PhD Prelim Exam
METHODS

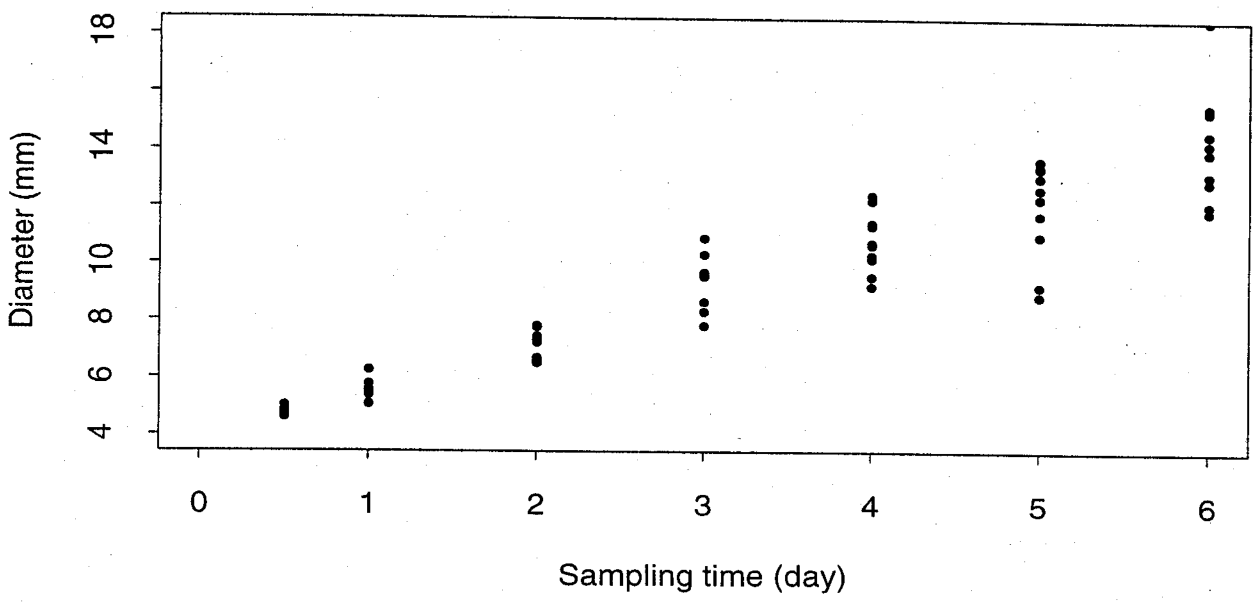
Spring 2004
(Given on 3/18/04)

A group on campus is studying growth rates of fungi that cause ugly looking spots on apples, which reduces their value. A fungus that grows faster is more of a problem than one that grows slowly. Ultimately, the group wants to compare growth rates between varieties collected from different parts of the country (VARIETIES). Because measuring fungal growth is laborious, they would like to design a statistically efficient experiment. The parts of this problem lead you through the stages of designing and analyzing such an experiment. The parts are generally independent, so if you get stuck on one part, start with another.

The fungi grow as dense colonies on a growth medium (agar with nutrients). A 'master colony' of fungus covering a large area is used to start the growth rate study. A plug of fungus, of known and fixed diameter, is transferred from a master colony to a small individual plate of agar with nutrients. The initial diameter is 4mm for every plug. As the fungus grows, it spreads horizontally from the starting plug. Growth rate is measured using increase in colony diameter. After X days of growth, the diameter of the colony is measured. The growth is reasonably circular, so the diameter is a reasonable measure of the size of the colony.

Part A:
The researcher has preliminary data on growth of one variety of fungus, using the method described in the previous paragraph. Ten plugs were measured after 1/2 day of growth, 10 were measured after 1 day, and 10 were measured each subsequent day until day 6. Each plug was measured only once, so a total of 70 plugs of fungus were used (7 times x 10 plugs per time). Measurement times were randomly assigned to plugs in a completely randomized design.

A plot of the data is:



1. The researcher is considering two possible models to estimate the growth rate:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \epsilon_i,\ \epsilon_i \sim \text{iid } N(0,\sigma^2) \qquad (1) \\ Y_i &= 4 + \beta_1 X_i + \epsilon_i,\ \epsilon_i \sim \text{iid } N(0,\sigma^2) \qquad (2) \end{aligned}$$

   Which of these two models is better for these data? Justify your choice.

2. The OLS estimate of β_1 in model 2 is $\hat{\beta}_{OLS} = \frac{\sum (Y_i - 4) X_i}{\sum X_i^2}$. The variance of the OLS estimator is $\text{Var}(\hat{\beta}_{OLS}) = \frac{\sigma^2}{\sum X_i^2}$. Some summary statistics computed from the data are:

n	$\sum X_i$	$\sum X_i^2$	$\sum Y_i$	$\sum Y_i^2$	$\sum (Y_i - \hat{Y}_i)^2$	$\sum X_i Y_i$
70	215	912.5	639.32	6649.85	86.81	2390.88

   Please estimate β_{OLS} and Var $(\hat{\beta}_{OLS})$.

3. Please estimate a 95% confidence interval for β_{OLS}. If you can not do the previous part, use $\hat{\beta}_{OLS} = 1.600$ and $\text{V}\hat{\text{a}}\text{r}(\hat{\beta}_{OLS}) = 0.0014$.

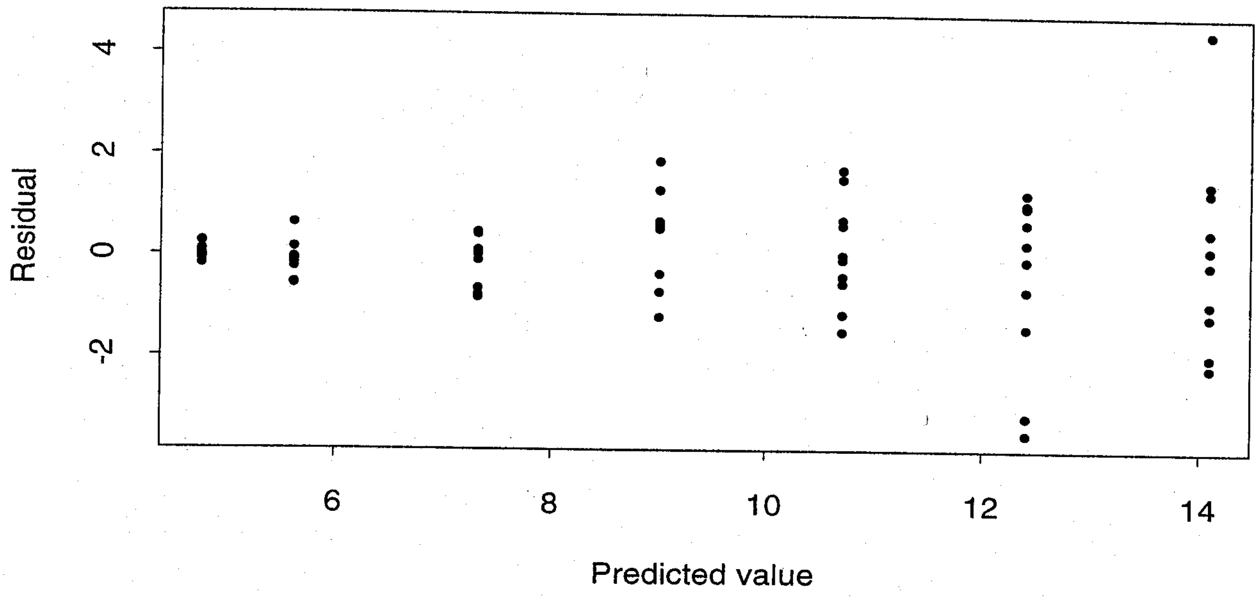
4. Derive $\hat{\beta}_{OLS}$ and Var $\hat{\beta}_{OLS}$ for model 2.

5. Using β as an estimate of growth rate assumes a linear relationship between Y (diameter) and X (day). The research wants to test for lack of fit to a linear relationship. They have fit 6 models (Table 2). Five are regressions; the sixth is the one-way ANOVA model with a separate mean, μ_j, for each day j. Sums-of-squared errors for each model are:

Model	SSE
$\text{E}(Y) = 4 + \beta_1 X$	86.909
$\text{E}(Y) = \beta_0 + \beta_1 X$	86.812
$\text{E}(Y) = 4 + \beta_1 X + \beta_2 X^2$	86.907
$\text{E}(Y) = 4 + \beta_1 X + \beta_2 X^2 + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4$	84.069
$\text{E}(Y) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_2 X^2 + \beta_3 X^3$	86.296
$\text{E}(Y) = \mu_j$	82.561

   Use the appropriate sums-of-squares to construct the most general possible test of lack of fit. Report your test statistic, a p-value (approximate from the tables), and a short conclusion.

Part B:
A residuals vs. predicted values plot (below) indicates that the error variance may not be constant.



One possible model for data with unequal error variance is

$$Y_i = 4 + \beta X_i + \epsilon_i,\ \epsilon_i \sim \text{indep } N(0, \sigma_i^2) \tag{3}$$

6. Indicate the properties (for example: bias, distribution, other properties) of $\hat{\beta}_{OLS}$ if model 2 (constant variance) is appropriate. Indicate which properties are maintained when model 3 (unequal variances) is correct.

