

Generalized Estimating Equations

Outline

- Review of Generalized Linear Models
 - General Linear Model, Generalized Linear Model
 - Exponential Family
 - Components of GLM
 - Types of Residuals
 - MLE for GLM, Iterative Weighted Least Squares
 - Measuring Goodness of Fit - Deviance and Pearson's χ^2
 - Over-Dispersion
- Quasi-Likelihood
 - Model Misspecification
 - Construction of Quasi-Likelihood
 - Q-L Estimating Equations
 - Optimality
 - Impact of Nuisance Parameters
- Generalized Estimating Equations

Review of Generalized Linear Models

Consider independent data (\mathbf{X}_i, y_i) , $i = 1, \dots, m$.

Linear Model

- **Random component:** normal errors.

$$Y_i | \mathbf{X}_i \sim \mathcal{N}(\mu_i, \sigma^2).$$

- **Systematic component:** linear combination of predictors.

$$\mu_i = \text{E}(Y_i | \mathbf{X}_i) = \mathbf{X}_i^T \boldsymbol{\beta}.$$

Generalized Linear Model

- **Random component:** exponential family f .

$$Y_i | \mathbf{X}_i \sim f(\theta_i, \phi),$$

where ϕ is the dispersion parameter.

- **Systematic component:** linear combination of predictors.

$$\eta_i = \mathbf{X}_i^T \boldsymbol{\beta}.$$

- **Link function:** associates the linear combination of predictors with the mean response.

$$\eta_i = g(\mu_i),$$

where $\mu_i = \text{E}(Y_i | \mathbf{X}_i)$.

Exponential Family

The density functions of the exponential family of distributions have this general form:

$$f(y; \theta, \phi) = \exp \left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}, \quad (1)$$

where θ is known as the *canonical* parameter and ϕ is a fixed (known) scale (dispersion) parameter.

Properties of Exponential Family

If $Y \sim f(y; \theta, \phi)$ in (??) then

$$\begin{aligned} E(Y) &= \mu = b'(\theta) \\ \text{Var}(Y) &= b''(\theta)a(\phi). \end{aligned}$$

Proof. The log-likelihood is

$$\begin{aligned} \ell(\theta, \phi) &= \log f(y; \theta, \phi) \\ &= \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi). \end{aligned}$$

Therefore

$$\begin{aligned} \frac{\partial \ell}{\partial \theta} &= \frac{y - b'(\theta)}{a(\phi)} \\ \frac{\partial^2 \ell}{\partial \theta^2} &= -\frac{b''(\theta)}{a(\phi)}. \end{aligned}$$

Using the fact that

$$\begin{aligned} \mathbb{E} \left(\frac{\partial l}{\partial \theta} \right) &= 0, \\ \mathbb{E} \left(\frac{\partial^2 l}{\partial \theta^2} \right) &= - \mathbb{E} \left(\frac{\partial l}{\partial \theta} \right)^2. \end{aligned}$$

We get

$$\begin{aligned} \mathbb{E} \left(\frac{y - b'(\theta)}{a(\phi)} \right) &= 0 \\ \Rightarrow \mathbb{E}(Y) &= b'(\theta) \\ \mathbb{E} \left(\frac{\partial l}{\partial \theta} \right)^2 &= \mathbb{E} \left\{ \frac{(y - b'(\theta))^2}{a^2(\phi)} \right\} \\ &= \frac{\text{Var}(Y)}{a^2(\phi)} \end{aligned}$$

So

$$\begin{aligned} \frac{\text{Var}(Y)}{a^2(\phi)} &= \frac{b''(\theta)}{a(\phi)} \\ \Rightarrow \text{Var}(Y) &= b''(\theta)a(\phi). \end{aligned}$$

Examples of Exponential Family

- Poisson: $Y =$ number of events (counts)

$$\begin{aligned} f(y; \theta, \phi) &= \frac{e^{-\lambda} \lambda^y}{y!} \\ &= \exp \{y \log \lambda - \lambda - \log(y!)\} \end{aligned}$$

so

$$\begin{aligned} \theta &= \log \lambda \\ b(\theta) &= \lambda = \exp(\theta) \\ c(y, \phi) &= -\log(y!) \\ a(\phi) &= 1 \\ \mu &= b'(\theta) = \exp(\theta) = \lambda \\ \text{Var}(Y) &= b''(\theta)a(\phi) = \exp(\theta) = \lambda \end{aligned}$$

- Binomial: $Y = s/m$, frequency of successes in m trials

$$\begin{aligned} f(y; \theta, \phi) &= \binom{m}{my} \pi^{my} (1 - \pi)^{m-my} \\ &= \exp \left\{ \frac{y \log \left(\frac{\pi}{1-\pi} \right) + \log(1 - \pi)}{1/m} + \log \binom{m}{my} \right\} \end{aligned}$$

so

$$\begin{aligned} \theta &= \log \left(\frac{\pi}{1 - \pi} \right) = \text{logit}(\pi) \\ b(\theta) &= -\log(1 - \pi) = \log[1 + \exp(\theta)] \\ c(y, \phi) &= \log \binom{m}{my} \\ a(\phi) &= \frac{1}{m} \\ \mu &= b'(\theta) = \frac{\exp(\theta)}{1 + \exp(\theta)} = \pi \\ \text{Var}(Y) &= b''(\theta)a(\phi) = ? \end{aligned}$$

- Gaussian

$$f(y; \theta, \phi) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(y - \mu)^2}{2\sigma^2} \right\}$$

$$\theta =$$

$$b(\theta) =$$

$$c(y, \phi) =$$

$$a(\phi) =$$

$$\mu =$$

$$\text{Var}(Y) =$$

Components of GLM

- **Canonical link function:** a function $g(\cdot)$ such that

$$\eta = g(\mu) = \theta$$

where θ is the canonical parameter.

- Gaussian: $g(\mu) = \mu$.
- Poisson: $g(\mu) =$
- Binomial: $g(\mu) =$

- **Variance function:** a function $V(\cdot)$ such that

$$\text{Var}(Y) = V(\mu)a(\phi).$$

Usually $a(\phi) = w\phi$ where ϕ is the scale parameter and w is a weight.

