2011 COMPREHENSIVE WRITTEN EXAMINATION, PAPER III

STOR 664 Questions

Problem 1 (55 points)

Answer Parts (a), (b), (c) based on the following:

Suppose we need to weigh four objects using a scale, which will produce measurements with uncorrelated, zero mean errors of common variance σ^2. Cost constraints will allow us to take up to 12 measurements in total. Comparing the following strategies:

(i) Weigh each object 3 times on its own.

(ii) Weigh each of the 6 possible combinations of pairs of objects together twice.

(a) (7 points) Express strategies (i) and (ii) in terms of general linear models. Explain clearly your notations and model assumptions.

(b) (20 points) Derive least squares estimators of each of the weights using both strategies. Compute the corresponding variances of the estimators and explain which strategy is more preferable.

(c) (13 points) Suppose instead of the stated assumption about σ^2, the standard deviation of a reading is proportional to the number of objects being weighed. Which of the two weighing strategies is then superior? Explain.

(d) (15 points) *This part is independent of the other parts.*
Let α, β, γ be the interior angles of a triangle so that $\alpha + \beta + \gamma = 180$ degrees. Independent measurements A, B, C are obtained for α, β, γ respectively. What is the "best" estimate of α? In what sense is it best? (Clearly state the model that you use and any assumptions that you make. For this question, you can use any result learnt in class without proof but state the result you use.)

Problem 2 (45 points)

Suppose

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \tag{1}$$

for $i = 1, \ldots, n$, where $\{\epsilon_i\}_{i=1}^n$ are i.i.d. $N(0, \sigma^2)$ random errors. Assume that the matrix X whose i-th row is $(1, x_{i1}, \ldots, x_{ip})$ has rank $(p+1) < n$.

(a) (7 points) Without proof, provide a testing procedure of the hypotheses

$$H_0 : \beta_q = \cdots = \beta_p = 0 \qquad \text{versus} \qquad H_a : \text{ not } H_0$$

where q satisfies $1 \leq q \leq p$.

(b) (17 points) Prove that the test in (a) remains unchanged if a fixed constant is added to each Y_i.

(c) (15 points) Let $\hat{\beta}_0, \hat{\beta}_1, \ldots \hat{\beta}_p$ be the least squares estimates of $\beta_0, \ldots, \beta_p$ based on model (1). The partial residual plot of the j-th regressor based on this model (j fixed, $1 \leq j \leq p$) is the scatterplot of the points $\{(x_{ij}, e_i^*)\}_{i=1}^n$, where

$$e_i^* = Y_i - \hat{\beta}_0 - \sum_{\ell=1, \ell \neq j}^{p} \hat{\beta}_\ell x_{i\ell}.$$

Show that if we consider the simple linear regression of the e_i^* on the x_{ij}, the least squares estimate of the slope of the fit is equal to $\hat{\beta}_j$.

(d) (6 points) Briefly explain the use of partial residual plots.

COMPREHENSIVE WRITTEN EXAMINATION, PAPER III

FRIDAY AUGUST 19, 2011, 9:00 A.M.–1:00 P.M.

STOR 665 QUESTIONS

1. **[60 points]** Suppose we observe $y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$, which is distributed according to the following multinomial distribution with probabilities

$$\left\{\frac{1}{2} + \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right\}.$$

   (a) **[5 points]** Write down the log-likelihood based on y.

   (b) **[15 points]** Derive the Fisher-Scoring algorithm for obtaining the maximum likelihood estimate (MLE) of θ.

   (c) **[40 points]** Now treat y as incomplete data from $x = (x_1, x_2, x_3, x_4, x_5)$ with the following multinomial probabilities

$$\left\{\frac{1}{2}, \frac{\theta}{4}, \frac{1-\theta}{4}, \frac{1-\theta}{4}, \frac{\theta}{4}\right\}.$$

   In addition, $y_1 = x_1 + x_2$, $y_2 = x_3$, $y_3 = x_4$, $y_4 = x_5$. The MLE of θ given y can be obtained using the EM algorithm.

   i. **[5 points]** State the incomplete data and the complete data.

   ii. **[5 points]** Derive the log-likelihood based on x.

   iii. **[5 points]** Calculate the corresponding MLE for θ given x.

   iv. **[15 points]** Explicitly derive the E-step.

   v. **[5 points]** Explicitly derive the M-step.

   vi. **[5 points]** Calculate the EM estimate for θ obtained in the first five iterations. Clearly state your initial estimate for θ, and calculations needed to obtain the EM estimate in each iteration.

2. **[40 points]** For each of the following problems, write down an appropriate model for the data, clearly indicating the parameters involved and your choice of method for estimation.

   (a) **[10 points]** The number of patients checking in at an emergency room is recorded for each hour j of every day i, denoted as N_{ij}. Exploratory data analysis suggests that each count follows a Poisson distribution with the rate λ_{ij}, depending on day i and hour j. State a two-way ANOVA model for the counts.

   (b) **[15 points]** For the above problem, suppose in addition that the daily effects are random and exhibit certain between-day dependence. State a reasonable choice of dependence, and modify the above model to incorporate such dependence.

(c) **[15 points]** The following histogram is about the waiting time (in minutes) of $N = 299$ consecutive eruptions of the Old Faithful geyser in Yellowstone National Park. The bimodal pattern suggests that the waiting times come from two populations, each of which can be modeled using a normal distribution.