EVALUATION OF GENERALIZED LINEAR MODEL ASSUMPTIONS USING RANDOMIZATION

Tony McCue, Erin Carruthers, Jenn Dawe, Shanshan Liu, Ashley Robar, Kelly Johnson

Introduction

Generalized linear models (GLMs) represent a class of regression models that allow us to generalize the linear regression approach to accommodate many types of response variables including count, binary, proportions and positive valued continuous distributions (Nelder and Wedderburn, 1972; Hilbe, 1994; Hoffman, 2004). Because of its flexibility in addressing a variety of statistical problems and the availability of software to fit the models, it is considered a valuable statistical tool and is widely used. In fact, the generalized linear model has been referred to as the most significant advance in regression analysis in the past twenty years (Hoffman 2004).

Generalized linear models include three components: 1) a random component which is the response and an associated probability distribution; 2) a systematic component, which includes explanatory variables and relationships among them (e.g., interaction terms): and 3) a link function, which specifies the relationship between the systematic component or linear predictor and the mean of the response. It is the link function that allows generalization of the linear models for count, binomial and percent data thus ensuring linearity and constraining the predictions to be within a range of possible values (Guisan, 2002). This ability to handle a larger class of error distributions and data types is a key improvement of GLMs over linear models. Stated formally, the three components of a GLM are:

(1) $g(\mu)=\mu$,

(2) and, $\mu=\beta_0 + \beta_i X_i + \beta_{i+1} X_{i+1} \ldots$
where, μ is the mean of the response.

(3) $g(\mu)$ shows that the mean of the response is linked to the structural model, thus $g(\mu)$ is the link function, and η is the structural component, with X_i denoting explanatory variables, and β_i the parameters to be estimated.

Assumptions and Diagnostics

Similar to the linear model approach, there are key assumptions that must be met when computing a p-value using the GLM approach and violation of any of these assumptions may compromise the interpretation of model results by producing biased standard errors and thus unreliable p-values. There are, however, disagreements in the literature on what constitutes key assumptions, decisions and checks for generalized linear modeling. Because the type I error (the p-value) on the improvement in fit with the GLM is calculated from the chi-square distribution which assumes homogenous, normal, and independent deviations centered on zero (Dobson, 2002), it follows that these are considered key assumptions for GLMs. There is a general consensus that the assumptions of homogeneity and independence of residuals must be met (Breslow, 1996; Lindsey, 1997; Cameron and Trivedi, 1998; Dobson, 2002; Hoffman, 2004). McCullogh and Nelder (1989), however, point out that the independence assumption can be relaxed to "at least uncorrelated". The importance of normality of residuals in GLMs, on the other hand, is debated. Some authors (e.g., Lindsey, 1997; Dobson, 2002; Hoffman, 2004) suggest that normality of the residuals must be met to correctly interpret the results while others (Gill, 2001) note that normally distributed errors are not a condition of GLM quality but simply a description of model behavior. In addition to the assumptions of the chi-square distribution stated above, Breslow (1996) also considers the correct specification of the variance function (ν), the overdispersion factor (θ) and the link function (g) to be critical assumptions underlying GLMs.

Homogeneity, normality and independence

The chi-square distribution assumes that the error term for all combinations of the independent variable is homoscedastic (i.e., same scatter) (Dobson, 2002). When faced with heteroscedastic errors, the standard errors of the coefficients are biased thus the significance tests are incorrect and the ability to make inferences from the model is compromised (Hoffman, 2004). Graphically, a post-model scatterplot of the residual and fitted values can indicate homoscedasticity. Often, the variability of the error term increases with larger values of the independent variables and this is shown by a cone or

fan in the residuals versus fitted values plot. An hourglass pattern, when there is a large deviance of residuals from the line, at low and high extremes of the independent variable may also be evident. These plots may also show outliers and inadequacy of the model (Seber, 1980). Formal diagnostic tests are based on statistical hypothesis testing; the null hypothesis (variances are equal) is tested against the alternate hypothesis that they are not. We chose not to consider formal statistical tests of the assumptions, prior to the main statistical test, because this returns to methodologies which the GLM approach was designed to avoid – "rattling through an extensive toolbox full of distinct and separate tests" (Gill, 2001, p.90). A description of several formal statistical tests for distribution assumptions and how to implement is provided in Greene (2000).

The chi-square distribution also assumes that the residuals are normally distributed with mean=0. As mentioned above, there is disagreement in the literature surrounding the importance of this assumption for GLMs. Graphical analysis of normality is generally performed using normal probability plots and histograms of residuals (Lindsey, 1997; Dobson, 2002; Hoffman, 2004). The points in the plot should lie on or near the straight line representing normality and systematic deviations or outlying observations indicate a departure from this distribution (Dobson, 2002).

Another assumption of the chi-square distribution is that of statistical independence of the errors indicating observations are random and there is no relationship in space or time. This assumption is in doubt whenever there is a natural grouping or clustering of the data. In this course we focused on graphical representations for testing the assumptions of homoscedasticity and normality of residuals but we were not exposed to the evaluation of the independence assumption. Graphical diagnosis of independent residuals is to plot each residual against a neighbouring value (e.g., using a lag plot). Residuals should fluctuate randomly with no pattern and an upward or downward trend indicates that the residuals may be related (Dobson, 2002; Hoffman, 2004).

Overdispersion and Link Functions

The assumption that the variance is equal to the mean is restrictive for most biological

data (Van Hoef and Boveng, 2007). Count data, in particular are often overdispersed, that is exhibiting more variation than given by the mean. However, GLM models typically used for binary or count data, (e.g., a logistic regression or log-link with Poisson error distribution) do not have a separate dispersion term. The variance is assumed equal to the mean. Because overdispersion is so common, models such as the quasi-poisson and negative binomial model have been developed for these data. The quasi-poisson model specifies the variance by adding an over dispersion parameter (θ) (i.e., specifies the relationship between the variance and the mean) while the negative binomial model assumes that the variance is larger than the mean (Hoffman, 2004; Van Hoef and Boveng, 2007). Overdispersion may also result from poor choice of link function, missing terms or interactions in the linear predictor, or outliers in the data (Myers et al., 2002). Determining the cause of a poorly fit model may be more difficult as the symptoms of a poorly specified model are often the same as an overdispersed model (Myers et al., 2002). In this paper, we calculate the dispersion as the ratio of the residual deviance over residual degrees of freedom. If the variance is equal to the mean, dispersion should be one.

Diagnosing Assumptions

There are two approaches available to examine the assumptions of homoscedasticity, normality and independence, these being informal graphical methods and formal test methods. Informal graphical methods involve visual inspections of residual plots. If the above mentioned assumptions of the chi-square distribution are satisfied, residuals should be independent, have a distribution which is approximately normal with a mean of zero and have a constant variance (Dobson, 2002). For each graphical plot of residuals, there is an associated formal statistical test which involves hypothesis testing (Seber, 1980). The main disadvantage of using a formal test is that sample size can largely affect the decision of whether the model fits the data or not (Cameron and Trivedi, 1998). For smaller sample sizes, formal tests lack power. With a large dataset, even mild deviations from non-normality may be detected, but there would be little reason to abandon the model because the effects of non-normality are mitigated. Cameron and Trivedi (1998) liken formal tests to black boxes which provide a single number compared to a critical

value. Furthermore, interpretation of the p-value does not indicate what action to take.

