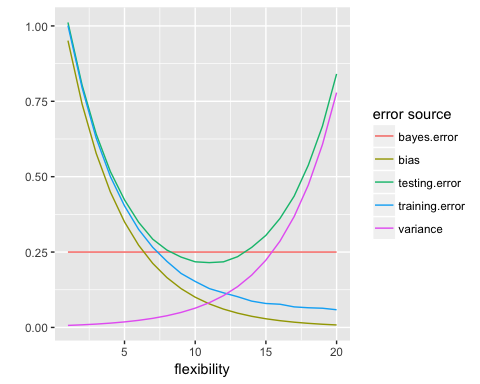
Week 1 Assignment

David Russo

1/15/2017

* 2.4 #1
  1. A flexible statistical model would perform **better** than an inflexible statistical model when *n* is extremely large and *p* is small. With a large sample size and small number of predictors, we are less likely to overfit the data. Furthermore, flexible models will have lower bias than inflexible models which will contribute to a lower overall test error.
  2. A flexible statistical model would perform **worse** than an inflexible statistical model when *n* is small and *p* is extremely large. Using a flexible model with a small sample size would introduce a large amount of variance due to overfitting.
  3. A flexible statistical model would perform **better** than an inflexible statistical model when the relationship between the predictors and the response is highly non-linear. Flexibile statistical models are better equipped to handle non-linearities than inflexible statistical models. Inflexible statistical models will often only detect linear relationships between predictors and the response.
  4. A flexible statistical model would perform **worse** than an inflexible statistical model when the variance of the error terms is high. If the variance of the error terms is high, a flexible model will be likely to overfit the data, resulting in higher variance and higher test error.
* 2.4 #2
  1. Regression setting; inference; n = 500, p = profit, number of employees, industry
  2. Classification setting; prediction; n = 20; p = price charged for the product, marketing budget, competition price
  3. Regression setting; prediction; n = 52; p = % change in US market, % change in British market, % change in German market
* 2.4 #3
  1. 
     1. The bias decreases with model flexibility because models that are more flexible are able to better fit the training data. As the fit improves, the bias decreases.
     2. The variance increases with model flexibility because more flexible models tend to overfit the training data. When the model is applied to new data that wasn't used to train the model, the model will have a high degree of variation. This is because the highly flexible model can only model the training data, and not new data.
     3. The training error decreases with model flexibility because more flexible models are able to model the training data well. As flexibility increases, data overfitting increases which decreases bias because the model is able to predict the training data well.
     4. The testing error has a U-shape because it demonstrates the bias-variance trade off. As model flexibility increases, bias decreases but variance increases. The minimum point of the U-shape represents the degree of flexibility that yields the best trade off between bias and variance in terms of predictive ability on new data. Models that are not flexible enough will have high bias but low variance when tested on new data. Models that are too flexible will overfit the data and have low bias but high variance when tested on new data.
     5. The Bayes error is a horizontal line because it is constant. It represents the true noise in the data that cannot be captured by the model.
* 2.4 #4
  1. Classification
     1. One example of a classification problem is predicting whether or not a customer will default on their mortgage based on their personal and financial data. The response is binary: 0 if they did not default and 1 if they did. The predictors are salary in dollars, down payment in dollars, age in years, gender, and savings account balance in dollars. The goal of this problem is prediction because the bank wants to make wise investment decisions and not give a mortgage to a risky customer. The bank doesn't necessarily care what causes a customer to default on their mortgage, but rather to be able to accurately predict if a customer will default on their mortgage.
     2. Another example of a classification problem is determining which health related attributes contribute to a patient's lung cancer status. The response is binary: 0 if they do not have lung cancer and 1 if they do have lung cancer. The predictors are age in years, gender, healthy eating status, exercise level, smoker status, and alcohol usage status. The goal of this problem is inference because a doctor will want to identify traits and habits that could lead to having lung cancer. It won't be useful to predict whether or not a given patient has cancer based on their attributes, but rather to instruct them to adopt different a different lifestyle to reduce their risk for cancer.
     3. A final example of a classification problem is determining whether or not an airplane part will fail within the first three months of usage. The response is binary: 0 if the part will not fail and 1 if the part will fail. The predictors are plant where the part was manufactured and the amount of time it has been sitting on the shelf. The goal of this problem is prediction because the airplane operator will not be concerned with why the part is failing but rather whether or not it will fail.
  2. Regression
     1. One example of a regression problem is predicting the amount of money that a customer will spend at a store based on their attributes. The response is dollars spent and the predictors are the gender of the customer, the occupation of the customer, the age of the customer, and the number of times that the customer goes shopping per month. The goal of the problem is inference so that the store can better understand how to market to their target audience.
     2. Another example of a regression problem is predicting how many pounds a patient will lose while participating in a diet and exercise regiment being offered by a fitness club. If the customer doesn't lose their target amount of weight, they will receive a portion of their money back. The response is weight loss in pounds and the predictors are height, current weight, gender, and daily activity level. The goal of the problem is prediction because the fitness club does not want to put risky customers on their diet and exercise regiment. Their model must accurately predict weight loss or else they will lose money.
     3. A final example of regression is predicting the amount of money for which a house will sell. The response is sale price in dollars and the predictors are square feet of the house, number of bathrooms, number of bedrooms, and distance from the nearest school. The goal of this problem is prediction because the seller wants to have accurate estimates of how much their house will sell for. They will not necessarily care what features contribute to a higher selling price.
  3. Cluster Analysis
     1. One example of a cluster analysis is an animal adoption agency wants to classify their animals into 4 categories based on their attributes: good for a family with young kids, good for a family without young kids, good for young individuals/couples with no kids, and good for older individuals/couples with no kids. The animals could then be marketed in such a way that they would be more likely to be adopted by a fitting owner. This is a clustering problem because there aren't existing category labels for the animals on which to train a model.
     2. Another example of a cluster analysis is a marketing company is trying to segment their audience into distinct groups to provide targeted advertising. This is a clustering problem because there aren't pre-existing labels for their target audience and they must instead be determined based on the attributes of their audience.
     3. A final example of a cluster analysis is determining which category a given investor falls into: risk averse, risk neutral, and willing to take investment risks. With this information, the bank could offer targeted services to a given individual based on their attributes. This is a clustering problem because there are no pre-existing labels and hence no training data on which to train a model.
* 2.4 #8
  1. Set working directory and load data

