PhD Prelim Exam
METHODS

Summer 2008
(Given on 7/8/08)

Data were collected to examine the effects of a drug on a behavior of adult male rats. The behavior was the rate at which rats deprived of water press a lever to obtain water. This study included 24 adult male rats of the same strain and approximately the same weight. Prior to the experiment, each rat was trained to press a lever to obtain water until a stable rate of pressing was reached. The rats were grouped into three blocks, with 8 rats in each block, based on how quickly they learned to press the lever (block 1 = learned slowly, block 2 = learned at a moderate rate, block 3 = learned quickly).

Two treatment factors were considered. One factor (**dosage**) was the dosage level of the drug. Four dosage levels were used. Each dosage level is specified in terms of milligrams of the drug per kilogram of body weight of the rat. The dosage levels were 0, 0.25, 0.35 and 0.45 mg/kg, respectively. The drug was administered by injection. The 0 mg/kg level corresponded to injection with a saline solution. One hour after the injection was administered, each rat was allowed to obtain water by pressing a lever. The second factor (**npress**) was the number of times a rat was required to press the lever to obtain water. This factor had two levels: 2 presses or 5 presses. Within each block, one rat was randomly assigned to each of the eight combinations of the two factors. The measured response (**Y**) was the number of lever presses divided by the length of the session (in seconds).

The data are shown below. The first column gives a rat identification number and the second column gives the block. The third and forth columns give the levels of the two treatment factors. The fifth column gives the observed response.

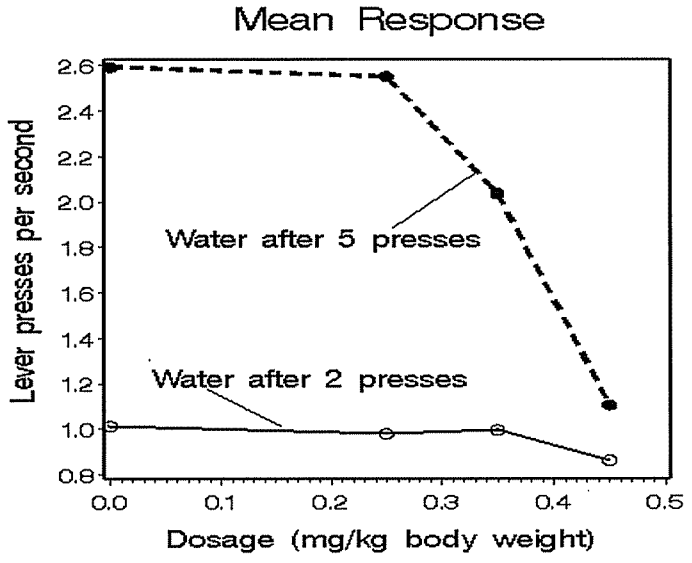
Id	block	npress	dosage	Y
1	1	2	0	0.81
2	1	2	0.25	0.78
3	1	2	0.35	0.83
4	1	2	0.45	0.60
5	1	5	0	2.18
6	1	5	0.25	2.20
7	1	5	0.35	1.86
8	1	5	0.45	0.90
9	2	2	0	1.03
10	2	2	0.25	0.93
11	2	2	0.35	0.98
12	2	2	0.45	0.91
13	2	5	0	2.62
14	2	5	0.25	2.60
15	2	5	0.35	2.09
16	2	5	0.45	1.02
17	3	2	0	1.20
18	3	2	0.25	1.23
19	3	2	0.35	1.18
20	3	2	0.45	1.08
21	3	5	0	2.98
22	3	5	0.25	2.85
23	3	5	0.35	2.16
24	3	5	0.45	1.40

Let Y_{ijk} denote the observed response for the rat assigned to the j-th level of the npress factor and the k-th dosage level in the i-th block. Sample means (averaging across blocks) are given below for the 8 combinations of levels of the two treatment factors, along with corresponding sample standard deviations. Each is computed from three observations.

Table 1. Sample Means and Sample Standard Deviations

	Dosage Level (mg/kg)				
npress level	0	0.25	0.35	0.45	
2	$\overline{Y}_{\cdot 11} = 1.013$ $s_{11} = 0.196$	$\overline{Y}_{\cdot 12} = 0.980$ $s_{12} = 0.229$	$\overline{Y}_{\cdot 13} = 0.997$ $s_{13} = 0.176$	$\overline{Y}_{\cdot 14} = 0.863$ $s_{14} = 0.243$	$\overline{Y}_{\cdot 1\cdot} = 0.963$
5	$\overline{Y}_{\cdot 21} = 2.593$ $s_{21} = 0.401$	$\overline{Y}_{\cdot 22} = 2.550$ $s_{22} = 0.328$	$\overline{Y}_{\cdot 23} = 2.037$ $s_{23} = 0.157$	$\overline{Y}_{\cdot 24} = 1.107$ $s_{24} = 0.261$	$\overline{Y}_{\cdot 2\cdot} = 2.072$
	$\overline{Y}_{\cdot\cdot 1} = 1.803$	$\overline{Y}_{\cdot\cdot 2} = 1.765$	$\overline{Y}_{\cdot\cdot 3} = 1.517$	$\overline{Y}_{\cdot\cdot 4} = 0.985$	$\overline{Y}_{\cdots} = 1.5175$

The sample means are also displayed on the following graph.



One model proposed for these data was

$$Y_{ijk} = \mu + \beta_i + \eta_j + \delta_k + \lambda_{jk} + \varepsilon_{ijk} \qquad \text{(model 1)}$$

where β_i denotes a fixed effect corresponding to the i-th block, η_j denotes a fixed effect corresponding to the j-th level of the npress factor, δ_k denotes a fixed effect corresponding to the k-th dosage level, λ_{jk} corresponds to interaction between levels of the npress and dosage factors, and ε_{ijk} is a random error term. It was also assumed that $\varepsilon_{ijk} \sim \text{NID}\left(0, \sigma_\varepsilon^2\right)$, which means that the random errors are independent and identically distributed according to a Gaussian distribution with mean zero and variance σ_ε^2.

Table 2 shows the ANOVA table based on this model.

Table 2. ANOVA for Model 1

Source of variation	df	Type I Sums of Squares	Mean Square	F-value	P-value
Blocks	2	0.96070	0.48035		
Npress	1	7.37042	7.37042	842.33	<0.0001
Dosage	3	2.55908	0.85303	97.49	<0.0001
Npress x dosage	3	1.78275	0.59425	67.91	<0.0001
Error	14	0.12250	0.00875		

1. Using model 1, construct a 95 percent confidence interval for $\delta_4 - \delta_1 + \lambda_{24} - \lambda_{21}$. In the context of this study, give an interpretation of your confidence interval.

2. Using orthogonal contrasts, partition the sum of squares for interaction between the levels of the npress and dosage factors into three components. In the context of this study, give an interpretation of each of your contrasts. Test the significance of each contrast and state your conclusions.

3. In this study, rats were randomly assigned to combinations of factor levels within each block. What is the purpose of this randomization?

A second model proposed for these data was

$$Y_{ijk} = \mu + \beta_i + \eta_j + \gamma_j X_k + \varepsilon_{ijk} \qquad \text{(model 2)}$$

where $X_1 = 0$, $X_2 = 0.25$, $X_3 = 0.35$, and $X_4 = 0.45$ represent the drug dosage levels used in the study. As in model 1, β_i denotes a fixed effect corresponding to the i-th block, η_j denotes a fixed effect corresponding to the j-th level of the npress factor, and ε_{ijk} denotes a random error term with $\varepsilon_{ijk} \sim \text{NID}\left(0, \sigma_\varepsilon^2\right)$. An analysis of variance table for this model is shown below.

Table 3. ANOVA for Model 2

Source of variation	Df	Type I Sum of Squares	Mean Square	F-value	P-value
Blocks	2	0.96070	0.48035		
Npress	1	7.37042	7.37042	84.20	<0.0001
Regression on Dosage	2	2.88872	1.44436	16.50	<0.0001
Error	18	1.57561	0.08753		

4. Give a definition for an estimable parameter in a linear model.
   (a) Using your definition, show that the parameter γ_2 in model 2 is estimable.
   (b) In the context of this study, carefully explain what γ_2 represents.

5. Does model 2 fit the data as well as model 1? Give some statistical justification for your answer.

An alternative study was proposed that would use only 6 rats. In this study the rats would not be grouped into blocks as in the original study. This study would use the same 2 levels of the npress factor and the same 4 drug dosage levels, but each rat would be exposed to all four dosage levels of the drug in successive time periods. A rat would be injected with one dosage level and the experiment would be conducted to measure the number of lever presses per second in a fixed time period beginning one hour after injection. Then the experimenters would wait for one week to allow the drug to be removed from the rat's body and inject the rat with a different dosage level and perform the experiment. After another one week washout period, the same rat would be injected with a third dosage level and the response would be recorded. Finally, after a third one week washout period, the same rat would be injected with the fourth dosage level and the response would be recorded. Each rat in the study would be randomly assigned to one of the of the 24 possible orderings of the four dosage levels. Then, three of the rats would be randomly assigned to each level on the npress factor. This study would produce a total of $6 \times 4 = 24$ observations.

One possible model for the data gathered from this alternative study is

$$Y_{ijk} = \mu + \eta_j + \delta_k + \lambda_{jk} + \tau_{ij} + \psi_{ijk} \qquad \text{(model 3)}$$

where η_j denotes a fixed effect corresponding to the j-th level of the npress factor, δ_k denotes a fixed effect corresponding to the k-th dosage level, λ_{jk} corresponds to interaction between the levels of the npress and dosage factors, and τ_{ij} is a random effect associated with the i-th rat assigned to the j-th level of the npress factor, and ψ_{ijk} is a random error term. It was also assumed that $\tau_{ij} \sim \text{NID}\left(0, \sigma_\tau^2\right)$ and $\psi_{ijk} \sim \text{NID}\left(0, \sigma_\psi^2\right)$, and all τ_{ij} are distributed independently of all ψ_{ijk}.

6. Consider the set of four observations taken on the i-th rat assigned to the j-th level of the npress factor, $\underset{\sim}{Y}_{ij} = (Y_{ij1}, Y_{ij2}, Y_{ij3}, Y_{ij4})'$.

   (a) Assuming that model 3 is correct for the alternative study, derive the covariance matrix for $\underset{\sim}{Y}_{ij}$.

   (b) Give a formula for the generalized least squares estimator of an estimable linear combination of parameters for model 3. Be sure to define any notation that you use.

   (c) What advantages, if any, are provided by using generalized least squares estimation instead of ordinary least squares estimation in part (b)?

