STATISTICS 174: APPLIED STATISTICS

FINAL EXAM

DECEMBER 12, 2003

Time allowed: 3 hours.

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 instructor for clarification if you feel the question is ambiguous.

Where questions require a numerical answer, it is more important to demonstrate precise understanding of the method of calculation than to get the actual numerical answer correct.

Statistical tables are not provided: except where explicitly indicated otherise, the exam does not require precise numerical knowledge of any distributions.

A provisional mark scheme is given in square brackets (total 100 marks).

1. Suppose you are fitting a quadratic regression

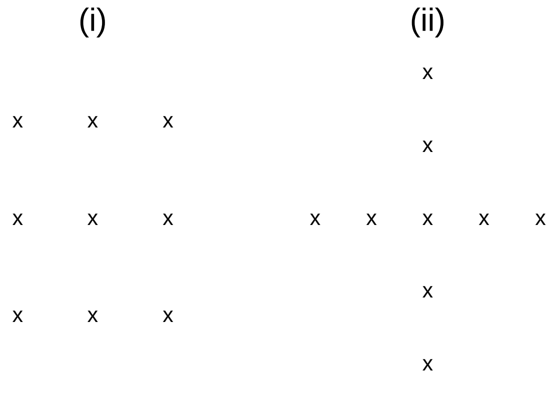
$$
\begin{aligned}
y_i &= f(x_i) + \epsilon_i,\ i = 1, 2, ..., n, & (1)\\
f(x) &= \beta_0 + \beta_1 x + \beta_2 x^2 & (2)
\end{aligned}
$$

to a set of covariates x_i lying in the interval $(0, 1)$. Suppose it is known that $\beta_0 < 0,\ \beta_0 + \beta_1 + \beta_2 > 0,\ \beta_2 > 0$. Also the least squares estimates, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, satisfy the same inequalities. Thus $f(x)$ is a convex function (increasing to $\pm\infty$ as $x \to \pm\infty$) and has a unique root $x^* \in (0, 1)$, satisfying $f(x^*) = 0$. Our objective is to estimate x^*, with a confidence interval.

Describe *two* methods of doing this, one based on the delta method and the other based on Fieller's method. Assume that point estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ are available, and have an estimated covariance matrix $s^2 A$, where $A = (X^T X)^{-1}$. Describe explicitly the procedure by which a point estimate and 95% confidence interval for x^* may be constructed, noting any difficulties that may occur. [20]

(You can assume that $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$, s^2 and A are given: no need to give formulae to derive these. In terms of those quantities, you are expected to give as much explicit detail as you can. There is no need to expend a lot of effort reducing algebraic expressions to their simplest form; it will suffice to give explicit expressions to be evaluated.)

2. Consider the following two possible arrangements for a response surface design:



Design (i) is one of the designs proposed in Section 7.3.1 of the course text (p. 358), while (ii) is an alternative which we will analyze here. In (i), it is assumed that the x_{i1} and x_{i2} coordinates of the sampling points are $-1,\ 0,\ 1$; in (ii), they are $-2,\ -1,\ 0,\ 1,\ 2$.

The main model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{12} x_{i1} x_{i2} + \beta_{22} x_{i2}^2 + \epsilon_i, \tag{3}$$

where the ϵ_i error terms are independent with mean 0 and common variance σ^2.

(a) Why cannot the model (3) be fitted directly from design (ii)? [4]

(b) Suppose now β_{12} is known to be 0. Write down the normal equations for $\beta_0,\ \beta_1,\ \beta_2,\ \beta_{11},\ \beta_{22}$ under design (ii), and solve them to get precise formulae for the estimators $\hat{\beta}_0,\ \hat{\beta}_1,\ \hat{\beta}_2,\ \hat{\beta}_{11},\ \hat{\beta}_{22}$. What are the variances of these estimators, as functions of σ^2? [15]

(c) Suppose $\hat{\beta}_0 = 1.36,\ \hat{\beta}_1 = 0.51,\ \hat{\beta}_2 = -0.79,\ \hat{\beta}_{11} = 0.23, \hat{\beta}_{22} = 0.97$, and we are trying to find (x_1^*, x_2^*) that minimize the expected response, still assuming design (ii) with $\beta_{12} = 0$. (Note that both $\hat{\beta}_{11}$ and $\hat{\beta}_{22}$ are positive, so any estimated stationary point will be a minimum of the estimated function.) We also assume the sample residual standard deviation is $s = 0.875$. Calculate the point estimates $\hat{x}_1^*$ and $\hat{x}_2^*$, and compute approximate 95% confidence intervals by both the delta method and the Fieller method, noting any difficulties that occur. (For the relevant percentage point of a t distribution, use $t^* = t_{4;0.975} = 2.776$.) [15]

(d) Briefly compare the virtues of the designs (i) and (ii). Which has the smaller variances for the parameter estimates? Why do you think design (ii) is little used in practice? [6]

Hints:

(i). You are free and encouraged to quote any results from the text (in particular, pages around 357–364) in the course of answering this question.

(ii). The inverse of the 3×3 matrix

$$\begin{pmatrix} a & c & c \\ c & b & d \\ c & d & b \end{pmatrix}$$

is another matrix of the same structure, so you can solve for $a,\ b,\ c,\ d$ to compute the inverse.

(iii). If you cannot get a complete answer to part (b) you should still attempt parts (c) and (d); credit will be give, as far as possible, for correct method, even if the numerical answers are incorrect.

3. Table 1 (see end of exam) is part of a data base documenting deaths in the salmon population along the Columbia River in Washington State, as a function of time and various pollutants. The observations are as follows:

Y1	Number of dead fish found
SP	Indicator for spring quarter
SU	Indicator for summer quarter
FA	Indicator for fall quarter
YR	Year since start of study
PH	Level of phosphorus in the river
NT	Level of nitrogen in the river
SO2	Level of atmospheric sulfur dioxide
PM10	Level of atmospheric particulate matter

There is no indicator for the winter quarter because this is collinear with the indicators for spring, summer and fall.

