ComS 573: Homework #1

Due on February 7, 2014

Professor De Brabanter at 10am

Josh Davis

Problem 1

Answer the following questions using the table below.

Observation	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Part A

Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

Solution

The equation for Euclidean distance is: $\text{dist} = \sqrt{(x_1 - 0)^2 + (x_2 - 0)^2 + (x_3 - 0)^2}$ Thus giving us:

Observation	Equation	Result
1	$\sqrt{0^2 + 3^2 + 0^2}$	3
2	$\sqrt{2^2 + 0^2 + 0^2}$	2
3	$\sqrt{0^2 + 1^2 + 3^2}$	3.16
4	$\sqrt{0^2 + 1^2 + 2^2}$	2.24
5	$\sqrt{-1^2 + 0^2 + 1^2}$	1.41
6	$\sqrt{1^2 + 1^2 + 1^2}$	1.73

Part B

Prediction with $k = 1$.

Solution

For $k = 1$, the prediction for our test point includes a single neighbor, thus it includes Observation 5 which is Green. Since the probability of being Green is 1, our test point should be Green as well.

Part C

Prediction with $k = 3$.

Solution

For $k = 3$, the prediction for our test includes three neighbors: Observation 5 (Green), Observation 6 (Red), and Observation 2 (Red). The probability for Green is then 1/3 and the probability of Red is 2/3. The test point should then be Red.

Part D

If the Bayes decision boundary (gold standard) is highly nonlinear in this problem, then would we expect the best value for k to be large or small?

Solution

If the Bayes decision boundary is highly nonlinear, then we would expect the best value for k to be small. This is because the larger the value of k, the less flexible our model becomes. The less flexible that it is, the more linear it gets.

Problem 2

Suppose we would like to fit a straight line through the origin, i.e., $Y_i = \beta_1 x_i + e_i$ with $i = 1, \ldots, n$, $\mathrm{E}[e_i] = 0$, and $\mathrm{Var}[e_i] = \sigma_e^2$ and $\mathrm{Cov}[e_i, e_j] = 0, \forall i \neq j$.

Part A

Find the least squares esimator for $\hat{\beta}_1$ for the slope β_1.

Solution

To find the least squares estimator, we should minimize our Residual Sum of Squares, RSS:

$$\begin{aligned} RSS &= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \\ &= \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 x_i\right)^2 \end{aligned}$$

By taking the partial derivative in respect to $\hat{\beta}_1$, we get:

$$\frac{\partial}{\partial \hat{\beta}_1}(RSS) = -2 \sum_{i=1}^{n} x_i (Y_i - \hat{\beta}_1 x_i) = 0$$

This gives us:

$$\begin{aligned} \sum_{i=1}^{n} x_i (Y_i - \hat{\beta}_1 x_i) &= \sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 \\ &= \sum_{i=1}^{n} x_i Y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \end{aligned}$$

Solving for $\hat{\beta}_1$ gives the final estimator for β_1:

$$\hat{\beta}_1 = \frac{\sum x_i Y_i}{\sum x_i^2}$$

Problem 2 continued on next page...

Part B

Calculate the bias and the variance for the estimated slope $\hat{\beta}_1$.

Solution

For the bias, we need to calculate the expected value $\mathrm{E}[\hat{\beta}_1]$:

$$\begin{aligned}
\mathrm{E}[\hat{\beta}_1] &= \mathrm{E}\left[\frac{\sum x_i Y_i}{\sum x_i^2}\right] \\
&= \frac{\sum x_i \mathrm{E}[Y_i]}{\sum x_i^2} \\
&= \frac{\sum x_i(\beta_1 x_i)}{\sum x_i^2} \\
&= \frac{\sum x_i^2 \beta_1}{\sum x_i^2} \\
&= \beta_1 \frac{\sum x_i^2 \beta_1}{\sum x_i^2} \\
&= \beta_1
\end{aligned}$$

Thus since our estimator's expected value is β_1, we can conclude that the bias of our estimator is 0.

For the variance:

$$\begin{aligned}
\mathrm{Var}[\hat{\beta}_1] &= \mathrm{Var}\left[\frac{\sum x_i Y_i}{\sum x_i^2}\right] \\
&= \frac{\sum x_i^2}{\sum x_i^2 \sum x_i^2} \mathrm{Var}[Y_i] \\
&= \frac{\sum x_i^2}{\sum x_i^2 \sum x_i^2} \mathrm{Var}[Y_i] \\
&= \frac{1}{\sum x_i^2} \mathrm{Var}[Y_i] \\
&= \frac{1}{\sum x_i^2} \sigma^2 \\
&= \frac{\sigma^2}{\sum x_i^2}
\end{aligned}$$

Problem 3

From ISLR: Chapter 2, Problem 8. Use the data set called *College* to answer the following questions.

