HW5 (Due Mar 4)

You may work in groups of two. Both names must appear on the turned-in assigment. Note that you may not work with someone who was in your group previously in this class.

1) Generate a dataset using the following code:

```
x1=rexp(4000,10)
x2=rgamma(4000,10)
x3=4-2*x2+rnorm(4000,sd=4)
y=1000-2*x1-5*x2^(-.2)-4*x3+rnorm(4000,sd=10)
data=data.frame(x1,x2,x3,y)
```

NOTE: YOU MUST PROVIDE ALL CODE, OUTPUT (INCLUDING GRAPHS AND OUTPUT OF RELEVANT R-FUNCTIONS) AS WELL AS ANY REQUIRED ANSWERS.

(a) Use an array of scatterplots (created using the par(mfrow()) function to search for a good power transformation on the x2 variable to make y linear with respect to the transformed variable. You should consider at least 10 powers between -2 and 2. Which power transformation do you think is best?

(b) Use the boxcox() function to identify a good parameter λ such that y is linear with respect to $\frac{x_2^{\lambda}-1}{\lambda}$. Estimate a regression model of y with respect to x1, the transformed variable, and x3.

(c) Check for correlations between the variables and use Principle Component Analysis to come up with 3 predictor variables which have no (or low) correlation with one another. Obtain an estimated regression model of y with respect to these principle components.

(d) Use a Scree plot to determine a good number of principle components to use. Obtain an estimated regression model of y with respect to these principle components.

2) Consider the dataset:

Observation	1	2	3	4	5	6
x_1	-8/3	-2/3	10/3	1/3	4/3	-5/3
x_2	37/6	7/6	-11/6	-23/6	-11/6	1/6
y	2	8	12	10	10	8

It can be shown that variables $x_1 and\ x_2$ have high correlation. To solve this issue princinple component analysis is performed in R, which produces the following output:

```
> pca$x
          PC1         PC2
[1,] -6.6988350 -0.51429387
[2,] -1.3429159  0.04617916
[3,]  3.1754553 -2.09492388
[4,]  3.5469963  1.49143322
[5,]  2.2434412 -0.32536184
[6,] -0.9241419  1.39696720
> pca$rotation
        PC1       PC2
x1  0.466007 -0.884781
x2 -0.884781 -0.466007
> pca$sdev
[1] 3.878907 1.336943
```

Note that the columns x_1 and x_2 have mean 0.

(a) State the formula that is used to come up with the new variables (Principle Components) based on the original predictor variables.

(b) State the mathematical expression that produces the number 3.1754553 (highlighted in green).

(c) State code that could be used to estimate the ordinary linear regression model

$$y = \alpha + \beta_1 PCA1 + \beta_2 PCA2 + \varepsilon$$

(d) A model for y using the first principle component is estimated as

$$y = 8.3333 + \ .8571 PCA1 + \varepsilon$$

Predict the y- value of an observation for which $x1 = 2$ and $x_2 = -2$

(e) Sketch by hand a screeplot. That is, a plot showing the variances of the two principle components.

3) Use the following code to generate data which follows a Logistic Regression Model:

```
x1=sample(1:10,200,replace=TRUE)
x2=sample(1:10,200,replace=TRUE)
prob=1/(1+exp(-(2+4*x1-5*x2)))
y=rbinom(n=200, size=1, prob=prob)
```

Answer each of the following and state all code that was used:

(a) Estimate a logistic regression model using R. Write out the complete estimated regression model (of how y relates to the x1 and x2 variables).
(b) Compute the misclassification rate for the estimated model.
(c) Use the estimated model to predict the whether y=0 or 1 for the first observation.
(d) Use the estimated model to predict the probability that y=1 for the second observation.