PhD Prelim Exam
METHODS

Summer 2004
(Given on 7/27/04)

An experiment was conducted to study the effectiveness of using red clover plant material as a fertilizer for corn. All possible combinations of incorporation date (fall or spring), crop (oats alone or oats with red clover), and nitrogen fertilizer amount (0, 70, or 140 kg N/ha) were randomly assigned to a total of 48 plots of land. A balanced design was used so that 4 plots were treated with each combination of the three factors. For the sake of brevity we will call the three factors *date*, *crop*, and *fert*.

In spring of 2001 oats or oats with red clover were planted in each plot depending on the assigned level of the factor *crop*. In fall of 2001 oats were harvested from the field leaving behind in each plot either oat residue or oat residue plus red clover. In either the fall of 2001 or spring of 2002, depending on the assigned level of the factor *date*, the leftover plant material in each plot was mixed in with the soil to provide nutrients for growing the next season's crop. Corn was then planted in the same field in the spring of 2002. Nitrogen fertilizer was applied to each plot at a rate depending on the assigned level of the factor *fert*. The yield of the corn in Mg/ha was recorded for each plot. The yield data was used to evaluate the effects of *date*, *crop*, *fert* and their interactions. The following model was fit to the 48 yield observations.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{1i} X_{2i} + \beta_5 X_{1i} X_{3i} + \beta_6 X_{2i} X_{3i} + \beta_7 X_{1i} X_{2i} X_{3i} + e_i \quad (1)$$

where

Y_i denotes the yield of corn planted in the ith plot;

$\beta_0, \beta_1, \ldots, \beta_7$ denote unknown parameters;

$X_{1i} = 0$ if the plant material on the ith plot was mixed into the soil (incorporated) in the fall, and $X_{1i} = 1$ if the plant material on the ith plot was incorporated in the spring;

$X_{2i} = 0$ if the ith plot had been planted with oats alone, and $X_{2i} = 1$ if the ith plot had been planted with oats plus red clover;

X_{3i} = the amount of nitrogen fertilizer assigned to the ith plot;

and $e_1, \ldots, e_{48}$ are independent and identically distributed normal random variables with mean 0 and unknown variance σ^2.

Alternatively, we can write model (1) more compactly using matrix notation as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

where $\mathbf{Y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\mathbf{e}$ are defined in the usual manner.
The mean square error from the fit of model (1) was 0.69529. The least squares estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7)'$ was $\hat{\boldsymbol{\beta}} = (7.079, -0.004, 3.196, 0.036, -1.133, 0.003, -0.021, 0.003)'$.
The inverse of the $\mathbf{X}'\mathbf{X}$ matrix was

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.20833 & -0.20833 & -0.20833 & -0.00179 & 0.20833 & 0.00179 & 0.00179 & -0.00179 \\ -0.20833 & 0.41667 & 0.20833 & 0.00179 & -0.41667 & -0.00357 & -0.00179 & 0.00357 \\ -0.20833 & 0.20833 & 0.41667 & 0.00179 & -0.41667 & -0.00179 & -0.00357 & 0.00357 \\ -0.00179 & 0.00179 & 0.00179 & 0.00003 & -0.00179 & -0.00003 & -0.00003 & 0.00003 \\ 0.20833 & -0.41667 & -0.41667 & -0.00179 & 0.83334 & 0.00357 & 0.00357 & -0.00714 \\ 0.00179 & -0.00357 & -0.00179 & -0.00003 & 0.00357 & 0.00005 & 0.00003 & -0.00005 \\ 0.00179 & -0.00179 & -0.00357 & -0.00003 & 0.00357 & 0.00003 & 0.00005 & -0.00005 \\ -0.00179 & 0.00357 & 0.00357 & 0.00003 & -0.00714 & -0.00005 & -0.00005 & 0.00010 \end{bmatrix}.$$

a) How many degrees of freedom are associated with the mean square error?

b) A second model was fit to the data. This second model excluded all interactions involving X_3 = amount of nitrogen fertilizer. The mean square error for this simpler model was 1.00786. Does it seem reasonable to exclude all interactions involving X_3 = fertilizer amount from the model? Explain.

Regardless of your answer to part (b), please use model (1) and the results from the fit of model (1) to the data when answering subsequent parts of this problem.

c) For each of the four combinations of the factors *date* and *crop* provide an estimated regression line relating yield to amount of nitrogen fertilizer.

d) Provide a 95% confidence interval for the slope of the regression line relating yield to amount of nitrogen for the treatment associated with fall incorporation of the "oats alone" crop.

e) Provide a 95% confidence interval for the slope of the regression line relating yield to amount of nitrogen for the treatment associated with fall incorporation of the "oats with red clover" crop.

f) Was there a significant difference between the two slopes mentioned in parts (d) and (e)? Conduct one test to answer this question. Provide a test statistic, its degrees of freedom, a relevant table value, and a conclusion at the 0.05 significance level.

g) The researchers were interested in estimating the *fertilizer replacement value* (FRV) associated with growing a combination of oats and red clover, and then incorporating (plowing under) the red clover into the soil (along with oat residue) after the harvest of oats in the fall. Red clover is a type of plant called a *legume*, which means essentially that the plants have bacteria associated with their roots that fix nitrogen in the soil where it can be utilized by other plants. Thus, plowing under red clover returns nitrogen to the soil and decreases the need for adding fertilizer in the spring. The FRV for the treatment involving fall incorporation of red clover is defined as the amount of fertilizer needed for the mean yield for fall incorporation of the oat alone crop to match the mean yield for fall incorporation of oats with red clover (without additional nitrogen fertilizer). Provide an estimate of the fertilizer replacement value for the treatment involving fall incorporation of red clover.

h) Provide an exact 95% confidence interval for the FRV estimated in part (g).
Hint: A confidence interval for a ratio of the form β_j/β_k can be obtained by constructing an F-distributed pivotal quantity that involves

$$\hat{\beta}_j - \left(\frac{\beta_j}{\beta_k}\right)\hat{\beta}_k.$$

i) The researchers wanted to know if there was a significant difference between the FRV for fall incorporation of red clover and the FRV for spring incorporation of red clover. (The FRV for spring incorporation of red clover is defined as the amount of fertilizer needed for the mean yield for spring incorporation of the oat alone crop to match the mean yield for spring incorporation of oats with red clover without additional nitrogen fertilizer.) Estimate the FRV for spring incorporation of red clover, and conduct one test to determine if the two FRV values differ significantly at the 0.05 level. To save time, do not bother to compute the value of your test statistic. Define your test statistic as a function of $\hat{\beta}$, $(\mathbf{X}\mathbf{X})^{-1}$, and MSE; and explain briefly how you would use its value to conduct an approximate test.

j) The analyses conducted in parts (a) through (i) would be appropriate if the experiment was conducted according to a completely randomized design. However, instead of a completely randomized design, the following design was used.

The field was divided into four blocks, each containing four plots. The four combinations of *date* and *crop* were randomly assigned to the four plots within each block. Each plot was subdivided into 3 subplots. The three fertilizer amounts were randomly assigned to the 3 subplots within each plot.

