On exploratory analytic method for multi-way contingency tables with an ordinal response variable and categorical explanatory variables

Zheng Wei and Daeyoung Kim

University of Maine and University of Massachusetts-Amherst

Supplementary Material

The supplementary material includes the tables and figures for Example 1, and the simulation design and simulation results of Section 4.1.

S1. Example 1

Fig.S1 shows the checkerboard copula density of X_1 and X_2 in Example 1 of the main paper.

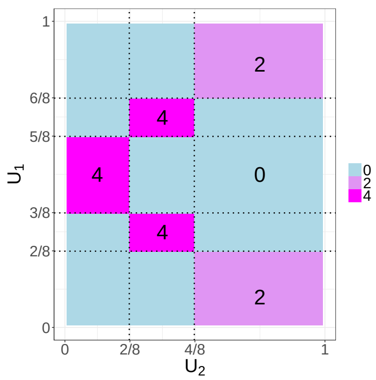

Fig. S1: Copula density $c^{+}(u_1, u_2)$

Fig.S2 shows the checkerboard copula regression function of U_2 on U_1 in Example 1 of the main paper.

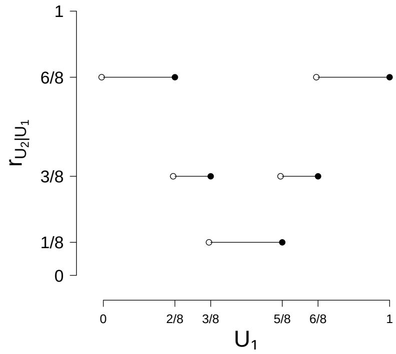

Fig. S2: $r_{U_2|U_1}(u_1)$ vs. u_1

Table S1 provides an example illustrating the invariant property of the checkerboard copula regression given in the Proposition 2 of the main paper. Let g_1 be a permutation of the categories of X_1 such that $\tilde{X}_1 = g_1(X_1)$ is a random variable with the categories $\{x_5^1, x_2^1, x_1^1, x_3^1, x_4^1\}$. Then the joint p.m.f. of $\tilde{X}_1$ and X_2 is given in Table S1 (a) with the p.m.f. of $\tilde{X}_1$, (2/8, 1/8, 2/8, 2/8, 1/8). Using the associated checkerboard copula regression of U_2 given $\tilde{U}_1$, we predict the categories of X_2 given the categories of $\tilde{X}_1$ in Table S1(b), which shows that the prediction of the category of X_2 is invariant over the permutation on the categories of X_1.

Table S1: **(a)** Joint p.m.f. of $(\tilde{X}_1, X_2)$; **(b)** Estimated category of X_2 for a category of $\tilde{X}_1$.

$\tilde{X}_1$ \ X_2	x_1^2	x_2^2	x_3^2
x_5^1	0	0	2/8
x_2^1	0	1/8	0
x_1^1	0	0	2/8
x_3^1	2/8	0	0
x_4^1	0	1/8	0

a

$\tilde{X}_1$	u_2^*	$f_{X_2\|\tilde{X}_1}$
x_5^1	6/8	x_3^2
x_2^1	3/8	x_2^2
x_1^1	6/8	x_3^2
x_3^1	1/8	x_1^2
x_4^1	3/8	x_2^2

b

Table S2 (a) shows the two-way ordinal table with the count $\{n_{ij}\}$ where $n_{ij} = 80 \times p_{ij}$ and p_{ij} is given in Table 1(a) of the main paper. In Table S2 (b) and S2 (c), we also predict the category of X_2 given a category of X_1 using the estimated checkerboard copula regression and quantify uncertainty of the predicted category of X_2 using the 1000 bootstrap resampling. These results support that the proposed method successfully identified the functional dependence in Table S2 (a).

Table S2: (a) Observed table $\{n_{ij}\}$; (b) $\hat{u}_2 = \hat{r}_{U_2|U_1}(\hat{u}_1)$ and the estimated category of X_2; (c) Proportion of each estimated category of X_2 for a category of X_1 using bootstrap resampling method.

X_1 \ X_2	x_1^2	x_2^2	x_3^2
x_1^1	0	0	20
x_2^1	0	10	0
x_3^1	20	0	0
x_4^1	0	10	0
x_5^1	0	0	20

a

X_1	$\hat{u}_2*$	$\hat{f}_{X_2\|X_1}$
x_1^1	6/8	x_3^2
x_2^1	3/8	x_2^2
x_3^1	1/8	x_1^2
x_4^1	3/8	x_2^2
x_5^1	6/8	x_3^2

b

X_1 \ X_2	x_1^2	x_2^2	x_3^2
x_1^1	0	0	100
x_2^1	0	100	0
x_3^1	100	0	0
x_4^1	0	100	0
x_5^1	0	0	100

c

S2. Simulation study

A simulation study is designed to evaluate the proposed checkerboard copula regression based association measure under different types of association scenarios in a $I_1 \times I_2$ ordinal contingency table with the (ordinal/nominal) explanatory variable X_1 and the ordinal response variable X_2.

S2.1. Simulation design for an ordinal explanatory variable X_1

We consider the following five simulation factors : (i) the type of association between X_1 and X_2, (ii) the magnitude of association, (iii) the marginal distributions of X_1 and X_2, (iv) the sample size n, and (v) the table size (values of I_1 and I_2). For the type of association, we considered four association patterns between X_1 with I_1 levels ($X_1 = 1, \ldots, I_1$) and X_2 with I_2 levels ($X_2 = 1, \ldots, I_2$):

- No association - X_1 and X_2 have no association;
- Linear pattern - X_2 increases linearly as X_1 increases linearly;
- Monotone nonlinear pattern - X_2 increases exponentially or logarithmically as X_1 increases linearly;
- Nonmonotone nonlinear pattern - X_2 increases quadratically as X_1 increases linearly.

To simulate the contingency table with the four types of association, we employed the proportional odds cumulative logit model (CLM), one of the widely used parametric ordinal response models:

$$
\begin{aligned}
\operatorname{logit}[P(X_2 \leq i_2 | X_1)] &= \alpha_{i_2}, \quad \text{for no association} \\
&= \alpha_{i_2} - \beta X_1, \quad \text{for linear and monotone nonlinear association} \\
&= \alpha_{i_2} - \beta_1 X_1 - \beta_2 X_1^2, \quad \text{for nonmonotone nonlinear association}
\end{aligned}
$$

where $i_2 = 1, 2, \ldots, I_2 - 1$, β, β_1 and β_2 are the regression coefficients, and α_{i_2}s are the intercepts satisfying $\alpha_1 < \ldots < \alpha_{i_2-1}$. As will be shown below, the regression coefficients and the intercepts will determine the second and third simulation factors, the magnitude of association and the marginal distribution of X_2, respectively. For the linear and monotone nonlinear association types, we used the CLM with a single predictor X_1 and the four values of β representing the four levels of magnitude of association, (weak, moderate, strong, very strong): $\beta = (0.25, 0.85, 1.4, 2)$. Note that $\exp(\beta) = \exp\left(\operatorname{logit}[P(X_2 \leq i_2 | X_1 = i_1)] -\right.$

$\text{logit}[P(X_2 \leq i_2 | X_1 = i_1 - 1)]$) represents the constant value of odds ratio for the $(I_1 - 1)(I_2 - 1)$ separate 2×2 tables obtained by taking all pairs of adjacent rows and all binary of collapsings of X_2 [1]. The zero value of β (i.e., $\beta = 0$) represents no association between X_1 and X_2. For the nonmonotone nonlinear association, we employed the CLM with linear and quadratic terms of X_1 and $(\beta_1, \beta_2) = (18, 3)$.

