Assignment and Take Home Exam

Biology 7932
Instructor D.C. Schneider

The Generalized Linear Model Approach to Data Analysis
Central Terms

1. Taylor's Power Law

Definition:

Taylor developed the idea that the variance to mean relationship conforms to a power of the mean. In ecological systems this gives a robust method of relating the variance to the mean population size. Taylor's power law is $S^2 = a\mu^b$, where a and b are constants, a is a sampling parameter while b is an index of aggregation characteristics of the species, s^2 is the variance, μ is the mean.

Connection to the Generalized Linear Model:

The generalized linear model is a statistical, linear model that generalizes the General Linear Model in the following ways:

- It permits the usage of error distributions from the exponential family (Poisson, binomial, negative binomial, gamma, etc.), whereas the general linear model only allows the normal distribution to be utilized
- The variance may depend on a known function of the mean (for example the binomial distribution). The dependent variable values are predicted from a linear combination of predictor variables, which are 'connected' to the dependent variable via a link function. (http://www.statsoftinc.com/textbook/stglz.html)

Values of Y or 1-1/2b obtained by fitting either the Box-Cox transformation or Taylors power law and their corresponding transformations.

Power Law	Transformation
1	linear y=ax + b
0.5	square root $y = x^{1/2}$
0	logarithmic
-0.5	reciprocal square root
-1	reciprocal y = 1/x

Perry, J.N. Taylor's Power Law for Dependence of Variance on Mean in Animal Populations. Applied Statistics 30(3): 254-263.

Power, J.H., and E.B. Moser. 1999. Linear model analysis of net catch data using the negative binomial distribution. Can. J. Fish. Aquat. Sci. 56. 191-200. (p192)

Southwood, R. and P.A. Henderson. 2000. Ecological Methods. Chapter 2. Blackwell Science Ltd. Malden, MA. (Online pages 12-13).

http://entomology.unl.edu/lgh/ent806/Lecture13_dispersion.htm
http://en.wikipedia.org/wiki/Generalized_linear_model

2. Scale Parameter, Dispersion Parameter, Variance Function.

Scaled Residuals: The two basic types of residuals are the so-called Pearson residuals and deviance residuals. Pearson residuals are based on the difference between observed responses and the predicted values; deviance residuals are based on the contribution of the observed responses to the log-likelihood statistic. It is possible to scale for the Pearson or the deviance residuals.

Scaled Pearson residuals are the raw residuals (data minus fitted values) divided by the standard deviation of the data according to the model mean variance relationship and estimated scale parameter. Pearson residuals are the raw residuals divided by the standard deviation of the data, and multiplied by the square root of the scale parameter. Pearson residuals are independent of the scale parameter.

Scaled Pearson chi-square is the sum of the squared scaled residuals and the scaled deviance. The formula to determine the Chi Square from Pearson residuals is:

$$\sum r_p^2 = X^2 .$$

If the deviance is used as a measure of discrepancy of a generalized linear model, then each unit contributes a quantity to the deviance, so that

$$D = \sum_i d_i$$

These summed residuals, when divided by the *df* should equal close to 1. If the goodness of fit value is high or low it is either over or underdispersed. If the goodness of fit is under or overdispersed the standard error is either under or over estimated. The error in estimation makes the confidence intervals and test statistics unusable. Residuals are scaled to correct the standard error, making the confidence intervals and test statistics usable.

The Pearson residual is computed as the raw residual (*y*-m*)*, scaled by the estimated standard deviation of *y*. The Pearson residuals are just rescaled versions of the raw or response residuals and are defined as:

$$r_P = \frac{(y_i - \hat{\mu}_i)}{\sqrt{V(\hat{\mu}_i)}} .$$

The scaled versions of the Pearson and deviance residuals are defined as:

$$r_P{}' = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\phi}\, V(\hat{\mu}_i)(1-h)}} \text{ and } r_D{}' = \frac{r_D}{\sqrt{\hat{\phi}(1-h)}}$$

The *Ø* is the calculated multiplier obtained to scale the Pearson or deviance residuals to one.

Heinzl,H. and M. Mittlbock. 2003. Psuedo R-squared measures for Poisson regression models with over- and under dispersed data. Computational Statistics and Data Analysis. 44:253-271

Piegorsch, W., and A.J., Bailer. 2005.Analyzing Environmental Data. John Wiley and Sons. pp. 112-3
http://stat.ethz.ch/R-manual/R-patched/library/mgcv/html/residuals.gam.html
http://mathstat.carleton.ca/~help/minitab/STREGRSN.pdf
http://www.statsoft.com/textbook/stglz.html
http://www.stats.ox.ac.uk/pub/bdr/IAUL/ModellingLecture5.pdf
http://www.csc.fi/cschelp/sovellukset/stat/sas/sasdoc/sashtml/stat/chap29/sect20.htm#idxgmo0215
http://www.cas.lancs.ac.uk/short_courses/notes/gen_models/session4.pdf

Scale parameter:
In probability theory and statistics, a scale parameter is a special kind of numerical parameter of a parametric family of probability distributions. The value of the scale parameter determines the scale of the probability distribution. If the scale parameter is large, then the distribution will be more spread out; if the scale parameter is small then it will be more concentrated.

