4. Quasi-Likelihood Methods

Review of Generalized Linear Models

Consider independent data (X_i, y_i), $i = 1, \cdots, m$.

Linear Model

- Random component: normal errors

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

- Systematic component: linear combination of predictors

$$\mu_i = E(Y_i | X_i) = X_i^T \beta.$$

Generalized Linear Model

- Random component: exponential family f

$$Y_i | X_i \sim f(\theta, \phi)$$

where ϕ is the dispersion parameter.

- Systematic component: linear combination of predictors

$$\eta_i = X_i^T \beta.$$

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

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

where $\mu_i = E(Y_i|X_i)$.

Exponential Family

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

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

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 (1), then

$$E(Y) = \mu = b^{'}(\theta),$$
$$var(Y) = b^{''}(\theta)a(\phi).$$

  Proof) The log-likelihood is

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

  Therefore,

$$\frac{\partial l(\theta,\phi)}{\partial \theta} = \frac{y - b^{'}(\theta)}{a(\phi)},$$
$$\frac{\partial^2 l(\theta,\phi)}{\partial \theta^2} = -\frac{b^{''}(\theta)}{a(\phi)}.$$

Using the fact that

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

we get

$$E\left(\frac{y - b^{'}(\theta)}{a(\phi)}\right) = 0 \quad \rightarrow \quad E(Y) = b^{'}(\theta),$$

$$E\left\{\left(\frac{\partial l}{\partial \theta}\right)^2\right\} = E\left\{\frac{(y - b^{'}(\theta))^2}{a^2(\phi)}\right\} = \frac{var(Y)}{a^2(\phi)}.$$

So,

$$\frac{var(Y)}{a^2(\phi)} = \frac{b^{''}(\theta)}{a(\phi)} \quad \rightarrow \quad var(Y) = b^{''}(\theta)a(\phi).$$

- Examples of exponential family

– 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

$$\begin{aligned} & \theta = \log\lambda, \quad b(\theta) = \lambda = \exp(\theta), \\ & c(y,\phi) = -\log(y!), \quad a(\phi) = 1. \\ \Longrightarrow \quad & \mu = b^{'}(\theta) = \exp(\theta) = \lambda, \\ & var(Y) = b^{''}(\theta)a(\phi) = \exp(\theta) = \lambda. \end{aligned}$$

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

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

So

$$\theta = \log\left(\frac{\pi}{1-\pi}\right) = \mathsf{logit}(\pi), \quad b(\theta) = -\log(1-\pi) = \log\{1+\exp(\theta)\},$$

$$c(y,\phi) = \log\begin{pmatrix} m \\ my \end{pmatrix}, \quad a(\phi) = \frac{1}{m}.$$

$$\Longrightarrow \qquad \mu = b^{'}(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} = \pi,$$

$$var(Y) = b^{''}(\theta)a(\phi) = \frac{\pi(1-\pi)}{m}.$$

– Gaussian:

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

So,

$$\theta = \mu, \quad b(\theta) = \mu^2/2,$$

$$a(\phi) = \sigma^2, \quad c(y,\phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2),$$

$$\Longrightarrow \qquad \mu = b^{'}(\theta) = \mu,$$

$$var(Y) = b^{''}(\theta)a(\phi) = \sigma^2.$$

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: $g(\mu) = \log(\mu)$,
  Binomial: $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$.

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

$$var(Y) = v(\mu)a(\phi).$$

  Usually $a(\phi) = \phi/w_i$ where ϕ is the scale parameter and w is a weight.
  Gaussian: $v(\mu) = 1$, Poisson: $v(\mu) = \lambda$,
  Binomial: $v(\mu) = \pi(1-\pi)$.

- 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 $Z > 0$).
- Complementary log-log: $g(\mu) = \log(-\log \mu)$, β is the log hazard ratio.

Example: Seizure Data

Model Diagnosis and Residuals

Like ordinary linear models, residuals can be used to assess model fit for GLMs:
Q-Q plots (sometimes hard to interpret. eg., for binary data)
Residuals vs fitted values, or omitted covariates
Systematic departure of the mean structure
Variance function

Types of Residuals

1. Response residuals: $r_R = y - \hat{\mu}$.

2. 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)

3. Working residuals

$$r_w = (y - \hat{\mu})\frac{\partial \eta}{\partial \mu} = Z - \hat{\eta}.$$

4. Deviance residuals: contribution of Y_i to the deviance

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

   - Close to a normal distribution (less skewed) than Pearson residuals.
   - Often better for spotting outliers.

