Analysis of Categorical Data
Chapter 1 and 2: Introduction and Contingency Table

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through these chapters, you should be able to

1. describe different sampling themes,
2. compute odds ratios and understand their implications,
3. describe confounding and Simpson's paradox,
4. construct partial table and marginal table,
5. evaluate and test associations.

Categorical Variable

A categorical variable has a measurement scale consisting of a set of categories.

- Binary: Yes/No
- Nominal: Volvo/Volkswagen/Toyota/BMW
- Ordinal: Disagree/Neutral/Agree
- Counts: 0, 1, 2, ...

A continuous variable has a measurement scale consisting of all real numbers in an interval.

Distributions of Categorical Data

- Bernoulli distribution $Y \sim \text{Bernoulli}\left(\pi\right)$:

$$P\left(Y=y\right)=\pi^{y}\left(1-\pi\right)^{n-y}, \quad y=0,1,$$

  where π is the success probability.

- Binomial distribution $Y \sim \text{Binomial}\left(n,\pi\right)$:

$$P\left(Y=y\right)=\left(\begin{array}{c} n \\ y \end{array}\right)\pi^{y}\left(1-\pi\right)^{n-y}, \quad y=0,1,...,n.$$

  where n is the total number of trials and π is the success probability.

- Multinomial distribution $Y \sim \text{Multinonial}\left(\boldsymbol{n},\boldsymbol{\pi}\right)$:

$$P\left(n_1,n_2,...,n_c\right)=\frac{n!}{n_1!n_2!\cdots n_c!}\pi_1^{n_1}\pi_2^{n_2}\cdots\pi_c^{n_c},$$

  where $\pi_i=P\left(\text{outcome } i\right)$, $\sum_{i=1}^{c}\pi_i=1$, and $\sum_{i=1}^{c}n_i=n$.

Distributions of Categorical Data

- Poisson distribution $Y \sim \text{Poi}\left(\mu\right)$:

$$P\left(Y=y\right)=\frac{\mu^{y}}{y!}\exp\left\{-\mu\right\}, \quad y=0,1,2,....$$

  where μ is the mean.

- Negative binomial distribution $Y \sim \text{NegBin}\left(\mu,\phi\right)$:

$$P\left(Y=y\right)=\frac{\Gamma\left(y+\phi\right)}{\Gamma\left(\phi\right)\Gamma\left(y+1\right)}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y}, \quad y=0,1,2,....$$

  where $\Gamma\left(\right)$ is the gamma function, μ is the mean, and ϕ is the dispersion parameter.

  - $\mathbb{E}\left(Y\right)=\mu$ and $\text{var}\left(Y\right)=\mu+\mu^{2}/\phi$.
  - If $\phi\to\infty$, the negative binomial distribution reduces to the Poisson distribution.

Apply Appropriate Methods

A variable's measurement scale determines which statistical methods are appropriate.

- Apply methods appropriate for the actual scale.
- Methods for variables of one type usually can be used with variables at higher levels, but usually not at lower levels.
    - e.g., if we ignore ordering, ordinal data become nominal data. But ordinal data methods cannot be used with nominal data.

In this course, we focus on the case where the response variable is categorical. The covariates/features can be continuous or categorical.

Contingency Table

Let X be a categorical variable with I categories, and Y be a categorical variable with J categories. An $I \times J$ contingency table having I rows for categories of X and J columns for categories of Y displays the frequency counts of outcomes of (X, Y).

	Y			
X	1	2	$\cdots$	J
1	n_{11}	n_{12}	$\cdots$	n_{1J}
2	n_{21}	n_{22}	$\cdots$	n_{2J}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}

Joint Distribution

We can also tabulate the joint distribution of (X, Y) as an $I \times J$ table. Let $\pi_{ij} = P(X = i, Y = j)$. Then,

	Y			
X	1	2	$\cdots$	J
1	π_{11}	π_{12}	$\cdots$	π_{1J}
2	π_{21}	π_{22}	$\cdots$	π_{2J}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I	π_{I1}	π_{I2}	$\cdots$	π_{IJ}

We must have

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\pi_{ij} = 1.$$

Marginal Distribution

	Y				
X	1	2	$\cdots$	J	Total
1	π_{11}	π_{12}	$\cdots$	π_{1J}	π_{1+}
2	π_{21}	π_{22}	$\cdots$	π_{2J}	π_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	π_{I1}	π_{I2}	$\cdots$	π_{IJ}	π_{I+}
Total	π_{+1}	π_{+2}	$\cdots$	π_{+J}	1

Here

$$\begin{aligned} i\text{th row total: } P\left(X=i\right)=\pi_{i+} &= \sum_{j=1}^{J}\pi_{ij}, \\ j\text{th column total: } P\left(Y=j\right)=\pi_{+j} &= \sum_{i=1}^{I}\pi_{ij}. \end{aligned}$$

Conditional Distribution

	Y				
X	1	2	$\cdots$	J	Total
1	π_{11}	π_{12}	$\cdots$	π_{1J}	π_{1+}
2	π_{21}	π_{22}	$\cdots$	π_{2J}	π_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	π_{I1}	π_{I2}	$\cdots$	π_{IJ}	π_{I+}
Total	π_{+1}	π_{+2}	$\cdots$	π_{+J}	1

Denote

$$\begin{aligned} \pi_{i|j} = P\left(X = i \mid Y = j\right) &= \frac{P\left(X = i, Y = j\right)}{P\left(Y = j\right)}, \\ \pi_{j|i} = P\left(Y = j \mid X = i\right) &= \frac{P\left(X = i, Y = j\right)}{P\left(X = i\right)}. \end{aligned}$$

Independence

Two categorical variables are independent if

$$\pi_{ij} = \pi_{i+}\pi_{+j}, \text{ for all } i \text{ and } j.$$

When X and Y are independent,

$$\begin{aligned} \pi_{i|j} = P\left(X = i \mid Y = j\right) &= P\left(X = i\right) = \pi_{i+}, \\ \pi_{j|i} = P\left(Y = j \mid X = i\right) &= P\left(Y = j\right) = \pi_{+j}. \end{aligned}$$

Sampling

When working with a contingency table, sampling theme is important.

1. In Poisson sampling, the cell counts $\{N_{ij}\}$ follow independent Poisson distributions, i.e., $N_{ij} \sim \text{Poisson}\left(\mu_{ij}\right)$. The joint probability mass function for outcomes $\{n_{ij}\}$ is

$$P\left(N_{11} = n_{11}, \cdots, N_{IJ} = n_{IJ}\right) = \prod_{i=1}^{I}\prod_{j=1}^{J} \frac{\mu_{ij}^{n_{ij}}}{n_{ij}!} \exp\left\{-\mu_{ij}\right\}.$$

   In Poisson sampling, the total sample size $n = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$ is a random variable.

2. In multinomial sampling, the total sample size n is fixed but the row and column totals are not fixed. The cell counts $\{N_{ij}\}$ follow a multinomial distribution

$$P\left(N_{11} = n_{11}, \cdots, N_{IJ} = n_{IJ}\right) = \frac{n!}{n_{11}!n_{12}!\cdots n_{IJ}!} \prod_{i=1}^{I}\prod_{j=1}^{J} \pi_{ij}^{n_{ij}}.$$

Independent Multinomial Sampling

Besides Poisson sampling and multinomial sampling, other sampling themes are possible.

In independent multinomial sampling, the row sums are fixed and the rows follow independent multinomial distributions. Then,

$$P\left(N_{11}=n_{11},\cdots,N_{IJ}=n_{IJ}\right) = \prod_{i=1}^{I}\left[\frac{n_{i+}!}{n_{i1}!n_{i2}!\cdots n_{iJ}!}\prod_{j=1}^{J}\pi_{j|i}^{n_{ij}}\right].$$

where $\pi_{j|i}=P\left(\text{column } j \mid \text{row } i\right)$.

Independent Multinomial Sampling

	Y			
X	1	$\cdots$	J	Total
1	$P(Y=1 \mid X=1)$	$\cdots$	$P(Y=J \mid X=1)$	1
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	$P(Y=1 \mid X=I)$	$\cdots$	$P(Y=J \mid X=I)$	1

	Y		
X	1	$\cdots$	J
1	$P(X=1 \mid Y=1)$	$\cdots$	$P(X=1 \mid Y=J)$
2	$P(X=2 \mid Y=1)$	$\cdots$	$P(X=2 \mid Y=J)$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
I	$P(X=I \mid Y=1)$	$\cdots$	$P(X=I \mid Y=J)$
Total	1	$\cdots$	1

Example: Sampling

Helmet use

Suppose that our data can be represented by the following 2×3 table

	Helmet Use		
Gender	No helmet	Traditional helmet	Airbag helmet
Female			
Male			

Which sampling theme is plausible for the following scenarios?

1. We take all cyclists passing Ångström,
2. We only choose 200 cyclists passing Ångström,
3. We take 100 female and 100 male passing Ångström.

Example: A Case-Control Study

Smoking and Lung Cancer

Suppose that 100 patients with lung cancer were admitted last year. For each patient, we record their past smoking behavior. We take another 100 patients without lung cancer and record their past smoking behavior.

	Lung Cancer	
Smoking	Yes	No
Yes		
No		
Total	100	100

Which sampling theme is plausible?

Example: Another Case-Control Study

Smoking and Lung Cancer

Suppose that 100 non-smokers and 100 smokers are recruited in a study. None of them has lung cancer. We will investigate how many will get lung cancer.

	Lung Cancer		
Smoking	Yes	No	Total
Yes			100
No			100

Which sampling theme is plausible?

Effects of Different Sampling

	Cancer	
Smoking	Yes	No
Yes		
No		

Smoking is a binary variable, and Cancer is also a binary variable.

1. Under multinomial sampling, we can obtain $P(\text{Smoking} \mid \text{Cancer})$ and $P(\text{Cancer} \mid \text{Smoking})$.
2. Under independent multinomial sampling with fixed row sums, we can obtain $P(\text{Cancer} \mid \text{Smoking})$.
3. Under independent multinomial sampling with fixed column sums, we can obtain $P(\text{Smoking} \mid \text{Cancer})$.

Quantity of Interest

Suppose that we have a 2×2 table. Let $\pi_{1|i}$ and $\pi_{1|j}$ be the success probability in row i and column j, respectively. We are often interested in

1. Difference of proportions: $\pi_{1|1} - \pi_{1|2}$,
2. Relative risk: $\pi_{1|1}/\pi_{1|2}$,
3. Odds:

$$\frac{\pi_{1|i}}{1-\pi_{1|i}} = \frac{\pi_{i1}}{\pi_{i2}},$$

4. Odds ratio (denoted by θ):

$$\theta = \frac{\mu_{11}/\mu_{12}}{\mu_{21}/\mu_{22}} = \frac{\pi_{1|1}/\left(1-\pi_{1|1}\right)}{\pi_{1|2}/\left(1-\pi_{1|2}\right)} = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}},$$

where μ_{ij} be the cell expected frequencies corresponding to $X = i$ and $Y = j$.

Sampling

The above quantities often depend on the sampling themes.

- Under multinomial sampling, we can obtain $P(\text{row } i, \text{column } j)$, $P(\text{column } j \mid \text{row } i)$, and $P(\text{row } i \mid \text{column } j)$.
- Under independent multinomial sampling with fixed row sums, we should work with $P(\text{column } j \mid \text{row } i)$, but not $P(\text{row } i \mid \text{column } j)$.

An interesting property of odds ratio is that different sampling themes lead to the same way of computing the odds ratio. That is, the odds ratio can always be computed.

Estimate Odds Ratio in 2 × 2 Table

The sample odds ratio is

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Estimate Odds Ratio

Table: Effect of planting time on the survival of plum root cuttings

	Survival	
Time	Dead	Alive
at once	217	263
in spring	365	115

Independence: Sufficient Condition

In a 2×2 case let $\pi_{ij} = P\left(X = i,\, Y = j\right)$. Suppose that the odds ratio is 1 as

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \quad = \quad 1.$$

Then, we must have

$$\pi_{12}\pi_{21} = \pi_{11}\pi_{22} \quad = \quad \pi_{11}\left(1 - \pi_{11} - \pi_{12} - \pi_{21}\right).$$

Hence,

$$\begin{aligned} \pi_{11} \quad &= \quad \pi_{12}\pi_{21} + \pi_{11}^2 + \pi_{11}\pi_{12} + \pi_{11}\pi_{21} \\ &= \quad \left(\pi_{12} + \pi_{11}\right)\left(\pi_{21} + \pi_{11}\right) \\ &= \quad \pi_{1+}\pi_{+1}. \end{aligned}$$

Likewise, we can show $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j. Hence, $\theta = 1$ implies independence of X and Y.

$\theta = 1$: Sufficient Condition

In a 2×2 table, suppose that X and Y are independent. Then,

$$\pi_{ij} = \pi_{i+}\pi_{+j},$$

for all i and j. Then, the odds ratio satisfies

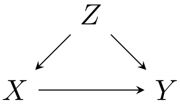
$$\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{\pi_{1+}\pi_{+1} \times \pi_{2+}\pi_{+2}}{\pi_{1+}\pi_{+2} \times \pi_{2+}\pi_{+1}} = 1.$$

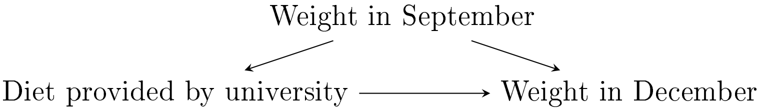
Hence, independence of X and Y implies $\theta = 1$.

Therefore, in a 2×2 table, the odds ratio equals one if and only if X and Y are independent. If $\theta > 1$ (< 1), then the first row is more (less) likely to have a success than the second row, implying dependence.

Confounding

Confounding means that the effect of X on Y depend on the effect of other variables that can influence both X and Y.




Weight in September

Diet provided by university

Weight in December

Simpson's Paradox

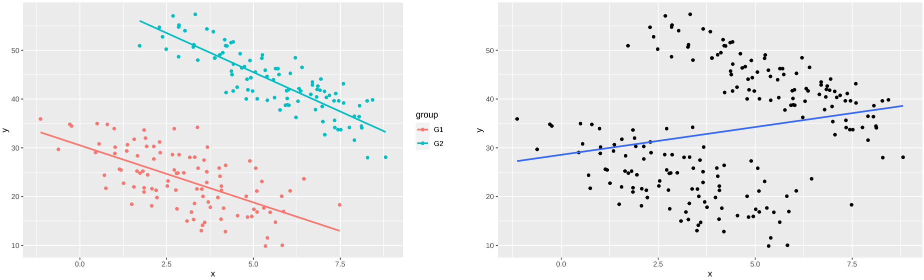
An example that has been analyzed to death is Berkeley university admission rate.

Male		Female	
Applicants	Admitted	Applicants	Admitted
8442	44%	4321	35%

	Male		Female	
Department	Applicants	Admitted	Applicants	Admitted
1	825	62%	108	82%
2	560	63%	25	68%
3	325	37%	493	34%
4	417	33%	375	35%
5	191	28%	393	24%
6	373	6%	341	7%

Women tend to apply to departments with low admission rates, but men tend to apply to departments with high admission rates.

Simpson's Paradox in Regression



Simpson's Paradox In Classification

	Classification	
Classifier	Correct	Incorrect
1	90	10
2	90	10

		Classification	
Observed	Classifier	Correct	Incorrect
Success	1	90	0
	2	81	9
Failure	1	0	10
	2	9	1

Partial Table

Suppose that Z is a confounder (or control variable) when studying the XY relationship. We can use the partial table that fixes the levels of Z. That is, for each level of Z, we make a contingency table for X and Y.

		Admission	
Department	Gender	Admitted	Not admitted
1	Male	512	314
	Female	88	19
2	Male	353	207
	Female	17	8
3	Male	120	205
	Female	168	325

Marginal Table

If we make a two-way contingency table by combining the partial tables, then it is a XY marginal table, ignoring Z.

		Admission	
Department	Gender	Admitted	Not admitted
1	Male	512	314
	Female	88	19
2	Male	353	207
	Female	17	8
3	Male	120	205
	Female	168	325

	Admission	
Gender	Admitted	Not admitted
Male	985	726
Female	273	352

Conditional Association

The associations in partial tables are called conditional associations. Suppose that we have (X, Y, Z) in a $2 \times 2 \times K$ table, where Z is a control variable. Let $\{\mu_{ijk}\}$ be the cell expected frequencies corresponding to $(X = i, Y = j, Z = k)$. Then,

$$\text{conditional odds ratio: } \theta_{XY(k)} = \frac{\mu_{11k}/\mu_{12k}}{\mu_{21k}/\mu_{22k}}, \text{ fixing } Z = k,$$

$$\text{marginal odds ratio: } \theta_{XY} = \frac{\mu_{11+}/\mu_{12+}}{\mu_{21+}/\mu_{22+}},$$

where $\mu_{ij+} = \sum_k \mu_{ijk}$.

Conditional Association

Sample values of $\theta_{XY(k)}$ and θ_{XY} replace μ by n as

$$\text{conditional odds ratio: } \hat{\theta}_{XY(k)} = \frac{n_{11k}/n_{12k}}{n_{21k}/n_{22k}}, \text{ fixing } Z = k,$$

$$\text{marginal odds ratio: } \hat{\theta}_{XY} = \frac{n_{11+}/n_{12+}}{n_{21+}/n_{22+}}.$$

Compute $\hat{\theta}_{XY(k)}$ and $\hat{\theta}_{XY}$

		Admission	
Dept.	Gender	Admitted	Not ad.
1	Male	512	314
	Female	88	19
2	Male	353	207
	Female	17	8
3	Male	120	205
	Female	168	325

	Admission	
Gender	Admitted	Not ad.
Male	985	726
Female	273	352

Different Types of Independence

Suppose that we have (X, Y, Z) in an $I \times J \times K$ table, where Z is a control variable.

- X and Y are conditionally independent at level k of Z if X and Y are independent when $Z = k$:

$$P\left(Y = j \mid X = i, Z = k\right) \quad = \quad P\left(Y = j \mid Z = k\right), \text{ for all } i, j.$$

- X and Y are conditionally independent given Z if X and Y are independent at every value of Z. It is often denoted by $X \perp Y \mid Z$. In other words, given Z, Y does not depend on X.
- X and Y are (marginally) independent if

$$P\left(Y = j \mid X = i\right) \quad = \quad P\left(Y = j\right), \text{ for all } i, j.$$

It is often denoted by $X \perp Y$.

Marginal and Conditional Independence

Suppose that X and Y are conditionally independent given Z. Let $\pi_{ijk} = P(X = i, Y = j, Z = k)$. Then, for all (i, j, k),

$$\begin{aligned}\pi_{ijk} &= P(X = i, Y = j \mid Z = k) P(Z = k) \\ &= P(X = i \mid Z = k) P(Y = j \mid Z = k) P(Z = k) \\ &= \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}.\end{aligned}$$

But,

$$P(X = i, Y = j) = \sum_{k=1}^{K} \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}} \neq \underbrace{\left(\sum_{k=1}^{K} \pi_{i+k}\right)}_{=P(X=i)} \underbrace{\left(\sum_{k=1}^{K} \pi_{+jk}\right)}_{=P(Y=j)}$$

Hence, conditional independence does not imply marginal independence.

Different Types of Independence

X, Y, and Z are mutually independent if

$$\pi_{ijk} = P(X=i)\,P(Y=j)\,P(Z=k)\,, \text{ for all } i,j,k.$$

1. Mutual independence implies marginal independence.

$$\begin{aligned}\pi_{ij+} &= \sum_{k=1}^{K}\pi_{ijk}\\ &= \sum_{k=1}^{K}\left(\pi_{i++}\pi_{+j+}\pi_{++k}\right)\\ &= \pi_{i++}\pi_{+j+}.\end{aligned}$$

2. Mutual independence implies conditional independence.

Back to Odds in $2 \times 2 \times K$ Table

Suppose that X and Y are conditionally independent given Z in a $2 \times 2 \times K$ table. Then, in any partial table with a fixed k, the conditional odds must be

$$\theta_{XY(k)} = \frac{\mu_{11k}/\mu_{12k}}{\mu_{21k}/\mu_{22k}} = 1.$$

Suppose that X and Y are marginally independent ignoring Z in a $2 \times 2 \times K$ table. Then, the marginal odds must be

$$\theta_{XY} = \frac{\mu_{11+}/\mu_{12+}}{\mu_{21+}/\mu_{22+}} = 1.$$

Homogeneous Association

A $2 \times 2 \times K$ table has homogeneous XY association when

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}.$$

That is, the effect of X on Y is the same at each category of Z. When this occurs, we say there is no interaction between two variables in their effects on the other variable.

- Suppose that $X \perp Y \mid Z$. Then the table has homogeneous XY association since $\theta_{XY(k)} = 1$ for all k.
- If there is interaction, the effect of X on Y depends on Z.

Test Homogeneous Association

For a $2 \times 2 \times K$ table, we can test homogeneous association using the Breslow-Day test.

$$\begin{aligned} H_0: &\quad \theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}. \\ H_1: &\quad H_0 \text{ is not true.} \end{aligned}$$

Keep in mind that homogeneous association does not mean that the marginal odds ratio is the same as the conditional odds ratio.

	$Z=1$		$Z=2$	
X	$Y=1$	$Y=2$	$Y=1$	$Y=2$
1	100	20	100	100
2	200	20	60	30

Odds Ratio to $I \times J$ Table

Suppose that we have an $I \times J$ table.

