At the age of six months, most infants begin to eat supplementary semi-solid foods, such as homogenized infant foods (or **beikosts**). At this stage, micronutrient deficiencies can severely limit the physical and intellectual capacity of children. Dietary antinutrients, such as phytic acid (a major component of all plant seeds), can reduce the bioavailability of certain micronutrients, such as zinc, iron, and copper. Thus, research should be addressed not only at ensuring an adequate micronutrient content in beikosts, but also at limiting antinutritive content of vegetable origin. In what follows, you will be presented with data from such a study performed by a food company, where the interest lies in relating the concentration of phytic acid to four key beikost ingredients, in an effort to determine the best formulation of different ingredients to enhance micronutrient bioavailability by ensuring the lowest antinutritive factor when infant foods are manufactured.

The first part of this problem will ask you to develop/prove a number of results which will be useful in answering the questions in the second part of the problem, involving the data from the study described above.

Part I

Consider a linear model written in the form

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \cdots + \boldsymbol{X}_m\boldsymbol{\beta}_m + \boldsymbol{\epsilon}, \tag{1}$$

where

- $\boldsymbol{y}$ is an $n \times 1$ response vector;
- $\boldsymbol{X}_i$ is an $n \times p_i$ known matrix of constants, for $1 \leq i \leq m$;
- $\boldsymbol{\beta}_i$ is a $p_i \times 1$ vector of unknown regression coefficients (fixed parameters), for $1 \leq i \leq m$;
- $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of random errors that has a multivariate normal distribution with mean vector $\mathbf{0}_n$ and covariance matrix $\sigma^2\boldsymbol{I}_n$, where $\mathbf{0}_n$ is an $n \times 1$ vector of 0's and $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

In addition, assume that

- the total number of regression coefficients, denoted by $p = \sum_{i=1}^{m} p_i$, is less than n (i.e., $p < n$),
- $\boldsymbol{X}_i^T\boldsymbol{X}_i$ is nonsingular for each $i = 1, \ldots, m$, and
- matrices $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ are mutually orthogonal, i.e., for each $1 \leq i \leq m$ and $1 \leq j \leq m$, we have
$$\boldsymbol{X}_i^T\boldsymbol{X}_j = \mathbf{0} \text{ whenever } i \neq j\,. \tag{2}$$

Finally, let $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^T, \widehat{\boldsymbol{\beta}}_2^T, \ldots, \widehat{\boldsymbol{\beta}}_m^T)^T$ be the least squares estimator of $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \ldots, \boldsymbol{\beta}_m^T)^T$ under the full model (1).

1. Show that, for all $i = 1, \ldots, m$, the least squares estimator of $\boldsymbol{\beta}_i$ under the full model (1) is the same as that under the model $\boldsymbol{y} = \boldsymbol{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}$.

2. Define the residual vector $\boldsymbol{e} = \boldsymbol{y} - \sum_{i=1}^m \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_i$. Show that

$$\boldsymbol{y}^T\boldsymbol{y} = \sum_{i=1}^m \widehat{\boldsymbol{\beta}}_i^T \boldsymbol{X}_i^T \boldsymbol{X}_i \widehat{\boldsymbol{\beta}}_i + \boldsymbol{e}^T\boldsymbol{e}. \tag{3}$$

3. Explain how to interpret expression (3) as an "analysis of variance" decomposition of the total sum of squares $\boldsymbol{y}^T\boldsymbol{y}$ into m sums of squares due to regression (one corresponding to each of the submatrices $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$) and the residual sum of squares. More specifically, in your answers, construct a table with the following columns:

Source	Sum of Squares	Degrees of Freedom	Mean Square

4. Suppose that we want to test the null hypothesis H_0: all of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_\ell$ are zero vectors for a given ℓ between 1 and m, against the alternative hypothesis H_a: not all of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_\ell$ are zero vectors. Show that, using the notation introduced previously, a suitable test statistic is

$$\frac{\sum_{i=1}^\ell \widehat{\boldsymbol{\beta}}_i^T \boldsymbol{X}_i^T \boldsymbol{X}_i \widehat{\boldsymbol{\beta}}_i}{\sum_{i=1}^\ell p_i} \times \frac{n-p}{\boldsymbol{e}^T\boldsymbol{e}}\,. \tag{4}$$

In addition, specify the distribution of the test statistic in expression (4) under H_0.

Part II

The data provided in Table 1 on page 5 are the result of an experiment in which y was the measured concentration of phytic acid and x_1, x_2, x_3, and x_4 were the weights of different ingredients of an infant food formula. Note that $x_1, \ldots, x_4$ were standardized to the values -1, 0, and 1 (i.e., they represent a coding of the natural variables into dimensionless variables).

The full model corresponding to this experiment, often referred to as a "quadratic" model, has the form

$$y_i = \gamma_0 + \sum_{j=1}^{4} \gamma_j x_{ji} + \sum_{j=1}^{4} \gamma_{jj} x_{ji}^2 + \sum_{j=1}^{3} \sum_{k=j+1}^{4} \gamma_{jk} x_{ji} x_{ki} + \epsilon_i, \tag{5}$$

where $i = 1, \ldots, 25$ is the "run" number, $\epsilon_1, \ldots, \epsilon_{25}$ are random errors, assumed to be independent and normally distributed with mean 0 and unknown common variance $\sigma^2 > 0$. We refer to γ_0, $\sum_{j=1}^{4} \gamma_j x_{ji}$, $\sum_{j=1}^{4} \gamma_{jj} x_{ji}^2$, and $\sum_{j=1}^{3} \sum_{k=j+1}^{4} \gamma_{jk} x_{ji} x_{ki}$ as the "intercept," "linear," "quadratic," and "cross-product" terms, respectively.

A common modification of the quadratic model (5) is to replace each of the covariates x_{ji}^2 by $x_{ji}^2 - c$, with all other terms remaining the same, where c is a constant to be determined below. (Note that we still refer to $\sum_{j=1}^{4} \gamma_{jj}(x_{ji}^2 - c)$ as the "quadratic" term.) The form of the modified full model is thus given by

$$y_i = \gamma_0 + \sum_{j=1}^{4} \gamma_j x_{ji} + \sum_{j=1}^{4} \gamma_{jj} (x_{ji}^2 - c) + \sum_{j=1}^{3} \sum_{k=j+1}^{4} \gamma_{jk} x_{ji} x_{ki} + \epsilon_i. \tag{6}$$

Let $\boldsymbol{y} = (y_1, \ldots, y_{25})^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_{25})^T$. Then model (6) can be written in the form of model (1) with $m = 4$, that is,

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{X}_3\boldsymbol{\beta}_3 + \boldsymbol{X}_4\boldsymbol{\beta}_4 + \boldsymbol{\epsilon}, \tag{7}$$

where $\boldsymbol{X}_1\boldsymbol{\beta}_1$, $\boldsymbol{X}_2\boldsymbol{\beta}_2$, $\boldsymbol{X}_3\boldsymbol{\beta}_3$, and $\boldsymbol{X}_4\boldsymbol{\beta}_4$ represent the intercept, linear, quadratic, and cross-product terms of model (6), respectively.

5. Show that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal, in the sense defined in expression (2).

6. Determine the value of c for which $\boldsymbol{X}_1$ and $\boldsymbol{X}_3$ are orthogonal, in the sense defined in expression (2).

Note that this c value will make $\boldsymbol{X}_1 \ldots, \boldsymbol{X}_4$ mutually orthogonal. You may use this fact without proof and any results from part I in answering the following questions.

7. Under the modified quadratic model (6), the parameter estimates, standard errors, and t-values are given in Table 2 on page 6. In addition, the sums of squares corresponding to the four terms and the residual sum of squares are given below.

Source	Sum of Squares
Intercept	37.6996
Linear	0.58455
Quadratic	0.10454
Cross-product	0.09905
Residual	0.22326

Test whether each of the intercept, linear, quadratic, and cross-product regression coefficients $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_4$ in model (7) is a zero vector. Give the test statistic, its distribution (with corresponding degrees of freedom) under the null hypothesis, and your conclusion for each of the four tests. Provide sufficient details to justify your conclusions.

8. Test whether both the quadratic and cross-product regression coefficients $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$ in model (7) are zero vectors. Give the test statistic, its distribution (with corresponding degrees of freedom) under the null hypothesis, and your conclusion. Provide sufficient details to justify your conclusion.

9. Test whether all the γ_{jk} $(1 \le j \le k \le 4)$ in model (5) are zero. Give the test statistic, its distribution (with corresponding degrees of freedom) under the null hypothesis, and your conclusion. Provide sufficient details to justify your conclusion.

10. Suppose that we fit the model containing just the intercept and linear terms, assuming the quadratic and cross-product terms are 0. By the result of question 1 in part I, the values of $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_4$ will be the same as those given in Table 2, but the standard errors will be different. Calculate the standard errors of $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_4$ under the model with only the intercept and linear terms. Determine which of the estimated parameters $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_4$ are statistically significant.

11. In your own words, explain the overall conclusions drawn from this experiment for the food company.

Consider fitting the model containing just the intercept and linear terms of x_1 and x_2,

$$y_i = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i} + \epsilon_i, \text{ for } \; i = 1, \ldots, 25, \tag{8}$$

to the data in Table 1 on page 5. Answer the following questions based on this model.

12. Show that the least squares estimators of γ_0, γ_1, and γ_2 are the same as those given in Table 2.

13. Derive a 95% confidence interval for the mean concentration of phytic acid when $x_1 = -1$ and $x_2 = -1$.

14. Derive a 95% prediction interval for the next measured concentration of phytic acid when $x_1 = -1$ and $x_2 = -1$.

15. Derive a 95% confidence interval for the difference between the mean concentration of phytic acid when $x_1 = -1$ and $x_2 = -1$ and the mean concentration when $x_1 = 1$ and $x_2 = 1$.

Table 1: Data set for the nutrition study.

Run	x_1	x_2	x_3	x_4	y
1	-1	-1	0	0	0.95
2	-1	0	-1	0	1.19
3	-1	0	0	-1	0.98
4	-1	0	0	1	1.06
5	-1	0	1	0	1.04
6	-1	1	0	0	1.21
7	0	-1	-1	0	1.12
8	0	-1	0	-1	1.03
9	0	-1	0	1	1.23
10	0	-1	1	0	1.10
11	0	0	-1	-1	1.25
12	0	0	-1	1	1.19
13	0	0	0	0	1.16
14	0	0	1	-1	0.96
15	0	0	1	1	1.40
16	0	1	-1	0	1.37
17	0	1	0	-1	1.53
18	0	1	0	1	1.87
19	0	1	1	0	1.22
20	1	-1	0	0	1.12
21	1	0	-1	0	1.20
22	1	0	0	-1	1.43
23	1	0	0	1	1.37
24	1	0	1	0	1.35
25	1	1	0	0	1.37

Table 2: Estimation summaries for the modified quadratic model (6) fit to the data in Table 1.

Parameter	Estimate	Standard Error	t Value
γ_0	1.22800	0.02988	41.09
γ_1	0.11750	0.04314	2.72
γ_2	0.16833	0.04314	3.90
γ_3	0.02083	0.04314	0.48
γ_4	0.07833	0.04314	1.82
γ_{11}	0.02708	0.08892	0.30
γ_{22}	0.07917	0.08892	0.89
γ_{33}	0.01208	0.08892	0.14
γ_{44}	0.10167	0.08892	1.14
γ_{12}	0.00250	0.07471	0.03
γ_{13}	0.07500	0.07471	1.00
γ_{14}	0.03500	0.07471	0.47
γ_{23}	0.03250	0.07471	0.44
γ_{24}	0.03500	0.07471	0.47
γ_{34}	0.12500	0.07471	1.67

1. Let $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)$. Since $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ are mutually orthogonal, we have

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1^T\boldsymbol{X}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{X}_2^T\boldsymbol{X}_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{X}_m^T\boldsymbol{X}_m \end{bmatrix}$$

and

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & (\boldsymbol{X}_m^T\boldsymbol{X}_m)^{-1} \end{bmatrix}.$$

Hence

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \begin{bmatrix} (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & (\boldsymbol{X}_m^T\boldsymbol{X}_m)^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_1^T\boldsymbol{y} \\ \boldsymbol{X}_2^T\boldsymbol{y} \\ \vdots \\ \boldsymbol{X}_m^T\boldsymbol{y} \end{bmatrix} \\ &= \begin{bmatrix} (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{y} \\ (\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T\boldsymbol{y} \\ \vdots \\ (\boldsymbol{X}_m^T\boldsymbol{X}_m)^{-1}\boldsymbol{X}_m^T\boldsymbol{y} \end{bmatrix}, \end{aligned}$$

that is, $\widehat{\boldsymbol{\beta}}_i = (\boldsymbol{X}_i^T\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^T\boldsymbol{y}$, for $i = 1, \ldots, m$.

2. By the mutual orthogonality of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ and the result in question 1, we have, for $i = 1, \ldots, m$,

$$\widehat{\boldsymbol{\beta}}_i^T\boldsymbol{X}_i^T\boldsymbol{e} = \widehat{\boldsymbol{\beta}}_i\boldsymbol{X}_i^T\left(\boldsymbol{y} - \sum_{j=1}^m \boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j\right) = \widehat{\boldsymbol{\beta}}_i^T\left(\boldsymbol{X}_i^T\boldsymbol{y} - \boldsymbol{X}_i^T\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_i\right) = 0.$$

Thus,

$$\begin{aligned} \boldsymbol{y}^T\boldsymbol{y} &= \left(\sum_{i=1}^m \boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_i + \boldsymbol{e}\right)^T\left(\sum_{j=1}^m \boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j + \boldsymbol{e}\right) \\ &= \sum_{i=1}^m\sum_{j=1}^m \widehat{\boldsymbol{\beta}}_i^T\boldsymbol{X}_i^T\boldsymbol{X}_j\widehat{\boldsymbol{\beta}}_j + 2\sum_{i=1}^m \widehat{\boldsymbol{\beta}}_i^T\boldsymbol{X}_i^T\boldsymbol{e} + \boldsymbol{e}^T\boldsymbol{e} \\ &= \sum_{i=1}^m \widehat{\boldsymbol{\beta}}_i^T\boldsymbol{X}_i^T\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_i + \boldsymbol{e}^T\boldsymbol{e}. \end{aligned}$$

3. The analysis-of-variance (ANOVA) table that explains the result in question 2 is given below.

Source	Sum of Squares	Degrees of Freedom	Mean Square
Regression on $\boldsymbol{X}_1$	$\widehat{\boldsymbol{\beta}}_1^T \boldsymbol{X}_1^T \boldsymbol{X}_1 \widehat{\boldsymbol{\beta}}_1$	p_1	$\frac{\widehat{\boldsymbol{\beta}}_1^T \boldsymbol{X}_1^T \boldsymbol{X}_1 \widehat{\boldsymbol{\beta}}_1}{p_1}$
Regression on $\boldsymbol{X}_2$	$\widehat{\boldsymbol{\beta}}_2^T \boldsymbol{X}_2^T \boldsymbol{X}_2 \widehat{\boldsymbol{\beta}}_2$	p_2	$\frac{\widehat{\boldsymbol{\beta}}_2^T \boldsymbol{X}_2^T \boldsymbol{X}_2 \widehat{\boldsymbol{\beta}}_2}{p_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
Regression on $\boldsymbol{X}_m$	$\widehat{\boldsymbol{\beta}}_m^T \boldsymbol{X}_m^T \boldsymbol{X}_m \widehat{\boldsymbol{\beta}}_m$	p_m	$\frac{\widehat{\boldsymbol{\beta}}_m^T \boldsymbol{X}_m^T \boldsymbol{X}_m \widehat{\boldsymbol{\beta}}_m}{p_m}$
Residual	$\boldsymbol{e}^T \boldsymbol{e}$	$n - p$	$\frac{\boldsymbol{e}^T \boldsymbol{e}}{n-p}$
Total	$\boldsymbol{y}^T \boldsymbol{y}$	n	

4. By the above ANOVA table, the F statistic for testing H_0 versus H_a is given by

$$\frac{(SSE_r - SSE_f)/\sum_{i=1}^{l} p_i}{SSE_f/(n-p)},$$

where SSE_r and SSE_f are the residual sums of squares under the reduced and full models. Using the result in question 2, we have

$$\begin{aligned} SSE_r &= \boldsymbol{y}^T \boldsymbol{y} - \sum_{i=\ell+1}^{m} \widehat{\boldsymbol{\beta}}_i^T \boldsymbol{X}_i^T \boldsymbol{X}_i \widehat{\boldsymbol{\beta}}_i; \\ SSE_f &= \boldsymbol{y}^T \boldsymbol{y} - \sum_{i=1}^{m} \widehat{\boldsymbol{\beta}}_i^T \boldsymbol{X}_i^T \boldsymbol{X}_i \widehat{\boldsymbol{\beta}}_i = \boldsymbol{e}^T \boldsymbol{e}. \end{aligned}$$

Thus, $SSE_r - SSE_f = \sum_{i=1}^{\ell} \widehat{\boldsymbol{\beta}}_i^T \boldsymbol{X}_i^T \boldsymbol{X}_i \widehat{\boldsymbol{\beta}}_i$, and the F statistic is indeed given by expression (4). Its distribution under H_0 is $F_{\sum_{i=1}^{\ell} p_i, n-p}$, the F distribution with numerator and denominator degrees of freedom $\sum_{i=1}^{\ell} p_i$ and $n - p$.