workDir <- "/Users/davidrusso/Documents/Classes/Applied Data Analysis/Assignments"  
  
college <- read.csv(paste(workDir, "College.csv", sep = "/"), header = TRUE)

* 1. Explore the data

# assign row names to the dataset college  
rownames(college) <- college[, 1]  
college[1:6, 1:6]

## X Private Apps  
## Abilene Christian University Abilene Christian University Yes 1660  
## Adelphi University Adelphi University Yes 2186  
## Adrian College Adrian College Yes 1428  
## Agnes Scott College Agnes Scott College Yes 417  
## Alaska Pacific University Alaska Pacific University Yes 193  
## Albertson College Albertson College Yes 587  
## Accept Enroll Top10perc  
## Abilene Christian University 1232 721 23  
## Adelphi University 1924 512 16  
## Adrian College 1097 336 22  
## Agnes Scott College 349 137 60  
## Alaska Pacific University 146 55 16  
## Albertson College 479 158 38

# remove the column containing the college names  
college <- college[, -1]  
college[1:6, 1:6]

## Private Apps Accept Enroll Top10perc  
## Abilene Christian University Yes 1660 1232 721 23  
## Adelphi University Yes 2186 1924 512 16  
## Adrian College Yes 1428 1097 336 22  
## Agnes Scott College Yes 417 349 137 60  
## Alaska Pacific University Yes 193 146 55 16  
## Albertson College Yes 587 479 158 38  
## Top25perc  
## Abilene Christian University 52  
## Adelphi University 29  
## Adrian College 50  
## Agnes Scott College 89  
## Alaska Pacific University 44  
## Albertson College 62

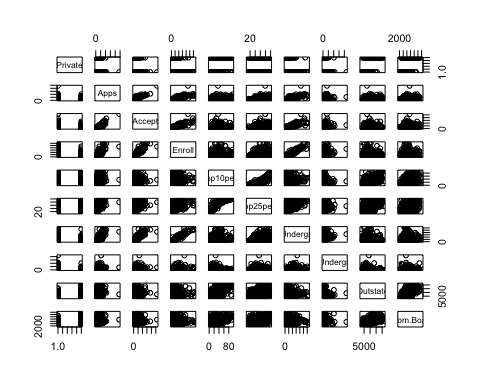
* 1. Perforom exploratory data analysis on *college* data set
     1. summary()

summary(college)

