Generalized Linear Models

Kerby Shedden

Department of Statistics, University of Michigan

December 6, 2021

Motivation for nonlinear models

The key properties of a linear model are that

$$E[y|X] = X\beta \quad \text{and} \quad \text{var}[y|X] \propto I.$$

In some cases where these conditions are not met, we can transform y so that the properties of a linear model are well-satisfied.

However it is often difficult to find a transformation that simultaneously linearizes the mean and gives constant variance.

Also, if y lies in a restricted domain (e.g. $y_i \in \{0, 1\}$), parameterizing $E[y|X]$ as a linear function of X violates the domain restriction.

Generalized linear models (GLM's) are a class of nonlinear regression models that can be used in certain cases where linear models do not fit well.

Logistic regression

Logistic regression is a specific type of GLM. We will develop logistic regression from first principles before discussing GLM's in general.

Logistic regression is used for binary outcome data, where $y_i = 0$ or $y_i = 1$. It is defined by the probability mass function

$$P(y_i = 1|x_i = x) = \frac{\exp(\beta' x)}{1 + \exp(\beta' x)} = \frac{1}{1 + \exp(-\beta' x)},$$

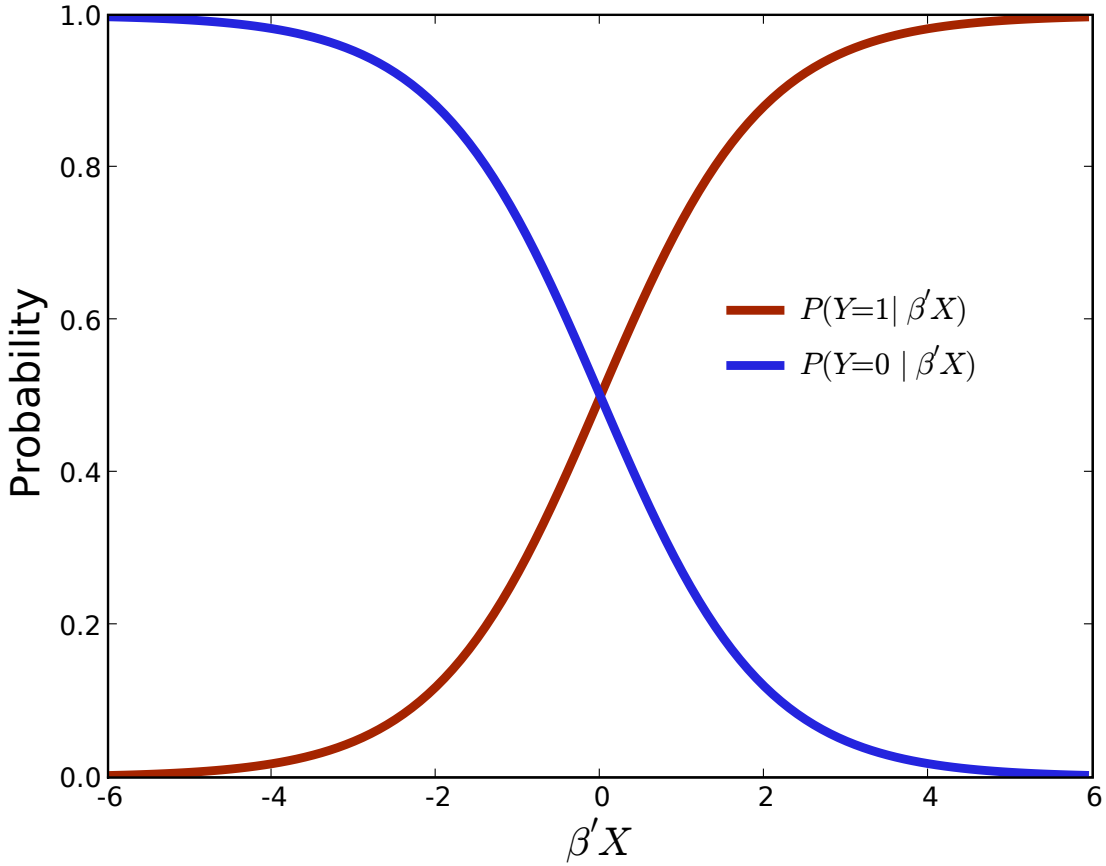
which implies that

$$P(y_i = 0|x_i = x) = 1 - P(y_i = 1|x_i = x) = \frac{1}{1 + \exp(\beta' x)},$$

where in most cases, $x_{i0} = 1$ so β_0 is the intercept.