Maximum Likelihood Estimation for GLMs

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

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

The log-likelihood is

$$l = \sum_{i=1}^{m} \left\{ \frac{y\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\} = \sum_{i=1}^{m} l_i,$$

$$S_j(\beta) = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{m} \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j},$$

where

$$\frac{\partial l_i}{\partial \theta_i} = \frac{1}{a_i(\phi)}(y_i - b'(\theta_i)) = \frac{1}{a_i(\phi)}(y_i - \mu_i),$$

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

$$\frac{\partial \mu_i}{\partial \eta_i} = \frac{1}{g'(\mu_i)}, \qquad \frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

Therefore,

$$S_j(\beta) = \sum_{i=1}^{m} \frac{x_{ij}}{g'(\mu_i)} \left[a_i(\phi) v(\mu_i)\right]^{-1} (y_i - \mu_i). \qquad (2)$$

- 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 (2) in matrix form

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

Hence

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

Moreover,

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

The information matrix is of the form

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

The MLEs $\hat{\beta}$ and $\hat{\phi}$ are asymptotically independent.

Iterative reweighted least squares

- When $g(\mu) = \mu = X\beta$, (2) immediately suggests an iterative weighted least squares (IWLS) algorithm for solving the score equation:
  - For given $\hat{\beta}$, calculate the weights

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

  - Solve

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

– Update W.

- For fixed ϕ, when g is nonlinear, the IWLS algorithm needs to be modified by constructing a working response

$$Z = \hat{\eta} + (Y - \hat{\mu})\frac{\partial \eta}{\partial \mu}$$

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 $\eta = X_i\beta$.

- The IWLS algorithm can be justified as an application of the Fisher scoring method.

Fisher Scoring

To solve the score equations

$$S(\beta) = 0,$$

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, $-I_n$).
The algorithm:

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

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

$$\hat{\beta}^{(j+1)} = \hat{\beta}^{(j)} + \left(\hat{I_n}^{(j)}\right)^{-1} S(\hat{\beta}^{(j)}).$$

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

4. Iterate until convergence criterion is satisfied.

Note that IWLS is equivalent to Fisher scoring.

Deviance

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

Idea:

- For μ_i, two approaches to estimate μ_i
  - from the fitted model: $\mu_i(\hat{\beta})$.
  - from the full (saturated) model: y_i.

- On 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

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

  - The deviance of the fitted model is defined as

$$D(\hat{\mu}; y) = -2 \log \left(\frac{L(\hat{\mu}; \phi)}{L(y; \phi)} \right).$$

ex) $y_i \sim N(\mu_i, \phi)$. Then the sum of residual squares is

$$D(\hat{\mu}, y) = \sum_i (y_i - \mu_i(\hat{\beta}))^2.$$

Note: except for normal responses, the distribution of D, even asymptotically, is unknown.

- The deviance is not useful for binary data with many covariate patterns or for any model with a scale parameter.

- Since the deviance is analogous to the RSS, it is possible to define an R^2 as

$$R^2 = 1 - \frac{\text{Deviance for fitted model}}{\text{Deviance for null model}}.$$

Despite commonly reported in GLM software, it is not very useful.

- Let $\hat{\theta} = \theta(\hat{\mu})$, $\tilde{\theta} = \theta(y)$. Assume $a_i(\phi) = \phi/w_i$, the deviance can be written

$$D(y, \hat{\mu}) = 2 \sum_{i=1}^{m} w_i \left\{ y_i[\tilde{\theta}_i - \hat{\theta}_i] - [b(\tilde{\theta}_i) - b(\hat{\theta}_i)] \right\} / \phi.$$

  For example,

  – Normal

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

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

  – Poisson

$$\log f(y_i; \theta, \phi) = y_i \log \mu_i - \mu_i,$$

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

  It is sometimes referred to as the G statistic.

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

– Binomial

$$\log f(y_i;\theta_i,\phi) = m_i\left\{y_i\log\mu_i + (1-y_i)\log(1-\mu_i)\right\},$$

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

- The deviance is the sum of the deviance residuals.

Pearson's χ^2

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

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

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

– Normal: $\chi^2 = SSE$
– Poisson: $\chi^2 = \sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2/\hat{\mu}_i$

– Binomial: $\chi^2 = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2/[\hat{\mu}_i(1-\hat{\mu}_i)]$

- For normal-theory linear models, both deviance and χ^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 is additive when comparing nested models with ML estimation while the generalized Pearson's χ^2 is sometimes preferred for easy interpretation.