Part A

Read in the data using **read.csv()**.

```
# Download data if it doesn't exist
if (!file.exists("./College.csv")) {
    download.file("http://www-bcf.usc.edu/~gareth/ISL/College.csv", destfile = "./College.csv")
}

# Read in Data
college <- read.csv("./College.csv", header = TRUE)
```

Part B

Look at the data and and remove the first column.

```
# View/edit the data fix(college)

# View the data View(college)

# Remove first column according to page 55
college <- college[, -1]
```

Part C

Part I

Use the **summary()** function to produce a numerical summary of the variables.

```
summary(college)

##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.0
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.0
##            Median : 1558   Median : 1110   Median : 434   Median :23.0
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.6
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.0
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.0
##    Top25perc      F.Undergrad     P.Undergrad        Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836   Max.   :21700
##    Room.Board       Books          Personal         PhD
##  Min.   :1780   Min.   :  96   Min.   : 250   Min.   :  8.0
##  1st Qu.:3597   1st Qu.: 470   1st Qu.: 850   1st Qu.: 62.0
##  Median :4200   Median : 500   Median :1200   Median : 75.0
```

Problem 3 continued on next page. . .

```
##  Mean   :4358   Mean   : 549   Mean   :1341   Mean   : 72.7  
##  3rd Qu.:5050   3rd Qu.: 600   3rd Qu.:1700   3rd Qu.: 85.0  
##  Max.   :8124   Max.   :2340   Max.   :6800   Max.   :103.0  
##     Terminal       S.F.Ratio     perc.alumni        Expend     
##  Min.   : 24.0   Min.   : 2.5   Min.   : 0.0   Min.   : 3186  
##  1st Qu.: 71.0   1st Qu.:11.5   1st Qu.:13.0   1st Qu.: 6751  
##  Median : 82.0   Median :13.6   Median :21.0   Median : 8377  
##  Mean   : 79.7   Mean   :14.1   Mean   :22.7   Mean   : 9660  
##  3rd Qu.: 92.0   3rd Qu.:16.5   3rd Qu.:31.0   3rd Qu.:10830  
##  Max.   :100.0   Max.   :39.8   Max.   :64.0   Max.   :56233  
##    Grad.Rate    
##  Min.   : 10.0  
##  1st Qu.: 53.0  
##  Median : 65.0  
##  Mean   : 65.5  
##  3rd Qu.: 78.0  
##  Max.   :118.0  
```

Problem 3 continued on next page. . .

Part II

Use the **pairs()** function to produce a scatterplot of the first 10 columns.

```
pairs(college[, 1:10])
```