(a) An initial analysis is performed of various transformations of Y_1 and all eight other variables as covariates. The transformations considered are: $Y_2 = Y_1^{0.75}$; $Y_3 = Y_1^{0.5}$; $Y_4 = Y_1^{0.25}$; $Y_5 = \log Y_1$. The respective residuals sums of squares (RSS) for response variables Y_1, Y_2, Y_3, Y_4, Y_5 are 13078, 604.26, 23.719, 0.68373 and 2.17356. The geometric mean of the Y_1 values is 27.729. Based on these numbers, which of the five transformations would you select? [4]

(b) Now suppose that the transformation you selected was $Y_3 = \sqrt{Y_1}$ (*not* necessarily the correct answer to the first part!). Various possible combinations of variables are considered, and the corresponding R^2 values calculated using the `SELECTION=RSQUARE` option in SAS. An edited version of the results is given as Table 2. (The full SAS output gives R^2 for each of the possible $2^8 = 256$ models. This has been shortened here so that only the leading candidates, for each possible model order, are included.) The ANOVA table and table of parameter estimates for the full model is at the beginning of Table 3. Based on these tables, give as complete a discussion as you can about the issues involved in selecting some subset of the eight covariates as a model in this instance. (You should consider at least the following model selection criteria: AIC, BIC, forward selection, stepwise selection and backward selection.) [15]

(c) For the model with all eight covariates and Y_3 as the response, a complete table of SAS output (mildly edited) is given as Table 3. Using this table, give as complete a discussion as you can about all the aspects of model diagnostics, including (i) points of high leverage, (ii) residuals and outliers, (iii) influential observations by all the standard criteria for assessing influence, (iv) multicollinearity. [12]

(d) The same model was also fitted in SPLUS and the regression output plotted using the "plot" command. The results are shown in Figure 1, also at the end of the exam. Discuss each of the six plots in this figure and explain its relevance to interpreting the fit of the regression model. [6]

(e) The last four rows in Table 1 are dummy rows (missing variable in Y_1). Suggest a reason why these rows might have been included, and explain how to interpret those parts of the output that relate to these four rows. [3]

TABLE 1

Y1	SP	SU	FA	YR	PH	NT	SO2	PM10
51	1	0	0	1	4.7	25.1	9.3	40.4
133	0	1	0	1	10.1	17.5	8.3	18.9
12	0	0	1	1	4.6	17.6	7.1	44.6
41	0	0	0	1	15.9	22.0	29.1	28.3
38	1	0	0	2	4.3	15.6	12.9	21.4
101	0	1	0	2	5.5	15.4	7.1	51.9
21	0	0	1	2	5.0	22.9	5.0	11.8
19	0	0	0	2	7.9	19.3	13.1	13.9
42	1	0	0	3	1.6	16.9	24.0	27.3
139	0	1	0	3	8.8	26.2	14.7	22.2
11	0	0	1	3	14.7	19.0	10.1	26.5
16	0	0	0	3	14.1	20.5	14.1	13.0
58	1	0	0	4	10.9	17.8	24.3	36.8
103	0	1	0	4	9.7	19.2	9.5	48.1
17	0	0	1	4	8.0	18.5	9.4	7.1
19	0	0	0	4	12.0	19.1	12.6	35.6
294	1	0	0	5	96.9	19.4	26.4	16.9
138	0	1	0	5	12.6	19.4	23.1	78.3
15	0	0	1	5	2.1	20.1	35.4	90.6
15	0	0	0	5	18.6	28.7	18.9	25.5
33	1	0	0	6	13.8	20.3	10.0	73.7
101	0	1	0	6	10.1	22.2	14.7	32.8
13	0	0	1	6	9.4	13.9	17.8	12.4
10	0	0	0	6	3.6	24.5	6.8	38.2
19	1	0	0	7	8.8	24.3	9.7	9.7
60	0	1	0	7	1.8	23.6	12.4	16.8
9	0	0	1	7	6.6	22.0	12.1	59.8
7	0	0	0	7	16.7	14.6	5.9	69.4
29	1	0	0	8	4.2	22.1	14.0	73.7
65	0	1	0	8	9.5	17.3	7.4	75.3
6	0	0	1	8	4.9	25.2	11.0	71.6
13	0	0	0	8	4.0	19.6	7.8	90.0
35	1	0	0	9	25.7	21.0	12.6	37.7
56	0	1	0	9	3.4	22.3	17.5	27.7
10	0	0	1	9	34.7	15.9	15.5	32.1
4	0	0	0	9	4.6	23.3	7.2	23.6
30	1	0	0	10	2.2	22.7	21.1	13.6
97	0	1	0	10	35.8	19.7	14.1	43.3
9	0	0	1	10	3.1	21.4	10.1	113.9
6	0	0	0	10	14.2	18.3	17.1	46.0
·	1	0	0	11	12.128	20.36	13.98	40.51
·	0	1	0	11	12.128	20.36	13.98	40.51
·	0	0	1	11	12.128	20.36	13.98	40.51
·	0	0	0	11	12.128	20.36	13.98	40.51

TABLE 2

```
Number in
  Model      R-Square    Variables in Model

      1        0.4415    su
      1        0.2460    ph
      1        0.2014    fa
      1        0.0626    so
      1        0.0597    yr
      1        0.0374    sp
      2        0.7233    su ph
      2        0.6351    sp su
      2        0.5398    su so
      2        0.5012    su yr
      2        0.4996    su fa
      2        0.4655    su pm
      3        0.8444    sp su ph
      3        0.8076    su yr ph
      3        0.7537    su fa ph
      3        0.7532    su ph so
      3        0.7278    su ph pm
      3        0.7243    su ph nt
      4        0.9255    sp su yr ph
      4        0.8597    sp su ph so
      4        0.8457    sp su ph pm
      4        0.8446    sp su ph nt
      4        0.8444    sp su fa ph
      4        0.8368    su fa yr ph
      5        0.9377    sp su yr ph so
      5        0.9308    sp su yr ph pm
      5        0.9284    sp su yr ph nt
      5        0.9255    sp su fa yr ph
      5        0.8613    sp su ph so pm
      5        0.8598    sp su ph nt so
      6        0.9421    sp su yr ph so pm
      6        0.9403    sp su yr ph nt so
      6        0.9378    sp su fa yr ph so
      6        0.9352    sp su yr ph nt pm
      6        0.9309    sp su fa yr ph pm
      6        0.9284    sp su fa yr ph nt
      7        0.9460    sp su yr ph nt so pm
      7        0.9423    sp su fa yr ph so pm
      7        0.9403    sp su fa yr ph nt so
      7        0.9352    sp su fa yr ph nt pm
      7        0.8653    su fa yr ph nt so pm
      7        0.8614    sp su fa ph nt so pm
      8        0.9460    sp su fa yr ph nt so pm
```

TABLE 3

```
                              The REG Procedure
                                Model: MODEL1
                           Dependent Variable: y3

                            Analysis of Variance

                                    Sum of          Mean
  Source                  DF       Squares        Square    F Value    Pr > F

  Model                    8     415.35894      51.91987      67.86    <.0001
  Error                   31      23.71925       0.76514
  Corrected Total         39     439.07819


            Root MSE              0.87472    R-Square     0.9460
            Dependent Mean        6.03308    Adj R-Sq     0.9320
            Coeff Var            14.49875

                             Parameter Estimates

                   Parameter      Standard                           Variance
Variable    DF      Estimate         Error   t Value   Pr > |t|     Inflation

Intercept    1       2.10450       1.03177      2.04     0.0500             0
sp           1       2.71702       0.39920      6.81     <.0001       1.56207
su           1       6.23606       0.39276     15.88     <.0001       1.51209
fa           1      -0.04979       0.39824     -0.13     0.9013       1.55456
yr           1      -0.36907       0.05297     -6.97     <.0001       1.21028
ph           1       0.10321       0.00971     10.63     <.0001       1.20760
nt           1       0.06366       0.04390      1.45     0.1571       1.08977
so           1       0.05367       0.02160      2.49     0.0186       1.12564
pm           1       0.01072       0.00592      1.81     0.0799       1.24430

                           Collinearity Diagnostics

                           Condition  -------Proportion of Variation-------
    Number   Eigenvalue        Index    Intercept           sp           su

         1      5.62789      1.00000   0.00054371      0.00437      0.00393
         2      1.07252      2.29071   0.00002461      0.23218      0.03133
         3      1.00128      2.37080   0.00000194      0.01221      0.30224
         4      0.55744      3.17741   0.00019241      0.09824      0.00146
         5      0.27442      4.52860   0.00064000      0.21246      0.27463
         6      0.19741      5.33930      0.00324      0.39651      0.26794
         7      0.16182      5.89733      0.00167   0.00082776      0.00276
         8      0.09649      7.63706      0.03534      0.02707      0.08028
```

