PhD Prelim Exam
METHODS

Summer 2009
(Given on 7/7/09)

An experiment was conducted to study the effects of an amino acid supplement on egg production of hens. A total of 55 hens were assigned to 11 treatment groups using a completely randomized design with 5 hens per treatment. All hens within a particular treatment group received daily feed containing the same amount of the amino acid supplement. The amount of the supplement varied across treatment groups as specified in Table 1.

The mass of the eggs produced by each hen was recorded daily. At the end of the study, the total mass of the eggs produced by each hen was computed and divided by the number of days of the study to obtain one egg-mass-per-day measurement (g/day) for each hen. Averages of these measurements for each treatment group are provided in Table 1.

Table 1. A summary of the experimental data.

Treatment Group	Amount of Supplement (g/day)	Number of Hens	Average Egg Mass Per Day (g/day)
1	0.0	5	17.32
2	0.5	5	23.52
3	1.0	5	35.70
4	1.5	5	36.82
5	2.0	5	40.94
6	2.5	5	41.74
7	3.0	5	42.56
8	3.5	5	49.24
9	4.0	5	56.42
10	4.5	5	57.90
11	5.0	5	54.46

1. A single-factor analysis of variance was used to analyze the data. Write down the complete model for the data that matches this analysis. Let y_{ij} denote the j^{th} observation from treatment group i, and denote the mean for treatment group i by μ_i ($i = 1, \ldots, 11$; $j = 1, \ldots, 5$). Denote the error variance by σ^2.

2. The mean squared error for the ANOVA analysis is 37.2. The sample variance of the 11 averages in the last column of Table 1 is 168.4. Compute the F-statistic for testing the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_{11},$$

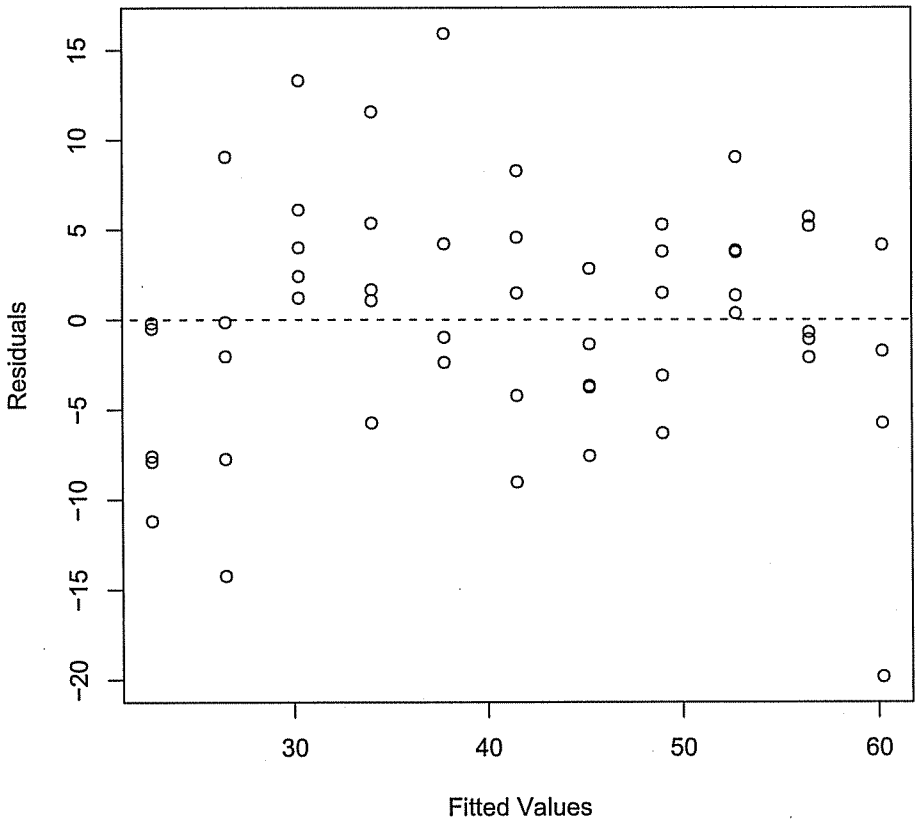
provide the degrees of freedom associated with the test statistic, and state the conclusion of the test at the 0.05 significance level.

3. Provide a 95% confidence interval for μ_5.

4. Provide a 95% prediction interval for an egg-mass-per-day measurement (obtained under the same conditions as the current study) from a new hen fed 2 grams of the supplement per day.

5. Provide a 95% prediction interval for the average of 10 egg-mass-per-day measurements (obtained under the same conditions as the current study) from 10 new hens fed 2 grams of the supplement per day.

6. The researchers wish to consider a simple linear regression model for the data with egg mass per day as the response and amount of the supplement as the explanatory variable. A plot of residuals vs. fitted values from the simple linear regression is provided on the next page. Explain, in general, the simple linear regression model assumptions that can be checked by examining a plot of residuals vs. fitted values. Based on this particular plot, do you have any concerns about the appropriateness of the simple linear regression model for this data set? Explain.



7. The mean squared error from the fit of the simple linear regression model is 44.3. Conduct a lack-of-fit test for the simple linear regression model. State the test statistic, its degrees of freedom, and a conclusion at the 0.05 significance level that indicates whether or not you think the linear regression model is appropriate for this data set.

8. Now suppose the researchers decide to fit the following model to the data.

$$y_i = \alpha_0 B_0(x_i) + \alpha_1 B_1(x_i) + \alpha_2 B_2(x_i) + \alpha_3 B_3(x_i) + \alpha_4 B_4(x_i) + \alpha_5 B_5(x_i) + \epsilon_i$$

for $i = 1, \ldots, 55$. Here, we assume that the observations are sorted in increasing order by treatment group with y_i the egg-mass-per-day measurement for hen i and x_i the amount of supplement for hen i ($x_1 = \cdots = x_5 = 0$, $x_6 = \cdots = x_{10} = 0.5$, ..., $x_{51} = \cdots = x_{55} = 5$). The random variables $\epsilon_1, \ldots, \epsilon_{55}$ are assumed to be independent and identically distributed as $N(0, \sigma_\epsilon^2)$. The values $\alpha_0, \alpha_1, \ldots, \alpha_5$ are unknown parameters. The functions $B_0, B_1, \ldots, B_5$ are defined as follows.

$$B_0(x) = \begin{cases} 1 - x & \text{for } x \in [0, 1] \\ 0 & \text{otherwise} \end{cases}, \qquad B_5(x) = \begin{cases} x - 4 & \text{for } x \in (4, 5] \\ 0 & \text{otherwise} \end{cases},$$

$$\text{and, for } k = 1, 2, 3, 4, \quad B_k(x) = \begin{cases} 1 - k + x & \text{for } x \in (k-1, k] \\ 1 + k - x & \text{for } x \in (k, k+1] \\ 0 & \text{otherwise.} \end{cases}$$

Plot the functions $B_0, \ldots, B_5$ on a single graph. Use a solid line for B_0 and B_3, a dashed line for B_1 and B_4, and a dotted line for B_2 and B_5.

[*Hint*: Each function of x on the right side of the equations above is linear with a slope of either 1 or -1. Examine the values of each of these linear functions at the endpoints of its interval.]

9. What do functions in the set

$$\{\alpha_0 B_0(\cdot) + \alpha_1 B_1(\cdot) + \alpha_2 B_2(\cdot) + \alpha_3 B_3(\cdot) + \alpha_4 B_4(\cdot) + \alpha_5 B_5(\cdot) : \alpha_0, \alpha_1, \ldots, \alpha_5 \in \mathbb{R}\}$$

look like? Provide a precise description of this set of functions.

[*Hint*: What are the values of

$$\alpha_0 B_0(k) + \alpha_1 B_1(k) + \alpha_2 B_2(k) + \alpha_3 B_3(k) + \alpha_4 B_4(k) + \alpha_5 B_5(k)$$

for $k \in \{0, 1, \ldots, 5\}$? What functional form does $\sum_{k=0}^{5} \alpha_k B_k(x)$ have for $x \in [0, 1]$? For $x \in [1, 2]$? Is $\sum_{k=0}^{5} \alpha_k B_k(x)$ continuous on $(0, 5)$?]

10. For what values of $\alpha_0, \alpha_1, \ldots, \alpha_5$ is the model in part 8 equivalent to the simple linear regression model discussed in part 6?

11. Let $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_5)'$, and let $\mathbf{0}$ denote a vector of zeros. Determine a matrix M such that $H_0 : M\boldsymbol{\alpha} = \mathbf{0}$ is equivalent to $H_0 : E(y|x) = \beta_0 + \beta_1 x$ for some $\beta_0, \beta_1 \in \mathbb{R}$.

12. Let $\boldsymbol{y} = (y_1, \ldots, y_{55})'$ and X be the matrix whose i, j^{th} element is $B_{j-1}(x_i)$ for $i = 1, \ldots, 55$ and $j = 1, \ldots, 6$. Write down an estimate of $M\boldsymbol{\alpha}$ in terms of $\boldsymbol{y}$, X, and M.

13. Provide a test statistic for testing $H_0 : M\boldsymbol{\alpha} = \mathbf{0}$ in terms of $\boldsymbol{y}$, X, and M.

14. State the distribution of the test statistic in part 13 under H_0.

15. When the model given in part 8 is fit to the data, the mean squared error is 34.5. Using this information together with information provided in other parts of this exam question, compute the value of the test statistic in part 13.

    [*Hint*: Do not try to evaluate the expression in part 13 directly. The statistic can be computed on a hand calculator in less than a minute with no matrix inversion if you know another expression for the statistic in part 13.]

16. Which of the three models considered in this exam question (see parts 1, 6, and 8) do you believe is most appropriate for the data?

17. Regardless of your answer to the previous question, assume that the researchers will rely on the model from part 8 for their analysis. When this model is fit to the data, the least squares estimate of α is

$$(16.38, 34.40, 40.68, 42.41, 57.04, 55.32)'.$$

The $(X'X)^{-1}$ matrix (with entries rounded to three decimal places) is

$$\begin{bmatrix} 0.166 & -0.028 & 0.005 & -0.001 & 0.000 & 0.000 \\ -0.028 & 0.142 & -0.024 & 0.004 & -0.001 & 0.000 \\ 0.005 & -0.024 & 0.141 & -0.024 & 0.004 & -0.001 \\ -0.001 & 0.004 & -0.024 & 0.141 & -0.024 & 0.005 \\ 0.000 & -0.001 & 0.004 & -0.024 & 0.142 & -0.028 \\ 0.000 & 0.000 & -0.001 & 0.005 & -0.028 & 0.166 \end{bmatrix}$$

The researchers would like to know if there is evidence to support the claim that too much of the dietary supplement will cause a decrease in mean egg mass per day. Provide a test to answer this question. Determine a test statistic, its degrees of freedom, and a conclusion at the 0.05 level.

18. Suppose that hen owners profit 2 cents per gram of egg mass per day for each hen. Suppose the supplement costs 20 cents per gram. Assuming the model in part 8 is correct, estimate the ideal amount of supplement in the interval 0 to 5 grams for an owner who would like to maximize expected profit.

19. Provide a confidence region to accompany your estimate in part 18. The confidence level of the confidence region should be at least 95%. You may assume that if multiple amounts of the supplement yield identical profit, the owner would prefer the least amount of the supplement that yields maximum profit. Thus, your confidence region should be for the smallest amount of the supplement that will provide maximum profit.

1.

$$y_{ij} = \mu_i + \varepsilon_{ij} \text{ for } i = 1, \ldots, 11 \text{ and } j = 1, \ldots, 5.$$

Here $\{\varepsilon_{ij} : i = 1, \ldots, 11; j = 1, \ldots, 5\}$ is a set of independent and identically distributed $N(0, \sigma^2)$ random variables.

2.

$$\text{MS}_{\text{treatment}} = \frac{\sum_{i=1}^{11} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{11-1} = 5\frac{\sum_{i=1}^{11}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2}{11-1} = 5 * 168.4 = 842.$$

$$F = 842/37.2 \approx 22.6.$$

This F statistic has 10 numerator and 44 denominator degrees of freedom. The 0.95 quantile of an F distribution with 10 and 44 degrees of freedom is approximately 2.05. Thus, we reject the null hypothesis at the 0.05 level.

3.

$$\bar{y}_{5\cdot} \pm t_{1-.05/2,44}\sqrt{\text{MSE}/5} \iff 40.94 \pm 2.02\sqrt{37.2/5} \iff (35.4, 46.4).$$

4.

$$\bar{y}_{5\cdot} \pm t_{1-.05/2,44}\sqrt{\text{MSE}(1+1/5)} \iff 40.94 \pm 2.02\sqrt{37.2(1+1/5)} \iff (27.4, 54.4).$$

5.

$$\bar{y}_{5\cdot} \pm t_{1-.05/2,44}\sqrt{\text{MSE}(1/10+1/5)} \iff 40.94 \pm 2.02\sqrt{37.2(1/10+1/5)} \iff (34.2, 47.7).$$

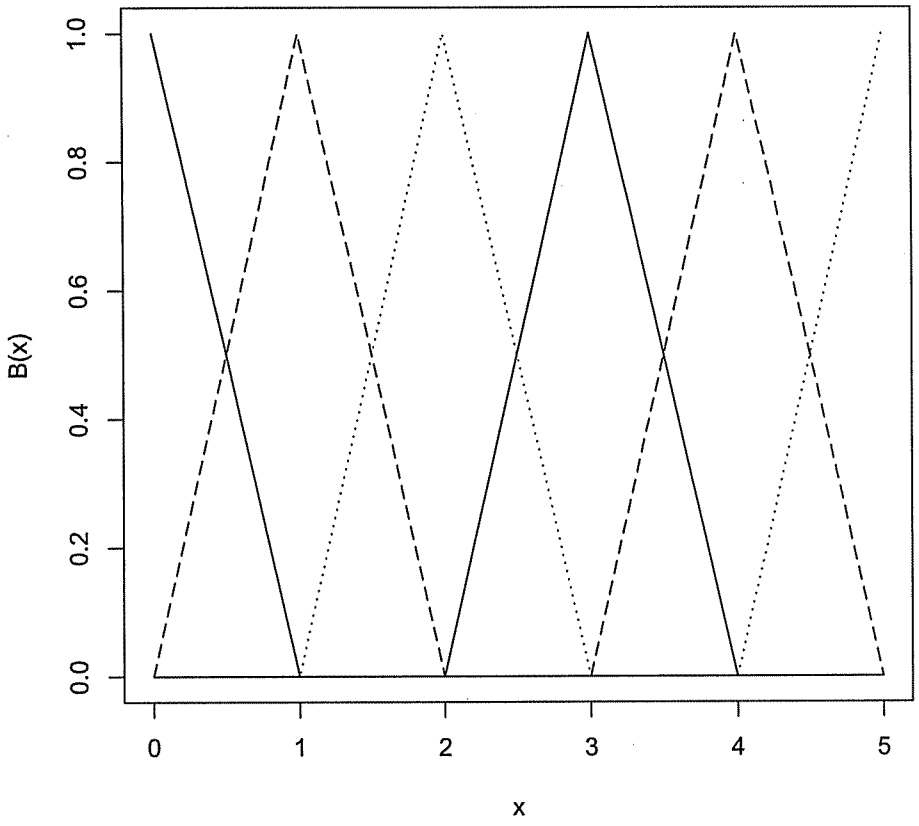
6. Residual plots can be used to visually assess whether there is a linear relationship between the mean of the response and the explanatory variable. They can also be used to check whether the assumption of constant residual variance is appropriate. Some information about normality of the error distribution can be obtained, but this assumption is easier to check with other plots. Independence among errors can be assessed, but evidence for a departure from the independence assumption is confounded with evidence for lack of linearity.

