

# 36-350: Data Mining

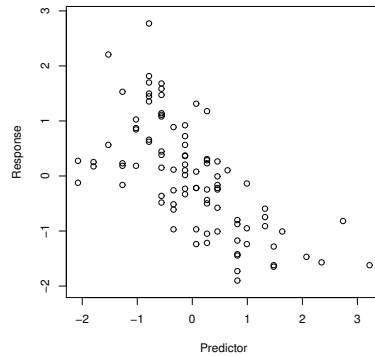
## Homework 7

Date: October 6, 2003

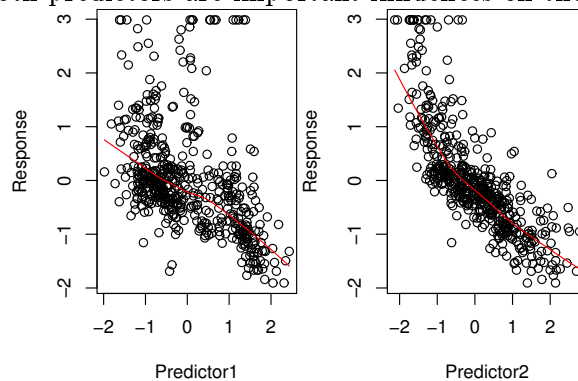
Due: start of class October 13, 2002

---

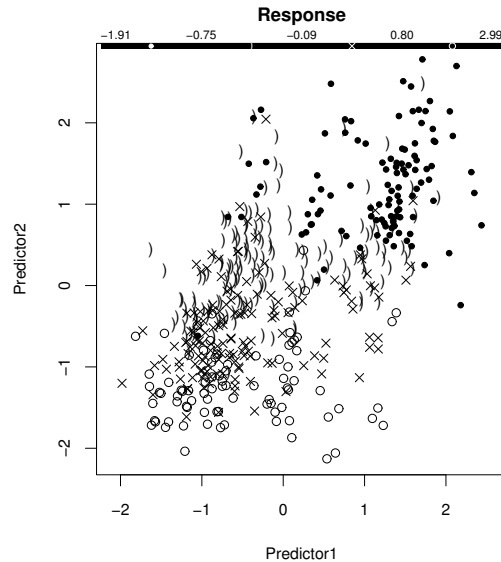
1. Below is a scatterplot of a predictor variable and a response variable.



- (a) Draw a lowess-type trend line through this data, showing the expected response for each value of the predictor.
- (b) As the predictor value increases, does the expected response always decrease? Explain.
2. In trying to predict a response variable, Louis Reasoner makes the following pairwise scatterplots, and concludes that both predictors are important influences on the response.



You, on the other hand, make a contour plot of the data, shown below. (The response has been divided into groups which are coded with different symbols.)



- (a) Draw approximate straight line boundaries between the response groups.
  - (b) Are both predictors important?
  - (c) Do the predictors interact?
  - (d) What important property of the variables has Louis failed to take into account?
3. In the computer lab, you made a PCA projection of the neighborhood data. The data should look like a tilted square, with the variable axes (arrows) clustered in such a way that they form a cross shape on the square. Each cluster contains variables which are highly correlated or anti-correlated. Call them Axis 1 and Axis 2.
- (a) What do you think Axis 1 and Axis 2 correspond to in real life? Can you explain the correlations between variables?
  - (b) Which Axis has the most to do with house price?
4. In the computer lab, you made a contour plot of house `Price` against `Distance` and `Low.Status`.
- (a) Are `Distance` and `Low.Status` correlated?
  - (b) Do `Distance` and `Low.Status` interact in predicting `Price`?
  - (c) In the pairwise scatterplot, `Price` always increases with `Distance`. What happens when `Low.Status` is taken into account? When does `Price` increase with `Distance`?
5. In the computer lab, you made a contour plot of house `Price` versus a projection of the predictors. The projection has two dimensions, `h1` and `h2`.
- (a) The contours show an interaction between `h1` and `h2`: in part of the plot, prices only depend on `h1`, while elsewhere prices depend on `h1` and `h2` equally. What parts of the plot are these and what types of neighborhoods do they correspond to?
  - (b) When prices only depend on `h1`, which three predictors have most weight? Which three have least weight?
  - (c) When prices depend on `h1` and `h2` equally, which three predictors have most weight? Which three have least weight? What are the major differences from part (b)?