Generalized linear models: model selection, diagnostics, and overdispersion

Erin Carruthers[1,2], Keith Lewis[1,2], Tony McCue[1,2], Peter Westley[1,2,3]

[1]Authorship order is alphabetical. All authors contributed equally
[2]Department of Biology, Memorial University of Newfoundland
[3]Ocean Sciences Centre, Memorial University of Newfoundland

March 4, 2008

Contents

Introduction

When one's scientific path and statistical education intersects with the generalized linear model (we will use the abbreviation GLM for generalized linear models and LM for general linear models following modern statistical conventions), it opens many doors and provides for a more holistic approach to analyzing data. Opening these doors comes at a price; however, because one leaves the relatively safe confines of the LM, where interpretation and diagnostics are well established, and steps onto ground that is less firm. In addition to the usual issues of study design, model building, model selection and diagnostics, there are also issues of overdispersion, choice of the link term, and clustered data. Many of these issues are more complicated with GLMs than with LMs. Some of us who have used GLMs for years have struggled to find a firm footing, i.e. find clear answers to the above issues. This has not always been easy since the literature does not always provide a clear guide. This paper is an attempt to find terra firma and to bring our previous best to new heights.

Linear models have been applied to an almost unimaginable range of problems in many different fields. A linear model essentially assumes a linear relationship between two or more variables (e.g. as X increases so does Y). Most introductory courses are taught, either explicitly or implicitly, within the framework of the General Linear Model (LM). Developed in the 1950s, LMs unified many existing linear models (e.g. t-tests, ANOVA, regression, ANCOVA) that assume the error term (i.e. the residuals) is normally distributed and are calculated using the method of least squares. In the 1970s, the Generalized Linear Model (GLM) was introduced, extending the LM to include models with non-normal errors (e.g. binomial, Poisson). As with LMs, the response variable is assumed to be independently distributed (although how one determines independence is anyone's guess). Such tests include log-linear models and logistic regression and are calculated using Maximum Likelihood Estimates (MLE). Quasi-likelihood, developed in the mid-1970s, allowed for GLMs to be used on a broader class of response variables. In the late 1980s, Generalized Estimating Equations (GEEs) were developed which allow for the analysis of non-normal, clustered data (e.g. repeated measures, Littel et al. 2002). These models are usually concerned with population-level inferences. Finally, Generalized Linear Mixed Models (GLMMs) have been developed more recently and extend GLMs to include random-effects (Agresti 2002). In contrast with GEEs, these models often have subject-specific interpretations. This paper focuses primarily on GLMs (with special reference to logistic regression) but GEEs are also discussed. Other areas of statistics such as multivariate, non-parametric, and Monte Carlo approaches, as well as other model families like Generalized Additive Models (GAMs) are beyond the scope of this paper.

We organized this document to reflect the process of building, evaluating, and testing statistical models which reflect *your understanding* of an ecological system. Specifically, this is the result of a search to answer the following questions:

1) What are the criteria for building models? What types of model selection criteria are available and which is most appropriate?
2) What are appropriate diagnostics for the GLM (with special reference to logistic regression)?
3) What is overdispersion, how is it detected, and how should it be dealt with?
4) What is clustered data and how it should be analyzed?
This summary has involved a literature search of both primary and grey literature. In each section, where possible/appropriate, we attempt to define the issue, describe how the issue should be dealt with, and offer practical suggestions for how this is to be accomplished in R and SAS. We have kept the style informal and where no resolution was achieved on an issue, we leave the question in the text with potential answers. This is meant to be a working document.
We assume the reader has a basic understanding of GLM and has taken Quantitative Methods in Biology or an equivalent course. While the goal of this paper has been to find *terra firma* and a 'code of rules' for GLM, remember:
"…the code is more what you'd call "guidelines" than actual rules" - *Captain Barbossa*.

Model Selection Procedures

This section provides an overview of model selection approaches with special emphasis on information-theoretic criteria (i.e. AIC, BIC and others). Selection among a suite of models has become a common approach to interpretation of complex biological systems. At some point when using statistical inference, a biologist must decide among the main approaches: frequentist (hypothetico-deductive or information-theoretic) or Bayesian; choosing the most appropriate method to the task at hand. If one chooses a model selection approach, be reminded that "Although selection procedures are helpful exploratory tools, the model-building process should utilize theory and common sense" (Agresti 2002, p. 217). Beware of model dredging with too many *a priori* models (Burnham and Anderson 2002; i.e. going on a fishing trip with model selection procedures).

Implicit in the approaches presented herein are development of *a priori* models developed as likely alternatives based on knowledge of the systems of interest (Burnham and Anderson 2002, Guthery et al. 2005). Development of ecologically plausible models is based on expert insight of the biological processes at work – this is the foundation of model selection. Kadane and Lazar (2004) pose the following basic questions as a guide: "Which measures are important to the outcome? Which are not? Are there interactions between the variables that need to be taken into account?" In addition to considering ecological plausibility, data collection, parameter estimation, and preliminary diagnostics are crucial in developing a suite of 'potential' models among which one, or a few, best models may be selected for inference.

