

Model Based Statistics in Biology.

Part III. The General Linear Model.

Chapter 11 Review of GLM, Single Explanatory Variable

ReCap.	Part I (Chapters 1,2,3,4)
ReCap	Part II (Ch 5, 6, 7)
ReCap	Part III (Ch 9, 10)
11.1	Model Based versus Voodoo Statistics
11.2	List of Terms
11.3	Commentary on Generic Recipe
11.4	Review Questions

Table 11.1 at end

on chalk board

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.

ReCap (Ch 9, 10) The General Linear Model is more useful and flexible than a collection of special cases.

Regression is a special case of the GLM. We saw examples where the explanatory variable X was fixed and where the explanatory was measured with error.

ANOVA is another special case of the general linear model.

The relation of the response to explanatory variable is expressed as set of means.

Factor consists of fixed effects or random effects. For the fixed effects, interest is in the source of the differences. For random effects, interest is in whether there is variance among groups, above and beyond variance within groups.

Today: Review of the GLM, Single Explanatory Variable.

Wrap-up.

The model based approach in this course fosters sound diagnosis (e.g. residual versus fit plots). It points to improved therapeutics (e.g., model revision). It avoids ill-founded 'cures' including that do more harm than good.

Commonly used analyses in biology are special cases of the GLM (See Table 11.1).

11.1 Model Based versus Voodoo Statistics

This course emphasizes model based statistics. Table 11.1 lists the commonly used statistical methods that are special cases of the general linear model. In Chapter 9 we covered the GLM for a single explanatory variable on a ratio scale (regression). In Chapter 10 we covered the GLM for a single explanatory variable on a nominal scale, *i.e.* a factor (categorical variables) with multiple classes. In Chapters 12 and 13 we will extend the GLM to two or more explanatory variables

Table 11.1 matches general linear models with the traditional names for special cases of the GLM in the literature (regression, ANOVA, t-tests, ANCOVA, etc).

Model based statistics free us from voodoo statistics, defined as following a ritualized activity in the hope of inducing sound analysis. Voodoo statistics are all too common.

Here are some examples of voodoo statistics.

‘Check assumptions before undertaking statistical analysis’	X
‘Use hypothesis testing to check assumptions.	X
‘ANOVA or regression cannot be used to analyze indices of species diversity’	X
‘Compute the power of a test if the null hypothesis is accepted’	X
‘Use a non-parametric test if your data do not meet the assumptions for ANOVA or regression’	BDC
‘Use an arcsin transform with percentages’	BDC
‘Use a square root transform for count data’	BDC
‘A two way ANOVA requires equal sample size in each cell’	BDC

Some of these prescriptions are completely wrong (marked X).

They are the equivalent of a physician prescribing substances that have been shown to be harmful rather than curative.

Some of these prescriptions (marked BDC, before digital computers) were sound advice in the days before digital computers, which vastly improved data medical diagnostics and therapeutics in the latter decades of the 20th century. BDC prescriptions are the equivalent of a physician using 19th century rules of thumb instead of science based diagnostics and therapeutics developed in the later half of the 20th century.

Voodoo statistics violate a principle that should guide data analysis as much as the practice of medicine: *primum non nocere* - above all do no harm (to the patient or the data). Voodoo statistics do harm by reducing the amount of information (reducing data to ranks, making sample sizes equal by throwing away data), by restricting the interpretive scope (eliminating interaction terms when adopting rank based tests), by creating uninterpretable models (as with arcsin transforms), or by discarding interpretable parameters (means, slopes, odds) in favour of less informative statistics (medians).

11.1 A Tour of Voodoo Statistics.

Exhibit A. ‘Check assumptions before undertaking statistical analysis’

This is one of the commonest and at the same time most harmful practices. It is erroneous because the assumptions for computing p-values depend on the distribution of the residuals, not the response variable. As we have seen the residuals can meet the assumptions even when 'the data' do not. Checking assumptions is good practice, but only if the correct assumption is checked.

