On connectivity of fibers with positive marginals in multiple logistic regression

Ruriko Yoshida
Dept. of Statistics, University of Kentucky

Joint work with H. Hara and A. Takemura

`polytopes.net`

	Blood Pressure	Serum Cholesterol (mg/100ml) 1 < 200	2 200-209	3 210-219	4 220-244	5 245-259	6 260-284	7 > 284
1	< 117	2/53	0/21	0/15	0/20	0/14	1/22	0/11
2	117-126	0/66	2/27	1/25	8/69	0/24	5/22	1/19
3	127-136	2/59	0/34	2/21	2/83	0/33	2/26	4/28
4	137-146	1/65	0/19	0/26	6/81	3/23	2/34	4/23
5	147-156	2/37	0/16	0/6	3/29	2/19	4/16	1/16
6	157-166	1/13	0/10	0/11	1/15	0/11	2/13	4/12
7	167-186	3/21	0/5	0/11	2/27	2/5	6/16	3/14
8	> 186	1/5	0/1	3/6	1/10	1/7	1/7	1/7

Source : [Cornfield, 1962]

Data on coronary heart disease incidence in Framingham, Massachusetts [Cornfield, 1962, Agresti, 1990]. A sample of male residents, aged 40 through 50, were classified on blood pressure and serum cholesterol concentration. $2/53$ in the (1,1) cell means that there are 53 cases, of whom 2 exhibited heart disease.

Data on occurrence of esophageal cancer

Table 1: Data on occurrence of esophageal cancer

	Alcohol Consumption	Age 1 25-34	2 35-44	3 45-54	4 55-64	5 65-74	6 75+
0	Low	0/106	5/169	21/159	34/173	36/124	8/39
1	High	1/10	4/30	25/54	42/69	19/37	5/5

Source : [Breslow and Day, 1980]

This table refers to the occurrence of esophageal cancer in Frenchmen which were classified on ages and dummy variable on alcohol consumption.

Logistic regression and positive margins

In most applications of the logistic regression model, for each combination of covariates, the number of "successes" and the number of "failures" are observed.

The number of trials (i.e. the sum of numbers of "successes" and "failures") for each combination of covariates is usually fixed by a sampling scheme and positive. We call this marginal **the response variable marginal**.

Therefore we are usually interested in the connectivity of fibers with positive response variable marginals for sampling tables via Monte Carlo Markov chain (MCMC).

Univariate Logistic Regression Model

Let $\{1, \ldots, J\}$ be the set levels of a covariate and let X_{1j} and X_{2j}, $j = 1, \ldots, J$, be the numbers of successes and failures, respectively. The probability for success p_j is modeled as

$$\text{logit}(p_j) = \log \frac{p_j}{1 - p_j} = \alpha + \beta j, \qquad j = 1, \ldots, J.$$

The sufficient statistics for the model is $(X_{1+}, X_{+1}, \ldots, X_{+J}, \sum_{j=1}^{J} j X_{1j})$.

A **move** z is a table such that $X + z$ satisfies the given margins.

Moves $z = (z_{ij})$ for the model satisfy $(z_{1+}, z_{+1}, \ldots, z_{+J}) = 0$ and

$$\sum_{j=1}^{J} j z_{1j} = 0.$$

Bivariate Logistic Regression Model

Let $\{1, \ldots, J\}$ and $\{1, \ldots, K\}$ be the sets levels of two covariates. Let X_{1jk} and X_{2jk}, $j = 1, \ldots, J$, $k = 1, \ldots, K$, be the numbers of "successes" and "failures", respectively, for level (j, k). The probability for "success" p_{1jk} is modeled as

$$\text{logit}(p_{1jk}) = \log\left(\frac{p_{1jk}}{1 - p_{1jk}}\right) = \mu + \alpha j + \beta k,$$

$$j = 1, \ldots, J, \quad k = 1, \ldots, K.$$

The sufficient statistics for this model is X_{1++}, $\sum_{j=1}^{J} jX_{1j+}$, $\sum_{k=1}^{K} kX_{1+k}$, X_{+jk}, $\forall j, k$.