An important point to remember concerning model selection is that all of the procedures select the 'best' fit models from the *a priori* options in a relative framework. Therefore, there is no

assurance that the 'best' model is actually a good predictor or explanation of the processes at work. Additionally, one must remember that the procedures presented here may not select a single 'best' model. In this situation Faraway (2002) asks: 1) do the models have similar qualitative consequences; 2) do they make similar predictions; 3) what is the cost of measuring the predictors; and 4) which has the best diagnostics?

Following is an overview of available methods for model selection with insights to application. We do not present full mechanics of implementation here, as this information is readily available elsewhere (e.g. Agresti 2002, Burnham and Anderson 2002, Faraway 2002). We refer the reader to Foster (2002), Burnham and Anderson (2002), and Ward (2007) for more in depth coverage of the field.

Stepwise Procedures

These procedures have been extensively used, but are often dissuaded for use in the selection process based on being judgment oriented, not rigorous, and not criterion-based (e.g. Burnham and Anderson 2002, Kadane and Lazar 2004, Whittingham et al. 2006). 'Significance' of parameters should not be the only criterion for inclusion. These procedures are: Backward Elimination (sequential deletion of terms until removal leads to "significantly" poorer fit), Forward Inclusion (addition of terms until no "significant" improvement in fit is detected), and Stepwise Regression (iterative addition/deletion of terms to include all terms that "significantly" affect fit and none that do not).

Criterion-Based Procedures (non-information-theoretic based)–

Predicted Residual Sum of Squares (PRESS):

This criterion selects the model with lowest sum of squared errors without inclusion of the *i*th case. PRESS tends toward larger models.

Adjusted R^2:

This criterion functions to minimize variance. Burnham and Anderson (2002) state that adjusted R^2 is a useful measure of proportion of variation, but not useful in model selection as many models will be nearly equal.

Mallows' C_p Statistic:

Mallows' C_p is easy to compute, closely related to adjusted R^2, and may be used to guide the researcher in the process of model subset selection. Choosing the model that minimizes C_p and estimating parameters with least squares is valid for normal errors, homogenous variance, but prone to bias and should be avoided in most other situations.

Information-Theoretic Procedures

Information-theory developed in the 1950s and was quantified in statistics with Akaike Information Criterion in the 1970s. No attempt will be made to detail the theory or procedures here. An extensive summary of information theoretic criteria involving model parsimony and the practical use of model inference can be found in Zellner et al. (2002) and Burnham and Anderson (2002) respectively.

Akaike Information Criterion (AIC):

AIC is a valid procedure to compare non-nested models. AIC is a better estimator of predictive accuracy, whereas BIC (see below) is a better criterion for determining process (Foster 2002, Ward 2007). Detractors contend that AIC tends to over fit the data (e.g. Kadane and Lazar 2004). Note if your model or data are severely overdispered AIC will result in biased outcomes and other model selection procedures are more appropriate.

AIC_C is a second-order AIC for small sample sizes. This modified criterion contains an additional bias adjustment, which will tend toward AIC in large samples. AIC_C is recommended over AIC any time (n/global K) < 40 (i.e. any time you have less than 40 observations per parameter; Burnham and Anderson 2002).

$QAIC_{(C)}$ is a modified AIC for models containing an overdispersion parameter. This can be computed in both large and small sample versions described above. The objective of this quasi-likelihood form is to balance over- and under-fitting of the models when overdispersion is present (Anderson et al. 1994).

It is recommended that researchers present AIC differences (ΔAIC) and Akaike weights (ω_i) as the "raw" AIC value is meaningless outside of the context of the other models under consideration. Akaike weights function as evidence ratios for interpreting results of multiple models. There is no advantage to bootstrapping models over using Akaike weights (Burnham and Anderson 2002).

Models with AIC differences <2 should be considered equivalent in strength of inference; those >10 can be omitted from further consideration (eliminate model with higher AIC values and lower weight).

Schwarz or Bayesian Information Criterion (BIC):

Tends to smaller models than AIC (due to an extra penalty for parameters when n<7.4) Kadane and Lazar (2004), proponents of BIC state, "In general, models chosen by BIC will be more parsimonious than those chosen by AIC." While detractors contend that BIC

underfits the data and introduces bias in the form of overestimating precision (e.g. Burnham and Anderson 2002).

The following are less commonly used information-theoretic criteria that the reader may encounter in literature.

Deviance Information Criterion (DIC):

DIC was proposed as an equivalent Bayesian method to AIC. Beware that many questions are still unanswered regarding its use (Ward 2007).

Takeuchi's Information Criterion (TIC):

TIC is more generalized than AIC (AIC is a special case) by containing a more general bias adjustment term. Burnham and Anderson (2002) summarize that matrix error estimation can cause instability with this procedure, although it is useful for large sample sizes with good estimates of matrix elements in the case of poor approximating models.

Focused Information Criterion (FIC), Network Information Criterion (NIC), and Risk Inflation Criterion (RIC) are included among other alternatives of information-theoretic model selection criteria that have not gained popular use in biological science.