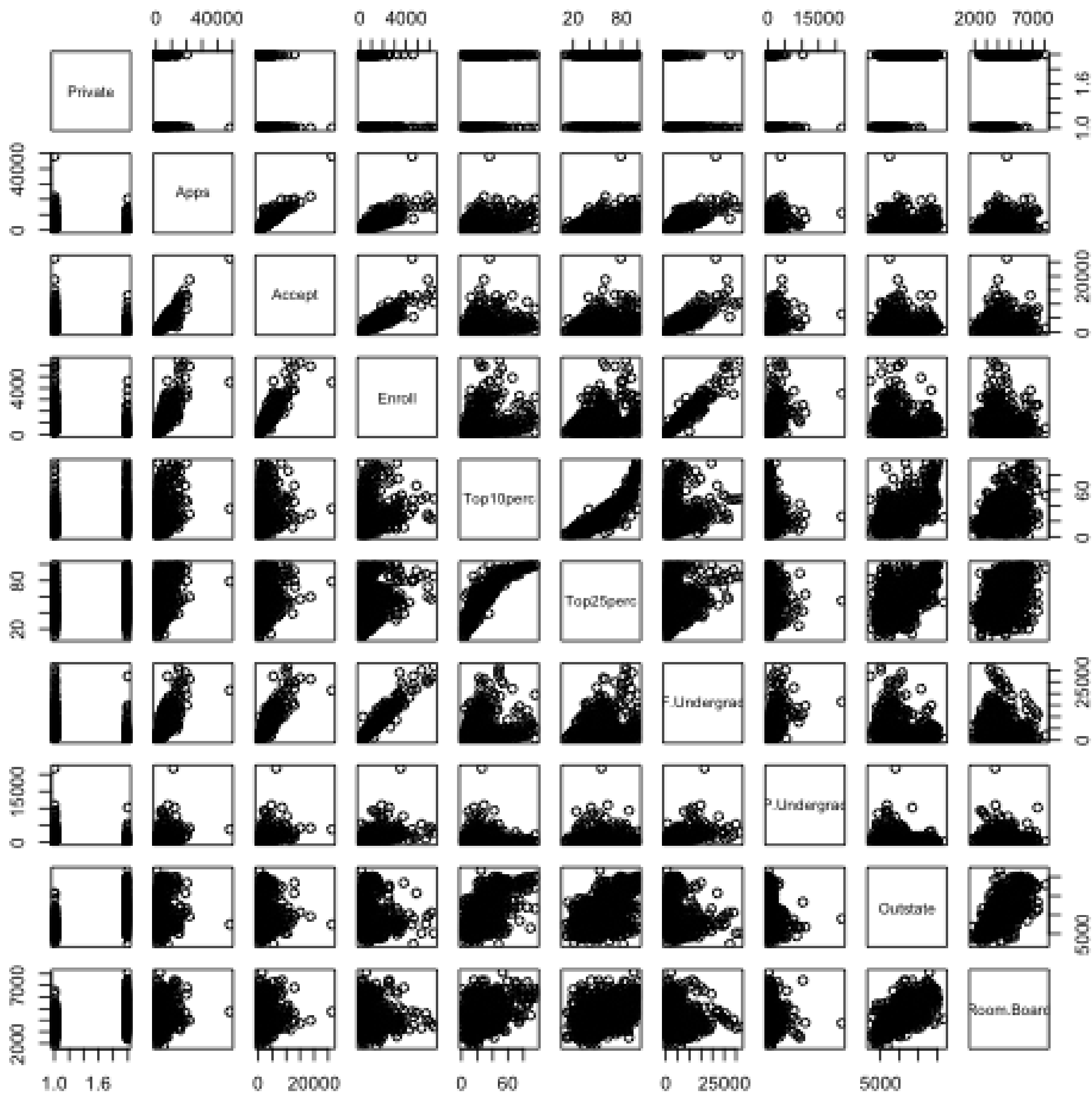

Figure 1: Part II plot.

Problem 3 continued on next page. . .

Part III

Use the **plot()** function to create boxplots of Outstate vs. Private.

```
plot(college$Private, college$Outstate, main = "Out of State Tuition vs Private Colleges",
    xlab = "Private College", ylab = "Out of State Tuition")
```