1. There are $\binom{I}{2}$ pairs of rows and $\binom{J}{2}$ pairs of columns. For rows a and b and columns c and d, there are $\binom{I}{2}\binom{J}{2}$ odds ratios of the form
$$\frac{\mu_{ac}/\mu_{ad}}{\mu_{bc}/\mu_{bd}} = \frac{\mu_{ac}\mu_{bd}}{\mu_{bc}\mu_{ad}}.$$
2. The local odds ratios are
$$\theta_{ij} = \frac{\pi_{ij}/\pi_{i+1,j}}{\pi_{i,j+1}/\pi_{i+1,j+1}} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i+1,j}\pi_{i,j+1}},$$
for $i = 1, ..., I-1$ and $j = 1, ..., J-1$. There are $(I-1)(J-1)$ local odds ratios. They determine all odds ratios formed from pairs of rows and pairs of columns.

Maximum Likelihood Estimator

We often estimate a parameter θ (not necessarily odds ratio) by maximum likelihood estimator. Under some regularity conditions, the distribution of the maximum likelihood estimator can be approximated by

$$N\left(\theta, \mathcal{I}^{-1}\left(\theta\right)\right),$$

where

$$\mathcal{I}\left(\theta\right) \quad = \quad \mathrm{var}\left[\frac{\partial \ell\left(\theta\right)}{\partial \theta}\right] = -\mathrm{E}\left[\frac{\partial^2 \ell\left(\theta\right)}{\partial \theta \partial \theta^T}\right]$$

is the Fisher information matrix and $\ell\left(\theta\right)$ is the log-likelihood function.

Wald Statistics and Delta Method

The Wald test statistic for a unidimensional parameter θ is

$$Z = \frac{\hat{\theta} - \theta_0}{\text{standard error of } \hat{\theta}},$$

where θ_0 is some hypothesized value of θ. If the true value of θ is θ_0 and $\hat{\theta}$ is asymptotically normal, then Z is approximately $N(0,1)$. That is,

$$\hat{\theta} - \theta_0 \approx N\left(0, \text{var}\left[\hat{\theta}\right]\right).$$

For a continuously differentiable function $g(\theta)$, the delta method implies that

$$g\left(\hat{\theta}\right) - g(\theta_0) \approx N\left(0, \left[\frac{\partial g(\theta_0)}{\partial \theta^T}\right] \text{var}\left[\hat{\theta}\right] \left[\frac{\partial g(\theta_0)}{\partial \theta^T}\right]^T\right).$$

Likelihood Ratio Test

Suppose that we want to test $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_1$. The likelihood ratio test statistic is

$$\lambda(x) = \frac{\sup\limits_{\theta \in \Theta_0} L(\theta)}{\sup\limits_{\theta \in \Theta_0 \cup \Theta_1} L(\theta)}.$$

Under some regularity conditions,

$$-2 \log \lambda(x) \approx \chi^2_v,$$

when sample size increases, where the degrees of freedom v is the number of free parameters when $\theta \in \Theta_0 \cup \Theta_1$ minus the number of free parameters when $\theta \in \Theta_0$.

Analysis of Categorical Data
Chapter 3: Inference for Contingency Table

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through this chapter, you should be able to

1. test independence in contingency table,
2. test monotone trend.

Odds Ratio

Suppose that we have observed a 2×2 table

	Y	
X	1	2
1	n_{11}	n_{12}
2	n_{21}	n_{22}

The sample odds ratio is

$$\hat{\theta} = \frac{n_{11} n_{22}}{n_{12} n_{21}} \geq 0.$$

If $\hat{\theta} > 0$, then we can consider

$$\log \hat{\theta} = \log n_{11} + \log n_{22} - \log n_{12} - \log n_{21}.$$

Wald Confidence Interval

An estimated standard error of $\log \hat{\theta}$ is

$$\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

Hence, a Wald confidence interval for $\log \theta$ is

$$\log \hat{\theta} \pm z_{\alpha/2} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

However,

$$\hat{\theta} = \frac{n_{11} n_{22}}{n_{12} n_{21}}$$

can be 0 (if $n_{11}n_{22} = 0$), ∞ ($n_{12}n_{21} = 0$), or undefined (if $n_{11}n_{22} = n_{12}n_{21} = 0$). Consequently, the Wald interval may not exist.

- An ad-hoc approach is to add 0.5 to n_{ij}.
- Use other approaches such as the score interval or the likelihood ratio confidence interval.

Example: Aspirin Use and Myocardial Infraction

Compute $\hat{\theta}$ and find a 95% confidence interval for θ

	Myocardial Infraction	
	Yes	No
Placebo	28	656
Aspirin	18	658

Independence

We have an $I \times J$ contingency table from multinomial sampling with probabilities $\{\pi_{ij}\}$.We want to test

$$\begin{aligned} H_0: &\quad \text{independence as } \pi_{ij} = \pi_{i+}\pi_{+j} \text{ for all } i, j, \\ H_1: &\quad H_0 \text{ is not true.} \end{aligned}$$

The log-likelihood under H_1 is

$$\ell_0\left(\pi_{i+}, \pi_{+j}\right) = \log\left(\frac{n!}{n_{11}! \cdots n_{IJ}!}\right) + \sum_i \sum_j n_{ij} \log\left(\pi_{i+}\pi_{+j}\right).$$

The log-likelihood under H_1 is

$$\ell_1\left(\pi_{ij}\right) = \log\left(\frac{n!}{n_{11}! \cdots n_{IJ}!}\right) + \sum_i \sum_j n_{ij} \log\left(\pi_{ij}\right).$$

Likelihood Ratio Test

The MLE under H_0 is

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n},\ \hat{\pi}_{+j} = \frac{n_{+j}}{n}.$$

The MLE under H_1 is

$$\hat{\pi}_{ij} \quad = \quad \frac{n_{ij}}{n}.$$

The likelihood ratio test statistic is

$$G^2 = -2\left[\ell_0\left(\hat{\pi}_{i+}, \hat{\pi}_{+j}\right) - \ell_1\left(\hat{\pi}_{ij}\right)\right] \quad = \quad -2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(\frac{n_{i+}n_{+j}/n}{n_{ij}}\right).$$

If H_0 holds, G^2 also converges in distribution to to chi-square with $(IJ-1)-(I-1)-(J-1) = (I-1)(J-1)$ degrees of freedom. A rule-of-thumb is that no more than 20% of $\hat{\mu}_{ij} < 5$.

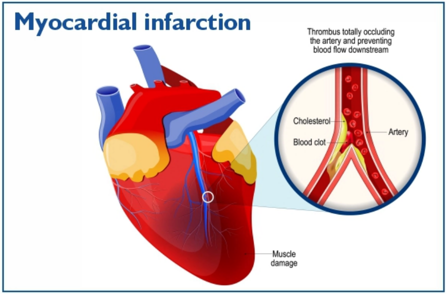
Pearson Chi-Square

The Pearson chi-square that tests H_0 : independence is

$$\begin{aligned} X^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(\text{observed frequency}_{ij} - \text{expected frequency}_{ij}\right)^2}{\text{expected frequency}_{ij}} \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - n\hat{\pi}_{i+}\hat{\pi}_{+j}\right)^2}{n\hat{\pi}_{i+}\hat{\pi}_{+j}}. \end{aligned}$$

If H_0 holds, X^2 converges in distribution to to chi-square with $\left(I-1\right)\left(J-1\right)$ degrees of freedom. A rule-of-thumb is still that no more than 20% of $\hat{\mu}_{ij} < 5$.

Aspirin Use and Myocardial Infarction


Myocardial infarction
Thrombus totally occluding the artery and preventing blood flow downstream
Cholesterol
Blood clot
Artery
Muscle damage

Test independence

	Myocardial Infarction	
	Yes	No
Placebo	28	656
Aspirin	18	658

Fisher's Exact Test

For 2×2 tables, regardless of sampling, under the independence assumption, conditioning on both sets of marginal totals, the only free cell is n_{11}. It follows the hypergeometric distribution

$$P\left(n_{11}=t\right) = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1}-t}}{\binom{n}{n_{+1}}}.$$

- For $H_0 : \theta = 1$ (independence) versus $H_1 : \theta > 1$, the Fisher's exact test uses the p-value $P\left(n_{11} \geq t_o\right)$ where t_o is the observed value of n_{11}.
- For $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$, there are different ways of computing the p-value. They lead to different p-values.

Example

Fisher's Tea Tasting Experiment

	Guess Poured First		
Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

Ordinality and Scoring

If our data are ordinal, using the above X^2 and G^2 are less ideal since they ignore ordinality of data.

To keep ordinality, many people choose to assign scores to the ordinal variables: $u_1 \leq u_2 \leq \cdots \leq u_I$ be the scores for the rows, and $v_1 \leq v_2 \leq \cdots \leq v_J$ be the scores for the columns. The scores are then treated as the values of the variables. However, this approach has several serious issues:

1. How shall we assign scores?
2. Are the distance between the assigned score actually reflect the "distance" between categories?

Ordinal Variables

Suppose that both X and Y are ordinal.

1. A pair of subjects is concordant if the subject ranked higher on X also ranks higher on Y.
2. A pair of subject is discordant if the subject ranking higher on X ranks lower on Y.

	Job satisfaction		
Age	1: Not satisfied	2: Satisfied	3: Very satisfied
1: < 30	34	53	88
2: $30 - 50$	80	174	304
3: > 50	29	75	172

Concordant/Discordant Pairs

	Job satisfaction		
Age	1: Not satisfied	2: Satisfied	3: Very satisfied
1: < 30	34	53	88
2: $30 - 50$	80	174	304
3: > 50	29	75	172

- Subject A belongs to $(1,1)$ and subject B belongs to $(2,2)$. The pair (A, B) is concordant.
- Subject A belongs to $(2,2)$ and subject B belongs to $(1,1)$. Also concordant.
- Subject A belongs to $(1,2)$ and subject B belongs to $(2,1)$. The pair (A, B) is discordant.
- Subject A belongs to $(2,1)$ and subject B belongs to $(1,2)$. Also discordant.

Probability of Concordant/Discordant

Suppose that we have two independent subjects A and B from a joint distribution $\{\pi_{ij}\}$.

1. The probability of a concordant pair is

$$\begin{aligned}\Pi_c &= \sum_{i,j}\{P[A=(i,j)]\,P[B=(h,k),\ h>i,\ k>j \mid A=(i,j)]\} \\ &\quad+\sum_{i,j}\{P[A=(i,j)]\,P[B=(h,k),\ h<i,\ k<j \mid A=(i,j)]\} \\ &= 2\sum_{i,j}\left\{\pi_{ij}\sum_{h>i}\sum_{k>j}\pi_{hk}\right\}.\end{aligned}$$

2. The probability of a discordant pair is

$$\Pi_d = 2\sum_{i,j}\left\{\pi_{ij}\sum_{h>i}\sum_{k<j}\pi_{hk}\right\}.$$

Gamma Coefficient

We define the Goodman-Kruskal's gamma as

$$\gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}.$$

1. γ has the range $-1 \leq r \leq 1$. It work in a similar way as the Pearson correlation coefficient.
2. If $\gamma > 0$ $(\Pi_c > \Pi_d)$, then it is more likely to have concordant pairs than discordant pairs (positive trend).
3. If $\gamma < 0$ $(\Pi_c < \Pi_d)$, then it is less likely to have concordant pairs than discordant pairs (negative trend).
4. If $\gamma = 0$, then no trend.
5. If X and Y are independent, then $\gamma = 0$. But $\gamma = 0$ does not mean independence.

Alternative Method

For ordinal data, we use the sample Goodman-Kruskal's gamma is

$$\hat{\gamma} = \frac{C-D}{C+D}$$

to check whether they have a monotone trend, where C is the total number of concordant pairs of observations, and D is the total number of discordant pairs of observations.

- If $\gamma = 0$, there is no trend between X and Y.
- For a large sample size, $\hat{\gamma}$ is approximately normal.

Example Sample Gamma Coefficient

The sample version of γ is

$$\hat{\gamma} = \frac{C - D}{C + D},$$

where C is the total number of concordant pairs and D is the total number of discordant pairs.

Compute $\hat{\gamma}$

	Job satisfaction		
Age	1: Not satisfied	2: Satisfied	3: Very satisfied
1: < 30	34	53	88
2: $30 - 50$	80	174	304
3: > 50	29	75	172

Wilcoxon Test or Mann-Whitney Test

Suppose that we have two random variables Y_0 and Y_1. The Wilcoxon test or the Mann-Whitney test tests whether

$$P(Y_0 > Y_1) \quad = \quad P(Y_0 < Y_1).$$

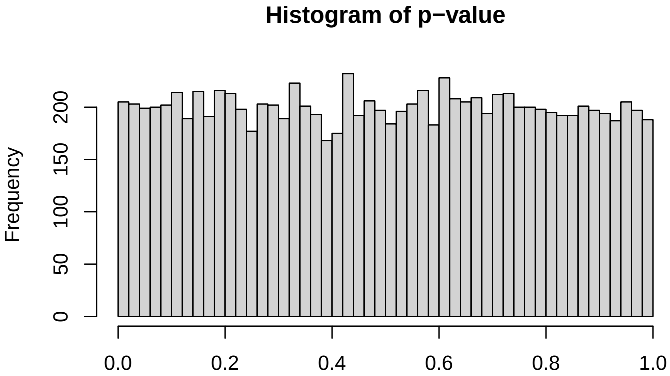
In the special case where we have a $2 \times J$ table $(I = 2)$ and the scores for X are $\{0, 1\}$. We have two groups, one group with $X = 0$ and another group with $X = 1$. Then the general idea is that

1. Assign ranks to the whole sample of size $n_{0+} + n_{1+}$.
2. Compute the sum of ranks assigned to the group $X = 0$.
3. If H_0 is not true, the sum of ranks tends to be either small or large.

The Kruskal-Wallis test generalizes the Mann-Whitney test to more than 2 groups. The Kruskal-Wallis test can be viewed as a non-parametric version of one-way ANOVA.

Be Careful With Their Hypotheses

	Y			
X	1	2	3	4
0	0.05	0.5	0.35302019	0.09697981
1	0.1666553	0.2833447	0.5000000	0.0500000



Analysis of Categorical Data
Chapter 4: Introduction to Generalized Linear Models

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through this chapter, you should be able to

1. verify exponential dispersion family,
2. describe the components of GLM,
3. fit GLMs,
4. perform model comparison,
5. perform residual analysis.

Exponential Dispersion Family

A random variable Y_i belongs to the exponential dispersion family if the pmf/pdf is of the form

$$f\left(y_i;\theta_i,\phi_i\right)=\exp\left\{\frac{y_i\theta_i-b\left(\theta_i\right)}{\phi_i}+c\left(y_i,\phi_i\right)\right\}.$$

- θ_i is the natural parameter.
- $\phi_i>0$ is the dispersion parameter, which can be either known or unknown. We often have $\phi_i=\phi$ or $\phi_i=\phi/w_i$ with a known w_i.
- No y_i can be included in $b\left(\theta_i\right)$.
- No θ_i can be included in $c\left(y_i,\phi_i\right)$.

Example: Poisson Distribution

- The pmf of a Poisson distribution $\text{Poisson}\left(\mu_i\right)$ is

$$P\left(Y_i=y_i\right)=\frac{\mu_i^{y_i}}{y_i!}\exp\left\{-\mu_i\right\}=\exp\left\{y_i\log\left(\mu_i\right)-\mu_i-\log\left(y_i!\right)\right\},$$

which does not directly fit into the exponential dispersion family

$$f\left(y_i;\theta_i,\phi\right)=\exp\left\{\frac{y_i\theta_i-b\left(\theta_i\right)}{\phi_i}+c\left(y_i,\phi_i\right)\right\}.$$

- However, if we define $\theta_i=\log\left(\mu_i\right)$, then

$$P\left(Y_i=y_i\right)=\exp\left\{\frac{y_i\theta_i-\exp\left(\theta_i\right)}{1}-\log\left(y_i!\right)\right\}.$$

Here $\phi_i=1$, which is a constant.

Example: Binomial Distribution

- The pmf of a binomial distribution $\text{Bin}(n_i, \pi_i)$ with n_i being the total number of trials and π_i being the success probability is

$$P(Z_i = z_i) = \begin{pmatrix} n_i \\ z_i \end{pmatrix} \pi_i^{z_i} (1-\pi_i)^{n_i - z_i}$$
$$= \exp\left\{ z_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \log(1-\pi_i) + \log\begin{pmatrix} n_i \\ z_i \end{pmatrix} \right\},$$

whose expectation depends on n_i.

- Define $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$ and consider $Y_i = Z_i / n_i$, then

$$P(Y_i = y_i) = \exp\left\{ \frac{y_i\theta_i - \log[1+\exp(\theta_i)]}{1/n_i} + \log\begin{pmatrix} n_i \\ n_i y_i \end{pmatrix} \right\}.$$

Here $\phi_i = \phi / w_i$ with $\phi = 1$ and $w_i = n_i$.

Moments of Exponential Family

For the exponential dispersion family,

$$\begin{aligned}\mathbb{E}\left(Y_i\right) &= b'\left(\theta_i\right), \\ \operatorname{var}\left(Y_i\right) &= \phi_i b''\left(\theta_i\right),\end{aligned}$$

where $V\left(\theta_i\right) = b''\left(\theta_i\right)$ is called the variance function.

Components of Generalized Linear Model

1. Random component: Response variable Y_i and its probability distribution from exponential dispersion family.
2. Linear predictor $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$: Model matrix $\boldsymbol{X}$ of size $n \times p$ and parameter vector $\boldsymbol{\beta}$ of size $p \times 1$. The linear predictor for y_i is

$$\eta_i \quad = \quad \boldsymbol{x}_i^T \boldsymbol{\beta} = \sum_{j=1}^{p} x_{ij}\beta_j,$$

   where $\boldsymbol{x}_i^T$ is the ith row of $\boldsymbol{X}$.
3. Link function $g\left(\right)$: $g\left(\right)$ transforms $\mu_i = \mathbb{E}\left(Y_i\right)$ to the linear predictor

$$g\left(\mu_i\right) \quad = \quad \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta},$$

   The link function must be monotonic and differentiable.

Examples of Link Functions

Suppose that Y_i follows a Bernoulli distribution $(n_i = 1)$ or a binomial distribution $(n_i \neq 1)$. The most common link function is the logit link (logistic model or logit model):

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

Suppose that Y_i follows a Poisson distribution. The link function is often the log-link $g(\mu) = \log \mu$.

Everything is Connected

A GLM transforms μ_i through the link function $g\left(\mu_i\right) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$.

θ_i, μ_i, η_i, $\boldsymbol{\beta}$ are all connected through $b\left(\theta_i\right)$ and $g\left(\mu_i\right)$.

$$\begin{tikzcd}
\theta_i \arrow[r, Leftrightarrow, "\mu_i = b'\left(\theta_i\right)"] & \mu_i \arrow[r, Leftrightarrow, "\eta_i = g\left(\mu_i\right)"] & \eta_i & \boldsymbol{\beta} \arrow[l, Rightarrow, "\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}"]
\end{tikzcd}$$

Suppose that $b\left(\theta_i\right) = \exp\left(\theta_i\right)$ and $g\left(\mu_i\right) = \mu_i^3$. Then,

$$\begin{tikzcd}
\theta_i \arrow[r, Leftrightarrow, "\mu_i = \exp\left(\theta_i\right)", "\theta_i = \log\left(\mu_i\right)"'] & \mu_i \arrow[r, Leftrightarrow, "\eta_i = \mu_i^3", "\mu_i = \eta_i^{1/3}"'] & \eta_i & \boldsymbol{\beta} \arrow[l, Rightarrow, "\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}"]
\end{tikzcd}$$

Canonical Link

- The link function of a GLM transforms the mean of the random component to the linear predictor $\eta_i = g\left(\mu_i\right)$.
- The link function that transforms the mean μ_i to the natural parameter θ_i is called the canonical link.

$$\begin{aligned} \theta_i &= g\left(\mu_i\right) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, && \text{canonical link}, \\ \theta_i &\neq g\left(\mu_i\right) = \eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, && \text{otherwise}. \end{aligned}$$

- For a Poisson distribution, the canonical link is the log link.
- For a binomial distribution, the canonical link is the logit link.

Likelihood in Exponential Family

- For n independent observations, the likelihood is the product of densities or mass functions:

$$\prod_{i=1}^{n} f\left(y_i;\theta_i,\phi_i\right) = \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b\left(\theta_i\right)}{\phi_i} + c\left(y_i,\phi_i\right)\right\}.$$

- The log-likelihood is

$$\sum_{i=1}^{n}\left\{\frac{y_i\theta_i - b\left(\theta_i\right)}{\phi_i} + c\left(y_i,\phi_i\right)\right\}.$$

The log-likelihood will be denoted by $\ell\left(\boldsymbol{\mu};\boldsymbol{y}\right)$, where the ith entry of $\boldsymbol{\mu}$ is $\mu_i = \mathbb{E}\left(Y_i\right)$ and the ith entry of $\boldsymbol{y}$ is y_i.