Summary

AIC and other information-theoretic approaches have largely replaced p-value based procedures in some fields of biology (Johnson and Omland 2004). Information-theory in general has been commonly promoted as a means to further scientific knowledge over hypothetico-deductive approaches (see review by Anderson et al. 2000). While this may not always be the best method of inference, we present a summary of the methods here, as they are now part of mainstream ecological literature (Hobbs and Hilborn 2006). However like null hypothesis testing, with which information-theoretic approaches are contrasted, information-theoretic model selection criteria can be sloppily applied (see Guthery et. al 2005 and citations thereafter).

Model Evaluation and Diagnostics

After you run a model, but before interpreting results, you will want to assess it. How well does your simplification of reality (your model) explain the ecological phenomena? There are two types of diagnostics: those that are used to evaluate model fit and those that identify where (which data points) the model fits poorly. Model fit reflects whether the appropriate link function and structural model (i.e., relevant predictors are included, predictors are not correlated and irrelevant predictors are excluded) have been specified (see Littel et al. 2002; Agresti 2002,

2007; Crawley 2002). Problems in the data can be typos, outliers and influential data points, but also aspects of the data set that affect model performance, such as sparse data and zero cell counts (Menard 1995). Because methods of evaluating logistic regression models are limited, special attentions should be paid to the effects of 'data problems'. Menard (1995) cautions that if statistical assumptions are violated, you may (1) systematically over- or under-estimate coefficients, (2) have large standard errors, or (3) have inaccurate statistical significance.
One of the reasons for choosing a GLM, instead of a LM, is to account for the error structure. By selecting an appropriate link function variance should be homogeneous and deviance residuals normal. Thus, by using the appropriate link, assumptions about the error structure are like that for LM. To visually assess the link function fit, plot the linear predictor against the estimated link function – the plot should be linear. For example, the code for evaluating the Poisson distribution in SAS is:

Code box 1:

```
plot (y*=xbeta+(y-pred)/pred) v xbeta
```

where y is the response variable and pred is predicted values, and xbeta is the estimate of the link function. All are available through OBSTATS in SAS. To assess whether a logit link is appropriate, plot the natural log of the odds (y =ln(p/(1-p))) versus each of the predictors.

To test homogeneity of variance assumption, plot the standardized residuals against fitted values: this serves the same purpose as LM. This graph could indicate poor model choice or a poor link function (Littel *et al.* 2002). Note that there seem to be variations on this theme – some recommend plotting deviance or Chi-square residuals v. fits. For a general discussion of diagnostic plots see McCullagh and Nelder (1989). Littel *et al.* (2002) provide examples for Poisson distributions. This should apply to all other distributions except binomial and beta-binomial. For binomial distributions, diagnostic plots of residuals do not seem useful; the 2-line residual plot is difficult to interpret. Instead Pregibon (1981) and Menard (1995) concentrate on diagnostics that identify where the model fits poorly.

Overdispersion is an issue that commonly arises in GLM model fitting and is a relatively complicated and contentious issue. We therefore have devoted a whole section to why and when it occurs and ways of handling it when it does arise (see below).

Key assumptions related to the structural model are that all relevant predictors are included, irrelevant ones excluded and that the predictors are not correlated. Choice of which predictors to include is based on theory, your understanding of the ecological system, and available resources. Menard (1995) stated that bias from excluding relevant variables is more important than inefficiency from including irrelevant ones. We discuss approaches for determining which predictors to include in the model selection section of this document. Here we will focus on identifying correlation among predictors.

Lack of correlation among predictors is one of the key assumptions of both LM and GLM. Ideally, one would check for association among predictors prior to building a model, using a contingency table approach for categorical variables and correlation for continuous variables. Menard (1995) recommends (re)considering correlation/collinearity among predictors, if the model is significant but parameters are not, or if the standard errors of the coefficient are very high. High standard errors could also be the result data problems, such as zero cell counts or outliers, described below.

Zero cell counts, having no observations for a level within a categorical predictor, leads to inefficient estimation of parameters (super-high covariates and whopper-sized standard errors). These should not affect overall model fit because those cell contain little information and, therefore are given little weight in the model fit. For categorical predictors, another consideration is the number of cases per level. Agresti (2002) made a distinction between *c* and *n*, where c is the number of cases (i.e., df = c-1), for large *n* the G^2 approximates Chi-squared distribution but usually poor when the n/c<5. Thus, with categorical predictors guidelines for the number of explanatory variables are based on the number of individuals per case – not sample sizes. Agresti (2007) states that even with sample sizes of 1000, if successful outcomes are rare, the number of explanatory variables should be limited to ~3 if there are only 30 times the result in one outcome (either successes or failure).

Menard (1995) suggests checking for coding errors when you find extreme values because they may skew regression estimates. However, if after scrutinizing extreme data points, there are no coding errors or other justifications for removing the data points, they must remain. This may indicate a missing explanatory variable. For example, Menard (1995) did not remove outliers from his 'pot-smoking teenagers example' – those who had delinquent friends but did not inhale – after scrutinizing the extreme values he found no errors or reasons to exclude those values from the model. There were likely some attributes (e.g. IQ) of the anomalous teens that Menard (1995) was not including in his model.

Influence measures compare models with and without extreme values to assess whether or not those values were driving the result or how much overall deviance in the model is attributable to those extreme values. Venables and Ripley (2004) suggest using the R function 'influence.measures.'

