Modeling Association Between Two or More Multiple-Response Categorical Variables

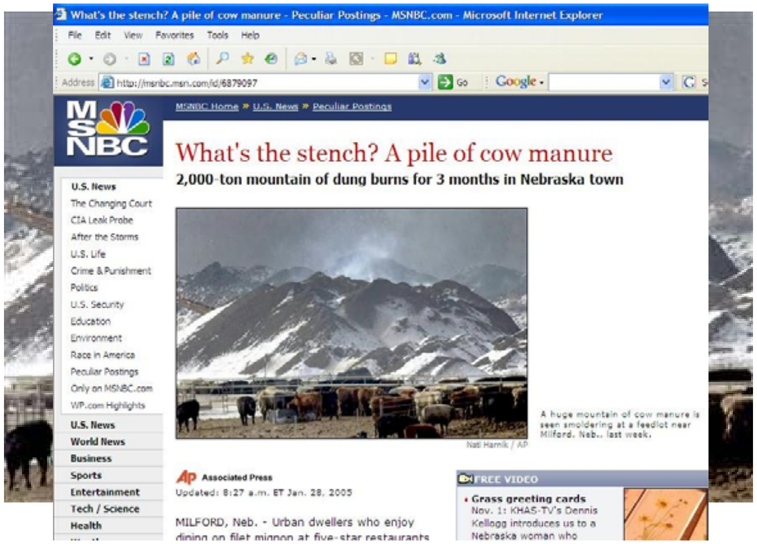
Christopher R. Bilder
Department of Statistics
University of Nebraska-Lincoln
www.chrisbilder.com
chris@chrisbilder.com

This research was supported in part by National Science Foundation grant SES-0207212

Introduction

Introduction

There is no place like Nebraska!



What's the stench? A pile of cow manure

2,000-ton mountain of dung burns for 3 months in Nebraska town

A huge mountain of cow manure is seen smoldering at a feedlot near Milford, Neb., last week.

Nati Harnik / AP

Associated Press

Updated: 8:27 a.m. ET Jan. 28, 2005

MILFORD, Neb. - Urban dwellers who enjoy dining on filet mignon at five-star restaurants

Grass greeting cards
Nov. 1: KHAS-TV's Dennis Kellogg introduces us to a Nebraska woman who

---

Multiple-response categorical variables

- "Choose all that apply" or "pick any" from a set of *items*
  - Lead to multiple-response categorical variables (MRCVs)
- Examples
  - 1997 new Federal standards for ethnicity reporting (*Federal register*, 1997, p. 58781)
    - Choose all that apply from these "items":
      - American Indian or Alaskan Native
      - Asian
      - Black or African American
      - Native Hawaiian or Other Pacific Islander
      - White
      - Some Other Race
    - **Individuals may choose more than one race!**
    - Census 2000

Multiple-response categorical variables

- Examples (continued)
  - Marketing research studies (Chambers and Skinner, 2003)
    - Consumer choices among pop (Holbrook et al., 1982)
      - Coke, Pepsi, Sprite, …
    - Perceptions about quality of car manufacturers (Umesh, 1995)
      - Toyota, GM, Ford, …
  - Contraceptive use studies (Foxman et al., 1997)
    - Examine urinary tract infection and contraception method used by women
- Positive/negative responses to each item
  - Correlated binary random variables

Kansas farmer example

- Survey of 279 Kansas farmers conducted by Kansas State University
- What swine waste disposal methods do you use? Pick all that apply:
  - Lagoon
  - Pit
  - Natural drainage
  - Holding tank
- What do you test swine waste for? Pick all that apply:
  - Nitrogen
  - Phosphorus
  - Salt

Kansas farmer example

- Observed counts

		Waste storage method chosen			
		Lagoon	Pit	Natural Drainage	Holding Tank
Test waste chosen	Nitrogen	27	16	2	2
	Phosphorus	22	12	1	1
	Salt	19	6	1	0

- Questions of interest**:**
  - Is waste storage independent of what the waste is tested for?
  - If they are dependent, what is the association structure?
    - Does some waste storage methods lead to more or less testing than others?
    - Are there particular storage/contaminant combinations for which there is more or less testing than for others?