Provide columns labeled SOURCE and DF from an ANOVA table for the analysis of this data. Circle any term in your ANOVA table that will serve as an error term for testing the significance of any of the fixed factors *date*, *crop*, *fert*, or their interactions. Draw an arrow from each fixed term to the appropriate error term for testing the significance of that fixed term. If you label any of your terms using a generic name like "error," please name the term or terms that make up that error.

a) $48 - 8 = 40$

b)

$df_{red} = 48 - 5 = 43 \quad SSE_{\text{red}} = (43)(1.00786) = 43.33798$
$SSE_{full} = (40)(0.69529) = 27.8116 \quad df_{full} = 40$

$$F = \frac{(43.33798 - 27.8116)/(43 - 40)}{0.69529} \approx 7.44 > 2.84 = F_{3,40}^{(0.05)}$$

The simpler model is inadequate.

c)

Fall/Oat Alone: $\hat{\beta}_0 + \hat{\beta}_3 X_3 = 7.079 + 0.036 X_3$
Fall/Oat with Red Clover: $\hat{\beta}_0 + \hat{\beta}_2 + (\hat{\beta}_3 + \hat{\beta}_6) X_3 = 10.275 + 0.015 X_3$
Spring/Oat Alone: $\hat{\beta}_0 + \hat{\beta}_1 + (\hat{\beta}_3 + \hat{\beta}_5) X_3 = 7.075 + 0.039 X_3$
Spring/Oat with Red Clover: $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_4 + (\hat{\beta}_3 + \hat{\beta}_5 + \hat{\beta}_6 + \hat{\beta}_7) X_3 = 9.138 + 0.021 X_3$

d)

$$\hat{\beta}_3 \pm t_{40}^{(0.025)} \text{SE}(\hat{\beta}_3) \qquad 0.036 \pm (2.021)\sqrt{(0.69529)(0.00003)} \qquad (0.027, 0.045)$$

e)

$$\hat{\beta}_3 + \hat{\beta}_6 \pm t_{40}^{(0.025)} \text{SE}(\hat{\beta}_3 + \hat{\beta}_6)$$

$$0.015 \pm (2.021)\sqrt{(0.69529)\{0.00003 + 0.00005 + 2(-0.00003)\}}$$

$$(0.007, 0.0230)$$

f)

$$|t| = \frac{|\hat{\beta}_6|}{\text{SE}(\hat{\beta}_6)} = \frac{0.021}{\sqrt{(0.69529)(0.00005)}} \approx 3.56 > 2.021 = t_{40}^{(0.025)}.$$

There was a significant difference between the slopes at the 0.05 level.

g) Find X_3 such that $\hat{\beta}_0 + \hat{\beta}_2 = \hat{\beta}_0 + \hat{\beta}_3 X_3$. The solutions is

$$\widehat{FRV} = \hat{\beta}_2 / \hat{\beta}_3 \approx 88.78 \text{ kg N/ha}$$

h) Let $\theta = \beta_2/\beta_3 = FRV$. Then

$$\left\{ \frac{\hat{\beta}_2 - \theta\hat{\beta}_3}{\mathrm{SE}(\hat{\beta}_2 - \theta\hat{\beta}_3)} \right\}^2 \sim F_{1,40}.$$

An exact 95% confidence interval for θ is given by the set of θ satisfying

$$\left\{ \frac{\hat{\beta}_2 - \theta\hat{\beta}_3}{\mathrm{SE}(\hat{\beta}_2 - \theta\hat{\beta}_3)} \right\}^2 \leq F_{1,40}^{(0.05)}.$$

The quantity to the left of the inequality has the form

$$\frac{a\theta^2 + b\theta + c}{d\theta^2 + e\theta + f}$$

where $a = \hat{\beta}_3^2$, $b = -2\hat{\beta}_2\hat{\beta}_3$, $c = \hat{\beta}_2^2$, $d = 0.00003 * \mathrm{MSE}$, $e = -2 * 0.00179 * \mathrm{MSE}$, and $f = 0.41667 * \mathrm{MSE}$. Rearranging terms yields

$$(dF_{1,40}^{(0.05)} - a)\theta^2 + (eF_{1,40}^{(0.05)} - b)\theta + fF_{1,40}^{(0.05)} - c \geq 0.$$

Solving for θ provides the confidence interval

$$\frac{b - eF_{1,40}^{(0.05)} \pm \sqrt{(eF_{1,40}^{(0.05)} - b)^2 - 4(dF_{1,40}^{(0.05)} - a)(fF_{1,40}^{(0.05)} - c)}}{2(dF_{1,40}^{(0.05)} - a)} \Longleftrightarrow (62.72, 118.92).$$

i) The fertilizer replacement value for spring is given by

$$\frac{\beta_2 + \beta_4}{\beta_3 + \beta_5}.$$

An estimate is 52.9 kg N/ha.

To construct a test, note that

$$H_0 : \frac{\beta_2}{\beta_3} = \frac{\beta_2 + \beta_4}{\beta_3 + \beta_5} \quad \Longleftrightarrow \quad H_0 : \beta_2\beta_5 - \beta_3\beta_4 = 0.$$

An approximate test can be obtained via the delta method. The test statistic

$$Z = \frac{\hat{\beta}_2\hat{\beta}_5 - \hat{\beta}_3\hat{\beta}_4}{\sqrt{\mathrm{MSE}d'(\hat{\beta})(\mathbf{XX})^{-1}d(\hat{\beta})}}$$

has an asymptotic standard normal distribution under H_0, where

$$d'(\beta) \equiv (0, 0, \beta_5, -\beta_4, -\beta_3, \beta_2, 0, 0).$$

j) This is a split-plot experiment where the whole-plot portion of the experiment is a RCBD design with a two-factor treatment structure. Note that *whole-plot error* in the table below consists of block*date, block*crop, and block*date*crop interactions. These interactions are combined to form a single error term which corresponds to whole-plot experimental units to which the combinations of date and crop were randomly assigned. The split-plot error term consists of block*fert, block*date*fert, block*crop*fert, and block*date*crop*fert interactions. These interactions are combined to form a single error term which corresponds to split-plot experimental units to which the amounts of nitrogen fertilizer were randomly assigned.

SOURCE	DF	Error Term
block	3	
date	1	whole-plot error
crop	1	whole-plot error
date*crop	1	whole-plot error
whole-plot error	9	
fert	2	split-plot error
date*fert	2	split-plot error
crop*fert	2	split-plot error
date*crop*fert	2	split-plot error
split-plot error	24	
c. total	47	

You may use any of the following results to answer this question.

Result 1. Let $\boldsymbol{A}$ be a symmetric matrix with $\text{rank}(\boldsymbol{A}) = k$, and let $\boldsymbol{Y} \sim N(\mu, \boldsymbol{\Sigma})$. If $\boldsymbol{A\Sigma}$ is idempotent, then

$$\boldsymbol{Y}^T\boldsymbol{AY} \sim \chi^2_k(\mu^T\boldsymbol{A}\mu).$$

Result 2. Let $\boldsymbol{Y} \sim N(\mu, \boldsymbol{\Sigma})$ and let $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ be symmetric matrices. If $\boldsymbol{A}_1\boldsymbol{\Sigma}\boldsymbol{A}_2 = \boldsymbol{0}$, then

$$\boldsymbol{Y}^T\boldsymbol{A}_1\boldsymbol{Y} \quad \text{and} \quad \boldsymbol{Y}^T\boldsymbol{A}_2\boldsymbol{Y}$$

are independent random variables.

Result 3. If $\boldsymbol{Y} \sim N(\mu, \boldsymbol{\Sigma})$ and $\boldsymbol{A}$ is a symmetric matrix, then

$$\begin{aligned} E(\boldsymbol{Y}^T\boldsymbol{AY}) &= \mu^T\boldsymbol{A}\mu + \text{trace}(\boldsymbol{A\Sigma}) \\ Var(\boldsymbol{Y}^T\boldsymbol{AY}) &= 4\,\mu^T\boldsymbol{A\Sigma A}\mu + 2\,\text{trace}(\boldsymbol{A\Sigma A\Sigma}) \end{aligned}$$

Result 4. If $\boldsymbol{Y} \sim N(\mu, \boldsymbol{\Sigma})$, then

$$\boldsymbol{AY} \sim N(\boldsymbol{A}\mu, \boldsymbol{A\Sigma}\boldsymbol{A}^T).$$

Result 5. If $\begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}\right)$, then $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent random vectors if and only if $Cov(\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

You may use definitions from the lecture notes and other results proven in the lecture notes if you clearly state the results or definitions.

1. Six pigs were used in an experiment to determine the effects of two dietary supplement additives (I and II) and three pig breeds on weight gain. Two six week old pigs were sampled from each of three breeds: Yorkshire (Breed 1), Duroc (Breed 2) and Landrace (Breed 3). Each pig was weighed at the beginning of the experiment and at 8 weeks after the beginning of the experiment, and the 8 week weight gains were recorded. The following table shows the treatment combinations used in the study.

