Model Based Statistics in Biology.
Part IV. The General Linear Model.
Chapter 15 GLM - Review

ReCap. Part I. Quantitative Background (Chapters 1,2,3,4)
ReCap Part II Quantifying Uncertainty (Ch 5, 6, 7)
ReCap Part III The General Linear Model
Single Explanatory Variable (Ch 9, 10, 11)
Part IV. GLM. Multiple Explanatory Variables.
12 Multiple Regression
13 GLM ANOVA
14 GLM ANCOVA
15 Review - GLM with multiple explanatory variables.

ReCap Part I (Chapters 1,2,3,4) Quantitative reasoning: Example of scallops, which combined models (what is the relation of scallop density to substrate?) with statistics (how certain can we be?)
ReCap Part II (Chapters 5,6,7)
Hypothesis testing uses the logic of the null hypothesis to make a decision about an unknown population parameter.
Estimation is concerned with the specific value of an unknown population parameter.

The General Linear Model.
Advantage of learning unifying concepts rather than list of statistical tests.
GLM a general procedure that is more useful and flexible than a collection of special cases.

Today: GLM Review
Concepts old and new
Table 15.1
Questions
Purpose --> Be able to...

on chalk board

Review session at last lecture before exam. Bring questions.

Today: Review of the General Linear Model

Review technique (1995-1998):
Start with general commentary, key points for each step of the Generic Recipe.
Then move to questions (brought in for answering).
Then at end do 5 minutes on purpose of course, moving to how this translates into testing (GLM exam).

Review technique (1994-1998):
Diagrammatic summaries in Lab 9 used for review in lab.
Move this material to review at end of section on exploratory analysis, because material in lab 9 includes exploratory, while lab 9 uses confirmatory data situations, no examples where exploratory would be appropriate.

Review technique (Fall 2000 and 2003):
Review lecture emphasized key concepts in course, brought together in execution of GLM.
Did not use commentary.

Other review techniques.
Ask for projects or theses, from people in class, then set up as GLM analysis via dialogue.

Introduction.
In this part of the course we looked first at a single response variable Y as a function of one or more explanatory variables X

GLM Y = f(X)
Regression: X is on a ratio scale
relation of Y to X expressed by a line
One-way ANOVA: X is on an nominal scale (categories)
relation of Y to X expressed as a set of means, one for each category

Then we moved to multiple explanatory variables. GLM Y = f(X1, X2, etc)
The explanatory variables can be regression variables . . . . . Multiple Regression
They can be categorical variables . . . . . Multiway ANOVA
They can be a mixture of
regression and categorical variables . . . Analysis of Covariance (ANCOVA)

Multiple explanatory variables have several uses in statistical analysis.
-Analyze the effects of one variables, controlling for another (statistical control)
-Efficient experimental design.
Multifactor experiments yield more
information per unit of effort than single factor experiments. This is important when the costs of experimental units are high. In agriculture, the spatial unit may be small (plot) but requires considerable preparation and time to obtain results.
In laboratory studies, we wish to minimize the number of experimental animals for ethical as well as economic reasons.
-Combine like studies to yield new insights.

Concepts old and new. List of Terms

Terms in bold on board
Tried in 2003, worked OK

Here are terms learned so far in the course. These terms cover most of the important concepts in this course.

Response (dependent) and **explanatory** (independent) variables.

GLM consists of a structural model (explanatory variables) and error model.

Regression (ratio or interval scale, continuous but also counts) variable
vs **Categorical** (nominal scale) variable (factors with levels)

Random vs **fixed** categorical variables.