Overdispersion

- For Poisson regression, it is expected that $var(Y_i) = \mu_i$. However this can be sometimes violated.

- Overdispersion describe the situation above. That is, the data are overdispersed when the actually $var(Y_i)$ exceeds the GLM variance $\phi v(\mu_i)$.

- For binomial and Poisson models we often find overdispersion:

- Binomial: $Y = s/m$, $E(Y) = \mu$, $var(Y) > \mu(1-\mu)/m$.
- Poisson: $E(Y) = \mu$, $var(Y) > \mu$.

- How does overdispersion arise?
  - If there is population heterogeneity, then overdispersion can be introduced.
    Suppose there exists a binary covariate Z_i and that

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

    Then

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

Therefore, if we do not observe Z_i, then the omitted factor leads to increased variation.

Impact of Model Misspecification

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

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

- Let F_θ be the assumed distribution family for independent data Y_i, $i = 1, \cdots, n$. Then the quasi-log likelihood of the data is given by

$$l_n(y, \theta) = n^{-1} \sum_{i=1}^{n} \log f(y_i, \theta).$$

- Let $\hat{\theta}_n$ be the quasi-maximum likelihood estimator (QMLE) for θ^* (based on n observations). That is, $\hat{\theta}_n$ solves the score equations that arise from the assumed F_θ:

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

- When F_θ contains the true structure G (i.e., $G(y) = F(y, \theta_0)$) for some θ_0 in Θ, the (Q)MLE $\hat{\theta}_n$ is consistent for θ_0 under suitable regularity conditions (Wald, 1949). It it does not hold, $\hat{\theta}_n$ is a natural estimator for θ^* which minimizes the Kullback-Leibler information criterion,

$$I(g : f, \theta) = E\left(g(y_i)/f(y_i, \theta)\right).$$

(because $l_n(y, \theta)$ is a natural estimator for $E(\log f(y_i, \theta))$.)

Result:

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

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

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

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

where

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

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

- A is the expected value of the observed (based on the assumed model) information (times $\frac{1}{n}$).

- 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 $S_i^F(\theta)$, because the assumed model is not true.

- In general $\hat{\theta}$ is not consistent for 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 parameter β, our estimator $\hat{\beta}$ will converge to whatever value solves

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

Recall that we have

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

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

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:

$$E(Y_i) = \mu_i(\beta),$$
$$var(Y_i) = \sigma^2 v(\mu_i)$$

where σ^2 is unknown and 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):

$$E(U) = 0, \quad var(U) = \frac{1}{\sigma^2 v(\mu)},$$
$$-E\left(\frac{\partial U}{\partial \mu}\right) = \frac{1}{\sigma^2 v(\mu)}.$$

Therefore the integral is

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

If it exists, it 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.

- The log likelihood function is identical to the quasi-likelihood if and only if it belongs to the exponential family.

Theorem: For one observation of y, the log likelihood function l has the property

$$\frac{\partial l(\mu; y)}{\partial \mu} = \frac{y - \mu}{v(\mu)}$$

where $\mu = E(Y)$ and $v(\mu) = var(Y)$ if and only if the density of Y can be written in the form

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

where

$$\theta = \int \frac{d\mu}{v(\mu)}$$

and g is some function of θ. ■

Note:

- The one-parameter exponential family is the weakest distribution assumption in that only the mean-variance relationship is specified.

- Note that not all possible choices of v lead to an authentic log likelihood function.

Quasi-Likelihood Estimating Equations

The quasi-likelihood regression parameter $\hat{\beta}$ for Y_i, $i = 1, \cdots, n$ is obtained as the solution to the quasi-score equations:

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

where $D_{ij} = \frac{\partial \mu_i}{\partial \beta_j}$ and $V = diag(\sigma^2 V(\mu_i))$.

Properties:

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

$$I_n = D^T V D,$$
$$var(\hat{\beta}) \approx I_n^{-1}.$$

- Asymptotic normality

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \to N(0, I_n^{-1}).$$

- 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. Methods of moments is used to estimate σ^2

$$\tilde{\sigma^2} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(Y_i - \hat{\mu}_i)^2}{v_i(\hat{\mu}_i)}.$$

Example: Seizure Data.

