PhD Prelim Exam
METHODS

(Majors & Co-Majors)

Summer 2013
(Given on 7/16/13)

Alfalfa is a perennial legume sometimes called the *queen of hay* due to its high nutritional value. Its primary use is as feed for dairy cows because of its high protein content (up to 20% to 22% depending on soil and other environmental factors) and its highly digestible fiber. A large proportion of the alfalfa produced in the United States is grown in California, Texas, Wisconsin and New York, states with high concentrations of dairy farms. Alfalfa can be harvested several times per year (up to 10 "cuttings" in warm, humid climates and up to four or five in cooler weathers) and in ideal conditions can yield up to 10 tons per acre per year.

An experiment conducted at the Cornell Experiment Station was designed to compare the yield in tons per acre (t/a) for six different varieties of alfalfa, under several fertilizer treatments. The six alfalfa varieties included in the experiment were: Atlantic (`A`), Grimm (`G`), K. Command (`C`), Narragannsut (`N`), Ontario (`O`) and Ranger (`R`). The fertilizer treatments consisted of all 2×2 combinations of high and low levels of potassium and phosphorous. We use `k` and `K` to denote the low and high levels of potassium, respectively. We use `p` and `P` to denote the low and high levels of phosphorous, respectively. The experiment was repeated in two fields or blocks, as illustrated in Figure 1. The data and an analysis of the experiment are discussed in Casella (2008).

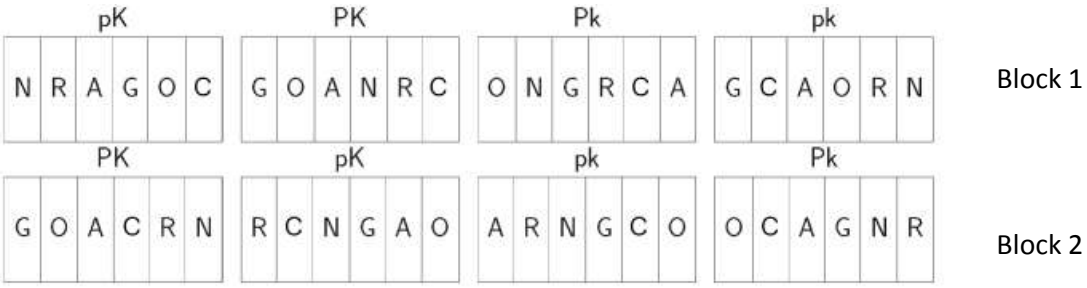

Figure 1: One possible physical layout for the alfalfa experiment

Operationally, each of the two fields was divided into four plots. The four fertilizer treatments (`pk`, `Pk`, `pK` or `PK`) were randomly assigned to plots. Plots were then divided into six sub-plots of equal size and the six varieties were planted in randomly assigned sub-plots in each plot. During the growing season, the alfalfa was cut three times; the yields reported in the experiment were the sum of the yields in the three cuttings in each subplot. A total of $N = 48$ alfalfa yield measurements were reported. The yield measurements in t/a are shown in Table 1 (page 3).

We use Y_{ijk} to denote the yield measurement in the ith fertilizer treatment in the jth block and for the kth variety, where $i = 1, ..., f$, $j = 1, ..., b$ and $k = 1, ..., g$. In this experiment, $f = 4$, $b = 2$ and $g = 6$. The observed mean alfalfa yields for each

fertilizer treatment and variety combination are denoted by $\bar{Y}_{i.k}$ and are computed as

$$\bar{Y}_{i.k} = \frac{1}{2}\sum_{j=1}^{2} Y_{ijk}, \quad i = 1, ..., 4; \;\; k = 1, ..., 6.$$

The observed mean alfalfa yield for fertilizer treatment i is denoted by $\bar{Y}_{i..}$ and is computed as

$$\bar{Y}_{i..} = \frac{1}{12}\sum_{j=1}^{2}\sum_{k=1}^{6} Y_{ijk}, \quad i = 1, ..., 4.$$

Similarly, the observed mean alfalfa yield for variety k is denoted by $\bar{Y}_{..k}$ and is computed as

$$\bar{Y}_{..k} = \frac{1}{8}\sum_{i=1}^{4}\sum_{j=1}^{2} Y_{ijk}, \quad k = 1, ..., 6.$$

Mean yields are shown in Table 2 (page 3).

A model that describes the experiment is

$$Y_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{ij} + \gamma_k + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + \delta_{ijk}, \qquad \text{(Model 1)}$$

$i = 1, ..., 4;\; j = 1, 2;\; k = 1, ..., 6$ where

$$\begin{aligned}
\mu &= \text{a fixed overall mean effect} \\
\tau_i &= \text{a fixed effect of fertilizer treatment } i \\
\beta_j &= \text{a random block effect for block } j \\
\epsilon_{ij} &= \text{a random whole plot error, for treatment } i \text{ in block } j \\
\gamma_k &= \text{a fixed variety effect for variety } k \\
(\tau\gamma)_{ik} &= \text{a fixed interaction effect of fertilizer treatment } i \text{ by variety } k \\
(\beta\gamma)_{jk} &= \text{a random interaction effect of block } j \text{ by variety } k \\
\delta_{ijk} &= \text{a random sub-plot error}
\end{aligned}$$

We assume that $\beta_j \sim iid\ \mathrm{N}(0, \sigma^2_\beta)$, $\epsilon_{ij} \sim iid\ \mathrm{N}(0, \sigma^2_\epsilon)$, $(\beta\gamma)_{jk} \sim iid\ \mathrm{N}(0, \sigma^2_{\beta\gamma})$ and $\delta_{ijk} \sim iid\ \mathrm{N}(0, \sigma^2_\delta)$. All random effects are assumed to be independent, and for identifiability reasons we impose the parameter restrictions

$$\bar{\tau} = \bar{\gamma} = (\overline{\tau\gamma})_{i.} = (\overline{\tau\gamma})_{.k} = 0,$$

where

$$\bar{\tau} = \frac{1}{4}\sum_{i=1}^{4}\tau_i, \quad \bar{\gamma} = \frac{1}{6}\sum_{k=1}^{6}\gamma_k, \quad (\overline{\tau\gamma})_{i.} = \frac{1}{6}\sum_{k=1}^{6}(\tau\gamma)_{ik} \;\text{ and }\; (\overline{\tau\gamma})_{.k} = \frac{1}{4}\sum_{i=1}^{4}(\tau\gamma)_{ik}\ .$$

Block 1			Block 2		
Treatment	Variety	Yield (t/a)	Treatment	Variety	Yield (t/a)
pk	A	3.52	pk	A	3.61
pk	G	3.13	pk	G	4.43
pk	C	3.03	pk	C	4.15
pk	N	3.74	pk	N	4.65
pk	O	3.21	pk	O	4.72
pk	R	3.24	pk	R	4.21
Pk	A	4.54	Pk	A	4.53
Pk	G	3.88	Pk	G	4.80
Pk	C	4.10	Pk	C	4.74
Pk	N	3.65	Pk	N	5.45
Pk	O	4.06	Pk	O	4.54
Pk	R	4.42	Pk	R	4.94
pK	A	3.74	pK	A	3.40
pK	G	2.86	pK	G	3.81
pK	C	3.42	pK	C	4.14
pK	N	3.98	pK	N	4.78
pK	O	4.32	pK	O	3.78
pK	R	3.31	pK	R	3.93
PK	A	4.95	PK	A	5.39
PK	G	4.05	PK	G	5.25
PK	C	4.84	PK	C	5.15
PK	N	4.82	PK	N	6.13
PK	O	5.29	PK	O	6.00
PK	R	3.94	PK	R	5.80

Table 1: Yields of alfalfa (t/a).

Fertilizer	Variety						
treatment	A	G	C	N	O	R	$\bar{Y}_{i\cdot\cdot}$
pk	3.56	3.78	3.59	4.20	3.97	3.73	3.83
Pk	4.54	4.34	4.42	4.55	4.30	4.68	4.47
pK	3.57	3.34	3.78	4.38	4.05	3.62	3.79
PK	5.17	4.65	5.00	5.48	5.65	4.87	5.13
$\bar{Y}_{\cdot\cdot k}$	4.21	4.03	4.20	4.65	4.49	4.22	

Table 2: Mean yields for fertilizer treatment by variety combinations (t/a) and their averages.

Part I

1. Given the design of the alfalfa experiment, which effects – fertilizer treatment effects or variety effects – do you think will be more precisely estimated? Why?

2. Under Model 1, give expressions for the covariances among observations within the same block with:

a. Different fertilizers, but the same variety

b. Same fertilizer, but different varieties

c. Different fertilizers, and different varieties

3. Give the degrees of freedom and compute the mean squares for the ANOVA table below:

Source	df	SS	MS
Block		6.961	
Fertilizer		14.775	
Fertilizer × Block		0.746	
Variety		2.071	
Fertilizer × Variety		1.526	
Variety × Block		1.849	
Fertilizer × Variety × Block		1.562	

4. The expectations of five of the mean squares in the ANOVA table from question **3** are given below. Derive the expectations of the mean square for Variety and of the mean square for the Fertilizer × Variety interaction, in terms of the model parameters. We use Fert, Var instead of Fertilizer, Variety to shorten the notation.

$$\begin{aligned}
E[MS(\text{Block})] &= gf\sigma^2_\beta + f\sigma^2_{\beta\gamma} + g\sigma^2_\epsilon + \sigma^2_\delta \\
E[MS(\text{Fert})] &= \frac{bg}{f-1}\sum_i \tau_i^2 + g\sigma^2_\epsilon + \sigma^2_\delta \\
E[MS(\text{Fert} \times \text{Block})] &= g\sigma^2_\epsilon + \sigma^2_\delta \\
E[MS(\text{Var} \times \text{Block})] &= f\sigma^2_{\beta\gamma} + \sigma^2_\delta \\
E[MS(\text{Fert} \times \text{Var} \times \text{Block})] &= \sigma^2_\delta
\end{aligned}$$

5. Test the hypothesis that there are no differences in mean yield due to fertilizer treatment. Use a 5% significance level ($\alpha = 0.05$) and state your conclusions.

6. Test the two hypotheses:

a. of no differences between varieties, and

b. of no interaction between fertilizer treatments and varieties.

In both tests, use a 5% significance level and state your conclusions.

7. Consider the comparison between the high and the low values of phosphorous (`P` versus `p`) when potassium is applied at its low level (`k`). Compute a point estimate of an appropriate contrast and obtain a 95% confidence interval for the corresponding true difference in model parameters. Interpret your results.

8. Suppose now that we wish to compare interactions $(\tau\gamma)_{ik}$ at different levels of the fertilizer treatment. The estimable contrast potentially of interest is

$$\eta = \sum_{ik} a_{ik}(\tau\gamma)_{ik},$$

which can be estimated unbiasedly by

$$\hat{\eta} = \sum_{ik} a_{ik}\bar{Y}_{ik} \ .$$

a. Derive an expression for the variance of $\hat{\eta}$.

b. Assume that $\sigma^2_{\beta\gamma} = 0$ and show that

$$\mathrm{Var}(\hat{\eta}) = \frac{\sigma^2_\delta}{b}\sum_{ik}(a_{ik} - \bar{a}_i)^2 + \frac{\sigma^2_\delta + g\sigma^2_\epsilon}{bg}\sum_i(\sum_k a_{ik})^2.$$

c. Refer to the observed mean yields in Table 2 (page 3). We wish to compare the mean yield of variety `R` in fertilizer treatment `PK` with the mean yield of variety `A` in fertilizer treatment `pk`. For this comparison, we have $a_{11} = -1$, $a_{46} = 1$ and all other $a_{ik} = 0$. Calculate a point estimate for the contrast and compute its variance. Assume that $\sigma^2_{\beta\gamma} = 0$ as in question **8b.**

d. Give the degrees of freedom for the Satterthwaite approximation to the sampling distribution of $\hat{\mathrm{Var}}(\hat{\eta})$ using the expression in question **8b.** that assumes that $\sigma^2_{\beta\gamma} = 0$.

Part II

Alfalfa can be harvested (cut) several times per year, during its growing season. In the experiment carried out at the Cornell Experiment Station,all sub-plots were harvested three times, at equally spaced two-month intervals. Therefore, the sub-plot yields Y_{ijk} shown in Table 1 are the sums of the yields obtained in the three cuttings in the sub-plots. We use t to index time, so that $t = 1, ..., 3$. Let Y_{ijkt} denote the yield observed at time t in a sub-plot in block j, under fertilizer treatment i and planted with variety k.

A model that can be used for the new response Y_{ijkt} is

$$\begin{aligned} Y_{ijkt} &= \mu + \tau_i + \beta_j + \epsilon_{ij} + \gamma_k + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + \delta_{ijk} \\ &\quad + C_k + (\tau C)_{it} + (\gamma C)_{kt} + (\tau\gamma C)_{ikt} + (\beta C)_{jt} + \theta_{ijt} + \phi_{ikt} + \psi_{ijkt}, \end{aligned} \quad \text{(Model 2)}$$

$i = 1, ..., 4;\ j = 1, 2;\ k = 1, ..., 6;\ t = 1, ..., 3$ where

$$\begin{aligned} \mu &= \text{a fixed overall mean effect} \\ \tau_i &= \text{a fixed effect of fertilizer treatment } i \\ \beta_j &= \text{a random block effect for block } j \\ \epsilon_{ij} &= \text{a random whole plot error, for treatment } i \text{ in block } j \\ \gamma_k &= \text{a fixed variety effect for variety } k \\ (\tau\gamma)_{ik} &= \text{a fixed interaction effect of fertilizer treatment } i \text{ by variety } k \\ (\beta\gamma)_{jk} &= \text{a random interaction effect of block } j \text{ by variety } k \\ \delta_{ijk} &= \text{a random sub-plot error} \\ C_t &= \text{a fixed cutting effect for cutting } t \\ (\beta C)_{jt} &= \text{a random interaction between block } j \text{ and cutting } t \\ \theta_{ijt} &= (\tau\beta C)_{ijt}, \text{ a random interaction between treatment } i, \text{ block } j, \text{ and cutting } t \\ \phi_{jkt} &= (\beta\gamma C)_{jkt}, \text{ a random interaction between block } j, \text{ variety } k, \text{ and cutting } t \\ \psi_{ijkt} &= (\tau\beta\gamma C)_{ijkt}, \text{ a random error} \end{aligned}$$

As before, we assume that $\beta_j \sim iid\ \mathrm{N}(0, \sigma^2_\beta)$, $\epsilon_{ij} \sim iid\ \mathrm{N}(0, \sigma^2_\epsilon)$, $(\beta\gamma)_{jk} \sim iid\ \mathrm{N}(0, \sigma^2_{\beta\gamma})$ and $\delta_{ijk} \sim iid\ \mathrm{N}(0, \sigma^2_\delta)$. Further, we now also assume that $(\beta C)_{jt} \sim iid\ \mathrm{N}(0, \sigma^2_{\beta C})$, $\theta_{ijt} \sim iid\ \mathrm{N}(0, \sigma^2_\theta)$, $\phi_{jkt} \sim iid\ \mathrm{N}(0, \sigma^2_\phi)$ and $\psi_{ijkt} \sim iid\ \mathrm{N}(0, \sigma^2_\psi)$. All random effects are assumed to be independent.

