Generalized Linear Models

Biostatistics 653

Applied Statistics III: Longitudinal Analysis

Generalized Linear Models

- We have focused on methods for analyzing longitudinal data in which the outcome was continuous and the vector of responses was assumed to follow the multivariate normal distribution. In addition, we fit a linear model to the repeated measurements.
- We now consider a more general class of regression models, the class of generalized linear models. The linear regression model is a member of this class, as are the logistic and Poisson regression models.
- We will review the generalized linear model (for univariate data) and its properties before applying it in the longitudinal data setting.

Generalized Linear Models

- In many applications, the distribution of a continuous response may be non-normal.
- In addition, the response may be discrete

  -- dichotomous or binary $(Y_i = 1, Y_i = 0)$

  -- ordered categorical $(Y_i \in \{1, \cdots, C\})$, with the ordering of the index important

  -- unordered categorical or nominal $(Y_i \in \{1, \cdots, C\})$, with the ordering of the index unimportant

  -- count $(Y_i \in \{0,1,2, \cdots, \infty\})$
- Finally, a non-linear regression model relating the predictors to the mean may be needed.

Generalized Linear Models

- The generalized linear model extends the methods of regression analysis to settings in which the outcome variable may follow a Bernoulli, Poisson, gamma, or other distribution in the exponential family.
- This model has some, but not all, properties of the linear model. One important shared property is that a parameter related to the expected value of the response is assumed to depend on a linear function of the covariates.
- The value of the GLM lies in the ability to develop techniques, statistics, and properties for the entire group simply based on the form of the likelihood.

Generalized Linear Models

- Let $Y_i, i = 1, \cdots, N$, be independent responses. The probability model for Y_i has three components:
  1. the distributional assumption,
  2. the systematic component, and
  3. the link function.
- Y_i is assumed to have a probability distribution that belongs to the exponential family. The general form for this family of distributions is given by

$$f(Y_i) = \exp[\frac{Y_i\theta_i - b(\theta_i)}{a(\phi)} - c(Y_i, \phi)]$$

where θ_i is the canonical parameter, ϕ is a scale parameter, and a(.), b(.) and c(.) are known functions.

The Distributional Assumption

- The exponential family of distributions includes the normal, Bernoulli, Poisson, gamma, beta, chi-square, multinomial, geometric, and others.

- Popular distributions not in the exponential family include the uniform distribution and Student's t distribution.

- To fix ideas, let’s consider a few examples of exponential family distributions.

Example: The Normal Distribution

- The normal distribution is in the exponential family, as it can be written as

$$f(Y_i|\mu,\sigma^2) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{(Y_i-\mu_i)^2}{2\sigma^2}\right]$$

$$= \exp[\frac{Y_i\mu_i - \frac{1}{2}\mu_i^2}{\sigma^2} - \frac{1}{2}(\frac{Y_i^2}{\sigma^2} + \log(2\pi\sigma^2))]$$

- So that $\theta_i = \mu_i, b(\theta_i) = \frac{\theta_i^2}{2}, a(\phi) = \sigma^2$, and $c(Y_i, \phi) = \frac{1}{2}(\frac{Y_i^2}{\sigma^2} + \log(2\pi\sigma^2))$.

Example: The Bernoulli Distribution

- The Bernoulli distribution is in the exponential family, as it can be written as

$$f(Y_i|\pi_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$
$$= \exp[Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i)]$$

- So that $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = logit(\pi_i), b(\theta_i) = -\log(1-\pi_i)$, $a(\phi) = 1$, and $c(Y_i, \phi) = 0$.

Example: The Poisson Distribution

- The Poisson distribution is in the exponential family, as it can be written as

$$f(Y_i|\lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{Y_i}}{Y_i!}$$
$$= \exp[Y_i \log \lambda_i - \lambda_i - \log(Y_i!)]$$

- So that $\theta_i = \log \lambda_i$, $b(\theta_i) = \lambda_i, a(\phi) = 1$, and $c(Y_i, \phi) = \log(Y_i!)$.