Hence moves $Z = (z_{ijk})$ for the model satisfy

$$z_{1++} = 0, \quad \sum_{j=1}^{J} j z_{1j+} = 0, \quad \sum_{k=1}^{K} k z_{1+k} = 0, \quad z_{+jk} = 0, \ \forall j, k.$$

A **Markov basis** is a set of **moves** which is guaranteed to connect all feasible contingency tables satisfying the given margins [Diaconis and Sturmfels, 1998].

Difficulty: the number of elements in a minimal Markov basis for a model can be exponentially many.

Question: Finding a set of Markov connecting moves that are much simpler than the full Markov basis with positive response variable marginals.

Such connecting sets are called **Markov subbases** [Chen et. al., 2006].

Markov subbasis for univariate logistic regression

Let $\boldsymbol{e}_j$ denote the contingency table with just 1 frequency in the j-th cell.

$\mathcal{B} = \{\pm(\boldsymbol{e}_{j_1}+\boldsymbol{e}_{j_4}-\boldsymbol{e}_{j_2}-\boldsymbol{e}_{j_3}) \mid 1 \leq j_1 < j_2 \leq j_3 < j_4 \leq J, \ \ j_2-j_1 = j_4-j_3\}$

Theorem: [Chen, Dinwoodie, Dobra, Huber, 2005]

The set of moves

$$\mathcal{B}_0 = \{z \in \mathcal{B} \mid j_2 = j_1 + 1, j_3 = j_4 - 1\}$$

connects every fiber satisfying $(X_{+1}, \ldots, X_{+J}) > 0$ for the univariate logistic regression model.

Markov subbasis for univariate logistic regression

if $j_2 \neq j_3$

$$\begin{array}{c|cccc|} & j_1 & j_2 & j_3 & j_4 \\ \hline i=1 & -1 & 1 & 1 & -1 \\ i=2 & 1 & -1 & -1 & 1 \\ \hline \end{array}$$

$$\begin{array}{c|cccc|} & j_1 & j_2 & j_3 & j_4 \\ \hline i=1 & 1 & -1 & -1 & 1 \\ i=2 & -1 & 1 & 1 & -1 \\ \hline \end{array}$$

if $j_2 = j_3$

$$\begin{array}{c|ccc|} & j_1 & j_2 & j_4 \\ \hline i=1 & 1 & -2 & 1 \\ i=2 & -1 & 2 & -1 \\ \hline \end{array}$$

$$\begin{array}{c|ccc|} & j_1 & j_2 & j_4 \\ \hline i=1 & -1 & 2 & -1 \\ i=2 & 1 & -2 & 1 \\ \hline \end{array}$$

Configuration for the bivariate logistic regression model

Consider two configurations $A = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_J)$ and $B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_K)$, where $\boldsymbol{a}_j$ and $\boldsymbol{b}_k$ are column vectors. We assume the homogeneity, i.e., there exist weight vectors w, v such that $\langle w, \boldsymbol{a}_j \rangle = 1, \ \forall j, \ \langle v, \boldsymbol{b}_k \rangle = 1, \ \forall k$.

The configuration $A \otimes B$ of the **Segre product** of A and B is defined as

$$A \otimes B = \Big(\boldsymbol{a}_j \oplus \boldsymbol{b}_k, \ j = 1, \ldots, J, k = 1, \ldots, K\Big), \qquad \boldsymbol{a}_j \oplus \boldsymbol{b}_k = \begin{pmatrix} \boldsymbol{a}_j \\ \boldsymbol{b}_k \end{pmatrix}.$$

Let

$$A = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 2 & \ldots & J \end{pmatrix} . B = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ 1 & 2 & \ldots & K \end{pmatrix}.$$

Fact: The configuration for the bivariate logistic regression model is the Lawrence lifting of Segre product $\Lambda(A \otimes B)$.

Markov subbasis

Consider a set of moves which connects every fiber satisfying $X_{+jk} > 0$, $\forall j, k$.

