Statistics 216
Homework 1 Solutions, total 70 points

1. *10 points*
Grading notes: Each part is worth 2 points, 1 for the correct answer and 1 for a valid explanation. Student explanations need not match the explanations provided here, but they must be correct, and they must match the answer given. If the incorrect answer is given, but the explanation matches the answer, award 1 point out of 2 for that part, even though the explanation cannot be correct because it led to the incorrect conclusion.

   (a) In this setting, I would expect a flexible method to perform **better** than an inflexible method. Because the number of observations is extremely large and the number of predictors is small, I would not be worried about the flexible method over-fitting the data, and a flexible method could capture more of the relationship between the predictors and the response than an inflexible method.

   (b) In this setting, I would expect a flexible method to perform **worse** than an inflexible method. Because the number of predictors is extremely large and the number of observations is small, I would expect a flexible method to over-fit the data.

   (c) In this setting, I would expect a flexible method to perform **worse** than an inflexible method. Because the variance of the error terms is extremely high, I would expect a flexible method to over-fit to the noise in the data.

   (d) In this setting, I would expect a flexible method to perform **better** than an inflexible method. Because σ^2 is small, there is a reduced risk of over-fitting to the data, and because the relationship between the predictors and the response is highly non-linear, I would not expect an inflexible method to capture this relationship.

   Grading notes for part (e): Any answer (better, worse or even "I'm not sure") can be valid for full credit here. Just make sure that the explanation matches the answer.

   (e) In this setting, I would expect a flexible method to perform **worse** than an inflexible method. The fact that the relationship between the predictors and the response is highly non-linear suggests a flexible method, but the fact that σ^2 is large suggests an inflexible method. In this case, I would choose an inflexible method because a model can still be informative even if it is wrong. Over-fitting is a worse sin.

2. *12 points*
Grading notes: Each part is worth 3 points, 1 for correctly identifying the type of problem (regression, classification or unsupervised learning), 1 for identifying the goal (inference or prediction) and 1 for correctly identifying n and p. Even if n and p are approximately correctly (off by one or two), award full credit for the third point. Deduct 1 point if the type of problem is not properly explained. n and p do not necessarily require

   (a) This scenario is a **regression** problem because the response, CEO salary, is a quantitative variable. We are most interested in **inference**. $n = 500$ (one data point for each firm) and $p = 3$ (profit, number of employees and industry).

   Grading notes for part (b): Anything goes. As long as the student specifies a type of learning problem, a goal, an n and a p, the student receives full credit. To clarify, that's 1 point for explaining any type of learning problem, 1 point for stating a goal and 1 point for any n and p.

   (b) This scenario is an **unsupervised learning** problem because we have a large dataset of customer-restaurant ratings, and we want to learn the underlying structure. We are most interested in **prediction**. $n = 10,000$ (number of customers) and $p = 1000$ (number of restaurants).

(c) This scenario is a **classification** problem because the response, success or failure, is a qualitative variable. We are most interested in **prediction**. $n = 20$ (one for each previous product) and $p = 13$ (price charged for the product, marketing budget, competition price and ten other variables).

(d) This scenario is a **regression** problem because the response, % change in the US dollar, is a quantitative variable. We are most interested in **prediction**. $n = 56$ (one data point for each week of the year 2012) and $p = 3$ (% change in the US market, % change in the British market and % change in the German market).

3. *15 points: (a) 6 points (b) 6 points (c) 3 points*
Grading notes: For parts (a) and (b), each example is worth 2 points: 1 point for description of response and predictors, and 1 point for explanation of the goal (inference or prediction). For part (c), each example is worth 1 point (award the point if the student makes it clear how the example is an unsupervised learning problem). Student answers will vary greatly; as such, no solutions are provided.

4. *13 points, all in part (c): i. 1 point ii. 1 point iii. 2 points iv. 3 points v. 6 points*
Grading notes: There is nothing to grade for parts (a) and (b).
Grading rules for the subsections of part (c) are given below.