Software box 1
Normality of Deviance Residuals: Deviance residuals should be normally distributed (Pierce and Schafer 1986) - this is produced in Software:
R: Rcmdr provides a panel of 4 diagnostic plots to test model assumptions. Diagnostics can be easily added to the data set and new graphs generated. The command plot (glm) is useful for plotting four typical model diagnostic plots.

The printout from R-help files states:
Plot(glm) produces four plots. The first is the jackknife deviance residuals against the fitted values. The second is a normal QQ plot of the standardized deviance residuals. The dotted line is the expected line if the standardized residuals are normally distributed, i.e. it is the line with intercept 0 and slope 1. The final two panels are plots of the Cook statistics. On the left is a plot of the Cook statistics against the standardized leverages. In general there will be two dotted lines on this plot. The horizontal line is at 8/(n-2p) where n is the number of observations and p is the number of parameters estimated. Points above this line may be points with high influence on the model. The vertical line is at 2p/(n-2p) and points to the right of this line have high leverage compared to the variance of the raw residual at that point. If all points are below the horizontal line or to the left of the vertical line then the line is not shown. The final plot again shows the Cook statistic this time plotted against case number enabling us to find which observations are influential. (R-help files)"
(http://stat.ethz.ch/R-manual/R-patched/library/boot/html/glm.diag.plots.html)

Most of these are easy to get in Rcmdr by adding the residuals to the data set.

SAS: provides all of the examples from Littel et al. (2002) quite easily (it is after all from a SAS manual)
There are high quality graphics available in SAS through the ODS - http://www.ats.ucla.edu/stat/SAS/faq/sas9_stat_plots.htm- see this site for PROC REG diagnostics. However, these only seem to be available for a limited number of PROCs - it's a bit annoying but indicates that consensus on diagnostics is not as concrete as it is for regressions. One could probably produce the diagnostic plots by requesting residuals and graphing them but it seems to be hard in SAS.

Overdispersion

What is overdispersion

The following paraphrases from the various cited references in an attempt to summarize the available literature, rather than attempting to write an original paper. As a warning, that this is a difficult section. The concept of overdispersion is considered, even by statisticians, easy to state but difficult to mitigate. There is little or no disagreement over the definition of overdispersion, the archetypical definition articulated by Crawley (2002) who defines overdispersion simply as the case where the residual deviance is greater than the residual degrees of freedom. In otherwords, if φ (ratio of deviance and df) > 1 one speaks of overdispersion because the data have larger variance than expected under the assumption of the chosen binomial distribution (Højsgaard and Halekoh). Crawley (2002) goes on to discuss overdispersion as the polite statistician's version of Murphy's Law. Overdispersion tends to arise because 'you' have not measured one or more of the factors that turned out to be important. It may also occur if you specify an incorrect error distribution. This means that the probability you are attempting to model is not constant within each cell, but behaves like a random variable. This in turn results in an inflated residual deviance.

Thus, by not accounting for overdispersion you may increase your type I error.

In the worst case, all the predictor variables you have measured may turn out to be unimportant so that you have no information at all on any of the genuinely important predictors. In this case the minimal adequate model is just the overall mean, and all your 'explanatory' variables provide no extra information. Young et al. (1999) agree:

"Overdispersion should be accounted for in a [GLM] analysis. Failure to do so in the presence of overdispersion results in type I error rates well above the nominal ones. When overdispersion is not present, the test for treatment effects is not negatively affected by considering overdispersion."

When is overdispersion a problem and how does it arise?

There are no clear cut decision rules, when to decide that there is sufficient overdispersion to be problematic in analyses. As mentioned previously, there are no adverse consequences of explicating considering overdispersion in analyses, therefore, if there is any semblance of inflated deviance, the investigator should definitely account for overdispersion (see subsequent sections for techniques on how to implement this).

Two completely different reasons can underlie the phenomenon of a Pearson statistic χ^2 from a fitted logistic model to be larger than expected. First, systematic deficiencies of the model (see GLM diagnostics). Second, unexplained random variation (i.e. overdispersion). These two explanations have different implications for the use of the fitted model (Højsgaard and Halekoh). Systematic deficiencies of the model may be due to specification of the wrong link function, missing covariate(s), failure to consider and implement transformations (e.g. the logarithm of a covariate may be better then the covariate itself), and outlying observations (Højsgaard and Halekoh 2005). Overdispersion does not exist in the circumstances where the dependent variable is a Bernoulli [0, 1] variable (e.g. happens for example in logistic regression with continuous covariates), or when the largest model under consideration is equal to the saturated model.
If even after investigating these potential pitfalls the variance is larger than explicable by the [binomial] assumption (which forces the dispersion parameter to be 1), there are some extra sources of variation, which must be addressed. There are two different causes for overdispersion which have the same statistical implications; random variation in response probabilities, and interaction (correlation) between [binary] responses (See Højsgaard and Halekoh 2005).

