Applied Multivariate and Longitudinal Data Analysis

Longitudinal Data Analysis: Review of Generalized Linear Model (scalar case)

Ana-Maria Staicu

SAS Hall 5220; 919-515-0644; astaicu@ncsu.edu

1 Introduction

In the previous chapters we focused on methods for analyzing longitudinal data where

- the response variable is continuous with values ranging over the real line;
- the vector of subject-level responses is assumed to have (exactly or approximately) a multivariate normal distribution.

In this chapter, we consider the case where the vector of subject-level responses cannot be modeled using a normal distribution.

Examples: the response is binary and takes only values 0("failure")/1 ("success"); the response is a "count" $(0, 1, 2, \ldots)$ but the values are relatively small, etc.

We refer to these types of responses by the name "generalized" responses. The models used to analyze generalized responses, analogous to linear models, are called *generalized linear models*. We begin with a review of the generalized linear models for scalar responses and then discuss these class of models for repeated measures.

Data Example. Consider the following well known data set (Hand et al. , 1994). The numbers of Prussian militiamen killed by being kicked by a horse in each of 10 separate corps of militiamen are measured between 1875 – 1894. Here is a snapshot of the data:

Table 1:

Obs	Year	Corps	Number of men killed
1	1875	1	0
2	1875	2	0
3	1875	3	0
4	1875	4	0
5	1875	5	1
6	1875	6	1
7	1875	7	0
8	1875	8	0
9	1875	9	1
10	1875	10	0
11	1876	1	0
⋮	⋮	⋮	⋮

Some questions of interest:

- Are the differences in the number of men killed attributed to systematic effects of year or corps?

- How to assess formally that two particular years, say 1875 and 1880, have the same effect on the number of horse kick deaths?

2 Generalized linear models: scalar response

Generalized linear models extend the methods of regression analysis to settings where the outcome is dichotomous (binary variable), count etc. They share many of the characteristics of linear models; most notably the fact that a linear combination of the covariates is related to the mean response. They differ from the linear model in couple a of ways including the fact that the distribution of the response is not normal. The distribution of the response is assumed to be in the *exponential family*. The exponential family class is a very wide class of distribution and includes the normal distribution, Bernoulli distribution, Poisson, Gamma etc.

Setting
Denote the observed data by: $[Y_i, X_{i1}, \ldots, X_{iK}]$ for $i = 1, \ldots, n$ where Y_i is the scalar response and $X_{i1}, \ldots, X_{iK}$ are covariates. It is assumed that Y_i is in the exponential family, and Y_i's are independent over i. The *generalized linear models* are specified by the following THREE main parts:

1) distributional assumption of Y_i

2) modeling the systematic component (i.e. it describes the manner in which the covariates affect the mean response)

3) specification of the link function (i.e. links the mean response to the systematic component).

1. The distributional assumption. It is assumed that the outcome Y_i follows a distribution that belongs to the exponential family. Examples:

- $Y_i \sim Bernoulli(p_i)$.

$$P(Y_i = y) = p_i^y (1 - p_i)^{1-y}.$$

What are the mean and variance?

- $Y_i \sim Poisson(\lambda_i)$.

$$P(Y_i = y) = \frac{e^{-\lambda_i} \lambda_i^y}{y!}$$

What are the mean and variance?

- $Y_i \sim N(\mu_i, \sigma^2)$.

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu_i)^2}$$

What are the mean and variance?

The exponential family models are denoted by $EF(\eta_i, \phi)$, where η_i is related to the mean of Y_i and ϕ_i is called *dispersion parameter* or *scale parameter* and is related to the variance of Y_i, that is not captured by η_i.

2. The systematic component. The systematic component specifies that the effect of the covariates $X_{i1}, \ldots, X_{iK}$ on the mean response Y_i can be expressed in terms of the following linear predictor

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_K X_{iK}$$

or in vector format $\eta_i = X_i^T \beta$, where X_i is the $(K+1)$-dimensional column vector of X_{il}'s with 1 as the first element, and $\beta = (\beta_0, \beta_1, \ldots, \beta_K)^T$ is the $(K+1)$-dimensional column vector of β_l's. The parameter η_i is called the *linear predictor*. The parameter β is called *regression parameter*.

3. The link function The *link function* is a function that links a transformation of the mean response to the linear predictor. Denote the mean response $\mu_i = E[Y_i]$; then the common notation for the link function is: $g(\mu_i) = \eta_i$.
The link function is known and assumed monotone and differentiable over the domain of μ_i.

