Generalized Linear Models

MIT 18.655

Dr. Kempthorne

Spring 2016

Outline

1 Generalized Linear Models
- Linear Predictors and Link Functions
- Maximum Likelihood Estimation
- Logistic Regression for Binary Responses
- Likelihood Ratio Tests
- Vector Generalized Linear Models

Generalized Linear Model

Data: $(y_i, \mathbf{x}_i), i = 1, \ldots, n$ where

y_i: response variable

$\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})^T$: p explanatory variables

Linear predictor: For $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p) \in R^p$:

$\mathbf{x}_i\boldsymbol{\beta} = \sum_{j=1}^{p} x_{i,j}\beta_j$

Probability Model: $\{y_i\}$ independent, canonical exponential r.v.'s:

- Density: $p(y_i \mid \eta_i) = e^{\eta_i y_i - A(\eta_i)} h(x)$
- Mean Function: $\mu_i = E[Y_i] = \dot{A}(\eta_i)$
- Link Function $g(\cdot)$: $g(\mu_i) = \mathbf{x}_i\boldsymbol{\beta}$

  With estimate $\hat{\boldsymbol{\beta}}$: $\mathbf{x}_i\hat{\boldsymbol{\beta}} = \hat{g}(\mu_i)$.

Canonical Link Function: $g(\mu_i) = \eta_i = [\dot{A}]^{-1}(\mu_i) = \mathbf{x}_i^T\beta$

Matrix Notation

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{p,n} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$E[\mathbf{y}] = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} \dot{A}(\eta_1) \\ \dot{A}(\eta_2) \\ \vdots \\ \dot{A}(\eta_n) \end{bmatrix} = \dot{A}(\boldsymbol{\eta})$$

Examples:

$$\begin{aligned} y_i &\sim Bernoulli(\theta_i): & \eta_i &= log(\tfrac{\theta_i}{1-\theta_i}) \\ y_i &\sim Poisson(\lambda_i): & \eta_i &= log(\lambda_i) \\ y_i &\sim Gaussian(\mu_i, 1): & \eta_i &= \mu_i \end{aligned}$$

Outline

1. Generalized Linear Models
   - Linear Predictors and Link Functions
   - Maximum Likelihood Estimation
   - Logistic Regression for Binary Responses
   - Likelihood Ratio Tests
   - Vector Generalized Linear Models

Log-Likelihood Function for Generalized Linear Model

$$\begin{aligned}
\ell(\beta) &= \sum_{i=1}^{n}(\eta_i y_i - A(\eta_i) + log[h(y_i)]) \\
&\propto \boldsymbol{\eta}^T \mathbf{y} - \mathbf{1}^T A(\boldsymbol{\eta}) \\
&= g(\boldsymbol{\mu})^T \mathbf{y} - \mathbf{1}^T A(\boldsymbol{\eta}) \\
&= [\mathbf{X}\beta]^T \mathbf{y} - \mathbf{1}^T A(\boldsymbol{\eta}) \text{ (for Canonical Link)}
\end{aligned}$$

Note: $T(\mathbf{y}) = \mathbf{X}^T\mathbf{y}$ is sufficient when $g(\cdot)$ is canonical

Maximum Likelihood Estimation of β

Solve for $\{\boldsymbol{\beta}_m, m = 1, 2, \ldots\}$ iteratively:

$$\begin{aligned}
0 = \dot{\ell}(\boldsymbol{\beta}_{m+1}) &= \dot{\ell}(\boldsymbol{\beta}_m) + (\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_m)\ddot{\ell}(\boldsymbol{\beta}_m) \\
\Longrightarrow \quad \boldsymbol{\beta}_{m+1} &= \boldsymbol{\beta}_m + [-\ddot{\ell}(\boldsymbol{\beta}_m)]^{-1}\dot{\ell}(\boldsymbol{\beta}_m)
\end{aligned}$$

"Fisher Scoring Algorithm" $\Longleftrightarrow$ Newton-Raphson

$$\boldsymbol{\beta}_m \xrightarrow{Pr} \hat{\boldsymbol{\beta}} \text{ (the MLE)}$$

