York SPIDA

John Fox

Notes

Generalized Linear Models

1. Topics

- The structure of generalized linear models
- Poisson and other generalized linear models for count data
- Diagnostics for generalized linear models (as time permits)
- Logit and Loglinear models for contingency tables (as time permits)
- Implementation of generalized linear models in R

2. The Structure of Generalized Linear Models

- A synthesis due to Nelder and Wedderburn, generalized linear models (GLMs) extend the range of application of linear statistical models by accommodating response variables with non-normal conditional distributions.
- Except for the error, the right-hand side of a generalized linear model is essentially the same as for a linear model.

- A generalized linear model consists of three components:

1. A *random component*, specifying the conditional distribution of the response variable, Y_i, given the explanatory variables.
   - Traditionally, the random component is a member of an "exponential family" — the normal (Gaussian), binomial, Poisson, gamma, or inverse-Gaussian families of distributions — but generalized linear models have been extended beyond the exponential families.
   - The Gaussian and binomial distributions are familiar.
   - Poisson distributions are often used in modeling count data. Poisson random variables take on non-negative integer values, $0, 1, 2, \ldots$. Some examples are shown in Figure 1.

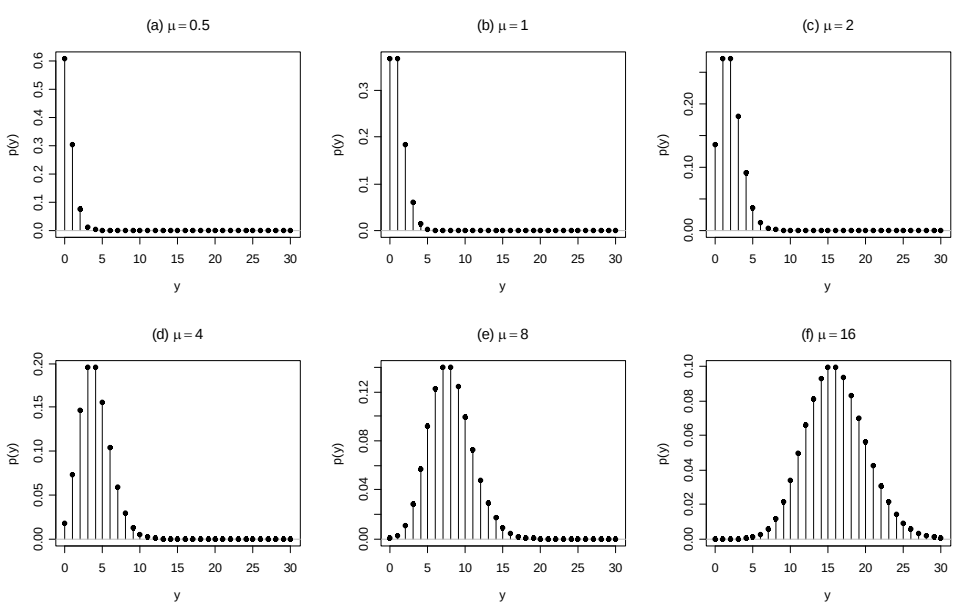

Figure 1. Poisson distributions for various values of the "rate" parameter (mean) μ.

- The gamma and inverse-Gaussian distributions are for positive continuous data; some examples are given in Figure 2.

2. A linear function of the regressors, called the *linear predictor*,

$$\eta_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} = \mathbf{x}_i' \boldsymbol{\beta}$$

on which the expected value μ_i of Y_i depends.

- The X's may include quantitative predictors, but they may also include transformations of predictors, polynomial terms, contrasts generated from factors, interaction regressors, etc.

3. An invertible *link function* $g(\mu_i) = \eta_i$, which transforms the expectation of the response to the linear predictor.

- The inverse of the link function is sometimes called the *mean function*: $g^{-1}(\eta_i) = \mu_i$.

