UPPSALA UNIVERSITET

Lecture Notes

Analysis of Categorical Data

Rami Abou Zahra

Inlämningsdatum
October 23, 2025

Contents

1. Chapter 1 & 2

- Nominal: no ordering behind the categories

Note: The features can be continuous, but in this course the categorical variable is discrete.

1.1. Slide 33.

One can always construct a table whose partial tables has odds ratio 1. For the Berkley data, looking at the universiuty as a whole we had independence but dependence when looking departmentwise. Just because the odds ratio is 1, does not mean that the marginal odds will also be 1.

		Y	
Z	X	0	1
Z_1	0	100	10
Z_2	1	200	20
Z_3	0	100	50
Z_4	1	60	30

Here, the odds ratio is $\frac{100 \cdot 20}{10 \cdot 200} = 1 = \frac{100 \cdot 30}{50 \cdot 60}$, but the marginal table looks like this:

	Y	
X	0	1
0	100+100	10+50
1	200+60	20+30

We can see that $\theta_{xy} = \frac{200 \cdot 50}{60 \cdot 260} \neq 1$

1.2. Slide 38.

Odds ratio can be computed by pairwise computation.
Local odds ratio: *only* adjacent, eg $X = 0$ and $Y = 2$ columns will not be included. Only this is needed.

1.3. Slide 41.

For the following table:

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

Assuming multinomial sampling with the total sum being fixed to nm we wish to find the distribution of all n_{ij}. Since n is known, we normalize:

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

Note that this is indeed a valid estimation, since they all sum to 1. Also, since they sum to 1, we only need to know three of them. When we want to estimate the distribution of $\frac{1}{n}\begin{bmatrix} n_{11} \\ n_{12} \\ n_{21} \end{bmatrix}$, we use the CLT:

$$\frac{1}{\sqrt{n}}\left(\frac{1}{n}\begin{bmatrix} n_{11} \\ n_{12} \\ n_{21} \end{bmatrix} - \begin{bmatrix} \pi_{11} \\ \pi_{12} \\ \pi_{21} \end{bmatrix}\right) \approx N\left(0, \begin{bmatrix} \pi_{11}(1-\pi_{11}) & -\pi_{11}\pi_{12} & \pi_{11}\pi_{21} \\ & \pi_{12}(1-\pi_{12}) & -\pi_{12}\pi_{21} \\ & & \pi_{21}(1-\pi_{21}) \end{bmatrix}\right)$$

Example: Consider $g(x_1, x_2, x_3) = \ln(x_1) - \ln(x_2) - \ln(x_3) + \ln(1 - x_1 - x_2 - x_3)$

If:

$$\left.\begin{array}{l} \dfrac{n_{11}}{n} = x_1 \\ \dfrac{n_{12}}{n} = x_2 \\ \dfrac{n_{21}}{n} = x_3 \end{array}\right\} \quad \ln\left(\frac{n_{11}}{n}\right) - \ln\left(\frac{n_{12}}{n}\right) - \ln\left(\frac{n_{21}}{n}\right) + \underbrace{\ln\left(1 - \frac{n_{11}}{n} - \frac{n_{12}}{n} - \frac{n_{21}}{n}\right)}_{\ln\left(n_{22}/n\right)} = \ln\left(\widehat{\theta}\right)$$

To find the distribution of $\ln\left(\widehat{\theta}\right)$, apply the delta method to g:

$$\frac{\partial g}{\begin{bmatrix} \partial x_1 \\ \partial x_2 \\ \partial x_3 \end{bmatrix}} = \begin{bmatrix} \dfrac{1}{x_1} - \dfrac{1}{1 - x_1 - x_2 - x_3} \\ \vdots \end{bmatrix}$$

$$\Rightarrow \ln\left(\widehat{\theta}\right) - \ln\left(\theta_0\right) \approx N(0, ?)$$

$$? = \left[\frac{1}{\pi_{11}} - \frac{1}{\pi_{22}}, -\frac{1}{\pi_{12}} - \frac{1}{\pi_{22}}, -\frac{1}{\pi_{21}} - \frac{1}{\pi_{22}}\right] \begin{bmatrix} \pi_{11}(1-\pi_{11}) & -\pi_{11}\pi_{12} & \pi_{11}\pi_{21} \\ & \pi_{12}(1-\pi_{12}) & -\pi_{12}\pi_{21} \\ & & \pi_{21}(1-\pi_{21}) \end{bmatrix}$$

$$\begin{bmatrix} \dfrac{1}{\pi_{11}} - \dfrac{1}{\pi_{22}} \\ -\dfrac{1}{\pi_{12}} - \dfrac{1}{\pi_{22}} \\ -\dfrac{1}{\pi_{21}} - \dfrac{1}{\pi_{22}} \end{bmatrix} = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

The last equality holds regardless of sampling method.

