Generalized Linear Models[1]

STA 2101/442: Fall 2012

[1] See last slide for copyright information.

Suggested Reading: Davison's *Statistical models*

- Exponential families of distributions Sec. 5.2
- Chapter 10 is on nonlinear regression models
- See pp. 468-492

Overview

Examples of Generalized Linear Models

- Normal regression
- Logistic regression
- Poisson regression

Components of a Generalized Linear Model

- **Random Component**: Probability distribution for Y
- **Systematic component**: Specifies explanatory variables in the form of a "linear predictor that looks like a regression equation.
- **Link function**: Connects $\mu = E(Y|\mathbf{X})$ to the linear predictor

Random Component: Distribution of Y

- Ordinary regression: Normal
- Logistic regression: Bernoulli
- Poisson regression: Poisson

- Other possibilities: Binomial, Exponential, Gamma, Geometric . . .

Systematic component: A regression-like equation called the *linear predictor*

$$\eta = \beta_0 + \beta_1 x_1 + \ldots, + \beta_{p-1} x_{p-1}$$

Link Function: The linear predictor is an increasing function of the expected value

$$g(\mu) = \beta_0 + \beta_1 x_1 + \ldots, +\beta_{p-1} x_{p-1}$$

- The function $g(x)$ is strictly increasing.
- The linear predictor is an increasing function of μ.
- So μ is an increasing function of the linear predictor.

Normal Distribution

Link function $g(\mu) = \beta_0 + \beta_1 x_1 + \ldots, +\beta_{p-1} x_{p-1}$

- $E(Y) = \mu$
- $g(\mu) = \mu$
- $\mu = \beta_0 + \beta_1 x_1 + \ldots, +\beta_{p-1} x_{p-1}$
- The identity link

Bernoulli Distribution

Link function $g(\mu) = \beta_0 + \beta_1 x_1 + \ldots, +\beta_{p-1} x_{p-1}$

- $E(Y) = \mu = \pi$
- $g(\mu) = \log \frac{\mu}{1-\mu}$
- $\log \frac{\mu}{1-\mu} = \beta_0 + \beta_1 x_1 + \ldots, +\beta_{p-1} x_{p-1}$
- The logit link

Poisson Distribution

Link function $g(\mu) = \beta_0 + \beta_1 x_1 + \ldots, +\beta_{p-1} x_{p-1}$

- $E(Y) = \mu = \lambda$
- $g(\mu) = \log(\mu)$
- $\log(\mu) = \beta_0 + \beta_1 x_1 + \ldots, +\beta_{p-1} x_{p-1}$
- The log link

"Natural" Exponential Family of Distributions

- Includes most of the familiar distributions
- Provides a unified theory for generalized linear models
- Leads to a general, highly efficient method for finding MLEs numerically
  - Iterative weighted least squares
  - Closely related to Newton-Raphson
- Points to a *natural* link function.
- The "natural" parameter of a one-parameter exponential family is $\theta = g(\mu)$.
- The link functions we have been using are natural links.

Natural exponential family of distributions

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}$$

- Support does not depend on θ or ϕ.
- θ is the natural parameter.
- ϕ is the dispersion parameter, often known.
- $\theta = g(\mu)$, where $\mu = E(Y)$
- $E(Y) = b'(\theta)$ gives $\mu = g^{-1}(\theta)$
- $Var(Y) = \phi\, b''(\theta) = \phi V(\mu)$
- $V(\mu)$ is called the *variance function.*

Normal

$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}$

$$\begin{aligned}
\frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(y-\mu)^2}{2\sigma^2}} &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{y^2 - 2y\mu + \mu^2}{2\sigma^2}\right\} \\
&= \exp\left\{\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} + \left(-\frac{y^2}{2\sigma^2} - \log\sqrt{\sigma^2} - \log 2\pi\right)\right\}
\end{aligned}$$

- Natural parameter is $\theta = \mu$
- Natural link is the identity function.
- Dispersion parameter is $\phi = \sigma^2$
- $b(\theta) = \frac{\theta^2}{2}$

Bernoulli

$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}$

$$\begin{aligned}
\pi^y(1-\pi)^{1-y} &= \exp\{y\log\pi + (1-y)\log(1-\pi)\} \\
&= \exp\{y(\log\pi - \log(1-\pi)) + \log(1-\pi)\} \\
&= \exp\left\{y\left(\log\frac{\pi}{1-\pi}\right) + \log(1-\pi)\right\} \\
&= \exp\left\{\frac{y\left(\log\frac{\pi}{1-\pi}\right) - (-\log(1-\pi))}{1} + 0\right\}
\end{aligned}$$

- Natural parameter is $\theta = \log\frac{\pi}{1-\pi} = \log\frac{\mu}{1-\mu}$
- Natural link is the logit function.
- Dispersion parameter is $\phi = 1$
- $b(\theta) = \log(1+e^\theta)$

Deviance

- Goal is to compare a model to a "Super" model that fits the data as well as possible.
- Example: If an experiment has c outcomes, you can't beat a multinomial with c categories.
- The $c-1$ parameters soak up all $c-1$ degrees of freedom, so in this case you could call the **S**uper model "**S**aturated."

