Stat 516 Homework 3 Due Tuesday, March 1

1. Examine and run the R code available at

   http://www.public.iastate.edu/ dnett/S516/hw3pr1.R

   This code simulates a data matrix y from a single-channel gene expression experiment. Suppose that the rows of y correspond to genes and that the columns correspond to experimental units, with one single-channel microarray slide per experimental unit. Suppose the first five columns represent the experimental units that received treatment 1 and the last five columns correspond to experimental units that received treatment 2 in a completely randomized experimental design.

   (a) Write a paragraph that explains how the data were simulated. Imagine that the paragraph is for a statistics paper that will be read by other statisticians who do not necessarily use R or know anything about computing. A reader should understand the exact distribution of the data from your description.

   (b) Without doing any preprocessing of the data, conduct a two-sample t-test for each gene. Assume that the variance within a gene is constant across observations in both treatment groups but might differ from gene to gene. How many genes have p-values ≤ 0.001? How many of the genes with p-values ≤ 0.001 were simulated to be differentially expressed?

   (c) Normalize the data by median centering. Provide the R code that you used and the data for the first gene after median centering.

   (d) Conduct a two-sample t-test for each gene using the median-centered data. Again assume that the variance within a gene is constant across observations in both treatment groups but might differ from gene to gene. How many genes have p-values ≤ 0.001? How many of the genes with p-values ≤ 0.001 were simulated to be differentially expressed?

   (e) Which data set (the original or normalized) would be more useful to a scientist who wants to identify differentially expressed genes?

   (f) What proportion of all differentially expressed genes had p-values ≤ 0.001 in the analysis of the normalized data?

2. Consider the mixed model for a single gene for the experiment depicted on page 49 of the notes "Introduction to Mixed Linear Models in Microarray Experiments." Let σ_m^2 and σ_e^2 denote the variance components for the mouse random effects and the error random effects, respectively. Determine the following in terms of these variance components.

   (a) $\mathrm{Var}(Y_{111})$.

   (b) The covariance between any two observations from a single mouse.

   (c) The correlation between any two observations from a single mouse.

   (d) The covariance between any two observations from different mice in the same treatment group.

   (e) The covariance between any two observations from mice in different treatment groups.

   (f) The variance of the average of the two observations from a single mouse.

   (g) The variance of the average of all the observations from a single treatment group.

   (h) $\mathrm{Var}(\bar{Y}_{1\cdot\cdot} - \bar{Y}_{2\cdot\cdot})$.

(i) Now suppose the data in the file

http://www.public.iastate.edu/ dnett/S516/hw3data.txt

has been observed for a single gene, where y denotes the log-scale normalized signal intensity. For each mouse, find the sample variance of the two observations for that mouse. This should give you 8 sample variances. Find the average of those sample variances. This provides an estimate of σ_e^2.

(j) Now find the average of the two observations for each mouse. Compute a two-sample t-test using these averages (4 for each treatment) to test for differential expression between treatments. Report a p-value and the degrees of freedom for your test.

(k) Within each treatment group, find the sample variance of the 4 averages used in the previous question. Average these two sample variances to obtain an estimate of your answer to part (f). Use this together with your answer to part (i) to find an estimate of σ_m^2.

(l) Now run the R code in the file

http://www.public.iastate.edu/ dnett/S516/hw3pr2.R

to fit the linear mixed model to the 16 data points. How does the output compare with your answers to parts (i), (j), and (k)? (Note that R reports the square roots of variance components rather than the variance components. These standard deviations are listed under headings *(Intercept)* and *Residual*. The *(Intercept)* heading corresponds to σ_m in this case.)

3. Suppose researchers were interested in studying the effects of two diets (A and B) on gene expression in muscle tissue of hogs. Eight pens containing four hogs each were available for an experiment. For logistical reasons, all hogs in any one pen had to be fed the same diet. The researchers randomly assigned 4 of the pens to diet A and the other 4 to diet B. At the time of slaughter, one RNA sample was taken from muscle tissue of each of the 32 hogs. These 32 samples were measured using 32 Affymetrix GeneChips (one for each sample).

   (a) Does this experiment involve blocking? If so, name the blocks.

   (b) What are the experimental units in this experiment?

   (c) What are the observational units in this experiment?

   (d) Write down a statistical model for this experiment based on the experimental design. You may use the abbreviated notation like that on slides 51, 56, 62, 63, 73 and 75 of "Introduction to Mixed Linear Models in Microarray Experiments."

   (e) Is this experiment best described as a completely randomized design, randomized complete block design, split-plot design, incomplete block design, or Latin square design? Explain.