Graphical methods are considered more informative and, further, that formal tests are unnecessary (Seber, 1980; Cameron and Trivedi, 1998; Gill, 2001; Dobson, 2002; Hoffman, 2004). Residual plots are relatively easy to construct and appropriate graphical tools exist in most statistical software (Feder, 1974; Carrol and Spiegelmann, 1992). Visual analysis of residuals can potentially detect violations, ways they can be corrected, as well as provide a feel for the effect of the violation (Cameron and Trivedi, 1998). Some skill is needed, however, to interpret graphical representations of residuals. For example, patterns are often overlooked in plots of residuals from large sample sizes (Seber, 1980). We chose to focus on graphical methods for evaluating model fit because these methods will provide the maximum amount of information from the residuals such as the nature of the misspecification, thus also aiding in identifying the appropriate ways to correct it. However, because we do not yet have a feel for the implications of violating the assumptions of chi-square distributions, we compared p-values based on chi-square distributions with those generated by randomization.

Randomization

Resampling methods can take on many forms in modern statistical analyses including randomization, bootstrap, jackknife, and Monte Carlo. Randomization (or permutation) tests involve reordering observed data values (i.e. reshuffling). Bootstrapping differs only with regard to replacement in the sampling procedure (Potvin and Roff 1993, Manly 2007). Monte Carlo methods are a more generalized approach within while the previous methods may be considered mode specific approaches (Crowley 1992, Manly 2007). Jackknife sampling involves iteratively removing a sample of observations, a technique which is relatively easy to compute, but crude and laden with more assumptions (Crowley 1992). Each of these methods has attributes best suited to specific applications (see Crowley [1992], Edgington [1995], and Manly [2007] for detailed coverage of random resampling methods). We employed permutation tests as a means to calculate distribution-free p-values for each dataset under consideration. This is the simplest method randomization, with the fewest assumptions, when explanatory variables are

fixed.

Randomization carries relatively few assumptions providing more power and accurate p-values compared with model-based methods (Crowley 1992, Manly 2007). The p-values from randomization should equal those of a model when assumptions are reasonable (Petraitis et al. 2001, Manly 2007). Edgington (1995) recommends computing a test statistic from the experimental data, then repeatedly randomizing outcomes and computing the statistic for comparison. This is recognized as the simplest method available. Alternatives, which may be more reliable, involve randomizing residuals, but are typically "not much better than the use of the t- and F-distributions" (Manly 2007, p.201). We randomized response values and calculated the G-statistic (likelihood-ratio χ^2) for GLMs. A minimum of 1000 permutations were computed for all datasets, as recommended by Manly (2007), with a significance level of 5%.

Methods

In this paper, we conducted GLM analyses on multiple datasets. Initial choice of link functions and error distributions were based on knowledge of the data set and error distributions. We evaluated the models using the dispersion parameters, and graphical methods to identify violations of assumptions of homogeneity of variance, normality of residuals, and independence of both explanatory variables and residuals. We evaluated the efficacy of various methods of evaluating assumptions for computing the Type I error using randomization. Brief descriptions, verbal models and variable definitions for the data sets used are detailed in Appendix A.

The open-source statistical package R, with 'MASS' and 'car' packages for GLM confidence intervals and diagnostics, was used to implement and evaluate the models (Venables and Ripley 2002; Fox, 2007; RTeam, 2007). Generic code is presented in Appendix B. We checked for linear correlations among continuous explanatory variables prior to running the analyses. Where linear correlations were found, we chose one of the explanatory variables (e.g., dataset 13, Table 1). After running the analyses, we examined plots of residuals versus fitted values for the assumptions of homogeneity of variance and

where appropriate, if the straight line assumption was met. We looked for well scattered plot with no cones or fans for homogeneity of variance. The same plot was examined for diagnostic arches or bowls, which would indicate that the straight line assumption was not met. To check for normality we used a qqplot, looking to see if residuals fell along the 1:1 line. The lag plot was used to examine independence of the residuals versus the lagged residuals. We looked for plots which contained no trends. Dispersion and the link function were checked with quick calculations instead of graphical methods. We checked the quality of the model fit by plotting the linear link function and considering whether the slope of the line is below 1, the linear link underfit cases larger observed values (Gill, 2001). Conversely, slopes greater than 1, overfit smaller observed values. Gill (2001), however, does not provide guidance on what indicates an inappropriate linear link. Decisions made by analysts based on our examination of residuals are detailed in results section.

Results

A variety of datasets were analyzed using both GLMs and randomizations. Calculated p-values from these analyses were compared. This exercise focused on the importance of meeting various assumptions of GLMs. Datasets analysed, using various GLMs, commonly violated assumptions during the process. One way of checking to see whether the model chosen is accurate is to run a randomization. If randomizations are used, there must be enough iterations run to detect the result (i.e. 5000 randomizations when $\alpha = 0.05$, 10000 randomizations when $\alpha = 0.01$).

The purpose of this report was to determine whether failure of key assumptions skewed resulting p-values. If p-values differed markedly between randomization and model fitting, we returned to the list of assumptions to determine where the violations occurred.

When comparing the two methods for determining p-values, any large deviations, which significantly affected results were flagged by highlighting them in green (Table 1). This indicated model analyses must not have produce accurate results and interpretation of

model results may lead to the wrong conclusion. Each model assumed certain criteria (e.g., normality, homogeneity, independence of explanatory variables).

Out of a total of 141 p-value comparisons across 16 different data sets, 39 comparisons had large deviations between the model results and the randomization results. Assumptions most commonly violated were homogeneity, normality and dispersion, respectively (Table 2).

Results for individual datasets

Dataset 1

I initially tried the Gaussian error structure, as that is one of the first things to look at with count data. Most assumptions were not met, due to the skewing of the data (this tends to occur with percentage data). The dispersion of the data, however, was close to zero, which is to be expected with Gaussian error. The next step was to change the error structure to Poisson (also suggested when using percentage data). Again, most of the assumptions were not met, probably again due to the skew of the data. The data was severely underdispersed. Changing the error structure again, this time to a Gamma structure, the assumptions were met. Dispersion was improved (1.03). Randomized p-values agreed with the chi-square distribution values.

Dataset 2

I started with a Gaussian error structure with an identity link to make sure my data was not normally distributed. When running the model homogeneity, normality, and dispersion were violated. The over-dispersion was huge. I switched to a Poisson error structure with a log link, as data are counts. This violated homogeneity and dispersion. However, the normality plots looked much better. The overdispersion was still huge. I then switched to a quasi-Poisson error structure with a log link to fix the over-dispersion. For this, all assumptions were met.

I focused on the interaction terms because their p-values were highly significant and I wanted to determine whether these values changed. The quasi-Poisson error structure with log link deemed the best model for interpretation. Based on the chi-square

distribution p-values, interactions F1*F2 and F2*F3 were highly significant, while F1*F3 was not significant. The randomized p-values show the same results.

Dataset 3

I started with a Gaussian error structure with an identity link to make sure my data was not normally distributed. Normality and dispersion were violated. Overdispersion was huge. I switched to a Poisson error structure with a log link because data are counts. All assumptions were met, but overdispersion was still a small issue. I then switched to a quasi-Poisson error structure with a log link to fix the over-dispersion. All assumptions were met.

I focused on the interaction terms in the first analysis because their p-values were highly significant however when the error structure was changed, the interaction terms became non-significant and I could interpret the main effects. quasi-Poisson error structure with log link deemed the best model for interpretation. Based on the chi-square distribution p-values, factors F1and X1 were highly significant, while factor F2 was not significant. The randomized p-values show the same results (factors F1 and X1 highly significant, F2 not).

Dataset 4

Initially I used a Gaussian error structure with an identity link. The normality assumption was violated and was the data was overdispersed. This violated normality and dispersion. The over-dispersion was huge. I then switched to a Poisson error structure with a log link because data are counts. The normality assumption was violated. I then switched to a quasi-Poisson error structure to see if the minor overdispersion problem was solved. Normality assumption was still violated and the overdispersion value remained the same. No improvement in assumptions, therefore Poisson error structure with log link is deemed the best model for interpretation.

