STAT 511 **Assignment 8** **Name _____**
Spring 2002

Reading Assignment: Rencher, Chapter 15
Neter, Kutner, Nachtsheim & Wasserman, *Applied Linear Statistical Models*, 4th Edition, Chapters 24, 28, 29., and Appendix D. Appendix D has rules for evaluating expectations of means squares for models with fixed and random effects and many types of balanced experiments.

Written Assignment: On-campus students: Due Wednesday, April 17, in class.
Distance students: Put solutions in the mail by April 25

Second Exam: This is a take home exam that will be distributed to the on-campus students In class on April 17 and it will be due in class on April 24. No late papers will be accepted. Arrangements with distance students will be made via e-mail.

1. Analyze the data from Problem 4 on Assignment 7, assuming that the block effects corresponding to the clinics are random effects. Assume the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \tau_k + e_{ijk}$$

where Y_{ijk} is the observed increase in systolic blood pressure dogs in the k-th clinic that are induced with the j-th disease and treated with the i-th drug. Here, we will consider the clinics as a random sample from the possible clinics that could have been used in this study, and assume that $\tau_k \sim \text{NID}(0, \sigma_t^2)$ and $e_{ijk} \sim \text{NID}(0, \sigma_e^2)$ are independent sets of random effects.

A. Obtain REML estimates of the variance components σ_t^2 and σ_e^2.

B. Use the lme( ) function in S-PLUS or Type III sums of squares from PROC MIXED in SAS to test the null hypothesis of no interaction between the drug and disease factors. State your conclusion.

C. Because of the imbalance caused by missing data the Type III sums of squares F-tests of the null hypothesis of no interaction between the drug and disease factors provided by PROC GLM and PROC MIXED in SAS may not be exactly the same (see the SAS output posted in the file **dogs.sas.output1** on the course web page). Which is the more appropriate test to use? Explain

D. Use the "slice" option and the LSMEANS option in PROC MIXED or write a function in S-PLUS to examine differences in estimated mean increases in systolic blood pressure for the four drugs for each disease. State your conclusions. Does the same drug provide the lowest mean increase in systolic blood pressure for each disease?

E. Predict the values of the random effects of the clinics selected for this study. Do the clinic effects appear to be a random sample from a normal distribution? Explain.

F. Do the random errors appear to be a random sample from a normal distribution? Explain.

2. Four plants of the same variety were randomly sampled from a large field of plants. Three leaves were randomly selected from each plant and three determinations of the concentration of a certain acid were made on each leaf using each of three different methods. These methods are labeled as method A, method B, and method C. The data are presented in the following table. Larger values correspond to higher concentrations of the acid.

		Method of Determination		
Plant	**Leaf**	**A**	**B**	**C**
1	1	11.2	11.6	12.0
	2	16.1	16.5	16.8
	3	18.3	18.7	19.0
2	1	14.1	13.8	14.2
	2	18.5	18.2	19.0
	3	11.9	12.1	12.4
3	1	15.3	15.9	16.0
	2	19.5	19.3	20.1
	3	16.9	16.5	17.2
4	1	7.3	7.0	7.8
	2	8.9	9.4	9.3
	3	10.9	10.5	11.3

These data are posted in the file macid.dat on the course web page. This file has four columns. The first column identifies plants, the second column identifies leaves within plants, The third column identifies methods for determination of acid levels, and the fourth column contains the observed acid concentrations.

Consider the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{jk} + \varepsilon_{ijk}$$

where Y_{ijk} is the observed acid concentration made by the i-th method on the k-th leaf of the j-th plant and

$$\beta_j \sim \text{NID}(0, \sigma^2_\beta) \qquad \gamma_{jk} \sim \text{NID}(0, \sigma^2_\gamma) \qquad \varepsilon_{ijk} \sim \text{NID}(0, \sigma^2_\varepsilon)$$

and all random effects are independent of each other. In this model, β_j represents variation in acid concentrations among plants, γ_{jk} represents variation in acid concentrations among leaves within plants, and ε_{ijk} represents and remaining random variation.

A. Using this model, obtain an ANOVA table for the observed data.

B. Report formulas for expectations of mean squares.

C. Obtain REML estimates of the variance components σ^2_β , σ^2_γ and σ^2_ε. What is the largest source of random variation in this study?

D. Construct a 95% confidence interval for the mean acid concentration for the population of plants as measured by method C,

E. Construct a 95% confidence interval for the difference in mean acid concentrations determined by methods C and A.

F. Examine differences in estimates of acid concentration means for all of the methods. State your conclusions.

G. Estimate the correlation between determinations of acid concentrations taken from two different leaves of the same plant.

H. Estimate the correlation between determinations of acid concentrations taken from the same leaf.

I. This experiment could have been done in other ways. For example, the researchers could have sampled 12 plants and sampled three leaves from each plant. Then, methods A, B, and C could be applied to different leaves from the same plant using a random assignment of methods to leaves for each plant selected for the study. This would also provide 36 observations on acid concentrations. Would this result in more precise or less precise comparisons of the three methods for determining acid concentrations? Estimate the gain or loss of precision.