How much overdispersion is too much? How do you know when you really must account for overdispersion in your model? The following are several opinions. First, if the ratio of the Pearson-statistic to its degrees of freedom is about 2 or larger. McCullagh and Nelder (1989, p.125) argue that unless there are good reasons for relying on the [binomial assumption], it seems to be wise to be cautious and to assume that overdispersion is present. Second, Crawley (2002) suggests testing (using quasi-likelihood and F vs. Chi-sq tests) for the affect of

overdispersion anytime the ratio of residual deviance to residual degrees of freedom is >1. Similarly, McCullagh and Nelder (1989) indicate that any level of dispersion >1 should be considered and modeled to prevent magnification of error rates and skewed confidence intervals. As a warning, overdispersion should be suspected for repeated counts but not frequencies (Lindsey 1999).

According to Anderson et al. (1994), once one has found an adequate model structure, overdispersion, values between one and three are typical. Sophisticated modeling of overdispersion may well be unnecessary at these low levels. Conversely, if overdispersion is as big as 10 (and perhaps if it is as much as 5), or more, important structural variation remains to be extracted from the data (i.e., the model selected is not structurally adequate; Anderson et al. 1994).

In summary, if there is large overdispersion (>5), you probably missed something and need to add it to the model. If there is overdispersion present but its low (<5), check the structure of the model. If it seems OK, then try to remedy overdispersion (see below).

How to deal with overdispersion, assuming that the structural model is acceptable?

There are several typical techniques employed when overdispersion is present to account for inflated deviance. It is important to note that these techniques do not rid the model of overdispersion, but rather embrace and account for it in the model structure. The first general technique is to take a quasi-likelihood approach. In this technique one uses the same error distribution (i.e. Poisson, binomial) but adjusts the standard errors and test statistics. Specifically, hypothesis testing with an F-test instead of Chi-squared is recommended. In other words, model overdispersion by letting the actual variance equal the assumed variance multiplied by an additional scale parameter that adjust for the discrepancy between assumed and actual (Littel et al. 2002). Another way of thinking of this is that using the F-test vs. Chi-squared (or any other 'solution') does not make overdispersion go away but simply takes it into account in our hypothesis testing.

The quasibinomial and quasipoisson families included in the R stats package are quite simple. Following McCullagh & Nelder, the quasibinomial model keeps the same structure, except that "the variance is inflated by an unknown factor "δ^2" (p. 126). The estimates of the b are not changed, but the estimates of the standard errors are changed. McCullagh & Nelder (1989) observe that the covariance matrix of the parameter estimates is inflated according to the estimated value of the dispersion coefficient."(GLM3)

It is comforting to know that the quasi-likelihood approach is quite robust as stated in the passage below:

"When you want something as good as a "sum of squares model" or a maximum likelihood model, but you don't have the tools for either, there is an alternative that is "almost as good, i.e. the quasi-liklihood approach."(GLM3)
Generally, quasi-likelihood adjustments (i.e., use of $\hat{c} > 1$) are made only if some distinct lack of fit has been found (for example, if the observed significance level $P \leq 0.15$ or 0.25) and the goodness-of-fit degrees of freedom ≥ 10, as rough guidelines (Anderson and Burnham 2002).

Selected quote on overdispersion:

"…and then the magical part of quasi-lielihood becomes apparent: When you want something as good as "sum of squares models" or a maximum likelihood model, but you don't have the tools for either, there is an alternative that is "almost as good" [quasi-likelihood].

- From Johnson (2006)

What are the costs of overdispersion, i.e. what does it cost to correct for it and what does it cost not to correct for it? e.g. why not just use quasibinomial?

The impact of estimating the dispersion parameter on the parameter estimates and the estimated variances are given in the following list:
• parameter estimates: they are unchanged,
• estimated (co)variances: multiplied by $\hat{\varphi}$,
• Log-likelihood, scaled deviance divided by $\hat{\varphi}$,
• Wald-CI-intervals: width of the interval sqrt$\hat{\varphi}$-times larger (Højsgaard and Halekoh 2005).

Ultimately, the cost of incorporating overdispersion estimates into your model is trivial compared to the cost of not incorporating when you really should have. Specifically, if you do not correct for overdispersion, the estimates of the standard errors are too small which leads to biased inferences, i.e. you will observer smaller p-values than you should and thus make more Type I errors. As a result, confidence intervals will also be incorrect (Oregon stats).

Model Validation

Once one has completed the above steps and satisfied themselves that their chosen model is the 'best', testing of the model with independent data becomes tantamount. Ground-truthing the predictive capability of the model should therefore be addressed and results reported (Burnham and Anderson 2002, Guthery et al. 2005). Various methods of model testing exist dependent on model objectives and available data (e.g. bootstrapping, randomization, prospective sampling). More robust estimates of prediction error will be provided by using independent data. Fielding and Bell (2002) provide a thorough summary of available methods for testing models with binary data (i.e. logistic regression and others) and the strengths and weakness of summary statistics of predictive power resulting from these validation procedures.

Software box 2:
R: specify the family (e.g. quasipoisson) (see Quick R for simple examples)
SAS: Use the dscale/pscale options in the model statement. This causes all standard errors and test statistics to be corrected for a scale parameter estimated using the deviance. No advice on which is better. Approach criticized as simplistic - a different distribution is likely bett boer (Littel et al. 2002).
Do dscale and pscale represent the estimators (resid deviance function/df) and (Pearson's χ^2 /df)? If so, Young et al. (1999) and Williams (1988b in Engel and Brake 1993) both recommend use of Pearson's χ^2 as a better estimator.
In SPlus, you correct for overdispersion with count data or proportion data very simply,
by stating test="F" instead of test ="Chi" in the anova function. See pp 518, 545, etc.