Maximum Likelihood Estimator

Since $g\left(\mu_i\right) = \boldsymbol{x}_i^T \boldsymbol{\beta}$ and $\mu_i = \mathbb{E}\left(Y_i\right) = b'\left(\theta_i\right)$, θ_i is a function of $\boldsymbol{\beta}$. We can maximize the log-likelihood

$$\ell = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b\left(\theta_i\right)}{\phi_i} + c\left(y_i, \phi_i\right) \right\}$$

to obtain the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$, denoted by $\hat{\boldsymbol{\beta}}$.

The gradient be expressed as

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{V}^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}\right),$$

where $\boldsymbol{X}_{n\times p}$ is the model matrix, $\boldsymbol{D}_{n\times n}$ is the diagonal matrix with (i,i)th element $\partial \mu_i / \partial \eta_i$, and $\boldsymbol{V}_{n\times n}$ is a diagonal matrix with (i,i)th element $\operatorname{var}\left(Y_i\right)$.

Example: Find Score Function

Gradient of Poisson regression

Consider the Poisson regression model, where $Y_i \sim \text{Poisson}\left(\mu_i\right)$ and $\log\left(\mu_i\right) = \eta_i = \beta_1 + \beta_2 x_i$. Show that

$$\begin{aligned} \boldsymbol{D} &= \operatorname{diag}\left\{\exp\left(\beta_1 + \beta_2 x_i\right)\right\}, \\ \boldsymbol{V} &= \operatorname{diag}\left\{\exp\left(\beta_1 + \beta_2 x_i\right)\right\}. \end{aligned}$$

General Problem

- Consider a general problem that, for a scalar-valued function $h(\boldsymbol{\beta})$, we need to find the solution of

$$\mathbf{0} = \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}.$$

- The solution is approximately the solution of

$$\mathbf{0} = \frac{\partial h(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \approx \frac{\partial h\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta}} + \frac{\partial^2 h\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \left(\boldsymbol{\beta} - \boldsymbol{\beta}^{(t)}\right)$$

for some known $\boldsymbol{\beta}^{(t)}$, which yields

$$\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)} - \left(\frac{\partial^2 h\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right)^{-1} \frac{\partial h\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta}},$$

if the Hessian matrix is invertible.

Newton-Raphson Method or Newton's Method

We can name a first guess of $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{(0)}$, and update parameter estimates using

$$\begin{aligned}
\boldsymbol{\beta}^{(1)} &\approx \boldsymbol{\beta}^{(0)} - \left(\frac{\partial^2 h\left(\boldsymbol{\beta}^{(0)}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)^{-1} \frac{\partial h\left(\boldsymbol{\beta}^{(0)}\right)}{\partial\boldsymbol{\beta}}, \\
\boldsymbol{\beta}^{(2)} &\approx \boldsymbol{\beta}^{(1)} - \left(\frac{\partial^2 h\left(\boldsymbol{\beta}^{(1)}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\right)^{-1} \frac{\partial h\left(\boldsymbol{\beta}^{(1)}\right)}{\partial\boldsymbol{\beta}}, \\
&\vdots
\end{aligned}$$

until $\frac{\partial h\left(\boldsymbol{\beta}^{(t+1)}\right)}{\partial\boldsymbol{\beta}}$ is sufficiently close to 0 or $\boldsymbol{\beta}^{(t+1)}$ and $\boldsymbol{\beta}^{(t)}$ are sufficiently close.

Newton-Raphson in GLM

In GLM, we need to find the solution of

$$\mathbf{0} = \frac{\partial \ell\left(\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T \boldsymbol{D} \boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}\right).$$

The Newton-Raphson in GLM updates the parameter estimator as

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} - \left(\frac{\partial^2 \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right)^{-1} \frac{\partial \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta}} \\ &= \boldsymbol{\beta}^{(t)} + \left(-\frac{\partial^2 \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right)^{-1} \frac{\partial \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta}} \end{aligned}$$

until convergence. Here, we are taking the inverse of the observed information matrix.

Newton-Raphson to Fisher Scoring

- The Newton-Raphson method updates the parameter estimator as

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left(-\frac{\partial^2 \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right)^{-1} \frac{\partial \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta}}.$$

- The Fisher scoring updates the parameter estimator as

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} + \left[E\left(-\frac{\partial^2 \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right) \right]^{-1} \frac{\partial \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta}} \\ &= \boldsymbol{\beta}^{(t)} + \left[\boldsymbol{\mathcal{I}}\left(\boldsymbol{\beta}^{(t)}\right) \right]^{-1} \frac{\partial \ell\left(\boldsymbol{\beta}^{(t)}\right)}{\partial \boldsymbol{\beta}}, \end{aligned}$$

where $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}$ for GLM with $\boldsymbol{W} = \boldsymbol{D} \boldsymbol{V}^{-1} \boldsymbol{D}$.

Iterative Reweighted Least Squares

- Plugging in the expression of information matrix and score function, Fisher scoring becomes

$$\begin{aligned}\boldsymbol{\beta}^{(t+1)} &= \left(\boldsymbol{X}^T \boldsymbol{W}^{(t)} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}^{(t)} \left[\boldsymbol{X}\boldsymbol{\beta}^{(t)} + \left(\boldsymbol{D}^{(t)}\right)^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}^{(t)}\right)\right] \\ &= \left(\boldsymbol{X}^T \boldsymbol{W}^{(t)} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}^{(t)} \boldsymbol{z}^{(t)}.\end{aligned}$$

- This means that, at each step, $\boldsymbol{\beta}$ is updated using weighted least squares with closed forms using the adjusted response variable $\boldsymbol{z}^{(t)}$.
- In other words, for GLM, estimators are obtained by an iterative reweighted least squares (IRLS) procedure.

Biproduct: Standard Error

- The IRLS procedure updates the parameter estimates by

$$\boldsymbol{\beta}^{(t+1)} = \left(\boldsymbol{X}^T\boldsymbol{W}^{(t)}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{W}^{(t)}\left[\boldsymbol{X}\boldsymbol{\beta}^{(t)} + \left(\boldsymbol{D}^{(t)}\right)^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}^{(t)}\right)\right].$$

- If n is large enough and all assumptions are correct, the distribution of $\hat{\boldsymbol{\beta}}$ can be approximated by

$$N\left(\boldsymbol{\beta}, \left(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\right).$$

  where $\hat{\boldsymbol{W}}$ is the latest $\boldsymbol{W}$ from IRLS.

- The standard error of $\hat{\beta}_j$ can be approximated by $\sqrt{\hat{\tau}_j}$, where $\hat{\tau}_j$ is the (j, j)th element of $\left(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}$.

Prediction

Once we have obtained $\hat{\boldsymbol{\beta}}$, we can predict η by $\hat{\eta} = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}$ and μ by $\hat{\mu} = g^{-1}\left(\boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}\right)$, where $\boldsymbol{x}_0$ is the vector of regressors/features, and $g^{-1}()$ is the inverse function of $g()$.

- The distribution of $\hat{\eta} = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}$ is then approximately
$$N\left(\boldsymbol{\eta},\ \boldsymbol{x}_0^T \left(\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X}\right)^{-1} \boldsymbol{x}_0\right).$$
- A $1-\alpha$ confidence interval for η is
$$\hat{\eta} \pm z_{1-\alpha/2} \sqrt{\boldsymbol{x}_0^T \left(\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X}\right)^{-1} \boldsymbol{x}_0},$$
where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of $N(0,1)$.
- The $1-\alpha$ confidence interval for μ is
$$g^{-1}\left(\hat{\eta} \pm z_{1-\alpha/2} \sqrt{\boldsymbol{x}_0^T \left(\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X}\right)^{-1} \boldsymbol{x}_0}\right).$$

Maximum log-Likelihood of Our Model

- Given $\hat{\boldsymbol{\beta}}$, the fitted μ_i is $\hat{\mu}_i = g^{-1}\left(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}\right)$, where $g^{-1}()$ is the inverse function of $g()$.
- The fitted θ_i, denoted by $\hat{\theta}_i$, is the solution of $\hat{\mu}_i = b'\left(\hat{\theta}_i\right)$.
- The likelihood of our model becomes

$$L\left(\hat{\boldsymbol{\mu}};\boldsymbol{y}\right) \equiv \prod_{i=1}^{n}\exp\left\{\frac{y_i\hat{\theta}_i - b\left(\hat{\theta}_i\right)}{\phi_i} + c\left(y_i,\phi_i\right)\right\}.$$

Saturated Model

The saturated model that fits the data "perfectly" uses y_i to estimate μ_i for all i, i.e., $\hat{\mu}_i = y_i$.

- Since $\mu_i = b'(\theta_i)$, the fitted θ_i is the solution of $\hat{\mu}_i = b'\left(\hat{\theta}_i\right)$.
- NOTE: there is no $\boldsymbol{\beta}$ directly involved here.

The likelihood of the saturated model is

$$L(\boldsymbol{y}; \boldsymbol{y}) \equiv \prod_{i=1}^{n} \exp\left\{ \frac{y_i \hat{\theta}_i^{(s)} - b\left(\hat{\theta}_i^{(s)}\right)}{\phi_i} + c(y_i, \phi_i) \right\},$$

where the superscript denotes that it is the saturated model.

(Residual) Deviance

Consider testing

$$\begin{aligned} H_0 &: \text{ The model fits the data as good as the saturated model} \\ H_1 &: \text{ The model fits the data worse than the saturated model} \end{aligned}$$

The likelihood ratio test statistic is $-2\log\left(\frac{L(\hat{\boldsymbol{\mu}};\boldsymbol{y})}{L(\boldsymbol{y};\boldsymbol{y})}\right)$.

In Poisson GLM or binomial GLM, the (residual) deviance is

$$D\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right) \quad = \quad -2\log\left(\frac{L\left(\hat{\boldsymbol{\mu}};\boldsymbol{y}\right)}{L\left(\boldsymbol{y};\boldsymbol{y}\right)}\right),$$

where ϕ_i is known in both models. If the model fits the data well, $D\left(\boldsymbol{y};\hat{\boldsymbol{\mu}}\right) \approx \chi^2\left(m-p\right)$, where *m is the number of parameters in the saturated model*, p is the number of parameters in the model of interest, and m should not increases as n increases.

Example: Deviance for Binomial model

In a binomial model,

$$P\left(Y_i = y_i\right) = \left(\begin{array}{c} n_i \\ y_i \end{array}\right) \pi_i^{y_i} \left(1 - \pi_i\right)^{n_i - y_i}.$$

Our model yields predicted probability $\hat{\pi}_i$. Hence, the deviance is

$$D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}\right) = -2 \log \left(\frac{\prod_{i=1}^{n} \left(\begin{array}{c} n_i \\ y_i \end{array}\right) \hat{\pi}_i^{y_i} \left(1 - \hat{\pi}_i\right)^{n_i - y_i}}{\prod_{i=1}^{n} \left(\begin{array}{c} n_i \\ y_i \end{array}\right) y_i^{y_i} \left(1 - y_i\right)^{n_i - y_i}} \right).$$

Grouped Data and Ungrouped Data

```
####  Ungrouped data
Ungroup

##     y x1         x2
## 1  0  0  0.8458632
## 2  0  0  0.6726630
## 3  1  0 -0.4372080
## 4  0  0 -1.4194868
## 5  1  0  0.8742662
## 6  1  1 -0.7330018
## 7  1  1 -0.8285645
## 8  0  1 -0.2341681
## 9  0  1  0.5203699
## 10 1  1  0.1571108
## 11 0  1  0.2665822
## 12 0  1  0.2124662
```

```
####  Grouped data
Group

##   fail success x1 x2
## 1    2       1  0  0
## 2    1       1  0  1
## 3    1       2  1  0
## 4    3       1  1  1
```

Grouped Data Expressed as Ungrouped

```
##    y x1 x2
## 1  0  0  0
## 2  0  0  0
## 3  1  0  0
## 4  0  0  1
## 5  1  0  1
## 6  1  1  0
## 7  1  1  0
## 8  0  1  0
## 9  0  1  1
## 10 1  1  1
## 11 0  1  1
## 12 0  1  1
```

```
##   fail success x1 x2
## 1    2       1  0  0
## 2    1       1  0  1
## 3    1       2  1  0
## 4    3       1  1  1
```

Grouped Data Expressed as Ungrouped

```
##
## Call:
## glm(formula = y ~ x1 + x2, family = binomial(), data = DF)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2310  -0.9793  -0.8850   1.1513   1.5585
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.1251     1.0238  -0.122    0.903
## x1            0.2502     1.2310   0.203    0.839
## x2           -0.7372     1.2141  -0.607    0.544
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 16.301  on 11  degrees of freedom
## Residual deviance: 15.914  on  9  degrees of freedom
## AIC: 21.914
##
```

Grouped Data Expressed as Ungrouped

```
##
## Call:
## glm(formula = cbind(success, fail) ~ x1 + x2, family = binomial(),
##     data = NewDF)
##
## Deviance Residuals:
##       1        2        3        4
## -0.4758   0.6007   0.4758  -0.4373
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.1251     1.0238  -0.122    0.903
## x1            0.2502     1.2310   0.203    0.839
## x2           -0.7372     1.2141  -0.607    0.544
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.3912  on 3  degrees of freedom
## Residual deviance: 1.0049  on 1  degrees of freedom
## AIC: 12.261
```

Null Model and Null Deviance

- Consider a special model where only the intercept is included

$$g\left(\mu_i\right) \quad = \quad \beta_0,$$

  with $p = 1$.
- The fitted mean for individual i is $\hat{\mu}_i = g^{-1}\left(\beta_0\right)$, which is the same for all i.
- The estimator of θ_i is obtained from $\hat{\mu}_i = b'\left(\theta_i\right)$, still the same for all i.
- This is called a null model and its residual deviance is called the null deviance.
  - The null model represents the worst model that we can build.
  - The null deviance compares the null model with the saturated model.

Compare Two Models

- Suppose that we have two models (M_0 and M_1) and that M_0 nested in M_1 with different $\boldsymbol{x}$. The deviances for M_0 and M_1 are

$$M_0: \quad D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0\right) \quad \text{and} \quad M_1: \quad D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1\right).$$

- In binomial GLM or Poisson GLM, the difference in the deviance is

$$G^2\left(M_0|M_1\right) \overset{\text{def}}{=} D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0\right) - D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1\right),$$

which is the test statistic for H_0: M_0 versus H_1: M_1.

- We reject H_0 if

$$G^2\left(M_0|M_1\right) \geq \chi^2_{1-\alpha}\left(p_1 - p_0 > 0\right),$$

where M_0 has p_0 parameters and M_1 has p_1 parameters.

AIC: Minimizing Distance of the Fit from the Truth

- The Akaike information criterion (AIC) is a nearly “unbiased” estimator of the “distance” between the assumed model and the unknown truth.
- It is a penalized log-likelihood

$$\text{AIC} = -2\ell\left(\hat{\boldsymbol{\beta}}_M\right) + 2 \cdot \text{number of parameters in model } M.$$

- AIC is NOT

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu\right)^2 + 2 \cdot \text{number of parameters in model } M.$$

- We prefer to model with the smallest AIC or a parsimonious model that has AIC near the minimum.
- In practice, AIC tends to be conservative, in the sense that it tends to select more explanatory variables.

BIC: Consistent Model Selection

- Bayesian information criterion penalizes a complex model much more than AIC.

$$\text{BIC} = -2\ell\left(\hat{\boldsymbol{\beta}}_M\right) + \log(n) \cdot \text{number of parameters in model } M.$$

- We prefer to model with the smallest BIC or a parsimonious model that has BIC near the minimum.
- BIC is consistent in model selection in the sense that

$$P\left(\text{Choose the true model if it is a candidate}\right) \to 1, \text{ as } n \to \infty.$$

- In contrast, AIC is not consistent.

Pearson Residual and Deviance Residual

- The Pearson residual is

$$\frac{y_i - \hat{\mu}_i}{\sqrt{V\left(\hat{\mu}_i\right)}}.$$

- The deviance residual is

$$\operatorname{sign}\left(y_i - \hat{\mu}_i\right) \cdot \sqrt{d_i},$$

  whose squared values sum to the deviance.

It is often suggested to perform residual checking to investigate whether any patterns can be observed for grouped data. If the model fits data well, we should not observe any trends.

Unfortunately...

Unfortunately, the residual plots for models fitted by glm() may not be useful, when we have ungrouped data.

- We generate binary data from a logistic model

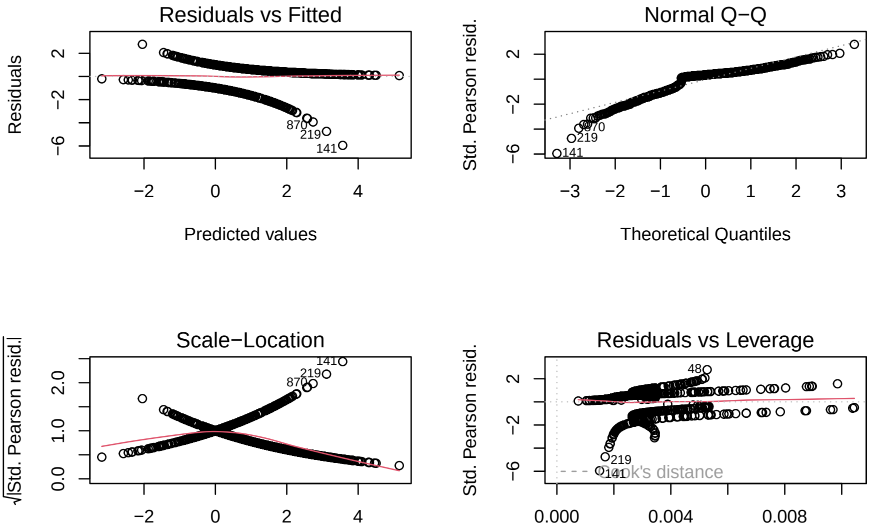
$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \beta_x x_2,$$

  where x_1 takes values 0 or 1, and x_2 is continuous.

- We fit the model using

```
logit <- glm(y ~ x1 + x2, data = Data, family = binomial())
```

Residual Plots

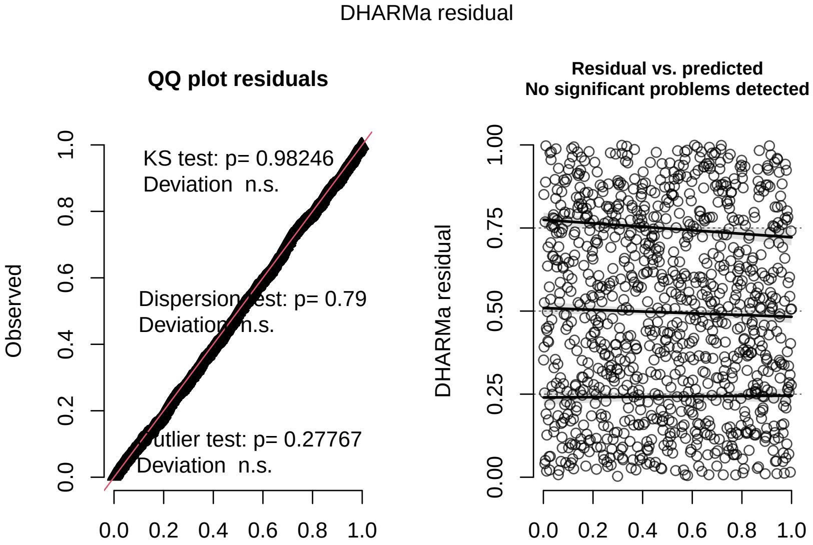


Alternative: Randomized Quantile Residuals

1. Fit your model using glm() or other functions
2. Simulate (randomized) quantile residuals using simulateResiduals()
    1. First, for each observation i, simulate q response variables using the predicted μ_i.
    2. Second, for each observation i, compute the percentage that simulated response less than y_i and the percentage that simulated response less than or equal to y_i.
    3. Third, if two percentages are the same, the quantile residual is the percentage. If not the same, the randomized quantile residual is draw from a uniform distribution between two percentages.
3. Plot the (randomized) quantile residuals using plot().

If your model is correct, the cdf of y_i follows a uniform distribution. Hence, we expect the quantile residuals to be uniform and spread out everywhere.

Randomized Quantile Residual Plots



Analysis of Categorical Data
Chapter 5 and 6: Logistic Regression

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through this chapter, you should be able to

1. make inference for a logistic model,
2. perform model diagnostic/selection,
3. estimate odds ratio from a logistic model,
4. test conditional independence,
5. test homogeneous association.

Logistic Regression

In general, a logistic regression model is of the form

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) \quad = \quad \boldsymbol{\beta}^T \boldsymbol{x}_i, \quad i = 1, ..., n,$$

where $\pi_i = P\left(Y_i = 1 \mid \boldsymbol{x}_i\right)$ and $Y_i \mid \boldsymbol{x}_i \sim \text{Binomial}\left(m_i, \pi_i\right)$.

- The link function is the logit link $g\left(\pi\right) = \log\left(\frac{\pi_i}{1-\pi_i}\right)$.
- We can fit the model by IRLS.

Interpretation of Logistic Model

Consider the logistic model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) \quad = \quad \beta_0+\beta_1 x_{i1}+\cdots+\beta_p x_{ip}.$$

Then,

$$\begin{aligned} & \frac{P\left(Y_i=1|x_{i1}=a+1,x_{i2},...,x_{ip}\right)/P\left(Y_i=0|x_{i1}=a+1,x_{i2},...,x_{ip}\right)}{P\left(Y_i=1|x_{i1}=a,x_{i2},...,x_{ip}\right)/P\left(Y_i=0|x_{i1}=a,x_{i2},...,x_{ip}\right)} \\ = \quad & \frac{\exp\left\{\beta_0+\beta_1\left(a+1\right)+\beta_2x_2\right\}}{\exp\left\{\beta_0+\beta_1a+\beta_2x_2\right\}}=\exp\left\{\beta_1\right\}\end{aligned}$$

is the (conditional) odds ratio, adjusting for other covariates.

