Name _____

Laboratory #7
Applying the General Linear Model - Multifactor ANOVA

The purpose of this lab is to increase your skill and facility in applying the general linear model to what the statistician John Tukey called "data situations." A data situation includes the accessory information that accompanies a set of measurements. For example, the designer of an experiment may already know that a convenient blocking factor, such as separate greenhouses, will have effects that are independent of treatments and controls. In this situation, the experimenter will carry out a randomized block design to control for the among greenhouse variation. Or to take another example, a physiologist may know that the effects of a drug on a rat can carry over for a day, and so not use any one rat on successive days. The term "data situation" is not in wide use, but is valuable in distinguishing "data" (a set of scaled numbers) from situational knowledge that accompanies any set of data.

In this lab we'll look at three "data situations." In the first, we'll construct a two-factor model based on the research question, which concerns guinea pig litter sizes (Box 13.12, Sokal and Rohlf 2012). In the second data situation, we'll use accessory information to reformulate the litter size model. In the third situation, which concerns lactic acid production in frog embryos (Box 11.7 in Sokal and Rohlf 1995) there is a missing value. This results in an unbalanced design that is not readily handled in ANOVA routines. In contrast, GLM routines automatically handle unbalanced data.

Once you have completed the lab write-up, you should have

- the ability to translate a data situation into a model-based analytic format

- the ability to use a statistical package to execute a two-factor ANOVA model

- increased experience in interpreting interactive effects

-familiarity with how your package handles missing data

At this point make sure that you have two data sets:

LitterSize.txt	Box 13.12 in Sokal and Rohlf 2012
FrogEmbryos.txt	Box 11.7 in Sokal and Rohlf 1995

Both are available on the course website: www.mun.ca/biology/schneider/b4605/Data/Labs

Data situation #1 Two factor mixed model ANOVA Name _____

StrainB	Strain13	Year
2.68	2.36	1916
2.60	2.41	1917
2.43	2.39	1918
2.90	2.85	1919
2.94	2.82	1920
2.70	2.73	1921
2.68	2.58	1922
2.98	2.89	1923
2.85	2.78	1924

The first example of the GLM for two-factor ANOVA will be a re-analysis of the LitterSize data from Sokal and Rohlf (2012). In the previous analysis we did not take into account the pairing of observations by years. The randomization test you carried out in Lab 4 was for an unpaired comparison of mean litter sizes in two strains of Guinea pig. The animals in both strains experienced similar conditions within a year, while showing difference in litter size from year to year. We can control for year to year variation by introducing year as a random factor. This is called a paired *t*-test or paired comparisons—we have two categories in the factor of interest, genetic strain. The pairing variable is usually treated as random—a sample from a population of years in this case. A paired comparisons analysis is a mixed model with a fixed type of factor having two levels, and a random type of factor having multiple levels.

Fill in the table of showing variable names, using *LSize*, *Str*, and *Yr* as symbols.

Variable Name	Symbol	Units	Scale Type	Role	Factor Type
				Response	
				Explanatory	
				Explanatory	

Here is GLM format for any number of explanatory variables.

$Y = \beta_o + \sum \beta_i X_i + \epsilon_{normal}$ β_o = grand mean, $\sum \beta_i X_i$ = sum of explanatory terms

$\sum \beta_i X_i = \beta_1 X_1 + \beta_2 X_2 + \beta_{1x2} X_1 X_2$ <- - Note the interactive effects term.

Now use the symbols from above to write a GLM relating the response variable to the two explanatory variables. Your model should have 1 variable on the left, and on the right 3 explanatory terms and the residual (error) term.

_____ $-\beta_o$ = _____

df: _____ = _____ + _____ + _____ + _____

Write in the degrees of freedom below each term in the model, taking both explanatory variables as categorical. *Year* is a categorical variable because we have no reason to expect a trend in litter size over 9 years.

When you include the interaction term in the model, how many degrees of freedom remain to estimate the error term? _____

Data situation #1 (continued) Name _____

Now re-write your model with interactive effect assumed to be zero.

_____ $-\beta_o$ = _____

df: _____ = _____ + _____ + _____

Fill in the degrees of freedom for your model.

Import the data into your statistical package, and then reorganize it from tabular to model format.
When you are done, you should have three variables in 3 columns.
Each column variable should have 18 rows.

