STATS 212: Generalized Linear Models
Lecture 6: Quasi-likelihood, overdispersion, and GLMM

Babak Shahbaba

UCI, Spring 2010

Quasi-likelihood

- Recall that for generalized linear models, the score function has the following general form

$$u_j(\beta) = \sum_i \frac{(y_i - \mu_i)x_{ij}}{Var(y_i)} \frac{\partial \mu_i}{\partial \eta_i}$$

- As we can see, the score function depends on the distribution of y only through μ_i and $Var(y_i)$ only.
- Moreover, $Var(\mu_i)$ itself is a function a μ_i such that $Var(y_i) = V(\mu_i)$, where V is called the variance function.
- For example, in the Poisson model, $V(\mu_i) = \mu_i$, in the Bernoulli model, $V(\mu_i) = \mu_i(1 - \mu_i)$, for the normal model $V(\mu_i) = \sigma^2$ (i.e., it is a constant function).
- Instead of fully specifying the distribution of y_i, we can specify the mean-variance relationship, i.e., the variance function, only. The resulting likelihood is called *quasi-likelihood* (QL).

Quasi-likelihood

- Note that we still need to define the link function $g(\mu_i) = \eta_i = x_i\beta$.
- The quasi-score function based on the quasi-lilihood model will be

$$u_j(\beta) = \sum_i \frac{(y_i - \mu_i)x_{ij}}{V(\mu_i)} \frac{\partial \mu_i}{\partial \eta_i}$$

- This means that the derivative of quasi-likelihood function (or more correctly, quasi-log-likelihood), ψ, should be defined as

$$\frac{\partial \psi(\mu, y)}{\partial \mu} = \frac{y - \mu}{V(\mu)}$$

- And the quasi-log-likelihood function for n independent observations is

$$Q(\mu) = \sum_i \psi(\mu_i, y_i)$$

Quasi-likelihood

- For example, if we define $V(\mu) = \mu$

$$\frac{\partial \psi(\mu, y)}{\partial \mu} = \frac{y - \mu}{\mu}$$

which means

$$\psi(\mu, y) = y \log(\mu) - \mu + c(y)$$

- In this case, the quasi-likelihood function has the same form as the likelihood for the Poisson model

$$L(\mu, y) = y \log(\mu) - \mu - \log(y!)$$

which was expected since we set the variance function equal to the mean.

Quasi-likelihood

- To estimate β based on QL function, we maximize the quasi-likelihood function.
- For this, similar to the ML approach, we set the partial derivatives of $Q(\beta)$ to 0.

$$\sum_i \frac{(y_i - x_i\hat{\beta})x_{ij}}{V(x_i\hat{\beta})} \frac{\partial \hat{\mu}_i}{\partial \hat{\eta}_i} = 0$$

- These are called *estimating equations*.
- The generalization of QL for multivariate responses is called generalized estimation equations (GEE).
- As before, we can use iterative weighted least squares to obtain $\hat{\beta}$.

Quasi-likelihood

- The QL estimators have the similar properties to those of *ML* estimators.
- Their sampling distribution is asymptotically normal with the following approximate covariance matrix

$$cov(\hat{\beta}) \quad = \quad (x'wx)^{-1}$$

where w is a diagonal matrix with diagonal elements

$$w_i \quad = \quad \frac{(\frac{\partial \mu_i}{\partial \eta_i})^2}{V(\mu_i)}$$

Overdispersion

- The mean-variance relationships defined by the Poisson and Bernoulli models are quite restrictive.
- In the Poisson model, the variance is restricted to be equal to the mean.
- In reality, however, the variance is usually larger than the mean for count data mainly due to the heterogeneity of observations.
- To account for this, we can utilize an additional parameter ν, called the *dispersion* parameter, and define the variance function as $Var(y) = \nu V(\mu)$, where ν is usually unknown.
- For example, in the Poisson model, $Var(y) = \nu\mu$, and in the Bernoulli model, $Var(y) = \nu\mu(1 - \mu)$.

