Laboratory #5a.
The General Linear Model: Regression

The purpose of this laboratory is to give you practice in using the General Linear Model in the analysis of data. The GLM includes many of the most frequently used procedures in statistical analysis--ANOVAs, t-tests, regressions, and analysis of covariance ANCOVA. The GLM applies in exploratory analysis (What is the best model?), in evidential analysis (What is the evidence?), and inferential analysis (What is probable, given a model?).

The GLM is a linear model (additive effects) with a normal probability model (Seber, G.A.F. 1966. The Linear Hypothesis: A General Theory. London, Griffin). Consequently, examination of residuals relative to the normal error assumption is an important part of the execution of the GLM in all three modes. In exploratory analysis, we examine residuals in order to diagnose whether our description of pattern in the data (the formal model) is adequate. If the residual versus fit plot shows a horizontal band, with no bowls or arches, then we have arrived at an adequate description of pattern in the data. If the residuals do show pattern, then we can use observed patterns in the residuals to construct a better model of the relation of the response variable to explanatory variables.

In an evidential analysis, we examine residuals in order to diagnose whether the probability model we are using is appropriate for the data. The general linear model assumes a normal (fixed error) model and so we will be looking at whether the residuals vary in a systematic way with the fitted values, or with any of the explanatory variables.

In inferential analysis, we examine residuals in order to diagnose whether our data meet the assumptions for computing long run probabilities from *F, t,* or χ^2 distributions. A *p*-value calculated from these distributions cannot be trusted if the ***residuals*** are correlated, heterogeneous, or non-normal. If the residuals deviate substantially from normality, we cannot make appeal to the law of large numbers when we calculate probabilities. Some people think the data must be checked for "normality" before undertaking analysis. This is a widespread misunderstanding of regression and ANOVA.

Once you have completed the lab, you should have

- capacity to undertake regression analysis using either a GLM or regression routine in a statistical package.

- a working knowledge of the mechanics of residual analysis in a statistical package

At this point make sure that you have two data sets

TriboliumWeights.txt	Box 14.1 in Sokal and Rohlf 2012
TriboliumSurvival.txt	Box 14.4 in Sokal and Rohlf 2012

The data sets can be found on the course website.
https://github.com/DavidCSchneider/StatisticalScience/tree/main/Data/Labs

Analysis #1. *Tribolium* weight in relation to humidity. Regression routine.

Most statistical packages contain separate routines for regression and for the general linear model. Analysis #1 demonstrates the general linear model and residual analysis with a regression routine. The example comes from a Box 14.1 in the text by Sokal and Rohlf (1995). The research questions are:

Does weight loss in the flour beetle *Tribolium* depend on humidity ?

If it does, what is the rate of weight loss in relation to humidity ?

To begin, we open the data file and look at the description of the data. From this information we list the response and explanatory variables along with units, the type of measurement scale, and a symbol for use in the model we will write. Both the response and the explanatory variable are on a ratio type of scale so our model will be a regression.

Variable	Symbol	Units	Type	Role
Weight Loss	*WLoss*	mg	ratio	Response
Percent humidity	*PctH*	%	ratio	Explanatory

The model is $WLoss = \beta_o + \beta\ PctH + \epsilon_{normal}$ β_o = grand mean, β = slope, normal error

The regression routine will report: $WLoss = \alpha + \beta\ PctH + \epsilon_{normal}$ α = *Y* intercept

In Lab 4 you learned how to import data from a text file. If you are using a package with a spreadsheet interface, go ahead and copy the data from the text file and paste it into your package. For users of R studio, an excel file version of the ASCII (text) file is available on the course website at the same location as the text file. After downloading the excel file, go to R-studio and import the data file as follows.

Define Data from file

Import DataSet (Environment window in RStudio)
Choose Excel
Browse to find data file
Examine dataset that appears
Import to a data object in R

Here is pseudocode for regression, analysis of a *Y*-variable against the *X* variable

Define the response variable, *Y*
Define the explanatory variable *X*.
Write the model.

Write the model

Define the data set in statistical software package
Run the analysis to obtain parameters, fitted values, residuals

Run regression

Save residuals and fitted values.
Plot residuals vs fitted values to check linear assumption.

Extract residual diagnostics

Revise model if linear assumption not met.
Evaluate residuals for homogeneity and normality.

Evaluate model

Analysis #1 (continued).