Unfortunately, many people are going to tell you (erroneously!) that "your data must be normal" when in fact it is the residuals that must be normal, in order carry out statistical analyses.

What do you do if someone insists that you check your assumptions before computing residuals? (Don't be surprised if this happens).

The first step is to state the assumptions correctly.

"Type I error calculated from an F or t-distribution of course assumes that the residuals from my model were normal. I examined the residuals, found they were normal, and so I used the p-values calculated by the statistical package."

Some people might still argue. If they do, try citing a text written by statisticians.

"According to Neter et al (1983) Seber (1966) and Sokal and Rohlf (2012) the assumptions for computing p-values from F, t, and Chisquare distributions are that the residuals are normal, homogeneous, independent, and sum to zero."

Neter, J., W. Wasserman, M.H. Kutner (1983). Applied Linear Regression Models. Homewood Illinois, Richard D. Irwin, Inc. (page 31, 32, 49)

Seber, G.A.F. 1966. The Linear Hypothesis: A General Theory. London, Griffin.

Sokal, R.R., F.J. Rohlf. 2012. Biometry. 4th edition. Freeman.

Note that some texts by non-statisticians fail to state assumptions clearly, or fail to state clearly that assumptions are checked by examining the residuals. The biometry text by Sokal and Rohlf states the assumptions clearly.

Exhibit C: Use a non-parametric test if your data do not meet the assumptions for ANOVA or regression.

Non-parametric tests based on reduction of data to ranks were sound practice before 1980, when personal computers made high speed computation widely available. Back then, the benefits (ease of computation by hand, assumptions in computing p-values met regardless of data) outweighed the disadvantages. Now that we have high speed computers, the benefits are gone and the disadvantages remain.

Disadvantages

- loss of information in reducing ratio scale data to ranks.
- concomitant loss of power and increase in Type II error.
- loss of interaction terms, which can be as informative as main effects.
- interpretation restricted to medians, rather than mean values.
- cannot compute confidence limits.

Ease of computation by hand no longer matters. With high speed computers it is no longer necessary to reduce data to ranks to devise a test that guarantees meeting the assumptions for computing a p-value. We can use randomization to compute a p-value that is as valid as a p-value from rank based tests.

Exhibit D: Don't use ANOVA or regression to analyze indices of species diversity. These indices summarize, as a single number, the information obtained by sorting a collection into n species, each with a count of N_i organisms, from which we obtain $p_i = N_i/\sum N_i$ the proportion of organisms in each species. Here are three common indices.

n species richness

$H = \sum_{i=1}^n p_i \ln p_i$ The Shannon-Weaver index

$D = \sum (p_i)^2$ The Simpson index

These measures are bounded at zero and so have the potential for non-symmetrical errors around means and regression lines. In practice, the deviations are not always large, and may not matter anyhow if the number of collections being compared is large. As with Exhibit A and B, sound statistical practice begins with diagnostics based on the residuals. If the residuals in an analysis are heterogeneous or non-normal, we can compute p-values by randomization, or by use a non-normal error structure. There is no reason to avoid ANOVA or regression for these measures.

Exhibit E: Use an arcsin transform with percentages.

Percentages are bounded at 0 and 1, so in principle residuals will not be normal. In practice, residuals tend to cluster around zero, and the transform usually has little effect on the distribution of residuals. Percentages should thus be assumed innocent until proven guilty. Best practice is to examine the residuals.

The arcsin transform has a higher voodoo score than rank based methods because there is no guarantee that the transform will produce homogeneous and normal residuals. The arcsin transform is more noxious than benign because the mean of arcsin transformed numbers is uninterpretable. If the residuals from the analysis percentages do prove to be unacceptable, randomization is a far better way of addressing the problem of assumptions for computing p-values and confidence limits.

See: Warton, David I., and Francis K. C. Hui. 2011. The arcsine is asinine: the analysis of proportions in ecology. *Ecology* 92: 3–10. [doi:10.1890/10-0340.1]

Exhibit G: Use a square root transform for count data.

