Local Odds Ratio Estimation for Stratified Contingency Tables with Multiple Responses

Thomas Suesse[a], Ivy Liu[b]

[a]*School of Mathematics and Applied Statistics, University of Wollongong, Australia*

[b]*School of Mathematics, Statistics and Operations Research, Victoria University of Wellington, New Zealand*

Abstract

For a two–way contingency table with categorical variables, local odds ratios are commonly used to describe the relationships between the row and column variables. An ordinary case has mutually exclusive cell counts, i.e., each subject must fit into one and only one cell. However, many surveys have a situation where respondents may select more than one outcome category. We discuss the maximum likelihood method and suggest the Mantel–Haenszel local odds ratio estimation for K such $2 \times c$ tables, treating the multiple responses as an extension of the multinomial sampling model. We derive new dually consistent (co)variance estimators and show their performance with a simulation study.

Key words: Consistency, Local odds ratio, Mantel–Haenszel estimator, Odds ratio, Multiple responses

1. Introduction

Many studies are designed to compare groups on a multi–level response variable. One often uses a two-way contingency table that cross–classifies subjects on both group and response variables to display relationships between them. A set of odds ratios, such as *local odds ratios* [1, p.55] that use four cells in adjacent rows and columns, can describe the associations. If a study attempts to control for other factors that might influence the relationships, a three-way contingency table can show the associations between the group and response variables controlling for a possibly confounding variable. The three–way contingency table consists of K 2-way partial tables, where K is the number of levels for the control variable.

For a simple case of K 2×2 tables, let $\pi_{j|ik}$ be the probability of selecting item $j = 1, 2$ for a subject in row $i = 1, 2$ and stratum $k = 1, \ldots, K$. The kth odds ratio is defined as $\Psi^k = (\pi_{1|1k}\pi_{2|2k})/(\pi_{1|2k}\pi_{2|1k})$. The Mantel–Haenszel (MH) [23] estimator is a popular way of summarizing a common odds ratio when the conditional association is assumed to remain the same given the control variable, i.e., $\Psi^1 = \cdots = \Psi^K$. It is used not only when the common odds ratio assumption

seems plausible, but also as a summary measure when the association varies only mildly across the tables. Greenland [13] extended the MH method to the common local odds ratios for K $2 \times c$ tables, where the response variable has c categories. The local odds ratios in stratum k have the form

$$\Psi_j^k = \frac{\pi_{j|1k}\pi_{(j+1)|2k}}{\pi_{j|2k}\pi_{(j+1)|1k}}, \quad j = 1, \ldots, c-1. \tag{1}$$

The assumption of common local odds ratios states $\Psi_j = \Psi_j^1 = \ldots = \Psi_j^K$ ($j = 1, \ldots, c-1$). The MH estimators are dually consistent, i.e. consistent under the large–stratum (K is bounded while the number of subjects per stratum goes to infinity) and sparse–data (K goes to infinity with sample size, but the number of subjects per stratum remains fixed) limiting models. It is efficient under the null of no association.

The cell counts in the contingency table described above are mutually exclusive, i.e., each of the subjects must fit into one and only one cell. Some of the sampling models satisfy this condition. For instance, Greenland [13] assumed for each stratum the following sampling situations: c independent binomials and two independent rows of multinomials with c outcome categories. However, the mutually exclusive property might not hold in a 3-way table but, the conditional associations between the group and response variables are still the main interest of the study. This situation occurs often in a survey when respondents may select any number out of *c* outcome categories. For instance, the respondents are often told to "mark all that apply". Categorical outcome variables for this type of data are called *pick any/c variables* or *multiple response variables*, where c is the number of outcome categories (called *items*) and "/" stands for "out of" [10].

Table 1 shows an example of this type of data, where students of a statistics course at the Victoria University of Wellington in New Zealand were asked to tick their favourite bar. The study recorded the features of the bars separately, treating each feature as an item. Each bar may have more than one feature. Table 1 lists $c = 3$ items: "drink deals" (item 1), "pool table" (item 2) and "sports tv" (item 3). We assign a positive response for item j (e.g. "drink deals"), when the student's favourite bar has that particular feature. Each student also answered some personal questions (such as major, gender, working status, smoking status, etc.). For this example, we are interested to find the association between working status and preferred features of the bars, controlling on students' majors. Let $Y_j = 1$ if a student's selected bar has feature j ($j = 1, 2, 3$) and let $Y_j = 0$ otherwise. Let $\mathbf{Y} = (Y_1, Y_2, Y_3)'$ denote the response profile on 3 categories with (Y_1, Y_2, Y_3) corresponding to the (yes, no) outcome of the selected bar features. For example, if a student's selected bar has feature "drink deals" only, $\mathbf{Y} = (1, 0, 0)'$. The table displays both the $2 \times 3 \times 6$ "marginal" contingency table (for 6 different majors) and the $2 \times 2^3 \times 6$ "complete" contingency table on the multiple response profile. The marginal table shows the response counts for the features, by cross–classifying students according to their working status and major. For instance, within the Marine Biology major, 4 students who had work selected bars with the "sports tv" feature. The complete table shows the counts of the possible profiles $\mathbf{Y}$ for each combination of work levels and majors.

The analysis of this type of data has received much attention since Loughin and

Scherer [22]. They proposed a weighted chi–square test and a bootstrap test for the hypothesis that the probability of selecting any given item is identical among levels of a predictor variable. A series of work by Decady and Thomas [11], Bilder et al. [9], and Bilder and Loughin [5, 6] focused on tests of various hypotheses for a single multiple response variable. Later on, Thomas and Decady [27] and Bilder and Loughin [7] considered the tests of independence between two multiple response variables cases. Besides the tests, Agresti and Liu [3, 2] discussed different strategies for modeling multiple response data. Bilder and Loughin [8] extended their earlier work to simultaneously model and estimate the association structure between two multiple response variables in complex survey sampling situations. In addition to the modeling and testing procedures, Liu and Suesse [21] derived a closed form of the odds ratio estimation by comparing the odds of each of the items being selected for different groups. The purpose of this article is to show how one can use the simple concept of the local odds ratios (1) and extend their inferences for multiple responses to describe the associations in the marginal $2 \times c \times K$ table.

In the next two sections, we introduce the maximum likelihood (ML) method and propose the new MH estimation for the multiple response case. In Section 4, we illustrate methods using two examples. Section 5 shows the performance of our new estimators in a simulation study. The paper finishes with comments and discussions.

2. The ML Method

We can express the model which assumes common local odds ratios for all strata as

$$\log\left(\frac{\pi_{j|1k}\pi_{j+1|2k}}{\pi_{j|2k}\pi_{j+1|1k}}\right) = \beta_j, \quad \text{for all } k = 1, \ldots, K, \tag{2}$$

where $\beta_j = \log \Psi_j$. This model is not a standard logit model.

The ML inference for the model requires that the cell probabilities of the complete table are estimated under the constraints imposed by the model. Assume that cell counts in each row of the complete table follow a multinomial distribution. The likelihood function refers to the multinomial cell probabilities $\{p_{ijk},\, i = 1, 2,\, j = 1, \ldots, 2^c,\, k = 1, \ldots, K\}$, but the model itself applies to the marginal probability $\{\pi_{j|ik},\, i = 1, 2,\, j = 1, \ldots, c-1,\, k = 1, \ldots, K\}$. Haber [14] and Lang and Agresti [18] presented numerical algorithms for maximizing multinomial likelihoods subject to constraints for generalized loglinear models having the matrix form

$$\mathbf{C} \log \mathbf{A}\mathbf{p} = \mathbf{X}\boldsymbol{\beta}, \tag{3}$$

where $\mathbf{p}$ refers to the vector of multinomial cell probabilities. The model (2) has the above form, where the matrix $\mathbf{A}$ contains 0 and 1 entries in such a pattern that when applied to $\mathbf{p}$ it forms the relevant marginal probabilities $\pi_{j|ik}$; the matrix $\mathbf{C}$ contains 0, 1, and -1 entries in such a pattern that when applied to the log marginal probabilities, it forms the log local odds ratios for Model (2); $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{c-1})'$ and $\mathbf{X}$ is simply a row vector with 1's.