7. Indicate the properties of $\hat{\text{Var}}\ \hat{\beta}_{OLS}$ if model 2 (constant variance) is appropriate. Indicate which properties are maintained when model 3 (unequal variances) is correct.

8. Because of the way fungal colonies grow, the researchers believe that the errors have constant coefficient of variation (c.v. = σ/μ). This model is

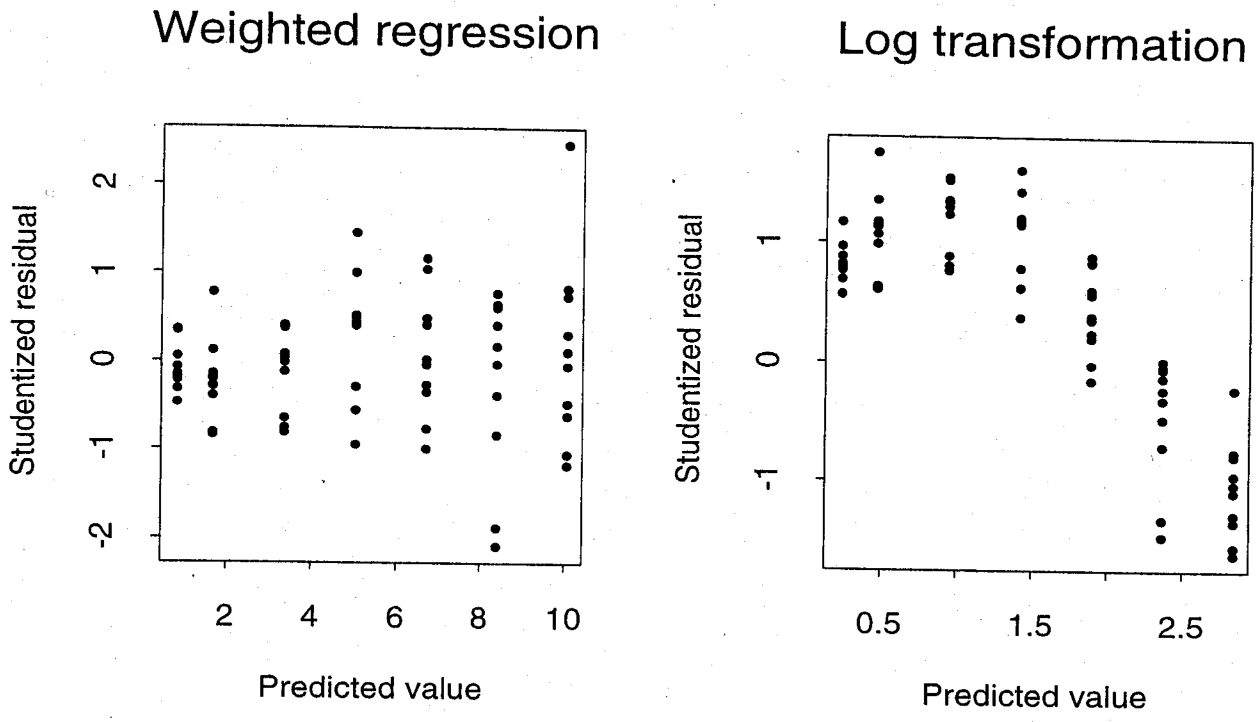
$$\begin{aligned} Y_i &= 4 + \beta X_i + \epsilon_i,\ \epsilon_i \sim \text{indep } N(0, \sigma_i^2) && (4) \\ \sigma_i^2 &= (4 + \beta X_i)^2 \sigma^2 && (5) \end{aligned}$$

Consider two approaches for dealing with unequal variances:
a) iteratively reweighted least squares, using weights $w_i = 1/(4 + \hat{\beta} X_i)^2$
b) log transforming the response and fitting the model

$$\log(Y_i - 3) = \beta X_i + \epsilon_i,\ \epsilon_i \sim \text{iid } N(0, \sigma^2) \tag{6}$$

The constant 3 in the log transformed model was chosen so that the median diameter =4 when $X_i = 0$, i.e. $\log(4 - 3) = \log(1) = 0$.

Plots of the residuals vs. predicted values for each approach are:



Which approach is more appropriate for these data? .Explain the advantages and disadvantages of each approach.

Part C:
For the second experiment (comparing growth rates among 6 varieties), the investigators and you decide to use 3 replicates at each of 4 sampling times (1 day, 2 days, 4 days, 6 days). The linear growth with constant coefficient of variation model (equations 4, 5) is reasonable for each variety.

9. If each plug is measured once, observations can be considered independent. The investigators want to test the null hypothesis that all 6 varieties have the same growth rate against an alternative that at least one variety has a different growth rate. How will you test this hypothesis?

   Describe your approach in sufficient detail that a reviewer of the manuscript can tell what was done. Simply saying 'use ANOVA or 'use a t test' is not sufficient.

10. After you get the data, you discover that the investigators didn't do the experiment as you originally thought. They only used a total of 18 plugs in the experiment (6 varieties x 3 plugs per variety). Each plug was measured four times, i.e. on each of the four days. Is assumption of independent errors reasonable? Explain why or why not.

11. One possible model for the actual data (see part 10) leads to the ANOVA table:

Source	d.f.
Day	1
Day*Variety	5
Plug(Variety)*Day	12
Error	54
Total	72

Write down a reasonable model that corresponds to this ANOVA table. Still assume that the measurement variance for each plug at each time has a constant coefficient of variation.

What, in terms of the parameter(s) in your model, is the mean growth rate for each variety?

Using your model and/or the ANOVA table, how can you test equality of growth rates among varieties?

1. Because the diameter at the start of the experiment is fixed at 4 mm, model 2 is more appropriate.

2. $\hat{\beta}_{OLS} = \frac{\sum Y_i X_i - 4 \sum X_i}{\sum X_i^2} = \frac{2390.88 - 4 \times 215}{912.5} = 1.678$
   $s^2 = \hat{\sigma}^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-1} = \frac{86.81}{69} = 1.258$
   $\text{Var } \hat{\beta} = \frac{s^2}{\sum X_i^2} = 1.258/912.5 = 0.001378$

3. A 95% c.i. for β_{OLS} is $\hat{\beta}_{OLS} \pm t_{0.075,df} \sqrt{\hat{\text{Var}}\ \hat{\beta}}$.
   df = 70 - 1 = 69, $t_{0.975,69} \approx 1.994$, $\sqrt{\hat{\text{Var}}\ \hat{\beta}} = \sqrt{0.00138} = 0.0371$.
   Using the estimates: $1.678 \pm 1.994 \times 0.037 = (1.60, 1.75)$.
   Using the provided values: $1.60 \pm 1.994 \times \sqrt{0.00140} = (1.52, 1.67)$

4. $\hat{\beta}_{OLS}$ is the value of β that minimizes $\text{SSE} = \sum (Y_i - (4 + \beta X_i))^2$.
   Setting the derivative to 0 and solving gives: $2 \sum (Y_i - 4 - \beta X_i)(-X_i) = 0$
   $\sum (Y_i - 4) = \hat{\beta}_{OLS} \sum X_i^2$, hence, $\hat{\beta}_{OLS} = \frac{\sum (Y_i - 4)}{\sum X_i^2}$.

   Var $\hat{\beta}_{OLS}$ can be obtained algebraically from $\hat{\beta}_{OLS}$ since Var $Y_i | X_i = \sigma^2$ and the X_i's are constants. A quicker derivation is to remember the matrix expression: $\text{Var } \hat{\boldsymbol{\beta}} = \sigma^2 (\boldsymbol{X'X})^{-1}$. Here the X matrix has only one column, so $\boldsymbol{X'X} = \sum X_i^2$, hence Var $\hat{\beta}_{OLS} = \sigma^2 / \sum X_i^2$. An unbiased estimate of σ^2 is the error mean square: $s^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-1}$. Plugging in gives the estimator $\hat{\text{Var}}\ \hat{\beta}_{OLS}$.

5. Since there are replicate values at at least one X value, the most general test of lack of fit is to compare the proposed regression to a means model (1-way ANOVA model). The regression has 1 parameter; the means model has 7. Using the model comparison approach,

$$\begin{aligned} F &= \frac{(SS_{regression} - SS_{meansmodel})/(change in d.f.)}{SS_{meansmodel}/(error d.f.)} \\ &= \frac{(86.909 - 82.561)/(7-1)}{82.561/(70-7)} \\ &= 0.724/1.31 = 0.55 \end{aligned}$$

   This is compared to quantiles of the F distribution with 6 and 63 d.f.. The F statistic is small, so $p > 0.5$. There is no evidence that the proposed regression does not fit the data. Or, there is no evidence of lack of fit.

   Note that it is not possible to claim that the proposed regression fits the data. Our test may not be very powerful.

6. Three of the following required for full credit.

Property	Holds with unequal variance?
Unbiased	Yes
Normally distributed	Yes
Minimum Variance	No
Var $= \sigma^2/\sum X_i^2$	No

7.

Property	Holds with unequal variance?
Unbiased	No
Proportional to Chi-square	No

8. Both approaches fit models with constant coefficient of variation.
Disadvantages:
approach a) more complex model, harder to implement,
residuals still give indication of unequal variances
approach b) Trend (E Y vs X) is no longer linear.
I believe a) is more appropriate because the trend appears linear. Log transformation (approach b) distorts the trend. This is apparent in the residual plot for b), which is not flat.