5. Let $\boldsymbol{x}_0$ be a 25×1 vector of 1's. For $j = 1, \ldots, 4$, let $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{j,25})^T$ be a 25×1 vector consisting of the values in the x_j column of Table 1. That is,

$$\begin{aligned} \boldsymbol{x}_1 &= (-1, -1, -1, -1, -1, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1)^T; \\ \boldsymbol{x}_2 &= (-1, 0, 0, 0, 0, 1, -1, -1, -1, -1, 0, 0, 0, 0, 0, 1, 1, 1, 1, -1, 0, 0, 0, 0, 1)^T; \\ \boldsymbol{x}_3 &= (0, -1, 0, 0, 1, 0, -1, 0, 0, 1, -1, -1, 0, 1, 1, -1, 0, 0, 1, 0, -1, 0, 0, 1, 0)^T; \\ \boldsymbol{x}_4 &= (0, 0, -1, 1, 0, 0, 0, -1, 1, 0, -1, 1, 0, -1, 1, 0, -1, 1, 0, 0, 0, -1, 1, 0, 0)^T. \end{aligned}$$

Then $\boldsymbol{X}_1 = \boldsymbol{x}_0$ and $\boldsymbol{X}_2 = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_4)$. For $j = 1, \dots, 4$, $\boldsymbol{x}_0^T \boldsymbol{x}_j$ is the sum of all the non-zero elements of $\boldsymbol{x}_j$, or the sum of 6 1's and 6 (-1)'s. Thus, $\boldsymbol{X}_1^T \boldsymbol{X}_2 = (0, 0, 0, 0)$, implying that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are orthogonal.

6. For $j = 1, \dots, 4$, the jth element of the 1×4 vector $\boldsymbol{X}_1^T \boldsymbol{X}_3$ is

$$\sum_{i=1}^{25} (x_{ji}^2 - c) = \sum_{i=1}^{25} x_{ji}^2 - 25c = 12 - 25c.$$

Thus, $\boldsymbol{X}_1$ and $\boldsymbol{X}_3$ are orthogonal if and only if $12 - 25c = 0$, or $c = 12/25 = 0.48$.

7. By the results in questions 2 to 4, we can use the following ANOVA table to test whether each of these coefficients is a zero vector.

Source	Sum of Squares	Degrees of Freedom	Mean Square	F
Intercept ($\boldsymbol{\beta}_1$)	37.6996	1	37.6996	1688.6
Linear ($\boldsymbol{\beta}_2$)	0.58455	4	0.14614	6.55
Quadratic ($\boldsymbol{\beta}_3$)	0.10454	4	0.02614	1.17
Cross-product ($\boldsymbol{\beta}_4$)	0.09905	6	0.01651	0.74
Residual	0.22326	10	0.02233	
Total	38.7110	25		

For testing whether each $\boldsymbol{\beta}_j$ ($j = 1, \dots, 4$) is a zero vector, the observed f values are 1688.6, 6.55, 1.17, and 0.74; the distributions of the F statistics under the null hypotheses are $F_{1,10}, F_{4,10}, F_{4,10}$, and $F_{6,10}$; the critical values at the 5% significance level are 4.96, 3.48, 3.48, and 3.22. Thus, there is strong evidence that $\boldsymbol{\beta}_j$ is a non-zero vector for $j = 1, 2$, but little evidence that $\boldsymbol{\beta}_j$ is a non-zero vector for $j = 3, 4$.

8. To test whether both $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$ are zero vectors, we can use the result in question 4 and the above ANOVA table to obtain the observed f as follows:

$$\frac{(0.10454 + 0.09905)/(4 + 6)}{0.22326/10} = 0.91.$$

The distribution of the F statistic under the null hypothesis is $F_{10,10}$, and the critical value at the 5% significance level is 2.98. Thus, there is little evidence that at least one of $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$ is a non-zero vector.

9. Note that model (5) becomes model (6) if we replace γ_0 by $\gamma_0^* = \gamma_0 - c \sum_{j=1}^4 \gamma_{jj}$. Thus, fitting models (5) and (6) to the data gives the same residual sum of squares (0.22326). In addition, the null hypothesis here is the same as that in question 8, so the residual sum of squares for the reduced model (under the null hypothesis) is the same as that in question 8, or $0.22326 + 0.09905 + 0.10454 = 0.42685$. Thus, the observed f is given by

$$\frac{(SSE_r - SSE_f)/10}{SSE_f/10} = \frac{(0.42685 - 0.22326)/10}{0.22326/10} = 0.91.$$

So the answers to this question are the same as those to question 8.

10. Note that $(\gamma_1, \ldots, \gamma_4)^T = \boldsymbol{\beta}_2$. By the solution to question 5,

$$\operatorname{Var}(\widehat{\boldsymbol{\beta}}_2) = \left(\boldsymbol{X}_2^T \boldsymbol{X}_2\right)^{-1} \sigma^2 = (12\boldsymbol{I}_4)^{-1} \sigma^2 = \left(\sigma^2/12\right) \boldsymbol{I}_4.$$

Thus, by the solution to question 9, the standard errors of $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_4$ are given by

$$\sqrt{\frac{\widehat{\sigma^2}}{12}} = \sqrt{\frac{0.42685/20}{12}} = 0.0422.$$

Together with the values of $\widehat{\gamma}_1, \ldots, \widehat{\gamma}_4$ in Table 2, this gives the observed t values 2.78, 3.99, 0.49, and 1.86. Using the critical value 2.086 (for t_{20}) at the 5% significance level, we conclude that $\widehat{\gamma}_1$ and $\widehat{\gamma}_2$ are statistically significant, but $\widehat{\gamma}_3$ and $\widehat{\gamma}_4$ are not.

11. The x_1 and x_2 ingredients should be kept low to reduce the concentration of phytic acid in the infant food formula; the other two ingredients do not appear to have any significant effect on the concentration.

12. By the solutions to questions 5 and 10, the model in question 10 and model (8) can be written in the form of model (1) as

$$\boldsymbol{y} = \sum_{j=0}^{5} \boldsymbol{x}_j \gamma_j + \boldsymbol{\epsilon} \quad \text{and} \quad \boldsymbol{y} = \sum_{j=0}^{2} \boldsymbol{x}_j \gamma_j + \boldsymbol{\epsilon}.$$

By the result of question 1, for $j = 0, 1, 2$, the least squares (LS) estimators of γ_j under both models are given by $\left(\boldsymbol{x}_j^T \boldsymbol{x}_j\right)^{-1} \boldsymbol{x}_j^T \boldsymbol{y}$, and are thus the same.

13. By the result in question 2 and the solutions to questions 7 and 12, the residual sum of squares for model (8) is given by

$$\boldsymbol{y}^T \boldsymbol{y} - \sum_{j=0}^{2} \left(\boldsymbol{x}_j^T \boldsymbol{x}_j\right) \widehat{\gamma}_j^2 = 38.711 - 25(1.228)^2 - 12(0.1175)^2 - 12(0.16833)^2 = 0.5057.$$

The mean concentration of phytic acid when $x_1 = -1$ and $x_2 = -1$ is $\mu_{(-1,-1)} = \gamma_0 - \gamma_1 - \gamma_2$; its LS estimator is given by $\widehat{\mu}_{(-1,-1)} = \widehat{\gamma}_0 - \widehat{\gamma}_1 - \widehat{\gamma}_2$. So the LS estimate of $\mu_{(-1,-1)}$ is $1.228 - 0.1175 - 0.16833 = 0.94217$.

Similar to the solution to question 10, the standard error of $\widehat{\mu}_{(-1,-1)}$ is given by

$$\sqrt{\widehat{\operatorname{Var}}\left(\widehat{\gamma}_0\right) + \widehat{\operatorname{Var}}\left(\widehat{\gamma}_1\right) + \widehat{\operatorname{Var}}\left(\widehat{\gamma}_2\right)} = \sqrt{\left(\frac{1}{25} + \frac{1}{12} + \frac{1}{12}\right) \widehat{\sigma}^2} = \sqrt{\frac{31}{150} \cdot \frac{0.5057}{22}} = 0.06892.$$

The 0.975 quantile of the t_{22} distribution is 2.074; thus, a 95% confidence interval for the mean concentration of phytic acid when $x_1 = -1$ and $x_2 = -1$ is given by

$$0.94217 \pm 2.074(0.06892), \quad \text{or} \quad (0.80, 1.09).$$

14. To obtain a 95% prediction interval for the next measured concentration of phytic acid when $x_1 = -1$ and $x_2 = -1$, we only need to replace the standard error 0.06892 in the solution to question 13 by

$$\sqrt{\left(1 + \frac{1}{25} + \frac{1}{12} + \frac{1}{12}\right)\widehat{\sigma}^2} = \sqrt{\frac{181}{150} \cdot \frac{0.5057}{22}} = 0.1665.$$

So the 95% prediction interval is given by

$$0.94217 \pm 2.074(0.1665), \quad \text{or} \quad (0.60, 1.29).$$

15. The difference between the mean concentration of phytic acid when $x_1 = -1$ and $x_2 = -1$ and the mean concentration when $x_1 = 1$ and $x_2 = 1$ is $\mu_d = \mu_{(-1,-1)} - \mu_{(1,1)} = -2\gamma_1 - 2\gamma_2$. Thus, the LS estimate of μ_d is $-2(0.1175 + 0.16833) = -0.5717$; the standard error of $\widehat{\mu}_d$ is given by

$$2\sqrt{\left(\frac{1}{12} + \frac{1}{12}\right)\widehat{\sigma}^2} = 2\sqrt{\frac{1}{6} \cdot \frac{0.5057}{22}} = 0.1238.$$

Thus, a 95% confidence interval for the difference in means μ_d is given by

$$-0.5717 \pm 2.074(0.1238), \quad \text{or} \quad (-0.83, -0.31).$$

Metal-oxide semiconductor (MOS) transistors are the basic building blocks of microelectronic devices. An analog MOS transistor is a voltage-controlled electrical switch that allows for different intensities of current to travel from a source to a destination, depending on the voltage to which the transistor is subjected. MOS transistors are tiny: a thousand of them fit within the width of a human hair. When several MOS transistors are placed on a thin silica wafer, they form a *circuit*, with many different "channels" through which current can travel.

Data from an experiment conducted in the Microelectronics Division of Lucent Technologies are described in Pinheiro and Bates (2000). The experiment was designed to study the variability in current between and within manufactured n-channel analog MOS circuits, at five different voltage levels. Ten wafers randomly selected from a large population of wafers were used in the experiment. We use $n_w = 10$ to denote the number of wafers. The response variable is the intensity of the current measured in milli-Amperes (mA) at eight randomly chosen sites on each wafer. We use $n_s = 8$ to denote the number of measurement sites on each wafer. Finally, we use $n_v = 5$ to denote the number of voltage (V) levels applied to each wafer. Therefore, the total number of measurements obtained on each wafer is $8 \times 5 = 40$, and the total number of measurements N in the experiment is $10 \times 40 = 400$.

Measurements taken at two sites labeled 1 and 2 on wafer labeled 1 are shown in Table 1 below.

Wafer	Site	Voltage (V)	Current (mA)
1	1	0.8	0.9009
1	1	1.2	3.8682
1	1	1.6	7.6406
1	1	2.0	11.7360
1	1	2.4	15.9340
1	2	0.8	1.0320
1	2	1.2	4.1022
1	2	1.6	7.9316
1	2	2.0	12.0640
1	2	2.4	16.2940

Table 1: *Current at five voltage levels, on wafer 1, sites 1 and 2.*

Figure 1 shows current plotted against voltage, for each wafer and site combination.

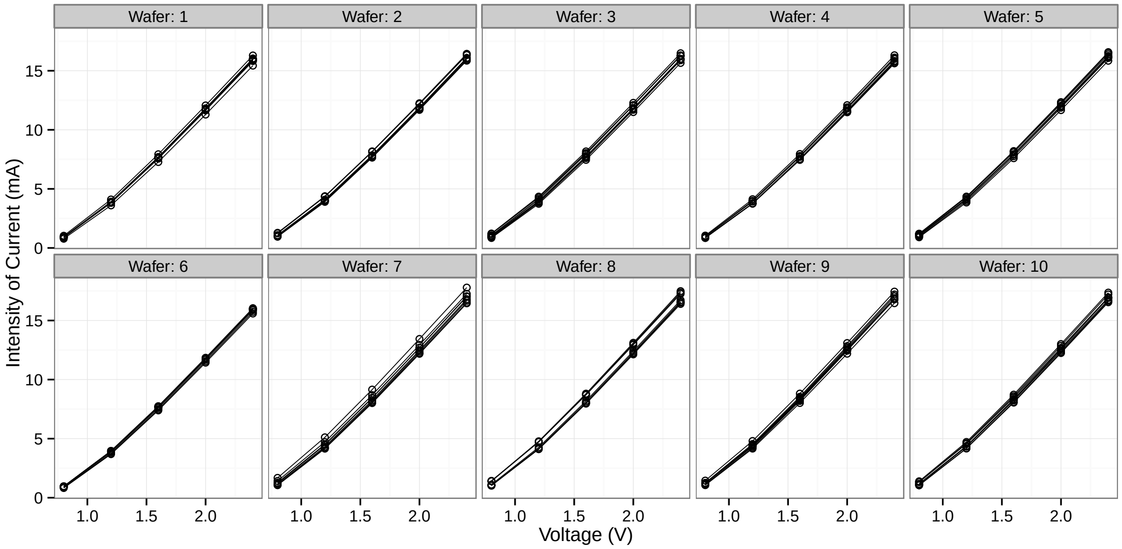

Figure 1: *Current (mA) by voltage (V). There is one panel for each wafer, and in each panel there are 8 curves, one for each measurement site.*

We use Y_{ijk} to denote the current measured on the jth site at the kth voltage level on the ith wafer, where $i = 1, ..., n_w = 10$, $j = 1, ..., n_s = 8$ and $k = 1, ..., n_v = 5$. The mean responses by wafer for each voltage level are given by

$$\bar{Y}_{i \cdot k} = \frac{1}{n_s} \sum_{j=1}^{n_s} Y_{ijk},$$

and are shown in Table 2.

