Laboratory #6.
The General Linear Model: Single Factor Analysis of Variance

The purpose of this laboratory is to demonstrate the use of computer packages to carry out ANOVA (Analysis of Variance) as a special case of the General Linear Model. In ANOVA, the explanatory variables are categorical (on a nominal scale). Each explanatory variable (called a factor) has two or more categories (also called classes or levels). ANOVA compares mean values among the categories within a factor.

Because we are using the General Linear Model, much of what you learned in the previous lab (regression) carries over to ANOVA. We will use the same steps to analyze the data and then evaluation the residuals.

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

- the capacity to re-organize data from tabular to model format
- greater facility in the diagnosis of residuals
- a working knowledge of ANOVA in the model format

At this point make sure that you have three data files:

Daphnia ages	DaphniaAges.txt	Box 9.5 in Sokal and Rohlf, 1995
Fly wing lengths	FlyWings.txt	Table 8.1 in Sokal and Rohlf, 2012
Cod movements	FishMove.txt	Cod movements by hour of day

These data files are available on the web at: www.mun.ca/biology/schneider/b4605/Data/Labs
Each file has a text version and an excel or comma save (csv) version.

Analysis #1 Fixed factor ANOVA. *Daphnia ages*.

The data for this example come from Box 9.5 (Sokal and Rohlf 1995).
The research question is:
Does average age at beginning of reproduction depend on strain in *Daphnia*?

To begin, we open the data file DaphniaAge.txt and look at the description of the data. At first glance we might think that each column is a variable. Looking more closely, we see that both columns list ages. We realize that we are looking at one variable (*Age*) arranged in two groups (2 strains).

The response variable (the one we plot on a Y axis) is: _____

What is the explanatory variable? _____ _____

Analysis #1 (continued) Fixed factor ANOVA. *Daphnia* Ages

Now list the variables along with a symbol and units, and the type of measurement scale (nominal, ordinal, interval, ratio) using *Age* and *Str* as symbols.

Variable	Symbol	Units	Scale Type	Role
				Response
				Explanatory

Now write the model in GLM format $\_\_\_\_ = \beta_o + \beta \_\_\_\_ + \epsilon_{normal}$

β_o = grand mean, β = contrast (Group mean $- \beta_o$)

Next import the Daphia Ages data to your statistical package and name each group (column) as *Strain1* or *Strain2*. In packages with a spreadsheet interface, this is readily done with copy and paste. In packages without a spreadsheet interface, the csv version is more practical.

See Lab 4

Define Data from file

Some packages allow analysis in this tabular format, a column of data for each group of a single explanatory variable.

Here is an example for two columns of data (factor with 2 levels).

```
MTB > aovoneway c1-c2
```

Run GLM Anova

ANOVA routines for tabular format are found in some but not all packages. To convert the tabular format to model based format we are going to collect all the values of the response level in a single column, then place labels next to each value in an adjacent column for the explanatory variable, which is a factor.

Here is pseudocode for re-organizing data from tabular to model format.

```
Place each value of the response variable into a single column.
Assign a name to this column (the response variable).
Label a new column for the first explanatory variable (X1).
For each value in Y, place the category label in the adjacent (X1)
column.
```

Reorganize from tabular to model format

This can be done in a spreadsheet before importing the data to the statistical package. It can also be done in the package after the data are imported. Here are two examples.

Line code in Minitab.

```
MTB > stack c1-c2 c3;
SUBC> subscripts c4.
MTB > name c3 'Age' c4 'Strains'
MTB > print c1-c4
```

Reorganize from tabular to model format

Analysis #1 (continued) Fixed factor ANOVA. *Daphnia* ages

R-code.

```
DaphniaAges   #This is the name of the input table
Daphnia<-stack(DaphniaAges)
names(Daphnia) <-c("Age","Strain")
Daphnia
```

Reorganize from tabular to model format

R-code yellow alert. Copying and running blocks of R-code short-circuits learning. After pasting a block of R-code into the script box in R-studio put the cursor on the first line of code. Then think about what that line of code does and what you expect to see in the console box. For example, with the first line of code you expect to see a display of the input table. With the second line of code you expect a new data object to appear in the environment. And you expect to see nothing appear in the console. With the final line of code you expect to see the new data object displayed, with names assigned to each column (vector) in the data object. You`ll need to understand the code because in Analysis #2 you will be applying the code to a new data set.

