STATISTICS 174: APPLIED STATISTICS

TAKE-HOME FINAL EXAM

POSTED ON WEBPAGE: 6:00 pm, DECEMBER 6, 2004

HAND IN BY: 6:00 pm, DECEMBER 7, 2004

This is a take-home exam. You are expected to work on it by yourself and to hand it in to my office (Smith 201) no later than 6:00 pm on Tuesday, December 7. The exam itself and both data sets are posted on the course web page.

You are allowed to consult all course materials and to use computer software, including Excel, SAS, R, S-PLUS and Matlab. Although some of the questions are "pocket calculator" exercises, in all parts of the exam, you are allowed and actively encouraged to use the computer if it simplifies your task.

You are allowed and actively encouraged to quote any formulas or results from the text or homework exercises without repeating the derivation. However if you do this, you should clearly cite what result you are quoting.

Question 1 is largely a theoretical exercise and you are expected to show full working. Questions 2 and 3 are more computational exercises in which you are expected to use one of SAS, R or S-PLUS. In these exercises, you should not hand in lengthy computer output, but only those parts of the output that are directly relevant to the question. However, in all parts of the exam, you are expected to describe what you did (and why) in enough detail that I could, if I wanted to, reproduce your calculations.

You are not allowed to consult with each other or with any other person other than myself. Below, I ask you to sign a "pledge" that you have abided with this rule. PLEASE MAKE SURE YOU SIGN THIS PLEDGE AND HAND IN THIS PAGE WITH YOUR SOLUTION. As with the previous exams in the course, I remind you that the university's Honor Code is in effect.

Grading: Question 1 is worth 40 points, questions 2 and 3 are worth 30 each. You are encouraged to tackle the whole exam, and all answers will be graded. [However, don't feel you have to answer everything! As a rough guide, I expect that over the whole course (average of homeworks, midterm and final), a score of 55–60% will suffice for a P and a score of 75–80% for an H.]

Good luck, and feel free to contact me if you have any queries!

IMPORTANT! Please sign the following and return it with your exam.

Pledge: I certify that this is my own work and I have not discussed this exam with any other person except the instructor.

Signed: Date:

1. A response surface design consists of nine design points (x_{i1}, x_{i2}) with $x_{i1}, x_{i2} \in \{-1, 0, 1\}$. In this question the objective is not to find the location at which the response is minimized or maximized, but to predict the response on a circle of radius 1 around the origin (Fig. 1).

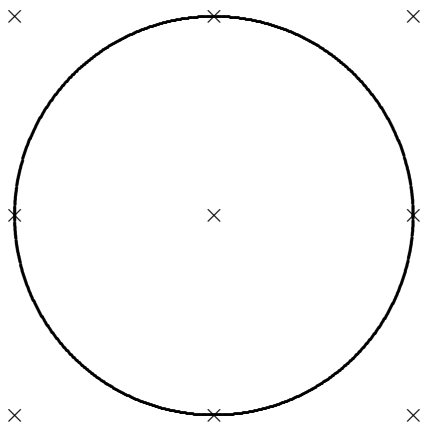

Fig. 1. Illustration of response surface design.

(a) Suppose we fit the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{12} x_{i1} x_{i2} + \beta_{22} x_{i2}^2 + \epsilon_i \qquad (1)$$

where $\epsilon_i,\ i = 1, ..., 9$ are independent $N(0, \sigma^2)$. Suppose also we try to predict the result of a future experiment

$$y^* = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \beta_{11} x_1^{*2} + \beta_{12} x_1^* x_2^* + \beta_{22} x_2^{*2} + \epsilon^* \qquad (2)$$

independent of $y_1, ..., y_9$, where $x_1^* = \cos\theta,\ x_2^* = \sin\theta$, some θ between 0 and 2π.

The obvious predictor is $\widehat{y}^* = \widehat{\beta}_0 + \widehat{\beta}_1 x_1^* + \widehat{\beta}_2 x_2^* + \widehat{\beta}_{11} x_1^{*2} + \widehat{\beta}_{12} x_1^* x_2^* + \widehat{\beta}_{22} x_2^{*2}$ where $\widehat{\beta}_0,\ \widehat{\beta}_1, ...,$ are least squares estimators under the model (1). In terms of θ and σ^2, find an expression for $E\left\{(\widehat{y}^* - y^*)^2\right\}$, and show that this must lie between $\frac{197}{144}\sigma^2$ and $\frac{14}{9}\sigma^2$.

