NEBRASKA ACADEMY FOR
METHODOLOGY, ANALYTICS & PSYCHOMETRICS

Introduction to Loglinear Models

Ann M. Arthur

Overview

- Considerations
- Parameters of contingency tables
- Loglinear model
  - Hypotheses to be tested
  - Interpretation of estimates
  - Model selection
- Useful parameterization for some categorical models

Considerations

- Categorical and discrete data
  - Poisson (count data)
  - Binomial (dichotomous data)
  - Multinomial (polytomous data)
- Research questions
  - All variables are categorical
  - Want to describe and understand associations between variables

Parameters

Categorical Data

- Frequencies or cell counts
- Compute probabilities
  - $p_i = f_i/n$
  - E.g., $p_{blue} = \frac{24}{93} = 0.258$

Color	Count	p
Blue	24	0.258
Brown	48	0.516
Green	15	0.161
Other	6	0.065
Total	93	1

Contingency Tables

- Assume two categorical variables, A with *i=1,…,I* categories and B with *j=1,…,J* categories
- Frequencies/cell counts can be arranged into an $I \times J$ contingency table

		B				
		1	2	…	J	
A	1	f_{11}	f_{12}	…	f_{1J}	f_{1+}
	2	f_{21}	f_{22}	…	f_{2J}	f_{2+}
	3	f_{31}	f_{32}	…	f_{3J}	f_{3+}
	⋮	…	…	…	…	f_{4+}
	I	f_{I1}	f_{I2}	…	f_{IJ}	f_{5+}
		f_{+1}	f_{+2}	f_{+3}	f_{+4}	f_{++} or n

2 x 2 Contingency Table

- Data from Sewell and Shah (1968) on 10,319 Wisconsin high school seniors
  - See also Fienberg (1977)
- Fundamental parameters
  - Probabilities
  - Odds
  - Odds Ratios

Plans to Attend College (Sewell & Shah, 1968)

		No	Yes	Total
Parental Encouragement	Low	4653	312	4965
	High	2290	3064	5354
	Total	6943	3376	10319

Probabilities

- Joint probabilities
  - Describe co-occurrence
  - $p_{ij} = \frac{f_{ij}}{n}$
  - $p_{Low,No} = \frac{4653}{10319} = 0.45$

Plans to Attend College (Sewell & Shah, 1968)

		No	Yes	Total
Parental Encouragement	Low	4653	312	4965
	High	2290	3064	5354
	Total	6943	3376	10319

Probabilities

- Joint probabilities
  - Describe co-occurrence
  - $p_{ij} = \frac{f_{ij}}{n}$
  - $p_{Low,No} = \frac{4653}{10319} = 0.45$

Plans to Attend College (Sewell & Shah, 1968)

		No	Yes	Total
Parental Encouragement	Low	4653	312	4965
	High	2290	3064	5354
	Total	6943	3376	10319

Probabilities (2)

- Joint probabilities
  - Describe co-occurrence
  - $p_{ij} = \frac{f_{ij}}{n}$
  - $p_{Low,No} = \frac{4653}{10319} = 0.45$

Plans to Attend College (Sewell & Shah, 1968)

		No	Yes	Total
Parental Encouragement	Low	4653	312	4965
		0.45	*0.03*	
	High	2290	3064	5354
		0.22	*0.30*	
	Total	6943	3376	10319

Probabilities (3)

- Marginal probabilities
  - $p_{i+} = \frac{f_{i+}}{n}$, also $\frac{f_{+j}}{n}$
  - $p_{Low} = \frac{4965}{10319} = 0.48$

Plans to Attend College (Sewell & Shah, 1968)

		No	Yes	Total
Parental Encouragement	Low	4653	312	4965
	High	2290	3064	5354
	Total	6943	3376	10319

Probabilities (3)

- Marginal probabilities
  - $p_{i+} = \frac{f_{i+}}{n}$, also $\frac{f_{+j}}{n}$
  - $p_{Low} = \frac{4965}{10319} = 0.48$

Plans to Attend College (Sewell & Shah, 1968)

		No	Yes	Total
Parental Encouragement	Low	4653	312	4965
	High	2290	3064	5354
	Total	6943	3376	10319

Probabilities (4)

- Marginal probabilities
  - $p_{i+} = \frac{f_{i+}}{n}$, also $\frac{f_{+j}}{n}$
  - $p_{Low} = \frac{4965}{10319} = 0.48$

Plans to Attend College (Sewell & Shah, 1968)

		No	Yes	Total
Parental Encouragement	Low	4653	312	4965
				0.48
	High	2290	3064	5354
				0.52
	Total	6943	3376	10319
		0.67	*0.33*	

Probabilities (5)

- Conditional probabilities
  - Implies causal structure (DV|IV)
  - $p_{j|i} = \frac{p_{ij}}{p_{i+}}$
  - E.g., What is the probability that they are not planning to attend college, given low parental encouragement?
    - $p_{No|Low} = \frac{0.45}{0.48} = 0.94$

Plans to Attend College (Sewell & Shah, 1968)

		No	Yes	Total
Parental Encouragement	Low	4653	312	4965
	High	2290	3064	5354
	Total	6943	3376	10319

Probabilities (5)