Generally speaking, β_j is the expected change in the log odds for one unit increase in x_{ij}, holding the other terms fixed.

Sampling: Prospective or Retrospective

Sometimes (e.g., a case-control study), X is random instead of Y.

Prospective study

	Cancer		
Smoking	Yes	No	Total
Yes	n_{11}	n_{12}	n_{1+}
No	n_{21}	n_{22}	n_{2+}

Case-Control study

	Cancer	
Smoking	Yes	No
Yes	n_{11}	n_{12}
No	n_{21}	n_{22}
Total	n_{+1}	n_{+2}

In a prospective study, we have $P(Y \mid X)$. In a case-control study, we have $P(X \mid Y)$. We can still build a logistic model to model $P(Y \mid X)$ among the selected subjects in a case-control study, the β can still be estimated. Hence, we can still estimate the odds ratio.

Qualitative Explanatory Variables

In our course, we mainly work with logistic models with categorical x.

Consider an $I \times 2$ table. In row i, let y_i be the number of successes out of n_i trials. We can treat y_i as $Y_i \sim \text{Bin}\left(n_i, \pi_i\right)$.

- The corresponding logit model is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_i,$$

  expressed as the model in one-way ANOVA.

- Using dummy variables, the model becomes

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_{I-1} x_{I-1} + \beta_I x_I,$$

  where x_j's are the dummy variables.

Identification and Interpretation

For identification, we need $\beta_1 = 0$ or $\beta_I = 0$, or other conditions.

- Suppose that $\beta_I = 0$ such that $x_i = i$ if the observations are in row i. Then,

$$\begin{aligned}\log\left(\frac{\pi_i}{1-\pi_i}\right) &= \alpha + \beta_i, \quad i = 1, ..., I-1,\\ \log\left(\frac{\pi_I}{1-\pi_I}\right) &= \alpha.\end{aligned}$$

- α is the log odds for row I, and $\alpha + \beta_i$ is the log odds for row i.
- β_i is the log odds ratio between row i and I.
- $\beta_i - \beta_j$ is the log odds ratio between row i and j.

Test β_i

We know from the general theory of GLM that

$$\hat{\boldsymbol{\beta}} \sim N\left(\boldsymbol{\beta}, \left(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\right).$$

- We can test individual β_i using

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\left[\left(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\right]_{ii}}} \sim N(0,\ 1).$$

- We can test a linear combination $\boldsymbol{c}^T\boldsymbol{\beta}$ using

$$\frac{\boldsymbol{c}^T\hat{\boldsymbol{\beta}} - \boldsymbol{c}^T\boldsymbol{\beta}}{\sqrt{\boldsymbol{c}^T\left(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}\boldsymbol{c}}} \sim N(0,\ 1).$$

Saturated Model and Null Model

Consider the model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) \quad = \quad \alpha+\beta_i,$$

where $i=1,2,...,I$.

- The model is saturated if the model has I parameters. The MLE satisfies $\hat{\pi}_i = Y_i/n_i$.
- In the null model $\beta_i=0$ for all i, then $\operatorname{logit}(\pi_i)=\alpha$ and

  $$P\left(Y=1\mid X=i\right) \quad = \quad \frac{\exp\{\alpha\}}{1+\exp\{\alpha\}},$$

  implying that X and Y are independent.

  - What can we use null deviance for?

Ordinal Predictor

If we formulate the model as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_i,$$

then we treat the categorical X as nominal.

If X is ordinal, then it is always difficult to treat. Two alternatives are

1. Assign scores and use scores as the continuous covariates. But the scores can affect the results.
2. Treat ordinal X as nominal X. But we have information loss.

Example: Heart Disease

We have a sample of males. The response variable is whether they developed coronary heart disease. The explanatory variable is the blood pressure level.

```
Data

##   with without pressure
## 1    3     153     <117
## 2   17     235  117-126
## 3   12     272  127-136
## 4   16     255  137-146
## 5   12     127  147-156
## 6    8      77  157-166
## 7   16      83  167-186
## 8    8      35     >186
```

Pressure as Ordinal: Model Fitting

If we treat pressure as an ordinal variable, then we can assign scores of your choice to it and fit a logistic model.

```
Data$score <- c(111.5, 121.5, 131.5, 141.5, 151.5, 161.5, 176.5, 191.5)
Logit <- glm(cbind(with, without) ~ score, family = binomial, data = Data)
summary(Logit)


##
## Call:
## glm(formula = cbind(with, without) ~ score, family = binomial,
##     data = Data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0617  -0.5977  -0.2245   0.2140   1.8501
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.082033   0.724320  -8.397  < 2e-16 ***
## score        0.024338   0.004843   5.025 5.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 30.0226  on 7  degrees of freedom
## Residual deviance:  5.9092  on 6  degrees of freedom
## AIC: 42.61
##
## Number of Fisher Scoring iterations: 4
```

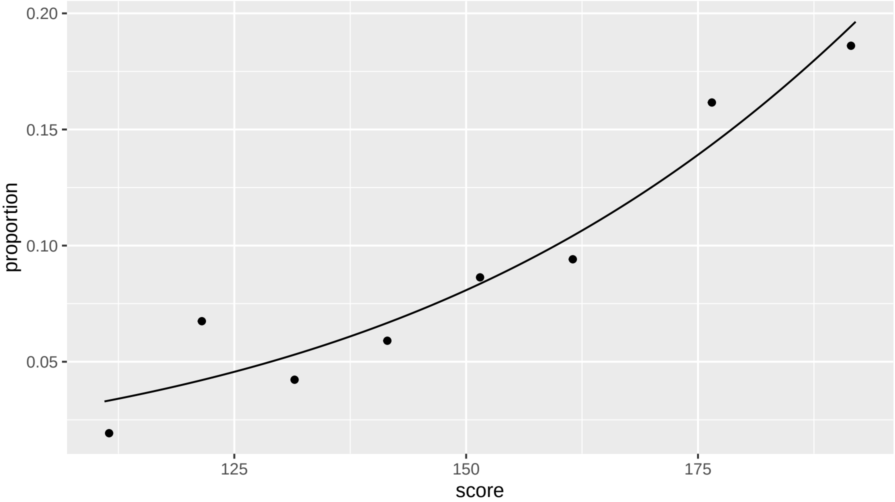
Pressure as Ordinal: Residuals

We can compute the Pearson residual and the standardized Pearson residual. The latter is closer to $N(0,1)$ if the model holds.

```
## Pearson residual
residuals(Logit, type = "pearson")

##          1          2          3          4          5          6          7
## -0.9794311  2.0057103 -0.8133348 -0.5067270  0.1175833 -0.3042459  0.5134721
##          8
## -0.1394648

## Standardized Pearson residual
rstandard(Logit, type = "pearson")

##          1          2          3          4          5          6          7
## -1.1057850  2.3746058 -0.9452701 -0.5727440  0.1260886 -0.3260730  0.6519547
##          8
## -0.1773473
```

Pressure as Ordinal: Plots

We can plot the observed proportions and compare them with the fitted curve.

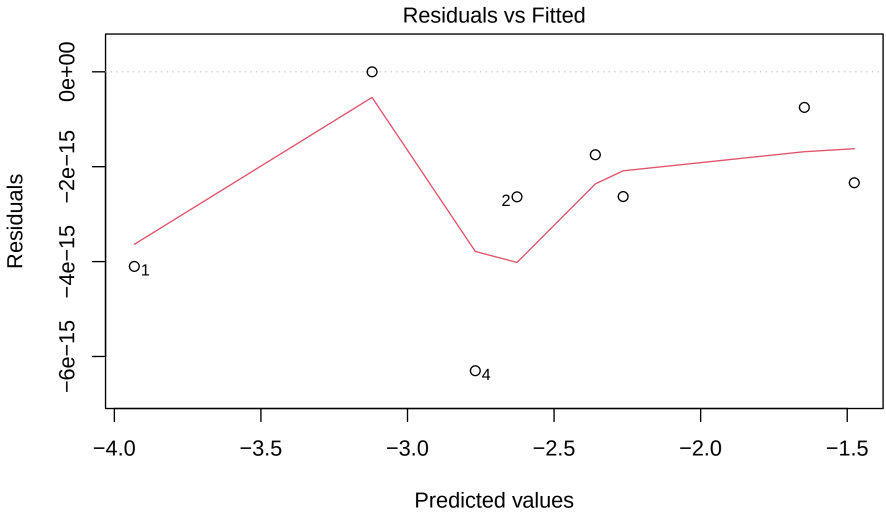


Pressure as Nominal: Model Fitting

If we treat pressure as an nominal variable, then the model fits the data perfectly.

```
##
## Call:
## glm(formula = cbind(with, without) ~ pressure, family = binomial,
##     data = Data)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0  0  0
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -3.9318     0.5830  -6.744 1.54e-11 ***
## pressure>186       2.4559     0.7025   3.496 0.000472 ***
## pressure117-126    1.3055     0.6348   2.057 0.039731 *
## pressure127-136    0.8109     0.6534   1.241 0.214543
## pressure137-146    1.1632     0.6374   1.825 0.068030 .
## pressure147-156    1.5725     0.6566   2.395 0.016615 *
## pressure157-166    1.6675     0.6913   2.412 0.015858 *
## pressure167-186    2.2856     0.6438   3.550 0.000385 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3.0023e+01  on 7  degrees of freedom
## Residual deviance: 3.2196e-14  on 0  degrees of freedom
## AIC: 48.701
##
```

Pressure as Nominal: Zero Residuals



Multiway Table

- The model

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + \beta_i$$

  can be used for a two-way table of size $I \times 2$.

- If we have a three-way table of size $I \times 2 \times K$, then we can consider the model

$$\begin{aligned} \log\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right) &= \alpha + \beta_i^X + \beta_k^Z, \\ \text{or} \quad \log\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right) &= \alpha + \beta_i^X + \beta_k^Z + \beta_{ik}^{XZ}, \end{aligned}$$

  where π_{ik} is the success probability when $X = i$ and $Z = k$.

Homogeneous Association and Conditional Independence

Consider the model

$$\log\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right) = \alpha + \beta_i^X + \beta_k^Z,$$

for a $2 \times 2 \times K$ model. At a fixed level of $Z = k$, the log odds ratio is

$$\left(\alpha + \beta_1^X + \beta_k^Z\right) - \left(\alpha + \beta_0^X + \beta_k^Z\right) = \beta_1^X - \beta_0^X = \beta_1^X,$$

if the identification restriction is $\beta_0^X = 0$.

- The conditional odds ratio is $\exp\left(\beta_1^X\right)$ for any $Z = k$, which means that the $2 \times 2 \times K$ table has homogeneous XY association.
- If we further have $\beta_1^X = 0$, then the conditional odds ratio is 1 and $X \perp Y \mid Z$ (conditional independence).

Logit Model to Test Conditional Independence

In a $2 \times 2 \times K$ table, the logistic model becomes

$$\text{logit}\left(\pi_{ik}\right) \quad = \quad \alpha + \beta x_i + \beta_k^Z, \quad k = 1, ..., K,$$

where $x_i = 0$ or 1. Testing conditional independence $X \perp Y \mid Z$ is equivalent to testing $H_0 : \beta = 0$ in the model.

1. If we assume homogeneous XY association, then
   - the Wald test statistic is $\hat{\beta}/\text{SE}$.
   - the likelihood ratio test compares the deviances between the model with $\beta = 0$ and the model with estimated β.
2. More generally, we can compare the model

$$\text{logit}\left(\pi_{ik}\right) \quad = \quad \alpha + \beta_k^Z, \quad k = 1, ..., K.$$

and the saturated model using the deviance as a goodness-of-fit test of the model.

Test Conditional Independence: Example

A clinical trial

		Response	
Study	Treatment	Success	Failure
1	Drug	11	25
	Placebo	10	27
2	Drug	16	4
	Placebo	22	10
3	Drug	14	5
	Placebo	7	12
4	Drug	2	14
	Placebo	1	16

Cochran-Mantel-Haenszel Test

The Cochran-Mantel-Haenszel test is a non-model-based test of conditional independence in a $2 \times 2 \times K$ table.

- When $K = 1$, regardless of sampling, under the independence assumption, conditioning on both sets of marginal totals, the only free cell is n_{11} that follows the hypergeometric distribution

$$P\left(n_{11} = t\right) = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1} - t}}{\binom{n_{++}}{n_{+1}}}.$$

  (Fisher's exact test).
- The mean and variance of the hypogeometric distribution are

$$\begin{aligned} \mathbb{E}\left(n_{11}\right) &= \frac{n_{1+}n_{+1}}{n_{++}}, \\ \operatorname{var}\left(n_{11}\right) &= \frac{n_{1+}n_{2+}n_{+1}n_{+2}}{n_{++}^2\left(n_{++} - 1\right)}. \end{aligned}$$

Partial Table

When $K > 1$, in each partial table k, we conditional on the row margins and column margins. When the conditional independence assumption holds, then n_{11k} follows a hypergeometric distribution with

$$\begin{aligned} \mu_{11k} = \mathbb{E}\left(n_{11k}\right) &= \frac{n_{1+k}n_{+1k}}{n_{++k}}, \\ \operatorname{var}\left(n_{11k}\right) &= \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_{++k}^2\left(n_{++k}-1\right)}. \end{aligned}$$

The Cochran-Mantel-Haenszel test statistic is

$$\text{CMH} = \frac{\left[\sum_k\left(n_{11k}-\mu_{11k}\right)\right]^2}{\sum_k \operatorname{var}\left(n_{11k}\right)},$$

which has a large-sample chi-squared null distribution with degree of freedom 1.

Common Odds Ratio

In the logit model

$$\operatorname{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z, \quad k = 1, ..., K,$$

the conditional odds ratio is $\exp(\beta)$ for any $Z = k$ (homogeneous association). The ML estimate of the common odds ratio is $\exp\left(\hat{\beta}\right)$, where $\hat{\beta}$ is the MLE of β.

The Mantel-Haenszel estimator is

$$\hat{\theta}_{MH} = \frac{\sum_k (n_{11k} n_{22k} / n_{++k})}{\sum_k (n_{12k} n_{21k} / n_{++k})}.$$

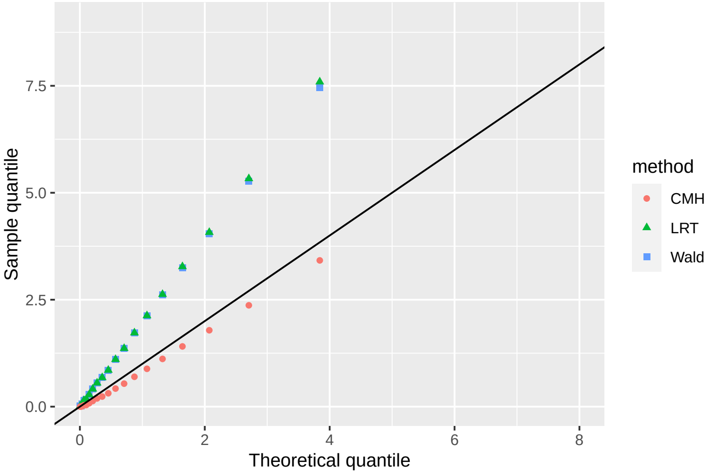
More on Cochran-Mantel-Haenszel

The CMH test can also work well when $K \to \infty$ as $n \to \infty$ (sparse table).

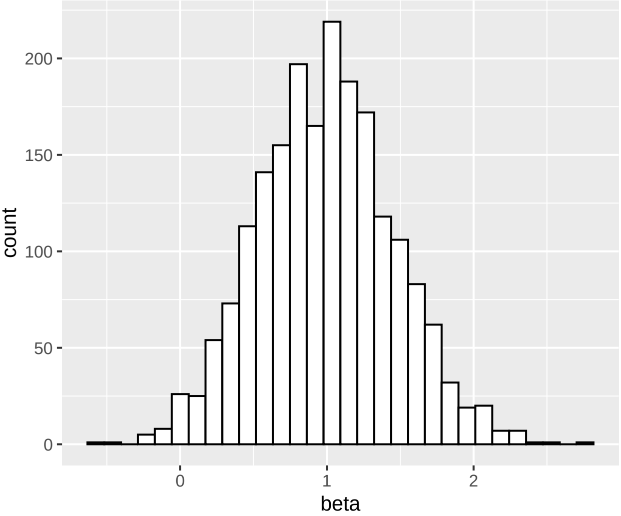
- This occurs for example for paired data: for each k, the treatment is offered only to one subject, and the control is offered only to one subject.
- In this case, $n = 2K$ and the number of observations in each partial table is 2.
- If a logistic model is fitted, the number of parameters is $1 + 1 + (K - 1)$.

The MH estimator of common odds ratio is generally preferred over the ML estimator if K is large and the tables are sparse.

What's The Problem? Simulation When $\beta = 0$



What's The Problem? Simulation of MLE When $\beta = 0.5$



Meta-Analysis

Suppose that we have K studies for the same research question. Each study yields a 2×2 table. We can combine information from all studies and refer analysis to the $2 \times 2 \times K$ table.

		Response	
Study	Treatment	Success	Failure
1	Drug	11	25
	Placebo	10	27
2	Drug	16	4
	Placebo	22	10
3	Drug	14	5
	Placebo	7	12
4	Drug	2	14
	Placebo	1	16

Conditional Association

Suppose that we have (X, Y, Z) in a $2 \times 2 \times K$ table, where Z is a control variable. Let $\{\mu_{ijk}\}$ be the cell expected frequencies corresponding to $(X = i, Y = j, Z = k)$. Then,

$$\text{conditional odds ratio: } \theta_{XY(k)} = \frac{\mu_{11k}/\mu_{12k}}{\mu_{21k}/\mu_{22k}}, \text{ fixing } Z = k,$$

$$\text{marginal odds ratio: } \theta_{XY} = \frac{\mu_{11+}/\mu_{12+}}{\mu_{21+}/\mu_{22+}},$$

are generally not the same, where $\mu_{ij+} = \sum_k \mu_{ijk}$.

However, they will be the same if

1. either Z and X are conditionally independent,
2. or Z and Y are conditionally independent.

These conditions are called the collapsibility conditions.

Back to Logit Models

Consider a $2 \times 2 \times K$ table. The logit model

$$\text{logit}\left(\pi_{ik}\right) = \alpha + \beta x_i + \beta_k^Z, \quad k = 1, ..., K,$$

has the same treatment effect β for each $Z = k$.

- The XY conditional odds ratio is $\exp\left(\beta\right)$.
- The marginal odds ratio θ_{XY} can be different from $\exp\left(\beta\right)$, since we do not have the collapsibility conditions.

Consider a $2 \times 2 \times K$ table. The logit model

$$\text{logit}\left(\pi_{ik}\right) = \alpha + \beta x_i,$$

satisfies the collapsibility condition $Y \perp Z \mid X$. Hence, the XY conditional odds ratio $\exp\left(\beta\right)$ is the same as the marginal odds ratio.

Test Homogeneous Association

The logit model

$$\operatorname{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z, \quad k = 1, ..., K,$$

has homogeneous association. Hence, we can test the goodness-of-fit of the model as a tool to test homogeneous association.

		Response	
Study	Treatment	Success	Failure
1	Drug	11	25
	Placebo	10	27
2	Drug	16	4
	Placebo	22	10
3	Drug	14	5
	Placebo	7	12
4	Drug	2	14
	Placebo	1	16

Analysis of Categorical Data
Chapter 7: Alternative Modeling of Binary Response Data

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through this chapter, you should be able to

1. fit binary data models other than the logit model,
2. describe conditional maximum inference,
3. apply conditional maximum likelihood.

Canonical Link

If the response variable Y_i belongs to exponential dispersion family, its pmf/pdf is of the form

$$f\left(y_i; \theta_i, \phi_i\right) = \exp\left\{\frac{y_i\theta_i - b\left(\theta_i\right)}{\phi_i} + c\left(y_i, \phi_i\right)\right\},$$

where θ_i is the natural parameter. The link function of a GLM transforms the mean μ_i to the linear predictor $\eta_i = g\left(\mu_i\right)$.
The canonical link function transforms the mean μ_i to the natural parameter θ_i. Hence,

$$\begin{aligned} \theta_i &= g\left(\mu_i\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}, && \text{canonical link,} \\ \theta_i &\neq g\left(\mu_i\right) = \boldsymbol{x}_i^T\boldsymbol{\beta}, && \text{otherwise.} \end{aligned}$$

Logit Link As Canonical Link

Let $Z_i \sim \text{Bin}\left(n_i, \pi_i\right)$ and $Y_i = Z_i / n_i$. The pmf of Y_i is

$$\begin{aligned} P\left(Y_i = y_i\right) &= \exp\left\{\frac{y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log\left(1-\pi_i\right)}{1/n_i} + \log\begin{pmatrix} n_i \\ n_i y_i \end{pmatrix}\right\} \\ &= \exp\left\{\frac{y_i \theta_i - \log\left[1 + \exp\left(\theta_i\right)\right]}{1/n_i} + \log\begin{pmatrix} n_i \\ n_i y_i \end{pmatrix}\right\}, \end{aligned}$$

where $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$ is the natural parameter.
The canonical link satisfies

$$\theta_i \quad = \quad g\left(\pi_i\right) = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

Hence, the logit link is the canonical link.