There was only one explanatory variable here so I focused on whether this p-value changed. Based on the chi-square distribution p-values, factor F1 was highly significant. The randomized p-values show the same results (factor F1 highly significant).

Dataset 5

Because the response variables were counts, Poisson distribution with a log link was chosen first, and 'concentration' was treated as ratio scale. All the assumptions were met except the homogeneity and normality assumptions. When Gaussian distribution with an identity link was used, errors were heterogeneous, non-normal and overdispersed. Next, both of these two distributions were tried again when 'concentration' was considered as a categorical variable, and the assumptions were still not met. Among these models, the Poisson distribution with 'concentration' as a categorical scale seemed to be the most improved one. Although the homogeneity and normality assumption were violated, the model-computed p-values agreed with the p-values that calculated by randomization.

Dataset 6

I chose the Poisson error structure because data were counted survivals, which is discrete numbers. Except homogeneity, all the assumptions were met when 'concentration' was category; when it was continuous, overdispersion occurred. I then changed the distribution to Gaussian, which did not improve the agreement of the assumptions. Almost all assumptions were met when quasi-Poisson distribution was used and 'concentration' was treated as a continuous variable. In the Poisson model, the types of compounds had a significant effect based on p-value calculated by model but no significant effect according to randomization p-value. By using the quasi-Poisson distribution, we obtained p-values which agreed with randomization p-values.

Dataset 7

Again, the Poisson distribution was chosen because the data were counted numbers. Errors were not homogeneous or normal, however the other assumptions were met. Among the seven p-values calculated by Poisson distribution, only two of them agreed with randomization p-values. For example, as for 'species', model-based p-value was

<0.0001, while randomization p-value was 0.996. Using quasi-Poisson, all the assumptions except homogeneity were met. The p-values calculated by quasi-Poisson distribution agreed with randomization p-values.

Dataset 8

Initially I chose a log-link model with Poisson error distribution because the response variables were counts. The model met most of the assumptions shown in plots. However, overdisperion was substantial, 16 times what the model assumes. Based on the chi-square distribution p-values, all explanatory variables appear significant. For this data set, this would mean that soak time and hook type significantly affect the number of tuna landed per longline set – if overdispersion is ignored. Randomized p-values, however, show none of the explanatory variables were significant. I re-ran the model using a negative binomial distribution to account for overdispersion. Only soak time was a significant predictor in this model. While the negative binomial error distribution corrected overdispersion, I was unable to run a randomization on this model. This may have been due to maximum likelihood estimation not coming to a single value (based on error messages provided by R). Unreasonably, wide confidence intervals indicated remaining problems with this model.

Dataset 9

I chose the Gaussian distribution for this dataset as the response is a continuous variable. All assumptions were met (independent errors, homogeneity, normal errors, appropriate link function) so Gaussian distribution was considered appropriate. The p-values obtained from randomization were very close to the Gaussian model p-values.

Dataset 10

I chose the Gaussian distribution for this dataset as the response is a continuous variable. All assumptions were met (independent errors, homogeneity, appropriate link function) except a few values on the normality plot were off the line so maybe applying a gamma distribution would straighten up my normality plot.

All assumptions were met using the gamma model (independent errors, homogeneity, appropriate link function) but the normality plot looked worse than with the Gaussian model. The normality plot for the Gaussian model isn't half bad but my sample size was small (n=35) and my p-values are close to the critical value (0.05) so I chose to randomize. I first did 1000 randomizations but my p-values were still close to 0.05, so I did 5000 and then 10,000 randomizations which still gave me p-values that were close to 0.05. The p-values obtained from randomization were very close to the accompanying model p-values. The Gaussian model (including randomizations) indicated no significance for any terms (although the p-values for two terms (site and the interaction term between site and age) are close to 0.05) but the gamma model indicated significance of two terms: site and the interaction term between site and age but the p-value is close to 0.05. The Gaussian model appears to be more appropriate as the normality plot is a little better than the gamma

Dataset 11

I chose the Poisson distribution for this dataset as the response variable is counts.

Graphical analysis after conducting the Poisson model indicated heterogeneous errors, non-normal errors and overdispersion. As well the link slope was 0.8 which is less than 1.0. The results of the Poisson model indicate that the three interaction terms are significant but randomization indicates no significance here so the Poisson model, based on assumption violation and overdispersion as well as comparison to randomization, is not an appropriate model. So, I tried a quasi-Poisson and the errors displayed heterogeneity, non-normality and the slope of the link was 0.8 (less than 1). However, despite this, the p-values from the quasi poisson and the randomization were similar. Because the assumptions of homogeneous and normal errors were still not met with the quasi-Poisson, I moved on to a negative binomial model. The assumptions here were all met but the slope link was 0.7 which is less than 1.

With the negative binomial model for this dataset, the confidence intervals were unreasonably wide (e.g., for Bluebell Island 0.02 to 20,000 parasites). Therefore, although dispersion is corrected, there is something wrong with the negative binomial

model for this dataset. It should be noted because the link slope was 0.7 for this model, I changed the link to square root and identity, but R reported an error. As well, when randomization was attempted on the negative binomial model with log link, R reported an error. Therefore, choice of model for this dataset was inconclusive and more consideration is required.

Dataset 12

I started with Poisson error with canonical log link due to count nature of response variable. Only 2 assumptions were violated: strong reverse cone exemplifying heterogeneity and overdispersion by factor of 2.5. To better the model for both the heterogeneous residuals and overdispersion I moved to a negative binomial error with canonical log link. This fixed the overdispersion, but there was still a reverse cone and now a sigmoid curve in the qqplot for normality. The third attempt focussed solely on overdispersion by using quasi-Poisson error with canonical log link. Now the only violation was heterogeneous errors (reverse cone). Randomization shows that the quasi-Poisson error model was the best of the 3, but still exhibited substantial differences in p-values, especially the interaction term of primary interest

Dataset 13

I started with Poisson error with canonical log link due to count nature of response variable. The correlation plot showed the effects of carapace width and weight were nearly equal. I continued with the Poisson error with canonical log link after removing one correlated explanatory variable, followed Agresti's choice of variable to drop. 3 assumptions were violated with Poisson error: reverse cone in res v. fit plot, extensive tail away from normality in lower end of qqplot, overdispersed by factor of 3.3. Attempting to better model both the heterogeneous residuals and overdispersion I moved to a negative binomial error with canonical log link. This fixed the overdispersion, but there was still a reverse cone and now tails on both ends of the qqplot, plus the lines of residuals in the res v. fit plot are curving. The third attempt focussed solely on overdispersion by using quasi-Poisson error with canonical log link as this improved the first dataset model. 2 of the original 3 violations are still present, but randomization

shows that this model error structure predicts 2 of the explanatory variables extremely well, while the regression variable is off by a factor of 1.

Dataset 14

For this data set I chose the binomial model as the data is binary, and could be biologically described in odds or odds ratios. When running the model all assumptions were met. Data was dispersed close to 1. The slope of the line was 1, proving that it was an appropriate model. Randomized p-values agreed with the chi-square distribution values.

Dataset 15

I chose logistic regression initially because the response variable is binary, and because I wanted to express results as odds ratios. All assumptions were met, including dispersion. The ratio of residual deviance over residual *df* was 0.992, unexpectedly good for binomial data. Randomized p-values agreed with those based on the chi-square distribution.