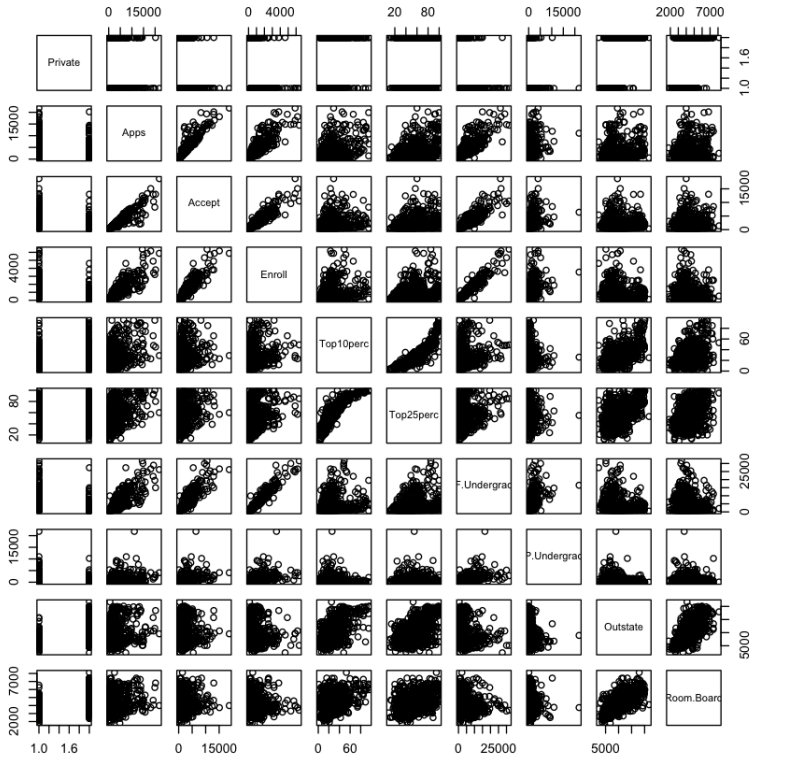
(c) *Grading notes for part (c)i: 1 point for displaying a summary like the one below, 0 points otherwise*

i.

```
Private        Apps           Accept          Enroll       Top10perc      Top25perc      F.Undergrad     P.Undergrad         Outstate
No :211   Min.   :   81   Min.   :   72.0   Min.   :  35.0   Min.   : 1.00   Min.   :  9.00   Min.   :  139   Min.   :    1.0   Min.   : 2340
Yes:565   1st Qu.:  776   1st Qu.:  603.2   1st Qu.: 242.0   1st Qu.:15.00   1st Qu.: 41.00   1st Qu.:  991   1st Qu.:   95.0   1st Qu.: 7305
          Median : 1558   Median : 1109.5   Median : 434.0   Median :23.00   Median : 54.00   Median : 1707   Median :  352.5   Median : 9990
          Mean   : 2944   Mean   : 1987.5   Mean   : 775.2   Mean   :27.55   Mean   : 55.77   Mean   : 3677   Mean   :  851.6   Mean   :10445
          3rd Qu.: 3603   3rd Qu.: 2407.5   3rd Qu.: 893.8   3rd Qu.:35.00   3rd Qu.: 69.00   3rd Qu.: 3969   3rd Qu.:  964.0   3rd Qu.:12931
          Max.   :21804   Max.   :18744.0   Max.   :6392.0   Max.   :96.00   Max.   :100.00   Max.   :31643   Max.   :21836.0   Max.   :21700
  Room.Board       Books           Personal         PhD           Terminal       S.F.Ratio      perc.alumni        Expend       Grad.Rate
Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.00   Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
1st Qu.:3596   1st Qu.: 469.5   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.00   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6749   1st Qu.: 53.00
Median :4198   Median : 500.0   Median :1200   Median : 75.00   Median : 82.00   Median :13.60   Median :21.00   Median : 8372   Median : 65.00
Mean   :4357   Mean   : 549.2   Mean   :1340   Mean   : 72.64   Mean   : 79.68   Mean   :14.08   Mean   :22.75   Mean   : 9659   Mean   : 65.45
3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1692   3rd Qu.: 85.00   3rd Qu.: 92.00   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10838   3rd Qu.: 78.00
Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.   :100.00   Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
```

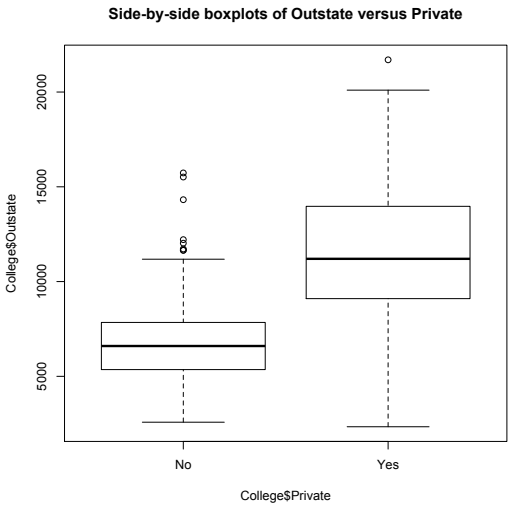
Grading notes for part (c)ii: 1 point for displaying a plot like the one below, 0 points otherwise

ii.