Count data are bounded at zero and so in principle residuals from the analysis of count data will deviate from normality and homogeneity. In practice count data usually do deviate from homogeneity, especially if zero counts are present. The variance will increase along with the mean of groups or fitted values, either in a 1:1 fashion (as with Poisson counts) or as a multiple of the mean (as with counts of highly aggregated objects). But while we can expect count data to have a heterogeneous rather than fixed variance, we cannot count on the square root transform to flatten the variance to a constant value. Thus best practice is to examine the residuals, rather than invoking the square root transform whenever count data are analyzed. If residuals are heterogeneous, then a square root transform might be tried if the data show the characteristics of Poisson counts: few counts above 10 or so and lots of zero counts. For counts with non-Poisson characteristics (zeros and counts above 10 or so) the square root transform will fail to impose homogeneity on the variance. The log transform is a better bet, although there are still no guarantees. The log transform has the advantage that back transformation of the average of logged data produces the geometric mean, which can be interpreted as a measure of central tendency on a multiplicative scale. Note however, that the log transform creates problems of its own (cf Ch9.4). For count data, best practice is to use an appropriate error structure (binomial, poisson, overdispersed poisson) within the computational framework of the generalized linear model. Some simple examples will be shown in Part 5 of this course.

Exhibit H: A two way ANOVA requires equal sample sizes in each cell.

The General Linear model allows correct estimates of unequal sample sizes in multiway designs.

Exhibit I. Use hypotheses tests to determine if your data are normal.

Tests of normality compound the problem because the sensitivity of these tests will often lead us to conclude that data are not normal when sample sizes are large, when violations are unlikely to affect the computation of p-values or confidence limits. Conversely, tests of normality will be insensitive to violations at small sample sizes, when violations can have a large effect on computation p-values and confidence limits. We have learned a better procedure: check the residuals. So this example of voodoo statistics is not only erroneous, but reliably wastes time by provoking unnecessary remedies.

Exhibit J: Compute the power of a test if the null hypothesis is accepted.

See:

John M. Hoenig and Dennis M. Heisey. 2001. The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis. *The American Statistician* 55(1); 19-24.

11.2 Review of GLM Concepts

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)**

Analysis of Variance: partition SS

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

Assumptions for p-values from cdf (4)

Assumptions not met

Hypothesis testing. H_A / H_0 for parameters,
for variance due to each term in model

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

Analyze parameters (best practice, but not widespread).

11.2 Review of Concepts

One-tail vs Two-tailed tests.

F-tests always one tailed (equivalent to two-tail test on means)

t-tests one or two tailed.

One-tail versus two tail tests. This question arises each year.

For H_A / H_0 pairs concerning terms in the model we are carrying out one-tailed tests using the F-distribution. We are testing whether the variance in the numerator of the F-ratio exceeds the variance in the denominator. We use the upper tail of the F-distribution because the numerator variance in an ANOVA table includes the variance due to the factor in addition to the residual variance used to form the ratio.

H_A / H_0 pairs concerning terms in the model are equivalent to two-tailed tests of the parameters in that term. For example:

$H_A: \text{Var}(\beta_X \cdot X) > 0$ Is equivalent to $H_A: \beta_X \neq 0$

In addition to these two-tailed tests concerning the parameter of a term, we can form our own one-tailed test about a parameter. For example we might expect a positive relation between height of sons and their fathers.

$H_A: \beta_{Hfather} > 0$

To calculate the p-value on a one-tailed test of the parameter, we cut the p-value (from the F-distribution) in half.

Two tailed test for slopes $H_A: \text{Var}(\beta_X \cdot X) > 0$ Equivalent to $H_A: \beta_X \neq 0$


$H_0: \text{Var}(\beta_X \cdot X) = 0$ Equivalent to $H_0: \beta_X = 0$

For this test we use the upper tail of the F-distribution

Hence $p = 1 - 0.9288 = 0.0712$

One-tailed test for slopes $H_A: \beta_x > 0$ one-tailed test

$H_0: \beta_x = 0$

MTB> cdf 3.5;	
SUBC> f 1 30.	
3.500	0.9288

For this test we again use the upper tail of the F-distribution, as above, then cut this in half. Hence $p = (1 - 0.9288)/2 = 0.0712/2 = 0.0356$

The one-tailed test is significant at the conventional 5% criterion.