9. What is the covariance between yields for two different cuttings in the same sub-plot based on Model 2?

10. There are two degrees of freedom associated with the effects of cutting. Construct two orthogonal one-degree of freedom contrasts to test for non-zero linear and quadratic effects, respectively, of cutting on yield for fertilizer treatment `pk`, in block 1, for variety `A`, and derive the variance of each of the contrasts under Model 2.

11. Suppose that you find that the effects of:

a. fertilizer treatments,

b. interactions between fertilizer treatment and the linear trend in cutting

are both significantly different from zero. In your own words, interpret those two results for someone who is not a statistician.

Alfalfa is a perennial legume sometimes called the *queen of hay* due to its high nutritional value. Its primary use is as feed for dairy cows because of its high protein content (up to 20% to 22% depending on soil and other environmental factors) and its highly digestible fiber. A large proportion of the alfalfa produced in the United States is grown in California, Texas, Wisconsin and New York, states with high concentrations of dairy farms. Alfalfa can be harvested several times per year (up to 10 "cuttings" in warm, humid climates and up to four or five in cooler weathers) and in ideal conditions can yield up to 10 tons per acre per year.

An experiment conducted at the Cornell Experiment Station was designed to compare the yield in tons per acre (t/a) for six different varieties of alfalfa, under several fertilizer treatments. The six alfalfa varieties included in the experiment were: Atlantic (`A`), Grimm (`G`), K. Command (`C`), Narragannsut (`N`), Ontario (`O`) and Ranger (`R`). The fertilizer treatments consisted of all 2×2 combinations of high and low levels of potassium and phosphorous. We use `k` and `K` to denote the low and high levels of potassium, respectively. We use `p` and `P` to denote the low and high levels of phosphorous, respectively. The experiment was repeated in two fields or blocks, as illustrated in Figure 1. The data and an analysis of the experiment are discussed in Casella (2008).

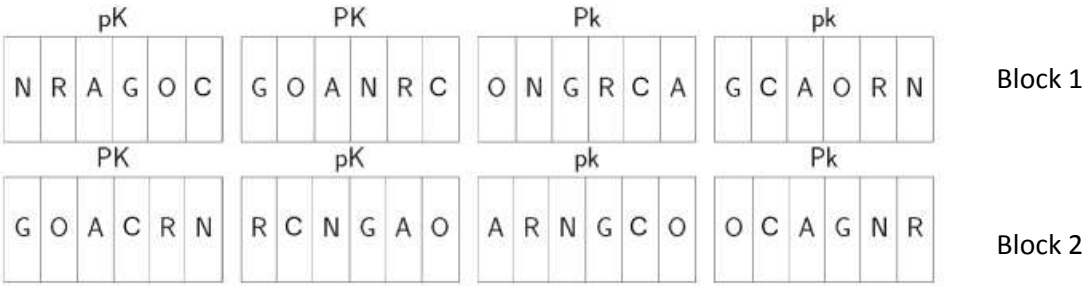

Figure 1: One possible physical layout for the alfalfa experiment

Operationally, each of the two fields was divided into four plots. The four fertilizer treatments (`pk`, `Pk`, `pK` or `PK`) were randomly assigned to plots. Plots were then divided into six sub-plots of equal size and the six varieties were planted in randomly assigned sub-plots in each plot. During the growing season, the alfalfa was cut three times; the yields reported in the experiment were the sum of the yields in the three cuttings in each subplot. A total of $N = 48$ alfalfa yield measurements were reported. The yield measurements in t/a are shown in Table 1 (page 3).

We use Y_{ijk} to denote the yield measurement in the ith fertilizer treatment in the jth block and for the kth variety, where $i = 1, ..., f$, $j = 1, ..., b$ and $k = 1, ..., g$. In this experiment, $f = 4$, $b = 2$ and $g = 6$. The observed mean alfalfa yields for each

fertilizer treatment and variety combination are denoted by $\bar{Y}_{i.k}$ and are computed as

$$\bar{Y}_{i.k} = \frac{1}{2}\sum_{j=1}^{2} Y_{ijk}, \quad i = 1, ..., 4; \;\; k = 1, ..., 6.$$

The observed mean alfalfa yield for fertilizer treatment i is denoted by $\bar{Y}_{i..}$ and is computed as

$$\bar{Y}_{i..} = \frac{1}{12}\sum_{j=1}^{2}\sum_{k=1}^{6} Y_{ijk}, \quad i = 1, ..., 4.$$

Similarly, the observed mean alfalfa yield for variety k is denoted by $\bar{Y}_{..k}$ and is computed as

$$\bar{Y}_{..k} = \frac{1}{8}\sum_{i=1}^{4}\sum_{j=1}^{2} Y_{ijk}, \quad k = 1, ..., 6.$$

Mean yields are shown in Table 2 (page 3).

A model that describes the experiment is

$$Y_{ijk} = \mu + \tau_i + \beta_j + \epsilon_{ij} + \gamma_k + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + \delta_{ijk}, \qquad \text{(Model 1)}$$

$i = 1, ..., 4;\; j = 1, 2;\; k = 1, ..., 6$ where

$$\begin{aligned}
\mu &= \text{a fixed overall mean effect} \\
\tau_i &= \text{a fixed effect of fertilizer treatment } i \\
\beta_j &= \text{a random block effect for block } j \\
\epsilon_{ij} &= \text{a random whole plot error, for treatment } i \text{ in block } j \\
\gamma_k &= \text{a fixed variety effect for variety } k \\
(\tau\gamma)_{ik} &= \text{a fixed interaction effect of fertilizer treatment } i \text{ by variety } k \\
(\beta\gamma)_{jk} &= \text{a random interaction effect of block } j \text{ by variety } k \\
\delta_{ijk} &= \text{a random sub-plot error}
\end{aligned}$$

We assume that $\beta_j \sim iid\ \mathrm{N}(0, \sigma^2_\beta)$, $\epsilon_{ij} \sim iid\ \mathrm{N}(0, \sigma^2_\epsilon)$, $(\beta\gamma)_{jk} \sim iid\ \mathrm{N}(0, \sigma^2_{\beta\gamma})$ and $\delta_{ijk} \sim iid\ \mathrm{N}(0, \sigma^2_\delta)$. All random effects are assumed to be independent, and for identifiability reasons we impose the parameter restrictions

$$\bar{\tau} = \bar{\gamma} = (\overline{\tau\gamma})_{i.} = (\overline{\tau\gamma})_{.k} = 0,$$

where

$$\bar{\tau} = \frac{1}{4}\sum_{i=1}^{4}\tau_i, \quad \bar{\gamma} = \frac{1}{6}\sum_{k=1}^{6}\gamma_k, \quad (\overline{\tau\gamma})_{i.} = \frac{1}{6}\sum_{k=1}^{6}(\tau\gamma)_{ik} \;\text{ and }\; (\overline{\tau\gamma})_{.k} = \frac{1}{4}\sum_{i=1}^{4}(\tau\gamma)_{ik}\ .$$

Block 1			Block 2		
Treatment	Variety	Yield (t/a)	Treatment	Variety	Yield (t/a)
pk	A	3.52	pk	A	3.61
pk	G	3.13	pk	G	4.43
pk	C	3.03	pk	C	4.15
pk	N	3.74	pk	N	4.65
pk	O	3.21	pk	O	4.72
pk	R	3.24	pk	R	4.21
Pk	A	4.54	Pk	A	4.53
Pk	G	3.88	Pk	G	4.80
Pk	C	4.10	Pk	C	4.74
Pk	N	3.65	Pk	N	5.45
Pk	O	4.06	Pk	O	4.54
Pk	R	4.42	Pk	R	4.94
pK	A	3.74	pK	A	3.40
pK	G	2.86	pK	G	3.81
pK	C	3.42	pK	C	4.14
pK	N	3.98	pK	N	4.78
pK	O	4.32	pK	O	3.78
pK	R	3.31	pK	R	3.93
PK	A	4.95	PK	A	5.39
PK	G	4.05	PK	G	5.25
PK	C	4.84	PK	C	5.15
PK	N	4.82	PK	N	6.13
PK	O	5.29	PK	O	6.00
PK	R	3.94	PK	R	5.80

Table 1: Yields of alfalfa (t/a).

Fertilizer	Variety						
treatment	A	G	C	N	O	R	$\bar{Y}_{i\cdot\cdot}$
pk	3.56	3.78	3.59	4.20	3.97	3.73	3.83
Pk	4.54	4.34	4.42	4.55	4.30	4.68	4.47
pK	3.57	3.34	3.78	4.38	4.05	3.62	3.79
PK	5.17	4.65	5.00	5.48	5.65	4.87	5.13
$\bar{Y}_{\cdot\cdot k}$	4.21	4.03	4.20	4.65	4.49	4.22	

Table 2: Mean yields for fertilizer treatment by variety combinations (t/a) and their averages.

Part I

1. Given the design of the alfalfa experiment, which effects – fertilizer treatment effects or variety effects – do you think will be more precisely estimated? Why?

Comparisons across varieties will have greater precision. These comparisons are carried out within fertilizer treatment levels, and therefore, each fertilizer treatment acts as its own control. There are more degrees of freedom associated with the factor assigned to the sub-plot units.

2. Under Model 1, give expressions for the covariances among observations within the same block with:

a. Different fertilizers, same variety: $\text{Cov}(Y_{ijk}, Y_{i'jk}) = \sigma^2_\beta + \sigma^2_{\beta\gamma}$

b. Same fertilizer, different varieties: $\text{Cov}(Y_{ijk}, Y_{ijk'}) = \sigma^2_\beta + \sigma^2_\epsilon$

c. Different fertilizers, different varieties: $\text{Cov}(Y_{ijk}, Y_{i'jk'}) = \sigma^2_\beta$

3. Give the degrees of freedom and compute the mean squares for the ANOVA table below:

Source	df	SS	MS
Block	1	6.961	6.961
Fertilizer	3	14.775	4.925
Fertilizer × Block	3	0.746	0.249
Variety	5	2.071	0.414
Fertilizer × Variety	15	1.526	0.102
Variety × Block	5	1.849	0.369
Fertilizer × Variety × Block	15	1.562	0.104

4. The expectations of five of the mean squares in the ANOVA table from question **3** are given below. Derive the expectations of the mean square for Variety and of the mean square for the Fertilizer × Variety interaction, in terms of the model parameters. We use Fert, Var instead of Fertilizer, Variety to shorten the notation.

$$\begin{aligned}
E[MS(\text{Block})] &= gf\sigma^2_\beta + f\sigma^2_{\beta\gamma} + g\sigma^2_\epsilon + \sigma^2_\delta \\
E[MS(\text{Fert})] &= \frac{bg}{f-1}\sum_i \tau_i^2 + g\sigma^2_\epsilon + \sigma^2_\delta \\
E[MS(\text{Fert} \times \text{Block})] &= g\sigma^2_\epsilon + \sigma^2_\delta \\
E[MS(\text{Var} \times \text{Block})] &= f\sigma^2_{\beta\gamma} + \sigma^2_\delta \\
E[MS(\text{Fert} \times \text{Var} \times \text{Block})] &= \sigma^2_\delta
\end{aligned}$$

The two expectations are as follow:

$$\begin{aligned} E[MS(\text{Var})] &= \frac{bf}{g-1}\sum_k \gamma_k^2 + f\sigma_{\beta\gamma}^2 + \sigma_\delta^2 \\ E[MS(\text{Fert} \times \text{Var})] &= \frac{b}{(g-1)(f-1)}\sum_{ik}(\tau\gamma)_{ik}^2 + \sigma_\delta^2. \end{aligned}$$

5. Test the hypothesis that there are no differences in mean yield due to fertilizer treatment. Use a 5% significance level ($\alpha = 0.05$) and state your conclusions.

We wish to test $H_0 : \tau_i = 0$ for all i versus the alternative hypothesis $H_a :$ at least one τ_i is different from 0. We construct an $F-$statistic using the MS for fertilizer treatment in the numerator and the whole-plot error MS as the denominator. Thus:

$$F_f^* = \frac{MS(\text{Fert})}{MS(\text{Fert} \times \text{Block})} = \frac{4.925}{0.249} = 19.811 \sim F_{3,3}.$$

The critical value for the test is 9.277. Since $F_f^* > 9.277$ we conclude that mean yields are significantly different across the four fertilizer treatments.

6. Test the two hypotheses:

a. of no differences between varieties, and

b. of no interaction between fertilizer treatments and varieties.

In both tests, use a 5% significance level and state your conclusions.

First we test $H_0 : \gamma_k = 0$ for all k against $H_a :$ at least one γ_k is different from 0. The test statistic is:

$$F_v^* = \frac{MS(\text{Var})}{MS(\text{Var} \times \text{Block})} = \frac{0.414}{0.369} = 1.122 \sim F_{5,5}.$$

The critical value for the test is 5.05. Since the statistic $F_v^* < 5.05$ we conclude that there are no significant mean yield differences that can be attributed to variety.

To test the null hypothesis of no interaction between fertilizer treatment and variety we use the $F-$ statistic:

$$F_{fv}^* = \frac{MS(\text{Fert} \times \text{Var})}{MS(\text{Fert} \times \text{Var} \times \text{Block})} = \frac{0.102}{0.104} = 0.977 \sim F_{15,15}.$$

The critical value for this test is 2.403. Since $F_{fv}^* < 2.403$ we conclude that there is no interaction of fertilizer treatment and variety on mean alfalfa yield.

7. Consider the comparison between the high and the low values of phosphorous (`P` versus `p`) when potassium is applied at its low level (`k`). Compute a point estimate of an appropriate contrast and obtain a 95% confidence interval for the corresponding true difference in model parameters. Interpret your results.

