STATS 415 HW 2

Jiahao Cheng

Feburary 2025

Problem 1

(a) As shown in plot 1, college data is loaded.

```
> setwd("/Users/jiatao/Desktop")
> # 1.1
> college <- read.csv("College.csv")
> college
                                  X Private  Apps Accept Enroll Top10perc Top25perc F.Undergrad
1             Abilene Christian University     Yes  1660   1232    721        23        52        2885
2                       Adelphi University     Yes  2186   1924    512        16        29        2683
3                           Adrian College     Yes  1428   1097    336        22        50        1036
4                     Agnes Scott College      Yes   417    349    137        60        89         510
5                 Alaska Pacific University    Yes   193    146     55        16        44         249
6                         Albertson College    Yes   587    479    158        38        62         678
7                   Albertus Magnus College    Yes   353    340    103        17        45         416
8                           Albion College     Yes  1899   1720    489        37        68        1594
9                         Albright College     Yes  1038    839    227        30        63         973
10                Alderson-Broaddus College    Yes   582    498    172        21        44         799
11                        Alfred University    Yes  1732   1425    472        37        75        1830
12                      Allegheny College      Yes  2652   1900    484        44        77        1707
13 Allentown Coll. of St. Francis de Sales     Yes  1179    780    290        38        64        1130
14                            Alma College     Yes  1267   1080    385        44        73        1306
15                         Alverno College     Yes   494    313    157        23        46        1317
16          American International College     Yes  1420   1093    220         9        22        1018
17                        Amherst College      Yes  4302    992    418        83        96        1593
18                      Anderson University    Yes  1216    908    423        19        40        1819
19                      Andrews University     Yes  1130    704    322        14        23        1586
20                  Angelo State University     No  3540   2001   1016        24        54        4190
```

Figure 1: Loading Data

(b) As shown in plot 2, the first column is eliminated.

(c) Plot 3 is the summary of the data.
Plot 4 uses the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data.
Plot 5 and 6 uses the plot() function to produce side-by-side boxplots of Outstate versus Private.
Plot 7 creates a new qualitative variable called Elite, shows the summary(), and the code of the boxplot, there are 78 Elite Universities. Plot 8 shows the boxplot.
Plot 9 uses the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. Plot 10 is the figure.
Plot 11 shows the linear regression on Accept with Apps. It shows that the Apps has a positive relationship with Accept.

```
> # 1.2
> rownames(college) = college[,1]
> college = college[,-1]
> View(college)
```

Figure 2: Eliminate the first column

```
> # 1.3
> summary(college)
   Private              Apps           Accept          Enroll       Top10perc       Top25perc
 Length:777         Min.   :   81   Min.   :   72   Min.   :  35   Min.   :  1.00   Min.   :  9.0
 Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.: 15.00   1st Qu.: 41.0
 Mode  :character   Median : 1558   Median : 1110   Median : 434   Median : 23.00   Median : 54.0
                    Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   : 27.56   Mean   : 55.8
                    3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.: 35.00   3rd Qu.: 69.0
                    Max.   :48094   Max.   :26330   Max.   :6392   Max.   : 96.00   Max.   :100.0
  F.Undergrad     P.Undergrad         Outstate       Room.Board       Books           Personal
 Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250
 1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850
 Median : 1707   Median :  353.0   Median : 9990   Median :4200   Median : 500.0   Median :1200
 Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341
 3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700
 Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800
      PhD            Terminal       S.F.Ratio      perc.alumni        Expend         Grad.Rate
 Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
 1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
 Median : 75.00   Median : 82.0   Median :13.60   Median :21.00   Median : 8377   Median : 65.00
 Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
 3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
 Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
```

Figure 3: Summary(college)

