Model Based Statistics in Biology.

Part III. The General Linear Model.

Chapter 8 Statistical Inference with the General Linear Model

Elementary statistics courses for biologists tend to lead to the use of a stereotyped set of tests:
1 without critical attention to the underlying model involved;
2 without due regard to the precise distribution of sampling errors;
3 with little concern for the scale of measurement;
4 careless of dimensional homogeneity;
5 without considering the ideal transformation;
6 without any attempt at model simplification;
7 with too much emphasis on hypothesis testing and too little emphasis on parameter estimation.
M.J. Crawley. 1993. GLIM for Ecologists. (London, Blackwell)

ReCap. Part I (Chapters 1,2,3,4)
ReCap Part II (Ch 5, 6, 7)
ReCap Part III
8.1 Introduction
8.2 Component concepts
8.3 Generic Recipe

on chalk board

Experimentation with order of presentation.
1994 Components concepts Lec13
ANOVA example Lec14
1995 Component concepts Lec13
regression example Lec14.
1996 Component concepts Lec13
(in 20 minutes) then
regression example Lec14
1997 Mon: Concepts L13 + ex L14
Wed: revisit L13 + ex L15
Went well.
1998 Same as 1997. Lec 13 in
15 minutes. Went well
2000 General material Lec 13 in
5 minutes.
Components of GLM in 15 min
Then to Lec 14. Went well
2002 Lec 13 General Intro and components
in 10 minutes, then Lec14
2018 Add likelihood

ReCap Part I (Chapters 1,2,3,4)
Quantitative reasoning: Example of fly heterozygosity, which combined models (what is the relation of fly heterozygosity to altitude?) with statistics (how certain can we be?)

ReCap Part II (Chapters 5,6,7)
Data equations summarize pattern in data as a series of parameters (means, slopes).
Frequency distributions, a key concept in statistics, are used to quantify uncertainty.
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.

We have now concluded the first third of course.
We move on to the second third, the General Linear Model

Today: Introduction to the General Linear Model
Begin with brief introduction to component concepts in a generic recipe.
Then work through an example, using the generic recipe.

Wrap-up
The general linear model has many advantages over learning a series of tests.
It lends itself naturally to a problem solving approach based on biological concepts.
We have already covered most of the component concepts.We will use a generic recipe that once learned, permits us to undertake a wide variety of analyses.

8.1 Introduction

Statistics are routinely presented as a collection of recipes in courses for scientists. The basic ingredients are null and alternative hypotheses, a statistic (F, t, or chisquare), a p-value, and the declaration of a decision. The recipes focus on the inverted logic of the null hypothesis rather than on the biological relevance of the model, focus on p-values rather than the interpretation of parameters or the degree of uncertainty associated with each parameter estimate. The recipes often ignore diagnosis of assumptions or evaluation of the sample relative to the population. The recipe collection is huge. Widely used texts typically cover the following tests: one-sample hypotheses, two sample hypotheses, paired sample hypotheses, one-way ANOVA, multiple comparisons, two-way ANOVA, hierarchical ANOVA, multiway ANOVA, regression, multiple regression, analysis of covariance (ANCOVA), polynomial regression, logistic regression, goodness of fit tests, and contingency tests. The menus of widely used statistical packages (Minitab, SPSS, SAS, Systat) contain even longer lists of tests. Choosing from such a long list is daunting, and as it turns out, unnecessary.