In order to obtain the intended types of association, the marginal distributions of X_1 and X_2 also need to be specified. For no association and linear/nonmonotone nonlinear association, the discrete uniform distribution was considered for X_1 and X_2, denoted as (X_1, X_2) = (Uniform$(1, I_1)$, Uniform$(1, I_2)$). For monotone nonlinear association, we considered the discrete uniform distribution for X_2 and two non-uniform distributions for X_1, the Binomial distribution with the number of trials equal to I_1 and the probability equal to $(0.2, 0.8)$: (X_1, X_2) = (Binomial$(I_1, 0.2)$, Uniform$(1, I_2)$) (X_2 increases logarithmically as X_1 increases) and (X_1, X_2) = (Binomial$(I_1, 0.8)$, Uniform$(1, I_2)$) (X_2 increases exponentially as X_1 increases). The reasons for such choice are twofold. First, we want to simulate data showing clear monotone nonlinear patterns. Second, we need to obtain reliable maximum likelihood (ML) estimation for the proportional odds CLM as the ML estimation of logit model parameters could break down for the contingency tables with sparse counts or certain patterns of sampling zero counts (see [2] and [3, 4]).

From the CLMs specified above, we simulated data of size n=(500,1000,2000) and created $I_1 \times I_2$ contingency tables whose sizes are 3×3, 3×5, 5×3, and 5×5 ($I_1 = 3, 5$ and $I_2 = 3, 5$).

The number of experimental conditions under each of four association types is

- No association:12 (= 1 association level $\times$ 1 marginal distribution of (X_1, X_2) $\times$ 3 sample sizes × 4 table sizes);

- Linear pattern: 48 (= 4 association levels × 1 marginal distribution of (X_1, X_2) × 3 sample sizes × 4 table sizes);
- Monotone nonlinear pattern: 96 (= 4 association levels × 2 marginal distributions of (X_1, X_2) × 3 sample sizes × 4 table sizes);
- Nonmonotone nonlinear pattern: 12 (= 1 association level × 1 marginal distribution of (X_1, X_2) × 3 sample sizes × 4 table sizes)

Under each experimental condition for each association type, we simulated 1000 contingency tables, and then calculated the proposed association measure in Eq. (13) of the main paper, $\hat{\rho}^2_{X_1 \to X_2}$. We presented the simulation results using the boxplots.

S2.1.1. Parameter values for no association and linear/monotone nonlinear association

To simulate the contingency tables with no association and linear/monotone nonlinear association, we apply the cumulative logit model (CLM) with a single predictor,

$$g_{i_2}(X_1) = \log\left[\frac{P(X_2 \le i_2|X_1)}{P(X_2 > i_2|X_1)}\right] = \alpha_{i_2} - \beta X_1, \qquad i_2 = 1, \ldots, I_2 - 1. \tag{S2.1}$$

For no association and linear association, we generate X_1 uniformly distributed over $1, \ldots, I_1$, denoted as Uniform$(1, I_1)$, and select the values of α_{i_2}s so that the marginal distribution of X_2 is uniformly distributed over $1, \ldots, I_2$, denoted as Uniform$(1, I_2)$. The uniform distribution of X_1 and X_2 provided the clear linear pattern in simulated $I_1 \times I_2$ contingency tables. In Table S3, we list the values of the parameters (α_{i_2}, β) of the proportional odds CLMs used for the simulation of the data with no association and linear pattern, and the distribution of X_1 for each table size.

Table S3: No association and linear association for the CLM with an ordinal X_1: parameters α_{i_2} and $\beta = (0, 0.25, 0.85, 1.4, 2)$ for no, weak, moderate, strong and very strong association, respectively, and the distribution of X_1 in each table size

.

Table size $I_1 \times I_2$	CLM model	Distribution of X_1	α_{i_2}s for No, Weak, Moderate, Strong, Very Strong association
3×3	$g_{i_2}(X_1) = \alpha_{i_2} - \beta X_1, i_2 = 1, 2.$	Uniform on $\{1, 2, 3\}$	$\alpha_{i_2} = [-0.69, 0.69]$, $\alpha_{i_2} = [-0.21, 1.2]$, $\alpha_{i_2} = [0.93, 2.48]$ $\alpha_{i_2} = [1.9, 3.7]$, $\alpha_{i_2} = [2.9, 5.1]$
3×5	$g_{i_2}(X_1) = \alpha_{i_2} - \beta X_1, i_2 = 1, 2, 3, 4.$	Uniform on $\{1, 2, 3\}$	$\alpha_{i_2} = [-1.39, -0.41, 0.41, 1.39]$, $\alpha_{i_2} = [-0.9, 0.09, 0.91, 1.9]$, $\alpha_{i_2} = [0.17, 1.25, 2.16, 3.23]$ $\alpha_{i_2} = [1.05, 2.27, 3.33, 4.55]$, $\alpha_{i_2} = [1.90, 3.34, 4.66, 6.10]$
5×3	$g_{i_2}(X_1) = \alpha_{i_2} - \beta X_1, i_2 = 1, 2.$	Uniform on $\{1, 2, 3, 4, 5\}$	$\alpha_{i_2} = [-0.69, 0.69]$, $\alpha_{i_2} = [0.04, 1.47]$, $\alpha_{i_2} = [1.63, 3.47]$ $\alpha_{i_2} = [2.95, 5.45]$, $\alpha_{i_2} = [4.30, 7.69]$
5×5	$g_{i_2}(X_1) = \alpha_{i_2} - \beta X_1, i_2 = 1, 2, 3, 4.$	Uniform on $\{1, 2, 3, 4, 5\}$	$\alpha_{i_2} = [-1.39, -0.41, 0.41, 1.39]$, $\alpha_{i_2} = [-0.67, 0.33, 1.17, 2.17]$, $\alpha_{i_2} = [0.76, 2, 3.1, 4.33]$ $\alpha_{i_2} = [1.84, 3.46, 4.95, 6.55]$, $\alpha_{i_2} = [2.9, 5, 7, 9]$

To simulate contingency tables with monotone nonlinear association, we apply the CLM with a single predictor in Eq. (S2.1) and consider the discrete uniform distribution for X_2 and two non-uniform distributions for X_1, the Binomial distribution with the number of trials equal to I_1 and the probability equal to $(0.2, 0.8)$: (X_1, X_2)=(Binomial$(I_1, 0.2)$, Uniform$(1, I_2)$) (X_2 increases logarithmically as X_1 increases) and (X_1, X_2) = (Binomial$(I_1, 0.8)$, Uniform$(1, I_2)$) (X_2 increases exponentially as X_1 increases). In Table S4, we list the values of the parameters (α_{i_2}, β) of the proportional odds CLMs used for the simulation of the data with monotone nonlinear association and (X_1, X_2)=(Binomial $(I_1, 0.8)$, Uniform$(1, I_2)$) in each table size.

