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

\[
y = x \beta + \varepsilon,
\]

where for \( \varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)' \) and \( \text{E} \varepsilon = 0 \). A particular calibration experiment produces \( n = 4 \) data points as per the following table:

<table>
<thead>
<tr>
<th>( x )</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>( y )</td>
<td>3</td>
<td>6</td>
<td>11</td>
<td>14</td>
</tr>
</tbody>
</table>

Initially suppose that \( \text{Var} \varepsilon = \sigma^2 I \).

\[
Y = \begin{pmatrix} 3 \\ 6 \\ 11 \\ 14 \end{pmatrix}, \quad X = \begin{pmatrix} 3 \\ 4 \\ 5 \\ 6 \end{pmatrix}
\]

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

\[
P_X = X (X'X)^{-1} X'
\]

where

\[
(X'X)^{-1} = \frac{1}{36 + 25 + 36} \begin{pmatrix} 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 6 \end{pmatrix}
\]

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

b) By the criterion of "size of the hats, \( h_a \)" which of the 4 observations is "most influential" in the fitting of the linear model here?

The \( h_a \) are the diagonal elements of \( P_X \), the largest of which is the last, which corresponds to the 4th observation \((x_4, y_4) = (16, 14)\).

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

\[
\hat{Y} = X (X'X)^{-1} X' Y
\]

\[
= \begin{pmatrix} 3 \\ 4 \\ 5 \\ 6 \end{pmatrix} \frac{1}{86} \begin{pmatrix} 3 & 4 & 5 & 6 \\ 4 & 5 & 6 & 6 \end{pmatrix} \begin{pmatrix} 3 \\ 4 \\ 5 \\ 6 \end{pmatrix}
\]

\[
= \frac{3 + 24 + 55 + 84}{86} \frac{1}{86} \begin{pmatrix} 3 \\ 4 \\ 5 \\ 6 \end{pmatrix}
\]

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

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

upper 5% pt of \( \chi^2_3 \)

lower 5% pt of \( \chi^2_3 \)

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

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

upper 5% pt of \( \chi^2_3 \)

lower 5% pt of \( \chi^2_3 \)

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

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

upper 5% pt of \( \chi^2_3 \)

lower 5% pt of \( \chi^2_3 \)
d) Give 90% two-sided prediction limits for a new $y$ for $x = 10$. (No need to simplify.)

This is prediction of $y^*$ with $E y^* = 10 \beta$ and $\text{Var} y^* = 10^2 - S_e^2$

limits are

$$10 \text{bols} \pm t \sqrt{\text{MSE}} \sqrt{1 + 10 \left(\frac{s_e}{10}\right)^2}$$

$$10/2 \pm 2.353 \sqrt{6} \sqrt{1 + \frac{100}{86}}$$

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 $\text{Var} \epsilon = \sigma^2 \text{diag}(9, 16, 25, 36)$.

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

$$T: \quad \text{Model Matrix: } \quad W = TX$$

$$T = (\text{diag}(9, 16, 25, 36))^{-1}$$

$$= \text{diag}(\frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6})$$

$$= \begin{pmatrix} 3 \\ 4 \\ 5 \\ 6 \end{pmatrix}$$

f) Evaluate an appropriate point estimate of $\beta$ under these model assumptions.

$$\hat{\beta} = (W'W)^{-1}W'U = \frac{1}{4} \begin{pmatrix} 1 & 1 & 1 & 1 \end{pmatrix} \text{diag}(\frac{3}{3}, \frac{4}{4}, \frac{5}{5}, \frac{6}{6}) \begin{pmatrix} 3 \\ 4 \\ 5 \\ 6 \end{pmatrix}$$

$$U = TY$$

$$= \frac{1}{4} \left( \frac{3}{3} + \frac{4}{4} + \frac{5}{5} + \frac{6}{6} \right)$$

$$= \frac{422}{240}$$

g) Give a standard error (an estimated standard deviation) for your estimate of $\beta$ in part f) under these heteroscedastic model assumptions.