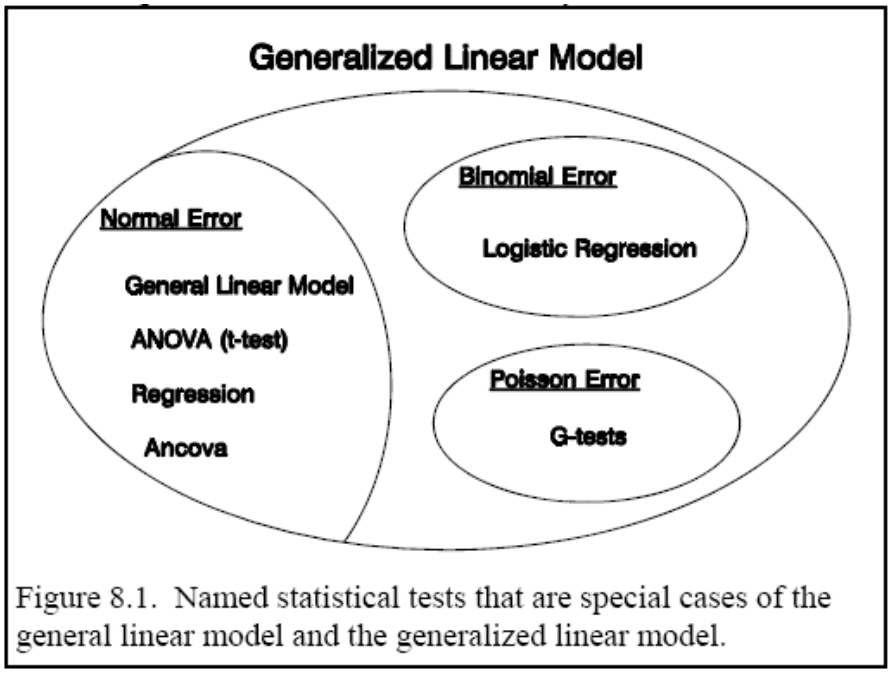


Figure 8.1. Named statistical tests that are special cases of the general linear model and the generalized linear model.

For problems in biology, statistical analysis will usually entail some form of functional relation: How does some quantity *Q* vary as a function of another set of quantities *X1, X2... etc* ? For these problems we can employ ***model-based statistics***, which focus on a response variable in relation to one or more explanatory variables. We will use the generalized linear model (Nelder and Wedderburn 1972, McCullagh and Nelder 1989), one of the major developments in statistics in the last quarter of the 20th century. It allows analysis based on any of several error distributions. We'll begin with the general linear model, which assumes a normal error (Figure 8.1). The general linear model (GLM) has been available in the SAS software package since at least 1980, and is now available in any reputable stat package. The generalized linear model (GzLM) has been available in SAS since the first decade of this century, and is now widely available in code based (SAS, R) as well as menu based (SPSS) software. This development of software allows the generalized linear model to be presented in introductory courses in statistics at the undergraduate level.

Advantages and disadvantages of learning model based statistics.
The GLM is a way of thinking in quantitative terms, using a simple model structure that relate one quantity to another.

The GLM has many advantages

Learn unifying concepts, rather than lots of special cases.
Can see relation of one test to another
No need to learn special procedures for each type of test
For example, no need to learn ANCOVA as two separate types of analysis:
-control for another variable
-comparison of two regression lines

More useful. Can accomplish far more with general approach than by using a series of boxes. Many data sets do not conform to the assumption for standard regressions and ANOVAs; it is more effective to learn a general procedure to handle these problems, rather than learning one set of remedies for problems with regression, another for problems with ANOVAs, etc. GLM is far more flexible in dealing with problems.

Easier to learn in the long-run. Though perhaps harder of first, because of the generality of the approach. A generic recipe is harder to learn than a specific procedure test. But a generic recipe is less work to learn than a whole collection of specific procedures.

Disadvantages
More abstract, harder to learn, at the outset.
Software in the past was difficult to use with steep learning curve. The GLIM package, for example could be described as "user-hostile." This is changing, as GLM and GzLM become available in packages that are intuitive and don't demand a steep learning curve.

