Exercise 1

Advanced Methods for Regression and Classification

Kai Thilenius

10/14/2025

0. Load and show data

```r
data(College,package="ISLR")
str(College)
```

```
## 'data.frame':    777 obs. of  18 variables:
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : num  1660 2186 1428 417 193 ...
##  $ Accept     : num  1232 1924 1097 349 146 ...
##  $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: num  2885 2683 1036 510 249 ...
##  $ P.Undergrad: num  537 1227 99 63 869 ...
##  $ Outstate   : num  7440 12280 11250 12960 7560 ...
##  $ Room.Board : num  3300 6450 3750 5450 4120 ...
##  $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : num  2200 1500 1165 875 1500 ...
##  $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
##  $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : num  7041 10527 8735 19016 10922 ...
##  $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
```

```r
# Check for missing data
print(paste("Number of missing values:", sum(is.na(College))))
```

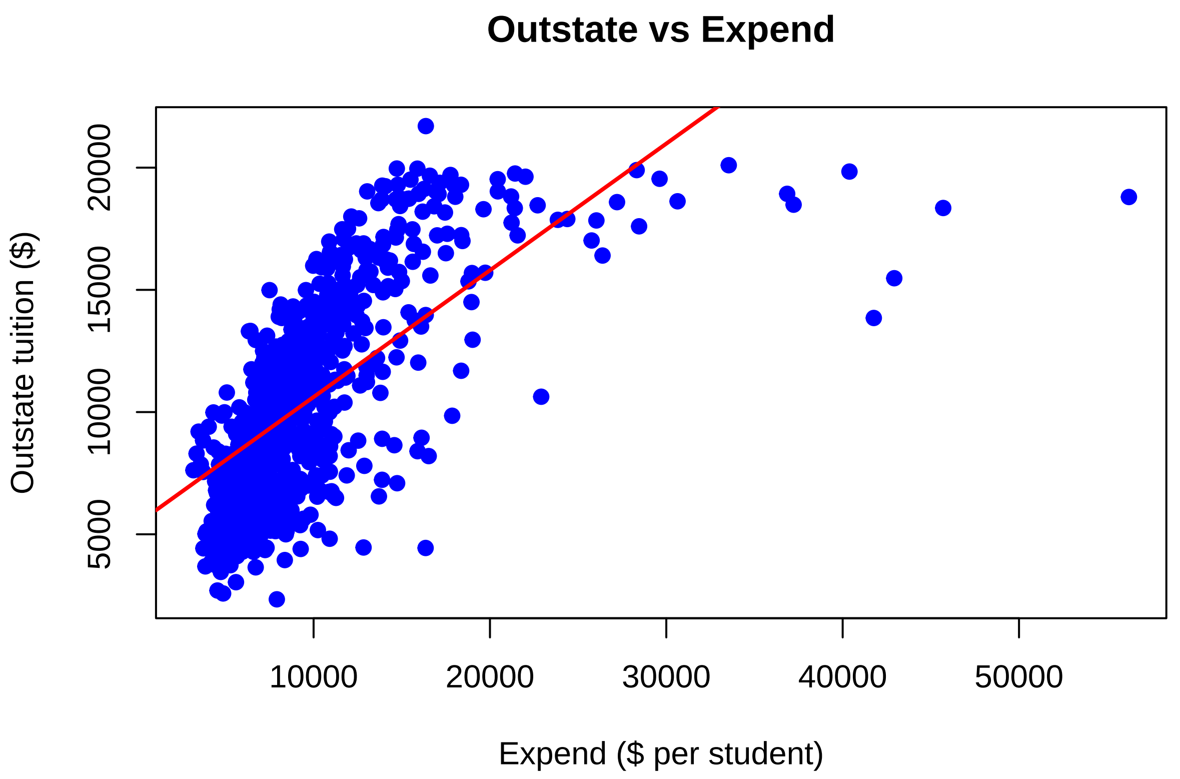
```
## [1] "Number of missing values: 0"
```

1. Predicting *Outstate* response by only using *Expend* as predictor

Outstate: Out-of-state tuition - how much a college charges students from outside the state

Expend: Instructional expenditure per student - how much the college spends on instruction for each student

```r
# linear regression based on those two only
lm1 <- lm(Outstate ~ Expend, data = College)

# Plot the data, and visualize the regression line
plot(College$Expend, College$Outstate,
     xlab = "Expend ($ per student)",
     ylab = "Outstate tuition ($)",
     main = "Outstate vs Expend",
     pch = 19, col = "blue")

abline(lm1, col = "red", lwd = 2)
```