9. One reasonable model is

$$\begin{aligned} Y_{ijk} &= 4 + \beta_i X_{ijk} + \epsilon_{ijk} \\ \epsilon_{ijk} &\sim N(0, \sigma^2_{ijk}) \\ \sigma^2_{ijk} &= (4 + \beta_i X_{ijk})^2 \sigma^2 \end{aligned} \tag{1}$$

where i indexes the variety, β_i is the growth rate for variety i, j indexes the replicate within each variety, k is the day.

To test whether all varieties have the same slope, you compare the fit of the above model to that of

$$\text{E } Y_{ij} = 4 + \beta X_{ij}$$

a model with single slope. If the variances were constant, you could use the typical model comparison approach, comparing error SS, that leads to an F statistic.

It is tempting to apply this approach to weighted SS, but a few minutes reflection suggests a difficulty: the comparison only works when the two models have the same weights. To see this, consider two models:

$$\begin{aligned} 1)Y_i &= 4 + \beta X_i + \epsilon_i,\ \epsilon_i\ N(0, \sigma^2) \\ 2)Y_i &= 4 + \beta X_i + \epsilon_i,\ \epsilon_i\ N(0, 2\sigma^2) \end{aligned}$$

The second is the first with a weight of 2 for each observation. The two models fit equally well, but the weighted error SS for model 2 will be 1/2 that from model 1. Likelihood's can be compared, since the values of the weights are included in the likelihood.

Some reasonable approaches are:
a) use the same weights for both models, i.e. compare the full model to

$$Y_{ij} = 4 + \beta X_{ij} + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma^2_{ij}),\ \sigma^2_{ij} = (4 + \beta_i X_{ij})^2 \sigma^2$$

where the β_i in the variance function are estimated from the full model.
b) construct a Wald test of $\beta_1 = \beta_2 = \beta_3 = \beta_4$. That F statistic would be of the form

$$F = (C\beta)'(C'\Sigma C)^{-1}(C\beta)$$

for an appropriate contrast matrix, C, and Σ is the Variance-covariance matrix of the β_i's.
c) construct a likelihood ratio test.

10. No, observations are not independent. The data are now an example of repeated measures. Observations from the same plug are likely correlated across sampling times. If one plug grows faster than the variety average, the observed diameter of that plug is likely to be higher than the average at all sampling times.

11. The ANOVA table corresponds to a generalization of the independence model, equation 1. One approach to repeated measures data is to consider that each plug has a (random) slope. The variability between plugs is described by the plug(variety)*day term in the ANOVA table. This is a random effect. So, a reasonable model is

$$Y_{ijk} = 4 + (\beta_i + u_{ik}) X_{ij} + \epsilon_{ijk},\ \epsilon_{ij} \sim \text{indep } N(0, \sigma^2_{ijk}),\ u_{ik} \sim \text{indep } N(0, \sigma^2_b) \qquad (2)$$

where i denotes the variety, j denotes the sampling time, and k denotes the specific plug. β_i is the mean growth rate for a variety; u_{ik} is the random deviation of the growth rate for a specific plug and the variety mean. σ^2_b is the between-plug variance in the growth rates.

There are a couple of possible models for the error variance, σ^2_{ijk}. It is unclear whether 'constant coefficient of variation' applies to the marginal mean for a variety or a conditional mean for a specific plug given the random effect. Either approach is acceptable. The distribution of the random effects is unspecified. A normal distribution is customary, but a more general answer is certainly acceptable.

The simpler model is that the c.v. and hence the variances depend on the marginal mean. That model is

$$\sigma^2_{ijk} = (4 + \beta_i X_{ijk})^2 \sigma^2$$

The conditional model is identical except that the variance of ϵ_{ijk} includes u_{ij}.

$$\sigma^2_{ijk} = (4 + (\beta_i + u_{ik}) X_{ijk})^2 \sigma^2$$

The estimates of the mean growth rate for each variety are the β_i from the model.

A test of equality of growth rates could be constructed using any of the approaches given in the answer to part 9, except that the appropriate variances should be used throughout.

The effects of adult striped cucumber beetles on leaf development in Black Satin summer squash plants were examined in an experiment performed at the Burden Research Station in Baton Rouge, Louisiana. Plants at one of three stages of development; cotyledon (Stage 1), first true leaf (Stage 2) and third true leaf (Stage 3), were exposed to one of four densities of cucumber beetle infestation, 0, 5, 10, or 15 adult beetles per plant. A total of 120 different plants were used; 10 plants were randomly assigned to each of the 12 combinations of plant stage and beetle density, yielding a 3 x 4 factorial arrangement. The number of fully developed leaves on each plant was recorded at 25 days after the plant emerged from the soil. This is the beginning of the heavy vegetative growth period that precedes fruit production. Some plants did not survive through 25 days after emergence, and no results were recorded for those plants. This mortality resulted in imbalanced data. The mean number of fully developed leaves per plant and the number of surviving plants at 25 days after emergence are shown in the following table.

Table 1. Sample means

	Adult beetle density				
Stage	0	5	10	15	
1	$\bar{Y}_{11\cdot} = 8.50$ $n_{11} = 10$	$\bar{Y}_{12\cdot} = 7.89$ $n_{12} = 9$	$\bar{Y}_{13\cdot} = 7.33$ $n_{13} = 6$	$\bar{Y}_{14\cdot} = 4.00$ $n_{14} = 1$	$\bar{Y}_{1\cdot\cdot} = 7.85$
2	$\bar{Y}_{21\cdot} = 9.20$ $n_{21} = 10$	$\bar{Y}_{22\cdot} = 8.80$ $n_{22} = 10$	$\bar{Y}_{23\cdot} = 7.56$ $n_{23} = 9$	$\bar{Y}_{24\cdot} - 6.40$ $n_{24} = 5$	$\bar{Y}_{2\cdot\cdot} = 8.24$
3	$\bar{Y}_{31\cdot} = 8.80$ $n_{31} = 10$	$\bar{Y}_{32\cdot} = 8.60$ $n_{32} = 10$	$\bar{Y}_{33\cdot} = 9.00$ $n_{33} = 9$	$\bar{Y}_{34\cdot} = 9.44$ $n_{34} = 9$	$\bar{Y}_{3\cdot\cdot} = 8.95$
	$\bar{Y}_{\cdot 1\cdot} = 8.83$	$\bar{Y}_{\cdot 2\cdot} = 8.45$	$\bar{Y}_{\cdot 3\cdot} = 8.04$	$\bar{Y}_{\cdot 4\cdot} = 8.06$	$\bar{Y}_{\cdots} = 8.41$

To be more explicit in discussing the experiment and potential models, some additional notation is needed. Let Y_{ijk} denote the observed number of fully developed leaves at 25 days after emergence for the k-th plant for which the j-th beetle density was introduced at the i-th stage of development, provided that the plant survived for at least 25 days after emergence. Let S_i denote an effect of the stage (cotyledon, first true leaf, and third true leaf for $i = 1,2,3$, respectively) at which the plants were first exposed to the beetles. Let D_j denote an effect of beetle density (0, 5, 10, 15 beetles for $j = 1,2,3,4$, respectively). Let P_{ijk} denote a random effect for the k-th plant exposed to the j-th density at the i-th stage of development. Then $k = 1,2,\ldots,n_{ij}$, where n_{ij} is the number of plants out of 10 that were alive at 25 days after emergence. A possible model is

$$Y_{ijk} = \mu + S_i \_ D_j + SD_{ij} + P_{ijk} \qquad \text{(model 1)}$$

Assume that $P_{ijk} \sim NID\left(0, \sigma_p^2\right)$.

The GLM procedure in SAS was used to compute Type I, Type II, and Type III sums of squares. These are shown in Table 2.

Table 2. Sums of Squares

Source of variation	df	Type I SS	MS	F-value	P-value
Stages	2	20.28	10.14	4.33	.0162
Densities	3	16.03	5.34	2.28	.0851
stage×density	6	41.80	6.97	2.97	.0110
Error	86	201.57	2.34		

Source of variation	df	Type II SS	MS	F-value	P-value
Stages	2	25.86	12.93	5.52	.0056
Densities	3	16.03	5.34	2.28	.0851
stage×density	6	41.80	6.97	2.97	.0110
Error	86	201.57	2.34		

Source of variation	df	Type III SS	MS	F-value	P-value
Stages	2	41.37	20.68	8.82	.0003
Densities	3	31.14	10.38	4.43	.0061
Stage×density	6	41.80	6.97	2.97	.0110
Error	86	201.57	2.34		

1. Describe the set of estimable functions of the interaction parameters $\left\{SD_{ij};\ i = 1,2,3\ ;\ j = 1,2,3,4\right\}$.

2. The value of the F-test for stage×density interaction is 2.97 with (6,97) degrees of freedom and p-value = .001 regardless of whether Type I, II or III sums of squares are used. Using the parameters in model 1, carefully state the null hypothesis and the alternative hypothesis for this F-test.

3. In a few sentences, give a description of the interaction between the stage of plant development when the beetles are introduced and the beetle density with respect to leaf development. Do <u>not</u> use any formulas or parameters in this description.

4. The researchers want to test the null hypothesis that the trends in the mean numbers of fully developed leaves for plants that are alive at 25 days after emergence are the same for stages 1 and 2. Show how to construct an appropriate test. You should <u>not</u> try to evaluate the test statistic. Just show how to construct it and give appropriate degrees of freedom.

5. Describe the null hypothesis corresponding to the Type III sum of squares F-test for "stages" (F = 8.82). How does this differ from the null hypotheses for the F-tests corresponding to the Type I and Type II sums of squares F-tests for "stages"?

