Categorical Data Analysis

Lecture Note 5

Instructor: Seokho Lee

Hankuk University of Foreign Studies

5. Three-Way Contingency Tables

In previous section we examined association between two categorical variables, X and Y. In many situations, there are other variables that can affect the relationship between the two variables of interest.

Example: A study was carried out to compare two treatments for a respiratory disorder. The goal was to compare the proportions of patients responding favorably to test and placebo. A confounding factor is that the study was carried out at two centers which had different patient populations. We wish to examine the association between treatment and response while adjusting for the effects of the centers.

		Respiratory Improvement		
Center	Treatment	Yes	No	Total
1	Test	29	16	45
	Placebo	14	31	45
	Total	43	47	90
2	Test	37	8	45
	Placebo	24	21	45
	Total	61	29	90

- In studying the effect of an explanatory variable X on a response variable Y, one should take into account other variables (covariates) that may influence the relationship. Otherwise, an observed effect of X on Y may simply reflect the associations of the covariates on X and Y.

- One strategy for examining the association between two variables while adjusting for the effect of others is *stratified analysis*.
  - In some cases the stratification can result from the study design. This is the case in a multicenter clinical trials.
  - In other cases, it may arise from a post-study stratification to control for the effects of certain explanatory variables. This is often used in observational studies where randomization cannot be used.

- Each two-way table corresponds to one stratum. The strata are determined by the unique combinations of the explanatory variables.

- The analysis of sets of tables addresses many of the same questions asked in the analysis of a single table:
  - Is there an association between rows and columns?
  - What is the strength of that association (odds ratio)?

  Here we are looking at the overall association instead of the association in just one table.

5.1. Partial Association

Multivariate Categorial Data: (X, Y, Z).

$$\begin{aligned} X &= \text{explanatory variable (predictor)} \\ Y &= \text{response} \\ Z &= \text{control variable (covariate)} \end{aligned}$$

5.1.1. Partial Tables and Marginal Table

- A **partial table** is a two-way cross-sectional table that classifies X and Y at the different levels of the control variable Z.
  - They display the relationship between X and Y at fixed levels of Z.
  - They control (remove) the effect of Z by holding its value constant.
- The X–Y **marginal table** is formed by combining the partial tables.
  - It contains no information about Z.
  - It ignores Z rather than controlling for it.

Example: The marginal table for the respiratory disease data follows:

	Respiratory Improvement		
Treatment	Yes	No	Total
Test	66	24	90
Placebo	38	52	90
Total	104	76	180

- The associations in partial tables are called *conditional associations*.
- The association in the marginal table is called the *marginal association*.
- One can describe these associations using odds ratios.

5.1.2. Conditional and Marginal Odds Ratios

Suppose we have $2 \times 2 \times K$ table. This corresponds to K different 2×2 tables, one for each of the K levels of Z,

- The *conditional odds ratios* are the odds ratios for the partial tables.
  - Let n_{ijk} denote the observed count of $X = i$, $Y = j$, and $Z = k$, and let $E[n_{ijk}] = e_{ijk}$.
  - The odds ratio and its estimate for the k^{th} partial table are
    $$\theta_{XY(k)} = \frac{e_{11k}e_{22k}}{e_{12k}e_{21k}}, \quad \text{and} \quad \hat{\theta}_{XY(k)} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}.$$
  - $\theta_{XY(k)}$ measure the conditional X–Y association when $Z = k$.
- The marginal odds ratio is the odds ratio for the marginal X–Y table.
  $$\theta_{XY} = \frac{e_{11+}e_{22+}}{e_{12+}e_{21+}}, \quad \text{and} \quad \hat{\theta}_{XY} = \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}}.$$

Example: Compute the conditional and marginal odds ratio for the respiratory disease data set.

$$\begin{aligned}
\hat{\theta}_{XY(1)} &= \frac{n_{111}n_{221}}{n_{121}n_{211}} = \frac{29 \cdot 31}{14 \cdot 16} = 4.013 \\
\hat{\theta}_{XY(2)} &= \frac{n_{112}n_{222}}{n_{122}n_{212}} = \frac{37 \cdot 21}{24 \cdot 8} = 4.047 \\
\hat{\theta}_{XY} &= \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}} = \frac{66 \cdot 52}{38 \cdot 24} = 3.763
\end{aligned}$$

5.1.3. Simpson’s Paradox

We will illustrate Simpson’s paradox with an example.

Example: Applicants can apply for either a sales position or an office position. For each position, each applicant was classified according to gender and according to whether the applicant was offered the position. The combined table follows:

	Result		
Gender	Offer	Deny	Total
Male	160	115	275
Female	160	165	325
Total	320	280	600

The resulting marginal odds ratio is

$$\hat{\theta}_{XY} = \frac{160 \cdot 165}{115 \cdot 160} = 1.435$$

The $2 \times 2 \times 2$ table follows:

		Result		
Type of Position	Gender	Offer	Deny	Total
Office	Male	15	60	75
	Female	75	150	225
Total		90	210	300
Sales	Male	145	55	200
	Female	85	15	100
Total		230	70	300

The resulting conditional odds ratios are

$$\begin{aligned} \hat{\theta}_{XY(1)} &= \frac{15 \cdot 150}{60 \cdot 75} = 0.5 \\ \hat{\theta}_{XY(2)} &= \frac{145 \cdot 15}{85 \cdot 55} = 0.465 \end{aligned}$$

Thus, the marginal association has a different direction than the conditional associations. This phenomenon is called *Simpson's paradox*.