Table S4: Monotone nonlinear association for CLM with an ordinal X_1: parameters α_{i_2} and $\beta = (0.25, 0.85, 1.4, 2)$ for weak, moderate, strong and very strong association levels, respectively, and (X_1, X_2)=(*Binomial* $(I_1, 0.8)$, Uniform$(1, I_2)$) in each table size.

Table size $I_1 \times I_2$	CLM model	Distribution of X_1	α_{i_2}s for Weak, Moderate, Strong, Very Strong association
3×3	$g_{i_2}(X_1) = \alpha_{i_2} - \beta X_1, i_2 = 1, 2.$	*Binomial*$(3, 0.8)$	$\alpha_{i_2} = [-0.05, 1.35]$, $\alpha_{i_2} = [1.49, 2.94]$ $\alpha_{i_2} = [2.88, 4.45]$, $\alpha_{i_2} = [4.39, 6.13]$
3×5	$g_{i_2}(X_1) = \alpha_{i_2} - \beta X_1, i_2 = 1, 2, 3, 4.$	*Binomial*$(3, 0.8)$	$\alpha_{i_2} = [-0.74, 0.24, 1.06, 2.04]$, $\alpha_{i_2} = [0.76, 1.79, 2.64, 3.66]$ $\alpha_{i_2} = [2.08, 3.21, 4.13, 5.2]$, $\alpha_{i_2} = [3.49, 4.76, 5.79, 6.91]$
5×3	$g_{i_2}(X_1) = \alpha_{i_2} - \beta X_1, i_2 = 1, 2.$	*Binomial*$(5, 0.8)$	$\alpha_{i_2} = [0.35, 1.75]$, $\alpha_{i_2} = [2.82, 4.34]$ $\alpha_{i_2} = [5.05, 6.79]$, $\alpha_{i_2} = [7.47, 9.51]$
5×5	$g_{i_2}(X_1) = \alpha_{i_2} - \beta X_1, i_2 = 1, 2, 3, 4.$	*Binomial*$(5, 0.8)$	$\alpha_{i_2} = [-0.35, 0.64, 1.46, 2.45]$, $\alpha_{i_2} = [2.05, 3.14, 4.03, 5.08]$ $\alpha_{i_2} = [4.18, 5.42, 6.44, 7.59]$, $\alpha_{i_2} = [6.44, 7.9, 9.11, 10.4]$

In Table S5, we list the values of the parameters (α_{i_2}, β) of the proportional odds CLMs used for the simulation of the data with monotone nonlinear association and (X_1, X_2)=(Binomial $(I_1, 0.2)$, Uniform$(1, I_2)$) in each table size.

Table S5: Monotone nonlinear association for CLM with an ordinal X_1 : parameters α_{i_2} and $\beta = (0.25, 0.85, 1.4, 2)$ for weak, moderate, strong and very strong association levels, respectively, and (X_1, X_2)=(*Binomial* $(I_1, 0.2)$, Uniform$(1, I_2)$) in each table size.

Table size $I_1 \times I_2$	CLM model	Distribution of X_1	α_{i_2}s for Weak, Moderate, Strong, Very Strong association
3×3	$g_{i_2}(X_1) = \alpha_{i_2} - \beta X_1, i_2 = 1, 2.$	*Binomial*$(3, 0.8)$	$\alpha_{i_2} = [-0.35, 1.05]$, $\alpha_{i_2} = [0.45, 1.91]$ $\alpha_{i_2} = [1.15, 2.73]$, $\alpha_{i_2} = [1.87, 3.61]$
3×5	$g_{i_2}(X_1) = \alpha_{i_2} - \beta X_1, i_2 = 1, 2, 3, 4.$	*Binomial*$(3, 0.8)$	$\alpha_{i_2} = [-1.04, -0.06, 0.76, 1.74]$, $\alpha_{i_2} = [-0.26, 0.76, 1.61, 2.64]$ $\alpha_{i_2} = [0.41, 1.47, 2.39, 3.52]$, $\alpha_{i_2} = [1.09, 2.21, 3.23, 4.51]$
5×3	$g_{i_2}(X_1) = \alpha_{i_2} - \beta X_1, i_2 = 1, 2.$	*Binomial*$(5, 0.8)$	$\alpha_{i_2} = [-0.25, 1.15]$, $\alpha_{i_2} = [0.75, 2.28]$ $\alpha_{i_2} = [1.61, 3.35]$, $\alpha_{i_2} = [2.5, 4.53]$
5×5	$g_{i_2}(X_1) = \alpha_{i_2} - \beta X_1, i_2 = 1, 2, 3, 4.$	*Binomial*$(5, 0.8)$	$\alpha_{i_2} = [-0.94, 0.04, 0.86, 1.85]$, $\alpha_{i_2} = [0.02, 1.07, 1.96, 3.04]$ $\alpha_{i_2} = [0.8, 1.96, 2.98, 4.22]$, $\alpha_{i_2} = [1.6, 2.9, 4.1, 5.57]$

S2.1.2. Parameter values for nonmonotone nonlinear association

To simulate the $I_1 \times I_2$ contingency table with clear nonmonotone nonlinear association, we consider the CLM with linear and quadratic terms of X_1 in Eq. (S2.2),

$$g_{i_2}(X_1) = \log\left[\frac{P(X_2 \leq i_2|X_1)}{P(X_2 > i_2|X_1)}\right] = \alpha_{i_2} - 18X_1 - 3X_1^2, \qquad i_2 = 1, \ldots, I_2 - 1. \tag{S2.2}$$

We generate $X_1 \sim$ Uniform$(1, I_1)$, and select the values of α_{i_2}s so that the marginal distribution of X_2 is Uniform$(1, I_2)$: $\alpha_{i_2} = [-10.92, -8.91]$ for the 3×3 table, $\alpha_{i_2} = [-11.98, -10.46, -9.28, -8.11]$ for the 3×5 table, $\alpha_{i_2} = [-25.49, -19.45]$ for the 5×3 table, and $\alpha_{i_2} = [-26.82, -24.84, -22.54, -15.42]$ for the 5×5 table.