Dataset 16

The model was rerun using a subset of data to see if the large sample size was behind the agreement between the two p-value calculations. Instead, I encountered a problem common in logistic regression with sparse data. Logistic regression was trying to estimate proportions yes or no for each level of categorical variables. Since it cannot estimate what happens between the categories, logistic regression will not work – it will produce inefficient parameter estimates – when there are too few instances in a category level (Menard, 1995). Agresti (2007) provides a guideline of at least 5 instances per level. Basically, I could not run a randomized logistic regression with the reduced data set but did not recognize the underlying problem from the ANODEV table. The problem was, however, evident in the confidence intervals, which were unreasonably wide.

Table 1. Generalized linear model structure, error, evaluation of assumptions, and randomization.

Dataset	Model Structure	Error Structure	Link Function	df (residual)	Deviance (residual)	Model Term	p (LR χ^2)	Assumptions	Slope of Link	No. Permutations	p (randomized)	Comments
1	μ = X1 + F1 + X1*F1	Gaussian	identity	47	0.15579	X1 F1 X1*F1	0.4418 <0.0001 0.4957	☐ Straight Line ☑ Independent ☐ Homogenous ☑ Normal Errors ☑ Linear Link ☐ Dispersion	0.983	1000	0.322 <0.001 0.436	Violated Straight Line, Homogeneity and Dispersion. Switched to a Poisson error structure with a log link.
1	μ = X1 + F1 + X1*F1	Poisson	log	47	7.7723	X1 F1 X1*F1	0.9428 <0.0001 0.2501	☑ Straight Line ☐ Independent ☐ Homogenous ☑ Normal Errors ☑ Linear Link ☐ Dispersion	0.974	1000	0.945 0.010 0.298	Violated Independence, Homogeneity and Dispersion. Switched to a Gamma error structure with a log link.
1	μ = X1 + F1 + X1*F1	Gamma	log	47	48.5342	X1 F1 X1*F1	0.7832 <0.0001 0.3401	☑ Straight Line ☑ Independent ☑ Homogenous ☑ Normal Errors ☑ Linear Link ☑ Dispersion	1.042	1000	0.753 <0.001 0.368	All assumptions met. Considered Best Model!
2	μ = F1 + F2 + F3 + F1*F2 + F1*F3 + F2*F3	Gaussian	identity	80	5589.9	F1*F2 F1*F3 F2*F3	0.0301 0.3206 0.1440	☑ Independent ☐ Homogenous ☐ Normal Errors ☑ Linear Link ☐ Dispersion	1.000	1000	0.034 0.329 0.154	Violated Homogeneity, Normality and Dispersion. Switched to a Poisson error structure with a log link.
2	μ = F1 + F2 + F3 + F1*F2 + F1*F3 + F2*F3	Poisson	log	80	459.64	F1*F2 F1*F3 F2*F3	<0.0001 0.0120 <0.0001	☑ Independent ☐ Homogenous ☑ Normal Errors ☑ Linear Link ☐ Dispersion	1.012	1000	0.094 0.683 0.185	Violated Homogeneity and Dispersion. Switched to a Quasi-poisson error structure with a log link.
2	μ = F1 + F2 + F3 + F1*F2 + F1*F3 + F2*F3	Quasi-poisson	log	80	459.64	F1*F2 F1*F3 F2*F3	0.0162 0.4965 0.0413	☑ Independent ☑ Homogenous ☑ Normal Errors ☑ Linear Link	1.012	1000	0.025 0.551 0.063	All assumptions met. Considered Best Model!
3	μ = F1 + F2 + X1 + F1*F2 + F1*X1 + F2*X1	Gaussian	identity	108	906.7	F1*F2 F1*X1 F2*X1	0.0482 0.0039 0.6957	☑ Straight Line ☑ Independent ☑ Homogenous ☐ Normal Errors ☑ Linear Link ☐ Dispersion	1.000	1000	0.051 0.001 0.662	Violated Normality and Dispersion. Switched to a Poisson error structure with a log link.

Table 1. Generalized linear model structure, error, evaluation of assumptions, and randomization (continued).

Dataset	Model Structure	Error Structure	Link Function	df (residual)	Deviance (residual)	Model Term	p (LR χ^2)	Assumptions	Slope of Link	No. Permutations	p (randomized)	Comments
3	μ = F1 + F2 + X1 + F1*F2 + F1*X1 + F2*X1	Poisson	log	108	164.77	F1 F2 X1	<0.0001 0.7978 <0.0001	☑ Straight Line ☑ Independent ☑ Homogenous ☑ Normal Errors ☑ Linear Link ☑ Dispersion	0.996	1000	<0.001 0.957 0.057	All assumptions met. Comparison with a Quasi-poisson error structure with a log link.
3	μ = F1 + F2 + X1 + F1*F2 + F1*X1 + F2*X1	Quasi-poisson	log	108	164.77	F1 F2 X1	<0.0001 0.8479 0.0003	☑ Straight Line ☑ Independent ☑ Homogenous ☑ Normal Errors ☑ Linear Link	0.996	1000	<0.001 0.851 <0.001	All assumptions met. Considered Best Model!
4	μ = F1	Gaussian	identity	10	58	F1	0.0002	☑ Independent ☑ Homogenous ☐ Normal Errors ☑ Linear Link ☐ Dispersion	1.000	10000	0.011	Violated Normality and Dispersion. Switched to a Poisson error structure with a log link.
4	μ = F1	Poisson	log	10	14.751	F1	<0.0001	☑ Independent ☑ Homogenous ☐ Normal Errors ☑ Linear Link ☑ Dispersion	1.000	10000	0.007	Violated Normality. Comparison with a Quasi-poisson error structure with a log link. Considered Best Model!
4	μ = F1	Quasi-poisson	log	10	14.751	F1	<0.0001	☑ Independent ☑ Homogenous ☐ Normal Errors ☑ Linear Link ☑ Dispersion	1.000	10000	0.007	Violated Normality.
5	μ = F1+F2+F3	Gaussian	indentity	571	4419.6	F1 F2 F3	<0.0001 0.0503 0.5861	☑ Independent ☐ Homogenous ☐ Normal Errors ☑ Linear Link ☐ Dispersion	1.000	5000	<0.0001 0.088 0.005	Violated Homogeneity, Normality and Dispersion. *P-value for F2 near 0.05 ran 5000 iterations.
5	μ = F1+F2+F3	Possion	log	571	364.45	F1 F2 F3	<0.0001 0.3006 0.9767	☑ Independent ☐ Homogenous ☐ Normal Errors ☑ Linear Link ☑ Dispersion	1.000	1000	<0.0001 0.990 0.959	Violated Homogeneity and Normality Change model structure (X1 instead of F1). Considered Best Model!

Table 1. Generalized linear model structure, error, evaluation of assumptions, and randomization (continued).

