Contingency tables having 3 and 4 dimensions
Aims of today’s lecture

In today’s lecture we apply Poisson regression to the analysis of contingency tables having 3 or 4 dimensions.

- Types of independence
- Mosaic plots for 3 and 4 factors
- Connection between independence and interactions for 3 and 4 dimensional tables
- Hierarchical models
- Graphical models

Reference: Coursebook, sections 5.3, 5.3.1
Example: the Florida murder data

- 675 convicted murderers in Florida, 1976-1977

- Classified by
  - Death penalty (n/y)
  - Victims race (black/white)
  - Defendants race (black/white)
Contingency table

<table>
<thead>
<tr>
<th>Race</th>
<th>Victim’s race</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>Black</td>
</tr>
<tr>
<td>Death Penalty</td>
<td>Yes</td>
</tr>
<tr>
<td>Black</td>
<td>6</td>
</tr>
<tr>
<td>Black</td>
<td>0</td>
</tr>
</tbody>
</table>
Getting the data into R

```r
> counts = c(6,0,97,9,11,19,52,132)
> murder.df = data.frame(counts=counts, expand.grid(Defendant =
+   DP = c("Yes","No"), Victim = c("B","W")))
>
> murder.df

<table>
<thead>
<tr>
<th>counts</th>
<th>Defendant</th>
<th>DP</th>
<th>Victim</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>B</td>
<td>Yes</td>
<td>B</td>
</tr>
<tr>
<td>0</td>
<td>W</td>
<td>Yes</td>
<td>B</td>
</tr>
<tr>
<td>97</td>
<td>B</td>
<td>No</td>
<td>B</td>
</tr>
<tr>
<td>9</td>
<td>W</td>
<td>No</td>
<td>B</td>
</tr>
<tr>
<td>11</td>
<td>B</td>
<td>Yes</td>
<td>W</td>
</tr>
<tr>
<td>19</td>
<td>W</td>
<td>Yes</td>
<td>W</td>
</tr>
<tr>
<td>52</td>
<td>B</td>
<td>No</td>
<td>W</td>
</tr>
<tr>
<td>132</td>
<td>W</td>
<td>No</td>
<td>W</td>
</tr>
</tbody>
</table>
```

Note use of `expand.grid` to generate all possible combinations of the factor levels (defendant varying most rapidly)
> murder.tab = array(counts, c(2,2,2),
    dimnames=list(Defendant = c("B","W"),
                  DP = c("Yes","No"), Victim = c("B","W")))
> murder.tab

, , Victim = B
   DP
Defendant Yes No
   B   6  97
   W   0  9

, , Victim = W
   DP
Defendant Yes No
   B  11  52
   W  19 132
Using `xtabs`

```r
> murder.tab = xtabs(counts ~ Defendant + DP + Victim, data = murder.df)
> murder.tab
, , Victim = B
   DP
Defendant Yes  No
  B   6  97
  W   0  9

, , Victim = W
   DP
Defendant Yes  No
  B  11  52
  W 19 132
```
Questions

- Is death penalty independent of race?
- What is the role of victim’s race?
- What does “independent” mean when we have three factors?
Types of independence

▶ Suppose we have 3 criteria (factors) $A$, $B$ and $C$

▶ In a multinomial sampling context, let

$$\pi_{ijk} = \Pr(A = i, B = j, C = k).$$

▶ Various forms of independence may be of interest. These can be expressed in terms of the probabilities $\pi_{ijk}$
Marginal probabilities

\[ \{ A = i, B = j \} = \bigcup_k \{ A = i, B = j, C = k \} \]

so

\[ \Pr(A = i, B = j, C = k) = \sum_k \Pr(A = i, B = j, C = k) \]

\[ = \sum_k \pi_{ijk} \]

\[ = \pi_{ij} + \]
Marginal probabilities (ii)

\{ A = i \} = \bigcup_k \{ A = i, B = j, C = k \}

So

\[ Pr(A = i) = \sum_j \sum_k Pr(A = i, B = j, C = k) \]

\[ = \sum_j \sum_k \pi_{ijk} \]

\[ = \pi_{i++} \]
All three factors independent

Expressed as

\[ \pi_{ijk} = Pr(A = i, B = j, C = k) = Pr(A = i) \times Pr(B = j) \times Pr(C = k) = \pi_{i++} \pi_{+j+} \pi_{++k} \]
One factor independent of the other two

Expressed as

\[ \pi_{ijk} = \Pr(A = i, B = j, C = k) = \Pr(A = i) \times \Pr(B = j, C = k) = \pi_{i++} \pi_{+jk} \]
Two factors conditionally independent, given a third