Deviance $= -2(\ell_M - \ell_S)$

ℓ is the maximized log likelihood

- Denote the parameter of the Model by θ and the parameter of the Supermodel by σ
- The models might look very different, including the parameter spaces.

$$
\begin{aligned}
-2(\ell_M - \ell_S) &= -2log \frac{\prod_{i=1}^{n} f(y_i|\widehat{\theta})}{\prod_{i=1}^{n} f(y_i|\widehat{\sigma})} \\
&= -2log \prod_{i=1}^{n} \frac{f(y_i|\widehat{\theta})}{f(y_i|\widehat{\sigma})} \\
&= \sum_{i=1}^{n} -2 \log \left(\frac{f(y_i|\widehat{\theta})}{f(y_i|\widehat{\sigma})} \right) \\
&= \sum_{i=1}^{n} d_i
\end{aligned}
$$

Deviance $= -2log\frac{\prod_{i=1}^{n} f(y_i|\widehat{\theta})}{\prod_{i=1}^{n} f(y_i|\widehat{\sigma})} = \sum_{i=1}^{n} d_i$

- The deviance terms d_i are contributions to a difference in fit (deviance) between the model and the best possible model.
- They are somewhat like residuals.
- Maybe big ones are worth investigating.
- Deviance residuals are defined as $r_i^D = \text{sign}(y_i - \widehat{\mu}_i)\sqrt{d_i}$

Deviance looks like the likelihood ratio statistic G^2

Deviance $= -2log\frac{\prod_{i=1}^{n} f(y_i|\widehat{\theta})}{\prod_{i=1}^{n} f(y_i|\widehat{\sigma})} = \sum_{i=1}^{n} d_i$

- Looks like the model represents a null hypothesis.
- The Supermodel is somehow less restricted.
- So *sometimes* it must be a chi-squared test for goodness of model fit.
- What is that ideal "Supermodel" that fits as well as possible?

What is the model that fits as well as possible?

- If just a few (c) categories and plenty of observations in each category (say at least 5), the best model is a multinomial.
  - Any model with $c-1$ parameters that are 1-1 with $\pi_1, \ldots, \pi_{c-1}$ will *soak up* all the degrees of freedom and is said to be "saturated."
  - For a saturated model, the deviance is zero.
  - A model with fewer than $c-1$ parameters cannot be saturated, and the deviance is a likelihood ratio test statistic, null hypothesis that the model is true.
- There are some other examples of super-models that are reasonable. In structural equation models, an example is the unrestricted multivariate normal.
- Often, the super-model is not a reasonable model.

An unreasonable model

Logistic regression with continuous explanatory variables

- One observation only in each of n combinations of explanatory, response variable values.
- One parameter for each observation.
- Model fits perfectly.
- Likelihood equals one.
- All parameter estimates on the boundary of the parameter space.
- *Not* chi-squared under H_0.
- Denominator of deviance equals one.
- Deviance is just -2 log likelihood of the model.
- Deviance is not a test of model fit, or anyway nobody knows the distribution under H_0.

What happens when there are *a few* ties in the explanatory variable values . . .

R's `help glm` defines the deviance as

"…up to a constant, minus twice the maximized log-likelihood. Where sensible, the constant is chosen so that a saturated model has deviance zero."

At least, Deviance $= -2(\ell_M - \ell_S)$ is -2log likelihood plus a constant, so the *difference* in deviance values between 2 nested models should be the large-sample likelihood ratio test of full *vs.* reduced.

One last scary question
And a reassuring answer

If you fit a full and a reduced model separately, might they use a different definition of the supermodel, and hence the deviance?

- I have tried unsuccessfully to make R misbehave this way.
- The null deviance is the deviance of a model with just an intercept.
- Compare the null deviance of your full and reduced models. If they are the same, both models are using the same definition of deviance and everything is okay.
- And in my experience with R's `glm` function, they are always the same.

Copyright Information