HOMEWORK 2

Andrew Guo

2/10/2022

PROBLEM 1

The table below provides a training data set containing six observations, three predictors, and one qualitative response variable. Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

Obs.	X_1	X_2	X_3	Y	Euclidean Dist.
1	0	3	0	Red	3
2	2	0	0	Red	2
3	0	1	3	Red	3.162
4	0	1	2	Green	2.236
5	-1	0	1	Green	1.414
6	1	1	1	Red	1.732

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$. Fill in the table above.

(b) What is our prediction with K = 1? Why?

With a value of k=1, this would mean that we are looking for the closest observation to the test point. The closest observation is 5 with a distance of 1.414 to the test point. Thus, our prediction is that our test observation will be green.

(c) What is our prediction with K = 3? Why?

With a value of k=3, this would mean that we are looking for the three closest observations to the test point. Those three are observations 5, 6, and 2. Two out of the three observations, observations 2 and 6 are red, and one out of the three observations, observation 5, is green. Thus, our prediction is that our test observation will be red.

(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the best value for K to be large or small? Why?

If the Bayes decision boundary is highly non-linear, this indicates that we have a low K value because the nonlinearity indicates that the boundary itself is highly flexible and finds patterns in the data that do not correspond to the decision boundary.

PROBLEM 2

This exercise relates to the `College` data set, which can be found in the file `College.csv`. It contains a number of variables for 777 different universities and colleges in the US. The variables are defined on pp. 54-55 of the text.

(a) Use the `read.csv()` function to read the data into `R`. Call the loaded data `college`.

```
college = read.csv('College.csv')
attach(college)
```

(b) Look at the data using the `View()` function. You should notice that the first column is just the name of each university. We don't really want `R` to treat this as data. However, it may be handy to have these names for later. Run the following in sequence: `rownames(college) <- college[, 1]` and then `college <- college[, -1]`. You should see that there is now a `row.names` column with the name of each university recorded. This means that `R` has given each row a name corresponding to the appropriate university. `R` will not try to perform calculations on the row names. We have also eliminated the first column in the data where the names were stored.

```
#View(college)
rownames(college) <- college[, 1]
college <- college[, -1]
```

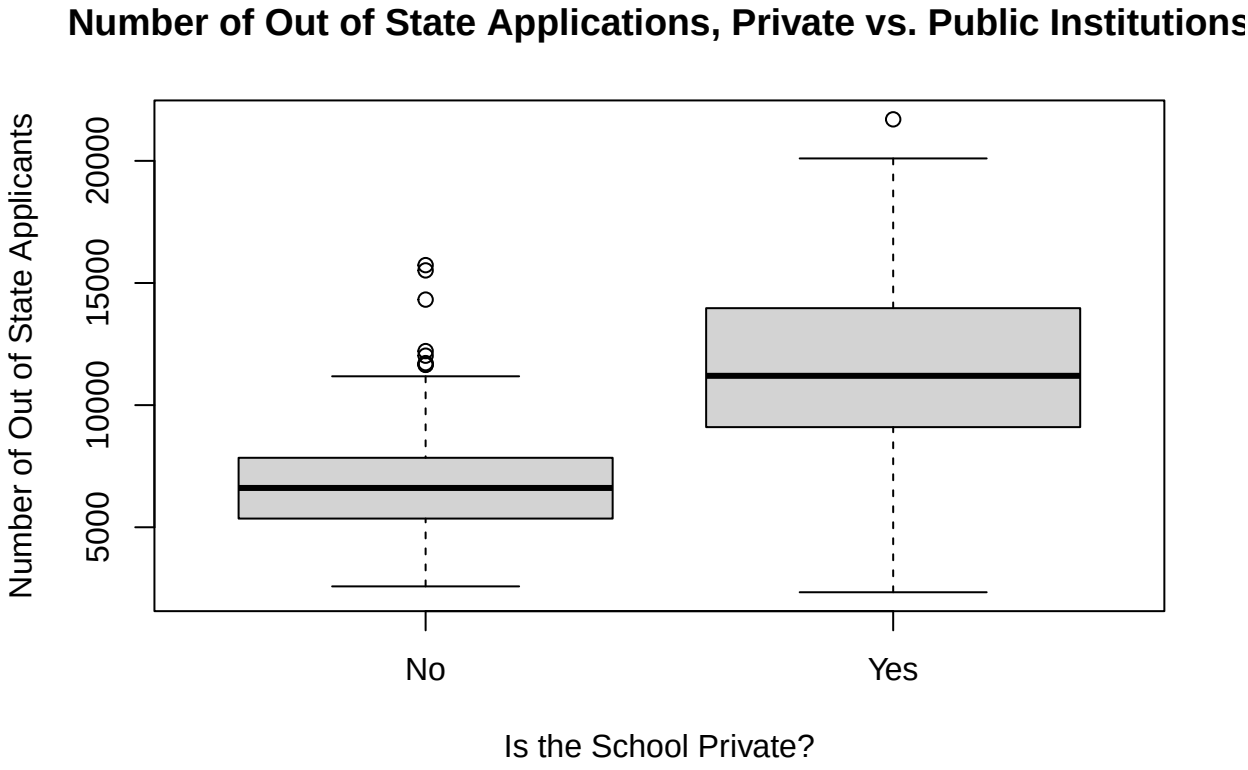
(c) Every variable in the data frame is either an `int` or `num` datatype, except for `Private` which is `chr`. You need to convert this to categorical factors by running `college$Private <- as.factor(college$Private)`. Then use the `summary()` function to produce a numerical summary of the variables in the data set.

```
college$Private <- as.factor(college$Private)
summary(college$Private)
```

```
##  No Yes
## 212 565
```