7. Suppose that model 1 is the correct model for the original study and model 3 is the correct model for the alternative study. For each model, $\mu + \eta_j + \delta_k + \lambda_{jk}$ represents a mean response for the j-th level of the npress factor and the k-th drug dosage. With respect to bias and variance, compare the advantages and disadvantages of the original and alternative studies for estimating each of the following contrasts:

   (a) $\eta_j - \eta_r + \lambda_{jk} - \lambda_{rk}$ for $j \neq r$: This is a difference between mean responses for the two levels of the npress factor at a particular dosage level. The generalized least squares estimator (and the mle) of this contrast is $\overline{Y}_{\cdot jk} - \overline{Y}_{\cdot rk}$ for both models 1 and 3.

   (b) $\delta_k - \delta_s + \lambda_{jk} - \lambda_{js}$ for $k \neq s$: This is a difference between mean responses for the two dosage levels at a particular level of the npress factor. The generalized least squares estimator (and the mle) for this contrast is $\overline{Y}_{\cdot jk} - \overline{Y}_{\cdot js}$ for both models 1 and 3.

   (c) $\lambda_{jk} - \lambda_{rk} - \lambda_{js} + \lambda_{rs}$ for $j \neq r$ and $k \neq s$: This is an interaction contrast. The generalized least squares estimator (and the mle) for this contrast is $\overline{Y}_{\cdot jk} - \overline{Y}_{\cdot rk} - \overline{Y}_{\cdot js} + \overline{Y}_{\cdot rs}$ for both models 1 and 3.

1. The least squares estimator (and mle) for $\delta_4 - \delta_1 + \lambda_{24} - \lambda_{21}$ is $\overline{Y}_{.24} - \overline{Y}_{.21}$, and a 95% confidence interval for $\delta_4 - \delta_1 + \lambda_{24} - \lambda_{21}$ is constructed as

$$(\overline{Y}_{.24} - \overline{Y}_{.21}) \pm t_{(14)\,0.025}\sqrt{\frac{2MS_{error}}{3}} \quad \Rightarrow \quad (-1.65,\ -1.32)$$

While it cannot be known if this interval actually contains the difference in mean rates at which rats press a lever to get water for rats treated with 0.45mg/kg dose of the drug and the rats treated with the placebo, when the rats are required to make 5 presses to obtain water, confidence intervals constructed in this manner would have a 95 percent chance of covering the difference in those means.

2. There are many possible sets of orthogonal contrasts that one could use. One set is:

Contrast	Estimate	Sum of Squares	PR>F
$(\lambda_{22} - \lambda_{21}) - (\lambda_{12} - \lambda_{11})$	-0.0100	.000075	0.9275
$(\lambda_{23} - \frac{\lambda_{22} + \lambda_{21}}{2}) - (\lambda_{13} - \frac{\lambda_{12} + \lambda_{11}}{2})$	-0.535	0.0935	<.0001
$(\lambda_{24} - \frac{\lambda_{32} + \lambda_{22} + \lambda_{21}}{3}) - (\lambda_{14} - \frac{\lambda_{13} + \lambda_{12} + \lambda_{11}}{3})$	-1.153	0.0882	<.0001

Whatever set of contrast are presented, interpretations must be given in the context of the study.

3. In this study, rats were randomly assigned to combinations of factor levels within each block. The use of randomization helps to eliminate potential sources of bias associated with differences in characteristics of individual rats by converting those differences to random errors with a known distribution. This also provides a basis for statistical inference.

4. A parameter is estimable if there is a linear combination of the observed responses, $\underset{\sim}{c}'\underset{\sim}{Y}$ such that $E(\underset{\sim}{c}'\underset{\sim}{Y})$ is equal to the true value of the parameter.

   (a) The parameter γ_2 in model 2 is estimable. Note that $E\left(\frac{\overline{Y}_{.24} - \overline{Y}_{.21}}{0.45}\right) = \gamma_2$

   (b) If model 2 is correct, then γ_2 represents the expected change in the rate at which rats press the lever to obtain water when the drug dosage is increased by one mg per kg of body weight and rats are required to press the lever 5 times to get water. If the block effects are additive as specified by model 2, this expected change in mean responses is consistent across blocks (learning ability levels).

5. Model 2 does not fit as well as model 1. Since model 2 is a special case of model 1, an F-test can be constructed as

$$F = \frac{\frac{SS_{error,model2} - SS_{error,model1}}{df_{error,model2} - df_{error,model1}}}{MS_{error,model1}} = \frac{\frac{1.57561 - 0.12250}{18 - 14}}{0.00875} = 41.5$$

Comparing this to the percentiles of a central F-distribution with (4,14) degrees of freedom, it is obvious that model 2 does not adequately describe the changes in the mean rates of pressing the lever as the dosage is increased.

6. Consider the set of four observations taken on the i-th rat assigned to the j-th level of the npress factor, $\underset{\sim}{Y}_{ij} = (Y_{ij1}, Y_{ij2}, Y_{ij3}, Y_{ij4})'$.

   (a) Model 3 impôses a compound symmetric covariance matrix for $\underset{\sim}{Y}_{ij}$;

$$\Sigma = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \end{pmatrix}$$

   where $\sigma^2 = \sigma_\tau^2 + \sigma_\psi^2$ is the common variance and $\rho = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\psi^2}$ is the common correlation.

   (b) Define the vector of 24 responses and the corresponding model matrix and parameter vector as follows:

$$\underset{\sim}{Y} = \begin{bmatrix} Y_{111} \\ Y_{112} \\ Y_{113} \\ Y_{114} \\ Y_{121} \\ Y_{122} \\ Y_{123} \\ Y_{124} \\ Y_{211} \\ Y_{212} \\ Y_{213} \\ Y_{214} \\ Y_{221} \\ Y_{222} \\ Y_{223} \\ Y_{224} \\ Y_{311} \\ Y_{312} \\ Y_{313} \\ Y_{314} \\ Y_{321} \\ Y_{322} \\ Y_{323} \\ Y_{324} \end{bmatrix} \quad X = \begin{bmatrix} 1\,1\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0 \\ 1\,1\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0 \\ 1\,1\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0 \\ 1\,1\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0 \\ 1\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0 \\ 1\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0 \\ 1\,0\,1\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0 \\ 1\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,1 \\ 1\,1\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0 \\ 1\,1\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0 \\ 1\,1\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0 \\ 1\,1\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0 \\ 1\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0 \\ 1\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0 \\ 1\,0\,1\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0 \\ 1\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,1 \\ 1\,1\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0 \\ 1\,1\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0 \\ 1\,1\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0 \\ 1\,1\,0\,0\,0\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0 \\ 1\,0\,1\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0\,0 \\ 1\,0\,1\,0\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0\,0 \\ 1\,0\,1\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,1\,0 \\ 1\,0\,1\,0\,0\,0\,1\,0\,0\,0\,0\,0\,0\,0\,1 \end{bmatrix} \quad \beta = \begin{bmatrix} \mu \\ \eta_1 \\ \eta_2 \\ \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \lambda_{11} \\ \lambda_{12} \\ \lambda_{13} \\ \lambda_{14} \\ \lambda_{21} \\ \lambda_{22} \\ \lambda_{23} \\ \lambda_{24} \end{bmatrix} \quad V = \begin{bmatrix} 1\,0\,0\,0\,0\,0 \\ 0\,1\,0\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0 \\ 0\,0\,0\,0\,1\,0 \\ 0\,0\,0\,0\,0\,1 \end{bmatrix} \otimes \Sigma$$

Using the covariance matrix Σ defined above, V is a covariance matrix constructed as the kronecker product of a 6x6 identity matrix with Σ. Then the generalized least squares estimator for an estimable linear combination of parameters $\underset{\sim}{c}'\beta$ is $\underset{\sim}{c}'(X'V^{-1}X)^{-}X'V^{-1}\underset{\sim}{Y}$, where $(X'V^{-1}X)^{-}$ denotes a generalized inverse.

(c) The generalized least squares estimator provides a best linear unbiased estimator for $\underset{\sim}{c}'\beta$. In this particular "balanced" situation, the least squares estimator is identical to the generalized least squares estimator. This would not necessarily be the case, however, if a complete set of four observations was not obtained on all 6 rats used in this experiment.

7. For each contrast, compare the variances of the estimators for model 3 with the alternative study and model 1 with the original study. Discuss situations in which a particular estimator would have the smaller variance.

   (a) Under model 1 for the original study, the variance of the estimator is $\mathrm{Var}(\overline{Y}_{\cdot jk} - \overline{Y}_{\cdot rk}) = \frac{2\sigma_\varepsilon^2}{3}$. Under model 3 for the alternative study the variance is $\mathrm{Var}(\overline{Y}_{\cdot jk} - \overline{Y}_{\cdot rk}) = \frac{2(\sigma_\tau^2 + \sigma_\psi^2)}{3}$. Since σ_ε^2 represents variation among rats within blocks and other random variation that should be represented by σ_ψ^2 in model 3 and σ_τ^2 represents variation among rats when no blocking is used, it is likely that $\sigma_\tau^2 + \sigma_\psi^2$ will be larger than σ_ε^2. The original experiment should be better for estimating a difference in mean responses for the two levels of the npress factor at a specific dosage level..

   (b) Under model 1 for the original study, the variance of the estimator is $\mathrm{Var}(\overline{Y}_{\cdot jk} - \overline{Y}_{\cdot js}) = \frac{2\sigma_\varepsilon^2}{3}$. Under model 3 for the alternative study the variance is $\mathrm{Var}(\overline{Y}_{\cdot jk} - \overline{Y}_{\cdot js}) = \frac{2\sigma_\psi^2}{3}$. Since σ_ψ^2 should be much smaller than σ_ε^2 because σ_ψ^2 represents only rat variability in behavior and does not have an additional between rat variability component that is represented by σ_ε^2 in model 1. The alternative design could be much better for estimating a difference in mean responses for different dosage levels estimating at a specific level of the npress factor.

   (c) Under model 1 for the original study, the variance of the estimator is $\mathrm{Var}(\overline{Y}_{\cdot jk} - \overline{Y}_{\cdot rk} - \overline{Y}_{\cdot js} + \overline{Y}_{\cdot rs}) = \frac{4\sigma_\varepsilon^2}{3}$. Under model 3 for the alternative study the variance is $\mathrm{Var}(\overline{Y}_{\cdot jk} - \overline{Y}_{\cdot rk} - \overline{Y}_{\cdot js} + \overline{Y}_{\cdot rs}) = \frac{4\sigma_\psi^2}{3}$. For the same reason indicated in part (b), the alternative design could be much better for estimating an interaction contrast between dosage levels and levels of the npress factor.