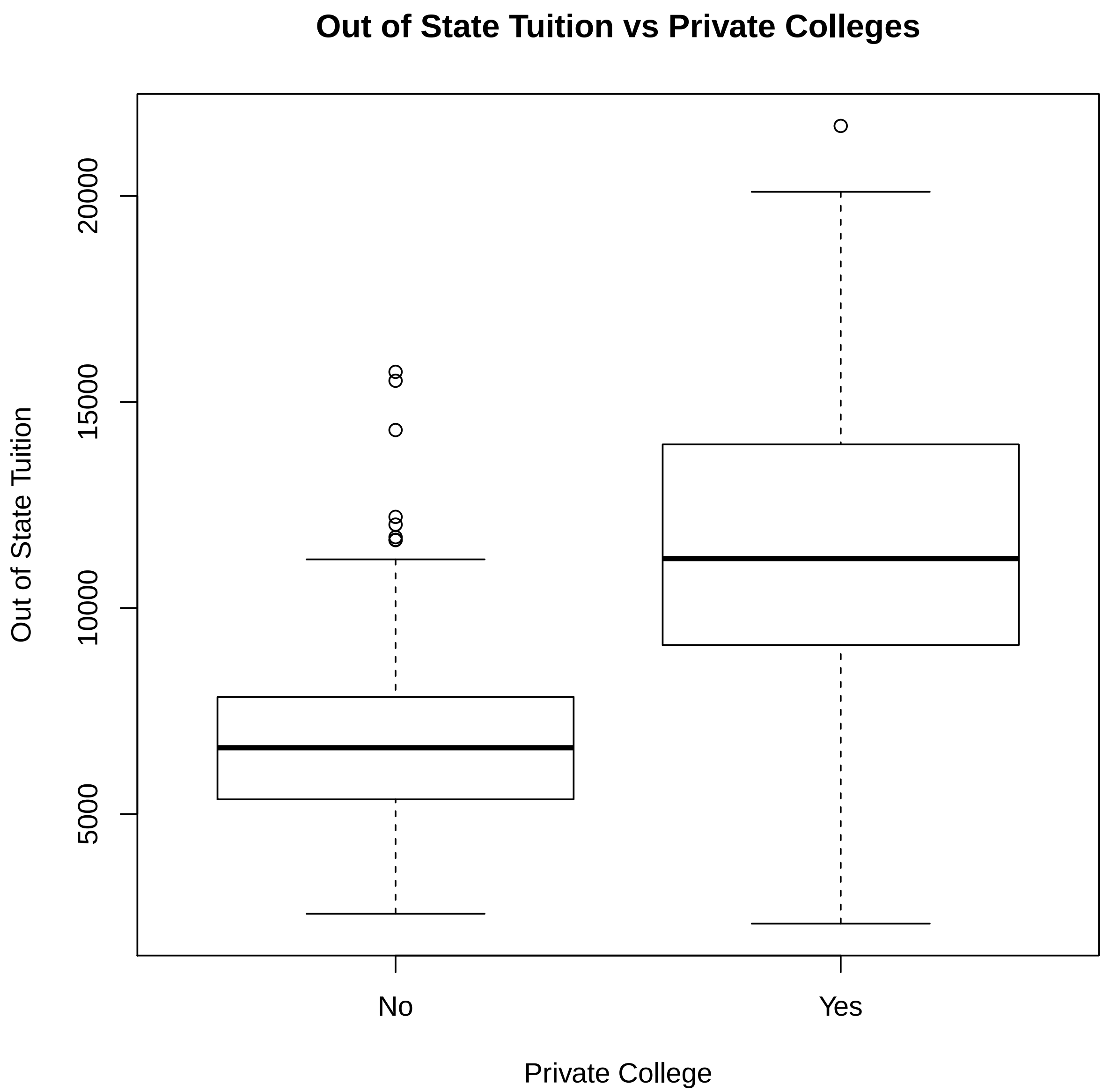

Figure 2: Part III plot.

Problem 3 continued on next page. . .

Part IV

Create a new qualitative variable called Elite. Use the **summary()** function and **plot()** function to display the info.

```
Elite <- rep("No", nrow(college))
Elite[college$Top10perc > 50] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college, Elite)

# Show number of elite vs non-elite colleges
summary(college)

##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.0
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.0
##            Median : 1558   Median : 1110   Median : 434   Median :23.0
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.6
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.0
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.0
##    Top25perc      F.Undergrad     P.Undergrad       Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836   Max.   :21700
##    Room.Board       Books         Personal         PhD
##  Min.   :1780   Min.   :  96   Min.   : 250   Min.   :  8.0
##  1st Qu.:3597   1st Qu.: 470   1st Qu.: 850   1st Qu.: 62.0
##  Median :4200   Median : 500   Median :1200   Median : 75.0
##  Mean   :4358   Mean   : 549   Mean   :1341   Mean   : 72.7
##  3rd Qu.:5050   3rd Qu.: 600   3rd Qu.:1700   3rd Qu.: 85.0
##  Max.   :8124   Max.   :2340   Max.   :6800   Max.   :103.0
##     Terminal       S.F.Ratio     perc.alumni       Expend
##  Min.   : 24.0   Min.   : 2.5   Min.   : 0.0   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.5   1st Qu.:13.0   1st Qu.: 6751
##  Median : 82.0   Median :13.6   Median :21.0   Median : 8377
##  Mean   : 79.7   Mean   :14.1   Mean   :22.7   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.5   3rd Qu.:31.0   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.8   Max.   :64.0   Max.   :56233
##    Grad.Rate      Elite
##  Min.   : 10.0   No :699
##  1st Qu.: 53.0   Yes: 78
##  Median : 65.0
##  Mean   : 65.5
##  3rd Qu.: 78.0
##  Max.   :118.0
```

Problem 3 continued on next page. . .

```r
# Show boxplot for Outstate vs Elite
plot(college$Elite, college$Outstate, main = "Out of State Tuition vs Elite Colleges",
    xlab = "Elite College", ylab = "Out of State Tuition")
```