Table 1: Marginal Table and Complete Information for the Bar Data

Marginal Table					Complete Table								
					Item			**Y**					
					1	0	0	0	0	1	1	1	1
	Item			total	2	0	0	1	1	0	0	1	1
major	1	2	3	students	3	0	1	0	1	0	1	0	1
Biology													
work													
yes	1	0	0	1		0	0	0	0	1	0	0	0
no	1	1	0	1		0	0	0	0	0	0	1	0
Marine Biology													
work													
yes	11	11	4	13		0	0	2	0	2	0	5	4
no	1	1	1	1		0	0	0	0	0	0	0	1
Ecology & Biodiversity													
work													
yes	7	8	4	10		1	0	2	0	1	0	2	4
no	3	4	5	6		1	1	0	1	0	0	0	3
Operations Research													
work													
yes	0	0	1	1		0	1	0	0	0	0	0	0
no	0	0	0	1		1	0	0	0	0	0	0	0
Psychology													
work													
yes	2	2	2	3		0	1	0	0	0	0	1	1
no	2	2	1	3		0	0	1	0	1	0	0	1
Statistics													
work													
yes	1	1	1	1		0	0	0	0	0	0	0	1
no	1	1	1	1		0	0	0	0	0	0	0	1

Items 1: drink deals; 2: pool table; and 3: sports tv.

A disadvantage of the ML approach is that it is only consistent under the large–stratum limiting model. Under a sparse–data limiting model, a simulation study given by [20] showed that the ML estimator of a common global odds ratio tends to over–estimate the true odds ratio. The common global odds ratio refers to K stratified $r \times c$ tables with multinomial sampling. Furthermore, for multiple responses the ML estimator of the common local odds ratio is not consistent because the number of parameters $K \times 2 \times (2^c - 1)$ that determine the multinomial distributions for all strata goes to infinity as $K \to \infty$.

An R function (mph.Rcode.R) for the algorithm may be obtained from Prof J. B. Lang of the Statistics Department, University of Iowa `http://www.stat.uiowa.edu/~jblang/`. The function is not only suitable for fitting generalized loglinear models (3), but also provides the algorithm for fitting Multinomial Poisson homogeneous (MPH) models [16]. Lang [17] considers homogeneous linear predictor models, a subclass of MPH models. Such models allow a broader class of link functions; for details of these models see [16, 17]. Bergsma et al. [4] proposed another fitting algorithm directly built on the work of [18, 15] and provided an R package called "cmm" for fitting such models. Their program is a modification of the Lang–Agresti algorithm.

3. The MH method

Under the common local odds ratio assumption, the MH estimator given by Greenland [13] is appropriate to summarize the conditional relationship. However, the variance estimator for the MH estimator proposed by Greenland [13] is not dually consistent anymore, but only consistent under the large-stratum limiting model. This article will propose new dually consistent variance and covariance estimators.

To be general, we consider the odds ratios for any two items (say, j and h) as follows:

$$\Psi_{jh}^k = \frac{\pi_{j|1k}\pi_{h|2k}}{\pi_{j|2k}\pi_{h|1k}}, \quad j < h = 2, \ldots, c,$$

where $\Psi_{jh}^k = \Psi_j^k \times \cdots \times \Psi_{(h-1)}^k$ if $h > j+1$ and $\Psi_{jh}^k = \Psi_j^k$ if $h = j+1$. For a $2 \times c \times K$ table, let $X_{j|ik}$ denote the number of subjects that choose item $j = 1, \ldots, c$ in row $i = 1, 2$ and stratum $k = 1, \ldots, K$. Also, let n_{ik} denote the number of subjects in row i and stratum k and let $N_k = n_{1k} + n_{2k}$ denote the totals for stratum k. For each stratum, there are two independent rows of c multiple response outcome categories. Under the common odds ratio assumption: $\Psi_{jh} = \Psi_{jh}^1 = \cdots = \Psi_{jh}^K$, the ordinary Mantel-Haenszel estimator $\hat{\Psi}_{jh}$ has the following form

$$\hat{\Psi}_{jh} = \frac{C_{jh}}{C_{hj}},$$

where $C_{jh} = \sum_{k=1}^K c_{jh|k}$ with $c_{jh|k} = X_{j|1k}X_{h|2k}/N_k$. The ordinary MH estimator $\hat{\Psi}_{jh}$ is dually consistent, i.e. consistent under the large–sample limiting model and the sparse–data limiting model, see Appendix A for the proof.

Let $L_{jh} = \log \hat{\Psi}_{jh}$. Greenland [13] proposed the following variance estimator U^{old}_{jhh} for Var(L_{jh}) and the following covariance estimator U^{old}_{jhs} for Cov(L_{jh}, L_{js}):

$$U^{old}_{jhh} := \frac{\sum_k c_{jh|k} d_{jh|k}}{2C^2_{jh}} + \frac{\sum_k c_{hj|k} d_{hj|k}}{2C^2_{hj}} + \frac{\sum_k c_{jh|k} d_{hj|k} + c_{hj|k} d_{jh|k}}{2C_{jh}C_{hj}}$$

$$U^{old}_{jhs} := \frac{\sum_k X_{j|1k} X_{h|2k} X_{s|2k}/N^2_k}{3C_{jh}C_{js}} + \frac{\sum_k X_{j|+k} X_{h|2k} X_{s|1k}/N^2_k}{3C_{jh}C_{sj}} + \frac{\sum_k X_{j|+k} X_{h|1k} X_{s|2k}/N^2_k}{3C_{hj}C_{js}} + \frac{\sum_k X_{j|2k} X_{h|1k} X_{s|1k}/N^2_k}{3C_{hj}C_{sj}}$$

with $d_{jh|k} := (X_{j|1k} + X_{h|2k})/N_k$, and $X_{j|+k} = \sum_i X_{j|ik}$. These estimators are dually consistent when there are two independent rows of multinomials in a contingency table where only one of the c outcome categories can be selected by each subject. There is no estimator for Cov(L_{jh}, L_{ts}) with $t \neq j$ and $s \neq h$, because Cov$(L_{jh}, L_{ts}) = 0$.

For the multiple response data, we need the complete information of the response profile on the c items (e.g. the right side of Table 1) to estimate the variance and covariance for $\{L_{jh}, j < h = 1, \ldots c\}$ because the row cell counts in the marginal contingency table do not follow a multinomial distribution anymore. However, multiple response data can be considered as an extension of multinomial data, since choosing exactly one category is obviously a sub–case of choosing any number of categories.

Let the pairwise probabilities for items j and h ($j, h \in \{1, \ldots, c\}$) be $\pi^{ab}_{jh|ik}$ with $a, b \in \{0, 1\}$, where $(0, 1)$ is the (no, yes) outcome for the selection of each item. Then $\pi^{ab}_{jh|ik}$ is the probability of observing the pairwise outcome (a, b) for items j and h. For instance, the notation $\pi^{11}_{jh|ik}$ represents the probability that a subject, who is in row i and stratum k, selects both items j and h. Similarly, define the pairwise observations as $\{X^{ab}_{jh|ik}\}$. The pairwise probabilities can be computed from the 2^c joint probabilities referring to 2^c response profiles. In a similar manner, the pairwise observations (e.g. Table 2 for the first three majors) can be obtained from the complete table. For instance, to obtain $X^{11}_{jh|ik}$ we sum over all those joint observations for which responses for items j and h are both positive.