Log-likelihood

- Based on the form of the distribution, it is easy to see that the log-likelihood is given by

$$l(\mu_i, \phi) = \frac{Y_i\theta_i - b(\theta_i)}{a(\phi)} - c(Y_i, \phi)$$

- It can be shown that

$$E(Y_i) = b'(\theta_i), V(Y_i) = b''(\theta_i)a(\phi)$$

where $b'(\theta_i) = \frac{\partial b(\theta_i)}{\partial \theta_i}$ and $b''(\theta_i) = \frac{\partial^2 b(\theta_i)}{\partial \theta_i^2}$

- Often, we define $\mu_i = E(Y_i) = b'(\theta_i)$ for the expected value

Log-likelihood

- It follows that the derivatives of the log-likelihood are given by

$$\frac{\partial l}{\partial \theta_i} = \frac{Y_i - b'(\theta_i)}{a(\phi)}, \frac{\partial l}{\partial \theta_i^2} = -\frac{b''(\theta_i)}{a(\phi)}$$

- To find the mean, we set

$$0 = E\left(\frac{\partial l}{\partial \theta_i}\right) = \frac{E(Y_i) - b'(\theta_i)}{a(\phi)}$$

which implies that $\mathrm{E}(\mathrm{Y_i}) = b'(\theta_i)$

Log-likelihood

- Similarly, we have

$$0 = E\left(\frac{\partial^2 l}{\partial \theta_i^2}\right) + E\left(\frac{\partial l}{\partial \theta_i}\right)^2 = -b''(\theta_i)a(\phi) + E\left(Y_i^2\right) - E(Y_i)^2$$

- Therefore, $V(Y_i) = b''(\theta_i)a(\phi)$

Log-likelihood

- For most commonly used exponential family distributions, $a(\phi) = \phi/w_i$, where ϕ is a dispersion parameter and w_i is a weight (typically equal to one).

- Hence, the mean and variance will typically follow the form:

$$\mu_i = b'(\theta_i), \sigma^2 = b''(\theta_i)\phi$$

- These simple formulae may be used to derive the mean and variance of exponential family distributions.

Log-likelihood

- For example, for the normal distribution, we have $\theta_i = \mu_i, b(\theta_i) = \mu_i^2/2$ and

$$E(Y_i) = b'(\theta_i) = \mu_i, V(Y_i) = b''(\theta_i)a(\phi) = 1 \times \sigma^2 = \sigma^2$$

- For the Bernoulli distribution, we have $\theta_i = logit(\pi_i), b(\theta_i) = -\log(1-\pi_i)$, $a(\phi) = 1$, and $\pi_i = \frac{e^{\theta_i}}{1+e^{\theta_i}}$. Therefore, we have

$$E(Y_i) = b'(\theta_i) = \pi_i, V(Y_i) = b''(\theta_i)a(\phi) = \pi_i(1-\pi_i)$$

Systematic Component

- The systematic component or linear predictor, η_i, describes the effect of the covariates on the expected value of Y_i. In particular, given covariates $X_{i,1}, \cdots, X_{i,p-1}$, the linear predictor is given by

$$\eta_i = \beta_0 + \sum_{k=1}^{p-1} X_{ik}\beta_k = X_i\beta$$

The Link Function

- Instead of modeling the mean, μ_i, as a linear function of predictors, X_i, we introduce on one-to-one continuously differentiable transformation g(.) and focus on

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

where g(.) is called the link function and η_i the linear predictor.

- We assume that the transformed mean follows a linear model,

$$\eta_i = X_i\beta$$

- Since the link function is invertible and one-to-one, we have

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(X_i\beta)$$

The Link Function

- Note that we are transforming the expected value, μ_i, instead of the raw data, Y_i.
- For classical linear models, the mean is the linear predictor. In this case, the identity link is reasonable since both μ_i and η_i can take any value on the real line.
- This is not the case in general.
- When $\theta_i = \eta_i$, then we say we are using the canonical link function.

Example: Normal Linear Regression

- The linear regression model assumes the normal distribution for Y_i, has linear predictor $\eta_i = X_i\beta$ (specification depends on covariates of interest), and assumes the identity link function $g(\mu) = \mu$, so that we have

$$E(Y_i) = X_i\beta$$

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

- The identify link is the canonical link here, as $E(Y_i) = \mu_i = \eta_i = X_i\beta$.