Here is line code for Minitab, then for R.

```
MTB > plot 'WLoss' * 'PctH'
MTB > regress 'WLoss' 1 'PctH';
SUBC> fits c4 ;
SUBC> residuals c5.
MTB > name c4 'fits' c5 'res'
```

Plot Data
Run regression
Obtain residuals and fits

```
View(TriboliumWeights)
names(TriboliumWeights)=c("WLoss","PctH")
with(TriboliumWeights,plot(PctH,WLoss,
  xlab="Relative humidity (%)",
  ylab="Weight loss (mg)",
  main="Figure 1. Tribolium weight loss relative to
          humidity"))
```

View imported data
Label columns
Plot Data

```
Wtmod<-lm(WLoss~PctH,data=TriboliumWeights)
```

Run regression

```
res1<-resid(Wtmod)
fit1<-fitted(Wtmod)
```

Obtain residuals and fits

Residuals and fits are saved for later use in both menu-based and code based packages.

When we use the general linear model (as in regression) we make assumptions about both the structural part of the model ($\hat{Y}$) and about the error distribution. $Y = \hat{Y} + \varepsilon_{normal}$
We will begin with the assumptions about the structural part of the model, then move to the assumptions about the error distribution.

Assumption 1. A straight line model is appropriate.
To check this assumption we look at the residual *vs* fit plot.
Some statistical packages produce this plot automatically as an option for regression output.

```
MTB > plot 'res' * 'fits'
```

Model linear?

```
plot(fit1,res1,
     ylab="residuals",
     xlab="fitted values",
     main="Figure 2. Tribolium weight loss
     residual vs fit plot")
```

Model linear?

There are no obvious bowls or arches in the plot. Assumption 1 is met.

Analysis #1 (continued).
So, we'll continue with the straight line regression model.
In most packages executing the regression routine produces the parameter estimates and a measure of fit r^2. With R we need to issue a command to extract coefficients and the ANOVA table from the model object.

```
summary.lm(wtmod)
anova(wtmod)
```

Obtain estimates
Obtain ANOVA table

Write the regression equation with parameter estimates, immediately below the structural model.
The structural model is: $\hat{Y} = \alpha + \beta_x X$

Regression equation: $WLoss =$ _____ + _____$PctH$

Statistical packages report results in the familiar slope intercept form, as above. The statistical package calculates the *Y*-intercept α from β_o. Try this calculation yourself. Use your package to calculate the means for *Y* and *X,* then the Y-intercept α.
$\bar{Y} = \beta_o =$ _____
$\bar{X} =$ _____
$\alpha = \beta_o - \beta_x \bar{X}$ _____

Next we move to the error component of the GLM. The normal error model rests on four assumptions. To diagnose these we use the residuals.

Assumption 2a. Homogeneous residuals.
This is the most important assumption. Violations of this assumption will have the greatest biasing effect estimates of parameters and on the *p*-value. This assumption is checked by going back to the residual *vs*. fit plot. We look at this graph in a new way—Do we see strong differences in vertical spread, from left to right in the plot?
Do you see any strong deviations from a homogeneous band in the residual *vs* fit plot for the *Tribolium* data ? _____
Assumption 2b. Residuals sum to zero.
Statistical packages estimate parameters in a way that makes this true, so we don't need to check this assumption.
Assumption 2c. Independent residuals.
One way to check the independence assumption is to plot the errors in the order in which the data were collected. Another way is to plot each residual against neighboring values in space. Use Help in you package to find the routine for a lag plot to check this assumption.

```
MTB > let c10 = lag('res')
MTB > plot c10 * 'res'
```

Errors independent?

```
lag.plot(res1,
         main="Figure 3. Lag plot. Tribolium weight
         loss residuals",
         diag=FALSE, do.lines=FALSE)
```

Errors independent?

The regression residuals from the analysis of *Tribolium* data show no obvious upward or downward trends, so residuals are taken to be independent.

Analysis #1 (continued)

Assumption 2d. Normal residuals.

This assumption tends to attract more attention than the homogeneity assumption, when in fact violations of the homogeneity assumption often have a greater distorting effect on parameter estiamtes and the *p*-value estimate. The two commonest graphical diagnostics for this assumption are histograms and normal probability plots.

Assumption 2d. Normal residuals. Checked with a histogram.

