Homework 3: Due Friday, November 19 at 11:59pm

For this assignment you will be analyzing the **`College`** data frame included in the **`ISLR`** package, which contains 18 features for 777 U.S. colleges obtained from the 1995 issue of U.S. News and World Reports. The goal of this assignment is to become more familiar with linear and logistic regression. All analyses must be performed in R using **`tidyverse`** and other packages discussed in class. Provide all responses in the designated spaces in this Word document, then save it as a pdf and upload it to Canvas.

1. **[5%]** Recode the binary feature **`Private`** with the value **`0`** in place of **`No`** and the value **`1`** in place of **`Yes`**, and store it in a new data frame called **`College.recoded`**.

   Provide code below:

```
College.recoded <- College %>%
mutate (Private = ifelse(Private == "Yes", 1,0))
head(College.recoded)
```

2. **[10%]** Fit a multiple linear regression model to predict **`Private`** from the 17 other features. Print a summary of the model output to the console. Which feature is most statistically important in this model, and how do you know?

   Provide code and console output below:

```
lm.fit <- lm (Private ~ Apps + Accept + Enroll + Top10perc + Top25perc + F.Undergrad + P.Undergrad + Outstate
+ Room.Board + Books + Personal + PhD + Terminal + S.F.Ratio + perc.alumni + Expend + Grad.Rate,
      College.recoded)
summary (lm.fit)
```

   OUTPUT:

```
Call:
lm(formula = Private ~ Apps + Accept + Enroll + Top10perc + Top25perc +
   F.Undergrad + P.Undergrad + Outstate + Room.Board + Books +
   Personal + PhD + Terminal + S.F.Ratio + perc.alumni + Expend +
   Grad.Rate, data = College.recoded)

Residuals:
    Min      1Q  Median      3Q     Max
-1.05640 -0.15757  0.02107  0.16986  1.42289


Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.799e-01  1.018e-01   8.640  < 2e-16 ***
Apps        -3.371e-05  9.401e-06  -3.586 0.000358 ***
Accept       4.259e-05  1.835e-05   2.320 0.020582 *
Enroll      -1.058e-05  4.928e-05  -0.215 0.830024
Top10perc    2.178e-03  1.530e-03   1.424 0.154973
Top25perc   -2.001e-04  1.178e-03  -0.170 0.865099
F.Undergrad -2.919e-05  8.496e-06  -3.435 0.000624 ***
P.Undergrad -9.345e-06  8.398e-06  -1.113 0.266212
Outstate     4.370e-05  4.787e-06   9.128  < 2e-16 ***
Room.Board   3.677e-05  1.262e-05   2.913 0.003685 **
Books        6.055e-05  6.223e-05   0.973 0.330844
Personal     3.199e-07  1.648e-05   0.019 0.984517
PhD         -4.052e-03  1.205e-03  -3.362 0.000812 ***
Terminal    -4.012e-03  1.323e-03  -3.032 0.002510 **
S.F.Ratio   -1.472e-02  3.358e-03  -4.384 1.33e-05 ***
perc.alumni  2.689e-03  1.067e-03   2.520 0.011931 *
Expend      -5.457e-06  3.303e-06  -1.652 0.098894 .
Grad.Rate    1.540e-03  7.726e-04   1.993 0.046605 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.272 on 759 degrees of freedom
Multiple R-squared:  0.6358,  Adjusted R-squared:  0.6276
F-statistic: 77.94 on 17 and 759 DF,  p-value: < 2.2e-16
```

Provide answers to questions below:

The feature with the smallest P value is the most statistically important. In this case, it is Outstate feature

3. **[6%]** Fit a simple linear regression model to predict **`Private`** from the most statistically important feature you identified in question 2. Print a summary of the model output to the console. Is the feature from question 2 still statistically important, and how do you know?

Provide code and console output below:

```
lm.fit <- lm (Private ~ Outstate, College.recoded)
summary (lm.fit)
```

OUTPUT:

```
Call:
lm(formula = Private ~ Outstate, data = College.recoded)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0511 -0.3506  0.1008  0.3004  0.7688
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.790e-02  3.711e-02   2.369   0.0181 *
Outstate    6.123e-05  3.317e-06  18.460   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3717 on 775 degrees of freedom
Multiple R-squared:  0.3054,  Adjusted R-squared:  0.3045
F-statistic: 340.8 on 1 and 775 DF,  p-value: < 2.2e-16
```

Provide answer to question below:

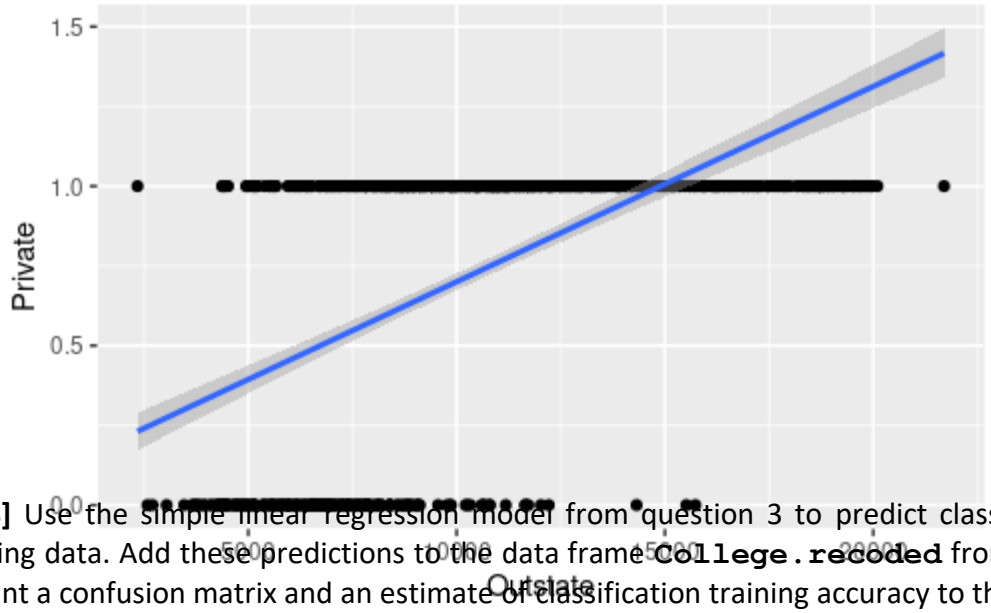
Yes, the Outstate feature is still statistically important because it has the smallest significance value.

4. **[10%]** Visualize the simple linear regression model from question 3 on a scatterplot.

Provide code below:

```
College.recoded %>%
ggplot(mapping = aes(x = Outstate, y = Private)) +
geom_point() +
geom_smooth(method = "lm")
```

Provide figure below:



5. **[10%]** Use the simple linear regression model from question 3 to predict classes for the training data. Add these predictions to the data frame `College.recoded` from question 1. Print a confusion matrix and an estimate of classification training accuracy to the console.

Provide code and console output below:

```
lm.probs <- predict(lm.fit, type = "response")
College.recoded <- College.recoded %>%
 mutate(probs = lm.probs,
       pred = ifelse(probs > 0.5,1,0))
College.recoded %>%
 select(Private, pred) %>%
 table()
College.recoded %>%
 summarize(accuracy = mean(pred == Private))
```

OUTPUT:

```
> College.recoded %>%
+   select(Private, pred) %>%
+   table()
      pred
Private   0   1
     0 111 101
     1  38 527
> College.recoded %>%
+   summarize(accuracy = mean(pred == Private))
  accuracy
1 0.8211068
```

6. **[12%]** Perform the same operations as in question 5, except using the multiple linear regression model from question 2. Is the performance of the multiple linear regression model better, worse, or the same as the simple linear regression model? Justify your answer, and then provide a reason for your finding.

Provide code and console output below:

```
lm.probs_multiple <- predict(lm.fit_multiple, type = "response")
College.recoded <- College.recoded %>%
 mutate(probs = lm.probs_multiple,
       pred = ifelse(probs > 0.5,1,0))
College.recoded %>%
 select(Private, pred) %>%
 table()
College.recoded %>%
 summarize(accuracy = mean(pred == Private))
```

OUTPUT:

```
> College.recoded %>%
+   select(Private, pred) %>%
+   table()
     pred
Private   0   1
    0 178  34
    1  13 552
> College.recoded %>%
+   summarize(accuracy = mean(pred == Private))
  accuracy
1 0.9395109
```

Provide answers to questions below:

The performance of the multiple linear regression model is better because of the higher accuracy of 93.9%.

7. [10%] Fit a multiple logistic regression model to predict **`Private`** from the 17 other features. Print a summary of the model output to the console. Which feature is most statistically important in this model, and how do you know?

Provide code and console output below:

```
glm.fit_multiple <- glm(Private ~ Apps + Accept + Enroll + Top10perc + Top25perc + F.Undergrad + P.Undergrad
+ Outstate + Room.Board + Books + Personal + PhD + Terminal + S.F.Ratio + perc.alumni + Expend + Grad.Rate,
                College.recoded, family= "binomial")
summary (glm.fit_multiple)
```