In this data set *Strain* is a fixed factor. It is fixed because the research question was whether these two groups, known to belong to known genetic strains, differed in age at first reproduction. The two strains were not being considered a random sample of strains from some population of strains. Inference is to only these two strains. In Analysis #2 we'll be looking at a random factor.

Here is pseudocode for GLM analysis.

Define the response variable, Y = age
Define the explanatory variable X = groups
Write the model and execute it.
Save residuals and fitted values.

Run GLM ANOVA with Residual diagnostics

Here are batches of code for Minitab, R, and SAS.

```
MTB > anova 'age' = 'groups';
SUBC> fits c10;
SUBC> residuals c11.
MTB > name c10 'fits'  c11 'res'
```

Run GLM ANOVA

```
MTB > glm 'age' = 'groups';
SUBC> fits....;
SUBC> residuals .....
```

Run GLM ANOVA

```
DModel<-lm(Age~Strain,data=Daphnia)
Dfit<-fitted(DModel)
Dres <-resid(DModel)
summary.lm(DModel)
```

Run GLM ANOVA

```
PROC GLM data=Daphnia;
  model Age=Strain; Class=Strain;
  output out=out1 r=res p=pred;
```

Run GLM ANOVA in SAS

Analysis #1 (continued) Fixed factor ANOVA. *Daphnia* ages

Having executed the model, we then evaluate the model.
Assumption 1. Straight line. No straight line model, so we skip this.

One of the advantages of the model-based style of analysis is that the four assumptions concerning the error distribution are the same for any GLM, so we will evaluate these just as we did with regression.

Assumption 2a and 2b. Residuals homogeneous and normal.

….Plot the residuals vs fitted values	***Errors homogeneous?***
….Plot normal scores	***Errors normal?***

Analysis #2. Random factor ANOVA. Fly winglengths

This analysis uses a new data set to extend what you have learned. The data are again in tabular format, so that you can strengthen your skill in converting data to model based format in your package. The model results are an opportunity to print out data equations, to widen your understanding of model based analysis of data. The data extend your experience with one-way ANOVA to the case of a random factor.

To begin, open the ASCII (text) file for the winglength dataset FlyWings.
The research question is: What is the variance in fly winglength among 7 random samples?
Fill in the table of variables, using *WL* and *Sample* as symbols.

Variable	Symbol	Units	Type	Role
				Response
				Explanatory

Use your symbols to write the model in GLM format. _____

Import the fly winglength data into your statistical package, then reorganize it from tabular to model format by modifying the code you developed in Analysis #1.

Analysis #2 (continued) Random factor ANOVA. Fly winglengths

Here is pseudocode for printing out data equations from a GLM.

Compute and display data equations

```
Place all values of the response variable in a single column Y
Place labels for each category in a column X (explanatory)
Calculate the mean for each category of X
Adjacent to each value of Y place the mean for its category
Label this column 'fits'
In a third column, calculate Y - 'fits'
Label this column 'resids'
Display the three columns and label them Data = Fits + Residuals
```

```
MTB > glm 'wlength' = 'sample';
SUBC> fits c10;
SUBC> residuals c11.
MTB > name c10 'fits'c11 'res'
MTB > print 'wlength' 'fits' 'res'.
```

```
# Input data reorganized to data object called Flies
FModel<-lm(WL~Sample,data=Flies)
Ffit<-fitted(FModel)
Fres <-resid(FModel)
WL<-Flies$WL
DataEq<-cbind(WL,Ffit,Fres)   #Data object with Data Eq
DataEq
```

Pick a winglength observation and write a data equation.

