Stat 511 Exam 1

February 24, 2003 Prof. Vardeman

1. In the calibration of a scientific instrument, "true" values x are known and produce experimental readings y on the instrument. Suppose that we are willing to assume that the mean value of y is proportional to x, so that

$$y_i = x_i\beta + \varepsilon_i$$

where for $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$, $\mathrm{E}\boldsymbol{\varepsilon} = \mathbf{0}$. A particular calibration experiment produces $n = 4$ data points as per

x	3	4	5	6
y	3	6	11	14

$$Y = \begin{pmatrix} 3 \\ 6 \\ 11 \\ 14 \end{pmatrix} \qquad X = \begin{pmatrix} 3 \\ 4 \\ 5 \\ 6 \end{pmatrix}$$

Initially suppose that $\mathrm{Var}\,\boldsymbol{\varepsilon} = \sigma^2\mathbf{I}$.

9 pts

a) Find a matrix $\mathbf{P_X}$ so that $\hat{\mathbf{Y}} = \mathbf{P_X Y}$.

$$P_X = X(X'X)^{-1}X' = \begin{pmatrix} 3 \\ 4 \\ 5 \\ 6 \end{pmatrix} \frac{1}{9+16+25+36} \begin{pmatrix} 3 & 4 & 5 & 6 \end{pmatrix}$$

$$= \frac{1}{86}\begin{pmatrix} 9 & 12 & 15 & 18 \\ 12 & 16 & 20 & 24 \\ 15 & 20 & 25 & 30 \\ 18 & 24 & 30 & 36 \end{pmatrix}$$

5 pts

b) By the criterion of "size of the hats, h_{ii}" which of the 4 observations is "most influential" in the fitting of the linear model here?

The h_{ii} are the diagonal elements of P_X, the largest of which is the last, which corresponds to the 4th observation $(x_4, y_4) = (6, 14)$

9 pts

c) Give 90% two-sided confidence limits for σ in the normal version of this model. (No need to simplify.)

$$\hat{Y} = X\left((X'X)^{-1}X'Y\right) = \begin{pmatrix} 3 \\ 4 \\ 5 \\ 6 \end{pmatrix} \frac{1}{86}\begin{pmatrix} 3 & 4 & 5 & 6 \end{pmatrix}\begin{pmatrix} 3 \\ 6 \\ 11 \\ 14 \end{pmatrix} = \begin{pmatrix} 3 \\ 4 \\ 5 \\ 6 \end{pmatrix} \frac{9+24+55+84}{86}$$

$$= 2\begin{pmatrix} 3 \\ 4 \\ 5 \\ 6 \end{pmatrix} = \begin{pmatrix} 6 \\ 8 \\ 10 \\ 12 \end{pmatrix}$$

So $SSE = (3-6)^2 + (6-8)^2 + (11-10)^2 + (14-12)^2 = 18$ and $MSE = \frac{18}{3}$

So 90% limits are $\left(\sqrt{\frac{18}{7.815}}, \sqrt{\frac{18}{.352}}\right)$

upper 5% pt of χ^2_3

lower 5% pt of χ^2_3

9pts d) Give 90% two-sided prediction limits for a new y for $x = 10$. (No need to simplify.)

This is prediction of y^* with $Ey^* = 10\beta$ and $\text{Var}\, y^* = 1\sigma^2$ - So limits are

$$10\, b_{OLS} \pm t\sqrt{MSE}\,\sqrt{1 + 10\left(\tfrac{1}{86}\right)10}$$

$$10(2) \pm 2.353\sqrt{6}\,\sqrt{1 + \tfrac{100}{86}}$$

Now suppose that it is plausible that not only is the mean value of y is proportional to x, but that so too is the standard deviation of y. That is, suppose that $\text{Var}\,\boldsymbol{\varepsilon} = \sigma^2\,\text{diag}(9,16,25,36)$.

5pts e) Give a matrix **T** such that **TY** follows a Gauss-Markov model. What is the model matrix for **TY** ?