   This residual plot shows some evidence against the linearity assumption. This can be seen by noting a "wavy" pattern in the residual mean across the horizontal axis and noting that three treatment groups have all residuals above 0 or all residuals below 0 – an unlikely occurrence under the assumptions of the model.

7. $$F = \frac{(53 * 44.3 - 44 * 37.2)/(53 - 44)}{37.2} \approx 2.12 > 2.10 \approx F_{0.95,9,44} \implies \text{reject } H_0.$$

   There is evidence of a lack of fit.

8.



9. Functions in the set are 0 outside the interval $[0, 5]$ and continuous and piecewise linear on the interval $(0, 5)$ with potential changes in slope at $x = 1, 2, 3$, and 4.

10. The function $\sum_{k=0}^{5} \alpha_k B_k(x)$ is continuous and piecewise linear. Thus, the function will be linear as long as the slopes of each piece are equal. The slope on the segment $(k-1, k]$ is $\alpha_k - \alpha_{k-1}$. Thus, the set of $\boldsymbol{\alpha}$ vectors that yield a linear function is

$$\{\boldsymbol{\alpha} \in I\!R^6 : \alpha_1 - \alpha_0 = \alpha_2 - \alpha_1 = \alpha_3 - \alpha_2 = \alpha_4 - \alpha_3 = \alpha_5 - \alpha_4\}.$$

11. From the previous part we know that $\sum_{k=0}^{5} \alpha_k B_k(x)$ will be linear if

$$\alpha_k - \alpha_{k-1} = \alpha_{k+1} - \alpha_k \quad \text{for all } k = 1, 2, 3, 4.$$

This condition is equivalent to

$$\alpha_{k-1} - 2\alpha_k + \alpha_{k+1} = 0 \quad \text{for all } k = 1, 2, 3, 4.$$

Thus,

$$M = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix}$$

is an appropriate choice.

12.

$$M\hat{\boldsymbol{\alpha}} = M(X'X)^{-1}X'\boldsymbol{y}$$

13.

$$\frac{\boldsymbol{y}'X(X'X)^{-1}M'(M(X'X)^{-1}M')^{-1}M(X'X)^{-1}X'\boldsymbol{y}/4}{(\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'X(X'X)^{-1}X'\boldsymbol{y})/49}$$

14. F with 4 anf 49 degrees of freedom.

15.

$$\frac{(53 * 44.3 - 49 * 34.5)/4}{34.5} \approx 4.76$$

16. We have already seen that the simple linear regression exhibits lack of fit relative to the ANOVA model. Also, the test statistic 4.76 is significant at below the 0.01 level. Thus, the linear regression model is not appropriate. It remains to compare the model in part 8 to the ANOVA model. The F-statistic comparing full and reduced models is

$$\frac{(49 * 34.5 - 44 * 37.2)/(49 - 44)}{37.2} \approx 0.29.$$

The p-value is quite large. Thus, there is no significant lack of fit for the model from part 8 relative to the ANOVA model. If there is a scientific reason behind fitting the continuous piecewise linear model, it would be preferable to the ANOVA model.

17. There would be evidence that too much of the supplement can cause a decrease in egg production if $\alpha_4 > \alpha_5$. Our test statistic is

$$\frac{57.04 - 55.32}{\sqrt{34.5(0.142 + .166 - 2 * (-0.028))}} \approx 0.49.$$

Compared to a t-distribution with 49 degrees of freedom, the p-value is large. Thus, there is no evidence that too much of the supplement leads to a decrease in egg production. Of course, we don't know anything about what happens after 5.0 grams.

18. The expected profit is given by

$$2[16.38B_0(x) + 34.40B_1(x) + 40.68B_2(x) + 42.41B_3(x) + 57.04B_4(x) + 55.32B_5(x)] - 20x$$

for $x \in [0, 5]$. This is a continuous piecewise linear function that connects the points $(0, 32.76)$, $(1, 48.80)$, $(2, 41.36)$, $(3, 24.82)$, $(4, 34.08)$, and $(5, 10.64)$. Thus, the expected profit is maximized at 1 gram of the supplement.

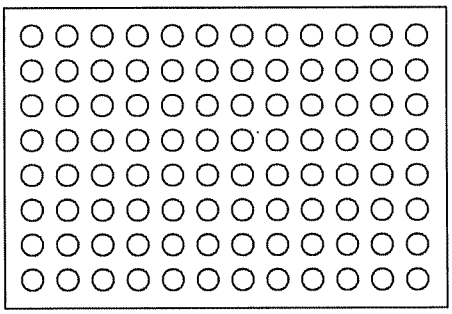
19. Because we are assuming that the true relationship between $E(y|x)$ and x is continuous and piecewise linear on $[0, 5]$ with potential slope changes at 1,2,3, or 4; it follows that the true expected profit function is continuous and piecewise linear on $[0, 5]$ with potential slope changes at 1,2,3, or 4. Because we are interested in the smallest x that maximizes the function, we need only consider $x = 0, 1, 2, 3, 4$, and 5. A confidence region with coverage at least 95% can be obtained by finding simultaneous confidence intervals for the profit function at $x = 0, 1, 2, 3, 4$, and 5. Bonferroni intervals are given by

$$\begin{aligned}
2 * (\hat{\alpha}_0 \pm t_{1-0.05/12,49}\sqrt{34.5 * 0.166}) &\iff (19.60, 45.92) \\
2 * (\hat{\alpha}_1 \pm t_{1-0.05/12,49}\sqrt{34.5 * 0.142}) - 20 &\iff (36.63, 60.97) \\
2 * (\hat{\alpha}_2 \pm t_{1-0.05/12,49}\sqrt{34.5 * 0.141}) - 40 &\iff (29.23, 53.49) \\
2 * (\hat{\alpha}_3 \pm t_{1-0.05/12,49}\sqrt{34.5 * 0.141}) - 60 &\iff (12.69, 36.95) \\
2 * (\hat{\alpha}_4 \pm t_{1-0.05/12,49}\sqrt{34.5 * 0.142}) - 80 &\iff (21.91, 46.25) \\
2 * (\hat{\alpha}_5 \pm t_{1-0.05/12,49}\sqrt{34.5 * 0.166}) - 100 &\iff (-2.52, 23.80)
\end{aligned}$$

Comparing each upper endpoint to all lower endpoints, we can see that only the last upper endpoint is not greater than all other lower endpoints. Thus, a confidence set for the amount of supplement that maximizes profit with coverage no less than 0.95 is $\{0, 1, 2, 3, 4\}$.

This question is motivated by issues that arise in the development of a type of so-called ELISA (Enzyme-Linked ImmunoSorbent Assay). This is a methodology widely used to determine the amount of an antibody or antigen in a biological specimen. In the version we will use as motivation here, liquid specimens containing an antibody of interest are serially diluted and (beginning at some appropriate level of dilution) successive dilutions of the original are placed into successive wells in a single row of the rectangular grid of wells on a polystyrene microtiter plate. For sake of concreteness, we will here assume that plates with 96 wells laid out in an "8 row by 12 column" rectangular array as illustrated in Figure 1 are used.

Figure 1: Schematic of Microtiter Plate



In the case we consider, after appropriate processing with biological agents, the material in a given well darkens and opacity of the well (as measured by the amount of light of a particular wave-length absorbed when the plate is placed into a "plate reader" capable of determining this absorbance) is an indicator of the amount of antibody present in that well. In particular, if a given row contains material from the serial dilution of a reference specimen, comparison of the amount of light absorbed on another row to that of the reference row provides a means of comparing the amounts of antibody in the original specimens (unknown and reference). In fact, methodology for the determination of the strength of an unknown specimen relative to that of a reference specimen was the primary goal in the scenario that motivated this question.

Notice that the concepts of the "top" row (row 1) and "left" column (column 1) of a plate are meaningful, as they are distinguishable on plates in terms of higher dilutions being placed on the right side and in terms of the left side of the plate (with lower dilutions) entering the reader before the right side.

PART 1

Standard practice in the lab developing this assay is to try to make every successive dilution of a specimen half the strength of the previous one (by, mixing nominally equal parts of the previous dilution and a clear liquid). A well on a plate then ideally has half as much antibody as the one immediately to its left.

The common expectation in the lab is then that successive measured absorbances (left to right across a row on a test plate) are approximately a sigmoid-shaped function of the logarithm of dilution number. For sake of concreteness, with

x = dilution number

($x = 2$ corresponding to half original specimen strength, $x = 4$ corresponding to one fourth original specimen strength, etc.) and

y = measured absorbance (in "machine" units)

assume that

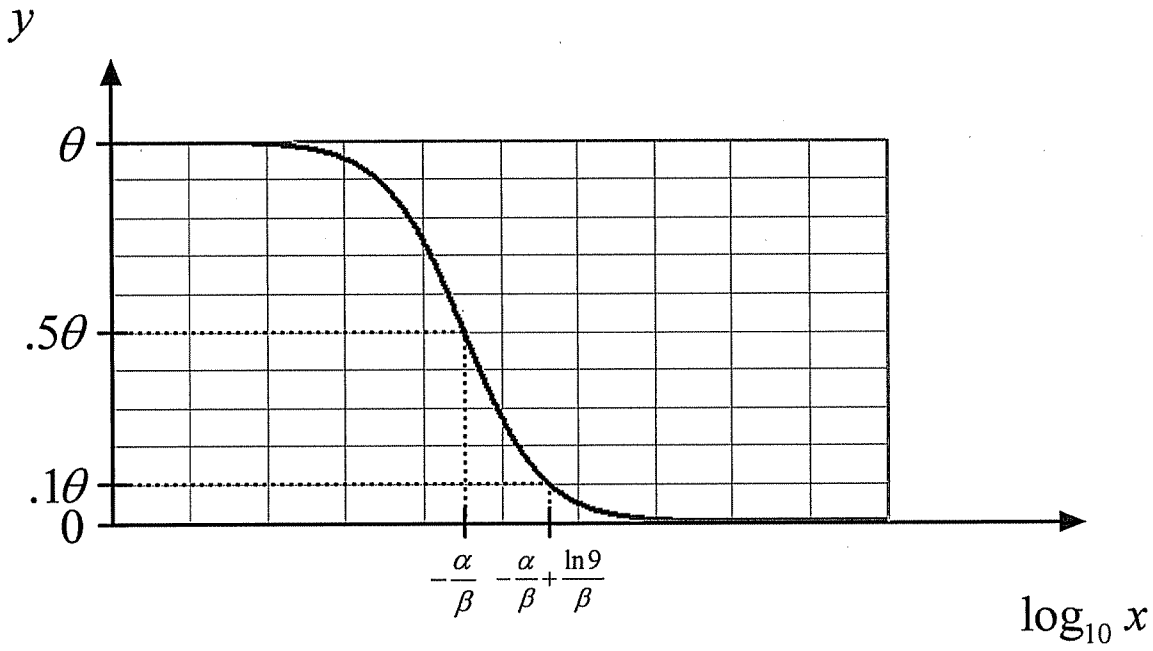
$$y \approx \frac{\theta}{1+\exp(\alpha+\beta\log_{10} x)} \tag{1}$$

Since absorbance decreases with x, in (1) the parameter β will typically be positive and its magnitude controls the rate of decline in absorbance with increasing x. θ plays the role of a maximum absorbance and when

$$\alpha+\beta\log_{10} x = 0 \text{ i.e. } \log_{10} x = -\frac{\alpha}{\beta}$$

(1) says that absorbance will be at half its maximum value. These features of the relationship (1) are illustrated in Figure 2.

Figure 2: Generic Ideal Absorbance Curve (1)



1. Suppose that ideally, successive dilutions of a specimen have numbers

$$2^0, 2, 2^2, 2^3, 2^4, \ldots$$

but small physical errors in processing work to replace the series of nominal log-numbers $\{i \cdot \log_{10} 2\}$ with the realized series of log-numbers

$$I, I+\eta_1+\log_{10} 2, I+\eta_1+\eta_2+2\cdot\log_{10} 2, I+\eta_1+\eta_2+\eta_3+3\cdot\log_{10} 2, \ldots$$

i.e. the series $\left\{ I+\sum_{j=0}^{i}\eta_j + i\cdot\log_{10} 2 \right\}$ (where $\eta_0 = 0$, the value I allows for variation in the preparation of the original material, and the η_j for $j>0$ represent small errors of not mixing exactly equal parts of the test material and clear diluent).

Suppose further that I and all the η_j's are independent with means 0, $\text{Var}\, I = \sigma_I^2$, and $\text{Var}\, \eta_j = \sigma_\eta^2$.

1a) Consider the realized dilution log-numbers corresponding to nominal dilution numbers 2^2 and 2^3 for a particular series of dilutions. What is the covariance matrix for this pair?

1b) Consider the realized dilution log-numbers corresponding to nominal dilution number 2^2 for two different series of dilutions that begin with the same physical specimen (and therefore share a common I). What is the correlation between these?

2. Consider the model

$$y = \frac{\theta}{1+\exp(\alpha+\beta\log_{10} x)} + \varepsilon \qquad (2)$$

(for realized dilution numbers, x) where we assume that ε's are mean 0, $\text{Var}\, \varepsilon = \sigma_\varepsilon^2$ random variables. In the context of question **1**, where the ith realized log-number is

$$I+\sum_{j=0}^{i}\eta_j + i\cdot\log_{10} 2$$

suppose that ε's, I, and η_j's are all independent. Find an approximate covariance matrix for absorbances corresponding to nominal dilution numbers 2^2 and 2^3 for a particular series of dilutions.