Behind Link Function: Distribution Assumption!

- Suppose that, in an ideal world, we could observe continuous y_i^* and we could use the linear model

$$y_i^* \quad = \quad \boldsymbol{x}_i^T \boldsymbol{\beta} - \varepsilon_i.$$

- However, in reality, we only observe y_i such that

$$y_i = \begin{cases} 0, & \text{if } y_i^* < 0, \\ 1, & \text{if } y_i^* \geq 0. \end{cases}$$

- In such a case, we often assume that $Y_i \sim \text{Bernoulli}\left(\pi_i\right)$.
- Note that

$$\pi_i = P\left(Y_i = 1\right) = P\left(Y_i^* \geq 0\right) = P\left(\varepsilon_i \leq \boldsymbol{x}_i^T \boldsymbol{\beta}\right) = F_{\varepsilon_i}\left(\boldsymbol{x}_i^T \boldsymbol{\beta}\right).$$

- The link function corresponds to the distribution assumption that we put on ε_i.

Link Functions For Binary Data

- logit link (logistic model) is inverse function of logistic cdf

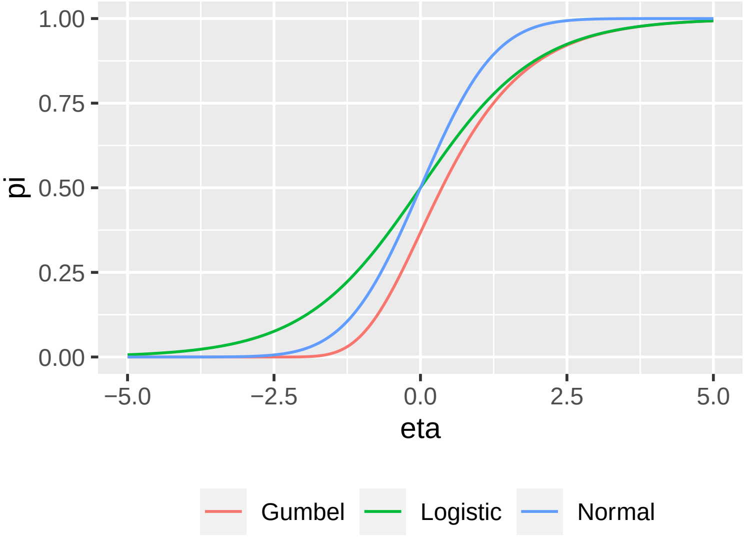
$$g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right).$$

- Probit link (probit model) is the inverse function of the standard normal cdf

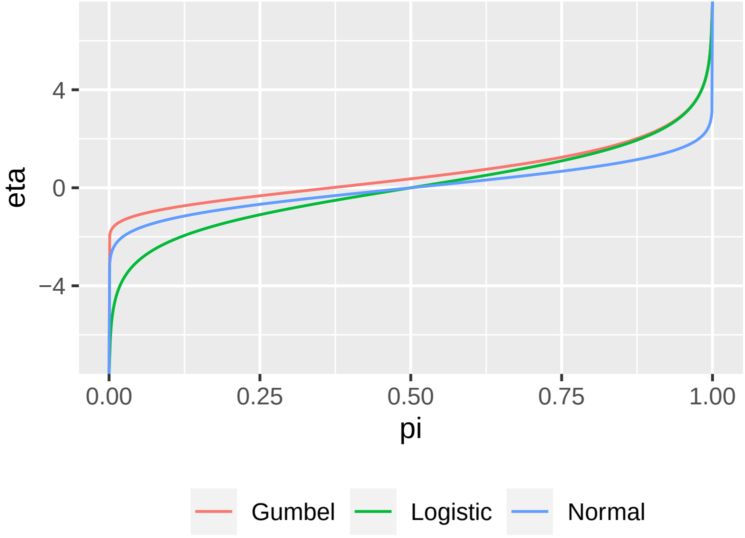
$$g(\pi_i) = \Phi^{-1}(\pi_i).$$

- Identity link (linear probability model) is the inverse function of the uniform distribution cdf $g(\pi_i) = \pi_i$.
- Log-log link is the inverse function of the Gumbel distribution cdf $g(\pi_i) = -\log[-\log(\pi_i)]$.
    - Complementary log-log link: $g(\pi_i) = \log[-\log(1-\pi_i)]$.

Different Distribution Functions



Different Link Functions



Symmetric Link Functions

A link function is symmetric about 0.5 if

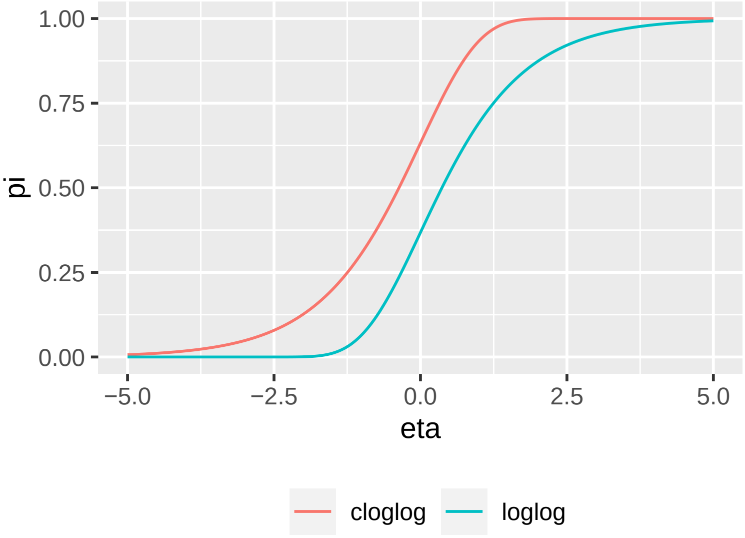
$$g(\pi) = -g(1-\pi).$$

It means that the response curve when π approaches 0 has a similar appearance to the response curve when π approaches 1.

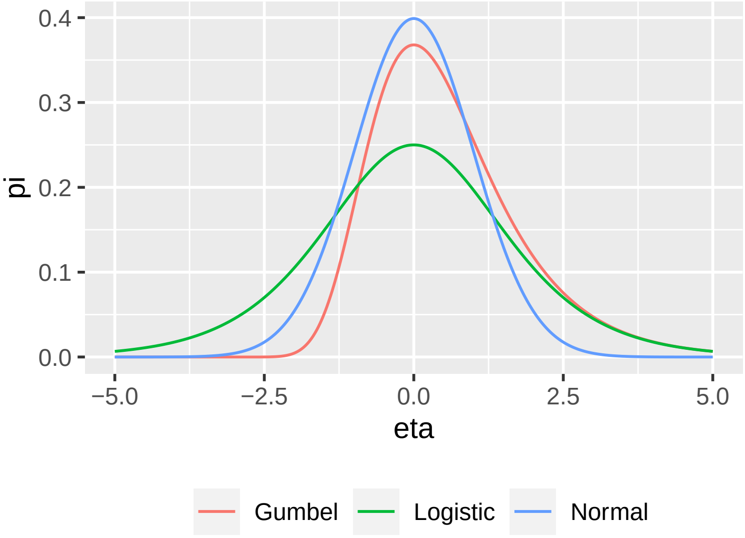
The log-log link and the clog-log link are not symmetric.

- For the clog-log link, π approaches 0 slowly, but approaches 1 quickly.
- For the log-log link, π approaches 1 slowly, but approaches 0 quickly.

Different Link Functions



Different Density Functions



Example: Different Link Functions

Number of beetles killed after exposure to gaseous carbon disulfide

	Response	
Dose	Killed	Not Killed
1.6907	6	53
1.7242	13	47
1.7552	18	44
1.7842	28	28
1.8113	52	11
1.8369	53	6
1.8610	61	1
1.8839	60	0

Likelihood

The model

$$\operatorname{logit} P\left(Y_i = 1\right) = \alpha + \sum_{j=1}^{p} \beta_j x_{ij}$$

is equivalent to

$$P\left(Y_i = y_i\right) = \frac{\exp\left[y_i\left(\alpha + \sum_{j=1}^{p} \beta_j x_{ij}\right)\right]}{1 + \exp\left(\alpha + \sum_{j=1}^{p} \beta_j x_{ij}\right)}.$$

For N independent observations, the likelihood is

$$P\left(Y_1 = y_1, \cdots, Y_n = y_n\right) = \frac{\exp\left[\left(\sum_{i=1}^{N} y_i\right)\alpha + \sum_{j=1}^{p}\left(\sum_{i=1}^{N} y_i x_{ij}\right)\beta_j\right]}{\prod_{i=1}^{N}\left[1 + \exp\left(\alpha + \sum_{j=1}^{p} \beta_j x_{ij}\right)\right]}.$$

Sufficient Statistics

By the factorization theorem of sufficient statistics,

$$P\left(Y_1 = y_1, \cdots, Y_n = y_n\right) = \frac{\exp\left[\left(\sum_{i=1}^N y_i\right)\alpha + \sum_{j=1}^p \left(\sum_{i=1}^N y_i x_{ij}\right)\beta_j\right]}{\prod_{i=1}^N \left[1 + \exp\left(\alpha + \sum_{j=1}^p \beta_j x_{ij}\right)\right]}$$

implies that $\left(\sum_{i=1}^N y_i, \sum_{i=1}^N y_i x_{i1}, \cdots, \sum_{i=1}^N y_i x_{ip}\right)$ is a sufficient statistic for $(\alpha, \beta_1, \cdots, \beta_p)$. In fact,

$$\frac{P\left(Y_1 = y_1, \cdots, Y_n = y_n\right)}{P\left(Y_1 = y_1', \cdots, Y_n = y_n'\right)} = \frac{\exp\left[\left(\sum_{i=1}^N y_i\right)\alpha + \sum_{j=1}^p \left(\sum_{i=1}^N y_i x_{ij}\right)\beta_j\right]}{\exp\left[\left(\sum_{i=1}^N y_i'\right)\alpha + \sum_{j=1}^p \left(\sum_{i=1}^N y_i' x_{ij}\right)\beta_j\right]}$$

does not depend on the parameters if and only if $\sum_{i=1}^N y_i = \sum_{i=1}^N y_i'$ and $\sum_{i=1}^N y_i x_{ij} = \sum_{i=1}^N y_i' x_{ij}$ for all j. Hence, they are also minimal sufficient.

Nuisance Parameter

Suppose that β_1 is the focus parameter and all others are nuisance parameters. Let

$$S = \left\{ (y_1^*, \cdots, y_N^*) : \sum_{i=1}^{N} y_i^* = t_0, \sum_{i=1}^{N} y_i^* x_{ij} = t_j, j = 2, .., p \right\}.$$

Then,

$$\begin{aligned} & P\left(Y_1 = y_1, \cdots, Y_n = y_n \mid \sum_{i=1}^{N} y_i = t_0, \sum_{i=1}^{N} y_i x_{ij} = t_j, j = 2, .., p \right) \\ = & \frac{P\left(Y_1 = y_1, \cdots, Y_n = y_n, \sum_{i=1}^{N} y_i = t_0, \sum_{i=1}^{N} y_i x_{ij} = t_j, j = 2, .., p \right)}{P\left(\sum_{i=1}^{N} y_i = t_0, \sum_{i=1}^{N} y_i x_{ij} = t_j, j = 2, .., p \right)} \\ = & \frac{\exp\left[\left(\sum_{i=1}^{N} y_i x_{i1} \right) \beta_1 \right]}{\sum_S \exp\left[\left(\sum_{i=1}^{N} y_i^* x_{i1} \right) \beta_1 \right]}, \end{aligned}$$

which depends only on β_1.

Conditional ML Estimator

The conditional ML estimator of β_1 maximizes the conditional likelihood

$$P\left(Y_1 = y_1, \cdots, Y_n = y_n \mid \sum_{i=1}^{N} y_i = t_0, \sum_{i=1}^{N} y_i x_{ij} = t_j, j = 2, .., p\right).$$

When the sample size n is not large enough and the number of nuisance parameters is large, ML for logistic regression may not perform so well. The conditional maximum likelihood tends to perform better.

Conditional Inference for 2×2 Tables

Consider the 2×2 table with independent binomial sampling

	Y		
X	1	0	Total
1	t	$n_1 - t$	n_1
0	s	$n_2 - s$	n_2

Suppose that

$$\operatorname{logit} P\left(Y_i = 1\right) = \alpha + \beta x_i,$$

where $x_1 = 1$ and $x_2 = 0$. To eliminate α, we conditional on its sufficient statistic $\sum_{i=1}^{N} y_i = s + t$. Then,

$$P\left(t \mid t + s, n_1, n_2\right) = \frac{\binom{n_1}{t}\binom{n_2}{s}\exp\left\{\beta t\right\}}{\sum_u \binom{n_1}{u}\binom{n_2}{s+t-u}\exp\left\{\beta u\right\}}.$$

If $\beta = 0$, we obtain the Fisher's exact test for 2×2 tables.

Conditional Inference for $2 \times 2 \times K$ Tables

In a $2 \times 2 \times K$ table, consider the logistic model

$$\text{logit}\pi_{ik} = \alpha + \beta x_i + \beta_k^Z,$$

where $x_1 = 1$ and $x_2 = 0$. Our focus parameter is often β. The sufficient statistics for $\left\{\beta_k^Z\right\}$ are $\{n_{+jk}\}$.

- When we treat n_{i+k} as fixed at each XZ combination in binomial sampling, small sample inference about β conditions on the row and column totals in each stratum.
- Conditional on the strata margins, an exact test uses $T = \sum_k n_{11k}$. The Cochran-Mantel-Haenszel test statistic

$$\text{CMH} = \frac{\left[\sum_k (n_{11k} - \mu_{11k})\right]^2}{\sum_k \text{var}(n_{11k})}$$

is based on $\sum_k n_{11k}$.

Sufficient Statistic For Sparse Tables

The idea of conditional ML is welcome especially when the contingency tables are sparse. Consider a $2 \times 2 \times K$ table and the model

$$\operatorname{logit} P\left(Y_{ik}=1\right) \quad = \quad \alpha_k+\beta x_i, \quad i=1,2,\ k=1,...,K,$$

where x_i is 0 or 1, and k means partial table k.

- In the extreme case where the row sums in each partial table are $(1,1)$, the joint likelihood is

$$\begin{aligned} & \prod_{k=1}^{K} P\left(Y_{1k}=y_{1k}, Y_{2k}=y_{2k}\right) \\ = \quad & \prod_{k=1}^{K}\left\{\frac{\exp\left[y_{1k}\left(\alpha_k+\beta\right)\right]}{1+\exp\left(\alpha_k+\beta x_i\right)} \times \frac{\exp\left[y_{2k}\alpha_k\right]}{1+\exp\left(\alpha_k\right)}\right\} \\ = \quad & \frac{\exp\left(\sum_k y_{1k}\beta\right)\exp\left[\sum_k\left(y_{1k}+y_{2k}\right)\alpha_k\right]}{\prod_{k=1}^{K}\left[1+\exp\left(\alpha_k+\beta\right)\right]\prod_{k=1}^{K}\left[1+\exp\left(\alpha_k\right)\right]}. \end{aligned}$$

- The sufficient statistics for $\{\alpha_k\}$ are $\{y_{1k}+y_{2k}\}$.

Conditional ML For Sparse Tables

Suppose that partial tables are independent of each other. Then

$$\begin{aligned} & P\left(Y_{11}=y_{11}, Y_{21}=y_{21} \cdots, Y_{2k}=y_{2K} \mid y_{1k}+y_{2k}=t_k, k=1,..,K\right) \\ = & \prod_{k=1}^{K} P\left(Y_{1k}=y_{1k}, Y_{2k}=y_{2k} \mid y_{1k}+y_{2k}=t_k\right) \\ = & \frac{\exp\left[\sum_{k=1}^{K} I\left(t_k=1, y_{1k}=1\right)\beta\right]}{\left[1+\exp\left(\beta\right)\right]^{\sum_{k=1}^{K} I\left(t_k=1\right)}}, \end{aligned}$$

which depends only on β. Its maximizer is the conditional MLE of β.

Even though $K \to \infty$, the number of parameters in the conditional likelihood is still 1. In contrast, the number of parameters in the likelihood is $K+1$.

Analysis of Categorical Data
Chapter 8: Multinomial Responses

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through this chapter, you should be able to

1. fit multinomial models for nominal responses,
2. fit multinomial models for ordinal responses,
3. test conditional independence.

Baseline Category Logit Models

Let Y be a categorical response with J nominal categories. Let $\pi_j\left(\boldsymbol{x}\right)=P\left(Y=j\mid\boldsymbol{x}\right)$ with $\sum_j\pi_j\left(\boldsymbol{x}\right)=1$. We treat Y as

$$Y \quad \sim \quad \text{Multinomial}\left(\pi_1\left(\boldsymbol{x}\right),...,\pi_J\left(\boldsymbol{x}\right)\right).$$

Let C be the baseline category, then the baseline-category logit model is

$$\log\left(\frac{\pi_j\left(\boldsymbol{x}\right)}{\pi_C\left(\boldsymbol{x}\right)}\right) \quad = \quad \alpha_j+\boldsymbol{\beta}_j^T\boldsymbol{x}, \quad j\neq C.$$

Response Probabilities

The baseline-category logit model implies that

$$\pi_j\left(\boldsymbol{x}\right) \;\;=\;\; \pi_C\left(\boldsymbol{x}\right)\exp\left\{\alpha_j + \boldsymbol{\beta}_j^T\boldsymbol{x}\right\}.$$

Hence, for $j \neq C$,

$$\pi_j\left(\boldsymbol{x}\right) \;\;=\;\; \frac{\exp\left\{\alpha_j + \boldsymbol{\beta}_j^T\boldsymbol{x}\right\}}{1 + \sum_{j\neq C}\exp\left\{\alpha_j + \boldsymbol{\beta}_j^T\boldsymbol{x}\right\}},$$

which is the softmax function

$$\frac{\exp\left(z_j\right)}{\sum_j \exp\left(z_j\right)},$$

with $z_j = \alpha_j + \boldsymbol{\beta}_j^T\boldsymbol{x}$. This in fact means that $\alpha_C = 0$ and $\boldsymbol{\beta}_C = \mathbf{0}$, which leads to the 1.

Maximum Likelihood

The baseline-category logit model is fitted by ML. The log-likelihood function for subject i is

$$\sum_{j=1}^{J} y_{ij} \log \pi_j \left(\boldsymbol{x}_i \right) = \sum_{j=1}^{J} y_{ij} \log \left[\frac{\exp \left\{ \alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x} \right\}}{1 + \sum_{j \neq C} \exp \left\{ \alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x} \right\}} \right],$$

where $\alpha_C = 0$ and $\boldsymbol{\beta}_C = \mathbf{0}$. $\left\{ \hat{\alpha}_j, \hat{\boldsymbol{\beta}}_j,\ j \neq C \right\}$ are obtained by numerical methods (e.g., Newton-Raphson method).

The choice of C will influence the parameter values, but not the response probabilities.

Latent Representation

Let U_j denote the latent utility of response outcome j. Suppose that

$$U_j = -\alpha_j - \boldsymbol{\beta}_j^T \boldsymbol{x} + e_j.$$

The response outcome is the value of j having maximum utility.

- Suppose that e_j are independent and have the Gumbel distribution $F(e) = \exp\{-\exp(-e)\}$. Then,

$$\pi_j(\boldsymbol{x}) = \frac{\exp\left\{\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x}\right\}}{1 + \sum_{j \neq C} \exp\left\{\alpha_j + \boldsymbol{\beta}_j^T \boldsymbol{x}\right\}}.$$

- Other distribution assumptions can be put on e_j. For example, e_j is independent $N(0,1)$.
- More generally, we can allow $\{e_j\}$ to be correlated.