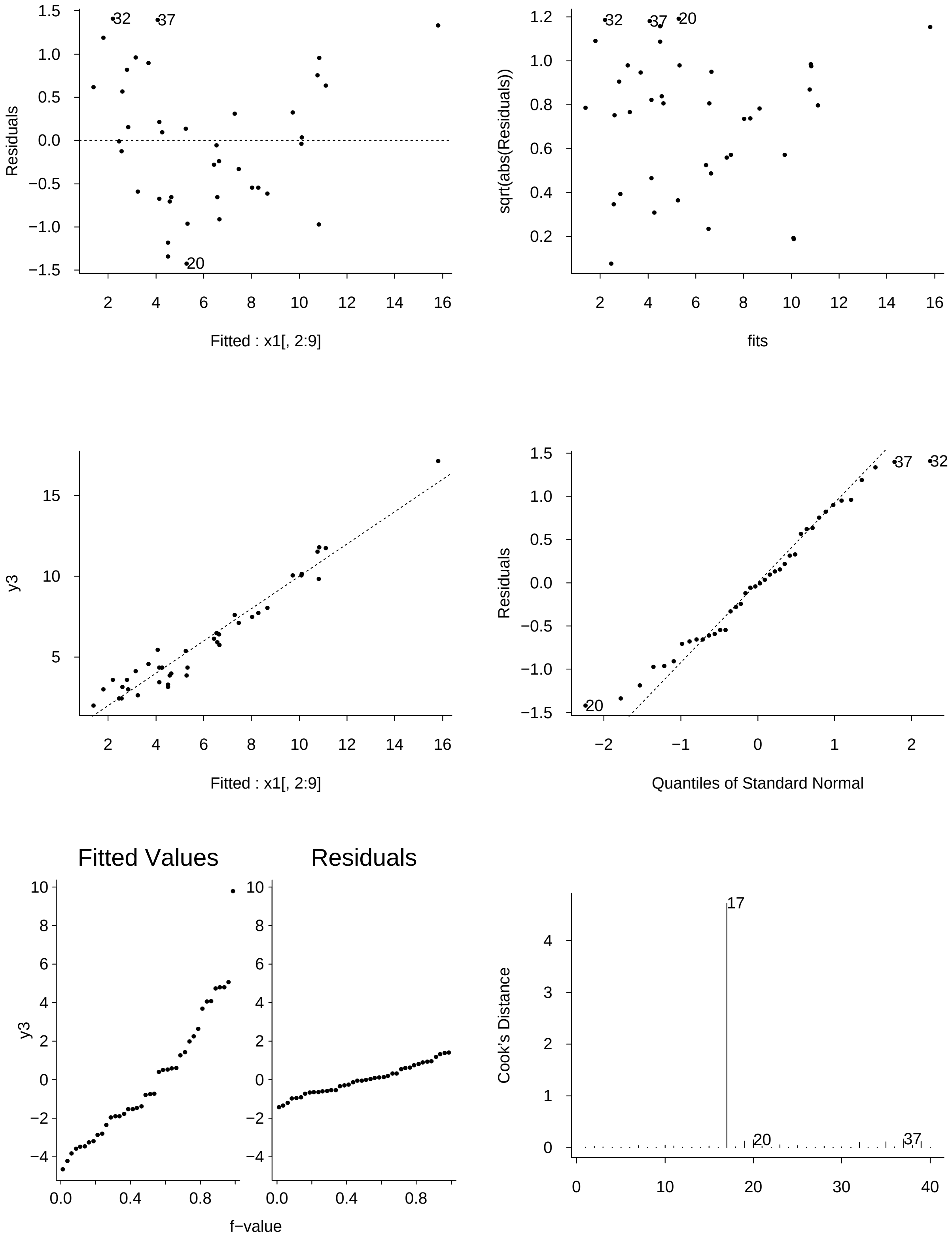
```
     9      0.01072     22.90872      0.95835      0.01612      0.03543

        --------------------Proportion of Variation--------------------
Number           fa           yr           ph           nt           so

     1      0.00391      0.00474      0.00768   0.00069027      0.00458
     2      0.12538   0.00041704      0.04415   0.00002084      0.00101
     3      0.17486   0.00001742  2.345115E-9   0.00000860   0.00008915
     4      0.01772      0.00392      0.69077   0.00052311   0.00060600

        --------------------Proportion of Variation--------------------
Number           fa           yr           ph           nt           so

     5      0.30949      0.12142      0.05795   0.00068498      0.02740
     6      0.25124      0.00901      0.08917      0.00819      0.33717
     7      0.00111      0.46006      0.01896      0.00728      0.13303
     8      0.05923      0.39510      0.05090      0.06537      0.48408
     9      0.05707      0.00530      0.04042      0.91723      0.01202

                                  -Proportion of Variation-
                          Number           pm

                               1      0.00567
                               2      0.00703
                               3   0.00009175
                               4      0.03670
                               5      0.23085
                               6      0.10289
                               7      0.55147
                               8      0.01707
                               9      0.04822

                          Output Statistics

         Dep Var    Predicted        Std Error
   Obs        y3        Value     Mean Predict          95% CL Mean

     1    7.1414       7.4678           0.4646       6.5204       8.4153
     2   11.5326      10.7762           0.3816       9.9979      11.5545
     3    3.4641       4.1402           0.3865       3.3520       4.9284
     4    6.4031       6.6424           0.4849       5.6535       7.6314
     5    6.1644       6.4422           0.4061       5.6139       7.2705
     6   10.0499      10.0881           0.4080       9.2560      10.9202
     7    4.5826       3.6854           0.4213       2.8261       4.5446
     8    4.3589       4.2626           0.3410       3.5671       4.9581
     9    6.4807       6.5362           0.4231       5.6733       7.3992
    10   11.7898      10.8366           0.4082      10.0042      11.6691
    11    3.3166       4.5006           0.3253       3.8372       5.1640
```

```
 12    4.0000       4.6539          0.3192       4.0029       5.3048
 13    7.6158       7.3023          0.3642       6.5595       8.0451
 14   10.1489      10.1135          0.3044       9.4926      10.7343
 15    4.1231       3.1626          0.3672       2.4136       3.9116
 16    4.3589       4.1408          0.2953       3.5385       4.7430
 17   17.1464      15.8109          0.7782      14.2237      17.3981
 18   11.7473      11.1102          0.4125      10.2690      11.9514
 19    3.8730       4.5773          0.6107       3.3316       5.8229
 20    3.8730       5.2939          0.4601       4.3555       6.2323
 21    5.7446       6.6508          0.3778       5.8803       7.4214
 22   10.0499       9.7226          0.2963       9.1184      10.3268
 23    3.6056       2.7838          0.4679       1.8294       3.7381
 24    3.1623       2.5960          0.3418       1.8990       3.2930
 25    4.3589       5.3179          0.3901       4.5224       6.1135
 26    7.7460       8.2910          0.3665       7.5436       9.0384
 27    3.0000       2.8437          0.3065       2.2187       3.4687
 28    2.6458       3.2350          0.4527       2.3117       4.1584
 29    5.3852       5.2511          0.3683       4.5000       6.0023
 30    8.0623       8.6745          0.3776       7.9045       9.4446
 31    2.4495       2.5704          0.3973       1.7600       3.3808
 32    3.6056       2.1963          0.4119       1.3562       3.0365
 33    5.9161       6.5700          0.3445       5.8673       7.2726
 34    7.4833       8.0258          0.3946       7.2211       8.8306
 35    3.1623       4.5030          0.4377       3.6104       5.3956
 36    2.0000       1.3805          0.3862       0.5928       2.1682
 37    5.4772       4.0814          0.4794       3.1037       5.0590
 38    9.8489      10.8202          0.4108       9.9823      11.6581
 39    3.0000       1.8098          0.4761       0.8389       2.7808
 40    2.4495       2.4556          0.3954       1.6492       3.2619
 41         .       4.4945          0.4035       3.6714       5.3175
 42         .       8.0135          0.4031       7.1914       8.8356
 43         .       1.7276          0.4087       0.8942       2.5611
 44         .       1.7774          0.4028       0.9560       2.5989

                                  Std Error  Student
Obs    95% CL Predict    Residual  Residual Residual   -2-1 0 1 2

  1    5.4478    9.4878   -0.3264      0.741   -0.440 |      |      |
  2    8.8298   12.7226    0.7564      0.787    0.961 |      |*     |
  3    2.1898    6.0906   -0.6761      0.785   -0.862 |     *|      |
  4    4.6027    8.6822   -0.2393      0.728   -0.329 |      |      |
  5    4.4753    8.4091   -0.2778      0.775   -0.359 |      |      |
  6    8.1196   12.0566   -0.0382      0.774  -0.0494 |      |      |
  7    1.7052    5.6655    0.8972      0.767    1.170 |      |**    |
  8    2.3478    6.1774    0.0963      0.806    0.120 |      |      |
  9    4.5545    8.5180   -0.0555      0.766  -0.0725 |      |      |
 10    8.8680   12.8053    0.9532      0.774    1.232 |      |**    |
 11    2.5972    6.4040   -1.1840      0.812   -1.458 |    **|      |
 12    2.7548    6.5529   -0.6539     0.814   -0.803 |     *|      |
 13    5.3699    9.2348    0.3134     0.795    0.394 |      |      |
 14    8.2245   12.0024    0.0354     0.820   0.0432 |      |      |
 15    1.2277    5.0974    0.9606     0.794    1.210 |      |**    |
