Assignment 3

Sanket Praveen Patil

2023-10-31

# Problem 1

library(psych)

# Importing data in R

df = read.csv("college.csv", header = T)  
dim(df)

## [1] 777 16

head(df)

## 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

describe(df$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

# Drop column with unique values

df <- df[, !colnames(df) %in% "school"]

# NA values check

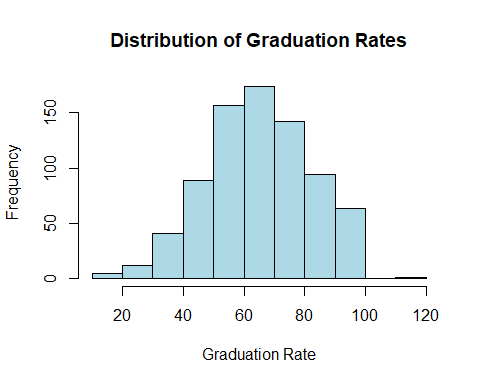
colMeans(is.na(df)) \* 100

## Private Accept.pct Elite10 F.Undergrad P.Undergrad Outstate   
## 0 0 0 0 0 0   
## Room.Board Books Personal PhD Terminal S.F.Ratio   
## 0 0 0 0 0 0   
## perc.alumni Expend Grad.Rate   
## 0 0 0

# Question 1:

# Plot the distribution

hist(df$Grad.Rate, main = "Distribution of Graduation Rates", xlab = "Graduation Rate", col = "lightblue", border = "black")



skew(df$Grad.Rate)

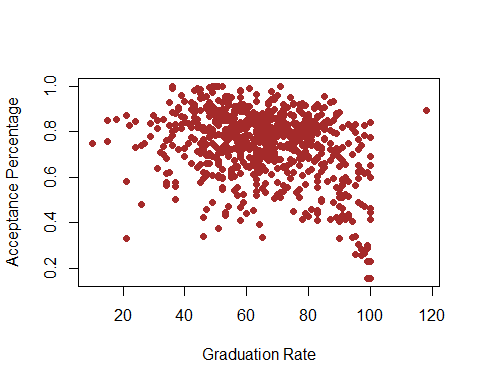
## [1] -0.1133384

# By seeing at the data, data looks normally distributed but slightly negatively skewed.

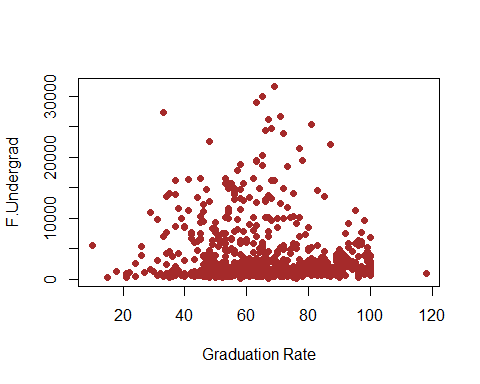
# Question 2 :

# Scatter plots

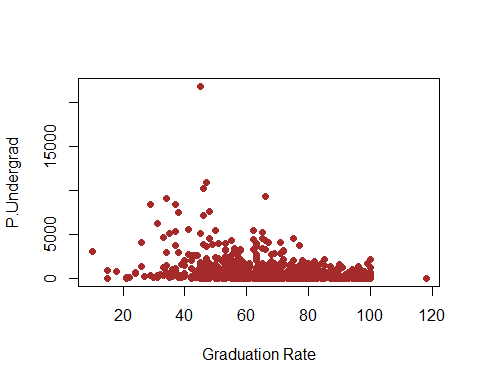
plot(df$Grad.Rate, df$Accept.pct, , xlab = "Graduation Rate", ylab = "Acceptance Percentage", col = "brown", pch = 16)



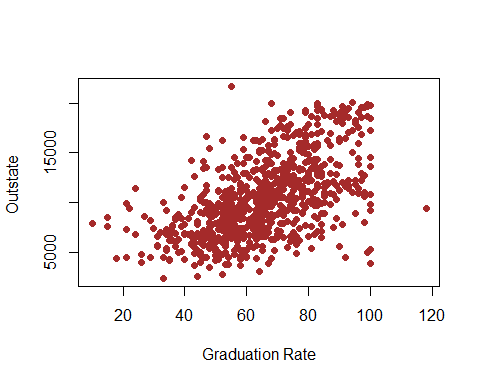
# Slight negative linear relation between Grad.Rate and Accept.pct  
  
plot(df$Grad.Rate, df$F.Undergrad, , xlab = "Graduation Rate", ylab = "F.Undergrad", col = "brown", pch = 16)



