GLM (Generalized Linear Model) #3 (version 2)

Paul Johnson

March 10, 2006

1 Why do you need a "quasi" model at all?

1. A fitted GLM appears to have "overdispersed" observations.
2. You want to estimate parameters without specifying the probability distribution of y_i in complete detail.
3. You want to estimate parameters for a model that is too complicated for the solution of the likelihood function

2 If you have the problem of overdispersion, what are the alternatives?

Overdispersion is a primarily a phenomenon of the one-parameter exponential distributions, Poisson and binomial, because those models restrict the dispersion parameter to be 1.0.

2.1 Assume a Mixed Model

If your model originally was this

$$g(\mu_i) = X_i b \tag{1}$$

then assume that there is an additional random variable causing the extra dispersion

$$g(\mu_i) = X_i b + e_i \tag{2}$$

There are some fabulously successful "special cases." Social scientists, following J. Scott Long's influential textbook, are inclined to assume that e_i is log Gamma, and that leads to y which is Negative Binomial. Its variance is larger than the Poisson. Similarly, in a logistic regression, one can add a random error and estimate a so-called "Beta-Binomial" model.

McCullagh & Nelder oppose the knee-jerk adoption of the Beta-Binomial (or Negative Binomial) models simply because they have dispersion patterns that match data. "Though this is an attractive option from a theoretical standpoint, in practice it seems unwise to rely on a specific form of over-dispersion, particularly where the assumed form has been chosen for mathematical convenience rather than scientific plausibility." (p. 126)

Bayesian modeling opens up major opportunities for fitting models with user-determined random errors.

2.2 Force a dispersion parameter into your GLM

The quasibinomial and quasipoisson families included in the R stats package are quite simple. Following McCullagh & Nelder, the quasibinomial model keeps the same structure, except that "the variance is inflated by an unknown factor σ^2" (p. 126). This "unknown factor" was called the dispersion parameter ϕ in my other handouts. The estimates of the $\hat{b}$ are not changed, but the estimates of the standard errors are changed. McCullagh & Nelder observe that the covariance matrix of the parameter estimates is inflated according to the estimated value of the dispersion coefficient

$$Cov(\hat{b}) = \phi\,(X'WX)^{-1} \tag{3}$$

There are other routines that have the same name, quasibinomial or quasipoisson, impose a little more structure on the way that ϕ is incorporated. I've seen at least one routine that fits a model called "quasipoisson" but it is mathematically equivalent to the more familiar Negative Binomial. So, when you find these things, you simply must read the manual.

2.3 Adopt a quasi-likelihood modeling strategy

Quasi-likelihood was proposed by Wedderburn and then popularized in the bible of generalized linear models, McCullagh & Nelder. The user is asked to specify the mean of y_i as a function of the inputs as well as the dependence of the variance of y_i on the mean (the so-called variance function).

The Generalized Estimating Equations (GEE) are an extension of the quasi-likelihood idea to estimate Generalized Linear Models with clustering of observations and inter-temporal correlations.

3 Insert GLS notes here

I have a brief handout on Generalized Least Squares. If you don't have it, get it! Or some book...

4 Quasi likelihood

As usual, the observations are collected into a column vector y. The version of quasi-likelihood that was presented by Wedderburn (and later McCullagh & Nelder) supposes that there are "true" mean values, μ_i. Recall that $\mu = (\mu_1, \mu_2, \cdots, \mu_N)'$ is a vector of mean values, one for each case. There are also observations $y = (y_1, y_2, \cdots, y_N)'$ drawn from a distribution. The original versions of this model, before the GEE was popularized, assumed that the y's were observed independently and that each y_i is affected only by "its own" mean μ_i.

The quasi-likelihood model consists of 3 parts.

1. A model for the mean of y_i for each case. We might as well think of those as predicted values, $\hat{\mu}_i$. The modeler is supposed to have in mind a relationship between the input variables, X and some parameters, b. Obviously, a link function must be specified, $g(\mu_i) = \eta_i = X_i b$

2. Variance of y_i as it depends on the means. V is a matrix of values which show ow the variance of y_i is affected by the average. This is a theoretically stated belief, not an empirically estimated function. For a first cut at the problem, the observations are not autocorrelated, so V is diagonal:

$$V(\mu) = \begin{bmatrix} V(\mu_1) & & & 0 & 0 \\ & V(\mu_2) & & 0 & 0 \\ & & \ddots & & \\ & & & & \\ 0 & & & V(\mu_{N-1}) & \\ 0 & 0 & & & V(\mu_N) \end{bmatrix} \tag{4}$$

3. Quasi-likelihood and Quasi-score functions.

See McCullagh & Nelder, 2ed, p. 327.