6. One researcher commented that the sample means in Table 1 and the sums of squares in Table 2 ignore the information in the plants that died within the first 25 days after emergence. A suggestion is made to enter an observation of zero leaves for each plant that died within the first 25 days after emergence. For the 10 plants introduced to 15 beetles at stage 1, for example, nine zeros would be included in the data set for the nine plants that died. This would change the sample mean for the number of fully developed leaves per plant from 4.0 to 0.4. Similar changes would occur for other stage ×density combinations where plants died. Is this a good suggestion? If you agree, explain why it is a good suggestion. If you disagree, explain why it is a poor suggestion and offer a better alternative for dealing with plant mortality.

7. All of the plants that were alive at 25 days after emergence were also alive at 37 days after emergence. For each of those plants, numbers of fully developed leaves were recorded at 25, 29, 33, and 37 days after emergence. Although one could do a separate analysis of leaf development at each time point, the researchers proposed the following model for the combined data set:

$$Y_{ijkm} = \mu + S_i + D_j + SD_{ij} + P_{ijk} + T_m + ST_{im} + DT_{jm} + SDT_{ijm} + \varepsilon_{ijkm} \quad \text{(model 2)}$$

where $P_{ijk} \sim NID\left(0, \sigma_p^2\right)$ are random plant effects, $\varepsilon_{ijkm} \sim NID\left(0, \sigma_\varepsilon^2\right)$ are random errors, any P_{ijk} is independent of any ε_{ijkm}, and

T_m is a time effect $m = 1,2,3,4$

S_i is a stage effect $i = 1,2,3$

D_j is a beetle density effect $j = 1,2,3,4$

Type III sums of squares are shown below with formulas for expected mean squares. In this table the $Q(\quad)$ notation indicates a quadratic form involving the indicated effects.

Source of variation	Df	Type III SS	Type III ∈ (MS)
stages	2	313.831	$\sigma_\varepsilon^2 + 4\sigma_p^2 + Q(S, SD, ST, SDT)$
densities	3	361.878	$\sigma_\varepsilon^2 + 4\sigma_p^2 + Q(D, SD, DT, SDT)$
stage * density	6	280.562	$\sigma_\varepsilon^2 + 4\sigma_p^2 + Q(SD, SDT)$
plants(stage, density)	86	1203.453	$\sigma_\varepsilon^2 + 4\sigma_p^2$
time	3	1278.205	$\sigma_\varepsilon^2 + Q(T, ST, DT, SDT)$
stage * time	6	11.552	$\sigma_\varepsilon^2 + Q(ST, SDT)$
density * time	9	25.194	$\sigma_\varepsilon^2 + Q(DT, SDT)$
stage * density * time	15	38.838	$\sigma_\varepsilon^2 + Q(SDT)$
error	258	303.981	σ_ε^2

Estimate standard errors for the following linear combinations of sample means.

(a) $\left(\overline{Y}_{24\cdot 4} - \overline{Y}_{21\cdot 4}\right) - \left(\overline{Y}_{34\cdot 4} - \overline{Y}_{31\cdot 4}\right)$

where $\overline{Y}_{ij\cdot 4} = \dfrac{1}{n_{ij4}} \sum_{k-1}^{n_{ij4}} Y_{ijk4}$ and

$n_{214} = 10,\ n_{244} = 5,\ n_{314} = 10,\ n_{344} = 9$ are the numbers of plants alive at 37 days.

(b) $\left(\overline{Y}_{14\cdot4}-\overline{Y}_{11\cdot4}\right)-\left(\overline{Y}_{14\cdot3}-\overline{Y}_{11\cdot3}\right)$

where $n_{144}=1$, $n_{114}=10$, $n_{143}=1$ and $n_{113}=10$ are the numbers of plants used to construct the four sample means.

(8) For the data in part (g), one could also consider models of the form

$$\begin{bmatrix} Y_{ijk1} \\ Y_{ijk2} \\ Y_{ijk3} \\ Y_{ijk4} \end{bmatrix} = \begin{bmatrix} \mu+S_i+D_j+SD_{ij}+T_1+ST_{i1}+DT_{j1}+SDT_{ij1} \\ \mu+S_i+D_j+SD_{ij}+T_2+ST_{i2}+DT_{j2}+SDT_{ij2} \\ \mu+S_i+D_j+SD_{ij}+T_3+ST_{i3}+DT_{j3}+SDT_{ij3} \\ \mu+S_i+D_j+SD_{ij}+T_4+ST_{i4}+DT_{j4}+SDT_{ij4} \end{bmatrix} + \begin{bmatrix} P_{ijk} \\ P_{ijk} \\ P_{ijk} \\ P_{ijk} \end{bmatrix} + \begin{bmatrix} e_{ijk1} \\ e_{ijk2} \\ e_{ijk3} \\ e_{ijk4} \end{bmatrix}$$

where $P_{ijk} \sim NID\left(0,\sigma_p^2\right)$ and $\underset{\sim}{e}_{ijk} = \begin{bmatrix} e_{ijk1} \\ e_{ijk2} \\ e_{ijk3} \\ e_{ijk4} \end{bmatrix} \sim NID\left(\underset{\sim}{0},\ V\right)$

and any P_{ijk} is independent of any $\underset{\sim}{e}_{ijk}$. REML estimation was used to estimate σ_p^2 and the elements of V for different forms of V. The values of the REML log-likelihoods are given in the following table:

Model for V	REML log-likelihood
Compound Symmetry model	-670.15
Equal correlations and heterogeneous variances	-642.35
AR(1) model with homogeneous variances	-667.25
Toeplitz model with heterogeneous variances	-640.7
Unstructured covariance model	-636.55

What can you conclude from this information?

(9) Are any of the four models considered in part (8) equivalent to the model considered in part (7)? If so, which one(s)?

(1) Any linear combination of the quantities $\mathrm{SD}_{ij} - \mathrm{SD}_{kj} - \mathrm{SD}_{i\ell} + \mathrm{SD}_{k\ell}$.

(2) The null hypothesis is that all quantities of the form $SD_{ij} - SD_{kj} - SD_{ie} + SD_{ke}$ are zero. The alternative is that at least one such quantity is non-zero.

(3) It would be wise to plot the sample means against beetle density with one set of points (or profile) for each of the three plant development stages, or you can examine the sample means. The profiles are approximately parallel for stages 1 and 2 with decreasing mean numbers of fully developed leaves as the beetle density increases. (Note that $\bar{Y}_{14\cdot}$ is based on only one observation.) The profile for stage 3 shows no decrease in the mean number of fully developed leaves as beetle density increases.

(4) Obtain a solution to the normal equations, e.g., $\underset{\sim}{b} = \left(X^T X\right)^{-} X^T \underset{\sim}{Y}$ where X is the 98 x 12 model matrix and $\underset{\sim}{Y}$ is the 98 x 1 vector of responses. Suppose the columns of X are arranged so that

$$\hat{b} = \left(\hat{\mu}, \hat{s}_1, \hat{s}_2, \hat{s}_3, \hat{D}_1, \hat{D}_2, \hat{D}_3, \hat{D}_4, SD_{11}, SD_{12}, SD_{13}, SD_{14}, SD_{21}, \ldots, SD_{33}, SD_{34}\right)^T$$

Since introducing no beetles at stage 1 is no different than introducing no beetles at stage 2, the mean response curves for the two stages exhibit the same trends across densities of beetles if the curves are identical. This can be tested by constructing

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

Since $C\underset{\sim}{\beta}$ is estimable, then $C\underset{\sim}{b}$ is the unique $b.l.u.e$ for $C\underset{\sim}{\beta}$. Reject the null hypothesis $H_o\!: C\underset{\sim}{\beta} = 0$ if

$$F = \frac{Y^T X\left(X^T X\right)^{-} C^T C\left(X^T X\right)^{-} X^T Y / 4}{Y^T\left(I - X\left(X^T X\right)^{-} X^T\right)Y / 86} > F_{(4,86).05} = 2.48$$

Alternatively, you could construct an F-test using contrasts of sample means. Which method is better?

If you interpreted the "trends are the same" to mean parallel response curves instead of identical mean response curves, you would need a different set of contrasts to specify the null hypothesis. There are an infinite number of equivalent choices. One possibility is to test the equivalence of linear, quadratic and cubic trends across densities. This is expressed as $H_o\colon C\beta = \underset{\sim}{0}$

where

$$C = \begin{bmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & -3 & -1 & 1 & 3 & 3 & 1 & -1 & -3 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & -1 & 3 & -3 & 1 & 1 & -3 & 3 & -1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Reject the null hypothesis $H_o\colon C\beta = \underset{\sim}{0}$ if

$$F = \frac{Y^T X \left(X^T X\right)^{-} C^T C \left(X^T X\right)^{-} X^T Y / 3}{Y^T \left(I - X\left(X^T X\right)^{-} X^T\right) Y / 86} > F_{(3,86)\,.05} = 2.72$$

(5) For type III sums of squares, the null hypothesis is

$$H_0 : \frac{1}{4}(\mu_{11} + \mu_{12} + \mu_{13} + \mu_{14}) = \frac{1}{4}(\mu_{21} + \mu_{22} + \mu_{23} + \mu_{24}) = \frac{1}{4}(\mu_{31} + \mu_{32} + \mu_{33} + \mu_{34})$$

where $\mu_{ij} = \mu + S_i + D_j + SD_{ij}$.

