Generalized Estimating Equations

Outline

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

Review of Generalized Linear Models (GLM)

Consider independent data Y_i, $i = 1, \ldots, m$ with the covariates of $\boldsymbol{X}_i$. In GLM, the probability model for Y_i has the following specification:

- **Random component**: $\boldsymbol{Y}_i$ is assumed to follow distribution that belongs to the exponential family.

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

  where ϕ is the dispersion parameter.

- **Systematic component**: given covariates $\boldsymbol{X}_i$, the mean of $\boldsymbol{Y}_i$ can be expressed in terms of the following linear combination of predictors.

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

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

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

  where $\mu_i = \mathrm{E}(Y_i \,|\, \boldsymbol{X}_i)$.

Exponential Family

In the random component of GLM, $\boldsymbol{Y}_i$ is assumed to follow a probability distribution that belongs to the exponential family.

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

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

where θ is called the *canonical* parameter and ϕ the scale (dispersion) parameter.

Note that $a(\cdot)$ and $b(\cdot)$ are some specific functions that distinguish one member of the exponential family from another. If ϕ is known, this is an exponential family model with only canonical parameter of θ.

The exponential family of distribution include the normal, Bernoulli, and Poisson distributions.

Properties of Exponential Family

If $Y \sim f(y; \theta, \phi)$ in (1) then

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

$< Proof >$

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

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

$< Proof(cont.) >$ Using the fact that

$$\mathrm{E}\left(\frac{\partial l}{\partial \theta}\right) = 0,$$
$$\mathrm{E}\left(\frac{\partial^2 \ell}{\partial \theta^2}\right) = -\mathrm{E}\left(\frac{\partial l}{\partial \theta}\right)^2,$$

we get

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

hence

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

Examples of Exponential Family

- Gaussian

$$
\begin{aligned}
f(y;\theta,\phi) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} \\
&= \exp\left\{\frac{y\mu-\mu^2/2}{\sigma^2} - \frac{1}{2}(y^2/\sigma^2+\log(2\pi\sigma^2))\right\}
\end{aligned}
$$

so

$$
\begin{aligned}
\theta &= \mu \\
b(\theta) &= \theta^2/2 \\
c(y,\phi) &= -\frac{1}{2}(y^2/\sigma^2+\log(2\pi\sigma^2)) \\
a(\phi) &= \phi = \sigma^2
\end{aligned}
$$

then

$$
\begin{aligned}
\mu &= b'(\theta) = \theta \\
\mathrm{Var}(Y) &= b''(\theta)a(\phi) = \sigma^2
\end{aligned}
$$

- 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

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

then

$$
\begin{aligned}
\mu &= b'(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} = \pi \\
\mathrm{Var}(Y) &= b''(\theta)a(\phi) = \pi(1-\pi)/m
\end{aligned}
$$

- Poisson: Y = number of events (counts)

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

so

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

then

$$
\begin{aligned}
\mu &= b'(\theta) = \exp(\theta) = \lambda \\
\mathrm{Var}(Y) &= b''(\theta)a(\phi) = \exp(\theta) = \lambda
\end{aligned}
$$

Components of GLM

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

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

  where θ is the canonical parameter.

  – Gaussian: $g(\mu) = \mu$.

  – Binomial: $g(\mu) = \text{logit}(\mu), \mu = \pi$.

  – Poisson: $g(\mu) = \log(\mu), \mu = \lambda$.

- **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: $V(\mu) = 1$.

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

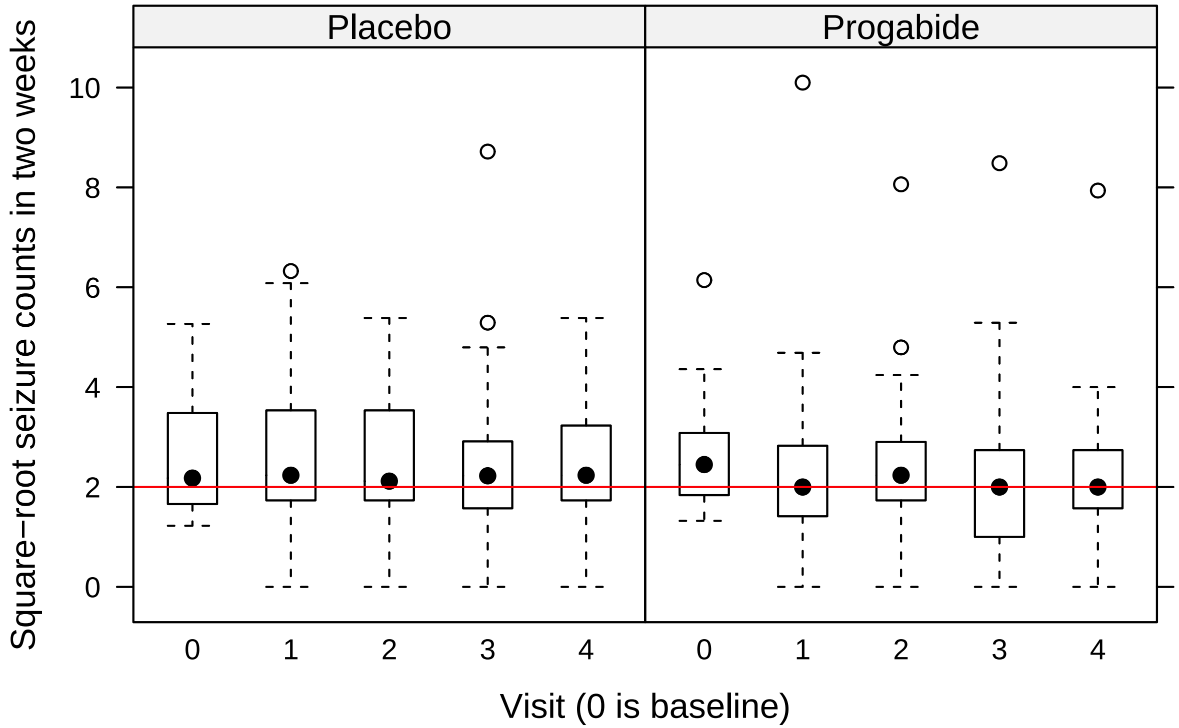
  – Poisson: $V(\mu) = \mu$.