					Wafer					
Voltage	1	2	3	4	5	6	7	8	9	10
0.8	0.907	1.083	1.023	0.948	1.074	0.937	1.212	1.172	1.234	1.228
1.2	3.865	4.106	4.038	3.924	4.127	3.903	4.404	4.355	4.46	4.452
1.6	7.629	7.876	7.817	7.687	7.927	7.661	8.324	8.279	8.408	8.402
2.0	11.718	11.93	11.902	11.769	12.031	11.739	12.537	12.502	12.648	12.643
2.4	15.915	16.1	16.089	15.954	16.23	15.92	16.839	16.816	16.974	16.973

Table 2: *Response averaged across sites for each wafer and voltage combination.*

Part I

Consider the linear random coefficients model

$$Y_{ijk} = (\beta_0 + b_{10i} + b_{20ij}) + (\beta_1 + b_{11i} + b_{21ij})x_k + (\beta_2 + b_{12i} + b_{22ij})x_k^2 + \epsilon_{ijk}, \qquad \text{(Model 1)}$$

where

$$\mathbf{b}_{1i} = \begin{bmatrix} b_{10i} \\ b_{11i} \\ b_{12i} \end{bmatrix} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_1), \qquad \mathbf{b}_{2ij} = \begin{bmatrix} b_{20ij} \\ b_{21ij} \\ b_{22ij} \end{bmatrix} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_2), \qquad \epsilon_{ijk} \sim \mathrm{N}(0, \sigma^2),$$

and $\mathbf{\Sigma}_1 = \{\sigma_{1lm}\}$, $\mathbf{\Sigma}_2 = \{\sigma_{2lm}\}$ are unknown symmetric positive definite matrices of dimensions 3×3, and σ^2 is an unknown scalar-valued parameter, and all the $\mathbf{b}_{1i}$, $\mathbf{b}_{2i}$ and ϵ_{ijk} are independent. In Model 1, $x_k = \text{voltage}_k - 1.6$, where 1.6 is the mean voltage level. Unknown parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ are fixed effects in the quadratic model. Random effects $\mathbf{b}_{1i}$ allow for wafer-to-wafer variability, and $\mathbf{b}_{2ij}$ allow for within-wafer (between-site) variability in the association between the response and voltage. Finally, $\mathbf{b}_1, \mathbf{b}_2$ denote the vectors with elements $\{\mathbf{b}_{1i}\}$ and $\{\mathbf{b}_{2ij}\}$, respectively.

1. For Model 1 with assumptions listed above, derive the following expectations and variances. Express your answers using matrix notation (i.e. $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$, σ^2), where appropriate.

a. $E[Y_{ijk}]$

b. $E[Y_{ijk}|\mathbf{b}_{1i}]$

c. $E[Y_{ijk} - Y_{ijk'}|\mathbf{b}_{1i}, \mathbf{b}_{2ij}]$, for $k \neq k'$

d. $\mathrm{var}(Y_{ijk})$

e. $\mathrm{var}(Y_{ijk} + Y_{ijk'})$, for $k \neq k'$

f. $\mathrm{var}(Y_{ijk} + Y_{ijk'}|\mathbf{b}_{1i}, \mathbf{b}_{2ij})$

2. Let $\mathbf{Y}_i$ denote the $(n_s * n_v) \times 1$ vector of measurements for the ith wafer and let $\boldsymbol{\epsilon}_i$ denote the corresponding vector of errors. We can re-express Model 1 as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_{1i}\mathbf{b}_{1i} + \mathbf{Z}_{2i}\mathbf{b}_{2ij} + \boldsymbol{\epsilon}_i,$$

where $\mathbf{X}_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i}$ are design matrices.

a. What are the dimensions of the matrices $\mathbf{X}_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i}$?

b. Write out the first five rows of the matrices $\mathbf{X}_i$ and $\mathbf{Z}_{1i}$ corresponding to measurements taken on site 1 in wafer 1. Assume that observations are ordered by voltage within site within wafer.

Part II

We fit Model 1 using the $N = 400$ measurements and assuming that the covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are diagonal, so that

$$\mathbf{\Sigma}_1 = \begin{bmatrix} \sigma_{10}^2 & 0 & 0 \\ 0 & \sigma_{11}^2 & 0 \\ 0 & 0 & \sigma_{12}^2 \end{bmatrix}, \quad \mathbf{\Sigma}_2 = \begin{bmatrix} \sigma_{20}^2 & 0 & 0 \\ 0 & \sigma_{21}^2 & 0 \\ 0 & 0 & \sigma_{22}^2 \end{bmatrix}.$$

Refer to the output labeled **Output for Model 1** shown below. The R command used to produce the output is

```
model1 = lme(current ~ voltage + I(voltage^2), data = Wafer2, random=
   list(Wafer=pdDiag(~voltage + I(voltage^2)), Site=pdDiag(~voltage + I(voltage^2))))
```

Output for Model 1

```
        AIC      BIC   logLik
  -198.7254 -158.886 109.3627

Random effects:
 Formula: ~x + I(x^2) | Wafer
 Structure: Diagonal
        (Intercept)      x         I(x^2)
StdDev:   0.30198     0.21142    1.9745e-06

 Formula: ~ x + I(x^2) | Site %in% Wafer
 Structure: Diagonal
        (Intercept)    voltage   I(voltage^2)  Residual
StdDev:   0.22372      0.04092    8.0651e-06   0.12492

Fixed effects: y ~ x + I(x^2)
                 Value  Std.Error  DF   t-value p-value
(Intercept)    7.9804   0.0991   318   80.4515       0
x              9.6486   0.0679   318  142.0645       0
I(x^2)         1.1704   0.0233   318   50.1660       0
 Correlation:
             (Intr)    x
x             0.000
I(x^2)       -0.075  0.000

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
     -1.6365     -0.8681     0.0480      0.6798     2.1269

Number of Observations: 400
Number of Groups:
          Wafer Site %in% Wafer
             10              80
```

3. Compute an estimate of the expected current intensity and its standard error for wafers to which we apply voltage equal to 1.5 V.

4. **a**. Find the REML estimates of $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$ and σ^2.

b. Comment on the relative sizes of the REML estimates of the standard deviations of the random effects, and interpret what they mean in the context of this problem.

5. Figure 2 shows the within-group residuals e_{ijk} plotted against centered voltage, by wafer and site combination. The within-group residuals are computed as the difference between the observed response and the within-group fitted value. What do you conclude from inspection of these residual plots?

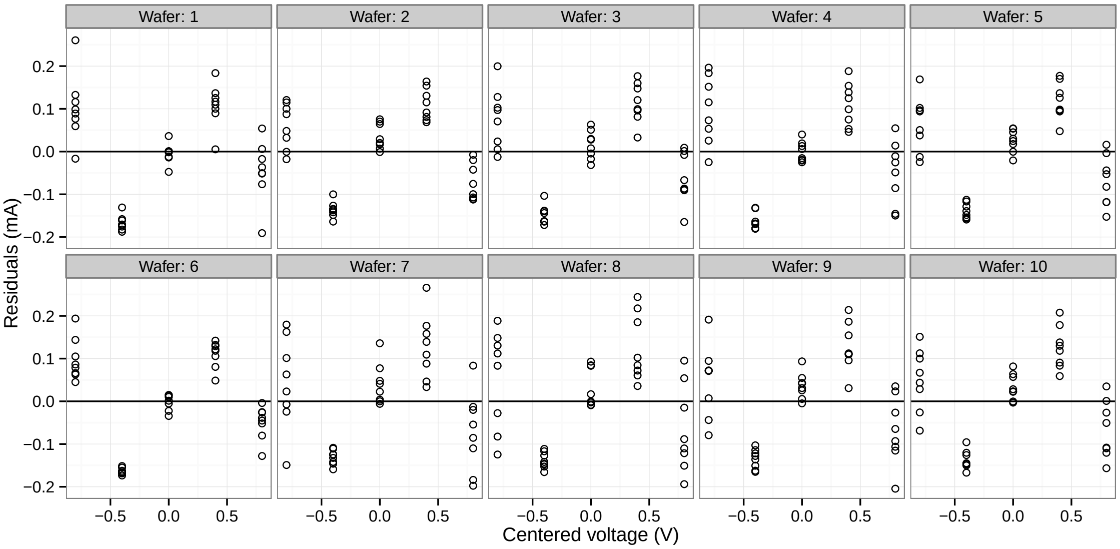

Figure 2: *Residuals for Model 1 plotted against centered voltage. There is one panel for each wafer, and in each panel there are 5 sets (each of 8) residuals, one set for each measurement site.*

Part III

We extended Model 1 by adding two more fixed effects to the model:

$$Y_{ijk} = (\beta_0 + b_{0i} + b_{0ij}) + (\beta_1 + b_{1i} + b_{1ij})x_k + (\beta_2 + b_{2i} + b_{2ij})x_k^2 + \beta_3 \cos(x_k) + \beta_4 \sin(x_k) + \epsilon_{ijk}, \tag{Model 2}$$

where β_3, β_4 are unknown regression coefficients. Once again we assume that the covariance matrices $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$ are diagonal, and fit the model to the vector of 400 responses. Refer to **Output for Model 2** on Page 6. The R command used to fit Model 2 is

```
model2 = lme(current ~ voltage + I(voltage^2) + cos(voltage) + sin(voltage),
    data = Wafer2, random=list(Wafer=pdDiag(~voltage + I(voltage^2)),
    Site=pdDiag(~voltage + I(voltage^2))))
```

Output for Model 2

```
        AIC       BIC    logLik
  -1189.308   -1141.562 606.6541

Random effects:
 Formula: ~x + I(x^2) | Wafer
 Structure: Diagonal
         (Intercept)    x     I(x^2)
StdDev:    0.3161    0.2114   0.0483

 Formula: ~x + I(x^2) | Site %in% Wafer
 Structure: Diagonal
        (Intercept)   x      I(x^2)    Residual
StdDev:   0.2491   0.1066   0.0611     0.00914

Fixed effects: y ~ x + I(x^2) + cos(x) + sin(x)
                 Value  Std.Error   DF   t-value    p-value
(Intercept)   -1.9954   0.3579     316  -5.5753       0
x              4.7621   0.0720     316  66.1136       0
I(x^2)         5.8802   0.1625     316  36.1653       0
cos(x)         9.9972   0.3432     316  29.1236       0
sin(x)         5.3576   0.0263     316 203.7107       0
 Correlation:
             (Intr)   x    I(x^2) cs(x)
x            0.000
I(x^2)      -0.952  0.000
cos(x)      -0.957  0.000  0.995
sin(x)       0.000 -0.333  0.000  0.000

Standardized Within-Group Residuals:
        Min          Q1         Med           Q3         Max
     -2.0932      -0.3116    -0.0175       0.3898     3.0480

Number of Observations: 400
Number of Groups:
          Wafer Site %in% Wafer
             10              80
```

6. **a**. Does adding the cosine and sine terms in centered voltage improve the fit of the model? Justify your answer.

b. Is it possible, given the current output, to conduct a likelihood ratio test (LRT) to compare Model 1 and Model 2? If not, explain why not.

7. Figure 3 shows the within-group residuals e_{ijk} for Model 2, plotted against centered voltage, by wafer and site combinations. What do you conclude from inspection of these residual plots? How do these residuals compare to the Model 1 residuals shown in Figure 2?

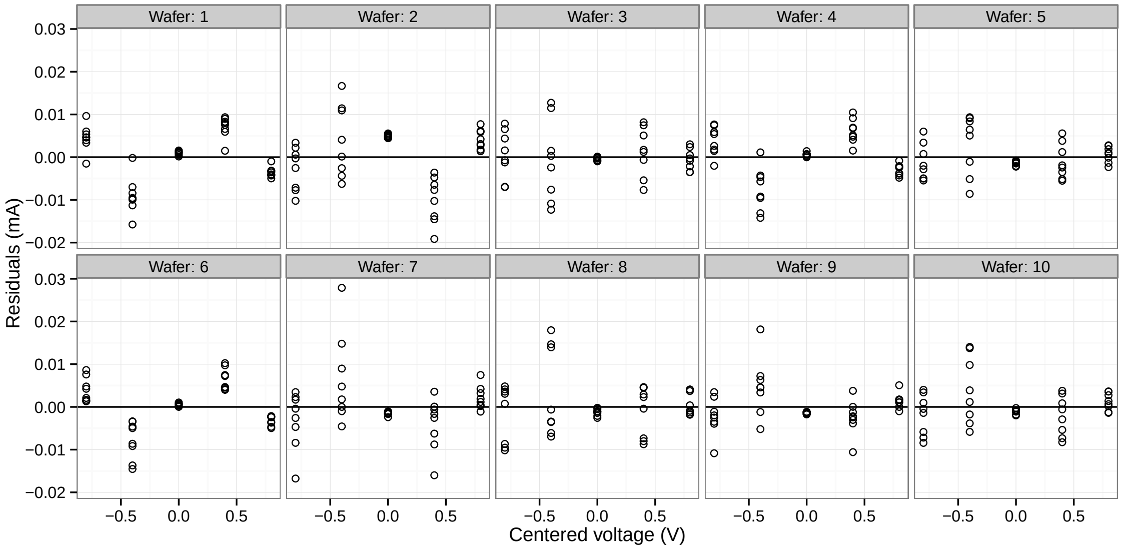

Figure 3: *Residuals for Model 2, plotted against voltage.There is one panel for each wafer, and in each panel there are 5 sets (each of 8) residuals, one set for each measurement site.*

8. Figure 4 shows the predicted values of the random coefficients $\hat{\mathbf{b}}_i$ associated with wafer, plotted against each other. Comment on what the plot indicates about the assumptions associated with Model 2.

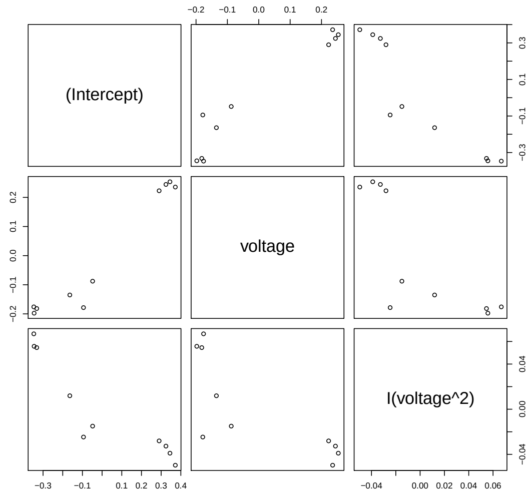

Figure 4: *Scatterplot matrix of predicted wafer-level random effects.*

Part IV

A third model fit to the 400 measurements was Model 2 modified in that the covariance matrix $\boldsymbol{\Sigma}_1$ is now assumed to be a general positive definite matrix with unknown elements σ^2_{1lm} for $l, m = 0, 1, 2$. We refer to this model as Model 3. The R command used to produce the output for Model 3 is

```
model3 = lme(current ~ voltage + I(voltage^2) + cos(voltage) + sin(voltage),
     data = Wafer2, random=list(Wafer=(~voltage + I(voltage^2)),
     Site=pdDiag(~voltage + I(voltage^2))), control=c1)
```

Output for Model 3

```
        AIC       BIC   logLik
  -1220.638 -1160.954 625.3188

Random effects:
 Formula: ~x + I(x^2) | Wafer
 Structure: General positive-definite, Log-Cholesky parametrization
             StdDev     Corr
(Intercept)  0.3240  (Intr)   x
x            0.2118  0.981
I(x^2)       0.0498 -0.944 -0.862

 Formula: ~x + I(x^2) | Site %in% Wafer
 Structure: Diagonal
        (Intercept)      x      I(x^2)        Residual
StdDev:   0.2397      0.1058   0.0600          0.0091

Fixed effects: y ~ x + I(x^2) + cos(x) + sin(x)
                  Value Std.Error  DF   t-value p-value
(Intercept)  -1.9954    0.3586   316  -5.5640       0
x             4.7621    0.0721   316  66.0108       0
I(x^2)        5.8802    0.1626   316  36.1459       0
cos(x)        9.9972    0.3433   316  29.1149       0
sin(x)        5.3576    0.0263   316 203.6501       0
Correlation:
             (Intr)     x   I(x^2)  cs(x)
x             0.260
I(x^2)       -0.976 -0.078
cos(x)       -0.955  0.000  0.994
sin(x)        0.000 -0.333  0.000  0.000

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
     -2.0717      -0.3288    -0.0182       0.4001     3.0491

Number of Observations: 400
Number of Groups:    Wafer Site %in% Wafer
                      10              80
```

9. Carry out a likelihood ratio test to compare Model 2 and Model 3. Use a Type I error probability $\alpha = 0.05$. Interpret the results of the test in the context of the problem.

10. Report what you have learned from your analysis to the Director of the Microelectronics Division at Lucent Technologies. In no more than 100 - 120 words, summarize what you have learned from this analysis regarding manufacturing variability in wafers.

Metal-oxide semiconductor (MOS) transistors are the basic building blocks of microelectronic devices. An analog MOS transistor is a voltage-controlled electrical switch that allows for different intensities of current to travel from a source to a destination, depending on the voltage to which the transistor is subjected. MOS transistors are tiny: a thousand of them fit within the width of a human hair. When several MOS transistors are placed on a thin silica wafer, they form a *circuit*, with many different "channels" through which current can travel.

