Generalized Linear Models

Kerby Shedden

Department of Statistics, University of Michigan

December 12, 2014

Motivation for nonlinear models

The key properties of a linear model are that

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

In some cases where these conditions are not met, we can transform Y so that these conditions are approximately satisfied.

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

Another challenge is that if Y lies in a restricted domain (e.g. $Y = 0, 1$), 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 are not appropriate.

Logistic regression

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

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

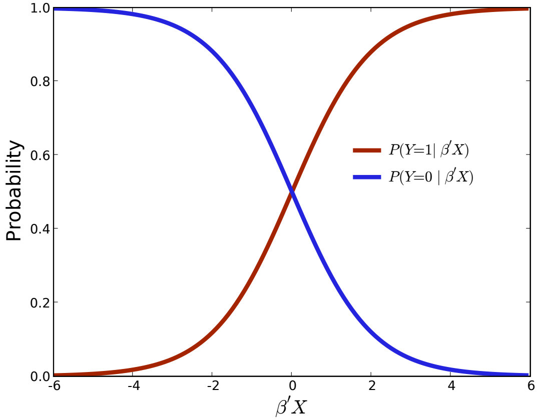
$$P(Y = 1|X = x) = \frac{\exp(\beta' x)}{1 + \exp(\beta' x)} = \frac{1}{1 + \exp(-\beta' x)},$$

which implies that

$$P(Y = 0|X = x) = 1 - P(Y = 1|X = x) = \frac{1}{1 + \exp(\beta' x)}.$$

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

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

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

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

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 results in an additive change of β_j in the conditional log odds, or 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 variance

$$\mathrm{var}(Y|X) = P(Y=1|X) \cdot P(Y=0|X) = \frac{1}{2+\exp(\beta' X)+\exp(-\beta' X)}.$$

Sicne 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 that may predict a person's disease status.

The prospective log odds

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

measures how informative the marker is for the disease.

Logistic regression and case-control studies

Suppose we model $M|D$ using logistic regression, so

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

The prospective log odds can be written

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

which equals

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

Logistic regression and case-control studies

If we had prospective data and used logistic regression to model the prosepctive relationship $D|M$, the log odds would have the form

$$\theta + \beta M.$$

Therefore we have shown that the coefficient β when we use logistic regression to regress M on D using case-control data is the same coefficient (in the population sense) as we would obtain from regressing D on M in a prospective study.

Note that the intercepts are not the same in general.

Estimation and inference for logistic regression

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

$$\begin{aligned} L(\beta|Y,X) &= \log \prod_i \frac{\exp(Y_i \cdot \beta' X_i)}{1+\exp(\beta' X_i)} \\ &= \beta' \sum_i Y_i X_i - \sum_i \log(1+\exp(\beta' X_i)). \end{aligned}$$

The statistic $\sum_i Y_i X_i$ is sufficient for β.

This likelihood is for the conditional distribution of Y given X. As in linear regression, we do not model the marginal distribution of X.

Estimation and inference for logistic regression

Logistic regression models are usually fit using maximum likelihood estimation.

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

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

$$G(\beta|Y,X) = \frac{\partial}{\partial\beta} L(\beta|Y,X) = \sum_i Y_i X_i - \sum_i \frac{\exp(\beta' X_i)}{1+\exp(\beta' X_i)} X_i.$$

Here $X_i \in \mathcal{R}^{p+1}$ is taken as a column vector, so $G(\beta|Y,X) \in \mathcal{R}^{p+1}$.

Estimation and inference for logistic regression

We can also write the gradient in the form

$$\sum_i X_i(Y_i - \mu_i),$$

where $\mu_i = \exp(\beta' X_i)/(1 + \exp(\beta' X_i))$.

At the MLE, every component of the gradient vector equals zero. Thus, as in OLS, the residuals are orthogonal to the covariates.

However the fitted values and residuals are not orthogonal in logistic regression.

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.

Here $X_i \in \mathcal{R}^{p+1}$ is taken as a column vector, so $H(\beta|Y,X) \in \mathcal{R}^{p+1\times p+1}$.