Cumulative Logits For Ordinal Response

Let Y be a categorical response with J ordinal categories and cell probabilities $\{\pi_j(\boldsymbol{x})\}$. Then,

$$P(Y \leq j \mid \boldsymbol{x}) = \pi_1(\boldsymbol{x}) + \cdots + \pi_j(\boldsymbol{x}).$$

The cumulative logits are defined as

$$\begin{aligned} \operatorname{logit} P(Y \leq j \mid \boldsymbol{x}) &= \log\left(\frac{P(Y \leq j \mid \boldsymbol{x})}{1 - P(Y \leq j \mid \boldsymbol{x})}\right) \\ &= \log\left(\frac{\pi_1(\boldsymbol{x}) + \cdots + \pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x}) + \cdots + \pi_J(\boldsymbol{x})}\right), \quad j = 1, ..., J-1. \end{aligned}$$

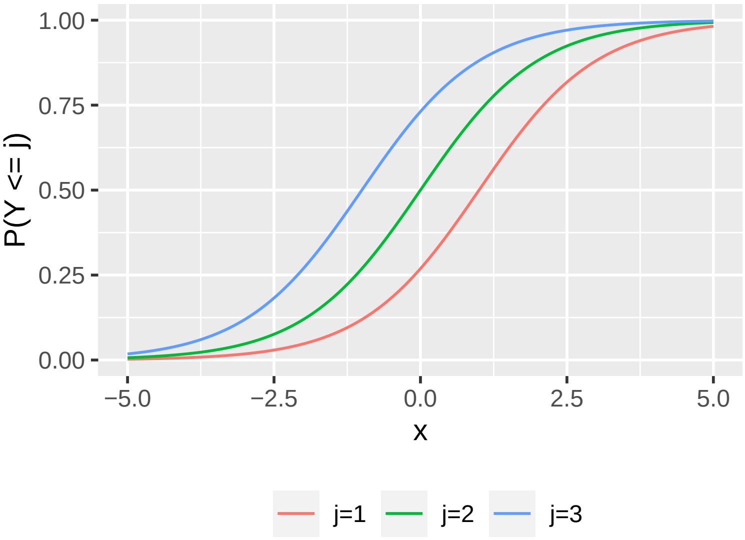
Cumulative Logit Model

The cumulative logit model is

$$\log\left(\frac{P\left(Y \leq j \mid \boldsymbol{x}\right)}{1 - P\left(Y \leq j \mid \boldsymbol{x}\right)}\right) \;=\; \alpha_j + \boldsymbol{x}^T\boldsymbol{\beta},\ j = 1, ..., J-1.$$

- Each model has its own intercept but share the same slopes. For each j, the model is an ordinary logistic model for a binary response.
- In this model α_j must be increasing in α_j, because $P\left(Y \leq j \mid \boldsymbol{x}\right)$ in increasing in j and $\log\left(\frac{P(Y\leq j|\boldsymbol{x})}{1-P(Y\leq j|\boldsymbol{x})}\right)$ is increasing in $P\left(Y \leq j \mid \boldsymbol{x}\right)$.

Increasing in α_j



Proportional Odds

The cumulative odds ratio is

$$\frac{P\left(Y\leq j\mid\boldsymbol{x}_1\right)/P\left(Y>j\mid\boldsymbol{x}_1\right)}{P\left(Y\leq j\mid\boldsymbol{x}_2\right)/P\left(Y>j\mid\boldsymbol{x}_2\right)}.$$

The cumulative logit model satisfies

$$\begin{aligned}
&\log\frac{P\left(Y\leq j\mid\boldsymbol{x}_1\right)/P\left(Y>j\mid\boldsymbol{x}_1\right)}{P\left(Y\leq j\mid\boldsymbol{x}_2\right)/P\left(Y>j\mid\boldsymbol{x}_2\right)}\\
=\ &\log\left(\frac{P\left(Y\leq j\mid\boldsymbol{x}_1\right)}{1-P\left(Y\leq j\mid\boldsymbol{x}_1\right)}\right)-\log\left(\frac{P\left(Y\leq j\mid\boldsymbol{x}_2\right)}{1-P\left(Y\leq j\mid\boldsymbol{x}_2\right)}\right)\\
=\ &\boldsymbol{\beta}^T\left(\boldsymbol{x}_1-\boldsymbol{x}_2\right).
\end{aligned}$$

- The odds of making response $Y\leq j$ at $\boldsymbol{x}_1$ is proportional to to the odds at $\boldsymbol{x}_2$. Hence, the model is also called the proportional odds model.
- Estimation is still ML with iterative numerical methods (e.g., based on the multinomial likelihood).

Latent Variable Motivation

Let Y^* denote the underlying continuous latent variable such that

$$Y^* \;\; = \;\; \boldsymbol{\beta}^T \boldsymbol{x} + e.$$

The thresholds $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_{J-1} < \alpha_J = \infty$ are cutpoints of the continuous scale. The observed response y satisfies

$$y = j \quad \text{if} \quad \alpha_{j-1} < y^* \leq \alpha_j.$$

Then,

$$\begin{aligned} P\left(Y \leq j \mid \boldsymbol{x}\right) \;&=\; P\left(Y^* \leq \alpha_j \mid \boldsymbol{x}\right) = P\left(\boldsymbol{\beta}^T \boldsymbol{x} + e \leq \alpha_j \mid \boldsymbol{x}\right) \\ &=\; P\left(e \leq \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x} \mid \boldsymbol{x}\right) = F\left(\alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}\right), \end{aligned}$$

where F is the conditional distribution of e given $\boldsymbol{x}$. Hence,

$$F^{-1}\left\{P\left(Y \leq j \mid \boldsymbol{x}\right)\right\} \;\; = \;\; \alpha_j - \boldsymbol{\beta}^T \boldsymbol{x}.$$

If we have logistic distribution, then we have a logit model. In general, we can use other distributions as well.

Cumulative Cloglog Model

Suppose that F is the cdf of the Gumbel distribution. Then, the cloglog link yields

$$\log\left\{-\log\left[1-P\left(Y\leq j\mid \boldsymbol{x}\right)\right]\right\} \quad = \quad \alpha_j-\boldsymbol{\beta}^T\boldsymbol{x}.$$

This is often called a proportional hazards model.

- The extreme value distribution is not symmetric. With this link $P\left(Y\leq j\mid \boldsymbol{x}\right)$ approaches 1 at a faster rate than it approaches 0.

The loglog link yields

$$\log\left\{-\log\left[P\left(Y\leq j\mid \boldsymbol{x}\right)\right]\right\} \quad = \quad \alpha_j-\boldsymbol{\beta}^T\boldsymbol{x}.$$

It is appropriate when the cloglog link holds for the categories listed in reverse order, i.e., approaching1 at a slower rate than approaching 0 .

Extension

We can make the slopes not the same, e.g.,

$$\log\left(\frac{P\left(Y_i \leq c\right)}{1-P\left(Y_i \leq c\right)}\right)=\begin{cases}\alpha_1+x_i\beta_1+z_i\gamma_1 & c=1\\ \alpha_2+x_i\beta_2+z_i\gamma_2 & c=2\end{cases},$$

or partially the same across groups, e.g.,

$$\log\left(\frac{P\left(Y_i \leq c\right)}{1-P\left(Y_i \leq c\right)}\right)=\begin{cases}\alpha_1+x_i\beta_1+z_i\gamma & c=1\\ \alpha_2+x_i\beta_2+z_i\gamma & c=2\end{cases}.$$

However, a problem is that the cumulative probabilities may be out of order.

Test Conditional Independence for Nominal Y

Suppose that Y is nominal, and that Z is nominal. XY conditional independence is equivalent to the baseline-category logit model

$$\log\left[\frac{P\left(Y=j \mid X=i, Z=k\right)}{P\left(Y=C \mid X=i, Z=k\right)}\right] \quad = \quad \alpha_{jk},$$

since $P\left(Y=j \mid X=i, Z=k\right)$ does not depend on X:

$$P\left(Y=j \mid X=i, Z=k\right) \quad = \quad \frac{\exp\left\{\alpha_{jk}\right\}}{1+\sum_{j\neq C}\exp\left\{\alpha_{jk}\right\}}.$$

Nominal X or Ordinal X

1. If X is nominal, an alternative to XY conditional independence is

$$\log\left[\frac{P\left(Y=j\mid X=i,Z=k\right)}{P\left(Y=C\mid X=i,Z=k\right)}\right] \;=\; \alpha_{jk}^{Z}+\beta_{ji}^{X},$$

with constraint $\beta_{jI}=0$ for each j. Then, conditional independence is to test $\beta_{j1}=\cdots=\beta_{jI}=0$ for all j. Large sample chi-squared tests have $(I-1)(J-1)$ df.

2. If X is ordinal and $\{x_i\}$ are the ordered scores, an alternative to XY conditional independence is

$$\log\left[\frac{P\left(Y=j\mid X=i,Z=k\right)}{P\left(Y=J\mid X=i,Z=k\right)}\right] \;=\; \alpha_{jk}^{Z}+\beta_{j}x_{i}.$$

Then, conditional independence is to test $\beta_j=0$ for all j. Large sample chi-squared tests have $J-1$ df.

Test Conditional Independence for Ordinal Y

Suppose that Y is ordinal with the cumulative logit models, XY conditional independence is equivalent to the model

$$\text{logit}\left[P\left(Y \leq j \mid X = i, Z = k\right)\right] = \alpha_{jk},$$

with $\alpha_{1k} < \alpha_{2k} < \cdots < \alpha_{J-1,k}$ for each k.

1. If X is nominal, an alternative to XY independence is

   $$\text{logit}\left[P\left(Y \leq j \mid X = i, Z = k\right)\right] = \alpha_{jk}^{Z} + \beta_i,$$

   where $\beta_I = 0$ for identification. The conditional independence is $\beta_i = 0$ for all i. Large sample chi-squared tests have $I - 1$ df.
2. If X is ordinal and $\{x_i\}$ are the ordered scores, an alternative to XY conditional independence is

   $$\text{logit}\left[P\left(Y \leq j \mid X = i, Z = k\right)\right] = \alpha_{jk}^{Z} + \beta x_i.$$

   Then XY conditional independence is $\beta = 0$. Large sample chi-squared tests have 1 df.

Cochran-Mantel-Haenszel Tests for $I \times J \times K$ Tables

The Cochran-Mantel-Haenszel test for $2 \times 2 \times K$ tables can be generalized to $I \times J \times K$ tables.

- Conditional on row and column totals, each stratum has $(I-1)(J-1)$ nonredundant cell counts. Let

$$\boldsymbol{n}_k = \begin{bmatrix} n_{11k} & n_{12k} & \cdots & n_{1,J-1,k} & \cdots & n_{I-1,J-1,k} \end{bmatrix}^T.$$

- If H_0 does not hold, then

$$\text{CMH} = \left(\sum_k \boldsymbol{n}_k - \sum_k \boldsymbol{\mu}_k\right)^T \left(\sum_k \boldsymbol{V}_k\right)^{-1} \left(\sum_k \boldsymbol{n}_k - \sum_k \boldsymbol{\mu}_k\right)$$

  should be large, where $\boldsymbol{\mu}_k = \mathbb{E}(\boldsymbol{n}_k)$ and $\boldsymbol{V}_k = \text{cov}(\boldsymbol{n}_k)$.
- If H_0 holds, its distribution can be approximated by a chi-square distribution with $(I-1)(J-1)$ df.

Analysis of Categorical Data
Chapter 9: Loglinear Model

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through this chapter, you should be able to

1. use loglinear models to analyze contingency tables,
2. understand different types of independence,
3. understand the connection between Poisson sampling and multinomial sampling,
4. understand the connection between loglinear model and logistic model,
5. obtain closed form expression of MLE.

Loglinear Model

Suppose that $\begin{pmatrix} N_1 & N_2 & \cdots & N_c \end{pmatrix}$ is a multinomial sample of n subjects. If π_i is the cell probability of cell i, the expected frequency is $\mu_i = n\pi_i$.

Suppose that $\begin{pmatrix} N_1 & N_2 & \cdots & N_c \end{pmatrix}$ is a Poisson sample. The count in each cell i follows a Poisson distribution with expected frequency μ_i.

The loglinear model builds a linear model on μ_i, and is applicable to both multinomial sampling and Poisson sampling. That is, the loglinear model is of the form

$$\log \mu_i \quad = \quad \lambda + \boldsymbol{\beta}^T \boldsymbol{x}_i.$$

Poisson and Multinomial Sampling

Suppose that there are c independent Poisson random variables N_i, $i = 1, 2, ..., c$, with mean μ_i. Their conditional distribution given $\sum_{i=1}^{c} N_i = n$ satisfies

$$\begin{aligned} P\left(N_1 = n_1, ..., N_c = n_c \mid \sum_{i=1}^{c} N_i = n\right) &= \frac{P\left(N_1 = n_1, ..., N_c = n_c\right)}{P\left(\sum_{i=1}^{c} N_i = n\right)} \\ &= \frac{n!}{\prod_{i=1}^{c} n_i!} \prod_{i=1}^{c} \left(\frac{\mu_i}{\sum_{j=1}^{c} \mu_j}\right)^{n_i}, \end{aligned}$$

which is a multinomial distribution with sample size n and outcome probabilities

$$\left\{\frac{\mu_i}{\sum_{j=1}^{c} \mu_j}\right\}.$$

Poisson Loglinear Model

Suppose that we have c independent Poisson random variables N_i, each with mean $\mu_i = \exp\left\{\lambda + \boldsymbol{\beta}^T \boldsymbol{x}_i\right\}$. The log-likelihood is

$$\ell = \sum_i \left[n_i \left(\lambda + \boldsymbol{\beta}^T \boldsymbol{x}_i\right) - \exp\left(\lambda + \boldsymbol{\beta}^T \boldsymbol{x}_i\right) - \log\left(n_i!\right)\right].$$

The first-order partial derivatives are

$$\begin{aligned} \frac{\partial \ell}{\partial \lambda} &= \sum_i n_i - \exp\left(\lambda\right) \sum_i \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right), \\ \frac{\partial \ell}{\partial \beta_k} &= \sum_i n_i x_{ik} - \exp\left(\lambda\right) \sum_i \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right) x_{ik}. \end{aligned}$$

Evaluated at the MLEs, we should have

$$\sum_i n_i x_{ik} = \frac{\sum_i n_i}{\sum_i \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right)} \sum_i \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right) x_{ik}.$$

Multinomial Loglinear Model

If we conditional on the sum $\sum_i n_i$, then $(N_1, ..., N_c)$ follows a multinomial distribution with cell probabilities

$$\frac{\mu_i}{\sum_j \mu_j} = \frac{\exp\left(\lambda + \boldsymbol{\beta}^T \boldsymbol{x}_i\right)}{\sum_j \exp\left(\lambda + \boldsymbol{\beta}^T \boldsymbol{x}_j\right)} = \frac{\exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right)}{\sum_j \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_j\right)}.$$

The multinomial log-likelihood is

$$\ell = \log\left[\frac{(\sum_i n_i)!}{\prod_i n_i!}\right] + \sum_i n_i \boldsymbol{\beta}^T \boldsymbol{x}_i - \left(\sum_i n_i\right) \log\left[\sum_j \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_j\right)\right].$$

The first-order partial derivative of multinomial log-likelihood is

$$\frac{\partial \ell}{\partial \beta_k} = \sum_i n_i x_{ik} - \left(\sum_i n_i\right) \frac{\sum_i \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right) x_{ik}}{\sum_i \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_i\right)}.$$

Hence, the Poisson loglinear model and the multinomial loglinear model should yield the same $\boldsymbol{\beta}$ estimators.

Multinomial loglinear model

We can rewrite the Poisson log-likelihood as

$$\begin{aligned}\ell &= \sum_i \left[n_i\left(\lambda+\boldsymbol{\beta}^T\boldsymbol{x}_i\right)-\exp\left\{\lambda+\boldsymbol{\beta}^T\boldsymbol{x}_i\right\}-\log\left(n_i!\right)\right]\\ &= \underbrace{\left\{\sum_i n_i\boldsymbol{\beta}^T\boldsymbol{x}_i - n\log\left[\sum_j \exp\left(\boldsymbol{\beta}^T\boldsymbol{x}_j\right)\right]\right\}}_{N_1,\ldots,N_c\mid\sum_i N_i=n} + \underbrace{\{n\log\mu-\mu\}}_{\sum_i N_i\sim\text{Poisson}(\mu)} + C,\end{aligned}$$

where $\mu=\sum_i \mu_i$. This means that we can reparametrize the Poisson model with parameters $(\lambda,\boldsymbol{\beta})$ to $(\mu,\boldsymbol{\beta})$.

- The multinomial loglinear model only takes the conditional multinomial distribution part.
- The Poisson loglinear model has one more parameter λ (or μ) than the multinomial logliner model, because of the random sample size.

Independent Model for Two-Way Tables

Consider an $I \times J$ contingency table with multinomial sampling of n subjects. If X and Y are independent, then

$$\pi_{ij} = \pi_{i+}\pi_{+j}.$$

The expected frequency of cell (i, j) is

$$\mu_{ij} = n\pi_{i+}\pi_{+j},$$

which is equivalent to

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y.$$

This is the loglinear model of independence.

Independence under Poisson sampling means independence under multinomial sampling when we conditional on n.

Identification

Consider the model

$$\log \mu_{ij} \quad = \quad \lambda + \lambda_i^X + \lambda_j^Y .$$

We can either let $\lambda_I^X = \lambda_J^Y = 0$ or $\sum_i \lambda_i^X = \sum_j \lambda_j^Y = 0$ for identification.

1. If $\lambda_I^X = \lambda_J^Y = 0$,

$$\lambda_i^X \quad = \quad \log \pi_{i+} - \log \pi_{I+}.$$

2. If $\sum_i \lambda_i^X = \sum_j \lambda_j^Y = 0$,

$$\lambda_i^X \quad = \quad \log \pi_{i+} - \frac{1}{I} \sum_{i=1}^{I} \log \pi_{i+}.$$

Saturated Model for Two-Way Table

If X and Y are dependent, the loglinear model becomes

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY},$$

with interaction λ_{ij}^{XY}.

- For identification, we need the constraints $\lambda_I^X = \lambda_J^Y = 0$ and $\lambda_{Ij}^{XY} = \lambda_{iJ}^{XY} = 0$ for all i and j.
- The total number of parameters in this model is

$$\underset{\lambda}{1} + \underset{\lambda_i^X}{I-1} + \underset{\lambda_j^Y}{J-1} + \underset{\lambda_{ij}^{XY}}{(I-1)(J-1)} = IJ.$$

- The number of cells is IJ. Hence, it is the saturated model.

Association in Two-Way Table

The saturated model and the independence model for a two-way table only differ in λ_{ij}^{XY}, which measures deviation from independence. In particular, the log local odds ratio satisfies

$$\begin{aligned}
\log\left(\frac{\mu_{ij}\mu_{i+1,j+1}}{\mu_{i,j+1}\mu_{i+1,j}}\right) &= \left(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}\right) \\
&\quad + \left(\lambda + \lambda_{i+1}^X + \lambda_{j+1}^Y + \lambda_{i+1,j+1}^{XY}\right) \\
&\quad - \left(\lambda + \lambda_i^X + \lambda_{j+1}^Y + \lambda_{i,j+1}^{XY}\right) \\
&\quad - \left(\lambda + \lambda_{i+1}^X + \lambda_j^Y + \lambda_{i+1,j}^{XY}\right) \\
&= \lambda_{ij}^{XY} + \lambda_{i+1,j+1}^{XY} - \lambda_{i,j+1}^{XY} - \lambda_{i+1,j}^{XY}.
\end{aligned}$$

Loglinear Model Under Different Sampling

The connection between the Poisson distribution and the multinomial distribution suggests that the loglinear models

$$\begin{aligned} \log \mu_{ij} &= \lambda + \lambda_i^X + \lambda_j^Y, \\ \log \mu_{ij} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \end{aligned}$$

can be used for both multinomial sampling and Poisson sampling.

- You can easily derive that $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$ in an independent model with both multinomial sampling and Poisson sampling.
- The saturated model implies $\hat{\mu}_{ij} = n_{ij}$.

Model Fit and Model Comparison

When we fit log-linear models, we often fit a Poisson log-linear model.

- Pearson χ^2 and deviance test whether a model holds by comparing cell fitted values to observed counts.
- If the model fits the data well, they can be approximated by a chi-square distribution with df being the number of cells minus the number of model parameters.

We can also use difference in deviance to compare different models, e.g., compare the model with interaction and the model without interaction. If models are not nested, we can use information criterion.

- Even though we have two models under multinomial sampling, we can still fit Poisson log-linear models and perform model selection via LRT or AIC.

Types of Independence

Consider a three-way $I \times J \times K$ table of variables X, Y, and Z. We can define different types of independence.

1. Mutual independence: the variables are mutually independent, denoted by (X, Y, Z) (generating class).
2. Joint independence: X and Z are jointly independent of Y, denoted by (XZ, Y).
3. Conditional independence: X is independent of Y given Z, denoted by (XZ, YZ) or $X \perp Y \mid Z$.
4. Marginal independence: X and Y are independent when ignoring Z, denoted by (X, Y).
5. Homogeneous association: the conditional odds ratio satisfies $\theta_{ij(1)} = \theta_{ij(2)} = \cdots = \theta_{ij(K)}$.

Marginality/Hierarchical Model

When you include higher-order terms into the model, always include all lower-order terms by the hierarchy principle. A loglinear model that follows this principle is a hierarchical loglinear model.

$$\begin{aligned}
\log \mu_{ij} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \\
\log \mu_{ijk} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \\
\log \mu_{ijk} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \\
\log \mu_{ijk} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}, \\
\log \mu_{ijk} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}.
\end{aligned}$$

Marginal Independence

X is marginally independent of Y, if

$$\pi_{ij+} = \pi_{i++}\pi_{+j+}, \text{ for all } i \text{ and } j.$$

This is denoted by (X, Y) or $X \perp Y$.

Conditional Independence

Categorical variables X and Y are conditionally independent given Z, if independence holds for each partial table within which Z is fixed;

$$\pi_{ij|k} \quad = \quad \pi_{i+|k}\pi_{+j|k}, \text{ for all } i,\ j, \text{ and } k,$$

where $\pi_{ij|k} = P\left(X = i, Y = j \mid Z = k\right)$. This is denoted by (XZ, YZ) or $X \perp Y \mid Z$.