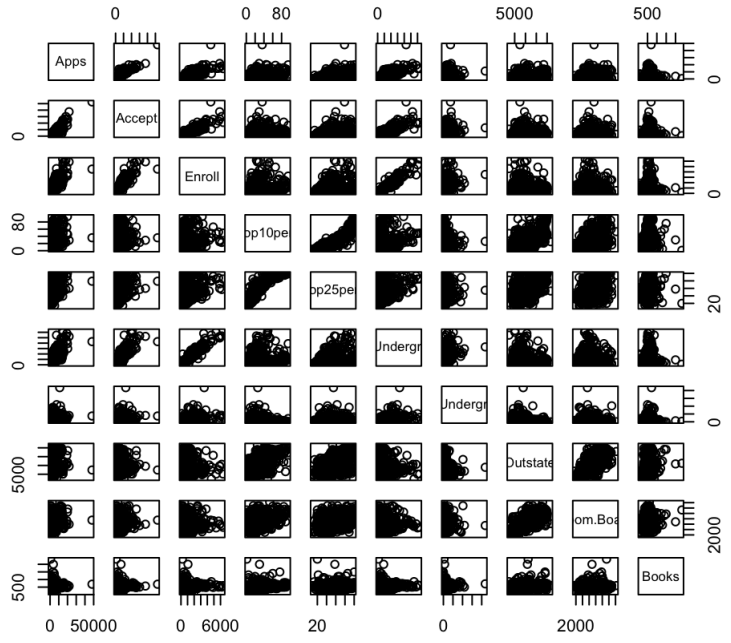

Figure 4: pairs(college[,2:11])

```
> college$Private <- factor(college$Private)
> plot(college$Outstate ~ college$Private,
+      xlab = "Private College",
+      ylab = "College Outstate",
+      main = "Outstae against Private")
```

Figure 5: boxplot code

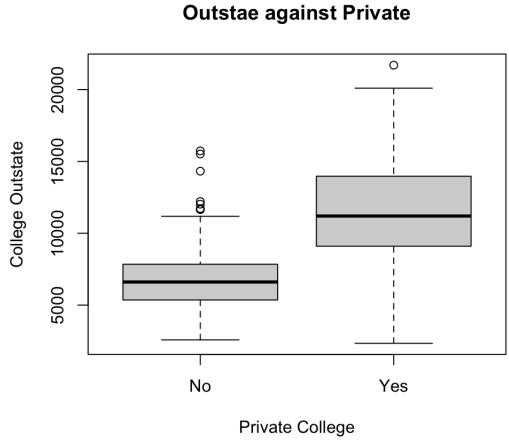

Figure 6: boxplot

```
> Elite = rep("No", nrow(college))
> Elite[college$Top10perc >50] = "Yes"
> Elite = as.factor(Elite)
> college = data.frame(college, Elite)
> summary(college$Elite)
 No Yes
699  78
> college$Elite <- factor(college$Elite)
> plot(college$Outstate ~ college$Elite,
+      xlab = "Elite College",
+      ylab = "College Outstate",
+      main = "Outstate against Elite")
```

Figure 7: Elite&summary()&boxplot

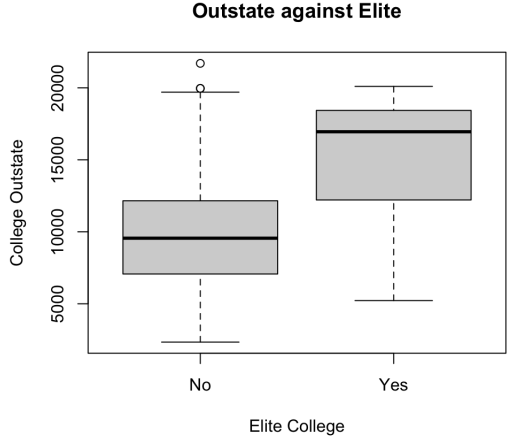

Figure 8: boxplot

```
> par(mfrow = c(2,2))
> hist(college$PhD, breaks = 50, main = "Histogram of PhD (50 bins)", xlab = "PhD")
> hist(college$PhD, breaks = 30, main = "Histogram of PhD (30 bins)", xlab = "PhD")
> hist(college$PhD, breaks = 10, main = "Histogram of PhD (10 bins)", xlab = "PhD")
> hist(college$PhD, breaks = 5, main = "Histogram of PhD (5 bins)", xlab = "PhD")
```