Estimating Functions

Note that 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

$$var(Y) = a(\phi)v(\mu).$$

We treat quasi-score equations as a special case of estimating equations where the variance can be

$$var(Y) = v(\mu, \phi).$$

An estimating function is a function of data and parameter, $g(Y, \theta)$. The function $g(Y, \theta)$ is unbiased if for any $\theta \in \Theta$,

$$E_\theta \left[g(Y, \theta) \right] = 0.$$

For an unbiased estimating function, the estimating equation

$$g(Y, \hat{\theta}) = 0$$

defines an estimator for θ.

Note Estimating functions form the basis of (almost ?) all of frequentist statistical estimation.

- Method of least squares (LS) (Gauss and Legendre): finite sample consideration

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

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

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

- Method of moments (K. Pearson)

Optimal Estimating function

- 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,

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

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

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

  - The numerator is the variance $var(g)$.
  - The denominator is square of the averaged gradient of g.

- We want that the optimal g has small variance and an 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:

$$\frac{g(y;\theta)}{E(\partial g/\partial\theta)}.$$

- Godambe (1960) showed that the score functions (even non-linear ones) for θ

$$i = \frac{\partial l(\theta)}{\partial\theta}$$

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

$$w^* = \frac{1}{E\left[i(\theta)\right]} \quad (\text{“Cramer-Rao lower bound”}).$$

- Godambe and Heyde (1987) proved that the quasi-

score function

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

is optimal among unbiased estimating functions which are linear in the data, that is, take the form

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

where $v_i = var(Y_i)$.
proof) Here is a sketch of the proof for the scalar case (Liang and Zeger, 1995). For an unbiased estimating function of the form (5). The optimality criterion (4) 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,

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

Using Schwarz's inequality

$$E(|XY|)^2 \leq EX^2 EY^2,$$

it follows immediately that

$$W_n^* \leq W_n$$

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 are not linear.

- When only the mean model is known, only the linear estimating function can be guaranteed to be unbiased.

- When there is a nuisance parameter ϕ, i.e., the likelihood is $f(y;\theta,\phi)$. Godame (1976) considered a complete and sufficient statistic T for ϕ and showed conditional score function

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

is the optimal estimating function for θ.

- What if there is a nuisance parameter but we cannot specify the likelihood?

Nuisance Parameter and Estimating Functions

- The existence of nuisance parameters is a nuisance indeed.
  - The conditional score function requires the existence of T, a complete and sufficient statistic for ϕ that does not depend on θ.
  - If $var(Y) = v(\mu, \phi) \neq a(\phi)v(\mu)$, the quasi-score function is no longer optimal.
  - If the dimension of ϕ increases with the sample size n, the MLE for θ may not even be consistent.

- 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.

Constructing Estimating Functions

- The data $y = (y_1, \cdots, y_n)$ is decomposed into n "strata" and the y_i's are uncorrelated with each other.

- Assuming the parameter θ are common to all n strata and the existence of an unbiased estimating function $g_i(y;\theta)$ for each of the n strata. i.e.,

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

- The unbiasedness of g_i is verified through defining a statistic A_i such that

$$E(g_i|A_i) = 0.$$

For example, $A_i = (y_1, \cdots, y_{i-1})$ (the history), then for $j < i$,

$$cov(g_i, g_j) = E(g_i g_j) = E(g_j E(g_i|A_i)) = 0.$$

Hence, the uncorrelated condition is automatically satisfied.

- We would like to combine the g_i using a weighted average where the weight is allowed to a function of A_i

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

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

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

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

- The term $var(g_i|A_i)^{-1}$ is used to down-weight those g_i with greater variance.

- The first term $E\left(\frac{\partial g_i}{\partial \theta} | A_i\right)^T$ transforms the space spanned by the data to the parameter space.

- In a regression setting

$$E(y_i) = \mu_i(\beta),$$

we can use

$$g_i = y_i - \mu_i(\beta)$$

and it leads to

$$g = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta} \right)^T var(y_i)^{-1}(y_i - \mu_i(\beta)). \qquad (6)$$

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

- Note that here we start to talk about clustered or correlated data.

- GEE1 is special case for this extremely flexible inference framework.

Examples

- Quasi-score function for independent over-dispersed data (Poisson or binomial).

- GEE1 for longitudinal data.

- Consider m 2×2 tables

y_{i1}	$n_{i1} - y_{i1}$	n_{i1}
y_{i2}	$n_{i2} - y_{i2}$	n_{i2}

  The parameter of interest is θ =common odds ratio.