Dataset	Model Structure	Error Structure	Link Function	df (residual)	Deviance (residual)	Model Term	p (LR χ^2)	Assumptions	Slope of Link	No. Permutations	p (randomized)	Comments
5	μ =X1+F1+F2	Gaussian	indentity	575	38831	X1 F1 F2	<0.0001 0.8965 1.0000	☑ Straight Line ☑ Independent ☐ Homogenous ☐ Normal Errors ☑ Linear Link ☐ Dispersion	1.000	1000	<0.0001 0.016 0.001	Violated Homogeneity, Normality and Dispersion. Switched to a Poisson error structure with a log link.
5	μ = X1+F1+F2	Possion	log	575	369.74	X1 F1 F2	<0.0001 0.3006 0.9767	☑ Straight Line ☑ Independent ☐ Homogenous ☐ Normal Errors ☑ Linear Link ☐ Dispersion	0.995	1000	<0.0001 0.454 0.288	Violated Homogeneity and Normality
6	μ = F1+F2+F1*F2	Gaussian	indentity	72	64252	F1 F2 F1*F2	<0.0001 0.0346 <0.0001	☑ Independent ☐ Homogenous ☐ Normal Errors ☑ Linear Link ☐ Dispersion	1.000	1000	<0.0001 0.248 <0.0001	Violated Homogeneity, Normality and Dispersion. Switched to a Poisson error structure with a log link.
6	μ = F1+F2+F1*F2	Possion	log	72	167.71	F1 F2 **F1*F2**	<0.0001 0.0024 <0.0001	☑ Independent ☐ Homogenous ☑ Normal Errors ☑ Linear Link ☑ Dispersion	1.000	1000	<0.0001 0.782 0.109	Violated Homogeneity and slight deviation in Normality plot. Change model structure (X1 instead of F1).
6	μ = X1+F1+X1*F1	Gaussian	indentity	84	2200881	X1 F1 X1*F1	<0.0001 0.3917 0.8197	☑ Straight Line ☑ Independent ☐ Homogenous ☐ Normal Errors ☑ Linear Link ☐ Dispersion	1.000	1000	<0.0001 0.202 0.555	Violated Homogeneity, Normality and Dispersion. Switched to a Poisson error structure with a log link.
6	μ = X1+F1+X1*F1	Possion	log	84	641.67	X1 F1 X1*F1	<0.0001 <0.0001 <0.0001	☑ Straight Line ☑ Independent ☐ Homogenous ☑ Normal Errors ☑ Linear Link ☑ Dispersion	0.985	1000	<0.0001 0.336 0.448	Violated Homogeneity and slight deviation in Normality plot Switched to a Quasi-poisson error structure with a log link.
6	μ = X1+F1+X1*F1	Quasi-poisson	log	84	641.67	X1 F1 X1*F1	<0.0001 <0.0001 0.0015	☑ Straight Line ☑ Independent ☐ Homogenous ☑ Normal Errors ☑ Linear Link	0.985	1000	<0.0001 <0.0001 0.004	Violated Homogeneity and Normality. Considered Best Model!

Table 1. Generalized linear model structure, error, evaluation of assumptions, and randomization (continued).

Dataset	Model Structure	Error Structure	Link Function	df (residual)	Deviance (residual)	Model Term	p (LR χ^2)	Assumptions	Slope of Link	No. Permutations	p (randomized)	Comments
7	μ = F1+F1+F3+ F1*F2+F1*F3+F2*F3+ F1*F2*F3	Gaussian	indentity	144	86556	F1	<0.0001		1.000	1000	<0.0001	Violated Homogeneity, Normality
						F2	0.0068	☑ Independent			0.081	and Dispersion.
						F3	<0.0001	☐ Homogenous			<0.0001	Switched to a Poisson error structure
						F1*F2	<0.0001	☐ Normal Errors			<0.0001	with a log link.
						F1*F3	<0.0001	☑ Linear Link			<0.0001	
						F2*F3	0.6138	☐ Dispersion			0.962	
						F1*F2*F3	<0.0001				<0.0001	
7	μ = F1+F1+F3+ F1*F2+F1*F3+F2*F3+ F1*F2*F3	Possion	log	144	251.87	F1	<0.0001		1.000	1000	<0.0001	Violated Homogeneity and Normality.
						F2	0.0024				<0.0001	Switched to a Quasi-poisson error
						F3	<0.0001	☑ Independent			0.996	structure with a log link.
						F1*F2	<0.0001	☐ Homogenous			0.706	
						F1*F3	<0.0001	☐ Normal Errors			0.526	
						F2*F3	0.6680	☑ Linear Link			<0.0001	
						F1*F2*F3	<0.0001	☑ Dispersion			0.826	
7	μ = F1+F1+F3+ F1*F2+F1*F3+F2*F3+ F1*F2*F3	Quasi-poisson	log	144	251.87	F1	<0.0001		1.000	1000	<0.0001	Violated Homogeneity.
						F2	0.0340				0.188	
						F3	<0.0001	☑ Independent			<0.0001	
						F1*F2	<0.0001	☐ Homogenous			<0.0001	
						F1*F3	<0.0001	☑ Normal Errors			<0.0001	
						F2*F3	0.7971	☑ Linear Link			0.994	Considered Best Model!
						F1*F2*F3	**<0.0001**				**<0.0001**	
8	μ = X1*F1	Poisson	log	108	1788.6	X1	<0.0001	☑ Straight Line	0.984	5000	0.244	Violated Normality and Dispersion
						F1	<0.0001	☑ Independent			0.331	Residuals increasing above the
						X1*F1	0.0004	☑ Homogenous			0.545	Normality line to the right
								☐ Normal Errors				
								☑ Linear Link				
								☐ Dispersion				
8	μ = X1*F1	Negative Binomial	log	108	101.54	X1	0.0190	☑ Straight Line	0.944	1000	Failed to compute	Violated Independence
						F1	0.0660	☐ Independent				Increasing trend in lag plot
						X1*F1	0.2860	☑ Homogenous				
								☐ Normal Errors				
								☑ Linear Link				
								☑ Dispersion				
9	μ = F1 + F2 + X1 + F1*F2 + F1*X1 + F2*X1	Gaussian	Identity	51	0.42	F_1	0.2300	☑ Straight Line	1.000	1000	0.210	All assumptions met.
						F_2	0.1600	☑ Independent			0.170	
						X_1	0.6200	☐ Homogenous			0.620	
						F_1*F_2	0.9900	☑ Normal Errors			0.990	
						F_1*X_1	0.2200	☐ Linear Link			0.210	
						X_2*X_1	0.1400				0.150	
10	μ = F1 + F2 + X1 + F1*F2 + F1*X1 + F2*X1	Gaussian	identity	25	0.0023	F_1	0.0500	☑ Straight Line	1.000	5000	0.070	Violated Normality.
						F_2	0.1700	☑ Independent			0.190	Switched to a Gamma error structure
						X_1	0.0800	☑ Homogenous			0.090	with an identity link.
						F_1*F_2	0.8400	☐ Normal Errors			0.850	
						F_1*X_1	0.0600	☑ Linear Link			0.070	Considered Best Model!
						X_2*X_1	0.1600				0.170	

Table 1. Generalized linear model structure, error, evaluation of assumptions, and randomization (continued).