4. Suppose researchers were interested in studying the effects of two diets (A and B) and four doses of a drug (0, 10, 20, and 30 mg/kg of body weight) on gene expression in muscle tissue of hogs. Eight pens containing four hogs each were available for an experiment. For logistical reasons, all hogs in any one pen had to be fed the same diet. The researchers randomly assigned 4 of the pens to diet A and the other 4 to diet B. Within each pen, the four hogs were randomly assigned to the four doses of the drug in a completely randomized manner with one hog for each dose. Each hog was injected with its assigned dose once each week prior to slaughter. At the time of slaughter, one RNA sample was taken from muscle tissue of each of the 32 hogs. These 32 samples were measured using 32 Affymetrix GeneChips (one for each sample).

(a) Name the treatment factors considered in this experiment.

(b) Name the levels of each treatment factor.

(c) Name the experimental units in this experiment.

(d) Is this experiment best described as a completely randomized design, randomized complete block design, split-plot design, incomplete block design, or Latin square design? Explain.

(e) Write down a model for the data based on this experimental design. You may use the abbreviated model notation discussed in class.

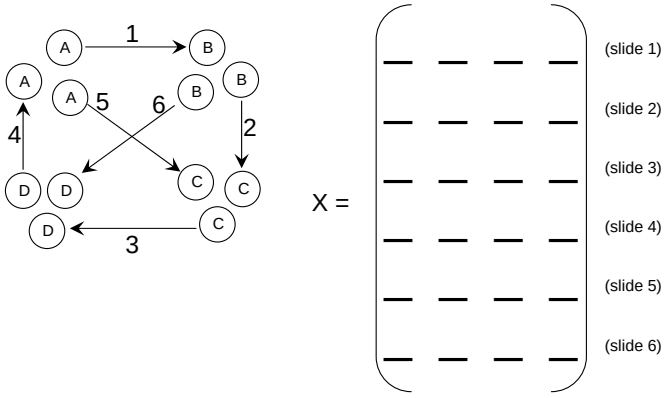
(f) Suppose that instead of using Affymetrix GeneChips, the researchers decided to measure expression using a total of 16 two-color microarray slides. Furthermore, suppose the researchers were primarily interested in understanding differences in gene expression between the two diets for each dose of the drug. Draw a picture (using the design notation that we have used in class) to illustrate how you would recommend pairing samples on slides and assigning dyes.

(g) Write down a model for the data based on this two-color microarray experimental design. You may use the abbreviated model notation discussed in class.

5. Suppose a two-color microarray experiment is to be conducted to compare the effect of four treatments (A, B, C, and D) on gene expression in maize. Suppose that treatment D is a control and that researchers are primarily interested in understanding which of the treatments A, B, and C differ from the control treatment D in terms of mean expression for each gene. The researchers have 12 experimental units and 6 slides available for the experiment. For any given gene, denote the mean expression of an observation as a function of dye and treatment according to the following table:

Treatment	Dye	Mean Expression
A	Cy3	$\mu + \delta_3 + \alpha$
B	Cy3	$\mu + \delta_3 + \beta$
C	Cy3	$\mu + \delta_3 + \gamma$
D	Cy3	$\mu + \delta_3$
A	Cy5	$\mu + \delta_5 + \alpha$
B	Cy5	$\mu + \delta_5 + \beta$
C	Cy5	$\mu + \delta_5 + \gamma$
D	Cy5	$\mu + \delta_5$

(a) Consider the balanced incomplete block design (depicted below) that compares each treatment to each other treatment on exactly one slide. Provide the appropriate X matrix for this design. Assume that we will use the Cy5-Cy3 difference on each slide as our response variable and that our parameter vector for this analysis is $(\delta_5 - \delta_3, \alpha, \beta, \gamma)'$. (Note that the numbers in the diagram below correspond to slide numbers. Please enter the rows of X accordingly.)