- Gaussian: $V(\mu) = 1$.
- Poisson: $V(\mu) =$
- Binomial: $V(\mu) =$

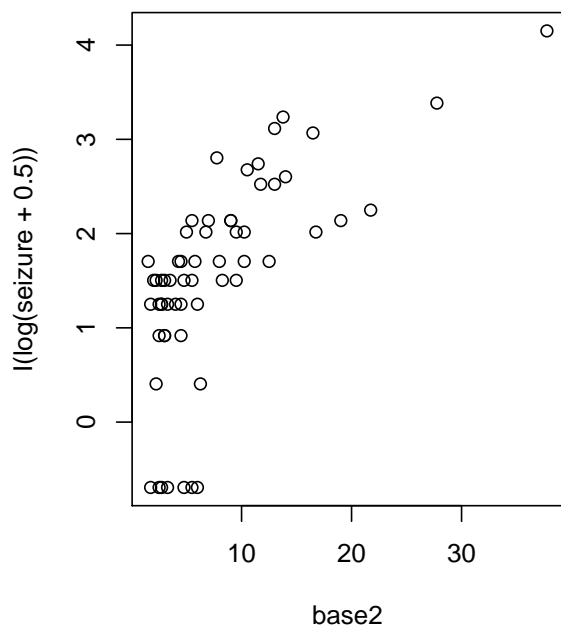
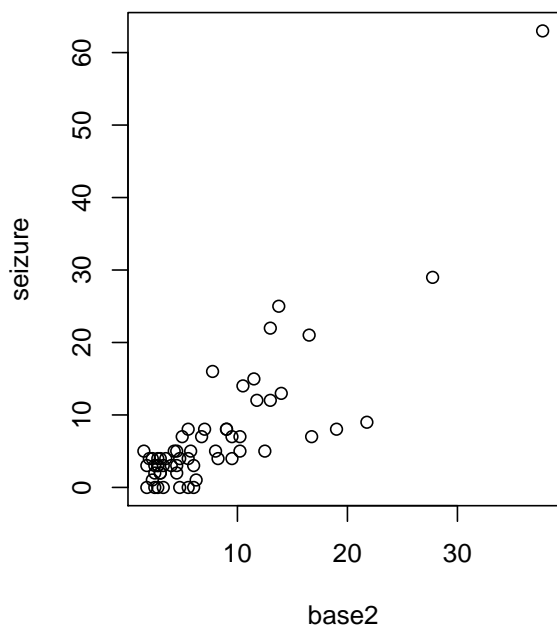
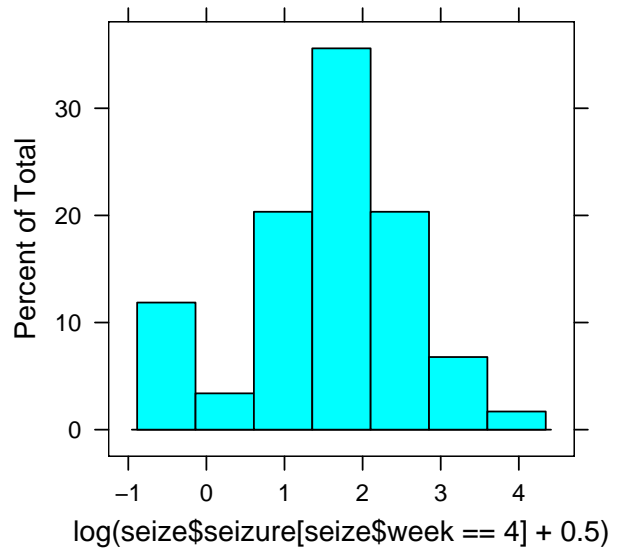
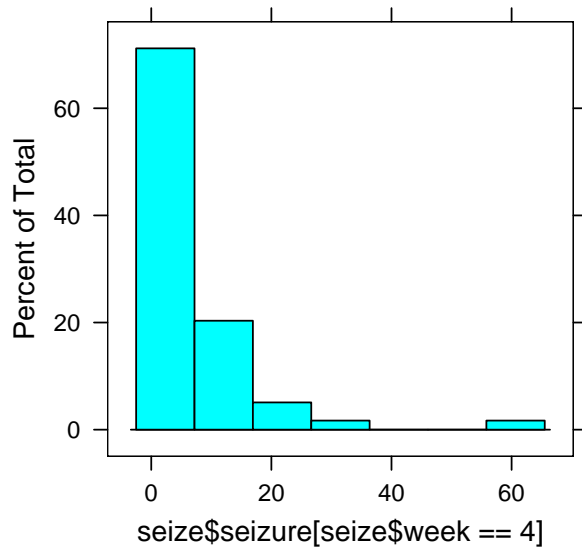
Alternative Link Functions

For binomial data,

- Logit: $g(\mu) = \log \frac{\mu}{1-\mu}$, β is the log-odds ratio.
- Probit: $g(\mu) = \Phi^{-1}(\mu)$ (threshold model: $Z = \mu + \epsilon$, $Y = 1$ if $z > k$.)
- Complementary log-log: $g(\mu) = \log(-\log \mu)$, β is the log hazard ratio.

Data Example: Seizure Data

- Using only the responses at week 4.



```

> library (lattice)
> seize <- read.table("C:/seize.data",
+                   col.names = c("id", "seizure", "week", "progabide",
+                   "baseline8", "age"))
> seize$base2 <- seize$baseline8 / 4
> seize.lm <- glm (I(log (seizure + 0.5)) ~ age + base2 + progabide,
+                 data = seize, subset = week == 4,
+                 family = gaussian)
> summary (seize.lm)

```

Call:

```

glm(formula = I(log(seizure + 0.5)) ~ age + base2 + progabide,
     family = gaussian, data = seize, subset = week == 4)

```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.9216	-0.3450	0.2560	0.5158	1.4711

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.698590	0.550308	1.269	0.2096
age	0.008016	0.016986	0.472	0.6389
base2	0.109705	0.015851	6.921	5.09e-09 ***
progabide	-0.457042	0.208729	-2.190	0.0328 *

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

(Dispersion parameter for gaussian family taken to be 0.634476)