If the normality assumption is met the residuals will cluster symmetrically around their mean (zero). If the assumption is not met the histogram will deviate substantially from a bell-shaped curve. Find the histogram routine in your package and check this assumption now.

```
MTB > histogram 'res'
```

Errors normal?

```
hist(res1,breaks=9,
     xlab="residuals",
     ylab="frequency",main="Figure 4. Tribolium
     weight loss")
```

Errors normal?

Evaluation of the histogram is difficult when there are few residuals, as in the *Tribolium* data. The visual impression from a histogram can depend very much on the number of classes used to construct the histogram. Use your package to replot the histogram with fewer classes. A symmetrical distribution is expected so use an odd number of classes (5 instead of 9).
To accomplish this, revise the code (SAS, Rstudio) or re-run the routine from a menu.
In the box below describe the differences in the two histograms.

Assumption 2d, Normal residuals. Checked with a normal probability plot.

We can check the normality error assumption by comparing the cumulative distribution of the residuals to the cumulative distribution of the normal distribution, which is sigmoid (S-shaped). Normal probability plots use a suitable transformation to straighten out the sigmoid curve into a straight line that rises diagonally. Normal residuals fall on the diagonally rising straight line, while non-normal residuals will deviate from the line. Find the normal plot routine in your package and check this assumption now. Here are line code versions.

```
MTB > nscores 'res' c30
MTB > plot c30 'res'
```

Errors normal?

```
qqnorm(res1)
```

The plot shows some deviation from the diagonal line, with too many residuals near zero.

Analysis #1 (continued)

Should we use hypothesis testing to check for normality?

Statistical packages contain tests of normality, which yield a *p*-value made against a 5% criterion. While a test of normality is useful in some circumstances, it guarantees a bad decision when examining residuals. A test of normality will reliably produce a small *p*-value when sample size is large, even though the deviations are small. A test of normality at small sample sizes will often produce a large *p*-value, even though there are substantial deviations. So a test of normality usually results in a decision that the assumption is not met when sample size is large, when deviations from normality become less important. A test of normality will usually fail to detect large deviations from normality when sample size is small, when deviations from normality are important. Statistical tests of the normality assumption can be relied upon to lead to the bad decisions.
For a fuller treatment of the topic see Johnson (1999) and more recently, Läärä (2009).
Johnson, D.H. 1999. The insignificance of statistical significance testing. *Journal of Wildlife Management* 63:763-772
Läärä, E. 2009. Statistics: reasoning on uncertainty, and the insignificance of testing null. — *Ann. Zool. Fennici* 46: 138–157.

Evidential strength and effect size.

Having evaluated the model, we then calculate a measure of the strength of the evidence for the model with the parameter $\hat{\beta} = -0.053$ compared to a model with $\beta = 0$. The likelihood can be calculated from the explained variance r^2 and the sample size *n*.

$$LR = (1 - r^2)^{-n/2}$$

From the regression output we see that $r^2 = 0.9787$ with $n = 10$.

$LR = (1-0.9787)^{-10/2} = 2.3 \times 10^8$

The estimated effect size was $\hat{\beta} = -0.053$ mg per 1% change in humidity. This effect size was ***far*** more likely than an effect size of $\beta = 0$. It was 10^8 times more likely than $\beta = 0$.

Inference from the likelihood ratio.

Hypothesis testing uses the likelihood ratio to make a decision about the null hypothesis, no change. Is hypothesis testing appropriate? In this example we have an experiment from a lab where beetle cultures were maintained in controlled temperature chambers. The experiment could have been repeated many times under nearly identical conditions. The population (target of inference) was a potentially large number of repeats, large enough to allow appeal to the law of large numbers and the normal distribution. The probability of the estimated slope of $\hat{\beta} = -0.053$ under conditions of fixed humidity (*PctH* fixed, hence $\beta = 0$) is calculated from the *F*-statistic, which in turn is based on the *LR*. From the regression routine output we see an *F*-ratio of 367, an extremely improbable result ($p = 5.7 \times 10^{-8}$) that we reject at the 5% criterion. When we use hypothesis testing we are restricted to a statement of rejection of the H_o. We cannot "accept the alternative" having rejected H_o. We cannot say we accept H_A when H_o is not rejected. And we cannot even say we accept the H_o when it is not rejected. If we want to say something about the alternative hypothesis, we use a likelihood ratio.

Do we need to recompute the p-value?