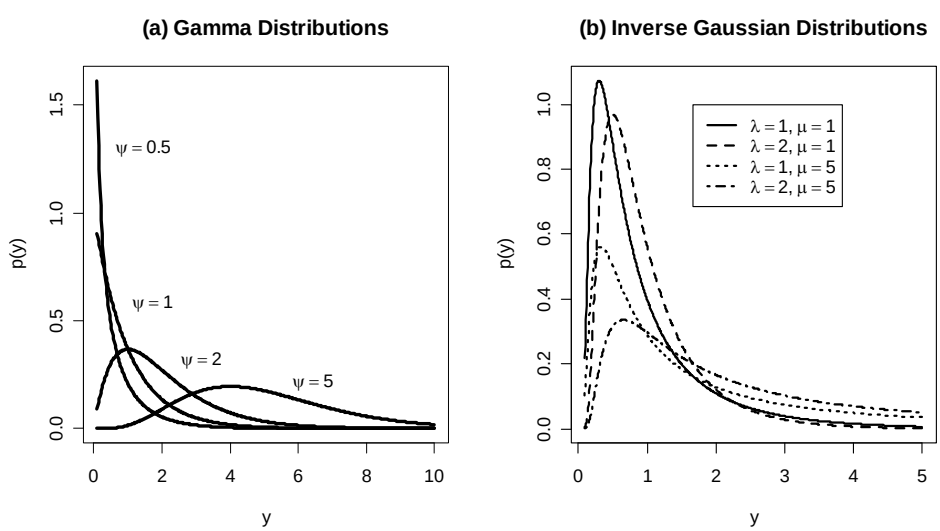

Figure 2. (a) Several gamma distributions for "scale" $\omega = 1$ and various values of the "shape" parameter ψ. (b) Inverse-Gaussian distributions for several combinations of values of the mean μ and "inverse-dispersion" λ.

- Standard link functions and their inverses are shown in the following table:

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
identity	μ_i	η_i
log	$\log_e \mu_i$	e^{η_i}
inverse	μ_i^{-1}	η_i^{-1}
inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
square-root	$\sqrt{\mu_i}$	η_i^2
logit	$\log_e \dfrac{\mu_i}{1-\mu_i}$	$\dfrac{1}{1+e^{-\eta_i}}$
probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

- The logit, probit, and complementary-log-log links are for *binomial data*, where Y_i represents the observed proportion and μ_i the expected proportion of "successes" in n_i binomial trials — that is, μ_i is the probability of a success.

  - For the probit link, Φ is the standard-normal cumulative distribution function, and Φ^{-1} is the standard-normal quantile function.
  - An important special case is *binary data*, where all of the binomial trials are 1, and therefore all of the observed proportions Y_i are either 0 or 1. This is the case that we examined in the previous session.

▶ Although the logit and probit links are familiar, the log-log and complementary log-log links for binomial data are not.

- These links are compared in Figure 3.
- The log-log or complementary log-log link may be appropriate when the probability of the response as a function of the linear predictor approaches 0 and 1 asymmetrically.

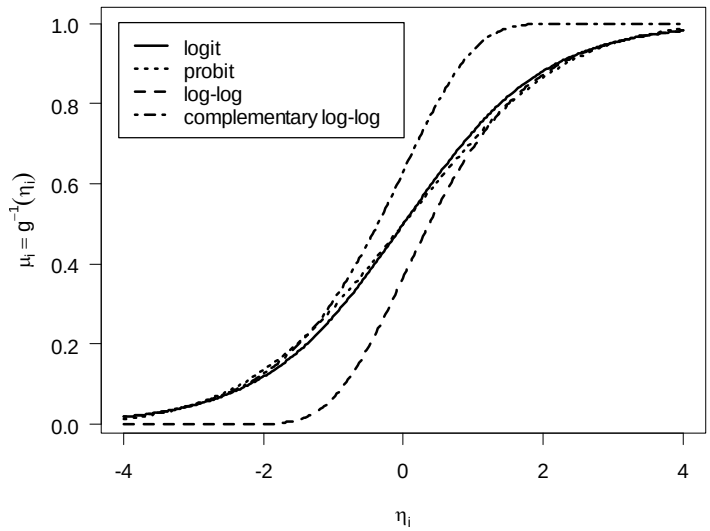

Figure 3. Comparison of logit, probit, and complementary log-log links. The probit link is rescaled to match the variance of the logistic distribution, $\pi^2/3$.

