ST3241 Categorical Data Analysis I
Generalized Linear Models

Some More Discussions

Deviance

- A saturated model has a separate parameter for each observation giving a perfect fit.
- Let $\tilde{\theta}$ denote the estimate of θ for the saturated model, corresponding to estimated means $\tilde{\mu_i} = y_i$ for all i.
- Let $\hat{\theta}$ denote the MLE of θ for the model under consideration.
- The *deviance* of the fitted model is defined as

$$\begin{aligned} D(y;\hat{\mu}) &= -2[L(\hat{\mu};y) - L(y;y)] \\ &= 2\sum_{i=1}^{N}[y_i\tilde{\theta}_i - b(\tilde{\theta}_i)]/a(\phi) - 2\sum_{i=1}^{N}[y_i\hat{\theta}_i - b(\hat{\theta}_i)]/a(\phi) \end{aligned}$$

Deviance

- Usually, $a(\phi)$ has the form $a(\phi) = \phi/w_i$, and this statistic equals

$$2\sum_{i=1}^{N} w_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]/\phi$$

- This is called *scaled deviance*.
- The greater the scaled deviance, the poorer the fit.
- For some GLMs, the scaled deviance has an approximate chi-squared distribution.

Deviance For Poisson Model

- For Poisson GLMs,
  $\hat{\theta}_i = \log \hat{\mu}_i$ and $b(\hat{\theta}_i) = \exp(\hat{\theta}_i) = \hat{\mu}_i$
- Similarly, for saturated model
  $\tilde{\theta}_i = \log y_i$ and $b(\tilde{\theta}_i) = y_i$
- Also, $a(\phi) = 1$ and the deviance is equal to

$$D(y; \hat{\mu}) = 2 \sum_{i=1}^{N} [y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i]$$

- When a model with log-link contains an intercept term, the deviance simplifies to

$$D(y; \hat{\mu}) = 2 \sum_{i=1}^{N} y_i \log(y_i/\hat{\mu}_i)$$

Deviance For Binomial Model

- Consider binomial GLMs with sample proportions $\{y_i\}$ based on $\{n_i\}$ trials. Then $\hat{\theta}_i = \log[\hat{\pi}_i/(1-\hat{\pi}_i)]$ and $b(\hat{\theta}_i) = \log[1+\exp(\hat{\theta}_i)] = -\log(1-\hat{\pi}_i)$
- Similarly, for the saturated model, $\tilde{\theta}_i = \log[y_i/(1-y_i)]$ and $b(\tilde{\theta}_i) = -\log(1-y_i)$
- Also, $a(\phi) = 1/n_i$, so $\phi = 1$ and $w_i = n_i$. The deviance equals $$2\sum_{i=1}^{N} n_i\{y_i(\log \tfrac{y_i}{1-y_i} - \log \tfrac{\hat{\pi}_i}{1-\hat{\pi}_i}) + \log(1-y_i) - \log(1-\hat{\pi}_i)\}$$ $$= 2\sum_{i=1}^{N} n_i y_i \log \tfrac{n_i y_i}{n_i \hat{\pi}_i} + 2\sum_{i=1}^{N} (n_i - n_i y_i) \log \tfrac{n_i - n_i y_i}{n_i - n_i \hat{\pi}_i}$$
- At setting i, $n_i y_i$ is the number of successes and $(n_i - n_i y_i)$ is the number of failures, $i = 1, \cdots, N$. Thus the deviance is $D(y;\hat{\mu}) = 2\sum \mathit{observed} \times \log(\mathit{observed}/\mathit{fitted})$

Deviance Residuals

- Define

$$d_i = 2w_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]$$

- The deviance residual for observation i is

$$\sqrt{d_i} \times sign(y_i - \hat{\mu}_i)$$

Some SAS Codes

```
data glm;
 input snoring disease total;
datalines;
0 24 1379
2 35 638
4 21 213
5 30 254
;
proc genmod; model disease/total = snoring / dist=bin
   link=identity;
proc genmod; model disease/total = snoring / dist=bin
   link=logit;
proc genmod; model disease/total = snoring / dist=bin
   link=probit;
run;
```

Some R Codes

```
snoring<-c(0,2,4,5)
disease<-c(24,35,21,30)
total<-c(1379,638,213,254)
glm(cbind(disease,total-disease) snoring,
  family=binomial(link="logit"))
glm(cbind(disease,total-disease) snoring,
  family=binomial(link=probit"))
```

Reference: McCullagh, P. and Nelder, J.A.
(1989).Generalized Linear Models. 2nd ed.
London: Chapman and Hall.