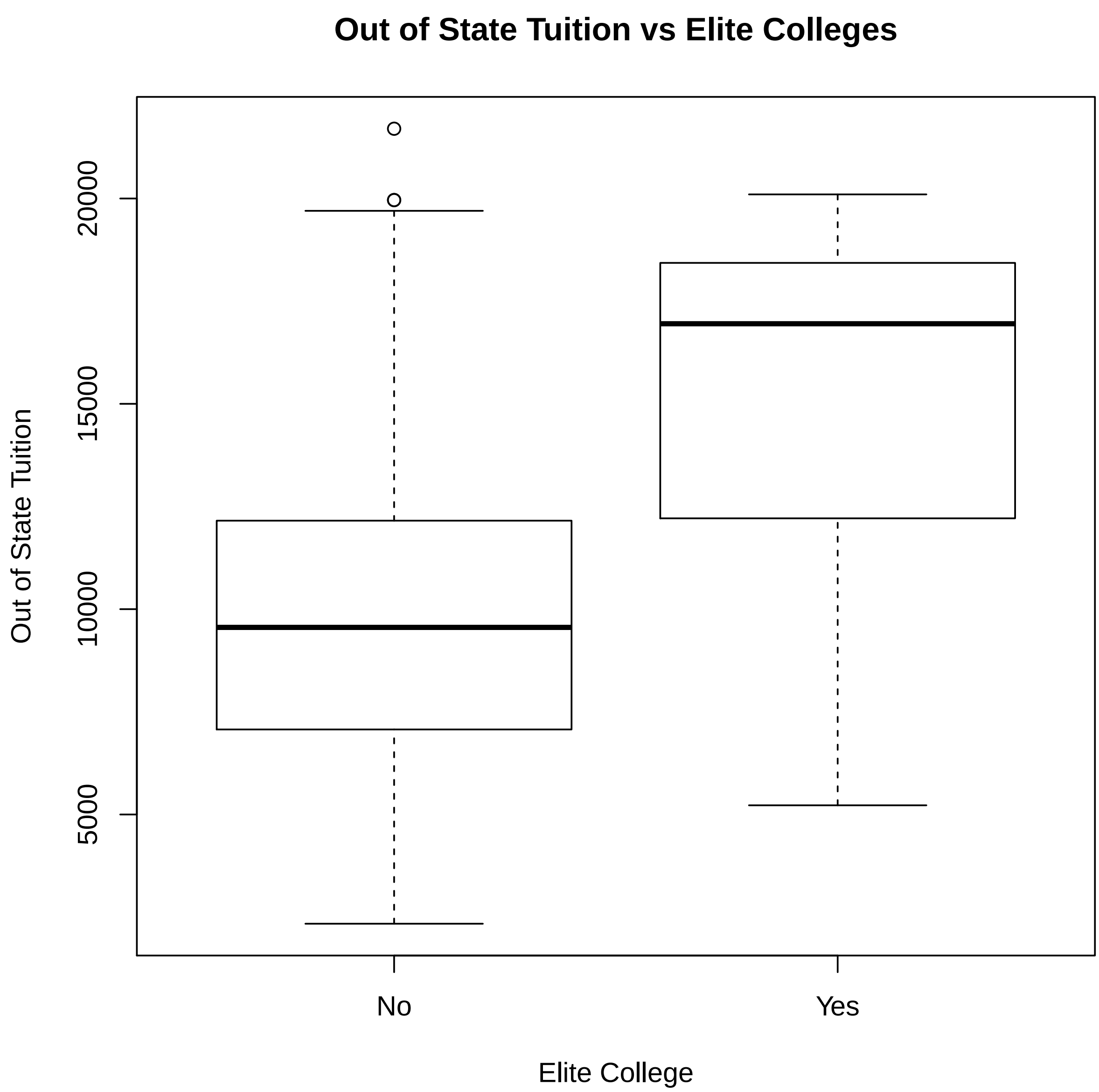

Figure 3: Part IV plot.

Problem 3 continued on next page. . .

