

0.1 GLM

Typical regression setup

$$E[Y_i] = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta} \text{ and}$$

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

Regression for non-normal data. General framework - *generalized linear models*

Components of a generalized linear model (glm)

1. random component
2. systematic component
3. link between the random and systematic components

the first two related to explained vs. unexplained variability

1. Random component

$$\mathbf{y} = (y_1, \dots, y_n)$$

pdf for y_i has form of exponential dispersion family,

$$f(y_i; \theta_i, \phi) = \exp\{[y_i \theta_i - b(\theta_i)]/a(\phi) + c(y_i; \phi)\}$$

- **NOTE:** in the text, $b(\theta_i) = \theta_i/a(\phi)$ and $c(\theta_i) = -b(\theta_i)/a(\phi)$ and $d(y_i) = c(y_i; \phi)$
- $\{y_i\}$ are jointly independent
- ϕ is scale or dispersion parameter
- usually $a(\phi) = \phi$ or ϕ/w_i for known weight w_i
- θ_i is the natural parameter
- general form has $a(y_i)\theta_i$; if $a(y_i) = y_i$, *canonical form*
- choices of $b(\theta)$ and $a(\phi)$ give rise to different pdf's

Examples

Poisson

$$\begin{aligned} f(y_i; \theta_i, \phi) &= \frac{\exp(-\lambda_i) \lambda_i^{y_i}}{y_i!} \\ &= \exp\{-\lambda_i + y_i \log \lambda_i + \log y_i!\} \\ &= \exp\{[y_i \log \lambda_i - \lambda_i] + \log y_i!\} \end{aligned}$$

where

$$\begin{aligned} \theta_i &= \log \lambda_i & a(\phi) &= 1 \\ b(\theta_i) &= \lambda_i = \exp(\theta_i) \\ c(y_i; \phi) &= \log y_i! \end{aligned}$$

Binomial

define as $n_i y_i$ where $y_i \sim Ber(\pi_i)$

$$\begin{aligned} f(y_i; \theta_i, \phi) &= \binom{n_i}{n_i y_i} \pi_i^{n_i y_i} (1 - \pi_i)^{n_i(1-y_i)} \\ &= \exp\{n_i y_i \log \pi_i + (n_i - n_i y_i) \log(1 - \pi_i) + \log \binom{n_i}{n_i y_i}\} \\ &= \exp\{[y_i \log \frac{\pi_i}{1 - \pi_i} + \log(1 - \pi_i)]/(1/n_i) + \log \binom{n_i}{n_i y_i}\} \end{aligned}$$

where

$$\begin{aligned} \theta_i &= \log \frac{\pi_i}{1 - \pi_i} \quad a(\phi) = 1/n_i \\ b(\theta_i) &= -\log(1 - \pi_i) = \log(1 + \exp\{\theta_i\}) \\ c(y_i; \phi) &= \log \binom{n_i}{n_i y_i} \end{aligned}$$

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

$$\begin{aligned} f(y_i; \theta_i, \phi) &= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\} \\ &= \exp\{-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} y_i^2 + \frac{1}{\sigma^2} y_i \mu_i - \frac{1}{2\sigma^2} \mu_i^2\} \\ &= \exp\{[y_i \mu_i - \mu_i^2/2]/\sigma^2 + [-\frac{1}{2}(\log(2\pi) + \log \sigma^2 + y_i^2/\sigma^2)]\} \end{aligned}$$

where

$$\begin{aligned} \theta_i &= \mu_i \quad a(\phi) = \sigma^2 \\ b(\theta_i) &= \frac{1}{2} \mu_i^2 = \frac{1}{2} \theta_i^2 \\ c(y_i; \phi) &= [-\frac{1}{2}(\log(2\pi) + \log \sigma^2 + y_i^2/\sigma^2)] \end{aligned}$$

Moments

define $l(\theta_i, \phi; y_i) = \log f(y_i; \theta_i, \phi) = [y_i \theta_i - b(\theta_i)]/a(\phi) + c(y_i; \phi)$

So, $\frac{\partial l}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)}$

Note:

$$\begin{aligned}
 E \left[\frac{\partial l}{\partial \theta_i} \right] &= E \left[\frac{f'(y_i; \theta_i, \phi)}{f(y_i; \theta_i, \phi)} \right] \\
 &= \int \frac{f'(y_i; \theta_i, \phi)}{f(y_i; \theta_i, \phi)} f(y_i; \theta_i, \phi) dy_i \\
 &= \frac{\partial}{\partial \theta} \int f(y_i; \theta_i, \phi) dy_i \\
 &= \frac{\partial}{\partial \theta} (1) = 0
 \end{aligned}$$

This implies

$$E \left[\frac{y_i - b'(\theta_i)}{a(\phi)} \right] = 0 \text{ and } E[y_i] = b'(\theta_i) (= \mu_i).$$

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

Note: $E \left[\frac{\partial^2 l}{\partial \theta_i^2} \right] = -E \left[\left(\frac{\partial l}{\partial \theta_i} \right)^2 \right]$ (under regularity conditions satisfied by exponential family).

This implies

$$E \left[-\frac{b''(\theta_i)}{a(\phi)} \right] = -E \left[\left(\frac{y_i - b'(\theta_i)}{a(\phi)} \right)^2 \right] = -\frac{1}{a(\phi)^2} \text{Var}(y_i)$$

and $\text{Var}(y_i) = a(\phi)b''(\theta_i)$ (variance may depend on mean!)

Examples

Poisson(μ_i): $E[Y_i] = \mu_i$

$$b(\theta_i) = \exp\{\theta_i\} \quad a(\phi) = 1$$

$$\begin{aligned}
 E(Y_i) &= b'(\theta_i) = \exp\{\theta_i\} = \mu_i \\
 \text{Var}(Y_i) &= b''(\theta_i) = \exp\{\theta_i\} = \mu_i
 \end{aligned}$$

$n_i y_i \sim \text{Binomial}(n_i, \pi_i)$: $E[Y_i] = \pi_i$

$$\begin{aligned}
 \theta_i &= \log \frac{\pi_i}{1 - \pi_i} = \text{logit}(\pi_i) \\
 b(\theta_i) &= \log(1 + \exp(\theta_i)) \\
 a(\phi) &= 1/n_i
 \end{aligned}$$

$$E(Y_i) = b'(\theta_i) = \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} = \pi_i$$

$$\text{Var}(Y_i) = b''(\theta_i)a(\phi) = \frac{\exp(\theta_i)}{(1 + \exp(\theta_i))^2} \frac{1}{n_i} = \frac{\exp(\theta_i)}{(1 + \exp(\theta_i))} \frac{1}{(1 + \exp(\theta_i))} \frac{1}{n_i} = \frac{\pi_i(1 - \pi_i)}{n_i}$$

2. Systematic component (linear predictor)

$\mathbf{x}_i = (x_{i0}, \dots, x_{ip})^T$ vector of explanatory variables for subject i

$$\eta_i = \mathbf{x}_i^T \boldsymbol{\beta} = \sum_{j=0}^p x_{ij} \beta_j$$

3. Link function: connect linear predictor to mean

$\mu_i = E[Y_i]$: linked to linear predictor by $\eta_i = g(\mu_i)$ where g (link function) is any monotone differentiable function; $g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$

The link g s.t. $g(\mu_i) = \theta_i$ is called the *canonical* link, $g^{-1} = b'$

Examples

Poisson

$$E[Y_i] = \mu_i = \exp(\theta_i) = b'(\theta_i)$$

so $g(\mu_i) = \log(\mu_i)$ is the *canonical* link

Binomial: $n_i Y_i \sim \text{Bin}(n_i \pi_i)$

$$E[Y_i] = \pi_i = \exp(\theta_i) / (1 + \exp(\theta_i))$$

So, $g(\pi_i) = \log(\pi_i / (1 - \pi_i))$ is the *canonical* link (logit link)

Normal

canonical link is the identity link

Score and Information

Score

$$U(\theta; \mathbf{y}) = \frac{\partial l(\theta; \mathbf{y})}{\partial \theta} = \frac{y - b'(\theta)}{a(\phi)}$$

Note:

$$\begin{aligned} E[U] &= \frac{b'(\theta) - b'(\theta)}{a(\phi)} \\ &= 0. \end{aligned}$$

The variance of U is the information, I ,

$$\begin{aligned} I[U] &= \text{Var}(U) \\ &= \frac{1}{a(\phi)^2} \text{Var}(Y) \\ &= \frac{1}{a(\phi)^2} a(\phi) b''(\theta) \\ &= \frac{1}{a(\phi)} b''(\theta) \end{aligned}$$

and

$$\text{Var}(U) = E(U^2) = -E(U')$$

2nd equality from

$$U' = \frac{\partial U}{\partial \theta} = -\frac{b''(\theta)}{a(\phi)}$$

So,

$$\begin{aligned} E(U') &= -\frac{b''(\theta)}{a(\phi)} \\ &= -\text{var}(U) = -I \end{aligned}$$

NOTE: the below rewrites what is above to match the text.

Score

$$U(\theta; \mathbf{y}) = \frac{\partial l(\theta; \mathbf{y})}{\partial \theta} = a(y)b'(\theta) + c'(\theta)$$

Note:

$$\begin{aligned} E[U] &= b'(\theta)\left[-\frac{c'(\theta)}{b'(\theta)}\right] + c'(\theta) \\ &= 0. \end{aligned}$$

The variance of U is the information, I ,

$$\begin{aligned} I[U] &= \text{Var}(U) \\ &= [b'(\theta)^2]\text{Var}[a(Y)] \\ &= [b'(\theta)^2]\frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3} \\ &= \frac{b''(\theta)c'(\theta)}{b'(\theta)} - c''(\theta) \end{aligned}$$

and

$$\text{Var}(U) = E(U^2) = -E(U')$$

2nd equality from

$$U' = \frac{\partial U}{\partial \theta} = a(Y)b''(\theta) + c''(\theta)$$

So,

$$\begin{aligned} E(U') &= b''(\theta)E[a(Y)] + c''(\theta) \\ &= b''(\theta)\left[-\frac{c'(\theta)}{b'(\theta)}\right] + c''(\theta) \\ &= -\text{var}(U) = -I \end{aligned}$$

0.2 Maximum likelihood parameter estimation

Ch 4 in book

For n independent observations

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{y}) &= \log \prod_{i=1}^n f(y_i; \theta_i, \phi) \\ &= \sum \log f(y_i; \theta_i, \phi) \\ &= \sum \{[y_i \theta_i - b(\theta_i)]/a(\phi) + c(y_i; \phi)\} \end{aligned}$$

For the canonical link,

$$\eta_i = \theta_i = \sum_{j=0}^p X_{ij} \beta_j$$

So,

$$\rightarrow L(\boldsymbol{\theta}; \mathbf{y}) = \sum \{[y_i \sum_{j=0}^p x_{ij} \beta_j - b(\sum_{j=0}^p x_{ij} \beta_j)]/a(\phi) + c(y_i; \phi)\}$$

Sufficient statistics are

$$\left\{ \sum_{i=1}^n y_i x_{ij} : j = 1, \dots, p \right\}$$

Now, let's derive the likelihood equations,

Contribution of observation i to the log likelihood is

$$\begin{aligned} l_i &= [y_i \theta_i - b(\theta_i)]/a(\phi) + c(y_i; \phi) \\ \frac{\partial l_i}{\partial \beta_j} &= \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\ \frac{\partial l_i}{\partial \theta_i} &= [y_i - b'(\theta_i)]/a(\phi) = [y_i - \mu_i]/a(\phi) \\ \frac{\partial \theta_i}{\partial \mu_i} &\rightarrow \frac{\partial \mu_i}{\partial \theta_i} = \frac{\partial b'(\theta_i)}{\partial \theta_i} = b''(\theta_i) = \text{Var}(Y_i)/a(\phi) \\ \frac{\partial \mu_i}{\partial \eta_i} &= \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \\ \frac{\partial \eta_i}{\partial \beta_j} &= x_{ij} \end{aligned}$$

So, putting all the pieces together,

$$\frac{\partial l_i}{\partial \beta_j} = \left(\frac{y_i - \mu_i}{a(\phi)} \right) \left(\frac{a(\phi)}{\text{Var}(Y_i)} \right) \left(\frac{\partial \mu_i}{\partial \eta_i} \right) x_{ij}$$

And the likelihood equations are

$$\sum_{i=1}^n \frac{(y_i - \mu_i)x_{ij}}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) = 0 \quad j = 0, \dots, p$$

Remarks:

- β implicitly through $\mu_i = g^{-1}(\sum x_{ij}\beta_j)$
- nonlinear functions of β (solve iteratively; more in a bit)
- depend on distribution of Y_i **only** through μ_i and $\text{Var}(Y_i)$ (which may depend on μ_i)
- when Y_i has distribution in natural exponential family, the relationship $\sigma_i^2 = V(\mu_i)$ between the mean and the variance characterizes the distribution (Jorgensen, 1987)

Asymptotic Covariance Matrix

Recall, MLE $\hat{\beta}$ is asymptotically $N(\beta, I^{-1})$ where I is the expected information matrix,

$$I_{jk} = E \left[-\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k} \right] = E \left[\frac{\partial l_i}{\partial \beta_j} \frac{\partial l_i}{\partial \beta_k} \right]$$

Recall,

$$\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^n \frac{\partial^2 l_i}{\partial \beta_j \partial \beta_k}$$

and

$$\frac{\partial l_i}{\partial \beta_j} = \frac{(y_i - \mu_i)x_{ij}}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) = 0.$$

Finally,

$$\begin{aligned} E \left[\frac{\partial l_i}{\partial \beta_j} \frac{\partial l_i}{\partial \beta_k} \right] &= E \left[\frac{(y_i - \mu_i)x_{ij}}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \frac{(y_i - \mu_i)x_{ik}}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right] \\ &= \frac{x_{ij}x_{ik}}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \\ &= w_i x_{ij} x_{ik}, \end{aligned}$$

where $w_i = \frac{1}{\text{Var}(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$

So,

$$E \left[-\frac{\partial^2 L(\beta)}{\partial \beta_j \partial \beta_k} \right] = \sum_{i=1}^n w_i x_{ij} x_{ik} = I_{jk}$$

This can be rewritten in matrix form as

$$I = X'WX,$$

where W is a diagonal matrix with elements w_i and X has rows X_i . Recall, $w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 / \text{Var}(Y_i)$.

An iterative estimation procedure we can use here is the method of scoring,

$$b^{(m)} = b^{(m-1)} + [I^{(m-1)}]^{-1}U^{(m-1)},$$

which can be rewritten as

$$I^{(m-1)}b^{(m)} = I^{(m-1)}b^{(m-1)} + U^{(m-1)},$$

and then plugging in the expressions for the information and score here, we obtain

$$X'WXb^{(m)} = X'Wz,$$

where $z_i = \sum_{j=0}^p x_{ij}b_j^{(m-1)} + (y_i - \mu_i) \left(\frac{\partial \eta_i}{\partial \mu_i}\right)$.

Remarks

- the 'response' z_i is a linearized version of $g(y_i)$ evaluated at μ_i ,

$$z_i = g(\mu_i^{(m-1)}) + (y_i - \mu_i^{(m-1)})g'(\mu_i^{(m-1)})$$

- looks like the normal equations for linear regression
- can solve using IRLS; initial value for β , solve, update weights, iterate until convergence
- estimate asymptotic covariance matrix of $\hat{\beta}$ by

$$\widehat{\text{Var}}(\hat{\beta}) = \hat{I}^{-1} = (X'\hat{W}X)^{-1}$$

where $\hat{W} = W(\hat{\beta})$.

- for the canonical link, expected and observed information are the same. This implies Fisher scoring and Newton Raphson (which uses the observed information) algorithms are identical (show for HW)

Illustration with Poisson model

$$Y_i \sim P(\mu_i)$$

Suppose $\eta_i = \log \mu_i = \mathbf{X}_i\boldsymbol{\beta}$, then

$\mu_i = \exp(\eta_i)$ and $\partial \mu_i / \partial \eta_i = \exp(\eta_i) = \mu_i$ and the lik eqs. become

$$\sum \frac{(y_i - \mu_i)x_{ij}}{\mu_i} \mu_i = 0$$

This implies $\sum y_i x_{ij} = \sum \mu_i x_{ij}$, i.e., equate sufficient statistics to their expected values. Expected information matrix is

$$I_{jk} = \sum x_{ij}x_{ik}\mu_i$$

where $w_i = \mu_i$. Estimate covariance matrix for $\hat{\beta}$ by plugging in $\hat{\mu}_i$'s.

0.3 Deviance, Goodness of fit, and Model Comparison

Ch 5 in book

For model of interest, $\{\hat{\mu}_i\}, \{\hat{\theta}_i\}$

For 'saturated' model, $\eta_i = \beta_i, i = 1, \dots, n$ and has ML estimate $\tilde{\mu}_i = y_i, i = 1, \dots, n$.

Clearly, all the variation in the y 's is assigned to the systematic component.

Now, recall the log likelihood (in terms of θ for ϕ fixed) is

$$L(\theta; \mathbf{y}) = \sum_i [y_i \theta_i - b(\theta_i)]/a(\phi) + c(y_i; \phi)$$

Let $L(\hat{\theta}; \mathbf{y}) = L(\hat{\mu}; \mathbf{y})$ = the log lik maximized over β (model of interest)

Let $L(\tilde{\theta}; \mathbf{y}) = L(\mathbf{y}; \mathbf{y})$ = the log lik for the saturated model

The likelihood ratio test for H_0 : [the model holds], is

$$\begin{aligned} -2 \log \left[\frac{\exp(L(\hat{\theta}; \mathbf{y}))}{\exp(L(\tilde{\theta}; \mathbf{y}))} \right] &= 2[L(\mathbf{y}; \mathbf{y}) - L(\hat{\mu}; \mathbf{y})] \\ &= 2 \sum_i [y_i \tilde{\theta}_i - b(\tilde{\theta}_i)]/a(\phi) \\ &\quad - 2 \sum_i [y_i \hat{\theta}_i - b(\hat{\theta}_i)]/a(\phi) \end{aligned}$$

Suppose $a(\phi) = \phi/w_i$

$$\begin{aligned} &= 2 \sum_i w_i [y_i (\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]/\phi \\ &= D(\mathbf{y}, \hat{\mu})/\phi \end{aligned}$$

The quantity above is the *scaled deviance*. The *deviance* is $D(\mathbf{y}, \hat{\mu})$.

Remarks

- for binomial and Poisson models, $\phi = 1$ so scaled deviance is equal to the deviance
- $D(\mathbf{y}, \hat{\mu}) \geq 0$; the greater the deviance, the poorer the fit; just the difference in log likelihood
- for normal model, scaled deviance is distributed as a χ^2 with $n - (p + 1)$ df
- for Poisson model (with large μ_i), scaled deviance is approximately distributed as a χ^2 with $n - (p + 1)$ df
- in general, when ϕ/w_i are small, Y_i approximately normal; e.g., binomial with large n_i 's
- for other cases, e.g., binary with continuous explanatory variables, NOT asymptotically χ^2 .

Examples

Poisson, $Y_i \sim P(\mu_i)$, with canonical link ($\theta_i = \log \mu_i$)

so, $\hat{\theta}_i = \log \hat{\mu}_i$ and $\tilde{\theta}_i = \log y_i$ and $\phi = 1$

The deviance is equal to the scaled deviance,

$$D = 2 \sum_i [y_i \log \frac{y_i}{\hat{\mu}_i} - y_i + \hat{\mu}_i]$$

Note: likelihood equations are $\sum_i y_i x_{ij} = \sum_i \hat{\mu}_i x_{ij} : j = 0, \dots, p$. If $x_{i0} = 1$ (an intercept), then the first likelihood equation is $\sum y_i = \sum \hat{\mu}_i$ and the deviance simplifies to

$$D = 2 \sum_i [y_i \log \frac{y_i}{\hat{\mu}_i}].$$

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

$D(\mathbf{y}; \boldsymbol{\mu}) = \sum_i (y_i - \hat{\mu}_i)^2$, the residual sum of squares.