Alternative Link Functions

For binomial data,

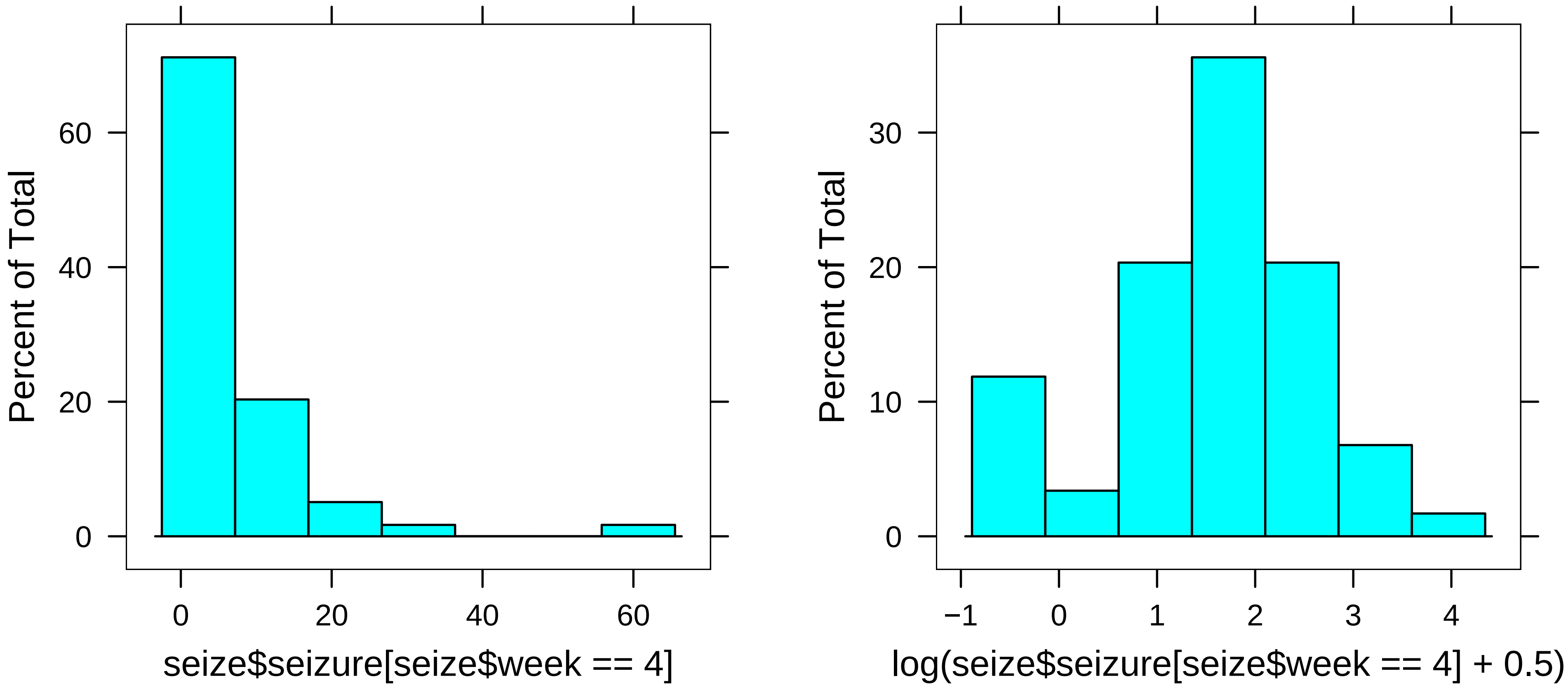
- Logit: $g(\mu) = \log \frac{\mu}{1-\mu}$, β is the log-odds ratio.
- Probit: $g(\mu) = \Phi^{-1}(\mu)$.
- Complementary log-log: $g(\mu) = \log(-\log(1 - \mu))$, β is the log hazard ratio.