Estimation and inference for logistic regression

From general theory about the MLE, the inverse of the Fisher information matrix

$$I(\beta)^{-1} = -[EH(\beta|Y, X)|X]^{-1}$$

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

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

Suppose we have a dependent variable that can be interpreted like a count. For example, we may be interested in risk factors for asthma attacks, with Y_i being the number of attacks experienced by the i^{th} subject in a sample.

Data like this are often right-skewed. If they truly reflect the number of occurences of an event that occurs like a Poisson process, then $Y|X$ should follow a Poisson distribution.

Poisson regression

If the individuals in the population are different in terms of their risk for asthma attacks, we may like to relate covariates to the number of attacks.

In Poisson regression, the distribution of $Y|X$ follows a Poisson distribution, with the mean response related to the covariates via

$$\log E[Y|X] = \beta' X.$$

This gives us a likelihood function

$$\prod_i \exp(-\lambda_i)\lambda_i^{Y_i}/Y_i!,$$

where $\lambda_i = \exp(\beta' X_i)$.

Poisson regression

The log-likelihood function is

$$-\sum_i \exp(\beta' X_i) + \beta' \sum_i Y_i X_i - \sum_i \log Y_i!,$$

so $\sum_i Y_i X_i$ is a sufficient statistic.

Due to the exponential structure of the mean function, increasing the covariate X_j by 1 unit is associated with an increase in the mean (and variance) by a factor of $\exp(\beta_j)$.

Poisson regression

The score function is

$$G(\beta|Y,X) = \sum_i Y_i X_i - \sum_i X_i \exp(\beta' X_i).$$

The mean response is $\mu_i = \exp(\beta' X_i)$. Thus when the score is zero we have

$$\sum_i X_i(Y_i - \mu_i) = 0,$$

so the residuals are orthongonal to every covariate.

Negative binomial regression

In a Poisson distribution, the variance is equal to the mean. In many real data sets, the variance is notably larger than the mean. This is called overdispersion.

The negative binomial distribution has an additional parameter that allows the variance to differ from the mean:

$$p(Y = y|\mu, \alpha) = \frac{\Gamma(y + \alpha)}{\Gamma(\alpha)y!} \cdot \frac{\mu^y \alpha^\alpha}{(\mu + \alpha)^{y+\alpha}}.$$

The mean is α, and the variance is $\mu + \mu^2/\alpha$.

We can define a regression model using the negative binomial distribution by setting $\mu = \exp(\beta' X)$, just as in Poisson regression.

General development of GLM's

In the GLM modeling framework:

- 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.

General development of GLM's

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

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

General development of GLM's

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.

General development of GLM's

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.

General development of GLM's

We can also write

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

where $\mu_i \equiv E[Y_i|X]$.

The composed function $g^{-1} \circ \gamma'^{-1}$ maps the expected value of the response variable μ_i to the linear predictor $\beta' X_i$. It is called the link function.

When $g(x) \equiv x$, the link function becomes γ'^{-1}, and it is called the canonical link function.

General development of GLM's

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

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

General development of GLM's

The expected value of the Hessian is

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

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 canonical GLM form by setting $g(x) = x$,
$\gamma(x) = x^2/2$, $w_i = 1$, $\phi = \sigma^2$, and
$\tau(Y_i, \phi) = -\log(2\pi\phi)/2 - Y_i^2/2\phi$.

The (canonical) link function is therefore $\gamma'^{-1}(x) = x$, that is,
$E[Y|X] = \beta' X$.

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

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

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

Examples of GLM's

Logistic regression (continued):

The canonical link function for logistic regression is therefore the logit function

$$\gamma'^{-1}(p) = \log(p/(1-p)).$$

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

The canonical link function is $\gamma'^{-1}(x) = \log(x)$.

Examples of GLM’s

Other examples of GLM’s include negative binomial GLM’s, Gamma GLM’s, and inverse Gaussian GLM’s.

