute error
mae_cv_P2 <- mean(abs(y_obs_cvP2 - y_pred_cvP2))

# Compute mean percentage absolute error
mape_cvP2 <- mean(abs((y_obs_cvP2 - y_pred_cvP2)/y_obs_cvP2))*100

# compute cross-validated R^2_pred
r2_pred_cvP2  <- cor(cbind(y_obs_cvP2, y_pred_cvP2))**2
r2_train_cvP2 <- summary(grad_model_2)$r.squared
diffr2_cvP2   <- abs(r2_train_cvP2 - r2_pred_cvP2)

# difference of cross-validate R2 and R2
dr2_cvP2 <- diffr2_cvP2[1,2]

# create dataframe
Model   <-  c("grad_model_1", "grad_model_2")
RMSE    <-  c(rmse_cv_P1, rmse_cv_P2)
MAE     <-  c(mae_cv_P1, mae_cv_P2)
MAPE    <-  c(mape_cvP1, mape_cvP2)
Diff_R2 <-  c(dr2_cvP1, dr2_cvP2)

df <- data.frame(Model, RMSE, MAE, MAPE, Diff_R2)

# print M! & M2 Model Info
df
```

```
##          Model     RMSE      MAE     MAPE      Diff_R2
## 1 grad_model_1 12.97300 9.586035 17.86888 1.387779e-15
## 2 grad_model_2 13.01482 9.647566 17.99663 4.996004e-16
```

```
# Problem 1 Final Model - full data
cv_model_1 <- lm(Grad.Rate ~ Private + Accept.pct + Elite10 +
                    F.Undergrad + P.Undergrad + Outstate + Room.Board +
                    Personal + PhD + perc.alumni + Expend, data=college)

results_cvM1 = cv.lm(data = college, form.lm = cv_model_1, m = 5, seed=11,
                    plotit = FALSE, printit = FALSE)

# validation statistics
# Root Mean Squared Error (RMSE)  value
rmse_cv_M1 <- sqrt((results_cvM1$Grad.Rate - results_cvM1$cvpred)%*%(results_cvM1$Grad.Rate - results_cvM1$cvpre
d)/nrow(results_cvM1))

# create cross validated values using full data
y_pred_cvM1 <- results_cvM1$Predicted
y_obs_cvM1  <- results_cvM1$Grad.Rate

# Compute mean absolute error
mae_cv_M1 <- mean(abs(y_obs_cvM1 - y_pred_cvM1))

# Compute mean percentage absolute error
mape_cvM1 <- mean(abs((y_obs_cvM1 - y_pred_cvM1)/y_obs_cvM1))*100

# compute cross-validated R^2_pred
r2_pred_cvM1  <- cor(cbind(y_obs_cvM1, y_pred_cvM1))**2
r2_train_cvM1 <- summary(cv_model_1)$r.squared
diffr2_cvM1   <- abs(r2_train_cvM1 - r2_pred_cvM1)

# difference of cross-validate R2 and R2
dr2_cvM1 <- diffr2_cvM1[1,2]

# Problem 2 Final Model
cv_model_2 <- lm(Grad.Rate ~ Elite10 + Accept.pct + Outstate +
                  perc.alumni + Expend + Elite10:Accept.pct +
                  Elite10:Outstate + Elite10:Expend, data=train.college)

# K-Fold cross-validation for multiple regression models
results_cvM2 = cv.lm(data = college, form.lm = cv_model_2, m = 5, seed = 11,
                    plotit = FALSE, printit = FALSE)


# validation statistics
# Root Mean Squared Error (RMSE)  value
rmse_cv_M2 <- sqrt((results_cvM2$Grad.Rate - results_cvM2$cvpred)%*%(results_cvM2$Grad.Rate - results_cvM2$cvpre
d)/nrow(results_cvM2))

# create cross validated values using full data
y_pred_cvM2 <- results_cvM2$Predicted
y_obs_cvM2  <- results_cvM2$Grad.Rate

# Compute mean absolute error
mae_cv_M2 <- mean(abs(y_obs_cvM2 - y_pred_cvM2))

# Compute mean percentage absolute error
mape_cvM2 <- mean(abs((y_obs_cvM2 - y_pred_cvM2)/y_obs_cvM2))*100

# compute cross-validated R^2_pred
r2_pred_cvM2  <- cor(cbind(y_obs_cvM2, y_pred_cvM2))**2
r2_train_cvM2 <- summary(cv_model_2)$r.squared
diffr2_cvM2   <- abs(r2_train_cvM2 - r2_pred_cvM2)

# difference of cross-validate R2 and R2
dr2_cvM2 <- diffr2_cvM2[1,2]


# create dataframe
Model   <-  c("Model 1", "Model 2")
RMSE    <-  c(rmse_cv_M1, rmse_cv_M2)
MAE     <-  c(mae_cv_M1, mae_cv_M2)
MAPE    <-  c(mape_cvM1, mape_cvM2)
Diff_R2 <-  c(dr2_cvM1, dr2_cvM2)

df <- data.frame(Model, RMSE, MAE, MAPE, Diff_R2)

# print M! & M2 Model Info
df
```

```
##     Model     RMSE      MAE     MAPE      Diff_R2
## 1 Model 1 12.97300 9.586035 17.86888 1.387779e-15
## 2 Model 2 13.16036 9.861168 18.57689 5.980538e-03
```