3. If measured absorbances, y, for the (8) nominal dilution numbers $x' = 2^9, 2^{10}, \ldots, 2^{16}$ are available from a single series of dilutions of a reference specimen, one might consider use of a standard Gaussian non-linear regression model

$$y_i = \frac{\theta}{1+\exp(\alpha+\beta\log_{10} x_i')} + \varepsilon_i \quad \text{for } i = 1,2,\ldots,8$$

(for iid $\text{N}(0,\sigma_\varepsilon^2)$ variables ε_i) in the analysis of such data. Discuss the appropriateness of inferences (tests and confidence intervals) for the parameters α, β, θ, and σ_ε^2 that

would be provided by standard "non-linear regression" software applied in this way. (For $\mathbf{Y} = (y_1, y_2, \ldots, y_8)'$ what does the modeling of question **2** suggest about $E\mathbf{Y}$ and $\mathrm{Cov}\,\mathbf{Y}$?)

4. The standard assumption for an assay of this type is that (unless something goes physically "wrong" in plate preparation or reading) regardless of what specimens (reference or unknown) are represented in the various rows on a plate, a relationship like (1) holds for measured absorbance and nominal dilution number for each row, where the parameters θ and β are common for all rows, and only α is potentially different row-to-row. This makes absorbance versus nominal log-dilution curves for specimens on different rows of a plate have the same shape, being only left-right shifts of one another. (However, changing lab conditions plate-to-plate allow values of θ and β to vary plate-to-plate.)

Three different specimens of material are used to make serial dilutions placed in rows 2 through 7 (top to bottom) of a plate with dilutions with nominal numbers $2^9, 2^{10}, \ldots, 2^{16}$ placed into columns 3 through 10 (left to right) of the plate. Rows 2 and 3 are used for Specimen #1, rows 4 and 5 are used for Specimen #2, and rows 6 and 7 are used for Specimen #3. (The first and last rows on each plate are avoided in an attempt to reduce measurement variation associated with potential "edge effects," where material in wells around the edges of a plate unfortunately behaves differently than material in the majority of the wells.) The two rows for a given specimen represent different dilution series begun from the same physical specimen.

There is some R output representing some least squares fitting of relationships like (1) to data derived from such a study. Treat this as a sensible descriptive analysis (i.e. don't worry much about formal inferences based on these data using this fitting) and use it as you answer the following.

4a) A study not detailed here (conducted by placing the same material in each well of several plates) indicates that (after accounting for edge effects) a standard deviation of around .009 machine units might be used to describe within-plate variation of measured absorbances of a fixed material. In light of this and the R output, do

- "sigmoidal shape", and
- "constant shape across specimens"

for absorbance versus dilution log-number curves seem sensible here? Explain.

4b) Under relationship (1) where α is allowed to vary specimen-to-specimen (but θ and β are not), a specimen with parameter α^* is judged to have

$$10^{-\left(\frac{\alpha^* - \alpha^{**}}{\beta}\right)}$$

times as much antibody as a specimen with parameter α^{**}. What then do the data from Specimens #1, #2, and #3 suggest about the relative amounts of antibody in those specimens?

PART 2

As an operational matter, the scientist developing this assay determines to do the following. Any plate run through the physical procedure will have material on it from one reference specimen represented in two rows of the plate, and material from at most two unknowns, each unknown represented in two rows of the plate. The reference and unknowns will be placed only in rows 2 through 7, and dilutions with nominal numbers $2^9, 2^{10}, \ldots, 2^{16}$ will be placed into columns 3 through 10. Least squares will be used to fit relationship (1) with a common (θ, β) to the (nominal dilution number, absorbance) data for the specimens on the plate, allowing a different α for each specimen. For each specimen (reference or unknown) the fitted value of the ratio

$$r = -\frac{\alpha}{\beta}$$

will be used to summarize its (pair of) dilution series absorbances. Then, for example, with reference and unknown values respectively r_{Ref} and r_{Unknown}, the unknown can be judged to have

$$10^{r_{\text{Unknown}} - r_{\text{Ref}}}$$

times as much antibody as the reference specimen.

Of serious concern to the scientist is the repeatability of the above-described assay. One does not expect values of r for a given specimen run on different plates to be consistent, but *differences* between a specimen's r and a reference r need to be relatively consistent across plates for the assay to have any practical value. So, one important activity is quantifying the consistency of results produced by this assay. Consider a study conducted for this purpose, organized as follows.

Plates can be prepared and run 2 per day. As many as 3 specimens will be represented on a plate. In an effort to understand the basic sources of variation in the production of the values r without the confounding issue of differences in the sources being measured, only reference specimens are used in a repeatability study. A pair of rows producing a given value of r come from (different) dilution series begun from the same original material (and in the event that a specimen is used one more than once on a plate or day, all dilution series for that specimen begin from the same material). So, a single plate could represent 3 different reference specimens. It could also represent a single reference specimen run 3 times (on successive pairs of rows).

Over a 5 day work week, 10 plates are prepared and tested, as indicated in Table 1 below.

Table 1: Design of the Repeatability Study

Day	Plate	Specimens
1	1	1,1,1
	2	1,1,1
2	3	2,3,4
	4	2,3,4
3	5	5,6,7
	6	5,8,9
4	7	10,11,12
	8	10,13,14
5	9	15,15,15
	10	15,15,15

Suppose that with

r_{ijkl} = ratio for the lth pair of series for specimen i run on day j on plate k

one models as

$$r_{ijk} = \mu + s_i + d_j + p_k + \varepsilon_{ijkl} \tag{3}$$

for μ an unknown constant, the s_i, d_j, p_k, and ε_{ijkl} independent mean 0 normal random variables, with Var $s_i = \sigma_s^2$, Var $d_j = \sigma_d^2$, Var $p_k = \sigma_p^2$, and Var $\varepsilon_{ijkl} = \sigma_\varepsilon^2$.

There is some R output summarizing an analysis of some hypothetical data for a study like that sketched above. Use it as needed in answering the questions below.

5. Write out the following in terms of the parameters of model (3):

5a) Corr(r_{1111}, r_{1112}) (the correlation between two ratios for two different dilution series pairs for the same specimen run on the same plate)

5b) Corr(r_{1111}, r_{1121}) (the correlation between two ratios for two different dilution series pairs for the same specimen run on the same day but on different plates)

5c) Corr(r_{2231}, r_{3231}) (the correlation between two ratios for two different specimens run on the same plate)

5d) Corr(r_{2231}, r_{3241}) (the correlation between two ratios for two different specimens run on different plates on the same day)

6. What seem to be the biggest contributors to variation in r? (Variation between dilution series pairs for a fixed specimen, variation between specimens, variation between plates on a given day, variation between days?) Explain.

7. Suppose that ratios for both the reference material and an unknown material follow the same model (3) except that the mean ratio (μ) is different for the two materials. At a future time, dilution series for a specimen of each will be run on a single plate producing

$$r_{\text{Ref}} = \mu_{\text{Ref}} + s_{16} + d_7 + p_{11} + \varepsilon_{\text{Ref}}$$

and

$$r_{\text{Unknown}} = \mu_{\text{Unknown}} + s_{17} + d_7 + p_{11} + \varepsilon_{\text{Unknown}}$$

Based on these, the difference

$$r_{\text{Unknown}} - r_{\text{Ref}} \tag{4}$$

will be reported as an estimate of $\mu_{\text{Unknown}} - \mu_{\text{Ref}}$ and

$$10^{r_{\text{Unknown}} - r_{\text{Ref}}} \tag{5}$$

will be reported as an estimate of the relative strength of the two materials, $10^{\mu_{\text{Unknown}} - \mu_{\text{Ref}}}$. What does the repeatability study (and the R output) suggest as standard errors for the estimates (4) and (5)? (Give a numerical value for the first and a function of $r_{\text{Unknown}} - r_{\text{Ref}}$ for the second.)

PART 1 R Output

```
> y1A
[1] 1.29 1.25 1.15 0.97 0.76 0.48 0.29 0.15
> y1B
[1] 1.31 1.25 1.14 1.00 0.75 0.49 0.28 0.17
> y2A
[1] 1.27 1.20 1.02 0.83 0.58 0.37 0.16 0.12
> y2B
[1] 1.25 1.18 1.06 0.84 0.59 0.37 0.20 0.10
> y3A
[1] 1.29 1.26 1.17 1.02 0.78 0.53 0.31 0.15
> y3B
[1] 1.31 1.26 1.15 1.02 0.79 0.53 0.33 0.17

> y1<-c(y1A,y1B)
> y2<-c(y2A,y2B)
> y3<-c(y3A,y3B)

> y<-c(y1,y2,y3)

> x
 [1] 2.70927 3.01030 3.31133 3.61236 3.91339 4.21442 4.51545 4.81648 2.70927
[10] 3.01030 3.31133 3.61236 3.91339 4.21442 4.51545 4.81648


> summary(nls(formula=y1~theta/(1+exp(alpha+beta*x)),start=c(theta=1.3,alpha=-
10.5,beta=2.5),trace=T))
0.1291273 :    1.3 -10.5   2.5
0.006488018 :   1.354697 -9.446136  2.375011
0.001612692 :    1.347793 -10.139099   2.534545
0.001587328 :    1.349757 -10.150707   2.538012
0.001587327 :    1.349742 -10.151144   2.538114

Formula: y1 ~ theta/(1 + exp(alpha + beta * x))

Parameters:
        Estimate Std. Error t value Pr(>|t|)
theta   1.349742   0.009003  149.92   <2e-16 ***
alpha -10.151144   0.197671  -51.35   <2e-16 ***
beta    2.538114   0.045786   55.43   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01105 on 13 degrees of freedom

Number of iterations to convergence: 4
Achieved convergence tolerance: 9.072e-06
```

```
> summary(nls(formula=y2~theta/(1+exp(alpha+beta*x)),start=c(theta=1.3,alpha=-
10.5,beta=2.5),trace=T))
0.5816205 :    1.3 -10.5   2.5
0.08180664 :   1.338513 -7.529886  2.020120
0.006602769 :   1.309612 -9.597931  2.506326
0.003433211 :   1.342925 -9.530394  2.500676
0.003431913 :   1.342973 -9.535173  2.501646

Formula: y2 ~ theta/(1 + exp(alpha + beta * x))

Parameters:
      Estimate Std. Error t value Pr(>|t|)
theta  1.34297    0.01678   80.05  < 2e-16 ***
alpha -9.53517    0.29426  -32.40 8.08e-14 ***
beta   2.50165    0.06970   35.89 2.17e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01625 on 13 degrees of freedom

Number of iterations to convergence: 4
Achieved convergence tolerance: 9.131e-06


> summary(nls(formula=y3~theta/(1+exp(alpha+beta*x)),start=c(theta=1.3,alpha=-
10.5,beta=2.5),trace=T))
0.07036615 :    1.3 -10.5   2.5
0.002373676 :    1.346124 -10.004627   2.476271
0.001283248 :    1.342302 -10.381621   2.562886
0.001281625 :    1.342856 -10.382187   2.563179
0.001281625 :    1.342857 -10.382179   2.563177

Formula: y3 ~ theta/(1 + exp(alpha + beta * x))


Parameters:
        Estimate Std. Error t value Pr(>|t|)
theta   1.342857   0.007593  176.85   <2e-16 ***
alpha -10.382179   0.179959  -57.69   <2e-16 ***
beta    2.563177   0.041449   61.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.009929 on 13 degrees of freedom

Number of iterations to convergence: 4
Achieved convergence tolerance: 5.971e-08
```

```
> ind1<-c(rep(1,16),rep(0,32))
> ind2<-c(rep(0,16),rep(1,16),rep(0,16))
> ind3<-c(rep(0,32),rep(1,16))
> xxx<-c(x,x,x)


> summary(nls(formula=y~theta/(1+exp(alpha1*ind1+alpha2*ind2+alpha3*ind3+beta*xxx)),
start=c(theta=1.35,alpha1=-10.15,alpha2=-9.5,alpha3=-10.4,beta=2.5),trace=T))
0.0903027 :    1.35 -10.15  -9.50 -10.40   2.50
0.006807985 :    1.347373 -10.038086  -9.552376 -10.151402   2.508605
0.006680612 :    1.346167 -10.126137  -9.636066 -10.240983   2.529904
0.006680592 :    1.346179 -10.126646  -9.636644 -10.241560   2.530048
0.006680592 :    1.346178 -10.126656  -9.636654 -10.241570   2.530050

Formula: y ~ theta/(1 + exp(alpha1 * ind1 + alpha2 * ind2 + alpha3 * ind3 +
    beta * xxx))

Parameters:
         Estimate Std. Error t value Pr(>|t|)
theta    1.346178   0.006124  219.82   <2e-16 ***
alpha1 -10.126656   0.130179  -77.79   <2e-16 ***
alpha2  -9.636654   0.124513  -77.39   <2e-16 ***
alpha3 -10.241570   0.131476  -77.90   <2e-16 ***
beta     2.530050   0.029903   84.61   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01246 on 43 degrees of freedom

Number of iterations to convergence: 4
Achieved convergence tolerance: 1.665e-07
```

PART 2 R Output

```
> rstudy
        r Specimen Day Plate
1  3.8381        1   1     1
2  3.8351        1   1     1
3  3.8371        1   1     1
4  3.8400        1   1     2
5  3.8407        1   1     2
6  3.8396        1   1     2
7  3.8403        2   2     1
8  3.8483        3   2     1
9  3.8405        4   2     1
10 3.8512        2   2     2
11 3.8567        3   2     2
12 3.8432        4   2     2
13 3.8502        5   3     1
14 3.8460        6   3     1
15 3.8441        7   3     1
16 3.8415        5   3     2
17 3.8517        8   3     2
18 3.8454        9   3     2
19 3.8413       10   4     1
20 3.8338       11   4     1
21 3.8477       12   4     1
22 3.8492       10   4     2
23 3.8336       13   4     2
24 3.8407       14   4     2
25 3.8576       15   5     1
26 3.8436       15   5     1
27 3.8513       15   5     1
28 3.8493       15   5     2
29 3.8576       15   5     2
30 3.8447       15   5     2

> repeatability<-lmer(r ~ 1 + (1|Specimen)+ (1|Day) + (1|Day:Plate))

> summary(repeatability)
Linear mixed model fit by REML
Formula: r ~ 1 + (1 | Specimen) + (1 | Day) + (1 | Day:Plate)
    AIC    BIC logLik deviance REMLdev
 -202.1 -195.1  106.1   -222.7  -212.1
Random effects:
 Groups     Name        Variance   Std.Dev.
 Specimen   (Intercept) 1.0140e-05 3.1843e-03
 Day:Plate  (Intercept) 3.1575e-15 5.6192e-08
 Day        (Intercept) 1.0697e-05 3.2706e-03
 Residual               2.3815e-05 4.8801e-03
Number of obs: 30, groups: Specimen, 15; Day:Plate, 10; Day, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)  3.84459    0.00198    1941
```