We assume $\mathbf{X}_{jh|ik} = (X^{00}_{jh|ik}, X^{01}_{jh|ik}, X^{10}_{jh|ik}, X^{11}_{jh|ik})$ follows a multinomial distribution with parameters n_{ik} and $\boldsymbol{\pi}_{jh|ik} = (\pi^{00}_{jh|ik}, \pi^{01}_{jh|ik}, \pi^{10}_{jh|ik}, \pi^{11}_{jh|ik})$ with $\pi^{00}_{jh|ik} + \pi^{01}_{jh|ik} + \pi^{10}_{jh|ik} + \pi^{11}_{jh|ik} = 1$. The marginal probabilities can be computed from the pairwise probabilities by $\pi_{j|ik} = \pi^{10}_{jh|ik} + \pi^{11}_{jh|ik}$ and $\pi_{h|ik} = \pi^{01}_{jh|ik} + \pi^{11}_{jh|ik}$. We can now show that

$$\mathbb{E} X_{j|ik} X_{h|ik} = n_{ik} n'_{ik} \pi_{j|ik} \pi_{h|ik} + n_{ik} \pi^{11}_{jh|ik} \tag{4}$$

Table 2: The Pairwise Observations for the Bar Data

Major	Pairwise Responses	Pair of Items 12	13	23	Total students
Biology					
Work					
Yes	Yes/Yes	0	0	0	
	Yes/No	1	1	0	
	No/Yes	0	0	0	
	No/No	0	0	1	1
No	Yes/Yes	1	0	0	
	Yes/No	0	1	1	
	No/Yes	0	0	0	
	No/No	0	0	0	1
Marine Biology					
Work					
Yes	Yes/Yes	9	4	4	
	Yes/No	2	7	7	
	No/Yes	2	0	0	
	No/No	0	2	2	13
No	Yes/Yes	1	1	1	
	Yes/No	0	0	0	
	No/Yes	0	0	0	
	No/No	0	0	0	1
Ecology & Biodiversity					
Work					
Yes	Yes/Yes	6	4	4	
	Yes/No	1	3	4	
	No/Yes	2	0	0	
	No/No	1	3	2	10
No	Yes/Yes	3	3	4	
	Yes/No	0	0	0	
	No/Yes	1	2	1	
	No/No	2	1	1	6

with $n'_{ik} = n_{ik} - 1$. If each subject can only choose one outcome category, following the multinomial samplings, we have $\text{Cov}(X_{j|ik}, X_{h|ik}) = -n_{ik}^2 \pi_{j|ik}\pi_{h|ik}$ and $\mathbb{E}X_{j|ik}X_{h|ik} = n_{ik}n'_{ik}\pi_{j|ik}\pi_{h|ik}$. So, the multinomial case is a special case of multiple responses. In Appendix B we use these results to present a sketch of the proof that the new estimators U_{jhh} for $\text{Var}(L_{jh})$, U_{jhs} for $\text{Cov}(L_{jh}, L_{js})$ and U_{jhts} for $\text{Cov}(L_{jh}, L_{ts})$ are dually consistent for multiple response data. For convenience, denote $X^{11}_{jh|ik}$ by $X_{jh|ik}$. The estimators U_{jhh}, U_{jhs} are defined as follows:

$$\begin{aligned} U_{jhh} &:= \widehat{\text{Var}}(L_{jh}) = U^{old}_{jhh} + U^{add}_{jhh} \\ U_{jhs} &:= \widehat{\text{Cov}}(L_{jh}, L_{js}) = U^{old}_{jhs} + U^{add}_{jhs}, \end{aligned} \tag{5}$$

where the additional terms U^{add} are given by

$$\begin{aligned} U^{add}_{jhh} = &- 4\frac{\sum_k X_{j|1k}X_{h|1k}X_{jh|2k}/N_k^2 + \sum_k X_{jh|1k}X_{j|2k}X_{h|2k}/N_k^2}{2C_{jh}C_{hj}} \\ &- \frac{\sum_k \{X_{jh|1k}(X_{j|2k} + X_{h|2k}) + X_{jh|2k}(X_{j|1k} + X_{h|1k})\}/N_k^2}{2C_{jh}C_{hj}} \\ &+ 4\frac{\sum_k X_{jh|2k}X_{jh|1k}/N_k^2}{2C_{jh}C_{hj}} \end{aligned}$$

and

$$\begin{aligned} U^{add}_{jhs} = &\frac{\hat{V}^A_{jhs|12}}{C_{jh}C_{js}} - \frac{\hat{V}_{jh,js}}{C_{hj}C_{js}} - \frac{\hat{V}_{js,jh}}{C_{jh}C_{sj}} + \frac{\hat{V}^A_{jhs|21}}{C_{hj}C_{sj}} \\ &+ \frac{\hat{V}^B_{jhs|12}}{3C_{jh}C_{js}} + \frac{\hat{V}^B_{hjs|12} + \hat{V}^B_{sjh|21}}{3C_{hj}C_{js}} + \frac{\hat{V}^B_{sjh|12} + \hat{V}^B_{hjs|21}}{3C_{jh}C_{sj}} + \frac{\hat{V}^B_{jhs|21}}{3C_{hj}C_{sj}} \end{aligned}$$

with

$$\hat{v}^A_{jhs|ijk} = \frac{1}{N_k^2}X^2_{j|ik}X_{hs|jk}, \qquad \hat{v}^B_{jhs|ijk} = -\frac{1}{N_k^2}X_{j|ik}X_{hs|jk}$$

$$\hat{v}_{jh,ts|k} = \frac{1}{N_k^2}\{X_{j|1k}X_{h|1k}X_{ts|2k} + X_{jh|1k}X_{t|2k}X_{s|2k} - X_{jh|1k}X_{ts|2k}\}$$

and $\hat{V}$ representing $\sum_k \hat{v}_k$. The estimator $U_{jhts} := \widehat{\text{Cov}}(L_{jh}, L_{ts})$ is given by

$$U_{jhts} := \frac{\hat{V}_{jt,hs}}{C_{jh}C_{ts}} - \frac{\hat{V}_{ht,js}}{C_{hj}C_{ts}} - \frac{\hat{V}_{js,ht}}{C_{jh}C_{st}} + \frac{\hat{V}_{hs,jt}}{C_{hj}C_{st}}.$$

When each subject can only choose one outcome category, the pairwise observations $X_{jh|ik}$ are all zero, because it is impossible to have both items chosen. Consequently, $U^{add}_{jhh} = U^{add}_{jhs} = U_{jhts} = 0$, such that $U_{jhh} \equiv U^{old}_{jhh}$ and $U_{jhs} \equiv U^{old}_{jhs}$.

This shows that our estimators are generalizations of Greenland's estimators and are also applicable for the multinomial sampling model in an ordinary case with only one response outcome for each subject.

4. Examples

We consider the data in Table 1. The stratification variable is the students' major and the row variable is the students' working status. Students are considered as working, if they either work part–time or full–time. The left side of Table 1 shows the positive responses for each item, and the right side shows the complete data.

Applying the MH method, we obtain $\{L_{12}, L_{13}, L_{23}\} = \{0.14,\ 0.60,\ 0.46\}$ with standard errors $\{0.24,\ 0.30,\ 0.28\}$ using the proposed new estimates (5). The bootstrap method gives standard errors $\{0.31,\ 0.38,\ 0.35\}$. The ML method fails to converge for this example due to the sparseness of the data under both Lang's algorithm and Bergsma's modified algorithm. When comparing the working effects for "drink deals" and "sports tv", the odds of choosing a favourite bar offering "drink deals" rather than "sports tv" for a student with part/full–time jobs are $\exp(0.60) = 1.83$ times the odds for a student without a job. The feature of "sports tv" is not as important as "drink deals" for students having work. This effect is significant when the new variance estimators are applied, but not detected when the numerical bootstrap method is used. In the next section, we will show that these new estimators have a better performance than the old estimators and the bootstrap variance. Unlike the ML method, our proposed new estimators have a closed form and can be obtained even for a highly sparse dataset.