- Conditional probabilities
  - Implies causal structure (DV|IV)
  - $p_{j|i} = \frac{p_{ij}}{p_{i+}}$
  - E.g., Given low parental encouragement, what is the probability that they do not plan to attend college?
    - $p_{No|Low} = \frac{0.45}{0.48} = 0.94$

Plans to Attend College (Sewell & Shah, 1968)

		No	Yes	Total
Parental Encouragement	Low	4653 *0.45* *0.94*	312	4965 *0.48*
	High	2290 *0.22*	3064 *0.30*	5354 *0.52*
	Total	6943 *0.67*	3376 *0.33*	10319

Odds

- $odds = p/(1-p)$
  - E.g., Odds of seniors not planning to attend college relative to those planning to attend
    - $\Omega_{1+} = p_{1+}/p_{2+} = \frac{6943}{3376} = 2.05$
    - Seniors are twice as likely to not plan to attend college, compared to those planning to attend

Plans to Attend College (Sewell & Shah, 1968)

		No	Yes	Total
Parental Encouragement	Low	4653	312	4965
	High	2290	3064	5354
	Total	6943	3376	10319

Odds

- $odds = p/(1-p)$
  - E.g., Odds of seniors not planning to attend college relative to those planning to attend
    - $\Omega_{1+} = p_{1+}/p_{2+} = \frac{6943}{3376} = 2.05$
    - Seniors are twice as likely to not plan to attend college, compared to those planning to attend
  - E.g., Odds of seniors planning to attend college compared to those planning to not attend
    - $\Omega_{2+} = \frac{3376}{6943} = 0.48$
    - Seniors are half as likely to plan to attend college, compared to those that are not planning to attend

Plans to Attend College (Sewell & Shah, 1968)

		No	Yes	Total
Parental Encouragement	Low	4653	312	4965
	High	2290	3064	5354
	Total	6943	3376	10319

Odds Ratio

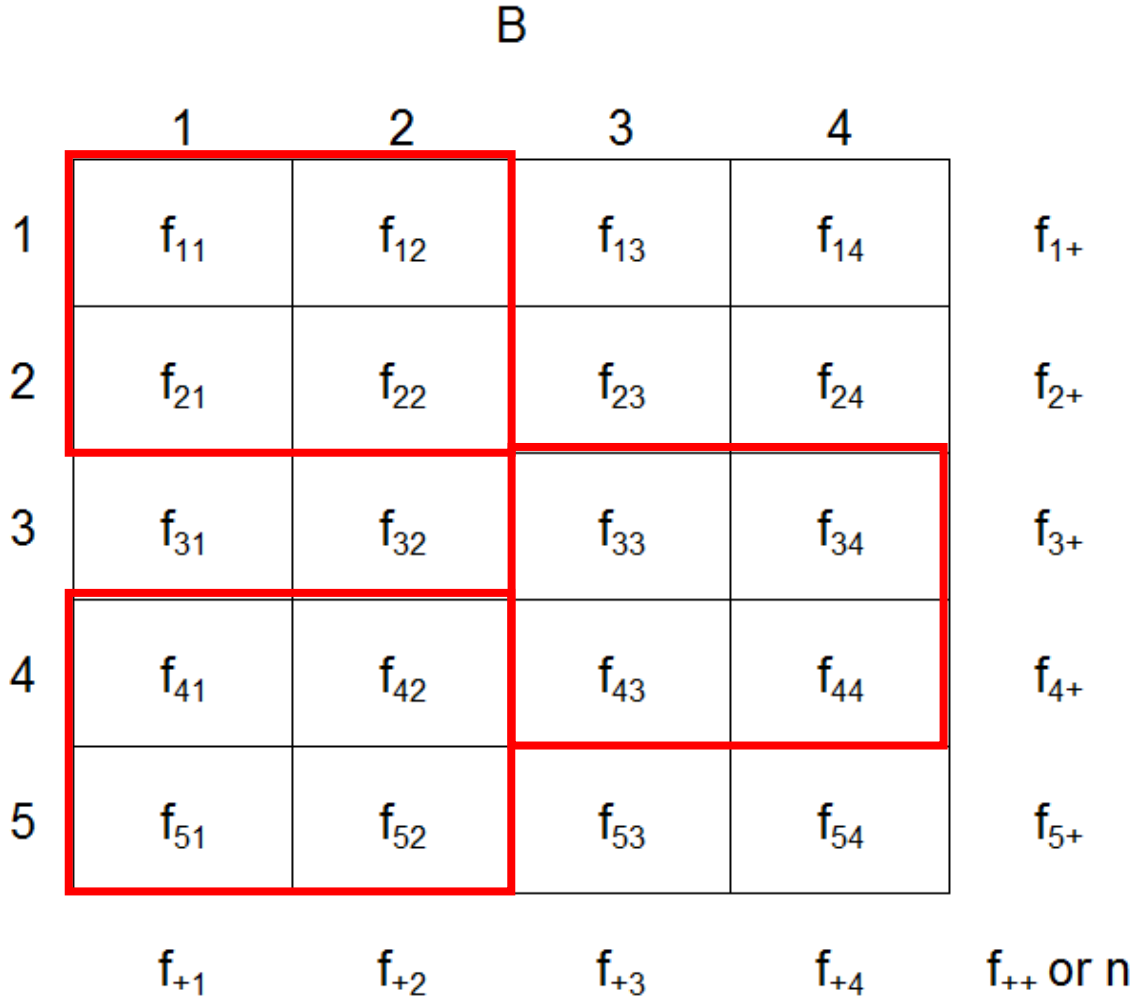
- Odds ratios
  - Compares two odds
    - $\theta = \frac{odds_1}{odds_2} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$
  - Ratio of cross-products
    - $\theta_{11} = \frac{(p_{11}/p_{12})}{(p_{21}/p_{22})} = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{f_{11}f_{22}}{f_{12}f_{21}}$
    - $\theta_{11} = \frac{(4653/312)}{(2290/3064)} = \frac{4653*3064}{2290*312} = 19.95$
      - Students with low parental encouragement have estimated odds of planning to not attend college that are 20 times the estimated odds of someone with high encouragement

