# 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: 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 $i = 1, \dots, n$, $\mathrm{E}[e_i] = 0$, 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 $\beta_1$.

## 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 $\beta_1$:

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

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 $\beta_1$, 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

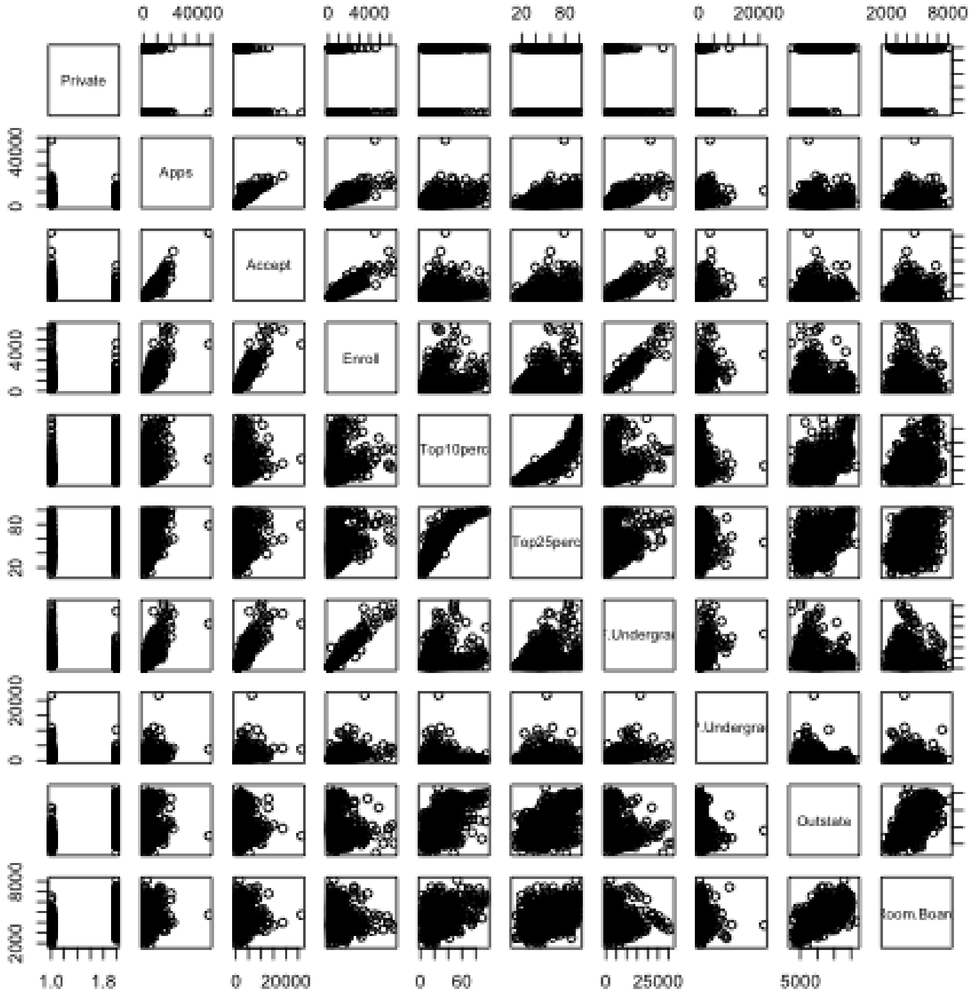


Figure 1: Pairs plot from Part II

Problem 3 continued on next page...

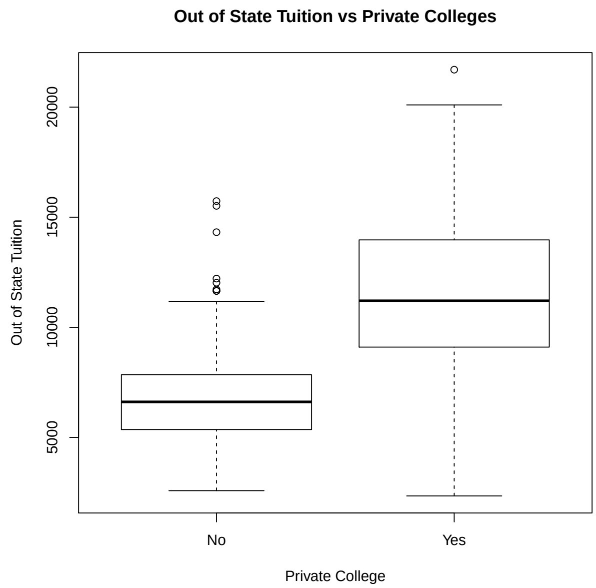


Figure 2: Boxplot of Outstate vs Private in Part III

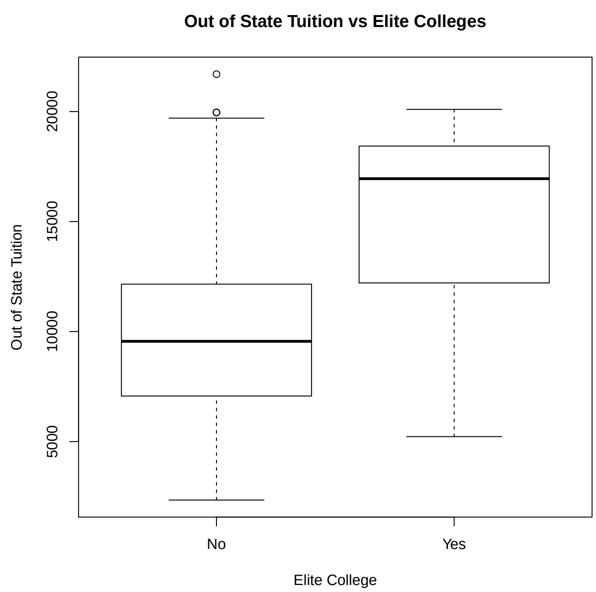


Figure 3: Boxplot of Elite colleges vs Outstate in Part IV

Problem 3 continued on next page. . .