For Canonical Link

$$
\begin{aligned}
\dot{\ell}(\beta) &= \frac{\partial}{\partial \boldsymbol{\beta}}[\textstyle\sum_{i=1}^n(\eta_i y_i - A(\eta_i) + log[h(y_i)])] \\
&= \textstyle\sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\beta}}[(\eta_i y_i - A(\eta_i) + log[h(y_i)])] \\
&= \textstyle\sum_{i=1}^n \left(\frac{\partial}{\partial \boldsymbol{\eta}_i}[(\eta_i y_i - A(\eta_i) + log[h(y_i)])]\right)\left(\frac{\partial \eta_i}{\partial \boldsymbol{\beta}}\right) \\
&= \textstyle\sum_{i=1}^n \left(y_i - \dot{A}(\eta_i)\right)\left(\frac{\partial \mathbf{x}_i^T \boldsymbol{\beta}}{\partial \boldsymbol{\beta}}\right) \\
&= \textstyle\sum_{i=1}^n (y_i - \mu_i)\,\mathbf{x}_i = \sum_{i=1}^n \mathbf{x}_i\,(y_i - \mu_i) = [\mathbf{X}]^T(\mathbf{y} - \boldsymbol{\mu}) \\
\ddot{\ell}(\beta) &= \frac{\partial}{\partial \boldsymbol{\beta}^T}[\dot{\ell}(\boldsymbol{\beta})] = \frac{\partial}{\partial \boldsymbol{\beta}^T}[\textstyle\sum_{i=1}^n \left(y_i - \dot{A}(\eta_i)\right)\left(\frac{\partial \mathbf{x}_i^T \boldsymbol{\beta}}{\partial \boldsymbol{\beta}}\right)] \\
&= \frac{\partial}{\partial \boldsymbol{\beta}^T}[\textstyle\sum_{i=1}^n \left(y_i - \dot{A}(\eta_i)\right)\mathbf{x}_i] \\
&= \textstyle\sum_{i=1}^n \left(-\frac{\partial}{\partial \boldsymbol{\beta}^T}[\dot{A}(\eta_i)]\right)\mathbf{x}_i] \\
&= \textstyle\sum_{i=1}^n \left(-\frac{\partial}{\partial \eta_i}[\dot{A}(\eta_i)]\frac{\partial \eta_i}{\partial \boldsymbol{\beta}^T}\right)\mathbf{x}_i] \\
&= \textstyle\sum_{i=1}^n \left(-[\ddot{A}(\eta_i)]\right)\mathbf{x}_i]\frac{\partial \eta_i}{\partial \boldsymbol{\beta}^T} = \sum_{i=1}^n \left(-[\ddot{A}(\eta_i)]\right)\mathbf{x}_i \mathbf{x}_i^T \\
&= \mathbf{X}^T \mathbf{W} \mathbf{X}
\end{aligned}
$$

For Canonical Link

$$\begin{aligned}\dot{\ell}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}}[\textstyle\sum_{i=1}^{n}(\eta_i y_i - A(\eta_i) + log[h(y_i)])] \\ &= \mathbf{X}^T(\mathbf{y} - \boldsymbol{\mu})\end{aligned}$$

$$\begin{aligned}\ddot{\ell}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}^T}[\dot{\ell}(\boldsymbol{\beta})] = \frac{\partial}{\partial \boldsymbol{\beta}^T}[\textstyle\sum_{i=1}^{n}\left(y_i - \dot{A}(\eta_i)\right)\left(\frac{\partial \mathbf{x}_i^T \boldsymbol{\beta}}{\partial \boldsymbol{\beta}}\right)] \\ &= \textstyle\sum_{i=1}^{n}\left(-[\ddot{A}(\eta_i)]\right)\mathbf{x}_i \mathbf{x}_i^T \\ &= \mathbf{X}^T \mathbf{W} \mathbf{X}\end{aligned}$$

where $\mathbf{W} = Cov(\mathbf{y})$ is diagonal with $\mathbf{W}_{i,i} = \ddot{A}(\eta_i) = Var[y_i]$

Iteratively Re-weighted Least Squares Interpretation

Given $\boldsymbol{\beta}_m$, and $\hat{\boldsymbol{\mu}}(\boldsymbol{\beta}_m) = \dot{A}(\mathbf{X}\boldsymbol{\beta}_m)$,

$$\begin{aligned}\boldsymbol{\beta}_{m+1} &= \boldsymbol{\beta}_m + [-\ddot{\ell}(\boldsymbol{\beta}_m)]^{-1}\dot{\ell}(\boldsymbol{\beta}_m) \\ &= \boldsymbol{\beta}_m + [\mathbf{X}^T\mathbf{W}\mathbf{X}]^{-1}[\mathbf{X}]^T(\mathbf{y} - \hat{\boldsymbol{\mu}}(\boldsymbol{\beta}_m))\end{aligned}$$