For type I sums of squares the null hypothesis is

$$H_0 : (\mu + S_1 + [10(D_1 + SD_{11}) + 9(D_2 + SD_{12}) + 6(D_3 + SD_{13}) + (D_4 + SD_{14})]/26)$$

$$= (\mu + S_2 + [10(D_2 + SD_{21}) + 10(D_2 + SD_{22}) + 9(D_3 + SD_{23}) + 5(D_{11} + SD_{24})]/34)$$

$$= (\mu + S_3 + [10(D_1 + SD_{31}) + 10(D_2 + SD_{32}) + 9(D_3 + SD_{33}) + 9(D_{11} + SD_{34})]/38)$$

(6) Discussion

(7) You can obtain method of moment estimates of variance components as follows:

$$\hat{\sigma}_{\epsilon}^{2} = \frac{303.981}{258} = 1.178$$

$$\hat{\sigma}_{P}^{2} = \frac{1}{4}\left(\frac{1203.453}{86} - \frac{303.981}{258}\right) = 3.204$$

These are also REML estimates in this case. Now, derive the variance of each linear combination of sample means.

(a) $Var\left(\overline{Y}_{24.4} - \overline{Y}_{21.4} - \overline{Y}_{34.4} + \overline{Y}_{31.4}\right) = \left(\sigma_{\epsilon}^{2} + \sigma_{P}^{2}\right)\left(\frac{1}{5} + \frac{1}{10} + \frac{1}{9} + \frac{1}{10}\right)$

and the standard error is

$$\sqrt{(1.178 + 3.204)\left(\frac{1}{5} + \frac{1}{10} + \frac{1}{9} + \frac{1}{10}\right)} = 1.50$$

(b) $Var\left(\overline{Y}_{24.4} - \overline{Y}_{11.4} - \overline{Y}_{14.3} + \overline{Y}_{11.3}\right) = 2.2\ \sigma_{\epsilon}^{2}$ and the standard error is $\sqrt{2.2\ (1.178)} = 1.61$

(8) Likelihood ratio tests can be used to compare some models, if one model can be obtained as a special case of a more general model. Values of the AIC or BIC criterion could be examined. Any of the listed models with heterogeneous variances provide a good fit to the data.

(9) Yes, the model with the compound symmetry covariance structure for the random errors is equivalent to the model in part (7).

PhD Preliminary Exam 2004

Methods Question III

Many problems involve what is sometimes called *compositional data*, which are data values that represent the proportions of a sample that belong to various categories. In one sample of sediment (e.g., from the bottom of a river), it might be recorded that 10% was sand, 25% was silt and the remainder was clay. In grading the condition of produce from a potential supplier, a restaurant buyer may record the composition of a sample as 40% excellent, 30% good, 17% fair, and 13% poor. Regardless of the problem, such data have the characteristic that values within a sample sum to 1.0.

An alternative structure that leads to similar data occurs in situations in which each of a number of discrete sampling units is placed into one of a number of mutually exclusive categories. This is common, for example, in studies that involve a number of species that occur in a group of organisms. In a sample of 50 organisms, 32 may be found to be "species A", 7 "species B", and 11 "species C". Here, the observed data values sum to the number of organisms examined.

One important problem that leads to data such as the example of *species composition* described immediately above involves sampling of the ocean to determine the proportion of marine fish that belong to various species groups, such as salmon, halibut, rockfish, lingcod, pollock, and so forth. These types of data are collected by research vessels either owned or contracted by the National Marine Fisheries Service to sample certain regions at particular times of the year. These vessels put out to sea, take a number of tows (or hauls), usually along a randomly selected transect within an area, and the number of fish belonging to a set of species categories are recorded. Scientists have a good idea of the total number of different species that should be caught in these research cruises (there may be 25 tows taken in one research cruise). Except in certain special studies, the entire haul contains too many fish to enumerate all of them. Thus, a sample is taken of the haul and the sample is then enumerated in species categories. For our purposes here, we will assume that

1. Hauls constitute a random sample of the area to be examined.

2. Samples from hauls are selected at random.

Note: while the hauls may constitute a random sample of the area, it is more difficult to claim this corresponds to a random sample of the fish under the water.

While the number of species categories of interest in these studies is typically around 120 to 150, we can consider a number of statistical issues that might be involved with such data in a reduced setting that is easier to deal with notationally, and we will use a much smaller number of categories below. Thus, consider a statistical analysis for a problem in which each of n organisms are recorded as belonging to one of $(k+1)$ categories, and define the random vector

$$\boldsymbol{X} \equiv (X_1, X_2, \ldots, X_{k+1})^T,$$

where, for $j = 1, \ldots, (k+1)$, $X_j \in \{0, 1, \ldots, n\}$ and $\sum X_j = n$.

A natural choice for modeling $\boldsymbol{X}$ is the multinomial distribution. Now define $\boldsymbol{Y} \equiv (Y_1, \ldots, Y_k)^T$, where $Y_j \equiv X_j$, $j = 1, \ldots, k$, and let $\boldsymbol{Y}$ have probability mass function (pmf),

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = K\, \theta_1^{y_1}\, \theta_2^{y_2} \ldots \theta_k^{y_k} \left(1 - \sum_{j=1}^{k} \theta_j\right)^{n-y_1-y_2-\ldots-y_k},$$

where

$$K \equiv \frac{n!}{y_1!\, y_2! \ldots\, y_k!\, (n - y_1 - y_2 - \ldots - y_k)!},$$

and

$$\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_k)^T.$$

1. What is the motivation for assigning the multinomial pmf given immediately above to $\boldsymbol{Y}$ rather than to $\boldsymbol{X}$?

2. Recall that univariate marginals from a multinomial are binomials, joint marginals are again multinomials, and conditionals are also multinomials. If $k = 3$, give the following in terms of n and $\boldsymbol{\theta}$. If you happen to remember these, that's fine (no need to show derivations). If you don't remember, derive them.

   (a) The expectations of $Y_1, \ldots, Y_3$.

   (b) The variances of $Y_1, \ldots, Y_3$.

   (c) The covariances of Y_i, Y_j for $i, j = 1, 2, 3$, $j \neq i$.

   Hint: to derive the covariances, recall that $var(X + Y) = var(X) + var(Y) - 2cov(X, Y)$ *and consider also the variance of the random variable defined as* $Z \equiv X + Y$.

3. Give the natural estimators of the expectations, variances, and covariances listed above. There is no need to show that these estimators have any properties or can be derived as maximum likelihood estimators or unbiased estimators and so forth. Just list the obvious choices.

4. Suppose for simplicity that $k = 3$ (so $k + 1 = 4$). Suppose further that a research cruise such as those described in the introduction resulted in data for 50 hauls and that a sample of 25 fish was taken from each haul. Let the species categories of interest be denoted as A, B, C, and D. Assume that the 25 fish were sampled independently. Consider the following subset of data from the 50 hauls:

Table 1. Partial data set for question 4.

Haul	A	B	C	D
1	2	14	7	2
2	4	13	6	2
3	2	12	8	3
4	3	15	5	2
5	1	17	7	0
6	5	11	6	3
7	1	16	7	1
8	3	14	5	3
9	1	15	7	2
10	2	18	2	3
11	5	16	4	0
12	2	12	10	1
13	0	11	9	5
14	3	15	5	2
15	1	16	7	1
16	3	16	6	0
17	1	18	4	2
18	4	13	8	0
19	1	14	8	2
20	6	11	6	2
21	3	19	2	1
22	1	12	10	2
23	2	14	7	2
24	2	13	10	0
25	2	13	10	0

The sum of values for each category over all 50 hauls was 121 of species A, 736 of species B, 323 of species C and 70 of species D. Suppose we have interest in estimating the multinomial parameters θ_1, θ_2 and θ_3.

(a) If you were to estimate the parameters for each haul individually, what might make you uneasy, given what you have seen in the partial data set tabled above? Would you be willing to calculate confidence intervals for the parameters for each haul?

(b) The obvious solution is to consider each haul as representing a multinomial vector $\boldsymbol{Y}_h$, $h = 1, \ldots, 50$, such that these vectors are independent and identically distributed. Then take $Y_j = \sum_h Y_{j,h}$. Although formal tests might be available, we have no reference books today. Based only on visual examination of the partial data set in Table 1, do you think that this assumption is likely to be reasonable, or not? If we make this assumption, what parameters remain to be estimated?

(c) Using the summary information for all 50 hauls given immediately after Table 1, estimate the expected values, variances, and covariances from question 2 (use the estimators you gave in question 3).

Taking all of the data from the complete set of 50 hauls, a scatterplot of values $Y_{3,h}$ against $Y_{1,h}$ is presented in Figure 1 as an example.