Arranging data into **Model Format**. **Data Equations**

Parameters: **Means, Slopes** μ notation versus β notation

Variability: **Variance, SS, df, MS**, Variance ratio (**F**)

Concepts old and new. List of Terms (continued)

Analysis of Variance: partition SS

Type I and II error. p-value (Type I error) from pdf, from cdf

Assumptions
for regression (straight line) (bowls/arches?)
for estimating parameters (fixed variance) (cones/spindles?)
for p-values from cdf (4)

Assumptions not met

Hypothesis testing. $\mathbf{H_A}$ / $\mathbf{H_0}$ for parameters,
for variance due to each term in model

Declare decision (conventional format: statistic, sample size or df, p-value).

Report evidence: **Likelihood ratios**

Analyze parameters (best practice, but not widespread).

One-tail vs Two-tailed tests.
F-tests always one tailed (equivalent to two-tail test on means)
t-tests one or two tailed.

Sequential vs adjusted sums of squares

If we have more than one explanatory variable, and the two explanatory variables are themselves correlated, then the partitioning of the SS will depend on the order in which the terms appear in the model. The first term in the list will capture the variance shared with the other term, in addition to the variance due only to the first term. The second term can only capture the variance due to that term. This problem arises as soon as we put a regression variable into our model. The problem does not arise if all of the explanatory variables are categorical variables (factors). ANOVA factors are independent of one another by definition. ANOVA factors are not necessarily independent of regression variables, however.

To solve the problem of results that depend on the order in which we list terms in our model statement, we use the SS assigned to a variable if it occurs *last* in the model. This is called the Type III or adjusted SS. It is the variance due to a term after we have corrected for the effects of all the other terms in the model. This contrasts with the Type I or Sequential SS, where partitioning of SS_{total} is according to the order specified in the model statement. Type I (Seq SS) add up to SS_{total}, while Type III (Adj SS) do not.

In any ANOVA table the SS should be clearly labelled as Type I (sequential) or Type III (adjusted) SS.

Concepts old and new. List of Terms (continued)

Interactive effects (interaction terms)

Forming the interaction terms
Number of interaction terms
Interaction terms first

Forming interaction terms. General linear models often include interaction terms in addition to main effects due to each explanatory variable. These terms describe the interactive effect of two (or more) terms on the response variable. The term is formed as the product of 2 or more explanatory variables. The degrees of freedom are computed as the product of the degrees of freedom of each main effect.

	Explanatory variables (Main effects)			interaction terms			
	A	B		A*B			
df	2	3		2*3			
	A	B	C	A*B	A*C	B*C	A*B*C
df	3	4	2	3*4	3*2	4*2	3*4*2

Number of terms. The number of possible terms will depend on the number of explanatory variables. We can have two-way interactions, three-way interactions, etc.

Number of Explanatory Variables	Number of two-way terms	Number of three-way terms	Number of four-way terms	Total
1	0			0
2	1	0		1
3	(3*2)/2 = 3	1	0	4
4	(4*3)/2 = 6	3	1	10

It is evident that interaction terms proliferate exponentially as we add explanatory variables. There are several criteria from omitting or removing interaction terms from our model.

Concepts old and new. List of Terms (continued)

Interactive effects (interaction terms)

Number of terms.

(1) If both factors are random, the interaction term is logically absent. It cannot be estimated because both factors are random and hence we have no way of matching levels in one factor with levels in another factor. In the flies within cages example, we cannot match flies across cages. This results in hierarchical ANOVA, in which the levels of one explanatory variables are nested within another.

(2) If one factor is random, then the interaction term can be estimated but is dropped if it is absent by design. We can estimate the interaction because we have enough information to match the levels of the fixed factor across levels of the random factor. In the randomized block example, we can match the genotypes (fixed) across the experiments (random factor). But while the interaction term can be estimated, it is usually dropped from the model if it is expected to be zero because levels of the fixed factor have been assigned randomly to units within the random factor (experiments, locations, time periods, etc). In the long run the differences among levels of the fixed factor will be independent of a factor if fixed effects are assigned randomly.

