Laboratory #5b.
The General Linear Model: Regression with Randomization

The purpose of this laboratory is to give you further practice in the model-based (GLM) approach to regression. The General Linear Model includes many well-known tests as special cases (ANOVAs, t-tests, regressions, and analysis of covariance ANCOVA).

In confirmatory analysis, where the emphasis is on correctly computing the p-value, we examine residuals in order to diagnose whether our data meet the assumptions for computing p-values using a theoretical distribution such as the F-distribution. p-values calculated from F, t, or chi-square distributions cannot be trusted if the ***residuals*** are correlated, heterogeneous, or non-normal. Some people think the data have to be "normal" before undertaking analysis, but this is incorrect.

When assumptions are not met, we make a decision about whether to undertake randomization methods to recompute a p-value free of assumptions.

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

- capacity to decide when to use a randomization to compute a p-value

- a working knowledge of the mechanics of computing a randomized p-value for a general linear model.

At this point make sure that you have the cod mortality data set. Garrod.txt
Garrod.csv

This can be found on the Biology Department server.

www.mun.ca/biology/schneider/b4605/data

Analysis #4. Cod mortality in relation to fishing effort. GLM routine.

The next example shows the application of a randomization test to regression. The data are cod mortality rates in relation to fishing effort, reported by D.J. Garrod in the 1960s. Begin by opening the data file Garrod.dat, examining its structure. and then moving the data from this file to the statistical package.

Define Data from file

Generic recipe for regression of Y-variable against X variable

```
Pseudocode (applies to any statistical package)
        Define the response variable, Y = mortality
        Define the explanatory variable X = effort
        Etc.
```

Run GLM regression with Residual diagnostics

Box 1 (Analysis 4)

Are assumptions met?

Is the straight line assumption valid ? (any bowls or arches in residual vs fit plot?)

Are the residuals homogeneous ? (cones or other patterns in residual vs fit plot?)

Display the residual histogram and normal score plot.

Are the residuals normal ?

Decision to recompute p-value

Is the sample size small (less than 30) ?

Is the p-value close to $\alpha = 5\%$?

Is a decision concerning significance ($p < \alpha$) likely to change if we obtain a more accurate p-value by randomization ?

To compute a better p-value, we keep track of how many randomized F-ratios exceed the observed F-ratio ($F_{observed} = 4.91$). We will do this by hand, to illustrate the procedure.

Table 1. Randomization tests, tallied by hand.

1. Write down the observed F-ratio, from the ANOVA table.
2. Randomize the data.
3. Compute a randomized F-ratio.
4. Keep a tally of the number of F-values less than or equal to $F_{observed}$, and the number greater than $F_{observed}$.

Analysis #4 (continued)

Here is a generic recipe for carrying a randomization test, using the GLM.

GLM p-values by randomization

```
Pseudocode (applies to any statistical package)
        Sample from Y variable  into a new column = ‘Yrandom’
        Regress ‘Yrandom’ against the explanatory variable X
        Record the F-ratio.
        Repeat, and tally the number of F-ratios greater the $F_{observed}$
        Divide by the number of F-ratios computed to obtain p-value.
```

Here are the specifics for obtaining a single randomized F-ratio in Minitab.

```
Minitab menu.
     Calculate
        Random data
              Sample from columns
                   Sample 13 rows
                   From column (mortality)
                   Store in c7
Name c7 ‘RanMort’
     Statistics
         ANOVA
                General Linear Model
                   Response ‘Ranmort’
                   Model  ‘effort’
                   Etc.
```

```
MTB > sample 13 'mort'  'Ranmort'
MTB > glm 'Ranmort' = 'effort';
SUBC> covariate 'effort'.
```

Now record (in Box 2) whether this ratio falls above or below the ratio (F=4.91) for the data before it was randomized.

Box 2 (Analysis 4)

# ≤ 4.91	# > 4.91

Record a tally in the appropriate column, you do not need to record the value of each F-ratio.

Analysis #4 (continued)

Next, we want to generate many of these randomized F-ratios. To do this, we run the same code repeatedly. Here is one way to run a sequence of commands repeatedly.

```
Minitab session
Scroll upward to find code that produced the randomized F-ratio
Highlight (select) code producing random F-ratio
Scroll down to active command line.
Paste code onto active command line.
Run the code by pressing the Return key.
```

Here is another way to run a sequence of commands repeatedly.

```
Minitab session
Scroll upward to find code that produced the randomized F-ratio
Highlight (select) code producing random F-ratio
Minitab menu
    Edit
        Command line (selected code should appear)
        Submit
```

Be sure to tally this randomized F-ratio in the table provided above.
Now run the batch file repeatedly and fill in the following tally sheet (Box 2, above) for 100 randomized F-ratios:

After 100 repeats, you will have a tally of the random F-ratios, above and below $F_{observed}$

Record the number (out of 100) that exceeded $F_{observed}$ ________

Record the percentage (out of 100) that exceeded $F_{observed}$ ________

This percentage is a rough estimate of the p-value for the randomization test.

