COMPREHENSIVE WRITTEN EXAMINATION, PAPER III

FRIDAY AUGUST 20, 2004, 9:00 A.M.–1:00 P.M.

STATISTICS 174 QUESTION

Answer all parts. Closed book, calculators allowed. It is important to show all working, especially with numerical calculations. Statistical tables are provided. You may freely quote results from the course notes or text without proof, but to the extent that it is feasible to do so, state precisely the result you are quoting.

Parts (a) through (f) are worth 15 points each, part (g) is worth 10 points; total 100.

Consider a linear model written in the form

$$Y = X_1\beta_1 + X_2\beta_2 + ... + X_k\beta_k + \epsilon \qquad (1)$$

where Y is $n \times 1$, X_i is $n \times p_i$ for $1 \le i \le k$, β_i is $p_i \times 1$, the total number of linear regression coefficients is $p = \sum_{i=1}^{k} p_i < n$, ϵ is an $n \times 1$ vector of errors that has a normal distribution with mean 0 and covariance matrix $\sigma^2 I_n$. We suppose also that $X_i^T X_i$ is nonsingular for each $i = 1, ..., k$, and

$$X_i^T X_j = 0 \text{ whenever } i \neq j. \qquad (2)$$

When (2) holds, the matrices $X_1, ..., X_k$ are said to be *mutually orthogonal.*

Suppose $\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_k \end{pmatrix}$ is the vector of least squares estimates under the full model (1).

(a) Show that $\widehat{\beta}_i = (X_i^T X_i)^{-1} X_i^T Y$, in other words, the estimate of β_i under the full model (1) is the same as under the model $Y = X_i\beta_i + \epsilon$.

(b) Define the residual vector $e = Y - \sum_{i=1}^{k} X_i \widehat{\beta}_i$. Show that

$$Y^T Y = \sum_{i=1}^{k} \widehat{\beta}_i^T X_i^T X_i \widehat{\beta}_i + e^T e. \qquad (3)$$

Explain how to interpret (3) as an "analysis of variance decomposition" of the total sum of squares $Y^T Y$ into k sums of squares due to regression (one corresponding to each of the submatrices $X_1, ..., X_k$) and the sum of squares due to error. What are the corresponding degrees of freedom?

(c) Suppose we want to test the hypothesis $H_0 : \ \beta_1 = ... = \beta_\ell = 0$ for some ℓ between 1 and k, against the alternative H_1 that not all of $\beta_1, ..., \beta_\ell$ are 0. (Note that by suitable reordering of $\beta_1, ..., \beta_k$, this may also be interpreted as a test that any given subset of $\beta_1, ..., \beta_k$ is 0.) Show that a suitable test statistic is

$$\frac{\sum_{i=1}^{\ell} \widehat{\beta}_i^T X_i^T X_i \widehat{\beta}_i}{\sum_{i=1}^{\ell} p_i} \cdot \frac{n-p}{e^T e}. \qquad (4)$$

What is the distribution of (4) when H_0 is true?

A recent paper[1] contained the following data matrix (among others):

Table 1

i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	Y_i
1	–1	–1	0	0	0.950
2	–1	0	–1	0	1.190
3	–1	0	0	–1	0.980
4	–1	0	0	1	1.060
5	–1	0	1	0	1.040
6	–1	1	0	0	1.210
7	0	–1	–1	0	1.120
8	0	–1	0	–1	1.030
9	0	–1	0	1	1.230
10	0	–1	1	0	1.100
11	0	0	–1	–1	1.250
12	0	0	–1	1	1.190
13	0	0	0	0	1.160
14	0	0	1	–1	0.960
15	0	0	1	1	1.400
16	0	1	–1	0	1.370
17	0	1	0	–1	1.530
18	0	1	0	1	1.870
19	0	1	1	0	1.220
20	1	–1	0	0	1.120
21	1	0	–1	0	1.200
22	1	0	0	–1	1.430
23	1	0	0	1	1.370
24	1	0	1	0	1.350
25	1	1	0	0	1.370

This table given the result of an experiment in which $X_{i1}, ..., X_{i4}$ were the weights of different ingredients of an infant food formula (standardized to the values –1, 0, 1) and Y_i was the measured concentration of phytic acid (considered an undesirable constituent).

