Lecture 16
PubH 7407: Analysis of Categorical Data
Spring 2011

Haitao Chu, M.D., Ph.D.

University of Minnesota

More on log-linear models...

On model building:

- Brown's tests of association (not discussed) give large models to start backwards elimination from. BMDP implements these.
- Another approach is to try backward elimination from models with all higher k-way interactions (e.g. 3-way).
- G^2 is model deviance, the drop in $-2 \log \mathcal{L}$ from reduced model to saturated model; Agresti uses G^2 for model building.

9.3.1: Model Diagnostics

Let's consider $I \times J \times K$ tables for illustration. The ideas immediately generalize.
A table has observed cell counts n_{ijk} and predicted under the model $n\hat{\pi}_{ijk}$ where π_{ijk} is given by, e.g.,

$$\log(n\pi_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ},$$

for model $[XY][XZ]$. The ijk^{th} raw residual is $n_{ijk} - n\hat{\pi}_{ijk}$. A standardized version *based on Poisson sampling* is given by

$$e_{ijk} = \frac{n_{ijk} - n\hat{\pi}_{ijk}}{\sqrt{n\hat{\pi}_{ijk}}}.$$

The standardized Pearson residual is $r_{ijk} = e_{ijk}/\sqrt{1-\hat{h}_{ijk}}$. One can find cells for which $|r_{ijk}| > 3$ and flag them as being ill-fit, or simply compare the raw counts n_{ijk} to the fitted values $n\hat{\pi}_{ijk}$.

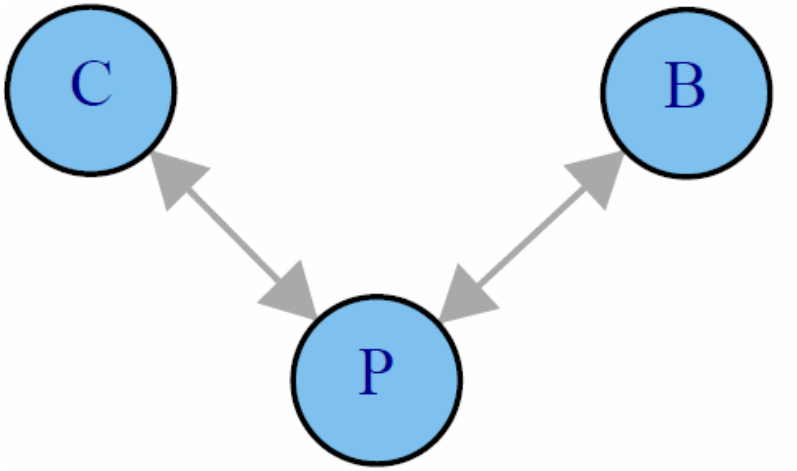
```
proc genmod order=data; class type chol bp;
 model count = type|chol type|bp / dist=poi link=log r;
```

Observation Statistics

Observation	Raw Residual	Pearson Residual	Deviance Residual	Std Deviance Residual	Std Pearson Residual	Likelihood Residual
1	1.5063291	0.0563535	0.0563337	0.3724586	0.3725894	0.3725864
2	−1.506329	−0.167882	−0.16841	−0.37376	−0.372589	−0.372827
3	−1.506329	−0.104318	−0.104444	−0.373039	−0.372589	−0.372625
4	1.5063291	0.3107738	0.3075386	0.3687106	0.3725894	0.3698952
5	5.0786106	0.1780138	0.1778292	1.4298604	1.431345	1.4313221
6	−5.078611	−0.598194	−0.605433	−1.448667	−1.431345	−1.434386
7	−5.078611	−0.3674	−0.369046	−1.437757	−1.431345	−1.431768
8	5.0786106	1.2346018	1.179489	1.3674495	1.431345	1.3840886

The `StReschi` have the r_{ijk}. All are within $|r_{ijk}| < 3$.

