ST3241 Categorical Data Analysis I
Loglinear Models

2×2 Models For Contingency Tables

Two-way Tables

- Consider an $I \times J$ contingency table that crossclassifies a sample of n subjects on two categorical responses.
- Let Y_{ij} be the observed cell frequency and μ_{ij} be the expected cell frequency of the (i, j)-th cell.
- Then we assume that the cell counts Y_{ij} are independent having Poisson(μ_{ij}) distribution.
- Note that, if π_{ij} is the cell probability, then $\mu_{ij} = n\pi_{ij}$.

Independence Model

- Under statistical independence of the row and column classifications, $\pi_{ij} = \pi_{i+}\pi_{+j}$ and hence $\mu_{ij} = n\pi_{i+}\pi_{+j}$.
- Denote the row variable by X and the column variable by Y.
- The formula expressing independence is multiplicative. Thus, $\log \mu_{ij}$ is additive

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$$

  for a row effect λ_i^X and a column effect λ_j^Y.
- This is the loglinear model of independence.
- The null hypothesis of independence between two categorical variables is simply the hypothesis that this model holds.

Example: Belief in Afterlife

Observed Frequency		Fitted Value		Log Fitted Value	
435	147	432.10	149.90	6.069	5.010
375	134	377.90	131.10	5.935	4.876

Example: Belief in Afterlife

Parameter	Set 1	Set 2	Set 3
λ	4.876	6.069	5.472
λ_1^X	0.134	0	0.0067
λ_2^X	0	-0.134	-0.067
λ_1^Y	1.059	0	0.529
λ_2^Y	0	-1.059	-0.529

Some SAS Codes

```
data after;
  input female $ belief $ count;
datalines;
  Female Yes 435
  Female No 147
  Male Yes 375
  Male No 134
;
run;
```

SAS Codes

```
proc genmod data=after order=data;
  class belief female;
  model count= female belief/ dist=poisson;
  output out=temp p=predict;
run;
proc print data=temp;
  var female belief count predict;
run;
```

Partial Output

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	1	0.1620	0.1620
Scaled Deviance	1	0.1620	0.1620
Pearson Chi-Square	1	0.1621	0.1621
Scaled Pearson X2	1	0.1621	0.1621
Log Likelihood	5164.1959		

Algorithm converged.

Partial Output

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Conf Limits		Chi-Square	Pr>ChiSq
Intercept	1	4.8760	0.0679	4.7429	5.0090	5160.87	<.0001
female Female	1	0.1340	0.0607	0.0151	0.2530	4.88	0.0272
female Male	0	0.0000	0.0000	0.0000	0.0000	-	-
belief Yes	1	1.0587	0.0692	0.9230	1.1944	233.83	<.0001
belief No	0	0.0000	0.0000	0.0000	0.0000	-	-
Scale	0	1.0000	0.0000	1.0000	1.0000		

Partial Output

Obs	female	belief	count	predict
1	Female	Yes	435	432.099
2	Female	No	147	149.901
3	Male	Yes	375	377.901
4	Male	No	134	131.099

Interpretations of Parameters

- For $I \times J$ tables, loglinear models treat the $N = IJ$ cell counts as N independent observations of a Poisson random component.
- The data are the N cell counts rather than the individual classifications of the n subjects.
- The model does not distinguish between response and explanatory variables.
- Differences between two parameters for a given variable relate to the log odds of making response, relative to the other, on that variable.
- If the response is binary, one can use logit models directly but loglinear models are useful for modeling relationships among two or more categorical response variables.

In $I \times 2$ Table

- Response Y has only 2 levels.
- In row i, the logit for the probability π that $Y = 1$ is:

$$\begin{aligned}\log(\frac{\pi_i}{1-\pi_i}) &= \log(\frac{\mu_{i1}}{\mu_{i2}}) = \log \mu_{i1} - \log \mu_{i2} \\ &= (\lambda + \lambda_i^X + \lambda_1^Y) - (\lambda + \lambda_i^X + \lambda_2^Y) \\ &= \lambda_1^Y - \lambda_2^Y\end{aligned}$$

- logit for Y does not depend on the levels of X.

Saturated Model

- A more complex model

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

- Here λ_{ij}^{XY} are the association terms reflecting deviation from the independence or interaction effect between i-th category of X and j-th category of Y.
- This is the saturated loglinear model in a two-way table.

Interpretation of Interaction

- There is a direct relationship between log odds ratios and $\{\lambda_{ij}^{XY}\}$ association parameters.
- In a 2×2 table,

$$\begin{aligned}
\log\theta &= \log(\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}) \\
&= \log\mu_{11} + \log\mu_{22} - \log\mu_{12} - \log\mu_{21} \\
&= (\lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}) + (\lambda + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}) \\
&- (\lambda + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}) - (\lambda + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY}) \\
&= \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}
\end{aligned}$$

Example: Belief in Afterlife

Association Parameter	Set 1	Set 2	Set 3
λ_{11}^{XY}	0.056	0.014	0.0
λ_{12}^{XY}	0.0	-0.014	0.0
λ_{21}^{XY}	0.0	-0.014	0.0
λ_{22}^{XY}	0.0	0.014	0.0

Example: Belief in Afterlife

- From the data, the odds ratio is
  $(435 \times 134)/(147 \times 375) = 1.057$.
- Therefore, the log odds ratio $= 0.056$
- From the association parameters

$$\hat{\lambda}_{11}^{XY} + \hat{\lambda}_{22}^{XY} - \hat{\lambda}_{12}^{XY} - \hat{\lambda}_{21}^{XY} = 0.056$$

Notes

- In $I \times J$ tables, only $(I-1)(J-1)$ parameters are non-redundant.
- These *interaction* parameters in the saturated model are coefficients of cross products of $(I-1)$ dummy variables for X with $(J-1)$ dummy variables for Y.
- Tests of independence analyze whether these $(I-1)(J-1)$ parameters equal 0, so they have residual $d.f. = (I-1)(J-1)$.

Notes

- The saturated model has as many parameters as it has Poisson observations.
- Thus, it gives a perfect fit.
- This model is also a hierarchical model as it includes all lower order terms composed from variables contained in a higher order term in the model

Three-way Tables

- Denote the cell expected frequencies in the contingency table by $\{\mu_{ijk}\}$.
- Single factor terms in loglinear models for $\{\mu_{ijk}\}$ represent marginal distributions.
- E.g. including $\hat{\lambda}_i^X$ in the model forces the fitted values to have the same totals at the various levels of X as do the observed data.
- Partial associations between variables correspond to two factor terms.

