Generalized Linear Models

STAT 526

Professor Olga Vitek

April 20, 2011

Specifying a Generalized Linear Model

Exponential Family of Distributions (EFD)

- A *exponential family distribution* has the probability mass/distribution function in the form of

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

  - θ *canonical parameter* representing location (also called *natural parameter*)
  - ϕ *dispersion parameter* representing the scale
  - $a(\cdot), b(\cdot), c(\cdot)$ known functions

- Usually, $a(\phi) = \phi/w$
  - w a known weight, varies between observations
  - ϕ can be known (one-parameter distribution) or unknown (two-parameter distribution)

Examples

<u>*Example:* $y_i \sim N(\mu_i, \sigma^2)$</u>

- $f(y_i) = \exp\left\{\frac{y_i\mu_i - \mu_i^2/2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\}$
- $\theta_i = \mu_i$, $\phi = \sigma^2$, $w_i = 1$.
- $a(\phi) = \phi$, $b(\theta_i) = \theta_i^2/2$, and
  $c(y_i, \phi) = -\frac{y_i^2}{2\phi} - \frac{1}{2}\log(2\pi\phi)$.

<u>*Example:* $y_i \sim Poisson(\lambda_i)$</u>

- $f(y_i) = \exp\{y_i \log(\lambda_i) - \lambda_i - \log(y_i!)\}$
- $\theta_i = \log(\lambda_i)$, $\phi = 1$, $w_i = 1$.
- $a(\phi) = 1$, $b(\theta_i) = \exp\{\theta_i\}$, and $c(y_i, \phi) = -\log(y_i!)$.

<u>*Example:* $y_i \sim Binomial(n_i, \pi_i)$</u>

- $f_i(\bar{y}) = \exp\left\{\frac{\bar{y}\log\frac{\pi_i}{1-\pi_i} + \log(1-\pi_i)}{1/n_i} + \log\begin{pmatrix} n_i \\ n_i\bar{y}\end{pmatrix}\right\}$
- $\theta_i = \log\frac{\pi_i}{1-\pi_i}$, $\phi = 1$, $w_i = n_i$.
- $a(\phi) = 1/n_i$, $b(\theta) = log[1 + exp(\theta)]$,
  $c(y, \phi) = \log\begin{pmatrix} n_i \\ n_i\bar{y}\end{pmatrix}$

Expected Value of EFD

- Assume $Y \sim EFD(\theta, \phi)$
  - Distribution $f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$
  - Log-likelihood $l(\theta, \phi; y) = log f(y; \theta, \phi)$

- $E\{Y\} = b'(\theta)$:
  - Since $\int f \partial y = 1$, under regularity conditions:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \theta} \int f \, \partial y = \int \frac{\partial}{\partial \theta} f \, \partial y \\
&= \int \frac{y - b'(\theta)}{a(\phi)} f \, \partial y = \frac{1}{a(\phi)} \left[\int y f \, dy - \int b'(\theta) f \, dy \right] \\
&= \frac{1}{a(\phi)} \left[E\{y\} - b'(\theta) \right]
\end{aligned}$$

Variance of EFD

- Assume $Y \sim EFD(\theta, \phi)$
  - Distribution $f(y; \theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\}$
  - Log-likelihood $l(\theta, \phi; y) = log f(y; \theta, \phi)$

- $Var\{Y\} = b''(\theta)a(\phi)$:
  - Since $\int f \partial y = 1$, under regularity conditions:

$$
\begin{aligned}
0 &= \frac{\partial^2}{\partial\theta^2}\int f\,\partial y = \int \frac{\partial^2}{\partial\theta^2} f\,\partial y \\
&= \int \frac{-b''(\theta)}{a(\phi)} f\,\partial y + \int \frac{[y - b'(\theta)]^2}{a(\phi)^2} f\,\partial y \\
&= \frac{-b''(\theta)}{a(\theta)} + \frac{Var\{y\}}{a(\phi)^2} \\
&= \frac{-b''(\theta)a(\phi) + Var\{y\}}{a(\phi)^2}
\end{aligned}
$$

Generalized Linear Models

- Data: $(y_i, \mathbf{x}_i) = (\ y_i, x_{i1}, x_{i2}, \cdots, x_{i,p-1}\ )$, $i = 1, 2, \cdots, n$

- Random component: $y_i \mid \mathbf{x}_i \overset{ind}{\sim} EFD(\theta_i, \phi)$
  - Counts (Poisson, Bernouilli, Binomial) or continuous (Gamma, Inverse Gaussian)
  - Assumptions: independent observations (exclude time series and spacial models)