(b) Suppose the true model is again (1), but that the data are analyzed under the incorrect assumption

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \epsilon_i. \qquad (3)$$

Suppose that under (3), the predictor of y^* is $\tilde{y}^*$. In terms of θ, σ^2 and the true value of β_{12}, find expressions for the mean and variance of $y^* - \tilde{y}^*$, and hence for $E\left\{(y^* - \tilde{y}^*)^2\right\}$.

(c) Show that a necessary and sufficient condition for $\tilde{y}^*$ to have smaller mean squared prediction error than $\widehat{y}^*$ is $4\beta_{12}^2 < \sigma^2$, except when (x_1^*, x_2^*) is one of $(1,0)$, $(0,1)$, $(-1,0)$ or $(0,-1)$. Why are those cases different, and what happens then?

The rest of this question is designed to explore how effective the PRESS statistic is in making the correct model choice.

(d) Compute the leverage values $\{h_{ii},\ i = 1, ..., 9\}$ for each of the models (1) and (3).

(e) For each of $i = 1, ..., 9$ and assuming model (1), calculate $E\{(y_i - \widehat{y}_{i(i)})^2\}$.

(f) For each of $i = 1, ..., 9$ and assuming that (1) is the true model, but (3) is used for the estimation, show that $E\{(y_i - \widehat{y}_{i(i)})^2\} = \left(\frac{x_{i1}^* x_{i2}^* \beta_{12}}{1-h_{ii}}\right)^2 + \frac{\sigma^2}{1-h_{ii}}$. Hence evaluate this expression for $i = 1, ..., 9$.

(g) By comparing the expected values of the PRESS statistics computed under each of the models (1) and (3), but assuming everywhere that (1) is the true model that generated the data, show that the PRESS criterion tends to favor model (3) if and only if $\frac{7}{4}\beta_{12}^2 < \sigma^2$. Comment briefly on the discrepancy between this answer and the one in (c).

2. A recent published paper concerning a response surface experiment included the following data set:

x1	x2	x3	y
-1.00000	-1.00000	-1.00000	0.926
-1.00000	-1.00000	1.00000	0.998
-1.00000	1.00000	-1.00000	1.072
-1.00000	1.00000	1.00000	1.091
1.00000	-1.00000	-1.00000	0.926
1.00000	-1.00000	1.00000	1.007
1.00000	1.00000	-1.00000	1.009
1.00000	1.00000	1.00000	1.058
-1.68179	0.00000	0.00000	1.232
1.68179	0.00000	0.00000	0.997
0.00000	-1.68179	0.00000	0.945
0.00000	1.68179	0.00000	1.231
0.00000	0.00000	-1.68179	0.927
0.00000	0.00000	1.68179	1.234
0.00000	0.00000	0.00000	1.245
0.00000	0.00000	0.00000	1.232
0.00000	0.00000	0.00000	1.212

```
0.00000   0.00000   0.00000   1.201
0.00000   0.00000   0.00000   1.222
0.00000   0.00000   0.00000   1.213
```

Here $x1, x2, x3$ were three variables whose optimal values the experimenter was trying to find, and y was the response variable.

(a) Fit the alternative models

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{33} x_{i3}^2 + \epsilon_i \quad (4)$$

and

$$\begin{aligned} y_i \quad = \quad & \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{33} x_{i3}^2 \\ & + \beta_{12} x_{i1} x_{i2} + \beta_{13} x_{i1} x_{i3} + \beta_{23} x_{i2} x_{i3} + \epsilon_i \end{aligned} \quad (5)$$

where x_{i1}, x_{i2}, x_{i3} are the observed values of $x1, x2, x3$ in row i. Which model do you prefer? Find the values of the coefficients, and their standard errors, and verify that each of $\widehat{\beta}_{11}, \widehat{\beta}_{22}, \widehat{\beta}_{33}$ is negative, under either model.

(b) In the published paper, the researchers favored model (4). Assuming this model, let (x_1^*, x_2^*, x_3^*) be the point at which the expected response is maximized. Find point estimates for $x_j^*,\ j = 1, 2, 3$.

(c) Is this a sound analysis? What things might be wrong with it and how might they be corrected?

[*Note:* I am expecting you to run some of the standard regression diagnostics, but unlike some of the earlier exercises, I don't want you to go systematically through all the diagnostics. There is a specific problem somewhere in this analysis; I want you to find it and say how you would deal with it.]