This article also considers another less sparse dataset to explore the difference between the ML and MH methods. Table 3 given by Bilder and Loughin [6] presents data where 239 sexually active college women were asked "What type of contraceptives have you used?". They could select any answer from the following: A-oral, B-condom, C-lubricated condom, D-spermicide, and E-diaphragm. The table contains information on selected items and whether or not a subject had a prior history of urinary tract infection (UTI). The stratification variable is age. The complete table is given in the original article [6]. For demonstration, we exclude item E due to zero cell counts in order to avoid adding a small count to implement both MH and ML methods. The MH approach gives $\{L_{AB}, L_{AC}, L_{AD}, L_{BC}, L_{BD}, L_{CD}\}$=$\{0.28$, -0.43, -0.45,-0.70, -0.73, $-0.02\}$ with standard errors $\{0.21,\ 0.25,\ 0.29,\ 0.13,$ $0.20,\ 0.21\}$ by applying formula (5). The bootstrap standard errors are $\{0.21,\ 0.25,$ $0.30,\ 0.14,\ 0.21,\ 0.22\}$. The ML approach gives estimates $\{0.28,\ -0.39,\ -0.46,$ $-0.67,\ -0.73,\ -0.07\}$ with standard errors $\{0.26,\ 0.25,\ 0.29,\ 0.13,\ 0.21,\ 0.21\}$. The results from both MH and ML methods are very similar for the these non–sparse data.

5. Simulation Study

We conduct a simulation study to investigate the performance of the proposed estimators U_{jhh} and U_{jhs}. For simplicity, we choose $c = 3$, so that it is possible

Table 3: The Marginal UTI Data

	Contraceptive					Total	Total
	A	B	C	D	E	responses	women
Age ≥ 24							
UTI							
No	18	9	8	7	0	42	24
Yes	8	9	2	3	2	24	14
Age < 24							
UTI							
No	55	41	37	27	0	160	85
Yes	75	68	33	22	5	203	116

to obtain both estimators U_{jhh} and U_{jhs}. For given $\{\Psi_{1h},\ h = 2, \ldots, c\}$, we fix the marginal probabilities of the first row by setting $\pi_{j|1k} = 0.50$ for all $j = 1, \ldots, c$. Then we set $\pi_{1|2k} = 1/(1 + \Psi_{1h})$ and $\pi_{h|2k} = \frac{\Psi_{1h}}{1+\Psi_{1h}}$ for $h = 2, \ldots, c$. This ensures that the probabilities of the second row are balanced around $1/2$, for example $\Psi_{12} = 1$ gives $\pi_{1|2k} = \pi_{2|2k} = 1/2$. We also set $\Psi = \Psi_{12} = \Psi_{13}$ and $N_k = N_1 = \cdots = N_K$ to ensure simplicity.

We define the pairwise dependency between items j and h in the form of an odds ratio $\theta_{jh|ik}$, following Bilder and Loughin [6]:

$$\theta_{jh|ik} = \frac{P(Y_j = 1, Y_h = 1|ik)P(Y_j = 0, Y_h = 0|ik)}{P(Y_j = 0, Y_h = 1|ik)P(Y_j = 1, Y_h = 0|ik)}.$$

From the marginal probabilities $\{\pi_{j|ik}, j = 1, \ldots, c\}$ and the odds ratios $\{\theta_{jh|ik}, j \neq h = 1, \ldots, c\}$, we can compute the unique set of pairwise probabilities $\{\boldsymbol{\pi}_{jh|ik}, j \neq h = 1, \ldots, c\}$. Then the 2^c joint probabilities $\{P(Y_1 = a_1, \ldots, Y_c = a_c|ik), a_j = 0, 1, j = 1, \ldots, c\}$ in the complete table (e.g. right side of Table 1) can be computed from the probabilities $\{\pi_{j|ik}, j = 1, \ldots, c\}$ and $\{\boldsymbol{\pi}_{jh|ik}, j \neq h = 1, \ldots, c\}$, if a feasible solution exists [19].

There are several approaches to computing such a solution of the joint probabilities for given pairwise and marginal probabilities. One approach is to use linear programming. Another is applying the iterative proportional fitting (IPF) algorithm as described by Gange [12]. The generation of the joint probabilities subject to $\{\pi_{j|ik}, j = 1, \ldots, c\}$ and $\{\theta_{jh|ik}, j \neq h = 1, \ldots, c\}$ is analogous to the one applied in the simulation study by Bilder et al. [9]. We prefer IPF over linear programming because it generates strictly positive (> 0) joint probabilities (assuming such a solution exists), in contrast to linear programming, which might produce zero joint probabilities. Consequently IPF does not exclude any of the 2^c outcomes of the joint distribution. For simplicity, we also assume a constant association $\theta = \theta_{jh|ik}$ for all items $j \neq h = 1, \ldots, c$, rows $i = 1, 2$ and strata

$k = 1, \ldots, K$.

For the simulation scheme, we include the sampling model of two independent rows of multinomials with c outcome categories to create a special case of multiple response data. Setting the covariance between two items to $\text{Cov}(Y_j, Y_h) = -\pi_{j|ik}\pi_{h|ik}$ yields $\pi_{jh|ik} = P(Y_j = 1, Y_h = 1|ik) = 0$ and consequently $\theta_{jh|ik} = 0$. Therefore, fixing the covariance in such a way for all pairs of items yields the multinomial distribution. Under $\theta = \theta_{jh|ik} = 0$, we sample from the multinomial distribution with probabilities $\{\pi_{1|ik}, \ldots, \pi_{c|ik}\}$ for row $i = 1, \ldots, r$ and stratum $k = 1, \ldots, K$, in which the probabilities need to satisfy the condition $\sum_{j=1}^{c} \pi_{j|ik} = 1$. Since the setup of $\pi_{j|1k} = 0.50$ for all $j = 1, \ldots, c$ does not meet this condition, for the multinomial case we set $\pi_{j|1k} = 1/c$, $\pi_{1|2k} = 1/[(c-1)\Psi + 1]$ and $\pi_{j|2k} = \Psi\pi_{1|2k}$ for $j \geq 2$. For both rows ($i = 1, 2$), $\sum_{j=1}^{c} \pi_{j|ik} = 1$ with $c = 3$.

The number of bootstrap simulations was chosen as $B = 400$ and the number of simulated datasets was $20,000$. We record the mean and m.s.e. (mean squared error) of the newly proposed (co)variance estimators (U), the "old" (co)variance estimators proposed by Greenland [13] based on multinomial sampling (U^{old}), and the bootstrap estimate of (co)variance. The empirical variance and covariance of the L's over all simulations are regarded as the "true" (co)variances. The number of simulations for which the MH estimates are undefined (NA) is also recorded. The simulation results are based only on those data sets for which the MH estimates are finite.

Table 4 shows the simulation results of the variance estimators for various scenarios. The newly proposed estimators, U_{122} and U_{123}, perform better than the bootstrap estimates of (co)variance except for $K = 20$ and $N_k = 5$. They are also superior to U_{122}^{old} and U_{123}^{old} for $\theta > 0$. For $\theta = 0$ (multinomial situations), U and U^{old} are identical, because $U^{add} = 0$ due to the impossible event of observing the pairwise observation $(1, 1)$. Furthermore, the larger θ is, the larger the difference between U_{jhs}^{old} and U_{jhs} becomes. Generally, U^{old} cannot be recommended for multiple responses ($\theta > 0$), because the U^{old}'s are severely biased. Under $\theta = 0$ for which each respondent can only select one outcome category, the old and new estimators are identical. Overall the newly proposed (co)variance estimators U_{jhh} and U_{jhs} perform very well for the general case of multiple responses in various levels of association between items. We assume that U_{jhts} behaves similarly to U_{jhs} and U_{jhh}, due to the similar construction of the estimator.