 16    2.2579    6.0237    0.2181     0.823    0.265 |      |      |
 17   13.4230   18.1988    1.3355     0.399    3.344 |      |******|
 18    9.1379   13.0826    0.6371     0.771    0.826 |      |*     |
 19    2.4014    6.7531   -0.7043     0.626   -1.125 |    **|      |
 20    3.2781    7.3096   -1.4209     0.744   -1.910 |   ***|      |
 21    4.7075    8.5942   -0.9063     0.789   -1.149 |    **|      |
 22    7.8391   11.6062    0.3273     0.823    0.398 |      |      |
 23    0.7605    4.8070    0.8218     0.739    1.112 |      |**    |
 24    0.6806    4.5113    0.5663     0.805    0.703 |      |*     |
 25    3.3646    7.2713   -0.9590     0.783   -1.225 |    **|      |
 26    6.3567   10.2252   -0.5450     0.794   -0.686 |     *|      |
 27    0.9534    4.7340    0.1563     0.819    0.191 |      |      |
 28    1.2263    5.2438   -0.5893     0.748   -0.787 |     *|      |
 29    3.3154    7.1868    0.1340     0.793    0.169 |      |      |
 30    6.7315   10.6176   -0.6123     0.789   -0.776 |     *|      |
 31    0.6109    4.5298   -0.1209     0.779   -0.155 |      |      |
 32    0.2244    4.1683    1.4092     0.772    1.826 |      |***   |
 33    4.6526    8.4874   -0.6539     0.804   -0.813 |     *|      |
 34    6.0687    9.9829   -0.5425     0.781   -0.695 |     *|      |
 35    2.5082    6.4979   -1.3407     0.757   -1.770 |   ***|      |
 36   -0.5697    3.3307    0.6195     0.785    0.789 |      |*     |
 37    2.0470    6.1157    1.3959     0.732    1.908 |      |***   |
 38    8.8492   12.7912   -0.9713     0.772   -1.258 |    **|      |
 39   -0.2213    3.8410    1.1902     0.734    1.622 |      |***   |
 40    0.4978    4.4133 -0.006062     0.780  -0.0078 |      |      |
 41    2.5298    6.4591         .         .        .
 42    6.0492    9.9778         .         .        .
 43   -0.2414    3.6967         .         .        .
 44   -0.1866    3.7415         .         .        .

      Cook's           Hat Diag      Cov          ------DFBETAS-----
Obs        D  RStudent        H    Ratio   DFFITS Intercept       sp

  1    0.008   -0.4346   0.2821   1.7683  -0.2724    0.0706  -0.1377
  2    0.024    0.9598   0.1903   1.2637   0.4653    0.2074   0.0095
  3    0.020   -0.8579   0.1952   1.3420  -0.4225   -0.1398  -0.0233
  4    0.005   -0.3240   0.3073   1.8792  -0.2158    0.0112   0.1068
  5    0.004   -0.3534   0.2156   1.6495  -0.1853   -0.1170  -0.0930
  6    0.000   -0.0486   0.2176   1.7156  -0.0256   -0.0135  -0.0011
  7    0.046    1.1777   0.2320   1.1645   0.6472   -0.0330   0.0445
  8    0.000    0.1176   0.1520   1.5775   0.0498    0.0274  -0.0283
  9    0.000   -0.0713   0.2340   1.7509  -0.0394   -0.0139  -0.0164
 10    0.047    1.2429   0.2177   1.0928   0.6557   -0.3230   0.0063
 11    0.038   -1.4863   0.1383   0.8224  -0.5954   -0.1173  -0.0030
```

```
12    0.011   -0.7982   0.1331   1.2826  -0.3128   -0.1222   0.1968
13    0.004    0.3887   0.1734   1.5531   0.1780    0.0394   0.0832
14    0.000    0.0425   0.1211   1.5275   0.0158    0.0027   0.0006
15    0.035    1.2194   0.1763   1.0551   0.5641    0.1934   0.0058
16    0.001    0.2609   0.1140   1.4854   0.0936    0.0491  -0.0612
17    4.719    4.1146   0.7916   0.1149   8.0182   -1.2901   0.9473
18    0.022    0.8216   0.2223   1.4138   0.4393   -0.0743  -0.0273
19    0.134   -1.1297   0.4875   1.8015  -1.1018    0.3046   0.0896
20    0.155   -2.0003   0.2767   0.6018  -1.2371    0.6561   0.5477
21    0.034   -1.1549   0.1866   1.1164  -0.5531    0.0179  -0.3230
22    0.002    0.3922   0.1147   1.4491   0.1412   -0.0331  -0.0018
23    0.055    1.1164   0.2861   1.3046   0.7068    0.3790  -0.0458
24    0.010    0.6975   0.1526   1.3716   0.2960   -0.0154  -0.1393
25    0.041   -1.2353   0.1988   1.0728  -0.6154    0.1010  -0.3453
26    0.011   -0.6802   0.1755   1.4196  -0.3138    0.0646  -0.0072
27    0.001    0.1878   0.1228   1.5151   0.0702   -0.0231   0.0026
28    0.025   -0.7824   0.2679   1.5298  -0.4733   -0.3102   0.1881
29    0.001    0.1663   0.1773   1.6192   0.0772   -0.0179   0.0458
30    0.015   -0.7709   0.1863   1.3837  -0.3689   -0.0873  -0.0170
31    0.001   -0.1527   0.2064   1.6807  -0.0778    0.0460  -0.0046
32    0.106    1.9018   0.2218   0.6197   1.0152    0.2147  -0.4023
33    0.013   -0.8087   0.1551   1.3094  -0.3465    0.0262  -0.1964
34    0.014   -0.6890   0.2035   1.4642  -0.3482    0.0628   0.0083
35    0.116   -1.8368   0.2504   0.6866  -1.0615   -0.2502   0.0939
36    0.017    0.7844   0.1949   1.3899   0.3860    0.0412  -0.1696
37    0.174    1.9977   0.3003   0.6239   1.3088   -0.1456   0.5094
38    0.050   -1.2702   0.2206   1.0756  -0.6758    0.0663   0.0454
39    0.123    1.6679   0.2962   0.8593   1.0821   -0.3312   0.0640
40    0.000 -0.007642   0.2043   1.6881  -0.0039   -0.0013   0.0021
41      .           .   0.2128     .        .         .        .
42      .           .   0.2124     .        .         .        .
43      .           .   0.2183     .        .         .        .
44      .           .   0.2120     .        .         .        .

                        Output Statistics

    ---------------------------DFBETAS--------------------------
Obs        su        fa        yr        ph        nt        so        pm

  1   -0.0053   -0.0107    0.1619   -0.0045   -0.1374    0.0879   -0.0795
  2    0.2293   -0.0069   -0.2071    0.0157   -0.1141   -0.1221   -0.0671
  3    0.0105   -0.1946    0.2373   -0.0162    0.0545    0.1474   -0.0688
  4    0.0818    0.0830    0.0925    0.0020   -0.0325   -0.1400   -0.0119
  5    0.0073    0.0132    0.0435    0.0590    0.0973    0.0187    0.0291
  6   -0.0109    0.0024    0.0106    0.0006    0.0110    0.0075   -0.0062
  7    0.0211    0.3541   -0.2101    0.0341    0.2265   -0.2700   -0.1753
  8   -0.0286   -0.0281   -0.0161   -0.0060   -0.0097    0.0006   -0.0133
  9    0.0008    0.0027    0.0041    0.0196    0.0165   -0.0191    0.0058
```