- For distributions in the exponential families, the conditional variance of Y is a function of the mean μ together with a dispersion parameter ϕ (as shown in the table below).
  - For the binomial and Poisson distributions, the dispersion parameter is fixed to 1.
  - For the Gaussian distribution, the dispersion parameter is the usual error variance, which we previously symbolized by σ_ε^2 (and which doesn't depend on μ).

Family	*Canonical Link*	*Range of* Y_i	$V(Y_i\|\eta_i)$
Gaussian	identity	$(-\infty, +\infty)$	ϕ
binomial	logit	$\frac{0, 1, ..., n_i}{n_i}$	$\frac{\mu_i(1-\mu_i)}{n_i}$
Poisson	log	$0, 1, 2, ...$	μ_i
gamma	inverse	$(0, \infty)$	$\phi\mu_i^2$
inverse-Gaussian	inverse-square	$(0, \infty)$	$\phi\mu_i^3$

- The *canonical link* for each familiy is not only the one most commonly used, but also arises naturally from the general formula for distributions in the exponential families.
  - Other links may be more appropriate for the specific problem at hand
  - One of the strengths of the GLM paradigm — in contrast, for example, to transformation of the response variable in a linear model — is the separation of the link function from the conditional distribution of the response.
- GLMs are typically fit to data by the method of maximum likelihood.
  - Denote the maximum-likelihood estimates of the regression parameters as $\widehat{\alpha}, \widehat{\beta}_1, ..., \widehat{\beta}_k$.
    - These imply an estimate of the mean of the response, $\widehat{\mu}_i = g^{-1}(\widehat{\alpha} + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_k x_{ik})$.

- The log-likelihood for the model, maximized over the regression coefficients, is

$$\log_e L_0 = \sum_{i=1}^{n} \log_e p(\widehat{\mu}_i, \phi; y_i)$$

  where $p(\cdot)$ is the probability or probability-density function corresponding to the family employed.
- A "saturated" model, which dedicates one parameter to each observation, and hence fits the data perfectly, has log-likelihood

$$\log_e L_1 = \sum_{i=1}^{n} \log_e p(y_i, \phi; y_i)$$

- Twice the difference between these log-likelihoods defines the *residual deviance* under the model, a generalization of the residual sum of squares for linear models:

$$D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = 2(\log_e L_1 - \log_e L_0)$$

  where $\mathbf{y} = \{Y_i\}$ and $\widehat{\boldsymbol{\mu}} = \{\widehat{\mu}_i\}$.

- Dividing the deviance by the estimated dispersion produces the *scaled deviance*: $D(\mathbf{y}; \widehat{\boldsymbol{\mu}})/\widehat{\phi}$.
- Likelihood-ratio tests can be formulated by taking differences in the residual deviance for nested models.
- For models with an estimated dispersion parameter, one can alternatively use incremental F-tests.
- Wald tests for individual coefficients are formulated using the estimated asymptotic standard errors of the coefficients.

▶ Some familiar examples:

- Combining the identity link with the Gaussian family produces the normal linear model.
  - The maximum-likelihood estimates for this model are the ordinary least-squares estimates.
- Combining the logit link with the binomial family produces the logistic-regression model (linear-logit model).

- Combining the probit link with the binomial family produces the linear probit model.

3. Poisson GLMs for Count Data

- Poisson generalized linear models arise in two common formally identical but substantively distinguishable contexts:

1. when the response variable in a regression model takes on non-negative integer values, such as a count;
2. to analyze associations among categorical variables in a contingency table of counts.