Breed	Pig Within Breed	Amount of Supplement I (g/day)	Amount of Supplement II (g/day)	Weight gain (kg/day)
1	1	1	1	Y_{11}
1	2	1	1	Y_{12}
2	1	0	2	Y_{21}
2	2	0	2	Y_{22}
3	1	2	0	Y_{31}
3	2	2	0	Y_{32}

The following model was proposed for the observed weight gain for the j-th pig from the i-th breed:

$$Y_{ij} = \mu + \alpha_i + \beta_1 Z_{1ij} + \beta_2 Z_{2ij} + \epsilon_{ij}$$

where α_i corresponds to the i-th breed, Z_{1ij} is the amount of supplement I and Z_{2ij} is the amount of supplement II used in the diets fed to the pigs, and the ϵ_{ij}'s are uncorrelated random errors that have mean zero and a common, unknown variance σ^2 that is strictly positive.

(a) In matrix form the model is

$$\boldsymbol{Y}_{6\times 1} = \boldsymbol{X}_{6\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{6\times 1}.$$

Write out $\boldsymbol{Y}, \boldsymbol{X}$ and $\boldsymbol{\beta}$ explicitly.

(b) Find the rank of $\boldsymbol{X}$.

(c) Define "estimability" for a function of $\boldsymbol{\beta}$.

(d) Give an interpretation of the quantity $\beta_1 - \beta_2$ in words that would be meaningful to a swine nutrition researcher. Show that $\beta_1 - \beta_2$ is <u>not</u> estimable.

(e) Suppose that you could add one more pig to the study. Give a combination of breed and levels of supplements I and II that would make $\beta_1 - \beta_2$ estimable.

2. Suppose the experiment in part 1 had been done with 6 Yorkshire pigs, instead of pigs from different breeds. Then, the combinations of dietary supplements I and II could be viewed as three treatment groups with two pigs in each group, as shown in the following table.

Treatment Group	Amount of Supplement I (g/day)	Amount of Supplement II (g/day)	Pig within treatment group	Weight gain (kg/day)
1	1	1	1	Y_{11}
1	1	1	2	Y_{12}
2	0	2	1	Y_{21}
2	0	2	2	Y_{22}
3	2	0	1	Y_{31}
3	2	0	2	Y_{32}

A model for the weight gains of these six pigs is

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

where the ϵ_{ij}'s are uncorrelated random errors, each with a normal distribution with mean zero and positive variance σ^2.

(a) Describe the set of estimable functions of $\mu_1, \tau_1, \tau_2, \tau_3$.

(b) The least squares estimator for $\mu + \tau_1$ is

$$\bar{Y}_{i.} = (Y_{i1} + Y_{i2})/2.$$

State conditions under which $\bar{Y}_{i.}$ is a best linear unbiased estimator.

(c) Define $MSE = \frac{1}{3}\sum_{i=1}^{3}\sum_{j=1}^{2}(Y_{ij} - \bar{Y}_{i.})^2$ and let $\mathbf{c} = (c_1, c_2, c_3)^T$ be a vector of constants. Show that

$$F = \frac{(c_1\bar{Y}_{1.} + c_2\bar{Y}_{2.} + c_3\bar{Y}_{3.})^2}{MSE}$$

has an F-distribution if appropriate conditions on c_1, c_2, c_3 are satisfied. Clearly state the conditions.

(d) Describe the potential benefits of assigning the six pigs to the three treatment groups completely at random (2 pigs to each group). Could you do something that would be even better than completely random assignment? Explain.

3. Suppose the experiment was was done with the six Yorkshire pigs as described in part 2, but now assume that the initial weight of each pig (when the experiment started) was also available. The data could be displayed as follows.

Pig	Amount of Supplement I (g/day)	Amount of Supplement II (g/day)	Initial weight (kg)	Weight gain (kg/day)
1	$X_{11} = 1$	$X_{21} = 1$	X_{31}	Y_1
2	$X_{12} = 1$	$X_{22} = 1$	X_{32}	Y_2
3	$X_{13} = 0$	$X_{23} = 2$	X_{33}	Y_3
4	$X_{14} = 0$	$X_{24} = 2$	X_{34}	Y_4
5	$X_{15} = 2$	$X_{25} = 0$	X_{35}	Y_5
6	$X_{16} = 2$	$X_{26} = 0$	X_{36}	Y_6

The regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i$$

was used. In this model, the ϵ_i's are uncorrelated random errors, each with a normal distribution with mean zero and unknown, positive variance σ^2. This model can be expressed in matrix form as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ for an appropriate $\boldsymbol{X}$ and the least squares estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$ is

$$\boldsymbol{b} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} \qquad (*)$$

(a) What condition must the values of X_3 satisfy for β_3 to be estimable? (Henceforth, assume this condition is satisfied.)

(b) To check the validity of the regression model, weight gains will be obtained on 8 new pigs with potentially different initial weights and different combinations of the levels of the two supplements. The data for these 8 pigs can be expressed as

Amount of Supplement I (g/day)	Amount of Supplement II (g/day)	Initial weight (kg)	Weight gain (kg/day)
Z_{11}	Z_{21}	Z_{31}	W_1
Z_{12}	Z_{22}	Z_{32}	W_2
Z_{13}	Z_{23}	Z_{33}	W_3
Z_{14}	Z_{24}	Z_{34}	W_4
Z_{15}	Z_{25}	Z_{35}	W_5
Z_{16}	Z_{26}	Z_{36}	W_6
Z_{17}	Z_{27}	Z_{37}	W_7
Z_{18}	Z_{28}	Z_{38}	W_8

Predicted weight gains for these 8 pigs will be computed as

$$\hat{W}_j = b_0 + b_1 Z_{1j} + b_2 Z_{2j} + b_3 Z_{3j} \ .$$

In fact, the entire vector of predictions $\hat{\boldsymbol{W}} = (\hat{W}_1, \ldots, \hat{W}_8)$ can be written as

$$\hat{\boldsymbol{W}} = \boldsymbol{Z}\boldsymbol{b}$$

for $\boldsymbol{Z}$ an appropriate matrix and $\boldsymbol{b}$ the OLS estimator $(*)$ from the original set of 6 pigs.

i. Obtain a standard error in general notation for the predicted weight gain for one of these new pigs.

ii. The researchers propose a prediction variance computed as

$$S_P^2 = \frac{1}{7}\sum_{j=1}^{8}(W_j - \hat{W}_j)^2$$

to be used in assessing the effectiveness of the regression model. The proposal is to reject the null hypothesis that the regression model is appropriate if

$$F = \frac{S_P^2}{S^2} > F_{(7,3),\alpha}$$

where $F_{(7,3),\alpha}$ is the upper α point of a central F-distribution with (7,3) d.f. and $S^2 = \frac{1}{3}\sum_{i=1}^{6}(Y_i - \hat{Y}_i)^2$ is the mean square of the residuals for the six original pigs. Critically comment on this proposed test. (Does this test, for example, have level α?)

There may be more than one correct or reasonable solution to some parts of this question. One solution is presented here. No attempt was made to discuss all possible solutions.

1. (a)

$$\boldsymbol{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 2 \\ 1 & 0 & 1 & 0 & 0 & 2 \\ 1 & 0 & 0 & 1 & 2 & 0 \\ 1 & 0 & 0 & 1 & 2 & 0 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

   (b) rank(A)=3

   (c) Definition

   (d) $\beta_1 - \beta_2$ represents the expected difference in weight gains for two pigs from the same breed when one pig is fed 1g/day of supplement I and no supplement II and the other pig is fed 1g/day of supplement II and no supplement I. Working from the definition, you could show that there is no linear combination of the Y_{ij}'s with expectation equal to $\beta_1 - \beta_2$. Since Y_{i1} and Y_{i2} have the same expectation we need only consider

$$\begin{aligned} E(c_1Y_{11} + c_2Y_{21} + c_3Y_{31}) &= c_1(\mu + \alpha_1 + \beta_1 + \beta_2) + c_2(\mu + \alpha_2 + 2\beta_2) + c_3(\mu + \alpha_3 + 2\beta_1) \\ &= \mu(c_1 + c_2 + c_3) + c_1\alpha_1 + c_2\alpha_2 + c_3\alpha_3 + \beta_1(c_1 + 2c_3) + \beta_2(c_1 + 2c_2) \end{aligned}$$

   which cannot equal $\beta_1 - \beta_2$ because we must have $c_1 = c_2 = c_3 = 0$ to delete the α_i's. Alternatively, you could show that (0 0 0 0 1 -1) is not in the row space of $\boldsymbol{X}$.