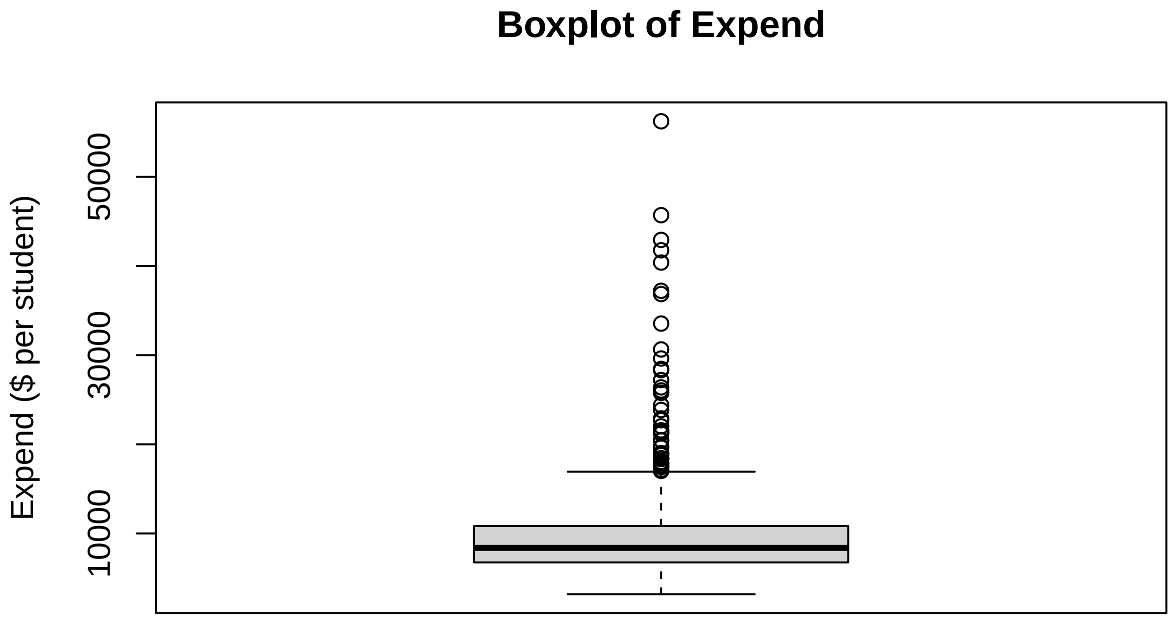
Conclusion:

- positive estimated slope - so the more the college pays per student, the higher the tuition
- BUT some "outliers" towards the right (way higher expends per student, than usual) give the predicted line a bias

2. More appropriate model against bias

As the plot before showed a few single points far away from the estimation, getting rid of those "outliers" would be the first idea to make the model "more appropriate". A boxplot of this one variable *Expend* helps visualizing this situation and also shows, that the **outliers are all above** the maxium. The common way of cutting both the low and the high quartile might be too drastic, so we **only take out the quartile of the highest expends**.