This means the two factors are independent in each of the subtables formed by fixing the level of the third factor e.g. Victim’s Race.

\[ \pi_{ijk} = \Pr(A = i, B = j \mid C = k) \]
\[ = \Pr(A = i \mid C = k) \times \Pr(B = j \mid C = k) \]

Equivalent to

\[ \pi_{ijk} = \frac{\pi_{i+k} \pi_{+jk}}{\pi_{++k}} \]
Parameterising tables of Poisson means with main effects and interactions

Recall (see Slides 54 of lecture 18) that in ordinary 3-way ANOVA we split up the table of means into main effects and interactions:

\[ \mu_{ijk} = \text{Int} + \alpha_i + \beta_j + \gamma_k + (\alpha \beta)_{ij} + (\beta \gamma)_{jk} + (\alpha \gamma)_{ik} + (\alpha \beta \gamma)_{ijk} \]

We can do exactly the same thing with the logs of the Poisson means in a 3-way table:

\[ \log \mu_{ijk} = \text{Int} + \alpha_i + \beta_j + \gamma_k + (\alpha \beta)_{ij} + (\beta \gamma)_{jk} + (\alpha \gamma)_{ik} + (\alpha \beta \gamma)_{ijk} \]
Parameterising tables of probabilities with main effects and interactions

Corresponding multinomial probabilities are given by the model

\[
\log \frac{\pi_{ijk}}{\pi_{111}} = \log \frac{\mu_{ijk}}{\mu_{111}} = \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\alpha\gamma)_{ik} + (\alpha\beta\gamma)_{ijk}
\]

Using these relationships we can work out the connection between the independence concepts and the interactions, like we did for the two-dimensional tables.
Parameterising tables of probabilities with main effects and interactions

The interactions are related to the different forms of independence:

- if all the interactions are zero, then the 3 factors are mutually independent;
- If the $ABC$ and $AB$ interactions are zero, then $A$ and $B$ are independent, given $C$;
- If the $ABC$, $AB$ and $AC$ interactions are zero, then $A$ is independent of $B$ and $C$. 
Summary of independence models

- All 3 factors independent in multinomial model
  - Equivalent to all interactions zero in Poisson model
    - Poisson Model is \( \text{count} \sim A + B + C \)

- \( A \) independent of \( B \) and \( C \)
  - Equivalent to all interactions between \( A \) and the others zero
    - Poisson Model is \( \text{count} \sim A + B + C + B : C \) or \( \text{count} \sim A + B \ast C \)
Factors $A$ and $B$ conditionally independent given $C$

- Equivalent to all interactions containing both $A$ and $B$ zero;

- Poisson Model is $\text{count} \sim A + B + C + A : C + B : C$ or $\text{count} \sim A \ast C + B \ast C$
Analysis strategy: Florida data

- We will fit some models to this data and investigate the pattern of independence/dependence between the factors.
- We fit a maximal model, and then investigate suitable submodels.
- We use, anova, AIC, stepwise
The Analysis: possible models

NB: ... = family=poisson, data=murder.df

model1 = glm(counts~Defendant+DP+Victim, ...)
model2 = glm(counts~Defendant*DP+Victim, ...)
model3 = glm(counts~Defendant+DP*Victim, ...)
model4 = glm(counts~Defendant*Victim + DP, ..)
model5 = glm(counts~Defendant*DP+Victim*DP, ...)
model6 = glm(counts~Defendant*DP+Victim*Defandent, ...)
model7 = glm(counts~Defendant*Victim+ DP*Victim, ...)
model8 = glm(counts~(Defandent+DP+Victim)^2, ...)
model9 = glm(counts~Dafendent*DP*Victim, ...)
model.list = list(model1, model2, model3, model4, model5, 
                   model6, model7, model8, model9)
result = matrix(0, 9, 4)
for (i in 1:8) {
  result[i, 1] = model.list[[i]]$deviance
  result[i, 2] = model.list[[i]]$df.residual
  result[i, 3] = 1 - pchisq(result[i, 1], result[i, 2])
  result[i, 4] = model.list[[i]]$aic
}
dimnames(result) = list(paste("Model", 1:9), c("Deviance", "df", "p-val", "AIC"))
Models

> round(result,4)