Plans to Attend College (Sewell & Shah, 1968)

		No	Yes	Total
Parental Encouragement	Low	4653	312	4965
	High	2290	3064	5354
	Total	6943	3376	10319

Odds Ratio (2)

- Local odds ratios
  - Adjacent cells
- Local odds ratios perfectly define all associations within the table!



Independence

- No association between two variables
  - Equal odds ratios
  - Joint probability is a product of the marginals
    - Not significantly different from expected values
- Can you collapse the table across a dimension?
  - To do this, the variable must have no significant interaction with the other variable

Multiway Tables

- Simpson's Paradox
  - When marginal tables leads to highly misleading inference
  - Specification problem
    - Correct functional form
    - All necessary variables
    - No unnecessary variables

Plans to Attend College (Sewell & Shah, 1968)

		No	Yes
Male	Low	1949	136
	High	1203	1703
Female	Low	2704	176
	High	1087	1361

Independence (2)

- Tests of independence
  - Pearson chi-square statistic, X^2
  - Likelihood ratio chi-square statistic, G^2
- Degrees of freedom
  - (I x J) - # of estimated parameters

Plans to Attend College (Sewell & Shah, 1968)

		No	Yes	Total
Parental Encouragement	Low	4653	312	4965
	High	2290	3064	5354
	Total	6943	3376	10319

Independence (2)

```
data college;
     input encouragement$ attend $ count @@;
     datalines;
     low no 4653   low yes 312
     high no 2290  high yes 3064
     ;
proc freq data=college order=data;
     weight count;
     tables encouragement*attend/chisq expected;
run;
```

Test of Independence (3)

Statistics for Table of encourage by college

Statistic	DF	Value	Prob
Chi-Square	1	3037.2184	<.0001
Likelihood Ratio Chi-Square	1	3405.7306	<.0001
Continuity Adj. Chi-Square	1	3034.9045	<.0001
Mantel-Haenszel Chi-Square	1	3036.9241	<.0001
Phi Coefficient		0.5425	
Contingency Coefficient		0.4769	
Cramer's V		0.5425	

- H_0: Independence
  - How plausible is it that the local odds ratio is 1?
- Larger values are the result of greater differences between expected and observed values
- Reject H_0

Loglinear Models

The Loglinear Model

- A type of generalized linear models (GLM), the family of models that extend ordinary least squares regression to non-normal distributions
- Models to describe the joint distributions
  - The dependent variable is a cell size (no distinction between dependent and independent variables)
  - Used to analyze cell counts in a more formal and complete manner (Hagenaars, 1993)
- The canonical link is the log link

Fundamental Parameters

- Odds and odds ratios
  - Range of $[0, \infty]$
  - Not symmetrical around 0
  - Value of 1 indicates equal odds and independence
- Logits
  - $Log(\theta)$
  - Range of $[-\infty, \infty]$
  - Symmetric around 0
  - Value of 0 indicates equal odds and no difference

Multiway Tables

- Higher-order odds ratio
  - $\theta_{111} = \frac{f_{111}f_{221}}{f_{121}f_{212}} / \frac{f_{112}f_{222}}{f_{122}f_{212}}$

- Partial odds ratio
  - Average conditional odds
    - Can't add and divide odds ratios
    - Geometric mean (multiply and take nth root)
  - $\theta_{11p} = \left(\prod_k^K \theta_{11k}\right)^{1/K} = \sqrt[K]{\theta_{111}\theta_{112} \dots \theta_{11K}}$

The Independence Model

- In probability form
  - Joint probabilities can determined by the marginals
  - $p_{ij} = p_{i+}p_{+j}$
- In expected frequency form
  - $\mu_{ij} = np_{i+}p_{+j}$
  - This form is multiplicative
- Take the natural log of expected frequencies
  - Yields the loglinear model
  - Additive

Multiplicative and Additive Models

- Taking the natural log yields the loglinear model of independence
  - $\mu_{ij} = np_{i+}p_{+j}$ (Multiplicative, expected frequencies)
  - $\log(\mu_{ij}) = \log(n) + \log(p_{i+}) + \log(p_{+j})$ (Additive)
  - $\log(\mu_{ij}) = \mu + \lambda_i^A + \lambda_j^B$ (Additive, loglinear notation)
    - Where A and B denote parental encouragement and college plans, respectively

Analogous to ANOVA

- $\log(\mu_{ij}) = \mu + \lambda_i^A + \lambda_j^B$

  μ is the average cell size, or "grand mean"