Partial Association Models

- Consider the loglinear model

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

- Since it contains an $X - Z$ two factor term, it permits association between X and Z controlling for Y.
- This model also permits a $Y - Z$ association, controlling for X.
- It does not contain a two-factor term for $X - Y$ association.
- It specifies conditional independence between X and Y, controlling for Z.

Partial Association Models

- We symbolize the previous model as (XZ, YZ).
- For $2 \times 2 \times k$ tables, this model corresponds to the hypothesis tested using the *Cochran – Mantel – Haenszel statistic*.
- The model that contains only single factor terms, denoted by (X, Y, Z), is called the *mutual independence model*.
- It treats each pair of variables as independent

Partial Association Models

- The model which permits all three pairs of variables to be conditionally dependent is

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

- For this model, the conditional odds ratios between any two variables are identical at each level of the third variable.
- We refer to this model as the homogeneous association model and symbolize it by (XY, YZ, XZ).

Saturated Model

- The model:

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$$

- The three factor term pertains to three-factor interaction.
- Denoted by (XYZ), this model permits the odds ratio between any two variables to vary across levels of the third variable.
- It provides a perfect fit in a three-way table.

Interpreting Model Parameters

- Interpretation of loglinear model parameters refer to their highest order terms.
- E.g. interpretations for the homogeneous association model use the two factor terms to describe associations.
- The two-factor parameters relate directly to conditional odds ratios.
- $X - Y$ conditional odds ratio $\theta_{XY(k)}$ in the homogeneous association model can be shown to be

$$\log \theta_{XY(k)} = \log(\frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}) = \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}$$

  which does not depend on k.

Example: Alcohol, Cigarette and Marijuana Use For High School Seniors

Alcohol Use	Cigarette Use	Marijuana Use: Yes	Marijuana Use: No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

Example: SAS Codes

```
data drug;
  input a $ c $ m $ count;
datalines;
Yes   Yes   Yes   911
Yes   Yes   No   538
Yes   No   Yes   44
Yes   No   No   456
No   Yes   Yes   3
No   Yes   No   43
No   No   Yes   2
No   No   No   279
;
run;
```

Example: SAS Codes

```
proc catmod data=drug order=data;
  weight count;
  model a*c*m =_response_/
    pred=freq;
  loglin a|m c|m;
run;
quit;
```

Partial Output

```
The CATMOD Procedure
                    Data Summary
 Response            a*c*m   Response Levels      8
 Weight Variable     count   Populations          1
 Data Set            DRUG    Total Frequency   2276
 Frequency Missing   0       Observations         8
Population Profiles
Sample   Sample Size
++++++++++++++++++++
1                 2276
```

Partial Output

Maximum Likelihood Analysis

Maximum likelihood computations converged.

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr> ChiSq
a	1	198.52	<.0001
m	1	155.20	<.0001
a*m	1	83.00	<.0001
c	1	292.68	<.0001
c*m	1	401.17	<.0001
Likelihood Ratio	2	187.75	<.0001

Partial Output

Analysis of Maximum Likelihood Estimates

Parameter		Estimate	Standard Error	Chi-Square	Pr>ChiSq
a	Yes	1.5949	0.1132	198.52	<.0001
m	Yes	-1.4731	0.1182	155.20	<.0001
a*m	Yes Yes	1.0313	0.1132	83.00	<.0001
c	Yes	0.6885	0.0402	292.68	<.0001
c*m	Yes Yes	0.8061	0.0402	401.17	<.0001

Partial Output

Maximum Likelihood Predicted Values for Frequencies

			Observed		Predicted		
a	c	m	Frequency	Standard Error	Frequency	Standard Error	Residual
Yes	Yes	Yes	911	23.37434	909.2396	23.36195	1.760417
Yes	Yes	No	538	20.26889	438.8404	17.15391	99.15957
Yes	No	Yes	44	6.568819	45.76042	6.679323	-1.76042
Yes	No	No	456	19.09554	555.1596	18.96775	-99.1596
No	Yes	Yes	3	1.730909	4.760418	2.126048	-1.76042
No	Yes	No	43	6.495199	142.1596	8.562114	-99.1596
No	No	Yes	2	1.413592	0.239583	0.112401	1.760417
No	No	No	279	15.64606	179.8404	10.27909	99.15957

Fitted Values for Loglinear Models

			Loglinear Models				
Alcohol Use	Cigarette Use	Marijuana Use	(A,C,M)	(AC,M)	(AM,CM)	(AC,AM,CM)	(ACM)
Yes	Yes	Yes	540.0	611.2	909.24	910.4	911
		No	740.2	837.8	438.84	538.6	538
	No	Yes	282.1	210.9	45.76	44.6	44
		No	386.7	289.1	555.16	455.4	456
No	Yes	Yes	90.6	19.4	4.76	3.6	3
		No	124.2	26.6	142.16	42.4	43
	No	Yes	47.3	118.5	0.24	1.4	2
		No	64.9	162.5	179.84	279.6	279

Estimated Odds Ratios

	Conditional Association			Marginal Association		
Model	A-C	A-M	C-M	A-C	A-M	C-M
(A,C,M)	1.0	1.0	1.0	1.0	1.0	1.0
(AC,M)	17.7	1.0	1.0	17.7	1.0	1.0
(AM,CM)	1.0	61.9	25.1	2.7	61.9	25.1
(AC,AM,CM)	7.8	19.8	17.3	17.7	61.9	25.1
(ACM) Level 1	13.8	24.3	17.5	17.7	61.9	25.1
(ACM) Level 2	7.7	13.5	9.7			

Example

- The fit for model (AC, AM, CM) is close to the observed data.
- The other models seem to fit poorly.
- The model (AM, CM) implies conditional independence between alcohol use and cigarette use, controlling for marijuana use, and yields odds ratios of 1.0 for the $A - C$ conditional association.
- The entry 2.7 for the $A - C$ marginal association for this model is the fitted odds ratio for the marginal $A - C$ fitted table.
- Model (AC, AM, CM) permits all pairwise associations but maintains homogeneous odds ratios between two variables at each level of the third variable.
- The estimated conditional odds ratios equal 1.0 for each pairwise term not appearing in a model.
- The estimates of conditional and marginal odds ratios are highly dependent on the model.