Dispersion parameter:
The dispersion parameter is the difference between the variance and the mean of the data. The dispersion parameter is introduced into the generalized linear model to lower the effect of overdispersion.

Using PScale and Dscale:
Using Pscale and Dscale affects the way in which the dispersion parameter is treated. If you specify DScale, the dispersion parameter is estimated by the deviance divided by its degrees of freedom. If you specify PScale, the dispersion parameter is estimated by Pearson's chi-square statistic divided by its degrees of freedom.

Dscale and Pscale fix the scale parameter at the value 1 in the estimation procedure. After the parameter estimates are determined, the exponential family dispersion parameter is assumed to be given by Pearson's chi-square statistic (for Pscale) or deviance (for Dscale) divided by the degrees of freedom, and all statistics such as standard errors and likelihood ratio statistics are adjusted appropriately.

Variance Function:
The variance function describes the relationship between the mean and the variance of the dependent variable. This allows the proper calculation of the variance (and everything that depends on it) under non-normal conditions. The variance function is used in generalized linear models to indicate the dependence of the variance of Y on location and scale parameters.

Variance functions and dispersion parameters for generalized linear models

Distribution	Variance Function V(μ)	Dispersion Parameter Φ
Normal	1	σ^2
Gamma	μ^2	1/α
Inverse Gaussian	μ^3	σ^2
Poisson	μ	1
Binomial	μ *(1- μ)/n	1

http://stat.ethz.ch/R-manual/R-patched/library/mgcv/html/residuals.gam.html
http://mathstat.carleton.ca/~help/minitab/STREGRSN.pdf
http://www.statsoft.com/textbook/stglz.html
http://www.stats.ox.ac.uk/pub/bdr/IAUL/ModellingLecture5.pdf
http://www.csc.fi/cschelp/sovellukset/stat/sas/sasdoc/sashtml/stat/chap29/sect20.htm#idxgmo0215
http://www.cas.lancs.ac.uk/short_courses/notes/gen_models/session4.pdf

3. Binomial, poisson, and non-poisson count data

Poisson data is often the starting point or benchmark for analysis and Poisson analysis is useful for real life count data (non-negative) that is unbounded. It has the property of equidistance, that the mean and the variance must be equal. Unfortunately Poisson data is often overdispersed or underdispersed. This means that the variance is either larger than the mean or smaller than the mean.

> Cameron, A.C., and Trivedi, P.K. 1998 Regression Analysis of Count Data. Cambridge University Press. pp3-5,7.

5 "postulates" of a Poisson model in spatial terms

"1. Start with no event occurrences in the region.
2. Occurrences in disjoint spatial sub-regions are independent
3. The number of occurrences in different sub-spatial regions depends only upon each sub-region's area.
4. Occurrence probability is proportional to spatial area of occurrence.
5. There are no exactly simultaneous occurrences."

> Bailer, J., Piegorsch, W. 1997. Statistics for Environmental Biology and Toxicology, CRC Press. pp 18-19

Binomial count data depends the total number of observations and the probability of success. This can be expressed as 1 or 0 or success or failure. Binomial count data becomes symmetrical as the number of counts rise. Binomial data with the same p will smooth itself out to a symmetrical curve as the count increases, thus becoming normal.

> D., Collett. 2002. Modelling Binary Data. CRC Press.chapter 2

Non-Poisson count data is better suited to analysis via other distributions such as negative binomial .

Negative binomial has a variance that is greater than the mean. The more that the variance differs from the mean the greater the dispersion, making Poisson less appropriate.

White, G.C. and R.E. Bennets. 1996. Analysis of frequency count data using the negative binomial distribution. Ecology. 77(8):2549-2557.

4. Odds and Odds Ratios

Odds

- The probability of success (p) divided by the probability of failure (1-p).

Odds Ratio (OR)

- Ratio of 2 odds and summary of the relationship between 2 variables
- Large OR, then large G, therefore larger difference in proportions
- OR 0-1 success less likely
- OR>1 success more likely
- The greater the OR the greater the association between variables, for example, the number of insects increase 4 times with every metre away from snowmelt.
- With OR only one link, since can only use binomial distribution. Logit link allow for useful interpretation of OR, with an exponential relationship the odd ratio increases multiplicatively with every unit increase
- OR are interchangeable, doesn't matter which side of the equation the response variable is on, the OR will be the same.
- Should fall within confidence limits, if limits include value of 1 then not a useful predictor because that means it is possible there is no change in odds (Agresti 1996; Cohen *et al.*2003; Garson 2005).

Agresti, A. 1996. An Introduction to Categorical Data Analysis. Wiley Series in Probability and Statistics. John Wiley and Sons, Inc., Toronto. Pages 22, 23, 107.