"Proc GENMOD fits generalized linear models (GLM) and handles modelling of overdispersed Poisson data using quasi-likelihood (e.g. SCALE=P) or the negative binomial distribution (DIST=NEGBIN), but not generalized linear mixed models (GLMM). In R a GLM is fitted using the glm function, and specifying family=quasipoisson is the equivalent of SCALE=P in SAS.

In contrast to the quasi-likelihood approach, one could assume a new distribution that does a better job a modelling variance (GLM3) - if Poisson, use Negative Binomial, if binomial, use beta-binomial (Lindsey 1999): if normal, use gamma (DCS pers. comm.). The negative binomial case is handled either by the negative.binomial family function (when the shape parameter is known) or the glm.nb function (if you want to estimate the shape by ML). Both negative.binomial and glm.nb are found in Venables and Ripley's MASS package." http://www.mail-archive.com/r-help@stat.math.ethz.ch/msg17801.html

Clustered Data: mixed effects models and GEEs

What is clustered data?
Clustered data is data that is correlated in some fashion. Common examples are:
-Longitudinal or repeated measures (same treatment many times on one subject). Note: some people distinguish between longitudinal and repeated measures - *I don't understand why*.
...maybe repeated measured are a subsection of longitudinal studies
-Cross-over (several treatments on one subject, e.g. one patient gets several types of drugs over the study period).
-clustered (correlated data - e.g. offspring from same litter, patients from same clinic, plants on the same plot)
Each of these examples clearly violate the statistical assumption of independence (of the errors)!!

What is the problem for ecological research?

As with logistic regression, the medical people seem to be way ahead of ecologists on this one. The standard medical experiment would be a control group and treatment group. A proportion of each group respond positively. Site effects either don't exist or are pooled (either explicitly or naively, i.e. ignored). Note: traditionally, proportions or counts were analyzed with a one- or two-way G-test. Once you reach 3 EVs, it became a matter of self preservation to use a

computer and do a GLM (i.e. logistic regression). Of course, GLM can now easily be used for any number of EVs.

However, more modern medical experiments are often not conducted in this fashion. Consider the following medical example. A study is conducted on the health of children in 16 communities. Half of the communities are control, half are treatment (depending on the question, it may not be possible to apply the treatment randomly to children - this is a community level question). One thousand kids are monitored in each community. Does n = 16 or 16,000. A pseudo-replication purist would probably suggest that n = 16. If the purist is correct, there is little advantage in this case to have 1000 subjects per community - 10 might work just as well as long as the mean is well estimated (of course the confidence intervals would be much smaller for the 1000 subject approach). However, running a model where n = 16,000, while it takes advantage of all of the information, is clearly inappropriate and is referred to as naive pooling (I suspect that these types of data were often explicitly pooled or analyzed with LM).

Now consider common ecological experiments where counts or proportions are measured on a per plot basis (e.g alive v. dead on multiple plots). Does one take the pseudo-replication approach or the naive pooling approach? A particularly thorny problem for ecologists is that we must consider how many plots are needed and how many measurements to make within the plot (no solution on this yet). Analyzing these types of data is not straightforward (nor is conducting a power analysis on them). Neither the psuedo-replication approach or the naive-pooling approach offer much guidance. Is there a better way? Yes!

Recent advances allow the analysis of such data in a flexible but valid manner. For an excellent reference, (see Burton et al. 1998). Panageas et al. (2008) demonstrate the need for clustering in medicine. Ying and Liu (2006) do the same and show relevant code for SAS. Bogarts (2004) shows a way to estimate sample sizes and gives a very good slide show presentation.

Note: Beware of “Mason’s diffusion process” where by new statistical techniques are touted as panaceas by proponents and disparaged as “nothing new” by critics (*I’m quite guilty of this and may be committing this error now*).

Recognize clustering when you encounter it (*Under development*)
“Observations within a cluster tend to be more alike than observations from a different cluster” (Agresti 2007). I’m not sure how to extend this but want to avoid a psuedo-replicationist witch hunt.

Understand why analysis needs to take clustering into account or Why do correlated errors matter?
• The key outputs of regression analysis are
estimates of

– β
– $ó^2$, variance of the errors, a key ingredient of how useful the model is, plus for predicting new values
– SE(β), standard error of estimated β, which determines the p-value and width of the confidence intervals

Failing to take clustered data into account makes estimated B, $ó^2$, SE(B) wrong; even in the best case, $ó^2$, SE(B) will be too small.....so this means a smaller p and CI than correct (Kleinman and Colegio 2007)

What to do about it? (Most of this is lifted from Burton et al. (1998) who give an excellent example data situation and explore different, valid models to analyze the same data).

1) Summarize data (data resolution) within cluster (e.g. take the mean of repeated measures) - one way to avoid the psuedo-replicationists.