While comparing variances is a worthwhile exercise, one should also be aware of other practical issues. Although the researchers propose to use a washout period in the alternative design, there could still be carry over effects across successive treatment periods on the same rat. As dosage levels are increased in successive treatment periods, for example, some of the drug from previous infections could accumulate and inflate the effects of the larger dosage levels used during later treatment periods. Since there are only 6 rats in the alternative study, all 24 possible orderings of the dosage levels cannot be used in a single study. The potential for biased comparisons of dosage levels and biased estimates of interaction contrasts is a major concern for the alternative study. Another concern with the alternative study is that the death of a rat in an early treatment period would result in a relatively big loss of information. A positive feature of the alternative study is that it would most likely be less expensive to work with 6 animals than 24 animals.

The following scenario is based on a real study, though most details have been changed and the data provided are not real. Attached to this question is a set of R output that you will want to consult at appropriate points in answering the following questions about this scenario and the analysis of data from it.

An experiment is done to determine whether the progressive effects of a particular disease can be detected in an electrical response of rat eyes to flashes of light. Four animals (rats) are infected with the disease and

y = a particular electrical response to a "standard" light flash (μV)

Two measurements are made on each infected rat at one month, two months, and then three months after infection, for a total of 6 observations per infected rat. To account for the fact that (unintended) changes in lab conditions and/or measurement equipment might occur over time, 2 measurements were also made on 2 uninfected animals at each measurement period. But, these were different animals each month. Thus there were a total of $4+2+2+2=10$ rats involved in the study. Four of these were infected and observed at multiple periods and 6 were control animals that were observed at only a single point in time.

The questions of greatest scientific interest in this study are questions such as "Is there a detectable difference in electrical response between rats that have been infected and those that have not?" and "Can changes in electrical response over time be detected for infected rats?"

1) In light of the basic goals of the study, discuss in qualitative terms any additional utility that would have been provided had the study

a) included measurement of the infected animals at month 0 (just before infection).

b) included measurement of each control animal at all of months $0,1,2,$ and 3.

Label the infected rats $i=1,2,3,4$ and the control rats $i=5,6,\ldots,9,10$. Suppose control rats 5 and 6 are tested at month 1, rats 7 and 8 are tested at month 2, and rats 9 and 10 are tested at month 3. Let

y_{ijk} = the kth response measured on rat i at period j

For

μ_0 = a mean response for an uninfected rat

μ_1 = a month 1 mean repsonse for an infected rat

μ_2 = a month 2 mean repsonse for an infected rat

μ_3 = a month 3 mean repsonse for an infected rat

and

r_i = a rat effect for rat i, $\quad i=1,2,\ldots,10$

m_j = a month effect for month j, $\quad j=1,2,3$

we will initially consider models for this situation of the basic form

$$y_{ijk}=\mu_0 I[i>4]+\mu_j I[i\le 4]+r_i+m_j+\varepsilon_{ijk} \qquad (1)$$

for all $n=36$ relevant combinations of $i, j,$ and k.

First, for the sake of simplicity, consider only the *first* of the two replicate measurements made on only rats $1, 2, 5, 7,$ and 9 at each time period, and an ordinary (fixed effects) Gauss-Markov linear model version of (1). Let

$$\mathbf{Y}' = \left(y_{111}, y_{121}, y_{131}, y_{211}, y_{221}, y_{231}, y_{511}, y_{721}, y_{931}\right)$$

and

$$\boldsymbol{\beta}' = \left(\mu_0, \mu_1, \mu_2, \mu_3, r_1, r_2, r_5, r_7, r_9, m_1, m_2, m_3\right)$$

2) Find a matrix $\mathbf{X}$ so that model (1) for these observations can be written in the usual linear model form

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

Is the resulting model of full rank? (Argue this carefully one way or the other.)

3) Is the quantity $\mu_1 - \mu_0$ estimable in this model? (Argue this carefully one way or another.)

It is probably both more reasonable and also more effective in terms of statistical inference to interpret the rat and month effects in equation (1) as random rather than fixed. So **henceforth** suppose that the r_i's in (1) are iid $\mathrm{N}\left(0, \sigma_r^2\right)$ independent of m_j's that are iid $\mathrm{N}\left(0, \sigma_m^2\right)$, all of which are independent of ε_{ijk}'s that are iid $\mathrm{N}\left(0, \sigma^2\right)$.

Continue for the present to consider only the *first* of the two replicate measurements made on only rats $1, 2, 5, 7,$ and 9 at each time period and $\mathbf{Y}$ exactly as listed at the top of this page.

4) Find matrices $\mathbf{X}$ and $\mathbf{Z}$ and vectors $\boldsymbol{\beta}$ and $\mathbf{u}$ so that model (1) for these observations can be written in the usual mixed linear model form

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

Is the resulting model of full rank? (Argue this carefully one way or the other.) Are the scientifically interesting quantities $\mu_i - \mu_0$ for $i = 1, 2, 3$ and $\mu_2 - \mu_1, \mu_3 - \mu_2,$ and $\mu_3 - \mu_1$ all estimable from these data? Why or why not?

Now begin to consider the whole data set from this study.

5) Based on the mixed effects model described by (1) and the distributional assumptions indicated just after **3)** above, evaluate the following in terms of model parameters:

a) $\mathrm{Var}\left(y_{ijk}\right)$ for any set of indices i, j, k in the data set

b) $\mathrm{Corr}\left(y_{111}, y_{112}\right)$

c) $\mathrm{Corr}\left(y_{111}, y_{211}\right)$ and $\mathrm{Corr}\left(y_{111}, y_{121}\right)$

d) $\mathrm{Corr}\left(y_{111}, y_{221}\right)$

6) Each pair $\left(y_{ij1}, y_{ij2}\right)$ in the data set can be used to compute a sample variance, say s_{ij}^2. How does this sample variance compare to

$$\frac{1}{2-1}\sum_{k=1}^{2}\left(\varepsilon_{ijk}-\bar{\varepsilon}_{ij.}\right)^2$$

(the "sample variance" of the pair $\left(\varepsilon_{ij1}, \varepsilon_{ij2}\right)$)? In light of this, what is the mean of the average sample variance, namely

$$\mathrm{E}\left(\frac{1}{18}\sum_{ij} s_{ij}^2\right) \quad ?$$

Using (1) write $\bar{y}_{1..}, \bar{y}_{2..}, \bar{y}_{3..}$, and $\bar{y}_{4..}$ (the simple averages of the rat 1,2,3, and 4 responses) in terms of averages of appropriate fixed and random effects and random errors. Based on this, what is the expected value of the sample variance of these four sample means? What is the expected value of the sample variance of rat 5 and 6 averages? Of rat 7 and 8 averages? Of rat 9 and 10 averages?

Call the sample variances for groups of rat means referred to above by the respective names $s^2(1,2,3,4), s^2(5,6), s^2(7,8)$, and $s^2(9,10)$. What are sufficient conditions on constants $c_0, c_{1234}, c_{56}, c_{78}, c_{9\ 10}$ under which

$$\mathrm{E}\left(c_0\left(\frac{1}{18}\sum_{ij} s_{ij}^2\right)+c_{1234}s^2(1,2,3,4)+c_{56}s^2(5,6)+c_{78}s^2(7,8)+c_{9\ 10}s^2(9,10)\right)=\sigma_r^2$$

7) The current version of the `lmer()` function in the `lme4` package of Bates produces point estimates of variance components (and their square roots) and Bayes credible intervals (presumably based on Jeffreys priors) for the "error" standard deviation in a mixed linear model and ratios of the other model standard deviations to the error standard deviation. Here the credible intervals are

```
$sigma
         lower    upper
[1,] 0.9412178 1.663813
attr(,"Probability")

$ST
     lower     upper
[1,]     0 0.8048939
[2,]     0 1.7889483
attr(,"Probability")
[1] 0.95
```

Discuss what these results tell you about sources of variability in measuring y on rat eyes.

8) Is there clear evidence of any difference in electrical response to light flash between uninfected and infected rat eyes (at any stage of the disease)? Explain carefully. Is there clear evidence that electrical response changes with time for an infected rat? Explain carefully.

9) No rats were actually tested at month 0 (just before rats 1 through 4 were infected). But based on the data in hand, you might predict what rat 1's responses would have been like at month 0. Give a sensible point prediction for a single test result for rat 1 at month 0. Then give a sensible point prediction *and standard error of that prediction* for a single test for an uninfected rat not included in the data set (say rat 11) at month 0.

Let

$$x_{ij} = \text{the number of months that rat } i \text{ has been infected when measurements are taken at month } j$$

An alternative to the mixed model version of (1) that assumes that change in electrical response is linear in time since infection, but that different rats have different rates of change (while still allowing for month effects on measurements) is

$$y_{ijk} = \beta_0 + (\beta_1 + r_i)x_{ij} + m_j + \varepsilon_{ijk} \qquad (2)$$

where we will assume that the r_i's in (2) are iid $\mathrm{N}(0, \sigma_r^2)$ independent of m_j's that are iid $\mathrm{N}(0, \sigma_m^2)$, all of which are independent of ε_{ijk}'s that are iid $\mathrm{N}(0, \sigma^2)$.

10) Evaluate $\mathrm{Var}(y_{111})$, $\mathrm{Var}(y_{121})$, and $\mathrm{Corr}(y_{111}, y_{121})$ in terms of parameters of model (2).

11) Find any unbiased estimator of σ_r^2 in model (2).