Dataset	Model Structure	Error Structure	Link Function	df (residual)	Deviance (residual)	Model Term	p (LR χ^2)	Assumptions	Slope of Link	No. Permutations	p (randomized)	Comments
10	μ = F1 + F2 + X1 + F1*F2 + F1*X1 + F2*X1	Gamma	identity	25	0.21	F_1	0.0290	☑ Straight Line	0.960	5000	0.040	Violated Normality and Dispersion.
						F_2	0.1400	☑ Independent			0.160	
						X_1	0.0600	☑ Homogenous			0.070	
						F_1*F_2	0.8300	☐ Normal Errors			0.830	
						F_1*X_1	0.0320	☑ Linear Link			0.050	Considered Best Model!
						X_2*X_1	0.1300	☐ Dispersion			0.150	
11	μ = F1 + F2 + X1 + F1*F2 + F1*X1 + F2*X1	Poisson	log	58	357	F_1	0.2100	☑ Straight Line	0.800	1000	0.880	Violated Homogeneity, Normality,
						F_2	0.5700	☑ Independent			0.870	Linear Link and Dispersion.
						X_1	0.8500	☑ Homogenous			0.950	Slope was less than 1.0.
						F_1*F_2	0.0000	☐ Normal Errors			0.440	Switched to a Quasi-poisson error
						F_1*X_1	0.0010	☐ Linear Link			0.560	structure with a log link.
						X_2*X_1	0.0080	☐ Dispersion			0.420	
11	μ = F1 + F2 + X1 + F1*F2 + F1*X1 + F2*X1	Quasi-poisson	log	58	357	F_1	0.8200	☑ Straight Line	0.800	1000	0.750	Violated Homogeneity, Normality
						F_2	0.8400	☑ Independent			0.790	and Linear Link.
						X_1	0.9500	☐ Homogenous			0.910	Slope was less than 1.0.
						F_1*F_2	0.1900	☐ Normal Errors			0.190	Switched to a Negative Binomial with a
						F_1*X_1	0.4400	☐ Linear Link			0.360	log link.
						X_2*X_1	0.3500				0.310	
11	μ = F1 + F2 + X1 + F1*F2 + F1*X1 + F2*X1	Negative Binomial	log	58	54	F_1	0.2800	☑ Straight Line	0.700	1000	Failed	Violated Linear Link.
						F_2	0.3100	☑ Independent			to	Slope was less than 1.0.
						X_1	0.8700	☑ Homogenous			compute	
						F_1*F_2	0.0200	☑ Normal Errors				
						F_1*X_1	0.0500	☐ Linear Link				
						X_2*X_1	0.0300					
12	$\mu = F_1 + F_2 + F_3 + F_4 + F_3*F_4$	Poisson	log	58	149.7314	F1	0.0055		0.996	1000	0.196	Violated Homogeneity and Dispersion.
						F2	0.0497	☑ Independent			0.354	Switched to a Negative Binomial error
						F3	0.3718	☐ Homogenous			0.896	structure with a log link.
						F4	<0.0001	☑ Normal Errors			0.003	
						F3*F4	<0.0001	☑ Linear Link			0.680	
								☐ Dispersion				
12	$\mu = F_1 + F_2 + F_3 + F_4 + F_3*F_4$	Negative Binomial	log	58	86.19038	F1	0.0477		0.978	1000	0.092	Violated Homogeneity and Normality.
						F2	0.1974	☑ Independent			0.289	Switched to a Quasi-poisson error
						F3	0.4515	☐ Homogenous			0.703	structure with a log link.
						F4	<0.0001	☐ Normal Errors			<0.001	
						F3*F4	0.0190	☑ Linear Link			0.259	
12	$\mu = F_1 + F_2 + F_3 + F_4 + F_3*F_4$	Quasi-poisson	log	58	149.7314	F1	0.0765		0.996	1000	0.142	Violated Homogeneity
						F2	0.2109	☑ Independent			0.281	
						F3	0.7356	☐ Homogenous			0.825	
						F4	0.0001	☑ Normal Errors			0.002	
						F3*F4	0.1298	☑ Linear Link			0.334	Considered Best Model!

Table 1. Generalized linear model structure, error, evaluation of assumptions, and randomization (continued).

Dataset	Model Structure	Error Structure	Link Function	df (residual)	Deviance (residual)	Model Term	p (LR χ^2)	Assumptions	Slope of Link	No. Permutations	p (randomized)	Comments
13	$\mu = F_1 + F_2 + X_1 + X_2$	Poisson	log	-	-	-	-	☐ Independent	-		-	Failed colinearity: X1=X2 Removed X2.
13	$\mu = F_1 + F_2 + X_1$	Poisson	log	166	558.6295	F1 F2 X1	0.0465 0.6993 <0.0001	☑ Straight Line ☑ Independent ☐ Homogenous ☐ Normal Errors ☑ Linear Link ☐ Dispersion	0.894	1000	0.551 0.897 <0.001	Violated Homogeneity, Normality, and Dispersion. Switched to a Negative Binomial error structure with a log link.
13	$\mu = F_1 + F_2 + X_1$	Negative Binomial	log	166	196.2019	F1 F2 X1	0.4788 0.8798 0.0004	☐ Straight Line ☑ Independent ☐ Homogenous ☐ Normal Errors ☐ Linear Link	0.709	1000	0.334 0.834 <0.001	Violated Straight Line, Homogeneity, Normality, and Linear Link. Switched to a Quasi-poisson error structure with a log link.
13	$\mu = F_1 + F_2 + X_1$	Quasi-poisson	log	166	558.6295	F1 F2 X1	0.4853 0.8962 0.0003	☑ Straight Line ☑ Independent ☐ Homogenous ☐ Normal Errors ☑ Linear Link	0.894	1000	0.485 0.910 0.002	Violated Homogeneity and Normality.
14	$\mu = X1 + F1 + X1*F1$	Binomial	logit	31	48.78	X1 F1 X1*F1	0.7337 0.9458 0.9942	☑ Independent ☑ Normal Errors ☑ Linear Link ☑ Dispersion	1.018	1000	0.745 0.956 0.999	All assumptions met. Considered Best Model!
15	$\mu = X1+X2+F2$	Binomial	logit	212	254.64	X1 X2 F1	0.4640 0.0180 0.0020	☑ Independent ☑ Normal Errors ☑ Linear Link ☑ Dispersion	0.985	1000	0.446 0.020 0.003	All assumptions met. Considered Best Model!
16	$\mu = X1+X2+F2$	Binomial	logit	18	17.81	X1 X2 F1	0.0230 0.6240 0.0530	☑ Independent ☑ Normal Errors ☑ Linear Link ☑ Dispersion	0.983	1000	Failed to compute	All assumptions met. Considered Best Model!

Table 2. Summary of influence of assumption violations

No. of data set	No. of p-value comparisons	No of p-value deviations (as a fraction)	Influence of assumption violations (as a fraction)						
			Straight line	Independent	Homogeneous	Normal	Linear link	Dispersion	Assuptions met
16	141	43/141 (30%)	0/2 (0%)	0/3 (0%)	17/24 (70.8%)	14/23 (60.9%)	1/5 (20%)	12/17 (70.6%)	1/7 (14.3)

Discussion

Using randomization provided a stark illustration of the dangers of simply using p-values and not considering model assumptions. Results were markedly different between p-values based on chi-square distributions and p-values based on randomization tests when assumptions were violated.

In total, 141 p-values were compared using two methods: generalized linear models and randomizations. The results of these comparisons are presented in Table 2. Marked differences between p-value comparisons were found 43 of 141 times (30.5%; green highlights in Table 1). Individual assumption violations were recorded to determine their influence on the p-value deviations. This was recorded as a fraction representing the number of violations when the p-value was different over the total number of violations. Three assumptions (homogeneity of residuals, dispersion parameter of 1.0, and normality of residuals) were most commonly found to be violated when the p-value comparisons had large deviations. Violations of the homogeneity of residuals assumption occurred 17 out of 24 times when the p-values differed substantially (70.8%). Extradispersion (i.e., dispersion parameter less than or greater than one) occurred in 12 out of 17 instances where the p-values were substantially different (70.6%). Violations of the normality of residuals assumption occurred 14 out of 23 times when the p-values were substantially different (60.9%). These results indicate that the assumptions of homogeneity of residuals, dispersion parameter of one and normality of residuals have heavier weights on the determination of the p-value, and when these assumptions are violated, the calculated p-value may be compromised.

Violations of the straight line assumption, independence of residuals assumption, and the appropriate link assumption did not appear to substantially influence changes in the p-value. Respective fractions for these assumptions are: 0/2 (0%), 0/3 (0%), 1/5 (20%). It is also noteworthy that substantially different p-values occurred in this report even when all assumptions were met. This occurred one out of seven times (14.3%). In these cases, there were particularly substantial differences found between the p-values from the

GLMs and the randomizations. Regardless of meeting assumptions, the p-values were found to be very different. In these cases, the randomization are considered the reliable method.