Cohen, J., S.G. West, L. Aiken and P. Cohen. 2003. Applied multiple regression/correlation analysis for the behavioural sciences. Lawrence Erlbaum Associates, Inc. Publishers. Mahwah, New Jersey. (online p. 490-492) http://www2.chass.ncsu.edu/garson/pa765/logistic.htm

Garson, D. Logistic Regression from PA 765: Stat Notes: An online textbook. Quantitative Research in Public Administration. NC State University. http://www2.chass.ncsu.edu/garson/pa765/logistic.htm

Examples:

http://www.cmh.edu/stats/definitions/or.htm

http://www.ats.ucla.edu/stat/sas/faq/oratio.htm

5. Overdispersion

Definition:

Overdispersion is when the variance of the data is greater than the expected variance. Specifically, with the Poisson model; when the variance exceeds the mean the variance is considered to be overdispersed (http://planetmath.org/encyclopedia/Overdispersion.html)

Sources:

Sources of overdispersion are highly random counts. Count data, such as 0,0,1,0,1,0,0,0,25,2,0,1 would cause overdispersion. The high value of the 25 would increase the variance and cause overdispersion.

Why it matters:

While parameter estimates are not altered by overdispersion, standard error is smaller. The smaller standard error creates errors in the confidence intervals and test statistics, making them unusable.

How Diagnosed:

Overdispersion is measured by the goodness of fit. The goodness of fit value is based upon how well the actual goodness of fit matches the *df* divided by the χ^2 with $\alpha = 0.5$. For example, if there are 30 *df* the experimental goodness of fit should be close to 1.02263.

DF	Probability 0.5	X2	DF/X2
1	0.5	0.455	***2.19811***
2	0.5	1.386	***1.4427***
3	0.5	2.366	***1.26798***
4	0.5	3.357	***1.19165***
5	0.5	4.351	***1.14904***
6	0.5	5.348	***1.12189***
7	0.5	6.346	***1.10309***
8	0.5	7.344	***1.08931***
9	0.5	8.343	***1.07877***
10	0.5	9.342	***1.07046***
20	0.5	19.337	***1.03426***
30	0.5	29.336	***1.02263***
40	0.5	39.335	***1.0169***
50	0.5	49.335	***1.01348***
60	0.5	59.335	***1.01121***
70	0.5	69.334	***1.0096***
80	0.5	79.334	***1.00839***
90	0.5	89.334	***1.00745***
100	0.5	99.334	***1.0067***
110	0.5	109.334	***1.00609***
120	0.5	119.334	***1.00558***

How Addressed:
Overdispersion can be addressed in a couple of ways. First, if appropriate, negative binomial can be used instead. It automatically scales down the overdispersion, as the variance is larger than the mean with negative binomial. The second method is scaling the overdispersion so it equals 1. Depending on *df* either the Pearson chi-square or the G-statistic can be scaled to 1. This corrects the small standard error and makes the confidence intervals and test statistics valid.

Heinzl,H. and M. Mittlbock. 2003. Psuedo R-squared measures for Poisson regression models with over- and under dispersed data. Computational Statistics and Data Analysis. 44:253-271.

http://www.id.unizh.ch/software/unix/statmath/sas/sasdoc/stat/chap29/sect27.htm#idxgmo0392
http://www.ats.ucla.edu/stat/sas/library/genmod.pdf
http://www.statsoft.com/textbook/gloso.html
http://www.uky.edu/ComputingCenter/SSTARS/P_NB_3.htm

6. Loglinear Models

Loglinear models are used for Poisson count data in the GzLM. An example of a loglinear model is $\mathrm{Ln}(F_{ij}) = m + l_i^A + l_j^B + l_{ij}^{AB}$. Loglinear models are used for categorical data.

$\mathrm{Ln}(F_{ij})$ = is the log of the expected cell frequency of the cases for cell ij in the contingency table.
μ = is the overall mean of the natural log of the expected frequencies
λ = terms each represent "effects" which the variables have on the cell frequencies
A and B = the variables
i and j = refer to the categories within the variables

Therefore:
λ_i^A = the main effect for variable A
λ_j^B = the main effect for variable B
λ_{ij}^{AB} = the interaction effect for variables A and B

This is a saturated model because all of the one-way and two way effects are included. If the model was rewritten, $\mathrm{Ln}(F_{ij}) = m + l_i^A + l_j^B$, we are assuming that the effects of the A and B are independent. This is called an independent model.

To test for fit of the loglinear model, the Pearson Chi-Squared and the log likelihood-ratio statistic are computed for goodness of fit. They compare the cell fitted values against the observed counts.

The larger the G^2 and the χ^2 and the smaller p-value, indicating a poor goodness of fit.

Example (http://www.socialresearchmethods.net/tutorial/Cho/logistic.htm)

For instance, we are interested in the relationship between smoking and lung cancer. The explanatory variable is whether to smoke (smoking or nonsmoking group), and the response variable is whether to have lung cancer. In this case, we have the 2 * 2 case-control design, because we have two levels in explanatory variables (smoking / nonsmoking) and two responses in response variables (cancer / no cancer). If we are also interested in the role of age, we can add "age" as continuous or categorical data. It will be easier to start with the data matrix we can have in either case.

If the age is the continuous explanatory variable, the data matrix looks like the following table. It is an ungrouped data set.

Age (Continuous)	Smoking (Yes=1/ No=0)	Cancer (Yes=1/ No=0)
36	1	0
47	0	1
49	1	0
29	1	1
60	0	1
55	1	1
65	1	0
38	1	1
56	0	1

On the other hand, if the age variable is categorized into three age groups, under 40, 41-60, over 61, we have three age group and the age variable is the categorical variable. In this case, it is possible to count the number of people in each cell of the contingency table. The following table summarizes the results of all three categorical variables. It is a grouped data set.