10	0.3503	0.0702	-0.2098	0.0315	0.4220	0.0484	-0.0518
11	-0.0020	-0.3694	0.1741	-0.1172	0.0154	0.1545	0.1079
12	0.1871	0.1775	0.0732	-0.0117	0.0138	-0.0115	0.0953
13	-0.0034	-0.0117	-0.0181	-0.0550	-0.0599	0.0907	-0.0006
14	0.0098	-0.0006	-0.0046	0.0012	-0.0013	-0.0040	0.0034
15	0.0043	0.3194	0.0267	-0.0553	-0.1052	-0.1078	-0.3235
16	-0.0636	-0.0641	-0.0191	0.0025	-0.0216	-0.0055	0.0016
17	0.1163	0.3675	-1.0871	7.0874	0.7728	0.1849	0.5057
18	0.2005	-0.0314	-0.0967	0.0050	-0.0093	0.2150	0.2260
19	0.0002	-0.2864	0.1485	0.2371	-0.0490	-0.8577	-0.4067
20	0.4424	0.3684	0.1756	-0.2470	-0.9150	-0.2421	0.0236
21	0.0168	0.0333	0.0926	-0.0667	-0.0273	0.2081	-0.3197
22	0.0970	0.0077	0.0165	-0.0081	0.0328	0.0201	-0.0255
23	-0.0143	0.2664	0.2408	-0.1834	-0.4403	0.1851	-0.3898
24	-0.1616	-0.1487	0.0016	-0.0110	0.1227	-0.1085	0.0037
25	-0.0228	-0.0584	-0.1760	0.0956	-0.1900	0.1827	0.2614
26	-0.1798	-0.0271	-0.1063	0.0885	-0.0803	-0.0179	0.1490
27	0.0010	0.0457	0.0078	0.0004	0.0219	-0.0055	0.0128
28	0.2341	0.2537	-0.0414	-0.0714	0.2626	0.1739	-0.1405
29	-0.0010	-0.0034	0.0108	-0.0158	0.0118	-0.0057	0.0342
30	-0.1723	0.0337	-0.0719	-0.0131	0.1204	0.1099	-0.1319
31	-0.0023	-0.0415	-0.0069	-0.0045	-0.0459	0.0099	-0.0243
32	-0.5236	-0.5669	0.0755	-0.0201	-0.1204	-0.2632	0.5878
33	-0.0009	-0.0090	-0.1553	-0.0762	-0.0078	0.0935	0.0303
34	-0.1817	-0.0158	-0.1950	0.1179	-0.0213	-0.1174	0.1432
35	0.0126	-0.4638	-0.5027	-0.3620	0.4077	-0.0186	0.3233
36	-0.1865	-0.1680	0.2026	-0.0638	0.0481	-0.0998	-0.1494
37	0.0345	0.0560	0.8310	-0.6223	0.0012	0.4383	-0.6652
38	-0.3226	-0.0155	-0.3212	-0.3110	0.0396	0.0338	0.0557
39	-0.0215	0.3762	0.1768	0.0316	0.1859	-0.1501	0.6974
40	0.0020	0.0021	-0.0024	0.0005	0.0015	-0.0009	0.0006