Kansas farmer example

- What makes this problem unique?
  - Both questions result in multiple-response categorical variables (MRCVs)
  - Farmers can be represented multiple times in the table
  - Usual independence testing or loglinear modeling methods should not be used on this type of data
    - Cell counts are correlated most likely in a non-multinomial way
    - Margins do not add to proper totals

		Waste storage method chosen			
		Lagoon	Pit	Natural Drainage	Holding Tank
Test waste chosen	Nitrogen	27	16	2	2
	Phosphorus	22	12	1	1
	Salt	19	6	1	0

Alternative Representation

- Item response table – Pairwise cross-classification of all item responses

			Waste storage methods							
			Lagoon		Pit		Natural Drainage		Holding Tank	
			1	0	1	0	1	0	1	0
Test waste for	Nitrogen	1	27	13	16	24	2	38	2	38
		0	116	123	64	175	83	156	11	228
	Phosphorus	1	22	8	12	18	1	29	1	29
		0	121	128	68	181	84	165	12	237
	Salt	1	19	2	6	15	1	20	0	21
		0	124	134	74	184	84	174	13	245

Responses:
1=positive
0=negative

  - Each blue 2×2 "subtable" represents all 279 farmers
  - Previous table
    - Reports just (1,1) cell
    - Leads to non-invariant statistics
  - Leads naturally to examination of associations between Waste Storage items and Test Waste items

Summary of Past Research on MRCVs

- Focus has been on testing independence
  - Loughin, T. M. and Scherer, P. N. (1998). Testing for association in contingency tables with multiple column responses. *Biometrics* 54, 630-637.
  - Bilder, C. R. and Loughin, T. M. (2002). Testing for Conditional Multiple Marginal Independence. *Biometrics* 58. 200-208.
  - Bilder, C. R. and Loughin, T. M. (2004). Testing for Marginal Independence Between Two Categorical Variables with Multiple Responses. *Biometrics* 60, 241-8.
- Limited efforts to *model* association
  - Agresti and Liu (*Biometrics*, 1999, and *Sociological Methods & Research*, 2001)
    - Suggest using generalized loglinear models fit via MLE (Lang and Agresti, *JASA,* 1994)
    - Problems with achieving convergence for parameter estimates

Goals

- Develop models to describe association between two MRCVs
  - "Association" is defined by odds ratios within the subtables of the item response table
  - Assign parameters to control odds ratios within subtables
  - Develop inference procedures for models
- Extend models to allow more than two MRCVs

Notation

			Waste storage methods							
			Lagoon		Pit		Natural Drainage		Holding Tank	
			1	0	1	0	1	0	1	0
Test waste for	Nitrogen	1	27	13	16	24	2	38	2	38
		0	116	123	64	175	83	156	11	228
	Phosphorus	1	22	8	12	18	1	29	1	29
		0	121	128	68	181	84	165	12	237
	Salt	1	19	2	6	15	1	20	0	21
		0	124	134	74	184	84	174	13	245

- Focus on 2 MRCVs
- W denotes the "row" MRCV
  - W = contaminants tested
- Y denotes the "column" MRCV
  - Y = waste storage method
- W_i for i=1,..,I denotes the row variable items (levels)
  - W_1 is nitrogen, W_2 is phosphorous, W_3 is salt
  - W_i = 1 if subject picks item (positive response)
    W_i = 0 if subject does not pick item (negative response)
- Y_j for j=1,…,J is similarly defined for the column items
- n denotes the number of subjects in a simple random sample

Notation

- $m_{ab(ij)}$ is the number of (W_i=a, Y_j=b) responses where a = 0 or 1 and b = 0 or 1
  - $m_{11(31)}$ = 19 farmers who test waste for salt and also use lagoon as their waste storage method
  - $E(m_{ab(ij)}) = \mu_{ab(ij)}$

			Waste storage methods							
			Lagoon		Pit		Natural Drainage		Holding Tank	
			1	0	1	0	1	0	1	0
Test waste for	Nitrogen	1	27	13	16	24	2	38	2	38
		0	116	123	64	175	83	156	11	228
	Phosphorus	1	22	8	12	18	1	29	1	29
		0	121	128	68	181	84	165	12	237
	Salt	1	19	2	6	15	1	20	0	21
		0	124	134	74	184	84	174	13	245

- $\theta_{ij} = \mu_{11(ij)}\mu_{00(ij)}/(\mu_{10(ij)}\,\mu_{01(ij)})$ is the population odds ratio in subtable (i,j)
  - $\tilde{\theta}_{ij} = m_{11(ij)}m_{00(ij)}/(m_{10(ij)}\,m_{01(ij)})$ is the empirical odds ratio in subtable (i,j)

