Lecture 11: Introduction to Generalized Linear Models

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011

Division of Biostatistics and Epidemiology

Medical University of South Carolina

Outline

1. Introduction (motivation and history)
2. Review ordinary linear regression
3. Components of a GLM
4. Natural Exponential family

Brief Overview of Statistics

	Response Variable		
Explanatory Variables	Binary	Nominal	Continuous
Binary	2×2 table logistic regression	Contingency tables log-linear models	t-tests
Nominal	Logistic regression Log-linear models	Contingency tables log-linear models	ANOVA
Continuous	Dose-response models logistic regression	It depends	Multiple regression
Some Continuous and some categorical	Logistic regression	It depends	ANCOVA Multiple regression

Note, in general, most common analyses can be approached from a “modelling” approach. Some such as the log-linear and logistic are topics for this class.

Motivation for Modeling

Why do we want to “model” data?

- The structural form of the model describes the patterns of interactions or associations in data.
- Inference for the model parameters provides a way to evaluate which explanatory variable(s) are related to the response variable(s) while statistically controlling for the other variables.
- Estimated model parameters provide measures of the strength and importance of effects.
- A model’s predicted values “smooth” the data - That is, they provide good estimates of the mean of the response variable.
- Modeling enables use to examine general extensions to the methods we have studied thus far.

Review of Ordinary Linear Regression

Suppose you have a continuous response variable (Y) and continuous and/or discrete explanatory variables (X's).

You want to model the responses as a function of the explanatory variables (covariates). For example,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

where

1. Y_i is the response for the i^{th} subject
2. $\vec{\beta} = (\beta_0, \beta_1, \beta_2)'$ is a (column) vector of constants (or parameters) that describe the shape of the regression "line" (line, curve, etc)
3. $\vec{X}_i = (1, x_{1i}, x_{2i})$ is the (row) vector of explanatory variables for the i^{th} subject.
4. e_i is the random error assumed to be distributed as $N(0, \sigma^2)$

In general, you can view the previous regression model as,

$$Y = E(Y) + \epsilon$$

Where

$$\begin{aligned} E(Y) &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \\ & \quad \text{or in more general terms} \\ &= X_{n\times p}\beta_{p\times 1} \end{aligned}$$

Thus, $E(Y)$ is the $n \times 1$ vector of expectations.

Note,

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_n \end{bmatrix} = \begin{bmatrix} x_{11}, x_{12}, \ldots, x_{1p} \\ x_{21}, x_{22}, \ldots, x_{2p} \\ \cdots \\ x_{n1}, x_{n2}, \ldots, x_{np} \end{bmatrix}$$

is called the design matrix.

Common Models of this Type

The analysis of continuous data has relied heavily on the linear model presented. These reflect just a few applications of the linear model.

1. Simple linear regression
2. Multiple regression
3. ANOVA
4. ANCOVA

Estimators

The least squares estimator for β is

$$\tilde{\beta} = (X'X)^{-1}X'Y$$

The predicted value of Y (denoted as $\hat{Y}$) is

$$\hat{Y} = X\tilde{\beta}$$

Diagnostic of the regression fit can be accomplished with the Hat Matrix

$$H = X(X'X)^{-1}X'$$

As we develop our ‘generalized’ approach, you will notice many similarities.

Movement towards a GLM

1. For OLS, we are dependent on the distribution of Y being normal.
2. For categorical data (by definition), the normality assumption is rarely feasible.
3. We may also be interested in other relations of the $X\beta$ with Y. Other mapping functions that ensure the range of Y remains valid is one of the key justifications.

In terms of a GLM, we have three components related to these limitations.

Three Components of a GLM

There are 3 components of a generalized linear model (or GLM)

1. Random Component (the outcome)
2. Systematic Component (the design matrix multiplied by the parameter vector)
3. Link function (the function, $g(\cdot)$ that “links” the systematic component to the random component)

Nelder & Wedderburn (1972) are generally given credit for unifying a broad class (of existing) models into the GLM definitions.

They showed that provided the random component was part of the ‘Exponential Class’, the MLEs for all of the models could be obtained using the same algorithm.