The scaled deviance is

$$D(\mathbf{y}; \boldsymbol{\mu})/\phi = \sum_i (y_i - \hat{\mu}_i)^2/\sigma^2.$$

Model Comparison

Consider model M_0 ($\hat{\boldsymbol{\mu}}_0$) nested within M_1 ($\hat{\boldsymbol{\mu}}_1$), where $p_0 < p_1$. So,

$$L(\hat{\boldsymbol{\mu}}_1; \mathbf{y}) \geq L(\hat{\boldsymbol{\mu}}_0; \mathbf{y})$$

LR statistic comparing models is

$$\begin{aligned} -2[L(\hat{\boldsymbol{\mu}}_0; \mathbf{y}) - L(\hat{\boldsymbol{\mu}}_1; \mathbf{y})] &= D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\mu}}_1) \geq 0 \\ &= 2 \sum w_i [y_i(\hat{\theta}_{i1} - \hat{\theta}_{i0}) - b(\hat{\theta}_{i1}) + b(\hat{\theta}_{i0})] \end{aligned}$$

Under regularity conditions, asymptotically χ^2 with $p_1 - p_0$ degrees of freedom.

Example

Poisson with log link and an intercept. Recall

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2 \sum_i y_i \log \frac{y_i}{\hat{\mu}_i}$$

LR statistics for comparing M_0 to M_1 is

$$\begin{aligned} D(\mathbf{y}; \hat{\boldsymbol{\mu}}_0) - D(\mathbf{y}; \hat{\boldsymbol{\mu}}_1) &= 2 \sum_i y_i \log \frac{\hat{\mu}_{i1}}{\hat{\mu}_{i0}} \\ &= 2 \sum_i \hat{\mu}_{i1} \log \frac{\hat{\mu}_{i1}}{\hat{\mu}_{i0}} \end{aligned}$$

The 2nd equality follows from likelihood equations with an intercept and log link. Show for HW.

Comparison of nested GLMs by LRTs is called *analysis of deviance*.

In models where ϕ is unknown, typically estimate with moment estimator,

$$\hat{\phi} = \frac{1}{n - (p + 1)} \sum \frac{w_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

which looks like a re-scaled *Pearson's χ^2 statistic*. Estimate μ_i from the larger model.

Even when estimate scale parameter, difference in scaled deviances still asymptotically χ^2 as long as consistent estimator of scale parameter by Slutsky's theorem.

Remark

- when estimate ϕ , often better to approximate with an F-distribution; recall this is exact for a normal GLM

Another measure of model fit

Pearson's χ^2 statistic,

$$\chi^2(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^n \frac{w_i(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

,

and the scaled Pearson's χ^2 statistic,

$$\chi^2(\mathbf{y}; \hat{\boldsymbol{\mu}})/\phi = \sum_{i=1}^n \frac{w_i(y_i - \hat{\mu}_i)^2}{\phi V(\hat{\mu}_i)}$$

,

For normal GLM, these are equivalent to the (scaled) deviance.

Asymptotics for Deviance and Connection to Pearson's

Suppose $Y \sim f(\cdot; \theta, \phi)$ and

$$\begin{aligned} D(y; \mu) &= 2[\log f(y, \phi; y) - \log f(\mu, \phi; y)]\phi \\ &= 2\{y[\theta(y) - \theta(\mu)] - [b(\theta(y)) - b(\theta(\mu))]\} \end{aligned}$$

where $\theta(\mu) = (b')^{-1}(\mu)$. If ϕ is small, $Var(Y)$ is small (recall: $Var(Y) = \phi V(\mu)$) and Y will be close to μ with high probability. Now we will examine what happens to $D(y; \mu)$ in this case. Again, recall that $\partial\mu/\partial\theta = b''(\theta) = V(\mu)$ (implies $\partial\theta/\partial\mu = 1/V(\mu)$). Now consider a two term expansion of $D(y; \mu)$ about $y = \mu$:

$$\begin{aligned} D(y; \mu) &= 2\{y[\theta'(\mu)(y - \mu) + \frac{1}{2}\theta''(\mu)(y - \mu)^2] - [b'(\theta(\mu))\theta'(\mu)(y - \mu) \\ &\quad \frac{1}{2}\{b''(\theta(\mu))\theta'(\mu)^2 + b'(\theta(\mu))\theta''(\mu)\}(y - \mu)^2]\} + O(|y - \mu|^3) \\ &= 2\left\{\frac{y(y - \mu)}{V(\mu)} + \frac{1}{2}\theta''(\mu)y(y - \mu)^2 - \frac{\mu(y - \mu)}{V(\mu)} - \frac{1}{2}\frac{(y - \mu)^2}{V(\mu)} - \frac{1}{2}\theta''(\mu)\mu(y - \mu)^2\right\} \\ &\quad O(|y - \mu|^3) \\ &= \frac{(y - \mu)^2}{V(\mu)} + O(|y - \mu|^3). \end{aligned}$$

So, for fixed n and small ϕ (or large w_1, \dots, w_n), the observed responses y_1, \dots, y_n satisfy

$$D(\mathbf{y}; \boldsymbol{\mu}) = \sum_i D(y_i; \mu_i) \approx \sum_i \frac{(y_i - \mu_i)^2}{V(\mu_i)} = \chi^2(\mathbf{y}; \boldsymbol{\mu}).$$

Assuming the model is correct, the MLEs $\hat{\mu}$ should be close to μ (true value), and we can obtain the approximate equivalence above by plugging in $\hat{\mu}$.

Examples

Binomial

Suppose $mY \sim \text{Bin}(m, \pi)$. So $\mu = \pi$ and $\phi = 1/m$.

$$\begin{aligned} D(y; \mu) &= 2\{my[\log \frac{y}{1-y} - \log \frac{\mu}{1-\mu}] - m[-\log(1-y) + \log(1-\mu)]\} \\ &= 2\{my \log \frac{y}{\mu} + m(1-y) \log \frac{1-y}{1-\mu}\}. \end{aligned}$$

For $y_i : i = 1, \dots, n$,

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^n m_i \left\{ y_i \log \frac{y_i}{\hat{\mu}_i} (1-y_i) \log \frac{1-y_i}{1-\hat{\mu}_i} \right\}$$

$$\chi^2(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^n \frac{m_i (y_i - \hat{\mu}_i)^2}{\hat{\mu}_i (1 - \hat{\mu}_i)}.$$

For hw, show that the standard observed minus expected squared over expected for the $2n$ cells of success and failures is equivalent to Pearson's χ^2 above.

Poisson

$Y \sim P(\mu)$

$$D(y; \mu) = 2\{y[\log y - \log \mu] - (y - \mu)\} = 2\{y \log \frac{y}{\mu} - (y - \mu)\}$$

Thus,

$$D(\mathbf{y}; \boldsymbol{\mu}) = 2 \sum_i y_i \log \frac{y_i}{\mu_i} - (y_i - \mu_i)$$

and

$$\chi^2(\mathbf{y}; \boldsymbol{\mu}) = \sum_i \frac{(y_i - \mu_i)^2}{\mu_i}$$

Note here (again) that observed minus expected squared over expected.

0.3.1 Residuals

Pearson residual

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\text{Var}}(y_i)}}$$

Examples:

Poisson

$$e_i = \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}$$

Note: $\sum e_i^2$ is Pearson's χ^2 statistic. It is also the score statistic for H_0 : models holds.
Hw problem.

Binomial: $n_i Y_i \sim \text{Bin}(n_i, \pi_i)$

$$e_i = \frac{n_i y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}}$$

Deviance residual

Recall,

$$\begin{aligned} D(\mathbf{y}, \hat{\boldsymbol{\mu}}) &= 2 \sum_i w_i [y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)] \\ &= \sum_i d_i^* \end{aligned}$$

The deviance residual is defined as

$$d_i = \sqrt{d_i^*} \text{sign}(y_i - \hat{\mu}_i)$$

Poisson

$$d_i = \sqrt{2[y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i] \text{sign}(y_i - \hat{\mu}_i)}$$

Standardizing the residuals

Note: when $\phi = 1$,

$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = \sum d_i^*$ is often distributed approximately χ_{n-p}^2 under H_0 : the model holds.
So,

$$E[D(\mathbf{y}; \hat{\boldsymbol{\mu}})] \approx n - p$$

so

$$E[\sum d_i^*/n] \approx (n - p)/n < 1.$$

Similarly for Pearson,

$$E[\sum e_i^2/n] < 1.$$

Also, under H_0 ,

$E[e_i] \approx 0$ and $E[d_i] = E[\sqrt{d_i^*} \text{sign}(y_i - \hat{\mu}_i)] \approx 0$ (since probability of a positive residual is independent of $|d_i|$). Thus, the average variance of the residuals is less than 1.0.

So, we need to properly standardize them. First, we will examine $\text{Var}(\mathbf{y} - \hat{\boldsymbol{\mu}})$ as $n \rightarrow \infty$. Recall (asymptotically),

$$\text{Var}(\hat{\beta}) = (X'WX)^{-1},$$

where

$$\begin{aligned} W &= \text{diag}\{(\partial\mu_i/\partial\eta_i)^2/V(\mu_i)\} \\ &= DV(\boldsymbol{\mu})^{-1}D \end{aligned}$$

with $V(\boldsymbol{\mu}) = \text{diag}(V(\mu_i))$ and $D = \text{diag}\{\partial\mu_i/\partial\eta_i\}$. Also, note that $V(\boldsymbol{\mu}) = DW^{-1}D$ (will use later). By the delta method,

$$\text{Var}(\hat{\boldsymbol{\mu}}) = DX(X'WX)^{-1}XD,$$

($\hat{\boldsymbol{\mu}} = g^{-1}(x\hat{\boldsymbol{\beta}})$ which implies $\hat{\boldsymbol{\eta}} = X\hat{\boldsymbol{\beta}} = g(\hat{\boldsymbol{\mu}})$).

So have $\text{Var}(Y) = V(\mu)$ and $\text{Var}(\hat{\boldsymbol{\mu}})$ (above). To obtain $\text{Var}(\mathbf{Y} - \hat{\boldsymbol{\mu}})$ need their joint distribution. This can be found from the following theorem (from Pierce, Annals of Statistics, 1982, pp. 475-478) which shows the asymptotic effect of substituting estimators for parameters in certain types of statistics:

Theorem: Let $\hat{T}_n = T(\mathbf{x}; \boldsymbol{\theta})$ and $\hat{\boldsymbol{\theta}}_n = T(\mathbf{x}; \hat{\boldsymbol{\theta}})$. Assume,

1.

$$\sqrt{n} \begin{bmatrix} T_n \\ \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta} \end{bmatrix} \rightarrow N \left[\mathbf{0}, \begin{bmatrix} V_{11} & V_{12} \\ V'_{12} & V_{22} \end{bmatrix} \right]$$

2. exists a matrix \mathbf{B} such that

$$\sqrt{n}\hat{T}_n = \sqrt{n}T_n + \mathbf{B}\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + o_p(1).$$

[Aside: when T_n is differentiable in $\boldsymbol{\theta}$, this follows from first order expansion and $\mathbf{B} = \lim E \left[\frac{\partial T_n}{\partial \boldsymbol{\theta}} \right]$.

3. $\hat{\boldsymbol{\theta}}_n$ is asymptotically efficient.

Then,

$$\sqrt{n}\hat{T}_n \rightarrow N(\mathbf{0}, V_{11} - \mathbf{B}V_{22}\mathbf{B}').$$

Remarks:

1. $\sqrt{n}\hat{T}_n$ asymptotically has distribution of $\sqrt{n}T_n + \mathbf{B}\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})$. So,

$$\begin{aligned} \text{Var}(\sqrt{n}\hat{T}_n) &= \text{Var}(\sqrt{n}T_n) + \mathbf{B}\text{Var}(\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}))\mathbf{B}' \\ &\quad + \text{Cov}(\sqrt{n}T_n, \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}))\mathbf{B}' \\ &\quad + \mathbf{B}\text{Cov}(\sqrt{n}T_n, \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})) \end{aligned}$$

which implies $\mathbf{B} = -V_{12}V_{22}^{-1}$. \mathbf{B} taking this form is suggested by equating the variance expression above to the one given in the Theorem.

2. \hat{T}_n and $\hat{\theta}_n$ are asymptotically independent. This is an asymptotic version of the result that minimum variance unbiased estimators (here, $\hat{\theta}_n$ are uncorrelated with statistics having constant expectation (here, \hat{T}_n). See Rao (1973; Section 5a.2).

We now apply this to our setting. Let

$$\begin{aligned}\sqrt{n}T_n &= \mathbf{y} - \boldsymbol{\mu} \\ \sqrt{n}\hat{T}_n &= \mathbf{y} - \hat{\boldsymbol{\mu}} \\ \sqrt{n}\hat{\boldsymbol{\theta}}_n &= \hat{\boldsymbol{\mu}}.\end{aligned}$$

$$\begin{pmatrix} \mathbf{y} - \boldsymbol{\mu} \\ \hat{\boldsymbol{\mu}} - \boldsymbol{\mu} \end{pmatrix} \sim N\left(\mathbf{0}, \begin{pmatrix} V(\boldsymbol{\mu}) & V_{12} \\ V_{21} & DX(X'WX)^{-1}X'D \end{pmatrix}\right).$$

Then from the expansion $\mathbf{y} - \hat{\boldsymbol{\mu}} = (\mathbf{y} - \boldsymbol{\mu}) - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})$ and $\mathbf{B} = -\mathbf{I}$, $(\mathbf{y} - \hat{\boldsymbol{\mu}})$ is asymptotically distribution as a normal random variable with variance

$$V(\boldsymbol{\mu}) - DX(X'WX)^{-1}X'D.$$

We can re-write this in more standard residual form as

$$\begin{aligned}Var(\mathbf{y} - \hat{\boldsymbol{\mu}}) &= DW^{-1}D - DW^{-1/2}W^{1/2}X(X'WX)^{-1}X'W^{1/2}W^{-1/2}D \\ &= DW^{-1/2}[I - W^{1/2}X(X'WX)^{-1}X'W^{1/2}]W^{-1/2}D \\ &= V(\boldsymbol{\mu})^{1/2}[I - H]V(\boldsymbol{\mu})^{1/2}\end{aligned}$$

where $H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$. These look similar to residuals for linear models.

To construct a standardized Pearson residual, we do

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{Var}(y_i - \hat{\mu}_i)}} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)(1 - \hat{h}_i)}} = \frac{e_i}{\sqrt{1 - \hat{h}_i}}.$$

For standardized deviance residual,

$$\sqrt{d_i^*} \text{sign}(y_i - \hat{\mu}_i) / \sqrt{1 - \hat{h}_i}.$$

When the model holds, $r_i \rightarrow N(0, 1)$ (asymptotics for which y_i is approximately normal).

Example

Poisson

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} \frac{1}{\sqrt{1 - \hat{h}_i}}.$$

Approximately normality holds as $\mu_i \rightarrow \infty$ NOT as $n \rightarrow \infty$ for fixed μ_i .

For HW, show that

$$\hat{\boldsymbol{\eta}} = g(\hat{\boldsymbol{\mu}}) = \hat{H} \times \{\text{linear approx for } g(\mathbf{y})\}$$

Remarks:

- Pierce and Shafer (1986, JASA) claim deviance residuals tends to be closer to normality than standardized Pearson since has skewness closer to zero.
- for small values of ϕ/w_i ,

$$D(y_i; \mu_i) \approx \frac{w_i(y_i - \mu_i)^2}{V(\mu_i)} \sim N(0, 1)$$

So, the residuals should be approximately $N(0, 1)$ and absolute values bigger than 2 or 3 suggest outliers. Care when ϕ/w_i not small.

- as in linear models, residuals are NOT independent
- in standardizing have not taken into account the fact that the weights in the W matrix are random as well.

0.4 GLMs for binary data

Ch 6 in book

(y_1, \dots, y_n) independent observations where
 $n_i Y_i \sim \text{Bin}(n_i, \pi_i)$.

Classify into

1. *Grouped data*: 'group' binary data into covariate classes (categorical covariates)
 asymptotics usually $n_i \rightarrow \infty$, for n fixed
 often summarized with a contingency table

Consider a single binary covariate, X

		Y	
		0	1
x	0		n ₁
	1		n ₂

2. *Ungrouped data*: $n_1 = n_2 = \dots = n_n = 1$. Bernoulli observations
 asymptotics refer to $n \rightarrow \infty$; typically see with continuous covariates

Details on the Binomial distribution

Derivation

1. sum of independent, homogeneous (same π) Bernoullis
2. Y_1, Y_2 independent Poisson μ_1, μ_2 , then

$$Y_1 | Y_1 + Y_2 = m \sim \text{Bin}$$

$$\begin{aligned}
P(Y_1 = y | Y_1 + Y_2 = m) &= \frac{P(Y_1 = y, Y_2 = m - y)}{P(Y_1 + Y_2 = m)} \\
&= \frac{P(Y_1 = y)P(Y_2 = m - y)}{P(Y_1 + Y_2 = m)} \\
&= \frac{(\exp(-\mu_1)\mu_1^y/y!)(\exp(-\mu_2)\mu_2^{m-y}/(m-y)!)}{\exp[-(\mu_1 + \mu_2)(\mu_1 + \mu_2)^m/m!]} \\
&= \binom{m-y}{y} \pi^y (1-\pi)^{m-y}
\end{aligned}$$

where $\pi = \frac{\mu_1}{\mu_1 + \mu_2}$.