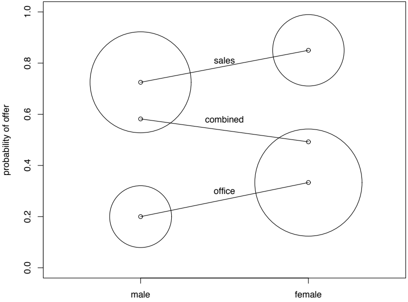
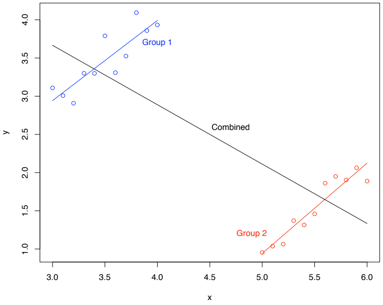


Illustration of Simpson's Paradox


Simpson's Paradox in Regression

5.1.4. Marginal versus Conditional Independence

- If X and Y are independent in each partial table, they are said to be *conditionally independent*. In this case,

$$\theta_{XY(k)} = 1, \quad \text{for all } k = 1, \cdots, K$$

- If X and Y are independent in the marginal table, they are said to be *marginally independent*. In this case,

$$\theta_{XY} = 1$$

- Conditional independence does not imply marginal independence. This is shown in the below table

		Response		
Clinic	Treatment	Success	Fail	
1	A	18	12	$\hat{\theta}_{XY(1)} = 1$
	B	12	8	
2	A	2	8	$\hat{\theta}_{XY(2)} = 1$
	B	8	32	
Total	A	20	20	$\hat{\theta}_{XY} = 2$
	B	20	40	

5.1.5. Homogeneous Association

- A $2 \times 2 \times K$ table has *homogeneous association* if

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

- Conditional independence is a special case of homogeneous association.

$$\theta_{XY(k)} = 1 \quad \text{for all } k = 1, 2, \cdots, K$$

- A general $I \times J \times K$ table has *homogeneous association* if any conditional odds ratio formed using two levels of X and two levels of Y is the same at any level of Z.
- Homogeneous association is a symmetric property. For variables X, Y, and Z, if there is a homogeneous X–Y association, there are also homogeneous X–Z and Y–Z associations.
- When homogeneous association occurs, there is no *interaction* between two variables in their effect on a third.

5.2. Cochran-Mantel-Haenszel Methods

We consider K sets of 2×2 tables where the k^{th} table, $k = 1, \cdots, K$, has counts

X\Y	1	2	Total
1	n_{11k}	n_{12k}	n_{1+k}
2	n_{21k}	n_{22k}	n_{2+k}
Total	n_{+1k}	n_{+2k}	n_{++k}

- We consider the null hypothesis that X and Y are conditionally independent; that is

$$H_0 \ : \ \theta_{XY(k)} = 1, \quad k = 1, \cdots, K$$

- Under the null hypothesis,

$$E[n_{11k}|H_0] \ = \ e_{11k} \ = \ \frac{n_{1+k}n_{+1k}}{n_{++k}}$$

and

$$Var(n_{11k}|H_0) \ = \ v_{11k} \ = \ \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_{++k}^2(n_{++k}-1)}$$

- the Cochran-Mantel-Haenszel statistic summarizes the information from the K partial tables:

$$CMH = Q_{CMH} = \frac{\left\{\sum_{k=1}^{K}(n_{11k} - e_{11k})\right\}^2}{\sum_{k=1}^{K} v_{11k}} = \frac{\left\{\sum_{k=1}^{K} \frac{n_{1+k}n_{2+k}}{n_{++k}}(p_{11k} - p_{21k})\right\}^2}{\sum_{k=1}^{K} v_{11k}}$$

  where $p_{i1k} = n_{i1k}/n_{i+k}$ is the proportion of subjects from the k^{th} stratum and i^{th} group to have a favorable response.

- When there is only one stratum ($K = 1$), the *CMH* statistic reduces to $(n-1)Q_p/n$ where Q_p is Pearson's chi-squared statistic. In this case, *CMH* equals the Mantel-Haenszel chi-squared statistic in `SAS PROC FREQ`.

- When there is more than one stratum, *CMH* becomes a stratum–adjusted chi-squared statistic.

