# Model Building for Logit and Log-linear Models

Carolyn J. Anderson

Department of Educational Psychology

University of Illinois at Urbana-Champaign

# Outline



- Association Graphs.
  1. Introduction.
  2. Collapsibility.
  3. Representing models.
- Modeling ordinal association.
  1. linear by linear association.
  2. ordinal tests of independence.
- Testing conditional independence.
- Effects of sparse data.
- Model fitting details.
- A hybrid model (log-linear with numerical predictors)

# Graphical Models



**Statistical Physics** (Gibbs, 1902). In large systems of particles, each particle occupies a site and can be in different states. The total energy of the system is composed of an external potential and a potential due to *interactions* of groups of particles. It is assumed that particles that are close to each other (i.e., they are "neighbors") interact while those that are not close to each other do not interaction.

**Genetics & Path Analysis.** (Wright, 1921, 1923, 1934). In studying the heritability of properties of natural species, graphs were used to represent *directed relations*. Arrows point from a "parent" to a "child". These ideas were taken up by Wold (1954) and Blalock (1971) in economics and social sciences and lead to what we know as path analysis.

**Interactions in 3–way contingency tables.** Barlett (1935). The notion of interaction in contingency tables studied by Barlett is formally identical to the notions used in statistical physics. The development of graphical models for multi-way contingency data stems from a paper by Darroch, J.N., Lauritzen, S.L., & Speed, T.P. (1980).

# Usefulness of Graphical Models



Graphical models are useful and are widely applicable because

1. Graphs visually represent scientific content of models and thus facilitate communication.
2. Graphs break down complex problems/models into smaller and simpler pieces that can be studied separately.
3. Graphs are natural data structures for digital computers.

Darroch, J.N., Lauritzen, S.L., & Speed, T.P. (1980). Markov fields and log-linear models for contingency tables. *Annals of Statistics, 8*, 522–539.

Edwards, D. (200). *Introduction to Graphical Modeling*, 2nd Edition. NY: Springer–Verlag.

Lauritzen, S.L. (1996). *Graphical Models*. NY: Oxford Science Publications.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*, 2nd Edition. Chichester: Wiley.

We'll be using graphs to

1. Help determine when marginal and partial associations are the same such that we can *collapse* a multi-way table into a smaller table (or tables) to study certain associations.

2. Represent substantive theories and hypotheses, which correspond to certain loglinear/logit models.

Some terminology & definitions (common to all graphical models)...

# Terminology & Definitions



Vertices (or “nodes”) are points that represent variables.

Edges are lines that connect two vertices.

The presence of an edge between two vertices indicates that an association exists between the two variables.


The absence of an edge between two vertices indicates that the two variables are independent.

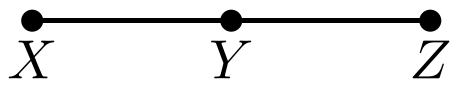

We will be restricting our attention to undirected relationships, so our lines won’t have arrows on them (lines with arrows represent directed relationships).

A Graph consists of a set of vertices and edges.

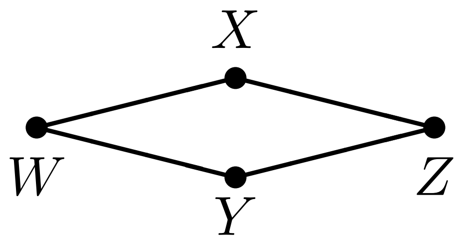
# More Terminology & Definitions



**Path** is a sequence of edges that go from one variable to another.



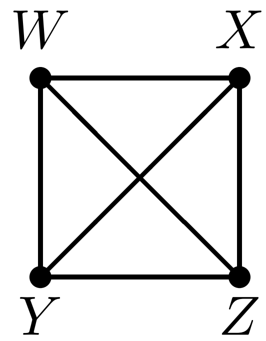

**Separated.** Two variables are said to be separated if all paths between the two variables intersect a third variable (or set of variables).



# Even More Definitions



A Clique is a set of vertices (variables) where each variable is connected to every other variable in the set.




This is also known as a “complete graph” and if this is part of a larger graph, a “complete subgraph”.

Fundamental Result (cornerstone of graphical modeling):

> Two variables are <u>conditionally independent</u> given any subset of variables that separates them.

# Graphs for Log-linear Models of...


Overview

Graphical Models

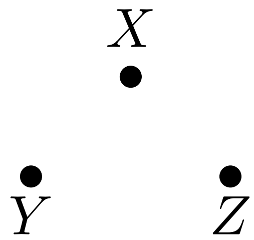
Collapsibility in 3–Way Tables

Modeling Ordinal Relationships in 2–Way Tables

Ordinal Tests of Independence

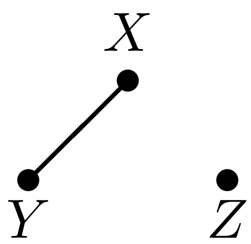
Tests of Conditional Independence


The graph for the Complete Independence, $(X, Y, Z)$

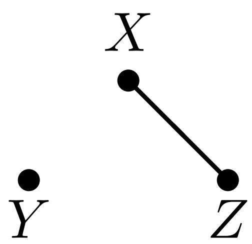



The graphs for Joint Independence,

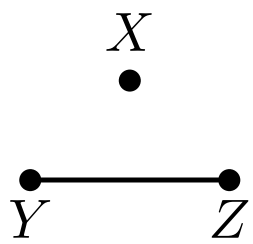
$(XY, Z)$




$(XZ, Y)$

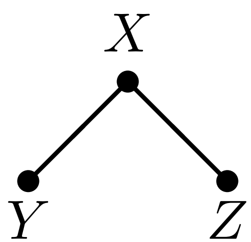



$(X, YZ)$

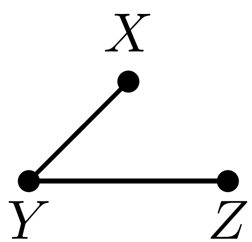



The graphs for Conditional Independence:

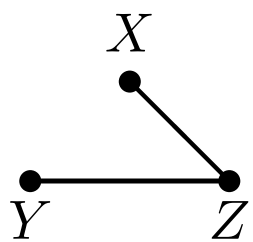
$(XY, XZ)$




$(XY, YZ)$



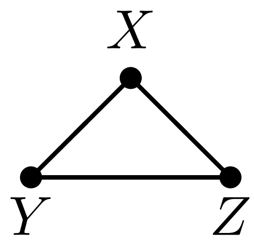

$(XZ, YZ)$



# Graphs for Log-linear Models for others



The graph for 3–Way Association model, $(XYZ)$:



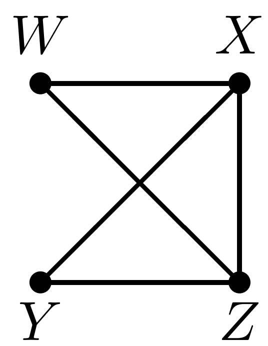

This is also a graph for Homogeneous Association, $(XY, XZ, YX)$, which is also a model of dependence.

# Association Graphs & Log-linear Models



- All Log-linear models have graphical representations.
- All independence log-linear models imply a unique graph, but not all dependence log-linear models have unique graphical representations.
- Each graph implies at least one log-linear model. Unless otherwise specified, the model "read" from a graph will the most complex one.

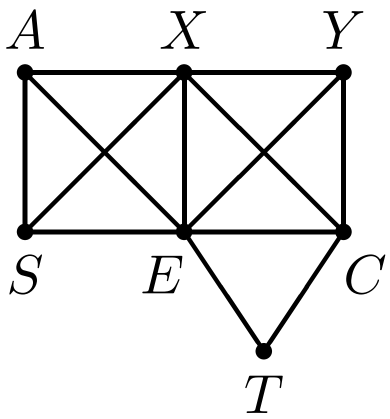
What is the log-linear model for this graph?



# More Association Graphs & Log-linear Models



What is the log-linear model for this graph?




What is the graph for this log-linear model?

$$(WY, YZ, ZX)$$

Are there other log-linear models with this graphical representation?

What is the graph for this log-linear model?

$$(WXY, WXZ)$$

# Collapsibility in 3–Way Tables



Under certain conditions, marginal associations and partial associations are the same (i.e., the partial odds ratios equal the marginal odds ratios).

The collapsibility condition for 3–way tables is

> For 3–way tables, $X$-$Y$ marginal and partial odds ratios are identical if either
>
> - $Z$ and $X$ are conditionally independent, or
> - $Z$ and $Y$ are conditionally independent.

In other words,

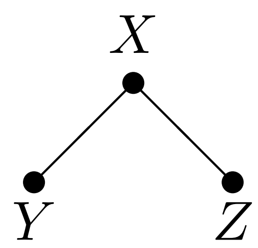
> The $X$-$Y$ marginal and partial odds ratios are identical if either the
>
> - Log-linear model $(XY, ZY)$ holds, or
> - Log-linear model $(XY, XZ)$ holds.