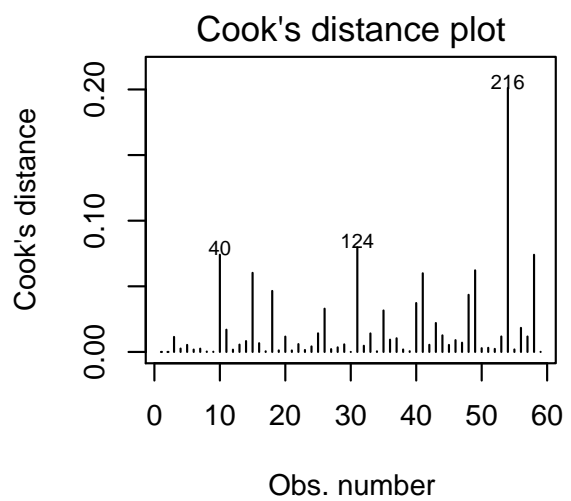
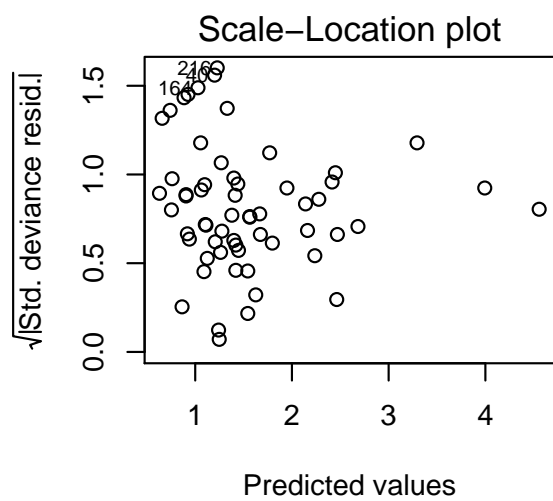
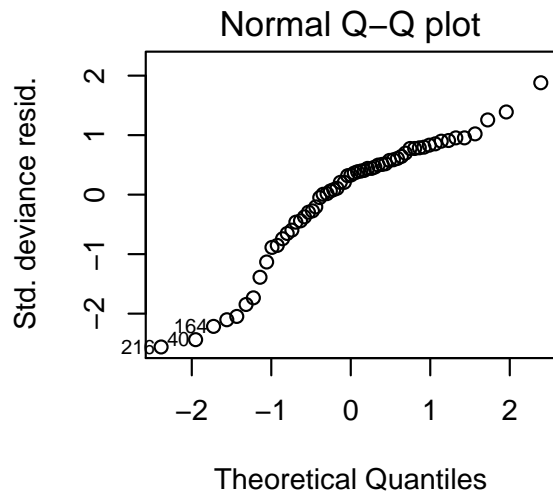
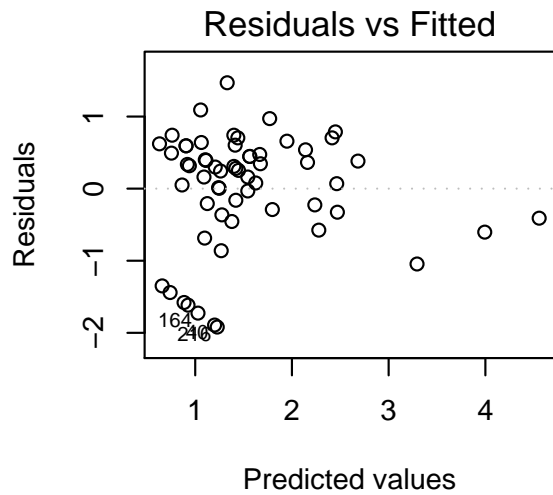
Null deviance: 68.647 on 58 degrees of freedom

Residual deviance: 34.896 on 55 degrees of freedom

AIC: 146.45

Number of Fisher Scoring iterations: 2

```
> par (mfrow = c(2, 2))
> plot (seize.lm)
```



The choice of scale for analysis is an important aspect of model selection.

- A common choice is between Y vs. $\log Y$.
- What characterizes a “good” scale? In classical linear regression analysis a good scale should combine
 - constancy of variance,
 - approximate Normality of errors, and
 - additivity of systematic effects.
- There is usually no *a priori* reason to believe that such a scale exists.
- For poisson distributed Y ,
 - $Y^{1/2}$ gives approximate constancy of variance,
 - $Y^{2/3}$ does better for approximate symmetry or Normality,
 - $\log Y$ produces additivity of the systematic effects,
 - no single scale will simultaneously produce all the desired properties.
- With the introduction of GLM, scaling problems are reduced.
 - normality and constancy of variance are no longer required,
 - however, the way in which the variance depends on the mean must be known.

```
> seize.glm <- glm (seizure ~ age + base2 + progabide,  
+                   data = seize, subset = week == 4,  
+                   family = poisson)  
> summary (seize.glm)
```

Call:

```
glm(formula = seizure ~ age + base2 + progabide, family = poisson,  
     data = seize, subset = week == 4)
```

Deviance Residuals:

	Min	1Q	Median	3Q	Max
	-3.1636	-1.0246	-0.1443	0.4865	3.8993

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.775574	0.284598	2.725	0.00643	**
age	0.014044	0.008580	1.637	0.10169	
base2	0.088228	0.004353	20.267	< 2e-16	***
progabide	-0.270482	0.101868	-2.655	0.00793	**

