Sociology 740

John Fox

Lecture Notes

11. Generalized Linear Models: An Introduction

1. Introduction

- 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.

2. Goals

- To introduce the format and structure of generalized linear models
- To show how the familiar linear, logit, and probit models fit into the GLM framework.
- To introduce Poisson generalized linear models for count data.
- To describe diagnostics for generalized linear models.

3. The Structure of Generalized Linear Models

- 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.

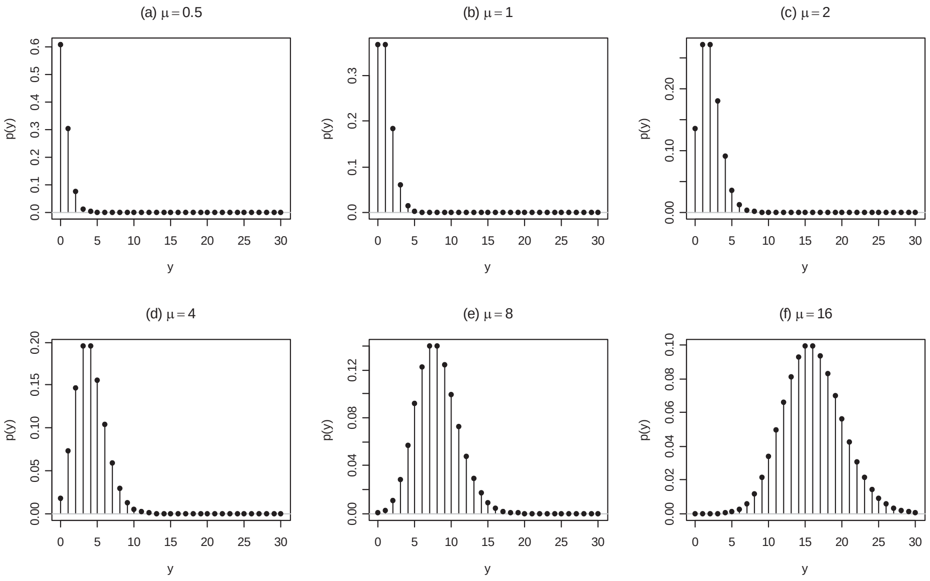

Figure 1. Poisson distributions for various values of the "rate" parameter (mean) μ.

- 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*,

$$\eta_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

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*: $g^{-1}(\eta_i) = \mu_i$.

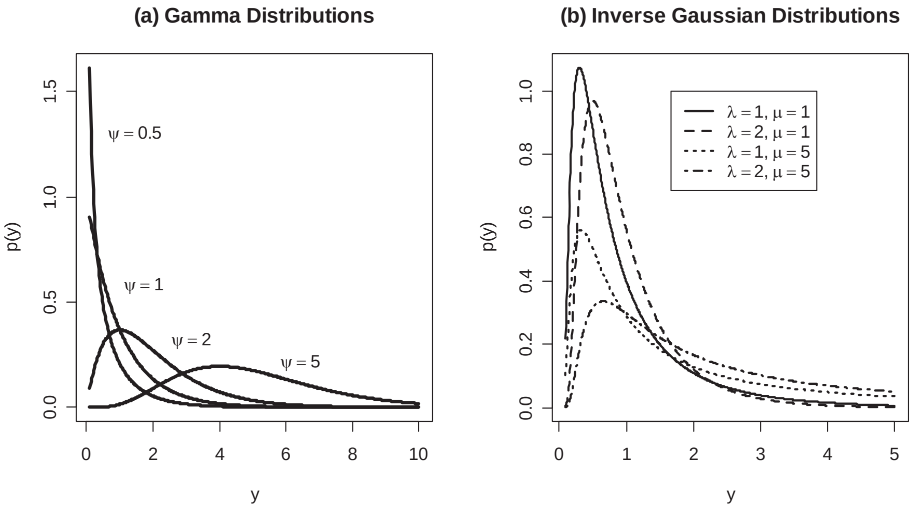

Figure 2. (a) Several gamma distributions for “scale” $\omega = 1$ and various values of the “shape” parameter ψ. (b) Inverse-Gaussian distributions for several combinations of values of the mean μ and “inverse-dispersion” λ.