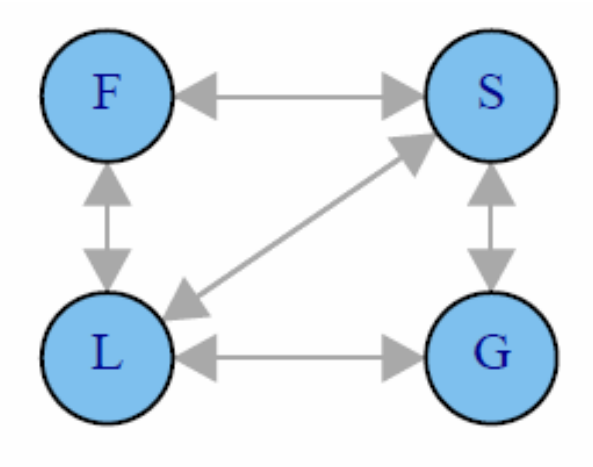
- An association graph plots each factor as a vertex and connects factors according to interaction terms in the log-linear model.
- Recall the the example that looked at personality type P, blood pressure B, and cholesterol C. We found the model $[PC][PB]$ fit. This has association graph:



- The two variables C and B are separated by P. All paths from C to B go through P. This implies that $C \perp B|P$.

From page 360: **Suppose that a model for a multiway table partitions variables into three mutually exclusive subsets A, B, and C such that B separates A and C. After collapsing the table over the variables in C, parameters relating to variables in A and parameters relating A to B are unchanged.** Also: $A \perp C|B$.

Alligator food example: the model $[GLS][SF][LF]$ fit the data. Then $A = \{G\}$, $C = \{F\}$ and $B = \{L, S\}$ from the association graph:



We can collapse the table over gender and examine associations among F, L, S without worrying about Simpson's paradox (recall we dropped gender from the model with food as the outcome). Also: $F \perp G | L, S$.
Example: Table 9.1 (p. 362). Five factors: M, C, A, G, R.
Model with all 10 3-factor interactions fits well with $G^2 = 5.3$ on 6 df p-value is 0.5. Reduced model with all 10 2-factor interactions also fits well with $G^2 = 15.3$ on 16 df and p-value is 0.5 (again).

```
data drug;
input g r a c m count @@;
datalines ;
0 1 1 1 1 405 0 1 1 1 0 268
0 1 1 0 1  13 0 1 1 0 0 218
0 1 0 1 1   1 0 1 0 1 0  17
0 1 0 0 1   1 0 1 0 0 0 117
1 1 1 1 1 453 1 1 1 1 0 228
1 1 1 0 1  28 1 1 1 0 0 201
1 1 0 1 1   1 1 1 0 1 0  17
1 1 0 0 1   1 1 1 0 0 0 133
0 0 1 1 1  23 0 0 1 1 0  23
0 0 1 0 1   2 0 0 1 0 0  19
0 0 0 1 1   0 0 0 0 1 0   1
0 0 0 0 1   0 0 0 0 0 0  12
1 0 1 1 1  30 1 0 1 1 0  19
1 0 1 0 1   1 1 0 1 0 0  18
1 0 0 1 1   1 1 0 0 1 0   8
1 0 0 0 1   0 1 0 0 0 0  17
;
```

```
proc genmod;
 class g r a c m;
 model count=g|r|a g|r|c g|r|m g|a|c g|a|m g|c|m r|a|c r|a|m r|c|m a|c|m
 / link=log dist=poi type3;
proc genmod;
 class g r a c m;
 model count=g|r g|a g|c g|m r|a r|c r|m a|c a|m c|m
 / link=log dist=poi type3;
```

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
g	1	5.98	0.0144
r	1	828.44	<.0001
g*r	1	0.84	0.3597
a	1	378.56	<.0001
g*a	1	3.38	0.0661
c	1	20.19	<.0001
g*c	1	0.98	0.3230
m	1	248.74	<.0001
g*m	1	9.82	0.0017
r*a	1	4.98	0.0256
r*c	1	0.44	0.5056
r*m	1	3.59	0.0582
a*c	1	185.86	<.0001
a*m	1	91.62	<.0001
c*m	1	498.13	<.0001

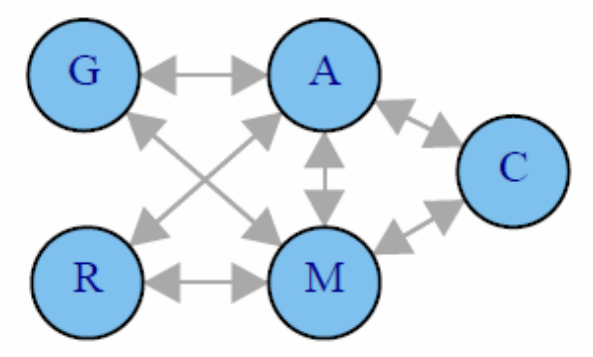
We can remove $[RC]$. Then $[GR]$. Then $[GC]$. (Not shown).

```
proc genmod;
 class g r a c m;
 model count=g|a g|m r|a r|m a|c a|m c|m / link=log dist=poi type3;
```