2) Pretend there are no clusters, i.e. naive pooling. This may work in some cases but you'll have to defend it (at least to yourself). Probably only if you can justify that within cluster correlation is really low!

3) Summary statistic or similar approach, i.e. paired t-test, random block (include site as a blocking variable and pretend its fixed so you don't have to do GLMMs), repeated measures ANOVA, MANOVA - limited to categorical predictors and normal errors. Doesn't work well for GLM unless you transform the data. For example, its easy to generate a Difference statistic for a paired t-test, just take the difference between the paired measurements. But how does one do this for a proportion??

KPL; Note that this approach may cause the model to crash - happened for my PhD.

4) GEE: Generalized Estimating Equations are a way to analyze clustered data within the GLM framework. These models are also called marginal models or population averaged models. The treatment is modeled separately from the within cluster correlation, i.e. treatment effects are averaged across clusters (Kuss no date).
Regression coefficients of a PA model describe the average population response curve. In PA, one explicitly models the marginal expectations while choosing a var/cov structure that adequately describes the correlation pattern among the repeated measurements (http://www.uoregon.edu/~robinh/gnmd13_rm_gee.txt) .

Note: The number of clusters is an issue (http://www.uoregon.edu/~robinh/gnmd13_rm_gee.txt) but see Haloekoh et al. 2006 for a solution.

Quote: “In general, I think of GEE as “taking account” of the correlation, treating it as an annoyance to be coped with en route to accurate inference. (This is reflected in estimation of the estimates.)”(REF?)

5) GLMM: GLMM (or random effects models or subject-specific (SS) models) treats the heterogenity among clinics as something of interest that can be modeled by a probability distribution (Kluss no date). In this case, one may be interested in individual responses. Regression coefficients of a SS model describe what the average individual's response curve looks like. In SS models, model individual heterogeneity using subject-specific random effects which partially determine the var/cov structure.
(http://www.uoregon.edu/~robinh/gnmd13_rm_gee.txt) .

For a comparison of the results of these approaches on the same data, see Table 1 (actually just the other attachment for now).

GEE v. GLMM - section *underdevelopment*

See Kluss (no date) for a good example of differences between GLMM and GEE.

Variance and correlation are mathematically identical (between GEE and GLMM)!! But: with GLMM, you can more easily get an estimate of $ó_b^2$, the variance between clusters, as well as estimated b_i within each cluster.

Table. Differences in the GEE and GLMM approach.

	GEE	GLMM
Variance	λ2	$\sigma_b^2+\sigma^2$
Correlation	A	$\sigma_b^2/(\sigma_b^2+\sigma^2)$
Covariance	α λ2	σ_b^2
	Correlation is a nuisance	Variance strucutre is interesting
	Account for wrong variance structure	Model assumptions (easier inference about variance, data can be MAR, fewer clusters OK
	Missing data must be MCAR	
	Needs 20-40 clusters	Assumptions must be true

Software box 2:

SAS: Use the REPEATED statement in PROC GENMOD. Specify the correlation matrix in the options. The GEE approach is generally robust to mis-specification of the correlation matrix (see (http://www.uoregon.edu/~robinh/gnmd13_rm_gee.txt for meaning of the different correlation matrixes). Bios 265 (no date) gives a good example of the influence of different correlation matrixes and what to do if you have lots of drop outs per cluster.

GLMM: Hallahan (2006) gives a good overview of how to implement GLMMs in SAS PROC GLIMMIX- it will not be considered further here.

R: Use package Geepack. This is probably better than SAS, especially for small cluster sizes, because it uses the jackknife variance estimator. See Haloekoh et al. (2006) for complete details on how to implement.

Anoymous (no date) gives a good summary of SAS and R code for clustered data.

References

Agresti, A. 2007. An introduction to categorical data analysis. Second Edition. John Wiley & Sons, Hoboken, NJ, USA.

Aitkin, M. 1996. A general maximum likelihood analysis of overdispersion in generalized linear models. Satatistics and Computing 6:251-262.

Anderson, D. R., K. P. Burnham, G. C. White. 1994. AIC model selection in overdispersed capture-recapture data. Ecology(75): 1780-1793.

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64: 912–923.

Burnham, K. P., and D. R. Anderson 2002. Model Selection and Multi-model Inference: A Practical Information-Theoretic Approach. Second Edition. Springer-Verlag, New York, NY, USA.

Burton, P., L. Gurrin, P. Sly. 1998. Tutorial in Biostatistics: extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modeling. Statistics in Medicine 17: 1261-1291.

Chen, J. J., H. Ahn, K. F. Cheng. Comparison of some homogeneity tests in analysis of over-dispersed binomial data. Environmental and Ecological Statistics 1: 315-324

Crawley, MJ. 2002. Statical computing: An introduction to data analysis using S-Plus. Wiley, New York.

Engel, B. and J. Brake. 1993. Analysis of embryonic development with a model for under-or overdispersion relative to binomial variation. Biometrics 49(1): 269-279.

Faraway, J. J. 2002. Practical Regression and Anova using R. <cran.r-project.org/doc/contrib/Faraway-PRA.pdf>

Fielding, A. H., and J. F. Bell. 2002. A review of methods for the assessment of prediction erros in conservation presence/absence models. Environmental Conservation 24: 38–49.

