COMPREHENSIVE WRITTEN EXAMINATION, PAPER III
FRIDAY AUGUST 22, 2003, 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. Some familiarity with the t and F distributions is assumed, but statistical tables are not required.

Tentative mark scheme: parts (a) and (f) are worth 10 points; (c) is worth 20 points; (b), (d), (e), (g), 15 points each, for a total of 100 points. Extra points may be given for meritorious work at the examiner's discretion.

(a) Define the *hat matrix* H and the *leverage* h_i associated with the ith observation in a linear regression. Give the algebraic formulae for H and h_i, and state *two* statistical interpretations of h_i.

(b) Consider a linear model with two covariates x_{i1}, x_{i2} and nine observations, arranged as follows $((x_{i1}, x_{i2}, y_i)$ coordinates given below each point — the y_i values are used later on):

x (−1,1,0)	x (0,1,1)	x (1,1,1)
x (−1,0,0)	x (0,0,−3)	x (1,0,1)
x (−1,−1,1)	x (0,−1,0)	x (1,−1,0)

The model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \tag{1}$$

with the usual linear model assumptions on the errors ϵ_i.

Calculate the leverages $h_i,\ i = 1, ..., 9$ associated with each of the nine observations. What is $\sum_{i=1}^{9} h_i$?

(c) Now suppose we have a sample of y_i values as shown in the diagram. Calculate directly, (i) the least squares estimates $\widehat{\beta}_j,\ j = 0, 1, 2$; (ii) the residual mean squared error s^2; (iii) the standard errors of the three parameter estimates. Which of the parameter estimates are significantly different from 0?

(d) It's possible that the central observation (–3) is an outlier. What are (i) the unstandardized residual e_i, (ii) the internally standardized residual e_i^*, (iii) the externally studentized residual d_i^*, for this observation? What are your conclusions about whether this observation is indeed an outlier?

(e) A further table of deletion diagnostics, derived from SAS output, is as follows.

Observation	CovRatio	DFFITS	DFBETAS (β_0)	DFBETAS (β_1)	DFBETAS (β_2)
1	1.5095	1.0441	0.5220	–0.6394	–0.6394
2	2.3902	0.0256	0.0162	–0.0198	0.0000
3	3.0939	–0.0842	–0.0421	0.0516	–0.0516
4	2.3902	0.0256	0.0162	0.0000	–0.0198
5	0.0045	–2.0207	–2.0207	0.0000	0.0000
6	2.0030	0.3426	0.2167	0.0000	0.2654
7	3.0939	–0.0842	–0.0421	–0.0516	0.0516
8	2.0030	0.3426	0.2167	0.2654	0.0000
9	2.7155	0.4303	0.2152	0.2635	0.2635

Based on this table, comment further on whether any of the observations are influential values. Note that observation 5 is the one in the center on the preceding diagram.

(f) Another possible explanation for the data is that we should be fitting a quadratic, instead of a linear, model. Suppose equation (1) is modified to

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \epsilon_i. \tag{2}$$

This model was run in SAS and produced a residual sum of squares (RSS) of 7.1111. Based on this, would you say that the quadratic model (2) is a significant improvement on the linear model (1)? (Use an F test.)

(g) Let us return to the situation of part (b). Suppose, instead of the given configuration, there are d non-constant covariates with the model

$$y_i = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} + \epsilon_i, \tag{3}$$

and there are $n = 2^d + 1$ data points arranged as follows: one observation for which $x_{i1} = ... = x_{id} = 0$ (the "center"), and one in each configuration for which $x_{ij} = \pm 1$ for each $j = 1, ..., d$ (the 2^d "corners").

Show that, in this configuration, any of the 2^d corner points has exactly $1 + d + d2^{-d}$ times the leverage of the center point.

SOLUTIONS

(a) Assuming the model $Y = X\beta + \epsilon$, with σ^2 the common variance of the error ϵ_i, we define $H = X(X^TX)^{-1}X^T$, h_i by the ith diagonal entry of H. Among the possible statistical interpretations are (i) $\sigma^2 h_i$ is the variance of the ith fitted value $\widehat{y}_i$, (ii) $\sigma^2(1-h_i)$ is the variance of the ith residual e_i.

(b) We have

$$X = \begin{pmatrix} 1 & -1 & -1 \\ 1 & -1 & 0 \\ 1 & -1 & 1 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{pmatrix}, \quad X^TX = \begin{pmatrix} 9 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 6 \end{pmatrix}, \quad (X^TX)^{-1} = \begin{pmatrix} \frac{1}{9} & 0 & 0 \\ 0 & \frac{1}{6} & 0 \\ 0 & 0 & \frac{1}{6} \end{pmatrix}.$$

The leverage associated with the ith row of the X matrix, $\begin{pmatrix} 1 & x_{i1} & x_{i2} \end{pmatrix}$ say, is

$$\begin{pmatrix} 1 & x_{i1} & x_{i2} \end{pmatrix}(X^TX)^{-1}\begin{pmatrix} 1 \\ x_{i1} \\ x_{i2} \end{pmatrix} = \frac{1}{9} + \frac{x_{i1}^2}{6} + \frac{x_{i2}^2}{6}.$$

This comes to $\frac{4}{9}$ for the four corner points, $\frac{5}{18}$ for the points of form $(0, \pm 1)$ or $(\pm 1, 0)$, and $\frac{1}{9}$ for the middle point (note that $\sum_i h_i = 3$, as it should).