6. Discussion

The article proposes an extension of the sampling model of two independent rows of multinomial responses to that of two independent rows of multiple response data per stratum. For surveys, it is very common to tick all that apply and not only one that applies. The cell counts in a 3-way contingency table are not necessarily mutually exclusive across response items. Greenland [13] proposed the MH estimators and their (co)variance estimators to summarize the conditional

Table 4: Simulation results for the variance and covariance estimators of the log odds ratio estimators when the true odds ratio $\Psi = 4$

				$\text{Var}(L_{12}), \text{Cov}(L_{12}, L_{13})$ Estimates			
				Empirical	New Proposed (U)	Greenland's (U^{old})	Bootstrap
K	N_k	θ	NA	100×mean	100×mean	100×mean	100×mean
					(10000×m.s.e.)	(10000×m.s.e.)	(10000×m.s.e.)
1	500	0	0	7.219, 5.062	7.056, 4.936	7.056, 4.936	7.338, 5.189
					(0.552, 0.568)	(0.552, 0.568)	(2.704, 3.616)
1	500	1	0	3.378, 2.474	3.354, 2.440	4.965, 3.251	3.448, 2.521
					(0.111, 0.0963)	(2.616, 0.690)	(0.489, 0.652)
1	500	10	0	2.392, 1.914	2.374, 1.905	4.964, 3.250	2.452, 1.980
					(0.0925, 0.0803)	(6.726, 1.870)	(0.230, 0.219)
5	20	0	103	49.87, 30.14	46.81, 29.26	46.81, 29.26	55.62, 23.21
					(532.0, 320.0)	(532.0, 320.0)	(364.5, 539.5)
5	20	1	1	23.10, 15.81	21.06, 14.05	30.48, 18.55	29.43, 16.48
					(75.37, 40.42)	(129.4, 45.19)	(201.6, 65.97)
5	20	10	2	16.56, 12.70	15.51, 11.79	31.07, 19.60	23.27, 16.22
					(54.12, 41.62)	(279.3, 86.20)	(162.2, 51.67)
20	5	0	2245	59.61, 15.57	71.55, 23.22	71.55, 23.22	53.97, 0.555
					(1309., 196.5)	(1309., 196.5)	(646.7, 2407.)
20	5	1	43	25.63, 15.66	23.73, 14.64	37.61, 20.16	29.61, 12.60
					(207.8, 75.65)	(419.2, 92.17)	(147.4, 151.4)
20	5	10	43	21.20, 16.07	19.48, 14.96	39.63, 23.49	25.52, 16.60
					(184.2, 115.3)	(635.1, 165.9)	(111.8, 52.88)
100	5	0	0	13.23, 4.025	12.69, 3.855	12.69, 3.855	18.95, 0.393
					(9.618, 0.734)	(9.618, 0.734)	(135.1, 114.0)
100	5	1	0	6.205, 3.163	6.015, 3.072	8.871, 4.097	8.247, 2.855
					(1.474, 0.518)	(8.859, 1.144)	(16.54, 8.23)
100	5	10	0	4.781, 3.253	4.722, 3.215	9.654, 5.248	6.972, 4.160
					(1.335, 0.767)	(25.929, 4.608)	(12.44, 2.107)

Note: For multiple responses, $\theta > 0$.
Define $\theta = 0$ for the cases where each subject can only select one item.

association between row and columns in such a 3-way table under the multinomial sampling case. We discuss both ML and MH methods applied to multiple response situations. Although the ML method is superior for the estimation, when data are sparse, it is not feasible to obtain the ML estimates using the fitting algorithms. On the other hand, the MH method is appropriate under both sparse–data and large–stratum cases. This article generalizes the MH (co)variance estimators to the multiple response situation in such a way, that under the multinomial sampling case, the Greenland [13] (co)variance estimator is a special case of our newly proposed estimator. Suesse [26] also considered the odds ratio estimation for K $2 \times c$ tables based on c dependent binomials, which is an extension of the independent binomial sampling model presented by Greenland.

Liu and Suesse [21] presented another MH estimator to analyze stratified multiple response data for K $2 \times c$ tables considering how each item being selected depends on the row variable. Compared to their MH estimator, the newly proposed estimator is more useful for various cases. For instance, when items represent different time points in a longitudinal study, we might be interested in the time effect as well. The local odds ratios provide a broader view on the association across different items than the the odds ratios described by Liu and Suesse [21] that considered each of the items separately.

The odds ratio has the following property $\Psi_{jh} = \Psi_{js}\Psi_{sh}$. Thus $\log \Psi_{jh}$ cannot only be estimated by L_{js} but also by $L_{js} + L_{sh}$. There is no unique estimator. Greenland [13] proposed the following generalized MH estimator following the Mickey and Elashoff [24] approach:

$$\widehat{\log \Psi_{jh}} := \bar{L}_{jh} := (L_{j+} - L_{h+})/c.$$

This approach is independent of the applied estimator and generally applicable to any estimator of $\log \Psi_{jh}$. Then, the generalized MH estimators $\{\bar{L}_{jh}\}$ have the property $\bar{L}_{jh} = \bar{L}_{js} + \bar{L}_{sh}$. Yanagawa and Fujii [28] also showed that their projection method yields the generalized MH estimator when applied to the ordinary MH estimator. If one chooses to use the generalized estimators, a dually consistent estimator for the covariance of $\bar{L}_{jh}$ and $\bar{L}_{ts}$ is given by:

$$\bar{U}_{jhts} := \widehat{\text{Cov}}(\bar{L}_{jh}, \bar{L}_{ts}) = \frac{1}{c^2}\{U^+_{jt} - U^+_{js} - U^+_{ht} + U^+_{hs}\} \tag{6}$$

with

$$U^+_{jh} = \begin{cases} U^+_{jj} = U_{j++} = \sum_{a,b} U_{jab} & , j = h \\ U^+_{jh} = U_{+jh} - U_{jh+} - U_{hj+} + U_{jh} + U^*_{jh} & , j \neq h \end{cases}, \tag{7}$$

where $U^*_{jh} = \sum_{\text{distinct } j,a,h,b} U_{jahb}$ if $j \neq h$, otherwise $U^*_{jh} = 0$. Greenland [13] proposed exactly the same formula as (6), but the term U^+_{jh} in equation (7) has, in contrast to Greenland's definition, an additional term U^*_{jh} due to $\text{Cov}(L_{jh}, L_{ts}) \neq 0$ for multiple response data. Appendix C shows the details. For non-distinct indices, we obtain the following formulae as sub–cases:

$$\bar{U}_{jhs} := \widehat{\text{Cov}}(\bar{L}_{jh}, \bar{L}_{js}) = \frac{1}{c^2}\{U_{j++} - U^+_{js} - U^+_{hj} + U^+_{hs}\}$$

and

$$\bar{U}_{jhh} := \widehat{\text{Var}}(\bar{L}_{jh}) = \frac{1}{c^2}\{U_{j++} - 2U^{+}_{jh} + U_{h++}\}.$$

The results shown for the examples in Section 4 are based on the generalized MH estimators.

This paper only considered K $2 \times c$ tables and could be further extended to K $r \times c$ (with $r > 2$) tables. This extension would lead to another generalized MH estimator and different formulae for the (co)variance estimators of these generalized MH estimators. These formulae also require additional unknown covariance estimators, which are subject to future research.

References

[1] Agresti, A.: 2002, *Categorical Data Analysis*, Wiley Series in Probability and Statistics, 2nd edn, Wiley.