OUTPUT:

```
Call:
glm(formula = Private ~ Apps + Accept + Enroll + Top10perc +
    Top25perc + F.Undergrad + P.Undergrad + Outstate + Room.Board +
    Books + Personal + PhD + Terminal + S.F.Ratio + perc.alumni +
    Expend + Grad.Rate, family = "binomial", data = College.recoded)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7673  -0.0318   0.0502   0.1717   4.2070

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.574e-02  1.860e+00  -0.014  0.98896
Apps        -5.138e-04  2.284e-04  -2.249  0.02452 *
Accept       9.328e-05  4.382e-04   0.213  0.83144
Enroll       1.331e-03  8.487e-04   1.568  0.11687
Top10perc    8.451e-03  2.841e-02   0.297  0.76614
Top25perc    7.305e-03  1.895e-02   0.385  0.69993
F.Undergrad -4.168e-04  1.472e-04  -2.832  0.00462 **
P.Undergrad  1.836e-05  1.348e-04   0.136  0.89164
Outstate     6.822e-04  1.099e-04   6.207  5.4e-10 ***
Room.Board   1.901e-04  2.575e-04   0.738  0.46053
Books        2.059e-03  1.318e-03   1.562  0.11837
Personal    -3.283e-04  2.700e-04  -1.216  0.22395
PhD         -6.027e-02  2.665e-02  -2.262  0.02371 *
Terminal    -3.590e-02  2.580e-02  -1.392  0.16402
S.F.Ratio   -8.461e-02  6.076e-02  -1.393  0.16372
perc.alumni  4.782e-02  2.097e-02   2.280  0.02260 *
Expend       2.077e-04  1.207e-04   1.721  0.08529 .
Grad.Rate    1.634e-02  1.171e-02   1.395  0.16294
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 910.75  on 776  degrees of freedom
Residual deviance: 239.50  on 759  degrees of freedom
AIC: 275.5

Number of Fisher Scoring iterations: 8
```

Provide answers to questions below:

The Outstate feature is the most statistically significant due to the smallest significance level.

8. **[5%]** Fit a simple logistic regression model to predict **`Private`** from the most statistically important feature you identified in question 7. Print a summary of the model output to the console. Is the feature from question 7 still statistically important, and how do you know?

Provide code below:

```
glm.fit <- glm (Private ~ Outstate, College.recoded, family = "binomial")
summary (glm.fit)
```

OUTPUT:

```
Call:
glm(formula = Private ~ Outstate, family = "binomial", data = College.recoded)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2535  -0.5405   0.2167   0.5774   2.4935

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.522e+00  4.094e-01  -11.04   <2e-16 ***
Outstate     6.235e-04  4.957e-05   12.58   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 910.75  on 776  degrees of freedom
Residual deviance: 574.89  on 775  degrees of freedom
AIC: 578.89

Number of Fisher Scoring iterations: 6
```

Provide answers to questions below:

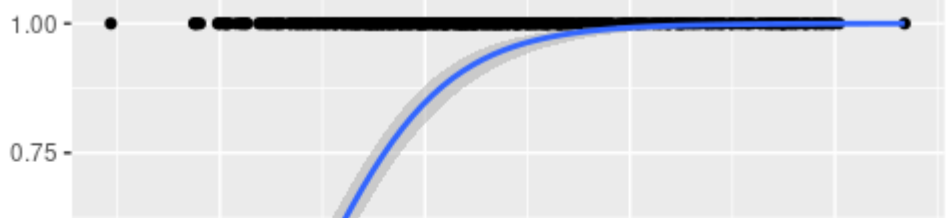
Yes, because Outstate still has the smallest significance value.

9. [10%] Visualize the simple logistic regression model from question 8 on a scatterplot.

Provide code below:

```
College.recoded %>%
 ggplot(mapping = aes(x = Outstate, y = Private)) +
 geom_point() +
 geom_smooth(method = "glm",
             method.args= c(family = "binomial"))
```

Provide figure below:



10. [10%] Use the simple logistic regression model from question 8 to predict classes for the training data. Add these predictions to the data frame **`College.recoded`** from question 1. Print a confusion matrix and an estimate of classification training accuracy to the console.

Provide code and console output below:

```
glm.probs <- predict(glm.fit, type = "response")
College.recoded <- College.recoded %>%
  mutate(probs = glm.probs,
         pred = ifelse(probs > 0.5,1,0))
College.recoded %>%
  select(Private, pred) %>%
  table()
College.recoded %>%
  summarize(accuracy = mean(pred == Private))
```

OUTPUT:

```
> College.recoded %>%
+   select(Private, pred) %>%
```

```
+   table()
      pred
Private   0   1
     0 140  72
     1  53 512
> College.recoded %>%
+   summarize(accuracy = mean(pred == Private))
   accuracy
1 0.8391248
```

11. [12%] Perform the same operations as in question 10, except using the multiple logistic regression model from question 7. Is the performance of the multiple logistic regression model better, worse, or the same as the multiple linear regression model? Justify your answer, and then provide a reason for your finding.

Provide code and console output below:

```
glm.probs_multiple <- predict(glm.fit_multiple, type = "response")
College.recoded <- College.recoded %>%
  mutate(probs = glm.probs_multiple,
         pred = ifelse(probs > 0.5,1,0))
College.recoded %>%
  select(Private, pred) %>%
  table()
College.recoded %>%
  summarize(accuracy = mean(pred == Private))
```

OUTPUT:

```
> College.recoded %>%
+   select(Private, pred) %>%
+   table()
      pred
Private   0   1
     0 191  21
     1  22 543
> College.recoded %>%
+   summarize(accuracy = mean(pred == Private))
   accuracy
1 0.9446589
```

Provide answers to questions below:

The performance of the multiple logistic regression is better than the multiple linear regression due to higher accuracy. The accuracy of multiple logistic regression is 94.5% and the accuracy of multiple linear regression is 93.9%. Although both have very close accuracy rates, multiple logistic regression is better.