DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST, 2002

Applied Statistics (Ph.D. Version)

Instructions to the Student

a. Answer all six questions. Each will be graded from 0 to 10.

b. Use a different booklet for each question. Write the problem number and your code number (**NOT YOUR NAME**) on the outside cover.

c. Keep scratch work on separate pages in the same booklet.

d. If you use a "well known" theorem in your solution to any problem, it is your responsibility to make clear which theorem you are using and to justify its use.

e. You may use calculators as needed.

---

1. Consider the general linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ where $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\mathbf{e}$ have dimensions $n \times 1$, $n \times p$, $p \times 1$, $n \times 1$, respectively, and $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{G})$. The matrices $\mathbf{G}$ and $\mathbf{X}$ have full rank.

(a) Derive the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$ and find an unbiased estimator of σ^2.

(b) Determine the joint distribution of the estimators found in (a).

(c) Show that the ordinary least squares estimator of $\boldsymbol{\beta}$ is unbiased. Can it be the BLUE when $\mathbf{G} \neq \mathbf{I}$?

2. Consider the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$$

where the e_i, $i = 1, \ldots, n$, are independent with mean 0 and variance σ^2. Let $\boldsymbol{\beta} = (\beta_1^T, \beta_2^T)^T$, and let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\beta}}_2^T)^T$ denote the vector of least squares estimates. Assume that $\mathbf{Y}$ is n-dimensional, that $\mathbf{X}_1$ has q columns and $\mathbf{X}_2$ has $p - q$ columns, and that $\mathbf{X}$, $\mathbf{X}_1$ and $\mathbf{X}_2$ have full rank.

(a) Suppose that one erroneously fits the model $\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{e_1}$ to the data, yielding $\tilde{\boldsymbol{\beta}}_1$ as the least squares estimate. Calculate the bias of $\tilde{\beta}_1$ and state the condition for the bias to be zero.

(b) Compute the variance-covariance matrix of $\tilde{\boldsymbol{\beta}}_1$.

(c) How would you test H_0: $\boldsymbol{\beta}_2 = \mathbf{0}$? (You do not need to provide a derivation of your test, but state clearly the formula for the test statistic and its distribution under H_0.)

3. Financial auditors regard the records of a firm as a population U consisting of N "accounts." Each account has a true value y_k and a recorded value x_k. The purpose of the audit is to estimate the total

$$t_{yU} = \sum_{k \in U} y_k = N\bar{y}_U,$$

while $t_{xU} = \sum_{k \in U} x_k$ is known. A simple random sample s, consisting of n accounts, is chosen and their y values are determined by the auditor. Two estimates of t_{yU} are possible: $\hat{t}_{y,srs} = N\bar{y}_s = (N/n)\sum_{k \in s} y_k$, and the difference estimator, $\hat{t}_D = t_{xU} + N\bar{d}_s$, where $d_k = y_k - x_k$ and $\bar{d}_s = \sum_{k \in s} d_k/n$.

(a) Show that $\hat{t}_D$ is an unbiased estimator of t_{yU}.

(b) Show that $\operatorname{Var}\hat{t}_D = N^2(1 - f)S_{dU}^2/n$, where $f = n/N$. Express S_{dU}^2 in terms of S_{xU}, S_{yU}, and

$$\rho = \frac{\sum_U (x_k - \bar{x}_U)(y_k - \bar{y}_U)}{(N - 1)S_{xU}S_{yU}}.$$

Here $S_{dU}^2 = \sum_{k \in U}(d_k - \bar{d}_U)^2/(N - 1)$, and S_{xU}^2 and S_{yU}^2 are defined similarly.

(c) Find conditions such that $\hat{t}_D$ has a smaller variance than $\hat{t}_{y,srs}$.

4. In an educational experiment, two teaching methods were compared. Four classrooms were selected at random from a population of first grade classrooms and assigned to method 1. An additional four classrooms were randomly selected and assigned to method 2. Each classroom teacher agreed to teach pre-reading skills to her students according to the method assigned to her class. The response was the child's reading readiness score after one month of instruction.

Let Y_{ijk}, $i = 1, 2$, $j = 1, \ldots, 4$, $k = 1, \ldots, 25$, denote the score of child k in classroom j receiving teaching method i.

(a) Write a model equation for Y_{ijk}, indicating clearly the side conditions on any fixed effect parameters and the assumed distributions for any random effects present.

(b) Write down the ANOVA table, including sums of squares, degrees of freedom and expected mean squares.

(c) How would you estimate the treatment means? What are the standard errors of your estimates?

5. A simple random sample s of size n is selected from a population U of size N. Let U_d be a domain of study in U with unknown size N_d. It is desired to estimate the domain total $t_d = \sum_{U_d} y_k$ from the sample. Let $s_d = s \cap U_d$ denote the set of sample elements belonging to the domain U_d and let n_d denote the number of elements in s_d.

(a) Show that the estimator $\hat{t}_d = (N/n) \sum_{s_d} y_k$ is unbiased and calculate its variance.

(b) Show that in large populations

$$\mathrm{Var}\,(\hat{t}_d) \doteq \frac{N^2}{n}\left(1 - \frac{n}{N}\right)(P_d S_d^2 + P_d Q_d \bar{y}_{U_d}^2),$$

neglecting terms in $1/N$ and $1/N_d$. Here $P_d = N_d/N$, $Q_d = 1 - P_d$, $\bar{y}_{U_d} = \sum_{U_d} y_k / N_d$ and $S_d^2 = \sum_{U_d} (y_k - \bar{y}_{U_d})^2/(N_d - 1)$.

6. Suppose that data (x_i, Y_{ij}), $i = 1, \ldots, k$, $j = 1, \ldots, n_i$, are available and the goal of the analysis is to fit a regression model

$$Y_{ij} = m(x_i) + e_{ij}$$

to the data. The error terms are assumed to be independent with a common $N(0, \sigma^2)$ distribution, and at least one of the n_i is greater than one.

(a) What statistic would you use to test H_0: $m(x) = \beta_0 + \beta_1 x$ against the general alternative? Give the formula for the test statistic and its distribution under H_0. You do not need to provide a derivation.

(b) If $n_i = 1$ for each i, how would you decide whether a straight line model fits the data?

(c) Suppose that the data are given in the following table:

x	10	10	15	20	20	25	25	25	30	35
y	73	78	85	90	91	87	86	91	75	65

A quadratic regression curve was fitted to the data, and the following partial ANOVA table was computed.

Source	Sum of Squares	d.f.	Mean Square
Linear regression	24.5	?	?
Addition of quadratic term	643.2	?	?
Lack of fit	?	?	?
Error	27.0	?	?
Corrected total	710.9	9	

The error sum of squares of the table is $\sum_i \sum_j (Y_{ij} - \bar{Y}_{i\cdot})^2$. Does the quadratic model seem to fit the data? Does it seem as if a linear model fits about as well?