		Lung	Cancer
Age Group	Smoking	Yes	No
Under 40	Smoking	15	4
41~60	Smoking	30	7
Over 60	Smoking	26	6
Under 40	No Smoking	8	2
41~60	No Smoking	14	2
Over 60	No Smoking	15	3

We call it the 2(Smoking)* 2(Lung Cancer)* 3(Age Groups) contingency table, because we have two levels of smoking, two levels of cancer, and three levels of age groups.

The logistic regression model tests whether smoking has an effect on lung cancer and whether the age effect on lung cancer exists and whether there is an interaction between smoking and age group and tries to find the best model which can predict the chance of lung cancer with the smoking and age variables.

In short, the logistic regression model is useful when the study is interested in the relationship between the categorical response variable and the categorical and/or continuous explanatory variables.

Agresti, A. 1996. *An Introduction to Categorical Data Analysis*. John Wiley, New York.pp 145-55.

Jeansonne, A. 2002. Loglinear Models.
(http://userwww.sfsu.edu/~efc/classes/biol710/loglinear/Log%20Linear%20Models.htm)
http://www.math.yorku.ca/SCS/Courses/grcat/grc8.html
http://www2.chass.ncsu.edu/garson/pa765/logit.htm

7. Saturated Model

A model with as many parameters as it has observations, giving the models perfect fit. The model has the maximum number of parameters possible. Such a model is sometimes useful as it serves as a benchmark to quantify how well a simpler model (one with fewer parameters) fits the data. A saturated model by itself isn't biologically very useful; the whole intent of a "model" is to achieve some synthetic simplification of set of observations (data points), whereas a saturated model isn't a simplification and provides no greater interpretability than just looking at the raw data points themselves.

Agresti, A. 1996. An Introduction to Categorical Data Analysis. Wiley Series in Probability and Statistics. John Wiley and Sons, Inc., Toronto. Pages 22, 23, 107.

Dobson, A.J. 2002. An Introduction to Generalised Linear Models. Chapman and Hall/CRC Press. (Online book)

Elston, R.C., J.M. Olson and L. Palmer. 2002. Biostatistical Genetics and Genetic Epidemiology. John Wiley and Sons Canada Ltd. Etobicoke, ON. (Online pages 315)

Lindsey, J.K. Applying Generalized Linear Models. Springer-Verlag, New York. P.23-24 (online book)

http://tecfa.unige.ch/~lemay/thesis/THX-Doctorat/node236.html
http://www.warnercnr.colostate.edu/~gwhite/mark/markhelp/saturatedmodel.htm
http://planetmath.org/encyclopedia/SaturatedModel.html
http://www.absc.usgs.gov/staff/WTEB/jschmutz/joel/snowfall_saturated.htm
http://www2.chass.ncsu.edu/garson/pa765/semAMOS1.htm
http://www.reference.com/browse/wiki/Saturated_model

8. Goodness of Fit

The Pearson chi-squared test is described by this formula:

$$\chi^2 = \sum\left(\frac{(O-E)^2}{E}\right)$$

where O is observed frequency and E is expected frequency. If the expected chi-square, (eg. χ^2= 2.5), is lower than the χ^2 obtained from the table based on the *df* and α, then the fit was good.

The Pearson chi-squared test is best used when the there is a larger sample size (n>9), three of more classes and all expected values are larger than 0.25.

Lee, C. F. 1998. Statistics for Business and Financial Economics, 2nd Ed World Scientific pp. 509-510.
http://www.statsdirect.com/help/chi_square_tests/chi_good.htm

The log likelihood ratio (G – Statistic) is calculated by this formula:

$$G = 2\sum f_i \ln \frac{f_i}{\hat{f}_i}$$

The G – Statistic (also called the deviance) is a better fit than the χ^2 when the sample sizes are smaller. Again, the G-statistic is used to test whether the data fits the error distribution (Poisson, negative binomial).

B. Gerstman. 2003. Epidemiology Kept Simple. Wiley. pp344-6.

The Wald statistic is z = β/ASE. The z value is compared to the β = 0, where 0 is standard normal. The second use squares the z value. This squared z can be compared to χ^2 table with a *df* = 1. The Wald statistics are widely used because it is easy to compute but there is evidence that they are not as reliable as the Pearson chi-squared or G-statistic.

Agresti, A. 1996. *An Introduction to Categorical Data Analysis*. John Wiley, New York. pp. 88-89.

MacKinnon, J.G., and R. Davidson. 2003. Econometric Theory and Methods. Oxford University Press. p. 422.

http://www.ats.ucla.edu/stat/sas/library/genmod.pdf
http://www.id.unizh.ch/software/unix/statmath/sas/sasdoc/stat/chap29/sect27.htm
http://en.wikipedia.org/wiki/Goodness_of_fit
http://www.personal.rdg.ac.uk/~snscolet/MScGLMs/Lecture5.pdf
http://www.uky.edu/ComputingCenter/SSTARS/P_NB_3.htm

9. Analysis of Deviance

For generalized linear models the terms in the model will in general no longer be orthogonal and also, sums of squares will for non-Normal distribution no longer appropriate measures of the contribution of a term to the total discrepancy. The analysis of deviance (AnoDev) table shows the deviance of the data from the model, for a sequence of models. It also shows the change in deviance (ΔG improvement in fit) due to each term in the model.