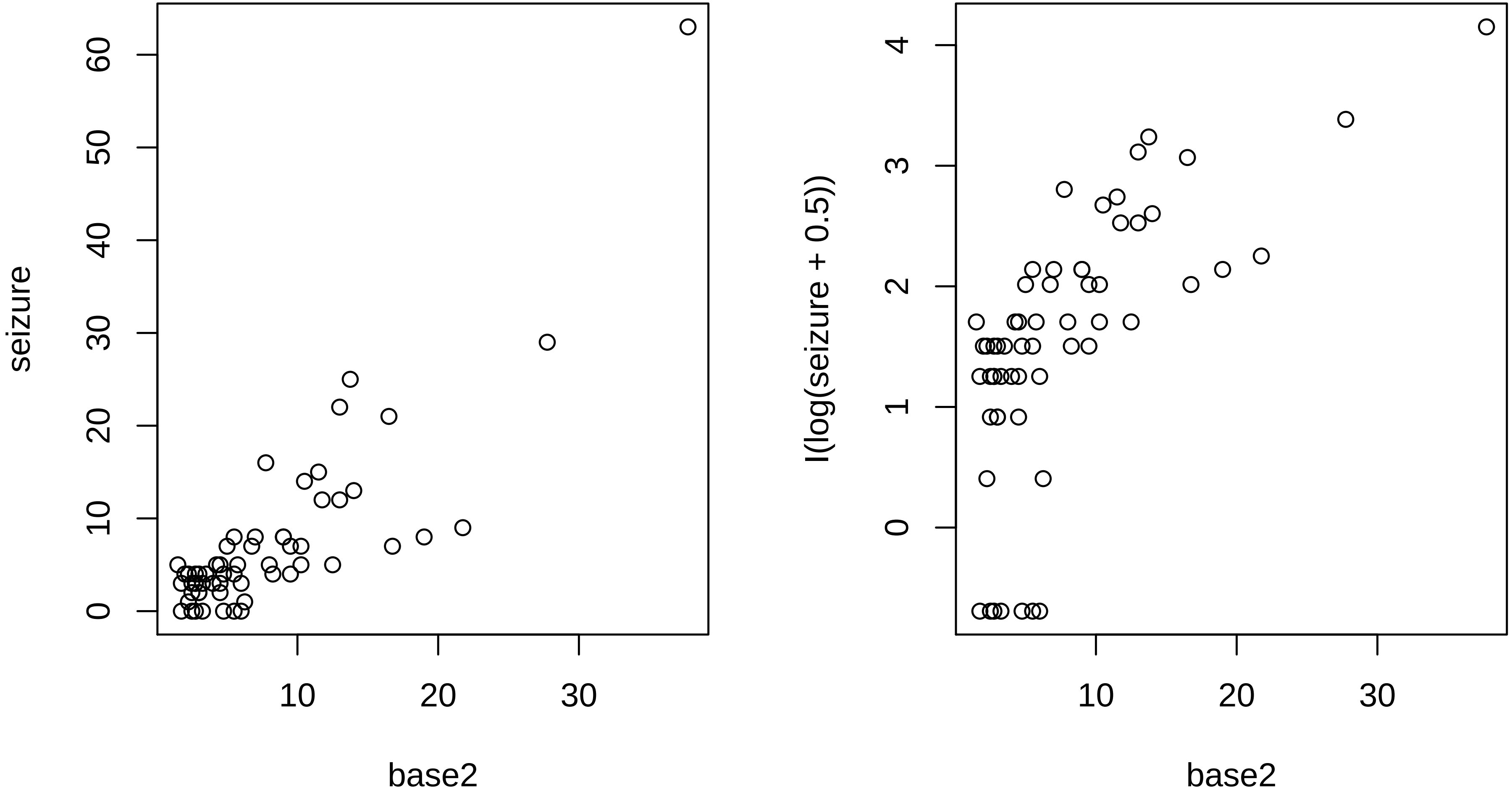
Data Example: Seizure Data Figure: Boxplots of square-root transformed seizure counts per two



weeks for epileptics at baseline (note: 8 weeks in total) and for four subsequent two-week periods after the patients were randomized to either placebo or progabide treatment.

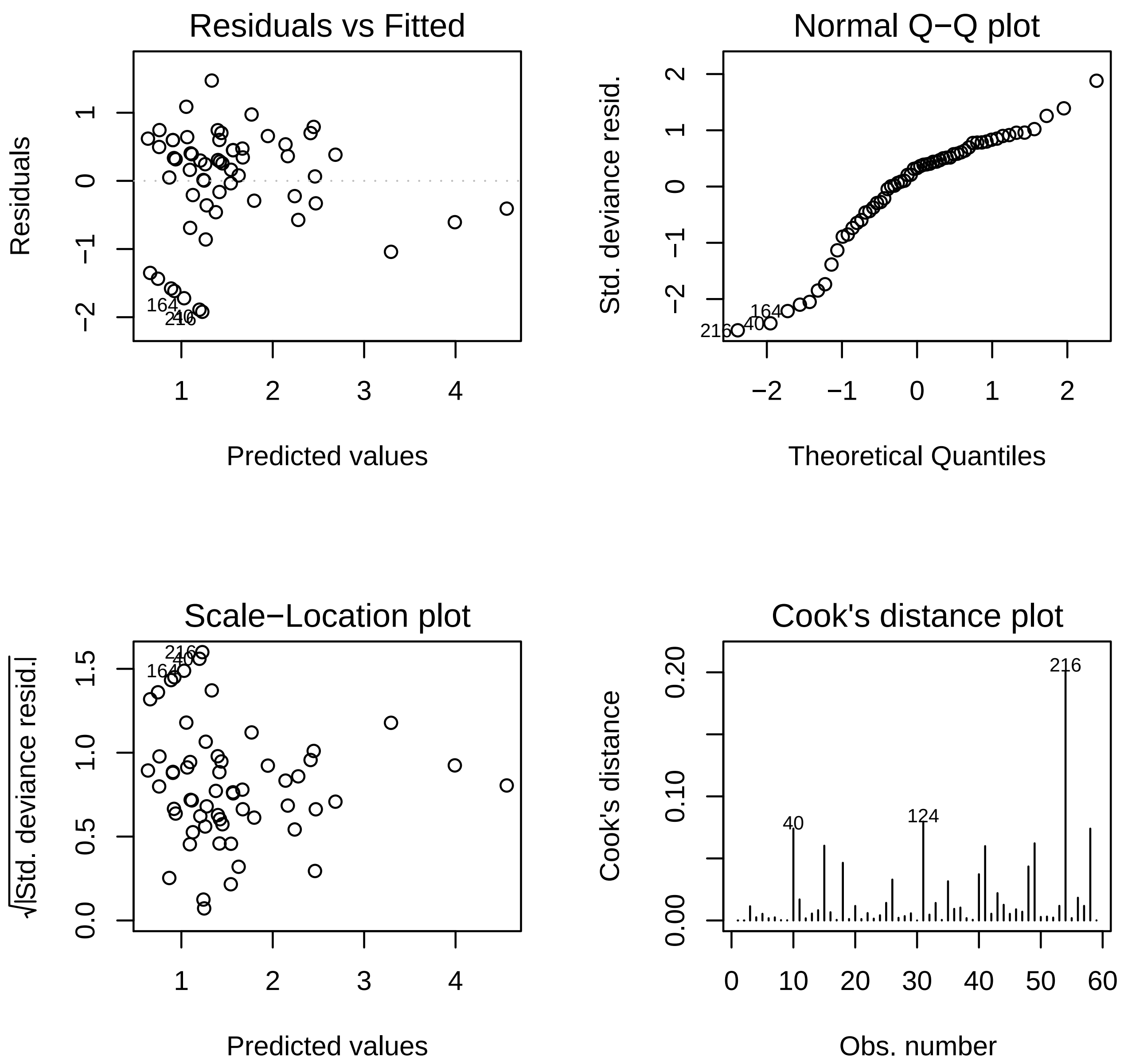
- Using only the responses at week 4.