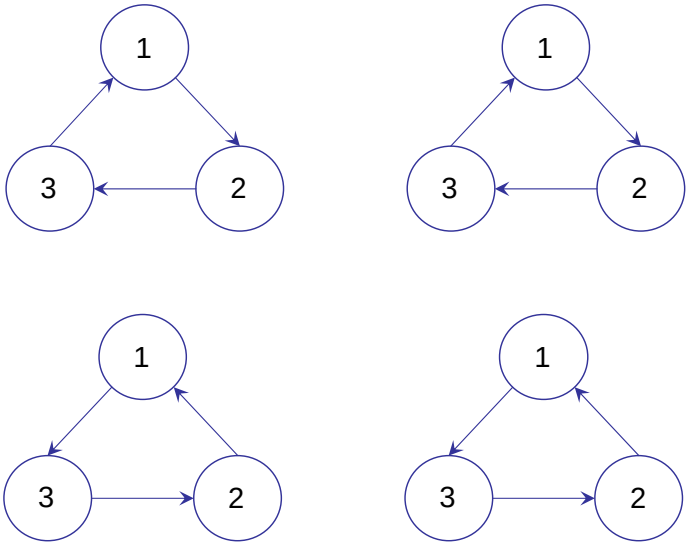


(b) Suppose that

$$(X'X)^{-1} = \begin{bmatrix} 0.2 & 0.10 & 0.10 & 0.00 \\ 0.1 & 0.55 & 0.30 & 0.25 \\ 0.1 & 0.30 & 0.55 & 0.25 \\ 0.0 & 0.25 & 0.25 & 0.50 \end{bmatrix}$$

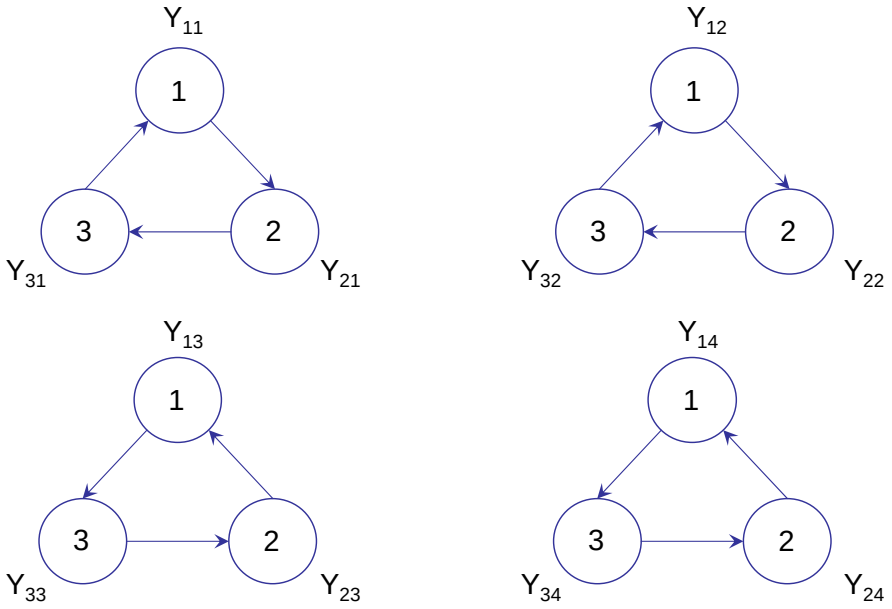
for the X matrix in part (a). Without using a computer or calculator, prove that the design in part (a) is inadmissible among the class of designs that use 6 slides and 12 experimental units to test for differences between the control treatment (D) and each of the other three treatments (A, B, and C). (Note that we are concerned about three tests: A vs. D, B vs. D, and C vs. D.)

6. Consider a completely randomized experiment with three treatments denoted 1, 2, and 3 and four experimental units per treatment. Suppose mRNA levels are measured with two-color microarrays using the following design.



In class we discussed a mixed linear model for the 24 observations obtained for a single gene. This model included fixed factors treatment and dye. Denote the treatment effects by τ_1, τ_2, and τ_3, and denote the dye effects by δ_1 and δ_2. The model we discussed included variance components for the random factors slide, experimental unit, and residual. Denote these variance components by σ_s^2, σ_u^2, and σ^2, respectively. Assume this model is correct throughout this problem.

Suppose that instead of analyzing the 24 observations or instead of taking red-green differences, the two observations for each experimental unit are averaged to obtain a total of 12 averages denoted by Y_{ij} in the figure below.



(a) Determine the variance of Y_{11} in terms of the variance components.

(b) Determine the covariance between Y_{11} and Y_{21} in terms of the variance components.

(c) Determine the covariance between Y_{21} and Y_{31} in terms of the variance components.

(d) Determine the covariance between Y_{11} and Y_{14} in terms of the variance components.

(e) Suppose the following model is fit to the 12 averages.

$$Y_{ij} = \mu + \alpha_i + b_j + e_{ij},$$

where μ and α_1, α_2, and α_3 are fixed effects; b_1, b_2, b_3, b_4 are independent and identically normally distributed block effects with mean 0 and variance σ_b^2; and e_{ij} are independent and identically distributed random errors with mean 0 and variance σ_e^2. As usual all random effects are assumed to be independent of one another. Note this model is just a standard model for an RCBD with random blocks. Find expressions for σ_b^2 and σ_e^2 in terms of the variance components σ_s^2, σ_u^2, and σ^2.