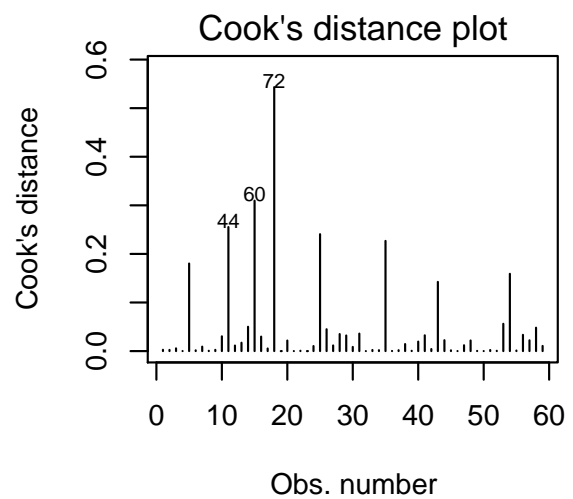
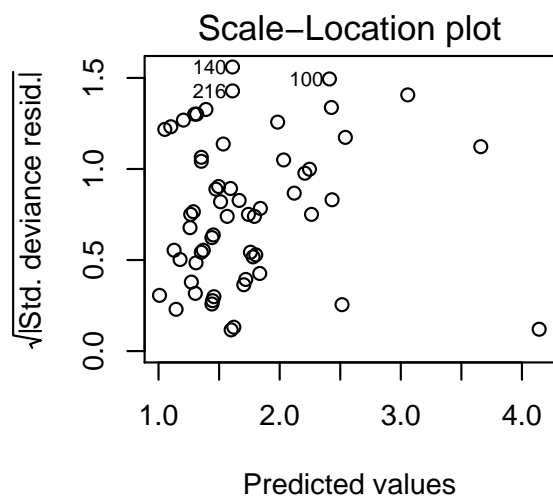
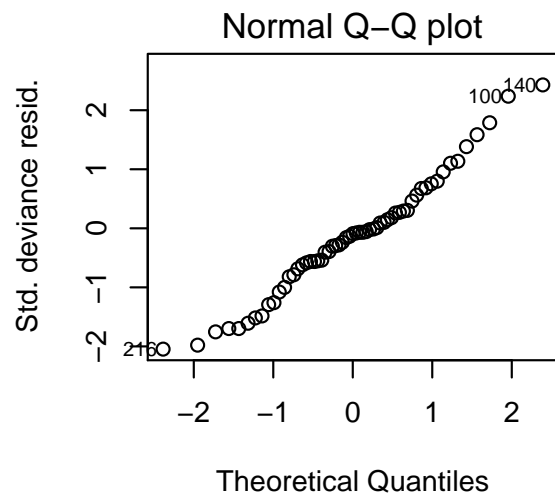
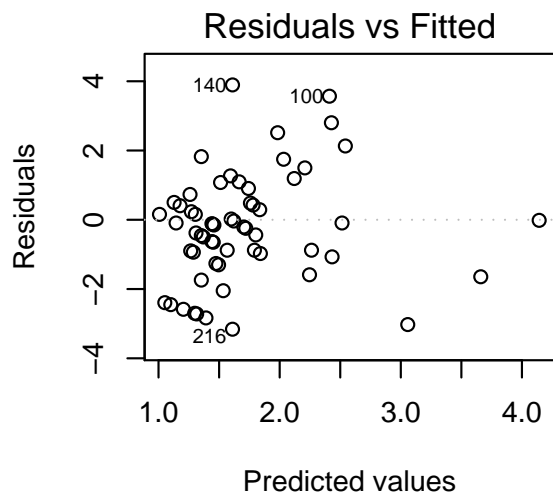
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 476.25 on 58 degrees of freedom
Residual deviance: 147.02 on 55 degrees of freedom
AIC: 342.79

Number of Fisher Scoring iterations: 5

```
> plot (seize.glm)
```



Model Diagnosis and Residuals

Like ordinary linear models, residuals can be used to assess model fit. For GLM, we require extended definitions of residuals.

Types of Residuals

- **Response residuals**

$$r_R = y - \hat{\mu}.$$

- **Pearson residuals** (standardized residuals)

$$r_P = \frac{y - \hat{\mu}}{\sqrt{V(\hat{\mu})}}.$$

- Constant variance and mean zero if the variance function is correctly specified.
- Useful for detecting variance misspecification (and autocorrelation).

- **Working residuals**

$$r_W = (y - \hat{\mu}) \cdot \left. \frac{\partial \eta}{\partial \mu} \right|_{\mu=\hat{\mu}} = Z - \hat{\eta},$$

where $Z = \hat{\eta} + (y - \hat{\mu}) \left. \frac{\partial \eta}{\partial \mu} \right|_{\mu=\hat{\mu}}$.

- **Deviance residuals**: contribution of Y_i to the deviance.

$$r_D = \text{sign}(y - \mu) \sqrt{d_i}$$

- Closer to a normal distribution (less skewed) than Pearson residuals.
 - Often better for spotting outliers.
- For more details in residuals in GLM, see McCullagh and Nelder (2nd Edition, Section 2.4) and Pierce and Schafer (JASA 1986).

Maximum Likelihood Estimation for GLMs

Solve score equations, for $j = 1, \dots, p$,

$$S_j(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \beta_j} = 0.$$

The log-likelihood:

$$\begin{aligned} \ell &= \sum_{i=1}^m \left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\} = \sum_i \ell_i \\ S_j(\boldsymbol{\beta}) &= \frac{\partial \ell}{\partial \beta_j} = \sum_i \frac{\partial \ell_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\ \frac{\partial \ell_i}{\partial \theta_i} &= \frac{1}{a_i(\phi)} (y_i - b'(\theta_i)) = \frac{1}{a_i(\phi)} (y_i - \mu_i) \\ \frac{\partial \theta_i}{\partial \mu_i} &= \left(\frac{\partial \mu_i}{\partial \theta_i} \right)^{-1} = \frac{1}{b''(\theta_i)} = \frac{1}{V(\mu_i)} \\ \frac{\partial \mu_i}{\partial \eta_i} &= \frac{1}{g'(\mu_i)} \\ \frac{\partial \eta_i}{\partial \beta_j} &= X_{ij} \end{aligned}$$

Therefore

$$S_j(\boldsymbol{\beta}) = \sum_{i=1}^m \frac{X_{ij}}{g'(\mu_i)} [a_i(\phi) V(\mu_i)]^{-1} (y_i - \mu_i). \quad (2)$$

- For fixed ϕ , the score function depends on μ_i and V_i only
- $\left(\frac{\partial \mu_i}{\partial \beta_j} \right) = \frac{X_{ij}}{g'(\mu_i)}$: Jacobian matrix.
- Weight $y_i - \mu_i$ by V_i^{-1} .

Fisher's Information

Write (??) in matrix form

$$S(\boldsymbol{\beta}) = \sum_{i=1}^m \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)^T [a_i(\phi)V(\mu_i)]^{-1} (y_i - \mu_i).$$

Hence,

$$\mathcal{I}(\boldsymbol{\beta}) = -\text{E} \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_i \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)^T [a_i(\phi)V(\mu_i)]^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right).$$

The observed counterpart is

$$-\partial S(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = \mathcal{I}(\boldsymbol{\beta}) - \sum_{i=1}^m \frac{\partial A_i}{\partial \boldsymbol{\beta}} (y_i - \mu_i(\boldsymbol{\beta})),$$

where $A_i = \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)^T [a_i(\phi)V(\mu_i)]^{-1}$. **For canonical links**, the observed one equals the expected one (exercise).

Moreover (Cox and Reid, 1987),

$$\mathcal{I}(\boldsymbol{\beta}, \phi) = \text{E} \left\{ -\frac{\partial^2 S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \phi} \right\} = 0.$$

The information matrix is of the form

$$\begin{pmatrix} \mathcal{I}(\boldsymbol{\beta}) & 0 \\ 0 & \mathcal{I}(\phi) \end{pmatrix}.$$

The MLEs $\hat{\boldsymbol{\beta}}$ and $\hat{\phi}$ are asymptotically independent, $\mathcal{I}^{-1}(\boldsymbol{\beta})$ is the asymptotic variance of $\hat{\boldsymbol{\beta}}$ and $\mathcal{I}^{-1}(\phi)$ is the asymptotic variance of $\hat{\phi}$.