Overdispersion

- Note that the dispersion parameter in the above setting drops out in the estimating equations, so the estimated $\hat{\beta}$ are the same as ML estimates.
- However, the covariance of $\hat{\beta}$ will be ν times bigger than what we previously defined since

$$w_i = \frac{(\frac{\partial \mu_i}{\partial \eta_i})^2}{\nu V(\mu_i)}$$

- To estimate ν, note that by definition $\nu = \frac{E[(y_i - \mu_i)^2]}{V(\mu_i)}$.
- Therefore, we can use the method of moments estimator as follows

$$\hat{\nu} = \frac{1}{n-p-1} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

Overdispersion

- The above estimate of dispersion parameter is the same as $X^2/(n-p-1)$ where X^2 is the general form of Pearson statistic.
- Therefore, we estimate the dispersion parameter by dividing Pearson statistic by the residual degrees of freedom, $n-p-1$.
- If $\hat{\nu} > 1$, we should adjust the standard error by multiplying it by $\sqrt{\hat{\nu}}$.

Random effects: Generalized Linear Mixed Models (GLMM)

- As we mentioned, the cause of overdispersion is usually population heterogeneity, which is not captured by the model.
- Generalized linear models we discussed so far do not take the heterogeneity of the population into account, since the effect of covariates are *fixed* across all observations.
- To address this issue, *generalized linear mixed models* are employed as a generalization of GLMs.
- The underlying assumption in these models is that the population is in fact comprised of several sub-groups.
- The sub-groups could be related subjects (e.g., siblings, matched pairs) or multiple observations for the same subject.
- As the results, one or more regression parameters (this may include the intercept) in these models may vary from one group to other. That is, the relationships between some of covariates and the response variable are group specific.

GLMM

- In GLMM, the linear predictor has the following form

$$g(\mu_{it}) \quad = \quad x_{it}\beta + z_{it}\gamma_i$$

  Here, i indexes the sub-groups, t is the index for observations in each subgroup, β are the *fixed effects*, and γ_i are the *random effects* (which is group specific).

- A simple form of GLMM is a *random intercept* model

$$g(\mu_{it}) \quad = \quad \alpha + \gamma_i + x_{it}\beta$$

GLMM

- The random effects are usually assumed to have a multivariate normal distribution, $\gamma_i \sim N(0, \Sigma)$
- Σ depends on few parameters defining the variance and possibly the correlation for the random effects.
- The likelihood based on GLMM is defined jointly based on β and Σ; that is, the model parameters for which the likelihood needs to be maximized is (β, Σ).
- To estimate the model parameters, we can use numerical methods such as *quadrature* method, *Monte Carlo EM*, and penalized quasi-likelihood approximation.
- In R, we can use the `lme4` package to fit a GLMM.

Logistic-normal model

- A simple GLMM for binary outcome is *logistic-normal*

$$\begin{aligned} logit[P(y_{it} = 1|\gamma_i)] &= x_{it}\beta + \gamma_i \\ \gamma_i &\sim N(0, \sigma^2) \end{aligned}$$

- In this model, sharing the common γ_i in each group creates a non-negative correlation for the response variable of observations in that group.
- The interpretation of β is different from what we had before.
- In this model, $(x_{it} - x_{is})\beta$ is the log odds ratio comparing cases in the same group.
- For cases in different group, the log odds ratio is $(x_{it} - x_{hs})\beta + (\gamma_i - \gamma_h)$, which of course depends on the random effects. Moreover

$$\begin{aligned} (\gamma_i - \gamma_h) &\sim N(0, 2\sigma^2) \\ \text{log-odds-ratio}|\beta &\sim N((x_{it} - x_{hs})\beta, 2\sigma^2) \end{aligned}$$

Quadrature method

- For simple models such as the logistic-normal model discussed above, we can integrate out the random effects parameter.
- In the logistic-normal, this leas to the following likelihood

$$\prod_i \left(\int_{-\infty}^{\infty} \prod_t [\frac{\exp(x_{it}\beta + \gamma_i)}{1 + \exp(x_{it}\beta + \gamma_i)}]^{y_i} [\frac{1}{1 + \exp(x_{it}\beta + \gamma_i)}]^{1-y_i} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\gamma_i^2/2\sigma^2) d\gamma_i \right)$$

- The integral can be approximated using *Gauss-Hermite quadrature* method.
- Gauss-Hermite quadrature method uses the following approximation

$$\int_{-\infty}^{\infty} f(\gamma) \exp(-\gamma^2) du \approx \sum_{m=1}^{M} w_m f(q_m)$$

Quadrature method

- You can use the gqz function from the R package `npmlreg` to obtain the quadrature points q_m and the corresponding weights w_m.
- After approximating the integral (which eliminates γ_i), we can estimate β and σ^2 using numerical methods such as Newton-Raphson.