Moments and Cumulants

moment generating function (mgf) for Bernoulli, $Y_i \sim Ber$

$$\begin{aligned}
M_Y(t) &= E[\exp(tY_i)] \\
&= \pi \exp(t * 1) + (1 - \pi) \exp(t * 0) \\
&= 1 - \pi(1 - \exp(t))
\end{aligned}$$

cumulant generating function (cgf) for Bernoulli, $Y_i \sim Ber$

$$K_Y(t) = \log[1 - \pi(1 - \exp(t))]$$

So, for $Y^* \sim Bin(m, \pi)$,

$$M_{Y^*}(t) = [1 - \pi(1 - \exp(t))]^m$$

$$K_{Y^*}(t) = m \log[1 - \pi(1 - \exp(t))]$$

and $\kappa_1 = m\pi$ and $\kappa_2 = m\pi(1 - \pi)$.

$$\begin{aligned}
\left. \frac{\partial K_{Y^*}(t)}{\partial t} \right|_{t=0} &= \left. \frac{m\pi \exp(t)}{1 - \pi(1 - \exp(t))} \right|_{t=0} \\
&= m\pi
\end{aligned}$$

$$\begin{aligned}
\left. \frac{\partial^2 K_{Y^*}(t)}{\partial t^2} \right|_{t=0} &= \left. \frac{(1 - \pi(1 - \exp(t)))m\pi \exp(t) - m\pi \exp(t)\pi \exp(t)}{(1 - \pi(1 - \exp(t)))^2} \right|_{t=0} \\
&= m\pi - m\pi^2 = m\pi(1 - \pi)
\end{aligned}$$

Link functions

recall $\eta = \sum_{j=0}^p x_j \beta_j$

Choose link functions

$g(\pi)$ which map $(0, 1) \rightarrow (-\infty, \infty)$ [i.e., unit interval to the real line]; so set $g = F^{-1}$ where F is a cdf

Common choices:

1. logit link (F from logistic distribution)

$$g(\pi) = \log \frac{\pi}{1-\pi} \text{ which implies } \pi = \frac{\exp \eta}{1+\exp \eta}$$

2. probit link (F from normal distribution)

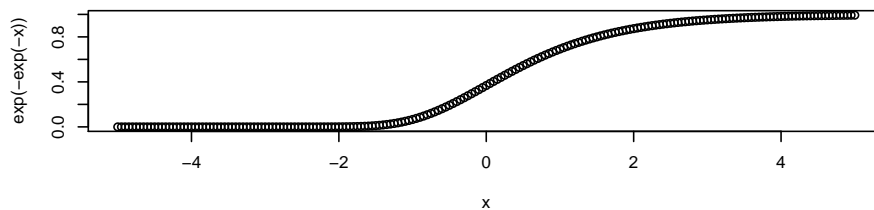
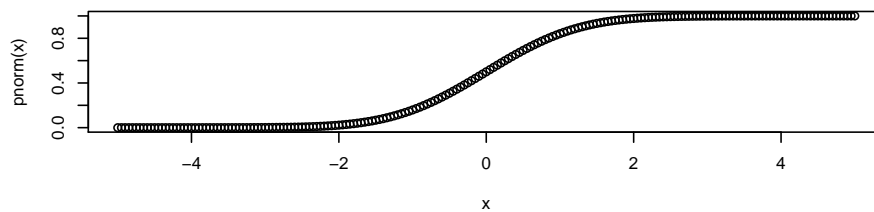
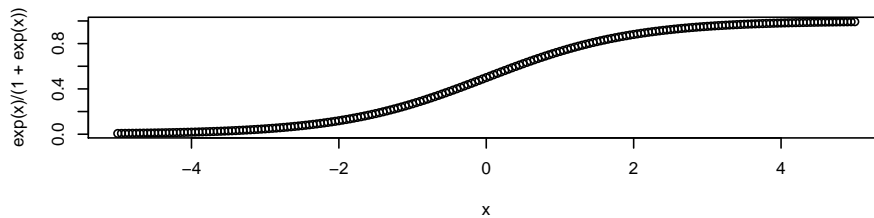
$$g(\pi) = \Phi^{-1}(\pi) \text{ which implies } \pi = \Phi(\eta)$$

3. complementary log-log link (F from extreme value distribution)

$$g(\pi) = \log[-\log(1-\pi)] \text{ which implies } \pi = 1 - \exp(-\exp(\eta))$$

4. log-log link (F from extreme value distribution)

$$g(\pi) = -\log[-\log(\pi)] \text{ which implies } \pi = \exp(-\exp(-\eta))$$



Some properties:

- logit and probit are symmetric: $g(\pi) = -g(1 - \pi)$
- continuous and increasing on $[0, 1]$

Remarks:

- logit and probit are almost linearly related over interval $[\cdot, \cdot]$; difficult to discriminate based on GOF
- complementary log-log: as $\pi \rightarrow 1$, approaches infinity much more slowly than logistic or probit; as $\pi \rightarrow 0$, more quickly

Details on logistic link

$$\log \frac{\pi}{1-\pi} = \sum_{j=0}^2 x_j \beta_j = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\text{implies } \pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}$$

Interpretation: β_1 is the log odds ratio for a unit change in X_1 holding X_2 constant; so coefficients are log odds ratios

Show this in class.

Retrospective sampling with the logit link

Very relevant for analysis of epidemiological data. In particular, case-control studies; idea is to find 'cases' of the disease (D) and look back in time for their exposure (X); then find corresponding controls

So, we sample $[X|D]$ but we want to make inference about $[D|X]$. Ok to do this for logit link but not for probit or log-log links. Details follow:

Suppose we are interested in estimating the parameters in the following regression model,

$$F^{-1}(P[D = 1 | x]) = \alpha + \beta^T x,$$

where F^{-1} is the inverse cdf of a standard normal (probit link) or the inverse cdf of a standard logistic distribution (logistic link) and D is the binary response of interest. However, suppose that instead of sampling D for given values of x , we sample x for given values of D . Let Z be a binary variable determining whether an individual is sampled or not. In particular, define

$$\pi_0 = P(Z = 1 | D = 1) \tag{1}$$

$$\pi_1 = P(Z = 1 | D = 0). \tag{2}$$

So the probability of being sampled conditional on D is independent of x (*). In this setup, we can estimate the regression

$$F^{-1}(P[D = 1 | Z = 1, x]) = \alpha^* + \beta^* x.$$

For the logistic and probit link determine whether $\beta = \beta^*$ and $\alpha = \alpha^*$? [Hint: use Bayes theorem]. Also, suppose π_0 and π_1 are functions of x . What happens in this case?

$$\begin{aligned}
P(D = 1|Z = 1, x) &= \frac{P(Z = 1|D = 1, x)P(D = 1|x)}{\sum_{d=0}^1 P(Z = 1|D = d, x)P(D = d|x)} \\
&\text{by (*) above} \\
&= \frac{P(Z = 1|D = 1)P(D = 1|x)}{\sum_{d=0}^1 P(Z = 1|D = d)P(D = d|x)} \\
&= \frac{\pi_0 F(\alpha + \beta'x)}{\pi_0 F(\alpha + \beta'x) + \pi_1 [1 - F(\alpha + \beta'x)]}
\end{aligned}$$

For logistic link, $F(x) = \exp(x)/(1 + \exp(x))$, so

$$\begin{aligned}
P(D = 1|Z = 1, x) &= \frac{\pi_0 \exp(\alpha + \beta x)/(1 + \exp(\alpha + \beta x))}{\pi_0 \exp(\alpha + \beta x)/(1 + \exp(\alpha + \beta x)) + \pi_1 1/(1 + \exp(\alpha + \beta x))} \\
&= \frac{\pi_0 \exp(\alpha + \beta x)}{\pi_0 \exp(\alpha + \beta x) + \pi_1} \\
&= \frac{\pi_0 \exp(\alpha + \beta x)}{\pi_0 \exp(\alpha + \beta x) + \pi_1} \\
&= \frac{1}{1 + (\pi_1/\pi_0) \exp(-\alpha - \beta x)} \\
&= \frac{\exp(\alpha^* + \beta x)}{1 + \exp((\alpha^* + \beta x))}
\end{aligned}$$

where $\alpha^* = \log \frac{\pi_1}{\pi_0} + \alpha$. So, the intercept changed, but the log odds ratio is the same. For the probit link, similar cancellation does not occur.

Note: if the selection probabilities, π_0 and π_1 are functions of x , then lose interpretation of log odds ratio.

Details on probit link

Motivation based on latent variable with a threshold.

Suppose an underlying normal response, Z and we observe

$$y_i = \begin{cases} 0 & \text{if } z_i \leq \tau \\ 1 & \text{if } z_i > \tau \end{cases}$$

WLOG, set $\tau = 0$.

Now, suppose $z_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where $\epsilon_i \sim N(0, 1)$. Note we set the variance to 1 as it only impacts the scale of β (and is not identified).

$$\begin{aligned}
P(Y_i = 1) &= P(Z_i > 0) \\
&= P(\epsilon_i > -(\beta_0 + \beta_1 x_i)) \\
&= 1 - \Phi[-(\beta_0 + \beta_1 x_i)] \\
&= \Phi(\beta_0 + \beta_1 x_i)
\end{aligned}$$

the probit link!

β_1 is equal to the change in $E[Z]$ for a one unit increase in x ; i.e., the expected number of standard deviation change in Z for a unit increase in x .

The idea of a (normal) tolerance distribution has been used frequently in toxicology. In principle, tolerance distribution can follow any parametric form.

Consider the following: $(X_i, Y_i) = (\text{dosage}, \text{response})$ for subject i ,

$$y_i = \begin{cases} 0 & \text{if survive} \\ 1 & \text{if die} \end{cases}$$

Assume $T \sim N(\mu, \sigma^2)$. Note $Y_i = 1$ provides the same information as $T_i \leq X_i$. Now,

$$\begin{aligned} \pi_i &= P(Y_i = 1) \\ &= P(T_i \leq X_i) \\ &= \Phi\left(\frac{X_i - \mu}{\sigma}\right) \\ &= \Phi(\beta_0 + \beta_1 X_i). \end{aligned}$$

So, $\beta_0 = -\frac{\mu}{\sigma}$ and $\beta_1 = \frac{1}{\sigma}$.

Maximum likelihood

assume $Y_i \sim \text{Bin}(m_i, \pi_i)$, $i = 1, \dots, n$ and $g(\pi_i) = \sum x_{ij} \beta_j = \eta_i$.

The log likelihood is given by

$$\begin{aligned} L(\boldsymbol{\beta}; \mathbf{y}) &= \sum_i [y_i \theta_i - m_i \log(1 + \exp(\theta_i)) + c(y_i; m_i)] \\ &= \sum_i \left[y_i \log \frac{\pi_i(\boldsymbol{\beta})}{1 - \pi_i(\boldsymbol{\beta})} + m_i \log(1 - \pi_i) + c(y_i; m_i) \right]. \end{aligned}$$

Recall, $\pi_i(\boldsymbol{\beta}) = g^{-1}(\sum_j x_{ij} \beta_j)$.

Note:

$$\begin{aligned} \frac{\partial l_i}{\partial \beta_j} &= \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \pi_i} \frac{\partial \pi_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\ \frac{\partial l_i}{\partial \theta_i} &= y_i - m_i \frac{\exp(\theta_i)}{1 + \exp(\theta_i)} = y_i - m_i \pi_i \\ \frac{\partial \theta_i}{\partial \pi_i} &= \frac{1}{m_i \pi_i (1 - \pi_i)} \text{ recall, } a(\phi) / \text{Var}(Y_i) \\ \frac{\partial \pi_i}{\partial \eta_i} &= \frac{\partial g^{-1}(\eta_i)}{\partial \eta_i} \text{ link specific} \\ \frac{\partial \eta_i}{\partial \beta_j} &= x_{ij} \end{aligned}$$

So the score is

$$U(\boldsymbol{\beta}; \mathbf{y}) = \frac{\partial L(\boldsymbol{\beta}; \mathbf{y})}{\partial \beta_j} = \sum_i \frac{y_i - m_i \pi_i}{m_i \pi_i (1 - \pi_i)} \frac{\partial \pi_i}{\partial \eta_i} x_{ij} = 0 : j = 1, \dots, p.$$

And the Fisher information is

$$\begin{aligned} E \left[-\frac{\partial^2 L}{\partial \beta_j \partial \beta_k} \right] &= \frac{1}{m_i \pi_i (1 - \pi_i)} \left(\frac{\partial \pi_i}{\partial \eta_i} \right)^2 x_{ij} x_{ik} \\ &= \{X'WX\}_{jk} \end{aligned}$$

where $W = \text{diag} \left\{ \left(\frac{\partial \pi_i}{\partial \eta_i} \right)^2 / (m_i \pi_i (1 - \pi_i)) \right\}$.

Remarks:

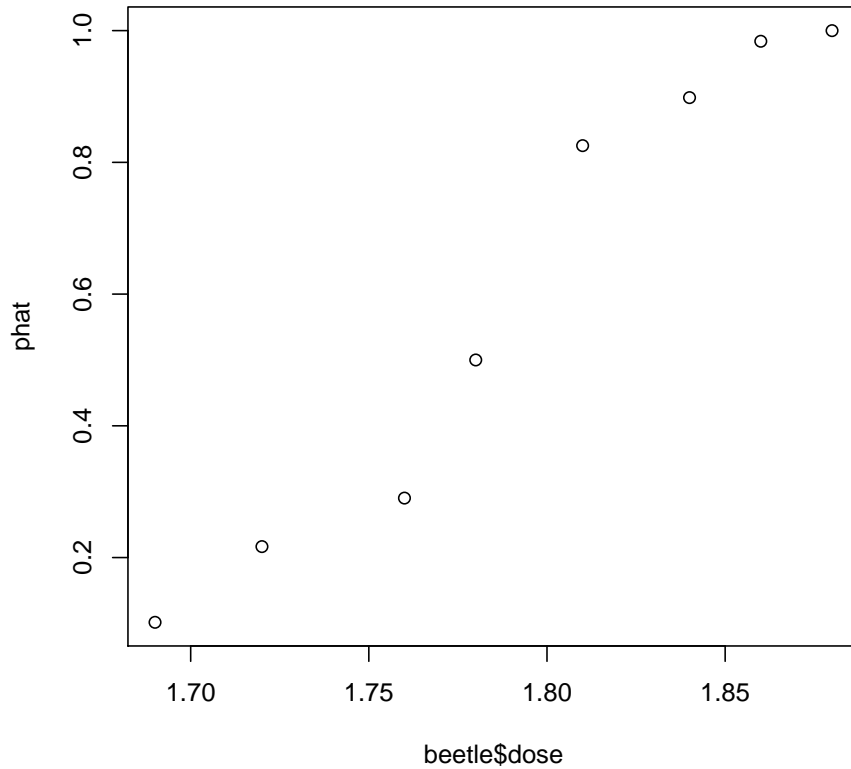
- for the logit link,
 $\frac{\partial \pi_i}{\partial \eta_i} = \pi_i(1 - \pi_i)$, so
 $\frac{\partial L}{\partial \beta_j} = \sum_i (y_i - m_i \pi_i) X_{ij}$, so
 and $W = \text{diag}\{\pi_i(1 - \pi_i)/m_i\}$.
- asymptotic distribution of $\hat{\beta}$: $\text{AsymVar}(\hat{\beta}) = (X'WX)^{-1}$
 note the behavior: as $m_i \rightarrow \infty$, the variance decreases; as $\pi \rightarrow 0$ (or 1), increases.
 However, the asymptotics we will consider here are as $n \rightarrow \infty$

Example: Beetle Mortality

From Ch 7 in book. Experiment to examine dose-response of gaseous carbon disulphide (from Bliss, 1935). For each of 8 doses, approximately 60 beetles were exposed. The number who died was recorded. Of interest to estimate the dose response relationship.

```
> beetle <- read.table("/home/mdaniels/classes/glm/spring2009.dir/bookdata.dir/GLM_data/bee
+   header = T)
> mortality <- beetle$y
> n <- beetle$n
> phat <- beetle$y/beetle$n

> plot(beetle$dose, phat)
```



```
> plot(beetle$dose, log(phat/(1 - phat)))
> beetle.reg <- glm(mortality/n ~ beetle$dose, family = binomial(link = "logit"),
+   weights = n)
> summary(beetle.reg)
```

Call:

```
glm(formula = mortality/n ~ beetle$dose, family = binomial(link = "logit"),
    weights = n)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.8986	-0.5475	0.9842	1.3315	1.7179

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-60.103	5.164	-11.64	<2e-16 ***
beetle\$dose	33.934	2.903	11.69	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.202 on 7 degrees of freedom
Residual deviance: 13.633 on 6 degrees of freedom
AIC: 43.831

Number of Fisher Scoring iterations: 4

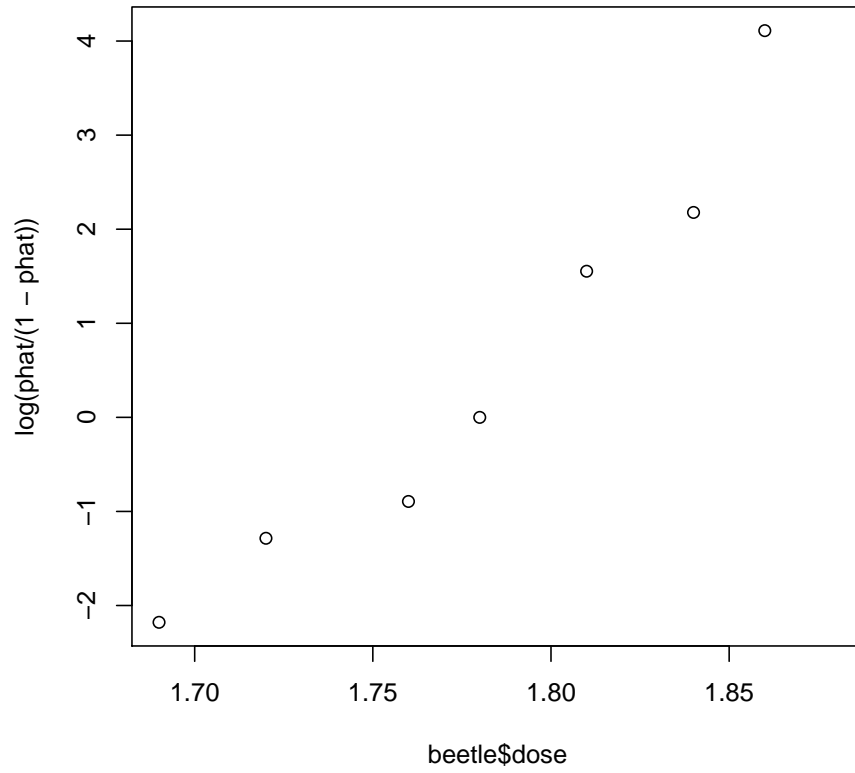
```
> coef(beetle.reg)
```

```
(Intercept) beetle$dose  
-60.10328    33.93416
```

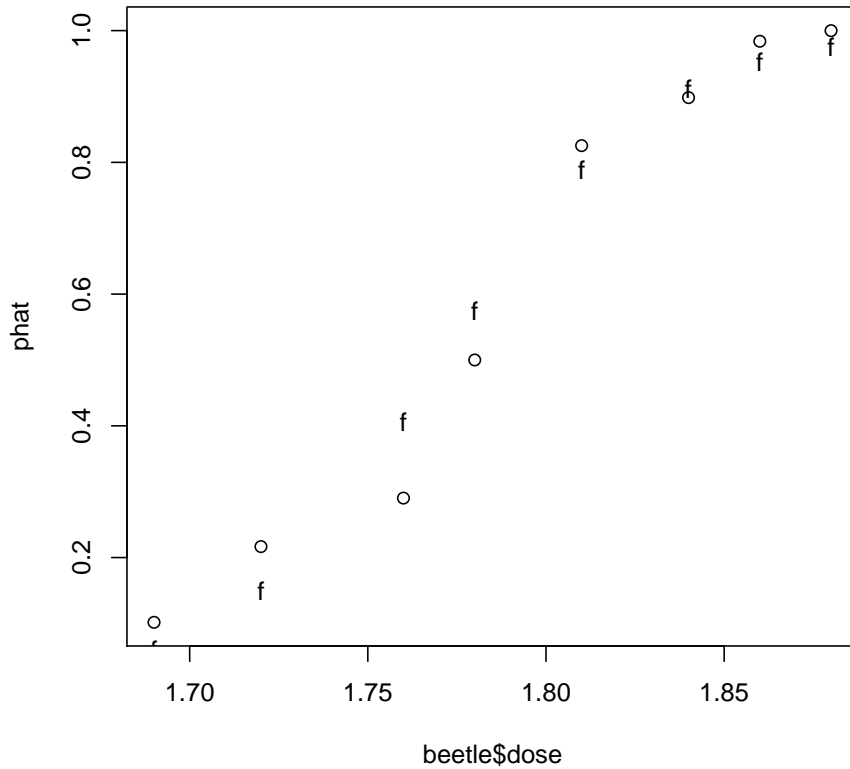
```
> vcov(beetle.reg)
```

```
(Intercept) beetle$dose  
(Intercept) 26.66859 -14.986119  
beetle$dose -14.98612  8.426637
```

```
> fit.p <- c(fitted.values(beetle.reg))  
> fit.y <- n * fit.p
```



```
> plot(beetle$dose, phat)
> points(beetle$dose, fit.p, pch = "f")
```



Let's look at residuals. The `residuals()` function gives the non-standardized deviance and Pearson residuals. The `rstandard()` function gives the standardized residuals