The *p*-value in hypothesis testing assumes that errors are independent, homogeneous, and normal. In the evaluation of the analysis of the *Tribolium* data, we found some evidence of deviations from normality. We have already learned a remedy: recompute the *p*-value by randomization. This produces a *p*-value free of the assumptions required for the *F*-distribution. In most packages this is a lot of work. So we ask at this point whether the extra work is necessary. If the p-value is close to the criterion for significance ($\alpha = 5\%$) then an incorrect *p*-value can lead to an erroneous decision. But if the *p*-value is far from the criterion (say by a factor of 5 or a factor of 1/5) then a better *p*-value won't change the decision. No matter how badly the assumptions are violated, *p*-values from the *F*-distribution almost never deviate from the randomization *p*-value by factors as large as 5 or 1/5. For the *Tribolium* data the *F*-statistic was huge ($F = 367$) and hence the *p*-value is minuscule ($p < 0.001$). If we recompute the p-value via randomization we get a more accurate p-value, we put time into the effort, but the decision (that the slope of the regression is not zero) will not change. So for this lab we are not going to put time into an effort that won't change our decision. We will use judgement to reject the null hypothesis, that weight loss does not depend on humidity.

Analysis #2. *Tribolium* weight loss in relation to humidity. GLM routine.

Regression routines assume all explanatory variables are continuous. GLM routines allow both continuous and categorical variables. The `lm` command in R is a GLM routine allowing both continuous and categorical variables. The `glm` command in R and Stata performs a Generalized Linear Model, which allows non-normal error structures. So if you are using R you have already used a GLM routine and you can skip to Analysis #3. If you are using SAS or Minitab, you can acquaint yourself with the GLM routine output by going back to the beginning of the analysis of the *Tribolium* data and repeating it with the `GLM` command. ..

Run GLM regression in SAS

```
PROC PLOT data=srbx14_1; plot WLoss*PctH;
PROC GLM data=srbx14_1;
  model WLoss=PctH;
  output out=out1 r=res p=pred;
PROC PLOT data=out1; plot res*pred/vref=0;
PROC UNIVARIATE data=out1 normal; var res;
```

Run GLM regression In Minitab

```
MTB > plot 'WLoss' * 'PctH'
MTB > glm 'WLoss' = 'PctH';
SUBC> covariate 'PctH';
SUBC> fits c4 ;
SUBC> residuals c5.
MTB > name c4 'fits' c5 'res'
```

Re-running the analysis to compare outputs should not take more than about 5 minutes, as the routines are very similar in structure and output in these two packages.
If you are using SPSS or STATA, you will be using the `anova` command where the explanatory variable is declared as continuous, as in the Minitab code shown here. This may take longer because of the differences between the regression and anova routines in these two packages. Because all of these routines store residuals, we can use the same simple graphical diagnostics shown already for any GLM.

.

Box 1 (Analysis 2)

After you have run the GLM routine make the following comparisons.

Display and compare the estimates of the Y-intercept α (GLM and regression routines)
Display and compare the estimates of the slope β_x.
Display and compare the ANOVA tables. Same structure? Same F-ratio ?

Analysis #3. *Tribolium* survival in relation to egg density. GLM routine.

The next example demonstrates regression analysis of more than one *Y*-value for each *X*-value. The data set will be another text example, taken from Box 14.4 in Sokal and Rohlf (1995). The research question is, Does *Tribolium* survival depend on density?

As before we open the data file TriboliumSurvival. We look at the description of the data. From this information we list the response and explanatory variables along with units, the type of measurement scale, and a symbol for use in the model we will write. Fill in the table.

Variable	Symbol	Units	Type	Role
	PctS			Response
	asinS			
	Eggs			Explanatory

Now, using the symbols, write the model for *PctS* in slope-intercept form.

At this point it is a good idea to close your previous session, saving any work you need, and start a new session. Then download the csv version of this file (on the course website) to your desktop. You can either open it to copy and paste, or browse for it and import it. Bring in both survival and arcsin(survival), as well as egg density.

Define Data from file

GLM routines allow categorical explanatory variables as well as regression variable. Minitab assumes categorical and so a special command is required to declare a regression variable.

```
MTB > glm 'PctS'  =  'Egg';
SUBC> covariate '[Egg]';
SUBC> fits c4;
SUBC> residuals c5.
MTB > name c4 'fits'  c5 'res'
```

Run GLM

The SAS GLM routine (shown in Analysis 1) assumes regression.

Analysis #3. (Continued).

R selects a data class when the data were imported. R tries in turn `logical, integer, numeric` and `complex,` moving on if any entry is not missing and cannot be converted. If all of these fail, the variable is converted to a `factor`.