It is also important to note that in some cases, a model was chosen to be the most appropriate for a data set, even though a substantial difference was observed in the calculated p-values (purple highlights in Table 1). This is true of three models in the report. This may be the a result of some violations of key assumptions, as it is sometimes common practice to accept a certain amount of violation as it is considered to be negligible on the result.

We were unable to run randomizations for two models with negative binomial distributions and one logistic regression. In each case, the reported confidence intervals were unreasonably wide thus indicating underlying problems in the models. For the logistic regression analysis, there were not enough instances for logistic regression to produce estimates for levels of the categorical variable indicating that the model was not adequate for the dataset. Shifting the error distribution to negative binomial resolved the overdispersion problem originally identified in the Poisson model, but did not produce a 'best model'. Therefore, additional analyses and consideration of alternate models is required prior to interpretation of these results.

Efficacy of current practices.

In this course, the key methods used for evaluating analysis models were graphical; specifically the straight line, homogeneity of residuals and normality of residuals assumptions. In our results, the straight line assumption was not considered critical when evaluating deviations of the randomized p-value and the chi-squared p-value. As previously discussed, when the randomized and chi-square p-values differed, there were often violations of the homogeneity of residuals assumption and the dispersion assumption (Table 2). These assumptions therefore had the greatest influence on the deviation of the p-values. From this, we can conclude that they are important assumptions when evaluating the efficacy of a model. We also evaluated normality plots of the

residuals from our models. Again, when the p-values differed substantially, the normality assumption was often violated. While not considered as critical as the assumptions of homogeneity of residuals or dispersion, normality seems to also be a key assumption in the evaluation of a model.

Although it was not discussed in this course, we also evaluated the appropriateness of the linear link function. Within our models, this link was not found to be very efficient in our analysis, as a violation of this assumption often was not related to substantial differences in the p-values between GLMs and randomizations. The slope of the link approached one in all datasets with the exception of two where the models indicated a slope of less than one. In one case there was a failure to compute, however, in the other the randomized and chi-square p-values matched, but other assumptions were violated. Therefore, the slope has an influence on the analysis, even when the model computed p-values agree with the randomization p-values.

In summary, several references revealed in the introduction bias the importance of some specific assumptions. After reviewing the analysis performed in this report, conclusions about these assumptions are as follows: violations of homogeneity, dispersion and normality are influential when calculating chi-square p-values. Straight line, independence and linear link were not found to be as influential.

References

Agresti, A. 2007. An introduction to categorical data analysis. Wiley-Interscience: Hoboken, NJ.

Breslow, N. 1996. Generalized linear models: checking assumptions and strengthening conclusions. *Statistica applicata*. 8: 23-41.

Cameron, A. and Trivedi, A. 1998. Regression analysis of count data. Cambridge University Press: Cambridge.

Carroll, R. and Spiegelman, C. 1992. Diagnostics for nonlinearity and heteroscedasticity in errors-in-variables regression. *Technometrics*. 34: 186-196.

Crowley, P. 1992. Resampling methods for computation-intensive data analysis in ecology and evolution. *Annual Review of Ecology and Systematics*. 23:405–447.

Dobson, A. 2002. An Introduction to Generalized Linear Models. Chapman & Hall/CRC: Boca Raton, FL.

Edgington, E. 1995. Randomization Tests. Third Edition. M. Dekker, New York.

Feder, P. 1974. Graphical techniques in statistical data analysis – tools for extracting information from data. *Technometrics*. 16:287-299.

Fox, J. 2007. car: Companion to Applied Regression R Foundation for Statistical Computing: Vienna.

Gill, J. 2001. Generalized linear models: a unified approach. Sage University Paper: London.

Greene, W. 2000. Econometric Analysis. Fourth Edition. Prentice Hall: Upper Saddle River, NJ.

Guisan, A. Edwards, T., Hastie, C. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modeling*. 157, 89-100.

Hilbe, J. 1994. Generalized linear models. *American Statistical Association*. 48: 255-265.

Hoffmann, J. P. 2004. Generalized linear models: an applied approach. Pearson: Boston.

Lindsey, J. 1997. Applying generalized linear models. Springer: New York.

McCullogh, P. And Nelder, J. 1989. Generalized linear models. Second edition. Chapman and Hall: New York.

Manly, B. 2007. Randomization, bootstrap and Monte Carlo methods in biology. Third Edition. Chapman & Hall/CRC: Boca Raton, FL.

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

Myers, R., Montgomery D., Vining, G. 2002. Generalized linear models: with applications in engineering and the sciences. Wiley-Interscience: New York.

Nelder, J, and Wedderburn, R. 1972. Generalized linear models. *J. R. Statist. Soc. A.* 135: 370-384.

Potvin, C. and Roff, D. 1993. Distribution-free and robust statistical method: viable alternatives to parametric statistics? *Ecology.* 74: 1617–1628.

RTeam: R Development Core Team. 2008. R: a language and environmental for statistical computing. Version 2.8.0. Vol. R Foundation for Statistical Computing: Vienna.

Seber, G. 1980. Linear regression analysis. John Wiley and Sons: New York.

Van Hoef, J. and Boveng, P. 2007. Quasi-poisson vs. negative binomial regression: how should we model overdispersion count data? *Ecology.* 88: 2766-2772.

Venables, W. and Ripley, B. 1999. Modern applied statistics with S (MASS). Fourth edition. Springer: New York.

Appendix A

Dataset 1

Purpose:
To determine the effects of population and total area of several US States on the percent of electoral votes they have.

Verbal model:
Do the population size and/or total area of land affect the percent of electoral votes given to each state?

	Variable name	Symbol	Units	Scale
Response Variable	Votes	vote	count	continuous
Explanatory Variables	Population size (X1)	pop	count	continuous
	Total area (X2)	area	m^3	continuous

Dataset 2

Purpose:
To determine whether sampling date (3), cage treatment (3) and segregation of cores into layers (2) affects the number of species present in each core.

Verbal model:
Is total species number a function of sampling date, cage treatment and core layer?

	Variable name	Symbol	Units	Scale
Response Variable	Total # of species	Total	count	continuous
Explanatory Variables	Sampling Date (F1)	Date	days	categorical
	Cage Treatment (F2)	Treat	-	categorical
	Core Layer (F3)	Deplr	-	categorical

Dataset 3

Purpose:
To determine whether sampling location (2), rock shape (3) and rock size affects the species richness on each rock.

Verbal model:
Is species richness a function of sampling location, rock shape and rock size?

	Variable name	Symbol	Units	Scale
Response Variable	Species Richness	Spri	count	continuous
Explanatory Variables	Sampling Location (F1)	Loc	-	categorical
	Rock Shape (F2)	Shape	-	categorical
	Rock Size (X1)	Size	cm	continuous

Dataset 4

Purpose:
To determine whether the type of bird call treatment (5) affects the number of Fork-Tailed Storm Petrels caught in a net.

Verbal model:
Is total species number a function of sampling date, cage treatment and core layer?

	Variable name	Symbol	Units	Scale
Response Variable	Birds Caught	Birdnum	count	continuous
Explanatory Variables	Treatment (F1)	Treat	-	categorical

Dataset 5

Purpose:
To determine the effects of a chemical compound on the survival of green alga *Enteromorpha linza* (Chlorophyta) spores. There are 6 concentration gradients including the control, 5 repeated experiments, and 20 field of view selected randomly for observation.

Verbal model:
Is the survival of spores of Enteromorpha linza depend on compound concentration, control for repeat and field of view?

	Variable name	Symbol	Units	Scale
Response Variable	Survival	Surv	count	continuous
Explanatory Variables	Concentration (F1)	Con	mg/ml	categorical
	Repeat (F2)	Rep	times	categorical
	Field of view (F3)	Fov	-	categorical

Dataset 6

Purpose:
To determine the relationship between the survival of green alga *Enteromorpha linza* spores and three chemical compounds. There are 6 concentration gradients, and three types of chemical compounds.