11.2 Review of Concepts. One-tail vs Two-tailed tests.

There is variation in whether p-values are reported for one or both tails of symmetrical probability density functions (t and normal distributions). Statistical packages usually report just the tail requested (upper or lower), while spreadsheets and printed tables often combine probabilities from both tails.


For example, the minitab cdf command always gives one tail. It always give the probability of obtaining the observed or greater value of the statistic.

For example

$$\text{Hence } p = 1 - 0.9703 = 0.03$$

This is for one tailed test (probability of t greater than 1.96). For two tailed test p-value is $2 \cdot 0.03 = 0.06$ (probability of t greater than 1.96 OR less than -1.96).

Spreadsheets and statistical tables often give the two tailed p-value, not the one tailed p-value for a t-distribution. For example, most spreadsheets report a p-value of 0.06 for $t = 1.96$ with 30 df. If in doubt, check against a known p-value versus t statistic. You can use $t = 1.96$ with 300 degrees of freedom (hence equivalent to a normal distribution). This will give a p-value of either $p = 0.025$ (upper tail) or $p = 0.05$ (both tails).

MTB> cdf 1.96;	
SUBC> t 30.	
1.9600 0.9703	

11.3 Commentary on the Generic Recipe for Hypothesis testing with GLM

1. Construct model

Statistics are a way of summarizing pattern in the form of a formal model. The simplest and most familiar is the computation of a mean--a single value that is taken as representative of a set of observations. The General Linear Model is one of the most used. It includes such familiar procedures as regression and ANOVA. The GLM relates one (or more!) response variables to one or more explanatory variables.

$$Y = \beta_o + \beta_{X1}X_1 + \dots + \text{residuals}$$

$$Y = \sum \beta_i X_i + \text{residuals}$$

The GLM is flexible. It allows the explanatory variables to be on a nominal (categorical) measurement scale, on a ratio type of scale, or on both.

Verbal and *graphical* models are useful in formulating a general linear model for use in analysis of data. A typical sequence is to go from data to a verbal model, then to a graphical model, and finally to a formal model.

1

In setting up an analysis with the GLM it is important to separate the response from explanatory variables. One of the best ways to help someone who is having trouble "analyzing their data" is to ask them to identify their response variables, separating these from explanatory variables.

Both *response* and *explanatory* variables are quantities. They should be defined by a procedural statement, assigned a name and a symbol, with units as well as numerical values.

In learning to use the GLM it is important to write out response and explanatory variables, then state the model in words and picture before trying to write it. With practice it is possible to write the model statement directly.

When the GLM is used in inferential statistics, the model is written for the population. The convention of writing parameters with a greek symbol is used to designate that the model applies to the population. The parameter that pertains to a ratio scale variable is a slope (regression analysis). The parameter that pertains to a categorical variable is a set of expected values, expressed as deviations from grand mean $E(Y) = \beta_o$.

2. Execute model

Place data in model format:

- one column for response variable
- one column for each explanatory variable.
- categorical variables use numbers (or letters) for each level of a factor.

Code the model statement,

Write model statement that follows sequence of terms in step 2.

Obtain fits and residuals

GLM routines use the data to make estimates of the model parameters.

These estimates are distinguished from the true (and unknown) values of the population parameters by placing a hat over the parameter.

β_o stands for the mean of the entire population.

$\hat{\beta}_o$ stands for the mean computed from the sample.

Estimates are made according to two criteria: either minimizing the squared residuals between the model and the data, or maximizing the likelihood of the estimate, given the data. In many cases the latter works out to be the same as the former estimate. The mean, computed according to the familiar formula, is both a maximum likelihood and minimum deviation estimate of the true value. Slopes for linear regression are estimated by minimizing the sum of the squared vertical deviations from the regression line. In simple cases this is accomplished in one step with a formula. In more complex cases, this is accomplished by iterative techniques such as curvilinear regression.