(d) Now return to the analysis of part (b), and suppose we want to find a setting of (x_1^*, x_2^*, x_3^*) which the company will use for future industrial production. For reasons having to do with the specifics of the industrial process, the company decides to consider only solutions for which $x_2^* = x_3^*$ (x_1^* is unconstrained). Denoting the common value of x_2^* and x_3^* by x_2^*, find an estimate, standard error and 95% confidence intervals for x_2^* using the delta method. Also find a 95% confidence interval using Fieller's method, and compare the two results.

3. This question is about an analysis of variance experiment, reinterpreted as a linear regression. It does not assume detailed knowledge about analysis of variance.

   A recent paper discussed the following experiment related to the extraction of juice from blueberries. Three control variables were considered: temperature, level of sulfur dioxide (SO_2) and citric acid (coded as 0 or 1). Two response variables were measured: ACY (anthocynanin) and TP (total phenolics), both of which are considered to have beneficial health effects. The data were as follows:

Number	Temp (deg C)	SO2 (ppm)	Citric Acid	ACY	TP
1	50	0	0	27.5	55.9
2	50	0	1	42.6	62.6
3	80	0	0	50.2	71.4
4	80	0	1	62.4	88.8
5	50	50	0	92.2	307.3
6	50	50	1	96.5	316.4
7	80	50	0	97.5	420.6
8	80	50	1	102.2	413.8
9	50	100	0	90.6	386.0
10	50	100	1	82.2	337.5
11	80	100	0	92.1	641.0
12	80	100	1	91.4	684.3

   Consider the model

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \eta_{ik} + \zeta_{jk} + \epsilon_{ijk}, \tag{6}$$

   where $\alpha_i,\ i = 1, 2,\ \beta_j,\ j = 1, 2, 3,\ \gamma_k,\ k = 1, 2$ are main effects due to temperature, SO_2 and citric acid respectively, $\delta_{ij}, \eta_{ik}, \zeta_{jk}$ are interaction terms, and ϵ_{ijk} are independent $N(0, \sigma^2)$ errors. To make the model identifiable, assume any of $\alpha_i, \beta_j, \gamma_k, \delta_{ij}, \eta_{ik}, \zeta_{jk}$ is 0 when any of i, j, k is 1 (note that this is a different identifiability condition from the ones assumed in most of the examples of Chapter 8).

   (a) Write the model (6) in the form $Y = X\beta + \epsilon$, where Y is the vector of responses (of dimension 12), the vector β consists of all the non-zero unknown parameters, and X is a design matrix of zeros and ones. (You should find that X is 12×10.)

   (b) Using SAS's PROC REG or the "lm" command in R or S-PLUS, fit the model (6) to the data, where temperature, SO_2 and citric

acid are the three factor variables and ACY is the response. Also consider possible transformations of the response and indicate which you prefer. (For example, you should consider both the square root and the log transformation, and others in the Box-Cox family if you have time. It is not necessary to give detailed tables of parameter values, but state the value of the residual sum of squares or the estimated s, and any other statistics that are directly relevant to the question.)

(c) Now using whatever transformation you selected in (b), decide which of the main effects and interactions is significant. (Again, I don't want very detailed regression output, but indicate the main steps of your analysis and how you did them.)

(d) Repeat the steps of (b) and (c) for the TP response variable. (It's not necessary that the transformation of TP be the same as that for ACY.)

(e) Write a short report on your conclusions for the company. Recall that the company's objective is to choose *one* setting of the three control variables so that both ACY and TP are high. Your report should indicate which settings you recommend, but should also make clear to what extent the differences among different possible settings are statistically significant, and whether you would recommend further experimentation.

SOLUTIONS (MARKS FOR EACH PART IN BRACKETS)

1. (a) {8} $\widehat{y}^* = c^T\widehat{\beta}$ where $c^T = (\ 1 \quad x_1^* \quad x_2^* \quad x_1^{*2} \quad x_1^*x_2^* \quad x_2^{*2}\ )$. Then $E\{(\widehat{y}^* - y^*)^2\} = \sigma^2\{c^T(X^TX)^{-1}c + 1\}$, by equation (3.25) (p. 124) of the course text.