  – Let $g_i(y_{i1}, y_{i2}; \theta) = y_{i1}(n_{i2} - y_{i2}) - \theta y_{i2}(n_{i1} - y_{i1})$.
  – It is unbiased (also true conditional on $A_i = y_{i1} + y_{i2}$)

$$E(g_i) = n_{i1}n_{i2}\pi_{i1}(1-\pi_{i2}) - \theta n_{i2}n_{i1}\pi_{i2}(1-\pi_{i1}) = 0.$$

– The estimating function is

$$\sum_{i=1}^{m} E\left[\frac{\partial g_i}{\partial \theta}\right]^T var(g_i)^{-1} g_i(y_{i1}, y_{i2}; \theta)$$

$$\approx \sum_{i=1}^{m} \frac{1}{n_{i1} + n_{i2}} g_i(y_{i1}, y_{i2}; \theta).$$

– The solution

$$\hat{\theta} = \frac{\sum_{i=1}^{m} y_{i1}(n_{i2} - y_{i2})/(n_{i1} + n_{i2})}{\sum_{i=1}^{m} y_{i2}(n_{i1} - y_{i1})/(n_{i1} + n_{i2})}$$

is the Mantel-Haenszel estimator which is known to be superior to the MLE of θ. In particular, if $n_{i1} = n_{i2} = 1$ for all i and $m \to \infty$, the MLE converges to θ^2 rather than θ while the Mantel-Haenszel estimator is still consistent.

Nuisance Parameter

- Even though we choose g_i that does not include ϕ in its functional form, the distribution of g_i generally depends on ϕ.

- Liang and Zeger (1995) argued that the impact of the nuisance parameters on g and on the corresponding solution of $g = 0$ is small because of three orthogonality properties

  1. $E(g(\theta, \phi^*); \theta, \phi) = 0$ for all θ, ϕ, and ϕ^* where ϕ^* is an incorrect value (estimate) for ϕ.
     This property states that the unbiasedness of the conditional score function is also preserved if evaluated at the incorrect value ϕ^* for nuisance parameter. Hence, for example, $\hat{\theta}(\phi^*)$ is consistent where $\hat{\theta}(\phi^*)$ is the root of $g(\theta, \phi^*) = 0$.
  2. $E\left(\frac{\partial g(\theta,\phi^*)}{\partial \phi^*}; \theta, \phi\right) = 0$ for all θ, ϕ, and ϕ^*.
     This property is an asymptotic version of the first one and it implies that $E(g(\theta, \tilde{\phi}); \theta, \phi) = 0$ for any values of $\tilde{\phi}$ which are $\sqrt{n}-$consistent. That is,
     $$\sqrt{n}(\tilde{\phi} - \phi) \xrightarrow{p} 0.$$
  3. $cov\left(g(\theta, \phi), \frac{\partial \log f(y;\theta,\phi)}{\partial \phi}\right) = 0$ for all θ and ϕ.
     This property implies that $E(g(\theta, \hat{\phi}_\theta); \theta, \phi) = 0$ for $\hat{\phi}_\theta$, the MLE of ϕ for fixed θ.

- Conclusion: when a $\sqrt{n}-$consistent estimator $\tilde{\phi}$ for

ϕ is used, the asymptotic variance of $\hat{\theta}$ (solution of $g(\theta, \tilde{\phi}) = 0$) is the same as if the true value of ϕ is known.

Asymptotic Properties of the Estimator (Fahrmeir and Kaufmann (1985))

- The optimal estimating function is defined in terms of finite sample properties (unbiasedness, minimal variance) but the derived estimator enjoys some nice asymptotic properties.

- Weak consistency: The sequence $X_1, X_2, \cdots, X_n$ converge in probability to X.

$$X_n \xrightarrow{p} X,$$

if for any $\epsilon > 0$,

$$P(|X_n - X| > \epsilon) \to 0 \text{ as } n \to \infty.$$

- Strong consistency: X_n converges to X almost

surely

$$X_n \overset{a.s.}{\to} X,$$

if for any $\epsilon > 0$,

$$P(Sup_{m>n}|X_m - X| > \epsilon) \to 0 \text{ as } n \to \infty.$$

- For classic linear model with iid errors,

$$\lim_{n\to\infty} \lambda_{min} \sum_{i=1}^{n} X_i X_i^T \to \infty$$

is the necessary and sufficient condition for either weak or strong consistency of the OLS estimator $\hat{\beta}$. λ_{min} is the smallest eigenvalue of a symmetric matrix.