2. Chapter 3

2.1. Slide 6.

It does not need to be from multinomial sampling, but then we would need to change the likelilhood-function.
Under H_0, we have $(I-1)+(J-1)$ parameters under H_1 we have $IJ-1$ parameters

3. Chapter 4

3.1. Slide 9.

$$\frac{\partial \ell}{\partial \beta} = \frac{\partial \ell}{\partial \theta}\frac{\partial \theta}{\partial \mu}\frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \beta}$$

With the following holding:

$$\mu = b'(\theta) \Rightarrow \frac{\partial \mu}{\partial \theta} = b''(\theta) = \frac{\operatorname{Var}(Y_i)}{\phi_i}$$
$$\Rightarrow \frac{\partial \theta}{\partial \mu} = \frac{\phi_i}{\operatorname{Var}(Y_i)}$$
$$\eta = x_i^T \beta \Rightarrow \frac{\partial \eta}{\partial \beta} = x_i$$

This yields for functions belonging to the exponential family:

$$\frac{\partial}{\partial \beta}\left[\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i,\phi_i)\right] = \overbrace{\frac{y-i-b'(\theta_i)}{\phi_i}}^{\partial \ell/\partial \theta} \cdot \frac{\phi_i}{\operatorname{Var}(Y_i)} \cdot \frac{\partial \mu}{\partial \eta} \cdot x_i$$

3.2. Slide 12.

In the Poisson case, we have the link-function $g(\mu) = \ln(\mu) \Rightarrow \mu = \exp\{\eta_i\}$, this yields the following matrices:

- $\boldsymbol{D} = \dfrac{\partial \mu_i}{\partial \eta} = \exp\{\eta_i\} = \exp\left\{x_i^T \beta\right\}$
- $\boldsymbol{V} = \mu_i = \exp\{\eta\}\exp\left\{x_i^T\beta\right\}$

3.3. Slide 13.

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

3.4. Slide 16.

The Fisher infortmation can be expressed in the following way:

$$I(\beta) = \operatorname{Var}\left(\frac{\partial \ell}{\partial \beta}\right) = -\mathbb{E}\left[\frac{\partial^2 \ell(\beta)}{\partial^2}\right] = \operatorname{Var}\left(\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right) = \boldsymbol{X}^T\boldsymbol{D}\boldsymbol{V}^{-1}\underbrace{\operatorname{Var}(\boldsymbol{y}-\boldsymbol{\mu})}_{=\boldsymbol{V}}\boldsymbol{V}^{-1}\boldsymbol{D}\boldsymbol{X}$$
$$= \boldsymbol{X}^T\boldsymbol{D}\boldsymbol{V}^{-1}\boldsymbol{D}\boldsymbol{X}$$

3.5. Slide 20.

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

3.6. Slide 21.

The likelihood function here is not the same the same as the one for β, since the MLE assumes β follows a certain model.

3.7. **Slide 25.**
Note that we have continuous data in the x_2 column, therefore it is automatically ungrouped

3.8. **Slide 32.**
Here $n =$ number of observations

3.9. **Slide 35.**
Note that we have patterns due to ungrouped data.

3.10. **Slide 37.**
In the second figure, we have the quantile regression (3 lines). Reading this, they should be straight in the quantiles .75, .5, .25

4. Chapter 5 & 6

4.1. **Slide 4.**

Say we have two models, $M_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ and $M_2 = \beta_0 + \beta_1 x_2 + \beta_2 x_2 + \beta_3 x_1 x_2$
In M_1, if there is change in x_1 there is no change in the other covariates not concerning x_1, however, in M_2, if we change x_1 then this also changes $\beta_1 \wedge \beta_3$

4.2. **Slide 5.**

Prospective: Look ahead in time (will smoking cause cancer?)
Case-control: Data already available (is probability of cancer higher given that you smoke?)

