COMPREHENSIVE WRITTEN EXAMINATION, PAPER III

FRIDAY AUGUST 17, 2001, 9:00 A.M.

STATISTICS 174 QUESTION

SECTION I (70% of credit)

A chemical experiment is performed in which the relationship between the concentration of a reactant x_i and the rate of reaction y_i is given by the formula

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \quad 1 \le i \le n, \tag{1}$$

in which $\{\epsilon_i\}$ are independent $N[0, \sigma^2]$ errors. Assume $\beta_2 < 0$ so that the function $y = \beta_0 + \beta_1 x + \beta_2 x^2$ has a unique maximum at $x = -\beta_1/(2\beta_2)$.

Assume the experiment is normalized so that $\sum x_i = \sum x_i^3 = 0$, $\sum x_i^2 = n$, $\sum x_i^4 = Cn$ for some $C > 1$.

1. Suppose the model (1) is fitted by ordinary least squares, producing estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$. Give explicit algebraic expressions for the estimators, $\widehat{\beta}_0$, $\widehat{\beta}_1$, $\widehat{\beta}_2$, and derive the variance-covariance matrix of these estimators as a function of σ^2.
2. Defining $\theta = -\beta_1/(2\beta_2)$, $\widehat{\theta} = -\widehat{\beta}_1/(2\widehat{\beta}_2)$, give an approximate expression for the variance of $\widehat{\theta}$, using the delta method.
3. Treating the approximation you derived in part 2 as exact, and writing s^2 as the usual unbiased estimator of σ^2 (you are not asked to write down an explicit algebraic expression for this), show how to derive an approximate $100(1-\alpha)\%$ confidence interval for θ, for given $\alpha \in (0,1)$.
4. A physical theory suggests $\theta = \frac{1}{2}$. By rewriting the model (1) in the form
$$y_i = \gamma_0 + \gamma_1(x_i - x_i^2) + \gamma_2 x_i + \epsilon_i, \tag{2}$$
show how the hypothesis $H_0 : \ \theta = \frac{1}{2}$ may be rewritten as a hypothesis about $(\gamma_0, \gamma_1, \gamma_2)$, and hence derive an *exact* test of H_0 against the alternative $H_1 : \ \theta \neq \frac{1}{2}$.

   Hint: You may find the following matrix identity useful. The inverse of the 3×3 matrix
$$\begin{pmatrix} 1 & -1 & 0 \\ -1 & x & 1 \\ 0 & 1 & 1 \end{pmatrix},$$
where $x \neq 2$, is
$$\frac{1}{x-2}\begin{pmatrix} x-1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & x-1 \end{pmatrix}.$$

5. Suppose now there are two regressions (corresponding to different experiments, e.g. two different chemicals) of the form

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \quad 1 \le i \le n, \\ y_i &= \delta_0 + \delta_1 x_{i-n} + \delta_2 x_{i-n}^2 + \epsilon_i, \quad n+1 \le i \le 2n, \end{aligned}$$

where $x_1, ..., x_n$ satisfy the same assumptions as before, and both β_2 and δ_2 are negative. In this case, the null hypothesis is that the maxima of the two curves $y = \beta_0 + \beta_1 x + \beta_2 x^2$, $y = \delta_0 + \delta_1 x + \delta_2 x^2$, occur for the same x. Is it possible to write this as a linear hypothesis which may be tested exactly (as in part 4) or is it necessary to use an approximate method (as in part 3)? In either case, give an outline of the proposed method of analysis (full algebraic details are not required for this part).

SECTION II (30% of credit)

Tables 1 and 2 (later) show measurements of four variables for 48 samples of rock (data taken from the book by Venables and Ripley). The variables represent the area, perimeter, shape and permeability; the intention is to be able to predict permeability from measurements of the other three variables. A regression analysis is considered in which area ($\times 10^{-3}$), perimeter ($\times 10^{-3}$) and shape are considered the three covariates denoted x_1, x_2, x_3 respectively, and the logarithm of permeability is the response variable. For various combinations of x_1, x_2, x_3, the model fits are represented by Table 3, assuming the standard linear model assumptions. The estimated residual standard deviation and associated degrees of freedom, for each of eight models, are shown in Table 3.

1. Based on the given table of residual standard errors, and making the standard linear model assumptions, describe which model (i.e. which combination of x_1, x_2 and x_3) you would select for these data. Be sure to indicate your rationale for this selection.

2. A plot of residuals *versus* original y values (i.e. the logarithms of permeability) is shown in Figure 1. Based on this plot, would you highlight any particular feature as indicating that the model is not fitting the stated assumptions in this instance?

3. Suggest an explanation for whatever you observed in part 2, and if you can, a possible alternative method of analysis. You are allowed to speculate about the motivations for conducting the experiment in the particular way that it appears to have been done.

