Handout 13

STA 138

Homogeneous Association

If we have three discrete variables X, Y and Z at levels 2, 2 and K respectively, then we have a $2 \times 2 \times K$ contingence table. For any given level of $Z = k$, the odds ratio of the 2×2 table is called the conditional odds ratio and is denoted by $\theta_{XY(k)}$. We say that there is **Homogeneous Association** if

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

Conditional Independence of X and Y is a special case of **Homogeneous Association** where the value of the common odds is equal to 1, ie,

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

Consider the following $2 \times 2 \times K$ contingency table. There are $K = 3$ departments and we have a 2×2 table for each department. This is a part of the 1973 Berkeley Admissions data, and the full data were analyzed by Bickel, Hammel and O'Connell (1975). If we consider the data aggregated over all the departments, then the odds for admission ratio comparing males to females is about 1.9. However, the odds ratios in the three departments do not look substantially differ from 1. This is known as Simpson's paradox discussed in Handout 7.

Aggregated Data

	Accept	Reject
Male	495	763
Female	243	716
Odds ratio	1.91156	

The following table gives the detailed admission data with counts from $K = 3$ departments.

	Department A		Department B		Department C	
	Accept	Reject	Accept	Reject	Accept	Reject
Male	353	207	120	205	22	351
Female	17	8	202	391	24	317
Odds ratio	$\hat{\theta}_{XY(1)} = 0.80250$		$\hat{\theta}_{XY(2)} = 1.13306$		$\hat{\theta}_{XY(3)} = 0.82787$	

Note that for the aggregate data, the odds for acceptance of males are estimated to be 1.9 times the odds for acceptance for females. If we test for independence using the aggregate data, we have

$$G^2 = 48.8494, \; df = 1.$$

The p-value is less than 0.0001. From the aggregate data, we find that gender and acceptance are not independent. We now look at association between gender and acceptance for individual departments.

Let

X be the gender ($X = 1$ for males, $X = 2$ for females),

Y be the admission status ($Y = 1$ for accepted, $Y = 2$ for rejected), and

Z be the department ($Z = 1$ for Department A, $Z = 2$ for Department B, and $Z = 3$ for Department C).

Denote the conditional odds ratio for acceptance for males to females given $Z = k$ by $\theta_{XY(k)}$, ie,

$$\begin{aligned}\theta_{XY(k)} &= \frac{\text{odds for accept for a male applicant in department } k}{\text{odds for accept for a female applicant in department } k} \\ &= \frac{P(Y=1|X=1,Z=k)/[1-P(Y=1|X=1,Z=k)]}{P(Y=1|X=2,Z=k)/[1-P(Y=1|X=2,Z=k)]} \\ &= \frac{\pi_{11k}/\pi_{12k}}{\pi_{21k}/\pi_{22k}} = \frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}}.\end{aligned}$$

If n_{ijk} is the observed count for gender i, admission status j and department k. The total number applicants in the k^{th} department is n_{++k}. Here we have

$$n_{++1} = 585, n_{++2} = 918, n_{++3} = 714.$$

Recall that

$$\begin{aligned}\hat{\theta}_{XY(k)} &= \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}, \\ \hat{\theta}_{XY(1)} &= 0.80250,\ \hat{\theta}_{XY(2)} = 1.13306,\ \hat{\theta}_{XY(3)} = 0.82787.\end{aligned}$$

Estimation of the Common Odds Ratio Under Homogeneous Association

If we assume that the conditional odds are the same over the $K = 3$ departments, ie, $\theta_{XY(1)} = \cdots = \theta_{XY(K)}$, how do we estimate the common odds θ? For notational simplicity we will denote $\hat{\theta}_{XY(k)}$ by $\hat{\theta}_k$.

We will write down two estimates: one due to Cochran-Mantel-Haenszel (CMH) and the other is a so-called exponential estimate.

Cochran-Mantel-Haenzsel (CMH) estimate

The CMH estimate for the common odds ratio is

$$\begin{aligned}\hat{\theta}_{CMH} &= \frac{\sum_{k=1}^K n_{11k}n_{22k}/n_{++k}}{\sum_{k=1}^K n_{12k}n_{21k}/n_{++k}} = \sum_{k=1}^K w_k\hat{\theta}_k/w_+, \text{ with} \\ w_k &= n_{12k}n_{21k}/n_{++k} \text{ and } w_+ = \sum_{k=1}^K w_k.\end{aligned}$$

Thus $\hat{\theta}_{CMH}$ is a weighted average of $\{\hat{\theta}_k\}$. For our data,

$$\begin{aligned}w_1 &= \frac{(207)(17)}{585} = 6.01538 \\ w_2 &= \frac{(205)(202)}{918} = 45.10893 \\ w_3 &= \frac{(351)(24)}{714} = 11.79832,\ w_+ = w_1 + w_2 + w_3 = 62.92263, \\ \hat{\theta}_{CMH} &= \sum_{k=1}^K w_k\hat{\theta}_k/w_+ = 1.04433.\end{aligned}$$