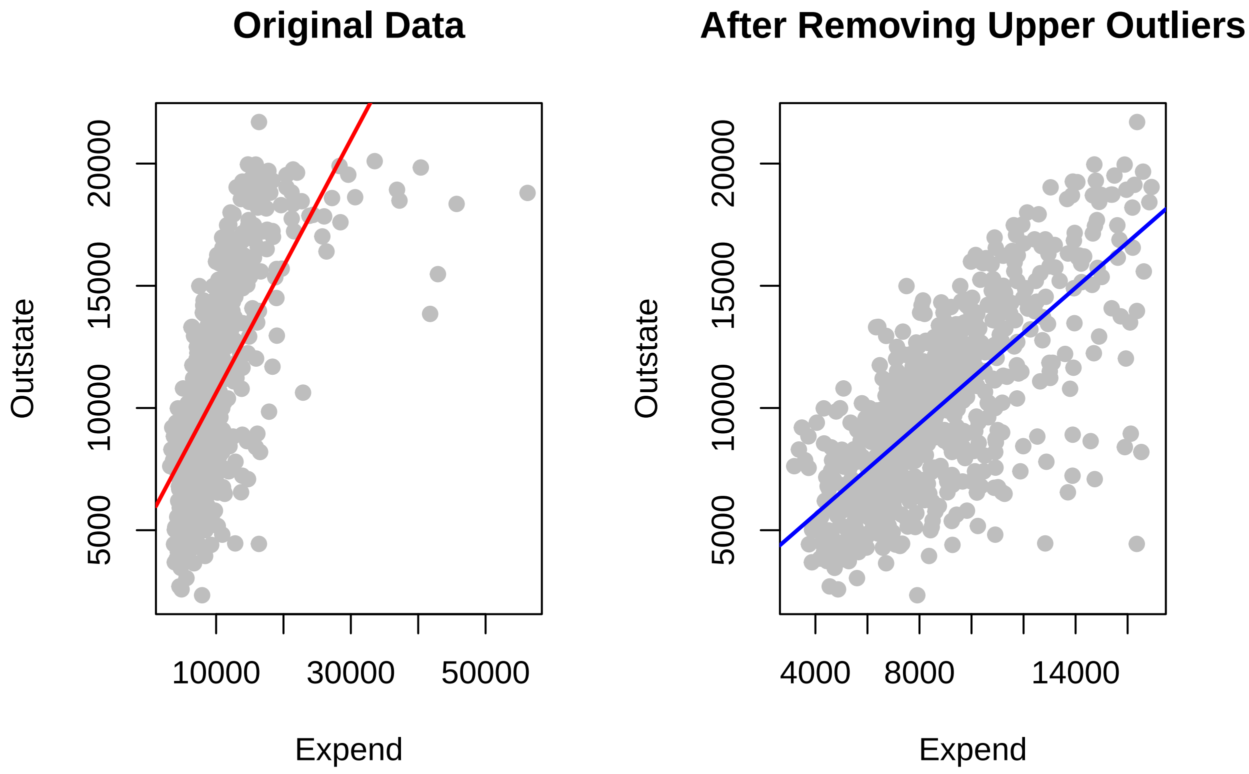
```r
# verifying outlier theory
boxplot(College$Expend,
        main = "Boxplot of Expend",
        ylab = "Expend ($ per student)")
```



```r
# Calculate quartiles and IQR
Q1 <- quantile(College$Expend, 0.25)
Q3 <- quantile(College$Expend, 0.75)
IQR <- Q3 - Q1

# Keep all rows except extreme high outliers
College_no_outliers <- subset(College, Expend < (Q3 + 1.5 * IQR))

# calculate new model after removing upper outliers
lm2 <- lm(Outstate ~ Expend, data = College_no_outliers)
```

Then we can compare both models visually and numerically (see table below). We seem to see a more appropriate 2nd model. We will continue using this “clened” dataset in the following questions:

```r
par(mfrow = c(1, 2))

# Original model
plot(College$Expend, College$Outstate,
     main = "Original Data",
     xlab = "Expend", ylab = "Outstate",
     pch = 19, col = "gray")
abline(lm1, col = "red", lwd = 2)

# After removing high outliers
plot(College_no_outliers$Expend, College_no_outliers$Outstate,
     main = "After Removing Upper Outliers",
     xlab = "Expend", ylab = "Outstate",
     pch = 19, col = "gray")
abline(lm2, col = "blue", lwd = 2)
```



```
par(mfrow = c(1, 1))

## Values copied in LateX table
# summary(lm1)$r.squared
# summary(lm2)$r.squared

# summary(lm1)$coefficients
# summary(lm2)$coefficients
```

3. Predict response on *Apps* with binary variable *Private*

Metric	Original Model	No-Outliers Model
R^2	0.4526	0.5156
Intercept	5433.51	1939.38
Slope (Expend)	0.518	0.927
p-value (Expend)	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$

Table 1: Comparison of linear regression models predicting *Outstate* from *Expend*, before and after removing upper outliers.

```r
lm3 <- lm(Apps ~ Private, data = College_no_outliers)
summary(lm3)
```

```
##
## Call:
## lm(formula = Apps ~ Private, data = College_no_outliers)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5497  -1170   -643    474  42364
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5729.9      225.3   25.43   <2e-16 ***
## PrivateYes   -4108.1      267.6  -15.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3281 on 727 degrees of freedom
## Multiple R-squared:  0.2448, Adjusted R-squared:  0.2438
## F-statistic: 235.7 on 1 and 727 DF,  p-value: < 2.2e-16
```

The intercept of 5729.9 here is the predicted average number of applications for colleges where *Private = "No"*. The *Private = "Yes"* signifies the difference in average applications between private and public colleges and is with -4108.1 negative, so: Private colleges=5729.9-4108.1=1621.8 applications (on average). As they both have small p-values, their differece is statistically significant.