Case	Area	Perimeter	Shape	Permeability
1	4990	2792	0.09	6.3
2	7002	3893	0.15	6.3
3	7558	3931	0.18	6.3
4	7352	3869	0.12	6.3
5	7943	3949	0.12	17.1
6	7979	4010	0.17	17.1
7	9333	4346	0.19	17.1
8	8209	4345	0.16	17.1
9	8393	3682	0.20	119.0
10	6425	3099	0.16	119.0
11	9364	4480	0.15	119.0
12	8624	3986	0.15	119.0
13	10651	4037	0.23	82.4
14	8868	3518	0.23	82.4
15	9417	3999	0.17	82.4
16	8874	3629	0.15	82.4
17	10962	4609	0.20	58.6
18	10743	4788	0.26	58.6
19	11878	4864	0.20	58.6
20	9867	4479	0.14	58.6
21	7838	3429	0.11	142.0
22	11876	4353	0.29	142.0
23	12212	4698	0.24	142.0
24	8233	3518	0.16	142.0
25	6360	1977	0.28	740.0
26	4193	1379	0.18	740.0
27	7416	1916	0.19	740.0
28	5246	1585	0.13	740.0
29	6509	1851	0.23	890.0
30	4895	1240	0.34	890.0
31	6775	1728	0.31	890.0
32	7894	1461	0.28	890.0
33	5980	1427	0.20	950.0
34	5318	991	0.33	950.0
35	7392	1351	0.15	950.0
36	7894	1461	0.28	950.0
37	3469	1377	0.18	100.0
38	1468	476	0.44	100.0
39	3524	1189	0.16	100.0
40	5267	1645	0.25	100.0

Table 1: Data for part II, cases 1–40.

Case	Area	Perimeter	Shape	Permeability
41	5048	942	0.33	1300.0
42	1016	309	0.23	1300.0
43	5605	1146	0.46	1300.0
44	8793	2280	0.42	1300.0
45	3475	1174	0.20	580.0
46	1651	598	0.26	580.0
47	5514	1456	0.18	580.0
48	9718	1486	0.20	580.0

Table 2: Data for part II, cases 41–48.

Variables Included	Residual SE	d.f.	Variables Included	Residual SE	d.f.
None	1.643376	47	$x_1 + x_2$	0.852043	45
x_1	1.574854	46	$x_1 + x_3$	1.381568	45
x_2	1.157668	46	$x_2 + x_3$	1.103856	45
x_3	1.416901	46	$x_1 + x_2 + x_3$	0.852752	44

Table 3: Results of various model fits.

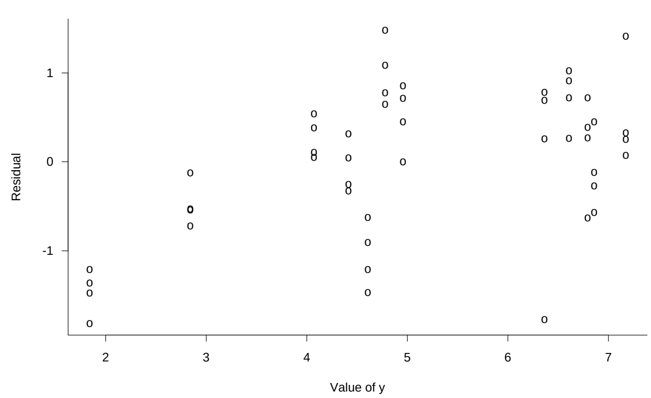

Figure 1. Residuals *vs.* original y values for model fit with all of x_1, x_2, x_3.

SOLUTION

SECTION I

1. The matrices X^TX and $(X^TX)^{-1}$ are given by

$$X^TX = n \left(\begin{array}{ccc} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & C \end{array} \right), \quad (X^TX)^{-1} = \frac{1}{n} \left(\begin{array}{ccc} \frac{C}{C-1} & 0 & -\frac{1}{C-1} \\ 0 & 1 & 0 \\ -\frac{1}{C-1} & 0 & \frac{1}{C-1} \end{array} \right).$$

Hence the point estimates are

$$\begin{aligned} \widehat{\beta}_0 &= \frac{1}{n(C-1)} \sum y_i(C - x_i^2), \\ \widehat{\beta}_1 &= \frac{1}{n} \sum y_i x_i, \\ \widehat{\beta}_2 &= \frac{1}{n(C-1)} \sum y_i(x_i^2 - 1), \end{aligned}$$

and the variance-covariance matrix is given by $(X^TX)^{-1}\sigma^2$.