If $X \perp Y \mid Z$, then

$$\begin{aligned} \pi_{ijk} &= P\left(X = i \mid Y = j, Z = k\right) P\left(Y = j, Z = k\right) \\ &= P\left(X = i \mid Z = k\right) P\left(Y = j, Z = k\right) \\ &= \pi_{i+k}\pi_{+jk}/\pi_{++k}. \end{aligned}$$

The corresponding loglinear model is

$$\log \mu_{ijk} \quad = \quad \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

Joint Independence

Y is jointly independent of X and Z, if

$$\pi_{ijk} = \pi_{i+k}\pi_{+j+}, \text{ for all } i,\ j, \text{ and } k.$$

This is denoted by (XZ, Y) or $(X, Z) \perp Y$. The corresponding loglinear model is

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}.$$

- (XZ, Y) means that (X, Y) and (Z, Y).
- Conditional independence is weaker than joint independence, since conditional independence with $\lambda_{jk}^{YZ} = 0$ leads to joint independence.

Mutual independence

X, Y, and Z are mutually independent, if

$$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}, \text{ for all } i, j, \text{ and } k.$$

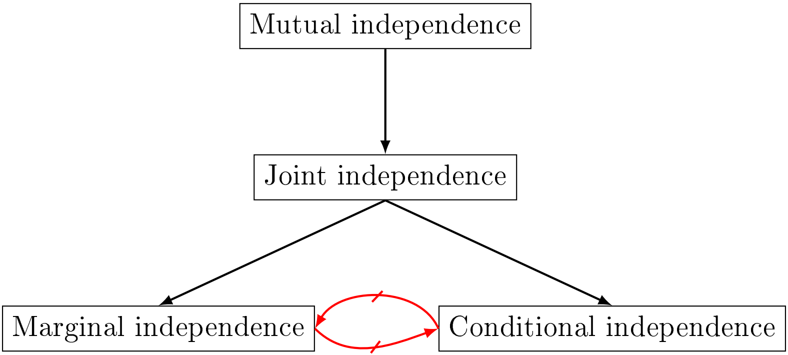
This is denoted by (X, Y, Z). The corresponding loglinear model is

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z.$$

- Joint independence is weaker than mutual independence, since joint independence with $\lambda_{jk}^{YZ} = 0$ leads to mutual independence.
- From probability theory perspective,

$$\pi_{ijk} = \underbrace{\pi_{i++}\pi_{++k}}_{\pi_{i+k}}\pi_{+j+}.$$

Relationships in Graph



Mutual independence

Joint independence

Marginal independence

Conditional independence

More General Models

Consider the models

$$
\begin{aligned}
(X,Y,Z):&\ \log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \\
(XZ,Y):&\ \log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} \\
(XZ,YZ):&\ \log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.
\end{aligned}
$$

They have three, two, and one pair of conditionally independent variables, respectively.

The saturated model is

$$(XYZ):\ \log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}.$$

Interpretation By Conditional Association

At a fixed level $Z = k$, the local odds ratios are

$$\theta_{ij(k)} = \frac{\pi_{ijk}\pi_{i+1,j+1,k}}{\pi_{i+1,j,k}\pi_{i,j+1,k}}, \quad 1 \leq i \leq I-1,\ 1 \leq j \leq J-1.$$

Suppose that $X \perp Y \mid Z$, where $\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k}$. Then,

$$\theta_{ij(k)} = \frac{\left(\pi_{i+k}\pi_{+jk}/\pi_{++k}\right)\left(\pi_{i+1,+,k}\pi_{+,j+1,k}/\pi_{++k}\right)}{\left(\pi_{i+1,+,k}\pi_{+jk}/\pi_{++k}\right)\left(\pi_{i+k}\pi_{+,j+1,k}/\pi_{++k}\right)} = 1.$$

Hence, $X \perp Y \mid Z$ is equivalent to $\theta_{ij(k)} = 1$ for all possible i, j, and k.

Interpretation By Conditional Association

Consider the model

$$(XY, XZ, YZ):\ \log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

At a fixed level $Z = k$, the local odds ratios satisfies

$$\begin{aligned} \log \theta_{ij(k)} &= \log\left(\frac{\mu_{ijk}\mu_{i+1,j+1,k}}{\mu_{i+1,j,k}\mu_{i,j+1,k}}\right) \\ &= \lambda_{ij}^{XY} + \lambda_{i+1,j+1}^{XY} - \lambda_{i+1,j}^{XY} - \lambda_{i,j+1}^{XY}, \end{aligned}$$

which implies homogeneous XY association

$$\theta_{ij(1)} = \theta_{ij(2)} = \cdots = \theta_{ij(K)},\ \text{for all } i \text{ and } j.$$

By symmetry, we also have homogeneous XZ association and homogeneous YZ association.

Inference of Conditional Association

The local odds ratios satisfies

$$\log \theta_{ij(k)} \;=\; \lambda_{ij}^{XY} + \lambda_{i+1,j+1}^{XY} - \lambda_{i+1,j}^{XY} - \lambda_{i,j+1}^{XY}.$$

When a Poisson log-linear model

$$\log \mu \;=\; \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$$

is fitted, we know that

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \;\approx\; N\left(\mathbf{0},\; \left(\boldsymbol{X}^T \boldsymbol{W}^{-1} \boldsymbol{X}\right)^{-1}\right).$$

Hence, for any constant vector $\boldsymbol{a}$,

$$\boldsymbol{a}^T \hat{\boldsymbol{\beta}} - \boldsymbol{a}^T \boldsymbol{\beta} \;\approx\; N\left(\mathbf{0},\; \boldsymbol{a}^T \left(\boldsymbol{X}^T \boldsymbol{W}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{a}\right).$$

Log-Likelihoood Under Poisson Sampling

Consider the log-linear model with Poisson sampling

$$\log \mu_i \quad = \quad \lambda + \boldsymbol{\beta}^T \boldsymbol{x}_i, \quad i = 1, ..., N.$$

The log-likelihood is

$$\ell \quad = \quad \sum_{i=1}^{N} \left[n_i \left(\lambda + \boldsymbol{\beta}^T \boldsymbol{x}_i \right) - \exp\left(\lambda + \boldsymbol{\beta}^T \boldsymbol{x}_i \right) - \log\left(n_i! \right) \right].$$

- The sufficient statistic of β_j is $\sum_i n_i x_{ij}$.
- The first-order partial derivatives are

$$\frac{\partial \ell}{\partial \beta_j} \quad = \quad \sum_{i=1}^{N} \left(n_i - \mu_i \right) x_{ij}.$$

Hence, the MLE must satisfy $\sum_{i=1}^{n} n_i x_{ij} = \sum_{i=1}^{n} \mu_i x_{ij}$.

Sufficient Statistic and Marginal Sum

For a contingency table, N is the number of independence cells, $\boldsymbol{\beta}$ is the vector of effects $\left\{\lambda_i^X, \lambda_j^Y, \cdots\right\}$, and x_{ij} is either 0 or 1.

- The sufficient statistics $\{\sum_i n_i x_{ij}\}$ become some marginal sums.
- $\sum_{i=1}^N n_i x_{ij} = \sum_{i=1}^N \mu_i x_{ij}$ means that the sufficient statistics are the same as the the expected value μ in the marginal tables.

Since minimal sufficient statistic can be expressed as a function of any other sufficient statistics, we can obtain the explicit expression of μ using the minimal sufficient statistics.

Three-Way Table Under Poisson Sampling

Under Poisson sampling, the log-likelihood of a three-way table satisfies

$$\ell \quad = \quad \text{constant} + \sum_{i,j,k} \left[n_{ijk} \log\left(\mu_{ijk}\right) - \mu_{ijk} \right].$$

For the saturated model,

$$\begin{aligned} \ell \quad = \quad & \text{constant} + n\lambda + \sum_i n_{i++}\lambda_i^X + \sum_j n_{+j+}\lambda_j^Y + \sum_k n_{++k}\lambda_k^Z \\ & + \sum_{i,j} n_{ij+}\lambda_{ij}^{XY} + \sum_{i,k} n_{i+k}\lambda_{ik}^{XZ} + \sum_{j,k} n_{+jk}\lambda_{jk}^{YZ} + \sum_{i,j,k} n_{ijk}\lambda_{ijk}^{XYZ} \\ & - \sum_{i,j,k} \exp\left(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}\right). \end{aligned}$$

The sufficient statistics are $\{n_{i++}\}$, $\{n_{+j+}\}$, $\{n_{++k}\}$, $\{n_{ij+}\}$, $\{n_{i+k}\}$, $\{n_{+jk}\}$, $\{n_{ijk}\}$. Minimal sufficient statistics are $\{n_{ijk}\}$.

Explicit MLE of μ

The minimal sufficient statistics of a given loglinear model are the MLEs for the corresponding marginal totals of μ.

- For example, consider the model (XZ, YZ). Its log-likelihood is

$$\begin{aligned} \ell &= \text{constant} + n\lambda + \sum_i n_{i++}\lambda_i^X + \sum_j n_{+j+}\lambda_j^Y + \sum_k n_{++k}\lambda_k^Z \\ &\quad +0 + \sum_{i,k} n_{i+k}\lambda_{ik}^{XZ} + \sum_{j,k} n_{+jk}\lambda_{jk}^{YZ} + 0 \\ &\quad -\sum_{i,j,k} \exp\left(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + 0 + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + 0\right). \end{aligned}$$

- The minimal sufficient statistics are $\{n_{i+k}\}$, $\{n_{+jk}\}$.
- Hence, we should have $\hat{\mu}_{i+k} = n_{i+k}$, $\hat{\mu}_{+jk} = n_{+jk}$.

MLE of μ: Another Example

Consider the model (X, Y, Z). Its log-likelihood is

$$\begin{aligned}\ell &= \text{constant} + n\lambda + \sum_i n_{i++}\lambda_i^X + \sum_j n_{+j+}\lambda_j^Y + \sum_k n_{++k}\lambda_k^Z \\ &\quad - \sum_{i,j,k} \exp\left(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z\right).\end{aligned}$$

The minimal sufficient statistics are $\{n_{i++}\}$, $\{n_{+j+}\}$, $\{n_{++k}\}$. Hence, we must have $\hat{\mu}_{i++} = n_{i++}$, $\hat{\mu}_{+j+} = n_{+j+}$, and $\hat{\mu}_{++k} = n_{++k}$.

Functions of Minimal Sufficient Statistic

MLE has an invariance property in the sense that a function of MLE is the MLE of the function, i.e., the MLE of $g(\theta)$ is $g\left(\hat{\theta}\right)$, where $\hat{\theta}$ is the MLE of θ. Hence, if we can express a μ as a closed form of expected values of minimal sufficient statistics, then we obtain the closed form expression of such μ.

- Consider the model (XZ, YZ). The MLEs are $\hat{\mu}_{i+k} = n_{i+k}$, $\hat{\mu}_{+jk} = n_{+jk}$. Since
$$\mu_{ijk} \quad = \quad n\pi_{ijk} = \frac{n\pi_{i+k}\pi_{+jk}}{\pi_{++k}} = \frac{\mu_{i+k}\mu_{+jk}}{\mu_{++k}},$$
we should have $\hat{\mu}_{ijk} = n_{i+k}n_{+jk}/n_{++k}$.
- Consider the model (X, Y, Z). The MLEs are $\hat{\mu}_{i++} = n_{i++}$, $\hat{\mu}_{+j+} = n_{+j+}$, and $\hat{\mu}_{++k} = n_{++k}$. Since
$$\mu_{ijk} \quad = \quad n\pi_{i++}\pi_{+j+}\pi_{++k} = \frac{\mu_{i++}\mu_{+j+}\mu_{++k}}{n^2},$$
we should have $\hat{\mu}_{ijk} = n_{i++}n_{+j+}n_{++k}/n^2$.

Decomposable Models

Different from other models in a three-way table, the model (XY, XZ, YZ)

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

has no closed form probabilistic expression in terms of π_{ijk} and its marginals. Hence, we don't have the closed form expression for μ_{ijk}, although we still have closed form expression for μ_{ij+}, μ_{i+k}, and μ_{+jk}.

A loglinear model having no closed form expression for the probabilistic model is nondecomposable. Otherwise, it has a closed form factorization of $\{\pi_{ijk}\}$ and it is called decomposable.

Different Sampling Designs

We said that different sampling designs are connected, e.g., the Poisson loglinear model and the multinomial loglinear model should yield the same $\boldsymbol{\beta}$ estimators.

But they should still be treated differently in model selection.

- If a marginal total is fixed by the sampling design, then the corresponding λ-term must be present in the loglinear model regardless of its lack of statistical significance.
- The reason is that we must keep the MLE of μ to be the same as the observed marginal total.
- For example, if n_{ij+} is fixed by sampling design, then we must pick models that satisfy $\hat{\mu}_{ij+} = n_{ij+}$.

An Example of Fixed Marginal Total

		Mortality		
Seedlings	Depth	Dead	Alive	Total
Longleaf	High	41	59	100
	Low	11	89	100
Slash	High	12	88	100
	Low	5	95	100

We treat Mortality as our response variable. By design, the row totals are fixed to 100. Our loglinear model should ensure $\hat{\mu}_{i+k} = 100$. Hence, λ_{ik}^{DS} must be included in our model.

Loglinear Model and Logistic Model

Consider an $I \times 2 \times K$ table. Suppose that we have fitted a (XY, XZ, YZ) loglinear model. Then,

$$
\begin{aligned}
& \log \frac{P\left(Y=1 \mid X=i, Z=k\right)}{P\left(Y=2 \mid X=i, Z=k\right)} \\
= \; & \log \frac{\pi_{i1k}}{\pi_{i2k}} = \log \frac{\mu_{i1k}}{\mu_{i2k}} \\
= \; & \left(\lambda + \lambda_i^X + \lambda_1^Y + \lambda_k^Z + \lambda_{i1}^{XY} + \lambda_{ik}^{XZ} + \lambda_{1k}^{YZ}\right) \\
& - \left(\lambda + \lambda_i^X + \lambda_2^Y + \lambda_k^Z + \lambda_{i2}^{XY} + \lambda_{ik}^{XZ} + \lambda_{2k}^{YZ}\right) \\
= \; & \underbrace{\left(\lambda_1^Y - \lambda_2^Y\right)}_{\alpha} + \underbrace{\left(\lambda_{i1}^{XY} - \lambda_{i2}^{XY}\right)}_{\beta_i^X} + \underbrace{\left(\lambda_{1k}^{YZ} - \lambda_{2k}^{YZ}\right)}_{\beta_k^Z},
\end{aligned}
$$

which is an additive logistic model.

In the log-linear model, we model the association λ_{ik}^{XZ}. But in the logistic model, we do not model λ_{ik}^{XZ}, as they cancel out.

Loglinear Model and Logistic Model

Many other loglinear models for an $I \times 2 \times K$ table is equivalent to a logistic model.

- For (Y, XZ),

$$\log \frac{P\left(Y=1 \mid X=i, Z=k\right)}{P\left(Y=2 \mid X=i, Z=k\right)} = \underbrace{\left(\lambda_1^Y - \lambda_2^Y\right)}_{\alpha}.$$

- For (XY, XZ),

$$\log \frac{P\left(Y=1 \mid X=i, Z=k\right)}{P\left(Y=2 \mid X=i, Z=k\right)} = \underbrace{\left(\lambda_1^Y - \lambda_2^Y\right)}_{\alpha} + \underbrace{\left(\lambda_{i1}^{XY} - \lambda_{i2}^{XY}\right)}_{\beta_i^X}.$$

- For (XZ, YZ),

$$\log \frac{P\left(Y=1 \mid X=i, Z=k\right)}{P\left(Y=2 \mid X=i, Z=k\right)} = \underbrace{\left(\lambda_1^Y - \lambda_2^Y\right)}_{\alpha} + \underbrace{\left(\lambda_{1k}^{YZ} - \lambda_{2k}^{YZ}\right)}_{\beta_k^Z}.$$

Higher Dimensions

We have introduced the loglinear models for three-way tables $(I \times J \times K)$. It can be generalized to contingency tables of higher dimensions.

$$
\begin{aligned}
\log \mu_{hijk} &= \lambda + \lambda_h^W + \lambda_i^X + \lambda_j^Y + \lambda_k^Z, \\
\log \mu_{hijk} &= \lambda + \lambda_h^W + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{hi}^{WX} + \lambda_{hj}^{WY} + \lambda_{hk}^{WZ}, \\
\log \mu_{hijk} &= \lambda + \lambda_h^W + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{hi}^{WX} + \lambda_{hj}^{WY} + \lambda_{hk}^{WZ} + \lambda_{ik}^{XZ} \\
&\quad + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ}, \\
\log \mu_{hijk} &= \lambda + \lambda_h^W + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{hi}^{WX} + \lambda_{hj}^{WY} + \lambda_{hk}^{WZ} + \lambda_{ik}^{XZ} \\
&\quad + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{hij}^{WXY}.
\end{aligned}
$$

Analysis of Categorical Data
Chapter 10: Build Log linear model

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through this chapter, you should be able to

1. Interpret the conditional independence graph,
2. Interpret the generator multigraph,
3. determine conditional independence and collapsibility from graphs,
4. determine decomposability,
5. obtain closed form expression of MLE.

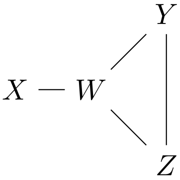
Hierarchical Loglinear Model

A loglinear model is a hierarchical loglinear model, if a higher-order term is in the model then all lower-order terms are also included in the model.

$$
\begin{aligned}
\log \mu_{ij} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}, \\
\log \mu_{ijk} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \\
\log \mu_{ijk} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \\
\log \mu_{ijk} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}, \\
\log \mu_{ijk} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}.
\end{aligned}
$$

Graph

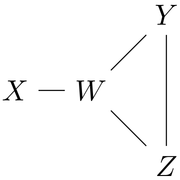
A graph consists of a set of vertices, and a set of edges connecting some vertices.



In a conditional independence graph for a hierarchical loglinear model, each vertex represents a variable.

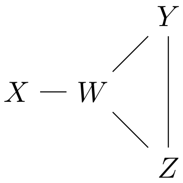
- The edges represent pairwise association, i.e., first-order interaction in the loglinear model.
- The absence of an edge connecting two variables represents conditional independence between them.

Reading Conditional Independence Graph



- The graph lacks XY and XZ.
- This can be a graph for (WX, WY, WZ, YZ).
- However, the loglinear model described by the graph is not unique. Two loglinear models with the same pairwise associations can have the same association graph. This can also be a graph for (WX, WYZ), since WYZ also includes WY, WZ, YZ.

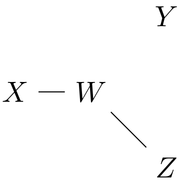
Path



- Two vertices in a graph are adjacent if an edge joint hem.
- A path from vertex X to vertex Y is a sequence of edges leading from X to Y.
- A set C of vertices separates the two sets A and B of vertices, if all paths from any vertex in A to any vertex in B pass through at least one vertex in C.
  - Consider $A = \{X\}$ and $B = \{Y\}$. Then $C = \{W\}$ separates A and B. $C = \{W, Z\}$ also separates A and B.

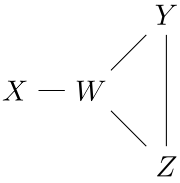
Independence

Unconditionally independent variables are said to lie in different connected components of the graph.



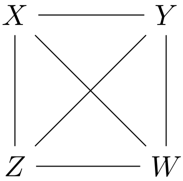
- (X, Z, W) is independent of Y.

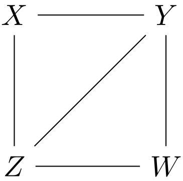
Global and Local Markov Properties



- Global Markov property: If a set C of variables separates two sets A and B of variables, then the variables in A are conditionally independent of the variables in B, given the variables in C.
  - $X \perp Y \mid W$, $X \perp Y \mid (W, Z)$.
- Local Markov property: a variable is conditionally independent of all other variables, given its adjacent neighbors to which it is connected with an edge.
  - $X \perp Y \mid W$, $X \perp (Y, Z) \mid W$.

Complete Graph

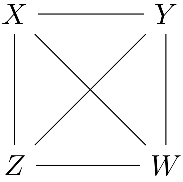


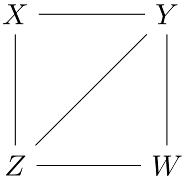


- A complete graph has an edge joining every pair of vertices.
  - LHS is a complete graph but RHS is not a complete graph.
- A maxclique in a graph is a set of vertices that form a complete graph that is not contained in a larger complete graph.
  - LHS has a maxclique: $(XYZW)$.
  - RHS has two maxcliques: (XYZ) and (YZW).

Graphical Models

A hierarchical loglinear model is called a graphical model when there is a one-to-one correspondence between the maxcliques in the conditional independence graph and the generating class of the loglinear model.

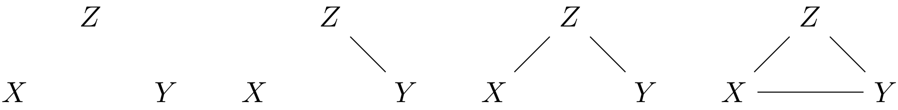




- Both $(XYZW)$ and (XYZ, WYZ, WX) have the same conditional independence graph on the left.
    - $(XYZW)$ is a graphical model. (XYZ, WYZ, WX) is not a graphical model since the generating class is not the same as the maxclique.
- The generating class (XYZ, YZW) yields the conditional independence graph on the right, and is graphical.