These estimates are used to calculate fitted values and from these the residuals. The residuals are plotted against the fitted values to evaluate the model (next step).

Output from GLM routines.

- Most routines provide residuals and fitted values as an output option.
- Most GLM routines provide the parameters for the GLM

These consist of slopes and means, the latter expressed as deviations from the grand mean β_o

Parameter estimates in general linear model format:

$\hat{\beta}_o$	overall mean
$\hat{\beta}_{group}$	deviations from overall mean
$\hat{\beta}_{regression}$	overall slope
$\hat{\beta}_o + \hat{\beta}_{group}$	means in each group

- Parameters can be estimated outside a GLM routine with functions that estimate slopes and means.

Residuals computed from fitted values are plotted against fitted values.

3. Use residuals vs fits to evaluate the model.

A GLM consists of the response variables, the structural model (consisting of explanatory variables and parameters), and the error.

We first evaluate the structural model.

A bowl or arch in the residual plot indicates that the relation of response to explanatory variable is not a straight line. If this proves to be the case the model needs to be revised so that the relation of response to explanatory variables is correctly represented. The straight line assumption does not need to be checked if the GLM consists entirely of categorical (ANOVA) variables.

Another way of looking at this assumption (extra).

Are the residuals associated with the model ? We want to ensure zero covariance between the model and the residuals.

$$\text{Var}(Y) = \text{Var}(\text{Model}) + \text{Var}(\text{Residuals}) + \text{Cov}(\text{Model}, \text{Residuals})$$

Covariance is detectable as a curved pattern in the plot of residuals against the fitted values, it does not occur as a simple positive or negative association. If there is covariance, then the model is inappropriate and the variance estimates used in hypothesis testing will be in error.

Another model should be used.

Next, we evaluate the error model. For the GLM, the error model is that of homogeneous, normal, and independent errors. If these are violated we have two choices – we can revise the model to an error structure that is more appropriate than normal (fixed) error. Examples are logistic and poisson regression, for which the error increases in step with the magnitude of the fitted values. These are special cases of the generalized linear model, which allows us to specify any of several models. The alternative course of action is to continue with the general linear model but use an empirical distribution of outcomes instead of a chisquare (or t or F) distribution to calculate p-values or confidence limits. There are many good reasons to adopt the first course of action (McCullagh and Nelder 1987, Myers et al 2002). The examples in this course have relied on the second course of action, where an empirical distribution is constructed via randomization.

3. Use residuals vs fits to evaluate the model.

Residuals and fits are used to evaluate distributional assumptions.

In this course we rely in the first instance on chisquare (or t or F) distributions to compute p-values and confidence limits. These distributions rely on four assumptions:

1. $\text{Var}(\text{res}) = \text{constant}$. Plot residuals versus fits, check for cones.
2. $E(\text{res}) = 0$ This will be automatically true for analyses in which parameters are estimated from data, as in most statistical packages, so no need to check.
3. $\text{Cov}(\text{res}_i, \text{res}_j) = 0$ i.e., residuals independent.

This is checked by plotting residuals in some logical order, such as order in which data were collected.

Equivalent check is to plot residuals against neighboring value

Create new column of residuals lagged by 1, then plot residuals vs lag(res)

This check can be extended to multiple lags, not just lag 1.

This is accomplished with ACF command in minitab.

4. Residuals normal. This is checked by
 - looking at histogram of residuals
 - checking the fit to normal distribution with rootogram
 - checking fit with nscore(res) vs residuals (straight line if normal)

In this course we use primarily graphic displays to evaluate the assumptions. The reason for this is that statistical tests of assumptions perform poorly. Statistical tests of assumptions are insensitive to violations at small sample sizes, which is precisely when violations can distort estimates of p-values. Tests of assumptions become increasingly sensitive to minor violations at large sample sizes, which is when violations no longer distort estimates of p-values. Statistical tests of assumptions seem like a good idea, but upon examination turn out to be a bad idea.