Model Development: Loglinear Models

- Consider a single subtable (items W_i and Y_j)
  - Loglinear model for counts in a table is $m_{ab} \sim \text{Poisson}(\mu_{ab})$, where
    $$\log(\mu_{ab}) = \gamma + \eta_a^W + \eta_b^Y + \lambda_{ab}^{WY}$$
    - Association controlled through λ_{ab}^{WY}
    - Other terms force predicted margins to match observed
  - Set-last-to-zero estimability restrictions $\Rightarrow \log(\theta) = \lambda_{00}^{WY}$ where θ is the odds ratio
  - Independence between W_i and $Y_j \Leftrightarrow \theta$ =1, or $\log(\mu_{ab}) = \gamma + \eta_a^W + \eta_b^Y$
- Extend this model to cover all subtables simultaneously
  - Estimate model parameters from entire item response table
  - Model association parameters according to effects of W-items, Y-items, and interactions
  - Like factorial ANOVA, except modeling log-odds-ratios instead of means

---

Generalized loglinear model

- First, consider the case where there is independence in each subtable (all θ_{ij}=1).
  - This is called Simultaneous Pairwise Marginal Independence (SPMI) – Agresti and Liu (1999)
- Model under SPMI: $\log(\mu_{ab(ij)}) = \gamma_{ij} + \eta_{a(ij)}^W + \eta_{b(ij)}^Y$ for a=0,1, b=0,1, i=1,…,I, and j=1,…,J
  - For the W_i and Y_j subtable, it is the "usual" loglinear model under independence - $\log(\mu_{ab}) = \gamma + \eta_a^W + \eta_b^Y$
    - No association parameters anywhere!
  - Predicted subtable count margins match the observed subtable margins

---

Generalized loglinear model

- Non-SPMI models:
  - $\log(\mu_{ab(ij)}) = \gamma_{ij} + \eta_{a(ij)}^W + \eta_{b(ij)}^Y + \lambda_{ab}$
    - Homogenous association model
    - Odds ratios between the W_i and Y_j items all the same: $\log(\theta_{ij}) = \lambda_{00}$ for all (i,j) pairs
  - $\log(\mu_{ab(ij)}) = \gamma_{ij} + \eta_{a(ij)}^W + \eta_{b(ij)}^Y + \lambda_{ab} + \lambda_{ab(j)}^Y$
    - W-homogenous association model
    - Odds ratios between (W_i,Y_j) vary across the Y_j items only
    - $\log(\theta_{ij}) = \lambda_{00} + \lambda_{00(j)}^Y$
  - $\log(\mu_{ab(ij)}) = \gamma_{ij} + \eta_{a(ij)}^W + \eta_{b(ij)}^Y + \lambda_{ab} + \lambda_{ab(i)}^W$
    - Y-homogenous association model
    - Odds ratios between (W_i,Y_j) vary across the W_i items only
    - $\log(\theta_{ij}) = \lambda_{00} + \lambda_{00(i)}^W$

Generalized loglinear model

- Non-SPMI models (continued):
  - $\log(\mu_{ab(ij)}) = \gamma_{ij} + \eta^{W}_{a(ij)} + \eta^{Y}_{b(ij)} + \lambda_{ab} + \lambda^{W}_{ab(i)} + \lambda^{Y}_{ab(j)}$
    - Main-effects association model
    - Main effects of both W and Y on the odds ratios
    - Differences between log odds ratios for any two items of Y are constant across W and vice versa
  - $\log(\mu_{ab(ij)}) = \gamma_{ij} + \eta^{W}_{a(ij)} + \eta^{Y}_{b(ij)} + \lambda_{ab} + \lambda^{W}_{ab(i)} + \lambda^{Y}_{ab(j)} + \lambda^{WY}_{ab(ij)}$
    - Saturated model
    - No constraints on the odds ratios for the W_i and Y_j combinations
    - Model-predicted odds ratios match observed odds ratios in each subtable

Fitting the models

- Maximum likelihood estimation
  - Observe a vector of binary responses for each subject
    - $(W_1, \ldots, W_I, Y_1, \ldots, Y_J)$ – 2^{I+J} possible
    - Counts for each response combination are multinomial
    - Kansas farmer data

W_1	W_2	W_3	Y_1	Y_2	Y_3	Y_4	Count
0	0	0	0	0	0	0	1
0	0	0	0	0	0	1	9
0	0	0	0	0	1	0	69
⋮							⋮
1	1	1	1	1	1	0	1
1	1	1	1	1	1	1	0