Example: Binary Regression

- The logistic regression model assumes the Bernoulli distribution for Y_i, has linear predictor $\eta_i = X_i\beta$, and assumes the logit link function $g(\mu_i) = \eta_i = logit(\pi_i)$, which is the canonical link function.
- For the binomial distribution, $0 < \mu_i < 1$ (mean of Y_i is $n_i\mu_i$). Therefore, the link function should map from $(0,1) \rightarrow R$.
- Other possibilities for the link function include the probit link, $g(\mu_i) = \eta_i = \Phi^{-1}(\pi_i)$ and the complementary log-log link $g(\mu_i) = \eta_i = \log[-\log(1-\pi_i)]$.
- The choice of link function can affect inference. Usually the logistic link and probit links look similar to each other. While the complementary log-log often differs from the other two, you would need a large sample size to discriminate between the links.

Example: Poisson Regression

- The Poisson regression model assumes the Poisson distribution for a count Y_i, has linear predictor $\eta_i = X_i\beta$, and often assumes the log link function $g(\mu_i) = \eta_i = \log(\lambda_i)$, which is the canonical link function. With this link function, additive effects contributing to η_i are acting multiplicatively on λ_i.
- As a short summary:

	Normal	Poisson	Binomial	Gamma
Notation	$N(\mu_i, \sigma^2)$	$\text{Pois}(\mu_i)$	$\text{Bin}(n_i, \pi_i)$	$G(\mu_i, \nu)$
Range of Y_i	$(-\infty, \infty)$	$[0, \infty)$	$[0, n_i]$	$(0, \infty)$
Dispersion, ϕ	σ^2	1	$1/n_i$	ν^{-1}
Cumulant: $b(\theta_i)$	$\theta_i^2/2$	$\exp(\theta_i)$	$\log(1 + e^{\theta_i})$	$-\log(-\theta_i)$
Mean function, $\mu(\theta_i)$	θ_i	$\exp(\theta_i)$	$1/(1 + e^{-\theta_i})$	$-1/\theta_i$
Canonical link: $\theta(\mu_i)$	identity	log	logit	reciprocal

Estimation by Maximum Likelihood

- Frequentist inference for GLM's typically relies on the MLE and asymptotic approximations. Recall that Y_i follows a distribution in the exponential family. Considering a sample of N subjects, the log-likelihood is given by

$$l(\beta, \phi) = \sum_{i=1}^{N} \{ \frac{Y_i \theta_i - b(\theta_i)}{a(\phi)} - c(Y_i, \phi) \}$$

- To find maximum likelihood estimates, we maximize this log-likelihood. To do so, we set the score equations, obtained as the first derivatives of the log-likelihood, equal to zero.

Estimation by Maximum Likelihood

- Recall the chain rule $\frac{\partial b(\theta_i)}{\partial \beta} = \frac{\partial b(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta}$. In the GLM, by definition, $b'(\theta) = \mu$, so that we have $\frac{\partial b(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta} = \mu_i \frac{\partial \theta_i}{\partial \beta}$.
- We have the score $u(\beta)$ given by

$$u(\beta) = \frac{\partial l(\beta, \phi)}{\partial \beta} \propto \sum_{i=1}^{N} Y_i \frac{\partial \theta_i}{\partial \beta} - \frac{\partial b(\theta_i)}{\partial \beta} = \sum_{i=1}^{N} (Y_i - \mu_i)\frac{\partial \theta_i}{\partial \beta}$$

- For the canonical link $\theta_i = \eta_i$, the above reduces to

$$\sum_{i=1}^{N} (Y_i - \mu_i)X_i$$

- To get the MLE of β, we set the score equations equal to zero and solve.