Random Component

- The random component of a GLM consists of a response variable Y with the independent observations $(y_1, y_2, \ldots, y_n)$.
- For a GLM, Y needs to have a distribution in the natural exponential family.
- Recall from theory, an exponential class distribution is of the form

$$f(y_i; \theta_i) = a(\theta_i) b(y_i) exp[y_i Q(\theta_i)]$$

This, in terms of common language, is

- $a(\theta_i)$ is a function only dependent on the unknown parameter
- $b(y_i)$ is a function of the observed sample
- $Q(\theta_i)$ is a function only dependent on the unknown parameter

Easy Example Exponential Class Variable

Suppose,

$$Y \sim \text{Poisson}(\lambda)$$

Then,

$$\begin{aligned} f(y, \lambda) &= \frac{e^{-\lambda}\lambda^y}{y!} \\ &= e^{-\lambda}(\frac{1}{y!})e^{y \log \lambda} \end{aligned}$$

Here $\theta = \lambda$, $a(\theta) = a(\lambda) = e^{-\lambda}$, $b(y) = 1/y!$ and $Q(\pi) = \log \lambda$
Thus,
A Poisson random variable is of the exponential class variable.

Slightly more complicated example

Suppose,

$$Y_i \sim Bern(\pi)$$

where $P(Y_i = 1) = \pi$ and $P(Y_i = 0) = 1 - \pi$

Then,

$$\begin{aligned} f(y_i;\pi) &= \pi^{y_i}(1-\pi)^{1-y_i} \\ &= \frac{\pi^{y_i}}{(1-\pi)^{y_i}(1-\pi)^{-1}} \\ &= (1-\pi)(\frac{\pi}{1-\pi})^{y_i} \\ &= (1-\pi)(1)e^{(y_i \log \frac{\pi}{1-\pi})} \end{aligned}$$

Here $\theta = \pi$, $a(\theta) = a(\pi) = (1-\pi)$, $b(y) = 1$ and $Q(\pi) = \log(\pi/(1-\pi))$

Thus, a Bernoulli random variable is a member of the exponential class.

Exponential Class Variables in General

In general,

- The majority of distributions we will be interested in are exponential class variables
- This includes the more common examples of
  1. Normal
  2. Poisson
  3. Binomial
  4. Multinomial
- It also includes the less common examples of
  1. Gamma
  2. Negative Binomial

Examples

1. Dichotomous (binary) with a fixed number of trials
    - MI / No MI
    - Success/Failure
    - Correct/Incorrect
    - Agree/Disagree

    These responses have a **Bernoulli** distribution.

2. Counts (including cells in a contingency table)
    - Number of babies born at MUSC daily
    - Number of car wrecks per year in Charleston County

    These responses have a **Poisson** distribution.

Most Common Distributions

Although, many distributions are members of the Exponential Class,

For the most part, we will focus on the

1. Binomial
2. Poisson

distributions.

However, the approach we will discuss works equally well for all exponential class distributions.

Systematic Component

Denote a new vector $(\eta_1, \eta_2, \ldots, \eta_n)$ such that

$$\begin{aligned} \eta_i &= \textstyle\sum_j \beta_j x_{ij}, \quad i = 1, \ldots, n \\ &= X_i \beta \end{aligned}$$

- Recall, previously, we let η_i be the $E(Y)$.
- However, this results in a linear regression model.
- If we want to minimize this dependency, let

$$\eta = f(E(Y))$$

  where $f(\cdot)$ is a function.

Link Component

Denote,

$$E(Y) = \mu_i$$

Let, $g(\cdot)$ be a monotonic and differentiable function such that

$$\eta_i = g(\mu_i)$$

Thus,

$$g(\mu_i) = \sum_j \beta_j x_{ij}, \quad i = 1, \ldots, N$$

In words, we are now modeling a function of the mean (or Expectation) as a combination of linear predictors.