4. State population and whether sample is representative.

Once we have an acceptable model, we move to statistical inference. In confirmatory statistics, an inference is made from a sample to a population. In text examples it is often difficult to identify the population to which an inference is being made. Often, the population is statistical: all possible measurements of the response variable, under some specified set of conditions. Inference to an enumerable population (e.g., biological population) is rare because there is usually insufficient information to state the probability with which each individual from the population was sampled.

In exploratory statistics no inference is made from sample to population. Instead "batches" of observations are examined for pattern, relative to some stated criterion.

5. Decide on mode of inference. Is hypothesis testing appropriate?

Hypothesis testing (H_A versus H_0) is usually considered mandatory. This view stems from R.A. Fisher.



Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis

--R.A. Fisher 1935

This view has come increasingly into question. By the mid 1990s M.R. Nester (1996) had collected over 125 quotes against the use of hypothesis tests. Many statisticians now recommend that hypothesis testing be replaced by parameter estimates accompanied some measure of error. This is more flexible and allows us to include biological reasoning in drawing conclusions. The generic recipe (Table 11.1) adopts a compromise. The logic of hypothesis testing is presented because it is so widely used. Its limitations have been pointed out along the way. The alternative (step 10 in the recipe) has been demonstrated whenever appropriate.

6. If hypothesis testing is appropriate, state H_A / H_0 pairs, with tolerance for Type I error.

The alternative hypothesis is that the response variable is related to an explanatory variable. This is usually the basis for undertaking an analysis, so it makes sense to write this first. In the analysis of variance the H_A / H_0 pair is expressed about the variance due to a term in the model. The hypothesis pair concerning a term in the model is equivalent to a general hypothesis about the parameters: that the means differ, the slopes are not zero, or that interactive effects are present.

The logic of hypothesis testing with inferential statistics is that all possible outcomes can be divided into two categories: those included under the alternative hypotheses, and those not included. These latter are labelled the null hypothesis. The H_A/H_0 pair should include all outcomes. For example, if the H_A were that the expected value for a treated group exceed the expected value for a control group, then the H_0 would be that the treated group was less than or equal to the control group.

$\alpha = 5\%$ This is the conventional criterion for statistics in biology. It is an arbitrary compromise between Type I and II statistical error. Reducing the Type I error by making α smaller will raise Type II error, the chance of missing a true effect. The criterion used should be stated before undertaking a statistical analysis. A convenient place to do this is in the material and methods section of a scientific report or thesis.

7. ANOVA - Calculate df and variance, partition according to model.

The total degrees of freedom (sample size minus 1) are partitioned according to the model statement. This is easily done by hand. It is also done by the statistical package. Computation of df by hand is a good way to check that the model statement is carrying out the analysis intended.

The sum of the squared deviation of the data from the mean is partitioned according to the GLM written for the analysis. The sum of squares to be partitioned is

$$SS_{\text{total}} = \sum Y^2 - n^{-1}(\sum Y)^2$$

In Minitab:

```
MTB> let k1 = ssq('Y')
MTB> print k1
```



Completion of table.

In learning to use the GLM it helps to use the following sequence:

Write out the headings of ANOVA table

Sources	df	SS	MS	F
---------	----	----	----	---

Fill in sources

Fill in df (these are partitioned according to the model)

Add SS_{tot} to bottom of table.

Add partitioned SS to the table from a computer print-out.

Compute MS from SS/df, if not already done by the computer.

Compute F from MS/MS, if not already computed.

Once this sequence becomes familiar, it is sufficient to undertake a quick partitioning of the degrees of freedom by hand, check these against the print-out, make sure the intended SS were printed, and compute MS and F-ratios if not already computed correctly by computer.

Calculate Type I error.

It is easier to calculate a p-value from a theoretical distribution than from an observed distribution that must be generated by randomization. p-values from F, t, chi-square and other distributions no longer need to be the approximate values obtained from tables. Any good statistical package (e.g. Minitab, SAS) will allow you to calculate exact p-values from F, t, chi-square and other distributions. These are more informative than critical or cut off values obtained from printed tables.