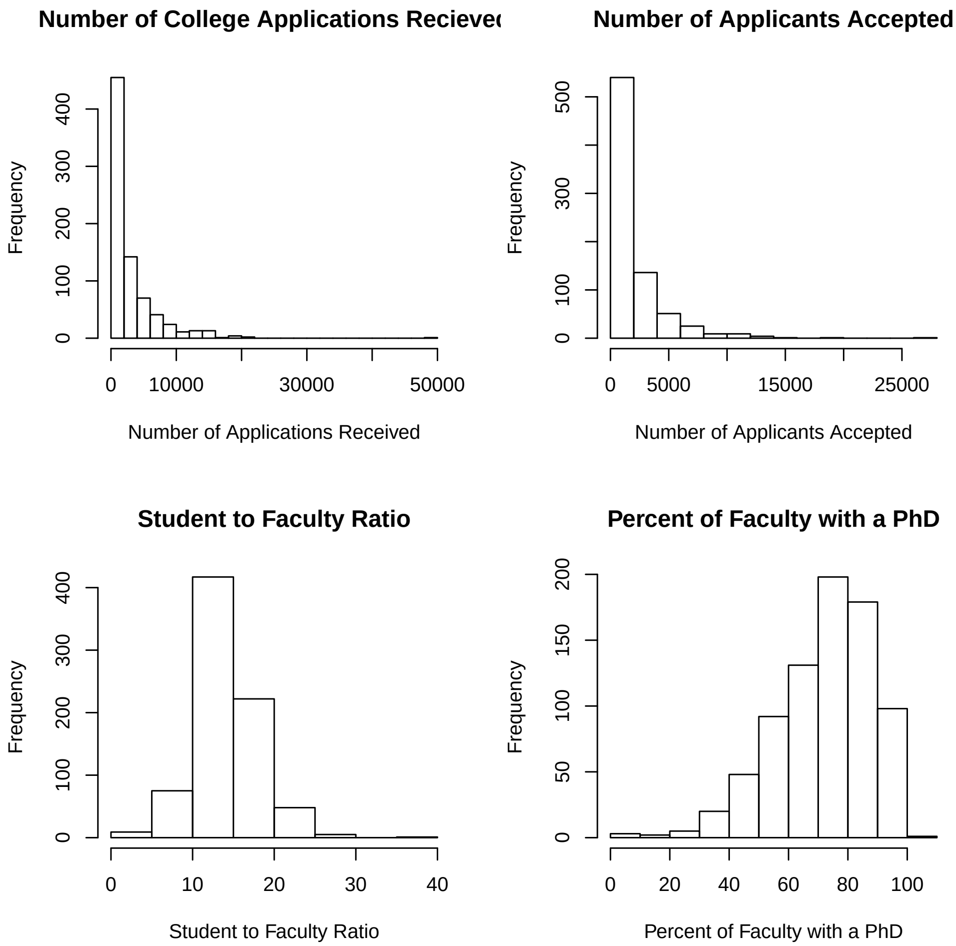


Figure 4: Various histograms from Part V

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.

Problem 3 continued on next page. . .

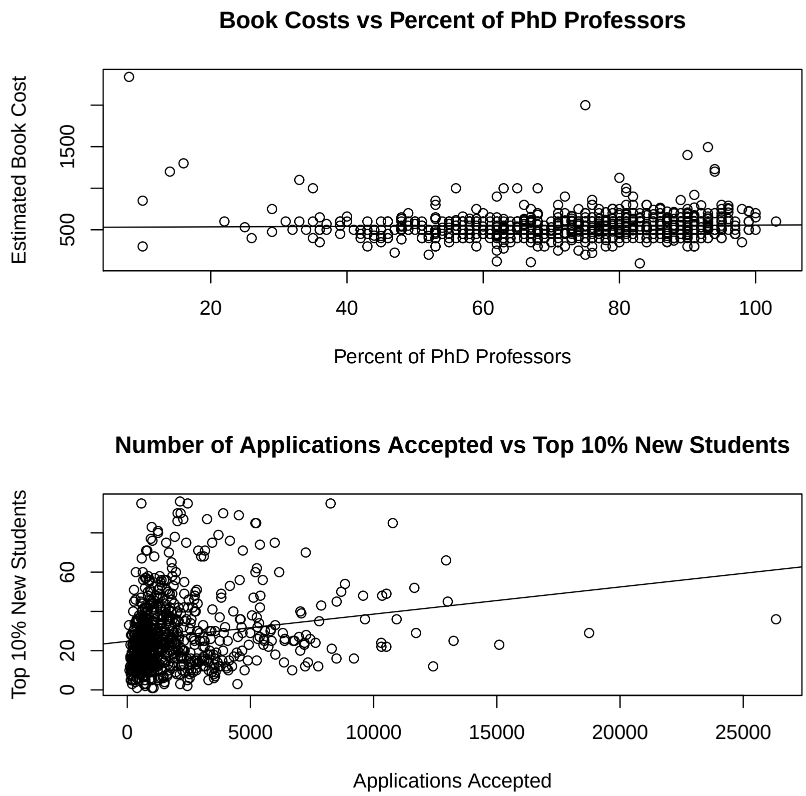


Figure 5: Part VI

# 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 $\beta_1$ 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

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

but by our assumption 3:

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

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