STAT347: Generalized Linear Models

Lecture 9

Winter, 2023
Jingshu Wang

Today’s topics:

- Poisson loglinear model
- Poisson modeling for contingency tables
- R example

Counts as the response in regression

In many applications, the response variables are counts

- Our example in Lecture 1: the number of male satellite for female horseshoe crabs
- Number of views for a YouTube video
- Number of mRNA copies measured for each gene in RNA sequencing experiments

Features of the counts:

- The response Y typically has a wide range of values
- Larger counts typically have higher randomness

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.

Poisson loglinear model

- Estimated variance of $\hat{\beta}$: $\widehat{\text{var}}(\hat{\beta}) = (X^T \hat{W} X)^{-1}$. Each diagonal element $w_{ii} = v_{ii} = \text{var}(y_i) = \mu_i$
- 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.

Modeling a 2×2 table by three different ways

Quality	No Particles	Particles	Total
Good	320	14	334
Bad	80	36	116
Total	400	50	450

Table 1: 2×2 table. A sample of wafers was drawn and cross-classified according to whether a particle was found on the die that produced the wafer and whether the wafer was good or bad.

Three different assumptions on data collection

- The data is obtained by randomly sample 400 wafers without particles and 50 with particles
  - This leads to a Binomial GLM where the grouped data has 2 samples (one for no particles, the other for particles)

Modeling a 2×2 table by three different ways

Quality	No Particles	Particles	Total
Good	320	14	334
Bad	80	36	116
Total	400	50	450

Three different assumptions on data collection

- The data is obtained by randomly sample 400 wafers without particles and 50 with particles
- We randomly sample 450 wafers and cross-classify them.
  - This leads to a multinomial model where the grouped level data only has one sample $y = (320, 80, 14, 36) \sim \text{Multinomial}(450, p)$
- The data is obtained from observations during a fixed period of time and we happen to observe 450 total observations.
  - This leads to a Poisson model where the data has 4 samples: $X_i = 00, 01, 10, 11$ and $Y_i = 320, 80, 14, 36$.

Modeling a 2×2 table by three different ways

Equivalence between the Poisson distribution and Multinomial distribution:

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.

Modeling for contingency tables

Table 7.1 Number of Deaths from Lung Cancer, by Histology, Stage of Disease, and Follow-up Time Interval[a]

Follow-up Time Interval (months)	Histology:	I			II			III		
	Disease Stage:	1	2	3	1	2	3	1	2	3
0–2		9	12	42	5	4	28	1	1	19
		(157	134	212	77	71	130	21	22	101)
2–4		2	7	26	2	3	19	1	1	11
		(139	110	136	68	63	72	17	18	63)
4–6		9	5	12	3	5	10	1	3	7
		(126	96	90	63	58	42	14	14	43)
6–8		10	10	10	2	4	5	1	1	6
		(102	86	64	55	42	21	12	10	32)
8–10		1	4	5	2	2	0	0	0	3
		(88	66	47	50	35	14	10	8	21)
10–12		3	3	4	2	1	3	1	0	3
		(82	59	39	45	32	13	8	8	14)
12+		1	4	1	2	4	2	0	2	3
		(76	51	29	42	28	7	6	6	10)

[a]Values in parentheses represent total follow-up months at risk.

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:

- Assume that two categorical variables are independent

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

- Allow two categorical variables to have an interaction

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

Two categorical variable are independent

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

with $\sum_i \phi_i = \sum_j \psi_i = 1$

Equivalently, we can assume that

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

(We may assume a different identification condition $\sum_i \beta_i^A = \sum_j \beta_j^B = 0$).

This model has $1 + (r - 1) + (c - 1)$ free parameters

Two categorical variable are independent

The non-constant part of the log-likelihood is

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

As we use 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 $\hat{\psi}_j = y_{+j}/y_{++}$.

Two categorical variable have 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 can be interpreted as 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 $e^{\gamma_{22}^{AB}}$.

Three-way contingency table

- Consider an $r \times c \times l$ table for three categorical variables (denote as A, B and C).
- The Poisson GLM assumes that the count y_{ijk} in each cell independently follows a Poisson distributions with mean μ_{ijk}.
- There are multiple scenarios for the dependence assumptions across the three variables

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

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

Homogenous association

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

Equivalently, the loglinear form is

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

- Any two pairs are dependent, but the dependence does not change with the value of the third variable.

- Given any fixed level k of C, the conditional association (conditional odds ratios) does not depend on k

- The saturated model allows any dependence structure

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

Connection with binomial/multinomial regressions

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

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

Equivalently, we have the model

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

which is a logistic regression model

Connection with binomial/multinomial regressions

- The log-linear model treat all categorical variables symmetrically as X and regard the counts in each cell as response y.
- The logistic models treat one of the categorical variables as response y and the remaining categorical variables as X.

- A three-term interaction in the Poisson model corresponds to the interaction term in the logistic regression.
- 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

R data example for contingency tables

- Check Example5 R notebook