[2] Agresti, A. and Liu, I.: 2001, Strategies for modelling a categorical variable allowing multiple category choices, *Sociological Methods & Research* **29**(4), 403–434.

[3] Agresti, A. and Liu, I. M.: 1999, Modelling a categorical variable allowing arbitrarily many category choices, *Biometrics* **55**(3), 936–943.

[4] Bergsma, W., Croon, M. and Hagenaars, J. A.: 2009, *Marginal Models; For Dependent, Clustered, and Longitudinal Categorical Data*, Springer, New York.

[5] Bilder, C. R. and Loughin, T. M.: 2001, On the first-order Rao-Scott correction of the Umesh-Loughin-Scherer statistic, *Biometrics* **57**(4), 1253–1255.

[6] Bilder, C. R. and Loughin, T. M.: 2002, Testing for conditional multiple marginal independence, *Biometrics* **58**(1), 200–208.

[7] Bilder, C. R. and Loughin, T. M.: 2004, Testing for marginal independence between two categorical variables with multiple responses, *Biometrics* **60**(1), 241–248.

[8] Bilder, C. R. and Loughin, T. M.: 2009, Modeling multiple-response categorical data from complex surveys, *Canadian Journal of Statistics* **37**(4), 553–570.

[9] Bilder, C. R., Loughin, T. M. and Nettleton, D.: 2000, Multiple marginal independence testing for pick any/c variables, *Communications in Statistics-Simulation and Computation* **29**(4), 1285–1316.

[10] Coombs, C.: 1964, *A Theory of Data*, Wiley, New York.

[11] Decady, Y. J. and Thomas, D. R.: 2000, A simple test of association for contingency tables with multiple column responses, *Biometrics* **56**(3), 893–896.

[12] Gange, S. J.: 1995, Generating multivariate categorical variates using the iterative proportional fitting algorithm, *American Statistician* **49**(2), 134–138.

[13] Greenland, S.: 1989, Generalized Mantel-Haenszel estimators for K 2xJ tables, *Biometrics* **45**(1), 183–191.

[14] Haber, M.: 1985, Maximum-likelihood methods for linear and log-linear models in categorical-data, *Computational Statistics & Data Analysis* **3**(1), 1–10.

[15] Lang, J. B.: 1996, Maximum likelihood methods for a generalized class of log-linear models, *Annals of Statistics* **24**(2), 726–752.

[16] Lang, J. B.: 2004, Multinomial-Poisson homogeneous models for contingency tables, *Annals of Statistics* **32**(1), 340–383.

[17] Lang, J. B.: 2005, Homogeneous linear predictor models for contingency tables, *Journal of the American Statistical Association* **100**(469), 121–134.

[18] Lang, J. B. and Agresti, A.: 1994, Simultaneously modelling joint and marginal distributions of multivariate categorical responses, *Journal of the American Statistical Association* **89**(426), 625–632.

[19] Lee, A. J.: 1993, Generating random binary deviates having fixed marginal distributions and specified degrees of association, *American Statistician* **47**(3), 209–215.

[20] Liu, I.: 2003, Describing ordinal odds ratios for stratified r x c tables, *Biometrical Journal* **45**(6), 730–750.

[21] Liu, I. and Suesse, T.: 2008, The analysis of stratified multiple responses, *Biometrical Journal* **50**(1), 135–149.

[22] Loughin, T. M. and Scherer, P.: 1998, Testing for association in contingency tables with multiple column responses, *Biometrics* **54**, 630–637.

[23] Mantel, N. and Haenszel, W.: 1959, Statistical aspects of the analysis of data from retrospective studies of disease, *Journal of the National Cancer Institute* **22**, 719–748.

[24] Mickey, R. M. and Elashoff, R. M.: 1985, A generalization of the Mantel-Haenszel estimator of partial association for 2xJxK-tables, *Biometrics* **41**(3), 623–635.

[25] Sen, P. K. and Singer, J. M.: 1993, *Large Sample Methods in Statistics: An Introduction with Applications*, Chapman & Hall, New York.

[26] Suesse, T.: 2008, *Analysis and Diagnostics of Categorical Variables with Multiple Outcomes*, PhD thesis, Victoria University of Wellington.

[27] Thomas, D. R. and Decady, Y. J.: 2004, Testing for association using multiple response survey data: Approximate procedures based on the rao-scott approach, *International Journal of Testing* **4**(1), 43–59.

[28] Yanagawa, T. and Fujii, Y.: 1995, Projection-method Mantel-Haenszel estimator for K 2xJ tables, *Journal of the American Statistical Association* **90**(430), 649–656.

A. Dually Consistency of Ordinary MH Estimator

A.1. Sparse Data Limiting Model

For the sparse data limiting model, the number of observations per stratum is bounded ($O(N_k) = 1$) and K approaches infinity.

From $\pi_{j|1k}\pi_{h|2k} = \Psi_{jh}\pi_{h|1k}\pi_{j|2k}$, which follows from the assumption of a common odds ratio, and equation (4), we derive

$$\begin{aligned}\mathbb{E}\omega_{jh|k} = \mathbb{E}(c_{jh|k} - \Psi_{jh}c_{hj|k}) =& \mathbb{E}c_{jh|k} - \Psi_{jh}\mathbb{E}c_{hj|k} \\ =& \{\mathbb{E}X_{j|1k}\mathbb{E}X_{h|2k} - \Psi_{jh}\mathbb{E}X_{h|1k}\mathbb{E}X_{j|2k}\}/N_k \\ =& \{n_{1k}n_{2k}\pi_{j|1k}\pi_{h|2k} - \Psi_{jh}n_{1k}n_{2k}\pi_{h|1k}\pi_{j|2k}\}/N_k \\ =& \{n_{1k}n_{2k}(\pi_{j|1k}\pi_{h|2k} - \pi_{j|1k}\pi_{h|2k})\}/N_k = 0\end{aligned}$$

We can write

$$\hat{\Psi}_{jh} - \Psi_{jh} = \frac{\sum_{k=1}^{K} c_{jh|k} - \Psi_{jh}c_{hj|k}}{\sum_{k=1}^{K} c_{hj|k}} = \frac{\sum_{k=1}^{K}(c_{jh|k} - \Psi_{jh}c_{hj|k})/K}{\sum_{k=1}^{K} c_{hj|k}/K} \tag{8}$$

$$= \frac{\sum_{k=1}^{K} \omega_{jh|k}/K}{\sum_{k=1}^{K} c_{hj|k}/K} = \frac{\Omega_{jh}/K}{C_{hj}/K} \quad . \tag{9}$$

with with $\omega_{jh|k} := c_{jh|k} - \Psi_{jh}c_{hj|k}$ and $\Omega_{jh} := \sum_k \omega_{jh|k}$.

The term $c_{jh|k}$ is a bounded random variable under model II, hence, the variance of C_{jh} is $o(K^2)$ and Chebyshev's weak law of large numbers states $(\Omega_{jh} - \mathbb{E}\Omega_{jh})/K \to_p 0$. Since $\mathbb{E}\omega_{jh|k} = 0$, the expression $(\Omega_{jh} - \mathbb{E}\Omega_{jh})/K \to_p 0$ reduces to $\Omega_{jh}/K \to_p 0$, that is, the numerator of $\hat{\Psi}_{jh} - \Psi_{jh}$ converges to zero in probability. Applying the Chebyshev weak law of large numbers again to the denominator yields

$$\sum_{k=1}^{K} c_{jh|k}/K \overset{K\to\infty}{\longrightarrow}_p \lim_{K\to\infty} \sum_{k=1}^{K} \mathbb{E}(c_{jh|k})/K < \infty.$$

This limit is finite and nonzero. Thus, we conclude $\hat{\Psi}_{jh} - \Psi_{jh} \to_p 0$ by Slutsky's theorem.

