

# Lecture 15: Logistic and Poisson Regression

Nancy R. Zhang

Statistics 203, Stanford University

March 1, 2010

## Review - Binary responses model

Model:  $Y \in \{0, 1\}$ ,

$$P(Y = 1 | X_1, \dots, X_p) = g^{-1}(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p).$$

Where

$$g(\pi) = \log\left(\frac{\pi}{1 - \pi}\right).$$

The inverse  $g^{-1}$  is

$$g^{-1}(z) = \frac{e^z}{1 + e^z}.$$

We have no choice but to accept non-constant variance,

$$\text{Var}(Y) = \pi(X)[1 - \pi(X)].$$

## Review - Model interpretation

An intuitive quantity to assess probabilities:

$$odds = \frac{P(Y = 1|X)}{P(Y = 0|X)}.$$

In the logistic regression model,

$$\log(odds) = \beta X.$$

The parameter  $\beta$  is the contribution of unit increase in  $X$  to the increase (decrease) in odds. For example, if  $X$  were binary as well,

$$\log\left(\frac{odds(X = 1)}{odds(X = 0)}\right) = \beta.$$

# Logit Model for Multinomial Response

If the response  $Y$  belong to  $K$  categories.

① Designate one category as the “base” category.

②

$$P(Y = k|X) = \frac{e^{X\beta_k}}{1 + \sum_{l=1}^{K-1} e^{X\beta_l}}$$

Here,  $\beta_k = (\beta_{k1}, \dots, \beta_{kp})$ .

$$P(Y = K|X) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{X\beta_l}}$$

③  $p \times (K - 1)$  parameters.

④  $\beta_{ki}$  for  $k$ -th category and  $i$ -th predictor interpreted as increase in log-odds from base category.

# Logit Model for Multinomial Response

Equivalent definition:

$$\log \frac{\pi_k(X)}{\pi_K(X)} = \alpha_k + X\beta_k, \quad k = 1, \dots, K - 1,$$

where

$$\pi_k(X) = P(Y = k|X).$$

# Alligator Food Example

Study on the primary food choice of alligators.

- 1 Data: 219 alligators captured in four Florida lakes.
- 2 Response variable: food type, in volume, found in the alligator's stomach. 5 categories:
  - 1 fish
  - 2 invertebrate
  - 3 reptile
  - 4 bird
  - 5 other
- 3 Predictors:
  - 1 Lake of capture (Hancock, Oklawaha, Trafford, George)
  - 2 Gender (M, F).
  - 3 Size ( $\leq 2.3m$ ,  $\geq 2.3m$ ).

# Alligator Food Choice Example

TABLE 7.1 Primary Food Choice of Alligators

Lake	Gender	Size (m)	Primary Food Choice				
			Fish	Invertebrate	Reptile	Bird	Other
Hancock	Male	≤ 2.3	7	1	0	0	5
		> 2.3	4	0	0	1	2
	Female	≤ 2.3	16	3	2	2	3
		> 2.3	3	0	1	2	3
Oklawaha	Male	≤ 2.3	2	2	0	0	1
		> 2.3	13	7	6	0	0
	Female	≤ 2.3	3	9	1	0	2
		> 2.3	0	1	0	1	0
Trafford	Male	≤ 2.3	3	7	1	0	1
		> 2.3	8	6	6	3	5
	Female	≤ 2.3	2	4	1	1	4
		> 2.3	0	1	0	0	0
George	Male	≤ 2.3	13	10	0	2	2
		> 2.3	9	0	0	1	2
	Female	≤ 2.3	3	9	1	0	1
		> 2.3	8	1	0	0	1

Source: Data courtesy of Clint Moore, from an unpublished manuscript by M. F. Delaney and C. T. Moore.

- 1 Do gender, size, or lake of capture influence food choice?
- 2 Are there interaction effects?
- 3 Obtain estimates of  $P(\text{food choice} = \text{fish} \mid \text{Gender, Size, Lake})$ .

Functions for multinomial fitting in R: `multinom` in library `nnet`.

## Fitted probabilities $\hat{\pi}$

If you had data about the size of the alligators (and not just the classification ( $\leq$  or  $\geq 2.3$  m)), then you can estimate a response curve like this:

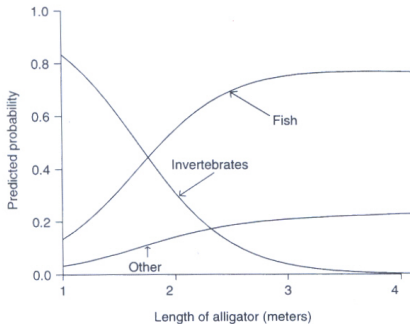


FIGURE 7.1 Estimated probabilities for primary food choice.

