STATISTICS 174: APPLIED STATISTICS

MIDTERM EXAM

OCTOBER 15, 2003

Time allowed: 75 minutes.

This is an open book exam: all course notes and the text are allowed, and you are expected to use your own calculator. Answers should preferably be written in a blue book.

The exam is expected to be your own work and no consultation during the exam is allowed. You are allowed to ask the teaching assistant for clarification if you feel the question is ambiguous.

Each question is worth 20 points. All answers will be graded, but the maximum total score is 100. Thus, you can obtain full marks by answering five of the six questions. Show all numerical and algebraic calculations.

Statistical tables are not provided: you do not need to know precise values for any distributions to be able to answer the following questions.

Smoothing is the generic name for a class of statistical procedures that aim to fit smooth curves through data. One example of a smoothing problem is to take a simple scatterplot of observations $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ and to construct a smooth curve $y = f(x)$ that passes approximately, but not exactly, through the n points of the scatterplot.

The purpose of this question is to develop detailed formulas and properties of a specific solution this problem: namely, to estimate the smooth curve for a given value of x, we take the five nearest data points to x and fit a quadratic regression through them. The problem is simplified by assuming that the data values of x are equally spaced.

We therefore consider the problem of estimating a smooth function $f(x)$, by quadratic regression based on five data points at $x-2h,\ x-h,\ x,\ x+h$ and $x+2h$, where h is the spacing between data points. Since the structure of the problem is invariant to location and scale changes, there is no loss of generality in assuming $x = 0$ and $h = 1$. This leads therefore to the following formulation of the problem:

Given x values $x_j = j$ for $j = -2,\ -1,\ 0,\ 1$ and 2, and corresponding y_j values written $y_{-2},\ y_{-1},\ y_0,\ y_1$ and y_2, find the optimal estimator of β_0 obtained through the regression

$$y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \epsilon_j,\ (j = -2, -1, 0, 1, 2), \tag{1}$$

where, as is usual in linear regression, the $\{\epsilon_j\}$ are assumed independent normally distributed random errors with mean 0 and variance σ^2.

1. Suppose first we assume $\beta_2 = 0$, i.e. we fit a linear regression instead of a quadratic regression. By writing the problem in the familiar $y = X\beta + \epsilon$ form, show how to derive the least squares estimates, and show in particular that
$$\tilde{\beta}_0 = \frac{y_{-2} + y_{-1} + y_0 + y_1 + y_2}{5}, \tag{2}$$
and state its variance. (We use $\tilde{\beta}_0$ rather than $\hat{\beta}_0$, to distinguish this from the alternative estimator of question 2.)

2. Now suppose we fit the full model (1) without assuming $\beta_2 = 0$. Derive the least squares estimators in this case, show that
$$\hat{\beta}_0 = \frac{-6y_{-2} + 24y_{-1} + 34y_0 + 24y_1 - 6y_2}{70}. \tag{3}$$
What is its variance?

3. A common way to compare two estimators is in terms of their *mean squared error*, which may be defined as $B^2 + V$, where B is the bias and V is the variance. The variances of $\tilde{\beta}_0$ and $\hat{\beta}_0$ were derived in questions 1 and 2; the bias of $\hat{\beta}_0$ is 0. If $\beta_2 \neq 0$, what is the bias of $\tilde{\beta}_0$? Hence show the mean squared error of $\tilde{\beta}_0$ is smaller than that of $\hat{\beta}_0$ if and only if $14\beta_2^2 < \sigma^2$.

4. In practice, if we wanted to decide which of $\tilde{\beta}_0$ or $\hat{\beta}_0$ to use, we would probably conduct a hypothesis test of $H_0 : \beta_2 = 0$ against the alternative $H_1 : \beta_2 \neq 0$. Assuming the full model (1), calculate the least squares estimator $\hat{\beta}_2$ and its standard deviation, and hence construct a specific test at significance level α. (Assume s^2, the usual unbiased estimator of σ^2, is known.)

5. Calculate the power of the test in question 4 for the case when $\beta_2 \neq 0$, assuming β_2 and σ^2 are known. (The answer will be given by the Pearson-Hartley procedure for certain α, ϕ, ν_1, ν_2. A complete answer to this question will give explicit expressions for ϕ, ν_1 and ν_2.)

6. Suppose we want to calculate simultaneous 95% confidence intervals for $E\{y_j\}$, $j = -2, -1, 0, 1, 2$. A natural way to do this is through a set of intervals of form
$$(\hat{y}_j - Gs_j,\ \hat{y}_j + Gs_j), \tag{4}$$
where $\hat{y}_j$ is the estimated value of $E\{y_j\}$ and s_j is the corresponding confidence interval standard error. Describe two methods of determining G, giving an an explicit formula in each case. (For this question, you do not need to determine formulas for $\hat{y}_j$ and s_j: assume these are known.)

SOLUTIONS

1. We have,

$$X = \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix}, \; X^T X = \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix}, \; (X^T X)^{-1} = \begin{bmatrix} \frac{1}{5} & 0 \\ 0 & \frac{1}{10} \end{bmatrix},$$

$$X^T Y = \begin{bmatrix} y_{-2} + y_{-1} + y_0 + y_1 + y_2 \\ -2y_{-2} - y_{-1} + y_1 + 2y_2 \end{bmatrix}, \qquad (5)$$

which leads quickly to the given expression for $\tilde{\beta}_0$, and variance $\frac{\sigma^2}{5}$.