  λ_i^A is the row effect for variable A, or deviation from the average cell size due to level i

  λ_j^B is the column effect of variable B, or deviation from the average cell size due to level j
- The equation for the expected values
- Sparseness
  - Can't take the natural log of 0
  - If there are cell counts of 0, they need to be adjusted
    - Create a new count variable with a very small amount added (e.g., 0.0001)

Loglinear Model of Independence

- The loglinear model of independence for three variables is:

  $\log(f_{ijk}^{ABC}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C$

- This model omits all higher-order terms
  - Assumes there are no interactions between variables (e.g., $\lambda_{ij}^{AB} = 0$)

Estimation

- SAS
  - PROC GENMOD
  - PROC CATMOD
- Lem
- R
  - loglin()
  - glm()

Example – Independence Loglinear Model

```
data college;
  input sex $ encouragement $ attend $ count;
datalines;
male      low   no   1949
male      low   yes   136
male      high  no   1203
male      high  yes  1703
female    low   no   2704
female    low   yes   176
female    high  no   1087
female    high  yes  1361
;
```

Example – Independence Loglinear Model

```
proc genmod data=college;
     class sex encouragement attend;
     model count =  sex encouragement attend /
          dist=poi link=log lrci type3 obstats;
run;
```

Output – Independence Loglinear Model

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	4	3567.0473	891.7618
Scaled Deviance	4	3567.0473	891.7618
Pearson Chi-Square	4	3268.5410	817.1353
Scaled Pearson X2	4	3268.5410	817.1353
Log Likelihood		64231.0306	
Full Log Likelihood		-1818.0268	
AIC (smaller is better)		3644.0537	
AICC (smaller is better)		3657.3870	
BIC (smaller is better)		3644.3714	

- H_0: Independence holds
- Overall fit
  - Large X^2 and G^2

Output – Independence Loglinear Model

	Analysis Of Maximum Likelihood Parameter Estimates							
	Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
μ	Intercept		1	6.6665	0.0225	6.6223	6.7103	88151.3	<.0001
λ_1^A	sex	female	1	0.0653	0.0197	0.0267	0.1040	11.00	0.0009
λ_2^A	sex	male	0	0.0000	0.0000	0.0000	0.0000	.	.
λ_1^B	encouragement	high	1	0.0754	0.0197	0.0368	0.1141	14.66	0.0001
λ_2^B	encouragement	low	0	0.0000	0.0000	0.0000	0.0000	.	.
λ_1^C	attend	no	1	0.7210	0.0210	0.6800	0.7623	1180.96	<.0001
λ_2^C	attend	yes	0	0.0000	0.0000	0.0000	0.0000	.	.
	Scale		0	1.0000	0.0000	1.0000	1.0000		

- Convert estimates back into cell counts (dummy-coding approach)
  - Males with low encouragement not planning to attend college
  - $\mu_{111} = \exp(\lambda + \lambda_2^A + \lambda_2^B + \lambda_1^C) = \exp(6.67 + 0 + 0 + 0.72) = 1615.662$

Output – Independence Loglinear Model

count	sex	encouragement	attend	Predicted Value	Linear Predictor	Error of the Linear Predictor	HessWgt	Lower	Upper	Raw Residual	Pearson Residual
1949	male	low	no	1615.7671	7.3875651	0.0187612	1615.7671	1557.4323	1676.2869	333.23289	8.2900754
136	male	low	yes	785.6589	6.6665227	0.0224535	785.6589	751.83327	821.00638	-649.6589	-23.1776
1203	male	high	no	1742.3599	7.4629958	0.0183671	1742.3599	1680.7526	1806.2255	-539.3599	-12.92141
1703	male	high	yes	847.21405	6.7419534	0.0221253	847.21405	811.26002	884.76152	855.78595	29.401438
2704	female	low	no	1724.8662	7.4529048	0.0184204	1724.8662	1663.7036	1788.2773	979.13381	23.575686
176	female	low	yes	838.7078	6.7318624	0.0221695	838.7078	803.04509	875.95427	-662.7078	-22.8832
1087	female	high	no	1860.0068	7.5283354	0.0180188	1860.0068	1795.4648	1926.8688	-773.0068	-17.92363
1361	female	high	yes	904.41925	6.807293	0.021837	904.41925	866.52699	943.96849	456.58075	15.18213

The Saturated Loglinear Model

- Models all possible associations between cell counts

$$\log(f_{ijk}^{ABC}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

μ is the average cell size
λ_i^A, λ_j^B, and λ_k^C are the main effects of variables *A*, *B*, and *C*
λ_{ij}^{AB}, λ_{ik}^{AC}, λ_{jk}^{BC}, and λ_{ijk}^{ABC} are higher-order terms

The Saturated Loglinear Model

- “Saturated” means the number of cells is equal to the number of parameters estimated
  - Just-identified model
- This model is often not of interest
  - No degrees of freedom available to test hypotheses
  - Does not simplify interpretation of the data

Example – Saturated Loglinear Model

```
proc genmod data=college;
     class sex encouragement attend;
     model count = sex|encouragement|attend /
          dist=poi link=log lrci type3 obstats;
run;
```

