STAT 770 Oct. 14 Lectures

Estimating Equations Without Likelihoods and some HW-related Topics

Reading and Topics for this lecture: Secs. 4.5-4.7, Ch. 5

(1) Using GLM Estimating Eq'n without the Likelihood!

(2) Other Model Examples: Noncanonical, Dispersion

(3) Topics for the HW: profile likelihood CI (Ch. 3, p. 80), and Fisher Scoring (Sec. 4.6)

(4) Logistic 'Model-Building' **(Ch. 5 material)**

(5) More on fitting `R` models using `glm`

Estimating Equation for GLM

Now work backwards. **Assume Y_i conditionally independent given X_i with (conditional) means μ_i satisfying $\mu_i = g(X_i' \beta)$, with β the same for $1 \le i \le n$ and g known.**

Assume (for now) that Var$(Y_i \,|\, X_i) = v(\mu_i)$, with $v(\cdot)$ known.

Idea: estimate β as the solution of

$$\sum_{i=1}^{n} X_i \frac{Y_i - g^{-1}(X_i^{tr}\beta)}{g'(g^{-1}(X_i^{tr}\beta))\, v(g^{-1}(X_i^{tr}\beta))} = \mathbf{0}$$

by writing $\mu_i = g^{-1}(X_i^{tr}\beta)$.

Recall $g^{-1} =$ logit for Logistic Regression, log for Poisson.

General Idea of Estimating Equations

Suppose *iid* (Y_i, X_i) satisfy $E_{\beta_0}\Big[\sum_{i=1}^n Q(Y_i, X_i, \beta_0)\Big] \equiv 0$ in model P_β with parameter $\beta = \beta_0$ (maybe + other nuisance parameters)

Law of Large Numbers implies

$$n^{-1} \sum_{i=1}^n Q(Y_i, X_i, \beta) \xrightarrow{P_\beta} 0$$

Assume regularity conditions (smoothness and moments) on Q

as in MLE-theory special case $Q(Y_i, X_i, \beta) = \nabla_\beta \log f(Y_i \,|\, X_i, \beta)$):

Q continuously differentiable in β, with the matrix

$E_\beta\Big(\sum_{i=1}^n \nabla_\beta \{Q(Y_1, X_1, \beta)\}^{tr}\Big)$ **nonsingular**

Further Steps in Estimating Equation Theory

$M(\beta) = n^{-1} \sum_{i=1}^n Q(Y_i, X_i, \beta)$ is random function for $\beta \in B_\epsilon(\beta_0)$

Under regularity conditions, uniformly $\approx E_{\beta_0}\Big(\frac{1}{n} \sum_{i=1}^n Q(Y_i, X_i, \beta)\Big)$ and

$\frac{1}{n} \sum_{i=1}^n \nabla\{Q(Y_i, X_i, \beta)\}^{tr} \approx E_{\beta_0}\Big(\frac{1}{n} \sum_{i=1}^n \nabla_\beta \{Q(Y_i, X_i, \beta)\}^{tr}\Big) = A_n(\beta)$

$M(\beta)$ is $\mathbf{0}$ at $\beta = \beta_0$, and conclude (via *empirical process theory*) that with prob. $\to 1$ there is solution in $B_\epsilon(\beta_0)$.

(Uniform in n, β) Nonsingularity of $A_n(\beta)$ near β_0 implies root of $M(\cdot)$ locally unique: if $\beta^*, \tilde{\beta}$ are solutions in $B_\epsilon(\beta_0)$, then

$$\mathbf{0} = M(\beta^*) - M(\tilde{\beta}) \approx \Big(A_n(\beta_0)\Big)^{tr} (\beta^* - \tilde{\beta}) + o\Big(\|\beta^* - \tilde{\beta}\|\Big)$$

Drawing Conclusions from Variance Expressions, II

General case (under regularity conditions): $\mu_i = g^{-1}(X_i^{tr}\beta)$, and

$$\widehat{\beta} \;\text{ solves }\; \sum_{i=1}^n X_i \frac{Y_i - \mu_i}{g'(\mu_i)\, v(\mu_i)} = 0 \quad , \qquad \widehat{\mu}_i = g^{-1}(X_i^{tr}\, \widehat{\beta})$$

$$\widehat{\beta} - \beta \overset{\mathcal{D}}{\approx} \mathcal{N}\Big(\mathbf{0},\, \Big[\sum_{i=1}^n X_i\, X_i^{tr} \left((g'(\widehat{\mu}_i))^2\, v(\widehat{\mu}_i)\right)^{-1}\Big]^{-1}\Big)$$

$$\text{Variance matrix} \quad = \quad \text{Information}^{-1} = \left(\mathbf{X}^{tr}\, W\, \mathbf{X}\right)^{-1}$$