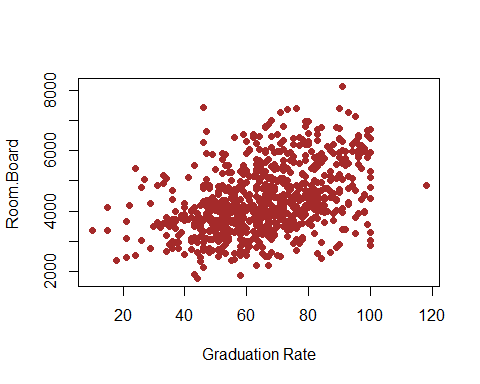
# No linear relation between Grad.Rate and F.Undergrad  
  
plot(df$Grad.Rate, df$P.Undergrad, , xlab = "Graduation Rate", ylab = "P.Undergrad", col = "brown", pch = 16)



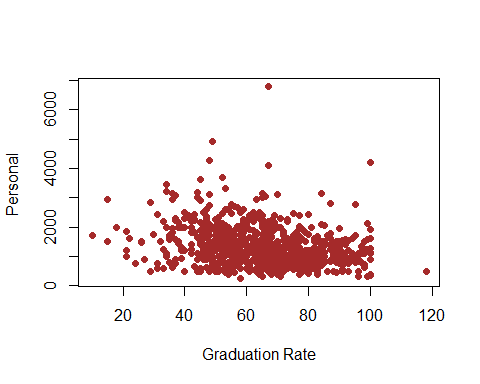
# No linear relation between Grad.Rate and P.Undergrad  
  
plot(df$Grad.Rate, df$Outstate, , xlab = "Graduation Rate", ylab = "Outstate ", col = "brown", pch = 16)



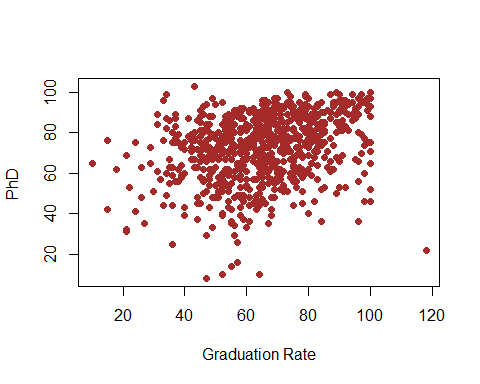
# There is positive linear relation between Grad>rate and Outstate  
  
plot(df$Grad.Rate, df$Room.Board, , xlab = "Graduation Rate", ylab = "Room.Board ", col = "brown", pch = 16)



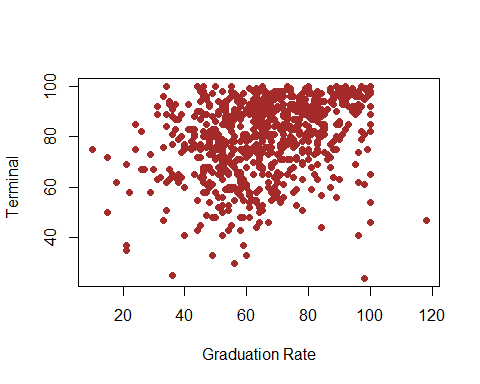
# There is positive linear relation between Grad>rate and Room.Board  
  
plot(df$Grad.Rate, df$Personal, , xlab = "Graduation Rate", ylab = "Personal ", col = "brown", pch = 16)



# There is no linear relation between Grad.Rate and Personal  
  
plot(df$Grad.Rate, df$PhD, , xlab = "Graduation Rate", ylab = "PhD ", col = "brown", pch = 16)



# There is slight linear relation between Grad.Rate and PhD  
  
plot(df$Grad.Rate, df$Terminal, , xlab = "Graduation Rate", ylab = "Terminal ", col = "brown", pch = 16)