Link Function

- The function $g(\cdot)$ is called the “link function” since it links the $E(Y_i)$ to the set of explanatory variables and their estimates.
- For example, if we let $g(x) = x$ (the identify link) and Y is distributed as a Normal R.V., then, we are back to the linear model (either simple linear or multiple linear regression)
- For GLM, you generally have the flexibility to choose what ever link you desire.
- However, there is a **Special** link that we need to consider

Canonical Link

$$f(y_i;\theta_i) = a(\theta_i)b(y_i)exp[y_iQ(\theta_i)]$$

If we revisit the density function for an exponential, we see a function $Q(\theta_i)$ that looks interesting.

$Q(\theta)$ is defined as the natural parameter of the distribution.

If we let $g(\cdot)$ be defined such that it transforms the mean to the natural parameter, we have the **Canonical link**

Example Canonical Link

Suppose

$$Y_i \sim Bern(\pi)$$

where $P(Y_i = 1) = \pi$ and $P(Y_i = 0) = 1 - \pi$

Then we previously showed that

$$f(y_i; \pi) \quad = \quad (1-\pi)(1)e^{(y_i \log \frac{\pi}{1-\pi})}$$

with $Q(\pi) = \log(\pi/(1-\pi))$

So, if we would let

$$g(\pi) = \log(\pi/(1-\pi)) = \sum_j \beta_j x_j$$

We would have the canonical link of a Bernoulli/Binomial distribution.

Recall, the function

$$g(\pi) = log(\pi/(1-\pi))$$

was previously introduced as the 'log odds' and was called the logit.

Lets recap what we have just accomplished.

If we let the random component be **Bernoulli/Binomial** and consider the linking function as **logit**, we can model the log odds ratio as a linear function of covariates using

$$g(\pi) = \sum_j \beta_j x_j$$

Since $g(\pi) = \log(\pi/1 - \pi)$, we can write the success probability as

$$\begin{array}{rcl} \frac{\pi}{1-\pi} & = & e^{X\beta} \\ \pi & = & e^{X\beta} - \pi e^{X\beta} \\ \pi(1 + e^{X\beta}) & = & e^{X\beta} \\ \pi & = & \frac{e^{X\beta}}{1+e^{X\beta}} \end{array}$$

What are the β's and the X's?

For the logistic model, we have

$$g(\pi) = \log(\pi/(1-\pi)) = \sum_j \beta_j x_j$$

To answer the question, consider a model in which you have one predictor (i.e., treatment) and you observe the response MI/No MI.

Let

$$x_{1i} = \begin{cases} 1 & \text{if subject i received the active treatment} \\ 0 & \text{else} \end{cases}$$

and

$$X_i = [1, x_{i1}]$$

Thus, if subject i is on active drug,

$$X_i = [1, 1]$$

and if on placebo

$$X_i = [1, 0]$$

Then, the odds for a person on placebo would be

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1\cdot 0} = e^{\beta_0}$$

and for a subject on active drug, the log odds would be

$$\frac{\pi}{1-\pi} = e^{\beta_0+\beta_1\cdot 1} = e^{\beta_0+\beta_1}$$

Thus, the odds ratio of a success for comparing active treatment to placebo could be written as

$$OR = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

or that $log(OR) = \beta_1$.

If you recall, we introduced this notation when we introduced RD, RR and OR.

Other Choices in the Link

An implied *advantage* of the GLM formulation is that you can specify other links to derive additional parameter interpretations.

For example, suppose you used the “log” link ($log(\pi) = X\beta$) instead of the “logit” link.

Now, the log link is not the canonical link, but that is OKAY. Then,

$$log(\pi) = \beta_0 + \beta_1 x_{1i}$$

or

$$\pi = e^{\beta_0 + \beta_1 x_{1i}}$$

Therefore, the RR could be viewed as

$$RR = \frac{\pi|x=1}{\pi|x=0} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

β_1 can now be interpreted as log Relative Risk.

example

Recall our famous MI example.

	Myocardial Infarction	
	Fatal Attack or Nonfatal attack	No Attack
Placebo	189	10845
Aspirin	104	10933

Previously, we estimated the OR to be

$$OR = (189 \cdot 10933)/(104 \cdot 10845) = 1.832$$

which indicates that subjects taking placebo had 1.8 times the odds of having an MI when compared to subjects taking aspirin.

Now using a GLM

For this analysis, we want to use the aspirin group as the reference group and estimate.

