

DATA SCI 415 - Homework 2

Due Wednesday Feb 5, 2025, 11:59 PM

1. This exercise relates to the College data set, which can be found in the file `College.csv` (on Canvas— Files — Data). It contains a number of variables for 777 different universities and colleges in the US. The variables are

- `Private` : Public/private indicator
- `Apps` : Number of applications received
- `Accept` : Number of applicants accepted
- `Enroll` : Number of new students enrolled
- `Top10perc` : New students from top 10% of high school class
- `Top25perc` : New students from top 25% of high school class
- `F.Undergrad` : Number of full-time undergraduates
- `P.Undergrad` : Number of part-time undergraduates
- `Outstate` : Out-of-state tuition
- `Room.Board` : Room and board costs
- `Books` : Estimated book costs
- `Personal` : Estimated personal spending
- `PhD` : Percent of faculty with Ph.D.'s
- `Terminal` : Percent of faculty with terminal degree
- `S.F.Ratio` : Student/faculty ratio
- `perc.alumni` : Percent of alumni who donate
- `Expend` : Instructional expenditure per student
- `Grad.Rate` : Graduation rate

Before reading the data into R, it can be viewed in Excel or a text editor.

- (a) Use the `read.csv()` function to read the data into R. Call the loaded data `college`. Make sure that you have the directory set to the correct location for the data.
- (b) Look at the data using the `fix()` 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. Try the following commands:

```
> rownames(college) = college[,1]
> fix(college)
```

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. However, we still need to eliminate the first column in the data where the names are stored. Try

```
> college = college[,-1]
> fix(college)
```

Now you should see that the first data column is `Private`. Note that another column labeled `row.names` now appears before the `Private` column. However, this is not a data column but rather the name that R is giving to each row.

- (c) Now do the following.
 - i. Use the `summary()` function to produce a numerical summary of the variables in the data set.
 - ii. Use the `pairs()` function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix *A* using `A[,1:10]`.

- iii. Use the `plot()` function to produce side-by-side boxplots of `Outstate` versus `Private`.
- iv. 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)
```

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`.

- v. Use the `hist()` function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command `par(mfrow=c(2,2))` useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.
 - vi. Continue exploring the data, and provide a brief summary of what you discover.
2. This question should be answered using the `Carseats` data set in the `ISLR` package. You may use `help(Carseats)` to learn more about the data set.
- (a) Fit a multiple regression model to predict `Sales` using `Price`, `Urban`, and `US`.
 - (b) Provide an interpretation of each coefficient in the model. Be careful - some of the variables in the model are qualitative!
 - (c) Write out the model in equation form, being careful to handle the qualitative variables properly.

- (d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?
- (e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.
- (f) How well do the models in (a) and (e) fit the data?