However, if fixed effects are assigned haphazardly rather than by deliberately random assignment then interactive effects can be present. For example, we might assume that time of day does not affect the experimentally induced differences in the fixed factor. But if in fact experimentally induced differences are greater near noon that at other times of the day, then a preponderance of experiments near noon (because we assumed time of day did not matter and hence did not deliberately select random times of day) could produce interactive effects that can affect our conclusion. Thus it is good idea to test for interactive effects in designs with haphazard rather than deliberately random assignment of treatments to experimental units.

(3) The interactive terms are sometimes absent for a practical reason - not enough data. An example is paired comparisons with only two measurements per unit. The interaction term is assumed to be zero so as to proceed with the analysis. In analyses with 3 or more terms, we often will need to choose which factors to include in a model simply because we do not have enough data to examine all the possible interaction terms. When forced by make choices, our best course of action is to use what we know of the experimental situation. We may know from previous studies that a particular interactive effect is unlikely to substantial. If forced to make a choice, we would include the interactive effect of two fixed factors, rather than the interactive effects of an random and a fixed factor.

Concepts old and new. List of Terms (continued)

Interactive effects (interaction terms)

Number of terms.

(4) Finally interaction terms are dropped according to statistical criteria. An example is dropping a statistically insignificant interaction term from an ANCOVA design, before testing for one main effect controlled for another. However, dropping terms for statistical reasons will generate uncertainty that is hidden from sight. When we drop terms for statistical reasons we increase our Type II error (erroneous acceptance of the null hypothesis)

Test interaction terms first

The presence of an interactive effect means that we cannot interpret the effects of one explanatory variable on the response variable in the absence of information about the other explanatory variable. Thus the logic of hypothesis testing is that we cannot test main effects (those due to a single explanatory variable) if the interaction term is significant. This logic extends to 3 way and higher interactions. A three way interaction (A*B*C) tells us whether the two way interactions (A*B) differ across levels of the third factor C. Thus we cannot test two-way interactions if the 3 way interaction is significant.

Action if cannot reject null interaction. A small interactive effect means that we can interpret the effects of one explanatory variable in the absence of information about the other explanatory variable. Thus the logic of hypothesis testing is that we move upward in the ANOVA table, to test the main effects. The same logic applies to 3-way and higher interaction terms. If the 3-way term is not significant we move to the 2-way terms.

Interactive effects (interaction term)

Action if reject the null. If an interaction term cannot be rejected then we cannot proceed to hypothesis testing of main effects in the ANOVA table. Nor can we proceed upward from 3-way to 2-way interaction terms. We must cease using the table. We have two options. We can simply report the model, with coefficients for each term in the model. This is relatively straightforward for multiple regression. The result is an equation that allows us to compute expected values of the response variable, including interactive effects. The equation describes a response surface whose curvature is described by the interaction term. Reporting the model will be cumbersome if the model contains factors, hence with at least two means for each factor, at least two additional means or slopes for each interaction term. Our second option, which is preferred when we have factors in the model, is to break the model down into levels of one of the factors. If we have a significant 2-way interaction term (A*B) we break the analysis down to a 1-way analysis (A) for each level of B. Such an analysis might consist of tests of slopes (variable A) at each level of B. Or it might consist of tests of differences among means (factor A) at each level of B. Either way, we obtain a separate ANOVA table for each level of B. The differences we observe among these tables are statistically significant (because the 2-way interaction term was significant). Similarly, we break down a 3-way interaction term (A*B*C) into 2-way analysis (A*B) at each level of C. We obtain a separate analysis for each level of C. The differences we observe among these tables are statistically significant (because the 3-way interaction term was significant). This logical procedure will provide insight into the sources of variation in our response variable where we have several explanatory variables. It enables us to pick our those slopes or means that differ, depending on the other explanatory variables.

The interactive effects of two explanatory variables on the response variable are handled differently, depending on the design.