# Collapsibility in Graphical Terms

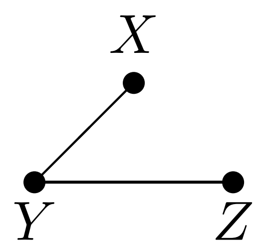


In terms of graphs,

The $X$-$Y$ marginal and partial odds ratios are identical if either of the following graphical models (or simpler ones) hold




or




Demonstration: On the next page are the partial (conditional) odds ratios and the marginal odds ratios computed based on fitted values from various log-linear models that we fit to the blue collar worker data.

# Example of Collapsibility



## Observed and fitted values from selected models:

| Manage | Super | Worker | $n_{ijk}$ | $M, S, W$ | $MS, W$ | $MS, MW$ | no $MSW$ |
|---|---|---|---|---|---|---|---|
| bad | low | low | 103 | 50.15 | 71.78 | 97.16 | 102.26 |
| bad | low | high | 87 | 82.59 | 118.22 | 92.84 | 87.74 |
| bad | high | low | 32 | 49.59 | 27.96 | 37.84 | 32.74 |
| bad | high | high | 42 | 81.67 | 46.04 | 36.16 | 41.26 |
| good | low | low | 59 | 85.10 | 63.47 | 51.03 | 59.74 |
| good | low | high | 109 | 140.15 | 104.53 | 116.97 | 108.26 |
| good | high | low | 78 | 84.15 | 105.79 | 85.97 | 77.26 |
| good | high | high | 205 | 138.59 | 174.21 | 197.28 | 205.74 |

## Partial and marginal odds ratios computed using fitted values.

| | Partial Odds Ratio | | | Marginal Odds Ratio | | |
|---|---|---|---|---|---|---|
| Model | W–S | M–W | M–S | W–S | M–W | M–S |
| $(M, S, W)$ | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| $(MS, W)$ | 1.00 | 1.00 | 4.28 | 1.00 | 1.00 | 4.28 |
| $(MS, MW)$ | 1.00 | 2.40 | 4.32 | 1.33 | 2.40 | 4.32 |
| $(MS, WS, MW)$ | 1.47 | 2.11 | 4.04 | 1.86 | 2.40 | 4.32 |
| $(MSW)$ level 1 | 1.55 | 2.19 | 4.26 | 1.86 | 2.40 | 4.32 |
| $(MSW)$ level 2 | 1.42 | 2.00 | 3.90 | | | |

# Collapsibility & Logit Models



The collapsibility condition for log-linear models applies to logit models as well.

Example: Problem 5.14 (page 138). Data from NCAA study of graduation rates of college athletes:

| Race | Sex | Graduates | Sample Size |
|---|---|---|---|
| White | women | 498 | 796 |
| White | men | 878 | 1625 |
| Black | women | 54 | 143 |
| Black | men | 197 | 660 |

The best logit model for these data is

$$\text{logit}(\pi_{ij}) = \alpha + \beta_i^R + \beta_j^S$$

Recall that $\exp(\beta_f^S - \beta_m^S)$ equals the odds ratio for graduation and gender of the athlete holding race fixed; that is,

$$\theta_{SG(i)} = \exp(\beta_f^S - \beta_m^S)$$

# Collapsibility & Logit Models (continued)

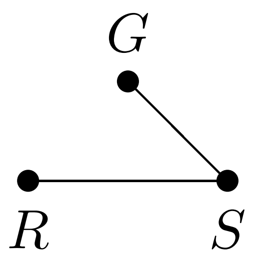


The logit model logit$(\pi_{ij}) = \alpha + \beta_i^R + \beta_j^S$ corresponds to the no 3–factor association log-linear models; that is, $(RS, RG, SG)$ where $G =$ whether the student athlete graduated or not.

If the logit model

$$\text{logit}(\pi_{ij}) = \alpha + \beta_j^S$$

had fit, which corresponds to the $(RS, SG)$ log-linear model, then we could have studied the gender–graduation relationship by looking at the gender $\times$ graduation marginal table.




According to the collapsibility condition, if the $(RS, SG)$ log-linear model fit, then the partial S-G odds ratio equals the marginal odds ratio; that is,

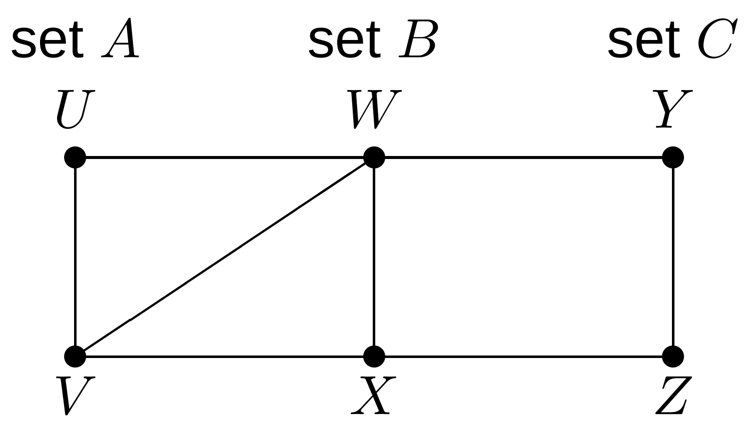
$$\theta_{SG(i)} = \theta_{SG}$$

# Collapsibility for Multiway Tables

from Agresti

> Suppose that variables in a model for a multiway table partition into three exclusive subsets, $A$, $B$, and $C$, such that $B$ separates $A$ and $C$; thus, the model does not contain parameters linking variables from $A$ with variables from $C$. When one collapses the table over the variables in $C$, model parameters relating variables in $A$ and model parameters relating variables in $A$ with variables in $B$ are unchanged.

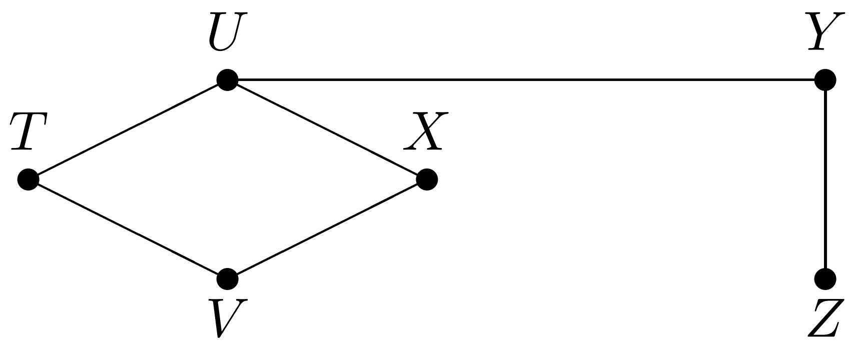
Graphically, each path between variables in set $A$ and variables in set $C$ involve at least 1 variable in set $B$.



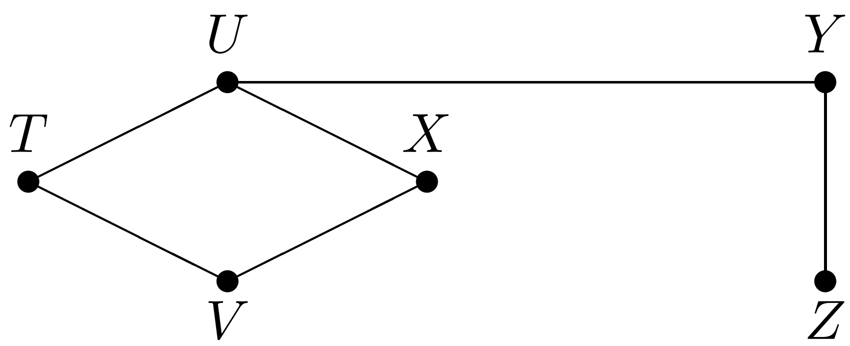
# Example of Collapsibility & Multiway Tables



One Possibility:




A 2nd Possibility:

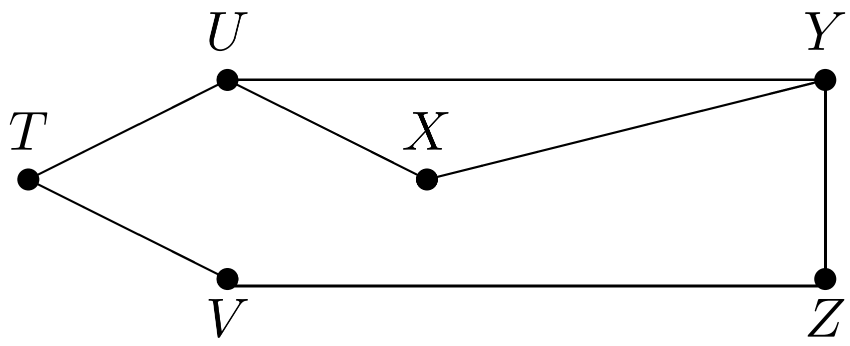



And others. . .