S2.2. Simulation design for a nominal explanatory variable X_1

In order to simulate a two-way contingency table for a nominal explanatory variable X_1 (with I_1 levels) and an ordinal response variable X_2 (with I_2 levels), we considered the proportional odds cumulative logit model (CLM) where X_1 is treated as a factor with $(I_1 - 1)$ indicator variables:

$$g_{i_2}(X_1) = \text{logit}[P(X_2 \le i_2|X_1)] = \alpha_{i_2} + \tau_1 w_1 + \ldots + \tau_{i_1} w_{i_1} + \ldots + \tau_{I_1-1} w_{I_1-1}$$

where $i_2 = 1, 2, \ldots, I_2 - 1$, α_{i_2}s are the intercepts satisfying $\alpha_1 < \ldots < \alpha_{I_2-1}$, $w_1, \ldots, w_{I_1-1}$ are the indicator variables for each level of X_i (i.e., $w_{i_1} = 1$ for an observation from the i_1-th level of X_1 and $w_{i_1} = 0$ otherwise), and $\tau_1, \ldots, \tau_{I_1-1}$ are the parameters determining the effect of each level of X_i. Note that, for identifiability, we employed the constraint $\tau_{I_1} = 0$ (i.e., the effect parameter for the last category of X_1 being equal to 0) and so the last category of X_1 was used as the reference category. As will be shown below, the effect parameters and the intercepts will determine the magnitude of association between X_1 and X_2, and the marginal distribution of X_2, respectively.

Under the CLM specified above, we considered the following three simulation factors : (i) the table size (values of I_1 and I_2), (ii) the sample size n, and (iii) the magnitude of association. We first used four different table sizes and three different sample sizes : $(I_1, I_2) = (3, 3), (3, 5), (5, 3), (5, 5)$ and n=(500,1000,2000). For the magnitude of association, we used five association levels, (no, weak, moderate, strong, very strong), as shown in Table S6. Note that τ_{i_1} determines the cumulative odds ratios comparing the i_1-th category and the reference category of X_1 : $\text{logit}[P(X_2 \le i_2|X_1 = i_1)] - \text{logit}[P(X_2 \le i_2|X_1 = I_1)] = \tau_{i_1} - \tau_{I_1} = \tau_{i_1}$. At last, we employed the discrete uniform distribution for the marginal distribution of X_1 and the values of α_{i_2} were chosen so that X_2 is also discretely uniform distributed in $[1, I_2]$.

For each combination of three simulation factors (sample size, table size and degree of association), we simulated 1000 contingency tables and computed the proposed association measure in Eq. (13) of the main paper, $\hat{\rho}^2_{X_1 \to X_2}$. We presented the simulation results using the boxplots.

Table S6: CLM with a nominal X_1 - the distribution of X_1 and (α_{i_2}, τ_{i_1}) for no, weak, moderate, strong and very strong association, respectively, in each table size

Table size $I_1 \times I_2$	CLM model	Distribution of X_1	α_{i_2}s for No, Weak, Moderate, Strong, Very Strong association
3×3	$g_{i_2}(X_1) = \alpha_{i_2} + z_1\tau_1 + z_2\tau_2, i_2 = 1, 2.$	Uniform on $\{1, 2, 3\}$	$\alpha_{i_2} = [-0.69, 0.69]$, $\tau_{i_1} = [0, 0]$ $\alpha_{i_2} = [-0.70, 0.70]$, $\tau_{i_1} = [0.25, -0.25]$ $\alpha_{i_2} = [-0.77, 0.77]$, $\tau_{i_1} = [0.85, -0.85]$ $\alpha_{i_2} = [-0.91, 0.91]$, $\tau_{i_1} = [1.4, -1.4]$ $\alpha_{i_2} = [-1.11, 1.11]$, $\tau_{i_1} = [2, -2]$
3×5	$g_{i_2}(X_1) = \alpha_{i_2} + z_1\tau_1 + z_2\tau_2, i_2 = 1, 2, 3, 4.$	Uniform on $\{1, 2, 3\}$	$\alpha_{i_2} = [-1.39, -0.41, 0.41, 1.39]$, $\tau_{i_1} = [0, 0]$ $\alpha_{i_2} = [-1.40, -0.41, 0.41, 1.40]$, $\tau_{i_1} = [0.25, -0.25]$ $\alpha_{i_2} = [-1.53, -0.45, 0.45, 1.53]$, $\tau_{i_1} = [0.85, -0.85]$ $\alpha_{i_2} = [-1.76, -0.53, 0.53, 1.76]$, $\tau_{i_1} = [1.4, -1.4]$ $\alpha_{i_2} = [-2.10, -0.66, 0.66, 2.10]$, $\tau_{i_1} = [2, -2]$
5×3	$g_{i_2}(X_1) = \alpha_{i_2} + z_1\tau_1 + z_2\tau_2 + z_3\tau_3 + z_4\tau_4, i_2 = 1, 2.$	Uniform on $\{1, 2, 3, 4, 5\}$	$\alpha_{i_2} = [-0.69, 0.69]$, $\tau_{i_1} = [0, 0, 0, 0]$ $\alpha_{i_2} = [-0.72, 0.72]$, $\tau_{i_1} = [0.55, 0.25, -0.25, -0.55]$ $\alpha_{i_2} = [-0.82, 0.82]$, $\tau_{i_1} = [1.1, 0.85, -0.85, -1.1]$ $\alpha_{i_2} = [-1.02, 1.02]$, $\tau_{i_1} = [1.7, 1.4, -1.4, -1.7]$ $\alpha_{i_2} = [-1.31, 1.31]$, $\tau_{i_1} = [2.3, 2, -2, -2.3]$
5×5	$g_{i_2}(X_1) = \alpha_{i_2} + z_1\tau_1 + z_2\tau_2 + z_3\tau_3 + z_4\tau_4, i_2 = 1, 2, 3, 4.$	Uniform on $\{1, 2, 3, 4, 5\}$	$\alpha_{i_2} = [-1.39, -0.41, 0.41, 1.39]$, $\tau_{i_1} = [0, 0, 0, 0]$ $\alpha_{i_2} = [-1.43, -0.42, 0.42, 1.43]$, $\tau_{i_1} = [0.55, 0.25, -0.25, -0.55]$ $\alpha_{i_2} = [-1.61, -0.48, 0.48, 1.61]$, $\tau_{i_1} = [1.1, 0.85, -0.85, -1.1]$ $\alpha_{i_2} = [-1.93, -0.61, 0.61, 1.93]$, $\tau_{i_1} = [1.7, 1.4, -1.4, -1.7]$ $\alpha_{i_2} = [-2.37, -0.79, 0.79, 2.37]$,$\tau_{i_1} = [2.3, 2, -2, -2.3]$

