Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 16 Introduction

ReCap. Parts I – IV. The General Linear Model
Part V. The Generalized Linear Model
16 Overview
Advantages
References
Texts
Generic recipe

ReCap Part I (Chapters 1,2,3,4) Quantitative reasoning (Models and Measurement): Example of scallops, which combined models (what is the relation of scallop density to substrate?) with statistics (how certain can we be?)

ReCap Part II (Chapters 5,6,7)

Estimation: Inference from data to a parameter value.

Likelihood ratios: A measure of the strength of the evidence.

Hypothesis testing: Inference from data to finite or infinite populations

ReCap Part III The General Linear Model.

Single explanatory variable: Regression and ANOVA

Multiple explanatory variables: Multiple regression, multiway ANOVA, ANCOVA

GLM a general procedure that is more flexible and useful than a collection of named tests.

Today: Overview of the Generalized Linear Model

Introduction to the generalized linear model.

In the previous sections we learned to write, execute, and interpret statistical models for linear model with normal (fixed) errors. In this section we extend what we have learned to linear models with errors that are not homogeneous. One important application will be count data, which is bounded at zero, and will show non-homogeneous errors because the variance rises with the mean. Count data were treated in 20th texts as separate topics, including logistic regression, contingency tests, multiway tables, analysis of frequencies, and log-linear models. The conceptual framework that extends the general linear model to these topics and other sources of non-normal errors was developed by McCullagh and Nelder (1972). The computational machinery was developed at the same time, but was daunting and limited in its earliest form (GLIM). In this century the necessary statistical software is now widely available in both code based (SAS, R, Minitab, MatLab) and pull-down menu form (SPSS, Minitab).

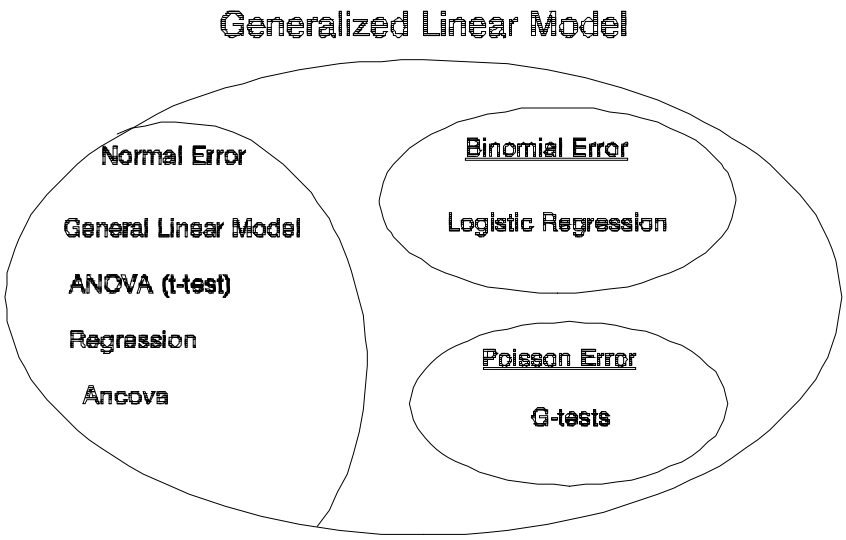
Observational data often do not meet the normal error assumptions for the General Linear Model. Count data often result in heterogeneous residuals. Data where extreme counts in the right tail (above the mean) similarly result in heterogeneous residuals. Data bounded at both 0 and 1, such as percent cover of an area, often result in heterogeneous residuals, with reduced dispersion near 0 and 1, resulting in a spindle shaped residual vs fit plot.

Transformations are traditionally recommended to remedy these problems.
This remedy sometimes cures the statistical problems, but the side effects are unhealthy. They include uninterpretable models of the relation of the response to explanatory variable (Warton and Hui 2011) and biased estimates of effect sizes (Packard 2009, Ballentyne 2013, St.-Pierre et al 2018).

We have already seen one remedy, a randomization test. With this remedy we don't have to maltreat the data with a transformation. An inevitable side effect of this remedy is that our estimates of parameters (means, slopes, *etc*) are still made with a normal error model. As a result a few large counts, as from a process that generates right skewed data, will have an undue influence on our estimate of the mean.

The Generalized Linear model (GzLM) leaves the response variable on its original scale. In science, analysis of data on its original scale, retaining units, is more appropriate than some arbitrary and often uninterpretable transformation chosen for statistical reasons.

