Biostat 651: MidTerm Concept and Formula Sheet

1 Generalized Linear Models: Overview

Generalized Linear Models are concerned with finding some function of the mean, μ_i, that is linearly related to parameters, β by way of fixed covariates $\boldsymbol{x}_i$. This allows for modeling non-Normal random variables, like those in the exponential family with potentially non-constant variance.

$$g(\mu_i) = x_i^T \beta \quad i = 1, ..., n$$

2 MLE

MLE and related functions The Likelihood function $L(\theta)$ is proportional to the probability distribution function denoted, $L(\theta) \propto f_X(x|\theta)$. This means that effectively, it is the same function up to a constant.
The log likelihood is the log of the likelihood function, it and several of its derivatives are useful in understanding and estimating the parameters of the distribution function. Typical notation can be noted as follows:

$$\begin{aligned} X_1, ...X_n &\overset{i.i.d}{\sim} f_X(x|\theta) \\ L(\theta) &\propto f_{\boldsymbol{X}}(x|\theta) = \prod_{i=1}^{n} f_X(x|\theta) \\ l(\theta) &= \ln L(\theta) \\ U(\theta) &= \frac{\partial l(\theta)}{\partial \theta} \\ J(\theta) &= -\frac{\partial U(\theta)}{\partial \theta} = -\frac{\partial^2 l(\theta)}{\partial \theta^2} \\ I(\theta) &= -E[J(\theta)] = -E[\frac{\partial l(\theta)^2}{\partial \theta^2}] \end{aligned}$$

Where
$U(\theta)$ is the score function.
$J(\theta)$ is the information function
$I(\theta)$ is the expected information
The Maximum Likelihood Estimate, often denoted MLE, is the θ at which the likelihood function (as well as log-likelihood function) is maximized. It is found analytically by solving the score function equal to zero for θ or via numerical methods.

Exponential Family

Distributions from the exponential family can be constructed in the following parameterization.

$$f_X(x) = \exp\left\{\frac{\theta t(x) - b(\theta)}{a(\phi)} + c(x,\phi)\right\}$$

or alternatively

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) &= \frac{1}{Z(\boldsymbol{\theta})} h(\boldsymbol{x}) \exp\left\{\eta(\boldsymbol{\theta})^T \phi(\boldsymbol{x})\right\} \\ &= h(\boldsymbol{x}) \exp\left\{\eta(\boldsymbol{\theta})^T \phi(\boldsymbol{x}) - A(\boldsymbol{\theta})\right\} \end{aligned}$$

Definitions for Exponential Family

- $g(\mu_i)$ is called the link function
  must be monotone, differentiable function
- $\eta_i = x_i^T \beta$
- $t(x) = x \Rightarrow$ the parameterization is called 'canonical', θ is called the canonical or natural parameter
- $\eta_i = x_i^T \beta = \theta_i \Rightarrow$ the link is called 'canonical'
- $E[X] \equiv \mu = b'(\theta)$
- $V[X] = a(\phi) b''(\theta)$
- Variance of X typically expressed $V[X] = v(\mu) a(\phi)$
  - $v(\mu) = b''(\theta)$
- $b'(\theta_i) = g^{-1}(\mu_i)$ inverse functions

GLM Estimation

MLE for Exponential Families occurs by way of chain rule $U_i(\beta) = \frac{\partial l_i(\theta)}{\theta_i} \frac{\partial \theta_i}{\mu i} \frac{\partial \mu_i}{\eta_i} \frac{\partial \eta_i}{\beta_i}$
Resulting in:

$$U(bm\beta) = \frac{1}{a(\phi)} X^T V^{-1} \Delta^{-1} (Y - \mu)$$

Where $V = \text{diag}\{v(\mu_1), ..., v(\mu_n)\}$, and $\Delta = \{g'(\mu_1), ..., g'(\mu_n)\}$

Canonical Link Score Function

$$U(bm\beta) = \frac{1}{a(\phi)} X^T (Y - \mu)$$

The above functions give allow us to derive the MLE of a vector form exponential family.