Example – Saturated Loglinear Model

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	0	0.0000	.
Scaled Deviance	0	0.0000	.
Pearson Chi-Square	.	0.0000	.
Scaled Pearson X2	.	0.0000	.
Log Likelihood		66014.5543	
Full Log Likelihood		-34.5032	
AIC (smaller is better)		85.0064	
AICC (smaller is better)		.	
BIC (smaller is better)		85.6420	

- No degrees of freedom to test model fit

Example – Saturated Loglinear Model

Analysis Of Maximum Likelihood Parameter Estimates										
Parameter				DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept				1	4.9127	0.0857	4.7397	5.0761	3282.25	<.0001
sex	female			1	0.2578	0.1142	0.0349	0.4829	5.10	0.0239
sex	male			0	0.0000	0.0000	0.0000	0.0000	.	.
encouragement	high			1	2.5275	0.0891	2.3570	2.7066	804.55	<.0001
encouragement	low			0	0.0000	0.0000	0.0000	0.0000	.	.
sex*encouragement	female	high		1	-0.4820	0.1198	-0.7180	-0.2479	16.18	<.0001
sex*encouragement	female	low		0	0.0000	0.0000	0.0000	0.0000	.	.
sex*encouragement	male	high		0	0.0000	0.0000	0.0000	0.0000	.	.
sex*encouragement	male	low		0	0.0000	0.0000	0.0000	0.0000	.	.
attend	no			1	2.6624	0.0887	2.4928	2.8408	901.15	<.0001
attend	yes			0	0.0000	0.0000	0.0000	0.0000	.	.
sex*attend	female	no		1	0.0696	0.1180	-0.1628	0.3000	0.35	0.5553
sex*attend	female	yes		0	0.0000	0.0000	0.0000	0.0000	.	.
sex*attend	male	no		0	0.0000	0.0000	0.0000	0.0000	.	.
sex*attend	male	yes		0	0.0000	0.0000	0.0000	0.0000	.	.
encouragement*attend	high	no		1	-3.0100	0.0964	-3.2027	-2.8247	975.83	<.0001
encouragement*attend	high	yes		0	0.0000	0.0000	0.0000	0.0000	.	.
encouragement*attend	low	no		0	0.0000	0.0000	0.0000	0.0000	.	.
encouragement*attend	low	yes		0	0.0000	0.0000	0.0000	0.0000	.	.
sex*encourage*attend	female	high	no	1	0.0532	0.1303	-0.2016	0.3096	0.17	0.6832
sex*encourage*attend	female	high	yes	0	0.0000	0.0000	0.0000	0.0000	.	.
sex*encourage*attend	female	low	no	0	0.0000	0.0000	0.0000	0.0000	.	.
sex*encourage*attend	female	low	yes	0	0.0000	0.0000	0.0000	0.0000	.	.

- All parameters estimated
- Some non-significant interactions

Example – Saturated Loglinear Model

count	sex	encouragement	attend	Predicted Value	Linear Predictor	Standard Error of the Linear Predictor	HessWgt	Lower	Upper
1949	male	low	no	1949	7.5750717	0.0226513	1949	1864.3651	2037.4769
136	male	low	yes	136	4.9126549	0.0857493	136	114.96059	160.88992
1203	male	high	no	1203	7.0925737	0.0288315	1203	1136.9051	1272.9374
1703	male	high	yes	1703	7.4401467	0.0242322	1703	1624.008	1785.8342
2704	female	low	no	2704	7.9024874	0.0192308	2704	2603.9787	2807.8632
176	female	low	yes	176	5.170484	0.0753778	176	151.82767	204.02078
1087	female	high	no	1087	6.9911769	0.0303309	1087	1024.2638	1153.5788
1361	female	high	yes	1361	7.215975	0.0271063	1361	1290.5807	1435.2617

- Predicted values perfectly represent the observed data

Reduced Loglinear Models

- Do you need higher-order terms, or can they be eliminated?
  - Reduced models with good fit greatly simplifies the interpretation
  - Parsimony
- Possible models
  - Model of all possible associations
  - Models with main effects and two-ways interactions
  - Models with main effects only
  - Model of independence

Reduced Loglinear Models

- A three-variable model that permits some two-way associations

$$\log(f_{ijk}^{ABC}) = \mu + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{jk}^{BC}$$

- Two factor terms describe conditional odds ratios
  - λ_{ij}^{AB} association between *A* and *B,* controlling for *C*
  - λ_{jk}^{BC} association between *B* and *C*, controlling for *A*
- This model is referred to as (AB, BC)

Reduced Loglinear Models

- Test whether there is conditional independence within the multiway table
- Compare the fit of various reduced loglinear models to the saturated model

Plans to Attend College (Sewell & Shah, 1968)

		No	Yes
Male	Low	1949	136
	High	1203	1703
Female	Low	2704	176
	High	1087	1361

Example – Reduced Loglinear Models

Sex	Encouragement	Attend	(S, E, A)	(SE, A)	(SA, EA)	(SE, SA, EA)	(SEA)
Male	Low	No	1615.8	1402.9	2112.4	1945.9	1949
		Yes	785.7	682.1	170.0	139.1	136
	High	No	1742.4	1955.3	1039.6	1206.1	1203
		Yes	847.2	950.7	1669.0	1699.9	1703
Female	Low	No	1724.9	1937.8	2540.6	2707.1	2704
		Yes	838.7	942.2	142.0	172.9	176
	High	No	1860.0	1647.1	1250.4	1083.9	1087
		Yes	904.4	800.9	1395.0	1364.1	1361

