STAT347: Generalized Linear Models
Lecture 9

Today's topics: Chapters 7.1 and 7.2

- Poisson loglinear model
- Poisson modeling for contingency tables

1 Poisson loglinear model

Poisson distribution density function is

$$f(y) = e^{-\mu}\mu^y/y! = e^{y\log\mu-\mu}/y!$$

Loglinear model: use the canonical link

$$\log\mu_i = X_i^T\beta$$

Or equivalently, $\mu_i = (e^{\beta_1})^{x_{i1}}\cdots(e^{\beta_p})^{x_{ip}}$, assuming that each x_{ij} has a multiplicative effect on y_i.

- Estimated variance of $\hat\beta$: $\widehat{\mathrm{var}}(\hat\beta) = (X^T\hat W X)^{-1}$. Each diagonal element $w_{ii} = v_{ii} = \mathrm{var}(y_i) = \mu_i$
- Residual deviance:
$$D_+(y,\hat\mu) = 2\sum_{i=1}^n\left[y_i\log\left(\frac{y_i}{\hat\mu_i}\right) - y_i + \hat\mu_i\right]$$
- Offset: forcing the coefficient of a variable to be 1.

  Example: modeling rates, y_i crime counts and t_i the total population within each county, and we assume
$$\log(\mu_i/t_i) = X_i^T\beta$$
  or equivalently $\log(\mu_i) = \log(t_i) + X_i^T\beta$. the adjustment term $\log(t_i)$ is called an offset as we do not need to estimate its coefficient.

2 Poisson modeling for contingency tables

For independent Poisson counts $(y_1,\cdots,y_c)$, the total $n = \sum_i y_i$ follows a Poisson distribution with mean $\sum_i\mu_i$. Conditional on the total n, the conditional joint distribution is

$$\frac{P(y_1 = n_1,\cdots,y_c = n_c)}{P(\sum_i y_i = n)} = \left(\frac{n!}{\prod_i n_i!}\right)\prod_{i=1}^c p_i^{n_i}$$

and it follows a multinomial distribution.

- This indicates that we can view the data equivalently as there are n i.i.d. samples and each sample follows a multinomial distribution to choose one of the cells.

2.1 one-way layout

Analogous to ANOVA, consider a one-way layout for the count response. Assume that each cell $i \in \{1, 2, \cdots, c\}$ has n_i repeated observations. Then the Poisson model is

$$\log(\mu_{ij}) = \beta_0 + \beta_i, \quad j = 1, 2, \cdots, n_i$$

2.2 Two-way contingency table

Consider an $r \times c$ table for two categorical variables (denote as A and B). The Poisson GLM assumes that the count y_{ij} in each cell independently follows a Poisson distributions with mean μ_{ij}. Consider two scenarios:

2.2.1 Two categorical variables are independent

If we assume that the two categorical variables are independent, then we can assume

$$\mu_{ij} = \mu\phi_i\psi_j$$

Equivalently, we can assume that

$$\log \mu_{ij} = \beta_0 + \beta_i^A + \beta_j^B$$

This model has a $[1 + (r-1) + (c-1)]$ free parameters (degree of freedom).
The non-constant part of the log-likelihood is

$$L(\mu) = \sum_{i=1}^{r}\sum_{j=1}^{c} y_{ij}\log\mu_{ij} - \sum_{i=1}^{r}\sum_{j=1}^{c}\mu_{ij}$$

The we used the canonical link, the score equations should be

$$\sum_{i,j}(y_{ij} - \mu_{ij}) = 0$$

$$\sum_{j}(y_{ij} - \mu_{ij}) = 0, \quad i = 1, 2, \cdots, r$$

$$\sum_{i}(y_{ij} - \mu_{ij}) = 0, \quad j = 1, 2, \cdots, c$$

Thus we get the MLE: $\hat{\mu} = y_{++}$, $\hat{\phi}_i = y_{i+}/y_{++}$ and $\hat{\psi}_j = y_{+j}/y_{++}$.
We can also write down the likelihood conditional on n, and we get the same MLE (Chapter 7.2.2).

2.2.2 Two categorical variables has an interaction

We can assume

$$\log \mu_{ij} = \beta_0 + \beta_i^A + \beta_j^B + \gamma_{ij}^{AB}$$

- We need identifiability conditions such as $\gamma_{1j}^{AB} = \gamma_{i1}^{AB} = 0$ for identifiability.
- In total adds $(r-1) \times c - 1$ more parameters
- This model is saturated
- The interactions pertain to odds ratios. For instance, $r = c = 2$

$$\log\frac{p_{11}/p_{12}}{p_{21}/p_{22}} = \log\frac{\mu_{11}/\mu_{12}}{\mu_{21}/\mu_{22}} = \gamma_{11}^{AB} + \gamma_{22}^{AB} - \gamma_{12}^{AB} - \gamma_{21}^{AB}$$

Under our previous identification condition, the odds ratio is $e^{\gamma_{22}^{AB}}$

2.3 Three-way contingency table

Consider an $r \times c \times l$ table. Assume that for an individual sample

- Mutual independence

$$P(A = i, B = j, C = k) = P(A = i)P(B = j)P(C = k)$$

  Equivalently, the loglinear form is

$$\log \mu_{ijk} = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C$$

- Joint independence

$$P(A = i, B = j, C = k) = P(A = i)P(B = j, C = k)$$

  Equivalently, the loglinear form is

$$\log \mu_{ijk} = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{jk}^{BC}$$

- Conditional independence

$$P(A = i, B = j \mid C = k) = P(A = i \mid C = k)P(B = j \mid C = k)$$

  Equivalently, the loglinear form is

$$\log \mu_{ijk} = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{ik}^{AC} + \gamma_{jk}^{BC}$$

- Homogeneous association

$$\log \mu_{ijk} = \beta_0 + \beta_i^A + \beta_j^B + \beta_k^C + \gamma_{ik}^{AC} + \gamma_{jk}^{BC} + \gamma_{ij}^{AB}$$

  An interpretation of this model is that any two pairs are dependent, but the dependence does not change with the value of the third variable.

3 Connection with binomial/multinomial regression models

- The log-linear model treat all categorical variables symmetrically and regard the cells as response
- The logistic models distinguish between response and categorical variables

Consider the case where $r = 2$ and treat it as the response variable for a logistic regression. Then start from the loglinear model, we have

$$\begin{aligned}\log \frac{P(A = 1 \mid B = j, C = k)}{P(A = 2 \mid B = j, C = k)} &= \log \mu_{1jk} - \log \mu_{2jk} \\ &= (\beta_1^A - \beta_2^A) + (\gamma_{1j}^{AB} - \gamma_{2j}^{AB}) + (\gamma_{1j}^{AC} - \gamma_{2j}^{AC})\end{aligned}$$

Equivalently, we have the model

$$\text{logit}[P(A = 1 \mid B = j, C = k)] = \lambda + \delta_j^B + \delta_k^C$$

which is a logistic regression model

- The Poisson loglinear model and binomial logistic model also have the same score equations
- The same results hold for the multinomial baseline-category logit model

Next time: Chapters 7.3-7.5