Logistic regression

This plot shows $P(y = 1|x)$ and $P(y = 0|x)$, plotted as functions of $\beta' x$:



Logistic regression

The logit function

$$\text{logit}(x) = \log(x/(1-x))$$

maps the unit interval onto the real line. The inverse logit function, or expit function

$$\text{expit}(x) = \text{logit}^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$$

maps the real line onto the unit interval.

In logistic regression, the logit function is used to map the linear predictor $\beta' x$ to a probability.

Logistic regression

The linear predictor in logistic regression is the conditional log odds:

$$\log\left[\frac{P(y=1|x)}{P(y=0|x)}\right] = \beta' x.$$

Thus one way to interpret a logistic regression model is that a one unit increase in x_j (the j^{th} covariate) results in a change of β_j in the conditional log odds.

Or, a one unit increase in x_j results in a multiplicative change of $\exp(\beta_j)$ in the conditional odds.

Latent variable model for logistic regression

It may make sense to view the binary outcome y as being a dichotomization of a latent continuous outcome y_c,

$$y = \mathcal{I}(y_c \geq 0).$$

Suppose $y_c|X$ follows a logistic distribution, with CDF

$$F(y_c|x) = \frac{\exp(y_c - \beta' x)}{1 + \exp(y_c - \beta' x)}.$$

In this case, $y|x$ follows the logistic regression model:

$$P(y = 1|x) = P(y_c \geq 0|x) = 1 - \frac{\exp(0 - \beta' x)}{1 + \exp(0 - \beta' x)} = \frac{\exp(\beta' x)}{1 + \exp(\beta' x)}.$$

Mean/variance relationship for logistic regression

Since the mean and variance of a Bernoulli trial are linked, the mean structure

$$E[y|x] = P(y = 1|x) = \frac{\exp(\beta' x)}{1 + \exp(\beta' x)}$$

also determines the variances

$$\mathrm{var}[y|x] = P(y = 1|x) \cdot P(y = 0|x) = \frac{1}{2 + \exp(\beta' x) + \exp(-\beta' x)}.$$

Since the variance depends on x, logistic regression models are always heteroscedastic.

Logistic regression and case-control studies

Suppose we sample people based on their disease status D ($D = 1$ is a case, $D = 0$ is a control).

We are interested in a binary marker $M \in \{0, 1\}$ that may predict a person's disease status.

The prospective log odds

$$\log \left[\frac{P(D = 1|M = m)}{P(D = 0|M = m)} \right]$$

measures how informative the marker is for the disease.

Logistic regression and case-control studies

Suppose we model $P(M = m|D = d)$ using logistic regression, so

$$P(M = 1|D = d) = \frac{\exp(\alpha + \beta d)}{1 + \exp(\alpha + \beta D)}$$

$$P(M = 0|D = d) = \frac{1}{1 + \exp(\alpha + \beta d)}.$$

More generally,

$$P(M = m|D = d) = \frac{\exp(m(\alpha + \beta d))}{1 + \exp(\alpha + \beta d)}.$$

Logistic regression and case-control studies

Since

$$\log \frac{P(M=1|D=d)}{P(M=0|D=d)} = \alpha + \beta d$$

we see that β is the coefficient of d in the retrospective log odds.