- The canonical link for the Poisson family is the log link.

3.1 Over-Dispersed Binomial and Poisson Models

- The binomial and Poisson GLMs fix the dispersion parameter ϕ to 1.
- It is possible to fit versions of these models in which the dispersion is a free parameter, to be estimated along with the coefficients of the linear predictor
  - The resulting error distribution is not an exponential family; the models are fit by "quasi-likelihood."
- The regression coefficients are unaffected by allowing dispersion different from 1, but the coefficient standard errors are multiplied by the square-root of $\widehat{\phi}$.
  - Because the estimated dispersion typically exceeds 1, this inflates the standard errors
  - That is, failing to account for "over-dispersion" produces misleadingly small standard errors.

- So-called *over-dispersed* binomial and Poisson models arise in several different circumstances.
  - For example, in modeling proportions, it is possible that
    - the probability of success μ varies for different individuals who share identical values of the predictors (this is called "unmodeled heterogeneity");
    - or the individual successes and failures for a "binomial" observation are not independent, as required by the binomial distribution.
- The negative-binomial distribution is also frequently used to model over-dispersed count data.

4. Diagnostics for GLMs

- Most regression diagnostics extend straightforwardly to generalized linear models.
- These extensions typically take advantage of the computation of maximum-likelihood estimates for generalized linear models by iterated weighted least squares (the procedure typically used to fit GLMs).

4.1 Outlier, Leverage, and Influence Diagnostics

4.1.1 Hat-Values

- Hat-values for a generalized linear model can be taken directly from the final iteration of the IWLS procedure
- They have the usual interpretation — except that the hat-values in a GLM depend on Y as well as on the configuration of the X's.

4.1.2 Residuals

- Several kinds of residuals can be defined for generalized linear models:
  - *Response residuals* are simply the differences between the observed response and its estimated expected value: $Y_i - \widehat{\mu}_i$.
  - *Working residuals* are the residuals from the final WLS fit.
    - These may be used to define partial residuals for component-plus-residual plots (see below).
  - *Pearson residuals* are case-wise components of the Pearson goodness-of-fit statistic for the model:

$$\frac{\widehat{\phi}^{1/2}(Y_i - \widehat{\mu}_i)}{\sqrt{\widehat{V}(Y_i|\eta_i)}}$$

where ϕ is the dispersion parameter for the model and $V(Y_i|\eta_i)$ is the variance of the response given the linear predictor.

- *Standardized Pearson residuals* correct for the conditional response variation and for the leverage of the observations:

$$R_{Pi} = \frac{Y_i - \widehat{\mu}_i}{\sqrt{\widehat{V}(Y_i|\eta_i)(1-h_i)}}$$

.

- *Deviance residuals*, D_i, are the square-roots of the case-wise components of the residual deviance, attaching the sign of $Y_i - \widehat{\mu}_i$.

▶ *Standardized deviance residuals* are

$$R_{Di} = \frac{D_i}{\sqrt{\widehat{\phi}(1-h_i)}}$$

▶ Several different approximations to *studentized residuals* have been suggested.

- To calculate exact studentized residuals would require literally refitting the model deleting each observation in turn, and noting the decline in the deviance.

- Here is an approximation due to Williams:

$$E_i^* = \sqrt{(1-h_i)R_{Di}^2 + h_i R_{Pi}^2}$$

where, once again, the sign is taken from $Y_i - \widehat{\mu}_i$.

- A Bonferroni outlier test using the standard normal distribution may be based on the largest absolute studentized residual.

4.1.3 Influence Measures

- An approximation to Cook's distance influence measure is

$$D_i = \frac{R^2_{Pi}}{\widehat{\phi}(k+1)} \times \frac{h_i}{1-h_i}$$

- Approximate values of dfbeta$_{ij}$ and dfbetas$_{ij}$ (influence and standardized influence on each coefficient) may be obtained directly from the final iteration of the IWLS procedure.
- There are two largely similar extensions of added-variable plots to generalized linear models, one due to Wang and another to Cook and Weisberg.