We use ζ to denote the contrast $\zeta = \sum_i a_i \tau_i$ for $a_i = (-1, 1, 0, 0)$. We estimate ζ as

$$\hat{\zeta} = \bar{Y}_{2..} - \bar{Y}_{1..}.$$

From Table 2 on page 3, the estimate is $\hat{\zeta} = 4.47 - 3.83 = 0.64$. The variance of the contrast is computed as

$$\text{Var}(\sum_i a_i \bar{Y}_{i..}) = \frac{\sigma_\delta^2 + g\sigma_\epsilon^2}{bg} \sum_i a_i^2.$$

From the expected mean squares given in question **4.** we find that the MS of `Fert`×`Block` is an estimator of $\sigma_\delta^2 + g\sigma_\epsilon^2$. Therefore,

$$\begin{aligned} \hat{\text{Var}}(\sum_i a_i \bar{Y}_{i..}) &= \frac{0.248}{12} \times 2 \\ &= 0.041. \end{aligned}$$

Therefore, a 95% confidence interval for ζ is computed as $\hat{\zeta} \pm 1.96\sqrt{\hat{\text{Var}}(\hat{\zeta})}$ and is equal to (0.242, 1.038). We conclude that when potassium is added to the field at the low level, increasing phosphorous will significantly increase mean alfalfa yield.

8. Suppose now that we wish to compare interactions $(\tau\gamma)_{ik}$ at different levels of the fertilizer treatment. The estimable contrast potentially of interest is

$$\eta = \sum_{ik} a_{ik} (\tau\gamma)_{ik},$$

which can be estimated unbiasedly by

$$\hat{\eta} = \sum_{ik} a_{ik} \bar{Y}_{ik} \ .$$

a. Derive an expression for the variance of $\hat{\eta}$.

The variance of $\hat{\eta}$ is

$$\text{Var}(\hat{\eta}) = \frac{\sigma_\delta^2}{b} \sum_{ik} a_{ik}^2 + \frac{\sigma_{\beta\gamma}^2}{b} \sum_k (\sum_i a_{ik})^2 + \frac{\sigma_\epsilon^2}{b} \sum_i (\sum_k a_{ik})^2.$$

b. Assume that $\sigma^2_{\beta\gamma} = 0$ and show that

$$\text{Var}(\hat{\eta}) = \frac{\sigma^2_\delta}{b}\sum_{ik}(a_{ik} - \bar{a}_i)^2 + \frac{\sigma^2_\delta + g\sigma^2_\epsilon}{bg}\sum_i(\sum_k a_{ik})^2.$$

We use the fact that

$$\sum_k a^2_{ik} = \sum_k(a_{ik} - \bar{a}_i + \bar{a}_i)^2 = \sum_k(a_{ik} - \bar{a}_i)^2 + g\bar{a}^2_i.$$

Then,

$$\begin{aligned}
\text{Var}(\hat{\eta}) &= \frac{\sigma^2_\delta}{b}\sum_i\left(\sum_k(a_{ik} - \bar{a}_i)^2 + g\bar{a}^2_i\right) + \frac{\sigma^2_\epsilon}{b}\sum_i(\sum_k a_{ik})^2 \\
&= \frac{\sigma^2_\delta}{b}\sum_{ik}(a_{ik} - \bar{a}_i)^2 + \frac{\sigma^2_\delta gb\sum_i(\sum_k a_{ik})^2 + \sigma^2_\epsilon}{b}\sum_i(\sum_k a_{ik})^2 \\
&= \frac{\sigma^2_\delta}{b}\sum_{ik}(a_{ik} - \bar{a}_i)^2 + \frac{\sigma^2_\delta + g\sigma^2_\epsilon}{bg}\sum_i(\sum_k a_{ik})^2.
\end{aligned}$$

c. Refer to the observed mean yields in Table 2 (page 3). We wish to compare the mean yield of variety `R` in fertilizer treatment `PK` with the mean yield of variety `A` in fertilizer treatment `pk`. For this comparison, we have $a_{11} = -1$, $a_{46} = 1$ and all other $a_{ik} = 0$. Calculate a point estimate for the contrast and compute its variance. Assume that $\sigma^2_{\beta\gamma} = 0$ as in question **8b.**

The point estimate of the contrast η is just the difference between $\bar{Y}_{46.}$ and $\bar{Y}_{11.}$. From Table 2 (page 3), we find that

$$\hat{\eta} = 4.87 - 3.56 = 1.31.$$

First, we compute

$$\begin{aligned}
\sum_{ik}(a_{ik} - \bar{a}_i)^2 &= (-1 + \frac{1}{6})^2 + 5(\frac{1}{6})^2 + 5(-\frac{1}{6})^2 + (1 - \frac{1}{6})^2 \\
&= 2 \times \left(\frac{5}{6}\right)^2 + 2 \times 5\left(\frac{1}{6}\right)^2 \\
&= \frac{5}{3}.
\end{aligned}$$

Similarly, we find that $\sum_i(\sum_k a_{ik})^2 = 2$.
To obtain an estimate of the variance, we note that an estimate of $\sigma^2_\delta + g\sigma^2_\epsilon$ is the $MS(\text{Fert} \times \text{Block})$, so then we get

$$\begin{aligned}
\hat{\text{Var}}(\hat{\eta}) &= \frac{5}{3}\frac{MS(\text{Fert} \times \text{Block} \times \text{Var})}{b} + 2\frac{MS(\text{Fert} \times \text{Block})}{bg} \\
&= \frac{5}{3}\frac{0.104}{2} + 2\frac{0.249}{12} \\
&= 0.9085.
\end{aligned}$$

d. Give the degrees of freedom for the Satterthwaite approximation to the sampling distribution of $\hat{\text{Var}}(\hat{\eta})$ using the expression in question **8b.** that assumes that $\sigma^2_{\beta\gamma} = 0$.

From part **b.** we note that $\hat{\text{Var}}(\hat{\eta})$ can be written as a linear combination of two independent χ^2 random variables. That is,

$$\hat{\text{Var}}(\hat{\eta}) = B_1 \times MS(\text{Fert} \times \text{Block} \times \text{Var}) + B_2 \times MS(\text{Fert} \times \text{Block}),$$

where B_1, B_2 are known constants:

$$\begin{aligned} B_1 &= \frac{1}{b}\sum_{ik}(a_{ik} - \bar{a}_i)^2 = \frac{5}{3 \times 2} \\ B_2 &= \frac{1}{bg}\sum_{i}(\sum_{k} a_{ik})^2 = \frac{1}{6}. \end{aligned}$$

A linear combination of two independent χ^2 distributions is not distributed as a χ^2 random variable (unless $B_1 = B_2 = 1$). Satterthwaite, however, proposed that its distribution could be approximated by a χ^2_ν distribution, where ν is estimated from the data. In our case:

$$\hat{\nu} = \frac{(B_1 \times MS(\text{Fert} \times \text{Block} \times \text{Var}) + B_2 \times MS(\text{Fert} \times \text{Block}))^2}{\frac{B_1^2}{(b-1)(g-1)(f-1)} MS(\text{Fert} \times \text{Block} \times \text{Var})^2 + \frac{B_2^2}{(b-1)(f-1)} MS(\text{Fert} \times \text{Block})^2}$$

From part **c.** we know that the numerator in the expression for $\hat{\nu}$ is 0.9085^2. The denominator equals

$$\left(\frac{5/6}{1/20}\right)^2 (0.104)^2 + \left(\frac{1/6}{1/3}\right)^2 (0.249)^2 = 0.00727.$$

The estimate of the degrees of freedom ν is therefore $\hat{\nu} = 114$, so we conclude that the estimated variance of the contrast between interaction means is approximately distributed as a χ^2_{114}.

Part II

Alfalfa can be harvested (cut) several times per year, during its growing season. In the experiment carried out at the Cornell Experiment Station,all sub-plots were harvested three times, at equally spaced two-month intervals. Therefore, the sub-plot yields Y_{ijk} shown in Table 1 are the sums of the yields obtained in the three cuttings in the sub-plots. We use t to index time, so that $t = 1, ..., 3$. Let Y_{ijkt} denote the yield observed at time t in a sub-plot in block j, under fertilizer treatment i and planted with variety k.

A model that can be used for the new response Y_{ijkt} is

$$\begin{aligned} Y_{ijkt} &= \mu + \tau_i + \beta_j + \epsilon_{ij} + \gamma_k + (\tau\gamma)_{ik} + (\beta\gamma)_{jk} + \delta_{ijk} \\ &\quad + C_k + (\tau C)_{it} + (\gamma C)_{kt} + (\tau\gamma C)_{ikt} + (\beta C)_{jt} + \theta_{ijt} + \phi_{ikt} + \psi_{ijkt}, \end{aligned} \quad \text{(Model 2)}$$

$i = 1, ..., 4;\ j = 1, 2;\ k = 1, ..., 6;\ t = 1, ..., 3$ where

$$\begin{aligned} \mu &= \text{a fixed overall mean effect} \\ \tau_i &= \text{a fixed effect of fertilizer treatment } i \\ \beta_j &= \text{a random block effect for block } j \\ \epsilon_{ij} &= \text{a random whole plot error, for treatment } i \text{ in block } j \\ \gamma_k &= \text{a fixed variety effect for variety } k \\ (\tau\gamma)_{ik} &= \text{a fixed interaction effect of fertilizer treatment } i \text{ by variety } k \\ (\beta\gamma)_{jk} &= \text{a random interaction effect of block } j \text{ by variety } k \\ \delta_{ijk} &= \text{a random sub-plot error} \\ C_t &= \text{a fixed cutting effect for cutting } t \\ (\beta C)_{jt} &= \text{a random interaction between block } j \text{ and cutting } t \\ \theta_{ijt} &= (\tau\beta C)_{ijt},\ \text{a random interaction between treatment } i, \text{ block } j, \text{ and cutting } t \\ \phi_{jkt} &= (\beta\gamma C)_{jkt},\ \text{a random interaction between block } j, \text{ variety } k, \text{ and cutting } t \\ \psi_{ijkt} &= (\tau\beta\gamma C)_{ijkt},\ \text{a random error} \end{aligned}$$

As before, we assume that $\beta_j \sim iid\ \mathrm{N}(0, \sigma^2_\beta)$, $\epsilon_{ij} \sim iid\ \mathrm{N}(0, \sigma^2_\epsilon)$, $(\beta\gamma)_{jk} \sim iid\ \mathrm{N}(0, \sigma^2_{\beta\gamma})$ and $\delta_{ijk} \sim iid\ \mathrm{N}(0, \sigma^2_\delta)$. Further, we now also assume that $(\beta C)_{jt} \sim iid\ \mathrm{N}(0, \sigma^2_{\beta C})$, $\theta_{ijt} \sim iid\ \mathrm{N}(0, \sigma^2_\theta)$, $\phi_{jkt} \sim iid\ \mathrm{N}(0, \sigma^2_\phi)$ and $\psi_{ijkt} \sim iid\ \mathrm{N}(0, \sigma^2_\psi)$. All random effects are assumed to be independent.

9. What is the covariance between yields for two different cuttings in the same sub-plot based on Model 2?

The covariance between two yields from the same sub-plot but different cuttings is

$$\mathrm{Cov}(Y_{ijkt}, Y_{ijkt'}) = \sigma^2_\beta + \sigma^2_\epsilon + \sigma^2_{\beta\gamma} + \sigma^2_\delta.$$

Because of the assumption of independence, every variance term indexed by t drops off the covariance.

10. There are two degrees of freedom associated with the effects of cutting. Construct two orthogonal one-degree of freedom contrasts to test for non-zero linear and quadratic effects, respectively, of cutting on yield for fertilizer treatment `pk`, in block 1, for variety `A`, and derive the variance of each of the contrasts under Model 2.

The two orthogonal contrasts are:

$$\begin{aligned} \text{Linear trend in yields} &= \hat{\omega}_1 = Y_{1111} - Y_{1113} \\ \text{Quadratic trend in yields} &= \hat{\omega}_2 = Y_{1111} - 2Y_{1112} + Y_{1113}, \end{aligned}$$

where the indices $i = 1; j = 1; k = 1$ refer to fertilizer treatment `pk`, block 1 and variety `A`, respectively.

The variance of $\hat{\omega}_1$ is:

$$\begin{aligned} \text{Var}(\hat{\omega}_1) &= \text{Var}(Y_{1111}) + \text{Var}(Y_{1113}) - 2\text{Cov}(Y_{1111}, Y_{1113}) \\ &= 2\sigma^2_\beta + 2\sigma^2_\epsilon + 2\sigma^2_\delta + 2\sigma^2_{\beta C} + 2\sigma^2_\theta \\ &\quad +2\sigma^2_\phi + 2\sigma^2_\psi - 2(\sigma^2_\beta + \sigma^2_\epsilon + \sigma^2_{\beta\gamma} + \sigma^2_\delta) \\ &= 2(\sigma^2_{\beta C} + \sigma^2_\theta + \sigma^2_\phi + \sigma^2_\psi). \end{aligned}$$

The variance of $\hat{\omega}_2$ is:

$$\begin{aligned} \text{Var}(\hat{\omega}_2) &= \text{Var}(Y_{1111}) + 4\text{Var}(Y_{1112}) + \text{Var}(Y_{1113}) - 4\text{Cov}(Y_{1111}, Y_{1112}) + \\ &\quad \text{Cov}(Y_{1111}, Y_{1113}) - 4\text{Cov}(Y_{1112}, Y_{1113}) \\ &= 6\text{Var}(Y_{111t}) - 7\text{Cov}(Y_{111t}, Y_{111t'}) \\ &= 6(\sigma^2_{\beta C} + \sigma^2_\theta + \sigma^2_\phi + \sigma^2_\psi) - (\sigma^2_\beta + \sigma^2_\epsilon + \sigma^2_{\beta\gamma} + \sigma^2_\delta). \end{aligned}$$

11. Suppose that you find that the effects of:

a. fertilizer treatments,

b. interactions between fertilizer treatment and the linear trend in cutting

are both significantly different from zero. In your own words, interpret those two results for someone who is not a statistician.

A significant effect of fertilizer treatment indicates that, when we average over all cuttings, varieties and blocks, any differences we observe in alfalfa yield can be attributable to differences in the fertilizer treatment.

A significant interaction between fertilizer treatment and the liner trend in cutting suggests that - averaging over varieties and blocks – the linear change in yield of alfalfa between cuttings is different under different fertilizer treatments.

Part I

Consider the general linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ is an $n \times 1$ response vector, $\boldsymbol{X}$ is an $n \times p$ model matrix of known constants, $\boldsymbol{X}$ is not necessarily of full rank, $\boldsymbol{\beta}$ is an unknown parameter vector in $\mathbb{R}^p$, and $\boldsymbol{\varepsilon}$ is a random vector with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$.