Methods 2 Key 2009 Statistics Prelim

Note Title 10/20/2008

1. a) These variables are $I + \eta_1 + \eta_2 + 2\log_{10} 2$ and $I + \eta_1 + \eta_2 + \eta_3 + 3\log_{10} 2$. The first has variance $\sigma_I^2 + 2\sigma_\eta^2$ and the 2nd has variance $\sigma_I^2 + 3\sigma_\eta^2$. They have covariance

$$E\left(I + \eta_1 + \eta_2\right)\left(I + \eta_1 + \eta_2 + \eta_3\right) = \sigma_I^2 + 2\sigma_\eta^2$$

so the covariance matrix is

$$\begin{pmatrix} \sigma_I^2 + 2\sigma_\eta^2 & \sigma_I^2 + 2\sigma_\eta^2 \\ \sigma_I^2 + 2\sigma_\eta^2 & \sigma_I^2 + 3\sigma_\eta^2 \end{pmatrix}$$

b) These are variables $I + \eta_{11} + \eta_{21} + 2\log_{10} 2$ and $I + \eta_{12} + \eta_{22} + 2\log_{10} 2$. They both have variance $\sigma_I^2 + 2\sigma_\eta^2$ and they have covariance $E(I + \eta_{11} + \eta_{21})(I + \eta_{12} + \eta_{22}) = \sigma_I^2$ so the correlation is

$$\frac{\sigma_I^2}{\sigma_I^2 + 2\sigma_\eta^2}$$

2. Write

$$m(z) = \theta / (1 + \exp(\alpha + \beta z))$$

and $m'(z) = \theta\beta \exp(\alpha + \beta z) / (1 + \exp(\alpha + \beta z))^2$ and approximate

$$y_1 \approx m(2\log_{10} 2) + m'(2\log_{10} 2)(I + \eta_1 + \eta_2) + \epsilon_1$$

and

$$y_2 \approx m(3\log_{10}2) + m'(3\log_{10}2)(I + \eta_1 + \eta_2 + \eta_3) + \epsilon_2$$

So (approximately) the covariance matrix for $\begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$ is

$$\begin{pmatrix} (m'(2\log_{10}2))^2(\sigma_I^2 + 2\sigma_\eta^2) & m'(2\log_{10}2)\,m'(3\log_{10}2)(\sigma_I^2 + 2\sigma_\eta^2) \\ \text{same} & (m'(3\log_{10}2))^2(\sigma_I^2 + 3\sigma_\eta^2) \end{pmatrix} + \sigma_\varepsilon^2 \frac{I}{2\times 2}$$

3. The iid errors assumption of the usual non-linear regression model is from 2 clearly not right... since $m'(z) < 0$ we see that for a given dilution series (no surprise) absorbances are positively correlated. So the correlation structure is not right.

The mean structure is only approximately right at best. That is, the nonlinear regression model says

$$E y_i = m(\log_{10} x_i') = m(E \log_{10} x_i)$$

$$= m(E(I + \eta_1 + \cdots + \eta_i + i \log_{10} 2))$$

while in fact

$$E y_i = E\, m(I + \eta_1 + \eta_2 + \cdots + \eta_i + i \log_{10} 2)$$

for the nonlinear function $m(z)$. So if the σ_I^2 or σ_η^2 are very big, the mean structure is also potentially quite inappropriate. So, I don't trust inferences based on the usual nonlinear regression model here.

4. a) As a purely descriptive matter, least squares fitting seems quite sensible here. Fitted θ's + β's for y_1, y_2, and y_3 separately are in substantial agreement with those from a simultaneous fit. SSE values are

y_1 fit alone	.00159
y_2 fit alone	.00343
y_3 fit alone	.00128

} total to about .0063

y_1, y_2, y_3 fit simultaneously .0067 — there is not much difference

Further, from the overall fit $\hat{\sigma} = .0125$ is not much larger than the standard deviation that accounts only for measurement error for opacities (and doesn't take into account any lack of fit of the "common sigmoid shape") model.

b) Take, for example, specimen 1 as a reference, it appears that specimen 2 has about

$$10^{-\frac{(-9.6366 + 10.1266)}{2.53005}} = .64 \text{ i.e. } 64\%$$

as much antibody as specimen 1 — Further, it appears that specimen 3 has about

$$10^{-\frac{(-10.2716 + 10.1266)}{2.53005}} = 1.11 \text{ i.e. } 111\%$$

as much antibody as specimen 1

5a) $$\frac{\sigma_s^2 + \sigma_d^2 + \sigma_p^2}{\sigma_s^2 + \sigma_d^2 + \sigma_p^2 + \sigma_\varepsilon^2}$$

b) $$\frac{\sigma_s^2 + \sigma_d^2}{\sigma_s^2 + \sigma_d^2 + \sigma_p^2 + \sigma_\varepsilon^2}$$

c) $$\frac{\sigma_d^2 + \sigma_p^2}{\sigma_s^2 + \sigma_d^2 + \sigma_p^2 + \sigma_\varepsilon^2}$$

d) $$\frac{\sigma_d^2}{\sigma_s^2 + \sigma_d^2 + \sigma_p^2 + \sigma_\varepsilon^2}$$

6. The estimated "residual," specimen and day variance components are roughly comparable (The corresponding sources appear to be roughly equally important). The "plate" component of variance does not seem to be large.

7. $r_{unknown} - r_{Ref} = s_{17} - s_{16} + \varepsilon_{unknown} - \varepsilon_{Ref}$ so that

$\text{Var}(\quad) = 2\sigma_s^2 + 2\sigma_\varepsilon^2$ and a standard error for the difference in r's is thus

$$\sqrt{2\hat{\sigma}_s^2 + 2\hat{\sigma}_\varepsilon^2} = \sqrt{2(1.014\times10^{-5}) + 2(2.3816\times10^{-5})}$$

$$= .00824$$

Then a standard error for $10^{r_{unknown} - r_{ref}}$ is

$$\underbrace{\ln 10 \cdot 10^{r_{unknown} - r_{ref}}}_{\text{The derivative of } y = 10^x \text{ at } x = r_{unknown} - r_{ref}} \cdot (.00824)$$

1 Problem Background

Growing concern with the pollution caused by runoff of nitrogen fertilizer from agricultural fields (primarily corn fields), combined with increasing nitrogen fertilizer costs have prompted numerous efforts to reduce the amount of fertilizer applied to corn crops. Many of the studies involved in these efforts have been traditional agricultural field studies with relatively small plots of land subjected to various treatments in some randomized manner. It has been difficult, however, to extend the results of such studies to larger scales such as multiple fields in one area or, especially, multiple farms across the midwest or even the state of Iowa. This is because the amount of natural nitrogen (without fertilizer) available to corn crops varies both temporally and spatially, and remains extremely difficult to forecast for any particular area. Thus, the effect of reducing nitrogen fertilizer is also variable in both time (e.g., years) and space (e.g., area of Iowa). More recently, a number of large-scale studies have been undertaken that attempt to determine the effect, if any, of reduced fertilization on overall crop yields in corn-producing regions as a whole. These studies also attempt to identify conditions that might be related to situations under which reduced fertilization does or does not lead to lower yields at finer levels of resolution, such as individual farms. This question concerns one such study, conducted by the Iowa Soy Bean Association (yes, they worry about corn as well as soy beans).

The study in question involved a cooperative effort among about 40 farmers and the Iowa Soybean Association. Each farmer cooperating in the study selected one field on which he or she operates and made use of two rates of nitrogen fertilizer application. One rate was the usual rate that would be applied, and the other rate was a reduction from the usual rate by 50 pounds per acre. These treatments were applied in long strips (1500 – 2000 feet long) in a portion of the field, as illustrated in Figure 1. The response variable of primary interest was yield, one value representing a sampling unit of 60 by 60 feet. In about one half of the fields, the strips to which fertilizer was applied were one sampling unit wide, illustrated in the left panel of Figure 2. In the other fields the strips were two sampling units wide, as illustrated in the right panel of Figure 2. Note that neither panel of Figure 2 shows the entire treated area of Figure 1, only enough to understand the arrangement of treatments and sampling units within strips. Fields typically contained 3 pairs of strips (e.g., 6 strips with two treatments in each pair). The same number of strips

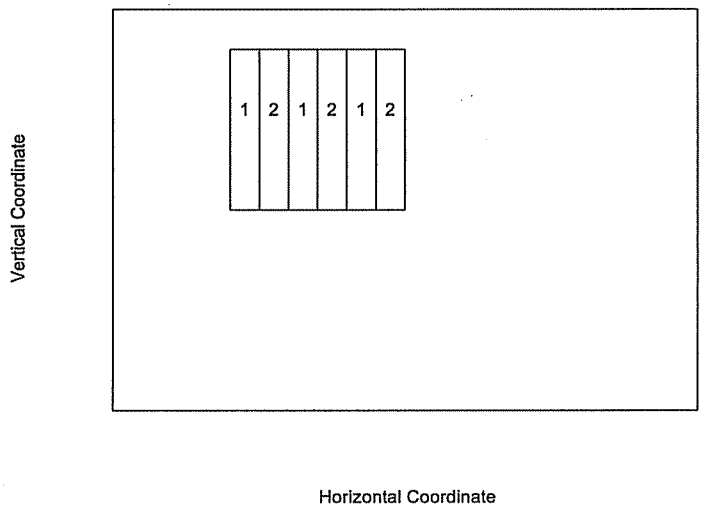


Figure 1: Overall layout of nitrogen fertilization studies. The figure depicts one field with strips of the two treatments, 1 =regular or normal rate of fertilizer and 2 =reduced rate of fertilizer.

in each field received each of the two treatments (again, the panels of Figure 2 show only a portion of treated area).

Along with the measured response (yield), a number of covariates were recorded. A list of these follows.

1. Time of fertilizer application – fall, spring, summer.
2. The type of fertilizer – ammonium nitrate, urea, manure.
3. The previous crop grown on the field – corn, beans.
4. Tillage time – fall, spring, fall and spring.
5. Rainfall.
6. Temperature.
7. Parent soil type – there are four basic parent soil types in Iowa.
8. Soil unit – a subdivision of parent soil type to include slope and aspect. Generally, there may be 1 to 6 soil units within an individual field.

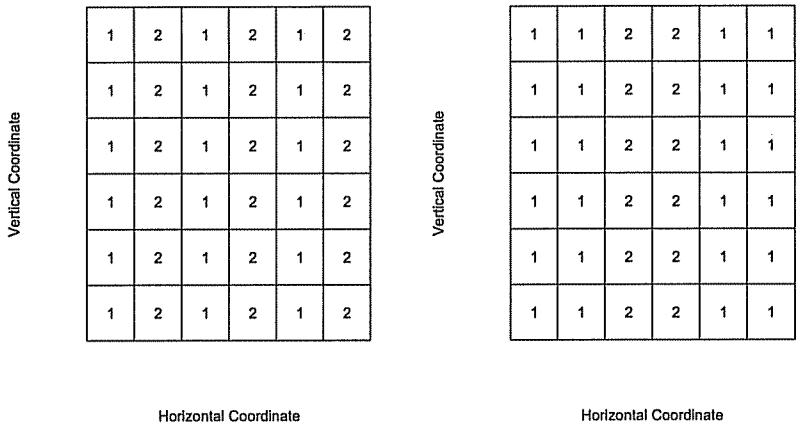


Figure 2: Detail of sampling units used in fertilization studies. Shown is a portion of several strips from Figure 1. The plot in the left panel displays a portion of the sampling units and treatment arrangements for the basic study configuration. The plot in the right panel displays a portion of the sampling units and treatment arrangements for the "double wide" strips. Each square represents one 60 by 60 foot sampling unit.

Most of these (all but perhaps soil unit) are recorded at the level of a field, so that the same value of the covariates will apply to each sampling unit within a given field.

The Iowa Soybean Association has asked the Department of Statistics at Iowa State University to prepare a proposal outlining possible strategies for the analysis of data from these studies. The primary objective of such analysis is to determine under what conditions, if any, observed yields from the reduced fertilizer rates are not decreased from the corresponding normal fertilizer yields.

2 Complications and Additional Information

1. What are being called the normal and reduced fertilizer rates, referred to previously as treatments, do not represent the same absolute amount of fertilizer on different fields. That is, the normal rate differs among fields, but is generally between 120 and 180 pounds per acre. The study protocol could not dictate the actual value of the normal rate to cooperating farmers. Thus, "normal rate" means whatever the farmer would have done without participating in the study. Concomitantly, what

absolute amount of fertilizer corresponds to the reduced rate also varies, because "reduced" means 50 pounds per acre less than normal.

2. Levels of type of fertilizer, time of application, and tillage are not crossed factors in the study. That is, not all fertilizer types are applied at all times or under all tillage strategies.

3. Empirical distributions of yield from sampling units within a field (at a given level of fertilizer) do not appear to be symmetric. Differences between adjacent sampling units, however, do appear to be unimodal and nearly symmetric in shape.

4. All of the potential covariates associated with management (e.g., fertilizer type, tillage, previous crop, and normal fertilizer rate) are under the control of the cooperating farmer, not investigators at the Iowa Soybean Association.

5. While there were typically 3 to 4 sets of paired strips in the study design, the placement of those strips in a field and the fertilizer applied to them (normal or reduced rate) were also up to the farmer and cannot be considered as resulting from a strategy of randomization.

6. Aside from previous crop (corn, beans) the possible effects, if any, of the potential covariates are largely unknown. It seems to be quite generally accepted, however, that the previous crop should be important.