Fitting Loglinear Models

- Some loglinear models have explicit formulas for the fitted values $\{\hat{\mu}_{ijk}\}$ in terms of $\{n_{ijk}\}$.
- For example, the model (XZ, YZ) of $X - Y$ conditional independence has

$$\hat{\mu}_{ijk} = \frac{n_{i+k} n_{+jk}}{n_{++k}}$$

- Many loglinear models do not have any direct estimates.
- ML estimation then uses iterative methods.

Chi-Square Goodness-of-Fit Tests

- Consider the null hypothesis that the expected frequencies for a three-way table satisfy a given loglinear model.
- The LR and Pearson Chi-square statistics are:

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}\log(\frac{n_{ijk}}{\hat{\mu}_{ijk}}),$$

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{(n_{ijk}-\hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}}$$

- The degrees of freedom equals the number of cell counts minus the number of non-redundant parameters in the model.
- The saturated model has $d.f. = 0$.

Example: Drug Use Data

Model	G^2	χ^2	df	P-value
(A,C,M)	1286.0	1411.4	4	<0.001
(A,CM)	534.2	505.6	3	<0.001
(C,AM)	939.6	824.2	3	<0.001
(M,AC)	843.8	704.9	3	<0.001
(AC,AM)	497.4	443.8	2	<0.001
(AC,CM)	92.0	80.8	2	<0.001
(AM,CM)	187.8	177.6	2	<0.001
(AC,AM,CM)	0.4	0.4	1	0.54
(ACM)	0.0	0.0	0	

Residuals

- One can study the quality of the fit more closely by studying cell residuals.
- They may indicate why a particular model does not fit well or highlight cells that display lack of fit.
- We may use *adjusted residuals* or *Pearson residuals*.
- When the model holds, the adjusted residuals have approximately standard normal distribution.
- So, absolute values of *adjusted residuals* larger than 2 when there are few cells and larger than 3 when there are many cells , indicate lack of fit.

Example: SAS Codes (AM,CM)

```
proc genmod data=drug ;
  class a c m;
  model count= a c m a*m c*m / dist=p;
  output out=temp reslik=reslik p=pred;
run;
data temp;
  set temp;
  id = _n_;
  rename reslik=adjres1;
  rename pred=fitted1;
run;
```

Example: SAS Codes (AC,AM,CM)

```
proc genmod data=drug;
  class a c m;
  model count= a c m a*m c*m a*c/ dist=p;
  output out=temp1 reslik=reslik p=pred;
run;
data temp1;
  set temp1;
  id = _n_;
  rename reslik=adjres2;
  rename pred=fitted2; run;
```

Example: SAS Codes (Print Output)

```
data combo;
  merge temp temp1;
  by id;
run;
proc print data=combo noobs;
  var a c m count fitted1 adjres1 fitted2
  adjres2;
run;
```

Partial Output

a	c	m	count	fitted1	adjres1	fitted2	adjres2
Yes	Yes	Yes	911	909.240	3.6955	910.383	0.63332
Yes	Yes	No	538	438.840	12.7451	538.617	-0.63333
Yes	No	Yes	44	45.760	-3.6956	44.617	-0.63336
Yes	No	No	456	555.160	-12.8498	455.383	0.63332
No	Yes	Yes	3	4.761	-3.7093	3.617	-0.63848
No	Yes	No	43	142.160	-13.7941	42.383	0.63329
No	No	Yes	2	0.240	2.3852	1.383	0.60617
No	No	No	279	179.840	12.4902	279.617	-0.63333

Tests About Partial Association

- Test about partial association by comparing different loglinear models.
- For model (AC, AM, CM), the null hypothesis of no partial association between alcohol use and cigarette smoking states that the λ^{AC} term equals zero.
- To test whether the simpler model (AM, CM) of $A - C$ conditional independence holds against the alternative that the model (AC, AM, CM) holds

Likelihood Ratio Tests

- The likelihood ratio statistic $-2(L_0 - L_1)$ is identical to the goodness-of-fit G^2 statistics (deviance) for the model without that term and the model containing that term.
- The $d.f.$ equals the difference between the corresponding $d.f.$ values.
- The test statistic for testing $\lambda^{AC} = 0$ in model (AC, AM, CM) is the difference
  $G^2[(AM, CM)|(AC, AM, CM)] =$
  $G^2(AM, CM) - G^2(AC, AM, CM)$
  between $G^2 = 187.8(d.f. = 2)$ for model (AM, CM) and
  $G^2 = 0.4(d.f. = 1)$ for model (AC, AM, CM).

Example: Drug Use Data

Model	G^2	χ^2	$d.f.$	$P-value$
(A,C,M)	1286.0	1411.4	4	<0.001
(A,CM)	534.2	505.6	3	<0.001
(C,AM)	939.6	824.2	3	<0.001
(M,AC)	843.8	704.9	3	<0.001
(AC,AM)	497.4	443.8	2	<0.001
(AC,CM)	92.0	80.8	2	<0.001
(AM,CM)	187.8	177.6	2	<0.001
(AC,AM,CM)	0.4	0.4	1	0.54
(ACM)	0.0	0.0	0	

Example

- The difference of 187.4 is based on $d.f. = 2 - 1 = 1$.
- The small P-value provides strong evidence against the null hypothesis and in favor of an $A - C$ partial association.
- The statistics comparing models (AC, CM) and (AC, AM) with model (AC, AM, CM) also provides strong evidence of $A - M$ and $C - M$ partial associations.
- We should use the model (AC, AM, CM) rather than any simpler models.

Notes

- The sample size can strongly influence results of any inferential procedure.
- One is more likely to detect an effect of given size as the sample size increases.
- For small sizes, reality may be much more complex than indicated by the simplest model that passes the goodness-of-fit test.
- For large sample sizes, statistically significant effects can be weak and unimportant.
- This is a limitation of hypothesis testing.

Confidence Intervals For Odds Ratios

- ML estimators of parameters have large sample normal distributions.
- For models in which the highest order terms are twofactor associations, the estimates refer to the conditional log odds ratios.
- One can use the estimates along with their standard errors to construct confidence intervals for true log odds ratios and then exponentiate them to form intervals for odds ratios.