# There is no linear relation between Grad.Rate and Terminal

# Separating numeric columns

numeric\_cols <- sapply(df, is.numeric)  
df\_numeric <- df[, numeric\_cols]  
correlation\_matrix <- cor(df\_numeric)  
correlation\_matrix

## 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  
## 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  
## S.F.Ratio perc.alumni Expend Grad.Rate  
## Accept.pct 0.10998188 -0.13210402 -0.40862232 -0.286971504  
## Elite10 -0.29349738 0.30259090 0.55977784 0.348732550  
## F.Undergrad 0.27970335 -0.22946222 0.01865162 -0.078773129  
## P.Undergrad 0.23253051 -0.28079236 -0.08356842 -0.257000991  
## Outstate -0.55482128 0.56626242 0.67277862 0.571289928  
## Room.Board -0.36262774 0.27236345 0.50173942 0.424941541  
## Books -0.03192927 -0.04020774 0.11240908 0.001060894  
## Personal 0.13634483 -0.28596808 -0.09789189 -0.269343964  
## PhD -0.13053011 0.24900866 0.43276168 0.305037850  
## Terminal -0.16010395 0.26713029 0.43879922 0.289527232  
## S.F.Ratio 1.00000000 -0.40292917 -0.58383204 -0.306710405  
## perc.alumni -0.40292917 1.00000000 0.41771172 0.490897562  
## Expend -0.58383204 0.41771172 1.00000000 0.390342696  
## Grad.Rate -0.30671041 0.49089756 0.39034270 1.000000000

# By looking at the correlation\_matrix, we can clearly see that variable Outstate and Grad.Rate have the highest correlation value of 0.5712.

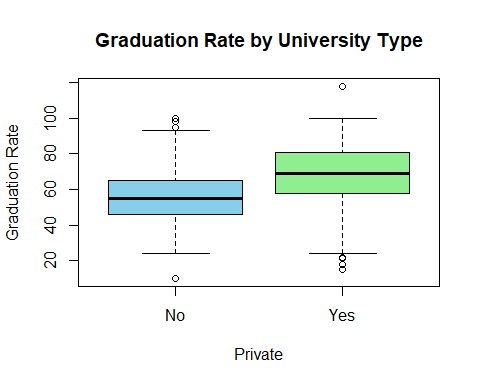
# Where as variables Books and Grad.Rate have the least correlation value of 0.001060894.

# No variable is highly correlated with Grad.Rate

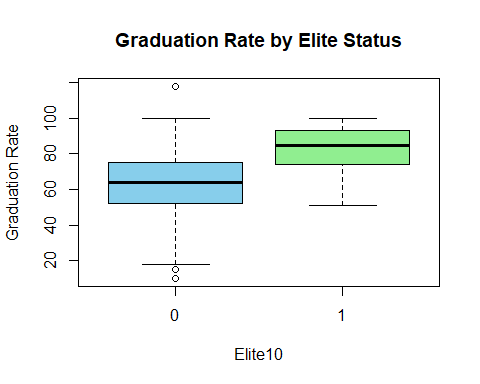
# Question c:

# Boxplot for graduation rates by university type

boxplot(Grad.Rate ~ Private, data = df, main = "Graduation Rate by University Type",  
 col = c("skyblue", "lightgreen"), ylab = "Graduation Rate")



# Boxplot for graduation rates by elite status  
boxplot(Grad.Rate ~ Elite10, data = df, main = "Graduation Rate by Elite Status",  
 col = c("skyblue", "lightgreen"), ylab = "Graduation Rate")



* By looking at the plots, we can see Graduation rate vary by status(elite vs non elite) as there is clear separation between two classes.
* By looking at the plots, we can see there is slight overlapping of Public and Private university. Hence Graduation rates vary by University Type but not as good as Status(elite vs non elite).