The Analysis of Deviance table summarizes information about the sources of variation in the response for the set of data. The ANODEV reports change in fit due to each model term. The G statistic is the fit of the model to the intercept, and replaces the sequential sums of squares seen in ANOVA. The ΔG is the change in fit associated with each term in the model.

Example of an ANODEV table:

LR Statistics For Type 1 Analysis

Source	2*LogLikelihood (G)	DF	Chi-Square(ΔG)	Pr > ChiSq
Intercept	557.5548			
YEAR	561.7104	5	4.16	0.5272
SURVEY	562.8268	5	1.12	0.9526
TRANSECT	746.1630	15	183.34	<.0001
YEAR*SURVEY	769.0660	8	22.90	0.0035
YEAR*TRANSECT	1152.0339	71	382.97	<.0001
SURVEY*TRANSECT	1353.5831	66	201.55	<.0001

http://www.csc.fi/cschelp/sovellukset/stat/sas/sasdoc/sashtml/insight/chap39/sect23.htm#idxfit0465
http://www.warnercnr.colostate.edu/~gwhite/mark/markhelp/anodev.htm
http://www.csc.fi/cschelp/sovellukset/stat/sas/sasdoc/sashtml/insight/chap16/sect4.htm#idxlog0030

10. Maximum Likelihood

- Estimates what the sample should likely be (likelihood) or as typical as possible given the observed values.
- Likelihood is calculated repeatedly as iterations until the iterations no longer differ greatly (or the algorithm converges).
- Predicts how likely the observed values of the response variable can be predicted from the observed values of the explanatory variables (Agresti 1996; Cohen *et al.*2003; Garson 2005; Johnston 2005).

Maximum Likelihood Method.

- The method of maximum likelihood (the term first used by Fisher, 1922a) is a general method of estimating parameters of a population by values that maximize the likelihood (L) of a sample. The likelihood L of a sample of n observations x1, x2, ..., xn, is the joint probability function p(x1, x2, ..., xn) when x1, x2, ..., xn are discrete random variables. If x1, x2, ..., xn are continuous random variables, then the likelihood L of a sample of n observations, x1, x2, ..., xn, is the joint density function f(x1, x2, ..., xn).
- Let L be the likelihood of a sample, where L is a function of the parameters θ1, θ2, ... θk. Then the maximum likelihood estimators of θ1, θ2, ... θk are the values of θ1, θ2, ... θk that maximize L.
- Let θbe an element of Ω. If Ω is an open interval, and if L(θ) is differentiable and assumes a maximum on W, then the MLE will be a solution of the following equation: (dL(θ))/dθ = 0 (http://www.statsoft.com/textbook/stathome.html click on generalized linear model: Maximum Likelihood method.

Agresti, A. 1996. An Introduction to Categorical Data Analysis. Wiley Series in Probability and Statistics. John Wiley and Sons, Inc., Toronto. Pages 8-10, 96.

Garson, D. Logistic Regression from PA 765: Stat Notes: An online textbook. Quantitative Research in Public Administration. NC State University. http://www2.chass.ncsu.edu/garson/pa765/logistic.htm

Cohen, J., S.G. West, L. Aiken and P. Cohen. 2003. Applied multiple regression/correlation analysis for the behavioural sciences. Lawrence Erlbaum Associates, Inc. Publishers. Mahwah, New Jersey. (online p. 498)

Johnston, G. SAS Software to Fit the Generalized Linear Model. SAS Institute Inc., Cary, NC http://www.ats.ucla.edu/stat/sas/library/genmod.pdf pages 1-8

11.Scaled Residuals

The two basic types of residuals are the so-called Pearson residuals and deviance residuals. Pearson residuals are based on the difference between observed responses and the predicted values; deviance residuals are based on the contribution of the observed responses to the log-likelihood statistic. It is possible to scale for the Pearson or the deviance residuals.

Scaled Pearson chi-square is the sum of the squared scaled residuals and the scaled deviance. The formula to determine the Chi Square from Pearson residuals is:

$$\sum r_p^2 = X^2 .$$

If the deviance is used as a measure of discrepancy of a generalized linear model, then each unit contributes a quantity to the deviance, so that

$$D = \sum_i d_i$$

These summed residuals, when divided by the *df* should equal close to 1. If the goodness of fit value is high or low it is either over or underdispersed. If the goodness of fit is under or overdispersed the standard error is either under or over estimated. The error in estimation makes the confidence intervals and test statistics unusable. Residuals are scaled to correct the standard error, making the confidence intervals and test statistics usable.

The Pearson residual is computed as the raw residual (*y*-m*),* scaled by the estimated standard deviation of *y*. The Pearson residuals are just rescaled versions of the raw or response residuals and are defined as:

$$r_P = \frac{(y_i - \hat{\mu}_i)}{\sqrt{V(\hat{\mu}_i)}}.$$

The scaled versions of the Pearson and deviance residuals are defined as:

$$r_P{}' = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\phi}\, V(\hat{\mu}_i)(1-h)}} \quad \text{and} \quad r_D{}' = \frac{r_D}{\sqrt{\hat{\phi}(1-h)}}$$

The Ø is the calculated multiplier obtained to scale the Pearson or deviance residuals to one.