Data from an experiment conducted in the Microelectronics Division of Lucent Technologies are described in Pinheiro and Bates (2000). The experiment was designed to study the variability in current between and within manufactured n-channel analog MOS circuits, at five different voltage levels. Ten wafers randomly selected from a large population of wafers were used in the experiment. We use $n_w = 10$ to denote the number of wafers. The response variable is the intensity of the current measured in milli-Amperes (mA) at eight randomly chosen sites on each wafer. We use $n_s = 8$ to denote the number of measurement sites on each wafer. Finally, we use $n_v = 5$ to denote the number of voltage (V) levels applied to each wafer. Therefore, the total number of measurements obtained on each wafer is $8 \times 5 = 40$, and the total number of measurements N in the experiment is $10 \times 40 = 400$.

Measurements taken at two sites labeled 1 and 2 on wafer labeled 1 are shown in Table 1 below.

Wafer	Site	Voltage (V)	Current (mA)
1	1	0.8	0.9009
1	1	1.2	3.8682
1	1	1.6	7.6406
1	1	2.0	11.7360
1	1	2.4	15.9340
1	2	0.8	1.0320
1	2	1.2	4.1022
1	2	1.6	7.9316
1	2	2.0	12.0640
1	2	2.4	16.2940

Table 1: *Current at five voltage levels, on wafer 1, sites 1 and 2.*

Figure 1 shows current plotted against voltage, for each wafer and site combination.

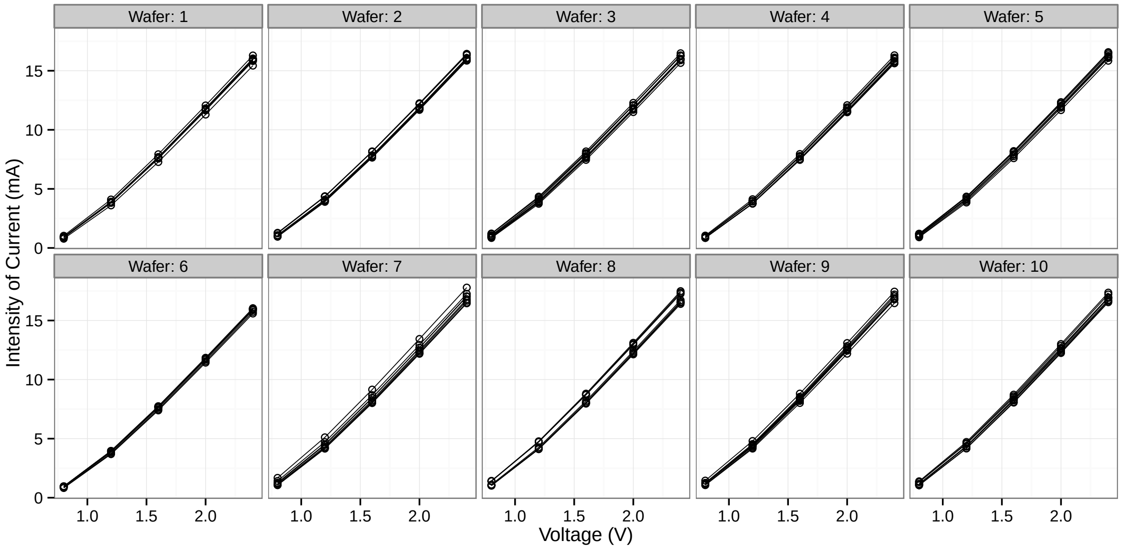

Figure 1: *Current (mA) by voltage (V). There is one panel for each wafer, and in each panel there are 8 curves, one for each measurement site.*

We use Y_{ijk} to denote the current measured on the jth site at the kth voltage level on the ith wafer, where $i = 1, ..., n_w = 10$, $j = 1, ..., n_s = 8$ and $k = 1, ..., n_v = 5$. The mean responses by wafer for each voltage level are given by

$$\bar{Y}_{i\cdot k} = \frac{1}{n_s} \sum_{j=1}^{n_s} Y_{ijk},$$

and are shown in Table 2.

	Wafer									
Voltage	1	2	3	4	5	6	7	8	9	10
0.8	0.907	1.083	1.023	0.948	1.074	0.937	1.212	1.172	1.234	1.228
1.2	3.865	4.106	4.038	3.924	4.127	3.903	4.404	4.355	4.46	4.452
1.6	7.629	7.876	7.817	7.687	7.927	7.661	8.324	8.279	8.408	8.402
2.0	11.718	11.93	11.902	11.769	12.031	11.739	12.537	12.502	12.648	12.643
2.4	15.915	16.1	16.089	15.954	16.23	15.92	16.839	16.816	16.974	16.973

Table 2: *Response averaged across sites for each wafer and voltage combination.*

Part I

Consider the linear random coefficients model

$$Y_{ijk} = (\beta_0 + b_{10i} + b_{20ij}) + (\beta_1 + b_{11i} + b_{21ij})x_k + (\beta_2 + b_{12i} + b_{22ij})x_k^2 + \epsilon_{ijk}, \qquad \text{(Model 1)}$$

where

$$\mathbf{b}_{1i} = \begin{bmatrix} b_{10i} \\ b_{11i} \\ b_{12i} \end{bmatrix} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_1), \qquad \mathbf{b}_{2ij} = \begin{bmatrix} b_{20ij} \\ b_{21ij} \\ b_{22ij} \end{bmatrix} \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_2), \qquad \epsilon_{ijk} \sim \mathrm{N}(0, \sigma^2),$$

and $\mathbf{\Sigma}_1 = \{\sigma_{1lm}\}$, $\mathbf{\Sigma}_2 = \{\sigma_{2lm}\}$ are unknown symmetric positive definite matrices of dimensions 3×3, and σ^2 is an unknown scalar-valued parameter, and all the $\mathbf{b}_{1i}$, $\mathbf{b}_{2i}$ and ϵ_{ijk} are independent. In Model 1, $x_k = \text{voltage}_k - 1.6$, where 1.6 is the mean voltage level. Unknown parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ are fixed effects in the quadratic model. Random effects $\mathbf{b}_{1i}$ allow for wafer-to-wafer variability, and $\mathbf{b}_{2ij}$ allow for within-wafer (between-site) variability in the association between the response and voltage. Finally, $\mathbf{b}_1, \mathbf{b}_2$ denote the vectors with elements $\{\mathbf{b}_{1i}\}$ and $\{\mathbf{b}_{2ij}\}$, respectively.

1. For Model 1 with assumptions listed above, derive the following expectations and variances. Express your answers using matrix notation (i.e. $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$, σ^2), where appropriate.

a. $E[Y_{ijk}] = \beta_0 + \beta_1 x_k + \beta_2 x_k^2$

b. $E[Y_{ijk}|\mathbf{b}_{1i}] = \beta_0 + b_{10i} + (\beta_1 + b_{11i})x_k + (\beta_2 + b_{12i})x_k^2$

c. $E[Y_{ijk} - Y_{ijk'}|\mathbf{b}_{1i}, \mathbf{b}_{2ij}]$, for $k \neq k' = (\beta_1 + b_{11i} + b_{21ij})(x_k - x_{k'}) + (\beta_2 + b_{12i} + b_{22ij})(x_k^2 - x_{k'}^2)$

d. $\mathrm{var}(Y_{ijk}) = \begin{bmatrix} 1 & x_k & x_k^2 \end{bmatrix} (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2) \begin{bmatrix} 1 \\ x_k \\ x_k^2 \end{bmatrix} + \sigma^2$

e. $\mathrm{var}(Y_{ijk}+Y_{ijk'})$, for $k \neq k' = \mathbf{z}'(\mathbf{\Sigma}_1+\mathbf{\Sigma}_2)\mathbf{z}+2\sigma^2$, where $\mathbf{z}' = \begin{bmatrix} 2 & x_k + x_{k'} & x_k^2 + x_{k'}^2 \end{bmatrix}$

f. $\mathrm{var}(Y_{ijk} + Y_{ijk'}|\mathbf{b}_{1i}, \mathbf{b}_{2ij}) = 2\sigma^2$

2. Let $\mathbf{Y}_i$ denote the $(n_s * n_v) \times 1$ vector of measurements for the ith wafer and let $\boldsymbol{\epsilon}_i$ denote the corresponding vector of errors. We can re-express Model 1 as

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_{1i}\mathbf{b}_{1i} + \mathbf{Z}_{2i}\mathbf{b}_{2ij} + \boldsymbol{\epsilon}_i,$$

where $\mathbf{X}_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i}$ are design matrices.

a. What are the dimensions of the matrices $\mathbf{X}_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i}$?

The matrices $\mathbf{X}_i, \mathbf{Z}_{1i}, \mathbf{Z}_{2i}$ have dimensions $40 * 3$, $40 * 30$ and $40 * 24$, respectively, where 3 is the number of fixed effects in the model, $30 = 3 * 10$ is the number of elements in $\mathbf{b}_1$ and $240 = 3 * 8 * 10$ is the number of elements in the vector $\mathbf{b}_2$.

b. Write out the first five rows of the matrices $\mathbf{X}_i$ and $\mathbf{Z}_{1i}$ corresponding to measurements taken on site 1 in wafer 1. Assume that observations are ordered by voltage within site within wafer.

The first five rows of $\mathbf{X}_i, \mathbf{Z}_{1i}$ are given by:

$$\mathbf{X}_i = \begin{bmatrix} 1 & -0.8 & 0.64 \\ 1 & -0.4 & 0.16 \\ 1 & 0.0 & 0.0 \\ 1 & 0.4 & 0.16 \\ 1 & 0.8 & 0.64 \end{bmatrix}, \quad \mathbf{Z}_{1i} = \begin{bmatrix} 1 & -0.8 & 0.64 & 0 & \cdots & 0 \\ 1 & -0.8 & 0.64 & 0 & \cdots & 0 \\ 1 & -0.8 & 0.64 & 0 & \cdots & 0 \\ 1 & -0.8 & 0.64 & 0 & \cdots & 0 \\ 1 & -0.8 & 0.64 & 0 & \cdots & 0 \end{bmatrix}.$$

Part II

We fit Model 1 using the $N = 400$ measurements and assuming that the covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are diagonal, so that

$$\mathbf{\Sigma}_1 = \begin{bmatrix} \sigma_{10}^2 & 0 & 0 \\ 0 & \sigma_{11}^2 & 0 \\ 0 & 0 & \sigma_{12}^2 \end{bmatrix}, \quad \mathbf{\Sigma}_2 = \begin{bmatrix} \sigma_{20}^2 & 0 & 0 \\ 0 & \sigma_{21}^2 & 0 \\ 0 & 0 & \sigma_{22}^2 \end{bmatrix}.$$

Refer to the output labeled **Output for Model 1** shown below. The R command used to produce the output is

```
model1 = lme(current ~ voltage + I(voltage^2), data = Wafer2, random=
   list(Wafer=pdDiag(~voltage + I(voltage^2)), Site=pdDiag(~voltage + I(voltage^2))))
```

Output for Model 1

```
        AIC      BIC   logLik
  -198.7254 -158.886 109.3627

Random effects:
 Formula: ~x + I(x^2) | Wafer
 Structure: Diagonal
        (Intercept)      x        I(x^2)
StdDev:   0.30198     0.21142    1.9745e-06

 Formula: ~ x + I(x^2) | Site %in% Wafer
 Structure: Diagonal
        (Intercept)    voltage   I(voltage^2)  Residual
StdDev:   0.22372      0.04092    8.0651e-06   0.12492
```

```
Fixed effects: y ~ x + I(x^2)
                Value  Std.Error  DF   t-value p-value
(Intercept)    7.9804   0.0991   318   80.4515       0
x              9.6486   0.0679   318  142.0645       0
I(x^2)         1.1704   0.0233   318   50.1660       0
 Correlation:
             (Intr)   x
x             0.000
I(x^2)       -0.075  0.000

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
    -1.6365     -0.8681      0.0480      0.6798      2.1269

Number of Observations: 400
Number of Groups:
          Wafer Site %in% Wafer
             10              80
```

3. Compute an estimate of the expected current intensity and its standard error for wafers to which we apply voltage equal to 1.5 V.

For $v = 1.5$, the centered predictor is $x = 1.5 - 1.6 = -0.1$. The MLE of the vector of fixed regression coefficients is $\hat{\boldsymbol{\beta}}$, with estimated covariance matrix $\hat{\boldsymbol{\Sigma}}_\beta$. From Output 1 above, we get

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 7.980 \\ 9.649 \\ 1.170 \end{bmatrix}, \quad \hat{\boldsymbol{\Sigma}}_\beta = \begin{bmatrix} 0.0098 & 0 & -0.0002 \\ 0 & 0.0046 & 0 \\ -0.0002 & 0 & 0.0005 \end{bmatrix}.$$

Therefore, the predicted mean response and SE for wafers subjected to 1.5 V is

$$\begin{aligned} \hat{y}(1.5V) &= 7.980 + 9.649(-0.1) + 1.170(-0.1)^2 = 7.027 \text{ mA}, \\ SE(\hat{y}(1.5V)) &= \begin{bmatrix} 1 & -0.1 & 0.01 \end{bmatrix} \begin{bmatrix} 0.0098 & 0 & -0.0002 \\ 0 & 0.0046 & 0 \\ -0.0002 & 0 & 0.0005 \end{bmatrix} \begin{bmatrix} 1 \\ -0.1 \\ 0.01 \end{bmatrix} \\ &= 0.099. \end{aligned}$$

4. **a**. Find the REML estimates of $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ and σ^2.

From Output 1, we find that:

$$\hat{\boldsymbol{\Sigma}}_1 = \begin{bmatrix} 0.30198^2 & 0 & 0 \\ 0 & 0.21142^2 & 0 \\ 0 & 0 & 1.97^2 \times 10^{-12} \end{bmatrix}, \; \hat{\boldsymbol{\Sigma}}_2 = \begin{bmatrix} 0.22372^2 & 0 & 0 \\ 0 & 0.04092^2 & 0 \\ 0 & 0 & 8.06^2 \times 10^{-12} \end{bmatrix},$$

and $\hat{\sigma}^2 = 0.124^2$.

b. Comment on the relative sizes of the REML estimates of the standard deviations of the random effects, and interpret what they mean in the context of this problem.

The covariance matrices are diagonal by assumption. $\hat{\boldsymbol{\Sigma}}_2$ represents the variability in the response between sites in the same wafer and $\hat{\boldsymbol{\Sigma}}_1$ represents the variability in the response between wafers. The variance of the random intercepts is higher than the site-to-site and wafer-to-wafer variability in the linear effect of voltage on the response. The variability between sites and between wafers in the quadratic effect of voltage appears to be negligibly small, at least in this model. The within-site error variance $\hat{\sigma}^2$ is moderately high, even though we have included multiple random terms in the model.

5. Figure 2 shows the within-group residuals e_{ijk} plotted against centered voltage, by wafer and site combination. The within-group residuals are computed as the difference between the observed response and the within-group fitted value. What do you conclude from inspection of these residual plots?

