ISyE 7406A: Homework # 1

Due date: Thursday, Jan 16

In your writeups, we expect clear explanations of models chosen, hypotheses tested, and findings analogous to what you would produce for a consulting project.

Problem 1. An ISyE 7406 student was asked to find the weights of two balls, A and B, using a scale with random measurement errors. When the student measured one ball at a time, the weights of A and B were 2 lbs and 1 lb, respectively. However, if the student measured both balls simultaneously, then the total weight $(A + B)$ were 4 lbs. The poor student was confused, and decided to repeat the above measurements. The new observed weights of A, B and $A + B$ are $2, 2$ and 5 lbs, respectively. For your information, the observed weights are summarized in the following table:

	A	B	$A + B$
First Time	2	1	4
Second Time	2	2	5

(a) There are 6 observations, and we can write the observed data in the matrix form $Y_{n\times 1} = X_{n\times p}\beta_{p\times 1}+\epsilon_{n\times 1}$ with $n = 6$ and $p = 2$. From this viewpoint, using the linear regression to

(i) estimate the weights of balls A and B; and

(ii) find a 70% **confidence interval** on the weight of ball A.

(iii) Suppose the student plans to measure the weight of ball A one more time. Find a 70% **prediction interval** on the new observed weight of ball A.

(b) Another approach is to consider the average weights, and this yields the following data table:

	A	B	$A + B$
"New Observed Average Weights"	2	1.5	4.5

Repeat part (a) by writing the "new observed average weights" in the matrix form $Y_{n\times 1} = X_{n\times p}\beta_{p\times 1}+ \epsilon_{n\times 1}$ with $n = 3$ and $p = 2$.

(c) Compare your results in (a) and (b).

Problem 2. Consider a simple linear regression model $Y = \beta_0 + \beta_1 x + \epsilon$. Suppose that we choose m different values of the independent variables x_i's, and each choice of x_i is duplicated, yielding k independent observations $Y_{i_1}, Y_{i_2}, \cdots, Y_{i_k}$. Is it true that the least squares estimates of the intercept and slope can be found by doing a regression of the mean responses, $\bar{Y}_i = (Y_{i_1} + Y_{i_2} + \cdots + Y_{i_k})/k$, on the x_i's? Why or why not? Explain.

Hints: this is a generalization of Problem 1. There are two kinds of linear regressions: one is based on a total of $n = mk$ "raw" observations (Y_i, x_i)'s, and the other is based on the m "average" observations $(\bar{Y}_i, x_i)$. See the hint pdf file for more details.

Problem 3 (R exercise). Consider the *zipcode* data, which are available from the book website: <www-stat.stanford.edu/ElemStatLearn>. You can also find it at Canvas. In the *zipcode* data, the first column stands for the response (Y) and the other columns stand for the independent variables $(X_i$'s). The detailed description can be found from

http://statweb.stanford.edu/~tibs/ElemStatLearn/datasets/zip.info.txt

Here we consider only the classification problem between 2's and 7's.

(1) Let us first obtain the training data. The following R code can yield the desired training data named as "ziptrain27"

```
ziptrain <- read.table(file="http://www.isye.gatech.edu/~ymei/7406/Handouts/zip.train.csv",
              sep = ",");
ziptrain27  <- subset(ziptrain, ziptrain[,1]==2 | ziptrain[,1]==7);
```

(2) **Exploratory Data Analysis**. Play with the training data "ziptrain27" and report some summary information/statistics of training data that you think are important or interesting. Some of the following R code might be useful, but please do not copy and paste the R results — you need to be selective, and use your own language to write up some sentences to summarize those important or interesting R results).

```
dim(ziptrain27);
sum(ziptrain27[,1] == 2);
summary(ziptrain27);
round(cor(ziptrain27),2);
## To see the letter picture of the 5-th row by changing the row observation to a matrix
rowindex = 5;  ## You can try other "rowindex" values to see other rows
ziptrain27[rowindex,1];
Xval = t(matrix(data.matrix(ziptrain27[,-1])[rowindex,],byrow=TRUE,16,16)[16:1,]);
image(Xval,col=gray(0:1),axes=FALSE) ## Also try "col=gray(0:32/32)"
```

(3) Using the training data "ziptrain27" to build the classification rule by (i) linear regression; and (ii) the KNN with $k = 1, 3, 5, 7$ and 15. Find the training errors of each choice.

```
### linear Regression
mod1 <- lm( V1 ~ . , data= ziptrain27);
pred1.train <- predict.lm(mod1, ziptrain27[,-1]);
y1pred.train <- 2 + 5*(pred1.train >= 4.5);
mean( y1pred.train  != ziptrain27[,1]);
## KNN
library(class);
kk <- 1;
xnew <- ziptrain27[,-1];
ypred2.train  <- knn(ziptrain27[,-1], xnew, ziptrain27[,1], k=kk);
mean( ypred2.train != ziptrain27[,1])
```

(4) Let us consider the testing data set, and derive the testing errors of each classification rule in (3).

```
ziptest <- read.table(file="http://www.isye.gatech.edu/~ymei/7406/Handouts/zip.test.csv",
              sep = ",");
ziptest27  <- subset(ziptest, ziptest[,1]==2 | ziptest[,1]==7);

## Testing error of KNN
kk <- 1;
xnew2 <- ziptest27[,-1];
ypred2.test  <- knn(ziptrain27[,-1], xnew2, ziptrain27[,1], k=kk);
mean( ypred2.test != ziptest27[,1])
```

Based on the above analysis, write some paragraph to provide a brief summary of what you discover. summarize your results.

the training data "ziptrain27" to build the classification rule by (i) linear regression; and (ii) the KNN with $k = 1, 3, 5, 7$ and 15.