T:

$$T = \left(\text{diag}(9,16,25,36)\right)^{-\frac{1}{2}} = \text{diag}\left(\tfrac{1}{3}, \tfrac{1}{4}, \tfrac{1}{5}, \tfrac{1}{6}\right)$$

Model Matrix: $W = TX$

$$= \text{diag}\left(\tfrac{1}{3}, \tfrac{1}{4}, \tfrac{1}{5}, \tfrac{1}{6}\right)\begin{pmatrix}3\\4\\5\\6\end{pmatrix} = \begin{pmatrix}1\\1\\1\\1\end{pmatrix}$$

5pts f) Evaluate an appropriate point estimate of β under these model assumptions.

$$\hat{\beta} = (W'W)^{-1}W'U = \tfrac{1}{4}(1\ 1\ 1\ 1)\,\text{diag}\left(\tfrac{1}{3}, \tfrac{1}{4}, \tfrac{1}{5}, \tfrac{1}{6}\right)\begin{pmatrix}3\\6\\11\\14\end{pmatrix}$$

$(U = TY)$

$$= \tfrac{1}{4}\left(\tfrac{3}{3} + \tfrac{6}{4} + \tfrac{11}{5} + \tfrac{14}{6}\right)$$

$$= \tfrac{422}{240}$$

5pts g) Give a standard error (an estimated standard deviation) for your estimate of β in part f) under these heteroscedastic model assumptions.

$$\text{Var}\, b_{OLS(U)} = \sigma^2 (W'W)^{-1} = \frac{\sigma^2}{4} \quad \text{So} \quad \widehat{\text{Var}\, b_{OLS(U)}} = \tfrac{1}{2}\sqrt{MSE_U}$$

$$MSE_U = \tfrac{1}{3}\left(\left(\tfrac{3}{3} - \tfrac{422}{240}\right)^2 + \left(\tfrac{6}{4} - \tfrac{422}{240}\right)^2 + \left(\tfrac{11}{5} - \tfrac{422}{240}\right)^2 + \left(\tfrac{14}{6} - \tfrac{422}{240}\right)^2\right)$$

$$= .3892$$

$$\text{So} \quad \widehat{\text{Var}\, b_{OLS(U)}} = \tfrac{1}{2}\sqrt{.3892} \doteq .3119$$

2. So-called "mixture experiments" are run to investigate how the composition of a substance (as measured by fractions of it that are of "pure component" types $i = 1, 2, \ldots, r$) affect some physical property y. For example, y might be an octane rating for a gasoline blended from r "pure" components like butane, alkylate, cat cracked, etc. Notice that in a mixture study

$$x_1 + x_2 + \cdots + x_r = 1$$

In this problem, we consider an $r = 4$ component mixture problem. Consider the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

for

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & .5 & .5 & 0 & 0 \\ 1 & .5 & 0 & .5 & 0 \\ 1 & .5 & 0 & 0 & .5 \\ 1 & 0 & .5 & .5 & 0 \\ 1 & 0 & .5 & 0 & .5 \\ 1 & 0 & 0 & .5 & .5 \\ 1 & .3\overline{3} & .3\overline{3} & .3\overline{3} & 0 \\ 1 & .3\overline{3} & .3\overline{3} & 0 & .3\overline{3} \\ 1 & .3\overline{3} & 0 & .3\overline{3} & .3\overline{3} \\ 1 & 0 & .3\overline{3} & .3\overline{3} & .3\overline{3} \\ 1 & .25 & .25 & .25 & .25 \end{pmatrix} = (\mathbf{1} \,|\, \mathbf{x}_1 \,|\, \mathbf{x}_2 \,|\, \mathbf{x}_3 \,|\, \mathbf{x}_4) \text{ and } \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

5pts a) For an arbitrary composition vector (x_1, x_2, x_3, x_4) (with each $x_i \geq 0$ and $\sum x_i = 1$) argue carefully that the corresponding mean response $\beta_0 + \sum \beta_i x_i$ is estimable.

Note that with $\underset{\sim}{v}_i$ the ith row of X and (x_1, x_2, x_3, x_4) a mixture vector

$$\beta_0 + \sum_i \beta_i x_i = \left(x_1 \underset{\sim}{V}_1 + x_2 \underset{\sim}{V}_2 + x_3 \underset{\sim}{V}_3 + x_4 \underset{\sim}{V}_4 \right) \underset{\sim}{\beta}$$

i.e. this is $\underset{\sim}{c}'\beta$ for $\underset{\sim}{c}'$ a linear combination of (the first 4) rows of X.