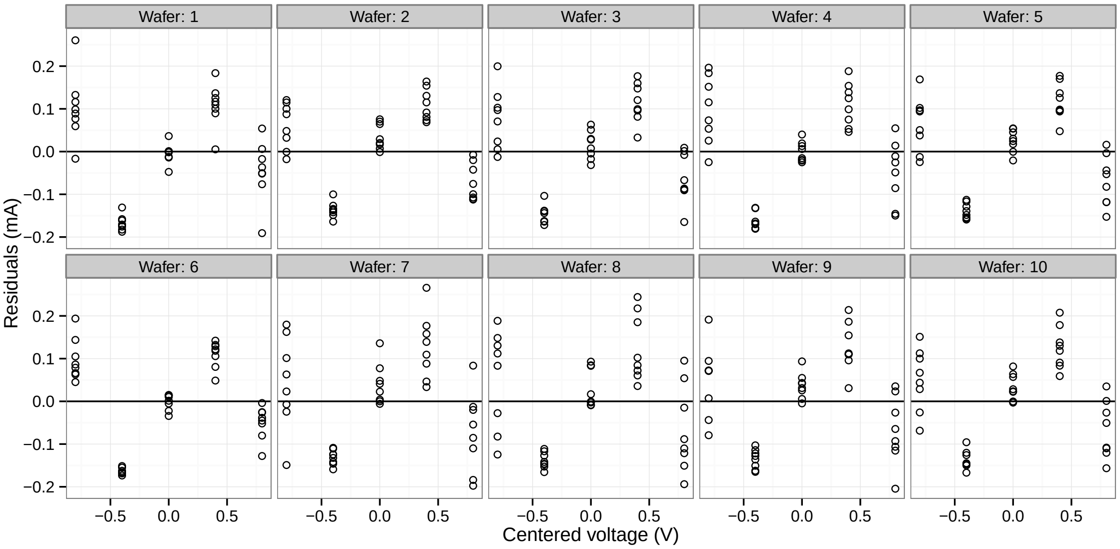

Figure 2: *Residuals for Model 1 plotted against centered voltage. There is one panel for each wafer, and in each panel there are 5 sets (each of 8) residuals, one set for each measurement site.*

It is clear from the plot that residuals are not randomly scattered about zero. There is a systematic wave-like association between the estimated residuals and voltage, suggesting that the regression model should include more than the linear and quadratic terms in voltage. One plausible approach would be to add terms in cosine and sine of voltage to the model that will permit accounting for the apparent periodicity in estimated residuals.

Part III

We extended Model 1 by adding two more fixed effects to the model:

$$Y_{ijk} = (\beta_0 + b_{0i} + b_{0ij}) + (\beta_1 + b_{1i} + b_{1ij})x_k + (\beta_2 + b_{2i} + b_{2ij})x_k^2 + \beta_3 \cos(x_k) + \beta_4 \sin(x_k) + \epsilon_{ijk}, \qquad \text{(Model 2)}$$

where β_3, β_4 are unknown regression coefficients. Once again we assume that the covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ are diagonal, and fit the model to the vector of 400 responses. Refer to **Output for Model 2** on Page 7. The R command used to fit Model 2 is

```
model2 = lme(current ~ voltage + I(voltage^2) + cos(voltage) + sin(voltage),
    data = Wafer2, random=list(Wafer=pdDiag(~voltage + I(voltage^2)),
    Site=pdDiag(~voltage + I(voltage^2))))
```

Output for Model 2

```
        AIC       BIC    logLik
  -1189.308   -1141.562 606.6541

Random effects:
 Formula: ~x + I(x^2) | Wafer
 Structure: Diagonal
        (Intercept)    x     I(x^2)
StdDev:    0.3161    0.2114   0.0483

 Formula: ~x + I(x^2) | Site %in% Wafer
 Structure: Diagonal
       (Intercept)   x      I(x^2)    Residual
StdDev:   0.2491   0.1066   0.0611     0.00914

Fixed effects: y ~ x + I(x^2) + cos(x) + sin(x)
                  Value  Std.Error   DF   t-value    p-value
(Intercept)   -1.9954   0.3579     316  -5.5753       0
x              4.7621   0.0720     316  66.1136       0
I(x^2)         5.8802   0.1625     316  36.1653       0
cos(x)         9.9972   0.3432     316  29.1236       0
sin(x)         5.3576   0.0263     316 203.7107       0
 Correlation:
             (Intr)   x    I(x^2) cs(x)
x             0.000
I(x^2)       -0.952  0.000
cos(x)       -0.957  0.000  0.995
sin(x)        0.000 -0.333  0.000  0.000

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3        Max
     -2.0932      -0.3116    -0.0175       0.3898     3.0480

Number of Observations: 400
Number of Groups:
          Wafer Site %in% Wafer
             10              80
```

6. **a**. Does adding the cosine and sine terms in centered voltage improve the fit of the model? Justify your answer.

Yes, adding the two new cosine and sine terms improves the fit of the model. There are several indications that this is the case:

- The marginal $t-$statistics for testing whether $\beta_3 = 0$ and $\beta_4 = 0$ given that all other terms are already in the model are both highly significant, meaning that we have strong evidence in favor of including the cosine and the sine of voltage in the model.

- The estimated variance of residuals is about 12 times smaller in Model 2 than in Model 1.
- Both AIC and BIC are smaller in Model 2 than in Model 1.

b. Is it possible, given the current output, to conduct a likelihood ratio test (LRT) to compare Model 1 and Model 2? If not, explain why not.

Even though the fixed effects structure of Model 1 is nested within that of Model 2, the LRT cannot be applied here. This is because the variance components in the model were estimated using REML rather than ML. The restricted likelihood that is maximized to obtain REML estimates of the variances corresponds to the likelihood associated with a set of $N - p$ linearly independent error contrasts, where p is the rank of the $\boldsymbol{X}$ matrix. In Model 1, $p = 2$ whereas in Model 2, $p = 4$. Therefore, for different number of fixed effects in the model, the vector of error contrasts on which the REML estimates of the variances are based also change.

7. Figure 3 shows the within-group residuals e_{ijk} for Model 2, plotted against centered voltage, by wafer and site combinations. What do you conclude from inspection of these residual plots? How do these residuals compare to the Model 1 residuals shown in Figure 2?

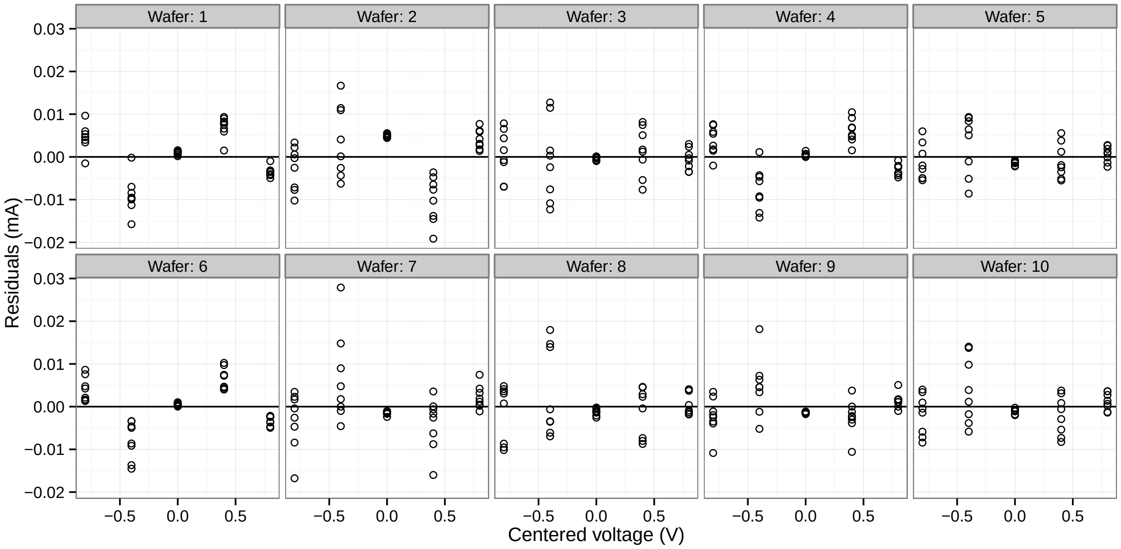

Figure 3: *Residuals for Model 2, plotted against voltage.There is one panel for each wafer, and in each panel there are 5 sets (each of 8) residuals, one set for each measurement site.*

The residuals in Fig. 3 still exhibit some systematic periodicity at least in some of the wafers, but overall they indicate that Model 2 fits the data better than Model 1. To improve the fit of the model we could consider two extensions to Model 2:

- Add wafer-level random effects b_{13i}, b_{14i} to accommodate the wafer-to-wafer variability in the periodicity of the response.

- Allowing for a less restrictive function to model the periodic behavior. For example, we could consider allowing the frequency of the cosine and sine waves to be different from 1 (which is what we assumed in Model 2). To do so, the cosine and sine terms would include an additional parameter ω that represents frequency so that the two additional terms in the model would look like $\beta_3 \cos(\omega x_k)$ and $\beta_4 \sin(\omega x_k)$. While the added flexibility gained by including ω might improve the fit of the model, estimation is now complicated by the fact that the new model would be non-linear in ω and voltage.

8. Figure 4 shows the predicted values of the random coefficients $\hat{\mathbf{b}}_i$ associated with wafer, plotted against each other. Comment on what the plot indicates about the assumptions associated with Model 2.

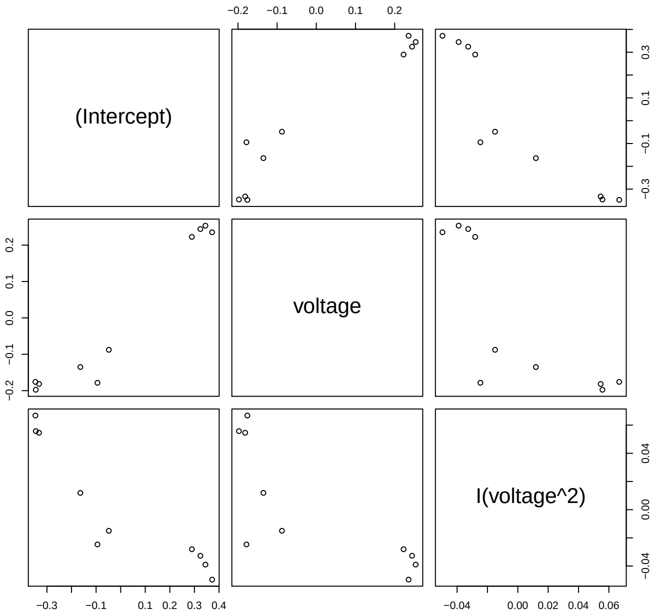

Figure 4: *Scatterplot matrix of predicted wafer-level random effects.*

Fig. 4 shows the estimated wafer-level random intercepts, linear terms and quadratic terms plotted against each other. It is apparent from the figure than the assumption of uncorrelated random effects at the level of wafer is not appropriate. From the figure, we learn that a general covariance matrix $\mathbf{\Sigma_1}$ rather than a diagonal matrix may be more plausible for these data.

Part IV

A third model fit to the 400 measurements was Model 2 modified in that the covariance matrix $\mathbf{\Sigma}_1$ is now assumed to be a general positive definite matrix with unknown elements σ^2_{1lm} for $l, m = 0, 1, 2$. We refer to this model as Model 3. The R command used to produce the output for Model 3 is

```
model3 = lme(current ~ voltage + I(voltage^2) + cos(voltage) + sin(voltage),
     data = Wafer2, random=list(Wafer=(~voltage + I(voltage^2)),
     Site=pdDiag(~voltage + I(voltage^2))), control=c1)
```

Output for Model 3

```
        AIC       BIC   logLik
  -1220.638 -1160.954 625.3188

Random effects:
 Formula: ~x + I(x^2) | Wafer
 Structure: General positive-definite, Log-Cholesky parametrization
             StdDev     Corr
(Intercept)  0.3240  (Intr)   x
x            0.2118  0.981
I(x^2)       0.0498 -0.944 -0.862

 Formula: ~x + I(x^2) | Site %in% Wafer
 Structure: Diagonal
        (Intercept)      x      I(x^2)        Residual
StdDev:   0.2397      0.1058   0.0600          0.0091

Fixed effects: y ~ x + I(x^2) + cos(x) + sin(x)
                 Value Std.Error  DF   t-value p-value
(Intercept)  -1.9954    0.3586   316  -5.5640       0
x             4.7621    0.0721   316  66.0108       0
I(x^2)        5.8802    0.1626   316  36.1459       0
cos(x)        9.9972    0.3433   316  29.1149       0
sin(x)        5.3576    0.0263   316 203.6501       0
Correlation:
             (Intr)     x   I(x^2)  cs(x)
x             0.260
I(x^2)       -0.976 -0.078
cos(x)       -0.955  0.000  0.994
sin(x)        0.000 -0.333  0.000  0.000

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
     -2.0717      -0.3288    -0.0182       0.4001     3.0491

Number of Observations: 400
Number of Groups:    Wafer Site %in% Wafer
                      10             80
```

9. Carry out a likelihood ratio test to compare Model 2 and Model 3. Use a Type I error probability $\alpha = 0.05$. Interpret the results of the test in the context of the problem.

Since Model 2 and Model 3 have the same fixed effects structure and differ only in the model for $\mathbf{\Sigma_1}$, and since Model 2 is nested within Model 3, we can conduct a LRT. From Output 2, we find that the value of the log-likelihood function at the MLE is $\log(L_2) = 606.6541$ and the number of parameters estimated in Model 2 is 12. For Model 3, the log-likelihood at the maximum is $\log(L_3) = 625.3188$ and the number of parameters in Model 3 is 15. We know that

$$2[\log(L_3) - \log(L_2)] \dot{\sim} \chi^2_{15-12}.$$

Here, $2[\log(L_3) - \log(L_2)] = 37.33$ larger than the upper $99th$ percentile of a χ^2 distribution with 3 degrees of freedom. Therefore, we conclude that Model 3, with a general covariance structure for the wafer-level random effects fits the data better than Model 2.

10. Report what you have learned from your analysis to the Director of the Microelectronics Division at Lucent Technologies. In no more than 100 - 120 words, summarize what you have learned from this analysis regarding manufacturing variability in wafers.

Results from the study conducted using a random sample of 10 n-channel devices (wafers) revealed several important issues in the manufacturing of these devices at Lucent Technologies. In summary, we found that

- As expected, there is increased intensity of current traveling through the device when the voltage applied to the wafer increases. While at first glance, the association between intensity of current and voltage appears to be quadratic, we found that voltage level induces a periodicity in the response. Our findings indicate that while this effect is significant, it can be accommodated by fitting a model that includes a cosine and a sine term in voltage.
- There is non-negligible variability between wafers and even withing wafers in the intensity of the current even at the same voltage level. While the variability between wafers is larger than the variability within wafers (between sites) both are several orders of magnitude larger than the random experimental variability. Therefore, this would suggest that the manufacturing process for the MOS devices should be reviewed with an eye towards increasing the homogeneity of the wafers and of the components of the wafers.

Since 1980, when President Jimmy Carter signed a law making small-scale brewing legal, the craft beer industry has exploded in the United States, and was estimated to have accounted for just over 14 billion dollars in retail sales in 2013. Most major universities now offer courses on one or more aspects of the brewing process. The University of California at Davis and Oregon State University both offer graduate degrees in the science of fermentation and brewing. Even ISU has a course titled "The Biochemistry of Beer." Despite this academic activity, however, quantitative analyses of various biological and physical processes involved in brewing have not been well developed. This question centers on the analysis of data from one study connected with the activity of *Saccharomyces cerevisiae*, the species of yeast that converts carbohydrates to carbon dioxide and alcohol (i.e., fermentation).

Efficient fermentation requires that the yeast added to fermentable liquid (called wert) contain an adequate supply of glycogen (the primary form of glucose that is used by living organisms to provide energy). If a brewer were to know the glycogen content of his or her yeast prior to adding it to the wert he or she would be able to adjust the quantity of yeast needed, saving time and money in the production process. But analyzing samples of yeast for glycogen concentration is a complicated bio-chemical procedure requiring special equipment, and is thus prohibitive in cost for most craft brewers. A much simpler measurement was proposed, based on iodine staining and measuring absorbtion spectra (Quain, D.E. and Tubb, R.S. 1983, A rapid and simple method for the determination of glycogen in yeast. *Journal of the Institute of Brewing* **89**:38-40). The experimental procedure was to grow yeast cultures for 48 h under various environmental conditions that should produce a range of glycogen synthesis by the yeast. Each culture was then measured with the complex but accepted procedure to determine glycogen concentration (in mg/ml) and also with the simple absorbance method being proposed in the paper. The data from this paper are shown in Figure 1 in which chemically determined glycogen concentrations (in mg/ml) are given on the horizontal axis and absorbance (at 660nm wavelength) from the proposed method is given on the vertical axis. Absorbance is technically unitless (it is a ratio of amounts of radiation) but can be reported relative to a reference solution, as is done in Figure 1, explaining why values greater than 1 appear. A total of 45 yeast cultures are represented in this figure.

Non-Bayesian Analysis

Let $Y_i;\ \ i = 1, \ldots, n$ be random variables associated with the values of absorbance read for yeast culture i, and let $X_i;\ \ i = 1, \ldots, n$ be random variables associated with the chemically determined glycogen concentrations for those same cultures. We model both absorbance and chemical glycogen as random variables because neither of these quantities were controlled in the experiment, both resulting from the various environmental conditions under which yeast cultures were grown.