2. (a) Any function of the form $c_1(\mu + \tau_1) + c_2(\mu + \tau_2) + c_3(\mu + \tau_3)$ is estimable.

   (b) By the Gauss-Markov theorem, if $E(Y_{ij}) = \mu + \tau_i$ for all (i, j) and the ϵ_{ij}'s are uncorrelated with mean zero and homogeneous variance σ^2.

   (c) You can show that

$$F = \frac{(c_1\bar{Y}_{1.} + c_2\bar{Y}_{2.} + c_3\bar{Y}_{3.})^2}{MSE}$$

   has an F-distribution by showing that $(c_1\bar{Y}_{1.} + c_2\bar{Y}_{2.} + c_3\bar{Y}_{3.})^2$ and MSE are independent and distributed as suitable multiples of chi-square random variables.
   The model can be written as $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where

$$\boldsymbol{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \tau_3 \end{bmatrix}$$

   Then $\frac{3MSE}{\sigma^2} = \frac{Y^T(I - P_X)Y}{\sigma^2}$ where $P_X = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}$ is an idempotent matrix and $\boldsymbol{Y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I)$. By result 1, $\frac{3\text{MSE}}{\sigma^2} \sim \chi^2_3(\mathbf{0})$ because $\frac{1}{\sigma^2}(I - P_X)(\sigma^2 I) = I - P_X$ is an idempotent matrix, and rank(X)=3 and $(X\beta)^T(\frac{1}{\sigma^2}(I - P_X))X\beta = \mathbf{0}$. Now determine conditions under which $(c_1\bar{Y}_{1.} + c_2\bar{Y}_{2.} + c_3\bar{Y}_{3.})^2$ has a chi-squared distribution. Note that

$$c_1\bar{Y}_{1.} + c_2\bar{Y}_{2.} + c_3\bar{Y}_{3.} = \boldsymbol{a}^T\boldsymbol{Y} \text{ where } \boldsymbol{a} = \frac{1}{2}(c_1\ c_1\ c_2\ c_2\ c_3\ c_3)^T.$$

To apply result 1 to

$$\frac{1}{\sigma^2}(c_1\bar{Y}_{1.} + c_2\bar{Y}_{2.} + c_3\bar{Y}_{3.})^2 = \frac{\boldsymbol{Y}^T\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Y}}{\sigma^2}.$$

We must have conditions on $\boldsymbol{a}$ so that $(\frac{1}{\sigma^2}\boldsymbol{a}\boldsymbol{a}^T)(\sigma^2 I) = \boldsymbol{a}\boldsymbol{a}^T$ is idempotent. This requires that $\boldsymbol{a}\boldsymbol{a}^T = (\boldsymbol{a}\boldsymbol{a}^T)(\boldsymbol{a}\boldsymbol{a}^T) = (\boldsymbol{a}^T\boldsymbol{a})\boldsymbol{a}\boldsymbol{a}^T$. Consequently we must have $1 = \boldsymbol{a}^T\boldsymbol{a} = (c_1^2 + c_2^2 + c_3^2)/2$ or $c_1^2 + c_2^2 + c_3^2 = 2$. Note that $\text{rank}(\boldsymbol{a}\boldsymbol{a}^T) = 1$. Consequently, $\frac{1}{\sigma^2}(c_1\bar{Y}_{1.} + c_2\bar{Y}_{2.} + c_3\bar{Y}_{3.})^2$ has a non-central chi-squared distribution with 1 df and non-centrality parameter $\delta^2 = \beta^T\boldsymbol{X}^T\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{X}\beta/\sigma^2$. Finally, we must show that $MSE = \boldsymbol{Y}^T(I - P_X)\boldsymbol{Y}$ is independent of $\boldsymbol{Y}^T\boldsymbol{a}\boldsymbol{a}^T\boldsymbol{Y}$. This follows from result 2, because $(I - P_X)(\sigma^2 I)\boldsymbol{a}\boldsymbol{a}^T = \sigma^2(I - P_X)\boldsymbol{a}\boldsymbol{a}^T = 0$. (Note that $\boldsymbol{a} = \boldsymbol{X}(0\ \frac{c_1}{2}\ \frac{c_2}{2}\ \frac{c_3}{2})^T$ and $(I - P_X)\boldsymbol{X} = 0$). Then,

$$F = \frac{(c_1\bar{Y}_{1.} + c_2\bar{Y}_{2.} + c_3\bar{Y}_{3.})^2}{MSE}$$

satisfies the definition of a random variable with a non-central F-distribution with (1,3) degrees of freedom and non-centrality parameter

$$\delta^2 = \frac{[c_1(\mu + \tau_1) + c_2(\mu + \tau_2) + c_3(\mu + \tau_3)]^2}{\sigma^2}.$$

This statistic can be used to test the null hypothesis

$$H_0 : c_1(\mu + \tau_1) + c_2(\mu + \tau_2) + c_3(\mu + \tau_3) = 0$$

against the alternative

$$H_A : c_1(\mu + \tau_1) + c_2(\mu + \tau_2) + c_3(\mu + \tau_3) \neq 0.$$

(d) Randomization eliminates possible sources of bias associated with differences among Yorkshire pigs. It also yields an approximate F-distribution for the statistic in part (c), regardless of distributional assumptions about variation in pigs. (The approximation is better for larger numbers of pigs.) Randomization does not necessarily reduce variability. You may be able to reduce variability and increase power of tests by blocking. Here you could form two blocks based on initial weight, with the three heaviest pigs in one block and the three lightest pigs in the other block. Then randomly assign pigs to the three supplement combinations without each block.

3. (a) Either $X_{31} \neq X_{32}$ or $X_{33} \neq X_{34}$ or $X_{35} \neq X_{36}$.

(b) i. $\hat{\boldsymbol{W}} = \boldsymbol{Z}\boldsymbol{b} = \boldsymbol{Z}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{Y}$ and the variance-covariance matrix $\hat{\boldsymbol{W}}$ is for

$$\sigma^2\boldsymbol{Z}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{Z}^T.$$

Using MSE from part (a) to estimate σ^2, the standard error of the i-th predicted weight gain is

$$S_i = \sqrt{MSE(1 + h_{ii})}$$

where h_{ii} is the (i,i) element of

$$\boldsymbol{Z}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{Z}^T.$$

ii. $F = S_P^2/S^2$ does not have an F-distribution. S_P^2 does not satisfy result 1.

Background

In chemical analyses of environmental samples, the variable of concern is often the concentration of some substance present in a sample of some medium, such as the concentration of mercury in fish tissue, the concentration of nitrogen in a soil sample, or the concentration of chlorophyll in a water sample. In studies that involve such chemical analyses, there is a good deal of concern with the effect of sample handling, preparation, and instrument calibration on the measured concentration. Sample handling, preparation, and instrument calibration may be considered to consist of three components, (1) sample collection and storage, (2) sample processing, and (3) instrumentation. Sample collection and storage involves the physical methods used to collect a sample from the environment in the first place (e.g., plastic versus metal devices for collection of soil), the physical properties of how a sample is brought from the field to the laboratory (e.g., plastic bag versus glass container), and the storage procedure used if the sample is not analyzed immediately (e.g., wet versus dry). Sample preparation involves any procedures used to produce material from a sample that can be processes by the chemical instruments used for the analysis, such as sample mixing, extraction with solvents, heating or drying, and so forth. Instrumentation involves preparing a laboratory instrument such as a gas chromatograph, mass spectrometer, or spectrophotometer for processing of samples. This usually consists of calibrating the instrumentation so that accurate values for chemical concentrations can be produced from instruments that physically measure some other property of a sample, such as fluorescence or optical density.

Our primary concern in this question will be the effect of storage time on the measured concentration of some chemical substance in a particular type of environmental sample. Exactly what substance or type of sample or storage technique we are dealing with has little bearing on what will be asked in this question, but

a brief description of the study design may be useful. The study under consideration was conducted by collecting one environmental sample, mixing thoroughly, and dividing it into 30 aliquots (subsamples) of equal volume. At each of 6 points in time, 5 of the aliquots are removed from storage and processed according to a given procedure for chemical analysis. Relative to the considerations discussed in the preceding paragraph, sample collection was identical for all 30 aliquots, and the effects of storage may be considered independent for all 30 aliquots (since subsampling was conducted prior to storage). Sample processing and instrumentation were necessarily conducted 6 separate times, for the analysis of the 5 aliquots to be assessed at those times. Time was measured in days of storage, and the 6 times at which aliquots were analyzed were 0, 3, 6, 12, 17, and 24 days in storage.