seizure
0
10
20
30
40
50
60
10
20
30
base2
I(log(seizure + 0.5))
0
1
2
3
4
10
20
30
base2

```
> library (lattice)
> seize <- read.table("data/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)
```

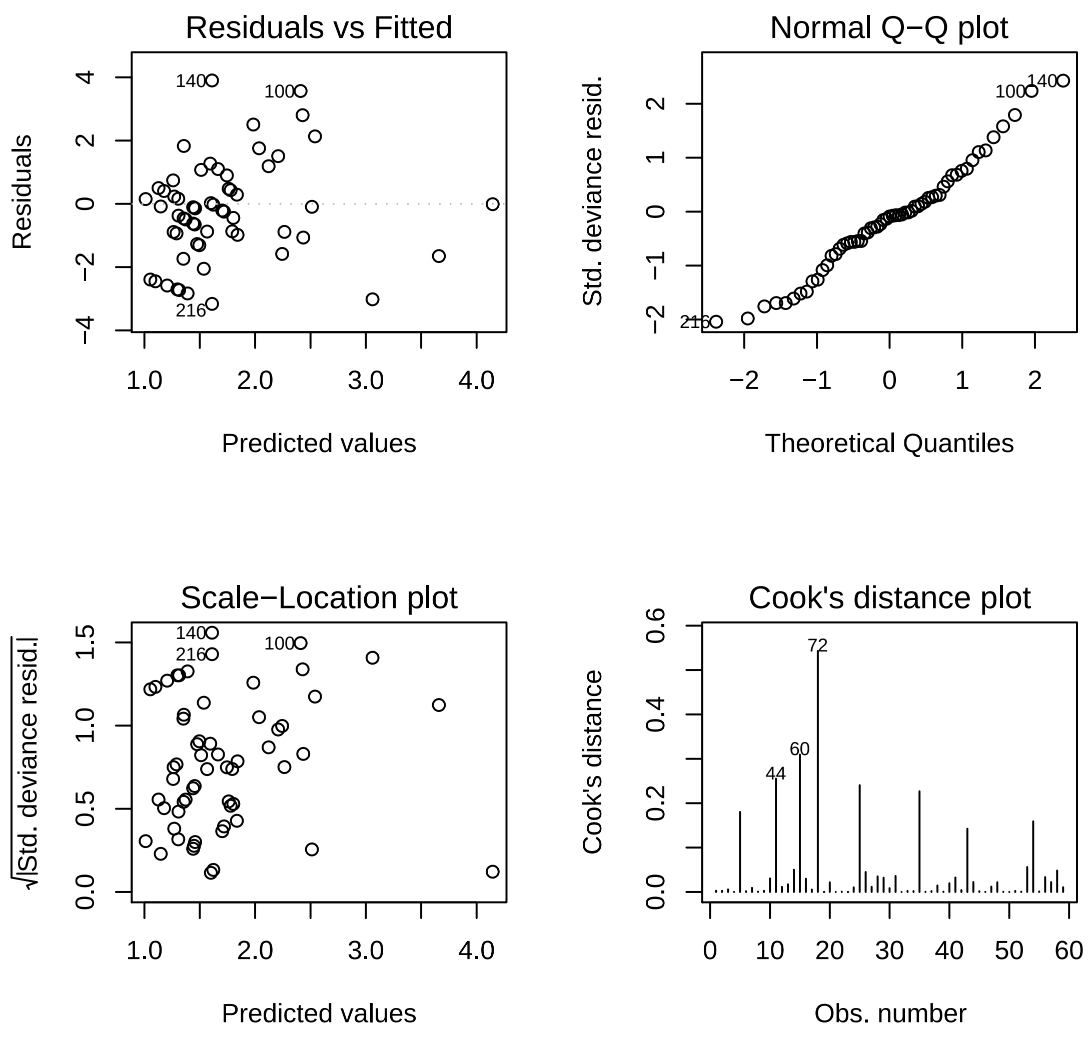

Residuals vs Fitted
Residuals
Predicted values
Normal Q−Q plot
Std. deviance resid.
Theoretical Quantiles
Scale−Location plot
√|Std. deviance resid.|
Predicted values
Cook's distance plot
Cook's distance
Obs. number
216
40
164
124