3 Questions

The setting of the questions presented in this section is one in which you have been asked to prepare a short proposal to the the Iowa Soybean Association to receive funding for conducting analyses of the data resulting from the study just described. While there is not time in this exam for you to actually do this, the questions that follow relate to various issues that would need to be addressed in such a proposal. Thus, your task is to discuss the issues identified in the questions intelligently. **The large number of specific questions contained here are meant to stimulate your thinking and should not necessarily be taken as points you must address separately in enumerated form. The key "questions" in what follows are identified *in emphasized text*, and must be addressed in some form.**

1. Some university agronomists have been critical of the value of studies such as that described in the previous sections. They point out that the study design is not really similar to a traditional randomized complete block design as often employed in more controlled studies. They also question whether any cause-and-effect relations could be uncovered by this type of broad-scale study design. *Comment on both the value and the limitations of the study described in this question.* In preparing your answer, consider the following issues.

   (a) Could this study be analyzed using the experimental approach to analysis? What would (or could) be defined as experimental units? Would it be possible to define a randomization test procedure to address the primary objective of the study?

   (b) Is the purpose of the study to determine cause-and-effect relations? If so, what might they be? Could the results of an analysis from this study lead to inferences about such cause-and-effect relations?

   (c) What is the purpose of using blocks in a randomized block study design? Do the multiple sets of strips in this study fulfill the same purpose?

   (d) What is the purpose of using a large number (relative to the number in more carefully controlled studies) of fields across the state of Iowa in this study?

   (e) Why have more carefully controlled traditional study designs failed (to date) to provide satisfactory answers to the primary objective of this study?

2. In developing a model-based approach to the analysis of this study, consider first an individual field that has sampling design as in the left panel of Figure 2. Make use of the following notation in your answer. First, define locations $\{s_i : i = 1, \ldots, n\}$ for the n sampling units in a field, where $s_i \equiv (u_i, v_i)$ with u_i being an integer-valued horizontal coordinate (across strips) and v_i an integer-valued vertical coordinate (along strips) for sampling unit i. That is, consider a regular lattice (grid) to be defined and laid over the treated portion of a given field (as illustrated in Figures 1 and 2). *Define random variables appropriate for the analysis of these data, both fundamental-level variables and, if desired, constructed variables that might be used in a model.* Define these variables for only a single field but keeping in mind the fact that we wish to do so for each of many fields. Refer to particular pieces of information given in the Problem Background (Section 1) or Complications and

Additional Information (Section 2) that motivate your answer. Here, you might consider the following.

(a) What is being directly observed in each field?

(b) Do the quantities being directly observed have a consistent meaning across different fields?

3. Continue to focus on a single field. Data are available from a small pilot study laid out exactly as in Figure 1 with 3 pairs of strips, and having 20 sampling units per strip. Measurements of yield for the three strips receiving normal fertilizer rate resulted in the following stem-and-leaf plots. Yields from the three strips receiving reduced fertilizer were visually quite similar.

```
The decimal point is 2 digit(s) to the right of the |

0 | 78899          0 | 99           0 | 8899
1 | 01234          1 | 01234        1 | 01234
1 | 578            1 | 5679         1 | 5689
2 | 023            2 | 023          2 | 124
2 | 578            2 | 579          2 | 68
3 | 1              3 | 03           3 | 02
```

Putting all three strips receiving normal fertilizer together provides enough values to construct a histogram, which is presented in Figure 3. To illustrate the point made in item 3 of the additional information given in Section 2 of the problem description, the histogram for differences between sampling units with normal and reduced fertilizer that are adjacent in the field is presented in Figure 4.

Note here that differences are computed as yield for normal minus yield for reduced fertilizer rates in bushels per acre. The mean difference among adjacent sampling units with different fertilizer rates was 1.936, the variance was 9.7944 and recall that there are 60 differences in the field. *Suppose that nitrogen fertilizer costs* 0.26 *dollars per pound, and corn prices are* 1.50 *dollars per bushel. Given that the yield values were recorded in bushels per acre, assume the simplest model possible for yield differences and compute a* 95% *interval estimate for the difference in profit (in dollars per acre) between reduced and normal fertilizer rates in this pilot study.*

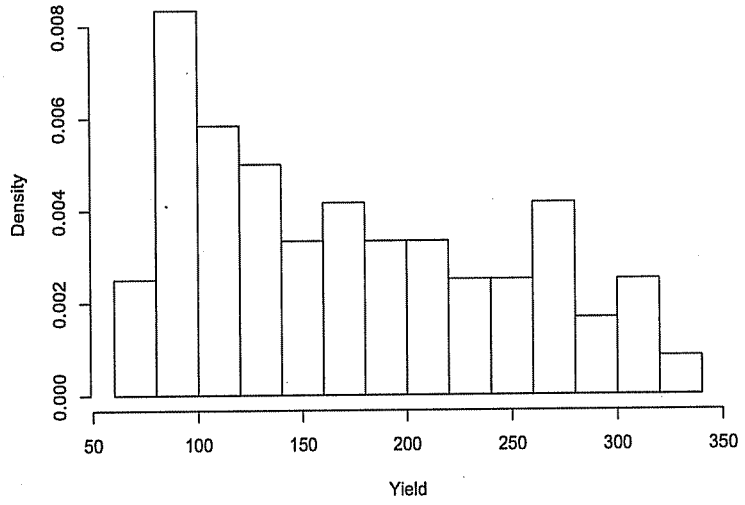


Figure 3: Histogram of all yield values from normal fertilizer in the pilot study.

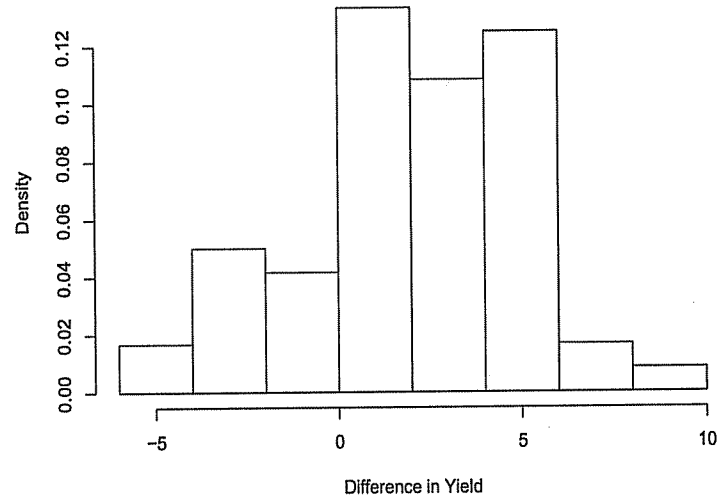


Figure 4: Histogram of all yield values from normal fertilizer in the pilot study.

4. Now consider the set of 40 fields for which data will be available, but do not consider covariate information at this point. *Suggest a potential model structure that might be appropriate for the problem.* At this point, and in all except the last question, you may make simplifying assumptions of conditional independence, regardless of whether you feel this is entirely appropriate or not. You may also wish to add an index to your random variables from the previous question to account for field identity. The following issues might be valuable to consider.

   (a) Would we expect different fields to all show the same relation between normal and reduced treatments, or would be expect each field to be a potentially unique realization of this relation?

   (b) Would we **necessarily** expect there to be some overall relation between normal and reduced treatments considered across the entire set of fields from which data are available?

5. *Comment on whether it would be necessary to separate fields having the two sampling designs of Figure 2, or whether they could both be described by one model structure.* Consider, specifically,

   (a) Could the same model you have proposed in question 3 be applied directly to both fields having the sampling design of the left panel and ones having the sampling design of the right panel of Figure 2?

   (b) Could the locations you were given or the random variables you defined in question 1 be modified to allow both sampling designs to be described by one model?

   (c) Could data from the sampling design of the right panel of Figure 2 be used to provide useful information for assessing the significance of the differences observed in data from the sampling design of the left side of Figure 2?

6. Because we do not yet have any data, typical tools such as scatterplots and other diagnostics cannot yet be used to examine possible effects of covariates. Thus, we must be rather generic in suggesting how covariate information might be used in an analysis. *Offer a plan or strategy for incorporating covariate information in your model structure.* Consider the following.

(a) Are there enough fields (40) to approach this as a typical variable selection problem?

(b) Are there exploratory approaches available that might be used in conjunction with the results of fitting your model from question 4 to identify the potential usefulness of specific covariates?

(c) How might you deal with the "unbalanced" nature of the possible covariates of fertilizer type, time of application, and tillage? (See Section 2 on complications and additional information.)

7. *Suggest an approach you would attempt for estimation and inference using your model. Identify problems or issues you anticipate will arise.* If you take a likelihood approach, consider the following.

(a) What will be the likelihood needed for the analysis?

(b) Does it seem that it will be straightforward to simultaneously maximize this likelihood in all relevant parameters? If not, what options might be available?

(c) What methods would you anticipate will be useful to produce inferential quantities (e.g., standard errors and/or tests)?

(d) What methods do you anticipate will be useful for model assessment?

If you take a Bayesian approach, consider the following.

(a) How might the posterior be computed?

(b) What distributions (posteriors and/or posterior predictives) would seem the most likely to be useful in inference?

(c) How might you approach model assessment if a Bayesian approach is taken?

8. Now consider the issue of item (c) in question 1 more explicitly, and the fact that you were instructed in question 3 to make use of independence assumptions to develop your basic model structure. *Suggest potential elaborations of your model structure that might deal with the underlying issue. Be as explicit as possible in terms of the model you have formulated.*

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

   The limitations of the study design are largely related to the fact that the study cannot be considered under the framework of what was called *scientific abstraction* in Stat 601. That is, the study is not an experiment. Some of the following points should be raised.

   (a) Randomization is lacking in selection of the fields, portions of the fields to be treated, and how treatments were assigned to those portions of the fields. This alone is sufficient to preculde the use of the experimental approach to analysis.

   (b) What are being called treatments do not have a constant definition in terms of the amount of fertilizer applied. Any difference among such treatments cannot be considered to apply in a relatively uniform manner for all "units" in some population (if these could even be defined – see the next point).

   (c) It would be difficult to define experimental units in any non-arbitrary manner, in part because of the previous point and in part because of the application method used. The only possibilities would seem to be fields or strips of the fields to which fertilizer was applied.

   (d) Clearly, no cause-and-effect relations could be demonstrated as a result of the study.

   The value of the study is actually related to its shortcomings in terms of experimentation. Because of the control necessary to conduct a typical randomized complete block design study with well-defined treatments, a study of that type cannot be conducted on a large-scale. Because of the complexity of the problem under investigation, studies conducted on a small scale cannot result in conclusions that are widely applicable. The following points are relevant.

(a) The study is not designed to allow inferences about cause-and-effect relations. Rather, the study is designed to determine if, in situations that farmers will actually face, there are factors that can be identified as potential indicators of when reduced fertilization is likely to result in only minimal loss of yield. If such factors are identified by the study, additional studies will be required to determine if there is any consistency in their value.

(b) Statistical analysis will necessarily involve modeling, and it is useful to consider how, in this context, one might deal with heterogeneity in uderlying soil conditions within fields. The primary purpose of blocks in an experimental design is to guard against effects caused by inherent differences in portions of a field that correspond to blocks. In this study, there is pairing of application strips, which is helpful, but we may still wish to examine the possiblity of spatial structure in the observed data.

2. Question 2

Random variables for the directly observed quantities may be defined for individual sampling units as shown in the left panel of Figure 1 in the question. In an individual field, let the sampling units be identified by the locations $\{\boldsymbol{s}_i : i = 1, \ldots, n\}$. Define random variables $Y(\boldsymbol{s}_i)$ to be connected with yield from units receiving normal fertilizer, and $X(\boldsymbol{s}_i)$ to be connected with yield from units receiving reduced fertilizer. The normal and reduced fertilizer rates do not have constant meaning across fields, but the objective is to determine conditions under which reducing a usual level of fertilizer by 50 pounds per acre will or will not reduce yield. Thus, for a given field, define response variables for differences among sampling units with normal fertilizer and adjacent units with reduced fertilizer as follows. Assume that, as shown on the left of Figure 1, there are an even number of sampling units in each row of the treated field, that the left-most strip was given normal fertilizer rate, and that $\boldsymbol{s}_1 = (1, 1)$, $\boldsymbol{s}_2 = (1, 2)$ etc. Then our existing random variables are $\{Y(\boldsymbol{s}_i) : i = 1, 3, 5, \ldots, n-1\}$ and $\{X(\boldsymbol{s}_i) : i = 2, 4, 6, \ldots, n\}$.

Define locations $\boldsymbol{w}_j$ by combining $\boldsymbol{s}_i$ and $\boldsymbol{s}_{i+1}$ so that $\boldsymbol{w}_j = (u_i, v_i + 0.5)$. Then

$\boldsymbol{w}_1 = (1, 1.5)$, $\boldsymbol{w}_2 = (1, 3.5)$ etc., and here the index j goes from 1 to $k = n/2$. Finally, define random variables

$$Z(\boldsymbol{w}_j) = Y(\boldsymbol{s}_i) - X(\boldsymbol{s}_{i+1}). \tag{1}$$

Note that, if the original sampling units correspond to a 60×60 foot square, and we take 60 foot as a unit distance, distances among both the $\boldsymbol{s}_i$ and the $\boldsymbol{w}_j$ can be computed using the Euclidean norm.

3. Question 3.

A simply model is to assume that the random variables $Z(\boldsymbol{w}_j)$; $j = 1, \ldots, n/2$ from question 2 (i.e., differences in yield) are independent and identically distributed according to a normal distribution with mean μ and variance σ^2. The reduced fertilizer costs $0.26(50) = 13.0$ dollars less per acre than the normal fertilizer rate. The difference in profit between reduced rate and normal rate fertilizer is then,

$$D(\boldsymbol{w}_j) = 13.0 - 1.5\, Z(\boldsymbol{w}_j).$$

The $D(\boldsymbol{w}_j)$ then are independent and identically distributed according to a normal distribution with mean $\mu_p = 13.0 - 1.5\mu$ and variance $\sigma_p^2 = 1.5^2\sigma^2 = 2.25\sigma^2$. Estimates of these parameters from the 60 observations are

$$\begin{aligned} \hat{\mu}_p &= 13.0 - 1.5\,\hat{\mu} = 13.0 - 1.5(1.936) = 10.096 \\ \hat{\sigma}_P^2 &= 2.25\,\hat{\sigma}^2 = 2.25(9.7944) = 22.0374 \end{aligned}$$

A resultant 95% interval estimate of μ_p is then

$$\hat{\mu}_p \pm 1.96\{\hat{\sigma}_p^2/60\}^{1/2} = (8.909,\ 11.284).$$

4. Question 4.

Given the indication from the section on Complications and Additional Information that differences in yield among adjacent sampling units appear to have empirical distributions that are unimodal and symmetric, it would be natural to assume that, within a field, the random variables $Z(\boldsymbol{w}_j)$ have normal distributions. To allow

for multiple fields at this point, extend the indexing of these random variables to be $\{Z(f, \boldsymbol{w}_j) : j = 1, \ldots, k_f;\ f = 1, \ldots, F\}$ where f indexes field and j indexes location within field. Assume that, given values $-\infty < \mu_f < \infty$ and $\sigma_f^2 > 0$, the $Z(f, \boldsymbol{w}_j)$ are conditionally independent such that, for $\boldsymbol{w}_j = 1, \ldots, k_f$,

$$Z(f, \boldsymbol{w}_j) \sim N(\mu_f, \sigma_f^2). \tag{2}$$

Take μ_f and σ_f^2 to be random variables, independent and identically distributed for $f = 1, \ldots, F$, such that

$$\begin{aligned} \mu_f &\sim N(\mu, \tau^2) \\ \sigma_f^2 &\sim IG(\alpha, \beta), \end{aligned} \tag{3}$$

where $IG(\alpha, \beta)$ denotes an inverse gamma distribution with parameters $\alpha > 0$ and $\beta > 0$.