Source	DF	Chi-Square	Pr > ChiSq
g	1	6.20	0.0127
a	1	428.92	<.0001
g*a	1	5.51	0.0189
m	1	264.33	<.0001
g*m	1	8.90	0.0029
r	1	834.63	<.0001
r*a	1	4.78	0.0288
r*m	1	2.99	0.0836
c	1	25.49	<.0001
a*c	1	187.38	<.0001
a*m	1	92.05	<.0001
c*m	1	497.00	<.0001

The final model is $[GA][GM][RA][RM][AC][AM][CM]$. This model has $G^2 = 17.54$ on 19 df for a p-value of 0.55.

The association graph looks like:



- We see that $C \perp G \perp R | M, A$. For example, cigarette use is independent of gender given marijuana and alcohol use.
- What if we accept that `r*m` is not needed above ($p = 0.083$)? Then race is connected to G, M, and C only through alcohol. We would have $R \perp (G, M, C)|A$, i.e. $R \perp G|A$, $R \perp M|A$, and $R \perp C|A$.

8.2.3: $I \times J \times K$ table interpretation for $[XY][XZ][YZ]$

For $1 \leq i \leq I-1$ and $1 \leq j \leq J-1$ define

$$\theta_{ij(k)} = \frac{\pi_{i,j,k}\ \pi_{i+1,j+1,k}}{\pi_{i,j+1,k}\ \pi_{i+1,j,k}} = \frac{\left[\frac{P(Y=j,X=i|Z=k)}{P(Y=j+1,X=i|Z=k)}\right]}{\left[\frac{P(Y=j,X=i+1|Z=k)}{P(Y=j+1,X=i+1|Z=k)}\right]}.$$

There are $(I-1)(J-1)$ local odds ratios at each level of $Z=k$. This completely determines the dependence structure among $X, Y|Z=k$.
For model $[XY][XZ][YZ]$ we have

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

This implies

$$\log \theta_{ij(k)} = \lambda^{XY}_{i,j} + \lambda^{XY}_{i+1,j+1} - \lambda^{XY}_{i,j+1} - \lambda^{XY}_{i+1,j}.$$

So $\theta_{ij(1)} = \theta_{ij(2)} = \cdots = \theta_{ij(K)}$ for all i and j, the model of homogeneous association.

Similarly, $[XY][XZ][YZ]$ implies $\theta_{(1)jk} = \theta_{(2)jk} = \cdots = \theta_{(I)jk}$ for all j and k, and $\theta_{i(1)k} = \theta_{i(2)k} = \cdots = \theta_{i(J)k}$ for all i and k. This is the difference between $[XY][XZ][YZ]$ and the saturated model $[XYZ]$ in which there is no homogeneous association.

Section 8.5.3: $[XY][XZ][YZ]$ and logistic regression

Now let's say Y is the outcome and is dichotomous. Then

$$\begin{aligned}
\log \frac{P(Y=1|X=i,Z=k)}{P(Y=2|X=i,Z=k)} &= \log \frac{P(Y=1,X=i,Z=k)}{P(Y=2,X=i,Z=k)} \\
&= \log\ n\pi_{i1k} - \log\ n\pi_{i2k} \\
&= \left[\lambda + \lambda_i^X + \lambda_1^Y + \lambda_k^Z + \lambda_{i1}^{XY} + \lambda_{ik}^{XZ} + \lambda_{1k}^{YZ}\right] \\
&\quad - \left[\lambda + \lambda_i^X + \lambda_2^Y + \lambda_k^Z + \lambda_{i2}^{XY} + \lambda_{ik}^{XZ} + \lambda_{2k}^{YZ}\right] \\
&= \left[\lambda_1^Y - \lambda_2^Y\right] + \left[\lambda_{i1}^{XY} - \lambda_{i2}^{XY}\right] + \left[\lambda_{1k}^{YZ} - \lambda_{2k}^{YZ}\right] \\
&\equiv \beta_0 + \beta_i^X + \beta_k^Z,
\end{aligned}$$

which corresponds to an additive logistic regression model.