Logistic regression and case-control studies

The prospective log odds can be written

$$\begin{aligned}\log \frac{P(D=1|M=m)}{P(D=0|M=m)} &= \log \frac{P(M=m|D=1)P(D=1)/P(M=m)}{P(M=m|D=0)P(D=0)/P(M=m)} \\ &= \log \frac{P(M=m|D=1)P(D=1)}{P(M=m|D=0)P(D=0)}\end{aligned}$$

Logistic regression and case-control studies

Continuing from the previous slide, we have

$$\log \frac{P(M = m|D = 1)P(D = 1)}{P(M = m|D = 0)P(D = 0)} =$$
$$\log \left[\frac{\exp(m \cdot (\alpha + \beta))/(1 + \exp(\alpha + \beta))}{\exp(m \cdot \alpha)/(1 + \exp(\alpha))} \cdot \frac{P(D = 1)}{P(D = 0)} \right],$$

which equals

$$\beta m + \log \left[\frac{1 + \exp(\alpha)}{1 + \exp(\alpha + \beta)} \cdot \frac{P(D = 1)}{P(D = 0)} \right].$$

Thus β is both the coefficient of d in the retrospective log odds, and it is the coefficient of m in the prospective log odds. This is sometimes called case/control convertibility.

Estimation and inference for logistic regression

Assuming independent cases, the log-likelihood for logistic regression is

$$\begin{aligned} L(\beta|y,X) &= \log \prod_i \frac{\exp(y_i \cdot \beta' x_i)}{1+\exp(\beta' x_i)} \\ &= \sum_{i:y_i=1} \beta' x_i - \sum_i \log(1+\exp(\beta' x_i)). \end{aligned}$$

This likelihood is for the conditional distribution of y given X.

As in linear regression, we do not model the marginal distribution of x (a row of X).

Estimation and inference for logistic regression

Logistic regression models are usually fit using maximum likelihood estimation.

This means that the parametric likelihood above is maximized as a function of β.

The gradient of the log-likelihood function (the score function) is

$$\begin{aligned} G(\beta|y,X) &= \frac{\partial}{\partial\beta}L(\beta|y,X) \\ &= \sum_{i:y_i=1} x_i - \sum_i \frac{\exp(\beta' x_i)}{1+\exp(\beta' x_i)} x_i \\ &= \sum_i \left(y_i - \frac{\exp(\beta' x_i)}{1+\exp(\beta' x_i)} \right) x_i. \end{aligned}$$

Estimation and inference for logistic regression

The Hessian of the log-likelihood is

$$H(\beta|y,X) = \frac{\partial^2}{\partial\beta\beta'}L(\beta|y,X) = -\sum_i \frac{\exp(\beta' x_i)}{(1+\exp(\beta' x_i))^2} x_i x_i'.$$

The Hessian is strictly negative definite as long as the design matrix has independent columns. Therefore $L(\beta|y,X)$ is a concave function of β, so has a unique maximizer, and hence the MLE is unique.

Estimation and inference for logistic regression

From the general theory of the MLE, the Fisher information

$$I(\beta) = -(E[H(\beta|y, X)|X])^{-1}$$

is the asymptotic sampling covariance matrix of the MLE $\hat{\beta}$. Since $H(\beta|y, X)$ does not depend on y, $I(\beta) = -H(\beta|y, X)^{-1}$.

Since $\hat{\beta}$ is an MLE for a regular problem, it is consistent, asymptotically unbiased, and asymptotically normal if the model is correctly specified.

Poisson regression

The Poisson distribution is a single-parameter family of distributions on the sample space $\{0, 1, 2, \ldots\}$.

A key property of the Poisson distribution is that the mean is equal to the variance.

The Poisson distribution is usually parameterized in terms of a parameter λ that is equal to the common mean and variance.

In regression, we don't want just a single distribution. Instead we want a family of distributions indexed by the covariate vector x.

To create a regression methodology based on the Poisson distribution, we can formulate a regression model in which $y|x$ is Poisson, with mean and variance equal to $\lambda(x) = \exp(\beta' x)$.

Poisson regression

Since the mean function in a Poisson distribution has an exponential form, the covariates are related multiplivatively to the mean.