8. **Recompute p-value if necessary**

If the assumptions are not met, it is the p-value (not the F-ratios or other components of the analysis) that cannot be trusted.

Decision to recompute based on 3 questions.

Assumptions not met ? If so, skip to next step.

df_{err} small ?

If $df > 100$ little need to recompute p-value, even if residuals terrible.
the p-value won't change by much.

If $30 < df < 100$ then may need to recompute if residuals terrible
the p-value may change

If $df < 30$ then recompute p-value if residuals terrible.

p close to α ?

If p not close (e.g. twice or half α) then recomputation unlikely to
change the decision, even though the p-value is incorrect.

If the assumptions are not met, the remedy is to compute the p-value by randomization. This results in an observed distribution of outcomes from the data, when the H_0 has been made true by randomizing the data so as to remove pattern. Outcomes tabulated as frequency distribution. Then compute p-value of observed statistic (data not randomized). Could this statistic, of this magnitude, have arisen by chance ?

9. **Declare decision, with evidence**

In this step we use the logic of the null hypothesis to declare decisions about terms in the model. We reject decisions that the parameters (means) are equal, or that a slope is equal to zero, or that interactive effects are absent.

In declaring any statistical decision it helps to remember that rejecting the H_0 eliminates chance as an explanation for a particular outcome. Eliminating chance does not, however, establish causality. The relation of the response variable to the explanatory variable may be due to some factor other than the explanatory variable. An example of this is a regression of food consumption against age. Food consumption appears to rise with age in animals that increase in size with age. Food consumption changes in more complex ways with age, after adjusting for the effects of size.

The conclusion from a statistical analysis with the GLM should contain, at a minimum, the F-ratio, df, p-value, and whether the p-value was obtained from a theoretical frequency distribution or from frequency distribution generated by randomization.

9. Declare decision, with evidence

It is a good idea to reflect at this point on the population from which inference is being made.

Statistical population: all possible measurements based on the procedural statement

Enumerable population: all units in a definable frame.

For example, individuals in a biological population.

Were the measurements or units representative of the population ?

10. Report and interpret parameters of biological interest.

This is at least as important as declaring decisions about terms in the model.

Often we want to go beyond a statement that an effect (term) is significant.

We want to know how large the effect was: how far apart were the means ? What is the slope of a linear relation ?

We want to know whether parameters were estimated with great uncertainty or with little uncertainty. So we look at some measure of uncertainty for each parameter. This can be a confidence limit, a standard error, or a standard deviation.

At this point entertain questions brought in for answering.

Table 11.1 Commonly Used Tests, Based on the General Linear Model.

Analysis	Response Variable	Explanatory Variable	Interaction	Comments
t-test	1 ratio	1 nominal	Absent	compares two means
1-way ANOVA	1 ratio	1 nominal	Absent	compares 3 or more means in 1 category
2-way ANOVA	1 ratio	2 nominal	Present	tests for interactive effects compares means in 2 categories, if no interaction
Paired Comparison	1 ratio	2 nominal	Absent too few df	compares 2 means in 1 category, controlled for 2nd category (blocks or units)
Randomized Blocks	1 ratio	2 nominal	Assumed Absent*	compares 3 or more means in 1 category, controlled for 2nd category (blocks or sampling units)
Hierarchical ANOVA	1 ratio	≥ 2 nominal	Absent	nested comparisons of means
ANCOVA	1 ratio	≥ 1 ratio ≥ 1 nominal	Present	compares two or more slopes
			Assumed Absent*	compares means, controlled for slopes
Regression	1 ratio	1 ratio	Absent	tests linear relation of response to explanatory
Multiple Regression	1 ratio	\geq ratio	Assumed Absent*	tests linear relation to 2 explanatory variables relation expressed as a plane

*The interaction term is often assumed to be absent. Including the interaction term allows us to check the assumption. This is a good idea provided there are at least as many df in the error term as the interaction term.