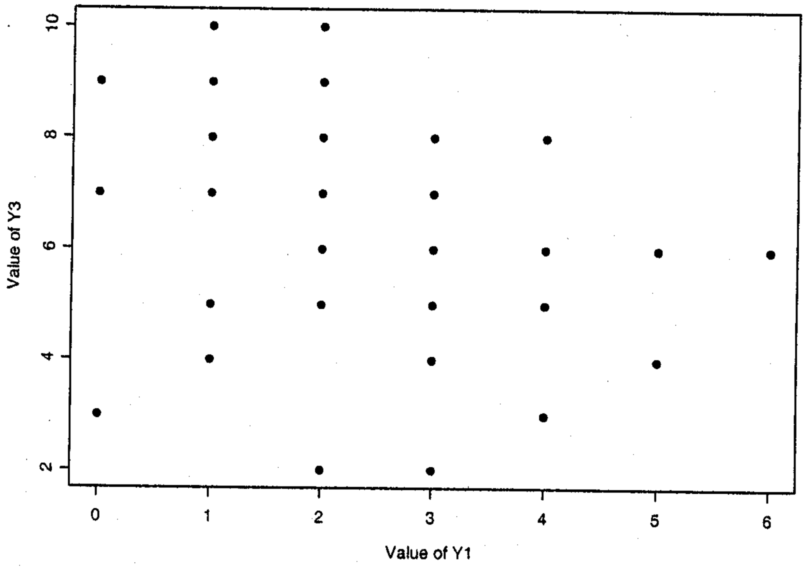


Figure 1. Scatterplot of $Y_{3,h}$ versus $Y_{1,h}$ from the data set of question 4.

The complete sample correlation matrix for Y_1, Y_2, and Y_3 was

$$\begin{pmatrix} 1 & -0.41 & -0.24 \\ -0.41 & 1 & -0.64 \\ -0.24 & -0.64 & 1 \end{pmatrix}$$

You can see the scatterplot that corresponds to the estimate −0.24 in Figure 1.

(d) The sample correlation matrix immediately above should be similar to your estimated values from 4c. Does similarity of the above correlation matrix with the matrix you estimated in question 4c support the assumption of *iid* multinomial vectors across hauls, or is it irrelevant to that assumption?

5. Now, another research cruise similar to that described in question 4 resulted in a different data set for the same four species (A, B, C, and D). A *portion of this data set* is given in the following table.

Table 2. Partial data set for question 5.

Haul	A	B	C	D	Haul	A	B	C	D
1	16	8	0	1	14	5	16	0	4
2	5	9	7	4	15	7	7	4	7
3	0	8	1	16	16	5	4	2	14
4	4	0	14	7	17	12	10	0	3
5	8	4	8	5	18	10	11	0	4
6	0	1	6	18	19	9	4	2	10
7	8	13	3	1	20	9	12	1	3
8	2	3	1	19	21	8	16	1	0
9	1	15	0	9	22	3	7	0	15
10	5	12	4	4	23	4	13	5	3
11	3	20	0	2	24	3	16	2	4
12	5	20	0	0	25	0	3	6	16
13	3	17	1	4					

The sum of values for each category over all 50 hauls was 270 of species A, 500 of species B, 138 of species C and 342 of species D. Suppose we have interest in estimating the multinomial parameters θ_1, θ_2 and θ_3. A scatterplot of the same variables as given in Figure 1 for the data of question 4 is presented in Figure 2 for the data of this question

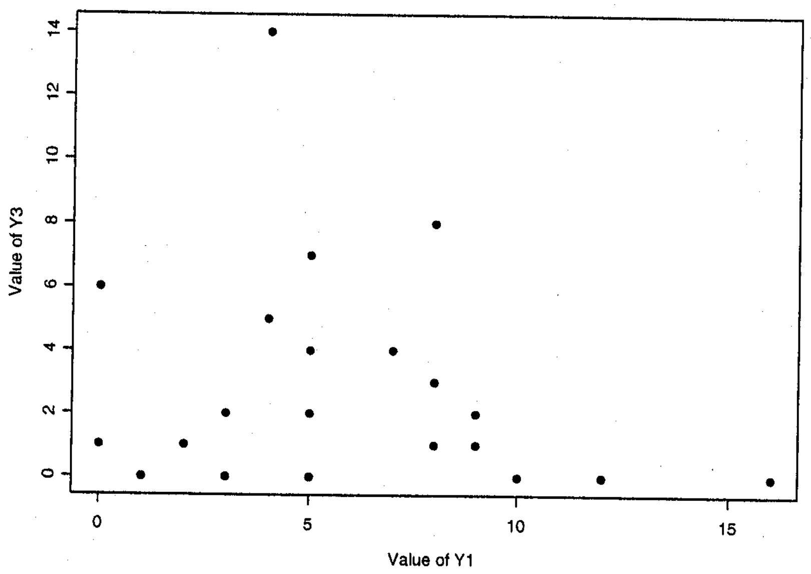


Figure 2. Scatterplot of $Y_{3,h}$ versus $Y_{1,h}$ from the data set of question 5.

The sample correlation matrix for the complete data set of 50 hauls is,

$$\begin{pmatrix} 1 & 0.06 & -0.21 \\ 0.06 & 1 & -0.55 \\ -0.21 & -0.55 & 1 \end{pmatrix}$$

Having seen a report you wrote describing the analysis of the data from question 4, a statistician with the National Marine Fisheries Service decides to do the same analysis with these data.

For your answer to question 5, comment on this plan, presenting any empirical evidence for your conclusions you feel is appropriate. Recall the various parts of question 4 when constructing your comment.

6. Finally, a third cruise similar to those of questions 4 and 5 resulted in a data set, a portion of which is given immediately below.

Table 3. Partial data set for question 6.

Haul	A	B	C	D	Haul	A	B	C	D
1	1	11	2	11	14	8	4	7	6
2	3	1	0	21	15	8	2	3	12
3	1	7	0	17	16	11	9	5	0
4	0	21	1	3	17	7	7	3	8
5	12	5	6	2	18	6	13	3	3
6	13	10	1	1	19	8	3	4	10
7	9	5	7	4	20	6	6	3	10
8	9	6	3	7	21	2	12	2	9
9	1	10	0	14	22	2	19	2	2
10	3	12	2	8	23	3	10	0	12
11	3	17	5	0	24	7	15	0	3
12	3	6	3	13	25	14	4	2	5
13	12	2	2	9					

The sum of values for each category over all 50 hauls was 285 of species A, 398 of species B, 144 of species C and 423 of species D. Interest is again in estimating the multinomial parameters θ_1, θ_2 and θ_3. A scatterplot of the same variables as given in questions 4 and 5 is presented in Figure 3 for the data of this question.

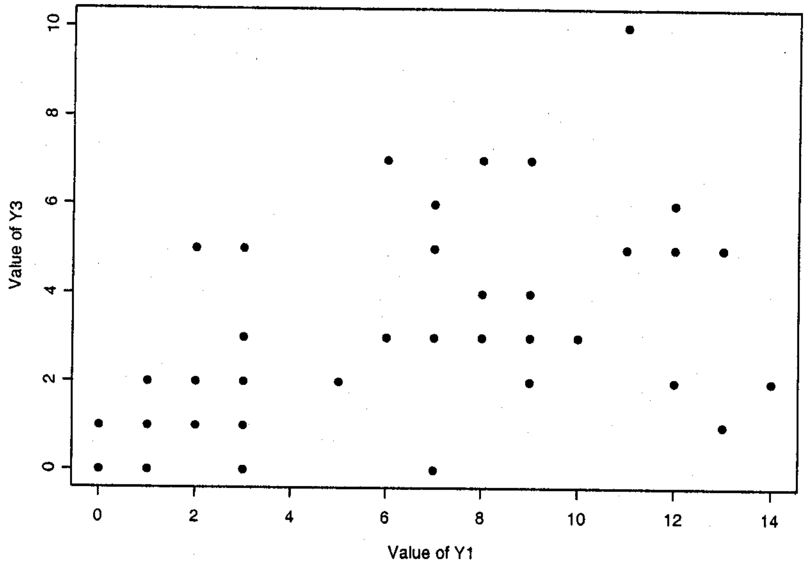


Figure 3. Scatterplot of $Y_{3,h}$ versus $Y_{1,h}$ from the data set of question 6.

The sample correlation matrix for the complete data set of 50 hauls is

$$\begin{pmatrix} 1 & -0.43 & 0.56 \\ -0.43 & 1 & -0.42 \\ 0.56 & -0.42 & 1 \end{pmatrix}$$

In the same way that you did for question 5, comment on a plan to conduct an analysis based on the assumption of *iid* multinomial vectors across hauls. Is there anything different here than for the situations of questions 4 or 5? Again present any empirical evidence for your conclusions you feel is appropriate, and recall the various parts of question 4 when constructing your comment.

7. In real studies of the type under discussion it is not possible to actually take a random sample of a fixed number of fish in a haul. Rather, what are taken are called "basket samples". A basket sample is a sample of roughly 500 kg of

weight from the mass of fish brought on board from a haul. The procedure is assumed to give a random sample of fish from the haul, but it does not contain a fixed number of fish.

It has been suggested that, in light of this fact, a more appropriate model for species composition in a sample would consist of a random variable for the number of fish in a sample and, conditional on this number, a multinomial for the count in each species group. It has also been suggested that a Poisson distribution might be appropriate for the total number of fish in a sample.

(a) Using N_h to denote the total number of fish in a sample from haul h, and $X_{1,h}$, $X_{2,h}$, $X_{3,h}$ and $X_{4,h}$ the number of fish in four species categories, formulate a model in which N_h has a Poisson distribution with parameter λ and, given that $N_h = n_h$, the four counts are multinomial. Assume all quantities are independent across hauls.

(b) Find the joint marginal distribution of $(X_{1,h}, X_{2,h}, X_{3,h}, X_{4,h})^T$ that results from this model. Find the univariate marginal distributions of each $X_{j,h}; j = 1, 2, 3, 4$.

(c) Could this model be used to address any of the issues that you identified in the previous questions? Consider, in particular, question 1, question 5, and question 6.