As discussed earlier, the usual link function for the negative binomial GLM is the log function, which is not the canonical link function.

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 $L_{\rm opt} - {\rm df}$, where $L_{\rm opt}$ 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).

Estimating Equations

The normal equations for OLS can be written in the form:

$$X'(Y - X\beta) = 0.$$

These are sometimes called estimating equations for the model parameters β.

If we observe data in groups, with $X_{(i)}\imath\mathcal{R}^{n_i \times p+1}$ and $Y_{(i)} \in \mathcal{R}^{p+1}$ denoting the data in the i^{th} group, then the estimating equations can be expressed

$$\sum_i X'_{(i)}(Y_{(i)} - X_{(i)}\beta) = 0.$$

Estimating Equations

If we model the errors in group i as having covariance matrix $\Sigma_{(i)}$, then the estimating equations become

$$\sum_i X'_{(i)} \Sigma_{(i)}^{-1} (Y_{(i)} - X_{(i)}\beta) = 0.$$

Estimating Equations

Next we derive the score function for GLM's when the data are observed in groups.

The linear predictor can be written $\eta_{(i)} = X_{(i)}\beta = g(\theta_{(i)})$ (with g applied element-wise). If we are using the canonical link, then $\theta_{(i)} \equiv \eta_{(i)}$.

Thus using the chain rule, the score function with respect to β is

$$\sum_i w_i X'_{(i)} d\theta_{(i)}/d\eta_{(i)} \cdot (Y_{(i)} - \gamma'(\theta_{(i)})).$$

Checking dimensions, $d\theta_{(i)}/d\eta_{(i)} \in \mathcal{R}^{p+1\times p+1}$ (will be the identity matrix when using the canonical link) and $Y_{(i)} - \gamma'(\theta_{(i)} \in \mathcal{R}^{p+1}$ is a vector of residuals for group i.

Estimating Equations

Let $\Delta_i = d\theta_{(i)}/d\eta_{(i)}$ and $S_i = Y_{(i)} - \gamma'(\theta_{(i)}) = Y_{(i)} - \mu_i$. The estimating equations become

$$\sum_i w_i X'_{(i)} \Delta_i S_i = 0.$$

Next we will derive standard errors for these estimates. We do this by linearizing the mean

$$\mu_i(\hat{\beta}) = \mu_i(\beta) + J_i(\beta)(\hat{\beta} - \beta),$$

where $J_i(\beta) = d\mu_i/d\beta \in \mathcal{R}^{n_i \times p+1}$ is the Jacobian matrix.

Estimating Equations

The linearized estimating equations are

$$\sum_i X'_{(i)}\Delta_i(Y_i - \mu_i(\beta) - J_i(\beta))(\hat{\beta} - \beta),$$

and the solution to the linearized estimating equations is

$$\hat{\beta} - \beta \approx [\sum_i X'_{(i)}\Delta_i J_i(\beta)]^{-1} \sum X'_{(i)}\Delta_i S_i.$$

Thus we obtain an approximate sampling covariance matrix for $\hat{\beta}$:

$$[\sum_i X'_{(i)}\Delta_i J_i(\beta)]^{-1} \cdot [\sum_i X'_{(i)}\Delta_i \mathrm{cov}(Y_{(i)})\Delta_i X_{(i)}] \cdot [\sum_i X'_{(i)}\Delta_i J(\beta)]^{-1}.$$

Estimating Equations

Using the chain rule, we can obtain

$$J_i(\beta) = A_i \Delta_i X_i,$$

where A_i is the $n_i \times n_i$ matrix with diagonal equal to $\gamma''(\theta_i)$. Note that these values are the variances of the components of $Y_{(i)}$, up to a factor of ϕ.