3. The following experiment was done to compare the effects of three different treatments (labeled A, B, and C) for controlling worms that live in the intestines of young pigs. Eliminating the worms from the intestines of the pigs, or at least inhibiting their development, will enable the pigs to gain weight more rapidly. Consequently, the measured response was the amount of weight a pig gained during the course of the study (in kg). Ten litters of pigs were selected from the many litters produced at a large farm. Three male pigs were randomly selected from each litter and those three pigs were randomly assigned to the three treatments. In this way, each of the three treatments was used on one pig from each litter. All pigs were the same age when they entered the study.

   A. Identify the following components of this study:
      i. Experimental units
      ii. Blocking factor (fixed or random?)
      iii. Treatment factor (fixed or random?)

   B. Write down the linear model that you would initially use to analyze these data.

C. Outline an ANOVA table for your model in part B. Include columns for the sources of variation, degrees of freedom, and expectations of mean squares.

D. Give a formula for a 95% confidence interval for the mean weight gain for treatment A. Include values for sample sizes and a value or formula for degrees of freedom.

E. Give a formula for a 95% confidence interval for the difference in the mean weight gains for treatments A and C. Include values for sample sizes and a values or formula for degrees of freedom.

4. An automobile manufacturer used four automobiles and four drivers in a study of the effects of four gasoline additives on reducing nitrogen oxide levels in automobile emissions. The additives are simply labeled A, B, C and D. The four automobiles were sampled from the automobiles of a specific model produced by the company. The drivers were sampled from a large group of test drivers that worked for the company. A Latin square design was used in an attempt to "balance out" the effects of automobile-to-automobile and driver-to-driver variation on the comparison of the gasoline additives. In this design, each driver drove each automobile once and used each additive once. Also, oxides are shown in the following table (a larger value indicates a greater reduction). These data are stored in the file **gas.additive.dat** posted on the course web page. (BHH)

	Auto 1	Auto2	Auto 3	Auto 4
Driver 1	A (21)	B (26)	D (20)	C (25)
Driver 2	D (23)	C (26)	A (20)	B (27)
Driver 3	B (15)	D (13)	C (16)	A (16)
Driver 4	C (17)	A (15)	B (20)	D (20)

A. For this experiment identify (if there are none, simply answer "none"):
   i. Fixed blocking factors
   ii. Random blocking factors
   iii. Fixed treatment factors
   iv. Random treatment factors

B. Using your answer to part A and assuming that the effects of the factors are strictly additive (no interaction) and that any random effect is distributed independently of any other random effect, write out a linear model for these data.

C. Evaluate the ANOVA table for your model in part B. Include a column to show expectations of mean squares.

D. Compute REML estimates of variance components, Are the REML estimates equal to the method of moments estimates in this case?

E. Construct a 95% confidence interval for the mean nitrogen oxide reduction provided by additive A.

F. Construct a 95 % confidence interval for the difference in the mean nitrogen oxide reductions for additives A and B.

G. Use the Tukey Honest Significant Difference (HSD) to determine which additive (or additives) provides the greatest reduction in the mean emission of nitrogen oxides (averaging across drivers and cars).

5. The data in the following table are from an experiment where the amount of dry matter was measured for wheat plants grown under conditions with different levels of moisture and different amounts of fertilizer. There were 48 pots and 12 plastic trays used in the experiment. The same soil mixture was used in each pot. Four pots were placed in each tray. The levels of the moisture factor corresponded to adding either 10, 20, 30, or 40 mℓ. of water per pot per day to the tray. The water was absorbed thorough holes in the bottom of the pots. Moisture levels were randomly assigned to trays with three trays assigned to each moisture level. There could be variation among trays assigned to the same moisture level because of the inability of the researchers to exactly maintain the desired moisture level in each tray. Furthermore, different trays may be subject to slightly different environmental conditions (temperature, humidity, light. etc…), but pots in the same tray would be subject to relatively similar conditions.

   Before planting the wheat seeds, fertilizer was added to the soil in the pots at levels of 2, 4, 6, or 8 mg. per pot. The four levels of fertilizer were randomly assigned to the four pots within each tray. An independent randomization was done within each tray. Then the same number wheat seeds were planted in each pot and after 30 days the wheat plants were removed from the pots and dried. The weight of the dry matter (in ounces) was recorded for each pot. The observed weights are shown in the following table. (M&J, 1984)

