Statistics 216
Homework 1, due Wednesday Jan 29, 2014.

1. For each of parts (a) through (d), indicate whether you would expect the performance of a flexible statistical learning method to perform better or worse than an inflexible method. Give reasons in each case.

   (a) The number of observations n is extremely large, and the number of predictors p is small.

   (b) The number of predictors p is extremely large, and the number of observations n is small.

   (c) The variance of the error terms, i.e. $\sigma^2 = \text{Var}(\epsilon)$, is extremely high.

   (d) The relationship between the predictors and response is highly non-linear, and σ^2 is small.

   (e) The relationship between the predictors and response is highly non-linear, and σ^2 is large.

2. Explain whether each scenario below is a regression, classification or unsupervised learning problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

   (a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

   (b) Our website has collected the ratings of 1000 different restaurants by $10,000$ customers. Each customer has rated about 100 restaurants, and we would like to recommend restaurants to customers who have not yet been there.

   (c) We are considering launching a new product and wish to know whether it will be a *success* or a *failure*. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

   (d) We are interesting in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

3. In this next question we consider some real-life applications of statistical learning:

   (a) Describe three real-life applications in which *classification* might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

   (b) Describe three real-life applications in which *regression* might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

   (c) Describe three real-life applications in which *unsupervised learning* might be useful.

4. 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:

   - `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) 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 10 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.

5. In this exercise, we will predict the number of applications received using the other variables in the `College` data set.

   (a) Split the data set into a training set and a test set of approximately equal size.

   (b) Fit a linear model using least squares on the training set, and report the training and test error obtained. Do not include the `Elite` predictor, or the `Accept` or `Enrol` predictors in the regression.

   (c) Comment on the results obtained. How accurately can we predict the number of college applications received? What are the most important predictors?

6. Using the same setup as in the previous question, form a new outcome variable Y which equals one if the number of applications is greater than or equal to the overall median and zero otherwise. Fit a logistic regression model to Y and report the training and test misclassification rates, and the most important predictors. As above, do not include the `Elite` predictor, or the `Accept` or `Enrol` predictors in the regression. Compare the results of this analysis to that of the linear regression approach in the previous question.