R Printout

```
> MoreRats
       y Rat Month Mu
1  50.49   1     1  1
2  51.87   1     1  1
3  47.70   1     2  2
4  45.85   1     2  2
5  44.79   1     3  3
6  44.75   1     3  3
7  49.82   2     1  1
8  49.56   2     1  1
9  46.76   2     2  2
10 46.39   2     2  2
11 40.80   2     3  3
12 40.36   2     3  3
13 48.74   3     1  1
14 50.60   3     1  1
15 48.55   3     2  2
16 48.82   3     2  2
17 45.16   3     3  3
18 43.95   3     3  3
19 51.59   4     1  1
20 50.70   4     1  1
21 47.76   4     2  2
22 47.15   4     2  2
23 42.96   4     3  3
24 44.69   4     3  3
25 48.92   5     1  0
26 49.46   5     1  0
27 52.11   6     1  0
28 51.18   6     1  0
29 47.50   7     2  0
30 49.23   7     2  0
31 46.73   8     2  0
32 49.63   8     2  0
33 48.50   9     3  0
34 47.31   9     3  0
35 49.23  10     3  0
36 48.04  10     3  0

> eyeout<-lmer(y~Mu+(1|Rat)+(1|Month))

> summary(eyeout)
Linear mixed model fit by REML
Formula: y ~ Mu + (1 | Rat) + (1 | Month)
   AIC   BIC logLik deviance REMLdev
 130.6 141.7 -58.29    121.8   116.6
```

```
Random effects:
 Groups   Name        Variance Std.Dev.
 Rat      (Intercept) 0.65612  0.81001
 Month    (Intercept) 0.89120  0.94403
 Residual             1.26390  1.12423
Number of obs: 36, groups: Rat, 10; Month, 3


Fixed effects:
            Estimate Std. Error t value
(Intercept)  48.9867     0.7154   68.48
Mu1           0.6040     0.8865    0.68
Mu2          -1.1996     0.8865   -1.35
Mu3          -5.1381     0.8865   -5.80

Correlation of Fixed Effects:
    (Intr) Mu1    Mu2
Mu1 -0.339
Mu2 -0.339  0.323
Mu3 -0.339  0.323  0.323

> vcov(eyeout)
4 x 4 Matrix of class "dpoMatrix"
           [,1]       [,2]       [,3]       [,4]
[1,]  0.5117438 -0.2146778 -0.2146778 -0.2146778
[2,] -0.2146778  0.7859343  0.2540876  0.2540876
[3,] -0.2146778  0.2540876  0.7859343  0.2540876
[4,] -0.2146778  0.2540876  0.2540876  0.7859343

> ranef(eyeout)
$Rat
   (Intercept)
1   0.37817047
2  -1.10549417
3   0.42485040
4   0.30247330
5  -0.31951647
6   0.93101854
7  -0.10549049
8  -0.19972633
9  -0.33906738
10  0.03278213

$Month
  (Intercept)
1   0.8305952
2  -0.4145720
3  -0.4160232
```

```
> fitted(eyeout)
 [1] 50.79942 50.79942 47.75067 47.75067 43.81067 43.81067 49.31576 49.31576
 [9] 46.26701 46.26701 42.32701 42.32701 50.84610 50.84610 47.79735 47.79735
[17] 43.85735 43.85735 50.72372 50.72372 47.67497 47.67497 43.73497 43.73497
[25] 49.49775 49.49775 50.74828 50.74828 48.46660 48.46660 48.37237 48.37237
[33] 48.23158 48.23158 48.60343 48.60343

> sim<-mcmcsamp(eyeout, 1000000)

> HPDinterval(sim)
$fixef
                lower      upper
(Intercept) 47.400550 50.5627802
Mu1         -1.036940  2.5609806
Mu2         -2.927035  0.3629514
Mu3         -6.871015 -3.5825394
attr(,"Probability")
[1] 0.95


$ST
     lower     upper
[1,]     0 0.8048939
[2,]     0 1.7889483
attr(,"Probability")
[1] 0.95

$sigma
         lower    upper
[1,] 0.9412178 1.663813
attr(,"Probability")
[1] 0.95
```

Methods II Key Statistics Ph.D. Prelim 2008

Note Title 7/2/2008

1) a) As it stands, any comparison between infection and non-infection mean responses must be based on a comparison of infected and control rats and is clouded by differences between rats (rat effects in modeling terms). If infected rats had been measured before infection, some part of the comparison between infection an non-infection could be done "within rats" and presumably provide more a precise view of differences between them.

b) As it stands, changes in the effects of infection over time are clouded by potential month-to-month measurement differences. Measuring the presumably physically stable control rats at all periods would have given a far firmer view of month-to-month changes in measurement and made it easier to know what are measurement changes and what are disease-induced changes.

2)

$$\begin{pmatrix} y_{111} \\ y_{121} \\ y_{131} \\ y_{211} \\ y_{221} \\ y_{231} \\ y_{511} \\ y_{721} \\ y_{931} \end{pmatrix} = \underbrace{\begin{pmatrix} 0&1&0&0&1&0&0&0&0&1&0&0 \\ 0&0&1&0&1&0&0&0&0&1&0 \\ 0&0&0&1&1&0&0&0&0&0&0&1 \\ 0&1&0&0&0&1&0&0&0&1&0&0 \\ 0&0&1&0&0&1&0&0&0&0&1&0 \\ 0&0&0&1&0&1&0&0&0&0&0&1 \\ 1&0&0&0&0&0&1&0&0&1&0&0 \\ 1&0&0&0&0&0&0&1&0&0&1&0 \\ 1&0&0&0&0&0&0&0&1&0&0&1 \end{pmatrix}}_{X} \begin{pmatrix} m_0 \\ m_1 \\ m_2 \\ m_3 \\ r_1 \\ r_2 \\ r_5 \\ r_7 \\ r_9 \\ m_1 \\ m_2 \\ m_3 \end{pmatrix}}_{\beta} + \underset{\sim}{\epsilon}$$

This model is not full rank, as

$$\underset{\sim}{1} = \text{sum of the 1st 4 columns of } \underset{\sim}{X} = \text{sum of the last 3 columns of } \underset{\sim}{X}$$

which implies that the columns of $\underset{\sim}{X}$ are not linearly independent.

3) N_0, $\mu_1 - \mu_0$ is not estimable in this model. If it were estimable, we would have to be able to find a linear combination of rows of X that is $(-1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0)$. But in order to get 0's in positions 7, 8 and 9 one would have to have coefficients 0 0 the last 3 rows of $\underset{\sim}{X}$, which makes it impossible to get a -1 in the 1st entry of the vector.

4)

$$\underbrace{\begin{pmatrix} y_{111} \\ y_{121} \\ y_{131} \\ y_{211} \\ y_{221} \\ y_{231} \\ y_{511} \\ y_{721} \\ y_{931} \end{pmatrix}}_{\underset{\sim}{Y}} = \underbrace{\begin{pmatrix} 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \\ 1&0&0&0 \\ 1&0&0&0 \\ 1&0&0&0 \end{pmatrix}}_{\underset{\sim}{X}} \begin{pmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \underbrace{\begin{pmatrix} 1&0&0&0&0&1&0&0 \\ 1&0&0&0&0&0&1&0 \\ 1&0&0&0&0&0&0&1 \\ 0&1&0&0&0&1&0&0 \\ 0&1&0&0&0&0&1&0 \\ 0&1&0&0&0&0&0&1 \\ 0&0&1&0&0&1&0&0 \\ 0&0&0&1&0&0&1&0 \\ 0&0&0&1&0&0&0&1 \end{pmatrix}}_{\underset{\sim}{Z}} \underbrace{\begin{pmatrix} r_1 \\ r_2 \\ r_5 \\ r_7 \\ r_9 \\ m_1 \\ m_2 \\ m_3 \end{pmatrix}}_{\underset{\sim}{U}} + \underset{\sim}{\epsilon}$$

This model is clearly of full rank. The 1st, 2nd, 3rd and 7th

rows of X are 4 standard unit vectors (which are linearly independent). So the rank of X is 4 (also the # of columns of X). The fact that X is of full rank guarantees the estimability of every linear combination of $\mu_0, \mu_1, \mu_2, \mu_3$, including those mentioned in this question.

5) a) $\text{Var } y_{ijk} = \sigma_r^2 + \sigma_m^2 + \sigma^2$

b) $\text{Corr}(y_{111}, y_{112}) = \dfrac{\sigma_r^2 + \sigma_m^2}{\sigma_r^2 + \sigma_m^2 + \sigma^2}$

c) $\text{Corr}(y_{111}, y_{211}) = \dfrac{\sigma_m^2}{\sigma_r^2 + \sigma_m^2 + \sigma^2}$; $\text{Corr}(y_{111}, y_{121}) = \dfrac{\sigma_r^2}{\sigma_r^2 + \sigma_m^2 + \sigma^2}$

d) $\text{Corr}(y_{111}, y_{221}) = 0$

6) According to (1), y_{ij1} and y_{ij2} differ only in that the first includes (as a summand) ϵ_{ij1} and the second includes (as a summand) ϵ_{ij2}. Thus the sample variance of y_{ij1} and y_{ij2} is numerically the same as the sample variance of ϵ_{ij1} and ϵ_{ij2} which are iid $N(0, \sigma^2)$. Thus each $E s_{ij}^2 = \sigma^2$ and $E \frac{1}{18} \sum_{ij} s_{ij}^2 = \sigma^2$. Then for $i = 1,2,3,4$

$$\bar{y}_{i\cdot\cdot} = \frac{1}{3}(\mu + \mu_2 + \mu_3) + r_i + \bar{m}_\cdot + \bar{\epsilon}_{i\cdot\cdot}$$

So the sample variance of these is numerically the same as the sample variance of

$$r_i + \bar{\epsilon}_{i\cdot\cdot} \qquad i = 1,2,3,4$$

which are iid $N\left(0, \sigma_r^2 + \frac{\sigma^2}{6}\right)$. So $E s^2(1,2,3,4) = \sigma_r^2 + \frac{\sigma^2}{6}$.

Parallel arguments show $E s^2(5,6) = E s^2(7,8) = E s^2(9,10) = \sigma_r^2 + \frac{\sigma^2}{2}$

So sufficient conditions to produce a l.c. of the sample variances of sample means with expected value σ_r^2 are

1) $c_{1234} + c_{56} + c_{78} + c_{910} = 1$

and 2) $c_0 = -\left(\frac{1}{6} c_{1234} + \frac{1}{2} c_{56} + \frac{1}{2} c_{78} + \frac{1}{2} c_{910}\right)$

7)

```
$sigma
          lower    upper
[1,] 0.9412178 1.663813
attr(,"Probability")
[1] 0.95

$SST
     lower    upper
[1,]     0 0.8048939
[2,]     0 1.7889483
attr(,"Probability")
[1] 0.95
```

credible interval for σ

σ_r is clearly smaller than σ (and in fact is potentially very small)

σ_m is far less precisely known than σ_r (and is potentially very small to nearly twice the size of σ)

These overall give the impression of a measurement system that is fairly imprecise in that its inherent "repeatability" variation is clearly larger than subject-to-subject variability. Its potential "changes with time" as measured by σ_m may be as large or even bigger than its repeatability variation.

8) The first of these questions can answered directly from what is printed out. The t-value for $\mu_3 - \mu_0$ is -5.80 and "clearly" statistically significant. There are clear differences between infected and non-infected responses, at least by the time of month 3.

$\hat{\mu}_0$ $\quad \hat{\mu}_1 - \hat{\mu}_0$ $\quad \hat{\mu}_2 - \hat{\mu}_0$ $\quad \hat{\mu}_3 - \hat{\mu}_0$

```
Fixed effects:
            Estimate Std. Error t value
(Intercept) -48.9867     0.7154   68.48
Mu1          -0.6040     0.8865    0.68
Mu2          -1.1996     0.8865   -1.35
Mu3          -5.1381     0.8865   -5.80
```

To look for progressive effects of the disease we could estimate differences like $\mu_3 - \mu_1$. Note that $\hat{\mu}_3 - \hat{\mu}_1 = \hat{\mu}_3 - \hat{\mu}_0 - (\hat{\mu}_1 - \hat{\mu}_0)$

$= -5.1381 - (-.6040)$

$= -5.7421$

A standard error for this difference is from

```
> vcov(eyeout)
4 x 4 Matrix of class "dpoMatrix"
           [,1]       [,2]       [,3]       [,4]
[1,]  0.5117438 -0.2146778 -0.2146778 -0.2146778
[2,] -0.2146778  0.7859343  0.2540876  0.2540876
[3,] -0.2146778  0.2540876  0.7859343  0.2540876
[4,] -0.2146778  0.2540876  0.2540876  0.7859343
```

given by

$$\sqrt{(-1,1)\begin{pmatrix} .786 & .254 \\ .254 & .786 \end{pmatrix}\begin{pmatrix} -1 \\ 1 \end{pmatrix}} \approx 1.03$$

and so there are clear differences in electrical response with disease progression.