$\Delta = (\boldsymbol{\beta}_{m+1} - \boldsymbol{\beta}_m)$ is solved as the WLS regression of

$$\mathbf{y}_* = [\mathbf{y} - \hat{\boldsymbol{\mu}}(\boldsymbol{\beta}_m)]$$

on $\mathbf{X}_* = \mathbf{W}\mathbf{X}$

using $\boldsymbol{\Sigma}_* = Cov(\mathbf{y}_*) = \mathbf{W}$

The WLS estimate of $\boldsymbol{\Delta}$ is given by:

$$\begin{aligned}\hat{\boldsymbol{\Delta}} &= [\mathbf{X}_*\boldsymbol{\Sigma}_*^{-1}\mathbf{X}_*]^{-1}\mathbf{X}_*^T\boldsymbol{\Sigma}_*^{-1}\mathbf{y}_* \\ &= [\mathbf{X}^T\mathbf{W}\mathbf{W}^{-1}\mathbf{W}\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{W}\mathbf{W}^{-1}[\mathbf{y} - \hat{\boldsymbol{\mu}}(\boldsymbol{\beta}_m)] \\ &= [\mathbf{X}^T\mathbf{W}\mathbf{X}]^{-1}\mathbf{X}^T[\mathbf{y} - \hat{\boldsymbol{\mu}}(\boldsymbol{\beta}_m)]\end{aligned}$$

Outline

1 Generalized Linear Models
- Linear Predictors and Link Functions
- Maximum Likelihood Estimation
- Logistic Regression for Binary Responses
- Likelihood Ratio Tests
- Vector Generalized Linear Models

Logistic Regression for Binary Responses

Binomial Data: $Y_i \sim Binomial(m_i, \pi_i),\ i = 1, \ldots, k.$

Log-Likelihood Function

$$\ell(\pi_1, \ldots, \pi_k) = \prod_{i=1}^{k}[(Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right)) + m_i \log(1-\pi_i)]$$

Covariates: $\{\mathbf{x}_i, i = 1, \ldots, k\}$

Logistic Regression Parameter:

$$\eta_i = \log[\pi_i/(1 - \pi - i)] = \mathbf{x}_i^T \beta$$

$$\mathbf{W} = Cov(\mathbf{Y}) = diag(m_i \pi_i (1-\pi_i)),\ (k \times k \text{ matrix})$$

Outline

1. Generalized Linear Models
   - Linear Predictors and Link Functions
   - Maximum Likelihood Estimation
   - Logistic Regression for Binary Responses
   - Likelihood Ratio Tests
   - Vector Generalized Linear Models

Likelihood Ratio Tests for Generalized Linear Models

Consider testing

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0 = \dot{A}(\mathbf{X}\boldsymbol{\beta}_0).$$

vs

$$H_{Alt} : \boldsymbol{\mu} = \boldsymbol{\mu}_* \text{ (general } n\text{-vector)}$$
$$\text{e.g., } \boldsymbol{\mu}_* = \dot{A}(\mathbf{X}_*\boldsymbol{\beta}_*) \text{ with } \mathbf{X}_* = I_n \text{ and } \boldsymbol{\beta}_* = \boldsymbol{\eta}.$$

Suppose $\mathbf{y}$ is in interior of convex support of $\{\mathbf{y} : p(\mathbf{y} \mid \boldsymbol{\eta}) > 0\}$.

Then $\hat{\boldsymbol{\eta}} = \eta(\mathbf{y}) = \dot{A}^{-1}(\mathbf{y})$ is the MLE under H_{Alt}

The Likelihood Ratio Test Statistic of H_0 vs H_{Alt}:

$$\begin{aligned} 2\log\lambda &= 2[\ell(\eta(Y)) - \ell(\eta(\boldsymbol{\mu}_0))] \\ &= 2\left([\eta(\mathbf{y}) - \eta(\boldsymbol{\mu}_0)]^T\mathbf{y} - [A(\boldsymbol{\eta}(\mathbf{y})) - A(\eta(\boldsymbol{\mu}_0))]\right) \\ &= \text{"}\textbf{Deviance}\text{" Between } \mathbf{y} \text{ and } \boldsymbol{\mu}_0 \end{aligned}$$