Iterative Weighted Least Squares

The MLE of β can be obtained by iterative weighted least squares (IWLS).

- When $g(\mu) = \mu = X\beta$, (??) immediately suggests an IWLS algorithm for solving the score equation:

1. For given $\hat{\beta}$, calculate the weights

$$w_i = V(\hat{\beta})^{-1}.$$

2. Solve

$$\sum_i X_i^T w_i (y_i - X_i \beta) = 0.$$

3. Update W .

- (For fixed ϕ) When g is non-linear, the IWLS algorithm needs to be modified by constructing a *working response*

$$Z = \hat{\eta} + (Y - \hat{\mu}) \left. \frac{\partial \eta}{\partial \mu} \right|_{\mu=\hat{\mu}}$$

and modifying the weights to account for the rescaling from Y to Z

$$w_i = \frac{1}{V(\hat{\mu})} \frac{1}{g'(\hat{\mu})^2}.$$

The score equation becomes

$$\sum_i X_i^T w_i (z_i - X_i \beta) = 0,$$

where $\eta_i = X_i \beta$. This has the same form as (??) if the $\hat{\mu}$ in w_i is replaced by μ (exercise).

- The IWLS algorithm can be justified as an application of the Fisher scoring method. (See McCullagh and Nelder, 2nd edition, pages 41-43.)

Fisher Scoring

To solve the score equations

$$S(\boldsymbol{\beta}) = 0,$$

iterative method is required for most GLMs. The Newton-Raphson algorithm uses the observed derivative of the core (gradient) and Fisher scoring method uses the expected derivative of the score (i.e., Fisher's information matrix, $-\mathcal{I}_n$)

The algorithm:

1. Find an initial value $\hat{\boldsymbol{\beta}}^{(0)}$.

2. For $j \rightarrow j + 1$ update $\hat{\boldsymbol{\beta}}^{(j)}$ via

$$\hat{\boldsymbol{\beta}}^{(j+1)} = \hat{\boldsymbol{\beta}}^{(j)} + (\hat{\mathcal{I}}_n^{(j)})^{-1} S(\hat{\boldsymbol{\beta}}^{(j)}).$$

3. Evaluate convergence using changes in $\log \mathcal{L}$ or $\|\hat{\boldsymbol{\beta}}^{(j+1)} - \hat{\boldsymbol{\beta}}^{(j)}\|$.

4. Iterate until convergence criterion is satisfied.

Measuring Goodness of Fit - Deviance and Pearson's X^2

Deviance

Deviance is a quantity to measure how well the model fits the data.

- For μ_i , two approaches to estimate μ_i
 - from the fitted model: $\mu_i(\hat{\beta})$, the likelihood of current model.
 - from the full (saturated) model: y_i , the maximum achievable likelihood.
- One can compare $\mu_i(\hat{\beta})$ with y_i through the likelihood function.
 - Express the likelihood as a function of μ_i 's and ϕ

$$\mathcal{L}(\mu, \phi) = \prod_{i=1}^m L_i = \prod_{i=1}^m f(y_i; \mu_i, \phi)$$

- The deviance of the fitted model is defined as

$$D(\hat{\mu}; y) = -2 \sum_{i=1}^m \{ \log L_i(\hat{\mu}; \phi) - \log L_i(y, \phi) \} a(\phi).$$

- Deviance examples:
 - Normal:

$$\log f(y_i; \theta_i, \phi) = \frac{(y_i - \mu_i)^2}{2\phi},$$

$$D(\mathbf{y}, \hat{\mu}) = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 = \text{SSE}.$$

The sum of residual squares! Note: except for normal responses, the distribution of D , even asymptotically, is unknown.

- Poisson:

$$\log f(y_i; \theta_i, \phi) = y_i \log(\mu) - \mu$$

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

It is sometimes referred to as the G statistic. Note the second term can be omitted as its sum is 0.

– Binomial:

$$\log f(\mathbf{y}_i; \theta_i, \phi) = m_i \{y_i \log \mu + (1 - y_i) \log(1 - \mu)\},$$

$$D(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^n \left\{ m_i y_i \log \left(\frac{y_i}{\hat{\mu}_i} \right) - m_i (1 - y_i) \log \left(\frac{1 - y_i}{1 - \hat{\mu}_i} \right) \right\}. \quad (3)$$

- The deviance is the sum of squared deviance residuals. $D = \sum_{i=1}^m r_{Di}^2$.

Pearson's X^2

- Another measure of discrepancy is the generalized Pearson's X^2 statistic

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

Note that it is the sum of the squared Pearson's residuals.

- Pearson's X^2 examples:
 - Normal: $X^2 = \text{SSE}$.
 - Poisson: $X^2 = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i$
 - Binomial: $X^2 = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 / [\hat{\mu}_i(1 - \hat{\mu}_i)]$.
- For normal-theory linear models, when the model is correct, both deviance and X^2 have exact χ^2 distribution. For other models both have (approximate) asymptotic χ^2 distribution (but the approximation may not be very good even when n is very large).
- The deviance has a general advantage as a measure of discrepancy in that it is additive when comparing nested models if ML estimates are used, while the generalized Pearson's X^2 is sometimes preferred for easy interpretation.

Overdispersion

- For Poisson regression, it is expected that $\text{Var}(Y_i) = \mu_i$. However this can be sometimes violated.
- **Overdispersion** describes the situation that the data are overdispersed when the actually $\text{Var}(Y_i)$ exceeds the GLM variance $\phi V(\mu)$.
- For Binomial and Poisson models we often find overdispersion:
 - Binomial: $Y = s/m$, $E(Y) = \mu$, $\text{Var}(Y) > \mu(1 - \mu)/m$.
 - Poisson: $E(Y) = \mu$, $\text{Var}(Y) > \mu$.

How Does Overdispersion Arise?

- If there is population *heterogeneity*, say, clustering in the population, then overdispersion can be introduced.

Suppose there exists a binary covariate, Z_i and that

$$\begin{aligned} Y_i | Z_i = 0 &\sim \text{Poisson}(\lambda_0) \\ Y_i | Z_i = 1 &\sim \text{Poisson}(\lambda_1) \\ \Pr(Z_i = 1) &= \pi \end{aligned}$$

Then

$$\begin{aligned} E(Y_i) &= \pi\lambda_1 + (1 - \pi)\lambda_0 = \mu \\ \text{Var}(Y_i) &= E(\text{Var}(Y_i | Z_i)) + \text{Var}(E(Y_i | Z_i)) \\ &= E(Y_i) + \text{Var}(\lambda_1 Z_i + \lambda_0(1 - Z_i)) \\ &= \mu + (\lambda_1 - \lambda_0)^2 \pi(1 - \pi) \end{aligned}$$

Therefore, if we do not observe Z_i (e.g. latent variable) then the omitted factor leads to increased variation.