If we contrast the mean value for two different covariate vectors, $x^{(1)}$ and $x^{(2)}$, such that $x_j^{(1)} - x_j^{(2)} = 1$, and $x_k^{(1)} = x_k^{(2)}$ for $k \neq j$, then the means at these two points are related through:

$$\lambda(x^{(1)}) = \exp(\beta_j)\lambda(x^{(2)}).$$

Poisson regression

Setting the mean to be $\lambda_i = \exp(\beta' x_i)$, the PMF for one observation in a Poisson regression model is

$$\exp^{-\lambda_i} \lambda_i^{y_i} / y_i!$$

The corresponding contribution to the log likelihood is

$$-\lambda_i + y_i \log(\lambda_i) - \log(y_i!) = -\exp(\beta' x_i) + y_i \cdot \beta' x_i - \log(y_i!),$$

and the contribution to the score function is

$$-x_i \exp(\beta' x_i) + y_i \cdot x_i = (y_i - \exp(\beta' x_i)) x_i.$$

Score equations

The MLE is a stationary point of the score function. Thus, for logistic regression, the following equation is satisfied at the MLE:

$$\sum_i \left(y_i - \frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)} \right) x_i = 0.$$

For Poisson regression, this equation is satisfied at the MLE:

$$\sum_i (y_i - \exp(\beta' x_i)) x_i$$

We also know that for OLS (viewed here as a Gaussian regression model), this equation is satisfied at the MLE

$$\sum_i (y_i - \beta' x_i) x_i = 0.$$

Score equations

Writing $\mu_i = E[y_i|x_i]$, we see that for all three types of regression models, the following equation is satisfied.

$$\sum_i (y_i - \mu_i)x_i = 0.$$

This shows that the residuals are orthogonal to each covariate in all of these models, and that achieving this orthogonality characterizes the MLE.

This turns out to be a useful generic framework for regression, as many different mean functions $\mu(\beta)$ can be substituted into this equation, and the solution of the equation can be used to estimate β.

Relationship between the mean and variance

We have seen three parametric regression models, each of which expresses the mean in terms of the linear predictor. The family is the distributional family used to form the log-likelihood and score functions.

For each of these models, the variance can also be related to the mean.

Family	Mean (μ)	Variance ($v(\mu)$)
Gaussian	$\beta' x$	1
Binomial	$1/(1+\exp(-\beta' x))$	$\mu(1-\mu)$
Poisson	$\exp(\beta' x)$	μ

Relationship between the mean and variance

The variance functions here are only specified up to a constant of proportionality. That is, $\mathrm{var}(y_i|x_i) = \phi v(\mu_i)$, where ϕ is the scale parameter.

In any single index model, $\mu_i = \mu(\beta' x_i)$, so $\partial \mu_i / \partial \beta = \mu'(\beta' x_i) \cdot x_i \propto x$.

Note that in each case above, $\partial \mu_i / \partial \beta$ is proportional to $v_i \cdot x_i$, where $v_i \in \mathcal{R}$ is the variance (but this is not always the case).

Estimating equations

For the three examples we are focusing on here, the MLE can be defined as the solution to the estimating equations:

$$\sum_i \partial\mu_i/\partial\beta \cdot (y_i - \mu_i(\beta))/v_i(\beta) = 0$$

which can also be expressed

$$\sum_i \mu'(\eta_i) \cdot (y_i - \mu_i(\beta)) \cdot x_i/v_i(\beta) = 0$$

where $\eta_i = \beta' x_i$ is the linear predictor.

Estimating equations

The score equations constitute a system of $p = \dim(x)$ equations in p unknowns. It should be solvable unless there is some degeneracy in the equations.

In the "canonical setting", $(\partial \mu_i / \partial \beta)/v_i(\beta) \propto x_i$, so these equations are equivalent to the orthogonality between residuals and covariates. But we will see below that the form given here extends to some "non-canonical" settings and hence is somewhat more general.