Heinzl,H. and M. Mittlbock. 2003. Psuedo R-squared measures for Poisson regression models with over- and under dispersed data. Computational Statistics and Data Analysis. 44:253-271

Piegorsch, W., and A.J., Bailer. 2005.Analyzing Environmental Data. John Wiley and Sons. pp. 112-3

http://stat.ethz.ch/R-manual/R-patched/library/mgcv/html/residuals.gam.html
http://mathstat.carleton.ca/~help/minitab/STREGRSN.pdf
http://www.statsoft.com/textbook/stglz.html
http://www.stats.ox.ac.uk/pub/bdr/IAUL/ModellingLecture5.pdf
http://www.csc.fi/cschelp/sovellukset/stat/sas/sasdoc/sashtml/stat/chap29/sect20.htm#idxgmo0215
http://www.cas.lancs.ac.uk/short_courses/notes/gen_models/session4.pdf

12. **Link**

The *link function* in *generalized linear models* specifies a nonlinear transformation of the *predicted* values so that the distribution of predicted values is one of several special members of the exponential family of distributions (e.g., gamma, Possion, binomial, etc.). The *link function* is therefore used to model responses when a dependent variable is assumed to be nonlinearly related to the predictors. The link function serves to link the random or stochastic component of the model, the probability distribution of the response variable, to the systematic component of the model (the linear predictor).

Formulas for common link functions

Link	**Formula**
Identity	M
Log	log μ
Inverse	1/
Square Root	$\sqrt{}\ \mu$
Logit	Log $\mu/1-\mu$
Probit	$\Phi^{-1}(\mu)$
Complementary log-log	Log(-log(1- μ))
Power	μ^k
Arcsine	Sin-1(2 μ-1) $0 \leq \mu \leq 1$
Box-Cox	$(\mu^{\lambda}-1)/\lambda$

Each distribution has a most commonly used link, called the canonical link.

Canonical link functions associated with common probability distributions

Probability Distribution	**Canonical Link Function**
Normal	Identity
Binomial	Logit
Poisson	Log
Gamma	Reciprocal

http://www.stat.uiowa.edu/~luke/xls/glim/glim/node7.html
http://userwww.sfsu.edu/~efc/classes/biol710/Glz/Generalized%20Linear%20Models.htm
http://www.statsoft.com/textbook/glosl.html
http://www.warnercnr.colostate.edu/~gwhite/mark/markhelp/linkfunctions.htm

Atkinson, A. and Riani, M. 2000. Robust Diagnositc Regression Analysis. Springer, New York.

13. Error Distibutions

The generalized linear model differs from the general linear model in two major aspects

1. The distribution of the dependent (or response) variables can be (explicitly) non-normal, and does not have to be continuous (ie may be binomial).

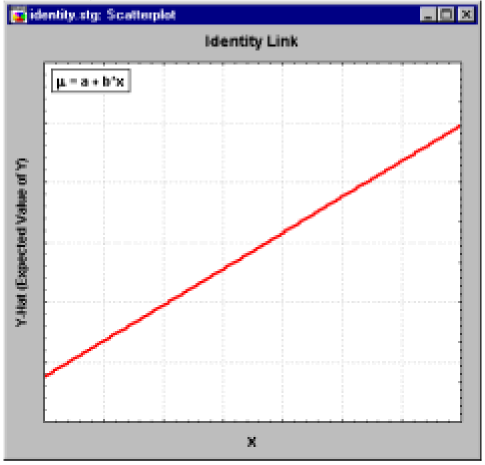
2. The dependent variable values are predicted from a linear combination of predictor variables, which are ‘connected’ to the dependent variable via a link function.

http://www.statsoftinc.com/textbook/stglz.html

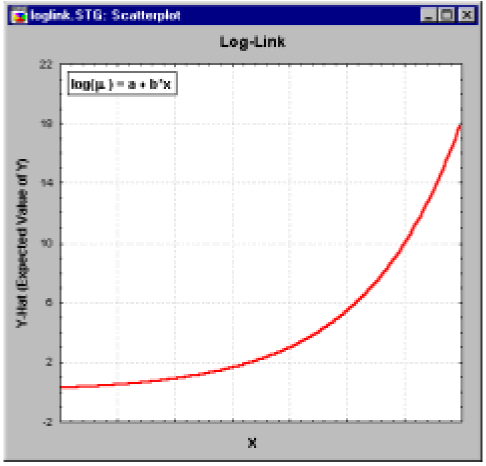
Distributions: Generalized linear models encompass the general linear model and enlarge the class of linear least-squares models in two ways: the distribution of *Y* for fixed *x* is merely assumed to be from the exponential family of distributions, which includes important distributions such as the binomial, Poisson, exponential, and gamma distributions, in addition to the normal distribution (http://userwww.sfsu.edu/~efc/classes/biol710/Glz/Generalized%20Linear%20Models.htm)