   For X, X^TX and $(X^TX)^{-1}$, refer to p. 358 of the course text. In particular,

$$(X^TX)^{-1} = \begin{pmatrix} \frac{5}{9} & 0 & 0 & -\frac{1}{3} & 0 & -\frac{1}{3} \\ 0 & \frac{1}{6} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{6} & 0 & 0 & 0 \\ -\frac{1}{3} & 0 & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{4} & 0 \\ -\frac{1}{3} & 0 & 0 & 0 & 0 & \frac{1}{2} \end{pmatrix}$$

   Hence with $x_1^* = \cos\theta,\ x_2^* = \sin\theta$,

$$\begin{aligned} c^T(X^TX)^{-1}c &= \sigma^2\left(\frac{5}{9} + \frac{1}{6}\cos^2\theta + \frac{1}{6}\sin^2\theta + \frac{1}{2}\cos^4\theta \right. \\ &\qquad \left. + \frac{1}{4}\cos^2\theta\sin^2\theta + \frac{1}{2}\sin^4\theta - \frac{2}{3}\cos^2\theta - \frac{2}{3}\sin^2\theta\right) \\ &= \sigma^2\left(\frac{5}{9} - \frac{3}{4}\cos^2\theta\sin^2\theta\right) \end{aligned}$$

   using $\cos^2\theta + \sin^2\theta = 1$ and $\frac{1}{2}\cos^4\theta + \cos^2\theta\sin^2\theta + \frac{1}{2}\sin^4\theta = \frac{1}{2}(\cos^2\theta + \sin^2\theta)^2 = \frac{1}{2}$.

   Hence

$$E\{(\widehat{y}^* - y^*)^2\} = \sigma^2\left(\frac{14}{9} - \frac{3}{4}\cos^2\theta\sin^2\theta\right). \qquad (7)$$

   But $\cos^2\theta\sin^2\theta$ lies between 0 and $\frac{1}{4}$ (the maximum is when $\cos^2\theta = \sin^2\theta = \frac{1}{2}$), so (7) lies between $\frac{197}{144}\sigma^2$ and $\frac{14}{9}\sigma^2$, as claimed.

   (b) {8} We may rewrite (1) in the form $Y = X_1\gamma + X_2\beta_{12} + \epsilon$ where $\gamma^T = (\ \beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_{11} \quad \beta_{22}\ )$, X_1 is the matrix X (as on page 358 of the course text) with the fifth column removed, and X_2 is the fifth column of X, in other words $X_2^T = (\ 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad -1 \quad 0 \quad -1 \quad 1\ )$. Note that because the fifth column of X is orthogonal to all the other columns, we have $X_1^TX_2 = 0$. We estimate $\widehat{\gamma} = (X_1^TX_1)^{-1}X_1^TY$. This has mean $E\{\widehat{\gamma}\} = (X_1^TX_1)^{-1}X_1^T(X_1\gamma + X_2\beta_{12}) = \gamma$ where we used the fact that $X_1^TX_2 = 0$. In other words, $\widehat{\gamma}$ is still an unbiased estimator of γ even under the wrong model when $\beta_{12} \neq 0$. Also the covariance matrix of $\widehat{\gamma}$ is $(X_1^TX_1)^{-1}\sigma^2$, as usual.

Write $c_1^T = (\ 1 \quad x_1^* \quad x_2^* \quad x_1^{*2} \quad x_2^{*2}\ )$, so that $y^* = c_1^T\gamma + x_1^*x_2^*\beta_{12} + \epsilon^*$, $\tilde{y}^* = c_1^T\widehat{\gamma}$, and hence

$$y^* - \tilde{y}^* = c_1^T(\gamma - \widehat{\gamma}) + x_1^*x_2^*\beta_{12} + \epsilon^*$$

has mean $x_1^*x_2^*\beta_{12}$ and variance $\sigma^2\{c_1^T(X_1^TX_1)^{-1}c_1 + 1\}$ as in part (a). To evaluate the latter expression, we note that $(X_1^TX_1)^{-1}$ is the same as $(X^TX)^{-1}$ except that the fifth column and fifth row are omitted (this follows again because X_1 and X_2 are orthogonal, or just repeat the computation in the text with $X_1^TX_1$ in place of X^TX), hence $\sigma^2\{c_1^T(X_1^TX_1)^{-1}c_1 + 1\}$ is the same as (7) but with a term $\frac{1}{4}\cos\theta\sin\theta$ subtracted.

Finally, since $E\{(\tilde{y}^* - y^*)^2\}$ is the sum of the squared bias and the variance, we have

$$E\{(\tilde{y}^* - y^*)^2\} = \cos^2\theta\sin^2\theta\beta_{12}^2 + \sigma^2\left(\frac{14}{9} - \cos^2\theta\sin^2\theta\right). \quad (8)$$

[Many students did not give a precise explanation of why $E\{y^* - \tilde{y}^*\} = \beta_{12}x_1^*x_2^*$. It is important that $E\{\widehat{\beta}_0\} = \beta_0, E\{\widehat{\beta}_1\} = \beta_1$, etc. This wouldn't be true without the orthogonality of X_1 and X_2.]