Examples:

- Identity link (common for normal responses), $g(x) = x$:
- Logistic link (common for binary responses $0/1$), $g(x) = log\frac{x}{1-x}$:
- Probit link (used for binary responses) $g(x) = \Phi^{-1}(x)$ where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of a standard normal variable $N(0,1)$.
- Log link (used for counts responses) $g(x) = log(x)$

Illustration:

A. *Logistic regression model*: The observed data are $\{Y_i, X_i\}_{i=1}^n$. Assume a logistic model for the response: $Y_i \sim Bernoulli(p_i)$ with $log\{p_i/(1-p_i)\} = \eta_i$ where $\eta_i = \beta_0 + \beta_1 X_i$.
With your stats buddy, identify the three model components of this GLM.

1. *Odds ratio of success* or simply *odds* are defined as:

$$\frac{P(Y_i = 1)}{P(Y_i = 0)}$$

   and describe the likelihood of success relative to failure. Eg: An odds ratio of 4 to 1 means that the probability of success is 4 times larger than the probability of failure. Thus, the logistic model relates the log odds to the covariates.

2. Draw a plot (approx) of the probability of success $P(Y_i = 1|X = x)$ versus x (assume $\beta_1 > 0$).

3. Assume X_i is continuous. Describe the interpretation of β_0 and β_1

B. *Log-linear model for counts*: This model is often used when the response is

- counts of the number of times some event occurs in either time or space (e.g. the number of militiamen killed by horse kicks during a year)
- rates at which an event occurs (e.g. the number of epileptic seizures in 4-weeks interval or in 8-weeks interval; or counts over different size of groups, or crops infected by virus over regions of different areas). The absolute number of events (count) is sometimes not satisfactorily because it refers to different 'times at risks' (interval of time, or size of groups, or areas of regions).

The primary objective of the log-linear regression is to relate the expected counts/rates to a set of covariates.

Consider the observed data: $\{Y_i, X_i, T_i\}_{i=1}^n$; the response Y_i has values $0, 1, 2, \ldots$ and X_i is scalar and T_i is related to the time during which Y_i is observed. Assume the model:

$$Y_i \sim Poisson(\mu_i) \text{ and } \log\left(\frac{\mu_i}{T_i}\right) = \beta_0 + \beta_1 X_i,$$

Identify the three model components of this GLM.

- The "covariate" T_i is the *time at risk*, and is known as *offset*.
- Normally we have $var(Y_i) = \mu_i$ (recall Y_i follows Poisson distribution). However in many biomedical applications the count data have variability that exceeds that which is predicted by Poisson model. To account for this, an overdispersion parameter is introduced:

$$var(Y_i) = \phi\mu_i$$

where ϕ is the overdispersion parameter. This parameter is estimated from the data.

Failure to account for overdispersion results in standard error for estimated parameters that are smaller than they should be, and implicitly in smaller (hence incorrectly specified) p-values.

In the following let's assume that the response is calculated over the same duration (time at risk); in this case it is common to assume $T_i = 1$ for all i. Interpret the regression coefficients β_0 and β_1. Assume that X_i is continuous.

3 Parameter estimation: Maximum Likelihood

Combining the parts 1, 2 and 3 the generalized linear model (GLM) is specified by:

$$Y_i \sim EF(\eta_i, \phi), \qquad g\left(E[Y_i]\right) = \eta_i, \qquad \eta_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_K X_{iK}$$

or in short $g\left(E[Y_i]\right) = X_i^T \beta$, for the known and monotone link function g and unknown regression parameter β; here ϕ is a dispersion parameter and is considered a nuisance parameter. In this section we discuss ways to estimate β.

As GLMs specify a distribution for the response Y_i, maximum likelihood estimation (MLE) is used to estimate the model parameters.

In-class: Assume a logistic model for the response: $Y_i \sim Bernoulli(p_i)$ with $log\{\mu_i/(1-\mu_i)\} = \eta_i$ where $\eta_i = X_i^T\beta$. Here we used the parameterization $\mu_i = P(Y_i = 1)$
Let's write the likelihood function of β. The maximum likelihood estimate (MLE) of β is obtained by setting the first derivative of the log-likelihood function to zero.