Let $\boldsymbol{e}_{jk} = (e_{ijk})$ be redefined as an integer array with 1 at the cell $(1jk)$, -1 at the cell $(2jk)$ and 0 everywhere else. Define $\mathcal{B}_{\Lambda(A \otimes B)}$ as the set of moves $z = (z_{ijk})$ satisfying the following conditions,

1. $z = \boldsymbol{e}_{j_1 k_1} - \boldsymbol{e}_{j_2 k_2} - \boldsymbol{e}_{j_3 k_3} + \boldsymbol{e}_{j_4 k_4}$;

2. $(j_1, k_1) - (j_2, k_2) = (j_3, k_3) - (j_4, k_4)$.

Theorem [Hara, Takemura, Y., 2009]

$\mathcal{B}_{\Lambda(A \otimes B)}$ connects every fiber satisfying $X_{+jk} > 0$, $\forall j, k$.

Examples of moves ($i = 1$ layer)

(1) $k_1 = \cdots = k_4$

$$\begin{array}{c|cccc|} & j_1 & j_2 & j_3 & j_4 \\ \cline{2-5} k_1 & 1 & -1 & -1 & 1 \\ \cline{2-5} \end{array}$$

(2) $k_1 = \cdots = k_4$ and $j_2 = j_3$

$$\begin{array}{c|ccc|} & j_1 & j_2 & j_4 \\ \cline{2-4} k_1 & 1 & -2 & 1 \\ \cline{2-4} \end{array}$$

(3) $k_1 = k_2$ and $j_2 = j_3$

$$\begin{array}{c|ccc|} & j_1 & j_2 & j_4 \\ \cline{2-4} k_1 & 1 & -1 & 0 \\ k_3 & 0 & -1 & 1 \\ \cline{2-4} \end{array}$$

(4) $(j_2, k_2) = (j_3, k_3)$

$$\begin{array}{c|ccc|} & j_1 & j_2 & j_4 \\ \cline{2-4} k_1 & 1 & 0 & 0 \\ k_2 & 0 & -2 & 0 \\ k_4 & 0 & 0 & 1 \\ \cline{2-4} \end{array}$$

(5) $k_1 = k_2$ $(k_3 = k_4)$

$$\begin{array}{c|cccc|} & j_1 & j_2 & j_3 & j_4 \\ \cline{2-5} k_1 & 1 & -1 & 0 & 0 \\ k_3 & 0 & 0 & -1 & 1 \\ \cline{2-5} \end{array}$$

(6) $k_1 = k_4$ and $j_2 = j_3$

$$\begin{array}{c|ccc|} & j_1 & j_2 & j_4 \\ \cline{2-4} k_2 & 0 & -1 & 0 \\ k_1 & 1 & 0 & 1 \\ k_3 & 0 & -1 & 0 \\ \cline{2-4} \end{array}$$

	Blood Pressure	Serum Cholesterol (mg/100ml) 1 < 200	2 200-209	3 210-219	4 220-244	5 245-259	6 260-284	7 > 284
1	< 117	2/53	0/21	0/15	0/20	0/14	1/22	0/11
2	117-126	0/66	2/27	1/25	8/69	0/24	5/22	1/19
3	127-136	2/59	0/34	2/21	2/83	0/33	2/26	4/28
4	137-146	1/65	0/19	0/26	6/81	3/23	2/34	4/23
5	147-156	2/37	0/16	0/6	3/29	2/19	4/16	1/16
6	157-166	1/13	0/10	0/11	1/15	0/11	2/13	4/12
7	167-186	3/21	0/5	0/11	2/27	2/5	6/16	3/14
8	> 186	1/5	0/1	3/6	1/10	1/7	1/7	1/7

Source : [Cornfield, 1962]

Data on coronary heart disease incidence in Framingham, Massachusetts [Cornfield, 1962, Agresti, 1990]. A sample of male residents, aged 40 through 50, were classified on blood pressure and serum cholesterol concentration. $2/53$ in the (1,1) cell means that there are 53 cases, of whom 2 exhibited heart disease.

Data on coronary heart disease incidence

We examine the goodness-of-fit of the model with $J = 7$ and $K = 8$ by likelihood ratio statistic L_0.

We test the bivariate logistic regression defined above as a null hypothesis vs. ANOVA type logit model, namely:

$$\mathrm{H}_0 : \mathrm{logit}(p_{1jk}) = \log\left(\frac{p_{1jk}}{1 - p_{1jk}}\right) = \mu + \alpha j + \beta k,$$

for $j = 1, \ldots, J, \quad k = 1, \ldots, K$.