FIGURE 1


Residuals
Fitted : x1[, 2:9]
32
37
20
sqrt(abs(Residuals))
fits
32
37
20
y3
Fitted : x1[, 2:9]
Residuals
Quantiles of Standard Normal
37
32
20
Fitted Values
Residuals
y3
f−value
Cook's Distance
17
20
37

SOLUTIONS

General comment about the exam. This was an exam with a lot of numerical computation if you tried to do everything, but I also emphasized that I didn't regard numerical computation (under exam conditions, with a strict time limit and only a pocket calculator for assistance) as the sole point of the exam and would give substantial partial credit for correct method even if the numerical answers were incorrect. Some students took this too literally and gave *only* a verbal description of the method without any attempt to carry out the computations on the data in hand. The was fine for question 1, where there was no specific data, but for the others, I did expect evidence of a serious attempt to get the numerical solution correct. Other students tried to compute every numerical detail and never finished the exam — while I can understand that it's difficult to know exactly how much detail to give, I did regard it as part of the test to be able to judge how much detail was needed to justify the answer, especially with question 3. In much of applied statistics using packages, one is overwhelmed with computer output; a large part of the skill is judging what information and detail is really needed to back up a proposed model or solution.

In grading the exam, I tried to give fair credit both to verbal descriptions and numerical computations, but it was sometimes difficult to decide exactly how much credit to give each type of answer. In the end, I felt all the students in the class had made a good attempt at the exam and deserved to pass the course, but the actual scores were lower than I would normally have expected.

1. Delta method: write the solution of the quadratic equation as

$$x^* = \frac{-\beta_1 + \sqrt{\beta_1^2 - 4\beta_0\beta_2}}{2\beta_2}.$$

(There are two roots; but the assumptions in the question make it clear that the larger of the two roots is the one required, and also that this is a real root.) The estimate $\hat{x}^*$ is obtained by substituting $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ for $\beta_0, \beta_1, \beta_2$.

Differentiating x^* with respect to each parameter,

$$\begin{aligned}
\frac{\partial x^*}{\partial \beta_0} &= -(\beta_1^2 - 4\beta_0\beta_2)^{-1/2}, \\
\frac{\partial x^*}{\partial \beta_1} &= \frac{\beta_1(\beta_1^2 - 4\beta_0\beta_2)^{-1/2} - 1}{2\beta_2}, \\
\frac{\partial x^*}{\partial \beta_2} &= -\frac{\beta_0(\beta_1^2 - 4\beta_0\beta_2)^{-1/2}}{\beta_2} + \frac{\beta_1 - \sqrt{\beta_1^2 - 4\beta_0\beta_2}}{2\beta_2^2}.
\end{aligned}$$

Substitute $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ for $\beta_0, \beta_1, \beta_2$, and let c_0, c_1, c_2 be the resulting approximations to $\frac{\partial x^*}{\beta_0}$, $\frac{\partial x^*}{\beta_1}$, $\frac{\partial x^*}{\beta_2}$. The estimated standard error of $\hat{x}^*$ is $s\sqrt{c^T AC}$ and the corresponding confidence interval is

$$\hat{x}^* \pm t^* s\sqrt{c^T AC},$$

where $t^* = t_{n-3;.975}$.

Alternatively: (this approach avoids the explicit solution of a quadratic equation and therefore could in principle be used for other types of equation for which there is not an explicit solution)

If $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 x^2$ and $\hat{x}^*$ solves $\hat{f}(\hat{x}^*) = 0$, we have

$$\begin{aligned} 0 &= f(x^*) - \hat{f}(\hat{x}^*) \\ &= f(x^*) - f(\hat{x}^*) + f(\hat{x}^*) - \hat{f}(\hat{x}^*) \\ &\approx (x^* - \hat{x}^*)f'(x^*) + (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)\hat{x}^* + (\beta_2 - \hat{\beta}_2)\hat{x^*}^2 \\ &\approx (x^* - \hat{x}^*)f'(x^*) + (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x^* + (\beta_2 - \hat{\beta}_2)x^{*2} \end{aligned}$$

where the last step follows because essentially this is only a first-order approximation and therefore terms like $(\beta_1 - \hat{\beta}_1)(x^* - \hat{x}^*)$ may be ignored.

Thus we are led to

$$\hat{x}^* - x^* \approx -\frac{1}{f'(x^*)} \left\{ (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)x^* + (\hat{\beta}_2 - \beta_2)x^{*2} \right\}.$$

The same method as given earlier therefore works, if we define $c_0 = -\frac{1}{\hat{f}'(\hat{x}^*)}$, $c_1 = -\frac{\hat{x}^*}{\hat{f}'(\hat{x}^*)}$, $c_2 = -\frac{\hat{x^*}^2}{\hat{f}'(\hat{x}^*)}$ (substituting estimates for unknown parameters throughout). With $f'(x) = \beta_1 + 2\beta_2 x$, this leads to the same expressions as given above.

Fieller's method: this method will include in the confidence interval all values of x for which the hypothesis $H_0: \ \beta_0 + \beta_1 x + \beta_2 x^2 = 0$ would be accepted, at two-sided significance level .05. Denoting the individual entries of A by $a_{ij}, \ i, j \in \{0, 1, 2\}$, the estimated standard error of $\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2$ is $s\sqrt{a_{00} + 2a_{01}x + (a_{11} + 2a_{02})x^2 + 2a_{12}x^3 + a_{22}x^4}$, hence the hypothesis test includes all x values for which

$$(\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2)^2 \le t^{*2} s^2 \{a_{00} + 2a_{01}x + (a_{11} + 2a_{02})x^2 + 2a_{12}x^3 + a_{22}x^4\}. \qquad (4)$$

Equality is satisfied in (4) if

$$\begin{aligned} &(\hat{\beta}_2^2 - t^{*2}s^2 a_{22})x^4 + 2(\hat{\beta}_1\hat{\beta}_2 - t^{*2}s^2 a_{12})x^3 + \{\hat{\beta}_1^2 - 2\hat{\beta}_0\beta_2 - t^{*2}s^2(a_{11} + 2a_{02})\}x^2 \\ &+2(\hat{\beta}_0\hat{\beta}_1 - t^{*2}s^2 a_{01})x + \hat{\beta}_0^2 - t^{*2}s^2 a_{00} = 0. \qquad (5) \end{aligned}$$

There are now various possibilities. The cleanest solution is if (5) has exactly two roots in the interval $(0, 1)$, and the inequality (4) is satisfied between those two roots. In that case, these two roots define the boundaries of a confidence interval for x^*. In other cases, it's possible the entire interval $(0, 1)$ might be included in the confidence interval, or an interval of the form $(0, x^\dagger)$ or $(x^\dagger, 1)$ (where $x^\dagger \in (0, 1)$ is one of the real roots of (5)), or, conceivably, (5) has either three or four roots within $(0, 1)$, and the confidence set consists of two disjoint intervals.