Monte Carlo EM

- For problems with higher dimension, the *EM* (expectation-maximization) algorithm would be more appropriate.
- The *EM* algorithm is widely used to solve complex maximum likelihood problems, especially when dealing with missing data and unobserved variables.
- Here, the random effects are treated as missing data.
- In each iteration, r, during the E-step the expectation is taken with respect of $(\gamma|y, \beta^{(r)}, \Sigma^{(r)})$.
- In the M-step, we maximize the likelihood with respect of (β, Σ) to obtain $(\beta^{(r+1)}, \Sigma^{(r+1)})$.
- The expectation in the E-step can be performed using Monte Carlo method (MCEM).

Other mixture models

- The generalized linear mixed models we discussed so far is based on normal mixtures of linear predictors.
- Non-normal mixture models could also be used to deal with overdisperssion and in general for more flexibility.
- Some of these models are
  - Beta-binomial
  - Negative-binomial
  - Zero-inflated Poisson

Beta-binomial

- The beta-binomial model is used to handle the overdisperssion issue for modeling binary response variables.
- As before, the response variable is assumed to have binomial(n, θ). However, in this model, θ itself is assumed to have beta(a, b) distribution.
- The results is a beta mixture of binomials.
- Beta is a *conjugate* prior for the parameter of binomial. Therefore, the above model is an example of conjugate mixture models.
- By integrating θ out, we obtain the following closed form

$$P(y|a,b) = \binom{n}{y} \frac{B(a+y, b+n-y)}{B(a,b)}$$

where B is the Beta function.

Beta-binomial

- The mean and variance of this distribution are

$$\begin{aligned} E(y) &= n\mu \\ var(y) &= n\mu[1 - \frac{(n-1)\nu}{1+\nu}] \\ \mu &= \frac{a}{a+b} \\ \nu &= \frac{1}{a+b} \end{aligned}$$

- We can use the beta-binomial as the distribution of response variable along with the following link function

$$logit(\mu_i) = x_i\beta$$

Negative binomial

- To deal with the overdispersion problem when modeling the count data, we can assume $y \sim \text{Poisson}(\lambda)$ distribution as before, but this time we also assume that $\lambda \sim \text{Gamma}(a, b)$.
- The result is a gamma mixture of Poisson distributions.
- We can marginalize over λ. The results is the negative binomial distribution for y.

Zero-inflated Poisson

- When modeling count response variables, it is not uncommon for the data to have excess zeros.
- Lambert (1992), discussed the application of this model for modeling defects in manufacturing. Another possible application is number of accidents per year for each driver.
- For such data, we can use zero-inflated Poisson (zip) distribution for the response variable.
- The probability mass function for this distribution is as follows

$$P(y|\theta,\lambda) = \begin{cases} \theta + (1-\theta)\exp(-\lambda), & \text{for } y = 0 \\ (1-\theta)\exp(-\lambda)\lambda^y/y!, & \text{for } y = 1, 2, ... \end{cases}$$

Zero-inflated Poisson

- The above model is in fact a mixture of a Poisson(λ) distribution and a point mass at 0. Or alternatively,

$$P(y|\theta,\lambda) = \theta\text{Poisson}(0) + (1-\theta)\text{Poisson}(\lambda)$$

- Model parameters (θ,λ) can depend on some covariates as follows

$$\begin{aligned} logit(\theta_i) &= x_i\beta \\ log(\lambda_i) &= x_i\gamma \end{aligned}$$

Zero-inflated Poisson

- To obtain the MLE of β and γ, we can use EM algorithms with *data augmentation*.
- Data augmentation refers to methods where we increase the dimensionality of the data by introducing latent variables.
- While increasing the dimensionality might seem counter-intuitive, in many situations, this makes the analysis of complex data easier.
- For the above model, we can introduce a latent variable $z_i \sim \text{Bernoulli}(\theta)$ such that $y_i = 0$ when $z_i = 1$, and $y_i \sim \text{Poisson}(\lambda_i)$ when $z_i = 0$.
- The complete data (i.e., observed and latent) is therefore (y, z).
- In R, you can use the `zeroinfl` function from the `pscl` packages to fit a zip model.