- For generalized linear models (Fahrmeir and Kaufmann, 1985):

(D) Divergence:

$$\lambda_{min} I_n \to \infty,$$

where $I_n = I_n(\beta_0)$ and β_0 is the true β.

(C) Boundedness from below: for all $\delta > 0$,

$$I_n(\beta) - cI_n \text{ positive semidefinite}$$

for any $\beta \in N_n(\delta)$, $n \geq n_1(\delta)$, $N_n(\delta)$ is a neighborhood of β_0 ($N_n(\delta) = \{\beta : ||I_n^{T/2}(\beta - \beta_0)|| < \delta\}$) and $c > 0$ is independent of δ.

(N) Convergence and continuity: for all $\delta > 0$,

$$max_{\beta \in N_n(\delta)} ||V_n(\beta) - I|| \to 0$$

where $V_n(\beta) = I_n^{-1/2} I_n(\beta) I_n^{-T/2}$ is the normalized information matrix and I is the identity matrix.

Note that (N) implies (C).

- Theorem: Under (D) and (C), the sequence of roots $\hat{\beta}_n$

  1. asymptotic existence

$$P(s_n(\hat{\beta}_n) = 0) \to 1.$$

2. Weak consistency

$$\hat{\beta}_n \xrightarrow{p} \beta_0.$$

- Theorem: Under (D) and (N), the normed score function is asymptotically normal:

$$I_n^{-1/2} S_n \xrightarrow{d} N(0, I).$$

It also implies that

$$I_n^{1/2}(\hat{\beta}_n - \beta_0) \xrightarrow{d} N(0, I).$$

(the normed MLE is asymptotically normal)

Note

- This is for canonical link.
- $I_n = I_n(\beta_0)$.

- In practice we need replace I_n with $I_n(\hat{\beta})$. The same results hold under a slightly stronger version of (N).

EE Asymptotics

- For general EE, the conditions for consistency are also expressed in terms of the information matrices (both model based and the "true" information). For details, see Crowder (1986).

- The asymptotic distribution for the root of an estimating equation can be derived based on Taylor's expansion of the estimating function:

$$\begin{aligned} 0 &= S_n(\hat{\theta}) = \sum_{i=1}^{n} S_i(\hat{\theta}) \\ &\approx S_n(\theta_0) + S_n^{'}(\theta_0)(\hat{\theta} - \theta_0) \\ &= \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} S_i(\theta_0) + \frac{1}{n}S_n^{'}(\theta_0)\sqrt{n}(\hat{\theta} - \theta_0). \end{aligned}$$

Here $\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} S_i(\theta_0) \xrightarrow{d} N(0, B)$ by C.L.T. and $-\frac{1}{n}S_n^{'}(\theta_0) \to A$ by L.L.N. Therefore,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, A^{-1}BA^{-1})$$

where $B = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} var_{\theta_0}(S_i(\theta))$ and $A = -\lim_{n\to\infty} \frac{1}{n}E_{\theta_0}(S_n^{'}(\theta_0))$.

Empirical Variance Estimators

- Recall that the EE estimator has variance with the sandwich form $A^{-1}BA^{-1}$ where

$$A = -\lim_{n\to} \frac{1}{n}\sum_{i=1}^{n} E(S_i^{'}(\theta_0)),$$

$$B = \lim_{n\to} \frac{1}{n}\sum_{i=1}^{n} E\left(S_i(\theta_0)S_i(\theta_0)^T\right).$$

The expectations are taken under the true model and they are evaluated at the true value θ_0.

- In practice we use $\hat{A}^{-1}\hat{B}\hat{A}^{-1}$ to estimate the variance where θ^* is substituted with $\hat{\theta}_n$ and do not take the expectations.

- The empirical variance estimator uses

$$\hat{A} = -\frac{1}{n}\sum_{i=1}^{n} S_i'(\hat{\theta}),$$

$$\hat{B} = \frac{1}{n}\sum_{i=1}^{n} S_i(\hat{\theta})S_i(\hat{\theta})^T.$$

- Under mild regularity conditions, this leads to a consistent variance estimate for $\hat{\theta}$ without adopting an explicit variance model.

- For small or moderate samples we may choose to adopt a variance model for efficient variance estimates.

- $\hat{A}^{-1}$ would be the model-based variance estimator because if the model is correct ($A = B$).