Part V

Use the **hist()** function to produce histograms.

```
par(mfrow = c(2, 2))
hist(college$Apps, 20, main = "Number of College Applications Recieved", xlab = "Number of Applications
hist(college$Accept, 10, main = "Number of Applicants Accepted", xlab = "Number of Applicants Accepted"
hist(college$S.F.Ratio, 10, main = "Student to Faculty Ratio", xlab = "Student to Faculty Ratio")
hist(college$PhD, 10, main = "Percent of Faculty with a PhD", xlab = "Percent of Faculty with a PhD")
```

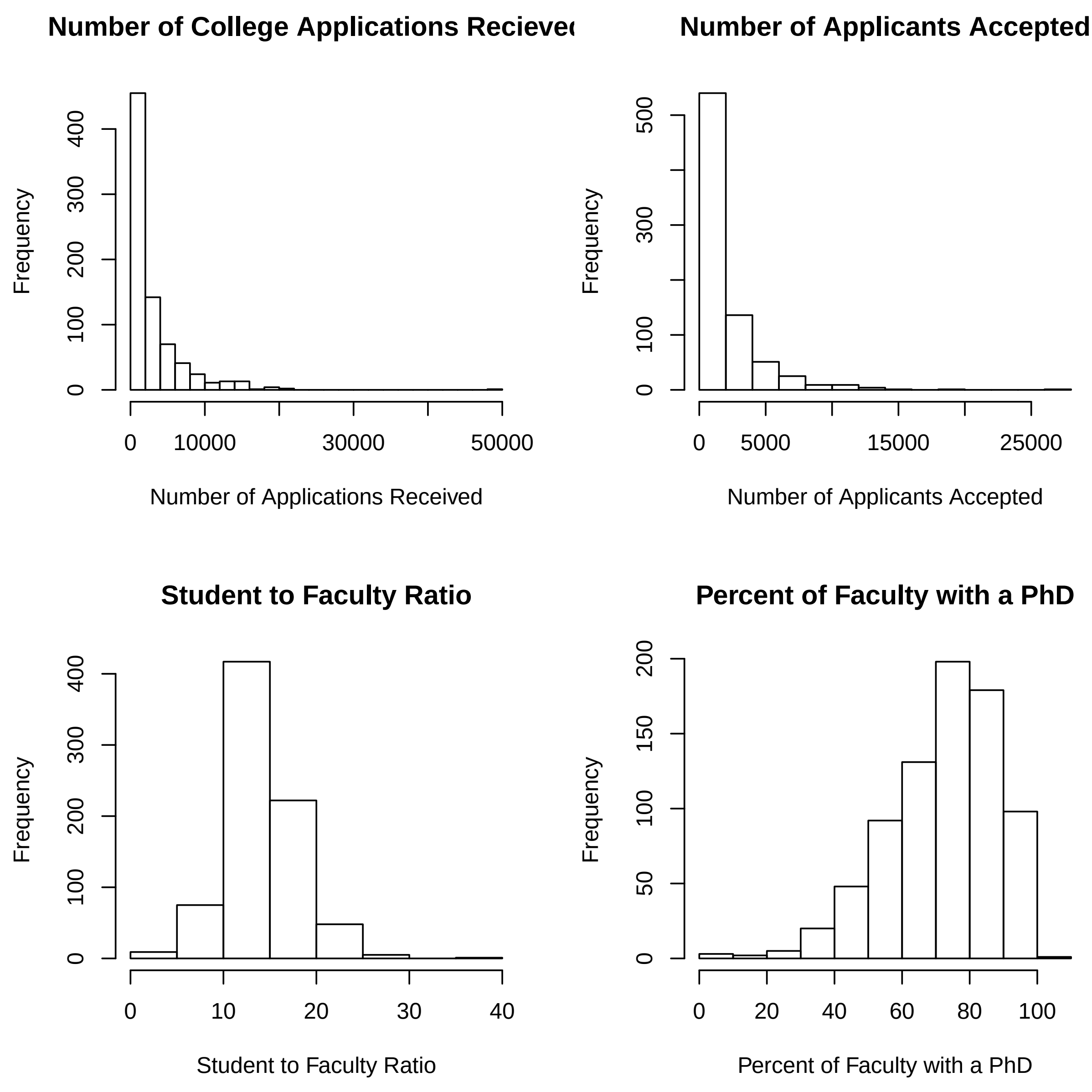

Figure 4: Part V plot.

Problem 3 continued on next page. . .

Part VI

Continue exploring and report your results.

Below is my exploration of the data. I wanted to see if there were a few different relationships between a few different sets of the data.

As you can see, there seems to be a relationship between number of applications accepted and the percent of students coming from the top 10% of high school class.