2. Define $f(\beta_1, \beta_2) = -\beta_1/(2\beta_2)$ with partial derivatives $f_1 = \partial f/\partial\beta_1 = -1/(2\beta_2)$, $f_2 = \partial f/\partial\beta_2 = \beta_1/(2\beta_2^2)$. By the delta method, the variance of $f(\widehat{\beta}_1, \widehat{\beta}_2)$ is given approximately by

$$f_1^2 \mathrm{Var}(\widehat{\beta}_1) + f_2^2 \mathrm{Var}(\widehat{\beta}_2) + 2 f_1 f_2 \mathrm{Cov}(\widehat{\beta}_1, \widehat{\beta}_2),$$

however, the third term is 0 and the remaining two evaluate to

$$\frac{1}{4\beta_2^2} \cdot \frac{\sigma^2}{n} + \frac{\beta_1^2}{4\beta_2^4} \cdot \frac{\sigma^2}{n(C-1)}.$$

3. Assuming s^2 is the usual unbiased estimate of σ^2 with $n-3$ d.f., we define the standard error

$$S.E. = \sqrt{\frac{1}{4\widehat{\beta}_2^2} \cdot \frac{s^2}{n} + \frac{\widehat{\beta}_1^2}{4\widehat{\beta}_2^4} \cdot \frac{s^2}{n(C-1)}},$$

and the desired approximate confidence interval is of the form

$$\widehat{\theta} \pm t_{n-3;1-\alpha/2} \cdot S.E.$$

or any equivalent form.

4. The null hypothesis corresponds to $\gamma_2 = 0$ in the rewritten form. The $X^T X$ matrix for this problem becomes

$$X^T X = n \begin{pmatrix} 1 & -1 & 0 \\ -1 & C+1 & 1 \\ 0 & 1 & 1 \end{pmatrix},$$

so applying the hint with $x = C + 1$,

$$(X^T X)^{-1} = \frac{1}{n} \cdot \frac{1}{C-1} \begin{pmatrix} C & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & C \end{pmatrix}.$$

Then

$$\widehat{\gamma}_2 = \frac{1}{n(C-1)} \sum y_i \{x_i^2 + (C-1)x_i - 1\},$$

and its variance is $C\sigma^2/((C-1)n)$. Hence a standard t-test would reject H_0 at level α if

$$|\widehat{\gamma}_2| > s\sqrt{\frac{C}{(C-1)n}} t_{n-3;1-\alpha/2}.$$

5. The null hypothesis corresponds to $\beta_1/\beta_2 = \gamma_1/\gamma_2$ and there are numerous ways of writing this hypothesis in different ways as functions of the parameters; unfortunately, none of them appears to reduce to a case in which an exact test can be constructed. Therefore, we use the delta method, one version of which is to test whether $\theta = 0$, where $\theta = \beta_1\gamma_2 - \beta_2\gamma_1$. The two halves of the experiment (first n and last n observations) are entirely independent and the estimates $\widehat{\beta}_1$ etc., and their standard errors, may be derived as before (with all four estimates mutually independent). A test of $H_0 : \theta = 0$ may be derived based on $\widehat{\theta} = \widehat{\beta}_1\widehat{\gamma}_2 - \widehat{\beta}_2\widehat{\gamma}_1$, with standard error

$$SE = \sqrt{\frac{s^2}{n(C-1)}\{(\widehat{\beta}_1^2 + \widehat{\gamma}_1^2) + C^2(\widehat{\beta}_2^2 + \widehat{\gamma}_2^2)\}},$$

s being estimated from the two samples combined (we are here assuming that the variance is common to both samples). Noting that s has $2n - 6$ d.f., the final (approximate) test is to reject H_0 at level α if

$$|\theta| > SE \cdot t_{2n-6;1-\alpha/2}.$$

There are numerous possible alternative solutions based on different ways of writing the null hypothesis; any such solution will be accepted.

SECTION II

1. The only relevant comparison is between the models $\{x_1, x_2\}$ and $\{x_1, x_2, x_3\}$ (all considerations involving either x_1 or x_2 lead to decisive evidence that both variables should be included in the model).

   For x_1, x_2 alone, one finds the residual sum of squares is $45 \times .854043^2 = 32.669$ with 45 d.f., while for $x_1 + x_2 + x_3$ it is $44 \times .852752^2 = 31.996$ with 44 d.f. The F statistic is

   $$\frac{32.669 - 31.996}{31.996} \times \frac{44}{1} = 0.925,$$

   with 44 and 1 degrees of freedom, and since $F < 1$, we conclude that x_3 is not significant. Therefore, the optimal model, under this analysis and with these assumptions, is that the best model includes x_1 and x_2 but does not include x_3.

2. We observe that y takes on only 12 distinct values, each value replicated four times, and the residuals appear grouped within each of the 12 clusters. Therefore, it appears that the assumption of independent errors is violated: there is a grouping (also interpretable as a correlation) within each of the 12 subgroups.

3. It seems likely that the data were collected from just 12 distinct samples of rock but that the rock samples were cut up in different ways to construct various samples of different dimensions. As for the analysis, there is no clear-cut answer to this but some possibilities include: (i) include a random (or non-random) effect in the model for each of the 12 subgroups, (ii) average over the four observations in each subgroup and just treat as 12 independent observations (the simplest solution, but suffers from the disadvantage that the resulting regression is based on averages over groups of four samples rather than single vectors of (x_1, x_2, x_3)), (iii) re-analyze the data as a calibration experiment (any further detail provided about this possibility will earn additional credit).