(c) (i) $\sum y_i = \sum y_i x_{i1} = \sum y_i x_{i2} = 1$ so $\widehat{\beta}_0 = \frac{1}{9}$, $\widehat{\beta}_1 = \frac{1}{6}$, $\widehat{\beta}_2 = \frac{1}{6}$. (ii) The residual sum of squares (RSS) is $(y - X\widehat{\beta})^T(y - X\widehat{\beta}) = y^Ty - \widehat{\beta}^TX^Ty = \sum y_i^2 - \begin{pmatrix} \frac{1}{9} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 13 - \left(\frac{1}{9} + \frac{1}{6} + \frac{1}{6}\right) = \frac{113}{9}$. Hence $s^2 = \frac{113}{54} = 2.0926$, $s = 1.4467$. (iii) The three standard errors are $\frac{s}{\sqrt{9}}$, $\frac{s}{\sqrt{6}}$, $\frac{s}{\sqrt{6}}$, or numerically, .4822, .5906, .5906. According to standard t statistics, none of the three parameter estimates $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$ is significantly different from 0.

(d) For the observation at (0,0), we have $y_i = -3$, $\widehat{y}_i = \frac{1}{9}$ and hence $e_i = y_i - \widehat{y}_i = -3.1111$. The formulae for the internally standardized and externally studentized residuals are

$$\begin{aligned} e_i^* &= \frac{e_i}{s\sqrt{1-h_i}} = -\frac{3.1111}{1.4467 \times \sqrt{\frac{8}{9}}} = -2.281, \\ d_i^* &= e_i\sqrt{\frac{n-p-1}{(1-h_i)(n-p)s^2 - e_i^2}} = -3.1111\sqrt{\frac{5}{\frac{8}{9} \times 6 \times 2.0926 - 3.1111^2}} = -5.715. \end{aligned}$$

Based on either of these, but especially the externally studentized residual, it does appear that this observation is an outlier.

(e) According to the standard criteria, the critical value for DFFITS is $2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{3}{9}} = 1.155$, and the critical value for DFBETAS is $\frac{2}{\sqrt{n}} = 0.667$ (or 1 since n is “small” in this instance).

COVRATIO is considered critical if $|\text{COVRATIO} - 1| > \frac{3p}{n} = 1$, i.e. if COVRATIO$< 0$ or > 2. By these criteria, observation 5 is influential both by DFFITS and for DFBETAS with β_0, i.e. the central observation has a big influence on the intercept but not on the slopes (since y_5 is not included in the sums that define $\widehat{\beta}_1$ and $\widehat{\beta}_2$, the influence is exactly zero in this case, as reflected by the table). For COVRATIO, it appears that all observations except 1 and 5 are critical. In this case a more plausible explanation is that the deletion-based residual standard deviation $s_{(i)}$ is very much reduced (compared with s) for $i = 5$, as reflected by the very small COVRATIO, but there is a compensating increase when any other observation is deleted; in other words, it still looks as though observation 5 is the one that is truly influential on the estimated residual variance.

(f) Under the linear model, the RSS is $\frac{113}{9} = 12.556$ (from **(c)**) with 6 d.f. Under the quadratic model, the RSS is 7.111 with 4 d.f. The F statistic is

$$\frac{12.556 - 7.111}{2} \cdot \frac{4}{7.111} = 1.53$$

which is not significant as a $F_{2,4}$ random variable (the p value is about 0.32 though you are not required to specify the exact value).

(g) The first row of X is $\begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}$ (1 followed by d zeros) and the remaining 2^d rows are all of the form $\begin{pmatrix} 1 & \pm 1 & \dots & \pm 1 \end{pmatrix}$. We have

$$X^T X = \begin{pmatrix} 2^d + 1 & 0 & \dots & 0 \\ 0 & 2^d & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & 2^d \end{pmatrix}, \quad (X^T X)^{-1} = \begin{pmatrix} \frac{1}{2^d+1} & 0 & \dots & 0 \\ 0 & \frac{1}{2^d} & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \dots & \frac{1}{2^d} \end{pmatrix}.$$

Then h_1 is of the form

$$\begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix} (X^T X)^{-1} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \frac{1}{2^d + 1},$$

and the remaining h_i are of the form

$$\begin{pmatrix} 1 & \pm 1 & \dots & \pm 1 \end{pmatrix} (X^T X)^{-1} \begin{pmatrix} 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{pmatrix} = \frac{1}{2^d + 1} + \frac{d}{2^d}.$$

The ratio of the last two expressions is $1 + d\frac{2^d+1}{2^d}$, as required.