One additional piece of background information is relevant for us. The physics of degradation for the type of compound being analyzed for (i.e., the response of interest) and the particular storage method used indicate that the relation between chemical concentration of the response variable (y say) and the time of storage (x say) should be of the form,

$$y = \frac{1}{\beta_0 + \beta_1 x}. \tag{1}$$

In (1) the values of the parameters β_0 and β_1 are unknown. In particular, it is not known whether or not $\beta_1 = 0$ for the particular compound and times of storage under study.

Questions

1. The following are basic summary values for the study, where the column labeled Time denotes the storage time of samples, and the columns labeled Mean and Variance give the first two sample moments for measured concentrations (in μ g /L) computed separately for each time.

Time	Mean	Variance
0	2.163	0.1348
3	0.996	0.1744
6	1.199	0.5139
12	1.255	0.3162
17	0.518	0.1396
24	1.238	0.1805

A statistician was consulted by the scientist in charge of the study, and this statistician, based on the information presented above, fit the following non-linear regression model,

$$Y_{i,j} = \frac{1}{\beta_0 + \beta_1 x_i} + \epsilon_{i,j}, \tag{2}$$

where i indexes time of storage, and j indexes replicate aliquot (so $Y_{i,j}$ represents concentration of the j^{th} "sample" analyzed at the i^{th} point in time. Also in (1) the assumption was made that the $\epsilon_{i,j}$ were independent and identically distributed with $E(\epsilon_{i,j}) = 0$ and $var(\epsilon_{i,j}) = \sigma^2$ for all i, j. This model was fit using a Gauss-Newton algorithm to compute generalized least squares estimates.

1A. The basic parameter update at a step of a generalized least squares algorithm may be written as

$$\boldsymbol{\beta}^{(j+1)} = \boldsymbol{\beta}^{(j)} + \boldsymbol{\delta}^{(j)},$$

where,

$$\boldsymbol{\delta}^{(j)} = \left(\boldsymbol{V}^{(j)T}\boldsymbol{W}^{(j)}\boldsymbol{V}^{(j)}\right)^{-1}\boldsymbol{V}^{(j)T}\tilde{\boldsymbol{Y}}^{(j)}.$$

Identify the forms of the matrices $\boldsymbol{V}$ and $\boldsymbol{W}$, and the vector $\tilde{\boldsymbol{Y}}$ for the model given in expression (2).

1B. The resultant estimates for fitting model (1) were $\hat{\beta}_0 = 0.553$, $\hat{\beta}_1 = 0.035$, $\hat{\sigma}^2 = 0.3619$. The scientist in charge of the study was particularly interested in the time at which the expected concentration (i.e., $E(Y_{i,j})$) reached a value of 1.2. Find a point estimate for this time.

1C. The regression curve resulting from a fit of model (2) for the data of this question is presented in Figure 1, and a corresponding standardized residual plot is presented in Figure 2.

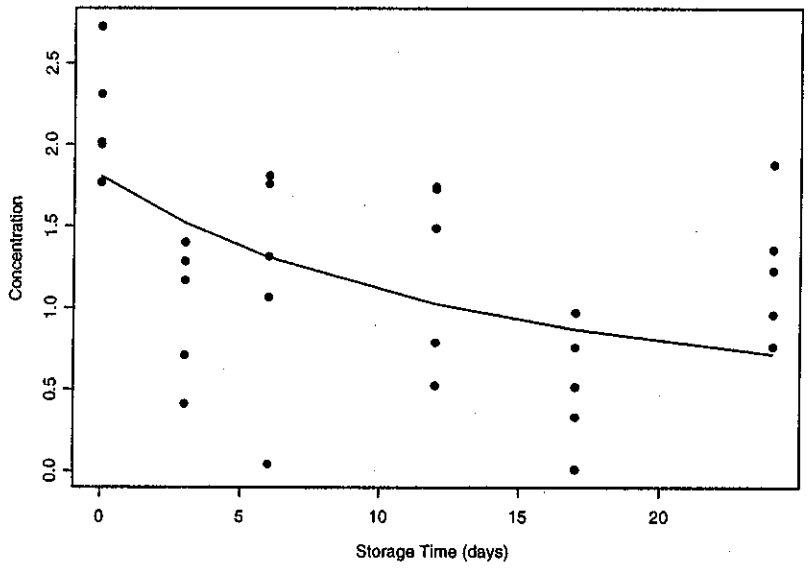

Figure 1. Scatterplot and Fitted Regression Curve.

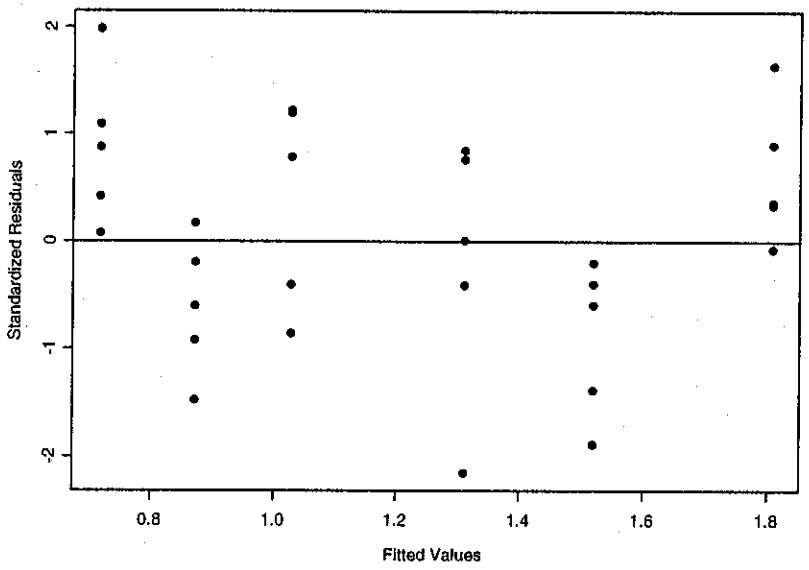

Figure 2. Standardized Residual Plot.

The consulting statistician was troubled by these plots, given the theoretical relation of expression (1) between concentration and storage time for this problem. Should this concern be focused on a misspecification of the systematic model component (i.e., the expectation function or solution locus) or on the assumed structure of constant error variance? Why?

1D. The scientist in charge of the study insists that the physical and chemical theory involved dictates the form of expression (1) without question and that, given a particular "run" in the laboratory (i.e., a setup of equipment and calibration of instrumentation) there is no reason to believe that observed values should have other than constant variance errors. Nevertheless, our statistician decides to produce a typical "Box-Cox" transformation plot, which is presented in Figure 3, in hopes of detecting some way to modify the model

that may help improve the disturbing patterns of Figure 1 and Figure 2.

Do you see anything in this plot that helps our consulting statistician? What type of model modification is this type of plot designed to suggest?

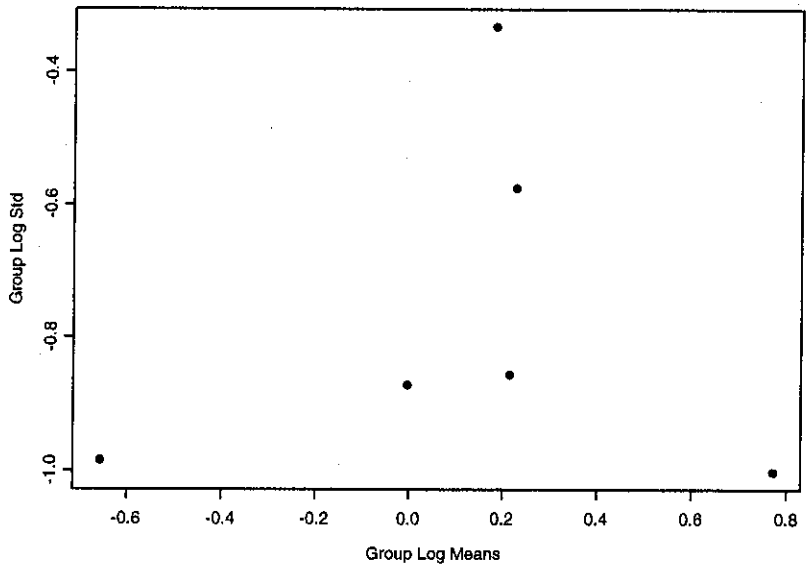

Figure 3. Box-Cox Transformation Plot.