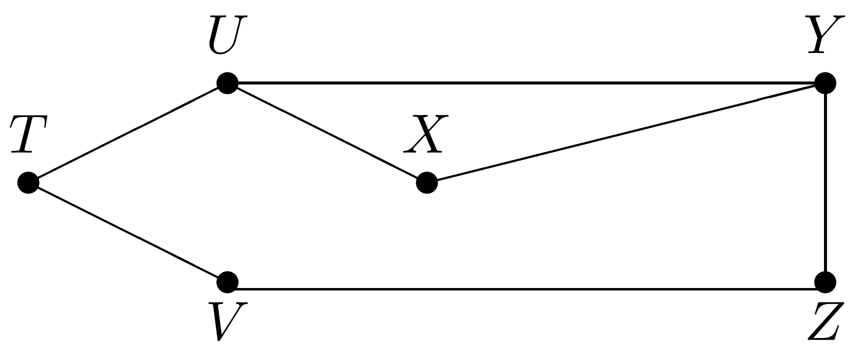
# Example 2 of Collapsibility & Multiway Tables



## One Possibility:




## A 2nd Possibility:



# Using Graphs to Guide Modeling



Example: Data from a concurrent-task detection experiment. (Olzak, 1981; Olzak & Wickens, 1983; Wickens, 1989; Anderson, 2000).

There are two signals (i.e., vertically oriented sin ways):

- $H$ — A high frequency one.
- $L$ — A low frequency one.

On each trial for each potential signal, subjects rated on a 1 to 6 scale whether a signal was present or not where 1 indicates they were sure that no signal was presented and 6 indicates that they were sure that a signal was presented. Each subject performed 2,000 trials where there were 500 consisting of $2 \times 2$ combinations of $H$ and $L$ signals being present or absent.

So there are 2 response variables:

$X$ for the rating of the $H$ signal
$Y$ for the rating of the $L$ signal.

. . . and there were 2 factors (conditions) were $L$ and $H$ present and/or absent

# Assumptions & Data



Concerned about the assumptions?

- Independence of observations?
- Homogeneity?

High Frequency Signal

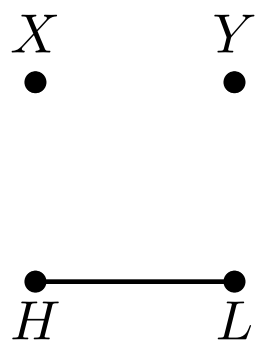
| Low Freq | $Y$ | Absent $X=1$ | 2 | 3 | 4 | 5 | 6 | Present 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Absent | 1 | 69 | 6 | 1 | 1 | 0 | 0 | 10 | 5 | 2 | 11 | 16 | 28 |
| | 2 | 34 | 20 | 10 | 3 | 1 | 0 | 8 | 5 | 11 | 43 | 27 | 38 |
| | 3 | 43 | 24 | 13 | 9 | 1 | 0 | 9 | 6 | 7 | 28 | 32 | 45 |
| | 4 | 78 | 40 | 20 | 6 | 0 | 1 | 8 | 6 | 14 | 19 | 23 | 22 |
| | 5 | 32 | 38 | 17 | 5 | 4 | 0 | 4 | 5 | 7 | 6 | 18 | 18 |
| | 6 | 5 | 14 | 3 | 2 | 0 | 0 | 0 | 1 | 2 | 3 | 5 | 8 |
| Present | 1 | 4 | 1 | 0 | 0 | 0 | 0 | 5 | 0 | 1 | 4 | 4 | 9 |
| | 2 | 5 | 3 | 2 | 1 | 0 | 0 | 0 | 1 | 3 | 6 | 9 | 27 |
| | 3 | 8 | 6 | 3 | 1 | 0 | 0 | 2 | 3 | 2 | 11 | 27 | 20 |
| | 4 | 36 | 25 | 18 | 3 | 1 | 0 | 9 | 12 | 11 | 10 | 23 | 31 |
| | 5 | 83 | 69 | 26 | 6 | 1 | 0 | 16 | 7 | 5 | 19 | 23 | 40 |
| | 6 | 127 | 50 | 12 | 7 | 2 | 0 | 21 | 14 | 13 | 20 | 21 | 61 |

With four variables, there are many possible models to fit. However, we don't need to consider all models that could be fit to the data.

# Random Responding



Since $H$ and $L$ were fixed by the experimenter (i.e., "fixed by design"), all models should include terms $\lambda^{HL}$ for the $HL$ association. The simplest model would be that a subject responds randomly



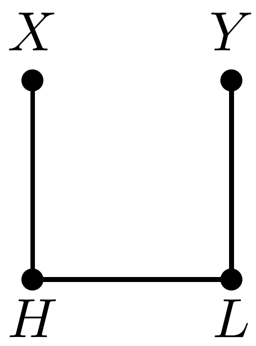

The log-linear model: $(HL, X, Y)$.

$G^2 = 2265.57$, $df = 130$, $p < .01$

# Detectable Signals



The subject can detect the signals & detecting one does not influence detection of the other (i.e., subject does what the experimenter asked).




Log-linear model:

$$(HL, XH, YL)$$

This is a "base" model to which we can add more complicated forms of associations.

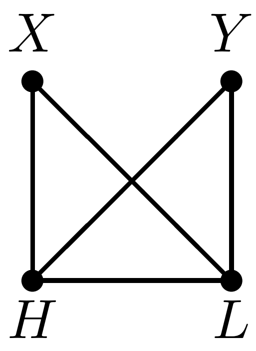
$G^2 = 2,152.26$, $df = 120$, $p < .01$

If one signal or the other was not detectable, then we might have another base model (e.g., $(HL, XH, Y)$ or $(HL, LY, X)$).

# Association to the unrelated signal



In this model, responses to one signal are influenced by whether both signals are present and/or absent (i.e., the appropriate and inappropriate signal).

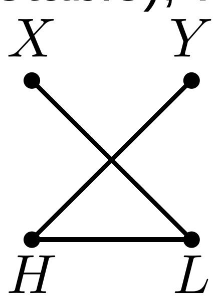



The log-linear model

$$(HLX, HLY)$$

$G^2 = 221.43$, $df = 100$, $p < .01$

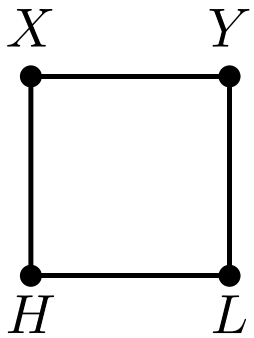
- $X$ and $Y$ are conditionally independent given $H$ and $L$.
- A more restricted alternative model that also has this graphical representation, $(HL, HX, HY, LX, LY)$.
- Since we're only considering models that "make sense" (i.e. that are interpretable), we wouldn't include a model such as



# Response-response association



We add to the base model (detectable signals) the possibility that a response regarding one signal is related to response to the other signal.




The log-linear model: $(HL, HX, LY, XY)$.

$G^2 = 1,944.84$, $df = 95$, $p < .01$

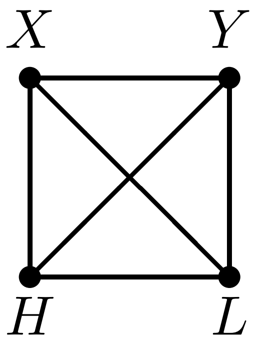
# All pairwise associations.



The log-linear model $(HL, HX, HY, LX, LY, XY)$.

$G^2 = 113.82$, $df = 85$, $p = .02$

It's graphical representation is




This is also the representation of many other log-linear models with dependencies, including model with 4-way interaction (i.e. saturated model).

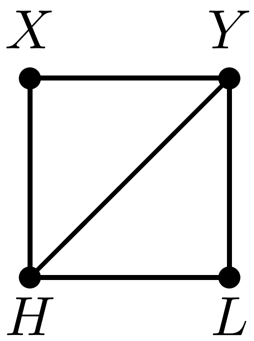
This is the most complex graph, but there are interesting log-linear models that have this representation.

# Another Model

We can add three–factor terms to the all pairwise association model. Some of these all have reasonable interpretations.

For example, consider the model that adds $\lambda^{HLY}$ (i.e., $(HLY, HY, LY, XY)$). The $\lambda^{HLY}$ terms imply that delectability of the $L$ signal (measured by $Y$) is affected by the presence of the $H$ signal.