# Question D:

df$Private <- ifelse(df$Private == "Yes", 1, 0)

# Applying linear regression model

model <- lm(Grad.Rate ~ ., data=df)  
summary(model)

##   
## Call:  
## lm(formula = Grad.Rate ~ ., data = df)  
##   
## 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 \*\*   
## Accept.pct -1.811e+01 3.843e+00 -4.712 2.91e-06 \*\*\*  
## Elite10 4.017e+00 2.003e+00 2.005 0.04527 \*   
## F.Undergrad 6.809e-04 1.429e-04 4.767 2.24e-06 \*\*\*  
## P.Undergrad -1.956e-03 3.904e-04 -5.009 6.80e-07 \*\*\*  
## Outstate 1.235e-03 2.286e-04 5.401 8.88e-08 \*\*\*  
## Room.Board 1.667e-03 5.944e-04 2.805 0.00517 \*\*   
## Books -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   
## S.F.Ratio 1.003e-03 1.619e-01 0.006 0.99506   
## perc.alumni 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

# Question e:

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

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

vif(model)

## Private Accept.pct Elite10 F.Undergrad P.Undergrad Outstate   
## 2.739521 1.486633 1.687903 2.233117 1.643393 3.935059   
## Room.Board Books Personal PhD Terminal S.F.Ratio   
## 1.976762 1.115823 1.290983 3.917716 3.946581 1.909722   
## perc.alumni Expend   
## 1.672367 2.922643

# With the help of vif values, we can see that there is no multi-collinearity between variables as all variables have value < 10.

# Question f:

# Using backward selection method and forward selection method for variable selection

backward\_model <- step(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

summary(backward\_model)

##   
## Call:  
## lm(formula = Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad +   
## P.Undergrad + Outstate + Room.Board + Personal + PhD + perc.alumni +   
## Expend, data = df)  
##   
## 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

forward\_model <- step(model, direction="forward")

## 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

summary(forward\_model)

##   
## Call:  
## lm(formula = Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad +   
## P.Undergrad + Outstate + Room.Board + Books + Personal +   
## PhD + Terminal + S.F.Ratio + perc.alumni + Expend, data = df)  
##   
## 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 \*\*   
## Accept.pct -1.811e+01 3.843e+00 -4.712 2.91e-06 \*\*\*  
## Elite10 4.017e+00 2.003e+00 2.005 0.04527 \*   
## F.Undergrad 6.809e-04 1.429e-04 4.767 2.24e-06 \*\*\*  
## P.Undergrad -1.956e-03 3.904e-04 -5.009 6.80e-07 \*\*\*  
## Outstate 1.235e-03 2.286e-04 5.401 8.88e-08 \*\*\*  
## Room.Board 1.667e-03 5.944e-04 2.805 0.00517 \*\*   
## Books -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   
## S.F.Ratio 1.003e-03 1.619e-01 0.006 0.99506   
## perc.alumni 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

# By seeing the results for both the models, we can see backward selection model gave more significant variables than forward selection. Also adjusted R-squared is slightly better in backward selection model.

# Question g:

# As explained above, we can clearly see backward selection model performed better in terms of variable significance and adjusted R-squared. Hence we will choose backward selection model.

# We will use the backward\_model from now onwards for analysis.

final\_model <- lm(formula = Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad +   
 P.Undergrad + Outstate + Room.Board + Personal + PhD + perc.alumni +   
 Expend, data = df)  
summary(final\_model)

##   
## Call:  
## lm(formula = Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad +   
## P.Undergrad + Outstate + Room.Board + Personal + PhD + perc.alumni +   
## Expend, data = df)  
##   
## 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

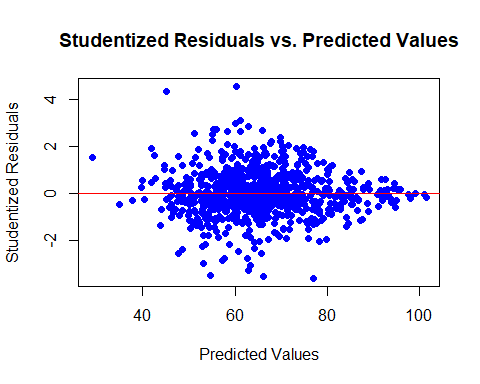
# Grad.Rate=Grad.Rate=48.40+4.77×Private−17.78×Accept.pct+4.02×Elite10+0.00066×F.Undergrad−0.00196×P.Undergrad+0.00122×Outstate+0.00153×Room.Board−0.00182×Personal+0.0842×PhD+0.306×perc.alumni−0.0004465×Expend

# Question h:

residuals <- rstudent(final\_model)  
predicted\_values <- predict(final\_model)