Example

- Assume the model (AC, AM, CM) holds.
- We estimate the conditional odds ratio between alcohol use and cigarette use.
- The software reports $\hat{\lambda}_{11}^{AC} = 2.054$ with $ASE = 0.174$
- The lone nonzero term equals the estimated conditional log odds ratio.
- A 95% C.I. for the true conditional log odds ratio is

$$2.054 \pm 1.96 \times (0.174) = (1.71, 2.39)$$

- Thus, a 95% C.I. for the true conditional odds ratio is

$$(e^{1.71}, e^{2.39}) = (5.5, 11.0)$$

Example

- For model (AC, AM, CM), the 95% confidence intervals are $(8.0, 49.2)$ for the $A - M$ conditional odds ratio and $(12.5, 23.8)$ for the $C - M$ conditional odds ratio.
- The intervals are wide, but these associations also are strong.
- In summary, there is a strong tendency for users of one drug to be users of a second drug, and this is true both for users and nonusers of the third drug.

Automobile Accident Example

			Injury	
Gender	Location	Seat Belt	No	Yes
Female	Urban	No	7287	996
		Yes	11587	759
	Rural	No	3246	973
		Yes	6134	757
Male	Urban	No	10381	812
		Yes	10969	380
	Rural	No	6123	1084
		Yes	6693	513

Example

- The table refers to observations of 68,694 passengers in autos and light trucks involved in accidents in the state of Maine in 1991.
- The table classifies passengers by *gender* (G), *location* of accident (L), *seat belt* use (S), and *injury* (I).

Four-way Tables

- Basic concepts of three-way tables extend readily to multi-way tables.
- We consider a four-way table with variables $W, X, Y,$ and Z.
- Interpretations are simplest when there are no three-factor interaction terms.
- The homogeneous association model is (WX, WY, WZ, XY, XZ, YZ).
- Here each pair of variables is conditionally dependent, with the same odds ratios at each combination of levels of the other two variables.

Four-way Tables

- An absence of a two factor term implies conditional independence for those variables.
- Model (WX, WY, WZ, XZ, YZ) does not contain an $X - Y$ term, so it treats X and Y as conditionally independent at each combination of levels of W and Z.
- A model could contain any of the four possible three factor interaction terms: WXY, WXZ, WYZ, XYZ.
- The saturated model contains all these terms plus a four factor interaction term.

Example: SAS Codes

```
data injury;
  input G L S I count;
datalines;
0 0 0 0 7287
0 0 0 1 996
...
1 1 1 1 5
;
run;
```

SAS Codes

```
ods listing close;
proc catmod data=injury;
  weight count;
  model G*I*L*S= _response_/ pred=freq;
  loglin g|i g|l g|s i|l i|s l|s;
  ods output predictedfreqs=temp1;
run;
quit;
data temp1 (keep=p1 functionnum);
  set temp1;
  rename predfunction=p1;
run;
```

SAS Codes

```
proc catmod data=injury;
  weight count;
  model G*I*L*S=_response_/ pred=freq;
  loglin g|l|s g|i i|l i|s;
  ods output predictedfreqs=temp2 anova=temp3;
run;
quit;
ods output close;
ods listing;
data temp2;
  set temp2;
  rename predfunction=p2;
run;
```

SAS Codes

```
data combo;
  merge temp1 temp2;
  by functionnum;
  Male=G+0;
  Location=L+0;
  Seat=S+0;
  Injury=I+0;
  rename obsfunction = observed;
run;
proc format;
  value male 0='Female' 1='Male';
  value location 0='Urban' 1='Rural';
  value Yesno 0='No' 1='Yes';
run;
```

SAS Codes

```
proc sort data=combo;
  by male location seat injury;
run;
proc print data= combo noobs;
  format male male.  location location.
  seat yesno.  injury yesno.;
  var male location seat injury observed
  p1 p2;
run;
```

Partial Output

Male	Location	Seat	Injury	observed	p1	p2
Female	Urban	No	No	7287	7166.369	7273.214
Female	Urban	No	Yes	996	993.0169	1009.786
Female	Urban	Yes	No	11587	11748.31	11632.62
Female	Urban	Yes	Yes	759	721.3055	713.3779
Female	Rural	No	No	3246	3353.829	3254.662
Female	Rural	No	Yes	973	988.7848	964.3382
Female	Rural	Yes	No	6134	5985.493	6093.502
Female	Rural	Yes	Yes	757	781.8927	797.4979
Male	Urban	No	No	10381	10471.5	10358.93
Male	Urban	No	Yes	812	845.1187	834.0683
Male	Urban	Yes	No	10969	10837.83	10959.23
Male	Urban	Yes	Yes	380	387.5588	389.7677
Male	Rural	No	No	6123	6045.306	6150.192
Male	Rural	No	Yes	1084	1038.08	1056.808
Male	Rural	Yes	No	6693	6811.371	6697.644
Male	Rural	Yes	Yes	513	518.2429	508.3564

Example: Automobile Accidents

			Injury		(GI,GL,GS,IL,IS,LS)		(GLS,GI,IL,IS)	
Gender	Location	seat Belt	No	Yes	No	Yes	No	Yes
Female	Urban	No	7287	996	7166.4	993.0	7273.2	1009.8
		Yes	11587	759	11748.3	721.3	11632.6	713.4
	Rural	No	3246	973	3353.8	988.8	3254.7	964.3
		Yes	6134	757	5985.5	781.9	6093.5	797.5
Male	Urban	No	10381	812	10471.5	845.1	10358.9	834.1
		Yes	10969	380	10837.8	387.6	10959.2	389.8
	Rural	No	6123	1084	6045.3	1038.1	6150.2	1056.8
		Yes	6693	513	6811.4	518.2	6697.6	508.4

Example: Goodness of Fit Tests

Model	G^2	$d.f.$	P-value
(G,I,L,S)	2792.8	11	<0.0001
(GI,GL,GS,IL,IS,LS)	23.4	5	<0.001
(GIL,GIS,GLS,ILS)	1.3	1	0.25
(GIS,GL,IL,LS)	22.8	4	<0.001
(GLS,GI,IL,IS)	7.5	5	0.11
(ILS,GI,GL,GS)	20.6	4	<0.001

16 cells - 4p - 1

Discussions

- Model (G, I, L, S), which implies mutual independence of the four variables, fits very poorly.
- Model (GI, GL, GS, IL, IS, LS) fits much better but still has lack of fit.
- Model (GIL, GIS, GLS, ILS) contains all three factor interactions seems to fit well, but is quite complex and difficult to interpret.
- This suggests studying models that are more complex than (GI, GL, GS, IL, IS, LS) but simpler than (GIL, GIS, GLS, ILS).