Data	=	Model	+	Residual
_____	=	_____		_____

The factor Sample in this data set is a random factor. It is random because the research question was whether groups differ. The groups were not formed according to named categories, as in an experiment. The samples were chosen without regard to any attribute of the group. Inference is to the population from which the samples were drawn. Inference is not restricted to the samples that were measured.

Evaluation.
Assumption 1. There is no straight line assumption with ANOVA
Assumptions 2a and 2b. Residuals homogeneous and normal.

….Plot the residuals vs fitted values

Errors homogeneous?

….Plot normal scores

Errors normal?

Do you see any strong heterogeneity in the residual vs fit plot ? _____
Do you see any strong deviation from normal distribution of residuals ? _____

Analysis #3. ANOVA in model format. Cod movements.

The next data set, in file 'FishMove.txt' was collected by Don Clark, a graduate student at Memorial University. The data are used with his permission. The research question was whether movements of juvenile cod *Gadus morhua* depended on time of day. The data set illustrates several perplexities that arise when analysing data. Should we treat time as aregression variable or a factor? Conversion of a numerical variable to a factor depends on the package. What if we cannot trust *p*-values because of gross violations of assumption? Is randomization worth the effort?

At this point you should close the session where you analyzed the fly winglength data, saving any material that you need later. Open a new session or worksheet in your package.

Open the ASCII data file FishMove.txt, and examine its structure. Information about the structure is at the end (bottom) of the file. The information you need to identify response and explanatory variables also occurs at the end of this file.
State a research question _____

In order to fill in the table of variables you will need to decide whether to model time as a trend (regression) or as a sequence of categories. To decide, plot the Distance again Time.

Define Data from file

Why is fitting a trend (regression line) inappropriate? _____
Now fill in the table of variables.

Variable	Symbol	Units	Type	Role
				Response
				Explanatory

Use your symbols to write the model in GLM format. _____

Pseudocode (applies to any statistical package)
Define the response variable, Y
Define the explanatory variable X
Execute model and save residuals and fitted values.
Check *p*-value assumptions of homogeneous normal residuals.

Run GLM ANOVA with residual diagnostics

This data set illustrates a common situation, numerical data for categories of a factor. GLM routines in Minitab and SPSS assume the explanatory variable is categorical unless it is declared as a regression variable (covariate). SAS assumes the explanatory variable is a regression variable, unless you declare otherwise in the GLM routine (shown in Analysis #1). R reads the explanatory variables as numeric, unless it encounters a non-numeric value. To convert a numeric variable to a factor in the data frame use the `as.factor` function.

```
FishMove$Time<-as.factor(FishMove$Time)
```

Write-up for this laboratory (two parts):

Lab 6a For the three data sets (DaphniaAges, FlyWings, CodMove), present your results using the following simplified format. Make sure you have everything you need before leaving the lab.

A. Write the model. State H_A/H_o pair about the model parameter, the difference in means.
B. State the test statistic for hypothesis testing
C. Show residuals vs fit plot, and comment on whether residuals are homogeneous.
D. Evaluate whether residuals are normal, with one graph (only) for evidence.
E. State whether a *p*-value via randomization needs to be computed. If it does, state whether randomization will change the decision to reject or not reject the null hypothesis.
F. Report the ANOVA table and state the decision about the H_o
G. Declare decision verbally with reference to the biology and goal of analysis.

Be sure to show at least one plot for each of 2 assumptions (homogeneous, and normal). Label each plot and refer to the plots in your evaluation of these two assumptions.

Lab 6b (extra)
Find data presented in at least three categories of a single explanatory variable. You can use data on the web. The course website has references to articles that display data in at least three categories.

Many students find it is quicker to skim a printed copy of a journal than to search on line. In general, articles published before 1980 display data more often than more recent articles. recent articles rarely display data but sometimes include online data.

Display the data in model format, showing variable names and first 10 lines of data. Be sure to list the full reference for the source of data.

Using the10-step generic recipe for hypothesis testing with the General Linear Model, present the results of your analysis of the data.