Estimation by Maximum Likelihood

- Sometimes, the score equations can be represented as

$$\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right)^T Var^{-1}(Y_i)(Y_i - \mu_i)$$

- The above follows from

$$\frac{\partial \mu_i}{\partial \beta} = \frac{\partial \mu_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta} = \frac{\partial b'(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta} = b''(\theta_i)\frac{\partial \theta_i}{\partial \beta} = \frac{Var(Y_i)}{a(\phi)}\frac{\partial \theta_i}{\partial \beta}$$

Estimation by Maximum Likelihood

- In general, these score equations need to be solved iteratively, using numerical algorithms such as iteratively reweighted least squares, Newton-Raphson, or Fisher scoring. In some cases (logistic regression), ϕ is a known constant. In other cases, including the normal linear model, estimation of ϕ may also be required.
- Also, note that score equations are an example of an estimating function (more later).

- The variances of the estimated regression coefficients can be readily obtained from the log-likelihood as

$$Var(\hat{\beta}_{ML}) = -E\left\{\frac{\partial^2 l}{\partial\beta\partial\beta'}\right\}^{-1}$$

$$= \left\{\sum_{i=1}^{N}\left(\frac{\partial\mu_i}{\partial\beta}\right)^T Var^{-1}(Y_i)\left(\frac{\partial\mu_i}{\partial\beta}\right)\right\}^{-1}$$

Quasi-Likelihood

- An important property of the GLM family is that the score function $u(\beta)$ depends only on the mean and variance of Y_i. Because of this, the estimating equation given by setting the score function equal to zero may be used to estimate the regression parameters for any choices of link and variance function (even if they do not correspond to a particular member of the exponential family).
- In this case, we refer to the score as a quasi-score function because its integral with respect to β may not correspond to a proper likelihood function.
- This suggests an estimation approach in which we make assumptions about the link and variance functions without trying to specify the entire distribution of Y_i.
- Solving these quasi-likelihood equations leads to an estimate $\hat{\beta}_{QL}$ that has variance of the same form as the ML estimate of β.

Model Selection and Uncertainty

- In most data analyses, there is uncertainty about the model & you need to do some form of model comparison.
  - -- Select q out of p predictors to form a parsimonious model
  - -- Select the link function (e.g., logit or probit)
  - -- Select the distributional form (normal, gamma)

- We can use the AIC or BIC criteria to select variables, the link function or distribution.

Model Selection and Uncertainty

- For example, suppose that we have binary outcome data. There are many possible link functions - any smooth, monotone function mapping from $R \to [0,1]$ (i.e., cumulative distribution functions for continuous densities).

- We simulated data from a logistic regression model with $X_i = (1, dose_i)$, where $dose_i \sim Uniform(0,1), i = 1, \cdots, 100.$
- The parameters were chosen to $\beta = (-3, 5)$.
- We considered the logistic model.
- As an alternative to the logistic model, we considered the probit:

$$P(Y_i = 1|X_i) = \Phi(X_i\beta)$$

where $\Phi(z) = \int_{-\infty}^{z}(2\pi)^{-\frac{1}{2}}\exp(-\frac{z^2}{2})$ is the standard normal cdf.

- We also considered the complementary log-log model:

$$P(Y_i = 1|X_i) = 1 - \exp\{-\exp(X_i\beta)\}$$

Model Selection and Uncertainty

- Logistic regression:
$$\hat{\beta} = (-3.67, 6.14), AIC = 94.0211, BIC = 99.23145$$

- Probit regression:
$$\hat{\beta} = (-2.14, 3.60), AIC = 93.9993, BIC = 99.20964$$

- Complementary log-log regression:
$$\hat{\beta} = (-3.14, 4.39), AIC = 94.21893, BIC = 99.42927$$

- Low values of AIC & BIC are preferred, so we select the probit model (which happens to be wrong)

Model Selection and Uncertainty

- It is very often the case that many models are consistent with the data. This is particularly true when there is a large number of models in the list of plausible models.
- Ideally, substantive information can be brought to bear to reduce the size of the list. For example, certain models may be more consistent with biology.
- However, one is typically still faced with many possible models & concerned about sensitivity of inferences to the model chosen or selected by some algorithm.
- Given the selection bias that occurs, it may be better to focus *a priori* on a single model rather than run stepwise selection.
- However, better yet would be to account formally for model uncertainty (e.g., through Bayesian model averaging).