Three-Factor Interaction

- Interpretations are more complicated when a model contains three-factor interaction terms.
- Of the four possible models,(GLS, GI, IL, IS) appears to fit best.
- For model (GLS, GI, IL, IS), each pair of variables is conditionally dependent.
- At each level of I the association between G and L or between G and S or between L and S varies across the levels of the remaining variables.
- For this model, it is inappropriate to interpret the $G - L$, $G - S$ and $L - S$ two factor terms on their own.

Three-Factor Interactions

- One would not convert $\hat{\lambda}^{GS}$ to a fitted $G - S$ odds ratio, because the presence of the GLS three-factor interaction term implies that the $G - S$ odds ratio varies across the levels of L.
- Since I does not occur in a three factor interaction, the conditional odds ratio between I and each variable is the same at each combination of levels of the other two variables.
- When a model has a three factor interaction term but no term of higher order than that, one can study the interaction by calculating fitted odds ratios between two variables at each level of the third.

Estimated Conditional Odds Ratios

	Loglinear Models	
Odds Ratio	(GI,GL,GS,IL,IS,LS)	(GLS,GI,IL,IS)
$G-I$	0.58	0.58
$I-L$	2.13	2.13
$I-S$	0.44	0.44
$G-L(S=no)$	1.23	1.33
$G-L(S=yes)$	1.23	1.17
$G-S(L=urban)$	0.63	0.66
$G-S(L=rural)$	0.63	0.58
$L-S(G=female)$	1.09	1.17
$L-S(G=male)$	1.09	1.03

Statistical vs. Practical Significance

- Model (GLS, GI, IL, IS) seems to fit much better than (GI, GL, GS, IL, IS, LS).
- The difference in G^2 values of $23.4 - 7.5 = 15.9$ being based on $d.f. = 5 - 4 = 1$.
- The fitted odds ratios, however, show that the degree of three-factor interaction is weak.
- The fitted odds ratio between any two of G,L, and$S$ is similar at both levels of the third variable.
- The significantly better fit of model (GLS, GI, IL, IS) mainly reflects the enormous sample size.

Effect of Large Sample Sizes

- Large sample provides small standard errors.
- A statistically significant effect need not be important in a practical sense.
- With huge samples, it is crucial to focus on estimation rather than hypothesis testing.
- e.g. the model (GI, GL, GS, IL, IS, LS) is adequate for practical purposes. .. Simpler models are easier to summarize.
- One should not use goodness-of-fit tests alone to select a final model.

Dissimilarity Index

- For a table of arbitrary dimension with cell counts $\{n_i = np_i\}$ and fitted values $\{\hat{\mu}_i = n\hat{\pi}_i\}$ one can summarize the closeness of the model fit to the sample data by the dissimilarity index

$$D = \sum |n_i - \hat{\mu}_i|/(2n) = \sum |p_i - \hat{\pi}_i|/2$$

- This index takes values between 0 and 1, with smaller values representing a better fit.
- It represents the proportion of sample cases that must move to different cells in order for the model to achieve a perfect fit.

Dissimilarity Index

- The dissimilarity index D estimates a corresponding index Δ that describes model lack-of-fit in the population sampled.
- The value $\Delta = 0$ occurs when the model holds perfectly.
- In that case D overestimates Δ, substantially so for small samples, because of sampling variation.
- When the model does not hold, for sufficiently large n, the goodness-of-fit statistics G^2 and χ^2 will be large, showing lack-of-fit.
- The estimator D then reveals whether the lack of fit suggested by those statistics is important in practical sense.
- $D < 0.03$ suggests that the sample data follow the model quite closely, even though the model is not $perfect$.

Some SAS Codes

- As a continuation of the previous SAS codes, we can add the following statements to compute the dissimilarity indices for those models.

```
proc sql;
  select sum(abs(observed-p1))/
    (2*sum(observed)) as d1,
  sum( abs(observed-p2) ) / (2*sum(observed) )
  as d2
  from combo;
quit;
```

Some SAS Codes

d1	d2
0.008219	0.002507

- For either model, moving less than 1% of the data yields a perfect fit.
- The small value of D for the model (GI, GL, GS, IL, IS, LS) suggests that in practical terms, this model provides a decent fit.

Loglinear-Logit Connection

- Consider the loglinear model of homogeneous association in three-way tables

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

- Suppose Y is binary, and we treat it as a response and X and Z as explanatory.
- Let π denote the probability that $Y = 1$, which depends on the levels of X and Z.

Loglinear-Logit Connection

- The logit for Y is

$$\begin{aligned}
\text{logit}(\pi_{ik}) &= \log(\frac{\pi_{ik}}{1-\pi_{ik}}) = \log(\frac{P(Y=1|X=i,Z=k)}{P(Y=0|X=i,Z=k)}) \\
&= \log(\frac{\mu_{i1k}}{\mu_{i2k}}) = \log(\mu_{i1k}) - \log(\mu_{i2k}) \\
&= (\lambda + \lambda_i^X + \lambda_1^Y + \lambda_k^Z + \lambda_{i1}^{XY} + \lambda_{ik}^{XZ} + \lambda_{1k}^{YZ}) \\
&\quad -(\lambda + \lambda_i^X + \lambda_2^Y + \lambda_k^Z + \lambda_{i2}^{XY} + \lambda_{ik}^{XZ} + \lambda_{2k}^{YZ}) \\
&= (\lambda_1^Y - \lambda_2^Y) + (\lambda_{i1}^{XY} - \lambda_{i2}^{XY}) + (\lambda_{1k}^{YZ} - \lambda_{2k}^{YZ})
\end{aligned}$$

Loglinear-Logit Connection

- The first term is a constant which does not depend on i or k.
- The second term depends on the level i of X.
- The third term depends on the level k of Z.
- The logit has the additive form