Quasi-Likelihood

Motivation - Impact of Model Misspecification

Huber (1967) and White (1982) studied the properties of MLEs when the model is misspecified.

Setup

- Let F_θ be the *assumed* distribution family for independent data $Y_i, i = 1, \dots, n$.
- Let $\hat{\theta}_n$ be the MLE (based on n observation). That is, $\hat{\theta}_n$ solves the score equations that arise from the assumed F_θ :

$$\sum_{i=1}^n S_i^F(\hat{\theta}_n) = 0.$$

- However the true distribution of Y_i is given by $Y_i \sim G$.

Result

- $\hat{\theta}_n \longrightarrow \theta^*$ such that

$$E_G \left[\sum_{i=1}^n S_i^F(\theta^*) \right] = 0.$$

- The estimator $\hat{\theta}_n$ is asymptotically normal:

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \longrightarrow \mathcal{N}(0, A^{-1}BA^{-1})$$

where

$$A = -\lim \frac{1}{n} \sum_{i=1}^n E_G \left[\frac{\partial}{\partial \theta} S_i^F(\theta) \Big|_{\theta^*} \right]$$

$$B = \lim \frac{1}{n} \sum_{i=1}^n \text{Var}_G [S_i^F(\theta) |_{\theta^*}]$$

$$= \lim \frac{1}{n} \sum_{i=1}^n E_G [S_i^F(\theta) |_{\theta^*}]^2$$

- A is the expected value of the observed (based on the assumed model) information (times $1/n$).
- B is the true variance of $S_i^F(\theta)$ which may no longer be equal to minus the expected (under the true model) derivative of $S_i^F(\theta)$ if the assumed model is not true.
- In general $\hat{\theta}$ is not consistent to θ_0 . But sometimes we get lucky and $\theta^* = \theta_0$ — the model misspecification does not hurt the consistency of $\hat{\theta}_n$.
- Sometimes we get even luckier and $\theta^* = \theta_0$ and $A = B$. The model misspecification does not hurt our standard error estimates either.
- For GLM where we are modeling the mean $E(Y_i) = \mu_i$ via a regression model with parameters β , our estimator, $\hat{\beta}$, will converge to whatever value solves

$$E_G[S(\beta)] = 0.$$

Recall that we have

$$S(\beta) = \sum_{i=1}^m \left(\frac{\partial \mu_i}{\partial \beta} \right)^T [a_i(\phi)V(\mu_i)]^{-1}(y_i - \mu_i).$$

As long as $Y_i \sim G$ such that $E_G(Y_i) = \mu$ then our estimator will be consistent! We do not need Poisson, or Binomial distribution for the GLM point estimate $\hat{\beta}$.

Construction of Quasi-Likelihood

McCullagh and Nelder, 1989, Chapter 9.

It is often possible to characterize the first two moments of the response variable with unknown distribution:

$$\begin{aligned} E(Y_i) &= \mu_i(\boldsymbol{\beta}) \\ \text{Var}(Y_i) &= \sigma^2 V(\mu_i) \end{aligned}$$

where σ^2 is unknown, V has known functional form.

The function

$$U(\mu; y) = \frac{Y - \mu}{\sigma^2 V(\mu)}$$

has the following properties in common with a score function (derivative of a log-likelihood function):

$$\begin{aligned} E(U) &= 0, \\ \text{Var}(U) &= \frac{1}{\sigma^2 V(\mu)}, \\ -E\left(\frac{\partial U}{\partial \mu}\right) &= \frac{1}{\sigma^2 V(\mu)}. \end{aligned}$$

Therefore the integral

$$Q(\mu; y) = \int_y^\mu \frac{y - t}{\sigma^2 V(t)} dt,$$

if exists, should behave like a log-likelihood function for μ and is referred as the **quasi-likelihood**, or the **log quasi-likelihood** for μ based on data y .

- Find the quasi-likelihoods for $V(\mu) = 1$ and $V(\mu) = \mu$. See Table 9.1 in McCullagh and Nelder for other variance functions. Note that not all possible choices of V lead to an authentic log likelihood function.

- It can be shown (Wedderburn, 1974) that the log likelihood function is identical to the quasi-likelihood if and only if it belongs to the exponential family.
- **Theorem** (Wedderburn, 1974): For one observation of y , the log likelihood function ℓ has the property

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

where $\mu = E(Y)$ and $V(\mu) = \text{Var}(Y)$, *if and only if* the density of Y can be written in the form

$$\exp \{y\theta - g(\theta)\},$$

where θ is some function of μ .

Assuming a one-parameter exponential family for z is equivalent to making no assumption other than the mean-variance relationship.

Quasi-Likelihood Estimating Equations

The quasi-likelihood regression parameter, $\hat{\beta}$ for Y_i , $i = 1, \dots, n$ is obtained as the solution to the **quasi-score equations**, $U(\hat{\beta}) = 0$, where

$$U(\beta) = D^T V^{-1}(Y - \mu)$$

$$D_{ij} = \frac{\partial \mu_i}{\partial \beta_j}$$

$$V = \text{diag}(\sigma^2 V(\mu_i))$$

Properties

- The covariance matrix of $U(\beta)$ plays the same role as Fisher information in the asymptotic variance of $\hat{\beta}$:

$$\mathcal{I}_n = D^T V^{-1} D,$$

$$\text{Var}(\hat{\beta}) \approx \mathcal{I}_n^{-1}.$$

- These properties are based **only** on the correct specification of the *mean* and *variance* of Y_i .
- Note that for the estimation of σ^2 , the quasi-likelihood does not behave like a log likelihood. Method of moments is used to estimate σ^2

$$\begin{aligned} \tilde{\sigma}^2 &= \frac{1}{n-p} \sum_{i=1}^n \frac{(Y_i - \hat{\mu}_i)^2}{V_i(\hat{\mu}_i)} \\ &= \frac{\chi^2}{n-p}, \end{aligned}$$

where χ^2 is the generalized Pearson statistics.