```
> residuals(beetle.reg)
```

```

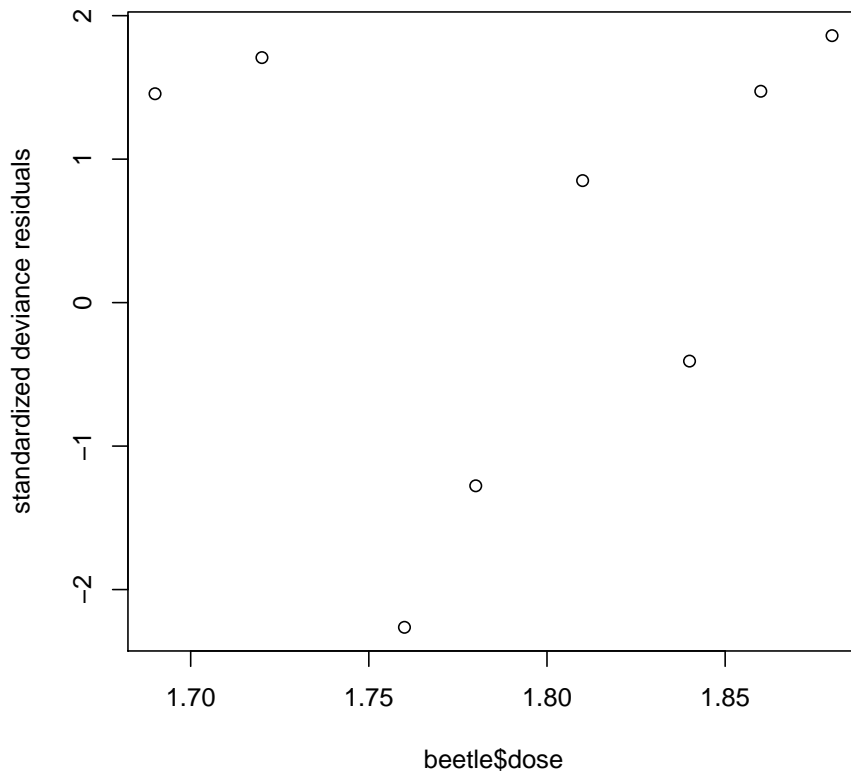
      1      2      3      4      5      6      7
1.2398294 1.3778449 -1.8986166 -1.1186690  0.7286035 -0.3571006  1.3160467
      8
1.7179121
```

```
> residuals(beetle.reg, type = "pearson")
```

```

      1      2      3      4      5      6      7
1.3558254 1.4524694 -1.8598462 -1.1249478  0.7121278 -0.3649854  1.1430541
      8
1.2222545
```

```
> plot(beetle$dose, rstandard(beetle.reg), ylab = "standardized deviance residuals")
```



```
> beetle.log.reg <- glm(mortality/n ~ beetle$dose, family = binomial(link = "cloglog"),
+   weights = n)
> summary(beetle.log.reg)
```

Call:

```
glm(formula = mortality/n ~ beetle$dose, family = binomial(link = "cloglog"),
     weights = n)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.37517	-0.36801	0.07958	0.54314	1.46367

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-38.889	3.127	-12.44	<2e-16 ***
beetle\$dose	21.664	1.736	12.48	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

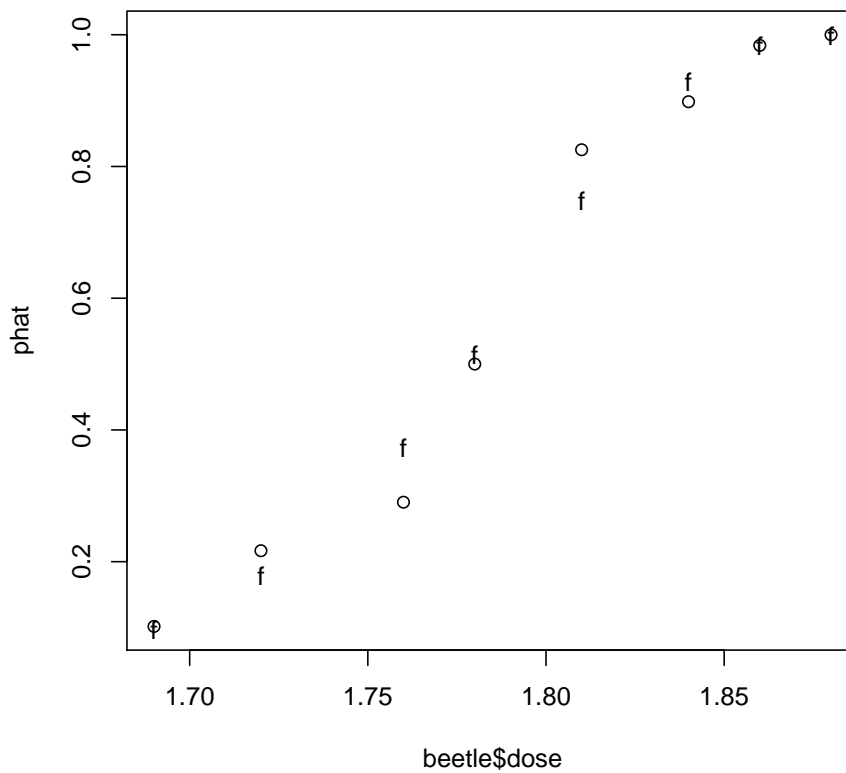
(Dispersion parameter for binomial family taken to be 1)

Null deviance: 284.2024 on 7 degrees of freedom
 Residual deviance: 5.6281 on 6 degrees of freedom
 AIC: 35.826

Number of Fisher Scoring iterations: 4

```
> fit.p <- c(fitted.values(beetle.log.reg))
> fit.y <- beetle$n * fit.p

> plot(beetle$dose, phat)
> points(beetle$dose, fit.p, pch = "f")
```



Let's look at residuals

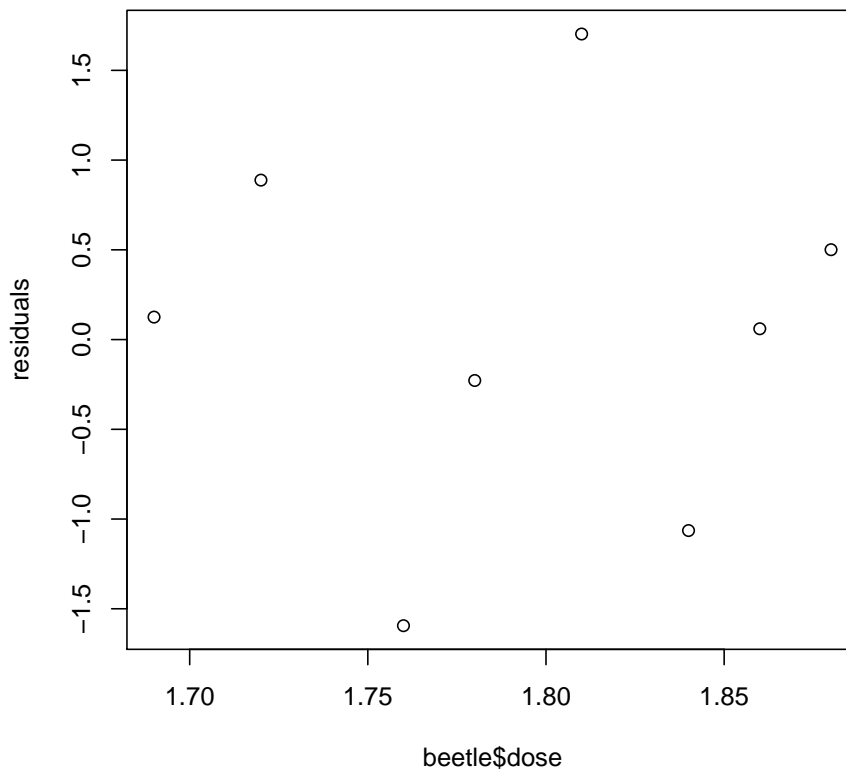
```
> residuals(beetle.log.reg)

      1      2      3      4      5      6      7
0.1086719 0.7556860 -1.3751704 -0.2029708 1.4636703 -0.8631398 0.0504797
```

```

      8
0.4722976
> residuals(beetle.log.reg, type = "pearson")
      1      2      3      4      5      6
0.10936338 0.77548323 -1.35045690 -0.20300811  1.40681946 -0.91471219
      7      8
0.05007421 0.33412012
> plot(beetle$dose, rstandard(beetle.log.reg), ylab = "standardized deviance\nresiduals")

```



Deviance

$$\begin{aligned}
 D(\mathbf{y}, \hat{\boldsymbol{\pi}}) &= 2L(\tilde{\boldsymbol{\pi}}; \mathbf{y}) - L(\hat{\boldsymbol{\pi}}; \mathbf{y}) \\
 &= 2 \sum_i \left\{ y_i \log \frac{y_i}{\hat{\mu}_i} + (m_i - y_i) \log \frac{m_i - y_i}{m_i - \hat{\mu}_i} \right\}
 \end{aligned}$$

As discussed earlier, we can use this as a goodness of fit (GOF) statistic for testing the adequacy of a given model. Asymptotically (details below) will approach a χ_{n-p}^2 . It is assumed that the Y_i are independent and the asymptotics are:

- n fixed with $m_i \rightarrow \infty$ for each i AND $m_i \pi_i (1 - \pi_i) \rightarrow \infty$.
- so clearly, the χ^2 approximation will *not* work for continuous covariates

But, for comparing nested models,

$D(\mathbf{y}; \boldsymbol{\mu}_0) - D(\mathbf{y}; \boldsymbol{\mu}_A)$ - just the likelihood ratio test statistic; asymptotically, $p_A - p_0$ under either the GOF asymptotic assumption above OR $n \rightarrow \infty$.

Exploration of asymptotics for Deviance GOF statistic

Data (grouped)

$$\begin{array}{c|c|c|c|c} y_1 & y_2 & y_3 & \cdots & y_{n/2} \\ \hline m_1 - y_1 & m_2 - y_2 & m_3 - y_3 & \cdots & m_{n/2} - y_{n/2} \end{array}$$

where $\sum_{i=1}^{n/2} m_i = N$. Assume $y_1, \dots, y_{n/2}$ are independent $\text{Bin}(m_i, \pi^*)$. In modified notation (counts in a $2 \times n/2$ contingency table),

$$\begin{array}{c|c|c|c|c} n_1 & n_2 & n_3 & \cdots & n_{n/2} \\ \hline n_{n/2+1} & n_{n/2+2} & n_{n/2+3} & \cdots & n_n \end{array}$$

and $\sum_{i=1}^n n_i = N$. Define $\hat{\mu}_i = m_i \hat{\pi}_i^* = N \hat{\pi}_i$, where $\sum_{i=1}^n \hat{\pi}_i = 1$. Also, let $p_i = n_i/N$. So,

$$\begin{aligned} D(\mathbf{n}, \boldsymbol{\pi}) &= 2 \sum n_i \log \frac{n_i}{\hat{\mu}_i} \\ &= 2N \sum p_i \log \left(1 + \frac{p_i - \hat{\pi}_i}{\hat{\pi}_i} \right) \end{aligned}$$

Asymptotically, as $N \rightarrow \infty$, $\pi_i > 0$.

Do the expansion,

$\log(1+x) = x - x^2/2 + x^3/3 - \dots$, for $|x| < 1$. And note:

$x = \frac{p_i - \hat{\pi}_i}{\hat{\pi}_i} \rightarrow 0$ when the model holds, so $|x| < 1$ is reasonable.

Rewrite the Deviance as

$$\begin{aligned} D(\mathbf{n}, \boldsymbol{\pi}) &= 2N \sum_i [\hat{\pi}_i + (p_i - \hat{\pi}_i)] \left[\frac{p_i - \hat{\pi}_i}{\hat{\pi}_i} - \frac{1}{2} \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i^2} + \dots \right] \\ &= 2N \sum_i (p_i - \hat{\pi}_i) + \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} - \frac{1}{2} \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} + O_p((p_i - \hat{\pi}_i)^3) \end{aligned}$$

Note: $\sum p_i = 1$, $\sum \hat{\pi}_i = 1$ and $\sum_i (p_i - \hat{\pi}_i) = 0$. Also,

$$p_i - \hat{\pi}_i = (p_i - \pi_i) - (\hat{\pi}_i - \pi_i).$$

The first term is $O_p(N^{-1/2})$; the second, $O_p(N^{-1/2})$ also. So,

$$\begin{aligned} D(\mathbf{n}, \boldsymbol{\pi}) &= N \sum_i \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} + 2N O_p(N^{-3/2}) \\ &= X^2 + O_p(N^{-1/2}) \\ &= X^2 + o_p(1) \end{aligned}$$

X^2 is Pearson's χ^2 statistic.

Example

The goal of this study was to identify risk factors associated with giving birth to a low birth weight baby (weighing less than 2500 grams). Data were collected on 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies. Five variables which were thought to be of importance were age, smoking during pregnancy (yes/no), weight of the subject at her last menstrual period (LWT), race (3 categories), and the number of physician visits during the first trimester of pregnancy (FTV).

```
> lbw <- read.table("/home/mdaniels/classes/glm/data.dir/binomial.dir/lbw.dat",
+   header = T)
> lbw.fit <- glm(LOW ~ AGE, family = binomial, data = lbw)
> summary.glm(lbw.fit)
```

Call:

```
glm(formula = LOW ~ AGE, family = binomial, data = lbw)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.0402	-0.9018	-0.7754	1.4119	1.7800

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.38458	0.73212	0.525	0.599
AGE	-0.05115	0.03151	-1.623	0.105

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 234.67 on 188 degrees of freedom
 Residual deviance: 231.91 on 187 degrees of freedom
 AIC: 235.91

Number of Fisher Scoring iterations: 4

```
> lbw.fit <- glm(LOW ~ AGE + SMOKE, family = binomial, data = lbw)
> summary.glm(lbw.fit)
```

Call:

```
glm(formula = LOW ~ AGE + SMOKE, family = binomial, data = lbw)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.1589	-0.8668	-0.7470	1.2821	1.7925

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
--	----------	------------	---------	----------

```
(Intercept) 0.06091    0.75732    0.080    0.9359
AGE          -0.04978    0.03197   -1.557    0.1195
SMOKE        0.69185    0.32181    2.150    0.0316 *
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 234.67 on 188 degrees of freedom
Residual deviance: 227.28 on 186 degrees of freedom
AIC: 233.28
```

Number of Fisher Scoring iterations: 4

```
> lbw.fit <- glm(LOW ~ AGE + SMOKE + factor(RACE) + LWT + FTV,
+   family = binomial, data = lbw)
> summary.glm(lbw.fit)
```

Call:

```
glm(formula = LOW ~ AGE + SMOKE + factor(RACE) + LWT + FTV, family = binomial,
    data = lbw)
```

Deviance Residuals:

```
      Min       1Q   Median       3Q      Max
-1.5185  -0.9082  -0.5879   1.3093   2.0428
```

Coefficients:

```
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.328778   1.111145   0.296   0.7673
AGE          -0.022205   0.034659  -0.641   0.5217
SMOKE        1.053215   0.380989   2.764   0.0057 **
factor(RACE)2 1.230961   0.517297   2.380   0.0173 *
factor(RACE)3 0.941545   0.417994   2.253   0.0243 *
LWT          -0.012489   0.006437  -1.940   0.0524 .
FTV          -0.007697   0.164075  -0.047   0.9626
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 234.67 on 188 degrees of freedom
Residual deviance: 214.58 on 182 degrees of freedom
AIC: 228.58
```

Number of Fisher Scoring iterations: 4

We can't approx residual deviance with χ_{n-p}^2 since binary ($m_i = 1$). However, can compare models by looking at difference in deviances : just LRT.

```
> pear.resid <- residuals(lbw.fit, type = "pearson")
> dev.resid <- residuals(lbw.fit, type = "deviance")
> par(mfrow = c(2, 1))
> hist(pear.resid)
> hist(dev.resid)
> lbw.fit <- glm(LOW ~ AGE, family = binomial(link = logit), data = lbw)
> summary.glm(lbw.fit)
```

Call:

```
glm(formula = LOW ~ AGE, family = binomial(link = logit), data = lbw)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.0402	-0.9018	-0.7754	1.4119	1.7800

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.38458	0.73212	0.525	0.599
AGE	-0.05115	0.03151	-1.623	0.105

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 234.67 on 188 degrees of freedom
 Residual deviance: 231.91 on 187 degrees of freedom
 AIC: 235.91

Number of Fisher Scoring iterations: 4

```
> lbw.fit <- glm(LOW ~ AGE, family = binomial(link = probit), data = lbw)
> summary.glm(lbw.fit)
```

Call:

```
glm(formula = LOW ~ AGE, family = binomial(link = probit), data = lbw)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.0413	-0.9030	-0.7737	1.4102	1.7903

Coefficients:

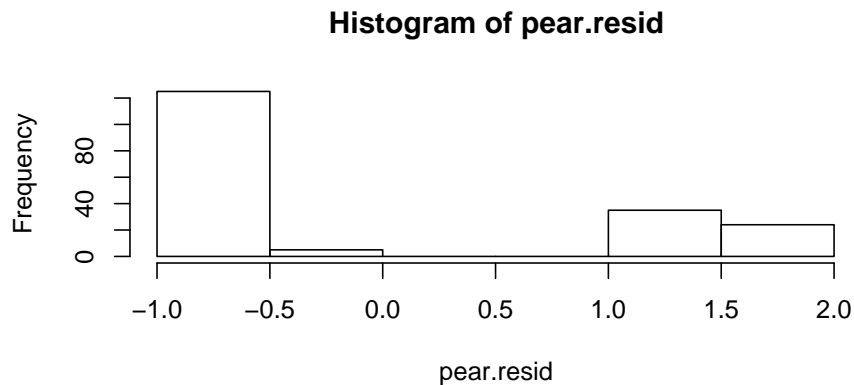
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	0.23590	0.43981	0.536	0.5917
AGE	-0.03155	0.01875	-1.682	0.0925 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 234.67 on 188 degrees of freedom
 Residual deviance: 231.85 on 187 degrees of freedom
 AIC: 235.85

Number of Fisher Scoring iterations: 4



0.4.1 Overdispersion

Clustering mechanism with binomials

$$\text{Var}(Y) > m\pi(1 - \pi).$$

This can arise from clustering (non-independence). Consider the following.

Assume n clusters of size k (so $m = nk$). Now assume in cluster i , $Z_i \sim \text{Bin}(k, \pi_i)$. So, $Y = \sum_{i=1}^n Z_i$.

Let

$$\begin{aligned} E[\pi_i] &= \pi \\ \text{Var}(\pi_i) &= \tau^2\pi(1 - \pi). \end{aligned}$$

In this situation, let's compute the marginal mean and variance of the 'overdispersed' binomial random variable Y ,

$$\begin{aligned} E[Y] &= EE[Y|\pi_i] \\ &= E\left[\sum_{i=1}^n \pi_i k\right] \\ &= m\pi \end{aligned}$$

$$\begin{aligned} \text{Var}[Y] &= \text{Var}E[Y|\pi_i] + E[\text{Var}(Y|\pi_i)] \\ &= \text{Var}\left[\sum \pi_i k\right] + E\left[\sum_i k\pi_i(1 - \pi_i)\right] \\ &= k^2 n \tau^2 \pi(1 - \pi) + kn\pi - kn[\tau^2\pi(1 - \pi) - \pi^2] \\ &= km\tau^2\pi(1 - \pi) + m\pi(1 - \pi) - m\tau^2\pi(1 - \pi) \\ &= [k\tau^2 + 1 - \tau^2]m\pi(1 - \pi) \\ &= (1 + (k - 1)\tau^2)m\pi(1 - \pi) \\ &= \sigma^2 m\pi(1 - \pi) \end{aligned}$$

Correlated binary data

In general, if $Y = \sum_{j=1}^m Z_j$ where Z_j are exchangeable binary random variables. Let $\pi = P(Z_j = 1)$. The correlation between Z_j and Z_k is

$$\begin{aligned} \rho &= \text{Corr}(Z_j, Z_k) \\ &= \frac{E(Z_j Z_k) - E(Z_j)E(Z_k)}{\sqrt{\text{Var}(Z_j)\text{Var}(Z_k)}} \\ &= \frac{P(Z_j = Z_k = 1) - \pi^2}{\pi(1 - \pi)}. \end{aligned}$$

This correlation is typically called the *intraclass correlation*. Note that this correlation is *not* restricted based on the means since the means are the same for all Z_j . For this case, the general form for the variance is

$$\begin{aligned} \text{Var}(Y) &= \sum \text{Var}(Z_j) + 2 \sum_{j < k} \text{Cov}(Z_j, Z_k) \\ &= m\pi(1 - \pi) + 2 \frac{m(m - 1)}{2} \pi(1 - \pi)\rho \\ &= m\pi(1 - \pi)[1 + (m - 1)\rho]. \end{aligned}$$

Mixture models for correlated binary data

Suppose conditional on p , Z_1, \dots, Z_n are independent Bernoulli random variables (so $Y \sim Bin(n, p)$). Assume p is random with mean π . The Z_j are correlated,

$$E(Z_j Z_k) = E[E[Z_j Z_k | p]] = E[E[Z_j | p] E[Z_k | p]] = E(p^2) \geq \pi^2$$

This implies that

$$0 \leq \rho = \frac{E(p^2) - \pi^2}{\pi(1 - \pi)} = \frac{Var(p)}{\pi(1 - \pi)} \leq 1.$$

So, $\rho \in (0, 1)$ if p is not degenerate at π and thus have overdispersion.