The full model assumed in the paper was

$$Y_i \quad = \quad \beta_0 + \sum_{j=1}^{4} \beta_j X_{ij} + \sum_{j=1}^{4} \beta_{jj} X_{ij}^2 + \sum_{j=1}^{3} \sum_{k=j+1}^{4} \beta_{jk} X_{ij} X_{ik} + \epsilon_i, \tag{5}$$

where ϵ_i is a random error, assumed to be independent normal with mean 0 and common variance σ^2. We interpret the different contributions β_0, $\sum_{j=1}^{4} \beta_j X_{ij}$, $\sum_{j=1}^{4} \beta_{jj} X_{ij}^2$ and $\sum_{j=1}^{3} \sum_{k=j+1}^{4} \beta_{jk} X_{ij} X_{ik}$ as the "intercept", "linear" "quadratic" and "cross-product" terms respectively.

Suppose we modify (5) to

$$Y_i \quad = \quad \beta_0 + \sum_{j=1}^{4} \beta_j X_{ij} + \sum_{j=1}^{4} \beta_{jj} (X_{ij}^2 - c) + \sum_{j=1}^{3} \sum_{k=j+1}^{4} \beta_{jk} X_{ij} X_{ik} + \epsilon_i, \tag{6}$$

in other words, we replace each of the covariates X_{ij}^2 by $X_{ij}^2 - c$ for some constant c, all other terms remaining the same. We still refer to $\sum_{j=1}^{4} \beta_{jj} (X_{ij}^2 - c)$ as the "quadratic" term in (6).

[1] B. Martínez, F. Rincón, M.V. Ibáñez, P. Abellán, Improving the nutritive value of homogenized infant foods using response surface methodology, *Journal of Food Science* **69**, SNQ38–SNQ43 (2004)

(d) For what value of c do the "intercept", "linear", "quadratic" and "cross-product" parts of the model become mutually orthogonal, in the sense defined after equation (2)?

Under the full model (6), the set of parameter estimates, standard errors and t values is

Table 2

Parameter	Estimate	S.E.	t value
β_0	1.22800	0.02988	41.09
β_1	0.11750	0.04314	2.72
β_2	0.16833	0.04314	3.90
β_3	–0.02083	0.04314	–0.48
β_4	0.07833	0.04314	1.82
β_{11}	–0.02708	0.08892	–0.30
β_{22}	0.07917	0.08892	0.89
β_{33}	–0.01208	0.08892	–0.14
β_{44}	0.10167	0.08892	1.14
β_{12}	–0.00250	0.07471	–0.03
β_{13}	0.07500	0.07471	1.00
β_{14}	–0.03500	0.07471	–0.47
β_{23}	–0.03250	0.07471	–0.44
β_{24}	0.03500	0.07471	0.47
β_{34}	0.12500	0.07471	1.67

The analysis of variance table corresponding to part (b) above is

Table 3

Component	Sum of squares	Degrees of freedom	Mean square
Intercept	37.6996		
Linear	0.58455		
Quadratic	0.10454		
Cross-product	0.09905		
Residual	0.22326		
Total $\sum Y_i^2$	38.711	25	

(e) Complete the "degrees of freedom" and "mean square" entries of the above table and hence test whether each of the intercept, linear, quadratic and cross-product components of the model is 0.

(f) Suppose we fit the model containing just the intercept and linear component, assuming the quadratic and cross-product components are 0. By the result of (a) above, the values of $\widehat{\beta}_1, ..., \widehat{\beta}_4$ will still be as given by Table 2, but the standard errors will be different. What are the standard errors in this case? Hence determine which of the parameters $\beta_1, ..., \beta_4$ is significant.

(g) What are your overall conclusions for the food company?

SOLUTIONS

(a) Let $S_i = X_i^T X_i$. Then the full $X^T X$ matrix and its inverse are given by

$$X^T X = \begin{pmatrix} S_1 & 0 & \dots & 0 \\ 0 & S_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & S_k \end{pmatrix}, \qquad (X^T X)^{-1} = \begin{pmatrix} S_1^{-1} & 0 & \dots & 0 \\ 0 & S_2^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & S_k^{-1} \end{pmatrix},$$

using (2). Hence the full linear model estimates are

$$(X^T X)^{-1} X^T Y = \begin{pmatrix} S_1^{-1} & 0 & \dots & 0 \\ 0 & S_2^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & S_k^{-1} \end{pmatrix} \begin{pmatrix} X_1^T Y \\ X_2^T Y \\ \vdots \\ X_k^T Y \end{pmatrix} = \begin{pmatrix} S_1^{-1} X_1^T Y \\ S_2^{-1} X_2^T Y \\ \vdots \\ S_k^{-1} X_k^T Y \end{pmatrix},$$

which is the required result.