Outline of Answer

1. Number of free random variables is k not $k+1$. Recognize the bounded sum constraint on $\boldsymbol{X}$.

2. Standard.

   (a) $E(Y_j) = n\theta_j$

   (b) $var(Y_j) = n\theta_j(1-\theta_j)$

   (c) $cov(Y_j, Y_k) = -n\theta_j\theta_k$

3. The obvious estimators are $\hat{\theta}_j = Y_j/n$ and estimates of $E(Y_j)$, $var(Y_j)$, and $cov(Y_j, Y_k)$ as above with substitution of $\hat{\theta}_j$ for θ_j.

4. (a) The presence of 0 values in records for individual hauls causes potential problems (estimates of 0 with 0 variance).

   (b) Examination of the data indicates a remarkable degree of consistency across hauls. Now want to estimate θ_1, θ_2, θ_3, which are the common parameters for each haul.

   (c) Should get

   $$\hat{\theta}_1 = 0.097, \quad \hat{\theta}_2 = 0.589, \quad \hat{\theta}_3 = 0.258$$

   Estimated correlation matrix from the aggregated multinomial is

   $$\begin{pmatrix} 1 & -0.39 & -0.19 \\ -0.39 & 1 & -0.70 \\ -0.19 & -0.70 & 1 \end{pmatrix}$$

   (d) The sample correlation matrix is quite similar to this estimate based on the aggregated data. Yes, this supports the assumption of *iid* multinomial vectors across hauls.

5. The answer to question 5 should include:

   - The partial data set given exhibits the occurrence of 0 values just like the data set of question 4, but lacks the consistency of those data. Here, the 0 values appear for different species groups in different hauls, and there is greater variability in values for the same species group across hauls.
   - Estimates here are

     $$\hat{\theta}_1 = 0.216, \quad \hat{\theta}_2 = 0.400, \quad \hat{\theta}_3 = 0.110$$

     Estimated correlation matrix is

     $$\begin{pmatrix} 1 & -0.43 & -0.18 \\ -0.43 & 1 & -0.29 \\ -0.18 & -0.29 & 1 \end{pmatrix}$$

     The sample correlation matrix from this question is not as close to this as was the case in question 4, particularly for $cov(Y_1, Y_2)$ which is essentially zero in the sample correlations.
   - Putting this together, one should question the assumption of *iid* multinomial vectors among hauls. The italicized note on page 2 of the introduction has relevance here. Specifically, one might surmise that, although hauls are randomly selected on a map, fish move, habitat choices for different species groups might differ, and so forth. That is, there are any number of biological reasons that the distributions of species groups may not be identically distributed among hauls.

6. The answer to question 6 should include:

   - The partial data set does include some 0 values and is more similar to that of question 5 than question 4.

- Estimates here are

$$\hat{\theta}_1 = 0.228, \quad \hat{\theta}_2 = 0.318, \quad \hat{\theta}_3 = 0.115$$

  Estimated correlation matrix is

$$\begin{pmatrix} 1 & -0.37 & -0.20 \\ -0.37 & 1 & -0.25 \\ -0.20 & -0.25 & 1 \end{pmatrix}$$

  There is a dramatic difference between this estimated correlation matrix and what results from the sample correlations given in the question. In particular, $\hat{cor}(Y_1, Y_3) = 0.56$ there while it is -0.20 here. In addition, the scatterplot of observed values Y_3 versus Y_1 reflects the positive sample correlation. If the *iid* multinomial model is appropriate here, this correlation should be negative.

- Here, one might question not only the identical distribution assumption, but whether it would be appropriate to assume independence among species groups. The negative correlation in a multinomial is produced by the bounded sum constraint. These data suggest that Y_1 and Y_3 tend to be both small or large, within the bounded sum condition. To get something like this we would need not only non-identically distributed multinomials, but ones in which there is a positive relation between $\theta_{1,h}$ and $\theta_{3,h}$ across hauls.

7. The answer to question 7 includes some derivations, and recognizing the connections among these derived results and pertinent aspects of the models in previous questions.

   (a) Write down multinomial and Poisson distributions in an organized fashion.

(b) One should find that the univariate marginals of the $X_{j,h}$ are Poisson with parameters $\lambda\theta_j$. The joint marginal of $\boldsymbol{X}$ is in the form of a product of these univariate marginals. From this, they should realize that the marginal distributions of the X_j are independent Poissons.

(c) Connections with questions 1, 5, and 6 are:

- Since the X_j are marginally independent, the connection with question 1 is that it now makes sense to model $(X_1, X_2, X_3, X_4)^T$ rather than $(Y_1, Y_2, Y_3)^T$. That is, there is no bounded sum constraint in the marginal model.
- Since Poisson pmfs have support that includes 0 and are additive, the problems identified in question 5 are nicely addressed with this model. Specifically, one could estimate from any group of hauls, regardless of whether there are zeros or not. One could estimate from the total sum under the knowledge that what is being estimated is the sum of non-iid random variables.
- But, since the marginal Poisson distributions for the X_j are independent, this model does not deal with the correlation suggested in question 6, since that correlation comes from other than the bounded sum property of multinomials.

Consider the following two-way crossed classification model:

$$y_{ijk} = \mu + \alpha_i + \gamma_j + \epsilon_{ijk}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, K,$$

where $I \geq 2$, $J \geq 2$, and $K \geq 2$; μ, the α_i's and the γ_j's are unknown fixed parameters; the y_{ijk}'s are observable variables; the ϵ_{ijk}'s are unobservable and independently distributed as $N(0, \sigma^2)$, where σ^2 is an unknown positive parameter.

Let

$$\underline{\beta} = (\mu, \alpha_1, \ldots, \alpha_I, \gamma_1, \ldots, \gamma_J)'$$

and

$$\overline{y}_{i..} = \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} y_{ijk}, \quad \overline{y}_{.j.} = \frac{1}{IK} \sum_{i=1}^{I} \sum_{k=1}^{K} y_{ijk}, \quad \overline{y}_{...} = \frac{1}{IJK} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} y_{ijk}.$$

Note: You may use matrices in your derivations, but no matrix should appear in your final answers to the following questions except for part (j). Note that δ given below is such that $0 < \delta < 1$.

(a) Show that $\mu + \alpha_i + \gamma_j$ is an estimable function and its best linear unbiased estimator (BLUE) is given by

$$\overline{y}_{i..} + \overline{y}_{.j.} - \overline{y}_{...}.$$

(b) Identify all estimable functions $\underline{\lambda}'\underline{\beta}$ and give their BLUE's.

(c) Show that the BLUE of $\alpha_1 - \alpha_2$ is $\overline{y}_{1..} - \overline{y}_{2..}$ and the distribution of this BLUE is normal with mean $\alpha_1 - \alpha_2$ and variance $2\sigma^2/(JK)$.

(d) Give an unbiased estimator, $\widehat{\sigma}^2$, of σ^2.

(e) Find a $(1 - \delta)$ confidence interval for $\alpha_1 - \alpha_2$.

(f) Show that the BLUE of $\alpha_1 - \alpha_2$ and the BLUE of $\gamma_1 - \gamma_2$ are independent.

(g) Find a $(1 - \delta)$ confidence set for the ratio $(\alpha_1 - \alpha_2)/(\gamma_1 - \gamma_2)$.

(h) Give confidence intervals for the differences $\alpha_i - \alpha_l$ $(i, l = 1, \ldots, I, \ i \neq l)$ such that the probability of simultaneous coverage is **exactly** $1 - \delta$.

(i) Derive a size-δ test of the null hypothesis $H_0 : \alpha_1 = \cdots = \alpha_I$ versus the alternative hypothesis H_a: not H_0.

(j) Suppose now that the γ_j's are independently distributed as $N(0, \sigma_\gamma^2)$ and are independent of the ϵ_{ijk}'s, where σ_γ^2 is an unknown positive parameter. All other assumptions are the same as above. Show that, under this new model, the simple least squares estimator of any estimable function is also a BLUE of this function.

Ph.D. Prelim Exam Solutions. Spring 2004 Linear Models

(a) Since $E(y_{ijk}) = \mu + \alpha_i + \gamma_j$, then $\mu+\alpha_i+\gamma_j$ is estimable. The BLUE of $\mu+\alpha_i+\gamma_j$ is its least squares estimator. Consider minimizing $\sum_{i,j,k}(y_{ijk} - \mu - \alpha_i - \gamma_j)^2$ with respect to β. This leads to the following normal equations:

$$\bar{y}_{\cdots} - \mu - \bar{\alpha} - \bar{\gamma} = 0 \quad (\text{where } \bar{\alpha} = \frac{1}{I}\sum_{i=1}^{I}\alpha_i,\ \bar{\gamma} = \frac{1}{J}\sum_{j=1}^{J}\gamma_j)$$

$$\bar{y}_{i\cdot\cdot} - \mu - \alpha_i - \bar{\gamma} = 0, \quad i = 1, \cdots, I$$

$$\bar{y}_{\cdot j\cdot} - \mu - \bar{\alpha} - \gamma_j = 0, \quad j = 1, \cdots, J.$$

One solution to these equations is given by

$$\hat{\mu} = \bar{y}_{\cdots}, \quad \hat{\alpha}_i = \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}, \ i = 1, \cdots, I, \ \hat{\gamma}_j = \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}, \ j = 1, \cdots, J.$$

Thus, the BLUE of $\mu+\alpha_i+\gamma_j$ is $\bar{y}_{i\cdot\cdot} + \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}$.

