DEPARTMENT OF MATHEMATICS
UNIVERSITY OF MARYLAND
GRADUATE WRITTEN EXAMINATION
AUGUST 2004

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 Appl Stat Problems

1.1 Estimation of the Angle of a Parallelogram

A surveyor measures once each of the angles of a parallelogram with angles $\theta, \pi - \theta, \theta, \pi - \theta$, and obtains one noisy observation on each angle. The noise components ϵ_i, $i = 1, 2, 3, 4$, have mean 0 and variance σ^2.

a. Estimate θ and σ^2.
b. Now, the surveyor wishes to estimate π in addition to θ. Is it possible? If it is, will this have any effect on the estimate of θ?

Solution

Write: $y_1 = \theta + \epsilon_1$, $y_2 = \pi - \theta + \epsilon_2$, $y_3 = \theta + \epsilon_3$, and $y_4 = \pi - \theta + \epsilon_4$, or

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix} \pi + \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} \theta + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \end{pmatrix} \tag{1}$$

Then the LSE of θ is:

$$\hat{\theta} = \mathbf{A}_{22}^{-1}\mathbf{X}_2'\mathbf{y}^* = \frac{1}{4}(1,-1,1,-1)\begin{pmatrix} y_1 \\ y_2 - \pi \\ y_3 \\ y_4 - \pi \end{pmatrix} = \frac{y_1 - y_2 + y_3 - y_4}{4} + \frac{\pi}{2} \tag{2}$$

and

$$\mathrm{Var}(\hat{\theta}) = \sigma^2 \mathbf{A}_{22}^{-1} = \sigma^2/4 \tag{3}$$

and $\hat{\sigma}^2 = \frac{1}{3}\sum_1^4 (y_i - \hat{\theta}_i)^2$.

If also π is estimated, then

$$\hat{\theta} = \frac{y_1 + y_3}{2} \tag{4}$$

with variance

$$\mathrm{Var}(\hat{\theta}) = \sigma^2 \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}_{11.2}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} + \sigma^2\mathbf{A}_{22}^{-1} = \sigma^2/2 \tag{5}$$

Substituting $y_1 \approx y_3 \approx \theta$, $y_2 \approx y_4 \approx \pi - \theta$, we see that both estimators (2) and (4) are sensible, $(y_1 - y_2 + y_3 - y_4)/4 + \pi/2 \approx \theta$ and $(y_1 + y_3)/2 \approx \theta$. We refer to (2) as the *standard* estimator and to (4) as the *redundant* estimator. Since both estimators are unbiased, we conclude from (3) and (5) that the standard estimator (2) is twice as efficient as the redundant one. Interestingly, $\hat{\pi} = (y_1 + y_2 + y_3 + y_4)/2 \approx \pi$.

1.2 Estimation of the Angles of a Triangle

A surveyor measures once each of the angles α, β, γ of an area that has the shape of a triangle, and obtains unbiased measurements Y_1, Y_2, Y_3 (in radians). It is known that $\mathrm{Var(Y_i)} = \sigma^2$, $i = 1, 2, 3$.

a. Estimate θ and σ^2.
b. Now, the surveyor wishes to estimate π in addition to θ. Is it possible? If it is, will this have any effect on the estimates of the angles?

Solution

The resulting linear model is

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 - \pi \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$

and the least squares estimates of the unknown angles are

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 2Y_1 - Y_2 - Y_3 + \pi \\ 2Y_2 - Y_1 - Y_3 + \pi \end{pmatrix}$$

with covariance matrix

$$\mathrm{Var} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \frac{\sigma^2}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$$

and $\hat{\sigma}^2 = (y_1 + y_2 + y_3 - \pi)^2/3$

On the other hand, if also π is estimated,

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \pi \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$

This time the least squares estimates are different yet sensible,

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \\ \hat{\pi} \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_1 + Y_2 + Y_3 \end{pmatrix}$$

and

$$\text{Var}\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\\ \hat{\pi}\end{pmatrix} = \sigma^2\begin{pmatrix}1 & 0 & 1\\ 0 & 1 & 1\\ 1 & 1 & 3\end{pmatrix}$$

It follows that if π is estimated then $\text{Var}(\hat{\alpha}) = \text{Var}(\hat{\beta}) = \sigma^2$, whereas if π is not estimated the estimates are more precise since $\text{Var}(\hat{\alpha}) = \text{Var}(\hat{\beta}) = 2\sigma^2/3$.

1.3 Simple Linear Regression

Consider the model $y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i$, where ϵ_i are independent $N(0, \sigma^2)$, and $E(y_i) = \eta_i$, $i = 1, ..., n$. Let $\hat{\alpha}, \hat{\beta}$ denote the least squares estimators. Define:

$$S^2 = \frac{1}{n-2}\sum_1^n (y_i - \hat{y}_i)^2$$

a. Obtain the least squares estimators $\hat{\alpha}, \hat{\beta}$.
b. Prove that $E(S) \leq \sigma$
c. Which of the three variables $\hat{\alpha}, \hat{\beta}, S^2$, are independent? Provide a rigorous argument.
d. What is the distribution of $(\hat{\beta} - \beta)^2 \sum (x_i - \bar{x})^2 / S^2$?