- 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 the previous lecture.

▶ 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:

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

  - 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.

- 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.

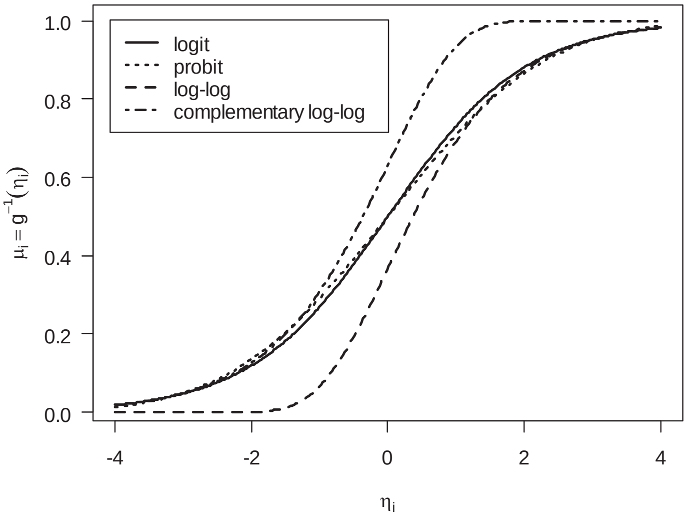

Figure 3. Comparison of logit, probit, and complementary log-log links. The probit link is rescaled to match the variance of the logistic distribution, $\pi^2/3$.

4. 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 (an application that I won't take up here).

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

4.1 Poisson Regression

- Recall Ornstein's data on interlocking director and top-executive positions among 248 major Canadian firms
  - Ornstein performed a least-squares regression of the number of interlocks maintained by each firm on the firm's assets, and dummy variables for the firm's nation of control and sector of operation.
  - I found that a square-root transformation of the response variable tends to stabilize residual variance and make the distribution of the residuals more symmetric.
- Because the response variable is a count, a Poisson linear model might be preferable.
  - The marginal distribution of number of interlocks, in Figure 4, shows many zero counts and an extreme positive skew.
  - Fitting a Poisson GLM with log link to Ornstein's data produces the following results:

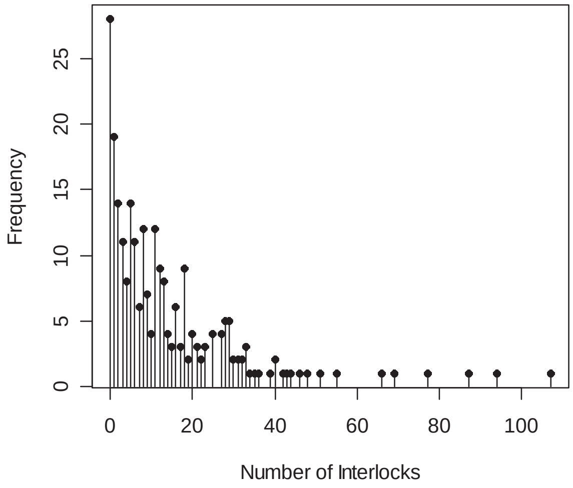

Figure 4. Distribution of number of interlocks maintained by 248 large Canadian corporations.