$$\text{Var } b_{\text{ols}}(u) = \frac{\sigma^2}{4} \quad \text{So } \quad \overline{\text{Var } b_{\text{ols}}(u)} = \frac{1}{2} \sqrt{\text{MSE}_U}$$

$$\text{MSE}_U = \frac{1}{3} \left( \left( \frac{3}{3} - \frac{422}{240} \right)^2 + \left( \frac{4}{4} - \frac{422}{240} \right)^2 + \left( \frac{5}{5} - \frac{422}{240} \right)^2 + \left( \frac{6}{6} - \frac{422}{240} \right)^2 \right)$$

$$= 0.3892$$

So $$\overline{\text{Var } b_{\text{ols}}(u)} = \frac{1}{2} \sqrt{0.3892} = 0.3119$$
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
\[
Y = X\beta + \varepsilon
\]
for
\[
X = \begin{pmatrix}
1 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 0 & 1 \\
1 & .5 & .5 & 0 & 0 \\
1 & .5 & 0 & .5 & 0 \\
1 & .5 & 0 & 0 & .5 \\
1 & .5 & 0 & .5 & 0 \\
1 & .33 & .33 & 0 & .33 \\
1 & .33 & .33 & 0 & .33 \\
1 & .33 & 0 & .33 & .33 \\
1 & 0 & .33 & .33 & .33 \\
1 & .25 & .25 & .25 & .25 \\
\end{pmatrix}
\]

**Exercise**

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 \( y_i \) the \( i \)th row of \( X \) and \((x_1, x_2, x_3, x_4)\) a mixture vector
\[
\beta_0 + \sum \beta_i x_i = (x_1y_1 + x_2y_2 + x_3y_3 + x_4y_4) \beta
\]
i.e. This is \( e'\beta \) for \( e \) a linear combination of (the first 4) rows of \( X \).

b) The parameter \( \beta_0 \) is not estimable in this model. Argue this point carefully.

\( \text{Rank}(X) \) is both the row rank and the column rank. (So \( \text{rank}(X) \) is at most 5.) The fact that the last 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 \( \beta_0 \) were estimable, then all of \( \beta_0, \beta_1 = (\beta_0 + \beta_1) - \beta_0, \beta_2, \beta_3, \beta_4 \) 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

\[ Y = X'\gamma = (x_1 | x_2 | x_3 | x_4)\gamma + \varepsilon \text{ for } \gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{pmatrix} \]

\[ x_1: \text{This is the mean response for a pure component of Type 1.} \]

\[ x_2-x_1: \text{This is the difference in mean responses for Type 1 and Type 2 pure components.} \]

\[ \text{Note the fact that } \Sigma x_i = 1 \text{ makes the usual regression interpretation of the rate of change of mean, } \frac{\partial \mu}{\partial x}, \text{ holding the other } x_i \text{'s fixed, impossible.} \]

\[ d) \text{ Give a matrix } C \text{ and a vector } d \text{ so that the hypothesis that “pure component #1 has mean response } 3 \text{ and simultaneously an “equal parts mixture of components” has mean response } 12 \text{” in the form } H_0 : C\gamma = d. \]

\[ C : \begin{pmatrix} 1 & 0 & 0 & 0 \\ .25 & .25 & .25 & .25 \end{pmatrix} \quad d : \begin{pmatrix} 3 \\ 12 \end{pmatrix} \]

\[ e) \text{ Is the hypothesis in d) testable? Explain.} \]

Yes it is. Both rows of } C } \text{ are rows of } X^* \text{ so both } \gamma_1 \text{ and } \gamma = \frac{1}{4}(\beta_1 + \beta_2 + \beta_3 + \beta_4) \text{ are estimable. The 2nd row of } C \text{ is not a multiple of the 1st, so Rank}(C) > 1 \text{ i.e. } C \text{ 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.

\[ f) \text{ For which of the } (x_1, x_2, x_3, x_4) \text{ mixtures in the data set is the mean of } y \text{ most precisely estimated? Say why your answer agrees with intuition.} \]

From the printout, the smallest diagonal entry of } H \text{ is the last one. Since } \text{Var}(\hat{\gamma}) = \sigma^2 H, \text{ it is then the last } (x_1, x_2, x_3, x_4), \text{ namely } (.25, .25, .25, .25) \text{ 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.
g) Give 90% two sided confidence limits for $\gamma_1 - \gamma_2$. (Plug in, but you need not simplify.)

We want

$$a_{1.05} \frac{Q}{b_{10}} \pm t_{1.05} \sqrt{\frac{16.48}{11}} \frac{1}{\sqrt{(1,-1)(.532,-.009)(-1)}}$$

Thus is

$$(10.31 - 11.87) \pm 1.796 \sqrt{\frac{16.48}{11}} \sqrt{(1,-1)(.532,-.009)(-1)}$$

$n-$rank$(x) = 15-4 = 11$ \hspace{1cm} t_{11} \text{ upper 5% pt}$

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 $E Y = \gamma 1$ for some $\gamma$, 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

$$F = \frac{SSH_0 / \chi}{SSE / n-$rank$(X)}$$

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

$F = 25.9$ \hspace{1cm} $df = 3, 11$