The quasi-log-likelihood for the i'th case has some properties of a log-likelihood function is defined as

$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)}\, dt \tag{5}$$

This might be described as a "variance weighted indicator of the gap between the observed value of y_i and the expected value." One should choose μ_i to make this as large as possible. If one chose μ_i to optimize Q_i,one would solve the first order condition

$$\frac{\partial Q_i}{\partial \mu_i} = \frac{y_i - \mu_i}{\phi V(\mu_i)} = 0 \tag{6}$$

This is the result of the application of a first-year calculus result about the differentiation of integrals with respect to their upper limit of integration.

In the regression context, the parameter of interest is a regression coefficient which is playing a role in predicting μ_i . The first order equation for the quasi-log-likelihood function is a quasi-score function. (Recall that a maximum likelihood score equation is a first order condition, one that sets the first derivative of each parameter equal to 0.) The quasi-score function "looks like" a score equation from the GLM.

$$U(b_k) = \sum \frac{1}{\phi} \frac{\partial \mu_i}{\partial b_k} \frac{1}{V(\mu_i)} (y_i - \mu_i) \tag{7}$$

It is "like" a score equation, in the sense that it satisfies many of the fundamental properties of score equations. Recall the ML Fact 1, $E(U(b_k)) = 0$, for example.

In matrices, it would be stated

$$U(b_k) = \frac{1}{\phi} D' V^{-1} (y - \mu) \tag{8}$$

D is a matrix of partial derivatives showing the impact of each coefficient on the predicted value for each case and D' is the transpose of D.

$$U(\hat{b}) = \frac{1}{\phi} \begin{bmatrix} \frac{\partial \mu_1}{\partial b_1} & & & \frac{\partial \mu_1}{\partial b_p} \\ \frac{\partial \mu_2}{\partial b_1} & \frac{\partial \mu_2}{\partial b_2} & \cdots & \\ & & & \\ \frac{\partial \mu_N}{\partial b_1} & & & \frac{\partial \mu_N}{\partial b_p} \end{bmatrix}' \begin{bmatrix} V(\mu_1) & & & 0 \\ & V(\mu_2) & & \\ & & \ddots & \\ 0 & & & V(\mu_N) \end{bmatrix}^{-1} \begin{bmatrix} y_1 & - & \mu_1 \\ y_2 & - & \mu_2 \\ \vdots & & \vdots \\ y_N & & \mu_N \end{bmatrix} \tag{9}$$

As long as we are working with the simple, nonautocorrelated case, the values of D' and V^T can be easily written down:

$$U(\hat{b}) = \frac{1}{\phi} \begin{bmatrix} \frac{\partial \mu_1}{\partial b_1} & \frac{\partial \mu_2}{\partial b_1} & & \frac{\partial \mu_N}{\partial b_1} \\ \frac{\partial \mu_1}{\partial b_2} & \frac{\partial \mu_2}{\partial b_2} & & \frac{\partial \mu_N}{\partial b_2} \\ & & \ddots & \\ \frac{\partial \mu_1}{\partial b_p} & \frac{\partial \mu_2}{\partial b_p} & & \frac{\partial \mu_N}{\partial b_p} \end{bmatrix} \begin{bmatrix} \frac{1}{V(\mu_1)} & & & \\ & \frac{1}{V(\mu_2)} & & \\ & & \ddots & \\ & & & \frac{1}{V(\mu_N)} \end{bmatrix}' \begin{bmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ \vdots \\ y_N \quad \mu_N \end{bmatrix} \tag{10}$$

This is not derived from a Likelihood equation, but McCullagh & Nelder use words like "related" or "connected".