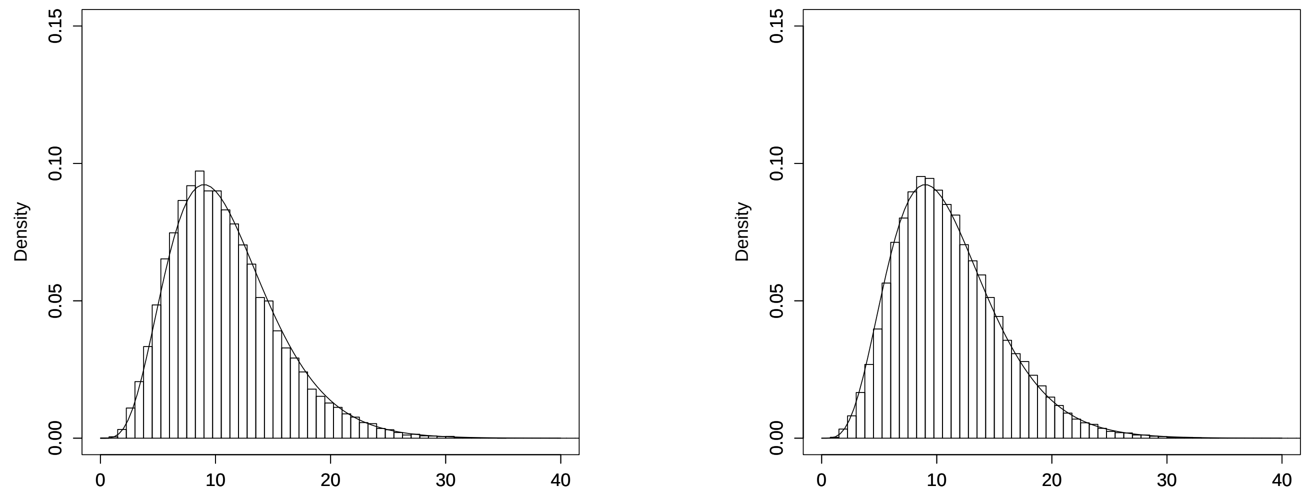
$$\mathrm{H}_1 : \mathrm{logit}(p_{1jk}) = \log\left(\frac{p_{1jk}}{1 - p_{1jk}}\right) = \mu + \alpha_j + \beta_k,$$

where $\sum_{j=1}^{J} \alpha_j = 0$ and $\sum_{k=1}^{K} \beta_k = 0$.

Data on coronary heart disease incidence

The value of L_0 is 13.07587 and the asymptotic p-value is 0.2884 from the asymptotic distribution χ^2_{11}. We computed the exact distribution of L_0 via MCMC with $\mathcal{B}_{\Gamma(A \otimes B)}$ defined. As an extension of $\mathcal{B}_0$ to the bivariate model, we define $\mathcal{B}_0^2$ by the set of moves $z = \boldsymbol{e}_{j_1k_1} - \boldsymbol{e}_{j_2k_2} - \boldsymbol{e}_{j_3k_3} + \boldsymbol{e}_{j_4k_4}$ satisfying $(j_1, k_1) - (j_2, k_2) = (j_3, k_3) - (j_4, k_4)$ is either of $(\pm 1, 0)$, $(0, \pm 1)$, $(\pm 1, \pm 1)$ or $(\pm 1, \mp 1)$.

The estimated p-values are 0.2706 with $\mathcal{B}_{\Gamma(A \otimes B)}$ and 0.2958 with $\mathcal{B}_0^2$. Therefore bivariate logistic regression model is accepted.