It's graphical representation is

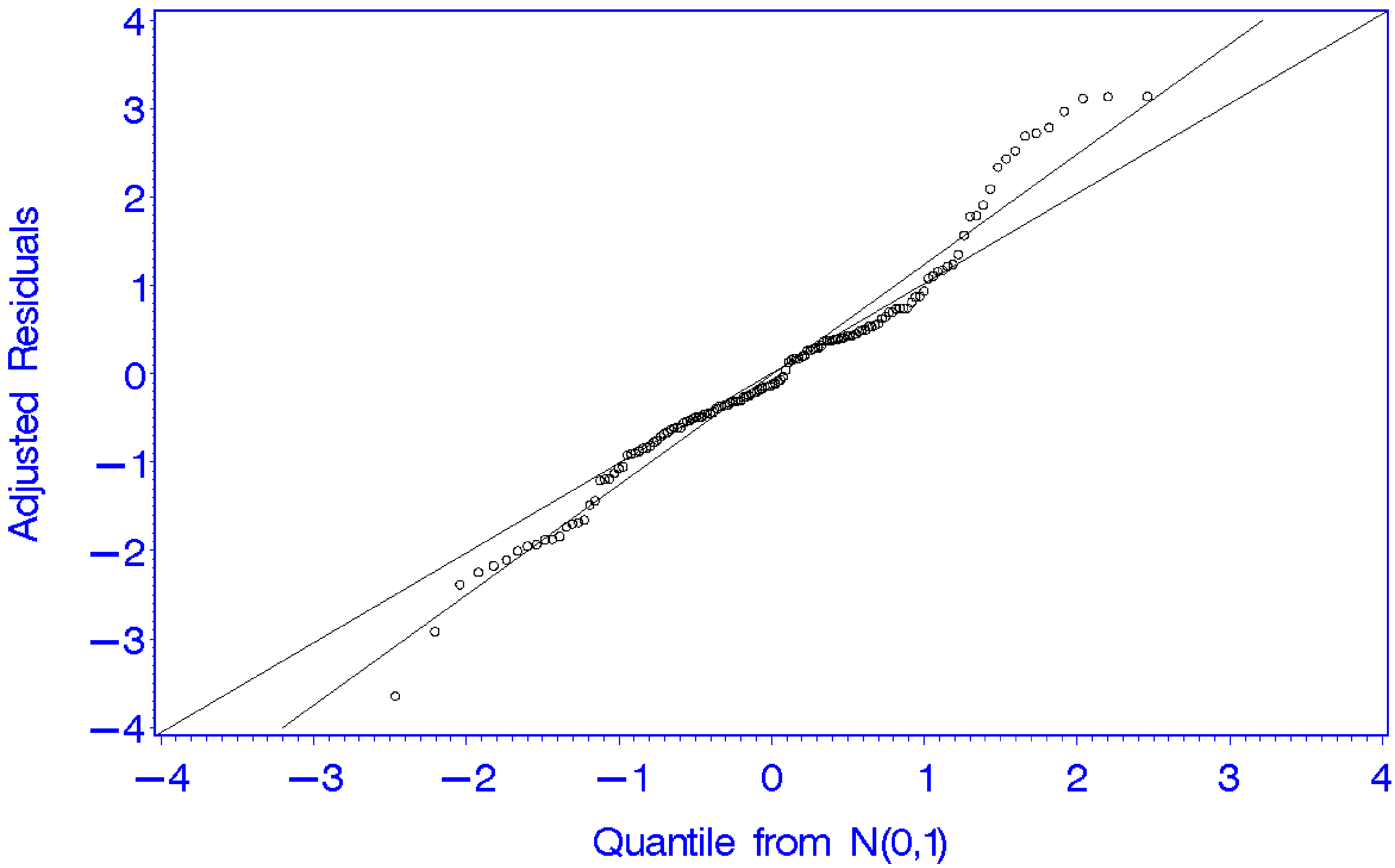



Fit of this model

$$G^2 = 98.06, \quad df = 80, \quad p = .08$$

# QQ Plot for (HLY,HY,LY,XY)





# Eg: 4-way Table with Time Ordering



Using a suggested ordering of the variables in terms of time and causal hypotheses and show how to “decompose” a model into smaller pieces.

Example from Agresti, 1990; The variables:

**G** for gender.

**PMS** for premarital sex.

**EMS** for extra martial sex.

**M** for marital status (divorced, still married).

We’ll depart somewhat from the graphical models that we’ve discussed so far and talk about directed relationships.

The point in time at which values of variables were determined:

**G**      **PMS**      **EMS**      **M**

Any variable to the right of others could be a response & those left of it explanatory.

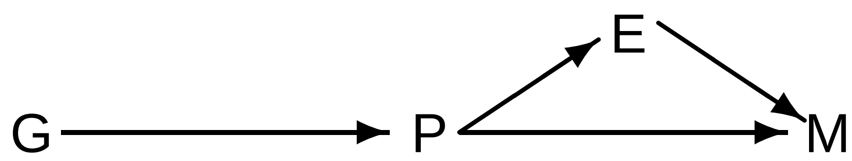
# Breaking down the analysis



We could analyze these data in three stages:

| Stage | Response | Explanatory |
|---|---|---|
| (1) | PMS | Gender |
| (2) | EMS | Gender, PMS |
| (3) | M | Gender, PMS, EMS |

To further guide the modeling consider the following figure, which might have been hypothesized as the existing causal structure for the variables.




Stage 1: PMS is the response & G explanatory.

$G^2\left[(G,P)\right] = 75.26$, $df = 1$, and $p < .0001$.

Sample (marginal) odds ratio $\hat{\theta}_{GP} = .27$ (or $1/.27 = 3.70$).

# Next Stage

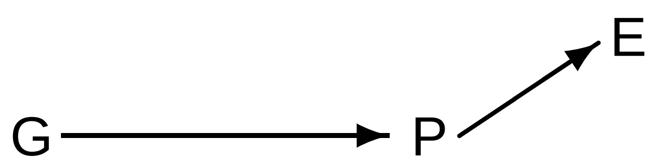
Stage 2: EMS is the response and G & PMS are possible explanatory variables.

| Model | *df* | $G^2$ | $p$ | $X^2$ | $p$ |
|---|---|---|---|---|---|
| $(GP, E)$ | 3 | 48.92 | $< .001$ | 56.77 | $< .001$ |
| $(GP, PE)$ | 2 | 2.91 | .23 | 2.95 | .23 |
| $(GP, GE, PE)$ | 1 | .00[a] | .98 | .00[a] | .98 |

*a.* Value $= .0008$.

Loglinear model $(GP, PE)$ fits pretty well.
The estimated $P$-$E$ odds ratio $\hat{\theta}_{EP} = 3.99$.

The marginal odds ratio is also equal to 3.99, and the reason why can be seen by looking at the figure for the model that fit:



# Last Stage

Stage 3: M is the response, while G, PMS and EMS are all explanatory variables.

| Model | $df$ | $G^2$ | $p$ |
|---|---|---|---|
| $(EGP, EM, PM)$ | 5 | 18.16 | $< .01$ |
| $(EGP, EMP)$ | 4 | 5.25 | .26 |
| $(EGP, EMP, GM)$ | 3 | .70 | .98 |

Notes:

- $(EGP, EM, PM)$ corresponds to the original figure.
- $(EGP, EMP)$ adds an interaction between EMS and PMS with respect to M, marital status (from previous analyses we knew that there was a strong & important EMS $\times$ PMS $\times$ M effect.
- $(EGP, EMP, GM)$ adds a main effect for Gender with respect to predicting M.
- $(EGP, EMP)$ and $(EGP, EMP, GM)$ are more complex than implied by original figure.

# Modeling Ordinal Relationships in 2–Way Tables



- Loglinear models for contingency tables treat all variables as nominal variables.
- If there is an ordering of the categories of the variables, this is not taken into account
- Could rearrange the rows and/or columns of a table and we would get the same fitted odds ratios for the data as we would given the ordinal ordering of the rows and/or columns.

High School and Beyond: Consider Program type (Vocational/technical, general and academic) and SES (low, middle, high).

| | Program Type | | |
|---|---|---|---|
| SES | Vo/Tech | General | Academic |
| Low | 45 | 50 | 44 |
| Middle | 82 | 70 | 147 |
| High | 20 | 25 | 117 |

# In between independence & saturated models



For the SES $\times$ Program type data, if the two variables are independent, then we have

$$\log(\mu_{ij}) = \lambda + \lambda_i^S + \lambda_j^P$$

This model fails to fit the data, $df = 4$, $G^2 = 53.72$, $p < .001$, which leaves us with the saturated model

$$\log(\mu_{ij}) = \lambda + \lambda_i^S + \lambda_j^P + \lambda_{ij}^{SP}$$

We can use ordering of SES levels and assign scores to them and we'll guess at the ordering of the program types, which we can use our model.

Given scores for the rows $\{u_1 \leq u_2 \ldots \leq u_I\}$ and scores for the columns $\{v_1 \leq v_2 \leq \ldots \leq v_J\}$, then we can model the dependency between the variables:

$$\log(\mu_{ij}) = \lambda + \lambda_i^S + \lambda_j^P + \beta u_i v_j$$

This model is know as the "linear by linear association model".

# Linear by Linear Association Model



$$\log(\mu_{ij}) = \lambda + \lambda_i^S + \lambda_j^P + \beta u_i v_j$$

- It's called the "linear by linear association model," because...

  For each row $i$, the association is a linear function of the columns,

$$\lambda_{ij}^{SP} = (\beta u_i) v_j$$

  For each column $j$, the association is a linear function of the rows.

$$\lambda_{ij}^{SP} = (\beta v_j) u_i$$

- Only has 1 more parameter than the independence model (i.e., $\beta$), so it is "in between" independence and the saturated models.
- If $\beta > 0$, then $X$ and $Y$ are positively associated (i.e., $X$ tends to go up as $Y$ goes up).
- If $\beta < 0$, the $X$ and $Y$ are negatively associated.