4.2 Nonlinearity Diagnostics

- Component-plus-residual plots also extend straightforwardly to generalized linear models.
  - Nonparametric smoothing of the resulting scatterplots can be important to interpretation, especially in models for binary responses, where the discreteness of the response makes the plots difficult to examine.
  - Similar effects can occur for binomial and Poisson data.
- Component-plus-residual plots use the linearized model from the last step of the IWLS fit.
  - For example, the partial residual for X_j adds the working residual to B_jX_{ij}.
  - The component-plus-residual plot graphs the partial residual against X_j.

5. Logit and Loglinear Models for Contingency Tables

5.1 The Binomial Logit Model for Contingency Tables

- When the explanatory variables — as well as the response variable — are discrete, the joint sample distribution of the variables defines a contingency table of counts.
- An example, drawn from *The American Voter* (Converse et al., 1960), appears below.
  - This table, based on data from a sample survey conducted after the 1956 U.S. presidential election, relates voting turnout in the election to strength of partisan preference, and perceived closeness of the election:

Perceived Closeness	*Intensity of Preference*	*Turnout*	
		Voted	Did Not Vote
One-Sided	Weak	91	39
	Medium	121	49
	Strong	64	24
Close	Weak	214	87
	Medium	284	76
	Strong	201	25

- The following table gives the *empirical logit* for the response variable,

$$\log_e \frac{\text{proportion voting}}{\text{proportion not voting}}$$

for each of the six combinations of categories of the explanatory variables:

Perceived Closeness	*Intensity of Preference*	$\log_e \frac{\text{Voted}}{\text{Did Not Vote}}$
One-Sided	Weak	0.847
	Medium	0.904
	Strong	0.981
Close	Weak	0.900
	Medium	1.318
	Strong	2.084

· For example,

$$\begin{aligned}
&\text{logit(voted}|\text{one-sided, weak preference)}\\
&= \log_e \frac{91/130}{39/130}\\
&= \log_e \frac{91}{39}\\
&= 0.847
\end{aligned}$$

· Because the conditional proportions voting and not voting share the same denominator, the empirical logit can also be written as

$$\log_e \frac{\text{number voting}}{\text{number not voting}}$$

· The empirical logits are graphed in Figure 4, much in the manner of profiles of cell means for a two-way analysis of variance.

► Logit models are fully appropriate for tabular data.

- When, as in the example, the explanatory variables are qualitative or ordinal, it is natural to use logit or probit models that are analogous to analysis-of-variance models.

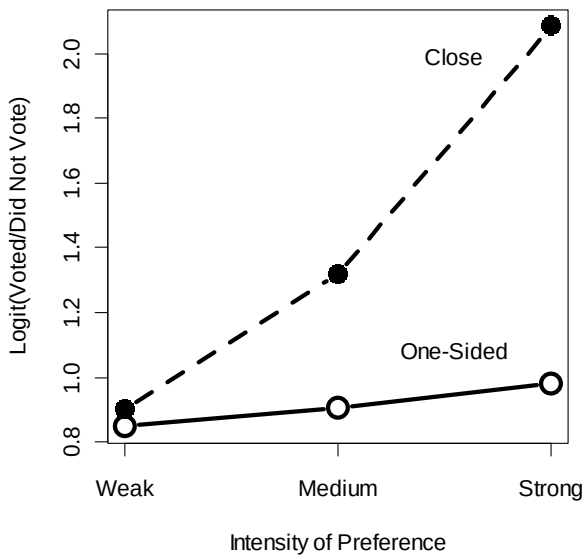

Figure 4. Empirical logits for the *American Voter* data.