		Level of fertilizer (mg)			
Moisture Level (mℓ/pot/day)	**Tray**	**2**	**4**	**6**	**8**
10	1	3.3458	4.3170	4.5572	5.8794
	2	4.0444	4.1413	6.5173	7.3776
	3	1.9758	3.8397	4.4730	5.1180
20	4	5.0490	7.9419	10.7697	13.5168
	5	5.9131	8.5129	10.3934	13.9157
	6	6.9511	7.0265	10.9334	15.2750
30	7	6.5693	10.7348	12.2626	15.7133
	8	8.2974	8.9081	13.4373	14.9575
	9	5.2785	8.6654	11.1372	15.6332
40	10	6.8393	9.0842	10.3654	12.5144
	11	6.4997	6.0702	10.7486	12.5034
	12	4.0482	3.8376	9.4367	10.2811

These data have been posted on the course web page as **wheatw.dat.**

A. Identify the following features of this experiment, if they exist.

primary (or whole plot) experimental units:

sub-plot units:

treatment factors:

blocking factors:

B. Consider the model

$$Yijk = \mu + \alpha_i + \gamma_{ij} + \tau_k + \delta_{ik} + e_{ijk}$$

where Y_{ijk} is the observed dry matter weight for the wheat grown in the pot assigned to the k-th level of fertilizer in the j-th tray assigned to the i-th level of the moisture factor. Here γ_{ij} and e_{ijk} are random terms with

$$e_{ijk} \sim NID(0, \sigma_e^2) \qquad \text{and} \qquad \gamma_{ij} \sim NID(0, \sigma_g^2)$$

and any e_{ijk} is independent of any γ_{ij}. Report an ANOVA table for this model and give formulas for the expectation of the mean squares. (SAS output from the GLM procedure is posted n the file wheatw.sas.output1. SAS and S-PLUS code for applying mixed linear models to these data are posted in the files **wheatw.sas** and **wheatw.ssc**, respectively.)

C. Use the mean squares from the ANOVA table in Part B to obtain method of moment estimates of the variance components σ_e^2 and σ_g^2.

D. With respect to the model in part B, which of the following are estimable quantities?

$\mu + \tau_1$, $\qquad \mu + \alpha_1 + \tau_1 + \delta_{11}$,

$\tau_1 - \tau_2$, $\qquad \alpha_1 + \delta_{11} - \alpha_2 - \delta_{21}$,

$\delta_{11} - \delta_{13} - \delta_{21} + \delta_{23}$, $\qquad \left(\alpha_1 + \frac{1}{4}\sum_{k=1}^{4} \delta_{1k}\right) - \left(\alpha_2 + \frac{1}{4}\sum_{k=1}^{4} \delta_{2k}\right)$

Give the value of the estimate of any quantity that is estimable. Report a standard error for each estimate and construct an appropriate 95% confidence interval.

E. Examine the profile plot of the sample means for the various combinations of the moisture and fertilizer factors with moisture level on the horizontal axis and one profile for each fertilizer level. Examine the corresponding plot with fertilizer levels on the

horizontal axis and one profile for each moisture level. Look for trends. Do not submit these plots. Given the results from the profile plots, the following quadratic model may provide a reasonable simplification of the model in part B:

$$Y_{ijk} = \beta_0 + \beta_1\left(\mathbf{X_{1i}} - \overline{\mathbf{X}}_{1.}\right) + \beta_2\left(\mathbf{X_{1i}} - \overline{\mathbf{X}}_{1.}\right)^2$$

$$+ \beta_3\left(X_{2k} - \overline{X}_{2.}\right) + \beta_4\left(\mathbf{X_{1i}} - \overline{\mathbf{X}}_{1.}\right)\left(\mathbf{X_{2k}} - \overline{\mathbf{X}}_{2.}\right)$$

$$+ \beta_5\left(X_{2k} - \overline{X}_{2.}\right)^2 + \gamma_{ij} + \varepsilon_{ijk}\,.$$

where X_{1i} denotes the value of the i-th moisture level.

X_{2k} denotes the value of the k-th fertilizer level

and $\gamma_{ij} \sim \text{NID}(0, \sigma_g^2)$ is independent of $e_{ijk} \sim \text{NID}(0, \sigma_{\mathbf{e}}^2)$.

Report REML estimates for $\sigma_{\mathbf{e}}^2$ and σ_g^2 for this model and report a table of estimates, standard errors, and tests of significance for β_0, β_1, β_2, β_3, β_4, β_5.

F. Use the model in part E to estimates the mean dry weight matter when fertilizer is applied at a level of 5 mg and the moisture level is 15 ml/pot/day. Provide a 95% confidence interval for your estimate. Note, that an easy way to do this in SAS is to add this case to the data set with a missing value for the response. Alternatively, you could use an ESTIMATE statement in the MIXED procedure. In S-PLUS, you can directly obtain the estimate from the estimates of the coefficients and variance components.

G. Is the model in part E appropriate for these data? You can partially answer this question by considering whether or not the model in Part B a significant improvement over the model in Part E? Give a value for your test statistic and state your conclusion.

One could spend more time and effort trying to find the most parsimonious model that provides an adequate description of the data for this study. You are not required to do this, but we will make some comments in the posted solution.