STATS 216 Introduction to Statistical Learning
Stanford University, Winter 2015

**Problem Set 3**

**Due:** Wednesday, February 25, 2015
Remember the university honor code. All work and answers must be your own.

---

1. Consider two curves, $\hat{g}_1$ and $\hat{g}_2$, defined by

$$\hat{g}_1 = \arg\min_g \left( \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[ g^{(3)}(x) \right]^2 dx \right),$$

$$\hat{g}_2 = \arg\min_g \left( \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[ g^{(4)}(x) \right]^2 dx \right),$$

where $g^{(m)}$ represents the $m$th derivative of $g$.

   (a) As $\lambda \to \infty$, will $\hat{g}_1$ or $\hat{g}_2$ have the smaller training RSS?

   (b) As $\lambda \to \infty$, will $\hat{g}_1$ or $\hat{g}_2$ have the smaller test RSS?

   (c) For $\lambda = 0$, will $\hat{g}_1$ or $\hat{g}_2$ have the smaller training and test RSS?

2. Suppose that we carry out backward stepwise, forward stepwise, and best subset all on the same data set. Each approach will yield a sequence of models with $k = 0$ up through $k = p$ predictors.

   (a) Which approach with $k$ predictors will have the smallest *test* residual sum of squares? Explain.

   (b) Which approach with $k$ predictors will have the smallest *training* residual sum of squares? Explain.

   (c) True or False:

      i. The predictors in the $k$-variable model identified by forward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by backward stepwise selection.

      ii. The predictors in the $k$-variable model identified by backward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by forward stepwise selection.

      iii. The predictors in the $k$-variable model identified by best subset are a subset of the predictors in the $(k+1)$-variable model identified by best subset selection.

      iv. The predictors in the $k$-variable model identified by backward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by backward stepwise selection.

v. The predictors in the $k$-variable model identified by forward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by forward stepwise selection.

Explain each answer.

3. **You may work in groups up to size 4 on this problem. If you do work in groups, write the names of all your group members on your problem set.**

   This question uses the variables `dis` (the weighted mean of distances to five Boston employment centers) and `nox` (nitrogen oxides concentration in parts per 10 million) from the `Boston` data. We will treat `dis` as the predictor and `nox` as the response.

   (a) Use the `poly()` function to fit a cubic polynomial regression to predict `nox` using `dis`. Report the regression output, and plot the resulting data and polynomial fits.

   (b) Plot the polynomial fits for a range of different polynomial degrees (say, from 1 to 10), and report the associated residual sum of squares.

   (c) Perform cross-validation or another approach to select the optimal degree for the polynomial, and explain your results.

   (d) Use the `bs()` function to fit a regression spline to predict `nox` using `dis`. Report the output for the fit using four degrees of freedom. How did you choose the knots? Plot the resulting fit.

   (e) Now fit a regression spline for a range of degrees of freedom, and plot the resulting fits and report the resulting RSS. Describe the results obtained.

   (f) Perform cross-validation or another approach in order to select the best degrees of freedom for a regression spline. Describe your results.

4. **You may work in groups up to size 4 on this problem. If you do work in groups, write the names of all your group members on your problem set.**

   This problem works with the `body` dataset used in the in class session from Feb 11. The goal of this problem is to perform and compare Principal Components Regression and Partial Least Squares on the problem of trying to predict someones weight. While you can use any R tools at your disposal to complete the problem, `library(pls)` and `Lab 3` from `ISLR` will probably be very helpful and the problem set was written with these approaches in mind. If you have not already downloaded the data, please go to the class coursework page and do so. More information about this dataset can be found at `http://www.amstat.org/publications/jse/v11n2/datasets.heinz.html`.

   (a) Read the `body` dataset into R using the `load()` function. This dataset contains:

   - `X`: A dataframe containing 21 different types of measurements on the human body.
   - `Y`: A dataframe that contains the age, weight (kg), height (cm), and the gender of each person in the sample.

Let's say we forgot how the gender is coded in this dataset. Using a simple visualization, explain how you can tell which gender is which.

(b) Reserve 200 observations from your dataset to act as a test set and use the remaining 307 as a training set. On the training set, use both `pcr` and `plsr` to fit models to predict a person's weight based on the variables in `X`. Use the options `scale = TRUE` and `validation='CV'`. Why does it make sense to scale our variables in this case?

(c) Run `summary()` on each of the objects calculated above, and compare the training % variance explained from the `pcr` output to the `plsr` output. Do you notice any consistent patterns (in comparing the two)? Is that pattern surprising? Explain why or why not.

(d) For each of models, pick a number of components that you would use to predict future values of weight from `X`. Please include any further analysis you use to decide on the number of components.

(e) Practically speaking, it might be nice if we could guess a person's weight without measuring 21 different quantities. Do either of the methods performed above allow us to do that? If not, pick another method that will and fit it on the training data.

(f) Compare all 3 methods in terms of performance on the test set. Keep in mind that you should only run one version of each model on the test set. Any necessary selection of parameters should be done only with the training set.