The Generalized Linear model was developed in 1972 (Nelder and Wedderburn) and presented in text form a decade later (McCullagh and Nelder 1983). The GzLM includes logistic regression (*e.g.* Menard 1993) and log linear models (Feinberg 1970, Bishop et al 1975, Agresti 1996) as special cases. Introductory texts in biology (*e.g.* Sokal and Rohlf 2012, Zar 2010) treat the analysis of count data under the heading of analysis of frequencies.



The analysis of count data is sometimes presented as 'non-parametric' tests, and hence free of assumptions (Siegel, S. 1956. *Nonparametric Statistics for the Behavioral Sciences*. New York, NY, US: McGraw-Hill). The analysis of count data employs the chi-squared distribution, which rests on the same assumptions as its daughters the *F* and *t*-distributions (Feller, W. 1971 *An Introduction to Probability Theory and its Applications.* Volume II, 2nd edition). While the text is long-gone the notion regularly surfaces in online advice forums (Quora, ResearchGate).

```
                                                                    rubrics
2003 Initial presentation GzLM - Problems

Contrast of GzLM and GLM not clearly stated
Notation of GzLM not clear
Concepts of error model, structural model, and link
   not clearly distinguished.
Links introduced (in quizzes) without explanation.
No hands-on examples in lab.

For binomial response
   Proportion, odds, logOdds, Odds ratio
   not clearly distinguished.

Distinction of binomial and poisson response not clearly distinguished in
lecture notes.
Clearly distinguished in class in 2003 by defining the unit:
   Score each unit as present/absent -> binomial
   Count in a defined unit —> poisson.

Too many concepts in the initial GzLM lecture.

2005-2017 GzLM reduced to writing the model only, no execution
Odds and odds ratio calculation to lectures and quizzes after 2008
2017 Likelihood to Lab 3
2018 Goodness of fit (Ch16.2) moved early in the course (Ch7.6)
2018 LR calculations and interpretation to generic recipe and quizzes
```

Advantages of GzLM.

Learning the GzLM has many advantages, compared to learning special cases such as logistic regression, log-linear models, *etc*.

Carryover. Concepts already learned apply. For example a contingency test has the same structural model as a two way ANOVA.

Improved quality of statistical analyses. Likelihood ratios and p-values are more reliable if based on an appropriate error structure. Parameter estimates (means, ratios, odds) are more accurate if based on an appropriate error structure.

Transformations become unnecessary. Transforming the response variable does not necessarily remedy failure to the meet assumptions of a normal error model. Transformation do change the relation of the response to explanatory variable, sometimes in uninterpretable ways (square root transform, arcsin transform). With the generalized linear model the response variable remains on its original scale. Effect sizes can be interpreted in the same units as the measurement of the response variable. Exponential, power law, and inverse functions (*e.g*. Michaelis-Menten) can be fit assuming additive errors on the measurements scale, not on the transformed scale.

Tacit assumptions disappear. The latter half of the 20th century saw a proliferation of methods for analysis as if transformation produce normal residuals. This assumption was rarely stated and almost never examined. The generalized linear model eliminates the assumption. The error model is explicit.

Greater flexibility in analysis. The generalized linear model provides many error structures, freeing us of shoehorning every statistical analysis into a normal error structre.

References

Ballentyne, F. 2013. *Journal of Theoretical Biology* 317: 418–421

Nelder, J.A., Wedderburn, R.W.M. 1972. Generalized Linear Models. J*ournal of the Royal Statistical Society* A 135: 370-84.
The publication that introduced the GzLM

Packard , J. 2009 . *Journal of Theoretical Biology* 257, 515–518: 317: 418–421)

St.Pierre, A.P. V. Shikon, D.C. Schneider. Count data in biology—Data transformation or model reformation? *Ecology and Evolution.* 2018;1–9.

Sokal, R.R. and F.J. Rohlf. 2012. Biometry. Freeman.

Zar, J.H. 2010. Biostatistical Analysis. Prentice-Hall.

Warton, D.I. and F. K. C. Hui.2011. *Ecology* 92:3–10.

Texts

The Generalized Linear Model (Nelder and Wedderburn 1972) is one of the most important contributions to statistics in the last half of the 20th century. Texts range from accessible to highly mathematical. Here is a lightly annotated list.

Agresti, A. 1996, 2007 *Introduction to Categorical Data Analysis*. NY: John Wiley and Sons.
Presents multiway tables and GzLM, examples from social science and biology.

Bishop, Y.M.M., Feinberg, S.E., Holland. 1975. *Discrete Multvariate Analysis*. MIT Press.
Extensive treatment of multiway tables, examples from social, health, and biological sciences.