Design	Interaction term
two-way ANOVA	Test before interpreting main effects
randomized blocks	Assumed zero because fixed*random = random
paired comparisons	Assumed zero because fixed*random = random
hierarchical ANOVA	Absent by definition
ANCOVA- heterogeneity of slopes	Focus of the analysis
ANCOVA- statistical control	Test before interpreting main effects
multiple regression	Can be evaluated but often ignored.

Output from GLM routines.

1. Most routines provide residuals and fitted values as an output option.
2. Most GLM routines provide the parameters for the GLM
   These consist of estimate of slopes and means, the latter expressed as deviations from the grand mean β_o
3. Parameters can be estimated outside a GLM routine with functions that estimate slopes and means.

```
 At this point entertain questions brought in for answering.

 Save last 5 minutes to run through the list of "be able
to..."

 Note that exams from previous years are available on
reserve.
```

GLM Exam

List of "be able to" at end of lecture.

The purpose of this course is not to learn lots of material, or memorize lots of statistical tests or terms.

The purpose of this course is to develop your skills in analyzing data, including statistical analysis. The GLM exam reflects this, as you can see from the review questions (on the web and on reserve in the library).

General Linear Model--Review from 4 Oct 1997 (Lec20), 6 Nov 1996 (Lec22)
revised 17 Oct 1998, 29 Oct 2000, 13 Nov 2002 (Lec 19)

```
Review material from 1995 (Lecture 22) did not connect the GLM to the
more familiar "named" techniques.
β notation (standing for both means and slopes) made it hard to
connect. Review session (50 min) in 1996 (Lecture 19) worked thru 9
named tests, showing each as a GLM
[t-test, 1-way, 2-way, paired comp, rand. blocks, hierarchical(nested),
regression, multiple regression, ANCOVA]

This went well, brought the concepts together.
In 1997 this overview moved earlier, before covering applications such as
statistical control, experimental design, model revision.

1997: full 50 minutes: named tests
    (t-test --->  ANCOVA) +key +summary
1998: same as 1997.
    Summary material relies on acquaintance with several cases
2000: Table 11.1 (old Table 9) introduced earlier,
    before hierarchical ANOVA (Nadine Simmonds),
    circa 10 minutes (went ok)
2001: Table 11.1 (old 9) reviewed after multiple regression (no GLM exam)
2002: Table 11.1 (old 9), with short review just before GLM exam.
2003: Introduce Table 11.1 after regression and ANOVA (single explanatory)
2012: In class worksheets begun, continued in subsequent years.
2019: Ch 11 to Ch 15, review before Exam 2. Table 11.1 to Table 15.1
```

‘Named test’ versus the GLM approach to statistics.
The ‘named test’ approach is standard in undergraduate courses. Unfortunately, it is rather like the reliance on roman numerals in Europe in the Middle Ages. Learning GLM after the standard “named test” approach of introductory statistics course is analogous to switching to arabic numerals after learning Roman numerals. The new method (arabic numerals, GLM) is unfamiliar. But with practice becomes familiar, with far greater capacity. Image trying to learn arabic numerals after 20 years using onlyl Roman numerals.

Learning the GLM, like learning arabic numerals, leads to greater capacity.
Example of greater capacity with arabic numerals:
Adding V + IX is hard enough, try dividing VIII by IV.
Example of greater capacity with model based statistics:
Analysis of data for which there is no name, such as one factor and two covariates.

Review technique. Open any text in statistics, find one of the problems at the end of a chapter, set up the GLM analysis for that problem.
Response variable (usually the hardest step to accomplish).
Explanatory variable(s)
Sketch relation of response variable to each explanatory (lines or set of means)
Write the GLM (add graph above each term).
Complete the Source and df columns in the ANOVA table
State name of test (after completing the set-up)

```
This works well with individuals, especially if they undertake
the narrative of what they are doing and why at each step.
It works less well in class, as there is no opportunity for
individual narrative, to identify those areas of understanding
that are sketchy.
```

Relation of 'Named tests' to GLM. See Table 15.1.
Work through 9 named tests, showing each as a GLM

1. Write name.
2. Write GLM $Y = \beta_O + \beta_{sp} X_{sp} + \varepsilon$,
   (notation emphasizes similarities)
   (because only subscripts change)
3. Draw picture above each β, showing relation of Y to X
   For each β express H_A as query,
   this links to ANOVA table
4. Write the Source df table from the model.
5. Compute df and fill this in, mentioning that SS partitioned by computer.