S2.3. Simulation results

S2.3.1. $I_1 \times I_2$ table with an ordinal explanatory variable

Fig.S3 presents the boxplots of the proposed measure $\hat{\rho}^2_{X_1 \to X_2}$ from the simulated $I_1 \times I_2$ tables with no association for three sample sizes (n=500 (white), 1000 (light grey), 2000 (dark grey) in each panel). Fig.S4-S6 show the boxplots of the proposed measure $\hat{\rho}^2_{X_1 \to X_2}$ from the simulated $I_1 \times I_2$ tables under four association levels (from weak (leftmost panel) to very strong (rightmost panel)) and the three sample sizes (n=500 (white), 1000 (light grey), 2000 (dark grey) in each panel). Fig.S4 is concerned with the results from the contingency tables with linear association pattern. Fig.S5 and S6 are concerned with the contingency tables showing monotone nonlinear pattern with the marginal distribution (X_1, X_2)=(Binomial$(I_1, 0.8)$, Uniform$(1, I_2)$) and (X_1, X_2)=(Binomial$(I_1, 0.2)$, Uniform$(1, I_2)$), respectively. Fig.S7 shows boxplots based on the simulated tables with the nonmonotone nonlinear pattern.

From the simulation results, we have made the following observations. First, the variability of the distribution for $\hat{\rho}^2_{X_1 \to X_2}$ decreases as the sample size increases and the center of the distribution is stable over different sample sizes, regardless of the table size and association pattern. Second, for the case of no association, the values of $\hat{\rho}^2_{X_1 \to X_2}$ are very close to zero (their ranges are less than 0.05 and their medians are less than 0.01), regardless of the table size and the sample size. The sampling distributions of $\hat{\rho}^2_{X_1 \to X_2}$ are right-skewed, but the amount of skewness decreases as the sample size increases. Third, for tables with linear and monotone nonlinear pattern, as the magnitude of association increase from weak to very strong (i.e., β in CLM increases from 0.25 to 2), the magnitude of $\hat{\rho}^2_{(X_1 \to X_2)}$ also increases. Note that for weak association tables with linear and monotone nonlinear pattern, $\hat{\rho}^2_{X_1 \to X_2}$ are small (their medians are less than 0.05), regardless of the table size and the sample size. Last, for tables with linear/monotone nonlinear/nonmonotone nonlinear patterns, the proposed measure $\hat{\rho}^2_{(X_1 \to X_2)}$ generally increases as the size of the table increases (i.e., either I_1 or I_2 increases) and the increment of I_1 leads to larger value of $\hat{\rho}^2_{(X_1 \to X_2)}$ than the increment of I_2 : $\hat{\rho}^2_{(X_1 \to X_2)}$ (3×3 table) < $\hat{\rho}^2_{(X_1 \to X_2)}$ (3×5 table) < $\hat{\rho}^2_{(X_1 \to X_2)}$ (5×3 table) < $\hat{\rho}^2_{(X_1 \to X_2)}$ (5×5 table).

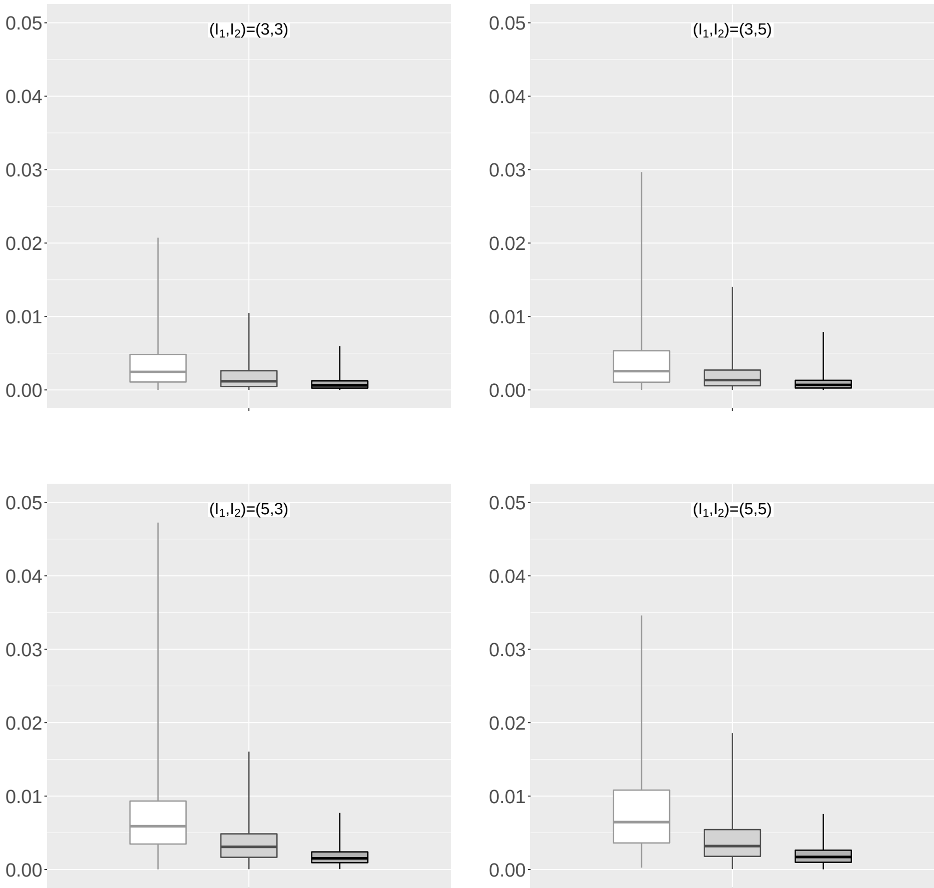

Fig. S3: Boxplots of $\hat{\rho}^2_{(X_1 \to X_2)}$ for $I_1 \times I_2$ tables simulated from the CLM with an ordinal predictor X_1 - no association and the sample sizes n = 500(white), 1000(light grey), 2000(dark grey).