9) The output includes point predictions for the random effects that are involved in the data set and in particular

$$\hat{r}_1 = .378$$

We can only predict m_0 as 0. So sensible predictor for rat 1 response at month 0 is

$$\hat{\mu}_0 + \hat{r}_1 = 48.987 + .378 = 49.365$$

A prediction for rat 11 at month 0 can only involve $\hat{\mu}_0$ and 0 predictions for random effects, so a point prediction is

$$\hat{\mu}_0 = 48.987$$

with corresponding standard error of prediction

$$\sqrt{(.7154)^2 + .656 + .891 + 1.264} = 1.8229$$

$(\text{std error for } \mu_0)^2$ $\quad \hat{\sigma}_r^2$ $\quad \hat{\sigma}_m^2$ $\quad \hat{\sigma}^2$

10) $\text{Var}\, y_{111} = \sigma_r^2(1)^2 + \sigma_m^2 + \sigma^2 \qquad \text{Var}\, y_{121} = \sigma_r^2(2)^2 + \sigma_m^2 + \sigma^2$

$$\text{Cov}(y_{111}, y_{121}) = \text{Cov}(r_1(1) + m_1 + \epsilon_{111},\ r_1(2) + m_2 + \epsilon_{121})$$
$$= E(2(1)r_1^2) = 2\sigma_r^2$$

So $\text{Corr}(y_{111}, y_{121}) \doteq \dfrac{2\sigma_r^2}{\sqrt{\sigma_r^2+\sigma_m^2+\sigma^2}\ \sqrt{4\sigma_r^2+\sigma_m^2+\sigma^2}}$

11) Consider $y_{111} = \beta_0 + (\beta_1+r_1)(1) + m_1 + \epsilon_{111}$

and $y_{211} = \beta_0 + (\beta_1+r_2)\,1 + m_1 + \epsilon_{211}$

The sample variance of these is numerically the same as the sample variance of $r_1+\epsilon_{111}$ and $r_2+\epsilon_{211}$ which are iid with mean 0 and variance $\sigma_r^2+\sigma^2$. Similarly, e.g.

$$y_{511} = \beta_0 + (\beta_1+r_5)\,0 + m_1 + \epsilon_{511}$$

and $y_{512} = \beta_0 + (\beta_1+r_5)\,0 + m_1 + \epsilon_{512}$

have a sample variance that is numerically the same as the sample variance of ϵ_{511} and ϵ_{512} which are iid with mean 0 and variance σ^2. So

$$E\left(\frac{1}{2-1} \sum_{i=1}^{2} \left(y_{i11} - \tfrac{1}{2}(y_{111}+y_{211})\right)^2 - s_{51}^2 \right) = \sigma_r^2$$

1 Problem Background

Agronomists at many land-grant institutions, including Iowa State, make recommendations on the amount of nitrogen fertilizer to apply to corn fields. Corn accumulates a huge amount of nitrogen, both in green plant material and in grain. Historically, the fertilizer strategy used by growers was simple. Apply more nitrogen fertilizer than corn could possibly use, so that there was always a ready supply to the plants. Unused nitrogen fertilizer washed into streams and rivers and was gone. Two things have happened to change this simple strategy. First, we are now aware that high levels of nitrogen fertilizer in streams and rivers is a type of pollution. Secondly, the price of nitrogen fertilizer has increased dramatically, making fertilizer one of the largest costs associated with growing corn. These two factors, environmental protection and economic reality have prompted a need to reassess the way in which fertilizer recommendations are made.

Recommending fertilizer levels is not simply a matter of determining how much nitrogen corn needs. That can be, and has been, done. To understand the complexities of fertilizer recommendations one must have some sense of how corn obtains the nitrogen it needs. Major sources, aside from fertilizer, are soil organic matter and crop residue (plant material left in the field from the previous year). The nitrogen in these sources is not readily available to corn which needs what is called *plant available nitrogen* (primarily ammonium and nitrate). Organic matter nitrogen in soil and crop residue can be converted to plant available nitrogen by the action of microorganisms in the soil, a process called nitrogen mineralization. The amount of mineralization that occurs in a year depends on soil and weather conditions and, because of this, it is impossible to predict with any degree of accuracy and precision how much nitrogen will be available to corn. To make this uncertain situation even more complex, corn needs different amounts of nitrogen at different stages of plant growth which, although somewhat predictable in time, does vary from year to year as well.

The upshot of this discussion is that determining an exact amount of non-fertilizer nitrogen that will be available to corn in a given area in a given year at just the right growth stage is impossible. Thus, so is recommending how much nitrogen fertilizer will be needed in a given area in a given year. As a result of this uncertainty in how much fertilizer will be needed, agronomists have adopted a regional strategy for fertilizer recommendations. The

attempt is to communicate to producers the relative risks involved in under-fertilization and over-fertilization. Under-fertilization results in economic loss due to decreased yield. Over-fertilization results in economic loss due to paying for more fertilizer than was needed, and greater environmental damage than would be necessary. To help communicate these risks to producers, agronomists have defined two quantities as follows.

1. Lowest level of nitrogen fertilizer that produces maximum yield.

   Similar to algal growth in lakes, corn growth and the production of grain requires a number of necessary inputs. Fertilization with nitrogen is based on the knowledge that nitrogen is a limiting factor for corn. Nitrogen can be added in the form of fertilizer until it is no longer the active limiting factor. After this point, continuing to increase nitrogen provides no additional increase in yield. The level of nitrogen fertilizer that just produces the maximum yield is the point at which this occurs.

2. Economic optimal nitrogen rate.

   Even before the lowest nitrogen rate that produces maximum yield is reached, increases in corn yield in response to fertilizer follows a law of diminishing returns. That is, for given increments of increased fertilizer, as more and more increments are added, the corresponding increases in yield become less and less. The economic optimal rate of fertilization is defined as the lowest nitrogen rate above which adding additional fertilizer will not produce an increase in yield great enough to cover the cost of the additional fertilizer. The economic optimal rate will always be no more than the lowest rate that produces maximum yield.

What agronomists would like to be able to provide to producers are quantities such as the following:

- The distribution of the economic optimal rate for a region such as Iowa over a span of years and locations.

- The probability that a given level of nitrogen fertilization will be greater than needed to produce maximum yield.

- The distribution of the difference between a given nitrogen rate and the economic optimal rate for a region over a span of years.

2 Modeling Individual Nitrogen Trials

The data used by agronomists to determine the response of corn to nitrogen fertilization come from small-scale field studies called nitrogen trials. A typical nitrogen trial is based on a randomized complete block design in which five or six nitrogen fertilizer rates are applied to plots in each of four blocks, designed to help alleviate bias that might be caused by small-scale spatial effects. It is not really known that soil within blocks is homogeneous, or that differences occur among blocks. The blocking design is simply used as a way to protect against the possibility that there are such differences. The data we will be concerned with in this question come from a set of 43 nitrogen trials conducted in Iowa in the years 2001, 2002 and 2003. Each of the 43 trials used 6 levels of nitrogen at 0, 45, 90, 135, 180 and 225 kilograms per hectare (kg/ha). The response variables of interest are taken to be corn yield in 1000 kg/ha, and observations consist of average yield over 4 replicates (the blocks in the experimental design). Plots of yield against nitrogen rate are presented for four nitrogen trials in Figure 1.

Agronomists have used a regression model to describe the relation between yield and nitrogen rate that contains a piecewise quadratic response or expectation function. Specifically, let x_j represent nitrogen fertilizer rate and let Y_j be associated with yield at nitrogen rate x_j; $j = 1, \ldots, n$. The model used is,

$$Y_j = g(x_j, \boldsymbol{\beta}) + \sigma\epsilon_j, \tag{1}$$

where,

$$g(x_j, \boldsymbol{\beta}) = \begin{cases} \beta_0 + \beta_1 x_j + \beta_2 x_j^2 & x_j \leq \psi \\ \beta_0 + \beta_1 \psi + \beta_2 \psi^2 & x_j > \psi. \end{cases}$$

In (1) the ϵ_j are assumed to be independent and identically distributed with a location-scale family distribution having expected value 0 and variance 1, and ψ is a parameter that represents the lowest nitrogen rate that produces maximum yield.

ANSWER QUESTION 1 NOW (Section 4, 10)

Unease with model (1) caused me to search for other potential expectation functions for use in a nonlinear regression model. One that I want to consider uses what is called a spherical response function and results in the model,

$$Y_j = g(x_j, \boldsymbol{\theta}) + \sigma\epsilon_j, \tag{2}$$

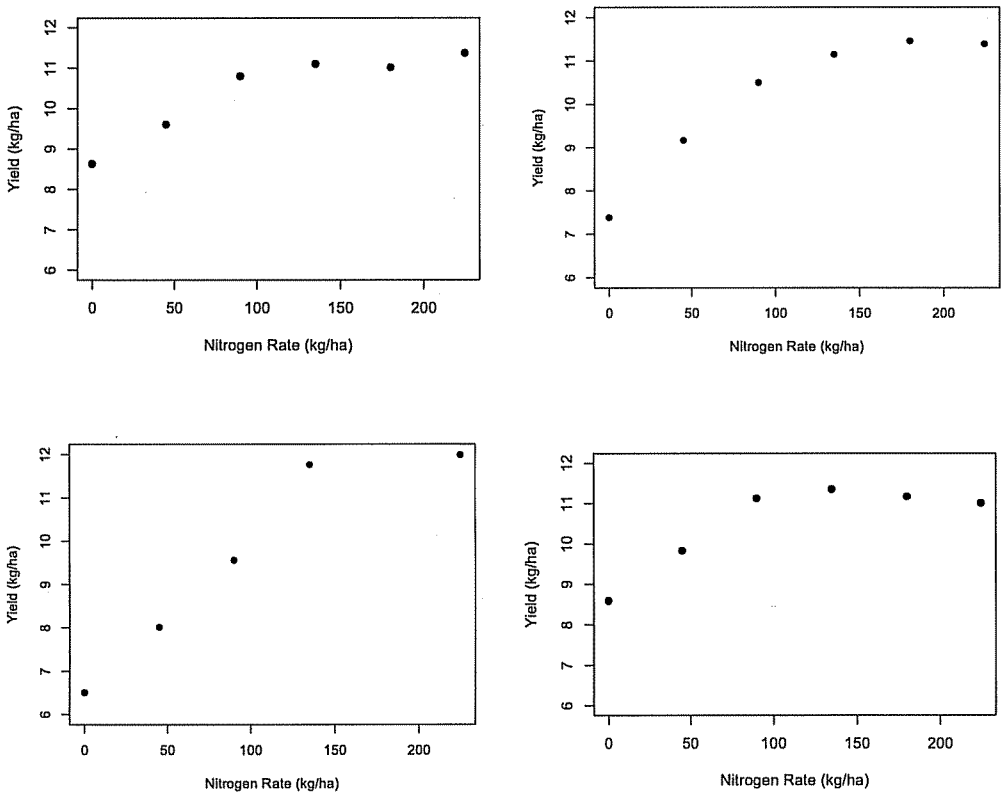


Figure 1: Examples of data from individual nitrogen trials.

