Homework 1

Stat 151A, Fall 2017

Due September 12, 2017

Definition 1 *Consider the linear model*

$$y \;=\; X\beta \;+\; e, \quad E(e) \;=\; 0,$$

$X \in \mathbb{R}^{n\times p}$, $\beta \in \mathbb{R}^p$. *A vector value linear function of* β, *say* $\Lambda^T\beta$, *is estimable if* $\Lambda^T\beta \;=\; P^T X\beta$ *for all* $\beta \in \mathbb{R}^p$, *for some matrix* P.

Notation Throughout if A is matrix $C(A)$ denotes its column space.

1. Consider the data: $Y_i = \beta_0 + \beta_1 + e_i$ where $e_1, \ldots, e_n$ are uncorrelated errors with mean zero and variance σ^2.

   (a) (.5 pts) Write this model in the form $Y = X\beta + e$ with $\beta = (\beta_0, \beta_1)^T$. Specify the matrix X.

   (b) (.5 pts) Write down the normal equations. Find a solution to them. Is the solution unique?

   (c) (.5 pts) What is the least squares estimate of $\beta_0 + \beta_1$?

   (d) (.5 pts) Is β_1 estimable?

   (e) (.5 pts) Consider now another observation $Y_{n+1} = \beta_0 + 2\beta_1 + e_{n+1}$ where $e_1, \ldots, e_{n+1}$ are uncorrelated errors with mean zero and variance σ^2. Write this model in the form $Y = X\beta + e$ and calculate the least squares estimate of β.

2. Suppose y_1, y_2, y_3, y_4 are uncorrelated random variables with common variance σ^2, and

   $$E(y_1) \;=\; a, \quad E(y_2) \;=\; a - b, \quad E(y_3) \;=\; a + b, \quad E(y_4) \;=\; b,$$

   for some real numbers a and b.

   (a) (.5 pts) Find a matrix X such that $E(y) \;=\; X\,\beta$, where $y \;=\; (y_1, y_2, y_3, y_4)^T$, and $\beta \;=\; (a, b)^T$.

   (b) (1 pts) Find a basis of $C(X)$. What is the rank of X?

   (c) (1 pts) Find the ordinary least square estimates of a and b.

3. (1.5pt)(Excersise B.14 from R. Christensen 2011) Let M_1 and M_2 be perpendicular (orthogonal) projection matrices, and let M_0 be a perpendicular projection operator onto $C(M_1) \cap C(M_2)$. Show that the following are equivalent:

   (a) $M_1 M_2 = M_2 M_1$.

   (b) $M_1 M_2 = M_0$.

4. Auto-mpg analysis: The data file `auto-mpg.data` is avaliable at `https://archive.ics.uci.edu/ml/datasets/auto+mpg`, and includes gas mileage info for a variety of cars from the 1980s, in addition to other features. In this problem we consider that 'mpg' is the response variable of interest. The attributes are (see read me file as well):

1. mpg:	continuous
2. cylinders:	multi-valued discrete
3. displacement:	continuous
4. horsepower:	continuous
5. weight:	continuous
6. acceleration:	continuous
7. model year:	multi-valued discrete
8. origin:	multi-valued discrete
9. car name:	string (unique for each instance)

   (a) (1 pts) Exploratory Data Analysis (EDA): Prepare scatterplots, boxplots, pairs plot with smoothing lines, co-plot, density estimators plot. Discuss what you observe.

   (b) (1 pts) Carry out an Ordinary Least Squares analysis with gas mileage as response variable and other features as explanatory variables (and include an intercept). Write an OLS program (your own code) using linear algebra as discussed in class. Your output should include: the coefficient estimates, the residual sum of squares, the SSreg, and the R^2. How do your results compare with those provided by the function *lm()*?

   (c) (1 pts) Create a residual versus fitted plot from the regression above. Discuss. Are there any outliers?

   (d) (.5 pts) What can you conclude from your overall analysis?