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)
```



Maximum Likelihood Estimation for GLMs

Solve score equations, for $j = 1, \ldots, p$, $S_j(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \beta_j} = 0$.

The log-likelihood:

$$\ell = \sum_{i=1}^{m} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\} = \sum_i \ell_i$$

$$S_j(\boldsymbol{\beta}) = \frac{\partial \ell}{\partial \beta_j} = \sum_i \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}$$

$$\frac{\partial \ell_i}{\partial \theta_i} = \frac{1}{a(\phi)}(y_i - b'(\theta_i)) = \frac{1}{a(\phi)}(y_i - \mu_i)$$

$$\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1} = \left(\frac{\partial b'(\theta_i)}{\partial \theta_i}\right)^{-1} = \frac{1}{b''(\theta_i)} = \frac{1}{V(\mu_i)}$$

$$\frac{\partial \mu_i}{\partial \eta_i} = \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^{-1} = \left(\frac{\partial g(\mu_i)}{\partial \mu_i}\right)^{-1} = \frac{1}{g'(\mu_i)}$$

$$\frac{\partial \eta_i}{\partial \beta_j} = X_{ij}$$

Therefore

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

- $(\frac{\partial \mu_i}{\partial \beta_j}) = \frac{X_{ij}}{g'(\mu_i)}$: Jacobian matrix.
- For fixed ϕ, the score function depends on μ_i and V_i only
- No knowledge on ϕ is needed for deriving the MLE of $\boldsymbol{\beta}$.

Write (2) in matrix form

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

Hence, the **Fisher's Information** is

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

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

Moreover (Cox and Reid, 1987),

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

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 ϕ.

Why Not Weighted Least Squares

The WLS approach need to minimize the following objective function

$$Q(\boldsymbol{\beta}, \phi) = \sum_{i=1}^{m} \frac{y_i - \mu_i(\boldsymbol{\beta})}{\mathrm{Var}(Y_i; \boldsymbol{\beta}, \phi)}.$$

Minimizing Q is equivalend to solving $\partial Q(\boldsymbol{\beta}, \phi)/\partial\boldsymbol{\beta} = 0$, where

$$\partial Q(\boldsymbol{\beta}, \phi)/\partial\boldsymbol{\beta} = \sum_{i=1}^{m} \left\{ -2 \left(\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^T \mathrm{Var}^{-1}(Y_i; \boldsymbol{\beta}, \phi)(y_i - \mu_i(\boldsymbol{\beta})) + \left(\frac{\partial}{\partial \boldsymbol{\beta}} \mathrm{Var}^{-1}(Y_i; \boldsymbol{\beta}, \phi) \cdot (y_i - \mu_i(\boldsymbol{\beta}))^2 \right) \right\}$$

- The first term is identical to $S(\boldsymbol{\beta})$.
- The second term has in general non-zero expectations. When $\mathrm{Var}^{-1}(Y_i)$ is free of $\boldsymbol{\beta}$ or $E[(y_i - \mu_i(\boldsymbol{\beta}))^2] = 0$, $\partial Q(\boldsymbol{\beta}, \phi)/\partial\boldsymbol{\beta} \equiv S(\boldsymbol{\beta}, \phi)$, and hence, the WLS estimator and the MLE are equivalent.
- In general, $\mathrm{E}[Q(\boldsymbol{\beta}, \phi)] \neq 0$; hence, the WLS estimator is generally inconsistent.

Iterative Weighted Least Squares

The MLE of $\boldsymbol{\beta}$ can be obtained by iterative weighted least squares (IWLS).

- When $g(\mu) = \mu = \boldsymbol{X}\boldsymbol{\beta}$, (2) immediately suggests an IWLS algorithm for solving the score equation:

  1. For given $\hat{\boldsymbol{\beta}}$, calculate the weights
  $$w_i = V(\mu_i; \hat{\boldsymbol{\beta}})^{-1}.$$
  2. Solve $\sum_i \boldsymbol{X}_i^T w_i(y_i - \boldsymbol{X}_i\boldsymbol{\beta}) = 0$ to get the next $\hat{\boldsymbol{\beta}}$.
  3. Go back to step 1 to update w_i's.

- (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

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

  – What is Z? $g(y) \simeq g(\mu) + (y - \mu)g'(\mu) = \eta + (y - \mu)\partial\eta/\partial\mu$

  – What is w_i? $\mathrm{Var}(Z) = \left(\left.\frac{\partial\eta}{\partial\mu}\right|_{\mu=\hat{\mu}} \right)^2 \mathrm{Var}(Y) = g'(\hat{\mu})^2 V(\hat{\mu})$.

  – What is $\sum_i \boldsymbol{X}_i^T w_i(z_i - \boldsymbol{X}_i\boldsymbol{\beta}) = 0$?

    This has the same form as (2) if the $\hat{\mu}$ in w_i is replaced by μ (excercise).

  – 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 score (gradient) and Fisher scoring method uses the expected derivative of the score (i.e., Fisher's information matrix, $-\mathcal{I}_m$)

The algorithm:

1. Find an initial value $\hat{\boldsymbol{\beta}}^{(0)}$.
2. For $j \to j+1$ update $\hat{\boldsymbol{\beta}}^{(j)}$ via
$$\hat{\boldsymbol{\beta}}^{(j+1)} = \hat{\boldsymbol{\beta}}^{(j)} + \left(\hat{\mathcal{I}}_m^{(j)}\right)^{-1} S(\hat{\boldsymbol{\beta}}^{(j)}).$$
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: $\mu_i(\hat{\beta})$,
  - from the full (saturated) model: y_i, the observed response.
- 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

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

  - Deviance is proportional to the likelihood ratio test statistic comparing the null hypothesis that the fitted model is adequate versus the saturated alternative.
  - A small value in D would indicate that the fitted model describes the data rather well.

Deviance examples:

- Normal:

$$\log f(y_i;\theta_i,\phi) = \frac{(y_i-\mu_i)^2}{2\sigma^2},$$