# Linear by Linear Association Model (continued)



$$\log(\mu_{ij}) = \lambda + \lambda_i^S + \lambda_j^P + \beta u_i v_j$$

- The odds ratio for any $2 \times 2$ sub-table is a direct function of the row and column scores and $\beta$.

$$\begin{aligned}\log\left(\frac{\mu_{ij}\mu_{i'j'}}{\mu_{i'j}\mu_{ij'}}\right) &= \log(\mu_{ij}) + \log(\mu_{i'j'}) - \log(\mu_{i'j}) - \log(\mu_{ij'}) \\ &= \beta(u_i v_j + u_{i'} v_{j'} - u_{i'} v_j - u_i v_{j'}) \\ &= \beta(u_i - u_{i'})(v_j - v_{j'})\end{aligned}$$

The strongest associations occur in the extreme corners of the table, which is where the differences between scores is the largest.

The smallest associations occur for rows and columns that have scores that are more nearly equal.

# Example of linear by linear model



For the high school data example, it seems reasonable to assign equally spaced scores for the levels of SES:

$$u_1 = 1, \qquad u_2 = 2, \qquad u_3 = 3$$

For the program types, it seems reasonable to order them as:

Vo/Tech General Academic

Guess that Vo/Tech and General should be closer together than are General and Academic; therefore, let's try

$$v_1 = 1, \qquad v_2 = 2 \qquad v_3 = 4$$

Using these scores

| Model | $df$ | $G^2$ | $p$ | $\Delta df$ | $\Delta G^2$ | $p$ |
|---|---|---|---|---|---|---|
| Independence | 4 | 53.715 | $< .001$ | — | — | — |
| L by L | 3 | 5.980 | .10 | 1 | 47.74 | $< .001$ |

# Estimated Parameters & Odds Ratios



$$\hat{\beta} = .32 \quad \text{and} \quad \exp(.32) = 1.38,$$

The odds ratio for a unit change in row and column scores equals 1.38 (e.g., odds ratio for low–middle SES and vo/tech–academic subtable).

The extreme corners of our table, which correspond to the low & high SES levels and program types vo/tech & academic:

$$\hat{\theta} = \exp\left[.3214(3-1)(4-1)\right] = \exp(.3214(6)) = 6.88$$

The odds of attending an academic versus a vo/tech program if you're high SES is 6.88 times larger than the odds if you're low SES.

# SAS/GENMOD and Fitting the L by L model



```
DATA hsb;
  input ses $ hsp $ count u v ;
  datalines;
 low    general     50   1   1
 low    academic    44   1   4
 low    votech      45   1   2
 mid    general     70   2   1
 mid    academic   147   2   4
 mid    votech      82   2   2
 hi     general     25   3   1
 hi     academic   117   3   4
 hi     votech      20   3   2
PROC GENMOD data=hsb;
  class ses hsp;
  model count = ses hsp u*v / link=log dist=poi;
  title 'Linear x Linear Association Model';
```

Linear x Linear Association Model

The GENMOD Procedure

Model Information

| | |
|---|---|
| Data Set | WORK.HSB |
| Distribution | Poisson |
| Link Function | Log |
| Dependent Variable | count |

| | |
|---|---|
| Number of Observations Read | 9 |
| Number of Observations Used | 9 |

Class Level Information

| Class | Levels | Values |
|---|---|---|
| ses | 3 | hi low mid |
| hsp | 3 | academic general votech |

Criteria For Assessing Goodness Of Fit

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| Deviance | 3 | 5.9798 | 1.9933 |
| Scaled Deviance | 3 | 5.9798 | 1.9933 |
| Pearson Chi-Square | 3 | 5.6845 | 1.8948 |
| Scaled Pearson X2 | 3 | 5.6845 | 1.8948 |
| Log Likelihood | | 2020.3156 | |

Algorithm converged.

Linear x Linear Association Model

The GENMOD Procedure

Analysis Of Parameter Estimates

| Parameter | | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|---|
| Intercept | | 1 | 3.04 | 0.21 | 2.63 | 3.45 | 216.20 | <.0001 |
| ses | hi | 1 | -1.59 | 0.19 | -1.95 | -1.22 | 72.06 | <.0001 |
| ses | low | 1 | 0.04 | 0.15 | -0.26 | 0.34 | 0.07 | .7903 |
| ses | mid | 0 | 0.00 | 0.00 | 0.00 | 0.00 | . | . |
| hsp | academic | 1 | -0.59 | 0.23 | -1.04 | -0.14 | 6.72 | .0095 |
| hsp | general | 1 | 0.58 | 0.14 | 0.30 | 0.86 | 16.44 | <.0001 |
| hsp | votech | 0 | 0.00 | 0.00 | 0.00 | 0.00 | . | . |
| u*v | | 1 | 0.32 | 0.05 | 0.23 | 0.42 | 43.71 | <.0001 |
| Scale | | 0 | 1.00 | 0.00 | 1.00 | 1.00 | | |

NOTE: The scale parameter was held fixed.

# Choice of Scores



- Sets of scores with the same spacing between them will lead to the same goodness-of-fit statistics, fitted counts, odds ratios, *and* $\hat{\beta}$.

  For HSB data, the following set of scores for the columns (hsp) would yield that same result: $v_1 = 0$, $v_2 = 1$, $v_3 = 3$.

- Two sets of scores with the same *relative* spacing will lead to the same goodness-of-fit statistics, fitted counts, and odds ratios, but different estimates of $\beta$. e.g.,

$$v_1 = 2, \qquad v_2 = 4 \qquad v_3 = 8$$

  With these column (HSP) scores, $\hat{\beta} = .1607$.

  - Odds ratio for low & middle (or middle & high) and vo/tech & general

$$\hat{\theta} = \exp[.1607(2-1)(4-2)] = \exp[.1607(2)] = \exp[.3214] = 1.38$$

  - Odds ratio for low & high SES and program types vo/tech & academic:

$$\hat{\theta} = \exp[.1607(3-1)(8-2)] = \exp[.1607(12)] = 6.88$$

# Uniform Association Model



When scores are consecutive integers (or equally spaced scores) are used, e.g.,

$$u_1 = 1, \quad u_2 = 2, \quad \ldots, u_I = I$$
$$v_1 = 1, \quad v_2 = 2, \quad \ldots, v_J = J$$

This special case of L by L model is the "Uniform Association Model."

The uniform association model for the HSB example:

| Model | $df$ | $G^2$ | $p$ |
|---|---|---|---|
| Independence | 4 | 53.715 | $< .01$ |
| L by L | 3 | 5.980 | .10 |
| Uniform Assoc | 3 | 11.74 | $< .01$ |

This model is called the Uniform Association Model, because the odds ratios for any two adjacent rows and any two adjacent columns equals

$$\theta = \exp\left[\beta(u_i - u_{(i-1)})(v_j - j_{(j-1)})\right] = \exp(\beta)$$

The "Local Odds Ratio" equals $\exp(\beta)$ and is the same for adjacent rows and columns.

# GSS example of Uniform Association Model



Recall. . .

**Item 1:** A working mother can establish just as warm and secure of a relationship with her children as a mother who does not work.

**Item 2:** Working women should have paid maternity leave.

| | Item 2 | | | | |
|---|---|---|---|---|---|
| | strongly agree | agree | neither | disagree | strongly disagree |
| Item 1 | 1 | 2 | 3 | 4 | 5 |
| 1 | 97 | 96 | 22 | 17 | 2 |
| 2 | 102 | 199 | 48 | 38 | 5 |
| 3 | 42 | 102 | 25 | 36 | 7 |
| 4 | 9 | 18 | 7 | 10 | 2 |

# GSS Results



| Model/Test | $df$ | $G^2$ | $p$ | Estimates |
|---|---|---|---|---|
| Independence | 12 | 44.96 | $< .001$ | |
| $M^2$ | 1 | 36.261 | $< .001$ | $r = .20$ |
| Uniform Assoc | 11 | 8.67 | $.65$ | $\hat{\beta} = .24$, $ASE = .0412$ |

$H_o : \beta = 0$ vs $H_a : \beta \neq 0$,
L.R. test: $G^2 = (44.96 - 8.67) = 36.29$, $df = 1$, $p < .01$

The estimated local odds ratio equals $e^{.24} = 1.28$.

For the extreme corners of the table, the estimated odds ratio equals $e^{.24(3)(4)} = 18.5$

Unlike the tests of ordinal association that are based on a correlation, these models provide us with estimated odds ratios for the table, as well as permit us to check residuals, etc.