(c) {4} Comparing (7) and (8), $E\{(\tilde{y}^* - y^*)^2\} < E\{(\widehat{y}^* - y^*)^2\}$ if and only if

$$\cos^2\theta\sin^2\theta\beta_{12}^2 < \frac{\sigma^2}{4}\cos^2\theta\sin^2\theta \quad (9)$$

The inequality (9) reduces to $4\beta_{12}^2 < \sigma^2$ except when $\cos\theta = 0$ or $\sin\theta = 0$ — in these cases, $\widehat{y}^*$ and $\tilde{y}^*$ are identical, so there is no distinction between the two methods.

(d) {5} Under (1) the h_{ii} values for $i = 1, ..., 9$ are $\frac{5}{9}, \frac{5}{9}, \frac{5}{9}, \frac{5}{9}, \frac{29}{26}, \frac{29}{26}, \frac{5}{9}, \frac{29}{26}, \frac{29}{26}$ while under (2) all nine values are $\frac{5}{9}$ (direct matrix calculation, or create an artificial data set $y_1, ..., y_9$ and run the influence diagnostics in SAS or S-PLUS).

(e) {6} By equations (4.4) and (4.5) of the text, $y_i - \widehat{y}_{i(i)} = \frac{e_i}{1-h_{ii}}$. Under model (1), this has mean 0 and variance $\frac{\sigma^2}{1-h_{ii}}$. This has values $\frac{9}{4}\sigma^2, \frac{9}{4}\sigma^2, \frac{9}{4}\sigma^2, \frac{9}{4}\sigma^2, \frac{36}{7}\sigma^2, \frac{36}{7}\sigma^2, \frac{9}{4}\sigma^2, \frac{36}{7}\sigma^2, \frac{36}{7}\sigma^2$, respectively for $i = 1, ..., 9$.

(f) {6} Writing the model in the form $Y = X_1\gamma + X_2\beta_{12} + \epsilon$ as in part (b), we have $e = (I - H_1)Y$ where $H_1 = X_1(X_1^TX_1)^{-1}X_1^T$, so e has covariance matrix $(I - H_1)\sigma^2$ as usual, and mean $(I - H_1)(X_1\gamma + X_2\beta_{12}) = X_2\beta_{12}$ since $(I - H_1)X_1 = 0$ and $H_1X_2 = 0$ by orthogonality of X_1 and X_2. Hence e_i, the ith component of e, has variance $(1 - h_{ii})\sigma^2$ and mean $x_{i1}x_{i2}\beta_{12}$. It then follows that $\frac{e_i}{1-h_{ii}}$ has variance $\frac{\sigma^2}{1-h_{ii}}$

and mean $\frac{x_{i1}x_{i2}\beta_{12}}{1-h_{ii}}$, from which the result follows. The individual values are $\frac{9}{4}\sigma^2$ for $i = 1, 2, 3, 4, 7$ and $\frac{81}{16}\beta_{12}^2 + \frac{9}{4}\sigma^2$ for $i = 5, 6, 8, 9$.

[As in part (b), many answers missed the point about orthogonality.]

(g) {3} From the results of (e) and (f), the mean value of the PRESS statistic under (3) is smaller than the mean value under (1) if $\frac{81}{16}\beta_{12}^2 + \frac{9}{4}\sigma^2 < \frac{36}{7}\sigma^2$. This quickly reduces to $\frac{7}{4}\beta_{12}^2 < \sigma^2$.

There are various comments you could make about this. If $\frac{7}{4}\beta_{12}^2 < \sigma^2 < 4\beta_{12}^2$ then PRESS will tend to select model (1) even though the prediction criterion would favor (3) — thus, at least in this instance, PRESS seems to over-favor the larger model. As for an explanation of this, it appears that the deletion aspect of PRESS is not accurately representing the biases and variances of the desired prediction experiment.

2. This exercise was based on the paper:

Kong, Q., He, G., Chen, Q. and Chen, F. (2004), Optimization of medium composition for cultivating *Clostridium butyricum* with response surface methodology. *Journal of Food Science* **69**, No. 7, M163–M168.

though the analysis given below differs substantially from that in the paper!

