LINEAR REGRESSION MODELS W4315

HOMEWORK 6 ANSWERS

March 9, 2010

Due: 03/09/2010
Instructor: Frank Wood

1. (15 points) Refer to the data 'hw6p1.dat' on the course website and read into MATLAB as the design matrix(the first column is already added as a **1** vector interpreted as the intercept). Check 'fread' command using the help system. What is the most complex model in terms of number of parameters that one could fit to this data?
Extra credits: Which parameters should be non-zero?
Hint: Consider the rank of the design matrix, and read about principle component analysis(PCA) online.

Answer:

The design matrix $\mathbf{X}$ is a $n \times p$ matrix, where n is the number of observations for each predictor variable and (p-1) is the number of predictor variables. In this problem, n=20 and p=61.

In ordinary linear regression analysis, one of the key assumptions is that the design matrix $\mathbf{X}$ must have full column rank p. Since $n < p$ for this data, $\mathbf{X}$ does not have full rank, some of the parameters are not identifiable. One way to see how many parameters could be identified is to use the matlab function "rank".

Matlab Code:
format long
$X = load('hw6p1.dat')$
$[n, p] = size(X)$
$rk = rank(X)$

This shows that the rank of X is 10. To fit more that 10 parameters, PCA or regularization of some form must be employed. These will be covered later in the class.

2. (45 points) Suppose $X_1, ..., X_n$ are i.i.d. samples from $N(0, \sigma^2)$. Denote $\bar{X}$ as the

sample mean. Prove $S = \sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \sigma^2\chi^2(n-1)$ following the steps below using Cochran's theorem:

a. Remember that we have the decomposition

$$\sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n\bar{X}^2 \tag{1}$$

Show the matrices corresponding to all the three quadratic terms in (3).

b. Derive the rank of each matrix above. (Hint: Recall problem 3 in homework 5.)

c. Use Cochran's theorem to prove $S \sim \sigma^2\chi^2(n-1)$.

Answer:

a. Denote $\mathbf{X} = (X_1, ..., X_n)'$, then we have the matrix form of (1) as

$$\mathbf{X}'\mathbf{X} = \mathbf{X}'(\mathbf{I} - \tfrac{1}{n}\mathbf{J})\mathbf{X} + \mathbf{X}'\tfrac{1}{n}\mathbf{J}\mathbf{X}$$

where $\mathbf{J}$ is an $n \times n$ matrix whose elements are all 1. So the corresponding matrics for (1) are respetively $\mathbf{I}, \mathbf{I} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{J}, \frac{\mathbf{1}}{\mathbf{n}}\mathbf{J}$.

b.

Since $\mathbf{I}$ is an $n \times n$ identity matrix, so $r(\mathbf{I}) = n$.

For the 2nd matrix, it's easy to verify that $\mathbf{I} - \frac{1}{n}\mathbf{J}$ is idempotent, since

$$\begin{aligned}(\mathbf{I} - \frac{1}{n}\mathbf{J})(\mathbf{I} - \frac{1}{n}\mathbf{J}) &= \mathbf{I} - \frac{2}{n}\mathbf{J} + \frac{1}{n^2}\mathbf{J}^2 \\ &= \mathbf{I} - \frac{2}{n}\mathbf{J} + \frac{1}{n^2}n\mathbf{J} \\ &= \mathbf{I} - \frac{2}{n}\mathbf{J} + \frac{1}{n}\mathbf{J} \\ &= \mathbf{I} - \frac{1}{n}\mathbf{J}\end{aligned}$$

The 2nd equation holds since $\mathbf{J}^2 = n\mathbf{J}$. Now that $\mathbf{I} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{J}$ is idempotent, we have

$$\begin{aligned}rk(\mathbf{I} - \frac{1}{n}\mathbf{J}) &= tr(\mathbf{I} - \frac{1}{n}\mathbf{J}) \\ &= tr(\mathbf{I}) - tr(\frac{1}{n}\mathbf{J}) \\ &= n - \frac{1}{n} \times n \\ &= n - 1\end{aligned}$$

where $tr(\mathbf{A})$ stands for the trace of matrix $\mathbf{A}$, $rk(\mathbf{A})$ for the rank of it. It's apparent from the definition of rank of a matrix that $rk(\frac{1}{n}\mathbf{J}) = 1$.

c. Since $rk(\mathbf{I}) = rk(\mathbf{I} - \frac{1}{n}\mathbf{J}) + rk(\frac{1}{n}\mathbf{J})$, we can directly apply Cochran's theorem, so $S \sim \sigma^2\chi^2(n-1)$.