  - Estimate the multinomial probability for each combination
    - Subject to marginal model constraints
      - Item response table is marginal summary of the multinomial counts
    - Lang and Agresti (*JASA*, 1994)
  - Large number of combinations (2^{I+J}) leads to sparseness
  - Convergence problems occur

Fitting the models

- Marginal estimation: estimating equations approach
  - Fit model directly to the item response table
    - Temporarily ignore that a subject contributes a response to EACH subtable
    - Treat the counts as coming from one multinomial distribution

			Waste storage methods							
			Lagoon		Pit		Natural Drainage		Holding Tank	
			1	0	1	0	1	0	1	0
Test waste for	Nitrogen	1	27	13	16	24	2	38	2	38
		0	116	123	64	175	83	156	11	228
	Phosphorus	1	22	8	12	18	1	29	1	29
		0	121	128	68	181	84	165	12	237
	Salt	1	19	2	6	15	1	20	0	21
		0	124	134	74	184	84	174	13	245

  - Parameter estimates result from maximizing the (incorrect) multinomial likelihood equations
    - $\mathbf{X}'\hat{\boldsymbol{\mu}} = \mathbf{X}'\mathbf{m}$
    - $\hat{\boldsymbol{\mu}}$ and **m** are $4IJ\times 1$ vectors of the corresponding $\hat{\mu}_{ab(ij)}$ and $m_{ab(ij)}$ quantities
    - **X** is a matrix of 0's and 1's relating the expected to the observed counts for a model

Fitting the models

- Marginal estimation (continued)
  - Fit the models using PROC GENMOD in SAS or glm in R
  - Parameter estimates
    - Called "pseudo" MLEs by Rao and Scott (*Annals of Statistics*, 1984) for a similar problem
      - Loglinear models for contingency table counts arising through complex survey sampling
      - True likelihood equations are not used
    - Consistent

Model comparison statistics

- Compare two nested models
  - H_o: Smaller model
  - H_a: Larger model
- Pearson and LRT like statistics
  - Pearson: $X^2 = \sum_{a,b,i,j}\left(\hat{\mu}^{(a)}_{ab(ij)} - \hat{\mu}^{(o)}_{ab(ij)}\right)^2 \Big/ \hat{\mu}^{(o)}_{ab(ij)}$
  - Generally will not have asymptotic χ^2 distributions because of the incorrect multinomial assumption
  - Asymptotic distribution is a linear combination of independent χ^2_1 random variables

Model comparison statistics

- Pearson and LRT statistics (continued)
  - First and second-order Rao-Scott (*Annals of Statistics*, 1984) adjustments can be applied
    - Adjusted statistics have asymptotic first and/or second order moments the same as a χ^2 random variable
    - Reject H_o if $X^2/d > \chi^2_{1-\alpha,\nu}$ where d is the adjustment
    - Past MRCV research has shown tests do not always hold the correct size
      - Especially for the first-order adjustment
      - Bilder, Loughin, and Nettleton (*Comm. in Stat.*, 2000) and Bilder and Loughin (*Biometrics*, 2002)

Model comparison statistics

- New bootstrap procedure
  1. Find predicted counts, $\hat{\mu}^{(o)}$and $\hat{\mu}^{(a)}$, from specified H_o and H_a models, respectively, and calculate the Pearson statistic, X^2
  2. Find observed 2×2 tables for each W_i & $W_{i'}$ (i<i′) and Y_j & $Y_{j'}$ (j<j′) response pair
  3. With $\hat{\mu}^{(o)}$and observed counts from 2., use the algorithm of Gange (*American Statistician*, 1995) to obtain the multinomial probability of each possible (W_1, …, W_I, Y_1, …, Y_J) combination under the H_o model
  4. Simulate B resamples of $(W_1^*,\ldots,W_I^*,Y_1^*,\ldots,Y_J^*)'$ using these multinomial probabilities
  5. Fit the models to each resample and calculate $X_b^{2^*}$ for b=1,…,B
  6. Calculate the p-value as $B^{-1}\sum_{b=1}^{B} I(X_b^{2^*} \geq X^2)$ where $I(\cdot)$ is the indicator function

Model comparison statistics

- What is the Gange algorithm?
  - Gange, S.J. (1995). Generating multivariate categorical variates using the iterative proportional fitting algorithm. *The American Statistician* 49, 134-138.
  - Method to generate vectors of correlated binary observations
  - Uses Iterative Proportional Fitting method
    - Fitting method for loglinear models
    - Specify marginal contingency tables – “configurations”
  - Model predicted sub-tables ($\hat{\mu}^{(o)}$) and observed 2×2 tables for each W_i & $W_{i'}$ (i<i′) and Y_j & $Y_{j'}$ (j<j′) response pair are used as the configurations
  - Obtain a 2^{I+J} vector of multinomial probabilities under the null hypothesis model