Although this figure is the central result of their paper, Quain and Tubb discuss it directly in only a few brief sentences (p. 39 of that paper):

> Analysis of least squares gave a correlation coefficient of 0.95 for the two methods of measurement. Therefore, staining yeast cells with iodine can be used as a rapid and simple method for determining glycogen in yeast. From the data in Fig. 1(a), glycogen concentrations ($x\, mg\, ml^{-1}$) can be calculated by substituting for y (absorbance at 660nm) in the following equation: $x = (y - 0.26)/1.48$.

ANSWER QUESTIONS 1 AND 2 NOW

(Questions begin on page 10.)

For the remainder of this entire prelim question, ignore any difficulties with the approach of Quain and Tubb that you may have identified in questions 1 and/or 2. Assume that we are able to appropriately formulate the problem in the following manner. Define random variables Y_i to be connected with the absorbance of yeast culture i; $i = 1, \ldots, n$ and covariates $x_i;\ \ i = 1, \ldots, n$ as the chemically obtained glycogen concentrations for those same cultures. Suppose our interest is in estimating a 90% prediction interval for absorbance at a given level of chemically determined glycogen, with special interest in the upper endpoint of such an interval. Assume further that we wish to model the relation between these variables with an additive error model of the form

$$Y_i = \mu_i + \sigma g(\mu_i, \theta)\epsilon_i; \quad i = 1, \ldots, n, \tag{1}$$

where $g(\cdot)$ is a known function, any parameters that may appear in this function besides those involved in μ_i (i.e., θ) are considered known, the ϵ_i are assumed to be independent and identically distributed with standard normal distributions, and

$$\mu_i = \beta_0 + \beta_1 x_i \tag{2}$$

where β_0 and β_1 are unknown parameters.

A question with this model is whether the data provide any evidence that $g(\mu_i, \theta)$ should be anything other than identically equal to 1, giving a model with constant variances. A standard Box-Cox plot was constructed by dividing the data of Figure 1 into 8 equally spaced bins defined on the basis of chemically determined glycogen concentrations, the covariates of model (2). This plot is presented in Figure 2 and an ordinary least squares slope value for this plot is -0.385. If we were to accept that a straight line reasonably describes the pattern of points in Figure 2 we might formulate a power-or-the-mean model as in (1) and (2) where

$$g(\mu_i, \theta) = \mu_i^{\theta}, \tag{3}$$

and with $\theta = -0.40$.

This power-of-the-mean model (with $\theta = -0.40$ fixed) was fit to the data, resulting in the estimates $\hat{\beta}_0 = 0.23$, $\hat{\beta}_1 = 1.61$, and $\hat{\sigma}^2 = 0.00506$. Estimates from a model with constant variance were $\hat{\beta}_0 = 0.24$, $\hat{\beta}_1 = 1.59$, and $\hat{\sigma}^2 = 0.00753$. NOTE: THESE ESTMATES DIFFER SLIGHTLY FROM THE VALUES REPORTED BY QUAIN AND TUBB (1983) BECAUSE THE DATA USED IN FIGURE 1 HAD TO BE RECONSTRUCTED FROM THE FIGURE IN THEIR PAPER (they didn't actually give the numerical values). BUT WE ARE NO LONGER CONCERNED WITH THAT PAPER, ONLY FITTING REGRESSION MODELS TO THE DATA OF FIGURE 1. A plot of standardized residuals for the power-of-the-mean model is shown in Figure 3A and one for the model with constant variance is shown in Figure 3B.

ANSWER QUESTION 3 NOW

(Questions begin on page 10.)

It may be instructive to more closely examine the values used to construct the Box-Cox plot of Figure 2, which serves as our primary evidence that a model with nonconstant variances should be considered. The sample means, variances and sizes of the binned values that were used to construct Figure 2 are reported in Table 1. Values for bins 5 and 7 were not used in Figure 2.

ANSWER QUESTION 4 NOW

(Questions begin on page 10.)

At this point our challenge is to determine whether we would prefer a model with constant variances or a model with non-constant variances and, in particular, a power-of-the-mean model. There are a number of approaches one might consider for meeting this challenge. One of these would be to design a simulation-based model assessment to choose between the competing models. Another would be to consider the power-of-the-mean model with an unknown value of the power θ in (4). The parameters to be estimated would then consist of β_0, β_1, σ^2, and θ.

ANSWER QUESTIONS 5, 6, 7 AND 8 NOW

(Questions begin on page 10.)

Bayesian Analysis

If we are willing to assume normal distributions for the response random variables we might also consider estimation and inference from a Bayesian perspective. Consider the model, for $i = 1, \ldots, n$ and assuming independence everywhere,

$$\begin{aligned} Y_i &= \mu_i + \sigma \mu_i^{\theta} \epsilon_i, \\ \mu_i &= \beta_0 + \beta_1 x_i \\ \epsilon_i &\sim N(0,1) \\ \beta_0 &\sim N(B_0, V_0) \\ \beta_1 &\sim N(B_1, V_1) \\ \sigma^2 &\sim \text{InvGamma}(a, b) \\ \theta &\sim \text{Uniform}(-c, c) \end{aligned} \tag{4}$$

A typical approach to analysis with this model would involve a Markov Chain Monte Carlo simulation procedure, most likely with the overall structure of a Gibbs Sampling algorithm. There might, however, arise the need for Metropolis-Hastings steps within this overall structure, or the use of some other procedure to sample from unnormalized distributional forms.

ANSWER QUESTIONS 9, 10 AND 11 NOW

(Questions begin on page 10.)

A Fundamental Issue

Regardless of whether one takes a Bayesian or a non-Bayesian approach to analysis, the problem described in this question can be used to motivate a consideration of a fundamental issue in model selection and assessment. Although there are no standard terms for different schools of thought relative to these issues, there are two threads of reasoning that we might call *Pure Model Assessment* and *Pure Model Selection.* Pure Model Assessment might be thought of in analogy with Ronald Fisher's *test of significance* in which there is a null hypothesis, but no alternative hypothesis is specified. In contrast, Pure Model Selection can only be put into practice if there are (at least) two definite models between which we must choose. This approach is perhaps most clearly illustrated by penalized likelihood criteria such as Akaike's Information Criterion. Likelihood ratio tests and Bayes Factors are technically really Pure Model Selection procedures, but can incorporate a bit of Pure Assessment flavor if the evidence in favor of a "alternative" model over a "null" model must be overwhelming before the null model is rejected in favor of the alternative model.

ANSWER QUESTIONS 12 AND 13 NOW

(Questions begin on page 10.)

Figures

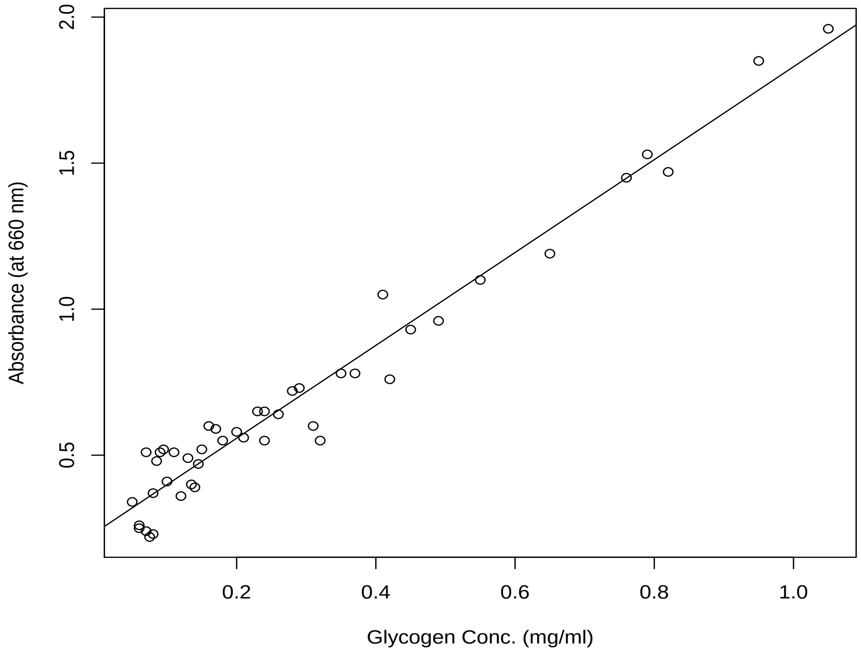

Figure 1: Scatterplot of iodine absorbance in yeast (vertical axis) and chemical glycogen concentration (horizontal axis). Data reconstructed from Figure 1(a) of Quain and Tubb, 1983. Line shown is the estimated expectation function from fitting the model of expressions (1), (2) and (3) to the reconstructed data.

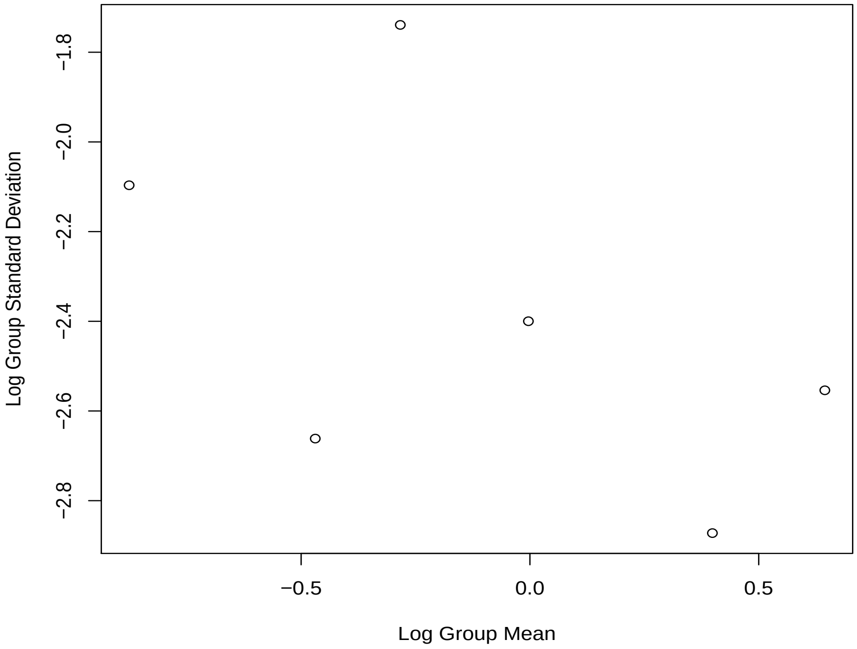

Figure 2: Box-Cox plot for the data of Figure 1.

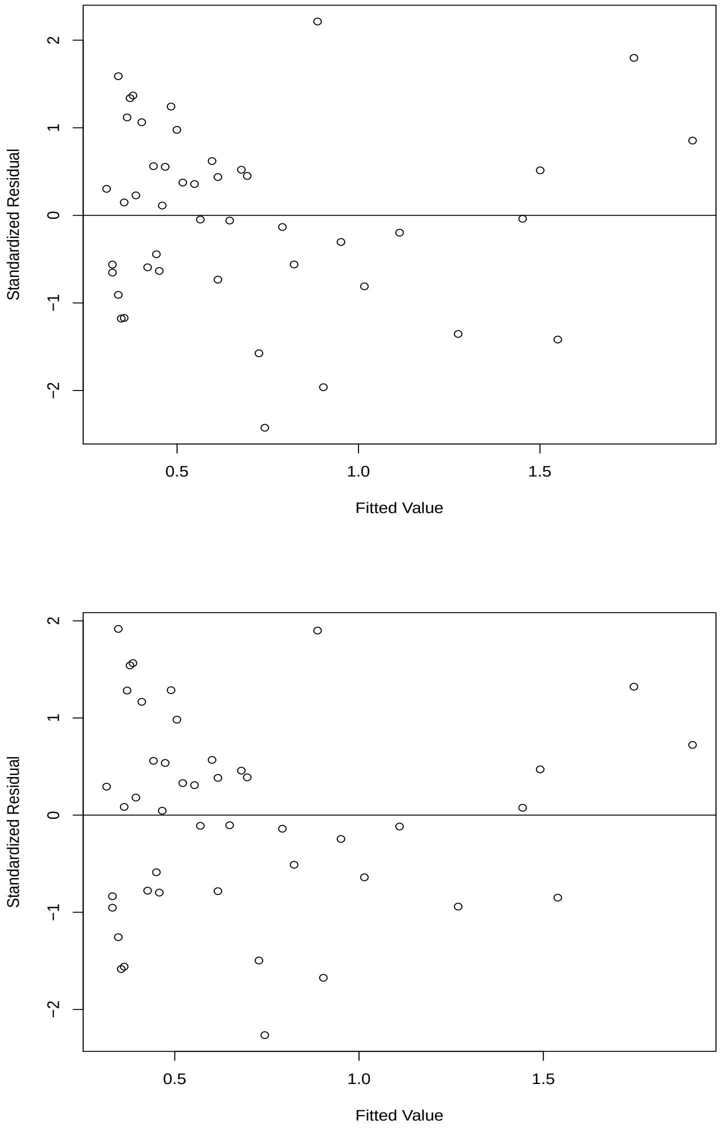

Figure 3: Residual plots for models with variances decreasing with mean (upper panel) and constant variances (lower panel).

Tables

Bin	Mean	Variance	Size
1	0.4165	0.0151	20
2	0.6255	0.0049	9
3	0.7533	0.0308	6
4	0.9967	0.0082	3
5	1.1900	NA	1
6	1.3900	0.0392	2
7	1.4700	NA	1
8	1.9050	0.0060	2

Table 1: Values of group means, variances, and sizes for bins used to construct a Box-Cox plot for the data of Figure 1.

Questions

1. In terms of the random variables Y_i and X_i defined at the start of the section titled "Non-Bayesian Analysis", give the model that was used by Quain and Tubb in their analysis of the data of Figure 1.

   Hint: While this is meant to be a rather straightforward question, be careful to use precise notation for dealing with random and non-random quantities and be clear if parts of an analysis are formulated only conditionally.

2. Comment on the recomendation of Quain and Tubb that glycogen concentration can be determined through the use of the expression $x = (y - 0.26)/1.48$. You may make use of any notation you introduced in Question 1 if you wish.

   Hint: If X and Y have a bivariate normal distribution with $E(X) = \mu_x$, $E(Y) = \mu_y$, $var(X) = \sigma_x^2$, $var(Y) = \sigma_y^2$, and $cov(X, Y) = \sigma_{xy}$, then $\rho = cor(X, Y) = \sigma_{xy}/\sqrt{\sigma_x^2\sigma_y^2}$ and the conditional expectation of Y may be written as $E(Y|X = x) = \mu_y + \rho\sigma_y/\sigma_x(x - \mu_x)$.

3. The estimates of regression coefficients β_0 and β_1 are quite similar between the model with nonconstant variances and the model with constant variances. The standardized residual plots of Figure 3 are close to visually identical (if you look closely you will be able to detect some minor differences, but that is all). This might lead us to the conclusion that it does not really matter which model we choose to make use of. Why is this not really the case?

   Hint: recall our current objective of obtaining prediction intervals as described in the paragraph preceeding expression (1). To make things concrete, consider addressing this objective for estimated expectations of $\hat{\mu}_i = 0.25$ and $\hat{\mu}_i = 1.5$.

4. What concern do the values in Table 1 raise relative to the motivation for formulating our power-of-the-mean model?

   Hint: It may also be helpful to re-examine the scatterplot of Figure 1.

5. Outline a simulation-based assessment procedure that would help determine whether we should prefer a model with constant variances or a power-of-the-mean model (with fixed $\theta = -0.40$). This should be done in a manner that would allow someone to take your outline and turn it into a practical computational procedure.

6. Briefly describe how we could estimate model parameters if θ is included in the set

to be estimated, but we are unwilling to assume a full distributional form for the problem. That is, describe in algorithmic form how the pseudo-likelihood approach of Carroll and Ruppert would be applied to this problem. Explain why we would not be able to examine uncertainty in the estimated value of θ under this procedure (if we follow the prescription of Carroll and Ruppert).

7. Briefly describe how we could estimate model parameters if θ is included in the set to be estimated if we are willing to assume normal distributions for the response variables $Y_i; \quad i = 1, \ldots, n$. Explain how we might then assess evidence against the possibility that $\theta = 0$.