Development of GLM's using likelihoods

A GLM is based on the following conditions:

- The y_i are conditionally independent given X.
- The probability mass function or density can be written

$$\log p(y_i|\theta_i, \phi, x_i) = w_i(y_i\theta_i - \gamma(\theta_i))/\phi + \tau(y_i, \phi/w_i),$$

where w_i is a known weight, $\theta_i = g(\beta' x_i)$ for an unknown vector of regression slopes β, $g(\cdot)$ and $\gamma(\cdot)$ are smooth functions, ϕ is the "scale parameter" (which may be either known or unknown), and $\tau(\cdot)$ is a known function.

Development of GLM's using likelihoods

The log-likelihood function is

$$L(\beta, \phi | y, X) = \sum_i w_i(y_i\theta_i - \gamma(\theta_i))/\phi + \tau(y_i, \phi/w_i).$$

The score function with respect to θ_i is

$$w_i(y_i - \gamma'(\theta_i))/\phi.$$

Development of GLM's using likelihoods

Next we need a fundamental fact about score functions.

Let $f_\theta(y)$ be a density in y with parameter θ. The score function is

$$\frac{\partial}{\partial\theta}\log f_\theta(y) = f_\theta(y)^{-1}\frac{\partial}{\partial\theta}f_\theta(y).$$

The expected value of the score function is

$$\begin{aligned} E\frac{\partial}{\partial\theta}\log f_\theta(y) &= \int f_\theta(y)^{-1}\left(\frac{\partial}{\partial\theta}f_\theta(y)\right)f_\theta(y)dy \\ &= \frac{\partial}{\partial\theta}\int f_\theta(y)dy \\ &= 0. \end{aligned}$$

Thus the score function has expected value 0 when θ is at its true value.

Development of GLM's using likelihoods

Since the expected value of the score function is zero, we can conclude that

$$E[w_i(y_i - \gamma'(\theta_i))/\phi|X] = 0,$$

so

$$E[y_i|X] = \gamma'(\theta_i) = \gamma'(g(\beta' x_i)).$$

Note that this relationship does not depend on ϕ or τ.

Development of GLM's using likelihoods

Using a similar approach, we can relate the variance to w_i, ϕ, and γ'. By direct calculation,

$$\partial^2 L(\theta_i | y_i, x_i, \phi) / \partial \theta_i^2 = -w_i \gamma''(\theta_i)/\phi.$$

Returning to the general density $f_\theta(y)$, we can write the Hessian as

$$\frac{\partial}{\partial \theta \theta'} \log f_\theta(y) = f_\theta(y)^{-2} \left(f_\theta(y) \frac{\partial^2}{\partial \theta \theta'} f_\theta(y) - \frac{\partial f_\theta(y)}{\partial \theta} \cdot \frac{\partial f_\theta(y)}{\partial \theta'} \right).$$

Development of GLM's using likelihoods

The expected value of the Hessian is

$$
\begin{aligned}
E\frac{\partial}{\partial\theta\theta'}\log f_\theta(y) &= \int \frac{\partial}{\partial\theta\theta'}\log f_\theta(y)\cdot f_\theta(y)dy \\
&= \frac{\partial}{\partial\theta\theta'}\int f_\theta(y)dy - \int\left(\frac{\partial f_\theta(y)/\partial\theta}{f_\theta(y)}\cdot\frac{\partial f_\theta(y)/\partial\theta'}{f_\theta(y)}\right) f_\theta(\\
&= -\mathrm{cov}\left(\frac{\partial}{\partial\theta}\log f_\theta(y)|X\right).
\end{aligned}
$$

Therefore

$$
w_i\gamma''(\theta_i)/\phi = \mathrm{var}\left(w_i(y_i-\gamma'(\theta_i))/\phi|X\right)
$$

so $\mathrm{var}[y_i|X] = \phi\gamma''(\theta_i)/w_i$.