$\mathbf{X}_{n\times p}$ has i'th row X_i , $\quad W_{n\times n} = \mathtt{diag}\Big(\big[g'(\widehat{\mu}_i))^2\, v(\widehat{\mu}_i)\big]^{-1}\Big)$

With canonical link: $J = \mathcal{I}\Big|_{\beta=\widehat{\beta}}$ and $W = \mathtt{diag}(v(\widehat{\mu}_i))$

Estimating Equation Interpretation (Sec. 4.7)

We just saw that* theory tells:

$\widehat{\beta}$ solving the Estimating Equation $\sum_{i=1}^{n} X_i \frac{Y_i - \mu_i}{g'(\mu_i)\, v(\mu_i)} = 0$

is asymptotically normal with variance $(\mathbf{X}^{tr} W \mathbf{X})^{-1}$ assuming only that Y_i are independent with (conditional given X_i) mean and variance $\mu_i = g^{-1}(X_i^{tr}\, \beta)$, $v(\mu_i)$.

Similar theory shows that solving $\sum_{i=1}^{n} h(\mu_i)\, X_i\, (Y_i - \mu_i)$ gives $\sqrt{n}$ consistent asymptotically normal estimator (**like weighted least squares!**) without the assumption on $v(\mu_i)$, but estimator is generally **not efficient**, and the variance expression is different.

Other GLMs and Extensions

Non-canonical Link Examples:

(I). Binomial outcome Y_i, $\mu_i \in (0,1)$, link g^{-1} any dist'n function other than `plogis`, e.g. $g^{-1} = \Phi$ for probit model.

(II). Poisson outcome Y_i, link g^{-1} any monotone map on $\mathbb{R}$, can be identity, e.g. for linear model, if $X_i'\beta$ all positive

Models with Overdispersion: will cover this extension next time

Profile Likelihood and Confidence Intervals

This is material from Ch. 3, p.80.

Let $\beta = (\gamma, \lambda)$ be the parameter (eg in a GLM) with MLE $\widehat{\beta}$ and restricted MLE $\widehat{\lambda}_r(\gamma_0)$ calculated under hypothesis $H_0 : \gamma = \gamma_0$

Then $2\Big[\log L(\widehat{\beta}) - \log L(\gamma_0, \widehat{\lambda}_r(\gamma_0))\Big] \sim \chi^2_{\text{dim}(\gamma)}$ (Wilks Thm)

inverted LRT Conf. Interval: $\Big\{\gamma_0 : -2 \log L_{prof}(\gamma_0) \leq \chi^2_{d,\alpha}\Big\}$

where $L_{prof}(\gamma_0) \equiv L(\gamma_0, \widehat{\lambda}_r(\gamma_0))/L(\widehat{\beta})$ (Profile Likelihood)

Can calculate the test-based CI's using `confint` in `R`.

Numerical Maximization and Fisher Scoring

$L(\beta)$ usually maximized by Newton-Raphson (NR) Iterations

to solve $\nabla \log L(\beta) = 0$

$$\boxed{\beta_{k+1} = \beta_k + \left\{ -\nabla^{\otimes 2} \log L(\beta_k) \right\}^{-1} \nabla \log L(\beta_k)}$$

Recall Observed Information $J = -\nabla^{\otimes 2} \log L(\widehat{\beta})$

So $\{ \cdot \}$ matrix in NR is a current-iterate version of J

Fisher Scoring uses iterates with Fisher Info matrix:

$$\beta_{k+1} = \beta_k + \mathcal{I}\Big|_{\beta=\beta_k} \nabla \log L(\beta_k)$$

Recall that $\mathcal{I}(\widehat{\beta}) = J$ [only] in canonical-link models

R Script with Illustrations of Methods

(i) Model-building: use of Deviances and
Standardized Coefficients in `glm`

(ii) Profile Likelihoods and `confint`

(iii) Likelihood Maximization and Fisher Scoring