Beta-Binomial models

The most common model for correlated binary data is *Beta-Binomial model*. Here,

$$\begin{aligned} Z_i | p &\sim Ber(p) \\ p &\sim Beta(\alpha, \beta). \end{aligned}$$

Let $Y = \sum_{i=1}^m Z_i$. The marginal distribution for Y is a beta-binomial distribution, $Y \sim BetaBin(m, \pi, \rho)$. Its probability mass function is given by

$$\begin{aligned} P(Y = y) &= E[P(Y = y | p)] \\ &= \int_0^1 \binom{m}{y} p^y (1-p)^{m-y} \frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1} dp \\ &= \binom{m}{y} \frac{B(\alpha + y, \beta + m - y)}{B(\alpha, \beta)} \int_0^1 \frac{1}{B(\alpha + y, \beta + m - y)} p^{\alpha+y-1} (1-p)^{\beta+m-y-1} dp \\ &= \binom{m}{y} \frac{B(\alpha + y, \beta + m - y)}{B(\alpha, \beta)} \end{aligned}$$

where $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$ is the beta function.

Beta-Binomial Regression

When have overdispersed binomial data, can use Beta-Binomial model for regression.

Assume Y_1, \dots, Y_n are overdispersed binomial counts.

$$\begin{aligned} Y_i &\sim BetaBin(m_i, \pi_i, \rho) \\ g(\pi_i) &= x'_i \boldsymbol{\beta}, i = 1, \dots, n \end{aligned}$$

where β and ρ are unknown. When no covariates, easy to show that

$$\pi = \frac{\alpha}{\alpha + \beta} \text{ and } \rho = \frac{1}{\alpha + \beta + 1}.$$

Derive log likelihood for HW.

Example: AVSS Data

This data is from a toxicological study in which the response was the number of dead fetuses in litters of mice (Brooks et al, 1997). There were 127 litters, ranging in size from 2 to 20. The concern is that each of these 127 binomials are composed of non-independent Bernoulli fetuses. We examine that next.

```
> avss <- read.table("/home/mdaniels/classes/glm/spring2009.dir/avss.data",
+   header = T)
> quantile(avss$littersize)
```

```
 0%  25%  50%  75% 100%
2.0 12.0 14.0 15.5 20.0
```

```
> quantile(avss$ndead/avss$littersize)
```

```
      0%      25%      50%      75%      100%
0.00000000 0.00000000 0.05555556 0.09808612 0.50000000
```

```
> avss.bin.fit <- glm(ndead/littersize ~ 1, family = binomial,
+   data = avss, weights = littersize)
> summary(avss.bin.fit)
```

Call:

```
glm(formula = ndead/littersize ~ 1, family = binomial, data = avss,
    weights = littersize)
```

Deviance Residuals:

```
      Min       1Q   Median       3Q      Max
-1.6766  -1.4028  -0.2140   0.4464   4.0609
```

Coefficients:

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.61995     0.09576  -27.36  <2e-16 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 213.51 on 126 degrees of freedom
Residual deviance: 213.51 on 126 degrees of freedom
AIC: 360.34
```

Number of Fisher Scoring iterations: 5

```
> library(VGAM)
> avss.bin.vgam.fit <- vglm(cbind(ndead, littersize - ndead) ~
+   1, family = binomialff, data = avss)
> summary(avss.bin.vgam.fit)
```

Call:

```
vglm(formula = cbind(ndead, littersize - ndead) ~ 1, family = binomialff,
    data = avss)
```

Pearson Residuals:

	Min	1Q	Median	3Q	Max
logit(mu)	-1.2067	-1.0096	-0.20765	0.476	5.9518

Coefficients:

	Value	Std. Error	t value
(Intercept)	-2.6200	0.095756	-27.361

Number of linear predictors: 1

Name of linear predictor: logit(mu)

(Default) Dispersion Parameter for binomialff family: 1

Residual Deviance: 213.5086 on 126 degrees of freedom

Log-likelihood: -427.6936 on 126 degrees of freedom

Number of Iterations: 5

```
> avss.betabin.fit <- vglm(cbind(ndeath, littersize - ndeath) ~ 1,
+   family = betabinomial, data = avss)
> summary(avss.betabin.fit)
```

Call:

```
vglm(formula = cbind(ndeath, littersize - ndeath) ~ 1, family = betabinomial,
     data = avss)
```

Pearson Residuals:

	Min	1Q	Median	3Q	Max
logit(mu)	-1.0258	-0.89816	0.075598	0.57684	3.3171
logit(rho)	-1.1221	-0.90483	0.384230	0.49801	6.7095

Coefficients:

	Value	Std. Error	t value
(Intercept):1	-2.5997	0.12617	-20.6045
(Intercept):2	-2.7689	0.34968	-7.9184

Number of linear predictors: 2

Names of linear predictors: logit(mu), logit(rho)

Dispersion Parameter for betabinomial family: 1

Log-likelihood: -417.4525 on 252 degrees of freedom

Number of Iterations: 6

Remark: Hauck-Donner Phenomenon

in a 1977 paper by Hauck and Donner, they showed

...for testing hypothesis regarding a single parameter in a binomial logit model, that Wald's test has the following undesirable features: (1) for any sample size, Wald's test statistic decreases to zero as the distance between the parameter estimate and the null value increases; (2) the power of Wald's test, based on its large sample distribution, decreases to the significance level for alternatives far from the null value.

As a result, prefer LRTs to Wald tests in logit models.

Quasi-likelihood

Recall we can estimate overdispersion using

$$\hat{\phi} = \frac{1}{n - (p + 1)} \sum \frac{w_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}.$$

And in principle, it would make sense to inflate the variance of $\hat{\beta}$ by

$$\text{Var}(\hat{\beta}) = \hat{\phi}(X'WX)^{-1}.$$

But how can we justify this? Can do it via estimating equations (M-estimation). Details follow.

We will keep our mean model, $g(\mu_i) = X_i\beta$ but replace the assumption about the distribution of y_i (i.e., the random component) with an assumption only about the variance. Specifically, we assume Y_1, \dots, Y_n are independent responses with means and variances given by

$$g(\mu_i) = x_i'\beta \tag{3}$$

$$\text{Var}(Y_i) = \frac{\phi}{w_i} V(\mu_i) \tag{4}$$

for some known function $V(\mu_i)$, known weights w_1, \dots, w_n , and unknown dispersion parameter ϕ .

Under these conditions, it can be shown that the optimal (to be defined later) estimating equations for β are given by

$$\sum_{i=1}^n \frac{w_i (y_i - \mu_i)}{V(\mu_i) g'(\mu_i)} x_i = 0.$$

If (3) holds but (4) does not (in that case, might call $V(\mu)$ the *working variance function*), the estimating equations are no longer optimal, but they are still *unbiased*; here *unbiased* refers to the LHS having expectation zero at the true value of β .

Now, define

$$\psi(y; \boldsymbol{\beta}, x, w) = \frac{w(y - \mu)}{V(\mu)g'(\mu)}x.$$

Then, at the true value of $\boldsymbol{\beta}$, the expectation is zero.

Now, observe,

$$E[\psi(y; \boldsymbol{\beta}, x, w)\psi(y; \boldsymbol{\beta}, x, w)'] = \frac{[Var(Y)]}{V(\mu)g'(\mu)/w]^2}xx'.$$

and

$$\begin{aligned} \psi'(y; \boldsymbol{\beta}, \mathbf{x}, w) &= \frac{\partial}{\partial \boldsymbol{\beta}'} \left[\frac{w(y - \mu)}{V(\mu)g'(\mu)} \mathbf{x} \right] \\ \frac{\partial}{\partial \mu} \left[\frac{w(y - \mu)}{V(\mu)g'(\mu)} \mathbf{x} \right] \frac{\partial \mu}{\partial \boldsymbol{\beta}'} &= \left\{ \frac{-w}{V(\mu)g'(\mu)} + w(y - \mu) \frac{\partial}{\partial \mu} \left[\frac{1}{V(\mu)g'(\mu)} \right] \right\} \mathbf{x} \frac{\partial \mu}{\partial \boldsymbol{\beta}'} \end{aligned}$$

From this, we can show that

$$\begin{aligned} -E[\psi'(y; \boldsymbol{\beta}, bx, w)] &= \frac{w}{V(\mu)g'(\mu)} \mathbf{x} \frac{\partial \mu}{\partial \boldsymbol{\beta}'} \\ &= \frac{w}{V(\mu)g'(\mu)^2} \mathbf{x} \mathbf{x}'. \end{aligned}$$

According to the theory of estimating equations, for large n , $\hat{\boldsymbol{\beta}}$ is approximately normally distributed with mean $\boldsymbol{\beta}$ and covariance matrix V_n/n where $V_n = A_n^{-1}B_nA_n'^{-1}$ (the *sandwich matrix*) with

$$\begin{aligned} A_n &= -\frac{1}{n}E[\psi'(y; \boldsymbol{\beta}, \mathbf{x}, w)] \\ &= -\frac{1}{n} \sum_{i=1}^n \frac{w_i}{V(\mu_i)g'(\mu_i)^2} \mathbf{x}_i \mathbf{x}_i' \\ &= \frac{1}{n} X' \Omega X. \end{aligned}$$

$$\begin{aligned} B_n &= \frac{1}{n} E[\psi(y; \boldsymbol{\beta}, \mathbf{x}, w)\psi(y; \boldsymbol{\beta}, \mathbf{x}, w)'] \\ &= \frac{1}{n} \sum_{i=1}^n \frac{Var(Y_i)}{[V(\mu_i)g'(\mu_i)/w_i]^2} \mathbf{x}_i \mathbf{x}_i' \\ &= \frac{1}{n} X' \Omega^* X. \end{aligned}$$

where

$$\Omega = \text{diag} \left\{ \frac{w_i}{V(\mu_i)g'(\mu_i)^2}, i = 1, \dots, n \right\}$$

and

$$\Omega^* = \text{diag} \left\{ \frac{\text{Var}(Y_i)}{[V(\mu_i)g'(\mu_i)/w_i]^2}, i = 1, \dots, n \right\}.$$

V_n can be consistently estimated by plugging in $\hat{\mu}_i$ (i.e., $\hat{\beta}$). Can construct Wald-type statistics in standard way.

Notice that if (4) holds, $\Omega^* = \phi\Omega$ and $V_n = \phi(X'\Omega X)^{-1}$ (same form as glm). In that case, estimate as before with moment estimator for ϕ .

Now we will explicitly discuss *quasi-likelihood*. Let

$$q(\mu; y, \phi) = \frac{y - \mu}{\phi V(\mu)}$$

and let

$$\begin{aligned} Q(\mu; y, \phi) &= \int_y^\mu q(t; y, \phi) dt \\ &= \int_y^\mu \frac{y - t}{\phi V(t)} dt, \end{aligned}$$

so that $Q'(\mu; y, \phi) = q(\mu; y, \phi)$. The function $Q(\mu; y, \phi)$ is called the *quasi-likelihood function* (really the quasi-loglikelihood) for a single observation and $q(\mu; y, \phi)$ is the *quasi-score*. For the complete data,

$$Q(\boldsymbol{\mu}; \mathbf{y}, \phi) = \sum_{i=1}^n Q(\mu_i; y_i, \phi/w_i).$$

The optimal estimating equations are

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = \sum_{i=1}^n q_i(\mu_i; y_i, \phi/w_i) \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = 0.$$

These are the likelihood equations if we pretend $Q(\boldsymbol{\mu}; \mathbf{y}, \phi)$ is the true log-likelihood. Define the *quasi-deviance* for a single observation as

$$D(y; \mu) = -2\phi Q(\mu; y, \phi) = 2 \int_\mu^y \frac{y - t}{V(t)} dt$$

and

$$D(\mathbf{y}; \boldsymbol{\mu}) = \sum_{i=1}^n w_i D(y_i; \mu_i).$$

For comparing nested models, twice quasi-likelihoods is χ^2 with difference in dimensions of $\boldsymbol{\beta}$'s as the degrees of freedom (McCullagh, 1983).

Estimating functions

Def'n of estimating function: a function $g(\mathbf{y}, \boldsymbol{\theta})$ of the data \mathbf{y} and the parameters $\boldsymbol{\theta}$ having zero mean for all parameter values

Obtain estimates as $g(\mathbf{y}; \hat{\boldsymbol{\theta}}) = \mathbf{0}$.

The score is an estimating function

Class of *linear estimating functions*

$$h(\mathbf{y}; \boldsymbol{\beta}) = H'(\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\beta}))$$

where H may be a function of $\boldsymbol{\beta}$ but not \mathbf{Y} . Linear in \mathbf{y} for each $\boldsymbol{\beta}$. It is *optimal* in the sense of having minimum variance of $\mathbf{a}'\boldsymbol{\beta}$ for all linear estimating functions.

Consider $\tilde{\boldsymbol{\beta}}$ the solution of $h(\mathbf{y}; \tilde{\boldsymbol{\beta}}) = \mathbf{0}$. Let $\hat{\boldsymbol{\beta}}$ be the solution of $h(\cdot)$ for $H' = D'V^{-1}$. Then can show

$$\text{var}(\mathbf{a}'\tilde{\boldsymbol{\beta}}) \geq \text{var}(\mathbf{a}'\hat{\boldsymbol{\beta}}).$$

Proof:

Expand $h(\mathbf{y}; \tilde{\boldsymbol{\beta}})$ in a Taylor series around the true parameter value $\boldsymbol{\beta}$.

$$h(\mathbf{y}; \tilde{\boldsymbol{\beta}}) = h(\mathbf{y}; \boldsymbol{\beta}) + \frac{\partial}{\partial \boldsymbol{\beta}} h(\mathbf{y}; \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \dots$$

$$0 = h(\mathbf{y}; \boldsymbol{\beta}) - H'D(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \dots$$

where $D = \partial\boldsymbol{\mu}/\partial\boldsymbol{\beta}$ So

$$(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx (H'D)^{-1}h(\mathbf{y}; \boldsymbol{\beta})$$

and

$$\text{Var}(\tilde{\boldsymbol{\beta}}) = (H'D)^{-1}H'VH(D'H)^{-1}$$

where $\text{Var}(h(\mathbf{y}; \boldsymbol{\beta})) = H'VH$ and $V = \text{Var}(Y) = \phi V(\mu)$.

And note that

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}) &= D'V^{-1}VV^{-1}D \\ &= D'V^{-1}D \end{aligned}$$

Now examine

$$\text{Var}(\tilde{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}}) = (H'D)^{-1}H'VH(D'H)^{-1} - (D'V^{-1}D)^{-1}.$$

Show that this is non negative definite for all H ; i.e.,

$$\text{Var}(\tilde{\boldsymbol{\beta}}) - \text{Var}(\hat{\boldsymbol{\beta}}) \geq 0,$$

corresponds to all the eigenvalues being greater than or equal to zero.

$$D'(V^{-1} - H(H'VH)^{-1}H')D \geq 0.$$

which is equivalent to

$$D'V^{-1/2}(I - V^{1/2}H(H'VH)^{-1}H'V^{1/2})V^{-1/2}D \geq 0.$$

true since this is the residual covariance matrix of $D'V^{-1}Y$ on $H'V^{1/2}$ (think of this as X)

Example: Seed germination data

data from a 2x2 factorial experiment from Crowder (1978). in each replication, a 'batch of tiny seeds is brushed onto a plate covered with a certain extract at a given dilution' and 'the numbers of germinated and ungerminated seeds are subsequently counted'. (Crowder, 1978). The experimental factors were the types of seeds (O. aegyptiaca 75 and 72) and the type of extract (bean root vs. cucumber root).

```
> seeds <- read.table("/home/mdaniels/classes/glm/spring2009.dir/data.dir/seeds.txt",
+   header = T)
> seeds
```

	seed	root	y	n
1	075	Bean	10	39
2	075	Bean	23	62
3	075	Bean	23	81
4	075	Bean	26	51
5	075	Bean	17	39
6	075	Cmbr	5	6
7	075	Cmbr	53	74
8	075	Cmbr	55	72
9	075	Cmbr	32	51
10	075	Cmbr	46	79
11	075	Cmbr	10	13
12	073	Bean	8	16
13	073	Bean	10	30
14	073	Bean	8	28
15	073	Bean	23	45
16	073	Bean	0	4
17	073	Cmbr	3	12
18	073	Cmbr	22	41
19	073	Cmbr	15	30
20	073	Cmbr	32	51
21	073	Cmbr	3	7

```
> seeds.binom <- glm(y/n ~ seed * root, family = binomial, data = seeds,
+   weights = n)
> summary(seeds.binom)
```

Call:

```
glm(formula = y/n ~ seed * root, family = binomial, data = seeds,
    weights = n)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-2.01617	-1.24398	0.05995	0.84695	2.12123

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.4122	0.1842	-2.238	0.0252 *
seed075	-0.1459	0.2232	-0.654	0.5132
rootCmbr	0.5401	0.2498	2.162	0.0306 *
seed075:rootCmbr	0.7781	0.3064	2.539	0.0111 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 98.719 on 20 degrees of freedom
 Residual deviance: 33.278 on 17 degrees of freedom
 AIC: 117.87

Number of Fisher Scoring iterations: 4

```
> seeds.qbinom <- glm(y/n ~ seed * root, family = quasibinomial,
+ weights = n, data = seeds)
> summary(seeds.qbinom)
```

Call:

```
glm(formula = y/n ~ seed * root, family = quasibinomial, data = seeds,
weights = n)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-2.01617	-1.24398	0.05995	0.84695	2.12123

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.4122	0.2513	-1.640	0.1193
seed075	-0.1459	0.3045	-0.479	0.6379
rootCmbr	0.5401	0.3409	1.584	0.1315
seed075:rootCmbr	0.7781	0.4181	1.861	0.0801 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 1.861832)

Null deviance: 98.719 on 20 degrees of freedom
 Residual deviance: 33.278 on 17 degrees of freedom

AIC: NA

Number of Fisher Scoring iterations: 4

0.5 Models for polytomous data

Two types:

1. nominal scale - unordered
2. ordinal scale - ordered

Ordinal Models

Assume K ordered categories

Can base model on the cumulative response probabilities,

$$\gamma_j = P(Y \leq j)$$

as opposed to $\pi_j = P(Y = j)$.

Begin with parallel regression models

$$\log \frac{\gamma_j(x)}{1 - \gamma_j(x)} = \theta_j - \beta'x \quad j = 1, \dots, K - 1, \gamma_K = 1$$

where $\gamma_j(x) = P(Y \leq j|x)$. This model is often called the *proportional odds model* or the *cumulative logit model*. The reason for the former name is the following:

The odds of event $\{Y \leq j\}$ is independent of j ,

$$\frac{\gamma_j(x_1)/(1 - \gamma_j(x_1))}{\gamma_j(x_2)/(1 - \gamma_j(x_2))} = \exp\{-\beta'(x_1 - x_2)\}.$$

θ must satisfy $\theta_1 \leq \theta_2 \leq \dots \leq \theta_{K-1}$.

Another common model using the log-log link on the cumulative probabilities, γ_j ,

$$\log[-\log(1 - \gamma_j(x))] = \theta_j - \beta'x \quad j = 1, \dots, K - 1.$$

This is often called the *proportional hazards model*. To see why,

$$\begin{aligned} -\log(1 - \gamma_j(x)) &= \exp(\theta_j - \beta'x) \\ 1 - \gamma_j(x) &= \exp\{-\exp(\theta_j - \beta'x)\} \end{aligned}$$

So,

$$\begin{aligned} (1 - \gamma_j(x_1))^{\exp(\beta'x_1)} &= \exp(-\exp(\theta_j)) \\ &\text{and} \\ (1 - \gamma_j(x_2))^{\exp(\beta'x_2)} &= \exp(-\exp(\theta_j)) \end{aligned}$$

So,

$$(1 - \gamma_j(x_1)) = (1 - \gamma_j(x_2))^{\exp\{\beta'(x_2 - x_1)\}},$$

and note that the survivor function is defined as 1-cdf.

These models can be motivated via a tolerance distribution (or underlying latent variable distribution)

Proportional odds (P.O.) model: $Z - \beta'x$ follows a standard logistic distribution

Proportional hazards (P.H.) model: $Z - \beta'x$ follows the extreme value distribution

Basic idea is if the unobserved variable in the interval, $\theta_{j-1} < Z \leq \theta_j$, then $Y = j$

Derive this for HW

There are also ordinal probit models where $Z - \beta'x$ follows a normal distribution

Examples

The data are the degree of pneumoconiosis in coal workers as a function of years of exposure (grouped into 8 categories).