1. State the definition of a linearly estimable function $\boldsymbol{c}'\boldsymbol{\beta}$.

2. Give the ordinary least squares estimator of a linearly estimable function $\boldsymbol{c}'\boldsymbol{\beta}$.

3. Give the normal equations.

4. Do the normal equations always have a solution? Prove that your answer is correct.

5. In the general linear model stated above, we have assumed that $\boldsymbol{\varepsilon}$ is a random vector with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$. What additional assumptions (if any) about $\boldsymbol{\varepsilon}$ guarantee that the ordinary least squares estimator of a linearly estimable function $\boldsymbol{c}'\boldsymbol{\beta}$ has minimum variance among all linear unbiased estimators of $\boldsymbol{c}'\boldsymbol{\beta}$?

6. Consider a special case of the general linear model where

$$\boldsymbol{\beta} = (\mu, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \gamma_1, \gamma_2, \gamma_3, \gamma_4)'$$

and

$$E(y_{ij}) = \mu + \alpha_i + \gamma_j \text{ for } i = 1, 2, 3, 4,\ j = 1, 2, 3, 4.$$

Suppose the cost of observing y_{ij} is $i(i-j)^2$ dollars for all $i = 1, 2, 3, 4$ and $j = 1, 2, 3, 4$.

a) State which of the responses $\{y_{ij} : i = 1, 2, 3, 4,\ j = 1, 2, 3, 4\}$ you would pay to observe if the goal is to spend as little as possible to make $\alpha_i - \alpha_{i^*}$ estimable for all $i \neq i^*$ and $\gamma_j - \gamma_{j^*}$ estimable for all $j \neq j^*$.

b) Order the responses you chose to observe in 6(a) in a response vector $\boldsymbol{y}$ and provide the corresponding model matrix $\boldsymbol{X}$ so that $E(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}$.

c) Prove that $\alpha_i - \alpha_{i^*}$ is estimable for all $i \neq i^*$ and $\gamma_j - \gamma_{j^*}$ is estimable for all $j \neq j^*$ from the responses you chose in 6(a) to observe.

Part II

Researchers were interested in studying the effects of 6 treatments on the grain yield of corn plots. A total of 18 plots were used in the experiment. The 18 plots were arranged in a 6 $\times$3 layout with 6 rows of 3 plots each as depicted in Figure 1. A randomized complete block design was used to assign treatments to plots, where the 6 plots in rows 1 and 2 were considered to be one block, the 6 plots in rows 3 and 4 were considered to be a second block, and the 6 plots in rows 5 and 6 were considered to be a third block. The plots in Figure 1 are labeled with their randomly assigned treatments.

Figure 1. Layout of plots labeled with their randomly assigned treatments.

Row				**Block**
1	6	1	4	**1**
2	2	3	5	**1**
3	1	3	4	**2**
4	5	6	2	**2**
5	1	2	6	**3**
6	5	4	3	**3**

For $i = 1, \ldots, 6$ and $j = 1, 2, 3$, let y_{ij} denote the grain yield for the plot that received treatment i in block j.

7. For $i = 1, \ldots, 6$ and $j = 1, 2, 3$, suppose

$$y_{ij} = \mu + \tau_i + b_j + e_{ij}, \tag{1}$$

where μ, $\tau_1, \ldots, \tau_6$ are unknown real-valued parameters and $b_1, b_2, b_3 \overset{iid}{\sim} N(0, \sigma_b^2)$ independent of $e_{11}, e_{21}, \ldots, e_{63} \overset{iid}{\sim} N(0, \sigma_e^2)$.

a) Find the mean and variance of $\bar{y}_{1\cdot} - \bar{y}_{2\cdot}$.

b) Define MSB (mean square for blocks) and MSE (error mean square) and provide an unbiased estimator of σ_b^2 based on these.

c) Prove that the estimator you provided in question 7(b) is unbiased for σ_b^2.

8. Again assume that Model (1) holds. Use the following partial `R` input and output to answer questions 8(a) through 8(e). If it is not possible to answer a question from the information provided, explain what additional information is needed to answer the question. (The `R` object `y` is an appropriately ordered 18-dimensional vector containing the grain yield data for the 18 plots.)

```
> treatment=factor(rep(1:6,3))
> block=factor(rep(1:3,each=6))
> library(nlme)
> o=lme(y~treatment,random=~1|block)
> summary(o)

Fixed effects: y ~ treatment
             Value StdErr DF t-value p-value
(Intercept) 15.733 0.9480 10  16.596  0.0000
treatment2  -1.067 1.2906 10  -0.826  0.4278
treatment3   2.000 1.2906 10   1.550  0.1523
treatment4   1.100 1.2906 10   0.852  0.4140
treatment5  -3.367 1.2906 10  -2.609  0.0261
treatment6  -0.333 1.2906 10  -0.258  0.8014
```

a) Provide the numerical value of $\bar{y}_{1\cdot} - \bar{y}_{2\cdot}$.

b) Provide the numerical value of a test statistic for testing $H_0 : \tau_1 = \tau_2$.

c) Precisely state the distribution of the test statistic (whether H_0 is true or not) whose observed value you reported in question 8(b).

d) Provide a p-value for the test of $H_0 : \tau_1 = \tau_2$ versus $H_a : \tau_1 \neq \tau_2$.

e) Provide the numerical value of the test statistic for testing $H_0 : \tau_3 = \tau_4$.

9. Now suppose that instead of Model (1), the true model is

$$y_{ij} = \mu + \tau_i + r_{a(ij)} + e_{ij}, \;\; (i = 1, \ldots, 6, \; j = 1, 2, 3) \tag{2}$$

where μ, $\tau_1, \ldots, \tau_6$ are unknown real-valued parameters, $a(ij)$ is the row number of the plot that received treatment i in block j, and $r_1, \ldots, r_6 \overset{iid}{\sim} N(0, \sigma_r^2)$ independent of $e_{11}, e_{21}, \ldots, e_{63} \overset{iid}{\sim} N(0, \sigma_e^2)$.

a) Derive an expression for the variance of $\bar{y}_{1\cdot} - \bar{y}_{2\cdot}$ in terms of σ_r^2 and σ_e^2.

b) Suppose Model (2) holds but that you are able to observe only the subset of the responses involving treatments 1 and 2; that is, you observe only

$$\boldsymbol{y} = (y_{11}, y_{21}, y_{12}, y_{22}, y_{13}, y_{23})'.$$

Given that $\sigma_r^2/\sigma_e^2 = 2$, derive an expression for the BLUE of $\tau_1 - \tau_2$.

Part I

1. $\boldsymbol{c}'\boldsymbol{\beta}$ is linearly estimable if and only if there exists a linear unbiased estimator of $\boldsymbol{c}'\boldsymbol{\beta}$.

2. $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ is the ordinary least squares estimator of $\boldsymbol{c}'\boldsymbol{\beta}$ if and only if $\hat{\boldsymbol{\beta}}$ satisfies $\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$.

3. The normal equations are $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}'\boldsymbol{y}$, where $\boldsymbol{b}$ is a vector to be solved for.

4. The normal equations always have a solution. We showed in our course notes that

$$\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}' = \boldsymbol{X}'.$$

Thus,

$$\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{y}$$

so that $\boldsymbol{b} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y}$ is always a solution to the normal equations.

5. The Gauss-Markov Theorem requires $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$ for some $\sigma^2 > 0$. Note that normality is not required.

6. **a**) The responses y_{11}, y_{22}, y_{33}, and y_{44} can be observed at no cost. If we also observe the next three least expensive responses (y_{12}, y_{23}, y_{34}), $\alpha_i - \alpha_{i^*}$ is estimable for all $i \neq i^*$ and $\gamma_j - \gamma_{j^*}$ is estimable for all $j \neq j^*$. This will be shown in 6(c).

b)

$$\boldsymbol{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{22} \\ y_{23} \\ y_{33} \\ y_{34} \\ y_{44} \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

c) It is simple to verify that

$$E(y_{12} - y_{22}) = \alpha_1 - \alpha_2,\ E(y_{23} - y_{33}) = \alpha_2 - \alpha_3,\ E(y_{34} - y_{44}) = \alpha_3 - \alpha_4,$$

$$E(y_{12} - y_{22} + y_{23} - y_{33}) = \alpha_1 - \alpha_3,\ E(y_{23} - y_{33} + y_{34} - y_{44}) = \alpha_2 - \alpha_4,$$

$$E(y_{12} - y_{22} + y_{23} - y_{33} + y_{34} - y_{44}) = \alpha_1 - \alpha_4$$

and

$$E(y_{11} - y_{12}) = \gamma_1 - \gamma_2,\ E(y_{22} - y_{23}) = \gamma_2 - \gamma_3,\ E(y_{33} - y_{34}) = \gamma_3 - \gamma_4,$$

$$E(y_{11} - y_{12} + y_{22} - y_{23}) = \gamma_1 - \gamma_3,\ E(y_{22} - y_{23} + y_{33} - y_{34}) = \gamma_2 - \gamma_4,$$

$$E(y_{11} - y_{12} + y_{22} - y_{23} + y_{33} - y_{34}) = \gamma_1 - \gamma_4.$$

Thus, $\alpha_i - \alpha_{i^*}$ is estimable for all $i \neq i^*$ and $\gamma_j - \gamma_{j^*}$ is estimable for all $j \neq j^*$.

7. **a**) Note that

$$\bar{y}_{1\cdot} - \bar{y}_{2\cdot} = \tau_1 - \tau_2 + \bar{e}_{1\cdot} - \bar{e}_{2\cdot}.$$

Thus,

$$E(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) = \tau_1 - \tau_2$$

and

$$\mathrm{Var}(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) = 2\sigma_e^2/3.$$

b) An unbiased estimator of σ_b^2 is $(\mathrm{MSB} - \mathrm{MSE})/6$, where

$$\mathrm{MSB} = \frac{6\sum_{j=1}^{3}(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2}{3-1}$$

and

$$\mathrm{MSE} = \frac{\sum_{i=1}^{6}\sum_{j=1}^{3}(y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2}{(6-1)(3-1)}.$$

c) First, note that

$$y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot} = e_{ij} - \bar{e}_{i\cdot} - \bar{e}_{\cdot j} + \bar{e}_{\cdot\cdot}.$$

Next, let $u_{ij} = e_{ij} - \bar{e}_{\cdot j}$ and note that, for any $i = 1, \ldots, 6$,

- u_{i1}, u_{i2}, u_{i3} are independent and identically distributed,
-

$$\begin{aligned}\mathrm{Var}(u_{ij}) &= \mathrm{Var}(e_{ij} - \bar{e}_{\cdot j}) = \mathrm{Var}(e_{ij}) + \mathrm{Var}(\bar{e}_{\cdot j}) - 2\mathrm{Cov}(e_{ij}, \bar{e}_{\cdot j}) \\ &= \sigma_e^2 + \sigma_e^2/6 - 2\sigma_e^2/6 = 5\sigma_e^2/6, \text{ and}\end{aligned}$$

- $\bar{u}_{i\cdot} = \bar{e}_{i\cdot} - \bar{e}_{\cdot\cdot}$ so that $u_{ij} - \bar{u}_{i\cdot} = e_{ij} - \bar{e}_{i\cdot} - \bar{e}_{\cdot j} + \bar{e}_{\cdot\cdot}$.

From the above points, it follows that, for any $i = 1, \ldots, 6$,

$$E\left(\sum_{j=1}^{3}(y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot})^2\right) = E\left(\sum_{j=1}^{3}(u_{ij} - \bar{u}_{i\cdot})^2\right) = (3-1)5\sigma_e^2/6.$$

Thus,

$$E(\mathrm{MSE}) = \frac{6(3-1)5\sigma_e^2/6}{(6-1)(3-1)} = \sigma_e^2.$$

Now note that

$$\begin{aligned}\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot} &= b_j + \bar{e}_{\cdot j} - \bar{b}_{\cdot} - \bar{e}_{\cdot\cdot} \\ &= v_j - \bar{v}_{\cdot},\end{aligned}$$

where $v_j = b_j + \bar{e}_{\cdot j}$. The random variables v_1, v_2, v_3 are iid with variance $\sigma_b^2 + \sigma_e^2/6$. Thus,

$$\begin{aligned} E(\text{MSB}) &= E\left(\frac{6\sum_{j=1}^3(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2}{3-1}\right) = \frac{6E\left(\sum_{j=1}^3(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})^2\right)}{3-1} \\ &= \frac{6E\left(\sum_{j=1}^3(v_j - \bar{v}_{\cdot})^2\right)}{3-1} = \frac{6(3-1)(\sigma_b^2 + \sigma_e^2/6)}{3-1} \\ &= 6\sigma_b^2 + \sigma_e^2. \end{aligned}$$

It follows that

$$E\left(\frac{\text{MSB} - \text{MSE}}{6}\right) = \sigma_b^2.$$

8. **a**) $15.733 - (15.733 - 1.067) = 1.067$

b) -0.826

c) The test statistic has a non-central t distribution with 10 degrees of freedom and non-centrality parameter

$$\frac{\tau_1 - \tau_2}{\sqrt{2\sigma_e^2/3}}.$$

d) 0.4278

e) Because this is a randomized complete block design with one experimental unit per treatment in each block, the standard error is the same for any difference of treatment means. Thus, the statistic for testing $H_0 : \tau_3 = \tau_4$ is

$$t = (2 - 1.1)/1.2906 \approx 0.697.$$

9. **a**)

$$\begin{aligned} \bar{y}_{1\cdot} - \bar{y}_{2\cdot} &= \tau_1 - \tau_2 + \bar{e}_{1\cdot} - \bar{e}_{2\cdot} + (r_1 + r_3 + r_5)/3 - (r_2 + r_4 + r_5)/3 \\ &= \tau_1 - \tau_2 + \bar{e}_{1\cdot} - \bar{e}_{2\cdot} + (r_1 - r_2 + r_3 - r_4)/3. \end{aligned}$$

Thus, $\text{Var}(\bar{y}_{1\cdot} - \bar{y}_{2\cdot}) = 2\sigma_e^2/3 + 4\sigma_r^2/9$.