(b) Note that $\underline{\lambda}'\beta$ is estimable if and only if, for some constants a_{ijk}'s, $\underline{\lambda}'\beta = E(\sum_{i,j,k} a_{ijk}\, y_{ijk})$, or

$\underline{\lambda}'\beta = \sum_{i,j} b_{ij}(\mu + \alpha_i + \gamma_j)$, for some constants b_{ij}'s.

Thus all estimable functions are given by $\sum_{i,j} b_{ij}(\mu+\alpha_i+\gamma_j)$ (where b_{ij}'s are any given constants) and their BLUE's are $\sum_{i,j} b_{ij}(\bar{y}_{i\cdot\cdot} + \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})$.

(c) By part (a), the BLUE of $\alpha_1-\alpha_2$ is its LSE $\bar{y}_{1..}-\bar{y}_{2..}$. Since y_{ijk}'s are independent and normally distributed, $\bar{y}_{1..}-\bar{y}_{2..}$ is normal, and its mean is $\alpha_1-\alpha_2$ and its variance is

$$Var(\bar{y}_{1..}-\bar{y}_{2..}) = Var(\bar{y}_{1..}) + Var(\bar{y}_{2..}) = \frac{2}{JK}\sigma^2.$$

(d) An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{IJK-I-J+1}\sum_{i,j,k}(y_{ijk}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2.$$

(e) By (c) and (d), a $(1-\delta)$ confidence interval for $\alpha_1-\alpha_2$ is

$$\bar{y}_{1..}-\bar{y}_{2..} \pm t_{\delta/2:(IJK-I-J+1)}\sqrt{\frac{2}{JK}}\cdot\hat{\sigma},$$

where $t_{\delta/2:(IJK-IJ+1)}$ is the upper $(\delta/2)$ point of a t distribution with $(IJK-I-J+1)$ degrees of freedom.

(f) $Cov(\bar{y}_{1..}-\bar{y}_{2..}, \bar{y}_{.1.}-\bar{y}_{.2.}) = \sum_{i=1}^{2}\sum_{j=1}^{2}(-1)^{i+j}Cov(\bar{y}_{i..}, \bar{y}_{.j.})$.

Since $Cov(\bar{y}_{i..}, \bar{y}_{.j.}) = cov\left(\frac{1}{JK}\sum_{j,k} y_{ijk}, \frac{1}{IK}\sum_{i,l} y_{ijl}\right)$

$$= \frac{1}{IJK^2}\sum_{k=1}^{K} Var(y_{ijk}) = \frac{\sigma^2}{IJK},$$ then

$Cov(\bar{y}_{1..}-\bar{y}_{2..}, \bar{y}_{.1.}-\bar{y}_{.2.}) = 0$. On the other hand, the joint distribution of $(\bar{y}_{1..}-\bar{y}_{2..}, \bar{y}_{.1.}-\bar{y}_{.2.})$ is multivariate normal. Thus, $\bar{y}_{1..}-\bar{y}_{2..}$ and $\bar{y}_{.1.}-\bar{y}_{.2.}$ are independent.

(g) Let $\theta = (\alpha_1 - \alpha_2)/(\gamma_1 - \gamma_2)$. Then $(\alpha_1 - \alpha_2) - \theta(\gamma_1 - \gamma_2) = 0$, and $(\bar{y}_{1..} - \bar{y}_{2..}) - \theta(\bar{y}_{.1.} - \bar{y}_{.2.}) \sim N(0, (\frac{1}{JK} + \frac{\theta^2}{IK}) \cdot 2\sigma^2)$ (by parts (c) and (f)). Thus

$$\frac{(\bar{y}_{1..} - \bar{y}_{2..}) - \theta(\bar{y}_{.1.} - \bar{y}_{.2.})}{\hat{\sigma} \cdot \sqrt{\frac{2}{JK} + \frac{2\theta^2}{IK}}} \sim t_{IJK-I-J+1}.$$

Hence a $(1-\delta)$ confidence interval set for $\theta = \frac{\alpha_1 - \alpha_2}{\gamma_1 - \gamma_2}$ is given by

$$\left\{\theta : \left|(\bar{y}_{1..} - \bar{y}_{2..}) - \theta(\bar{y}_{.1.} - \bar{y}_{.2.})\right| < t_{\delta/2 : IJK-I-J+1} \cdot \hat{\sigma} \sqrt{\frac{2}{JK} + \frac{2\theta^2}{IK}}\right\}.$$

(h) Note that $E(\bar{y}_{i..}) = \mu + \bar{\gamma} + \alpha_i$, and $\bar{y}_{i..}$ is the BLUE of $\mu + \bar{\gamma} + \alpha_i$. Also $\alpha_i - \alpha_l = (\mu + \bar{\gamma} + \alpha_i) - (\mu + \bar{\gamma} + \alpha_l)$. Thus, Tukey's method applies and we have the following confidence intervals for $(\alpha_i - \alpha_l)$'s with exact coverage probability of $(1-\delta)$:

$$(\bar{y}_{i..} - \bar{y}_{l..}) \pm (q^*_{\delta : I, IJK-I-J+1}) \cdot \sqrt{\frac{1}{JK}} \cdot \hat{\sigma},$$

where $\hat{\sigma}$ is given in part (d).

(i) Under H_0, we have $y_{ijk} = \mu + \alpha_i + \gamma_j + \varepsilon_{ijk}$, which is a one-way model. The fitted values for y_{ijk} are $\bar{y}_{\cdot j \cdot}$. Thus a size-δ F test of H_0 vs H_a rejects H_0 if

$$F = \frac{(SSR_f - SSR_r)/(I-1)}{SSE_f/(IJK-I-J+1)} > F_{\delta: I-1, IJK-I-J+1}$$

where

$$SSR_f = \sum_{i,j,k} (y_{i\cdot\cdot} + y_{\cdot j \cdot} - \bar{y}_{\cdot\cdot\cdot})^2 = K \sum_{i,j} (\bar{y}_{i\cdot\cdot} + \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot\cdot})^2,$$

$$SSR_r = IK \cdot \sum_{j=1}^{J} \bar{y}_{\cdot j \cdot}^2,$$

$$SSE_f = \sum_{i,j,k} (y_{ijk} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot\cdot})^2,$$

and $F_{\delta: I-1, IJK-I-J+1}$ is the upper δ point of an F distribution with $(I-1, IJK-I-J+1)$ degrees of freedom.

(j) Let $\underline{y}_{ij} = (y_{ij1}, \cdots, y_{ijK})'$, $\underline{1}_K$ be a $K\times 1$ vector of 1's, $\underline{\beta} = (\mu, \alpha_1, \cdots, \alpha_I)'$, $\underline{\gamma} = (\gamma_1, \cdots, \gamma_J)'$, and write the new model as

$$\underline{y} = X\underline{\beta} + Z\underline{\gamma} + \underline{\varepsilon},$$

where

$$\underline{y} = \begin{pmatrix} \underline{y}_{11} \\ \vdots \\ \underline{y}_{1J} \\ \underline{y}_{21} \\ \vdots \\ \underline{y}_{2J} \\ \vdots \\ \underline{y}_{I1} \\ \vdots \\ \underline{y}_{IJ} \end{pmatrix}, \quad X = \begin{pmatrix} \underline{1}_{JK} & \underline{1}_{JK} & & \\ \vdots & & \ddots & \\ \vdots & & & \ddots \\ \underline{1}_{JK} & & & \underline{1}_{JK} \end{pmatrix}_{(IJK)\times(I+1)}.$$

$$Z = \begin{pmatrix} U \\ \vdots \\ U \end{pmatrix}_{(IJK)\times J} \quad \text{with } U = \begin{pmatrix} \underline{1}_K & & \\ & \ddots & \\ & & \underline{1}_K \end{pmatrix}_{(JK)\times J}, \quad \text{and } \underline{\varepsilon} = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1J} \\ \vdots \\ \varepsilon_{I1} \\ \vdots \\ \varepsilon_{IJ} \end{pmatrix}.$$

Thus, $V = Var(\underline{y}) = \sigma_r^2 \cdot Z \cdot Z' + \sigma^2 \cdot I_{(IJK)}$

To show that the simple least squares estimator is also a BLUE for any estimable function, it suffices to show that $VX = XQ$ for some Q.

First note that $$Z \cdot Z' \cdot X = \begin{pmatrix} D & \cdots & D \\ \vdots & & \vdots \\ D & \cdots & D \end{pmatrix} \cdot \begin{pmatrix} \underline{1}_{JK} & \underline{1}_{JK} & & \\ \vdots & & \ddots & \\ \underline{1}_{JK} & & & \underline{1}_{JK} \end{pmatrix}$$

$$= \begin{pmatrix} I \cdot A & A & \cdots & A \\ \vdots & \vdots & & \vdots \\ I \cdot A & A & \cdots & A \end{pmatrix},$$ where $D = UU'$. and $A = D \cdot \underline{1}_{JK}$.

After simplification, we have $A = K \cdot \underline{1}_{JK}$. Thus,

$ZZ'X = K \cdot X \cdot Q_1$, where $$Q_1 = \begin{pmatrix} I & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 1 \end{pmatrix}_{(I+1)\times(I+1)}.$$

In summary, we have $VX = XQ$ for

$$Q = K\sigma_r^2 \cdot Q_1 + \sigma^2 I_{(I+1)}.$$