- The *CMH* statistic is large when $(n_{11k} - e_{11k})$ is consistently positive or consistently negative for all the tables.
  - When the true odds ratio $\theta_{XY(k)} > 1$ in partial table k, we expect to observe $(n_{11k} - e_{11k}) > 0$.
  - When the true odds ratio $\theta_{XY(k)} > 1$ in every table k, the sum of these difference tends to be a relatively large positive number.
  - When the true odds ratio $\theta_{XY(k)} < 1$ in every table k, the sum of these difference tends to be a relatively large negative number.

- The *CMH* statistic has approximately a chi-squared distribution with $df = 1$ when H_0 is true. This approximation holds when the combined row sample sizes ($\sum_{k=1}^{K} n_{i+k} = n_{i++}$) are large enough ($> 30$).

- The *CMH* statistic is effective for detecting patterns across K strata when there is a strong tendency to expect the majority of the differences ($p_{11k} - p_{21k}$) to have the same sign.

    - The *CMH* statistic is often called an *averaged partial association statistic*.

    - *CMH* may fail to detect association when the differences are in opposite directions and of similar magnitude.

- The *CMH* statistic potentially removes the confounding effect of the explanatory variables that define the strata and can provide an increase in the power for detecting association. This strategy is similar to adjustment using blocks in the randomized block design.

- The *CMH* test works best when the X–Y association is similar in each partial table. It may not work well when the association varies dramatically among the partial tables.

    - Usually when there is an association, it is usually in the same direction across tables, although to a varying degree.

    - The *CMH* test has good power against the alternative of consistent patterns of association. It has low power for detecting association in opposite directions.

    - Regardless of power, the method always has the desired level of significance under H_0, so it is always a valid method.

    - Always examine the partial tables to determine if you have a situation in which the association is inconsistent and the *CMH* statistic is not very powerful.

5.2.1. Estimation of a Common Odds Ratio

In a $2 \times 2 \times K$ table, when the association seems stable across the K partial tables, we can estimate an assumed common value for the K true odds ratio.

- Recall that the odds ratio for the k^{th} partial table is estimated by

$$\hat{\theta}_k = \hat{\theta}_{XY(k)} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}$$

The standard error of $\log(\hat{\theta}_k)$ is estimated by

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

- Suppose $\theta_1 = \theta_2 = \cdots = \theta_K$. The *Mantel-Haenszel estimator* of the common odds ratio is

$$\hat{\theta}_{MH} = \frac{\sum_{k=1}^{K} \frac{n_{11k}n_{22k}}{n_{++k}}}{\sum_{k=1}^{K} \frac{n_{12k}n_{21k}}{n_{++k}}}$$

The formula for the standard error of $\log \hat{\theta}_{MH}$ is complicated.

- Another estimator for the common odds ratio is the *logit estimator*:

$$\hat{\theta}_L = \exp\left\{ \frac{\sum_{k=1}^K w_k \log \hat{\theta}_k}{\sum_{k=1}^K w_k} \right\}$$

  The weights are given by

$$w_k = \left(\frac{1}{n_{11k}} + \frac{1}{n_{12k}} + \frac{1}{n_{21k}} + \frac{1}{n_{22k}} \right)^{-1}$$

- A $100(1-\alpha)\%$ confidence interval for θ based on $\hat{\theta}_L$ is

$$\exp\left\{ \log \hat{\theta}_L \pm z_{\alpha/2} \left[\sum_{k=1}^K w_k \right]^{-1/2} \right\}$$

- The logit estimator is also a reasonable estimator, but it requires adequate sample sizes (all $n_{ijk} > 5$). The Mantel-Haenszel estimator is not as sensitive to sample size.

- When the counts are small and you want to find an exact confidence interval for the common odds ratio, you need to use logistic regression.

- If the true odds ratios are not identical but do not vary drastically (in a single direction), $\hat{\theta}_{MH}$ or $\hat{\theta}_L$ provide a useful summary of the K conditional associations.
- If the odds ratios are not homogeneous, then the common odds ratio should be viewed cautiously. One should emphasize the within-strata odds ratios.

5.2.2. Testing Homogeneity of the Odds Ratios

We consider testing the null hypothesis of equality of the odds ratios across the K strata:

$$H_0 \ : \ \theta_1 = \theta_2 = \cdots = \theta_K$$

The *Breslow-Day* test statistic has the Pearson's chi-squared form:

$$Q_{BD} = \sum_{k=1}^{K} \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(n_{ijk} - e_{ijk})^2}{e_{ijk}}$$

where e_{ijk} is the estimated cell count in the k^{th} partial table under the null hypothesis of equal odds ratios. It is calculated assuming

1. $\hat{\theta}_{MH}$ is the odds ratio for each stratum.
2. Each partial table has the same marginal totals as the observed tables.

- The Breslow-Day statistic is approximately chi-squared with $df = K - 1$. The sample sizes need to be relatively large in each partial table with $e_{ijk} > 5$ in at least 80% of the cells. See Agresti, p.64, for an adjustment that improves the chi-squared approximation.

- The *CMH* and Breslow-Day tests are large sample tests. Agresti discusses exact tests in Section 3.3.