```
> pn <- read.table("pneumo.txt", header = T)
> pn
```

	exposure.time	normal	mild	severe
1	5.8	98	0	0
2	15.0	51	2	1
3	21.5	34	6	3
4	27.5	35	5	8
5	33.5	32	10	9
6	39.5	23	7	8
7	46.0	12	6	10
8	51.5	4	2	5

```
> library(VGAM)
> pn.lin.polr <- vglm(formula = cbind(normal, mild, severe) ~ (exposure.time),
+   cumulative(parallel = TRUE), pn)
> summary(pn.lin.polr)
```

Call:

```
vglm(formula = cbind(normal, mild, severe) ~ (exposure.time),
     family = cumulative(parallel = TRUE), data = pn)
```

Pearson Residuals:

	Min	1Q	Median	3Q	Max
logit(P[Y<=1])	-1.7113	-0.52588	0.38765	0.65896	1.6519
logit(P[Y<=2])	-1.5734	0.06058	0.20508	0.42445	0.7370

Coefficients:

	Value	Std. Error	t value
(Intercept):1	3.955844	0.409646	9.6567

```
(Intercept):2  4.869049  0.443680 10.9742
exposure.time -0.095904  0.011940 -8.0324
```

Number of linear predictors: 2

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])

Dispersion Parameter for cumulative family: 1

Residual Deviance: 13.39733 on 13 degrees of freedom

Log-likelihood: -208.4594 on 13 degrees of freedom

Number of Iterations: 5

```
> pn.polr <- vglm(formula = cbind(normal, mild, severe) ~ log(exposure.time),
+   cumulative(parallel = TRUE), pn)
> summary(pn.polr)
```

Call:

```
vglm(formula = cbind(normal, mild, severe) ~ log(exposure.time),
     family = cumulative(parallel = TRUE), data = pn)
```

Pearson Residuals:

	Min	1Q	Median	3Q	Max
logit(P[Y<=1])	-1.2479	-0.071638	0.14410	0.30860	0.77144
logit(P[Y<=2])	-1.0444	-0.184145	0.30926	0.33528	0.50476

Coefficients:

	Value	Std. Error	t value
(Intercept):1	9.6761	1.3241	7.3078
(Intercept):2	10.5817	1.3454	7.8649
log(exposure.time)	-2.5968	0.3811	-6.8139

Number of linear predictors: 2

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])

Dispersion Parameter for cumulative family: 1

Residual Deviance: 5.02683 on 13 degrees of freedom

Log-likelihood: -204.2742 on 13 degrees of freedom

Number of Iterations: 4

```
> pn.npolr <- vglm(formula = cbind(normal, mild, severe) ~ log(exposure.time),
+   cumulative(parallel = FALSE), pn)
> summary(pn.npolr)
```

Call:

```
vglm(formula = cbind(normal, mild, severe) ~ log(exposure.time),
     family = cumulative(parallel = FALSE), data = pn)
```

Pearson Residuals:

	Min	1Q	Median	3Q	Max
logit(P[Y<=1])	-1.1501	-0.145714	0.12492	0.38245	0.72881
logit(P[Y<=2])	-1.1502	-0.050604	0.18858	0.28637	0.56586

Coefficients:

	Value	Std. Error	t value
(Intercept):1	9.5933	1.33084	7.2085
(Intercept):2	11.1048	1.89296	5.8664
log(exposure.time):1	-2.5713	0.38387	-6.6983
log(exposure.time):2	-2.7435	0.53225	-5.1546

Number of linear predictors: 2

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])

Dispersion Parameter for cumulative family: 1

Residual Deviance: 4.8844 on 12 degrees of freedom

Log-likelihood: -204.2029 on 12 degrees of freedom

Number of Iterations: 6

Nominal Models

Multinomial logit model

$$\begin{aligned} \log \frac{\pi_j(x)}{\pi_K(x)} &= \eta_j(x) \\ &= \beta'_j x \quad j = 1, \dots, K-1 \end{aligned}$$

Sometimes called baseline category logit model. above is equal to

$$\text{logit}[P(Y = j|Y = j \text{ or } Y = K)]$$

Note that other logits

$$\begin{aligned}
\log \frac{\pi_a(x)}{\pi_b(x)} &= \log \frac{\pi_a(x)}{\pi_K(x)} - \log \frac{\pi_b(x)}{\pi_K(x)} \\
&= \eta_a(x) - \eta_b(x) \\
&= (\boldsymbol{\beta}_a - \boldsymbol{\beta}_b)'x
\end{aligned}$$

This is a generalization of logistic regression. Based on multivariate logistic distribution.
Multinomial probit model

Tricky to express without latent variables, $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{i,K-1})$. Assume

$$\mathbf{Z}_i \sim N(\boldsymbol{\beta}'x_i, \Sigma),$$

and the elements of $\Sigma = \{\sigma_{lm}\}$. Set $\sigma_{11} = 1$ for identifiability (similar to binomial probit).
The multinomial response Y_i is

$$Y_i(\mathbf{Z}_i) = \begin{cases} 0 & \text{if } \max(\mathbf{Z}_i) < 0 \\ j & \text{if } \max(\mathbf{Z}_i) = Z_{ij} > 0 \end{cases}$$

These models are hard to fit; high dimensional integration. Fit using Bayesian MCMC (Imai and van Dyk, 2005; Journal of Econometrics)

Discrete choice models

model choice one of a discrete set of options

e.g., type of housing (buy house, buy condo, rent house, rent apartment..); where to shop (Mall A, Mall B, catalog...); transportation to work (car, bus, subway, walk, bike)

Two types of explanatory variables

1. characteristics of the chooser - constant across the choice set; e.g., income, education, demographics
2. characteristics of the choices - take on different values for each response choice; e.g., cost of getting to work, time needed

let \mathbf{x}_{ij} denote values of covariates for response choice j ($j = 1, \dots, K$ for subject i ,
 $\mathbf{x}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{iK})$)

$\pi_j(\mathbf{x}_i)$ probability of choice j for subject i

Let $C_i =$ set of possible response choices for subject i

McFadden (1974) proposed model for covariates that are characteristics of the choice

$$\pi_j(\mathbf{x}_i) = \frac{\exp(\boldsymbol{\beta}'\mathbf{x}_{ij})}{\sum_{h \in C_i} \exp(\boldsymbol{\beta}'\mathbf{x}_{ih})}$$

For this model, for choices a and b ,

$$\log \frac{\pi_a(\mathbf{x}_i)}{\pi_b(\mathbf{x}_i)} = \boldsymbol{\beta}'(\mathbf{x}_{ia} - \mathbf{x}_{ib})$$

Remarks

- influence of particular covariate on subject's choice between a and b depends on the distance between subject's values of that variable for these choices
- if $\mathbf{x}_{ia} = \mathbf{x}_{ib}$ then $\pi_a = \pi_b$
- odds of a being chosen over b does *not* depend on which other choices are in the choice set or on what covariates are for those choices (independence from irrelevant alternatives; Luce, 1959)
- baseline category logit model has the form

$$\pi_j(\mathbf{x}_i) = \frac{\exp(\boldsymbol{\beta}'_j \mathbf{x}_i)}{\sum_{h=1}^K \exp(\boldsymbol{\beta}'_h \mathbf{x}_i)}$$

with $\boldsymbol{\beta}_k = \mathbf{0}$, for 'characteristics of the chooser'

McFadden's discrete choice model,

$$\pi_j(\mathbf{x}_i) = \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_{ij})}{\sum_{h \in C_i} \exp(\boldsymbol{\beta}' \mathbf{x}_{ih})}$$

for 'characteristics of choices'. It can also handle 'characteristics of the chooser' and has baseline category logit as a special case

Convert $\boldsymbol{\beta}'_j \mathbf{x}_i$ to $\boldsymbol{\beta}' \mathbf{x}_{ij}$.

Reference for all this, McFadden (1974) *Frontiers in Econometrics*

Features and Properties of Multinomial Distribution

pmf

$$P(Y_1 = y_1, Y_2 = y_2, \dots, Y_K = y_K; m, \boldsymbol{\pi}) = \binom{m}{\mathbf{y}} \pi_1^{y_1} \pi_2^{y_2} \dots \pi_K^{y_K}$$

Moment generating function

$$\begin{aligned} M_Y(t) &= E[\exp(t'Y)] \\ &= E[\exp(\sum t_j Y_j)] \\ &= \left\{ \sum \pi_j e^{t_j} \right\}^m \end{aligned}$$

Cumulant generating function

$$K_Y(t) = m \log \sum \pi_j e^{t_j}$$

Moments

$$E[Y_r] = m\pi_r$$

$$Cov(Y_r, Y_s) = \begin{cases} m\pi_r(1 - \pi_r) & r = s \\ -m\pi_r\pi_s & r \neq s \end{cases}$$

Also, marginally, the probability of being in category r is binomial,

$$Y_r|m \sim \text{Bin}(m, \pi_r)$$

The multinomial covariance matrix

$$\begin{aligned} \text{Var}(\mathbf{Y}) &= \Sigma_Y \\ &= m\{\text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'\} \end{aligned}$$

Generalized inverse etc.

It is easy to show that

$$\Sigma_Y^- = \frac{1}{m} \text{diag}(\boldsymbol{\pi})^{-1} = \text{diag}\left(\frac{1}{m\pi_j}\right), 1 \leq j \leq K$$

is the generalized inverse of Σ

Maximum likelihood

i th observation y_i contribute

$$\begin{aligned} l(\boldsymbol{\pi}_i, \mathbf{y}_i) &= \sum_j y_{ij} \log \pi_{ij} \\ &\quad \sum_j y_{ij} = m_i, \sum_j \pi_{ij} = 1 \\ l(\boldsymbol{\pi}, \mathbf{y}) &= \sum_i \sum_j y_{ij} \log \pi_{ij} \end{aligned}$$

The j th component of the score for observation i (derived using Lagrange multipliers) can be shown to be

$$\frac{\partial l(\boldsymbol{\pi}_i; \mathbf{y}_i)}{\partial \pi_{ij}} = \frac{y_{ij} - m_i \pi_{ij}}{\pi_{ij}}$$

(re: $\sum_j \pi_{ij} = 1$).

Derivation

Maximize $\sum \sum y_{ij} \log \pi_{ij}$ subject to the constraint $\sum_j \pi_{ij} = 1$.

So (for observation i), maximize

$$\sum_j y_{ij} \log \pi_{ij} - \lambda(\sum_j \pi_{ij} - 1).$$

$$\begin{aligned} \frac{\partial}{\partial \pi_{ij}} &= \frac{y_{ij}}{\pi_{ij}} - \lambda = 0 \\ &\text{multiply both sides by } \pi_{ij} \\ &\equiv y_{ij} - \lambda \pi_{ij} = 0 \\ &\equiv \text{now sum over } j; \text{ ok since lik eqs. for all } \pi_{ij} \text{ are zero} \\ &\equiv \sum_j y_{ij} - \lambda \sum_j \pi_{ij} = 0 \end{aligned}$$

So, $\lambda = m$.

The lik eqs. can be re-expressed using the generalized inverse of the covariance matrix of Y_i as

$$\frac{\partial l(\boldsymbol{\pi}; \mathbf{y})}{\partial \boldsymbol{\pi}_i} = m_i \Sigma_{Y_i}^{-1} (\mathbf{y}_i - m \boldsymbol{\pi}_i)$$

Deviance

$$\begin{aligned} D(\mathbf{y}, \hat{\boldsymbol{\pi}}) &= 2[l(\tilde{\boldsymbol{\pi}}; \mathbf{y}) - l(\hat{\boldsymbol{\pi}}; \mathbf{y})] \\ &= 2 \sum_i \sum_j y_{ij} \log \tilde{\boldsymbol{\pi}} - 2 \sum_i \sum_j y_{ij} \log \hat{\boldsymbol{\pi}} \\ &= 2 \sum_i \sum_j y_{ij} \log \frac{m_i \tilde{p}_{ij}}{m_i \hat{\pi}_{ij}} \\ &= 2 \sum_i \sum_j y_{ij} \log \frac{y_{ij}}{\hat{\mu}_{ij}} \\ &\rightarrow \chi^2 \text{ under } H_0: \text{ model holds as } \hat{\mu}_{ij} \rightarrow \infty \end{aligned}$$

Overdispersion

in presence of overdispersion, i.e.,

$$\text{Var}(\mathbf{Y}) = \sigma^2 \Sigma$$

using quasilielihood, can estimate

$$\hat{\sigma}^2 = X^2 / (n(k-1) - p)$$

Or be fully parametric using mixtures, Dirichlet-Multinomial model

Dirichlet distribution

Generalization of the Beta distribution; has K parameters; conjugate prior for multinomial probability vector

$$\boldsymbol{\pi} \sim D(\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1 + \dots + \alpha_K)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \pi_1^{\alpha_1 - 1} \dots \pi_K^{\alpha_K - 1}.$$

Let $\alpha_0 = \sum_{j=1}^K \alpha_j = m^*$.

$$\begin{aligned} E[\pi_j] &= \frac{\alpha_j}{\alpha_0} = \pi_j^* \\ \text{Var}(\pi_j) &= \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)} \end{aligned}$$

We end up doing the regression on π_j^* .

Details for Multinomial-Dirichlet model

$$\begin{aligned}
E[Y_j] &= EE[Y_j|\pi_j] \\
&= E[m\pi_j] \\
&= m\alpha_j/\alpha_0
\end{aligned}$$

$$\begin{aligned}
Var[Y_j] &= E[Var[Y_j|\pi_j]] + Var[E[Y_j|\pi_j]] \\
&= E[m\pi_j(1 - \pi_j)] + Var(m\pi_j) \\
&= mE[\pi_j] - mE[\pi_j^2] + Var(m\pi_j) \\
&= m\frac{\alpha_j}{\alpha_0} - m\left[\frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0\alpha_0} \frac{1}{\alpha_0 + 1} + \left(\frac{\alpha_j}{\alpha_0}\right)^2\right] + m^2\frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0\alpha_0} \frac{1}{\alpha_0 + 1} \\
&\quad \text{plug in } \pi^* \text{ for } \alpha_j/\alpha_0 \\
&= m\pi_j^* - m\left[\frac{\pi_j^*(1 - \pi_j^*)}{m^* + 1} + \pi_j^{*2}\right] + m^2\frac{\pi_j^*(1 - \pi_j^*)}{m^* + 1} \\
&= m[\pi_j^* - \pi_j^{*2}] - m\frac{\pi_j^*(1 - \pi_j^*)}{m^* + 1}(1 - m) \\
&= m\pi_j^*(1 - \pi_j^*) - m\pi_j^*(1 - \pi_j^*)\frac{1 - m}{m^* + 1} \\
&= m\pi_j^*(1 - \pi_j^*)\left[\frac{m^* + m}{m^* + 1}\right]
\end{aligned}$$

Remarks:

- as $m^* \rightarrow \infty$, $Var(\pi_j) \rightarrow 0$ (no overdispersion).
- if $m = 1$, no overdispersion (factor equals 1)
- largest overdispersion factor is m

Examples

This data was collected as part of the High School and Beyond study conducted by the National Education Longitudinal Studies program of the National Center for Education Statistics. One question of interest was determining which factors are related to the choice of the type of high school program (academic, vocational, or general). Factors include gender, race, socioeconomic status, school type, and scores on subject area tests.