- Goal: Model $\mu_i = E\{y_i|\mathbf{x}_i\}$

- Systematic component: Joint effects of $\mathbf{x}_i$ through their linear combination $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} = \mathbf{x}_i' \beta$

- Link function: Function $g(\mu_i)$ that links $\mu_i = E\{y_i\}$ and $\eta_i = \mathbf{x}_i' \beta$
  - $g(\mu_i) = \eta_i$

Link Functions

- The link function $g(\mu_i)$ defines a specific probability model.
  - Logistic regression:
    $g(\mu_i) = logit(\mu_i) = log(\frac{\mu_i}{1-\mu_i}) \overset{model}{=} \mathbf{x}_i'\beta$
  - Probit regression:
    $g(\mu_i) = \Phi^{-1}(\mu_i) \overset{model}{=} \mathbf{x}_i'\beta$
- The link function also defines the mean function $g^{-1}(\mathbf{x}_i'\beta)$.
  - Logistic regression: $\mu \overset{model}{=} \frac{1}{1+exp(-\mathbf{x}_i'\beta)}$
  - Probit regression: $\mu \overset{model}{=} \Phi(\mathbf{x}_i'\beta)$
- If specify $\theta \overset{model}{=} \mathbf{x}_i'\beta$, i.e. $g(\mu_i) = \theta_i = \eta_i$, $g(\mu_i)$ is the *canonical link*.
  - Remember that in EFD $\mu_i = b'(\theta_i)$
  - With the canonical link, $g(\mu_i) = g(b'(\theta_i)) = \theta_i$
  - Therefore $g(\cdot)$ is the inverse function of $b'(\cdot)$

GLMs with Canonical Links

	Normal	Poisson	Binomial
Notation	$N(\mu, \sigma^2)$	$P(\lambda)$	$B(n, \pi)$
Range of y	$(-\infty, \infty)$	$0 : \infty$	$0 : n$
ϕ	σ^2	1	1
$b(\theta)$	$\theta^2/2$	e^θ	$\log(1 + e^\theta)$
Expected value $\mu(\theta)$	θ	e^θ	$\frac{e^\theta}{1+e^\theta}$
Canonical link $\theta = g(\mu)$	identity	log	logit
Variance function $V(\mu)$	1	μ	$\mu(1-\mu)$

- Specifying distribution in R

```
    glm(formula, family = gaussian, data,...)
```

```
family = binomial(link = "logit")
family = gaussian(link = "identity")
family = Gamma(link = "inverse")
family = inverse.gaussian(link = "1/mu^2")
family = poisson(link = "log")
```

Fitting a GLM and Assessing the Quality of Fit

Likelihood Equations

- Log-likelihood:

$$L(\beta) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi)$$

  - θ depends on model parameters β

- Likelihood equations (Agresti Sec. 4.4.4)

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{var(y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0,\ j = 1, \ldots, p$$

  - depends on β through $\mu_i = g^{-1}(\mathbf{x}_i'\beta)$
  - depends on distr. of y_i through μ_i and $var(y_i)$

- Using the canonical link $\theta = g(\mu) \overset{model}{=} \mathbf{x}_i'\beta$:

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{var(y_i)} b'' x_{ij} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{a(\phi)} = 0$$

  If $a(\phi)$ same for all i : $\sum_{i=1}^{n} x_{ij}y_i - \sum_{i=1}^{n} x_{ij}\mu_i = 0$

  - unifies several model fitting algorithms
  - Normal equations for Normal distribution

Iterative Weighted Least Squares: The Algorithm

- Transform nonlinear optimization problem into a series of (weighted) least squares fits

- Step 1: Given a working value $\widehat{\beta}^{(k)}$
  - Calculate $\widehat{\mu}_i^{(k)} = g^{-1}(\ \mathbf{x}_i'\widehat{\beta}^{(k)}\ )$

- Step 2: Approximate $g(y_i)$ by its lineariza-
tion
in the neighborhood around $\widehat{\mu}_i^{(k)}$
  - $g(y_i) \approx z_i^{(k)} = g(\widehat{\mu}_i^{(k)}) + (y_i - \widehat{\mu}_i^{(k)}) g'(\widehat{\mu}_i^{(k)})$
  - Note that $Var\{z_i^{(k)}\} = [g'(\widehat{\mu}_i^{(k)})]^2\ Var\{y_i\}_{\widehat{\mu}_i^{(k)}}$