5pts b) The parameter β_0 is not estimable in this model. Argue this point carefully.

Rank(X) is both the row rank and the column rank. (So rank(X) is at most 5.) The fact that the 1st column is the sum of the last 4 means that X is not of full rank. Now looking at the first 4 rows of X we see that all of $\beta_0+\beta_1$, $\beta_0+\beta_2$, $\beta_0+\beta_3$ and $\beta_0+\beta_4$ are estimable. If β_0 were estimable, then all of β_0, $\beta_1 = (\beta_0+\beta_1) - \beta_0$, $\beta_2, \beta_3, \beta_4$ would be estimable and X would be of full rank.

Now consider a full rank restricted version of the original mixture model of the form

$$\mathbf{Y} = \mathbf{X}^*\boldsymbol{\gamma} = (\mathbf{x}_1 \mid \mathbf{x}_2 \mid \mathbf{x}_3 \mid \mathbf{x}_4)\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \text{ for } \boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{pmatrix}$$

6pts c) In this restricted model and the mixture context, what is the interpretation of the parameter γ_1? What is $\gamma_1 - \gamma_2$?

γ_1: This is the mean response for a pure component of Type 1.

$\gamma_1 - \gamma_2$: This is the difference in mean responses for Type 1 and Type 2 pure components

(Note the fact that $\Sigma x_i = 1$ makes the usual regression interpretation of "rate of change of mean y wrt x_1 holding the other x's fixed" impossible)

7pts d) Give a matrix **C** and a vector **d** so that the hypothesis that "pure component #1 has mean response 3 and simultaneously an 'equal parts mixture of components' has mean response 12" in the form $\mathrm{H}_0 : \mathbf{C}\boldsymbol{\gamma} = \mathbf{d}$.

C: $\begin{pmatrix} 1 & 0 & 0 & 0 \\ .25 & .25 & .25 & .25 \end{pmatrix}$

d: $\begin{pmatrix} 3 \\ 12 \end{pmatrix}$

6pts e) Is the hypothesis in d) testable? Explain.

Yes it is. Both rows of C are rows of X^* so both β_1 and $\frac{1}{4}(\beta_1+\beta_2+\beta_3+\beta_4)$ are estimable. The 2nd row of C is not a multiple of the 1st, so $\mathrm{Rank}(C) > 1$ i.e. C is of full rank (2).

There is some R output attached to this exam. The first part of it concerns this mixture problem. Use it to help you answer the following questions.

6pts f) For which of the (x_1, x_2, x_3, x_4) mixtures in the data set is the mean of y most precisely estimated? Say why your answer agrees with intuition.

From the printout, the smallest diagonal entry of H is the last one. Since $\mathrm{Var}\,\hat{Y} = \sigma^2 H$, it is then the last (x_1, x_2, x_3, x_4), namely $(.25, .25, .25, .25)$ that has the most precisely estimated mean response. This is a set of conditions at the "center" of the experimental region, where one should expect to be best informed.

9pts g) Give 90% two sided confidence limits for $\gamma_1 - \gamma_2$. (Plug in, but you need not simplify.)

We want $\underset{\sim}{c}'\underset{\sim}{b}_{OLS} \pm t\sqrt{MSE}\sqrt{\underset{\sim}{c}'(X^{*\prime}X^*)^{-1}\underset{\sim}{c}}$ for $\underset{\sim}{c}' = (1,-1,0,0)$

This is $(10.81 - 11.87) \pm 1.796\sqrt{\frac{16.48}{11}}\sqrt{(1,-1)\begin{pmatrix} .532 & -.009 \\ -.009 & .532 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix}}$

$n - \text{rank}(X) = 15 - 4 = 11$

t_{11} upper 5% pt

9pts h) Notice in this model that if $H_0\!:\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4$ is true, then the fact that in the mixture context $\sum x_i = 1$ implies that $E\,Y = \gamma\mathbf{1}$ for some γ, that is, the mean response is constant. Give the value of and degrees of freedom for an F statistic for testing this hypothesis.

We want
$$F = \frac{SS_{H_0}/l}{SSE/n-\text{rank}(X^*)}$$

$$= \frac{116.49/3}{16.48/11} = 25.9$$

$F =$ 25.9 $df =$ 3 , 11