From Agresti, Categorical Data Analysis

# Count data

- 1 Men and women were asked whether they believed in the after life (1991 General Social Survey).

- 2 Results:

	Y	N or U	
M	435	147	582
F	375	134	509
Total	810	281	1091

- 3 Question: is belief in afterlife independent of gender?

# Contingency Tables

	Y	N or U	
M	435	147	582
F	375	134	509
Total	810	281	1091

- 1 Model:  $Y_{ij} \sim \text{Poisson}(\lambda_{ij})$ .
- 2  $H_0$ : Independence. i.e.  $\lambda_{ij} = \lambda \alpha_i \beta_j$ .
- 3  $H_A$ :  $\lambda_{ij}$  arbitrary.
- 4 Under independence:

$$\log E(Y_{ij}) = \log \lambda_{ij} = \log \lambda + \log \alpha_i + \log \beta_j.$$

# Poisson Regression

- 1 Model fitting: Newton Raphson.
- 2 Confidence intervals: same as for Binomial, use local Gaussianity.
- 3 Assessment of model fit: Deviance residuals.

## Loglinear versus Logit Models

	Y	N or U	
M	435	147	582
F	375	134	509
Total	810	281	1091

Model:  $Y_{ij} \sim \text{Poisson}(\lambda_{ij})$ .

If you have two Poissons,  $\text{Poiss}(\lambda_{i1})$  and  $\text{Poiss}(\lambda_{i2})$ , then conditioned on their sum, each count is a binomial.

$$Y_{i,1} | Y_{i,1} + Y_{i,2} \sim \text{Binomial} \left( Y_{i,1} + Y_{i,2}, \frac{\lambda_{i1}}{\lambda_{i1} + \lambda_{i2}} \right)$$

Then,

$$\begin{aligned} \text{logit}P(1 | \text{row} = i, \text{row sum} = n_i) &= \log \frac{P(1 | \text{row} = i, \text{row sum} = n_i)}{P(2 | \text{row} = i, \text{row sum} = n_i)} \\ &= \log \frac{\lambda_{i1}}{\lambda_{i2}} \\ &= \log \lambda_{i1} - \log \lambda_{i2}. \end{aligned}$$

## 2 × 2 tables

$$\text{logit}P(1 | \text{row} = i, \text{row sum} = n_i) = \log \lambda_{i1} - \log \lambda_{i2}.$$

Under the null hypothesis:

$$H_0 : \lambda_{ij} = \lambda * \alpha_i * \beta_j,$$

$$\log \lambda_{ij} = \log \lambda + \log \alpha_i + \log \beta_j.$$

$$\text{logit}P(1 | \text{row} = i, \text{row sum} = n_i) = \log \beta_1 - \log \beta_2 \equiv \delta$$

The key is that the above logit does not depend on  $i$ . In binomial regression, we are modeling

$$\text{logit}P(1 | X) = \beta_0 + \beta_1 X.$$

So testing  $H_0$  is equivalent to testing  $\beta_1 = 0$  in logistic regression.

## 2 × 2 tables

Thus...

- 1 Testing the hypothesis  $H_0 : \lambda_{ij} = \lambda * \alpha_i * \beta_j$  in the Poisson model is the same as testing independence in the logistic model.
- 2 To test this hypothesis, you fit the model with  $\lambda_{ij}$  arbitrary, and then use Chi-square test on the difference of deviances.
- 3 The difference of deviances will be the same as the logit model, but the absolute deviances will be different.

## 3-way tables: Alcohol Cigarette, and Marijuana Use

Survey asked 2276 students in their final year of high school in a nonurban area near Dayton, Ohio whether they ever used alcohol, cigarettes, or marijuana.

Alcohol Use	Cigarette Use	Marijuana Use	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

This is example of a  $2 \times 2 \times 2$  contingency table. Shorthand: A=alcohol, C=cigarette, M=marijuana.

## 3-way tables: Types of Interaction

Alcohol Use	Cigarette Use	Marijuana Use	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

$$Y_{ijk} \sim \text{Poisson}(\lambda_{ijk})$$

Conditioned on total ( $N$ )  $Y_{ijk} \sim \text{Multinom}(N, \pi_{ijk})$ .

$\pi_{i++}$  be probability of row  $A = i$ ,

$\pi_{ij+}$  be probability of  $A = i, C = j$ , etc.