Summary of advantages. First, students learn unifying concepts rather than a sequence of apparently unrelated procedures. Students can see the relation of one test to another, rather than having to learn special procedures for each topic. For example, ANCOVA can be presented as two applications of the same model, rather than as two separate procedures, one for comparing slopes and one for statistical control of a regression variable. Remedies for recurring problems (*e.g.*, heterogeneous variances) are presented once, rather than several times in different guises. The approach means that remedies can be learned, instead of memorizing specific remedies for each test. The mechanics of analysis are presented once, rather than different procedures for each test. Students are able to accomplish more with the general linear model than by learning statistics as a set of named procedures. For example, with this approach students can set up and execute the analysis of a response variable in relation to two categorical and a single regression variable. There is no name for this analysis, and hence it is outside any list of tests. This greater flexibility leads to better quantitative work in biology. The GLM is a way of thinking in quantitative terms, using formal models that relate one quantity to another.

The material in the next two weeks will use the same generic recipe, applying it to special cases such as regression, ANOVA, t-tests.
The GLM will become familiar through repetition.

From over 20 years experience (B4605/B7220 at MUN) the model-based approach is readily grasped and executed by 3rd and 4th year undergraduate biology majors. The presentation here begins by assembling the components learned so far – quantities, data equations, computing the fit of the model to the data, computing the improvement in fit due to an explanatory variable, either categorical or regression. After a discussion of assumptions for computing p-values in an ANOVA table (which tracks the improvement in fit due to explanatory variables), it moves to a generic recipe that will be applied first to regression (Chapter 9), the ANOVA (Chapters 10, 11) and ANCOVA (Chapter 14).

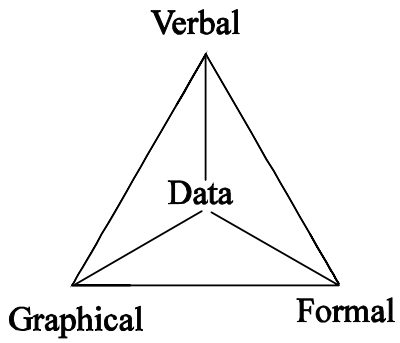
> Another look at 8.1
>
> Biologists agree that the list of current bird species is finite and rapidly approaching completion. Do you think that a list of statistical tests is finite or could ever be complete ? Why or why not ?

8.2 Component Concepts

The General Linear Model is a sophisticated concept that substantially improves the quality of statistical analysis by non-statisticians. We will be using component concepts that have already been covered in this course.

Model based Statistics

Data at the centre, three forms of summarization at the apices of the triangle. The GLM summarizes data as a formal model. To arrive at this model we begin with verbal model, often in the form of a question. We can use graphical display to express our model and as aid in constructing the formal model. We use the formal model (GLM) to undertake the statistical analysis, which quantifies uncertainty.



8.2 Component Concepts

Quantity

A well-defined quantity has 5 parts:

procedural statement, name, symbol, values, and units.

The GLM relates one variable quantity (response variable) to one or more other quantities (explanatory variables).

Example:	Plant growth	is	a function of nutrients and sunlight
	Growth Rate	=	f (nutrients, sunlight)
	G	=	f (N, PAR)

Each symbol stands for a variable, which can take on several values via direct measurement, via calculation from direct measurements, or (in the case of an explanatory variable) via categories values fixed by experimental design (e.g. control, treatment).

Variance of a quantity

We have already met the variance of a quantity. It is the mean squared deviation from the average value of the quantity. The true value of the variance of a quantity is often unknown, so an estimate is made. The estimate is:

$$\mathrm{Var}(Q) = (n-1)^{-1} \Sigma(Q - \mathrm{mean}(Q))^2$$

In this formula, you will recognize the sum of the squared deviations

$$SS = \Sigma(Q - \mathrm{mean}(Q))^2$$

This is the fit to the simplest of all models, mean(Q)

Model components and data equations

The general linear model has three components: a ***response variable*** Y, a ***structural model*** consisting of one or more explanatory variables X1, X2, etc., and an ***error term***. Table 8.1 shows equivalent expressions. Each term in the model (response variable, explanatory variables, error) represents a vertical string (vector) of numbers. Consequently the symbolic expressions in Table 8.1 represents a series of ***data equations*** (see Chapter 5).

Table 8.1 Equivalent expressions of the general linear model.