(d) Use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Private`. Label the x and y axes appropriately. What does the plot tell you?

```
boxplot(college$Outstate~college$Private, main = 'Number of Out of State Applications, P
```



The data reveals that private institutions yield a far greater number of applications from out of state students compared to public institutions. We see that for public schools, the mean number of out of state applications is rouhghly 7,000 with 50% of the applications ranging from around 5,000 to 8,000 applicants. Meanwhile, private institutions see a mean number of out of state applications of roughly 12,000, with 50% of all applications ranging from roughly 10,000 to 14,000 applications.

(e) Create a new qualitative variable, called `Elite`, by binning the `Top10perc` variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

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

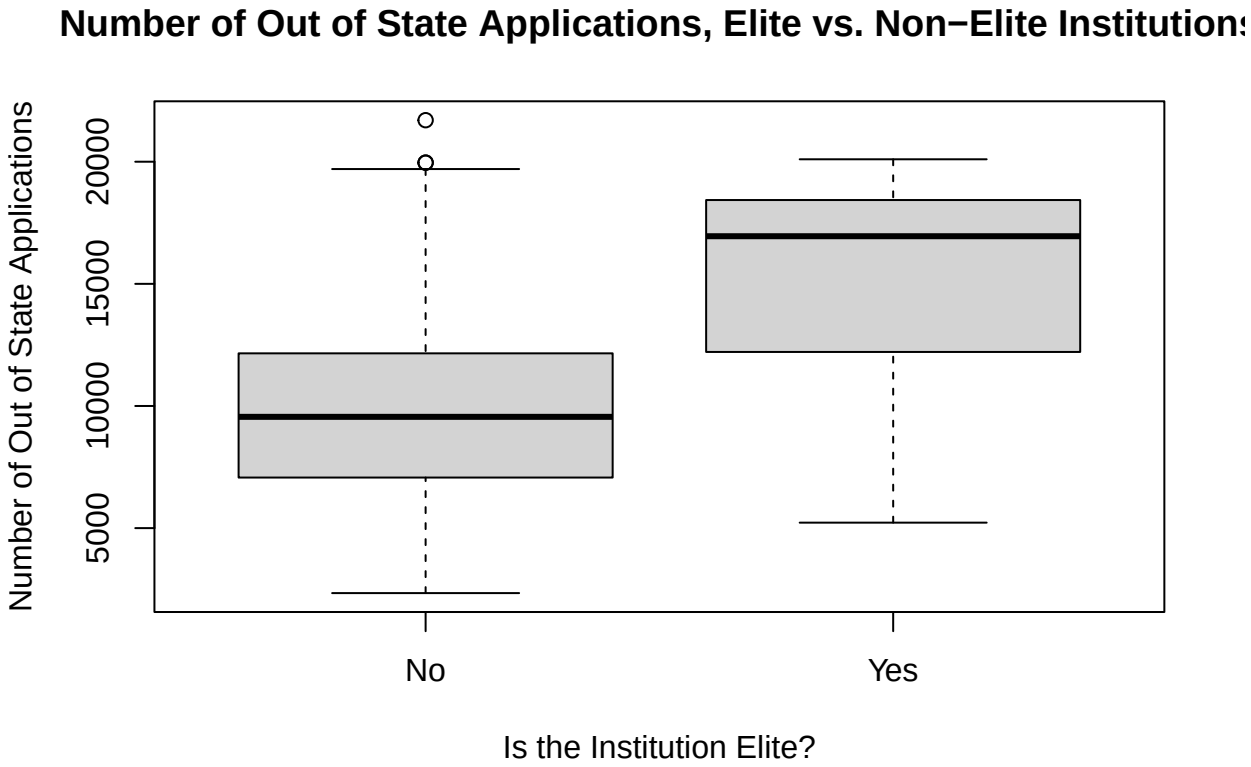
```
Elite <- rep("No", nrow(college))
Elite[college$Top10perc >50] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college, Elite)
```