1. Normal, Gamma and Poisson Distribution:

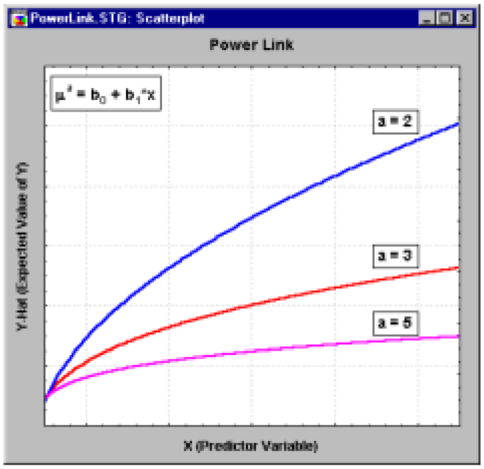
Identity link f(z)=z



Log link f(z) = log (z)



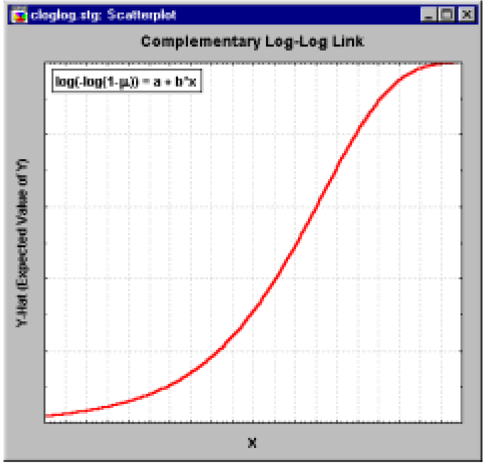
Power link f(z) = z^a, for a given a



http://www.statsoftinc.com/textbook/stglz.html

2. Binomial distribution:

Logit link



StatSoft, Inc. 1984-2003. Electronic textbook Statsoft. STATISTICA, StatSoft, Inc. http://www.statsoftinc.com/textbook/stglz.html

Deviance:
The table below displays deviance for each of the probability distributions available in PROC GENMOD, an estimate of the scaled deviance and Pearson's chi-square statistic which are then divided by the degrees of freedom for the model. The deviance is used for the goodness of fit of the model.

Distribution	Deviance
Normal	$\sum_i w_i (y_i - \mu_i)^2$

Poisson	$2\sum_i w_i \left[y_i \log\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i)\right]$
Binomial	$2\sum_i w_i m_i \left[y_i \log\left(\frac{y_i}{\mu_i}\right) + (1 - y_i)\log\left(\frac{1-y_i}{1-\mu_i}\right)\right]$
Inverse Gaussian	$\sum_i \frac{w_i(y_i - \mu_i)^2}{\mu_i^2 y_i}$
Gamma	$2\sum_i w_i \left[-\log\left(\frac{y_i}{\mu_i}\right) + \frac{y_i - \mu_i}{\mu_i}\right]$
Multinomial	$\sum_i \sum_j w_i y_{ij} \log\left(\frac{y_{ij}}{p_{ij} m_i}\right)$
Negative Binomial	$\sum_i \sum_j w_i y_{ij} \log\left(\frac{y_{ij}}{p_{ij} m_i}\right)$

SAS Institute Inc., SAS OnlineDoc®, Version 8, Cary, NC: SAS Institute Inc., 1999.
http://www.id.unizh.ch/software/unix/statmath/sas/sasdoc/stat/chap29/sect27.htm

Canonical Link:

The link function serves to link the random or stochastic component of the model, the probability distribution of the response variable, to the systematic component of the model (the linear predictor):

$$E(Y) = g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j \ , \qquad \textbf{(2)}$$

Where g(μ) is a non-linear link function that links the random component, $E(Y)$, to the systematic component $(\beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j)$. For traditional linear models in which the random component consists of the assumption that the response variable follows the Normal distribution, the canonical link function is the identity link. The identity link specifies that the expected mean of the response variable is identical to the linear predictor, rather than to a non-linear function of the linear predictor. The canonical link functions for a variety of probability distribution are given below.

Probablility Distribution Type	**Canonical Link Function**
Normal	**Identity**
Poisson	Log
Binomial	Logit
Inverse Gaussian	Power^{-2}
Gamma	Power^{-1}
Negative Binomial	Log

Agresti, A. 1996. *An Introduction to Categorical Data Analysis*. John Wiley, New York.

Johnston, G. SAS Software to Fit the Generalized Linear Model. SAS Institute Inc., Cary, NC http://www.ats.ucla.edu/stat/sas/library/genmod.pdf pages 1-8

14. Parameter Estimation

OLS: Ordinary least squares parameters estimates are created with this equation:

$$Yi = a0 + a1\ X1i + a2\ X2i + ... + ak\ Xki + ei,$$

where Y_i is related to each accompanying X_i variable.
Ordinary least squares (OLS) parameter estimates are accurate and unbiased if the data follows these rules provides by Dr. D.J.C. Smant:

1. There is no correlation between explanatory variables and residuals (no simultaneity), i.e. cov(Xji,ei) = 0. Failure of this assumption results in biased estimates of the coefficients of explanatory variables.