A.2. Large Stratum Limiting Model

Let us consider the case $N \to \infty$ with $N\alpha_{ik} = n_{ik}$ and $0 < \alpha_{ik} < 1$, that is, as N approaches infinity the number of subjects n_{ik}, for all rows i and strata k, also approaches infinity. Note $N_k = n_{1k} + n_{2k} = N \sum_i \alpha_{ik}$.

We have

$$\begin{aligned} C_{jh}/N &= \sum_{k=1}^{K} c_{jh|k}/N = \sum_{k=1}^{K} X_{j|1k} X_{h|2k}/(N_k N) \\ &= \sum_{k=1}^{K} \frac{n_{1k} n_{2k}}{N_k N} \frac{X_{j|1k}}{n_{1k}} \frac{X_{h|2k}}{n_{2k}} = \sum_{k=1}^{K} \frac{n_{1k} n_{2k}}{NN} \frac{N}{N_k} \frac{X_{j|1k}}{n_{1k}} \frac{X_{h|2k}}{n_{2k}} \\ &\overset{N\to\infty}{\longrightarrow}_p \sum_{k=1}^{K} \alpha_{1k}\alpha_{2k} (\sum_i \alpha_{ik})^{-1} \pi_{j|1k}\pi_{h|2k} = \sum_{k=1}^{K} (\sum_i \alpha_{ik}^{-1})^{-1} \pi_{j|1k}\pi_{h|2k}. \end{aligned}$$

Therefore

$$\begin{aligned} \hat{\Psi}_{jh} = \frac{C_{jh}}{C_{hj}} = \frac{C_{jh}/N}{C_{hj}/N} &\overset{N\to\infty}{\longrightarrow}_p \frac{\sum_{k=1}^{K} (\sum_i \alpha_{ik}^{-1})^{-1} \pi_{j|1k}\pi_{h|2k}}{\sum_{k=1}^{K} (\sum_i \alpha_{ik}^{-1})^{-1} \pi_{h|1k}\pi_{j|2k}} \\ &= \Psi_{jh} \frac{\sum_{k=1}^{K} (\sum_i \alpha_{ik}^{-1})^{-1} \pi_{h|1k}\pi_{j|2k}}{\sum_{k=1}^{K} (\sum_i \alpha_{ik}^{-1})^{-1} \pi_{h|1k}\pi_{j|2k}} = \Psi_{jh}. \end{aligned}$$

B. Asymptotic Covariances

B.1. Sparse-data Limiting Model

Let $\text{Var}^a(\cdot)$ and $\text{Cov}^a(\cdot)$ refer to the asymptotic variance and covariance. From above $\hat{\Psi}_{jh} - \Psi_{jh} = \frac{\Omega_{jh}/K}{C_{hj}/K} = \frac{\sum_k \omega_{jh|k}/K}{C_{hj}/K}$.

First by independence of rows $\text{Cov}(\Omega_{jh}, \Omega_{ts}) = \sum_{k=1}^{K} \text{Cov}(\omega_{jh|k}, \omega_{ts|k})$. Note that $\mathbb{E}|\omega_{jh|k} - \mathbb{E}\omega_{jh|k}|^3 = \mathbb{E}|\omega_{jh|k}|^3 = O(1)$, because $c_{jh|k}$ is a bounded random variable under the sparse-data limiting model. By setting $\delta = 1$, we conclude from the Multivariate Central Limit Theorem [25, p.123] that $K^{-1/2}\,(\Omega_{jh}, \Omega_{ts}) = \sqrt{K}(\Omega_{jh}/K,\, \Omega_{ts}\,/K)$ converges to a zero mean multivariate normal distribution with covariance $\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \text{Cov}(\omega_{jh|k}, \omega_{ts|k})$, by noting that $\mathbb{E}\omega_{jh|k} = 0$ and $\text{Cov}(\omega_{jh},\, \omega_{ts})$ exists. We conclude the asymptotic covariance between Ω_{jh} and Ω_{ts} is $\lim_{K\to\infty} K \cdot \text{Cov}^a(\Omega_{jh}, \Omega_{ts}) = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \text{Cov}(\omega_{jh|k}, \omega_{ts|k})$.

Therefore by the delta method, Slutsky's theorem, equation (8), and using that the denominator terms $\lim_K \mathbb{E}C_{hj}/K$ are finite we obtain

$$\begin{aligned}
&\lim_{K\to\infty} K \cdot \text{Cov}^a(\log \hat{\Psi}_{jh}, \log \hat{\Psi}_{ts}) \\
=&1/(\Psi_{jh}\Psi_{ts}) \lim_{K\to\infty} K \cdot \text{Cov}^a(\hat{\Psi}_{jh}, \hat{\Psi}_{ts}) \\
=&1/(\Psi_{jh}\Psi_{ts}) \frac{\lim_{K\to\infty} K \cdot \text{Cov}^a(\Omega_{jh}, \Omega_{ts})}{(\lim_K \mathbb{E}C_{hj}/K)(\lim_K \mathbb{E}C_{st}/K)} \\
\\
=&1/(\Psi_{jh}\Psi_{ts}) \frac{\lim_{K\to\infty} 1/K \cdot \sum_k \text{Cov}(\omega_{jh|k}, \omega_{ts|k})}{(\lim_K \mathbb{E}C_{hj}/K)(\lim_K \mathbb{E}C_{st}/K)}
\end{aligned}$$

for arbitrary indices $j, h, s, t \in \{1, \ldots, c\}$ with $j \neq h$ and $s \neq t$.

Now we obtain the following variance

$$\text{Var}(\omega_{jh|k}) = v^1_{jh|k} - 2\Psi_{jh} v^2_{jh|k} + \Psi^2_{jh} v^3_{jh|k}$$

and covariances

$$\begin{aligned}
\text{Cov}(\omega_{jh|k}, \omega_{js|k}) &= v_{jhs|12,k} - \Psi_{jh} v_{jh,js|k} - \Psi_{js} v_{js,jh|k} + \Psi_{jh}\Psi_{js} v_{jhs|21,k} \\
\text{Cov}(\omega_{jh|k}, \omega_{ts|k}) &= v_{jt,hs|k} - \Psi_{jh} v_{ht,js|k} - \Psi_{ts} v_{js,ht|k} + \Psi_{jh}\Psi_{ts} v_{hs,st|k}
\end{aligned}$$

with

$$\begin{aligned}
v^1_{jh|k} &= \frac{n_1 n_2}{N^2}(\pi_{j|1}\pi_{h|2} + n_1'\pi^2_{j|1}\pi_{h|2} + n_2'\pi_{j|1}\pi^2_{h|2}) \\
v^2_{jh|k} &= \frac{n_1 n_2}{N^2}(n_1'\pi_{j|1}\pi_{h|1}\pi_{jh|2} + n_2'\pi_{j|2}\pi_{h|2}\pi_{jh|1} + \pi_{jh|1}\pi_{jh|2}) \\
v^3_{jh|k} &= \frac{n_1 n_2}{N^2}(\pi_{h|1}\pi_{j|2} + n_1'\pi^2_{h|1}\pi_{j|2} + n_2'\pi_{h|1}\pi^2_{j|2}) \\
v_{jh,ts|k} &= \frac{n_1 n_2}{N^2}(\pi_{jh|1}\pi_{ts|2} + n_1'\pi_{j|1}\pi_{h|1}\pi_{ts|2} + n_2'\pi_{jh|1}\pi_{t|2}\pi_{s|2}) \\
v_{jhs|abk} &= \frac{n_1 n_2}{N^2} v^A_{jhs|abk} + v^B_{jhs|abk} \ (a \neq b) \\
v^A_{jhs|abk} &= \frac{n_1 n_2}{N^2}\pi_{hs|bk}(\pi_{j|ak} + n_a'\pi^2_{j|ak}) \ (a \neq b) \\
v^B_{jhs|abk} &= \frac{n_1 n_2}{N^2} n_b'\pi_{j|ak}\pi_{h|bk}\pi_{s|bk} \ (a \neq b).
\end{aligned}$$

The subscript k is often suppressed for convenience only.