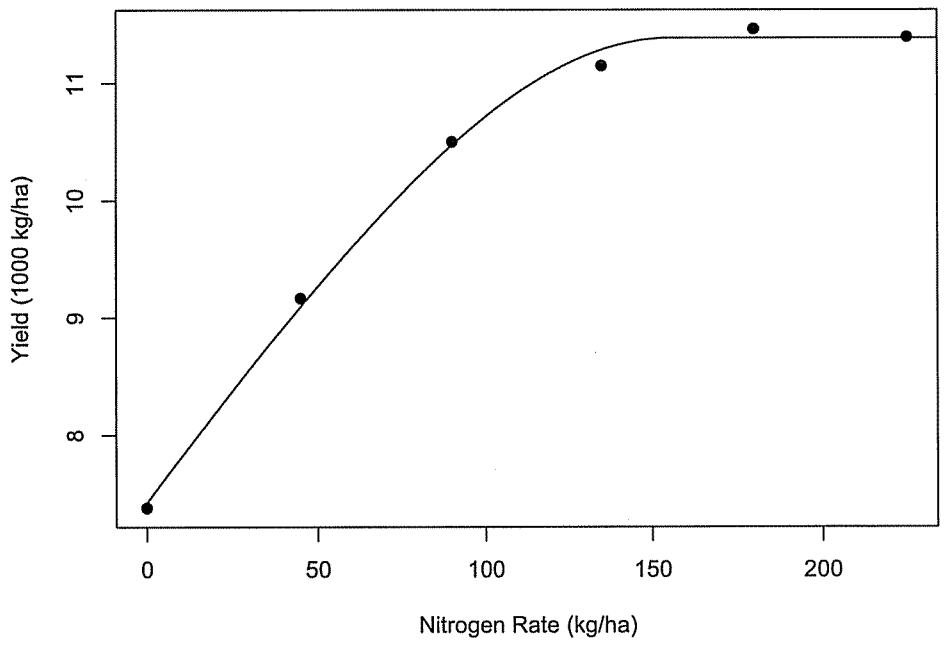


Figure 2: One nitrogen trial with fitted response function from model (2).

where, ϵ_j are assumed to be independent and identically distributed with a location-scale family distribution having expected value 0 and variance 1. In (2) we will let $\boldsymbol{\theta} = (c_0, c_1, a)^T$ and,

$$g(x_j, \boldsymbol{\theta}) = \begin{cases} c_0 + c_1 \left\{ \frac{3}{2} \left(\frac{x_j}{a} \right) - \frac{1}{2} \left(\frac{x_j}{a} \right)^3 \right\} & 0 < x_j \leq a \\ c_0 + c_1 & x_j > a \end{cases}$$

This response function increases fairly linearly to a plateau or maximum value at $c_0 + c_1$. That is, c_0 is an intercept term that represents yield with no fertilizer, $c_0 + c_1$ is the maximum yield, and a is the minimum nitrogen rate at which the yield reaches its maximum. To give you a visual impression of this response function, a plot of the model fitted to data from the nitrogen trial (estimation by generalized least squares) in the upper right panel of Figure 1 is presented in Figure 2.

Unfortunately, not all nitrogen trials produce data as pleasing as the trials exhibited in Figure 1. Data from four additional trials are presented in Figure 3 to illustrate this fact.

ANSWER QUESTION 2 NOW (Section 4, page11)

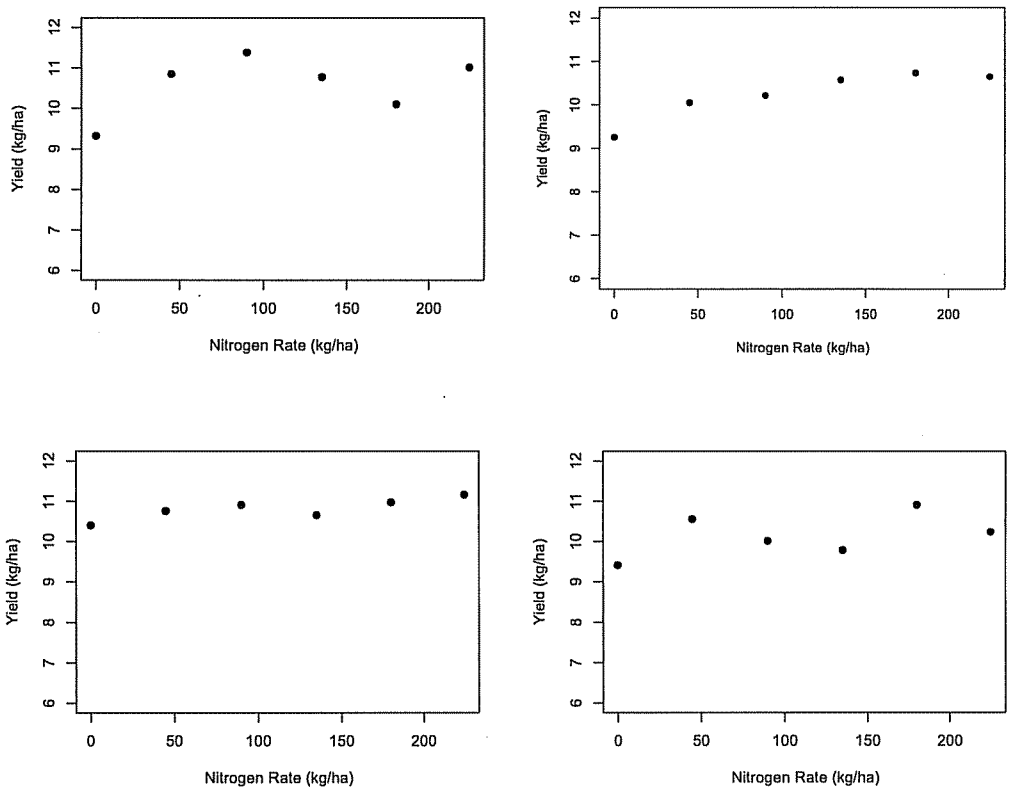


Figure 3: More examples of data from individual nitrogen trials.

I selected a subset of 28 nitrogen trials on which to conduct model development. This selection was by visual inspection of scatterplots such as those in Figure 1 and Figure 3 to produce what I would call a "nice" data set. The spherical model (2) was fit to each of these 28 trials individually using generalized least squares estimation. Stem-and-leaf plots of the parameter estimates are presented immediately below.

For estimated values $\hat{c}_0$:

```
The decimal point is at the |

 5 | 9
 6 | 39
 7 | 34499
 8 | 0566888
 9 | 33468
10 | 129
11 | 1135
12 | 5
```

For estimated values $\hat{c}_1$:

```
The decimal point is at the |

0 | 77
1 | 133688
2 | 567778
3 | 122567
4 | 0378
5 | 55
6 | 03
```

For estimated values $\hat{a}$:

```
The decimal point is 1 digit(s) to the right of the |

  4 | 6
  6 | 5749
  8 | 4478
 10 | 979
 12 | 2359
 14 | 36566
 16 | 0787
 18 | 0
 20 | 3
 22 | 7
```

Not surprisingly, these plots show a fair amount of variability in estimates of the parameters from model (2) when fitted to data from individual nitrogen trials. This prompted the development of a hierarchical model as described in the next section.

3 A Hierarchical Model for Nitrogen Trials

The starting point for development of a hierarchical model was to modify the notation of model (2) to allow consideration of more than one nitrogen trial at a time. Thus, now consider $x_{i,j}$ as nitrogen rate j in trial i, and associate the random variable $Y_{i,j}$ with yield for this nitrogen rate j in trial i. Generalize model (2) as,

$$Y_{i,j} = g(x_{i,j}, \boldsymbol{\theta}_i) + \sigma\epsilon_{i,j}, \tag{3}$$

and now take $\epsilon_{i,j} \sim iid\, N(0,1)$. In (3) we will let $\boldsymbol{\theta}_i = (c_{0i}, c_{1i}, a_i)^T$ and,

$$g(x_{i,j}, \boldsymbol{\theta}_i) = \begin{cases} c_{0i} + c_{1i}\left\{\frac{3}{2}\left(\frac{x_{i,j}}{a_i}\right) - \frac{1}{2}\left(\frac{x_{i,j}}{a_i}\right)^3\right\} & 0 < x_{i,j} \le a_i \\ c_{0i} + c_{1i} & x_{i,j} > a_i \end{cases}$$

A hierarchical model is completed by assigning distributions to the $\{c_{0i}, c_{1i}, a_i : i = 1, \ldots, T\}$ and σ^2, and also any parameters that appear in these distributions. We will use generic notation $\pi(\cdot)$ for a density, interpreted to mean the density of its argument, so

that $\pi(x)$ is the density of X and $\pi(y)$ is the density of Y (although π is not necessarily the same function). Thus, what is needed are distributions specified for $\pi(c_{0i}|\lambda_0)$, $\pi(c_{1i}|\lambda_1)$, $\pi(a_i|\lambda_a)$, these for $i = 1, \ldots, T$, and also $\pi(\sigma^2|\psi)$, $\pi(\lambda_0)$, $\pi(\lambda_1)$, and $\pi(\lambda_a)$. Note that in this generic notation $\psi, \lambda_0, \lambda_1$ and λ_a may be vectors. Clearly, we are now headed for a Bayesian analysis.

ANSWER QUESTION 3 NOW

The hierarchical model we have formulated (no matter what version of it you ended up with) is well suited to analysis through the use of posterior simulation via Gibbs Sampling. For this we will need full conditional posterior distributions, and we will use a similar generic notation here, with $p(x|\cdot)$ denoting the posterior density of X given all other quantities involved. For example, $p(c_{0i}|\cdot)$ will denote

$$p(c_{0i}|\boldsymbol{y}, \{c_{0j} : j \neq i\}, \{c_{1k} : k = 1, \ldots, T\}, \{a_k : k = 1, \ldots, T\}, \sigma^2, \psi, \lambda_0, \lambda_1, \lambda_a)$$

ANSWER QUESTION 4 NOW

Running a Gibbs algorithm using all of the full conditional posteriors produces a sample from the joint posterior

$$p(\{c_{0i}, c_{1i}, a_i : i = 1, \ldots, T\}, \sigma^2, \psi, \lambda_0, \lambda_1, \lambda_a|\boldsymbol{y}) \tag{4}$$

and any of the associated marginal distributions we desire, such as $p(\sigma^2|\boldsymbol{y})$. We now must determine which of these possibilities we should use for the purpose of inference, and this depends on the viewpoint we have about the manner in which our hierarchical structure is representing the problem.