<table>
<thead>
<tr>
<th>Model</th>
<th>Deviance</th>
<th>df</th>
<th>p-val</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>137.9294</td>
<td>4</td>
<td>0.0000</td>
<td>181.6109</td>
</tr>
<tr>
<td>Model 2</td>
<td>137.7079</td>
<td>3</td>
<td>0.0000</td>
<td>183.3895</td>
</tr>
<tr>
<td>Model 3</td>
<td>131.6796</td>
<td>3</td>
<td>0.0000</td>
<td>177.3612</td>
</tr>
<tr>
<td>Model 4</td>
<td>8.1316</td>
<td>3</td>
<td>0.0434</td>
<td>53.8132</td>
</tr>
<tr>
<td>Model 5</td>
<td>131.4582</td>
<td>2</td>
<td>0.0000</td>
<td>179.1398</td>
</tr>
<tr>
<td>Model 6</td>
<td>7.9102</td>
<td>2</td>
<td>0.0192</td>
<td>55.5917</td>
</tr>
<tr>
<td>Model 7</td>
<td>1.8819</td>
<td>2</td>
<td>0.3903</td>
<td>49.5635</td>
</tr>
<tr>
<td>Model 8</td>
<td>0.7007</td>
<td>1</td>
<td>0.4025</td>
<td>50.3823</td>
</tr>
<tr>
<td>Model 9</td>
<td>0.0000</td>
<td>0</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Looks like Model 7 is good: Defendant’s race and DP independent given victim’s race.
Testing for a specific form of independence

- Fit the model corresponding to the form of independence being tested.
- Get the residual deviance
- Calculate a p-value.

e.g. to see if all three factors are independent: fit the model
counts \sim \text{Defendant} + \text{DP} + \text{Victim}, and get the deviance and p-value:

```r
> model1 = glm(counts~Defendant+DP+Victim, family=poisson, 
data=murder.df)
> summary(model1)
...
Residual deviance: 137.93 on 4 degrees of freedom
> 1-pchisq(137.93,4)
[1] 0
```
p-value very small, so very strong evidence against the independence model
The Copenhagen housing data

317 apartment residents in Copenhagen were surveyed on their housing. Three variables were measured:

sat: satisfaction with housing (Low, medium, high)

cont: amount of contact with other residents (Low, high)

infl: influence on management decisions (Low, medium, high)
The data

<table>
<thead>
<tr>
<th>sat</th>
<th>infl</th>
<th>cont</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>61</td>
</tr>
<tr>
<td>Medium</td>
<td>Low</td>
<td>Low</td>
<td>23</td>
</tr>
<tr>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>17</td>
</tr>
<tr>
<td>Low</td>
<td>Medium</td>
<td>Low</td>
<td>43</td>
</tr>
<tr>
<td>Medium</td>
<td>Medium</td>
<td>Low</td>
<td>35</td>
</tr>
<tr>
<td>High</td>
<td>Medium</td>
<td>Low</td>
<td>40</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>26</td>
</tr>
<tr>
<td>Medium</td>
<td>High</td>
<td>Low</td>
<td>18</td>
</tr>
<tr>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>54</td>
</tr>
</tbody>
</table>

9 more lines...
# Tabular form

<table>
<thead>
<tr>
<th>Influence = LOW</th>
<th>Cont</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sat</td>
<td>Low</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>78</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>Low</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>46</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Influence = MED</th>
<th>Cont</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sat</td>
<td>Low</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>48</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>Low</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>45</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Influence = HIGH</th>
<th>Cont</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sat</td>
<td>Low</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>15</td>
</tr>
<tr>
<td>Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>Low</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>45</td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Page 27/41
The analysis

> housing(glm <- glm(count ~ sat * infl * cont, family = poisson, data = housing.df)
> anova(housing(glm, test = "Chisq")

Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL 17 166.757
sat 2 26.191 15 140.566 2.054e-06
infl 2 20.040 13 120.526 4.451e-05
cont 1 22.544 12 97.983 2.054e-06
sat:infl 4 75.577 8 22.406 1.504e-15
sat:cont 2 7.745 6 14.661 0.021
infl:cont 2 11.986 4 2.675 0.002
sat:infl:cont 4 2.675 0 9.546e-15 0.614

This time, only the 3-factor interaction is insignificant. This is the “homogeneous association model”.
The Homogeneous Association Model

Association between two factors is measured by sets of odds ratios

$$OR_{ij} = \frac{\pi_{ij} \pi_{11}}{\pi_{i1} \pi_{1j}}.$$
## Copenhagen conditional OR’s

### Influence = LOW

<table>
<thead>
<tr>
<th>Cont</th>
<th>Sat</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>*</td>
<td>1.56</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>*</td>
<td>1.97</td>
<td></td>
</tr>
</tbody>
</table>

### Influence = MED

<table>
<thead>
<tr>
<th>Cont</th>
<th>Sat</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>*</td>
<td>1.15</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>*</td>
<td>1.92</td>
<td></td>
</tr>
</tbody>
</table>