4.Convert Private as variable with level +/- 1

The leveling the following ruleset is expected:

	-1	+1
No	1	0
Yes	0	1

Table 2: Your table caption here

With this transformation the intercept now represents the overall mean of the Data.

```r
College_no_outliers$Private_leveled <- ifelse(College_no_outliers$Private == "Yes", 1, -1)

table(College_no_outliers$Private, College_no_outliers$Private_leveled)
```

```
##
##         -1   1
##   No  212   0
##   Yes   0 517
```

```r
lm4 <- lm(Apps ~ Private_leveled, data = College_no_outliers)
summary(lm4)
```

```
##
## Call:
## lm(formula = Apps ~ Private_leveled, data = College_no_outliers)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5497  -1170   -643    474  42364
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3675.9      133.8   27.48   <2e-16 ***
## Private_leveled  -2054.0      133.8  -15.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3281 on 727 degrees of freedom
## Multiple R-squared:  0.2448, Adjusted R-squared:  0.2438
## F-statistic: 235.7 on 1 and 727 DF,  p-value: < 2.2e-16
```

5. Predict Apps response by all variables

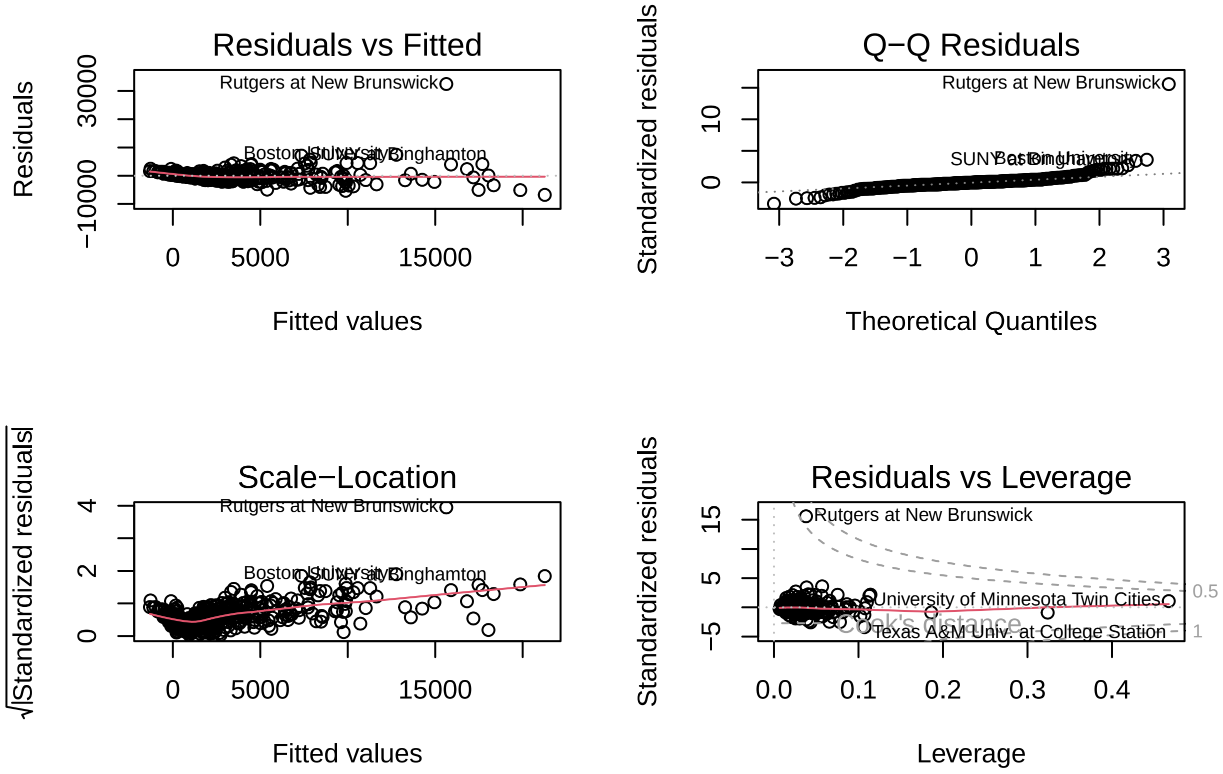
Doing that only makes sense with variables that make sense. This “content-wise” excludes:

- Accept: is determined by **Apps** and can only be affects after the application happened
- Enroll: also depends on the application process
- Personal: college internal spendings