b) Let $a = (y_{11} + y_{12} - y_{21} - y_{22})/2$ and $b = y_{13} - y_{23}$. It is straightforward to show that a is the BLUE of $\tau_1 - \tau_2$ using the first four observations and that b is the BLUE of $\tau_1 - \tau_2$ using the last two observations. The estimators a and b are clearly independent of each other. Also,

$$\text{Var}(a) = 2(\sigma_r^2 + \sigma_e^2)/2 = \sigma_r^2 + \sigma_e^2 = \sigma_e^2(\sigma_r^2/\sigma_e^2 + 1) = 3\sigma_e^2$$

and

$$\mathrm{Var}(b) = 2\sigma_e^2.$$

Based on the optimal (inverse-variance) weighting, the BLUE of $\tau_1 - \tau_2$ using all the data in $\boldsymbol{y}$ is

$$\begin{aligned}
&\frac{1/(3\sigma_e^2)}{1/(3\sigma_e^2) + 1/(2\sigma_e^2)}a + \frac{1/(2\sigma_e^2)}{1/(3\sigma_e^2) + 1/(2\sigma_e^2)}b \\
&= \frac{2}{5}a + \frac{3}{5}b \\
&= \frac{1}{5}y_{11} + \frac{1}{5}y_{12} - \frac{1}{5}y_{21} - \frac{1}{5}y_{22} + \frac{3}{5}y_{13} - \frac{3}{5}y_{23}.
\end{aligned}$$

This same answer can be obtained as follows. Note that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{I}_{2\times 2} \\ \boldsymbol{I}_{2\times 2} \\ \boldsymbol{I}_{2\times 2} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu + \tau_1 \\ \mu + \tau_2 \end{bmatrix}, \quad \text{and } \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma_e^2 \boldsymbol{V}) \text{ with}$$

$$\begin{aligned}
\boldsymbol{V} &= \begin{bmatrix}
\sigma_r^2/\sigma_e^2 + 1 & 0 & 0 & 0 & 0 & 0 \\
0 & \sigma_r^2/\sigma_e^2 + 1 & 0 & 0 & 0 & 0 \\
0 & 0 & \sigma_r^2/\sigma_e^2 + 1 & 0 & 0 & 0 \\
0 & 0 & 0 & \sigma_r^2/\sigma_e^2 + 1 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_r^2/\sigma_e^2 + 1 & \sigma_r^2/\sigma_e^2 \\
0 & 0 & 0 & 0 & \sigma_r^2/\sigma_e^2 & \sigma_r^2/\sigma_e^2 + 1
\end{bmatrix} \\
&= \begin{bmatrix}
3 & 0 & 0 & 0 & 0 & 0 \\
0 & 3 & 0 & 0 & 0 & 0 \\
0 & 0 & 3 & 0 & 0 & 0 \\
0 & 0 & 0 & 3 & 0 & 0 \\
0 & 0 & 0 & 0 & 3 & 2 \\
0 & 0 & 0 & 0 & 2 & 3
\end{bmatrix}.
\end{aligned}$$

Thus, the BLUE of $\tau_1 - \tau_2$ is $[1, -1](\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}$, which simplifies to

$$\frac{1}{5}y_{11} + \frac{1}{5}y_{12} - \frac{1}{5}y_{21} - \frac{1}{5}y_{22} + \frac{3}{5}y_{13} - \frac{3}{5}y_{23}.$$

Ecologists and wildlife managers want to know the relation between the use of areas by animals and the characteristics of those areas (i.e., the type of habitat available). This is particularly true in terms of the status of predators that have historically been present in a region and the types of habitats currently existing in those regions. For example, the bobcat (*Lynx rufus*) has historically been a top predator in the Ozarks region of Missouri and Arkansas. Bobcats are larger than house cats but smaller than cougars, typically about $0.75 - 1.00$ meters ($30 - 50$ inches) in length and about $18 - 22$ kilograms ($40 - 50$ pounds). While still present in the Ozarks, bobcats are certainly more rare than they used to be, and a question of interest is what habitat characteristics are crucial for the survival of these animals. There is some evidence that the proportion of an area in forest is important, but there is some question about how much forest is optimal for bobcats since they prey largely on rabbits and deer, both of which prefer more open areas.

Bobcats are territorial but the sizes of their home ranges are quite variable ($0.05 - 300$ km^2 or $0.02 - 120$ mi^2), and may change between summer and winter as well. This makes determination of whether bobcats are using a particular area difficult, as does the fact that this species is highly secretive and difficult to observe. A reasonably effective sampling method involves preparation of "track monitoring sites" as circular areas of fine sand or sifted dirt about 1 meter in radius. Such monitoring sites are "baited" (usually with a scent post treated with bobcat urine) in the evening and revisited in the morning and the presence or absence of bobcat tracks noted. Bobcats mark their territories with urine and feces, so the scent of a "foreign" bobcat should attract the attention of any "resident" bobcat. This method does suffer from the possibility that a monitoring site can be destroyed by weather or various nocturnal animals other than bobcats, so setting out m monitoring sites may result in only $w < m$ that provide useful data.

The data for this question come from a large effort to determine the use of 15 different areas in the Ozarks of Missouri and Arkansas by bobcats. A number of track monitoring sites (from 1 to 5 sites provided useful data) were prepared in each area on different occasions. Most areas (13 of the 15) were sampled on 4 occasions, but one was sampled on 3 occasions and one on only 2 occasions. The proportion of each of the 15 areas that would be considered to be forest was also recorded. The end result is a total of 57 observations that each consist of values for sampling occasion, proportion forest (one value for each area), the number of track monitoring sites providing data, and presence/absence

of bobcats in the area (presence at any site implies presence in the area).

The primary objective in the analysis of these data is to determine the relation, if any, between the proportion of forest in an area and the probability the area is used by bobcats. Two fundamental assumptions were that (1) if bobcats were using an area they were using it for the entire study period, and (2) bobcats are not falsely detected. This second assumption implies that, if bobcats are not using an area, there is no chance of seeing their tracks at a monitoring site. (This is not unreasonable as long as the data are recorded by trained biologists, since there are no other animals with tracks similar to those of a bobcat.)

Notation

The notation to be used throughout this question is as follows. Let areas be indexed by $i = 1, \ldots, N$ ($N = 15$ in this study) and sampling occasions within an area by $j = 1, \ldots, n_i$. Define random variables connected with the observation of bobcat presence in an area at a given sampling occasion as

$$Y_{i,j} = \begin{cases} 1 & \text{if bobcat tracks are detected in area } i \text{ on occasion } j \\ 0 & \text{otherwise.} \end{cases}$$

Note that the presence/absence of bobcats is considered over monitoring sites for a sampling occasion. That is, $Y_{i,j} = 1$ if any of the monitoring sites in area i on occasion j detected bobcat activity. Assume that the $Y_{i,j}$ are independent. Also, let x_i denote the proportion of area i considered to be forest, and let $w_{i,j}$ denote the number of monitoring sites used in area i on sampling occasion j.

Part I: A Simple Model

A simple model for this problem is a basic generalized linear model with a binomial random component, logit link, and covariate of proportion forest. The binomial responses would be constructed as the sum of the binary $Y_{i,j}$ over sampling occasions $j = 1, \ldots, n_i$, for each area, written in the usual manner for generalized linear models as proportions. That is, define response variables

$$R_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{i,j} = \frac{1}{n_i} S_i, \tag{1}$$

and, for $0 < p_i < 1$ and $i = 1, \ldots, N$, write the probability mass function for the random model component as,

$$f_i(r_i|p_i) = \frac{n_i!}{(n_i - n_i r_i)!\,(n_i r_i)!}\, p_i^{n_i r_i}\,(1-p_i)^{n_i - n_i r_i};\;\; r_i = (0/n_i), (1/n_i), \ldots, (n_i/n_i). \quad (2)$$

Define the systematic model component through the link function $g(\cdot)$ and linear term η_i as

$$g(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i = \eta_i, \quad (3)$$

where, as indicated in the section on Notation, x_i is the proportion of area i in forest. A scatterplot of the observed proportions r_i against proportions of forest is presented in Figure 1 on page 6.

ANSWER QUESTIONS 1, 2, 3 and 4 NOW

(Questions begin on page 10)

Plots of the fitted regression function and standardized deviance residuals are presented in Figure 2. Neither of these plots are visually pleasing. The lack of an increase in the estimated probability that an area is used by bobcats for higher values of forest is due to the fact that zero values exist across the range of the covariate (proportion forest).

Part II: Modeling Detection

Recall from the introduction to this question that bobcats are secretive animals, and that a fundamental assumption of the study was that bobcats were not falsely detected. On the other hand, it is entirely possible that bobcats use a given area but are not detected by the sampling protocol used in the study.

The original data may provide some information about the issue of detection because of the repeated sampling of the $N = 15$ areas included in the study. A model that attempts to account for imperfect detection is as follows. Define the latent random variables, for $i = 1, \ldots, N$,

$$Z_i = \begin{cases} 1 & \text{if bobcats are using area } i \\ 0 & \text{otherwise.} \end{cases}$$

and, assuming independence of the Z_i's, assign these binary variables distributions with parameters $0 < \psi_i < 1$,

$$h(z_i|\psi_i) = \psi_i^{z_i}(1-\psi_i)^{1-z_i};\quad z_i = 0, 1. \quad (4)$$

To incorporate the influence of forest on use of an area by bobcats, we might further model the ψ_i as,

$$m(\psi_i) = \log\left(\frac{\psi_i}{1-\psi_i}\right) = \gamma_0 + \gamma_1 x_i; \quad i = 1, \ldots, N, \tag{5}$$

where, as before, x_i is the proportion of area i in forest.

The observable random variables $Y_{i,j}$ continue to be defined as in the Notation section, the variables S_i continue to represent the aggregation of the $Y_{i,j}$ as given in expression (1), and the S_i are now assigned conditional probability mass functions, for $0 < p < 1$ and $i = 1, \ldots, N$,

$$f_i(s_i|z_i, p) = \frac{n_i!}{s_i!\,(n_i - s_i)!}\,(pz_i)^{s_i}\,(1 - pz_i)^{n_i - s_i}; \quad s_i = 0, 1, \ldots, n_i. \tag{6}$$

Note that we could write (6) in terms of the proportions $R_i = S_i/n_i$ as was done in expression (2), but we are no longer attempting to write these probability mass functions in the form of exponential dispersion families for direct application of basic generalized linear models, so it is just as convenient to leave them in more standard form. Fitted functions of response proportion to proportion of forest based on maximum likelihood estimation applied to the model of expressions (4), (5) and (6) is presented in Figure 3.

ANSWER QUESTIONS 5, 6, 7, 8 AND 9 NOW

(Questions begin on page 10)

Part III: Allowing Variable Detection

The model of Part II takes the probability of detecting the use of an area by bobcats to be constant for all sampling occasions within each area. But we know that the sampling effort was not constant for all areas and times. As indicated in the introductory portion of this question, the number of useable track monitoring sites varied from 1 to 5 for different areas and sampling occasions. Let $w_{i,j}$ be the number of useable monitoring sites in area i at sampling occasion j, $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. A model more detailed than those of either Part I or Part II would then treat the $Y_{i,j}$ directly rather than the aggregated values S_i. We continue to assume that the total study duration was short enough so that bobcat use of an area did not change over the course of the study. Thus, the Z_i are still meaningful, as is the specification of the distribution of the $Y_{i,j}$ as conditional on the value of the Z_i. With some repetition, but to be complete for this Part of the question, the model under consideration is formulated as follows.

Random Variables

$$Z_i = \begin{cases} 1 & \text{if bobcats are using area } i \\ 0 & \text{otherwise.} \end{cases},$$

$$Y_{i,j} = \begin{cases} 1 & \text{if bobcat tracks are detected in area } i \text{ at time } j \\ 0 & \text{otherwise.} \end{cases}$$

Model for Bobcat Use (Z_i)

$$h(z_i|\psi_i) = \psi_i^{z_i}(1-\psi_i)^{1-z_i}; \quad z_i = 0, 1. \tag{7}$$

$$m(\psi_i) = \log\left(\frac{\psi_i}{1-\psi_i}\right) = \gamma_0 + \gamma_1 x_i, \tag{8}$$

where x_i is the proportion of area i in forest.

Model for Detection of Use ($Y_{i,j}$)

$$f(y_{i,j}|z_i, p_{i,j}) = (z_i p_{i,j})^{y_{i,j}}(1 - z_i p_{i,j})^{1-y_{i,j}}; \quad y_{i,j} = 0, 1, \tag{9}$$

$$q(p_{i,j}) = \log\left(\frac{p_{i,j}}{1-p_{i,j}}\right) = \beta_0 + \beta_1 w_{i,j}, \tag{10}$$

where $w_{i,j}$ is the number of track monitoring stations in area i on sampling occasion j. Figure 4 displays estimated probabilities from Part II and from Part III. Maximum likelihood produces estimates of the ψ_i from the model of Part III as shown by the solid curve in Figure 4. On this graph, circles correspond to the observed $y_{i,j}$. Symbols plotted as "X" correspond to the $R_i = S_i/n_i$ used in the model of Part II, and the dashed curve gives estimates of the ψ_i from that model (and this is the same as the solid curve in Figure 3).

ANSWER QUESTIONS 10, 11 AND 12 NOW

(Questions begin on page 10)

Figures

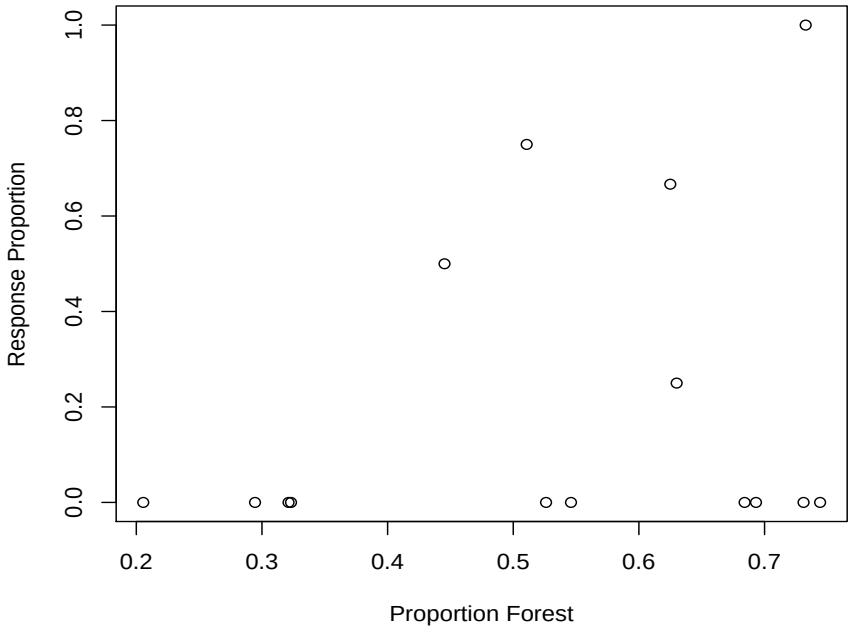

Figure 1: Scatterplot of proportion of sampling occasions when bobcat activity was detected versus proportion of area in forest.