Information Matrix for Exponential Families Also used for estimation (with IRWLS algorithm) as well as inference.
Canonical Link only requires the information function

$$J(\beta) = \frac{1}{a(\phi)} X^T V X$$

Non-canonical link does require the expected information

$$I(\beta) = X^T \{a(\phi) \Delta V \Delta\}^{-1} X$$

Estimation Algorithms

- Newton Raphson
  i. $\hat{\beta}_{(j+1)} = \hat{\beta}_j + J^{-1}(\beta_j)U(\hat{\beta}_j)$
  ii. Stopping criterion $||\hat{\beta_{(j+1)}} - \hat{\beta_{(j)}}|| < \epsilon$
  iii. Start $\hat{\beta_{(0)}} = \mathbf{0}$ usually
- Fisher Scoring- replace $J^{-1}(\beta_j)$ with I^{-1}
  i. Longer to converge
  ii. More robust to initial choice of $\hat{\beta_{(0)}}$
  iii. Equivalent to NR if canonical link used
- Iteratively Re-Weighted Least Squares
  i. Derived from Fisher Scoring
  ii. Uses initial guess $\hat{\beta}_0 = 0$ along with initial computations of weight matrix, and weighted response matrix to eventually scale $\hat{\beta}$
  iii. Like others, repeats until desired precision is gained.
  $\hat{\beta}^{(j+1)} = (X^T V_j X)^{-1} X^T V_j Z_j$
  where $Z_j = \eta_j + V_j^{-1}(Y - \mu_j)$

```
## Sample IRWLS
import numpy as np
nu = 3
ix = 0
max_iter = 1000
beta_ = np.array([-1,-1,-1]).reshape(3,1)
epsilon = 99
tolerance = 0.0001
inv_y = -nu/y ## link function
while epsilon > tolerance and ix <= max_iter:
   eta = X.dot(beta_)
   mu = -1/eta
   v = mu**2 ## b''(theta)
   V = np.diag(np.array(v).flatten())
   Z = eta + np.dot(np.linalg.inv(V), (np.array(y)-np.array(mu)))
   beta_1 = np.linalg.inv(np.dot(X.T.dot(V),X)).dot(np.dot(X.T.dot(V),Z))
   epsilon = np.sqrt( (beta_1 - beta_).T.dot(beta_1 - beta_))
   beta_ = beta_1
   ix +=1
```

3 GLM inference

3.1 Overview

- Asymptotically equivalent; not numerically equivalent
- LRT tends to perform better than Wald or Score in smaller samples
- Note: Wald is not invariant under reparameterization, Score and LR are invariant
  - e.g. $H_0 : \beta_1 = 0$ could yield different result than $H_0 : e^{\beta_1} = 1$
- assume $\beta \in \mathbb{R}^{q \times 1}$ for results below

3.2 Wald Test

Fit full model, $H_0 : \beta_1 = d$, test statistic takes form

$$\text{Test stat } = (C\hat{\beta} - d)^T \{\hat{V}(C\hat{\beta} - d)\}^{-1}(C\hat{\beta} - d) \sim \chi^2_r$$

where rank(C)=r and $\hat{V}[C\hat{\beta} - d] = CI^{-1}(\hat{\beta})C^T$

3.3 Score Test

Fit restricted model; maximize likelihood under null hypothesis constraints

$$\text{Test stat } = U(\hat{\beta}_H)^T I(\hat{\beta}_H)^{-1} U(\hat{\beta}_H) \sim \chi^2_{q1}$$

$$\hat{B}_H = \begin{bmatrix} d \\ \hat{\beta}_{2H} \end{bmatrix}$$

3.4 Likelihood Ratio Test

Fit full model, fit the reduced model then compute the LRT statistic

$$LRT = 2\{l(\hat{\beta}) - l(\hat{\beta}_H)\} \sim \chi^2_{q1}$$

4 GLM Diagnostics

4.1 Goodness of Fit

(Judge Current Model)