However, there doesn't seem to be a relationship between book costs vs the percent of professors with a PhD.

```
par(mfrow = c(2, 1))
df <- data.frame(x = college$PhD)
df$y <- college$Books

T <- lm(y ~ x, data = df)

plot(college$PhD, college$Books, main = "Book Costs vs Percent of PhD Professors",
    xlab = "Percent of PhD Professors", ylab = "Estimated Book Cost")
abline(T)

df <- data.frame(x = college$Accept)
df$y <- college$Top10perc

T <- lm(y ~ x, data = df)

plot(college$Accept, college$Top10perc, main = "Number of Applications Accepted vs Top 10% New Students
    xlab = "Applications Accepted", ylab = "Top 10% New Students")
abline(T)
```

Problem 3 continued on next page. . .

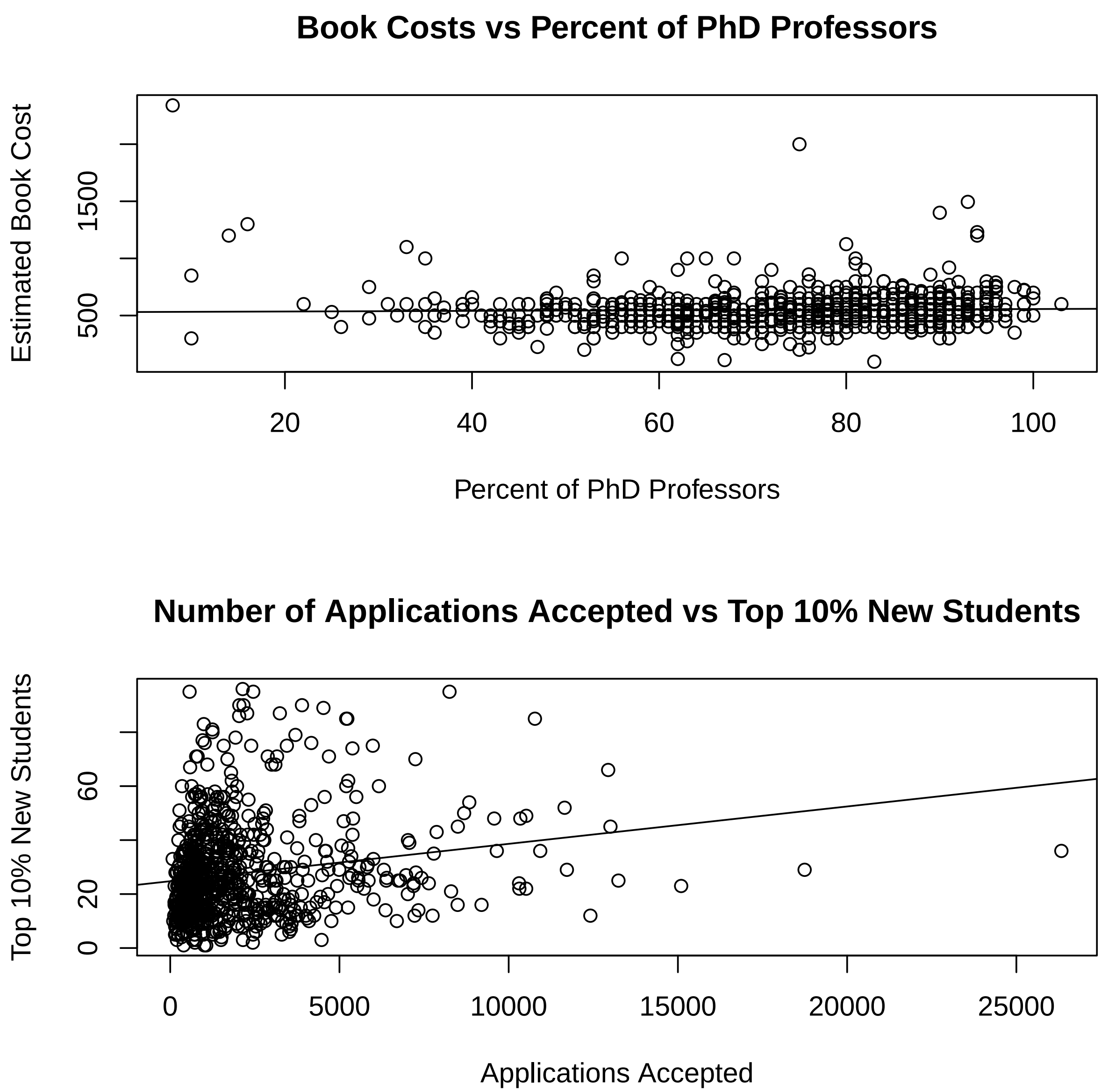

Figure 5: Part VI plot.