2. Our statistician finally comes to the conclusion that, if the relation of expression (1) cannot be changed, the plots shown in Figures 1 and 2 suggest an effect of analyses conducted on different days (as distinct from the effect of storage time). He or she thus suggests that an appropriate model might be of the form,

$$Y_{i,j} = \frac{1}{\beta_0 + \beta_1 x_i} + \gamma_i + \epsilon_{i,j}, \tag{3}$$

where $\epsilon_{i,j} \sim iidN(0, \sigma^2)$ independent of $\gamma_i \sim iidN(0, \tau^2)$.

2A. Is there anything in what we know about the way the study was conducted that would make model (3) plausible? If so, what?

2B. Derive expressions, in the notation of this question, for the conditional (on values of γ_i) and marginal models corresponding to (3). Give expectations, variances, and covariances.

2C. Name one approach that might be taken for estimation of the parameters β_0, and β_1 in model (3).

3. Our statistician knows several colleagues who have some familiarity with the type of random effects model given in expression (3) and with the type of software that can be used to fit such models. He consults one of these colleagues (C1), and receives a suggestion for modifying the model. The suggestion is that, since expression (1) indicates an inverse relation between storage time and concentration, one might consider the model,

$$Y_{i,j} = \alpha_0 + \alpha_1 \frac{1}{x_i} + \gamma_i + \epsilon_{i,j}, \tag{4}$$

where in (4) we have denoted fixed parameters as α_0 and α_1 to emphasize that they are not exactly the same as β_0 and β_1 in model (3).

The statistician we have called C1 suggests this modification because it allows software developed for linear mixed effects models to easily produce estimates of model parameters.

3A. Derive expressions, in the notation of this question, for conditional and marginal models corresponding to (4). As in question 2B, include expectations, variances, and covariances.

3B. Compare (and contrast) your answer to 3A with that of 2B. Do these models differ? If so, how?

3C. Why might one prefer model (3) to model (4) in this situation?
Hint: consider the range of storage times contained in the data.

3D. Identify a potential difficulty with model (3) that might make it suspect for general use in this type of problem.

4. Another colleague (C2) consulted by our statistician friend suggests a different alternative, this using the idea of a generalized linear mixed model,

$$\begin{aligned} [Y_{i,j}|\gamma_i] &\sim indepN(\mu_i, \sigma^2) \\ \frac{1}{\mu_i} &= \eta_0 + \eta_1 + \gamma_i, \end{aligned} \tag{5}$$

with $\gamma_i \sim iidN(0, \tau^2)$.

Derive the conditional and marginal models corresponding to (5), again with expectations, variances, and covariances. Some of these expressions may require the use of unevaluated integrals.

5. The scientist in charge of this study is disturbed by the implications of any model that contains random effects for the day of analysis (again, not the day effect that corresponds to storage time) because of the implications this might have for the types of comparisons and inferences that could be made based on the chemical analytical procedure used. Explain this concern.

6. As a result of the suggestion that there might be an effect of the combination of factors that correspond to an analytical "run" (day of chemical analysis, calibration, technician effects, etc.) the scientist conducts analyses of additional samples following the same protocol used previously. The basic study described previously is repeated three additional times. Summary statistics of these three additional repetitions of the study are given in the following table.

	Sample 2		Sample 3		Sample 4	
Time	Mean	Var	Mean	Var	Mean	Var
0	2.275	0.9311	2.256	0.6372	2.553	0.1951
3	1.665	0.4809	1.377	0.2286	2.942	0.0865
6	1.039	0.2997	1.597	0.0512	2.397	0.0985
12	1.435	0.7600	1.236	0.2173	1.488	0.6604
17	0.945	0.1285	1.717	0.1037	1.361	1.1408
24	0.627	0.2455	1.482	0.2624	1.271	0.1323

Scatterplots of the full data sets, along with fitted regression equations from model (2) are presented in Figures 4-6 and the corresponding standardized residual plots are presented in Figures 7-9.

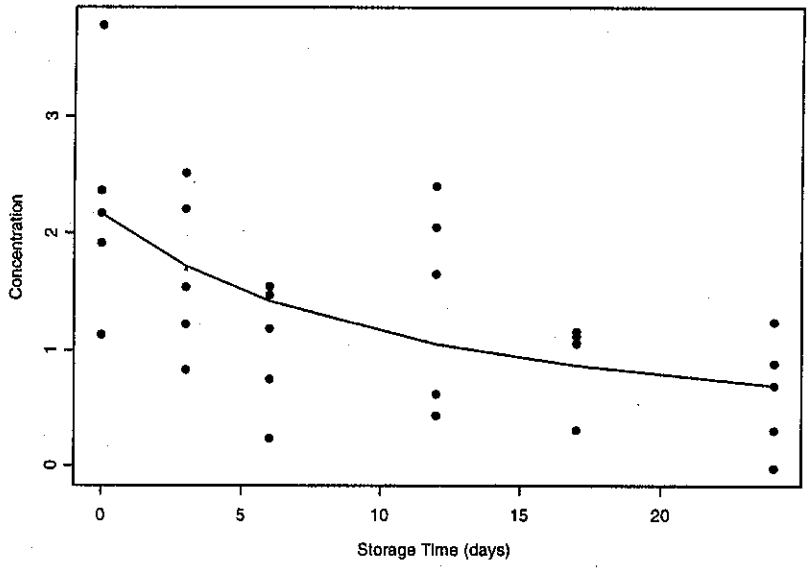

Figure 4. Scatterplot and Fitted Regression Curve for Sample 2.

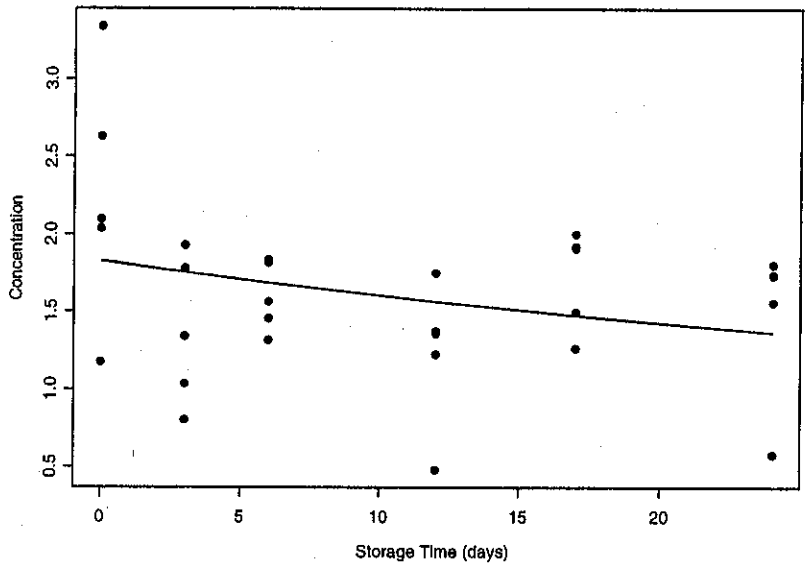

Figure 5. Scatterplot and Fitted Regression Curve for Sample 3.