# Scatter plot residuals vs predicted\_values

plot(predicted\_values, residuals, main="Studentized Residuals vs. Predicted Values",  
 xlab="Predicted Values", ylab="Studentized Residuals", col="blue", pch=16)  
abline(h=0, col="red")



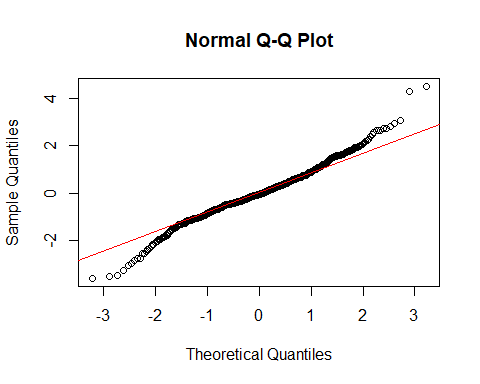
* By seeing at the scatter plot, we can see variation in studentized residuals first increased and then decreased.
* Hence our model didn’t performed better as constant variation is absent. Also there are some outliers which has studentized residuals > 3.

# Question i:

residuals <- residuals(final\_model)

# Create a normal probability plot

qqnorm(rstandard(final\_model), main="Normal Q-Q Plot")  
qqline(rstandard(final\_model), col="red")



* By seeing at the normal probability plot, we can see that there are many points which do not follow the line. Hence model does not fit well.

# Question j:

# Finding outliers

residuals\_standardized <- rstandard(final\_model)  
outliers\_indices <- which(residuals\_standardized > 3)

# Extract values with standardized residuals greater than 3

outliers\_values <- residuals\_standardized[outliers\_indices]

# Show the indices and values of outliers

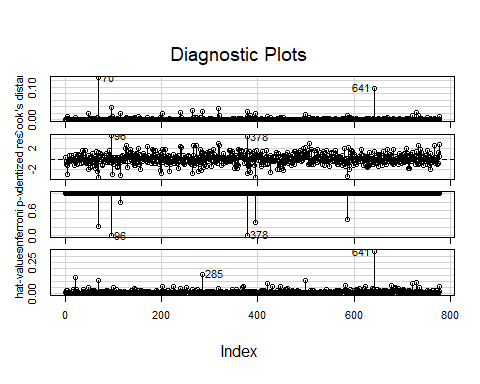
cat("Indices of outliers:", outliers\_indices, "\n")

## Indices of outliers: 96 318 378

cat("Values of outliers:", outliers\_values, "\n")

## Values of outliers: 4.509045 3.072519 4.28339

cooksd <- cooks.distance(final\_model)  
# Find indices of influential points with Cook's distance > 1  
influential\_indices <- which(cooksd > 1)  
  
influenceIndexPlot(final\_model)



* We can see three indexes which has outliers i.e. Standardized Residuals>3.

# Question k:

summary(final\_model)

##   
## Call:  
## lm(formula = Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad +   
## P.Undergrad + Outstate + Room.Board + Personal + PhD + perc.alumni +   
## Expend, data = df)  
##   
## 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

# The final model’s R-squared value is 0.4431, indicating that it explains approximately 44.31% of the variation in graduation rates among universities. The Adjusted R-squared is 0.4351.

# The model provides a moderate level of explanatory power for predicting graduation rates.

# The F-statistic is 55.33, and the p-value is very low (2.2e-16), indicating that the overall model is statistically significant.

# Question l:

# Based on the summary of the model,

The R-squared score of 0.4431 indicates that the model appears to fit the data quite well. This suggests that the model explains approximately 44.31% of the variation in graduation rates.

**Top 3 most Important Predictors:**

**1. perc.alumni** - A positive coefficient indicates that a higher percentage of alumni donations is associated with higher graduation rates.

**2. P.Undergrad** - A negative coefficient indicates that increasing the number of part-time undergraduate students coincides with a decrease in graduation rates.

**3. Outstate** - A positive coefficient indicates that a higher percentage of outstate is associated with higher graduation rates.

The variable ‘Private’ has a positive coefficient (4.770) and is statistically significant (p=0.00486). According to the model, private universities have greater graduation rates than public universities.