Solution

a. $\hat{\alpha} = \bar{y}$ and

$$\hat{\beta} = \sum (x_i - \bar{x})(y_i - \bar{y}) / \sum (x_k - \bar{x})^2$$

b. Use Cauchy-Schwarz: $E^2(S \times 1) \leq E(S^2)$
c. We have

$$\frac{\sum (y_i - \eta_i)^2}{\sigma^2} = \frac{n(\hat{\alpha} - \alpha)^2}{\sigma^2} + \frac{(\hat{\beta} - \beta)^2 \sum (x_i - \bar{x})^2}{\sigma^2} + \frac{\sum (y_i - \hat{y}_i)^2}{\sigma^2}$$

Thus by Cochran's Theorem, $\hat{\alpha}, \hat{\beta}, S^2$ are all independent.
d. From c we have, respectively,

$$\chi_n^2 = \chi_1^2 + \chi_1^2 + \chi_{n-2}^2$$

so that $(\hat{\beta} - \beta)^2 \sum (x_i - \bar{x})^2 / S^2 \sim F(1, n-2)$.

1.4 Variance Stabilization

We wish to find an increasing function f such that in the model

$$f(y_i) = \mathbf{x}_i'\hat{\boldsymbol{\beta}} + \epsilon_i$$

the ϵ_i have approximately the same variance.

Assume $\mu_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$, and for a known w,

$$Var(y_i) = w(\mu_i)$$

a. Argue that the ϵ_i will have approximately the same variance when

$$f(\mu) = \int \frac{d\mu}{w(\mu)^{1/2}}$$

b. Suppose the responses y_i are Poisson(μ_i), what is the form of f?
c. Suppose the responses y_i are Binomial(m, p_i), what is the form of f?

Solution

a. Clearly, by a Taylor approx.

$$Var[f(y)] \approx \left(\frac{df}{d\mu}\right)^2 w(\mu)$$

b. $w(\mu) = \mu$, $f(\mu) \equiv \mu^{1/2}$
c. $w(\mu) = \mu(1 - \mu/m)$, $f(\mu) \equiv \sin^{-1}[(\mu/m)^{1/2}]$

1.5 Nonstandard sampling problem

From a population $\mathcal{U}$ of size $N = 3$, a simple random sample $\mathcal{S}$ of size $n = 2$ is selected. The goal is to estimate the population total

$$t_{y\mathcal{U}} = y_1 + y_2 + y_3.$$

Prove that the estimator $\hat{t}(\mathcal{S})$ defined by

$$\hat{t}(\mathcal{S}) = \begin{cases} (3/2)y_1 + (3/2)y_2 & \text{if } \mathcal{S} = \{1,2\} \\ (3/2)y_1 + 2y_3 & \text{if } \mathcal{S} = \{1,3\} \\ (3/2)y_2 + y_3 & \text{if } \mathcal{S} = \{2,3\} \end{cases}$$

is unbiased. In addition, prove that Var $[\hat{t}(\mathcal{S})]$ is smaller than the the variance of the equally weighted estimator of the total $3\bar{y}_{\mathcal{S}}$ if $y_3(3y_2 - 3y_1 - y_2) > 0$.

Solution. Each of the possible simple random samples has probability $P(\mathcal{S}) = 1/3$. Therefore

$$E[\hat{t}] = (1/3)[(3/2)y_1 + (3/2)y_2 + (3/2)y_1 + 2y_3 + (3/2)y_2 + y_3] = y_1 + y_2 + y_3$$

so $\hat{t}$ is unbiased. Moreover

$$\begin{aligned} E[\hat{t}^2] &= (1/3)[((3/2)y_1 + (3/2)y_2)^2 + ((3/2)y_1 + 2y_3)^2 + ((3/2)y_2 + y_3)^2] \\ &= 3[y_1^2/2 + y_2^2/2 + 5y_3^2/9 + y_1y_2/2 + 2y_1y_3/3 + y_2y_3/3] \end{aligned}$$

so that

$$\text{Var}\,[\hat{t}(\mathcal{S})] = y_1^2/2 + y_2^2/2 + 2y_3^2/3 - y_1y_2/2 - y_2y_3$$

By contrast,

$$\begin{aligned} \text{Var}\,[3\bar{y}_{\mathcal{S}}] &= 9(1/2 - 1/3)S^2_{y\mathcal{U}} \\ &= (1/2)[y_1^2 + y_2^2 + y_3^2 - y_1y_2 - y_1y_3 - y_2y_3] \end{aligned}$$

The difference in variances is

$$\text{Var}\,[\hat{t}(\mathcal{S})] - \text{Var}\,[3\bar{y}_{\mathcal{S}}] = y_3^2/6 - y_1y_3/2 + y_2y_3/2 = -y_3(3y_2 - 3y_1 - y_3)/6.$$