8. Recall the stated objective of constructing a prediction interval for a new value of Y_j at a particular covariate level x_j (which may or may not be included in our set of observed values). Consider this problem using the power-of-the-mean model with an estimated value of θ. What would be a simple, but naive, procedure for determining a 95% prediction interval at covariate value x_j? Briefly explain how one could determine a true coverage rate for this method of arriving at a prediction interval.

9. For a model with constant variances, normal priors for β_0 and β_1 would lead to conditional conjugacy. That is, the full conditional posteriors for β_0 and β_1 would also have the form of normal distributions. By deriving the full conditional posterior for β_0 demonstrate that this is no longer true for model (4).

10. For a model with constant variances, an Inverse Gamma prior for σ^2 would also lead to conditional conjugacy. Demonstrate that this remains true for model (4).
Hint: If a random variable X has a Gamma distribution with parameters a and b, then $Y = 1/X$ has an Inverse Gamma distribution with parameters a and b.

11. Consider basing the choice of the model with constant variances versus the power-of-the-mean model on a Bayes factor. The densities needed to form a Bayes Factor for evidence in favor of model (4) relative to a model with constant variances are not readily derived in closed form. Describe how you would compute a Bayes Factor for selecting between these two models using simulation. In your notation, let $\boldsymbol{y}^* = y_1^*, \ldots, y_n^*$ denote the actual data values.

12. In a short paragraph, support the Pure Model Assessment viewpoint for this problem

and argue that unless a simple linear regression with constant variances can be shown demonstrably inadequate for representing the data, the power-of-the-mean model should not be considered.

13. In a short paragraph, support the Pure Model Selection viewpoint for this problem and argue that if the power-of-the mean model with fixed power θ can be shown at all superior to the simple linear regression it should be preferred.

These are a sketch of the answers hoped for. Other possibilities might exist for some of the questions that would be entirely adequate if they are both technically correct and logically consistent.

Question 1. Using Y_i as a random variable associated with absorbance and X_i a random variable associate with chemically determined glycogen concentration for yeast culture $i = 1, \ldots, n$, the model used by Quain an Tubb may be writen as, for $i = 1, \ldots, n$,

$$\begin{aligned} E(Y_i|X_i = x_i) &= \beta_0 + \beta_1 x_i \\ var(Y_i) &= \sigma^2. \end{aligned}$$

Question 2. In the notation of Question 1, Quain and Tubb obtained ordinary least squares estimates $\hat{\beta}_) = 0.26$ and $\hat{\beta}_1 = 1.48$. Their recommendation is then to predict X_i given that $Y_i = y_i$ as

$$E(X_i|\hat{Y_i} = y_i) = [y_i - \hat{\beta}_0]/\hat{\beta}_1.$$

The minimum mean squared error predictor will be the expected value $E(X_i|Y_i = y_i)$, while what has been estimated is the expected value $E(Y_i|X_i = x_i)$. From the hint,

$$E(Y_i|X_i = x_i) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x_i - \mu_x),$$

so that, in terms of the parameters of the model of Question 1,

$$\begin{aligned} \beta_0 &= \mu_y - \rho\frac{\sigma_y}{\sigma_x}\mu_x \\ \beta_1 &= \rho\frac{\sigma_y}{\sigma_x} \end{aligned}$$

and, again from the bivariate normal form,

$$\begin{aligned} E(X_i|Y_i = y_i) &= \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y_i - \mu_y) \\ &= [\beta_0 - \mu_y]/\beta_1 + \beta_1\frac{\sigma_x^2}{\sigma_y^2}(y_i - \mu_y), \end{aligned}$$

which cannot be equated with the form of the estimator of Quain and Tubb.

The error of Quain and Tubb was to assume that the model of Question 1 (which they never actually write) was

$$E(Y_i|X_i = x_i) = \beta_0 + \beta_1 E(X_i|Y_i = y_i),$$

which is a nonsensical statement.

Question 3. The stated objective is to formulate prediction intervals for the responses at given values of the covariate. Prediction intervals will depend on the estimated distributions of the responses at various levels of the covariate. According to the model given in (1), (2), and (3), the response variables Y_i have normal distributions with

$$\begin{aligned} E(Y_i) &= \beta_0 + \beta_1 x_i \\ var(Y_i) &= \sigma^2 \mu_i^{2\theta} \end{aligned}$$

While the expected values are modeled in the same way as those in the model with constant variances, the variances differ substantially. For example, at the estimated expectations $\hat{\mu}_i = 0.25$ and $\hat{\mu}_i = 1.5$ given in the hint, and with $\hat{\sigma}^2 = 0.00506$ and $\theta = -0.40$, the variances would be estimated as

$$\begin{aligned} v\hat{a}r(Y_i) = 0.00881 \quad \text{for } \hat{\mu}_i = 0.25 \\ v\hat{a}r(Y_i) = 0.00430 \quad \text{for } \hat{\mu}_i = 1.50, \end{aligned}$$

which would lead to substantial differences in the width of prediction intervals at these points.

Question 4. Concern is caused by the small number of observations falling into bins with larger means. Variance estimates for all of the bins with mean greater than about 0.75 (bin number 3) cannot be precise. This is caused by the sparse occurence of observations for covariate values greater than about 0.50, which is evident upon re-examination of the scatterplot in Figure 1. It is possible that this pattern of observations along with chance has produced what appears to be decreasing variability in Figure 1.

Question 5. Any number of simulation-based assessments could be constructed. The key elements are (1) determining whether one wishes to simulate from both models or just the model with constant variances - see Questions 12 and 13 – and (2) identifying a suitable measure to address the question of interest. The following are two possibilities:

(a) Simulate only from the fitted model with constant variances and using the observed values of covariates.

(b) For each simulated data set compute the slope for a Box-Cox plot constructed with the same bin definitions as used for the actual data. Denote these values for M simulated data sets as s_k^*; $k = 1, \ldots, M$. Denote the value for the actual data set as s_a^*.

(c) Compute a simulation-based $p-$value as

$$\begin{aligned} p_L &= \frac{1}{M} \sum_{k=1}^{M} I(s_k^* < s_a^*) \\ p_U &= \frac{1}{M} \sum_{k=1}^{M} I(s_a^* < s_k^*) \\ p &= \min\{p_l, p_U\}, \end{aligned}$$

where $I(A)$ is the indicator function that assumes a value of 1 if the statement A is true, and 0 otherwise.

Here, we would decline to reject the model with constant variances unless the $p-$value is suitably small (e.g., less than 0.05). Alternatively,

(a) Simulate from the fitted model with constant variances and using the observed values of covariates.

(b) For M simualted data sets, fit both the model with constant variances and the power-of-the mean model with fixed $\theta = -0.40$. Compute the ratio of estimated value of σ^2 as

$$r_k^* = \frac{\hat{\sigma}^2_{pom}}{\hat{\sigma}^2_{cv}},$$

where $\hat{\sigma}^2_{pom}$ is the value estimated from the power-of-the-mean model, and $\hat{\sigma}^2_{cv}$ is the value estimated from the model with constant variances. Let r_a^* denote the value of the ratio computed from the actual data.

(c) Compute a simulation-based $p-$value as

$$\begin{aligned} p_L &= \frac{1}{M}\sum_{k=1}^{M} I(r_k^* < r_a^*) \\ p_U &= \frac{1}{M}\sum_{k=1}^{M} I(r_a^* < r_k^*) \\ p_{cv} &= \min\{p_l,\, p_U\}, \end{aligned}$$

where $I(A)$ is the indicator function that assumes a value of 1 if the statement A is true, and 0 otherwise.

(d) Simulate from the fitted power-of-the-mean model, again using the observed covariate values.

(e) Repeat steps 2 and 3, denoting the resultant $p-$value p_{pom}.

Here, we would choose the model with constant variances if $p_{cv} > p_{pom}$ and the power-of-the-mean model if $p_{pom} > p_{cv}$.

Question 6. If we are unwilling to make a full distributional assumption we could use the pseudo-likelihood procedure of Carroll and Ruppert. In this procedure we alternate betweeen the following two steps:

(a) Using the likelihood that would correspond to a model based on normal distributions, update θ and σ^2 through maximization with regression parameters held fixed. If based on an iterative process, one step could be deemed sufficient.

(b) With updated values of θ and σ^2, update values of the regression parameters using a generalized least squares procedure.

(c) Repeat steps (a) and (b) until convergence.

The typical procedure is to then ignore uncertainty in the estimated value of θ and make inference based on the Fundamental Theorem of Generalized Least Squares.

Question 7. With a full distributional assumption (most likely normal) we could estimate all parameters simultaneously through an iterative procedure such as Newton-Raphson.

If we did so, one option for making inference about θ would be based on Wald theory using the estimated inverse observed information matrix. We could either construct a test of $\theta = 0$ or compute an interval estimate. Simultaneous maximum likelihood estimates could also be determined through the use of unscaled profile likelihood – where θ would be the natural choice for profiling. Inference could then be based on a normed profile likelihood. As usual, if 0 lies outside the interval we would reject the hypothesis that $\theta = 0$.

Question 8. Under the power-of-the-mean model the response random variables Y_i; $i = 1, \ldots, n$ are independent and have normal distributions with means $\mu_i = \beta_0 + \beta_1 x_i$ and variances $v_i = \sigma^2 \mu_i^{2\theta}$. Let $F(\psi)$ denote this distribution, where $\psi = (\beta_0, \beta_1, \sigma^2, \theta)$. Endpoints of a naive prediction interval would then be the 0.025 and 0.975 quantiles of the estimated distribution, $F(\hat{\psi})$ which we might denote as $F_{0.025}(\hat{\psi})$ and $F_{0.975}(\hat{\psi})$, respectively. This interval ignores uncertainty in parameter estimates, but the coverage could be adjusted through the use of a parametric bootstrap as follows.

(a) Simulate data from the fitted model, using the observed covariate values. Also simulate a new response value at a chosen covariate value, observed or unobserved. This value will be called the predictand and for a given simulaed data set will be denoted as $y_k^{(0)}$.

(b) For simulated data set $k = 1, \ldots, M$, estimate parameters as ψ_k^*, and compute a prediction interval with enpoints

$$\begin{aligned} L_k &= F_{0.025}(\psi_k^*) \\ U_k &= F_{0.975}(\psi_k^*) \end{aligned}$$

(c) Compute the actual coverage as

$$1 - \alpha' = \frac{1}{M} \sum_{k=1}^{M} I(L_K < y_k^{(0)} < U_k),$$

where $I(A)$ is the indicator function as in the solution to Question 5.

We would then report $(1-\alpha')100\%$ as the coverage of our prediction interval. Alternatively, especially if coverage differs markedly from 0.95 and computation is fast, we could adjust the nominal level of interval formulation to make the observed level close to 0.95.

Question 9. The full conditional posterior distribution of β_0 for model (5) may be derived as follows.

$$\begin{aligned} p(\beta_0|\cdot) &\propto \pi(\beta_0)\, f(\boldsymbol{y}|\beta_0,\beta_1,\sigma^2,\theta) \\ &\propto \exp\left[-\frac{1}{2V_0}(\beta_0-B_0)^2\right]\left[\prod_{i=1}^{n}(\beta_0+\beta_1 x_i)^{\theta}\right]^{-1}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{\frac{y_i-\beta_0-\beta_1x_i}{(\beta_0+\beta_1x_i)^{\theta}}\right\}^2\right] \end{aligned}$$

Because β_0 appears in the denominator outside the exponential and/or because there is no way to isolate β_0 in the second exponential expression above, it will not be possible to complete the square and obtain a form that is recognizable as a normal density.

Question 10. The full conditional posterior distribution for σ^2 in model (5) may be derived as follows.

$$\begin{aligned} p(\sigma^2|\cdot) &\propto \pi(\sigma^2)\, f(\boldsymbol{y}|\beta_0,\beta_1,\sigma^2,\theta) \\ &\propto \frac{1}{(\sigma^2)^{a+1}}\exp(-b/\sigma^2)\left[\prod_{i=1}^{n}\frac{1}{(\sigma^2)^{1/2}}\right]\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{\frac{y_i-\beta_0-\beta_1x_i}{(\beta_0+\beta_1x_i)^{\theta}}\right\}^2\right] \\ &\propto \frac{1}{(\sigma^2)^{a+(n/2)+1}}\exp\left[\frac{1}{\sigma^2}\left(-b-\frac{1}{2}\sum_{i=1}^{n}\left\{\frac{y_i-\beta_0-\beta_1x_i}{(\beta_0+\beta_1x_i)^{\theta}}\right\}^2\right)\right], \end{aligned}$$

which is in the form of the kernel of an Inverse Gamma density with parameters

$$a+\frac{n}{2} \qquad \text{and} \qquad b+\frac{1}{2}\sum_{i=1}^{n}\left\{\frac{y_i-\beta_0-\beta_1x_i}{(\beta_0+\beta_1x_i)^{\theta}}\right\}^2.$$

Question 11. Let $f_{cv}(\boldsymbol{y}|\beta_0,\beta_1,\sigma^2)$ denote the likelihood for the model with constant variances; call this model M_1. Let $f_{pom}(\boldsymbol{y}|\beta_0,\beta_1,\sigma^2,\theta)$ denote the likelihood for the power-of-the-mean model; call this model M_2. The joint prior for model M_1 is then

$$\pi_1(\beta_0,\beta_1,\sigma^2)=\pi(\beta_0)\pi(\beta_1)\pi(\sigma^2),$$

while that for model M_2 can be written as

$$\pi_2(\beta_0, \beta_1, \sigma^2, \theta) = \pi(\beta_0)\pi(\beta_1)\pi(\sigma^2)\pi(\theta).$$

With observed data $\boldsymbol{y}^*$ the Bayes factor in favor of the power-of-the-mean model (in favor of M_2) can be written as

$$BF(M_2, M_1) = \frac{h_1(\boldsymbol{y}^*)}{h_2(\boldsymbol{y}^*)} = \frac{\displaystyle\int f_{pom}(\boldsymbol{y}^*|\beta_0, \beta_1, \sigma^2, \theta)\, \pi_2(\beta_0, \beta_1, \sigma^2, \theta)\, d\beta_0\, d\beta_1\, d\sigma^2\, d\theta}{\displaystyle\int f_{cv}(\boldsymbol{y}^*|\beta_0, \beta_1, \sigma^2)\, \pi_1(\beta_0, \beta_1, \sigma^2)\, d\beta_0\, d\beta_1\, d\sigma^2}.$$

This Bayes Factor can be computed numerically using Monte Carlo approximations to both the numerator and denominator. This could be accomplished as follows.

(a) For $k = 1, \ldots, M$ independently simulate values $\beta_{0,k}$, $\beta_{1,k}$, σ_k^2, and θ_k from their individual priors.

(b) Approximate the numerator of the Bayes Factor as

$$h_{2,M}(\boldsymbol{y}^*) = \frac{1}{M}\sum_{k=1}^{M} f_{pom}(\boldsymbol{y}^*|\beta_{0,k}, \beta_{1,k}, \sigma_k^2, \theta_k)$$

and the denominator as

$$h_{1,M}(\boldsymbol{y}^*) = \frac{1}{M}\sum_{k=1}^{M} f_{cv}(\boldsymbol{y}^*|\beta_{0,k}, \beta_{1,k}, \sigma_k^2).$$

(c) The Bayes Factor is then computed as

$$BF_M(M_2, M_1) = \frac{h_{2,M}(\boldsymbol{y}^*)}{h_{1,M}(\boldsymbol{y}^*)}.$$

Note that there is no need for concern in using the same samples of the parameters in evaluation of the integrals as long as the Monte Carlo sample size M is sufficiently large.

Question 12. An argument in favor of Pure Model Assessment follows.

The central question is whether a model can adequately represent the observed data. If more than one such model exists, preference is given *a priori* to the simplest one based on the principle of parsimony (Note: this is not really Occam's Razor, although it is often presented as such). Thus, we should focus on the simple model, here the constant variance model, and only reject it if it can be shown to be inadequate. A more complex alternative should be considered only if the simple model is decidedly unable to represent the data, since making models more complex (within a common framework) cannot decrease the ability of models to fit data. If one considers alternatives, there is little reason to restrict attention to a power of the mean model. A host of alternatives could perhaps be constructed and choosing among them to find the "best" is an endless exercise.