R-studio might (or might not) tell you what it did. To find out what R did, issue the command:

```
str(TriboliumSurvival)
```

Alternatively you can go to the environment window, click list, and see the same thing. You'll see that two of the variables were read in as `numeric`, one was read as `integer`, and none were read in as `factor`, *i.e.* as a categorical variable. For now, because we are using regression, `integer` is ok for the response variable. In Lab6 we'll use the functions `as.factor` and `as.numeric` to manage categorical variables in R.

Using what you learned in Analysis #1, run GLM regression for the Tribolium survival data.

	Run GLM
	Model Linear?
	Residuals homogeneous?
	Residuals independent?
	Residuals normal?

The response variable is a percentage, and hence we might expect the residuals to be non-normal. Instead, we found that the residuals were normal but not homogeneous.
For decades, textbooks have recommended the arcsin transformation of response variables that are percentages. The arcsin transform became a ritual, apparently no-one checked to see if there was a problem or if the arcsin ritual eliminated the problem. Let's see if the arcsin changes the residual diagnostics. Redo the analysis, using the column of arcsin transformed data. This will produce the ANOVA table shown in Box 14.4 of the Sokal and Rohlf 1995 text.

Box 2 (Analysis 3)

For the analysis of arcsine transformed survival:

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

_____ Are the residuals homogeneous ? (No cones or other patterns in residual vs fit plot?)

_____ Are the residuals normal ? (use normal scores/qqplot)

_____ Did the arcsin transform change the residual diagnostics concerning normality ?

For more on the topic of the arcsin transform see Warton and Hui 2011 *Ecology* 92: 3-10.

When assumptions are not met, we next ask whether it is worth taking the time to recompute a p-value by randomization.

Box 3 (Analysis 3)

Is the sample size small (less than 30) ? _____.

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

Given your answers, is the decision based on this p-value likely to change if we obtain a more accurate p-value by randomization ? _____.

Extra. ANOVA tables and F-ratios for regression with several Y-values for each X-value.

The recommended procedure with several Y-values is to form the F-ratio based on all of the observations, not on just the means for each group (Freund, J. 1971. *Mathematical Statistics*. Prentice-Hall). Using all of the data gives proper weight to each group–the larger the sample size, the greater the weight for that group. Proper weighting is also achieved by using the sample size when using the means. The two methods (all of the data, or means weighted by sample size) should produce the same ANOVA table and parameter estimates. Try carrying out the regression analysis using means weighted by n = sample size. This will produce the correctly weighted ANOVA table in Box 14.4 of Sokal and Rohlf 1995.

Write-up for this laboratory.

Please do not hand unlabelled computer output! Instead, cut sections of output to a document and label each section in the document. If you paste tabular results such as ANOVA tables into your document, use a non-scalable font (`such as Courier 10`) to display this material correctly. Otherwise the material will be distorted and nearly unreadable. Make sure to label all plots on both axes with name of variable and units where appropriate, and to add a caption with source of data and type of plot. To be sure they have everything, students often make a rough draft of their write-up before leaving the lab.

Analysis #1

Present your results from the analysis of the *Tribolium* weight data, using the following simplified format.

A. Write the statistical model, state H_A/H_o pair about the explanatory variable.
B. Present the ANOVA table with F-statistic.
C. Show residuals *vs* fit plot, and comment on whether straight line assumption is met.
D. State whether residuals are homogeneous, normal, and independent. Include appropriate plots with comments on each assumption.
E. Report decision- H_o, reject or not rejected, with report of statistic, n, and p-value.
F. Report effect size (slope parameter with units) If H_o rejected interpret effect size with reference to the research question.

Analysis #2

If appropriate (packages other than R), complete Box 1 and comment on the advantages and disadvantages of the regression routine compared to the GLM routine.

Analysis #3 - *Tribolium* survival.

Calculate the *LR* for untransformed and for transformed data. Comment on the difference (Small? Large?), with respect to strength of evidence (*LR*), homogeneity of residuals, and normality of residuals.

Support your comments on the differences by referring to labelled residual plots for both survival and arcsin(survival) in relation to density.

Complete Boxes 2 & 3. Refer to residual plots in completing Box 2.

Show residual plots for untransformed and arcsin transformed data as in step D in Analysis #1. Make sure the plots are clearly labelled as untransformed or arcsin transformed data.