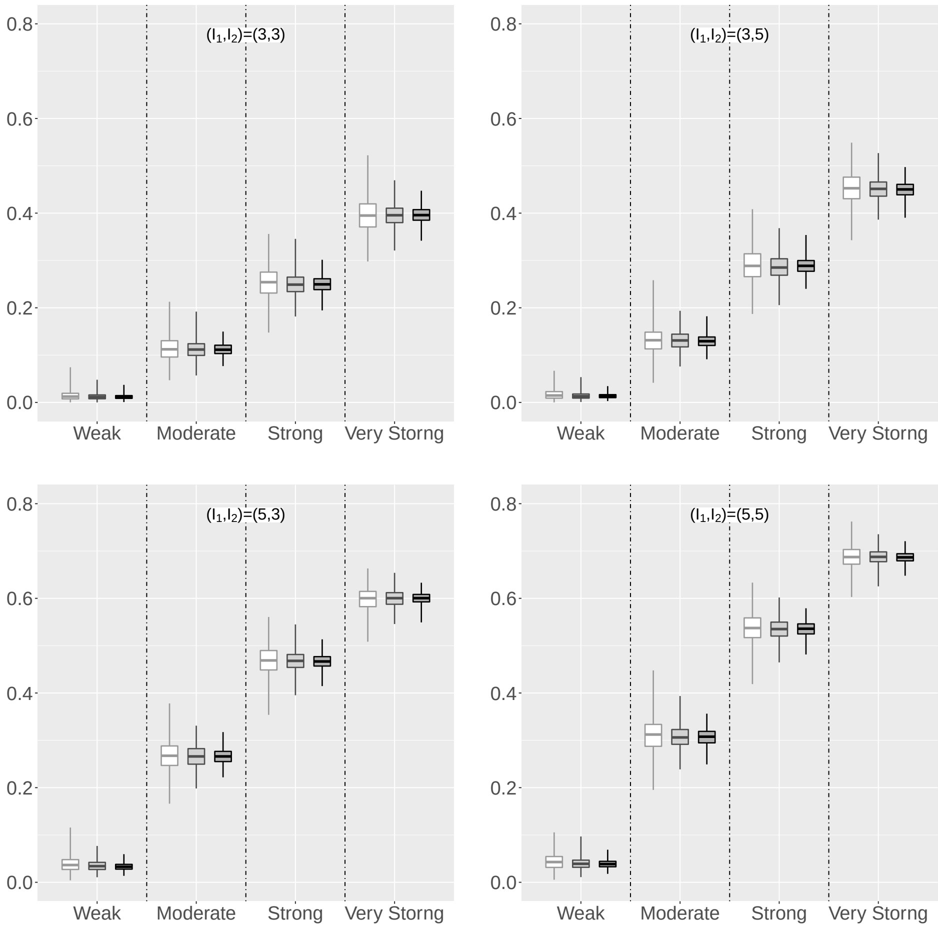

Fig. S4: Boxplots of $\hat{\rho}^2_{(X_1 \to X_2)}$ for $I_1 \times I_2$ tables simulated from the CLM with an ordinal predictor X_1 - the weak(leftmost panel), moderate(left middle panel), strong(right middle panel), and very strong(rightmost panel) linear association levels and the sample sizes $n = 500$(white), 1000(light grey), 2000(dark grey).

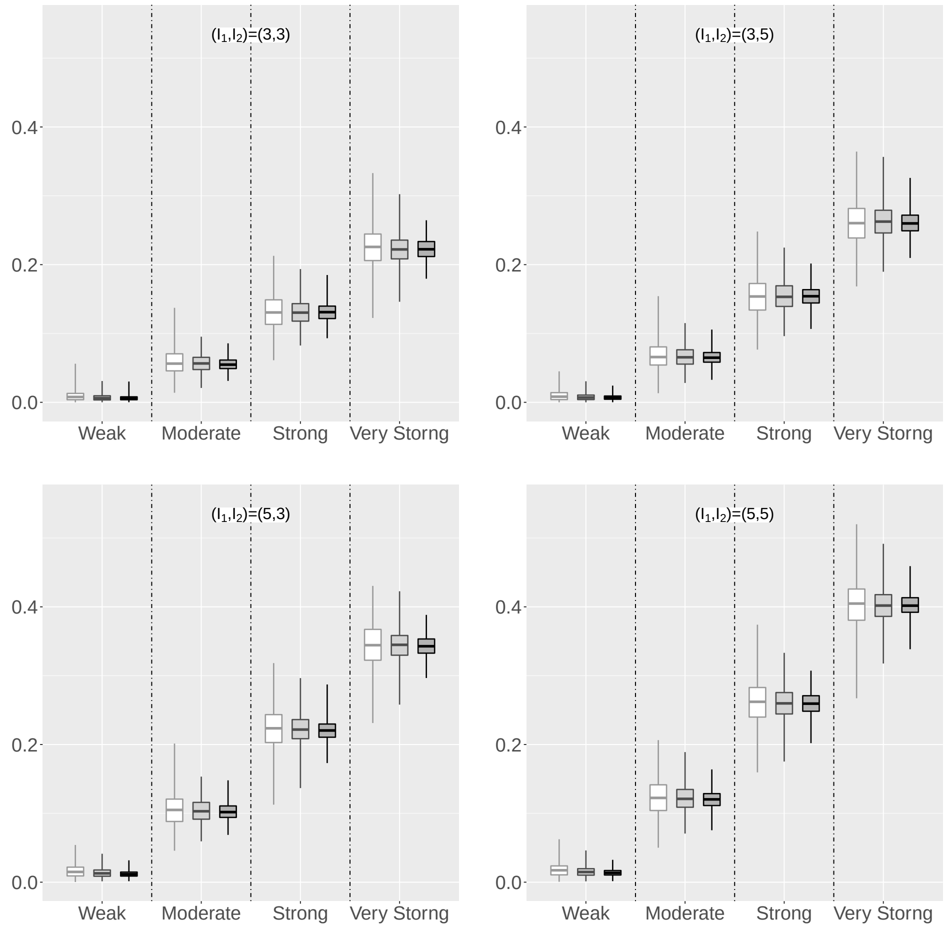

Fig. S5: Boxplots of $\hat{\rho}^2_{(X_1\to X_2)}$ for $I_1 \times I_2$ tables simulated from the CLM with an ordinal predictor X_1 and the marginal distribution (X_1, X_2)=(Binomial$(I_1, 0.8)$, Uniform$(1, I_2)$) - the weak(leftmost panel), moderate(left middle panel), strong(right middle panel), and very strong(rightmost panel) monotone non-linear association levels and the sample sizes $n = 500$(white), 1000(light grey), 2000(dark grey).