In order to calculate a confidence interval for θ, we first obtain a confidence interval for $\log(\theta)$ and then exponentiate it. An approximate 95% confidence interval for $\log(\theta)$ is given by $\log(\hat{\theta}_{CMH}) \pm 1.96SE$, where

$$\begin{aligned} SE^2 &= \frac{1}{w_+^2}\sum_{k=1}^{K} w_k^2(1/n_{11k} + 1/n_{12k} + 1/n_{21k} + 1/n_{22k}) = 0.015673, \\ SE &= 0.12519, \end{aligned}$$

Thus a 95% confidence interval for $\log(\theta)$ is

$$\log(1.04433) \pm (1.96)(0.12519), \text{ ie, } 0.04338 \pm 0.24537, \text{ ie,}$$
$$(-0.20199, 0.28875).$$

Thus a 95% confidence interval for θ is $(e^{-0.20199}, e^{0.28875}) = (0.8171, 1.3348)$.

This tells us that we cannot reject $H_0 : \theta = 1$ in favor of $H_1 : \theta \neq 1$ at level $\alpha = 0.05$. Conclusion: we cannot reject the null hypothesis of conditional independence of gender and admission status (given department)

Exponential estimator

In this method, an estimator of $\log(\theta)$ is obtained first using a weighted average of $\{\log(\hat{\theta}_k)\}$, and then it is exponentiated in order to get estimate of the common odds ratio. The estimator is

$$\begin{aligned} \hat{\theta}_{\exp} &= \exp\left(\sum w_k^* \log(\hat{\theta}_k)/w_+^*\right), \text{ with} \\ w_k^* &= 1/[1/n_{11k} + 1/n_{12k} + 1/n_{21k} + 1/n_{22k}] \text{ and } w_+^* = \sum w_k^*. \end{aligned}$$

For our data,

$$\begin{aligned} w_1^* &= 5.2223, w_2^* = 48.2639, w_3^* = 10.7383, w_+^* = 64.2245, \\ \log(\hat{\theta}_{\exp}) &= [(5.2223)\log(0.8025) + (48.2639)\log(1.1331) + (10.7383)\log(0.8279)]/64.2245 \\ &= 0.0462, \\ \hat{\theta}_{\exp} &= \exp(0.0462) = 1.0473. \end{aligned}$$

Note that the values of $\hat{\theta}_{CMH}$ and $\hat{\theta}_{\exp}$ are quite close.

An approximate 95% confidence interval for $\log(\theta)$ is $\log(\hat{\theta}_{\exp}) \pm 1.96SE$, where $SE = 1/\sqrt{w_+^*}$. Thus an approximate 95% confidence interval for $\log(\theta)$ is

$$0.0462 \pm (1.96)(1/\sqrt{64.2245}), \text{ ie, } 0.0462 \pm 0.2446, \text{ ie,}$$
$$(-0.1984, 0.2908).$$

Thus an approximate 95% confidence interval for θ is $(e^{-0.1984}, e^{0.2908}) = (0.8200, 1.3374)$.

Note that the two confidence intervals for θ using $\hat{\theta}_{CMH}$ and $\hat{\theta}_{\exp}$ are quite similar.

Test for Conditional Independence.

Recall that conditional independence is a special case of homogeneous association. How does one test that X and Y are conditionally independent under the assumption of homogeneous association? Let θ be

the common value of $\theta_{XY(1)}, \ldots, \theta_{XY(K)}$, then conditional independence is equivalent to $\theta = 1$. We can test $H_0 : \theta = 1$ vs $H_1 : \theta \neq 1$ or a one-sided tests such as $H_0 : \theta = 1$ vs $H_1 : \theta > 1$.

Assuming homegeneous association, we wish to test if admission status and gender are conditionally independent given the department, ie, test $H_0 : \theta = 1$ vs $H_1 : \theta \neq 1$, which is equivalent to testing $H_0 : \log(\theta) = 0$ vs $H_1 : \log(\theta) \neq 0$.

If we use the CMH estimate of θ, then we can calculate the z-statistic $z = \log(\hat{\theta}_{CMH})/SE$, where SE is the standard error of $\log(\hat{\theta}_{CMH})$. Here

$$z = \frac{\log(1.04433)}{0.12519} = 0.3465.$$

The p-value is 2 times the area to the area of 0.3465 under the standard normal curve, and this area is about 0.729. Thus we conclude that we cannot reject the null. We cannot reject the hypothesis that gender and admission status are conditionally independent (assuming homegeneus association).

We can also use $\hat{\theta}_{\text{exp}}$ to carry out the test.

Test for Homogeneous Association.