```

> hsb <- read.table("/home/mdaniels/classes/glm/spring2009.dir/hsb.txt",
+   header = T)
> hsb[1:10, ]

```

	id	gender	race	ses	schtyp	prog	read	write	math	science	socst
1	70	male	white	low	public	general	57	52	41	47	57
2	121	female	white	middle	public	vocation	68	59	53	63	61
3	86	male	white	high	public	general	44	33	54	58	31
4	141	male	white	high	public	vocation	63	44	47	53	56

5	172	male	white	middle	public	academic	47	52	57	53	61
6	113	male	white	middle	public	academic	44	52	51	63	61
7	50	male	african-amer	middle	public	general	50	59	42	53	61
8	11	male	hispanic	middle	public	academic	34	46	45	39	36
9	84	male	white	middle	public	general	63	57	54	58	51
10	48	male	african-amer	middle	public	academic	57	55	52	50	51

```
> library(VGAM)
> prog <- vglm(formula = prog ~ gender + as.factor(race) + as.factor(ses) +
+   as.factor(schtyp) + read + write + math + science + socst,
+   family = multinomial(), data = hsb)
> summary(prog)
```

Call:

```
vglm(formula = prog ~ gender + as.factor(race) + as.factor(ses) +
+   as.factor(schtyp) + read + write + math + science + socst,
+   family = multinomial(), data = hsb)
```

Pearson Residuals:

	Min	1Q	Median	3Q	Max
log(mu[,1]/mu[,3])	-5.4640	-0.51707	0.17121	0.554379	3.2906
log(mu[,2]/mu[,3])	-5.0498	-0.44109	-0.20940	-0.078704	2.9835

Coefficients:

	Value	Std. Error	t value
(Intercept):1	-7.4816368	2.104634	-3.554840
(Intercept):2	-3.8494902	2.148607	-1.791621
gendermale:1	0.3210247	0.502104	0.639358
gendermale:2	0.2283564	0.527437	0.432955
as.factor(race)asian:1	0.6994617	1.470012	0.475821
as.factor(race)asian:2	2.0524960	1.437003	1.428317
as.factor(race)hispanic:1	0.1991482	0.839365	0.237260
as.factor(race)hispanic:2	-0.4328644	0.917063	-0.472011
as.factor(race)white:1	-0.3361177	0.748056	-0.449321
as.factor(race)white:2	-0.0394299	0.763317	-0.051656
as.factor(ses)low:1	-0.0474770	0.704564	-0.067385
as.factor(ses)low:2	1.0511645	0.731242	1.437506
as.factor(ses)middle:1	-1.1815758	0.570067	-2.072697
as.factor(ses)middle:2	-0.4786224	0.622528	-0.768837
as.factor(schtyp)public:1	-2.0552546	0.834734	-2.462167
as.factor(schtyp)public:2	-1.4707368	0.889453	-1.653529
read:1	0.0348108	0.034224	1.017157
read:2	-0.0093712	0.034424	-0.272230
write:1	0.0316619	0.035857	0.883009
write:2	-0.0046150	0.037108	-0.124366
math:1	0.1139920	0.038851	2.934101

math:2	0.0047003	0.038137	0.123247
science:1	-0.0522961	0.034247	-1.527020
science:2	0.0496391	0.033667	1.474429
socst:1	0.0804008	0.029382	2.736435
socst:2	0.0606308	0.028819	2.103824

Number of linear predictors: 2

Names of linear predictors: $\log(\mu[,1]/\mu[,3])$, $\log(\mu[,2]/\mu[,3])$

Dispersion Parameter for multinomial family: 1

Residual Deviance: 305.8705 on 374 degrees of freedom

Log-likelihood: -152.9353 on 374 degrees of freedom

Number of Iterations: 5

```
> prog <- vglm(formula = prog ~ as.factor(schtyp) + math + socst,
+   family = multinomial(), data = hsb)
> summary(prog)
```

Call:

```
vglm(formula = prog ~ as.factor(schtyp) + math + socst, family = multinomial(),
     data = hsb)
```

Pearson Residuals:

	Min	1Q	Median	3Q	Max
$\log(\mu[,1]/\mu[,3])$	-7.2261	-0.60856	0.23299	0.62678	3.0640
$\log(\mu[,2]/\mu[,3])$	-5.0981	-0.41570	-0.26718	-0.14628	2.2416

Coefficients:

	Value	Std. Error	t value
(Intercept):1	-7.334858	1.712305	-4.2836
(Intercept):2	-2.718936	1.709884	-1.5901
as.factor(schtyp)public:1	-1.767736	0.796361	-2.2198
as.factor(schtyp)public:2	-1.119714	0.854171	-1.3109
math:1	0.109957	0.029497	3.7277
math:2	0.035508	0.030745	1.1549
socst:1	0.078127	0.023399	3.3390
socst:2	0.040484	0.023351	1.7337

Number of linear predictors: 2

Names of linear predictors: $\log(\mu[,1]/\mu[,3])$, $\log(\mu[,2]/\mu[,3])$

Dispersion Parameter for multinomial family: 1

Residual Deviance: 335.6515 on 392 degrees of freedom

Log-likelihood: -167.8258 on 392 degrees of freedom

Number of Iterations: 5

```
> pred.data <- data.frame(math = mean(hsb$math), socst = mean(hsb$socst),
+   schtyp = c("public", "private"))
> predicted <- predict(prog, newdata = pred.data, type = "response")
> cbind(c("public", "private"), predicted)
```

	academic	general	vocation
1 "public"	"0.502049504289693"	"0.267890605685388"	"0.230059890024919"
2 "private"	"0.736733688797799"	"0.205631228914954"	"0.0576350822872472"

0.6 Count data: log linear models

start with the Poisson distribution, $P(\mu)$ with mean μ

$$P(Y = y) = \frac{\exp(-\mu)\mu^y}{y!} : y = 0, 1, 2, \dots$$

this can be derived as the limit of a binomial with large n and small π (HW problem)
 similarly, if set $\mu = mp$, can think of an m independent $Pois(p)$ (aggregation)
 recall the mgf for a Poisson r.v. is

$$\begin{aligned} m_Y(t) &= E[\exp(tY)] \\ &= \sum_{y=1}^{\infty} \frac{e^{ty} e^{-\mu} \mu^y}{y!} \\ &= \sum_{y=1}^{\infty} \frac{e^{-\mu} (e^t \mu)^y}{y!} \\ &= e^{-\mu} e^{\mu \exp(t)} \sum_{y=1}^{\infty} \frac{e^{-\mu \exp(t)} (e^t \mu)^y}{y!} \\ &= \exp(\mu(e^t - 1)) \end{aligned}$$

so, the cumulant generating function is

$$K_Y(t) = \mu(\exp(t) - 1)$$

all the cumulants are equal to μ
 It is also easy to show that

$$\frac{Y - \mu}{\mu^{1/2}} \sim N(0, 1) + O_p(\mu^{-1/2})$$

so approach normality as $\mu \rightarrow \infty$ (show for HW; use mgf or cgf).

The variance stabilizing transform is the square root (though sometime use logarithm transformation),

$$\begin{aligned} E[Y^{1/2}] &\approx \mu^{1/2} \left[1 - \frac{1}{8\mu}\right] \\ &= \mu^{1/2} \left[1 - O_p\left(\frac{1}{\mu}\right)\right] \\ &\approx \mu^{1/2} \\ \text{Var}[Y^{1/2}] &\approx \left\{1 + \frac{3}{8\mu}\right\} / 4 \\ &= \{1 + O_p(1/\mu)\} / 4 \\ &\approx 1/4 \end{aligned}$$

Variance does NOT depend on mean. Derive for HW. So, for large μ can model $y^{1/2}$ using normal methods. The error terms here are $O_p(\mu^{-1})$ (i.e., asymptotics are $\mu \rightarrow \infty$).

Log likelihood

$$\ell(\boldsymbol{\mu}, \mathbf{y}) = \sum (y_i \log \mu_i - \mu_i).$$

and deviance is

$$D(\mathbf{y}; \boldsymbol{\mu}) = 2 \sum (y_i \log(y_i/\mu_i) - (y_i - \mu_i)).$$

the second term drops out if an intercept. And similar to the binomial, if expand in a Taylor series, can show asymptotically equivalent to Pearson's χ^2 .

The likelihood equations are

$$\sum_i \frac{(y_i - \mu_i)}{\text{Var}(Y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_j} = \mathbf{0} : j = 0, 1, \dots, p.$$

With the canonical link (log),

$$\sum (y_i - \mu_i) x_{ij} = 0 : j = 0, 1, \dots, p.$$

equate parameters to their sufficient statistics.

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (X'WX)^{-1}$$

where $W = \text{diag} \left(\left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 / \text{Var}(Y_i) \right)$. Under the canonical link, $W = \mu_i$.

Overdispersion $\text{Var}(Y) > E(Y)$; very common

e.g., if $Y_i \sim P(\mu_i)$ when all relevant covariates are in the model, probably NOT Poisson when only some of these covariates (b/c variance is wrong). How deal with this? Similar to binominal

- parametric mixture approach - negative binomial (Gamma mixture of Poissons)
- quasi-likelihood: $Var(Y) = \sigma^2 \mu$ (not assume parametric distribution)

Negative binomial

$$\begin{aligned} Y_i | \mu_i &\sim P(\mu_i) \\ \mu_i &\sim \text{Gamma}(\phi\mu, \phi) \end{aligned}$$

where $E[\mu_i] = \mu$ and $Var(\mu_i) = \phi\mu/\phi^2 = \mu/\phi$.

Easy to show that the marginal variance of Y_i is

$$\begin{aligned} Var(Y_i) &= Var[E(Y_i | \mu_i)] + E[Var(Y_i | \mu_i)] \\ &= \mu/\phi + \mu \\ &= \mu(1 + 1/\phi) \end{aligned}$$

So the overdispersion parameter is $(1 + 1/\phi)$. The probability mass function is

$$P(Y = y) = \frac{\Gamma(y + \phi\mu)\phi^{\phi\mu}}{y!\Gamma(\phi\mu)(1 + \phi)^{y+\phi\mu}}.$$

Derive for HW. Not follow exponential family form. Can maximize likelihood directly on this.

There are other ways to parameterize that provide different assumptions about the marginal variance (either linear or quadratic in the mean; more soon). A mixture that doesn't have a closed form (but is somewhat common) is a Poisson normal mixture with $\log \mu_i \sim N(\lambda_i, \sigma^2)$.

Another parameterization for the negative binomial is

$$\begin{aligned} Y_i | \mu_i &\sim P(\mu_i) \\ \mu_i &\sim \text{Gamma}(\phi, \phi/\mu) \end{aligned}$$

such that $E[\mu_i] = \mu$ and $Var(\mu_i) = \mu^2/\phi$, where $\phi > 0$ is the shape parameter. As $\phi \rightarrow \infty$,

it is less skewed. The marginal distribution for Y is

$$\begin{aligned}
p(y_i) &= \int p(y_i|\mu_i)p(\mu_i)d\mu_i \\
&= \int \frac{\exp(-\mu_i)\mu_i^{y_i}}{y!} \frac{\exp(-\mu_i\phi/\mu)\mu_i^{\phi-1}}{\Gamma(\phi)(\phi/\mu)^{-\phi}} \\
&= \int \frac{(\phi/\mu)^\phi}{\Gamma(y_i+1)\Gamma(\phi)} \exp(-\mu_i[1+\phi/\mu])\mu_i^{y_i+\phi-1}d\mu_i \\
&= \int \frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)} \left(\frac{\phi}{\mu}\right)^\phi \left(1+\frac{\phi}{\mu}\right)^{-(y+\phi)} \\
&= \int \frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)} \left(\frac{\phi}{\mu}\right)^\phi \left(\frac{\mu}{\phi+\mu}\right)^{(y+\phi)} \\
&= \int \frac{\Gamma(y+\phi)}{\Gamma(y+1)\Gamma(\phi)} \left(\frac{\phi}{\mu+\phi}\right)^\phi \left(\frac{\mu}{\phi+\mu}\right)^{(y_i)} : y = 0, 1, 2, \dots
\end{aligned}$$

Via iterated expectations, easy to show $E(Y) = \mu$ and $Var(Y) = \mu + \mu^2/\phi$ (quadratic variance function).

As $\phi \rightarrow \infty$, $Var(\mu_i) \rightarrow 0$ and $Var(Y) \rightarrow \mu$, and negative binomial goes to a Poisson.

For fixed ϕ (dispersion parameter), exponential family,

$$\begin{aligned}
P(Y = y) &= \exp\left\{y \log \frac{\mu}{\mu + \phi} + \phi \log \frac{\phi}{\phi + \mu} + \log \frac{\Gamma(y + \phi)}{\Gamma(\phi)\Gamma(y + 1)}\right\} \\
&= \exp\left\{[y\theta + \log(1 - \exp(\phi\theta))]/(1/\phi) + \log \frac{\Gamma(y + \phi)}{\Gamma(\phi)\Gamma(y + 1)}\right\}
\end{aligned}$$

where $\theta = (1/\phi) \log \frac{\mu}{\mu + \phi}$, $b(\theta) = \log(1 - \exp(\phi\theta))$, $a(\phi) = 1/\phi$.

Can fit via Newton-Raphson. For link function g (typically use log instead of canonical link),

$$\frac{\partial^2 L}{\partial \beta_j \partial \phi} = \sum_i \frac{(y_i - \mu_i)x_{ij}}{(\phi + \mu_i)^2} \frac{\partial \mu_i}{\partial \eta_i}$$

and

$$E \left[\frac{\partial^2 L}{\partial \beta_j \partial \phi} \right] = \mathbf{0}$$

so $\hat{\beta}$ and $\hat{\phi}$ are orthogonal (asymptotically independent).

0.6.1 Poisson regression in epidemiological studies

Interest in impact of some exposure on disease. For example, prospective cohort studies (follow group of exposed and unexposed individuals over time and see who develops disease).

$$\begin{aligned}
Y &\sim P(\mu) \\
\log \mu &= \log E + \mathbf{x}\beta
\end{aligned}$$

where E is the expected number of deaths (constructed from standardized rate times person years of follow-up OR just person years); even if the former, this is treated as known.

So $\mu = E \exp(\mathbf{x}\boldsymbol{\beta})$ and $\exp(\mathbf{x}\boldsymbol{\beta})$ is the *rate*.

How do in R? “offset=logE”.

0.6.2 Connection between log-linear models and multinomial response models

Based on the fact that can derive multinomial from a set of independent Poissons if condition on their total.

$Y_1, \dots, Y_k \sim P(\mu_k)$ with $H_0 : \mu_1 = \dots = \mu_k = \exp(\beta_0)$ with alternative $\log \mu_j = \beta_0 + \beta_1 x_j$.

Poisson log likelihood

$$l(\beta_0, \beta_1) = \beta_0 \sum y_j + \beta_1 \sum x_j y_j - \sum \exp(\beta_0 + \beta_1 x_j)$$

Transform to (τ, β_1) where $\tau = \sum \exp(\beta_0 + \beta_1 x_j)$. Let $y. = \sum_j y_j$.

Note that

$$\tau = \exp(\beta_0) \sum \exp(\beta_1 x_j) \leftrightarrow \beta_0 = \log \tau - \log \sum \exp(\beta_1 x_j)$$

So,

$$\begin{aligned} l(\tau, \beta_1) &= y. \log \tau - y. \log \sum \exp(\beta_1 x_j) + \beta_1 \sum x_j y_j - \tau \\ &= y. \log \tau - \tau + \beta_1 \sum x_j y_j - y. \log \sum \exp(\beta_1 x_j) \\ &= l_m(\tau; m) + l_{Y|m}(\beta_1; y) \end{aligned}$$

where $m = y.$ and

$$\begin{aligned} l_m(\tau; m) &= y. \log \tau - \tau \\ l_{Y|m}(\beta_1; y) &= \beta_1 \sum x_j y_j - y. \log \sum \exp(\beta_1 x_j) \end{aligned}$$

$l_{Y|m}(\beta_1; y)$ is the multinomial likelihood with $\pi_j = \frac{\exp(\beta_1 x_j)}{\sum_i \exp(\beta_1 x_i)}$. All of the information on β_1 is from $l_{Y|m}(\beta_1; y)$.

Similar connections with more general models to address with conditional likelihood.

For example,

$$\log \mu_{ij} = \phi_i + x'_{ij} \boldsymbol{\beta}.$$

Asymptotic distribution of Pearson's χ^2

Recall, $X^2 = \sum_{i=1}^N \frac{(y_i - \mu_i)^2}{\mu_i}$.

Key Result: Rao (1973) [Linear statistical inference and applications; Wiley]

Let Y be a multivariate normal with mean $\boldsymbol{\nu}$ and covariance matrix \mathbf{B} . A necessary and sufficient condition for $(\mathbf{Y} - \boldsymbol{\nu})' \mathbf{C} (\mathbf{Y} - \boldsymbol{\nu})$ to have a χ^2 distribution is $\mathbf{BCBCB} = \mathbf{BCB}$. The degrees of freedom is equal to the rank of \mathbf{CB} .

Note: when \mathbf{B} is nonsingular, the condition simplifies to $\mathbf{CBC} = \mathbf{C}$.

Also, note:

$$\begin{aligned} X^2 &= \sum_{i=1}^N \left(\frac{(y_i - \hat{\mu}_i)}{\hat{\mu}_i^{1/2}} \right)^2 \\ &= \sum_{i=1}^N \left(\frac{\sqrt{n}(p_i - \hat{\pi}_i)}{\hat{\pi}_i^{1/2}} \right)^2 \\ &= \mathbf{e}'\mathbf{e} \end{aligned}$$

where $\mathbf{e} = (e_1, \dots, e_n)$ with $e_i = \frac{\sqrt{n}(p_i - \hat{\pi}_i)}{\hat{\pi}_i^{1/2}}$ and $p_i = \frac{y_i}{n}$ where $n = \sum_{i=1}^N y_i$ and $\hat{\pi}_i = \frac{\hat{\mu}_i}{n}$.

So, connecting this to the result from Rao,

$$\begin{aligned} \mathbf{Y} &= \mathbf{e}, \boldsymbol{\nu} = \mathbf{0}, \mathbf{C} = \mathbf{I} \\ \mathbf{B} &= \mathbf{I} - \boldsymbol{\pi}_0^{1/2} \boldsymbol{\pi}_0^{1/2'} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}' \end{aligned}$$

where $\mathbf{A} = \text{diag}(\boldsymbol{\pi}_0)^{-1/2} (\partial\boldsymbol{\pi}/\partial\boldsymbol{\beta}_0)$ where the subscript 0 denotes the true value of the parameters. The form for \mathbf{B} was derived when we standardized the Pearson residuals earlier in the semester.

Recall the condition in the Result is $\mathbf{BCBCB} = \mathbf{BCB}$. Here $\mathbf{C} = \mathbf{I}$ so the condition reduces to $\mathbf{BBB} = \mathbf{BB}$. This will hold if \mathbf{B} is idempotent ($\mathbf{BB} = \mathbf{B}$). Check this.

$$(\mathbf{I} - \boldsymbol{\pi}_0^{1/2} \boldsymbol{\pi}_0^{1/2'} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}')(\mathbf{I} - \boldsymbol{\pi}_0^{1/2} \boldsymbol{\pi}_0^{1/2'} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}') = ?.$$

Note:

$$\begin{aligned} \mathbf{A}'\boldsymbol{\pi}_0^{1/2} &= \left(\frac{\partial\boldsymbol{\pi}}{\partial\boldsymbol{\beta}_0} \right) \text{diag}(\boldsymbol{\pi}_0)^{-1/2} \boldsymbol{\pi}_0^{1/2} \\ \mathbf{A}'\boldsymbol{\pi}_0^{1/2} &= \left(\frac{\partial\boldsymbol{\pi}}{\partial\boldsymbol{\beta}_0} \right) \mathbf{1}' \\ &\text{note that } \mathbf{1}' \text{ is } N \times 1 \\ &= \sum_{i=1}^N \frac{\partial\pi_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} \\ &= \frac{\partial}{\partial\boldsymbol{\beta}} \sum_{i=1}^N \pi_i(\boldsymbol{\beta}) \\ &= \frac{\partial}{\partial\boldsymbol{\beta}} (1) \\ &= 0. \end{aligned}$$

We can now use this to simplify \mathbf{BB} ,

$$\begin{aligned} \mathbf{BB} &= \mathbf{I} - \boldsymbol{\pi}_0^{1/2} \boldsymbol{\pi}_0^{1/2'} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}' - \boldsymbol{\pi}_0^{1/2} \boldsymbol{\pi}_0^{1/2'} + \boldsymbol{\pi}_0^{1/2} \boldsymbol{\pi}_0^{1/2'} \boldsymbol{\pi}_0^{1/2} \boldsymbol{\pi}_0^{1/2'} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}' + \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}' \\ &= \mathbf{I} - \boldsymbol{\pi}_0^{1/2} \boldsymbol{\pi}_0^{1/2'} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}' - \boldsymbol{\pi}_0^{1/2} \boldsymbol{\pi}_0^{1/2'} + \boldsymbol{\pi}_0^{1/2} (1) \boldsymbol{\pi}_0^{1/2'} \\ &= \mathbf{I} - \boldsymbol{\pi}_0^{1/2} \boldsymbol{\pi}_0^{1/2'} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} \mathbf{A}' \\ &= \mathbf{B}. \end{aligned}$$

And since \mathbf{e} is asymptotically multivariate normal (Poisson case for HW; did before for binomial), X^2 is asymptotically χ^2 .

What about degrees of freedom? Let $\dim(\beta) = p$. For a symmetric, idempotent matrix, the rank is equal to the trace. So

$$\begin{array}{rcl} I & - & \pi_0^{1/2} \pi_0^{1/2'} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' \\ N & - & 1 - p. \end{array}$$

Thus, the df is $N - p - 1$.

Examples Famous study conducted by Sir Richard Doll. In 1951, all British doctors sent a brief questionnaire about whether they smoked tobacco. Data is the number of deaths from coronary heart disease among male doctors 10 years after the survey. It also has the total number of person years of observation.

