

Generalized Linear Models¹

STA 312: Fall 2012

¹See last slide for copyright information.

Suggested Reading: Chapter 3

- Read Pages 65-77
- Section 3.4.3, Pages 85-86 on the deviance is important. The rest of the chapter is optional.

Overview

- 1 Basics
- 2 The Exponential Family of Distributions
- 3 Deviance

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*

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

Or

$$\eta = \alpha + \beta_1 x_1 + \dots, + \beta_k x_k$$

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

$$g(\mu) = \alpha + \beta_1 x_1 + \dots + \beta_k x_k$$

- 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 $g(\mu) = \alpha + \beta_1 x_1 + \dots, + \beta_k x_k$

- $E(Y) = \mu$
- $g(\mu) = \mu$
- $\mu = \alpha + \beta_1 x_1 + \dots, + \beta_k x_k$
- The identity link

Bernoulli Distribution

Link function $g(\mu) = \alpha + \beta_1 x_1 + \dots, + \beta_k x_k$

- $E(Y) = \mu = \pi$
- $g(\mu) = \log \frac{\mu}{1-\mu}$
- $\log \frac{\mu}{1-\mu} = \alpha + \beta_1 x_1 + \dots, + \beta_k x_k$
- The logit link

Poisson Distribution

Link function $g(\mu) = \alpha + \beta_1 x_1 + \dots, + \beta_k x_k$

- $E(Y) = \mu = \lambda$
- $g(\mu) = \log(\mu)$
- $\log(\mu) = \alpha + \beta_1 x_1 + \dots, + \beta_k x_k$
- 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.

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 **Super** model “**Saturated.**”

$$\text{Deviance} = -2(L_M - L_S)$$

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.

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

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

- 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 $r_i^D = \text{sign}(y_i - \hat{\mu}_i) \sqrt{d_i}$

Deviance looks like the likelihood ratio statistic G^2

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

- 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), it's a multinomial.
 - Any model with $c - 1$ parameters that are 1-1 with π_1, \dots, π_{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, $\text{Deviance} = -2(L_M - L_S)$ is $-2\log$ 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

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

This slide show was prepared by **Jerry Brunner**, Department of Statistics, University of Toronto. It is licensed under a **Creative Commons Attribution - ShareAlike 3.0 Unported License**. Use any part of it as you like and share the result freely. The \LaTeX source code is available from the course website:
<http://www.utstat.toronto.edu/~brunner/oldclass/312f12>