Example – Reduced Loglinear Models

Sex	Encouragement	Attend	(S, E, A)	(SE, A)	(SA, EA)	(SE, SA, EA)	(SEA)
Male	Low	No	1615.8	1402.9	2112.4	1945.9	1949
		Yes	785.7	682.1	170.0	139.1	136
	High	No	1742.4	1955.3	1039.6	1206.1	1203
		Yes	847.2	950.7	1669.0	1699.9	1703
Female	Low	No	1724.9	1937.8	2540.6	2707.1	2704
		Yes	838.7	942.2	142.0	172.9	176
	High	No	1860.0	1647.1	1250.4	1083.9	1087
		Yes	904.4	800.9	1395.0	1364.1	1361

Example – Reduced Loglinear Models

Sex	Encouragement	Attend	(S, E, A)	(SE, A)	(SA, EA)	(SE, SA, EA)	(SEA)
Male	Low	No	1615.8	1402.9	2112.4	1945.9	1949
		Yes	785.7	682.1	170.0	139.1	136
	High	No	1742.4	1955.3	1039.6	1206.1	1203
		Yes	847.2	950.7	1669.0	1699.9	1703
Female	Low	No	1724.9	1937.8	2540.6	2707.1	2704
		Yes	838.7	942.2	142.0	172.9	176
	High	No	1860.0	1647.1	1250.4	1083.9	1087
		Yes	904.4	800.9	1395.0	1364.1	1361

Example – Reduced Loglinear Models

Sex	Encouragement	Attend	(S, E, A)	(SE, A)	(SA, EA)	(SE, SA, EA)	(SEA)
Male	Low	No	1615.8	1402.9	2112.4	1945.9	1949
		Yes	785.7	682.1	170.0	139.1	136
	High	No	1742.4	1955.3	1039.6	1206.1	1203
		Yes	847.2	950.7	1669.0	1699.9	1703
Female	Low	No	1724.9	1937.8	2540.6	2707.1	2704
		Yes	838.7	942.2	142.0	172.9	176
	High	No	1860.0	1647.1	1250.4	1083.9	1087
		Yes	904.4	800.9	1395.0	1364.1	1361

- Cell counts – how do they compared to the saturated model?
- Model (SE, SA, EA) comes the closest to the observed data

Model Selection

Model	X^2	G^2	DF	*p*
(S, E, A)	3268.54	3567.05	4	<.0001
(S, EA)	160.851	161.32	3	<.0001
(SA, E)	3104.86	3492.11	3	<.0001
(SE, A)	3042.77	3410.98	3	<.0001
(SE, SA)	2993.7	3336.04	2	<.0001
(SE, EA)	5.2546	5.25	2	0.0724
(SA, EA)	86.5847	86.38	2	<.0001
(SE, SA, EA)	0.1665	0.17	1	0.683
(SEA)	0	0	0	0

Note. P-values are for G^2 statistic.

Loglinear Models

- Hypotheses to be tested
  - Independence
  - Reduced models
- Interpretation
  - For dummy-coding approach, ANOVA-style decomposition
  - Convert estimates into expected cell counts
    - Males with low encouragement not planning to attend college
    - $\mu_{111} = \exp\left(\lambda + \lambda_2^A + \lambda_2^B + \lambda_1^C\right) = \exp(6.67 + 0 + 0 + 0.72) = 1615.662$
- Model selection
  - Retain the model that fits well and represents the observed data well

Loglinear Parameterization of Common Categorical Models

The Logistic Model

- Special case of the generalized linear model
  - Regresses a binary dependent variable on 1+ independent variables
  - Models the log of the odds of the dependent variable
  - The canonical link function is the logit
  - Does not describe relationships among independent variables
- When one variable is binary, the logistic models for that response are equal to certain loglinear models
  - Construct logits for one variable to help interpret loglinear models (Bishop, 1969)

Using Logistic Models to Interpret

- Two types of coding yield identical estimates
  - Dummy coding (0, 1)
  - Effect coding (-1, 1)
- Identifying constraints and power rules
  - $\sum \lambda_i^A = 0$
  - Lambdas sum to zero in effect coding approach
  - When you change an odd number, change the sign
    - $\lambda_1^A = -\lambda_2^A$
  - When you change an even number, same sign
    - $\lambda_{11}^{AB} = -\lambda_{12}^{AB} = -\lambda_{21}^{AB} = \lambda_{22}^{AB}$

Using Logistic Models to Interpret

- Odds ratios relate to two-factor loglinear parameters and main effects
  - The log odds ratio for the effect of *A* on *C*

Logit	Loglinear
$\beta_1^A - \beta_2^A$	$\lambda_{11}^{AC} + \lambda_{22}^{AC} - \lambda_{12}^{AC} - \lambda_{21}^{AC}$

1) Specify a logit for one variable
2) Substitute the loglinear parameterization for the odds
3) Use power rules to substitute and solve