(a) {7} In model (5), none of the variables $\beta_{12}, \beta_{13}, \beta_{23}$ is significant (all have p-values bigger than 0.5) so we prefer model (4), in which all parameters are significant except possibly β_1 (p-value .060). The coefficients (from SAS) are:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.22328	0.02578	47.44	<.0001
x1	1	-0.03531	0.01711	-2.06	0.0596
x2	1	0.06253	0.01711	3.66	0.0029
x3	1	0.05399	0.01711	3.16	0.0076
x11	1	-0.05359	0.01665	-3.22	0.0067
x22	1	-0.06296	0.01665	-3.78	0.0023
x33	1	-0.06561	0.01665	-3.94	0.0017

The covariances of $(\widehat{\beta}_1, \widehat{\beta}_{11})$, $(\widehat{\beta}_2, \widehat{\beta}_{22})$, $(\widehat{\beta}_3, \widehat{\beta}_{33})$, are each 0.

[Some students did a formal F-test of model (4) against (5). This results in $F = 0.17$: not significant.]

(b) {3} $x_j^* = -\frac{\beta_j}{2\beta_{jj}}$, $j = 1, 2, 3$ so substituting the point estimates from (a), we find $\widehat{x}_1^* = -.3294$, $\widehat{x}_2^* = .4966$, $\widehat{x}_3^* = .4114$.

(c) {7} There is no problem with multicollinearity etc., but the influence diagnostics show that there are two large outliers in rows 9 and 14, each of which is highly significant (externally studentized values of 3.46 and 3.79; DFFITS of 4.30 and 4.71; $2\sqrt{\frac{p}{n}} = 1.18$). Omitting these values and rerunning the regression, we get results

Intercept	1	1.22089	0.01146	106.57	<.0001
x1	1	-0.01047	0.00950	-1.10	0.2940
x2	1	0.06253	0.00760	8.23	<.0001
x3	1	0.02722	0.00950	2.86	0.0154
x11	1	-0.07361	0.01025	-7.18	<.0001
x22	1	-0.04732	0.00769	-6.15	<.0001
x33	1	-0.08840	0.01025	-8.62	<.0001

with point estimates of $x_1^* = -.0711,\ x_2^* = .661,\ x_3^* = .154$, substantially different from the above.

(d) {13} Fixing $x_2 = x_3$, the response surface includes terms $(\beta_2+\beta_3)x_2+(\beta_{22}+\beta_{33})x_2^2$ which is maximized by $x_2^* = -(\beta_2+\beta_3)/(2(\beta_{22}+\beta_{33}))$. Writing $\theta_1 = \beta_2 + \beta_3,\ \theta_2 = \beta_{22} + \beta_{33}$, the quantity of interest is $g(\theta_1, \theta_2) = -\frac{\theta_1}{2\theta_2}$ for which we also have $g_1 = \frac{\partial g}{\partial \theta_1} = -\frac{1}{2\theta_2},\ g_2 = \frac{\partial g}{\partial \theta_2} = \frac{\theta_1}{2\theta_2^2}$.

The covariance matrix of $\widehat{\beta}$ shows among other things that (i) $\widehat{\beta}_2$ and $\widehat{\beta}_3$ each has estimated variance .0002926338, and they are independent; (ii) the estimated variances of $\widehat{\beta}_{22}$ and $\widehat{\beta}_{33}$ are each .0002773153 and their estimated covariance is .0000275354; (iii) $\widehat{\beta}_2$ and $\widehat{\beta}_3$ are independent of $\widehat{\beta}_{22}$ and $\widehat{\beta}_{33}$. By (i), the standard error of $\widehat{\theta}_1$ is $\sqrt{2 \times .00029262338} = .0242$. By (ii), the standard error of $\widehat{\theta}_2$ is $\sqrt{2 \times (.0002773153 + .0000275354)} = .0247$, and by (iii), $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are independent.

Substituting the estimates for $\beta_2, \beta_3, \beta_{22}, \beta_{33}$, the estimated values of $\theta_1, \theta_2, x_2^*, g_1$ and g_2 are .11652, –.12857, .45314, 3.89 and 3.52. By the delta method, the estimated standard error of $\widehat{x}_2^*$ is $\sqrt{3.89^2 \times .0242^2 + 3.52^2 \times .0247^2} = .1281$. The RSS has 13 degrees of freedom and $t_{13;.975} = 2.16$, so an approximate 95% confidence interval is $.45314 \pm 2.16 \times .1281 = (.1764, .7298)$.