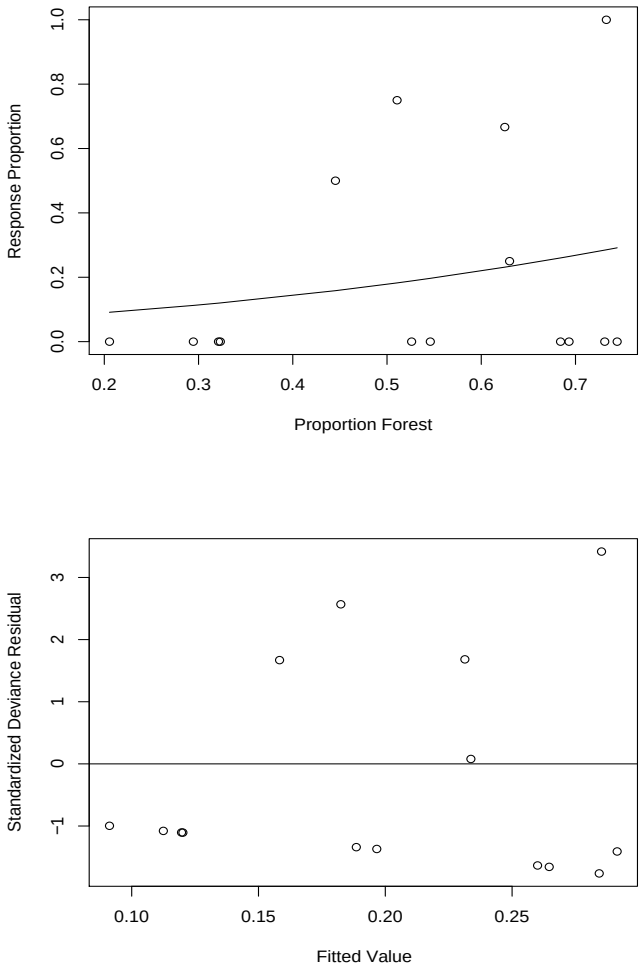

Figure 2: Fitted regression function (upper) and standardized deviance residuals (lower) for the model of expressions (2) and (3) fit to the data from Figure 1.

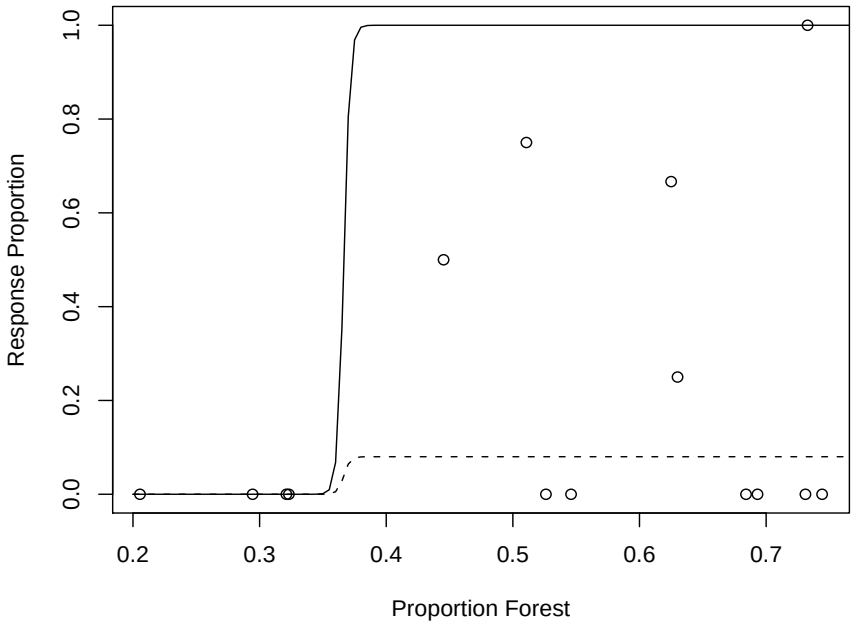

Figure 3: Fitted regression functions for the model of expressions (4), (5) and (6). The solid curve gives estimated values of ψ_i, while dashed curve gives estimated values of $p\psi_i$.

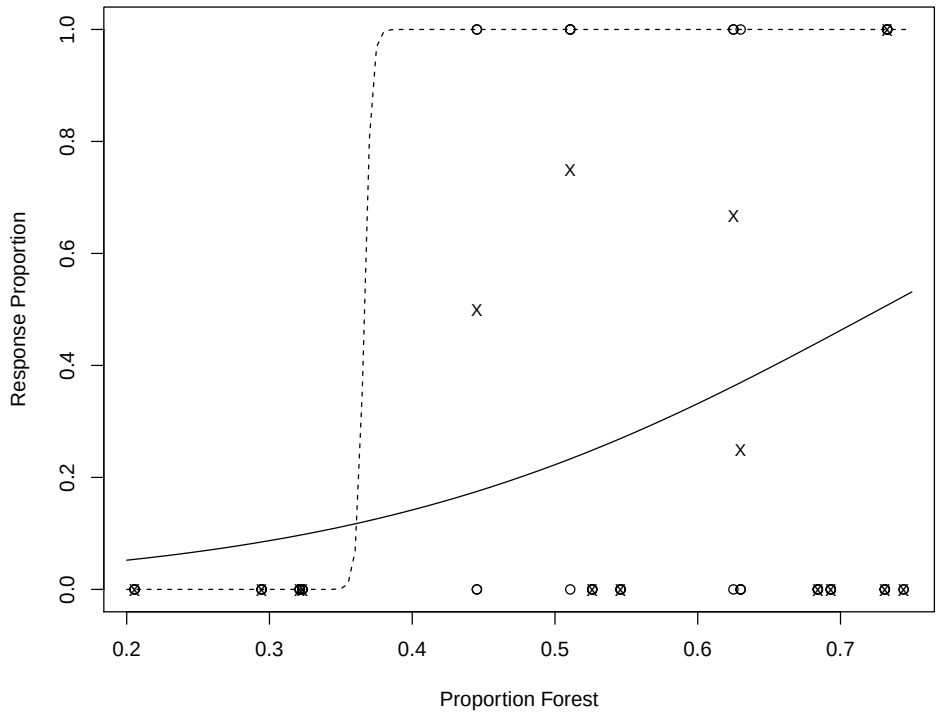

Figure 4: Estimated probabilities of use of areas by bobcats under two different models. Circles are observed binary responses $y_{i,j}$, X-symbols are $R_i = S_i/n_i$, dashed curve gives estimates of ψ_i from the model of Part II and solid curve gives estimates of ψ_i from the model of Part III.

Questions

1. Write the random model component of expression (2) in the form of an exponential dispersion family,

$$f(s_i|\theta_i, \phi) = \exp\left[a_i(\phi)\{r_i\theta_i - b(\theta_i)\} + c(r_i, \phi)\right].$$

   Identify θ_i, ϕ, and $a_i(\phi)$ in terms of the original quantities in (2). Identify $b(\theta_i)$ in terms of θ_i. Note that the use of $a_i(\phi)$ here is slightly more general than what is often written for an exponential dispersion family with only ϕ. Here, $a_i(\phi)$ is a function of ϕ but may include other constants that could depend on the index $i = 1, \ldots, N$.

2. For the model of expressions (2) and (3), maximum likelihood estimates of the regression parameters were $\hat{\beta}_0 = -2.836$ and $\hat{\beta}_1 = 2.618$. The estimated inverse expected information was

$$\hat{I}^{-1} = \begin{pmatrix} 23.860 & -37.615 \\ -37.615 & 63.368 \end{pmatrix}.$$

   The deviance associated with this estimated model was 37.332.

   Use these values to assess the effect of the proportion of forested area on the presence or absence of bobcats under the assumed model using $\alpha = 0.10$.

3. Fitting a model with constant binomial parameter ($p_i = p$ for all i in (2)) results in $\hat{p} = 0.207$ and a deviance of 40.302. Use this information in combination with values from question 2 to assess the effect of the proportion of forested area, still using $\alpha = 0.10$.

4. Comment on the results of questions 2 and 3. As part of your considerations, use any appropriate information from questions 2 and 3 to assess the appropriateness of the model of expressions (2) and (3) for assessing the effect of forested area on bobcat presence/absence.

5. Consider an area with proportion of forest $x_i = 0.50$. At this value of the covariate, how do the models of Part I and Part II represent the probability that the area is used by bobcats? For each model, what are expressions for the probability that bobcats are detected as using the area?

6. At a fixed covariate value, if estimated versions of these models gave $\hat{p}_i > \hat{\psi}_i$ where p_i is from (2) and ψ_i is from (4), what would you conclude?

   Hint: To simplify this, consider only one sampling occasion, and define events U_i as use of area i and D_i as detection. Relate p_i from (2) and ψ_i from (4) to the probabilities of these events. Also consider what events the data provide information about (which should be the same for the two models).

7. For the model of Part II, derive the marginal probability mass function of S_i; $i = 1, \ldots, N$.

8. Maximum likelihood estimates for the model of Part II are $\hat{\gamma}_0 = -147.956$, $\hat{\gamma}_1 = 403.709$ and $\hat{p} = 0.077$. The inverse observed information is (the order is γ_0, γ_1, p),

$$\hat{I}^{-1} = \begin{pmatrix} 7.7 \times 10^8 & -2.2 \times 10^9 & -0.0153 \\ -2.2 \times 10^9 & 7.0 \times 10^9 & -0.146 \\ -0.0153 & -0.0146 & 0.0017 \end{pmatrix}.$$

   (a) Use these values to assess the effect of forest on the use of areas by bobcats under this model.

   (b) Compute both 90% and 95% intervals for the parameter p. These intervals illustrate a potential deficiency of Wald theory in developing interval estimators. Breifly (one sentence) describe this potential deficiency.

   (c) What options might be available for application to this problem that would not suffer the deficiency of Wald intervals?

9. Unlike the model of Part I, the marginal response distribution for the model of Part II (which you derived for question 5) cannot be represented in the form of an exponential dispersion family with known dispersion parameter. As a result, direct computation of a deviance that has known asymptotic properties is not possible. Outline a procedure you might use to assess the effectiveness of the model of Part II in representing the data.

10. In Figure 4 the solid curve corresponds to the model of Part III fit to the circles (the $y_{i,j}$) while the dashed curve corresponds to the model of Part II for the X symbols (the R_i). Suppose that an assessment of these models (perhaps similar to the procedure you outlined in question 9) suggests that the model of Part III, while superior to that of Part II, tends to underestimate the probability of use for areas

with larger proportions of forest. Suggest a modification that might improve the model.

Hint: focus on expression (8).

11. While frequentist estimation of parameters in the model of Part III is not difficult, we might choose to take a Bayesian approach to analysis. What parameters would need to be assigned prior distributions? What prior distributions might you think of using and why?

12. The marginal likelihood for the model of Part III can be obtained by summing over values of Z_i and is not overly complex. But suppose one wished instead to "integrate out" the z_i as part of an overall Gibbs Sampling algorithm. Using generic notation so that $\pi(\theta)$ represents the prior distribution for any quantity θ in the model, and $p(\theta|\cdot)$ represents the full conditional density or mass function of any quantity θ given all other quantitites in the model, write down a list of the full conditional posteriors needed to implement a Gibbs algorithm for this problem. Do not try to use specific distributional forms.

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

Question 1. The random model component may be written in the form of an exponential dispersion family by taking

$$\begin{aligned} \theta_i &= \log\left(\frac{p_i}{1-p_i}\right) \\ \phi &= 1 \\ a_i(\phi) &= n_i \\ b(\theta_i) &= \log\{1+\exp(\theta_i)\} \\ c(r_i,\phi) &= log(n_i!)\log\{(n_i - n_i r_i)!\} - \log\{(n_i r_i)!\} \end{aligned}$$

Question 2. To assess the effect of proportion of forest on the presence/absence of bobcats we could compute a Wald theory interval estimate for β_1. A 90% interval is $(-10.477,\ 15.713)$ from which we would conclude that the proportion of the study area in forest has no effect.

Question 3. Based on the information in this question we could construct a likelihood ratio test of a reduced model with a constant binomial parameter against the regression model by taking the difference of deviances. That is, deviance is defined as

$$D_m = -2(\ell_m - \ell_s)$$

where ℓ_m is the maximized log likelihood for the model under consideration and ℓ_s is the log likelihood for the saturated model. Letting D_F denote the deviance for the (full) regression model and D_R the deviance for the (reduced) model with a constant binomial parameter

$$T = D_R - D_F = -2(\ell_R - \ell_s - \ell_F + \ell_s) = -2(\ell_R - \ell_F)$$

For the values given in this question, $T = 2.97$. With 2 parameters in the full model and 1 in the reduced, comparison with a chi-squared distribution with 1 degree of freedom results in $p = 0.0848$. Using $\alpha = 0.10$ this would lead to a decision to reject the model with a constant binomial parameter in favor of the full model with proportion forest as a covariate.

Question 4. In question 2 the conclusion was that there is no effect of proportion forest on bobcat presence or absence. In question 3 a model with proportion forest as a covariate was selected over a reduced model without that covariate, suggesting that knowledge of the proportion of forest increases our ability to determine the probability that bobcats use an area. This is a seeming contradiction, for which there may be any number of causes. For one, the assessment of question 2 was based on asymptotic normality of maximum likelihood estimators while that of question 3 was based on asymptotic behavior of maximized likelihoods, which are different results. The sample size is only $N = 15$ so neither of these results are certain to give good approximations. Regardless, the deviance of the model given in question 2 may be used as a likelihood ratio goodness of fit test statistic because the dispersion parameter is fixed (i.e., $\phi = 1$). Comparing this deviance of 37.332 to a chi-squared distribution with 13 degrees of freedom results in $p = 0.0004$ from which we would conclude that the regression model does not adequately describe the data.