(b) We have

$$\begin{aligned} Y^T Y &= (X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + ... + X_k\widehat{\beta}_k + e)^T (X_1\widehat{\beta}_1 + X_2\widehat{\beta}_2 + ... + X_k\widehat{\beta}_k + e) \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} \widehat{\beta}_i X_i^T X_j \widehat{\beta}_j + 2\sum_{i=1}^{k} \widehat{\beta}_i X_i^T e + e^T e \\ &= \sum_{i=1}^{k} \widehat{\beta}_i X_i^T X_i \widehat{\beta}_i + e^T e \end{aligned}$$

where we have used (2) and also

$$\widehat{\beta}_i X_i^T e \;=\; \widehat{\beta}_i X_i^T (Y - \sum_{j=1}^{k} X_j \widehat{\beta}_j) \;=\; \widehat{\beta}_i X_i^T (Y - X_i \widehat{\beta}_i) \;=\; 0$$

which follows from the normal equations for $\widehat{\beta}_i$. This proves (3). The corresponding "analysis of variance table" (including the degrees of freedom) is

Component	Sum of squares	Degrees of freedom	Mean square
Regression on X_1	$\widehat{\beta}_1^T X_1^T X_1 \widehat{\beta}_1$	p_1	$\frac{\widehat{\beta}_1^T X_1^T X_1 \widehat{\beta}_1}{p_1}$
Regression on X_2	$\widehat{\beta}_2^T X_2^T X_2 \widehat{\beta}_2$	p_2	$\frac{\widehat{\beta}_2^T X_2^T X_2 \widehat{\beta}_2}{p_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Regression on X_k	$\widehat{\beta}_k^T X_k^T X_k \widehat{\beta}_k$	p_k	$\frac{\widehat{\beta}_k^T X_k^T X_k \widehat{\beta}_k}{p_k}$
Residual	$e^T e$	$n - p$	$\frac{e^T e}{n-p}$
Total	$Y^T Y$	n	

(c) By Theorem 3.3 from the course notes, the F test is of the form

$$\frac{(SSE_0 - SSE_1)/q}{SSE_1/(n-p)} \sim F_{q,n-p} \text{ under } H_0.$$

Here

$$\begin{aligned} SSE_0 &= Y^T Y - \sum_{i=\ell+1}^{k} \widehat{\beta}_i^T X_i^T X_i \widehat{\beta}_i, \\ SSE_1 &= Y^T Y - \sum_{i=1}^{k} \widehat{\beta}_i^T X_i^T X_i \widehat{\beta}_i \;=\; e^T e, \\ SSE_0 - SSE_1 &= \sum_{i=1}^{\ell} \widehat{\beta}_i^T X_i^T X_i \widehat{\beta}_i. \end{aligned}$$

We also have $q = \sum_{i=1}^{k} p_i$ (the difference in the number of parameters in the models H_0 and H_1). It follows that the F statistic is of the form (4), and its distribution under H_0 is $F_{\sum_{i=1}^{\ell} p_i, n-p}$.

(d) The different model components are already orthogonal except that $\sum_i X_{ij}^2 = 12$ for each $j = 1, 2, 3, 4$; in other words, the quadratic term is not orthogonal with the intercept. However we do have $\sum_i \left(X_{ij}^2 - \frac{12}{25} \right) = 0$; therefore, defining $c = \frac{12}{25}$ makes the problem orthogonal.

(e) Degrees of freedom are 1, 4, 4, 6, 10 and the mean square components are 37.6996, 0.14614, 0.02614, .01651, .02233. Dividing each of the first four number by the last one, we get F ratios 1688, 6.54, 1.17, 0.74. Comparing with the 5% critical values of 4.96, 3.48, 3.48, 3.22, we conclude that the intercept and linear components are significant, the quadratic and cross-product components are not.

(f) If we use the model with just linear and intercept components instead of the full model, the value of s (estimated standard deviation of residuals) changes from $\sqrt{\frac{.22326}{10}} = .14942$ to $\sqrt{\frac{.22326+.09905+.10454}{20}} = .14609$. All the standard errors are proportional to s, so the standard error of $\widehat{\beta}_1, ..., \widehat{\beta}_4$ changes from .04314 to $.04314 \times \frac{.14609}{.14942} = .04218$. The t statistics do not change much; for instance, the t statistic for $\widehat{\beta}_4$ is now $\frac{.07833}{.04218} = 1.86$, which is less than the critical value (for t_{20}) of 2.086. We conclude that β_1 and β_2 are significant, β_3 and β_4 are not.

(g) The X_1 and X_2 components of the food mix should be kept low to ensure that phytic acid is not too high. The other components do not appear to have a significant effect.