Data	=	Model	+	residual
Observed	=	Expected	+	residual
Response	=	f(explanatory variables)	+	residual
Y	=	f(X1 + X2 +...)	+	error
Y	=	$\Sigma\ \beta i$ Xi	+	ε

The explanatory variables can be on a nominal type of scale (ANOVA), on a ratio type of scale (regression), or both (ANCOVA). The residuals are distributed normally.

Data Equations – Simplest model.

A data equation is written for each value of the response variable. The model in Table 8.2 is the simplest possible: the variable of interest M is equal to fixed value β_o the average value for the population. Of course, we almost always do not know the true value of β_o. Our best estimate of β_o is the mean $\hat{\beta}_o$ = 59 g, computed from the data we have.

Table 8.2 Data equations for measurement of the mass of 3 juvenile cod *Gadus morhua*.

Data	=	Fitted values	+	Residual
M	=	mean(M)	+	ε
55 g	=	59 g	+	–4 g
60 g	=	59 g	+	+1 g
62 g	=	59 g	+	+3 g
		H_o		

Data Equations –Comparison of models. Null and Alternative model

As scientists we are often interested in comparing models. For example, does juvenile cod mass differ in vegetated and unvegetated habitats? Chapter 5 showed another example - does genetic variability decrease with altitude? We might expect that harsh environmental conditions (such as we encounter as we climb a mountain) would reduce variability. This leads to a research hypothesis, that heterozygosity decreases with altitude. The research hypothesis, in statistical jargon, is called the "alternative" model H_A. It is compared to a "null" model which summarizes the current state of science knowledge. In the analysis of the fly heterozygosity data (Chapter 5) the null model (current state of knowledge) was no change in heterozygosity with change in altititude. The alternative model was that heterozygosity H decrease as a function of elevation E.

H	=	E(H)	+	ε	population
H	=	$\hat{H}$	+	*residual*	*sample*
H	=	*Intercept* + $\hat{\beta}_E \cdot E$	+	*residual*	
H	=	*0.63* *–0.1298* $\cdot$ E	+	*residual*	

Heterozygosity is the response variable (on the left), elevation is the explanatory variable (on the right), and there is a residual calculated for each observation of heterozygosity. This model represents a series of data equations. Each data equation consists of a single observation (on the left), an expected value (parameters and explanatory variables on the right), and the residual or left over part, to make the equation balance out.

Data Equations – Comparison of models

In the analysis of the oat yield data we used a model to describe oat yield Y as a function of group X (treated versus untreated).

$$Y = \beta_0 + \beta_x \cdot X$$

In this example the explanatory variable is on nominal scale. It consists of categories.

Equivalent notation

Measurements	=	Model	+	residual
Q	=	f(X)	+	residual

f(X) means "function of X"

Q	=	E(Q)	+	residual

E(Q) is estimate of the true value f(X)

Q	=	$b_o + b_X X$	+	residual

b_o b_1 are estimates of parameters

Q	=	$\hat{\beta}_o + \hat{\beta}_X X$	+	residual

hats over parameters also signify estimates

Estimates of parameters

We have already encountered the concept of estimation.

It can refer to informal estimates, such as an order magnitude estimate of mass of an elephant 10^3 kg ? 10^4 kg?

It can refer to an estimate from a formula, such as calculating a mean, a variance, or a standard deviation.

It can refer to an iterative procedure, such as a least squares estimate or the maximum likelihood estimate we saw in Lab 3.

In statistical analysis with the GLM, we will be using the formal machinery of estimation (least squares) to calculate the "best" estimates of means and slopes, according to a least squares criterion. For non-normal error structures (GzLM) we will be using iteratively reweighted least squares or maximum likelihood estimates. We will let the statistical package do the work for us.

The most commonly estimated parameters are means, slopes, and proportions and odds ratios.