### Influence = HIGH

<table>
<thead>
<tr>
<th>Cont</th>
<th>Sat</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>*</td>
<td>2.40</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>*</td>
<td>1.99</td>
<td></td>
</tr>
</tbody>
</table>
Calculating the conditional OR’s

<table>
<thead>
<tr>
<th>Influence = <strong>HIGH</strong>: OR’s</th>
<th>Cont</th>
<th>Sat</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>*</td>
<td>2.40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>*</td>
<td>1.99</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Influence = <strong>MED</strong>: Counts</th>
<th>Cont</th>
<th>Sat</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>25</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>18</td>
<td>25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>54</td>
<td>62</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
2.40 = \frac{26 \times 25}{18 \times 15} \\
1.99 = \frac{26 \times 62}{54 \times 15}
\]
For the homogeneous association model, the conditional odds ratios for $A$ and $B$ (i.e. using the conditional distributions of $A$ and $B$ given $C = k$) do not depend on $k$. That is, the pattern of association between $A$ and $B$ is the same for all levels of $C$.

The common values of the conditional Log OR’s in the separate $A \times B$ tables are estimated by the $A : B$ interactions.
Estimating the conditional OR’s

The estimated Sat-cont conditional log ORs are estimated with Sat-cont interactions in homogeneous association model:

```r
homogen.glm<-glm(count~sat*infl*cont-sat:infl:cont, family=poisson, data=housing.df)
> summary(homogen.glm)

Coefficients:

            Estimate Std. Error z value Pr(>|z|)     
satMedium:contHigh  0.4157   0.1948  2.134  0.03281 *
satHigh:contHigh   0.6496   0.1823  3.563  0.00037 ***
```

Page 33/41
Estimating the conditional OR’s (cont)

The estimated Sat-cont conditional log ORs are estimated with Sat-cont interactions in homogeneous association model:

```
> 0.4157 + c(-1,1)*1.96*0.1948
[1] 0.033892 0.797508
> exp(0.4157 + c(-1,1)*1.96*0.1948)
[1] 1.034473 2.220002
> exp(0.4157)
[1] 1.515431
> exp(0.6496)
[1] 1.914775
```

For the OR for Satisfaction medium/contact High OR, conditional on Influence, the estimate for the log OR is 0.4157, std error is 0.1948. The odds ratio is \(\exp(0.4157) = 1.52\) and the CI is \(\exp(0.4157 \pm 1.96 \times 0.1948)\) i.e. \((1.034, 2.220)\). The estimated OR for Satisfaction High/contact High is \(\exp(0.6496) = 1.914\). Thus, the odds of satisfaction = Low (vs satisfaction = Med) is about 50% higher for contact = Low than it is for contact = High.
Similar results apply for 4-dimensional tables.

For example, models for 4 factors A, B, C and D:

- A, B, C and D all independent: $A + B + C + D$
- A, B independent of C and D: $A*B + C*D$
- D conditionally independent of C, given A and B: $A*B*C + A*B*D$
Hierarchical Models

- We will assume that all models are hierarchical: if the model includes an interaction with factors $A_1, \ldots A_k$, then it includes all main effects and interactions that can be formed from $A_1, \ldots A_k$.

- We can represent hierarchical models by listing these “maximal interactions“.
Examples

- 2 factor model $A + B + A:B$ is hierarchical
  Hierarchical notation: [AB]

- 3 factor model $A + B + C + A:B$ is hierarchical
  Hierarchical notation: [AB][C]

- 3 factor model $A + B + C + A:B + A:C$ is hierarchical
  Hierarchical notation: [AB][AC]

- 3 factor model $A + B + C + A:B + A:B:C$ is not hierarchical (Doesn’t contain B:C or A:C)
Graphical Models

A way of visualising independence patterns:

- A subset of hierarchical models
- Each factor represented by the vertex of an “association” graph
- Two vertices connected by edges if they have a non-zero interaction
- Then
  - A vertex not connected to any other vertex is independent of the other vertices
  - Two vertices not directly connected are conditionally independent, given the connecting vertices
Examples

- **2 factor model**
  \[ A + B + A:B \text{ or } [AB] \]

- **3 factor model**
  \[ A + B + C + AB \text{ or } [AB][C] \]
  C independent of A and B

- **3 factor model**
  \[ A + B + C + A:B + A:C + B:C \text{ or } [AB][AC][BC] \]
More examples

- **3 factor model**
  \[A + B + C + AB + AC \text{ or } [AB][AC]\]

- **4 factor model**
  \[A + B + C + D + A:B + A:C + B:C \text{ or } [AB][AC][BC][D]\]
More examples

4 factor model
A + B + C + D + A:B + B:C + A:D
or [AB][BC][AD]

C and D are conditionally independent given A and B