LINEAR REGRESSION MODELS W4315

Midterm Examination QUESTIONS

November 23, 2009

Instructor: Frank Wood (10:35-11:50)

1. (40 points) You are a statistical analysis consultant who has been asked to develop a regression model to analyze data collected in the following way: a team of two food scientists was tasked with developing a new super-fluffy pancake recipe (fluffiness is a made up concept but for our purposes will be defined as measurable scalar quantity denoted f) based on experimenting with including varying levels of a secret new ingredient (the concentration of which is denoted by i). Nearly everything went well, each scientist prepared different recipes with varying levels of the secret new ingredient (i_s^r where $s = 1$ indicates the first scientist and $s = 2$ indicates the second scientist and $r \in [1, \ldots, n_s]$ is an integer that indicates which of the n_s pancake recipes tried by scientist s the input quantity corresponds to. Each scientist measured the fluffiness using their own machine, producing fluffiness values f_s^r. Unfortunately the variance of the fluffiness measurement produced by the scientists two machines was different (this is the reason why you were called). Each scientist produced measurement whose errors where identically and independently distributed but whose variance was different than the other (σ_s^2 is the fluffiness measurement variance for scientist s).

(a) Using matrix notation ($\vec{f} = [f_1^1, \ldots, f_1^{n_1}, f_2^1, \ldots, f_2^{n_2}]^T$, etc.) set up a normal regression problem and derive the maximum likelihood estimates for the regression coefficients $\vec{\beta} = [\beta_0, \beta_1]^T$ under the given assumptions. *Hint : use a vector like* $\vec{\sigma^2} = [\sigma_1^2, \ldots, \sigma_1^2, \sigma_2^2, \ldots, \sigma_2^2]^T$ *where* σ_s^2 *is repeated* n_s *times.*

(b) Provide the maximum likelihood estimators for both σ_1^2 and σ_2^2.

2. (30 points) Once again you are a statistical consultant sent in to help solve a difficult regression analysis problem. You have been asked to fit a second order polynomial regression through the origin model to the following data

$$\begin{bmatrix} X_1 & Y \\ 1 & 1 \\ 1 & 2 \end{bmatrix}$$

(a) Using matrix notation, set up the specified regression model. Show the design and response matrices explicitly.

(b) Using matrix rank identities, prove that the regression problem as specified lacks a unique solution. Explain the intuition behind your answer in no more than two sentences.

(c) In one sentence each, explain two options for solving this problem that don't involve changing the number of parameters in the model (i.e. keeping the model a 2^{nd} order polynomial regression model).

3. (30 points) The F^* test statistic for the general linear test given by

$$F^* = \frac{\frac{SSE(R)-SSE(F)}{df_R-df_F}}{\frac{SSE(F)}{df_F}} \sim F(df_R - df_F, df_F) \tag{1}$$

can be used in tests of whether or not a full model (denoted by F) versus a reduced model (denoted by R) would best explain the data. This test statistic can also be used to decide whether or not input variables should be included in a linear regression model.

In multiple regression (where $E\{Y\} = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1}$) the test statistic

$$F^* = \frac{\frac{SSR(X_1,\ldots,X_{p-1})}{p-1}}{\frac{SSE(X_1,\ldots,X_{p-1})}{n-p}} \tag{2}$$

is used to test whether there is a regression relation between the response variable Y and the set of X variables. In other words this test statistic can be used to perform (among others) the hypothesis test

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0 \\ H_a &: \text{not all } \beta_k (k = 1, \ldots, p-1) \text{ equal } 0 \end{aligned} \tag{3}$$

(a) State the hypothesis test in (3) in the form of the general linear test in (1). In other words, identify $SSE(R)$, $SSE(F)$, df_R, and df_F for the hypothesis test.

(b) Show that the result for (a) is equivalent to the test statistic given in (2).

(c) If $p = 5$ and $n = 60$ and $F^* = 5.78$ what should you conclude? State your conclusion in words (1 sentence).