- Treating perceived closeness of the election as the 'row' factor and intensity of partisan preference as the 'column' factor, for example, yields the model

$$\text{logit } \pi_{jk} = \mu + \alpha_j + \beta_k + \gamma_{jk}$$

where

· π_{jk} is the conditional probability of voting in combination of levels j of perceived closeness and k of preference;

· μ is the general mean of turnout in the population;

· α_j is the main effect on turnout of membership in the jth level of perceived closeness;

· β_k is the main effect on turnout of membership in the kth levels of preference; and

· γ_{jk} is the interaction effect on turnout of simultaneous membership in levels j of perceived closeness and k of preference.

- Under the usual sigma constraints, this model leads to deviation-coded regressors (`contr.sum` in R), as in the analysis of variance.
- Likelihood-ratio tests for main-effects and interactions can be constructed in close analogy to the incremental F-tests for the two-way ANOVA model.

5.2 Loglinear Models

- Poisson GLMs may also be used to fit loglinear models to a contingency table of frequency counts, where the object is to model association among the variables in the table.
- The variables constituting the classifications of the table are treated as 'explanatory variables' in the Poisson model, while the cell count plays the role of the 'response.'
- We previously examined Campbell et al.'s data on voter turnout in the 1956 U. S. presidential election
  - We used a binomial logit model to analyze a three-way contingency table for turnout by perceived closeness of the election and intensity of partisan preference.
  - The binomial logit model treats turnout as the response.
- An alternative is to construct a log-linear model for the expected cell count.

- This model looks very much like a three-way ANOVA model, where in place of the cell mean we have the log cell expected count:

$$\begin{aligned}\log \mu_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_k \\ &\quad + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk}\end{aligned}$$

- Here, variable 1 is perceived closeness of the election; variable 2 is intensity of preference; and variable 3 is turnout.
- Although a term such as $\alpha\beta_{ij}$ looks like an 'interaction,' it actually models the association between variables 1 and 2.
- The three-way term $\alpha\beta\gamma_{ijk}$ allows the association between any pair of variables to be different in different categories of the third variable; it thus represents an interaction in the usual sense of that concept.

▶ In fitting the log-linear model to data, we can use sigma-constraints on the parameters, much as we would for an ANOVA model.

▶ In the context of a three-way contingency table, the loglinear model above is a saturated model, because it has as many independent parameters (12) as there are cells in the table.

▶ The likelihood-ratio test for the three-way term Closeness $\times$ Preference $\times$ Turnout is identical to the test for the Closeness $\times$ Preference interaction in the logit model in which Turnout is the response variable.

▶ In general, as long as we fit the parameters for the associations among the explanatory variable (here Closeness$\times$Preference and, of course, its lower-order relatives, Closeness and Preference) and for the marginal distribution of the response (Turnout), the loglinear model for a contingency table is equivalent to a logit model.
- There is, therefore, no real advantage to using a loglinear model in this setting.
- Loglinear models, however, can be used to model association in other contexts.

6. Implementation of GLMs in R

- The `glm()` function in R is very similar in use to `lm()`,

```
glm(formula, family, data, subset,
      weights, na.action, contrasts)
```

- The `family` argument is one of `gaussian` (the default), `binomial`, `poisson`, `Gamma`, `inverse.gaussian`, `quasi`, `quasibinomial`, or `quasipoisson`.
  - It is possible to write functions for additional families (e.g., the `negative.binomial` family for count data in the **MASS** package).
- The "family-generator" function specified as the value of the `family` argument can itself take a link argument (and possibly other arguments); in each case there is a default link.
  - The available links for each family (○) and the default link (●) are given in the following table:

	link			
family	identity	inverse	sqrt	1/mu^2
gaussian	●	○		
binomial				
poisson	○		○	
Gamma	○	●		
inverse. gaussian	○	○		●
quasi	●	○	○	○
quasibinomial				
quasipoisson	○		○	

family	link			
	log	logit	probit	cloglog
gaussian	○			
binomial	○	●	○	○
poisson	●			
Gamma	○			
inverse. gaussian	○			
quasi	○	○	○	○
quasibinomial		●	○	○
quasipoisson	●			