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

Accident Data Revisited

- The model (GLS, GI, IL, IS) has the form

$$\begin{aligned}\log \mu_{gils} &= \lambda + \lambda_g^G + \lambda_i^I + \lambda_l^L + \lambda_s^S + \\ &\quad \lambda_{gi}^{GI} + \lambda_{gl}^{GL} + \lambda_{gs}^{GS} + \lambda_{il}^{IL} + \lambda_{is}^{IS} + \lambda_{ls}^{LS} + \lambda_{gls}^{GLS}\end{aligned}$$

- One can treat injury (I) as a response variable and gender (G), location (L) and seat belt use (S) as explanatory variables.
- Let π denote the probability of injury.
- Forming $\text{logit}(\pi_{gls})$ at each combination of levels of $G, L,$ and S, one can show that the above loglinear model is equivalent to logit model

$$\text{logit}(\pi_{gls}) = \alpha + \beta_g^G + \beta_l^L + \beta_s^S$$

Interpretations

- Here, $G, L,$ and S all affect I, but without interacting.
- The parameters in the two models are related by

$$\alpha = \lambda_1^I - \lambda_2^I, \beta_g^G = \lambda_{g1}^{GI} - \lambda_{g2}^{GI}, \beta_l^L = \lambda_{1l}^{IL} - \lambda_{2l}^{IL}, \beta_s^S = \lambda_{1s}^{IS} - \lambda_{2s}^{IS}$$

- In the logit calculation, all terms in the loglinear model not having the injury index i in the subscript are cancelled.

Example: Birth Control Data

	Teenage Birth Control			
Premarital Sex	Strongly Disagree	Disagree	Agree	Strongly Agree
Always wrong	81	68	60	38
	(42.4)	(51.2)	(86.4)	(67.0)
	7.6	3.1	-4.1	-4.8
Almost always wrong	24	26	29	14
	(16.0)	(19.3)	(32.5)	(25.2)
	2.3	1.8	-0.8	-2.8
Wrong only sometimes	18	41	74	42
	(30.1)	(36.3)	(61.2)	(47.4)
	-2.7	1.0	2.2	-1.0
Not wrong at all	36	57	161	157
	(70.6)	(85.2)	(143.8)	(111.4)
	-6.1	-4.6	2.4	6.8

Example: Birth Control Data

- The loglinear model of independence has goodness-of-fit statistic $G^2 = 127.6$ based on $d.f. = 9$.
- The model fits poorly, providing strong evidence of dependence.
- But, adding the interaction term makes the model saturated and of little use.

Example: Birth Control Data

- The table also contains fitted values and adjusted residuals for the independence model.
- The residuals in the corners of the table are very large.
- Observed counts are much larger than the independence model predicts in the corners where both responses are the most negative possible or the most positive possible.
- Cross-classifications of ordinal variables often exhibit their greatest deviations from independence in the corner cells.
- This pattern indicates lack of fit in the form of a positive trend.
- Subjects who feel more favorable to making birth control available to teenagers also tend to feel more tolerant about premarital sex.

Linear-by-Linear Association

- Assign scores u_i to the I rows and v_j to the J columns.
- We must have $u_1 \leq u_2 \leq \cdots \leq u_I$ and $v_1 \leq v_2 \leq \cdots \leq v_J$ to reflect the category ordering.
- The model is

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

- The independence model is the special case $\beta = 0$. The final term represents the deviation from independence.

Notes

- The deviation is linear in the Y scores at a fixed level of X and linear in the X scores at a fixed level of Y.
- In column j, the deviation is a linear function of X, having form (slope)$\times$(score for X), with slope βv_j
- This model is called the linear-by-linear association model.

Interpretations

- The parameter β refers to the direction and strength of association.
- When $\beta > 0$, there is a tendency for Y to increase as X increases.
- When $\beta < 0$, there is a tendency for Y to decrease as X increases.
- When the data display a positive or negative trend, this model usually fits much better than the independence model.

Describing Associations

- For the 2×2 table using the cells intersecting rows a and c with columns b and d, the model has odds ratio equal to

$$\frac{\mu_{ab}\mu_{cd}}{\mu_{ad}\mu_{cb}} = \exp[\beta(u_c - u_a)(v_d - v_b)]$$

- The association is stronger as $|\beta|$ increases.
- For given β pairs of categories that are farther apart have greater differences between their scores and odds ratios farther from 1.

Further Comments

- In practice, the most common choice of scores is $u_i = i$ and $v_j = j$, simply the row and column numbers.
- The odds ratios formed using adjacent rows and adjacent columns are called *local odds ratios*.
- For these unit spaced scores, the local odds ratios simplifies so that e^{β} is the common value of all the local odds ratios.
- Any set of equally-spaced row and column scores has the property of uniform local odds ratios.
- This special case of the model is called *uniform association*.

Example: SAS Codes (Inputting the Data)

```
data sex;
  input premar birth count @@;
  assoc = premar*birth;
datalines;
1 4 38   1 3 60   1 2 68   1 1 81
2 4 14   2 3 29   2 2 26   2 1 24
3 4 42   3 3 74   3 2 41   3 1 18
4 4 157 4 3 161   4 2 57   4 1 36
;
run;
```

SAS Codes: Fitting Independence Model

```
proc genmod data = sex order=data;
  class premar birth;
  model count = premar birth /dist = poi
  link = log ;
  output out = table7_3 pred=pred STDRESCHI = r;
run;
proc sort data=table7_3;
  by birth;
run;
```

SAS Codes: Printing Some Output

```
proc format;
  value premar 1='Always wrong' 2='Usually Wrong'
  3='Sometimes Wrong' 4='Never wrong';
  value birth 1='Strongly Disagree' 2='Disagree'
  3='Agree' 4='Strongly Agree';
run;
proc print data = table7_3 noobs;
  format premar premar.  birth birth.;
run;
```