Examples of GLM's

Gaussian linear model: The density of $y|X$ can be written

$$\begin{aligned}\log p(y_i|x_i) &= -\log(2\pi\sigma^2)/2 - \frac{1}{2\sigma^2}(y_i - \beta' x_i)^2 \\ &= -\log(2\pi\sigma^2)/2 - y_i^2/2\sigma^2 + (y_i\beta' x_i - (\beta' x_i)^2/2)/\sigma^2.\end{aligned}$$

This can be put into GLM form by setting $g(x) = x$, $\gamma(x) = x^2/2$, $w_i = 1$, $\phi = \sigma^2$, and $\tau(y_i, \phi) = -\log(2\pi\phi)/2 - y_i^2/2\phi$.

Examples of GLM's

Logistic regression: The mass function of $y|x$ can be written

$$\begin{aligned} \log p(y_i|x_i) &= y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \\ &= y_i \log(p_i/(1 - p_i)) + \log(1 - p_i), \end{aligned}$$

where

$$p_i = \text{logit}^{-1}(\beta' x_i) = \frac{\exp(\beta' x_i)}{1 + \exp(\beta' x_i)}.$$

Since $\log(p_i/(1 - p_i)) = \beta' x$, this can be put into GLM form by setting $g(x) = x$, $\gamma(x) = -\log(1 - \text{logit}^{-1}(x)) = \log(1 + \exp(x))$, $\tau(y_i, \phi) \equiv 0$, $w = 1$, and $\phi = 1$.

Examples of GLM's

Poisson regression: In Poisson regression, the distribution of $y|x$ follows a Poisson distribution, with the mean response related to the covariates via

$$\log E[y|x] = \beta' x.$$

It follows that $\log \mathrm{var}[y|x] = \beta' x$ as well. The mass function can be written

$$\log p(y_i|x_i) \quad = \quad y_i \beta' x_i - \exp(\beta' x_i) - \log(y_i!),$$

so in GLM form, $g(x) = x$, $\gamma(x) = \exp(x)$, $w = 1$, $\tau(y_i) = -\log(y_i!)$, and $\phi = 1$.

Examples of GLM's

Negative binomial regression: In negative binomial regression, the probability mass function for the dependent variable Y is

$$P(y_i = y|x) = \frac{\Gamma(y + 1/\alpha)}{\Gamma(y + 1)\Gamma(1/\alpha)} \left(\frac{1}{1 + \alpha\mu_i}\right)^{1/\alpha} \left(\frac{\alpha\mu_i}{1 + \alpha\mu_i}\right)^{y}.$$

The mean of this distribution is μ_i and the variance is $\mu_i + \alpha\mu_i^2$. If $\alpha = 0$ we get the same mean/variance relationship as the Poisson model. As α increases, we get increasingly more overdispersion.

Examples of GLM's

Negative binomial regression (continued):

The log-likelihood (dropping terms that do not involve μ) is

$$\log P(y_i = y|x_i) = y \log(\frac{\alpha\mu_i}{1+\alpha\mu_i}) - \alpha^{-1}\log(1+\alpha\mu_i)$$

Suppose we model the mean as $\mu_i = \exp(\beta' x_i)$. Then in the standard GLM notation, we have

$$\theta_i = \log\left(\frac{\alpha\exp(\beta' X_i)}{1+\alpha\exp(\beta' x_i)}\right),$$

so $g(x) = \log(\alpha) + x - \log(1+\alpha\exp(x))$, and $\gamma(x) = -\alpha^{-1}\log(1-\exp(x))$.

Link functions

In a GLM, the link function maps the mean to the linear predictor $\eta_i = x_i'\beta$. Since

$$E[y_i|x_i] = \gamma'(g(\eta)),$$

it follows that the link function is the inverse of $\gamma' \circ g$.

For example, in the case of logistic regression,

$$\gamma'(g(\eta)) = \exp(\eta)/(1+\exp(\eta)),$$

which is the expit function. The inverse of this function is the logit function $\log(p/(1-p))$, so the logit function is the link in this case.

Link functions

When $g(x) = x$, the resulting link function is called the canonical link function.