Here is the pseudocode for reorganizing data to model format.

```
Read tabular data into the package spreadsheet, one column per class
Insert a new column into the spreadsheet.  Assign it a name (Y)
Insert a new column adjacent to Y.  Assign it a name (factor X1)
Paste in the appropriate values of factor X1
Insert a new column adjacent to X1.  Assign it a name (factor X2)
Paste in appropriate values of factor X2
```

Define Data from file

Reorganize from tabular to model format

This is easily done in statistical packages that have built in spreadsheets, such as Minitab, SPSS, and SPlus. If you are using R, you will find it *far* easier to reorganize the data in a spreadsheet than in R. Save the reorganized data as a txt, csv, or excel file to read into R.
Next, we execute the analysis.

The pseudocode is nearly the same as for regression and oneway ANOVA

```
Define the response variable, Y
Define the explanatory variables X1 and X2
Write the model.  Then execute the model
Save residuals and fitted values.
Check assumptions of homogeneous normal residuals.
```

Run GLM ANOVA with residual diagnostics

The model can be executed from the pull-down menus in SPSS, SPlus, and Minitab.

Here are the Minitab command lines

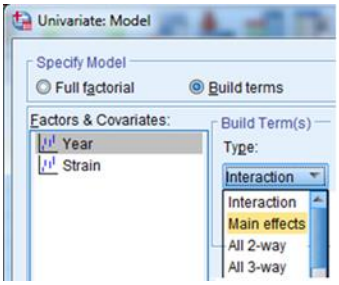
```
MTB > glm 'lsize' = 'yr' 'strain' ;
SUBC> residuals RES1 ;
SUBC> fits FITS1.
```

Run GLM ANOVA with residual diagnostics

Here is the R code.

```
LsizeMod <- lm(ls~strain+factor(year), data = CavyData)
Lsizefit <- fitted(LsizeMod)
Lsizeres <- resid(LsizeMod)
Anova(LsizeMod)
Plot(y=Lsizeres, x=Lsizefit)
```

SPSS uses a graphic interface to build the main effect only model -- >



Data situation #1 (continued) Name _____

To review the concept behind this analysis, write the revised model then fill in both the df and SS below each term in the model expressed by the command line.

Model _____

df: _____ = _____ + _____ + _____

SS: _____ = _____ + _____ + _____

Print out your ANOVA table and place it here. Be sure it is adequately labelled (caption for table, source of data, name of the response variable).

Comment on whether the residuals are homogeneous and normal, referring to diagnostic plots that you have attached to the end of this report. Be sure plots are adequately labelled (caption, axis labels, source of data).

State whether you would trust the p-value computed from the F-distribution, referring to your evaluation of homogeneity and normality of the residuals.

How does this paired comparison analysis compare to that for the unpaired analysis?
Unpaired: ($F_{1,16} = 1.29$, $p = 0.273$)

Paired: F_____ p =_____

Data situation #2 Two factor ANOVA with fixed effects Name _____

Year was taken as a random factor in the previous analysis. Inference was thus from a sample of 9 years to a population of many years. The first 3 years of data were a period of stress (R.R. Sokal, pers.comm.) in the guinea pig colony maintained by Sewall Wright, one of the founders of population genetics. Wright was Sokal's thesis supervisor.

Guinea pigs require fresh vegetables to stay healthy. Like us and unlike most animals, Guinea pigs cannot produce their own Vitamin C. The price of vegetables rose in the US during World War I, which erupted in July 1914 and ended on 11 November 1918. Write a GLM for analysis of the litter size data using two fixed factors: *Strain* (2 levels) and *Period* (War, Postwar).

_____ $-\beta_o$ = _____

df: _____ = _____ + _____ + _____ + _____

To execute the model add a new factor *Period* having two categories (War,Postwar) with 1916, 17, and 18 labelled War. With only 9 cases, the easiest way to do this is by typing directly into a spreadsheet – either the one for your package or the csv or excel file for importing into R.

( ) Print and tape (or copy and paste) your ANOVA table here.The GLM term of interest is the interaction term.
Did the 2 strains respond in the same way to change in conditions after 1918?

Data situation #3

The paired comparison in Analysis #1 (fixed factor with 2 levels) is a special case of a complete block design (fixed factor with 2 or more levels). The next example is a complete block design with 6 levels of the fixed factor and 4 blocks (clutches). The analysis will be the same as the previous one, except that this time a missing value creates unbalanced data that cannot be handled by many ANOVA routines. GLM routines automatically adjust for missing data.