ANSWER QUESTIONS 5 AND 6 NOW

4 Questions

1. Consider fitting model (1) to data from a single nitrogen trial. I picked what looked like the "nicest" nitrogen trial from Figure 1, which I decided was data from the upper right plot of that figure, and I tried to estimate parameters β_0, β_1, β_2 and ψ of the piecewise quadratic model in expression (1) using the R function *nls*. There seemed to be no problem in doing this. Taking the fitted object and running the *summary* function produced the following:

```
Formula: 0 ~ piecewise(x, y, b0, b1, b2, psi)


Parameters:
      Estimate Std. Error t value Pr(>|t|)
b0   7.382e+00  1.643e-01  44.926 0.000495 ***
b1   4.463e-02  9.309e-03   4.794 0.040859 *
b2  -1.109e-04  9.938e-05  -1.116 0.380448
psi  1.314e+02  4.274e+01   3.075 0.091485 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


Residual standard error: 0.1643 on 2 degrees of freedom


Number of iterations to convergence: 4
Achieved convergence tolerance: 2.234e-10
```

The estimated parameters do not look unreasonable (the intercept should be about 7 and the lowest nitrogen at maximum yield about 131, which match ok with the plot of Figure 2).

(a) Should model (1) be considered a linear model or a nonlinear model (justify your response)?

(b) The default algorithm for the R function *nls* is Gauss-Newton. Should I accept the point estimates in the above table as valid generalized least squares estimates? Should I accept the estimated standard errors in the above table as valid estimates of uncertainty in the sampling distributions of the parameter estimates?

NOTE: This question has to do with validity of estimates, not quality.

(c) Assuming I am willing to accept the estimates returned by R as valid (whether or not this is true), what is a concern with the quality of the estimates? If the agronomist I am working with wants to adhere to a strict 0.05 level for declaring "significance," how can I explain that yield with non-zero nitrogen significantly increases over yield at zero nitrogen (i.e., the test for $\beta_1 = 0$ is significant) and yet the nitrogen level at which yield first reaches its maximum, ψ, is not significantly greater than 0?

(d) Assuming I am not willing to accept the estimates returned by R as valid (again, whether or not this is a correct decision), suggest and outline an alternative procedure for estimation. You may assume that the ϵ_j in model (1) are independent and identically distributed with $N(0, 1)$ distributions if you wish.

(e) Regardless of whether I accept the R estimates as valid, suggest and outline an alternative procedure for finding standard errors or confidence intervals for the parameters of model (1). You may make the same distributional assumption as in question 1(d) if you wish.

2. The agronomist involved in this project indicated that some nitrogen trials in the data we were given would be considered "non-responsive" to fertilizer. I do not know how this determination is made, but let's assume it is simply a matter of judgment and visual inspection of the data on the part of agronomists. It would be pleasing to have a more formal process by which to declare a given nitrogen trial as responsive or not responsive to fertilizer.

(a) Argue that it would be unwise to demand a completely formal probability-based decision rule, and that we should either rely on the judgment of the agronomists here, or at least allow the possibility of some ad-hoc reasoning in the way we declare nitrogen trials responsive or non-responsive to fertilizer. You may want to review your answer to question 1(c) before you write your answer to this

question.

(b) One possibility for a totally ad-hoc decision rule would be to declare a trial non-responsive if any of the observed yields at non-zero nitrogen rates were less than the yield with no fertilizer (nitrogen rate 0). Another possibility that would fall into the completely ad-hoc category would be to declare a trial as non responsive only if none of the observed yields for non-zero nitrogen rates were increased over the yield with no fertilizer. What is the common weakness of these rather extreme ad-hoc procedures?

(c) What would you propose for declaring individual nitrogen trials as responsive or non-responsive? This may be any one of the possibilities already described, or may be something else that you want to propose.

3. Using any information given to this point, including plots and the description of the problem, assign specific distributional forms to $\pi(c_{0i}|\lambda_0)$, $\pi(c_{1i}|\lambda_1)$, $\pi(a_i|\lambda_a)$, these for $i = 1, \ldots, T$, and also $\pi(\sigma^2|\psi)$, $\pi(\lambda_0)$, $\pi(\lambda_1)$, and $\pi(\lambda_a)$, keeping in mind that ψ, λ_0, λ_1 and λ_a may be vectors. If the distributions you choose for $\pi(\lambda_0)$, $\pi(\lambda_1)$ and $\pi(\lambda_a)$ are themselves parameterized distributions you DO NOT need to indicate what values you might choose for those parameters.

   For each distribution you specify, indicate your motivation for the choice made, such as empirical evidence from a plot (indicate which plot and what aspect of it was important), mathematical convenience (keep in mind that we will be deriving posterior distributions), or lack of information. There is no one correct answer to this, but keep in mind ensuring proper support or approximation for the distributions you choose. The one stipulation given to you is that down the road we may want to compare regions (e.g., parent soil types) and would plan to do so through the use of Bayes Factors.

   HINT: Some of you are aware that I generally have discouraged the use of proper uniform priors on arbitrary intervals. For this problem ignore my concerns with that technique. In fact, in the model I actually used there are three such distributions.

4. Suppose that $\pi(c_{0i}|\lambda_0)$ has been specified as $N(\mu_0, \tau_0^2)$ (if you have something else in your answer to question 3 that's fine as long as you motivated it). Demonstrate that this results in conditional conjugacy in that $p(c_{0i}|\cdot)$ is also a normal distribution.

5. The major issue in inference involves the use of the conditional distributions $\{p(c_{0i}|\boldsymbol{y})$:

$i = 1, \ldots, T\}$, $\{p(c_{1i}|\boldsymbol{y}) : i = 1, \ldots, T\}$ and $\{p(a_i|\boldsymbol{y}) : i = 1, \ldots, T\}$ and, in turn, how we are thinking about the quantities $\{c_{0i}, c_{1i}, a_i : i = 1, \ldots, T\}$ within the context of the problem. *HINT: Re-read the problem background section.*

(a) Do you think that the distributions $p(c_{0i}|\boldsymbol{y})$, $p(c_{1i}|\boldsymbol{y})$ and $p(a_i|\boldsymbol{y})$ listed at the beginning of this question can be used directly to provide inferential statements of the type listed at the end of the problem background section? If yes, how should they be used to produce such statements? If no, why not, and are there any alternatives?

(b) Consider the first bulleted item at the end of the problem background section, the desire to provide a distribution for the economic optimal nitrogen rate, let's call it W. The agronomist tells us that this can be calculated for a given nitrogen trial as

$$W = \begin{cases} \left[\frac{1.5c_{1i}a_i^2 - a_i^3/R}{1.5c_{1i}}\right]^{1/2} & R > \frac{a_i}{1.5c_{1i}} \\ 0 & R \leq \frac{a_i}{1.5c_{1i}} \end{cases}, \tag{5}$$

where R is a ratio of corn price to nitrogen price. Assume you are given a value of R. Outline steps you would use to produce a distribution to *predict* W for new situations (assumed to follow the same model that we have analyzed).

6. Model assessment in this problem is an issue with many parts, but let's just focus on one particular aspect of the overall issue. In the model, we took c_{0i}, c_{1i} and a_i to be independent of each other for a given nitrogen trial (indexed by i). In a highly nonlinear response function such as used in model (3) this might not be entirely adequate. Briefly outline a process you might use to assess whether the values of c_{0i}, c_{1i} and a_i you used in question 5 to produce a predictive distribution for W are reflective of the association (or lack thereof) that we might expect among these values.

HINT: Recall our initial modeling of data from individual nitrogen trials.

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.

1. Question 1

   (a) Model (1) is a nonlinear model. Derivatives of the response function with respect to the parameters are,

$$\begin{aligned}
\frac{\partial}{\partial\beta_0}g(x_j,\boldsymbol{\beta}) &= 1 \\
\frac{\partial}{\partial\beta_1}g(x_j,\boldsymbol{\beta}) &= \begin{cases} x_j & x_j \le \psi \\ \psi & x_j > \psi \end{cases} \\
\frac{\partial}{\partial\beta_2}g(x_j,\boldsymbol{\beta}) &= \begin{cases} x_j^2 & x_j \le \psi \\ \psi^2 & x_j > \psi \end{cases} \\
\frac{\partial}{\partial\psi}g(x_j,\boldsymbol{\beta}) &= \begin{cases} 0 & x_j \le \psi \\ \beta_1 + 2\beta_2\psi & x_j > \psi \end{cases}
\end{aligned}$$

   These derivatives depend on the parameter values, making this a nonlinear model.

   (b) I should not uncritically accept these values as valid generalized least squares estimates, despite the fact that the point estimates look reasonable. I certainly should not accept the estimated standard errors. A Gauss-Newton algorithm is developed from a Taylor series expansion of the expectation function. While the expectation function in model (1) is continuous, its derivatives are not (as shown above), invalidating such an expansion.

   (c) The obvious concern is that we are estimating 5 parameters (β_0, β_1, β_2, ψ and σ^2) on the basis of 6 data values. Regardless of the fact that the R output contains a column labeled "t value," inference for this analysis of a nonlinear model is based on asymptotic results. Although there may be various views

about what sample size is needed for the asymptotics to produce reasonable approximations, 6 is not even in the ballpark. As a result, I cannot explain the paradox in test results other than by indicating such results are meaningless.