In the examples above, linear regression, logistic regression, and Poisson regression all used the canonical link function, but negative binomial regression did not.

The canonical link function for negative binomial regression is $1/x$, but this does not respect the domain and is harder to interpret than the usual log link.

Another setting where non-canonical links arise is the use of the log link function for logistic regression. In this case, the coefficients β are related to the log relative risk rather than to the log odds.

Estimating equations and quasi-likelihood

As noted above, the regression parameters in a GLM can be estimated by solving these estimating equations:

$$\sum_i \partial \mu_i / \partial \beta \cdot (y_i - \mu_i(\beta)) / v_i(\beta) = 0$$

Note that we only need to correctly specify $v_i(\beta)$ up to a constant. For example, in the Gaussian case, we can set $v_i(\beta) = 1$.

Estimating equations and quasi-likelihood

This opens up the possibility of specifying a large class of regression models through their first two moments, represented by the functions $\mu(\beta)$ and $v(\beta)$.

For example, we can get quasi-Poisson regression by specifying $\mu_i(\beta) = \exp(x_i'\beta)$ and $v_i(\beta) = \mu_i(\beta)$. This formulation of quasi-Poisson regression never refers to the Poisson distribution directly, it only depends on moments.

It can be shown that solving the quasi-likelihood equations generally gives consistent estimates of β, as long as the data are sample from a population in which the specified mean and variance functions are correct.

Estimating equations and quasi-likelihood

Wedderburn introduced a "quasi-likelihood" function that can be used when working with estimating equations. It has the form

$$Q(y;\mu,v) = \int_0^\mu \frac{y-u}{v(u)}du$$

Since

$$\partial Q/\partial\beta = \partial\mu/\partial\beta \cdot \partial Q/\partial\mu,$$

and $\partial Q/\partial\mu = (y-\mu)/v(\mu)$ by the fundamental theorem of calculus, we ssee that $\partial Q/\partial\beta$ gives the estimating equations discussed above.

Estimating equations and quasi-likelihood

In some cases the quasi-likelihood is an actual likelihood, but even if it is not, we can use it in place of a likelihood for many purposes.

For example, we can define a "Quasi Information Criterion" QIC, analogous to AIC for model selection as

$$\sum_i Q(y_i, \hat{\mu}_i, v) - p,$$

where $p = \dim(\beta)$. This quantity is to be maximized, or alternatively we can minimize

$$-2\sum_i Q(y_i, \hat{\mu}_i, v) + 2p.$$

Scale parameters and quasi-likelihood

Since the quasi-likelihood estimating equations are homogeneous, we can estimate the mean structure $\mu = \exp(\beta' x)$ in a setting where the specified variance is off by a multiplicative constant. For example, these estimating equations can be used to consistently estimate β in a quasi-Poisson model where $\mathrm{Var}[y_i|x_i] = \phi E[y_i|x_i]$.

This is a quasi-likelihood estimator, because there is no single "quasi-Poisson distribution". There are many distributions that have this variance structure, but the solution to these estimating equations is not the MLE for a specific distribution.

This can be viewed as a way to construct a consistent estimator for β that can be used for any distribution where the conditional variance has this structure.

Scale parameters and quasi-likelihood

In a quasi-likelihood analysis, the scale parameter is usually estimated in a separate step, after the regression parameters (β) are estimated by solving the estimating equations.

There are several related ways to estimate the scale parameter. A common approach is to use

$$\hat{\phi} = \frac{\sum_i (y_i - \hat{\mu}_i)^2 / \hat{v}_i}{n - p}.$$

Overdispersion

Under the Poisson model, $\mathrm{var}[y|x] = E[y|x]$. A Poisson model results from using the Poisson GLM with the scale parameter ϕ fixed at 1.

The quasi-Poisson model is the Poisson model with a scale parameter that may be any non-negative value. Under the quasi-Poisson model, $\mathrm{var}[y|x] \propto E[y|x]$.

The negative binomial GLM allows the variance to be non-proportional to the mean.