- Saturated Model: n parameters

  $\tilde{\theta}$

- Null model, $\hat{\mu_i} = \frac{1}{n}\sum_{i=1}^n g(\mu_i) \quad \forall i$

- **Deviance**: Generalized sum of squares - compare fitted and sautrated

  $D = 2\sum_{i=1}^n [Y_i(\tilde{\theta_i} - \hat{\theta_i}) - \{b(\tilde{\theta_i}) - b(\hat{\theta_i}))\}]$

  i. **Scaled Deviance**: $D^* = \frac{D}{a(\phi)}$ which $\sim \chi^2_{n-q}$ asymptotically with good model fit

  ii. Deviance decreases when covariates increase (applies to nested models)

  iii. Deviance can be used to carry out hypothesis tests in a manner equivalent to LRT

  $$D^* = 2 \times \{l(\tilde{\beta}) - l(\hat{\beta})\}$$

  iii. D_0^*, D_1^* denote scaled deviances under H_0, H_1 then allows

  $$\text{LRT Test Stat} = D_0^* - D_1^*$$

- *Pearson Chi-Square Statistic*

$$\text{test stat} = \sum_{i=1}^n \frac{(Y_i - \hat{\mu_i})^2}{\hat{V}(Y_i)} \sim \chi^2_{n-q} \text{ as } n \to \infty$$

- Comments:

  i. Scaled Deviance and Pearson statistics do not always work for logistic regression.

  ii. Pearson χ^2 has 'intuitive' appeal

4.2 Residuals

- Pearson Residuals:

$$\hat{r_i}^P = \frac{Y_i - \hat{\mu}_i}{\hat{V}_i^{1/2}}$$

- Deviance Residuals

$$\hat{r_i}^D = \text{sign}(Y_i - \hat{\mu})\sqrt{|D_i|}$$

$$D = \sum_{i=1}^{n} (\hat{r_i}^D)^2$$

- Standardized Residuals

$$\hat{r_i}^{PS} = \frac{\hat{r_i}^P}{\sqrt{1 - h_{ii}}}$$

$$\hat{r_i}^{DS} = \frac{\hat{r_i}^D}{\sqrt{1 - h_{ii}}}$$

- Short aside: Hat matrix takes following form for GLM (from IRWLS)

  i. leverage still h_{ii}

$$H = V^{1/2} X (X^T V X)^{-1} X^T V^{1/2}$$

4.3 Influence Measure

- One-step approximation used to avoid n refittings of cook's distance

$$D_i = \frac{1}{q} (\frac{h_{ii}}{1 - h_{ii}}) (r_i^{PS})^2$$

- Multicollinearity - detect high correlation in *weighted* predictors

$$VIF_j = \frac{1}{1 - R^2_{(j)}}$$

  i. where $R^2_{(j)} = R^2$ obtained from regressing the jth covariates against all other covariates

  ii. VIF = 1 : not correlated ;

  iii. $1 < VIF < 5$: moderately correlated ;

  iv. $VIF > 5$ to 10: high correlation

5 Helpful Hints/notes

$$E[U(\theta)] = 0$$
$$V[U(\theta)] = E[U(\theta)^T U(\theta)] = I(\theta)$$

6 Exponential Family- Wortwhile Derviations

All derivations are canonical

6.1 Bernoulli

6.1.1 Parameterization

$$Y \sim BERN(p) \Rightarrow f_Y(y) = p^y(1-p)^{1-y}$$
$$f_Y(y;p) = \exp\left\{y\ln(\frac{p}{1-p}) + \ln(1-p)\right\}$$

Table 1: Bernoulli

function	value
$a(\phi)$	1
$t(y)$	y
θ	$\ln(\frac{p}{1-p})$
p	$\frac{e^\theta}{1+e^\theta}$
$b(\theta)$	$\ln(1-p)$
$b'(\theta)$	$\frac{-1}{1-p}$
$b''(\theta)$	$\frac{-1}{(1-p)^2}$
$c(y,\phi)$	0
$g(\mu)$	$\frac{e^\theta}{1+e^\theta}$