### 1 A,C, and M mutually independent

$$\log \lambda_{ijk} = \lambda + \lambda_i^A + \lambda_j^C + \lambda_k^M$$

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$$

## 3-way tables: Types of Interaction

Alcohol Use	Cigarette Use	Marijuana Use	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

$$Y_{ijk} \sim \text{Poisson}(\lambda_{ijk})$$

Conditioned on total ( $N$ )  $Y_{ijk} \sim \text{Multinom}(N, \pi_{ijk})$ .

$\pi_{i++}$  be probability of row  $A = i$ ,

$\pi_{ij+}$  be probability of  $A = i, C = j$ , etc.

①  $M$  is **jointly independent** of  $A, C$

$$\log \lambda_{ijk} = \lambda + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ij}^{AC}$$

$$\pi_{ijk} = \pi_{ij+} \pi_{++k}$$

## 3-way tables: Types of Interaction

Alcohol Use	Cigarette Use	Marijuana Use	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

$$Y_{ijk} \sim \text{Poisson}(\lambda_{ijk})$$

Conditioned on total ( $N$ )  $Y_{ijk} \sim \text{Multinom}(N, \pi_{ijk})$ .

$\pi_{i++}$  be probability of row  $A = i$ ,

$\pi_{ij+}$  be probability of  $A = i, C = j$ , etc.

### 1 C and M **conditionally independent** given A

$$\log \lambda_{ijk} = \lambda + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ij}^{AC} + \lambda_{ik}^{AM}$$

$$\pi_{jk|i} = \pi_{j+|i} \pi_{+k|i}$$

## 3-way tables: Types of Interaction

Alcohol Use	Cigarette Use	Marijuana Use	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

$$Y_{ijk} \sim \text{Poisson}(\lambda_{ijk})$$

Conditioned on total ( $N$ )  $Y_{ijk} \sim \text{Multinom}(N, \pi_{ijk})$ .

$\pi_{i++}$  be probability of row  $A = i$ ,

$\pi_{ij+}$  be probability of  $A = i, C = j$ , etc.

- 1 Each pair of A,C, and M has **homogeneous association**.

$$\log \lambda_{ijk} = \lambda + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ij}^{AC} + \lambda_{ik}^{AM} + \lambda_{ik}^{CM}.$$

e.g. the dependence relationship of A, C does not depend on M.

## 3-way tables: Types of Interaction

Alcohol Use	Cigarette Use	Marijuana Use	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

$$Y_{ijk} \sim \text{Poisson}(\lambda_{ijk})$$

Conditioned on total ( $N$ )  $Y_{ijk} \sim \text{Multinom}(N, \pi_{ijk})$ .

$\pi_{i++}$  be probability of row  $A = i$ ,

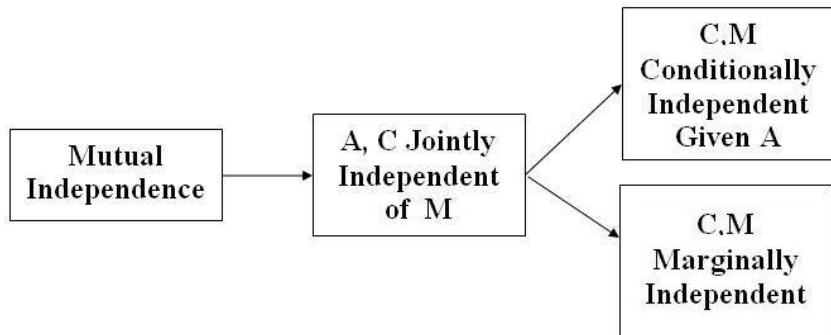
$\pi_{ij+}$  be probability of  $A = i, C = j$ , etc.

### 1 Saturated Model.

$$\log \lambda_{ijk} = \lambda + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ij}^{AC} + \lambda_{ik}^{AM} + \lambda_{jk}^{CM} + \lambda_{ijk}^{ACM}.$$

## 3-way tables: Types of Interaction

Symbol	Interpretation
(A,C,M)	Mutual Independence
(AC,M)	AC jointly independent of M
(AC,AM)	M, C conditionally independent given A
(AC,AM,CM)	Homogeneous association of each pair.



Marginal independence: fit  $2 \times 2$  table.

# Analysis of 3-way tables

- 1 Fit log-linear model (Poisson GLM) for each of the models.
  - 1 Criterion: maximum likelihood.
  - 2 Fitting method: Newton Raphson.
- 2 Use a model selection criterion to choose the best one.
  - 1 AIC, BIC.
  - 2 Use Deviance  $\chi^2$  test to choose between nested models.

`glm(..., family=poisson), loglm(MASS)` in R.