The other variables might “content-wise” influence applicants by having an effect on the colleges reputation, popularity, costs, . . . and we get some good looking graphs Diagnostic Plots:

```r
set.seed(123)   # same randomness seed for reproducibility
n <- nrow(College_no_outliers)
train_index <- sample(1:n, size = 2/3 * n)
train <- College_no_outliers[train_index, ]
test <- College_no_outliers[-train_index, ]

lm5 <- lm(Apps ~ Private + Top10perc + Top25perc + F.Undergrad +
                P.Undergrad + Outstate + Room.Board + Books + PhD +
                Terminal + S.F.Ratio + perc.alumni + Expend + Grad.Rate,
              data = train)

# summary(lm5)
```

```r
par(mfrow = c(2, 2)) # arrange in a 2x2 grid
plot(lm5)
```



```r
# pairs(College_no_outliers[, c("Apps", "F.Undergrad", "Expend", "Grad.Rate", "Outstate")]) # for pairw
```

6. Same model, scaled variables

The standardized coefficients allow direct comparison of predictor importance:

```r
lm6 <- lm(Apps ~ scale(Top10perc) + scale(Top25perc) + scale(F.Undergrad) +
                      scale(P.Undergrad) + scale(Outstate) + scale(Room.Board) +
                      scale(Books) + scale(PhD) + scale(Terminal) +
                      scale(S.F.Ratio) + scale(perc.alumni) + scale(Expend) +
                      scale(Grad.Rate) + Private,
                data = train)
summary(lm6)
```

```
##
## Call:
## lm(formula = Apps ~ scale(Top10perc) + scale(Top25perc) + scale(F.Undergrad) +
##     scale(P.Undergrad) + scale(Outstate) + scale(Room.Board) +
##     scale(Books) + scale(PhD) + scale(Terminal) + scale(S.F.Ratio) +
##     scale(perc.alumni) + scale(Expend) + scale(Grad.Rate) + Private,
##     data = train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6789   -741    -48    574  32461
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3356.22     259.07  12.955  < 2e-16 ***
## scale(Top10perc)      -73.93     214.30  -0.345 0.730254
## scale(Top25perc)      310.82     209.18   1.486 0.137982
## scale(F.Undergrad)  3284.44     155.52  21.119  < 2e-16 ***
## scale(P.Undergrad)  -159.06     124.23  -1.280 0.201023
## scale(Outstate)        80.99     191.24   0.424 0.672118
## scale(Room.Board)     434.57     127.99   3.395 0.000743 ***
## scale(Books)           12.55     103.61   0.121 0.903643
## scale(PhD)            -93.99     186.76  -0.503 0.615014
## scale(Terminal)      -108.70     182.89  -0.594 0.552560
## scale(S.F.Ratio)      207.63     131.12   1.583 0.113985
## scale(perc.alumni)  -237.24     124.00  -1.913 0.056330 .
## scale(Expend)         456.31     181.28   2.517 0.012163 *
## scale(Grad.Rate)      396.66     125.56   3.159 0.001684 **
## PrivateYes           -554.89     347.82  -1.595 0.111312
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2125 on 471 degrees of freedom
## Multiple R-squared:  0.7389, Adjusted R-squared:  0.7311
## F-statistic:  95.2 on 14 and 471 DF,  p-value: < 2.2e-16
```

Because variables are standardized, we see that **F.Undergrad** has by far the **largest effect** on `Apps`, followed by **Room.Board**, **Grad.Rate**, and **Expend**.

7. RMSEs of 5. and 6. and their comparison

with the formula

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$$

we can compare the performace of the previous models.

```r
# --- Predictions for lm5 (unscaled model) ---
pred_train_lm5 <- predict(lm5, newdata = train)
pred_test_lm5  <- predict(lm5, newdata = test)

# --- Compute RMSE for lm5 ---
rmse_train_lm5 <- sqrt(mean((train$Apps - pred_train_lm5)^2))
rmse_test_lm5  <- sqrt(mean((test$Apps - pred_test_lm5)^2))

# --- Predictions for lm6 (scaled model) ---
pred_train_lm6 <- predict(lm6, newdata = train)
pred_test_lm6  <- predict(lm6, newdata = test)
```