In the prospective case,

$$P(Y = 1 \mid X) = \frac{\exp\{\alpha + \beta x\}}{1 + \exp\{\alpha + \beta x\}}$$

Since we are using the logit model, $\ln\left(\frac{\pi}{1+\pi}\right) = \alpha + \beta x$
In the case-control case it depends on the sampling. It yields slightly different models and probabilities

$$P(Z = 1 \mid Y = 1), \quad P(Z = 1 \mid Y = 0)$$

If we take a study of cancer vs driving+smoking, then we look at the subset we are interested in. This Z variable is that, if what we are interested in is sampled in the observation.

$$P(Y = 1 \mid Z = 1, X) = \frac{P(Z = 1 \mid Y = 1, X)P(Y = 1 \mid X)}{P(Z = 1 \mid Y = 1, X)P(Y = 1 \mid X) + P(Z = 1 \mid Y = 0, X)P(Y = 0 \mid X)}$$

$$\begin{aligned}
&= \frac{P(Z = 1 \mid Y = 1)\dfrac{\exp\{\alpha + \beta x\}}{1 + \exp\{\alpha + \beta x\}}}{P(Z = 1 \mid Y = 1)\dfrac{\exp\{\alpha + \beta x\}}{1 + \exp\{\alpha + \beta x\}} + \dfrac{P(Z = 1 \mid Y = 0)}{1 + \exp\{\alpha + \beta x\}}} \\
&= \frac{P(Z = 1 \mid Y = 1)\exp\{\alpha + \beta x\}}{P(Z = 1 \mid Y = 1)\exp\{\alpha + \beta x\} + P(Z = 1 \mid Y = 0)} \\
&= \frac{P(Z = 1 \mid Y = 1)\exp\{\alpha + \beta x\}/P(Z = 1 \mid Y = 0)}{1 + \dfrac{P(Z = 1 \mid Y = 1)\exp\{\alpha + \beta x\}}{P(Z = 1 \mid Y = 0)}} \\
&\Rightarrow \alpha = \alpha + \ln\left(\frac{P(Z = 1 \mid Y = 1)}{P(Z = 1 \mid Y = 0)}\right)
\end{aligned}$$

4.3. **Slide 17.**

If X is the gender, and Z is the department (from the Berkley admission data example), then

- $\beta_i^X = \begin{cases} 1 & \text{female} \\ 0 & \text{male} \end{cases}$
- $\beta_k^Z = \begin{cases} 1 & \text{dept. k} \\ 0 & \text{else} \end{cases}$
- $\beta_{ik}^{XZ} = \begin{cases} 1 & \text{dept. k} \wedge \text{female} \\ 0 & \text{else} \end{cases}$

4.4. **Slide 20.**

Residual deviance tells us if we have a good model by comparing to χ^2. If it is less than χ^2, then the model assuming conditional independence is good.

In order to compare models, we compare their residuals. Say we have two models M_1 and M_2 with respective residuals R_1 and R_2, then we compare $R_1 - R_2$ with $\chi^2(1)$ (one-degree of freedom). If the difference is larger than the χ^2, we reject the null-hypothesis that we have conditional independence.

4.5. **Slide 21.**
This is like the Fisher-exact test, but without confounding factos.

4.6. **Slide 22.**
If we have confounding factors, we do the Fisher-exact on each partial table. If they are independent, then all are hypergeometrically distributed.

4.7. **Slide 23.**
We need homogeneous association here!

4.8. **Slide 24.**
As $k \to \infty$, we get more partial tables.

$$\underbrace{1}_{\text{intcpt.}} + \underbrace{1}_{\beta_1} + \underbrace{(k-1)}_{\text{deg. free.}}$$

4.9. **Slide 26.**
We should be seeing convergence to $\beta = 0.5$, but it is centered around 1 instead. This is because $n = 2k$ and the MLE converges to 2β

4.10. **Slide 27.**
CMH is popular for Meta-analysis.

4.11. **Slide 28.**
If $Z \wedge X \mid Y$ are conditionally independent, then they are conditionally independent given Y, i.e $Z \perp X \mid Y \Rightarrow P(X = 1 \mid Y = j, Z = k) \overset{\text{cond. indep}}{=} P(X = 1 \mid Y = j)$
Odds-ratio between X, Y for some level of Z is given by:

$$\begin{aligned}
&\frac{P(X = 1 \mid Y = 1, Z = k)P(X = 2 \mid Y = 2, Z = k)}{P(X = 1 \mid Y = 2, Z = k)P(X = 2 \mid Y = 1, Z = k)} \\
= &\underbrace{\frac{P(X = 1 \mid Y = 1)P(X = 2 \mid Y = 2)}{P(X = 1 \mid Y = 2)P(X = 2 \mid Y = 1)}}_{\text{marginal table}} = \theta(X, Y, Z = k) \leftarrow \quad \text{partial table}
\end{aligned}$$

4.12. **Slide 30.**
Just use residual deviance, if larger than χ^2, then bad & switch to saturated model.