# Back to HSB



For the HSB data, using equally spaced scores we find that

$$M^2 = 40.87, \qquad df = 1, \qquad p < .001, \qquad \text{and} \qquad r = .26$$

However, when we fit the linear by linear association model with equal scores it did not fit the data (this is shown in the residuals, as well).

| Model | | $df$ | $G^2$ | $p$ |
|---|---|---|---|---|
| Independence | | 4 | 53.715 | $< .01$ |
| L by L | (unequal spacing) | 3 | 5.980 | .10 |
| Uniform Assoc | (equal spacing) | 3 | 11.74 | $< .01$ |

# Ordinal Tests of Independence



CMH test was one way to test of ordinal association (or independence), but now we have a model based method. Using the linear by linear association model

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \beta u_i v_j$$

The likelihood ratio test and the Wald test of the hypothesis

$$H_o : \beta = 0$$

is the same as testing

$$H_o : \text{independence}$$

Using the likelihood ratio test,

$$G^2(I|L \times L) = G^2(I) - G^2(L \times L)$$

For the HSB data: $G^2(I|L \times L) = 53.715 - 5.98 = 47.73$ with $df = 4 - 3 = 1$, and $p < .001$.

The Wald test: $\left(\frac{\hat{\beta}}{ASE}\right)^2 = \left(\frac{.3199}{.0485}\right)^2 = 43.55$

The CMH test is the efficient score test for this same hypothesis (i.e., $M^2 = 40.87$, $df = 1$, $p < .001$).

# More Association Models for HSB



| Model | | $df$ | $G^2$ | $p$ |
|---|---|---|---|---|
| Independence | | 4 | 53.715 | $< .01$ |
| Uniform Assoc | (equal spacing) | 3 | 11.74 | $< .01$ |
| L by L | (unequal spacing) | 3 | 5.980 | .10 |
| Nominal HSP $\times$ Ordinal SES | (equal spaced SES) | 2 | 2.30 | .32 |
| $RC$ association | (scores estimated) | 1 | 1.74 | .19 |

- The estimated parameters for the SES $\times$ HSP association in the nominal $\times$ ordinal model

$$\hat{\beta}_{\text{votech}} = .000, \qquad \hat{\beta}_{\text{general}} = -.005 \qquad \hat{\beta}_{\text{academic}} = .864$$

- $RC$ association model estimates the scores for both SES and HSP, as well as $\beta$ (the "association parameter").

| HSP | est. score | SES | est. score |
|---|---|---|---|
| VoTech | -.423 | Low | -.669 |
| General | -.393 | Middle | -.071 |
| Academic | .816 | High | .740 |

and $\hat{\beta} = 1.000$

# Comments on models for ordinal variables



- This approach is not restricted to models for 2–way tables and log-linear models. You add use scores in log-linear and/or logit model for higher–way tables.

- There are more general models where the scores are estimated from the data. For 2–way tables, this includes Goodman's "row effects" model ($R$), "column effects" model ($C$), "row $+$ column" effects model ($R + C$), and the row–column model $RC$). There are generalizations of these models to multiple dimensions and higher–way tables.

- There are also models for ordinal *response* variables that take into account the ordering of the categories.

- Other ordinal models (Vermunt, J.K. (2001). *Sociological Methodology*).

- Linear by linear with latent variable interpretations (Anderson & Vermunt, 2000; Anderson, 2002; etc).

# Wickens & Olzak revisited



A good model for Wickens & Olzak data is $(HLY, HY, LY, XY)$,

$$\log(\mu_{ijkl}) = \lambda + \lambda_i^H + \lambda_j^L + \lambda_{ij}^{HL} + \lambda_k^X + \lambda^Y + \lambda_{kl}^{XY} + \lambda_{ik}^{HX} + \lambda_{jl}^{LY} + \lambda_{ijl}^{HLY}$$

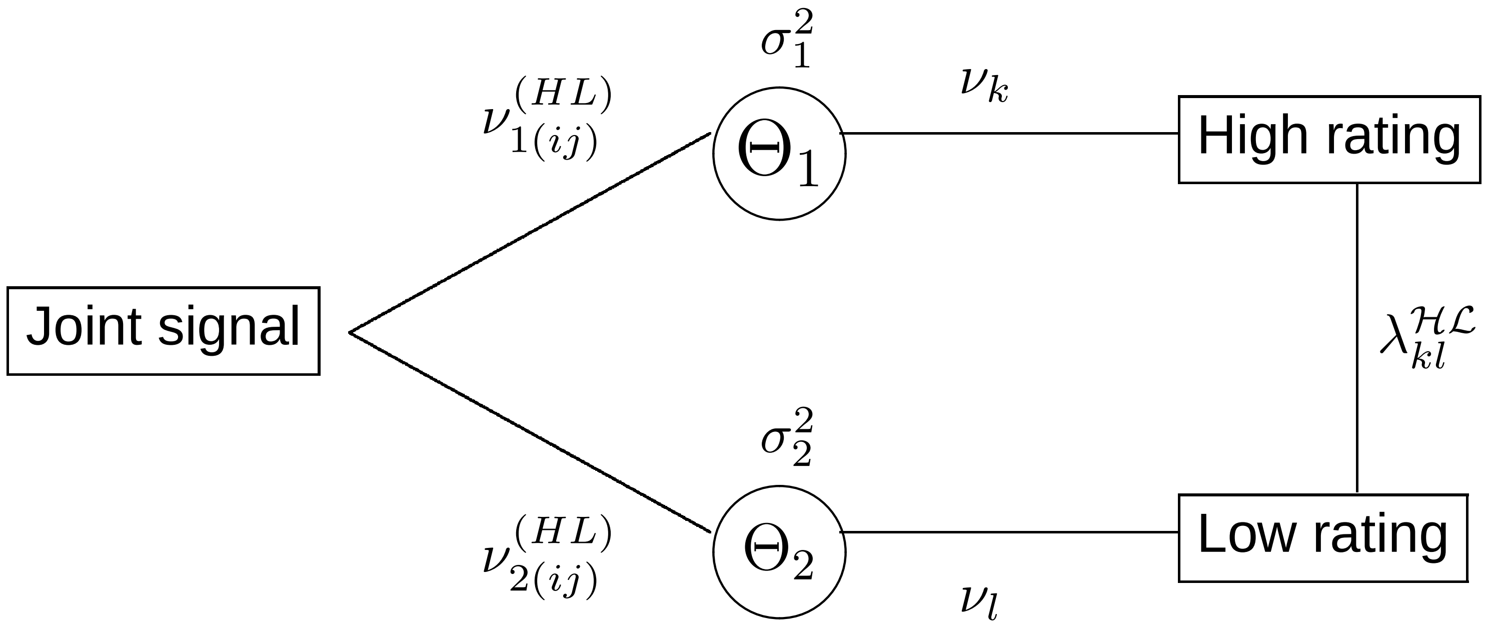
Let $u_k = 1, \ldots, 6$ and $v_l = 1, \ldots, 6$ be scores for the high and low responses, respectively. We can use these instead of nominal responses:

$$\log(\mu_{ijkl}) = \lambda + \lambda_i^H + \lambda_j^L + \lambda_{ij}^{HL} + \lambda_k^X + \lambda^Y + \lambda_{kl}^{XY} + \lambda_i^H u_k + \lambda_j^L v_l + \lambda_{ij}^{HL} v_l$$

This model doesn't fit particularly well ($G^2 = 259.1267$, $df = 100$, $p < .01$), but one with estimated scores does.