- The likelihood function for this model, $L(\beta) = \prod_{i=1}^n P(Y_i = y_i|x_i) = \prod_{i=1}^n \mu_i^{y_i}(1-\mu_i)^{(1-y_i)}$. Recall that

$$\mu_i = \frac{\exp\{X_i^T\beta\}}{1+\exp\{X_i^T\beta\}}$$

- The log-likelihood function for this model is:

- Denote by $\mathcal{D}(\beta)$ the gradient of the log-likelihood function (the vector of the first derivatives of the likelihood function with respect to all components of β). The MLE of β are among the solutions of $\mathcal{D}(\beta) = 0$. Calculate $\mathcal{D}(\beta)$.

The equations described by $\mathcal{D}(\beta) = 0$ can be written for the general case where the data are $\{Y_i, X_i \sim R^k\}$ and $Y_i \sim EF(\mu_i, \phi)$ as:

$$\sum_{i=1}^{n} \frac{1}{v(\mu_i)}(y_i - \mu_i)\left(\frac{\partial \mu_i}{\partial \beta}\right) = 0;$$

they are also know as the *estimating equations* for the regression parameter β. This is because they are used to estimate the unknown parameter.

Remarks. Before discussing how these estimating equations are solved, let's analyze them:

- Term $(y_i - \mu_i)$ represents the deviation of y_i from its mean;
- Term $v(\mu_i)$ is the variance of y_i by ignoring the scale parameter.
- The regression parameter β appears in both the mean and the weight $(1/v(\mu_i))$.
- Even when $\partial \mu_i / \partial \beta$ is constant as function of β it may be still very complicated to get close form expression for the estimate $\widehat{\beta}$.

This optimization problem is not easy to solve, because it involves (in general) a nonlinear system of equations in β. For that reason, an iterative method is needed. One simple way to understand this procedure is from the sequence of iterations called *iterative re-weighted least squares* (IRWLS), Nelder and Wedderburn (JRSSA 1972):

- Start with an initial value (guess) $\widehat{\beta}^{(0)}$ using the data and the distribution Y_i;
- Compute the weights $v_i^{(0)} = v(\mu_i(X_i^T \widehat{\beta}^0))$

- Estimate $\widehat{\beta}^{(1)}$ such that
$$\sum_{i=1}^{m} \frac{1}{v_i^0}\{y_i - \mu_i(X_i^T\beta)\}\frac{\partial \mu_i}{\partial \beta} = 0$$
- Compute new weights $v_i^{(1)} = v\{\mu_i(X_i^T\widehat{\beta}^{(1)})\}$ and repeat until there is not much variation in the estimates $\widehat{\beta}^{(l)}$'s.

To gain insight into the solution, we consider alternatives which explicitly provide the solution at each iteration. One common method is the *Newton-Raphson algorithm*, a generalization of the Newton algorithm for the multidimensional parameter. Denote $\mathcal{H}(\beta)$ the Hessian of the log-likelihood, i.e. the matrix of second derivatives with respect to all components of β. Then, one Newton-Raphson iteration step is

$$\widehat{\beta}^{new} = \widehat{\beta}^{old} - \{\mathcal{H}(\widehat{\beta}^{old})\}^{-1}\mathcal{D}(\widehat{\beta}^{old}).$$

A variant of the Newton-Raphson is the *Fisher scoring algorithm* which replaces the Hessian by its expectation (w.r.t. the observations)

$$\widehat{\beta}^{new} = \widehat{\beta}^{old} + [E\{\mathcal{H}(\widehat{\beta}^{old})\}]^{-1}\mathcal{D}(\widehat{\beta}^{old}).$$

In particular while $\mathcal{H}(\beta)$ has a cumbersome expression, note that its expected value is

$$E\mathcal{H}(\beta) = \sum_i \left[\frac{\{g'(X_i\beta)\}^2}{v(\mu_i)}\right] X_i^T X_i.$$

Define $W = diag[\{g'(X_1^T\beta)\}^2/v(\mu_1), \ldots, \{g'(X_n^T\beta)\}^2/v(\mu_n)]$ and

$$\tilde{Y} = \left(\frac{Y_1 - \mu_1}{g'(X_1^T\beta)}, \ldots, \frac{Y_n - \mu_n}{g'(X_n^T\beta)}\right)^T.$$

Also let X be the $n \times K$ matrix obtained by row-stacking X_i^T's. The solution provided by the *Fisher scoring algorithm* at the $(l+1)$ step is

$$\widehat{\beta}^{(l+1)} = \widehat{\beta}^{(l)} + (X^TWX)^{-1}X^TW\tilde{Y} = (X^TWX)^{-1}X^TWZ$$

where $Z = (Z_1, \ldots, Z_m)$ is the vector of adjusted dependent variables $Z_i = X_i^T\widehat{\beta}^{(l)} + \widetilde{Y}_i$

$$Z_i = X_i^T\widehat{\beta}^{(l)} + \frac{Y_i - \mu_i}{g'(X_i^T\widehat{\beta}^{(l)})}$$

and Z_i is calculated for the current value of β, i.e. $\widehat{\beta}^{(l)}$; thus is adjusted at each iteration.