Under Fieller's method, we have to evaluate $Q(x) = (\widehat{\theta}_1 - \widehat{\theta}_2 x)^2 - t^{*2}s^2(a - 2cx + bx^2)$. Here $as^2,\ bs^2,\ cs^2$ are respectively the estimated variances of $\widehat{\theta}_1$ and $\widehat{\theta}_2$, and the estimated covariance. We also assume $t^* = 2.16$, as in the previous paragraph. Therefore, $Q(x) = (.11652 - .12857x)^2 - 2.16^2(.0242^2 + .0247^2x^2) = .01085 - .02996x + .01368x^2$. This quadratic equation has roots at –1.732 and –.4576. This is

an exact 95% confidence interval for $\frac{\theta_1}{\theta_2}$; therefore, an exact 95% confidence interval for $x_2^* = -\frac{\theta_1}{2\theta_2}$ is $(.2288, .8660)$.

[I made a slight correction to the Fieller interval on the first version of these solutions. Many students made numerical slips with both the delta and Fieller methods, but I tried not to penalize these too much if the basic method was right.]

The two confidence intervals (delta method and Fieller) overlap substantially, but they are not identical, with Fieller giving a more right-skewed interval.

3. This exercise was based on the paper:

   Lee, J. and Wrolstad, R.E. (2004), Extraction of anthocyanins and polyphenolics from blueberry-processing waste. *Journal of Food Science* **69**, No. 7, C564–C573.

   though the analysis given below differs substantially from that in the paper!

   (a) {6} We write

$$Y = \begin{pmatrix} y_{111} \\ y_{112} \\ y_{211} \\ y_{212} \\ y_{121} \\ y_{122} \\ y_{221} \\ y_{222} \\ y_{131} \\ y_{132} \\ y_{231} \\ y_{232} \end{pmatrix}, \beta = \begin{pmatrix} \mu \\ \alpha_2 \\ \beta_2 \\ \beta_3 \\ \gamma_2 \\ \delta_{22} \\ \delta_{23} \\ \eta_{22} \\ \zeta_{22} \\ \zeta_{32} \end{pmatrix}, X = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 \end{pmatrix}.$$

   (b) {4} With the y variable representing ACY, $\dot{y} = 72.29848$ so we consider transformations $h(y_i) = y_i, 2\sqrt{\dot{y}y_i}, \dot{y}\log(y_i)$ with respective residual sums of squares (RSS) of 14.71, 28.33, 66.58. Among these three, the favored transformation is the identity. (If you extend to $h_\lambda(y_i) = \dot{y}^{1-\lambda}\frac{y_i^\lambda - 1}{\lambda}$ it turns out the preferred value is about $\lambda = 1.9$, with RSS=9.9, but this is not a statistically significant reduction, compared with $\lambda = 1$.)

   [It seems that not everyone understands the need to rescale after a transformation! Some students just compared with raw RSSs without

rescaling, with results in favor of the log transformation, not correct in this instance. However there are other statistics you could look at, e.g. the adjusted or unadjusted R^2 statistic, or the F statistic for the overall model fit. All three of these — which are really different ways of measuring the same thing — point towards the model with ACY as the response as being superior to the ones based on logs or square roots.]

(c) {5} An initial table of results shows α_2, β_2, β_3 highly significant, ζ_{32} significant with $p = .042$ and γ_2, δ_{22}, δ_{23} having p-values just over .05. There are various intermediate tests you can do (credit for careful searching!) but if we fit the model containing just α_2, β_2, β_3, we have $RSS = 411.205$ with 8 DF, compared with $RSS = 14.71167$ with 2 DF for the full model. The F statistic is

$$\frac{411.205 - 14.71167}{6} \cdot \frac{2}{14.71167} = 8.984$$

which has p-value 0.104 against the $F_{6,2}$ distribution. In other words, to the extent that we can determine based on this experiment, only the main effects due to temperature and SO_2 are significant. Moreover, the difference between β_2 and β_3 is not significant (though $\widehat{\beta}_3 < \widehat{\beta}_2$).

[There are a number of other solutions which seem acceptable. For instance, if you use backward elimination, removing the η_{22} and ζ_{22} terms, it seems all the rest are significant at the .05 level, in contradiction to the above! Also, criteria such as C_p, AIC and even BIC seems to favor larger models. So if you said that I was willing to give full credit for it, though I still wanted to see some comparison of different models, not just a single model.]