The (co)variance estimators were constructed in such a way that they converge exactly to the asymptotic (co)variance(s). We can also express U_{jhs} as $U_{jhs} =$

U^{add}_{jhs} omitting U^{old}_{jhs} but only if $\hat{v}^B_{jhs|abk}$ is amended to $\hat{v}^B_{jhs|abk} = \frac{1}{N_k^2} X_{j|ak}\{X_{h|bk}X_{s|bk} - X_{hs|bk}\}$. Then for the covariance estimators we have $\sum_k \hat{v}_k/K \overset{K\to\infty}{\longrightarrow} \sum_k \mathbb{E}\hat{v}_k/K = \lim_K \sum_k v_k/K$ and $\sum_k c_{jh|k}/K \overset{K\to\infty}{\longrightarrow} \sum_k \mathbb{E}c_{jh|k}/K$ by Chebyshev's weak law of large numbers.

B.2. Large-stratum Limiting Model

By the delta method, the large stratum limiting variance is

$$\begin{aligned}
&\lim_{N\to\infty} N \cdot \text{Var}^a(\log \hat{\Psi}_{jh}) \\
&= \frac{\sum_k \frac{\alpha_1^2\alpha_2}{(\sum_i \alpha_{ik})^2}\{\pi^2_{j|1}\pi_{h|2} + \Psi^2_{jh}\pi^2_{h|1}\pi_{j|2} - 2\Psi_{jh}\pi_{j|1}\pi_{h|1}\pi_{jh|2}\}}{(\sum_k(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{h|1k}\pi_{j|2k})^2} \\
&\quad + \frac{\sum_k \frac{\alpha_1\alpha_2^2}{(\sum_i \alpha_{ik})^2}\{\pi_{j|1}\pi^2_{h|2} + \Psi^2_{jh}\pi_{h|1}\pi^2_{j|2} - 2\Psi_{jh}\pi_{jh|1}\pi_{j|2}\pi_{h|2}\}}{(\sum_k(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{h|1k}\pi_{j|2k})^2}
\end{aligned}$$

and the limiting covariances are

$$\begin{aligned}
&\lim_{N\to\infty} N \cdot \text{Cov}^a(\log \hat{\Psi}_{jh}, \log \hat{\Psi}_{js}) \\
&= \frac{\sum_k \frac{\alpha_1^2\alpha_2}{(\sum_i \alpha_{ik})^2}\{\pi^2_{j|1}\pi_{hs|2} - \Psi_{jh}\pi_{j|1}\pi_{h|1}\pi_{js|2} - \Psi_{js}\pi_{j|1}\pi_{s|1}\pi_{jh|2} + \Psi_{jh}\Psi_{js}\pi_{h|1}\pi_{s|1}\pi_{j|2}\}}{(\sum_k(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{h|1k}\pi_{j|2k})^2} \\
&\quad + \frac{\sum_k \frac{\alpha_1\alpha_2^2}{(\sum_i \alpha_{ik})^2}\{\pi_{j|1}\pi_{h|2}\pi_{s|2} - \Psi_{jh}\pi_{jh|1}\pi_{j|2}\pi_{s|2} - \Psi_{js}\pi_{js|1}\pi_{j|2}\pi_{h|2} + \Psi_{jh}\Psi_{js}\pi_{hs|1}\pi^2_{j|2}\}}{(\sum_k(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{h|1k}\pi_{j|2k})^2}
\end{aligned}$$

$$\begin{aligned}
&\lim_{N\to\infty} N \cdot \text{Cov}^a(\log \hat{\Psi}_{jh}, \log \hat{\Psi}_{ts}) \\
&= \frac{\sum_k \frac{\alpha_1^2\alpha_2}{(\sum_i \alpha_{ik})^2}\{\pi_{j|1}\pi_{t|1}\pi_{hs|2} - \Psi_{jh}\pi_{h|1}\pi_{t|1}\pi_{js|2} - \Psi_{ts}\pi_{j|1}\pi_{s|1}\pi_{ht|2} + \Psi_{jh}\Psi_{ts}\pi_{h|1}\pi_{s|1}\pi_{jt|2}\}}{(\sum_k(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{h|1k}\pi_{j|2k})^2} \\
&\quad + \frac{\sum_k \frac{\alpha_1\alpha_2^2}{(\sum_i \alpha_{ik})^2}\{\pi_{jt|1}\pi_{h|2}\pi_{s|2} - \Psi_{jh}\pi_{ht|1}\pi_{j|2}\pi_{s|2} - \Psi_{ts}\pi_{js|1}\pi_{h|2}\pi_{t|2} + \Psi_{jh}\Psi_{ts}\pi_{hs|1}\pi_{j|2}\pi_{t|2}\}}{(\sum_k(\sum_i \alpha_{ik}^{-1})^{-1}\pi_{h|1k}\pi_{j|2k})^2}.
\end{aligned}$$

The estimators were constructed such that

$$\begin{aligned}
\lim_{N\to\infty} N \cdot \text{Var}^a(\log \hat{\Psi}_{jh}) &= \lim_N N \cdot U_{jhh} \\
\lim_{N\to\infty} N \cdot \text{Cov}^a(\log \hat{\Psi}_{jh}, \log \hat{\Psi}_{js}) &= \lim_{N\to\infty} N \cdot U_{jhs} \\
\lim_{N\to\infty} N \cdot \text{Cov}^a(\log \hat{\Psi}_{jh}, \log \hat{\Psi}_{ts}) &= \lim_{N\to\infty} N \cdot U_{jhts}.
\end{aligned}$$

C. Generalized Covariance Estimators

We can write

$$\text{Cov}(\bar{L}_{jh}, \bar{L}_{ts}) = \text{Cov}\left(1/c\sum_{a=1}^{c}(L_{ja} - L_{ha}), 1/c\sum_{a=1}^{c}(L_{ta} - L_{sa})\right)$$
$$= \frac{1}{c^2}\sum_{a}\{\text{Cov}(L_{ja}, L_{ta}) + \text{Cov}(L_{ha}, L_{sa}) - \text{Cov}(L_{ja}, L_{sa}) - \text{Cov}(L_{ha}, L_{ta})\}$$
$$+ \frac{1}{c^2}\sum_{a\neq b}\{\text{Cov}(L_{jb}, L_{ta}) + \text{Cov}(L_{hb}, L_{sa}) - \text{Cov}(L_{jb}, L_{sa}) - \text{Cov}(L_{hb}, L_{ta})\}$$

and express $\sum_{a\neq b}\text{Cov}(L_{jb}, L_{ta})$ as

$$\begin{aligned}\sum_{a\neq b}\text{Cov}(L_{jb}, L_{ta}) = &\sum_{\substack{b\\(a=j)}}\text{Cov}(L_{jb}, L_{tj}) + \sum_{\substack{a\\(b=t)}}\text{Cov}(L_{jt}, L_{ta})\\ &- \text{Cov}(L_{jt}, L_{tj}) + \sum_{\text{distinct } j,b,t,a}\text{Cov}(L_{jb}, L_{ta})\\ = &-\sum_{a}\text{Cov}(L_{jt}, L_{ja}) - \sum_{a}\text{Cov}(L_{tj}, L_{ta})\\ &+ \text{Cov}(L_{jt}, L_{jt}) + \sum_{\text{distinct } j,b,t,a}\text{Cov}(L_{jb}, L_{ta})\end{aligned}$$

Now it is clear how we derived the equations (6) and (7). For a more detailed proof of the dually consistency of the proposed estimators, we refer to Suesse [26].