Exercise 1

Advanced Methods for Regression and Classification

Kai Thilenius

10/14/2025

Load and show data

```
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 ...
```

```
# 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

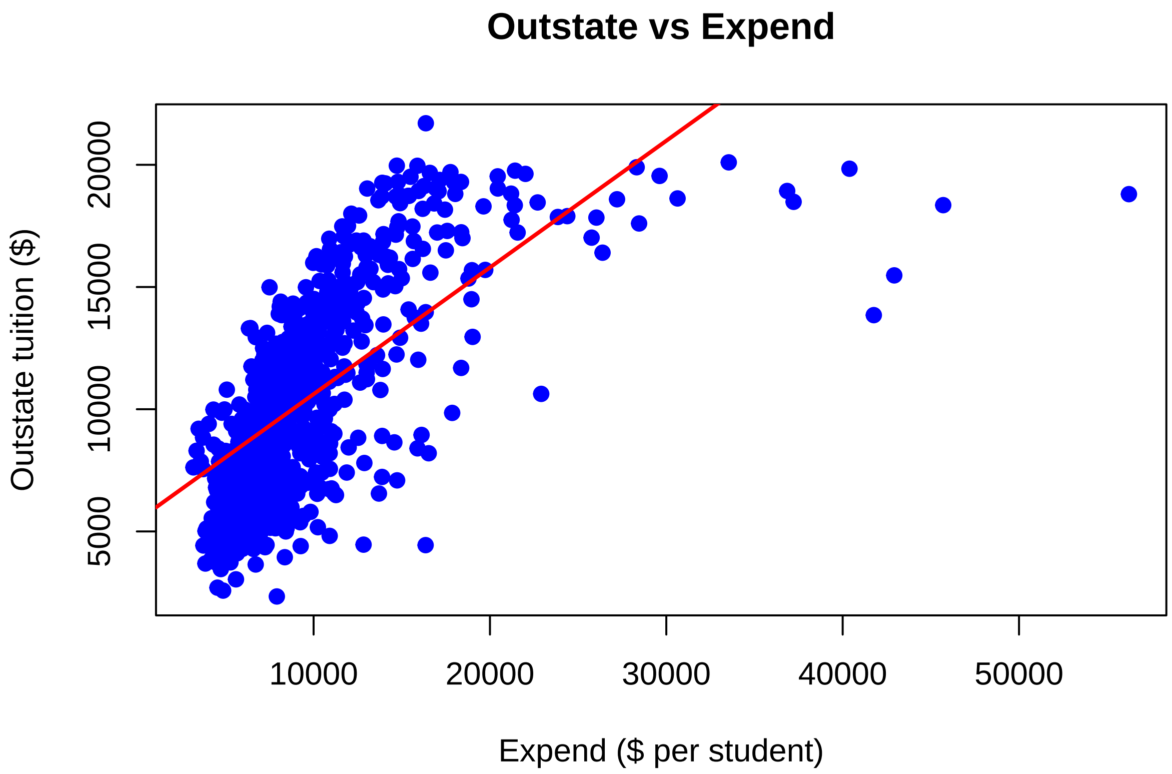
Outstate: Out-of-state tuition - how much a college charges students from outside the state

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)
```



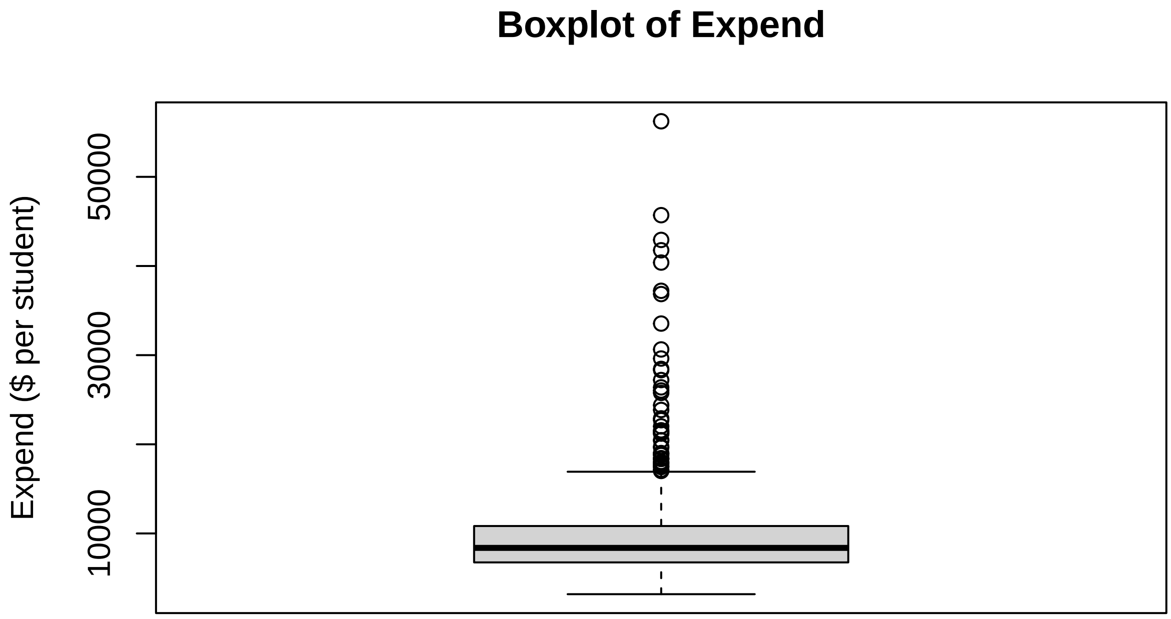
Conclusion:

- 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 agiant 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 helps visualizing this situation and also shows, that the outliers are all above the common area. 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)
```

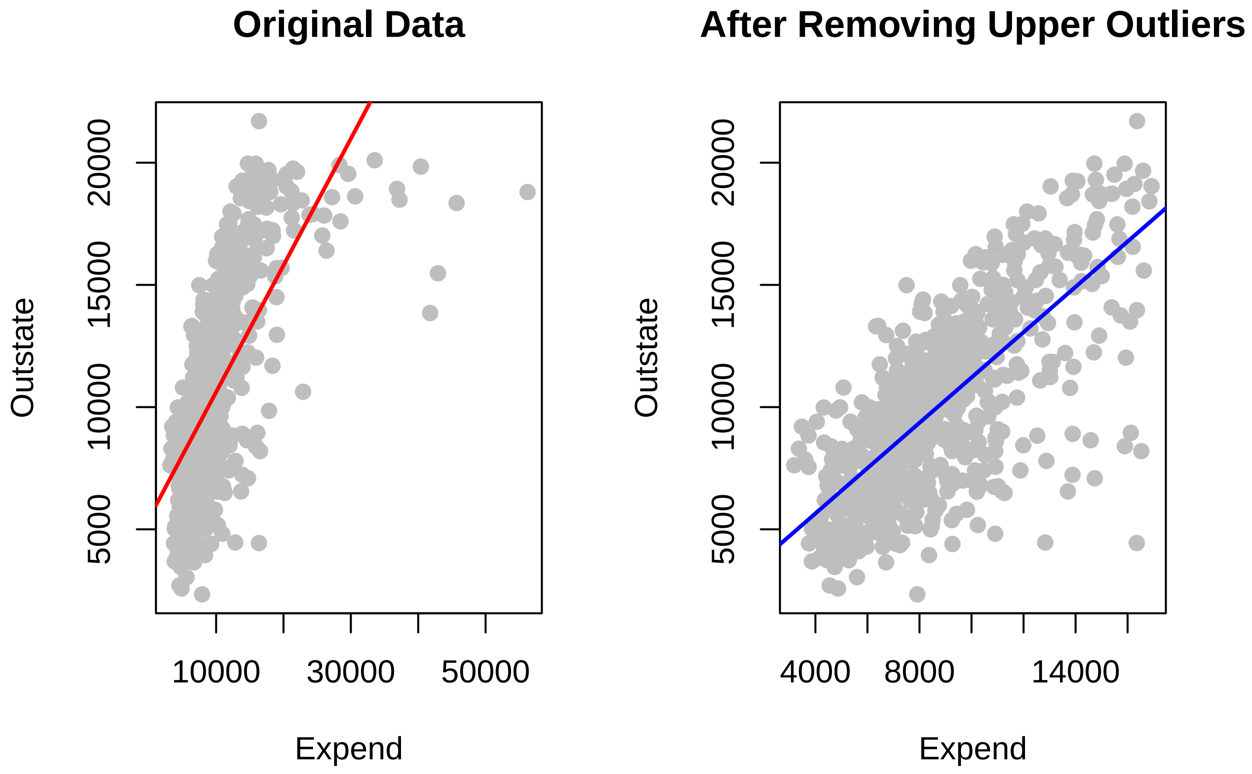
Then we can compare both models visually and numerically. We seem to see a more appropriate 2nd model:

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

# Original model
plot(College$Expend, College$Outstate,
     main = "Original Data",
     xlab = "Expend", ylab = "Outstate",
     pch = 19, col = "gray")
```

```r
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)
```



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

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

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

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.