$$D(\boldsymbol{y},\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{m}(y_i-\hat{\mu}_i)^2 = \text{SSE}.$$

 The sum of residual squares!

- Binomial:

$$\log f(y_i;\theta_i,\phi) = m_i\left\{y_i\log\mu + (1-y_i)\log(1-\mu)\right\},$$
$$D(\boldsymbol{y},\hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{m}\left\{m_i y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right) - m_i(1-y_i)\log\left(\frac{1-y_i}{1-\hat{\mu}_i}\right)\right\}.$$

- Poisson:

$$\log f(y_i;\theta_i,\phi) = y_i\log(\mu) - \mu$$
$$D(\boldsymbol{y},\hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{m}\left\{y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i-\hat{\mu}_i)\right\},$$

 where the second term can be omitted as its sum is 0.

The deviance is the sum of squared **deviance residuals**. $D=\sum_{i=1}^{m} {r_D}_i^2$. For Poisson,

$$r_{D_i} = \text{sign}(y_i-\hat{\mu}_i)\left\{2\left(y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i-\hat{\mu}_i)\right)\right\}^{1/2}$$

Pearson's X^2

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

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

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

- Pearson's X^2 examples:
  - Normal: X^2 = residual sum of squares.
  - Poisson: $X^2 = \sum_{i=1}^{m} (y_i - \hat{\mu}_i)^2 / \hat{\mu}_i$
  - Binomial: $X^2 = \sum_{i=1}^{m} (y_i - \hat{\mu}_i)^2 / [\hat{\mu}_i(1 - \hat{\mu}_i)]$.

- For normal responses, when the model is correct, both D 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 m 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.

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.

$$r_D = \text{sign}(y - \mu)\sqrt{d_i}, \text{ where } \sum_{i=1}^{m} d_i = D.$$

  – 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).

Overdispersion

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

How Does Overdispersion Arise?

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

Suppose there exists a binary covariate, Z_i and that

$$Y_i \,|\, Z_i = 0 \sim \text{Poisson}\,(\lambda_0)$$

$$Y_i \,|\, Z_i = 1 \sim \text{Poisson}\,(\lambda_1)$$

$$\Pr(Z_i = 1) = \pi$$

Then

$$\begin{aligned} \mathrm{E}(Y_i) &= \\ \mathrm{Var}(Y_i) &= \\ &= \\ &= \end{aligned}$$

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, $i = 1, \ldots, m$.
- Let $\hat{\theta}_m$ be the MLE (based on m observations). That is, $\hat{\theta}_m$ solves the score equations that arise from the assumed F_θ:

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

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

Result

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

$$\mathrm{E}_G \left[\sum_{i=1}^{m} S_i^F(\theta^*) \right] = 0.$$

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

$$\sqrt{m}(\hat{\theta}_m - \theta^*) \longrightarrow \mathcal{N}\left(0, A^{-1}BA^{-1}\right)$$

where

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

$$B = \lim \frac{1}{m} \sum_{i=1}^{m} \mathrm{Var}_G \left[S_i^F(\theta) \big|_{\theta^*} \right]$$

$$= \lim \frac{1}{m} \sum_{i=1}^{m} \mathrm{E}_G \left[S_i^F(\theta) \big|_{\theta^*} \right]^2$$

- A is the expected value of the observed (based on the assumed model) information (times $1/m$).

- 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 $\hat{\theta}_m$.

- 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 $\mathrm{E}(Y_i) = \mu_i$ via a regression model with parameters $\boldsymbol{\beta}$, our estimator, $\hat{\boldsymbol{\beta}}$, will converge to whatever value solves

$$\mathrm{E}_G[S(\boldsymbol{\beta})] = 0.$$

Recall that we have

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

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

Motivation - in practice

- There are situations where the investigators are uncertain about the probability mechanism by which the data are generated
  - underlying biologic theory is not fully understood
  - no substantial (empirical) experience of similar data from previous studies is available

- Nevertheless, the scientific objective can often be adequately characterized through regression:
  - Systematic component

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

  - Variances specification

$$Var(y) = a(\phi)V(\mu)$$

- Least square is a special case for

$$y_i = x_i'\beta + \epsilon_i$$

  - Systematic component: $\mu_i = E(y_i \mid x_i) = x_i'\beta$
  - Variances specification: $Var(y_i) \equiv a(\phi)$

    Distribution of ϵ_i is **unspecified**.

Construction of quasi-likelihood

McCullagh and Nelder, 1989, Chapter 9. Wedderburn (1974) Biometrika.

Wedderburn (1974) proposed to use the **quasi-score function** to estimate $\boldsymbol{\beta}$, i.e. by solving

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{m} S_i(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^T \mathrm{Var}^{-1}(Y_i; \boldsymbol{\beta}, \phi)(y_i - \mu_i(\boldsymbol{\beta})) = \boldsymbol{0}.$$

- The random component in the generalized linear models is replaced by the following assumptions:

$$\mathrm{E}[Y_i] = \mu_i(\boldsymbol{\beta}) \text{ and } \mathrm{Var}[Y_i] = V_i = a(\phi)V(\mu_i).$$

- The **quasi-likelihood function** is

$$Q(\mu; y) = \sum_{i=1}^{m} \int_{y_i}^{\mu_i} \frac{y_i - t}{a(\phi)V(t)} dt$$