1.6 Two stage sampling

A population $\mathcal{U}$ consists of N clusters, each of which contains M elements. Let y_{ij} denote the value of a variable y for the jth element in the ith cluster (or primary sampling unit). A simple random sample of n psu's is chosen, and from each psu a simple random sample of m elements is selected. Sampling from different psu's is performed independently. The statistic

$$\bar{y}_{\mathcal{S}} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij}$$

is used to estimate the population average

$$\bar{y}_{\mathcal{U}} = \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}.$$

(a) Write down the ANOVA table for the population (on an element basis), showing explicit formulas for the sums of squares between and within psu's.

(b) In terms of MSB and MSW, the mean squares in the population ANOVA table, prove that

$$\mathrm{Var}\left[\bar{y}_{S}\right] = \left(1 - \frac{n}{N}\right)\frac{\mathrm{MSB}}{nM} + \left(1 - \frac{m}{M}\right)\frac{\mathrm{MSW}}{mn}.$$

1.7 Two Way ANOVA

A researcher comes to you for advice about the analysis of an experiment she has conducted. She has used 5 treatments and she has 9 observations of some response variable Y for each treatment. She shows you the following ANOVA from a computer printout:

Source	d.f.	S.S.	F statistic	$P[> F]$
Treatments	4	100	8.33	0.0001
Error	40	120		

The researcher wants to know from you whether the analysis is correct, and if so what it means. For each of the three scenarios below answer the following questions:

(a) What is the appropriate model? Write a model equation, including any and all effects. Explain which effects are fixed and which are random, and state any distributional assumptions.

(b) Based on the model in (a), what is the corresponding ANOVA (include sources of variation, d.f., formulas for sums of squares, and the statistic for testing $H_0: \ \tau_1 = \tau_2 = \cdots = \tau_5$).

(c) Is the ANOVA from the printout above appropriate for testing $H_0: \ \tau_1 = \tau_2 = \cdots = \tau_5$?

Scenario I: 45 animals were used for the study and each treatment was applied to 9 animals selected at random.

Scenario II: 45 animals were used but they came from 9 different litters of size 5 each, and each treatment was assigned to one animal selected at random from each litter.

Scenario III: 15 animals were used, each treatment was assigned to 3 animals selected at random, and 3 observations were made on each animal.

Solution

Scenario I: This is a completely randomized design with $Y_{ij} = \tau_i + e_{ij}$, $i = 1, \ldots, 5$, $j = 1, \ldots, 9$ and the e_{ij} are i.i.d. $N(0, \sigma_e^2)$. The τ_i are fixed effects. The ANOVA table is as follows.

Source	d.f.	S.S.	F statistic
Treatments	4	$9\sum_{i=1}^{5}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	MST/MSE
Error	40	$\sum_{i=1}^{5}\sum_{j=1}^{9}(Y_{ij} - \bar{Y}_{i\cdot})^2$	

This is the model used by the researcher, so her ANOVA table and test statistic are correct.

Scenario II: This is a complete randomized block design with litters as blocks. The model is $Y_{ij} = \tau_i + b_j + e_{ij}$, $i = 1, \ldots, 5$, $j = 1, \ldots, 9$. The b_j are i.i.d. random litter effects with $N(0, \sigma_b^2)$ and the e_{ij} are i.i.d. $N(0, \sigma_e^2)$. The correct ANOVA table is

Source	d.f.	S.S.	F statistic
Treatments	4	$9\sum_{i=1}^{5}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	MST/MSE
Litters	8	$5\sum_{j=1}^{9}(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2$	
Error	32	$\sum_{i=1}^{5}\sum_{j=1}^{9}(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2$	

Note that if the litter effects are assumed to be fixed, the analysis is the same. In either case, the researcher's analysis is wrong.

Scenario III: This is a design with nested random effects. The model is $Y_{ijk} = \tau_i + a_{ij} + e_{ijk}$, $i = 1, \ldots, 5$, $j = 1, 2, 3$, $k = 1, 2, 3$. Here a_{ij} is the random effect of the jth animal in the ith treatment group and e_{ijk} is the measurement error on the kth observation of animal ij. It is assumed that the a_{ij} are i.i.d. $N(0, \sigma_a^2)$ and that the e_{ijk} are i.i.d. $N(0, \sigma_e^2)$. The correct ANOVA table is as follows.

Source	d.f.	S.S.	F statistic
Treatments	4	$9\sum_{i=1}^{5}(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot\cdot\cdot})^2$	MST/MSA
Animals	10	$5\sum_{j=1}^{3}(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot})^2$	
Error	30	$\sum_{i=1}^{5}\sum_{j=1}^{9}(Y_{ijk} - \bar{Y}_{ij\cdot})^2$	

Once again the researcher's analysis is wrong.

Note that in all three scenarios the researcher's SST is correct.