(f) Use the `summary()` function to see how many elite universities there are. Now use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Elite`. What does the plot tell you?

```
summary(Elite)
```

```
##  No Yes
## 699  78
```

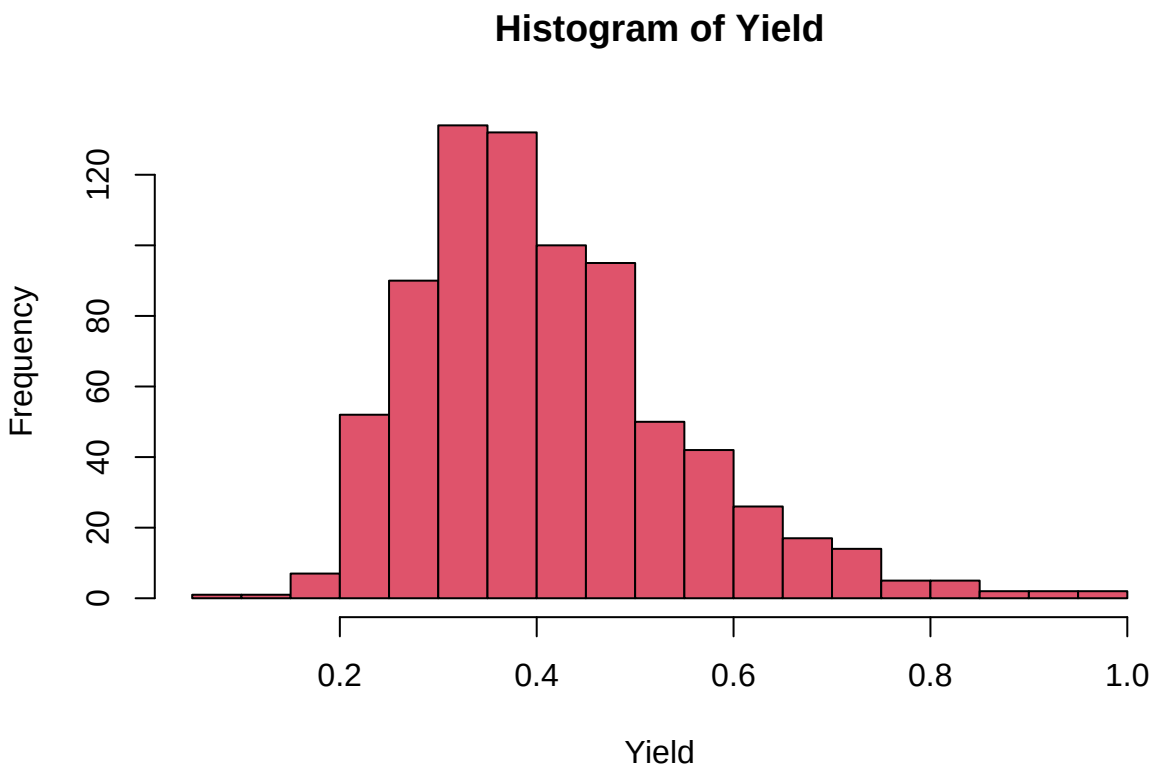
```
boxplot(college$Outstate~Elite, main = 'Number of Out of State Applications, Elite vs. N
```