Order:
t-test
1-way ANOVA
2-way ANOVA
paired comparisons
rand. blocks
hierarchical(nested)
regresssion
multiple regression
ANCOVA

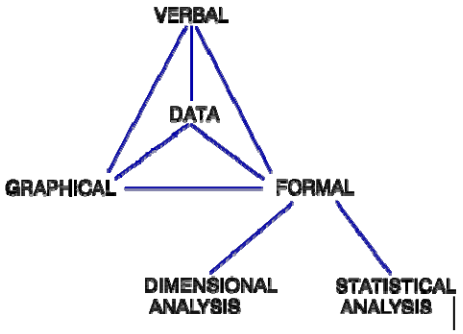
For models with interaction show how the term is handled (assumed zero, logically zero, etc).

Key to "named" tests.
The test that is used in any particular situation depends on the nature of the response variable, it depends on the number of explanatory variables, and it depends on the whether the explanatory variables are nominal or interval scale.

Here is a logical framework for deciding on which analysis to use. It has been set up in the form of a key, much like the keys used to identify an animal to family, genus, and species.

1. Is the response variable Y a count ?
   - If Y is poisson----> G-statistics, log-linear models
   - If Y is binomial----> Generalized linear model with binomial error.
     (includes logistic regression)
2. Is the response variable Y continuous?
   - If so, is the explanatory variable nominal or ratio ?
     - If X nominal ----> ANOVA
       - If one X ----> one way ANOVA
       - If several X----> multiway ANOVA (2-way, 3 way etc)
         - Nested design-----> no interaction term
         - Crossed design--> interaction term (if enough df)
     - If interval------> Regression
       - If one X-----> simple regression
       - If several X-----> multiple regression
     - If nominal X and interval X------> ANCOVA

Diagrammatic Summary of the GLM (formerly Lab 9)



"is a function of"

2 98
25 361
45 7600

$Y=ß_o+ß_1X+,$

Exploratory | Confirmatory

Y = f(?) | Y = f(x)