2. The expected or mean value of the residuals equals zero, i.e. E(ei) = 0. Failure of this assumption results in a biased estimate of the constant term.

3. Residuals are homoskedastic (no heteroskedasticity = no cones), i.e. E(e i2) = s2 = constant.Failure of this assumption results in inefficient estimates and biased tests of hypotheses.

4.Residuals are independently distributed (no serial correlation), i.e. E(eiej) = 0. Failure of this assumption results in inefficient estimates and biased tests of hypotheses.

5. Explanatory variables are independent (no multicollinearity), i.e. cov(Xi,Xj) = 0. Failure of this assumption results in inefficient estimates and biased tests of hypotheses.

In addition to these well-known standard assumptions we also have:

6. Residuals are normally distributed, i.e. e ~ N(0, s2) (combining assumptions 2 and 3 and 5). Failure of this assumption invalidates the use of the Student t-distribution in coefficient t-tests.

7. Explanatory variables are measured without error (no errors in variables). Failure of this assumption results in biased estimates of the coefficients.

8. Variables that are time series must be stationary (no unit roots), i.e.well-defined mean andvariance. Failure of this assumption results in spurious regressions (except in the special case of cointegration).

The narrow constraints of the OLS make it unsuitable for most biological situations.
http://www.few.eur.nl/few/people/smant/econometrics/intro_pr_ectr_2.pdf
http://econ.la.psu.edu/~hbierens/EasyRegTours/OLS.HTM

WLS: Weighted Least Squares is a popular method for estimating parameters and has been around since the late 1700’s. It deals with heteroskedasticity (cones) much better than OLS by weighting each variable. It uses the same equation as the OLS but there is a w_i beside each term deciding how much influence each variable should have:

$$Yi = a0 + a1\, w_i(X1i) + a2\, w_i(X2i) + ... + ak\, w_i(Xki) + ei,$$

WLS is also not very useful for biological applications. It is easy to put bias into the model from the weighting.

> http://www.utdallas.edu/~herve/Abdi-LeastSquares-pretty.pdf
> Schabenberger, O., and F.J., Pierce. 2001. Contemporary Statistical Models for the Plant and Soil Sciences. CRC Press. pp. 592.

Iteratively Reweighted Least Squares (IRLS): The OLS and WLS involves a variance as a function of the value of the observed data point. It also ties the weight to the variance of the observed data points. However, when a weighting scheme is applied to a series of data points, data points with very low values may be given more emphasis than is appropriate. An alternative weighting has been used to try to overcome this disadvantage. This method is the iteratively reweighted least squares method. Effectively, this is identical to the WLS except that observed data is replaced with calculated data. Thus, the weight is recalculated during each phase of the optimization process. Thus very low observed data points would not have the emphasis on the overall analysis. However, it is possible that the optimization may drive calculated values low, giving these points more emphasis and potentially distorting the final analysis. (http://www.boomer.org/c/p3/c13/c1306.html)

> Garthwaite, P.H., Jolliffe, I.T., and Jones. B. 2002. Statistical Inference.Oxford University Press. p. 63.
>
> Gibbons, R.D. and D.E.Coleman. 2001 Statistical Methods for Detection and Quantification of Environmental Contamination. Wiley. pp. 39-40.

http://userwww.sfsu.edu/~efc/classes/biol710/Glz/Generalized%20Linear%20Models.htm
http://www.itl.nist.gov/div898/handbook/pmd/section4/pmd432.htm

15. Model diagnosis

- Residuals (error) are analyzed in linear modeling to identify poorly fitted values. If there a large number of poorly fitted values exist then often the fit is determined to be inappropriate for the data. Residuals are also used for: looking for signs of nonlinearity, evaluating the effect of new explanatory variables, creating goodness of fit stats and evaluating leverage and influence for individual data points (Gill 2000)

- If you are doing a regression the model does require linearity, ie. no bowls or arches in the residuals versus fits. But if you are not using a regression then GzLM doesn't require linear relationship (Garson 2005)

- Even though the residuals in the GzLM we would like evenly distributed around zero (Gill 2000)

- We look at the residuals plot to investigate potential outliers and other interesting behavior of the data in using that particular model. Residual versus fits looking for striping to indicate zero's in data, cones for model fit and homogeneity, bowls if using a regression analysis

- We look at the deviance residuals to describe the stochastic behavior of the data relative to a constructed GzLM in a format that closely resembles the normal theory analysis of standard linear model residuals.

Agresti, A. 1996. An Introduction to Categorical Data Analysis. Wiley Series in Probability and Statistics. John Wiley and Sons, Inc., Toronto. Pages 109, 88-91

Garson, D. Logistic Regression from PA 765: Stat Notes: An online textbook. Quantitative Research in Public Administration. NC State University. http://www2.chass.ncsu.edu/garson/pa765/logistic.htm

Gill, J. 2000. Generalized Linear Models: a unified approach. Sage university Papers Series on Quantitative Applications in the Social Sciences. 07-134. Thousand Oaks, CA: Sage. Pages 51-66.

Johnston, G. SAS Software to Fit the Generalized Linear Model. SAS Institute Inc., Cary, NC http://www.ats.ucla.edu/stat/sas/library/genmod.pdf pages 1-8