On average, "Elite" institutions receive more out of state applications than institutions not labeled as "Elite". The mean Elite institution out of state application amount is roughly 17,000, with 50% of out of state applications ranging from roughly 13,000 to 18,000 applications, compared to the "non-elite" institutions having a mean application amount of roughly 10,000, with 50% of out of state applications ranging from about 7,000 to 12,000 applications.

(g) Create a new variable called `Yield`, which is the percentage of accepted students that enroll, and add it to the data frame `college`. Then create a histogram of `Yield` with 15 breaks using `hist()`. Describe the shape of the histogram. What is a practical interpretation of the shape?

```
Yield <- college$Enroll/college$Accept
college <- data.frame(college, Yield)
hist(Yield, col = 2, breaks = 15)
```




```
#View(college)
```

The data contains no gaps, the shape is roughly unimodal, contains no visible outliers, the center is beteween 0.2 and 0.4, and the distribution is slightly skewed to the right. This tells us that for a large number of cases, only twenty to fifty percent of students will actually enroll in any school that they are accepted by.

(h) Use the `mean()` and `sd()` functions to compute the mean and standard deviation of the `Yield` for elite versus non-elite schools. Which type of school has a higher `Yield`? Which type of school has greater variability in `Yield`? You can access the values of `Yield` for elite schools using `Yield[Elite == "Yes"]`.

```
mean(Yield[Elite == 'Yes'])
```

```
## [1] 0.3774866
```

```
mean(Yield[Elite == 'No'])
```

```
## [1] 0.4158676
```

```
sd(Yield[Elite =='Yes'])
```

```
## [1] 0.1101219
```

```
sd(Yield[Elite =='No'])
```

```
## [1] 0.1359164
```

The school with the larger mean value of "Yield" are schools that are not labeled as "Elite";

this same type of institution also yields greater variability in "Yield".

The lower yield for "Elite" instituitons can likely be attributed to the fact that students who are applying to "Elite" institutions have a variety of institutions that they are applying for (considering they may have a more "impressive" resume), which could range anywhere from 10-20 schools. Thus, "Elite" schools may have less of a yield because its applicants are applying to so many schools. However, non-elite schools may have more of a yield since these schools are often labeled as "safety" or "target" schools by college applicants, so students are more likely to settle for these schools if they are not accepted by an "Elite" institution. The variability can be explained similarly: "non-Elite" institutions may have a higher variability in yield because of the typology of its applicants. "Elite" institution applicants are likely those who have a 'decent/good' high school resume and less likely to have applications from those who don't have a similarly 'impressive' resume. However, "non-Elite' institutions see applications from the latter, as these students may see these schools as more feasible options, as well as from"elite" institution applicants looking for a "safety" option.

For instance, a school like the College of William and Mary may see a more stable yield over a large number of years considering that it is an "Elite" school, but a school like James Madison University, which woud not be considered "Elite", may see more variability in its results because it is in the same vicinity as other top Virginia schools like W&M, UVA, Washington & Lee, etc; some years, it may see a smaller yield if those top schools are accepting many applicants, some years, it may be much higher because those top schools are not accepting that many applicants (Same premise could possibly apply to Holy Cross and WPI vs. Worcester State).