Question 5. At a fixed level of the covariate, both the model of Part I and that of Part II represent the probability that an area is used by bobcats in the same way. In the model of Part I this is the expected value of R_i (which is also the expected value of any of the $Y_{i,j}$), while in the model of Part II this is the expected value of Z_i, namely,

$$\begin{aligned} E(R_i) &= \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \\ E(Z_i) &= \frac{\exp(\gamma_0 + \gamma_1 x_i)}{1 + \exp(\gamma_0 + \gamma_1 x_i)} \end{aligned}$$

The models differ in how they treat the probability of detection, however. In the model of Part I the probability of detection given use is 1, this is also the expected

value of R_i. In the model of Part II, the probability of detection depends on the use (i.e., the value of Z_i). If the area is not used ($Z_i = 0$) then the probability of detection is 0. If the area is used ($Z_i = 1$) the probability of detection in any one sampling occasion is p in expression (6). Over the n_i sampling occasions in area i, the probability of detection is then

$$Pr(S_i > 0|Z_i = 1) = 1 - Pr(S_i = 0|Z_i = 1) = 1 - (1-p)^{n_i}.$$

Question 6. Following the hint, let U_i denote the event that area i is used and D_i the event that use is detected. In the model from Part I $p_i = Pr(U_i)$, and in the model of Part II $\psi_i = Pr(U_i)$. Now, the model from Part I assumes that $Pr(D_i|U_i) = 1$ so that

$$p_i = Pr(U_i) = Pr(D_i \cap U_i)$$

In the model from Part II, $Pr(D_i|U_i) = p$ so that

$$pPr(U_i) = Pr(D_i \cap U_i) \Rightarrow \psi_i = Pr(U_i) = \frac{1}{p} Pr(D_i \cap U_i)$$

Now, the data provide information about $Pr(D_i \cap U_i)$ in both models and $0 < p < 1$ so that $\hat{p}_i$ should not be greater than $\hat{\psi}$. If this would occur, we would conclude that the assumption, common to both models, that bobcats are never falsely detected must be in error.

Question 7. First note that $Pr(S_i = 0|Z_i = 0) = 1$. Thus,

$$\begin{aligned} Pr(S_i = 0) &= Pr(S_i = 0|Z_i = 0)Pr(Z_i = 0) + Pr(S_i = 0|Z_i = 1)Pr(Z_i = 1) \\ &= (1 - \psi_i) + (1-p)^{n_i}\psi. \end{aligned}$$

Similarly, for $x = 1, 2, \ldots, n_i$,

$$\begin{aligned} Pr(S_i = x) &= Pr(S_i = x|Z_i = 1)Pr(Z_i = 1) \\ &= \psi \frac{n_i!}{x!(n_i - x)!} p^x (1-p)^{n_i - x}. \end{aligned}$$

The complete marginal probability mass function for S_i is then

$$g(s_i|p, \psi) = \begin{cases} (1 - \psi_i) + (1-p)^{n_i}\psi & s_i = 0 \\ \psi \frac{n_i!}{x!(n_i - x)!} p^x (1-p)^{n_i - x} & s_i = 1, 2, \ldots, n_i \end{cases}$$

Question 8. (a) A 90% interval estimate of the regression coefficient for proportion forest is $(-137226.9, 138034.3)$, from which we would conclude that there is no evidence that forest has an effect on the presence or absence of bobcats.

(b) A 90% interval estimate for p is $(0.0092, 0.1448)$, and a 95% interval is $(-0.0038, 0.1578)$. This illustrates that Wald theory intervals do not automatically obey parameter space constraints.

(c) Intervals computed from normed profile likelihoods or using the percentile parametric bootstrap method are guaranteed to conform to the parameter space.

Question 9. As is evident in Figure 3 of the question, the primary feature of these data that make modeling difficult is the observed values of 0 at higher values of the proportion of forest covariate. One might then select such a feature as, for example, the number of zeros for proportion of forest greater than 50% for use in simulation-based model assessment. An outline of an algorithm would be as follows:

i Compute the number of zeros at covariate values $x_i > 0.5$ in the actual data and denote this as t^*.

ii Using the estimated parameter values and the covariate values in the actual data, simulate M data sets from the model. Compute the number of zeros at covariate values $x_i > 50$ for each data set and denote these values as t_m; $m = 1, \ldots, M$.

iii A $p-$value for assessing the ability of the model to reflect this feature of the data is then

$$p = \frac{1}{M} \sum_{m=1}^{M} I(t_m \geq t^*),$$

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

Question 10. One could change the link function $m(\psi_i)$ in expression (8). A parameterized family of link functions could provide more flexibility in how slowly or rapidly the fitted curve increases with the proportion of forest.

Question 11. A Bayesian analysis of the model of Part III would require the specification of prior distributions for γ_0, γ_1, β_0 and β_1. One might consider diffuse normal priors for each of these parameters because each can take any value on the real line.

Question 12. Let $p(\theta|\cdot)$ be generic notation for the full conditional posterior of θ and $\pi(\theta)$ generic notation for the prior of θ. If the Z_i; $i = 1, \ldots, N$ are included as "parameters" in a Gibbs algorithm, the full conditional distributions needed would be

$$\begin{aligned}
p(\gamma_0|\cdot) &\propto \pi(\gamma_0) \prod_{i=1}^{N} h(z_i|\psi_i) \\
p(\gamma_1|\cdot) &\propto \pi(\gamma_1) \prod_{i=1}^{N} h(z_i|\psi_i) \\
p(\beta_0|\cdot) &\propto \pi(\beta_0) \prod_{i=1}^{N} \prod_{j=1}^{n_i} f(y_{i,j}|z_i, p_{i,j}) \\
p(\beta_1|\cdot) &\propto \pi(\beta_1) \prod_{i=1}^{N} \prod_{j=1}^{n_i} f(y_{i,j}|z_i, p_{i,j})
\end{aligned}$$

and, for $i = 1, \ldots, N$

$$p(z_i) \propto h(z_i|\psi_i) \prod_{j=1}^{n_i} f(y_{i,j}|z_i, p_{i,j})$$

Background

Inspections of critical rotating components of aircraft engines are done at the time of manufacturing and periodically when in service to assure the needed high reliability of the engine system. An accurate assessment of probability of detection (POD) of flaws (e.g., fatigue cracks) as a function of size is needed to make decisions on how often to inspect. Typically POD is estimated on the basis of laboratory experiments that contain seeded flaws (i.e., flaws purposely placed into test specimens).

Part I

Engineers conducted an experiment, generating 96 observations from an automated eddy current inspection system providing signal strength (units of volts) on flaws of a known size (units of mils, where one mil is 1/1000 of an inch). A physics-based model suggests that the logarithm of signal strength should be linearly related to the logarithm of flaw size over the range of sizes that are of interest. Based on this knowledge, the standard statistical model fit to such data is

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where

$$Y = \log(\text{signal strength})$$

and

$$x = \log(\text{flaw size})$$

and $\epsilon \sim \mathrm{N}(0, \sigma^2)$ is an error term corresponding to sources of variability other than flaw size that are not accounted for and are independent of flaw size. `R` output for the model fit using the function `lm` (implementing ordinary least squares) is shown below.

```
> summary(lm.out)

Call:
lm(formula = log(Response) ~ log(FlawSize), data = InspectionData)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99276 -0.17471  0.09239  0.23030  0.88527

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.77306    0.18822  -20.05   <2e-16 ***
log(FlawSize)  1.19683    0.06933   17.26   <2e-16 ***
---


Residual standard error: 0.3296 on 94 degrees of freedom
Multiple R-squared: 0.7602,     Adjusted R-squared: 0.7577
```

```
F-statistic:   298 on 1 and 94 DF,  p-value: < 2.2e-16

> vcov(lm.out)
               (Intercept) log(FlawSize)
(Intercept)     0.03542502  -0.012838572
log(FlawSize) -0.01283857   0.004806409
```

1. Derive expressions for the conditional (on x) mean and variance of Y.

2. Maximum likelihood (ML) estimation is usually used in the estimation of POD and associated quantities. [1] For the current data, there is no censoring. The ML estimates of the model parameters are closely related to the estimates provided by the `R` function `lm`. Give expressions for the ML estimates of the model parameters as a function of the estimates from the `R` output above. No derivations are required.

3. The threshold for detection for this type of experiment is 0.18 volts. That is, if the signal strength is greater than 0.18 volts, there is a decision that a flaw has been detected.

a) Using this criterion, derive an expression for POD as a function of the flaw size.

b) Give an expression for the ML estimate of $\mathrm{POD}(x)$ and show explicitly how the `R` output given above can be used to compute a ML estimate of POD as a function of flaw size. Recall that the model is in terms of $x = \log(\text{flaw size})$ and $y = \log(\text{signal strength})$.

4. Using the delta method, derive an expression for variance of the ML estimator of the POD (or some appropriate one-to-one function of POD) at a given flaw size from the answer to question **3**. Show how this expression, along with above `R` output, can be used to obtain a standard error for the POD estimate (or some appropriate one-to-one function of POD).

5. A widely used metric for inspection capability is the flaw size that can be detected with probability p (where p is commonly taken to be 0.90). This quantity is denoted by a_p. Give an expression for the ML estimator of a_p.

6. Give an expression that can be used to compute the standard error of the estimator of a_p (or some appropriate one-to-one function of a_p) as a function of the estimated variances and covariances between the model parameter estimators. Show how this expression, along with above `R` output, can be used to obtain an upper 95% confidence bound for a_p.

Part II

Planning a POD experiment involves choosing the number of specimens with seeded flaws and the sizes of the seeded flaws. One approach for doing this is to use a computer program that will allow the test planner to compute and compare properties of proposed test plans. Such a program

[1] This is because in some cases left and right censoring can arise due to non-detects (left censoring) and strong signals that saturate the measuring system (right censoring).

would require an evaluation of approximate variances (based on large-sample approximations) of estimators of quantities of interest [e.g., a_p or $\log(a_p)$].

7. Using the model from Part I, give an expression for the log likelihood of a proposed experiment as a function of the set of n values of x and the model parameters.

8. Derive an expression for the contribution of each (x_i, y_i) pair to the Fisher information matrix for the model parameters.

9. Give an expression for large-sample approximate variance-covariance matrix of the ML estimators of the parameters.

10. Give an expression for the large-sample approximate variance of the ML estimator of $\log(a_p)$.

Part III

A statistician who had been asked to analyze the data from Part I noticed when plotting the data that there were only 16 unique flaw sizes and that each of these flaw sizes had exactly six corresponding responses. After inquiring about this she was told that indeed there were only 16 flaw specimens and that each had been inspected 6 times. Each of three operators inspected each specimen two times.

11. After consulting further with the engineers about the purpose of the study and how it would be used to describe future inspection scenarios, the statistician correctly decided to treat operator effects as fixed and flaw effects as random. Describe the possible reasons for these decisions.

12. Write down an appropriate linear statistical model without interactions for analyzing the data from this experiment. Clearly state any assumptions that you need to make.

13. Suppose that the data indicated that there was not an important operator effect (something that had been expected because the inspection is mostly automated). Derive expressions for POD and a_p as a functions of the model parameters.

14. What would be the practical effect of using the naive analysis, ignoring the repeated measures structure of the experiment, relative to the correct analysis based on a repeated-measures model?

15. Suppose that you have software that will provide estimates of the model parameters and other information that you might need. Provide suggestions for at least two different methods that you could use to obtain confidence intervals for POD for a given flaw size. Describe the trade-offs between these methods. For one of these methods, list the steps in an algorithmic fashion such that a computer programmer who does not know much calculus, but who is otherwise competent, could implement the method.

1. The mean of Y is

$$\begin{aligned}\mathrm{E(Y)} &= \mathrm{E}(\beta_0 + \beta_1 \mathrm{x} + \epsilon) \\ &= \mathrm{E}(\beta_0) + \mathrm{E}(\beta_1 \mathrm{x}) + \mathrm{E}(\epsilon) \\ &= \beta_0 + \beta_1 x.\end{aligned}$$

The variance of Y is

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{Var}(\beta_0 + \beta_1 x + \epsilon) \\ &= \mathrm{Var}(\beta_0) + \mathrm{Var}(\beta_1 x) + \mathrm{Var}(\epsilon) \\ &= \sigma^2.\end{aligned}$$

2. In the case of regression (linear or nonlinear) with independent normally distributed errors, it is easy to show that maximum likelihood (ML) estimators of the regression coefficients are equivalent to the least squares estimators of the regression coefficients. Let $\widehat{\beta}_0$ and $\widehat{\beta}_1$ denote these estimators. Then it is also easy to show that the ML estimate of σ is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}[y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 x_i)]^2 = s^2 \times \frac{n-p}{n}$$

where s^2 is the usual unbiased estimate of σ^2 and $p = 2$ is the number of parameters in the regression model. Thus the ML estimates are

$$\begin{aligned}\widehat{\beta}_0 &= -3.77306 \\ \widehat{\beta}_1 &= 1.19683 \\ \widehat{\sigma} &= s \times \sqrt{\frac{n-p}{n}} = 0.3296 \times \sqrt{\frac{94}{96}} = 0.3261.\end{aligned}$$

3. Let $y_{th} = \log(0.18)$ be the detection threshold. Then the probability of detection is

$$\begin{aligned}\mathrm{POD}(x) = \Pr(Y > y_{th}) &= \Pr\left(Z > \frac{y_{th} - (\beta_0 + \beta_1 x)}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{y_{th} - (\beta_0 + \beta_1 x)}{\sigma}\right) \\ &= \Phi\left(\frac{\beta_0 + \beta_1 x - y_{th}}{\sigma}\right).\end{aligned}$$

The probability of detection can be estimated by evaluating $\mathrm{POD}(x)$ at values of the estimates of the model parameters. That is,

$$\widehat{\mathrm{POD}}(x) = \Phi\left(\frac{\widehat{\beta}_0 + \widehat{\beta}_1 x - y_{th}}{\widehat{\sigma}}\right).$$

4. A confidence interval based on inverting the Wald statistics is most computationally convenient but has the disadvantage of not being transformation invariant. Thus one needs to think carefully about the transformation to use to construct an interval. One approach is to base the interval procedure on the studentization of $\widehat{\mathrm{POD}}(x)$. In some circles, it is recognized that using an appropriate one-to-one transformation of of the quantity of interest provides a better alternative. In particular we can studentize

$$\widehat{z} = \frac{\widehat{\beta}_0 + \widehat{\beta}_1 x - y_{th}}{\widehat{\sigma}}.$$

That is, letting

$$z = \frac{\beta_0 + \beta_1 x - y_{th}}{\sigma}$$

we can construct a confidence interval based on the approximate $\mathrm{N}(0,1)$ distribution of

$$\frac{\widehat{z} - z}{\widehat{\sigma}_{\widehat{z}}}$$

where $\widehat{\sigma}_{\widehat{z}}$ is a standard error of $\widehat{z}$. In particular, an approximate $100(1-\alpha)\%$ confidence interval for z can be computed as

$$\underset{\sim}{[z}, \quad \tilde{z}] = \widehat{z} \pm \Phi^{-1}_{(1-\alpha/2)} \times \widehat{\sigma}_{\widehat{z}}$$

where $\Phi^{-1}_{(1-\alpha/2)}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Then an approximate $100(1-\alpha)\%$ confidence interval for $\mathrm{POD}(x)$ would be

$$\left[\Phi(\underset{\sim}{z}), \quad \Phi(\tilde{z})\right].$$

Letting $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma)$ and using the delta method,

$$\widehat{\sigma}_{\widehat{z}} = \left[\mathbf{g}_1' \; \widehat{\Sigma}_{\widehat{\boldsymbol{\theta}}} \; \mathbf{g}_1\right]^{\frac{1}{2}}$$

where $\widehat{\Sigma}_{\widehat{\boldsymbol{\theta}}}$ is an estimate of the covariance matrix of $\widehat{\boldsymbol{\theta}}$ and

$$\mathbf{g}_1 = \left(\frac{\partial z}{\partial \beta_0}, \frac{\partial z}{\partial \beta_1}, \frac{\partial z}{\partial \sigma}\right)' = \left(\frac{1}{\sigma}, \frac{x}{\sigma}, -\frac{z}{\sigma^2}\right)'$$

and the partial derivatives are evaluated at the ML estimates of $\boldsymbol{\theta}$.