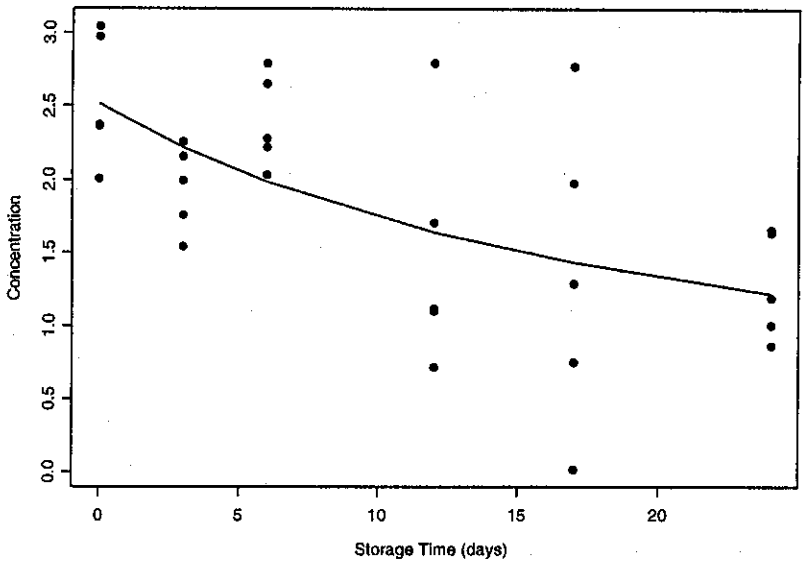

Figure 6. Scatterplot and Fitted Regression Curve for Sample 4.

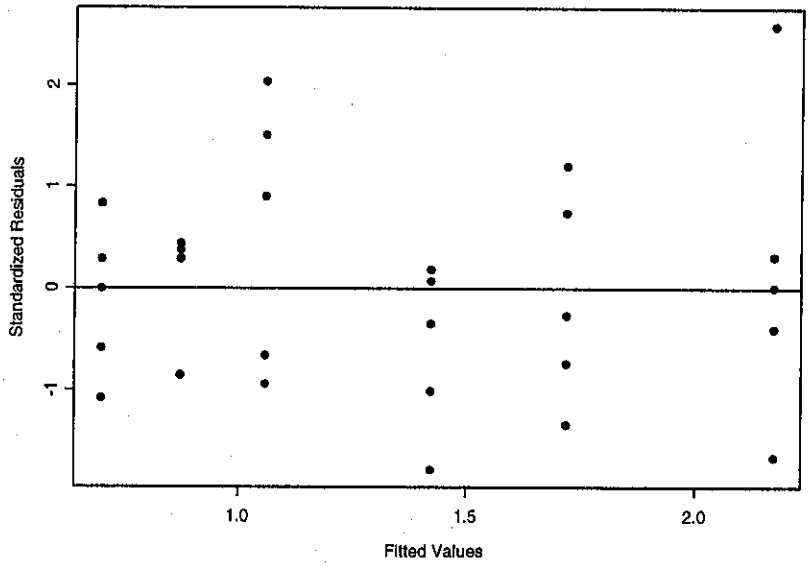

Figure 7. Standardized Residuals for Sample 2.

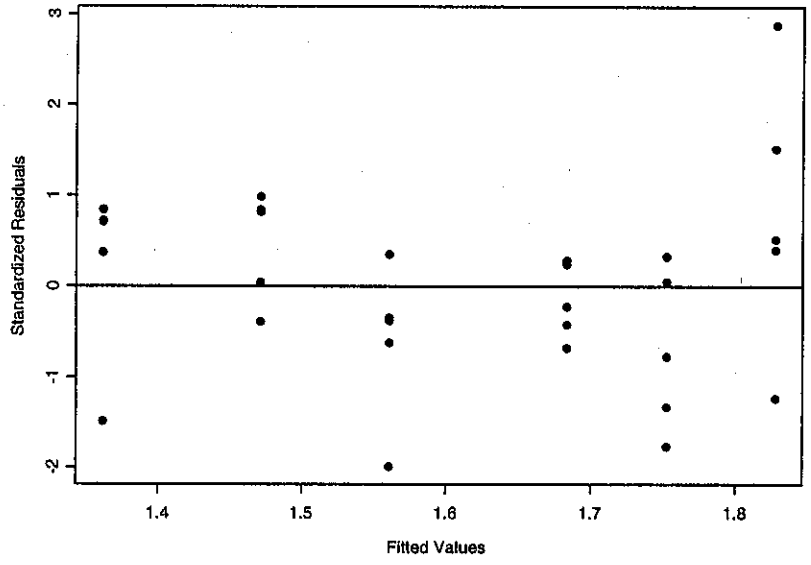

Figure 8. Standardized Residuals for Sample 3.

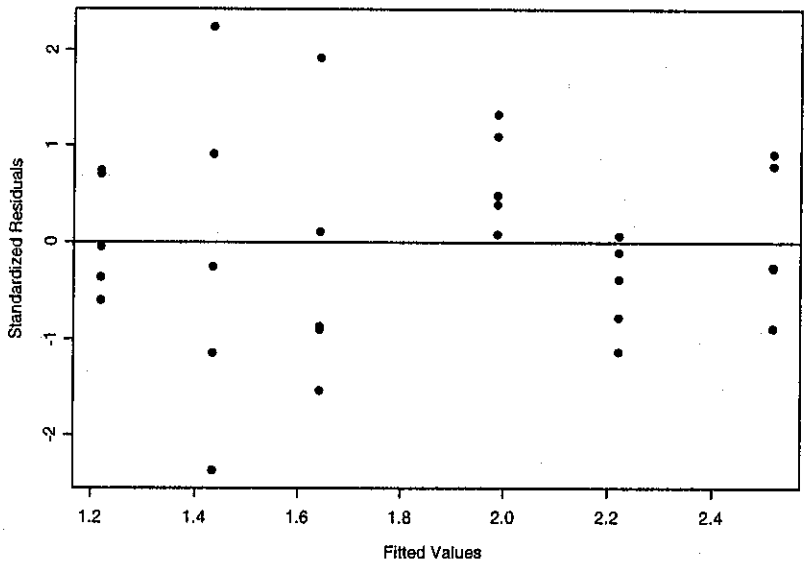

Figure 9. Standardized Residuals for Sample 4.

Parameter estimates for these regressions are given in the table below.

	Estimated Value		
Quantity	Sample 2	Sample 3	Sample 4
β_0	0.461	0.547	0.398
β_1	0.040	0.008	0.018
σ^2	0.462	0.306	0.380
95% Interval for β_0	(0.348, 0.574)	(0.443, 0.652)	(0.328, 0.467)
95% Interval for β_1	(0.014, 0.066)	(−0.002, 0.018)	(0.007, 0.028)

6A. Under the scientific assumption that the basic relation given in (1) between concentration and storage time is correct, these results seem to support the conclusion that there is an effect of day of analysis on the measured concentrations. In addition, there is a suggestion in the residual plots that this effect may be both in terms of level (a typical additive random effect) and also in terms of variance. Without introducing any new models, describe a procedure that could be used to investigate whether there is or is not an effect of day of analysis on the variance of measured concentrations.

6B. Assuming that there is evidence for an effect of day of analysis on both mean and variance, formulate a model that would describe this situation.

6C. Outline a statistical analysis of the model you formulated in question 6B. Use general notation to give form to the basic quantities you will need to consider in your analysis For example, if you denote the densities functions of a set of n independent random variables as $f(y_i|\theta)$ the log likelihood could be written as $L(\theta) = \prod_{i=1}^{n} f(y_i|\theta)$.

1A. If we let n denote the number of observations and p the number of parameters in the expectation function (here 2), and

$$f(x_i, \boldsymbol{\beta}) = \frac{1}{\beta_0 + \beta_1 x_i},$$

then

- $\tilde{\boldsymbol{Y}}^{(j)}$ is a vector of length n with i^{th} element

$$\tilde{Y}_i = y_i - f(x_i, \boldsymbol{\beta}^{(j)})$$

- $\boldsymbol{V}^{(j)}$ is an $n \times p$ matrix with ik^{th} element

$$V_{i,k} = \frac{\partial}{\partial \beta_k} f(x_i, \boldsymbol{\beta})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{(j)}}; \quad i = 1, \ldots, n; \; k = 1, \ldots, p$$

and

- In this model, $\boldsymbol{W}^{(j)}$ is the $n \times n$ identity matrix.

1B. Let T denote the storage time at which the expected concentration is 1.2. A point estimator of T is

$$\hat{T} = \frac{1}{\hat{\beta}_1}\left(\frac{1}{1.2} - \hat{\beta}_0\right),$$

which here takes the value $\hat{T} = 8.009$ days.

1C. Concern should be focused primarily on the systematic model component, since the residuals for days of chemical analysis do not appear to be centered at zero, indicating that the means of this groups of values are not well described by the model. On the other hand, there does not appear to be much evidence for nonconstant variance as the spreads of the groups of residuals are roughly similar.