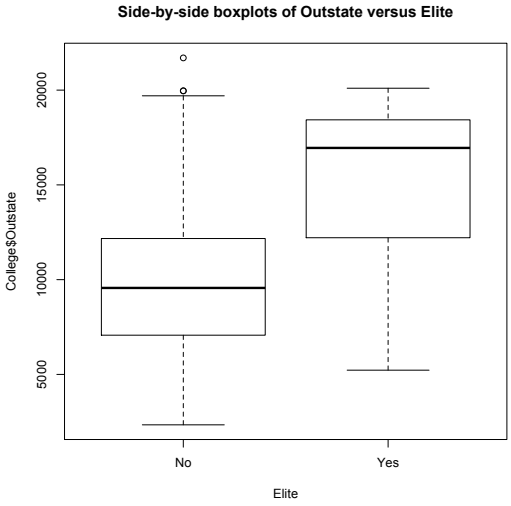
Grading notes for part (c)iii: 2 points for displaying a plot matching the one below, 1 point for a plot not matching the one below and 0 points otherwise

iii.



Grading notes for part (c)iv: The number of elite universities is worth 1 point, and then the other 2 points are earned as for the box plot in part (c)iii.

iv. There are 78 elite universities.



v. *Grading notes for part (c)v: For each variable, award 1 point for each histogram with a unique number of bins, up to 2 points. Award credit for up to three variables, for a total of up to 6 points. Student answers will vary greatly; as such, no solutions are provided.*

5. *10(+1) points: (b) 6(+1) points (c) 4 points*
Grading notes: This will be a very difficult question to grade because student solutions will vary greatly. Below is a sample solution, and each section has grading guidelines and some idea of what to expect from the students. There is nothing to grade for part (a).

Grading notes for part (b): The R output of a linear model is worth 2 points, but award only 1 point if it seems the student has improperly fitted the model. The training error and test error are worth 2 points <u>each</u>, but award only 1 point for each value that seems to have been calculated incorrectly. Additionally, award 1 point of extra credit if the student has done something beyond what was asked of them in this problem, perhaps by observing that there is one extreme outlier (Rutgers) and removing it from the dataset, or by addressing deficiencies of the model with interaction terms or a transformation of the response.

Comment: Because students are randomly splitting into training and test sets, expect wide variation in solutions. It is up to you to determine whether the student has fitted the model properly. Note that due to one extreme outlier (Rutgers) in the dataset, some solutions might look funny. For example, below the training MSE is larger than the test MSE. If Rutgers were randomly assigned to the test set, then the test MSE would be must larger than the training MSE. <u>Also, it is acceptable to report $RSE = \sqrt{MSE}$ instead of MSE.</u>

(b) The summary of the fit of the linear model is given below:

```
Call:
lm(formula = Apps ~ . - Accept - Enroll, data = College, subset = training)

Residuals:
   Min     1Q Median     3Q    Max
 -6459   -785   -128    573  31861

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.954e+03  1.241e+03  -3.185  0.00157 **
PrivateYes   5.253e+01  4.326e+02   0.121  0.90343
Top10perc    2.817e+01  1.652e+01   1.705  0.08905 .
Top25perc    3.722e+00  1.396e+01   0.267  0.78993
F.Undergrad  7.229e-01  3.953e-02  18.289  < 2e-16 ***
P.Undergrad -2.296e-01  1.444e-01  -1.590  0.11257
Outstate    -1.277e-02  5.625e-02  -0.227  0.82050
Room.Board   4.189e-01  1.452e-01   2.884  0.00415 **
Books        5.590e-01  7.249e-01   0.771  0.44110
Personal    -2.855e-01  1.888e-01  -1.512  0.13134
PhD         -9.786e+00  1.559e+01  -0.628  0.53059
Terminal    -5.178e+00  1.623e+01  -0.319  0.74993
S.F.Ratio    8.744e+01  4.248e+01   2.058  0.04024 *
perc.alumni -2.523e+01  1.242e+01  -2.032  0.04288 *
Expend       1.198e-01  3.431e-02   3.491  0.00054 ***
Grad.Rate    1.879e+01  8.851e+00   2.124  0.03436 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2202 on 372 degrees of freedom
Multiple R-squared:  0.746, Adjusted R-squared:  0.7357
F-statistic: 72.83 on 15 and 372 DF,  p-value: < 2.2e-16
```

The training MSE is 4648050, and the test MSE is 3055057.

Grading notes for part (c): Making some comment about how accurately we can predict numbers of applications based on this model is worth 2 points. Partial credit for these 2 points is at your discretion. In addition, award 1 point for each predictor that the student states is important, up to 2 points. Of course, in order to get credit for that point, the conclusion must be consistent with the linear model fit summary output provided by the student.

(c) It is very curious that the training MSE is larger (and extremely so) than the test MSE, this should not usually be the case. The residual standard error in the training set is 2202, while the median number of applications received is 1558, so this is a very large error. An R^2 of 0.746 is reasonably good, but I wouldn't say that this model very accurately predicts numbers of applications received.