2. The corresponding results for quadratic regression give

$$X = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}, \; X^T X = \begin{bmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{bmatrix}, \; (X^T X)^{-1} = \frac{1}{70} \begin{bmatrix} 34 & 0 & -10 \\ 0 & 7 & 0 \\ -10 & 0 & 5 \end{bmatrix}. \qquad (6)$$

We also have

$$X^T Y = \begin{bmatrix} y_{-2} + y_{-1} + y_0 + y_1 + y_2 \\ -2y_{-2} - y_{-1} + y_1 + 2y_2 \\ 4y_{-2} + y_{-1} + y_1 + 4y_2 \end{bmatrix}. \qquad (7)$$

Hence

$$\hat{\beta}_0 = \frac{1}{70}\{34(y_{-2} + y_{-1} + y_0 + y_1 + y_2) - 10(4y_{-2} + y_{-1} + y_1 + 4y_2)\}, \quad (8)$$

which quickly reduces to the form given, and the corresponding variance is σ^2 times the first diagonal entry of $(X^T X)^{-1}$, i.e. $\frac{17}{35}\sigma^2$.

3. The expected value of $\tilde{\beta}_0$ is

$$\frac{1}{5}(5\beta_0 + \beta_1 \sum x_j + \beta_2 \sum x_j^2) = \beta_0 + 2\beta_2,$$

so the bias is $2\beta_2$. The mean squared errors of $\hat{\beta}_0$ and $\tilde{\beta}_0$ are respectively $\frac{17}{35}\sigma^2$ and $\frac{1}{5}\sigma^2 + 4\beta_2^2$. We should therefore prefer $\tilde{\beta}_0$ if

$$\frac{1}{5}\sigma^2 + 4\beta_2^2 < \frac{17}{35}\sigma^2,$$

which quickly reduces to the form given.

4. We have

$$\begin{aligned}\hat\beta_2 &= \frac{1}{70}\{-10(y_{-2}+y_{-1}+y_0+y_1+y_2)+5(4y_{-2}+y_{-1}+y_1+4y_2)\}\\ &= \frac{2y_{-2}-y_{-1}+2y_0-y_1+2y_2}{14} \qquad (9)\end{aligned}$$

with variance $\frac{5\sigma^2}{70}=\frac{\sigma^2}{14}$, so a $100(1-\alpha)\%$ hypothesis test for $H_0:\ \beta_2=0$ will reject H_0 if

$$|\hat\beta_2| > \frac{s}{\sqrt{14}}t_{2;1-\alpha/2}. \qquad (10)$$

(The degrees of freedom are $n-p=5-3=2$.)

5. The test of question 4 may be expressed in the form $C\beta=h$ where $C=\begin{pmatrix}0 & 0 & 1\end{pmatrix}$ and $h=0$. The alternative has $h'=\beta_2$. Thus

$$\begin{aligned}\sigma^2\delta^2 &= (h-h')^T\left\{C(X^TX)^{-1}C^T\right\}^{-1}(h-h')\\ &= 14\beta_2^2 \qquad (11)\end{aligned}$$

and the degrees of freedom of the F test of question 4 are 1 and 2. (Although this is not essential to the answer of the question, it might be worth pointing out that the answer to question 4 may be reformulated as an F test because $14\hat\beta_2^2/s^2$ is distributed as the square of a t_2 random variable, which is $F_{1,2}$.) Therefore the parameters of the Pearson-Hartley procedure may be written as $\phi=\frac{\delta}{\sqrt{2}}=\frac{\sqrt{7}\beta_2}{\sigma}$, $\nu_1=1$ and $\nu_2=2$, along with α which is the significance level of the test.

Alternatively: It is possible to do this question by the substitution rule, though in this case, I think this is harder than the solution just given. The substitution rule says that $\sigma^2\delta^2$ is the value of SSE_0-SSE_1 when the values y_j are substituted by their true values under H_1: in other words, $y_j=\beta_0+\beta_1 j+\beta_2 j^2$. If we fit that data set under H_1, then the choice $\hat\beta_i=\beta_i$ $(i=0,1,2)$ exactly fits the data, so $SSE_1=0$. Under H_0, we estimate $\hat\beta_0=\bar y=\beta_0+2\beta_2$, while $\hat\beta_1=\beta_1$ (e.g. use the expression for $\hat\beta_1$ that is derived from the answer to question 1). Hence $e_j=y_j-\hat y_j=(\beta_0+\beta_1 j+\beta_2 j^2)-(\beta_0+\beta_1 j+2\beta_2)=(j^2-2)\beta_2$ and $SSE_0=\beta_2^2\sum\{(j^2-2)^2\}=\beta_2^2\{2^2+(-1)^2+(-2)^2+(-1)^2+2^2\}=14\beta_2^2$ and the rest follows the first solution given in the preceding paragraph. In the actual exam, several people tried it this way but nobody got the correct answer, whereas several people got the correct answer using the first method.

6. Bonferroni procedure: $G=t_{2;1-\alpha/10}$.

Scheffé procedure: $G^2=3F_{3,2;1-\alpha}$. Here $q=3$ because all linear combinations of $(\beta_0,\beta_1,\beta_2)$ lie in the same 3-dimensional subspace defined by the three parameter values.

For 95% confidence intervals, $\alpha=.05$.