Any situation in which $\mathrm{var}[y|x] > E[y|x]$ is called overdispersion. Overdispersion is often seen in practice.

One mechanism that may give rise to overdispersion is heterogeneity. Suppose we have a hierarchical model in which λ follows a Γ distribution, and $y|\lambda$ is Poisson with mean parameter λ. Then marginally, y is negative binomial.

Shape and other auxiliary parameters

We have seen that the scale parameter can be estimated intependently of the regression parameters (β). Some GLM's (or quasi-GLM's) contain additional parameters that cannot be estimated independently of β.

Once example of this is the shape parameter α in the negative binomial GLM. The shape parameter can be estimated by maximum likelihood, together with β (using a profile likelihood technique), or can be selected using QIC.

Gamma and beta GLM's also have auxiliary parameters that are estimated in this way.

Model comparison for GLM's

If ϕ is held fixed across models, then twice the log-likelihood ratio between two nested models $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ is

$$L \equiv 2\sum_i (y_i\hat{\theta}_i^{(1)} - \gamma(\hat{\theta}_i^{(1)}))/\phi - 2\sum_i (y_i\hat{\theta}_i^{(2)} - \gamma(\hat{\theta}_i^{(2)}))/\phi,$$

where $\hat{\theta}^{(2)}$ is nested within $\hat{\theta}^{(1)}$, so $L \geq 0$. This is called the scaled deviance.

The statistic $D = \phi L$, which does not depend explicitly on ϕ, is called the deviance.

Model comparison for GLM's

Suppose that $\hat{\theta}^{(1)}$ is the saturated model, in which $\theta_i = y_i$. If the GLM is Gaussian and $g(x) \equiv x$, as discussed above, the deviance is

$$\begin{aligned} D &= 2\sum_i (y_i^2 - y_i^2/2) - 2\sum_i (y_i \hat{\theta}_i^{(2)} - \hat{\theta}_i^{(2)\,2}/2) \\ &= \sum_i y_i^2 - 2y_i\hat{\theta}_i^{(2)} + \hat{\theta}_i^{(2)\,2} \\ &= \sum_i (y_i - \hat{\theta}_i^{(2)})^2. \end{aligned}$$

Model comparison for GLM's

Thus in the Gaussian case, the deviance is the residual sum of squares for the smaller model ($\hat{\theta}^{(2)}$).

In the Gaussian case, $D/\phi = L \sim \chi^2_{n-p-1}$.

When ϕ is unknown, we can turn this around to produce an estimate of the scale parameter

$$\hat{\phi} = \frac{D}{n-p-1}.$$

This is an unbiased estimate in the Gaussian case, but is useful for any GLM.

Model comparison for GLM's

Now suppose we want to compare two nested generalized linear models with deviances $D_1 < D_2$. Let $p_1 > p_2$ be the number of covariates in each model. The likelihood ratio test statistic is

$$L_2 - L_1 = \frac{D_2 - D_1}{\phi}$$

which asymptotically has a $\chi^2_{p_1-p_2}$ distribution.

If ϕ is unknown, we can estimate it as described above (using the larger of the two models).

The "plug-in" likelihood ratio statistic $(D_2 - D_1)/\hat{\phi}$ is still asymptotically $\chi^2_{p_1-p_2}$, as long as $\hat{\phi}$ is consistent.

The finite sample distribution may be better approximated using

$$\frac{D_2 - D_1}{\hat{\phi}(p_1 - p_2)} \approx F_{p_1-p_2, n-p_1},$$

Model comparison for GLM's

We can compare any two fitted GLM's using model selection statistics like AIC or BIC.

AIC favors models having small values of $L_{\text{opt}} - \text{df}$, where L_{opt} is the maximized log-likelihood, and df is the degrees of freedom. Equivalently, the AIC can be expressed

$$-D/2\hat{\phi} - p - 1.$$

The same $\hat{\phi}$ value should be used for all models being compared (i.e. by using the one from the largest model).