- If all's we care about is how (X, Z) relates to outcome Y, then logistic regression model is *okay*.
- If we are concerned with dependence structure among (X, Y, Z), then log-linear modeling is appropriate.
- Table 8.11 gives the equivalent logistic regression model to several log-linear models:

log-linear model	logit model with outcome Y
$[Y][XZ]$	logit $P(Y=1) = \alpha$
$[XY][XZ]$	logit $P(Y=1) = \beta_i^X$
$[YZ][XZ]$	logit $P(Y=1) = \beta_k^Z$
$[XY][XZ][YZ]$	logit $P(Y=1) = \beta_i^X + \beta_k^Z$
$[XYZ]$	logit $P(Y=1) = \beta_i^X + \beta_k^Z + \beta_{ik}^{XZ}$

- Question: where are $[X][Y][Z]$, $[X][YZ]$, $[Z][XY]$, and $[XY][YZ]$?

More on 'collapsibility'

Recall "personality type" data, which had three factors: P, C, and B. We decided $[PC][PB]$ fit the data.

- Fitting $[PC][PB]$ yields $\lambda_{11}^{PC} = -0.2176$ and $\lambda_{11}^{PB} = -0.2409$.
- Fitting $[PC]$, i.e. *collapsing over blood pressure*, yields $\lambda_{11}^{PC} = -0.2176$ (same as above).
- Fitting $[PB]$, i.e. *collapsing over cholesterol*, yields $\lambda_{11}^{PB} = -0.2409$ (same as above).
- In model $[PC][PB]$ we have

$$\theta_{11(k)} = \frac{P(P=1, C=1|B=k)P(P=2, C=2|B=k)}{P(P=1, C=2|B=k)P(P=2, C=1|B=k)}.$$

- In terms of the log-linear model parameters,

$$\log \theta_{11(k)} = \left[\lambda_{11}^{PC} + \lambda_{1k}^{PB}\right] + \left[\lambda_{22}^{PC} + \lambda_{2k}^{PB}\right] - \left[\lambda_{12}^{PC} + \lambda_{1k}^{PB}\right] - \left[\lambda_{21}^{PC} + \lambda_{2k}^{PB}\right] = \lambda_{11}^{PC},$$

  which is independent of k!
- This is because $\lambda_{12}^{PC} = \lambda_{21}^{PC} = \lambda_{22}^{PC} = 0$ for identifiability.
- So $\hat{\theta}_{11(k)} = e^{-0.2176} = 0.80$. The odds of having normal cholesterol is reduced 20% for personality type A (within each level of blood pressure).
- Collapsing over blood pressure yielding model $[PC]$ gives $\theta_{11} = \lambda_{11}^{PC}$ from the reduced model, which has *exactly the same outcome* $\hat{\theta}_{11} = 0.80$.
- As required by the collapsibility theorem, the marginal and conditional interpretations are the same. No information is lost by collapsing the table.

Seat belt example revisited

The final model was $[GLB][LBI][GI]$. Can we say *anything* succinctly here? Let's see how the gender/injury odds ratio changes with levels of location and belt use. Define

$$\theta_{11(kl)} = \frac{P(G=1, I=1|L=k, B=l)P(G=2, I=2|L=k, B=l)}{P(G=1, I=2|L=k, B=l)P(G=2, I=1|L=k, B=l)}.$$

In terms of log-linear model parameters,

$$\begin{aligned}
\log \theta_{11(kl)} &= \left[\lambda_{11l}^{GLB} + \lambda_{1kl}^{ILB} + \lambda_{11}^{GI}\right] + \left[\lambda_{21l}^{GLB} + \lambda_{2kl}^{ILB} + \lambda_{22}^{GI}\right] \\
&\quad - \left[\lambda_{11l}^{GLB} + \lambda_{2kl}^{ILB} + \lambda_{12}^{GI}\right] - \left[\lambda_{21l}^{GLB} + \lambda_{1kl}^{ILB} + \lambda_{21}^{GI}\right] \\
&= \lambda_{11}^{GI},
\end{aligned}$$

independent of $L = k$ and $B = l$, the model of *homogeneous association*.

What is the association graph for $[GLB][LBI][GI]$?

- From the output (last set of slides), $\hat{\theta}_{11(kl)} = e^{-0.5459} = 0.58$. The odds of not being injured for females is 0.58 times the odds for males within each (B, L) strata.
- Fitting the table *collapsed over* B and L, i.e. fitting $[GI]$, we obtain the *marginal* odds ratio $\hat{\theta}_{11} = e^{-0.4128} = 0.66$.
- The marginal interpretation is not the same (but not *that* different!) as the conditional interpretation. The conditions of the collapsibility theorem are not satisfied here, and so the interpretation changes upon collapsing the table.