36-350: Data Mining

Homework 5
Date: September 22, 2003 **Due: start of class September 29, 2003**

---

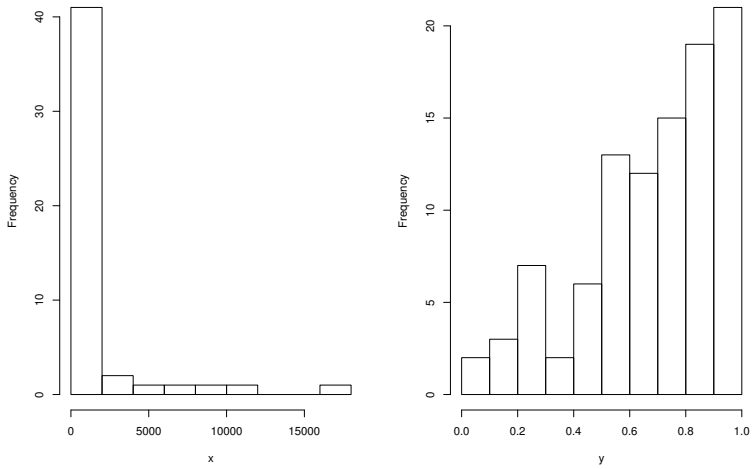
1. Below is the PCA projection matrix for the car data shown in class. All of the variables were standardized to have zero mean and unit variance.

	h1	h2
Price	-0.3	-0.5
MPG.highway	0.3	0.0
EngineSize	-0.3	-0.1
Horsepower	-0.3	-0.5
Passengers	-0.2	0.6
Length	-0.3	0.1
Wheelbase	-0.3	0.2
Width	-0.4	0.1
Turn.circle	-0.3	0.1
Weight	-0.4	0.0

   (a) If a car is one standard deviation above average in `Price`, one standard deviation below average in `MPG.highway`, and average on all other variables, what are its coordinates in the `(h1,h2)` projection?

   (b) If a car is one standard deviation above average in `Passengers`, one standard deviation above average in `Weight`, and average on all other variables, what are its coordinates in the `(h1,h2)` projection?

2. Below is plotted the distributions of two variables, x and y. Both are never equal to zero. For each variable, suggest a transformation in the power family which will make its distribution more symmetric.



3. In the computer lab, you made a matrix of scatterplots involving `Income`, `Illiteracy`, `HS.Grad`, and `Density`. In the plot of `Illiteracy` versus `Density`, there is an outlier point, which has an unusually high `Illiteracy` for its `Density`. Tracing this point to the other plots, what other unusual properties does it have?

4. (Long answer question) In the computer lab, you made a PCA projection of the States data. In this question, you will interpret it.

   (a) According to the plot, which variables are positively correlated with graduating high school (`HS.Grad`)? Which are negatively correlated? In each case, give a possible explanation.

   (b) Within the range of the data, you should find a few voids, where relatively few states appear. Where are they and what does each one mean?

   (c) The data distribution also has a few "funnels," directions where the states become increasingly compact. What are they and what does each one mean?

   (d) If the R^2 value of the projection was much lower, what would be the danger in interpreting the plot, as you did above?