Figure 9: Histogram code

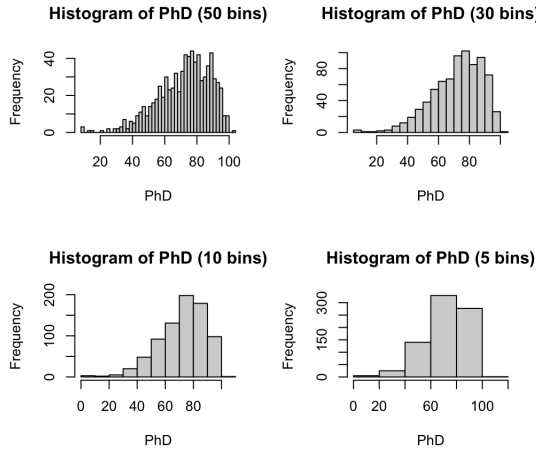

Figure 10: Histogram

```
> summary(lm(college$Accept~college$Apps))

Call:
lm(formula = college$Accept ~ college$Apps)

Residuals:
    Min      1Q  Median      3Q     Max
-6344.8  -154.2   -35.2   184.7  5490.5

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.253e+02  3.692e+01   6.101 1.66e-09 ***
college$Apps 5.975e-01  7.542e-03  79.226  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 813.1 on 775 degrees of freedom
Multiple R-squared:  0.8901,    Adjusted R-squared:   0.89
F-statistic:  6277 on 1 and 775 DF,  p-value: < 2.2e-16
```

Figure 11: lm(AcceptÃpp)

Problem 2

(a) Plot 12 shows the regression result.

```
> library(ISLR)
> model <- lm(Sales ~ Price + Urban + US, data=Carseats)
> summary(model)

Call:
lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9206 -1.6220 -0.0564  1.5786  7.0581

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.043469   0.651012  20.036  < 2e-16 ***
Price       -0.054459   0.005242 -10.389  < 2e-16 ***
UrbanYes    -0.021916   0.271650  -0.081    0.936
USYes        1.200573   0.259042   4.635 4.86e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.472 on 396 degrees of freedom
Multiple R-squared:  0.2393,    Adjusted R-squared:  0.2335
F-statistic: 41.52 on 3 and 396 DF,  p-value: < 2.2e-16
```

Figure 12: lm(AcceptÃpp)

(b) Interpretation:
$\beta_{Intercept} = 13.04$: the sale value when $Price = 0$, $Urban = No$, $US = No$. Since $Price$ can't be zero, the coefficient is just the proper vertical placement for $E(y_i|x_i)$. (Significant:$p < 0.05$)
$\beta_{Price} = -0.05$: other conditions are same, two individuals who differ in variable $Price$ by 1 unit are expected to differ in $Sales$ by β_{Price} units. (Significant:$p < 0.05$)
$\beta_{UrbanYes} = -0.02$: an $Urban$ individual tends to have lower $Sale$ than a $NonUrban$ individual. (Non-Significant:$p > 0.05$)
$\beta_{USYes} = 1.2$: an US individual tends to have higher $Sale$ than a $NonUS$ individual. (Significant:$p < 0.05$)

(c) Formula:
$UrbanYes = 1, USYes = 1$: $\hat{Sales} = 14.22 - 0.05 * Price$
$UrbanYes = 1, USYes = 0$: $\hat{Sales} = 13.02 - 0.05 * Price$
$UrbanYes = 0, USYes = 1$: $\hat{Sales} = 14.24 - 0.05 * Price$
$UrbanYes = 0, USYes = 0$: $\hat{Sales} = 13.04 - 0.05 * Price$

(d) I would reject the null hyphothesis for $\beta_{UrbanYes}$. Since the p-value is greater than 0.05.

(e) Plot 13 shows the smaller model.

```
> model1 <- lm(Sales ~ Price +  US, data=Carseats)
> summary(model1)

Call:
lm(formula = Sales ~ Price + US, data = Carseats)

Residuals:
    Min      1Q  Median      3Q     Max
-6.9269 -1.6286 -0.0574  1.5766  7.0515

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.03079    0.63098  20.652  < 2e-16 ***
Price       -0.05448    0.00523 -10.416  < 2e-16 ***
USYes        1.19964    0.25846   4.641 4.71e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.469 on 397 degrees of freedom
Multiple R-squared:  0.2393,    Adjusted R-squared:  0.2354
F-statistic: 62.43 on 2 and 397 DF,  p-value: < 2.2e-16
```

Figure 13: Smaller Model

(f) Compare to the previous model, every variables in the smaller one are significant while the Multiple R-squared values are similar. Therefore, the smaller model performs better.
This also shows that $UrbanYes$ doesn't contribute much to the model.