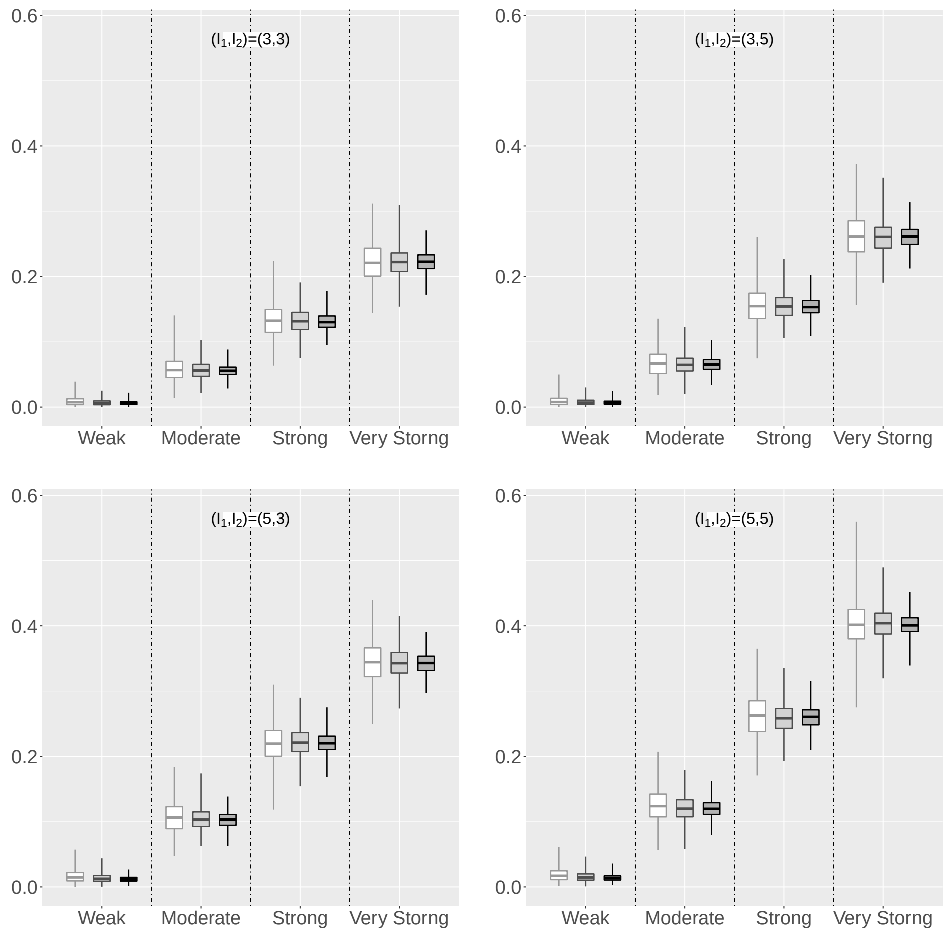

Fig. S6: Boxplots of $\hat{\rho}^2_{(X_1 \to X_2)}$ for $I_1 \times I_2$ tables simulated from the CLM with an ordinal predictor X_1 and the marginal distribution (X_1, X_2)=(Binomial(I_1, 0.2), Uniform(1, I_2)) - the weak(leftmost panel), moderate(left middle panel), strong(right middle panel), and very strong(rightmost panel) monotone non-linear association levels and the sample sizes n = 500(white), 1000(light grey), 2000(dark grey).

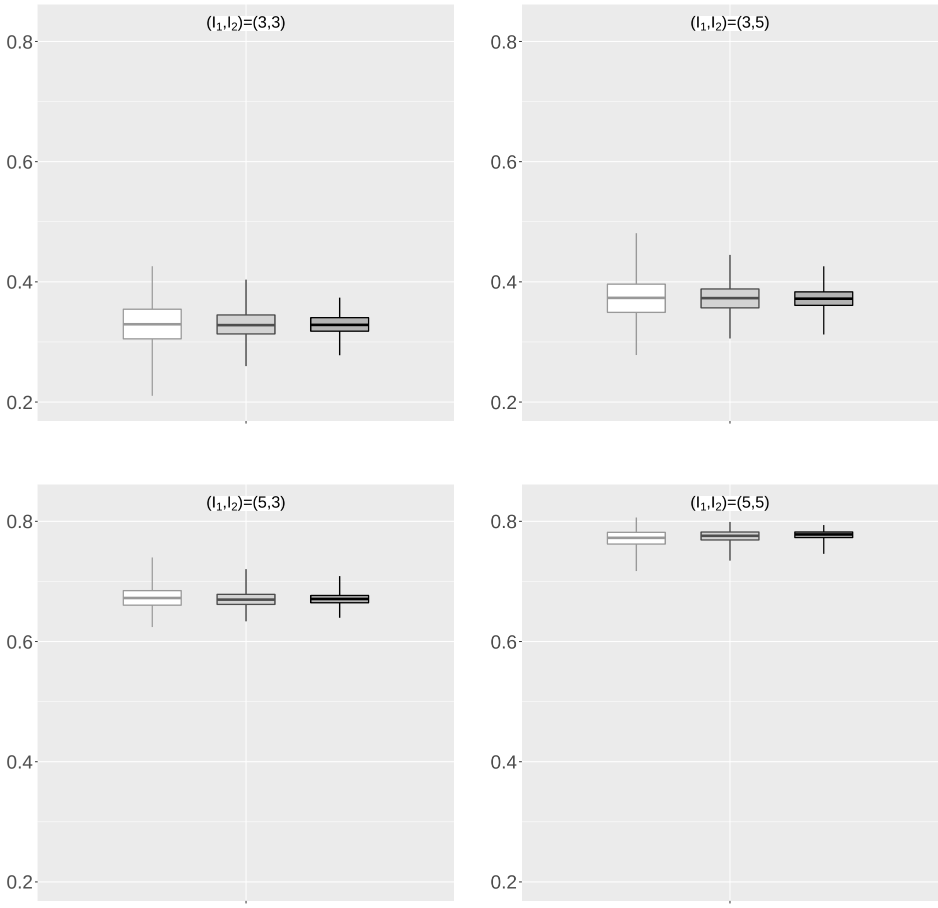

Fig. S7: Boxplots of $\hat{\rho}^2_{(X_1\to X_2)}$ for $I_1 \times I_2$ tables simulated from the CLM with an ordinal predictor X_1 - the nonmonotone non-linear association levels and the sample sizes n = 500(white), 1000(light grey), 2000(dark grey).

S2.3.2. $I_1 \times I_2$ table with a nominal explanatory variable

Fig.S8 presents the boxplots of the proposed measure $\hat{\rho}^2_{X_1 \to X_2}$ from the simulated $I_1 \times I_2$ tables with a nominal X_1 and no association for three sample sizes (n=500 (white), 1000 (light grey), 2000 (dark grey) in each panel). Fig.S9 shows the boxplots of the proposed measure $\hat{\rho}^2_{X_1 \to X_2}$ from the simulated $I_1 \times I_2$ tables under four association levels (from weak (leftmost panel) to very strong (rightmost panel)) and the three sample sizes (n=500 (white), 1000 (light grey), 2000 (dark grey)).

We first observe that the variance of the distribution for $\hat{\rho}^2_{X_1 \to X_2}$ decreases as the sample size increases and the center of the distribution is stable over different sample sizes, regardless of the table size and the magnitude of association. When there is no association, Fig.S8 shows that the sampling distributions of $\hat{\rho}^2_{X_1 \to X_2}$ are right-skewed, but the degree of skewness decreases as the sample size increases. Note that the values of $\hat{\rho}^2_{X_1 \to X_2}$ are very close to zero (their ranges are less than 0.05), regardless of the table size and the sample size. As shown in Fig.S9, the magnitude of $\hat{\rho}^2_{(X_1 \to X_2)}$ increases as the magnitude of association increase from weak to very strong. Notice that the proposed measure $\hat{\rho}^2_{(X_1 \to X_2)}$ tends to increase as the size of the table increases (i.e., either I_1 or I_2 increases).