Example: Seizure Data

In R, by specifying `family = quasi (link = log, variance = "mu")` or `family = quasipoisson`, `glm` will give the same results.

```
> seize.glm2 <- glm (seizure ~ age + base2 + progabide,
+                   data = seize, subset = week == 4,
+                   family = quasi (link = log, variance = "mu"))
> summary (seize.glm2)
```

Call:

```
glm(formula = seizure ~ age + base2 + progabide,
     family = quasi(link = log, variance = "mu"),
     data = seize, subset = week == 4)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.1636	-1.0246	-0.1443	0.4865	3.8993

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.775574	0.448580	1.729	0.0894 .
age	0.014044	0.013524	1.038	0.3036
base2	0.088228	0.006862	12.858	<2e-16 ***
progabide	-0.270482	0.160563	-1.685	0.0977 .

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

(Dispersion parameter for quasi family taken to be 2.484377)

Null deviance: 476.25 on 58 degrees of freedom
 Residual deviance: 147.02 on 55 degrees of freedom
 AIC: NA

Number of Fisher Scoring iterations: 5

- The standard error estimates for the regression parameters are smaller.
- Note that AIC is no longer available for quasi-likelihood.

Review of Estimating Functions

- Note: quasi-likelihood is also used more generally to refer to estimating functions (Hedyde, 1997) but we use it in a narrower sense in GLM with variance function being

$$\text{Var}(Y) = a(\phi)V(\mu).$$

- We treat quasi-score equations as a special case of estimating equations. The previous variance of Y is a special case of

$$\text{Var}(Y) = V(\mu, \phi).$$

- An **estimating function** is a function of data and parameter, $g(Y, \theta)$, such that an estimator $\hat{\theta}$ of θ is obtained as its root, that is $g(Y, \hat{\theta}) = 0$.
- An **unbiased estimating function (UEF)** has the property

$$E_{\theta}[g(Y, \theta)] = 0.$$

for any $\theta \in \Theta$. Role of unbiasedness: under regularity conditions, unbiased equations have roots which are consistent estimators.

- Estimating functions form the basis of (almost?) all of frequentist statistical estimation.
 - Method of least squares (LS) (Gauss and Legendre): finite sample consideration

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0.$$

- Maximum likelihood (ML) (Fisher): asymptotic property

$$\sum_i \frac{\partial}{\partial \theta} \log f(y_i; \theta) = 0.$$

- The well-known Mantel-Haenszel estimator for odds ratio (See Breslow, 1981 for properties).
- Method of moments (K. Pearson).

Optimality

- For linear models, Gauss-Markov theorem says that the LS estimate is the linear unbiased minimal variance (UMV) estimate for β , for fixed (finite) sample size.
- We know that the MLE is asymptotically unbiased and efficient (has minimal asymptotic variance among asymptotically unbiased estimators).
- Consider a class of *unbiased* estimating functions,

$$\mathcal{G} = \{g(y; \theta) : E_{\theta}[g(y; \theta)] = 0\}.$$

Godambe (1960) defined $g^* \in \mathcal{G}$ as an *optimal estimating function* among \mathcal{G} if it minimizes

$$W = \frac{E[g(y, \theta)^2]}{[E(\partial g / \partial \theta)]^2} = E\left[\frac{g(y, \theta)}{E(\partial g / \partial \theta)}\right]^2. \quad (4)$$

- The numerator is the variance $\text{Var}(g)$
 - The denominator: square of the averaged gradient of g .
 - We want the optimal g has small variance and on average as steep as possible near the true θ , which are related to the asymptotic variance of $\hat{\theta}$.
 - This is a finite sample criterion!
 - **Standardization:** $g(y, \theta) / E(\partial g / \partial \theta)$.
- Godambe (1960) showed that the score functions (even non-linear ones) for θ ,

$$\dot{\ell}(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}$$

where $\ell(\theta)$ is the log-likelihood function, are optimal estimating functions. Here

$$W^* = \frac{1}{E[\ddot{\ell}(\theta)]}$$

(“Cramer-Rao lower bound”)

- Godambe and Heyde (1987) proved that **quasi-score function**,

$$\sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \beta} \right)^T v_i^{-1} (y_i - \mu_i(\beta)),$$

where $v_i = \text{Var}(y_i) = a(\phi)V(\mu_i)$ is optimal among *unbiased* estimation functions **which are linear in the data**, that is, take the form

$$\sum_{i=1}^n a_i(\beta)(y_i - \mu_i(\beta)). \tag{5}$$

Proof. Here is a sketch of the proof for the scalar case (Liang and Zeger, 1995).

For an unbiased estimating function of the form (??, The optimality criterion (??) reduces to

$$W_n = \frac{\sum_{i=1}^n a_i^2 v_i}{\left\{ \sum_{i=1}^n a_i \frac{\partial \mu_i}{\partial \beta} \right\}^2} = \frac{\sum_{i=1}^n (a_i \sqrt{v_i})^2}{\left\{ \sum_{i=1}^n (a_i \sqrt{v_i}) \left(\frac{\partial \mu_i}{\partial \beta} \frac{1}{\sqrt{v_i}} \right) \right\}^2}$$

For the quasi-score function,

$$a^*(\beta) = \left(\frac{\partial \mu_i}{\partial \beta} \right) v_i^{-1},$$

hence

$$\begin{aligned} W_n^* &= \frac{\sum_{i=1}^n \left[\left(\frac{\partial \mu_i}{\partial \beta} \right) v_i^{-1} \sqrt{v_i} \right]^2}{\left\{ \sum_{i=1}^n \left[\left(\frac{\partial \mu_i}{\partial \beta} \right) v_i^{-1} \frac{1}{\sqrt{v_i}} \right] \left(\frac{\partial \mu_i}{\partial \beta} \sqrt{v_i} \right) \right\}^2} \\ &= \frac{\sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \beta} \right)^2 \frac{1}{v_i}}{\left\{ \sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \beta} \right)^2 \frac{1}{v_i} \right\}^2} \\ &= \frac{1}{\sum_{i=1}^n \left(\frac{\partial \mu_i}{\partial \beta} \frac{1}{\sqrt{v_i}} \right)^2} \end{aligned}$$

Using Cauchy-Schwarz's inequality

$$(\sum_i x_i y_i)^2 \leq (\sum_i x_i^2)(\sum_i y_i^2)$$

it follows immediately that

$$W_n^* < W_n$$

for any choice of $a_i(\beta)$.

- The best unbiased linear estimating functions are not necessarily very good — there could be better estimating functions that aren't linear.
- When only the mean model is known, only the linear estimating function can be guaranteed to be unbiased.
- Very often it is easier to verify the unbiasedness of g_i through defining a statistic \mathbf{A}_i such that