# Eg. of a Latent Variable Model





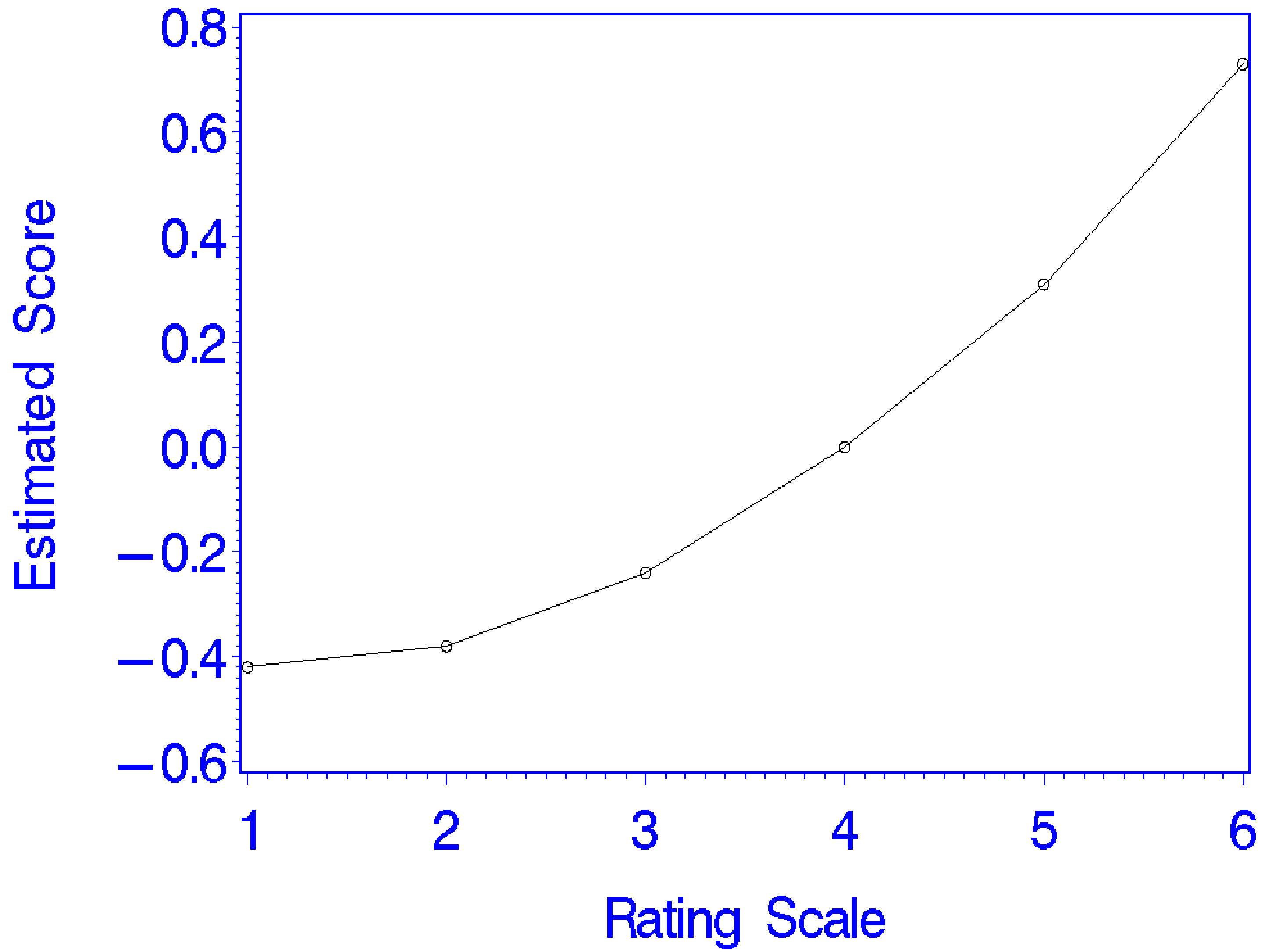

$$\log(\mu_{ijkl}) = \lambda + \lambda_{ij}^{HL} + \lambda_k^X + \lambda^Y + \lambda_{kl}^{XY} + \sigma_1^2\nu_{1(ij)}^{HL}\nu_k + \sigma_2^2\nu_{2(ij)}^{HL}\nu_l$$

$G^2 = 138.35$, $df = 98$, $p = .01$, $D = .082$.

But there were 2 subjects and this graph describes both. For the other subject (“subject A”), $G^2 = 111.12$, $df = 97$, $p = .15$, $D = .086$).

# Estimated Scores





# Parameter Estimates— the low signal



For both subjects

| Variable | Level | Parameter estimate | Standard error |
|---|---|---|---|
| rating | 1 | $\hat{\nu}_1^{rating} = -.42$ | (.01) |
| rating | 2 | $\hat{\nu}_2^{rating} = -.38$ | (.01) |
| rating | 3 | $\hat{\nu}_3^{rating} = -.24$ | (.01) |
| rating | 4 | $\hat{\nu}_4^{rating} = .00$ | (.02) |
| rating | 5 | $\hat{\nu}_5^{rating} = .31$ | (.02) |
| rating | 6 | $\hat{\nu}_6^{rating} = .73$ | (.02) |
| signal | high absent/low absent | $\hat{\nu}_{2(11)}^{(HL)} = -.49$ | (.00) |
| signal | high present/low absent | $\hat{\nu}_{2(21)}^{(HL)} = -.49$ | (.00) |
| signal | high absent/low present | $\hat{\nu}_{2(12)}^{(HL)} = .61$ | (.02) |
| signal | high present/low present | $\hat{\nu}_{2(22)}^{(HL} = .37$ | (.03) |
| $\Theta_2$ | | $\hat{\sigma}_2^2 = 3.30$ | (.12) |

# Parameter Estimates— the high signal



The $\nu$'s for ratings are the same as previous slide.

| | *Subject* $A$ | | *Subject* $B$ | |
|---|---|---|---|---|
| high absent/low absent | $\hat{\nu}^{(HL)}_{1(11)A} = -.58$ | (.00) | $\hat{\nu}^{(HL)}_{1(11)B} = -.50$ | (n.a.) |
| high present/low absent | $\hat{\nu}^{(HL)}_{1(21)A} = .41$ | (.00) | $\hat{\nu}^{(HL)}_{1(21)B} = .50$ | (n.a.) |
| high absent/low present | $\hat{\nu}^{(HL)}_{1(12)A} = -.39$ | (.00) | $\hat{\nu}^{(HL)}_{1(12)B} = -.50$ | (n.a.) |
| high present/low present | $\hat{\nu}^{(HL)}_{1(22)A} = .58$ | (.00) | $\hat{\nu}^{(HL)}_{1(22)B} = .50$ | (n.a.) |
| $\Theta_1$ | $\hat{\sigma}^2_{1A} = 2.69$ | (.18) | $\hat{\sigma}^2_{1B} = 7.11$ | (.49) |

# Tests of Conditional Independence



General terms of testing whether row ($X$) and column ($Y$) classifications are independent conditioning on levels of a third variable ($Z$).

There are 3 kinds of tests:

1. Likelihood ratio tests ("LR" for short).
   (a) Comparing conditional independence model to homogeneous association model.
   (b) Comparing conditional independence model to saturated model.
2. Wald tests.
3. Efficient score tests, i.e. Generalized CMH.

The LR and Wald tests require the estimation of (model) parameters, while the Efficient score tests do not.

# Nature of Variables: Ordinal &/or Nominal



We have 3 cases:

1. Nominal-Nominal
2. Ordinal-Ordinal
3. Nominal-Ordinal

So the possibilities are:

| Variable | | Type of Test | | |
|---|---|---|---|---|
| | | Likelihood | | (Generalized) |
| Row | Column | Ratio | Wald | CMH |
| Nominal | Nominal | | | |
| Nominal | Ordinal | | | |
| Ordinal | Ordinal | | | |

# Some Data that We'll Use



For illustration, we'll use some High School & Beyond data, i.e., the cross-classification of gender (**G**), SES (**S**) and high school program type (**P**).

| Females | High School Program | | | |
|---|---|---|---|---|
| SES | VoTech | General | Academic | Total |
| low | 15 | 19 | 16 | 50 |
| middle | 44 | 30 | 70 | 144 |
| high | 12 | 11 | 56 | 79 |
| Total | 71 | 60 | 142 | 273 |

| Males | High School Program | | | |
|---|---|---|---|---|
| SES | VoTech | General | Academic | Total |
| low | 30 | 31 | 28 | 89 |
| middle | 38 | 40 | 77 | 155 |
| high | 8 | 14 | 61 | 83 |
| Total | 76 | 85 | 166 | 327 |

# Model Based Tests of Conditional Independence



The likelihood ratio test. We compare the fit of the conditional independence model and comparing it to the homogeneous association model.

For example to test whether $X$ and $Y$ are conditionally independent given $Z$, i.e.,

$$H_O : \quad \text{all} \quad \lambda_{ij}^{XY} = 0$$

The likelihood ratio test statistic is

$$G^2\left[(XZ,YZ)|(XY,XZ,YZ)\right] = G^2(XZ,YZ) - G^2(XY,XZ,YZ)$$

with $df = df(XZ,YZ) - df(XY,XZ,YZ)$.

# The likelihood ratio test (example)



Example: **G**= Gender , **S**= SES, and **P**= Program type. Testing whether SES and program type are independent given gender,

$$H_O : \quad \text{all} \quad \lambda_{ij}^{SP} = 0$$

| | Goodness-of-fit Test | | | **Likelihood Ratio Test** | | |
|---|---|---|---|---|---|---|
| Model | $df$ | $G^2$ | $p$ | $\Delta df$ | $\Delta G^2$ | $p$ |
| $(GS, GP, SP)$ | 4 | 1.970 | .74 | — | — | — |
| $(GS, GP)$ | 8 | 55.519 | $< .0001$ | 4 | 53.548 | $< .001$ |

# Notes Regarding Likelihood Ratio Test



- This test assumes that $(XY, XZ, YZ)$ holds.
- This single test is preferable to conducting $(I-1)(J-1)$ Wald tests, one for each of the non-redundant $\lambda_{ij}^{XY}$'s. For our example, the result is pretty unambiguous; that is,

| Parameter | Estimate | ASE | Wald | $p$ |
|---|---|---|---|---|
| $\lambda_{lv}^{SP}$ | 1.8133 | .3233 | 31.450 | $< .0001$ |
| $\lambda_{lg}^{SP}$ | 1.6600 | .3033 | 29.952 | $< .0001$ |
| $\lambda_{mv}^{SP}$ | 1.1848 | .2786 | 18.079 | $< .0001$ |
| $\lambda_{mg}^{SP}$ | .8004 | .2639 | 9.198 | $.0024$ |

- For binary $Y$, this is the same as performing the likelihood ratio test of whether $H_O$ : all $\beta_i^X = 0$ in the logit model

$$\text{logit}(\pi_{ik}) = \alpha + \beta_i^X + \beta_k^Z$$

which corresponds to the $(XY, XZ, YZ)$ log-linear model.