Foster, M. R. 2002. The new science of simplicity. Pages 83–119 *in* Zellner, A., H. A. Keuzenkamp, and M. McAleer, eds. Simplicity, Inference and Modelling: Keeping It Sophisticatedly Simple. Cambridge University Press, Cambridge, UK.

Guthery, F. S., L. A. Brennan, M. J. Peterson, and J. J. Lusk. 2005. Information theory in wildlife science: critique and viewpoint. Journal of Wildlife Management 69: 457–465.

Haloekoh, Højsgaard, and Yan. 2006. The R package geepack for Generalized Estimating Equations. Journal of Statistical Software 15(2) 1-11.

Hobbs, N.T., Hilborn, R. 2006 Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. Ecological Applications 16: 5-19.

Horton, N. J., and S. R. Lipsitz. 1999. Review of Software to Fit Generalized Estimating Euqation Regression Models. The American Statistician 53: 160-169.

Johnson, J. B., and K. S. Omland. 2004. Model selection in ecology and evolution. Trends in Ecology and Evolution 19: 101–108.

Kadane, J. B., and N. A. Lazar. 2004. Methods and criteria for model selection. Journal of the American Statistical Association 99: 279–290.

Lindsey, J. K. 1999. On the use of corrections for overdispersion. Appl. Statist. 48(4): 553-561.

Littell, , R. C., W. W. Stroup, R. J. Feund. 2002. SAS for Linear Models 4th edition. Cary, NC: Sas Institute Inc.

McCullagh, P, and JA Nelder. 1989. Generalized linear models. Chapman Hall, London.

Menard, S. 1995. Applied logistic regression analysis. Sage Publications, Thousand Oaks, CA.

Paul, S. R. and A. S. Islam. 1998. Joint estimation of the mean and dispersion parameters in the analysis of proportions: a comparison of efficiency and bias. The Canadian Journal of Statistics 26(1): 83-94.

Ward, E. J. 2007. A review and comparison of four commonly used Bayesian and maximum likelihood model selection tools. Ecological Modelling 211: 1–10.

Whittingham, M. J., P. A. Stephens, R. B. Bradbury, and R. P. Freckleton. 2006. Why do we still use stepwise modeling in ecology and behavior? Journal of Animal Ecology 75: 1182–1189.

Williams, D. A. (1988b). Overdispersion in logistic-linear models. In Proceedings of' the Third International Workshop on Statistical Modelling, Vienna, 165- 174.

Young, L. J., N. L. Campbell, G. A. Capuano. 1999. Analysis of overdispersed count data from single-factor experiments: a comparative study. Journal of Agricultural, Biological, and Environmental Statistics 4(3): 258-275.

Zellner, A. ., H. A. Keuzenkamp, and M. McAleer (editors). 2002. Simplicity, Inference and Modelling: Keeping It Sophisticatedly Simple. Cambridge University Press, Cambridge, UK.

Useful websites

Abzug, R. and J. S. Simonoff. 2004. S-PLUS/R Code to Perform the Analyses inthe Book Nonprofit Trusteeship in DifferentContexts. http://pages.stern.nyu.edu/~jsimonof/NonprofitTrusteeship/splus.r.code.pdf

Johnson,P. 2006. GLM3 GLM (Generalized Linear Model) #3 (version 2) http://pj.freefaculty.org/stat/GLM/GLM3_v1.pdf

Højsgaard and Halekoh 2005. Overdispersion Søren Højsgaard and Ulrich Halekoh; Biometry Research Unit Danish Institute of Agricultural Sciences; June 1, 2005 Printed: June 1, 2005 File: overdispersion

Hρjsagaard, S., and U. Halekoh. 2005. Overdispersion. http://genetics.agrsci.dk/statistics/courses/phd05/material/src/overdispersion

Thompson, L. A. 2007. R (and S-PLUS) Manual to Accompany Agresti's Categorical Data Analysis (2002) 2nd edition.

Kuss, O. How to use SAS for logistic regression with Correlated Data. SUGI 27

Panageas, K. S. et al. 2008. The effects of clustering of outcomes on the association of prodedure volume and surgical outcomes. Ann Intern Med. 139: 658-665.

Ying, G-s, and C. Liu, 2006. Statistical analysis of clustered data using SAS system. NESUG 2006.

Weblinks

A cool lab webpage with intro to R

http://ecology.msu.montana.edu/labdsv/R/

Quick R

http://www.statmethods.net/

Oregon stats page - takes a bit of navigating but LOADED with good stats info and SAS code (this guy knows his stuff!!!)
http://www.uoregon.edu/~robinh/statistics.html

Grey Literature

Anonymous (no date) Longitudinal data.

Bios 265 (no date). R. A. GEE Analysis Using SAS PROC GENMOD of Cigarette Smoking Trends among Young Adults: 1986-1993.

Bogarts, K. 2004. An introduction to the analysis of cluster randomized trials. Power Point.

Hallahan, C. 2006. Proc Glimmix. Power Point Presentation.

Kleinman K. and R. Colegio. 2007. Clustered Data, mixed effects models, and GEE's. Power point presentation.