Denote by $\widehat{\beta}$ the resulting parameter estimate.

An estimate for the scale (dispersion) parameter ϕ is:

$$\widehat{\phi} = \frac{1}{n-K}\sum_i \frac{(y_i - \widehat{\mu}_i)^2}{v(\widehat{\mu}_i)};$$

this estimate is often referred to as the Pearson chi-square estimate divided by the degrees of freedom.

4 Inference for the regression parameter

For large n, the estimator satisfies

$$\widehat{\beta} \sim N_K\{\beta, \phi(\Delta^T V^{-1}\Delta)^{-1}\}$$

- ϕ is the scale parameter
- Δ is $n \times K$ matrix with rows equal to $\partial\mu_i/\partial\beta$
- V is $n \times n$ diagonal matrix with elements $v(\mu_i)$

The variance-covariance matrix of the estimator $\widehat{\beta}$ can be approximated by plugging in estimates for the unknown parameters,

$$\widehat{V}_{\widehat{\beta}} = \widehat{\phi}(\widehat{\Delta}^T\widehat{V}^{-1}\widehat{\Delta})^{-1}.$$

Hypotheses tests of the form $H_0 : L\beta = h$ can be carried using Wald test. In particular notice that

$$L\widehat{\beta} \sim N_r(L\beta, L\widehat{V}_{\widehat{\beta}}L^T);$$

thus the construction of the test statistics is similar to that discussed in the previous chapters.

5 Goodness of fit

Historically the *deviance* has played a major role in assessing the goodness of fit of the models, in particular in generalized models. In generalized linear models it has a similar role to the residual variance from ANOVA in linear models (residual sum of square, RSS). It can be used to test the fit of the link function and linear predictor to the data, or to test the significance of a particular predictor variable (or variables) in the model.

The deviance compares the likelihood of a "saturated" model, which is defined using the same distribution and link function as the model of interest but with $g(\mu_i) = \psi_i$ for all i, and the current model, where $g(\mu_i) = X_i^T\beta$. We can think of the saturated model as having the most general possible mean structure for the data since the means $|_i$ are unconstrained; hence the names "full model" or "maximal model" that the saturated model is also referred by. Mathematically, the deviance is defined as

$$D = 2\widehat{\ell}_S - 2\ell(\widehat{\beta});$$

where $\widehat{\beta}$ are the maximum likelihood estimates of the regression parameters. Under some regularity conditions, if the proposed model describes the data nearly as well as the saturated model, then asymptotically

$$D \sim \chi^2_{n-K}$$

where n is the number of subjects in the data and K is the number of regression parameters. If the proposed model is poor, D will be larger than predicted by the χ^2_{n-K} distribution. A version of deviance, the *scaled deviance* is also used: D/ϕ, where ϕ is the dispersion parameter.

Another statistic used to assess the goodness of fit is Pearson's chi-square (χ^2) statistic.

$$\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)^2}{v(\mu_i)};$$

where μ_i is the mean of Y_i and $v(\mu_i)$ is the variance function,

Residuals represent the difference between the data and the model. For Gaussian data the residuals are defined as $y - \widehat{\mu}$; in GLM the residuals are defined in several ways. Pearson residuals are calculated as

$$r_P = \frac{y - \widehat{\mu}}{\sqrt{V(\widehat{\mu})}}$$

Deviance residuals are defined as $sign(y - \widehat{\mu})\sqrt{d_i}$, where $D = \sum_i^n d_i^2$ is the deviance. Plotting the residuals versus the fitted values allows to make sure that there is no deterministic part left unexplained in the data.

6 Fitting GLM in R

Generalized linear models are fit in `R` using the function `glm()`. The formula is:

```
glm ( formula, family=binomial, data, offset)
```

Extracting the deviance residuals in R is using the function residuals().