If none of your randomized F-ratios were greater than 4.91, then your p value is less than 1 in 100 ($p < 1/100 = 0.01$). **It is not zero.**

To obtain a really good estimate of the p-value, we need more than just 100 randomized F-ratios. This is a lot of work unless we can use the statistical package to accumulate randomized F-ratios for us. To accomplish this we save a series of commands in a file (called a batch file), run the file to make sure it works, then turn the batch file over to a command that runs the file hundreds or even thousands of times. Here is the pseudo code that can be run to accumulate randomization results automatically.

```
Pseudocode (applies to any statistical package)
        Run a sequence of commands to produce randomized p-value.
        Store commands in a batch file.
        Run the batch file (correct the file if it doesn't work).
        Use special command to run batch file repeatedly.
```

Here is the Minitab procedure.

```
Minitab session
Scroll upward to find code that produced the randomized F-ratio
Highlight (select) code producing random F-ratio
Minitab menu
   Edit
      Command line (selected code should appear)
      Save 'batch.txt'   (to file to the  desktop)
   File
      Other
         Files
            Run an Exec
```

Here is the batch file to produce randomized F-ratios for the cod mortality data. Note that the minitab prompts do not appear in the batch file.

```
sample 13 'mort'   'Ranmort'
glm c7 = 'effort';
covariate 'effort'.
```

Try generating this file on your desktop and running it using the minitab Execute command.

This file produces a randomized F-ratio each time it is run, but fails to accumulate the randomized F-ratios. Here is a sequence of commands that will calculate the F-ratio and place it into a column for later use.

```
MTB > glm 'mort' = 'effort';
SUBC> covariate 'effort';
SUBC> fits c24;
SUBC> residuals c25.
MTB > let c24 = c24 - mean('mort')
MTB > let k1 = ssq(c24)/1           MSE for model
MTB > let k2 = ssq(c25)/(N('mort')-2)   MSE error
MTB > let k3 = k1/k2                       F-ratio
MTB > stack k3 c26 c26
```

Once we have a sequence of commands that will store the F-ratio, we can then generate and store random F-ratios.

Here is the batch file to obtain randomized F-ratios for the cod mortality data, then store these ratios in a column.

C22 is the randomized response variable (Y = Ranmort).
Randomized F-ratios accumulate in c26. As with all batch files, the minitab prompts are omitted.

```
sample 13 'mort' c22
glm c22 = ‘effort’;
covariate ‘effort’;
fits c24;
residuals c25.
let c24 = c24 - mean(c22)
let k1 = ssq(c24)/1                   MSE for model
let k2 = ssq(c25)/(N(c22)-2)             MSE error
let k3 = k1/k2                              F-ratio
stack k3 c26 c26
```

While this batch file can be generated by typing the commands in minitab, then copying and saving them, it is usually easier to use a text editor to create the batch file directly on your desktop. The Notepad program automatically creates a text file. If you use a word processor, you will need to use ‘Save As’ to create a text file (ascii format under file type during save routine).

After you have created the batch file in text format on your desktop, try running it once, using execute command. If it fails to execute, you will need to edit corrections into the file until it does work. Once it does work, try executing it 10 times, using the Execute command. Once the file runs safely 10 times, go for it and accumulate several hundred randomized F- ratios. Once you have at least 500, scroll down and find out how many you have

Record the number of randomized F-ratios ________

Next, find the number of randomized F-ratios that are less than F = 4.91, the observed F-ratio.
One way to do this is to sort the randomized F-ratio into a new column.
Then scroll down the sorted column to $F_{observed}$
Then note the line number on which it appears. This gives you the number of F-ratios that were less then $F_{observed}$

Record the number of randomized ratios that were less than $F_{observed}$ ________

Calculate the number of randomized ratios that exceeded $F_{observed}$ ________

Record the percentage that exceeded $F_{observed}$ ________

Now compare the rough estimate (Box 2, around 100 F-ratios) to this improved estimate based on at least 500 ratios.

Box 3 (Analysis 4)

	Rough (About 100)	Improved (>500)
Number of randomized F-ratios	_______	_______
Number of randomized ratios less than $F_{observed}$	_______	_______
Number of randomized ratios that exceed $F_{observed}$	_______	_______
Calculate the percentage that exceed $F_{observed}$	_______	_______

Write-up for laboratory 5b.

Please do not hand unlabelled computer output! Instead, cut sections of output to a document and label each section in the document. When pasting tabular input into your document, use a non-scalable font (`such as Courier 10`) to display this material. Otherwise the material will be distorted and nearly unreadable. Box 3 is best completed during the lab session.

Analysis #4

Complete Box 3.

Use the generic recipe for hypothesis testing with the general linear model (see lecture notes) to write up the analysis of the cod mortality data (Garrod.dat). Include appropriate plots with comments on whether a straight line is an appropriate model and whether residuals are homogeneous and normal. State the population (target of inference) as best you can.