Evaluation of residuals. We have seen that to use a statistical distribution (*t, F*, chisquare, normal) to calculate Type I error (the p-value) we need to make assumptions. We will rely on graphical displays to evaluate the assumptions (Chatfield 1998; Gelman, Pasarica & Dodhia 2002). The statistical literature warns against statistical tests to evaluate assumptions and advocates graphical tools (Montgomery & Peck 1992; Draper & Smith 1998, Quinn & Keough 2002). La¨a¨ ra¨ (2009) gives several reasons for not applying preliminary tests for normality, including: most statistical techniques based on normal errors are robust against violation; for larger data sets the central limit theory implies approximate normality; for small samples the power of the tests is low; and for larger data sets the tests are sensitive to small deviations (contradicting the central limit theory). In particular we will not adopt the mistaken practice of checking the response variable for normality. Instead we will obtain residuals to evaluate assumptions. Refs in Zuur et al 2010.

Units, Dimensions, and Model Interpretation

Units and dimensions are typically not considered in the statistical analysis of data. They should be. The parameters (means and slopes) that result from statistical analyses are usually parametric quantities, with units and dimensions that depend on the units and dimensions of the measured variables being analyzed. They are not simply numbers, which is how they are often reported. A glance at the set of the three data equations for cod weights (Table 8.2) will reveal that the mean has the same units and dimensions as the response variable, which appears on the left side of the equality sign. In a regression equation ($Y = \alpha + \beta_x X$ + residual) the intercept α must have the same units and dimensions as the response variable Y. The residual term must also have the same units and dimensions as the response variable Y. The regression coefficient β_x will have the same units and dimensions as the ratio Y/X, in order for the equation to be dimensionally consistent. In the heterozygosity example (Box 8.1), the slope β_x quantifies the altitudinal gradient in genetic variability in units of %/km.

There are several reasons why GLM parameters should be recognized as scaled quantities, rather than treated as simply numbers. First, the rules for operations on scaled quantities, which differ from those for numbers, apply to parameters. Two means can be added only if they have the same units. The rules for rigid and elastic rescaling apply to parameters, a fact that is not evident if parameters are treated as mere numbers. Erroneous calculations result if a parameter is treated as a number. A regression coefficient that is an estimate of a spatial gradient at a scale of 100 m cannot be used to calculate a gradient at another scale, unless that coefficient is rescaled according to its units and dimensions.

From likelihood ratios to hypothesis testing

Once we have a null and alternative model consistent with the data we can compare them as a likelihood ratio. In chapter 5.4 we compared the likelihood of the gradient to the no-gradient model. The measure of deviation was $SS_{tot} = 0.117$ for the null model. The fit improved for the alternative model. The deviation measure dropped to $SS_{res} = 0.0214$. The observed improvement in fit was $SS_{model} = 0.1174 - 0.0214 = 0.096$. The likelihood ratio was $(0.0214/1.1174)^{-7/2} = 453$. The alternative model (gradient of -0.1273/km) was 450 times more likely than then zero gradient model. Statistical practice in some areas science is moving toward reports such as this—a likelihood ratio with an effect size, such as the gradient in this case. Likelihood inference (Edwards 1972, Royall 1977) is common in genetics and some areas of ecology (Burnham and Anderson 1998).

From likelihood ratios to hypothesis testing (continued)

Hypotheses tests use the likelihood ratio to calculate a p-value that is routinely used in the Neyman-Pearson sense (decision for are against the null) and rarely if ever in the Fisher sense of a flexible guide for discarding the null. So when should we use hypothesis testing? There are two reasons. The first is when we need to consider Type II as well as Type I error, and thus, the balance between the two in designing an experiment. The second reason is prevailing practice ***IF*** we can justify the use of hypothesis testing. The justification is whether we can define a population to which we are inferring (Fisher) or whether we can define chance from a repeatable measurement procedure (Hacking). In experimental work, with a well defined protocol repeatable by others we have a justification. In observational studies where the number of uncontrolled variables is large we may well choose to likelihood inference rather than try to defend our measurement protocol as repeatable (Hacking) or as a sample from a population of infinite repeats of the study. In this course we will make that choice early in our analytic procedure.