H_0/H_A: Nominal, ordinal, interval ratio

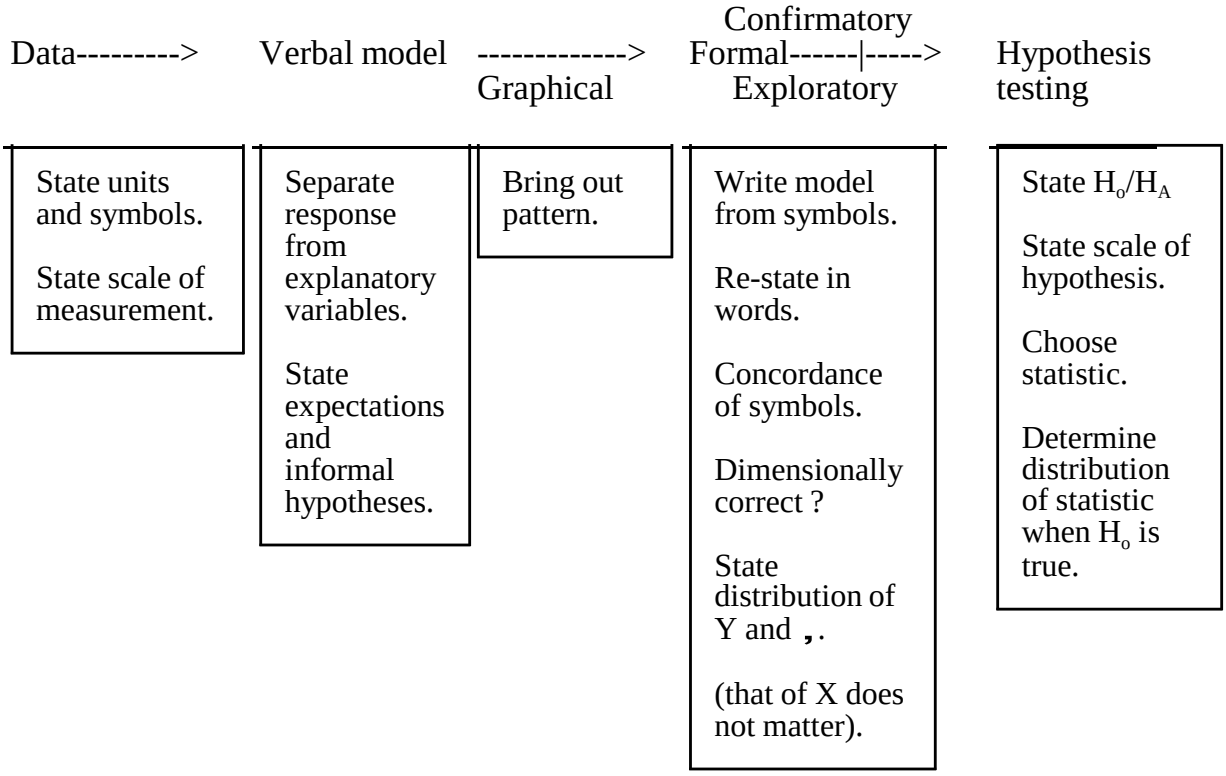
These diagrams combine, as mnemonic elements, the concepts presented thus far in the course. At the upper left is the familiar triangle, relating data to three forms of summarization. The first is verbal: one quantity is said to be a function of another quantity. For example, oxygen consumption is a function of cell size. The second form of summarization is graphical. For example, a series of measurements of oxygen consumption are plotted relative to cell size. A line drawn through the resulting cloud of points will represent the relation graphically. Finally, the relation can be expressed formally, as an equation. In the example at hand, oxygen consumption is the response variable, expressed as a function of cell size, the explanatory variable.

Once data have been summarized in a formal manner, this relation can be analyzed further. Dimensional analysis uses reasoning about quantities based on the principle of similarity. This principle is used extensively in physiology, and to an increasing degree in ecology and environmental biology.

The second form, statistical analysis, is far more prevalent in biology. It is used in either a confirmatory or in an exploratory fashion. In an exploratory analysis, the emphasis is on identifying pattern. The goal is discovery of functional relationships. A convenient mnemonic symbol is Y = f(?), as in the diagram. In a confirmatory analysis the emphasis is on determining whether a relation (of known form) is present or not. In confirmatory analysis, the emphasis is on hypothesis testing, using a null/alternative pair. The H_o/H_A pair can be on a nominal, ordinal, or interval scale.

Diagrammatic Summary (continued, formerly Lab 9)

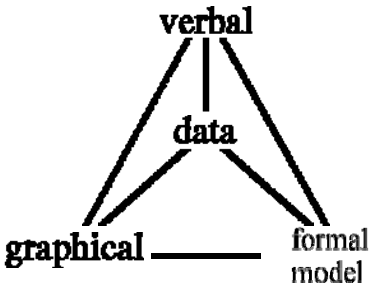
The most common analytic route in setting up an analysis is from data to a verbal model, then to a graphical model, and then to a formal model. This then becomes the basis for statistical analysis in either a confirmatory or exploratory fashion. The following diagram shows this route. The diagram shows one of many possible routes through the diagram on the previous page.