```
> doll <- read.table("smoke.txt", header = T)
> doll
```

	age.cat	age	smoke	y	py
1	35-44	1	1	32	52407
2	45-55	2	1	104	43248
3	55-64	3	1	206	28612
4	65-74	4	1	186	12663
5	75-84	5	1	102	5317
6	35-44	1	0	2	18790
7	45-55	2	0	12	10673
8	55-64	3	0	28	5710
9	65-74	4	0	28	2585
10	75-84	5	0	31	1462

```
> brit <- glm(y ~ age + smoke + offset(log(py)), family = poisson,
+ data = doll)
> summary(brit)
```

Call:

```
glm(formula = y ~ age + smoke + offset(log(py)), family = poisson,
    data = doll)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-4.5712	-2.7562	0.2857	1.4261	3.7183

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-8.11833	0.13929	-58.282	< 2e-16 ***
age	0.83583	0.02904	28.777	< 2e-16 ***
smoke	0.40637	0.10720	3.791	0.00015 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 935.067 on 9 degrees of freedom
 Residual deviance: 69.182 on 7 degrees of freedom
 AIC: 130.25

Number of Fisher Scoring iterations: 4

```
> rate <- (doll$y/doll$py) * 1000
> plot(doll$age[doll$smoke == 0], rate[doll$smoke == 0], lty = 2)
> lines(doll$age[doll$smoke == 1], rate[doll$smoke == 1], lty = 3)
> doll$agesq <- doll$age * doll$age
> brit.glm <- glm(y ~ age + smoke + agesq + age * smoke + offset(log(py)),
+   family = poisson, data = doll)
> summary(brit.glm)
```

Call:

```
glm(formula = y ~ age + smoke + agesq + age * smoke + offset(log(py)),
    family = poisson, data = doll)
```

Deviance Residuals:

1	2	3	4	5	6	7	8
0.43820	-0.27329	-0.15265	0.23393	-0.05700	-0.83049	0.13404	0.64107
9	10						
-0.41058	-0.01275						

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-10.79176	0.45008	-23.978	< 2e-16 ***
age	2.37648	0.20795	11.428	< 2e-16 ***
smoke	1.44097	0.37220	3.872	0.000108 ***
agesq	-0.19768	0.02737	-7.223	5.08e-13 ***
age:smoke	-0.30755	0.09704	-3.169	0.001528 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

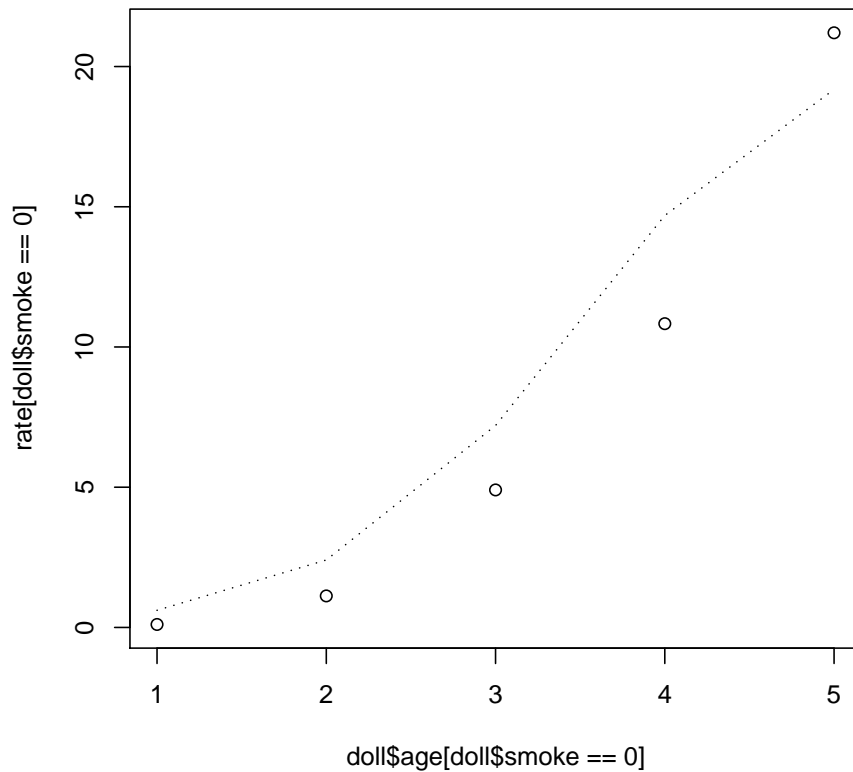
(Dispersion parameter for poisson family taken to be 1)

Null deviance: 935.0673 on 9 degrees of freedom
 Residual deviance: 1.6354 on 5 degrees of freedom
 AIC: 66.703

Number of Fisher Scoring iterations: 4

```
> cbind(brit.glm$fitted.values, doll$y)
```

```
      [,1] [,2]
1  29.584734  32
2  106.811960 104
3  208.198646 206
4  182.827893 186
5  102.576767 102
6   3.414801   2
7  11.541629  12
8  24.743377  28
9  30.229155  28
10 31.071038  31
```



Example 2

For the 30 Galapagos islands, this data consists of the count of the number of species of tortoise found on each island and the number that are endemic to the island. We have geographic variables for each island.

```
> gala <- read.table("gala.txt", header = T)
> gala <- gala[, -3]
```

```
> tort <- glm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
+   family = poisson, gala)
> summary(tort)
```

Call:

```
glm(formula = Species ~ Area + Elevation + Nearest + Scruz +
     Adjacent, family = poisson, data = gala)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-8.2752	-4.4966	-0.9443	1.9168	10.1849

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	3.155e+00	5.175e-02	60.963	< 2e-16 ***
Area	-5.799e-04	2.627e-05	-22.074	< 2e-16 ***
Elevation	3.541e-03	8.741e-05	40.507	< 2e-16 ***
Nearest	8.826e-03	1.821e-03	4.846	1.26e-06 ***
Scruz	-5.709e-03	6.256e-04	-9.126	< 2e-16 ***
Adjacent	-6.630e-04	2.933e-05	-22.608	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 3510.73 on 29 degrees of freedom
 Residual deviance: 716.85 on 24 degrees of freedom
 AIC: 889.68

Number of Fisher Scoring iterations: 5

```
> plot(log(tort$fitted.values), log((gala$Species - tort$fitted.values)^2),
+   xlab = expression(hat(mu)), ylab = expression((y - hat(mu))^2))
> abline(0, 1)
> library(MASS)
> tort.nb <- glm.nb(Species ~ Area + Elevation + Nearest + Scruz +
+   Adjacent, gala)
> summary(tort.nb)
```

Call:

```
glm.nb(formula = Species ~ Area + Elevation + Nearest + Scruz +
        Adjacent, data = gala, init.theta = 1.67460228613663, link = log)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.1344	-0.8597	-0.1476	0.4576	1.8416

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	2.9065247	0.2510344	11.578	< 2e-16	***
Area	-0.0006336	0.0002865	-2.211	0.027009	*
Elevation	0.0038551	0.0006916	5.574	2.49e-08	***
Nearest	0.0028264	0.0136618	0.207	0.836100	
Scruz	-0.0018976	0.0028096	-0.675	0.499426	
Adjacent	-0.0007605	0.0002278	-3.338	0.000842	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

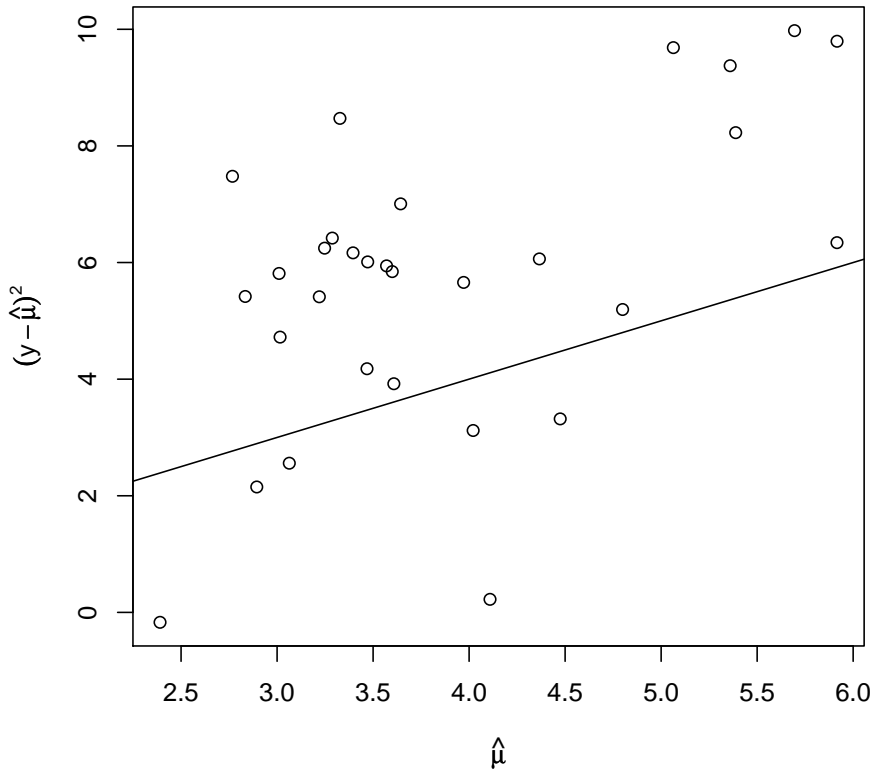
(Dispersion parameter for Negative Binomial(1.6746) family taken to be 1)

Null deviance: 88.431 on 29 degrees of freedom
 Residual deviance: 33.196 on 24 degrees of freedom
 AIC: 304.22

Number of Fisher Scoring iterations: 1

Theta: 1.675
 Std. Err.: 0.442

2 x log-likelihood: -290.223



0.7 Nuisance parameters and conditional likelihood

Consider the proportional odds model,

$$\log \frac{\gamma_{ij}}{1 - \gamma_{ij}} = \theta_j - \boldsymbol{\beta} \mathbf{x}_i.$$

Often, only $\boldsymbol{\beta}$ of interest. Can we construct a (conditional) likelihood that removes the nuisance parameters?

There are major problems when many nuisance parameters:

1. consistency of parameters of interest. when $p/n \rightarrow k > 0$
2. maximizing a function of many variables

Let $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$, where $\boldsymbol{\psi}$ is the parameter of interest and $\boldsymbol{\lambda}$ is the nuisance parameter. Suppose for each fixed value of $\boldsymbol{\psi}_0$ of $\boldsymbol{\psi}$, $S_{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0)$ is sufficient for $\boldsymbol{\lambda}$ and complete. There are two cases here:

1. $S_{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0)$ depends on $\boldsymbol{\psi}_0$

2. $S_{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0)$ is independent of $\boldsymbol{\psi}_0$

Case 1.

$[\mathbf{Y}|S_{\boldsymbol{\lambda}}(\boldsymbol{\psi}), \boldsymbol{\lambda}] = [\mathbf{Y}|S_{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0)]$ with the equality holding for $\boldsymbol{\psi} = \boldsymbol{\psi}_0$. Try to use

$$f_{\mathbf{Y}|S_{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0)}(\mathbf{y}|S_{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0); \boldsymbol{\psi}, \boldsymbol{\lambda}).$$

Using the above conditional distribution, we cannot get a true likelihood function (more later), but we can use profile likelihood in this case. Details follow.

Let $\hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}$ be the mle of $\boldsymbol{\lambda}$ for fixed $\boldsymbol{\psi}$. Define

$$\begin{aligned} l^p(\boldsymbol{\psi}; \mathbf{y}) &= l(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}; \mathbf{y}) \\ &= \sup_{\boldsymbol{\lambda}} l(\boldsymbol{\psi}, \boldsymbol{\lambda}; \mathbf{y}) \end{aligned}$$

This is the *profile likelihood*.

Properties:

1. max of $l^p(\boldsymbol{\psi}; \mathbf{y})$ same as overall mle
2. approximate confidence sets

$$\{\boldsymbol{\psi} : 2l^p(\hat{\boldsymbol{\psi}}; \mathbf{y}) - 2l^p(\boldsymbol{\psi}; \mathbf{y}) \leq \chi_{q, 1-\alpha}^2\}$$

where $q = \dim(\boldsymbol{\psi})$.

3. confidence interval based on second derivative of $l^p(\boldsymbol{\psi}; \mathbf{y})$ at the maximum NOT reliable if $\dim(\boldsymbol{\lambda})$ is not small
4. NOT a log likelihood function
 - (a) $E \left[\frac{\partial}{\partial \boldsymbol{\psi}} l^p(\boldsymbol{\psi}; \mathbf{y}) \right] \neq \mathbf{0}$ (hw problem)
 - (b) if dimension of $\boldsymbol{\lambda}$ is not small relative to n , the above expectation can be far from zero.

Case 2.

$S_{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0) = S_{\boldsymbol{\lambda}}$ and $f_{\mathbf{y}|S_{\boldsymbol{\lambda}}}(\mathbf{y}|S_{\boldsymbol{\lambda}}; \boldsymbol{\psi})$ so

$$\begin{aligned} l_c(\boldsymbol{\psi}) &= \log f_{\mathbf{y}|S_{\boldsymbol{\lambda}}}(\mathbf{y}|S_{\boldsymbol{\lambda}}; \boldsymbol{\psi}) \\ &= \log f_{\mathbf{y}}(\mathbf{y}; \boldsymbol{\psi}, \boldsymbol{\lambda}) - \log f_{S_{\boldsymbol{\lambda}}}(\mathbf{s}_{\boldsymbol{\lambda}}; \boldsymbol{\psi}, \boldsymbol{\lambda}). \end{aligned}$$

is the conditional log likelihood for $\boldsymbol{\psi}$. This is a log likelihood.

Example: Exponential family model

Let $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$. $l(\boldsymbol{\theta}; \mathbf{y}) = \boldsymbol{\theta}^T \mathbf{s} - b(\boldsymbol{\theta})$ and can be decomposed as

$$l(\boldsymbol{\theta}; \mathbf{y}) = \boldsymbol{\psi}^T \mathbf{s}_1 - \boldsymbol{\lambda}^T \mathbf{s}_2 - b(\boldsymbol{\psi}, \boldsymbol{\lambda})$$

where s_1 and s_2 are functions of the data. Typically of this form when independent observations and model linear in the canonical parameters, $\boldsymbol{\theta}$. Sufficient statistics are $s = \mathbf{x}^T \mathbf{y}$.

Consider $Y_i \sim P(\mu_i)$ $i = 1, 2$. Suppose interested in the ratio $\psi' = \mu_1/\mu_2$. Canonical parameters are $\theta_i = \log \mu_i$. So, $\psi = \log \psi' = \theta_1 - \theta_2$. For nuisance parameters, any of the following work, $\lambda'_1 = \mu_1$, $\lambda'_2 = \mu_2$, $\lambda'_3 = \mu_1 + \mu_2$, $\lambda'_4 = \mu_1 \mu_2$.

$$\begin{aligned} l(\boldsymbol{\theta}; \mathbf{y}) &= y_1 \theta_1 + y_2 \theta_2 - \exp(\theta_1) - \exp(\theta_2) \\ &= (y_1 + y_2) \lambda_1 - y_2 \psi - \exp(\lambda_1)(1 + \exp(-\psi)) \end{aligned}$$

where $\lambda_1 = \log \lambda'_1$. This can be written in terms of any of the λ_j , $j = 1, \dots, 4$.

Let $y = (y_1 + y_2) = s_2$. y is sufficient for λ_1 whatever value ψ takes. So $Y|S_2$ is independent of λ . So conditional likelihood for inference on $\boldsymbol{\psi}$,

$$l(\boldsymbol{\psi}|s_2) = \boldsymbol{\psi}^T s_1 - b^*(\boldsymbol{\psi}; s_2).$$

Here,

$$\begin{aligned} l(\boldsymbol{\psi}|s_2) &= \log f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\psi}, \lambda) - \log f_{S_\lambda}(s_\lambda; \boldsymbol{\psi}, \lambda) \\ &= \log f_{S_2}(s_2; \boldsymbol{\psi}, \lambda) - \log f_{S_\lambda}(s_\lambda; \boldsymbol{\psi}, \lambda) \\ &= \log P(\mu_1 + \mu_2) - \log P(\lambda'_1 + \lambda'_1/\psi') = \log(\exp(\lambda_1) + \exp(\lambda_1)/\exp(\psi)) \\ &= (y_1 + y_2) \lambda_1 - y_2 \psi - \exp(\lambda_1)(1 + \exp(-\psi)) \\ &\quad - \log[(\exp(\lambda_1) + \exp(\lambda_1)/\exp(\psi))^{s_2} \exp(-(\exp(\lambda_1) + \exp(\lambda_1)/\exp(\psi)))] \\ &= \psi s_1 + \lambda s_2 - \exp(\lambda_1)(1 + \exp(-\psi)) - s_2 \log[\exp(\lambda_1) \\ &\quad + \frac{\exp(\lambda_1)}{\exp(\psi)}] - \exp(\lambda_1) + \exp(\lambda_1 - \psi) \\ &= \psi s_1 + \lambda s_2 - \exp(\lambda_1) - \exp(\lambda_1 - \psi) - s_2 \log \exp(\lambda_1) \\ &\quad - s_2 \log(1 + \exp(-\psi)) \exp(\lambda_1) + \exp(\lambda_1 - \psi) \\ &= \psi s_1 - s_2 \log(1 + \exp(-\psi)) \\ &= \psi s_1 - b^*(\boldsymbol{\psi}; s_2) \end{aligned}$$

Successfully conditioned out the nuisance parameter λ_1 . Now maximize this conditional likelihood.

Another Example

Multicenter randomized clinical trial. Account for differential control response rates in different centers. But assume a common odds ratio. And this common odds ratio is the parameter of interest.

Assume $i = 1, \dots, n$ centers. Data are (Y_{1i}, Y_{2i}) with

$$Y_{ji} \sim \text{Bin}(m_{ji}, \pi_{ji}), j = 1, 2$$

with $\pi_{1i} = P(Y = 1|\text{Tx})$ and $\pi_{2i} = P(Y = 1|\text{control})$ and

$$\begin{aligned} \text{logit} \pi_{1i} &= \lambda_i + \Delta \\ \text{logit} \pi_{2i} &= \lambda_i \end{aligned}$$

Note: $\Delta = \log \frac{\pi_{1i}/(1-\pi_{1i})}{\pi_{2i}/(1-\pi_{2i})}$ i.e., log odds ratio.

Problem: $n + 1$ parameters based on $2n$ binomial proportion (extreme case, $m_{ij} = 1$)

Regular mle may NOT be consistent or efficient. Use conditional likelihood here.

Set $\psi = \exp(\Delta)$. Nuisance parameters are λ_i . A sufficient statistic for λ_i is $y_{\cdot i} = \sum Y_{1i} + Y_{2i}$. Set $S_{\lambda_i} = y_{\cdot i}$ (which is independent of Δ).

Conditional likelihood is

$$\begin{aligned}
 \exp(l_c(\Delta)) &= \prod_i f_{y_i|S_{\lambda_i}}(y_i|S_{\lambda_i}, \Delta) \\
 &\propto \prod_i f(y_{1i}|S_{\lambda_i}, \Delta) \\
 &= \text{product of density of non-central hypergeometric} \\
 &\propto \prod_i \psi^{y_{1i}} / \left[\sum_{j=\max(0, S_{\lambda_i}-m_{2i})}^{\min(m_{1i}, S_{\lambda_i})} \binom{m_{1i}}{j} \binom{m_{2i}}{S_{\lambda_i}-j} \psi^j \right] \\
 &\text{take logs} \\
 l_c(\Delta) &= \sum_i \left\{ y_{1i}\Delta - \log \left[\sum_{j=\max(0, S_{\lambda_i}-m_{2i})}^{\min(m_{1i}, S_{\lambda_i})} \binom{m_{1i}}{j} \binom{m_{2i}}{S_{\lambda_i}-j} \psi^j \right] \right\}.
 \end{aligned}$$

The score statistic can be shown to be

$$\begin{aligned}
 U(\Delta) &= \left. \frac{\partial l_c}{\partial \Delta} \right|_{\Delta=0} \\
 &= \sum_i \{Y_{1i} - E(Y_{1i})\} \\
 &= \sum_i (Y_{1i} - m_{1i}y_{\cdot i}/m_{\cdot i})
 \end{aligned}$$

The numerator of the Mantel-Haenszel test (M-H test is just the score test from this conditional likelihood).