Follow-up analysis

- Absolute value of standardized Pearson residuals
  - Check fit of model
  - Asymptotic standard normal distribution approximation
- Model predicted odds ratios
  - One odds ratio per subtable
  - Asymptotic distribution and standard error can be derived

---

Kansas farmer example

- Goodness-of-fit results where H_a model is the saturated:

H_o Model	Pearson statistic	Bootstrap p-value	2nd-order Rao-Scott adj. p-value
SPMI	64.03	0.0006	<0.0001
Homogenous association	62.76	0.0004	<0.0001
W-homogenous association	5.34	0.0412	0.0691
Y-homogenous association	62.68	0.0002	<0.0001
Main-effects association	5.28	0.0306	0.0690

  - B = 5,000 resamples
- H_o: W-homogenous association
  H_a: Main-effects association
  - Bootstrap p-value = 0.5036
  - Consider W-homogenous association model further

$$\log(\mu_{ab(ij)}) = \gamma_{ij} + \eta^{W}_{a(ij)} + \eta^{Y}_{b(ij)} + \lambda_{ab} + \lambda^{Y}_{ab(j)}$$

Kansas farmer example

- Further investigation of W-homogenous association model:

			Waste storage methods			
			Lagoon	Pit	Natural Drainage	Holding Tank
Test waste for	Nitrogen	$\lvert r_{ab(ij)}\rvert =$	2.41	1.04	0.28	0.97
		$\tilde{\theta}_{obs,ij} =$	2.20	1.82	0.10	1.09
		$C.I._{obs} =$	(1.22, 3.99)	(1.02, 3.26)	(0.03, 0.33)	(0.30, 3.99)
		$\hat{\theta}_{mod,ij} =$	3.18	1.57	0.09	0.79
		$C.I._{mod} =$	(1.73, 5.85)	(0.87, 2.84)	(0.02, 0.34)	(0.22, 2.85)
	Phosphorus	$\lvert r_{ab(ij)}\rvert =$	0.53	0.92	0.94	0.31
		$\tilde{\theta}_{obs,ij} =$	2.91	1.77	0.07	0.68
		$C.I._{obs} =$	(1.43, 5.92)	(0.92, 3.42)	(0.01, 0.37)	(0.12, 3.89)
		$\hat{\theta}_{mod,ij} =$	3.18	1.57	0.09	0.79
		$C.I._{mod} =$	(1.73, 5.85)	(0.87, 2.84)	(0.02, 0.34)	(0.22, 2.85)
	Salt	$\lvert r_{ab(ij)}\rvert =$	2.93	1.7	0.32	1.27
		$\tilde{\theta}_{obs,ij} =$	10.27	0.99	0.10	0.45
		$C.I._{obs} =$	(2.97, 35.47)	(0.44, 2.27)	(0.02, 0.57)	(0.04, 4.95)
		$\hat{\theta}_{mod,ij} =$	3.18	1.57	0.09	0.79
		$C.I._{mod} =$	(1.73, 5.85)	(0.87, 2.84)	(0.02, 0.34)	(0.22, 2.85)

where $r_{ab(ij)}$ is a standardized Pearson residual, $\tilde{\theta}_{obs,ij} = m_{11(ij)}m_{00(ij)}/(m_{01(ij)}m_{10(ij)})$ with 95% confidence intervals, $\hat{\theta}_{mod,ij} = \hat{\mu}_{11(ij)}\hat{\mu}_{00(ij)}/(\hat{\mu}_{01(ij)}\hat{\mu}_{10(ij)})$ with 95% confidence intervals

---

Kansas farmer example

- Possible lack of fit indicated for salt-testing with lagoon storage
  - Add a new model parameter
    - Indicate whether or not the subtable count is for testing waste for salt and lagoon waste storage
    - Forces a perfect fit to the corresponding subtable
  - Test new model versus saturated
    - Pearson statistic = 1.81
    - Bootstrap p-value is 0.3952 with B=5,000 resamples
    - Second-order Rao-Scott adjustment p-value is 0.5325

Kansas farmer example

- Results from model
  - Allows researchers to better understand the association structure between testing waste and waste storage