5. Question 5.

While the model as formulated in the previous two questions would be difficult to apply directly to a combined set of data from both sampling designs, it might be possible to combine these designs into one model with modified meanings for locations. Eventually this would be desirable to increase overall sample size in terms of the number of fields. Initially, however, it would most likely be advisable to keep these types of layouts separate because we do not understand any spatial effects that might be present in the fields.

The model as formulated could be applied to each sampling design separately by re-defining the random variables $X(\boldsymbol{s}_i)$ and $Y(\boldsymbol{s}_i)$ to be the same type of fertilizer application (either both normal rate or both reduced rate) for fields with the double-wide strip design (i.e., the right panel of Figure 2). This would provide valuable information because in this case we should have $\mu = 0$ in expression (3) above *a priori*. If the estimate of μ (either Bayesian or non-Bayesian) differs from 0, then this would indicate that the model applied to fields with single-wide strips will fail to necessarily provide accurate information about whether or not fertilization rates

differ on average for the entire state. It would then be doubtful that any useful information can be gained from the analysis. If, on the other hand, the estimated value of μ from applying the model when both $Y(\boldsymbol{s}_i)$ and $X(\boldsymbol{s}_i)$ correspond to equivalent fertilization rates does not differ (in a meaningful or significant way) from 0, then this would provide verification that the estimated μ when Y and X correspond to different fertilization rates does have substantive meaning.

6. Question 6.

   There will not be a sufficient number of fields to incorporate all of the potential covariates into a single model through the μ_f. Thus, approaching incorporation of covariates as a typical variable selection problem will prove ineffective. The fact that the levels of fertilizer types, time of appliation, and time of tilling are not balanced suggests that a combined covariate consisting of unique combinations of these factors be created. An initial approach to the examination of covariate information could be based on exploratory techniques. For example, either estimates of $\mu_f - \mu$ (e.g., posterior expectations from a Bayesian analysis) or surrogate values for $\mu_f - \mu$ (e.g., differences between average responses in a field and the maximum likelihood estimate of μ from a non-Bayesian analysis) could be plotted against covariates to determine if any patterns are discernable. In fact, this type of exploration may be all that is possible in terms of realizing the overall objective of the study. In the fortunate circumstance that patterns could be identified we might attempt to incorporate interval-scale (or continuous) covariates x such as weather variables through $\mu_f = g(x, \gamma)$ for some empirically determined function $g(\cdot)$ and parameters γ. Covariates that define factors could be considered as defining groups to which models would be fit separately.

7. Question 7.

   The details of answers to this question will depend on the model proposed in question 4. For the model suggested in this solution, the following would apply.

   (a) For a likelihood analysis we would need to obtain the joint marginal density

of the response variables for all fields, $\boldsymbol{Z} \equiv \{Z(f, \boldsymbol{w}_j) : f = 1, \ldots, F;\ j = 1, \ldots, k_f\}$. Independence among fields would allow this to be written as a product,

$$h(\boldsymbol{z}|\mu, \tau^2, \alpha, \beta) = \prod_{f=1}^{F} h(\boldsymbol{z}_f|\mu, \tau^2, \alpha, \beta),$$

where

$$h(\boldsymbol{z}_f|\mu, \tau^2, \alpha, \beta) = \int\int f(\boldsymbol{z}_f|\mu_f, \sigma_f^2)\, g(\mu_f, \sigma_f^2|\mu, \tau^2, \alpha, \beta)\, d\mu_f\, d\sigma_f^2,$$

and, by conditional independence,

$$f(\boldsymbol{z}_f|\mu_f, \sigma_f^2) = \prod_{j=1}^{k_f} f(z(f, \boldsymbol{w}_j)|\mu_f, \sigma_f^2).$$

The obvious statistical hurdle to be overcome is evaluation of the necessary integrals. Since these are only two-dimensional integrals one might attempt numerical integration, passing derivatives under the integrals for evaluation of first and second derivatives that might be needed in an iterative algorithm. This would result in a simultaneous maximization of the marginal log likelihood and calcualtion of observed information. Inference could then proceed using Wald theory. If this proves difficult, particularly in the case that continuous covariates have been included in the model, an alternative would be to employ unscaled profile likelihoods for a partition of the parameter vector. If successful, inference could then proceed through the calculation of normed profile likelihood intervals or, barring that, parametric bootstrap intervals.

(b) For a Bayesian analysis we would need to additionally specify prior distributions for μ, τ^2, α and β. If covariates are to be used, priors for μ and τ^2 would be replaced by a set of priors for the regression parameters. The difficulty in this approach would be computation of the joint posterior, which would most likely be attempted through the use of a Markov Chain sampler. An overall Gibbs algorithm would be a likely choice, as the full conditional posteriors can be determined in a straightforward manner. For example, in the model

of expresssions (2) and (3) of this solution, the full conditional posterior of μ, denoted as $p(\mu|\cdot)$, would be

$$p(\mu|\cdot) \propto \pi(\mu) \prod_{f=1}^{F} \left[g(\mu_f|\mu,\tau^2) \prod_{j=1}^{k_f} f(z(f,\boldsymbol{w}_j)|\mu_f,\sigma_f^2) \right],$$

where $\pi(\mu)$ is the prior distribution assigned to μ. Many of the full conditional posteriors may be known only up to a constant, so simulation from these distributions would likely involve sampling by rejection, adaptive rejection, ratio of uniforms, or adaptive ratio of uniforms.

Inference for a region as a whole would be based on the marginal posteriors $p(\mu|\boldsymbol{z})$ and $p(\tau^2|\boldsymbol{z})$. Because even a given field is not expected to show exactly the same relation from year to year, inferences about what might happen in a given field would most likely be based on the predictive posterior for new values of μ_f and σ_f^2, μ_f^* and σ_f^{2*}. For example,

$$p(\mu_f^*|\boldsymbol{z}) = \int\int g(\mu_f|\mu,\tau^2)\, p(\mu,\tau^2|\boldsymbol{z})\, d\mu\, d\tau^2.$$

Values from this predictive posterior would be generated by simulation of additional values from the model distribution $g(\mu_f|\mu,\tau^2)$ for each sampled pair (μ,τ^2) in the overall algorithm. Hopefully, such predictive distributions could be defined for several groups of fields based on factors identified as described in question 6.

8. Question 8.

To deal with the issue of heterogeneity within individual fields, we might formulate the data model for the $\{Z(f,\boldsymbol{w}_j) : j = 1,\ldots,k_f\}$ to include statistical dependence following a spatial structure. This could be accomplished by defining a neighborhood N_j for each location $\boldsymbol{w}_j$, for example a four-nearest neighbor structure. Let

$$\begin{aligned} N_j &\equiv \{\boldsymbol{w}_h : \boldsymbol{w}_h \text{ is a neighbor of } \boldsymbol{w}_j\} \\ \boldsymbol{z}(N_j) &\equiv \{z(f,\boldsymbol{w}_h) : \boldsymbol{w}_h \in N_j\}. \end{aligned}$$

Formulate a model for the $\{Z(f, \boldsymbol{w}_j) : j = 1, \ldots, k_f\}$ by specifying a conditional distributon for each variable as $N(\mu_{f,j}, \sigma_f^2)$ where

$$\mu_{f,j} = \mu_f + \eta \sum_{h \in N_j} \{z(f, \boldsymbol{w}_h) - \mu_f\}.$$

We would then assign distributions to the μ_f and σ_f^2 as in question 3. Analysis of this type of model using either Bayesian or non-Bayesian approaches would become more complex although, in principle, the major distinction with the previous material presented in the solution for question 6 is that the joint density for a given field could no longer be derived as a product of conditionally independent distributons. If we adhere to the use of Gaussian distributions, however, we know that the joint data model for observations in a field would be $Gau(\boldsymbol{\mu}_f, (I - C)^{-1}M)$, where $\boldsymbol{\mu}_f$ is a vector of length k_f having all elements equal to μ_f, C is a $k_f \times k_f$ symmetric matrix with j, hth element equal to η if $\boldsymbol{w}_j$ and $\boldsymbol{w}_h$ are neighbors and 0 otherwise, M is a $k_f \times k_f$ diagonal matrix with elements equal to σ_f^2, and I is the $k_f \times k_f$ identity matrix.

Company A manufactures a solution used to develop photographic film. The solution has a preservative that is chemically unstable. As time goes by, the preservative expires, and the developing solution becomes useless. The stability of the preservative is measured by means of a response that takes on values between 0 and 10. A value of 10 indicates that the preservative has suffered no decay since manufacturing, and the developing solution is "fresh." A value of 0 indicates complete decay of the preservative.

In order to assist Company A's management in answering questions such as

- When should inventories at the warehouse be renewed?
- What is the pattern of decay of the preservative in the developing solution under warehouse conditions?

chemists at the company conducted an experiment in which samples of developing solution were tested over a period of 8 months. Table 1 includes the data obtained in the experiment at the end of each month, as well as the partial derivatives of the mean response (see Model 1 below) with respect to α and β, evaluated at the least squares estimates (LSE) of the parameters. A plot of the data with the mean function evaluated at the LSE of α and β is included as Figure 1 on page 5.

Table 1: Data from 8 Month Experiment on Preservative

Month (x)	Response (y)	Derivative w.r.t α	Derivative w.r.t. β
1	10, 9.5, 10	0.15	−2.05
2	8.5, 8.0, 7.5	0.39	−4.44
3	7.3	0.56	−5.34
4	6.4	0.68	−5.40
5	6.6, 5.9	0.77	−5.02
6	6.0	0.83	−4.43
7	5.8	0.88	−3.78
8	5.8, 5.6	0.91	−3.15

The chemists had the following information prior to conducting their experiment:

- Little noticeable decay was expected during the first half month after manufacturing, i.e., it is expected that for $x = \frac{1}{2}$, $y = 10$.
- Decay of the preservative occurs after $x = \frac{1}{2}$.
- After enough time has elapsed, eventually the preservative will decay up to the point where the response is about 4. That is, as $x \to \infty$, y is close to 4.

It was postulated that a nonlinear model of the form

$$y_{ij} = \alpha + (10 - \alpha) \exp\left(-\beta(x_i - \frac{1}{2})\right) + \epsilon_{ij} \tag{1}$$

would be appropriate for $x_i \geq \frac{1}{2}$. Here, y_{ij} denotes the value of the response of the jth sample taken at the end of the ith month, x_i ($= i$ in the present context) is the number of months between

manufacture and testing, α and β are unknown, scalar-valued parameters, and ϵ_{ij} is a random error, where $\epsilon_{ij} \sim \text{Normal}(0, \sigma^2)$, and $\text{Cov}(\epsilon_{ij}, \epsilon_{kl}) = 0$, $(i,j) \neq (k,l)$. The main objective of the study was to estimate the parameters α and β of the model and thus be able to predict, for any given storage time $x > \frac{1}{2}$, the response value for the preservative.

1. Obtain the set of normal equations (NE) whose solutions are the least squares estimates (LSE) of α and β. **Do not attempt to solve the equations at this step.**

2. Approximate LS estimates of α and β can be obtained by a graphical method, without actually computing a solution to the NE. Explain how you would go about getting these approximate solutions.

3. One popular one-dimensional root-finding algorithm is the *Newton-Raphson* method. This technique finds values of z such that $f(z) = 0$. This method requires that both $f(z)$ and $f'(z)$ can be evaluated at arbitrary z. The Newton-Raphson formula consists geometrically of extending the tangent line to the curve $y = f(z)$ in the (z, y) plane at a current approximate root z_i until it crosses 0, then taking as the next approximate root z_{i+1} the abscissa (z-coordinate) of that zero-crossing.

   Discuss how to apply the Newton-Raphson method to solving the NE. Derive a general expression for z_{i+1} in terms of z_i, $f(z)$, and $f'(z)$, but **do not specifically evaluate the functions.**

4. Suppose you have chosen to iteratively solve the normal equations. Using the information provided by the chemists, find reasonable starting values for the unknown parameters α and β. Denote those initial values by $\alpha^{[0]}$ and $\beta^{[0]}$. Explain your selection.

5. The technique of linearizing a nonlinear mean function via a Taylor series expansion around some parameter value permits the use of iterative linear LS to estimate α and β.

   (a) Derive the approximating linear expansion for the mean function of Model 1 in α and β.

   (b) Derive the NE from the approximating linear expansion.

   (c) Explain how the solution would proceed.

6. LS estimates $\hat{\alpha}$ and $\hat{\beta}$ for α and β were obtained using the R function `nls`. Using information in the data set and in the attached output, perform an approximate test for lack of fit. Note that for some months there are replicate observations so that it is possible to calculate a pure error sum of squares.

7. (a) Obtain a standard error for the estimated mean response at $x = 2$. Recall that derivatives of the mean function of Model 1 evaluated at the LSE of α and β are available in Table 1.

   (b) Obtain a standard error of prediction for a single additional response at $x = 2$.

8. (a) Obtain a point estimate for the value of x at which the mean response first falls to 8.

   (b) Obtain an approximate confidence interval for the value of x at which the mean response first falls to 8.

Now we will consider a Bayesian approach to estimating the parameters α and β. Suppose that $\sigma = 0.4$. Suppose that we specify a prior distribution as follows. $\alpha \sim \text{Uniform}(0, 10)$, $\beta \sim \text{Gamma}(1, 2)$, α and β are independent *a priori*.

9. Use Bayes' Theorem to derive an expression for the joint posterior distribution for α and β. **Do not attempt to explicitly compute the normalizing constant for the joint posterior distribution**, but do provide an expression.