Problem 4

Consider the following equation of a straight line $Y_i = \beta_0 + \beta_1 x_i + e_i$ with $i = 1, \ldots, n$, $\mathrm{E}[e_i] = 0$, $\mathrm{Var}[e_i] = \sigma_e^2$, and $\mathrm{Cov}[e_i, e_j] = 0, \forall i \neq j$.

As in class, our estimator for β_1 is:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

which gives us:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1$$

as the two estimators for our line as given in the book and in lecture.

Part A

Calculate the bias for the estimator of the intercept $\hat{\beta}_0$.

Solution

In class, we determined that $\hat{\beta}_1$ is unbiased and thus $\mathrm{E}[\hat{\beta}_1] = \beta_1$.

Our expectation for $\hat{\beta}_0$ is thus:

$$\begin{aligned}
\mathrm{E}[\hat{\beta}_0] &= \mathrm{E}[\bar{y} - \hat{\beta}_1 \bar{x}] \\
&= \mathrm{E}[\bar{y}] - \mathrm{E}[\hat{\beta}_1 \bar{x}] \\
&= \frac{1}{n} \sum \mathrm{E}[y_i] - \mathrm{E}[\hat{\beta}_1 \bar{x}] \\
&= \frac{1}{n} \sum (\beta_0 + \beta_1 x_i) - \mathrm{E}[\hat{\beta}_1 \bar{x}] \\
&= \beta_0 + \frac{1}{n} \sum (\beta_1 x_i) - \mathrm{E}[\hat{\beta}_1 \bar{x}] \\
&= \beta_0 + \beta_1 \frac{1}{n} \sum (x_i) - \mathrm{E}[\hat{\beta}_1 \bar{x}] \\
&= \beta_0 + \beta_1 \bar{x} - \mathrm{E}[\hat{\beta}_1 \bar{x}] \\
&= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} \\
&= \beta_0
\end{aligned}$$

which shows that our estimator $\hat{\beta}_1$ is unbiased.

Problem 4 continued on next page. . .

Part B

Calculate the variance for the estimator of the intercept $\hat{\beta}_0$.

Solution

$$\begin{aligned}\text{Var}[\hat{\beta}_0] &= \text{Var}[\bar{y} - \hat{\beta}_1\bar{x}] \\ &= \text{Var}[\bar{y}] + \text{Var}[-\hat{\beta}_1\bar{x}] + 2\text{Cov}[\bar{y}, -\hat{\beta}_1\bar{x}]\end{aligned}$$

but by our assumption 3:

$$\begin{aligned}\text{Var}[\hat{\beta}_0] &= \text{Var}[\bar{y}] + \text{Var}[-\hat{\beta}_1\bar{x}] + 0 \\ &= \frac{1}{n^2}\sum(\text{Var}[y_i]) + \text{Var}[-\hat{\beta}_1\bar{x}] \\ &= \frac{n\sigma^2}{n^2} + \text{Var}[-\hat{\beta}_1\bar{x}] \\ &= \frac{\sigma^2}{n} + \text{Var}[\hat{\beta}_1\bar{x}] \\ &= \frac{\sigma^2}{n} + \bar{x}^2\text{Var}[\hat{\beta}_1] \\ &= \frac{\sigma^2}{n} + \bar{x}^2\text{Var}\left[\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}\right] \\ &= \frac{\sigma^2}{n} + \bar{x}^2\left(\frac{1}{\sum(x_i - \bar{x})^2}\right)^2 \text{Var}\left[\sum(x_i - \bar{x})(y_i - \bar{y})\right] \\ &= \frac{\sigma^2}{n} + \bar{x}^2\left(\frac{1}{\sum(x_i - \bar{x})^2}\right)^2 \left(\sum(x_i - \bar{x})^2\right)(\text{Var}[y_i - \bar{y}]) \\ &= \frac{\sigma^2}{n} + \bar{x}^2\left(\frac{1}{\sum(x_i - \bar{x})^2}\right)^2 \left(\sum(x_i - \bar{x})^2\right)\sigma^2 \\ &= \frac{\sigma^2}{n} + \bar{x}^2\left(\frac{1}{\sum(x_i - \bar{x})^2}\right)\sigma^2 \\ &= \frac{\sigma^2}{n} + \frac{\bar{x}^2\sigma^2}{\sum(x_i - \bar{x})^2} \\ &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)\end{aligned}$$

which agrees with equations 3.8 on page 66 of the textbook.