PROBLEM 3

Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of `sales`, `TV`, `radio`, and `newspaper`, rather than in terms of the coefficients of the linear model.

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	2.939	0.3119	9.42	< 0.0001
`TV`	0.046	0.0014	32.81	< 0.0001
`radio`	0.189	0.0086	21.89	< 0.0001
`newspaper`	−0.001	0.0059	−0.18	0.8599

Assume: TV parameter = b1, radio parameter = b2, newspaper parameter = b3

Ho: b1 = 0

Ha: b1 != 0

t(TV) = b1/SE(b1) = 0.046/0.0014 = 32.81

2[p(t > 32.81)] = 0.0001

0.0001 < 0.05

Reject the null; with a p-value smaller than a = 0.05, we reject the null. There is convincing evidence that 'TV' is a statistically significant predictor for 'sales' while holding the other predictors constant.

Ho: b2 = 0

Ha: b2 != 0

t(radio) = b2/SE(b2) = 0.189/0.0086 = 21.89

2[p(t > 21.89)] = 0.0001

0.0001 < 0.05

With a p-value smaller than a = 0.05, we reject the null. There is convincing evidence that 'radio' is a statistically significant predictor for 'sales' while holding the other predictors constant.

Ho: b3 = 0

Ha: b3 != 0

t(newspaper) = b3/SE(b3) = -0.0001/0.0059 = -0.18

2[p(t < -0.18)] = 0.8599

0.8599 > 0.05

With a p-value greater than a = 0.05, we fail reject the null.There is convincing evidence that 'newspaper' is not a statistically significant predictor for 'sales' while holding the other predictors constant.

This method of individually calculating the significance of the predictors is faulty, as by doing so many different t-test, the likelihood of a Type I error (false positive) increases

PROBLEM 4

I collect a set of data ($n = 100$ observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$.

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \epsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

The RSS would be lower for the cubic regression model because the presence of more parameters means that the model will be far more adept at "training" itself to fit the predesignated data the best. The presence of more parameters will mean that the R-Squared value will increase as more parameters will be used to calculate the supposed explanation in the variability in Y on the Xs. The linear model may display a similarly low RSS due to the true linear relationship between X and Y, but because the training data's actual shape and distribution is not given, a model with more parameters would be more apt at fitting itself within a given and known data set.

(b) Answer (a) using test rather than training RSS.

The RSS would be lower for the linear example because the cubic model will be prone to trying to find nonlinear patters and relationships in the test data. Because the true relationship between x and y is linear, the cubic regression model may yield an RSS that is higher because of the presence of a variety of nonlinear parameters that seek to model curvature. Additionally, when considering that the cubic regression model's shape and form is informed by the training data, new test data on Y that is linear may induce more error as the new test data may not exactly replicate the patterns, behaviors, and norms of the training data.

A linear regression model will have a lower RSS if the true relationship is linear because the model will fit the true nature of the relationship and will be far less prone to be influenced by data points that would otherwise fluctuate the model's accuracy.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

The RSS would be lower for the cubic regression model because the presence of more parameters means that the model will be far more adept at "training" itself to fit the predesignated data the best. The presence of more parameters will mean that the R-Squared value will increase as more parameters will be used to calculate the supposed explanation in

the variability in Y on the Xs. A model with more parameters is always more apt at fitting itself within a given and known data set.

(d) Answer (c) using test rather than training RSS.

Because we don't know the extent to which the true relationship is not linear, there is not enough information to make a definitive expectation. If the true relation is closer to linear than it is nonlinear, then we could expect the RSS in both the cubic and linear models to be roughly the same. The linear model would be able to approximate the relationship decently since the true relationship is roughly linear. The cubi regression model will be similarly low (and around the same as the linear model RSS) because the true relationship is definitively NOT linear, meaning that the cubic regression model will be able to account for any elements that seem to deviate from the overall linear shape. However, if the model is more nonlinear than linear, then we would expect the cubic regression model to have a smaller RSS due to cubic models in general being more adept at accounting for nonlinear patterns.