## Private Apps Accept Enroll Top10perc   
## No :212 Min. : 81 Min. : 72 Min. : 35 Min. : 1.00   
## Yes:565 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242 1st Qu.:15.00   
## Median : 1558 Median : 1110 Median : 434 Median :23.00   
## Mean : 3002 Mean : 2019 Mean : 780 Mean :27.56   
## 3rd Qu.: 3624 3rd Qu.: 2424 3rd Qu.: 902 3rd Qu.:35.00   
## Max. :48094 Max. :26330 Max. :6392 Max. :96.00   
## Top25perc F.Undergrad P.Undergrad Outstate   
## Min. : 9.0 Min. : 139 Min. : 1.0 Min. : 2340   
## 1st Qu.: 41.0 1st Qu.: 992 1st Qu.: 95.0 1st Qu.: 7320   
## Median : 54.0 Median : 1707 Median : 353.0 Median : 9990   
## Mean : 55.8 Mean : 3700 Mean : 855.3 Mean :10441   
## 3rd Qu.: 69.0 3rd Qu.: 4005 3rd Qu.: 967.0 3rd Qu.:12925   
## Max. :100.0 Max. :31643 Max. :21836.0 Max. :21700   
## Room.Board Books Personal PhD   
## Min. :1780 Min. : 96.0 Min. : 250 Min. : 8.00   
## 1st Qu.:3597 1st Qu.: 470.0 1st Qu.: 850 1st Qu.: 62.00   
## Median :4200 Median : 500.0 Median :1200 Median : 75.00   
## Mean :4358 Mean : 549.4 Mean :1341 Mean : 72.66   
## 3rd Qu.:5050 3rd Qu.: 600.0 3rd Qu.:1700 3rd Qu.: 85.00   
## Max. :8124 Max. :2340.0 Max. :6800 Max. :103.00   
## Terminal S.F.Ratio perc.alumni Expend   
## Min. : 24.0 Min. : 2.50 Min. : 0.00 Min. : 3186   
## 1st Qu.: 71.0 1st Qu.:11.50 1st Qu.:13.00 1st Qu.: 6751   
## Median : 82.0 Median :13.60 Median :21.00 Median : 8377   
## Mean : 79.7 Mean :14.09 Mean :22.74 Mean : 9660   
## 3rd Qu.: 92.0 3rd Qu.:16.50 3rd Qu.:31.00 3rd Qu.:10830   
## Max. :100.0 Max. :39.80 Max. :64.00 Max. :56233   
## Grad.Rate   
## Min. : 10.00   
## 1st Qu.: 53.00   
## Median : 65.00   
## Mean : 65.46   
## 3rd Qu.: 78.00   
## Max. :118.00

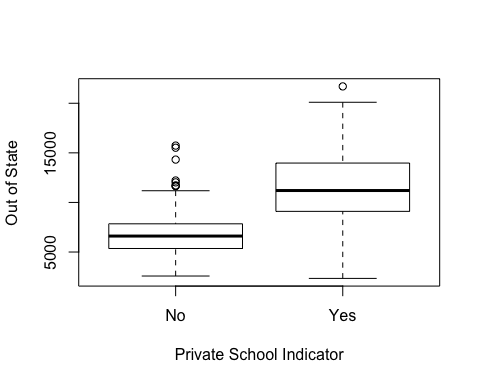
+ ii. pairs()

pairs(college[, 1:10])



+ iii. plot()

plot(college$Private, college$Outstate, xlab = "Private School Indicator", ylab = "Out of State")

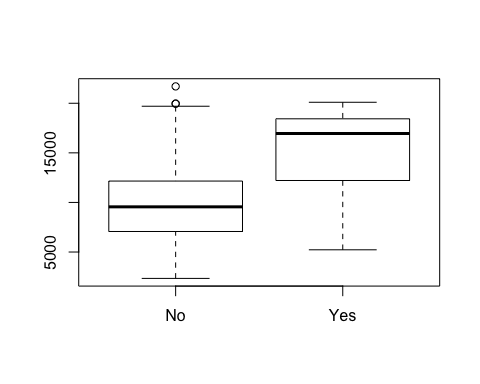


+ iv. Create new variable

# create Elite indicator for colleges where more than 50% of the students are from the top 10% of their high schoo class  
Elite <- rep("No", nrow(college))  
Elite[college$Top10perc > 50] <- "Yes"  
Elite <- as.factor(Elite)  
college <- data.frame(college, Elite)  
  