Using Logistic Models to Interpret

1) Form a logit for the loglinear model
   - $\log(\mu_{ijk}) = \mu + \lambda_i^S + \lambda_j^E + \lambda_k^A + \lambda_{ij}^{SE} + \lambda_{ik}^{SA} + \lambda_{jk}^{EA}$
   - Suppose A is the dependent variable and E and A are explanatory variables
   - $logit[P(A=1) = log\left[\frac{P(A=1)}{1-P(A=1}\right] = \log\left[\frac{P(A=1|S=i,E=j)}{P(A=2|S=i,E=j)}\right]$

$$= log\left(\frac{f_{ij1}}{f_{ij2}}\right) = \log(f_{ij1}) - \log(f_{ij2})$$

Using Logistic Models to Interpret

1) Form a logit for the loglinear model
   - $\log(\mu_{ijk}) = \mu + \lambda_i^S + \lambda_j^E + \lambda_k^A + \lambda_{ij}^{SE} + \lambda_{ik}^{SA} + \lambda_{jk}^{EA}$
   - Suppose A is the dependent variable and S and E are explanatory variables
   - $logit[P(A=1) = log\left[\frac{P(A=1)}{1-P(A=1}\right] = \log\left[\frac{P(A=1|S=i,E=j)}{P(A=2|S=i,E=j)}\right]$

$$= log\left(\frac{f_{ij1}}{f_{ij2}}\right) = \log(f_{ij1}) - \log(f_{ij2})$$

2) Substitute the loglinear parameterization for the odds

$$= \left(\mu + \lambda_i^S + \lambda_j^E + \lambda_1^A + \lambda_{ij}^{SE} + \lambda_{i1}^{SA} + \lambda_{j1}^{EA}\right)$$
$$-(\mu + \lambda_i^S + \lambda_j^E + \lambda_2^A + \lambda_{ij}^{SE} + \lambda_{i2}^{SA} + \lambda_{j2}^{EA})$$

Using Logistic Models to Interpret

1) Form a logit for the loglinear model
   - $\log(\mu_{ijk}) = \mu + \lambda_i^S + \lambda_j^E + \lambda_k^A + \lambda_{ij}^{SE} + \lambda_{ik}^{SA} + \lambda_{jk}^{EA}$
   - Suppose A is the dependent variable and E and A are explanatory variables
   - $logit[P(A=1) = log\left[\frac{P(A=1)}{1-P(A=1}\right] = \log\left[\frac{P(A=1|S=i,E=j)}{P(A=2|S=i,E=j)}\right]$

$$= log\left(\frac{f_{ij1}}{f_{ij2}}\right) = \log(f_{ij1}) - \log(f_{ij2})$$

2) Substitute the loglinear parameterization for the odds

$$= \left(\mu + \lambda_i^S + \lambda_j^E + \lambda_1^A + \lambda_{ij}^{SE} + \lambda_{i1}^{SA} + \lambda_{j1}^{EA}\right)$$
$$-(\mu + \lambda_i^S + \lambda_j^E + \lambda_2^A + \lambda_{ij}^{SE} + \lambda_{i2}^{SA} + \lambda_{j2}^{EA})$$
$$= (\lambda_1^A - \lambda_2^A) + (\lambda_{i1}^{SA} - \lambda_{i2}^{SA}) + (\lambda_{j1}^{EA} - \lambda_{j2}^{EA})$$

Using Logistic Models to Interpret

$$= (\lambda_1^A - \lambda_2^A) + (\lambda_{i1}^{SA} - \lambda_{i2}^{SA}) + (\lambda_{j1}^{EA} - \lambda_{j2}^{EA})$$

Using Logistic Models to Interpret

$$= (\lambda_1^A - \lambda_2^A) + (\lambda_{i1}^{SA} - \lambda_{i2}^{SA}) + (\lambda_{j1}^{EA} - \lambda_{j2}^{EA})$$