Therefore the approximate covariance matrix of $\hat{\beta}$ becomes

$$C^{-1} \cdot [\sum_i X'_{(i)} \Delta_i \mathrm{cov}(Y_{(i)}) \Delta_i X_{(i)}] \cdot C^{-1}.$$

where

$$C \equiv \sum_i X'_{(i)} \Delta_i A_i \Delta_i X_{(i)}^{-1}.$$

Estimating Equations

We only have a single residual vector $\hat{S}_i = Y_i - \hat{\mu}_i$ for each group, but we can use this to estimate the covariance of $Y_{(i)}$ as $\hat{S}_i\hat{S}_i'$. If we set

$$B \equiv \sum_i X_{(i)}'\Delta_i\hat{S}_i\hat{S}_i'\Delta_i X_{(i)}$$

then we obtain the approximate covariance matrix

$$\mathrm{cov}(\hat{\beta}) \approx C^{-1}BC^{-1}.$$

Note that this expression for $\mathrm{cov}(\hat{\beta})$ holds even when there are correlations among the cases within a group, so it is a "robust" covariance.

Generalized Estimating Equations (GEE)

The robust covariance derived above allows us to use standard GLM estimates of β when the data are dependent. We then use the robust variances in place of the usual variances obtained from the Fisher information matrix.

However GLM may not give us the most efficient possible estimates of β when the data are correlated. We can recover much of the lost efficiency by using a technique analogous to GLS called “Generalized Estimating Equations” (GEE).

Generalized Estimating Equations (GEE)

Suppose we wish to minimize the quadratic loss function

$$(Y-\mu)'\Sigma^{-1}(Y-\mu).$$

The gradient of this expression is

$$\mu'\Sigma^{-1}(Y-\mu),$$

where $\mu' = d\mu/d\beta \in \mathcal{R}^{p\times n}$ is the Jacobian obtained when differentiating μ with respect to β.

For grouped data, this leads to the estimating equations

$$\sum_i \mu_i'\Sigma_i^{-1}(Y_{(i)}-\mu_i).$$

Generalized Estimating Equations (GEE)

For a GLM, we know that $\mu(x) = \gamma'(g(\beta' x))$. Therefore,

$$\mu' = \gamma''(g(\beta' x))g'(\beta' x)x.$$

In the canonical case, this simplifies to

$$\mu' = \gamma''(\beta' x)x = X_{(i)}' \mathrm{Var}[Y_{(i)}|X_{(i)}]/\phi,$$

so our estimating equations for β become

$$\sum_i X_{(i)}' A_i \Sigma_i^{-1}(Y_{(i)} - \mu_i),$$

where A_i is the diagonal matrix whose diagonal elements are $\mathrm{Var}[Y_{(i)}|X_{(i)}]$.

Generalized Estimating Equations (GEE)

In the non-canonical case, the estimating equations are

$$\sum_i X'_{(i)} \Delta_i A_i \Sigma_i^{-1} (Y_{(i)} - \mu_i),$$

where $\Delta_i = d\theta_{(i)}/d\eta_{(i)} = g'(\eta_{(i)})$ as defined above.

In both the canonical and non-canonical cases, the estimating equations are easily solved using the Gauss-Seidel algorithm.

Generalized Estimating Equations (GEE)

The remaining issue is that we don't know the Σ_i.

To address this, we specify a "working model" $R_i(\alpha) \in \mathcal{R}^{n_i \times n_i}$ for the correlation matrix of $Y_{(i)}$. Common choices for $R(\alpha)$ would be exchangeable or autoregressive correlation matrices.

The variances are determined by the particular GLM family that we are using, so we have

$$V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2}$$

as a "working covariance matrix" for the data. We then substitute V_i for Σ_i when solving the estimating equations.

As in GLS, we may alternate several times between solving the estimating equations, and updating the parameter α that determines R_i.

Generalized Estimating Equations (GEE)

The variance estimator for GEE is $C^{-1}BC^{-1}$, where

$$B = \sum_i X'_{(i)} \Delta_i A_i V_i^{-1} \hat{S}_i \hat{S}'_i V_i^{-1} A_i \Delta_i X_{(i)},$$

$$C = \sum_i X'_{(i)} \Delta_i A_i V_i^{-1} A_i \Delta_i X_{(i)}.$$