The elements in $\partial u_i / \partial b_j$ carry along with them the information that is stated in the link function as well as the impact of b_j on η_j . Assuming that the predictive equation for the i'th case with the j'th variable is $b_j x_{ij}$,

$$\frac{\partial \mu_i}{\partial b_j} = \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial b_j} = \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \cdot x_{ij} \tag{11}$$

Since the link function is continuous, differentiable, and monotonic, this is the same as

$$\frac{1}{\partial g / \partial \mu_i} x_{ij} = \frac{1}{g'(\mu_i)} x_{ij} \tag{12}$$

The quasi-score equation looks almost exactly like the first order condition (the normal equations) from GLS. It also looks a lot like the first order condition of the GLM.

The covariance matrix of $U(\hat{b})$ is a familiar looking thing,

$$\frac{1}{\phi} D' V^{-1} D \tag{13}$$

Please note that we did NOT assume anything about the precise distribution of y_i. We have only assumed the predictive formula for μ and the variance function.

McCullagh & Nelder observe that the Newton-Raphson algorithm (using Fisher's Information matrix) can be used to solve the score equation. "Approximate unbiasedness and asymptotic Normality of $\hat{b}$ follow directly from (the algorithm) under the second-moment assumptions made in this chapter" (p. 328). "In all of the above respects the quasi-likelihood behaves just like an ordinary log likelihood" (p. 328).

4.1 Correlated cases

When the cases are not independent from one another, the most important change is in the matrix V^{-1}. Now it is a general matrix of weights indicating the interdependence of observed values y_i on the means of one or more observations.

You can write it all out verbosely, suppose the coefficients are $b = (b_1, b_2, ..., b_p)'$

$$\begin{bmatrix} \frac{d\mu_1}{db_1} & \frac{d\mu_2}{db_1} & \cdots & \frac{d\mu_N}{db_1} \\ \frac{d\mu_1}{db_2} & \frac{d\mu_2}{db_2} & & \frac{d\mu_N}{db_2} \\ & & & \\ \frac{d\mu_1}{db_p} & \frac{d\mu_2}{db_p} & & \frac{d\mu_N}{db_p} \end{bmatrix} \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1N} \\ w_{12} & w_{22} & & w_{2N} \\ \vdots & & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{NN} \end{bmatrix} \begin{bmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ \vdots \\ y_N - \mu_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{14}$$

In a quasi-likelihood framework, one must begin with an estimate of the coefficients $\hat{b}^1$and then iteratively calculate values for $\hat{\mu}$and $\hat{W}$ and $\hat{D}'$ and when the number smashing stops, one has obtained a quasi-likelihood estimator.

Liang and Zeger pioneered this. They claim that the parameter estimates $\hat{b}$ are consistent and have many of the properties of maximum likelihood.

5 It's the same as OLS/GLS when the models "coincide."

Suppose for a moment that we have p variables and p parameters to estimate. We only assume that because I want to show the preceding is the same as GLS if you assume the predictive model is linear. The data for the input variables has N rows and p columns (one for each parameter), including a "constant" column of 1's if desired:

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1p} \\ x_{21} & x_{22} & & & \\ \vdots & & & & \\ & & & & \\ x_{N1} & x_{N2} & \cdots & x_{(N-1)(p-1)} & x_{Np} \end{bmatrix} \tag{15}$$

Note that if you have a linear model, $\mu = Xb$, then $\frac{d\mu_i}{db_j} = x_{ij}$, and $D' = X'$.

6 What good could it possibly be?

It leads to the GEE framework for GLM applications to panel studies. In other words, longitudinal data analysis with non-Normal variables.

7 Why is Quasi-Likelihood needed?

If you read books on GLM, they go through all the general stuff about the exponential models and then, at the end, they say "hey, look here, we can get something almost as good, but with weaker assumptions. And its inspired by GLS."

Maybe you are like me and you think to yourself, "Who the heck cares? If I want something 'like GLS,' I'll just use GLS. I don't see why I need quasi-likelihood." I read article after article and I did not see the point. But then it hit me.

OLS is meaningful for "symmetrical" error terms, especially the Normal.

The motivation for GLS is minimizing the sum of squared errors. That concept works for Normal distributions, and possibly other symmetric distributions. But its not going to help much for asymmetric distributions.

And maximum likelihood is not an answer because the quasi development starts by presupposing that you don't know $f(y_i|X_i, b)$.

And then the magical part of quasi-likelihood becomes apparent:

When you want something as good as a "sum of squares model" or a maximum likelihood model, but you don't have the tools for either, there is an alternative that is "almost as good."