Comment. Many students misinterpreted this question and assumed that it was about finding the maximum or minimum of f, not the solution of $f(x) = 0$. Most students who interpreted the question this way did give a fairly complete and correct answer; but this case was covered in detail in class and in the lecture notes that were accessible during the exam, so I didn't feel I could give better than 50% credit for this solution. Well spotted to one student who noticed that a very similar question was asked on the 2001 final exam (question 2(b)), but they weren't exactly the same because the 2001 question made certain assumptions about the form of the $X^T X$ matrix that were not made here.

2. (a) The full X matrix and corresponding X^TX for design (ii) are

$$X = \begin{pmatrix} 1 & -2 & 0 & 4 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 2 & 0 & 4 & 0 & 0 \\ 1 & 0 & -2 & 0 & 0 & 4 \\ 1 & 0 & -1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 2 & 0 & 0 & 4 \end{pmatrix}, \qquad X^TX = \begin{pmatrix} 9 & 0 & 0 & 10 & 0 & 10 \\ 0 & 10 & 0 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 & 0 & 0 \\ 10 & 0 & 0 & 34 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 10 & 0 & 0 & 0 & 0 & 34 \end{pmatrix}.$$

However, the fifth row and column of X^TX are entirely zeros, meaning that the fifth parameter (β_{12}) cannot possibly be estimated, e.g. in the normal equations $X^TX\hat{\beta} = X^TY$, the coefficient of $\hat{\beta}_{12}$ is 0, meaning any value of $\hat{\beta}_{12}$ is consistent with the normal equations.

(b) Now omit the fifth row and column from the previous X^TX matrix. The matrix to be inverted is

$$X^TX = \begin{pmatrix} 9 & 0 & 0 & 10 & 10 \\ 0 & 10 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 & 0 \\ 10 & 0 & 0 & 34 & 0 \\ 10 & 0 & 0 & 0 & 34 \end{pmatrix}.$$

The second and third rows and columns are orthogonal to the rest; the first, fourth and fifth rows and columns form the matrix $X^TX = \begin{pmatrix} 9 & 10 & 10 \\ 10 & 34 & 0 \\ 10 & 0 & 34 \end{pmatrix}$. Using the hint, the inverse is also of the structure

$$\begin{pmatrix} a & c & c \\ c & b & d \\ c & d & b \end{pmatrix}$$

and, after solving for a, b, c, d, we find $a = \frac{17}{53}$, $b = \frac{103}{1802}$, $c = -\frac{5}{53}$, $d = \frac{25}{901}$. Putting this together with the second and third rows and columns of X^TX,

$$(X^TX)^{-1} = \begin{pmatrix} \frac{17}{53} & 0 & 0 & -\frac{5}{53} & -\frac{5}{53} \\ 0 & \frac{1}{10} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{10} & 0 & 0 \\ -\frac{5}{53} & 0 & 0 & \frac{103}{1802} & \frac{25}{901} \\ -\frac{5}{53} & 0 & 0 & \frac{25}{901} & \frac{103}{1802} \end{pmatrix}.$$

The estimators satisfy $\hat{\beta} = (X^TX)^{-1}X^TY$, in other words,

$$\begin{aligned} \hat{\beta}_0 &= \frac{17}{53}\sum y_i - \frac{5}{53}\sum y_i x_{i1}^2 - \frac{5}{53}\sum y_i x_{i2}^2, \\ \hat{\beta}_1 &= \frac{\sum y_i x_{i1}}{10}, \\ \hat{\beta}_2 &= \frac{\sum y_i x_{i2}}{10}, \end{aligned}$$

$$\begin{aligned}\hat{\beta}_{11} &= -\frac{5}{53}\sum y_i + \frac{103}{1802}\sum y_i x_{i1}^2 + \frac{25}{901}\sum y_i x_{i2}^2, \\ \hat{\beta}_{22} &= -\frac{5}{53}\sum y_i + \frac{25}{901}\sum y_i x_{i1}^2 + \frac{103}{1802}\sum y_i x_{i2}^2,\end{aligned}$$

and their variances are respectively $\frac{17}{53}\sigma^2$, $\frac{1}{10}\sigma^2$, $\frac{1}{10}\sigma^2$, $\frac{103}{1802}\sigma^2$, $\frac{103}{1802}\sigma^2$.

(c) We have $x_1^* = -\frac{\beta_1}{2\beta_{11}}$, $x_2^* = -\frac{\beta_2}{2\beta_{22}}$ from equation (7.23) (page 359 of the text) in which we assume $\beta_{12} = 0$. Estimates $\hat{x}_1^*$, $\hat{x}_2^*$ are obtained by substituting $\hat{\beta}_1$ etc. This leads to $\hat{x}_1^* = -1.109,\ \hat{x}_2^* = 0.407$.

Delta method: since $\frac{\partial x_1^*}{\partial \beta_1} = -\frac{1}{2\beta_{11}}$, $\frac{\partial x_2^*}{\partial \beta_{11}} = \frac{\beta_1}{2\beta_{11}^2}$, substitute $\hat{\beta}_1$, $\hat{\beta}_{11}$ to get multipliers $c_1 = -2.174,\ c_{11} = 4.820$. The standard error is then $s\sqrt{\frac{c_1^2}{10} + \frac{103c_{11}^2}{1802}} = 1.174$, and the approximate 95% confidence interval is $\hat{x}_1^* \pm 2.776 \times 1.174 = (-4.368, 2.151)$.

The corresponding results for x_2^* are: $c_2 = -\frac{1}{2\hat{\beta}_{22}} = -.515$, $c_{22} = \frac{\hat{\beta}_2}{2\hat{\beta}_{22}^2} = -.420$, standard error $s\sqrt{\frac{c_2^2}{10} + \frac{103c_{22}^2}{1802}} = 0.167$, and the approximate 95% confidence interval is $\hat{x}_2^* \pm 2.776 \times 0.167 = (-.058, 0.872)$.

Fieller method: by equation (7.27), page 363, we have to solve

$$(\hat{\theta}_2^2 - t^{*2}s^2 b)x^2 - 2\hat{\theta}_1\hat{\theta}_2 x + (\hat{\theta}_1^2 - t^{*2}s^2 a) = 0$$

since in this problem the covariance parameter c is 0. We have $t^* = 2.776,\ s = .875,\ b = \frac{103}{1802},\ a = \frac{1}{10}$. First do this for $\hat{\theta}_1 = \hat{\beta}_1 = 0.51$, $\hat{\theta}_2 = \hat{\beta}_{11} = 0.23$; the above quadratic equation becomes

$$-0.28434x^2 - 0.23460x - 0.32990 = 0$$

which has no real roots. Thus in this case, the Fieller interval consists of the whole real line.

For $\hat{\theta}_1 = \hat{\beta}_2 = -0.79$, $\hat{\theta}_2 = \hat{\beta}_{22} = 0.97$, the corresponding quadratic equation becomes

$$.60366x^2 + 1.53260x + 0.03410 = 0$$

which does have real roots, at –2.5164 and –0.0224. Multiplying each by $-\frac{1}{2}$, the endpoints of the 95% confidence interval are (0.0112,1.258).