$$E(g_i | \mathbf{A}_i) = 0.$$

One advantage of the conditional unbiasedness is that we may consider a broader class of UEFs in which the weight associated with g_i can be a function of \mathbf{A}_i ,

$$\sum_{i=1}^n a_i(\theta, \mathbf{A}_i) g_i.$$

Follow the proof for quasi-score function, the optimal linear combination is

$$g = \sum_{i=1}^n E \left(\frac{\partial g_i}{\partial \theta} | A_i \right)^T \text{Var}(g_i | A_i)^{-1} g_i. \quad (6)$$

Nuisance Parameter and Estimating Functions

When there is a nuisance parameter ϕ , i.e., the likelihood is $f(y; \theta, \phi)$, if the dimension of ϕ increases with the sample size n , the MLE for θ may not even be consistent.

- Godambe (1976) considered a complete and sufficient statistic T for ϕ for fixed θ , and showed conditional score function

$$\frac{\partial \log f(y | T = t; \theta)}{\partial \theta}$$

is the optimal estimating function for θ .

- The conditional score function requires the existence of T , a complete and sufficient statistic for ϕ that does not depend on θ . Such a statistic can be found for exponential family distributions, but more generally $t = t(\theta)$. In the later case, $\partial/\partial\theta[\log f(y | T = t; \theta)]$ depends on ϕ and hence is only locally optimal at the true ϕ (Lindsay, 1982). Quasi-likelihood can also suffer this limitation.
- If $\text{Var}(Y) = V(\mu, \phi) \neq a(\phi)V(\mu)$, the quasi-score function is no longer optimal.
- Liang and Zeger (1995) considered how to construct estimating functions for parameters of interest in the presence of nuisance parameter and the absence of fully specified likelihood.

Generalized Estimating Equations

- The data $\mathbf{y} = (\mathbf{y}_1, \dots, \mathbf{y}_m)$ is decomposed into m “strata” and the \mathbf{y}_i ’s are uncorrelated with each other. The dimensions of \mathbf{y}_i ’s are not required to be the same.
- Assuming the parameter, θ , is common to all m strata and the existence of an unbiased estimating function, $g_i(\mathbf{y}_i; \theta)$, for each of the m strata, i.e.,

$$E(g_i; \theta, \phi) = 0 \quad \forall \theta, \phi, i.$$

- In a regression setting

$$E(\mathbf{y}_i) = \boldsymbol{\mu}_i(\boldsymbol{\beta}),$$

where \mathbf{y}_i is an $n_i \times 1$ vector of responses, we can use

$$\mathbf{g}_i = \mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})$$

and it leads to the optimal \mathbf{g} among (??), namely

$$\mathbf{g} = \sum_{i=1}^m \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^T \text{Var}(\mathbf{y}_i)^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})). \quad (7)$$

This is referred to as the **generalized estimating equation** (GEE1).

- Note that the dimension of \mathbf{g}_i varies from stratum to stratum and when $n_i = 1$ for all i , \mathbf{g} reduces to the quasi-score function.
- It is a special case of (??) and (??).
- Quasi-score function is for independent, over-dispersed data (Poisson or binomial) while GEE1 is for correlated data.

Nuisance Parameter: Hello? I am Still Here.

- Even though we choose \mathbf{g}_i that does not include ϕ in its functional form, in general the distribution of \mathbf{g}_i depends on ϕ .
- Liang and Zeger (1995) argued that the impact of the nuisance parameters on \mathbf{g} and on the corresponding solution of $\mathbf{g} = 0$ is small, because it shares the orthogonality properties enjoyed by the conditional score function.

1. $E(\mathbf{g}(\theta, \phi^*); \theta, \phi) = 0$ for all θ , ϕ , and ϕ^* where ϕ^* is an incorrect value (estimate) for ϕ .

This property implies that $E(\mathbf{g}(\theta, \hat{\phi}_\theta); \theta, \phi) = 0$ for $\hat{\phi}_\theta$, the MLE of ϕ for fixed θ .

2. $E(\partial \mathbf{g}(\theta, \phi^*) / \partial \phi^*; \theta, \phi) = 0$ for all θ , ϕ , and ϕ^* .

3. $\text{Cov}(\mathbf{g}(\theta, \phi), \partial \log f(\mathbf{y}; \theta, \phi) / \partial \phi) = 0$ for all θ and ϕ .

- Implications:

– when a \sqrt{m} -consistent estimator $\hat{\phi}_\theta$ for ϕ is used, the asymptotic variance of $\hat{\theta}$ (solution to $\mathbf{g}(\theta, \hat{\phi}_\theta) = 0$) is the same as if the true value of ϕ is known. Hence, the choice among \sqrt{m} -consistent estimators is irrelevant, at least **when m is large**.

– the bias of $\mathbf{g}(\theta, \hat{\phi}_\theta)$ with $\hat{\phi}_\theta$ plugged into the EF is diminished at a faster rate than that of $\mathbf{S}_\theta(\theta, \hat{\phi}_\theta)$, the ordinary score function evaluated at $\hat{\phi}_\theta$.

– **robust** - even if the assumption on how ϕ describes the distribution of the \mathbf{y} 's is misspecified, the solution remains consistent and its asymptotic variance is unaltered.

Further Reading

- Chapter 2 of McCullagh and Nelder, 2nd edition.
- Chapter 10 of Fitzmaurice et al (2004).

References

- Breslow NE (1981). Odds ratio estimators when the data are sparse. *Biometrika* **68**:73-84.
- Cox DR and Reid N (1987) Parameter orthogonality and approximate conditional inference, *JRSSB* **49**:1-39.
- Crowder M (1986) On the consistency and inconsistency of estimating equations, *Econometric Theory* **2**:305-30.
- Fahrmeri L and Kaufmann H (1985) Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Annals of Statistics* **13**:342-68.
- Godambe VP (1960) An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics* **27**:357-72.
- - (1976) Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63**:277-84.
- Godambe VP and Heyde CC (1987) Quasi-likelihood an optimal estimation. *International Statistical Review* **55**:231-4.
- Heyde CC (1997) Quasi-likelihood and its applications. Springer-Verlag.
- Liang KY and Zeger SL (1995) Inference based on estimating functions in the presence of nuisance parameters (with discussion). *Statistical Science* **10**:158-73.
- Lindsay B (1982). Conditional score functions: some optimality results. *Biometrika* **69**:503-12.
- Wedderburn RWM (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**:439-47.