and

$$S(\boldsymbol{\beta}) = \partial Q / \partial \boldsymbol{\beta}.$$

- $S(\boldsymbol{\beta})$ possesses key properties of a score function

$$\mathrm{E}[S_i] = 0$$

$$\mathrm{Var}[S_i] = -\,\mathrm{E}[\partial S_i / \partial \mu_i]$$

- $S(\boldsymbol{\beta})$ would be the true score function for $\boldsymbol{\beta}$ if the Y_i's are indeed from an exponential family distribution.

How to assess precision of $\hat{\boldsymbol{\beta}}$

Taylor expansion gives $S(\hat{\boldsymbol{\beta}}) \doteq$

$$\sqrt{m}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \;\doteq\; \left\{ \sum_{i=1}^{m} \frac{(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}})' V_i^{-1} (\frac{\partial \mu_i}{\partial \boldsymbol{\beta}})}{m} \right\}^{-1} \left\{ \sum_{i=1}^{m} \frac{(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}})' V_i^{-1} (y_i - \mu_i(\boldsymbol{\beta}))}{\sqrt{m}} \right\}$$

$$\downarrow m \to \infty \qquad\qquad \downarrow m \to \infty$$

$$a^{-1} \qquad\qquad MVN(0, a)$$

"Model-based" variance estimate of $\hat{\beta}$

$$\left\{ \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)^T V_i^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right) \right\}_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}^{-1} \equiv A^{-1}$$

"Robust" variance estimate of $\hat{\beta}$

$$A^{-1} \left\{ \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)^T V_i^{-1} (y_i - \mu_i(\boldsymbol{\beta}))^2 V_i^{-1} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right) \right\}_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} A^{-1}$$

Summary for quasi-likelihood estimating equations

The quasi-likelihood regression parameter, $\hat{\boldsymbol{\beta}}$ for Y_i, $i = 1, \ldots, m$ is obtained as the solution to the quasi-score equations, $S(\boldsymbol{\beta}) = 0$, where

$$
\begin{aligned}
S(\boldsymbol{\beta}) &= \mathsf{D}^T \mathsf{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) \\
D_{ij} &= \frac{\partial \mu_i}{\partial \beta_j} \\
\mathsf{V} &= \text{diag}(a(\phi)V(\mu_i))
\end{aligned}
$$

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

$$
\begin{aligned}
\mathcal{I}_m &= \mathsf{D}^T \mathsf{V}^{-1} \mathsf{D}, \\
\text{Var}(\hat{\boldsymbol{\beta}}) &\approx \mathcal{I}_m^{-1}.
\end{aligned}
$$

- These properties are based **only** on the correct specification of the *mean* and *variance* of Y_i.
- Note that for the estimation of $a(\phi)$, the quasi-likelihood does not behave like a log likelihood. Method of moments is used.

$$
\tilde{a}(\phi) = \frac{1}{m-p} \sum_{i=1}^{m} \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} = \frac{\chi^2}{m-p},
$$

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 quasi-likelihood regression parameters are larger than that of GLM.
- 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 (Heyde, 1997) but we use it in a narrower sense in GLM with variance function being

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

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

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

- 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

$$\mathrm{E}_\theta[g(Y, \theta)] = 0, \text{ for any } \theta \in \Theta.$$

Role of unbiasedness: under regularity conditions, unbiased estimating 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

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

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

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

  - Method of moments (K. Pearson).

$$\mu_r(\theta) = E(\boldsymbol{Y}^r), r = 1, 2, \ldots \quad \text{and} \quad \hat{\mu}_r = \frac{1}{m}\sum_{i=1}^{m} y_i^r; \qquad \text{solve the equations of} \quad \mu_r(\theta) = \hat{\mu}_r.$$

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,

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

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

$$W = \frac{\mathrm{E}\left[g(y,\theta)^2\right]}{\left[\mathrm{E}(\partial g/\partial\theta)\right]^2} = \mathrm{E}\left[\frac{g(y,\theta)}{\mathrm{E}(\partial g/\partial\theta)}\right]^2. \tag{3}$$

  – The numerator is the variance, $\mathrm{Var}(g)$.

  – 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.

– W is the variance of the standardized estimating function: $g(y, \theta)/|\operatorname{E}(\partial g/\partial\theta)|$.

- 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

$$W^* = \frac{1}{-\operatorname{E}[\ddot{\ell}(\theta)]} = \frac{1}{\operatorname{E}[\dot{\ell}(\theta)^2]},$$

where $\ddot{\ell}(\theta) = \partial^2 l(\theta)/\partial\theta^2$ ("Cramer-Rao lower bound"). The denominator is Fisher's information.

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

$$\sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right)^T V_i^{-1}\big(y_i - \mu_i(\boldsymbol{\beta})\big),$$

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

$$\sum_{i=1}^{m} d_i(\boldsymbol{\beta}, \phi)(y_i - \mu_i(\boldsymbol{\beta})). \tag{4}$$

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

For an unbiased estimating function of the form (4), the optimality criterion (3) reduces to