Verbal model:
Does the survival of Enteromorpha linza spores depend on the concentration and types of compound?

	Variable name	Symbol	Units	Scale
Response Variable	Survival	Surv	count	continuous
Explanatory Variables	Concentration (F1)	Con	mg/ml	categorical
	Compound (F2)	Com	-	categorical

Dataset 7

Purpose:
To determine the effect of three chemical compounds on the survival of two species of green algae spores: *Enteromorpha linza* and *Ulva fasciata.*

Verbal model:
Does the survival of green algae spores depend on the species, the concentration and types of compound?

	Variable name	Symbol	Units	Scale
Response Variable	Survival	Surv	count	continuous
Explanatory Variables	Concentration (F1)	Con	mg/ml	categorical
	Compound (F2)	Com	-	categorical
	Species (F3)	Sp	-	categorical

Dataset 8

Purpose:
To determine possible effects on landed value from pelagic longline sets, I modeled number of tuna landed as a function of the same fishing variables.

Verbal model:
Is the number of tuna landed from each set a result of hook type used and soak time? Is there an interactive effect between the two explanatory variables?

	Variable name	Symbol	Units	Scale
Response Variable	Tuna number	num_tun	count	continuous
Explanatory Variables	Hook type (F1)	hookcd	-	categorical
	Soak time (X1)	time	hours	continuous

Dataset 9

Purpose:
To determine the effects of polychlorinated biphenyls (PCBs) on the activity of the hepatic phase II enzyme UDP-glucuronyltransferase in shorthorn sculpin (*Myoxocephalus scorpius*) at Saglek, Labrador.

Verbal model:
Is the activity of the hepatic phase II enzyme UDP-glucuronyltransferase in shorthorn sculpin related to PCB exposure, fish body mass or sex of the fish?

	Variable name	Symbol	Units	Scale
Response Variable	Enzyme activity	Act	nmol/min/mg protein	continuous
Explanatory Variables	Site (F1)	S1	-	categorical
	Sex (F2)	X1	-	categorical
	Body mass (X1)	M1	grams	continuous

Dataset 10

Purpose:
To determine the effects of PCBs on bone mineral density (BMD) of Black guillemot (*Cepphus grylle*) nestlings at Saglek, Labrador.

Verbal model:
Is the bone mineral density in guillemot nestlings related to PCB exposure, sex or bird age?

	Variable name	Symbol	Units	Scale
Response Variable	Bone mineral density	BMD	g/cm^2	continuous
Explanatory Variables	Site (F1)	S1	-	categorical
	Sex (F2)	X1	-	categorical
	Age (X1)	A1	days	continuous

Dataset 11

Purpose:
To determine the effects of PCB exposure on the abundance of gastrointestinal parasites in shorthorn sculpin at Saglek, Labrador. In particular, I model the abundance of an acanthocephalan, *Corynosema magdaleni* as the response variable.

Verbal model:
Is the abundance of C. magdaleni in shothorn sculpin at Saglek related to PCB exposure, fish bodymass or sex of the fish?

	Variable name	Symbol	Units	Scale
Response Variable	*C. magdaleni* abundance	C1	count	continuous
Explanatory Variables	Site (F1)	S1	-	categorical
	Sex (F2)	X1	-	categorical
	Mass (X1)	M1	grams	continuous

Dataset 12
<data from Agresti 2002, Table 7.1>

Purpose:
Determine if there is selective feeding on the variety of available food items.

Verbal Model:
The number of alligators in 4 Florida lakes select among 5 classes of food. Tests are controlled for gender and size (≤ or > 2.3m) of the alligator.

	Variable name	Symbol	Units	Scale
Response Variable:	Alligators	-	count	continuous
Explanatory Variables:	Gender	F1	-	categorical
	Size	F2	-	categorical
	Lake	F3	-	categorical
	Food	F4	-	categorical

Dataset 13
<data from Agresti 2002, Table 4.3>

Purpose:
Determine if there are physical attributes of female horseshoe crabs that make them more appealing to males during breeding.

Verbal Model:
The number of satellite male horseshoe crabs attending breeding females is dependent on colour, spine condition, weight and/or carapace width

	Variable name	Symbol	Units	Scale
Response Variable:	Satellite males	-	count	continuous
Explanatory Variables:	Colour	F1	-	categorical
	Spine condition	F2	-	categorical
	Carapace width	X1	cm	continuous
	Weight	X2	kg	continuous

Dataset 14

Purpose:
To determine the effects of location and size on the presence or absence of decoration found on decorator crabs (*Hyas araneus*) found in Bay Bulls.

Verbal model:
Are the odds of decoration of decorator crabs a function of location and/or size?

	Variable name	Symbol	Units	Scale
Response Variable	Decoration	dec	yes/no	categorical
Explanatory Variables	Location (F1)	loc	-	categorical
	Size (X1)	size	mm	continuous

Datasets 15 & 16

Purpose:
The purpose of these models (using datasets 15&16) is to determine the effects of fishing factors and fish length on whether or not a common bycatch species, longnose lancet fish (*Alepisaurus ferox*), survives the capture process to be released alive from pelagic longline fishing gear.

Verbal model:
Are the odds of survival of longnose lancetfish a function of hook type, soak time and fish length?

	Variable name	Symbol	Units	Scale
Response Variable	Survival	surv	yes/no	categorical
Explanatory Variables	Hook type (F1)	hookcd	-	categorical

Fish length (X1)	flen	cm	continuous
Soak time (X2)	time	hours	continuous

The difference between the two data sets is simply that the first contains lancetfish bycatch from the fisheries observer data collected between 2001 and 2004 (217 fish). The second contains 23 lancetfish observed in 2003. I used a subset to determine if sample size influenced the randomized p-values.

Appendix B

R_code - GLM Randomization

```
data.name<-read.delim("filename.txt")
names(data.name)
library(car)
pairs(with(data.name,cbind(var.1,var.2,var.3,var.4,...)))
model.name<-with(data.name,glm(response.variable~var.1+var.2+var.3+var.4+...,family=family))
plot(fitted(model.name),resid(model.name))
lag.plot(resid(model.name),diag=FALSE,do.lines=FALSE)
qqnorm(resid(model.name))
qqline(resid(model.name))
plot(fitted(model.name),with(data.name,response.variable))
abline(lm(with(data.name,response.variable)~fitted(model.name)))
coef(lm(with(data.name,response.variable)~fitted(model.name)))
deviance(model.name)
df.residual(model.name)
deviance(model.name)/df.residual(model.name)
Anova(model.name,type="III")
exp.chi<-data.frame(data.frame(Anova(model.name,type="III"))[,1])
rand.chi<-
data.frame(rbind(replicate(####,c(data.frame(with(data.name,Anova(glm(sample(response.variable,##,FALSE)~va
r.1+var.2+var.3+var.4+...,family=family),type="III")))[,1]))))
summary(c(rand.chi[1,])>exp.chi[1,])
summary(c(rand.chi[2,])>exp.chi[2,])
summary(c(rand.chi[...,])>exp.chi[...,])
# NEGATIVE BINOMIAL VARIANT
library(MASS)
model.name<-with(data.name,glm.nb(response.variable~var.1+var.2+var.3+var.4+...,link=link))
rand.chi<-
data.frame(rbind(replicate(####,c(data.frame(with(data.name,Anova(glm.nb(sample(response.variable,##,FALSE)
~var.1+var.2+var.3+var.4+..., link=link),type="III")))[,1]))))
# RANDOMIZED SELECTION OF AUTHORSHIP
sample(c(1:6),6,FALSE)
```