    — subscript $\widehat{\mu}_i^{(k)}$ means
    "variance evaluated at" $\widehat{\mu}_i^{(k)}$

Iterative Weighted Least Squares: The Algorithm

- Step 3: Estimate $\widehat{\beta}^{k+1}$
  - Use linear regression model $z_i = \mathbf{x}_i'\beta + \epsilon_i$
  - $E\{\epsilon_i\} = 0$
  - $Var\{\epsilon_i\} = \phi\ Var\{z_i^{(k)}\}$
  - $\widehat{\beta}^{(k+1)} = (X'WX)^{-1}X'WZ$

    where
    $Z = (z_1, \cdots, z_n)'$,
    $X = (\mathbf{x}_1', \cdots, \mathbf{x}_n')$,
    $W = \text{diag}\left\{Var\{\epsilon_1\}^{-1}, \cdots, Var\{\epsilon_n\}^{-1}\right\}$.

- Calculate $\widehat{\beta}^{(k+1)}$ iteratively until it converges to $\widehat{\beta}$.

Measure of Goodness of Fit: Deviance

- Current GLM, exponential family:

  $\theta = \theta(\mu)$; $\widehat{\mu} = g^{-1}(\mathbf{x}_i'\widehat{\beta})$; $\rightarrow$ $\widehat{\theta} = \theta(\widehat{\mu})$

  - log-likelihood $l(\widehat{\beta}; y, \phi) = \sum_{i=1}^{n} w_i \frac{y_i\widehat{\theta}_i - b(\widehat{\theta}_i)}{\phi}$

- Saturated model, n parameters:

  $\theta = \theta(\mu)$; $\tilde{\mu} = y_i$; $\rightarrow$ $\tilde{\theta} = \theta(\tilde{\mu})$

  - log-likelihood $l(y; y, \phi) = \sum_{i=1}^{n} w_i \frac{y_i\widehat{\theta}_i - b(\widehat{\theta}_i)}{\phi}$

- The deviance of the current GLM (called `residual deviance` in R):

$$\begin{aligned} D(\widehat{\beta}) &= 2\phi\ \{l(y; y, \phi) - l(\widehat{\beta}; y, \phi)\} \\ &= \sum_{i=1}^{n} 2w_i\{\ y_i(\tilde{\theta}_i - \widehat{\theta}_i) - b(\tilde{\theta}_i) + b(\widehat{\theta}_i)\ \} \end{aligned}$$

  - $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, $\rightarrow$ residual SS

Measure of Goodness of Fit: Generalized Pearson X^2

- Generalized Pearson X^2:

  $$X^2 = \sum_{i=1}^{n} (y_i - \widehat{\mu}_i)^2 / V(\widehat{\mu}_i)$$

  - $V(\widehat{\mu}_i) = b''(\widehat{\theta})$ is the variance function
  - $y_i \sim \mathcal{N}(\mu_i, \sigma^2)$, $\rightarrow$ residual SS
  - $y_i \sim Binomial(\mu_i, n_i)$, $\rightarrow$ original Pearson X^2

- Testing the quality of fit, known ϕ:

  Scaled deviance $D/\phi \overset{assympt.}{\sim} \chi^2_{n-p}$
  Scaled $X^2/\phi \overset{assympt.}{\sim} \chi^2_{n-p}$

  - Deviance is additive for nested sets of models, when using $\widehat{\beta}_{ML}$,
    – but poor approximation of χ^2, approximation does not improve as $n \rightarrow \infty$
  - X^2 has a more direct interpretation,
    – better approximation of χ^2

Diagnostics: Residuals

- Inspired from weighted linear regression
- Use last iteration of the IWLS algorithm
  - $Z = X\beta + \epsilon, \quad E[\epsilon] = 0, \quad var(\epsilon) = \phi \; Var\{z_i^{(k)}\}$
  - $\widehat{\beta} = (X'WX)^{-1}X'WZ$
  - $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$

- Response residuals: $e_{i,R} = y_i - \widehat{\mu}_i$
  - do not have constant variance
  - Analogue to simple residuals in regression
  - In R: `residuals(glmfit, type="response")`

- Pearson residuals: $e_{i,P} = \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}}$.
  - The denominator $\neq$ the variance of the residual
  - Analogue to standardize residuals in regression
  - In R: `residuals(glmfit, type="pearson")`