5. First solve for x_p in $\mathrm{POD}(x_p) = p$. That is,

$$\Phi\left[\frac{\beta_0 + \beta_1 x_p - y_{th}}{\sigma}\right] = p$$

$$x_p = \frac{y_{th} + \Phi^{-1}(p)\sigma - \beta_0}{\beta_1}.$$

Then

$$a_p = \exp\left[\frac{y_{th} + \Phi^{-1}(p)\sigma - \beta_0}{\beta_1}\right].$$

ML estimates of these quantities can be obtained by evaluating these expressions at the ML estimates of β_0, β_1, and σ.

6. Again, we can use the delta method

$$\widehat{\sigma}_{\widehat{x}_p} = \left[\mathbf{g}_2' \, \widehat{\Sigma}_{\widehat{\boldsymbol{\theta}}} \, \mathbf{g}_2\right]^{\frac{1}{2}}$$

where

$$\mathbf{g}_2 = \left(\frac{\partial x_p}{\partial \beta_0}, \frac{\partial x_p}{\partial \beta_1}, \frac{\partial x_p}{\partial \sigma}\right)' = \left(\frac{1}{\beta_1}, -\frac{x_p}{\beta_1^2}, -\frac{\Phi^{-1}(p)}{\beta_1}\right)' \tag{1}$$

and again the partial derivatives are evaluated at the ML estimates of $\boldsymbol{\theta}$.

As with POD, an approximate $100(1-\alpha)\%$ confidence interval for x_p can be computed as

$$[\underset{\sim}{x}_p, \quad \tilde{x}_p] = \widehat{x}_p \pm \Phi^{-1}_{(1-\alpha/2)} \times \widehat{\sigma}_{\widehat{x}_p}$$

Then an approximate $100(1-\alpha)\%$ confidence interval for a_p would be

$$[\underset{\sim}{a}_p, \quad \tilde{a}_p] = [\exp(\underset{\sim}{x}_p), \quad \exp(\tilde{x}_p)].$$

An upper 95% confidence bound for x_p could be obtained by using $\tilde{a}_p = \exp(\tilde{x}_p)$ with $\alpha = 0.10$.

Part II

7. The log likelihood for observation i with log flaw size x_i is

$$\mathcal{L}_i = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\epsilon_i)^2$$

where Y_i is the signal response that corresponds with each flaw size, x_i and

$$\epsilon_i = Y_i - \beta_0 - \beta_1 x_i.$$

The total log likelihood for all n observations is

$$\mathcal{L}_T = \sum_{i=1}^{n} \mathcal{L}_i.$$

8. The partial derivatives of the log likelihood for observation i are

$$\begin{aligned}\frac{\partial \mathcal{L}_i}{\partial \beta_0} &= \frac{\epsilon_i}{\sigma^2} \\ \frac{\partial \mathcal{L}_i}{\partial \beta_1} &= \frac{x\epsilon_i}{\sigma^2} \\ \frac{\partial \mathcal{L}_i}{\partial \sigma} &= \frac{\epsilon_i^2}{\sigma^3} - \frac{1}{\sigma}.\end{aligned}$$

The Hessian matrix of second partial derivatives of the log likelihood for observation i is,

$$H_i = \begin{bmatrix} \frac{\partial^2 \mathcal{L}_i}{\partial \beta_0^2} & \frac{\partial^2 \mathcal{L}_i}{\partial \beta_0 \beta_1} & \frac{\partial^2 \mathcal{L}_i}{\partial \beta_0 \sigma} \\ \frac{\partial^2 \mathcal{L}_i}{\partial \beta_1 \beta_0} & \frac{\partial^2 \mathcal{L}_i}{\partial \beta_1^2} & \frac{\partial^2 \mathcal{L}_i}{\partial \beta_1 \sigma} \\ \frac{\partial^2 \mathcal{L}_i}{\partial \sigma \beta_0} & \frac{\partial^2 \mathcal{L}_i}{\partial \sigma \beta_1} & \frac{\partial^2 \mathcal{L}_i}{\partial \sigma^2} \end{bmatrix} = \begin{bmatrix} -\frac{1}{\sigma^2} & -\frac{x_i}{\sigma^2} & -\frac{2\epsilon_i}{\sigma^3} \\ -\frac{x_i}{\sigma^2} & -\frac{x_i^2}{\sigma^2} & -\frac{2\epsilon_i x_i}{\sigma^3} \\ -\frac{2\epsilon_i}{\sigma^3} & -\frac{2\epsilon_i x_i}{\sigma^3} & \frac{1}{\sigma^2} - \frac{3\epsilon_i^2}{\sigma^4} \end{bmatrix}.$$

Noting that $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$, the Fisher information matrix for observation i can be expressed as

$$F_i = E \begin{bmatrix} -\frac{\partial^2 \mathcal{L}_i}{\partial \beta_0^2} & -\frac{\partial^2 \mathcal{L}_i}{\partial \beta_0 \beta_1} & -\frac{\partial^2 \mathcal{L}_i}{\partial \beta_0 \sigma} \\ -\frac{\partial^2 \mathcal{L}_i}{\partial \beta_1 \beta_0} & -\frac{\partial^2 \mathcal{L}_i}{\partial \beta_1^2} & -\frac{\partial^2 \mathcal{L}_i}{\partial \beta_1 \sigma} \\ -\frac{\partial^2 \mathcal{L}_i}{\partial \sigma \beta_0} & -\frac{\partial^2 \mathcal{L}_i}{\partial \sigma \beta_1} & -\frac{\partial^2 \mathcal{L}_i}{\partial \sigma^2} \end{bmatrix} = \frac{1}{\sigma^2} \begin{bmatrix} 1 & x_i & 0 \\ x_i & x_i^2 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

9. The large-sample approximate covariance matrix of $\widehat{\boldsymbol{\theta}}$ is

$$\begin{aligned}
\Sigma_{\widehat{\boldsymbol{\theta}}} &= \left(\sum_{i=1}^{n} F_i \right)^{-1} \\
&= \left(\frac{1}{\sigma^2} \begin{bmatrix} n & \sum_{i=1}^n x_i & 0 \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 & 0 \\ 0 & 0 & 2n \end{bmatrix} \right)^{-1} \\
&= \sigma^2 \begin{bmatrix} \frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} & \frac{-\sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} & 0 \\ \frac{-\sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} & \frac{n}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} & 0 \\ 0 & 0 & \frac{1}{2n} \end{bmatrix} \\
&= \frac{\sigma^2}{n} \begin{bmatrix} \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2/n} & \frac{-\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2/n} & 0 \\ \frac{-\sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2/n} & \frac{n}{\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2/n} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}.
\end{aligned}$$

10. The large-sample approximate variance of the ML estimator of $x_p = \log(a_p)$ can be obtained again by using the delta method. In particular,

$$\mathrm{AVar}(\widehat{x}_p) = \mathbf{g}_2' \; \Sigma_{\widehat{\boldsymbol{\theta}}} \; \mathbf{g}_2$$

where $\mathbf{g}_2$ is defined in (1) and the partial derivatives are evaluated at the estimates of the elements of $\boldsymbol{\theta}$.

Part III

11. It would be appropriate to consider the operators as fixed effects if these operators are the operators that would be doing inspections in the future and if the operator effect is not important. If the operator effect is important (unlikely in an automated inspection with experienced operators), it would imply separate POD functions for each operator.

It would be appropriate to consider flaw effects as random if seeded flaws, conditional on their nominal size, tend to produce different responses in a consistent manner. The flaws seeded in the test specimens are samples with response variability that is assumed to be representative of the response variability in the actual flaws that are to be detected.

12. Assuming that there is no interaction between operator and flaw effects,

$$Y = \beta_0 + \beta_1 x + \beta_2 w_1 + \beta_3 w_2 + \gamma_{\text{Flaw}} + \epsilon$$

where, among other ways, the dummy variables can be coded as

$$w_1 = \begin{cases} 1 & \text{for operator A} \\ 0 & \text{for for operator B} \\ -1 & \text{for for operator C} \end{cases}$$

$$w_2 = \begin{cases} 0 & \text{for operator A} \\ 1 & \text{for for operator B} \\ -1 & \text{for for operator C} \end{cases}$$

and $\gamma_{\text{Flaw}} \sim \text{N}(0, \sigma^2_{\text{Flaw}})$ and $\epsilon \sim \text{N}(0, \sigma^2_\epsilon)$ where γ_{Flaw} and ϵ are assumed to be independent.

13. First note if the operator effects are negligible (as they were in the actual application) or if each operator has the same probability of doing a given inspection

$$Y \sim \text{N}(\beta_0 + \beta_1 \text{x}, \sigma^2_{\text{Flaw}} + \sigma^2_\epsilon).$$

Then

$$\begin{aligned} \text{POD}(x) = \Pr(Y > y_{th}) &= \Pr\left(Z > \frac{y_{th} - (\beta_0 + \beta_1 x)}{\sqrt{\sigma^2_{\text{Flaw}} + \sigma^2_\epsilon}} \right) \\ &= 1 - \Phi\left(\frac{y_{th} - (\beta_0 + \beta_1 x)}{\sqrt{\sigma^2_{\text{Flaw}} + \sigma^2_\epsilon}} \right) \\ &= \Phi\left(\frac{\beta_0 + \beta_1 x - y_{th}}{\sqrt{\sigma^2_{\text{Flaw}} + \sigma^2_\epsilon}} \right). \end{aligned}$$

Again, one can estimate $\text{POD}(x)$ by evaluating this expression at the ML (or REML) estimates of the model parameters.

14. Using the naive analysis without considering the repeated-measures nature of the experiment could be expected to seriously underestimate the standard errors of the model parameters, leading to confidence intervals that are too narrow.

15. There are several different methods that could be used to construct approximate confidence intervals for this kind of model. These include

- Using a Wald statistic, leading to a "normal approximation" interval.
- Basing the desired interval on the likelihood, by inverting a likelihood ratio test.
- A bootstrap or simulation-based method.

The simplest procedure would use the Wald approach in a manner similar to that used in Question 4 of Part I. The estimate of POD can be expressed as

$$\widehat{\mathrm{POD}(x)} = \Phi(\widehat{z})$$

where

$$\widehat{z} = \frac{\widehat{\beta}_0 + \widehat{\beta}_1 x - y_{th}}{\sqrt{\widehat{\sigma}^2_{\mathrm{Flaw}} + \widehat{\sigma}^2_{\epsilon}}}.$$

Then an approximate $100(1-\alpha)\%$ confidence interval for z can be computed as

$$[\underset{\sim}{z}, \quad \tilde{z}] = \widehat{z} \pm \Phi^{-1}_{(1-\alpha/2)} \times \widehat{\sigma}_{\widehat{z}}$$

Letting $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma_{\mathrm{Flaw}}, \widehat{\sigma}_{\epsilon})$ and using the delta method,

$$\widehat{\sigma}_{\widehat{z}} = \left[\mathbf{g}_3' \; \widehat{\Sigma}_{\widehat{\boldsymbol{\theta}}} \; \mathbf{g}_3\right]^{\frac{1}{2}}$$

where $\widehat{\Sigma}_{\widehat{\boldsymbol{\theta}}}$ is an estimate of the covariance matrix of $\widehat{\boldsymbol{\theta}}$ and

$$\begin{aligned}
\mathbf{g}_3 &= \left(\frac{\partial z}{\partial \beta_0}, \frac{\partial z}{\partial \beta_1}, \frac{\partial z}{\partial \sigma_{\mathrm{Flaw}}}, \frac{\partial z}{\partial \sigma_{\epsilon}}\right)' \\
&= \left(\frac{1}{(\sigma^2_{\mathrm{Flaw}} + \sigma^2_{\epsilon})^{1/2}}, \frac{x}{(\sigma^2_{\mathrm{Flaw}} + \sigma^2_{\epsilon})^{1/2}}, -\frac{(\widehat{\beta}_0 + \widehat{\beta}_1 x - y_{th})\sigma_{\mathrm{Flaw}}}{(\sigma^2_{\mathrm{Flaw}} + \sigma^2_{\epsilon})^{3/2}}, -\frac{(\widehat{\beta}_0 + \widehat{\beta}_1 x - y_{th})\sigma_{\epsilon}}{(\sigma^2_{\mathrm{Flaw}} + \sigma^2_{\epsilon})^{3/2}}\right)'
\end{aligned}$$

and the partial derivatives are again evaluated at the estimates of the elements of $\boldsymbol{\theta}$. Finally, the confidence interval for POD is, as with the simpler model in Part I,

$$[\Phi(\underset{\sim}{z}), \quad \Phi(\tilde{z})]\,.$$

The Wald method is easy to implement, but with a small number of observational units such a simple first-order accurate method could turn out to be in adequate in that the actual coverage probability could be far from the nominal $100(1-\alpha)\%$ (as could be demonstrated by simulation). Likelihood-based methods are also first-order accurate, but generally out-perform Wald procedures. A second-order correct bootstrap or simulation-based method (such as bootstrap-t could be expected to do better. Bootstrap samples could be generated by a resampling scheme or through a full parametric simulation (i.e., simulate data from the fitted model). The latter might be expected to do better in smaller samples, but be less robust to departures from model assumptions.

For example, the normal approximation used in the Wald-based intervals could be improved by using simulation (parametric bootstrap) to obtain the distribution of $\widehat{z}$ and then replacing $\Phi^{-1}_{(1-\alpha/2)}$ by the $\alpha/2$ and $1-\alpha/2$ quantiles of this distribution. This is a bootstrap-t method that will be second-order accurate.