Therefore in terms of our regression dummy codes, we want

$$x_{1i} = \begin{cases} 1 & \text{if subject i received PLACEBO} \\ 0 & \text{if subject i received ASPIRIN} \end{cases}$$

with the response coding of

$$Y_i = \begin{cases} 1 & \text{if subject i has either a Fatal MI or a Non Fatal MI} \\ 0 & \text{if subject i does not have an MI} \end{cases}$$

Inputting Data

And we could input this data into SAS as

```
data one;
 input y x1 count;
 cards;
1 1 189
1 0 104
0 1 10845
0 0 10933
;
run;
```

Modeling Fitting Using SAS

And use PROC GENMOD (**gen**eralized linear **mod**els) to fit the data

```
proc genmod descending;
 freq count;
 model y = x1 /dist = bin link=logit;
 estimate ’X1’ x1 1 /exp;
run;
```

Notes:

1. We have specified our design matrix to include just X1. GENMOD automatically includes an intercept unless you tell it not to.
2. We used “dist=bin” to specify the distribution of Y as binomial
3. We used “link = logit” to fit the canonical link (logistic link)
4. The estimate statement invokes a contrast and exponentiates the parameter estimates

Don’t worry, we’ll become very familiar with GENMOD over the next few weeks.

Selected Results

```
     Response Profile

 Ordered             Total
   Value    y    Frequency

       1    1          293
       2    0        21778

PROC GENMOD is modeling the probability that y=’1’.
```

Note: The most important line is the one that indicates what level of the response is considered a success. In this case, we used “DESCENDING” to specify y=1 as the success. (by default, SAS takes the first sorted response category as the success)

```
                            Analysis Of Parameter Estimates

                               Standard     Wald 95% Confidence
Parameter    DF    Estimate       Error             Limits

Intercept     1     -4.6552      0.0985     -4.8483     -4.4620
x1            1      0.6054      0.1228      0.3647      0.8462

                                Contrast Estimate Results

                       Standard
Label      Estimate       Error     Alpha      Confidence Limits

X1           0.6054      0.1228      0.05      0.3647      0.8462
Exp(X1)      1.8321      0.2251      0.05      1.4400      2.3308
```

Therefore our estimate of OR = 1.832 with a 95% CI of (1.44, 2.33).

Using PROC LOGISTIC

```
proc logistic descending; freq count; model y = x1;
*estimate X1 x1 1 /exp;
run;

THE OUTPUT

Analysis of Maximum Likelihood Estimates

                                  Standard   Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept     1      -4.6551      0.0985     2232.4885        <.0001
x1            1       0.6054      0.1228       24.2911        <.0001


Odds Ratio Estimates

           Point          95% Wald
Effect    Estimate      Confidence Limits

x1            1.832       1.440       2.331
```

Difference? GENMOD is moment-based, but LOGISTIC uses ML!

Deviance

For a particular GLM for observations $y = (y_1, \ldots, y_n)$, let

$$l(\mu, y)$$

denote the log likelihood function expressed in terms of the means $\mu = (\mu_1, \ldots, \mu_n)$

Let

$$l(\hat{\mu}, y)$$

denote the maximum of the log likelihood for the model.

If we have n observations and fit a model with n parameters, we have a saturated model.

We then have a ‘perfect fit’ of the data (i.e., no degrees of freedom).

Denote the likelihood under the saturated model as

$$l(y, y)$$

Then, the DEVIANCE of a model is

$$D = -2[l(\hat{\mu}, y) - l(y, y)]$$

Note:

1. D is distributed as χ^2 with $df = N - p$
2. p is the number of parameters estimated under the alternative (or fitted model).
3. Recall, N in the saturated model is the number of parameters included (one for each observation).
4. Therefore, using the rules for calculating the df of a contingency table we developed earlier, df equals the difference in parameters estimated under the null (saturated model) and the alternative (at least one β not equal to zero)
5. For contingency tables $D \equiv G^2$
6. As we proceed, the Deviance will be used to provide a measure of model fit.

Calculation of Deviance by Hand -Okay, not really

If you recall, we used a **Poisson log linear** model to calculate the Pearson residuals for a contingency table.