Diagnostics: Standardized Pearson Residuals

- Standardized Pearson residuals:
  $e_{i,SP} = \frac{y_i - \widehat{\mu}_i}{\sqrt{\widehat{\phi}\ V(\widehat{\mu}_i)\ (1-h_{ii})}}$
  - h_{ii} is the i-th diagonal element of H
  - The denominator = the variance of the residual
  - Have constant variance and mean zero if $V(\mu)$ is correctly specified
  - Analogue to studentized residuals in regression
  - Useful for detecting variance misspecification or outlier detection

- In R: Original: `residuals(glmfit, type="pearson")`
  Standardized: `library(boot); glm.diag(glmfit)$rp`
  also see
  `glm.diag.plots(glmfit)`

Diagnostics: Deviance Residuals

- Deviance residuals:

  $e_{i,D} = sign(y_i - \widehat{\mu}_i)\sqrt{d_i}$

  – d_i is the contribution to the model deviance from the i-th observation

- Standardized deviance residuals:

  $e_{i,SD} = \frac{sign(y_i - \widehat{\mu}_i)\ \sqrt{d_i}}{\sqrt{\widehat{\phi}(1-h_{ii})}}$

  – Deviance residuals may be closer to Normal distribution (or at least less skewed) than the Pearson residuals

    * Not when y is binary!

  – When less skewed, may be better than Pearson residuals for outlier detection

- In R:
  Original: `residuals(glmfit, type="deviance")`
  Standardized: `library(boot); glm.diag(glmfit)$rd`

Diagnostics: Jacknife Residuals

- Jacknife residuals: approximated by

$$e_{i,J} = sign(y_i - \widehat{\mu}_i)\sqrt{(1 - h_{ii})e_{i,SD}^2 + h_{ii}e_{i,SP}^2}$$

  - the difference between the observed ith response, and predicted from the data without ith case
  - has an intermediate value between $e_{i,SD}$ and $e_{i,SP}$
  - usually closer to $e_{i,SD}$ than to $e_{i,SP}$, since the average value of h_{ii} is small
  - a good choice for diagnostics

- In R:
`library(boot); glm.diag(glmfit)$res`
or
`rstudent(glmfit)`

Diagnostics: Influential Points with Cook's Distance

- The Cook's distance statistics:

$$C_i = \frac{(\widehat{\beta}_{(i)} - \widehat{\beta})^T X^T W X (\widehat{\beta}_{(i)} - \widehat{\beta})}{p \ \widehat{\phi}}$$

  - $\widehat{\beta}_{(i)}$ is an estimate of β when excluding case i
  - p is the number of parameters
  - Measures the standardized change in linear predictor when the ith case is deleted
    * a standardized sum of squared $\Delta\beta$
  - Requires n maximizations
  - Can be approximated by a one-step procedure

- In R:

`library(boot); glm.diag(glmfit)$cook`
or
`cooks.distance(glmfit)`

Inference:
Testing and Prediction

Inference: Wald Test for β_j

- $g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}$

- H_0: $\beta_j = \beta_j^{H_0}$ versus H_a: $\beta_j \neq \beta_j^{H_0}$

- $\beta - \widehat{\beta} \overset{asympt.}{\sim} \mathcal{N}\left(\mathbf{0}, I(\widehat{\beta})^{-1}\right)$
  - $I(\beta)$ is the Fisher Information matrix $\left[-\frac{\partial^2 l(\beta,\phi;\mathbf{y})}{\partial\beta_i \partial\beta_j}\right]_{ij}$
  - $I(\widehat{\beta})$ denoted the matrix evaluated at $\beta = \widehat{\beta}$

- Test statistic $z = \frac{\widehat{\beta}_j - \beta_j^{H_0}}{se(\widehat{\beta}_j)} \overset{asympt.}{\sim} N(0,1)$
  - Based on asymptotic normality of the MLE

- Confidence interval for β_j: $\widehat{\beta}_j \pm z^{1-\alpha/2} SE\{\widehat{\beta}_j\}$

- Multivariate extension

$$W = (\widehat{\beta} - \beta_0)' \left[Cov(\widehat{\beta})\right]^{-1} (\widehat{\beta} - \beta_0) \overset{asympt}{\sim} \chi^2_{df}$$

  - df is the rank of $Cov(\widehat{\beta})$
    (i.e. # of nonredundant parameters in β)

Inference: Score Test

- $g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}$
- H_0: $\beta_j = \beta_j^{H_0}$ versus H_a: $\beta_j \neq \beta_j^{H_0}$
- Utilizes the score function $u(\beta) = \partial L(\beta)/\partial(\beta)$ evaluated at β_0
  - $|u(\beta)|$ is larger when β is further from β_0
- Test statistic: ratio of $u(\beta)$ to its SE evaluated under H_0:

$$z = \sqrt{\frac{\left[\partial L(\beta)/\partial\beta_0\right]^2}{-E\left[\partial^2 L(\beta)/\partial\beta_0^2\right]}} \overset{H_0,\ asympt.}{\sim} N(0,1)$$