# Notes Regarding Likelihood Ratio Test



For $2 \times 2 \times K$ tables, this likelihood ratio test of conditional independence has the same purpose as the Cochran–Mantel–Haenszel (CMH) test. For the CMH test,

- It works the best when the partial odds ratios are similar in each of the partial tables.
- It's natural alternative (implicit) hypothesis is that of homogeneous association.
- CMH is the efficient score tests of $H_O : \lambda_{ij}^{XY} = 0$ in the log-linear model.

# Direct Goodness-of-Fit Test



We compare the fit of the conditional independence model to the saturated model; that is,

$$G^2\left[(XZ,YZ)|(XYZ)\right] = G^2(XZ,YZ) - G^2(XYZ)$$

The null hypothesis for this test statistic is

$$H_O: \quad \text{all} \quad \lambda_{ij}^{XY} = 0 \qquad \text{and} \quad \text{all} \quad \lambda_{ijk}^{XYZ} = 0$$

Example: **G**= Gender , **S**= SES, and **P**= Program type. Testing whether SES and program type are independent given gender,

$$H_O: \quad \text{all} \quad \lambda_{ij}^{SP} = 0 \qquad \text{and} \quad \text{all} \quad \lambda_{ijk}^{GSP} = 0$$

| | **Goodness-of-fit Test** | | | Likelihood Ratio Test | | |
|---|---|---|---|---|---|---|
| Model | $df$ | $G^2$ | $p$ | $\Delta df$ | $\Delta G^2$ | $p$ |
| $(GS,GP,SP)$ | 4 | 1.970 | .74 | — | — | — |
| $(GS,GP)$ | 8 | 55.519 | $< .0001$ | 4 | 53.548 | $< .001$ |

# Notes on Direct Goodness-of-Fit Test



A direct goodness-of-fit test does not assume that $(XY, XZ, YZ)$ holds, while using $G^2\left[(XZ, YZ)|(XY, XZ, YZ)\right]$ does assume that the model of homogeneous association holds.

Disadvantages of the goodness-of-fit test as a test of conditional independence

1. It has lower power.

2. It has more $df$ than the Wald test, the CMH, and the LR test (i.e.,$G^2\left[(XZ, YZ)|(XY, XZ, YZ)\right]$).

# Ordinal Conditional Association



If the categories of one or both variables are ordered, then there are more powerful ways of testing for conditional independence.

With respect to models, we can use a generalized linear by linear model, more specifically a "homogeneous linear by linear association" model.

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \beta u_i v_j + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

where $u_i$ are scores for the levels of variable $X$, and $v_j$ are scores for the levels of variable $Y$.

Notes:

- The model of conditional independence is a special case of this model; that is, $\beta = 0$
- This model is a special case of the homogeneous association model.

# Ordinal Conditional Association



Example: Using as equally spaced scores for SES (i.e., $u_1 = 1$, $u_2 = 2$, and $u_3 = 3$), and unequally spaced scores for program type (i.e., $v_1 = 1$, $v_2 = 2$, and $v_3 = 4$), we fit the model

$$\log(\mu_{ijk}) = \lambda + \lambda_i^S + \lambda_j^P + \lambda_k^G + \beta u_i v_j + \lambda_{ik}^{SG} + \lambda_{jk}^{PG}$$

| | Goodness-of-fit Test | | | Likelihood Ratio Test | | |
|---|---|---|---|---|---|---|
| Model | $df$ | $G^2$ | $p$ | $\Delta df$ | $\Delta G^2$ | $p$ |
| $(GS, GP, SP)$ | 4 | 1.970 | .74 | — | — | — |
| $(GS, GP, SP)$–L$\times L$ | 7 | 7.476 | .38 | 3 | 5.505 | .138 |
| $(GS, GP)$ | 8 | 55.519 | $< .0001$ | 1 | 48.043 | $< .001$ |

From before...

| | Goodness-of-fit Test | | | Likelihood Ratio Test | | |
|---|---|---|---|---|---|---|
| Model | $df$ | $G^2$ | $p$ | $\Delta df$ | $\Delta G^2$ | $p$ |
| $(GS, GP, SP)$ | 4 | 1.970 | .74 | — | — | — |
| $(GS, GP)$ | 8 | 55.519 | $< .0001$ | 4 | 53.548 | $< .001$ |

# Example Continued...



- The null hypothesis for the likelihood ratio test statistic (in the last row of top table) is $H_O : \beta = 0$ with $df = 1$; whereas, in the lower table, it is

$$H_O : \quad \text{all} \quad \lambda_{ij}^{SP} = 0 \quad \text{with} \quad df = 4$$

- Comparing $G^2/df$ for the two tests,

$$53.548/4 = 13.387 \qquad \text{versus} \qquad 48.043/1 = 48.043$$

- Conclusion: If data exhibit linear by linear partial association, then using scores gives you a stronger (more powerful) test of conditional independence.

- The Wald statistic for $\beta$ equals $43.939$, $df = 1$, and $p < .0001$. This is comparable to the new likelihood ratio test statistic.

# Estimated Partial Odds Ratios



$\hat{\beta} = .3234$. The estimated partial odds ratio equals

$$\hat{\theta}_{SP(k)} = \exp\left[.3234(u_i - u_{i'})(v_j - v_{j'})\right]$$

For example, the smallest partial odds ratio is for low and middle SES and votech and general programs,

$$\hat{\theta}_{SP(k)} = \exp\left[.3234(2-1)(2-1)\right] = \exp(.3234) = 1.38$$

The largest partial odds ratio is for low and high SES and votech and academic programs equals

$$\hat{\theta}_{SP(k)} = \exp\left[.3234(3-1)(4-1)\right] = \exp(1.9404) = 6.96$$

# So far. . .



| Variable | | Type of Test | | |
|---|---|---|---|---|
| | | Likelihood | | (Generalized) |
| Row | Column | Ratio | Wald | CMH |
| Nominal | Nominal | X | X | |
| Nominal | Ordinal | | | |
| Ordinal | Ordinal | X | X | |

Next, the model based nominal–ordinal case.

For the nominal–ordinal case, we only put in scores for the categories of the ordinal variable and estimate a $\beta$ for each category of the nominal variable.

# Nominal–Ordinal Case



For example, if only have put in scores for SES, we fitting the model

$$\log(\mu_{ijk}) = \lambda + \lambda_i^S + \lambda_j^P + \lambda_k^G + \beta_j^P u_i + \lambda_{ik}^{SG} + \lambda_{jk}^{PG}$$

where $u_i$ are scores for SES (i.e., $u_1 = 1$, $u_2 = 2$, and $u_3 = 3$), and $\beta_j^P$ are estimated parameters.

| | Goodness-of-fit Test | | | Likelihood Ratio Test | | |
|---|---|---|---|---|---|---|
| Model | $df$ | $G^2$ | $p$ | $\Delta df$ | $\Delta G^2$ | $p$ |
| $(GS, GP, SP)$ | 4 | 1.970 | .74 | — | — | — |
| $(GS, GP)$ | 8 | 55.519 | $< .0001$ | 4 | 53.548 | $< .001$ |
| $(GS, GP, SP)$–L×$L$ | 7 | 7.476 | .38 | 3 | 5.505 | .14 |
| $(GS, GP)$ | 8 | 55.519 | $< .0001$ | 1 | 48.043 | $< .001$ |
| $(GS, GP, SP)$ with $u_i$ | 6 | 4.076 | .62 | 2 | 2.106 | .35 |
| $(GS, GP)$ | 8 | 55.519 | $< .0001$ | 2 | 51.443 | $< .001$ |

For the nominal–ordinal model, $\hat{\beta}^P_{votech} = -.8784$ & $\hat{\beta}^P_{gen} = -.8614$,
$\Longrightarrow$ the "best" scores for VoTech and General programs are much closer together than we had been assuming.

# So we have now discussed,

| Variable | | Type of Test | | |
|---|---|---|---|---|
| | | Likelihood | | (Generalized) |
| Row | Column | Ratio | Wald | CMH |
| Nominal | Nominal | X | X | |
| Nominal | Ordinal | X | X | |
| Ordinal | Ordinal | X | X | |

To complete our table, we need to talk about efficient score tests for testing conditional independence for each of the three cases.

The efficient score test of conditional independence of $X$ and $Y$ given $Z$ for an $I \times J \times K$ cross-classification is a generalization of the Cochran-Mantel-Haenszel statistic, which we discussed as a way to test conditional independence in $2 \times 2 \times K$ tables.

For each of three cases, the test statistic is a Generalized CMH Statistic.