A saturated model for a contingency table is one that contains an interaction term. For

example;

```
proc genmod;
  model count = x1 y x1*y /dist=poi link = log;
run;
```

produces a saturated model.

Proof

The design matrix for this model would be

$$X = [1, x_{1i}, y_i, x_{1i}y_i]$$

Since $x_{1i} = y_i = (0, 1)$
cell (1,1) has $X = (1, 1, 1, 1)$
cell (1,2) has $X = (1, 1, 0, 0)$
cell (2,1) has $x = (1, 0, 1, 0)$
cell (2,2) has $x = (1, 0, 0, 0)$

That is, counts for all cells are determined by a combination of $X\beta$.

GENMOD Results

```
           Criteria For Assessing Goodness Of Fit

Criterion                 DF             Value          Value/DF

Deviance                   0            0.0000               .
Scaled Deviance            0            0.0000               .
Pearson Chi-Square         0            0.0000               .
Scaled Pearson X2          0            0.0000               .
Log Likelihood                     181840.4662
```

Note: $df = 0$ since we are fitting the saturated model ($df = N - N$)

$l(y, y) = 181840.4662$

```
                              Analysis Of Parameter Estimates

                                         Standard     Wald 95% Confidence
Parameter     DF     Estimate            Error              Limits

Intercept      1       9.2995           0.0096       9.2808        9.3183
x1             1      -0.0081           0.0136      -0.0346        0.0185
y              1      -4.6552           0.0985      -4.8483       -4.4620
x1*y           1       0.6054           0.1228       0.3647        0.8462
```

Therefore, cell (1,1)'s count would be

$$\begin{aligned} log(\text{count cell 1,1}) &= 9.2995 - 0.0081 - 4.6552 + 0.6054 \\ &= 5.2416 \\ & \quad \text{or} \\ \text{count cell 1,1} &= e^{5.2416} \\ &= 189 \end{aligned}$$

The Alternative Model

```
proc genmod;
 model count = x1 y /dist=poi link = log;
run;
```

```
           Criteria For Assessing Goodness Of Fit

Criterion                    DF            Value        Value/DF

Deviance                      1          25.3720         25.3720
Scaled Deviance               1          25.3720         25.3720
Pearson Chi-Square            1          25.0139         25.0139
Scaled Pearson X2             1          25.0139         25.0139
Log Likelihood                       181827.7802
```

$l(\hat{\mu}, y) = 181827.7802$

```
                            Analysis Of Parameter Estimates

                               Standard     Wald 95% Confidence
Parameter    DF    Estimate       Error           Limits

Intercept     1      9.2956      0.0096      9.2769      9.3144
x1            1     -0.0003      0.0135     -0.0267      0.0261
y             1     -4.3085      0.0588     -4.4238     -4.1932
```

Thus, the expected cell count for cell (1,1) would be

$$\begin{aligned} log(\text{expected count cell 1,1}) &= 9.2956 - 0.0003 - 4.3085 \\ &= 4.9868 \\ &\quad \text{or} \\ \text{count cell 1,1} &= e^{4.9868} \\ &= 146.467 \end{aligned}$$

This is the same values as (with some rounding error) $293 * 11034/22071 = 146.48$

Therefore,

$$\begin{aligned} D &= -2(181827.7802 - 181840.4662) \\ &= 25.37 \end{aligned}$$

on $df = 4 - 3 = 1$

Summary of Canonical Links

Distribution	Natural Parameter	Canonical Link
Poisson	log(λ)	log
Normal	μ	identity
Binomial	log($\pi/(1-\pi)$)	logit

As stated before, just because the natural parameter suggests a certain link, there is no requirement to model only using the canonical link.

Recap

Some key summary points:

1. We generalized linear models by allowing for the specification of the distribution of Y and the relationship of the expectation to the design matrix
2. We do not need normality for the regression model
3. GLMs provide a unified theory of modeling that encompasses most of the important models for continuous and discrete variables
4. As we will see next, model parameters can be estimated by ML
5. By restricting the distributions to only exponential class distributions, we can use the same algorithm for ML estimation