Output

premar	birth	count	assoc	pred	r
Always wrong	Strongly Disagree	81	1	42.411	7.60318
Usually Wrong	Strongly Disagree	24	2	15.969	2.32832
Sometimes Wrong	Strongly Disagree	18	3	30.049	-2.68175
Never wrong	Strongly Disagree	36	4	70.571	-6.06333
Always wrong	Disagree	68	2	51.214	3.07671
Usually Wrong	Disagree	26	4	19.283	1.81148
Sometimes Wrong	Disagree	41	6	36.285	0.97623
Never wrong	Disagree	57	8	85.218	-4.60386
Always wrong	Agree	60	3	86.423	-4.11671
Usually Wrong	Agree	29	6	32.540	-0.81148
Sometimes Wrong	Agree	74	9	61.231	2.24730
Never wrong	Agree	161	12	143.806	2.38456
Always wrong	Strongly Agree	38	4	66.951	-4.83965
Usually Wrong	Strongly Agree	14	8	25.208	-2.75682
Sometimes Wrong	Strongly Agree	42	12	47.435	-1.02637
Never wrong	Strongly Agree	157	16	111.405	6.78455

SAS Codes: PROC CATMOD

```
proc catmod data=sex;
  weight count;
  model premar*birth= _response_/
    predict=freq noprofile noiter
    noresponse ;
  loglin premar birth;
run;
quit;
```

Partial Output

Maximum Likelihood Predicted Values for Frequencies

		-------Observed------		------Predicted------		
premar	birth	Frequency	Standard Error	Frequency	Standard Error	Residual
1	1	81	8.597365	42.41145	3.835381	38.58855
1	2	68	7.937662	51.21382	4.314471	16.78618
1	3	60	7.490815	86.42333	6.095867	-26.4233
1	4	38	6.036605	66.9514	5.130779	-28.9514
2	1	24	4.835077	15.96868	1.94804	8.031317
2	2	26	5.026925	19.28294	2.265306	6.717063
2	3	29	5.300169	32.53996	3.516572	-3.53996
2	4	14	3.713265	25.20842	2.827013	-11.2084
3	1	18	4.201203	30.0486	2.981299	-12.0486
3	2	41	6.259766	36.2851	3.396646	4.714903
3	3	74	8.251448	61.2311	4.989831	12.7689
3	4	42	6.332064	47.43521	4.118554	-5.43521
4	1	36	5.882213	70.57127	5.716958	-34.5713
4	2	57	7.313779	85.21814	6.309364	-28.2181
4	3	161	11.53289	143.8056	8.335372	17.19438
4	4	157	11.41846	111.405	7.268973	45.59503

SAS Codes: Fitting L by L Model

```
proc genmod data=sex order=data;
  format premar premar.  birth birth.;
  class premar birth;
  model count = premar birth assoc/
  dist=poi link=log ;
  output out=temp p=predict ;
run;
```

Output

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	8	11.5337	1.4417
Scaled Deviance	8	11.5337	1.4417
Pearson Chi-Square	8	11.5085	1.4386
Scaled Pearson X2	8	11.5085	1.4386
Log Likelihood		3041.7446	

Output

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square	Pr> ChiSq
Intercept		1	2.3532	0.2258	1.9106	2.7957	108.59	<.0001
premar	Always wrong	1	1.7537	0.2343	1.2944	2.2129	56.01	<.0001
premar	Usually Wrong	1	0.1077	0.1988	-0.2820	0.4974	0.29	0.5880
premar	Sometimes Wrong	1	-0.0163	0.1264	-0.2641	0.2314	0.02	0.8972
premar	Never wrong	0	0.0000	0.0000	0.0000	0.0000	-	-
birth	Strongly Agree	1	-1.8797	0.2491	-2.3679	-1.3914	56.94	<.0001
birth	Agree	1	-0.7245	0.1620	-1.0420	-0.4070	20.00	<.0001
birth	Disagree	1	-0.4641	0.1195	-0.6984	-0.2298	15.08	0.0001
birth	Strongly Disagree	0	0.0000	0.0000	0.0000	0.0000	-	-
assoc		1	0.2858	0.0282	0.2305	0.3412	102.46	<.0001
Scale		0	1.0000	0.0000	1.0000	1.0000		

SAS Codes: Printing Predicted Values

```
proc sort data = temp;
  by birth;
run;
proc print data = temp;
  format premar premar.  birth birth.;
  var premar birth count predict;
run;
```

Output

premar	birth	count	predict
Always wrong	Strongly Disagree	81	80.857
Usually Wrong	Strongly Disagree	24	20.750
Sometimes Wrong	Strongly Disagree	18	24.394
Never wrong	Strongly Disagree	36	33.000
Always wrong	Disagree	68	67.654
Usually Wrong	Disagree	26	23.107
Sometimes Wrong	Disagree	41	36.152
Never wrong	Disagree	57	65.088
Always wrong	Agree	60	69.396
Usually Wrong	Agree	29	31.543
Sometimes Wrong	Agree	74	65.681
Never wrong	Agree	161	157.379
Always wrong	Strongly Agree	38	29.094
Usually Wrong	Strongly Agree	14	17.600
Sometimes Wrong	Strongly Agree	42	48.773
Never wrong	Strongly Agree	157	155.533

Fit of The Linear Association Model

	Teenage Birth Control			
Premarital Sex	Strongly Disagree	Disagree	Agree	Strong Agree
Always wrong	81	68	60	38
	(80.91)	(67.6)	(69.4)	(29.1)
Almost always wrong	24	26	29	14
	(20.8)	(23.1)	(31.5)	(17.6)
Wrong only sometimes	18	41	74	42
	(24.4)	(36.1)	(65.7)	(48.8)
Not wrong at all	36	57	161	157
	(33)	(65.1)	(157.4)	(155.5)

Example: Continued

- ML estimate of $\beta = 0.286$, with $ASE = 0.028$.
- The positive estimate suggests that subjects having more favorable attitudes about availability of teen birth control also tend have more tolerant attitudes about premarital sex.
- Goodness-of-fit statistics are: $G^2 = 11.5$ and $\chi^2 = 11.5$ with $d.f. = 8$.
- P-value is 0.1749 suggesting the model to be a good fit.

Example: Continued

- The estimated local odds ratio is

$$\exp(\hat{\beta}) = \exp(0.286) = 1.33$$

- The strength of association seems weak.
- Non-local odds ratios are stronger.
- The estimated odds ratio for the four corner cells equals

$$\exp[\hat{\beta}(u_4 - u_1)(v_4 - v_1)] = \exp[0.286(4-1)(4-1)] = \exp(2.57) = 13.1$$

Notes

- Two sets of scores having the same spacing yield the same estimate of β and the same fit.
- Any other set of equally spaced scores yield the same fit but an appropriately rescaled estimate of β, so that the fitted odds ratios do not change.
- It is not necessary to use equally-spaced scores in the $L \times L$ model.