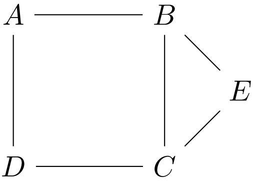
Maxclique and Minimal Sufficiency

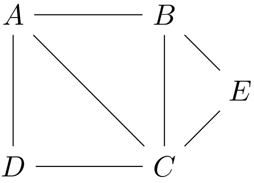
In a graphical model, the maxcliques are always related to the minimal sufficient statistics.



1. The maxclique is (X, Y, Z), same as generating class. The minimal sufficient statistic is $\{n_{i++}\}$, $\{n_{+j+}\}$, $\{n_{++k}\}$.
2. The maxclique is (X, YZ), same as generating class. The minimal sufficient statistic is $\{n_{i++}\}$ and $\{n_{+jk}\}$.
3. The maxclique is (XZ, YZ), same as generating class. The minimal sufficient statistic is $\{n_{i+k}\}$and $\{n_{+jk}\}$
4. The maxclique is (XYZ). If the generating class is (XY, XZ, YZ), then the model is not graphical.
    - The graphical model has minimal sufficient statistic $\{n_{ijk}\}$.

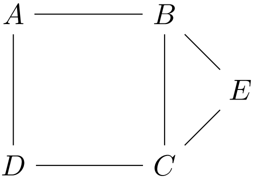
Chordal Graph

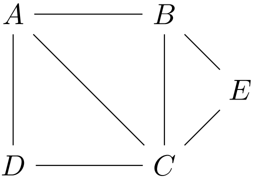




- A cycle is a sequence of edges that begins and ends at a given vertex, e.g., $B - C - E - B$ is a cycle.
- A chord is an edge between nonconsecutive vertices along a cycle.
  - Graph on the right: $A - B - C - D - A$ is a cycle, and the edge $A - C$ is a chord.
- A chordal graph is a graph where there exists a chord in every cycle of length four or more. If all cycles have length less than four, then the graph is also chordal.
  - Graph on the left is not chordal.
  - Graph on the right is chordal.

Decomposable Model





A decomposable model is defined to be a graphical model whose conditional independence graph is a chordal graph.

- The generating class (AB, BC, CD, AD, BCE) yields the graph on the left, and is a graphical model. The model is not decomposable because it is not chordal.
- The generating class (ABC, ACD, BCE) yields the graph on the right, and is a graphical model. The model is decomposable because it is also chordal.
- The generating class $(AB, BC, AC, AD, CD, BCE)$ yields the graph on the right, and is a chordal model, but not graphical.

Benefit of Decomposable Model

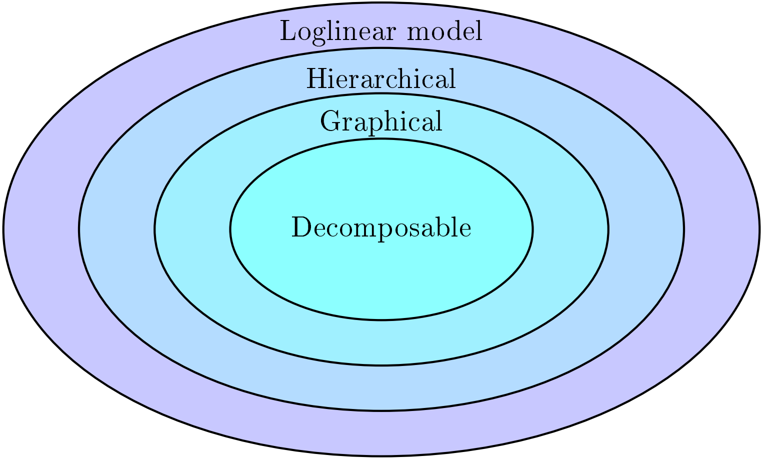
We have seen in Chapter 9 that a benefit of a decomposable model is that the joint probability of the contingency table can be factored in closed form with respect to the indices in the generating class, and the maximum likelihood estimators of μ are available in closed form only for decomposable models.

- The model (XZ, YZ) is decomposable. The MLEs are $\hat{\mu}_{i+k} = n_{i+k}$, $\hat{\mu}_{+jk} = n_{+jk}$. We also have

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}.$$

  Hence, $\hat{\mu}_{ijk} = n_{i+k}n_{+jk}/n_{++k}$.
- The model (XY, XZ, YZ) is not decomposable because it is not graphical. Hence, it has no closed form probabilistic expression in terms of π_{ijk} and its marginals.

Summary of Loglinear Models



Collapsibility in Three-way Contingency Table

Collapsibility condition for three-way table: In a three-way table, a variable is collapsible with respect to the interaction between the other two variables if it is at least conditionally independent of one of the other two variables given the third variable.

The XY marginal odds ratio and the XY conditional odds ratio are identical if

1. either $X \perp Z \mid Y$,
2. or $Y \perp Z \mid X$,
3. or both.

For example, these conditions occur for loglinear models (XY, YZ), (XY, XZ), and (XY, Z). We say that Z is collapsible with respect to the XY association.

Collapsibility in the (XY, XZ) Model

Consider the (XY, XZ) model

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}.$$

The XY marginal table satisfies

$$\begin{aligned} \log \mu_{ij+} &= \log \left[\sum_k \exp\left(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} \right) \right] \\ &= \lambda + \underbrace{\lambda_i^X + \log \left[\sum_k \exp\left(\lambda_k^Z + \lambda_{ik}^{XZ} \right) \right]}_{\text{new main effect } \lambda_i^X} + \lambda_j^Y + \lambda_{ij}^{XY}, \end{aligned}$$

where the interaction λ_{ij}^{XY} remains the same as in the partial table. Since the XY odds ratios are functions of λ_{ij}^{XY}, they are the same in both marginal and partial tables.

An Example

From the conditional independence graph, we can tell whether we can collapse a variable. Consider the conditional independence graph

$$A - M - C$$

The model corresponding model is (AM, CM), i.e., $A \perp C \mid M$. Hence, by the collapsibility condition,

- the AM conditional odds ratio is identical to the AM marginal odds ratio collapsed over C,
- the CM conditional odds ratio is identical to the CM marginal odds ratio collapsed over A,
- the AC conditional odds ratio is not necessarily identical to the AC marginal odds ratio collapsed over M.

Collapsibility in Multiway Contingency Tables

Collapsibility condition for multiway table: Suppose that a model for a multiway table partitions variables into three mutually exclusive subsets, A, B, C, such that B separates A and C. After collapsing the table over the variables in C,

1. all λ terms relating variables in A are unchanged,
2. all λ terms relating variables in A to variables in B are unchanged,
3. the λ terms relating variables in B are not invariant to collapsing.

A motivation is that, by the Markov property, $A \perp C \mid B$. In other words, conditional independence yields collapsibility.

- But such collapsibility condition is only a sufficient condition, not a necessary condition.
- Their numerical estimates may differ slightly.

Consequence of Collapsibility Theorem

Recall that the conditional odds ratio depend on λ. If B separates A and C, then one can collapse over the factors in C

- without distorting the the association between a factor in A and a factor in B,
- without distorting the the association between factors in A.

Consider the (XY, XZ) model, where X separates Y and Z. Hence,

$$\begin{aligned} \log \mu_{ijk} &= \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ}, \\ \log \mu_{ij+} &= \lambda + \beta_i^X + \lambda_j^Y + \lambda_{ij}^{XY}. \end{aligned}$$

That is, λ_j^Y and λ_{ij}^{XY} remain the same, but λ_i^X may not. At a fixed level $Z = k$, the local odds ratios satisfies

$$\log \theta_{ij(k)} = \log \left(\frac{\mu_{ijk}\mu_{i+1,j+1,k}}{\mu_{i+1,j,k}\mu_{i,j+1,k}} \right) = \lambda_{ij}^{XY} + \lambda_{i+1,j+1}^{XY} - \lambda_{i+1,j}^{XY} - \lambda_{i,j+1}^{XY}.$$

which remains unchanged.

Example: Collapsibility in Multiway Tables

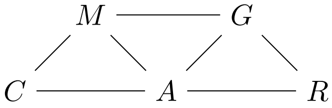
(WX, WY, WZ, YZ) model

1. Let $A = X$, $B = W$, and $C = \{Y, Z\}$. If we collapse over Y and Z, the WX association is unchanged.
2. Let $A = \{Y, Z\}$, $B = \{W\}$, and $C = \{X\}$. Associations among W, Y, Z remain the same after collapsing over X.
3. Let $A = \{X\}$, $B = \{W, Z\}$, and $C = \{Y\}$. Associations among W, Y, Z remain the same after collapsing over X. The association between W and Z is also unchanged. But if we collapse Y, the WZ association is changed.

(WX, WY, XY, Z) model

Let $A = \{X, Y\}$, $B = W$, and $C = \{Z\}$. If we collapse over Z, the associations among W, X, Y are unchanged.

Another Example



- The model is $(AC, AM, CM, AG, AR, GM, GR)$.
- $\{A, M\}$ separates C and $\{G, R\}$.
- $C \perp (G, R) \mid (A, M)$.
- The conditional associations between C and A, and between C and M are the same with the model (AC, AM, CM).

Generator Multigraph

The generator multigraph, or simply multigraph, is an alternative to the conditional independence graph for hierarchical loglinear models.

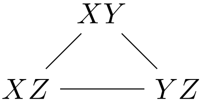
- Different from a conditional independence graph, a multigraph allows a vertex with more than one variable.

A multigraph consists of vertices and multiedges.

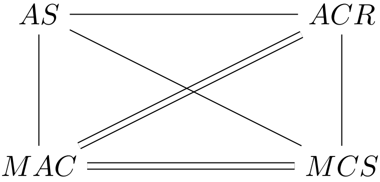
$$\begin{aligned} \text{vertex} &= \text{generators in the generating class} \\ \text{multiedge} &= \text{edges equal in number to the number of indices common} \\ &\quad\ \text{to the two generators being joined.} \end{aligned}$$

Examples of Multigraph

Consider the model with generating class (XY, XZ, YZ). The vertices are XY, XZ, and YZ. Each pair of vertices share a single index, so there is a single edge joining each pair. The multigraph is

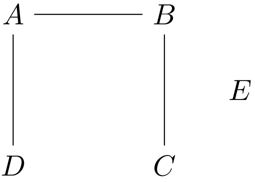


Consider the model with generating class (AS, ACR, MCS, MAC). The vertices are AS, ACR, MCS, and MAC. The pair (ACR, MAC) share two indices, so is the pair (MCS, MAC). The multigraph is

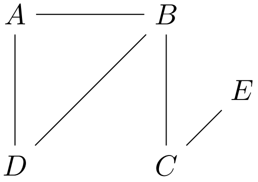


Tree

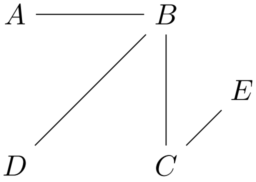
A connected graph is any graph for which there is at least one path from any vertex to any other vertex in the graph. A tree is a connected graph with no circuit (closed loop) that includes each vertex of the graph.



Not connected.



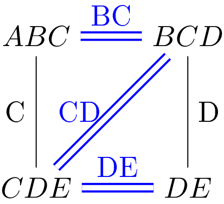
Connected graph, but not a tree because of the circuit including A, B, and D.



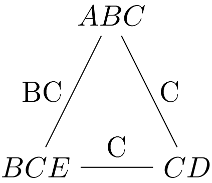
A tree.

Maximum Spanning Tree

For a multigraph, a maximum spanning tree is a tree where the sum of all the edges is maximum. The indices common to two vertices being joined in the maximum spanning tree are called the branches of the tree. The maximum spanning tree is not unique, neither is the branch set.



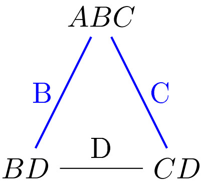
The maximum spanning tree is $[ABC]\,[BCD]\,[CDE]\,[DE]$. The branch set is $\{(BC)\,,(CD)\,,(DE)\}$



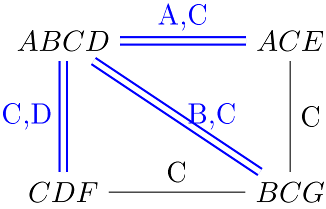
The maximum spanning tree can be $[ABC]\,[BCE]\,[CD]$ or $[CD]\,[ABC]\,[BCE]$. The branch set is $\{BC, C\}$ for both cases.

Sufficient and Necessary Condition

A hierarchical loglinear model is decomposable if and only if, in any maximum spanning tree of the multigraph, the number of factors equals to the number of indices added over the vertices minus the number of indices added over the branches of the maximum spanning tree.



Number of indices added over vertices is 7. Number of indices added over branches is 2. Number of factors is 4. Hence, not decomposable.



Number of indices added over vertices is 13. Number of indices added over branches is 6. Number of factors is 7. Hence, decomposable.

Joint Probability of Decomposable Model

Once we know that a loglinear model is decomposable, we want to factorize the joint probability of the contingency table in closed form with respect to the indices in the generating class.

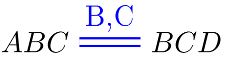
- Let $P(\ell_1, \ell_2, ..., \ell_d)$ be the probability that a subject is in level ℓ_1 of the first factor, level ℓ_2 of the second factor, ..., and level ℓ_d of the dth factor.
- Let S be a subset of indices from the set $\{1, 2, ..., d\}$. We use p_S to denote the marginal probability having indices contained in S while all other indices are summed over.

Then, for a decomposable model,

$$P(\ell_1, \ell_2, ..., \ell_d) = \frac{\prod_{S \in V} p_S}{\prod_{S \in B} p_S},$$

where V is the set of indices in the vertex set of the maximum spanning tree and B represents the multiset of indices in the branch set.

Find Joint Probability: One Example



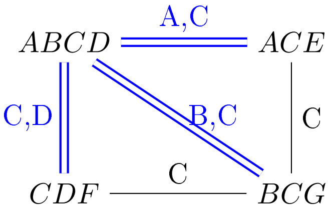
The maximum spanning tree is $[ABC]\,[BCD]$, and the branch set is $\{BC\}$.

- Number of indices added over vertices is 6.
- Number of indices added over branches is 2.
- Number of factors is 4. Hence, the loglinear model is decomposable.

Hence,

$$\pi_{ijkl} \quad = \quad \frac{\pi_{ijk+}\pi_{+jkl}}{\pi_{+jk+}}.$$

Find Joint Probability: Another Example



The maximum spanning tree is $[ABCD]\,[CDF]\,[ACE]\,[BCG]$, and the branch set is $\{(AC)\,,(BC)\,,(CD)\}$.

- Number of indices added over vertices is 13.
- Number of indices added over branches is 6.
- Number of factors is 7. Hence, the loglinear model is decomposable.

Hence,

$$\pi_{abcdefg} = \frac{\pi_{abcd+++}\pi_{++cd+f+}\pi_{a+c+e++}\pi_{+bc+++g}}{\pi_{a+c++++}\pi_{+bc++++}\pi_{++cd+++}}.$$

Fundamental Conditional Independence Set

Given a multigraph, we would like to find a fundamental conditional independence set.

- Suppose that the fundamental conditional independence set is $[C_1 \otimes C_2 \otimes \cdots \otimes C_k \mid S]$, where d factors in a contingency table have been partitioned by the $k+1$ sets of factors C_1,C_2, ..., C_k, and S.
- We can replace S with S' such that $S \subseteq S'$, and replace each C_i with C_i' such that $C_i' \subseteq C_i$, subject to

$$\left(C_1' \cup C_2' \cup \cdots \cup C_k'\right) \cap S' \quad = \quad \emptyset.$$

- Then, $[C_1' \otimes C_2' \otimes \cdots \otimes C_k' \mid S']$.

For example, if the fundamental conditional independence set is $[A, B \otimes D \mid E]$, then $[A \otimes D \mid B, E]$ and $[B \otimes D \mid A, E]$.

Decomposable Model

Denote a multigraph by M. Suppose that the resulting loglinear model is decomposable. We can obtain the fundamental conditional independence set using the following steps.

- Suppose that there are d factors in the contingency table. Let S be a subset of these factors.
- We construct a new multigraph M/S by removing each factor of S from each vertex in the multigraph, and removing each edge corresponding to that factor.
- Then, we have mutual independence of the sets of factors in the disconnected components of M/S, conditional on S.

We can simply let S be the branches of the maximum spanning tree. The resulting conditional independence sets

Conditional Independence: One Example

$$ABC \xlongequal{\text{B,C}} BCD$$

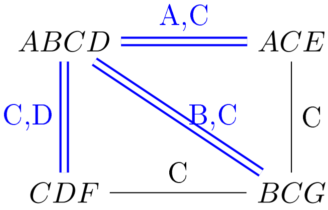
We have shown that the model is decomposable. The branch set is $\{BC\}$

- Let $S = \{B, C\}$, the branch set.
- To construct M/S, we remove all indices in S from each vertex and the edge corresponding to them.

$$A \qquad\qquad\qquad D$$

- M/S results in two disconnected components, i.e., $\{A\}$ and $\{D\}$.
- Hence, $A \perp D \mid (B, C)$. We often write $[A \otimes D \mid B, C]$ as well, in case there are more than two disconnected components.

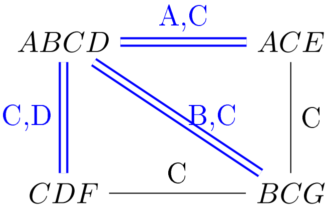
Conditional Independence: Another Example

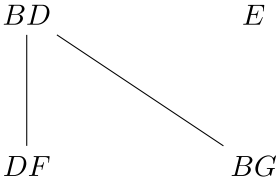


We have shown that the model is decomposable. The branch set is $\{(AC), (BC), (CD)\}$.

- We can choose $S = \{A, C\}$, or $S = \{B, C\}$, or $S = \{C, D\}$.

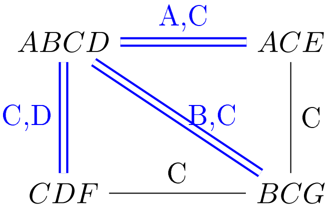
Conditional Independence: Another Example (2)

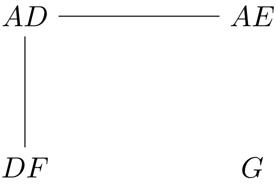




- If we choose $S = \{A, C\}$ and removing the factors in S from each vertex, then M/S is the graph on the right.
- M/S results in two disconnected components, i.e., $\{B, D, F, G\}$ and $\{E\}$.
- Hence, $[B, D, F, G \otimes E \mid A, C]$, or $(B, D, F, G) \perp E \mid (A, C)$.

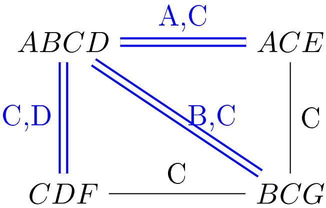
Conditional Independence: Another Example (3)

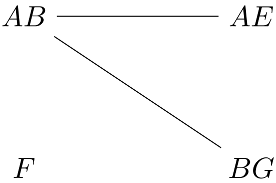




- If we choose $S = \{B, C\}$ and removing the factors in S from each vertex, then M/S is the graph on the right.
- M/S results in two disconnected components, i.e., $\{A, D, E, F\}$ and $\{G\}$.
- Hence, $[A, D, E, F \otimes G \mid B, C]$, or $(A, D, E, F) \perp G \mid (B, C)$.

Conditional Independence: Another Example (4)





- If we choose $S = \{C, D\}$ and removing the factors in S from each vertex, then M/S is the graph on the right.
- M/S results in two disconnected components, i.e., $\{A, B, E, G\}$ and $\{F\}$.
- Hence, $[A, B, E, G \otimes F \mid C, D]$, or $(A, B, E, G) \perp F \mid (C, D)$.

Nondecomposable Model

To find the fundamental conditional independence set for a nondecomposable model, we need to use a new concept called edge cutsets.

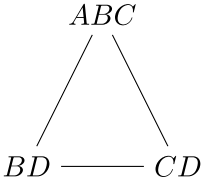
- An edge cutset of a multigraph is an inclusion-minimal set of multiedges whose removal disconnects the multigraph.

After removing Z from the graph, we obtain disconnected components. Hence, the edge cutset is $\{Z\}$.

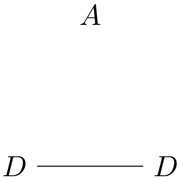
- For decomposable models, the edge cutsets are the branches of the maximum spanning tree.

Edge Cutset: Example

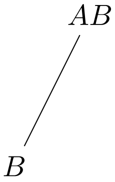
Consider the nondecomposable model on the right.



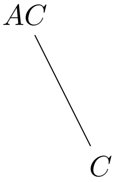
Edge cutset $S_1 = \{B, C\}$



Edge cutset $S_2 = \{C, D\}$



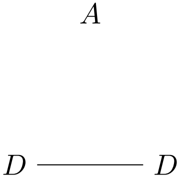
Edge cutset $S_3 = \{B, D\}$



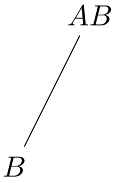
Identify Fundamental Conditional Independence Set

To identify the fundamental conditional independence set for nondecomposable models, we follow the same steps as in decomposable models, except that the indices in S come from the edge cutsets instead of the branches of a maximum spanning tree.

Edge cutset $S_1 = \{B, C\}$. Hence, $[A \otimes D \mid B, C]$.



Edge cutset $S_2 = \{C, D\}$. Only one component left, so no conditional independence set.



Edge cutset $S_3 = \{B, D\}$. Only one component left, so no conditional independence set.