Question 13. An argument in favor of Pure Model Selection follows.

The objective in this problem is prediction. If uncertainty in predictions is to be quantified, results might differ quite widely among alternatives. There is no logical force in favor of simplicity in this setting, and requiring that an alternative to a simple linear regression with constant variance must be overwhelmingly superior before being preferred carries a high risk of producing bad prediction intervals. Setting a reasonable, not overly stringent, threshold for how much better a power of the mean model must be before it would be preferred to the constant variance model serves as a compromise between avoiding needless complexity and identifying what might be important data patterns that can influence predictions.

A process engineer wishes to determine whether a change made to a chemical process has an important impact on the mean yield associated with a run of the process. A complicating issue in this regard is that each batch of raw material is sufficient to make only a few process runs, and different batches can be expected to have different characteristic yields.

Throughout this question, we will use a model for

$$y_{ijk} = \text{yield for run } k \text{ made using process } i \text{ and raw material batch } j$$

of the form

$$y_{ijk} = \mu_i + \beta_j + \varepsilon_{ijk} \qquad (*)$$

where the μ_i $(i = 1, 2)$ are unknown constants, the β_j are iid $\mathrm{N}\left(0, \sigma_\beta^2\right)$ independent of the ε_{ijk} that are themselves iid $\mathrm{N}\left(0, \sigma^2\right)$, and the variance components σ_β^2 and σ^2 are unknown constants.

Suppose initially that each raw material batch is sufficient to make 2 runs, and that 4 batches of raw material are available to the engineer for the study. Two possible plans for data collection are:

Plan I	**Plan II**
1 run from each raw material batch is made with each process (a total of 4 runs are made with each process)	2 raw material batches are dedicated to each process (a total of 4 runs are made with each process)

1. For

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

and

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix}$$

write out matrices **X** and **Z** so that the model (*) for 8 observations can be represented in usual matrix form

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

Do this first for **Plan I** and then for **Plan II**. In the first case, write the observations in the order

$$\mathbf{Y}=\begin{pmatrix} y_{111}\\ y_{211}\\ y_{121}\\ y_{221}\\ y_{131}\\ y_{231}\\ y_{141}\\ y_{241}\end{pmatrix}$$

In the second case, write the observations in the order

$$\mathbf{Y}=\begin{pmatrix} y_{111}\\ y_{112}\\ y_{121}\\ y_{122}\\ y_{231}\\ y_{232}\\ y_{241}\\ y_{242}\end{pmatrix}$$

2. What is the covariance matrix for $\mathbf{Y}$ (in the order indicated above) under **Plan 1**? Under **Plan 2**?

3. If all 4 unknown parameters were of some interest one might consider comparing **Plan 1** and **Plan 2** using appropriate 4×4 Fisher information matrices. Use the notation

$$\mathbf{D}\left(\mu_1,\mu_2,\sigma_\beta^2,\sigma^2\right) = \text{the } 4\times 4 \text{ Fisher Information matrix for } \mathbf{U} \sim \text{MVN}_2\left(\begin{pmatrix}\mu_1\\ \mu_1\end{pmatrix},\begin{pmatrix}\sigma^2+\sigma_\beta^2 & \sigma_\beta^2\\ \sigma_\beta^2 & \sigma^2+\sigma_\beta^2\end{pmatrix}\right)$$

$$\mathbf{E}\left(\mu_1,\mu_2,\sigma_\beta^2,\sigma^2\right) = \text{the } 4\times 4 \text{ Fisher Information matrix for } \mathbf{V} \sim \text{MVN}_2\left(\begin{pmatrix}\mu_2\\ \mu_2\end{pmatrix},\begin{pmatrix}\sigma^2+\sigma_\beta^2 & \sigma_\beta^2\\ \sigma_\beta^2 & \sigma^2+\sigma_\beta^2\end{pmatrix}\right)$$

$$\mathbf{F}\left(\mu_1,\mu_2,\sigma_\beta^2,\sigma^2\right) = \text{the } 4\times 4 \text{ Fisher Information matrix for } \mathbf{W} \sim \text{MVN}_2\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix},\begin{pmatrix}\sigma^2+\sigma_\beta^2 & \sigma_\beta^2\\ \sigma_\beta^2 & \sigma^2+\sigma_\beta^2\end{pmatrix}\right)$$

and write the Fisher Information matrices associated with **Plan I** and with **Plan II** in terms of these matrices.

4. Let $\bar{y}_{1..}$ be the arithmetic average of the process 1 observations and $\bar{y}_{2..}$ be the arithmetic average of the process 2 observations. What are the mean and standard deviation of

$$\bar{y}_{1..} - \bar{y}_{2..}$$

under **Plan 1**? Under **Plan 2**? (These can be found without appeal to matrix representations.)

5. For purposes of comparing μ_1 and μ_2, what does your answer to question **4** indicate about which of the two plans will typically be most effective?

Below are two hypothetical data sets, one corresponding to **Plan 1** and one corresponding to **Plan 2**.

Plan I		
Process	Batch	Yield
1	1	82.0
2	1	78.6
1	2	71.8
2	2	75.6
1	3	80.0
2	3	78.8
1	4	77.6
2	4	77.8

Plan II		
Process	Batch	Yield
1	1	82.0
1	1	79.2
1	2	71.9
1	2	76.3
2	3	79.3
2	3	78.9
2	4	77.0
2	4	77.8

6. For **both plans**, show the simple "by hand" calculations necessary to make valid/exact 95% two-sided t confidence intervals for $\mu_1 - \mu_2$.

7. Simple valid/exact 95% two-sided χ^2 confidence limits for σ^2 can be made from either set of hypothetical data above. Choose one of the plans and show the "by hand" calculations needed.

In the real application motivating this problem, practical constraints dictated that all runs from a given raw material batch had to be made consecutively, batches were of different sizes, and all runs from process 1 had to be made before runs from process 2. In fact, 4 small batches were dedicated to process 1, 1 larger batch was split between the two processes, and 1 batch of moderate size was dedicated to process 2. Attached to this question is an `R` printout useful in the analysis of the engineer’s data. Use it in answering the following questions.

8. Is there a statistically significant difference between the process means? Explain, referring carefully to appropriate items on the printout.

9. How does run-to-run variability in yield appear to compare with batch-to-batch variability? Explain, again referring to appropriate items on the printout.

10. The engineer in charge of this study says to you "We need to redo this study. We'll need to run process 1 before process 2. I can get raw material batches big enough to make as many as $r = 10$ runs per batch. We'll run the same number of batches, l, with each process (splitting no batch between processes). I want to estimate $\mu_1 - \mu_2$ to within .5. I'd like to minimize the total number of runs made

$$\text{total runs made} = 2lr$$

in meeting this goal. How many batches should we use for this study, and how many runs per batch should we make?"

Find this person appropriate values of r and l on the basis of the estimates on the printout.

R Printout

```
> data
   process batch     y
1        1     1 82.72
2        1     1 78.31
3        1     1 82.20
4        1     1 81.18
5        1     2 80.06
6        1     2 81.09
7        1     3 78.71
8        1     3 77.48
9        1     3 76.06
10       1     4 87.77
11       1     4 84.42
12       1     4 84.82
13       1     5 78.61
14       1     5 77.47
15       1     5 77.80
16       1     5 81.58
17       1     5 77.50
18       2     5 78.73
19       2     5 78.23
20       2     5 76.40
21       2     6 81.64
22       2     6 83.04
23       2     6 82.40
24       2     6 81.93
25       2     6 82.96

> Process<-as.factor(process)

> Batch<-as.factor(batch)

> output.1<-lme(y~1+Process,random=~1|Batch)

> summary(output.1)
Linear mixed-effects model fit by REML
 Data: NULL
       AIC      BIC    logLik
  108.6438 113.1858 -50.32191

Random effects:
 Formula: ~1 | Batch
        (Intercept) Residual
StdDev:    2.927192 1.467032

Fixed effects: y ~ 1 + Process
               Value Std.Error DF  t-value p-value
(Intercept) 81.05442  1.260345 18 64.31128  0.0000
Process2    -0.67123  1.019483 18 -0.65841  0.5186
```

```
 Correlation: 
         (Intr)
Process2 -0.19 


Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max 
-1.901566862 -0.557726122 -0.005590905  0.505906835  2.018904889 

Number of Observations: 25
Number of Groups: 6 

> intervals(output.1)
Approximate 95% confidence intervals

 Fixed effects:
                lower       est.     upper
(Intercept) 78.406534 81.0544212 83.702308
Process2    -2.813087 -0.6712329  1.470621
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Batch 
                   lower     est.    upper
sd((Intercept)) 1.501453 2.927192 5.706776

 Within-group standard error:
   lower     est.    upper 
1.059645 1.467032 2.031042 


> predict(output.1,level=0:1)
   Batch predict.fixed predict.Batch
1      1      81.05442      81.09966
2      1      81.05442      81.09966
3      1      81.05442      81.09966
4      1      81.05442      81.09966
5      2      81.05442      80.62849
6      2      81.05442      80.62849
7      3      81.05442      77.69771
8      3      81.05442      77.69771
9      3      81.05442      77.69771
10     4      81.05442      85.31342
11     4      81.05442      85.31342
12     4      81.05442      85.31342
13     5      81.05442      78.61820
14     5      81.05442      78.61820
15     5      81.05442      78.61820
16     5      81.05442      78.61820
17     5      81.05442      78.61820
18     5      80.38319      77.94697
```

```
19     5      80.38319      77.94697
20     5      80.38319      77.94697
21     6      80.38319      82.29782
22     6      80.38319      82.29782
23     6      80.38319      82.29782
24     6      80.38319      82.29782
25     6      80.38319      82.29782
```

1. Plan I

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad Z = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Plan II

$$X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \qquad Z = \text{as above}$$

2. In both cases

$$\text{Cov}\,Y = Z\,\sigma_\beta^2 \underset{4\times4}{I}\,Z' + \sigma^2 \underset{8\times8}{I} = \sigma_\beta^2 \underset{4\times4}{I} \otimes \underset{2\times2}{J} + \sigma^2 I$$

(variances are all $\sigma_\beta^2 + \sigma^2$, covariances for observations from the same lot are σ_β^2, and all other covariances are 0)

3. Fisher Information for independent observations adds so for Plan I the FI is

$$4\,F(\mu_1, \mu_2, \sigma_\beta^2, \sigma^2)$$

while for Plan II the FI is

$$2\,D(\mu_1, \mu_2, \sigma_\beta^2, \sigma^2) + 2\,E(\mu_1, \mu_2, \sigma_\beta^2, \sigma^2)$$

4. For Plan I

$$\bar{y}_{1..} - \bar{y}_{2..} = \left(\mu_1 + \bar{\beta}_. + \tfrac{1}{4}(\epsilon_{111} + \epsilon_{121} + \epsilon_{131} + \epsilon_{141})\right)$$

$$- \left(\mu_2 + \bar{\beta}_. + \tfrac{1}{4}(\epsilon_{211} + \epsilon_{221} + \epsilon_{231} + \epsilon_{241})\right)$$

$$= \mu_1 - \mu_2 + \epsilon_{1..} - \epsilon_{2..}$$

So $E(\bar{y}_{1..} - \bar{y}_{2..}) = \mu_1 - \mu_2$ and

$$Var(\bar{y}_{1..} - \bar{y}_{2..}) = 2\left(\frac{\sigma^2}{4}\right) = \frac{1}{2}\sigma^2$$

So $\sqrt{\text{above}} = \frac{\sigma}{\sqrt{2}}$

For Plan II

$$\bar{y}_{1..} - \bar{y}_{2..} = \left(\mu_1 + \tfrac{1}{2}(\beta_1 + \beta_2) + \epsilon_{1..}\right)$$

$$- \left(\mu_2 + \tfrac{1}{2}(\beta_3 + \beta_4) + \epsilon_{2..}\right)$$

$$= \mu_1 - \mu_2 + \tfrac{1}{2}(\beta_1 + \beta_2 - \beta_3 - \beta_4) + \epsilon_{1..} - \epsilon_{2..}$$

So $E(\bar{y}_{1..} - \bar{y}_{2..}) = \mu_1 - \mu_2$

$$Var(\bar{y}_{1..} - \bar{y}_{2..}) = \left(\frac{1}{2}\right)^2 4(\sigma_\beta^2) + \frac{1}{2}\sigma^2$$

$$= \sigma_\beta^2 + \frac{1}{2}\sigma^2$$

So $\sqrt{\text{above}} = \sqrt{\sigma_\beta^2 + \frac{1}{2}\sigma^2}$

5. Plan I is better, producing the smaller variance for the estimated difference in means

6. For Plan I, we can base an interval on 4 paired differences. These are

lot 1: 82.0 − 78.6 = 3.4
lot 2: 71.8 − 75.6 = −3.8

lot 3 : $80.0 - 78.8 = 1.2$

lot 4 : $77.6 - 77.8 = -.2$

So $\bar{d} = .15$, $s_d = 3.0216$ and the (3 d.f.) t interval is

$$.15 \pm 3.182 \frac{3.0216}{\sqrt{4}} \qquad i.e. \quad .15 \pm 4.807$$

For Plan II, we may make 4 batch mean responses, The first 2 with mean μ_1 and the 2nd 2 with mean μ_2. These can be used to make a difference in 2 sample means and a pooled (2 d.f.) estimate of variance (of a single lot mean) and thus a 2-sample interval

batch means are:

lot 1 : 80.6	lot 3 : 79.1
lot 2 : 74.1	lot 4 : 77.4
mean 77.35	mean 78.25
variance 21.125	variance 1.445

So a (2 d.f.) interval is

$$(77.35 - 78.25) \pm 4.303 \sqrt{\frac{21.125 + 1.445}{2}} \sqrt{\frac{1}{2} + \frac{1}{2}}$$

$$-.9 \pm 14.46$$

7. The 4 within batch differences are independent and have variance $2\sigma^2$. For Plan I these have 0 mean, while for Plan II they each have mean $\mu_1 - \mu_2$. So for Plan I, a 4 df interval for $2\sigma^2$ can be made based on the

sum of squares of these differences. For Plan II a 3 df interval for $2\sigma^2$ can be made based on their sample variance. In either case, then dividing limits by 2 gives limits for σ^2.

8. There is no statistically significant difference. The interval

$$(-2.813, 1.471)$$

is for $\mu_2 - \mu_1$ (The R convention sets level 1 as a baseline, so the "effect" of "process 2" is the incremental effect.) The interval covers 0.

9. $\hat{\sigma}_\beta = 2.93$ and $\hat{\sigma} = 1.47$ so appearances are that batch-to-batch variability is larger than the run-to-run variability

10. r runs from a given batch produce a sample mean with variance

$$\sigma_\beta^2 + \frac{\sigma^2}{r}$$

An average of ℓ (independent) such sample means has variance

$$\frac{1}{\ell}\left(\sigma_\beta^2 + \frac{\sigma^2}{r}\right)$$

So, $\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}$ will have mean $\mu_1 - \mu_2$ and variance

$$\frac{2}{\ell}\left(\sigma_\beta^2 + \frac{\sigma^2}{r}\right)$$

Then, in rough terms, the engineer wants

$$2\sqrt{\frac{2}{\ell}\left(\sigma_\beta^2+\frac{\sigma^2}{r}\right)} \approx .5$$

That is,

$$\frac{2}{\ell}\left(\sigma_\beta^2+\frac{\sigma^2}{r}\right) \approx \frac{1}{16}$$

$$\ell = 32\left(\sigma_\beta^2+\frac{\sigma^2}{r}\right)$$

And the engineer wants to minimize (over choice of $r=1,2,\ldots,10$)

$$r\ell = 32\left(\sigma_\beta^2+\frac{\sigma^2}{r}\right)r$$

$$= 32\left(r\sigma_\beta^2+\sigma^2\right)$$

Obviously, $r=1$ is best. So then, what remains is the choice of

$$\ell = 32\left(\sigma_\beta^2+\sigma^2\right)$$

and plugging in the estimates from R, I'd suggest

$$\ell = 32\left((2.93)^2+(1.47)^2\right) = 344$$

ouch! If this is "too big" something will have to give ... either the roughly 95% "confidence level" or the "to within .5" requirement.