(d) Making a full distributional assumption on the ϵ_j in model (1) allows maximum likelihood as an estimation procedure. One could employ an unnormed profile likelihood procedure to estimate ψ. This could be accomplished by taking finely spaced values of ψ and maximizing the likelihood in the other parameters at each fixed value of ψ, then selecting that value that gives the maximum of these profile values. This is outlined as follows:

i. Specify a finely spaced set of values $\boldsymbol{\psi} = \{\psi 1, \ldots, \psi_K\}$.

ii. Let $\theta = (\beta_0, \beta_1, \beta_2, \sigma^2)^T$ and let $f(y_j|\theta, \psi)$ denote the density of Y_j evaluated at the observed value y_j with parameters θ and ψ. Define $L(\theta, \psi) = \sum \log\{f(y_j|\theta, \psi)\}$ and

$$L_p(\psi_k) = \max_{\theta} L(\theta, \psi_k); \quad k = 1, \ldots, K$$

iii. The maximum likelihood estimate of ψ is then

$$\hat{\psi} = \max_{\boldsymbol{\psi}} L_p(\psi_k)$$

(e) Inference based on either generalized least squares or maximum likelihood estimates is asymptotic so that, regardless of whether I accept the R output as valid generalized least squares point estimates or I produce maximum likelihood estimates through profiling, I would like a non-asymptotic procedure for producing standard errors. If we are willing to specify that the ϵ_j are normal, this might be accomplished through the use of parametric bootstrap as follows:

i. Using point estimates of the parameters β_0, β_1, β_2, ψ and σ^2, simulate a set of data from the model using nitrogen rates (covariate values) as contained in the actual data.

ii. Estimate parameters for the simulated data using the same method that was applied to the actual data.

iii. Repeat steps 1 and 2 a large number of times, collecting bootstrap estimates over the repetitions.

iv. Estimate standard errors as the square root of the sample variance of the bootstrap estimates.

v. Produce confidence intervals as either basic bootstrap intervals or percentile bootstrap intervals as discussed in class.

2. Question 2.

(a) A fully probability-based decision rule is unlikely to be reliable here because of the small number of observations available for each trial. With limited information available in the data from an individual trial we are better off relying on the expert judgement of the agronomists.

(b) The common weakness of suggestions such as contained in the question is that variability is being assessed in an inconsistent manner across nitrogen rates. Both of these suggestions allow for variability in responses that receive non-zero nitrogen (although in different ways) but both take the value at zero nitrogen to be essentially observed without error.

(c) One possiblity is to define a trial as non-responsive if it does not seem possible or is at least extremely difficult to locate parameter estimates. One cannot make this a completely formal rule, because an inability to find starting values that will cause an iterative estimation algorithm to converge does not imply that such values fail to exist.

3. Question 3.

Correctness of answers here will depend on motivations given for various choices. My selections would be:

- Take $c_{01}, \ldots, c_{0T}$ to be independent and identically distributed such that $\pi(c_{0i}|\lambda_0)$ is normal $N(\mu_0, \tau_0^2)$ for $i = 1, \ldots, T$.

  Motivation comes from the stem-and-leaf plot for values of $\hat{c}_0$ on page 8 of the

question, which is unimodal, roughly symmetric, and appears to have small enough variance so that representation by a normal distribution would avoid placing substantial probability on the negative line.

- Take c_{1i} to be independent and identically distributed such that $pi(c_{1i}|\lambda_1)$ is gamma $Ga(\alpha, \beta)$ for $i = 1, \ldots, T$.

  Motivation comes from the stem and leaf plot for values of $\hat{c}_1$ on page 8 of the question which contains small values near zero and appears somewhat skewed right.

- Take a_i to be independent and identically distributed such that $\pi(a_i|\lambda_a)$ is normal $N(\mu_a, \tau_a^2)$ for $i = 1, \ldots, T$.

  Motivation comes partly from the stem-and-leaf plot for values of $\hat{a}$ on page 9 of the question, and partly from a lack of obviously better ideas. This distribution is located high on the number line and has a large variance. With a lack of strong evidence that a normal distribution is not reasonable, it becomes the default choice.

- Take $\sigma^2 \sim InvGa(\psi_1, \psi_2)$. The motivation for this choice is conditional conjugacy.

- With the selections above, $\lambda_0 = (\mu_0, \tau_0^2)$. Form a joint distribution as $\pi(\mu_0, \tau_0^2) = \pi(\mu_0)\pi(\tau_0^2)$. We might take $\mu_0 \sim N(M_0, V_0)$ by conditional conjugacy for combining with the normally distributed c_{0i}, and $\tau_0^2 \sim Unif(0, K_0)$ for some selected value K_0. This is motivated through a lack of knowledge about what would work "well" for τ_0^2 and is specified as a proper distribution so that Bayes factors will exist, as stipulated in the question.

- With the selections for data model parameter distributions, $\lambda_1 = (\alpha, \beta)$. Form a joint distribution $\pi(\alpha, \beta) = \pi(\alpha)\pi(\beta)$. We might take $\beta \sim Ga(\eta, \phi)$ based on conditional conjugacy, and $\alpha \sim Unif(0, K_a)$ for some specified value K_a based on a lack of knowledge for what else would be better. Note that this is not the same as attempting to place a non-informative prior on α. The reason for keeping this distribution proper is the same as for the distribution on τ_0^2.

- With the selections for distributions of data model parameters, $\lambda_a = (\mu_a, \tau_a^2)$. Form a joint distribution as $\pi(\mu_0, \tau_0^2) = \pi(\mu_0)\pi(\tau_0^2)$. We might take $\mu_a \sim N(M_a, V_a)$ and $\tau_a^2 \sim Unif(0, K_a)$ for exactly the same reasons we took μ_0 normal and τ_0^2 uniform.

4. Question 4.

Let $\boldsymbol{y}_i = (y_{i,1}, \ldots, y_{i,n_i})$ be the response observations from nitrogen trial i, corresponding to the random variables $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{in_i})$; $i = 1, \ldots, T$. Let $f(\boldsymbol{y}_i | c_{0i}, c_{1i}, a_i, \sigma^2)$ be the joint density of $\boldsymbol{Y}_i$ evaluated at the observed $\boldsymbol{y}_i$. Using the result that for hierarchical models such as this, a conditional posterior is proportional to the way it is specified in the model times anything specified as conditional on it,

$$\begin{aligned} p(c_{0i}|\cdot) &\propto \pi(c_{0i}|\mu_0, \tau_0^2)\, f(\boldsymbol{y}_i|c_{0i}, c_{1i}, a_i, \sigma^2) \\ &\propto \exp\left[-\frac{1}{2\tau_0^2}(c_{0i} - \mu_0)^2\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}\{y_{i,j} - g(x_{i,j}, c_{0i}, c_{1i}, a_i)\}^2\right]. \end{aligned} \tag{1}$$

From the model,

$$g(x_{i,j}, c_{0i}, c_{1i}, a_i) = \begin{cases} c_{0i} + c_{1i}\left\{\frac{3}{2}\left(\frac{x_{i,j}}{a_i}\right) - \frac{1}{2}\left(\frac{x_{i,j}}{a_i}\right)^3\right\} & 0 < x_{i,j} \le a_i \\ c_{0i} + c_{1i} & x_{i,j} > a_i \end{cases}$$

Let

$$t_{i,j} = \begin{cases} y_{ij} - c_{1i}\left\{\frac{3}{2}\left(\frac{x_{i,j}}{a_i}\right) - \frac{1}{2}\left(\frac{x_{i,j}}{a_i}\right)^3\right\} & 0 < x_{i,j} \le a_i \\ y_{i,j} - c_{1i} & x_{i,j} > a_i \end{cases}$$

Then (1) can be written as,

$$p(c_{0i}|\cdot) \propto \exp\left[-\frac{1}{2\tau_0^2}(c_{0i} - \mu_0)^2\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n_i}\{t_{i,j} - c_{0i}\}^2\right]. \tag{2}$$

Combining terms in the exponent of (2) and completing the square shows that $p(c_{0i}|\cdot)$ is $N(A, B)$ where

$$\begin{aligned} A &= \frac{\sigma^2\mu_0 + \tau_0^2\sum_j t_{i,j}}{\sigma^2 + n_i\tau_0^2} \\ B &= \frac{\sigma^2\tau_0^2}{\sigma^2 + n_i\tau_0^2} \end{aligned}$$

5. Question 5.

   (a) No. The individual posteriors $p(c_{0i}|\boldsymbol{y})$, $p(c_{1i}|\boldsymbol{y})$ and $p(a_i|\boldsymbol{y})$ for $i = 1, \ldots, T$ are not useful to directly make inferential statements such as those given at the end of the problem description. For a given i these posteriors are particular to that nitrogen trial which, from the description of the problem, are unique events. It is not possible even in principle to ever obtain another observation from the model with the same values of c_{0i}, c_{1i} and a_i. Thus, while these posteriors reflect what we believe about the values of these parameters for nitrogen trial i, they are not useful in making inference aobut distributions over regions or time spans such as those given in the problem background section. One way to make this clear is to consider what would be possible if we had perfect knowledge and could say exactly what the values of c_{0i}, c_{1i} and a_i were for nitrogen trial i. The collection of such values across trials $i = 1, \ldots, T$ would not allow inferential statements such as those desired unless we embedded them in a larger structure (i.e., a distribution over a set or population of nitrogen trials from which the observed trials have been drawn).

   The clearest alternative is to make use of posterior predictive distributions for c_{0i}, c_{1i} and a_i. The posterior predictive for a_i is, in fact, the predictive distribution of the lowest nitrogen rate needed to obtain maximum yield. This distribution could be used to make inferences such as that of the second bullet at the end of the problem background section, namely the probability that a given fertilizer level will exceed that needed to produce maximum yield.

   (b) A predictive distribution for W (or any other quantity that can be expressed as a function of c_{0i}, c_{1i} and a_i for that matter) can be produced as follows:

      i. In the Gibbs algorithm, for each draw of λ_0, λ_1 and λ_a from the joint posterior distribution (i.e., after burn-in), simulate values of c_{0i}, c_{1i} and a_i from $\pi(c_{0i}|\lambda_0)$, $\pi(c_{1i}|\lambda_1)$ and $\pi(a_i|\lambda_a)$. Call these values c_{0k}, c_{1k} and a_k.

      ii. Using c_{0k}, c_{1k} and a_k, compute a value of W, say W_k.

iii. The collection of values $\{W_k : k = 1, \ldots, M\}$ is a sample from the posterior predictive distribution of W.

6. Question 6.

Along with quantities such as W, the posterior predictive distribution of (c_{0i}, c_{1i}, a_i) can be used to assess the correlation among these values. That is, with $p(c_{0k}, c_{1k}, a_k|\boldsymbol{y})$ denoting the posterior predictive distribution of the data model parameters. Let $\hat{\theta}_i$ denote the set of generalized least squares estimates of model (2) fit to nitrogen trial i separately, $i = 1, \ldots, T$, and let $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_T)$. Conduct the following steps:

(a) Compute correlations C^{act} among elements of $\hat{\theta}$.

(b) Simulate M sets of T values each from $p(c_{0k}, c_{1k}, a_k|\boldsymbol{y})$, where T is the actual number of nitrogen trials in the data set,

$$\theta^*_m = \{(c_{0k}, c_{1k}, a_k) : k = 1, \ldots, T\}; \quad m = 1, \ldots, M.$$

(c) Compute correlations C^{sim}_m among elements of θ^*_m; $m = 1, \ldots, M$.

(d) Compute a posterior predictive p-value for correlation between any two elements (c_0, c_1, a) as,

$$P^* = \sum_{m=1}^{M} I\left(C^{act} \leq C^{sim}_m\right)/M.$$