Crawley, M.J. 1993. *GLIM for Ecologists.* Blackwell.
Stresses parameter estimation and model evaluation instead of hypothesis testing.
The GLIM programming language has been replaced by R/S+

Dobson, A.J. 1990. *An Introduction to Generalized Linear Models.* Chapman and Hall.

Feinberg, S.E. 1977, 1983. *The Analysis of Cross-Classified Categorical Data.* MIT Press.
Early treatment of log-linear models (Poisson error, log link) as multiway tables.
Examples mostly social science.

Feller, W. 1971 *An Introduction to Probability Theory and its Applications.* Volume II, 2nd edition

Hardin, J.W. and J.M. Hilbe. 2012. *Generalized Linear Models and Extensions*. Stata Press, College Station, Texas.
Treatment from the point of view of liklihood. Stat code throughout.

Hoffmann, J.P. *Generalized Linear Models. An Applied Approach*.
Pearson Education Inc (Allyn and Bacon).
Highly readable, includes SPSS, SAS and Stat code. Examples mostly from social science.

Menard, S. 1993. *Applied Logistic Regression Analysis*. London: Sage Publications.

Lindsey, J.K. *Applying Generalized Linear Models*. NY: Springer Texts in Statistics.
Extensive data sets and exercises from biology, social science, health science, engineering.

Madsen, H. and Thyregood, P. 2011. Introduction to General and Generalized Linear Models. CRC Press.
Emphasis on math, circa 40-100 numbered equations per chapter. Extensive R code.
Question guided analysis of 5 data situations in Chapter 7.

McCullagh, P. and J.A. Nelder. 1983 (1st edition) 1989 (2nd edition). Generalized Linear Models. Chapman and Hall.
The 2nd edition remains the authoritative text. Each application is illustrated by a simple data set.

Myers, R.H. Montgomery, D.C. Vining, G.G. 2002. Generalized Linear Models with Applications in Engineering and Sciences. Wiley.
Extensive use of math, many data sets and exercises, mostly from engineering.

Smithson, M., Merkle, E.C. 2014. Generalized LinearModels for Categorical and Continuous Limited Dependent Variables. CRC Press.
Data files available online, mostly social science. *Ca* 7 exercises per chapter.
Detailed description of complex analyses, in R and Stata.

In order to apply the generalized linear model, we need to make a few modifications of the generic recipe for applying the general linear model.

Table 16.1 Generic Recipe for Statistical Inference with the Generalized Linear Model.

Introduction. Data set with context and goals of analysis.

1. Construct model. Begin with verbal and graphical model.
   - Distinguish response from explanatory variables
   - Assign symbols, state units and type of measurement scale for each.
   - Make preliminary choice of error model.
   - Write out statistical model.
2. Execute model — Place data in model format, code model statement.
   - Compute fitted values from parameter estimates.
   - Compute residuals and plot against fitted values.
3. Evaluate the model, using residuals.
   - If fitted line inappropriate, revise the model (back to step 1).
   - If errors not homogeneous, revise error model (step 1).
   - If heterogeneity due to influential outliers, revise error model or link function.
   - Residuals independent ? (plot residuals versus residuals at lag 1)
     - If not add term to capture non-independence (often spatial or temporal).
   - If using chisquare, t, or F distribution to estimate Type I error, check normality.
     - Evaluate residuals with histogram and quantile or normal score plot.
   - If not met, check p-value with empirical distribution (by randomization).
4. Report evidence. Calculate omnibus likelihood ratio from ANOVA or ANODEV table.
   - If negligible, then skip to step 10.
5. Choose mode of inference: evidentialist, frequentist, priorist.
   - Priorist: Give probable cause for prior distribution.
   - Frequentist: What is the target of inference?
     - Hypothesis testing? If so, state test statistic, its distribution (t or F).
     - Fixed Type I error required? If so, state α.
6. Statistical analysis.
   - If evidentialist, report *LR* for fixed terms, then Step 7.
   - If frequentist Table Source, *df, SS*. Calculate t or F from *MS*.
     - If normal error, p from probability model. Otherwise, p by randomization.
     - If fixed α (Type I error) calculate confidence limits as appropriate, then Step 7.
     - If α not fixed, report p by category, then Step7.
   - If priorist, state and justify prior probability. Compute posterior probability, then to Step7.
7. Report science conclusions. Interpret parameters of biological interest (means, slopes, odds ratios) along with one measure of uncertainty (R^2, st. error, st. dev., or confidence intervals) or posterior probability.