5. Chapter 7

5.1. **Slide 5.**

For F_ε, what distribution you might ask? Logistic distribution:

$$\varepsilon \sim F(X) = \frac{\exp\{x\}}{1+\exp\{x\}} \quad \pi \Rightarrow \frac{\exp\left\{x^T\beta\right\}}{1+\exp\{x^T\beta\}} \Leftrightarrow \ln\left(\frac{\pi}{1-\pi}\right) = x^T\beta$$

Choice of link-function needs to be motivated for the error.

5.2. **Slide 13.**

Here it is assumed $y_i \in \{0, 1\}$

5.3. **Slide 14.**

This is for the Frequentist only. We assume we have a distribution with known θ. Y_i is observed from distribution.
A *statistic* $T = T(Y_1, \ldots, Y_n)$, conditional distribution of data given sufficient statistic does not depend on θ, all information about θ is contained in the sufficient statistic.
Minimal sufficient statistic uses the smallest dimension.
As an example, assume we have two samples X_i, Y_j. Take the ratio $\dfrac{f(Y \mid \theta)}{f(X \mid \theta)}$. A statistic $T(Y)$ is minimal sufficient if the ratio does not depend on $\theta \Leftrightarrow T(X) = T(Y)$
In order to not be dependent on α,, we need

$$\sum_i y_i = \sum_i y_i^*$$

so that they cancel. Thus, $\sum_i y_i$ is a minimal sufficient statistic.

5.4. **Slide 15.**

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2$$

Here, α and β_2 are nuisance parameters, and the minimal sufficent statistic for

- α: $\sum y$
- β_j: $\sum y_i x_{ij}$

6. Chapter 8

6.1. Slide 4.

In order to get rid of π_C:

$$1 = \sum_{j=1}^{J} \pi_j(x) = \sum_j \pi_C(x) \exp\left\{\alpha_j + \beta_j^T x\right\} \qquad \left.\begin{array}{r} \alpha_C = 0 \\ \beta_C = 0 \end{array}\right\} \Rightarrow \exp\{0\} = 1$$

$$\Rightarrow \pi_C = \frac{1}{\sum_j \exp\left\{\alpha_j + \beta_j^T x\right\}}$$

6.2. Slide 5.

We use maximum likelihood to find α, β. C can be freely chosen and $= 0$, this does not change probability due to softmax since:

$$\frac{\exp\{z_1\} / \exp\{z_C\}}{\left(\sum_i \exp\{z_i\}\right) \exp\{z_C\}}$$

$$\ln\left(\frac{\pi_j}{\pi_C}\right) = \alpha_j + \beta_j^T x$$

$$\ln\left(\frac{\pi_j}{\pi_A}\right) = \ln(\pi_j) - \ln(\pi_A) = \ln\left(\frac{\pi_j}{\pi_C}\right) - \ln\left(\frac{\pi_A}{\pi_C}\right)$$

$$= (\alpha_j - \alpha_A) + \underbrace{(\beta_j - \beta_A)}_{\substack{\text{normal since} \\ \begin{bmatrix} \widehat{\beta}_j \\ \widehat{\beta}_k \end{bmatrix} \text{ joint normal}}}^{T} x$$

6.3. Slide 6.

$$P(Y = j) P(U_j > U_k \quad \forall k \neq j) = \int P(U_j > U_k \quad \forall k \neq j) f(U_j) dU_j$$

By the law of total probability

6.4. Slide 7.

Ordering, eg grades, if you get a 4, then you also get a 3. We do cummulative probabilities.

6.5. Slide 10.

$$\ln\left(\frac{P(Y < j)}{1 - P(Y \leq j)}\right) = \alpha_j \beta^T x$$

(same as the binary outcome)

$$\frac{P(Y \leq j \mid x_1)}{1 - P(Y \leq j \mid x = x_1)} = \exp\left\{\beta^T (x_1 - x_2)\right\} \frac{P(Y \leq j \mid x_2)}{1 - P(Y \leq j \mid x = x_2)}$$

6.6. Slide 11.

Ordinal data assumes $\beta_j = \beta_i \ \forall i, j$. Because of total representation.

6.7. Slide 13.

Can happen here that as we include more categories, our probabilities decrease.