			Waste storage methods			
			Lagoon	Pit	Natural Drainage	Holding Tank
Test waste for	Nitrogen	$\hat{\theta}_{\mathrm{mod},ij}$ =	2.48	1.57	0.09	0.79
		$\mathrm{C.I.}_{\mathrm{mod}}$ =	(1.35, 4.54)	(0.87, 2.84)	(0.02, 0.34)	(0.22, 2.85)
	Phosphorus	$\hat{\theta}_{\mathrm{mod},ij}$ =	2.48	1.57	0.09	0.79
		$\mathrm{C.I.}_{\mathrm{mod}}$ =	(1.35, 4.54)	(0.87, 2.84)	(0.02, 0.34)	(0.22, 2.85)
	Salt	$\hat{\theta}_{\mathrm{mod},ij}$ =	10.27	1.57	0.09	0.79
		$\mathrm{C.I.}_{\mathrm{mod}}$ =	(2.97, 35.47)	(0.87, 2.84)	(0.02, 0.34)	(0.22, 2.85)

    - Lagoon waste storage has the strongest positive association with the waste testing
    - Natural drainage waste storage is negatively associated with testing waste for the three contaminants
      - Waste management implications for the farmers?

---

3 or more MRCVs

- Subtables are a 2^d-cell representation of the cross-classified individual item responses
  - d = number of MRCVs
- One subtable for each combination of items from the different MRCVs
  - When d = 3, there are IJK different 2×2×2 subtables where K is the number of items for a third MRCV
- Many different possible models!
  - Association structure can be modelled to vary according to items of MRCVs

---

3 or more MRCVs

- Example
  - Kansas farmer survey example also had a question about “sources of veterinary information”
    - Represent as a MRCV, Z, with 5 items
  - Best model for all three MRCVs:

$$\begin{aligned}\log(\mu_{abc(ijk)}) &= \gamma_{ijk} + \eta^{W}_{a(ijk)} + \eta^{Y}_{b(ijk)} + \eta^{Z}_{c(ijk)} + \lambda_{ab} + \lambda^{W}_{ab(i)} + \lambda^{Y}_{ab(j)} + \lambda^{WY}_{ab(ij)} \\ &+ \delta_{bc} + \delta^{Y}_{bc(j)} + \delta^{Z}_{bc(k)} + \delta^{YZ}_{bc(jk)}\end{aligned}$$

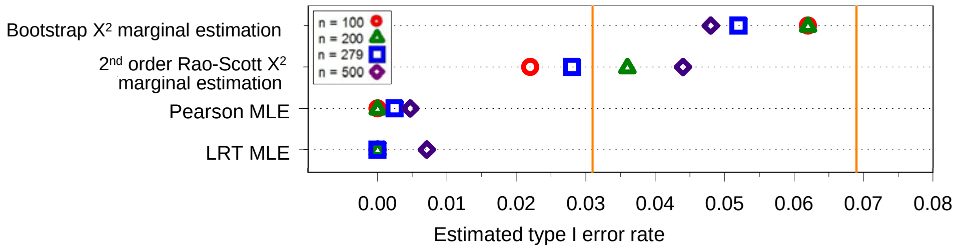
    - H_o: Model above vs. H_a: Saturated
      - Pearson statistic = 72.01
      - Bootstrap p-value of 0.8906 with B=5,000 resamples
      - 2nd-order Rao-Scott adjustment p-value is 0.8354
    - No significant standardized Pearson residuals

Simulations

- Investigate type I error
  - H_o:SPMI model, H_a:Saturated model
- Settings:
  - 2 MRCVs
  - 500 simulated data sets for each simulation
  - Nominal level = 0.05
  - B = 1,000 resamples
  - 150 iterations for MLE (convergence: 69% to 95%)
  - 95% expected range of estimated type I error rates: (0.031, 0.069)
  - Emulate observed values from the Kansas farmer data
    - I = 3 and J = 4

Simulations

- ☐ Dot plot



Summary

- ☐ New modeling procedure for MRCVs
  - Flexible and interpretable marginal models
  - Bootstrap testing procedures hold the correct size for simulations examined
  - Substantial improvement in computational ease and performance over other suggested methods
- ☐ Complex survey sampling data
  - NSF grant SES-0418632
- ☐ There is no place like Nebraska!
  - Official song of U. of Nebraska-Lincoln

Summary

There is no place like Nebraska!

Modeling Association Between Two or More Multiple-Response Categorical Variables

Christopher R. Bilder
Department of Statistics
University of Nebraska-Lincoln
www.chrisbilder.com
chris@chrisbilder.com

This research was supported in part by National Science Foundation grant SES-0207212