The estimators $\hat{\theta}_{CMH}$ and $\hat{\theta}_{\text{exp}}$ for the common odds ratio θ were obtained under the assumption of homogeneous association, ie, $\theta_1 = \cdots = \theta_K$. How does one test if the assumption is valid. There are different tests for this: one involves $\hat{\theta}_{\text{exp}}$, and the others are Pearson's chi-square or likelihood ratio tests. Here we describe a test using $\hat{\theta}_{\text{exp}}$.

The null and the alternative hypotheses are:

$H_0 : \theta_1 = \cdots = \theta_K$, H_1 :not all $\theta_1, \ldots, \theta_K$ are the same.

We can use the following test statistic

$$X^2 = \sum_{k=1}^{K} w_k^* \left[\log(\hat{\theta}_k) - \log(\hat{\theta}_{\text{exp}})\right]^2.$$

Under H_0, X^2 has a chi-square distribution with df=$K - 1$. Here

$$\begin{aligned} X^2 &= (5.2223)[\log(0.8025) - \log(1.0473)]^2 \\ &\quad + \cdots + (10.7383)[\log(0.8279) - \log(1.0473)]^2 \\ &= 1.2044, \\ df &= K - 1 = 2, \end{aligned}$$

The p-value is area to the right of 1.2044 under the χ_2^2 curve and this area is 0.5476. We cannot reject H_0. Thus the null hypothesis of homogeneous association, ie, $\theta_1 = \cdots = \theta_K$, cannot be rejected.

Connection with logistic regression.

In order to see the connection with logistic regression, let us change the notation for the variable 'admission status'. Let $Y = 1$ for acceptance and $Y = 0$ for rejection, and $\pi_k = P(Y = 1|X, Z = k)$, and π'_k be the logit of π_k. Then the model for homogeneous association is

$$\pi'_k = \beta_0 + \beta_1 X + \gamma_k,$$

where $\{\gamma_k\}$ are the department effects. Identifiability condition require one constraint on $\{\gamma_k\}$, which can be $\sum \gamma_k = 0$ or $\gamma_1 = 0$. When the gender is X, the odds for acceptance are

$$\pi_k/(1-\pi_k) = \exp(\beta_0 + \beta_1 X + \gamma_k).$$

For any department k, the odds ratio for acceptance for males to females is

$$\frac{\text{odds for acceptance for males}}{\text{odds for acceptance for females}} = \frac{\exp(\beta_0 + \beta_1 1 + \gamma_k)}{\exp(\beta_0 + \beta_1 2 + \gamma_k)} = \exp(-\beta_1).$$

For nonhomogeneous association, the model for π'_k must also contain the interaction term between gender and department, in addition to the terms X and γ_k.

Thus the test for homogeneous association is equivalent to testing if the interaction term can be dropped from the model which contains X, γ_k and the interaction term. Note that the model which contains gender, department and the interaction term gender*department is the saturated model.

The testing can be done by a chi-square test known as Breslow-Day test (same as Pearson's chi-square) or the likelihood ratio test, or the test using the test described $\log(\hat{\theta}_{\exp})$ as described in the section "Exponential Estimator".

In order to carry out the likelihood ratio test, we fit the model without the interaction term. **Both gender and department have been declared as factors in carrying out the analysis**. The data is given in a tabular form below. Here is a summary of the R outputs.

Call:

glm(formula = cbind(accept, reject) ~gender + dept, family = "binomial")

Deviance	Residuals				
1	2	3	4	5	6
−0.1253	0.5970	0.4534	−0.3392	−0.5303	0.5464

Coefficients:	Estimate	Std.Error	z value	Pr(>|z|)
Intercept	0.50138	0.14693	3.412	0.000644^{***}
genderM	0.04334	0.12469	0.348	0.728136
deptB	−1.13248	0.13316	−8.504	$< 2e-16^{***}$
deptC	−3.19988	0.18275	−17.509	$< 2e-16^{***}$

—

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 522.0563 on 5 degrees of freedom

Residual deviance: 1.2725 on 2 degrees of freedom

AIC: 42.261

Number of Fisher Scoring iterations: 3

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx

Note that we are testing $H_0 : \pi'_k = \beta_0 + \beta_1 X + \gamma_k$ for all k, vs $H_1 : \pi'_k \neq \beta_0 + \beta_1 X + \gamma_k$ for some k.

Residual deviance= 1.2725, with $df = 2$.

The p-value is area to the right of 1.2725 under the χ^2_2 curve and this area is about 0.529.

Recall that Residual Deviance equals $G^2 = -2[\log(L_M) - \log(L_S)]$, where L_S is the likelihood under the saturated model and L_M is the likelihood under the model without the interaction term.

Berkeley Admission Data

Accept	Reject	Gender	Department
353	207	M	A
17	8	F	A
120	205	M	B
202	391	F	B
22	351	M	C
24	317	F	C