3) Use power rules to substitute again

$$= 2\lambda_1^A + 2\lambda_{i1}^{SA} + 2\lambda_{j1}^{EA}$$

Loglinear parameters have corresponding logit parameters

$$logit[P(A = 1)] = \alpha + \beta_1^S + \beta_1^E$$

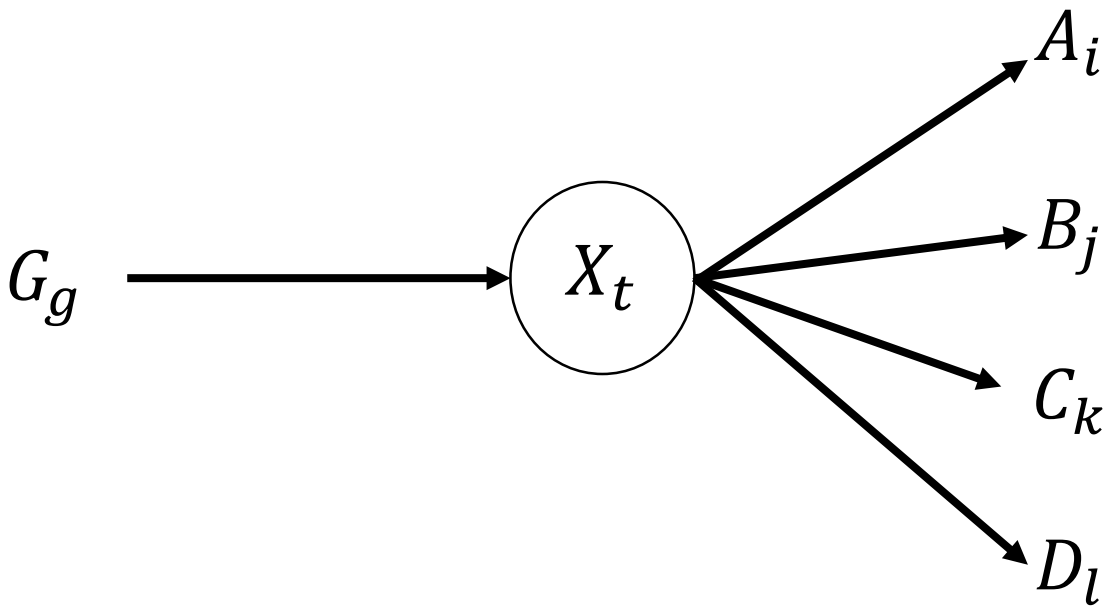
The Logistic Model

- The equivalent parameterizations enhance interpretation
- Historical breakthrough
  - Logistic models could be solved using iterative proportional fitting, which was previously used to solve loglinear models (Goodman, 1964; 1968)

The Latent Class Model

- Latent class analysis (LCA) is a special case of discrete (finite) mixture models (McLachlan & Peel, 2000; Newcomb, 2000)
  - Used to identify unobserved or latent groups
  - Assumes conditional independence
    - Controlling for the latent variable, all manifest variables are independent
- Two equivalent parameterizations (Goodman, 1974; Haberman, 1979)
  - Probabilistic
  - Loglinear

The Latent Class Model (3)



LCA Parameterizations

- The basic LCA model, assuming 3 manifest variables and 1 latent variable
- Probabilistic parameterization

$$\pi_{ijkt}^{ABCX} = \pi_t^X \pi_{it}^{A|X} \pi_{jt}^{B|X} \pi_{kt}^{C|X}$$

  - π_t^X is the latent class probability or “mixing” probability that a given member of the sample is in latent class *t*
  - $\pi_{it}^{A|X}$, $\pi_{jt}^{B|X}$ and $\pi_{kt}^{C|X}$ are conditional probabilities that the respondent in latent class *t* responds with 0 or 1 for each manifest indicator variable

LCA Parameterizations

- To obtain the loglinear form of the model, take the natural log of the probabilistic model

$$\ln\left(f_{ijkt}^{ABCX}\right) = \lambda + \lambda_t^X + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{it}^{AX} + \lambda_{jt}^{BX} + \lambda_{kt}^{CX}$$

- Includes only the higher-order terms that include the latent variable
- No interaction terms (e.g., $\lambda_{ijt}^{ABX} = 0$) because the model specifies conditional independence (McCutcheon, 1987; 2002)

LCA Parameterizations

- The two parameterizations are equivalent (Haberman, 1979)
  - Same number of parameters
  - Same expected values
- Some restrictions can only be imposed in one parameterization
  - "Reduced" latent class models

    $\pi_{ijkls}^{ABCDG} = \pi_{ts}^{X|G} \pi_{its}^{A|XG} \pi_{jts}^{B|XG} \pi_{kts}^{C|XG} \pi_{lts}^{D|XG} \pi_{s}^{G}$
  - Test hypotheses using loglinear parameterization
    - Use power rules to obtain "reduced" model in loglinear form
    - Does conditional independence hold across groups?

Summary

- Loglinear models are an essential method for understanding categorical data
  - Taking the natural log of cell counts yields an ANOVA decomposition
    - Log odds of cell sizes
    - Convert lambda parameters back into odds and odds ratios
- The two parameterizations permit ANOVA-style decomposition to contingency tables
  - Aid interpretability
    - Can estimate equivalent logistic models
  - Added flexibility in the types of restrictions that can be imposed
    - Conditional independence in latent class analysis models

References

Agresti, A. (2007). An introduction to categorical data analysis. (2nd ed.). Hoboken, NJ: Wiley.

Bishop, Y. M. (1969). Full contingency tables, logits, and split contingency tables. *Biometrics*, 383-399.

Fienberg, S. (1977) The analysis of cross-classified categorical data. Cambridge, MA: MIT Press

Goodman, L. A. (1964). A short computer program for the analysis of transaction flows. *Behavioral Science, 9,* 176-186.

Goodman, L. A. (1968). The analysis of cross-classified data: Independence, quasi-independence, and interactions in contingency tables with or without missing entries. *Journal of the American Statistical Associate, 63*(324), 1091-1131.

Hagenaars, J. (1993). *Loglinear models with latent variables (*Sage University Paper series on quantitative applications in the social sciences, series no. 07-094). Newbury Park, CA: Sage.

McCutcheon, A. L. (1987). *Latent class analysis*. Newbury Park, CA: Sage.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Model.* Hoboken, NJ: John Wiley & Sons. doi: 10.1002/0471721182

Newcomb, S. (1886). A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics, 8*(4), 343-366.

Sewell, W. H., & Shah, V. P. (1968). Social class, parental encouragement, and educational aspirations. *American Journal of Sociology*, *73*, 559-572.

Questions?

Ann.Arthur@huskers.unl.edu