Ordinal Tests of Independence

- To test independence, $H_0 : \beta = 0$.
- The LR test equals the reduction in G^2 between the independence (I) model and $L \times L$ model.
- $G^2(I|L \times L) = G^2(I) - G^2(L \times L)$
- This statistic refers to a single parameter β, and is based on $d.f. = 1$.
- For the example, this statistic equals $127.6 - 11.5 = 116.1$ with P-value < 0.0001, which shows extremely strong association.
- The Wald's statistic provides an alternative to test this hypothesis.

Tests of Conditional Independence

- We have seen how to test a partial association by comparing two loglinear models that contain or omit that association.
- The LR test compares the models by the difference of the G^2 statistics, which is identical to the difference of deviances.
- An important application of this test refers to the null hypothesis of X-Y conditional independence.
- One compares the model (XZ, YZ) of $X - Y$ conditional independence to the more complex model (XY, XZ, YZ) of homogeneous association.
- The test statistic is $G^2[(XZ, YZ)|(XY, XZ, YZ)] = G^2(XZ, YZ) - G^2(XY, XZ, YZ)$.
- This test assumes that the homogeneous association model holds.

Continued $\cdots$

- For $2 \times 2 \times K$ tables, the test of conditional independence comparing two loglinear or logit models has the same purpose as the Cochran- Mantel-Haenszel (CMH) test.
- The CMH test works well when the $X - Y$ odds ratio is similar in each partial table.
- It is also naturally directed toward the alternative of homogeneous association.
- For large samples, the model based LR test usually gives similar results as the CMH test.

Direct Goodness-of-Fit Test

- A statistic of the form $G^2(XY, XZ)$ (*deviance*) does not require an assumption about homogeneous association.
- The statistic could be large if there are three-factor interactions, or if there is no three-factor interaction but conditional dependence.
- A disadvantage in using this test is that it often has low power.

Detecting Ordinal Conditional Association

- A useful model

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

  The model is called a *homogeneous linear − by − linear association* model.

- The conditional independence model (XZ, YZ) is the special case of this model with $\beta = 0$.

- Unless this models fits very poorly, the tests comparing this model are more powerful than tests that ignore the ordering.

Cochran-Mantel-Haenszel Test for $2 \times 2 \times K$ table

- To Test: X and Y are conditionally independent given Z.
- So, $H_0 : \theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)} = 1.0$.
- In the k-th partial table, the row totals are n_{1+k}, n_{2+k} and column totals are n_{+1k}, n_{+2k}.
- Given both these totals, n_{11k} has a hypergeometric distribution and that determines all other cell counts in the k-th partial table.

Cochran-Mantel-Haenszel Test (Continued)

- Under the null hypothesis of independence,

$$\mu_{11k} = E(n_{11k}) = \frac{n_{1+k}n_{+1k}}{n},$$

$$Var(n_{11k}) = \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{{n_{++k}}^2(n_{++k}-1)}, k = 1, \cdots, K$$

- The test statistic is given by

$$CMH = \frac{[\sum_{k=1}^{K}(n_{11k} - \mu_{11k})]^2}{\sum_{k=1}^{K} Var(n_{11k})}$$

- This is called the $Cochran - Mantel - Haenszel$ (CMH) statistic.
- It has a large sample chi-squared distribution with $df = 1$.

Generalized Cochran-Mantel-Haenszel Test

- We can generalize CMH test to $I \times J \times K$ tables.
- When X and Y are *ordinal*, the test statistic generalizes the correlation statistic for two-way tables.
- It is designed to detect a *linear trend* in the $X - Y$ association that has the same direction in each partial table.
- The generalized correlation statistic has approximately a chi-squared distribution with $df = 1$.

Generalized CMH Test

- The generalized correlation statistic is:

$$M^2 = \frac{[\sum_{k=1}^{K}\{\sum_{i=1}^{I}\sum_{j=1}^{J} u_i v_j n_{ijk} - E(\sum_{i=1}^{I}\sum_{j=1}^{J} u_i v_j n_{ijk})\}]^2}{\sum_{k=1}^{K} var(\sum_{i=1}^{I}\sum_{j=1}^{J} u_i v_j n_{ijk})}$$

$$E(\sum_{i=1}^{I}\sum_{j=1}^{J} u_i v_j n_{ijk}) = \frac{(\sum_{i=1}^{I} u_i n_{i+k})(\sum_{j=1}^{J} v_j n_{+jk})}{n_{++k}}$$

$$var(\sum_{i=1}^{I}\sum_{j=1}^{J}) = \frac{1}{(n_{++k}-1)}[\sum_{i=1}^{I} u_i^2 n_{i+k} - (\sum_{i=1}^{I} u_i n_{i+k})^2/n_{++k}]$$

$$\times [\sum_{j=1}^{J} v_j^2 n_{+jk} - (\sum_{j=1}^{J} v_j n_{+jk})^2/n_{++k}]$$

Generalized CMH Test

- X is nominal and Y is ordinal
- The test of conditional independence compares the I rows using a statistic based on the variation in those I averaged row mean responses that is designed to detect differences among their true values.
- It has a large sample chi-squared distribution with $df = I - 1$.
- The formula for this statistic is complex and for $K = 1$, it is

$$(n-1)\frac{\sum_{i=1}^{I} n_{i+}(\sum_{j=1}^{J} v_j n_{ij}/n_{i+} - \sum_{j=1}^{J} v_j n_{+j}/n)^2}{\sum_{j=1}^{J} v_j^2 n_{+j} - (\sum_{j=1}^{J} v_j n_{+j})^2/n}$$

Generalized CMH Test

- X and Y both are nominal.
- Another CMH-type statistic based on $df = (I-1)(J-1)$, provides a *general association test*.
- It is designed to detect any type of association that is similar in each partial tables.
- In SAS, the *general association* alternative treats both X and Y as nominal and has $df = (I-1)(J-1)$.
- The *row mean scores differ* alternative treats the rows of X as nominal and the columns of Y as ordinal and has $df = I-1$.
- The *nonzero correlation* alternative treats both X and Y as ordinal and has $df = 1$.