(d) {9} With TP as the response, we have $\dot{y} = 229.5938$ and the RSS values for $h(y_i) = y_i, 2\sqrt{\dot{y}y_i}, \dot{y}\log(y_i)$ are 1574, 756, 403 suggesting the log transform as the best. (A Box-Cox transformation suggests $\lambda \approx -0.6$, with RSS about 295, but this would be an unusual transformation to adopt in practice.) Henceforth we use a log transformation, without the scaling by $\dot{y}$.

The full model has $RSS = .00764$ with DF=2. Initial analysis suggests γ_2, η_{22}, ζ_{22}, ζ_{32} could all be dropped, and if we refit the model without these parameters we get $RSS = .08356$ with DF=6. The $F_{6,2}$ statistic is 3.31; the p-value associated with that is 0.25. Thus we accept the second model as correct. (In this model we could also drop δ_{22}, but it is not conventional to drop one component of an interaction term without the other.)

[Here again, some answers missed the point about scaling of transformations. Also, those who applied forward/backward elimination often got different answers from the above, though in this case, I am less sure these are acceptable alternatives. The defaults in SAS use rather liberal criteria for retaining variables — it doesn't really make sense to retain variables with P-values over 0.1, though the small sample size makes interpretation of all these statistics problematic.]

(e) {6} For ACY, it seems we should take temperature at 80 and SO_2 at either 50 or 100. For TP, the main effects together with the strongly significant positive value of δ_{23} suggests that the optimal combination is temperature 80, SO_2=100. However, the fact that the experiment was small and there are several other effects that could be tested (e.g. should SO_2 have been tried at some levels other than 50 or 100?) suggests that there is scope for further experimentation.

Most answers to this question did not do an adequate job of addressing whether other settings of the control variables are significantly worse than the ones recommended. Here are two analytical approaches that could be used for this question.

One of them is to use SAS's PROC REG, with the "clm" option to produce confidence limits for the mean responses at each of the 12 possible settings. (Not "cli" in this instance, because the objective is to study the long-run output of the system, not the result of a single experiment.) When this is done for ACY as the response, the confidence intervals for TEMP=80 and SO2=50 or 100 nearly all overlap, whichever regression model was used, implying that the differences among these settings are not significant. However running the regression for log TP as the response, using the model recommended in (d), gives the confidence intervals:

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	4.0236	4.0817	0.0474	3.9696	4.1937	-0.0581
2	4.1368	4.0817	0.0474	3.9696	4.1937	0.0551
3	4.2683	4.3758	0.0474	4.2638	4.4879	-0.1075
4	4.4864	4.3758	0.0474	4.2638	4.4879	0.1106
5	5.7278	5.7409	0.0474	5.6288	5.8530	-0.0131
6	5.7570	5.7409	0.0474	5.6288	5.8530	0.0161
7	6.0417	6.0350	0.0474	5.9230	6.1471	0.006635
8	6.0254	6.0350	0.0474	5.9230	6.1471	-0.009665
9	5.9558	5.8887	0.0547	5.7593	6.0181	0.0671

```
 10     5.8216     5.8887     0.0547     5.7593     6.0181    -0.0671
 11     6.4630     6.4957     0.0547     6.3663     6.6251    -0.0327
 12     6.5284     6.4957     0.0547     6.3663     6.6251     0.0327
```

We can see that observations 11 or 12 (implying TEMP=80, SO2=100) have not only a higher predicted value than all the others, but also a confidence interval that is disjoint from the confidence intervals for all other settings, implying that the differences are truly significant.

The second method is based on an adaptation of Tukey's studentized range procedure, discussed in the very last class. Consider the following SAS code:

```
options ls=77 ps=58 nonumber label;
data rs1;
input x1 y1 y2;
ly2=log(y2);
datalines;
1  27.5 55.9
1  42.6 62.6
2  50.2 71.4
2  62.4 88.8
3  92.2 307.3
3  96.5 316.4
4  97.5 420.6
4  102.2 413.8
5  90.6 386.0
5  82.2 337.5
6  92.1 641.0
6  91.4 684.3
;
run;
;
proc anova;
class x1;
model y1=x1;
means x1 /tukey;
run;
;
proc anova;
class x1;
model ly2=x1;
means x1 /tukey;
run;
;
```

This combines temperature and SO_2 into a single factor variable with six levels, and ignores the possible effect of citric acid. The results show that for ACY (variable $y1$ in the SAS code), the four best values of $x1$ (4,3,6,5 in that order) are statistically indistinguishable. However, with log of TP ($ly2$) as the response, $x1 = 6$ is the best, superior to every other level according to the Tukey test. Combining the two results, we should take temperature at 80, SO_2 at 100.