- Multivariate extension
  - z^2 is a quadratic form based on $\partial^2 L(\beta)/\partial\beta_j\partial\beta_{j'}$ and the inverse of the Information matrix evaluated at β_0, compared to χ^2

Inference: Likelihood Ratio Test for Nested Models

- $g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}$
- $H_0: \ \beta_1 = \beta_2 = \cdots = \beta_q = 0$
  versus H_a: not all $\beta_1, \ \beta_2, \ \cdots, \ \beta_q = 0$
- LR test comparing scaled log-likelihoods
  - Reduced model: log-likelihood $l(\beta; \phi y, H_0)$
  - Full model: log-likelihood $l(\beta; \phi, y, H_a)$
  - $G^2 = -2\frac{\log_e l(\text{reduced}) - \log_e l(\text{full})}{\phi} \overset{assympt.}{\sim} \chi^2_q$
- LR test comparing scaled deviances
  - Reduced: $D(reduced) = -2[l(\beta; \phi, y, H_0) - l(y; y)]$
  - Full model: $D(full) = -2[l(\beta; \phi, y, H_a) - l(y; y)]$
  - $G^2 = \frac{D(reduced) - D(full)}{\phi} \overset{assympt.}{\sim} \chi^2_q$
  - Better approximation of χ^2 than model-specific deviances

Prediction at New Data $\mathbf{x}$

- On the link scale: easy
  (CI for a linear comb. of $\widehat{\beta}$)

  – $\widehat{\beta} \overset{asympt.}{\sim} \mathcal{N}(\beta, V(\widehat{\beta})) \rightarrow \mathbf{x}'\widehat{\beta} \overset{asympt.}{\sim} \mathcal{N}(\mathbf{x}'\beta, \mathbf{x}'\,V(\widehat{\beta})\,x)$

  – CI for $\widehat{\eta}(x) = \mathbf{x}'\widehat{\beta}$: $\mathbf{x}'\widehat{\beta} \ \pm \ z_{1-\alpha/2}\sqrt{\mathbf{x}'V(\widehat{\beta})\mathbf{x}}$

- On the mean scale: approximate

  – CI for $\widehat{\mu}(\mathbf{x}) = g^{-1}(\widehat{\eta}(\mathbf{x}))$: $g^{-1}\left(\mathbf{x}'\widehat{\beta} \ \pm \ z_{1-\alpha/2}\sqrt{\mathbf{x}'V(\widehat{\beta})\mathbf{x}}\right)$

  – approximate CI since applying a non-linear transformation

- If $g(\cdot)$ is a decreasing function, the upper and lower bounds for the CI of $\widehat{\mu}(x)$ are switched.

Overdispersion

- Assume GLM $y \overset{ind}{\sim} EFD(\theta, \phi)$
  - Implies $Var\{y\} = b''(\theta)\ a(\phi) = V(\mu)\ \phi/w$

- Overdispersion:

  $Var\{y\} \neq$ variance in the model

  - Do not include the right predictors
  - Response variables are positively correlated or clustered (overdispersion)
  - Response variables are negatively correlated (underdispersion)

- Solution: view ϕ as an unknown dispersion parameter
  - Estimate ϕ from the data: $\widehat{\phi} = X^2/(n-p)$, where p is the number of parameters in the model.

Inference in Presence of Overdispersion

- Solutions of the likelihood equations do not depend on ϕ

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{Var(y_i)} \frac{\partial \mu_i}{\partial \eta_i} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{ij}}{V(\mu_i)\ \phi / w_i} \frac{\partial \mu_i}{\partial \eta_i} = 0$$

  - $E\{y_i\}$ and $\hat{\beta}$ are not affected by $\hat{\phi}$

- Can fit the model without overdispersion, and adjust afterwards
- The standard error of $\hat{\beta}$ scales by $\sqrt{\hat{\phi}}$
- When testing $H_0 : \beta_1 = \cdots = \beta_q = 0$:
  - Likelihood-based approaches are not valid
  - Use F test: $\frac{D_0 - D_1}{q\ \hat{\phi}} \overset{asympt.}{\sim} F_{q,n-p}$
  - Caution: D_0 and D_1 are deviances but not scaled deviances