10. Describe how you would use the joint posterior distribution for (α, β) to derive point and interval estimates/predictions for

    (a) α

    (b) the mean response at $x = 2$

    (c) a new measured value of the response at $x = 2$

    Do not attempt to explicitly compute the integrals in these expressions.

```
> chem <- nls(index ~ a + (10 - a)*exp(-b*(month - 0.5)),
              start=list(a = a0,b = b0),trace=T)
#
# Part of the documentation of the nls() function includes:
# trace   value indicating if a trace of the iteration progress
#         should be printed. Default is FALSE. If TRUE the
#         residual (weighted) sum-of-squares and the parameter
#         values are printed at the conclusion of each iteration.

#sum-of-squares      alpha      beta
3.131861 :          4.0000000 0.1899635
2.802963 :          4.5635543 0.2245193
2.452439 :          5.1018572 0.2779595
2.069208 :          5.2696883 0.3306077
2.050609 :          5.1797727 0.3245042
2.050583 :          5.1831009 0.3252552
2.050583 :          5.1826175 0.3251754
2.050583 :          5.1826678 0.3251839

> summary(chem)

Formula: index ~ a + (10 - a) * exp(-b * (month - 0.5))

Parameters:
  Estimate Std. Error t value Pr(>|t|)
a  5.18267    0.43971  11.787  5.9e-08 ***
b  0.32518    0.06809   4.776 0.000452 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4134 on 12 degrees of freedom

Number of iterations to convergence: 7
Achieved convergence tolerance: 3.702e-06

> vcov(chem)
            a           b
a 0.19334157 0.027477003
b 0.02747700 0.004636724

> plot(month,index,xlim=c(0,12),ylim=c(5,10))
> a <- coef(chem)[1]
> b <- coef(chem)[2]
> m <- seq(0.5,12,length=1000)
> lines(m,a + (10 - a)*exp(-b*(m - 0.5)),type="l")
```

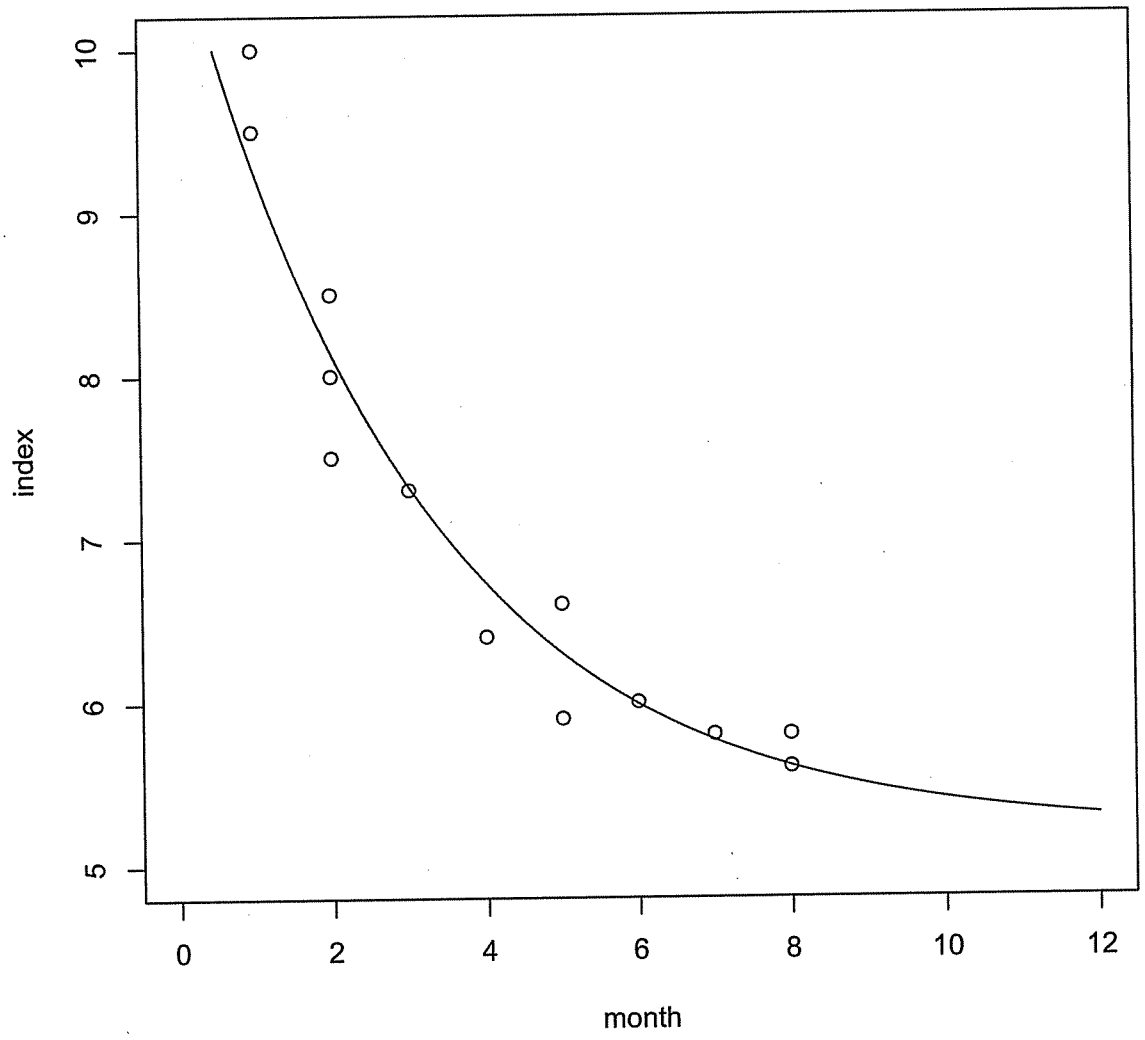


Figure 1: Data from Table 1 with the mean function of Model 1 evaluated at the LSE of α and β.

It was postulated that a nonlinear model of the form

$$y_{ij} = \alpha + (10 - \alpha)\exp\left(-\beta(x_i - \frac{1}{2})\right) + \epsilon_{ij} \tag{1}$$

would be appropriate for $x_i \geq \frac{1}{2}$. Here, y_{ij} denotes the value of the index of the jth sample taken at the end of the ith month, x_i $(= i$ in the present context) is the number of months between manufacture and testing, α and β are unknown, scalar-valued parameters, and ϵ_{ij} is a random error, where $\epsilon_{ij} \sim \text{Normal}(0, \sigma^2)$, and $\text{Cov}(\epsilon_{ij}, \epsilon_{kl}) = 0$, $(i, j) \neq (k, l)$.

1. Obtain the set of normal equations (NE) whose solutions are the least squares estimates (LSE) of α and β.

   Let

   $$Q(\alpha, \beta) = \sum_{ij}\left[y_{ij} - \alpha - (10 - \alpha)\exp\left(-\beta(x_i - \frac{1}{2})\right)\right]^2.$$

   We want to minimize $Q(\alpha, \beta)$ with respect to α, β.

$$\begin{aligned}
\frac{dQ}{d\alpha} &= 2\sum_{ij}\left[y_{ij} - \alpha - (10 - \alpha)\exp\left(-\beta(x_i - \frac{1}{2})\right)\right]\left[\exp\left(-\beta(x_i - \frac{1}{2})\right) - 1\right] \\
&= 2(\sum_{ij} y_{ij}\exp\left(-\beta(x_i - \frac{1}{2})\right) - \sum_{ij} y_{ij} + n\alpha + \\
&\quad (10 - 2\alpha)\sum_{ij}\exp\left(-\beta(x_i - \frac{1}{2})\right) - (10 - \alpha)\sum_{ij}\exp\left(-2\beta(x_i - \frac{1}{2})\right))
\end{aligned}$$

$$\begin{aligned}
\frac{dQ}{d\beta} &= 2\sum_{ij}\left[y_{ij} - \alpha - (10 - \alpha)\exp\left(-\beta(x_i - \frac{1}{2})\right)\right]\left[(10 - \alpha)(x_i - \frac{1}{2})\exp\left(-\beta(x_i - \frac{1}{2})\right)\right] \\
&= 2(10 - \alpha)(\sum_{ij} y_{ij}(x_i - \frac{1}{2})\exp\left(-\beta(x_i - \frac{1}{2})\right) - \alpha\sum_{ij}(x_i - \frac{1}{2})\exp\left(-\beta(x_i - \frac{1}{2})\right) - \\
&\quad (10 - \alpha)\sum_{ij}(x_i - \frac{1}{2})\exp\left(-2\beta(x_i - \frac{1}{2})\right))
\end{aligned}$$

   Setting $\frac{dQ}{d\alpha} = 0$ and $\frac{dQ}{d\beta} = 0$, we have the normal equations.

2. Approximate LS estimates of α and β can be obtained by a graphical method, without actually computing a solution to the NE. Explain how you would go about getting these approximate solutions.

   There are a variety of graphical approaches. Consider the following.

   (a) Rearrange the NE to solve for α.

$$\begin{aligned}
0 &= 2(\sum_{ij} y_{ij}\exp\left(-\beta(x_i - \frac{1}{2})\right) - \sum_{ij} y_{ij} + n\alpha + \\
&\quad (10 - 2\alpha)\sum_{ij}\exp\left(-\beta(x_i - \frac{1}{2})\right) - (10 - \alpha)\sum_{ij}\exp\left(-2\beta(x_i - \frac{1}{2})\right))
\end{aligned}$$

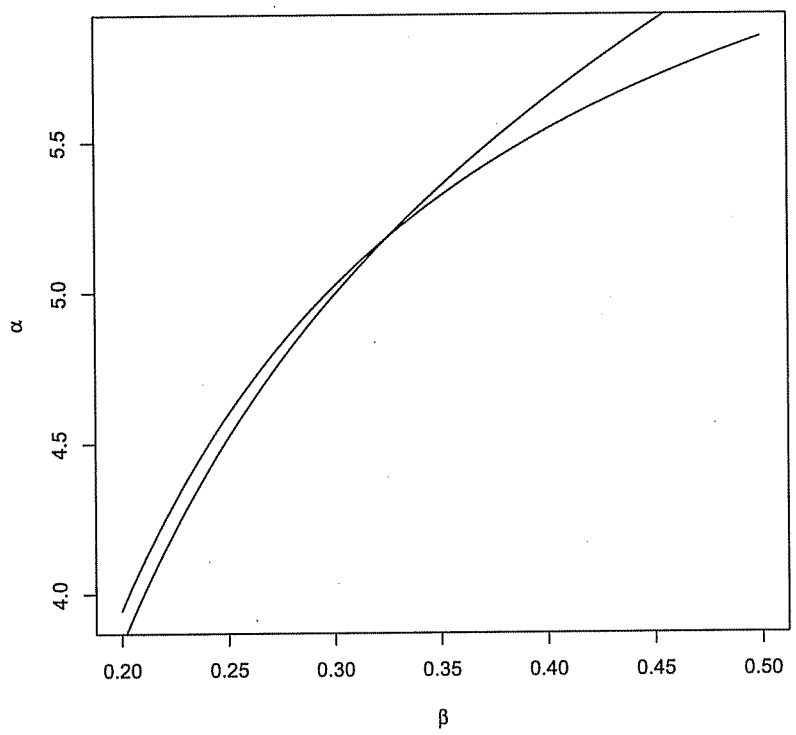


Figure 1: Approximate solution for NE: Solving both NE for α

$$\begin{aligned}\alpha &= \frac{\sum_{ij} y_{ij} - \sum_{ij} y_{ij} \exp\left(-\beta(x_i - \frac{1}{2})\right) - 10\sum_{ij} \exp\left(-\beta(x_i - \frac{1}{2})\right) + 10\exp\left(-2\beta(x_i - \frac{1}{2})\right)}{n - 2\sum_{ij} \exp\left(-\beta(x_i - \frac{1}{2})\right) + \sum_{ij} \exp\left(-2\beta(x_i - \frac{1}{2})\right)} \\ &= f_1(\beta)\end{aligned}$$

$$\begin{aligned}0 &= 2(10-\alpha)(\sum_{ij} y_{ij}(x_i - \frac{1}{2})\exp\left(-\beta(x_i - \frac{1}{2})\right) - \alpha(x_i - \frac{1}{2})\exp\left(-\beta(x_i - \frac{1}{2})\right) - \\ &\quad (10-\alpha)(x_i - \frac{1}{2})\exp\left(-2\beta(x_i - \frac{1}{2})\right)) \\ \alpha &= \frac{\sum_{ij} y_{ij}(x_i - \frac{1}{2})\exp\left(-\beta(x_i - \frac{1}{2})\right) - 10\sum_{ij}(x_i - \frac{1}{2})\exp\left(-2\beta(x_i - \frac{1}{2})\right)}{\sum_{ij}(x_i - \frac{1}{2})\exp\left(-\beta(x_i - \frac{1}{2})\right) - \sum_{ij}(x_i - \frac{1}{2})\exp\left(-2\beta(x_i - \frac{1}{2})\right)} \\ &= f_2(\beta)\end{aligned}$$

We can plot $(\beta, f_1(\beta) = \alpha)$ and $(\beta, f_2(\beta) = \alpha)$ on the same axes. Our approximate solution is found where these curves intersect. See Figure 1.

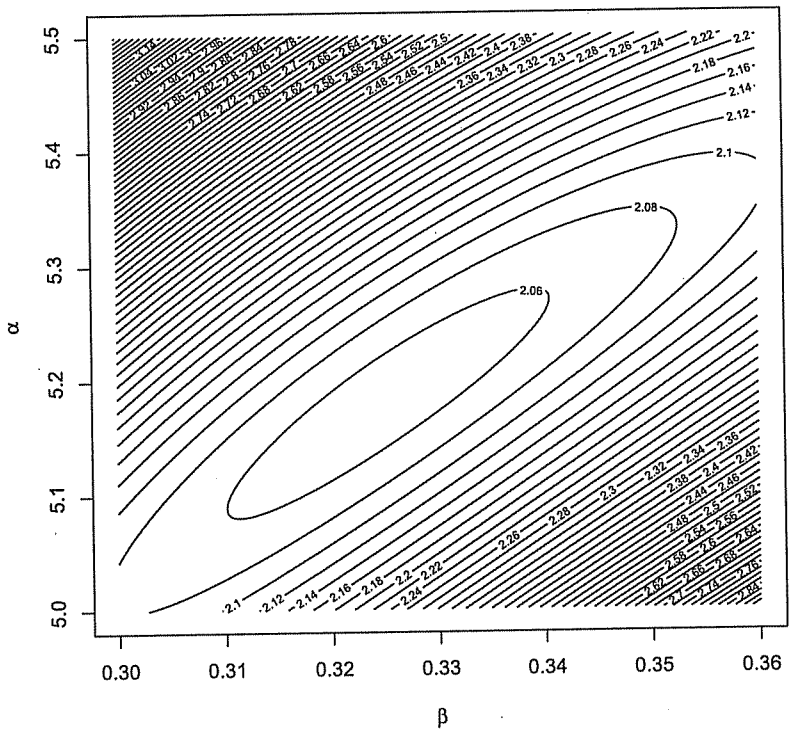


Figure 2: Approximate solution for NE: Contour plot of $Q(\alpha, \beta)$

(b) We can make a contour plot of $Q(\alpha, \beta)$. See Figure 2.