The “Elite10” variable has a positive coefficient, showing that top universities have higher graduation rates. However, the significance level is minimal (p-value = 0.04492).

# Problem 2:

# Question 1

# Creating an interaction model

interaction\_model <- lm(Grad.Rate ~ Elite10+ Accept.pct+ Outstate+ perc.alumni + Expend + Elite10 \* Accept.pct + Elite10 \* Outstate + Elite10 \* perc.alumni + Elite10 \* Expend, data = df)  
summary(interaction\_model)

##   
## Call:  
## lm(formula = Grad.Rate ~ Elite10 + Accept.pct + Outstate + perc.alumni +   
## Expend + Elite10 \* Accept.pct + Elite10 \* Outstate + Elite10 \*   
## perc.alumni + Elite10 \* Expend, data = df)  
##   
## 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

# The Elite101:perc.alumni interaction term is not statistically significant (p-value = 0.362485).

# Other interaction terms have p-values below 0.05, indicating significance.

# Question 2

interaction\_modelM2 <- lm(Grad.Rate ~ Elite10+ Accept.pct+ Outstate+ perc.alumni + Expend + Elite10 \* Accept.pct + Elite10 \* Outstate + Elite10 \* Expend, data = df)  
summary(interaction\_modelM2)

##   
## Call:  
## lm(formula = Grad.Rate ~ Elite10 + Accept.pct + Outstate + perc.alumni +   
## Expend + Elite10 \* Accept.pct + Elite10 \* Outstate + Elite10 \*   
## Expend, data = df)  
##   
## 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

# Grad.Rate=53.14+35.85×Elite10−15.05×Accept.pct+0.00232×Outstate+0.3334×perc.alumni−0.000951×Expend−21.64×Elite10×Accept.pct−0.002253×Elite10×Outstate+0.001057×Elite10×Expend

# Question 3

# The coefficient for Elite101 is 35.85, indicating that, on average, being an “Elite10” University is associated with an increase of 35.85 units in the Graduation Rate, holding other variables constant.

# The coefficient for Accept.pct is -15.05, suggesting that as the percentage of applicants accepted increases by one unit, the Graduation Rate is expected to decrease by 15.05 units, holding other variables constant.

# The coefficient for Outstate is 0.00232, suggesting that for each additional unit in out-of-state tuition, the Graduation Rate is expected to increase by 0.00232 units.

# The coefficient for perc.alumni is 0.3334, indicating that as the percentage of alumni who donate increases by one unit, the Graduation Rate is expected to increase by 0.3334 units.

# The coefficient for Expend is -0.000951, suggesting that for each additional unit in instructional expenditure per student, the Graduation Rate is expected to decrease by 0.000951 units, holding other variables constant.

# Question 4

# Load required libraries

library(caret)

## Loading required package: ggplot2

##   
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

## Loading required package: lattice

library(Metrics)

##   
## Attaching package: 'Metrics'

## The following objects are masked from 'package:caret':  
##   
## precision, recall

set.seed(123)  
# Define the control parameters for cross-validation  
ctrl <- trainControl(method = "cv", number = 5)  
  
final\_model\_cv <- train(Grad.Rate ~ Private + Accept.pct + Elite10 + F.Undergrad +   
 P.Undergrad + Outstate + Room.Board + Personal + PhD + perc.alumni +   
 Expend, data = df,method = "lm", trControl = ctrl)

# Making predictions

predictions\_M1 <- predict(final\_model\_cv)  
mape\_M1 <- mape(df$Grad.Rate, predictions\_M1)  
cat("MAPE for Model M1:", mape\_M1, "\n")

## MAPE for Model M1: 0.1786888

# Question 5:

interaction\_modelM2\_cv <- train(Grad.Rate ~ Elite10+ Accept.pct+ Outstate+ perc.alumni + Expend + Elite10 \* Accept.pct + Elite10 \* Outstate + Elite10 \* Expend,   
 data = df,method = "lm", trControl = ctrl)  
  
summary(interaction\_modelM2\_cv)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## 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

predictions\_M2 <- predict(interaction\_modelM2\_cv, newdata = df)  
  
mape\_M2 <- mape(df$Grad.Rate, predictions\_M2)  
cat("MAPE for Model M2:", mape\_M2, "\n")

## MAPE for Model M2: 0.1857689