I	II	III	IV
21.4	*	7.0	9.5
14.3	13.5	5.4	6.6
23.4	14.1	5.9	7.1
29.1	8.2	4.2	3.2
26.6	13.5	4.9	6.0
21.7	5.2	6.6	5.9

Data from Cohen 1954 *Physiol. Zool*., 27, 128-141 shown in Box 11.7 Sokal and Rohlf 1995.

Lactic acid production
(micromoles lactate x 10 per 12.25 embryos) in frog embryos from 4 clutches (4 columns) at 6 stages (6 rows) after 1st cleavage (0 min = 1st row, 360 = 2nd, 720 = 3rd, 1200 = 4th, 1600 = 5th, 2000 = 6th). * = missing value.

The research question is whether lactic acid production depends on stage.
Looking at the data, comment on the amount of variation across the random factor, clutches.

( )

Data situation #3 (Continued) Name _____

Fill in the table of variables, using *LAP*, *Stage*, and *Clutch* as symbols.

Variable	Symbol	Units	Scale Type	Role	Factor Type
				Response	
				Explanatory	
				Explanatory	

Write a GLM that relates the response variable to both explanatory variables. You should have 4 terms in addition to β_o on the right side of the equality sign.

_____ = β_o + _____

_____ =

Write in the degrees of freedom below each term in your model.
Write the model assuming no interaction term..

_____ = β_o + _____

_____ =

Write in the degrees of freedom below each term in your model.

Next, read the data (Sokal and Rohlf Box 11.7) into your statistical package.
Refer to Data situation #1 for the pseudocode or specific sequence of actions

Read tabular data into the package spreadsheet, one column per class *Etc.*

Define Data from file

Statistical packages handle missing values in different ways. Some packages treat * as missing. Other packages treat a blank as missing. Some use NA as missing. Make sure the missing value in Box 11.7 is coded as missing according to your package. This may mean changing the * in the data set to some other character such as a blank space.

Refer to Data situation #1 for the generic recipe (pseudocode) or specific sequence of actions

Reorganize to model format

Pseudocode (applies to any statistical package) Define the response variable, Y = LAP Define the explanatory variable X1 = Clutch X2 = Stage Save residuals and fitted values. Check p-value assumptions of homogeneous normal residuals.

Run GLM ANOVA with residual diagnostics

If your package has a 2-wayANOVA routine try running it with the FrogEmbryos data.
In R, use the aov( ) function.

What happens?_____

Then execute the analysis with a GLM command (be sure to compute fits and residuals).

Data Situation #3 (continued) Name _____

Comment on whether the residuals are acceptable (homogeneous and normal), referring to individually labelled diagnostic plots attached to this report. Be sure each plot is adequately labelled (caption, axes labelled, source of data).

State your evaluation of the assumptions from the residuals. Make a judgement on whether a randomized p-value would alter a Neyman-Pearson decision on a fixed criterion of $\alpha = 5\%$ and state your reasoning.

Display the calculation of the *F*-ratio from your GLM analysis, by filling in the blanks.

$$MS_{clutches}/MS_{error} = \_\_\_\_\_ / \_\_\_\_\_ = \_\_\_\_\_$$

Sokal and Rohlf (1995, p362) show how to adjust for missing values within an ANOVA table. Prior to GLM routines, the calculation of MS_{error} was laborious and time consuming. Here is the adjusted ANOVA table from Sokal and Rohlf (1995, Box 11.7).

Source of variation	*df*	*SS*	*MS*	*F*
A (columns: blocks: clutches)	3	1121.39	373.797	28.34**
B (rows: treatments	5	34.94	6.988	
Error	14	184.72	13.19	

Compare your table from the GLM routine in your package to the adjusted ANOVA table value in Box 11.7 of Sokal and Rohlf (1995). Confirm whether the MS_{error} you obtained via the GLM routine is the same as the laboriously calculated MS_{error} in Sokal and Rohlf (1995)

Write-up for this lab.

1. Fill in the blank spaces as requested throughout Lab 7 (Data situations #1 #2 and #3) and submit to Brightspace.

NOTE: Fill in the blanks can be annotated using Adobe Reader (free to download).

2. Submit your labelled plots as a separate pdf file in Brightspace.