3. One popular one-dimensional root-finding algorithm is the *Newton-Raphson* method. This technique finds values of z such that $f(z) = 0$. This method requires that both $f(z)$ and $f'(z)$ can be evaluated at arbitrary z. The Newton-Raphson formula consists geometrically of extending the tangent line to the curve $y = f(z)$ in the (z, y) plane at a current point z_i until it crosses 0, then setting the next guess z_{i+1} to the abscissa (z-coordinate) of that zero-crossing.

   Discuss how to apply the Newton-Raphson method to solving the NE. Derive a general expression for z_{i+1} in terms of z_i, $f(z)$, and $f'(z)$, but *do not specifically evaluate the functions.*

   Above, we showed how to rearrange the NE to solve for α. Since $f_1(\beta) = f_2(\beta)$, we have $f_1(\beta) - f_2(\beta) = f(\beta) = 0$. We can solve this using Newton-Raphson.

   From the geometry of Newton-Raphson, we know that

$$\begin{aligned} f(\beta_i) &= f'(\beta_i)\beta_i + b \\ 0 &= f'(\beta_i)\beta_{i+1} + b \end{aligned}$$

   Subtracting, we have $f(\beta_i) = f'(\beta_i)(\beta_i - \beta_{i+1})$, which implies $\beta_{i+1} = \beta_i - \frac{f(\beta_i)}{f'(\beta_i)}$. Iterate this to convergence, then substitute into either NE to solve for an estimate of α.

4. Suppose you have chosen to iteratively solve the normal equations. Using the information provided by the chemists, find reasonable starting values for the unknown parameters α and β. Denote those initial values by $\alpha^{[0]}$ and $\beta^{[0]}$. Explain your selection.

   From the information given in the problem, we know that $y \to 4$ as $x \to \infty$. Since as $x \to \infty$, $y \to \alpha$, we choose $\alpha^{[0]} = 4$.

   There are several reasonable approaches for choosing $\beta^{[0]}$. From the nonlinear model, we know that

$$\beta = \frac{-\log(\frac{y_{ij}-\alpha}{10-\alpha})}{x_i - \frac{1}{2}}$$

   Suppose that for each y_{ij} we evaluate this expression using $\alpha = \alpha^{[0]}$ to calculate a "β_{ij}". We could then set $\beta^{[0]} = \frac{1}{n}\sum_{ij}\beta_{ij} = 0.19$.

5. The technique of linearizing a nonlinear mean function via a Taylor series expansion around some parameter value permits the use of iterative linear LS to estimate α and β in Model 1.

   (a) Derive the approximating linear expansion for the mean function of Model 1 in α and β.

   (b) Derive the NE from the approximating linear expansion.

   (c) Explain how the solution would proceed.

   Let $g(\alpha, \beta, x) = \alpha + (10 - \alpha)\exp\left(-\beta(x - \frac{1}{2})\right)$. The linear expansion is

$$\begin{aligned} g(\alpha,\beta,x) &\approx g(\alpha_0,\beta_0,x) + \frac{dg}{d\alpha}|_{\alpha=\alpha_0}(\alpha-\alpha_0) + \frac{dg}{d\beta}|_{\beta=\beta_0}(\beta-\alpha_0) \\ &= \alpha_0 + (10-\alpha_0)\exp\left(-\beta_0(x-\frac{1}{2})\right) + \alpha - \alpha_0 - \exp\left(-\beta_0(x-\frac{1}{2})\right)(\alpha-\alpha_0) \\ &\quad -(10-\alpha_0)(x-\frac{1}{2})\exp\left(-\beta_0(x-\frac{1}{2})\right)(\beta-\beta_0) \\ &= 10\exp\left(-\beta_0(x-\frac{1}{2})\right) + \beta_0(10-\alpha_0)(x-\frac{1}{2})\exp\left(-\beta_0(x-\frac{1}{2})\right) \\ &\quad +\alpha(1-\exp\left(-\beta_0(x-\frac{1}{2})\right)) - \beta(10-\alpha_0)(x-\frac{1}{2})\exp\left(-\beta_0(x-\frac{1}{2})\right) \end{aligned}$$

   Let

$$Q(\alpha,\beta) = \sum_{ij}\left[y_{ij} - g(\alpha,\beta,x_i)\right]^2.$$

   We want to minimize $Q(\alpha, \beta)$ with respect to α, β. Let

$$\begin{aligned} k_{1i} &= 10\exp\left(-\beta_0(x_i-\frac{1}{2})\right) + \beta_0(10-\alpha_0)(x_i-\frac{1}{2})\exp\left(-\beta_0(x_i-\frac{1}{2})\right) \\ k_{2i} &= 1-\exp\left(-\beta_0(x_i-\frac{1}{2})\right) \\ k_{3i} &= (10-\alpha_0)(x_i-\frac{1}{2})\exp\left(-\beta_0(x_i-\frac{1}{2})\right) \end{aligned}$$

$$\frac{dQ}{d\alpha} = -2\sum_{ij}(y_{ij} - k_{1i} - k_{2i}\alpha + k_{3i}\beta)k_{2i}$$

$$\frac{dQ}{d\beta} = 2\sum_{ij}(y_{ij} - k_{1i} - k_{2i}\alpha + k_{3i}\beta)k_{3i}$$

Setting $\frac{dQ}{d\alpha} = 0$ and $\frac{dQ}{d\beta} = 0$, we have the normal equations.

Solving the normal equations, we have

$$\begin{aligned}\hat{\beta} &= \frac{(\sum_{ij} k_{2i}k_{3i})(\sum_{ij}(y_{ij} - k_{1i})k_{2i}) - (\sum_{ij} k_{2i}^2)(\sum_{ij}(y_{ij} - k_{1i})k_{3i})}{(\sum_{ij} k_{3i}^2)(\sum_{ij} k_{2i}^2) - (\sum_{ij} k_{3i}k_{2i})^2} \\ \hat{\alpha} &= \frac{\sum_{ij}(y_{ij} - k_{1i})k_{2i} + \hat{\beta}\sum_{ij} k_{3i}k_{2i}}{\sum_{ij} k_{2i}^2}\end{aligned}$$

Now choose $\alpha_0 = \alpha^{[0]}$ and $\beta_0 = \beta^{[0]}$. Solve for $\hat{\alpha} = \alpha^{[1]}$ and $\hat{\beta} = \beta^{[1]}$. Set $\alpha_0 = \alpha^{[1]}$ and $\beta_0 = \beta^{[1]}$. Iterate until convergence.

6. LS estimates $\hat{\alpha}$ and $\hat{\beta}$ for α and β were obtained using the R function `nls`. Using information in the data set and in the attached output, perform an approximate test for lack of fit. Note that for some months there are replicate observations so that it is possible to calculate a pure error sum of squares.

   To calculate the approximate test, we need to calculate $F^* = \frac{\text{MSLF}}{\text{MSPE}}$, with $df_{LF} = df_{PE} = 6$.

   - Let c = levels of $x = 8$, p = total number of parameters = 2, n = total number of observations = 14
   - MSLF = $\frac{\text{SSE–SSPE}}{c-p}$
   - MSPE = $\frac{\text{SSPE}}{n-c}$
   - SSE = $\sum_{ij}(Y_{ij} - \hat{Y}_{ij})^2$
   - SSPE = $\sum_{i}(Y_{ij} - \bar{Y}_{j})^2$

   We have SSE = 2.051, SSPE = 0.932, and $F^* = \frac{2.051-0.932}{0.932} = 1.20$, which indicates no lack of fit.

7. (a) Obtain a standard error for the estimated mean index at $x = 2$. Recall that derivatives of the mean function of Model 1 evaluated at the LSE of α and β are available in Table 1.

   Denote the estimated mean response at $x = 2$ as $\hat{Y}_2$. Let $g(\alpha, \beta) = \alpha + (10-\alpha)\exp\left(-\beta(x_i - \frac{1}{2})\right)$.

   $$s^2(\hat{Y}_2) \approx \hat{G}'\text{Var}(\alpha, \beta)\hat{G}$$

   where

   $$\hat{G} = \begin{pmatrix} \frac{dg}{d\alpha} \\ \frac{dg}{d\beta} \end{pmatrix}|_{\alpha=\hat{\alpha},\beta=\hat{\beta}}$$

   $\text{Var}(\alpha, \beta)$ is available from the R output, and $\hat{G}$ is available at $x = 2$ in the data table. Calculating, we find $s^2(\hat{Y}_2) \approx 0.026$.

(b) Obtain a standard error of prediction for a single additional index at $x = 2$.
We also have $s^2(\hat{Y}_{pred}) = \text{MSE} + s^2(\hat{Y}_2)$. SSE and the residual standard error are available from the R output. Calculating, we have $s^2(\hat{Y}_{pred}) \approx 0.026 + 0.171 = 0.197$.

8. (a) Obtain a point estimate for the value of x at which the mean index first falls to 8.
Setting the expression for the mean index to 8,

$$\begin{aligned} 8 &= \alpha + (10 - \alpha)\exp(-\beta(x - \frac{1}{2})) \\ x &= h(\alpha, \beta) = -\frac{1}{\beta}\log(8 - \alpha) + \frac{1}{\beta}\log(10 - \alpha) + \frac{1}{2} \end{aligned}$$

Given least squares estimates $\hat{\alpha} = 5.183$ and $\hat{\beta} = 0.325$, we calculate our point estimate $\hat{x} = h(\hat{\alpha}, \hat{\beta}) = 2.150$.

(b) Obtain an approximate confidence interval for the value of x at which the mean index first falls to 8.
Taking partial derivatives of $h(\alpha, \beta)$,

$$\begin{aligned} \frac{\partial h}{\partial \alpha} &= \frac{1}{\beta(8 - \alpha)} - \frac{1}{\beta(10 - \alpha)} \\ \frac{\partial h}{\partial \beta} &= \frac{1}{\beta^2}\log(8 - \alpha) - \frac{1}{\beta^2}\log(10 - \alpha) \end{aligned}$$

We know that

$$s^2(\hat{x}) \approx \hat{G}' \text{Var}(\alpha, \beta) \hat{G}$$

where

$$\hat{G} = \begin{pmatrix} \frac{dh}{d\alpha} \\ \frac{dh}{d\beta} \end{pmatrix} \Big|_{\alpha=\hat{\alpha}, \beta=\hat{\beta}}$$

$\text{Var}(\alpha, \beta)$ is available from the R output. We can evaluate

$$\begin{aligned} \hat{G} &= \begin{pmatrix} \frac{dh}{d\alpha} \\ \frac{dh}{d\beta} \end{pmatrix} \Big|_{\alpha=\hat{\alpha}, \beta=\hat{\beta}} \\ &= \begin{pmatrix} 0.453 \\ -5.072 \end{pmatrix} \end{aligned}$$

which means that $s^2(\hat{x}) = 0.0327$. An approximate 95% confidence interval for the number of months before the mean index degrades below 8 is $\hat{x} \pm t_{12,\alpha/2}\sqrt{s^2(\hat{x})} = 2.150 \pm 0.394 = (1.756, 2.394)$.

Now we will consider a Bayesian approach to estimating the parameters α and β. Suppose that $\sigma = 0.4$. Suppose that we specify a prior distribution as follows. $\alpha \sim \text{Uniform}(0, 10)$, $\beta \sim \text{Gamma}(1, 2)$, α and β are independent *a priori*.

9. Use Bayes Theorem to derive an expression for the joint posterior distribution for α and β.
Let $g(\alpha, \beta, x) = \alpha + (10 - \alpha)\exp\left(-\beta(x - \frac{1}{2})\right)$. We know that

$$Y_{ij} \sim \text{Normal}(g(\alpha, \beta, x_i), \sigma^2)$$

Bayes Theorem gives

$$\pi(\alpha, \beta \,|\, \mathbf{y}) = \frac{f(\mathbf{y} \,|\, \alpha, \beta)\pi(\alpha, \beta)}{\int\int f(\mathbf{y} \,|\, \alpha, \beta)\pi(\alpha, \beta) d\alpha d\beta}$$

where

$$f(\mathbf{y} \,|\, \alpha, \beta) = \prod_{ij} \frac{1}{\sqrt{2\pi}(0.4)} \exp(-\frac{1}{2(0.4)^2}(y_{ij} - g(\alpha, \beta, x_i))^2)$$

and

$$\pi(\alpha, \beta) = 0.2 \exp(-2\beta), \quad 0 \leq \alpha \leq 10, \beta > 0$$

10. Describe how you would use the joint posterior distribution to derive point and interval estimates for

    (a) α

    (b) the mean index at $x = 2$

    (c) a new measured value of the index at $x = 2$

    In each case, we can derive the posterior distribution of the quantities of interest and use the posterior mean as a point estimate and a posterior credible interval with appropriate probability as the interval estimate. Let $\pi(\alpha, \beta \,|\, \mathbf{y})$ denote the joint posterior distribution of α and β.

    (a)

    $$\pi(\alpha \,|\, \mathbf{y}) = \int_0^\infty \pi(\alpha, \beta \,|\, \mathbf{y}) d\beta$$

    (b) The mean index at $x = 2$ is $\alpha + (10 - \alpha)\exp(-1.5\beta)$. This is a function of α and β, and we use the change-of-variables theorem starting with the joint posterior distribution $\pi(\alpha, \beta)$.

    (c)

    $$\pi(y_{new} \,|\, \mathbf{y}) = \int_0^{10} \int_0^\infty \frac{1}{\sqrt{2\pi}(0.4)} \exp(-\frac{1}{2(0.4)^2}(y_{new} - g(\alpha, \beta, 2))^2)\pi(\alpha, \beta \,|\, \mathbf{y}) d\beta d\alpha$$