6.1.2 Derivations

$$Y_1, ...Y_n \overset{i.i.d}{\sim} \Rightarrow$$
$$f_{\boldsymbol{Y}}(y_i;p) = \exp\left\{\ln\left(\frac{p}{1-p}\right)\sum_{i=1}^n y_i + n\ln(1-p))\right\} = \prod_{i=1}^n f_Y(y;p)$$
$$l(p) = \ln\left(\frac{p}{1-p}\right)\sum_{i=1}^n y_i + n\ln(1-p)$$
$$U(p) = [p^{-1} - (1-p)^{-1}]\sum_{i=1}^n y_i - \frac{n}{1-p}$$
$$= \frac{np - \sum_{i=1}^n y_i}{p(1-p)}$$
$$J(p) = -p^{-2}\sum_{i=1}^n y_i - \frac{(2p-2)\sum_{i=1}^n y_i}{(1-p)^4} - \frac{n}{(1-p)^2}$$

6.2 Negative Binomial

6.2.1 Parameterization

$$Y \sim NBIN(p,r) \text{ with r known} \Rightarrow$$
$$f_Y(y) = \binom{y+r-1}{y}p^y(1-p)^r$$

$$f_Y(y|p,r) = \exp\left\{y\ln(p) + r\ln(1-p) + \ln\binom{y+r-1}{y}\right\}$$

Table 2: Neg-Binom

function	value
$a(\phi)$	1
$t(y)$	y
θ	$\ln(p)$
p	e^{θ}
$b(\theta)$	$r\ln(1-p)$
$b'(\theta)$	$\frac{rp}{1-p}$
$b''(\theta)$	$\frac{rp}{(1-p)^2}$
$c(y,\phi)$	$\ln\binom{y+r-1}{y}$
$g(\mu)$	$-\ln\frac{r+\mu}{\mu}$

$$Y_1,...Y_n \overset{i.i.d}{\sim} \Rightarrow$$

$$f_{\boldsymbol{Y}}(y_i,p,r) = \exp\Big\{\ln(p)\sum_{i=1}^{n} y_i + nr\ln(1-p) + \sum_{i=1}^{n}\ln\binom{y_i+r-1}{y}\Big\} = \prod_{i=1}^{n} f_Y(y|p,r)$$

$$l(p) = \ln(p)\sum_{i=1}^{n} y_i + n\ln(1-p) + \sum_{i=1}^{n}\ln\binom{y_i+r-1}{y_i}$$

$$U(p) = \frac{\partial l(p)}{\partial p} = \frac{\sum_{i=1}^{n} y_i}{p} - \frac{n}{1-p}$$

$$I(p) = -E[\frac{\partial U(p)}{\partial p}] = -E[-p^{-2}\sum_{i=1}^{n} y_i - \frac{n}{(1-p)^2}]$$

$$= \frac{n}{p} + \frac{n}{(1-p)^2}$$

$$\mu = \frac{\partial b(\theta)}{\partial\theta} = \frac{\partial}{\partial\theta}(r\ln(1-e^{\theta}))$$

$$= \frac{re^{\theta}}{1-e^{\theta}} = \frac{rp}{(1-p)}$$

$$\mu = \frac{rp}{(1-p)} = \frac{re^{\theta}}{1-e^{\theta}}$$

$$\eta = x_i^T\beta = \theta_i \Rightarrow \mu = \frac{re^{\eta}}{1-e^{\eta}}$$

$$\Rightarrow \frac{1-e^{\eta}}{e^{\eta}} = \frac{r}{\mu}$$

$$e^{-\eta} = \frac{r}{\mu} + 1$$

$$\eta = -\ln(\frac{r+\mu}{\mu})$$

$$g(\mu) = -\ln(\frac{r+\mu}{\mu})$$

$$
\begin{aligned}
v(\mu) = b''(\theta) &= \frac{\partial}{\partial\theta}\left(\frac{re^{\theta}}{1-e^{\theta}}\right) \\
&= \frac{re^{\theta}(1-e^{\theta}) + re^{2\theta}}{(1-e^{\theta})^2} \\
&= \frac{re^{\theta}}{(1-e^{\theta})^2} = \frac{rp}{(1-p)^2}
\end{aligned}
$$