- Academic honesty is important. Strict plagiarism policy will be enforced.
- You can work in a group of 2–3 people. (Group with different people in each HW.)
- Attach your R code after your typed-homework.
- Send also your R code to sta208.spring2015.ucd@gmail.com in a zipped file.

1. Consider a linear regression model with p parameters, fit by least squares to a set of training data $(x_1, y_1), \ldots, (x_N, y_N)$ drawn at random from a population. Let $\hat{\beta}$ be the least squares estimate. Suppose we have some test data $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_M, \tilde{y}_M)$ drawn at random from the same population as the training data. If $R_{tr}(\beta) = \frac{1}{N}\sum_{i=1}^{n}(y_i - \beta^T x_i)^2$ and $R_{te}(\beta) = \frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i - \beta^T \tilde{x}_i)^2$.

   (a) Show that $ER_{tr}(\hat{\beta}) < \sigma^2$, where σ^2 is the noise variance of the linear regression model.

   (b) Show that $ER_{te}(\hat{\beta}) > \sigma^2$.

   (c) What can you say about (a) and (b)?

   (d) Perform a simulation study to validate the above results. (Hint: repeat the experiment 100 times, in each trial, record $R_{tr}(\hat{\beta})$ and $R_{te}(\hat{\beta})$. Choose $N = 100$, $M = 100$ and $p = 10$. The predictors and errors are i.i.d. standard normal.)

2. Given data on two variables X and Y, consider fitting a cubic polynomial regression model $f(X) = \sum_{j=0}^{3}\beta_j X^j$. In addition to plotting the fitted curve, you would like a 95% confidence band about the curve. Consider the following tow approaches:

   (a) At each point x_0, form a 95% confidence interval for the linear function $\sum_{j=0}^{3}\beta_j x_0^j$.

   (b) Form a 95% confidence set of β as

   $$\{\beta : (\hat{\beta} - \beta)^T X^T X(\hat{\beta} - \beta) \leq \hat{\sigma}^2 \chi^2_{p+1,1-\alpha}\},$$

   then generate a confidence interval of $f(x_0)$.

   How do these approaches differ? Which band is likely to be wider? Conduct a small simulation experiment to compare the two methods by using the `Income` data with $Y =$ `Income` and $X =$`Years of Education`.

3. Download the file "hw1prob3.Rdata" from the course website and load it into your R session. Now you should have the predictor matrix x, which contains $n = 800$ points in $p = 2$ dimensions. Each point falls into either class 0 or 1. There are two scenarios for the class labels, given by y_1 and y_2.

(a) Plot the data in x with the class labels given by y_1. (Use the option `col` or `pch` or both to distinguish between the classes.) Run logistic regression, using the `glm` function with `family="binomial"`, to build a prediction rule. What is the training misclassification rate of this rule?

(b) Draw the decision boundary in $\mathbb{R}^2$ of the logistic regression rule from (a), on top of your plot from (a). What shape is it? Does this boundary look like it adequately separates the classes?

(c) Run logistic regression on the predictors in x, as well as the predictor x_1^2 (square of the first column of x). This is analogous to adding a quadratic term to a linear regression. What is the training misclassification rate of this rule? Why is this better than the rule from (a)?

(d) What is the shape of the decision boundary of the logistic regression rule from (c)? Draw this decision boundary on top of a plot of the (appropriately color-coded or pch-coded) data x. What shape is it?

(e) Plot the data in x with the labels given by y_2. Try a running logistic regression of y_2 on x, and also on $[x, x_1^2]$. What are the training misclassification rates of these rules? Draw the decision boundaries of each rule on top of a plot of the data.

(f) Why are neither of the decision boundaries from (e) adequate? What additional predictors can you pass to logistic regression in order for it to do a better job of separating the classes? (Hint: draw a curve between the classes by eye... what shape does this have?) Run a logistic regression with these additional predictors, report the training misclassification rate, and draw the new decision boundary. What shape is it?

(g) If adding polynomial terms seems to improve the training misclassification rate of the logistic regression rule, why dont we generally just keep including polynomial terms of higher and higher order? How could we choose how many polynomial terms to include in a principled manner?