7. Chapter 9

7.1. Slide 4.

$N_1, \cdots, N_C$ are independent Poisson with mean μ_i . $P(N_1 = n_1, \cdots, N_C = n_C \mid \sum_{i=1}^{C} N_i = n)$ for Poisson, total number of observations in random variable but we condition on it so we know it.

$$= \frac{P(N_1 = n_1 \cdots, N_C = n_C)}{P(\sum_{i=1}^{C} N_i = n)} \overset{\text{indep.}}{=} \frac{\prod_{i=1}^{C} \frac{\mu_i^{n_i}}{n_i!} \exp\{-\mu_i\}}{\sum N_i \sim \mathrm{Po}(\sum \mu_i)}$$

$$\Rightarrow \frac{\prod_{i=1}^{C} \frac{\mu_i^{n_i}}{n_i!} \exp\{-\mu_i\}}{\frac{(\sum \mu_i)^n}{n!} \exp\{-\sum \mu_i\}} = \frac{n! \prod_{i=1}^{C} \mu_i^{n_i}}{\prod_{i=1}^{C} n_i! \, (\sum \mu_i)^n} = \frac{n!}{\prod_{i=1}^{C} n_i!} \prod_{i=1}^{C} \left(\frac{\mu_i}{\sum \mu_i} \right)^{n_i}$$

$$\Rightarrow \pi_i = \frac{\mu_i}{\sum \mu_i} \leftarrow \quad \text{multinomial}$$

7.2. Slide 5.

Log-likelihood for Poisson:

$$\frac{\mu^y}{y!} \exp\{-\mu\} \Rightarrow \exp\{y \ln(\mu) - \mu - \ln(y!)\}$$

Maximizing this like-likehood, i.e $\frac{\partial \ell}{\partial \lambda} = 0$ yields

$$\frac{\partial \ell}{\partial \lambda} = \exp\{\lambda\} \sum_i \exp\{\beta^T x\} = \sum_i \underbrace{\exp\{\lambda + \beta^T x\}}_{\substack{\text{exponential}=\mu_i \\ \text{transformed linear} \\ \text{predictor}}}$$

7.3. Slide 6.

β:s are almost the same, even if we start with Poisson or multinomial, since λ in Poisson but multinomial cancels.
Given contingency table, just build loglinear and find β (simplest and fastest way)

7.4. Slide 11.

For association, we only care about the interaction term λ^{XY}

7.5. Slide 14.

- Do not inference each other at all. Eg, BMW prices in USA vs weather
- X is blood pressure, Z is the disease, Y is BMW prices

7.6. Slide 17.

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

$$\Rightarrow \ln(\pi_{ijk}) = \underbrace{\ln(n)}_{\lambda} + \underbrace{\ln(\pi_{i+k})}_{\lambda^{XZ}} + \underbrace{\ln(\pi_{+jk})}_{\lambda^{YZ}} - \underbrace{\ln(\pi_{++k})}_{\lambda^{Z}}$$

Now, by the hierarchical principle, we need λ^X, λ^Y as well, which is given thanks to the generative class.

If $\ln\left(\theta_{ij(k)}\right)$ expanded does not depend on $k, C \Rightarrow$ homogeneous association.

7.7. Slide 25.

Minimal sufficient statistic: $\frac{f(y)}{f(y')}$ does not depend on the parameter $\Leftrightarrow T(y) = T(y')$
In order to find sufficient statistic, use factorization theorem:

$$\ln(L(\theta)) = g(T(y), \theta) + h(y) \Rightarrow T(y) \text{ is sufficient stat.}$$

8. Course Summary

1. Contingency Table
- Able to identify sampling themes and what conclusion we can draw from this
- Odds ratio:
  - Independence
  - Pairwise OR, local OR, conditional OR
- Partial Table, marginal Table
- Simpsons paradox (marginal $\neq$ conditional)
- Homogeneous association
  - Test homogeneous association
  - Breslow day Test
  - Modelling to test
- Inference of OR or conditional OR
- Conf. intervals for log-OR and their derivation
- Test independence:
  - Pearson χ^2, likelihood ratio, Fishers exact
- Ordinal data:
  - Concordant & discordant Pairwise
  - Goodman-Kruskals γ
  - Wilcoxon test

2. Logistic Regression
- Express model (link functions, logit link)
- Interpret β, log Ordinal
- Interpret R outputs, residual deviance, null deviance, test homogeneous association
- Test conditional independence
- Logistic regression for case-control study
- Meaning of link function: probit, cloglog
- Conditional MLE with large k (for binary data, other models aswell)

3. Multinomial Data
- Baseline category model
- Cummulative logit model/proportional Odds
- Test conditional independence
- Model-free test: CMH (Multinomial + binomial data)

4. Log-linear model
- Multinomial sampling vs Poisson sampling (dff. likelihood same β)
- Different types of independence
- Interpret R ouputs (estimate deviance, residual deviance, AIC)
- Relation between log-linear model and Logistic
- Generating class $\Leftrightarrow$ conditional independence graph
- Graphical model and chordal graph
- Decomposable model - express point probability using sufficient statistic
- Multigraph - maximum spannin tree + branch set to determine decomposability