$$W_m = \frac{\sum_{i=1}^m d_i^2 V_i}{\left(\sum_{i=1}^m d_i \frac{\partial \mu_i}{\partial \beta}\right)^2} = \frac{\sum_{i=1}^m (d_i\sqrt{V_i})^2}{\left\{\sum_{i=1}^m (d_i\sqrt{V_i})\left(\frac{\partial \mu_i}{\partial \beta}\frac{1}{\sqrt{V_i}}\right)\right\}^2},$$

since...

$< Proof(cont.) >$

For the quasi-score function,

$$d_i^*(\beta, \phi) = \left(\frac{\partial \mu_i}{\partial \beta}\right) V_i^{-1},$$

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

Using Cauchy - Schwarz's inequality

$$(\Sigma_i x_i y_i)^2 \leq (\Sigma_i x_i^2)(\Sigma_i y_i^2),$$

it follows immediately that

$$W_m^* < W_m$$

for any choice of $d_i(\beta, \phi)$.$\square$

- 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 some statistic $\boldsymbol{T}_i$ such that

$$\mathrm{E}(g_i \mid \boldsymbol{T}_i) = 0.$$

  – 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 $\boldsymbol{T}_i$,

$$\sum_{i=1}^{m} d_i(\theta, \boldsymbol{T}_i) g_i.$$

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

$$g = \sum_{i=1}^{m} \mathrm{E}\left(\frac{\partial g_i}{\partial \theta} \mid \boldsymbol{T}_i\right)^T \mathrm{Var}(g_i \mid \boldsymbol{T}_i)^{-1} g_i. \tag{5}$$

such that the solution to the estimating equation has minimal asymptotic variance.

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 m, 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
$$\frac{\partial \log f(y \mid T = t; \theta)}{\partial \theta}$$
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, $\partial/\partial\theta[\log f(y \mid T = t; \theta)]$ 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 $\boldsymbol{y} = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_m)$ is decomposed into m "strata" and the $\boldsymbol{y}_i$'s are uncorrelated with each other. The dimensions of $\boldsymbol{y}_i$'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, $g_i(\boldsymbol{y}_i; \theta, \phi)$, for each of the m strata, i.e.,

$$\mathrm{E}(g_i; \theta, \phi) = 0 \qquad \forall \theta, \phi, i.$$

- In a regression setting

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

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

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

and it leads to the optimal $\boldsymbol{g}$ among (4), namely

$$\boldsymbol{g} = \sum_{i=1}^{m} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} \right)^T \mathrm{Var}(\boldsymbol{Y}_i)^{-1} \big(\boldsymbol{y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta}) \big). \tag{6}$$

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

- Note that the dimension of $\boldsymbol{g}_i$ varies from stratum to stratum and when $n_i = 1$ for all i, $\boldsymbol{g}$ reduces to the quasi-score function.
- It is a special case of (4) and (5).
- 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 $\boldsymbol{g}_i$ that does not include ϕ in its functional form, in general the distribution of $\boldsymbol{g}_i$ depends on ϕ.
- Liang and Zeger (1995) argued that the impact of the nuisance parameters on $\boldsymbol{g}$ and on the corresponding solution of $\boldsymbol{g} = 0$ is small, because it shares the orthogonality properties enjoyed by the conditional score function.

    1. $\mathrm{E}\left(\boldsymbol{g}(\theta, \phi^*); \theta, \phi\right) = 0$ for all θ, ϕ, and ϕ^* where ϕ^* is an incorrect value (estimate) for ϕ.
    2. $\mathrm{E}\left(\partial \boldsymbol{g}(\theta, \phi^*)/\partial \phi^*; \theta, \phi\right) = 0$ for all θ, ϕ, and ϕ^*.
    3. $\mathrm{Cov}\left(\boldsymbol{g}(\theta, \phi), \partial \log f(y; \theta, \phi)/\partial \phi\right) = 0$ 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 $\boldsymbol{g}(\theta, \hat{\phi}_\theta) = 0$) 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 $\boldsymbol{g}(\theta, \hat{\phi}_\theta)$ with $\hat{\phi}_\theta$ plugged into the EF is diminished at a faster rate than that of $\boldsymbol{S}_\theta(\theta, \hat{\phi}_\theta)$, the ordinary score function evaluated at $\hat{\phi}_\theta$.
  - **robust** - even if the assumption on how ϕ describes the distribution of the $\boldsymbol{y}$'s is misspecified, the solution remains consistent and its asymptotic variance is unaltered.

Further Reading

- Chapter 2 and 9 of McCullagh and Nelder, 2nd edition.

References (The highlighted papers will be distributed in class)

- Godambe VP (1960) An optimum property of regular maximum likelihood estimation. *Annals of Mathematical Statistics* **27**:357-72.
- Godambe VP (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.
- Huber, P. (1967), The behavior of the maximum likelihood estimates under nonstandard conditions, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley. **1**221233.
- **Liang KY and Zeger SL (1995) Inference based on estimating functions in the presence of nuisance parameters (with discussion). *Statistical Science* 10158-73.**
- Lindsay B (1982). Conditional score functions: some optimality results. *Biometrika* **69**:503-12.
- Pierce, D. and Schafer D. (1986). Residuals in Generalized Linear Models. *JASA* **81**:977-86.
- **Wedderburn RWM (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 61:439-47.**
- White, H. (1982), Maximum likelihood estimation of misspecified models, *Econometrica* **50**(1), 126.