(a) A histogram with $\mathcal{B}_{\Lambda(A\otimes B)}$ (b) A histogram with $\mathcal{B}_0^2$

Figure 1: Histograms of L_0 via MCMC with $\mathcal{B}_{\Lambda(A\otimes B)}$ and $\mathcal{B}_0^2$

Data on occurrence of esophageal cancer

Table 2: Data on occurrence of esophageal cancer

	Alcohol Consumption	Age 1 25-34	2 35-44	3 45-54	4 55-64	5 65-74	6 75+
0	Low	0/106	5/169	21/159	34/173	36/124	8/39
1	High	1/10	4/30	25/54	42/69	19/37	5/5

Source : [Breslow and Day, 1980]

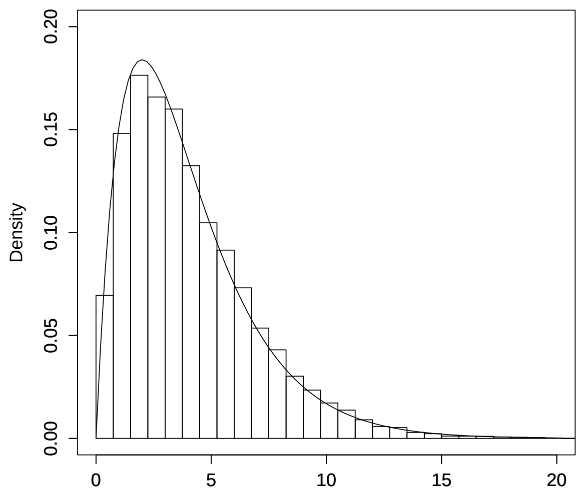
This table refers to the occurrence of esophageal cancer in Frenchmen which were classified on ages and dummy variable on alcohol consumption.

Data on occurrence of esophageal cancer

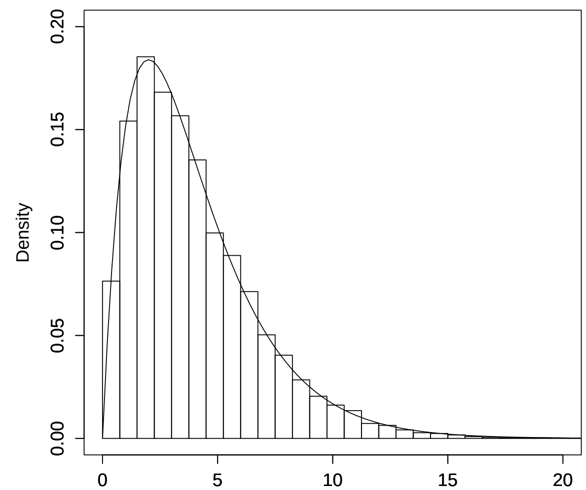
We test the goodness-of-fit of the bivariate logistic regression model with $J = 6$ and $K = 2$ by likelihood ratio statistics L_0 via MCMC. Then the value of L_0 is 20.89 and the asymptotic p-value is 0.0003330 from the asymptotic distribution χ^2_4.

We computed the exact distribution of L_0 via MCMC with $\mathcal{B}_{\Gamma(A \otimes B)}$ and $\mathcal{B}_0^2$. Figure 2 represents the histograms of L_0. The estimated p-values are 0.00011 with $\mathcal{B}_{\Gamma(A \otimes B)}$ and 0.00055 with $\mathcal{B}_0^2$. Therefore the model is rejected at the significance level of 1%.

Data on occurrence of esophageal cancer



(a) a histogram with $\mathcal{B}_{\Lambda(A \otimes B)}$



(b) a histogram with $\mathcal{B}_0^2$

Figure 2: Histograms of L_0 via MCMC with $\mathcal{B}_{\Lambda(A \otimes B)}$ and $\mathcal{B}_0^2$

The smooth line is asymptotic chi-square density, which shows a good fit.

Conjectures

The current proof for bivariate case is already very difficult and the general multivariate case remains to be a conjecture.

Conjecture: The set of moves $\mathcal{B}_{\Lambda(A_1 \otimes \cdots \otimes A_m)}$ connects every fiber with positive response marginals for the logistic regression with m covariates.

Conjecture: The subset of moves from $\mathcal{B}_{\Lambda(A_1 \otimes \cdots \otimes A_m)}$ such that the elements of $\boldsymbol{j}_1 - \boldsymbol{j}_2 = \boldsymbol{j}_3 - \boldsymbol{j}_4$ are ± 1 or 0 connects every fiber with positive response marginals for the logistic regression with m covariates. This is still conjecture for even $m = 2$.

Thank you....

The paper is available at http://arxiv.org/abs/0810.1793.