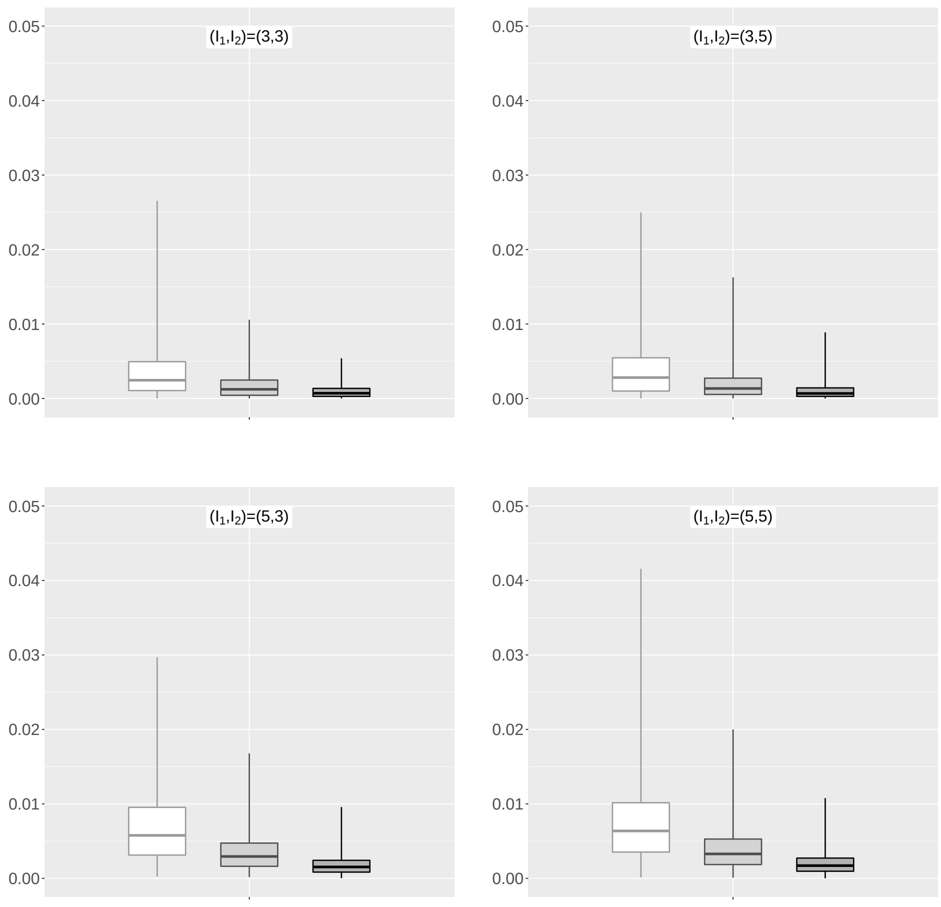

Fig. S8: Boxplots of $\hat{\rho}^2_{(X_1\rightarrow X_2)}$ for $I_1 \times I_2$ tables simulated from the CLM with a nominal predictor X_1 - no association and the sample sizes n = 500(white), 1000(light grey), 2000(dark grey).

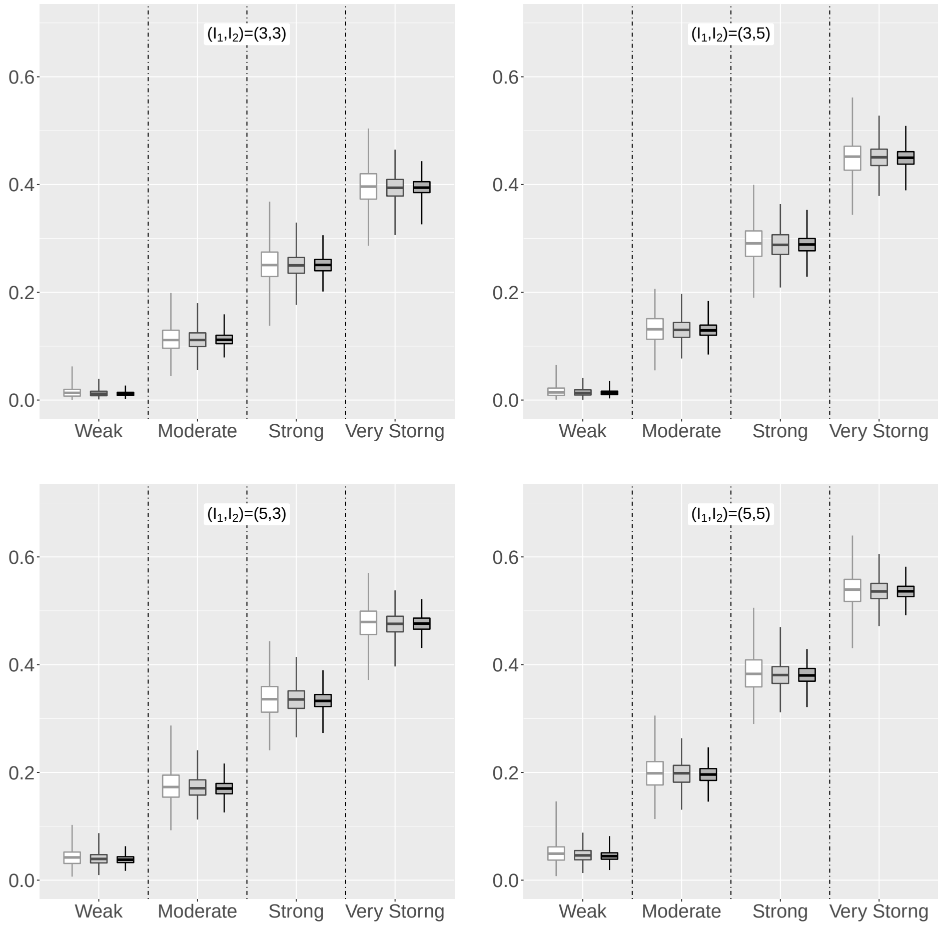

Fig. S9: Boxplots of $\hat{\rho}^2_{(X_1 \to X_2)}$ for $I_1 \times I_2$ tables simulated from the CLM with a nominal predictor X_1 and the marginal distribution (X_1, X_2)=(Uniform$(1, I_1)$, Uniform$(1, I_2)$) - the weak(leftmost panel), moderate(left middle panel), strong(right middle panel), and very strong(rightmost panel) association levels and the sample sizes n = 500(white), 1000(light grey), 2000(dark grey).

References

[1] A. Agresti, Analysis of Ordinal Categorical Data, Wiley, New York, 2 edition, 2010.

[2] A. Agresti, M. Kateri, Categorical Data Analysis, Springer, Berlin, 2011.

[3] S. J. Haberman, The analysis of residuals in cross-classified tables, Biometrics 29 (1973) 205–220.

[4] S. J. Haberman, Log-linear models and frequency tables with small expected cell counts, The Annals of Statistics 5 (1977) 1148–1169.