Conclusions: For x_2^*, both forms of confidence interval are computable and the difference between them represents the greater accuracy of the Fieller interval. For x_1^*, both confidence intervals are very wide so it is likely that neither is of much practical value.

(d) Comparing the diagonal entries of the two $(X^TX)^{-1}$ matrices, it appears that design (ii) has smaller variances for all five parameters. However, this could be misleading, because in reality, both designs are adapted to the size of the sampling region (i.e. the spacing between design points is really a variable parameter though this has not been indicated explicitly). When this is taken into account, the apparently smaller variances for design (ii) are not really important, but it is important that design (ii) only allows us to explore variation in directions parallel to the axes, whereas design (i), by allowing us to fit a full quadratic model, is more adaptable to different shapes of surface. More simply, the nonestimability of β_{12} in design (ii) is a more important issue than the apparently smaller variances of the parameter estimates.

3. (a) For transformation $h(y) = Cy^{\lambda}$, the condition $\prod h'(y_i) = 1$ forces $C = \dot{y}^{1-\lambda}/\lambda$; for $h(y) = \log y$, the corresponding condition is $C = \dot{y}$. When the given transformation is modified to include C, the residual sum of squares (RSS) is multiplied by C^2. Applying this with $\dot{y} = 27.729$, the rescaled RSS values are 13078, 5656.8, 2630.8, 1597.4 and 1671.2. Best among these five is Y_4.

   (b) Since $SSTO = 439.078$ (from the ANOVA table in Table 3), the respective values of $SSE = (1 - R^2)SSTO$ for the leading model of each model order are 245.23, 121.49, 68.32, 32.71, 27.35, 25.42, 23.71, 23.71, with respectively 38, 37,...,31 degrees of freedom. Computing $AIC = n \log SSE + 2p$, $BIC = n \log SSE + p \log n$ with $n = 40$ and $p = 2, 3, ..., 9$ for the eight model orders, the AIC values are 224.09, 197.99, 176.97, 149.51, 144.36, 143.43, 142.64, 144.64, and the BIC values are 227.46, 203.06, 183.72, 157.95, 154.49, 155.25, 156.15, 159.84. Best model by AIC has $p = 8$, i.e. all variables except FA; best model by BIC has $p = 6$, i.e. SP, SU, YR, PH, SO (with an intercept, of course).

   Backward selection; begin with model with 8 variables and successively drop FA, NT, PM, SO..., respective F ratios are 0.00, 2.31, 2.51, 6.66,... the first three are insignificant (p values 1, 0.139, 0.124) while the fourth is significant (p value 0.015). (You are not expected to be able to calculate the p-values but from general knowledge of F statistics, you should be able to say that $F \approx 2$ is most likely not significant while $F \approx 6$ is significant.) Thus, backward selection chooses the model with variables SP, SU, YR, PH, SO.

   Forward selection: successively select SU, PH, SP, YR, SO, each of which is highly significant; the next variable to be selected would be PM, but this is not significant (see previous paragraph). Forward selection would not be improved by dropping any variables at an intermediate stage, so stepwise selection leads to the same answer as forward or backward selection.

   (c) (i) Leverage: criterion is $h_i > 2p/n = .45$ (assuming $p = 9,\ n = 40$). Exceeded by observation 17 (by a large margin) and 19 (just).

   (ii) Outliers: Looking at RStudent (the externally studentized residual, and therefore more meaningful than "Student Residual"), the only value that seems even of concern is observation 17, for which the RStudent value is very large (4.11). (The actual two-sided p-value is .00027 but it suffices to note that it is almost certainly significant.) *However*, the actual residual for observation 17 is not so large: it is only the fifth largest in magnitude, behind observations 20, 32, 37, 34; hence it must be some combination of leverage and outlyingness that makes observation 17 so unusual in terms of RStudent.

   (iii) For DFFITS: critical value is $2\sqrt{p/n} = .949$. This is exceeded by observations 17, 19, 20, 32, 35, 37, 39 though all are slight exceedances with the exception of observation 17.

   DFBETAS: critical value is $2/\sqrt{n} = .317$. Exceeded by several parameters with observation 17 (*especially*, the parameter PH) and by at least one parameter in observations 7, 10, 15, 20, 23, 32, 35, 37, 39.

   Cook's D: observation 17 ($D = 4.7$) is enormously larger than any other (next largest is observation 37 with $D = .174$).

   COVRATIO: critical values are $1 \pm 3p/n = (0.325, 1.675)$. The upper bound is violated by numerous observations; the lower bound is violated by observation 17, suggesting that if this observation were omitted, the residual variance would substantially decrease.

MULTICOLLINEARITY: The largest VIF is 1.56, definitely not cause for concern (usual criterion says VIF > 10 is a problem). The largest condition index is 22.9, also not a cause of real concern (a condition index of 30–100 is usually taken to indicate moderate to strong collinearity). The two largest proportions of variance associated with the smallest eigenvalue are .958 for the intercept and .917 for NT; neither of these is very close to 1. *Conclusion:* The is no particular problem with multicollinearity in this data set.

(d) Reading from top left, top right, middle left and so on, the plots are, (i) Residual v. fitted values; no obvious departure from randomness except possibly for the observation at the extreme right hand side (observation 17 has fitted value 15.8 so this must be it); (ii) Square root of absolute residual v. fitted value, used as a graphic of whether σ^2 is constant (it appears this is OK); (iii) plot of observations v. fitted values (stays close to straight line); (iv) QQ plot of residuals (some suggestion of departure from normality at either end); (v) ordered fitted values and residuals (visually confirms substantial reduction in variance due to model fit; note $R^2 = .946$ is close to 1); (vi) Cook's D statistic (confirms observation 17 is very influential).

The overall conclusion from parts (c) and (d) of this question is that the only observation that causes serious difficulties is number 17, but looking at the original data, the real problem is not an anomalous value of Y_1 (or Y_3) but a very unusual value of PH. It seems quite plausible that this value of PH could have been an error and in that case, it would be appropriate to omit observation 17 from the analysis.

(e) The values for PH, NT, SO2 and PM10 are actually the means of the first 40 values, but even if you didn't guess that, it seems clear that the intention is to project a time trend under "typical" values of the pollution parameters. The predicted values are 4.4945, 8.0135, 1.7276, 1.7774; the prediction intervals (given under "95% CL Predict") are $(2.2598, 6.4591)$, $(6.0492, 9.9788)$, $(-.2414, 3.6967)$, $(-.1866, 3.7415)$. (These are prediction intervals for $\sqrt{Y_1}$; square both endpoints to make a prediction interval for Y_1. Values for which the lower prediction limit is < 0 can effectively be treated as 0.)

Comment on student solutions. Many students went through all the diagnostics and gave all the right answers, including identifying observation 17 as one that should be omitted, but still failed to state what was strange about this observation (two points specifically: that it was an outlier in either the internally standardized residual or the externally studentized residual, but *not* in the raw residuals; and that the problem seems to be in the value of PH, not that of the response variable Y1). It's one thing to be able to say that a certain test led to a certain conclusion, but you still have to be able to interpret the results!