Hypothesis testing

In this course we will be using a generic recipe for GLM based statistical analysis. This recipe incorporates the generic recipe for hypothesis testing. The test statistic will be the F-statistic, the ratio of two variances. These variances will be obtained by partitioning the response variable *Q* into two components, that due to the model, and that remaining (the residual)

Data	=	Model	+	Residual
Var(data)	=	Var(Model)	+	Var(Residual)

F = var(model) / var(residual)

We will use the F distribution to calculate the long run probability of any value of the F-ratio, given the degrees of freedom in the model and the degrees of freedom remaining in the residual.

Estimation and Confidence Limits

Hypothesis testing is the prevalent mode of statistical analysis in biology.
Estimation and confidence limits are more informative.

Restore attention to structure of model, including units and dimensions.
Allow exclusion of multiple hypotheses, not just the null hypothesis.

For example, in examining the relation of metabolic rate to body size, the null hypothesis is biologically irrelevant. We are more interested in excluding a 1:1 scaling than we are statistical rejection of the hypothesis of no relation.

8.3 A Generic Recipe for Applying the General Linear Model

The general linear model is not part of the traditional undergraduate curriculum for biology students. However, it is readily grasped by third and fourth year undergraduates in biology when presented as a procedure for analysis, rather than a set of formulas to memorize. Students with limited backgrounds in mathematics and statistics can successfully apply the following generic recipe (Table 8.3) to novel data sets and to their own data.

Table 8.3 Generic Recipe for Statistical Inference with the General Linear Model.

1. Construct model. Begin with verbal and graphical model.
   Distinguish response from explanatory variables
   Assign symbols, state units and type of measurement scale for each.
   Write out statistical model.
2. Execute model Place data in model format, code model statement.
   Compute fitted values from parameter estimates.
   Compute residuals and plot against fitted values.
3. Evaluate the model, using residuals.
   If straight line inappropriate, revise the model (back to step 1).
   If errors not homogeneous, consider using generalized linear model (step 1)
   If n small, evaluate assumptions for using chisquare, t, or F distribution.
   residuals homogeneous ? (residual versus fit plot)
   residuals independent ? (plot residuals versus residuals at lag 1)
   residuals normal ? (histogram of residuals, quantile or normal score plot)
   If not met, empirical distribution (by randomization) may be necessary
4. Partition df and SS according to model. Write SS and df for each term in model.
   State the full (null) and reduced (alternative) model
   Calculate likelihood ratio for omnibus model.
   If sufficient evidence for omnibus model Step 5, otherwise step 10.
5. Define target of inference. Choose mode of inference: evidentialist, frequentist, priorist.
   If priorist, see recipe. If evidentialist, step 9.
6. State test statistic, and sampling distribution (t, F. χ^2, or Monte Carlo).
   Fixed Type I error or Fisher sorting?
7. ANOVA: Table Source, SS, and df. Calculate MS, F-ratio.
   Obtain Type I error (p-value) from distribution (F or t).
8. Recompute Type I error if necessary.
   If assumptions not met compute Type I error by randomization if:
   sample small ($n < 30$) and if Type I error near fixed α.
9. Report statistical conclusion about fixed terms and factor contrasts in the model.
   For frequentist inference report either the ANOVA table, or F-ratio (df1,df2),
   or t-statistics (df=1) and Type I error (not α) for fixed terms and factor contrasts.
10. Report science conclusions. Interpret parameters of biological interest (means, slopes)
    along with one measure of uncertainty (st. error, st. dev., or conf. intervals).
    Use t or Monte Carlo distribution to compute confidence limits as needed.

The next chapters work through the generic recipe step by step for commonly used analyses in biology

Exercises

1. List of key concepts for review and future reference.

____model-based statistics	____general linear model
____response variable	____structural model
____data equations	____expected value
____true (population) value	____estimate
____null model	____alternative model
____goodness of fit	____analysis of variance
____degrees of freedom	____p-value of a variance ratio
____hypothesis testing	____assumptions for p-values
____randomized p-value	____generalized linear model
____ANCOVA	____link functions

2.