The most important predictor (not surprisingly) is the number of full-time undergrads, followed by expenditures per student and the cost of room and board. Also somewhat important are student-to-faculty ratio, percentage of alumni giving donations and graduate rate.

6. *10 points*
Grading notes: Grade this problem similarly to the previous one. That's 2 points for the R output of the logistic regression model (deduct 1 point if the variable Apps is included in the model), 2 points for training misclassification rate, 2 points for test misclassification rate and 4 points for comparing the results of the logistic regression analysis to the results of the linear regression. In order to receive all 4 of those points, the student should either compare the predictions made by the two models (similarly to how I do so with the table below) or compare the predictors that turn up as significant in the two models. Award partial credit at your discretion.

The summary of the fit of the logistic regression model is given below:

```
Call:
glm(formula = y ~ . - Apps - Accept - Enroll, family = binomial,
    data = College, subset = training)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.86151  -0.27319  -0.02912   0.03261   2.99554

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.570e+01  3.390e+00  -4.630 3.65e-06 ***
PrivateYes   1.648e+00  1.135e+00   1.452   0.1466
Top10perc   -9.774e-03  4.219e-02  -0.232   0.8168
Top25perc    2.872e-02  3.553e-02   0.808   0.4188
F.Undergrad  3.361e-03  4.961e-04   6.776 1.24e-11 ***
P.Undergrad -1.223e-04  5.842e-04  -0.209   0.8342
Outstate     1.295e-04  1.060e-04   1.222   0.2216
Room.Board   4.549e-04  2.783e-04   1.634   0.1022
Books        1.799e-04  1.668e-03   0.108   0.9141
Personal     7.764e-05  4.172e-04   0.186   0.8524
PhD         -2.288e-02  3.132e-02  -0.730   0.4651
Terminal     1.630e-02  3.130e-02   0.521   0.6025
S.F.Ratio    3.238e-02  9.962e-02   0.325   0.7452
perc.alumni -1.092e-02  2.646e-02  -0.413   0.6799
Expend       1.131e-04  6.753e-05   1.675   0.0940 .
Grad.Rate    3.120e-02  1.817e-02   1.717   0.0861 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 537.84  on 387  degrees of freedom
Residual deviance: 145.81  on 372  degrees of freedom
AIC: 177.81

Number of Fisher Scoring iterations: 9
```

The training misclassification rate is 6.7%, and the test misclassification rate is 11.3%.

We can compare the predictions on the test set from this model against whether the prediction from the linear model is above the median. That results in the table below. For the most part, the predictions agree, but there are 40 test points for which the linear model makes a prediction above the median while the logistic model makes a prediction below the median.

	$\hat{Y}_{LM} < 1558$	$\hat{Y}_{LM} \geq 1558$
$\hat{Y}_{LR} = 0$	151	40
$\hat{Y}_{LR} = 1$	5	193

We can also compare which predictors are important in the two models. The three most important predictors in the logistic regression model are the number of full-time undergrads, followed by expenditures per student and graduation rate. All three of these are predictors that I mentioned in part 5(c). It is reassuring that the two models discovered similar important predictors.

Appendix: Relevant R code

```
library(ISLR)

#4.
# (a) & (b): Using the College dataset from the library ISLR makes parts (a) and (b) unnecessary
# (c)
#               i.
summary(College)
#               ii.
pairs(College[, 1:10])
#               iii.
plot(College$Outstate ~ College$Private, main = 'Side-by-side boxplots of Outstate versus Private')
#               iv.
Elite = rep("No", nrow(College))
Elite[College$Top10perc > 50] = "Yes"
Elite = as.factor(Elite)
college = data.frame(College, Elite)
summary(Elite)
plot(College$Outstate ~ Elite, main = 'Side-by-side boxplots of Outstate versus Elite')

#5
# a
set.seed(309)
training = sample(1:777, 388)
#b
modelLM = lm(Apps ~ . - Accept - Enroll, data = College, subset = training)
summary(modelLM)
trainingError = mean(modelLM$residuals^2)
trainingError
predictionLM = predict(modelLM, newdata = College[-training, ])
testError = mean((predictionLM - College$Apps[-training])^2)
testError

#6
y = as.numeric(College$Apps > median(College$Apps))
modelLR = glm(y ~ . - Apps - Accept - Enroll, data = College, subset = training, family = binomial)
summary(modelLR)
trainingError = mean((modelLR$fitted > 0.5) != y[training])
trainingError
predictionLR = predict(modelLR, newdata = College[-training, ], type = 'response')
testError = mean((predictionLR > 0.5) != y[-training])
testError
table(predictionLR > 0.5, predictionLM > median(College$Apps))
```