1D. There is nothing in Figure 3 that helps with a modification of the model. This procedure is designed to determine a power transformation that can be applied to response variables to stabilize variance.

2A. Yes, we know that on each day of chemical analysis any sample preparation that must take place in the laboratory is started from scratch. Also, the instrumentation used to determine chemical concentrations is calibrated anew.

2B. For the conditional model,

$$\begin{aligned} E(Y_{i,j}|\gamma_i) &= \frac{1}{\beta_0 + \beta_1 x_i} + \gamma_i, \\ var(Y_{i,j}|\gamma_i) &= \sigma^2, \\ cov(Y_{i,j}, Y_{k,l}|\gamma_i, \gamma_k) &= 0; \quad \text{all } i, k, j, l. \end{aligned}$$

For the marginal model,

$$\begin{aligned} E(Y_{i,j}) &= \frac{1}{\beta_0 + \beta_1 x_i}, \\ var(Y_{i,j}) &= \tau^2 + \sigma^2 \\ cov(Y_{i,j}, Y_{k,l}) &= \begin{cases} \tau^2 & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases} \end{aligned}$$

2C. Possibilities include REML, maximum likelihood (if the model is extended to include normal distributions for γ_i and $\epsilon_{i,j}$), and quasi-likelihood or generalized estimating equations.

3A. For the conditional model,

$$\begin{aligned} E(Y_{i,j}|\gamma_i) &= \alpha_0 + \alpha_1 \frac{1}{x_i} + \gamma_i, \\ var(Y_{i,j}|\gamma_i) &= \sigma^2, \\ cov(Y_{i,j}, Y_{k,l}|\gamma_i, \gamma_k) &= 0; \quad \text{all } i, k, j, l. \end{aligned}$$

For the marginal model,

$$\begin{aligned} E(Y_{i,j}) &= \alpha_0 + \alpha_1 \frac{1}{x_i}, \\ var(Y_{i,j}) &= \tau^2 + \sigma^2 \\ cov(Y_{i,j}, Y_{k,l}) &= \begin{cases} \tau^2 & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases} \end{aligned}$$

3B. The only difference in these models is in the expectation function. The random effects γ_i are additive in both cases, and all variances and covariances, both conditional and marginal, are the same.

3C. The storage times include values of 0, so that model (4) would need to be modified to have a finite expectation (either conditionally or marginally) for this value of the covariate of storage time.

3D. Model (3) goes to zero as storage time (x_i) goes to infinity. If the random effects γ_i are taken to be $N(0, \tau^2)$ variates, then conditional expectations could well be placed on the negative line. This is not a difficulty for the marginal model, although even here substantial probability (but not expectation) could be placed on the negative line.

4. Here, let

$$h(x_i, \boldsymbol{\eta}, \gamma_i) = \frac{1}{\eta_0 + \eta_1 x_i + \gamma_i},$$

and let $f(\cdot|\sigma^2)$ denote a normal probability density function with variance parameter σ^2. Then, for the conditional model,

$$\begin{aligned} E(Y_{i,j}|\gamma_i) &= h(x_i, \boldsymbol{\eta}, \gamma_i), \\ var(Y_{i,j}|\gamma_i) &= \sigma^2, \\ cov(Y_{i,j}, Y_{k,l}|\gamma_i, \gamma_k) &= 0; \quad \text{all } i, k, j, l. \end{aligned}$$

For the marginal model,

$$\begin{aligned} E(Y_{i,j}) &= \int h(x_i, \boldsymbol{\eta}, \gamma_i) f(\gamma_i|\tau^2)\, d\gamma_i, \\ var(Y_{i,j}) &= \int \{h(x_i, \boldsymbol{\eta}, \gamma_i) - E(Y_{i,j})\}^2 f(\gamma_i|\tau^2)\, d\gamma_i + \sigma^2 \\ cov(Y_{i,j}, Y_{k,l}) &= \begin{cases} \int\int \{h_i - E(Y_{i,j})\}\{h_k - E(Y_{k,l})\}\, f(\gamma_i|\tau^2)\, f(\gamma_k|\tau^2)\, d\gamma_i\, d\gamma_k & \text{if } i = k \\ 0 & \text{if } i \neq k \end{cases} \end{aligned}$$

where in the last expression above we have also used the reduced notation $h_i = h(x_i, \boldsymbol{\eta}, \gamma_i)$, and the product $f(\gamma_i|\tau^2)\, f(\gamma_k|\tau^2)$ follows from independence of the γ_i.

5. The concern is that, if there is a substantial effect due to day of analysis, comparisons among samples or groups of samples that are analyzed on the same day (or laboratory analysis "run") remain legitimate, but the same may not be true for samples or groups of samples analyzed on different days in the laboratory. In addition, in neither case could inferences be made about absolute concentrations, only about relative levels of concentrations.

6A. There would be a number of possibilities here. One would be to make use of group sample variances, such as those reported in the table of page 10, and computing tests of homogeneity of variance among independent groups. This would make no use of any models for the mean structure.

Alternatively, one could define a measure of nonconstancy for the variances of groups, such as the range or variance of group sample variances. Then one might simulate data from model (3) with fixed τ^2 and σ^2 corresponding to estimated values from the data, and compute the chosen measure of nonconstancy for each simulated data set. A type of Monte Carlo test would result from repeating this a large number of times, say M, and determining the percentile rank of the measure from the actual data among the entire list of $M+1$ values.

6B. The following model is one example of an adequate answer.

Let $i = 1, \ldots, T$ index days and $j = 1, \ldots, n_i$ denote observations on day i. Let x_i denote the time of storage in days, and $Y_{i,j}$ the concentration of analysis j on day i. Let,

$$Y_{i,j} = \frac{1}{\beta_0 + \beta_1 x_i} + \gamma_i + \epsilon_{i,j},$$

where,

$$\epsilon_{i,j} \sim iid N(0, \sigma^2) \quad \gamma_i | \tau_i^2 \sim indep\, N(0, \tau_i^2)$$
$$\tau_i^2 \sim iid\, G(\theta)$$

where G is some appropriate distribution such as an inverse gamma, depending on an unknown parameter θ.

6C. This answer will depend on the model formulated in 6B. Here, let the notation $[X]$ stand for "the density of a random variable X". For the example given, one could find the conditional joint density of $\boldsymbol{Y}_i \equiv \{Y_{i,j} : j = 1, \ldots, n_i\}$ for each i as

$$[\boldsymbol{Y}_i | \gamma_i, \tau_i^2] = \prod_{j=1}^{n_i} [Y_{i,j} | \gamma_i, \tau_i^2],$$

where here $[Y_{i,j} | \gamma_i, \tau_i^2] = [Y_{i,j} | \gamma_i]$. Then the marginal density of $\boldsymbol{Y}_i$ would be

$$[\boldsymbol{Y}_i] = \int \int [\boldsymbol{Y}_i | \gamma_i, \tau_i^2]\, [\gamma_i | \tau_i^2]\, [\tau_i^2]\, d\gamma_i\, d\tau_i^2.$$

The log likelihood would then be formed as,

$$L(\boldsymbol{\beta}, \sigma^2, \theta) = \sum_{i=1}^{T} \log \{[\boldsymbol{Y}_i]\},$$

and this function maximized in $\boldsymbol{\beta}$, σ^2 and θ. Such maximization would likely require numerical evaluation of the integrals given immediately above, and possibly also their derivatives (depending on the maximization algorithm used).

Alternatively, one could assign prior distributions to $\boldsymbol{\beta}$, σ^2, and θ, and form the joint posterior as,

$$p(\boldsymbol{\beta}, \sigma^2, \tau_1^2, \ldots, \tau_T^2, \gamma_1, \ldots, \gamma_T, \theta | \{y_{i,j} : i = 1, \ldots, T; j = 1, \ldots, n_i\}) \propto$$

$$\prod_{i=1}^{T} \left\{ [\boldsymbol{Y}_i | \gamma_i, \tau_i^2] \, [\gamma_i | \tau_i^2] \, [\tau_i^2] \right\} \pi(\sigma^2) \pi(\boldsymbol{\beta}) \pi(\theta).$$

The analysis might then proceed by simulating values from this joint posterior through the use of Markov Chain Monte Carlo methods.