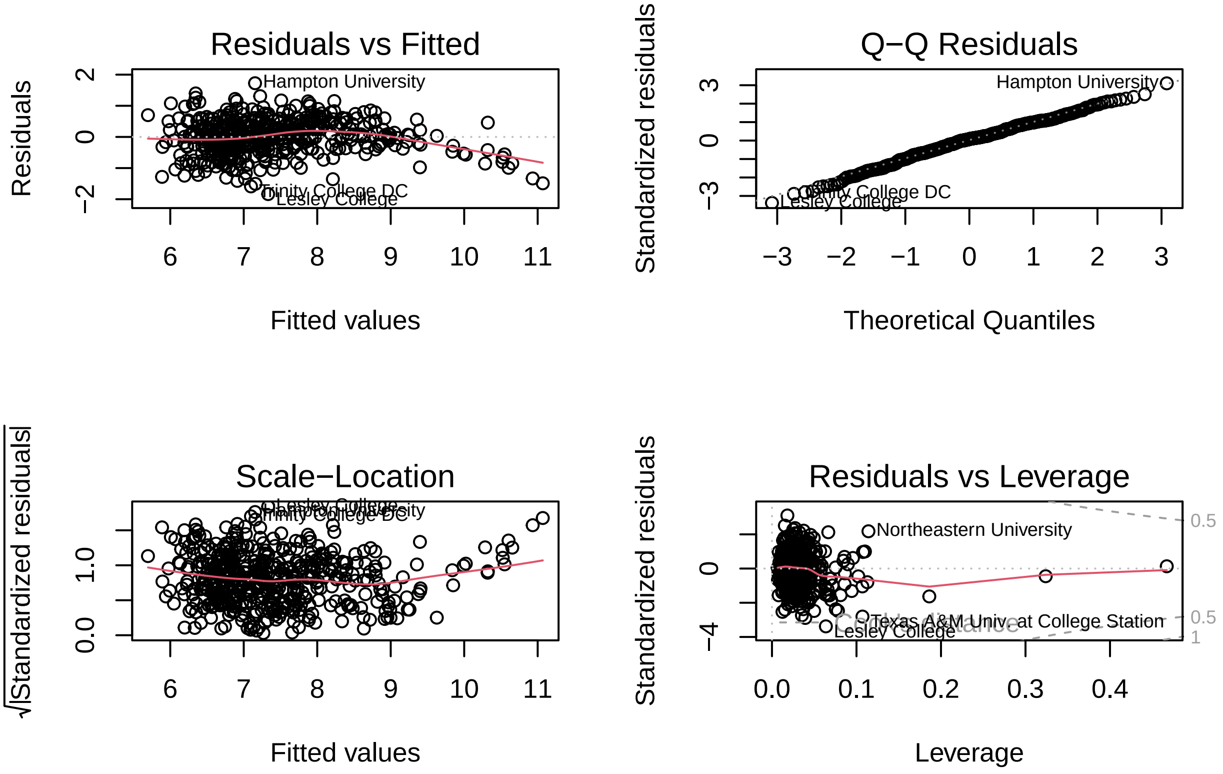
```
# --- Compute RMSE for lm6 ---
rmse_train_lm6 <- sqrt(mean((train$Apps - pred_train_lm6)^2))
rmse_test_lm6  <- sqrt(mean((test$Apps - pred_test_lm6)^2))

# --- Combine results into a table ---
rmse_results <- data.frame(
  Model = c("lm5 (unscaled)", "lm6 (scaled)"),
  RMSE_Train = c(rmse_train_lm5, rmse_train_lm6),
  RMSE_Test  = c(rmse_test_lm5, rmse_test_lm6)
)

rmse_results
```

```
##            Model RMSE_Train RMSE_Test
## 1 lm5 (unscaled)   2091.902  1430.797
## 2   lm6 (scaled)   2091.902  1430.797
```

Well … something went wrong obvioulsy. You should not trust (Chat GPT) on that one …

8. Lets see if we get any further with the log-transformed response

```
lm8 <- lm(log(Apps) ~ Private + Top10perc + Top25perc + F.Undergrad +
                        P.Undergrad + Outstate + Room.Board + Books + PhD +
                        Terminal + S.F.Ratio + perc.alumni + Expend + Grad.Rate,
              data = train)
par(mfrow = c(2, 2))
plot(lm8)
```



Diagnostic plots look good actually!

9. Yeah ... obviously we can not compare the RMSEs

Though i actually tried it lol. Here is a better solution probably - the Akaike's information criterion:

```
AIC(lm5, lm8)
```

```
##     df       AIC
## lm5 16 8842.9542
## lm8 16  834.3953
```

AIC of lm8 (which is the log transformed model) is way better than of lm5, so it is performing better!