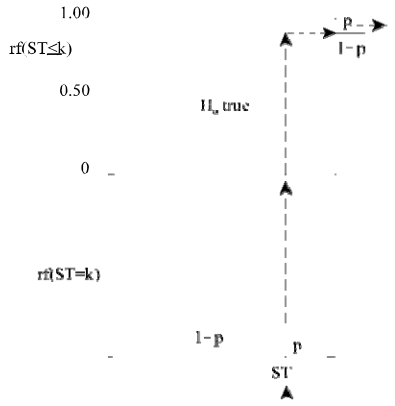


Special cases of the GLM

Many of the statistical tests used in biology are special cases of the general linear model. Included in this family of models are the most commonly encountered statistical tests in biology: regressions, ANOVAs, t-tests, ANCOVAs, etc. The general linear model is, in turn, a special case of the generalized linear model. Special cases of the generalized linear model include G-statistics, logistic regression, probit analysis, and of course the general linear model. The generalized linear model, like the general linear model, relates a response variable to one or more explanatory variables.

To illustrate flow of computations, combine several figures.
Triangle ---> formal model
Formal model ---> ANOVA table.
Detour from ANOVA table to cdf (p-value), back to table.





$[Y] = \sum X \cdot \beta + \epsilon$

Source	df	SS	MS	F	p
X1					
X2					
res					
total					

Once the model has been written, and the H_A/H_o pair established, the computations rely upon comparison of the variance due to the model with the error variance V(**,**).

The total variance is partitioned according to the model that was written.

This partitioning of the variance (together with the degrees of freedom) is used to compute Means Squares and ratios of mean squares, or F-ratios.

This partitioning is tabled.

p-values are then computed for F-ratios. These are computed from a theoretical F-distribution, if the residuals are normal. If the residuals are not normal, then p-values can be computed from an observed distribution generated by randomizing the data.

GLM exam "be able to"

A Purpose of analysis.

- determine whether one variable is related to another
- evaluate relation of a single variable to several explanatory variables.
- statistical control with a covariate
- statistical control with random factors.

B Why use the GLM ?

The general linear model is an effective way to solve problems in biology

Is plant growth related to fertilizer composition?

Is related variable related to another variable ?

C Mechanics (roughly in order of generic recipe)

-Should be able to assign symbols and units to variable quantities

-Able to separate response from explanatory variable

-Able to identify each explanatory variable as categorical (ANOVA type) or continuous (regression)

-Should be able to write a general linear model for each of these cases covered so far.

-Should be able to write a general linear model for data situations where these cases are appropriate

-Should be able to form H_A/H_o pair about response variables in relation to explanatory variables

Model I : means equal, slope = zero.

Model II : variance in Y due to variable X

-Should be able to set up ANOVA table from GLM

-Should be able to compute degrees of freedom from GLM plus structure of data (number of total df, number of groups, whether group (ANOVA) or slope (regression) variable

-Should be able to compute degrees of freedom for the named tests (special cases) covered thus far

C Mechanics (continued)

-Should know how numbers in ANOVA Table are related, to the point where one number can be computed from another if missing.
relation of df to one another
relation of SS to one another
relation of MS to SS and df
relation of F to MS

-Which MS to use in forming F-ratios ?
in general test MS due to a factor over MS residual
except Hierarchical, where MS is tested relative
to next lowest level

-Should know order in which F-ratios are tested
1 Begin with highest order interaction
if significant, stop. Cannot proceed
cannot declare decision about components
then move to lower order interactions
if significant, stop.
continue, until main effects are reached.

2 In Hierarchical, start with highest level

3 In cases of statistical control, no need to
test control variable e.g., randomized blocks

-Should be able to evaluate residuals, based on experience thus far.

-Should know what to do if residuals cannot be defended
-either due to being very bad
-or due to importance of Type I error
as determined by $, time, lives, etc

-Should know how to declare a statistical decision relative to one of the purposes above,
-from information in ANOVA table
-from ANOVA table + cdf

-Should understand cdf command in spreadsheet or statistical package.