Deviance Formulas for Distributions

$$\begin{array}{ll} \textit{Gaussian}: & \sum_i (y_i - \mu_i)^2/\sigma_0^2 \\ \textit{Poisson}: & 2\sum_i [y_i \log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)] \\ \textit{Binomial}: & 2\sum_i [y_i \log(y_i/\hat{\mu}_i) + (m_i - y_i)\log[(m_i - y_i)/(m_i - \hat{\mu}_i)] \end{array}$$

Outline

1. Generalized Linear Models
   - Linear Predictors and Link Functions
   - Maximum Likelihood Estimation
   - Logistic Regression for Binary Responses
   - Likelihood Ratio Tests
   - Vector Generalized Linear Models

Vector Generalized Linear Models

Data: $(\mathbf{y}_i, \mathbf{x}_i), i = 1, \ldots, n$ where

$\mathbf{y}_i$: a q-dimensional response vector variable

$\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})^T$: p explanatory variables

Probability Model: The conditional distributions of each $\mathbf{y}_i$ given $\mathbf{x}_i$ is of the form

$$p(\mathbf{y} \mid \mathbf{x}; \mathbf{B}, \phi) = f(\mathbf{y}, \eta_1, \ldots, \eta_M, \phi)$$

for some known function $f(\cdot)$, where $\mathbf{B} = [\boldsymbol{\beta}_1 \boldsymbol{\beta}_2 \cdots \boldsymbol{\beta}_M]$ is a $p \times M$ matrix of unknown regression coefficients.

M Linear Predictors: For $j = 1, \ldots, M$, the jth linear predictor is

$$\eta_j = \eta_j(\mathbf{x}) = \boldsymbol{\beta}_j^T \mathbf{x} = \sum_{k=1}^{p} B_{kj} x_k,$$

where $\mathbf{x} = (x_1, \ldots, x_p)^T$ with $x_1 = 1$ when there is an intercept.

Matrix Notation

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^T \\ \mathbf{y}_2^T \\ \vdots \\ \mathbf{y}_n^T \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{p,n} \end{bmatrix}$$

$$\mathbf{B} = [\boldsymbol{\beta}_1 | \boldsymbol{\beta}_2 | \cdots \boldsymbol{\beta}_M]$$

Link Function: For each observation i

$$E[\mathbf{y}_i] = \boldsymbol{\mu}_i \qquad (q \times 1 \text{ vectors}) \text{ and}$$
$$g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i = \mathbf{B}^T \mathbf{x}_i \qquad (M \times 1 \text{ vectors})$$

where:

$$\boldsymbol{\eta}_i = \boldsymbol{\eta}(\mathbf{x}_i) = \begin{pmatrix} \eta_1(\mathbf{x}_i) \\ \vdots \\ \eta_M(\mathbf{x}_i) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_1^T \mathbf{x}_i \\ \vdots \\ \boldsymbol{\beta}_m^T \mathbf{x}_i \end{pmatrix} = \mathbf{B}^T \mathbf{x}_i$$

Multivariate Exponential Family Models

- Density: $p(\mathbf{y}_i \mid \boldsymbol{\eta}_i) = e^{\boldsymbol{\eta}_i^T \mathbf{y}_i - A(\boldsymbol{\eta}_i)} h(\mathbf{y}_i)$
- Mean Function: $\boldsymbol{\mu}_i = E[\mathbf{y}_i] = \dot{A}(\boldsymbol{\eta}_i)$
- Link Function $g(\cdot)$: $g(\boldsymbol{\mu}_i) = \mathbf{B}^T \mathbf{x}_i$

Canonical Link Function: $g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i = [\dot{A}]^{-1}(\boldsymbol{\mu}_i) = \mathbf{B}^T \mathbf{x}_i$

Examples:

$$\mathbf{y}_i \sim Multinomial(m_i, \pi_{i,1}, \ldots, \pi_{i,M+1})$$
$$\pi_{i,j} \geq 0,\ j = 1, \ldots, M+1 \text{ and } \textstyle\sum_{j=1}^{M+1} \pi_{i,j} = 1.$$

$$\mathbf{y}_i \sim \text{M-Variate-Gaussian}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_0)$$
$$\boldsymbol{\mu}_i \in R^M \text{ and } \boldsymbol{\Sigma}_0 \text{ known } (M \times M)$$

Case Study: Applying Generalized Linear Models.

Note: Reference Yee (2010) on the VGAM Package for R

18.655 Mathematical Statistics
Spring 2016