# determine counts for elite colleges  
summary(college$Elite)

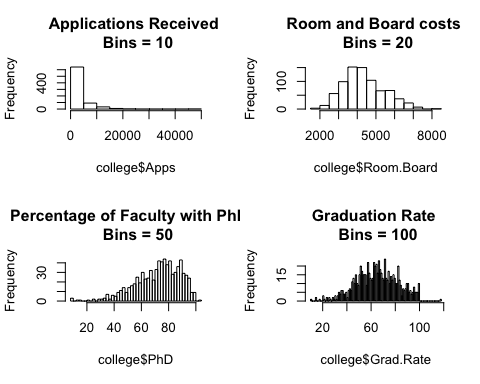
## No Yes   
## 699 78

# produce boxplots for Elite and Outstate  
plot(college$Elite, college$Outstate)



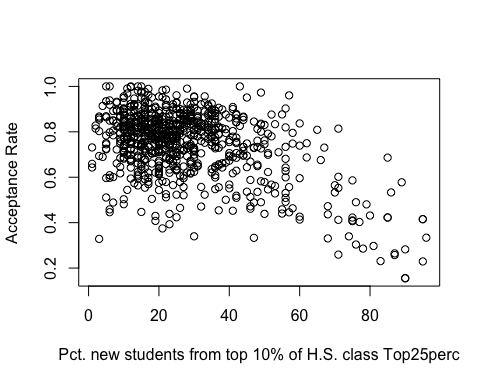
+ v. hist()

par(mfrow = c(2,2))  
hist(college$Apps, breaks = 10, main = "Applications Received \n Bins = 10")  
hist(college$Room.Board, breaks = 20, main = "Room and Board costs \n Bins = 20")  
hist(college$PhD, breaks = 50, main = "Percentage of Faculty with PhDs \n Bins = 50")  
hist(college$Grad.Rate, breaks = 100, main = "Graduation Rate \n Bins = 100")

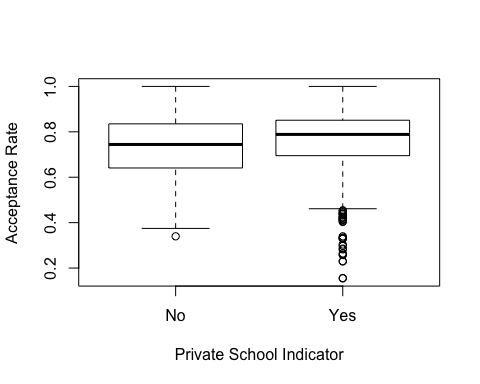


+ vi. Further exploration

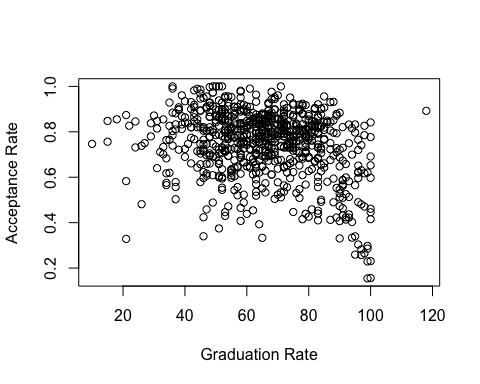
plot(college$Top10perc,  
 college$Accept/college$Apps,  
 xlab = "Pct. new students from top 10% of H.S. class Top25perc",  
 ylab = "Acceptance Rate")



plot(college$Private,  
 college$Accept/college$Apps,  
 xlab = "Private School Indicator",  
 ylab = "Acceptance Rate")



plot(college$Grad.Rate,  
 college$Accept/college$Apps,  
 xlab = "Graduation Rate",  
 ylab = "Acceptance Rate")

 As the number of students from the top 10% of their high school class increases, the acceptance rate decreases. This would indicate that colleges that receive more competitive applications are able to be more selective. The median acceptance rate is slightly higher for private schools but they have many more outlier examples of lower acceptance rates. Graduation rate is generally higher for schools with lower acceptance rates. In general, it appears that stronger students tend to go to more selective schools and that selective schools tend to have higher graduation rates. The most selective schools tend to be private.