	Coefficient	*Standard Error*
Constant	2.32	0.052
Assets	0.0000209	0.0000012
Nation of Control (baseline: Canada)		
Other	-0.163	0.073
United Kingdom	-0.577	0.089
United States	-0.826	0.049
Sector (baseline: Agriculture and Food)		
Banking	-0.409	0.156
Construction	-0.620	0.211
Finance	0.677	0.069
Holding Company	0.208	0.119
Manufacturing	0.0527	0.0752
Merchandizing	0.178	0.087
Mining	0.621	0.069
Transportation	0.678	0.075
Wood and Forest Products	0.712	0.075

– An analysis of deviance table for the model shows that all three explanatory variables have highly statistically significant effects:

Source	G^2	df	p
Assets	390.90	1	$\ll .0001$
Nation of Control	328.94	3	$\ll .0001$
Sector	361.46	9	$\ll .0001$

– The deviance for the null model (with only a constant) is 3737.0, and 1887.4 for the full model; thus an analog of the squared multiple correlation is

$$R^2 = 1 - \frac{1887.4}{3737.0} = .495$$

– Effect displays for the model are shown in Figure 5.

4.2 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.

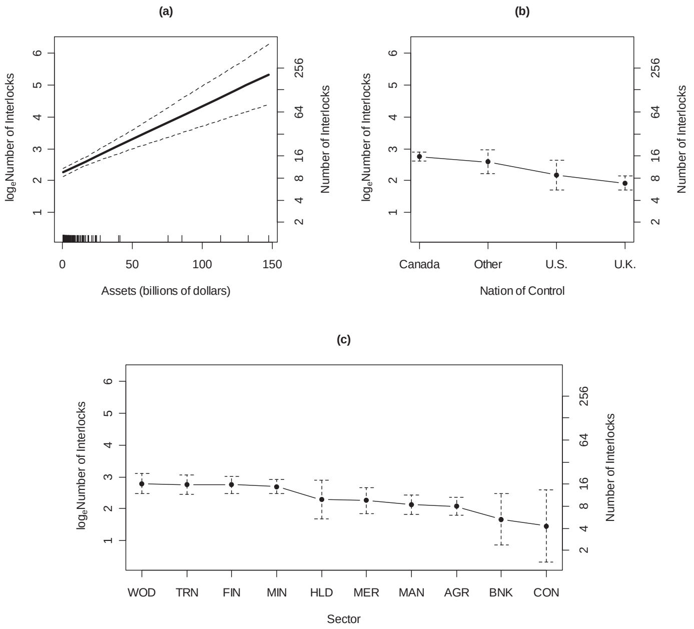

Figure 5. Effect displays for the (over-dispersed) Poisson regression model fit to Ornstein's interlocking-directorate data.

- 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.

5. 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).

5.1 Outlier, Leverage, and Influence Diagnostics

5.1.1 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.

5.1.2 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.

5.1.3 Influence Measures

- An approximation to Cook's distance influence measure is
$$D_i = \frac{R_{Pi}^2}{\widehat{\phi}(k+1)} \times \frac{h_i}{1-h_i}$$
- 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.

5.2 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 $B_j X_{ij}$.
  - The component-plus-residual plot graphs the partial residual against X_j.

- An illustrative component+residual plot, for assets in the over-dispersed Poisson regression fit to Ornstein's interlocking-directorate data appears in Figure 6.
  - This plot is difficult to examine because of the large positive skew in assets, but it appears as if the assets slope is a good deal steeper at the left than at the right.
  - I therefore investigated transforming assets down the ladder of powers and roots, eventually arriving at the log transformation, the component+residual plot for which appears quite straight (Figure 7).

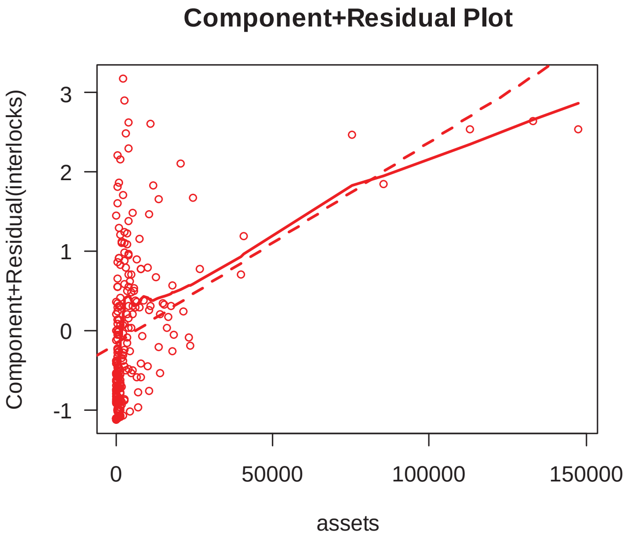

Figure 6. Component+residual plot for assets in the over-dispersed Poisson regression for Ornstein's data.

