1. In a study to examine the effect of 4 drugs on 3 experimentally induced diseases in dogs, each drug-disease combination was given to six randomly selected dogs. The measurement (Y) to be analyzed was the increase in systolic blood pressure (mm Hg) due to treatment. Unfortunately, some dogs were unable to complete the experiment. The data (Kutner, 1974) are shown in the following table.

	Disease (j)		
Drug (i)	1	2	3
1	42, 44, 36, 13, 19, 22	33, 26, 33, 21	31, -3, 25, 25, 24
2	28, 23, 24, 42, 13,	34, 33, 31, 36	3, 26, 28, 32, 3, 16
3	1, 29, 19	11, 9, 7, 1, -6	21, 1, 9, 3
4	24, 9, 22, -2, 15	27, 12, 12, -5, 16, 15	22, 7, 25, 5, 12

Consider the model $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$, where $\epsilon_{ijk} \sim NID(0, \sigma^2)$ and Y_{ijk} denotes the change in systolic blood pressure (mm Hg) for the k-th dog given the j-th disease and treated with the i-th drug.

(a) The file **dogs.r** contains R code for applying *lm( )* and *anova( )* functions to these data. Data are posted in the file **dogs.dat**. Note that this application of the *lm ( )* function imposes some restrictions to solve the normal equations. What are the restrictions?

(b) Using the solution to the normal equations provided by this application of the of *lm( )* function, report estimates of the following quantities:

$$\mu,\ \ \alpha_1,\ \ \beta_3,\ \ \gamma_{23},\ \ \alpha_2 - \alpha_3,\ \ \gamma_{22} - \gamma_{23} - \gamma_{32} + \gamma_{33},$$

$$\mu + \alpha_2 + \beta_3 + \gamma_{23},\ \ (\alpha_2 - \alpha_3) + \frac{1}{3}(\gamma_{21} + \gamma_{22} + \gamma_{23} - \gamma_{31} - \gamma_{32} - \gamma_{33})$$

Give an interpretation of each quantity with respect to the restricted model and the mean change in systolic blood pressure.

(c) The are many ways to put linear restrictions on parameters in the original model to obtain a solution to the normal equations. Would the least squares estimates of any of the linear combinations of parameters in part (b) have the same value for all such solutions to the normal equations? Which ones? Explain.

(d) Examine two ANOVA tables, one corresponding to the $R(\mu),\ R(\alpha|\mu),\ R(\beta|\mu,\alpha),\ R(\gamma|\mu,\alpha,\beta)$ partition of the sums of squares and another ANOVA table corresponding to the $R(\mu),\ R(\beta|\mu),\ R(\alpha|\mu,\beta),$ $R(\gamma|\mu,\alpha,\beta)$ partition. State any useful inferences that can be obtained from these two ANOVA tables. (Do not report the ANOVA tables.)

(e) Denote the mean change in systolic blood pressure for the i-th drug used with the j-th induced disease (a cell mean) as

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}.$$

Let $\bar{Y}_{ij\cdot}$ denote the average of the n_{ij} observations obtained from dogs induced with the j-th disease and treated with the i-th drug, and let S^2 denote the sum of squared residuals divided by its degrees of freedom. Use Results 4.7 and 4.8 from the course notes to show that

$$F = \frac{\sum_{j=1}^{3}\left(n_{1j}^{-1} + n_{3j}^{-1}\right)^{-1}\left(\bar{Y}_{1j\cdot} - \bar{Y}_{3j\cdot}\right)^2}{3S^2}$$

has a non-central F-distribution. Report the degrees of freedom for this distribution and describe the null hypothesis that can be tested with this statistic in terms of the cell means.

(f) Evaluate the test statistic in part (e). Report a p-value and state your conclusion.

(g) Examine type III sums of squares for these data (do not submit these sums of squares).

i. State the null hypothesis associated with the F-test for interaction. What can you conclude from the results of this test?

ii. Specify the C matrix needed to write the null hypothesis associated with the F-test for drug effects in the form $H_0 : C\boldsymbol{\beta} = \mathbf{0}$, where

$$\boldsymbol{\beta} = (\mu,\ \alpha_1,\ \alpha_2,\ \alpha_3,\ \alpha_4,\ \beta_1,\ \beta_2,\ \beta_3,\ \gamma_{11},\ \gamma_{21},\ \gamma_{31},\ \gamma_{41},\ \gamma_{12},\ \gamma_{22},\ \gamma_{32},\ \gamma_{42},\ \gamma_{13},\ \gamma_{23},\ \gamma_{33},\ \gamma_{43})^T$$

What can you conclude from the results of this test?

iii. Specify the C matrix needed to write the null hypothesis associated with the F-test for disease effects in the form $H_0 : C\mu = \mathbf{0}$, where

$$\boldsymbol{\mu} = (\mu_{11},\ \mu_{21},\ \mu_{31},\ \mu_{41},\ \mu_{12},\ \mu_{22},\ \mu_{32},\ \mu_{42},\ \mu_{13},\ \mu_{23},\ \mu_{33},\ \mu_{43})^T$$

is the vector of cell means, i.e., $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} = E(Y_{ijk})$. What can you conclude from the results of this test?

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**. 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

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

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

$$\beta_j \sim NID(0, \sigma^2_\beta), \;\; \gamma_{jk} \sim NID(0, \sigma^2_\gamma), \;\; \epsilon_{ijk} \sim NID(0, \sigma^2_\epsilon)$$

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. 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 Blackboard.

	Auto 1	Auto 2	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 the followings: (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).

4. 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.

Moisture Level (ml/pot/day)	Tray	Level of fertilizer (mg) 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 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

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

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

$$e_{ijk} \sim NID(0, \sigma_e^2) \quad \text{and} \quad \gamma_{ij} \sim NID(0, \sigma_\gamma^2)$$

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 code for applying mixed linear models to these data are posted in the files **wheatw.sas**)

(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_γ^2.

(d) With respect to the model in part (b), which of the following are estimable quantities?

$$\mu + \tau_1, \;\; \mu + \alpha_1 + \tau_1 + \delta_{11}, \;\; \tau_1 - \tau_2,$$

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

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 7 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):

$$\begin{aligned} Y_{ijk} = \beta_0 + \beta_1(X_{1i} - \bar{X}_{1\cdot}) + \beta_2(X_{1i} - \bar{X}_{1\cdot})^2 + \beta_3(X_{2k} - \bar{X}_{2\cdot}) \\ + \beta_4(X_{1i} - \bar{X}_{1\cdot})(X_{2k} - \bar{X}_{2\cdot}) + \beta_5(X_{2k} - \bar{X}_{2\cdot})^2 + \gamma_{ij} + e_{ijk}, \end{aligned}$$

where X_{1i} denotes the value of the i-th moisture level. X_{2k} denotes the value of the k-th fertilizer level and $\gamma_{ij} \sim NID(0, \sigma^2_\gamma)$ is independent of $e_{ijk} \sim NID(0, \sigma^2_e)$. Report REML estimates for σ^2_γ and σ^2_e 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.

(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.