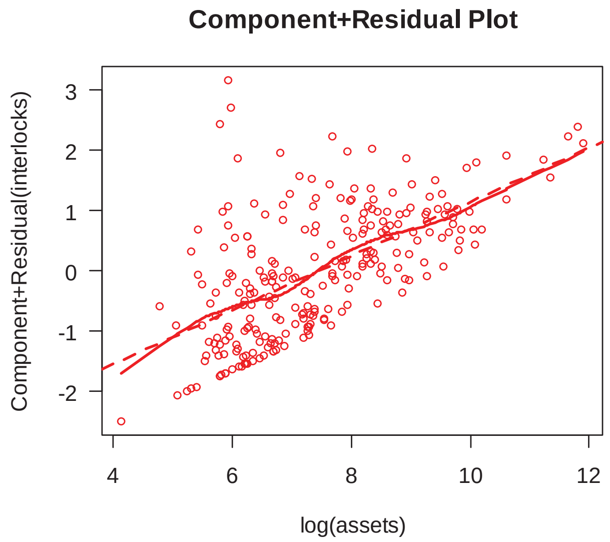

Figure 7. Component+residual plot for log(assets) in the respecified over-dispersed Poisson regression model for Ornstein's data.

6. Summary

▶ Generalized linear models (GLMs) consist of three components:

(a) A random component specifying the conditional distribution of the response variable Y given the explanatory variables, traditionally a member of an exponential family — the normal (Gaussian), binomial, Poisson, gamma, or inverse-Gaussian families of distributions.

– For distributions in exponential families, the conditional variance of Y is a function of μ, the mean of Y, and of a dispersion parameter ϕ; in the binomial and Poisson families, ϕ is fixed to 1.

(b) A linear predictor, $\eta_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$.

(c) A link function $g(\mu_i) = \eta_i$, which transforms the expectation of the response to the linear predictor; the inverse of the link is the mean function, $g^{-1}(\eta_i) = \mu_i$.

▶ Traditional GLMs are fit to data by maximum likelihood.

- The deviance under a fitted model is $D(\mathbf{y}; \widehat{\boldsymbol{\mu}}) = 2(\log_e L_1 - \log_e L_0)$, where $\mathbf{y}$ contains the observed values of the response variable, $\widehat{\boldsymbol{\mu}}$ contains the fitted values of the response, L_1 is the maximized likelihood for a saturated model that dedicates one parameter to each observation, and L_0 is the maximized likelihood under the model in question.
- The scaled deviance is $D(\mathbf{y}; \widehat{\boldsymbol{\mu}})/\widehat{\phi}$, where $\widehat{\phi}$ is an estimate of the dispersion.
- In analogy to incremental F-tests in an analysis of variance for linear models, differences in deviance may be used for likelihood-ratio tests in GLMs; for models with a dispersion parameter, F-tests are also available.
- Wald tests for individual coefficients are produced by dividing the estimated coefficients by their standard errors.

▶ The binomial family is used for dichotomous response variables. Pairing the binomial family with the logit link produces the logistic-regression model; pairing the binomial family with the probit link produces the probit model.

▶ The Poisson family is often used to analyze count data. The canonical link for the Poisson family is the log link.

▶ Over-dispersed binomial and Poisson models introduce a dispersion parameter ϕ that is not fixed to 1; these models are fit by quasi-likelihood.

▶ Most standard linear-model diagnostics may be generalized to GLMs. These include hat-values, studentized residuals, Cook's distances, added-variable plots, and component-plus-residual plots (among others).